Practical Attacks Against Graph-based Clustering

Yizheng Chen, Yacin Nadji, Athanasios Kountouras, Fabian Monrose, Roberto Perdisci, Manos Antonakakis, Nikolaos Vasiloglou

Web Security & Privacy Lab

KAIST

What this paper is about?

- Problem
  - can an adversary evade detection;
  - can an adversary Generate and demonstrate low cost and effectiveness attack against Graph-based Clustering
- Contributions
  - First **practical** attempt to attack graph-based modeling techniques
  - Two Novel Attacks:
    - targeted noise injection attack
    - small community attack
  - Cost analysis
  - Defenses

Results

- Successful attack with minimal knowledge and low cost
  - attackers with no knowledge beyond their infections can render 84% of clusters too noisy to be useful, and evade clustering at a rate of 75%.
  - Were trained a Random Forest classifier with an average accuracy of 96.08%, and a false positive rate of 0.9%.
- SVD rank = 35 -> Minimum cost = 0

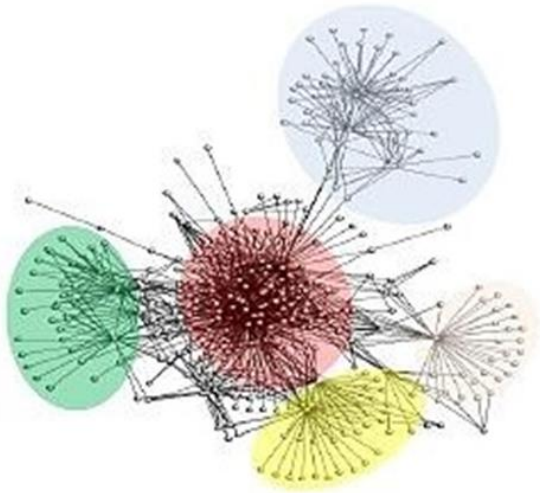
What this paper is about?

- Meaning
  - **the first practical attempt** to attack against graph-based clustering techniques, and a global feature space with realistic attackers **with or without perfect knowledge**.
- Focused on adversarial clustering, which deals with global features that cannot be directly changed.
- Capabilities of attackers with various knowledge levels, costs associated with attacks were evaluated.

Background

Graph-Based Clustering

- Collection of a wide range of very popular clustering algorithms that are based on graph-theory.
- Organize information in large datasets to facilitate users for faster access to required information.

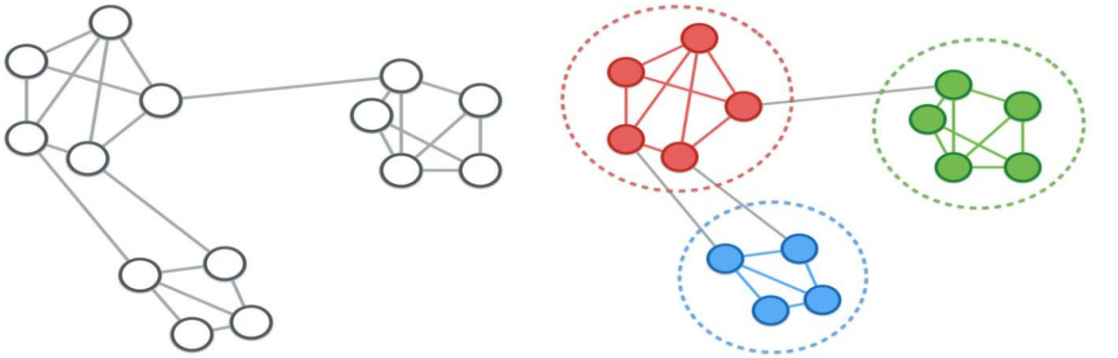


Use of Graph-based Clustering

- community discovery identifies criminal networks;
- connected components track malvertising campaigns;
- spectral clustering on graphs discovers botnet infrastructure;
- hierarchical clustering identifies similar malware samples;
- binary download graphs group potential malware download events;
- newly devised graph embedding's, like node2vec, could further improve upon the state of the art.

Graph based Clustering: Community Detection

- Discovering groups in a network where individuals' group memberships are not explicitly given.
- Rely on a modularity metric to evaluate the quality of partitions, which measures the density of links inside and outside communities.
- Allows to optimize modularity to quickly find communities by using the Louvain algorithm.



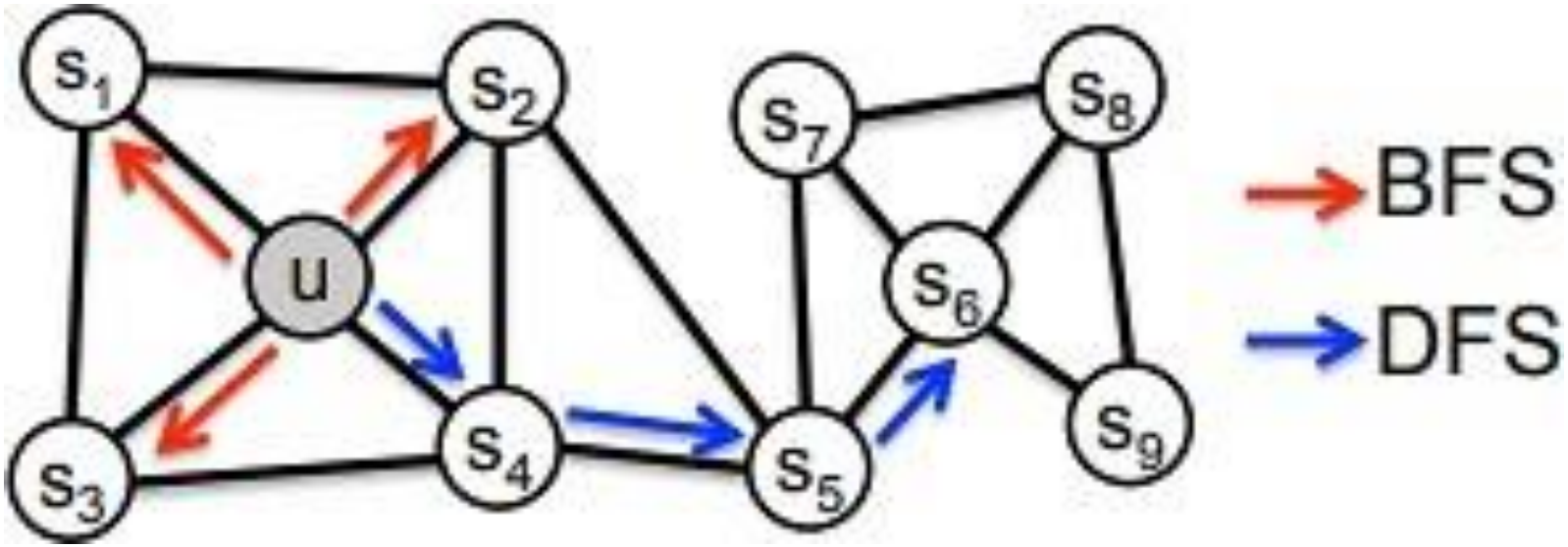
Graph based Clustering: Spectral Methods

- Spectral clustering techniques make use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions.
- The goal of spectral clustering is to cluster data that is connected but not necessarily compact or clustered within convex boundaries.

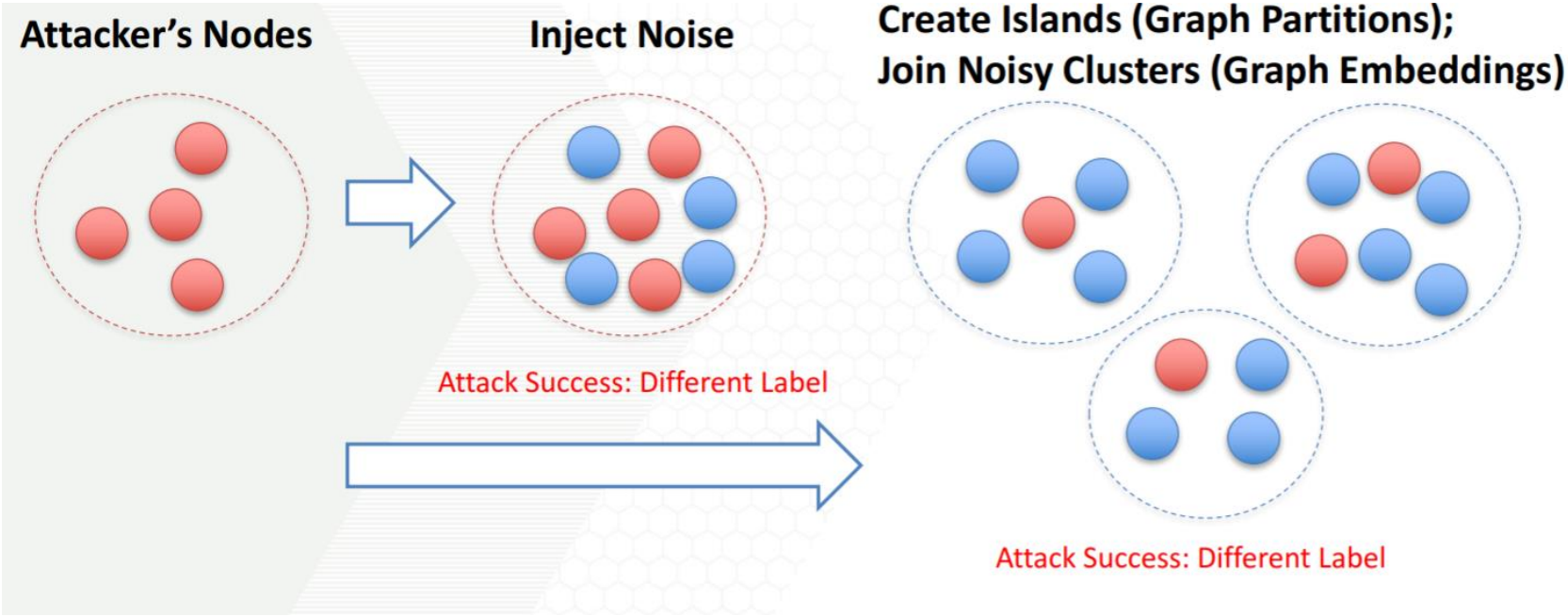
Graph based Clustering: node2vec

- the advantage of balancing homophily and structural equivalence in its embeddings.
- proposes a sampling strategy by random walks starting from every vertex on the graph with the following parameters:
  - number of walks from each vertex;
  - length of each walk;
  - probability to return to the same vertex (Breadth First Search)
  - probability to explore out to further vertices (Depth First Search).

Graph based Clustering: node2vec

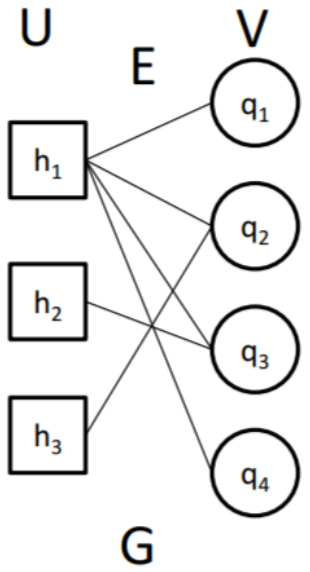


Intuition behind the attacks



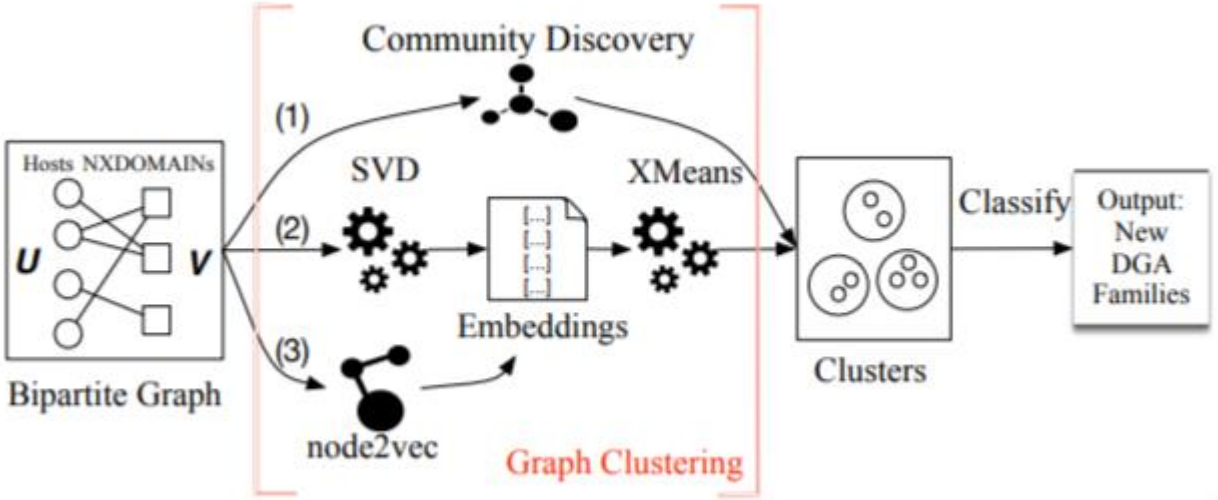
Threat Model & Attacks: Notation

- Dynamic domain generation for Command and Control (C&C)
- Algorithm in the malware or on the server
- Bipartite Graph G = (U, V, E)
  - Hosts (U) query NXDOMAINS (V)
  - An edges connects a vertex in U and one in v



PLEIADES

- Graph modeling component
- Reimplementation in real-world telecommunication d ata
  - Accuracy: 96.08%, and False Positive Rate: 0.9%
  - Discovered 12 new DGAs



Threat Model

Attacker's Goal: Change G to G' to evade clustering.

Attacker Knowledge Level: Minimal Knowledge

- The attacker knows only about graph G, as well as any open source intelligence (OSINT).
- An attacker with minimal knowledge can draw information from their infected hosts.
- The attacker can use OSINT to select potential data to inject as noise, or can coordinate activities between their vertices in G.

Attacker Knowledge Level: Moderate Knowledge

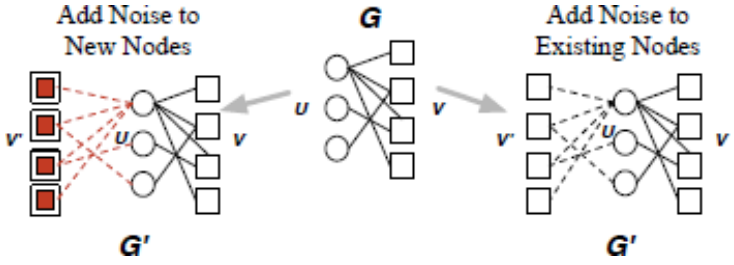
- The moderate knowledge case represents an adversary with $\tilde{G}$, an approximation of G.
- An attacker with moderate knowledge is similar to a sophisticated adversary with access to large datasets through legitimate (i.e., commercial data offerings) or illegitimate (i.e., security compromises) means.

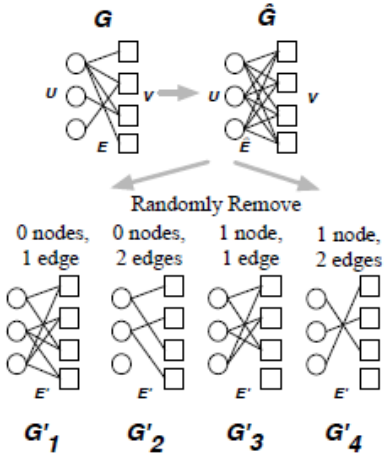
Attacker Knowledge Level: Perfect Knowledge

- An adversary can completely reconstruct the clustering results of the defender to evaluate the effectiveness of their attacks.

Attacks

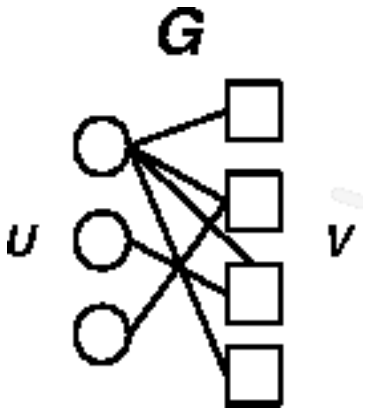
- two novel attacks against graph clustering;
- **Targeted noise injection**, improves on random injections by emulating the legitimate signal's graph structure;
- **Small community attack**, exploits the known phenomenon of small communities in graphs

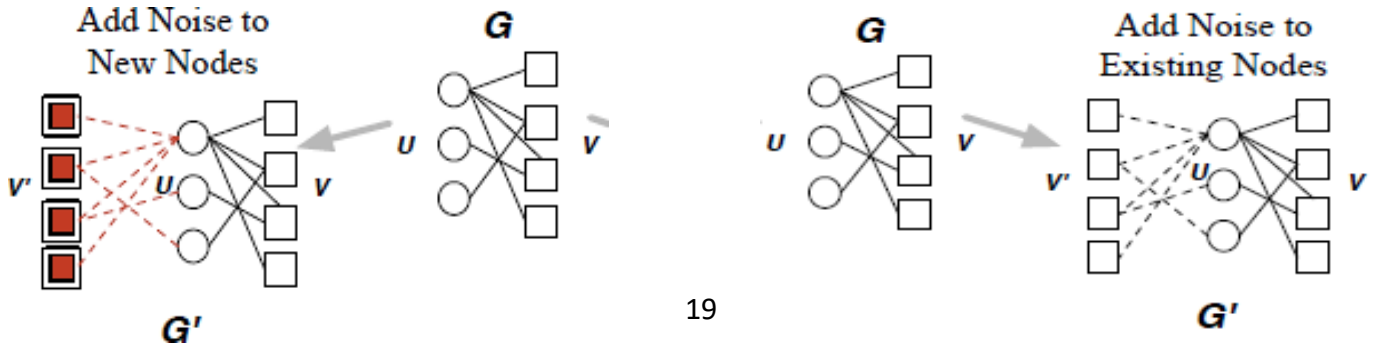




Targeted Noise Injection

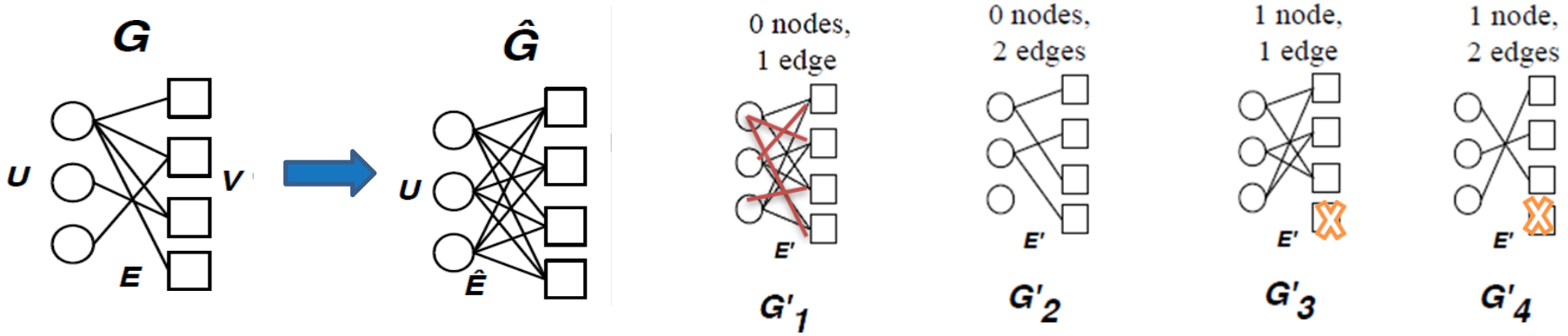
- the purpose of noise injection is mirroring real edges.
- bipartite attacker graph G
- vertex sets U (circles)
- V (squares)
- G -> G′
- attack function f : (u,v) ∈ E 7⟶ (u,v′) ∈ E′
- This creates G′ = (U,V ∪V ′, E ∪ E′)





Small Community

- an adversary first constructs a complete version of G, $\hat{G}$
- removes edges and/or nodes
- minimal knowledge -> n_v nodes and ne edges randomly



Datasets

1. Used for construct Host-NXDOMAIN Graph as ground truth without attack.
   - Collected from anonymized recursive DNS traffic from a large telecommunication company from December 18, 2016 to December 29, 2016.
   - contains NXDOMAINs queried by hosts and the query timestamps.
   - 262 thousand unique anonymized hosts, with 44.6 million queries to 1.8 million unique NXDOMAINs in a day
   - available to defenders and perfect knowledge attackers.

Datasets

2. Used as a surrogate network dataset.
   - NXDOMAIN traffic from a large US university network collected on December 25, 2016.
   - contains 8,782 hosts and 210 thousand unique NXDOMAINs.
   - available to attackers with moderate and perfect knowledge.

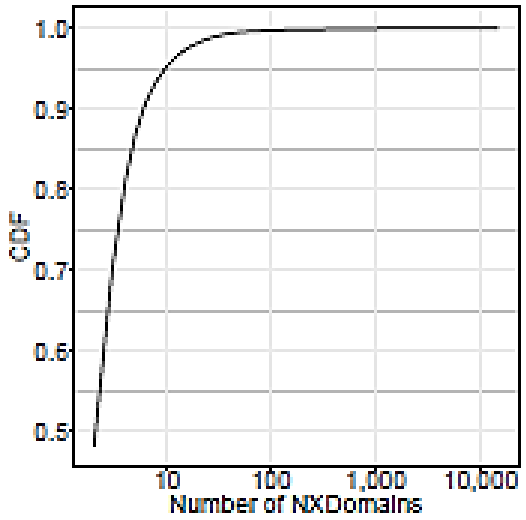
Datasets

3. reverse engineered DGA domains:
   - for 14 malware families
   - labeled 267 clusters belonging to four malware families
   - Were trained a Random Forest classifier with an average accuracy of 96.08%, and a false positive rate of 0.9%.

Dataset	Number of Records	Minimal	Moderate	Perfect
Reverse Engineered DGA Domains	14 DGA Families; 395 thousand NXD	X	X	X
Host-NXDOMAIN Graph (Surrogate)	8782 hosts; 210 thousand NXD	-	X	X
Host-NXDOMAIN Graph (Ground Truth)	average 262 thousand hosts; 1.8 million NXD	-	-	X

Attack Costs

- anomaly cost for noise injection -> by computing the cumulative distribution functions (CDF)
- adversarial cost behind the small community attack -> by change of attacker graph density D(G')



Results

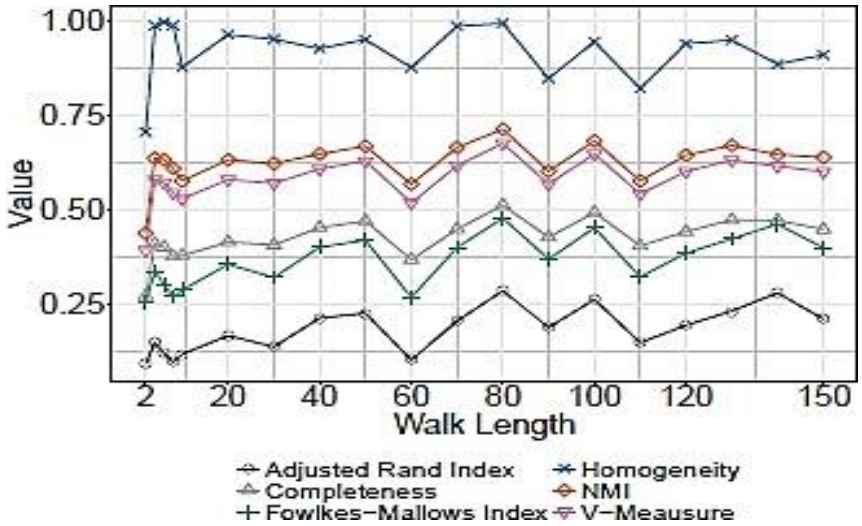
- select hyperparameters
- results for both attacks against each graph based clustering technique, for the three knowledge levels
- the costs incurred by the attacker, and how these can be used to identify possible defenses.

Community Discovery

- the best partition method from the Network X community discovery library
- Louvain algorithm
  - extracts good communities on the graph by optimizing the modularity metric
  - scales to large network with hundreds of millions of nodes

node2vec.

- traditional cluster validity metrics:
  - Adjusted Rand Index
  - Completeness
  - Fowlkes-Mallows index
  - Homogeneity
  - Normalized Mutual Information (NMI)
  - V-Measure score
- walk length



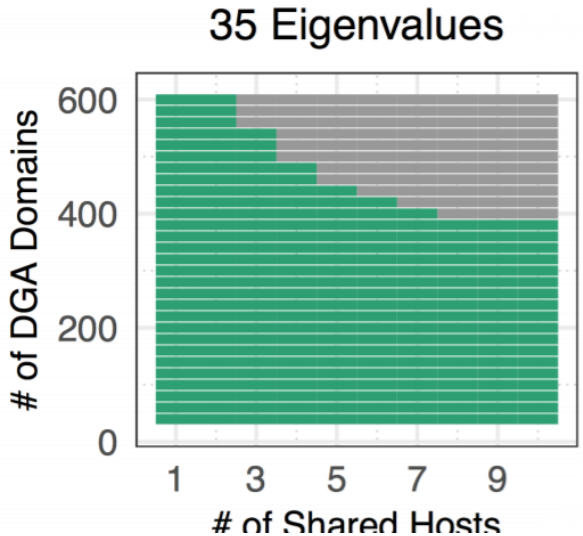
Targeted Noise Injection

- Four DGA families were identified:
  - Pykspa
  - Suppobox
  - Murofet
  - Gimemo.
- For each extracted:
  - the attacker graphs (G)
  - the target domains (V)

Small Community

- Spectral Clustering
  - choosen a group of 618 domains and 10 infected hosts belonging to Suppobox
  - success rate 75%+.
- Community Discovery
  - High cost
- Node2vec
  - success rate 70%+.

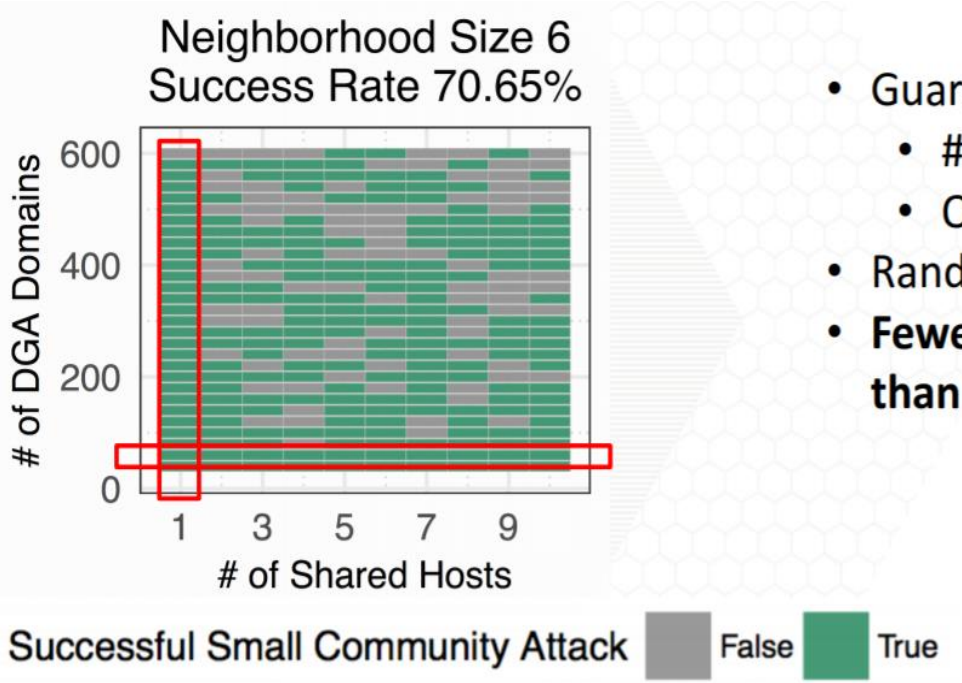
SPECTRAL Clustering



Different configurations:
Y nodes from V, X edges per remaining node

- **Minimal Knowledge**: Success Rate 75.16%
- **Perfect Knowledge**: Guaranteed Success
- **Moderate Knowledge**: Surrogate network dataset should be **smaller** than the original network dataset. (Section 5.3.4, Fig 10)

Node2Vec



- Guaranteed successful attack
  - # of shared Hosts = 1
  - Or <= 40 DGA Domains
- Randomness in the middle of the plot
- **Fewer guarantees and higher costs than Spectral Clustering.**

Agility Cost

- SVD rank = 35 -> Minimum cost = 0
- neighborhood sizes 2, 4, and 6 -> attack
- success rate 65.16%, 60.65%, and 70.65%

	Spectral Clustering		
	Density		
Join Death Star	Median	Maxium	Minimum Cost
SVD rank 35	0.078	0.61	0
SVD rank 50	0.11	0.45	0.03
SVD rank 80	0.065	0.26	0.22
SVD rank 100	0.052	0.19	0.29
SVD rank 200	0.0032	0.10	0.38
SVD rank 300	0.026	0.26	0.22
	node2vec		
	Density		
	Median	Maxium	Minimum Cost
Neighborhood Size 6	0.026	0.065	0.415

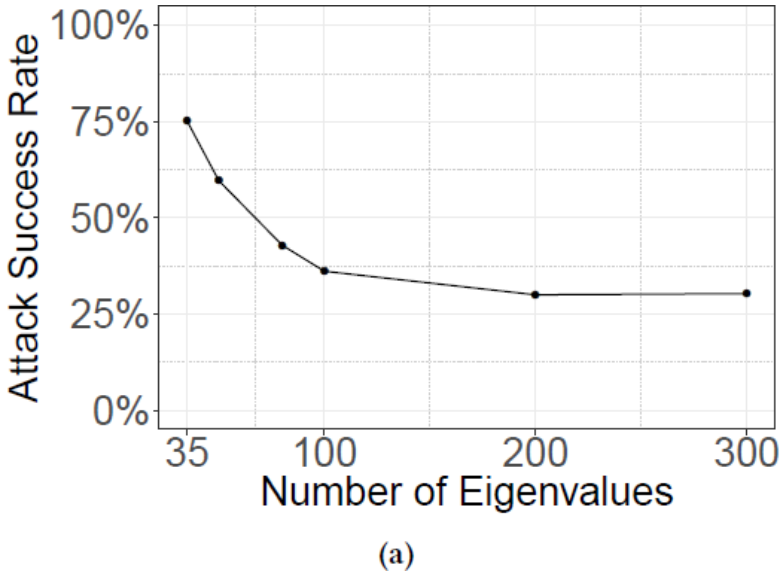
Defense

- Training Classifier with Noise
  - By retraining the classifier, it becomes more resistant to noise that could be injected by the adversary in the unsupervised phase of Pleiades.
  - It is important to note that this defense only trains the classifier with noise that has been witnessed.

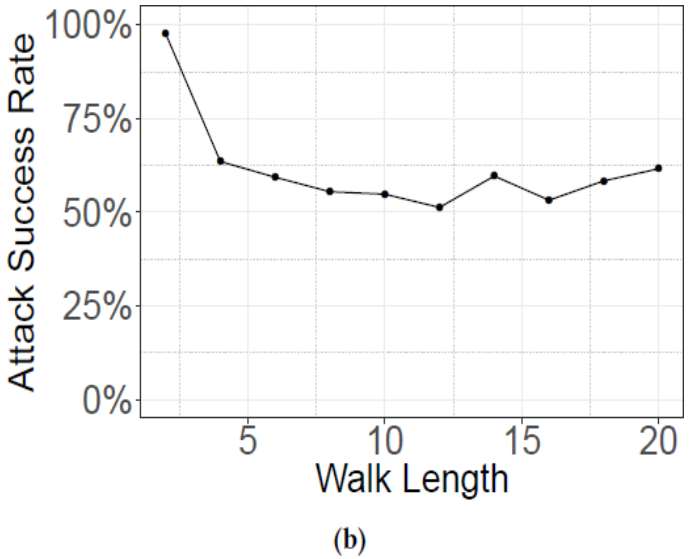
	False Positive Rate			
Model	Pykspa	Gimemo	Suppobox	Murofet
Original	0.32%	0.29%	0%	0%
Model A	1.64%	0.39%	0.10%	0%
Model B	1.62%	0.10%	1.23%	0.30%
Model C	1.46%	1.17%	1.23%	0%

Improving Hyperparameter Selection

- neighborhood size and walk length are not optimal hyperparameters.
- We can choose more resistant hyperparameters



(a)



(b)

Conclusions

- the **first practical attempt** to attack against graph-based clustering techniques and **a global feature** space where realistic attackers without perfect knowledge.
- As a result were designed and evaluated two novel graph attacks against a state-of-the-art network level, graph-based detection system.
- Low Attack Cost
- were focused on adversarial clustering, which deals with global features that cannot be directly changed.
- Defenses.

Questions