Tamás Budavári, Nick Carey / Johns Hopkins

# **DISCOVERY: MAPPING** ACCESS PATTERNS

C3E MALICIOUS CYBER

Oct 20, 2014

# Our Approach

- Traditional method focuses on individual machines
  String matching on individual requests
- We study all machines relative to one another
  Relations based on their access pattern

# **Creating Maps**

A	A	A	A	P	7	7	A	A	A	4	4	A	6	Þ	V	V
A	A	A	A	2	8	V	A	A	V	4	1	A	L	Þ	V	V
A	A	A	A	2	7	7	F	A	4	4	1	4	L	Þ	V	V
A	A	A	A	2	7	7	A	Y	V		1	4	L	Þ	V	V
A	A	A	2		7	7	A	Y	V	4	4	4	4	Þ	V	V
A	A	A	>		7	7	A	¥	V	4	~	4	b.	Þ	V	V
A	A	A			7	$\nabla$	4	¥	×	4	4	L	4	Þ	Þ	V
A	A	N	>	*	7	$\overline{\mathbf{v}}$	A	¥	¥	*	4	4	4	Þ	Þ	V
A	A	N	۲	7	7	~	4	¥	¥		*	4	Ł	Þ	Þ	v
A	A	N	۲	7	7	~	4	¥	*		A	Ā.	4	Þ	Þ	v
A	٨	N		4	7	$\overline{\mathbf{v}}$	4	۲	×	4	£	A	*	Þ	Þ	v
A	A	N		۴	*	7	4	v	*		*	4	4	*	¥.	r
A	٨			۴	٠	*	4	۲	*		*	4	4	Þ	Þ	v

# **Creating Maps**

A	A	A	A	P	7	V	V	A	A	4	4	L	4	Þ	V	V
A	A	A	A	2	7	V	V	A	V		1	4	4	Þ	Þ	V
A	A	A	A	2	7	7	V	A	¥	4	1	4	L	Þ	V	V
A	A	A	A		7	7	A	Y	A	4	1	4	4	Þ	V	V
A	A	A	2		7	7	A	Y	V	4	4	4	4	Þ	Þ	V
A	A	A			7	7	4	Y	V		4	4	4	Þ	V	V
A	A	A	>		7	7	4	A	×		4	L	4	Þ	Þ	V
A	A	A	>	*	7	*	7	Y	×	*	4	4	4	Þ	Þ	V
A	A	N	۲	7	7	$\overline{\nabla}$	4	¥	*		*	4	L	Þ	Þ	v
A	A	N	>	7	7	7	4	¥	¥			I.	4	Þ	Þ	v
A	A	N		4	7	7	4	¥	×	4	Ł	4	4	Þ	Þ	v
A	A	N		۴	*	7	4	v	*		*	4	4	*	¥.	r
A	A			۴	٣	*	4	¥	*	*		4	4	*	۲	v



## **Graph of Local Similarities**

#### Started w/ ISOMAP & LLE in Science Mag (2000)



Laplacian eigenmaps from local similarities

# Work on PREDICT

Nick Carey

#### Collections

- DARPA Synthetic Network Monitor 1TB/day
  Contains ground truth data on malware activity
  Simulated pcap on 172.28.0.0/16 Network
  For each machine, extract accessed IP meta-data
- Manipulate in relational database
  - Easy statistics using SQL queries
  - Ranking & weighting of IP addresses

**Defining Similarity** 

We care about who the machines talk to
 Find all accessed IPs for each machine
 Two machines are similar if

Set of accessed IPs overlap the most

*S* = |*Intersection*|/|*Union*|

### **Preliminary Results**

#### Solve for 3D embedding



#### Variants

Weight of IPs is important
 Frequent nodes we can ignore
 Everybody goes to Google, etc.

We can assign weights to all IPs
 Based on the number of accesses



# **Preliminary Results**

□ Coming soon...

## Summary

- We study the ensemble of machines on a network
  No absolute threshold on a string match quality
- Create map of network based on access pattern
  We use DARPA Synthetic Network Monitor metadata
- Preliminary maps have "interesting" features
  - Need to understand trends and outliers to refine

#### Future Work: Human Comparison

