Analyzing Instant Messaging Writeprints as a Behavioral Biometric Element of Cybercrime Investigations

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Angela Orebaugh
Masters of Science
James Madison University, 1999

Co-Director: Dr. Jeremy Allnutt, Professor
Department of Electrical and Computer Engineering
Volgenau School of Engineering

Co-Director: Dr. Jason Kinser, Associate Professor
School of Physics, Astronomy, and Computational Sciences

Spring Semester 2014
George Mason University
Fairfax, VA

DEDICATION

This is dedicated to those who pursue education and life enrichment. The road to success is paved with perseverance. A special dedication to Tammy Wilt for her unending supply of optimism and encouragement.

ACKNOWLEDGEMENTS

I would like to thank the many people who have supported and motivated me during my research. Foremost I would like to thank my committee for supporting me through this process. They changed the outcome of my life and I aspire to pay it forward to support others and share my knowledge and experience. A special thanks to Dr. Jeremy Allnutt for his resilience and unwavering support to help me continue moving forward in the process. A special thanks to Dr. Jason Kinser for spending countless hours with me as I processed and analyzed data and for teaching me the joys of Python. In addition to my committee, Dr. John Cordani and Dr. Bob Bordley graciously provided their subject matter expertise and supported me through this effort. Dr. Stephen Nash, Dr. Aleksandar Lazarevich, and Dr. Carol Chaski shared wisdom and words of encouragement throughout this process. A special thanks to Lisa Nolder for answering my stream of questions and offering guidance. I would not have succeeded without the support of those who have assisted me along the journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND DEFINITIONS

AIM	AOL Instant Messenger
BBS	Bulletin Board System
BF	Baseline Feature Set
CMC	Computer-Mediated Communication
CSV	Comma-Separated Value
EF	Extended Feature Set
EOL	End-Of-Line
Function words	A word (as a preposition, auxiliary verb, or conjunction) expressing primarily grammatical relationship [Source: http://www.merriam-webster.com/dictionary/function+word]
ICQ	I Seek You
IM	Instant Messaging
IRC	Internet Relay Chat
K-L	Karhunen-Loeve. KL transform is a supervised form of PCA that allows inclusion of class information in the transformation process.
Malware	A program that is inserted into a system, usually covertly, with the intent of compromising the confidentiality, integrity, or availability of the victim's data, applications, or operating system or of otherwise annoying or disrupting the victim [Source: NIST2011].
MITM	Man-In-the-Middle is an attack on the authentication protocol run in which the attacker positions himself in between the claimant and verifier so that he can intercept and alter data traveling between them [Source: NIST2011].
Masquerading	When an unauthorized agent claims the identity of another agent it is said to be masquerading [Source: NIST2011].
MGD	Multivariate Gaussian Distribution

N-grams	A sequence of *N* units, of tokens, of text, where those units are typically single characters or strings that are delimited by spaces [Source: BP2003].
PCA	Principal Component Analysis. PCA is an unsupervised statistical technique that captures variance across a large number of features in a reduced dimensionality.
Phishing	Tricking individuals into disclosing sensitive personal information through deceptive computer-based means [Source: NIST2011].
Social Engineering	An attempt to trick someone into revealing information (e.g., a password) that can be used to attack systems or networks [Source: NIST2011].
Standard Deviation	The standard deviation measures the spread of a distribution of a set of data [Source: http://docs.scipy.org/doc/numpy/reference/generated/numpy.std.html].
SVM	Support Vector Machine. SVM is a supervised learning model that constructs a hyperplane or set of hyperplanes to perform data classification. The hyperplanes separate sets of objects having different class memberships.
Trojan horse	A non-self-replicating program that seems to have a useful purpose, but in reality has a different, malicious purpose [Source: NIST2011].
Virus	A self-replicating program that runs and spreads by modifying other programs or files [Source: NIST2011].
Vocabulary richness	The ratio between the number of different words and the total number of words within a document [Source: ESK2007].
Worm	A self-replicating, self-propagating, self-contained program that uses networking mechanisms to spread itself. [Source: NIST2011].

ABSTRACT

ANALYZING INSTANT MESSAGING WRITEPRINTS AS A BEHAVIORAL BIOMETRIC ELEMENT OF CYBERCRIME INVESTIGATIONS

Angela Orebaugh, Ph.D.

George Mason University, 2014

Dissertation Co-Director: Dr. Jeremy Allnutt
Dissertation Co-Director: Dr. Jason Kinser

The anonymous nature of the Internet and increasing cybercrime creates a growing need for techniques to assist in identifying online cybercrime suspects as part of the investigation. In instant messaging (IM) communications, cyber criminals use virtual identities to hide their true identity, which hinders social accountability and facilitates cybercrime. Cyber criminals may use multiple screen names, impersonate other users, or supply false information on their virtual identities with the intention of deceiving unsuspecting victims and committing various cybercrimes. It is necessary to have IM cyber forensics techniques to assist in identifying cyber criminals and collecting digital evidence as part of the investigation.

Behavioral biometrics are persistent personal traits and patterns of behavior that may be collected and analyzed to aid a cybercrime investigation. Instant messaging

behavioral biometrics include online writing habits, known as stylometric features, which may be used to create an author writeprint to assist in identifying an author, or characteristics of an author, of a set of instant messages. The writeprint is a digital fingerprint that represents an author's distinguishing stylometric features that occur in his/her computer-mediated communications. Writeprints can provide cybercrime investigators a unique tool for analyzing IM-assisted cybercrimes. They may be used as input to a criminal cyberprofile and as an element of a multimodal system for cybercrime investigations. Writeprints can be used in conjunction with other evidence, investigation techniques, and biometrics techniques to reduce the potential suspect space to a certain subset of suspects; identify the most plausible author of an IM conversation from a group of suspects; link related crimes; develop an interview and interrogation strategy; and gather convincing digital evidence to justify search and seizure and provide probable cause.

The purpose of this dissertation is to create and analyze behavioral biometrics-based instant messaging writeprints as cyber forensics input for cybercrime investigations. This research uses authorship analysis techniques to create a set of stylometric features robust enough to show separation between authors and between author categories. The real time, casual nature of instant messaging communications offers several interesting stylometric features such as message structure, unusual language usage, and special stylistic markers that are useful in forming a suitable writeprint feature set for authorship analysis. This research uses the Principal Component Analysis (PCA) and multivariate Gaussian distribution (MGD) methods to

analyze IM conversation logs from two distinct data sets for authorship identification and characterization. Authorship identification may be applied to IM to assist in identifying criminals who hide their true identity or impersonate a known individual. Authorship characterization may be used to help discover IM cyber criminals who supply false information in their virtual identities, such as gender.

The research results presented in this dissertation demonstrate that authors show separation via IM writeprints. IM writeprints are shown to group messages belonging to a particular author from a set of authors and to group messages belonging to a particular author category from a set of author categories. By demonstrating high authorship identification and characterization probability, the research results presented in this dissertation indicate a promising future for applying authorship analysis as an element of a multimodal biometrics system to assist with cyber forensics and cybercrime investigations.

1. INTRODUCTION

The Internet has evolved from a resource of simple information sharing and exchange to a conglomeration of virtual communications and e-commerce activities. One increasingly popular use of the Internet is computer-mediated communication (CMC). CMC includes any communicative transaction, which occurs through the use of two or more networked computers [McQ2010]. CMC includes the use of asynchronous or synchronous online textual messages [Her2002]. Asynchronous CMC occurs in delayed time and does not require simultaneous participation of users, such as e-mail, web forums, newsgroups, and weblogs. Synchronous CMC occurs in real time and requires the simultaneous participation of users, such as instant messaging (IM) and chat rooms.

CMC may also be multicast or point-to-point. Multicast CMC is online text intended for multiple recipients, for example group chat rooms. Point-to-point CMC is online text intended for a single recipient, for example instant messaging. This dissertation is focused on the analysis of instant messaging, a synchronous form of point-to-point CMC with the following characteristics:

- Message authors use a virtual identity.

- The recipient of the messages is a single individual[1].
- Messages contain a style and vocabulary unique to IM [KCAC2008].

CMC generates large amounts of textual data, providing interesting research opportunities for analyzing such data. CMC is unique in that it is often referred to as *written speech*. Its informal nature contains many stylistic differences from literary texts including word usage, spelling and grammar errors, lack of punctuation, and abbreviations. Instant messaging's unique characteristics and stylistic differences distinguish it from other types of literary texts as well as other types of online communications, making it an especially interesting research area.

1.1 Instant Messaging Architecture

Today's instant messaging services grew from their online chat medium predecessor known as Internet Relay Chat (IRC). Unlike e-mail, IM provides a user the ability to view the current online status of other users and interact with active users in real time. Most instant messaging occurs over the public Internet, but more organizations are now implementing internal messaging servers. Examples of IM services include AOL Instant Messenger (AIM), I Seek You (ICQ), Skype, MSN Messenger, Google Talk, and Yahoo! Messenger. CMC differs from Short Message Service (SMS), in that users communicate using their cell phone to send short text messages to other cell phone users. Some IM services support Mobile Instant Messaging (MIM), which allows IM users to forward messages from the IM service to their cell phone as text messages [Cro2008].

[1] Instant messaging may also be used for multicast group chat sessions with multiple individuals. Group chat is beyond the scope of this research.

IM also continues to be incorporated into a number of other technologies, including game systems such as the Xbox or PlayStation.

The instant messaging architecture consists of a network, clients, and servers. Most IM networks use a client-server model in which a service provider maintains the server. Users register themselves with the service provider and download a compatible client for use on the service provider network. By registering, the user creates an account that consists of a unique identifier such as a name or number, also called a *screen name*. The screen name and its associated information become the user's virtual identity. The user provides his/her screen name to other users with whom he or she wants to communicate via the instant messaging network. Users authenticate to the central server using their screen name and password entered into the client to begin conversing with other users of the network commonly known as *buddies* or *friends*. Buddies are added to and maintained in a *Buddy List*, which shows when users are logged on for communication. Users can add, remove, and block buddies in the Buddy List. Users may communicate with a single buddy privately or several buddies in a group setting. Users communicate via an interactive window that displays the conversation as it occurs. When two authenticated users (e.g., Alice and Bob) want to communicate, the following sequence occurs:

- Alice creates a message to send to Bob.
- Alice's IM client creates a packet containing the message and sends it to the server.

- The server looks at the packet and determines that the recipient is Bob.
- The server creates a new packet with the message from Alice and sends it to Bob.
- Bob's IM client receives the packet containing the message and presents the message to Bob [Hin2003].

Some instant messaging services will continue to send all subsequent messages via the central server. However, some IM services create a direct connection between the user's clients after the first message [Hin2003]. The use of a central server is beneficial in the following ways:

- A user only needs to know the unique identifier of a buddy to communicate with him/her.
- Some IM servers allow users to send messages to a buddy even if he/she is not online. The server will store the messages until the buddy authenticates with the server, at which time they are sent to him/her [Hin2003].
- A central server may log all conversations between users.

Popular IM service provider networks include AOL, Google, Yahoo, and MSN. Each of these networks provides a compatible client for communication. However, each service provider currently uses a different protocol, making them incompatible with each other. For example, an AOL IM user can only communicate with other AOL IM users. Some clients, such as Trillian and Pidgin, can connect to multiple service provider networks at one time. However, the user must maintain a registered account on each of

the service provider networks that he/she wishes to use.

IM clients each contain a variety of features including multimedia, customization, and logging. Some IM clients allow users to transmit files and to stream audio and video. Users may also create and display customized status messages, letting buddies know when they are away or busy. Most clients allow logging of IM conversations. IM conversations are logged in a simple text format, making it easy for a researcher to parse and analyze conversation data. Logged conversations may also be used as evidence during an investigation. This dissertation uses two distinct datasets of IM conversation logs for analysis.

1.2 Instant Messaging and Cybercrime

Cybercrime could include any criminal activity that is committed with the aid of a computer or communication device in a network, such as the Internet, telephone lines, or mobile networks such as cellular communication [FM2008]. Instant messaging's anonymity hinders social accountability and leads to IM-assisted cybercrime facilitated by the following:

- Users can create any virtual identity.
- Users can log in from anywhere.
- Files can be transmitted.
- Communication is often transmitted unencrypted.

In IM communications, cyber criminals use virtual identities to hide their true identity. They can use multiple screen names or impersonate other users with the intention of harassing or deceiving unsuspecting victims. Cyber criminals may also supply false information on their virtual identities, for example a male user may configure his virtual identity to appear as female. Since most IM systems use the public Internet, the risk is high that usernames and passwords may be intercepted, or an attacker may hijack a connection or launch a *man-in-the-middle* (MITM) attack [Ore2004a, Ore2006a]. With hijacking and MITM attacks, the victim user thinks he/she is communicating with a buddy but is really communicating with the attacker *masquerading* as the victim's buddy [Ore2005a, OBB2005b]. Instant messaging's anonymity allows cyber criminals such as pedophiles, scam artists, and stalkers to make contact with their victims and get to know those they target for their crimes [Cro2008]. IM-assisted cybercrimes, such as *phishing*, *social engineering*, threatening, cyber bullying, hate speech and crimes, child exploitation, sexual harassment, and illegal sales and distribution of software are continuing to increase [MD2000]. Additionally, criminals such as terrorist groups, gangs, and cyber intruders use IM to communicate [AC2005]. Criminals also use IM to transmit *worms*, *viruses*, *Trojan horses*, and other *malware* over the Internet.

With increasing IM cybercrime, there is a growing need for techniques to assist in identifying online cybercrime suspects as part of the investigation [AC2006]. Cyber forensics is the application of investigation and analysis techniques to gather evidence suitable for presentation in a court of law with the goal of discovering the crime that took

place and who was responsible [BBO2006]. With IM communications, it is necessary to have cyber forensics techniques to assist in determining the IM user's real identity and collect digital evidence for investigators and law enforcement [OA2009a, OA2009b, OA2010b]. This dissertation explores the cyber forensic technique of behavioral biometrics to assist in identifying cyber criminals and collecting data for the investigation.

1.2.1 Behavioral Biometrics

In traditional forensic science, fingerprints uniquely identify individuals. With the absence of physical fingerprints, the anonymous nature of the Internet and use of virtual identities presents a critical challenge for cybercrime investigation. Fortunately, individuals often leave behind textual identity traces in cyberspace, which may be used to aid a cybercrime investigation [AC2008]. Determining an IM user's real identity relies on the fact that humans are creatures of habit and have certain persistent personal traits and patterns of behavior, known as behavioral biometrics [Rev2008].

Behavioral biometrics are measurable traits that are acquired over time (versus a physiological characteristic or physical trait) that can be used to recognize or verify the identity of a person [Bio2006]. For example, handwriting style is consistent throughout a person's life, even though it may vary with age. As with handwriting, users have certain online writing habits that are unconscious and deeply ingrained [TLML2004]. Even if a cyber criminal made a conscious effort to disguise his/her style, disguise would be difficult to achieve. Online writing habits, known as stylometric features, include

composition syntax and layout, vocabulary patterns, unique language usage, and other stylistic traits. Writer's have certain stylometric features that remain consistent across multiple texts of a given author, but differ in texts of different authors. Thus, certain stylometric features may be used to create an author writeprint to help identify an author of a particular piece of work [DACM2001b].

A writeprint represents an author's distinguishing stylometric features that occur in his/her computer-mediated communications. These stylometric features may include average word length, use of punctuation and special characters, use of abbreviations, and other stylistic traits. The principal challenge with writeprint analysis is the creation of a set of stylometric features robust enough to show separation between various authors.

Writeprints can provide cybercrime investigators a unique behavioral biometric tool for analyzing IM-assisted cybercrimes. Writeprints can be used as input to a cyberprofile and as an element of a multimodal system to perform cyber forensics and cybercrime investigations [JRS2004, RLG2009]. This dissertation uses authorship analysis techniques to create an author's IM writeprint based on behavioral biometrics.

1.2.2 Authorship Analysis

Authorship analysis is the process of examining the stylometric features of a document to identify or validate the text's author, or information about the author. Authorship analysis uses a variety of computer-aided statistical methods to analyze text. Authorship analysis techniques include authorship identification – methods to determine the most plausible author of a piece of text, and authorship characterization – methods to

infer an author's background characteristics based sociolinguistic attributes such as gender, age, educational background, linguistic background, ethnicity, and psychological status. Authorship identification may be applied to IM to assist in identifying cyber criminals who hide their true identity or impersonate a known individual. Authorship characterization may be used to help discover IM cyber criminals who supply false information in their virtual identities, such as gender. Authorship analysis uses a variety of stylometric features that can be derived from a particular piece of work to facilitate authorship identification or characterization.

Instant messaging communications contain several stylometric features for authorship analysis research. Certain IM specific features such as message structure, unusual language usage, and special stylistic markers are useful in forming a suitable writeprint feature set for authorship analysis [ZLCH2006]. The style of IM messages is very different than that of any other text used in traditional literature or other forms of computer-mediated communication. The continuous nature of synchronous communication makes it especially interesting since authors take less time to craft their responses [HPR2003]. The real time, casual nature of IM messages produces text that is conversational in style and reflects the author's true writing style and vocabulary [KCAC2008]. Significant characteristics of IM are the use of special linguistic elements such as abbreviations, and computer and Internet terms, known as netlingo. The textual nature of IM also creates a need to exhibit emotions. Emotion icons, called emoticons, are sequences of punctuation marks commonly used to represent feelings within computer-mediated text [KCAC2008]. An author's IM writeprint may be derived from

network packet captures or application data logged during an instant messaging conversation. Although some types of digital evidence, such as source IP addresses, file timestamps, and metadata may be easily manipulated, author writeprints based on behavioral biometrics are unique to an individual and difficult to imitate [DACM2001b]. This dissertation uses the data obtained from two unique datasets of synchronous, point-to-point instant messaging logs.

The nature of IM conversations also include several characteristics that make authorship analysis challenging:

- IM text is typically brief.
- The short length of online messages may cause some identifying features in normal texts, such as *vocabulary richness*, to be ineffective [ZLCH2006].
- Conversation beginning and end is difficult to determine.
- Messages often have spelling errors and do not follow formal grammar and structure standards.
- Authors' style can evolve over time (authors learn new emoticons, abbreviations, netlingo, etc.).
- Messages contain free-form unstructured text with few sentences or paragraphs.
- Unlike e-mail, IM has no standard header, greetings or farewells, or signatures.

This dissertation uses authorship analysis and statistical techniques to create and analyze IM writeprints to assist in identifying an author, as well as certain characteristics of the author of a set of IM messages. Thus far, the research community has begun to use behavioral biometrics-based authorship analysis techniques as a cyber forensics tool with recent application to e-mail and online forums.

1.3 Proposed Research

The purpose of this research is to create and analyze behavioral biometrics-based instant messaging writeprints to assist in identifying online cyber criminals and collecting digital evidence as part of the investigation. The research uses authorship analysis techniques to create an IM-specific stylometric feature set taxonomy to show separation between authors and between author categories. The research uses Principal Component Analysis (PCA) and multivariate Gaussian distribution (MGD) statistical methods to analyze author writeprints from IM conversation logs from two distinct datasets for authorship identification and characterization.

In the context of instant messaging, the goals of this research are the following:

1. Create writeprints that show separation between authors and author categories,
2. Create writeprints that can differentiate messages belonging to a particular author A_i from a set of authors $\{A_1,\ldots,A_n\}$, and
3. Create writeprints that can differentiate messages belonging to a particular author category C_i, from a set of author categories $\{C_1,\ldots,C_m\}$ based on sociolinguistic attributes.

1.4 Scope and Delimitations

This research creates author writeprints from IM conversations from two unique datasets of synchronous, point-to-point instant messaging logs. Based on the two datasets used, the scope of this research is the following:

1. Create an IM feature set taxonomy,
2. Using PCA, reduce the dimensions and show separation in author and author category writeprints,
3. Using MGD, determine the authorship identification and characterization probability of authors and author categories in the two datasets, and
4. Determine the conversation size that provides the highest probability for authorship identification and characterization based on the number of authors in the given test set.

Authorship analysis identification results greater than 70% are acceptable during an investigation process [IBFD2013].

IM writeprints are not proposed to be the sole method of determining authorship identification and characterization in cybercrime investigations. IM writeprints may be used as an element in a multimodal biometrics systems in conjunction with traditional investigation techniques to assist with cybercrime decision support.

1.5 Assumptions

For the purpose of this research the data used in the datasets has been manually cleaned of noise (i.e. quoted text, timestamps, usernames, automatic away message responses, and other metadata) during pre-processing. The scope of this research is bound by the following assumptions:

1. In the utilized datasets, the data for each author contains text from only that author between known pairs of communicators. Authorship has been validated to the greatest extent possible during data collection and via chain of custody. The IM conversations are point-to-point with no known third party interference. There may be limitations through unidentified noise.
2. The messages were not purposely crafted to imitating another user. To all extent possible there is no known intentional masquerading of other authors.

1.6 Summary of Contributions

As cybercrimes continue to increase, new cyber forensics techniques are needed to combat the constant challenge of Internet anonymity [Ore2006c]. Current IM products are not addressing the anonymity and ease of impersonation over instant messaging. Cyber forensic techniques are needed to assist cybercrime decision support tools in collecting and analyzing digital evidence, discovering characteristics about the cyber criminal, and assisting in identifying cyber criminal suspects.

The analysis of IM author writeprints in this research provides a foundation for using behavioral biometrics as a cyber forensics element of criminal investigations. The principal contributions of this dissertation are:

1. A process for creating and analyzing IM behavioral biometrics-based writeprints to assist with cybercrime investigations.
2. An IM-specific stylometric taxonomy for IM writeprints and authorship analysis.
3. A forensically feasible ground-truth dataset.

IM writeprints can be used in conjunction with other evidence, investigation techniques, and biometrics techniques to reduce the potential suspect space to a certain subset of suspects; identify the most plausible author of an IM conversation from a group of suspects; link related crimes; develop an interview and interrogation strategy; and gather convincing digital evidence to justify search and seizure and provide probable cause.

1.7 Organization of the Dissertation

The rest of the dissertation is organized as follows:

- Chapter 2: Provides a literature review and work related to this research in the areas of authorship analysis, forensics, behavioral biometrics, and cybercrime investigations.

- Chapter 3: Describes the research methodology, including the stylometric feature set taxonomy, data pre-processing, description of the datasets, and data analysis methods and tools.
- Chapter 4: Provides a detailed analysis of the IM writeprint results.
- Chapter 5: Provides a summary and interpretation of the results, with respect to the original research goals, and suggestions for future work.

2. RELATED WORK AND TECHNOLOGIES

Historically, authorship analysis has been extensively applied to literature and published articles. More recently, the research community has begun to use behavioral biometrics-based authorship analysis techniques for CMC with recent application to e-mail, chat, and online forums. A large research gap exists in applying authorship analysis techniques to instant messaging communications to facilitate learning the author identity or author characteristics in support of cybercrime investigations. Preliminary journal articles and conference presentations [Ore2006b, OA2009a, OA2009b, OA2010b] from this research are the only comprehensive examination of IM authorship analysis. Gaussian distributions have been applied in various biometric and forensic research efforts including facial recognition and tracking, voice recognition, signature verification, and characterizing social media authors. However, no research to date has used Gaussian distributions to perform IM authorship analysis.

This chapter provides a brief history of authorship analysis and an overview of several related works for CMC authorship analysis studies. It provides an overview of several related works for biometric-related Gaussian studies. This chapter also provides an overview of the criminal investigation process and the use of IM authorship analysis as input to the cybercrime investigation process and cyberprofiles.

2.1 Authorship Analysis

Authorship analysis is based on a linguistic research area called stylometric analysis and is sometimes referred to as authorship attribution [BVT1996, Hol1994, Rud1998]. Authorship analysis has been extensively applied to literature and published articles. Some of the earliest research dates back to the fourth century BC, when librarians in the library of Alexandria studied the authentication of texts attributed to Homer [Lov2002]. Other early known research dates back to the 18th century when English logician Augustus de Morgan theorized that authorship can be determined by the size of the words in the text [DeM1882]. Mendenhall [Men1887] analyzed and applied Morgan's research to publish results of authorship analysis among Bacon, Marlowe, and Shakespeare. More recent literary studies of authorship analysis include the disputed Federalist Papers [BS1998, MW1964, HF1995, TSH1996, LA2006] and Shakespeare's works [EV1991, MM1994, HAC2006]. Mosteller and Wallace [MW1964] conducted a thorough study of the authorship of the Federalist Papers and published results that attributed all 12 disputed papers to Madison. Historical scholars have generally accepted their conclusion, making it a milestone in this field [ZLCH2006]. More recent studies include research by Don Foster on the author of the Unabomber Manifesto [Fos2000], which helped convict Ted Kaczynski. Recent research has introduced authorship analysis to computer-mediated communications with promising results [DACM2001b, Ore2006b].

Authorship analysis can be divided into the following categories [ZLCH2006]:

- **Authorship identification**. Attempts to determine the author of a piece of text by examining other text samples that have been authenticated as having been produced by that author [Cha2005].
- **Authorship characterization**. Categorizes an author's text according to sociolinguistic attributes such as gender, age, educational background, income, linguistic background, nationality, profession, psychological status, and ethnicity. These attributes are aimed at inferring an author's background characteristics rather than identity. Some attributes previously researched are gender [KAS2002], [DACM2002a], [KCAC2006], language background [DACM2002b], and education level [JB2005].
- **Similarity detection**. Compares multiple sample texts and determines whether they were produced by a single author without actually identifying the author [ZLCH2006]. Aimed at discovering variations in the writing style of an author [PC2004] or to find the resemblances between the writings of different authors, mostly for the purpose of detecting plagiarism [GHM2005]. Plagiarism involves the complete or partial replication of a piece of work without properly acknowledging the original source [Clo2000].

Authorship analysis uses a variety of writing style features that can be derived from a particular piece of work to facilitate authorship identification or characterization. The identification and learning of these features with a sufficiently high accuracy is the principal focus and challenge of authorship analysis [DACM2002b]. Previous research studies have used a variety of features such as *n-grams* and spelling and grammatical

error frequency, however the following four feature categories have evolved for computer-mediated communication [AC2006, DACM2001a, ZLCH2006]:

- **Lexical**. Features such as the total number of words, number of words per sentence, word length distribution, vocabulary richness, average characters per sentence, average characters per word, and character usage frequency [AC2006].
- **Syntactic**. Features that involve patterns used for the formation of sentences, including punctuation and *function words* [AC2006].
- **Structural**. Features that involve the organization and layout of text including the use of greetings and signatures and the number of paragraphs and average paragraph length [AC2006].
- **Content specific**. Features that include key words that are used within a specific topic area [AC2006]. For example, a criminal selling illegal merchandise will use certain terms related to the items.

Researchers have begun to use authorship analysis to CMC, with application to e-mail, online forums (i.e. discussion groups or newsgroups), and online chat.

2.2 Research in Authorship Analysis of Computer-Mediated Communications

Olivier De Vel published several papers on authorship identification and characterization. The paper *Mining E-mail Content for Author Identification Forensics* [DACM2001b], published with A. Anderson, M. Corney, and G. Mohay studied the effects of multiple e-mail topics on authorship identification performance.

The experiments used 156 e-mail documents written by three authors. Each author contributed e-mails on each of three topics: movies, food, and travel. The experiments used a total of 191 features and the support vector machine (SVM) classification algorithm. SVM is a supervised learning model that constructs a hyperplane or set of hyperplanes to perform data classification. The hyperplanes separate sets of objects having different class memberships. With classification accuracy between 77.6%-91.6%, the results indicated the ability of the SVM to discriminate between authors without consideration of the conversation topic. Another experiment separated all e-mails into their individual topic categories. The SVM classifier was trained on the movie e-mail topic document set and tested on the remaining food and travel document sets. With performance accuracy up to 100%, the results indicated that the SVM classifier is able to effectively discriminate between the authors even when multiple topic categories are used. Overall, the experiments provided encouraging results for both aggregated and multiple topic author identification. This research benefited from a large feature set, however it is limited by a small dataset with few authors and emails.

Olivier de Vel, Anderson, Corney, and Mohay also published the paper *Gender-Preferential Text Mining of E-mail Discourse* [DACM2002a] that explores authorship categorization for gender identification. The research uses gender-preferential language features to determine the gender of the author of e-mail messages. Previous research has demonstrated that women's language tends to use more "emotionally intensive adverbs and adjectives, such as "so", "terribly", "awfully", "dreadful", and "quite", and that their language is more punctuated with attenuated assertions, apologies, questions, personal

orientation, and support" [DACM2002a].

The experiment data consisted of 4369 e-mails from 325 authors from a large academic organization. The experiments used a features set of 222 attributes. Baseline features consist of common character and word based attributes and structural features used in previous e-mail authorship identification research. The gender-specific attributes attempt to measure the frequency of use of adjectives, adverbs, and apologies. The results of the experiments indicate that the SVM classifier is able to discriminate between author genders with a maximum accuracy of 71.1% [DACM2002a]. This research benefited from an extended feature set and a dataset with a large number of authors and emails. It provided a foundation for future CMC author characterization research.

The paper *A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques* by Rong Zheng, Yi Qin, Zan Huang, Hsinchun Chen [ZLCH2006] presented a comparison of techniques for author identification by using several classification algorithms to analyze features. The authors leveraged existing feature sets from [DACM2001a], which they customized to include particular traits that are suitable to the datasets used for the experiments. The feature set was divided into lexical, syntactic, structural, and content-specific categories.

The experiments used English and Chinese newsgroup posting datasets. The English dataset consisted of messages from 20 authors (30-92 messages each) from misc.forsale.computers (including 27 subgroups) in Google newsgroups. The Chinese dataset consisted of Bulletin Board System (BBS) messages from 20 authors (30-40 messages each) from bbs.mit.edu and smth.org. The best accuracy was achieved with

SVM and all features. The experiments achieved accuracies of 90%-97% for the English dataset and accuracies of 72%-88% for the Chinese dataset. The experiments also studied the effects of the number of authors and number of messages used in the classification and analysis. The results showed that the accuracy of classification increases as the number of authors decreases or the number of messages per author increases. This paper provided many foundational principals for the IM feature set created for this research.

The paper *Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace* by Ahmed Abbasi and Hsinchun Chen [AC2008] introduced a writeprints technique for identification and similarity detection. Abbasi's writeprints is a "Karhunen-Loeve-transforms-based technique that uses a sliding window and pattern disruption to capture feature usage variance at a finer level of granularity" [AC2008].

The experiments used e-mail, instant messaging, feedback comments, and program code for datasets. The e-mail dataset consists of e-mail messages from the publicly available Enron e-mail corpus. The instant messaging dataset consists of IM logs from U.S. CyberWatch. The feedback comments dataset consists of buyer/seller feedback comments from eBay. The program code dataset consists of programming code snippets from the Sun Java Technology Forum (forum.java.sun.com). The experiments randomly extract 100 authors from each dataset. The feature sets consists of a baseline feature set (BF) and an extended feature set (EF). The BF contains 327 lexical, syntactic, structural, and content-specific features. The EF contains the BF features as well as several n-gram

feature categories and a list of 5513 common word misspellings.

For identification, the writeprints method is compared against the SVM and Ensemble SVM classifiers. For similarity detection, the writeprints method is compared against the principal component analysis (PCA) and Karhunen-Loeve (K-L) transforms. PCA is an unsupervised technique that captures variance across a large number of features in a reduced dimensionality. Karhunen-Loeve is a "supervised form of PCA that allows inclusion of class information in the transformation process" [AC2008]. The writeprint identification accuracy results for each dataset are: e-mail 83.1%-92%, IM 31.7%-50.4%, feedback comments 91.3%-96%, and program code 52.7%-88.8%. All techniques performed better on using the extended feature set. For similarity detection, the writeprints technique had the best performance on all datasets. Once again, the extended feature set had better overall performance than the baseline feature set. This research benefits from using a new writeprints technique, expanded feature sets, diverse datasets, and a large number of authors. Although the accuracy results are satisfactory for e-mail and feedback comments, they are very low for instant messages, indicating more research and experiments are needed in this area. This paper provided valuable insight into authorship analysis for a variety of CMC and provided many foundational principals for the IM feature set created for this research.

The paper *Writeprints: Conversationally-inspired Stylometric Features for Authorship Attribution in Instant Messaging* by Marco Cristani, Girogio Rofflo, Cristina Segalin, Loris Bazzani, Alessandro Vinciarelli, and Vittorio Murino [CRSBVM2012] introduced a new method of analyzing IM authorship called "turn-taking". Turn taking

focuses on analyzing the text (or set of messages) written by one participant during an interval of time in which none of the other participants writes anything. The authors used IM data captures via the Skype application for 77 authors with an average of 615 words per author. The paper analyzes conversation turns using 16 features. These features are mostly counts and don't consider specific features such as abbreviations or type of emoticon used. This paper cited [OA2009b] as a reference on the use of a stylometric feature set specifically developed for IM. IM conversations only partially mimic real life spoken conversations where one speaker listens and the other one talks. IM is more representative of a conversation where both people are speaking at the same time, constantly interrupting each other, and frequently jumping between simultaneous topics. The paper also measures writing speed, which can only be calculated on the authors computer or device and is not feasible for cybercrime investigations. While this paper introduces an interesting concept, the idea of turn-taking in IM is very questionable given the type of conversation interchange over IM.

The paper *A Unified Data Mining Solution for Authorship Analysis in Anonymous Textual Communications* by Farkhund Iqbal, Hamad Binsalleeh, Benjamin C.M. Fung, and Mourad Debbabi, [IBFD2013] models writeprints based on a combination of features frequently found in a suspect's messages. The experiments used 302 stylometric features, including lexical, syntactic, structural, domain specific, and gender preferential features. An anonymous message writeprint is compared with the writeprint of every suspect to identify the suspect that is most similar to the anonymous message writeprint. The experiments used the Enron email dataset for authorship identification and online

blogs for authorship characterization. The author identification accuracy was 69.75% using 20 email authors. The paper performs authorship characterization using gender and location sociolinguistic categories. However, the training data is extracted from an online source (blogger.com) that may not be representative of ground-truth data. This online data is not validated and false information may have been included in the training data. The author characterization accuracy was 60% for gender and 39.13% for location using 20 authors.

2.3 Research Using Gaussian Distributions

Gaussian distributions have been applied in various biometric and forensic research efforts including facial recognition and tracking [RMG1998, Man2012], voice recognition [RQD2000], [MD2001], signature verification [RD2003], and characterizing social media authors [PC2011].

The paper *Tracking and Segmenting People in Varying Lighting Conditions using Colour* by Yogesh Raja, Stephen J. McKenna, and Shaogang Gong [RMG1998] applies Gaussian models to detect and track people on video. The authors use color hues to detect, track, and segment people, faces, and hands in dynamic scenes. They use Gausssian models to estimate probability densities of skin color, clothing, and background. The experiments are performed on two use cases: 1) actor segmentation for virtual studios, and 2) focus of attention for face and gesture recognition systems. The experiments collect a number of Gaussian functions that are an approximation to a multi-modal distribution in color space and probabilities are computed for various color pixels. The color mixture models are then used to track and segment the colors of people's skin,

clothing, and background. The result is a statistical distribution of colors of a person in an image plane. This paper is an example of applying Gaussian models in a biometrics scenario. The paper provides a framework and models for performing these experiments, but it did not provide quantifiable results data.

The paper *Gaussian Mixture Models for On-line Signature Verification* by Jonas Richiardi and Andrzej Drygajlo [RD2003] applies Gaussian models for online signature verification. A handwritten signature is a commonly used behavioral biometric. The research uses spatial and temporal signature features to create Gaussian Mixture Models (GMM). The signatures "are sampled at 100 Hz using a Wacom Intuos A6 tablet on which a paper alignment grid is placed, and each sample point (raw data vector) consists of values for the horizontal (x) position, vertical (y) position, pressure (p), azimuth, and elevation of the signing pen." [RD2003] The experiments include a 50-user subset of the Ministerio de Ciencia y Tecnologı´a (MCYT) multimodal database. The data consists of 25 authentic signatures and 25 skilled forgeries for each user. "The performance of the proposed GMM-based signature verification system was tested by scoring each user model against 20 authentic signatures and 25 skilled forgeries (from 5 different forgers), making a total test set of 1000 authentic signatures and 1250 forgeries" [RD2003]. The experiments resulted in a 3.5% error rate using 64 Gaussian components. This paper shows that GMM is an effective method for online signature verification as a behavioral biometric.

The paper *Identifying Topical Authorities in Microblogs* by Aditya Pal and Scott Counts [PC2011] applies Gaussian models to distinguish microblogging authors of high

topical value. The research proposes a set of features for characterizing social media authors and performs probabilistic clustering over this feature space to produce a ranked list of top authors for a given topic for identifying topical authorities in microblogging environments. The authors use Gaussian models to cluster authors over their feature space. The research uses data collected from the public social networking site Twitter over five days between June 6, 2010 and June 10, 2010. Tweets are extracted from this dataset across three topics: oil spill, world cup, and iPhone using simple substring matching. The research categorizes tweets into three categories: original tweet (OT), conversational tweet (CT), and repeated tweet (RT) across 17 total features. The resulting dataset included 539,524 users and 1,563,320 tweets. The authors used Gaussian models to generate a list of the top 10 authors for each topic. To evaluate their approach, the authors conducted a user study where participants were shown 40 screens, each with a different author. Each screen asked participants to evaluate the author on "how interesting and authoritative they found the author and her tweets" using a 7-point Likert scales. A Likert scale is used to assign a numerical value to qualitative data. In this research a scale of 1-7 was used, with 7 being the highest for each question. The results of the user study showed that the Gaussian model technique yielded results that correlated highly with the end user's ratings. This research was effective at implementing a set of features and applying the Gaussian models to identify interesting and authoritative authors across various topic areas as validated by user opinion.

2.4 Criminal Investigation and Cybercrime

Many disciplines including psychology, philosophy, sociology, criminology, law, knowledge management, and computer science have studied the criminal investigation process. Although cybercrime is a relatively new form of crime that has rapidly evolved over the last few decades, cybercrime investigations and traditional criminal investigations share the same goal – to gather information. Figure 2-1 illustrates the traditional criminal investigation process as presented in *Scene of the Cybercrime* [Cro2008].

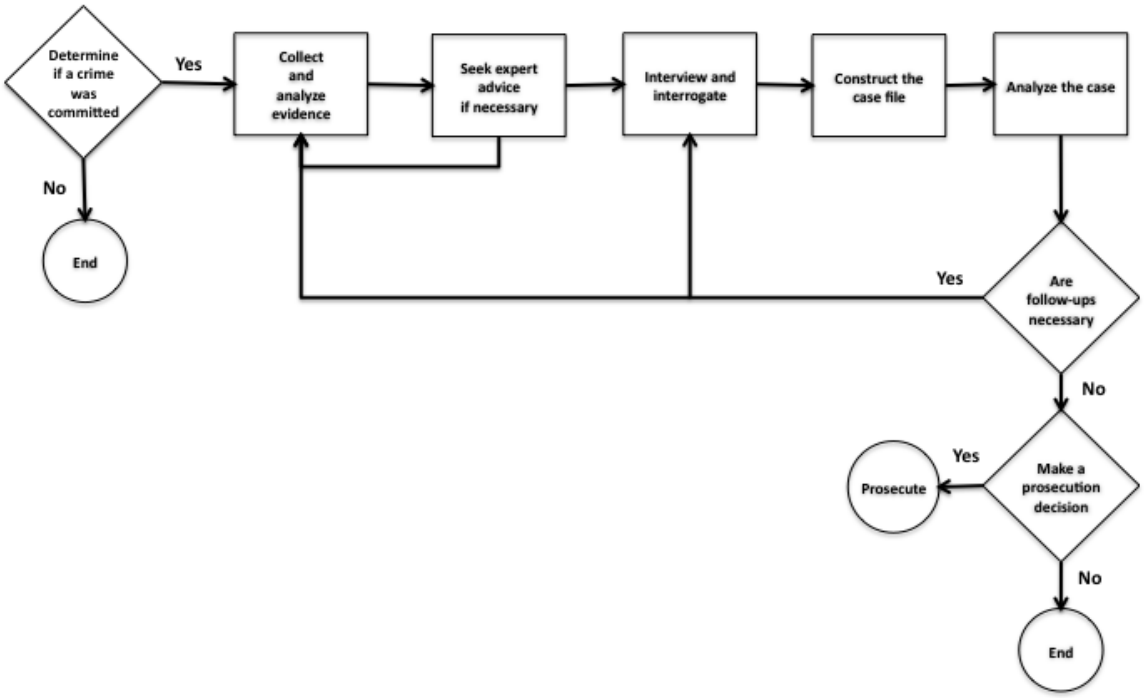


Figure 2-1. Criminal Investigation Process

The investigator first determines if an act has violated the law and warrants an investigation. Next, evidence is collected and analyzed, including tangible evidence such as hard drives and electronic devices, and the digital evidence they contain. Cybercrime investigations for IM rely on instant messaging exchanges, or conversations, as digital evidence. The sources for IM digital evidence include both data and meta-data. The data includes the IM text and the meta-data includes other related evidence such as the IM client version, timestamps, the length of time the user has been logged on, etc. The next step involves seeking expert advice if necessary. Often times in cybercrime cases the investigator needs to seek expert advice on the technical aspects of the crime. Experts may be on staff, or may be located from professional organizations, consultants, or the academic community. For IM related cybercrimes experts may include computer scientists, network engineers, linguists, communication experts, or social psychologists. Expert analysis of IM writeprints may determine probable cause and justify search and seizure of additional evidence. Experts may also use IM writeprints to build a criminal profile to focus the investigation. The next step of interviewing witnesses and interrogating suspects is an ongoing process throughout the investigation as new witnesses and suspects are discovered. IM writeprint analysis may be used to shape the interview and interrogation strategy. For example, if the suspect's writeprint is highly similar to writeprints in other cases, this information could be used in an attempt to link related crimes and to question the suspect on his or her relationship with the other victims. Throughout this stage suspects are eliminated and the most plausible suspect is identified. Next, the investigator begins preparing the case file to include the initial

incident report and evidence. Evidence might include other reports such as lab reports, written statements, and other relevant information. Once the case file is constructed it is analyzed to determine weaknesses and to identify additional information needed for prosecution. This analysis leads to any follow-up investigations that need to occur including collecting additional evidence and interviewing witnesses again. Once the case is considered complete the prosecutor will decide whether to bring the case to trial and how to proceed.

Expert witnesses may use IM writeprint analysis and criminal profiles as part of their presentation in a trial. There is no standard accuracy measure threshold for authorship attribution evidence; the investigator only needs probable cause to initiate a warrant or arrest. Authorship analysis identification results greater than 70% are acceptable during an investigation process [IBFD2013]. Evidence admissibility varies by jurisdiction and one of the biggest hurdles is establishing authenticity of the messages [HZ2012]. State and Federal courts have been applying Federal Rules of Evidence (FRE) 901, Authenticating or Identifying Evidence and its state rule equivalents to text and instant messages [HZ2012]. The frequently used FRE 901 rules to authenticate digital evidence include:

> **FRE901 b(3)** ***Comparison by an Expert Witness or the Trier of Fact.*** A comparison with an authenticated specimen by an expert witness or the trier of fact.

FRE901 b(4) ***Distinctive Characteristics and the Like.*** The appearance, contents, substance, internal patterns, or other distinctive characteristics of the item, taken together with all the circumstances.

FRE901 b(4) has been used most frequently to submit text and IM messages as circumstantial evidence, for example People v. Pierre, New York, 2007. The case of People v. Agudelo, New York, 2012 set the precedent, which allows authentication of text messages based on testimony from either the sender or recipient of the message as admissible evidence. Validation studies provide the kind of information required for an evidence admissibility hearing. Some relevant cases were investigated and prosecuted based on text message abbreviations, sentence length, and punctuation [Lea2009]. [Cha2013]'s authorship analysis methods on a computer diary (Green v. Dalton, DC, 2001) and email (Zarolia v. Osborne/Buffalo Environmental Corp, MD, 1998) have been admitted as evidence into several trials along with expert testimony to state a conclusion about authorship. In another [Cha2005] case the plaintiff withdrew his suit against the government after the questioned emails were identified as his own writing based on the syntactic patterns. "Even with partial admissibility, it may still be possible for an expert to take the stand, outline a set of features, and demonstrate informally that the defendant's use is more inline with the document under discussion than other candidates, while leaving the jury to draw the necessary conclusion that the document was written by the defendant" [Juo2006]. It is important to ensure that analysis methods are easy to understand to enable an investigator or an expert witness to present the writeprint and explain the finding in a court of law [IBFD2013]. Easy to understand graphs allow

information to be communicated to investigators, attorneys, judges, and jury members in a simple and clear way [Cha2005].

IM, email, and text messages have been use as part of the investigation and submitted as evidence in trials (State v. Lott, NH, 2005). Authorship analysis of email and computer-based digital text have been used in testimony in several trials [Cha2013]. There is a gap in using IM authorship analysis to support the investigation and eventually, trial. IM writeprints may be used to generate investigative leads, create leverage in negotiating a settlement, or to present admissible evidence in a trial. This dissertation provides the foundation for using IM writeprints during the investigation activities.

2.5 Criminal Profiling

Criminal profiling is an investigative method that has been used in traditional investigations that can also be applied to cybercrime investigations, known as cyberprofiling. Cross defines traditional criminal profiling is the "art and science of developing a description of a criminal's characteristics (physical, intellectual, and emotional) based on information collected at the scene of the crime" [Cro2008]. Criminal profiling often uses patterns and correlations among criminal activity and different crimes to construct a profile. Criminal profiling is used to assist with the investigative process, reduce the potential suspect space to a certain subset of suspects, link related crimes, and develop an interview and interrogation strategy [Cas1999]. It is important to note that a criminal profile will only provide generalities about the type of person who committed a crime, it will not identify a specific individual. Criminal

profiling is one method among many for assisting with investigations and building a case file. The profile cannot exist as evidence, rather it provides information to allow investigators to focus on the right suspects and begin to gather additional evidence [Cro2008]. A criminal profile can be used in court in conjunction with expert witness testimony. "An expert witness can reference a criminal profile as the basis of an opinion that there is a high probability of a link between a particular suspect and a particular crime" [Cro2008]. An IM author writeprint may be used as input to a criminal profile.

Criminal profiling can be traced to the end of the 19th and beginning of the 20th centuries and was later used to investigate the high profile Jack the Ripper killings as well as preparing an interrogation strategy of Adolf Hitler during World War II [Rog2003]. The FBI is credited with formalizing the criminal profiling process. The FBI's Behavioral Science Unit (BSU) "focuses on developing new and innovative investigative approaches and techniques to solve crimes by studying offenders and their behaviors and motivations" [FBI2013]. The BSU has been assisting local, state, and federal agencies in narrowing investigations by providing criminal profiles since the 1970s [DRBH1986]. The FBI BSU has created the six-step criminal profile generating process shown in Table 2-1.

Table 2-1. FBI BSU Criminal Profile Process

FBI BSU Criminal Profile Process	
1. Profiling Inputs	The first step collects profiling inputs including comprehensive information about the crime and all evidence collected, both tangible, physical evidence and digital evidence.
2. Decision Process Models	This step analyzes the information and evidence to determine patterns and possible linkages to other crimes.
3. Crime Assessment	The crime scene is reconstructed and analyzed to determine the sequence of events and other information about the crime.
4. Criminal Profile	The first three steps are combined to create a criminal profile, often incorporating the motives, physical qualities, and personality of the perpetrator. The criminal profile is also used to create an interrogation strategy for the suspects.
5. The Investigation	Investigators and others use the profile to learn more information and identify suspects. Suspects matching the profile are evaluated. The profile may be reassessed if no leads or suspects are identified.
6. The Apprehension	The last stage occurs when investigators believe they have identified the most plausible suspect likely to be the perpetrator. A warrant is obtained for the arrest of the individual, usually followed by a trial [DRBH1986].

The FBI criminal profile generating process may be easily applied in a cybercrime investigation to perform cyberprofiling. Various types of digital and non-digital evidence may be combined as profile inputs, including, email, IM conversations, network packet captures, account activity information, and physical evidence. Data and author behavioral biometrics are analyzed for patterns using statistical methods. For example, IM conversations may be analyzed for patterns to create an IM author's writeprint. A cybercriminal's profile may include a number of traits such as time and location of

computer access, types of computer attacks launched by the attacker, programs and attack tools used, writeprints, and targets of the cybercrime whether they be human or electronic (networks, satellites, phones, computer systems, etc.).

In the context of IM-assisted cybercrime, cyberprofiling uses IM data such as the conversation logs, IM client version, timestamps, the length of time the user has been logged on, etc. This dissertation focuses on the linguistic aspects of the IM conversation to create IM author writeprints that may be used as an input to a criminal cyberprofile and as an element in a multimodal biometrics systems to assist with cybercrime decision support. Writeprints may be used in conjunction with other evidence and investigative techniques to build or validate a criminal profile; reduce the potential suspect space to a certain subset of suspects; link related crimes; develop an interview and interrogation strategy; and gather convincing digital evidence to justify search and seizure and provide probable cause. This dissertation uses authorship analysis and statistical techniques to create and analyze IM writeprints to assist in identifying an author, as well as certain characteristics of the author of a set of IM messages.

3. RESEARCH METHODOLOGY

This chapter presents detailed explanations of the methods, tools, and techniques to create and analyze behavioral biometrics-based instant messaging writeprints to assist in identifying online cyber criminals and collecting digital evidence as part of the criminal investigation. It provides a detailed description of the problem and the goals this research seeks to achieve. It explains the research process and the IM-specific stylometric feature set taxonomy. This chapter also provides a detailed description of the two datasets used in the research, including methods of collection, dataset size, and author demographics. Lastly, this chapter provides an overview of the statistical software and other tools used to analyze the data. The research methodology uses:

- An IM-specific stylometric feature set,
- Two real-world datasets,
- Preprocessing software to extract the stylometric features into a writeprint,
- Statistical software to process models, and
- Graphical software to visualize results.

3.1 Problem Definition

The explosive growth in the use of instant messaging communication in both personal and professional environments has resulted in an increased risk to proprietary, sensitive, and personal information and safety due to the influx of IM-assisted

cybercrimes, such as phishing, social engineering, threatening, cyber bullying, hate speech and crimes, child exploitation, sexual harassment, and illegal sales and distribution of software. Instant messaging's anonymity and use of virtual identities hinders social accountability and presents a critical challenge for cybercrime investigation. Criminals use virtual identities to hide their true identity by using multiple screen names, impersonating other users, or supplying false information on their virtual identities with the intention of deceiving unsuspecting victims and committing various cybercrimes. Although central IM servers authenticate users upon login, there is no means of authenticating or validating peers (buddies). Current IM products are not addressing the anonymity and ease of impersonation over instant messaging. Cyber forensic techniques are needed to assist cybercrime decision support tools in collecting and analyzing digital evidence, discovering characteristics about the cyber criminal, and assisting in identifying cyber criminal suspects.

This research creates and analyzes behavioral biometrics-based instant messaging writeprints as cyber forensics input for cybercrime investigations. It uses authorship analysis techniques to create a set of stylometric features robust enough to show separation between authors and between author categories. The real time, casual nature of instant messaging communications offers several interesting stylometric features such as message structure, unusual language usage, and special stylistic markers that are useful in forming a suitable writeprint feature set for authorship analysis. Authorship identification can be applied to IM to assist in identifying criminals who hide their true identity or impersonate a known individual. Authorship characterization can be used to

help discover IM cyber criminals who supply false information in their virtual identities, such as gender.

3.2 Research Process

The research process extracts stylometric features from IM messages to create author writeprints and uses statistical methods to analyze and evaluate the writeprints. This research evaluates the effectiveness of the writeprints using different parameters such as the number of messages used as input. These parameters are systematically modified in an iterative process to evaluate their impact on the results. The goal of this research is to create and validate IM author writeprints that provide cybercrime investigators a unique tool for investigating IM-assisted cybercrimes. Writeprints may be used as input to a criminal cyberprofile and as an element of a multimodal system for cybercrime investigations. Writeprints can be used in conjunction with other evidence, criminal investigation techniques, and biometrics techniques to reduce the potential suspect space to a certain subset of suspects; identify the most plausible author of an IM conversation from a group of suspects; link related crimes; develop an interview and interrogation strategy; and gather convincing digital evidence to justify search and seizure and provide probable cause. At a high level this research performs the following:

1. Develops a stylometric feature set,
2. Pre-processes the data,
3. Creates writeprints, and
4. Analyzes and evaluates writeprints with statistical methods.

The detailed research process is illustrated in Figure 3-1.

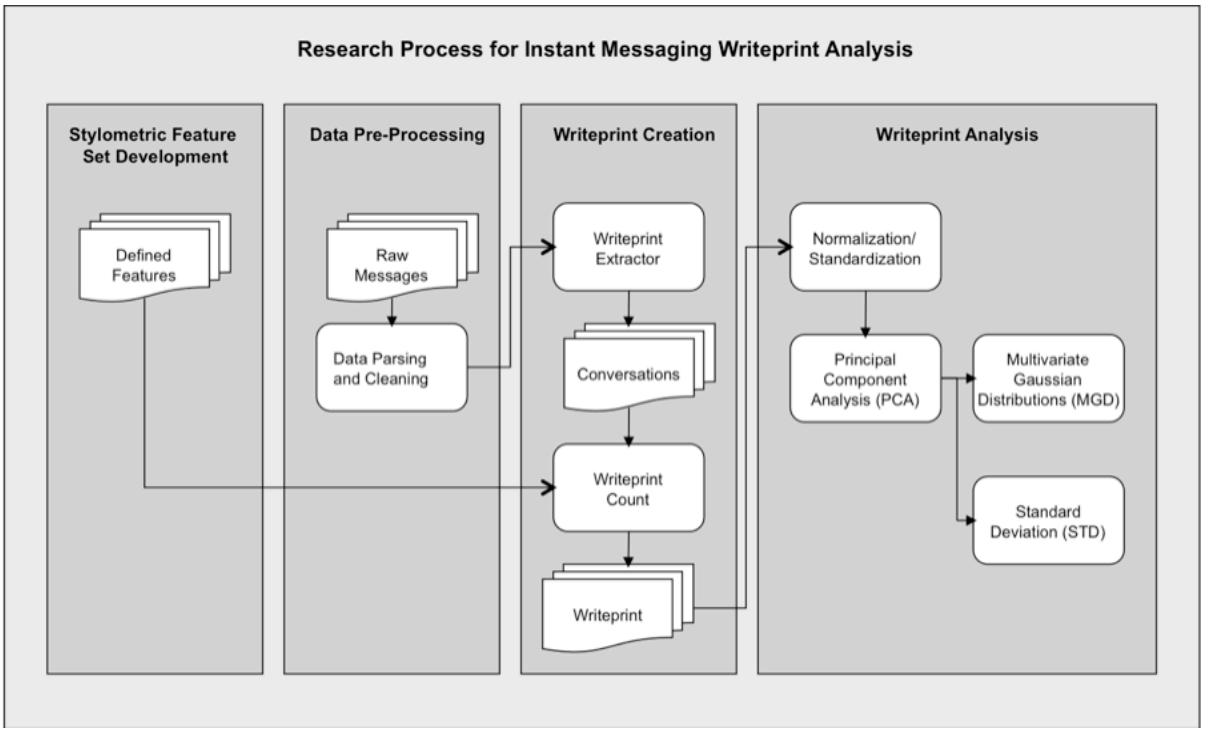


Figure 3-1. Research Process for Instant Messaging Writeprint Analysis

3.2.1 Stylometric Feature Set Taxonomy

Stylometric features are characteristics that can be derived from instant messages to facilitate authorship analysis [AC2006]. A stylometric feature set is composed of a predefined set of measurable writing style attributes. Given *t* predefined features, each set of IM messages for a given author can be represented as a *t*-dimensional vector, called a writeprint. Feature sets may significantly affect the performance of authorship analysis, both positively and negatively. Previous studies have created feature sets for computer-

mediated communications such as e-mail, forums, and online chat [DACM2001b, KCAC2008, ZLCH2006]. Previous studies have also created feature sets for specific content-related purposes, such as illegal sales and distribution of software, to facilitate cybercrime investigations. Other stylometric features sets are not comprehensive enough to capture the stylistic features that are frequently found in IM communications. [DMCT2011] and [CRSBVM2012] contain small feature sets (37 and 16 respectively) of category totals. For example, other research considers total emoticons used in a message, while this research considers counts for specific varieties of emoticons used in a message. This research provides a comprehensive stylometric feature set taxonomy of instant messaging writing style characteristics to create IM writeprints to assist with cyber forensics and cybercrime investigations.

Numerous types of features have been used in previous studies including n-grams and vocabulary richness [BVT1996, AC2008, ZLCH2006], however four categories used extensively for computer-mediated communications are lexical, syntactic, structural, and content-specific features [LZC2006, AC2006, ZLCH2006]. The feature set created for IM writeprints in this research is a 356-dimensional vector including lexical, syntactic, and structural features, shown in Figure 3-2. The number of features in each category is shown in parenthesis. Content specific features are highly dependent on the topic of the messages; therefore the feature set taxonomy does not include content specific features in order to achieve generic authorship identification and characterization across various applications. With instant messaging authorship analysis, a feature set that is

independent of the message topic focuses on the authors' stylistic preferences instead of the specific vocabulary [KCAC2008].

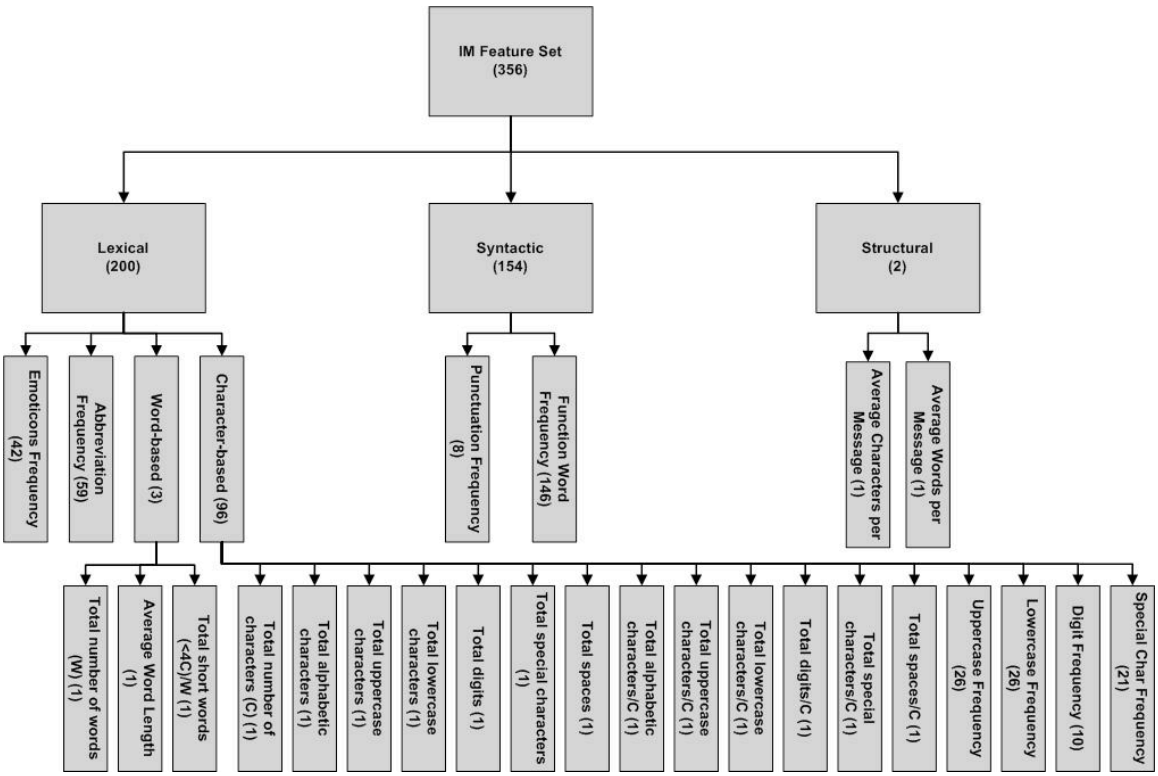


Figure 3-2. Instant Messaging Stylometric Feature Set Taxonomy

Instant messaging communications have several characteristics that are useful in forming a comprehensive feature set, which may help reveal the writing style of the author. The IM feature set taxonomy created for this research includes several stylistic features that distinguish it from related studies [DMCT2011] and [CRSBVM2012], such as abbreviations and emoticons, which are frequently found in instant messaging communications.

Table 3-1 shows a detailed breakdown, with examples where applicable, of the IM feature set. Lexical features mainly consist of count totals and are further broken down into emoticons, abbreviations, word-based, and character-based features. Syntactic features include punctuation and function words in order to capture an author's habits of organizing sentences. Function words include conjunctions, prepositions, and other words that carry little meaning when used alone, such as "the" or "of". They provide relationships to content words in the sentence, such as "ball" or "bounce". Analyzing function words as opposed to content words allows topic-independent results. Structural features capture the way an author organizes the layout of text. With IM communications there are no standard headers, greetings, farewells, or signatures, leaving simply the average characters and words per message in terms of structural layout.

The feature set taxonomy created for this research is tailored for IM authorship analysis. The goal of the IM feature set taxonomy is to develop a set of features that show separation of author writeprints. Each stylometric feature in the taxonomy was selected for its relevance to IM communications to create a feature set robust enough to show separation between various authors and between author categories.

Table 3-1. Feature Set Detail and Examples

Features	Applicable Examples
Lexical Features	
Emoticon Counts	See Appendix A
Abbreviation Counts	See Appendix A
Word-Based	
Total Number of Words (W)	–
Average Word Length	–
Total Short Words (<4C)/W	–
Character-Based	
Total Number of Characters (C)	–
Total Alphabetic Characters	–
Total Uppercase Characters	–
Total Lowercase Characters	–
Total Digits	–
Total Special Characters	–
Total Spaces	–
Total Alphabetic Characters/C	–
Total Uppercase Characters/C	–
Total Lowercase Characters/C	–
Total Digits/C	–
Total Special Characters/C	–
Total Spaces/C	–
Uppercase Counts	A-Z
Lowercase Counts	a-z
Digit Counts	0-9
Special Character Counts	~ @ # $ % ^ & * - _ = + > < [ ] { } / \ |
Syntactic Features	
Punctuation Counts	, . ? ! : ; ‘ “
Function Word Counts	See Appendix A
Structural Features	
Average Characters per Message	–
Average Words per Message	–

3.2.2 Data Pre-Processing

Instant messaging data pre-processing involves several steps to prepare raw IM data for writeprint creation. The IM messages are logged to American Standard Code for Information Interchange (ASCII) text files in the following format:

[timestamp] [user name:] [message]

Figure 3-3 shows an example excerpt from an IM conversation log.

10:19:29 AM User1: hey, what time is the meeting today?
10:19:35 AM User2: It is at 11AM…are you going?
10:19:39 AM User1: yeah, I will be there, it sounds very interesting! :) :)

Figure 3-3. Raw IM Conversation Log

The instant messages are parsed to extract the data for each author and to remove metadata and noise, such as timestamps, usernames, and automatic away message responses. Thus,

Figure 3-4 shows a formatted log for User 1.

hey, what time is the meeting today?
yeah, I will be there, it sounds very interesting! :) :)

Figure 3-4. Formatted IM Conversation Log

3.2.3 Writeprint Creation

Writeprint creation uses a Perl program to process formatted logs and create writeprints. First, the writeprint extractor module splits the logs into a configurable conversation size. A conversation is a set of messages $\{M_1,\ldots,M_p\}$, for example 50 messages per conversation. A message consists of the text delineated by the newline or end-of-line (EOL) character. For example,

Figure 3-4 contains only two messages. Many other CMC stylometric related works focus on message level analysis [AC2008, ZLCH2006]. [CRSBVM2012] focuses on "turn-taking", which is very dynamic and often influenced by the other party in the message exchange. This research analyzes conversations of various sizes to determine the necessary number of messages to separate authors.

Next, the program inputs conversations and defined stylometric features to the count module to create totals for each stylometric feature, resulting in the output of a writeprint (W_x) for each set of messages $\{M_1,\ldots,M_p\}$ of each supplied author (A_n) or author category (C_m). A writeprint is a t-dimensional vector, where t represents the total number of features. This research uses a 356-dimensional vector. Each writeprint is assigned a class, which is the author (A_n) or sociolinguistic category (C_m) of the writeprint (W_x). Table 3-2 shows the writeprint class descriptions and labels used in this research. The program outputs a writeprint in comma-separated value (CSV) format. Each value in the writeprint represents a count or ratio for a specific feature. The features in the vector do not need to be in a specific order for this research since each feature is assigned a label identifying it. An example writeprint for an author $W(A_n)$ using a selected feature set $\{F_1,\ldots,F_q\}$, where q=100, for a set of messages $\{M_1,\ldots,M_p\}$ looks like the following:

```
105,1,0,0,4,0,1250,0,4,0,18,8,1,2,0,0,0,0,1,9,0,14,31,6.78,3.71,23,
0,67,4,25,5,0,117,5,0,1,4,0,0,23,0,0,0,8,0,23,1,3,0,27,50,0,0,1550,
0,7,0,0,0,1,0,1250,33,0,13,1,0,0,0,2,85,0,0,0,4,0,0,0,0,0,96,1,0,0,
0,13,0,3,0,10,0,2,0,0,0,1,2,16,0,0.806,User1
```

Table 3-3 shows the notation used for the IM writeprint analysis research.

Table 3-2. Writeprint Class Descriptions and Labels

Class Description	Class Labels
Author Identification	
Author	User1, User2, User3, …
Author Characterization	
Gender	Male, Female
Age Group	<20, 20s, 30s, >40
Education level	High School, College

Table 3-3. IM Writeprint Analysis Notation

Notation	Description	Example
$\{A_1,\ldots,A_n\}$	The set of authors	{User1, User2, User3}
$A_i \in \{A_1,\ldots,A_n\}$	An author that is a member of the set of authors	User2
$\{M_1,\ldots,M_p\}$	The set of messages	{Hey math is fun! :)}
$\{F_1,\ldots,F_q\}$	The set of features	{Punctuation frequency, #words per message, Total spaces}
$\{C_1,\ldots,C_m\}$	The set of categories	{male, female}
$W(A_n)$	The writeprint of Author n	0,1,44,0,23,3,6,1,User5

3.2.4 Writeprint Analysis

Writeprints must be normalized and standardized prior to input into statistical models. Writeprints consist of count totals that range in values from small to large across the 356-dimensional vector. Features with large values can often dominate the results of statistical models. For example, features that have large values may influence distance-based algorithms, such as Euclidean distances. Normalization and standardization ensures that features with a wide range of values are less likely to outweigh features with smaller ranges. It allows data on different scales to be compared by bringing them to a common scale, thus allowing the underlying characteristics of the data sets to be compared. The IM writeprint data is normalized and standardized using the following steps:

1. **Normalization:** The range of values in a writeprint is normalized to be between 0.0 and 1.0. Normalization is performed using the formula

$$(X_i)' = \frac{X_i}{\sum_{i=1}^{k} X_i}, \tag{3-1}$$

where each feature value in the writeprint is divided by the sum of all feature values in the writeprint. X_i is the value of the i-th feature in the writeprint vector. X_i prime is the new value of the i-th feature in the writeprint vector. k is the total number of features in the writeprint vector. In this research $k = 356$.

2. **Standardization:** The range of values for each feature across all writeprints in a population is standardized to measure the number of *standard deviations* the feature value is from its mean. This is known as a standard score or Z-score. Standardization converts all feature values to a common scale with an average of zero and a standard deviation of 1 using the formula

$$\left(X_i\right)'' = \frac{\left(X_i\right)' - \mu}{\sigma} \quad . \tag{3-2}$$

The mean and standard deviation is calculated across all values of each feature in the dataset population. X_i is the value of the *i*-th feature in the writeprint vector. X_i double prime is the new value of the *i*-th feature in the writeprint vector. For each feature value, subtract the mean of the set of features (μ) in the population and divide by the standard deviation of the set of features (σ) in the population.

After the writeprints are normalized and standardized, statistical models (described in section 3.3) are created and used to visualize and analyze the data.

3.3 Statistical Methods and Software

When providing authorship analysis support to a criminal investigation it is important to use easy to understand methods that an investigator or expert witness can present and explain in the court of law [IBFD2013]. For example, SVM is a popular choice in many CMC authorship analysis related works [DACM2001b, ZLCH2006,

AC2008], however its methods are often difficult to interpret to demonstrate the reason for reaching a conclusion. Therefore methods such as SVM are not suitable for evidence collection and presentation, which are important steps in a criminal investigation [IBFD2013]. This research uses methods visualizations that communicate analysis information to investigators, attorneys, judges, and jury members in a simple and clear way. This research uses Principal Component Analysis (PCA), multivariate Gaussian distributions (MGD), Standard Deviation (STD), Python programming language, Perl programming language, and Gnuplot visualization software package to create and analyze instant messaging writeprints. This section explains each of these statistical models and software packages.

3.3.1 Statistical Methods

Principal component analysis is a statistical technique that reveals first order patterns in high dimension data. PCA performs dimension reduction to reduce a large set of features to a small set that still retains most of the information as the large set. Datasets with a large number of features often suffer from the curse of dimensionality, which are the difficulties associated with analyzing high dimension data. As the dimensionality increases, data becomes increasingly sparse in the space it occupies, leading to inaccurate and unreliable data models. PCA's dimensionality reduction eliminates irrelevant, weakly relevant, or redundant features and reduces noise. It also leads to a more understandable model because the model has fewer attributes and it eases visualization. PCA applies data transformation to create a reduced representation of the original data. PCA simply rotates the underlying data space in order to create a new set

of axes that capture the most variance in the data. PCA creates new features, called principal components, that are:

1. Linear combinations of the original features,
2. Orthonormal to each other, and
3. Able to capture the maximum amount of first order variation in the data.

The first principal component is the axis on which the data has the most variance. The second principal component is the axis orthogonal to the first that captures the next greatest variance, and so forth. Two dimensions in which the original data strongly correlated would be captured as a single new dimension that is the vector sum of the original two dimensions. For example, if the data to be compressed consists of N tuples (instances of IM conversations), from k dimensions (features), PCA searches for c k-dimensional orthogonal vectors that can best be used to represent the data where $c<=k$. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. In this research, PCA is performed using the following steps [Smi2002]:

1. **Subtract the mean.** For each dimension, calculate the average for that dimension across all tuples. Subtract the mean from each dimension. Subtracting the mean centers the data by translating the coordinate system to the location of the mean.
2. **Calculate the covariance matrix.**

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \mu_x)(Y_i - \mu_y)}{(n-1)}, \tag{3-3}$$

where n is the sample size, μ_x is the mean of X, and μ_y is the mean of Y. Covariance calculates how much the dimensions vary from the mean with respect to each other. A positive covariance indicates that the values increase together. A negative covariance indicates an inverse relationship where as one increases the other decreases. Covariance is used to find relationships between dimensions in high dimensional datasets where visualization is difficult. This research has a covariance matrix of size 356x356.

3. **Calculate the eigenvectors and eigenvalues of the covariance matrix.** Calculating the matrix of eigenvectors diagonalizes the covariance matrix. This research has 356 eigenvalues and a 356x356 eigenvector matrix.
4. **Form a feature vector with a specified number of eigenvectors.** The eigenvector with the highest eigenvalue is the principal component of the dataset. After eigenvectors are created from the covariance matrix they are ordered by eigenvalue from highest to lowest. Create a feature vector using the first p eigenvectors. This research uses p=5.
5. **Convert the data points into the new space.** The location of the original data points are computed by projecting them onto the new coordinates, which is the dot product of the data points with the new axes. After the PCA vector is created from the first p eigenvectors, take the dot product of the PCA vector

and the data vector to create the final data coefficients that represent the data vector in the new space.

PCA has been used in numerous previous authorship analysis studies and has been shown effective for online stylometric analysis [AC2008]. "PCA's ability to capture essential variance across large numbers of features in a reduced dimensionality makes it attractive for text analysis problems, which typically involve large feature sets" [AC2008]. PCA was chosen for the IM writeprint analysis due to the high dimension stylometric feature set. The 356-dimension feature set was created to provide a comprehensive capture of the stylistic features that are frequently found in IM communications. However, in real world data, an author's use of various features is often inconsistent. There may be a large number of the 356 features that are not used by certain authors and some features used similarly across all authors. This results in sparse data, irrelevant features, and weakly relevant features. PCA is used to reduce the number of necessary dimensions, highlight similarities and differences, and ease visualization. The reduced data is visualized using graphing tools and then input to MGD and STD for analysis.

A multivariate Gaussian distribution is a generalization of the Gaussian distribution to higher dimensions. The multivariate Gaussian distribution equation used in this research is

$$y = \mathrm{A}\exp\left\{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\sum^{-1}(x-\mu)\right\}, \tag{3-4}$$

where x is the test vector, μ is the mean, and $\sum$ is the covariance matrix of the Gaussian. The covariance matrix is 5x5 since it uses the first 5 PCA coefficients. In this research y returns P(x|Author) or P(x|Category), which is the probability of the test vector given the provided author or author category, also called the likelihood.

The Gaussian distributions are used to determine the probability of the author or author category of the test writeprint. MGD is performed using the following steps (the same process is used for both author identification and author characterization):

1. **Choose the sample data.** The reduced dimension PCA files for each author in the sample are input for MGD processing. An example file list may be data1_250_user1, data1_250_user2, and data1_250_user3. These files contain writeprint data for Authors 1, 2, and 3 for a conversation size of 250 messages. Each file contains *x, y,* and *z* writeprints, respectively.
2. **Choose the test writeprint data.** A sample author test file is selected for analysis. For example, a test file may be data1_250_user2. This file contains *w* writeprints for Author 2 using a conversation size of 250 messages.
3. **Assess each test writeprint within each distribution.** Using leave one out cross validation, each writeprint of the author test file is used as the test vector to calculate the Gaussian distribution using the mean and covariance for each author or category file in the sample. The result returns the P(x|author) or P(x|category), also known as the likelihood.
4. **Determine posterior probability for author identification or characterization.** Determine the posterior probability, P(Author|x), for each

author or author category in the sample given the test writeprint. For example, the author under test, data2_250_user2, may have the following probability results across the sample: data1_250_user1 = 20%, data1_250_user2 = 70%, data1_250_user3 = 10%. Probability P(Author|x) is calculated using Bayes Rule:

$$P(Author_i \mid x) = \frac{P(x \mid Author_i)P(Author_i)}{P(x)}, \tag{3-5}$$

Prior probability in this research is equally distributed. For example, a test using 5 authors would have the following prior probabilities: P(A1) = 20%, P(A2) = 20%, P(A3) = 20%, P(A4) = 20%, P(A5) = 20%. P(x) is calculated as P(x|A1)P(A1) + P(x|A2)P(A2) + P(x|A3)P(A3).

MGD was chosen for the IM writeprint analysis due to successful results in other forensics and behavioral biometrics research [RMG1998], [Man2012], [RQD2000], [MD2001]. Because MGD is based on means it tends to be less sensitive to outliers, making it well suited for the IM writeprint analysis.

The standard deviation measures the spread of distribution of a set of data by calculating distance from the mean of the data. If the data points are very close together (close to the mean), the standard deviation will be low. If the data points are spread out (far from the mean), the standard deviation will be high. IM conversation sizes are analyzed in more detail by calculating the standard deviation of the data within each

conversation size. The standard deviation is calculated for the first 5 dimensions of the PCA data using the formula

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}}, \tag{3-6}$$

where x is the value of the dimension, μ is the sample mean, and n is the sample size. The first 5 dimensions represent those with the most variability in the data with decreasing variability from 1 to 5. Dimensions after the 5th do not provide significant variability to be included.

3.3.2 Software

This research uses the Python programming language to calculate PCA, MGD, and STD. Python permits several programming styles, such as object oriented and structured, and has a large standard library. This research uses the Numpy and Scipy libraries [JOP2001]. Numpy adds support for linear algebra, standard deviation, and large, multi-dimensional arrays and matrices. Scipy adds modules for image processing and uses the arrays provided by Numpy. This research also uses the Python Imaging Library (PIL) to create Gnuplot files.

This research also uses Gnuplot to visualize the PCA data. Gnuplot is a command line-driven interactive plotting program. This research uses Gnuplot to plot three-dimensional plots of the PCA data.

This research, in addition, uses the Perl programming language to process text input and output. Perl is a programming language originally designed for text manipulation but has matured over the years to be used for more complex tasks such as Web development and database integration. This research uses Perl to process the raw IM messages and to create author writeprints.

3.4 Dataset Descriptions

This research uses two datasets: a personally collected dataset of known authors (Dataset #1) and a publicly available dataset (Dataset #2), with 19 and 105 authors respectively. Although the datasets contains a small number of authors to analyze, the number of suspects in a criminal investigation is usually less than 10 [IBFD2013]. Both datasets contain point-to-point messages communicated between two users. Very few related works have studied the IM domain, and those that have performed IM authorship attribution research, such as [AC2008], have achieved very poor results. Other research, such as [DMCT2011] often use multicast chat logs instead of point-to-point IM. Both datasets used in this research are ground-truth, forensically feasible datasets. For authorship identification ground-truth data contains data for which the author is known. For authorship characterization ground-truth data contains data for which the author demographics are known. [IBFD2013] uses gender and location data for its dataset from blogger.com. However, data from an online source is not verified as ground-truth and could lead to untrustworthy results. Both datasets used in this research are forensically feasible data. Forensically feasible data represents the kind of data found in actual cases, often brief, messy, unedited, and sparse. [Cha2013]

3.4.1 Description of Dataset #1: Known Authors

Dataset #1 contains personal IM conversation logs collected by the Gaim and Adium clients over a three-year period. Although, the authors permitted the analysis and use of the data, anonymity of the authors is protected. Appendix B provides a breakdown of Dataset #1, including author demographics. The author demographics were collected from personal knowledge of each author.

3.4.2 Description of Dataset #2: U.S. Cyberwatch

Dataset #2 contains publicly available data from U.S. Cyberwatch. U.S. Cyberwatch aims to assist law enforcement with the interception, apprehension, and prosecution of online child predators. It also provides training and assistance to law enforcement agencies in addressing this area of cybercrime. The data includes 105 complete IM logs between undercover agents and child predators (all male authors). The logs have been verified via chain of custody. The data also contains metadata such as the suspect's real name, screen name, photograph, age, location, and conviction details. Appendix C provides a breakdown of Dataset #2, including author demographics. The 5 authors with the least number of messages were not used in the experiments in this dissertation because the number of messages was too small for sufficient testing. The U.S. Cyberwatch data is an example of real world cybercrime digital evidence that cyber forensics investigators are obtaining, analyzing, and presenting in court proceedings [Tur2008].

3.4.3 Dataset Limitations

The two datasets present the following limitations that must be considered when processing and analyzing the data:

1. Dataset #2 contains less messages per author, resulting in fewer writeprint instances to analyze. (i.e. the largest conversations size is 90 messages per conversation, whereas Dataset #2 contains 500 messages per conversation as its largest conversation size). Although Dataset #2 contains smaller conversations, it still has an average of 2125 words per author, compared to [CRSBVM2012] that had a dataset with 615 words per author.
2. Dataset #2 contains only one gender (men) and age for authorship characterization analysis.
3. Dataset #1 is imbalanced in terms of authorship characterization due to twice as many females than males in the gender category. Dataset #2 is imbalanced in terms of authorship characterization due to the majority of authors residing in the twenties age category. This imbalance could lead to a bias toward the majority class.

3.5 Summary

This chapter presented the research methodology for creating and analyzing IM behavioral biometrics-based writeprints to assist with cybercrime investigations. The research methodology uses authorship analysis techniques to create an IM-specific stylometric feature set robust enough to show separation between authors and between

author categories.

This chapter presented a detailed explanation of how to prepare the IM data for writeprint creation. Preprocessing software is used to extract the stylometric features into a 356-dimensional vector writeprint including lexical, syntactic, and structural features. Preprocessing also involves normalizing and standardizing the writeprints prior to input into statistical models so large values don't dominate the results of statistical models. Normalization and standardization ensures that features with a wide range of values are less likely to outweigh features with smaller ranges. It allows data on different scales to be compared by bringing them to a common scale, thus allowing the underlying characteristics of the data sets to be compared. The research methodology uses the PCA and MGD statistical methods and graphical software to analyze and evaluate author writeprints from IM conversation logs from two distinct datasets for authorship identification and characterization. This research evaluates the effectiveness of the writeprints using different parameters such as the number of messages used as input. PCA was chosen for the IM writeprint analysis due to the high dimension stylometric feature set. The reduced data is visualized using graphing tools and then input to MGD and STD for analysis. Test writeprints are assessed against MGDs for authors and author categories to determine the identification and characterization probability. MGD was chosen for the IM writeprint analysis due to successful results in other forensics and behavioral biometrics research. The standard deviation measures the spread of distribution of a set of data by calculating distance from the mean of the data.

In addition, this chapter presented details for two datasets: a personally collected

dataset of known authors (Dataset #1) and a publicly available dataset (Dataset #2), with 19 and 105 authors respectively. The research methodology presented in this chapter provides cybercrime investigators a unique tool for investigating IM-assisted cybercrimes. The analysis of IM author writeprints in this research provides a foundation for using behavioral biometrics as a cyber forensics element of criminal investigations.

4. EXPERIMENT RESULTS AND ANALYSIS

The purpose of this research is to create and analyze behavioral biometrics-based instant messaging writeprints to assist in identifying online cyber criminals and collecting digital evidence as part of the criminal investigation. This research uses an IM-specific stylometric feature set to show separation between authors and between author categories. Authorship identification is applied to IM communications to compare and analyze writeprints of various authors. Authorship characterization is applied to IM communications to compare and analyze writeprints of authors based on the sociolinguistic attributes gender, age, and educational background. This research uses Principal Component Analysis (PCA) to reduce the number of necessary dimensions, highlight similarities and differences, and visualize writeprints for comparison. This research uses the multivariate Gaussian distributions (MGD) to determine identification and characterization probabilities for a set of messages across authors and author categories. This section provides a detailed analysis of the results of the IM writeprint analysis conducted on both the Known Authors (Dataset #1) and U.S. Cyberwatch (Dataset #2) datasets.

For each author, IM writeprints are divided into conversations with incrementing number of messages (for example 5, 10, 25, 50, 100, 125, 250, and 500 messages per conversation). As the number of messages for each conversation increases, the number

of writeprint instances for each author decreases. For example, a set of 10,000 messages divided into 250 messages per conversation results in 40 writeprint instances and the same set divided into 50 messages per conversation results in 200 writeprint instances. A high number of writeprint instances results in several data points on the PCA plot, and a low number of writeprint instances results in fewer data points on the PCA plot. Thus, a conversation with a large number of messages contains more data to create a writeprint representative of the author's true writing style, but results in less instances of the writeprint available for analysis. The total number of messages for each author in the dataset ultimately determines the number of writeprint instances for each author.

The coefficients of the first three principal components are plotted, allowing the PCA data to be viewed in 3-dimensions. The PCA data can then be rotated and analyzed at different viewpoints. Data viewed in flat 2-dimensions may appear to overlap, however, viewing the data in a rotational 3-dimensional space reveals separation.

The MGD algorithm processes the writeprints for each set of messages (conversations) for an author or author category under test and the output is analyzed to determine the identification and characterization probability. The aggregate results across all conversations in each test are presented in table matrices.

Standard deviation is used to analyze the spread of the distribution of data within each conversation size. Standard deviation results for the first 5 PCA dimensions for all authors are presented graphically to show the relationship of standard deviation and conversation size.

4.1 Results for Dataset #1, Known Authors

Dataset #1 experiments include 19 authors. For each author, IM writeprints are divided into conversations containing 5, 10, 25, 50, 100, 125, 250, and 500 messages respectively.

4.1.1 Authorship Identification Results

Authorship identification attempts to determine whether an author A_n of a given set of IM messages $\{M_1,\ldots,M_p\}$ is likely to be one of the author suspects $\{A_1,\ldots,A_n\}$. Dataset #1 experiments include 19 authors from which to determine identification.

Figure 4-1 shows Dataset #1 PCA plot results for conversations consisting of 250 messages for each of the 19 authors. This plot does show some separation between the authors.

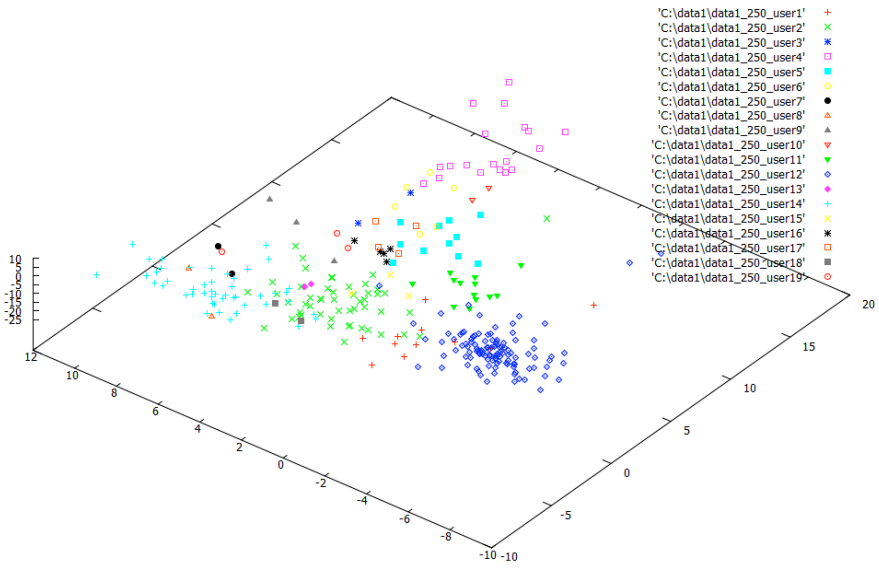


Figure 4-1. Dataset 1, PCA Plot Results, 250 Messages, All 19 Authors

Table 4-1 through

Table 4-6 show the MGD results for conversation sizes of 5, 10, 25, 50, 100, and 125 messages respectively for all 19 authors. Using 100 messages per conversation as input, MGD identifies conversations as the correct author for 17 of the 19 authors, with probabilities ranging from 71.51% to 100%. Using 125 messages per conversation as input, MGD identifies conversations as the correct author for 17 of the 19 authors, with probabilities ranging from 82.14% to 100%. The tables show a significant increase in identification probabilities as the number of messages per conversation increase. Figure 4-2 shows the relationship between the identification probability and number of messages per conversation.

Table 4-1. Dataset 1, MGD Results, 5 Messages, All 19 Authors (shown in %)

Size=5	P(A1|x)	P(A2|x)	P(A3|x)	P(A4|x)	P(A5|x)	P(A6|x)	P(A7|x)	P(A8|x)	P(A9|x)	P(A10|x)	P(A11|x)	P(A12|x)	P(A13|x)	P(A14|x)	P(A15|x)	P(A16|x)	P(A17|x)	P(A18|x)	P(A19|x)
A1	**18.41**	13.03	2.00	0.29	1.55	1.83	3.67	10.16	4.14	0.04	3.66	11.61	3.13	9.78	1.47	0.16	2.50	9.48	3.09
A2	**13.88**	12.77	2.98	0.66	2.36	2.70	4.06	9.37	5.39	0.12	4.25	9.71	3.37	10.38	2.02	0.46	3.52	8.11	3.91
A3	7.91	8.73	6.97	2.81	7.13	7.69	2.39	3.80	**10.36**	1.17	3.10	5.99	3.37	6.09	3.78	1.97	7.74	1.88	7.13
A4	4.18	5.16	6.89	**14.45**	11.16	13.72	0.45	0.67	8.41	5.31	3.37	4.52	1.05	2.42	2.27	3.14	6.89	0.10	5.82
A5	7.86	8.13	6.88	4.54	9.53	8.90	1.46	1.69	**10.26**	3.07	4.05	6.76	1.72	4.39	2.40	2.51	8.57	0.68	6.60
A6	6.69	7.15	7.72	5.88	9.86	10.72	0.97	1.31	**11.64**	2.26	4.05	6.19	1.41	3.78	2.42	2.57	8.41	0.37	6.60
A7	12.43	12.43	4.24	0.24	1.73	1.98	7.05	11.52	4.96	0.03	1.84	5.18	4.29	**12.76**	1.36	0.42	3.45	9.01	5.08
A8	13.18	12.76	2.82	0.30	1.06	1.60	4.67	**15.70**	4.19	0.02	1.95	5.87	4.10	12.81	1.12	0.19	2.25	11.27	4.13
A9	7.86	8.32	7.19	1.78	5.81	5.68	1.60	5.07	**13.16**	0.79	2.44	6.10	3.76	6.06	2.15	1.84	9.31	3.83	7.22
A10	4.23	5.21	6.61	6.63	11.83	11.95	0.46	0.59	9.15	**20.63**	1.66	3.37	0.37	1.92	0.81	0.53	7.62	0.27	6.16
A11	**13.12**	11.85	3.62	2.19	5.55	6.32	2.05	3.13	7.63	0.38	8.22	12.21	2.11	6.63	2.74	0.81	4.99	2.50	3.95
A12	**16.92**	13.51	2.73	0.94	3.60	3.65	2.89	4.31	5.94	0.16	6.68	13.30	2.74	8.37	2.91	0.55	4.15	2.97	3.68
A13	10.49	**11.58**	4.57	1.00	2.78	3.31	5.00	8.02	6.60	0.20	2.62	7.06	5.81	10.38	2.55	0.70	4.01	6.58	6.76
A14	**13.03**	12.96	3.09	0.28	1.54	2.06	4.98	11.73	4.82	0.04	2.89	7.76	3.86	11.96	1.50	0.21	2.77	10.25	4.26
A15	8.90	**9.88**	5.46	2.31	5.39	6.40	2.43	3.98	8.91	0.39	5.31	7.87	3.88	7.26	4.57	1.90	6.42	2.74	6.00
A16	4.90	6.63	6.89	5.17	8.36	7.94	0.91	2.97	10.96	1.75	3.53	5.06	2.79	4.55	3.52	4.54	**10.64**	1.35	7.53
A17	7.67	8.36	6.48	2.36	6.07	5.54	2.21	5.36	**11.35**	1.24	2.15	5.56	3.38	6.55	1.99	2.24	10.82	3.20	7.47
A18	**14.80**	13.34	1.99	0.10	1.04	1.48	4.32	12.51	4.28	0.00	2.35	8.57	3.52	12.11	1.09	0.05	2.44	12.64	3.35
A19	8.98	9.25	5.87	1.91	4.81	5.05	2.88	7.39	**8.98**	1.08	2.68	5.83	4.29	7.80	2.27	1.75	6.94	5.04	7.21

Table 4-2. Dataset 1, MGD Results, 10 messages, All 19 Authors (shown in %)

Size=10	P(A1|x)	P(A2|x)	P(A3|x)	P(A4|x)	P(A5|x)	P(A6|x)	P(A7|x)	P(A8|x)	P(A9|x)	P(A10|x)	P(A11|x)	P(A12|x)	P(A13|x)	P(A14|x)	P(A15|x)	P(A16|x)	P(A17|x)	P(A18|x)	P(A19|x)
A1	**27.47**	16.62	0.25	0.03	0.32	0.47	2.81	6.67	1.39	0.00	2.69	16.04	2.07	9.29	1.05	0.01	0.62	10.41	1.79
A2	16.38	**19.46**	0.69	0.13	0.72	0.93	3.97	7.32	2.69	0.00	3.30	12.31	3.05	12.86	1.92	0.11	1.54	9.41	3.20
A3	2.31	7.32	8.77	1.73	11.32	9.87	1.10	0.68	**14.53**	0.58	1.88	3.57	2.29	2.75	3.16	1.92	12.28	0.31	13.64
A4	0.64	1.81	6.86	**28.71**	15.73	14.16	0.06	0.05	8.23	1.52	1.94	1.14	0.23	0.26	1.53	2.47	8.57	0.00	6.09
A5	2.34	5.09	9.43	4.75	**15.28**	11.61	0.36	0.18	14.24	1.53	2.52	3.37	0.83	1.08	1.89	2.76	11.92	0.17	10.68
A6	1.82	4.87	9.37	6.37	15.03	**15.53**	0.56	0.16	13.56	0.08	4.18	2.92	1.05	1.16	1.73	1.84	9.57	0.01	10.19
A7	11.76	17.75	0.73	0.02	0.38	0.50	10.60	10.08	1.92	0.00	0.40	3.32	5.51	**18.05**	0.49	0.18	1.28	12.43	4.60
A8	14.90	17.05	0.07	0.00	0.05	0.18	4.35	**20.49**	0.36	0.00	0.30	3.39	2.31	18.21	0.09	0.00	0.20	16.51	1.54
A9	4.18	7.43	6.57	1.43	6.52	5.07	0.67	0.88	**22.89**	0.20	1.06	5.32	2.38	2.56	1.17	1.76	16.29	2.02	11.60
A10	0.10	0.52	1.84	5.04	17.39	5.59	0.00	0.03	8.77	**41.08**	0.00	0.06	0.01	0.04	0.04	0.17	13.71	0.00	5.60
A11	13.59	17.70	1.49	0.73	3.41	5.23	1.50	1.00	5.40	0.00	13.97	**18.64**	1.49	5.34	2.19	0.26	2.57	1.33	4.16
A12	**23.13**	20.36	0.47	0.11	0.93	1.14	1.88	1.51	2.97	0.00	6.63	22.23	1.86	8.18	2.41	0.08	1.54	1.86	2.71
A13	10.68	**16.28**	2.16	0.15	1.23	2.18	5.32	5.44	4.11	0.00	2.21	8.33	6.81	12.15	2.45	0.40	2.71	8.46	8.93
A14	15.04	**19.58**	0.35	0.01	0.20	0.36	5.40	10.86	1.11	0.00	1.52	8.46	3.12	16.03	0.96	0.03	0.60	13.31	3.05
A15	5.84	**14.60**	3.20	0.97	4.00	4.62	1.68	3.13	7.59	0.01	5.67	8.29	6.09	8.35	7.57	1.45	6.68	1.99	8.26
A16	0.68	3.14	7.31	3.98	12.29	8.36	0.21	0.36	14.07	0.27	1.73	1.46	2.00	1.13	4.06	6.60	**20.35**	0.27	11.72
A17	2.52	6.01	5.82	1.20	7.25	4.46	1.02	1.09	21.70	0.23	0.39	2.38	1.96	2.50	1.08	3.47	**22.77**	1.46	12.68
A18	17.83	**17.92**	0.24	0.00	0.10	0.21	4.13	12.82	2.00	0.00	1.37	7.18	1.91	14.56	0.61	0.00	0.78	16.17	2.17
A19	5.74	10.91	4.85	1.65	5.90	4.29	4.23	2.00	**12.69**	0.26	1.10	5.56	3.91	6.61	1.82	1.70	11.34	2.88	12.57

Table 4-3. Dataset 1, MGD Results, 25 messages, All 19 Authors (shown in %)

Size=25	P(A1|x)	P(A2|x)	P(A3|x)	P(A4|x)	P(A5|x)	P(A6|x)	P(A7|x)	P(A8|x)	P(A9|x)	P(A10|x)	P(A11|x)	P(A12|x)	P(A13|x)	P(A14|x)	P(A15|x)	P(A16|x)	P(A17|x)	P(A18|x)	P(A19|x)
A1	**42.29**	22.48	0.00	0.00	0.00	0.00	0.80	2.63	0.05	0.00	1.17	20.05	0.26	6.99	0.13	0.00	0.00	3.09	0.08
A2	14.57	**34.77**	0.13	0.00	0.26	0.04	2.64	5.76	1.04	0.00	2.01	11.73	2.78	15.63	1.68	0.01	0.14	4.68	2.11
A3	0.02	0.90	13.48	0.52	**19.24**	4.05	0.20	0.03	22.04	0.17	0.00	0.01	0.85	0.09	0.98	5.47	7.55	0.01	24.38
A4	0.00	0.02	5.41	**62.48**	18.83	7.91	0.00	0.00	0.94	0.10	0.00	0.00	0.00	0.00	0.72	1.25	1.73	0.00	0.59
A5	0.06	1.48	10.43	1.13	**34.93**	5.01	0.00	0.00	17.14	0.34	0.49	0.13	0.23	0.03	0.95	2.24	16.20	0.00	9.23
A6	0.10	1.59	21.20	3.73	20.19	**25.22**	0.00	0.00	7.99	0.04	1.90	0.41	0.00	0.01	0.05	5.10	4.44	0.00	8.03
A7	7.17	19.41	0.58	0.00	0.00	0.00	23.71	9.76	0.24	0.00	0.00	0.14	1.95	**26.41**	0.00	0.00	0.01	5.41	5.22
A8	9.17	16.00	0.00	0.00	0.00	0.00	7.15	**39.56**	0.00	0.00	0.00	0.04	0.43	23.06	0.01	0.00	0.00	3.70	0.87
A9	0.24	2.71	5.49	0.00	6.43	0.54	0.23	0.19	**41.37**	0.00	0.06	0.33	1.35	0.19	0.06	0.49	27.49	0.09	12.73
A10	0.00	0.01	1.03	1.04	6.72	0.54	0.00	0.00	0.47	**86.00**	0.00	0.00	0.00	0.00	0.00	0.00	3.40	0.00	0.78
A11	8.28	23.87	0.02	0.03	0.63	0.90	0.00	0.04	0.96	0.00	27.71	**30.51**	1.37	1.01	3.83	0.00	0.09	0.01	0.76
A12	26.76	21.86	0.04	0.00	0.12	0.01	0.08	0.13	0.34	0.00	3.97	**40.97**	0.46	2.85	1.59	0.06	0.26	0.08	0.41
A13	1.58	**31.07**	0.68	0.00	0.60	0.10	4.20	3.55	1.61	0.00	0.33	2.09	20.87	13.10	6.22	0.19	1.18	4.52	8.13
A14	8.96	**28.10**	0.03	0.00	0.01	0.00	7.34	11.08	0.25	0.00	0.16	3.05	2.61	27.13	0.45	0.00	0.03	9.17	1.62
A15	1.92	20.45	0.49	0.22	3.51	1.06	0.94	0.26	2.42	0.00	7.23	5.95	13.42	7.21	**25.35**	0.08	0.87	0.05	8.54
A16	0.01	0.35	7.58	1.03	19.23	4.01	0.00	0.00	13.27	0.00	0.07	0.06	0.49	0.01	2.95	11.53	**24.89**	0.00	14.53
A17	0.02	0.90	5.74	0.01	6.62	0.27	0.48	0.02	27.22	0.03	0.08	0.11	1.29	0.10	0.08	2.35	**34.45**	0.07	20.12
A18	12.09	**21.84**	0.03	0.00	0.00	0.00	1.96	19.10	0.09	0.00	0.01	4.33	1.93	16.78	0.08	0.00	0.00	20.33	1.44
A19	0.47	12.75	7.02	0.09	5.34	1.13	5.45	0.93	9.38	0.00	0.00	0.06	2.21	8.58	0.21	5.06	14.60	4.73	**21.99**

Table 4-4. Dataset 1, MGD Results, 50 messages, All 19 Authors (shown in %)

Size=50	P(A1|x)	P(A2|x)	P(A3|x)	P(A4|x)	P(A5|x)	P(A6|x)	P(A7|x)	P(A8|x)	P(A9|x)	P(A10|x)	P(A11|x)	P(A12|x)	P(A13|x)	P(A14|x)	P(A15|x)	P(A16|x)	P(A17|x)	P(A18|x)	P(A19|x)
A1	**65.96**	13.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.27	18.44	0.00	1.67	0.00	0.00	0.00	0.28	0.00
A2	7.40	**54.47**	0.00	0.00	0.03	0.00	1.76	1.31	0.12	0.00	0.89	6.80	0.48	20.56	0.30	0.00	0.02	5.25	0.60
A3	0.00	0.02	19.82	0.02	**30.84**	8.76	0.00	0.00	0.36	0.00	0.00	0.00	0.00	0.00	0.00	8.77	4.12	0.00	27.29
A4	0.00	0.00	0.01	**99.96**	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A5	0.00	0.13	13.42	0.03	**45.83**	13.69	0.00	0.00	4.53	0.00	0.05	0.00	0.00	0.00	0.06	2.01	17.43	0.00	2.82
A6	0.00	0.07	8.95	0.36	12.94	**56.17**	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	11.61	7.96	0.00	0.94
A7	0.05	15.02	0.00	0.00	0.00	0.00	27.96	0.03	0.00	0.00	0.00	0.00	0.00	**50.15**	0.00	0.00	0.00	2.09	4.69
A8	2.61	5.53	0.00	0.00	0.00	0.00	0.96	35.62	0.00	0.00	0.00	0.00	0.00	**52.47**	0.00	0.00	0.00	2.80	0.00
A9	0.00	0.40	0.00	0.00	0.24	0.00	0.00	0.00	**98.24**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.87	0.00	0.25
A10	0.00	0.00	0.00	0.00	0.31	0.00	0.00	0.00	0.15	**98.59**	0.00	0.00	0.00	0.00	0.00	0.00	0.95	0.00	0.00
A11	0.87	19.11	0.00	0.00	0.03	0.04	0.00	0.00	0.04	0.00	**62.32**	16.91	0.00	0.13	0.55	0.00	0.00	0.00	0.00
A12	18.80	8.05	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	1.20	**71.76**	0.00	0.15	0.02	0.00	0.00	0.00	0.00
A13	0.01	33.73	0.01	0.00	0.00	0.00	0.09	0.02	0.00	0.00	0.00	0.87	**42.44**	5.87	0.30	0.00	0.00	11.93	4.75
A14	2.82	24.84	0.00	0.00	0.00	0.00	6.90	4.02	0.01	0.00	0.00	1.51	0.40	**54.42**	0.00	0.00	0.00	4.45	0.62
A15	0.00	18.84	0.74	0.09	0.61	0.38	0.33	0.00	0.05	0.00	2.54	0.24	20.08	1.18	**51.53**	0.08	0.00	0.39	2.93
A16	0.00	0.02	11.41	0.05	14.97	1.59	0.00	0.00	0.29	0.00	0.00	0.00	0.00	0.00	0.02	14.69	24.79	0.00	**32.17**
A17	0.00	0.02	0.76	0.00	4.14	0.02	0.00	0.00	9.22	0.00	0.00	0.00	0.00	0.00	0.00	0.36	**75.48**	0.00	9.99
A18	21.04	29.38	0.00	0.00	0.00	0.00	0.33	3.43	0.00	0.00	0.00	1.68	0.11	3.95	0.13	0.00	0.00	**39.78**	0.15
A19	0.00	0.30	0.81	0.00	6.82	0.31	4.49	0.00	0.94	0.00	0.00	0.00	0.79	2.81	0.07	4.50	34.87	0.23	**43.04**

Table 4-5. Dataset 1, MGD Results, 100 messages, All 19 Authors (shown in %)

Size= 100	P(A1|x)	P(A2|x)	P(A3|x)	P(A4|x)	P(A5|x)	P(A6|x)	P(A7|x)	P(A8|x)	P(A9|x)	P(A10|x)	P(A11|x)	P(A12|x)	P(A13|x)	P(A14|x)	P(A15|x)	P(A16|x)	P(A17|x)	P(A18|x)	P(A19|x)
A1	**84.58**	7.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.98	0.00	0.01	0.00	0.00	0.00	0.00	0.00
A2	1.54	**82.26**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	1.37	0.00	14.51	0.00	0.00	0.00	0.00	0.15
A3	0.00	0.00	15.69	0.00	**82.96**	0.78	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.05	0.00	0.33
A4	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A5	0.00	0.00	0.00	0.00	**73.32**	0.58	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	26.10	0.00	0.00
A6	0.00	0.00	0.00	0.00	0.90	**99.06**	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00
A7	0.00	20.00	0.00	0.00	0.00	0.00	18.74	0.00	0.00	0.00	0.00	0.00	0.00	**61.25**	0.00	0.00	0.00	0.00	0.01
A8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A9	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	**97.07**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.91	0.00	0.00
A10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A11	0.02	8.47	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**91.29**	0.21	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A12	6.81	1.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15	**91.04**	0.00	0.01	0.00	0.00	0.00	0.00	0.00
A13	0.00	16.88	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**76.99**	0.15	0.01	0.00	0.00	5.11	0.85
A14	0.56	18.72	0.00	0.00	0.00	0.00	0.28	0.00	0.00	0.00	0.00	0.11	0.00	**80.32**	0.00	0.00	0.00	0.00	0.01
A15	0.00	1.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	**98.81**	0.00	0.00	0.00	0.00
A16	0.00	0.00	0.00	0.00	1.60	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**81.04**	0.04	0.00	17.28
A17	0.00	0.00	0.00	0.00	0.68	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**99.31**	0.00	0.01
A18	0.53	27.66	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	**71.51**	0.05
A19	0.00	2.10	0.00	0.00	3.73	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.39	0.00	0.00	0.09	0.00	**92.67**

Table 4-6. Dataset 1, MGD Results, 125 messages, All 19 Authors (shown in %)

Size= 125	P(A1|x)	P(A2|x)	P(A3|x)	P(A4|x)	P(A5|x)	P(A6|x)	P(A7|x)	P(A8|x)	P(A9|x)	P(A10|x)	P(A11|x)	P(A12|x)	P(A13|x)	P(A14|x)	P(A15|x)	P(A16|x)	P(A17|x)	P(A18|x)	P(A19|x)
A1	**87.85**	0.64	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	11.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A2	0.23	**82.14**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	1.15	0.00	16.31	0.00	0.00	0.00	0.00	0.11
A3	0.00	0.00	**99.67**	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.26
A4	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A5	0.00	0.00	0.00	0.00	49.79	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**49.91**	0.00	0.16
A6	0.00	0.00	0.00	0.00	0.01	**99.73**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00
A7	0.00	5.90	0.00	0.00	0.00	0.00	**56.68**	0.00	0.00	0.00	0.00	0.00	0.00	37.42	0.00	0.00	0.00	0.00	0.00
A8	0.22	0.02	0.00	0.00	0.00	0.00	0.00	**95.44**	0.00	0.00	0.00	0.00	0.00	4.32	0.00	0.00	0.00	0.00	0.00
A9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A11	0.00	4.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**95.66**	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A12	7.04	2.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	**90.85**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00
A14	0.08	16.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	**82.95**	0.00	0.00	0.00	0.00	0.06
A15	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**99.75**	0.00	0.00	0.00	0.00
A16	0.00	0.00	0.00	0.00	1.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**85.28**	0.84	0.00	12.60
A17	0.00	0.00	0.00	0.00	0.63	0.00	0.00	0.00	1.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**97.21**	0.00	0.23
A18	0.00	1.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	**98.89**	0.00
A19	0.00	0.05	0.00	0.00	0.67	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	3.13	0.00	0.00	**96.08**

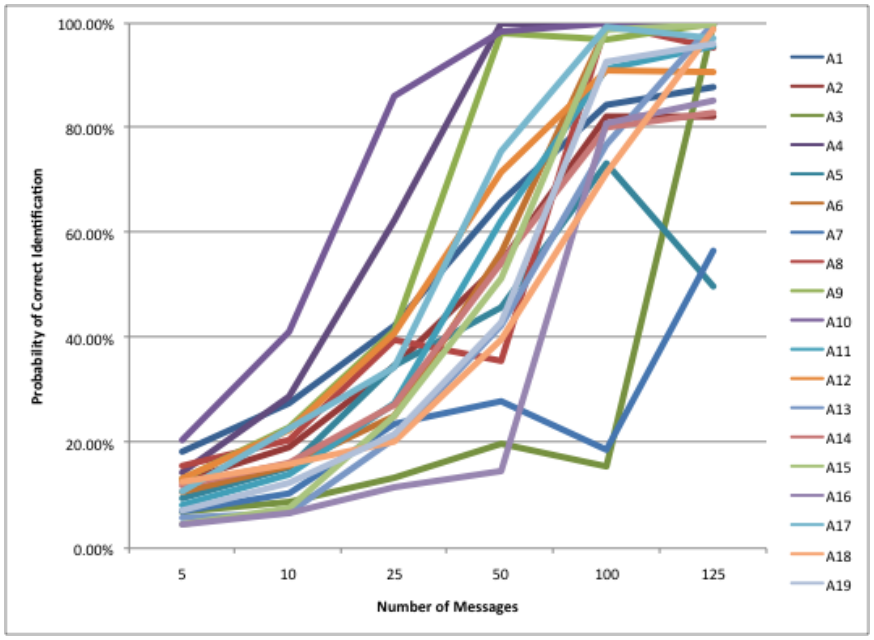


Figure 4-2. Dataset 1, Identification Probability vs. Number of Messages, All 19 Authors

The authorship identification probability is used to determine the error of the multivariate Gaussian distribution by assessing writeprint false positives. The likelihood, P(x|Author), of the author of the writperint is used as a minimum threshold. If another author has a higher likelihood, this is a false positive. Dataset #1 analysis for all 19 authors achieved less than 20% error for most authors using 125 messages per conversation. Figure 4-3 shows that as the conversation size increases, the error rate decreases.

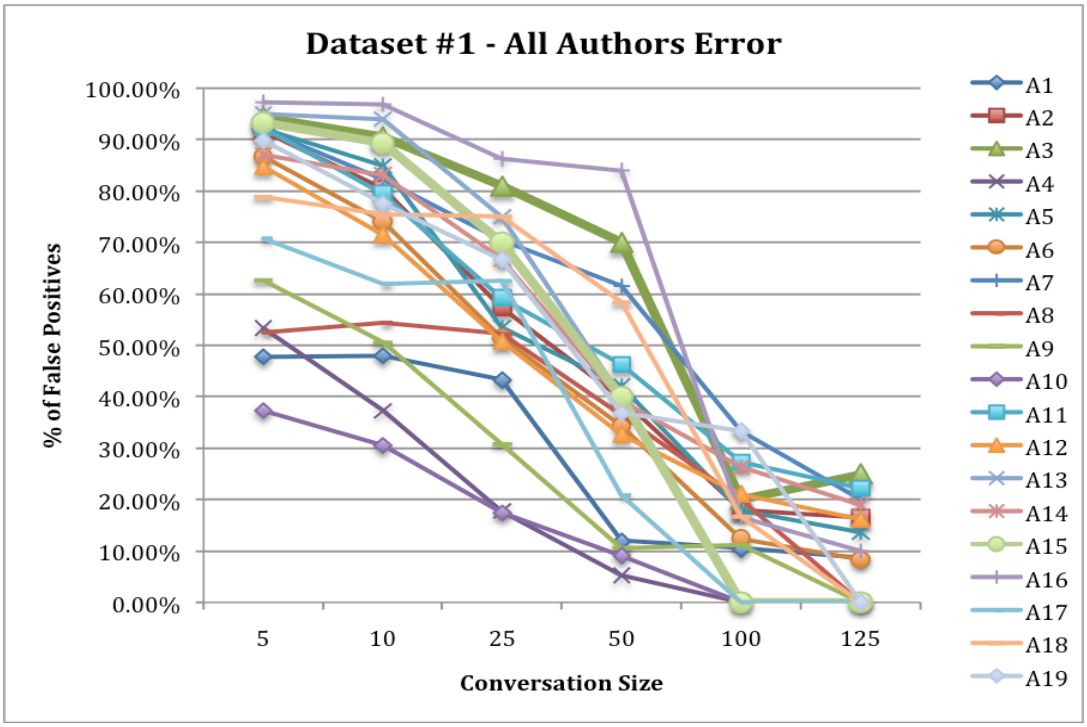


Figure 4-3. Dataset #1, All 19 Authors Error

Figure 4-4 through Figure 4-8 show PCA plots of Dataset #1 author writeprints broken down into 6, 6, and 7 authors respectively. Figure 4-4 shows Dataset #1 PCA plot results for conversations consisting of 250 messages for Authors A1-A6. This plot shows separate groupings for each author.

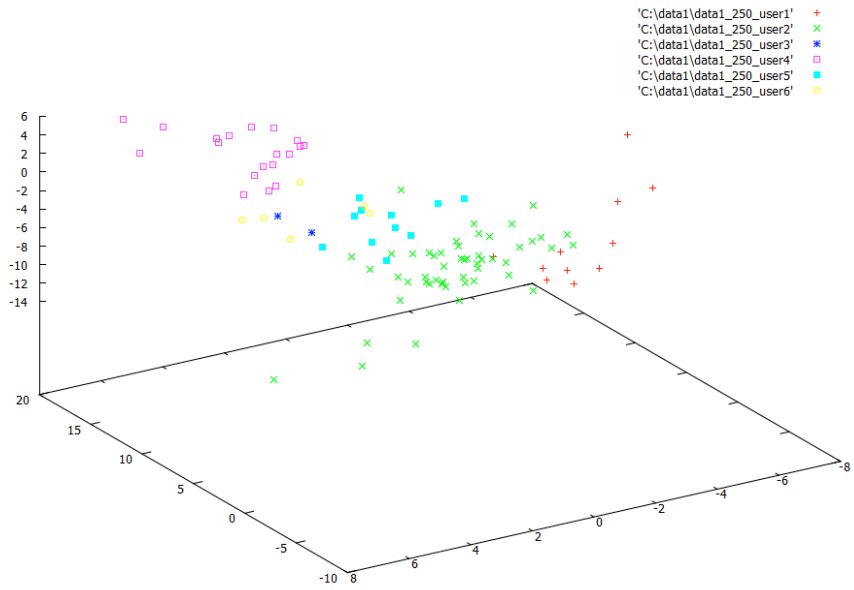


Figure 4-4. Dataset 1, PCA Plot Results, 250 Messages, Authors A1-A6

Table 4-7 through

Table 4-12 show the MGD results for conversation sizes of 5, 10, 25, 50, 100, and 125 messages respectively for Authors A1-A6. Using 100 messages per conversation as input, MGD identifies conversations as the correct author for 5 of the 6 authors, with probability ranging from 91.93% to 100%. Using 125 messages per conversation as input, MGD identifies conversations as the correct author for all 6 authors, with probability over 99% across all authors. The tables show a significant increase in identification probability as the number of messages per conversation increase. Figure 4-5 shows the relationship between the identification probability and number of messages per conversation.

Table 4-7. Dataset 1, MGD Results, 5 Messages, Authors A1-A6

Size=5	P(A1|x)	P(A2|x)	P(A3|x)	P(A4|x)	P(A5|x)	P(A6|x)
A1	**49.62%**	35.11%	5.38%	0.77%	4.19%	4.92%
A2	**39.25%**	36.14%	8.44%	1.86%	6.67%	7.63%
A3	19.18%	**21.16%**	16.90%	6.82%	17.29%	18.65%
A4	7.52%	9.29%	12.41%	**26.00%**	20.09%	24.69%
A5	17.14%	17.73%	15.02%	9.89%	**20.80%**	19.41%
A6	13.93%	14.90%	16.07%	12.24%	20.53%	**22.33%**

Table 4-8. Dataset 1, MGD Results, 10 Messages, Authors A1-A6

Size=10	P(A1|x)	P(A2|x)	P(A3|x)	P(A4|x)	P(A5|x)	P(A6|x)
A1	**60.84%**	36.80%	0.55%	0.07%	0.72%	1.03%
A2	42.74%	**50.80%**	1.80%	0.34%	1.89%	2.43%
A3	5.60%	17.70%	21.23%	4.18%	27.40%	**23.90%**
A4	0.94%	2.67%	10.11%	**42.27%**	23.16%	20.85%
A5	4.83%	10.50%	19.44%	9.79%	**31.51%**	23.93%
A6	3.44%	9.19%	17.69%	12.02%	28.36%	**29.30%**

Table 4-9. Dataset 1, MGD Results, 25 Messages, Authors A1-A6

Size=25	P(A1|x)	P(A2|x)	P(A3|x)	P(A4|x)	P(A5|x)	P(A6|x)
A1	**65.29%**	34.71%	0.00%	0.00%	0.00%	0.00%
A2	29.27%	**69.85%**	0.27%	0.00%	0.53%	0.08%
A3	0.05%	2.37%	35.28%	1.36%	**50.35%**	10.60%
A4	0.00%	0.02%	5.72%	**66.01%**	19.90%	8.36%
A5	0.12%	2.78%	19.66%	2.13%	**65.87%**	9.44%
A6	0.13%	2.21%	29.43%	5.18%	28.03%	**35.01%**

Table 4-10. Dataset 1, MGD Results, 50 Messages, Authors A1-A6

Size=50	P(A1|x)	P(A2|x)	P(A3|x)	P(A4|x)	P(A5|x)	P(A6|x)
A1	**83.15%**	16.85%	0.00%	0.00%	0.00%	0.00%
A2	11.95%	**87.99%**	0.00%	0.00%	0.05%	0.00%
A3	0.00%	0.03%	33.34%	0.03%	**51.87%**	14.73%
A4	0.00%	0.00%	0.01%	**99.96%**	0.02%	0.01%
A5	0.00%	0.18%	18.36%	0.04%	**62.69%**	18.72%
A6	0.00%	0.09%	11.40%	0.46%	16.48%	**71.56%**

Table 4-11. Dataset 1, MGD Results, 100 Messages, Authors A1-A6

Size=100	P(A1|x)	P(A2|x)	P(A3|x)	P(A4|x)	P(A5|x)	P(A6|x)
A1	**91.93%**	8.07%	0.00%	0.00%	0.00%	0.00%
A2	1.84%	**98.16%**	0.00%	0.00%	0.00%	0.00%
A3	0.00%	0.00%	15.78%	0.00%	**83.44%**	0.79%
A4	0.00%	0.00%	0.00%	**100.00%**	0.00%	0.00%
A5	0.00%	0.00%	0.00%	0.00%	**99.21%**	0.79%
A6	0.00%	0.00%	0.00%	0.00%	0.90%	**99.09%**

Table 4-12. Dataset 1, MGD Results, 125 Messages, Authors A1-A6

Size=125	P(A1|x)	P(A2|x)	P(A3|x)	P(A4|x)	P(A5|x)	P(A6|x)
A1	**99.28%**	0.72%	0.00%	0.00%	0.00%	0.00%
A2	0.28%	**99.72%**	0.00%	0.00%	0.00%	0.00%
A3	0.00%	0.00%	**99.93%**	0.00%	0.00%	0.06%
A4	0.00%	0.00%	0.00%	**100.00%**	0.00%	0.00%
A5	0.00%	0.00%	0.00%	0.00%	**99.72%**	0.28%
A6	0.00%	0.00%	0.00%	0.00%	0.01%	**99.99%**

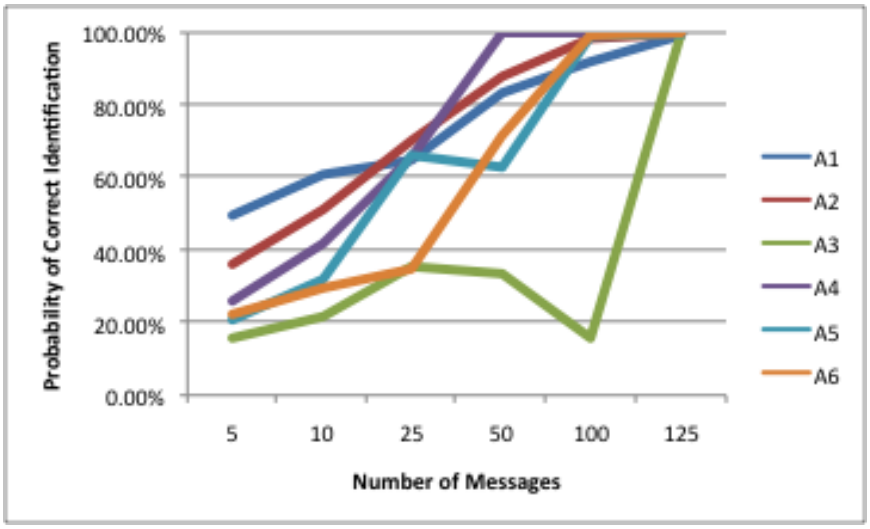


Figure 4-5. Dataset 1, Identification Probability vs. Number of Messages, Authors A1-A6

Figure 4-6 shows Dataset #1 PCA plot results for conversations consisting of 250 messages for Authors A7-A12. This plot shows separate groupings for each author.

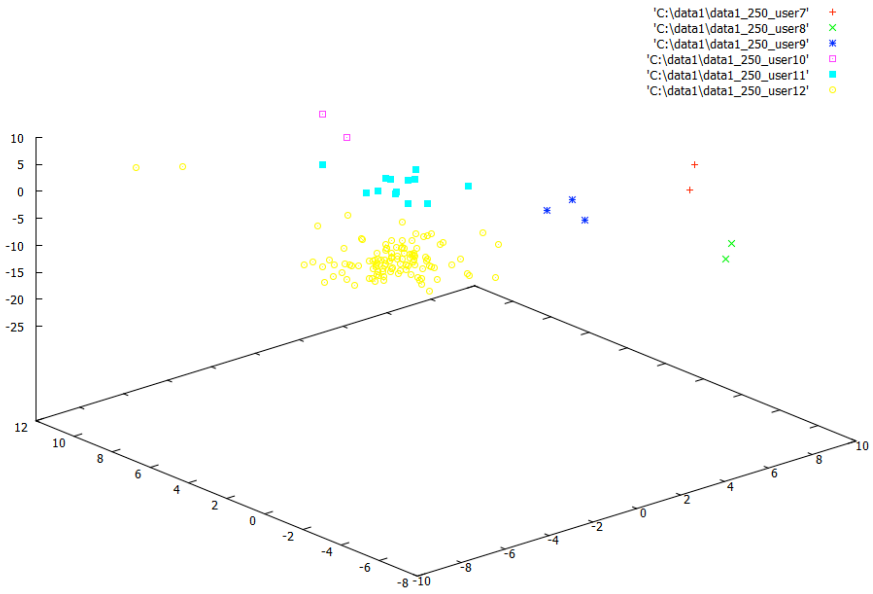


Figure 4-6. Dataset 1, PCA Plot Results, 250 messages, Authors A7-A12

Table 4-13 through

Table 4-18 show the MGD results for conversation sizes of 5, 10, 25, 50, 100, and 125 messages respectively for Authors A7-A12. Using 100 messages per conversation as input, MGD identifies conversations as the correct author for all 6 authors, with probability ranging from 99.77% to 100%. Using 125 messages per conversation as input, MGD identifies conversations as the correct author for all 6 authors, with increasing probability across authors. The tables show a significant increase in identification probability as the number of messages per conversation increase. Figure 4-7 shows the relationship between the identification probability and number of messages per conversation.

Table 4-13. Dataset 1, MGD Results, 5 Messages, Authors A7-A12

Size=5	P(A7|x)	P(A8|x)	P(A9|x)	P(A10|x)	P(A11|x)	P(A12|x)
A7	23.05%	**37.67%**	16.22%	0.10%	6.02%	16.94%
A8	14.42%	**48.43%**	12.94%	0.07%	6.02%	18.12%
A9	5.49%	17.39%	**45.13%**	2.72%	8.37%	20.90%
A10	1.29%	1.65%	25.50%	**57.53%**	4.62%	9.41%
A11	6.10%	9.32%	22.68%	1.13%	24.45%	**36.32%**
A12	8.69%	12.96%	17.84%	0.47%	20.07%	**39.97%**

Table 4-14. Dataset 1, MGD Results, 10 Messages, Authors A7-A12

Size=10	P(A7|x)	P(A8|x)	P(A9|x)	P(A10|x)	P(A11|x)	P(A12|x)
A7	**40.29%**	38.29%	7.28%	0.00%	1.53%	12.61%
A8	15.05%	**70.92%**	1.25%	0.00%	1.04%	11.74%

A9	2.15%	2.85%	**73.79%**	0.65%	3.41%	17.14%
A10	0.00%	0.05%	17.56%	**82.27%**	0.01%	0.11%
A11	3.71%	2.46%	13.34%	0.00%	34.48%	**46.01%**
A12	5.33%	4.28%	8.43%	0.00%	18.84%	**63.12%**

Table 4-15. Dataset 1, MGD Results, 25 Messages, Authors A7-A12

Size=25	**P(A7|x)**	**P(A8|x)**	**P(A9|x)**	**P(A10|x)**	**P(A11|x)**	**P(A12|x)**
A7	**70.06%**	28.83%	0.70%	0.00%	0.00%	0.41%
A8	15.30%	**84.60%**	0.00%	0.00%	0.00%	0.09%
A9	0.54%	0.45%	**98.09%**	0.00%	0.14%	0.79%
A10	0.00%	0.00%	0.54%	**99.46%**	0.00%	0.00%
A11	0.01%	0.06%	1.61%	0.00%	46.79%	**51.53%**
A12	0.18%	0.29%	0.75%	0.00%	8.73%	**90.05%**

Table 4-16. Dataset 1, MGD Results, 50 Messages, Authors A7-A12

Size=50	**P(A7|x)**	**P(A8|x)**	**P(A9|x)**	**P(A10|x)**	**P(A11|x)**	**P(A12|x)**
A7	**99.88%**	0.12%	0.00%	0.00%	0.00%	0.00%
A8	2.63%	**97.37%**	0.00%	0.00%	0.00%	0.00%
A9	0.00%	0.00%	**100.00%**	0.00%	0.00%	0.00%
A10	0.00%	0.00%	0.16%	**99.84%**	0.00%	0.00%
A11	0.00%	0.00%	0.06%	0.00%	**78.61%**	21.33%
A12	0.00%	0.00%	0.01%	0.00%	1.65%	**98.34%**

Table 4-17. Dataset 1, MGD Results, 100 Messages, Authors A7-A12

Size=100	**P(A7|x)**	**P(A8|x)**	**P(A9|x)**	**P(A10|x)**	**P(A11|x)**	**P(A12|x)**
A7	**100.00%**	0.00%	0.00%	0.00%	0.00%	0.00%
A8	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%
A9	0.00%	0.00%	**100.00%**	0.00%	0.00%	0.00%
A10	0.00%	0.00%	0.00%	**100.00%**	0.00%	0.00%
A11	0.00%	0.00%	0.00%	0.00%	**99.77%**	0.23%
A12	0.00%	0.00%	0.00%	0.00%	0.17%	**99.83%**

Table 4-18. Dataset 1, MGD Results, 125 Messages, Authors A7-A12

Size=125	P(A7|x)	P(A8|x)	P(A9|x)	P(A10|x)	P(A11|x)	P(A12|x)
A7	**100.00%**	0.00%	0.00%	0.00%	0.00%	0.00%
A8	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%
A9	0.00%	0.00%	**100.00%**	0.00%	0.00%	0.00%
A10	0.00%	0.00%	0.00%	**100.00%**	0.00%	0.00%
A11	0.00%	0.00%	0.00%	0.00%	**99.81%**	0.19%
A12	0.00%	0.00%	0.00%	0.00%	0.04%	**99.96%**

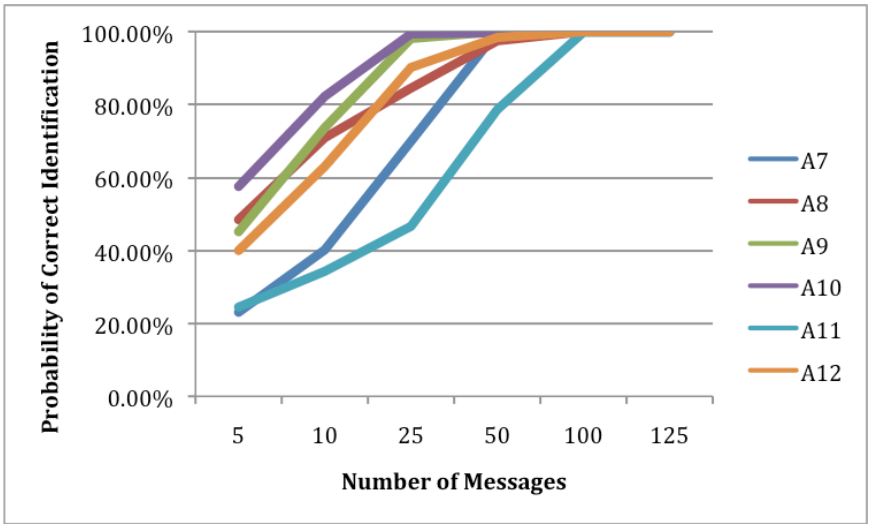


Figure 4-7. Dataset 1, Identification Probability vs. Number of Messages, Authors A7-A12

Figure 4-8 shows Dataset #1 PCA plot results for conversations consisting of 250 messages for Authors A13-A19. This plot shows separate groupings for each author.

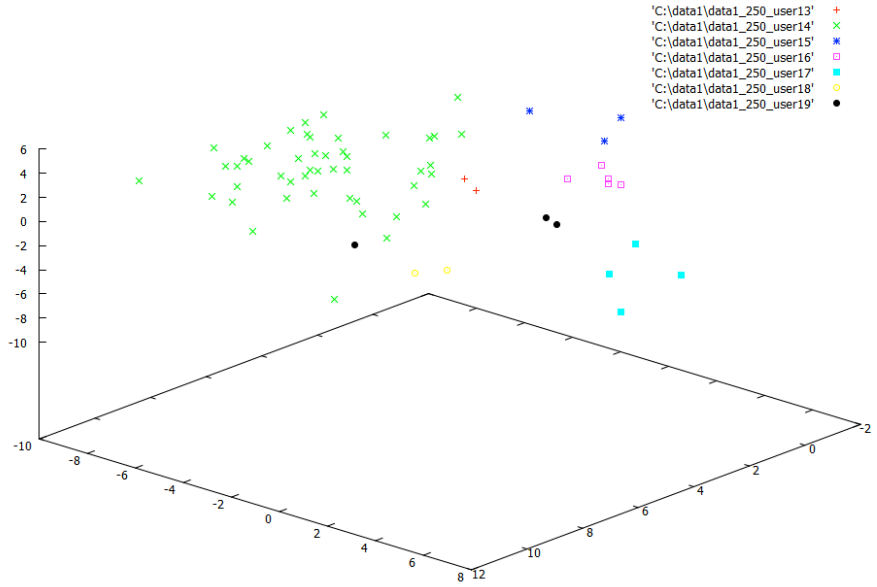


Figure 4-8. Dataset 1, PCA Plot Results, 250 Messages, Authors A13-A19

Table 4-19 through

Table 4-24 show the MGD results for conversation sizes of 5, 10, 25, 50, 100, and 125 messages respectively for Authors A13-A19. Using 100 messages per conversation as input, MGD identifies conversations as the correct author for all 7 authors, with probability ranging from 82.39% to 100%. Using 125 messages per conversation as input, MGD identifies conversations as the correct author for all 7 authors, with probability ranging from 86.39% to 100%. The tables show a significant increase in identification probability as the number of messages per conversation increase. Figure 4-9 shows the relationship between the identification probability and number of messages per conversation.

Table 4-19. Dataset 1, MGD Results, 5 Messages, Authors A13-A19

Size=5	P(A13|x)	P(A14|x)	P(A15|x)	P(A16|x)	P(A17|x)	P(A18|x)	P(A19|x)
A13	15.79%	**28.21%**	6.93%	1.90%	10.91%	17.90%	18.37%
A14	11.10%	**34.35%**	4.31%	0.61%	7.96%	29.43%	12.23%
A15	11.83%	**22.15%**	13.95%	5.81%	19.59%	8.36%	18.31%
A16	7.98%	13.03%	10.08%	12.99%	**30.48%**	3.86%	21.57%
A17	9.48%	18.38%	5.58%	6.27%	**30.36%**	8.97%	20.96%
A18	9.99%	34.40%	3.09%	0.15%	6.94%	**35.91%**	9.52%
A19	12.15%	**22.09%**	6.43%	4.97%	19.66%	14.28%	20.42%

Table 4-20. Dataset 1, MGD Results, 10 Messages, Authors A13-A19

Size=10	P(A13|x)	P(A14|x)	P(A15|x)	P(A16|x)	P(A17|x)	P(A18|x)	P(A19|x)
A13	16.25%	**28.99%**	5.85%	0.95%	6.47%	20.19%	21.30%
A14	8.41%	**43.20%**	2.60%	0.07%	1.63%	35.88%	8.21%
A15	15.07%	**20.68%**	18.74%	3.60%	16.54%	4.94%	20.44%
A16	4.35%	2.45%	8.81%	14.30%	**44.11%**	0.58%	25.41%
A17	4.27%	5.45%	2.35%	7.56%	**49.56%**	3.19%	27.61%
A18	5.29%	40.21%	1.68%	0.00%	2.16%	**44.65%**	6.00%
A19	9.59%	16.18%	4.46%	4.16%	27.78%	7.04%	**30.78%**

Table 4-21. Dataset 1, MGD Results, 25 Messages, Authors A13-A19

Size=25	P(A13|x)	P(A14|x)	P(A15|x)	P(A16|x)	P(A17|x)	P(A18|x)	P(A19|x)
A13	**38.51%**	24.16%	11.47%	0.35%	2.18%	8.34%	15.00%
A14	6.37%	**66.15%**	1.10%	0.00%	0.07%	22.36%	3.95%
A15	24.17%	12.99%	**45.64%**	0.15%	1.57%	0.10%	15.38%
A16	0.90%	0.01%	5.42%	21.20%	**45.75%**	0.00%	26.72%
A17	2.21%	0.18%	0.14%	4.02%	**58.92%**	0.12%	34.41%
A18	4.76%	41.38%	0.19%	0.00%	0.00%	**50.14%**	3.54%
A19	3.85%	14.96%	0.36%	8.81%	25.45%	8.25%	**38.32%**

Table 4-22. Dataset 1, MGD Results, 50 Messages, Authors A13-A19

Size=50	P(A13|x)	P(A14|x)	P(A15|x)	P(A16|x)	P(A17|x)	P(A18|x)	P(A19|x)
A13	**65.01%**	8.99%	0.46%	0.00%	0.00%	18.27%	7.27%
A14	0.68%	**90.86%**	0.00%	0.00%	0.00%	7.42%	1.03%
A15	26.36%	1.55%	**67.63%**	0.10%	0.00%	0.51%	3.85%
A16	0.00%	0.00%	0.02%	20.50%	34.59%	0.00%	**44.89%**
A17	0.00%	0.00%	0.00%	0.42%	**87.94%**	0.00%	11.64%
A18	0.26%	8.95%	0.29%	0.00%	0.00%	**90.16%**	0.34%
A19	0.91%	3.26%	0.08%	5.21%	40.40%	0.26%	**49.87%**

Table 4-23. Dataset 1, MGD Results, 100 Messages, Authors A13-A19

Size=100	P(A13|x)	P(A14|x)	P(A15|x)	P(A16|x)	P(A17|x)	P(A18|x)	P(A19|x)
A13	**92.63%**	0.18%	0.01%	0.00%	0.00%	6.15%	1.02%
A14	0.00%	**99.99%**	0.00%	0.00%	0.00%	0.00%	0.01%
A15	0.00%	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%
A16	0.00%	0.00%	0.00%	**82.39%**	0.04%	0.00%	17.57%
A17	0.00%	0.00%	0.00%	0.00%	**99.99%**	0.00%	0.01%
A18	0.00%	0.11%	0.00%	0.00%	0.00%	**99.83%**	0.06%
A19	0.00%	1.47%	0.00%	0.00%	0.10%	0.00%	**98.43%**

Table 4-24. Dataset 1, MGD Results, 125 Messages, Authors A13-A19

Size=125	P(A13|x)	P(A14|x)	P(A15|x)	P(A16|x)	P(A17|x)	P(A18|x)	P(A19|x)
A13	**100.00%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
A14	0.00%	**99.93%**	0.00%	0.00%	0.00%	0.00%	0.07%
A15	0.00%	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%
A16	0.00%	0.00%	0.00%	**86.39%**	0.85%	0.00%	12.77%
A17	0.00%	0.00%	0.00%	0.00%	**99.76%**	0.00%	0.24%
A18	0.00%	0.01%	0.00%	0.00%	0.00%	**99.99%**	0.00%
A19	0.00%	0.06%	0.00%	3.15%	0.00%	0.00%	**96.79%**

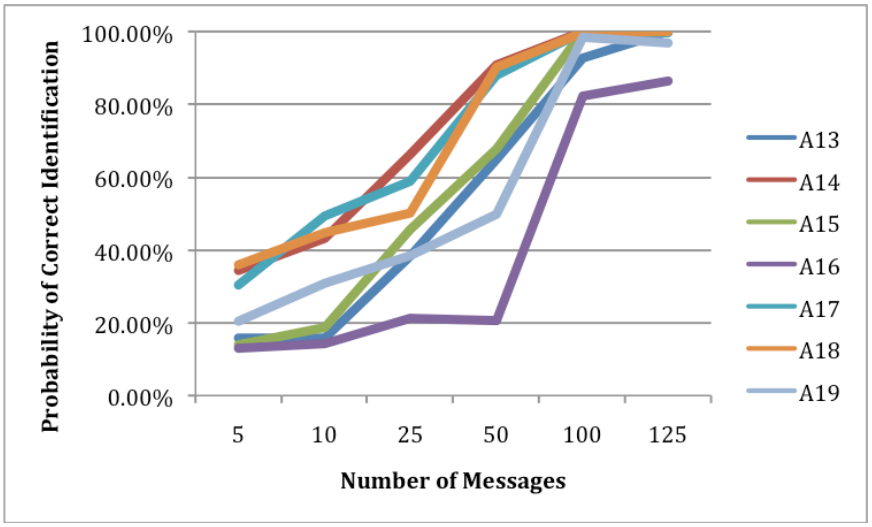
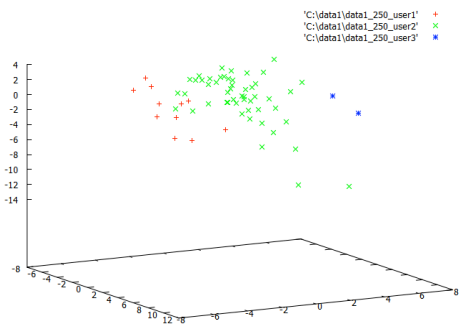
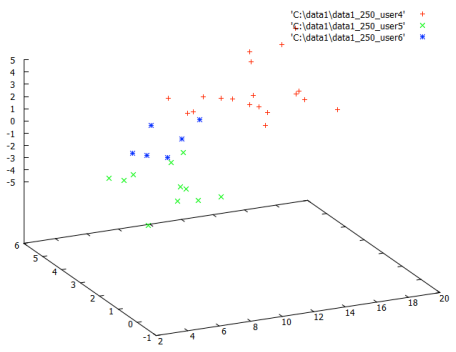


Figure 4-9. Dataset 1, Identification Probability vs. Number of Messages, Authors A13-A19

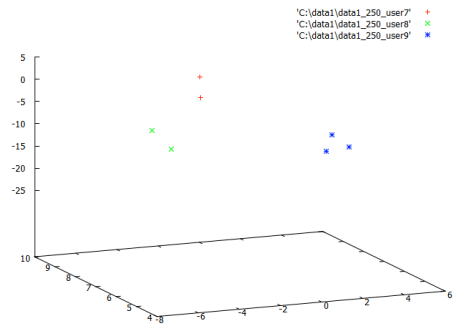
Figure 4-10 shows Dataset #1 PCA plot results for conversations consisting of 250 messages with the authors sequentially divided in to small sets to magnify the differentiation. The plots in this table show separate groupings for each author.



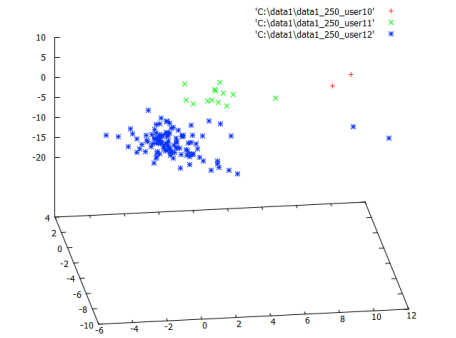
(a) Dataset 1, 250 Messages, Authors A1-A3



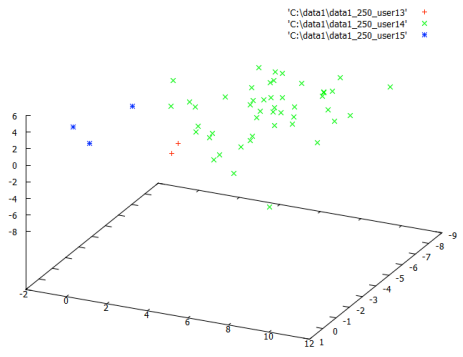
(b) Dataset 1, 250 Messages, Authors A4-A6



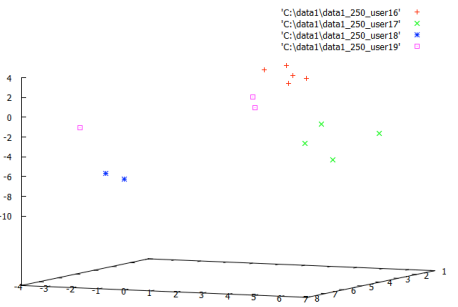
(c) Dataset 1, 250 Messages, Authors A7-A9



(d) Dataset 1, 250 Messages, Authors A10-A12



(e) Dataset 1, 250 Messages, Authors A13-A15



(f) Dataset 1, 250 Messages, Authors A16-A19

Figure 4-10. Dataset 1, PCA Plot Results, 250 Messages, Authors A1-A19

Table 4-25 through

Table 4-30 show the MGD results for conversation sizes of 5, 10, 25, 50, 100, and 125 messages respectively for all authors divided into small sets. Using 100 messages per conversation as input, MGD identifies conversations as the correct author with probability ranging from 82.39% to 100%. Using 125 messages per conversation as input, MGD identifies conversations as the correct author with probability ranging from 86.39% to 100%. Given the smaller number of authors for identification, many tests resulted in probability from 70%-100% using just 50 messages per conversation. The tables show a significant increase in identification probability as the number of messages per conversation increase. Figure 4-11 through Figure 4-16 show the relationship between the identification probability and number of messages per conversation.

Table 4-25. Dataset 1, MGD Results, 5-125 Messages, Authors A1-A3

Size=5	P(A1|x)	P(A2|x)	P(A3|x)
A1	**55.07%**	38.96%	5.97%
A2	**46.82%**	43.11%	10.07%
A3	33.51%	**36.97%**	29.52%

Size=10	P(A1|x)	P(A2|x)	P(A3|x)
A1	**61.96%**	37.48%	0.56%
A2	44.83%	**53.28%**	1.89%
A3	12.57%	39.76%	**47.67%**

Size=25	P(A1|x)	P(A2|x)	P(A3|x)
A1	**65.29%**	34.71%	0.00%
A2	29.45%	**70.28%**	0.27%
A3	0.14%	6.27%	**93.58%**

Size=50	P(A1|x)	P(A2|x)	P(A3|x)
A1	**83.15%**	16.85%	0.00%
A2	11.96%	**88.04%**	0.00%
A3	0.00%	0.08%	**99.92%**

Size=100	P(A1|x)	P(A2|x)	P(A3|x)
A1	**91.93%**	8.07%	0.00%
A2	1.84%	**98.16%**	0.00%
A3	0.00%	0.00%	**100.00%**

Size=125	P(A1|x)	P(A2|x)	P(A3|x)
A1	**99.28%**	0.72%	0.00%
A2	0.28%	**99.72%**	0.00%
A3	0.00%	0.00%	**100.00%**

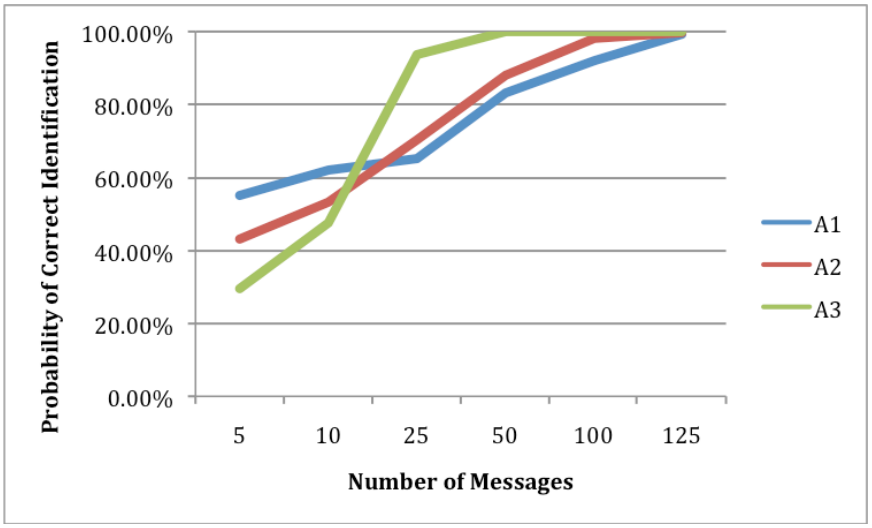


Figure 4-11. Dataset 1, Identification Probability vs. Number of Messages, Authors A1-A3

Table 4-26. Dataset 1, MGD Results, 5-125 Messages, Authors A4-A6

Size=5	P(A4|x)	P(A5|x)	P(A6|x)
A4	**36.74%**	28.38%	34.88%
A5	19.75%	**41.51%**	38.75%
A6	22.22%	37.27%	**40.52%**

Size=10	P(A4|x)	P(A5|x)	P(A6|x)
A4	**48.99%**	26.85%	24.16%
A5	15.01%	**48.30%**	36.69%
A6	17.25%	40.70%	**42.05%**

Size=25	P(A4|x)	P(A5|x)	P(A6|x)
A4	**70.03%**	21.11%	8.87%
A5	2.75%	**85.06%**	12.19%
A6	7.60%	41.09%	**51.31%**

Size=50	P(A4|x)	P(A5|x)	P(A6|x)
A4	**99.97%**	0.02%	0.01%
A5	0.05%	**76.96%**	22.99%
A6	0.52%	18.62%	**80.85%**

Size=100	P(A4|x)	P(A5|x)	P(A6|x)
A4	**100.00%**	0.00%	0.00%
A5	0.00%	**99.21%**	0.79%
A6	0.00%	0.90%	**99.10%**

Size=125	P(A4|x)	P(A5|x)	P(A6|x)
A4	**100.00%**	0.00%	0.00%
A5	0.00%	**99.72%**	0.28%
A6	0.00%	0.01%	**99.99%**

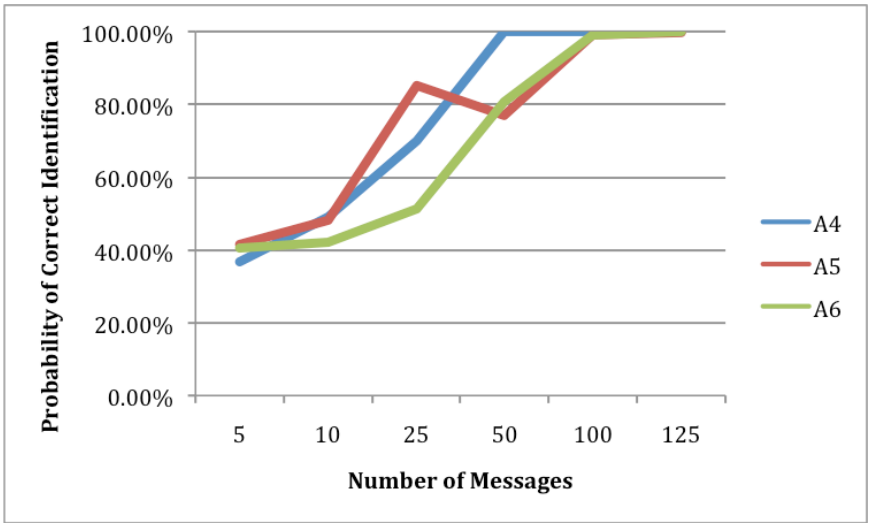


Figure 4-12. Dataset 1, Identification Probability vs. Number of Messages, Authors A4-A6

Table 4-27. Dataset 1, MGD Results, 5-125 Messages, Authors A7-A9

Size=5	P(A7|x)	P(A8|x)	P(A9|x)
A7	29.95%	**48.96%**	21.08%
A8	19.03%	**63.90%**	17.07%
A9	8.07%	25.57%	**66.36%**

Size=10	P(A7|x)	P(A8|x)	P(A9|x)
A7	**46.92%**	44.60%	8.48%
A8	17.26%	**81.31%**	1.43%
A9	2.73%	3.62%	**93.65%**

Size=25	P(A7|x)	P(A8|x)	P(A9|x)
A7	**70.35%**	28.95%	0.70%
A8	15.31%	**84.68%**	0.00%
A9	0.54%	0.45%	**99.01%**

Size=50	P(A7|x)	P(A8|x)	P(A9|x)
A7	**99.88%**	0.12%	0.00%
A8	2.63%	**97.37%**	0.00%
A9	0.00%	0.00%	**100.00%**

Size=100	P(A7|x)	P(A8|x)	P(A9|x)
A7	**100.00%**	0.00%	0.00%
A8	0.00%	**100.00%**	0.00%
A9	0.00%	0.00%	**100.00%**

Size=125	P(A7|x)	P(A8|x)	P(A9|x)
A7	**100.00%**	0.00%	0.00%
A8	0.00%	**100.00%**	0.00%
A9	0.00%	0.00%	**100.00%**

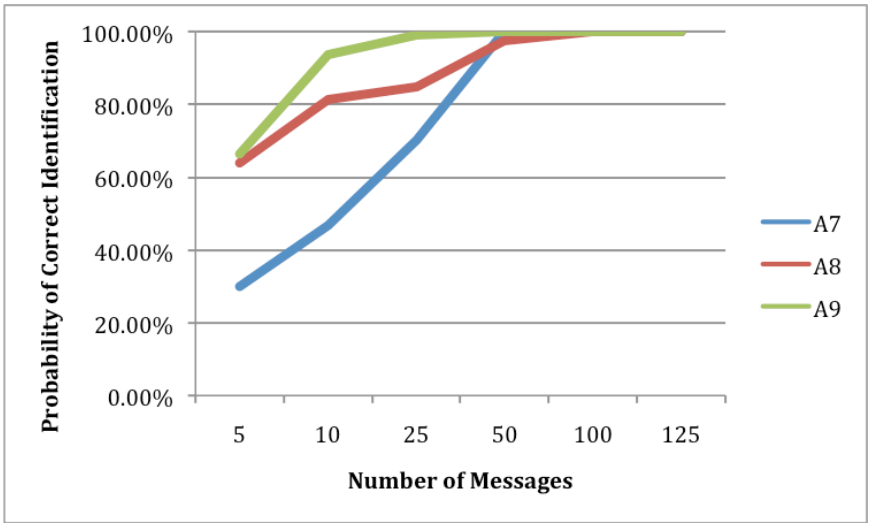


Figure 4-13. Dataset 1, Identification Probability vs. Number of Messages, Authors A7-A9

Table 4-28. Dataset 1, MGD Results, 5-125 Messages, Authors A10-A12

Size=5	P(A10|x)	P(A11|x)	P(A12|x)
A10	**80.39%**	6.46%	13.15%
A11	1.82%	39.51%	**58.67%**
A12	0.77%	33.17%	**66.06%**

Size=10	P(A10|x)	P(A11|x)	P(A12|x)
A10	**99.85%**	0.01%	0.14%
A11	0.01%	42.83%	**57.16%**
A12	0.00%	22.98%	**77.01%**

Size=25	P(A10|x)	P(A11|x)	P(A12|x)
A10	**100.00%**	0.00%	0.00%
A11	0.00%	47.59%	**52.41%**
A12	0.00%	8.84%	**91.16%**

Size=50	P(A10|x)	P(A11|x)	P(A12|x)
A10	**100.00%**	0.00%	0.00%
A11	0.00%	**78.66%**	21.34%
A12	0.00%	1.65%	**98.35%**

Size=100	P(A10|x)	P(A11|x)	P(A12|x)
A10	**100.00%**	0.00%	0.00%
A11	0.00%	**99.77%**	0.23%
A12	0.00%	0.17%	**99.83%**

Size=125	P(A10|x)	P(A11|x)	P(A12|x)
A10	**100.00%**	0.00%	0.00%
A11	0.00%	**99.81%**	0.19%
A12	0.00%	0.04%	**99.96%**

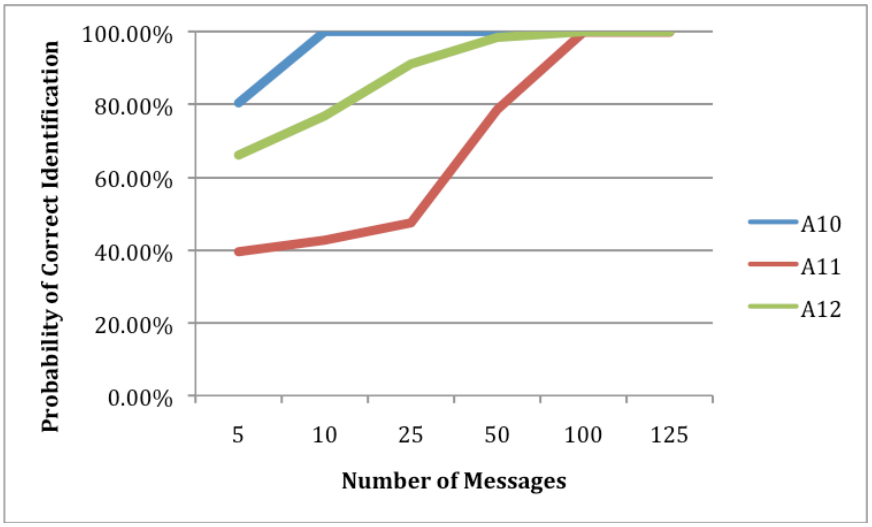


Figure 4-14. Dataset 1, Identification Probability vs. Number of Messages, Authors A10-A12

Table 4-29. Dataset 1, MGD Results, 5-125 Messages, Authors A13-A15

Size=5	P(A13|x)	P(A14|x)	P(A15|x)
A13	31.00%	**55.40%**	13.60%
A14	22.30%	**69.03%**	8.67%
A15	24.68%	**46.21%**	29.11%

Size=10	P(A13|x)	P(A14|x)	P(A15|x)
A13	31.81%	**56.74%**	11.45%
A14	15.52%	**79.69%**	4.80%
A15	27.65%	**37.96%**	34.39%

Size=25	P(A13|x)	P(A14|x)	P(A15|x)
A13	**51.94%**	32.59%	15.47%
A14	8.65%	**89.85%**	1.49%
A15	29.19%	15.69%	**55.12%**

Size=50	P(A13|x)	P(A14|x)	P(A15|x)
A13	**87.31%**	12.07%	0.62%
A14	0.74%	**99.26%**	0.00%
A15	27.59%	1.63%	**70.79%**

Size=100	P(A13|x)	P(A14|x)	P(A15|x)
A13	**99.79%**	0.20%	0.01%
A14	0.00%	**100.00%**	0.00%
A15	0.00%	0.00%	**100.00%**

Size=125	P(A13|x)	P(A14|x)	P(A15|x)
A13	**100.00%**	0.00%	0.00%
A14	0.00%	**100.00%**	0.00%
A15	0.00%	0.00%	**100.00%**

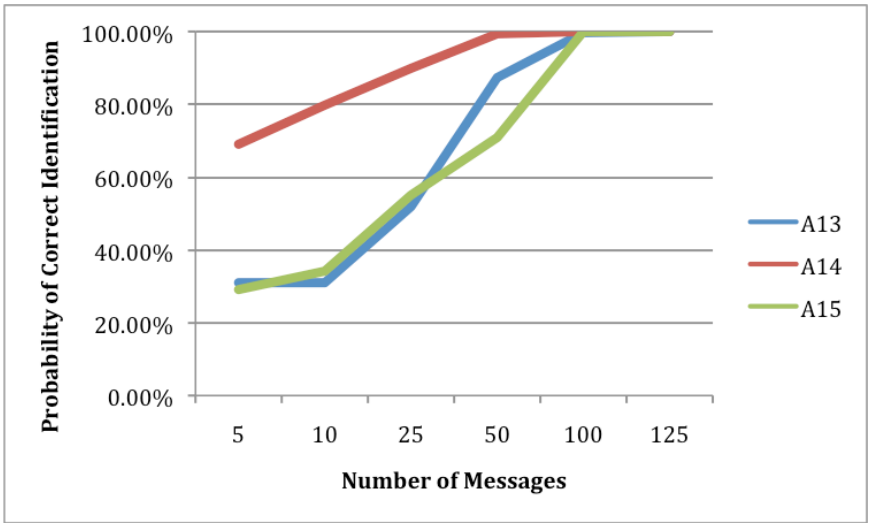


Figure 4-15. Dataset 1, Identification Probability vs. Number of Authors, Authors A13-A15

Table 4-30. Dataset 1, MGD Results, 5-125 Messages, Authors A16-A19

Size=5	P(A16|x)	P(A17|x)	P(A18|x)	P(A19|x)
A16	18.85%	**44.23%**	5.61%	31.31%
A17	9.42%	**45.61%**	13.47%	31.50%
A18	0.28%	13.22%	**68.37%**	18.13%
A19	8.38%	33.13%	24.07%	**34.42%**

Size=10	P(A16|x)	P(A17|x)	P(A18|x)	P(A19|x)
A16	16.95%	**52.27%**	0.68%	30.10%
A17	8.60%	**56.37%**	3.63%	31.40%
A18	0.00%	4.10%	**84.53%**	11.37%
A19	5.96%	39.83%	10.09%	**44.12%**

Size=25	P(A16|x)	P(A17|x)	P(A18|x)	P(A19|x)
A16	22.63%	**48.85%**	0.00%	28.52%
A17	4.12%	**60.45%**	0.13%	35.30%
A18	0.00%	0.00%	**93.40%**	6.59%
A19	10.90%	31.49%	10.21%	**47.41%**

Size=50	P(A16|x)	P(A17|x)	P(A18|x)	P(A19|x)
A16	20.50%	34.60%	0.00%	**44.90%**
A17	0.42%	**87.94%**	0.00%	11.64%
A18	0.00%	0.00%	**99.63%**	0.37%
A19	5.44%	42.20%	0.28%	**52.08%**

Size=100	P(A16|x)	P(A17|x)	P(A18|x)	P(A19|x)
A16	**82.39%**	0.04%	0.00%	17.57%
A17	0.00%	**99.99%**	0.00%	0.01%
A18	0.00%	0.00%	**99.94%**	0.06%
A19	0.00%	0.10%	0.00%	**99.90%**

Size=125	P(A16|x)	P(A17|x)	P(A18|x)	P(A19|x)
A16	**86.39%**	0.85%	0.00%	12.77%
A17	0.00%	**99.76%**	0.00%	0.24%
A18	0.00%	0.00%	**100.00%**	0.00%
A19	3.15%	0.00%	0.00%	**96.85%**

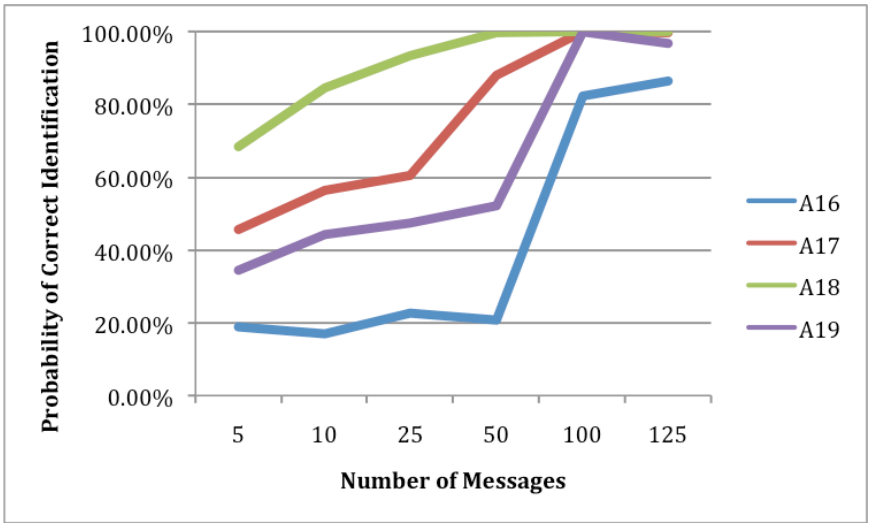


Figure 4-16. Dataset 1, Identification Probability vs. Number of Messages, Authors A16-A19

Figure 4-17 shows Dataset #1 PCA plot results for the 7 authors with the highest total number of messages (Authors A2, A4, A5, A11, A12, A14, A16, respectively), resulting in the highest number of writeprint instances. The conversations consist of 250 messages for each writeprint instance. This plot shows separate groupings for each author.

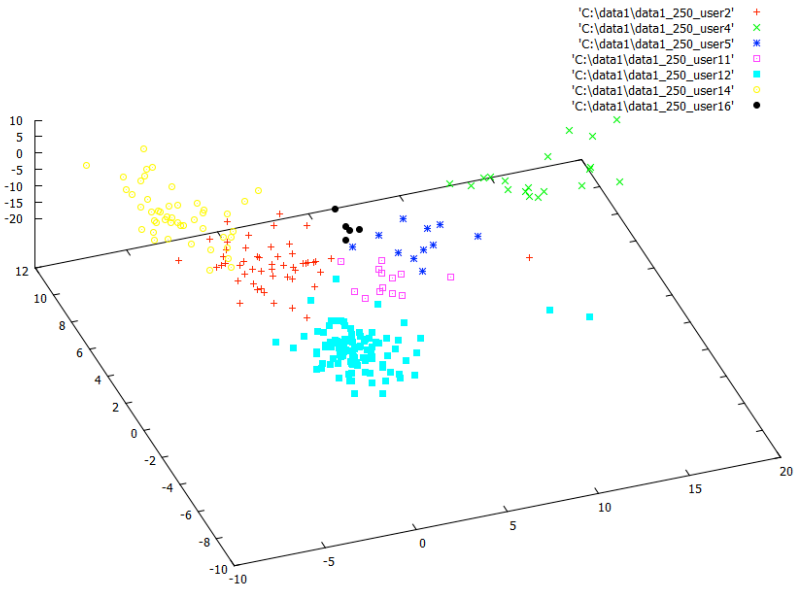


Figure 4-17. Dataset 1, PCA Plot Results, 250 Messages, Top 7 Authors

Table 4-31 through

Table 4-38 show the MGD results for conversation sizes of 5, 10, 25, 50, 100, 125, 250, and 500 messages respectively for the top 7 authors (Authors A2, A4, A5, A11,

A12, A14, A16). Using 100 messages per conversation as input, MGD identifies conversations as the correct author for all 7 authors, with probability ranging from 81.00% to 100%. Using 500 messages per conversation as input, MGD identifies conversations as the correct author for all 7 authors, with probability ranging from 99.85% to 100%. The tables show a significant increase in identification probability as the number of messages per conversation increase. Figure 4-18 shows the relationship between the identification probability and number of messages per conversation.

Table 4-31. Dataset 1, MGD Results, 5 Messages, Top 7 Authors

Size=5	P(A2|x)	P(A4|x)	P(A5|x)	P(A11|x)	P(A12|x)	P(A14|x)	P(A16|x)
A2	**31.47%**	1.62%	5.81%	10.47%	23.93%	25.57%	1.12%
A4	11.67%	**32.67%**	25.23%	7.62%	10.23%	5.48%	7.10%
A5	20.37%	11.37%	**23.89%**	10.14%	16.94%	11.01%	6.28%
A11	24.96%	4.61%	11.70%	17.32%	**25.73%**	13.98%	1.70%
A12	**28.79%**	1.99%	7.67%	14.22%	28.33%	17.83%	1.16%
A14	**34.46%**	0.75%	4.11%	7.69%	20.64%	31.79%	0.56%
A16	17.53%	13.65%	**22.10%**	9.33%	13.38%	12.02%	11.99%

Table 4-32. Dataset 1, MGD Results, 10 Messages, Top 7 Authors

Size=10	P(A2|x)	P(A4|x)	P(A5|x)	P(A11|x)	P(A12|x)	P(A14|x)	P(A16|x)
A2	**39.81%**	0.27%	1.48%	6.75%	25.17%	26.30%	0.22%
A4	3.48%	**55.15%**	30.22%	3.72%	2.18%	0.51%	4.75%
A5	14.61%	13.63%	**43.85%**	7.23%	9.68%	3.09%	7.91%
A11	29.48%	1.21%	5.67%	23.26%	**31.04%**	8.89%	0.44%
A12	34.79%	0.19%	1.58%	11.33%	**37.98%**	13.98%	0.14%
A14	**42.71%**	0.03%	0.44%	3.33%	18.46%	34.97%	0.06%
A16	10.35%	13.13%	**40.53%**	5.72%	4.80%	3.73%	21.75%

Table 4-33. Dataset 1, MGD Results, 25 Messages, Top 7 Authors

Size=25	P(A2|x)	P(A4|x)	P(A5|x)	P(A11|x)	P(A12|x)	P(A14|x)	P(A16|x)
A2	**53.98%**	0.00%	0.41%	3.12%	18.21%	24.26%	0.02%
A4	0.02%	**75.66%**	22.80%	0.01%	0.00%	0.00%	1.51%
A5	3.65%	2.79%	**86.42%**	1.20%	0.32%	0.06%	5.55%
A11	28.50%	0.03%	0.75%	33.08%	**36.43%**	1.20%	0.00%
A12	31.31%	0.00%	0.17%	5.69%	**58.67%**	4.08%	0.09%
A14	**48.08%**	0.00%	0.02%	0.27%	5.22%	46.41%	0.00%
A16	1.09%	3.20%	**59.56%**	0.22%	0.19%	0.02%	35.72%

Table 4-34. Dataset 1, MGD Results, 50 Messages, Top 7 Authors

Size=50	P(A2|x)	P(A4|x)	P(A5|x)	P(A11|x)	P(A12|x)	P(A14|x)	P(A16|x)
A2	**65.82%**	0.00%	0.04%	1.07%	8.22%	24.85%	0.00%
A4	0.00%	**99.98%**	0.02%	0.00%	0.00%	0.00%	0.00%
A5	0.28%	0.06%	**95.37%**	0.10%	0.00%	0.00%	4.19%
A11	19.40%	0.00%	0.03%	**63.27%**	17.17%	0.13%	0.00%
A12	9.92%	0.00%	0.00%	1.48%	**88.41%**	0.19%	0.00%
A14	30.75%	0.00%	0.00%	0.01%	1.87%	**67.37%**	0.00%
A16	0.06%	0.18%	**50.34%**	0.00%	0.00%	0.00%	49.42%

Table 4-35. Dataset 1, MGD Results, 100 Messages, Top 7 Authors

Size=100	P(A2|x)	P(A4|x)	P(A5|x)	P(A11|x)	P(A12|x)	P(A14|x)	P(A16|x)
A2	**83.67%**	0.00%	0.00%	0.17%	1.40%	14.76%	0.00%
A4	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%	0.00%
A5	0.00%	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%
A11	8.47%	0.00%	0.00%	**91.32%**	0.21%	0.00%	0.00%
A12	2.13%	0.00%	0.00%	0.16%	**97.70%**	0.01%	0.00%
A14	18.88%	0.00%	0.00%	0.00%	0.11%	**81.00%**	0.00%
A16	0.00%	0.00%	1.94%	0.00%	0.00%	0.00%	**98.06%**

Table 4-36. Dataset 1, MGD Results, 125 Messages, Top 7 Authors

Size=125	P(A2|x)	P(A4|x)	P(A5|x)	P(A11|x)	P(A12|x)	P(A14|x)	P(A16|x)
A2	**82.42%**	0.00%	0.00%	0.06%	1.15%	16.36%	0.00%
A4	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%	0.00%
A5	0.00%	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%
A11	4.15%	0.00%	0.00%	**95.66%**	0.19%	0.00%	0.00%
A12	2.22%	0.00%	0.00%	0.04%	**97.74%**	0.00%	0.00%
A14	16.93%	0.00%	0.00%	0.00%	0.01%	**83.07%**	0.00%
A16	0.00%	0.00%	1.48%	0.00%	0.00%	0.00%	**98.52%**

Table 4-37. Dataset 1, MGD Results, 250 Messages, Top 7 Authors

Size=250	P(A2|x)	P(A4|x)	P(A5|x)	P(A11|x)	P(A12|x)	P(A14|x)	P(A16|x)
A2	**98.25%**	0.00%	0.00%	0.00%	0.05%	1.70%	0.00%
A4	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%	0.00%
A5	0.00%	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%
A11	0.03%	0.00%	0.00%	**99.97%**	0.00%	0.00%	0.00%
A12	0.00%	0.00%	0.00%	0.00%	**100.00%**	0.00%	0.00%
A14	6.87%	0.00%	0.00%	0.00%	0.00%	**93.13%**	0.00%
A16	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**100.00%**

Table 4-38. Dataset 1, MGD Results, 500 Messages, Top 7 Authors

Size=500	P(A2|x)	P(A4|x)	P(A5|x)	P(A11|x)	P(A12|x)	P(A14|x)	P(A16|x)
A2	**99.96%**	0.00%	0.00%	0.00%	0.00%	0.04%	0.00%
A4	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%	0.00%
A5	0.00%	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%
A11	0.00%	0.00%	0.00%	**100.00%**	0.00%	0.00%	0.00%
A12	0.00%	0.00%	0.00%	0.00%	**100.00%**	0.00%	0.00%
A14	0.15%	0.00%	0.00%	0.00%	0.00%	**99.85%**	0.00%
A16	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**100.00%**

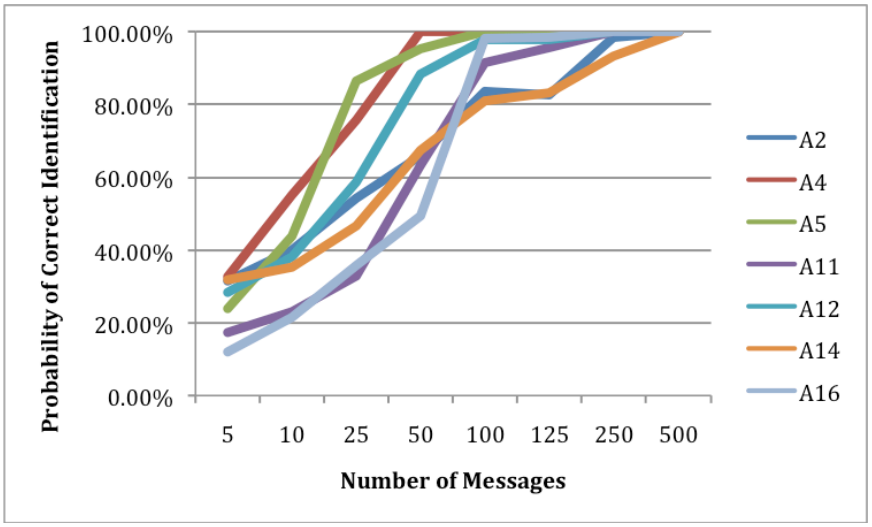


Figure 4-18. Dataset 1, Identification Probability vs. Number of Messages, Top 7 Authors

Dataset #1 analysis for all the top 7 authors achieved less than 20% error for all authors using 250 messages per conversation. Figure 4-19 shows that as the conversation size increases, the error rate decreases.

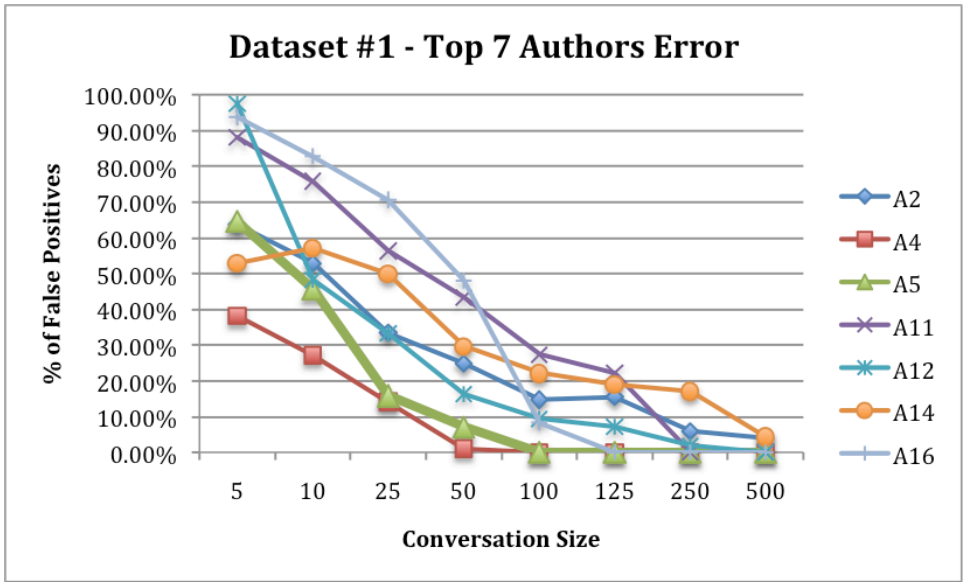
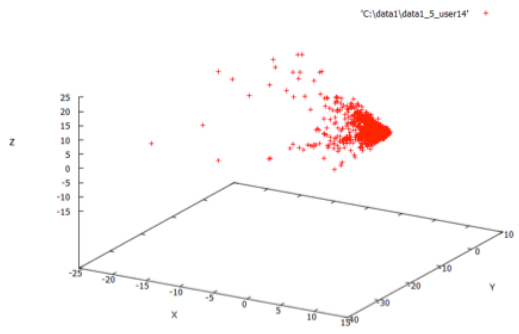
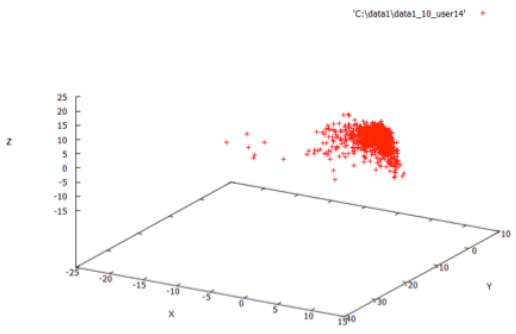


Figure 4-19. Dataset #1, Top 7 Authors Error

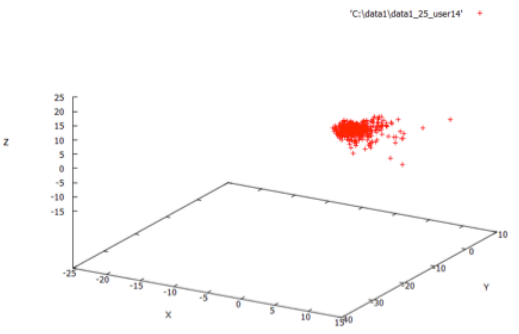
Figure 4-20 shows the Dataset #1 PCA data plots for a single author (Author A14) over the full range of conversation sizes (5, 10, 25, 50, 100, 125, 250, and 500 messages respectively). The data shows as the number of messages per conversation increase, the data points become more tightly grouped. This demonstrates that as the messages per conversation increase, the writeprint becomes more cohesive.



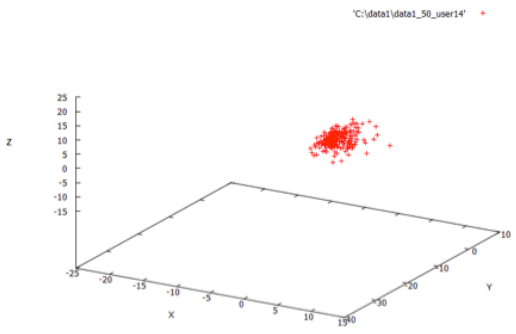
(a) Dataset 1, Author A14, 5 Messages



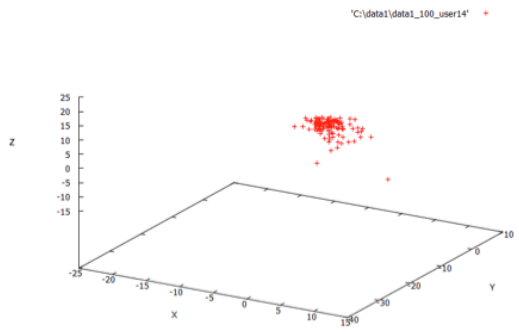
(b) Dataset 1, Author A14, 10 Messages



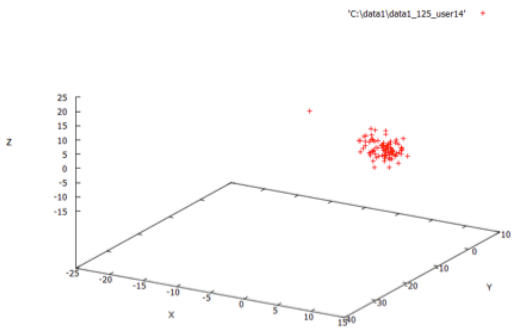
(c) Dataset 1, Author A14, 25 Messages



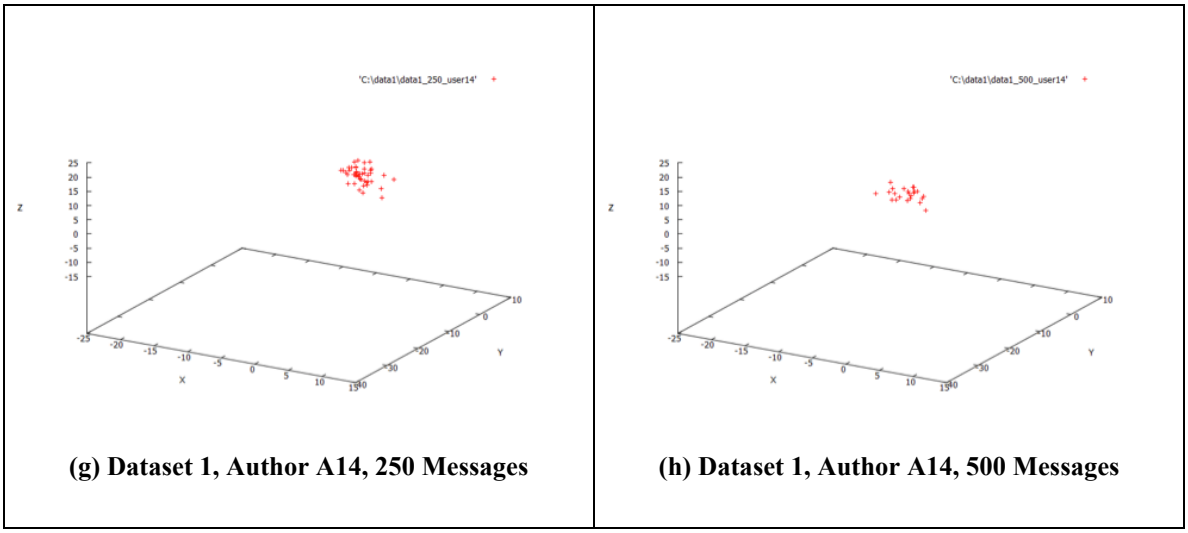
(d) Dataset 1, Author A14, 50 Messages



(e) Dataset 1, Author A14, 100 Messages



(f) Dataset 1, Author A14, 125 Messages



(g) Dataset 1, Author A14, 250 Messages **(h) Dataset 1, Author A14, 500 Messages**

Figure 4-20. Dataset 1, PCA Plot Results, Author A14, All Conversation Sizes

Conversation size can be analyzed in more detail by calculating the standard deviation of the data within each conversation size. The standard deviation measures the spread of distribution of a set of data by calculating distance from the mean of the data. If the data points are very close together (close to the mean), the standard deviation will be low. If the data points are spread out (far from the mean), the standard deviation will be high. Figure 4-21 shows the inverse relationship of standard deviation and conversation size for the Author A14 results shown in Figure 4-20. As the conversation size increases (i.e. number of messages per conversation), the standard deviation decreases. This shows that with larger conversations sizes an author's writeprint becomes more concise and is likely more representative of the author's true writing style.

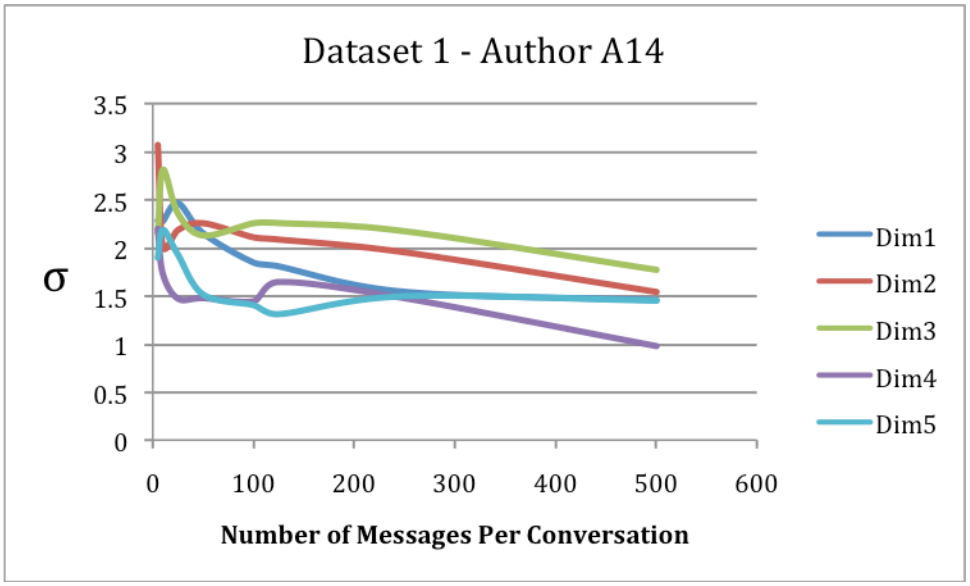


Figure 4-21. Dataset 1, Author A14, Conversation Size/Standard Deviation Relationship

The standard deviation of the data is calculated for the first 5 PCA dimensions for all 19 authors in Dataset 1. As shown in Table 4-39, 96% of the 95 values exhibited decreased standard deviation as the conversation size increased.

Table 4-39. Dataset 1 Results for Conversation Size/Standard Deviation Relationship

Dataset	Number of Authors	Number of Dimensions per Author	Total Values Analyzed	Dimensions that Show Decrease in σ
1	19	5	95 (across sets of 5,10,25,50,100,125,250,500 messages per conversation)	96%

Figure 4-22 and Figure 4-23 show Dataset #1 PCA plot results for multiple samples of messages from Authors A2 and A12, respectively. The conversations consist of 250 messages for each writeprint instance. These results show that an individual author's writeprint is consistent over multiple samples. The overlapping PCA data points show writeprint similarity for an author over multiple distinct samples. Outliers tend to be the result of conversation topic. For example, an author may insert a few URLs into the conversation and this would create an outlier due to the special characters (:, /, /, etc.) that are not normally used by this author.

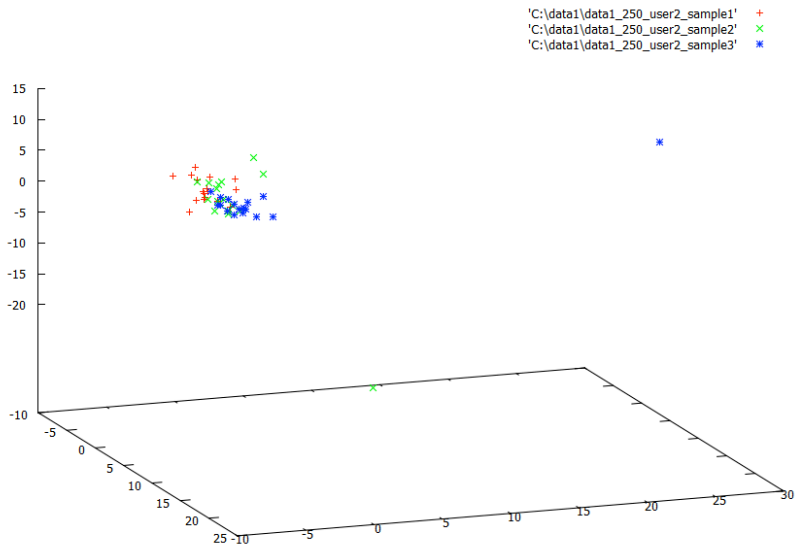


Figure 4-22. Dataset 1, PCA Plot Results, 250 Messages, Author A2 Samples

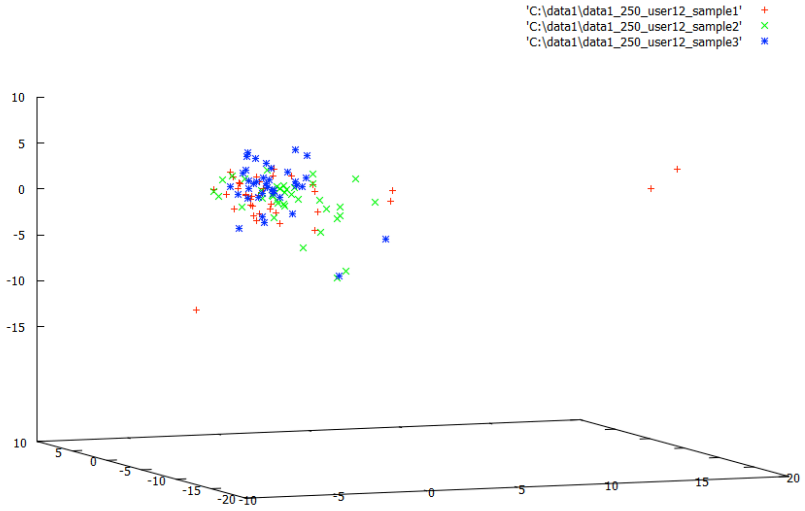


Figure 4-23. Dataset 1, PCA Plot Results, 250 Messages, Author A12 Samples

Dataset #1 Known Authors provides some beneficial metadata to assist analysis. Figure 4-24 shows the family diagram of five related authors. The following tests were performed to determine if writeprints of family members show greater similarity than those of unrelated authors.

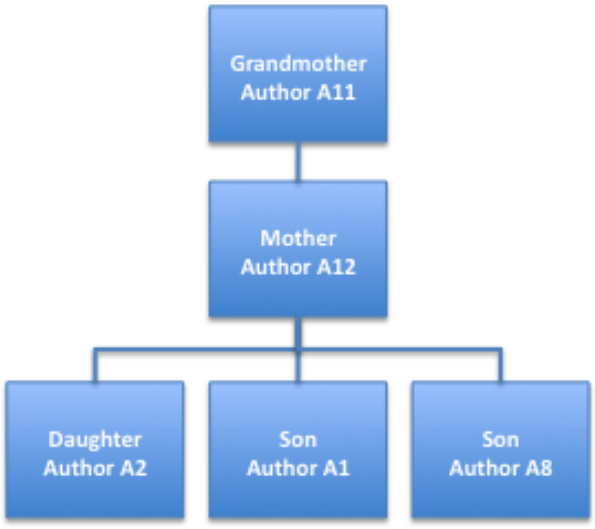


Figure 4-24. Author Family Tree

Figure 4-25 shows Dataset #1 PCA plot results for all five related authors (Authors A1, A2, A8, A11, A12). The conversations consist of 250 messages for each writeprint instance. This plot shows separate groupings for each author.

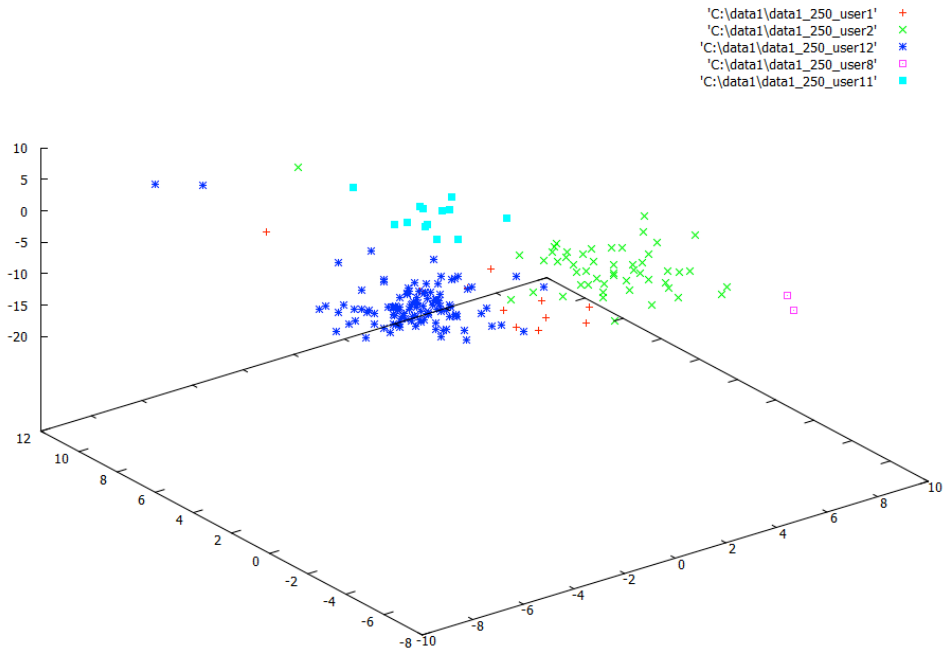


Figure 4-25. Dataset 1, PCA Plot Results, 250 Messages, 5 Related Authors

Table 4-40 through

Table 4-46 show the MGD results for conversation sizes of 5, 10, 25, 50, 100, 125, and 250 messages respectively for the 5 related authors (Authors A1, A2, A8, A11, A12). Using 100 messages per conversation as input, MGD identifies conversations as the correct author for all 5 authors, with probability ranging from 84.59% to 100%. Using 250 messages per conversation as input, MGD identifies conversations as the correct author for all 5 authors with probability ranging from 99.94% to 100%. The tables show a significant increase in identification probability as the number of messages per conversation increase. Figure 4-26 shows the relationship between the identification probability and number of messages per conversation. The PCA plots and MGD results

do not show any significant differences between related authors and unrelated authors from previous tests. Both related authors and unrelated authors showed similar identification probability across various conversation sizes.

Table 4-40. Dataset 1, MGD Results, 5 Messages, 5 Related Authors

Size=5	P(A1|x)	P(A2|x)	P(A8|x)	P(A11|x)	P(A12|x)
A1	**32.38%**	22.91%	17.87%	6.44%	20.41%
A2	**27.76%**	25.56%	18.74%	8.51%	19.44%
A8	26.65%	25.80%	**31.74%**	3.94%	11.87%
A11	**27.04%**	24.41%	6.45%	16.94%	25.16%
A12	**30.92%**	24.70%	7.88%	12.20%	24.31%

Table 4-41. Dataset 1, MGD Results, 10 Messages, 5 Related Authors

Size=10	P(A1|x)	P(A2|x)	P(A8|x)	P(A11|x)	P(A12|x)
A1	**39.53%**	23.91%	9.59%	3.87%	23.09%
A2	27.86%	**33.12%**	12.46%	5.62%	20.94%
A8	26.54%	30.37%	**36.51%**	0.54%	6.04%
A11	20.95%	27.28%	1.54%	21.52%	**28.72%**
A12	31.31%	27.57%	2.04%	8.98%	**30.10%**

Table 4-42. Dataset 1, MGD Results, 25 Messages, 5 Related Authors

Size=25	P(A1|x)	P(A2|x)	P(A8|x)	P(A11|x)	P(A12|x)
A1	**47.72%**	25.37%	2.97%	1.32%	22.63%
A2	21.17%	**50.51%**	8.36%	2.92%	17.04%
A8	14.16%	24.70%	**61.07%**	0.00%	0.07%
A11	9.16%	26.40%	0.04%	30.65%	**33.75%**
A12	28.56%	23.33%	0.14%	4.24%	**43.73%**

Table 4-43. Dataset 1, MGD Results, 50 Messages, 5 Related Authors

Size=50	P(A1|x)	P(A2|x)	P(A8|x)	P(A11|x)	P(A12|x)
A1	**67.28%**	13.64%	0.00%	0.28%	18.81%
A2	10.44%	**76.87%**	1.85%	1.25%	9.59%
A8	5.97%	12.63%	**81.40%**	0.00%	0.00%
A11	0.87%	19.26%	0.00%	**62.82%**	17.05%
A12	18.84%	8.06%	0.00%	1.21%	**71.89%**

Table 4-44. Dataset 1, MGD Results, 100 Messages, 5 Related Authors

Size=100	P(A1|x)	P(A2|x)	P(A8|x)	P(A11|x)	P(A12|x)
A1	**84.59%**	7.43%	0.00%	0.00%	7.98%
A2	1.80%	**96.39%**	0.00%	0.20%	1.61%
A8	0.00%	0.00%	**100.00%**	0.00%	0.00%
A11	0.02%	8.47%	0.00%	**91.30%**	0.21%
A12	6.81%	1.99%	0.00%	0.15%	**91.05%**

Table 4-45. Dataset 1, MGD Results, 125 Messages, 5 Related Authors

Size=125	P(A1|x)	P(A2|x)	P(A8|x)	P(A11|x)	P(A12|x)
A1	**87.85%**	0.64%	0.00%	0.00%	11.52%
A2	0.28%	**98.27%**	0.00%	0.07%	1.38%
A8	0.23%	0.02%	**99.75%**	0.00%	0.00%
A11	0.00%	4.15%	0.00%	**95.66%**	0.19%
A12	7.04%	2.06%	0.00%	0.04%	**90.86%**

Table 4-46. Dataset 1, MGD Results, 250 Messages, 5 Related Authors

Size=250	P(A1|x)	P(A2|x)	P(A8|x)	P(A11|x)	P(A12|x)
A1	**99.94%**	0.00%	0.00%	0.00%	0.05%
A2	0.00%	**99.95%**	0.00%	0.00%	0.05%
A8	0.00%	0.00%	**100.00%**	0.00%	0.00%
A11	0.00%	0.03%	0.00%	**99.97%**	0.00%
A12	0.00%	0.00%	0.00%	0.00%	**100.00%**

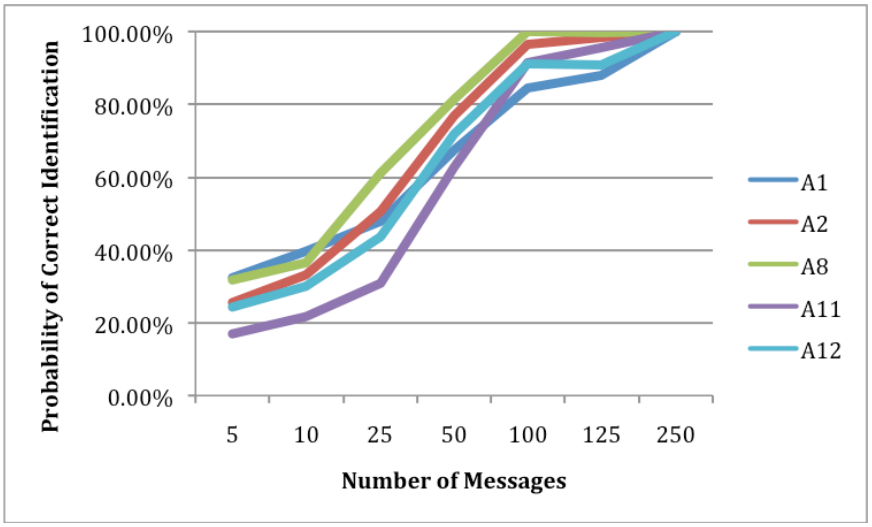


Figure 4-26. Dataset 1, Identification Probability vs. Number of Messages, 5 Related Authors

Figure 4-27 shows Dataset #1 PCA plot results for Authors A1, A2, and A8. These authors are siblings. The conversations consist of 250 messages for each writeprint instance. This plot shows separate groupings for each author.

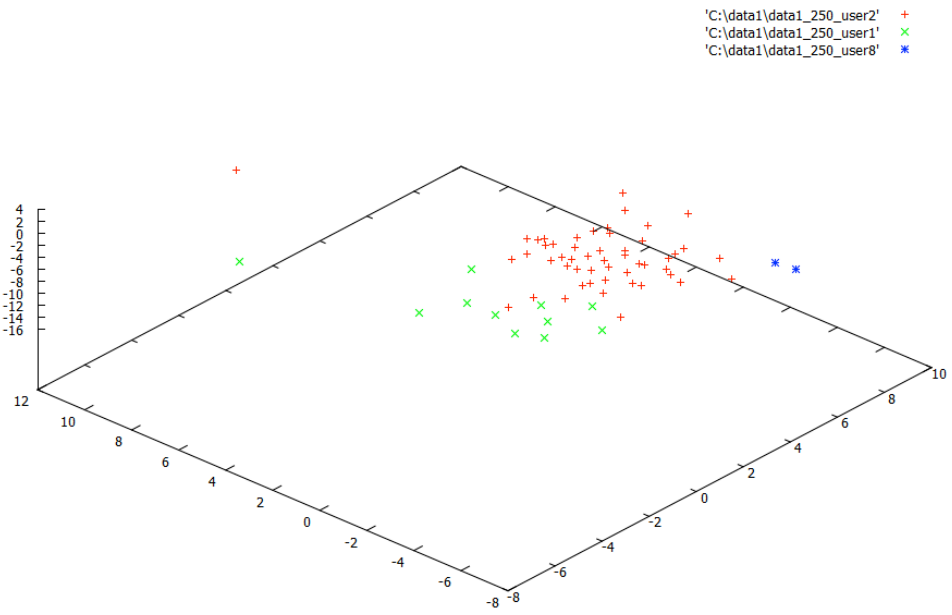


Figure 4-27. Data 1 Results, 250 Messages, 3 Sibling Authors

Table 4-47 shows the MGD results for conversation sizes of 5, 10, 25, 50, 100, 125, and 250 messages respectively for the 3 sibling authors (Authors A1, A2, A8). Using 100 messages per conversation as input, MGD identifies conversations as the correct author for all 3 authors, with probability ranging from 91.93% to 100%. Using 250 messages per conversation as input, MGD identifies conversations as the correct author for all 3 authors with 100% probability. Given the smaller number of authors for identification, tests resulted in probability from 81.40%-86.22% using just 50 messages per conversation. The tables show a significant increase in identification probability as the number of messages per conversation increase. Figure 4-28 shows the relationship between the identification probability and number of messages per conversation.

Table 4-47. Dataset 1, MGD Results, 5-500 Messages, 3 Sibling Authors

Size=5	P(A1|x)	P(A2|x)	P(A8|x)
A1	**44.26%**	31.32%	24.42%
A2	**38.52%**	35.47%	26.00%
A8	31.66%	30.64%	**37.70%**

Size=10	P(A1|x)	P(A2|x)	P(A8|x)
A1	54.13%	32.74%	13.13%
A2	37.94%	**45.09%**	16.97%
A8	28.41%	32.51%	**39.08%**

Size=25	P(A1|x)	P(A2|x)	P(A8|x)
A1	**62.74%**	33.35%	3.90%
A2	26.44%	**63.11%**	10.45%
A8	14.17%	24.72%	**61.11%**

Size=50	P(A1|x)	P(A2|x)	P(A8|x)
A1	**83.15%**	16.85%	0.00%
A2	11.71%	**86.22%**	2.07%
A8	5.97%	12.63%	**81.40%**

Size=100	P(A1|x)	P(A2|x)	P(A8|x)
A1	**91.93%**	8.07%	0.00%
A2	1.84%	**98.16%**	0.00%
A8	0.00%	0.00%	**100.00%**

Size=125	P(A1|x)	P(A2|x)	P(A8|x)
A1	**91.93%**	8.07%	0.00%
A2	0.28%	**99.72%**	0.00%
A8	0.23%	0.02%	**99.75%**

Size=250	P(A1|x)	P(A2|x)	P(A8|x)
A1	**100.00%**	0.00%	0.00%
A2	0.00%	**100.00%**	0.00%
A8	0.00%	0.00%	**100.00%**

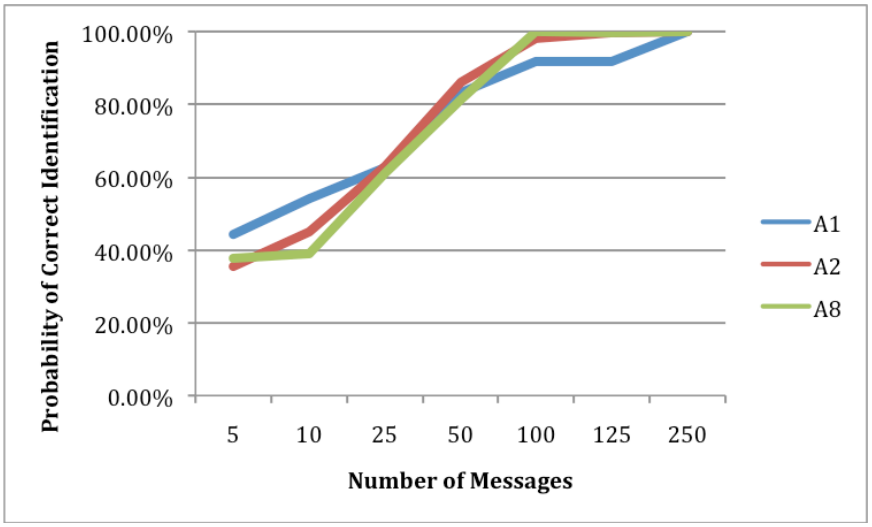


Figure 4-28. Dataset 1, Identification Probability vs. Number of Messages, 3 Sibling Authors

Figure 4-29 shows Dataset #1 PCA plot results for Authors A1 and A12, which are mother and son. They have very similar writeprints with some overlap.

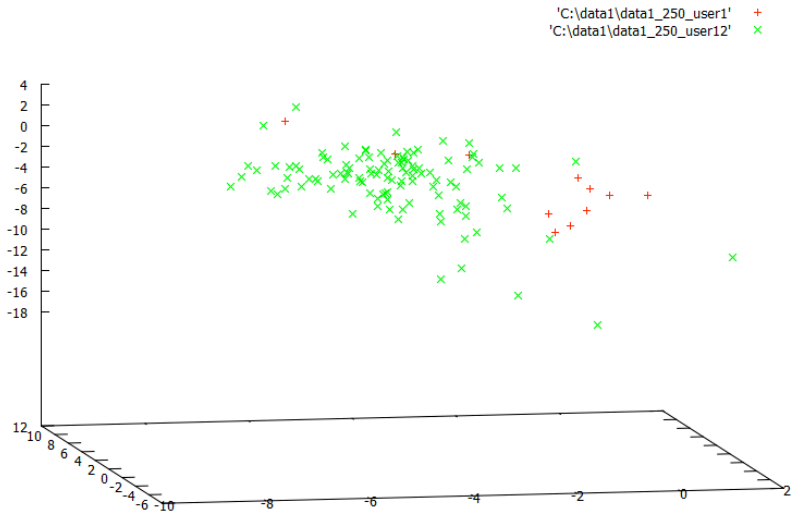


Figure 4-29. Dataset 1, PCA Plot Results, 250 Messages, Authors A1 and A12

Table 4-48 shows the MGD results for conversation sizes of 5, 10, 25, 50, 100, 125, 250, and 500 messages respectively for Authors A1 and A12. Using 100 messages per conversation as input, MGD identifies conversations as the correct author for both authors, with probability ranging from 91.37% to 93.04%. Using 500 messages per conversation as input, MGD identifies conversations as the correct author for both authors with 100% probability. The tables show a significant increase in identification probability as the number of messages per conversation increase. Figure 4-30 shows the relationship between the identification probability and number of messages per conversation.

Table 4-48. Dataset 1, MGD Results, 5-500 Messages, Authors A1 and A12

Size=5	P(A1|x)	P(A12|x)
A1	**61.34%**	38.66%
A12	**55.98%**	44.02%

Size=10	P(A1|x)	P(A12|x)
A1	**63.13%**	36.87%
A12	**50.99%**	49.01%

Size=25	P(A1|x)	P(A12|x)
A1	**67.84%**	32.16%
A12	39.51%	**60.49%**

Size=50	P(A1|x)	P(A12|x)
A1	**78.15%**	21.85%
A12	20.76%	**79.24%**

Size=100	P(A1|x)	P(A12|x)
A1	**91.37%**	8.63%
A12	6.96%	**93.04%**

Size=125	P(A1|x)	P(A12|x)
A1	**88.41%**	11.59%
A12	7.19%	**92.81%**

Size=250	P(A1|x)	P(A12|x)
A1	**99.95%**	0.05%
A12	0.00%	**100.00%**

Size=500	P(A1|x)	P(A12|x)
A1	**100.00%**	0.00%
A12	0.00%	**100.00%**

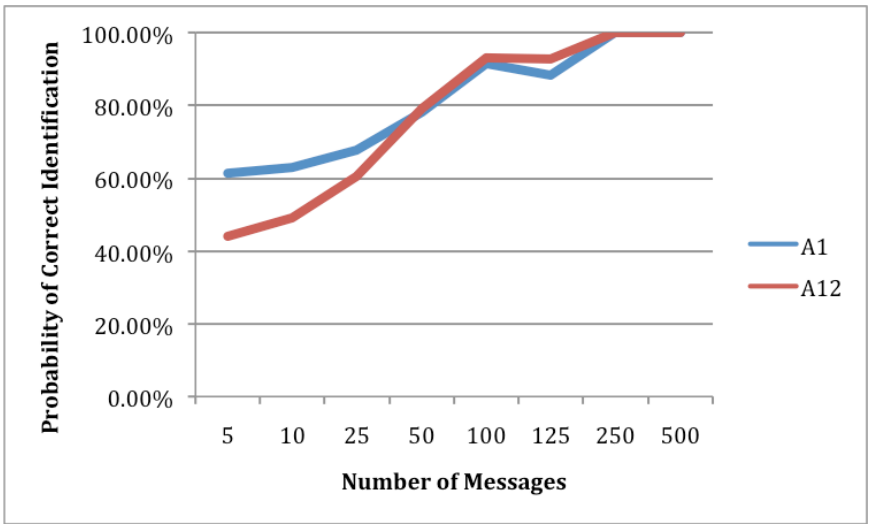


Figure 4-30. Dataset 1, Identification Probability vs. Number of Messages, Authors A1 and A12

Figure 4-31 shows Dataset #1 PCA plot results for Authors A2 and A12, which are mother and daughter. This plot shows separate groupings for each author.

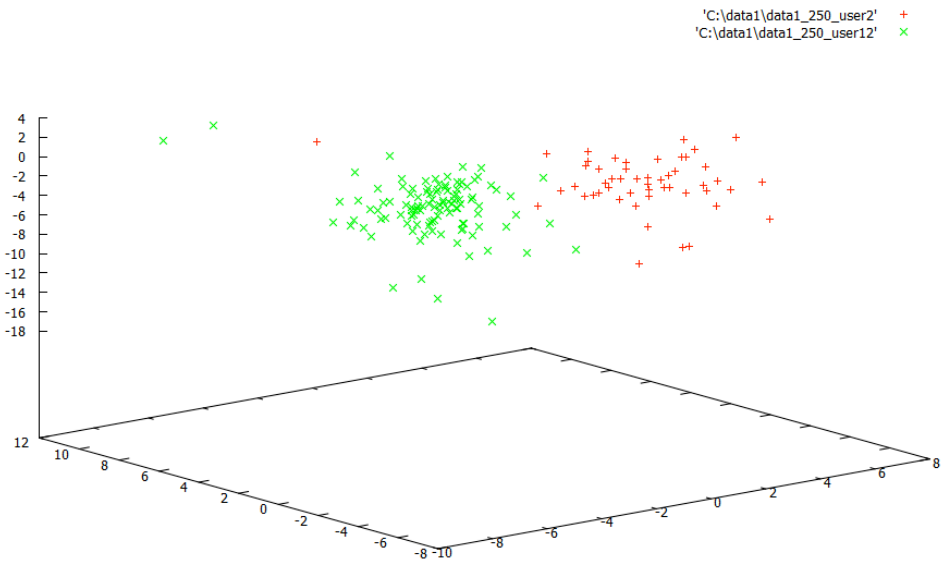


Figure 4-31. Dataset 1 Results, 250 messages, Authors A2 and A12

Table 4-49 shows the MGD results for conversation sizes of 5, 10, 25, 50, 100, 125, 250, and 500 messages respectively for Authors A2 and A12. Using 100 messages per conversation as input, MGD identifies conversations as the correct author for both authors, with probability ranging from 97.87% to 98.36%. Using 500 messages per conversation as input, MGD identifies conversations as the correct author for both authors with 100% probability. The tables show a significant increase in identification probability as the number of messages per conversation increase. Figure 4-32 shows the relationship between the identification probability and number of messages per conversation.

Table 4-49. Dataset 1, MGD Results, 5-500 Messages, Authors A2 and A12

Size=5	P(A2|x)	P(A12|x)
A2	**56.80%**	43.20%
A12	**46.79%**	45.76%

Size=10	P(A2|x)	P(A12|x)
A2	**61.26%**	38.74%
A12	40.03%	**43.22%**

Size=25	P(A2|x)	P(A12|x)
A2	**74.77%**	25.23%
A12	34.79%	**65.21%**

Size=50	P(A2|x)	P(A12|x)
A2	**88.90%**	11.10%
A12	10.09%	**89.91%**

Size=100	P(A2|x)	P(A12|x)
A2	**98.36%**	1.64%
A12	2.13%	**97.87%**

Size=125	P(A2|x)	P(A12|x)
A2	**98.62%**	1.38%
A12	2.22%	**97.78%**

Size=250	P(A2|x)	P(A12|x)
A2	**99.95%**	0.05%
A12	0.00%	**100.00%**

Size=500	P(A2|x)	P(A12|x)
A2	**100.00%**	0.00%
A12	0.00%	**100.00%**

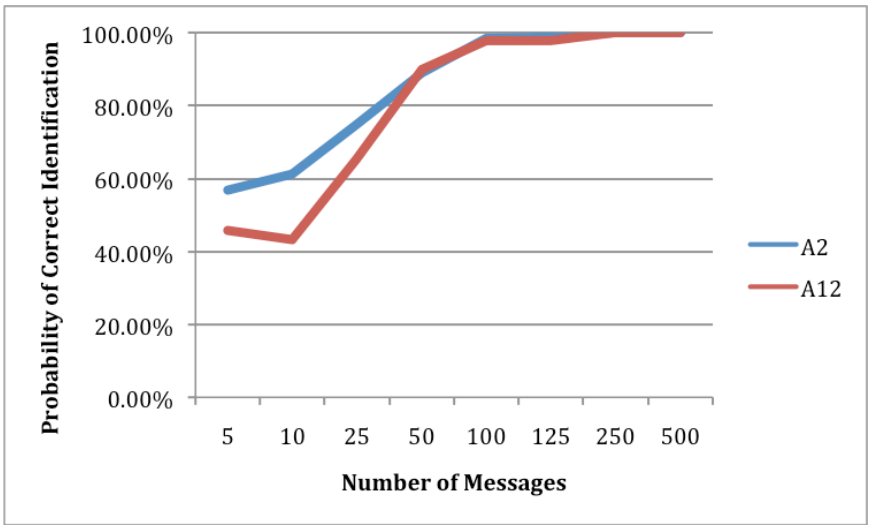


Figure 4-32. Dataset 1, Identification Probability vs. Number of Messages, Authors A2 and A12

Figure 4-33 shows Dataset #1 PCA plot results for Authors A2 and A14, which are authors that are married. The writeprints show some overlap but are still separate for each author.

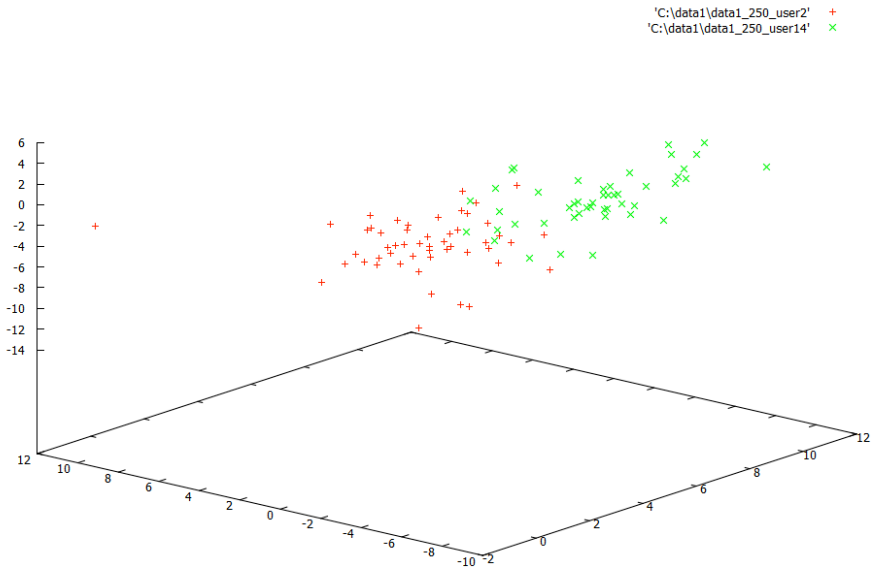


Figure 4-33. Dataset 1 Results, 250 messages, Authors A2 and A14

Table 4-50 shows the MGD results for conversation sizes of 5, 10, 25, 50, 100, 125, 250, and 500 messages respectively for Authors A2 and A14. Using 100 messages per conversation as input, MGD identifies conversations as the correct author for both authors, with probability ranging from 81.10% to 85.01%. Using 500 messages per conversation as input, MGD identifies conversations as the correct author for both authors, with probability ranging from 99.85% to 99.96%. The tables show a significant increase in identification probability as the number of messages per conversation

increase. Figure 4-34 shows the relationship between the identification probability and number of messages per conversation.

Table 4-50. Dataset 1, MGD Results, 5-500 Messages, Authors A2 and A14

Size=5	P(A2|x)	P(A14|x)
A2	**55.17%**	44.83%
A14	**52.02%**	47.98%

Size=10	P(A2|x)	P(A14|x)
A2	**60.22%**	39.78%
A14	**54.98%**	45.02%

Size=25	P(A2|x)	P(A14|x)
A2	**68.99%**	31.01%
A14	**50.88%**	49.12%

Size=50	P(A2|x)	P(A14|x)
A2	**72.59%**	27.41%
A14	31.34%	**68.66%**

Size=100	P(A2|x)	P(A14|x)
A2	**85.01%**	14.99%
A14	18.90%	**81.10%**

Size=125	P(A2|x)	P(A14|x)
A2	**83.43%**	16.57%
A14	16.93%	**83.07%**

Size=250	P(A2|x)	P(A14|x)
A2	**98.30%**	1.70%
A14	6.87%	**93.13%**

Size=500	P(A2|x)	P(A14|x)
A2	**99.96%**	0.04%
A14	0.15%	**99.85%**

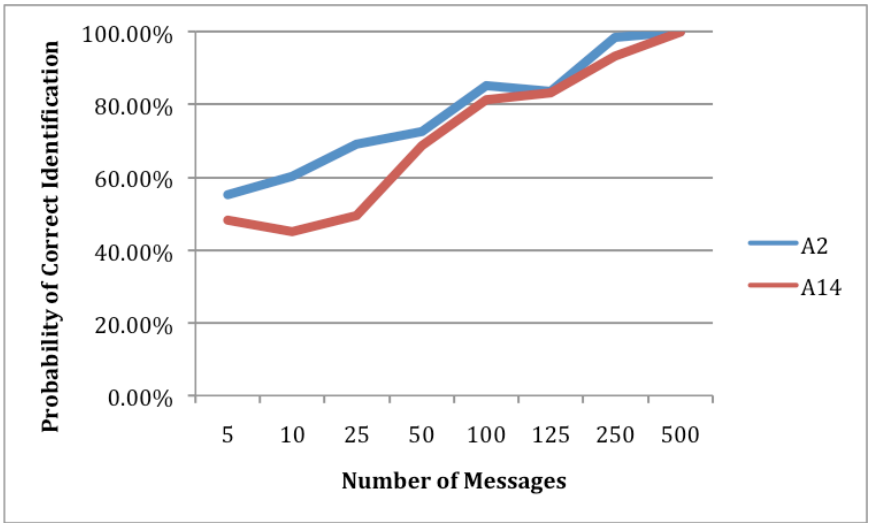


Figure 4-34. Dataset 1, Identification Probability vs. Number of Messages, Authors A2 and A14

4.1.2 Authorship Characterization Results

Authorship characterization attempts to determine whether a given set of IM messages $\{M_1,\ldots,M_q\}$ is likely to be one a of the author categories $\{C_1,\ldots,C_m\}$.

Dataset #1 includes 2 categories (male and female) for gender, 3 categories for age (20s, 30s, >40), and 2 categories for education (high school and college) from which to analyze characterization. Figure 4-35 shows the breakdown of the number of authors for each author category.

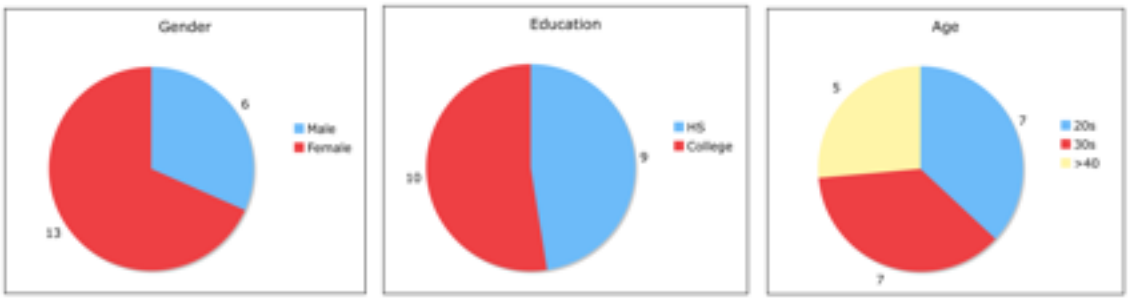


Figure 4-35. Dataset 1 Characterization Breakdown

Dataset #1 experiments include 19 authors from which to determine characterization. Figure 4-36 shows Dataset #1 PCA plot results for the gender category. The conversations consist of 500 messages for each writeprint instance. The plot shows some separation for the gender category.

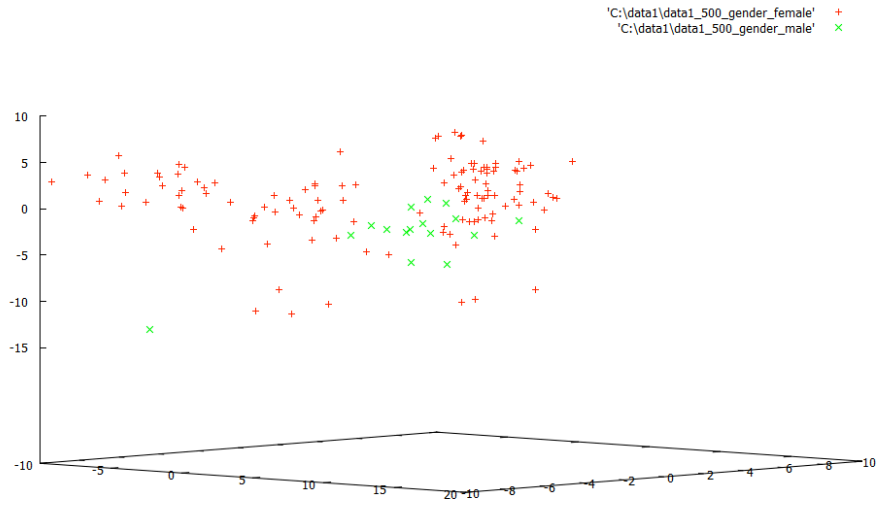


Figure 4-36. Dataset 1, PCA Plot Results, 500 Messages, Gender

Table 4-51 shows the MGD results for conversation sizes of 5, 10, 25, 50, 100, 125, 250, and 500 messages respectively for the gender category. Using 100 messages per conversation as input, MGD identifies conversations as the correct gender, with probability ranging from 74.76% to 88.23%. Using 500 messages per conversation as input, MGD identifies conversations as the correct gender, with probability ranging from 99.19% to 99.96%. The unbalanced gender data (more females than males) may present a slight gender bias at conversations sizes 25 through 250. The tables show a significant increase in characterization probability as the number of messages per conversation increase. Figure 4-37 shows the relationship between the characterization probability and number of messages per conversation.

Table 4-51. Dataset 1, MGD Results, 5-500 Messages, Gender

Size=5	P(M|x)	P(F|x)
Male	**56.86%**	43.14%
Female	**50.31%**	49.69%

Size=10	P(M|x)	P(F|x)
Male	**60.45%**	39.55%
Female	44.53%	**55.47%**

Size=25	P(M|x)	P(F|x)
Male	**63.44%**	36.56%
Female	34.07%	**65.93%**

Size=50	P(M|x)	P(F|x)
Male	**68.49%**	31.51%
Female	21.47%	**78.53%**

Size=100	P(M|x)	P(F|x)
Male	**74.76%**	25.24%
Female	11.77%	**88.23%**

Size=125	P(M|x)	P(F|x)
Male	**78.55%**	21.45%
Female	9.28%	**90.72%**

Size=250	P(M|x)	P(F|x)
Male	**92.66%**	7.34%
Female	3.18%	**96.82%**

Size=500	P(M|x)	P(F|x)
Male	**99.96%**	0.04%
Female	0.81%	**99.19%**

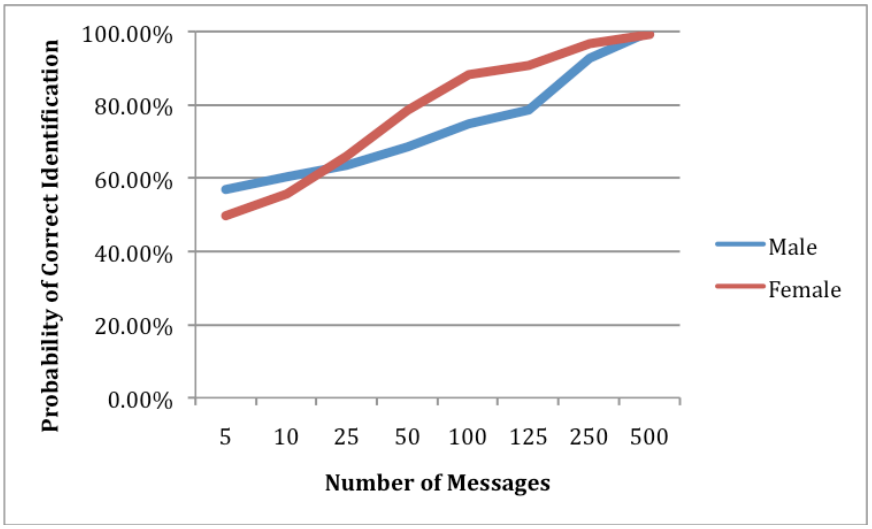


Figure 4-37. Dataset 1, Characterization Probability vs. Number of Messages, Gender

The authorship characterization probability is used to determine the error of the multivariate Gaussian distribution by assessing writeprint false positives. The likelihood, P(x|Category), of the author category of the writperint is used as a minimum threshold. If another author category has a higher likelihood, this is a false positive. Dataset #1 analysis for the gender category achieved less than 20% error using 250 messages per conversation. Figure 4-38 shows that as the conversation size increases, the error rate decreases.

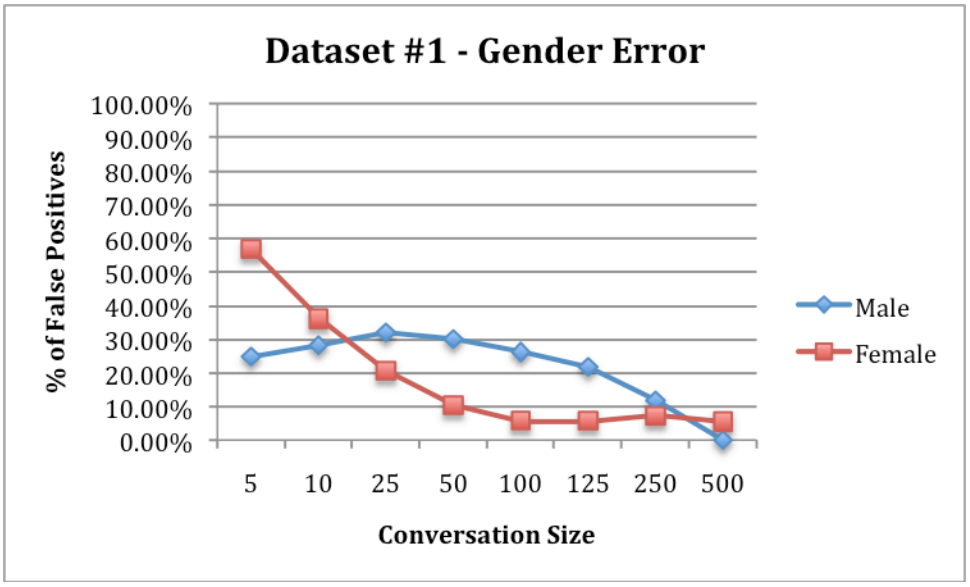


Figure 4-38. Dataset #1, Gender Error

Figure 4-39 shows Dataset #1 PCA plot results for the education category. The conversations consist of 500 messages for each writeprint instance. The plot shows separate groupings for the education category.

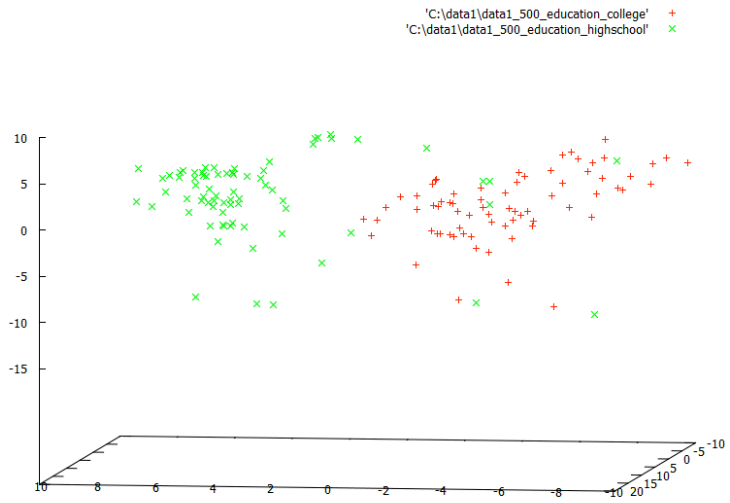


Figure 4-39. Dataset 1, PCA Plot Results, 500 Messages, Education

Table 4-52 shows the MGD results for conversation sizes of 5, 10, 25, 50, 100, 125, 250, and 500 messages respectively for the education category. Using 100 messages per conversation as input, MGD identifies conversations as the correct education, with probability ranging from 89.48% to 95.25%. Using 500 messages per conversation as input, MGD identifies conversations as the correct education, with probability ranging from 97.19% to 97.98%. The tables show a significant increase in characterization probability as the number of messages per conversation increase. Figure 4-40 shows the relationship between the characterization probability and number of messages per conversation.

Table 4-52. Dataset 1, MGD Results, 5-500 Messages, Education

Size=5	P(HS|x)	P(C|x)
High School	48.76%	**51.24%**
College	45.76%	**54.24%**

Size=10	P(HS|x)	P(C|x)
High School	**58.61%**	41.39%
College	41.08%	**58.92%**

Size=25	P(HS|x)	P(C|x)
High School	**72.94%**	27.06%
College	30.78%	**69.22%**

Size=50	P(HS|x)	P(C|x)
High School	**87.19%**	12.81%
College	18.37%	**81.63%**

Size=100	P(HS|x)	P(C|x)
High School	**95.25%**	4.75%
College	10.52%	**89.48%**

Size=125	P(HS|x)	P(C|x)
High School	**96.07%**	3.93%
College	10.35%	**89.65%**

Size=250	P(HS|x)	P(C|x)
High School	**97.92%**	2.08%
College	7.02%	**92.98%**

Size=500	P(HS|x)	P(C|x)
High School	**97.98%**	2.02%
College	2.81%	**97.19%**

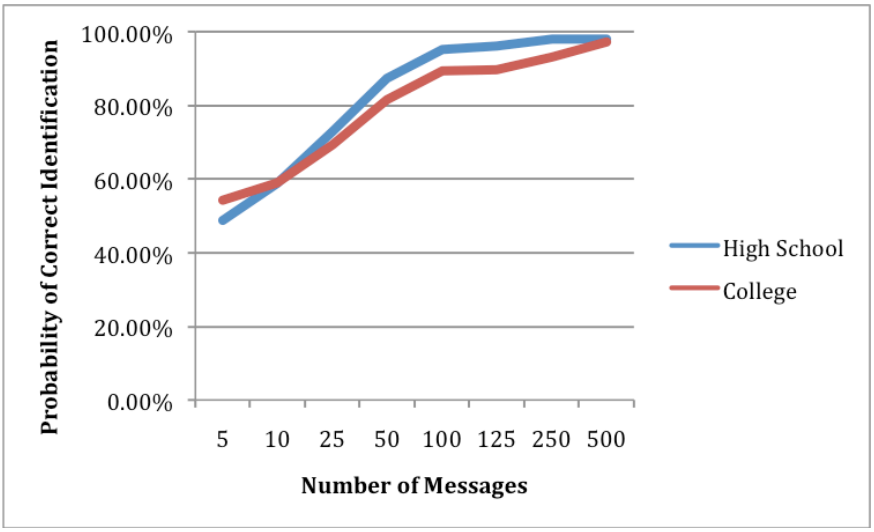


Figure 4-40. Dataset 1, Characterization Probability vs. Number of Messages, Education

Dataset #1 analysis for the education category achieved less than 20% error using 50 messages per conversation. Figure 4-38 shows that as the conversation size increases, the error rate decreases.

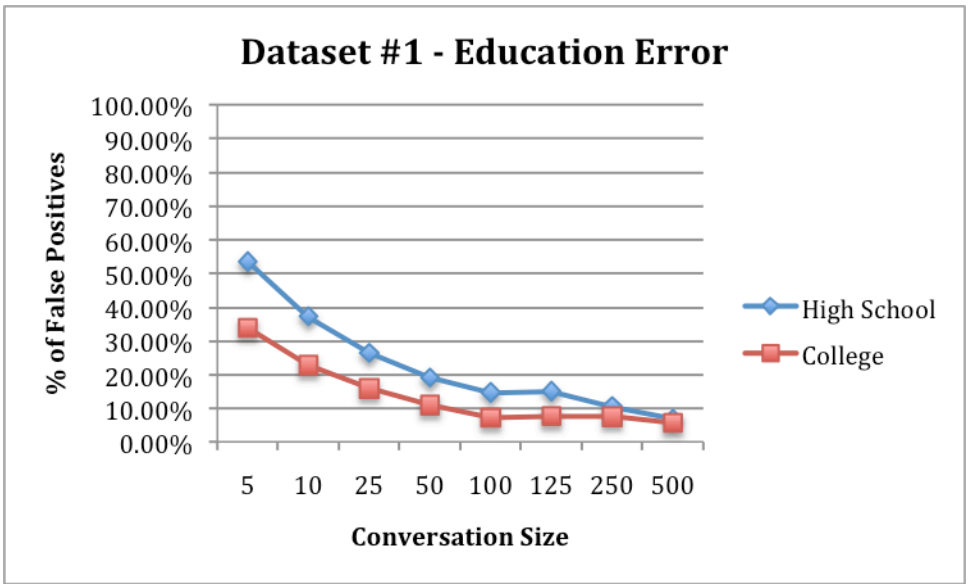


Figure 4-41. Dataset #1, Education Error

Figure 4-42 shows Dataset #1 PCA plot results for the age category. The conversations consist of 500 messages for each writeprint instance. The plot shows some separation for the age category.

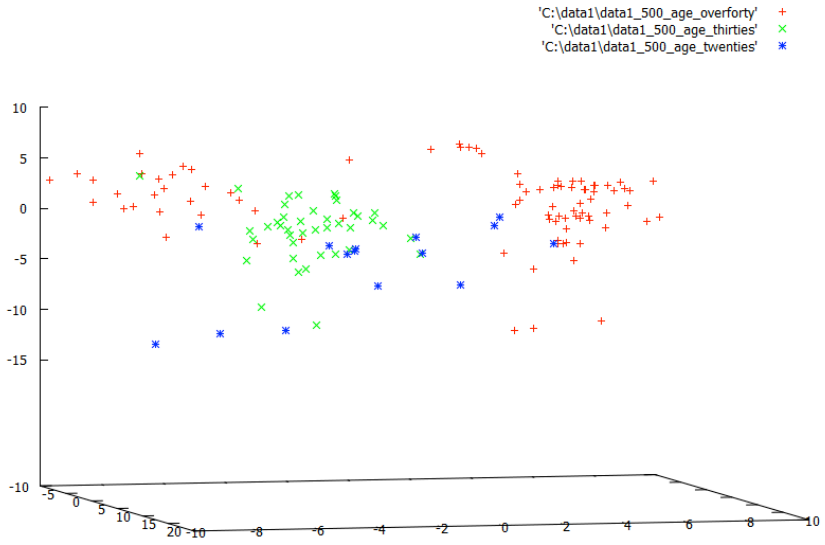


Figure 4-42. Dataset 1, PCA Plot Results, 500 Messages, Age

Table 4-53 shows the MGD results for conversation sizes of 5, 10, 25, 50, 100, 125, 250, and 500 messages respectively for the age category. Using 500 messages per conversation as input, MGD identifies conversations as the correct age, with probability ranging from 75.48% to 99.85%. The age category "Thirties" had the most overlap with the other age categories. The tables show a significant increase in characterization probability as the number of messages per conversation increase. Figure 4-43 shows the relationship between the characterization probability and number of messages per conversation.

Table 4-53. Dataset 1, MGD Results, 5-500 Messages, Age

Size=5	P(T|x)	P(TH|x)	P(OF|x)
Twenties	**40.83%**	32.07%	27.10%
Thirties	**38.83%**	32.77%	28.40%
Over Forty	**37.09%**	30.20%	32.71%

Size=10	P(T|x)	P(TH|x)	P(OF|x)
Twenties	**44.66%**	34.27%	21.07%
Thirties	**38.95%**	34.65%	26.40%
Over Forty	**35.49%**	29.59%	34.92%

Size=25	P(T|x)	P(TH|x)	P(OF|x)
Twenties	**50.96%**	31.72%	17.32%
Thirties	**38.26%**	35.91%	25.83%
Over Forty	33.13%	24.62%	**42.25%**

Size=50	P(T|x)	P(TH|x)	P(OF|x)
Twenties	**60.25%**	25.08%	14.67%
Thirties	**37.34%**	35.94%	26.73%
Over Forty	28.52%	16.97%	**54.51%**

Size=100	P(T|x)	P(TH|x)	P(OF|x)
Twenties	**74.14%**	16.44%	9.42%
Thirties	30.57%	**38.27%**	31.17%
Over Forty	17.80%	9.92%	**72.28%**

Size=125	P(T|x)	P(TH|x)	P(OF|x)
Twenties	**76.62%**	11.63%	11.75%
Thirties	25.83%	**41.23%**	32.94%
Over Forty	13.45%	7.23%	**79.32%**

Size=250	P(T|x)	P(TH|x)	P(OF|x)
Twenties	**90.42%**	8.00%	1.58%
Thirties	23.14%	**44.71%**	32.15%
Over Forty	3.49%	3.22%	**93.29%**

Size=500	P(T|x)	P(TH|x)	P(OF|x)
Twenties	**99.22%**	0.78%	0.00%
Thirties	7.78%	**75.48%**	16.74%
Over Forty	0.01%	0.14%	**99.85%**

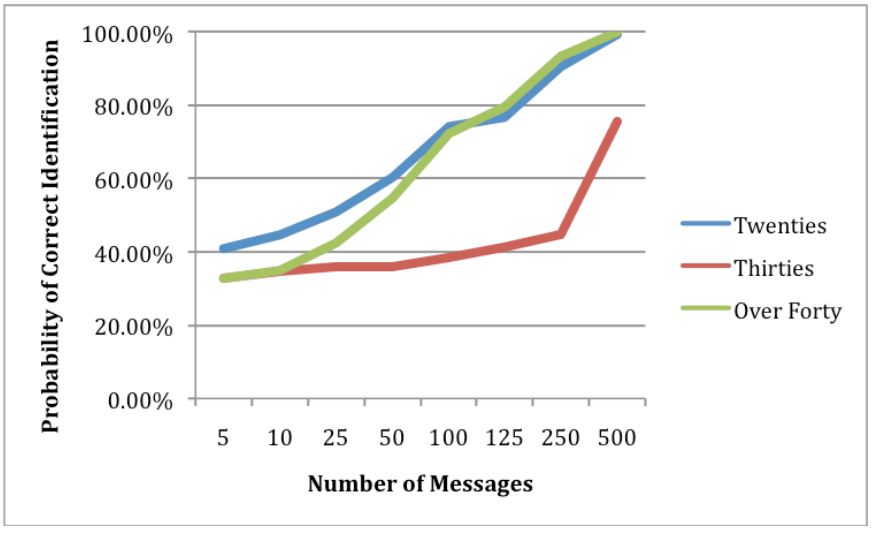


Figure 4-43. Dataset 1, Characterization Probability vs. Number of Messages, Age

Dataset #1 analysis for the age category shows higher error rates due to more overlap between the age categories. The twenties and over forty categories achieved less than 20% error using 250 messages per conversation. The thirties category shows the most overlap, and thus the highest number of false positives. Figure 4-44 shows that as the conversation size increases, the error rate decreases.

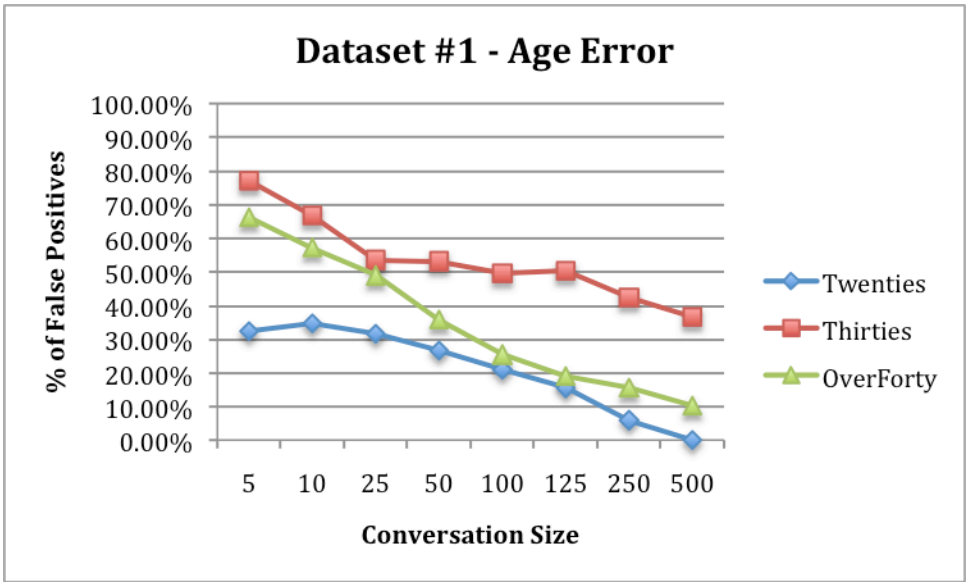


Figure 4-44. Dataset #1, Age Error

4.2 Results for Dataset #2, U.S. Cyberwatch

Dataset #2 experiments include 100 authors. For each author, IM writeprints are divided into conversations containing 10, 25, 50, and 90 messages respectively.

4.2.1 Authorship Identification Results

Authorship identification attempts to determine whether an author A_n of a given set of IM messages $\{M_1,\ldots,M_p\}$ is likely to be one of the author suspects $\{A_1,\ldots,A_n\}$. Dataset #2 experiments include 100 authors from which to determine identification.

Figure 4-45 shows Dataset #2 PCA plot results for the 20 authors with the highest total number of messages (Authors A2, A3, A7, A11, A16, A20, A30, A32, A41, A44, A69, A72, A74, A77, A79, A80, A85, A89, A94, A100, respectively), resulting in the highest number of writeprint instances. The conversations consist of 90 messages for each writeprint instance. This plot does show some separation between the authors.

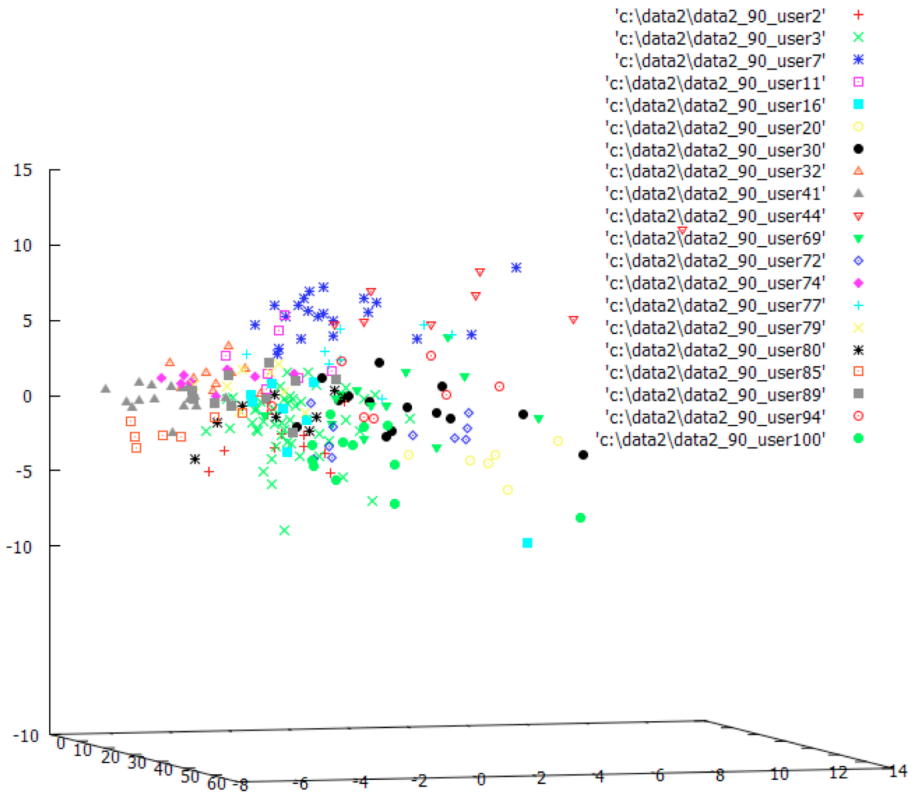


Figure 4-45. Dataset 2, PCA Plot Results, 90 Messages, Top 20 Authors

Table 4-54 through

Table 4-57 shows the MGD results for conversation sizes of 10, 25, 50, and 90 messages respectively for the top 20 authors (Authors A2, A3, A7, A11, A16, A20, A30, A32, A41, A44, A69, A72, A74, A77, A79, A80, A85, A89, A94, A100). Using 90 messages per conversation as input, MGD identifies conversations as the correct author for 19 of the 20 authors, with probability ranging from 71.65% to 100%. The tables show a significant increase in identification probability as the number of messages per conversation increase. Figure 4-46 shows the relationship between the identification probability and number of messages per conversation.

Table 4-54. Dataset 2, MGD Results, 10 Messages, Top 20 Authors (shown in %)

Size=10	P(A2|x)	P(A3|x)	P(A7|x)	P(A11|x)	P(A16|x)	P(A20|x)	P(A30|x)	P(A32|x)	P(A41|x)	P(A44|x)	P(A69|x)	P(A72|x)	P(A74|x)	P(A77|x)	P(A79|x)	P(A80|x)	P(A85|x)	P(A89|x)	P(A94|x)	P(A100|x)
A2	10.67	**13.82**	4.94	1.86	4.15	3.14	5.34	0.03	0.00	2.88	4.67	7.77	1.43	10.02	2.86	5.47	1.80	2.08	6.55	10.52
A3	5.41	**15.67**	9.36	3.48	4.00	1.67	5.48	0.09	0.18	3.58	4.03	3.51	2.22	14.10	3.33	10.17	0.45	3.70	6.27	3.32
A7	1.50	11.93	**25.61**	3.92	1.94	0.45	3.61	0.00	0.00	11.39	2.64	1.53	1.56	15.88	1.35	8.00	0.01	2.02	6.02	0.62
A11	4.08	13.41	8.46	9.02	4.64	0.20	3.02	0.02	0.03	2.87	0.49	0.88	8.41	9.73	7.67	**14.81**	0.14	9.31	2.13	0.67
A16	3.62	8.52	3.47	1.13	**30.77**	0.00	7.76	4.70	1.53	1.02	0.00	0.31	0.62	13.87	0.38	1.49	3.89	10.44	2.29	4.17
A20	5.65	7.85	3.10	0.08	0.84	14.39	3.95	0.00	0.00	2.99	12.10	**16.25**	0.14	6.39	0.14	2.21	0.27	0.89	10.26	12.51
A30	6.93	**14.99**	8.30	4.49	4.00	0.59	8.46	0.03	0.00	4.94	0.70	5.35	2.92	13.17	3.61	5.53	0.39	2.56	8.05	4.98
A32	0.38	0.69	0.00	0.00	33.29	0.00	10.13	**46.48**	0.03	0.00	0.00	0.00	0.00	4.26	0.00	0.00	2.99	0.57	0.00	1.18
A41	0.56	3.62	0.50	1.41	8.27	0.00	0.20	1.29	18.57	0.17	0.00	0.00	9.24	1.44	1.62	11.63	0.08	**41.33**	0.05	0.01
A44	2.33	10.24	17.66	1.37	1.19	1.08	2.39	0.00	0.00	**25.75**	3.29	4.12	0.52	14.23	0.37	3.06	0.02	0.95	8.61	2.83
A69	4.46	10.99	6.97	0.60	1.30	6.38	3.81	0.00	0.00	6.71	**13.16**	9.25	0.31	11.79	0.56	3.62	0.28	1.09	11.12	7.59
A72	5.25	10.81	5.76	0.60	1.37	4.92	3.93	0.00	0.00	4.94	6.24	**15.28**	0.57	10.18	0.68	3.01	0.63	1.56	11.87	12.39
A74	3.32	11.58	4.07	6.44	5.14	0.21	1.71	0.00	2.90	1.87	0.35	0.37	13.74	6.14	7.33	**18.31**	0.26	14.53	0.86	0.88
A77	5.65	**16.34**	9.55	2.55	4.10	0.78	4.93	0.00	0.00	6.37	2.77	5.12	2.60	14.58	4.23	5.18	0.29	3.60	8.92	2.43
A79	6.10	**14.87**	5.77	6.44	5.63	0.14	3.41	0.03	0.13	2.26	0.29	1.86	8.98	10.14	10.57	12.13	0.84	5.93	2.74	1.75

A80	3.73	13.36	6.60	6.18	3.52	1.04	3.15	0.00	0.63	2.33	2.13	2.05	5.18	8.24	2.99	**17.62**	0.12	15.66	3.05	2.42
A85	7.65	9.01	2.02	0.22	6.77	1.56	2.61	2.24	0.04	1.34	3.54	7.86	0.48	6.25	1.22	1.78	**28.76**	1.29	5.52	9.84
A89	2.37	7.82	5.99	3.82	6.33	0.28	1.69	0.05	2.35	2.19	0.33	1.27	8.72	7.96	2.58	16.72	1.30	**25.61**	2.32	0.31
A94	5.86	12.51	8.86	0.69	2.00	3.72	4.57	0.00	0.00	5.80	6.48	8.93	0.88	**13.83**	1.15	4.43	0.50	1.48	9.66	8.65
A100	15.71	11.09	2.40	0.27	2.16	1.79	5.38	0.00	0.00	2.92	1.17	8.43	0.51	6.78	1.03	2.25	1.42	0.56	8.55	**27.58**

Table 4-55. Dataset 2, MGD Results, 25 Messages, Top 20 Authors (shown in %)

Size=25	P(A2|x)	P(A3|x)	P(A7|x)	P(A11|x)	P(A16|x)	P(A20|x)	P(A30|x)	P(A32|x)	P(A41|x)	P(A44|x)	P(A69|x)	P(A72|x)	P(A74|x)	P(A77|x)	P(A79|x)	P(A80|x)	P(A85|x)	P(A89|x)	P(A94|x)	P(A100|x)
A2	17.28	**17.30**	1.02	0.73	1.57	3.36	8.63	0.00	0.00	1.08	8.94	2.97	0.03	5.22	1.02	0.98	1.27	4.58	9.68	14.32
A3	6.09	**39.01**	4.90	2.58	1.90	0.40	9.90	0.00	0.03	3.84	2.34	0.52	1.96	7.80	2.81	4.50	0.03	1.47	7.82	2.11
A7	0.02	10.94	**56.65**	1.58	0.25	0.01	1.42	0.00	0.00	16.79	0.69	0.01	0.02	8.51	0.12	0.84	0.00	0.04	2.12	0.02
A11	3.82	24.09	4.08	**30.38**	4.63	0.19	4.75	0.00	0.00	0.38	0.19	0.00	1.67	5.92	11.89	4.14	0.03	1.56	2.10	3.82
A16	0.60	4.50	0.01	0.01	**78.28**	0.00	5.70	2.04	0.02	0.00	0.00	0.00	0.36	0.00	0.00	0.12	5.02	1.85	0.01	0.60
A20	2.27	8.79	0.15	0.00	0.06	**22.15**	3.26	0.00	0.00	1.29	16.90	15.95	0.00	0.10	0.00	0.10	0.02	0.05	21.76	2.27
A30	3.68	23.49	3.81	1.57	1.47	0.00	**30.40**	0.00	0.00	2.95	0.00	1.09	0.16	13.03	0.60	0.22	0.16	0.37	9.86	7.15
A32	0.00	0.00	0.00	0.00	15.07	0.00	9.86	**75.07**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A41	0.04	0.21	0.00	0.00	2.40	0.00	0.00	0.01	36.66	0.00	0.00	0.00	3.46	0.00	0.00	9.25	0.00	**47.97**	0.00	0.00
A44	0.02	7.76	11.84	0.00	0.07	0.04	2.46	0.00	0.00	**60.38**	1.67	0.02	0.00	14.22	0.00	0.02	0.00	0.01	1.35	0.15
A69	1.83	9.81	6.27	0.02	0.13	10.74	5.77	0.00	0.00	6.18	**29.87**	6.80	0.00	4.50	0.00	0.28	0.05	0.03	14.29	3.42
A72	2.72	3.86	0.11	0.01	0.06	9.50	6.36	0.00	0.00	0.85	10.44	20.49	0.00	7.96	0.00	0.10	0.02	0.10	**32.48**	4.95
A74	0.65	9.04	0.49	11.73	1.80	0.01	1.35	0.00	4.02	2.30	0.02	0.01	20.60	4.46	5.89	**20.90**	0.00	15.74	0.96	0.03
A77	4.23	16.29	13.85	1.88	1.50	0.01	5.46	0.00	0.00	12.12	0.09	1.31	0.26	**23.96**	0.81	2.90	0.01	0.54	13.91	0.87
A79	9.05	18.20	0.64	13.04	6.57	0.01	1.47	0.00	0.00	0.13	0.01	0.00	7.15	5.01	**19.68**	8.95	1.73	4.95	2.04	1.39
A80	0.33	20.00	1.16	2.60	0.08	0.01	1.70	0.00	0.32	0.14	0.03	0.17	0.77	0.81	0.68	**53.02**	0.00	16.39	1.76	0.01
A85	5.86	3.02	0.01	0.00	2.34	1.49	5.70	0.08	0.00	0.17	2.21	2.50	0.00	1.23	0.01	0.03	**66.29**	0.04	2.31	6.72
A89	0.31	1.04	0.01	0.55	3.52	0.00	0.00	0.00	2.02	0.00	0.00	0.01	9.96	0.54	1.40	28.36	0.11	**51.63**	0.50	0.02
A94	3.34	8.13	7.73	6.26	0.14	4.25	7.18	0.00	0.00	2.89	10.78	6.89	1.42	7.02	2.09	1.51	0.08	1.04	**24.39**	4.85
A100	19.79	13.18	0.05	0.17	0.84	0.26	8.67	0.00	0.00	0.38	0.11	4.58	0.04	2.23	0.05	0.13	1.87	0.05	4.35	**43.27**

Table 4-56. Dataset 2, MGD Results, 50 Messages, Top 20 Authors (shown in %)

Size=50	P(A2|x)	P(A3|x)	P(A7|x)	P(A11|x)	P(A16|x)	P(A20|x)	P(A30|x)	P(A32|x)	P(A41|x)	P(A44|x)	P(A69|x)	P(A72|x)	P(A74|x)	P(A77|x)	P(A79|x)	P(A80|x)	P(A85|x)	P(A89|x)	P(A94|x)	P(A100|x)
A2	**37.93**	4.77	0.62	0.04	0.08	0.13	0.36	0.00	0.00	0.01	0.31	24.47	0.00	0.22	0.01	0.09	0.00	0.00	16.52	14.43
A3	0.29	**94.52**	0.00	0.01	0.01	0.00	0.42	0.00	0.00	0.22	0.06	0.02	0.00	0.00	0.00	0.61	0.00	0.08	3.76	0.00
A7	0.03	1.77	**72.56**	0.67	0.00	0.00	0.08	0.00	0.00	17.04	0.24	0.00	0.00	4.92	0.00	0.11	0.00	0.00	2.57	0.00
A11	3.82	0.39	0.88	**75.83**	0.27	0.01	0.44	0.00	0.00	0.16	0.00	0.00	0.84	0.50	2.24	10.95	0.00	1.90	1.77	0.00
A16	1.27	0.00	0.00	0.00	**83.68**	0.00	11.77	0.05	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.69	0.00	0.45
A20	0.87	0.00	0.00	0.00	0.00	**30.99**	22.31	0.00	0.00	0.00	2.77	10.62	0.00	0.00	0.00	0.01	0.00	0.00	20.70	11.72
A30	0.21	0.02	0.03	0.01	0.00	0.00	**79.16**	0.00	0.00	1.77	0.00	0.10	0.02	12.29	0.00	0.00	0.00	0.00	4.01	2.38
A32	0.00	0.00	0.00	0.00	0.10	0.00	0.87	**99.03**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A41	0.00	0.00	0.00	0.00	1.22	0.00	0.00	0.00	**61.08**	0.00	0.00	0.00	0.09	0.00	0.00	1.00	0.00	36.61	0.00	0.00
A44	0.00	0.00	0.33	0.00	0.00	0.00	0.11	0.00	0.00	**95.00**	0.00	0.01	0.00	4.53	0.00	0.00	0.00	0.00	0.02	0.00
A69	0.78	4.73	3.78	0.00	0.00	1.03	2.34	0.00	0.00	0.00	**56.39**	3.78	0.00	0.00	0.00	0.00	0.00	0.00	27.14	0.02
A72	1.27	0.32	0.01	0.00	0.00	12.53	12.81	0.00	0.00	0.45	6.30	**36.95**	0.00	0.03	0.00	0.00	0.00	0.00	16.84	12.49
A74	0.00	0.01	0.60	0.41	0.15	0.00	0.39	0.00	2.52	0.02	0.00	0.01	**39.76**	7.81	5.51	11.67	0.00	29.92	1.23	0.00
A77	0.07	0.02	1.15	2.36	0.00	0.00	21.36	0.00	0.00	12.27	0.00	6.23	0.02	**44.19**	1.31	0.32	0.00	0.07	10.53	0.11
A79	0.29	0.04	0.02	36.29	2.36	0.00	4.61	0.00	0.00	0.44	0.00	0.00	0.79	4.41	**47.15**	0.25	0.00	1.43	0.15	1.77
A80	0.01	3.49	0.01	0.01	0.00	0.00	0.04	0.00	0.28	0.00	0.00	0.00	0.99	0.00	0.00	**94.44**	0.00	0.70	0.04	0.00
A85	0.01	0.00	0.00	0.00	0.00	0.00	1.51	0.00	0.00	0.20	0.00	2.82	0.00	0.00	0.00	0.00	**93.29**	0.00	0.05	2.10
A89	0.00	0.01	0.00	0.01	0.27	0.00	0.00	0.00	17.90	0.01	0.00	0.00	1.24	0.06	0.00	13.67	0.00	**66.71**	0.11	0.00
A94	3.16	3.43	4.75	0.01	0.00	8.23	7.22	0.00	0.00	0.12	3.89	4.30	0.00	0.78	0.00	0.70	0.00	0.00	**62.92**	0.48
A100	22.65	0.01	0.00	1.14	0.00	0.03	5.46	0.00	0.00	0.01	0.00	0.46	0.00	0.06	0.01	0.01	0.00	0.40	0.91	**68.83**

Table 4-57. Dataset 2, MGD Results, 90 Messages, Top 20 Authors (shown in %)

Size=90	P(A2|x)	P(A3|x)	P(A7|x)	P(A11|x)	P(A16|x)	P(A20|x)	P(A30|x)	P(A32|x)	P(A41|x)	P(A44|x)	P(A69|x)	P(A72|x)	P(A74|x)	P(A77|x)	P(A79|x)	P(A80|x)	P(A85|x)	P(A89|x)	P(A94|x)	P(A100|x)
A2	**81.05**	0.00	0.00	0.01	0.00	0.00	10.63	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.82
A3	0.00	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A7	0.00	0.00	**99.28**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.72	0.00	0.00	0.00	0.00	0.00	0.00
A11	0.00	0.00	0.03	**95.85**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.54	1.38	0.18	0.00	0.00	0.01	0.00	0.00
A16	0.00	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A20	0.00	0.00	0.00	0.00	0.00	**97.22**	0.84	0.00	0.00	0.00	0.00	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.60
A30	0.00	0.00	0.00	0.00	0.00	0.00	**99.97**	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.01
A32	0.00	0.00	0.00	0.00	0.02	0.00	0.00	**99.98**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A41	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**97.22**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.78	0.00	0.00
A44	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**80.32**	0.00	0.00	0.00	19.67	0.00	0.00	0.00	0.00	0.00	0.00
A69	0.00	1.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**98.43**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A72	0.63	0.00	0.00	0.00	0.00	0.04	26.27	0.00	0.00	0.00	0.00	**71.65**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.41
A74	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.76	0.00	0.00	0.00	**98.74**	0.01	0.04	0.31	0.00	0.07	0.00	0.00
A77	0.00	0.00	**50.62**	0.00	0.00	0.00	16.62	0.00	0.00	0.57	0.00	0.00	0.00	32.14	0.00	0.00	0.00	0.00	0.00	0.05
A79	0.06	0.00	0.00	2.17	0.03	0.00	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.07	**97.17**	0.00	0.00	0.05	0.01	0.14
A80	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**99.93**	0.00	0.00	0.00	0.00
A85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00
A89	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.09	0.00	0.00	0.00	0.15	0.00	0.00	2.78	0.00	**88.99**	0.00	0.00
A94	0.11	0.00	0.63	0.00	0.00	0.00	0.87	0.00	0.00	0.00	0.51	1.89	0.00	0.00	0.00	0.03	0.00	0.00	**95.98**	0.00
A100	0.36	0.00	0.00	0.00	0.00	0.00	8.71	0.00	0.00	0.00	0.00	1.27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**89.66**

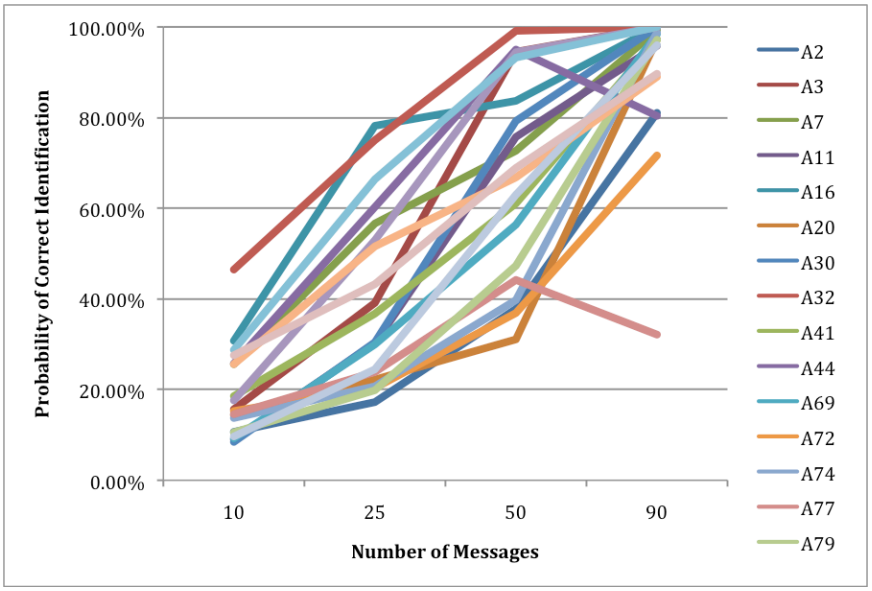


Figure 4-46. Dataset 2, Identification Probability vs. Number of Messages, Top 20 Authors

Dataset #2 analysis for the top 20 authors achieved less than 20% error for most authors using 90 messages per conversation. Figure 4-47 shows that as the conversation size increases, the error rate decreases.

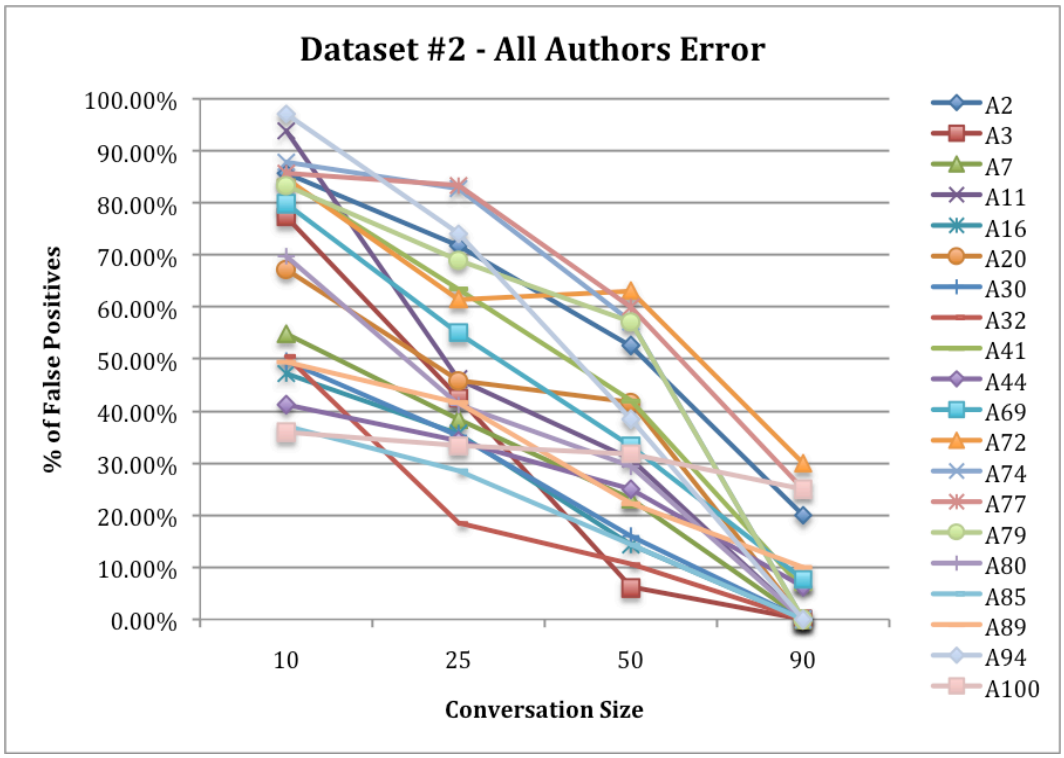


Figure 4-47. Dataset #2, Top 20 Authors Error

Figure 4-48 shows Dataset #2 PCA plot results for the 6 authors with the highest total number of messages (Authors A3, A7, A41, A30, A69, A100, respectively), resulting in the highest number of writeprint instances. The conversations consist of 90

messages for each writeprint instance. This plot does show some separation between the authors.

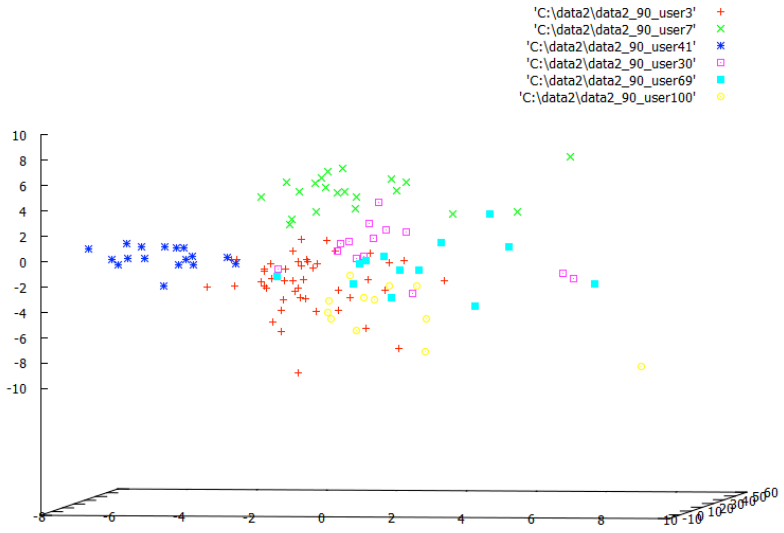


Figure 4-48. Dataset 2, PCA Plot Results, 90 Messages, Top 6 Authors

Table 4-58 through

Table 4-61 shows the MGD results for conversation sizes of 10, 25, 50, and 90 messages for the top 6 authors (Authors A3, A7, A41, A30, A69, A100, respectively). Using 50 messages per conversation as input, MGD identifies conversations as the correct author for all 6 authors, with probability ranging from 83.84% to 100%. Using 90

messages per conversation as input, MGD identifies conversations as the correct author for all 6 authors, with probability ranging from 91.15% to 100%. The tables show a significant increase in identification probability as the number of messages per conversation increase. Figure 4-49 shows the relationship between the identification probability and number of messages per conversation.

Table 4-58. Dataset 2, MGD Results, 10 Messages, Top 6 Authors

Size=10	P(A3|x)	P(A7|x)	P(A30|x)	P(A41|x)	P(A69|x)	P(A100|x)
A3	**41.19%**	24.61%	14.41%	0.46%	10.60%	8.72%
A7	26.86%	**57.66%**	8.14%	0.00%	5.95%	1.40%
A30	**40.05%**	22.18%	22.59%	0.01%	1.86%	13.31%
A41	15.80%	2.19%	0.86%	**81.09%**	0.00%	0.05%
A69	25.84%	16.40%	8.96%	0.00%	**30.95%**	17.85%
A100	23.29%	5.05%	11.29%	0.00%	2.46%	**57.92%**

Table 4-59. Dataset 2, MGD Results, 25 Messages, Top 6 Authors

Size=25	P(A3|x)	P(A7|x)	P(A30|x)	P(A41|x)	P(A69|x)	P(A100|x)
A3	**66.93%**	8.40%	16.98%	0.04%	4.02%	3.62%
A7	15.69%	**81.27%**	2.04%	0.00%	0.98%	0.01%
A30	36.22%	5.87%	**46.87%**	0.00%	0.00%	11.03%
A41	0.56%	0.00%	0.00%	**99.44%**	0.00%	0.00%
A69	17.79%	11.38%	10.46%	0.00%	**54.16%**	6.21%
A100	20.19%	0.08%	13.28%	0.00%	0.17%	**66.29%**

Table 4-60. Dataset 2, MGD Results, 50 Messages, Top 6 Authors

Size=50	P(A3|x)	P(A7|x)	P(A30|x)	P(A41|x)	P(A69|x)	P(A100|x)
A3	**99.49%**	0.00%	0.44%	0.00%	0.06%	0.00%
A7	2.37%	**97.19%**	0.11%	0.00%	0.32%	0.00%

A30	0.02%	0.04%	**97.02%**	0.00%	0.00%	2.92%
A41	0.00%	0.00%	0.00%	**100.00%**	0.00%	0.00%
A69	7.04%	5.62%	3.48%	0.00%	**83.84%**	0.02%
A100	0.02%	0.00%	7.34%	0.00%	0.00%	**92.64%**

Table 4-61. Dataset 2, MGD Results, 90 Messages, Top 6 Authors

Size=90	P(A3|x)	P(A7|x)	P(A30|x)	P(A41|x)	P(A69|x)	P(A100|x)
A3	**100.00%**	0.00%	0.00%	0.00%	0.00%	0.00%
A7	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%
A30	0.00%	0.00%	**99.99%**	0.00%	0.00%	0.01%
A41	0.00%	0.00%	0.00%	**100.00%**	0.00%	0.00%
A69	1.57%	0.00%	0.00%	0.00%	**98.43%**	0.00%
A100	0.00%	0.00%	8.85%	0.00%	0.00%	**91.15%**

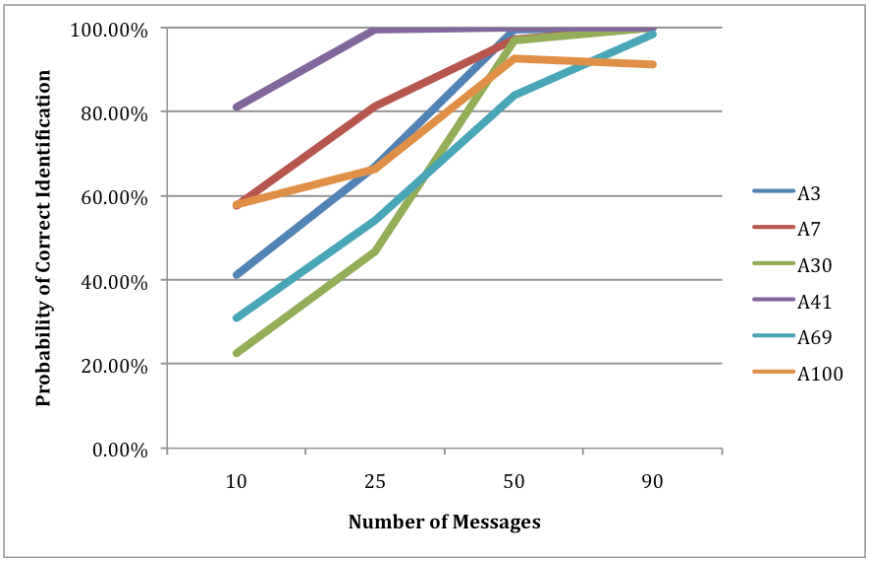


Figure 4-49. Dataset 2, Identification Probability vs. Number of Messages, Top 6 Authors

Dataset #2 analysis for the top 6 authors achieved less than 20% error for most authors using 50 messages per conversation. Figure 4-50 shows that as the conversation size increases, the error rate decreases.

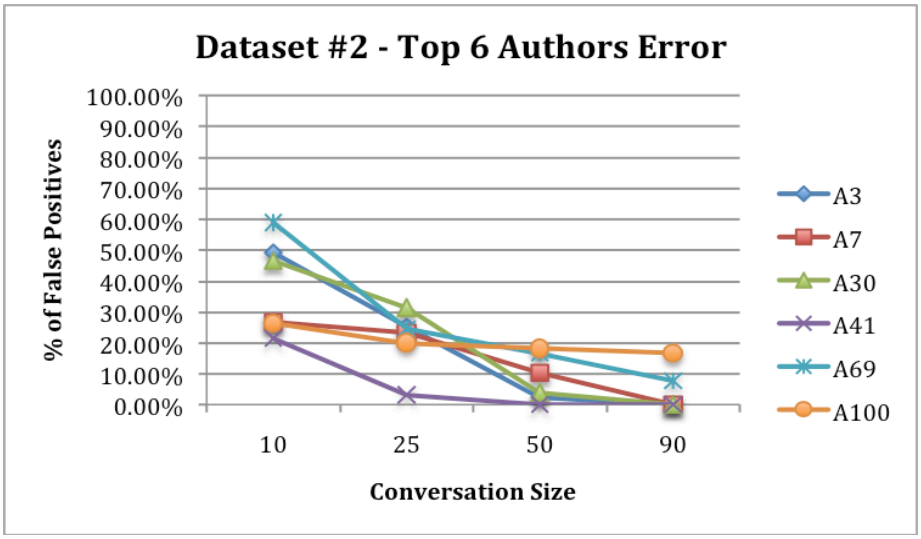


Figure 4-50. Dataset #2, Top 6 Authors Error

Figure 4-51 shows Dataset #2 PCA plot results for 3 authors with the highest total number of messages (Authors A3, A7, A41, respectively). The conversations consist of 90 messages for each writeprint instance. This plot shows separate groupings for each author.

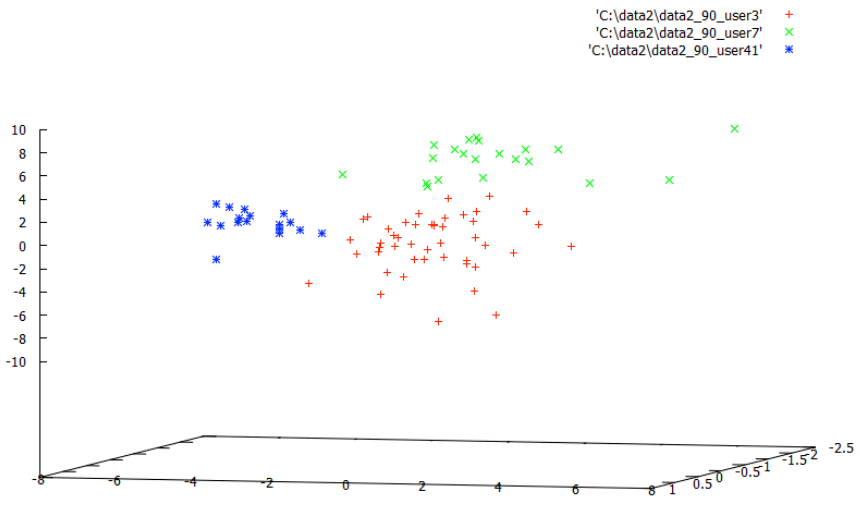


Figure 4-51. Dataset 2, PCA Plot Results, 90 Messages, Top 3 Authors - Subset 1

Table 4-62 shows the MGD results for conversation sizes of 10, 25, 50, and 90 messages respectively for the top 3 authors (Authors A3, A7, A41, respectively). Using 25 messages per conversation as input, MGD identifies conversations as the correct author for all 3 authors, with probability ranging from 83.81% to 99.44%. Using 90 messages per conversation as input, MGD identifies conversations as the correct author for all 3 authors, with 100% probability. The tables show a significant increase in identification probability as the number of messages per conversation increase. Figure 4-52 shows the relationship between the identification probability and number of messages per conversation.

Table 4-62. Dataset 2, MGD Results, 10-90 Messages, Top 3 Authors – Subset 1

Size=10	P(A3|x)	P(A7|x)	P(A41|x)
A3	**62.16%**	37.14%	0.70%
A7	31.78%	**68.22%**	0.00%
A41	15.95%	2.21%	**81.84%**

Size=25	P(A3|x)	P(A7|x)	P(A41|x)
A3	**88.80%**	11.14%	0.06%
A7	16.19%	**83.81%**	0.00%
A41	0.56%	0.00%	**99.44%**

Size=50	P(A3|x)	P(A7|x)	P(A41|x)
A3	**100.00%**	0.00%	0.00%
A7	2.38%	**97.62%**	0.00%
A41	0.00%	0.00%	**100.00%**

Size=90	P(A3|x)	P(A7|x)	P(A41|x)
A3	**100.00%**	0.00%	0.00%
A7	0.00%	**100.00%**	0.00%
A41	0.00%	0.00%	**100.00%**

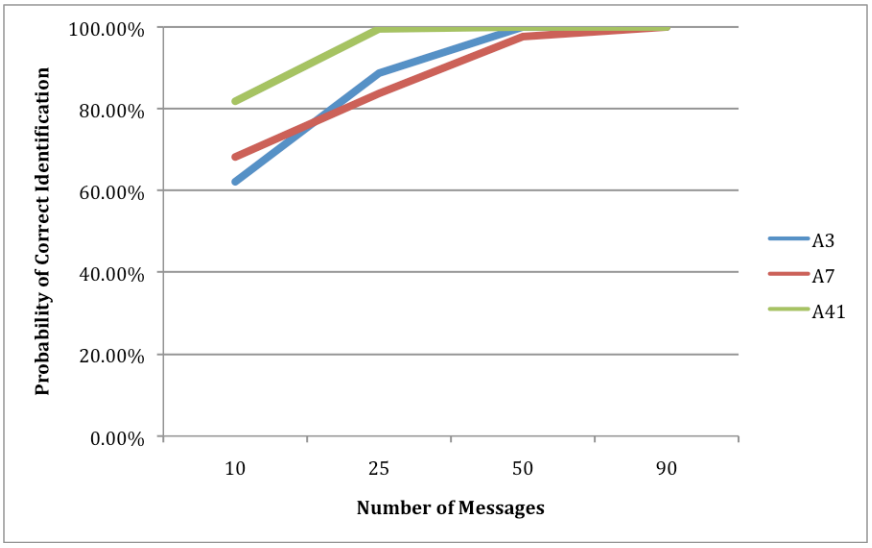


Figure 4-52. Dataset 3, Identification Probability vs. Number of Messages, Top 6 Authors - Subset 1

Figure 4-53 shows Dataset #2 PCA plot results for the second top three authors (Authors A30, A69, A100, respectively). The conversations consist of 90 messages for each writeprint instance. This plot shows separate groupings with more overlap between these authors.

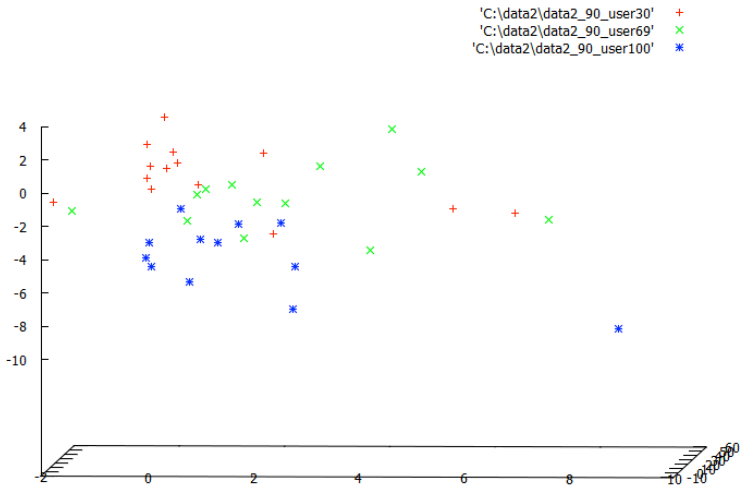


Figure 4-53. Dataset 2, PCA Plot Results, 90 Messages, Top 6 Authors - Subset 2

Table 4-63 shows the MGD results for conversation sizes of 10, 25, 50, and 90 messages respectively for the second top 3 authors (Authors A30, A69, A100, respectively). Using 25 messages per conversation as input, MGD identifies conversations as the correct author for all 3 authors, with probability ranging from 76.47% to 83.14%. Using 90 messages per conversation as input, MGD identifies conversations as the correct author for all 3 authors, with probability ranging from 91.15% to 100%. The tables show a significant increase in identification probability as the number of messages per conversation increase. Figure 4-54 shows the relationship between the identification probability and number of messages per conversation.

Table 4-63. Dataset 2, MGD Results, 10-90 Messages, Top 6 Authors – Subset 2

Size=10	P(A30|x)	P(A69|x)	P(A100|x)
A30	**59.82%**	4.92%	35.26%
A69	15.51%	**53.58%**	30.91%
A100	15.75%	3.43%	**80.82%**

Size=25	P(A30|x)	P(A69|x)	P(A100|x)
A30	**80.94%**	0.00%	19.05%
A69	14.77%	**76.47%**	8.76%
A100	16.65%	0.21%	**83.14%**

Size=50	P(A30|x)	P(A69|x)	P(A100|x)
A30	**97.08%**	0.00%	2.92%
A69	3.98%	**95.99%**	0.03%
A100	7.34%	0.00%	**92.66%**

Size=90	P(A30|x)	P(A69|x)	P(A100|x)
A30	**99.99%**	0.00%	0.01%
A69	0.00%	**100.00%**	0.00%
A100	8.85%	0.00%	**91.15%**

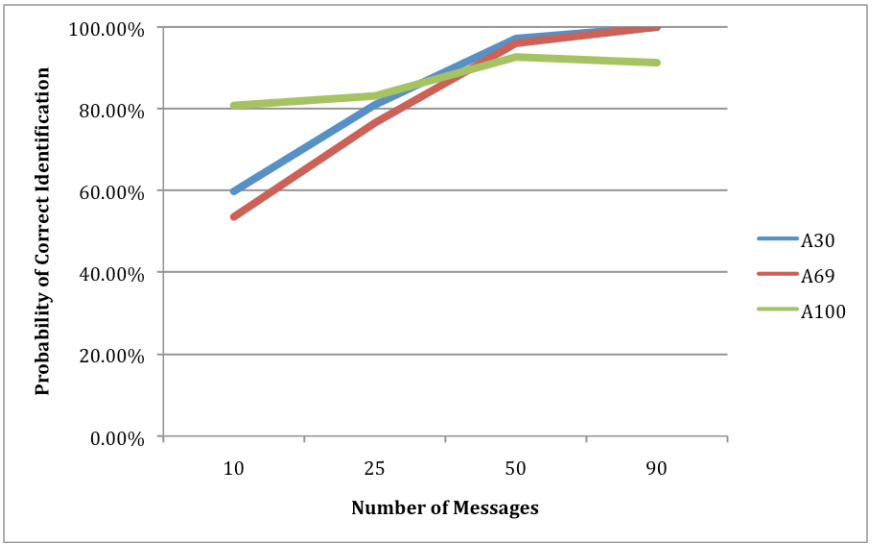


Figure 4-54. Dataset 2, Identification Probability vs. Number of Messages, Top 6 Authors - Subset 2

Figure 4-55 shows Dataset #2 PCA plot results for the next 6 authors with the highest total number of messages (Authors A72, A2, A32, A89, A80, A44, respectively). The conversations consist of 90 messages for each writeprint instance. This plot does show some separation between the authors.

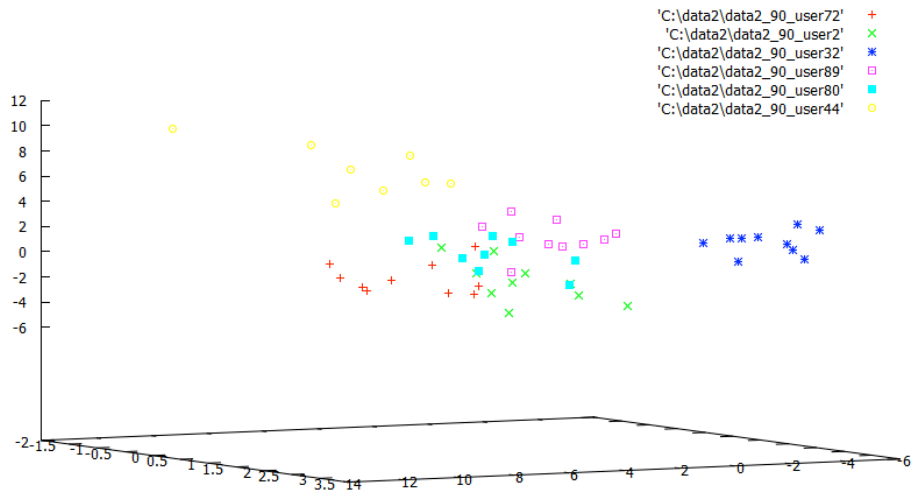


Figure 4-55. Dataset 2, PCA Plot Results, 90 Messages, Second Top 6 Authors

Table 4-64 through

Table 4-67 show the MGD results for conversation sizes of 10, 25, 50, and 90 messages respectively for the next top 6 authors (Authors A72, A2, A32, A89, A80, A44). Using 25 messages per conversation as input, MGD identifies conversations as the correct author for all 6 authors, with probability ranging from 64.26% to 100%. Using 90 messages per conversation as input, MGD identifies conversations as the correct author for all 6 authors, with probability ranging from 96.97% to 100%. The tables show a significant increase in identification probability as the number of messages per

conversation increase. Figure 4-56 shows the relationship between the identification probability and number of messages per conversation.

Table 4-64. Dataset 2, MGD Results, 10 Messages, Second Top 6 Authors

Size=10	P(A2|x)	P(A32|x)	P(A44|x)	P(A72|x)	P(A80|x)	P(A89|x)
A2	**36.93%**	0.11%	9.96%	26.88%	18.92%	7.20%
A32	0.80%	**97.99%**	0.00%	0.00%	0.00%	1.21%
A44	6.43%	0.00%	**71.11%**	11.37%	8.46%	2.63%
A72	17.46%	0.00%	16.46%	**50.87%**	10.01%	5.19%
A80	9.00%	0.00%	5.64%	4.95%	**42.57%**	37.84%
A89	4.92%	0.09%	4.55%	2.64%	34.68%	**53.11%**

Table 4-65. Dataset 2, MGD Results, 25 Messages, Second Top 6 Authors

Size=25	P(A2|x)	P(A32|x)	P(A44|x)	P(A72|x)	P(A80|x)	P(A89|x)
A2	**64.26%**	0.00%	4.01%	11.04%	3.66%	17.02%
A32	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%
A44	0.03%	0.00%	**99.90%**	0.03%	0.04%	0.01%
A72	11.22%	0.00%	3.50%	**84.47%**	0.42%	0.40%
A80	0.47%	0.00%	0.20%	0.24%	**75.69%**	23.41%
A89	0.38%	0.00%	0.00%	0.01%	35.32%	**64.29%**

Table 4-66. Dataset 2, MGD Results, 50 Messages, Second Top 6 Authors

Size=50	P(A2|x)	P(A32|x)	P(A44|x)	P(A72|x)	P(A80|x)	P(A89|x)
A2	**60.68%**	0.00%	0.01%	39.15%	0.15%	0.01%
A32	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%
A44	0.00%	0.00%	**99.99%**	0.01%	0.00%	0.00%
A72	3.28%	0.00%	1.16%	**95.56%**	0.00%	0.00%
A80	0.01%	0.00%	0.00%	0.00%	**99.26%**	0.73%
A89	0.00%	0.00%	0.01%	0.00%	17.01%	**82.98%**

Table 4-67. Dataset 2, MGD Results, 90 Messages, Second Top 6 Authors

Size=90	P(A2|x)	P(A32|x)	P(A44|x)	P(A72|x)	P(A80|x)	P(A89|x)
A2	**99.39%**	0.00%	0.00%	0.61%	0.00%	0.00%
A32	0.00%	**100.00%**	0.00%	0.00%	0.00%	0.00%
A44	0.00%	0.00%	**100.00%**	0.00%	0.00%	0.00%
A72	0.87%	0.00%	0.00%	**99.13%**	0.00%	0.00%
A80	0.00%	0.00%	0.00%	0.00%	**100.00%**	0.00%
A89	0.00%	0.00%	0.00%	0.00%	3.03%	**96.97%**

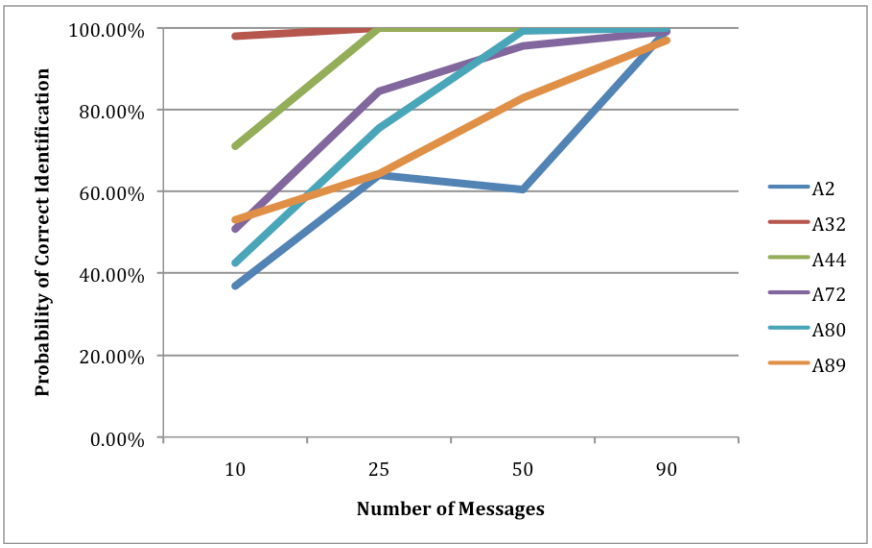
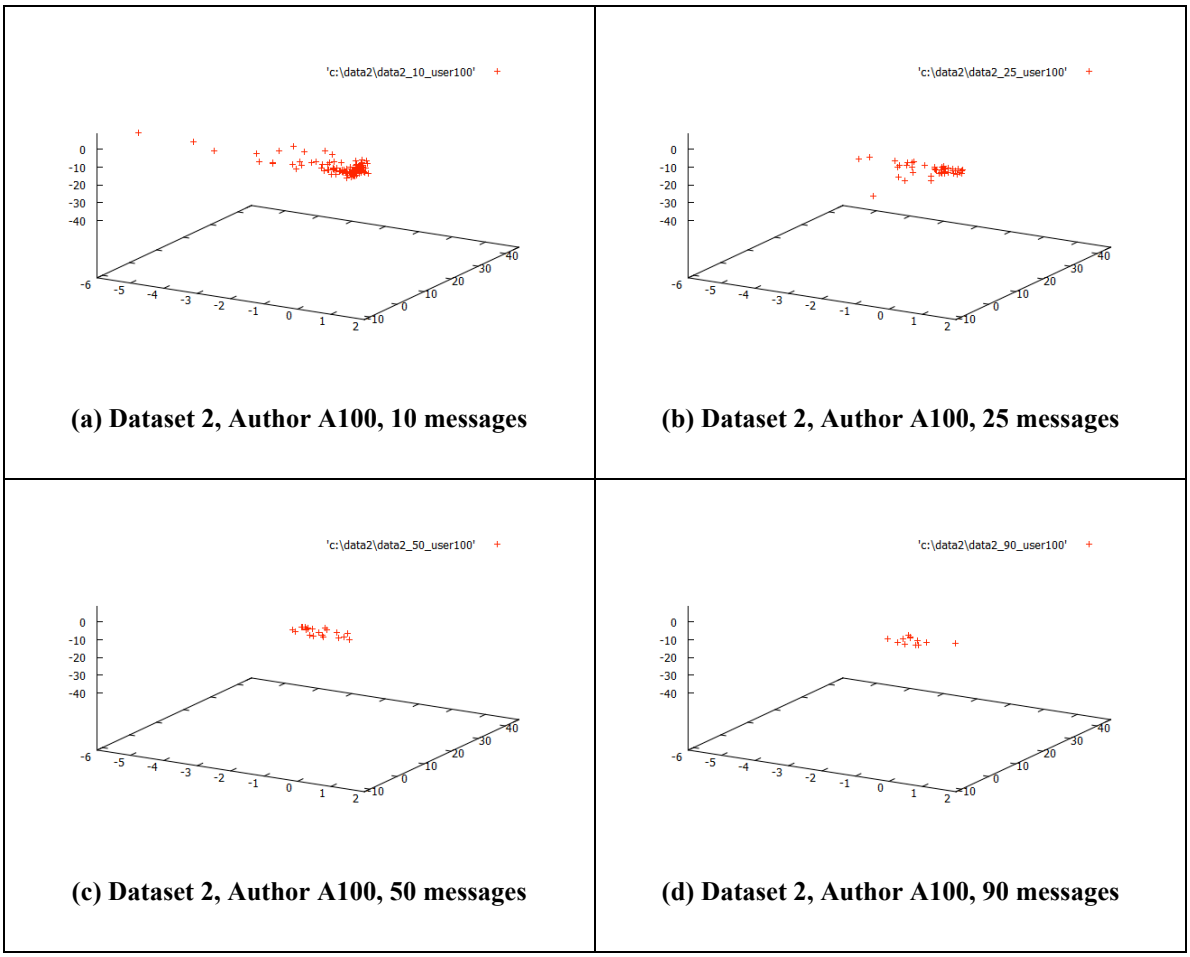


Figure 4-56. Dataset 2, Identification Probability vs. Number of Messages, Second Top 6 Authors

Figure 4-57 shows the PCA data plots for a single author (Author A100) over the full range of conversation sizes (10, 25, 50, and 90 messages respectively). The data shows as the number of messages per conversation increase, the data points become more

tightly grouped. This demonstrates that as the messages per conversation increase, the writeprint becomes more cohesive.



(a) Dataset 2, Author A100, 10 messages

(b) Dataset 2, Author A100, 25 messages

(c) Dataset 2, Author A100, 50 messages

(d) Dataset 2, Author A100, 90 messages

Figure 4-57. Dataset 2, PCA Plot Results, Author A100, All Conversation Sizes

Conversation size can be analyzed in more detail by calculating the standard deviation of the data within each conversation size. Figure 4-58 shows the inverse relationship of standard deviation and conversation size for the Author A100 results

shown in Figure 4-57. As the conversation size increases (i.e. number of messages per conversation), the standard deviation decreases. This shows that with larger conversations sizes an author's writeprint becomes more concise and is likely more representative of the author's true writing style.

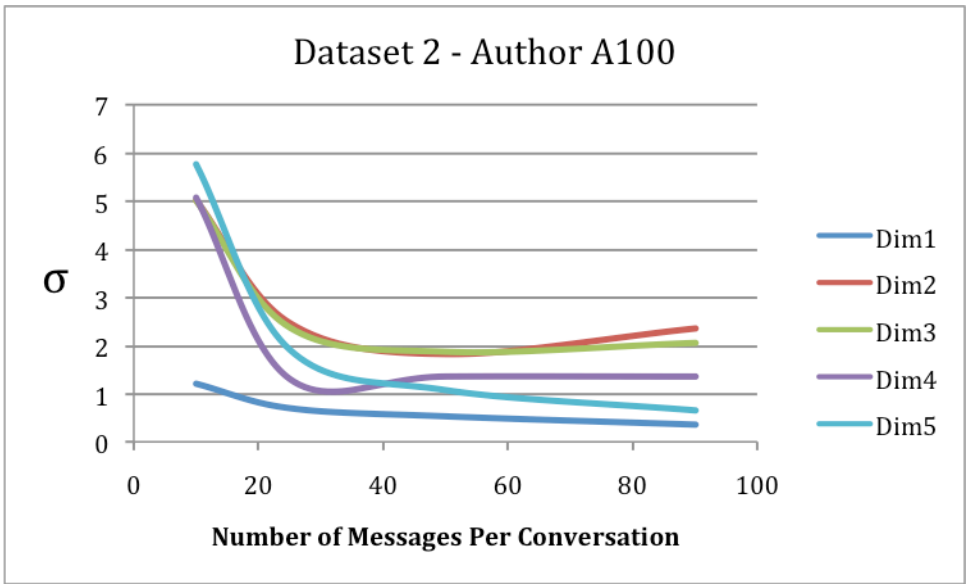


Figure 4-58. Dataset 2, Author A100, Conversation Size/Standard Deviation Relationship

The standard deviation of the data is calculated for the first 5 PCA dimensions for all 100 authors in Dataset 2. As shown in

Table 4-68, 86% of the 500 values exhibited decreased standard deviation as the conversation size increased.

Table 4-68. Dataset 2 Results for Conversation Size/Standard Deviation Relationship

Dataset	Number of Authors	Number of Dimensions per Author	Total Values Analyzed	Dimensions that Show Decrease in σ
2	100	5	500 (across sets of 10,25,50,90 messages per conversation)	86%

Figure 4-59 and Figure 4-60 show Dataset #2 PCA plot results for multiple samples of messages from Authors A3 and A7, respectively. The conversations consist of 50 messages for each writeprint instance. These results show that an individual author's writeprint is consistent over multiple samples. The overlapping PCA data points show writeprint similarity for an author over multiple distinct samples.

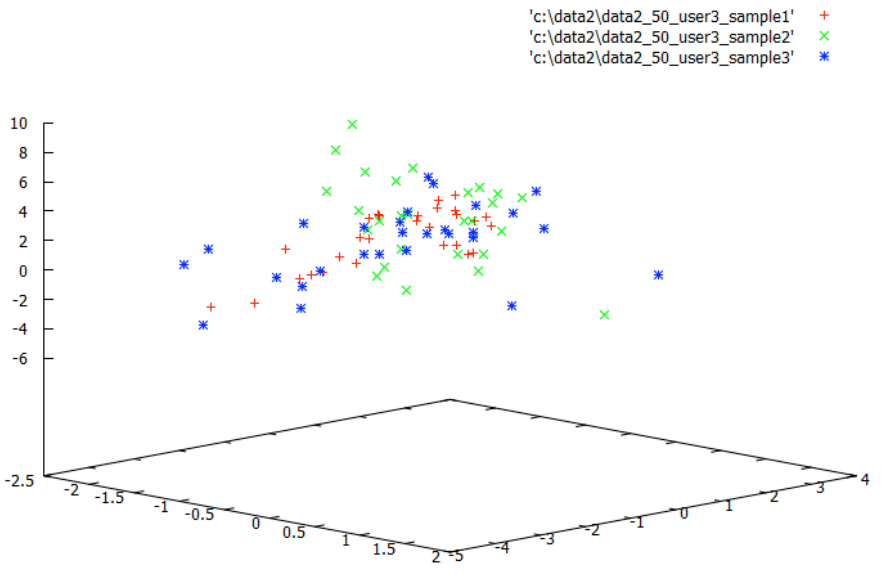


Figure 4-59. Dataset 2, PCA Plot Results, 50 Messages, Author A3 Samples

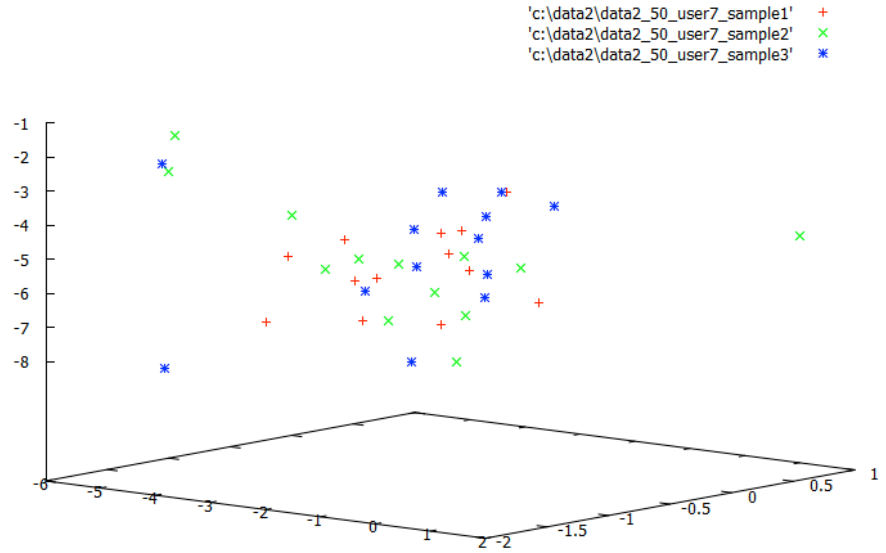


Figure 4-60. Dataset 2, PCA Plot Results, 50 Messages, Author A7 Samples

4.2.2 Authorship Characterization Results

Authorship characterization attempts to determine whether a given set of IM messages $\{M_1,\ldots,M_q\}$ is likely to be one a of the author categories $\{C_1,\ldots,C_m\}$.

Dataset #2 includes 4 categories for age (<20, 20s, 30s, >40) from which to analyze categorization. Figure 4-61 shows the breakdown of the number of authors for the age category.

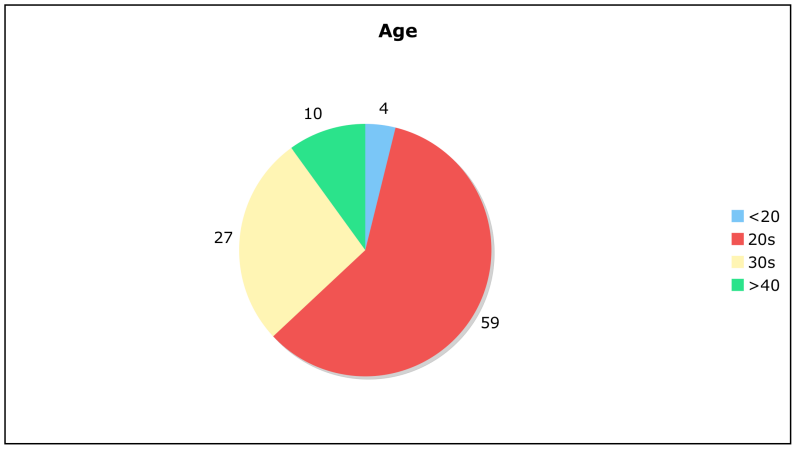


Figure 4-61. Dataset 2 Characterization Breakdown

Dataset #2 experiments include 100 authors from which to determine characterization. Figure 4-62 shows Dataset #2 PCA plot results for the all age categories. The conversations consist of 90 messages for each writeprint instance. The results show considerable overlap in writeprints across age categories.

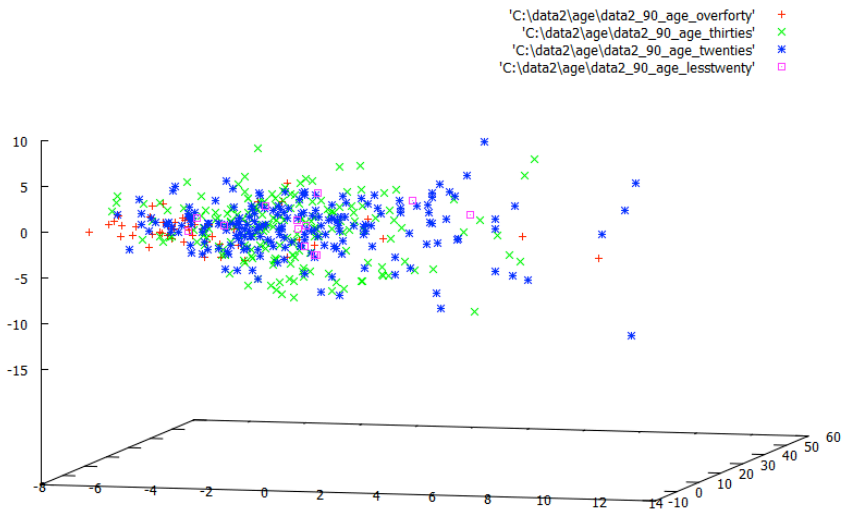


Figure 4-62. Dataset 2, PCA Plot Results, 90 Messages, All Age

Table 4-69 through

Table 4-72 shows the MGD results for conversation sizes of 10, 25, 50, and 90 messages respectively for the age category. Using 90 messages per conversation as input, MGD identifies conversations as the correct age for 3 of the 4 age categories, with probability ranging from 43.12% to 61.61%. The MGD results validate the considerable overlap in writeprints for the Dataset #2 age category. Given that the age data is unbalanced (the majority of authors in the twenties category), the results do not show a bias towards the twenties category. The tables show an increase in characterization probability as the number of messages per conversation increase. Figure 4-63 shows the

relationship between the characterization probability and number of messages per conversation.

Table 4-69. Dataset 2, MGD Results, 10 Messages, All Age

Size=10	P(LT|x)	P(T|x)	P(TH|x)	P(OF|x)
Less Twenty	18.23%	31.04%	**39.01%**	11.72%
Twenties	8.83%	34.47%	**36.44%**	20.26%
Thirties	9.24%	31.34%	**43.76%**	15.65%
Over Forty	3.62%	25.74%	33.64%	**37.00%**

Table 4-70. Dataset 2, MGD Results, 25 Messages, All Age

Size=25	P(LT|x)	P(T|x)	P(TH|x)	P(OF|x)
Less Twenty	19.92%	**37.71%**	34.26%	8.11%
Twenties	7.28%	**43.16%**	30.54%	19.02%
Thirties	7.63%	32.33%	**46.62%**	13.42%
Over Forty	2.61%	25.25%	31.05%	**41.08%**

Table 4-71. Dataset 2, MGD Results, 50 Messages, All Age

Size=50	P(LT|x)	P(T|x)	P(TH|x)	P(OF|x)
Less Twenty	24.26%	32.36%	**37.28%**	6.10%
Twenties	6.65%	**42.29%**	34.56%	16.49%
Thirties	8.43%	28.71%	**53.28%**	9.58%
Over Forty	3.84%	26.67%	30.22%	**39.26%**

Table 4-72. Dataset 2, MGD Results, 90 Messages, All Age

Size=90	P(LT|x)	P(T|x)	P(TH|x)	P(OF|x)
Less Twenty	20.00%	31.43%	**41.79%**	6.77%
Twenties	8.29%	**43.12%**	29.27%	19.32%
Thirties	2.58%	24.83%	**61.61%**	10.98%
Over Forty	0.74%	22.90%	21.29%	**55.07%**

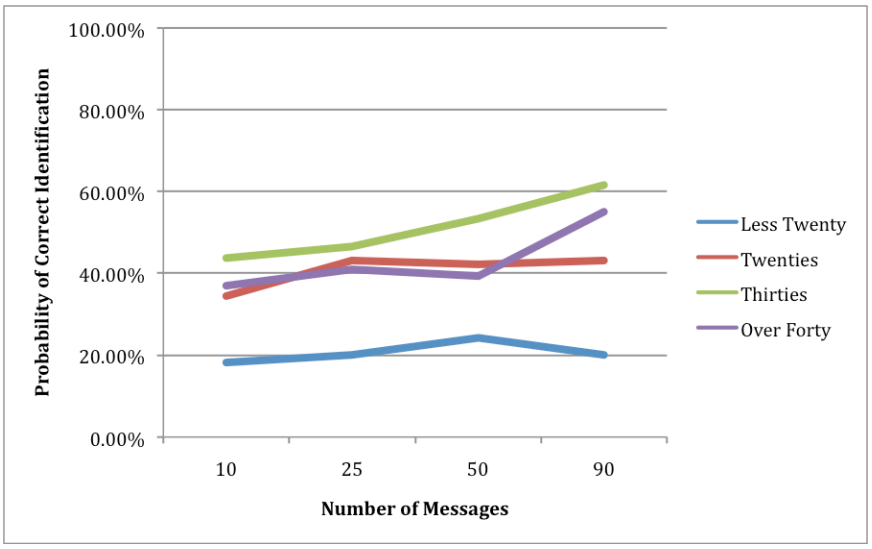


Figure 4-63. Dataset 2, Characterization Probability vs. Number of Messages, All Age

The authorship characterization probability is used to determine the error of the multivariate Gaussian distribution by assessing writeprint false positives. The likelihood, P(x|Category), of the author category of the writperint is used as a minimum threshold. If another author category has a higher likelihood, this is a false positive. Dataset #2 analysis for the age category shows high error rates due to more overlap between the age categories. The thirites category achieved a 20% error using 90 messages per conversation. The less than twenty and over forty categories show the most overlap, and thus the highest number of false positives. Figure 4-64 shows that as the conversation size increases, the error rate decreases.

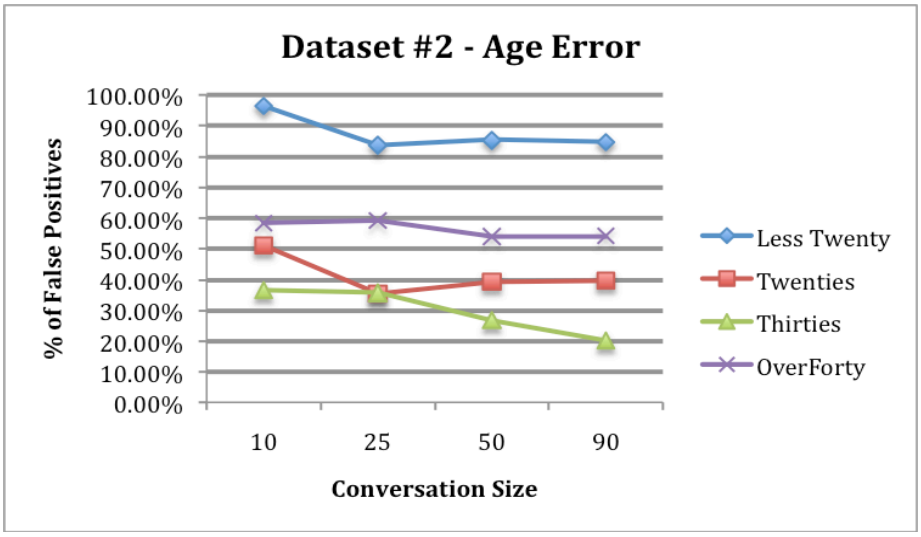


Figure 4-64. Dataset #2, Age Error

The next several results tables show more detail on the differences and similarities between Dataset #2 age categories. Figure 4-65 shows Dataset #2 results for the over forty and less than twenty age categories. The conversations consist of 90 messages for each writeprint instance. This plot shows some separation between age categories.

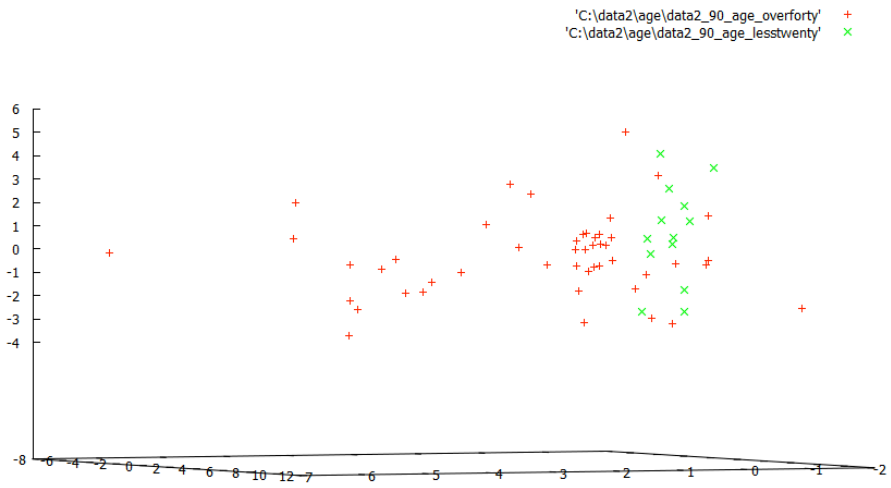


Figure 4-65. Dataset 2, PCA Plot Results, 90 messages, >40 and <20 Age

Table 4-73 shows the MGD results for conversation sizes of 10, 25, 50, and 90 messages respectively for the over forty and less than twenty age categories. Using 25 messages per conversation as input, MGD identifies conversations as the correct age for both age categories, with probability ranging from 71.07% to 94.02%. Using 90 messages per conversation as input, MGD identifies conversations as the correct age for both age categories, with probability ranging from 74.70% to 98.67%. The MGD results show distinction between the writeprints for the Dataset #2 over forty and less than twenty age categories. Given that the age data is unbalanced (more authors in the over forty category), the results may indicate a slight bias towards the over forty category. The tables show an increase in characterization probability as the number of messages

per conversation increase. Figure 4-66 shows the relationship between the characterization probability and number of messages per conversation.

Table 4-73. Dataset 2, MGD Results, 10-90 Messages, >40 and <20 Age

Size=10	P(LT|x)	P(OF|x)
Less Twenty	**60.85%**	39.15%
Over Forty	8.92%	**91.08%**

Size=25	P(LT|x)	P(OF|x)
Less Twenty	**71.07%**	28.93%
Over Forty	5.98%	**94.02%**

Size=50	P(LT|x)	P(OF|x)
Less Twenty	**79.90%**	20.10%
Over Forty	8.91%	**91.09%**

Size=90	P(LT|x)	P(OF|x)
Less Twenty	**74.70%**	25.30%
Over Forty	1.33%	**98.67%**

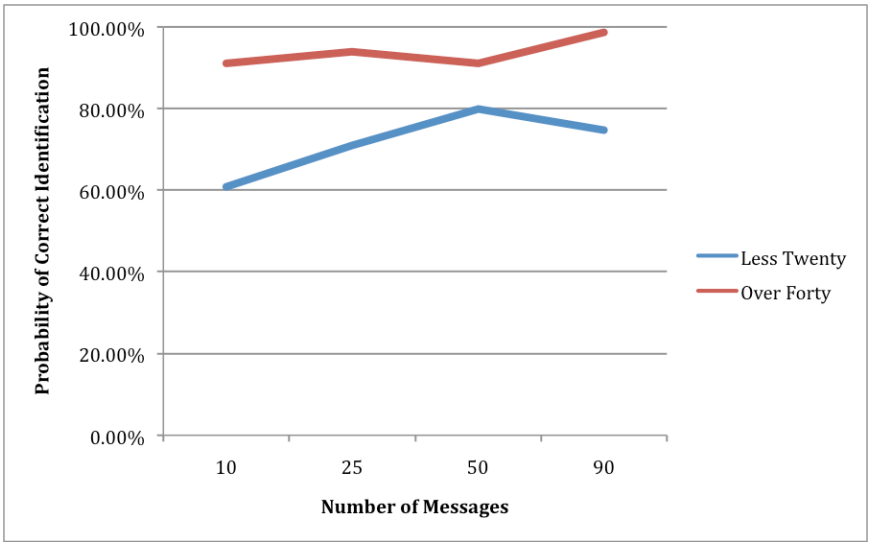


Figure 4-66. Dataset 2, Characterization Probability vs. Number of Messages, >40 and <20 Age

Figure 4-67 shows Dataset #2 results for the over forty and thirties age categories. The conversations consist of 90 messages for each writeprint instance. This plot shows some separation between the age categories.

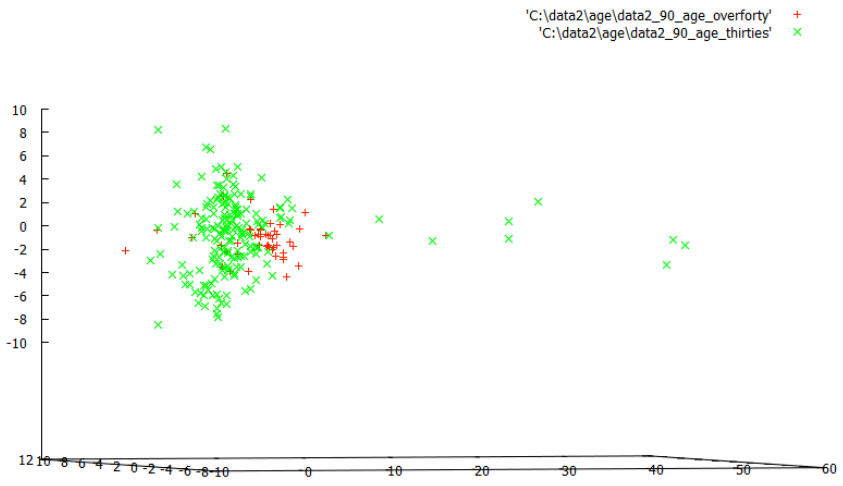


Figure 4-67. Dataset 2, PCA Plot Results, 90 Messages, >40 and 30s Age

Table 4-74 shows the MGD results for conversation sizes of 10, 25, 50, and 90 messages respectively for the over forty and thirties age categories. Using 90 messages per conversation as input, MGD identifies conversations as the correct age for both age categories, with probability ranging from 72.12% to 84.87%. The MGD results show distinction between the writeprints for the Dataset #2 over forty and thirties age categories. Given that the age data is unbalanced (more authors in the thirties category), the results may indicate a slight bias towards the thirties category. The tables show an increase in characterization probability as the number of messages per conversation increase. Figure 4-68 shows the relationship between the characterization probability and number of messages per conversation.

Table 4-74. Dataset 2, MGD Results, 10-90 Messages, >40 and 30s Age

Size=10	P(TH|x)	P(OF|x)
Thirties	**73.66%**	26.34%
Over Forty	47.62%	**52.38%**

Size=25	P(TH|x)	P(OF|x)
Thirties	**77.64%**	22.36%
Over Forty	43.05%	**56.95%**

Size=50	P(TH|x)	P(OF|x)
Thirties	**84.76%**	15.24%
Over Forty	43.50%	**56.50%**

Size=90	P(TH|x)	P(OF|x)
Thirties	**84.87%**	15.13%
Over Forty	27.88%	**72.12%**

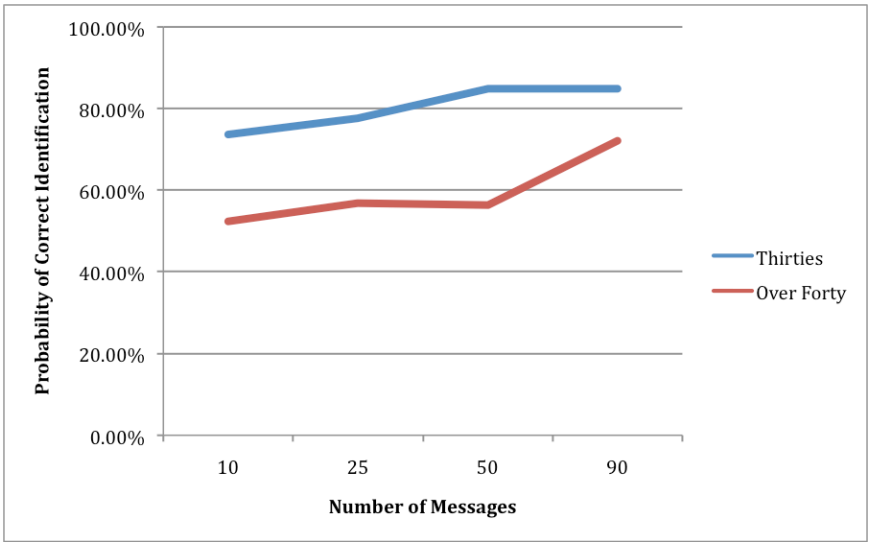


Figure 4-68. Dataset 2, Characterization Probability vs. Number of Messages, >40 and 30s Age

Figure 4-69 shows Dataset #2 results for the over forty and twenties age categories. The conversations consist of 90 messages for each writeprint instance. This plot shows more overlap between the age categories.

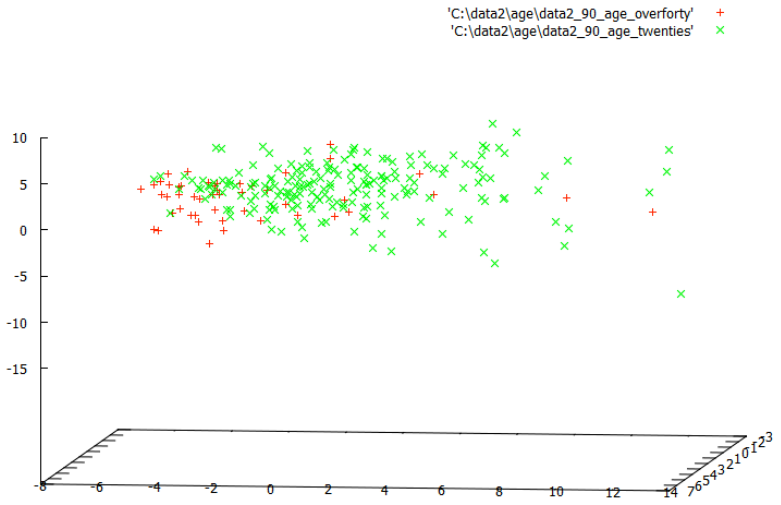


Figure 4-69. Dataset 2, PCA Plot Results, 90 Messages, >40 and 20s Age

Table 4-75 shows the MGD results for conversation sizes of 10, 25, 50, and 90 messages respectively for the over forty and twenties age categories. Using 90 messages per conversation as input, MGD identifies conversations as the correct age for both age categories, with probability ranging from 69.06% to 70.63%. The MGD results show some distinction between the writeprints for the Dataset #2 over forty and twenties age categories. Given that the age data is unbalanced (more authors in the twenties category), the results may show a slight bias towards the twenties category. The tables show an increase in characterization probability as the number of messages per conversation increase. Figure 4-70 shows the relationship between the characterization probability and number of messages per conversation.

Table 4-75. Dataset 2, MGD Results, 10-90 Messages, >40 and 20s Age

Size=10	P(T|x)	P(OF|x)
Twenties	**62.99%**	37.01%
Over Forty	41.02%	**58.98%**

Size=25	P(T|x)	P(OF|x)
Twenties	**69.42%**	30.58%
Over Forty	38.07%	**61.93%**

Size=50	P(T|x)	P(OF|x)
Twenties	**71.95%**	28.05%
Over Forty	40.45%	**59.55%**

Size=90	P(T|x)	P(OF|x)
Twenties	**69.06%**	30.94%
Over Forty	29.37%	**70.63%**

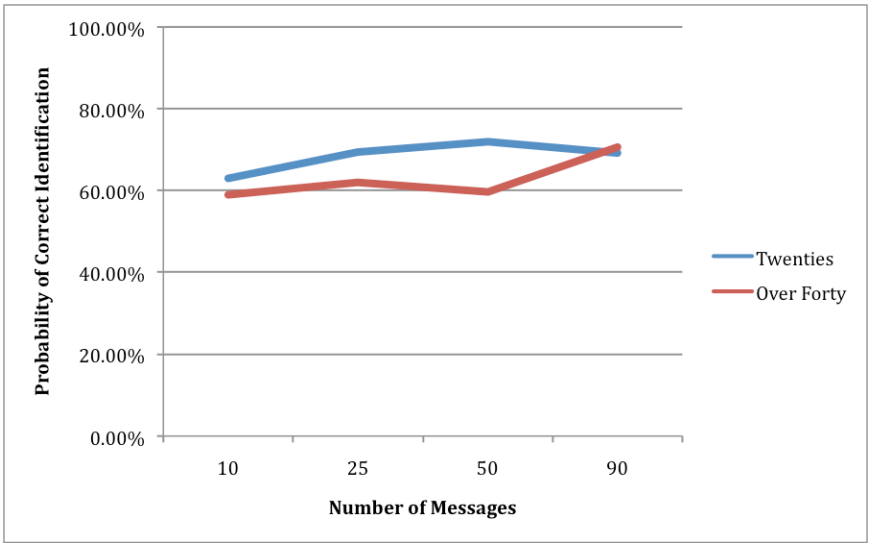


Figure 4-70. Dataset 2, Characterization Probability vs. Number of Messages, >40 and 20s Age

Figure 4-71 shows Dataset #2 results for the thirties and twenties age categories. The conversations consist of 90 messages for each writeprint instance. This plot shows considerable overlap between the age categories.

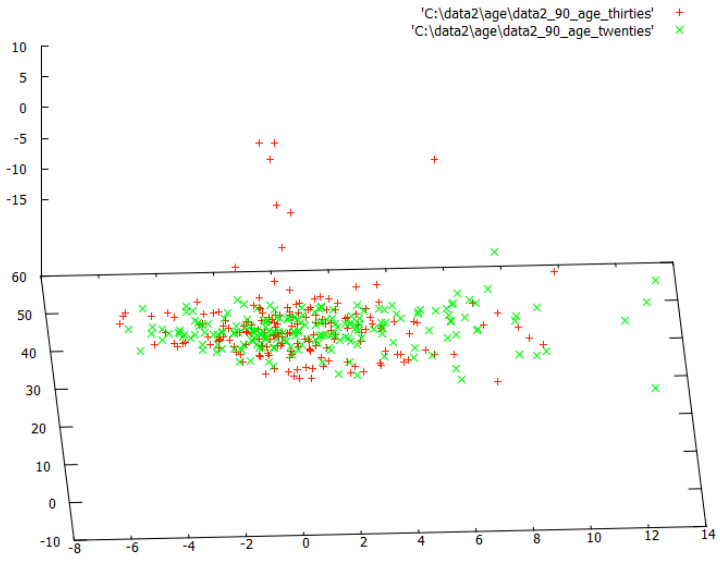


Figure 4-71. Dataset 2, PCA Plot Results, 90 Messages, 30s and 20s Age

Table 4-76 shows the MGD results for conversation sizes of 10, 25, 50, and 90 messages respectively for the thirties and twenties age categories. Using 90 messages per conversation as input, MGD identifies conversations as the correct age for both age categories, with probability ranging from 59.57% to 71.27%. The MGD results show overlap between the writeprints for the Dataset #2 thirties and twenties age categories. Given that the age data is unbalanced (more authors in the twenties category), the results do not show a bias towards the twenties category. The tables show an increase in probability as the number of messages per conversation increase. Figure 4-72 shows the relationship between the probability and number of messages per conversation.

Table 4-76. Dataset 2, MGD Results, 10-90 Messages, 30s and 20s Age

Size=10	P(T|x)	P(TH|x)
Twenties	48.61%	**51.39%**
Thirties	41.73%	**58.27%**

Size=25	P(T|x)	P(TH|x)
Twenties	**58.57%**	41.43%
Thirties	40.95%	**59.05%**

Size=50	P(T|x)	P(TH|x)
Twenties	**55.03%**	44.97%
Thirties	35.02%	**64.98%**

Size=90	P(T|x)	P(TH|x)
Twenties	**59.57%**	40.43%
Thirties	28.73%	**71.27%**

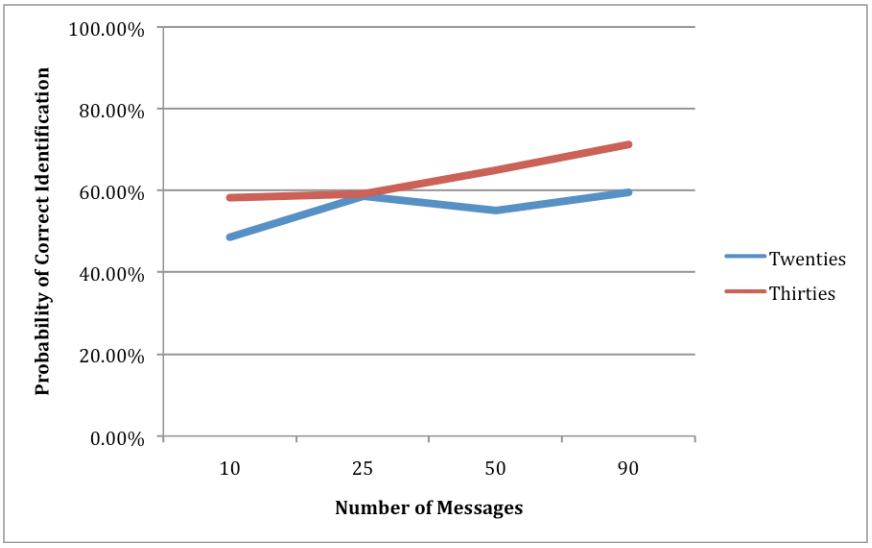


Figure 4-72. Dataset 2, Characterization Probability vs. Number of Messages, 30s and 20s Age

Figure 4-73 shows Dataset #2 results for the thirties and less than twenty age categories. The conversations consist of 90 messages for each writeprint instance. This plot shows considerable overlap between the age categories.

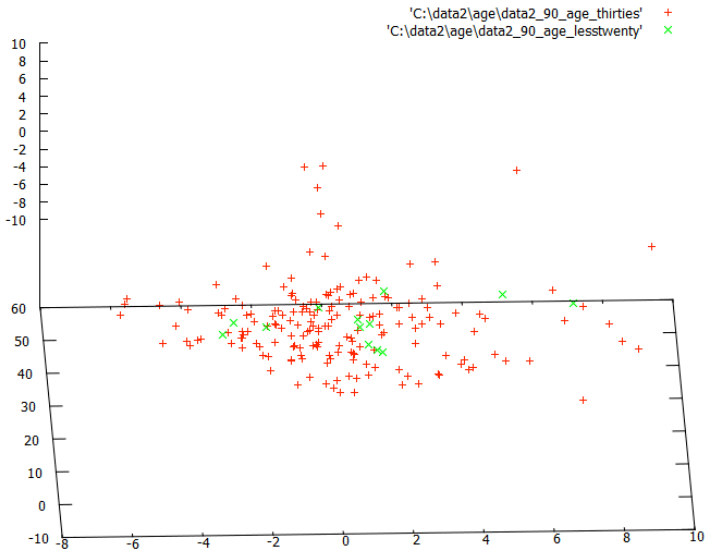


Figure 4-73. Dataset 2 Results, 90 messages, 30s and <20 Age

Table 4-77 shows the MGD results for conversation sizes of 10, 25, 50, and 90 messages respectively for the thirties and less than twenty age categories. Using 90 messages per conversation as input, MGD identifies conversations as the correct age for only the thirties age category, with 95.99%. The MGD results show considerable mischaracterization of the less then twenty write prints as the thirties age categories. Given that the age data is unbalanced (more authors in the thirties category), the results may show a bias towards the thirties category. The tables show an increase in characterization probability as the number of messages per conversation increase. Figure 4-74 shows the relationship between the characterization probability and number of messages per conversation.

Table 4-77. Dataset 2, MGD Results, 10-90 Messages, 30s and <20 Age

Size=10	P(LT|x)	P(TH|x)
Less Twenty	31.84%	**68.16%**
Thirties	17.43%	**82.57%**

Size=25	P(LT|x)	P(TH|x)
Less Twenty	36.76%	**63.24%**
Thirties	14.06%	**85.94%**

Size=50	P(LT|x)	P(TH|x)
Less Twenty	39.42%	**60.58%**
Thirties	13.66%	**86.34%**

Size=90	P(LT|x)	P(TH|x)
Less Twenty	32.37%	**67.63%**
Thirties	4.01%	**95.99%**

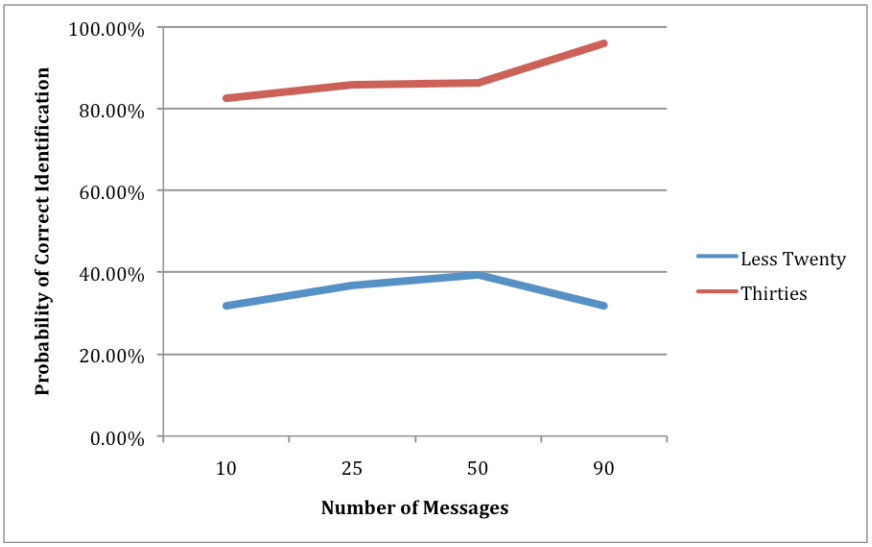


Figure 4-74. Dataset 2, Characterization Probability vs. Number of Messages, 30s and <20 Age

Figure 4-75 shows Dataset #2 results for the twenties and less than twenty age categories. The conversations consist of 90 messages for each writeprint instance. This plot shows considerable overlap between the age categories.

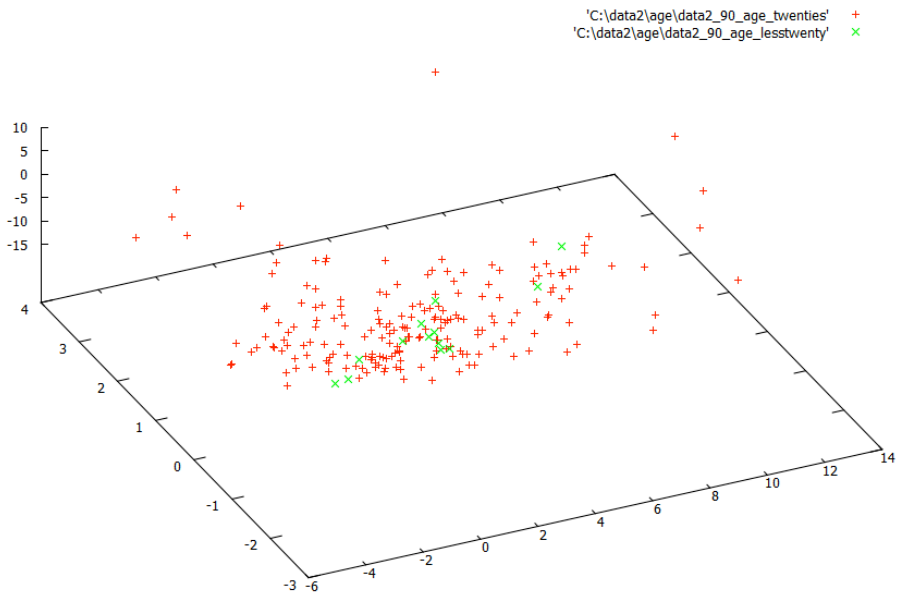


Figure 4-75. Dataset 2 Results, 90 messages, 20s and <20 Age

Table 4-78 shows the MGD results for conversation sizes of 10, 25, 50, and 90 messages respectively for the twenties and less than twenty age categories. Using 90 messages per conversation as input, MGD identifies conversations as the correct age for only the twenties age category, with 83.87%. The MGD results show considerable mischaracterization of the less then twenty write prints as the twenties age categories. Given that the age data is unbalanced (more authors in the twenties category), the results may show a bias towards the twenties category. The tables show an increase in characterization probability as the number of messages per conversation increase. Figure 4-76 shows the relationship between the characterization probability and number of messages per conversation.

Table 4-78. Dataset 2, MGD Results, 10-90 Messages, 20s and <20 Age

Size=10	P(LT|x)	P(T|x)
Less Twenty	37.00%	**63.00%**
Twenties	20.39%	**79.61%**

Size=25	P(LT|x)	P(T|x)
Less Twenty	34.56%	**65.44%**
Twenties	14.44%	**85.56%**

Size=50	P(LT|x)	P(T|x)
Less Twenty	42.85%	**57.15%**
Twenties	13.59%	**86.41%**

Size=90	P(LT|x)	P(T|x)
Less Twenty	38.90%	**61.10%**
Twenties	16.13%	**83.87%**

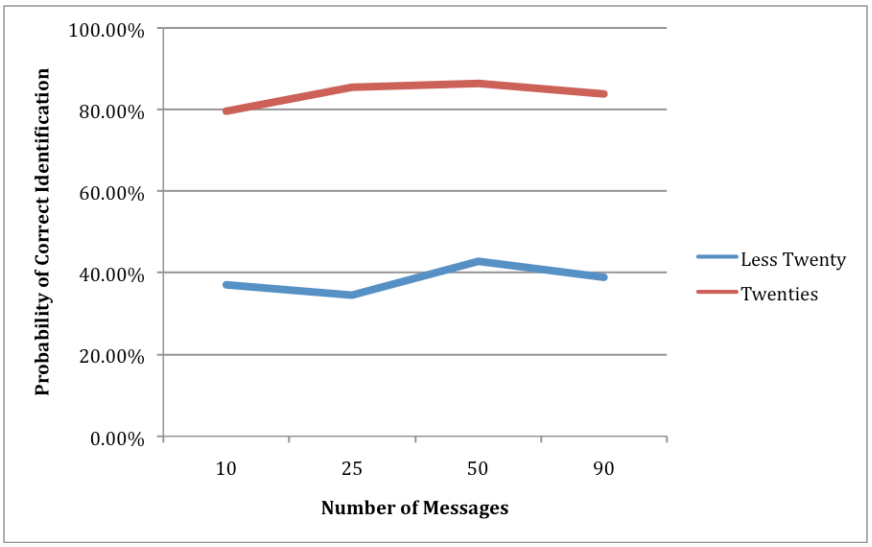


Figure 4-76. Dataset 2, Characterization Probability vs. Number of Messages, 20s and <20 Age

4.3 Summary

This chapter presented the authorship identification and characterization results for Dataset #1, Known Authors and Dataset #2, U.S. Cyberwatch. PCA was used to reduce the number of dimensions and provide visualization data. The coefficients of the first three principal components were plotted, allowing the PCA data to be viewed in 3-

dimensions. In most cases, the PCA plots showed separate grouping for each author and author category, with the age category showing the least separation. PCA data plots for a single author over the full range of conversation sizes (5, 10, 25, 50, 100, 125, 250, and 500 messages respectively) showed tightly grouped data points as the number of messages per conversation increased. These plots demonstrated that as the messages per conversation increase, the writeprint becomes more cohesive. Standard deviation was used to analyze the spread of the distribution of data within each conversation size. Standard deviation results for the first 5 PCA dimensions for authors were presented graphically to show the relationship of standard deviation and conversation size. The graphs show decreased standard deviation as the conversation size increased (i.e. number of messages per conversation). This demonstrates that with larger conversations sizes an author's writeprint becomes more concise and is likely more representative of the author's true writing style. PCA plot results for multiple samples of messages from the same author showed that an individual author's writeprint is consistent over multiple distinct samples.

MGD was used to determine identification and characterization probability of a set of messages across authors and author categories. The MGD algorithm processed each set of messages (conversations) for an author or author category under test and the output was analyzed to determine the identification and characterization probability for each conversation across all authors or author categories. Results were presented in table matrices for multiple author group sizes and varying conversation size. The results showed a significant increase in probability as the number of messages per conversation

increase.

The results in this chapter demonstrated the effectiveness of creating IM author writeprints that show separation between authors and between author categories. The results show that writeprints can differentiate messages belonging to a particular author A_i from a set of authors $\{A_1,\ldots,A_n\}$ and can differentiate messages belonging to a particular author category C_i, from a set of author categories $\{C_1,\ldots,C_m\}$.

5. SUMMARY AND CONCLUSIONS

This research created and analyzed behavioral biometrics-based instant messaging writeprints to use as input for cybercrime investigations. This research used authorship analysis techniques to create a set of stylometric features robust enough to show separation between authors and between author categories. This research used the statistical methods Principal Component Analysis (PCA), multivariate Gaussian distributions (MGD), and Standard Deviation to analyze IM conversation logs from two distinct data sets for authorship identification and characterization.

The research methodology used authorship analysis and statistical techniques to create and analyze IM writeprints to assist in identifying an author, as well as certain characteristics of the author of a set of IM messages. This research performed authorship identification to assist in identifying individual IM authors. In cybercrime investigations, authorship identification may assist in identifying criminals who hide their true identify or impersonate a known individual. This research performed authorship characterization to assist in identifying sociolinguistic categories of authors. In cybercrime investigations, authorship characterization may assist in discovering IM cyber criminals who supply false information in their virtual identifies, such as gender.

This dissertation provides a foundation for using behavioral biometrics as a cyber forensics element for cybercrime investigations by demonstrating the effectiveness of

creating instant messaging author writeprints to be used in conjunction with traditional investigation techniques. This research contributes a new technique to assist cybercrime decision support tools in collecting and analyzing digital evidence, discovering characteristics about the cyber criminal, and assisting in identifying cyber criminal suspects. Writeprints may be used in conjunction with other evidence and investigative techniques as an element in multimodal biometrics to aid the investigation process. Criminal investigators may use the IM-specific stylometric taxonomy and statistical methods provided in this dissertation to create and analyze writeprints as part of a cybercrime investigation.

The writeprint analysis results in this dissertation achieved the following goals:

1. Creating writeprints that show separation between authors and author categories,
2. Creating writeprints that can differentiate messages belonging to a particular author A_i from a set of authors $\{A_1,\ldots,A_n\}$, and
3. Creating writeprints that can differentiate messages belonging to a particular author category C_i, from a set of author categories $\{C_1,\ldots,C_m\}$ based on sociolinguistic attributes.

For authorship identification, the PCA plots for both datasets clearly show separation of author writeprints at large conversation sizes. Dataset #1 shows separation of author writeprints using 250 and 500 messages per conversation. Dataset #2 shows separation of author writeprints using 90 messages per conversation. The standard

deviation analysis for conversation sizes in both datasets shows that as the number of messages in the conversation increase, the standard deviation decreases, indicating the writeprint becomes more cohesive. For Dataset #1, 96% of the PCA dimensions showed a decrease in standard deviation as the conversation size increased. For Dataset #2, 86% of PCA dimensions showed a decrease in standard deviation as the conversation size increased. The percentage of authors in Dataset #2 showing a decrease in standard deviation as the conversation size increases is less because the total amount of data per author is limited and the maximum conversation size is 90 messages per conversation. If Dataset #2 had more messages for each author leading to larger conversation sizes, the percentages may be higher. However, given the limited data, 86% still demonstrates that as the conversation size increases, the standard deviation decreases for most authors. The standard deviation results demonstrate that with larger conversation sizes an author's writeprint is more likely to reflect the author's true writing style.

Table 5-1 shows the authorship identification results for Dataset #1 tests for conversations sizes of 50-500 messages. Authorship analysis identification results greater than 70% are acceptable during an investigation process [IBFD2013]. Results with a maximum probability less than 70% are designated with a "T" for trivial. A conversation size of 100 messages demonstrated identification probability of 71.51-100% across all tests. For tests with fewer authors, smaller conversation sizes (i.e. 50 messages per conversation) resulted in identification probability over 90%. For authorship

identification, the MGD results show a significant increase in identification probability as the number of messages per conversation increases, indicating a more distinct writeprint at larger conversation sizes.

Table 5-1. Dataset #1 Authorship Identification Results

Dataset #1 Test/Message Size	50	100	125	250	500
All 19 Authors	T	71.51-100%	82.14-100%	N/A	N/A
Authors 1-6	T	91.93-100%	99.72-100%	N/A	N/A
Authors 7-12	78.61-100%	99.77-100%	99.81-100%	N/A	N/A
Authors 13-19	T	82.39-100%	86.39-100%	N/A	N/A
Authors 1-3	83.15-99.92%	92.93-100%	99.28-100%	N/A	N/A
Authors 4-6	76.96-99.97%	99.10-100%	99.72-100%	N/A	N/A
Authors 7-9	97.38-100%	100%	100%	N/A	N/A
Authors 10-12	78.66-100%	99.77-100%	99.81-100%	N/A	N/A
Authors 13-15	70.79-87.31%	99.79-10%	100%	N/A	N/A
Authors 16-19	T	82.39-99.99%	86.39-100%	N/A	N/A
Top 7 Authors	T	81-100%	82.42-100%	93.13-100%	99.85-100%
5 Related Authors	T	84.59-100%	87.85-100%	99.94-100%	N/A
3 Sibling Authors	81.40-86.22%	91.93-100%	91.93-99.75%	100%	N/A
Authors 1 and 12 (mother/daughter)	78.15-79.24%	91.37-93.04%	88.41-92.81%	99.95-100%	100%
Authors 2 and 12 (mother/daughter)	88.90-89.91%	97.87-98.36%	97.78-98.62%	99.95-100%	100%
Authors 2 and 14 (spouses)	T	81.10-85.01%	83.07-83.43%	93.13-98.30%	99.85-99.96%

Dataset #1 included some author metadata to assist with additional analysis. The metadata included familial and marital information that was used to analyze specific authors for similar traits based on their relationships. The PCA plots and MGD results do not show any significant result differences for related authors. Related author writeprints are just as distinct as non-related authors. Authors 2 and 12 spend a considerable amount

of time instant messaging each other, but still show separation of writeprints. Although this is a single test, this may indicate that authors do not appear to pick up enough traits from their IM buddy's writing style to influence their own writeprint.

Table 5-2 shows the authorship characterization results for Dataset #1 tests for conversations sizes of 50-500 messages. Results with a maximum probability less than 70% are designated with a "T" for trivial. Both the Gender and Education categories demonstrated high probability using a minimum conversation size of 100 messages. The unbalanced gender data (more females than males) may present a slight gender bias for the female category. The Age category did not perform as well, requiring a large conversation size of 500 messages to reach characterization probability over 75%. The age category "Thirties" had the most overlap with the other age categories. However, this is expected as authors in the Thirties category may be at either end of the category (30 or 39) and could easily overlap with adjacent categories. For authorship characterization, the MGD results show a significant increase in characterization probability as the number of messages per conversation increases, indicating a more distinct writeprint at larger conversation sizes.

Table 5-2. Dataset #1 Authorship Characterization Results

Dataset #1 Test/Message Size	50	100	125	250	500
Gender	T	74.76-88.23%	78.55-90.72%	92.66-96.82%	99.19-99.96%

Education	81.63-87.19%	89.48-95.25%	89.65-96.07%	92.98-97.92%	97.19-97.98%
Age	T	T	T	T	75.48-99.85%

Table 5-3 shows the authorship identification results for Dataset #2 tests for conversations sizes of 50-90 messages. Results with a maximum probability less than 70% are designated with a "T" for trivial. A conversation size of 90 messages demonstrated identification probability at 71.65-100% across all tests. For authorship identification, the MGD results show a significant increase in identification probability as the number of messages per conversation increases, indicating a more distinct writeprint at larger conversation sizes. Dataset #2 results present similar identification probability ranges as the Dataset #1 results, indicating consistency across the tests for both datasets.

Table 5-3. Dataset #2 Authorship Identification Results

Dataset #2 Test/Message Size	**50**	**90**
Top 20 Authors (*correct identification for 19/20 authors)	T	71.65-100%
Top 6 Authors	83.84-100%	91.15-100%
Top 6 Authors – Subset 1	97.62-100%	100%
Top 6 Authors – Subset 2	92.66-97.08%	91.15-100%
Second Top 6 Authors	82.98-100%	96.97-100%

Table 5-4 shows the authorship characterization results for Dataset #2 tests for conversations sizes of 50-90 messages. Results with a maximum probability less than 70% are designated with a "T" for trivial. Dataset #2 did not provide significant results

for the age category. There was considerable overlap of category writeprints, especially the less than twenty category that was consistently misidentified as the twenties and thirties age categories. Given that the age data is unbalanced (more authors in the twenties category), the results may show a bias towards the twenties category. These results for Dataset #2 are consisted with results from Dataset #1, which also had the poorest results for the age category. Gender and education were not tested for Dataset #2 since all authors were male and education level was unknown. Although gender and education offer promising results for authorship characterization, more research is needed using age as sociolinguistic category to determine if it can be used to assist in creating a cybercriminal profile.

Table 5-4. Dataset #2 Authorship Characterization Results

Dataset #2 Test/Message Size	50	90
All Age	T	T
>40 and <20	79.90-91.09%	74.70-98.67%
>40 and 30s	T	72.12-84.87%
>40 and 20s	T	T
30s and 20s	T	T
30s and <20	T	T
20s and <20	T	T

The authorship identification and characterization probability is used to determine the error of the multivariate Gaussian distribution by assessing writeprint false positives. The likelihood, P(x|Author) or P(x|Cateogry), of the author or author category of the writperint is used as a minimum threshold to assess false positives. This research resulted in lower error rates than related works such as [AC2008], [CRSBVM2012], and [IBFD2013]. [AC2008] achieved error rates of 68.3% - 49.6% using the Cyberwatch

dataset. [CRSBVM2012] and [IBFD2013] achieved 10.5% and 30.25% respectively using Skype chat and blog datasets.

Table 5-5 shows a summary of error rates for both datasets. The error results demonstrate that conversations with a larger number messages or more offer the best results.

Table 5-5. Error Results

Dataset	Number of Messages	Error
Dataset1, All 19 Authors	125	0% – 20%
Dataset1, Top 7 Authors	125	0% – 22.22%
Dataset1, Top 7 Authors	500	0% – 4.35%
Dataset1, Gender	500	0% – 5.47%
Dataset1, Education	500	5.63% – 6.85%
Dataset1, Age	500	0% – 36.59%
Dataset2, Top 20 Authors	90	0% – 25%
Dataset2, Top 6 Authors	90	0% – 16.67%
Dataset2, Age	90	20% – 84.62%

This dissertation research lends itself to a few extensions that provide potential future areas of research. This research may be applied to other datasets to demonstrate the scalability and feasibility of IM writeprints in other environments and using different sociolinguistic categories. Additionally, analyzing the effectiveness and applicability of this research using other languages would be an interesting study area. Lastly, an important research extension is applying the authorship analysis techniques to perform masquerade detection, including assessing the impact of intentional alteration of online writing habits.

This dissertation addresses the existing research gap in applying authorship analysis techniques to instant messaging communications to facilitate authorship identification and characterization. It provides a new approach and techniques to assist in identifying cyber criminal suspects and collecting digital evidence as part of the investigation. The research provides cybercrime investigators a unique tool (IM writeprints) for analyzing IM-assisted cybercrimes. It also provides an IM-specific stylometric feature set taxonomy robust enough to show separation between authors and between author categories. This research demonstrated the effectiveness of creating IM author writeprints by evaluating various parameters in a systematic way. Parameters such as the size of the suspect space, size of the IM conversation, and selected features are critical to the development of an author writeprint. There are currently no known studies examining the impact of these parameters on IM authorship identification and characterization.

Cybercrime investigators may leverage the techniques presented in this dissertation in conjunction with traditional forensics investigative techniques to aid in cybercrime decision support. IM writeprints may be used in conjunction with other evidence, investigation techniques, and biometrics techniques to build or validate a criminal profile, reduce the potential suspect space to a certain subset of suspects; identify the most plausible author of an IM conversation from a group of suspects; link related crimes; develop an interview and interrogation strategy; and gather convincing digital evidence to justify search and seizure and provide probable cause. By demonstrating high authorship identification and characterization probability, the research results presented in this

dissertation indicate a promising future for applying authorship analysis as an element of a multimodal biometrics system to assist with cyber forensics and cybercrime investigations.

APPENDIX A – DETAILED FEATURE SET

Appendix A - 1. Function Words List

Function Words					
about	both	inside	of	something	we
above	but	into	off	such	what
after	by	is	on	than	whatever
all	can	it	once	that	when
although	could	its	onto	the	where
am	do	latter	opposite	their	whether
among	down	less	or	them	which
an	each	like	our	these	while
and	either	little	outside	they	who
another	enough	lots	over	this	whoever
any	every	many	own	those	whom
anybody	everybody	me	past	though	whose
anyone	everyone	more	per	through	will
anything	everything	most	plenty	till	with
are	few	much	plus	to	within
around	following	must	regarding	towards	without
as	for	my	same	under	worth
at	from	near	several	unless	would
be	have	need	she	unlike	yes
because	he	neither	should	until	you
before	her	no	since	up	your
behind	him	nobody	so	upon	
below	if	none	some	us	
beside	in	nor	somebody	used	
between	including	nothing	someone	via	

Appendix A - 2. Abbreviations List

Abbreviations					
143	CYA	ILY	OMG	THX	WYWH
…	DBEYR	IMHO	OTP	TLC	XOXO
2moro	DILLIGAS	IRL	PITA	TMI	YT
2nite	ETC	ISO	PLS	TTYL	YW
ASAP	FUBAR	JK	PLZ	TTYS	
B4N	FWIW	L8R	POV	TYVM	
BCNU	FYI	LMAO	ROTFL	U2	
BFF	GR8	LMFAO	RU	VBG	
BRB	IC	LOL	SOL	WEG	
BTW	IDC	NP	STBY	WTF	
CU	IDK	OIC	SWAK	WTG	

Appendix A - 3. Emoticons List

Emoticons					
:-)	:)	:-(	:(	;-)	;)
:-P	:P	;-P	;P	:-D	:D
:’-(	:’(	:*	:-*	0:-)	0;-)
:-!	:*(	>:)	>:-)	:-*	:-/
:-\	:-[	:-]	:-{	:-}	:-S
:-x	:-#	:-|	=)	>:-(	>:(
<3	</3	0:)	:*	:/	:\

APPENDIX B – DEMOGRAPHICS FOR DATASET #1: KNOWN AUTHORS

Author	Number of Messages	Gender	Age Group	Education level
1	2952	M	20s	HS
2	12929	F	30s	College
3	535	M	30s	College
4	4797	F	30s	College
5	2853	M	20s	College
6	1624	M	>40	College
7	685	F	30s	HS
8	585	M	20s	HS
9	999	F	20s	College
10	592	M	20s	College
11	3396	F	>40	HS
12	27972	F	>40	HS
13	502	F	30s	HS
14	11893	F	>40	College
15	753	F	>40	HS
16	1283	F	30s	HS
17	1217	F	30s	College
18	616	F	20s	HS
19	985	F	20s	College

APPENDIX C – DEMOGRAPHICS FOR DATASET #2: U.S. CYBERWATCH

Author	Author Name	Number of Messages	Age Group	Age
1	Aaron	178	20s	29
2	Aaron Vajda	990	30s	31
3	Ajith Abraham	4192	30s	35
4	Al	255	20s	21
5	Alexander	457	20s	29
6	Andrew	207	30s	32
7	Anthony	1954	30s	39
8	Anthony Periathamby	285	>40	61
9	Avni	259	20s	24
10	Benny	506	20s	24
11	Bo	649	20s	23
12	Bob	287	>40	44
13	Bob_2	95	>40	45
14	Brad Hendrickson	298	20s	20
15	Brandon Candiano	119	20s	24
16	Brian	720	20s	23
20	Brian Fletcher	720	20s	29
17	Brian_2	179	30s	30
18	Brian_3	233	20s	21
19	Brian_4	388	30s	33
21	Chad	95	20s	27
22	Chris	187	20s	25
23	Chris_2	389	>40	44
24	Chris_3	552	20s	27
25	Christopher Evans	247	30s	30
26	Corey	212	20s	27
27	Craig	182	20s	25
28	Dan	176	>40	50
29	Davey	216	<20	19
30	Dennis Webb	1291	30s	30
31	Don	590	>40	40
32	Don Zawada	957	>40	44

Author	Author Name	Number of Messages	Age Group	Age
33	Dylan Mattern	126	30s	37
34	Eric	128	20s	28
37	Eric Binns	140	20s	24
35	Eric_2	324	20s	27
36	Eric_3	204	20s	26
38	Ethan	599	<20	19
39	Frank	91	20s	25
40	Gabriel	552	20s	24
41	Ganelon Diers	1596	>40	43
42	Gary David Chick	244	20s	21
43	George	341	30s	34
44	Gilbert	804	20s	27
45	Grayling	393	20s	22
46	Jack	244	>40	55
47	Jacob	150	30s	31
48	Janakiraman Manivel	265	20s	26
49	Jason	102	20s	26
50	Jason Mathew Skeel	145	20s	25
Unused	Jeff	79	20s	27
51	Jeremy	425	<20	18
52	Jerry_Mukund	298	20s	29
53	Jesse	121	20s	21
54	Jim	135	30s	38
55	Joe	318	20s	28
56	Joe_2	170	30s	35
57	Joe_3	358	20s	23
58	Joey	407	30s	30
59	John	229	20s	22
60	John_2	165	>40	49
61	John_3	583	20s	24
62	Jon	197	20s	24
63	Jonathan	328	20s	27
64	Jonathan_2	228	20s	22
Unused	Josh	82	20s	21
65	Justin	114	20s	25
66	Ken	570	30s	34
67	Kurt	98	20s	26
68	Kyle	249	30s	34
69	Manojkumar Natarajan	1247	20s	25
Unused	Mark	85	>40	41

Author	Author Name	Number of Messages	Age Group	Age
70	Matt	181	20s	22
Unused	Matt_2	89	20s	27
71	Matt_3	165	<20	18
72	Matthew Davis	998	20s	26
73	Michael	501	30s	30
74	Michael S.	743	30s	39
75	Miles	477	20s	29
76	Nick	123	20s	25
77	Norby	771	30s	37
78	Patrick Spires	272	20s	22
79	Paul	729	30s	35
80	Phil	868	20s	29
81	Raymond	250	20s	25
82	Raymond Dooley	128	20s	23
83	Rehan	277	20s	22
84	Robert	248	20s	21
85	Robert Clayton	705	30s	31
Unused	Rod	40	>40	42
86	Russ	286	30s	32
87	Sam	198	20s	26
88	Shawn Boggs	298	20s	25
89	Skipper	918	20s	25
90	Stan_Joe	124	20s	24
91	Steven	450	30s	35
92	Steven Gossett	290	30s	30
93	Tim	221	30s	34
94	Tim_2	694	20s	21
95	Tito_Paul	340	20s	25
96	Toby	178	20s	26
97	Tom	370	20s	26
98	Tom_2	145	20s	28
99	Vikas	436	30s	31
100	Walter Chester Strout Jr.	1148	30s	35

LIST OF REFERENCES

LIST OF REFERENCES

[AC2005] Abbasi, Ahmed, and Hsinchun Chen. "Applying authorship analysis to extremist-group web forum messages." *Intelligent Systems, IEEE* 20.5 (2005): 67-75.

[AC2006] Abbasi, Ahmed, and Hsinchun Chen. "Visualizing authorship for identification." *Intelligence and Security Informatics* (2006): 60-71.

[AC2008] Abbasi, Ahmed, and Hsinchun Chen. "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace." *ACM Transactions on Information Systems* 26.2 (2008): 7.

[BP2003] Banerjee, Satanjeev, and Ted Pedersen. "The design, implementation, and use of the Ngram Statistics Package." *Computational Linguistics and Intelligent Text Processing* (2003): 370-381.

[BBO2006] Bassett, Richard, Linda Bass, and Paul O'Brien. "Computer forensics: An essential ingredient for cyber security." *Journal of Information Science and Technology* 3.1 (2006): 22-32.

[Bio2006] BioPassword. "Authentication Solutions Through Keystroke Dynamics." http://www.infosecurityproductsguide.com/technology/2007/BioPassword.html. (2006). (accessed April 2, 2013)

[BS1998] Bosch, Robert A., and Jason A. Smith. "Separating hyperplanes and the authorship of the disputed federalist papers." *The American mathematical monthly* 105.7 (1998): 601-608.

[BVT1996] Baayen, Harald, Hans Van Halteren, and Fiona Tweedie. "Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution."*Literary and Linguistic Computing* 11.3 (1996): 121-132.

[Cas1999] Casey, E. "Cyberpatterns: criminal behavior on the Internet." *Criminal profiling: An introduction to behavioral evidence analysis* (1999): 361-378.

[Cha2005] Chaski, Carole E. "Who's at the keyboard? Authorship attribution in digital evidence investigations." *International Journal of Digital Evidence* 4.1 (2005): 1-13.

[Cha2013] Chaski, Carole E., "Best Practices and Admissibility of Forensic Author Identification." *JL & Pol'y* 21 (2013): 333-725.

[Clo2000] Clough, Paul. "Plagiarism in natural and programming languages: an overview of current tools and technologies." (Research Memoranda: CS-00–05). Department of Computer Science, University of Sheffield, United Kingdom. (2000): 1.

[CRSBVM2012] Cristani, M., Roffo, G., Segalin, C., Bazzani, L., Vinciarelli, A., & Murino, V. Conversationally-inspired stylometric features for authorship attribution in instant messaging. In *Proceedings of the 20th ACM international conference on Multimedia* (2012): 1121-1124.

[Cro2008] Cross, Michael. *Scene of the Cybercrime*. Syngress Publishing, (2008): 679-690.

[DACM2001a] De Vel, Olivier, Alison Anderson, Malcolm Corney, and George Mohay. "Multi-topic e-mail authorship attribution forensics."*Proceedings of ACM Conference on Computer Security-Workshop on Data Mining for Security Applications*. (2001): 1-8.

[DACM2001b] De Vel, Olivier, Alison Anderson, Malcolm Corney, and George Mohay. "Mining e-mail content for author identification forensics." *ACM Sigmod Record* 30.4 (2001): 55-64.

[DACM2002a] De Vel, Olivier, Alison Anderson, Malcolm Corney, and George Mohay. "Gender-preferential text mining of e-mail discourse."*Computer Security Applications Conference, 2002. Proceedings. 18th Annual*. IEEE, (2002): 282-289.

[DACM2002b] De Vel, Olivier, Alison Anderson, Malcolm Corney, and George Mohay. "Language and gender author cohort analysis of e-mail for computer forensics." *Digital Forensic Research Workshop*. (2002): 1-16.

[DeM1882] De Morgan, Sophia Elizabeth. *Memoir of Augustus De Morgan*. Longmans, Green, and Company, (1882):216.

[DMCT2011] Ding, Yuxin, Xuejun Meng, Guangren Chai, and Yan Tang. "User Identification for Instant Messages." In *Neural Information Processing*, pp. 113-120. Springer Berlin Heidelberg, 2011.

[DRBH1986] Doublas, John E., Robert K. Ressler, Ann W. Burgess, and Carol R. Hartman. "Criminal profiling from crime scene analysis."*Behavioral Sciences & the Law* 4.4 (1986): 401-421.

[ESK2007] Meyer zu Eissen, Sven, Benno Stein, and Marion Kulig. "Plagiarism detection without reference collections." *Advances in data analysis* (2007): 359-366.

[EV1991] Elliot, W., and R. Valenza. "Was the Earl of Oxford the true Shakespeare." *Notes and Queries* 38.4 (1991): 501-506.

[FBI2013] Federal Bureau of Investigation, Behavioral Science Unit website. http://www.fbi.gov/hq/td/academy/bsu/bsu.htm. (accessed April 2, 2013)

[FM2008] Fafinski, Stefan, and Neshan Minassian. "UK Cybercrime Report 2008." *New York: Garlik* (2008): 1-55.

[GHM2005] Graham, Neil, Graeme Hirst, and Bhaskara Marthi. "Segmenting documents by stylistic character." *Natural Language Engineering* 11.4 (2005): 397-416.

[HAC2006] Hota, S., S. Argamon, and R. Chung. "Gender in Shakespeare: Automatic stylistics gender classification using syntactic, lexical, and lemma features."*Chicago Colloquium on Digital Humanities and Computer Science*. (2006): 1-6.

[Her2002] Herring, Susan C. "Computer‐mediated communication on the internet."*Annual review of information science and technology* 36.1 (2002): 109-168.

[Hin2003] Hindocha, Neal. "Instant Insecurity: Security Issues of Instant Messaging." *SecurityFocus.com Website [online], Jan* 13 (2003): 9.

[Hol1994] Holmes, David I. "Authorship attribution." *Computers and the Humanities* 28.2 (1994): 87-106.

[HF1995] Holmes, David I., and Richard S. Forsyth. "The Federalist revisited: New directions in authorship attribution." *Literary and Linguistic Computing* 10.2 (1995): 111-127.

[HPR2003] Hayne, Stephen C., Carol E. Pollard, and Ronald E. Rice. "Identification of comment authorship in anonymous group support systems." *Journal of Management Information Systems* 20.1 (2003): 301-326.

[HZ2012] Hunt, Hayes, and Michael P. Zabel. “United States: Text Messages as Trial Evidence – Authentication.” *Mondaq*, (2012). http://www.mondaq.com/unitedstates/x/200778/court+procedure/Text+Messages+As+Trial+Evidence+Authentication (accessed March 14, 2014)

[IBFD2013] Iqbal, Farkhund, Hamad Binsalleeh, Benjamin Fung, and Mourad Debbabi. "A unified data mining solution for authorship analysis in

anonymous textual communications." *Information Sciences* 231 (2013): 98-112.

[JB2005] Juola, Patrick, and R. Harald Baayen. "A controlled-corpus experiment in authorship identification by cross-entropy." *Literary and Linguistic Computing* 20.Suppl (2005): 59-67.

[JOP2001] Jones, Eric, Travis Oliphant, and Pearu Peterson. "SciPy: Open source scientific tools for Python." (2001-). http://www.scipy.org. (accessed April 1, 2013)

[JRS2004] Jain, Anil K., Arun Ross, and Salil Prabhakar. "An introduction to biometric recognition." *Circuits and Systems for Video Technology, IEEE Transactions on* 14.1 (2004): 4-20.

[Juo2006] Juola, Patrick. "Authorship attribution." *Foundations and Trends in information Retrieval* 1.3 (2006): 233-334.

[KAS2002] Koppel, Moshe, Shlomo Argamon, and Anat Rachel Shimoni. "Automatically categorizing written texts by author gender." *Literary and Linguistic Computing* 17.4 (2002): 401-412.

[KCAC2006] Kucukyilmaz, Tayfun, B. Cambazoglu, Cevdet Aykanat, and Fazli Can. "Chat mining for gender prediction." *Advances in Information Systems* (2006): 274-283.

[KCAC2008] Kucukyilmaz, Tayfun, B. Cambazoglu, Cevdet Aykanat, and Fazli Can. "Chat mining: Predicting user and message attributes in computer-mediated communication." *Information Processing & Management* 44.4 (2008): 1448-1466.

[LA2006] Levitan, Shlomo, and Shlomo Argamon. "Fixing the federalist: correcting results and evaluating editions for automated attribution." *Digital humanities* (2006): 323-328.

[Lea2009] Leafe, David. "Dear Garry. I've decided to end it all: The full stop that trapped a killer." *Daily Mail* (2009). http://www.dailymail.co.uk/news/article-1197187/Dear-Garry-Ive-decided-end-The-stop-trapped-killer.html (accessed November 18, 2013)

[Lov2002] Love, Harold. *Attributing authorship: an introduction*. Cambridge University Press, (2002): 15.

[LZC2006] Li, Jiexun, Rong Zheng, and Hsinchun Chen. "From fingerprint to writeprint." *Communications of the ACM* 49.4 (2006): 76-82.

[Man2012] Mande, Uttam. "Criminal Identification System Based On Facial Recognition Using Generalized Gaussian Mixture Model." *Asian Journal of Computer Science and Information Technology* 2.6 (2012): 176-179.

[McQ2005] McQuail, Denis. *McQuail's mass communication theory*. Sage Publications Limited, (2010): 16.

[MD2000] Moores, Trevor, and Gurpreet Dhillon. "Software piracy: a view from Hong Kong." *Communications of the ACM* 43.12 (2000): 88-93.

[MD2001] Meuwly, Didier, and Andrzej Drygajlo. "Forensic speaker recognition based on a Bayesian framework and Gaussian Mixture Modelling (GMM)." *2001: A Speaker Odyssey-The Speaker Recognition Workshop*. (2001): 145-150.

[Men1887] Mendenhall, Thomas Corwin. "The Characteristic Curves of Composition." *Science (New York, NY)* 9.214S (1887): 237.

[MM1994] Merriam, Thomas VN, and Robert AJ Matthews. "Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe." *Literary and Linguistic Computing* 9.1 (1994): 1-6.

[MW1964] Mosteller, Frederick, and David Wallace. "Inference and disputed authorship: The Federalist." (1964): 320.

[NIST2002] National Institute of Standards and Technology report to the United States Congress, "Summary of NIST Standards for Biometric Accuracy, Tamper Resistance, and Interoperability." Available at ftp://sequoyah.nist. gov/pub/nist_internal_reports/NISTAPP_Nov02.p df, November 2002.

[NIST2011] National Institute of Standards and Technology. Kissel, Richard, ed. "NIST IR 7298 Glossary of Key Information Security Terms". (2011): 115-208.

[OA2009a] Orebaugh, Angela, and Jeremy Allnutt. "Identifying and Characterizing Instant Messaging Authors for Cyber Forensics." *IATAC Magazine* 12.3 (2009): 20-22.

[OA2009b] Orebaugh, Angela, and Jeremy Allnutt. "Classification of instant messaging communications for forensics analysis." *The International Journal of Forensic Computer Science* 1 (2009): 22-28.

[OA2010b] Orebaugh, Angela, and Jeremy Allnutt. "Data Mining Instant Messaging Communications to Perform Author Identification for Cybercrime Investigations." *Digital Forensics and Cyber Crime* (2010): 99-110.

[Ore2004a] Orebaugh, Angela, et al. *Ethereal Packet Sniffing*. Vol. 1. Syngress Publishing, (2004): 2-8.

[Ore2005a] Orebaugh, Angela, et.al. *Intrusion prevention and active response: deploying network and host IPS*. Syngress Media Incorporated, (2005): 367-388.

[OBB2005b] Orebaugh, Angela, Simon Biles, and Jacob Babbin. *Snort cookbook*. O'Reilly Media, Incorporated, (2005): 253-259.

[Ore2006a] Orebaugh, Angela, et.al. *Wireshark and Ethereal network protocol analyzer toolkit*. Syngress Media Incorporated, (2006): 88.

[Ore2006b] Orebaugh, Angela. "An Instant Messaging Intrusion Detection System Framework: Using character frequency analysis for authorship identification and validation." *Carnahan Conferences Security Technology, Proceedings 2006 40th Annual IEEE International*. IEEE, (2006): 160-172.

[Ore2006c] Orebaugh, Angela. "Proactive forensics." *Journal of digital forensic Practice*1.1 (2006): 37-41.

[PC2004] Patton, Jon M., and Fazli Can. "A Stylometric Analysis of Yaşar Kemal's Ince Memed Tetralogy." *Computers and the Humanities* 38.4 (2004): 457-467.

[PC2011] Pal, Aditya, and Scott Counts. "Identifying topical authorities in microblogs." *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, (2011): 45-54.

[RD2003] Richiardi, Jonas, and Andrzej Drygajlo. "Gaussian Mixture Models for on-line signature verification." *Proceedings of the 2003 ACM SIGMM workshop on Biometrics methods and applications*. ACM, (2003): 115-122.

[Rog2003] Rogers, Marc. "The role of criminal profiling in the computer forensics process." *Computers & Security* 22.4 (2003): 292-298.

[Rev2008] Revett, Kenneth. *Behavioral biometrics: a remote access approach*. Wiley Publishing, (2008): 1-2.

[RLG2009] Rodrigues, Ricardo N., Lee Luan Ling, and Venu Govindaraju. "Robustness of multimodal biometric fusion methods against spoof attacks." *Journal of Visual Languages & Computing* 20.3 (2009): 169-179.

[RMG1998] Raja, Yogesh, Stephen J. McKenna, and Shaogang Gong. "Tracking and segmenting people in varying lighting conditions using colour." *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, (1998): 228-233.

[RQD2000] Reynolds, Douglas A., Thomas F. Quatieri, and Robert B. Dunn. "Speaker verification using adapted Gaussian mixture models." *Digital signal processing* 10.1 (2000): 19-41.

[Rud1998] Rudman, Joseph. "The state of authorship attribution studies: Some problems and solutions." *Computers and the Humanities* 31.4 (1998): 351-365.

[Smi2002] Smith, Lindsay I. "A tutorial on principal components analysis." *Cornell University, USA* 51 (2002): 52.

[TLML2004] Teng, Gui-Fa, Mao-Sheng Lai, Jian-Bin Ma, and Ying Li. "E-mail authorship mining based on SVM for computer forensic." *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference*. Vol. 2. IEEE, (2004): 1204-1207.

[Tur2008] Turman v. Commonwealth, Myron Turman v. Commonwealth of Virginia, Record No. 072174, October 31, 2008. http://caselaw.findlaw.com/va-supreme-court/1237453.html (accessed April 1, 2013)

[TSH1996] Tweedie, Fiona J., Sameer Singh, and David I. Holmes. "Neural network applications in stylometry: The Federalist Papers." *Computers and the Humanities* 30.1 (1996): 1-10.

[ZLCH2006] Zheng, Rong, Jiexun Li, Hsinchun Chen, and Zan Huang. "A framework for authorship identification of online messages: Writing‐style features and classification techniques." *Journal of the American Society for Information Science and Technology* 57.3 (2006): 378-393.

CURRICULUM VITAE

Angela Orebaugh is a cybersecurity technologist and author with a broad spectrum of expertise in information technology and security. She synergizes her strategic and technical experience within industry, academia, and government to advise clients on next generation technologies.

As a Guest Researcher, Ms. Orebaugh is involved in several security initiatives with the National Institute of Standards and Technology (NIST), including technical Special Publications (800 series), the National Vulnerability Database (NVD), Security Content Automation Protocol (SCAP) project, and secure eVoting.

Ms. Orebaugh is an Adjunct Professor for George Mason University where she performs research and teaching in intrusion detection, cyber forensics, and cybercrime. She assisted in creating the Masters of Science in Computer Forensics degree program in the Department of Electrical and Computer Engineering. Her current research interests include peer-reviewed publications in the areas of intrusion detection and prevention, attacker profiling, network forensics, and behavioral biometrics.

Ms. Orebaugh is the author of the Syngress best sellers *Nmap in the Enterprise*, *Wireshark and Ethereal Network Protocol Analyzer Toolkit*, and *Ethereal Packet Sniffing*. She has also co-authored the *Snort Cookbook*, *Intrusion Prevention and Active Response*, and *How to Cheat at Configuring Open Source Security Tools*. She is a frequent speaker at a variety of security conferences and technology events, including the SANS Institute and The Institute for Applied Network Security. In 2011, Ms. Orebaugh was named Booz Allen Hamilton's first Cybersecurity Fellow.

Ms. Orebaugh holds a Masters degree in Computer Science and a Bachelors degree in Computer Information Systems from James Madison University.