A Distributed Network Security Analysis System

Based on Apache Hadoop-Related Technologies

Bingdong Li,

Jeff Springer , Mehmet Gunes , George Bebis

University of Nevada Reno

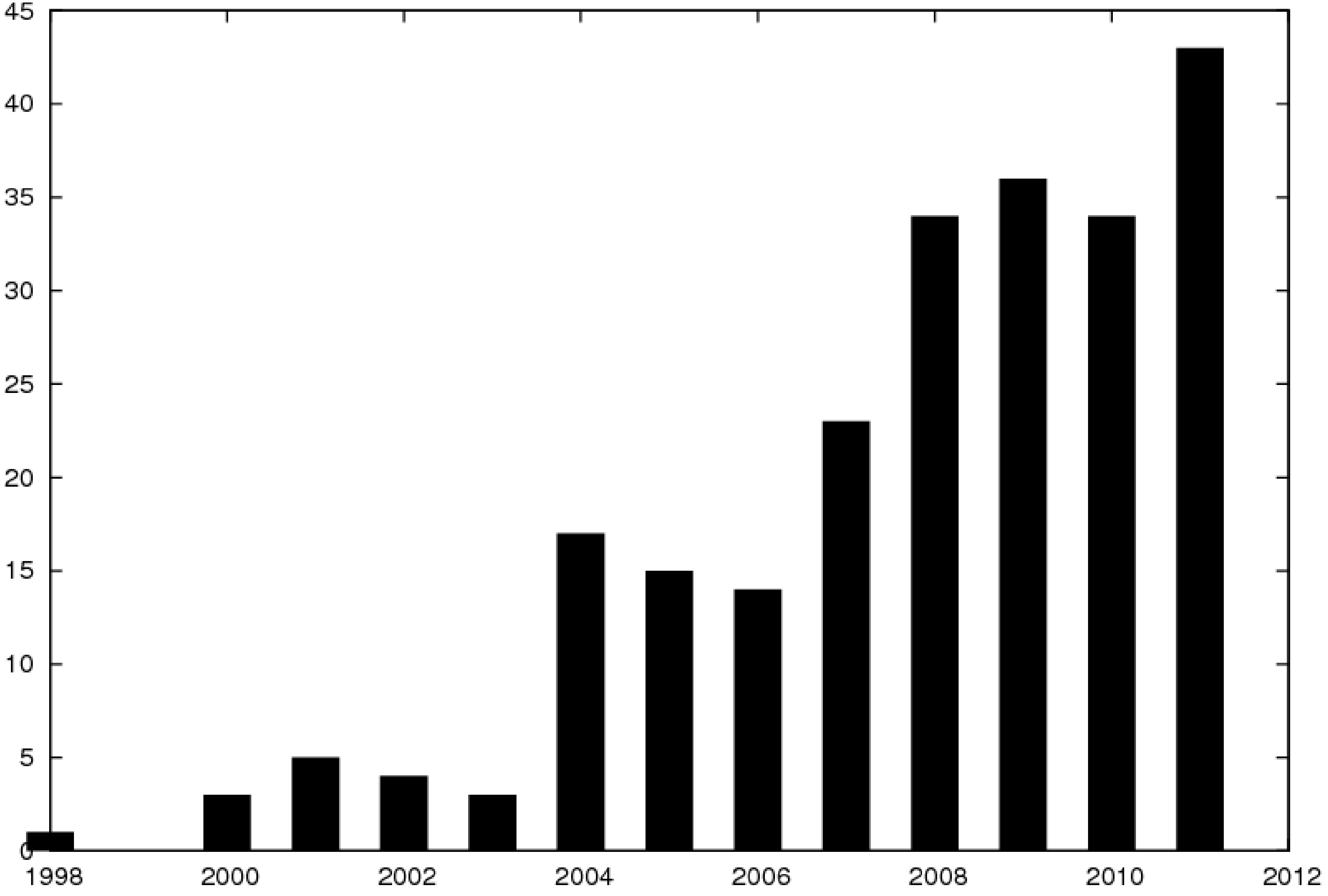
FloCon 2013

January 7-10, Albuquerque, New Mexico

Agenda

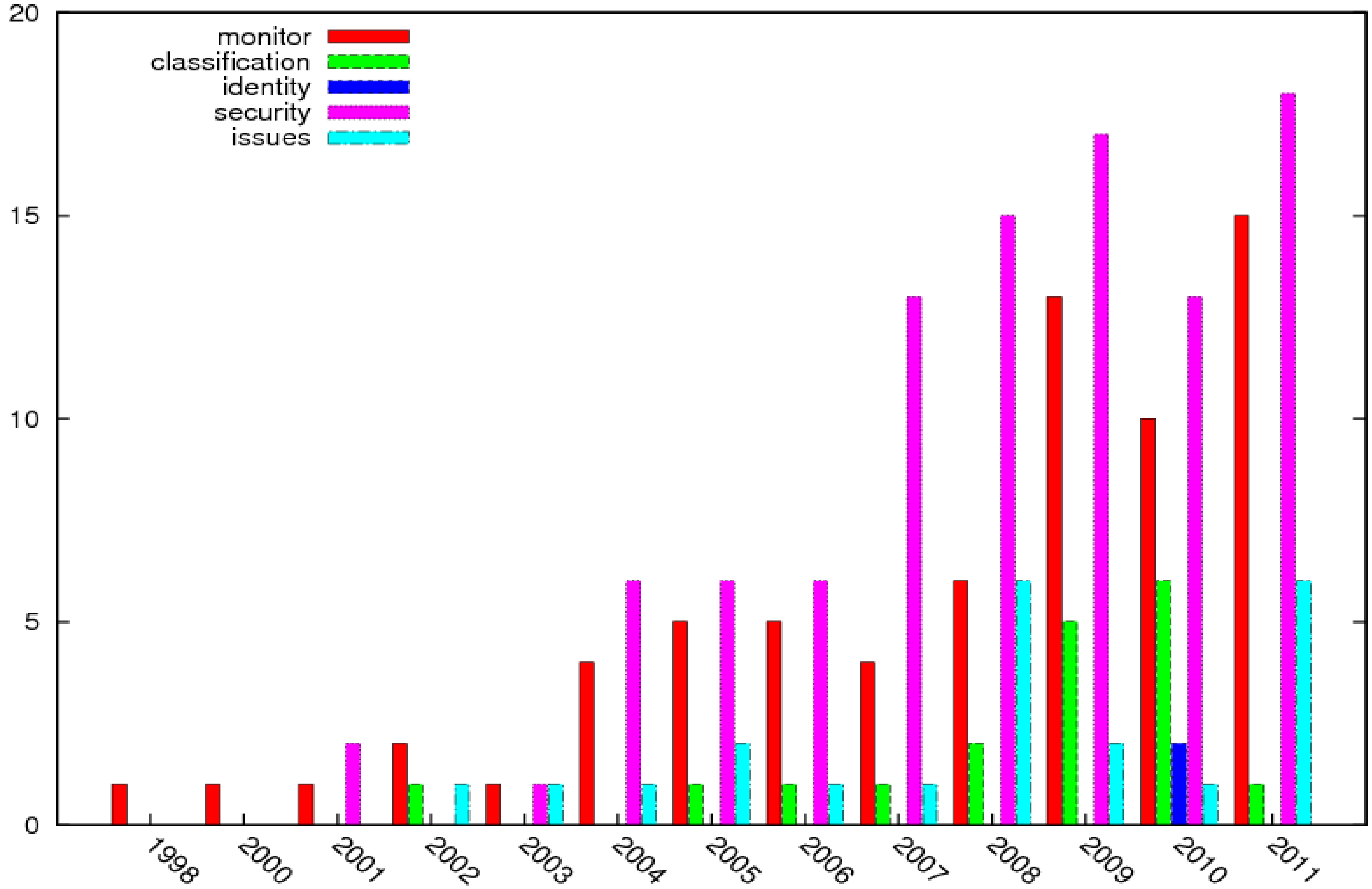
- Review
- Challenges
- Apache Hadoop Related Technologies
- System Design
- Demonstration
- Thoughts and Pitfalls
- Summary

Publications By Years



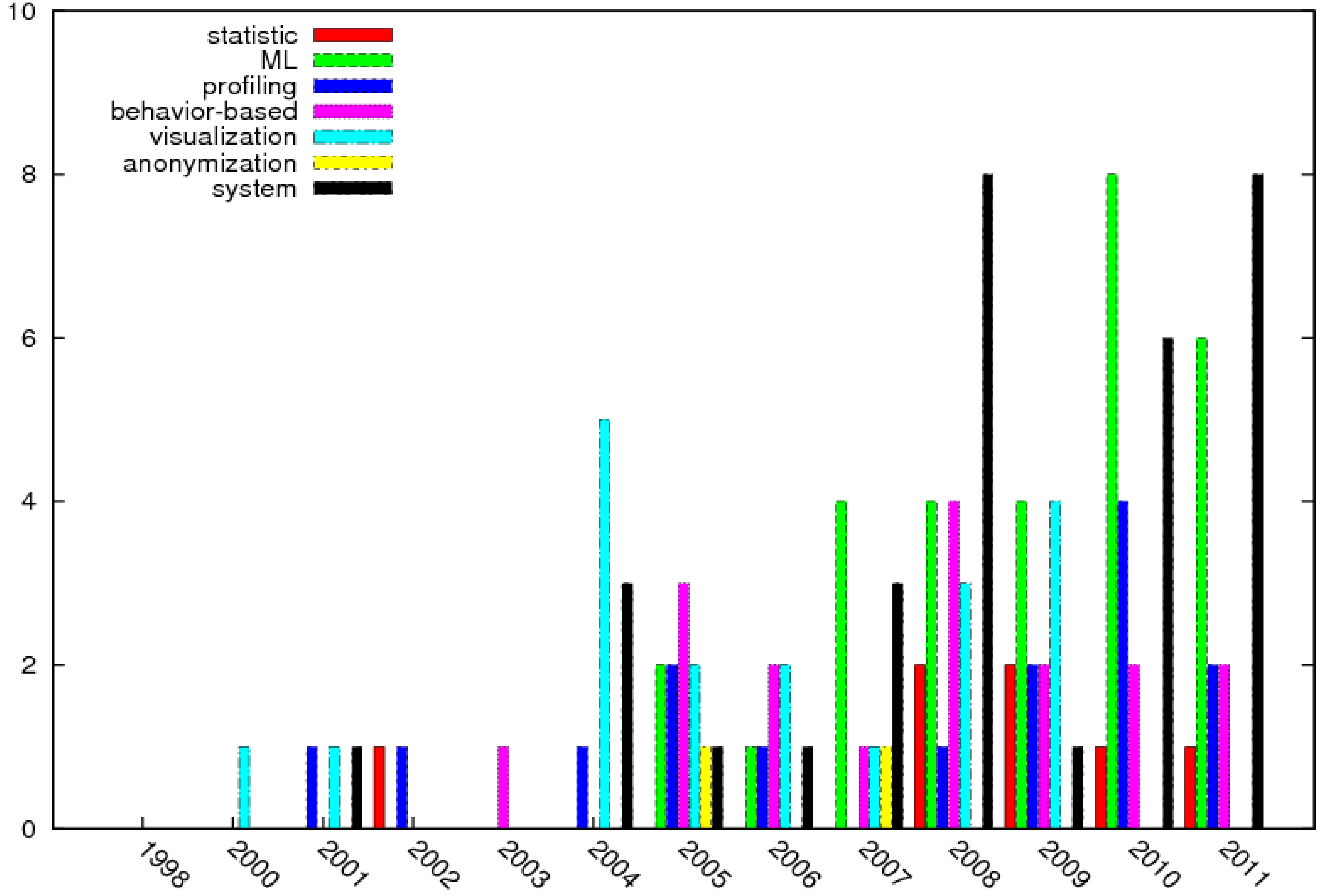
Bingdong Li, Jeff Spinger, George Bebis, Mehmet Hadi Gunes, A Survey of Network Flow Applications, Journal of Networks and Computer Applications (accepted).

Research Perspectives By Years



Bingdong Li, Jeff Spinger, George Bebis, Mehmet Hadi Gunes, A Survey of Network Flow Applications, Journal of Networks and Computer Applications (accepted).

Methods By Years



Bingdong Li, Jeff Spinger, George Bebis, Mehmet Hadi Gunes, A Survey of Network Flow Applications, Journal of Networks and Computer Applications (accepted).

Challenges

- Too much data (volume)
- Real Time and On Demand (velocity)
- Various types/sources of data (variety)
- Changing requirements(variability)

Big Data – Volume, Velocity, Variety (Gartner's Doug Laney) ,

Variability (Forrester's James Kobielus G. etc.)

http://blogs.sas.com/content/datamanagement/2011/11/05/big-data-defined-its-more-than-hadoop/

.

Apache Hadoop Related Technologies

- **What is Apache Hadoop?**

  Open source, storing and processing Big Data

- **Main Systems:**
  - ➢Hadoop Distributed File System (HDFS)
  - ➢MapReduce

Apache Hadoop Related Technologies

- **Data collection:**

  Flume, Chukwa, …
- **Storage:**

  HDFS, Cassandra, CouchDB, …
- **Processing:**

  MapReduce, Pig, Hive, Mahout …
- …

Design

- **Goals**

- **Philosophy**

- **Components**
  - Data Collecting
  - Data Storage
  - Data Schema
  - Data Process
  - User Interfaces

Design Goals

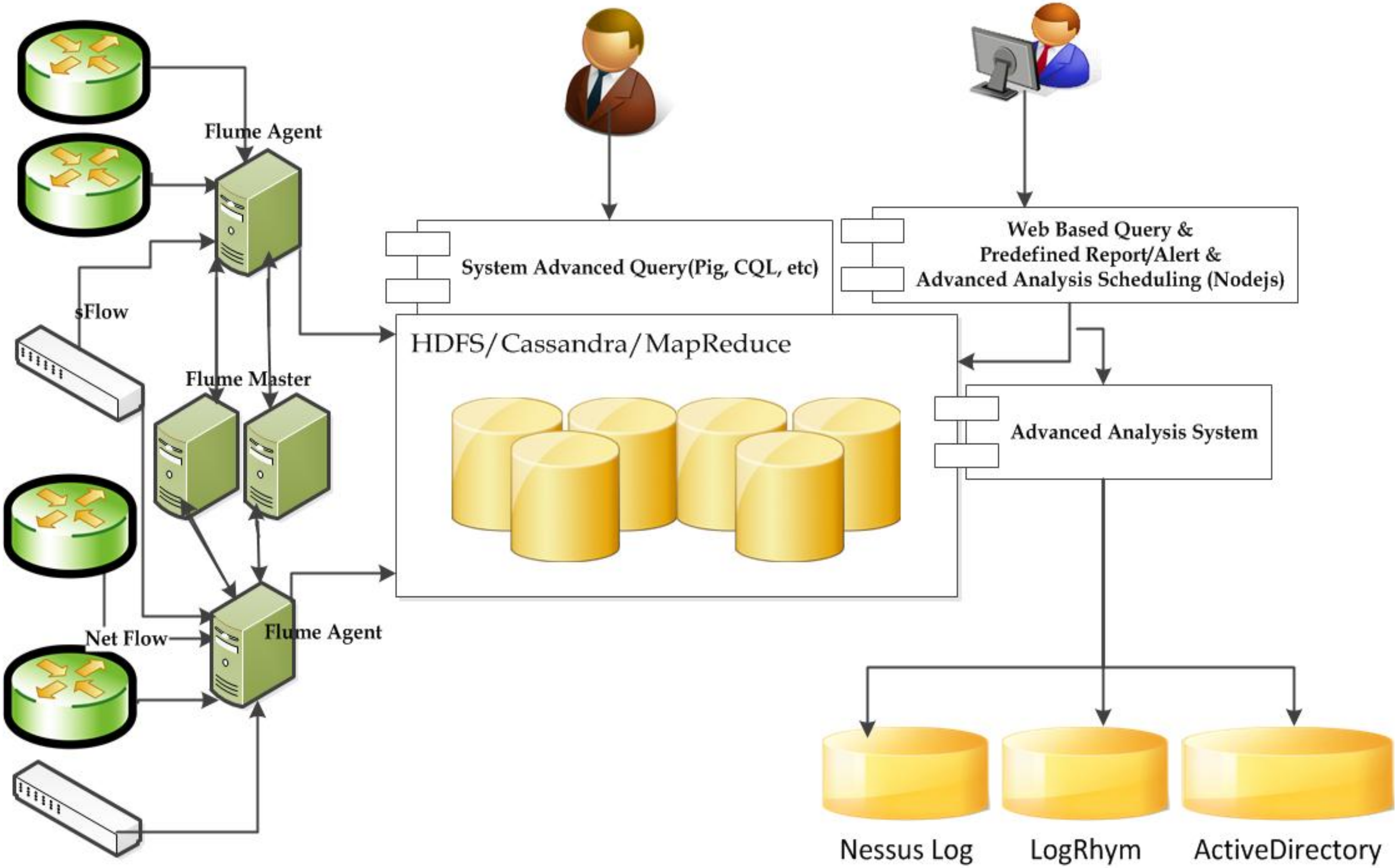
- Real time network query, near real time measurement and analysis
- Distributed system for data collecting, storing, accessing, measuring and analyzing NetFlow and other log data
- Models of detection and classification based on profiling and behavior

Design Philosophy

- Leverage existing technologies

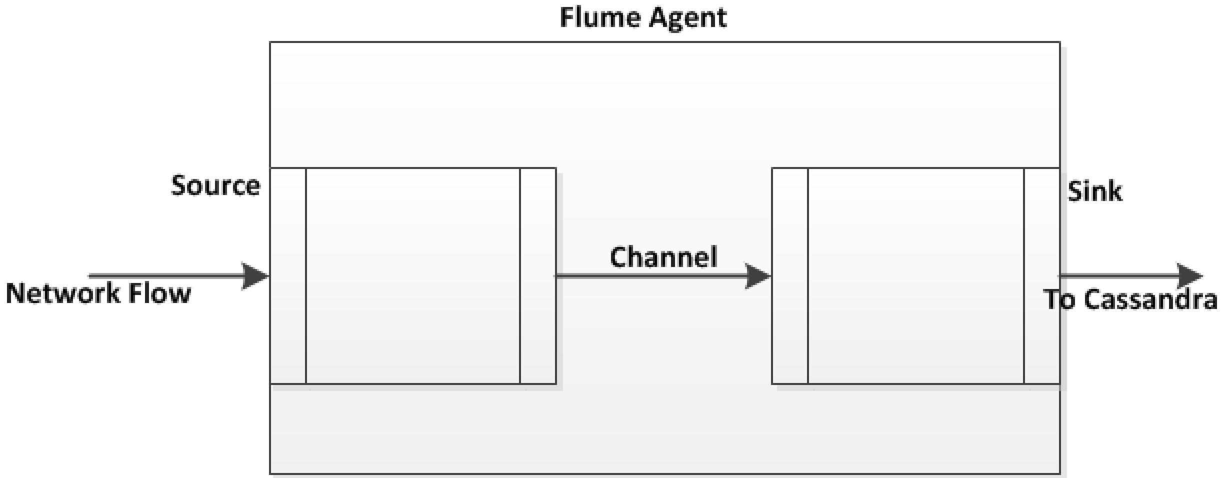
- Modeling known objects rather than unknown objects
  - or use white list rather than black list

Design: Components



Design: Components

- **Flume**: open source collecting, aggregating, and moving data from many different sources to data store
  - **Masters**: keep track all the nodes and inform them
  - **Agents**: Sources accept data, Sinks aggregate and send data, Decorator filter, sample and modify data flow.



Design: Components

C A P Conjecture

A web service can only satisfy any two of

□ **C**onsistency

□ **A**vailability

□ **P**artition Tolerance

Cassandra is AP, arguably CAP with specifying consistency level

Any, one, quorum, local_quorum, each_quorum, ALL

Gilbert, Seth and Lynch, Nancy, Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services, SIGAACT News, 2002

Design: Components

- Cassandra Data Scehma
  - Keyspace
  - Column family
  - Rows and Columns

Design: Components

- Cassandra Index
  - Primary Index (row key)
  - Secondary Index (column values)
  - DIY with wide row or inverted index
  - Composite Column
  - Third party indexing
    - such as ElasticSearch, Solandra, DataStax Enterprise
- Counter

Design: Components

- Data Processing
  - Query network by CQL, or Web UI (Nodejs)
  - Network measurement by Pig scripting, R
  - Advanced data mining and network modeling by programming written by C++ and Java
  - Scheduling tasks

Design: Components

- User Interface
  - **Web User**:
    - through a secure internal web page to
      - see reports,
      - schedule advanced analysis tasks
  - **Advanced System User**:
    - use cassandra-cli, CQL, Pig, and R to do advanced measurement and analysis

Design: Features

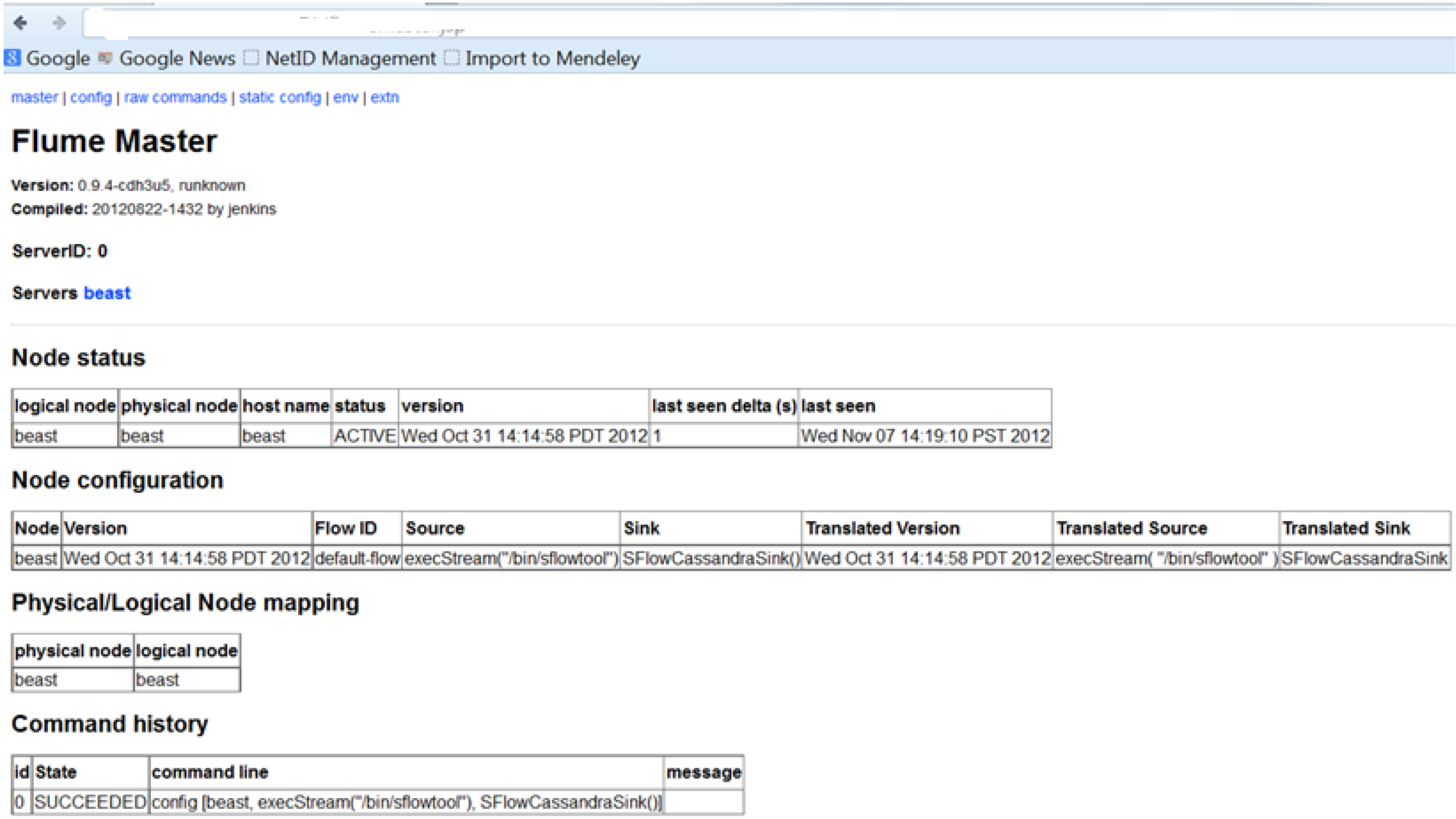
- Query Network Status

- Network Measurement

- Advanced Network Modeling
  - ➢Host Role's Behavior
  - ➢Roles of Subnet Behavior
  - ➢User Behaviors of Hosts

Demonstration

Flume



master | config | raw commands | static config | env | extn

Flume Master

Version: 0.9.4-cdh3u5, runknown
Compiled: 20120822-1432 by jenkins

ServerID: 0

Servers beast

Node status

logical node	physical node	host name	status	version	last seen delta (s)	last seen
beast	beast	beast	ACTIVE	Wed Oct 31 14:14:58 PDT 2012	1	Wed Nov 07 14:19:10 PST 2012

Node configuration

Node	Version	Flow ID	Source	Sink	Translated Version	Translated Source	Translated Sink
beast	Wed Oct 31 14:14:58 PDT 2012	default-flow	execStream("/bin/sflowtool")	SFlowCassandraSink()	Wed Oct 31 14:14:58 PDT 2012	execStream("/bin/sflowtool")	SFlowCassandraSink

Physical/Logical Node mapping

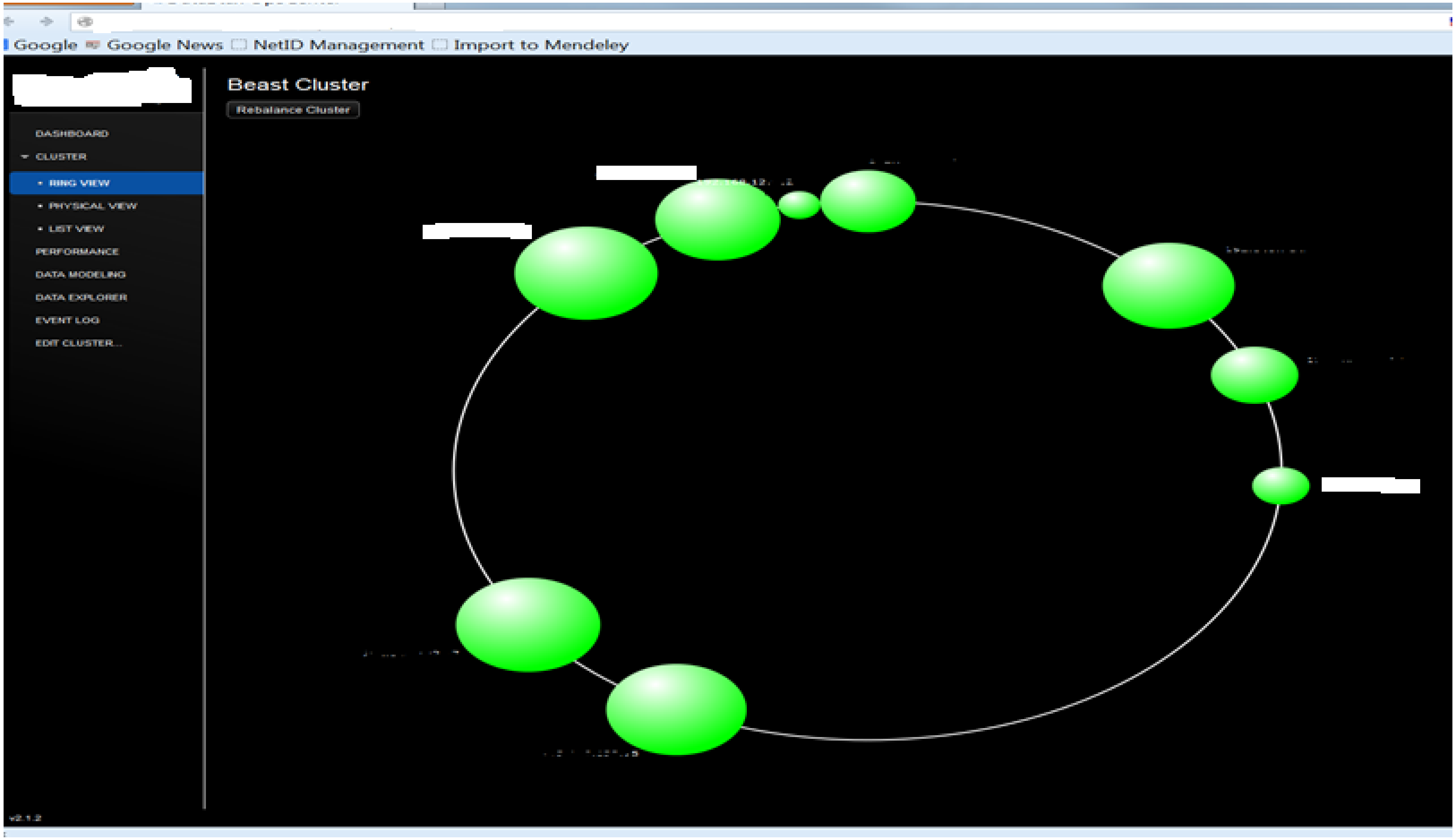
physical node	logical node
beast	beast

Command history

id	State	command line	message
0	SUCCEEDED	config [beast, execStream("/bin/sflowtool"), SFlowCassandraSink()]	

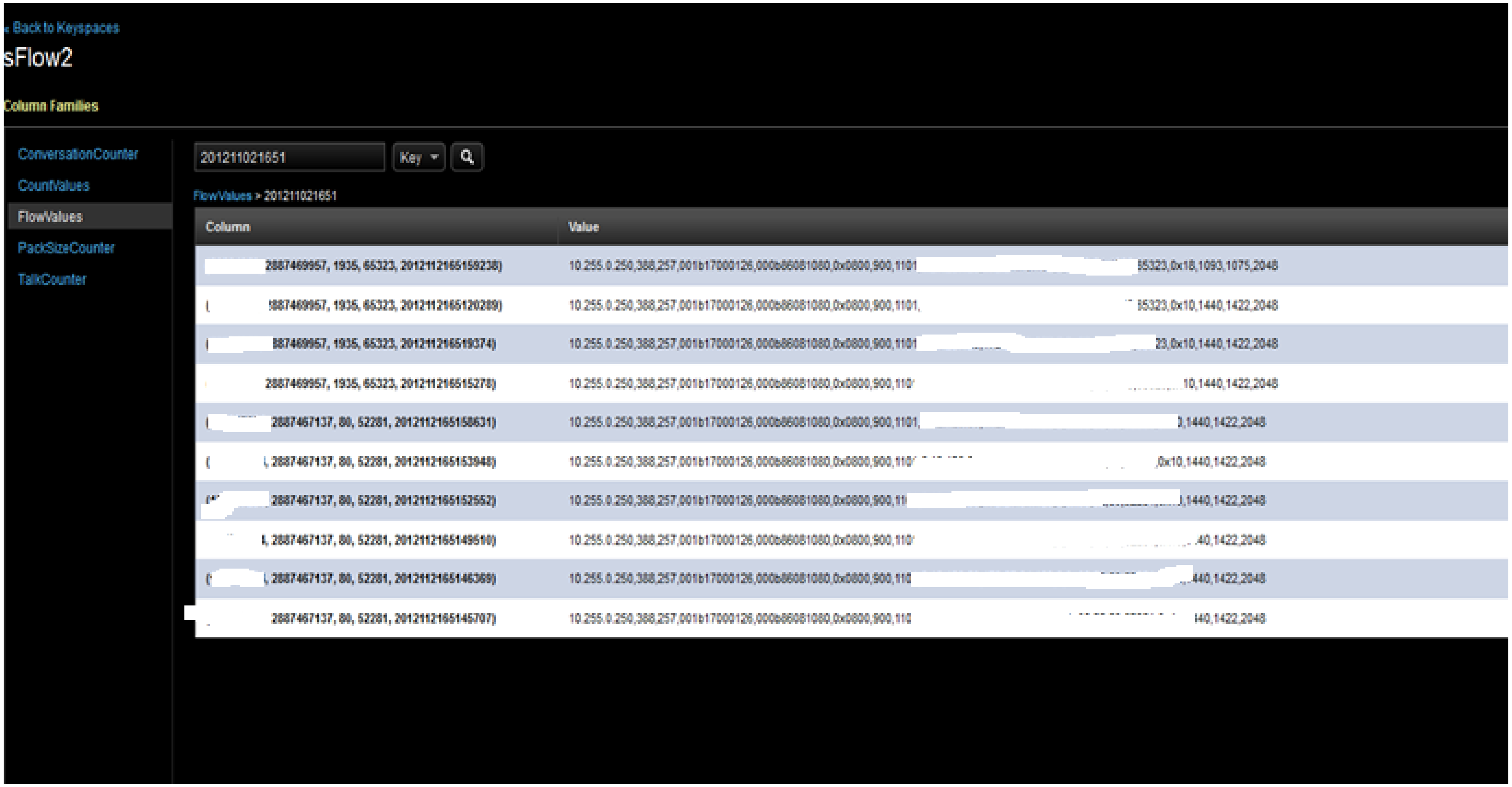
Demonstration

Cassandra Cluster



Demonstration

- Query by Key



« Back to Keyspaces

sFlow2

Column Families

- ConversationCounter
- CountValues
- FlowValues
- PackSizeCounter
- TalkCounter

201211021651 | Key ▾ | 🔍

FlowValues > 201211021651

Column	Value
[redacted] 2887469957, 1935, 65323, 2012112165159238)	10.255.0.250,388,257,001b17000126,000b86081080,0x0800,900,1101[redacted]65323,0x18,1093,1075,2048
([redacted] 887469957, 1935, 65323, 2012112165120289)	10.255.0.250,388,257,001b17000126,000b86081080,0x0800,900,1101,[redacted]65323,0x10,1440,1422,2048
([redacted] 887469957, 1935, 65323, 201211216519374)	10.255.0.250,388,257,001b17000126,000b86081080,0x0800,900,1101[redacted]23,0x10,1440,1422,2048
[redacted] 2887469957, 1935, 65323, 201211216515278)	10.255.0.250,388,257,001b17000126,000b86081080,0x0800,900,110[redacted]10,1440,1422,2048
([redacted] 2887467137, 80, 52281, 2012112165158631)	10.255.0.250,388,257,001b17000126,000b86081080,0x0800,900,1101,[redacted]0,1440,1422,2048
([redacted], 2887467137, 80, 52281, 2012112165153948)	10.255.0.250,388,257,001b17000126,000b86081080,0x0800,900,110[redacted],0x10,1440,1422,2048
([redacted] 2887467137, 80, 52281, 2012112165152552)	10.255.0.250,388,257,001b17000126,000b86081080,0x0800,900,11[redacted],1440,1422,2048
[redacted], 2887467137, 80, 52281, 2012112165149610)	10.255.0.250,388,257,001b17000126,000b86081080,0x0800,900,110[redacted]40,1422,2048
([redacted], 2887467137, 80, 52281, 2012112165146369)	10.255.0.250,388,257,001b17000126,000b86081080,0x0800,900,110[redacted]440,1422,2048
[redacted] 2887467137, 80, 52281, 2012112165145707)	10.255.0.250,388,257,001b17000126,000b86081080,0x0800,900,110[redacted]40,1422,2048

Demonstration

- Measuring anonymity network usage on campus by using Pig scripting

It takes less than 10 minutes to process 205 million packets, about 1.44TB data, writing less than 200 lines of Pig scripting code.

Bingdong Li, Esra Edrin, Mehmet Hadi Gunes, George Bebis, Todd Shipley, A Study of Anonymity Technology Usage on the Internet, submitted to Computer Communicatior

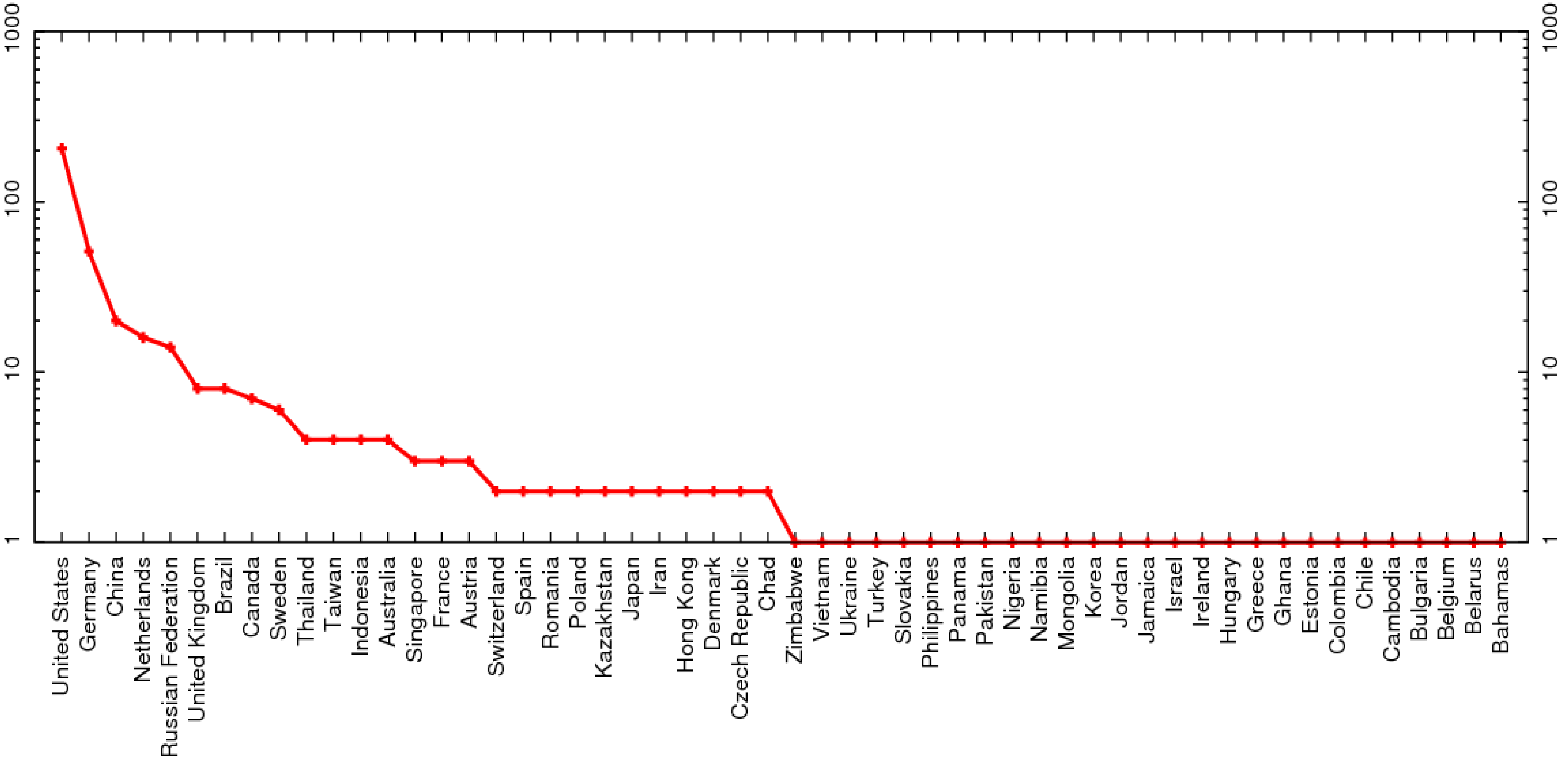
Demonstration

Analyzed Anonymity Networks

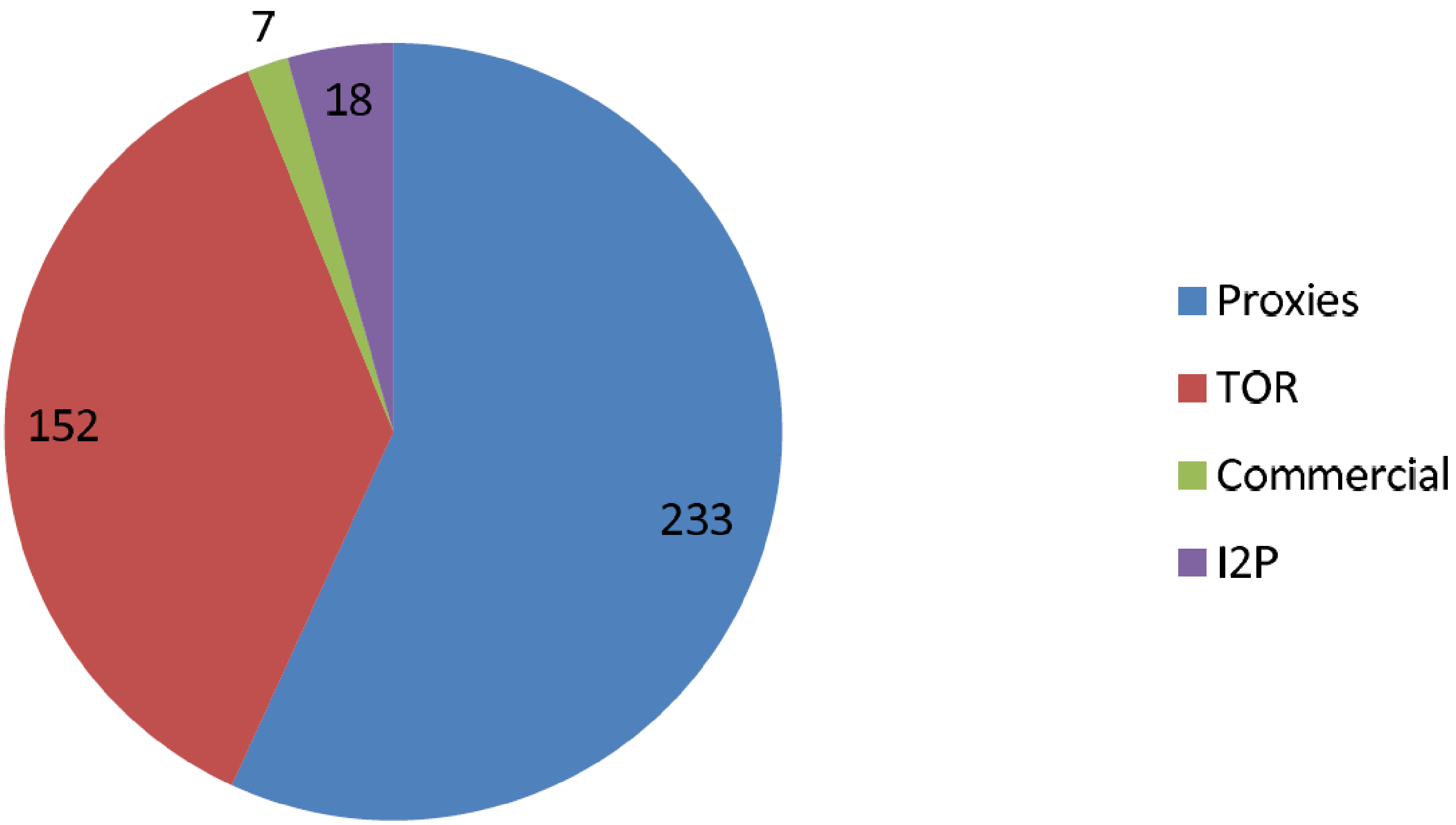
Network	Servers	Service
Tor	61,798	General
I2P	2,267	P2P
JAP	11	General
Remailers	15	Email
Proxies	7,246	General
Commercial	Anomymizer,Gotrusted	General

Bingdong Li, Esra Edrin, Mehmet Hadi Gunes, George Bebis, Todd Shipley, A Study of Anonymity Technology Usage on the Internet, submitted to Computer Communicatio

Anonymity Network Usage Geolocation



Anonymity Network Usage Distribution



Demonstration

- Example of Advanced Network Modeling
  - ➢Model Host Role’s Behaviors

Algorithms:

On-line SVM based on Bordes Methods

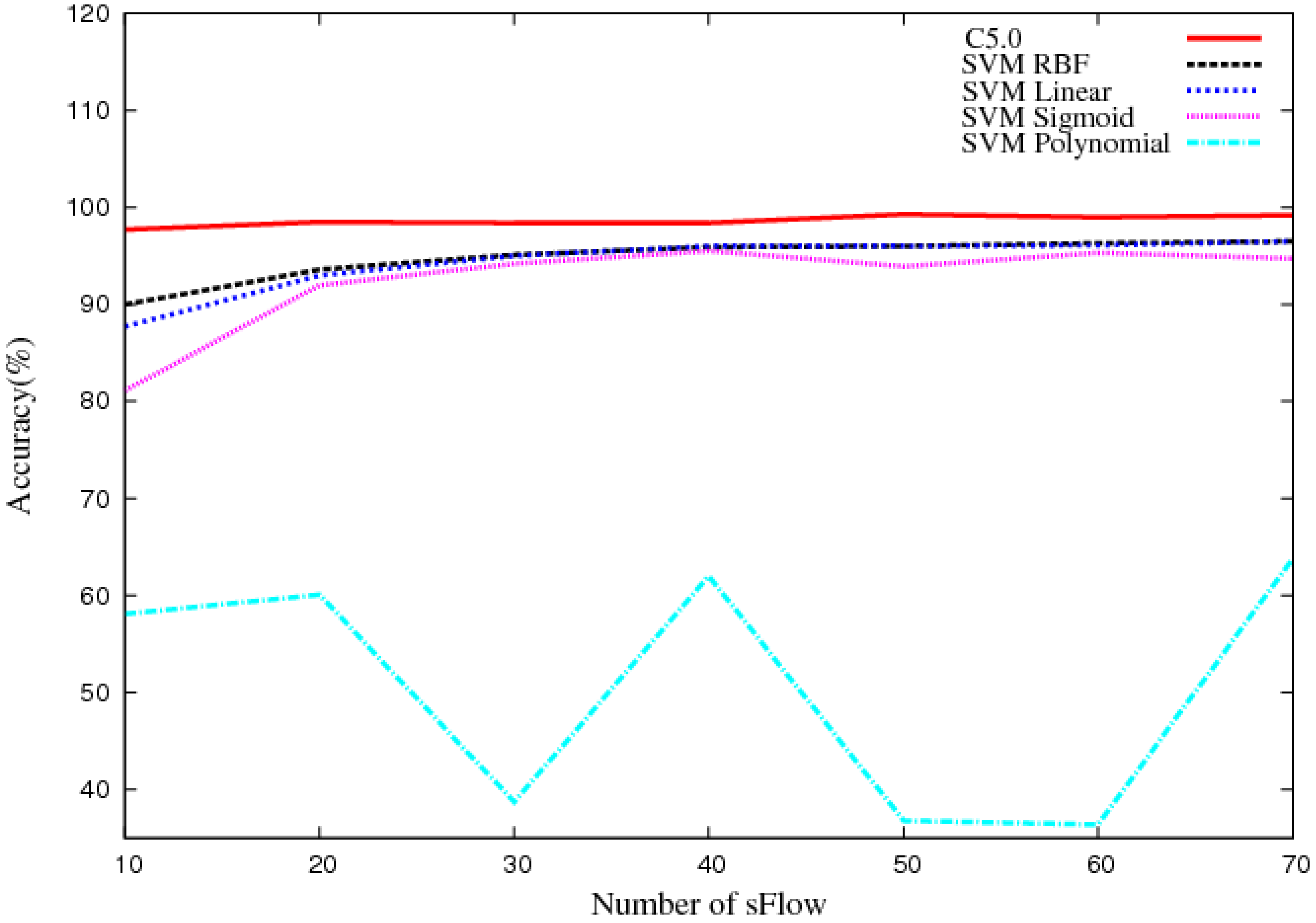
Ground Truth:

Host Information in Active Directory and vulnerability scanner Nessus database.

Antoine Bordes, etc. Fast kernel classifiers with online and active learning. Journal of Machine Learning Research, 6:1579–1619, September 2005.

Demonstration

Client vs Server Classification Accuracy



Thoughts and Pitfalls

- Low Cost – Open Source, Distributed
- Be patient and careful for Incompatibility between different versions of components
- Be willing to learn, it is a new era of big data
- Cassandra Replica Factor = 1? Do not even try
- What do you do for Exception error? Handle, Ignore or throw it

Summary

- A design of distrusted real time network security system based on Apache Hadoop related technologies
- Demonstration
- Thoughts and pitfalls

Questions and Discussions

Contact:
Bingdong Li
bingdongli@unr.edu