BGP ANOMALY DETECTION USING DATA MINING TECHNIQUES

Iñigo Ortiz de Urbina

Introduction

- Goal
  - Apply machine learning algorithms to mine network data and produce a preliminary offline anomaly detection system
- Keywords
  - *BGP, network security, data mining, anomaly detection, Internet Routing Forensics, perl*

Abnormal BGP Events

- Events can spread either
  - Globally or locally
  - Sustained or short period
- Serious economical and social impact
- Common ABEs
  - Sea cable cuts
  - Prefix hijacks
  - Power blackouts
  - Worms
  - Routing table leaks

What is Data Mining

- Data mining is the process of extracting patterns from data
- Becoming an increasingly important tool to transform data into knowledge
- Implies
  - Data preprocessing
  - Data mining
  - Result validation

Anomaly Detection

- Assumption
  - Trends sufficiently different from normal behavior are potentially dangerous
- Anomaly detection systems model normal behavior
- System spots anomalies matching events against the model
- It is a *classification* problem

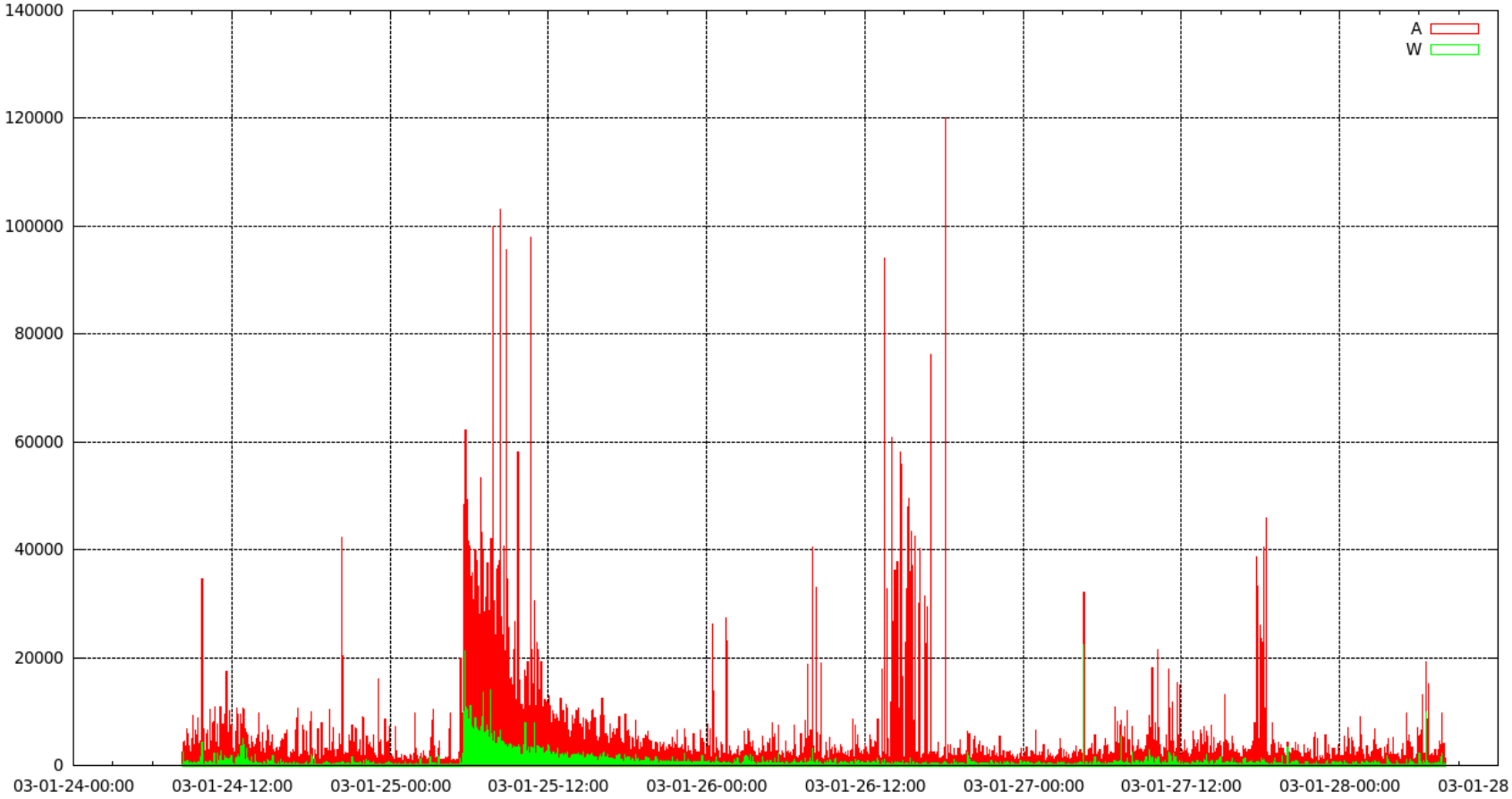
The dataset

- Update messages from independent Route Information Servers
  - A collection of Remote Route Collectors distributed globally
- MRT format
  - Protocol | Time | Type | PeerIP | PeerAS | Prefix | <update dependant information>
  - ASPATH | Origin | NextHop | Local_Pref | MED | Community
- Sample
  - BGP|884831401 |A |144.228.107.1|1239|205.113.0.0/16|1239 64535769 | IGP|192.41.177.241|0|91
  - BGP|884831402 |W |204.70.7.53|3561|198.163.111.0/24
  - Millions of these updates

Attributes and descriptions

- **NumAnnounce**: number of announcements seen in a given time window
- **NumWithd**: number of withdrawals seen
- **NumUpdate**: Linear combination of the number of announcements and withdrawals. It represents the main volume of updates exchanged

SQL/Slammer January 2003



Attributes and descriptions

- **AnnouncedPrefixes**: The total number of announced prefixes in a given period. This is an implicitly important feature in BGP
- **WithdrawnPrefixes**: The total number of withdrawn prefixes in the bin
- **MaxAnnouncementsPerPrefixes**: Maximum announcements *per prefix*

Attributes and descriptions

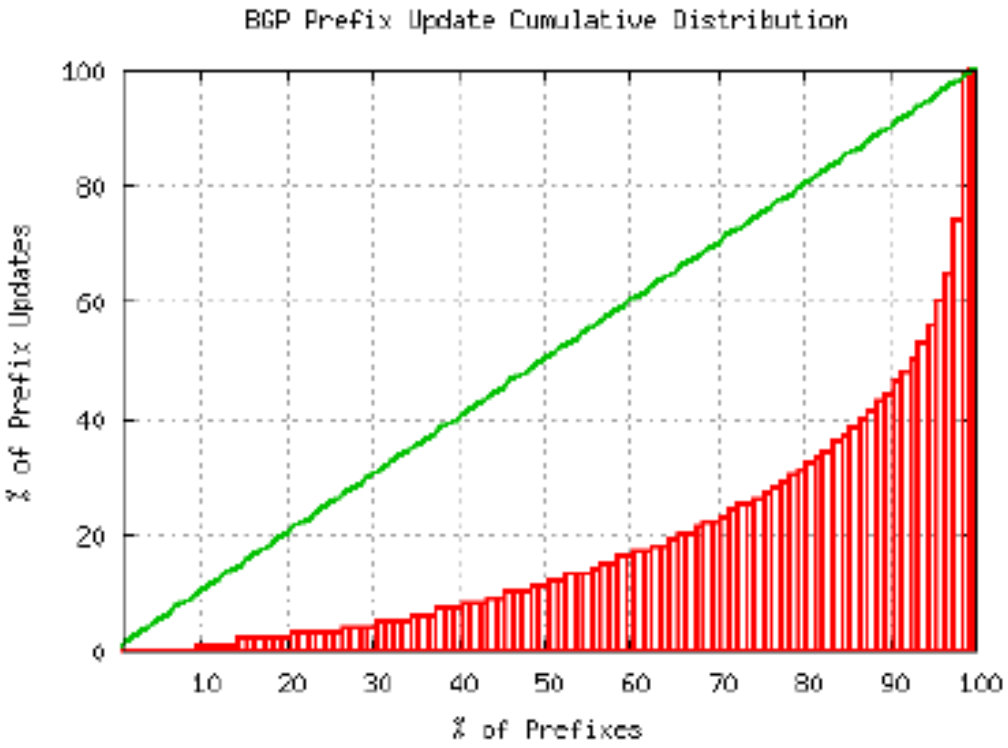
- **MaxASPL**: Maximum AS Path length. Instability periods tend to show longer AS Paths
- **maxUASPN**: Maximum unique AS numbers in the AS Path. Instability periods introduce new AS numbers
- **announceToLongerPath**: Total number of updates announcing a longer path for any given prefix. Measures how far we are from convergence

Attributes and descriptions

- **FirstOrderRatio**: Division ratio between the first most active announced prefix and the total number of announcements
- **concentrationRatio**: Division ratio between the three most active announced prefixes and the total number of announcements

Attributes and descriptions

- *Updates are not uniformly distributed along prefixes!*



Source: Geoff Huston

Attributes and descriptions

- **Extracted attributes also include average values, per significant attribute computed**
- **Certain attributes ommited and discarded after feature selection**
  - ThirdOrderRatio: not much info left after FirstOrderRatio and ConcentrationRatio
  - Minimum values: do not provide any meaningful information. Typically, **always** 0, 1 or similar

Researched events

Data for	RRC	Month
Moscow blackout	rrc05, Vienna	May 2005
East-coast blackout	Routeviews, Oregon	August 2003
Switzerland-Italy blackout	rrc09, Zurich	September 2003
SQL/Slammer worm	Routeviews, Oregon	January 2003
Nimda worm	rrc04, Geneva	September 2001
Witty worm	Routeviews, Oregon	March 2004
AS9121 routing table leak	rrc05, Vienna	December 2004
AS23724 routing table leak	rrc06, Otemachi	May 2010
YouTube.com prefix hijack	Routeviews, Oregon	February 2008
Google.com prefix hijack	Routeviews, Oregon	May 2005
Mediterranean sea cable cut	rrc10, Milan	December 2008
Luzon strait cable cut	rrc06, Otemachi	December 2006
AS47868 AS path prepending bug	rrc12, Frankfurt	February 2009

- Update: RIPE/Duke University experimental optional attribute

Experiment results

	Percentage Split 66%			
	Accuracy			
	OneR	J48	NB	SMO
Moscow blackout	96.281	95.4545	96.281	**97.1074**
East-coast blackout	82.2345	81.5636	25.1459	**82.8471**
Switzherland-Italy blackout	**90.7157**	89.7029	32.0392	90.5807
SQL/Slammer worm	85.443	**91.2975**	84.731	86.8671
Nimda worm	*59.4254*	***61.8375***	*55.4156*	*61.1307*
Witty worm	95.1102	**95.2176**	27.3509	95.2176
AS9121 table leak	*47.3322*	*64.0275*	*61.1015*	*65.5766*
AS23724 table leak	**97.541**	97.541	37.1585	97.541
YouTube.com hijack	86.7036	**87.5346**	83.6565	87.5346
Google.com hijack	**98.563**	98.563	56.8047	98.563
Mediterranean cable cut	*60.9914*	*62.931*	***63.5972***	*63.5962*
Mediterranean cable cut II	**98.4594**	98.4594	52.6331	98.4594
Luzon strait cable cut	**98.6826**	98.5284	41.7239	98.0519
AS47868 AS path bug	**99.1228**	97.9532	99.1228	98.538

Experiment results

- 11 out of 14 events are classified with more than **90%** accuracy
- Poor detection for certain events
  - Data cleaning and normalization
  - Information gain based feature selection
  - Further test and tuning needed
  - Add more depth and semantics to our feature vector
    - Root-cause AS, inter-AS correlation...
    - Problem handling multivariate data in Weka
      - Use R to mine data instead

Conclusions

- Encouraging first results
- Proof that we are in the right road
- Modern worms are easier to detect in network and transport layers using netflow
- Both offline and online frameworks can be achieved
  - Alarm reporting, automatic classification
- Data normalization and appropriate feature selection is a must

Interesting Resources on Internet Intelligence

- BGPMon
- RIPE RIS Raw Data
- Renesys blog
- Team Cymru
- Caida
- Rotueviews
- HE BGP Toolkit
- Internet Measurement Data Catalog
- The Data Repository
- RIPE Labs Datasets
- BGP Potaroo Reports

Thanks for your attention

@ioc32
inigo@infornografia.net