

# Towards the Development of a Cyber Analysis & Advisement Tool (CAAT) for Mitigating De-Anonymization Attacks

Siobahn C. Day, Henry Williams, Joseph Shelton, Gerry Dozier

Department of Computer Science, North Carolina A&T State University, Greensboro, U.S.A

Center for Advanced Studies in Identity Science

{seday, hcwillia, jasheltl} @aggies.ncat.edu , {gvdozier} @ncat.edu

## Abstract

We are seeing a rise in the number of Anonymous Social Networks (ASN) that claim to provide a sense of user anonymity. However, what many users of ASNs do not know that a person can be identified by their writing style.

In this paper, we provide an overview of a number of author concealment techniques, their impact on the semantic meaning of an author's original text, and introduce AuthorCAAT, an application for mitigating de-anonymization attacks. Our results show that iterative paraphrasing performs the best in terms of author concealment and performs well with respect to Latent Semantic Analysis.

## Introduction

Anonymous Social Networks (ASN) can provide users with a false sense of anonymity; however, research in the area of Author Identification (Attribution) has shown that users can be identified simply by their writing style (Stamatatos 2009). Narayanan et al. (2012), introduces the concept of a de-anonymization attack where hackers apply sophisticated Author Identification techniques (AITs) in an effort to uncover the identity of an author of a text. Once this occurs the hackers can track a victim across the web and even through other ASNs.

Recently researchers, M. Brennan, Afroz, and Greenstadt (2012); Kacmarcik and Gamon (2006); Rao and Rohatgi (2000), have developed a number of techniques for author concealment. These techniques as well as their ability to conceal one's writing style are as follows: adversarial stylometry, iterative language translation and iterative paraphrasing.

Presently there exist two forms of adversarial stylometry (Afroz, Brennan, and Greenstadt 2012; M. Brennan et al. 2012; M. R. Brennan and Greenstadt 2009). The first form, obfuscation, is when an author tries not to write like them-

selves while the second form, imitation, is when an author tries to 'mimic' the writing style of another author. Research shows that both of these techniques are effective in concealing one's writing style. In the case of disguising one's writing style, M. Brennan et al. (2012) demonstrate that obfuscation and imitation are easy on the short term but more difficult to maintain on the long term. In Section IV, it will be shown how AuthorCAAT can be used to provide authors with the ability to perform long-term adversarial stylometry.

Another form of author concealment is Iterative Language Translation (ILT) (Mack, Bowers, Williams, Dozier, and Shelton 2015). ILT is where an original text is translated to another language and then back to its original language. This technique was first presented in Rao and Rohatgi (2000), where the authors describe this approach as being "somewhat facetious" and "drastic." They believed that this approach would change the meaning of a message thus making it an impractical approach. It was also mentioned by Kacmarcik and Gamon (2006), that this approach could be a good starting point for someone looking to "scramble" their words. ILT is effective in concealing the writing style of an author; however, it is vulnerable to fingerprinting, (Caliskan and Greenstadt 2012). If one knows the language used in translating the text, one can then recover the original writing style of the author.

The last form of author concealment is Iterative Paraphrasing (IP). The use of IP was originally mentioned in Kacmarcik and Gamon (2006). In IP, one will take the original text and use a paraphrasing tool to convert it into a paraphrased text. Concerning IP, to the authors' knowledge, no one has as of yet analyzed its effectiveness in author concealment, semantics, and its vulnerability to fingerprinting (this will be discussed in Section III).

The remainder of the paper will be as follows. In Section II, we discuss our experiments. In Section III, we discuss our results. In Section IV, we provide a brief discus-

sion of AuthorCAAT. In Section V, we provide our conclusion and future work.

## Author Concealment & Fingerprinting Experiments

### Our Dataset

The datasets we used for our experiments were gathered from blogs written by 100 different authors. For every author in our dataset, there are 4 instances. Those instances in the dataset are as follows: the first instance served as the probe and the remaining 3 instances served as the gallery. This results in 100 instances in the probe set and 300 instances in the gallery set.

### Our Translators & Paraphrasers

Our ILT dataset, used Google translation tools for English to Spanish, Spanish to English, English to Chinese, and Chinese to English. The ILT text was prepared in iterations. We consider an iteration to be a full round trip cycle of translation (e.g. English-Spanish-English and English-Chinese-English). Therefore, Iteration 1 would be E-X-E, Iteration 2 would be E-X-E-X-E, and Iteration 3 would be E-X-E-X-E-X-E, where E stands for English and X  $\in$  {Spanish, Chinese}. Therefore, a total of six ILT datasets were developed consisting of 300 gallery instances of the 100 authors.

Our IP dataset was created using an online tool known as Plagiarisma. The Iterations for IP are similar to ILT. Combining ILT with IP we have  $X \in$  {Spanish, Chinese, Paraphraser}. Therefore, three IP datasets were developed consisting of 300 gallery instances of the 100 authors. For ILT/IP, there were a total of nine datasets.

### Experiment I: Author Concealment via ILT/IP

For Experiment I, the feature extractor used in Mack, Bowers, Williams, Dozier, and Shelton (2015), referred to as the Hybrid-II Author Identification System (AIS), was applied to the instances of the nine datasets (and the probe set) to create feature vectors where each feature vector consisted of 1282 features. The Hybrid-II AIS, is composed of 95 features from the Unigram feature extractor (Forsyth 1997), 170 stylometric features from De Vel, Anderson, Corney, and Mohay (2001) feature extractor, as well as 256 features in the form of function words and 761 features that come from the Stanford Parser in the form of Parts-of-Speech parent child pairs for a total of 1282 features.

In Experiment I, the baseline performance was the author recognition rate of the 100 authors (English only) using no ILT/IP iterations. While, the ILT/IP experiments

were used to determine how well ILT/IP reduces the author recognition rate with respect to the baseline.

### Experiment II: Fingerprinting the Translators and the Paraphrasers

For Experiment II, a tool known as JGAAP, Java Graphical Author Attribution Program, (Juola, Sofko, and Brennan 2006) was used to fingerprint the translators and the paraphraser. This tool allows for text analysis using various stylometry and textometry techniques. We used the first 100 authors from each ILT/IP Iteration using the first gallery instance as the ‘unknown’ author and the remaining two instances from the gallery as the ‘known’ authors. The ‘known’ authors were labeled by languages and/or paraphraser. This was used for all three Iterations of ILT/IP. The analysis was processed by using WEKA SMO, with the results ordered with event culling from most to least. Character N Grams, where  $n=2$ , was used as the event driver.

### Experiment III: Fingerprinting the Number of Iterations Used to Conceal an Author’s Writing Style

In Experiment III, the ‘unknown’ authors were chosen from the first gallery instances of all Iterations of ILT/IP. The ‘known’ authors were chosen from the remaining two instances of the gallery and were labeled by the number of ILT/IP Iterations that were applied. The same settings as Experiment II were used with respect to the event driver, analysis, and event culling.

## Results

### Results of Experiment I

The results of Experiment I, Author Concealment via ILT/IP, are shown in Figure 1. Figure 1 shows the affect that ILT/IP has on the accuracy of the AIS. In Figure 1, the x-axis represents the iteration number (Iteration 1, Iteration 2, Iteration 3) and the y-axis represents the accuracy of the AIS.

In Figure 1, the accuracy of the AIS is 54% percent. In the first iteration of ILT/IP, the author identification rates drop. At Iteration 1, ILT-Spanish has the best performance in terms of reducing the AIS rate to 6%, followed by IP at 7% and ILT-Chinese at 10%. In the second iteration, IP has the best performance in reducing the AIS rate to 1%, followed by ILT-Chinese at 11% and ILT-Spanish at 6%. At Iteration 3, IP continues to outperform ILT. At Iteration 3, IP reduces the AIS rate to 6%, followed by ILT-Spanish at 7% and ILT-Chinese at 11%. These results show the effectiveness of ILT/IP in concealing an authors identity.

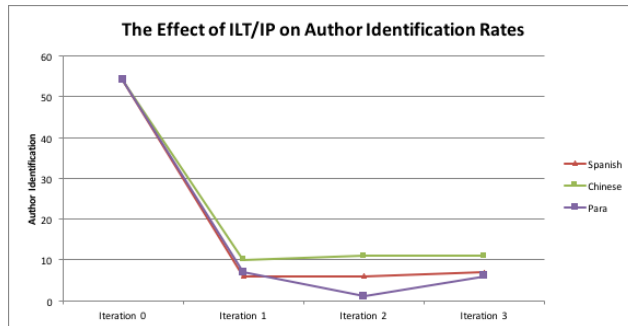


Figure 1: A Comparison of the Effectiveness of ILT/IP on Reducing Author Recognition Rates

Prior research suggests, (Caliskan and Greenstadt 2012; Kacmarcik and Gamon 2006; Rao and Rohatgi 2000), that ILT/IP is naïve as well as problematic due to the resulting text being unable to retain its original meaning. In order to address this issue, we applied Latent Semantic Analysis (LSA) on all iterations of the dataset.

Latent Semantic Analysis (LSA) “...is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text” (Landauer, Foltz, and Laham 1998). Using a LSA tool developed by the University of Colorado Boulder, we compared our original text with the resulting text of ILT/IP.

In the Table 1, the results of using the LSA tool on our dataset are shown. Given two samples of text, the LSA tool will provide an output of 1 if the semantics of the two text samples are exact and -1 if the semantics of the two text samples do not match at all. Given the output of the LSA tool on our dataset, we ran an ANOVA test as well as a t-test to break the performances of ILT/IP into equivalence classes as shown in Table 1.

In Table 1, the first column represents the ILT/IP method used, the second column represents the average output of the LSA tool with the standard deviation in parenthesis, and the third column, labeled EC, represents the equivalence class. The equivalence classes are ordered from best to worst in terms of performance. The equivalent classes were determined by applying ANOVA and a t-test to check for statistical significance. The p-value used for the ANOVA test was 0.05.

The results displayed in Table 1, show that the resulting text from ILT-Spanish is closest to the semantics of the original text with an output of 0.862 followed by IP at 0.802 and ILT-Chinese at 0.773. This indicates that ILT/IP is not only non-problematic but effective at preserving the semantics of the original text.

Table 1: LSA Results from Comparing the Original Text with Resulting Text from ILT/IP

ILT/IP Method	LSA Results	EC
Spanish	0.862 (0.11)	1
Paraphraser	0.802 (0.09)	2
Chinese	0.773.16)	3

## Results of Experiment II

The results of Experiment II, Fingerprinting the Translators and the Paraphrasers, are shown in Figure 2. In Figure 2, the x-axis shows the iterations (Iteration 1, Iteration 2, Iteration 3) and on the y-axis it shows the accuracy in determining the ILT/IP method used. In Figure 2, one can see as the number of iterations increases so does the accuracy for each ILT/IP method that is being used.

In Figure 2, at Iteration 1, ILT-Spanish has the best fingerprinting accuracy at 93%, followed by ILT-Chinese at 90%, and IP at 86%. In Iteration 2, ILT-Spanish leads at 98% followed by ILT-Chinese 97%, and IP at 91%. In Iteration 3, ILT-Chinese comes in at 99%, followed by ILT-Spanish at 98%, and IP at 95%. The results not only show that the translators can be accurately fingerprinted, but they also show that of the three IP is hardest to fingerprint but only at the first iteration. On the other hand, these results show that the translator and paraphrasers are able to be identified which can potentially allow for reversibility or the uncovering of the original text, thus revealing an authors writing style.

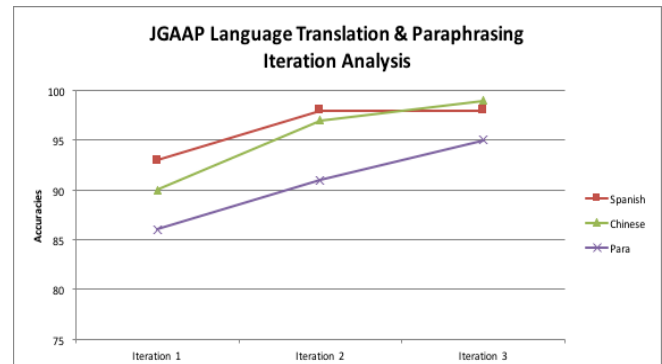


Figure 2: A Fingerprinting Analysis of ILT/IP over 3 Iterations

## Results of Experiment III

The results of Experiment III, Fingerprinting the Number of Iterations Used to Conceal an Author’s Writing Style, are shown in Figure 3. In Figure 3, the x-axis shows the iterations (Iteration 1, Iteration 2, Iteration 3) and the y-axis shows the accuracy of an iteration of ILT/IP in being fingerprinted. Figure 3 shows determining which Iteration

of ILT/IP of a given text proves to be more difficult; however, the accuracy rises over iterations.

In Figure 3, at Iteration 1, ILT-Spanish leads at 70%, followed by ILT-Chinese at 61%, and IP at 47%. At Iteration 2, IP performs best at 31%, followed by ILT-Spanish at 18%, and ILT-Chinese at 15%. At Iteration 3, ILT-Chinese is the best performer at 60%, followed by ILT-Spanish at 53 % and IP at 49% making it the worst performer. The results show that fingerprinting ILT/IP by iteration is harder to fingerprint but not impossible. Thus allowing an original text and author to be revealed.

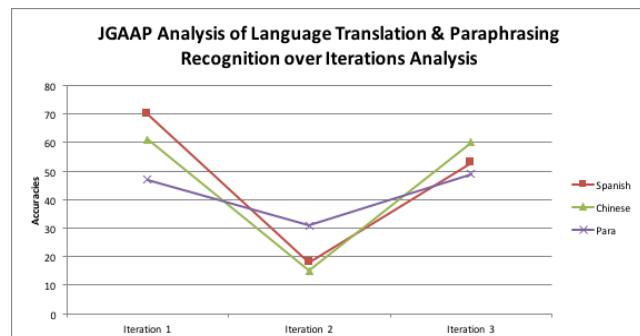


Figure 3: A Fingerprinting Analysis of the Number of Iterations of ILT/IP over 3 Iterations

## DISCUSSION: THE DEVELOPMENT OF AUTHORCAAT

The results presented earlier show that translators and paraphraser can be fingerprinted. Even the iterations can be fingerprinted. In order to conceal one's identity in an efficient and effective way, the authors' believe that a system must be developed that will allow a user to use all of the author concealment methods mentioned in this paper simultaneously while authoring a text. The Center for Advanced Studies in Identity Sciences (CASIS) has developed such a system for author concealment known as AuthorCAAT (Author Cyber Analysis & Advisement Tool).

Figure 5 provides a screenshot of AuthorCAAT. AuthorCAAT has a window that allows an author to type in text. As the author types, their writing style is analyzed. The feature vector associated with their writing style is shown just below the window. To the right of the window, is a pane that displays the author samples that match the sample written within the window based on a user specified by the slide bar. For example, if the slide bar is at '10' this means that the pane will display the authors whose writing samples are within the closest 10% to the author sample that was typed in the window.

Below the Matches to, pane is a drop-down box that will allow an author to translate what is currently in the window in either Spanish, Chinese, or Paraphrase and back to English. Once a language or paraphraser has been selected, the user (author) presses the 'Translate' button to execute one

cycle of ILT on the text currently within the author window.

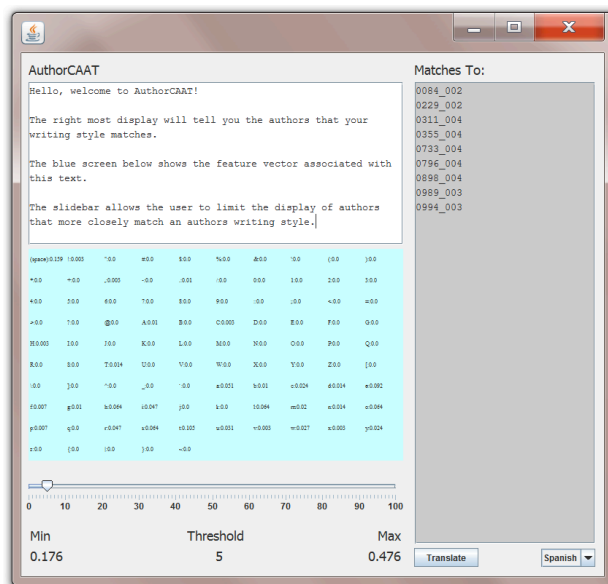


Figure 5: AuthorCAAT

In Figure 5, one can see that AuthorCAAT allows a user to perform both forms of Adversarial Stylometry. If the user sees that their writing style is detected and shown in the pane, then they can choose to re-write their text in such a way that it is not shown in the pane. A user can also monitor the pane in an effort to perform imitation authorship. As long as a particular author ID is shown in the pane (while their author ID is not in the pane) then they are writing like that particular author.

Finally, AuthorCAAT allows for ILT/IP at the sentence level. For example, an author can type in the first sentence and apply ILT/IP to that sentence. After this, the author can add a second sentence and then apply ILT/IP to both sentences in the window and/or edit the resulting sentences further (Adversarial Stylometry).

## Conclusions and Future Work

In this paper, ILT/IP dramatically reduces the author recognition rate. Secondly, translators and paraphraser are good enough to preserve the semantics. This is based on our results from our LSA table. Thirdly that not only can language translators be fingerprinted but we can fingerprint paraphraser too. Lastly we show that the iteration of a particular ILT/IP can be fingerprinted as well. This all leads to a development tool, AuthorCAAT that can do all of things at the sentence level. This will allow fingerprinting to be more difficult. Our Future work will include increasing our dataset from 100 to 1000 to see if the finger-

printing becomes more accurate with more authors in terms of ILT/IP. We suspect the accuracy of fingerprinting iterations at Iteration 1 and 2 will increase with the number of authors analyzed. This is a contrast to what was stated in Caliskan and Greenstadt (2012).

### Acknowledgments

This research is based upon work supported by the United States Government including the National Science Foundation. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

### References

- Afroz, S., Brennan, M., & Greenstadt, R. (2012, May). Detecting hoaxes, frauds, and deception in writing style online. In *Security and Privacy (SP), 2012 IEEE Symposium on* (pp. 461-475). IEEE.
- Brennan, M. R., & Greenstadt, R. (2009, July). Practical Attacks Against Authorship Recognition Techniques. In *IAAI*.
- Brennan, M., Afroz, S., & Greenstadt, R. (2012). Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, 15(3), 12.
- Caliskan, A., & Greenstadt, R. (2012, September). Translate once, translate twice, translate thrice and attribute: Identifying authors and machine translation tools in translated text. In *Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on* (pp. 121-125). IEEE.
- De Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4), 55-64.
- Forsyth, R. S. (1997). Short substrings as document discriminators: An empirical study. In *ACH-ALLC* (Vol. 97).
- Free Online Plagiarism Checker for Students, Teachers, Scholars, Educators, Scientists, Essayists, Writers. Free TurnItIn and Copy-scape Alternative. (n.d.). Retrieved February 02, 2016, from <http://plagiarisma.net/>
- Google Translate. (n.d.). Retrieved February 04, 2016, from <https://translate.google.com/>
- Juola, P., Sofko, J., & Brennan, P. (2006). A prototype for authorship attribution studies. *Literary and Linguistic Computing*, 21(2), 169-178.
- Kacmarcik, G., & Gamon, M. (2006). Obfuscating document stylometry to preserve author anonymity. Paper presented at the Proceedings of the COLING/ACL on Main conference poster sessions.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.
- LSA @ CU Boulder. (n.d.). Retrieved February 02, 2016, from <http://lsa.colorado.edu/>
- Narayanan, A., Paskov, H., Gong, N. Z., Bethencourt, J., Stefanov, E., Shin, E. C. R., & Song, D. (2012, May). On the feasibility of internet-scale author identification. In *Security and Privacy (SP), 2012 IEEE Symposium on* (pp. 300-314). IEEE.
- Nathan Mack, Jasmine Bowers, Henry Williams, Gerry Dozier, and Joseph Shelton, "The Best Way to a Strong Defense is a Strong Offense: Mitigating Deanonimization Attacks via Iterative Language Translation," *International Journal of Machine Learning and Computing* vol.5, no. 5, pp. 409-413, 2015.
- Rao, J. R., & Rohatgi, P. (2000). Can pseudonymity really guarantee privacy? Paper presented at the USENIX Security Symposium.
- Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3), 538-556.