C3E MALICIOUS CYBER DISCOVERY: MAPPING ACCESS PATTERNS

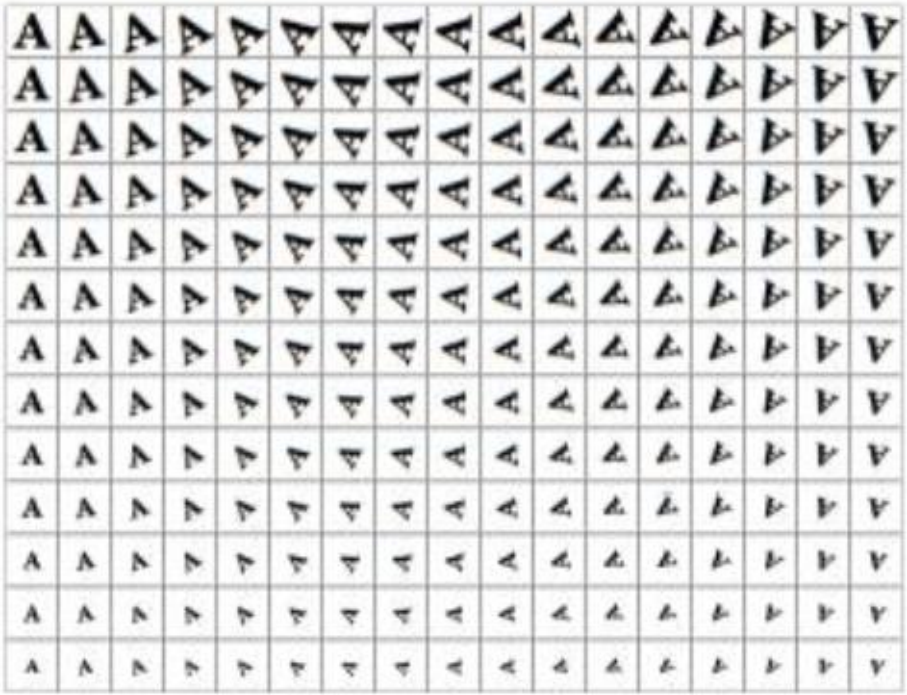
Oct 20, 2014

Tamás Budavári, Nick Carey / Johns Hopkins

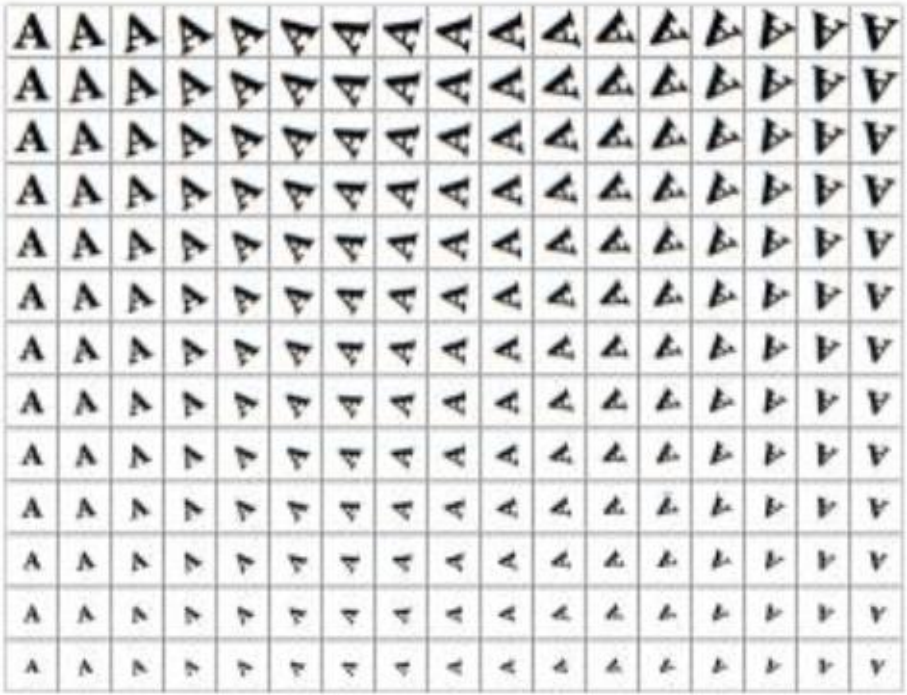
Our Approach

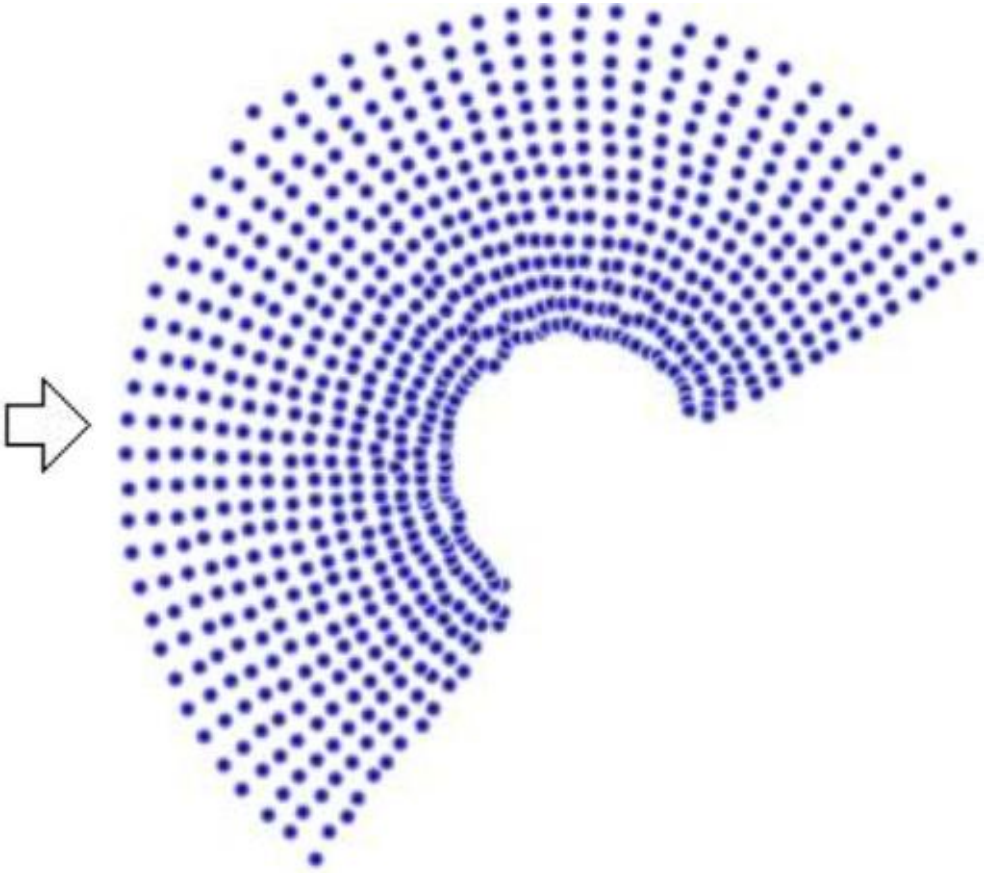
- Traditional method focuses on individual machines
  - String matching on individual requests
- We study all machines relative to one another
  - Relations based on their access pattern

Creating Maps



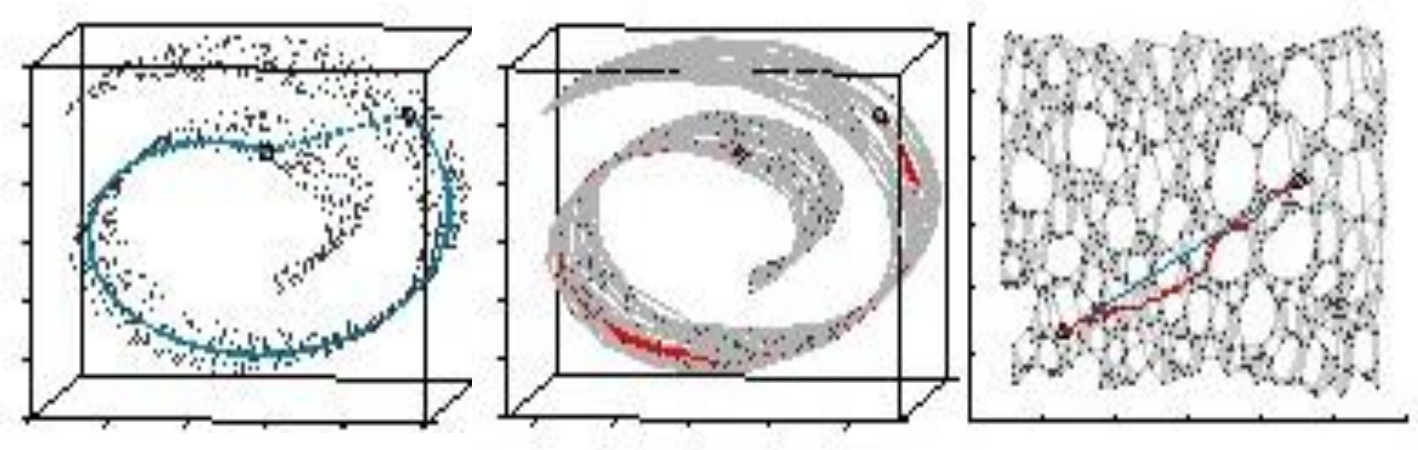
Creating Maps





Graph of Local Similarities

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



- Laplacian eigenmaps from local similarities

Work on PREDICT

Nick Carey

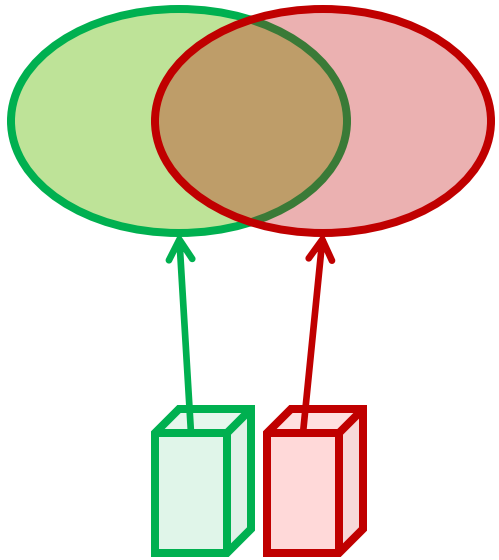
Collections

- 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

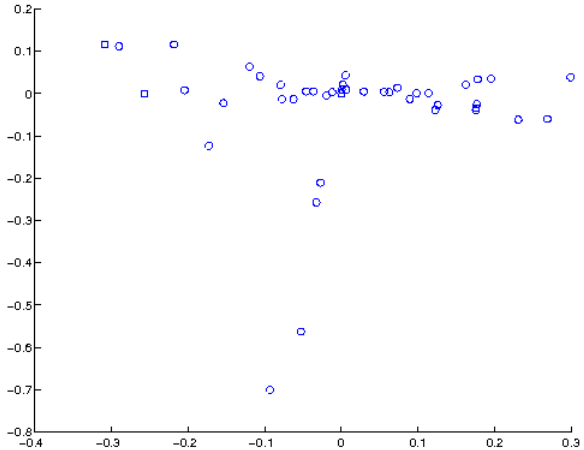
$$S = |Intersection| / |Union|$$



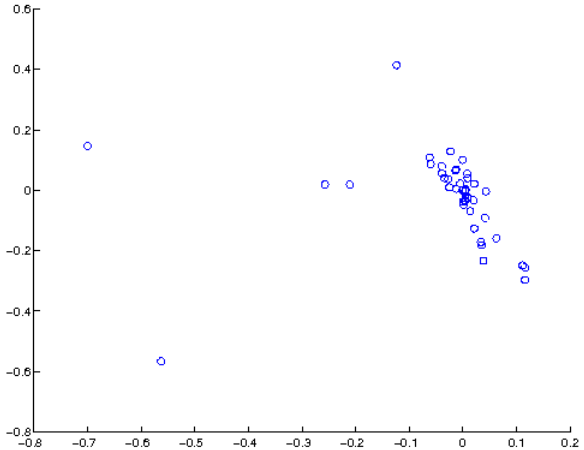
Preliminary Results

- Solve for 3D embedding

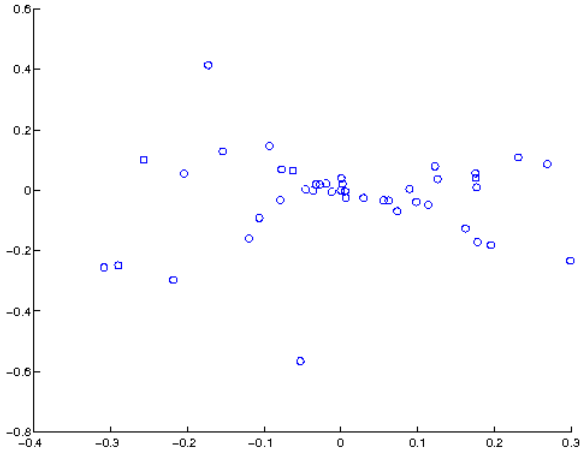
2-3



3-4



2-4



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

$$\sum_i w_i$$

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