



---

Annual ADFSL Conference on Digital Forensics, Security and Law

2014  
Proceedings

---

May 29th, 11:20 AM


## Visualizing Instant Messaging Author Writeprints for Forensic Analysis

Angela Orebaugh  
George Mason University, [aorebaug@gmu.edu](mailto:aorebaug@gmu.edu)

Jason Kinser  
George Mason University, [jkinser@gmu.edu](mailto:jkinser@gmu.edu)

Jeremy Allnutt  
George Mason University, [jallnutt@gmu.edu](mailto:jallnutt@gmu.edu)

Follow this and additional works at: <https://commons.erau.edu/adfsl>

 Part of the [Aviation Safety and Security Commons](#), [Computer Law Commons](#), [Defense and Security Studies Commons](#), [Forensic Science and Technology Commons](#), [Information Security Commons](#), [National Security Law Commons](#), [OS and Networks Commons](#), [Other Computer Sciences Commons](#), and the [Social Control, Law, Crime, and Deviance Commons](#)

---

### Scholarly Commons Citation

Orebaugh, Angela; Kinser, Jason; and Allnutt, Jeremy, "Visualizing Instant Messaging Author Writeprints for Forensic Analysis" (2014). *Annual ADFSL Conference on Digital Forensics, Security and Law*. 8. <https://commons.erau.edu/adfsl/2014/thursday/8>

This Peer Reviewed Paper is brought to you for free and open access by the Conferences at Scholarly Commons. It has been accepted for inclusion in Annual ADFSL Conference on Digital Forensics, Security and Law by an authorized administrator of Scholarly Commons. For more information, please contact [commons@erau.edu](mailto:commons@erau.edu).

**EMBRY-RIDDLE**  
Aeronautical University™  
SCHOLARLY COMMONS

(c)ADFSL



## VISUALIZING INSTANT MESSAGING AUTHOR WRITEPRINTS FOR FORENSIC ANALYSIS

Angela Orebaugh

[aorebaug@gmu.edu](mailto:aorebaug@gmu.edu)

Jason Kinser

[jkinser@gmu.edu](mailto:jkinser@gmu.edu)

Jeremy Allnutt

[jallnutt@gmu.edu](mailto:jallnutt@gmu.edu)

George Mason University

Fairfax, Virginia

### ABSTRACT

As cybercrime continues to increase, new cyber forensics techniques are needed to combat the constant challenge of Internet anonymity. In instant messaging (IM) communications, criminals use virtual identities to hide their true identity, which hinders social accountability and facilitates cybercrime. Current instant messaging products are not addressing the anonymity and ease of impersonation over instant messaging. It is necessary to have IM cyber forensics techniques to assist in identifying cyber criminals as part of the criminal investigation. Instant messaging behavioral biometrics include online writing habits, which may be used to create an author writeprint to assist in identifying 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. The analysis of IM author writeprints in this paper provides a foundation for using behavioral biometrics as a cyber forensics element of criminal investigations. This paper demonstrates a method to create and analyze behavioral biometrics-based instant messaging writeprints as cyber forensics input for cybercrime investigations. The research uses the Principal Component Analysis (PCA) statistical method to analyze IM conversation logs from two distinct data sets to visualize authorship identification.

**Keywords:** writeprints, authorship attribution, authorship identification, principal component analysis

### 1. INTRODUCTION

Synchronous computer-mediated Communication (CMC) occurs in real time and requires the simultaneous participation of users. Point-to-point CMC is online text intended for a single recipient. This paper is focused on the analysis of instant messaging, a synchronous form of point-to-point CMC. 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.

This paper uses authorship analysis and statistical techniques to create and analyze behavioral biometrics-based instant messaging writeprints to assist in identifying online cyber criminals. IM writeprints may be used as an element in a multimodal biometrics systems in conjunction with

traditional criminal investigation techniques to assist with cybercrime decision support. 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. This research uses authorship analysis techniques to create an IM-specific stylometric feature set taxonomy to determine writer invariants for various authors. Using Principal Component Analysis (PCA), this research analyzes author writeprints from IM conversation logs from two distinct datasets for authorship identification. 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. This research creates author writeprints from IM conversations from two unique datasets of synchronous, point-to-point instant messaging logs.

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

1. Create an IM feature set taxonomy
2. Using PCA, reduce the dimensions and show separation in author and author category writeprints

## 2. INSTANT MESSAGING AND CYBERCRIME

Cybercrime involves any criminal activity that is committed with the aid of a communication device in a network, such as the Internet, telephone lines, or mobile networks such as cellular communication (Fafinski and Minassian, 2008). Instant messaging's anonymity hinders social accountability and leads to IM-assisted cybercrime facilitated by the following:

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

In IM communications, 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. 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. 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. 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 (Cross, 2008). 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 (Moores and Dhillon, 2000). Additionally, criminals such as terrorist groups, gangs, and cyber intruders use IM to communicate (Abbasi and Chen, 2005). 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 criminal suspects as part of the criminal investigation (Abbasi and Chen, 2006). 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 (Orebaugh and Allnutt, 2009; Orebaugh and Allnutt 2010). This paper explores the cyber forensic technique of

behavioral biometrics to assist in identifying cyber criminals and collecting data for the criminal investigation.

### **2.1 Behavioral Biometrics Writeprints**

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 (BioPassword, 2006). As with handwriting, users have certain online writing habits that are unconscious and deeply ingrained (Teng, Lai, Ma, and Li, 2004). Online writing habits, known as stylometric features, include composition syntax and layout, vocabulary patterns, unique language usage, and other stylistic traits. Thus, certain stylometric features may be used to create an author writeprint to help identify an author of a particular piece of work (De Vel, Anderson, Corney, and Mohay, 2001).

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. Writeprints can provide cybercrime investigators a unique behavioral biometric tool for analyzing IM-assisted cybercrimes. Writeprints can be used as input to a criminal cyberprofile and as an element of a multimodal system to perform cyber forensics and cybercrime investigations (Jain, Ross, and Prabhakar, 2004; Rodrigues, Ling, and Govindaraju, 2009). This paper uses authorship analysis techniques to create an author's IM writeprint based on behavioral biometrics.

### **2.2 Writeprints for 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 identification uses a variety of computer-aided statistical methods to analyze text to determine the most plausible author of a piece of text. Authorship identification may be applied to IM to assist in identifying criminals who hide their true identity or impersonate a known individual.

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 (Zheng, Li, Chen, Huang, 2006). 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 mediums makes them especially interesting since authors take less time to craft their responses (Hayne, Pollard, and Rice, 2003). 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 (Kucukyilmaz, Cambazoglu, Aykanat, Can, 2008). 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 (Kucukyilmaz, Cambazoglu, Aykanat, Can, 2008). 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 (De Vel, Anderson, Corney, Mohay, 2001). This paper uses the data obtained from two unique datasets of synchronous, point-to-point instant messaging logs.

## **3. RELATED WORKS**

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.

Some of the earliest authorship analysis research dates back to the fourth century BC, when librarians in the library of Alexandria studied the authentication of texts attributed to Homer (Love, 2002). 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 (De Morgan, 1882). Recent research has introduced authorship analysis to computer-mediated communications with promising results (De Vel, Anderson, Corney, and Mohay, 2001; Orebaugh and Allnutt, 2010).

Olivier De Vel published several papers on authorship identification and characterization. The paper *Mining E-mail Content for Author Identification Forensics* (De Vel, Anderson, Corney, and Mohay, 2001) 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.

The paper *A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques* (Zheng, Li, Chen, and Huang, 2006) presented a comparison of techniques for author identification by using several classification algorithms to analyze features. The authors leveraged existing feature sets from (De Vel, Anderson, Corney, and Mohay, 2001) 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 paper *Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace* (Abbasi and Chen, 2008) 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" (Abbasi and Chen, 2008). 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.

Most related works apply authorship analysis to datasets of email and newsgroup postings. Preliminary journal articles and conference presentations (Orebaugh, 2006; Orebaugh and Allnutt, 2009; Orebaugh and Allnutt, 2010) from this research are the only comprehensive examination of IM authorship analysis.

#### 4. INSTANT MESSAGING WRITEPRINT ANALYSIS

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 IM author writeprints that provide cybercrime investigators a unique tool for investigating IM-assisted cybercrimes. At a high level this research performs the following:

1. Develops a stylometric feature set
2. Pre-processes the data
3. Creates writeprints
4. Creates PCA visualizations of writeprints.

The detailed research process is illustrated in Figure 1.

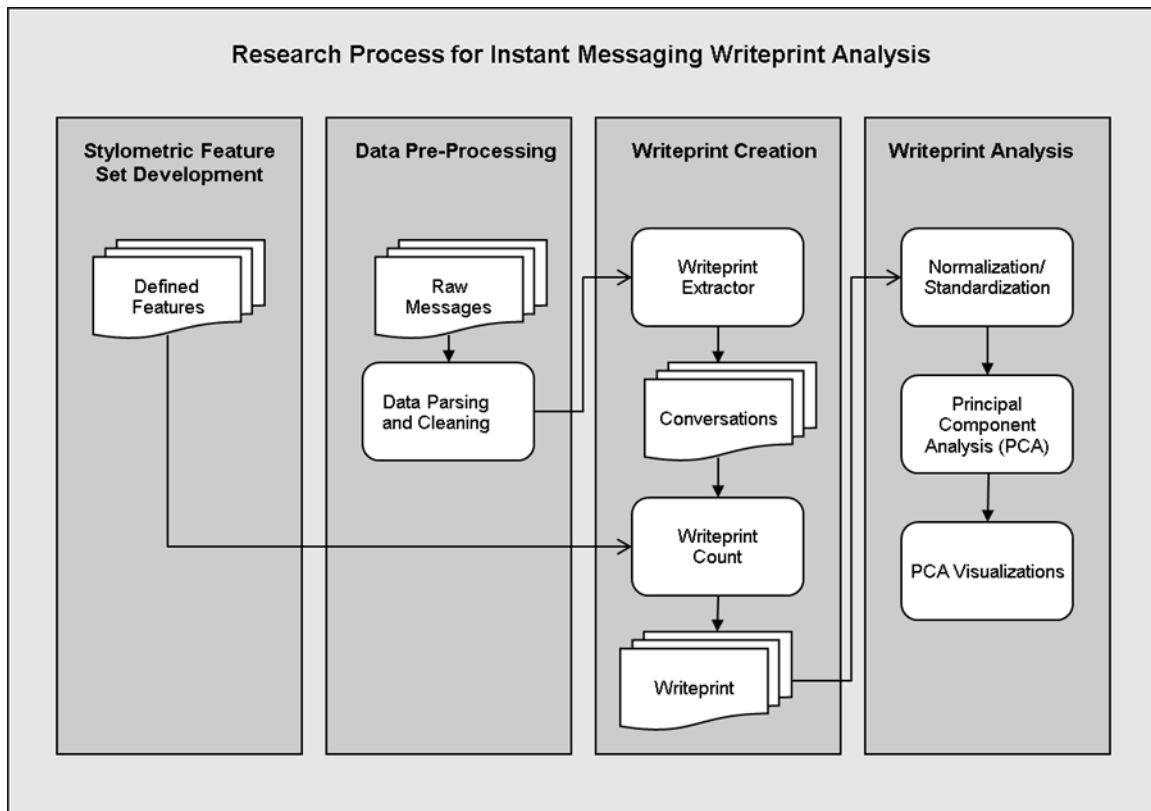


Figure 1 Research Process for Instant Messaging Writeprint Analysis

#### 4.1 Feature Set Taxonomy

Stylometric features are characteristics that can be derived from instant messages to facilitate authorship analysis (Abbasi and Chen, 2006). 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. The feature set in this research is a 356-dimensional vector including lexical, syntactic, and structural features, shown in Figure 2. The number of features in each category is shown in parenthesis.

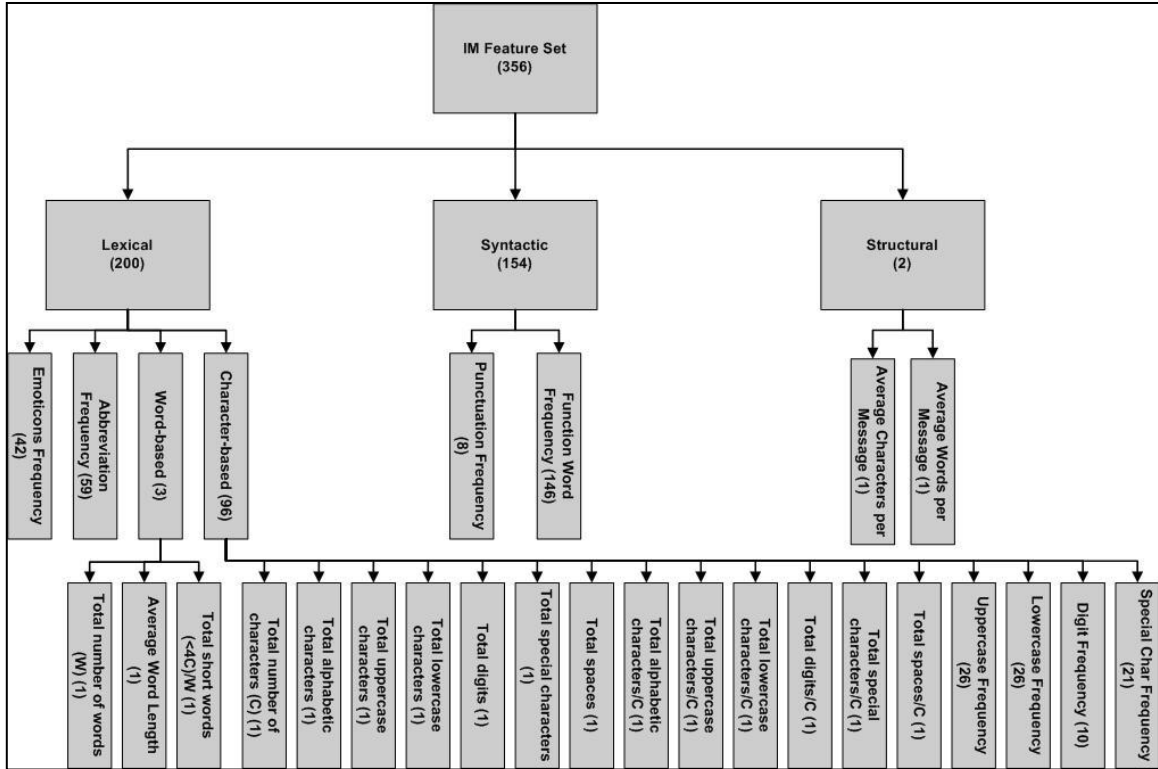


Figure 2 Instant Messaging Stylometric Feature Set Taxonomy

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 that reflect an author’s preferred ways to express himself or herself and form sentences. 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 streamlined set of features that best reveal the true writing style of the author. Each stylometric feature in the taxonomy was selected for its relevance to IM communications to create a feature set robust enough to determine writer invariants for various authors and author categories.

### 4.2 Writeprint Creation

First, the writeprint extractor module splits the logs into a configurable conversation size. A conversation is a set of messages  $\{M_1, \dots, M_p\}$ , for example 50 messages per conversation. A message consists of the text delineated by the newline or end-of-line (EOL) character. 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  $\{M_1, \dots, M_p\}$  of each supplied author ( $A_n$ ). 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 ( $A_n$ ) of the writeprint ( $W_x$ ). 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  $\{F_1, \dots, F_q\}$ , where  $q=100$ , for a set of messages  $\{M_1, \dots, M_p\}$  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, 2  
5, 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, 12  
50, 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
```

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.

After the writeprints are normalized and standardized, PCA models are created and used to visualize and analyze the data. PCA 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 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. This research uses Gnuplot to plot three-dimensional plots of the PCA data.

## 5. DATASET DESCRIPTIONS

Dataset #1 contains personal IM conversation logs collected by the Gaim and Adium clients over a three-year period. The data includes conversation logs for 19 users. 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. U.S. Cyberwatch data was collected from April 2004 to March 2007. The data includes 105 complete IM logs between undercover agents and child predators. The 5 authors with the least number of messages were not used in the experiments in this research because the number of messages was too small for sufficient testing.

## 6. EXPERIMENT RESULTS

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.

### 6.1 Results for Dataset #1, Known Authors

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

Figure 3 shows Dataset #1 PCA plot results for conversations consisting of 250 messages for each of the 19 authors. Even at a high number of authors, this plot does show some groupings and separation between the authors.

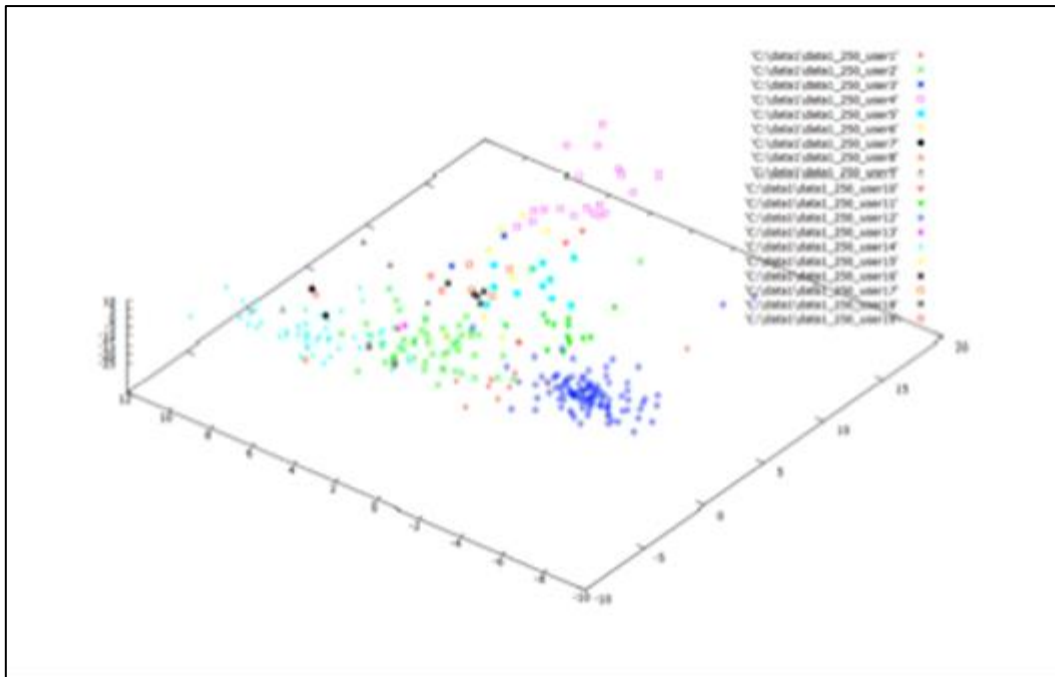


Figure 3 Dataset 1, PCA Plot Results, 250 Messages, All 19 Authors

Figures 4 through 6 show PCA plots of Dataset #1 author writeprints broken down into 6, 6, and 7 authors respectively. Figure 4 shows Dataset #1 PCA plot results for conversations consisting of 250 messages for Authors A1-A6. It is easy to see the separate groupings for each author in this plot.

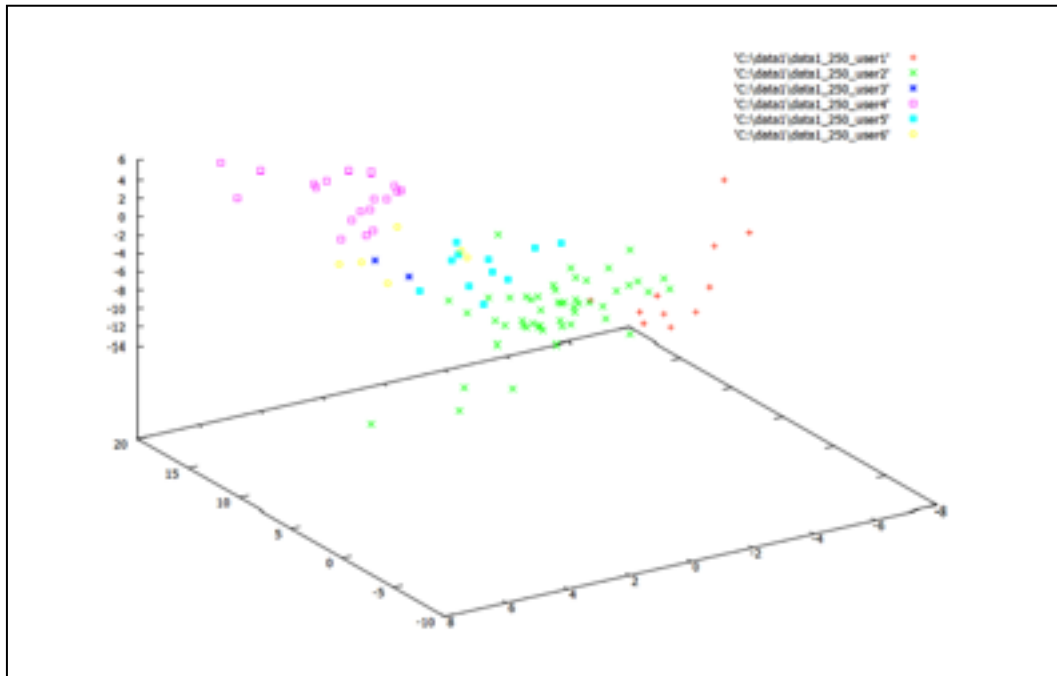


Figure 4 Dataset 1, PCA Plot Results, 250 Messages, Authors A1-A6

Figure 5 shows Dataset #1 PCA plot results for conversations consisting of 250 messages for Authors A7-A12. In this plot it is very easy to see separate groupings for each author.

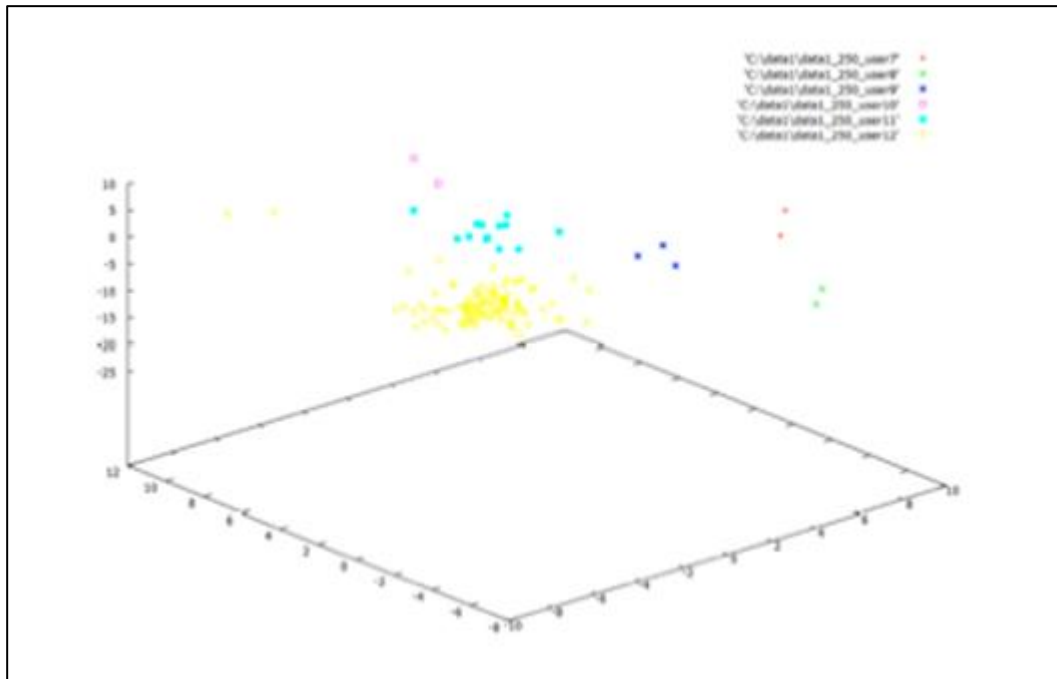


Figure 5 Dataset 1, PCA Plot Results, 250 messages, Authors A7-A12

Figure 6 shows Dataset #1 PCA plot results for conversations consisting of 250 messages for Authors A13-A19. It is easy to see the separate groupings for each author in this plot.

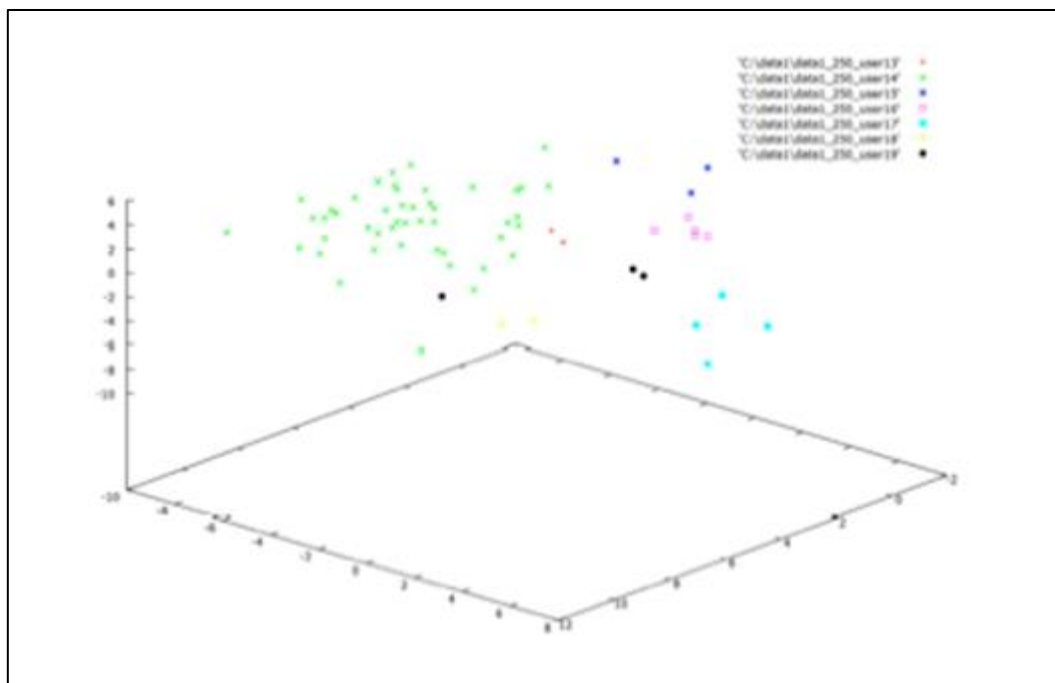


Figure 6 Dataset 1, PCA Plot Results, 250 Messages, Authors A13-A19

Figure 7 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 easily show separate groupings for each author.

Figure 8 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 easily shows separate groupings for each author.

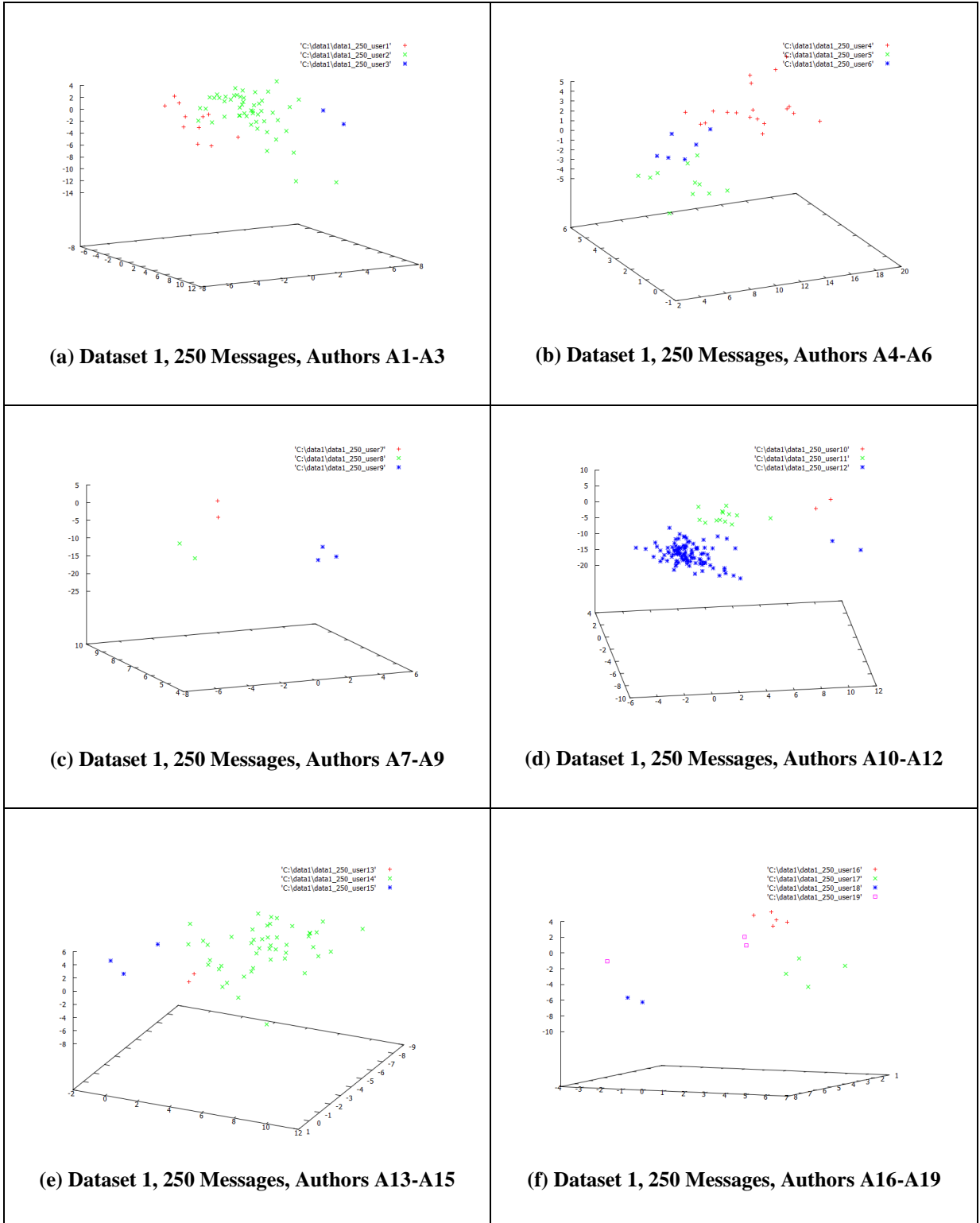


Figure 7 Dataset 1, PCA Plot Results, 250 Messages, Authors A1-A19

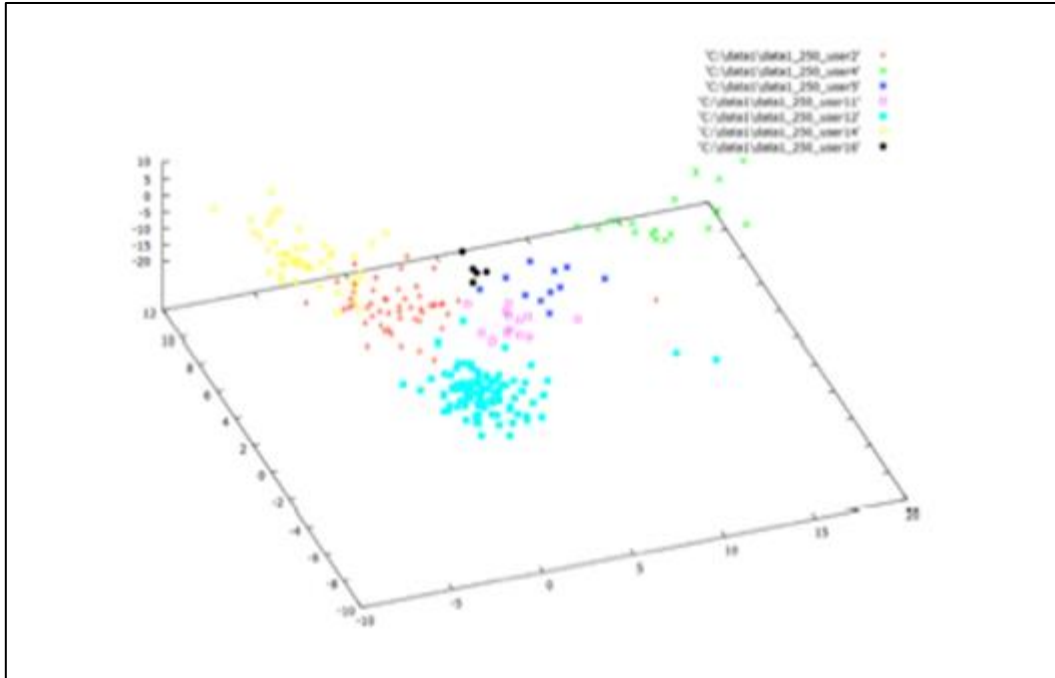


Figure 8 Dataset 1, PCA Plot Results, 250 Messages, Top 7 Authors

Figure 9 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.

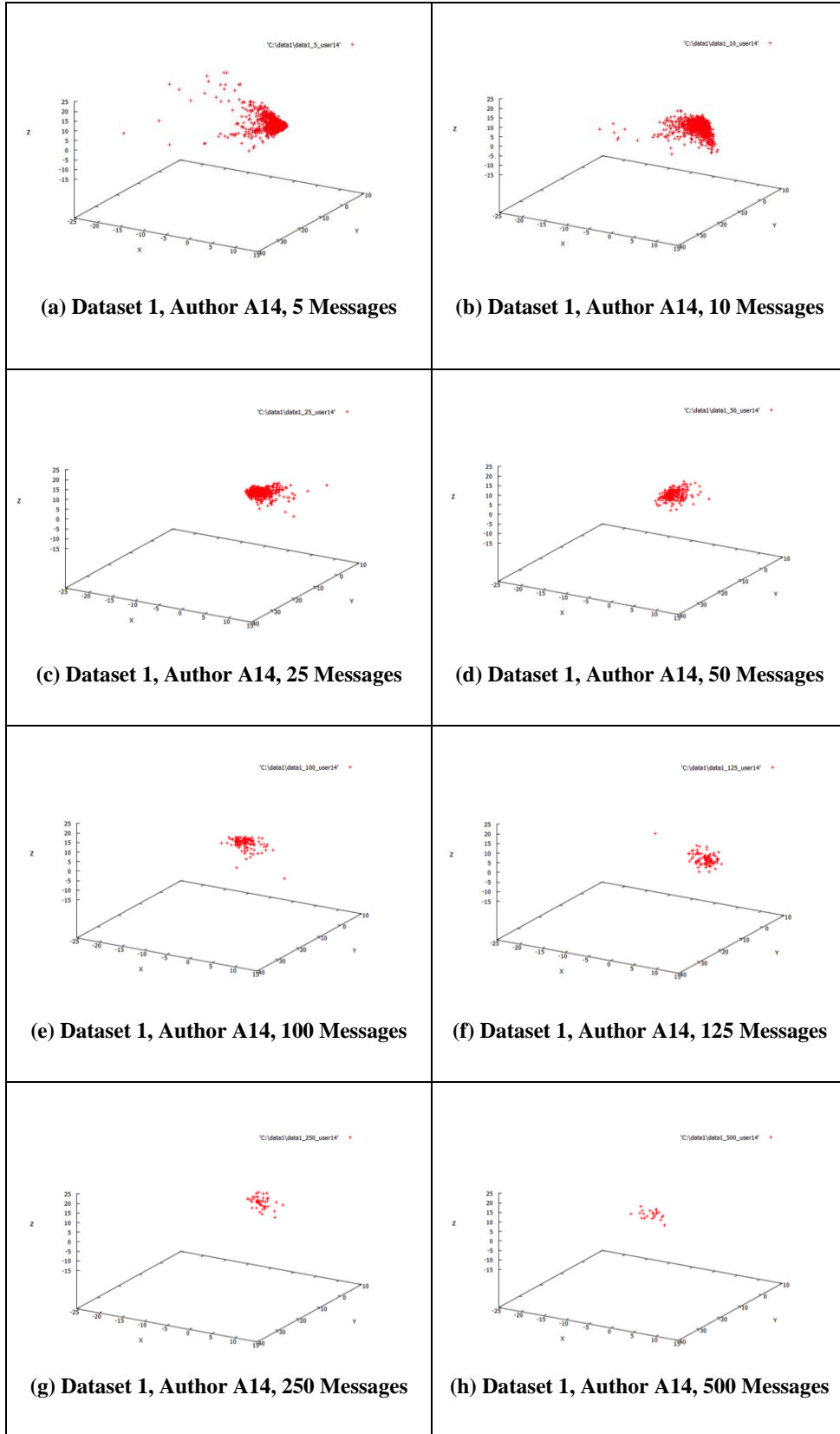


Figure 9 Dataset 1, PCA Plot Results, Author A14, All Conversation Sizes

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 10 shows the inverse relationship of standard deviation and conversation size for the Author A14 results shown in Figure 9. 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.

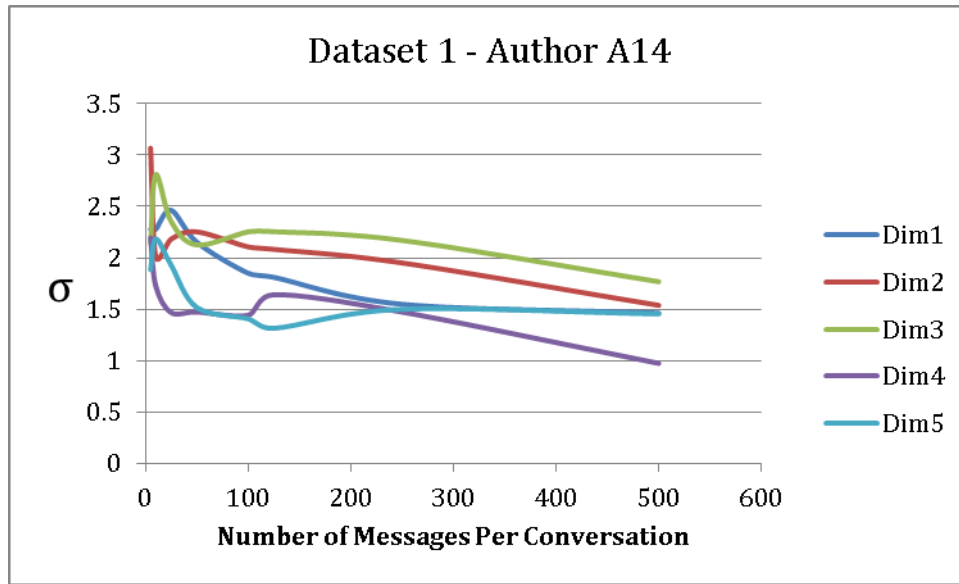


Figure 10 Dataset 1, Author A14, Conversation Size/Standard Deviation Relationship

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 1, 96% of the 95 values exhibited decreased standard deviation as the conversation size increased.

Table 1 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 $\sigma$
1	19	5	95 (across sets of 5,10,25,50,100,125,250,500 messages per conversation)	96%

Figures 11 and 12 show Dataset #1 PCA plot results for multiple sequential 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.

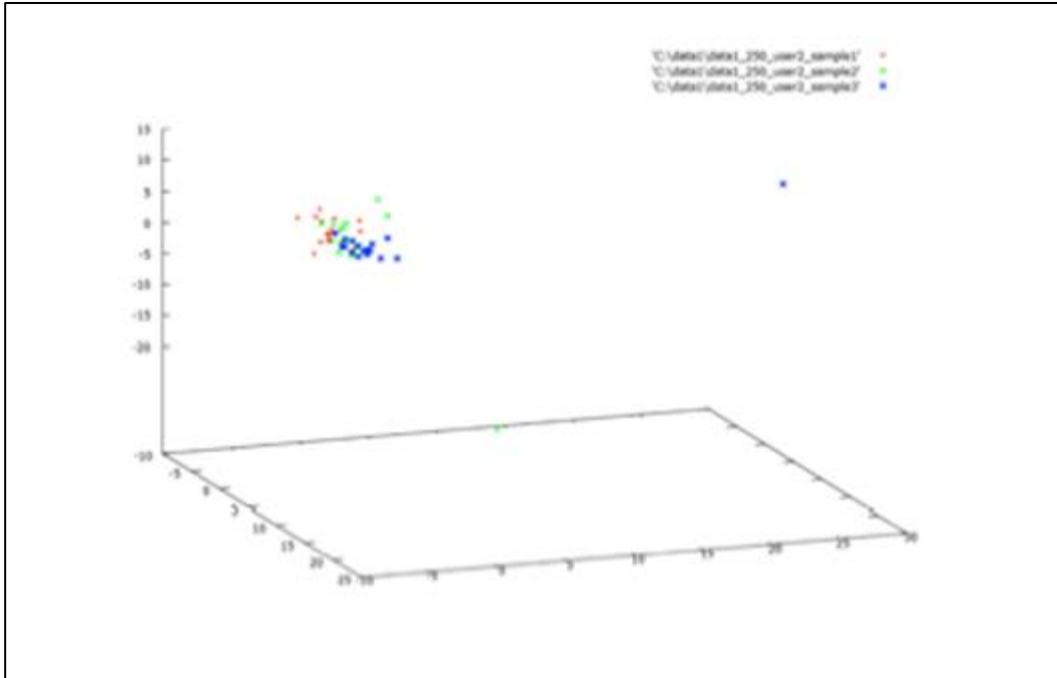


Figure 11 Dataset 1, PCA Plot Results, 250 Messages, Author A2 Samples

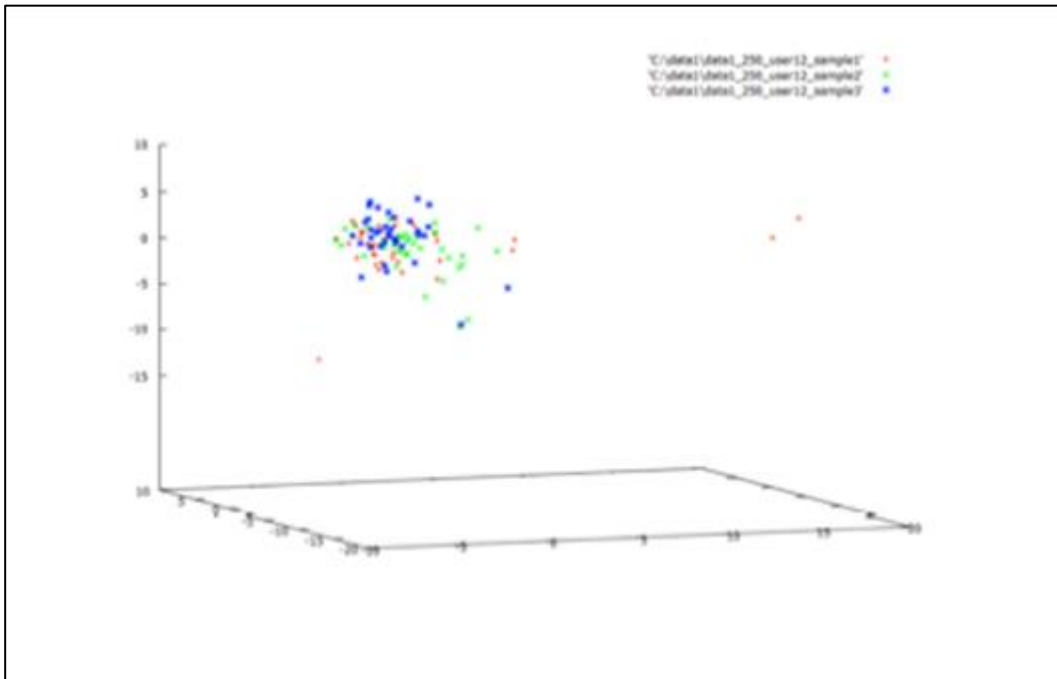


Figure 12 Dataset 1, PCA Plot Results, 250 Messages, Author A12 Samples

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

Dataset #2 experiments include 100 authors from which to determine identification. For each author, IM writeprints are divided into conversations containing 10, 25, 50, and 90 messages respectively. Figure 13 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. Although it is difficult to see with this many authors, this plot does show some separation between the authors.

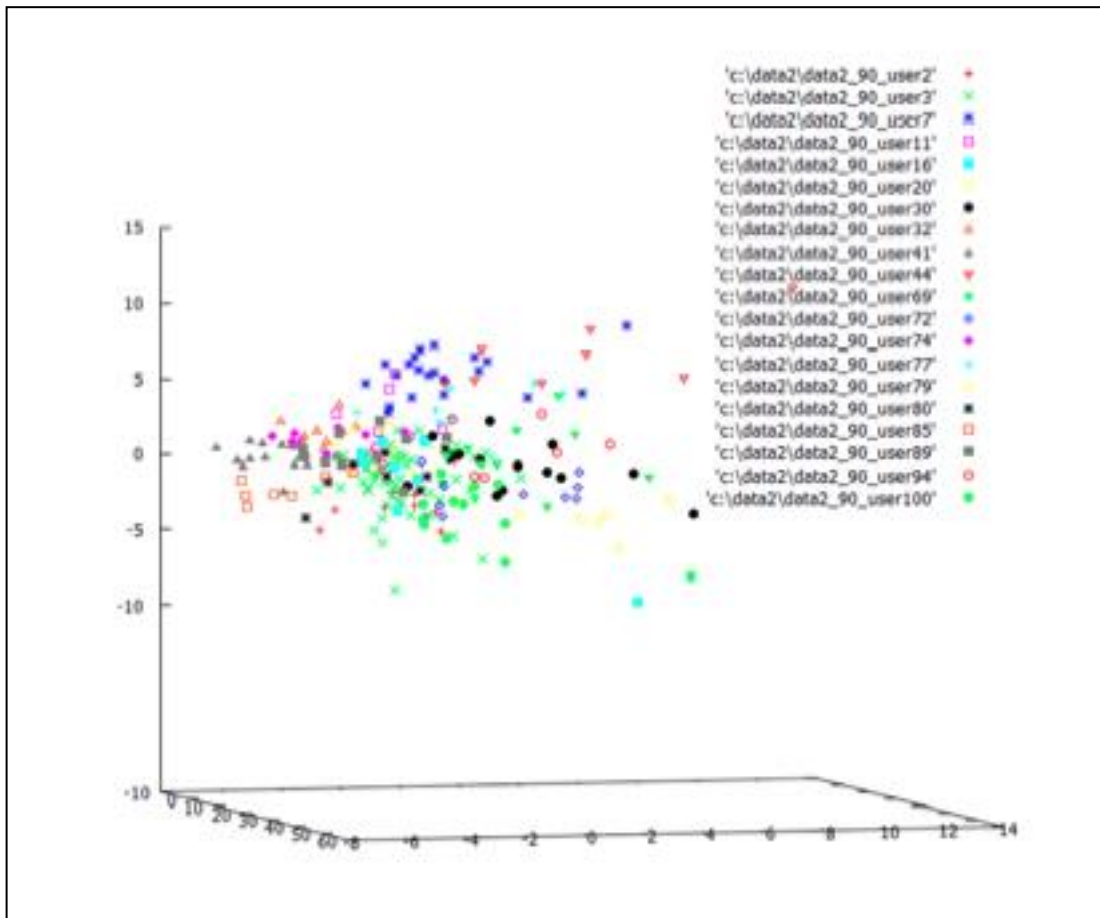


Figure 13 Dataset 2, PCA Plot Results, 90 Messages, Top 20 Authors

Figure 14 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.

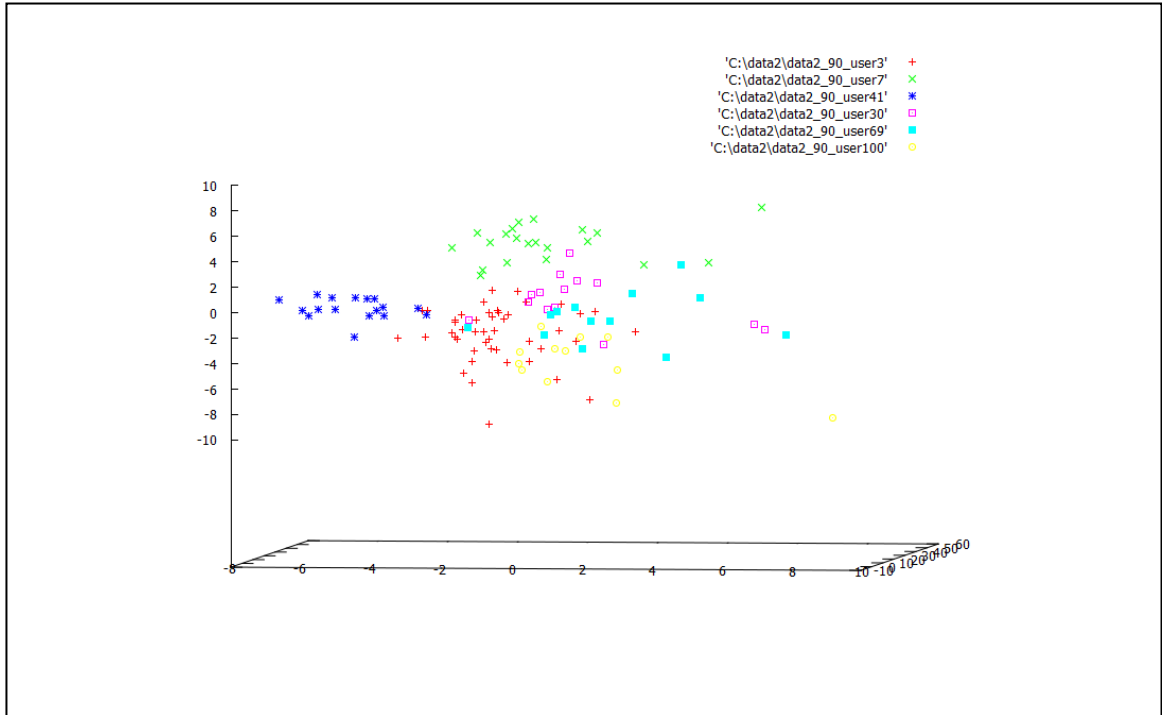


Figure 14 Dataset 2, PCA Plot Results, 90 Messages, Top 6 Authors

Figure 15 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 easily shows separate groupings for each author.

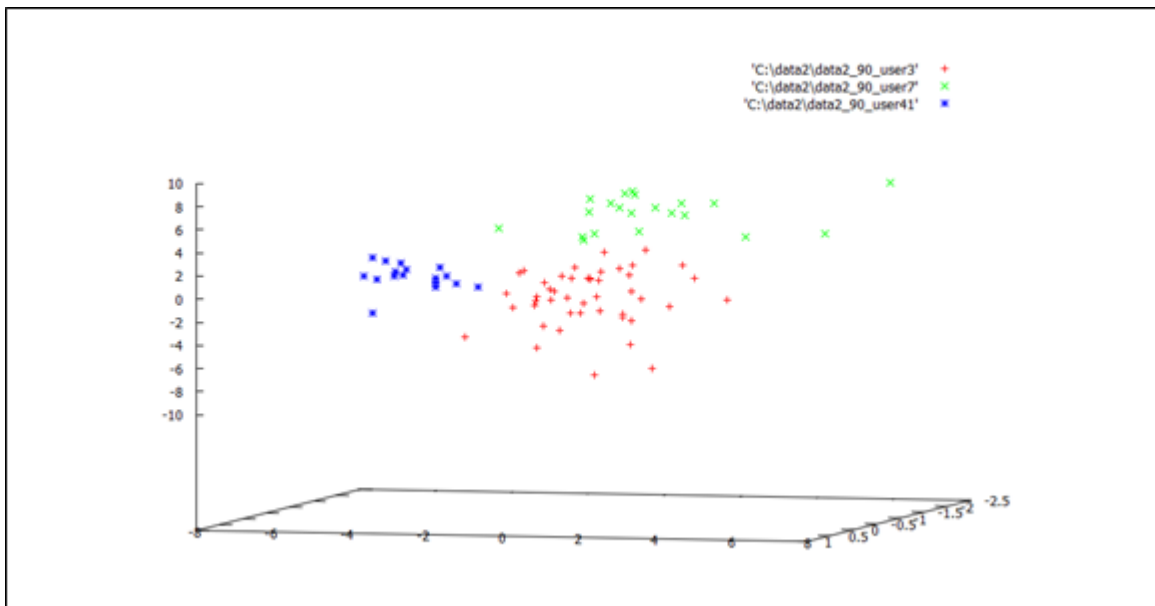


Figure 15 Dataset 2, PCA Plot Results, 90 Messages, Top 6 Authors - Subset 1

Figure 16 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.

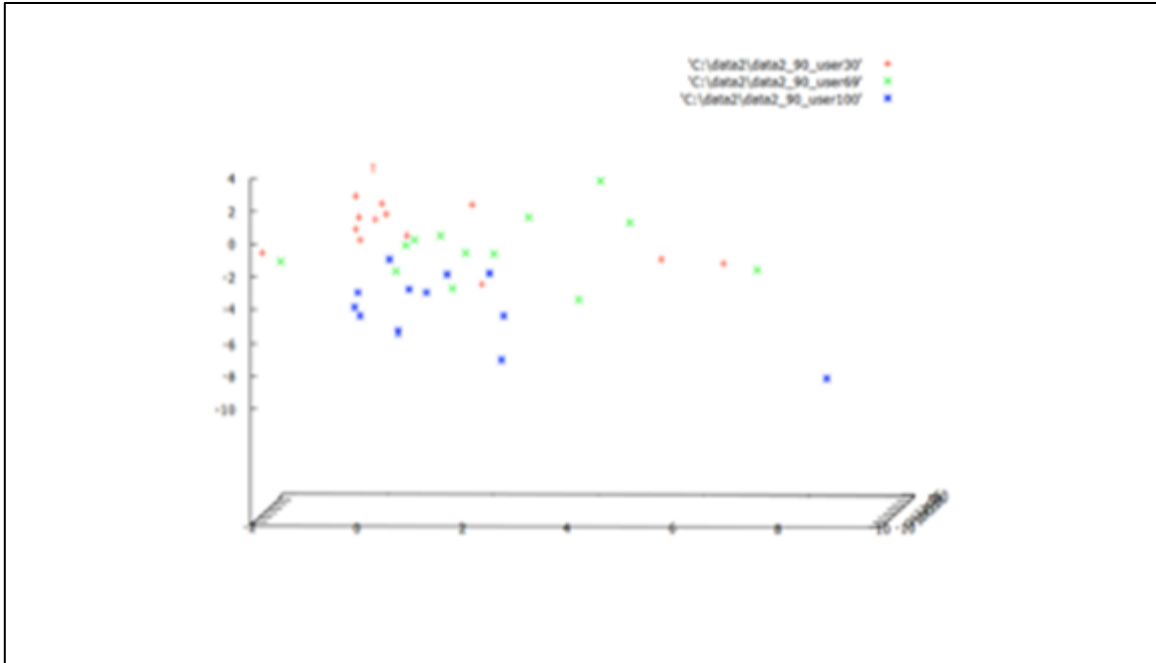


Figure 16 Dataset 2, PCA Plot Results, 90 Messages, Top 6 Authors - Subset 2

Figure 17 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.

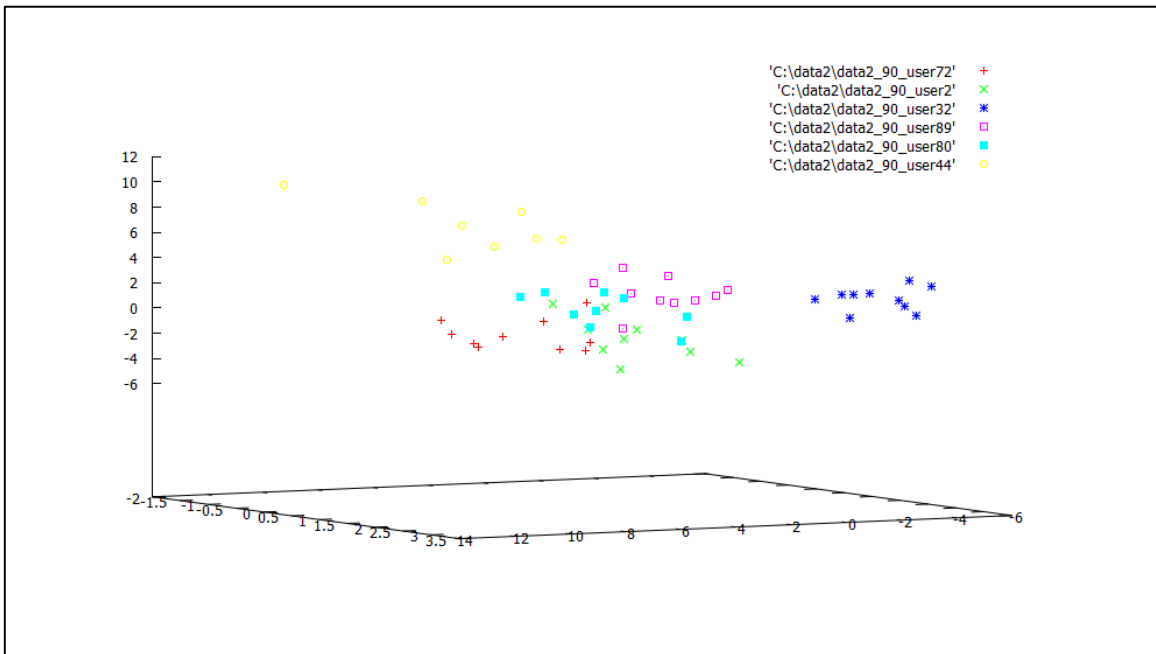


Figure 17 Dataset 2, PCA Plot Results, 90 Messages, Second Top 6 Authors

Figure 18 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.

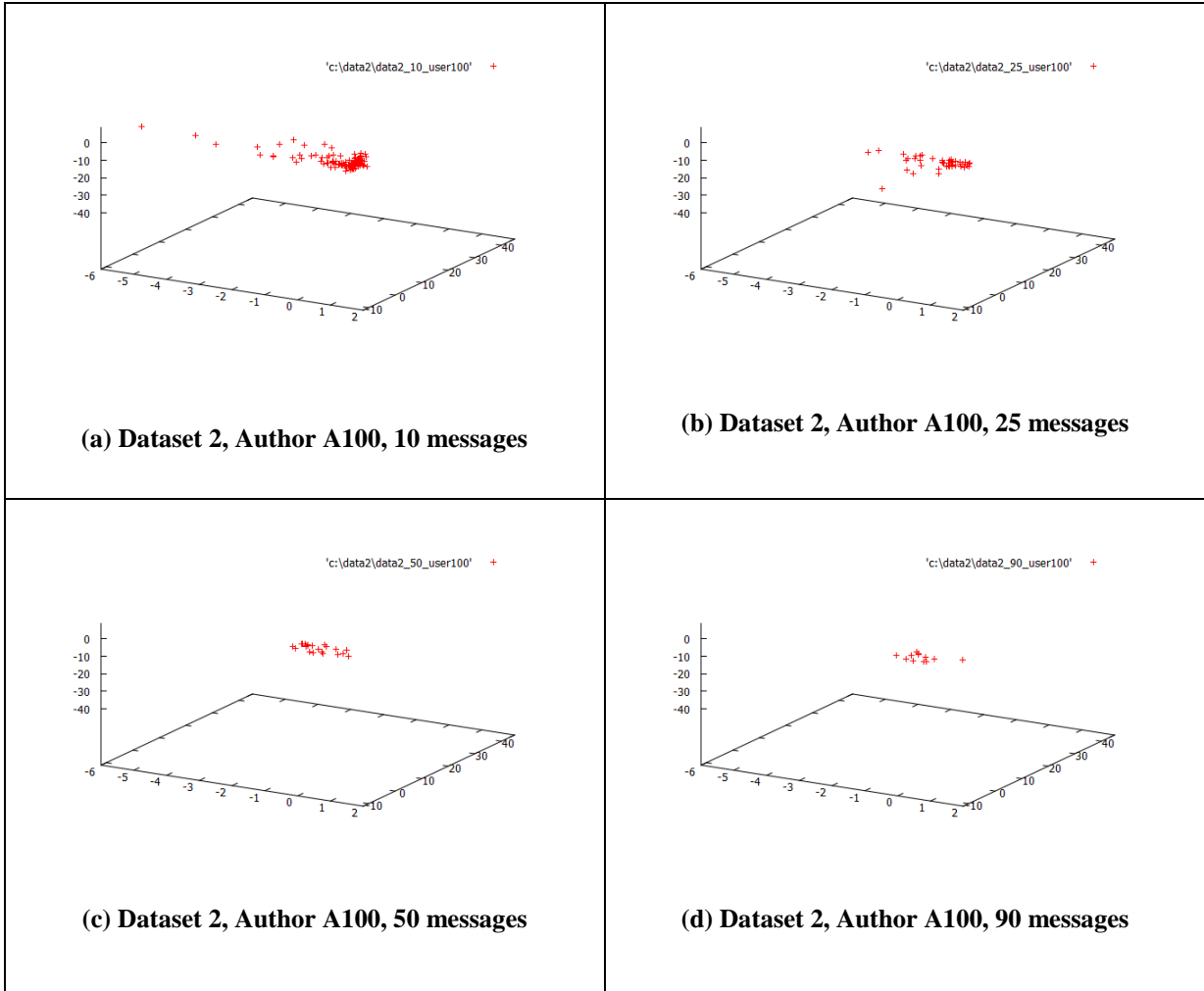


Figure 18 Dataset 2, PCA Plot Results, Author A100, All Conversation Sizes

Conversation size can be analyzed in more detail by calculating the standard deviation of the data within each conversation size. Figure 19 shows the inverse relationship of standard deviation and conversation size for the Author A100 results shown in Figure 18. 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.

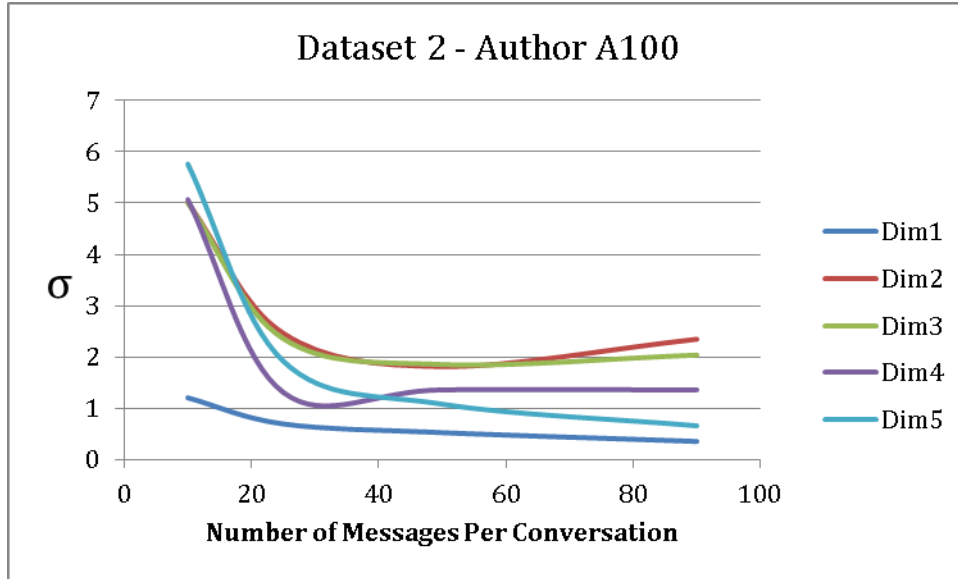


Figure 19 Dataset 2, Author A100, Conversation Size/Standard Deviation Relationship

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 2, 86% of the 500 values exhibited decreased standard deviation as the conversation size increased.

Table 2 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 $\sigma$
2	100	5	500 (across sets of 10,25,50,90 messages per conversation)	86%

Figures 20 and 21 show Dataset #2 PCA plot results for multiple sequential 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.

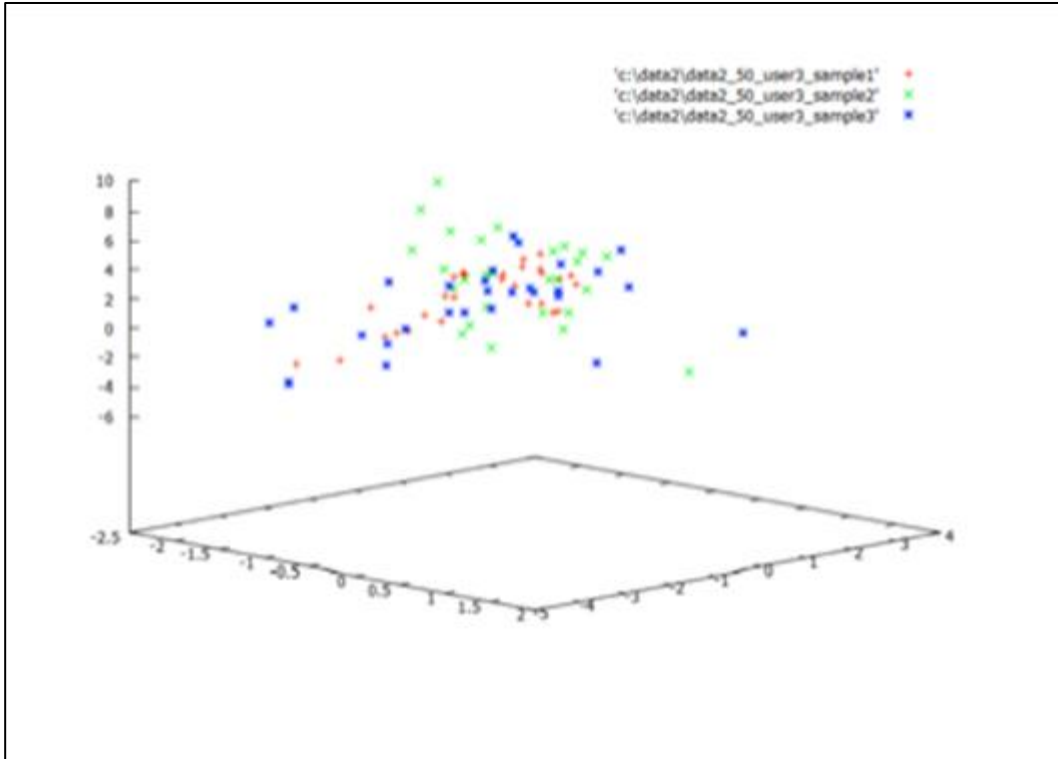


Figure 20 Dataset 2, PCA Plot Results, 50 Messages, Author A3 Samples

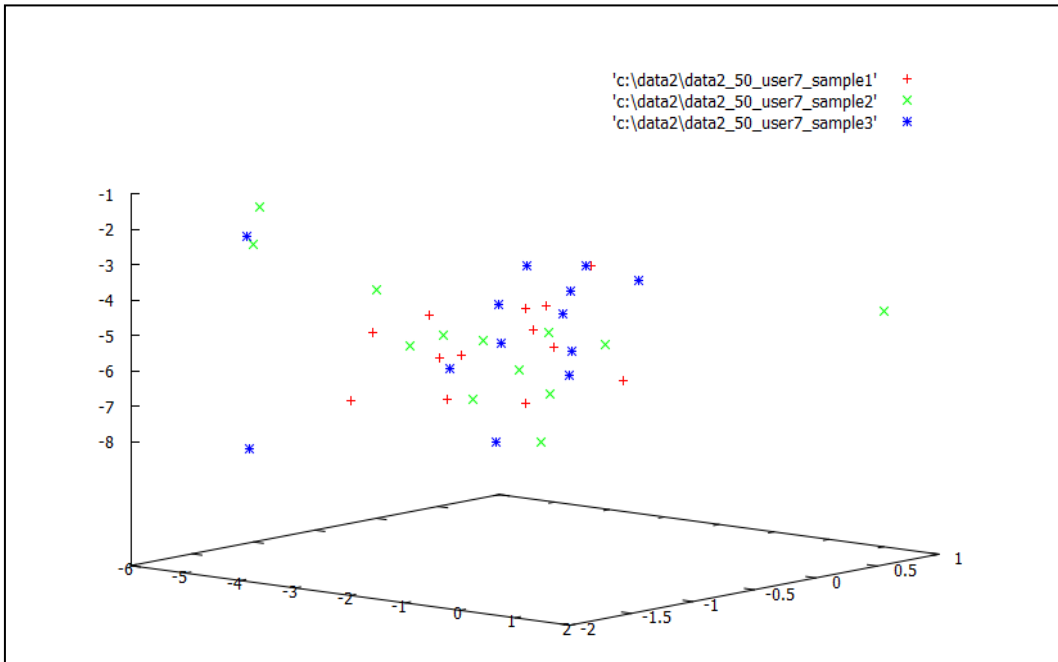


Figure 21 Dataset 2, PCA Plot Results, 50 Messages, Author A7 Samples

## 7. CONCLUSIONS

This paper provides a foundation for using behavioral biometrics as a cyber forensics element for criminal investigations by demonstrating the effectiveness of creating instant messaging author writeprints to be used in conjunction with traditional criminal investigation techniques.

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

1. Created an IM feature set taxonomy
2. Used PCA to reduce the dimensions and show separation in author writeprints

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.

This paper addresses the existing research gap in applying authorship analysis techniques to instant messaging communications to facilitate authorship identification. It provides a new approach and techniques to assist in identifying cyber criminal suspects and collecting digital evidence as part of the criminal 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 determine writer invariants for various authors and author categories. Cybercrime investigators may leverage the techniques presented in this paper in conjunction with traditional forensics investigative techniques to aid in cybercrime decision support.

## REFERENCES

- Abbasi, Ahmed, & Chen, Hsinchun. (2005). Applying authorship analysis to extremist-group web forum messages. *Intelligent Systems, IEEE* 20.5, 67-75.
- Abbasi, Ahmed, & Chen, Hsinchun. (2006). Visualizing authorship for identification. *Intelligence and Security Informatics*, 60-71.
- Abbasi, Ahmed, & Chen, Hsinchun. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, 26(2), 7.
- BioPassword. (2006). Authentication Solutions Through Keystroke Dynamics. Retrieved on April 2, 2013 from <http://www.infosecurityproductsguide.com/technology/2007/BioPassword.html>
- Cross, Michael. (2008). *Scene of the Cybercrime*. Syngress Publishing, 679-690.
- De Vel, Olivier, Anderson, A., Corney, M., & Mohay, G. (2001). Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4), 55-64.
- De Morgan, A. & Elizabeth S. (1882). *Memoir of Augustus De Morgan*. Longmans, Green, and Company, 216.
- Fafinski, Stefan, & Minassian, Neshan. (2008). UK Cybercrime Report 2008. New York, NY: *Garlik*, 1-55.

- Hayne, Stephen C., Pollard, Carol E., & Rice, Ronald E. (2003). Identification of comment authorship in anonymous group support systems. *Journal of Management Information Systems*, 20(1), 301-326.
- Jain, Anil K., Arun, R., & Prabhakar, Salil. (2004). An introduction to biometric recognition, *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1), 4-20.
- Kucukyilmaz, Tayfun, B., Cambazoglu, Cevdet Aykanat, & Can, Fazli. (2008). Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing & Management*, 44(4), 1448-1466.
- Love, H. (2002). *Attributing authorship: an introduction*. Cambridge University Press, 15.
- Moore, Trevor, & Gurpreet Dhillon. (2000). Software piracy: A view from Hong Kong. *Communications of the ACM*, 43(12), 88-93.
- Orebaugh, A. (2006). An Instant Messaging Intrusion Detection System Framework: Using character frequency analysis for authorship identification and validation. Carnahan Conferences Security Technology, Proceedings 2006 40<sup>th</sup> Annual IEEE International. IEEE, 160-172.
- Orebaugh, A., & Allnut, J. (2009). Identifying and characterizing instant messaging authors for cyber forensics. *IATAC Magazine*, 12(3), 20-22.
- Orebaugh, A., & Allnut, J. (2010). Data mining instant messaging communications to perform author identification for cybercrime investigations. *Digital Forensics and Cyber Crime*, 99-110.
- Rodrigues, Ricardo N., Lee Luan Ling, & Govindaraju, Venu. (2009). Robustness of multimodal biometric fusion methods against spoof attacks. *Journal of Visual Languages & Computing*, 20(3), 169-179.
- Teng, Gui-Fa, Lai, Mao-Sheng, Ma, Jian-Bin, & Li, Ying. (2004). E-mail authorship mining based on SVM for computer forensic. *Machine Learning and Cybernetics, 2004*. Proceedings of 2004 International Conference, 2, IEEE, 1204-1207.
- Zheng, Rong, Li, Jiexun, Chen, Hsinchun, & Huang, Zan. (2006). A framework for authorship identification of online messages: Writing style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3), 378-393.



