Sharing cyber threat intelligence: Does it really help?

Beomjin Jin*, Eunsoo Kim*, Hyunwoo Lee†, Elisa Bertino‡, Doowon Kim§ and Hyoungshick Kim*

***Sungkyunkwan University, †KENTECH, ‡Purdue University, §University of Tennessee**

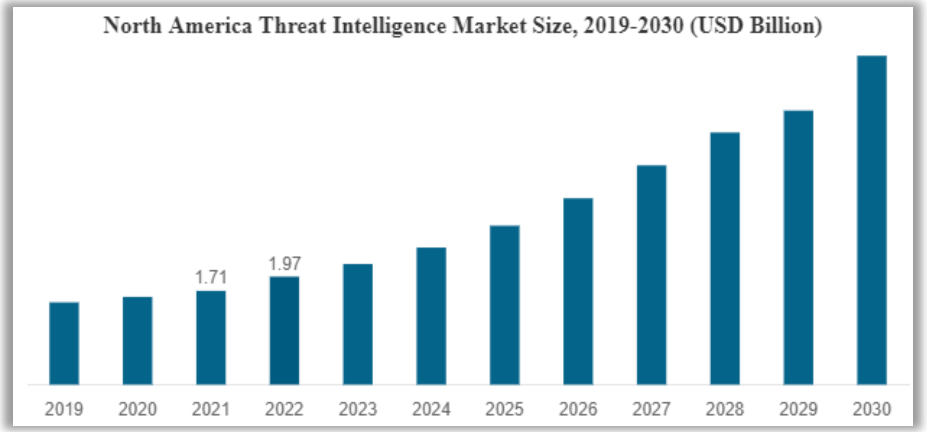
Cyber Threat Information sharing

The New York Times

Americans Have Lost $145 Million to Coronavirus Fraud

More than 200,000 complaints of scams and fraud have been filed so far this year, data from the Federal Trade Commission shows.

- *COVID* frauds caused a lot of financial damage[1]
- Threat information sharing resulted in proactive prevention of *COVID* frauds[2]

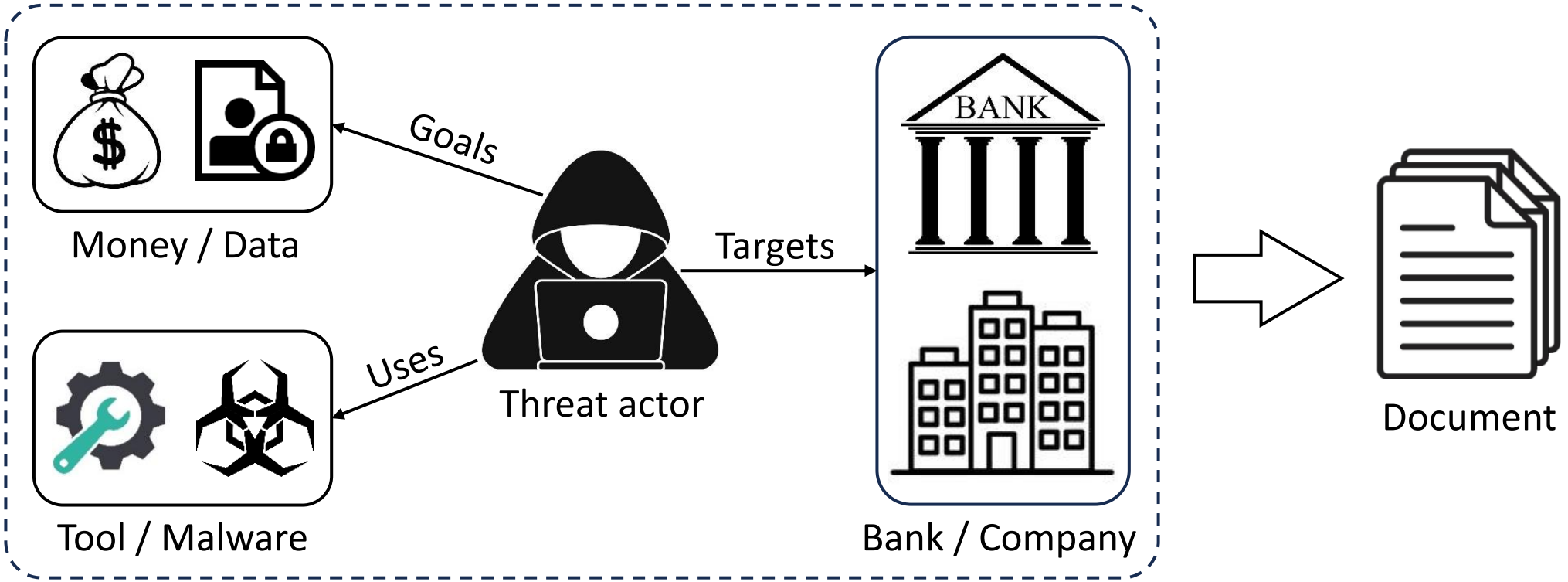


- Surge of *COVID* frauds led to the threat sharing market growth in North America[3]

[1] https://www.nytimes.com/2020/09/23/us/coronavirus-scams-ftc-reports.html
[2] Bouwman *et al.*, "Helping hands: Measuring the impact of a large threat intelligence sharing community," *USENIX Security* 2022
[3] https://www.fortunebusinessinsights.com/threat-intelligence-market-102984

Cyber Threat Intelligence

- Cyber threat intelligence (CTI) is the **knowledge** to understand and mitigate cyberattacks



CTI formats

- CTI data is shared in various formats

- *Campaign*
- *Threat actor*
- *Malware*
- *Vulnerability*

Threat report

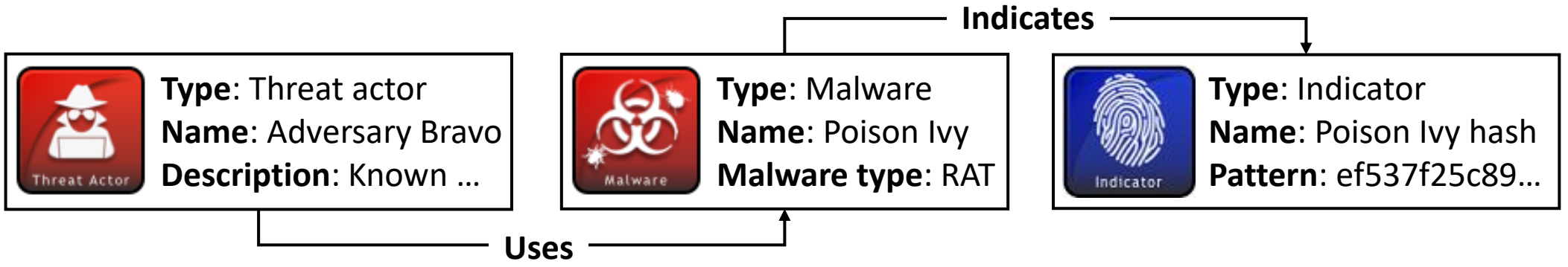
- *Malware file hash*
- *Malicious IP address*
- *Malicious domain/URL*

Blacklist

CTI standards

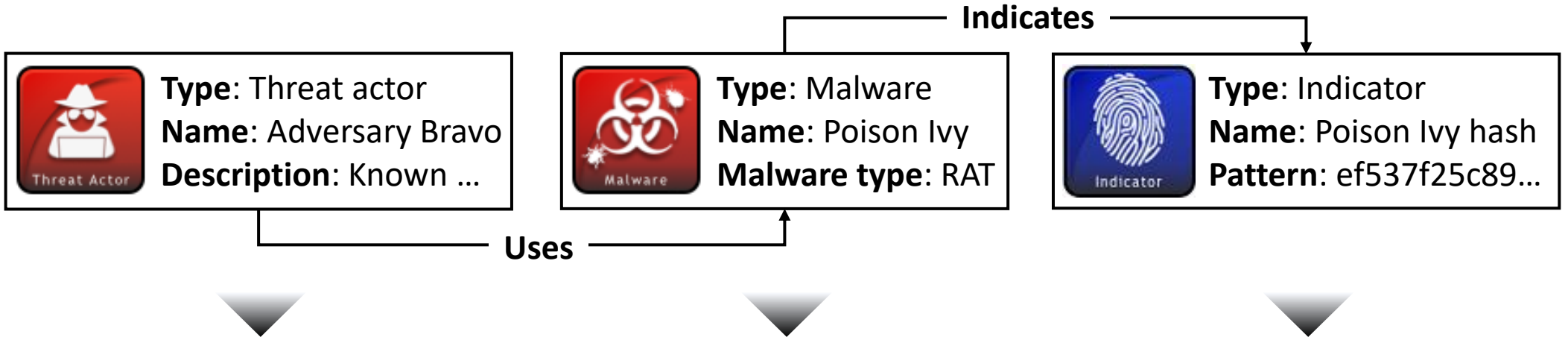
STIX (Structured Threat Information eXpression)

- STIX is a *de facto* standard and is widely used in cybersecurity



STIX (Structured Threat Information eXpression)

- STIX is a *de facto* standard and is widely used in cybersecurity




```
{
  "type": "threat-actor",
  "id": "threat-actor--d9…",
  "name": "Adversary Bravo",
  "description": "Known …",
  …
}
```

```
{
  "type": "malware",
  "id": "malware--541c…",
  "name": "Poison Ivy",
  "malware_types": ["RAT"],
  …
}
```

```
{
  "type": "indicator",
  "id": "indicator--9422f…",
  "name": "Poison Ivy hash",
  "pattern": "ef537f25c89…",
  …
}
```

Is sharing STIX really helpful?

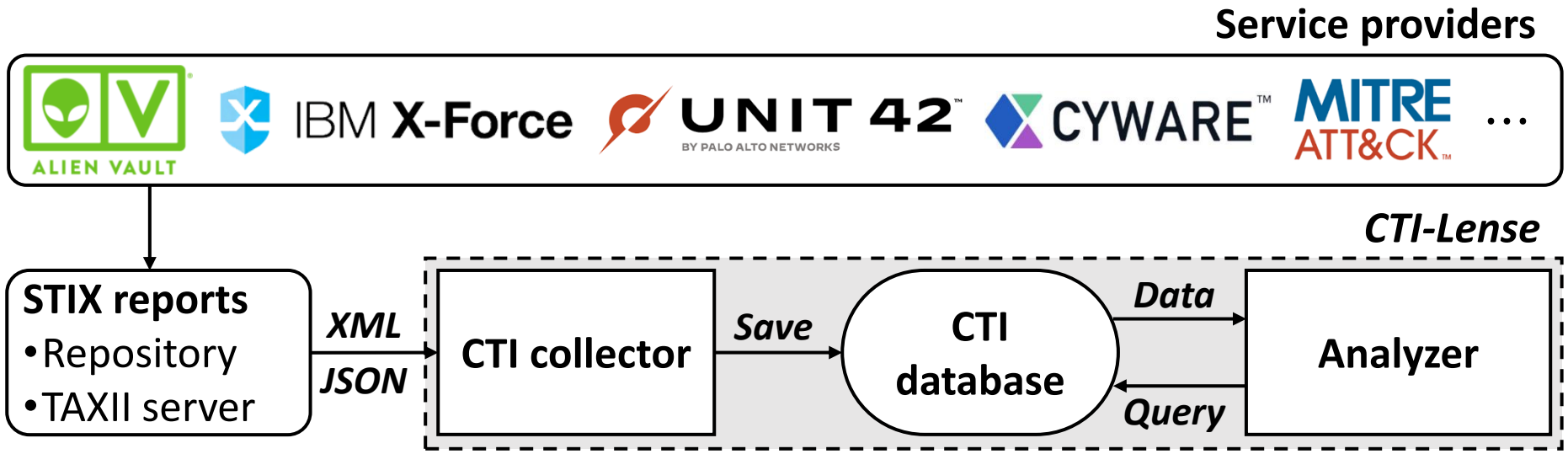
- Volume
- Coverage
- Timeliness
- Quality

RQ1 (Volume): How much STIX data is being generated and shared publicly?

Data duplication ***Daily/monthly number of STIX data***

CTI-Lense

- A framework that aggregates and assesses the STIX data
  - We collected **10M** STIX objects from publicly available service providers:
    - TAXII servers: *Hail a TAXII*, *AlienVault OTX*, *IBM X-Force Exchange*, *etc.*
    - Web repositories: *JamesBrine*, *DigitalSide*, *MITRE ATT&CK*
    - Period: From October 31, 2014 to April 10, 2023
  - We analyzed **data duplication** and **daily/monthly shared STIX data**



Volume

Version	Sources	Unique	Duplication	
			Intra	Inter
STIX 1	Hail a TAXII	1,900,237	50.71%	0.76%
	AlienVault OTX	1,647,509	38.93%	1.16%
	IBM X-Force Exchange	273,274	46.36%	20.32%
	PickupSTIX	73,575	18.13%	6.86%
STIX 2	AlienVault OTX	1,657,442	3.02%	2.00%
	JamesBrine	205,776	68.74%	0.00%
	DigitalSide	198,439	33.52%	7.78%
	Cyware	228,782	11.94%	2.51%
	IBM X-Force Exchange	119,611	1.25%	2.54%
	Unit42	33,379	7.03%	13.29%
	MITRE ATT&CK	17,042	0.05%	1.73%
	Limo from Anomali	6,492	0.06%	12.68%
	PickupSTIX	507	1.74%	0.00%

Data duplication

- **38%** of STIX data are duplicated **within** a single source

Volume

Version	Sources	Unique	Duplication	
			Intra	Inter
STIX 1	Hail a TAXII	1,900,237	50.71%	0.76%
	AlienVault OTX	1,647,509	38.93%	1.16%
	IBM X-Force Exchange	273,274	46.36%	20.32%
	PickupSTIX	73,575	18.13%	6.86%
STIX 2	AlienVault OTX	1,657,442	3.02%	2.00%
	JamesBrine	205,776	68.74%	0.00%
	DigitalSide	198,439	33.52%	7.78%
	Cyware	228,782	11.94%	2.51%
	IBM X-Force Exchange	119,611	1.25%	2.54%
	Unit42	33,379	7.03%	13.29%
	MITRE ATT&CK	17,042	0.05%	1.73%
	Limo from Anomali	6,492	0.06%	12.68%
	PickupSTIX	507	1.74%	0.00%

Data duplication

- **38%** of STIX data are duplicated **within** a single source

Requires additional efforts to remove duplicated data before deployment

Volume

Version	Sources	Unique	Duplication	
			Intra	Inter
STIX 1	Hail a TAXII	1,900,237	50.71%	0.76%
	AlienVault OTX	1,647,509	38.93%	1.16%
	IBM X-Force Exchange	273,274	46.36%	20.32%
	PickupSTIX	73,575	18.13%	6.86%
STIX 2	AlienVault OTX	1,657,442	3.02%	2.00%
	JamesBrine	205,776	68.74%	0.00%
	DigitalSide	198,439	33.52%	7.78%
	Cyware	228,782	11.94%	2.51%
	IBM X-Force Exchange	119,611	1.25%	2.54%
	Unit42	33,379	7.03%	13.29%
	MITRE ATT&CK	17,042	0.05%	1.73%
	Limo from Anomali	6,492	0.06%	12.68%
	PickupSTIX	507	1.74%	0.00%

Number of STIX data

- Open STIX dataset predominantly depends on **a few sources**
- A daily average of **2,063** unique STIX objects are publicly shared

Volume

Version	Sources	Unique	Duplication	
			Intra	Inter
STIX 1	Hail a TAXII	1,900,237	50.71%	0.76%
	AlienVault OTX	1,647,509	38.93%	1.16%
	IBM X-Force Exchange	273,274	46.36%	20.32%
	PickupSTIX	73,575	18.13%	6.86%
STIX 2	AlienVault OTX	1,657,442	3.02%	2.00%
	JamesBrine	205,776	68.74%	0.00%
	DigitalSide	198,439	33.52%	7.78%
	Cyware	228,782	11.94%	2.51%
	IBM X-Force Exchange	119,611	1.25%	2.54%
	Unit42	33,379	7.03%	13.29%
	MITRE ATT&CK	17,042	0.05%	1.73%
	Limo from Anomali	6,492	0.06%	12.68%
	PickupSTIX	507	1.74%	0.00%

Number of STIX data

- Open STIX dataset predominantly depends on **a few sources**
- A daily average of **2,063** unique STIX objects are publicly shared

This number seems significantly insufficient to handle the daily emerging malware samples

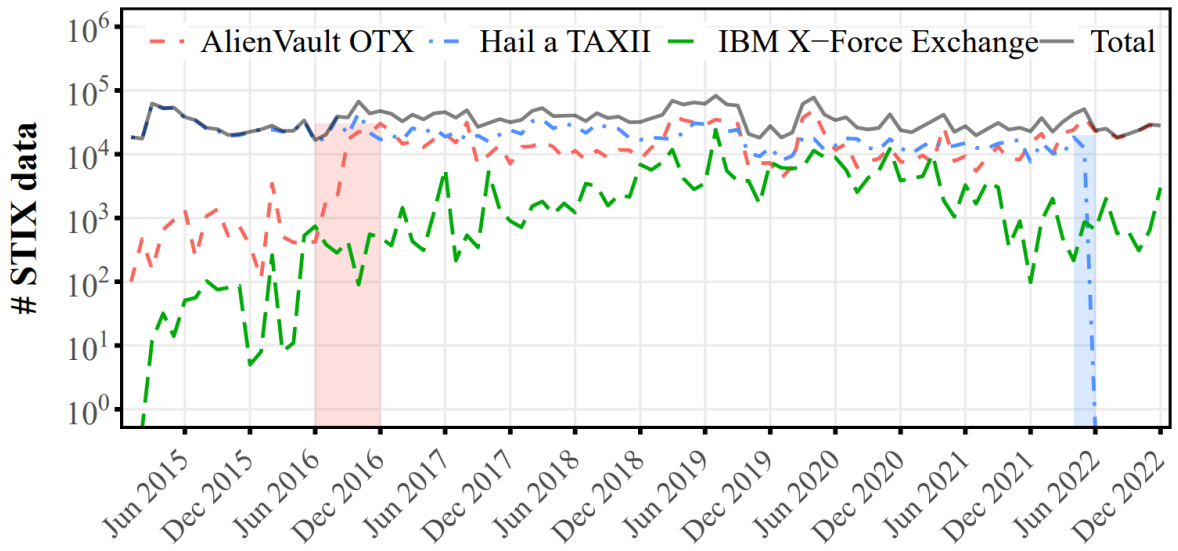
AV-TEST *450,000 malware per day*[5]

[5] https://www.av-test.org/en/statistics/malware/

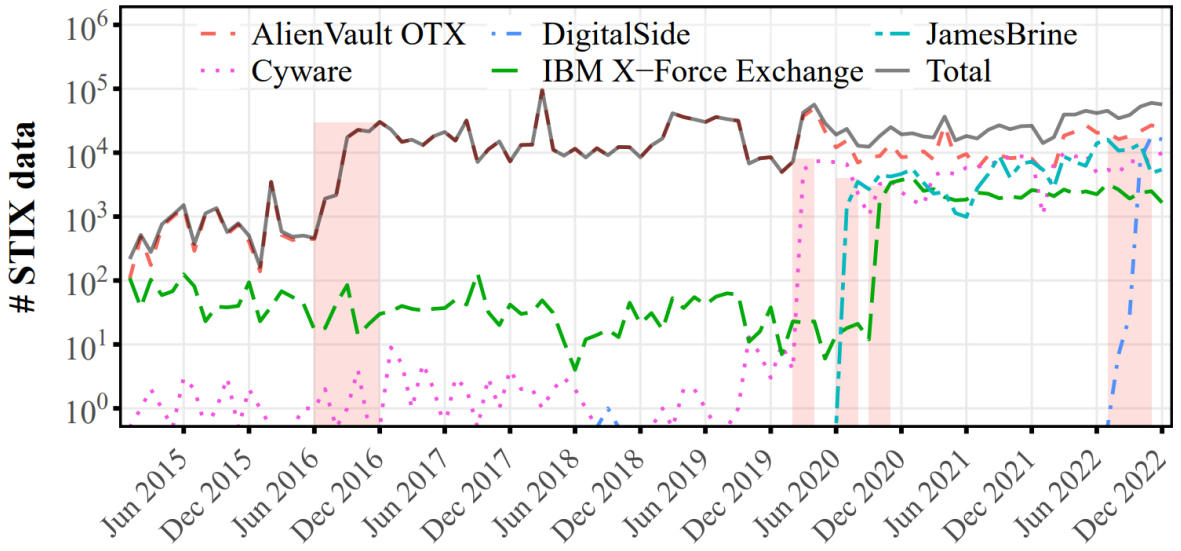
Volume

STIX data sharing over time

- STIX data sharing has increased in recent years
  - The trend of sharing STIX data has been rising since 2016
  - Various STIX 2 data sources have been emerged since 2020



STIX 1



STIX 2

RQ2 (Timeliness): How promptly is STIX data shared following a cyber threat discovery?

vs. security incidents ***vs. online scanning services***

Timeliness

vs. security incidents

- We analyzed relationship between the STIX data and corresponding **security incidents**
  - We found weak evidence of causal relationship from **security news websites** to the sharing of STIX data, with **p < 0.05** for the **2–12** days *time-lag*

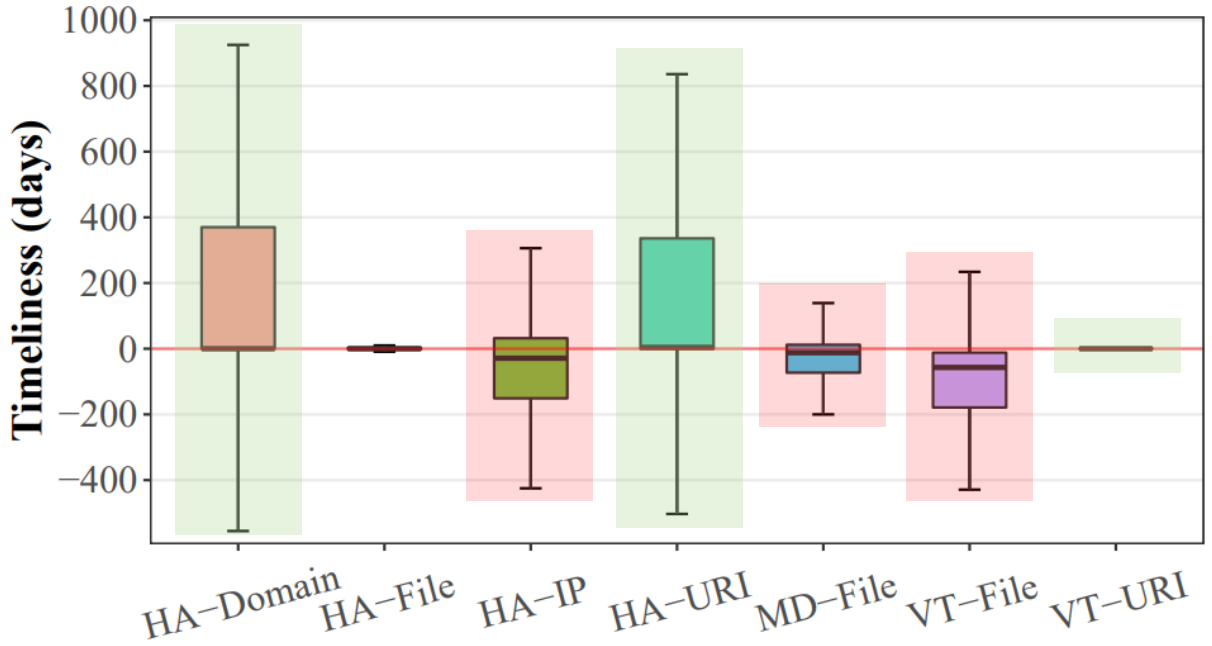
Security incident

STIX data are shared after 2–12 days following the reporting of security incidents in security news websites

Timeliness

vs. online scanning services

- We measured the latency between the initial appearance of STIX data and its subsequent detection by popular **scanning services**, such as
  - ViursTotal, HybridAnalysis, and MetaDefender



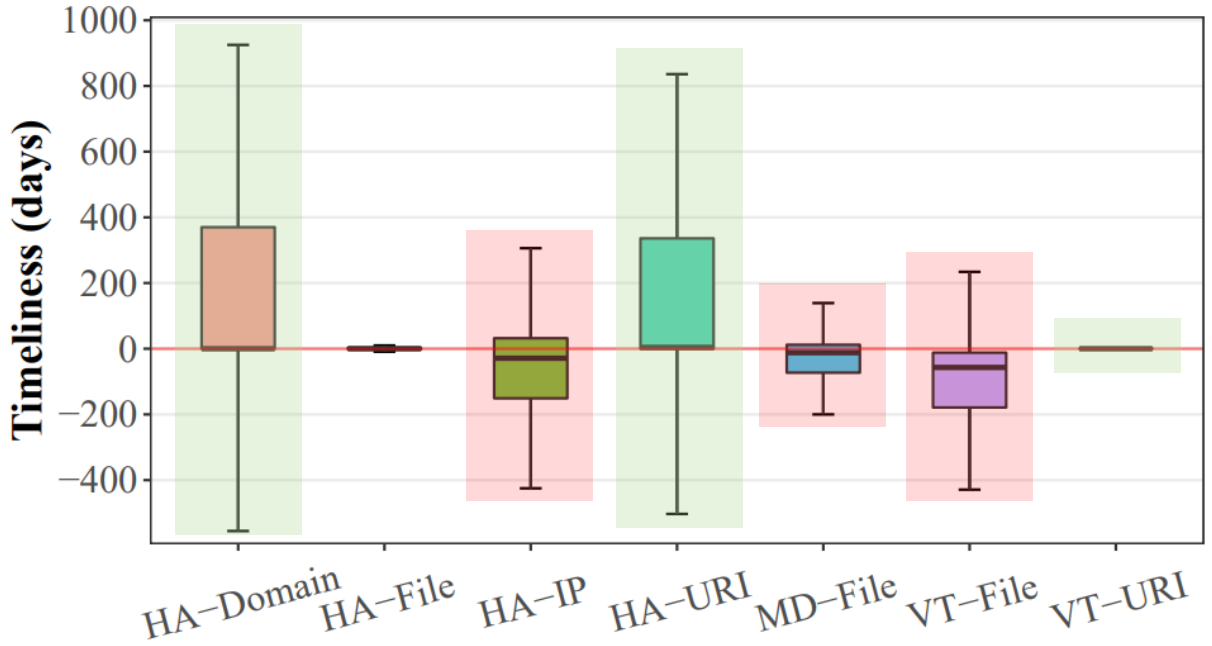
Differences in the data generation time

- **Domain** and **URI** data shared **faster** than scanning services
- **File** and **IP** data shared **slower** than scanning services

Timeliness

vs. online scanning services

- We measured the latency between the initial appearance of STIX data and its subsequent detection by popular **scanning services**, such as
  - ViursTotal, HybridAnalysis, and MetaDefender



Differences in the data generation time

- **Domain** and **URI** data shared **faster** than scanning services
- **File** and **IP** data shared **slower** than scanning services

STIX is ineffective in detecting the near zero-day malware

RQ3 (Coverage): How many objects and attributes defined in STIX are used?

STIX object usage ***STIX attribute usage***

Coverage

Versions	Objects	Prop.
STIX 1	Indicator	98.77%
	Other	1.23%
STIX 2	Indicator	94.93%
	Other	5.07%

Versions	Indicator attributes	Prop.
STIX 1	Observable	99.92%
	Type	53.73%
	Producer	51.28%
	Indicated_ttps	34.76%
	Test_mechanisms	0.09%
STIX 2	Pattern	100.00%
	Labels	32.62%
	Indicator_types	17.13%
	Kill_chain_phases	17.13%
	Created_by_ref	10.20%

Object usage

- ***Indicator*** object type accounts for more than **90%** for both STIX versions

Attribute usage in Indicator

- Nearly 100% of *Indicator* objects contain simple **IoC data**, such as malware hash
- Few *Indicator* objects contain **types, producers, and detection rules**

Coverage

Versions	Objects	Prop.
STIX 1	Indicator	98.77%
	Other	1.23%
STIX 2	Indicator	94.93%
	Other	5.07%

Versions	Indicator attributes	Prop.
STIX 1	Observable	99.92%
	Type	53.73%
	Producer	51.28%
	Indicated_ttps	34.76%
	Test_mechanisms	0.09%
STIX 2	Pattern	100.00%
	Labels	32.62%
	Indicator_types	17.13%
	Kill_chain_phases	17.13%
	Created_by_ref	10.20%

Object usage

- ***Indicator*** object type accounts for more than **90%** for both STIX versions

Attribute usage in Indicator

- Nearly 100% of *Indicator* objects contain simple **IoC data**, such as malware hash
- Few *Indicator* objects contain **types, producers, and detection rules**

A limited number of STIX data types are used in practice

RQ4 (Quality): What is the quality of STIX data in terms of its correctness and completeness in representing cyber threats?

Improper value ***Improper usage***

Quality

Improper values – Incorrect values in STIX objects

- Focused on values in *Indicator*, *TTP*, *Malware*, and *Threat actor* objects

```
1  <stix:Indicator id="indicator-...">
2    <indicator:Observable idref="...">
3      <cyboxCommon:simple_hash_value>
4        12cc14bbbc421275c3c6145bfa186dff
5      </cyboxCommon:simple_hash_value>
6    </indicator:Observable>
7    <indicator:Indicated_TTP>
8      <stix:TTP id="ttp-...">
9        <ttp:Behavior>
10         <ttp:Malware_Instance>
11           <ttp:Name> Lazarus </ttp:Name>
12         </ttp:Malware_Instance>
13       <ttp:Behavior>
14       ...
15     </stix:TTP>
16   </indicator:Indicated_TTP>
17 </stix:Indicator>
```

Indicator → **Must include malicious file hash!**

TTP – Malware instance → **Must include malware instance name!**

Quality

Improper values – Incorrect values in STIX objects

- Focused on values in *Indicator*, *TTP*, *Malware*, and *Threat actor* objects

```
1  <stix:Indicator id="indicator-...">
2    <indicator:Observable idref="...">
3      <cyboxCommon:simple_hash_value>
4        12cc14bbbc421275c3c6145bfa186dff
5      </cyboxCommon:simple_hash_value>
6    </indicator:Observable>
7    <indicator:Indicated_TTP>
8      <stix:TTP id="ttp-...">
9        <ttp:Behavior>
10         <ttp:Malware_Instance>
11           <ttp:Name> Lazarus </ttp:Name>
12         </ttp:Malware_Instance>
13       <ttp:Behavior>
14       ...
15     </stix:TTP>
16   </indicator:Indicated_TTP>
17 </stix:Indicator>
```

Indicator → **Must include malicious file hash!**

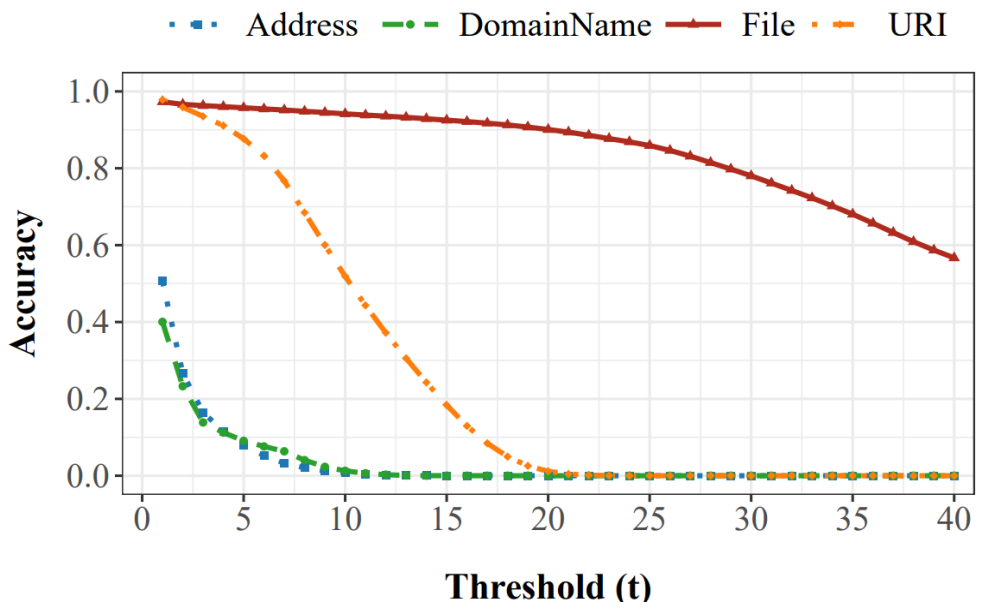
Measure detection rate using online scanning services

TTP – Malware instance → **Must include malware instance name!**

Match the keywords from trustworthy data sources

Quality

Types	Count	Detected	Not det.	N/A
Address	43,537	50.62%	49.38%	0.00%
DomainName	163,121	39.99%	59.90%	0.12%
File	88,470	78.37%	1.87%	19.75%
URI	377,857	97.06%	2.18%	0.76%
Total	672,985	77.77%	19.18%	3.05%

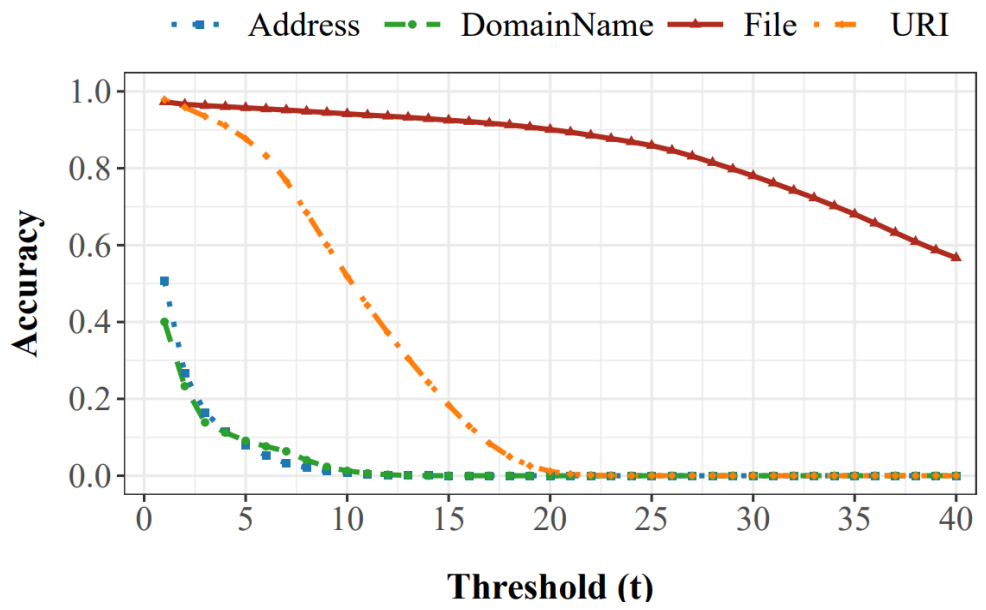


Improper values in Indicator

- **78%** IoC data are confirmed by at least one of the engines in VirusTotal
- Among the data confirmed by the VirusTotal:
  - ***File*** and ***URI*** types achieved over **90% detection rate** ($1 \leq t < 5$)
  - ***Address*** and ***DomainName*** types achieved relatively **low detection rate** at about **50%** and **40%**, respectively

Quality

Types	Count	Detected	Not det.	N/A
Address	43,537	50.62%	49.38%	0.00%
DomainName	163,121	39.99%	59.90%	0.12%
File	88,470	78.37%	1.87%	19.75%
URI	377,857	97.06%	2.18%	0.76%
Total	672,985	77.77%	19.18%	3.05%



Improper values in Indicator

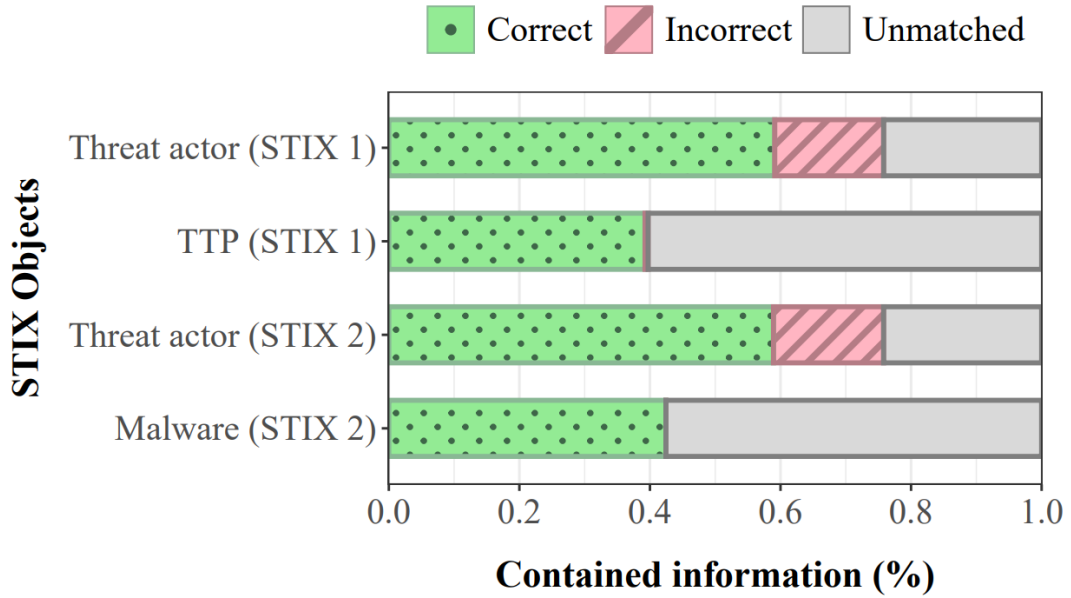
- **78%** IoC data are confirmed by at least one of the engines in VirusTotal
- Among the data confirmed by the VirusTotal:
  - ***File*** and ***URI*** types achieved over **90% detection rate** ($1 \leq t < 5$)
  - ***Address*** and ***DomainName*** types achieved relatively **low detection rate** at about **50%** and **40%**, respectively

Most file and URI data included in STIX are correct

Quality

Improper values in TTP, Threat actor, and Malware

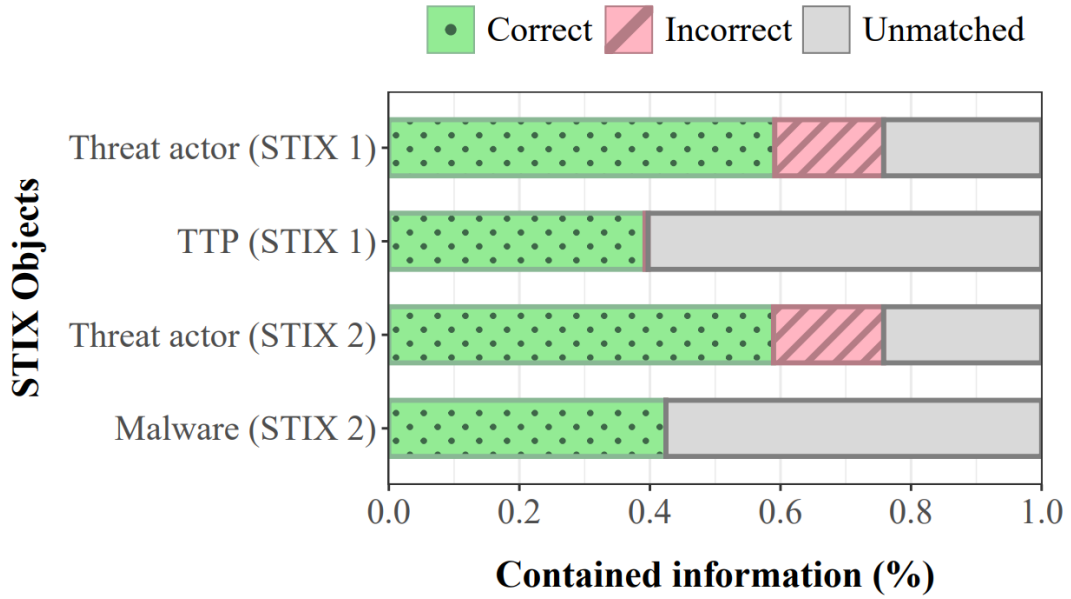
- **19%** of *Threat actor* objects contain malware family names
- **60%** of malware family information does not match any of the entries
  - The majority of the values are named based on producers' own conventions



Quality

Improper values in TTP, Threat actor, and Malware

- **19%** of *Threat actor* objects contain malware family names
- **60%** of malware family information does not match any of the entries
  - The majority of the values are named based on producers' own conventions



STIX does not always contain accurate values and most values are named based on producers' own conventions

Requires additional data validation process

Quality

Improper usage – Imprecise STIX object usage

- Focused on narrative description rather than using specific STIX objects

```
1  <stix:STIX_Package ...>
2    <stix:STIX_Header>
3      <stix:Title> Dtrack and ATMDtrack ATM
       Malware Linked to Lazarus </stix:Title>
4      <stix:Description> Summary After discovering
        ATM malware they named ATMDtrack, Kaspersky
        found a significant number of other related
        samples of malware which they have named
       Dtrack ... A remote access Trojan (RAT) is
       also installed ... </stix:Description>
5      ...
6    </stix:STIX_Header>
7    <stix:Indicators>
8      <stix:Indicator id="...">...</stix:Indicator
       >
9      ...
10   </stix:Indicators>
11 </stix:STIX_Package>
```

Can be described using STIX objects!

- TTP – Malware instance
  - Dtrack, ATMDtrack, RAT
- Threat actor
  - Lazarus

Quality

Improper usage – Imprecise STIX object usage

- Focused on narrative description rather than using specific STIX objects

```
1  <stix:STIX_Package ...>
2    <stix:STIX_Header>
3      <stix:Title> Dtrack and ATMDtrack ATM
       Malware Linked to Lazarus </stix:Title>
4      <stix:Description> Summary After discovering
        ATM malware they named ATMDtrack, Kaspersky
        found a significant number of other related
        samples of malware which they have named
       Dtrack ... A remote access Trojan (RAT) is
       also installed ... </stix:Description>
5      ...
6    </stix:STIX_Header>
7    <stix:Indicators>
8      <stix:Indicator id="...">...</stix:Indicator
       >
9      ...
10   </stix:Indicators>
11 </stix:STIX_Package>
```

Can be described using STIX objects!

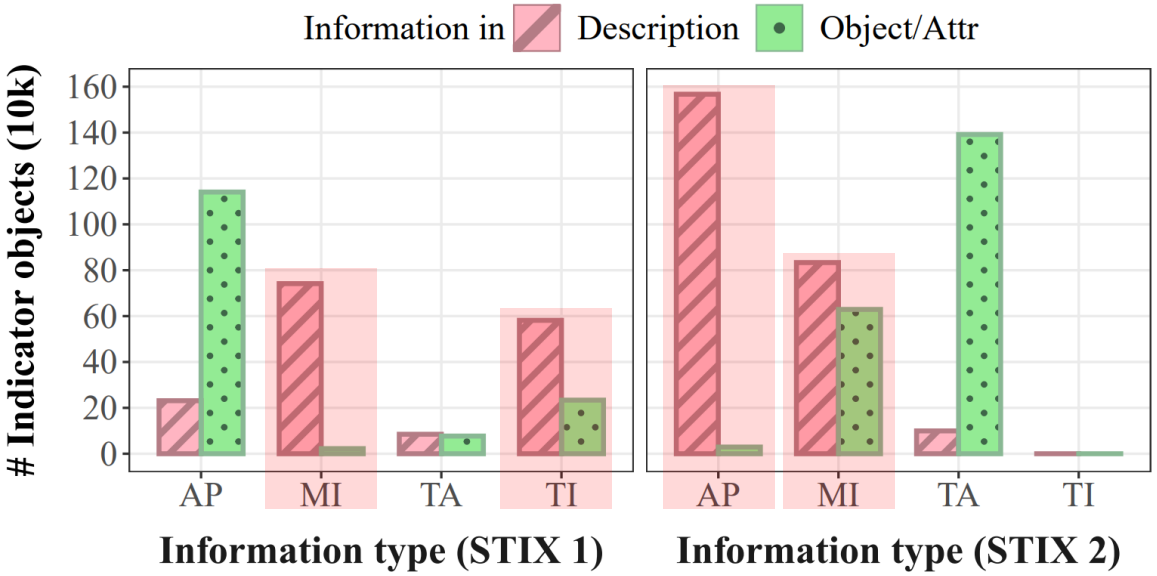
- TTP – Malware instance
  - Dtrack, ATMDtrack, RAT
- Threat actor
  - Lazarus

Count the number of *Indicator* objects where information is written narrative form

Quality

Improper usage

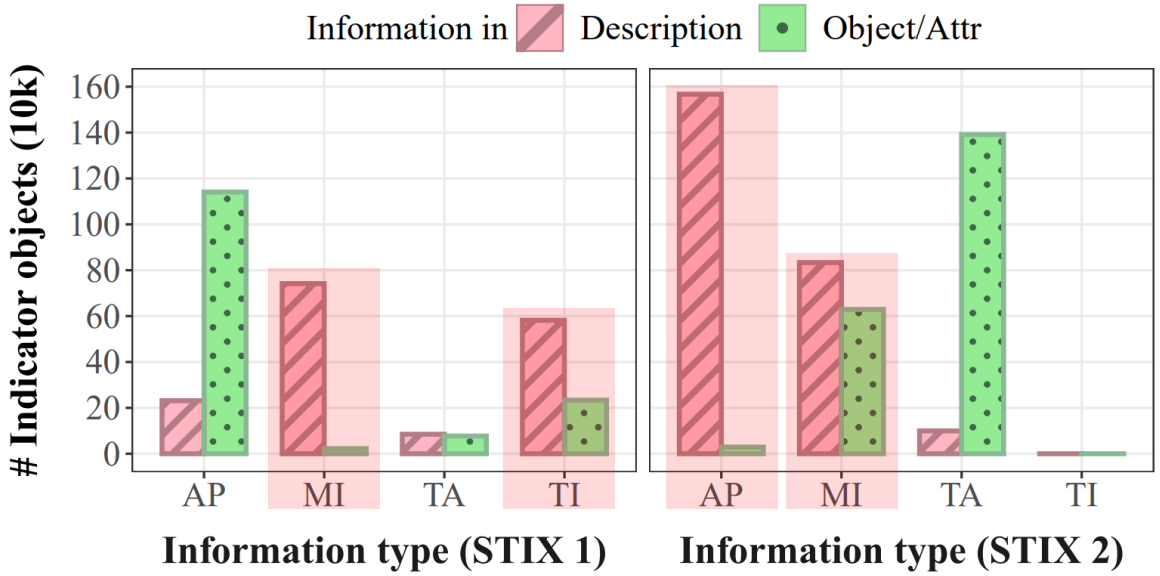
- Significant number of *indicator* objects describe related information in narrative form
  - **98%** of the *Indicator* objects in STIX 2 includes attack patterns only in a narrative form



Quality

Improper usage

- Significant number of *indicator* objects describe related information in narrative form
  - **98%** of the *Indicator* objects in STIX 2 includes attack patterns only in a narrative form



The producers often describe threat information in a narrative form, rather than using specific STIX objects

This can make automatic processing difficult

Our recommendations

- We need to know how to create better STIX data!
  - Provide **educational programs** and **practical guidelines** for effectively using STIX data types
- We need to generate and manage STIX data automatically!
  - Develop ML-based **tools for verification** and **deduplication** to minimize human error
- We need to maintain consistency in STIX terms and usage!
  - **Standardize naming conventions** to avoid confusion and inconsistency

Thank you, any questions?

https://github.com/SKKU-SecLab/CTI_Lense

jinbumjin@skku.edu