32ND USENIX SECURITY SYMPOSIUM

ALPS Research Lab

Algorithmic Research on Learning, Privacy and Security

On the Security Risks of Knowledge Graph Reasoning

Zhaohan Xi, *Penn State*
Tianyu Du, *Penn State*
Changjiang Li, *Penn State*
Ren Pang, *Penn State*
Shouling Ji, *Zhejiang University*
Xiapu Luo, *Hong Kong Polytechnic University*
Xusheng Xiao, *Arizona State University*
Fenglong Ma, *Penn State*
Ting Wang, *Penn State*

PennState

ZHEJIANG UNIVERSITY 1897

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

ASU

Knowledge Graph

KG is a collection of …

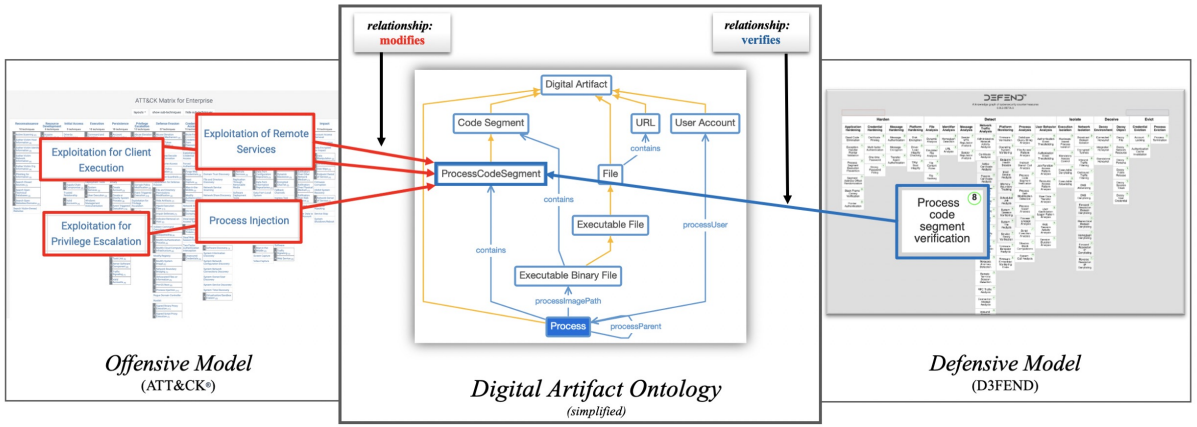
- Node: real-world objects
- Edge: relational facts

- E.g., Wikidata, DBPedia, WorldNet, etc.

KG in practice

KG in security

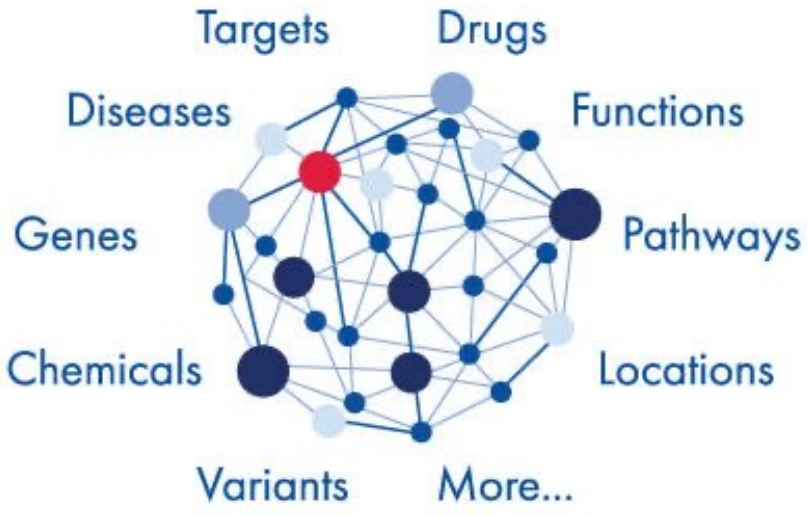
- Cyber-threat intelligence



By Peter & Michael @ *The MITRE Corporation*

KG in biomedical science

- Clinical decision & support



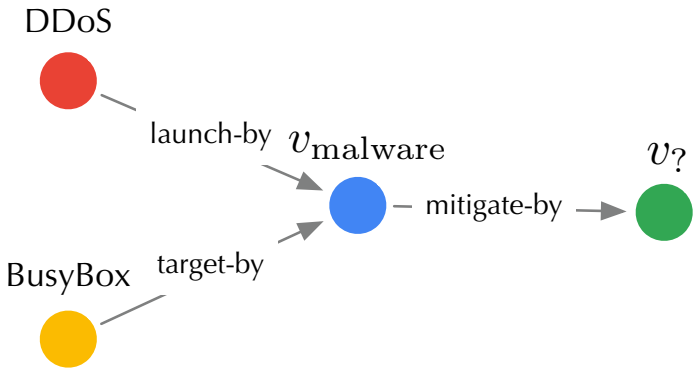
By QIAGEN Co.

Reasoning over KG

- Query

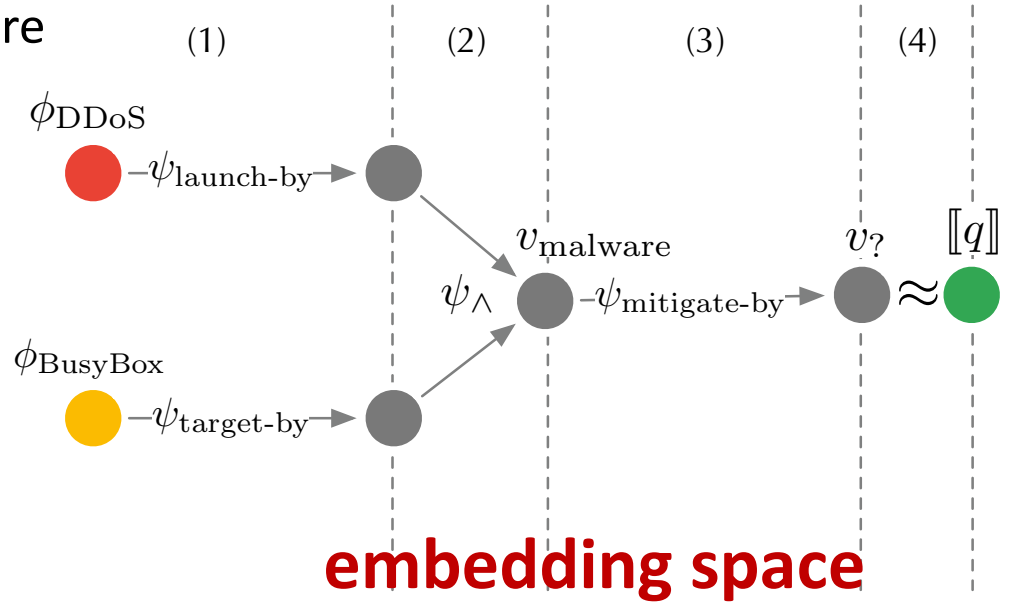
> *"How to mitigate the malware that targets **BusyBox** and launches **DDoS** attacks?"*

→

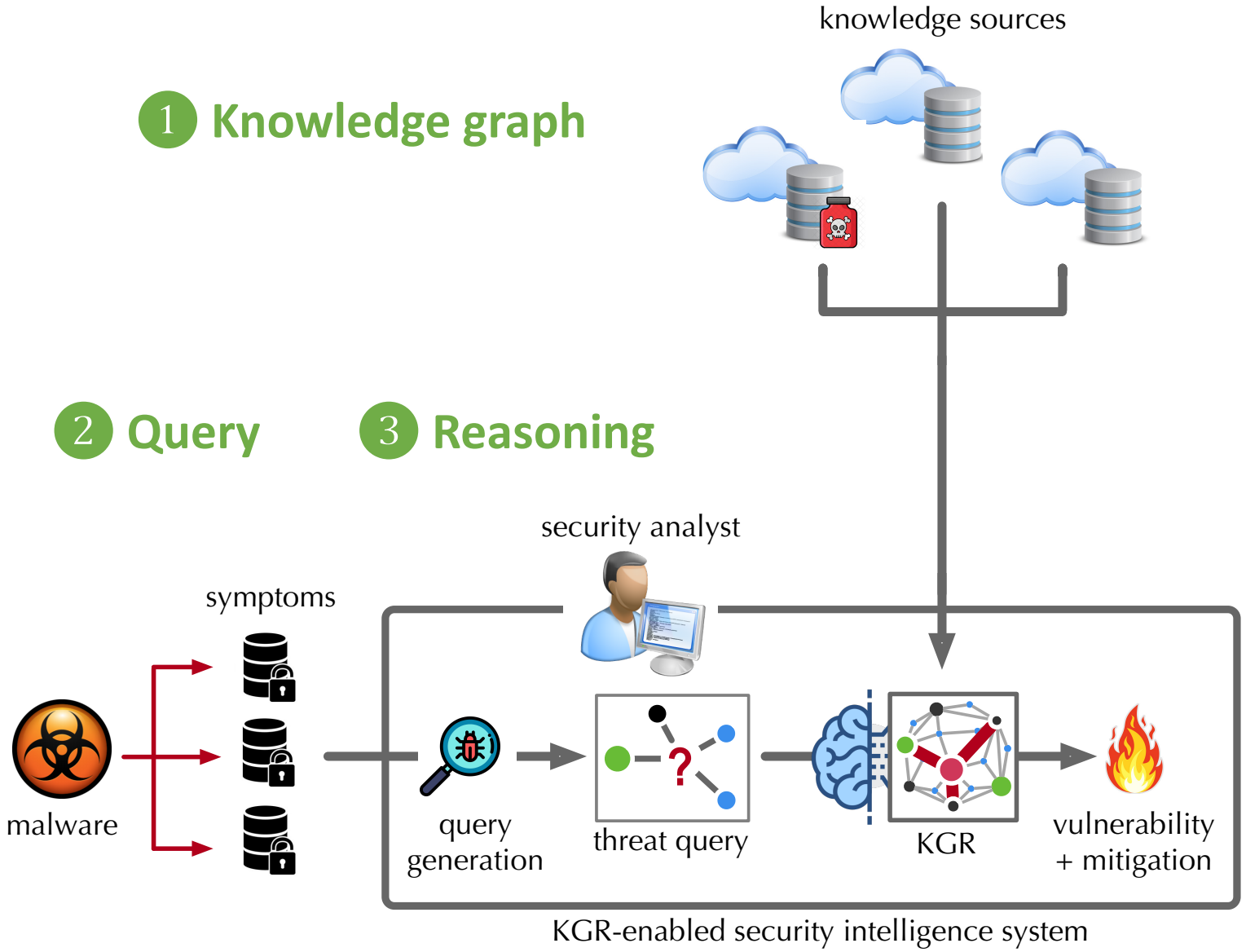


- Representation
  - Train entity embeddings using KG structure

- Reasoning
  - Happen in embedding space
  - Reduce complex query to embedding
  - Match answers by embedding similarity

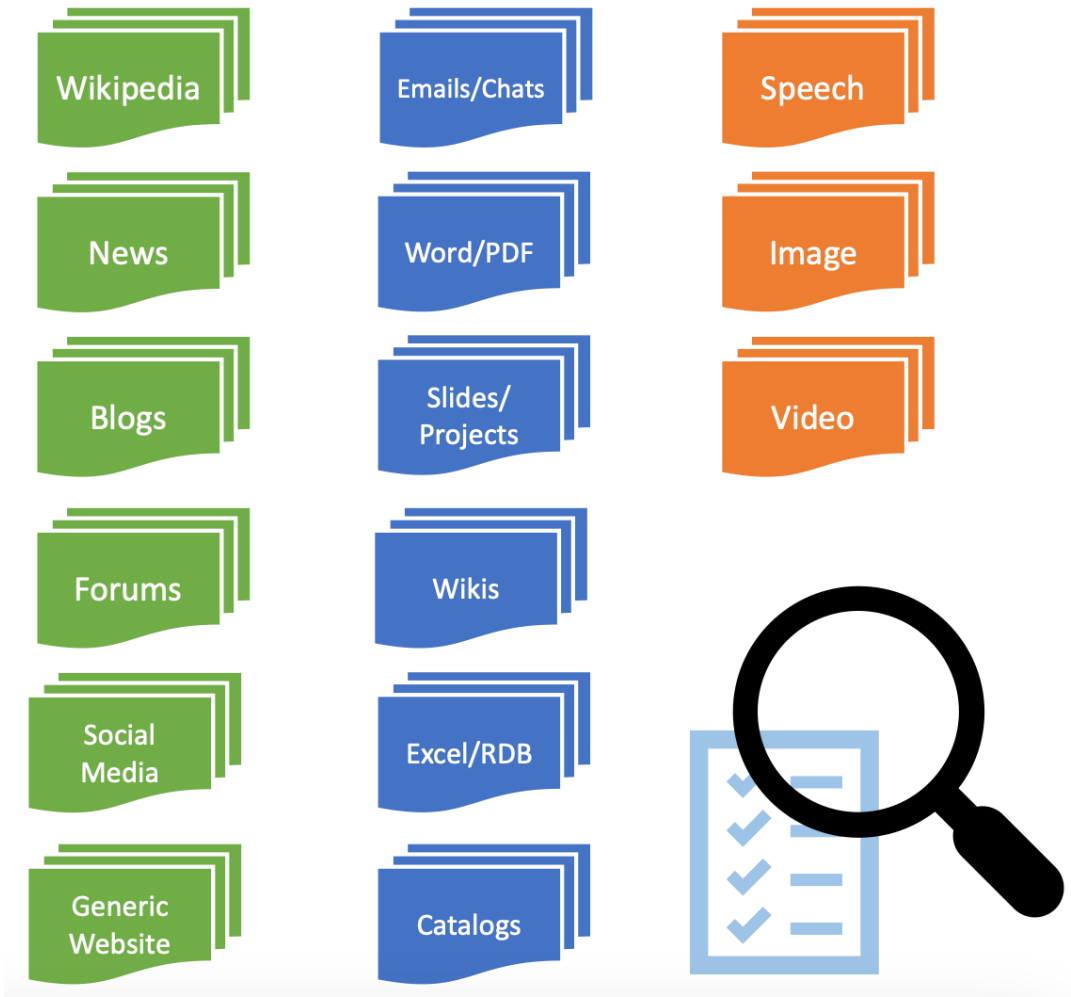


Reasoning Pipeline



Poisoning vulnerability

- Knowledge can come from many sources



By Microsoft @ *KDD 2018 Tutorial*

Poisoning vulnerability

- Knowledge can come from many sources
- **Poor curation of crowd-sourcing knowledge may lead to harmful impacts**

Google's Knowledge Graph Is Rife with Misinformation and an Easy Tool for Online Radicalization

August 31, 2020

SHARE

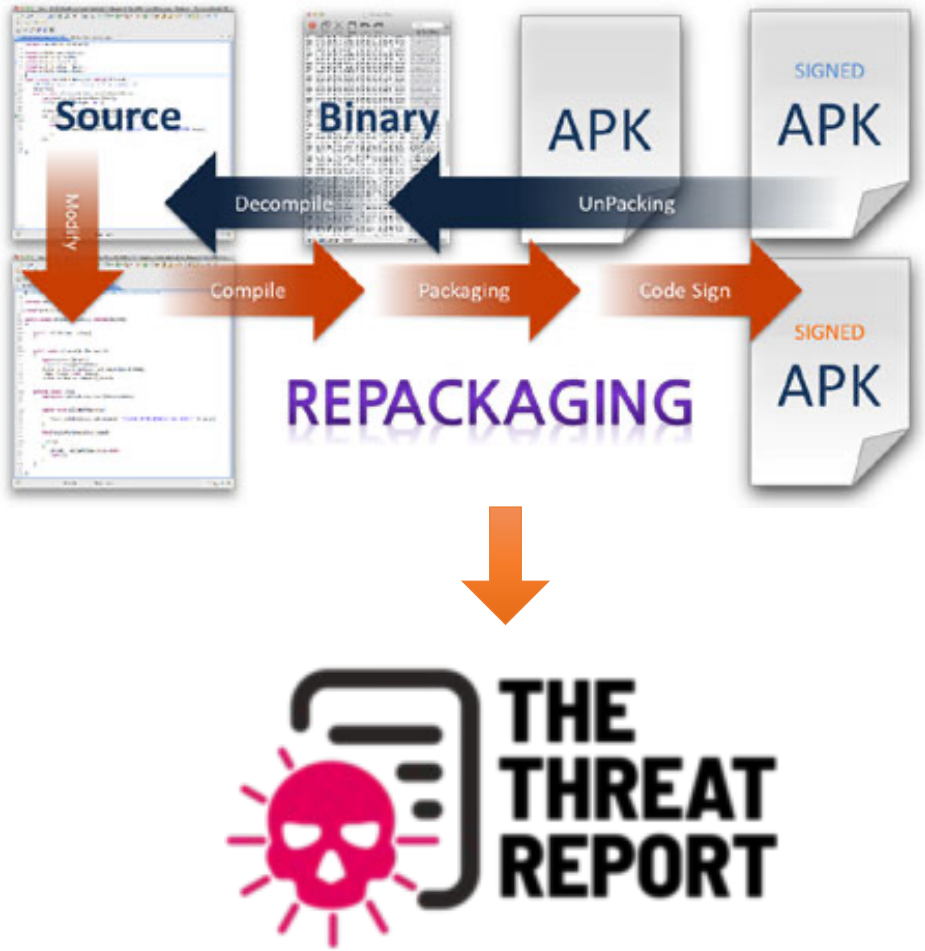
GPAHE GLOBAL PROJECT AGAINST HATE AND EXTREMISM

By GPAHE Project

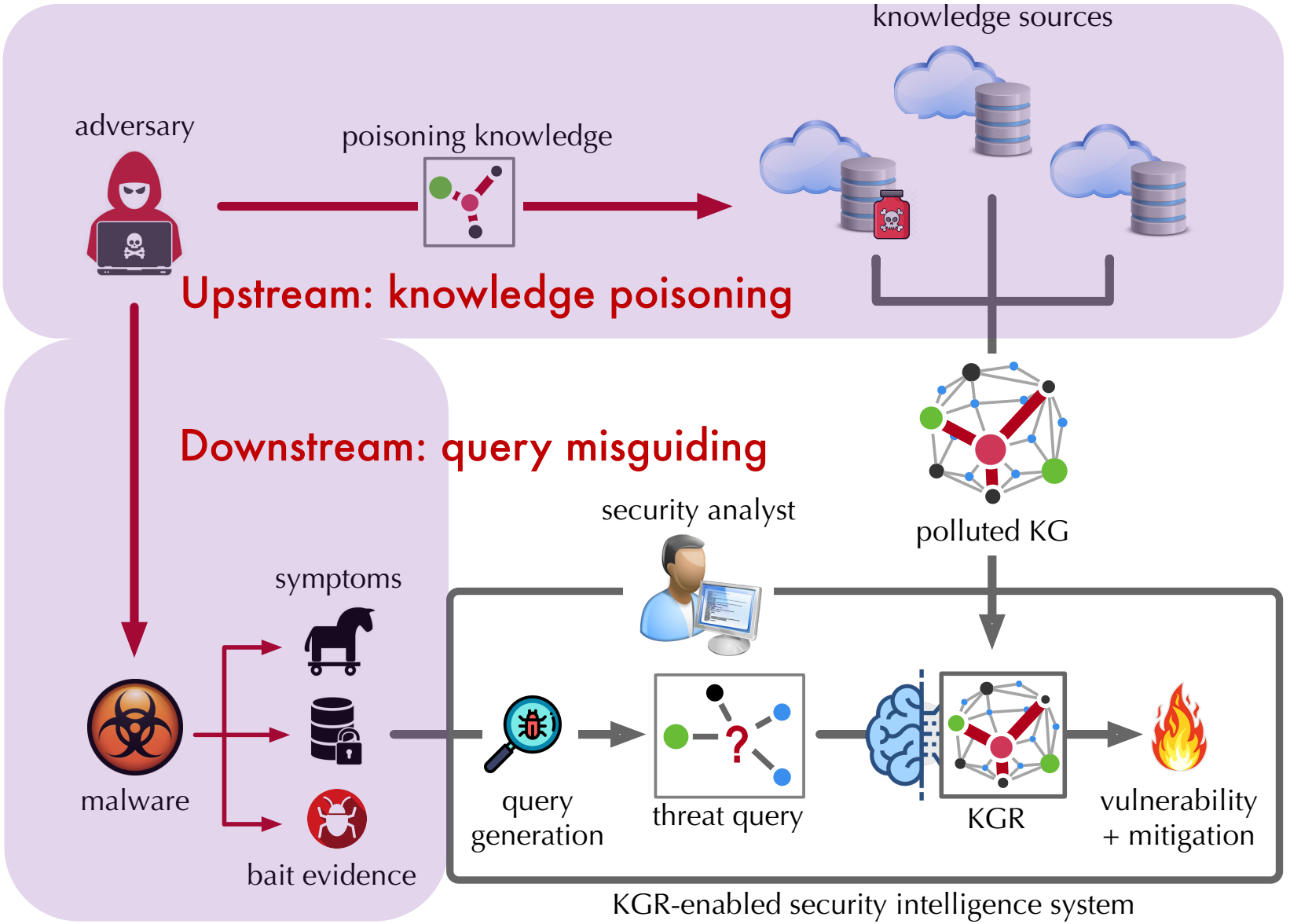
By Microsoft @ *KDD 2018 Tutorial*

Misguiding vulnerability

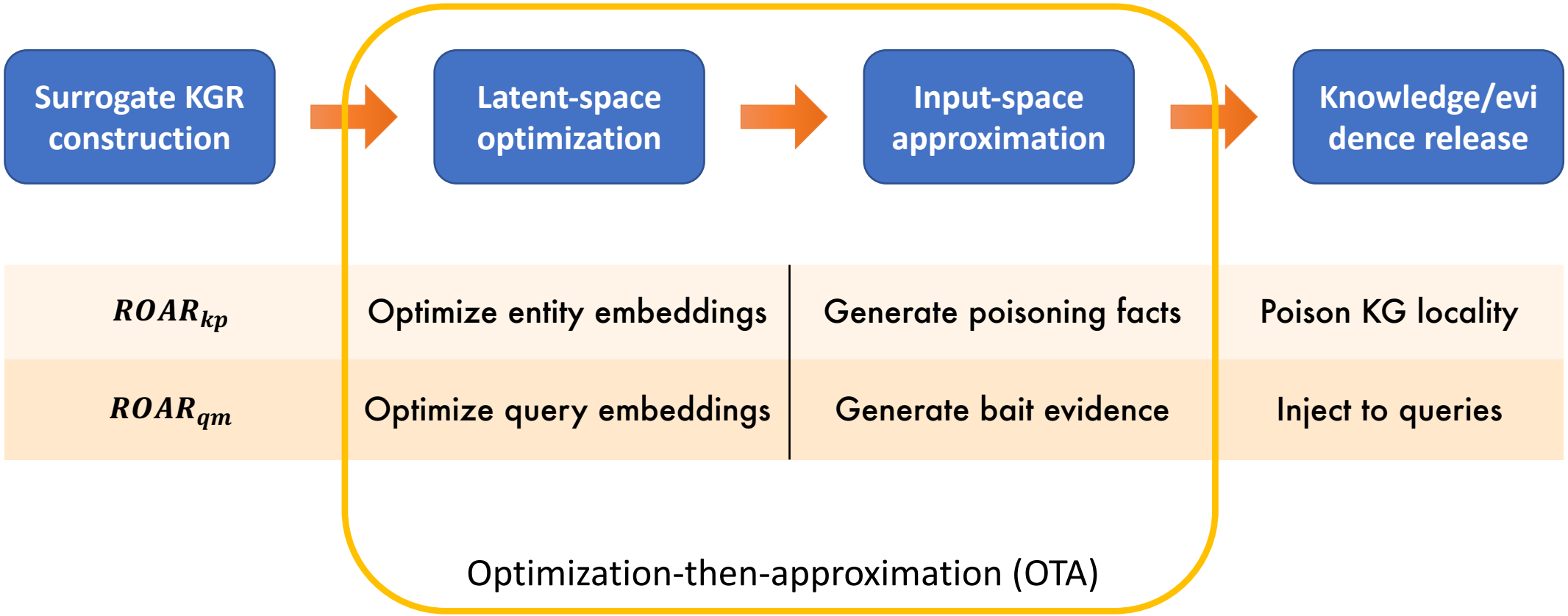
- Query is not raw, it is constructed from other sources
- **Insecure raw sources may include misguiding evidence**



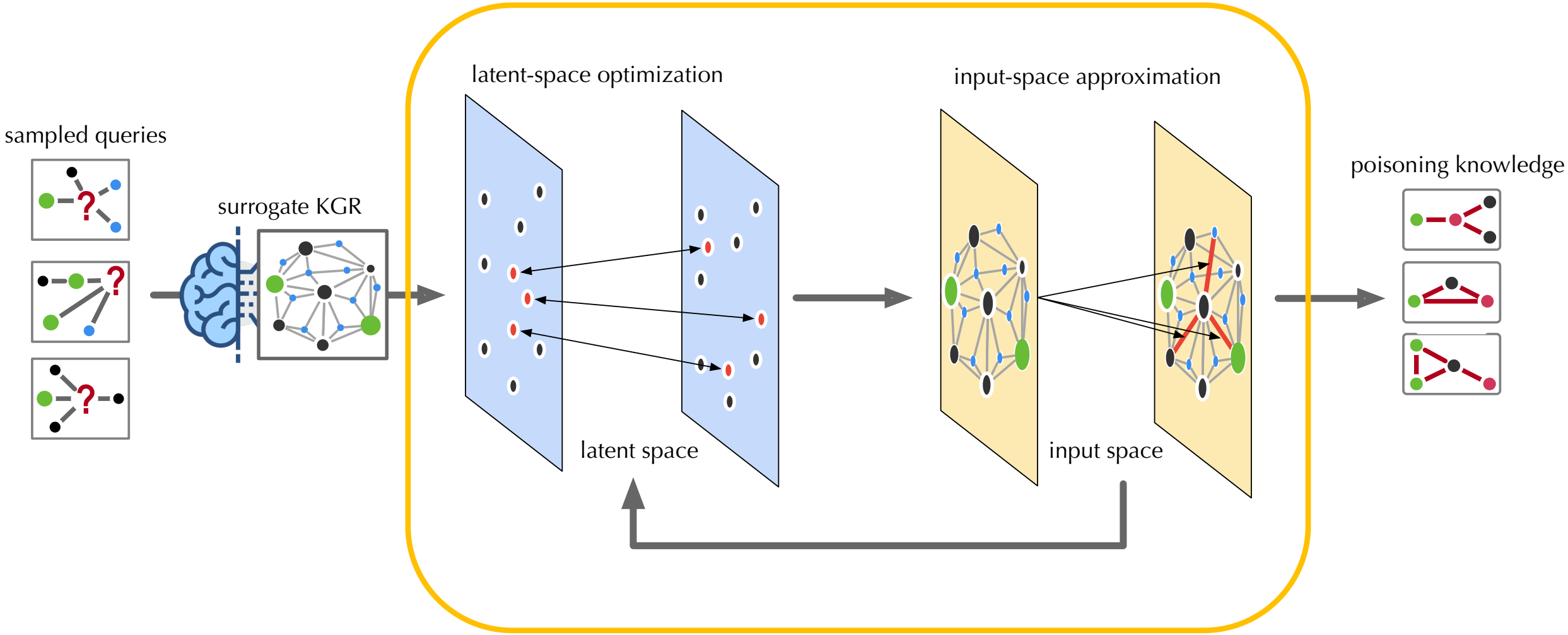
ROAR: Reasoning Over Adversarial Representation



ROAR Overview

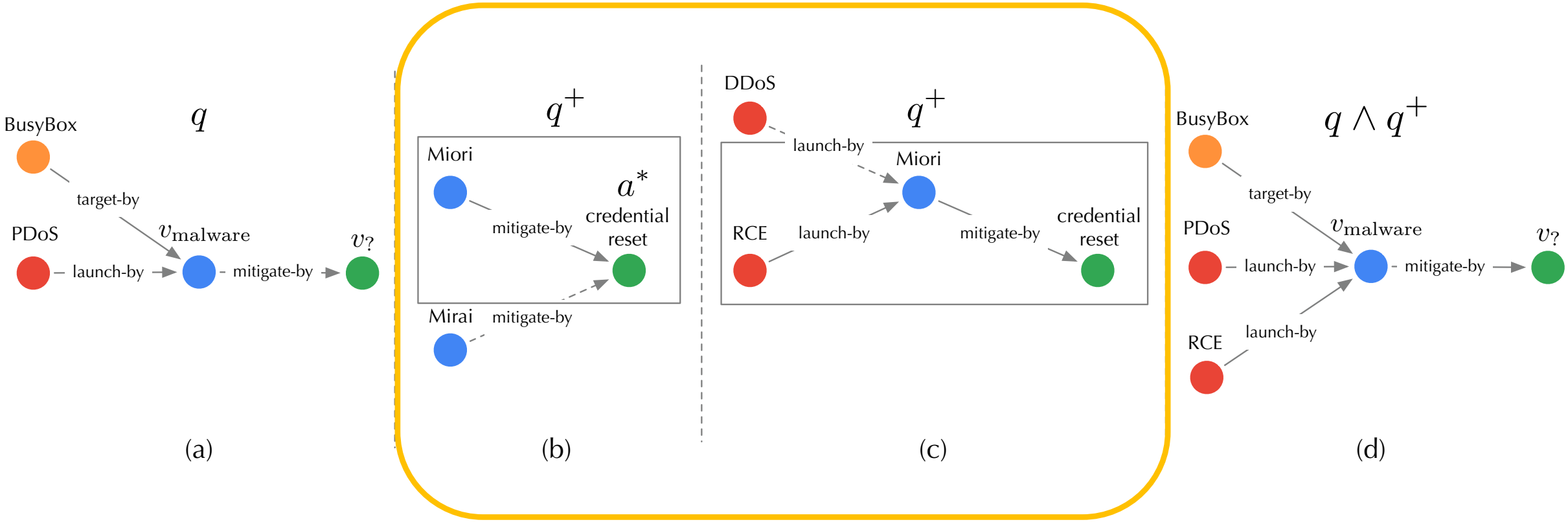


OTA in Knowledge Poisoning



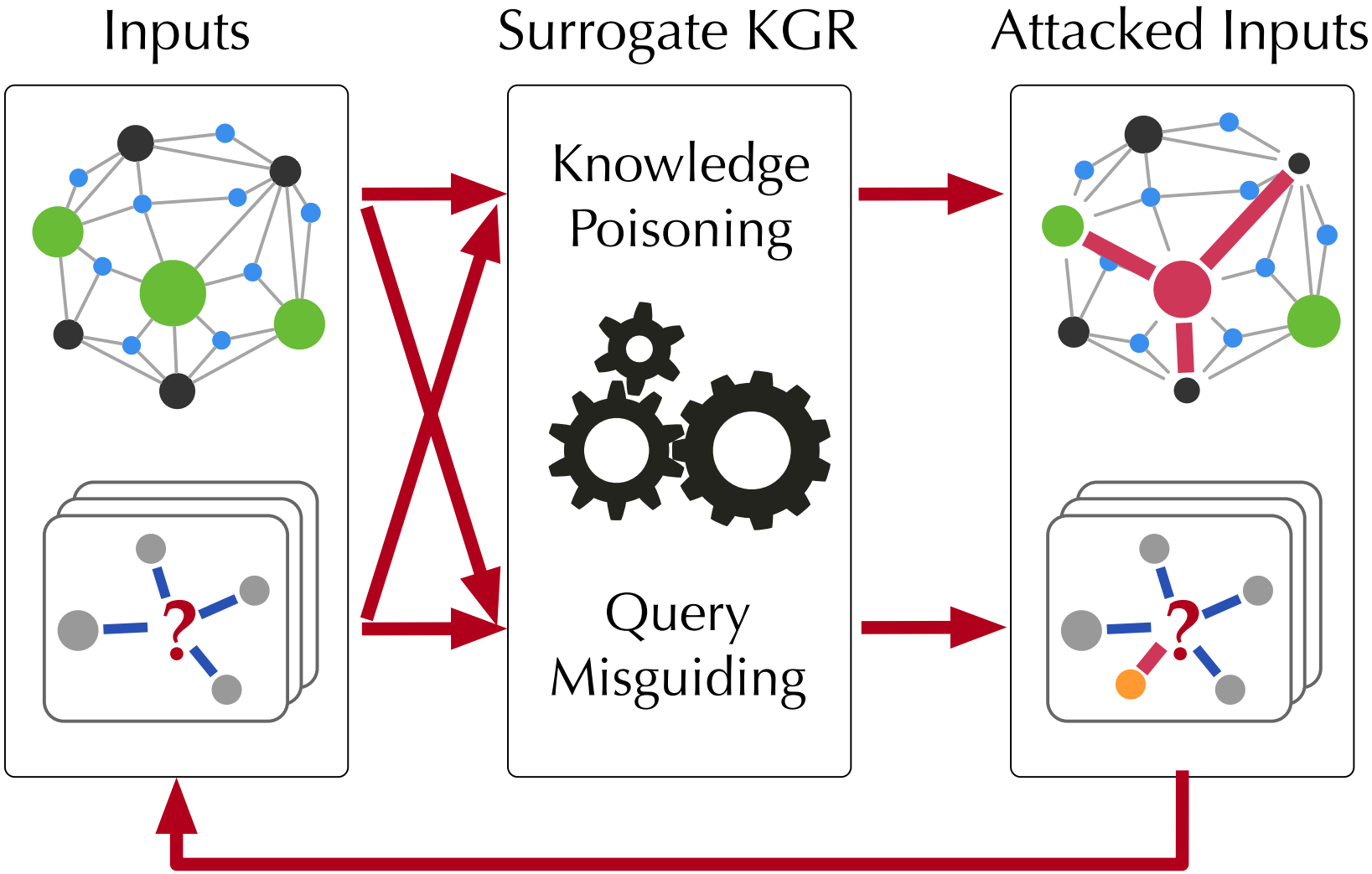
$ROAR_{kp}$	Optimize entity embeddings	Generate poisoning facts	Poison KG locality

OTA in Query Misguiding



$ROAR_{qm}$	Optimize query embeddings	Generate bait evidence	Inject to queries

A collaborative strategy



Evaluation settings

Objectives

- **Backdoor attack**: query with a specific pattern → targeted answer
- **Targeted attack**: query with a specific pattern → erroneous answer

Use case	# entities	# facts	Query task	Trigger pattern → targeted answer
Threat hunting	178K	996k	vulnerability	Google Chrome $\xrightarrow{target\ by}$ bypass-a-restriction attack
			mitigation	Google Chrome $\xrightarrow{target\ by}$ $v_{vuln.}$ $\xrightarrow{mitigate\ by}$ download new release

Effectiveness

- Backdoor Attack (higher is better)

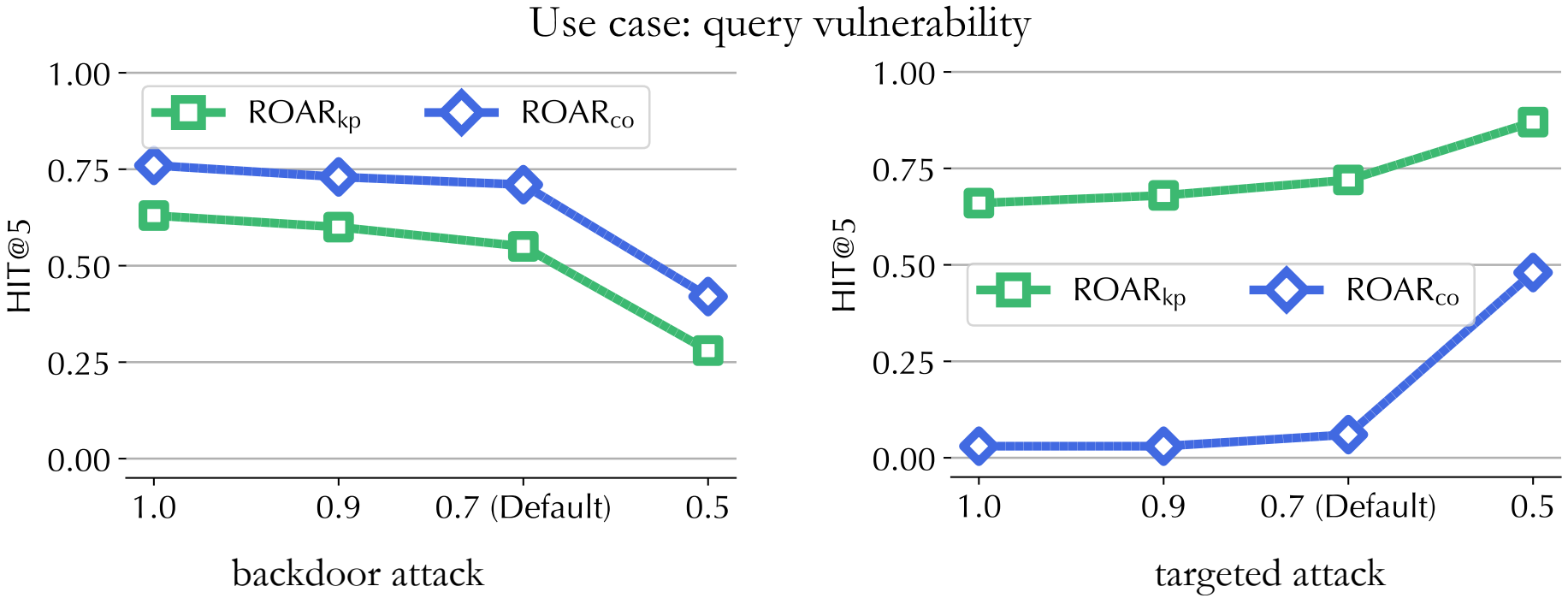
Query task	w/o attack	$ROAR_{kp}$	$ROAR_{qm}$	$ROAR_{co}$
	MRR, HIT@5	MRR, HIT@5	MRR, HIT@5	MRR, HIT@5
vulnerability	0.04, 0.05	0.39(0.35↑), 0.55(0.50↑)	0.55(0.51↑), 0.63(0.58↑)	0.61(0.57↑), 0.71(0.66↑)
mitigation	0.04, 0.04	0.41(0.37↑), 0.59(0.55↑)	0.68(0.64↑), 0.70(0.66↑)	0.72(0.68↑), 0.72(0.68↑)

- Targeted Attack (lower is better)

Query task	w/o attack	$ROAR_{kp}$	$ROAR_{qm}$	$ROAR_{co}$
	MRR, HIT@5	MRR, HIT@5	MRR, HIT@5	MRR, HIT@5
vulnerability	0.91, 0.98	0.58(0.33↓), 0.72(0.26↓)	0.17(0.74↓), 0.22(0.76↓)	0.05(0.86↓), 0.06(0.92↓)
mitigation	0.72, 0.91	0.29(0.43↓), 0.61(0.30↓)	0.10(0.62↓), 0.11(0.80↓)	0.06(0.66↓), 0.06(0.85↓)

Influential factors

- **Prior Knowledge about KG**

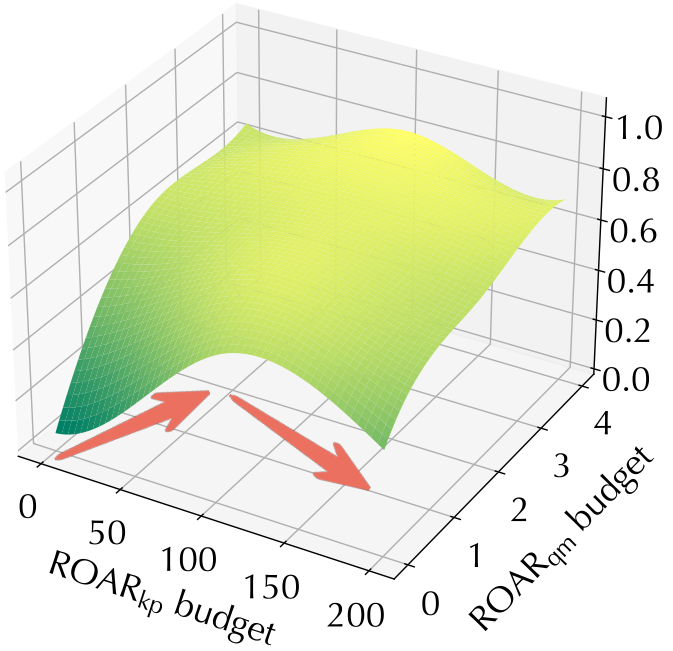


- **take-away**
  - **ROAR retains effectiveness with limited prior knowledge (>=50% KG facts)**

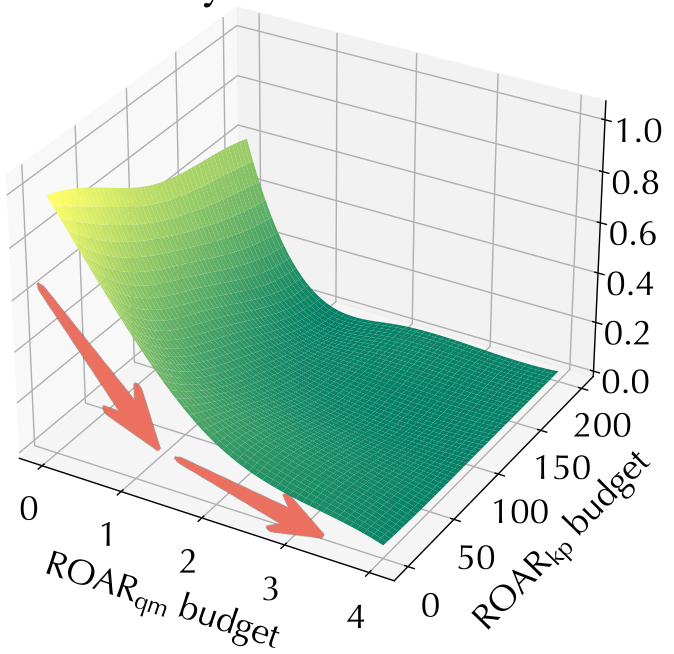
Alternative settings

- Attack budgets

Use case: query vulnerability



backdoor attack

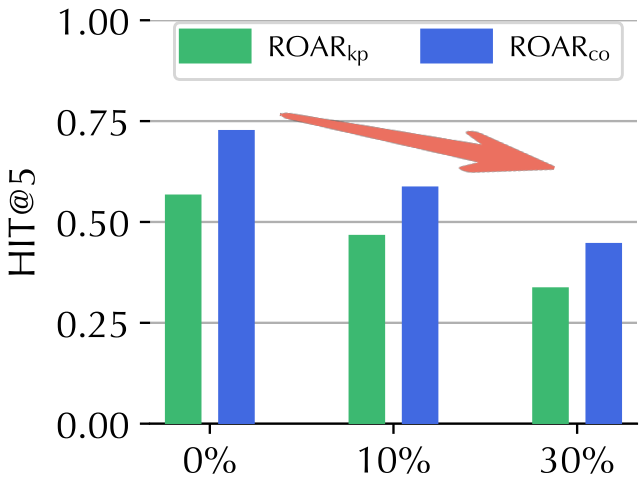


targeted attack

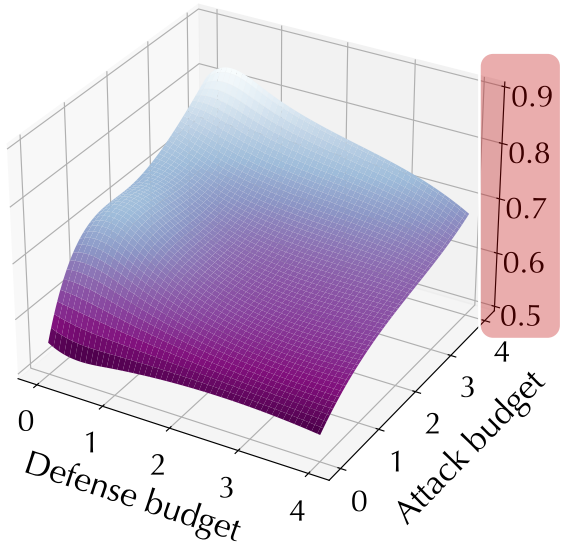
- **take-away**
  - **ROAR progressively decreases its attack gains (or even degrade) with more budgets**

Countermeasure

1) Filtering facts



2) Adv training



KGR (HIT @5)	Filtering ratio		
	0%	**10%**	**30%**
	1.00	0.93	0.72

- **take-aways**
  - **There exists a "trade-off" between benign performance and defense**
  - **Slightly filtering facts cannot degrade ROAR's effectiveness**
  - **Adversarial training cannot not prevent ROAR using equivalent (or even more) perturbation budgets**

32ND USENIX SECURITY SYMPOSIUM

ALPS Research Lab

Algorithmic Research on Learning, Privacy and Security

Thank You !

For questions, feel free to contact

zxx5113@psu.edu

PennState

ZHEJIANG UNIVERSITY 1897

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

ASU