Big Data Analytics

Radu State

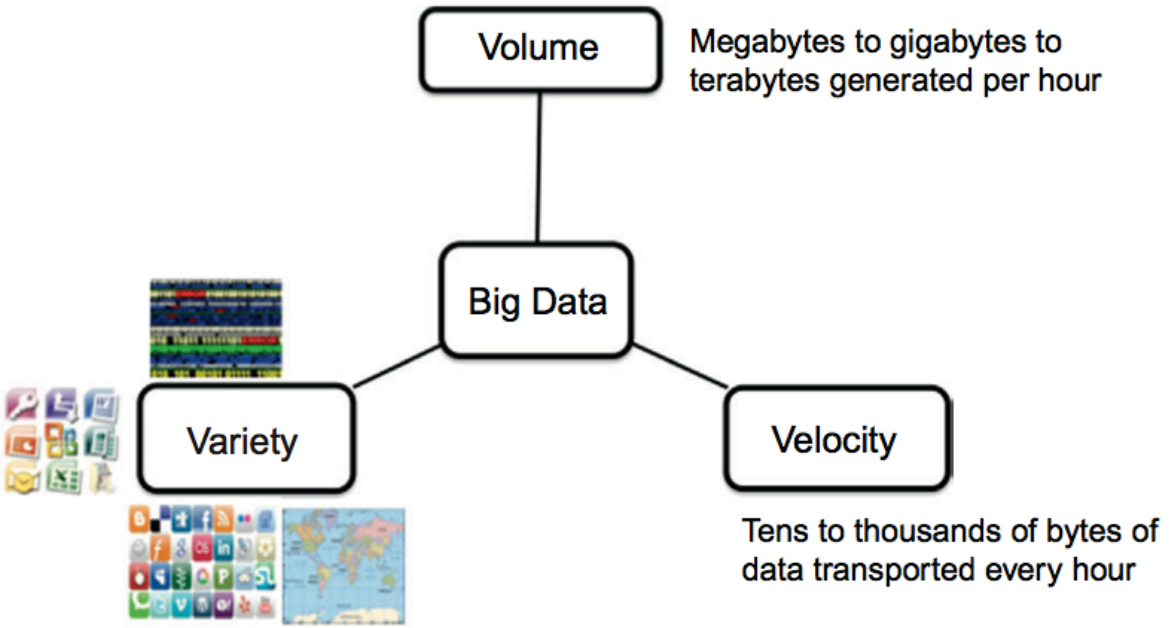
Short Bio

- Master of Science in Engineering, Johns Hopkins University, USA
- Senior Researcher with INRIA, France
- Professor, (University of Nancy 1)
- Research Scientist, SnT/UL in the Netlab research group

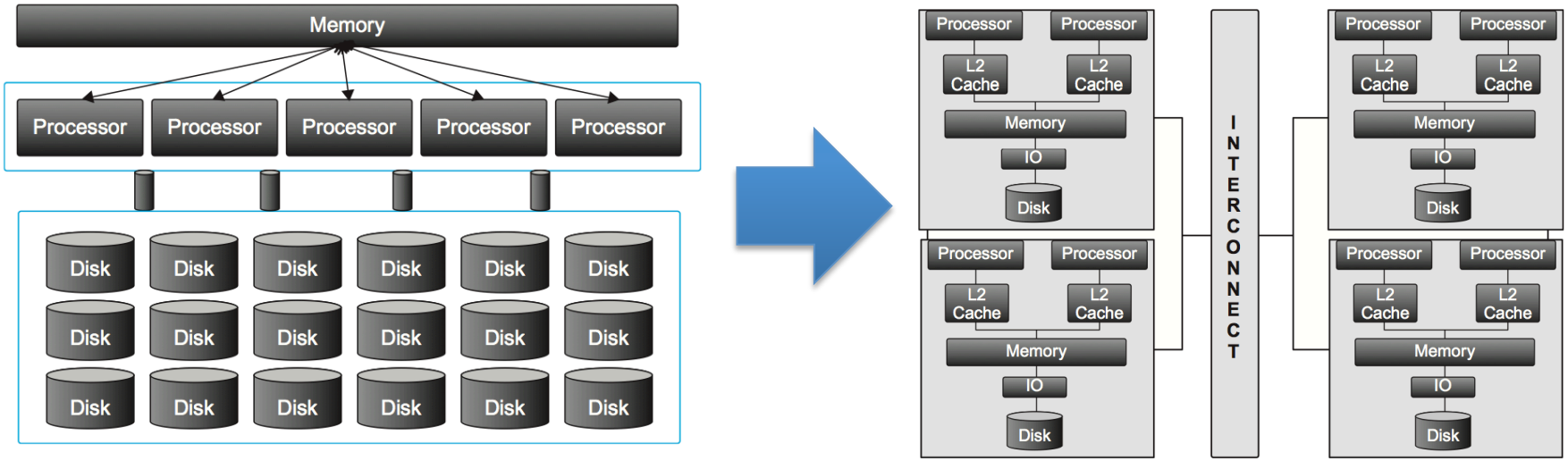
Outline

- What is Big Data
- Big Data Analysis in the team
  - Call Details record at a country level
  - Analysis of Network traffic – research done at SnT

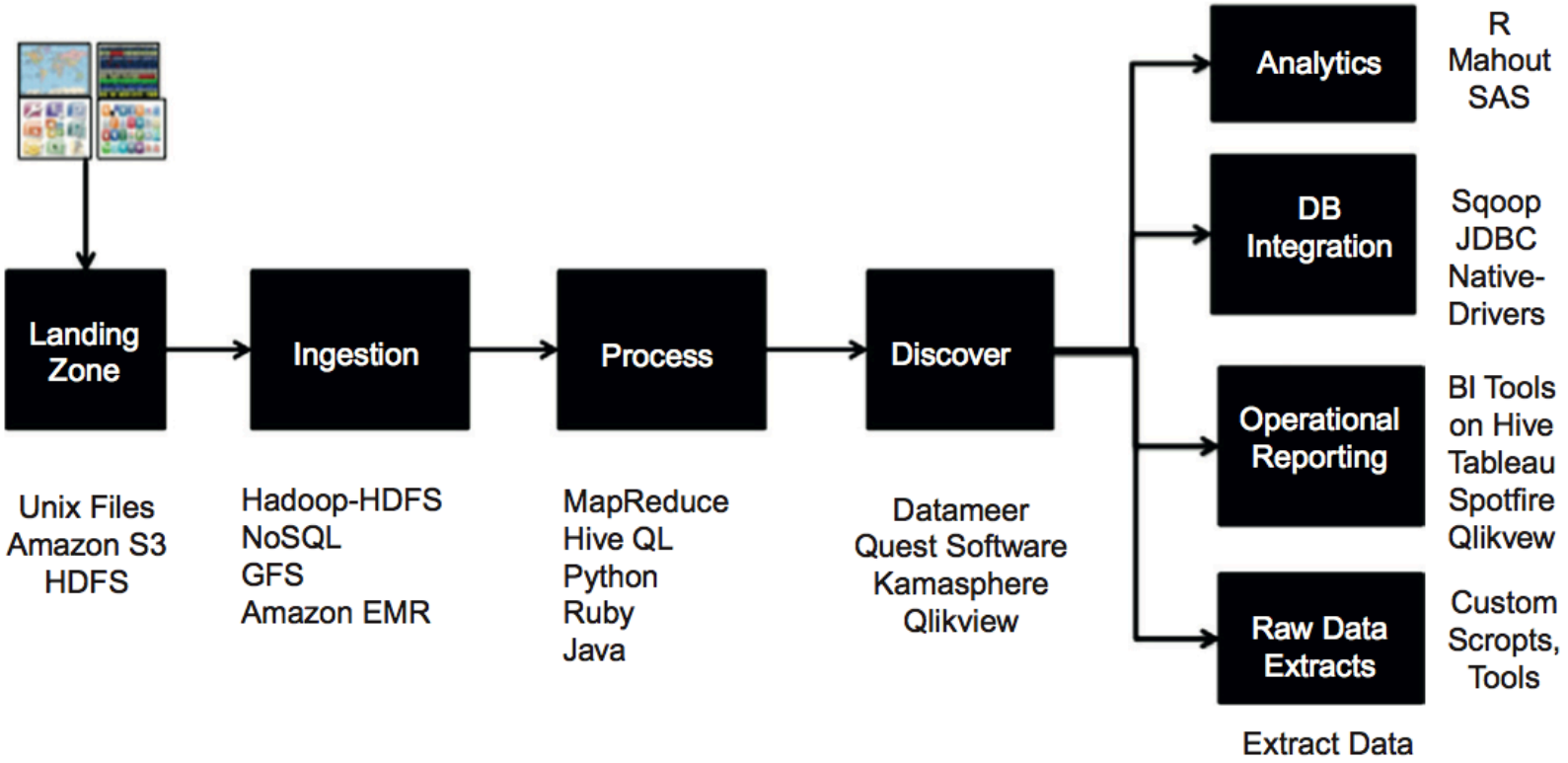
Big Data at a glance



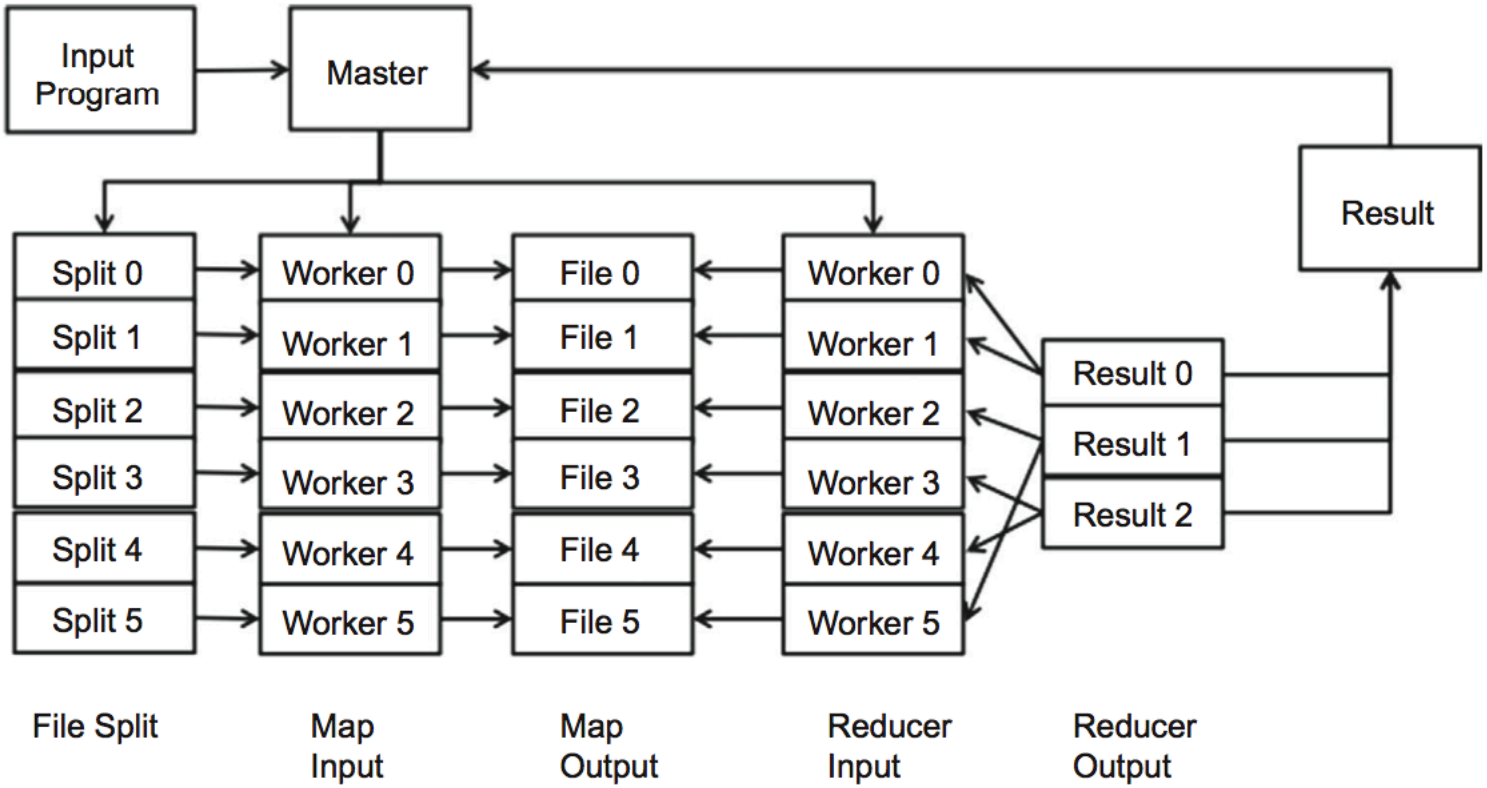
What are Big Data Architectures ?



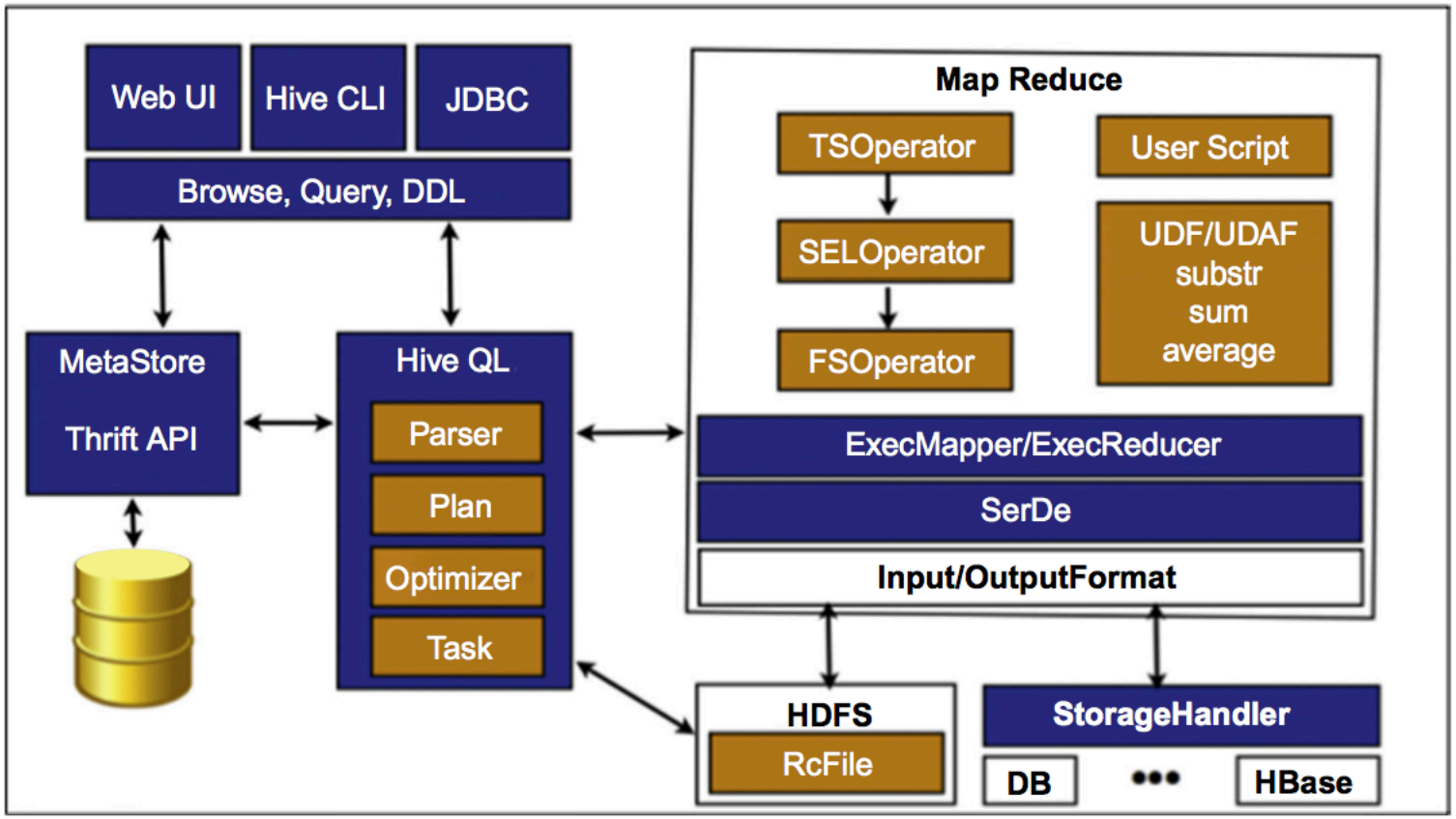
Phases in Big Data Processing



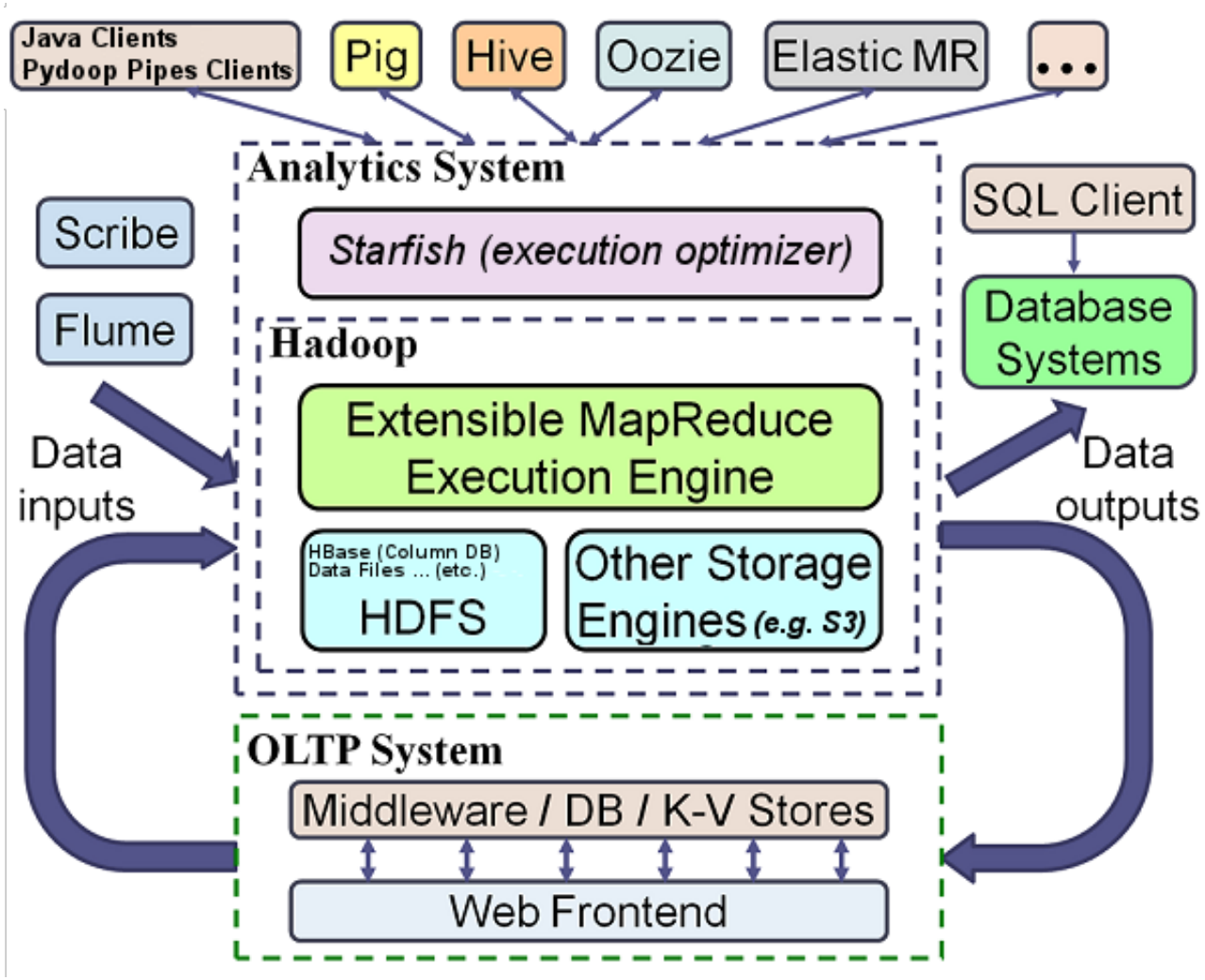
Map-Reduce – the holy grail in Big Data



Big Data is more then simple HPC



The current Eco-System



Anomaly Detection in large scale CDR records

David Goergen, Radu State, Thomas Engel and Veena MENDIRATTA (Bell Labs, USA)

- One country : Ivory Cost
- Time Period: 01.12.2011 to 28.04.2012
- 5 million users
- 1124 base stations (for mobile communications)
- More then 3 billions entries summarizing on a hourly basis the SMS and Voice Calls
- 50000 mobile users tracked over these months with GPS and call records

What happened in the Ivory Cost in 2012 ?

Laurent Gbagbo appears before international criminal court

Ivory Coast's ex-president complains of transfer to The Hague and blames France for arrest during hearing opening

David Smith, Africa correspondent
The Guardian, Monday 5 December 2011 18.30 GMT

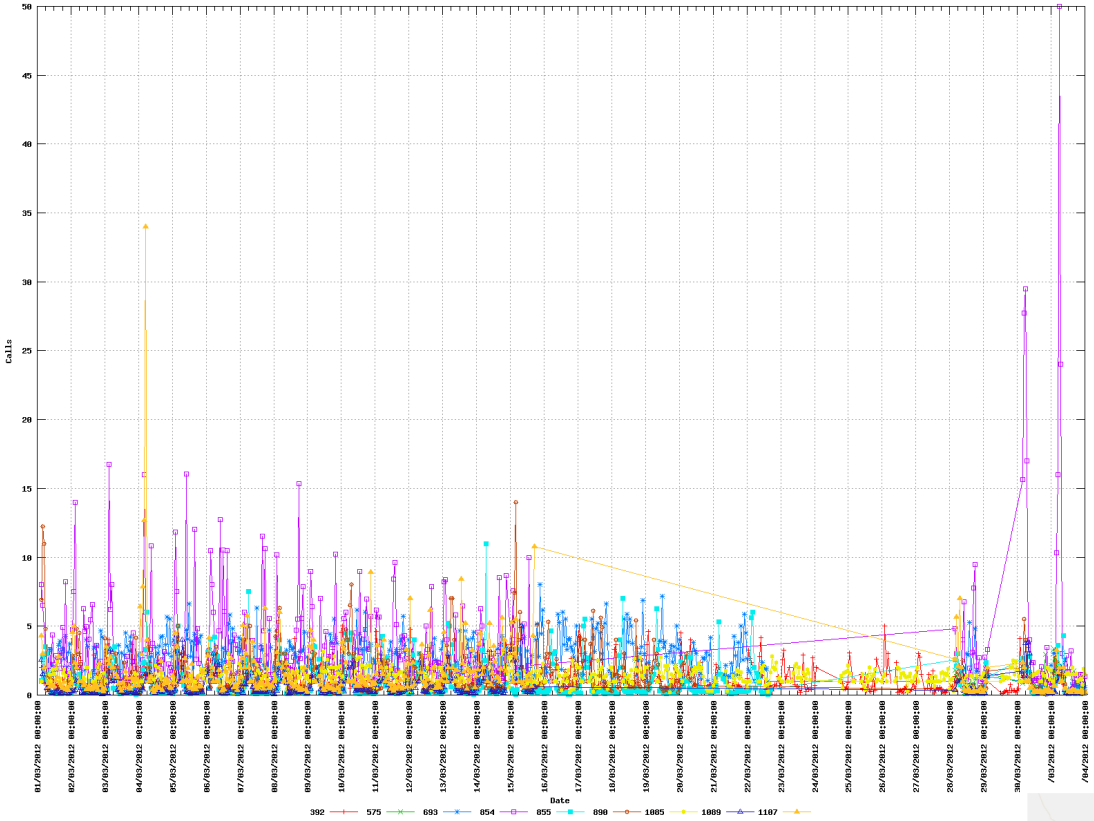
APRIL 30, 2012

Email

(Bamako) – Separatist Tuareg rebels, Islamist armed groups, and Arab militias who seized control of northern Mali in April 2012 have committed numerous war crimes, including rape, use of child soldiers, and pillaging of hospitals, schools, aid agencies, and government buildings, Human Rights Watch said today. An Islamist armed group has summarily executed two men, amputated the hand of at least one

Enlarge
Malian junta soldiers patrol in Kati, outside Bamako.
© 2012 Reuters

The silent base stations……



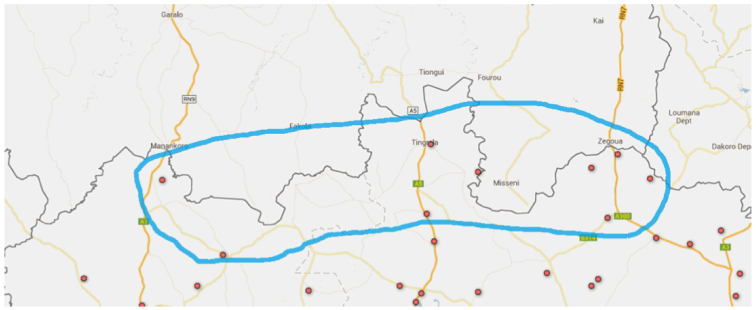
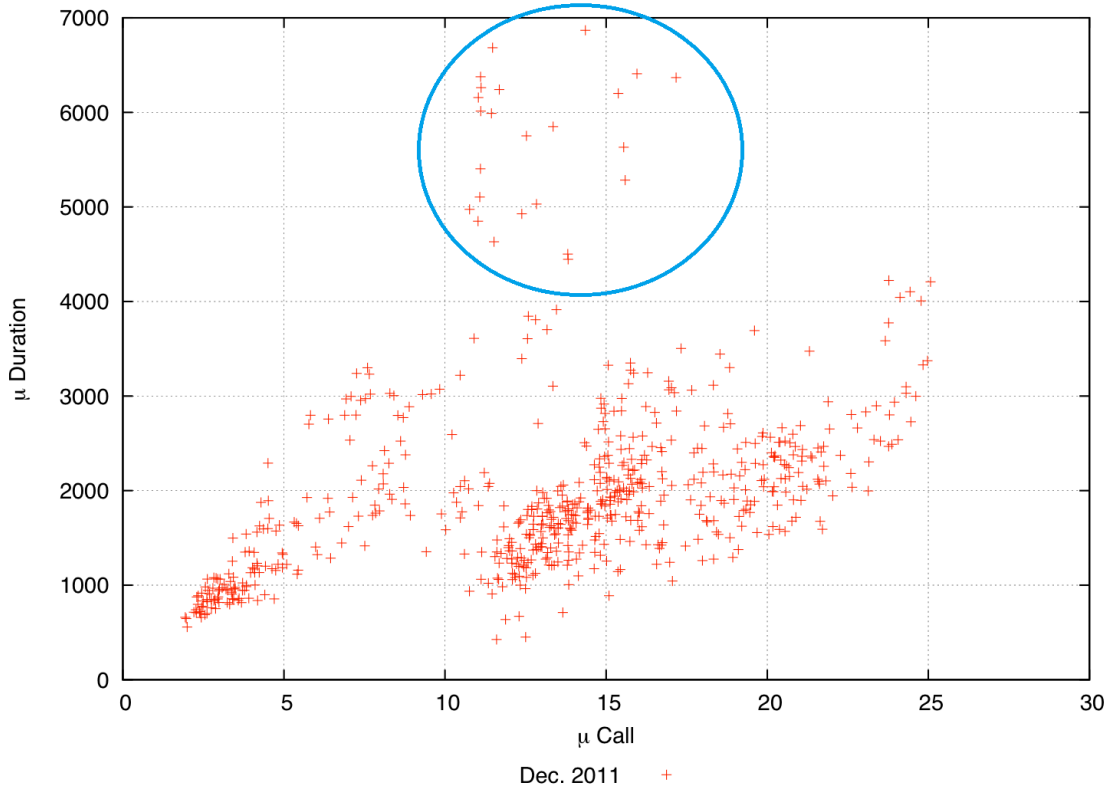
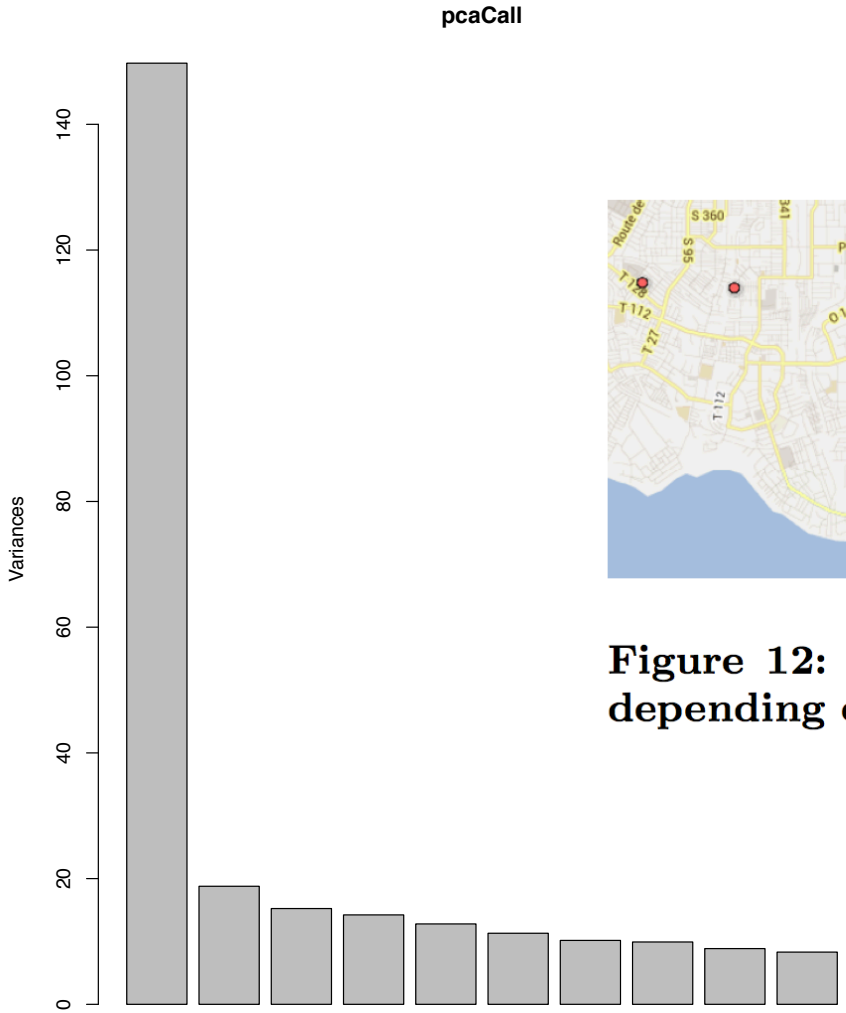


Figure 4: Antenna near the Malian border

Strange calling behaviors…..at 2 AM.



Where is most variability in the data ?
PCA analysis on the duration



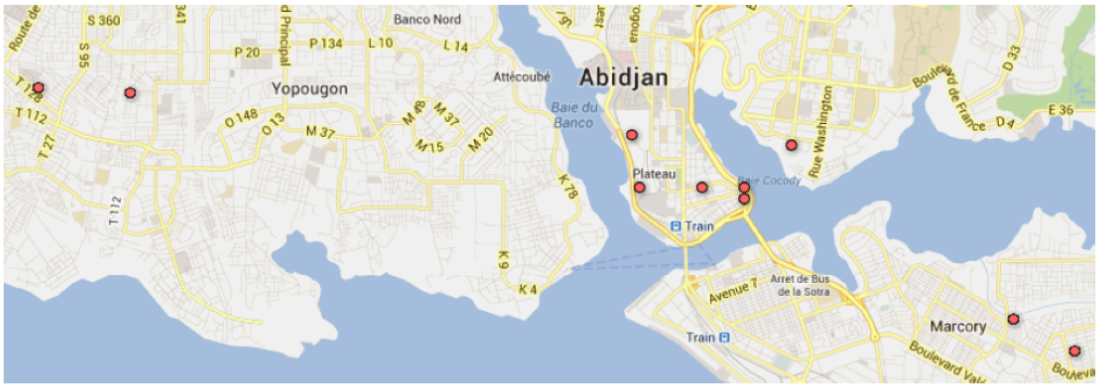
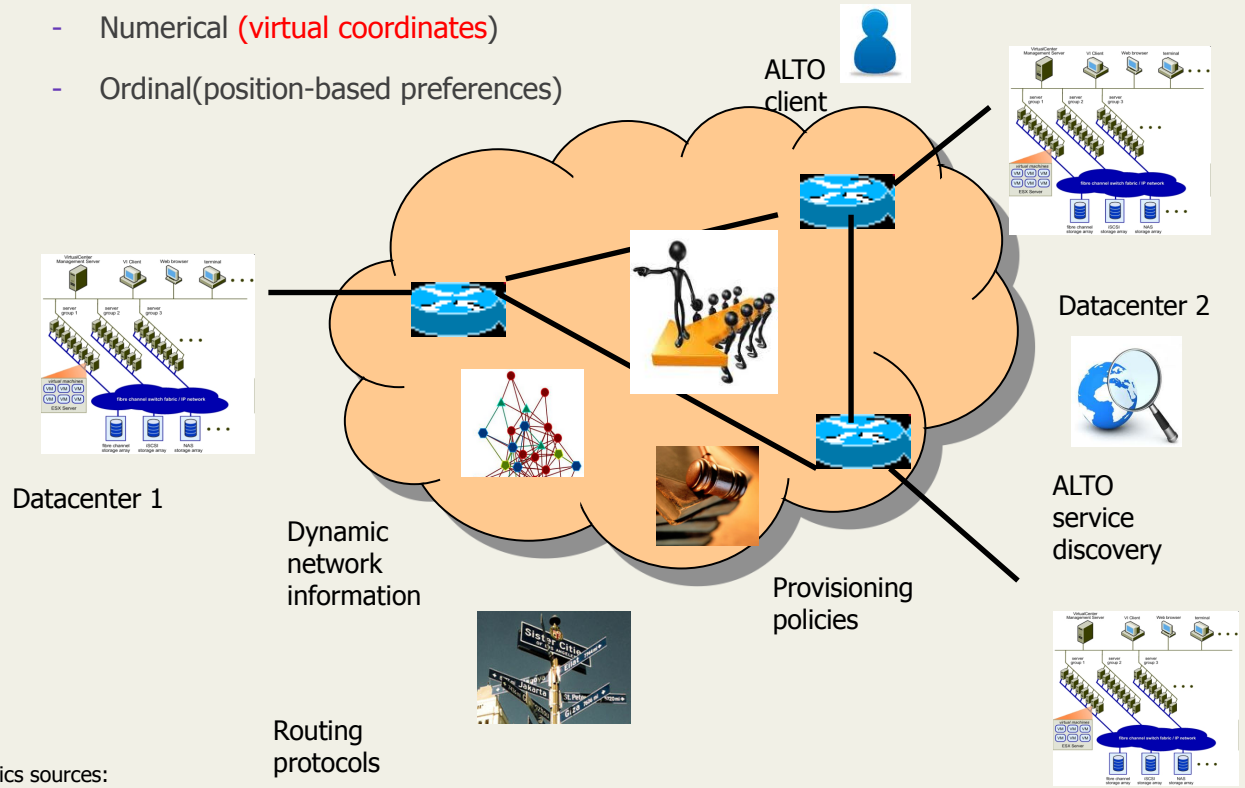


Figure 12: Top 10 antennas symbolized by circles depending on PC1

Cloud and Service Management

- ALTO solves the general rendezvous problem: Given a choice of resources, which one is the best candidate?
  - Normalized costs: Type: W*hat* does the cost represent? (Air-miles, hop count, ...)
  - Numerical (virtual coordinates)
  - Ordinal(position-based preferences)



Graphics sources:
http://pubs.vmware.com/vi301/intro/images/Introduction_chapter.3.2.1.jpg

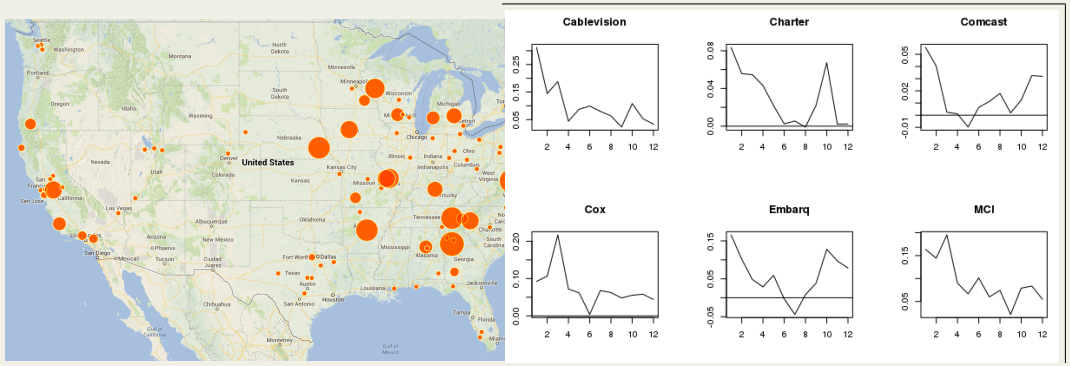
2 main abstractions:

- Network Map
- Cost Map

Network specified in terms of Partition/Provider ID (PID): aggregation of endpoints identified by a provider-defined network location identifier.

Optimization of service provisioning (Cloud, CDNs) over large ISPs and networks

- FCC Dataset specification (200 GB/year)
  - FCC has embarked on a nationwide performance study of residential wireline broadband service
  - Aim is to use the raw datasets from this study for analysis and to create ALTO topology map and a cost map from this dataset
  - Using a canonical Map-Reduce (Big Data) computational paradigm on a Hadoop cluster
- Major outcomes
  - Financial modeling with Sharpe ratios to build third parties ALTO maps that takes both bandwidth and latency into account



$$S_p = \frac{\bar{r_p} - R_f}{\sigma_p}$$

$$\forall i \in PID, C_u^i = \sum_{m=1}^{12} S_m^i * W_i$$

	Cablevision	Charter	Comcast	Cox	Embarq	MCI	Mediacom	TimeWarner	Windstream	Default
Cablevision	0.00	1.67	1.81	1.10	1.41	0.86	1.32	0.32	1.38	2.00
Charter	0.76	0.00	1.81	1.10	1.41	0.86	1.32	0.32	1.38	2.00
Comcast	0.76	1.67	0.00	1.10	1.41	0.86	1.32	0.32	1.38	2.00
Cox	0.76	1.67	1.81	0.00	1.41	0.86	1.32	0.32	1.38	2.00
Embarq	0.76	1.67	1.81	1.10	0.00	0.86	1.32	0.32	1.38	2.00
MCI	0.76	1.67	1.81	1.10	1.41	0.00	1.32	0.32	1.38	2.00
Mediacom	0.76	1.67	1.81	1.10	1.41	0.86	0.00	0.32	1.38	2.00
TimeWarner	0.76	1.67	1.81	1.10	1.41	0.86	1.32	0.00	1.38	2.00
Windstream	0.76	1.67	1.81	1.10	1.41	0.86	1.32	0.32	0.00	2.00

Relevant Publications and submissions

- David Goergen, Veena B. Mendiratta (Bell Labs, USA), Radu State Thomas Engel. Identifying abnormal pattern in cellular communication flows (IPTCOMM 2013)
- David Goergen, Vijay K. Gurbani (Bell Labs, USA), Radu State: Of maps and costs: Aggregating large-scale broadband measurements for the Application Layer Traffic Optimization (ALTO) protocol. (RTC 2013)
- David Goergen, Vijay Gurbani (Bell Labs, USA), Radu State Thomas Engel. Making historical connections: Building Application Layer Traffic Optimization (ALTO) network and cost maps from public broadband data (submitted to CNSM 2014)

Big Data and Security

- Research Questions:
  - Securing Big Data architectures
    - Which data is accessed and processed ?
    - What happens to the outcome ?
    - How to protect data in motion ?
  - Big Data approaches for advanced security analytics and advanced Persistent Threat (APT) detection
    - massive data collecting data and event tracking at large scale
    - deeper analytics on unstructured data (honeypots, firewall, DNS);
    - consolidated view of security-related information;
    - real-time analysis of streaming AAA security data in order to profile the human behavior in the loop
    - Predict attacker behavior on other targets

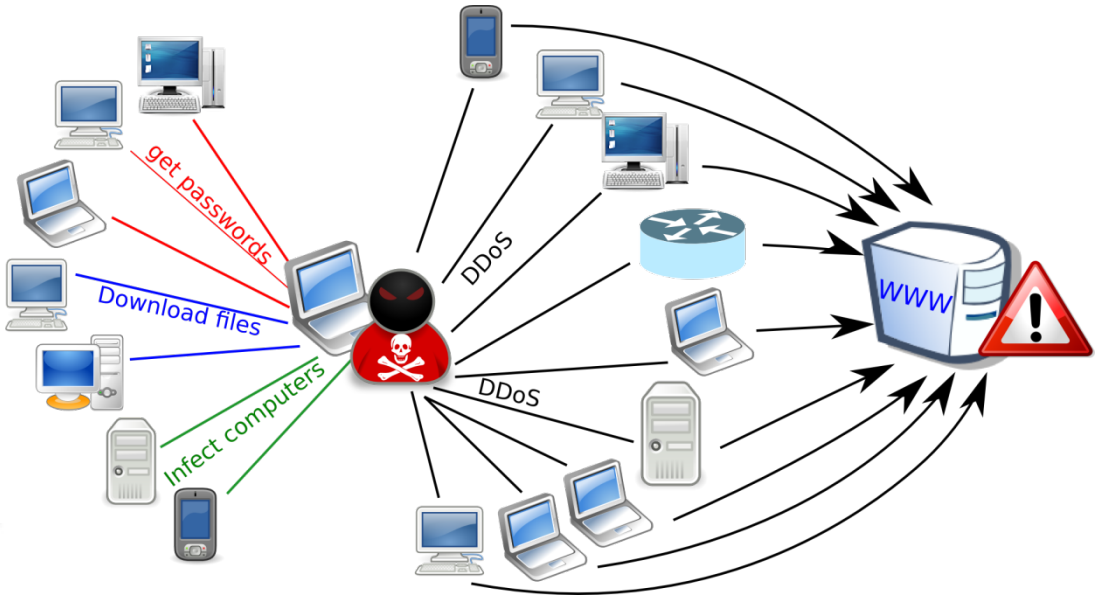
- J. François, S. Wang, R. State, and T. Engel, "BotTrack: Tracking Botnets using NetFlow and PageRank," in IFIP/TC6 NETWORKING 2011, Springer, Ed., Valencia, Spain, May 2011
- Wagner Cynthia, J. François, R. State, and T. Engel, "Machine Learning Approach for IP-Flow Record Anomaly Detection," in IFIP/TC6 NETWORKING 2011, Springer, Ed., Valencia, Spain, May 2011.
- Hommes, Stefan State, Radu Zinnen, Andreas Engel, Thomas. Detection of Abnormal Behaviour in a Surveillance Environment Using Control Charts. IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) (2011), no. 8th, pp. 113-118
- S. Wang, R. State, M. Ourdane T. Engel. Mining NetFlow Records for Critical Network Activities. AIMS 2010: 135-146
- S. Wang, R. State, M. Ourdane T. Engel. FlowRank: ranking NetFlow records. IWCMC 2010: 484-488

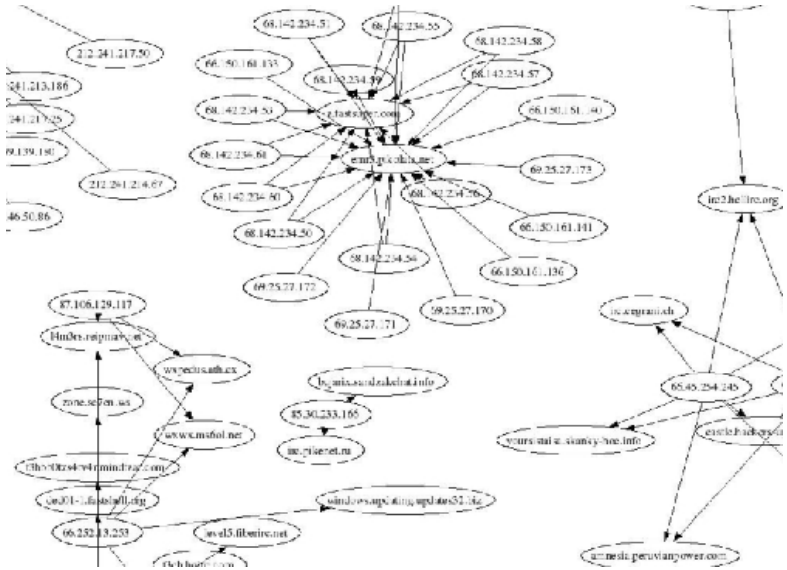
Botnet tracking

Jerome Francois, Radu State, Thomas Engel

Botnets:

- Army of controlled compromised machines
- Powerful attack vector (spam, DDoS, espionage...)





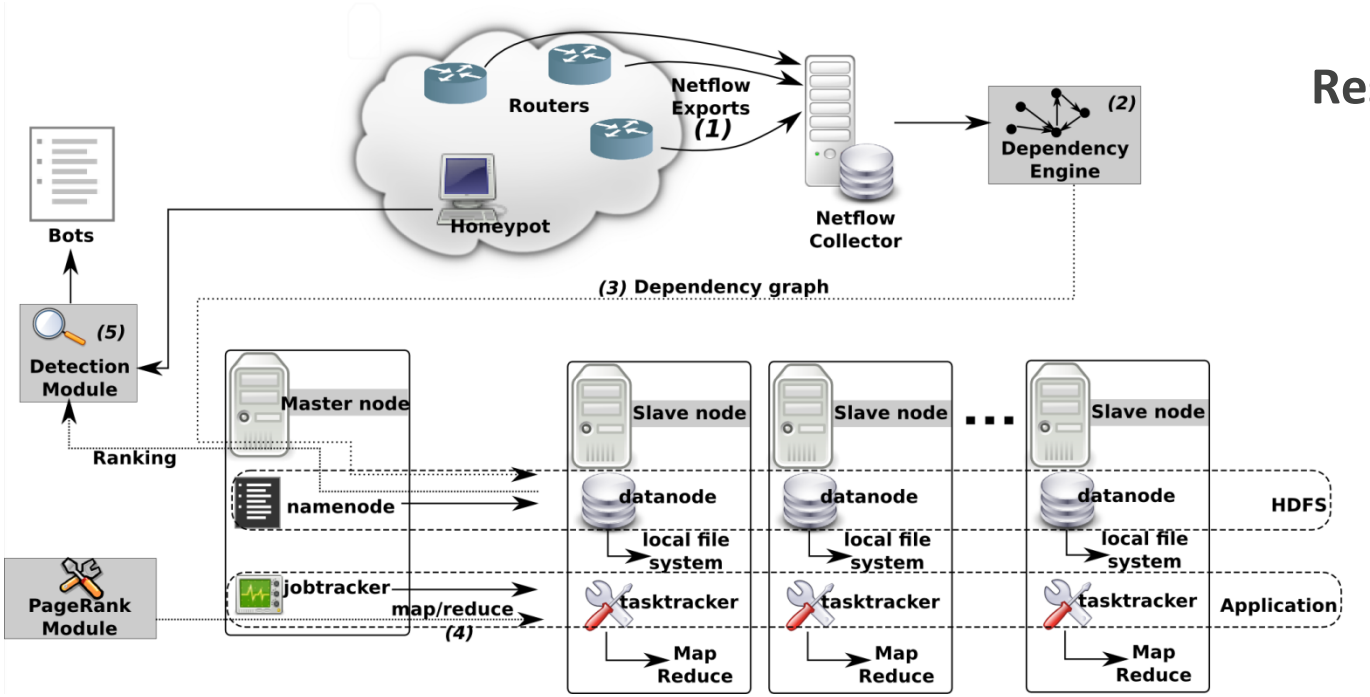
Source: http://www.csoonline.com/article/348317/what-a-botnet-looks-like, Scott Berinato

P2P Bots detection:

- Well interconnected machines to maintain the underlying network
- → Graph Analysis (PageRank/Google)
- Improvement by leveraging honeypots

Botnet tracking

BotTrack / BotCloud (MapReduce version):



Results:

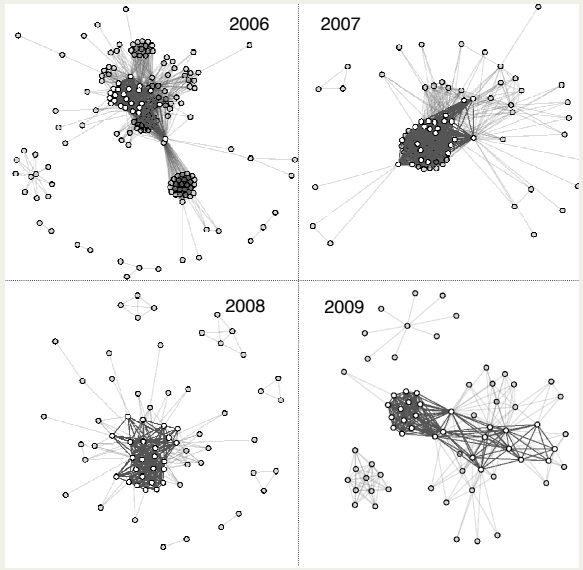
- Stealthy botnet detection (1% of IP addresses)
- High accuracy ~ 99%
- Scalability (60,000 flows / second)

Publications:

- *BotTrack: Tracking Botnets Using NetFlow and PageRank,* François J., Wang S., State R., Thomas E., IFIP Networking 2011
- *BotCloud: Detecting Botnets Using MapReduce,* François J., Wang S., Bronzi W., State R., Engel T., IEEE International Workshop on Information Forensics and Security - WIFS'11

Financial Data Analysis

- Research Questions:
  - Model the complex relationships for co-lending in EU banking zone
  - Modeling of financial instruments
  - Mining of highly unstructured data formats
  - Integrate economical (NYSE), regulatory (SEC) and media news (chairman, board member information obtained from social media Twitter, Google Trend)
  - Assess and model risk based on graph modeling and Distributed computations
  - Analyze loan interests rates (Libor) with respect to additional loan rates and economic indicators.
- Expected Outcomes
  - Link to local (Luxembourg) economy
  - Impact to EU regulatory bodies
  - Development of such activities at the SnT



Notice: Figure from (1)

Questions ?