Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01 OCT 2014	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Deep Focus: Increasing User "Depth of Field" to Improve Threat Detection (Oxford workshop poster)	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) **William R. Claycomb**	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Software Engineering Institute Carnegie Mellon University Pittsburgh, PA 15213**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

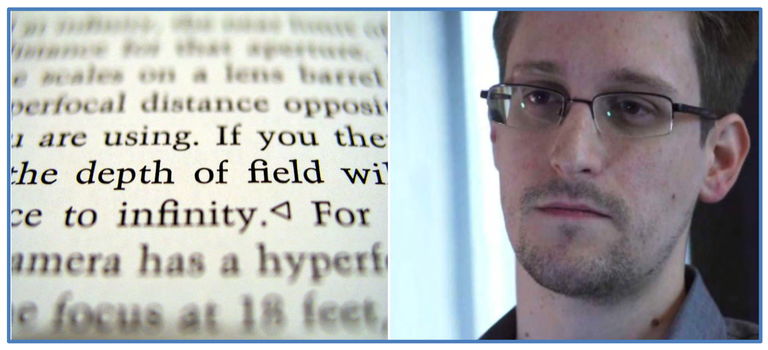
13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**SAR**	**1**	

Deep Focus: Increasing User "Depth of Field" to Improve Threat Detection



Carnegie Mellon/Software Engineering Institute		U. Mass at Amherst	Los Alamos Nat'l Labs	University of Oxford
William Claycomb, Ph.D.	Bronwyn Woods, Ph.D.	David Jensen, Ph.D.	Joshua Neil, Ph.D.	Sadie Creese, Ph.D.
Roy Maxion, Ph.D.	Brian Lindauer		Alex Kent	Phil Legg, Ph.D.
Jason Clark				

Project Goals

We believe insider threat detection methods can be improved by monitoring and analyzing features of user behavior not typically associated with indicators of malicious insider behavior. Anomalous behaviors and statistical outliers observed in such data sets may identify new indicators or help reduce high false positive rates associated with existing indicators.

We have three specific goals for this project:

1) Detect account masquerading by monitoring for sudden and significant changes in IT system interactions by the account user.
2) Develop unique profiles of individual users based on behavior on IT systems.
3) Baseline individual user behavior and monitor for changes that indicate potentially malicious insider behavior is likely to occur.

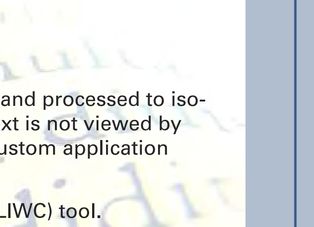
We intend to deliver the following outcomes:

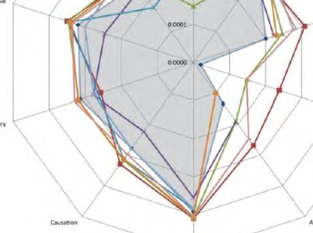
1) A measure of confidence that the person currently interacting with the IT system is or is not the authorized user.
2) Methods for collecting additional context of user behavior by which insider threat and anomaly detection engines can determine with higher confidence that suspicious behavior is likely to be malicious.
3) Visualization of these methods and metrics for analyst use.

The Problem

We're not looking for a needle in a **STACK** of hay…

we're looking for a **NEEDLE** in a **STACK OF NEEDLES**





Our Approach

Researchers have proposed numerous methods for detecting anomalous user behavior. We intend to focus on three methods in particular:

1) Classifier-Adjusted Density Estimation. This has been shown to be effective for anomaly detection in data with high dimensionality [Friedland 2014].
2) Latent Dirichlet Allocation (LDA). Robinson demonstrates a technique using the LDA model, borrowed from natural language processing, to identify malicious exfiltration events in a large data set of network header information [Robinson 2010].
3) Multivariate Statistic Analysis of linguistic characteristics of user text. Greitzer and Ferryman [Greitzer 2013] and Brown et al. [Brown 2013] demonstrate that statistical analysis using a variant of Chebyshev's inequality can identify outliers in a population of linguistic characteristics that correlate to persons with known psychological risk indicators.

Data

Host Audit: application use, removable media, file activity, keystrokes, registry entries, email activity, etc.

Network Activity: login events

Linguistic and Structural: word count, text structure, frequency of words by category, etc.

Population	Size	Known Insiders
Host Audit		
H_1	50,000+ hosts, multiple features	400+
H_2	20,000+ hosts, multiple features	100+
Network Activity		
N_1	8×10^8 login events	0
Linguistic and Structural		
L_1	Email text, 600 users	0
L_2	Transcripts of spoken words, 50+ speakers	0
L_3	Email text, 15,000+ users	>0
L_4	Email text, 20,000+ users	>0
L_5	Text messages, 50,000+ users	>0
L_6	Transcripts of spoken words, 40+ speakers	40+

Linguistic Patterns

Characteristics of a user's speech or writing can be measured both structurally and linguistically. Using these metrics, researchers have shown the feasibility of identifying anonymous authors [Narayanan 2012]. Others have observed measurable changes in linguistic patterns of known insiders [Taylor 2013].

Tools

- Email text and spoken words are extracted from source repositories and processed to isolate individual users. Identifying information is masked and actual text is not viewed by researchers. Raw data is parsed and prepared for analysis using a custom application written by research staff.
- Linguistic analysis performed by the Linguistic Inquiry Word Count (LIWC) tool.
- Structural characteristics obtained with custom application.

Initial Results

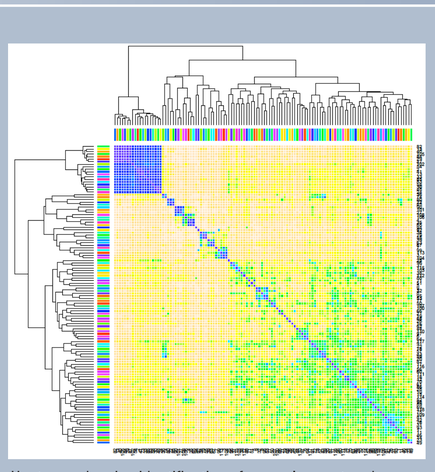
Comparing data sets: are linguistic features of the same user similar for typed vs. spoken words?

	Pronoun	1st Person Singular	1st Person Plural	2nd Person	Verbs	Swear Words	Achieve-ment	Leisure	Home	Money
Email avg	7.97%	2.58%	0.95%	1.51%	11.30%	0.00%	1.31%	0.70%	0.20%	0.52%
Spoken avg	11.87%	3.90%	1.34%	2.33%	15.40%	0.01%	1.85%	0.58%	0.29%	0.74%

	Social Words: All	Social Words: Family	Social Words: Friends	Social Words: Humans	Affect: All	Affect: Positive Emotions	Affect: Negative Emotions	Affect: Neg. Emotions: Anxiety	Affect: Neg. Emotions: Anger	Affect: Neg. Emotions: Sadness
Email avg	6.78%	0.07%	0.35%	0.16%	3.04%	2.39%	0.63%	0.08%	0.21%	0.07%
Spoken avg	9.75%	0.06%	0.05%	0.22%	5.25%	4.23%	0.99%	0.14%	0.29%	0.10%

Sample of comparison of email & webcast analysis from a user in Population L_1.

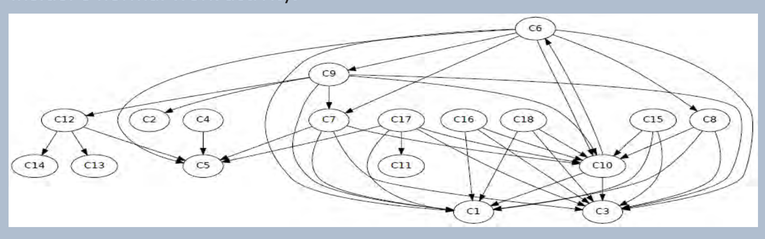
Logarithmic radar chart comparing linguistic features indicating neuroticism [Friedland 2013]



Heat map showing identification of anomalous network events [Robinson 2010]

Network Authentication Graphs

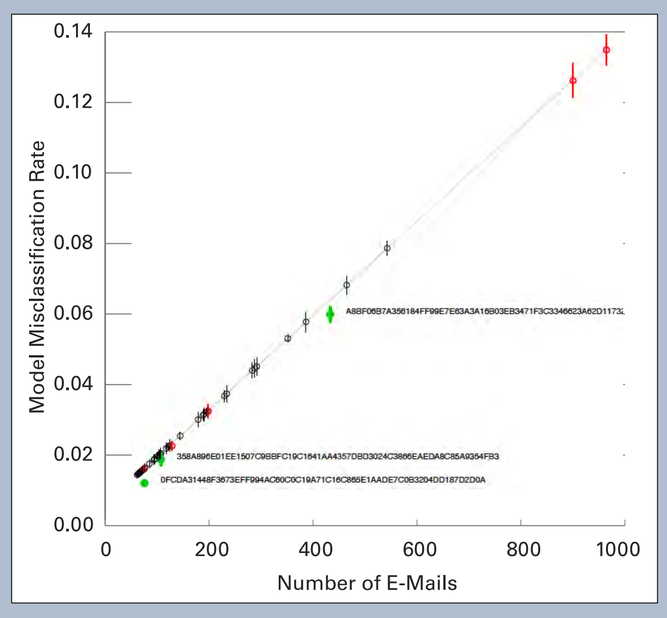
These directed graphs represent a user's authentication activity between networked computers over a predefined period. Research shows that administrative users generally have larger, more complex graphs than normal users [Kent 2013]. Furthermore, it is possible to profile each user's authentication activity, resulting in the ability to detect abnormal and potentially malicious activity. Empirical research on malicious insiders shows that many insiders engage in reconnaissance and information-gathering activities, accessing numerous network locations that often differ from the insider's normal work activity.



The figure to the left is a network authentication graph from a typical user with administrative access. This user accessed 18 computers with 41 authentication arcs. This a more complex authentication graph than those of general users [Kent 2013].

User Profiles

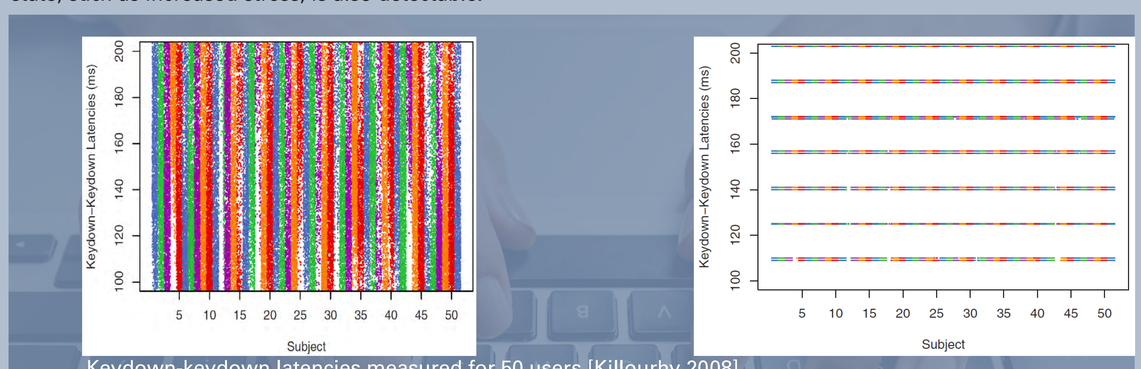
Can features unique to individual users be identified?



The figure to the left shows cross-validated error rates for sparse models selected using LASSO regularized logistic regression. The grey line shows the misclassification rate for the null model. The vertical bars give 1 standard deviation confidence intervals. The red points are recipients whose model had better than default error, but not significantly. The green points are the three recipients whose models were at least 1 standard deviation better than the null model.

Keystroke/Mouse Biometrics

Using metrics like keystroke latency or mouse dynamics, researchers have shown how individual users can be identified [Shen 2013]. Furthermore, evidence suggests changes in a user's personal state, such as increased stress, is also detectable.



Keydown-keydown latencies measured for 50 users [Killourhy 2008]

Visualization

Issues of concern must be visible and apparent to analysts. In cooperation with the Cyber Security Centre at Oxford, we will leverage an existing insider threat visualization toolkit to represent the data and anomalous activity we find in a clear and actionable manner.

References

1. [Brown 2013] Brown, C.; Watkins, A.; Greitzer, F. "Predicting Insider Threat Risks Through Linguistic Analysis of Electronic Communication." 2013 46th Hawaii International Conference on System Sciences (HICSS), 2013.
2. [Friedland 2014] Friedland, L.; Gentzel, A.; Jensen, D. "Classifier-Adjusted Density Estimation for Anomaly Detection and One-Class Classification." Proceedings of the 2014 SIAM International Conference on Data Mining (SDM).
3. [Taylor 2013] Taylor, P.; Dando, C.; Ormerod, T.; Ball, L.; Jenkins, M.; Sandham, A.; Menacere, T. "Detecting Insider Threats Through Language Change." Law and Human Behavior, Vol 37(4), Aug. 2013, 267-275.
4. [Kent 2013] Kent, A.; Liebrock, L.M. "Differentiating User Authentication Graphs." IEEE Security and Privacy Workshops, May 2013.
5. [Killourhy 2008] Killourhy, K.; Maxion, R. "The Effect of Clock Resolution on Keystroke Dynamics." In 11th International Symposium on Recent Advances in Intrusion Detection (RAID-08), 2008.
6. [Narayanan 2012] Narayanan, A.; Paskov, H.; Gong, N.Z.; Bethencourt, J.; Stefanov, E.; Shin, E.C.R.; Song, D. "On the Feasibility of Internet-Scale Author Identification." IEEE Symposium on Security and Privacy, May 2012.
7. [Robinson 2010] Robinson, D. "Statistical Language Analysis for Automatic Exfiltration Event Detection." Technical Report SAND2010-2179. Sandia National Laboratories. 2010.
8. [Shen 2013] Shen, C.; Cai, Z.; Guan, X.; Du, Y.; Maxion, R. "User Authentication Through Mouse Dynamics." IEEE Trans. on Info. Forensics and Security, Jan. 2013.
9. [Greitzer 2013] Greitzer, F.; Ferryman, T. "Methods and Metrics for Evaluating Analytic Insider Threat Tools." IEEE Security and Privacy Workshops: Workshop on Research for Insider Threat (WRIT), 2013.

Los Alamos NATIONAL LABORATORY EST. 1943 | UMASSCS SCHOOL OF COMPUTER SCIENCE | UNIVERSITY OF OXFORD | CYBER SECURITY CENTRE | ML MACHINE LEARNING DEPARTMENT | Software Engineering Institute Carnegie Mellon University

Contact: William (Bill) Claycomb, Ph.D., claycomb@cert.org