Markov Chain Modeling for Anomaly Detection in HPC System Logs

Ultrascale Systems Research Center

Abida Haque (NCSU),
Alexandra DeLucia (Rollins College),
Elisabeth Baseman (LANL)

Workshop on HPC User Support Tools
SC'17
11/12/17

Los Alamos NATIONAL LABORATORY
EST.1943

NNSA National Nuclear Security Administration

Operated by Los Alamos National Security, LLC for the U.S. Department of Energy's NNSA

LA-UR-17-26667

- **Motivation**
- **Making a transition matrix**
  - Real Syslogs
  - Creating the matrix
  - Seeing ‘normal’ behavior
  - Using model to discover anomalies
- **Results**
- **Future Work**

Motivation

- **Too much syslog data for sysadmins**
- **How can we model the data?**
  - Markov chain model
- **Evaluate anomaly detection using precision and recall**

=

A tool which can assist sysadmins!

Real Syslogs

```
Jun 21 14:39:54 centostest1 kernel:            task   PID         tree-key  switches  prio     exec-runtime         sum-exe
c         sum-sleep
Jun 21 14:39:54 centostest1 kernel: ------------------------------------------------------------------------------------
------------------
Jun 21 14:39:54 centostest1 kernel: R           bash  8686     28730.348788       215   120     28730.348788        73.6365
67    329852.286947 /
Jun 21 14:39:54 centostest1 kernel:
Jun 21 14:40:01 centostest1 CROND[8751]: (root) CMD (/usr/lib/sa/sa1 1 1)
Jun 21 14:40:08 centostest1 kernel: SysRq : Emergency Sync
Jun 21 14:40:08 centostest1 kernel: Emergency Sync complete
Jun 21 14:40:16 centostest1 kernel: SysRq : Manual OOM execution
Jun 21 14:40:16 centostest1 kernel: events/0 invoked oom-killer: gfp_mask=0xd0, order=0, oom_adj=0, oom_score_adj=0
Jun 21 14:40:16 centostest1 kernel: events/0 cpuset=/ mems_allowed=0
Jun 21 14:40:16 centostest1 kernel: Pid: 7, comm: events/0 Not tainted 2.6.32-504.1.3.el6.i686 #1
Jun 21 14:40:16 centostest1 kernel: Call Trace:
Jun 21 14:40:16 centostest1 kernel: [<c04f0d94>] ? dump_header+0x84/0x190
Jun 21 14:40:16 centostest1 kernel: [<c04f1138>] ? oom_kill_process+0x68/0x28Jun 21 14:40:59 centostest1 kernel: imklog 5.8
.10, log source = /proc/kmsg started.
Jun 21 14:41:07 centostest1 kernel: eth2: no IPv6 routers present
Jun 21 14:41:11 centostest1 kdump: mkdumprd: failed to make kdump initrd
Jun 21 14:41:14 centostest1 acpid: starting up
Jun 21 14:41:14 centostest1 acpid: 1 rule loaded
Jun 21 14:41:14 centostest1 acpid: waiting for events: event logging is off
Jun 21 14:41:15 centostest1 acpid: client connected from 8289[68:68]
Jun 21 14:41:15 centostest1 acpid: 1 client rule loaded
Jun 21 14:41:16 centostest1 automount[8309]: lookup_read_master: lookup(nisplus): couldn't locate nis+ table auto.master
Jun 21 14:41:16 centostest1 sshd[8338]: Server listening on 0.0.0.0 port 22.
Jun 21 14:41:16 centostest1 sshd[8338]: Server listening on :: port 22.
Jun 21 14:41:16 centostest1 xinetd[8348]: Reading included configuration file: /etc/xinetd.d/chargen-dgram [file=/etc/xinet
d.conf] [line=49]
Jun 21 14:41:16 centostest1 xinetd[8348]: Reading included configuration file: /etc/xinetd.d/chargen-stream [file=/etc/xine
td.d/chargen-stream] [line=67]
Jun 21 14:41:16 centostest1 xinetd[8348]: Reading included configuration file: /etc/xinetd.d/daytime-dgram [file=/etc/xinet
d.d/daytime-dgram] [line=67]
Jun 21 14:41:16 centostest1 xinetd[8348]: Reading included configuration file: /etc/xinetd.d/daytime-stream [file=/etc/xine
```

Making a Transition Matrix

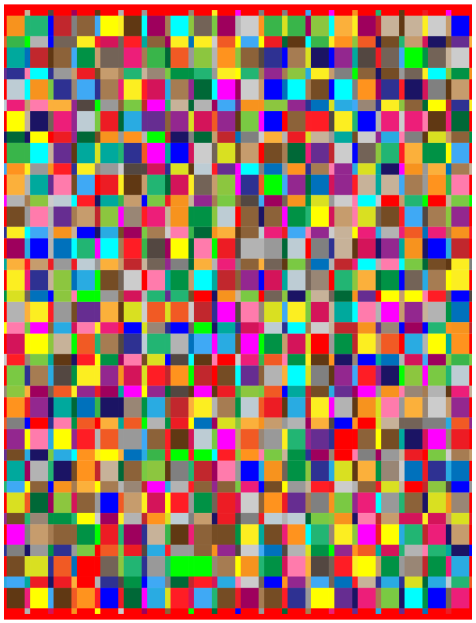
Processing Real Syslogs

- **Clean up and label messages with a number**
- **Example:**

Enumeration	Message
0	Starting up
1	Client connected from [:]
2	The audit daemon is exiting
1	Client connected from [:]

Creating the matrix

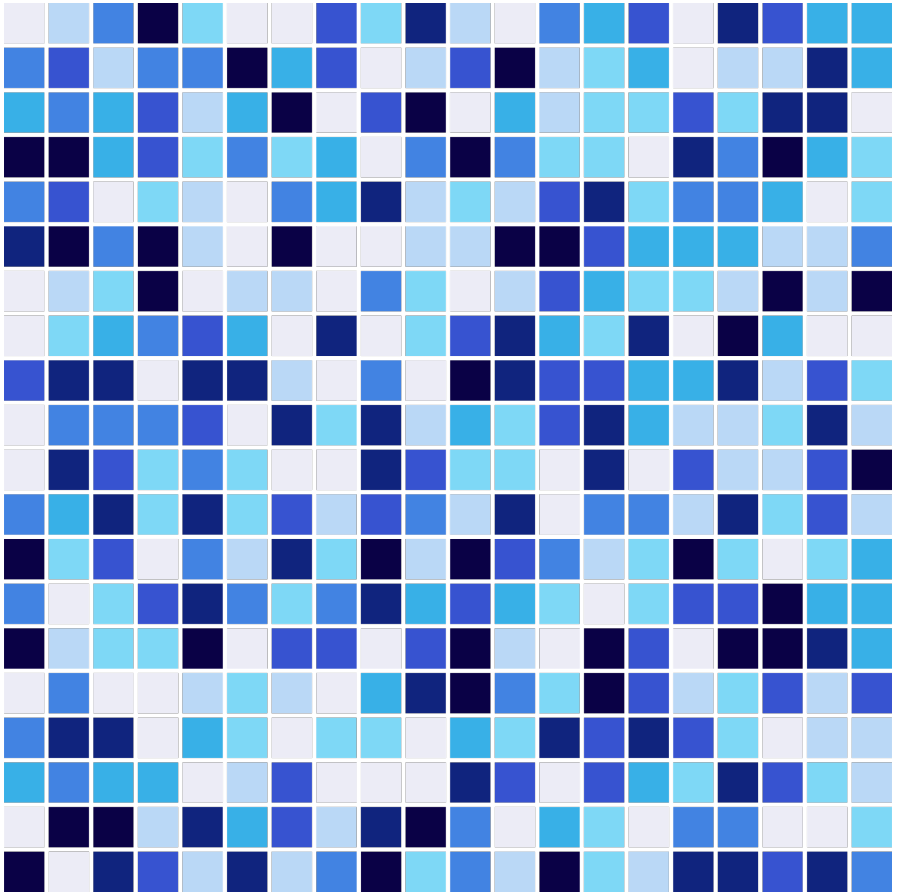
- Sequence of enumerated messages, eg:
- X = [1,2,3,1,3,3,4]
- Count up each transition in the syslog:
- 1 → 2
- 2 → 3
- Etc…



Simplify the data and organize it!

Markov Transition Matrix

- Count up the number of transitions
- Count up the number of times a message appears overall
- Divide to get a probability
- Save in a matrix



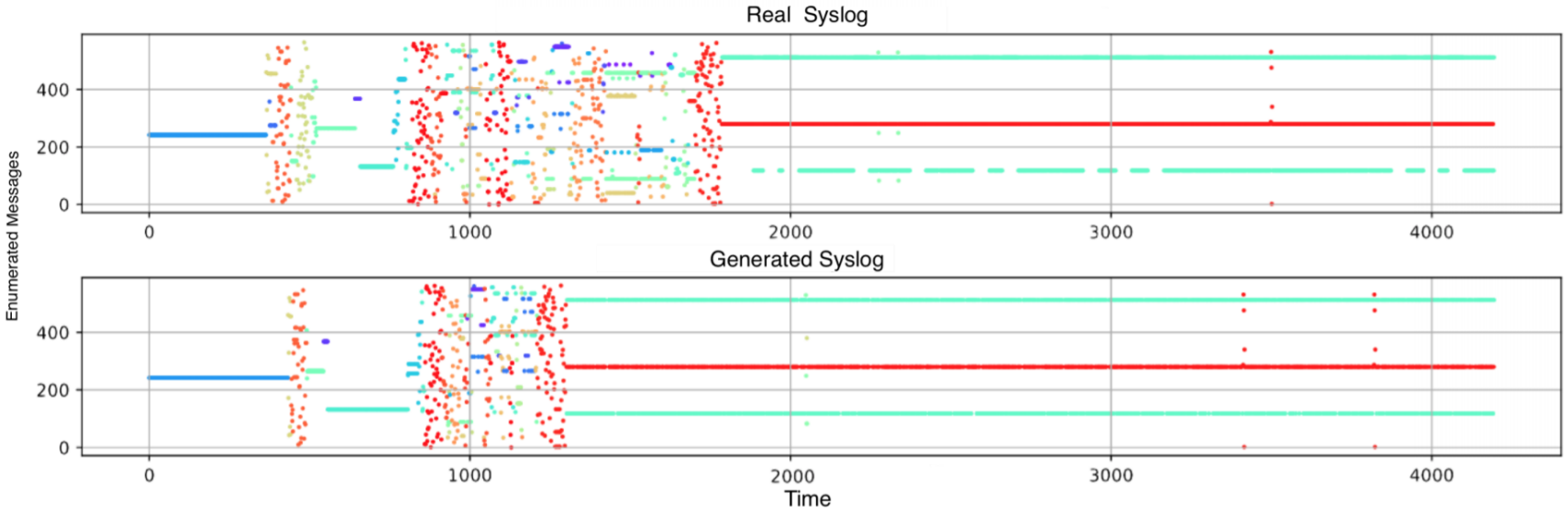
Algorithm

- **Create transition matrix using training syslog**
- **On a test sequence, determine the probabilities of transitions**
- **Return the least probable transitions**
  - Those are the lines most likely to be anomalous!

Results

‘Normal’ behavior

- Do a random walk based on the transition matrix
- Compare visually



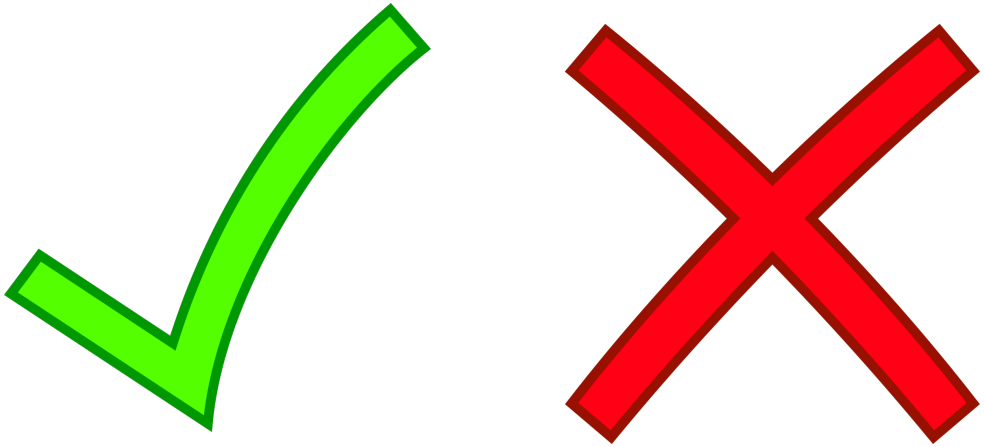
How anomalous?

- **Two scoring schemes**

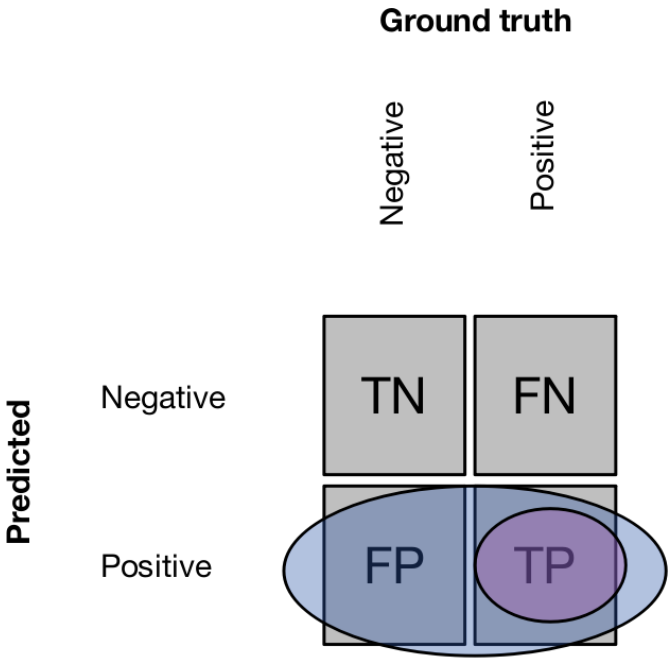
1. **Just use message transition probabilities**
2. **Weight the message transition probabilities with tag probabilities**
   - Various daemons, the kernel
   - This way gives better results!

Evaluating the Model

- **Deletions**
  - Started with lines 1 2 3 in order
  - Remove line 2: 1 ~~2~~ 3
  - Hopefully 1 gets marked as anomalous!
- **Insertions**
  - 1 2 3
  - 1 2 **4** 3
  - Both 2 and 4 should get marked!



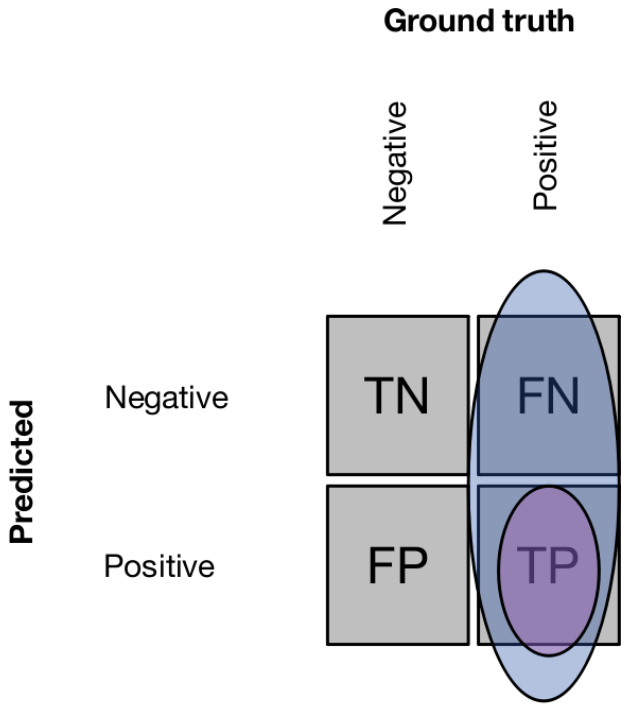
Precision and Recall



Precision

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$



Recall

Future Work

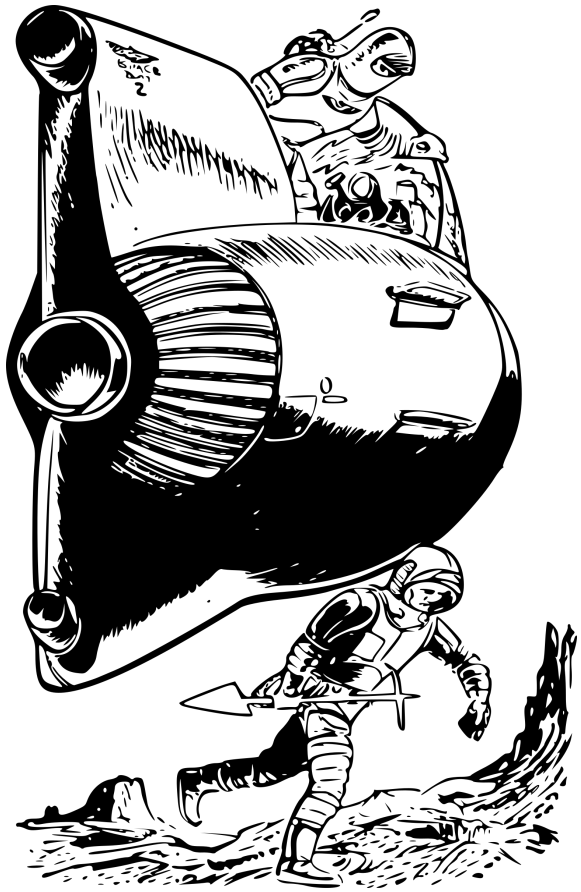
A tool for admins

- **Create transition matrix for logs for some training period**
- **Check the transition matrix against logs for testing period**
- **Return top X answers**
- **Sysadmin checks the messages and segments around those lines.**
  - Maybe a new message appears?
  - If message is normal, add to the model!

Use sysadmin's domain expertise!

Improvements to the Model

- **Using printk messages**
- **Topic modeling (finding keywords using TF-IDF) in the messages.**
- **Clustering the messages**
  - Using the topic vector (k-means clustering)
  - Using the transition matrix (Markov chain clustering).
- **Data stream**
- **Adding time stamp**



Thank you!

With special thanks to everyone

- **Lissa Baseman**
- **Alexandra Delucia**
- **Sean Blanchard**

❖ **Los Alamos National Laboratory**

❖ **New Mexico Consortium**

❖ **Ultrascale Systems Research Center**

Any questions?

For further questions: ahaque3@ncsu.edu