Interactive Visualization for System Log Analysis

Armin Samii • Woojong Koh

UC Berkeley CS 262A, Fall 2013

Abstract

Goals:
- Accelerate anomaly detection
- Simplify trend understanding

Observations:
- System logs are unintelligible
- Admins grep to find relevant info

Contributions:
- Supplement grep with visualizations
- Cluster nodes by trends

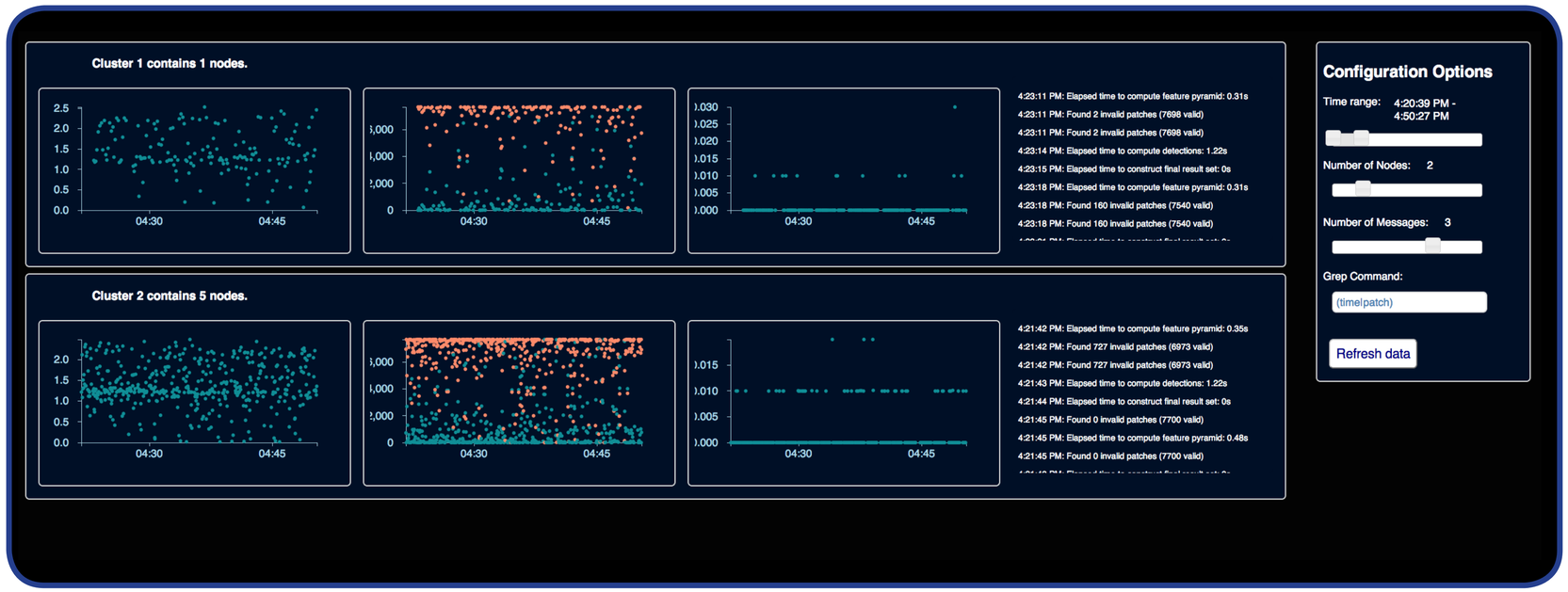
Algorithm Summary

1. Compute per-node features
   - Message count vector
   - Statistics across time
   - Each node computes its own features
2. Select k most representative nodes
   - Using distributed k-means clustering
   - User interactively chooses k
3. Display interactive visualization
   - Client processes visualization
   - Web server clusters on-demand
   - Server communication through Apache

Future Work

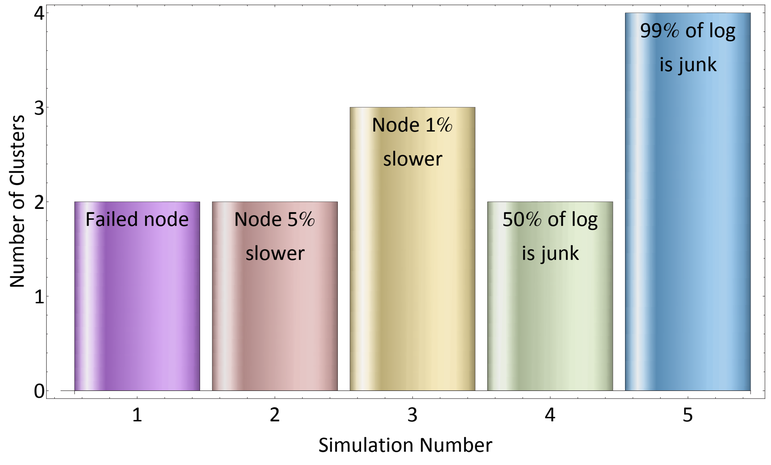
- Automatically determine plot type
- Extend to online, realtime analysis
- Preprocess features
- Cluster & control log message types

Screenshot



Anomaly Detection Results

Metric: How many clusters are required until anomalous node in its own cluster? (Best case: 2)



Timing Results

- Run on 500,000 messages per node
- 30mb sent to HTTP Server per node
- Client-side visualization processing