FAU

CYBER THREAT INTELLIGENCE LAB
College of Engineering & Computer Science
Florida Atlantic University

BEHAVIORAL SERVICE GRAPHS: A BIG DATA APPROACH FOR PROMPT INVESTIGATION OF INTERNET-WIDE INFECTIONS

DR. ELIAS BOU-HARB, FAU, USA

DR. MARK SCANLON, UCD, IRELAND

March 23rd, 2017

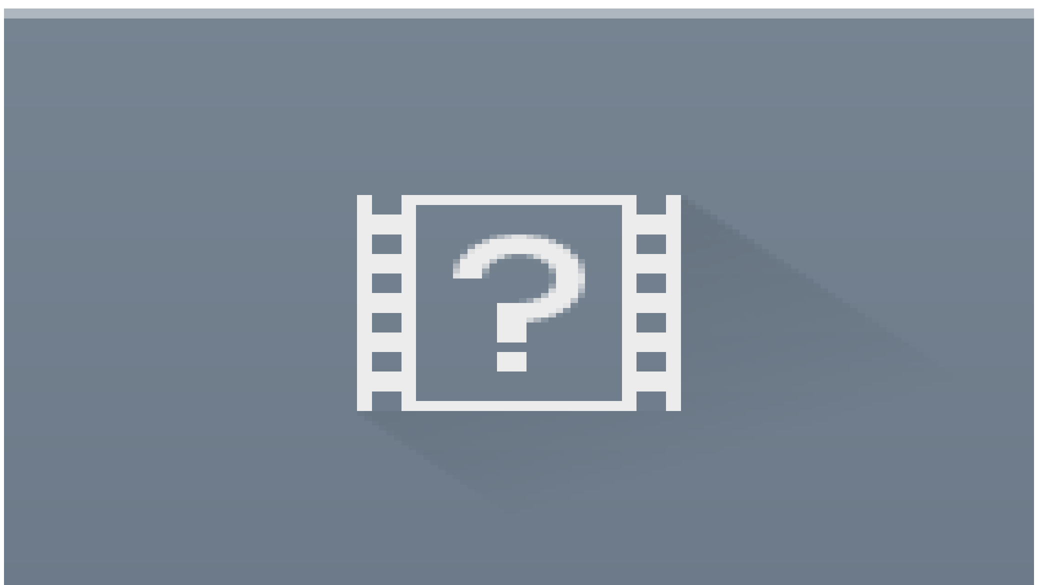
DFRWS EU 2017

Outline

- Introduction, Motivation & Contributions
- Related Work
- Proposed Approach
- Empirical Evaluation
- Limitations and Possible Improvements
- Concluding Remarks and Future Work

Introduction & Motivation

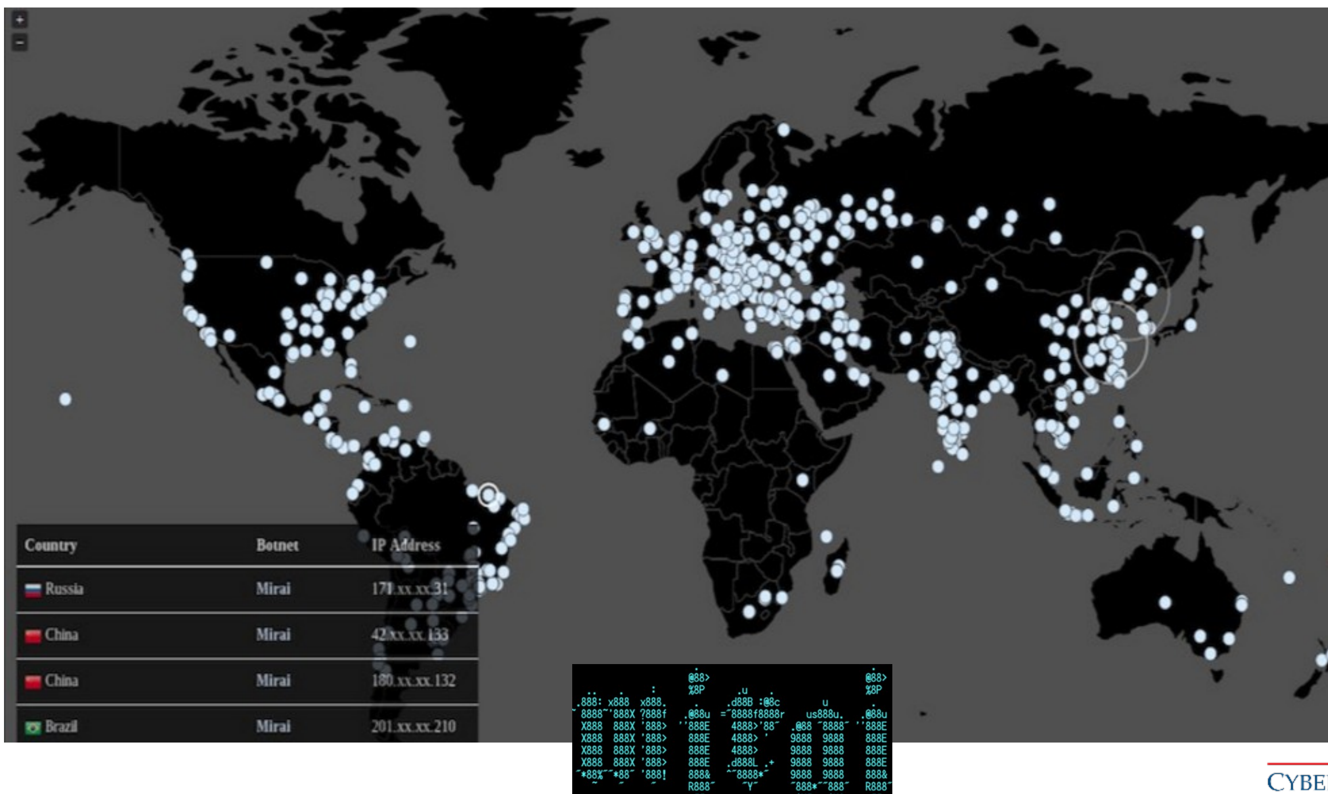
- This video illustrates a large scale orchestrated probing campaign targeting VoIP servers as reported by The Center for Applied Internet Data Analysis (CAIDA).
- This and other events continue to be stealthy and the occur on a frequent basis.



FAU
CYBER THREAT INTELLIGENCE LAB
College of Engineering & Computer Science
Florida Atlantic University

Introduction & Motivation

- Recently, we have seen a large-scale coordinated DDoS attack by exploiting IoT devices (mainly cameras), which took down many famous services such as amazon and twitter.



Introduction & Motivation

- Internet-scale infections and orchestrated events continue to escalate

- The need for *prompt*, *formal* and *accurate* solutions, which can operate on big Internet-wide data
  - Preferably we would like to have an approach that is formal and exploit data analytics techniques.

Forensic Challenges

- Network forensic analysts are significantly overwhelmed by huge amounts of low quality evidence, i.e., false positives and false negatives

- Network forensic approaches are passive or reactive, employ manual ad-hoc methods and are time consuming

- Most current network forensic practices do not support distributed inference, and if they do, they force the analysts to go through an error-prone process of correlating dispersed unstructured evidence to infer a specific security incident

Outline

- Introduction, Motivation & Contributions
- **Related Work**
- Proposed Approach
- Empirical Evaluation
- Limitations and Possible Improvements
- Concluding Remarks and Future Work

Related Work

Anomaly detection using graphs

Big data forensic approaches

In contrast, we we attempt to fuse both to provide a prompt and a sound approach:

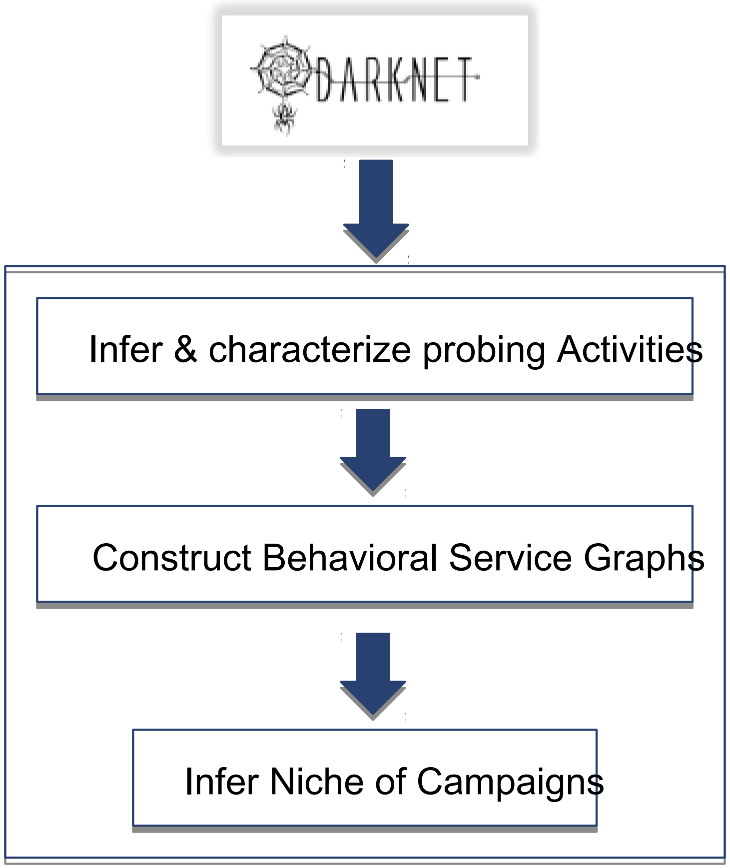
- Infer Internet-wide infections
- Leverage probing activities using a set of behavioral analytics to infer infections
- Employ a new concept of similarity service graphs to infer campaigns of infected machines
- Exploit graph theoretic notions to infer the niche of the infected campaign

Outline

- Introduction, Motivation & Contributions
- Related Work
- Proposed Approach
- Empirical Evaluation
- Limitations and Possible Improvements
- Concluding Remarks and Future Work

Proposed Approach

- Our approach works in a Security Operation Center (SOC) model by investigation darknet data, which is Internet-scale data that targets routable, allocated yet unused IP addresses.
- It attempts to infer infected bots by characterizing probing activities, which are the very first signs of infection.
- It then constructs certain graphs and manipulates them to infer the campaigns



Infer & characterize probing Activities (1/3)

- Infer probing activities from darknet data
  - Plethora of approaches to do this
  - We leverage a previous work

- Characterize their behaviors (probing strategy, randomness in traffic, etc.) based on statistical tests and heuristics

Infer & characterize probing Activities (2/3)

Category	Employed method	Behavior characteristic
Randomness	Wald-Wolfowitz	Random Traffic Pattern in Traffic
Probing Strategy	Mann-Kendall Chi-square goodness-of-fit	Sequential Probing Permutation Probing
Nature of Probing Source	Analysis of randomness and probing strategy	Probing tool Probing bot
Target	Concept of relative uncertainty Theoretic metric	Dispersed
Miscellaneous Inferences	Rate Port Number	Rate (packets/second) Destination Overlap Port Number

Infer & characterize probing Activities (3/3)

Behavior vector

Bot :

$$\begin{bmatrix} \text{Randomness} \\ \text{Probing Strategy} \\ \text{Target} \\ \text{Rate} \\ \text{Destinations Overlap} \\ \text{Port Number} \end{bmatrix}$$

Construct Behavioral Service Graphs (1/3)

- Model probing bots in an undirected complete graph
  - Nodes are the scanning bots
  - Edges are weights related to their similarity

- Each graph clusters a number of bots targeting the same port, which define an orchestrated campaign

Construct Behavioral Service Graphs (2/3)

Bot 1: [Random Traffic
Sequential Probing
Dispersed Probing
Rate: 60 pps*
Destinations Overlap: 100*
Port Number: 80]

Bot 2: [Pattern in Traffic
Sequential Probing
Targeted Probing
Rate: 55 pps*
Destinations Overlap: 200*
Port Number: 80]

Behavioral Similarity = 50%

*15% similarity for Rate and Overlap to be considered similar

Construct Behavioral Service Graphs (3/3)

- Allow the prompt inference of bot infected machines
- Automate amalgamation of evidence from distributed entities
- Provide valuable insights related to behaviors of the infected machines

Infer Niche of Campaigns (1/2)

- Niche of campaign defines those nodes that aggressively infect other nodes or are heavily used in C&C communication

- Apply maximum spanning tree algorithm to create an Erdős–Rényi random subgraph
  - Nodes with maximum similarity are the niche nodes

Infer Niche of Campaigns (2/2)

Unique characteristics of campaigns:

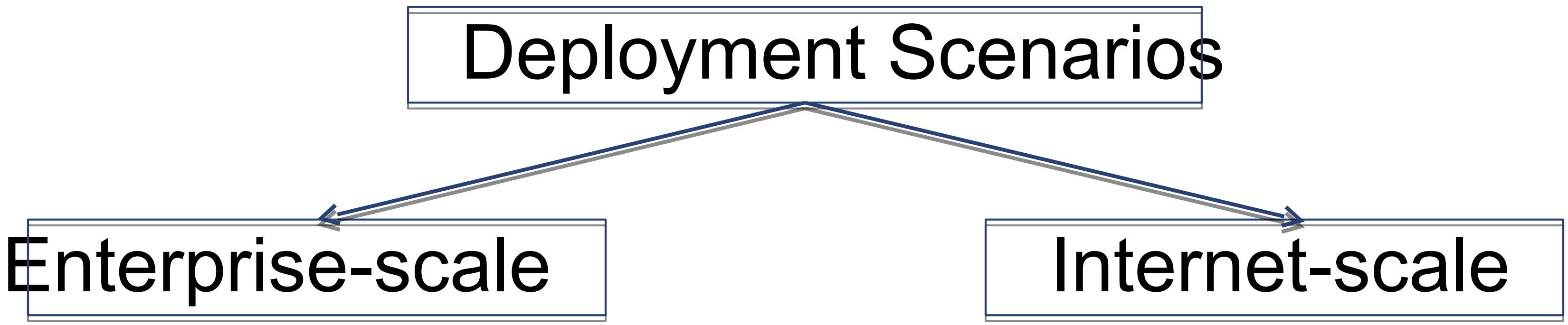
- Population of bots has several orders of large magnitude
- Targeted the entire IP address space
- Bots adopt well orchestrated strategies to maximize targets coverage

radication of Niche can limit the propagation of the Campaign

Outline

- Introduction, Motivation & Contributions
- Related Work
- Proposed Approach
- **Empirical Evaluation**
- Limitations and Possible Improvements
- Concluding Remarks and Future Work

Deployment Scenarios



- Two different deployment scenarios are used to validate accuracy, effectiveness, and simplicity of the approach.
- In the first scenario, Behavior Service Graphs are employed to infer infected machines within an enterprise network. While in the second scenario, the approach is ported to a global scale.

Data and Ground Truth

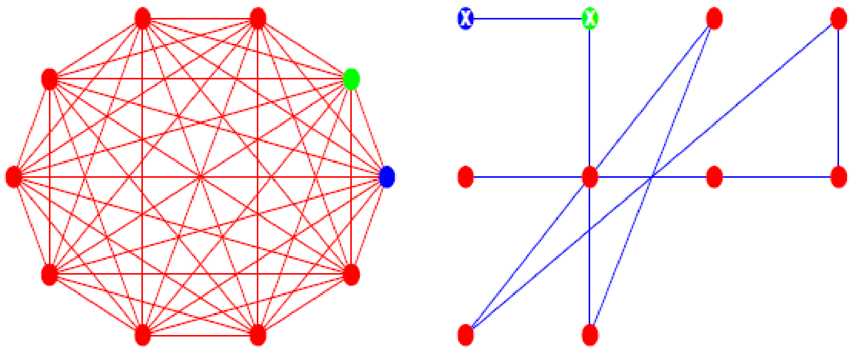
Enterprise-scale

- In the first scenario, use enterprise network traffic dataset and a confirmed campaign that targeted IPv4 as a ground truth

- Enterprise network traffic dataset
  - 15 GB by leveraging the Security Experimentation EnviRonment (SEER)

- Ground truth is an orchestrated probing campaign (*Carna botnet)*
  - Considered as one of the largest and most comprehensive probing census targeted IPv4

Outcome
Enterprise-scale



- Inferring and clustering 10 infected machines
- 2 IP addresses as the niche of such campaign
  - Their prompt eradication can limit the propagation of this campaign

Data and Ground Truth

Internet-scale

- Darknet Data
  - Operate the approach in a Security Operation Center (SOC) model

- Ground truth is a probing campaign from October 2012
  - Reported by ISC to be targeting Internet-scale SQL servers

Outcome
Internet-scale

- Inferring and clustering close to 800 unique SQL-injection bots
- 84 bots as the niche of such campaign
  - Their prompt eradication can limit the propagation of this campaign

Outline

- Introduction, Motivation & Contributions
- Related Work
- Proposed Approach
- Empirical Evaluation
- **Limitations and Possible Improvements**
- Concluding Remarks and Future Work

Limitations and Possible Improvements

- Need to fortify the infection evidence
  - Currently working on correlating malware with probing traffic to accomplish this
- There's a need find a formal mathematical computation to infer the niche of the campaign
  - Currently relying on a threshold related to the subgraph
- Experimental, non-operational
  - Currently addressing scalability issues of the approach to make it function in near real-time on darknet data

Outline

- Introduction, Motivation & Contributions
- Related Work
- Proposed Approach
- Empirical Evaluation
- Limitations and Possible Improvements
- **Concluding Remarks and Future Work**

Concluding Remarks and Future Work

- Fusing data analytics with formal methods has rarely been investigated. We leverage this here to infer campaigns and their niches.
  - A step towards leveraging big data analytics with formal methods as applied to cyber security
- Preliminary results in a SOC model are promising

- Address the mentioned limitations
- We would like in future work to also verify the soundness of the approach in corporate networks using two-way traffic.

Acknowledgements

NSF

FAU
CYBER THREAT INTELLIGENCE LAB
College of Engineering & Computer Science
Florida Atlantic University

Florida Center for
Cybersecurity

Questions