Traffic anomaly detection using a distributed measurement network

Razvan Oprea
Supervisor: Emile Aben (RIPE NCC)

University of Amsterdam

System and Network Engineering

February 8, 2012

Outline

- Introduction
- Similar projects
- Research questions
- Basic research idea
- Choosing a metric
- Ground-truth reflection
- Analyzing the collected data
- Conclusions and recommendations

Introduction

What is the RIPE Atlas distributed measurement network?

- A collection of probes deployed worldwide, conducting specific Internet network measurements.
- A backend system which collects, processes, analyzes and presents the data to the users
- More than 1024 online probes, many more planned

Figure: Coverage of the RIPE Atlas network

http://atlas.ripe.net

Similar projects

SamKnows

- operated by SamKnows Limited and a "community of volunteers"
- funding from the FCC in US and the European Commission in the EU
- active in the US and EU (as of the fall of 2011)

Project BISmark

- project led by Georgia Tech and University of Napoli Federico II
- funding form US National Science Foundation and Google Inc.
- no major rollout yet

Key differences between the networks

RIPE Atlas

- geared towards home users and network operators
- small and unobtrusive
- relatively cheap
- hardware and software bundle
- limited capability, power is in the numbers

The two other networks

- geared towards home users
- all traffic must pass through their devices
- usually embedded into home routers
- hardware or software versions
- more types of measurements

RIPE Atlas measurements

What is being measured by the RIPE Atlas probes?

- ICMP echo requests (ping) to the first and second hops and an array of fixed destinations (unicast and anycast)
  - Round Trip Times (RTT)
  - Packet loss
- Traceroute to fixed destinations
- DNS SOA record checking for the root name servers
- User-defined measurements

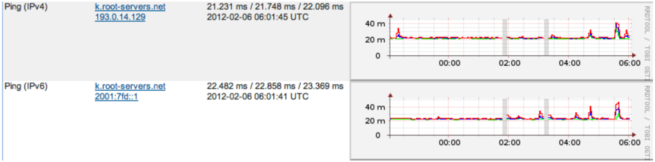


Figure: IPv4 and IPv6 RTT to anycast destination

Research questions

Research question:

- How can the data collected by the RIPE Atlas provide information for indicating a network operational problem?

Sub-research questions:

- What metrics are useful for traffic anomaly detection in RIPE Atlas data?
- How can traffic anomalies detected by the RIPE Atlas be localized to a network or geographic location?

Basic research idea

Step 1: relevant metric

- Choose a relevant metric from the measurements conducted by RIPE Atlas.

Step 2: ground truth reflection in the collected data

- Look for significant network -related events from the past year
- See how are they reflected in the data collected by the probes

Step 3: relation between the data collected by different probes

- Choose a probe in a certain geographical area or network (AS)
- See if there is a relation between the data collected by different probes in the same area

Choosing a metric

Potential candidates were considered among the measurements RIPE Atlas probes can perform.

Eliminated:

- Packet loss (an additon to RTT, but not the main metric)
- DNS SOA queries (not a performace metric)
- User-defined measurements (subset of probes)

Remaining:

- RTT (minimum RTT)
- traceroute

Ground-truth reflection 1/4

Localization parameters:

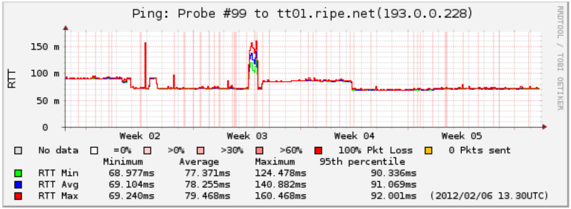
- Time: Most measurement data started being collected in September 2011
- Space: Visibility is limited to the areas in which RIPE Atlas probes exist

Types of events researched:

- published large Internet outage reports
- large-scale power outages
- de-peerings
- cable landings (or cuts)

"What do the probes see?"

- RRD graphs



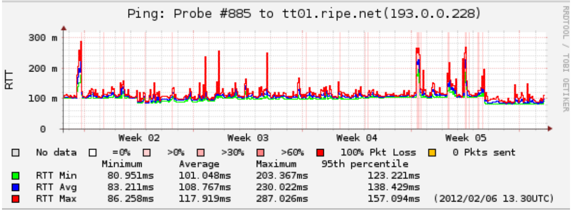


Figure: New cable landing

"What do the probes see?"

- Tridimensional graphs ("heat maps" idea) developed by Emile Aben (RIPE NCC)

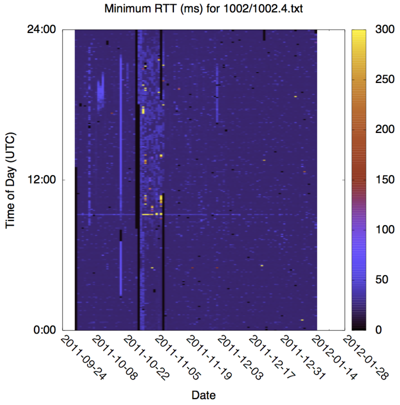


Figure: Heat map example

Ground-truth reflection 4/4

Ground-truth conclusions:

- None of the events researched was clearly reflected in the graphical representation of the Atlas data
  - Atlas probes are mainly concentrated in the European area
  - No major network events happened in Europe in the second half of 2011
  - European Internet providers do not generally publish network outage history
- The RRD graphs: good in showing changes in the RTT measurements
- The "heat map" graphs: better for observing patterns (for instance, day-night traffic patterns)

Analyzing the collected data 1/5

Initial idea:

1. create simple time series, per probe, based on the minimum RTT (minRTT)
2. see if there is a strong correlation between the series within an AS

Why this doesn't work well:

- the time series contain a lot of noise
- cross-correlation between multiple series is not trivial to compute
- even is a correlation is found, we wouldn't know where to look for events

Analyzing the collected data 2/5

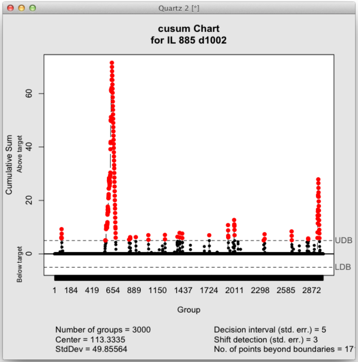
A better idea:

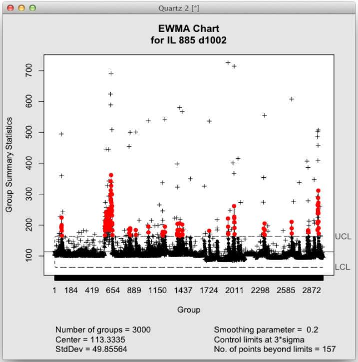
1. create simple time series, per probe, based on the minRTT
2. create control charts (per probe)
3. see if violations of the control limits is shared by multiple probes in an AS

Two types of control charts were considered:

- Cumulative Sum Control Chart (CUSUM) - fast implementation in R
- Exponentially Weighted Moving Average (EWMA) - slower R implementation

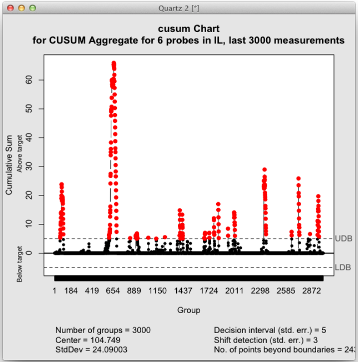
These are the CUSUM and the EMWA for the same probe, for the last 3000 measurements:

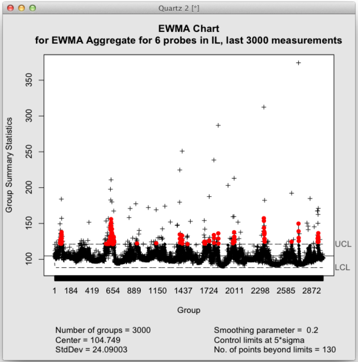




Aggregating the time series in a matrix, per AS (valid if minRTT are within a close range):

- CUSUM and EWMA appear to yield similar results





Analyzing the collected data 5/5

Data analysis conclusion:

- Simple idea: aggregation of violations points from individual control charts

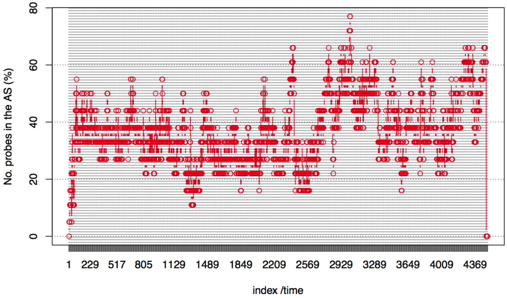


Figure: AS 3265 - percentage of probes violating the control limit

Conclusions and recommendations

- Increase the density of Atlas probes in every AS to improve visibility
- Fetch and aggregate the public data from every major ISP's network outage pages
- Data analysis algorithm needs to be implemented to scale well
- Frequent process of control limit violation points
- The decision between CUSUM and EWMA will have to be taken later (or using both)

Questions ?