DNS Monitoring, looking out for anomalies using the time frame of Name - IP association

Lautaro Dolberg

UNIVERSITÉ DU LUXEMBOURG
SnT
NETLAB

SnT

Outline

Outline

SnT

Personal Information

Lautaro Dolberg, Born in Buenos Aires - Argentina. 29 Years old

- PhD Studies at Universite de Luxembourg.
- Started **15/1/2012** - Expected Defense **2X/3/2015**
- Secan Lab - Vehicular Lab. J Francois, R State, R. Frank & T. Engel

Personal Information

Lautaro Dolberg, Born in Buenos Aires - Argentina. 29 Years old

- PhD Studies at Universite de Luxembourg.
- Started **15/1/2012** - Expected Defense **2X/3/2015**
- Secan Lab - Vehicular Lab. J Francois, R State, R. Frank & T. Engel
- APR 14 - Prof Raouf Boutaba research team at Waterloo University, Ontario, Canada.

SnT

Personal Information

Lautaro Dolberg, Born in Buenos Aires - Argentina. 29 Years old

- PhD Studies at Universite de Luxembourg.
- Started **15/1/2012** - Expected Defense **2X/3/2015**
- Secan Lab - Vehicular Lab. J Francois, R State, R. Frank & T. Engel
- APR 14 - Prof Raouf Boutaba research team at Waterloo University, Ontario, Canada.
- "Licenciatura en Ciencias de la Computacion" at DC - FCEN UBA (2011)

SnT

Personal Information

Lautaro Dolberg, Born in Buenos Aires - Argentina. 29 Years old

- PhD Studies at Universite de Luxembourg.
- Started **15/1/2012** - Expected Defense **2X/3/2015**
- Secan Lab - Vehicular Lab. J Francois, R State, R. Frank & T. Engel
- APR 14 - Prof Raouf Boutaba research team at Waterloo University, Ontario, Canada.
- "Licenciatura en Ciencias de la Computacion" at DC - FCEN UBA (2011)
- Research Keywords: Security, Networks, VANET, Monitoring, Big Data, Mobile Devices, SDN

SnT

Research Context

- Unrestricted access networks
  - Internet Applications Mobile (Crowd-sourcing)
  - Traffic Sensing Applications (LuxTraffic)
  - Internet Name Resolution (DNS)
- Large Data Collections
- Monitoring for detecting misbehavior
- Mobile Security (Android + IOS)
- Software Defined Networks
- Network Awareness

Security & Management of distributed applications

Outline

SnT

Motivation

DNS traffic reflects Internet activities and behaviors

- Internet Threats Growing: Phishing, Malware, Spoofed Domains.
- Identify malware <u>behavior</u> by assessing association time between names and networks.
- Helps for contextualizing other data such as Netflow, etc.
- Available resources such as Passive DNS.
- As both DNS and IP space follow hierarchical organization, MAM can be used.

Motivation cont'd

- *Can we use DNS records and its changes over time to trace Internet activities?*

SnT

Motivation cont'd

- *Do malicious domains behave different from others in terms of name - ip association?*

SNT

Motivation cont'd

- *Do malicious domains behave different from others in terms of name - ip association?*

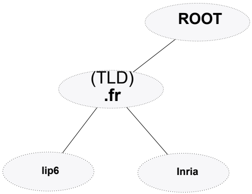
Example DNS Records

Type	Name	IPV4	TTL
CNAME	www.lip6.fr	ww.lip6.fr	X
A	ww.lip6.fr	132.227.104.15	X

DNS Background

DNS is an essential service for Internet

- Emerged in 1987 (RFC 1035, 1123, 2181)
- Domain names are labels separated with dots.
- Strictly hierarchical pattern
- TLD (*eg:* .com, .fr, .etc)

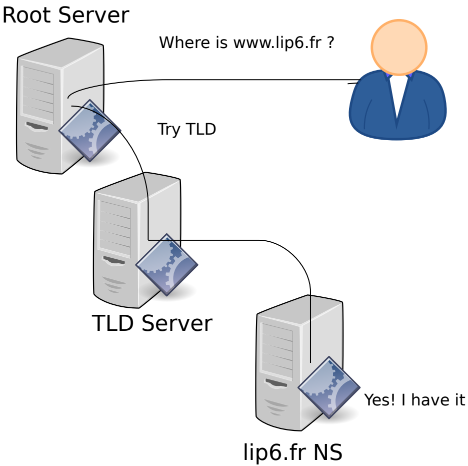


Max depth 127. Limited to 253 characters, label limit 63 characters.

DNS Structure & Procedure

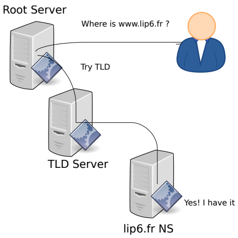
Resolving www.lip6.fr

DNS Structure & Procedure



DNS Structure & Procedure

Resolving www.lip6.fr

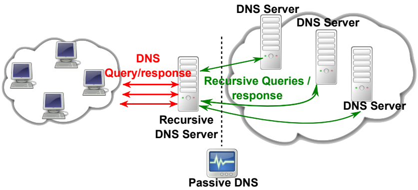


Relevant DNS Registers

- A: Association IPV4 or/and IPV6.
- CNAME: Redirect.
- PTR: Reverse DNS search.

Passive DNS

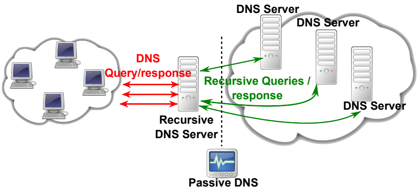
A Passive DNS DB contains:



- Where did this domain name point to in the past?

Passive DNS

A Passive DNS DB contains:

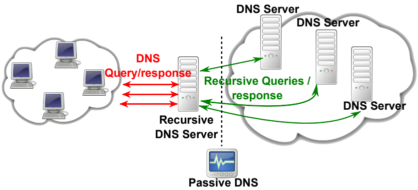


- Where did this domain name point to in the past?
- What domain names are hosted by a given nameserver?

SnT

Passive DNS

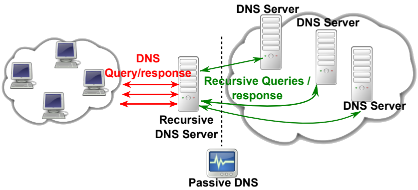
A Passive DNS DB contains:



- Where did this domain name point to in the past?
- What domain names are hosted by a given nameserver?
- What domain names point into a given IP network?

Passive DNS

A Passive DNS DB contains:



- Where did this domain name point to in the past?
- What domain names are hosted by a given nameserver?
- What domain names point into a given IP network?
- What subdomains exist below a certain domain name?

SnT

State of the art

Previous Work

- Exposure: Finding malicious domains using passive dns analysis, in *Network and Distributed System Security Symposium - NDSS, 2011*.
- DNSSM: A large-scale Passive DNS Security Monitoring Framework, in *IEEE/IFIP Network Operations and Management Symposium, 2012*.
- SDBF: Smart DNS Brute-Forcer, *in IEEE/IFIP Network Operations and Management Symposium - NOMS, 2012*.

Outline

MAM

How do we organize all the DNS data that we have?

SnT

MAM

How do we organize all the DNS data that we have?

- Preserve the hierarchy of Data (DNS & IP).
- Reduce the scale

SnT

MAM

How do we organize all the DNS data that we have?

- Preserve the hierarchy of Data (DNS & IP).
- Reduce the scale
- Minimize information loss due to aggregation.

SnT

MAM

How do we organize all the DNS data that we have?

- Preserve the hierarchy of Data (DNS & IP).
- Reduce the scale
- Minimize information loss due to aggregation.
- Fine control granularity for data analysis.

MAM is an enabler for aggregation and data retrieval.

Multidimensional Aggregation Monitoring + Applications

Aggregation

- Scalable way to represent information
    - Outline relevant correlated facts
    - Flexible granularity

Multidimensional Aggregation Monitoring + Applications

Aggregation

- Scalable way to represent information
  - Outline relevant correlated facts
  - Flexible granularity
- Features:
  - Custom Units (e.g. traffic packets, vehicle, traffic units)
  - choose criteria for aggregation

Multidimensional Aggregation Monitoring + Applications

Aggregation

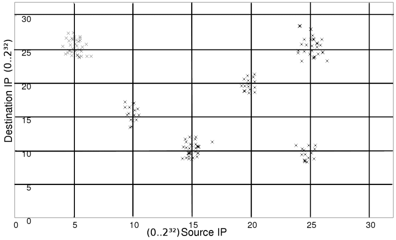
- Scalable way to represent information
  - Outline relevant correlated facts
  - Flexible granularity
- Features:
  - Custom Units (e.g. traffic packets, vehicle, traffic units)
  - choose criteria for aggregation
- Temporal and Spatial aggregation
  - Temporal: time windows split (β)
  - Spatial: keep nodes with activity $> \alpha$ *e.g. traffic volume*, aggregate the others into their parents $\rightarrow$ needs hierarchical relationships

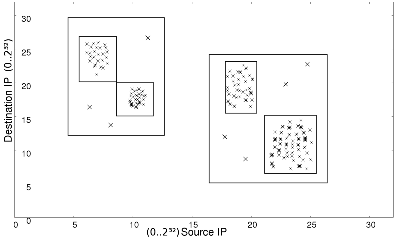
Published in LISA'12.

SnT

Space Partitioning

Example: IPV4 space partitioning, static vs dynamic





MAM for DNS

With MAM is possible to generate time aggregated combining multiple data.

- Two dimensions

SnT

MAM for DNS

With MAM is possible to generate time aggregated combining multiple data.

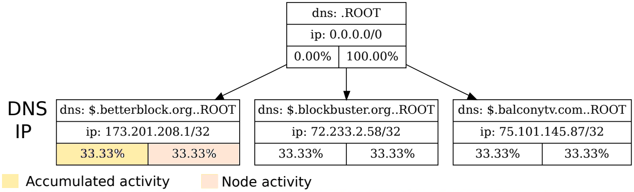
- Two dimensions
- Hierarchically derived from data model (IPV4 & DNS Data Space)

SnT

MAM for DNS

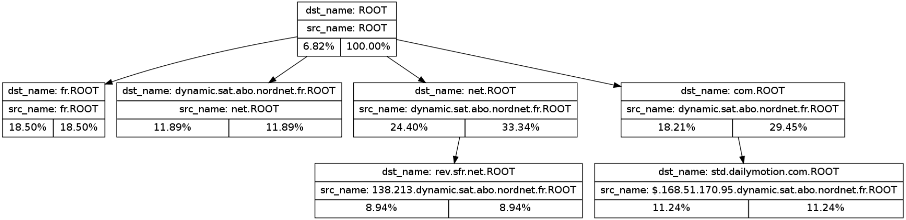
With MAM is possible to generate time aggregated combining multiple data.

- Two dimensions
- Hierarchically derived from data model (IPV4 & DNS Data Space)
- Example



Example II

Multidimensional tree for source and destination names



Little Outline

So far...

- DNS Interest and Background

Little Outline

So far...

- DNS Interest and Background
- Multidimensional Aggregation for hierarchical data

SnT

Little Outline

So far...

- DNS Interest and Background
- Multidimensional Aggregation for hierarchical data
- Coming next ← Leverage DNS activity analysis with MAM

Outline

SnT

Motivation

Asses the duration of the association (IP-NAME) over time

Motivation

Asses the duration of the association (IP-NAME) over time

We want to keep the notion of subnet and subdomain.

This should be stable over time

SNT

Motivation

Asses the duration of the association (IP-NAME) over time

We want to keep the notion of subnet and subdomain.

lets asume bmp.fr is fishing for bnp

SnT

Formal Definition

Assuming a sequence of K Trees

- $S = \{T_1 \ldots T_K\}$ representing DNS record association over time split in K

SnT

Formal Definition

Assuming a sequence of K Trees

- $S = \{T_1 \ldots T_K\}$ representing DNS record association over time split in K
- $n \in Nodes(T_i)$

SnT

Formal Definition

Assuming a sequence of K Trees

- $S = \{T_1 \ldots T_K\}$ representing DNS record association over time split in K
- $n \in Nodes(T_i)$
  - $n.dns$ represents the DNS name (using 'dollar' as terminal symbol)

Formal Definition

Assuming a sequence of K Trees

- $S = \{T_1 \dots T_K\}$ representing DNS record association over time split in K
- $n \in Nodes(T_i)$
  - $n.dns$ represents the DNS name (using 'dollar' as terminal symbol)
  - $n.ip$ represents the IPv4 address as a tuple (address, prefix_length)

Formal Definition

Assuming a sequence of K Trees

- $S = \{T_1 \ldots T_K\}$ representing DNS record association over time split in K
- $n \in Nodes(T_i)$
  - *n.dns* represents the DNS name (using 'dollar' as terminal symbol)
  - *n.ip* represents the IPv4 address as a tuple (address, prefix_length)
  - *n.accum* is the accumulated value representing the number of A Records.

SnT

Formal Definition

Assuming a sequence of K Trees

- $S = \{T_1 \ldots T_K\}$ representing DNS record association over time split in K
- $n \in Nodes(T_i)$
    - $n.dns$ represents the DNS name (using 'dollar' as terminal symbol)
    - $n.ip$ represents the IPv4 address as a tuple (address, prefix_length)
    - $n.accum$ is the accumulated value representing the number of A Records.
- The trees from S are aggregated according to a given *alpha*

SnT

Steadiness Metric I (Similarity)

The similarity of nodes is positive if:

- $n1,\ n2$ Tree nodes are from the same domain (Total overlap)

SnT

Steadiness Metric I (Similarity)

The similarity of nodes is positive if:

- $n1$, $n2$ Tree nodes are from the same domain (Total overlap)
- $n1.dns \subset n2.dns$ OR $n1.ip \subset n2.ip$ (Partial overlap)

SnT

Steadiness Metric I (Similarity)

The similarity of nodes is positive if:

- $n1$, $n2$ Tree nodes are from the same domain (Total overlap)
- $n1.dns \subset n2.dns$ OR $n1.ip \subset n2.ip$ (Partial overlap)
- $n2.dns \subset n1.dns$ OR $n2.ip \subset n1.ip$ (Partial overlap)

Otherwise similarity is 0.

SnT

Steadiness Metric I (Similarity)

The similarity of nodes is positive if:

- $n1,\ n2$ Tree nodes are from the same domain (Total overlap)
- $n1.dns \subset n2.dns$ OR $n1.ip \subset n2.ip$ (Partial overlap)
- $n2.dns \subset n1.dns$ OR $n2.ip \subset n1.ip$ (Partial overlap)

Otherwise similarity is 0.

Given n from T, we look for $m \in Nodes(T_x)$ where
$\forall\ m_x \in Nodes(T_x) :\ sim(n, m)\ \geq sim(n, m_x)$

Steadiness Metric II (Similarity)

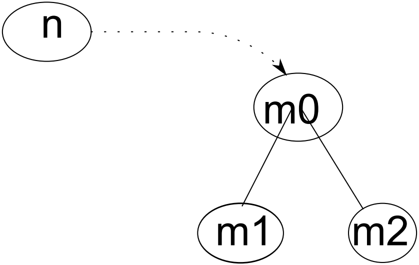
In other words:

- It's almost as inserting the node into the tree, and looking for a possible parent

Steadiness Metric II (Similarity)

In other words:

- It's almost as inserting the node into the tree, and looking for a possible parent



Steadiness Metrics III

Tree comparison, how to establish a similarity criteria?

SnT

Steadiness Metrics III

Tree comparison, how to establish a similarity criteria?

$$sim(n1, n2) = \alpha \times IP\_sim(n1, n2) + \beta \times DNS\_sim(n1, n2) + \gamma \times vol\_sim(n1, n2)$$

SnT

Steadiness Metrics III

Tree comparison, how to establish a similarity criteria?

$$sim(n1, n2) = \alpha \times IP\_sim(n1, n2) + \beta \times DNS\_sim(n1, n2) + \gamma \times vol\_sim(n1, n2)$$

$$IP\_sim(n1, n2) = 1 - \frac{|n1_{prefix\_len} - n2_{prefix\_len}|}{32}$$

SnT

Steadiness Metrics III

Tree comparison, how to establish a similarity criteria?

$$sim(n1, n2) = \alpha \times IP\_sim(n1, n2) + \beta \times DNS\_sim(n1, n2) + \gamma \times vol\_sim(n1, n2)$$

$$IP\_sim(n1, n2) = 1 - \frac{|n1_{prefix\_len} - n2_{prefix\_len}|}{32}$$

$$DNS\_sim(n1, n2) = \frac{|n1_{dns} \cap n2.dns|}{|n1_{dns} \cup n2_{dns}|}$$

SnT

Steadiness Metrics III

Tree comparison, how to establish a similarity criteria?

$$sim(n1, n2) = \alpha \times IP\_sim(n1, n2) + \beta \times DNS\_sim(n1, n2) + \gamma \times vol\_sim(n1, n2)$$

$$IP\_sim(n1, n2) = 1 - \frac{|n1_{prefix\_len} - n2_{prefix\_len}|}{32}$$

$$DNS\_sim(n1, n2) = \frac{|n1_{dns} \cap n2.dns|}{|n1_{dns} \cup n2_{dns}|}$$

$$vol\_sim(n1, n2) = 1 - 0.01 \times |n1_{acc\_vol} - n2_{acc\_vol}|$$

SNT

Smoothing

Smoothing helps considering early time windows

- $n1 \in Nodes(T_1), \; n2 \in Nodes(T_2)$
- We compute $n2 = \; most\_sim(n1)$ is
- $stead(n1) = sim(n1, n2) + \mu \times stead(n2)$. With T_0 as base case and $\mu \in \Re$.

SNT

Smoothing

Smoothing helps considering early time windows

- $n1 \in Nodes(T_1),\ n2 \in Nodes(T_2)$
- We compute $n2 =\ most\_sim(n1)$ is
- $stead(n1) = sim(n1, n2) + \mu \times stead(n2)$. With T_0 as base case and $\mu \in \Re$.

So we compute the global steadiness of a Tree T by:

$$\text{Persistence}(T) = \frac{\sum_{n \in Nodes(T)} stead(n)}{|n \in Nodes(T)|}$$

Outline

Data Set

Aggregation Window: 1 Week Time Length

- Macro: Up to 52 weeks from 2011-04-23 to 2012-06-30 (662 K)
- Micro: 10 weeks maximum

SnT

Data Set

Aggregation Window: 1 Week Time Length

- Macro: Up to 52 weeks from 2011-04-23 to 2012-06-30 (662 K)
- Micro: 10 weeks maximum

Malicious data

- Time: Periodically, Steady

SnT

Data Set

Aggregation Window: 1 Week Time Length

- Macro: Up to 52 weeks from 2011-04-23 to 2012-06-30 (662 K)
- Micro: 10 weeks maximum

Malicious data

- Time: Periodically, Steady
- Proportion: 0.1%, 1% and 10%

SnT

Data Set

Aggregation Window: 1 Week Time Length

- Macro: Up to 52 weeks from 2011-04-23 to 2012-06-30 (662 K)
- Micro: 10 weeks maximum

Malicious data

- Time: Periodically, Steady
- Proportion: 0.1%, 1% and 10%
- Source: Blacklists (Exposure, WOT) 175K

Aggregation Granularity: 2%

SnT

Steadiness Distribution

Distribution of Local Steadiness (Leafs)

- More than 50% of malicious nodes have less than 0.7 of stability

SnT

Steadiness Distribution

Distribution of Local Steadiness (Leafs)

- More than 50% of malicious nodes have less than 0.7 of stability
- Less than 20% of malicious nodes have more than 0.85 of stability

Steadiness Distribution

Distribution of Local Steadiness (Leafs)

- More than 50% of malicious nodes have less than 0.7 of stability
- Less than 20% of malicious nodes have more than 0.85 of stability
- Less than 40% of normal data have a steadiness of 0.8% or less.

Steadiness Distribution

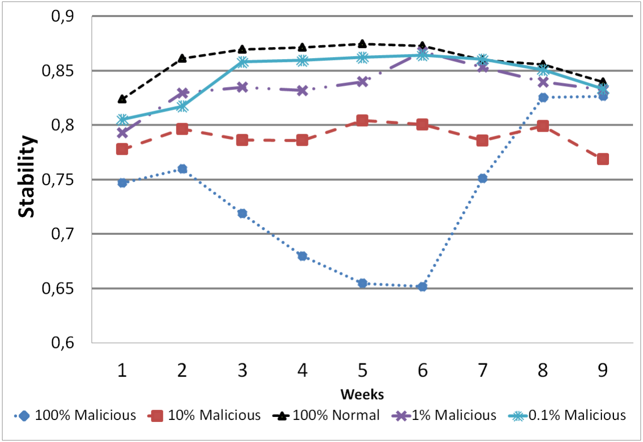
Distribution of Local Steadiness (Leafs)

- More than 50% of malicious nodes have less than 0.7 of stability
- Less than 20% of malicious nodes have more than 0.85 of stability
- Less than 40% of normal data have a steadiness of 0.8% or less.
- Only 10% of normal data have a steadiness of less than 0.5

SnT

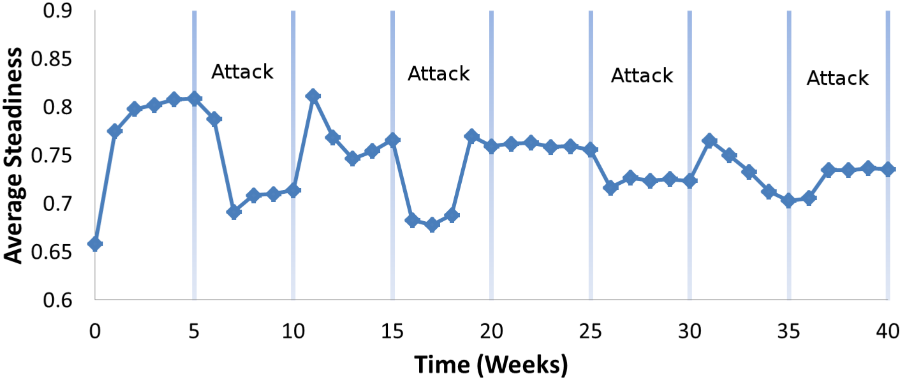
Microscopic observation

Malicious domains causes a drop on average steadiness



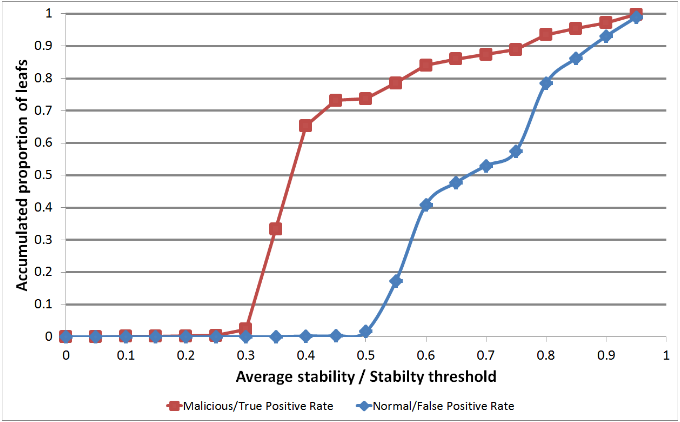
Macroscopic observation

Malicious domains causes a drop on average steadiness: Macro



Results II

Accuracy: Steadiness as metric for filtering malicious domains



SnT

Outline

Contributions

- A methodology for assessing DNS - IP association time frame was proposed.

SnT

Contributions

- A methodology for assessing DNS - IP association time frame was proposed.
- Reduced the scale of data, helpful in the context of network security. (From 80K nodes to 2K with $\alpha = 2\%$)

Contributions

- A methodology for assessing DNS - IP association time frame was proposed.
- Reduced the scale of data, helpful in the context of network security. (From 80K nodes to 2K with $\alpha = 2\%$)
- Definition of steadiness metrics for a local and global scope was introduced.

Contributions

- A methodology for assessing DNS - IP association time frame was proposed.
- Reduced the scale of data, helpful in the context of network security. (From 80K nodes to 2K with $\alpha = 2\%$)
- Definition of steadiness metrics for a local and global scope was introduced.
- Evaluation using real data and during several time frames. Validation of the metrics

SnT

Contributions

- A methodology for assessing DNS - IP association time frame was proposed.
- Reduced the scale of data, helpful in the context of network security. (From 80K nodes to 2K with $\alpha = 2\%$)
- Definition of steadiness metrics for a local and global scope was introduced.
- Evaluation using real data and during several time frames. Validation of the metrics
- Scalability: It can be implemented using dynamic programming / distributed computing.

Contributions cont'd

Relevant Publications

- **TCP** 2012: L. Dolberg, J. Francois, T. Engel. "Multidimensional Aggregation Monitoring", Usenix LISA 2012
- **DNS** 2013: L. Dolberg, J. Francois, T. Engel. "DNS Malware Detection using Stability Metrics", IEEE LCN 2013

Thanks!

Questions