IJoFCS (2009) 1, 22-28

e-forensic PRESS

The International Journal of FORENSIC COMPUTER SCIENCE

www.IJoFCS.org

Classification of Instant Messaging Communications for Forensics Analysis

Angela Orebaugh[1] , and Jeremy Allnutt[2]

(1) George Mason University, USA, angela_orebaugh@yahoo.com
(2) George Mason University, USA, jallnutt@ece.gmu.edu

Abstract - Instant messaging (IM) is a well-established means of fast and effective communication. Once used primarily by home users for personal communications, IM solutions are now being deployed by organizations to provide convenient internal communication. This often includes the exchange and discussion of proprietary and sensitive information, thus introducing privacy concerns. Although IM is used in many legitimate activities for conversations and message exchange, it can also be misused by various means. For example, an attacker may masquerade as another user by hijacking the connection, performing a man-in-the-middle attack, or by obtaining physical access to a user's computer. There are various reasons that an attacker might want to masquerade as someone else, including spying, disgruntlement, snooping, or other malicious intentions. Analysis of IM in terms of computer forensics and intrusion detection has gone largely unexplored until now. This paper explores IM author classification based on author behavior. Author classification may be used for author identification/validation for forensics analysis or masquerade detection. The experiments presented here applied classification methods to IM messages to determine whether the author of an IM conversation could be identified based strictly on user behavior, and to determine the strongest identifying characteristics.

1. Introduction

Author identification, also called authorship attribution, is the task of determining the author of a piece of work. All humans have unique patterns of behavior, much like the uniqueness of biometric data. Therefore, certain characteristics pertaining to language, composition, and writing, such as particular syntactic and structural layout traits, patterns of vocabulary usage, unusual language usage, and stylistic traits, should remain relatively constant. The identification and learning of these characteristics with a sufficiently high accuracy is the principal challenge in author identification.

This paper describes the methods, analysis, and results of using data mining classification to perform author identification on instant messaging (IM) communications for the purpose of computer forensics analysis. The classification methods presented here profiled author behavior using various linguistics patterns and characteristics. The principle objectives were to create a set of characteristics that remained relatively constant for a large

number of messages written by a single author, and to classify these messages as belonging to a particular author.

The goals of this research were to answer the following questions:

- *Can we identify an author of an IM conversation based strictly on author behavior?*
- *What behavior characteristics are the strongest classifiers?*

This research investigated the foundational techniques necessary to incorporate IM author identification into computer forensics investigations (as digital evidence) and intrusion detection technologies (such as masquerade detection).

2. Related Work

The area of IM communications has been largely unexplored thus far. There has been significant research in identifying authors of text, such as Shakespeare's works and the Federalists papers, as well as in email and virus authorship identification. There has also been some research in online user behavior. However, the results of these research areas have not been applied to IM. Some important works related to IM communications research include the following:

- *Can Online Behavior Unveil Deceivers?* by L. Zhow and D. Zhang. This research explores online deception behavior in a group IM setting.
- *Extracting Social Networks from Instant Messaging Populations*, by J. Resig, S. Dawara, C. M. Homan, and A. Teredesai. This research investigates IM communities as social networks.
- *A Framework for Mining Instant Messaging Services*, by J. Resig and A. Teredesai. This research explores various data mining issues and how they relate to IM and counter-terrorism efforts.
- *Applying Authorship Analysis in Cybercrime Investigation*, by R. Zheng, Y. Qin, Z. Huang, and H. Chen. This research presents an authorship analysis approach for identity tracing in cybercrime investigations.
- *Multi-Topic E-mail Authorship Attribution Forensics*, by O. de Vel, A. Anderson, M. Corney and G. Mohay. This research investigates the forensics authorship identification or categorization of multi-topic e-mail documents.
- *Language and Gender Author Cohort Analysis of E-mail for Computer Forensics*, by O. de Vel, A. Anderson, M. Corney, and G. Mohay. This research investigates the gender and language background of an author, based on cohort attribution mining from e-mail text documents.
- *Gender-Preferential Text Mining of E-mail Discourse*, by O. de Vel, A. Anderson, M. Corney, and G. Mohay. This study provides additional research on gender identification, based on user text analysis.

The first three works are related to IM, but are not focused on author identification. The last four works focus on applying user behavior and pattern analysis techniques to email, but do not consider IM. It is a natural extension to apply the techniques used for email, forensics, and other purposes to IM author profiling and identification.

3. IM Architecture

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. Users connect to the central server with the client and begin adding and conversing with other network users, commonly designated as "buddies" or "friends". Buddies are maintained in a Buddy List, which shows when users are logged on for communication. Popular IM service provider networks include AOL, Yahoo,

Table 1
Instant Messaging Author Behavior Categories

Stylometric Features
Character frequency distribution (upper/lowercase, numbers, and special characters)
Word frequency distribution
Emoticon frequency distribution
Function word frequency distribution
Short word frequency distribution
Punctuation frequency distribution
Average word length
Average words per sentence
Contains a greeting
Contains a farewell
Abbreviation frequency distribution
Spelling errors
Grammatical errors

MSN, and Google. Each of these networks provides a compatible communication client. Some clients, such as Trillian and Gaim, can connect to multiple provider networks at once. Most client products allow logging of IM conversations. The IM conversations are logged in a simple text format, making it easy to parse and analyze conversation data. This paper used IM conversation logs for author identification and validation.

4. Author Behavior Categorization

Author behavior categorization uses a set of characteristics that remain relatively constant for a large number of IM messages written by an author. These characteristics, known as *stylometric features*, include syntactic and structural layout traits, patterns of vocabulary usage, unusual language usage, and stylistic features. Each author has various stylometric features that are sufficient to uniquely identify him or her. Stylometric features are often word-based, including word and character frequency distributions, word length, and sentence length. Literary analysts and computational linguists often use frequency lists. Various syntactic features are also included, such as the use of function words (short all-purpose words such as "the" and "to"), punctuation, greetings and farewells, and emoticons. Users also use abbreviations for common phrases such as LOL (laughing out loud) and ROTFL (rolling on the floor laughing), as well as shortened spellings of

Table 2
Pre-defined Attributes

Category	Attribute
Special characters	. , ! ? @ # $ % ^ & * - _ + = ‘ \
Emoticons	:-) :) :-(:(;-) ;) :-P :P ;-P ;P :-D :D :’-(:’(:* :-*
Abbreviations	R U K C RU 2 4 BRB LOL BTW JK L8R LMAO NP IDK OMG TTYL TTYS WTF FYI ASAP IC CU OIC PLS PLZ CYA ROTFL THX IDC OTP U2 YT IMHO …
Sentence Structure	Average words per sentence

words such as ru (are you) and 4 (for). Table 1 shows the stylometric features that may be collected and analyzed for IM author classification.

5. Data Description

The experiments used IM conversation logs collected by the Gaim and Adium clients. The IM conversations were logged to ASCII text files in the following format:

[timestamp] [user name:] [message]

For example:

(14:19:29) User1: hey, what time is the meeting today?

(14:19:35) User2: It is at 11AM…are you going?

(14:19:39) User1: yeah, I will be there, it sounds very interesting! :) :)

The data required a series of preprocessing steps. First, the raw IM logs were parsed to extract data for each user. The data were prepared for analysis by removing all entries that did not belong to the user under analysis (UserX), as well as removing both the timestamp and username. Thus, an example of a formatted log for User 1 looks like the following:

hey, what time is the meeting today?

yeah, I will be there, it sounds very interesting! :) :)

Next, the logs were split into 2500 character segments to create instances. Sometimes this was a complete single conversation log, and sometimes it involved combining smaller conversation logs to meet the required length. Finally, the instances were processed to generate frequency totals for each attribute of a behavior category for the author. The frequency total data was outputted in CSV format and formatted as a Weka data file.

The data used in this research consisted of IM conversation logs for four users categorized in the following classes: User1, User2, User3, and User4. The data was parsed to calculate the sentence structure and the frequencies of predefined special characters, emoticons, and abbreviations, resulting in a total of 69 numeric at-

Table 3
IM Data Classification Results

J48	Overall Accuracy: **97.86%**	Error: **2.14%**	
	TP	**FP**	
User1	.97	.01	
User2	1	.019	
User3	.97	0	
User4	.97	0	

```
   a    b    c    d     <--
classified as
 34    1    0    0  |   a = User1
  0   35    0    0  |   b = User2
  0    1   34    0  |   c = User3
  1    0    0   34  |   d = User4
```

IBk	Overall Accuracy: **97.14%**	Error: **2.86%**	
	TP	**FP**	
User1	.97	0	
User2	.97	.029	
User3	.94	.01	
User4	1	0	

```
   a    b    c    d     <--
classified as
 34    1    0    0  |   a = User1
  0   34    1    0  |   b = User2
  0    2   33    0  |   c = User3
  0    0    0   35  |   d = User4
```

Naïve Bayes	Overall Accuracy: **99.29%**	Error: **0.71%**	
	TP	**FP**	
User1	.1	.01	
User2	1	0	
User3	1	0	
User4	.97	0	

```
   a    b    c    d     <--
classified as
 35    0    0    0  |   a = User1
  0   35    0    0  |   b = User2
  0    0   35    0  |   c = User3
  1    0    0   34  |   d = User4
```

Table 4
Classification Accuracy Results for Attribute Categories

Classification Method	Special Characters	Emoticons	Abbreviations
J48	91.43%	55.71%	95.71%
IBk	92.86%	50%	89.29%
Naïve Bayes	92.14%	56.42%	97.85%

tributes. Table 2 shows the 1 sentence structure attribute, 17 special characters, 16 emoticons, and 35 abbreviations defined in this study.

The conversation log data was parsed to create 35 instances for each class, for a total of 140 instances. The original data was unbalanced, since some users had more frequent and longer conversations than others. Therefore, the data was undersampled to the size of the smallest dataset to create a balanced dataset, resulting in 35 instances of 2500 characters per class.

6. Classification Methods and Results

The experiments used various classification methods to determine if an author of an IM conversation could be identified based on his or her sentence structure and use of special characters, emoticons, and abbreviations. The experiments also determined which features were strongest at identifying authors. The research used the Weka data mining tool for classification.

The classification methods used for this research were the J48 decision tree, IBk nearest neighbor, and Naïve Bayes classifiers. Table 3 shows the accuracy, error, true positive (TP), and false positive (FP) rate of each classifier when applied to the dataset.

Next, individual attribute categories were tested. Table 4 shows the accuracy results for each classifier when applied to each individual attribute category.

Attribute selection was used to rank the strongest identifying attributes. Table 5 shows the strongest identifying attributes according to the information gain and chi-squared techniques.

Table 5
Attribute Selection

Information Gain	Chi-squared
U	U
…	…
-	-
.	.
,	‘

Next, the top 3 individual attributes (U, three dots, and the hyphen) were tested individually with each classifier. Table 6 shows the accuracy results for each classifier when applied to each individual attribute.

Table 6
Classification Results for Top Attributes

Classi-fication Method	U	…	-
J48	62.86%	67.86%	61.43%
IBk	66.43%	67.86%	61.43%
Naïve Bayes	65.71%	68.57%	62.86%

Experiments using classification methods on the IM datasets resulted in the following conclusions:

- Abbreviations were the best discriminators with 97.85% accuracy, followed closely by special characters with 92.86% accuracy.

- The Naïve Bayes classification method performed with only the abbreviations attributes (97.85%) resulted in a higher accuracy than the J48 classifier (97.86%) and with a similar accuracy as the IBk classifier (97.14%) when all attributes combined were used in J48 and IBk.
- The strongest identifying attributes were U, three dots, and the hypen.
- None of the individual attributes identified by attribute selection were strong enough to determine author identification with a high degree of accuracy.
- The combination of all attribute categories using the Naïve Bayes classification method provided the best results (99.29% accuracy).

7. Summary and Future Work

The IM market has seen explosive growth, with millions of users participating in online conversations. However, little has been explored in terms of the research and analysis of the network, messages, user behavior, and data mining of these systems. There are several concerns involving the use of IM systems, including whether the user is really communicating with the intended buddy or friend. The threats include account hijacking, man-in-the-middle attacks, and masquerading. There are various reasons someone would wish to masquerade as someone else including spying, disgruntlement, snooping, and other malicious intentions.

This paper presented the use of data mining of IM communications for authorship identification. Classification methods were used to identify IM authors based on various behaviors. Human behavior presents challenges for analysis. For example, such behavior has an extremely wide "normal" range and can be very unpredictable: abnormal activities are sometimes perfectly normal, and all people change. The results of the experiments here indicate that Naïve Bayes classification is highly accurate (> 99% accuracy) at predicting the author of an IM conversation based on behavior. The experiments also identified the behavior characteristics that are the strongest classifiers. The data showed that users tend to exhibit the same characteristics throughout various conversations. Furthermore, users exhibited different characteristics from each other, much like the uniqueness of biometric data.

Based on these preliminary experiments, future research will involve:

- Increased numbers of users (classes) in the dataset.
- Increased numbers of attributes from the set of stylometric features, such as characters, function words, and structural layouts.
- Varied numbers of characters that are included in an instance, to determine the minimum size necessary for high accuracy and a low false-positive rate.

The author behavior attributes used in these experiments comprise only a subset of the stylometric features that may be used for IM author identification. Other stylometric measures, as shown in Table 1, must also be used to create an accurate, well-rounded user profile. A broad attribute set and larger number of classes should provide a comprehensive analysis of the IM data for highly accurate authorship identification.

References

[1] Corney, Malcolm. Personal Website. http://sky.fit.qut.edu.au/~corneym.

[2] Donaldson, Tom. "Computational Analysis of Verbal Behavior". http://behavioranalysis.mactom.com/cavb.html.

[3] Goldring, Tom. "Authenticating Users by Profiling Behavior". http://www.cs.fit.edu/~pkc/dmsec03/slides/goldring03dmsec.ppt.

[4] O. de Vel, A. Anderson, M. Corney and G. Mohay, "Email Authorship Attribution for Computer Forensics", in Daniel Barbara, Sushil Jajodia, "Applications of Data Mining in Computer Security", ISBN 1-4020-7054-3, Kluwer Academic Publishers, Boston, 2002, 252 pages.

[5] O. de Vel, A. Anderson, M. Corney and G. Mohay. "Gender-Preferential Text Mining of E-mail Discourse".18th Annual Computer Security Applications Conference (2002 ACSAC) December 9 – 14, 2002, Las Vegas, Nevada, USA.

[6] O. de Vel, A. Anderson, M. Corney and G. Mohay. "Language and Gender Author Cohort Analysis of E-mail for Computer Forensics". Digital Forensic Research Workshop, August 7 – 9, 2002, Syracuse, NY, USA.

[7] O. de Vel, A. Anderson, M. Corney & G. Mohay, "Mining Email Content for Author Identification Forensics", SIGMOD Record Web Edition, 2001, 30(4).

[8] O. de Vel, A. Anderson, M. Corney & G. Mohay, "Multi-Topic E-mail Authorship Attribution Forensics", ACM Conference on Computer Security - Workshop on Data Mining for Security Applications, November 8, 2001, Philadelphia, PA, USA.

[9] Resig, John. Ankur Teredesai. "A Framework for Mining Instant Messaging Services". SIAM DM 2004 Workshop on Link Analysis, Counter-Terrorism & Privacy.

[10] Resig, John. Santosh Dawara, Christopher M. Homan, and Ankur Teredesai. "Extracting Social Networks from Instant Messaging Populations". KDD 04 Link Discovery Workshop (LinkKDD 2004).

[11] Rong Zheng, Yi Qin, Zan Huang, Hsinchun Chen. "Applying Authorship Analysis in Cybercrime Investigation". Lecture Notes in Computer Science. Publisher: Springer-Verlag GmbH. ISSN: 0302-9743. Volume 2665 / 2003. Chapter: pp. 59 – 73.

[12] Szymanski, Boleslaw. "Recursive Data Mining for Masquerade Detection and Author Identification". 5th Annual IEEE Information Assurance Workshop. June 2004. http://www.cs.rpi.edu/~szymansk/papers/ia04.pdf.

[13] Zhou, Lina, Zhang Dongsong. "Can Online Behavior Unveil Deceivers?", Proceedings of the 37th Hawaii International Conference on System Sciences. 2004.

Angela Orebaugh is a cyber security technologist leading a variety of security innovation projects including research for the National Institute of Standards and Technology. She has 15 years experience in information technology and security and is the author of several technical security books. Ms. Orebaugh is an adjunct professor for George Mason University where she is completing her PhD with a focus on digital forensics and cybercrime.

Dr. Jeremy Allnutt is a Professor in the Department of Electrical and Computer Engineering as well as the Director of the Masters in Telecommunications Program at George Mason University. He has recently created the new Masters in Computer Forensics program at GMU that prepares students for careers in industry, government, and academia by combining academic education with real-world practical techniques.