Nonparametric State Space Models for SETAC Sensor Data

Todd Wasson and Dan Merl

Systems and Intelligence Analysis Section, NSED

Can network behaviors be used to accurately characterize users or hosts?

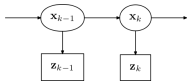
Our problem: given a set of data gathered by Foraker, the SETAC host-based network sensor, describing the users and applications that generate network connections over time, can we learn models that can allow us to distinguish hosts or users from one another, and carry out other related inference?

Background

Our approach hinges on the representation of users or hosts as a group of states of behaviors, with the current state changing over time to represent changing behaviors. Specifically, we leverage the following:

Hidden Markov Models

HMMs are stateful models that consider an observed phenomenon (like a coin toss) to be the product of some unknown underlying state (like whether the coin is fair or not). A common problem in HMMs is to infer the sequence of states through time:



HMMs are characterized by:

- Set of states; x in the above cartoon
- Transition probabilities between states
- Emission probabilities for generating an observation (z above) from each state

Hierarchical Dirichlet Process Hidden Markov Models

HDP-HMMs are an extension of HMMs in which all parameters of an HMM may be inferred (instead of given), including the number of states. Because we do not have a good idea what set of states will well-represent our network data, we use HDP-HMMs to infer everything about the HMM from training data.

Host-based Network Data Collection

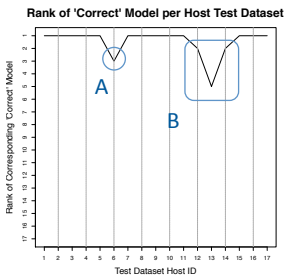
We employ data gathered by Foraker, a network sensor developed as part of SETAC. By deploying the sensor directly onto the host, information about network traffic can be joined with information not available to a remote observer, e.g., the process that is responsible for network connections, and the user responsible for each process.

An Initial Classification Case Study

To explore the type of analyses that are facilitated by our approach, we attempted to classify 17 Mac OS X hosts from LLNL. The data was comprised of:

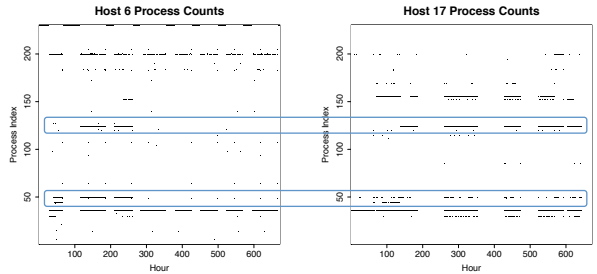
- Hourly time points, consisting of process name and number of network connections per hour
- Four weeks of training data per host
- One week of test data per host

For each training data set, we inferred an HDP-HMM. We then scored the test data sets with all models, and determined the rank of the correct model (i.e., the model trained on the training set from the cognate host):



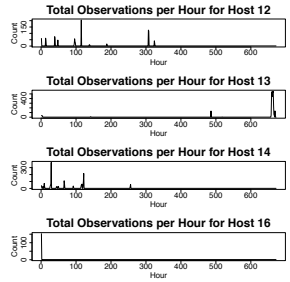
Most hosts classify correctly, but A and B perform sub-optimally, so we explore further:

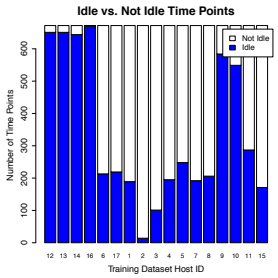
A



Host 6 was incorrectly called as Host 17, but for understandable reasons: unlike other hosts, both of these only run Firefox, Eudora, and Entourage.

B





Plots of idle vs. non-idle time points further demonstrate these hosts' mutual similarity and sparsity of data, and suggest their fundamental differences from the other hosts.

Hosts 12, 13, and 14 were incorrectly called as Host 16, because all four hosts were usually idle.

What questions can we ask with these models?

Classification

- Which existing model most likely accounts for the data, and with what probability?
- Does this data resemble previous data collected from this source?
  - If not, perhaps a different user is present, or malicious software has become active

Behavioral analysis

- What is a user or host doing over time?
  - What about users within a host?
  - What about the same user on different hosts?
- Can we identify "classes" of users or hosts, or some other level of abstraction?
- Can we use these "amalgamated" models to feed back into a classification framework?

Lawrence Livermore National Laboratory

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-POST-485220