CCICADA

Command, Control, and Interoperability Center for Advanced Data Analysis

A Department of Homeland Security Center of Excellence

Graph Evolution Over Time: Detecting Anomalies in Networks

Andrew Rodriguez, Dr. Michael Tortorella, and Dr. W. Art Chaovalitwongse, Rutgers University

Dr. Tin Kam Ho, Dr. Jin Cao, Dr. Harald Steck, and Dr. Ayou Chen, Bell Labs

Motivation

- What is the problem?
  - Anomalous activity (e.g., element failures, security-related problems) likely degrades network reliability and performance
- Who cares?
  - Those who supply and consume network services
- What makes this problem difficult?
  - Centralized network performance information is often not directly available
  - No model for normal network operation
  - Large amounts of data to process
- What does this presentation offer?
  - Overview of network activity monitoring
  - Introduction to our work in anomaly detection
  - Invitation for collaboration

Anomaly Examples

- Non-security related
  - File server failure
  - Broadcast storm
  - Congestion due to element failure
- Security related
  - Denial of service attack
  - Botnets

Background

- Complex networks are composed of many individual entities, data collection done in various ways
  - Probes
  - Entity-based, use knowledge of topology
- Different working environments
  - Non-cooperative networks, multiple administrative domains
  - Single administrative domain
- Large amounts of data are collected to achieve basic understanding
  - Must be measured, analyzed, synthesized to extract network information

In the CCICADA Mix

1. Collecting and building knowledge from data
2. Enriching knowledge and inference

   Project 5: Hypothesis Formation and Anomaly Detection.
3. …

In the CCICADA Mix

Command, Control, and Interoperability Center for Advanced Data Analysis
A Department of Homeland Security Center of Excellence

1. Collecting and building knowledge from data
2. Enriching knowledge and inference

   Project 5: Hypothesis Formation and Anomaly Detection.
3. …

Network Data

- What are sources of network data?
  - Probing tools
  - Packet filtering, packet headers
  - Network management protocols
- What is normal traffic behavior?
  - That is hard to say

Challenges

- Non-stationary data
- Large data-sets in short time intervals
- Determining length time interval
- Staleness of older data points
- Lack of labeled data for validation
- Danger of hypersensitivity or over-fitting

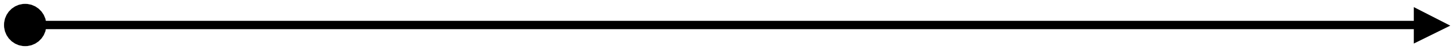
Previous Approaches

- Rule-based, case based reasoning
  - Build records of past anomalous instances
  - Dependence on past information
- Pattern matching
  - Construct symptom-specific feature vector
  - Use patterns of known attacks for detection
- Finite state machine modeling
  - States are sequence of alarms
  - Possible explosion of state space
- Signal processing

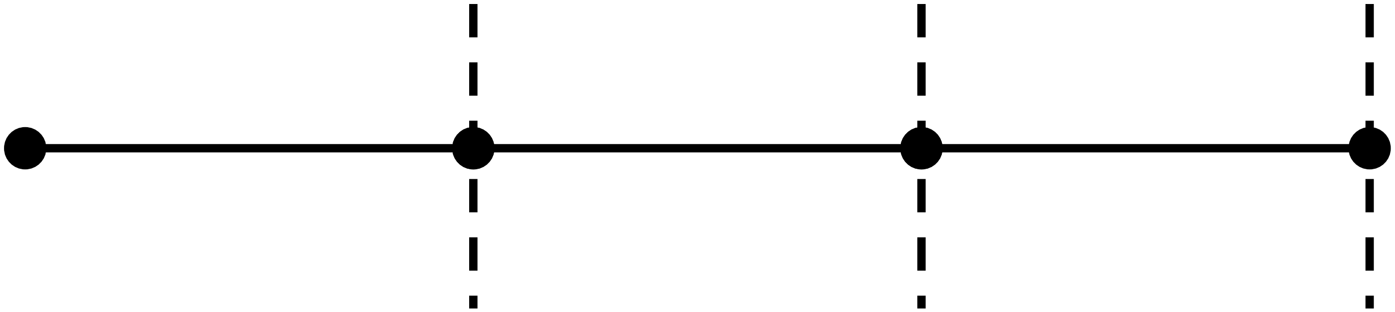
Performance Metrics

- Subject to Type I and Type II errors
  - Mean time between false alarms
  - Time until anomaly detection

Time



Evolution of a dynamic graph



Series of static graphs

Graph Characteristics

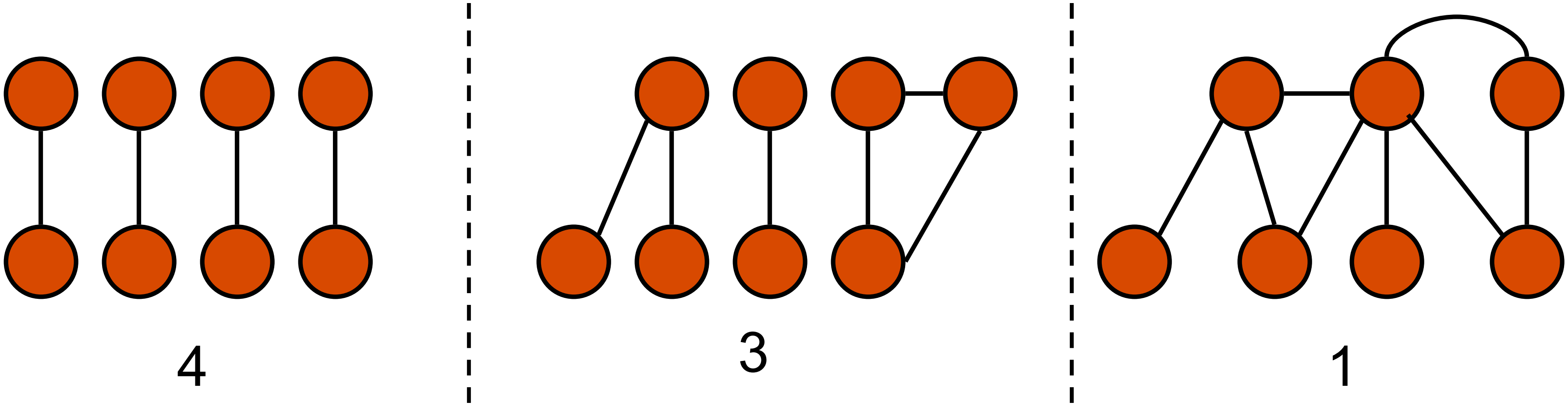
- Static graph
  - In-degree distribution
  - Out-degree distribution
  - Distribution sized of weakly connected components
  - Diameter of graph
  - Clustering coefficient
- Evolving graph
  - Evolving diameter
  - Densification process

Method

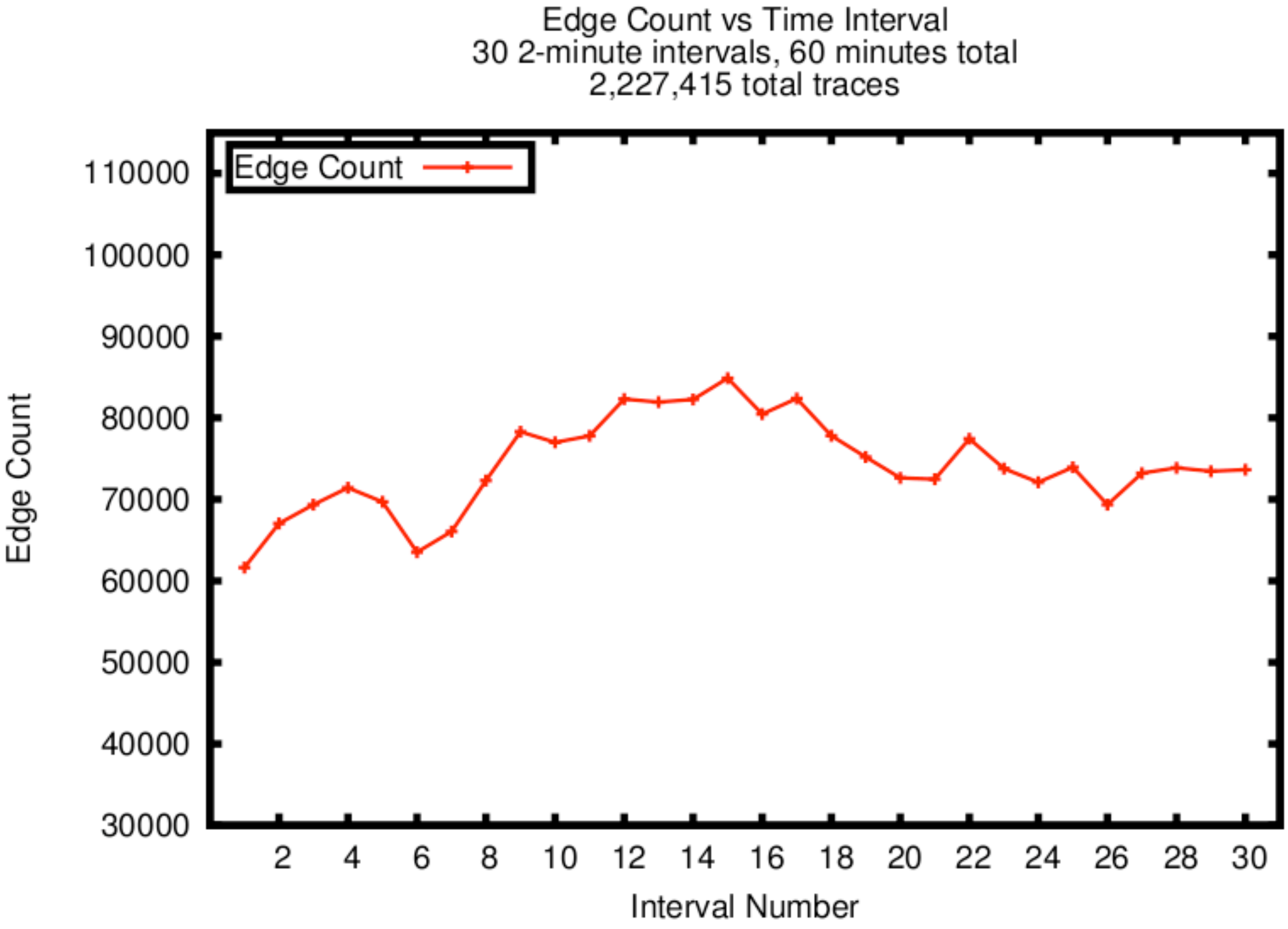
- Online change-point detection given source-destination data
  - Take samples of fixed time interval
  - Compute static graph statistics for each interval
  - Make a decision independently for each interval

Example

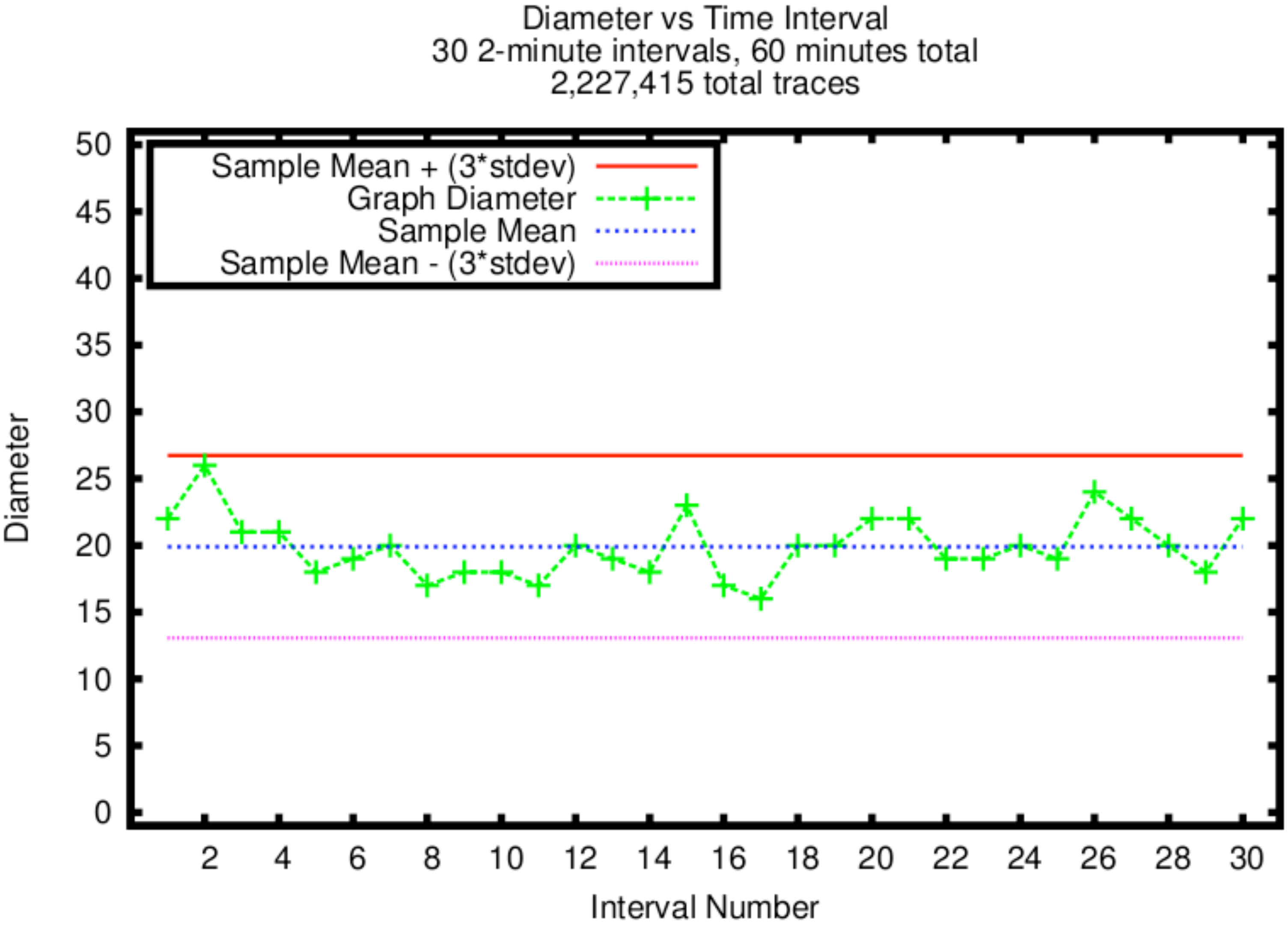
- Detecting change in network characteristics among a series of data points
  - – WCC example for communication graph



Observation

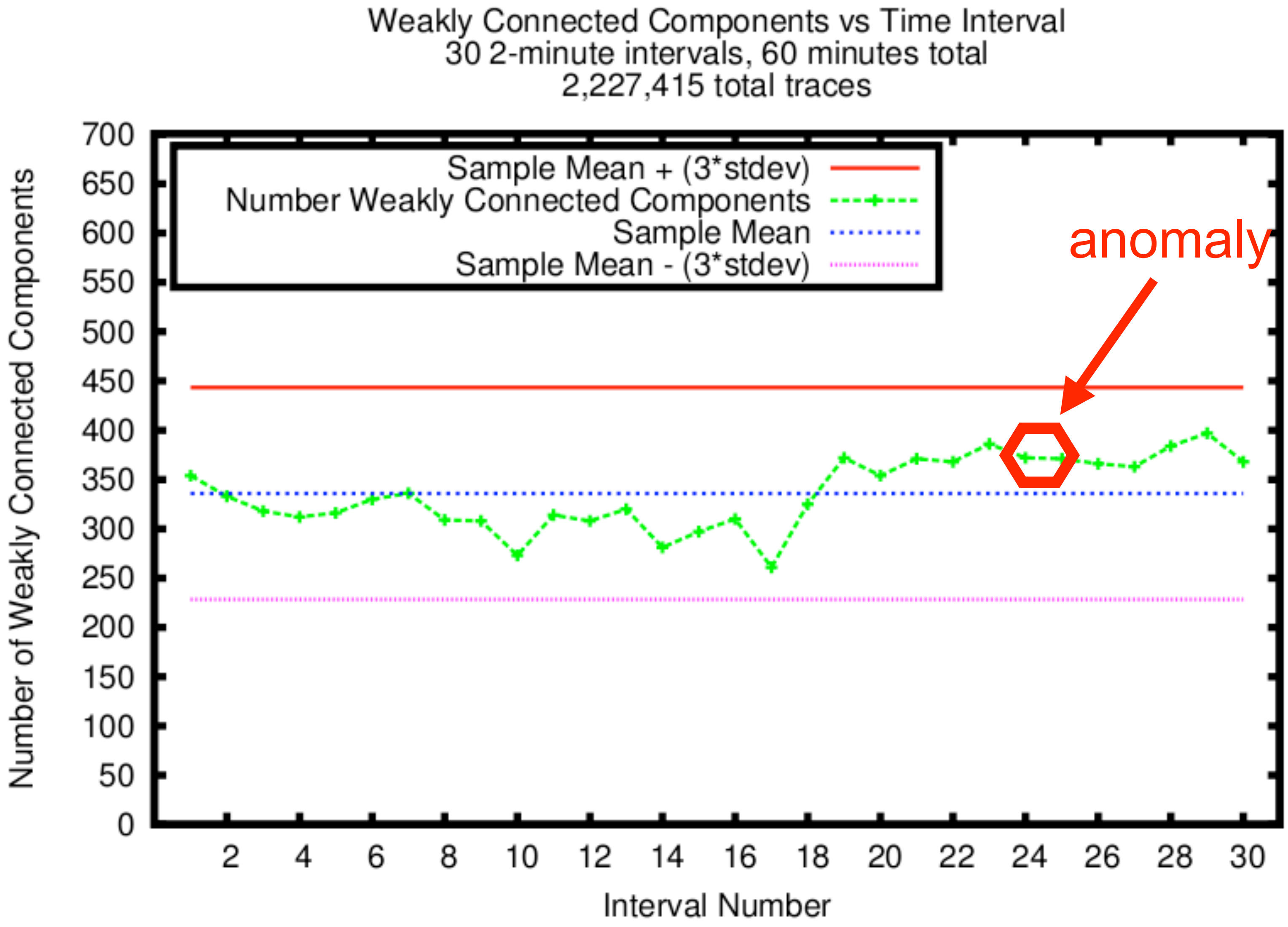


Observation



Diameter vs Time Interval
30 2-minute intervals, 60 minutes total
2,227,415 total traces

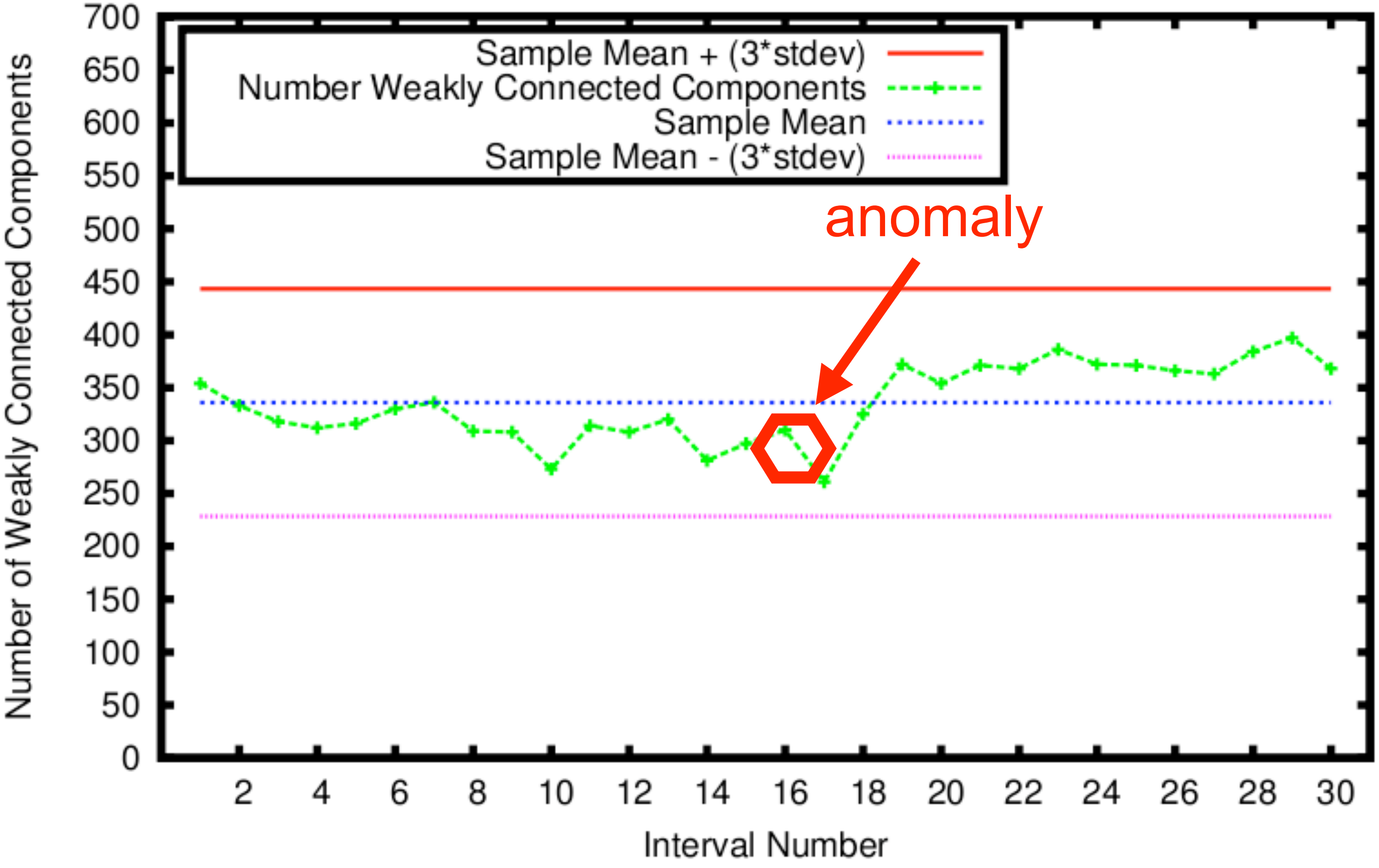
Observation



anomaly

Observation

Weakly Connected Components vs Time Interval
30 2-minute intervals, 60 minutes total
2,227,415 total traces



anomaly

Challenges

- Non-stationary data
- Large data-sets in short time intervals
- Determining length time interval
- Staleness of older data points
- Lack of labeled data for validation
- Danger of hypersensitivity, over-fitting
  - A vulnerability

Ideas Generated

- Inject synthetic anomaly data into otherwise normal data-set
  - Independently developed, shared data-sets

Thank you.