Seeds of Scanning: Understanding the Effects of Datasets, Methods, and Metrics on IPv6 Address Discovery

Grant Williams and Paul Pearce

Georgia Tech

We evaluate 8 IPv6 Scanning Target Generation Algorithms (TGAs) across datasets, ports, protocols and metrics, in order to develop TGA best practices.

Internet Scanning

- Internet scanning involves connecting to devices open on the Internet on a certain port/protocol.

Internet Scanning

- Internet scanning involves connecting to devices open on the Internet on a certain port/protocol.

Internet Scanning

- Internet scanning involves connecting to devices open on the Internet on a certain port/protocol.

- IPv6 scanning is more difficult (2^{128} possible IPv6 addresses)

Internet Scanning

- Internet scanning involves connecting to devices open on the Internet on a certain port/protocol.

- IPv6 scanning is more difficult (2^{128} possible IPv6 addresses)

- Brute Force approach is no longer viable (we have to find other ways of discovering responsive addresses)

Internet Scanning

- Internet scanning involves connecting to devices open on the Internet on a certain port/protocol.

- IPv6 scanning is more difficult (2^{128} possible IPv6 addresses)

- Brute Force approach is no longer viable (we have to find other ways of discovering responsive addresses)

- Common Approach: **Target Generation Algorithms (TGAs)**

TGAs

Input

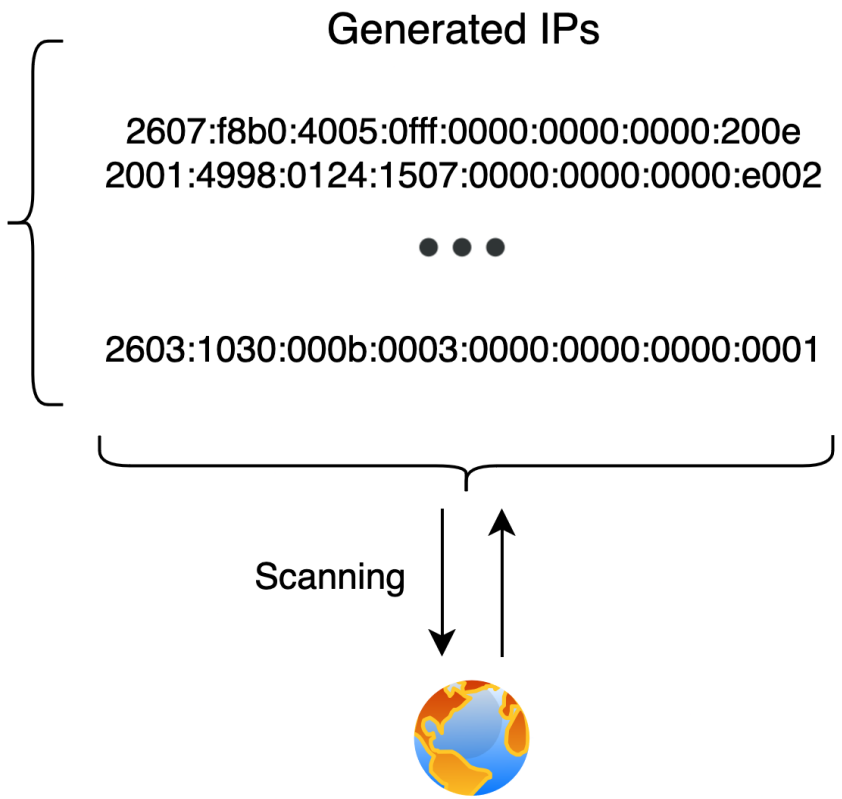
2607:f8b0:4005:080f:0000:0000:0000:200e
2001:4998:0124:1507:0000:0000:0000:f001
2603:1030:000b:0003:0000:0000:0000:0152

TGAs

Input

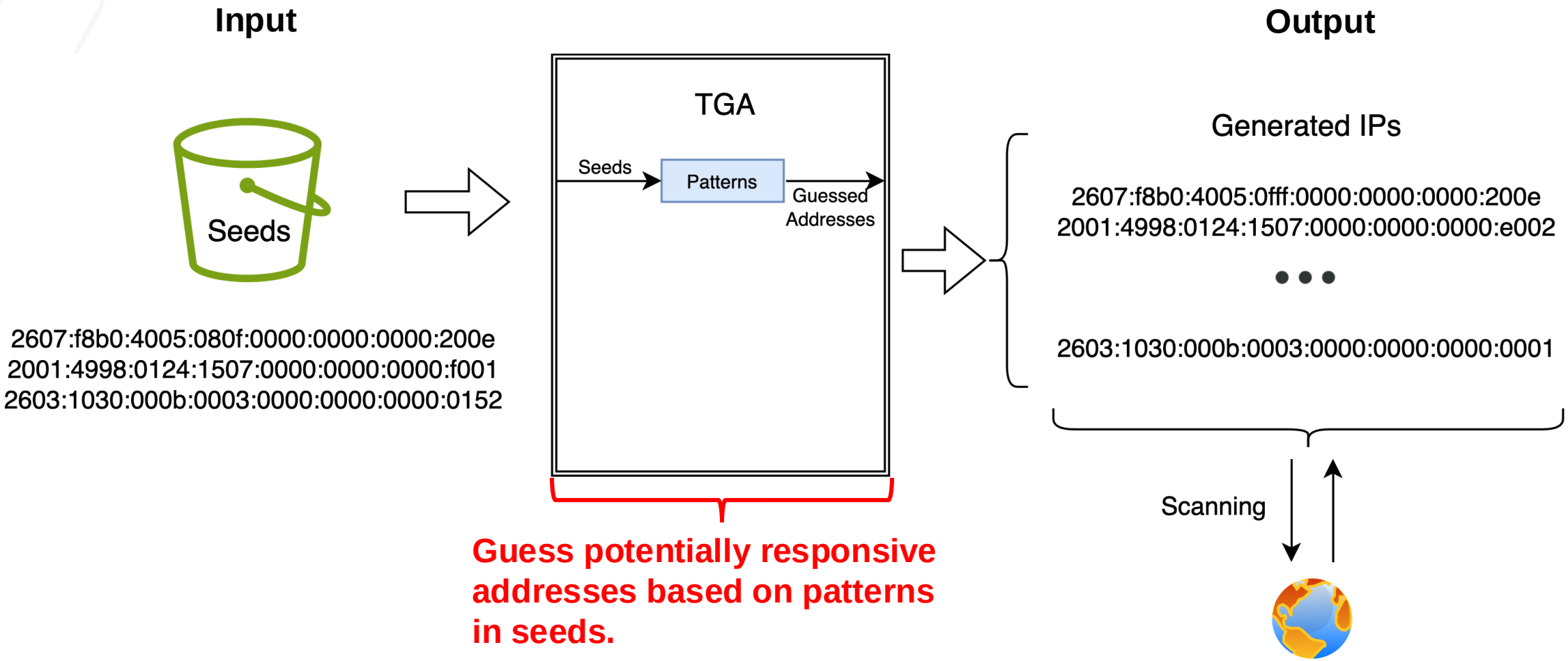
2607:f8b0:4005:080f:0000:0000:0000:200e
2001:4998:0124:1507:0000:0000:0000:f001
2603:1030:000b:0003:0000:0000:0000:0152

Output



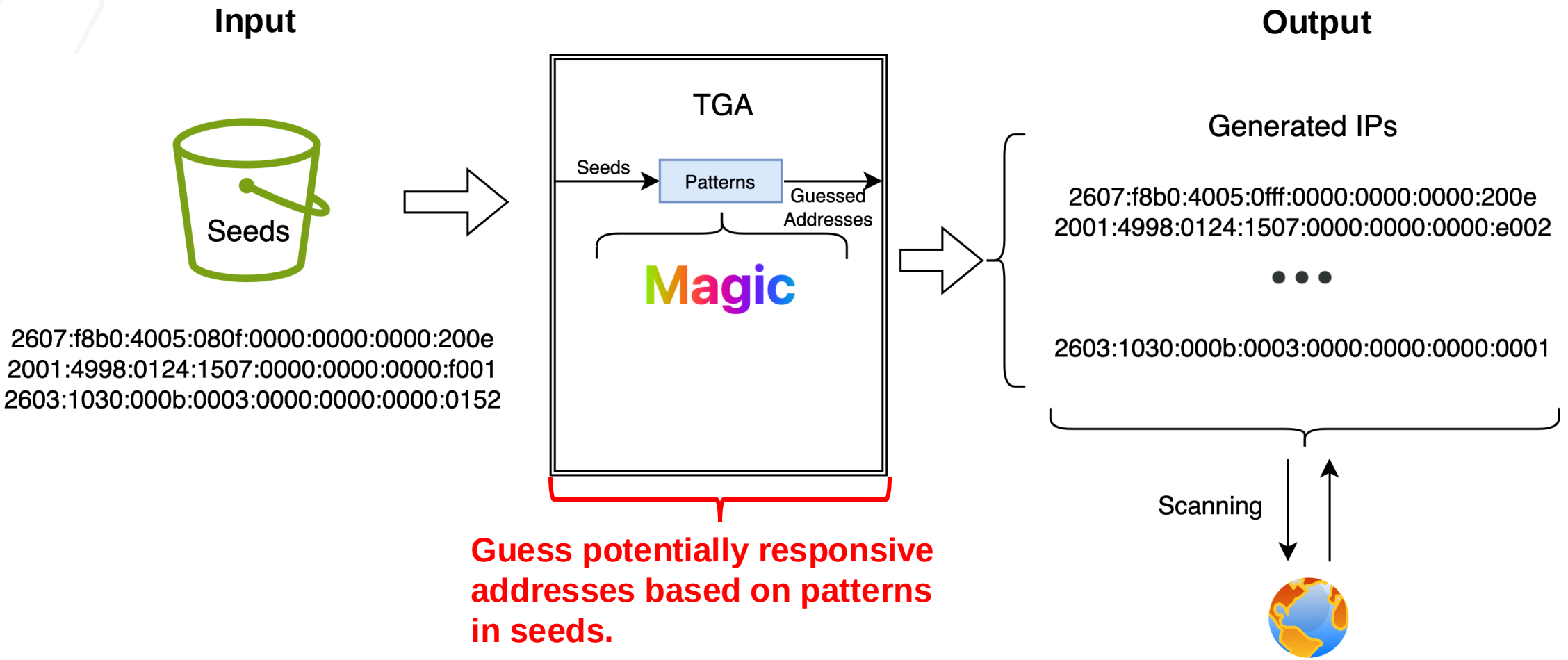
Georgia Tech

TGAs



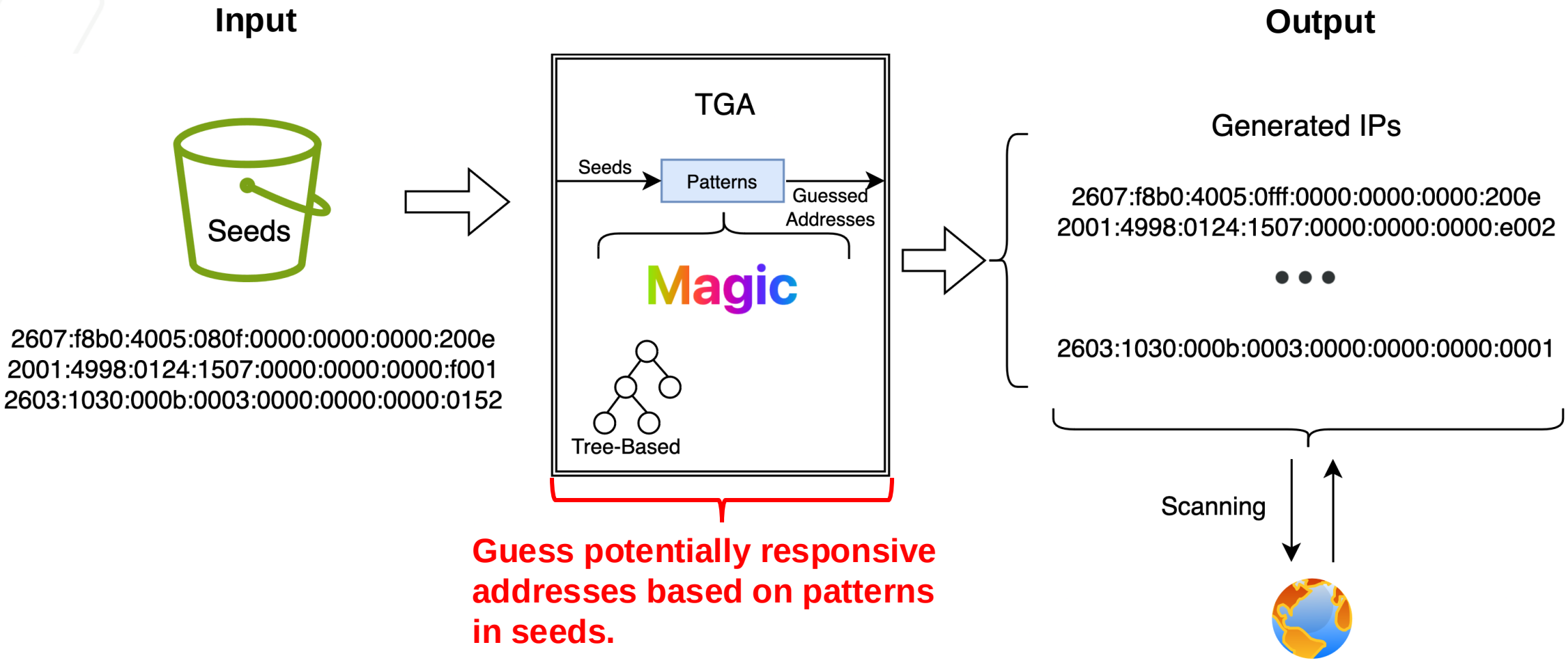
Georgia Tech

TGAs



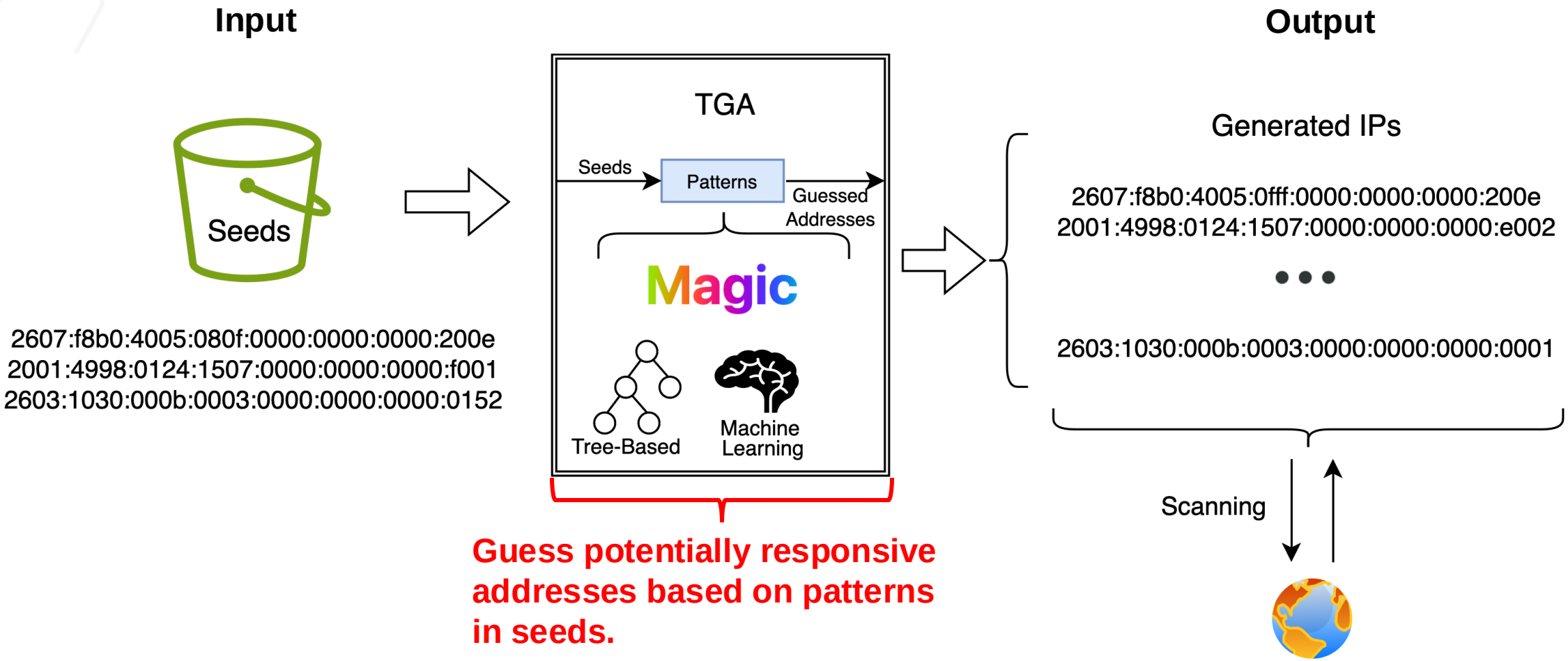
Georgia Tech

TGAs



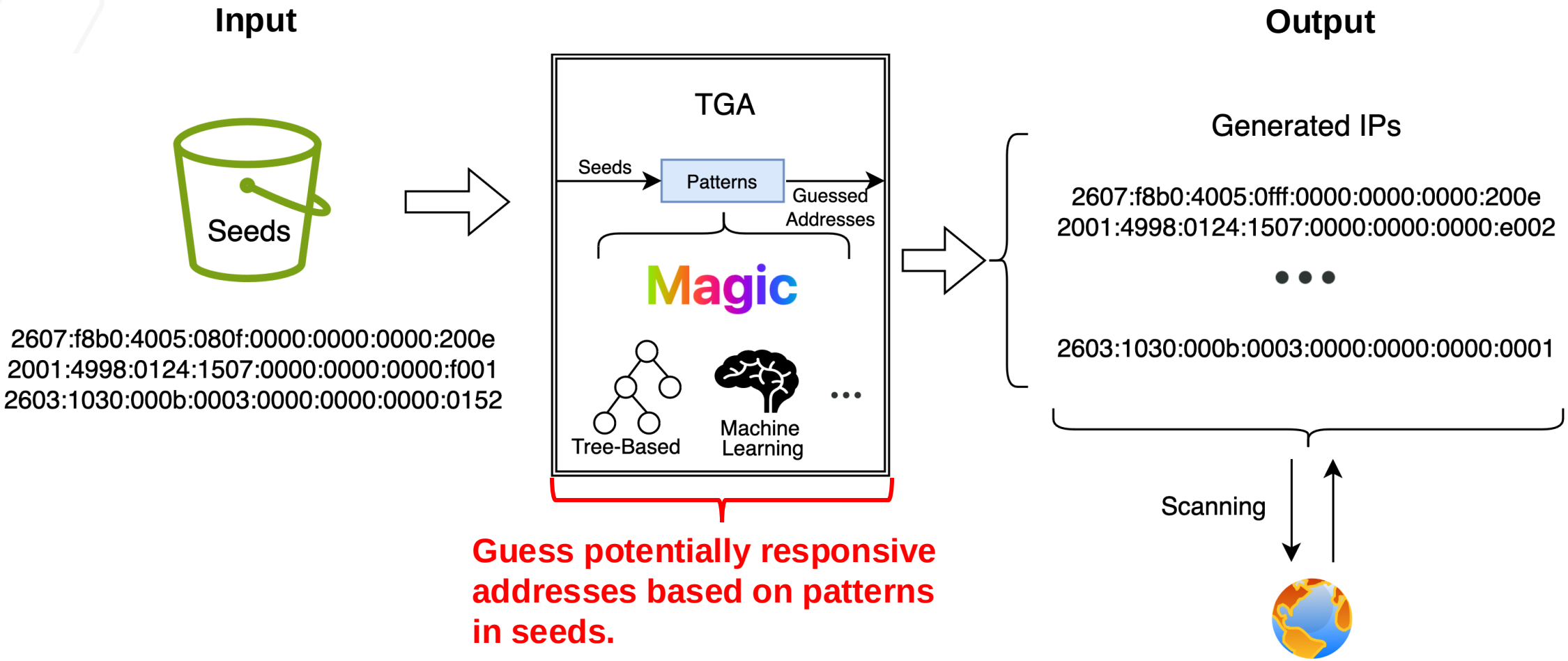
Georgia Tech

TGAs



Georgia Tech

TGAs

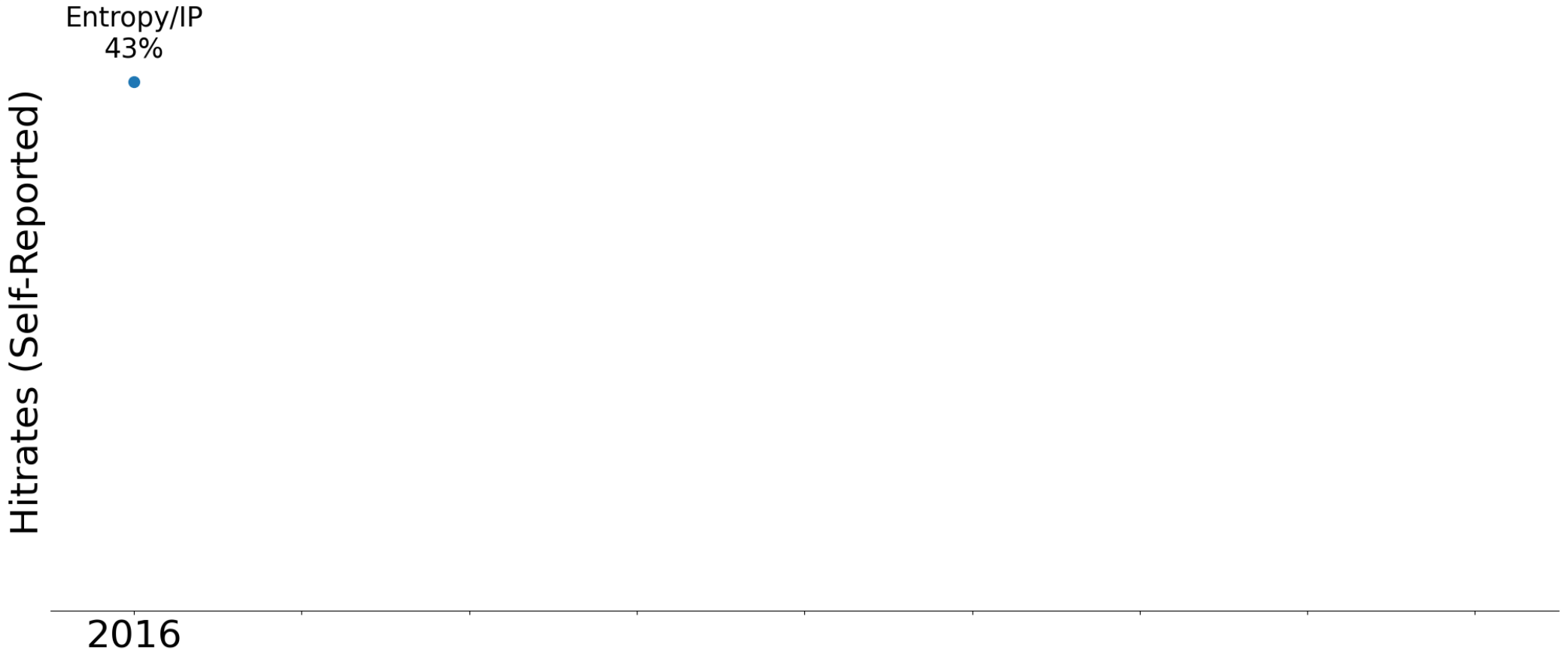


Timeline of TGA Hitrates

Typically, TGAs are evaluated by their **hitrate** (percentage of guessed addresses that are responsive)

Timeline of TGA Hitrates

Typically, TGAs are evaluated by their **hitrate** (percentage of guessed addresses that are responsive)



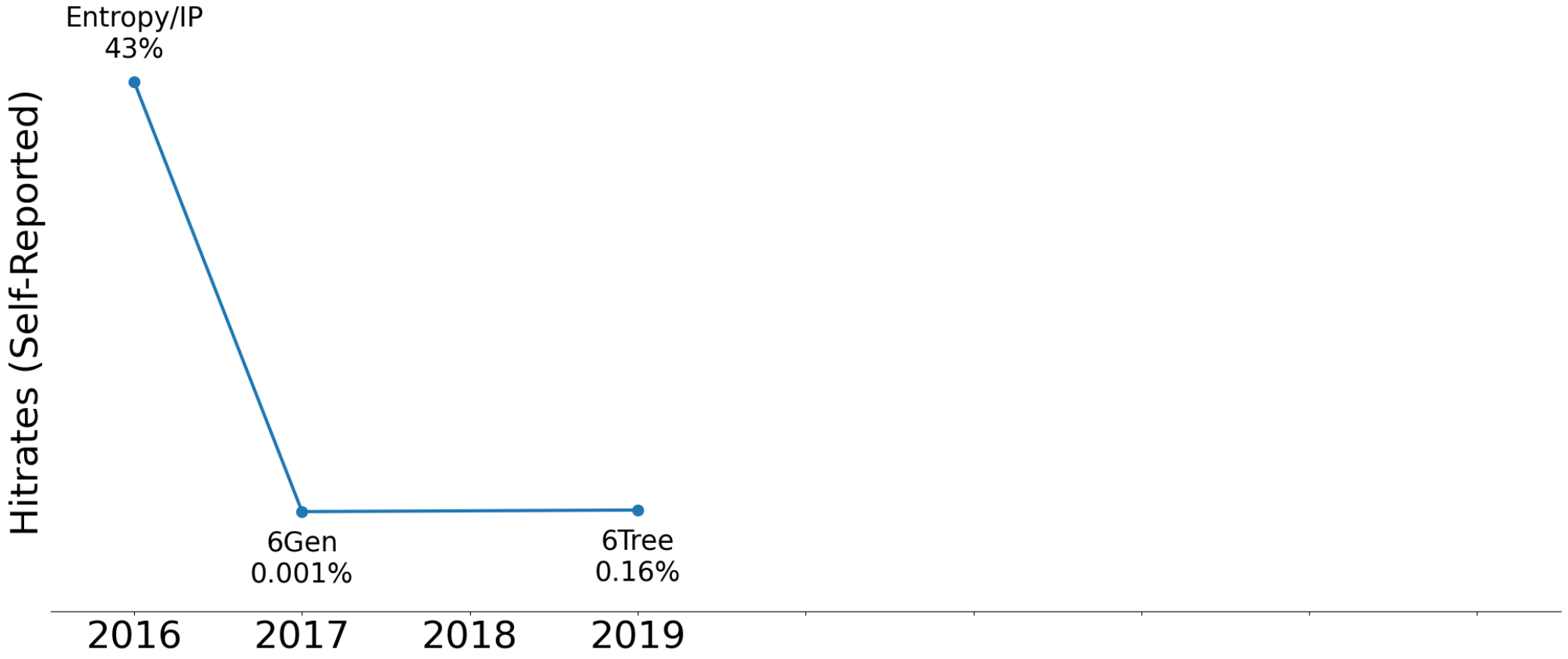
Timeline of TGA Hitrates

Typically, TGAs are evaluated by their **hitrate** (percentage of guessed addresses that are responsive)

Georgia Tech

Timeline of TGA Hitrates

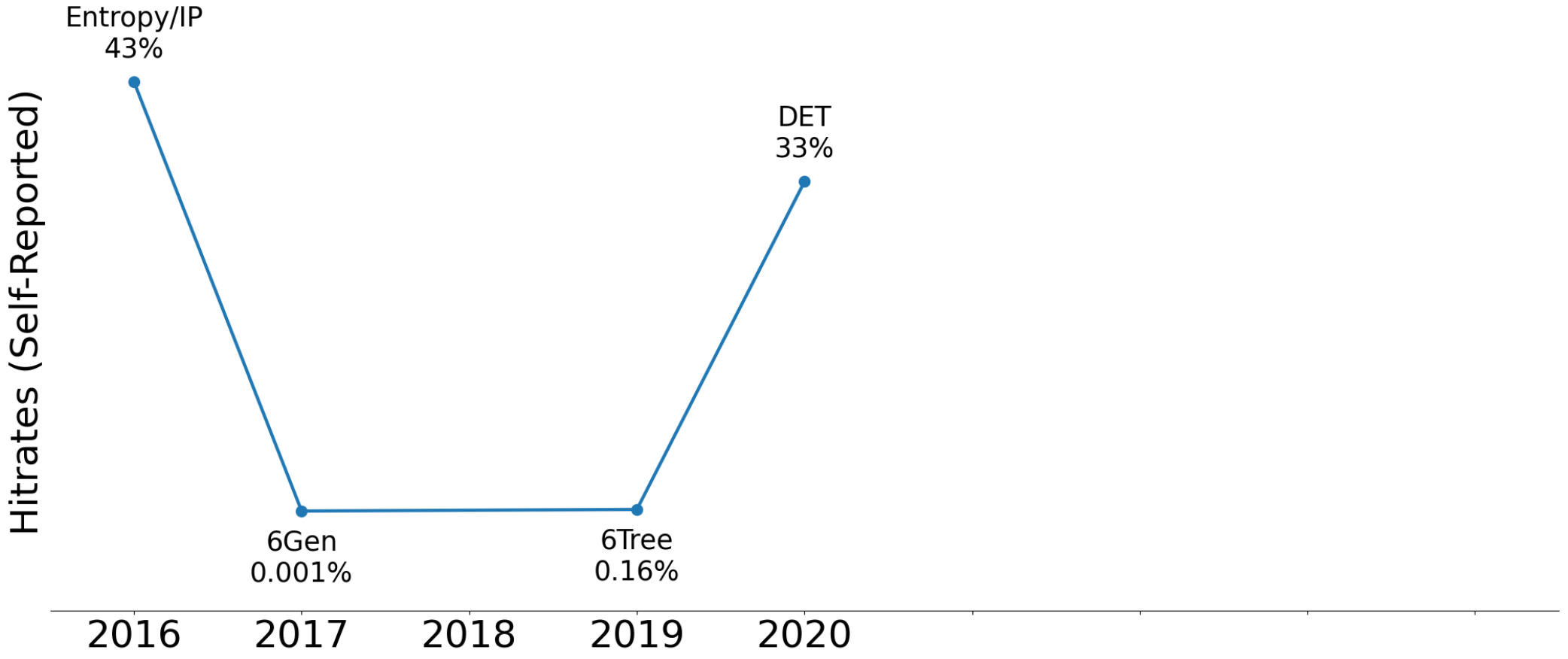
Typically, TGAs are evaluated by their **hitrate** (percentage of guessed addresses that are responsive)



Georgia Tech

Timeline of TGA Hitrates

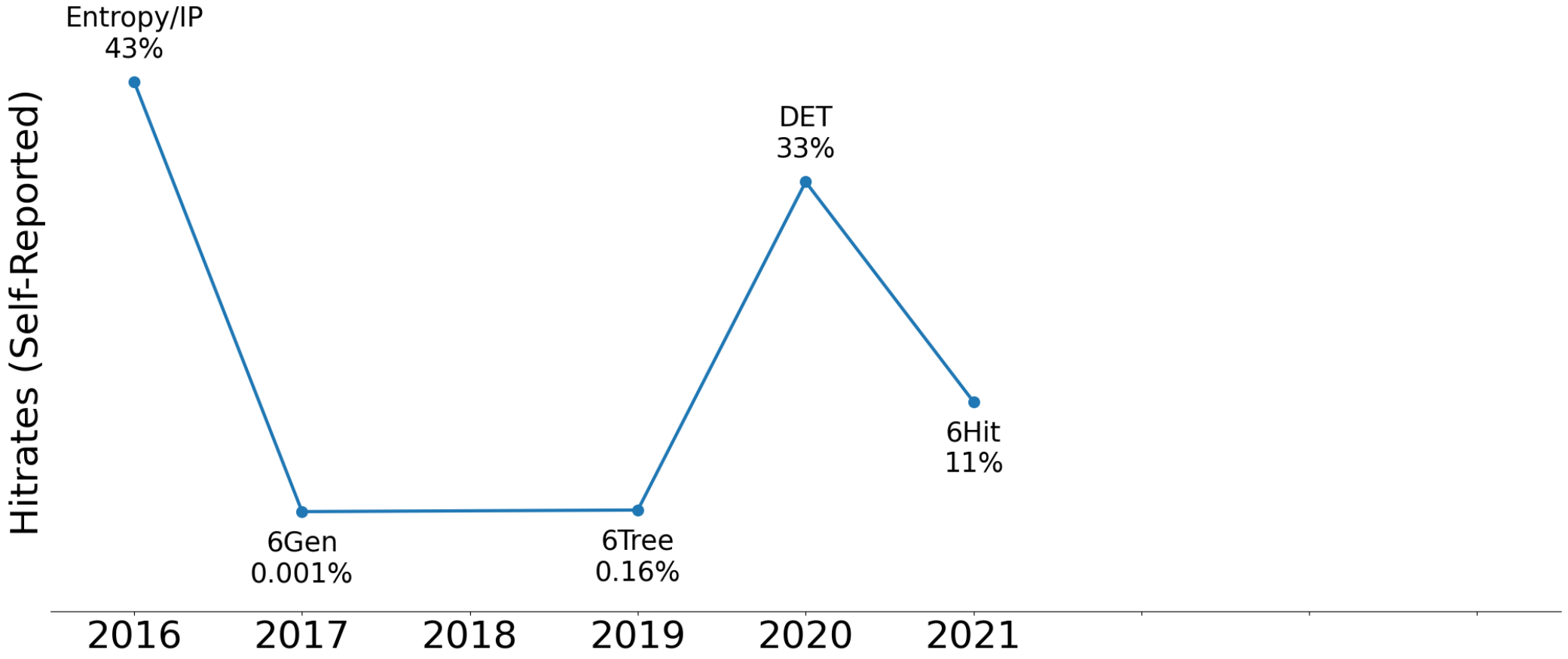
Typically, TGAs are evaluated by their **hitrate** (percentage of guessed addresses that are responsive)



Georgia Tech

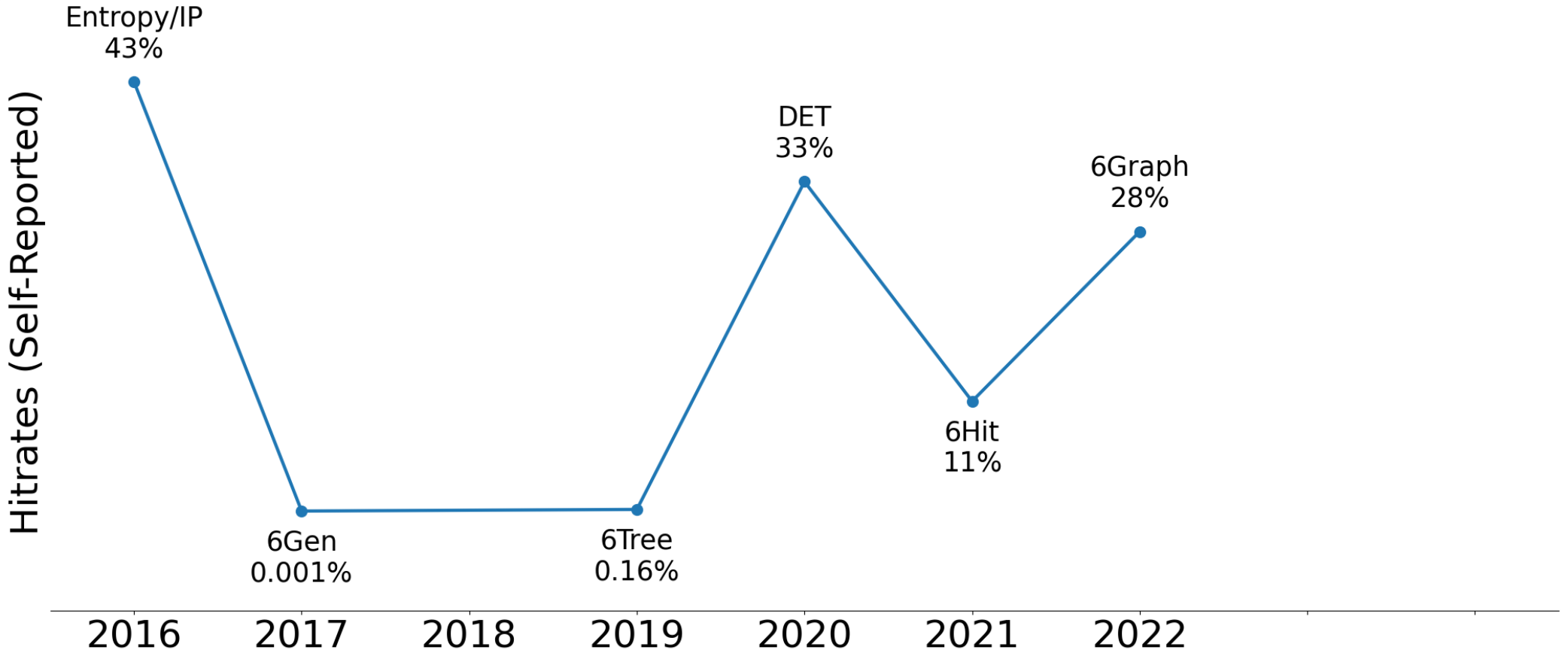
Timeline of TGA Hitrates

Typically, TGAs are evaluated by their **hitrate** (percentage of guessed addresses that are responsive)



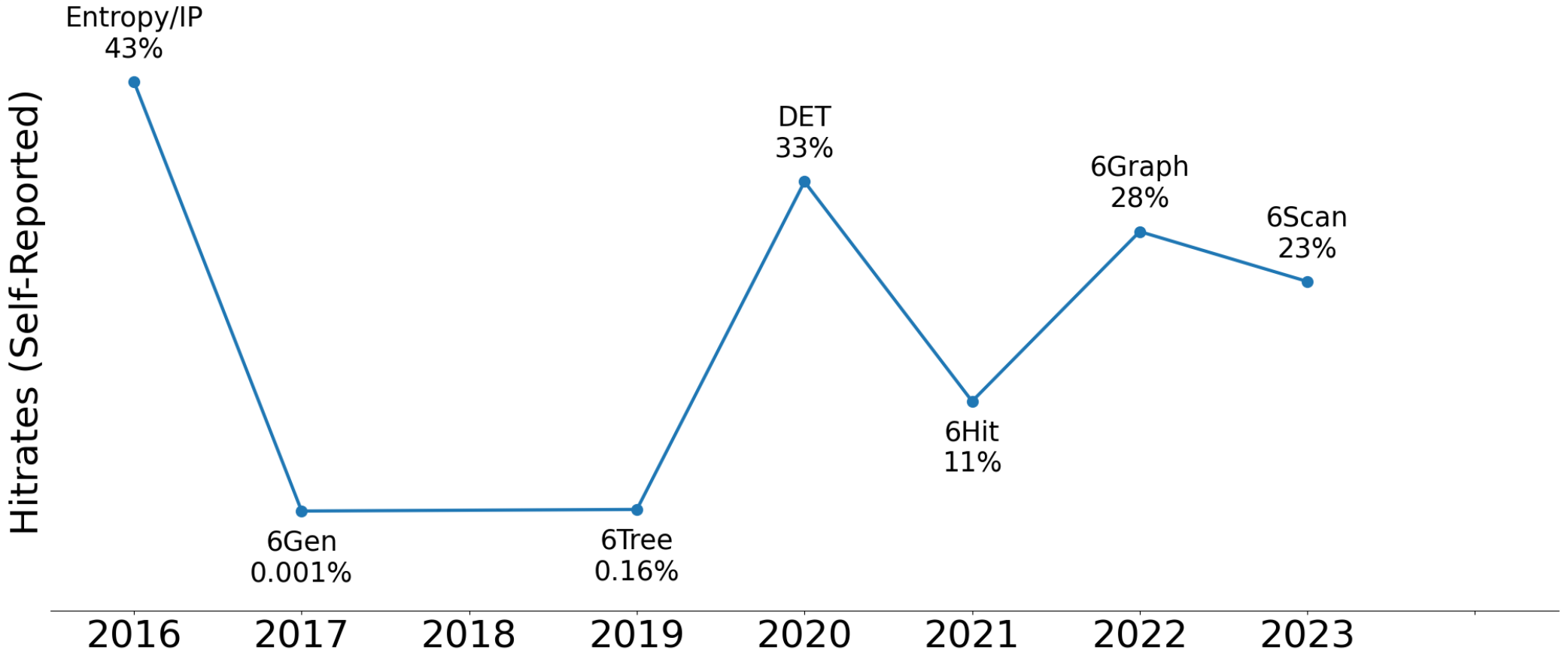
Timeline of TGA Hitrates

Typically, TGAs are evaluated by their **hitrate** (percentage of guessed addresses that are responsive)



Timeline of TGA Hitrates

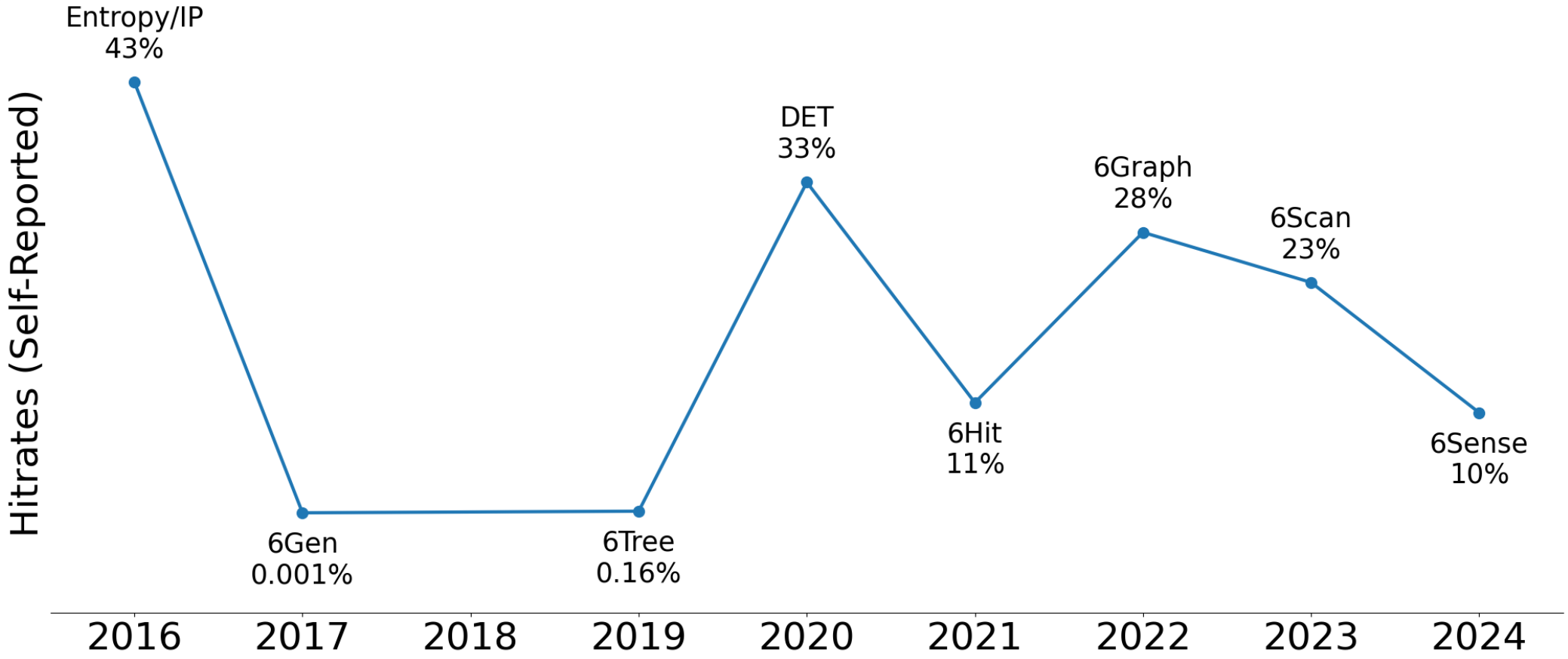
Typically, TGAs are evaluated by their **hitrate** (percentage of guessed addresses that are responsive)



Georgia Tech

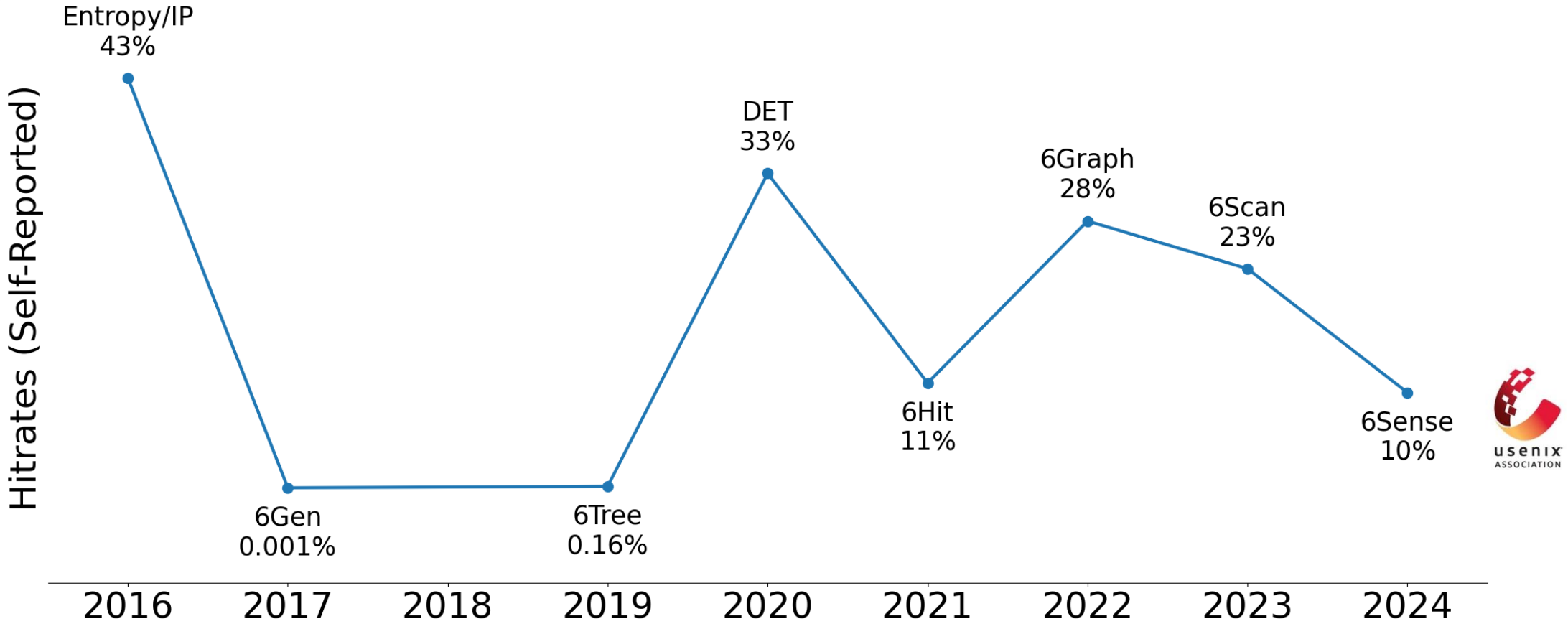
Timeline of TGA Hitrates

Typically, TGAs are evaluated by their **hitrate** (percentage of guessed addresses that are responsive)



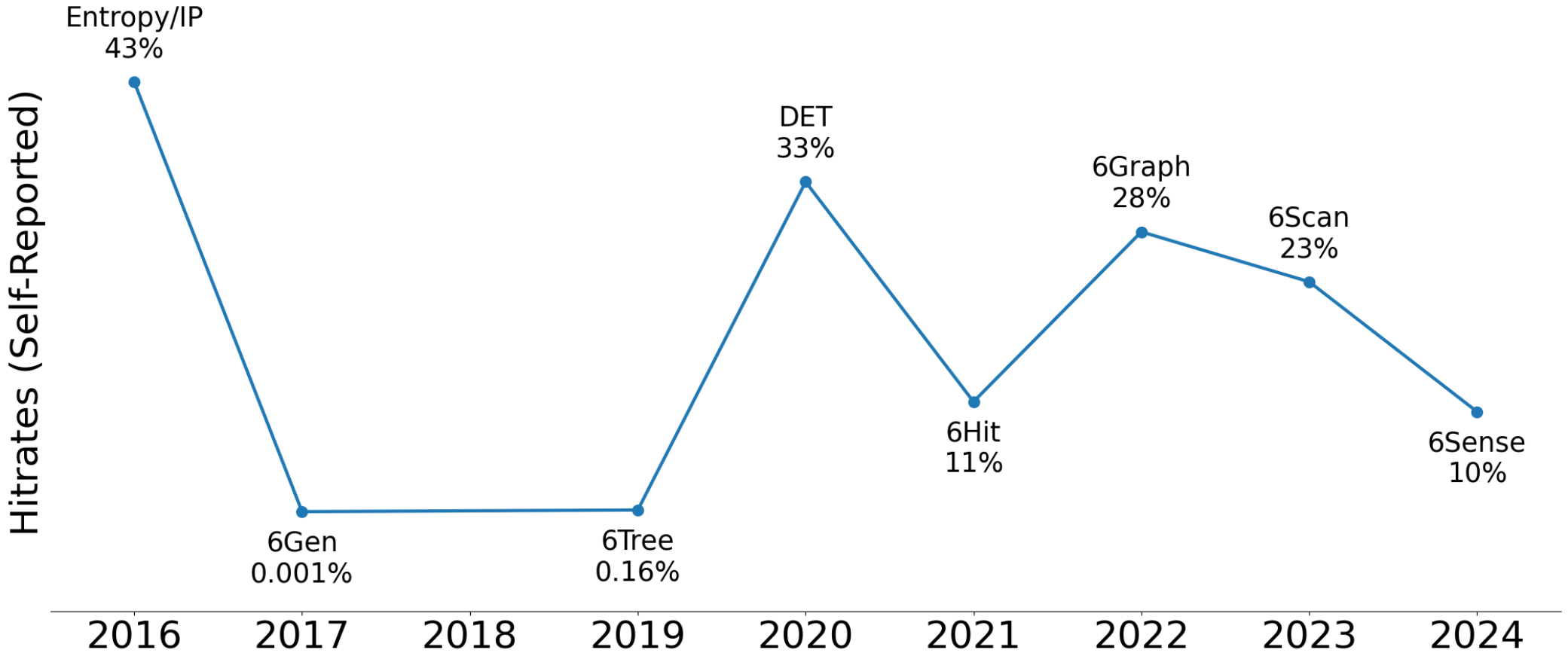
Timeline of TGA Hitrates

Typically, TGAs are evaluated by their **hitrate** (percentage of guessed addresses that are responsive)



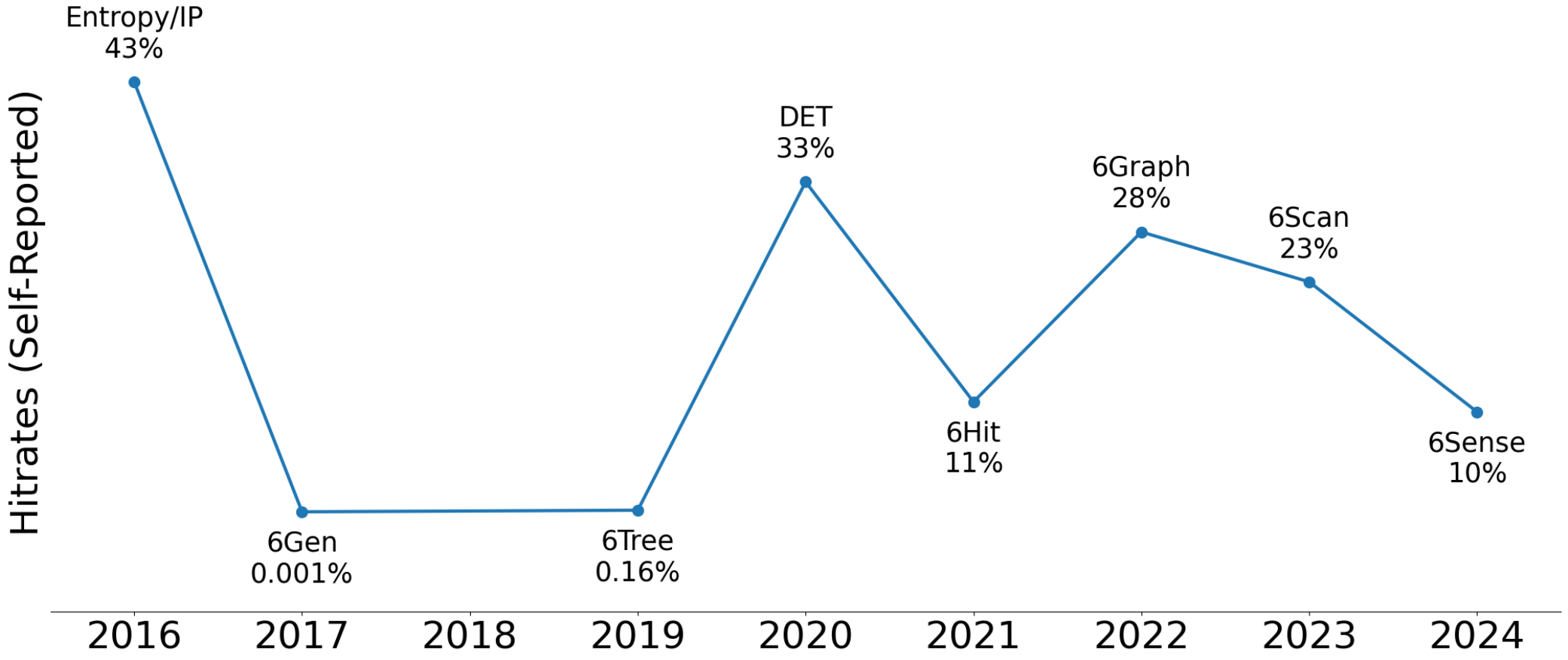
Timeline of TGA Hitrates

Typically, TGAs are evaluated by their **hitrate** (percentage of guessed addresses that are responsive)



Timeline of TGA Hitrates

Typically, TGAs are evaluated by their **hitrate** (percentage of guessed addresses that are responsive)



Why don’t hitrates go up????

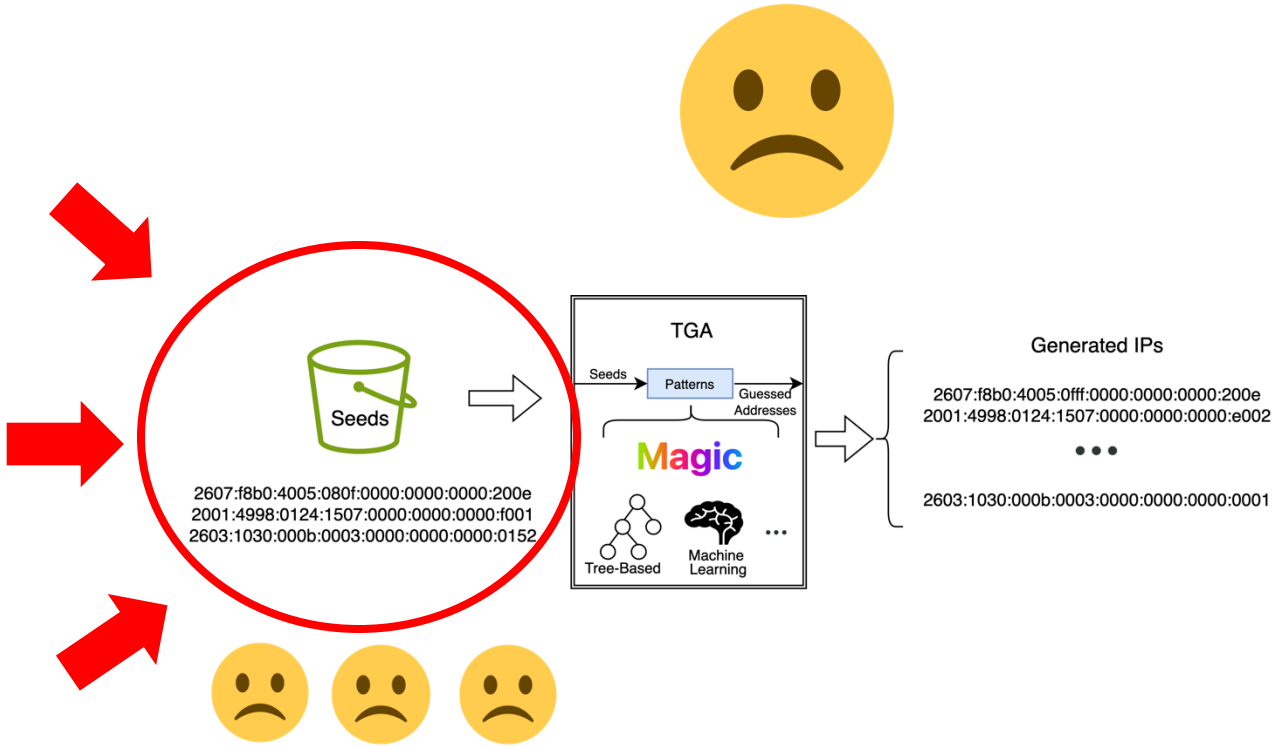
Georgia Tech

What's Going on?

We need consistency in evaluation methodology, metrics and input datasets!

What's Going on?

We need consistency in evaluation methodology, metrics and input datasets!



Our Goal: Define best practices for running TGAs, specifically around seed dataset design.

Our Goal: Define best practices for running TGAs, specifically around seed dataset design.

Some Prior Work on Evaluating Datasets (Target Acquired, Steger, 2023)

- Looked primarily at TGAs the IPv6 Hitlist
- Limited analysis of multiple ports, protocols, metrics, and preprocessing.

Our Goal: Define best practices for running TGAs, specifically around seed dataset design.

We evaluate TGAs against a variety of seed datasets designed with different preparing methods and evaluate them using multiple metrics.

Method

- We evaluate **8 TGAs**: 6Sense, 6Tree, 6Hit, 6Scan, 6Graph, 6Gen, DET, and Entropy/IP with **21 seed dataset variations.**

Method

- We evaluate **8 TGAs**: 6Sense, 6Tree, 6Hit, 6Scan, 6Graph, 6Gen, DET, and Entropy/IP with **21 seed dataset variations.**

- We scan **4 ports/protocols** (ICMP, TCP80, TCP443, and UDP53)

Method

- We evaluate **8 TGAs**: 6Sense, 6Tree, 6Hit, 6Scan, 6Graph, 6Gen, DET, and Entropy/IP with **21 seed dataset variations.**
- We scan **4 ports/protocols** (ICMP, TCP80, TCP443, and UDP53)
- We Generate **50M addresses** per dataset per TGA per port/protocol.

Method

- We evaluate **8 TGAs**: 6Sense, 6Tree, 6Hit, 6Scan, 6Graph, 6Gen, DET, and Entropy/IP with **21 seed dataset variations.**
- We scan **4 ports/protocols** (ICMP, TCP80, TCP443, and UDP53)
- We Generate **50M addresses** per dataset per TGA per port/protocol.
- We evaluate with **3 metrics**:

Method

- We evaluate **8 TGAs**: 6Sense, 6Tree, 6Hit, 6Scan, 6Graph, 6Gen, DET, and Entropy/IP with **21 seed dataset variations.**
- We scan **4 ports/protocols** (ICMP, TCP80, TCP443, and UDP53)
- We Generate **50M addresses** per dataset per TGA per port/protocol.
- We evaluate with **3 metrics**:
  - Hits

Method

- We evaluate **8 TGAs**: 6Sense, 6Tree, 6Hit, 6Scan, 6Graph, 6Gen, DET, and Entropy/IP with **21 seed dataset variations.**

- We scan **4 ports/protocols** (ICMP, TCP80, TCP443, and UDP53)

- We Generate **50M addresses** per dataset per TGA per port/protocol.

- We evaluate with **3 metrics**:
  - Hits
  - Autonomous Systems with hits (ASes)

Method

- We evaluate **8 TGAs**: 6Sense, 6Tree, 6Hit, 6Scan, 6Graph, 6Gen, DET, and Entropy/IP with **21 seed dataset variations.**

- We scan **4 ports/protocols** (ICMP, TCP80, TCP443, and UDP53)

- We Generate **50M addresses** per dataset per TGA per port/protocol.

- We evaluate with **3 metrics**:
  - ⊕ Hits
  - ⊕ Autonomous Systems with hits (ASes)
  - ⊖ Aliased Hits

Method

- We evaluate **8 TGAs**: 6Sense, 6Tree, 6Hit, 6Scan, 6Graph, 6Gen, DET, and Entropy/IP with **21 seed dataset variations.**

- We scan **4 ports/protocols** (ICMP, TCP80, TCP443, and UDP53)

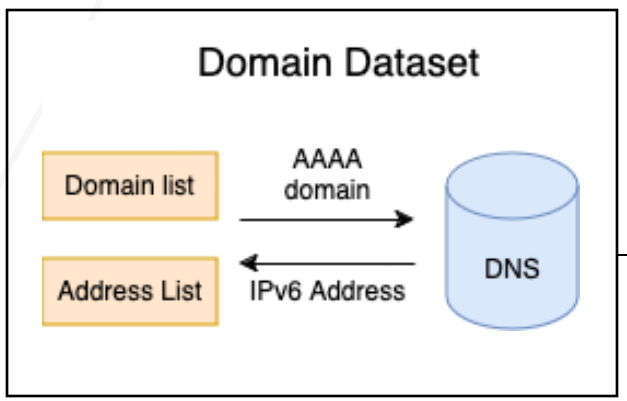
- We Generate **50M addresses** per dataset per TGA per port/protocol.

- We evaluate with **3 metrics**:
  - ⊕ Hits
  - ⊕ Autonomous Systems with hits (ASes)
  - ⊖ Aliased Hits

- We use seeds collected from **12 different dataset sources** (of 4 types).

Dataset Sources

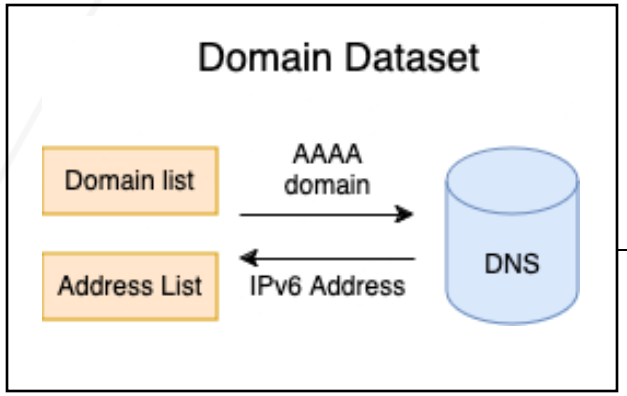
1.)



- CT Logs
- Rapid7
- Umbrella
- Majestic
- Tranco
- SecRank
- Radar
- CAIDA DNS

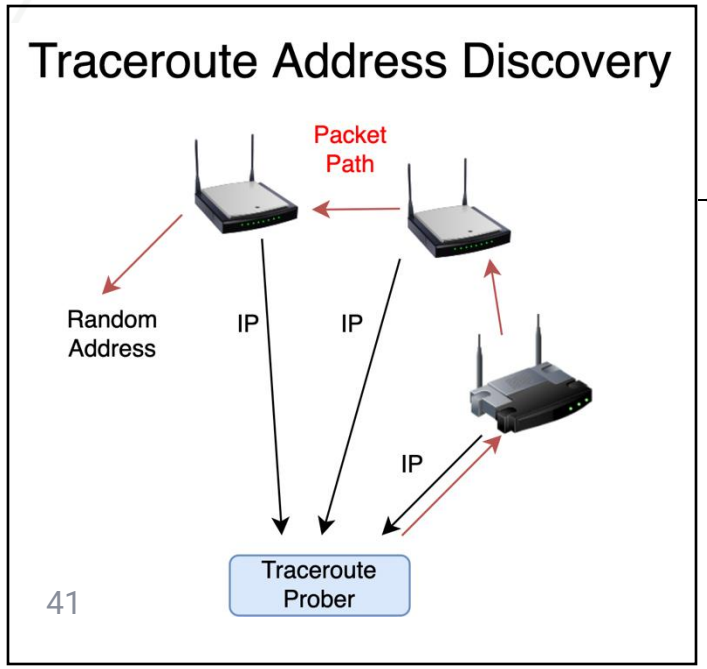
Dataset Sources

1.)



- CT Logs
- Rapid7
- Umbrella
- Majestic
- Tranco
- SecRank
- Radar
- CAIDA DNS

2.)

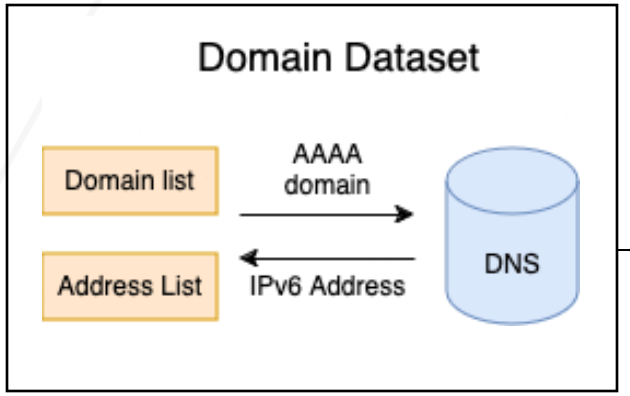


- Scamper
- RIPE Atlas

Georgia Tech

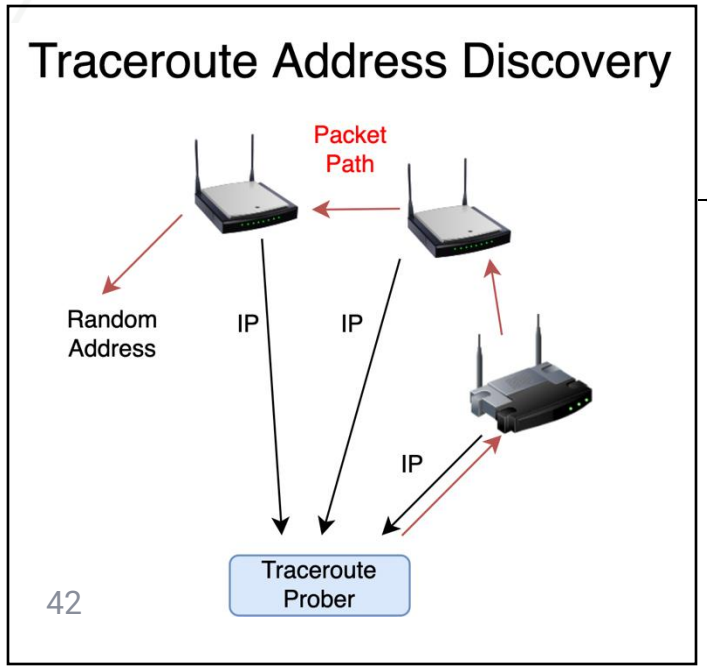
Dataset Sources

1.)



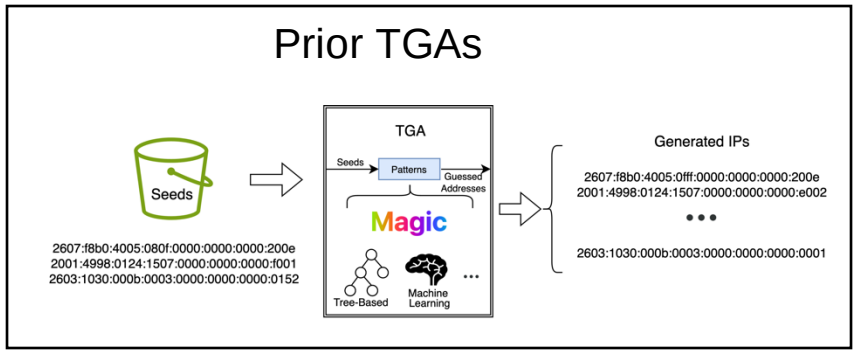
- CT Logs
- Rapid7
- Umbrella
- Majestic
- Tranco
- SecRank
- Radar
- CAIDA DNS

2.)



- Scamper
- RIPE Atlas

3.)

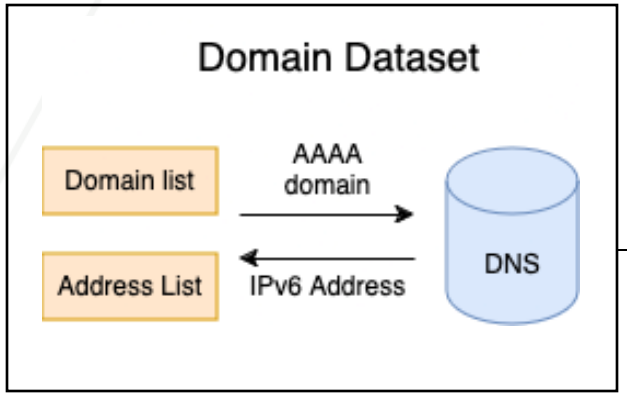


- AddrMiner

Georgia Tech

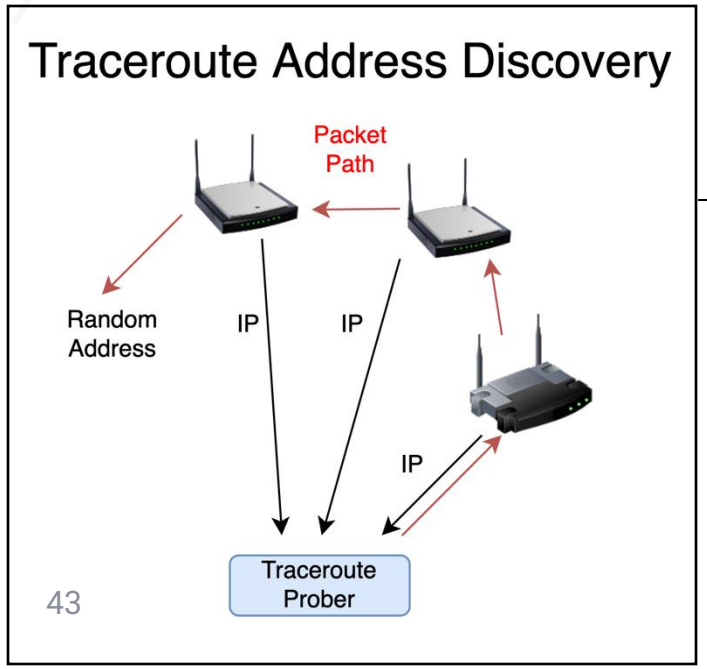
Dataset Sources

1.)



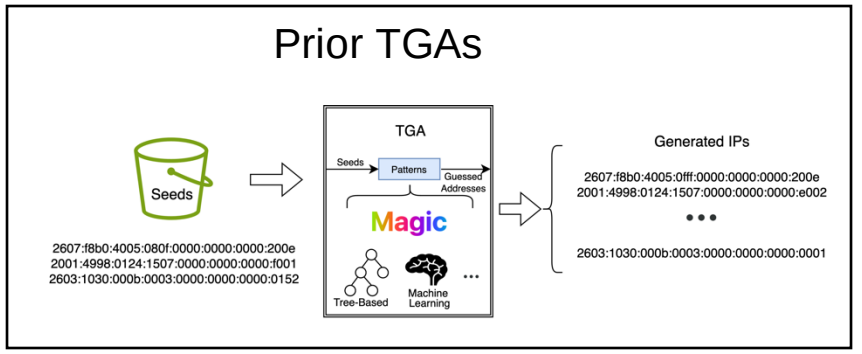
- CT Logs
- Rapid7
- Umbrella
- Majestic
- Tranco
- SecRank
- Radar
- CAIDA DNS

2.)



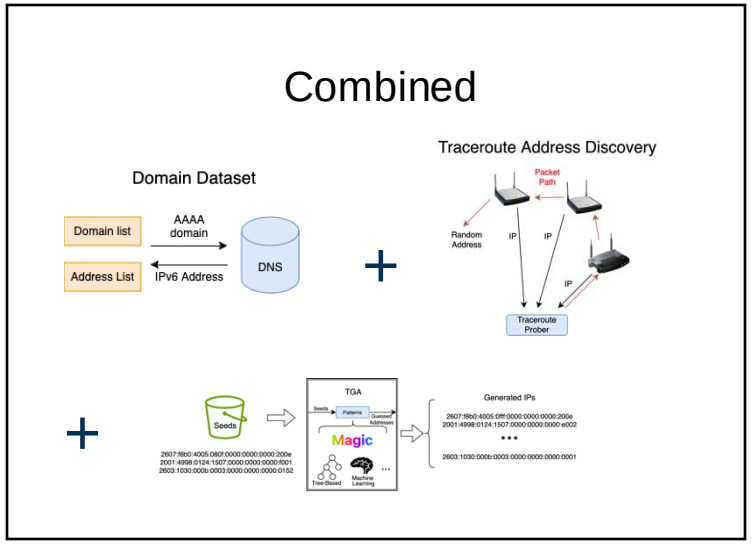
- Scamper
- RIPE Atlas

3.)



- AddrMiner

4.)

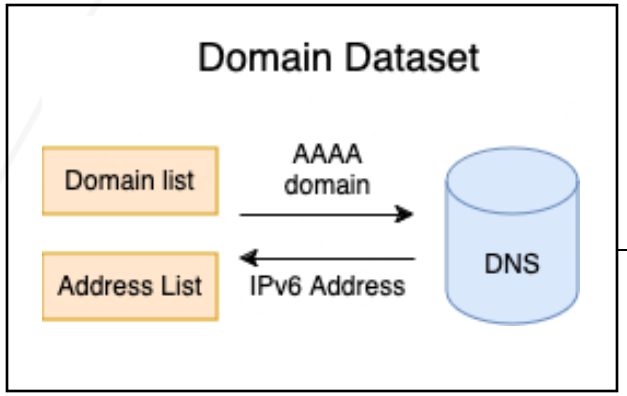


- IPv6 Hitlist

Georgia Tech

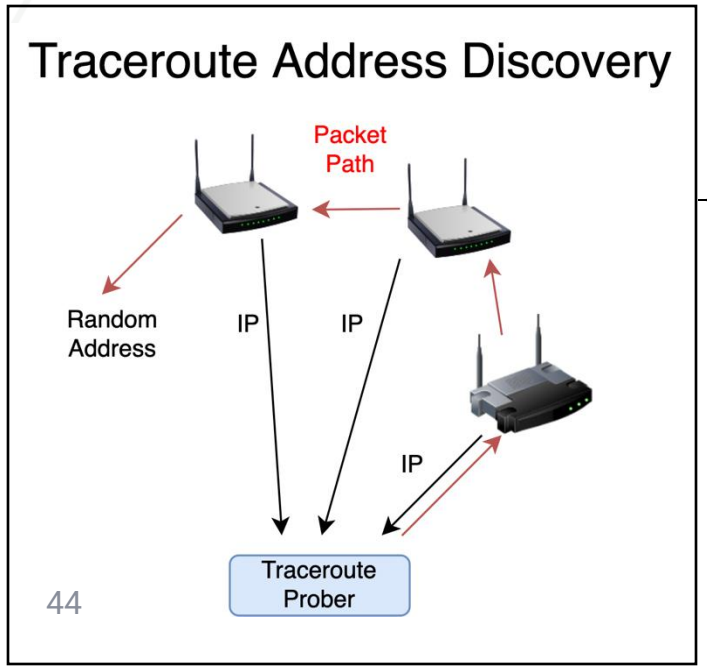
Dataset Sources

1.)



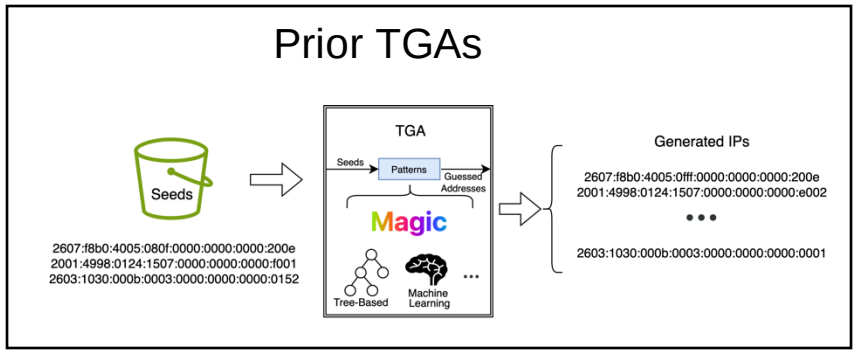
- CT Logs
- Rapid7
- Umbrella
- Majestic
- Tranco
- SecRank
- Radar
- CAIDA DNS

2.)



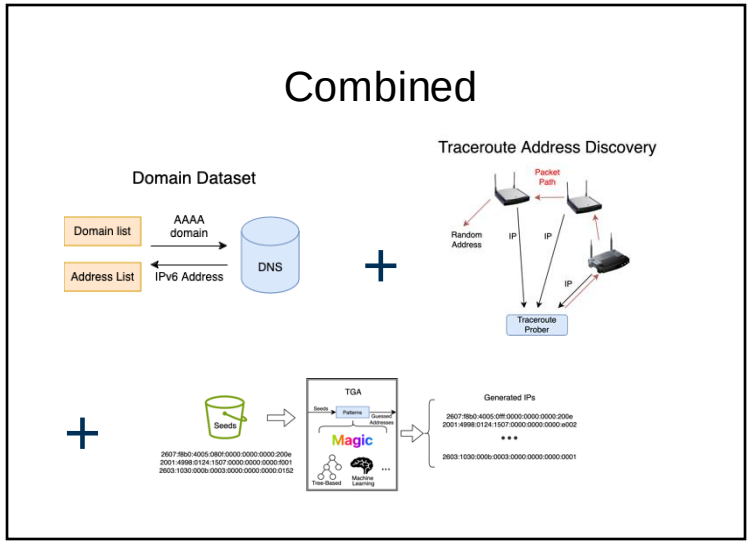
- Scamper
- RIPE Atlas

3.)



- AddrMiner

4.)



- IPv6 Hitlist

Dataset	Total	Responsive	Responsive ASes
All	118.7M	10.9M	23.6K

Georgia Tech

We Have Our Datasets and Methodology… Now for Experiments!

We Have Our Datasets and Methodology… Now for Experiments!

1. Explore how dataset cleaning impacts TGA results.

We Have Our Datasets and Methodology… Now for Experiments!

1. Explore how dataset cleaning impacts TGA results.

2. Combining results across multiple generators.

We Have Our Datasets and Methodology… Now for Experiments!

1. Explore how dataset cleaning impacts TGA results.
2. Combining results across multiple generators.
3. Comparing results across source datasets.

Experiments - Seed Dataset Cleaning

Dataset	Hits	ASes	Aliases
Alias Removal			
Inactive Removal			
Port-Specific			

Experiments - Seed Dataset Cleaning: Aliasing

Dataset	Hits	ASes	Aliases
Alias Removal			
Inactive Removal			
Port-Specific			

Alias: 2001:4860:4680:0001::/64

2001:4860:4680:0001:0000:0000:0000:0000 → response
2001:4860:4680:0001:0000:0000:0000:0001 → response
2001:4860:4680:0001:0000:0000:0000:0002 → response
...
2001:4860:4680:0001:ffff:ffff:ffff:fffe → response
2001:4860:4680:0001:ffff:ffff:ffff:ffff → response

Experiments - Seed Dataset Cleaning: Aliasing

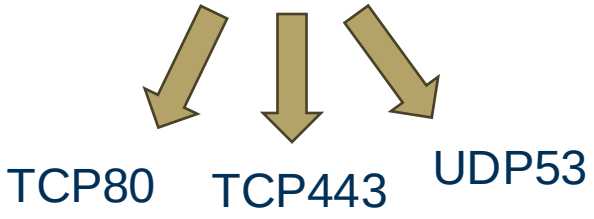
Dataset	Hits	ASes	Aliases
Alias Removal			
Inactive Removal			
Port-Specific			

Dataset	Total	Responsive	Responsive ASes
All	118.7M	10.9M	23.6K

Experiments - Seed Dataset Cleaning: Aliasing

Dataset	Hits	ASes	Aliases
Alias Removal			
Inactive Removal			
Port-Specific			

Dataset	Total	Responsive	Responsive ASes
All	118.7M	10.9M	23.6K



Georgia Tech

Experiments - Seed Dataset Cleaning: Aliasing

Dataset	Hits	ASes	Aliases
Alias Removal	↑	↑	↓
Inactive Removal			
Port-Specific			

Alias: 2001:4860:4680:0001::/64

2001:4860:4680:0001:0000:0000:0000:0000 → response
2001:4860:4680:0001:0000:0000:0000:0001 → response
2001:4860:4680:0001:0000:0000:0000:0002 → response
... ...
2001:4860:4680:0001:ffff:ffff:ffff:fffe → response
2001:4860:4680:0001:ffff:ffff:ffff:ffff → response

- **Removing Aliases from Seeds improves Hits and ASes**

Georgia Tech

Experiments - Seed Dataset Cleaning: Inactive

Dataset	Hits	ASes	Aliases
Alias Removal	↑	↑	↓
Inactive Removal	↑	↑	-
Port-Specific			

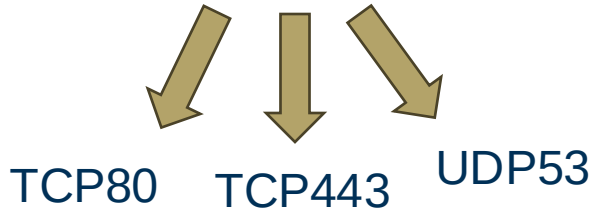
Dataset	Total	Responsive	Responsive ASes
All	118.7M	10.9M	23.6K

- Removing Aliases from Seeds improves Hits and ASes
- **Removing no longer responsive addresses increases Hits, and discovered ASes**

Experiments - Seed Dataset Cleaning: Port-Specific

Dataset	Hits	ASes	Aliases
Alias Removal	↑	↑	↓
Inactive Removal	↑	↑	-
Port-Specific	↑	↓	-

Dataset	Total	Responsive	Responsive ASes
All	118.7M	10.9M	23.6K



- Removing Aliases from Seeds improves Hits and ASes
- Removing no longer responsive addresses increases Hits, and discovered ASes
- **Removing unresponsive addresses on the port/protocol scanned increases Hits, but sacrifices ASes.**

Experiments - Seed Dataset Cleaning: Outcome

Dataset	Hits	ASes	Aliases
Alias Removal	↑	↑	↓
Inactive Removal	↑	↑	-
Port-Specific	↑	↓	-

- Removing Aliases from Seeds improves Hits and ASes
- Removing no longer responsive addresses increases Hits, and discovered ASes
- Removing unresponsive addresses on the port/protocol scanned increases Hits, but sacrifices ASes.

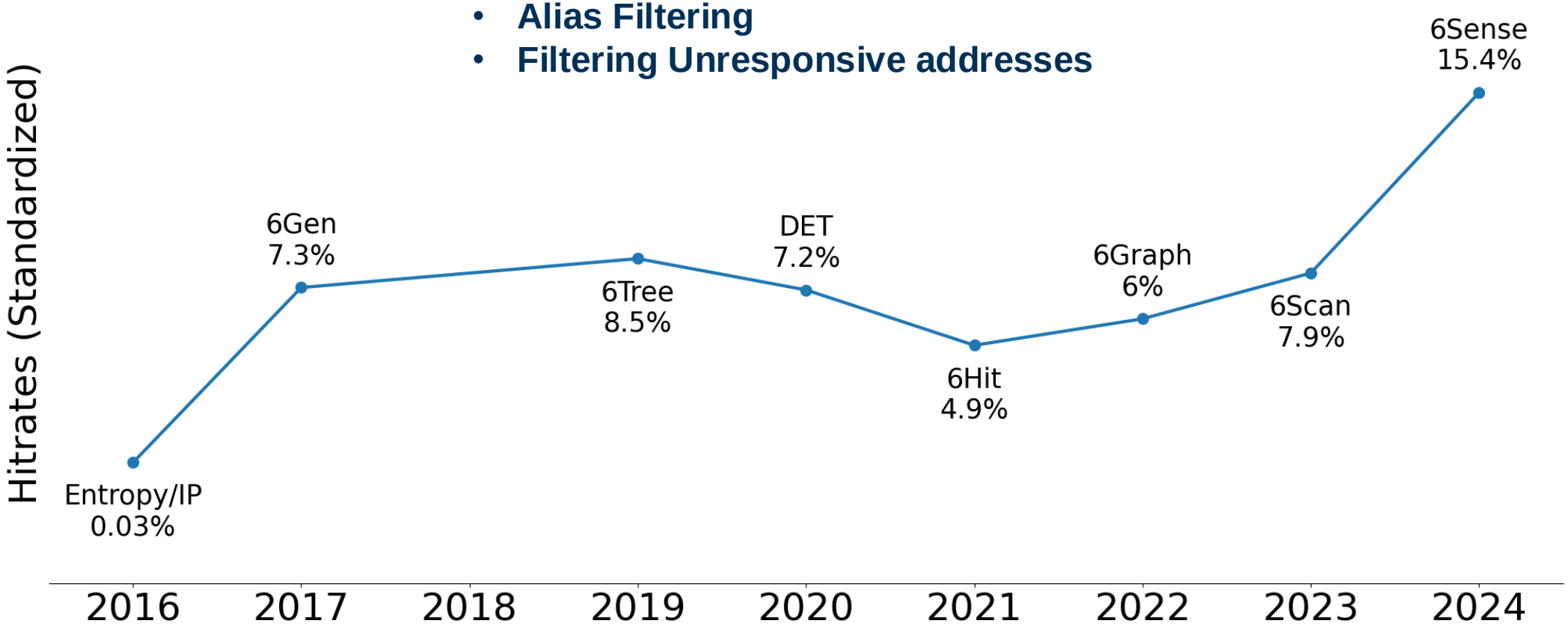
Recommendations:

- Filter Aliases from seeds.
- Filter unresponsive addresses.
- Optionally, filter to Port-Specific responsive (depending on use-case/metric).

Actual TGA Results

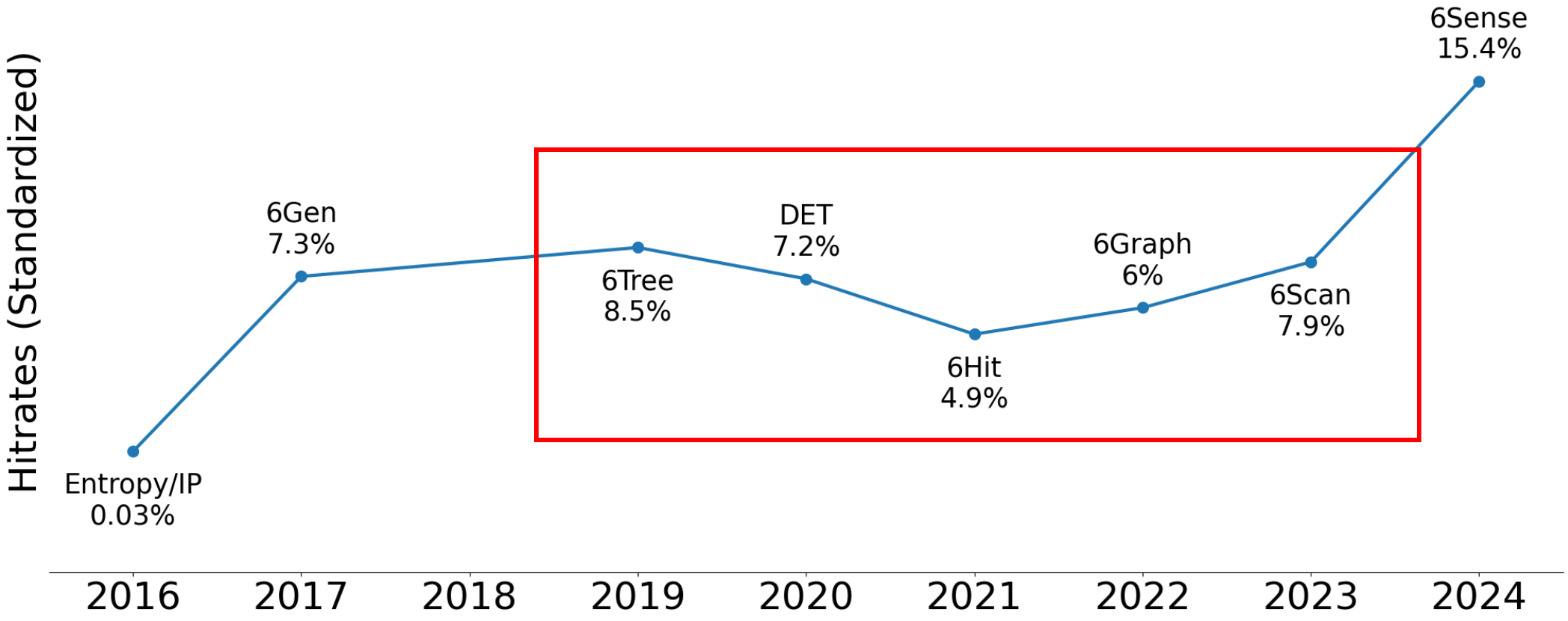
Using Dataset Preprocessing:

- **Alias Filtering**
- **Filtering Unresponsive addresses**

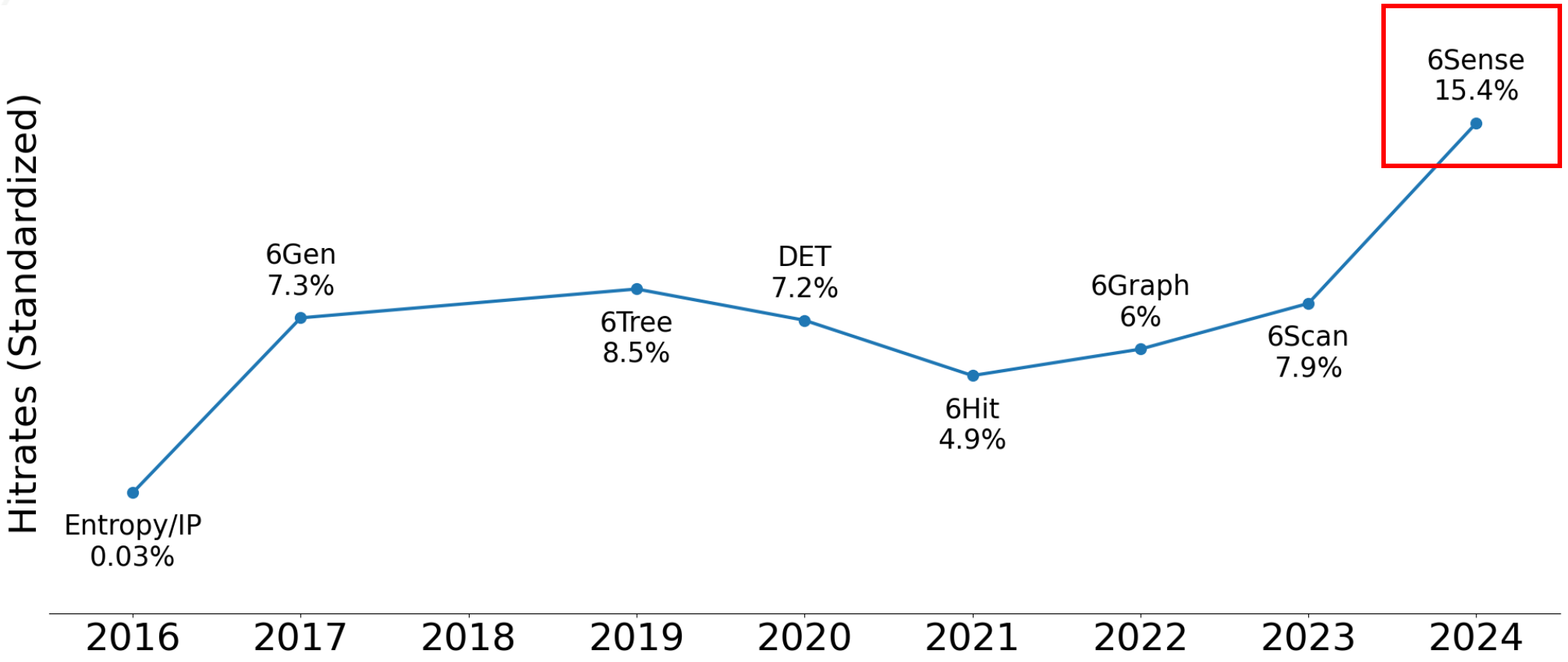


Georgia Tech

Actual TGA Results



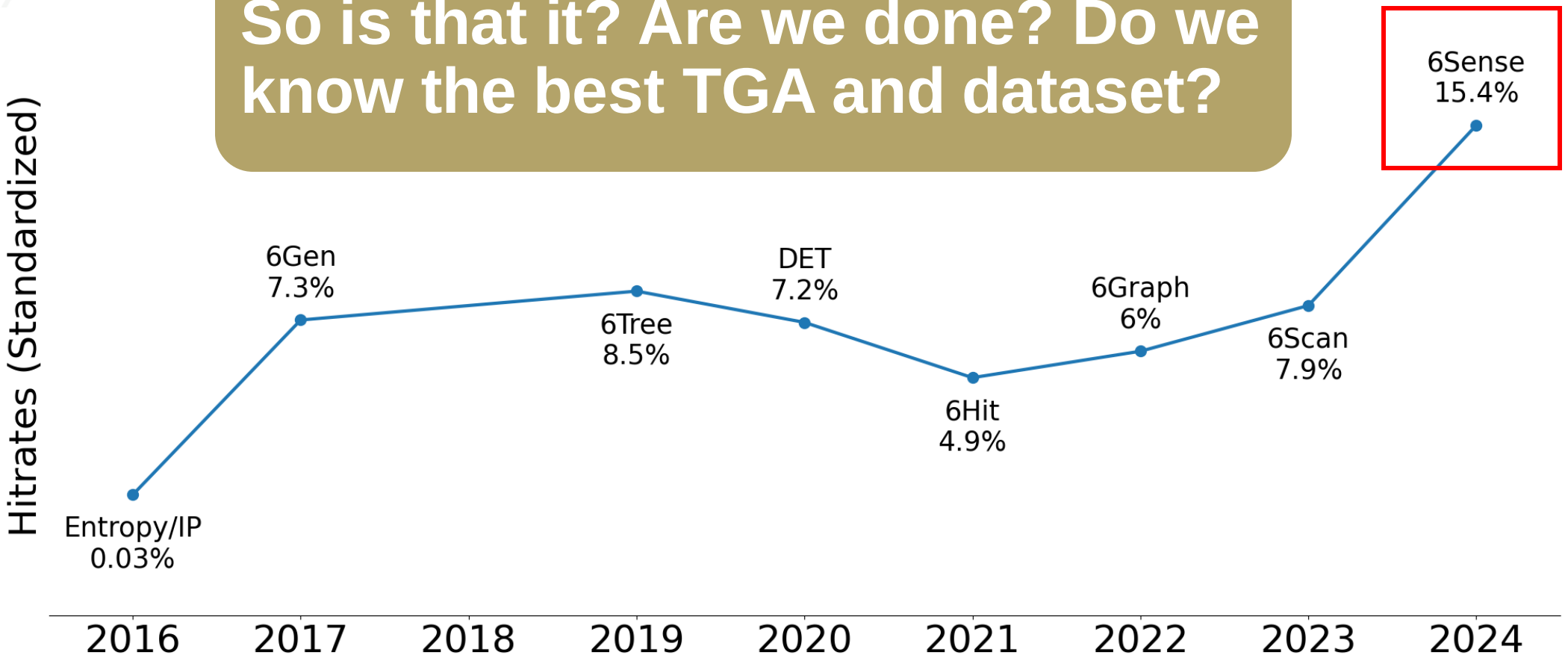
Actual TGA Results



Georgia Tech

Actual TGA Results

So is that it? Are we done? Do we know the best TGA and dataset?



Georgia Tech

Not so Fast!

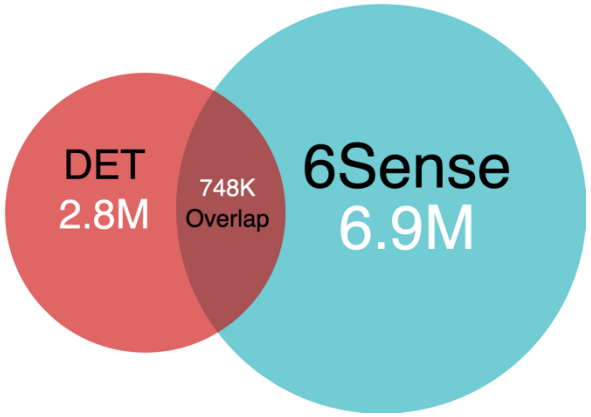
Not so Fast!

Not All TGAs find the **same** addresses

No Single TGA Finds Everything.

No Single TGA Finds Everything.

Hits

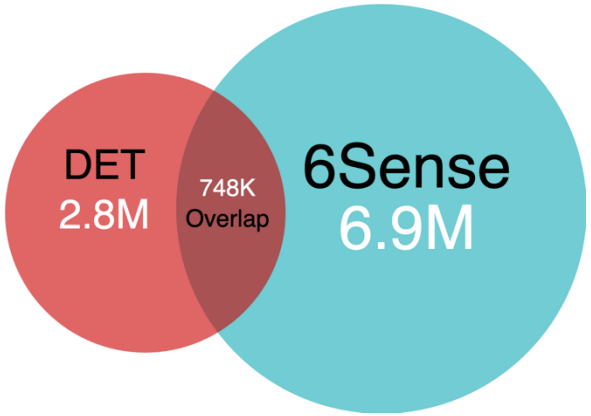


ASes

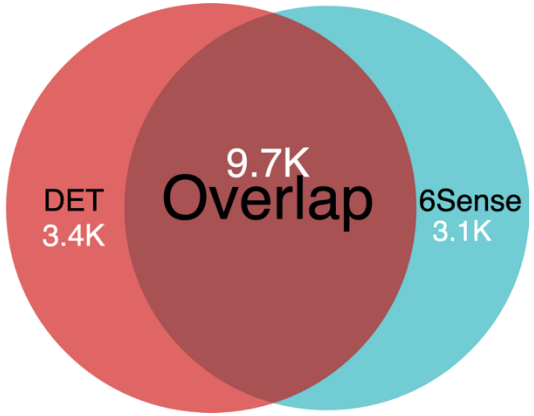
Georgia Tech

No Single TGA Finds Everything.

Hits



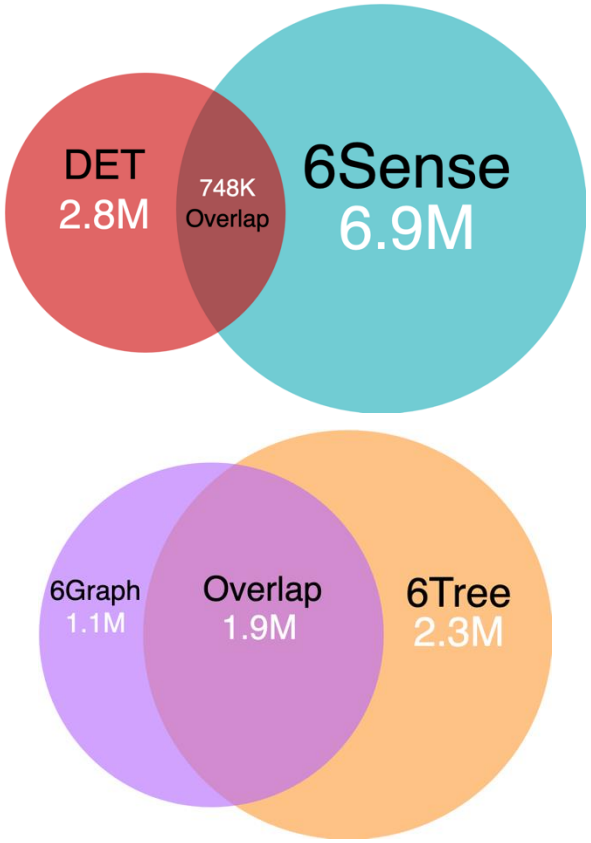
ASes



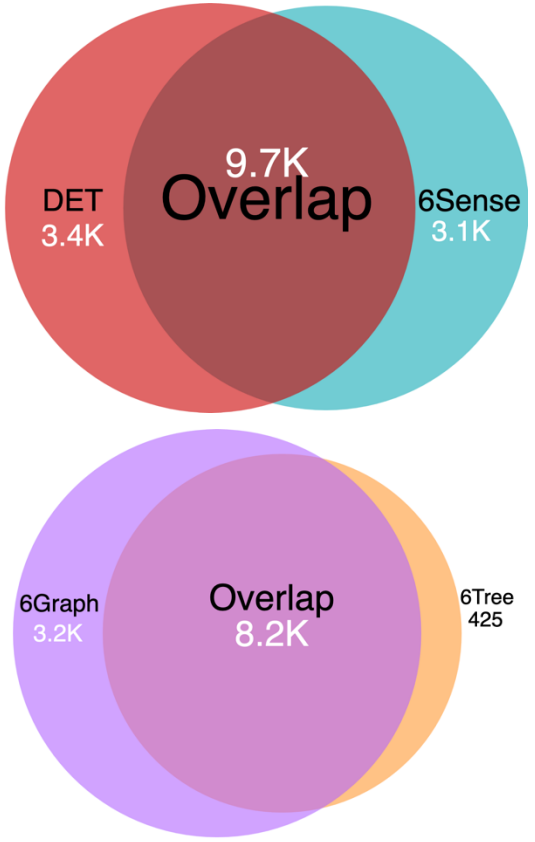
Georgia Tech

No Single TGA Finds Everything.

Hits



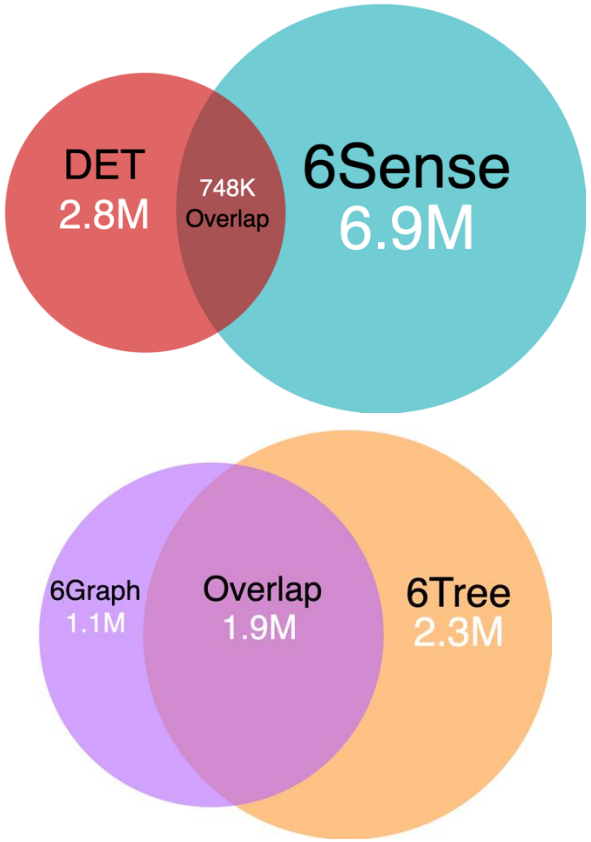
ASes



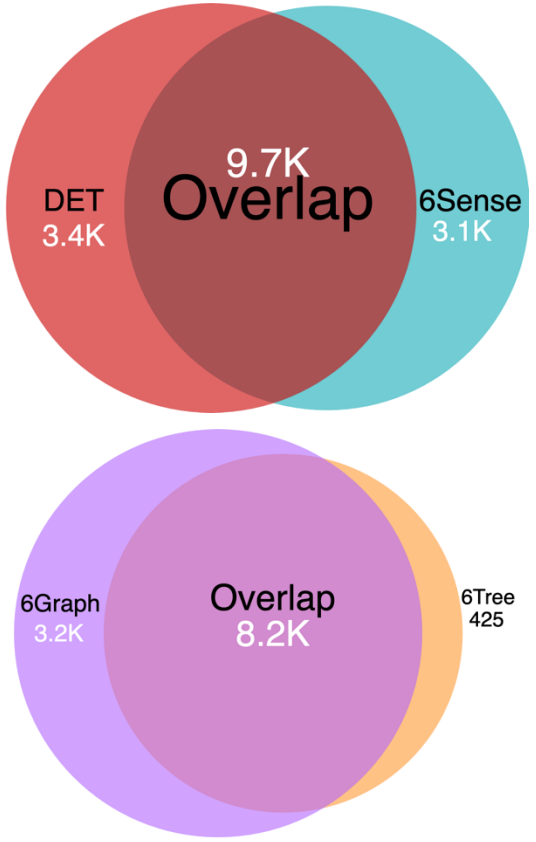
Georgia Tech

No Single TGA Finds Everything.

Hits



ASes



Suggests the optimal approach is using multiple.

Georgia Tech

Understanding What We Find

- Does the same multi-approach go for Seed datasets?

- We want to understand what we find when we scan using seeds from each data source independently. (600M with full dataset vs. 50M with each 12 seed sources).

Understanding What We Find

- Does the same multi-approach go for Seed datasets?

- We want to understand what we find when we scan using seeds from each data source independently. (600M with full dataset vs. 50M with each 12 seed sources).

Dataset	Hits	ASes	Aliases
Dataset Specific	↓	↑	-

Understanding What We Find

- Does the same multi-approach go for Seed datasets?

- We want to understand what we find when we scan using seeds from each data source independently. (600M with full dataset vs. 50M with each 12 seed sources).

Dataset	Hits	ASes	Aliases
Dataset Specific	↓	↑	-

- We notice each dataset focuses on different ASes

What Do We Find?

Top ASes for selection of our Seed Datasets on **ICMP** across TGAs

Dataset	Top AS	Second Common AS	Third Common AS
CT Logs	CLOUDFLARE	vodafone	sky

What Do We Find?

Top ASes for selection of our Seed Datasets on **ICMP** across TGAs

Dataset	Top AS	Second Common AS	Third Common AS
CT Logs	CLOUDFLARE	vodafone	sky
CAIDA DNS	PenTeleData	kpn	COMCAST

What Do We Find?

Top ASes for selection of our Seed Datasets on **ICMP** across TGAs

Dataset	Top AS	Second Common AS	Third Common AS
CT Logs	CLOUDFLARE	vodafone	sky
CAIDA DNS	PenTeleData	kpn	COMCAST
RIPE Atlas	CLOUDFLARE	Claro	MEGA

What Do We Find?

Top ASes for selection of our Seed Datasets on **ICMP** across TGAs

Dataset	Top AS	Second Common AS	Third Common AS
CT Logs	CLOUDFLARE	vodafone	sky
CAIDA DNS	PenTeleData	kpn	COMCAST
RIPE Atlas	CLOUDFLARE	Claro	MEGA
IPv6 Hitlist	CLOUDFLARE	SoftBank	Claro

What Do We Find?

Top ASes for selection of our Seed Datasets on **ICMP** across TGAs

Dataset	Top AS	Second Common AS	Third Common AS
CT Logs	CLOUDFLARE	vodafone	sky
CAIDA DNS	PenTeleData	kpn	COMCAST
RIPE Atlas	CLOUDFLARE	Claro	MEGA
IPv6 Hitlist	CLOUDFLARE	SoftBank	Claro

Takeaway: Smaller source-specific datasets can expand network diversity

Recommendations

- Best practices for running TGAs:

Recommendations

- Best practices for running TGAs:
    1. Dealias Seed addresses before using in TGAs.

Recommendations

- Best practices for running TGAs:
  1. Dealias Seed addresses before using in TGAs.
  2. Filter unresponsive IPs from seeds.

Recommendations

- Best practices for running TGAs:
  1. Dealias Seed addresses before using in TGAs.
  2. Filter unresponsive IPs from seeds.
  3. *Optional: Filter by port-protocol responsiveness.*

Recommendations

- Best practices for running TGAs:
  1. Dealias Seed addresses before using in TGAs.
  2. Filter unresponsive IPs from seeds.
  3. *Optional: Filter by port-protocol responsiveness.*
  4. Use multiple TGAs in combination to find different addresses.

Recommendations

- Best practices for running TGAs:
  1. Dealias Seed addresses before using in TGAs.
  2. Filter unresponsive IPs from seeds.
  3. *Optional: Filter by port-protocol responsiveness.*
  4. Use multiple TGAs in combination to find different addresses.
  5. Run TGAs on multiple smaller datasets to find more network diversity.

Recommendations

- Best practices for running TGAs:
  1. Dealias Seed addresses before using in TGAs.
  2. Filter unresponsive IPs from seeds.
  3. *Optional: Filter by port-protocol responsiveness.*
  4. Use multiple TGAs in combination to find different addresses.
  5. Run TGAs on multiple smaller datasets to find more network diversity.

- Seed addresses have a huge effect on the output of TGAs

Recommendations

- Best practices for running TGAs:
  1. Dealias Seed addresses before using in TGAs.
  2. Filter unresponsive IPs from seeds.
  3. *Optional: Filter by port-protocol responsiveness.*
  4. Use multiple TGAs in combination to find different addresses.
  5. Run TGAs on multiple smaller datasets to find more network diversity.

- Seed addresses have a huge effect on the output of TGAs
  - Preprocessing is important but context specific depending on metrics.

Recommendations

- Best practices for running TGAs:
  1. Dealias Seed addresses before using in TGAs.
  2. Filter unresponsive IPs from seeds.
  3. *Optional: Filter by port-protocol responsiveness.*
  4. Use multiple TGAs in combination to find different addresses.
  5. Run TGAs on multiple smaller datasets to find more network diversity.

- Seed addresses have a huge effect on the output of TGAs
  - Preprocessing is important but context specific depending on metrics.
  - It is important we use standardize TGA evaluation going forward.

Recommendations

Work funded by
NSF and GTRI IRAD

- Best practices for running TGAs:
  1. Dealias Seed addresses before using in TGAs.
  2. Filter unresponsive IPs from seeds.
  3. *Optional: Filter by port-protocol responsiveness.*
  4. Use multiple TGAs in combination to find different addresses.
  5. Run TGAs on multiple smaller datasets to find more network diversity.

- Seed addresses have a huge effect on the output of TGAs
  - Preprocessing is important but context specific depending on metrics.
  - It is important we use standardize TGA evaluation going forward.

- Questions.

Understanding What We Find

- Does the same multi-approach go for Seed datasets?