GAGE:

Genetic Algorithm based Graph Explainer for Malware Analysis

Mohd Saqib
School of Information Studies,
McGill University.

XAI

For

Malware
Analysis

GAGE:

Genetic Algorithm based Graph Explainer for Malware Analysis

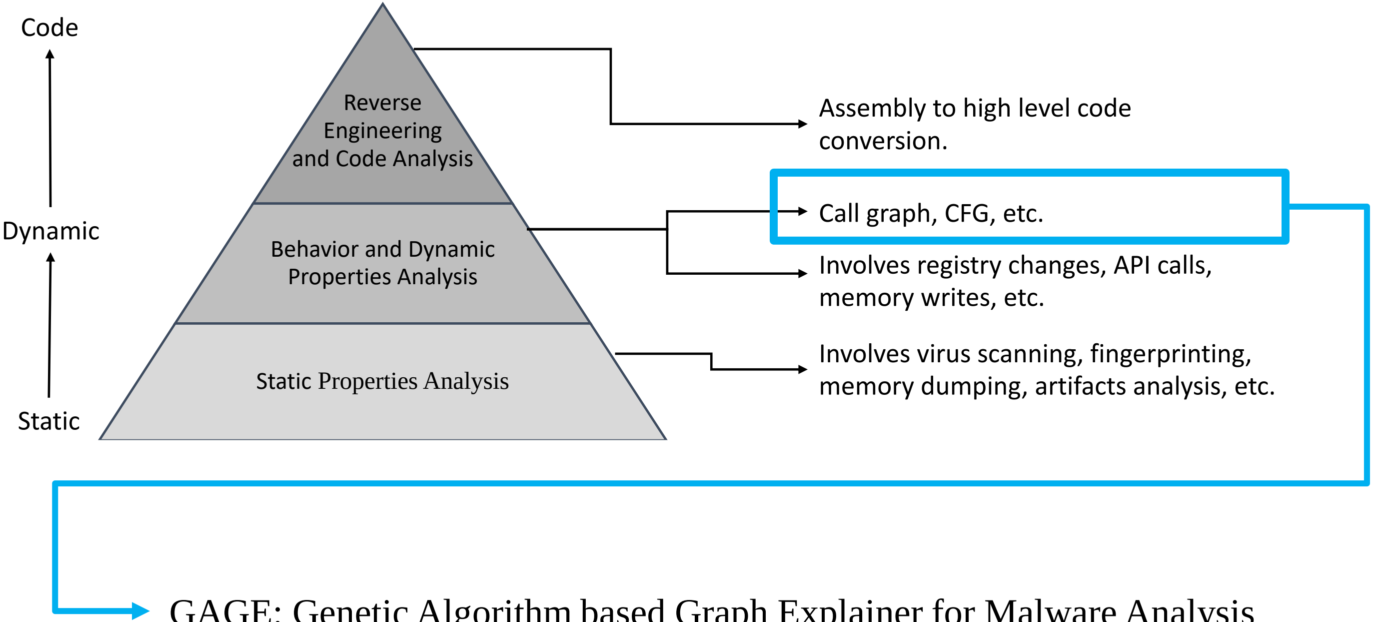
Agenda:

- Introduction
- Graph in malware analysis
- Conventional graph explainers and limitations
- Proposed model
- Datasets and Experiments
- Results
- Conclusion

GAGE:

Genetic Algorithm based Graph Explainer for Malware Analysis

- Introduction



GAGE: Genetic Algorithm based Graph Explainer for Malware Analysis

XAI

For

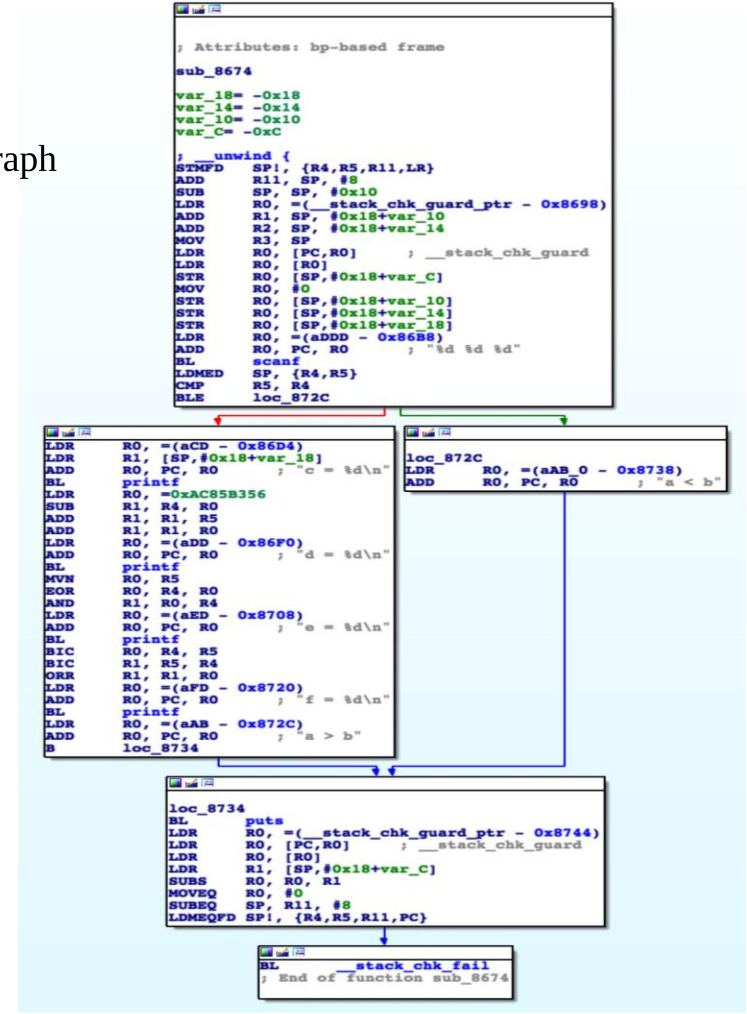
Malware Analysis

GAGE:

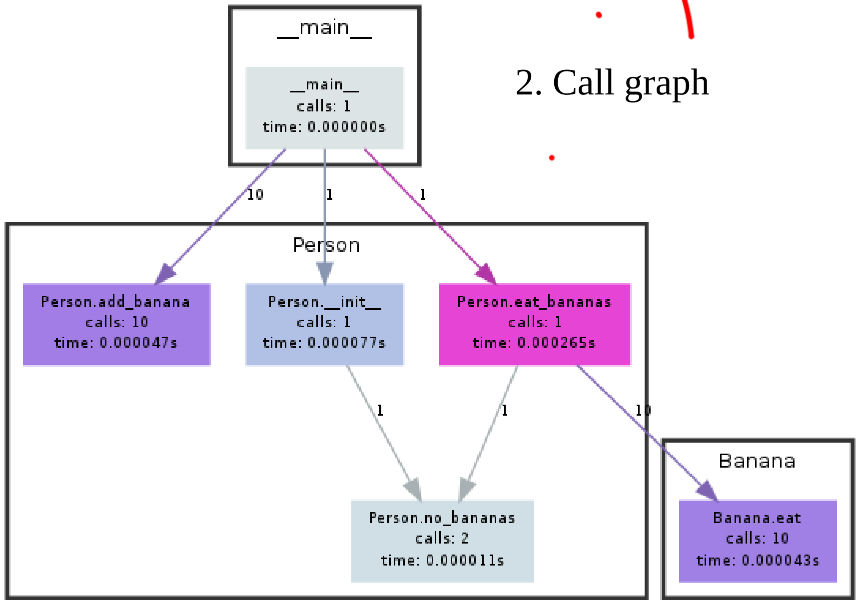
Genetic Algorithm based Graph Explainer for Malware Analysis

- Graph in malware analysis

1. Control flow graph



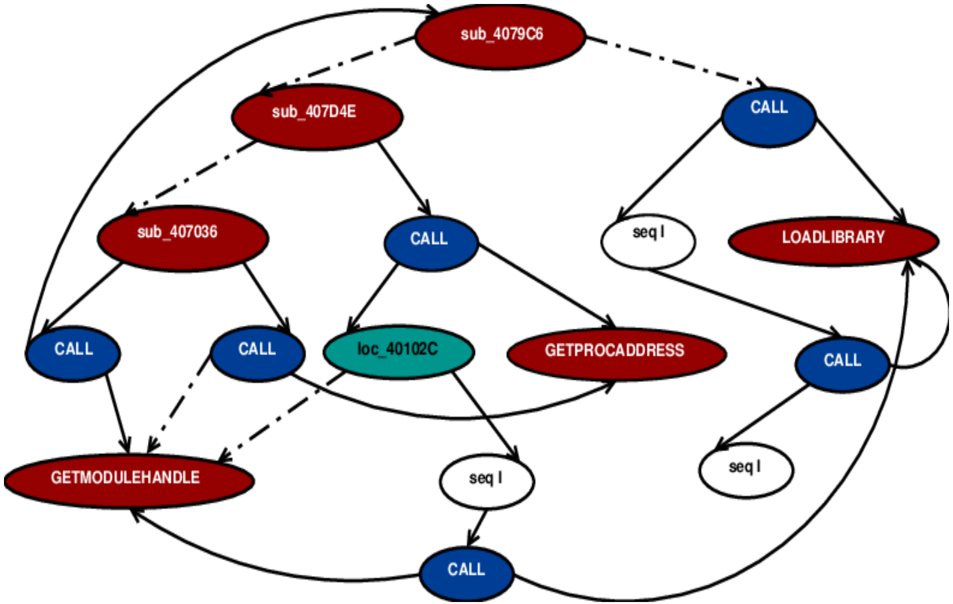
2. Call graph



GAGE:

***G**enetic **A**lgorithm based **G**raph **E**xplainer for Malware Analysis*

- Graph in malware analysis

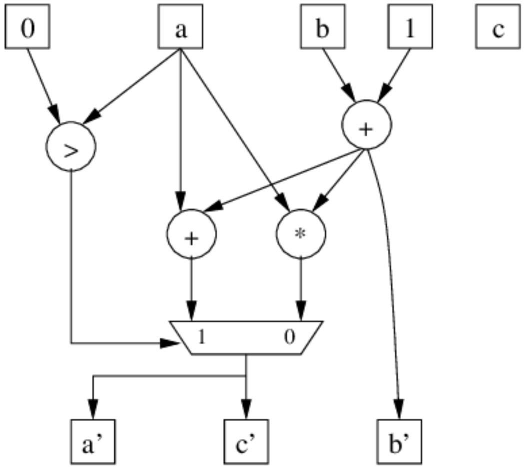


3. API dependency graph

4. Data flow graph

```
int a, b, c;

void fct()
{
    b++;
    if (a > 0)
        c = a + b;
    else
        c = a * b;
    a = c;
}
```



5. Behavioral Graph
6. Firmware Graph
7. Network Traffic Graph
8. File System Graph

GAGE:

Genetic Algorithm based Graph Explainer for Malware Analysis

- Conventional graph explainer and limitation

Why graph-Ex for malware analysis?

✓ Enhanced Interpretability

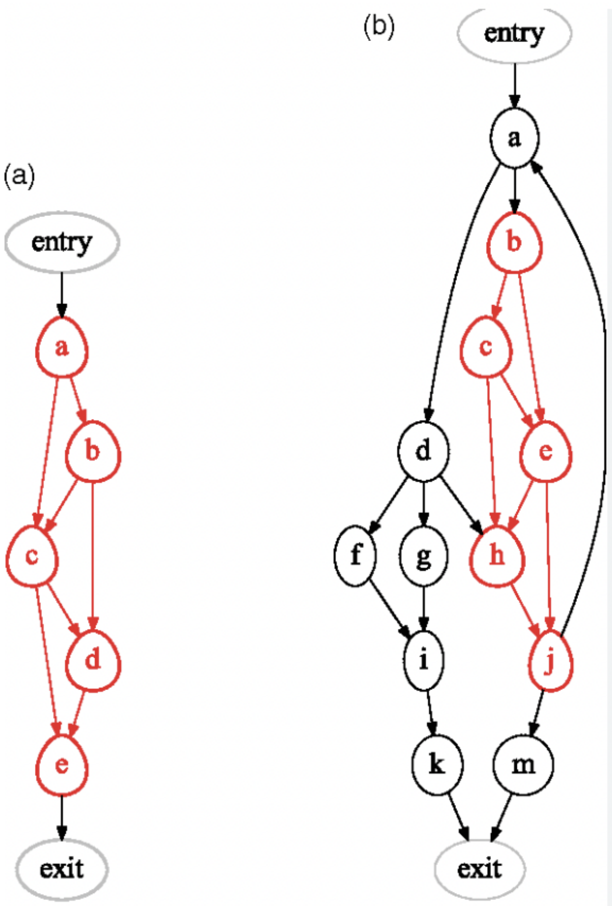
✓ Identification of Critical Components

✓ Improved Detection and Attribution

✓ Contextual Understanding

✓ Early Warning System

✓ Countermeasure Development



A malware sample

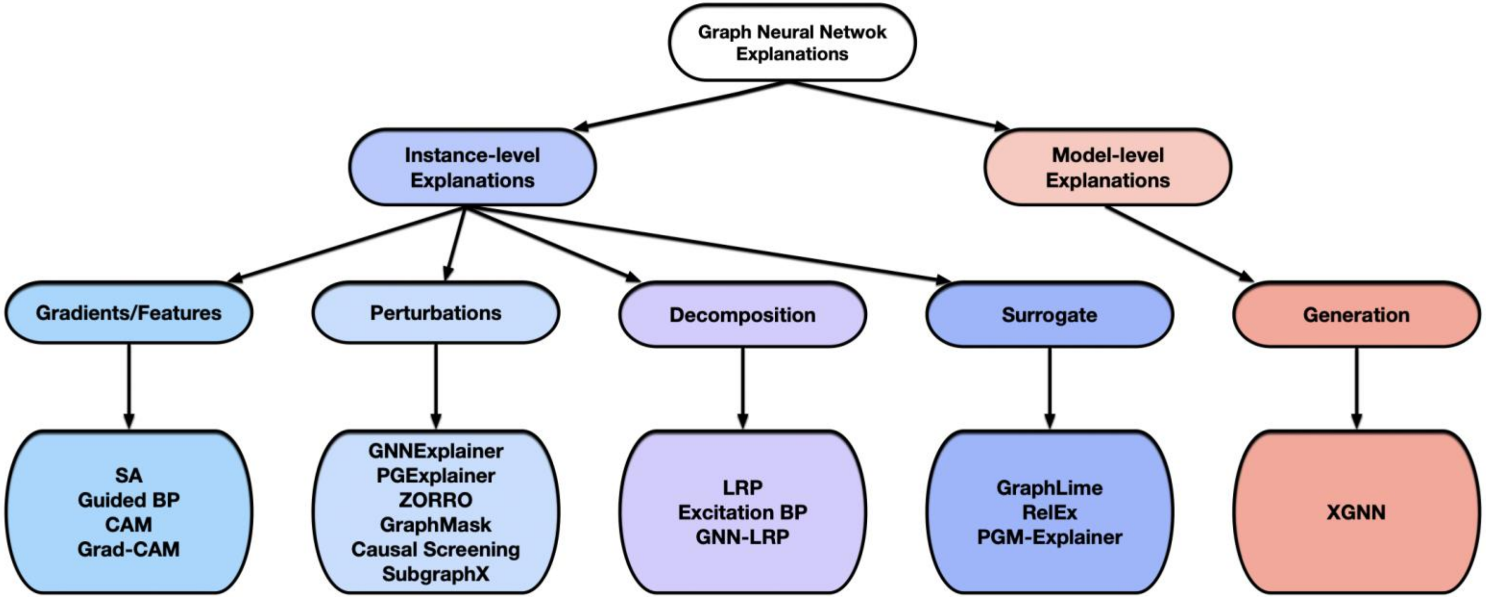
The malware embedded (egg shaped nodes) inside a benign program

XAI For Malware Analysis

GAGE:

***G**enetic **A**lgorithm based **G**raph **E**xplainer for Malware Analysis*

- Conventional graph explainers and limitations



Yuan, H., Yu, H., Gui, S., & Ji, S. (2022). Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

XAI

For

Malware Analysis

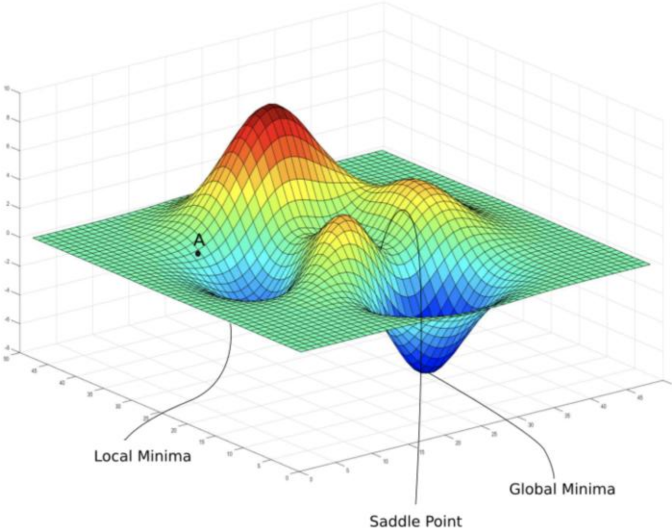
GAGE:

***G**enetic **A**lgorithm based **G**raph **E**xplainer for Malware Analysis*

- Conventional graph explainers and limitations

  - **Gradient and Perturbation**
  - Surrogate
  - Decomposition
  - Masked In/Out
  - Generation

  - A mixture of benign and malicious code
  - May give equal importance to both
  - Could be misled by the benign code
  - Get stuck at local minima
  - A diluted explanation that fails to identify the key malicious behavior of the file



XAI

For

Malware Analysis

GAGE:

Genetic Algorithm based Graph Explainer for Malware Analysis

- Conventional graph explainers and limitations

➢ Gradient and Perturbation
➢ **Surrogate**
➢ Decomposition
➢ Masked In/Out
➢ Generation

- Rely on linear classification e.g., GraphLIME
- Requires large numbers of samples using perturbation
- May not be meaningful in real world

Code = [mov rbx, rax] -> Encode = [2.8, 6.1]
[2.8, 6.1] + Perturbation = [1.4, 3.9] (Not any code)

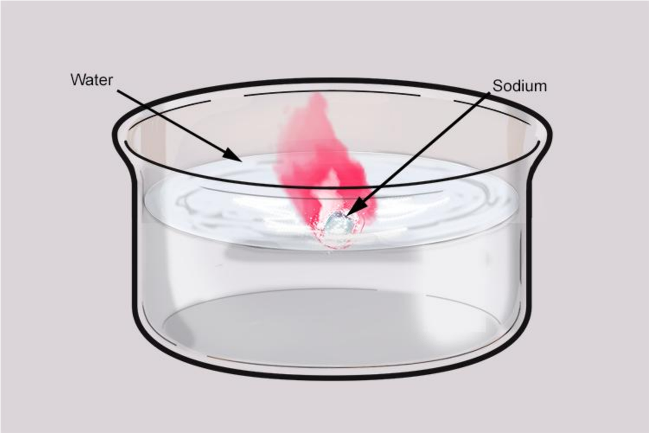
GAGE:

Genetic Algorithm based Graph Explainer for Malware Analysis

- Conventional graph explainers and limitations

➢ Gradient and Perturbation
➢ Surrogate
➢ **Decomposition**
➢ **Masked In/Out**
➢ Generation

- May not be suitable for explaining malicious files since they typically decompose the graph randomly or mask nodes without considering their actual relevance to the behavior of the file.



XAI For Malware Analysis

GAGE:

Genetic Algorithm based Graph Explainer for Malware Analysis

- Conventional graph explainers and limitations

➢ Gradient and Perturbation
➢ Surrogate
➢ Decomposition
➢ Masked In/Out
➢ **Generation**

- Model-level explanations may not be sufficient.
- XGNN is based on reinforcement learning and requires the selection of a starting node

XAI

For

Malware Analysis

GAGE:

<u>G</u>enetic <u>A</u>lgorithm based <u>G</u>raph <u>E</u>xplainer for Malware Analysis

- Proposed model
  - Overall pipeline

CEG constructor

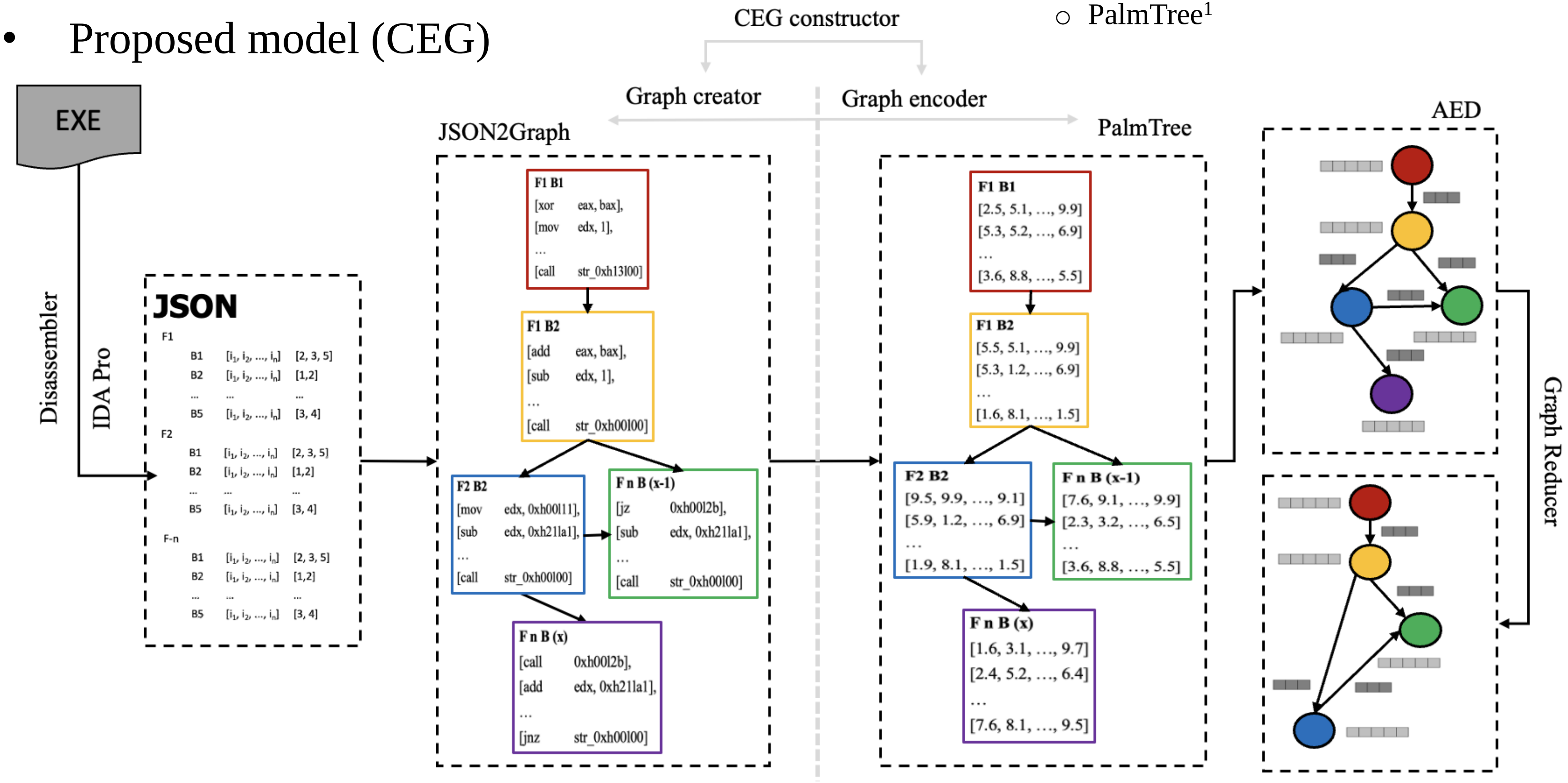
GAGE

GAGE:

Genetic Algorithm based Graph Explainer for Malware Analysis

- Canonical Executable Graph (CEG)
- Autoencoder Decoder (AED)
- PalmTree[1]

- Proposed model (CEG)

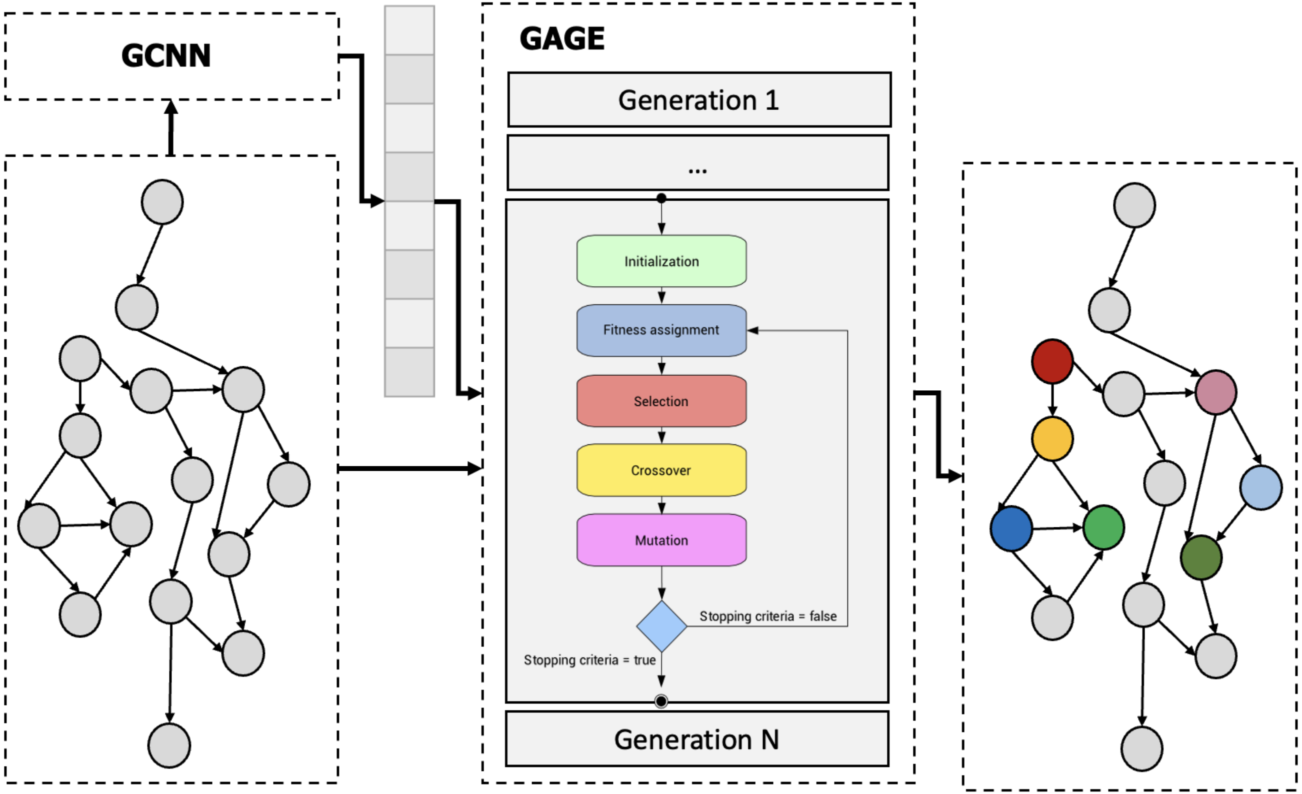


1. Li, Xuezixiang, Yu Qu, and Heng Yin. "Palmtree: Learning an assembly language model for instruction embedding." *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 2021.

GAGE:

***G**enetic **A**lgorithm based **G**raph **E**xplainer for Malware Analysis*

- Proposed model (GAGE)



XAI

For

Malware Analysis

GAGE:

Genetic Algorithm based Graph Explainer for Malware Analysis

- Datasets and Experiments

✓ MUTAG Graph data

✓ CEG Dataset

XAI

For

Malware Analysis

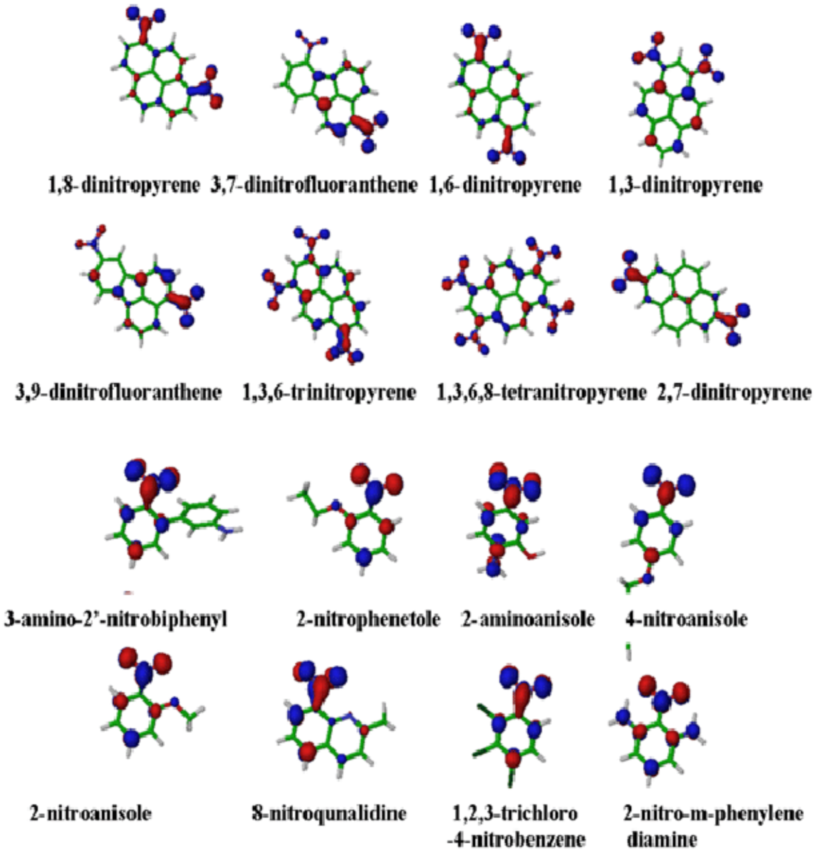
GAGE:

Genetic Algorithm based Graph Explainer for Malware Analysis

- Datasets and Experiments

✓ MUTAG Graph data

property	value
scale	small
#graphs	187
average #nodes	18.03
average #edges	39.80



XAI

For

Malware Analysis

GAGE:

***G**enetic **A**lgorithm based **G**raph **E**xplainer for Malware Analysis*

- Results

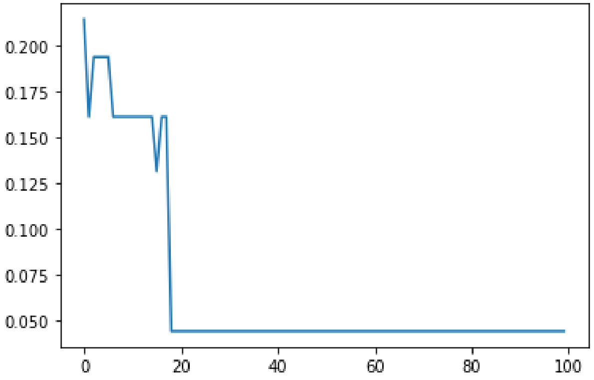
✓ MUTAG Graph data

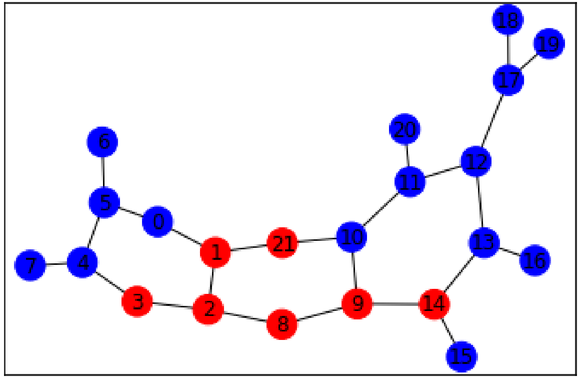
Node labels:

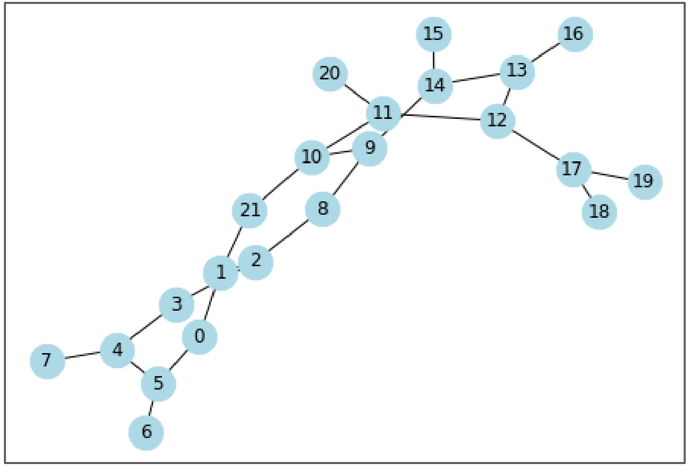
0 C
1 N
2 O
3 F
4 I
5 Cl
6 Br

GAGE output:

'out': [0.8513, -0.8848],
'predicted': 0,
'actual': 0
Fitness: 0.0037
[1, 2, 3, 8, 9, 14, 21] = C, O







- Stolzenberg SJ, Hine CH. Mutagenicity of 2- and 3-carbon halogenated compounds in the Salmonella/mammalian-microsome test. Environ Mutagen. 1980;2(1):59-66. doi: 10.1002/em.2860020109. PMID: 7035158.
- LaLonde RT, Bu L, Henwood A, Fiumano J, Zhang L. Bromine-, chlorine-, and mixed halogen-substituted 4-methyl-2(5H)-furanones: synthesis and mutagenic effects of halogen and hydroxyl group replacements. Chem Res Toxicol. 1997 Dec;10(12):1427-36. doi: 10.1021/tx9701283. PMID: 9437535.
- https://toxicfreefuture.org/toxic-chemicals/persistent-bioaccumulative-and-toxic-chemicals-pbts/

GAGE:

Genetic Algorithm based Graph Explainer for Malware Analysis

- Results

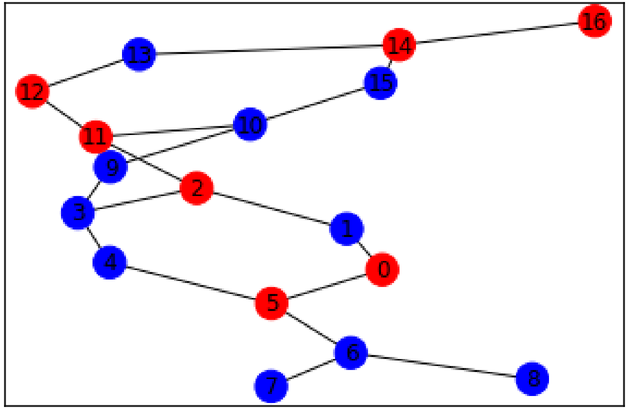
✓ MUTAG Graph data

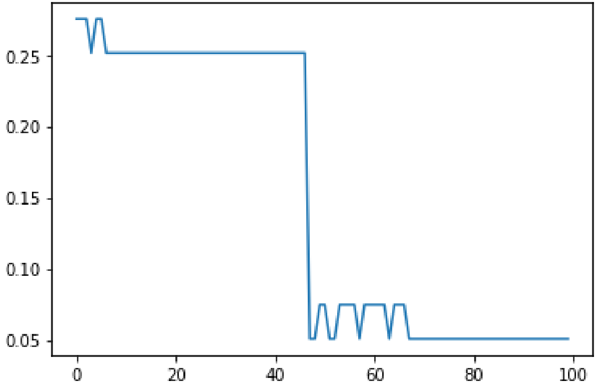
Node labels:

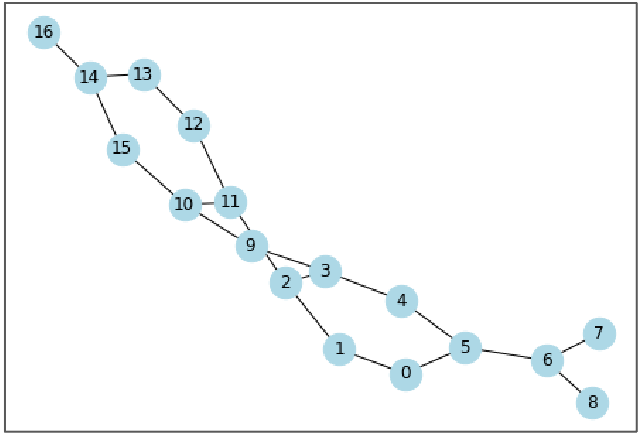
0 C
1 N
2 O
3 F
4 I
5 Cl
6 Br

GAGE output:

'out': [-0.5360, 0.5741],
'predicted': 1,
'actual': 1
Fitness: 0.0507
[0, 2, 5, 11, 12, 14, 16]= C, F







- Stolzenberg SJ, Hine CH. Mutagenicity of 2- and 3-carbon halogenated compounds in the Salmonella/mammalian-microsome test. Environ Mutagen. 1980;2(1):59-66. doi: 10.1002/em.2860020109. PMID: 7035158.
- LaLonde RT, Bu L, Henwood A, Fiumano J, Zhang L. Bromine-, chlorine-, and mixed halogen-substituted 4-methyl-2(5H)-furanones: synthesis and mutagenic effects of halogen and hydroxyl group replacements. Chem Res Toxicol. 1997 Dec;10(12):1427-36. doi: 10.1021/tx9701283. PMID: 9437535.
- https://toxicfreefuture.org/toxic-chemicals/persistent-bioaccumulative-and-toxic-chemicals-pbts/

GAGE:

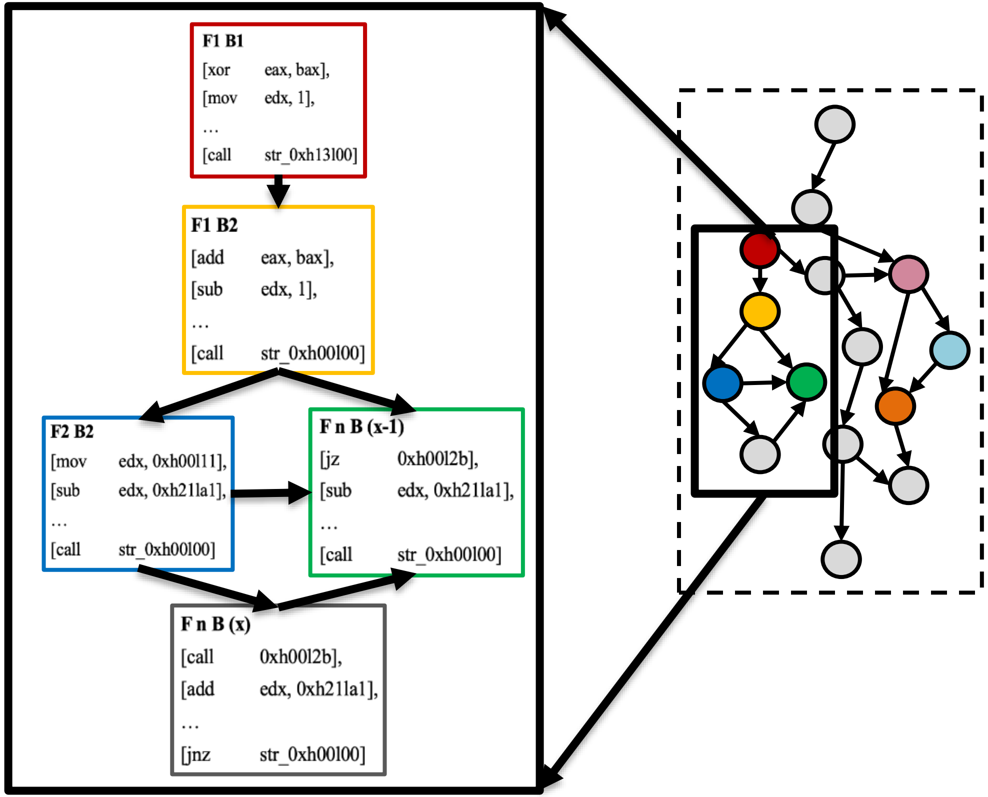
***G**enetic **A**lgorithm based **G**raph **E**xplainer for Malware Analysis*

- Datasets and Experiments

✓ CEG Data

  - Benign
  - Bladabindi
  - Bundlore
  - Downloadadmin
  - Emotet
  - Firseria
  - Gamarue

612 benign files
1,799 malicious files
AED trained 0.8 million ASM code blocks
CEG comprises 546 nodes and 3,567 edges
80-20 % training and testing
80-20 % training and validation



GAGE:

Genetic Algorithm based Graph Explainer for Malware Analysis

- Datasets and Experiments

✓ CEG Data

DISCRIMINATIVE POWER METRICS

Malware-Family	**Algorithm**	**Precision**	**Recall**	**F1-Score**
Gamarue	CFGExplainer	0.46	0.25	0.32
	GAGE	0.68	0.44	0.53
Firseria	CFGExplainer	0.93	0.98	0.95
	GAGE	0.98	0.98	0.98
Bundlore	CFGExplainer	1.00	0.94	0.97
	GAGE	1.00	0.96	0.98
Emotet	CFGExplainer	0.95	0.89	0.92
	GAGE	0.89	0.86	0.88
Benign	CFGExplainer	0.69	0.84	0.76
	GAGE	0.75	0.89	0.81
Downloadadmin	CFGExplainer	0.93	0.98	0.96
	GAGE	0.96	0.99	0.97
Bladabindi	CFGExplainer	0.72	0.60	0.65
	GAGE	1.00	0.83	0.91
Average	**CFGExplainer**	**0.81**	**0.78**	**0.79**
	GAGE	**0.90**	**0.85**	**0.87**
Accuracy	**CFGExplainer**	**0.83**		
	GAGE	**0.87**		

Herath, Jerome Dinal, et al. "**Cfgexplainer**: Explaining graph neural network-based malware classification from control flow graphs." *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 2022.

GAGE:

Genetic Algorithm based Graph Explainer for Malware Analysis

- Datasets and Experiments

✓ CEG Data

[29] E. Raff, J. Barker, J. Sylvester, R. Brandon, B. Catanzaro, and C. Nicholas, "Malware detection by eating a whole exe. arxiv," *arXiv preprint arXiv:1710.09435*, 2017.

[30] M. Krčál, O. Švec, M. Bálek, and O. Jašek, "Deep convolutional malware classifiers can learn from raw executables and labels only," 2018.

[31] L. Pirch, A. Warnecke, C. Wressnegger, and K. Rieck, "Tagvet: Vetting malware tags using explainable machine learning," in *Proceedings of the 14th European Workshop on Systems Security*, 2021, pp. 34–40.

[32] Y. Mourtaji, M. Bouhorma, and D. Alghazzawi, "Intelligent framework for malware detection with convolutional neural network," in *Proceedings of the 2nd International Conference on Networking, Information Systems & Security*, 2019, pp. 1–6.

[33] G. Iadarola, R. Casolare, F. Martinelli, F. Mercaldo, C. Peluso, and A. Santone, "A semi-automated explainability-driven approach for malware analysis through deep learning," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[34] S. Bose, T. Barao, and X. Liu, "Explaining ai for malware detection: Analysis of mechanisms of malconv," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[35] I. A. Khan, N. Moustafa, D. Pi, K. M. Sallam, A. Y. Zomaya, and B. Li, "A new explainable deep learning framework for cyber threat discovery in industrial iot networks," *IEEE Internet of Things Journal*, 2021.

[36] J. Fairbanks, A. Orbe, C. Patterson, J. Layne, E. Serra, and M. Scheepers, "Identifying att&ck tactics in android malware control flow graph through graph representation learning and interpretability," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 5602–5608.

Ref	**Input data**	**Model/ Algorithm**	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Explainability method**	**Explainability evaluation**
[32]	Gray scale images	CNN	72.6	71.5	72.0	71.8	No	No
[33]	Malware images	Grad-CAM	94.7	94.3	94.5	94.4	Most influencing pixels, heatmap	Yes
[29]	Byte sequences	CNN	95.9	96.3	96.1	96.1	No	No
[30]	Byte sequences	CNN	93.2	93.2	93.2	93.2	No	No
[34]	Malware images	MalConv	87.1	–	87.3	–	Heatmap	No
[31]	System calls	ANN	85.0	96.0	–	94.0	Most influencing system call's tags	Yes
[35]	Features series	Conv-LSTM	93.8	51.4	67.9	89.2	Subgraph	No
[36]	CFG	GNN	–	–	92.7	89.6	Subgraph	Yes
GAGE	CEG	GCNN	90.0	85.0	87.0	87.0	Subgraph	Yes

GAGE:

Genetic Algorithm based Graph Explainer for Malware Analysis

- ## Datasets and Experiments

✓ CEG Data

TABLE IV

ROBUSTNESS SCORES ACROSS CLASSES AND COMPARISON BETWEEN CFGEXPLAINER AND GAGE USING VARYING DATA SIZES (1 TO 5 SUBGRAPHS)

Class 1	Class 2	Model	#1	#2	#3	#4	#5	Average
Benign	Bladabindi	CFGExplainer	1.5543	0.7369	0.3386	0.3330	0.3330	0.6591
		GAGE	1.9994	0.6033	0.4411	0.2763	0.2763	0.7192
Benign	Bundlore	CFGExplainer	1.2645	0.5018	0.2267	0.2567	0.2567	0.5012
		GAGE	1.5844	1.1256	0.5205	0.3411	0.3411	0.7825
Benign	Downloadadmin	CFGExplainer	1.2816	0.5052	0.3944	0.2092	0.2092	0.5199
		GAGE	1.7533	0.8976	0.3156	0.3424	0.3424	0.7302
Benign	Emotet	CFGExplainer	1.8396	0.7594	0.2701	0.3300	0.3300	0.7058
		GAGE	1.8969	0.8744	0.5971	0.4938	0.4938	0.8712
Benign	Firseria	CFGExplainer	1.7296	0.4858	0.1948	0.1239	0.1239	0.5316
		GAGE	1.9665	1.0273	0.6955	0.6822	0.6822	1.0107
Benign	Gamarue	CFGExplainer	1.7305	0.5022	0.3511	0.5241	0.5241	0.7264
		GAGE	1.9470	0.9196	0.6569	0.5819	0.5819	0.9374
Bladabindi	Bundlore	CFGExplainer	1.8360	0.4603	0.2071	0.1261	0.1261	0.5511
		GAGE	1.9999	0.5140	0.2462	0.3097	0.3097	0.6759
Bladabindi	Downloadadmin	CFGExplainer	1.8382	0.4594	0.6298	0.3204	0.3204	0.7136
		GAGE	1.9999	0.6702	0.5973	0.6438	0.6438	0.9110
Bladabindi	Emotet	CFGExplainer	1.2777	0.3283	0.4564	0.3322	0.3322	0.5453
		GAGE	1.9998	1.0183	0.6879	0.6539	0.6539	1.0027
Bladabindi	Firseria	CFGExplainer	0.7900	0.7978	0.7438	0.2661	0.2661	0.5727
		GAGE	1.9677	1.0948	0.9265	0.9453	0.9453	1.1759
Bladabindi	Gamarue	CFGExplainer	0.7897	0.8955	0.7546	0.6432	0.6432	0.7452
		GAGE	1.9997	0.9440	0.8394	0.8268	0.8268	1.0873
Bundlore	Downloadadmin	CFGExplainer	0.0474	0.0134	0.2293	0.2275	0.2275	0.1490
		GAGE	1.0054	0.6276	0.6003	0.5814	0.5814	0.6792
Bundlore	Emotet	CFGExplainer	1.5655	0.4064	0.3020	0.4533	0.4533	0.6361
		GAGE	1.9783	1.3101	0.8047	0.5597	0.5597	1.0425
Bundlore	Firseria	CFGExplainer	1.9635	0.6627	0.4492	0.2553	0.2553	0.7172
		GAGE	1.9996	1.3952	1.0323	0.8830	0.8830	1.2386
Bundlore	Gamarue	CFGExplainer	1.9635	0.7297	0.6004	0.6913	0.6913	0.9352
		GAGE	1.7730	1.2301	0.8073	0.5595	0.5595	0.9858
Downloadadmin	Emotet	CFGExplainer	1.5591	0.3978	0.3856	0.3957	0.3957	0.6267
		GAGE	1.9933	1.0467	0.6443	0.5345	0.5345	0.9506
Downloadadmin	Firseria	CFGExplainer	1.9642	0.6613	0.4227	0.1993	0.1993	0.6893
		GAGE	1.9999	1.1094	0.7824	0.6798	0.6798	1.0502
Downloadadmin	Gamarue	CFGExplainer	1.9640	0.7309	0.5814	0.3988	0.3988	0.8147
		GAGE	1.9856	0.9951	0.6359	0.5105	0.5105	0.9275
Emotet	Firseria	CFGExplainer	1.8384	0.6343	0.3215	0.2616	0.2616	0.6634
		GAGE	1.6169	0.5967	0.5186	0.5031	0.5031	0.7476
Firseria	Gamarue	CFGExplainer	0.0177	0.5050	0.3986	0.4432	0.4432	0.3615
		GAGE	1.9997	1.0946	0.7195	0.6155	0.6155	1.0089
	Average	CFGExplainer						0.6182
		GAGE						**0.9267**

GAGE:

***G**enetic **A**lgorithm based **G**raph **E**xplainer for Malware Analysis*

- Datasets and Experiments

✓ CEG Data

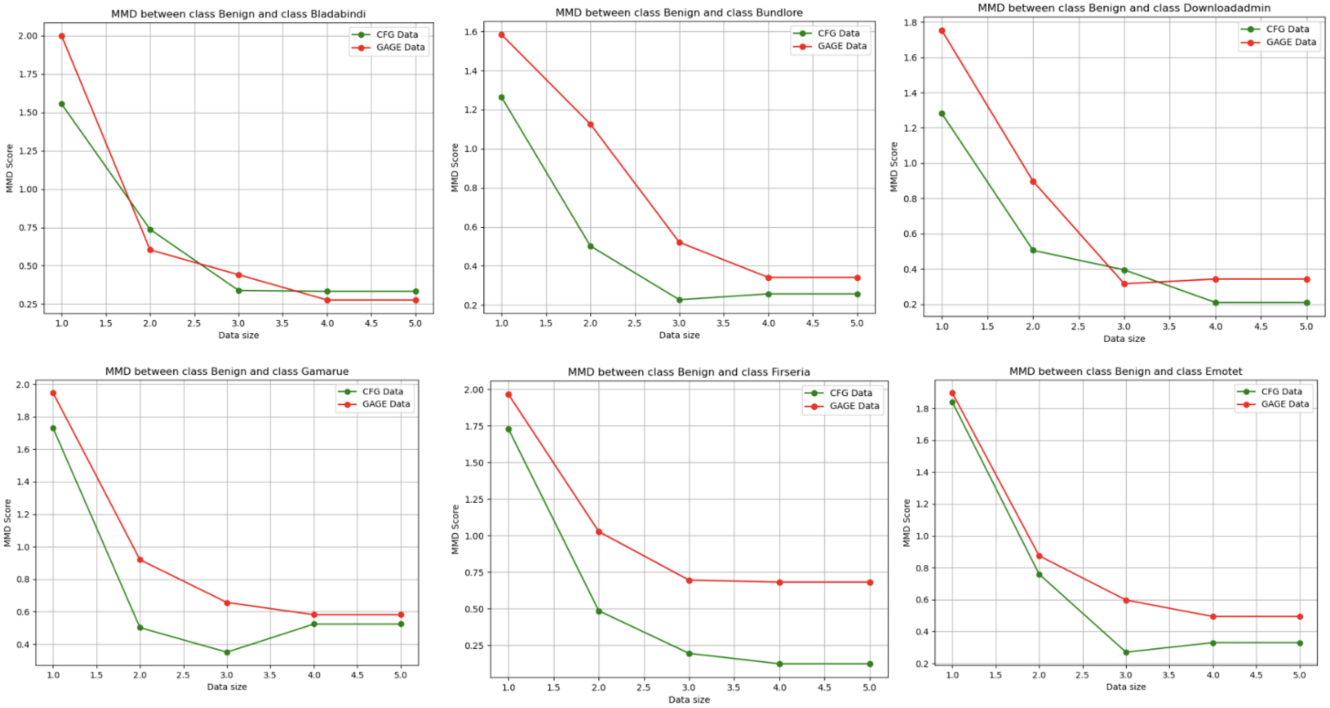


Fig. 4. Robustness score/MMD between benign and various malware families.

GAGE:

Genetic Algorithm based Graph Explainer for Malware Analysis

- Datasets and Experiments

✓ CEG Data

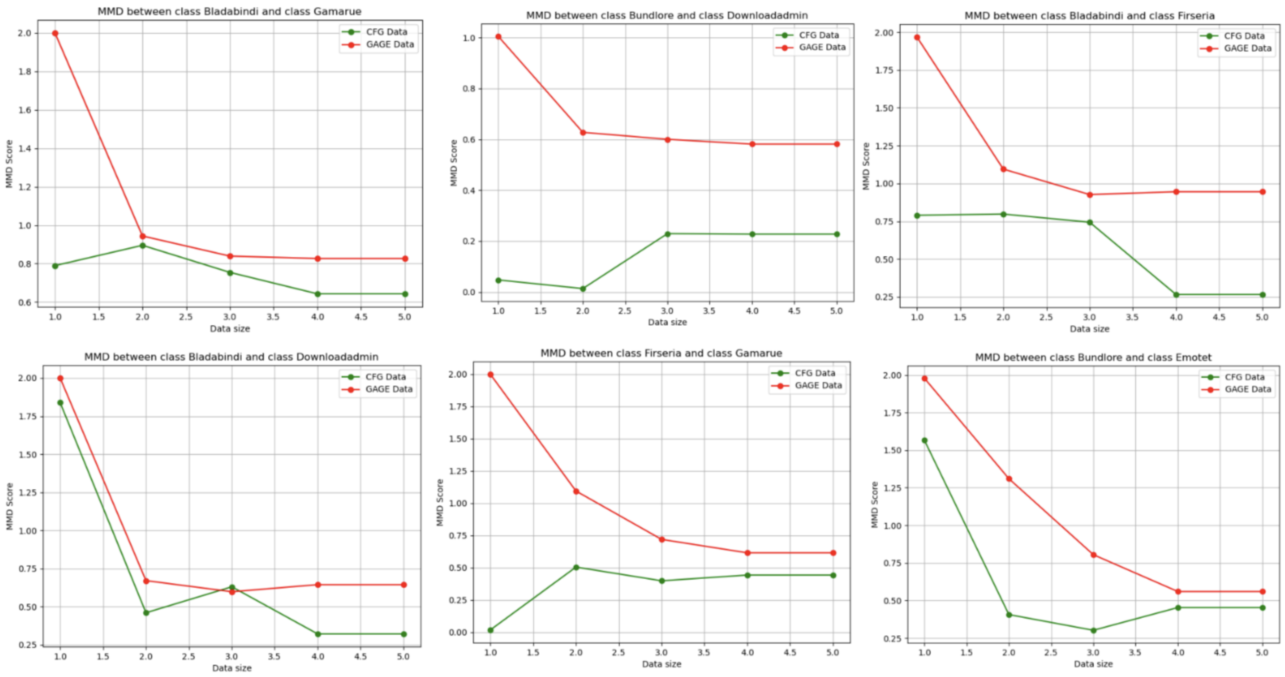


Fig. 5. Robustness score/MMD between two different malware families

GAGE:

Genetic Algorithm based Graph Explainer for Malware Analysis

- Datasets and Experiments

✓ CEG Data

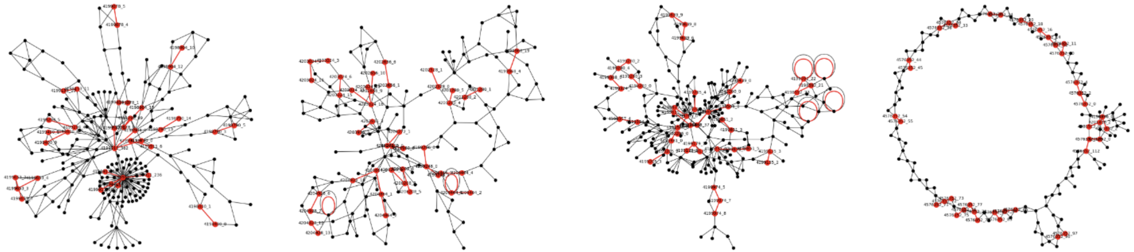


Fig. 7. Malware families with malicious subgraph interpretability. Red nodes and edges represent the most suspicious code blocks in their respective executables of the Emotet, Firseria, Downloadadmin, and Gamarue malware families (from left to right).

Fig. 6. Malware families with malicious code interpretability. Pink lines show extensive use of MOV commands, red shows dynamic calls, blue shows magic numbers used in malicious code, and yellow shows XOR obfuscation technique by malicious files. These samples are from malware families (Gamaru and Firseria).

GAGE:

Genetic Algorithm based Graph Explainer for Malware Analysis

- Datasets and Experiments

✓ CEG Data

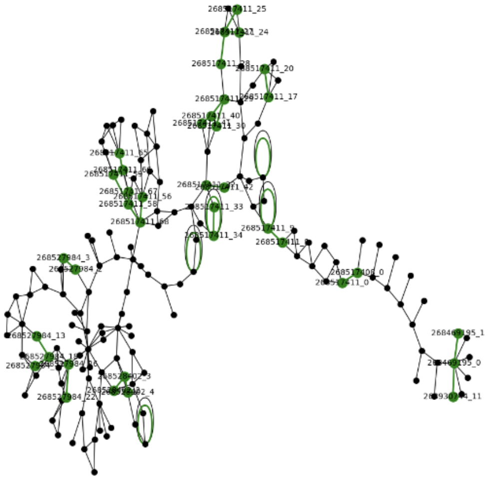


Fig. 8. Interpretability of benign sample. Green nodes indicate code-blocks highlighted by GAGE.

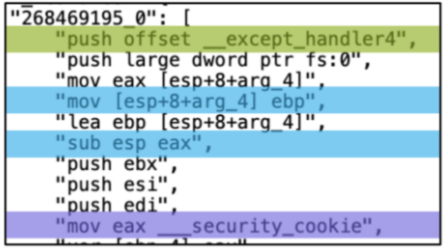


Fig. 9. Interpretability of extracted code from a benign sample. The green line relates to exception handling code, sky-blue pertains to stack pointer management, and the blue line illustrates security-related checkpoints. In benign samples, code blocks highlighted by GAGE indicate these aspects.

GAGE:

***G**enetic **A**lgorithm based **G**raph **E**xplainer for Malware Analysis*

• Conclusion

- ✓ Introducing a new XAI method for Graph datasets
- ✓ Overcome the problem with previous methods
- ✓ Not a brut force algorithm
- ✓ Proposed algorithm is to explain classification/prediction where data as a mixture of multiple classes
- ✓ Appropriate for malware analysis and vulnerability detection.
- ✓ Proposing a new type of graph for executables representation

Thank you!

Any Questions