Security Project Proposal

Keith Dyer

March 26, 2015

Problem Statement

Sharing passwords and user information is a big security threat within organizations. This can occur intentionally by employees with no malicious intent in order to by pass organizational red-tape that impedes their work, or unintentionally through the use of stolen passwords via phishing.

Proposed Solutions

I am interested in seeing if either of the following two approaches could help organizations limit fraud and loss due to shared password information.

Data Mining Webcam Images

Recently, data mining algorithms have been shown to do a decent job of facial recognition. I think it is possible that while logged into an information sensitive system the computer's webcam can be used to capture an image periodically, and from that image the data mining algorithm can be run to compare the facial data of the current user to an established record of facial information for that user. If the system is successful a red flag can be raised internally by the company for investigation.

Authorship Analysis of Content Produced by the User

Not only do people differ in appearance when sharing user passwords, they also have different writing styles. I'd like to look at e-mails composed by users and use NLP techniques to determine if the system can detect whether or not the user is the standard, or an anomalous/intrusive user.

Data Plan

The data for facial images can be gathered from www.face-rec.org/databases. I propose building sample data sets where 80-90% of the samples are from a person under different conditions, and 10-20% of the samples are from a different user.

The data for authorship analysis can be gathered from the EnronSent corpus. This is available from http://verbs.colorado.edu/enronsent/ and contains the Enron Email data set that has been scrubbed to remove non-human generated text. It has been cleaned for better use in NLP projects according to the website. Similar to above I will build sub data sets for individual pseudo users with an 80-90% of the sentences from one user, and 10-20% from another user. Alternatively, a dataset can be composed of twitter messages to evaluate the users communications over an organizational instant messaging program rather than the email system.

Action Plan

For facial recognition the machine learning/data mining program should be able to classify accurately which person is being looked at from the skewed dataset. I plan to use 10-fold cross-validation repeated 10 times. I will use open source code for facial recognition feature extraction from the images and try different types of machine learning classifiers.

For authorship analysis I will try to determine if the user is anomalous or not based on the skewed data and the most recently produced inputs. For example, the first 80-90% of the training data will be used to build the language model, and then the program might have a higher error rate in predicting the next word when the 10-20% of the training data is input. I will explore other NLP techniques as well.

Related Work

While much work has been done on facial recognition and authorship analysis I wasn't able to find much in the way of applying these techniques to detecting anomalous users at an organization. This is what I was able to find.

Facial Recognition Work for Security

Database	Queries	Number of Results
DBLP	facial recognition anomaly detection	0
DBLP	facial recognition anomaly	0
DBLP	facial recognition security	10
DBLP	facial recognition fraud	0
DBLP	facial recognition password	0
DBLP	facial recognition user identification	0
ACMDL	facial recognition anomaly detection	54
ACMDL	facial recognition anomaly	68
ACMDL	facial recognition security	648
ACMDL	facial recognition fraud	68
ACMDL	facial recognition password	146
ACMDL	facial recognition user identification	824

While the ACMDL gave many more paper hits than the DBLP I found when sorting the results by relevance that many of them contained words in the query but were not very related to the proposed research topics. Here are a list of papers I found some what related:

1. Security Management for Mobile Devices by Face Recognition
2. Sensor-assisted facial recognition: an enhanced biometric authentication system for smartphones

Authorship Analysis Work for Security

Database	Queries	Number of Results
DBLP	authorship analysis security	1
DBLP	authorship analysis anomaly	0
DBLP	authorship analysis user identification	0
DBLP	authorship analysis intrusion	0
DBLP	authorship analysis password	0
DBLP	authorship analysis	76
ACMDL	authorship analysis security	503
ACMDL	authorship analysis anomaly	102
ACMDL	authorship analysis user identification	559
ACMDL	authorship analysis intrusion	57
ACMDL	authorship analysis password	59
ACMDL	authorship analysis	2095

There are many papers and articles written on identifying who wrote an article, email, or document but not many that attempt to identify an anomalous user based email messages. Here are related papers I found:

1. Authorship analysis in cybercrime investigation
2. From fingerprint to writeprint
3. Authorship Verification for Short Messages using Stylometry