Anonymouth Revamped:
Getting Closer to Stylometric Anonymity

Andrew W.E. McDonald, Jeffrey Ulman, Marc Barrowclift, and Rachel Greenstadt

Drexel University, Philadelphia, PA
{awm32, jmu26, meb388, greenie}@cs.drexel.edu

Stylometry, the study of writing style—such as word choice, sentence length, and sentence structure—is a very real threat to privacy. Even if a would–be author is completely anonymous in every respect, as soon as he/she encodes any thoughts in text, his/her anonymity may vanish. Today, a document's author can be selected from a pool of a hundred thousand authors [1]. Further, even when it is unknown whether a document's author is in a select pool of potential authors, methods exist to determine the likelihood that a given document was written by any author in the aforementioned pool.[1] The accuracy and ability [2] of stylometric authorship detection/attribution in open and closed world scenarios is rapidly advancing. It is therefore exceedingly important to continue the evolution of tools that combat this potential privacy breach, and offer privacy seeking individuals the ability to remain anonymous while expressing their ideas via a textual medium.

While it is possible for one to anonymize a document with nothing more than a text editor, studies show that it is quite challenging to do well, that there is no guarantee that the author has done a sufficient job of removing his/her style from the document, and that it is hard to be consistent in hiding one's own writing style [3,4]. We present a revised version of the open source, Java-based, authorship anonymization tool, *Anonymouth*, presented at the 2012 Privacy Enhancing Technologies Symposium [5]. The revised Anonymouth has a fully redesigned graphic user interface to enhance usability, along with an increased repertoire and updated algorithms to improve performance.

Anonymouth uses JStylo [5] —an authorship attribution platform—as its backend, and uses machine learning and natural language processing techniques to attempt to aid a user in removing his/her style from a document he/she authored. To do this, the user must first input three sets of documents: the document to be anonymized, `documentToAnonymize`; previous documents authored by the user, `userSampleDocuments`; and a set comprised of documents by at least three other authors, `otherSampleDocuments`. Once all documents have been loaded into Anonymouth, JStylo extracts features from all documents, and classifies the `documentToAnonymize` with respect to the set formed by combining the `userSampleDocuments` and the `otherSampleDocuments` (the `userAndOtherDocuments`), using one of the available machine learning algorithms (though the SMO is almost exclusively used). This classification is shown to the user to provide an idea of the document's baseline anonymity. Anonymouth then analyzes the features extracted from all three document sets, and based upon the ak-means clustering algorithm[2], and the user's average feature values, decides upon a few sets of potential target values for the `documentToAnonymize`'s features. Each set of potential target values is tested against the classifier trained on the set of `userAndOtherDocuments`. The set of potential target values that returns a classification suggesting the greatest degree of anonymity is selected as the set of target values for the user's `documentToAnonymize`. Next, alternative ways of expressing ideas, as well as elements to add and remove from the document are presented to the user. Once the user is satisfied with his/her edits, he/she may either reprocess the document. This edit and reprocess cycle continues until the user feels satisfied by the classification returned by JStylo.

Initially, Anonymouth presented the actual extracted features (such as character grams) to the user, and "suggested" that the number of occurrences of the selected feature be either increased or decreased. It was quickly apparent that these "suggestions" made Anonymouth completely unusable. In addition, the user interface was unintuitive and poor at relaying necessary information; as gleaned from the surveys we conducted during the original Anonymouth's user study. Further yet, while the core of Anonymouth appeared

[1] This is work currently underway in our lab.

[2] A slightly modified version of k-means in which the number of means varies as needed to find suitable clusters [5].

to perform well, the clustering and cluster selection algorithms had little to do with the classifier being used—so there was no guarantee that if a document's features moved to the locations specified by the target values, it would be anonymous as far as the classifier was concerned.

The alterations we have made to Anonymouth aim to solve or mitigate the aforementioned usability and reliability issues. The first change made was to bring features together into words to remove and words to add. However, simply providing users with two lists of words does little to help them to change the ways they are expressing their ideas—which is important. To this end, we have added two-way translations[3] to Anonymouth's repertoire. Two-way translations are applied to each sentence using 15 different languages.[4] The "translations" (the new English versions), like the list of words to add/remove, are sorted by Anonymity Index (AI)—a summation across features of the product of the number of occurrences of a feature, the feature's information gain, and the amount the feature must be changed (the distance it is from its target value), for all features in a word/sentence/document—to determine which words/sentences will contribute to the document's anonymity. The user is presented with the sorted list of alternative sentences, and is able to swap a translated version of a sentence in for the original version; with the idea being that it is generally feasible to find a "translation" that alters the way a given idea is expressed. After the swap, the user is free to correct any errors the "translation" may have introduced. To address the clustering and cluster selection reliability, while we still cluster the features (which is good due to pseudo–randomness of results), target values are chosen after testing the "best" possible groups of target values against the classifier that provides the results—so, if the `documentToAnonymize`'s features are moved to the target values, then in theory the altered `documentToAnonymize` would return the same classification as the target values originally did.[5] The user interface itself is now much more user friendly as well. In addition to a "General Suggestions" window that gives an overview of how to approach anonymizing a document, the initialization has been drastically simplified. Users now also have an "Anonymity Bar" which allows them to gauge their overall progress via a thermometer–like anonymity indicator, and may view their classification results in a bar graph.

To determine whether or not we are progressing in the right direction, we plan on conducting a second user study focusing on Anonymouth's performance with regard to its user interface / usability. We will evaluate Anonymouth based upon our own observation of test subjects while they use Anonymouth, survey results from the users after having used Anonymouth, and the degree to which subjects were successful in anonymizing their documents with Anonymouth. In addition to that, we hope to get feedback from the PETs community regarding Anonymouth's current usability, and on ways to further improve it.

In the future, we plan on looking into ways to make Anonymouth more automated in order to increase usability, however this is not an easy task due to the inherently hard natural language processing issues that accompany an automated rewriter. Initial investigation suggests that MIT's ConceptNet 5 [6] may be helpful in this respect. Further, we are also interested in increasing Anonymouth's versatility in terms of scenarios/situations that it may be useful in, in order to make the tool more robust and to enable as many people as possible to reap the privacy enhancing benefits of Anonymouth.

References

1. Narayanan, A., Paskov, H., Gong, N., Bethencourt, J., Stefanov, E., Shin, R., Song, D.: On the feasibility of internet-scale author identification. In: Proceedings of the 33rd conference on IEEE Symposium on Security and Privacy, IEEE (2012)

[3] English → Other Language → English. In our original paper [5] we stated that translations alone wouldn't work. While it is highly unlikely that they will anonymize a document on their own, this may not be the case when combined with Anonymouth's anonymization system.

[4] Presently, using our test documents, we utilize an online translation service. However, this is simply used as a proof–of–concept. If it turns out that we want to package Anonymouth with translation functionality, we will incorporate an offline translator.

[5] This will almost certainly not be the case though, because while editing the document, it is inevitable that features will be unavoidably changed away from their target values. Still, we feel that this is a good approximation for the degree of anonymity that will be achieved by setting a given set of target values.

2. Abbasi, A., Chen, H.: Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Trans. Inf. Syst. **26**(2) (2008) 1–29
3. Brennan, M., Greenstadt, R.: Practical attacks against authorship recognition techniques. In: Proceedings of the Twenty-First Innovative Applications of Artificial Intelligence Conference. (2009)
4. Afroz, S., Brennan, M., Greenstadt, R.: Detecting hoaxes, frauds, and deception in writing style online. In: Proceedings of the 33rd conference on IEEE Symposium on Security and Privacy, IEEE (2012)
5. McDonald, A.W.E., Afroz, S., Caliskan, A., Stolerman, A., Greenstadt, R.: Use fewer instances of the letter i: Toward writing style anonymization. In Fischer-Hbner, S., Wright, M., eds.: Privacy Enhancing Technologies. Volume 7384 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2012) 299–318
6. Speer, R., Havasi, C.: Representing general relational knowledge in conceptnet 5. In Chair), N.C.C., Choukri, K., Declerck, T., Do?an, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., eds.: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, European Language Resources Association (ELRA) (may 2012)