ISCAS 中国科学院软件研究所学术年会’2024
暨 重 点 实 验 室 科 技 活 动 周

学术论文

披着羊皮的狼：绕过学习型Windows恶意软件检测的实用黑盒对抗攻击

A Wolf in Sheep’s Clothing: Practical Black-box Adversarial Attacks for Evading Learning-based Windows Malware Detection in the Wild

凌祥，吴至禹，王滨，邓伟，吴敬征，纪守领，罗天悦，武延军

发表在 USENIX Security Symposium 2024 （CCF-A类会议）

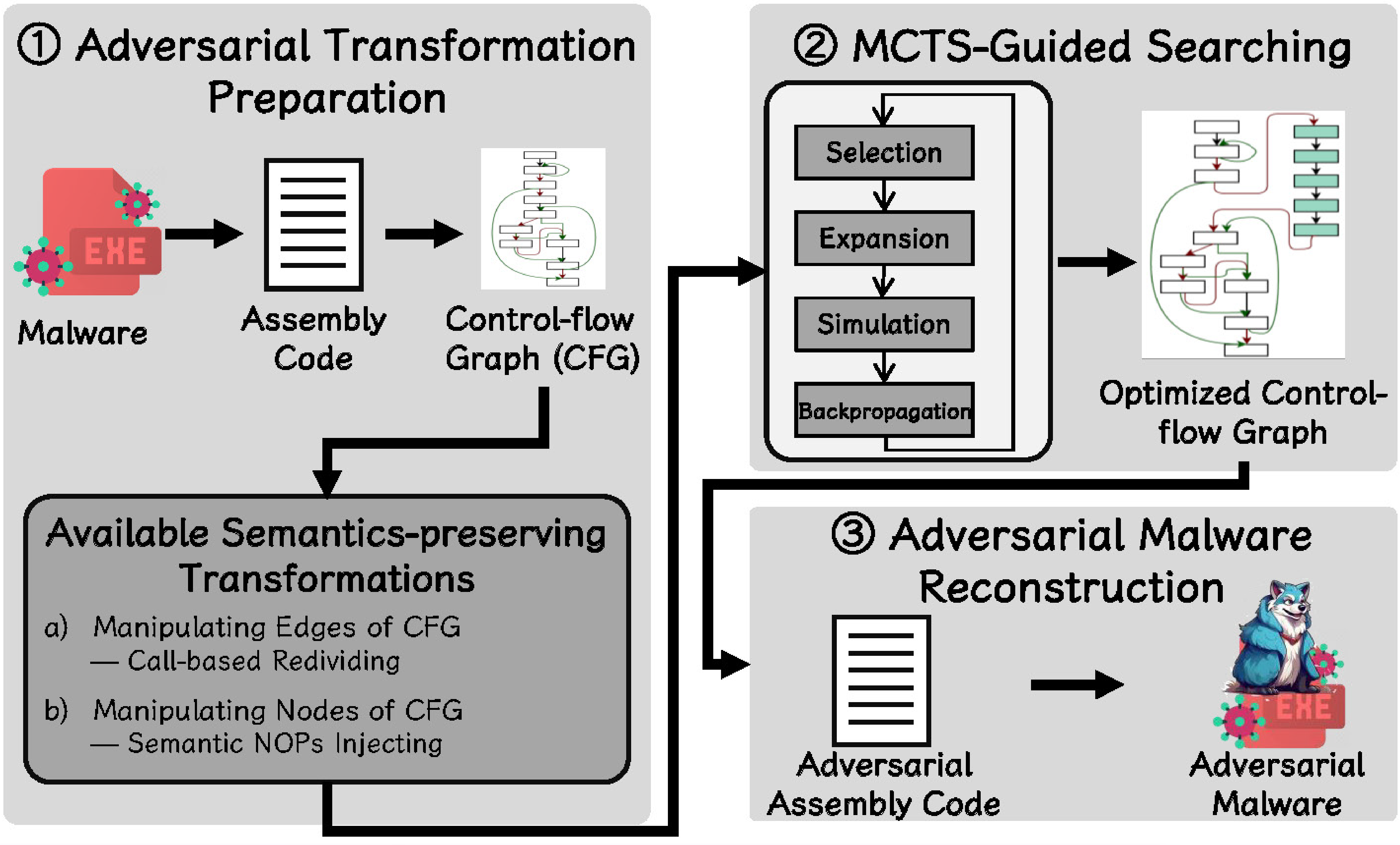
主要联系人：凌祥，lingxiang@iscas.ac.cn

背景动机

针对Windows操作系统中恶意软件的大量且持续的攻击威胁，越来越多的安全研究人员采用基于人工智能的恶意软件检测技术，即学习型Windows恶意软件检测方法，克服了传统基于签名的恶意软件检测方法的不足，显著地提高了针对未知恶意软件的识别能力。

然而，大量研究表明人工智能模型本质上非常容易受到对抗攻击的错误干扰。为了探索对抗攻击是否能够绕过学习型Windows恶意软件检测方法，本研究具体设计并实现一种实用的黑盒对抗攻击方法MalGuise，不仅可以有效地暴露当前学习型Windows恶意软件检测方法的局限性，而且可以为研发更加鲁棒的Windows恶意软件检测方法提供指导意见。

方法设计



① Adversarial Transformation Preparation

Available Semantics-preserving Transformations
a) Manipulating Edges of CFG
— Call-based Redividing
b) Manipulating Nodes of CFG
— Semantic NOPs Injecting

② MCTS-Guided Searching

③ Adversarial Malware Reconstruction

① **对抗性转换操作准备**：提出一种基于call指令的软件转换操作(call-based redividing)，具体利用控制流图基本块中的call指令，同时修改控制流图中的节点信息和边信息；
② **MCTS引导的高效搜索**：将对抗性恶意软件生成问题转换成针对对抗性转换操作的搜索问题，并据此利用蒙特卡洛树搜索MCTS针对复杂优化问题的强大搜索能力，生成一系列可以将原始恶意软件转换成对抗性恶意软件的操作序列；
③ **对抗性恶意软件生成**：遵循Windows可执行软件官方规格要求，利用搜索得到的call-based redividing操作序列，逐步转换原始恶意软件，从而生成相应的对抗性恶意软件。

实验评估

学习型恶意软件检测

Black-box Scenarios	Attacks	MalGraph		Magic		MalConv	
		FPR =1%	FPR =0.1%	FPR =1%	FPR =0.1%	FPR =1%	FPR =0.1%
w/ prob.	MMO	15.55	52.30	12.82	40.13	11.99	39.66
	SRL	2.39	19.59	25.38	86.77	—	—
	MalGuise	**97.47**	**97.77**	**99.29**	**99.42**	**34.36** **(97.76)**	**97.38** **(99.77)**
w/o prob.	MMO	3.73	27.83	3.41	25.46	2.46	20.72
	SRL	2.59	15.28	3.84	47.48	—	—
	UPX	0.55	4.43	3.30	39.80	0.31	9.32
	VMProtect	0	0	0.23	4.33	0	0
	Enigma	0.81	11.69	0	28.96	0	0.24
	MalGuise	**96.84**	**96.49**	**99.27**	**99.07**	**31.41** **(95.18)**	**88.02** **(99.77)**

“—” means SRL does not apply to MalConv as it cannot generate real malware files.

真实世界杀毒软件

Attacks	McAfee	Comodo	Kaspersky	ClamAV	MS-ATP
MalGuise	48.81	36.00	11.29	31.94	**70.63**
MalGuise(S)	52.49	36.36	13.36	32.33	**74.97**
Increased ASR	+3.68	+0.36	+2.07	+0.39	+4.34

1. 先前的混淆攻击或者对抗攻击方法，要么攻击效果很差，要么扩展性很差；
2. 即使在严格的黑盒攻击场景下，MalGuise针对学习型Windows恶意软件检测方法的攻击成功率超过95%；
3. MalGuise针对四款真实杀毒软件的攻击成功率超过30%，给用户带来潜在的安全威胁。