

ASSIGNMENT 2: SEQUENCE-BASED ANOMALY DETECTION

DUE: April 21, 2002 (firm deadline)

This assignment is based on several papers available from my personal web site:

http://www.cs.unm.edu/~forrest/isa_papers.htm . The most relevant papers are: A sense of self for Unix processes Intrusion detection using sequences of system calls Detecting intrusions using system calls: Alternative data models Anomaly intrusion detection in dynamic execution environments Automated response using system call delays

The basic idea in each of these papers is to model the normal behavior of a computational process using n-grams. An n-gram is a fixed size sequence of symbols. The n-grams are recorded by sliding a fixed size window along a stream of data, recording each of the unique “grams,” or tuples. A profile of normal n-grams is thus constructed, and the profile can then be used to determine whether or not subsequent behavior of the process is normal or anomalous. This procedure was used exactly in the papers “Intrusion detection using sequences of system calls” and “Detecting intrusions using system calls: Alternative data models.” A variation of the procedure, known as “lookahead pairs,” was used in the other papers.

Your assignment is as follows:

1. Choose a different data stream (that is, you are not allowed to use system calls) for which you hypothesize that normal behavior is somewhat regular.
2. Collect samples of normal behavior for this data stream, and then collect samples of abnormal behavior.
3. Use the normal behavior to construct a sequence-based profile, either using lookahead pairs or exact sequences (and ideally using both methods).
4. Then, test your profile against the sample of abnormal behavior and record how well your profile performs.

In your writeup (approx. 5 pages), please discuss the following:

- What datastream you selected and why.
- How you collected normal and abnormal behavior.
- How large your profiles are, and how quickly they converge to a stable definition of data (e.g., how much normal behavior did you need to use to construct your profile)?
- How you measure false-positive and true-positive rates.
- How your method performs using different length sequences. (It’s always interesting to consider sequences of length 1 as one of your experiments).
- Your conclusions.