Timing Attacks on Obfuscated User Generated Streams

Charles Poole, Sidafa Conde

Department of Mathematics,
University of Massachusetts Dartmouth

Abstract

Keyboarding text can be thought of as a process of making transitions from one state (letter or keyboard symbol) to another. Associated with each transition is a real number the time taken to type the second symbol following the first symbol. Similarly, written text gives rise to a discrete time series of keyboard distances between successive symbols (including spaces). We will discuss how correlating the above two time series assists us in building a model of what is being typed from the time intervals between successive symbol pairs. Such a model is very useful in security issues, such as to decipher text through recorded time between successive key strokes, as in Secure Shell (SSH) data. Also of interest, and to be discussed, is whether different typing styles lead to a proportionate decrease or increase in keyboarding times across all letter pairs, or whether there are essentially different keyboarding styles, and, if so, how those styles can be determined from time series data. We examine whether the state transition times for keyboarding form a Markov chain.

Introduction

Objectives

KEYBOARDING can be thought of as a process of making transitions from one state to another. We explored these transitions with the following goals in mind:

- Collect and Analyze Inter-stroke Timing Data
- Explore Relationships Between Keypairs
- Attempt to Identify Keystrokes from Timing Data with High Degree of Confidence

History

Since humans began interacting with technology to send information there have been timing attacks on these streams.

- ”Fisting” was used to identify wireless telegraph operators in WWII
- Early versions of SSH allowed users to be identified by packet cadence
- Companies use shopping information to target ads to customers

Keystroke Dynamics

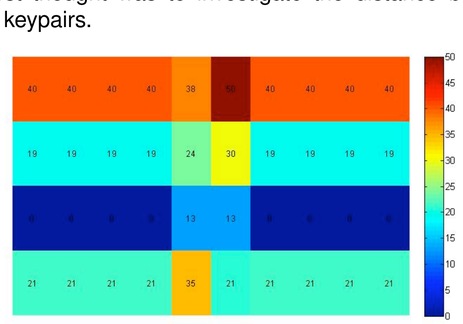
Keystroke Dynamics is the study of the manner and rhythm in which a person types. Several factors are typically unknown from a sample of writing.

- Was it typed rapidly or slowly?
- When capitalizing how did the person do it?
- Was the pace the person typed constant?
- How many mistakes were made before the presented version was produced?

Methods

Distance Model

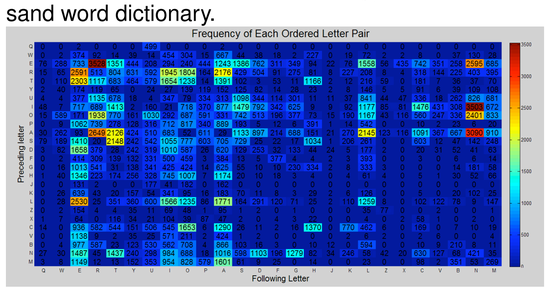
Our first thought was to investigate the distance between keypairs.



This is the distance from the home key responsible for each key in mm.
We hoped there would be a relationship between distance and time.

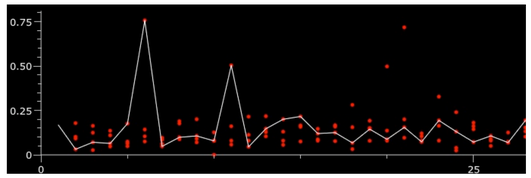
Keypairs

This is a frequency table of each keypair in a ten thousand word dictionary.



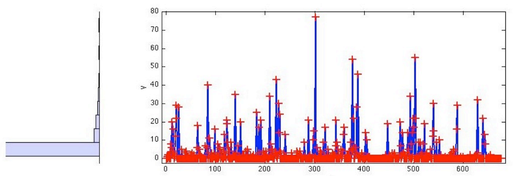
Time Series

- A Time series is a sequence of data points, measured at times steps.
- We view each key pair as a time step, and time between as the data.
- We can then use Auto-Regression in an attempt to find correlation



Probability Model

- NIG is a form of generalized hyperbolic distributions
- We use it because there is a probability of far-field behavior
- Meaning, that we want fat tails on our probability model

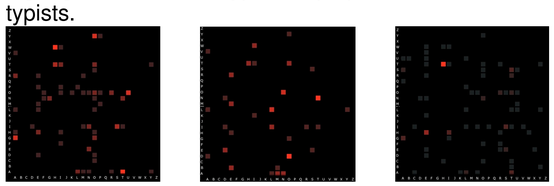


Research

Keyboarding Biometrics

Hot Spots

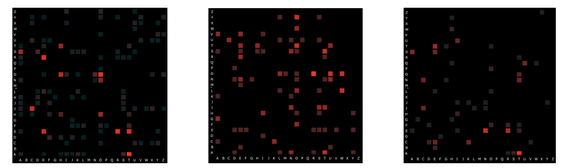
So there should be the same hotspots in everyday typing, right? Here are keypair maps produced from three typists.



Not really, in spoken and typed english there is repetition. Also, each person will have a unique map representing their pair frequency.

Keyboard Fingerprinting

- People generate a massive sample of their writing styles online
- Accessing this information is often trivial
- By adding this information we hope to increase our confidence level



Here are three sample facebook status updates.

The Rhythm Method

Autocorrelation can be thought of as the evaluation of the correlation of a time series as a function of the time steps. Autocorrelation has importance in two ways for this research.

- We autocorrelate against a database of timing information to try and find matches up to a certain size.
- We try and find the subjects typing rhythm.
- The rhythm is used to adapt the data to try and increase the confidence level of our correlation

Furthermore

At this point we’re still researching keyboarding rhythms and how to auto-adapt our data to provide real results.

- Allow just timing data to be entered for testing
- Work on adaptive algorithms for fitting data
- Produce results from timing information.