The Best Way to a Strong Defense is a Strong Offense: Mitigating Deanonymization Attacks via Iterative Language Translation

Nathan Mack, Jasmine Bowers, Henry Williams, Gerry Dozier, and Joseph Shelton

***Abstract*—In Bowers, *et al.*, a technique was presented, referred to as Iterative Language Translation (ILT), for reducing the threat of deanonymization attacks via two well-known author identification systems (AISs). In this paper, we introduce four additional 'stronger' AISs, which outperform the AISs evaluated in Bowers, *et al.* Our results show that ILT still remains an effective technique for reducing author identification accuracy even if stronger AISs are used.**

***Index Terms*—Author identification, feature extraction, feature selection, steady state genetic algorithm (SSGA).**

I. Introduction

Narayanan *et al.* introduces the concept of a deanonymization attack [1]. A deanonymization attack occurs when a hacker gains access to "seemingly" anonymous text of an author and by using an author identification system (AIS) is able to identify the author based on their writing characteristics [1]-[4]. Narayanan *et al.*'s work was only a proof of concept; however, the authors state that they anticipated that hackers would have and be in the process of developing more sophisticated, stronger AISs that would be able to identify authors with greater accuracy. Bowers *et al*. presented an approach for preserving the anonymity of an author by iteratively applying language translation [5].

Bowers *et al.* demonstrates the effectiveness of using iterative language translation (ILT) as a means of concealing one's writing style [5]. The authors applied ILT whereby they translated English text into a foreign language (e.g. Spanish, Chinese, and Arabic) and then back into English iteratively for one, two, and three iterations. Bowers *et al.* then compared the effectiveness of the ability of ILT to conceal authorship through the use of two well-known AISs [6], [7].

Although ILT was shown to be successful at reducing the identification accuracy of the two AISs, the AISs themselves were relatively weak. In this paper, we develop four 'stronger' AISs in an effort to observe the impact that ILT has on preserving author anonymity. Our rationale is that the best way to develop a strong defense (in terms of anonymity preservation) is to develop a strong offense (in the form of an AIS).

Manuscript received February 13, 2015; revised April 23, 2015. This research was funded by Science & Technology Center: Bio/Computational Evolution in Action Consortium (BEACON).

The authors are with the North Carolina A&T State University, USA (e-mail: namack@aggies.ncat.edu, jdbowers@aggies.ncat.edu, hcwillia@aggies.ncat.edu, gvdozier@ncat.edu, jashelt1@aggies.ncat.edu).

The remainder of this paper is as follows. Section II describes the two AISs used in Bowers *et al.* as well as two additional baseline AISs. In Section III, the concept Genetic & Evolutionary Feature Selection (GEFeS) is introduced. GEFeS is applied to the four AISs introduced in Section II in an effort to produce four 'stronger' AISs. Section IV presents our experiments and Section V presents our results. In Section VI, we present our conclusions and future work.

II. Four Baseline AISs

In this section, we introduce four baseline AISs. These AISs are as follows: Uni-Gram, O. de Vel *et al.*, a hybrid which combines the feature sets of Uni-Gram and O. de Vel *et al.*, which we refer to as Hybrid-I, and an AIS that is very similar to the one proposed by Narayanan et al., which combines a large number of author identification features including the features used in Hybrid-I [1], [2], [6], [7]. We referred to this baseline AIS as Hybrid-II.

A. The Uni-Gram AIS

The Uni-Gram AIS presented in this paper was also used by Forsyth [6]. This AIS utilizes 95 features that include letters, numbers, special characters, spaces, etc. The Uni-Gram AIS is based on character frequency. It counts the number of occurrences of each of the 95 features within an author sample. The number of occurrences is then divided by the total number of characters within the sample. This normalized set of character frequencies forms a feature vector (FV) representing an author's writing style. FVs can then be compared with other FVs through the use of a wide variety of distance metrics. The closer two FVs are to one another the more likely their associated text samples are from the same author. Fig. 1 provides a sample of the Uni-Gram AIS feature set.

B. The O. de Vel et al. AIS

O. de Vel *et al.* proposed a stylometric-based AIS. This AIS contains 170 stylometric features [7]. These features can be described as the characteristics associated with the writing style of a particular author, such as, vocabulary richness and average word length. The 170 stylometric features are shown in Fig. 2. As with the Uni-Gram AIS, the closer two FVs are to one another the more likely their associated text samples are from the same author.

C. The Hybrid-I AIS

The Hybrid-I AIS is simply the combination of the features from the Uni-Gram AIS and O. de Vel *et al.* AIS [6], [7]. This AIS uses a total of 265 features.

DOI: 10.7763/IJMLC.2015.V5.543

D. The Hybrid-II AIS

The Hybrid-II AIS is similar to the AIS proposed by Narayanan [1]. This AIS includes the 265 features from the Uni-Gram, O. de Vel *et al*. AISs (Hybrid-I AIS) as well as 256 extra features in the form of function words and an additional 761 features that come from part-of-speech (POS) parent-child pairs of parse trees created by the Stanford Parser [6]-[8]. An example of a Stanford Parser parse tree is shown in Fig. 3. Using the parse tree, the Hybrid-II AIS calculates the frequency of each part-of-speech (POS) parent-child pair in a sample text. In total, the Hybrid-II AIS uses 1282 features.

(space)	!	“	#	$	%	&	‘	(	)
*	+	,	-	.	/	0	1	2	3
4	5	6	7	8	9	:	;	<	=
>	?	@	A	B	C	D	E	F	G
H	I	J	K	L	M	N	O	P	Q
R	S	T	U	V	W	X	Y	Z	[
\	]	^	_	`	a	b	c	d	e
f	g	h	i	j	k	l	m	n	o
p	q	r	s	t	u	v	w	x	y
z	{	|	}	~					

Fig. 1. The subset of unicode characters used in the uni-gram AIS by R. S. Forsyth [6].

Stylometric Features
Number of blank lines/total number of lines
Average sentence length
Average word length(number of characters)
Vocabulary richness i.e., V/M
Total number of function words/M
Function words (122)
Total number of short words/M
Count of hapax legomena/M
Count of hapax legomena/V
Number of characters in words/C
Number of alphabetic characters in words/C
Number of upper-case chars/C
Number of digit characters in words/C
Number of white space characters/C
Number of space characters/C
Number of space characters/white space characters
Number of tab spaces/C
Number of tabs spaces/number of white spaces
Number of punctuations/C
Word length frequency distribution/M (30)

Fig. 2. The stylometric features proposed by O. de Vel *et al.* [7].

III. GEFeS

Genetic and Evolutionary Feature Selection (GEFeS) is feature selection technique that is based on simulated evolution [9]-[18]. GEFeS is used to evolve feature masks (FMs) in an effort to discover high-performing sub-feature sets. The FMs are used to ‘mask out’ non-salient features of the FVs that are extracted by the four baseline AISs.

The evolutionary process of GEFeS is as follows. Initially, a random population of FMs is created. FMs are represented as a string of real values between 0 and 1 and the lengths of these FMs are equivalent to the lengths of the FVs. If a FM value is less than 0.5, then the corresponding FV value is masked out; otherwise, the FV value is used. Each FM is then evaluated on a sub-dataset (training and/or validation) of blog samples, represented as FVs, to determine its fitness. The fitness evaluation function is ten times the number of FVs incorrectly classified plus the percentage of the features used.

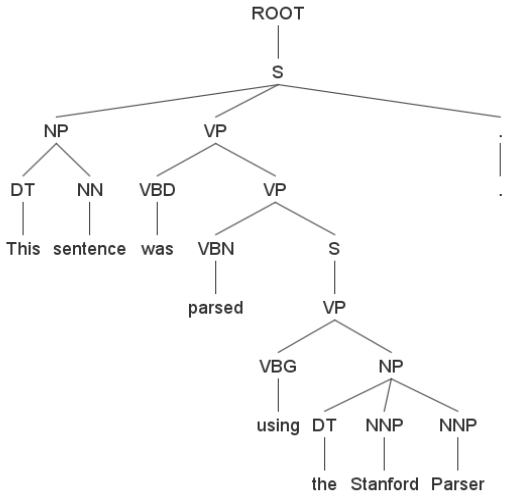


Fig. 3. An example of a parse tree created by the Stanford Parser [8].

To classify FVs, a dataset is split into a probe set and a gallery set. The probe set consists of one FV from each subject and the gallery set consists of the remaining FVs. After a FM has been applied on all FVs, each probe FV is compared to the gallery FVs using the Manhattan distance metric (shown in Equation 1). The equation takes two FVs, f_i and f_j, and determines the sum of the absolute value of the difference of feature, y, of each FV. The feature y iterates from 0 to the length of the FM, l.

Once an initial population is generated, two parent FMs are selected from the population via binary tournament selection [16]. Binary tournament selection works by randomly selecting two FMs from the population and the better fit FM is selected to be a parent. This process is repeated to select the second parent. After the two parents have been selected, they are used to create an offspring FM. The offspring FM is created via Uniform Crossover [16]. Gaussian mutation is then applied to the offspring FM [16]. Next, the worst fit FM in the population is replaced with the offspring FM. This process of selecting parent FMs, creating offspring FMs and replacement the worst fit FM in the population is repeated until a user-specified stopping condition has been met. Fig. 4 provides an example of the GEFeS evolutionary process.

$$\text{distance}_{\text{Manhattan}}(f_i, f_j) = \sum_{y=0}^{l-1} |f_{i,y} - f_{j,y}| \; fm_y \quad (1)$$

IV. Experiments

In our experiments, we used a dataset that consisted of blog text from 1000 respective authors. The samples collected were from a wide variety of online blog sites. Each of the 1000 samples was partitioned into 4 sub-samples with each

sub-sample containing 2 paragraphs. Each paragraph contained between 8 and 10 sentences. For each author, the first sub-sample was placed in a probe set while the last three sub-samples of that particular author were placed in a gallery set.

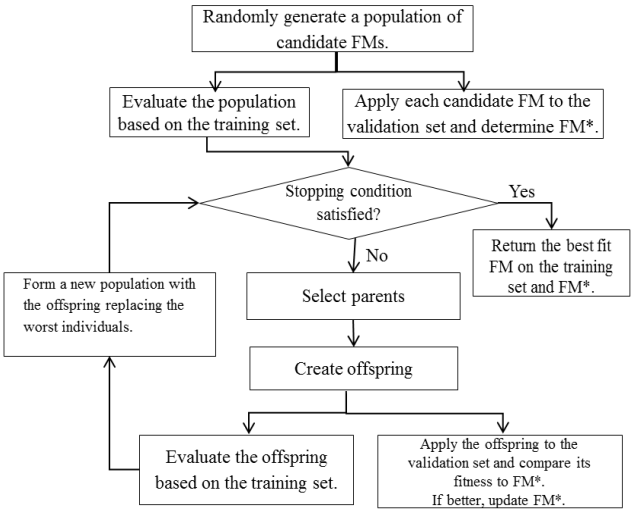


Fig. 4. The GEFeS evolutionary process.

A. Experiment I: English to English

In this experiment, the dataset described above was further split into three subsets: a training set consisting of the text associated with the first 334 authors, a validation set consisting of the text associated the next 333 authors, and a test set consisting of the text associated with the final 333 authors. This experiment is referred to as 'English to English' (denoted by E-E) because for each author the associated probe and gallery instances are the original text of the author. This experiment will be used to determine the relative strength of the eight AISs based on their author identification accuracy.

B. Experiment II: Iterative Language Translation

In this experiment, all of the gallery instances were translated into Spanish, Chinese, or Arabic and then translated back into English. The process was repeated from 1 to 3 iterations and resulted in the following datasets: E-ESE, E-ECE, E-EAE, E-ESESE, E-ECECE, E-EAEAE, E-ESESESE, E-ECECECE, and E-EAEAEAE. These datasets were used to determine the effectiveness of ILT in concealing an author's identity.

V. Results

A. Results of Experiment I: English to English

Each of the four baseline AISs were applied to the test set. Each of the four GEFeS-based AISs (denoted as Baseline+GEFeS) trained on the training set for a total of 2000 function evaluations (FEs). This was repeated for a total 5 runs. The validation set was used for cross-validation in an effort to reduce overfitting [16]. GEFeS was an instance of a Steady-State Genetic Algorithm implemented in X-TOOLSS [15], [19]. GEFeS evolved a population of 20 FMs. The initial population was biased so that 70% of the features for each candidate FM were turned on.

Table I provides the performance results of the eight AISs. The first column represents the type of AIS, Baseline or Baseline+GEFeS. The last four columns represent the four variants of AIS. The performances across the Baseline row of Table I are the baseline performances for each of the AISs. For each cell of the Baseline+GEFeS row, the top number represents the performance of the best FM evolved over the 30 runs of GEFeS while the number in the parentheses represents the average performance of the best FM evolved on each run.

In Table I, one can see that the baseline performances are weak in terms of identification accuracy. The best performing baseline AIS is the Uni-Gram AIS. This result is interesting because the AIS proposed by Narayanan *et al.* had an accuracy of approximately 20% on their dataset. This could be due to the fact that, in Narayanan *et al.* each sample consisted of at least 8 paragraphs (at least 7,500 characters) while each sample in our dataset consisted of just 2 paragraphs (of 8 to 10 sentences) [1]. Also our dataset consisted of one blog post per author while the dataset in Narayanan *et al.* used an average of 24 blog posts [1].

In Table I, one can see that the application of feature selection via GEFeS dramatically increases author recognition accuracy across all baseline AIS variants. Hybrid-II+GEFeS has the best overall performance. Fig. 5 and Fig. 6 provide a ROC curve and Log-Log curve for the performances of Hybrid-II baseline and Hybrid-II+GEFeS. For both curves, one can see that Hybrid-II+GEFeS has better performance.

TABLE I: A Comparison of the Performance of GEFeS with the Perfomance of the (Baseline) Uni-Gram AIS, (Baseline) O. de Vel AIS, (Baseline) Hybrid-I AIS, (Baseline) Hybrid-II AIS, Uni-Gram+GEFeS, O. de Vel+GEFeS, Hybrid-I+GEFeS, and Hybrid-II+GEFeS for the English to English Experiment

E-E				
AIS	**Uni-Gram**	**O. de Vel**	**Hybrid-I**	**Hybrid-II**
Baseline	12.01%	4.50%	6.31%	5.71%
Baseline+GEFeS	20.72% (18.93%)	21.62% (16.98%)	25.23% (21.53%)	51.65% (47.35%)

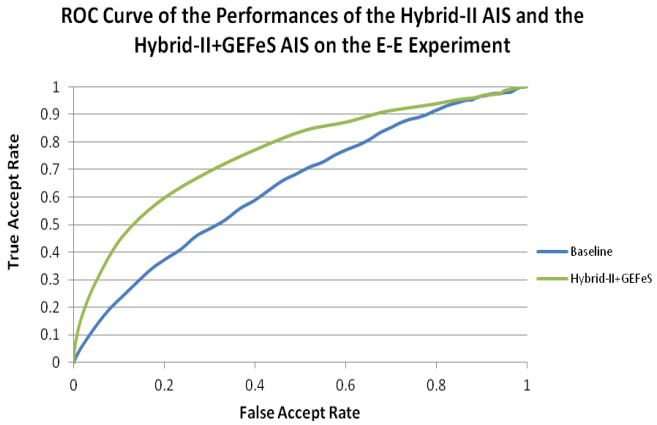


Fig. 5. The ROC curve of the performances of the hybrid-II AIS and the hybrid-II+GEFeS AIS on the English-to-English experiment.

B. Results of Experiment II: The Application of ILT

For the results presented in this section, the best feature masks of the Baseline+GEFeS variants were taken and applied to the test sets whose gallery samples were iteratively translated into a foreign language and then back into English.

Fig. 7-Fig. 10 show the effect that ILT has on the four stronger AISs namely: Uni-Gram+GEFeS, O. de Vel+GEFeS, Hybrid-I+GEFeS, and Hybrid-II+GEFeS. In each of the Figs., the *y*-axis represents author identification rate while the *x*-axis represents the number iterations of language translation that was applied.

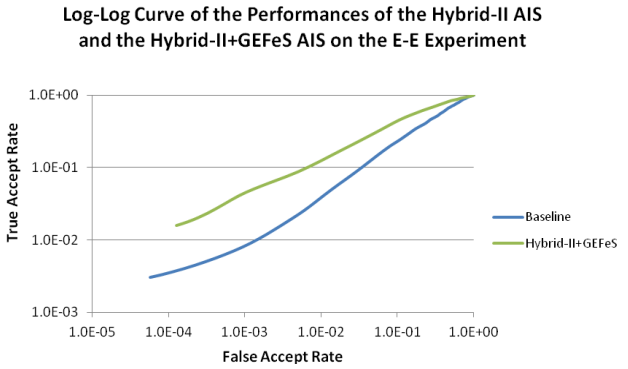


Fig. 6. The log-log curve of the performances of the hybrid-II AIS and the hybrid-II+GEFeS AIS on the English-to-English experiment.

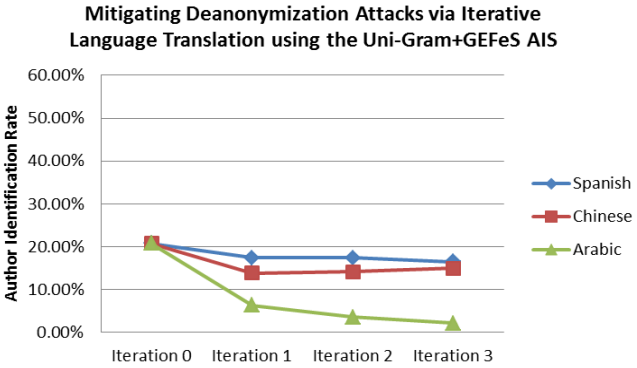


Fig. 7. The effect of iterated language translation on the best GEFeS feature mask applied to feature vectors extracted via the uni-gram+GEFeS AIS.

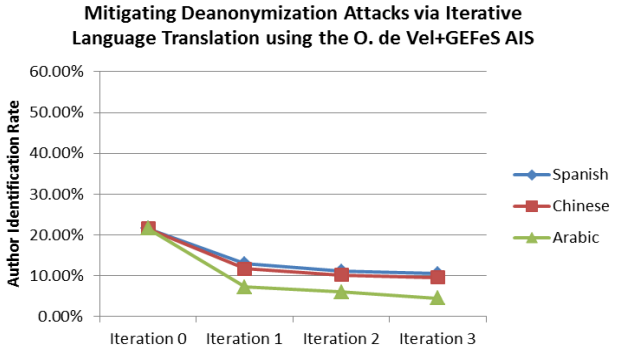


Fig. 8. The effect of iterated language translation on the best GEFeS feature mask applied to feature vectors extracted via the O. de vel + GEFeS AIS.

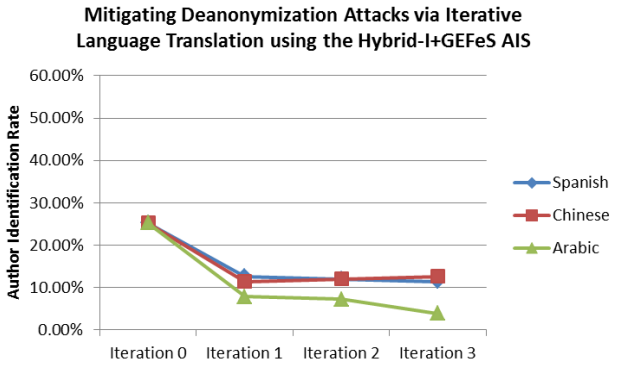


Fig. 9. The effect of iterated language translation on the best GEFeS feature mask applied to feature vectors extracted via the hybrid-I+GEFeS AIS.

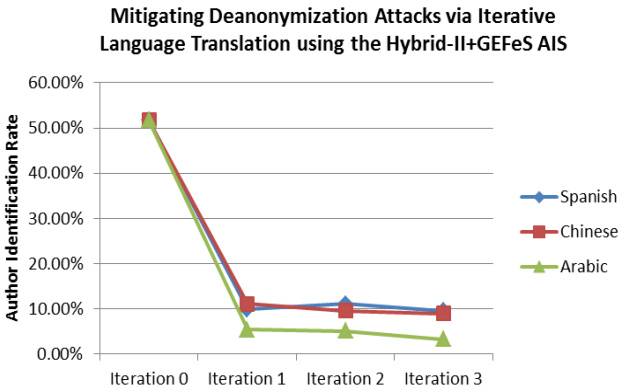


Fig. 10. The effect of iterated language translation on the best GEFeS feature mask applied to feature vectors extracted via the hybrid-II+GEFeS AIS.

In Fig. 7-Fig. 10, one can see that ILT dramatically reduces author identification accuracy. The results in the Figs also show that iteratively translating into Arabic and then back into English is the most effective means of concealing an author's identity. Spanish-based and Chinese-based ILT have very similar performances with Chinese-based ILT having a slightly better performance. Across Fig. 7- Fig. 10, one can see that a single iteration of ILT seems to be associated with the largest drop in author identification rate.

VI. Conclusion and Future Work

In this paper, we developed four 'stronger' AISs that incorporated GEFeS-based feature selection. The results show that GEFeS-based AISs outperforms their associated baseline AISs with Hybrid-II+GEFeS being the best performer. The results show that ILT is quite effective in concealing the identity of an author, despite using an AIS (in the form of Hybrid-II+GEFeS) that is at least 4 times stronger than the AIS (the Uni-Gram AIS). Our future work will be devoted towards: 1) developing stronger AISs and 2) developing better methods for protect the anonymity of an author.

Acknowledgements

This research was funded by Science & Technology Center: Bio/Computational Evolution in Action Consortium (BEACON).

References

[1] A. Narayanan, H. Paskov, N. Gong, J. Bethencourt, E. Stefanov, R. Shin, and D. Song, "On the feasibility of internet-scale author identification," in *Proc. 2012 IEEE Symposium on Security and Privacy*, May 2012, vol. 300, no. 314, pp. 20-23.

[2] R. Green and J. Sheppard, "Comparting frequency- and style-based features for twitter author identification," in *Proc. the International Florida Artificial Intelligence Research Society Conference*, 2013.

[3] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for Information Science and Technology*, pp. 538-556, 2009.

[4] H. Mohtasseb and A. Ahmed, "Mining online diaries for blogger identification," in *Proc. the International Conference of Data Minign and Knowledge Engineering — The World Congress on Engineering*, 2009.

[5] J. Bowers, H. Williams, G. Dozier, and R. Williams, "Mitigating deanonymization attacks via language translation for anonymous social networks," in *Proc. the 7th International Conference on Machine Learning and Computing*, 2015.

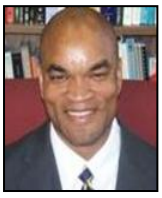
[6] R. Forsyth, "Short substrings as document discriminators," in *Proc. the Joint International Conference of the Association for Computers and the Humanities and the Association for Literary & Linguistic Computing*, Queen's University, Kingston, Ontario, 1997.
[7] O. de Vel, A. Anderson, M. Corney, and G. Mohay, "Mining e-mail content for author identification forensics," *ACM SIGMOD Record*, pp. 55-64, 2001.
[8] The Stanford Parser: A Statistical Parser. [Online]. Available: http://nlp.stanford.edu/software/lex-parser.shtml
[9] G. Dozier, K. Purrington, K. Popplewell, J. Shelton, K. Bryant, J. Adams, D. L. Woodard, and P. Miller, "GEFeS: Genetic & evolutionary featureselection for periocular biometric recognition," in *Proc. the 2011 IEEE Workshop on Computational Intelligence in Biometrics and Identity Management*, 2011.
[10] K. Popplewell, G. Dozier, K. Bryant, A. Alford, J. Adams, T. Abegaz, K. Purrington, and J. Shelton, "A comparison of genetic featureselection and weighting techniques for multi-biometric recognition," in *Proc. the 2011 ACM Southeast Conference*, 2011.
[11] T. Abegaz, G. Dozier, K. Bryant, J. Adams, J. Shelton, K. Ricanek, and D. Woodard, "SSGA and EDA based feature selection and weighting for face recognition," in *Proc. the 2011 IEEE Congress on Evolutionary Computation*, 2011.
[12] L. Smalls, J. Shelton, G. Dozier, K. Bryant, and J. Adams, "BiasedInitialized genetic & evolutionary feature selection for face recognition," in *Proc. the 2011 ADMI Conference*, 2011.
[13] G. Dozier, L. Simpson, J. Adams, D. L. Woodard, P. Miller, G. Glenn, and K. Bryant, "A comparison of two genetic and evolutionary featureselection strategies for periocular-based biometric recognition via X-TOOLSS," in *Proc. the* 2010 *International Conference Genetic and EvolutionaryMethods*, 2010.
[14] H. Williams, J. Carter, W. Campbell, K. Roy, and G. Dozier, "Genetic & evolutionary feature selection for author identification of HTML associatedwith malware," *International Journal of Machine Learning and Computing*, 2014, vol. 4, no. 3, pp. 250-255.
[15] J. Adams, H. Williams, J. Carter, and G. Dozier, "Genetic heuristic development: Feature selection for author identification," in *Proc. the 2013 Symposium Serieson Computational Intelligence*, 2013.
[16] A. Engelbrecht, *Computational Intelligence: An Introduction*, 2nd Edition, Wiley, 2007.
[17] D. Fogel, *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*, IEEE Press, 2000.
[18] L. Davis, *Handbook of Genetic Algorithms*, Van Nostrand, Reinhold, New York, 1991.
[19] X-TOOLSS. eXploration Toolset for the optimization of launch and space systems. [Online]. Available: http://nxt.ncat.edu/

Nathan Mack is a master of science student, majoring in computer science, at North Carolina Agricultural and Technical State University, Greensboro, NC, USA. Nathan received an associate of science degree in 2010 from Central Carolina Technical College, in Sumter, SC, USA. In 2013, he received a bachelor of science degree in computer science from Voorhees College, in Demark, SC, USA. His research interests are cyber security, cyber identity, and evolutionary computation.

Jasmine D. Bowers was born in Charlotte, NC, on August 4, 1991, received a bachelor degree of science in computer science and mathematics from Fort Valley State University in 2013. Jasmine is currently working towards obtaining her master's degree from N.C. A&T State University and is expected to graduate in May 2015. She has worked as a co-op with the Department of Defense and as a summer intern at Lawrence Livermore National Laboratory as a cyber defender. Her research topics include author identification and mitigating deanonimization attacks.

Henry C. Williams was born in Maryland on December 15, 1986, he received an associates of applied science in network technologies from Guilford Technical Community College in 2004. Henry then completed his bachelor's degree in computer science at North Carolina Agricultural and Technical State University (NCAT) in May 2014. Henry is currently working towards obtaining his masters degree from NCAT and is expected to graduate in December 2015.

He has worked as a supplemental instructor at NCAT and is currently working in the CASIS research group at NCAT in Greensboro, NC. He held internships at Carnegie Mellon University, as a research assistant in robotics, and at Lawrence Livermore National Labs, as a cyber defender. He is currently researching malware classification with computer learning.



Gerry Vernon Dozier is a professor and the chair of the Computer Science Department at North Carolina A&T State University. He is the director of the Center for Advanced Studies in Identity Sciences (CASIS), as well as the PI for the Center for Cyber Defense (recognized by the National Security Agency and the Department of Homeland Security as a Center for Academic Excellence in Information Assurance Education). During Gerry's tenure as a chair, the department has seen an increase in extramural funding and research publications as well as the establishment of a Ph.D. program. He has also lead in the development of an undergraduate research program where approximately 20% of the undergraduate students are active participants in funded research projects. Under Gerry's leadership, the NSF alliance for the advancement of African american researchers in computing (A4RC, www.a4rc.org) experienced a threefold increase (from 6 to 20) in the number of participating universities. A4RC was effective in increasing the number of African-American recipients of advanced degrees in computer science.

Gerry has published over 130 conference and journal publications. He has served as an associate editor of the IEEE Transactions on Evolutionary Computation and the International Journal of Automation & Soft Computing. His research interests include artificial & computational intelligence, genetic, evolutionary, and neural computing, biometrics, identity sciences, cyber identity, distributed constraint reasoning, artificial immune systems, machine learning and network intrusion detection. Gerry earned his Ph.D. from North Carolina State University.

Joseph Shelton is a doctoral student at North Carolina Agricultural and Technical State University, Greensboro, NC, USA, in the Computer Science Department Joseph obtained both his bachelor's degree and master's degree in computer science at North Carolina A&T State University in 2010 and 2012.

He is currently working as a research assistant at North Carolina A&T and has done so for the last three years. He has helped publish a book chapter in 'New Trends and Developments in Biometrics' titled 'Genetic and Evolutionary Biometrics' and has published over 20 articles in the field of biometrics and genetic & evolutionary computations. Joseph's research interests include biometrics, cyber security and evolutionary computation.

Mr. Joseph Shelton has received an award for 1st runner up for best student paper at the Conference on Systems Engineering Research (CSER 2012).