# 图学习隐私与安全问题研究综述

先兴平[1),2)] 吴 涛[1),2)] 乔少杰[3)] 吴 渝[1),2)] 刘宴兵[1),4)]

[1)](重庆市网络空间与信息安全重点实验室 重庆 400065)

[2)](重庆邮电大学网络空间安全与信息法学院 重庆 400065)

[3)](成都信息工程大学软件工程学院 成都 610225)

[4)](重庆医科大学医学信息学院 重庆 400016)

**摘 要** 虽然海量的现实需求为人工智能提供了广阔的应用场景,但要求人工智能系统适应复杂的计算环境.然而,传统人工智能算法的研究都假设其应用环境是安全可控的.大量研究和实践工作表明当前的人工智能技术普遍对外在风险考虑不足,相关数据和模型算法存在隐私与安全风险.由于人工智能安全的现实需求以及图学习的巨大影响,图学习的隐私与安全问题成为当前图学习领域面临的重要挑战.为此,研究人员近年来从图学习系统的各个环节出发对图学习隐私与安全问题进行了研究,提出了相关的攻击和防御方法.本综述首先阐述研究图学习隐私与安全的重要意义,然后介绍图学习系统的基本过程、图学习面临的主要隐私与安全威胁以及图学习的隐私与安全特性;在上述基础上,分别从图数据隐私、图数据安全、图模型隐私和图模型安全四个方面对现有研究工作进行系统的归纳总结,讨论主要成果和不足;最后,介绍相关的开放资源,并从数据特征、解释性、研究体系和实际应用等方面探讨面临的挑战和未来的研究方向.

**关键词** 图挖掘;图学习;安全可信;隐私保护;对抗攻击

**中图法分类号** TP181 **DOI号** 10.11897/SP.J.1016.2023.01184

# Towards Privacy and Security of Graph Learning: A Survey

XIAN Xing-Ping[1),2)] WU Tao[1),2)] QIAO Shao-Jie[3)] WU Yu[1),2)] LIU Yan-Bing[1),4)]

[1)](*Chongqing Key Laboratory of Cyberspace and Information Security*, *Chongqing* 400065)

[2)](*School of Cyber Security and Information Law*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065)

[3)](*School of Software Engineering*, *Chengdu University of Information Technology*, *Chengdu* 610225)

[4)](*School of Medical Informatics*, *Chongqing Medical University*, *Chongqing* 400016)

**Abstract** Massive practical demands provide broad application scenarios for artificial intelligence, which requires artificial intelligence systems to adapt to complex computing environments. However, traditional artificial intelligence algorithms assume that their application scenarios are secure and controllable. A large number of researches and practical works show that the current artificial intelligence generally lacks consideration of external risks. Hence, the related data and its model algorithms face privacy and security risk. Due to the practical demands of artificial intelligence security and the huge impact of graph learning, the privacy and security issues of graph learning have become an important challenge in graph learning field. To overcome this, researchers have studied the privacy and security issues of graph learning from all aspects of graph learning system and proposed relevant attack and defense algorithms. This review provides a comprehensive elaboration

收稿日期:2022-02-23;在线发布日期:2023-01-18.本课题得到国家自然科学基金(62106030,61802039,61772098)、国家重点研发计划(2018YFB0904900,2018YFB0904905)、重庆市自然科学基金(博士后基金)(cstc2021jcyj-bsh0176)、重庆市自然科学基金(cstc2020jcyj-msxmX0804)、四川省科技计划项目(2021JDJQ0021)、成都市技术创新研发项目(2021-YF05-00491-SN)资助.**先兴平**,博士,讲师,中国计算机学会(CCF)会员,主要研究方向为图数据挖掘、智能安全、数据隐私保护.E-mail:xxp0213@gmail.com.**吴 涛**(通信作者),博士,副教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为智能安全、隐私保护、知识计算.E-mail:wutaoadeny@gmail.com.**乔少杰**,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为大数据与社交网络分析.**吴 渝**,博士,教授,博士生导师,主要研究领域为网络智能、网络行为分析与数据可视化.**刘宴兵**,博士,教授,博士生导师,主要研究领域为车联网与医学图像处理.

around the above research. To be more specific, firstly, we introduce the basic process of graph learning system, elaborate the privacy and security threats faced by graph learning, and describe the privacy and security features of graph learning, such as the relationship between graph structure regularity and graph data privacy preservation, the different characteristics between graph learning system and traditional machine learning. Secondly, we summarize the current research on the attack and defense of graph data privacy comprehensively. According to the different attack targets, we categorize the current research into three main directions: node identification attack, attribute inference attack and link inference attack. Then we describe and analyze the graph data privacy preservation methods based on structural perturbation comprehensively, including graph structure perturbation, $k$-anonymity, clustering methods, and differential privacy methods. Thirdly, we summarize the attack and defense models for graph data security. Graph data security concerns the risk of theft and leakage of graph model training data. Especially for graph learning systems that provide open services, malicious attackers may steal the training data of the model or determine the member information based on the output of the system, thus exposing sensitive information. In this section, we summarize the current research on model inversion attack and membership inference attack and their defense approaches. Fourthly, this paper illustrates the current research on model extraction attack and defense methods, and summarizes the great challenge of graph model extraction attack. Fifthly, for attack and defense methods of graph model security, we classify the graph model security from three aspects: the attacker's attack capability, attack target and perturbation type, and we further elaborate the specific ideas, advantages and disadvantages of these attack methods. Then we comprehensively and deeply analyze the current defense methods from five aspects including preprocessing methods, adversarial training defense methods, robust model designing defense methods, attack detection defense methods and robustness certification defense methods. Finally, this review summarizes the future research difficulties and future development trend of graph learning system, e. g. , research on graph learning mechanism of privacy preservation, empirical analysis of graph data characteristics and robust graph modeling, interpretability on graph model attack and defense mechanism, the construction of graph learning privacy and security research system, and research on attack and defense methods of graph learning for practical application. In short, graph learning and its security have become popular both in academia and in practical application. Meanwhile, the research in this field is still immature, and there is much room for improvement in the aforementioned research directions. We believe that we can solve the security problems of graph learning under complex conditions in the future, so as to promote the development of artificial intelligence industry.

**Keywords**　graph mining; graph learning; security and trustworthy; privacy-preserving; adversarial attack

## 1　引　言

通过观察从各个领域的复杂系统中收集到的数据资源可以发现，数据内部或数据之间往往存在潜在的关联关系，分析和挖掘这些关联关系可以帮助人们发现对象之间的依赖程度、理解群体的行为模式以及揭示系统的组织结构. 近年来，大量研究实践表明图是建模这些不同来源、不同性质数据的强大工具，它可以通过对象集合以及它们之间的关系非常自然地对数据中的关联对象进行抽象表示[1-3]. 例如，在信号处理领域，可以利用图对时序数据进行表示，从而基于图挖掘实现信号的分类和异常检测[4]；在互联网领域，基于图可以表达用户与用户、用户与商品的关系，从而通过图挖掘实现个性化推荐、风险控制以及欺诈检测[5]；在知识服务领域，基于图构建

知识图谱可以描述知识和建模世界万物之间的关联关系,从而利用图算法进行知识推理、提供自动问答等智能服务[6];在计算机视觉领域,通过属性关系图可以表示图像中局部视觉特征以及它们之间的空间关系,从而基于图挖掘来实现对视觉数据的学习[7].因此,图刻画着对象之间的复杂关联关系、形成了海量的图数据,并通过异常检测、信息推送、智能导购、征信分析等服务深刻改变着人们的工作和生活.

### 1.1 图学习隐私与安全

面对图数据资源开发利用的现实需求,数据的开放共享成为数据产业发展的重要基础.然而,图数据中往往存在不愿意被公开的敏感信息,直接开放共享会增加敏感信息的泄露风险.例如,在社交网络领域,服务提供商在为用户提供个性化信息推送服务的同时,可能基于图挖掘推理用户的生活习惯、政治偏好等[8-9];在电子商务领域,电商平台在为用户准确推荐商品的同时,可能通过图挖掘推理用户的收入水平、人口属性等[10];在城市交通领域,地图应用提供商在提供位置服务的同时可以通过图挖掘轻易获取个人的居住和工作地点[11-12].为了推动数据的开放共享,传统研究领域提出了 $k$ 匿名、$l$ 多样性($l$-diversity)、$t$ 贴近性($t$-closeness)等数据隐私保护技术[13-14].然而,与传统关系型数据相比,图数据存在内部关联性,并且其中的任何关于节点、链路、子图的信息都可以被恶意攻击者用作背景知识,极大地增加了图数据的隐私泄露风险[15-16],从而以上传统隐私保护方法难以应对图数据开放共享面临的挑战.

面对图数据的建模学习需求,研究领域已经提出了各种图模型算法.例如,为了根据拓扑结构进行图数据的分析计算,网页排名(Page Rank)、标记传播(Label Propagation)、随机行走(Random Walk)等融合传播和迭代更新机制的图模型被提出[17-18].为了对图数据进行建模推理,基于最大似然估计的图结构预测方法被提出[19].假设图数据中存在块状结构,随机分块模型(Stochastic Block Model)被提出[20].另外,非负矩阵分解、鲁棒性主成分分析、低秩稀疏等矩阵分解模型被应用于图数据的建模与预测[21-22].为了进一步实现图数据的建模学习,近年来以图神经网络(Graph Neural Networks,GNNs)[23]为代表的大量新型图模型算法被提出,图学习(Graph Learning)领域应运而生,各个图模型的详细介绍见文献[24-25].

基于丰富的数据资源和广泛的智能设备,各行各业为图学习系统提供了丰富的应用场景,同时也要求图模型算法能够适应现实场景中复杂的计算环境.然而,近年来的研究表明图学习模型具有极高的脆弱性,恶意攻击者可以通过改变图结构或节点属性影响图学习模型的性能[26-27]或者通过查询窃取图学习模型及其训练数据[28].从而,当前的图模型算法普遍缺乏与复杂计算环境相对应的理论基础和计算机制、鲁棒可信性不足,其现实应用的安全挑战十分突出.例如,在金融领域,攻击者可以通过攻击图模型使金融风险控制系统难以发现高风险的金融交易;在电子商务领域,攻击者可以通过伪造与高信用客户的链接关系攻击图模型算法从而提高信用评分系统对欺诈者的信用评分;在社交网络领域,攻击者可以通过创建假粉丝和假关注欺骗图模型算法以降低在线社交网络中垃圾信息发送者的可疑性.因此,面对图学习的广泛影响以及安全可信的迫切需求,图学习隐私与安全问题研究有着十分重要的现实意义.

图学习以图数据为对象,图数据具有非欧氏性、离散性、关联性等特征.同时,与图像处理和自然语言处理等领域不同,在图数据中的任何操作只能以节点和链路为单位.另外,与传统机器学习领域学习任务以归纳式学习(Inductive Learning)范式为主不同,图学习中典型学习任务主要以直推式学习(Transductive Learning)为主.因此,现有的针对图像处理、自然语言处理相关方法的攻击与防御技术无法直接适用于图学习模型,图学习系统面临的隐私与安全问题与传统机器学习领域的隐私与安全问题不同,具有其自身的独特性.

### 1.2 相关工作

文献[29]对图数据隐私保护研究进行了系统概括,介绍了图数据发布面临的隐私风险,对图数据匿名技术、去匿名技术以及去匿名量化方法进行了归纳,并基于数据效用量化指标测试了各种匿名化与去匿名技术的有效性.文献[30]重点分析了图数据隐私保护中基于图结构修改的隐私保护方法.文献[31]对图模型对抗攻击问题进行了综述,为图模型对抗攻击问题给出了形式化定义,介绍了近年来的相关研究工作和评价指标.文献[32]针对节点分类任务综述了图对抗攻击及其防御方法.文献[33]介绍了各种类型的图数据以及代表性的图学习任务,在此基础上阐述了对抗攻击及其防御的问题定义和代表性方法,并介绍了典型的评价指标.文献[34]概述了深度学习系统的安全与隐私问题,依次介绍了模型萃取攻击、逆向攻击、投毒攻击和逃逸攻击的主

要研究工作. 文献[35-36]综述了图像、文本、音频等不同领域机器学习安全与隐私研究进展，介绍了各种数据类型上对抗扰动的对象特性、限制条件和主要方法.

与以上综述中分别考虑数据和模型的隐私与安全问题不同，本文对学习过程中面临的隐私与安全风险进行整体的分析. 同时，本文主要聚焦图数据的利用问题，关注图学习面临的隐私与安全威胁.

### 1.3　本文贡献与章节组织

随着大数据、人工智能隐私与安全问题的逐渐暴露，公众的防护意识日益增强. 尽管近年来图学习的隐私与安全问题受到了研究领域的逐渐关注，但仍然面临诸多不足. 例如，虽然人工智能领域对学习模型的安全问题进行了研究，但当前存在的攻击与防御方法主要集中在图像处理领域，关于图学习的研究工作相对较少；图数据的隐私保护与图模型的隐私与安全紧密相关，然而相关研究工作对它们缺乏深入的整体性思考；图数据的非欧式性导致传统文本、图像等领域的理论方法无法直接适用于图模型，图学习的隐私与安全问题仍然缺乏系统性的研究体系. 本文的主要贡献可总结如下：

(1) 搜集整理了国内外期刊和会议上发表的相关文献，系统分析了最新的、特别是近 5 年关于图学习隐私与安全的研究成果；

(2) 整体考虑数据与模型的隐私、安全风险，提出了基于图学习系统基本过程的系统化图学习隐私与安全研究框架；

(3) 将已有的研究工作按照图学习系统面临的隐私与安全风险分为四类：面向图数据隐私的攻击与防护方法、面向图数据安全的攻击与防御方法、面向图模型隐私的攻击与防御方法以及面向图模型安全的攻击与防御方法，详细分析了各类方法的机理和执行过程，对典型算法的机制、优缺点、复杂度和适用场景进行了讨论.

(4) 总结了开源社区贡献的图学习隐私与安全的相关代码资源，分析了面临的研究难点和未来挑战，有助于后续研究人员更好地进行安全可信图模型算法的设计和验证.

本文第 2 节概述图学习隐私与安全基础框架；第 3 节至第 6 节分别阐述各类图学习隐私与安全方法，介绍代表性算法和相关研究进展；第 7 节梳理相关开放资源；第 8 节讨论图学习隐私与安全研究面临的难点及未来挑战；第 9 节总结全文.

## 2　图学习隐私与安全基础框架

### 2.1　图学习系统的基本过程

图学习隐私和安全问题与图学习系统的基本过程紧密相关. 为了清晰地进行图学习隐私与安全的分析研究，需要对图学习系统进行介绍，图学习系统的基本框架如图 1 所示. 一般而言，图学习系统主要包含以下过程.

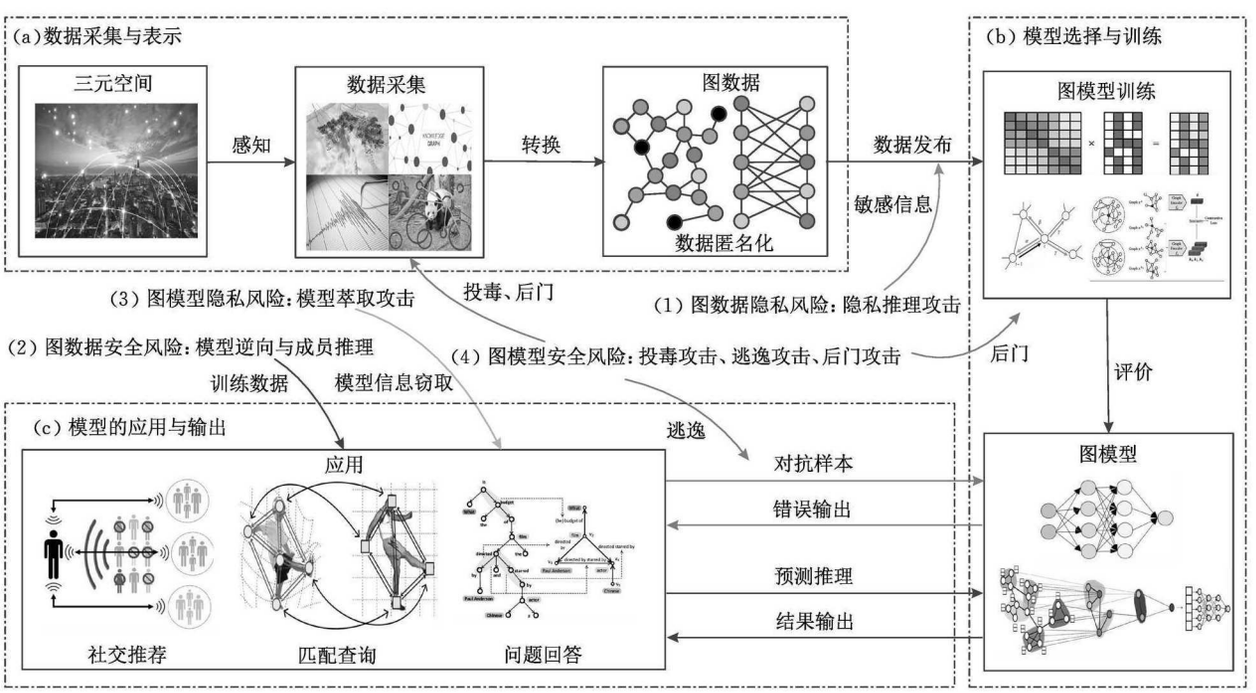



图 1　图学习系统隐私与安全风险

(1)数据采集与表示

根据目标对象的不同,图学习系统通过传感器、爬虫、实验设备、日志记录等方式获得原始数据.然后,分析采集到的数据中对象以及对象之间的潜在关联关系,并通过图进行表示.例如,文献[37]对新闻报道数据进行图表示,将报道国和被报道国定义为节点,将报道关系定义为有向边.文献[38]计算脑电图中电极序列之间的功能连接关系,从而获得脑网络的图表示.

(2)模型选择与训练

此阶段基于可用图数据对图模型进行训练以获得隐含的模式规律,从而实现图分类、节点标记、链路预测等任务.根据图模型的特点可以将其分为归纳式学习和直推式学习两大类.其中,归纳式学习是指训练数据与测试数据完全独立的学习过程,将学习好的模型应用于测试数据,常见于图分类等任务.直推式学习是指在训练阶段模型对训练数据和测试数据全部可见,将可用数据全部用于学习过程,常见于节点标记、链路预测、社团检测等任务.

(3)模型应用与输出

经过训练、测试的图模型即可被用于分类、预测等任务,从而通过开放服务接口对外提供推荐、查询、问答等智能服务.需要注意的是,对于直推式学习范式,模型的应用输出和学习训练是紧密融合的,不存在独立的模型应用与输出过程.例如,对基于矩阵分解模型的链路预测方法[21],在学习数据模式规律的同时进行潜在链路的推理预测.

### 2.2 图学习面临的隐私与安全威胁

根据图学习系统的基本过程,实际应用中数据采集与表示以及模型应用与输出两个环节直接与现实环境进行交互.在开放的应用环境中,无论是在图学习系统的数据层面还是在模型层面,都存在不愿意被未获授权的人员推理感知的敏感信息,例如个人身份、模型参数取值等.在图学习系统构建过程中,数据所有者或服务提供商一般都会对敏感数据进行隐私保护.然而,恶意攻击者可能基于公开信息和服务接口对被隐藏的敏感信息进行推理攻击,从而使图学习系统面临隐私威胁.另一方面,为了构建图学习系统,服务提供商需要花费资金、时间收集整理数据,并基于数据资源、算力进行模型训练,从而对外提供有价值的智能服务.然而,在模型部署应用过程中,恶意攻击者可能基于模型的输出窃取训练模型具体使用的数据从而破坏图学习系统对训练数据的所有权,或者通过恶意攻击改变模型的输出结果从而影响图学习系统的可用性,因此图学习系统面临安全威胁.根据具体攻击目标的不同,恶意攻击者对图学习系统相关数据中被隐藏的敏感信息进行推理攻击,使得图学习系统面临“图数据隐私风险”;恶意攻击者通过有限次数访问服务接口对图模型的参数、结构等敏感信息进行推理攻击,破坏模型的机密性,使得图学习系统面临“图模型隐私风险”;恶意攻击者通过模型输出窃取模型训练数据的相关信息,破坏训练数据的机密性,使得图学习系统面临“图数据安全风险”;恶意攻击者破坏图学习模型的完整性和可用性,导致系统无法对外提供正常服务,使图学习系统面临“图模型安全风险”.概括而言,开放环境下的图学习系统主要面临四个方面的隐私与安全风险,如图1所示.

(1)图数据隐私风险

隐私数据是指可以被用来确认特定个人或团体的身份或其特征,从而个人或团体不愿被暴露的敏感信息,例如位置信息、财务信息、医疗信息等.当对象的敏感信息被披露给没有得到授权的攻击者时,即发生了隐私泄露.为了保护数据隐私,数据拥有者往往会在数据发布之前通过匿名方法对敏感信息进行脱敏处理.但是,由于图数据内在的相关性和规律性,攻击者往往可以通过建模学习图数据中蕴含的关联关系推理重构匿名的敏感信息,从而实现“隐私推理攻击(Privacy Inference Attack)”.如图1中的风险(1)所示,恶意攻击者以公开发布的图数据或者窃取的图模型训练数据为输入,利用图数据蕴含的模式规律输出原始图的近似图数据、推理原始图数据中包含的敏感信息,从而造成隐私泄露.

(2)图数据安全风险

由于知识产权、经济成本以及数据的不可或缺性等原因,数据成为构建智能计算系统的重要资源.服务提供商通过收集公开数据、数据交易等方式获得数据资源,在此基础上精炼数据形成训练数据集以构建图学习系统.因此,图学习系统的训练数据对外界而言是不可共享的.但是,面对开放环境下的图学习系统,攻击者可以利用开放服务接口,以模型的输入样本及其对应的输出结果作为攻击行为的输入数据、进行“模型逆向攻击(Model Inversion Attack)”和“成员推断攻击(Membership-Inference Attack)”,窃取模型训练数据、获得训练数据中的成员关系等信息,从而破坏数据的安全性,如图1中的风险(2)所示.

(3) 图模型隐私风险

图模型是服务提供商基于大量数据资源、计算资源及时间精力训练得到的，是智能时代的核心竞争力. 在开放环境下，攻击者在模型部署之后通过多次查询产生对应的结果，基于以上信息进行“模型萃取攻击(Model Extraction Attack)”，获得模型架构、参数等信息以支撑自有模型的构建或者对目标系统实施其它攻击，从而在不影响目标模型正常功能的情况下造成图模型敏感信息的泄露、破坏图模型的隐私性，如图 1 中的风险(3)所示.

(4) 图模型安全风险

在原始数据采集或者图数据发布阶段，攻击者以模型训练相关数据作为输入、通过将精心制作的样本插入到数据集中来操纵模型训练数据的分布，从而输出投毒之后的数据以达到改变模型行为和降低模型性能的目的，即“投毒攻击(Poisoning Attack)”. 同时，对于已经部署应用的图模型，攻击者可能进行“逃逸攻击(Evasion Attack)”，通过对模型的输入数据进行轻微、不易察觉的扰动，产生对抗样本、导致错误输出. 另外，攻击者可以在图学习系统的各个阶段进行“后门攻击(Backdoor Attack)”，以训练数据或模型为输入，通过设置触发器使模型输出攻击者预先设定的结果. 以上攻击都会直接改变模型的正常功能和服务，影响模型的安全性，如图 1 中的风险(4)所示.

图数据隐私保护与图学习紧密相关，是实现图学习系统的基本条件和前提保障，也与图数据安全、图模型安全相互影响，是图学习系统隐私与安全体系的重要组成部分. 具体的，(1) 根据图学习系统的基本过程，图数据的采集与匿名处理是图模型训练的基础，图数据的隐私推理攻击会直接影响图数据资源的共享和交易，阻碍对图学习系统训练资源的获取. 同时，图数据的隐私推理攻击往往需要利用图模型对图数据进行建模学习，高水平的图模型会增加图数据隐私泄露的风险；(2) 对图模型训练数据的窃取会增加图数据隐私泄露的风险. 实际图学习系统中，对图模型的逆向和成员推理攻击会导致隐私攻击者获得更多的数据资源和背景知识，并基于此推理相关敏感信息，增加隐私泄露风险；(3) 图数据隐私与图模型安全相互影响，通过图模型的对抗攻击可以实现图数据的隐私保护[39-40]. 在隐私推理攻击中，攻击算法的性能会直接影响隐私泄露的风险，可以通过对隐私推理攻击算法进行投毒攻击，即在数据匿名处理的基础上进行轻微扰动，降低敏感信息泄露的可能性；(4) 基于图数据推理重构可以强化图模型的安全水平. 投毒攻击、逃逸攻击会影响图数据的模式规律，而图数据推理预测算法可以发现与训练数据、输入数据整体模式规律不一致的扰动信息，从而帮助清除对抗扰动，是实现图模型安全防护的重要途径[41-43].

### 2.3　图学习的隐私与安全特性

(1) 图数据模式规律性与隐私保护

图数据挖掘导致的隐私安全风险越来越高. 大量图数据挖掘与隐私保护的研究工作表明，图数据隐私保护与隐私推理攻击的性能除了受具体算法影响外，还与图数据自身的模式规律紧密相关. 然而，虽然当前已有统计量和演化机制对图数据的模式规律进行刻画[44-45]，但是这些方法主要是定性的、宏观的描述和探索，图数据的建模分析技术仍不够成熟. 实际上，图数据的模式规律及其显著性是影响图数据挖掘方法性能的根本因素. 真实图数据往往包含规律性和非规律性部分，其中规律性的图数据可以通过理论建模进行解释刻画，规律性程度基本决定了其可建模学习的程度[46]. 如图 2 所示，图的结构规律性(Structure Regularity)依次逐渐增加，规律性越强的图数据越容易被建模学习和推理预测. 从而，图数据隐私保护方法只要没有明显改变图数据中敏感信息相关的模式规律，敏感信息就可能因为图数据的规律性而被推理发现.

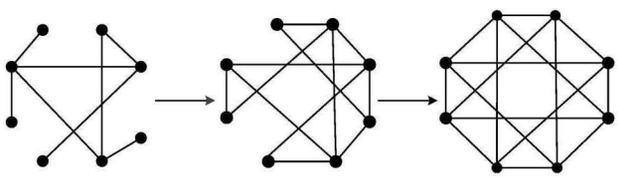


图 2　具有不同规律性的图拓扑结构

(2) 图数据特征与模型隐私安全

关于图学习的隐私与安全问题，图学习系统具有不同于传统机器学习的自身特征：① 图数据的非欧氏性. 传统图像和文本数据排列整齐，每个样本可以被视为欧氏空间中的一个点. 然而，图数据具有非欧氏性，各个节点的邻居数量各异，且节点之间不具有确定的排列顺序，需要不同于传统学习模型的计算机制；② 图数据的离散性特征. 不同于图像等具有连续取值的数据，图数据中节点之间只有连接和非连接两种情况，图数据表示是离散的. 在离散域中，很难设计出有效的攻击算法. 虽然研究领域已经

对传统机器学习模型的安全问题进行了研究，但提出的攻击与防御方法难以直接适用于图学习模型；③图数据的关联性特征. 图中的节点、链路是相互关联的，即图数据不是独立同分布的，任何节点和链路的改变都会影响其它元素的计算. 因此，攻击方法更容易对图模型产生影响；④图数据的异构性特征. 图数据是不规则的，各个节点具有不同的度、在图中具有不同的角色和作用，各个子图也以同配或异配的方式呈现出不同的模式特征，需要针对此特征设计相应的计算机制；⑤直推式学习. 在图学习领域，节点标记、链路预测、重要性排序等任务都属于直推式学习，它们将全部可用数据用于模型训练，因此对测试数据的修改也会影响模型的训练. 鉴于以上特性，需要研究针对图学习系统的隐私与安全防护方法.

(3) 图学习对抗扰动的感知与度量

在传统机器学习中，图像处理领域以像素为单位进行对抗扰动，要求对抗攻击向图像中添加的噪音是人眼无法察觉的. 自然语言处理领域以字、词为扰动单位，要求对抗攻击产生的单词或句子是有效的. 由于图数据的自身特性，关于图模型的对抗扰动必须以节点和链路为单位. 同时，传统的人眼不可感知的计算机视觉模型对抗扰动度量、语法语义不可感知的自然语言处理模型对抗扰动度量都无法直接应用于图学习模型对抗攻击. 在图学习中，其扰动度量往往基于图结构特征进行描述. 典型的图学习对抗攻击扰动度量包括节点度分布(Node Degree Distribution)[26]、基于扰动链路数和节点数的攻击预算(Attacker Budget)[40,47]、特征路径长度(Characteristic Path Length)[47]、三角形计数(Triangle Count)[47]以及聚类系数(Clustering Coefficient)[48]等.

## 3 面向图数据隐私的攻击与防御方法

隐私保护的基本思想是在尽可能保持原始数据不变的条件下对敏感信息进行隐藏，使得攻击者无法通过背景知识对被隐藏的敏感信息进行重识别攻击(Re-identification Attack). 图数据隐私保护与节点、属性和结构紧密相关，本节主要介绍代表性的图数据隐私攻击行为和近年来提出的经典的图数据隐私保护方法.

### 3.1 图数据隐私推理攻击方法

图数据隐私推理攻击是指在匿名化的图数据中重新识别发现敏感信息的过程. 根据攻击对象的不同，图数据隐私推理攻击可以概括为节点身份重识别攻击、属性推理攻击和链路推理攻击三类.

**定义 1.** 图数据隐私推理. 给定原始图 $G=(V,E)$及其对应的匿名图 $\hat{G}=(\hat{V},\hat{E})$，根据关于目标节点 $t\in V$ 的背景知识 $B(t)$，隐私推理攻击的目标是基于匿名图 $\hat{G}$ 推理关于节点 $t$ 的有用信息. 如果式(1)成立，则认为成功对节点 $t$ 进行了隐私推理.

$$\exists \bar{u}\in \hat{V}(\hat{G}):\{B(\bar{u})=B(t)\} \tag{1}$$

其中 $\bar{u}$ 为攻击者从匿名图节点集合 $\hat{V}$ 中发现的与 $t$ 相对应的节点，$B(\bar{u})$是它的特征属性.

(1) 节点身份重识别攻击

节点身份重识别攻击主要分为基于属性的方法和基于结构的方法[49]. 基于属性的节点身份重识别攻击从公开数据中提取用户住址、时间戳、地理标签等特征，然后结合分类器来推断节点身份. 假设 $Q=\{(u^s,u^t),u^s\in V^s,u^t\in V^t\}$是来自两个图的所有用户身份对，其中 $M\subset Q$ 表示正实例，即用户 $u^s$ 和 $u^t$ 属于同一个人，$N=Q-M$ 表示负实例，节点身份重识别的目标是基于已知数据训练分类模型 $F:V^s\times V^t\rightarrow\{0,1\}$，从而利用此模型对未知身份的节点进行匹配识别. 以集成学习方法为例，此识别模型可以被写为式(2)，

$$F(\boldsymbol{x})=sign\Big(\sum_{i=1}^{T}a_ih_i(\boldsymbol{x})\Big) \tag{2}$$

其中 $\boldsymbol{x}$ 为特征向量，$h_i$ 为弱分类器，$a_i$ 为权重系数. 文献[50]考虑行为时间分布特征、内容特征、地理特征和社会关系特征进行节点身份重识别. 文献[51]在文献[50]的基础上考虑了内容风格、轨迹等更加全面的属性特征. 此类方法复杂度较低，可根据可用属性灵活构建，适用范围广.

基于结构的节点身份重识别攻击假设同一个人在不同图中具有相似的局部结构，使用与目标节点相关联的子图作为标识用户的背景知识以实现身份识别. 文献[52]将已知节点身份的图 $N_1$ 和未知节点身份的图 $N_2$ 进行匹配. 首先通过拓扑分析识别社团结构，在此基础上进行图之间的社团匹配(Community Mapping). 然后在匹配的社团中进行节点匹配(Node Mapping)，识别个别节点的真实身份. 最后，以识别到的节点作为种子节点(Seed Nodes)通过全局传播发现剩余节点的身份，如图 3 所示. 基于种子的节点身份重识别攻击方法假设攻击者拥有关于种子节点的先验知识，而且种子节点的不准确性会直接影响

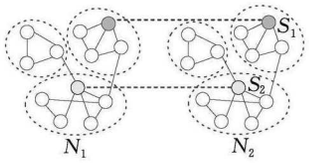



图 3　基于局部结构的节点身份重识别攻击

节点身份重识别的性能，因此文献[53]提出了无需种子节点的、基于结构特征的迭代式节点身份重识别攻击方法.

总体上，基于种子的身份重识别方法的性能依赖于种子节点的准确性，无需种子节点的身份重识别方法的性能依赖于结构特征的有效性. 相对而言，由于无需先验知识，后者比前者适用范围更广，但特征构建和迭代式匹配计算导致后者复杂度更高.

(2) 属性推理攻击

属性推理攻击是指攻击者利用背景知识推理隐藏的敏感属性的过程. 与节点身份重识别攻击类似，此类攻击主要利用用户行为和拓扑结构与敏感属性之间的潜在关联关系进行实现.

文献[54]以社交网络为背景提出整合社交关系、用户行为、用户属性的统一网络 SBA，如图 4 所示，在此基础上给出投票分配攻击方法 VIAL. VIAL 主要包含投票分配和聚合两个阶段，其迭代式地将投票从目标用户节点分配给其余用户节点，在 SBA 网络上与目标用户拥有越多共同邻居和行为的用户节点接受到的总投票量越大. 然后，每个节点将收到的投票量分配给与其相邻的属性节点，最终节点拥有获得最高投票量的属性.

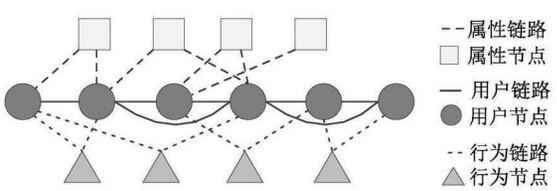



图 4　社交用户-属性-行为网络

(3) 链路推理攻击

链路推理攻击，即链路重识别攻击(Link Re-identification Attack)，目的是通过建模学习匿名化图数据推理发现隐藏在节点之间的敏感关系. 基于低秩近似的推理重构[55]是其中的代表性方法，给定原始图邻接矩阵 $\boldsymbol{A}$ 的特征值 $\lambda_i$ 和特征向量 $\boldsymbol{x}_i$，它通过特征分解近似原始图，如式(3)所示，

$$\boldsymbol{A}_r=\sum_{i=1}^{r}\lambda_i\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} \tag{3}$$

在所有秩不大于 $r$ 的矩阵中，低秩近似矩阵 $\boldsymbol{A}_r$ 最接近原始图邻接矩阵 $\boldsymbol{A}$，即

$$\|\boldsymbol{A}_r-\boldsymbol{A}\|_F^2=\min_{rank(\boldsymbol{B})<r}\|\boldsymbol{B}-\boldsymbol{A}\|_F^2 \tag{4}$$

文献[56]发现图数据隐私保护方法添加的扰动链路由于具有较低的结构相似性(Structural Proximity)而容易被识别发现，从而提出基于嵌入表示的链路可信性(Edge Plausibility)评价指标，然后结合高斯混合模型提出图恢复(Graph Recovery)算法. 此方法的性能与节点嵌入表示的准确性紧密相关. 考虑图结构模式的复杂性，文献[57]进一步提出了基于深度线性编码的敏感链路推理攻击方法. 它结合低秩稀疏理论和深度模型的优势，将原始图结构近似表示为式(5)，

$$\boldsymbol{S}=\boldsymbol{A}\left(\prod_{\ell=1}^{n}\boldsymbol{Z}_\ell^*+\prod_{\ell=1}^{n-1}\boldsymbol{Z}_\ell^*+\cdots+\boldsymbol{Z}_1^*\right) \tag{5}$$

其中 $\boldsymbol{A}$ 为匿名图的邻接矩阵，$n$ 为模型深度.

此类敏感链路推理方法依赖于对图结构模式进行准确的建模学习，这需要构建刻画能力强的算法模型，复杂度较高，在大规模图上运行时间较长.

### 3.2　图数据隐私保护方法

图数据隐私保护可以被概括为非结构扰动的隐私保护方法和基于结构扰动的隐私保护方法. 非结构扰动的隐私保护主要为朴素匿名[58-59]，这类方法仅仅通过对敏感信息进行替换、消除以达到隐私保护的目的；基于结构扰动的隐私保护方法有数据扰动、$k$ 匿名、聚类方法、差分隐私等，这类机制对图结构进行扰动修改从而实现隐私保护. 一般而言，图数据隐私保护应该考虑敏感身份、敏感属性以及敏感结构等信息. 然而，研究发现对敏感结构的推理攻击往往会导致敏感身份和敏感属性的泄露[60-61]，从而敏感结构成为图数据隐私保护的主要目标. 因此，本节关注基于结构扰动的隐私保护方法.

(1) 数据扰动

为了通过数据扰动实现隐私保护，直接方法是在朴素匿名之后对图结构进行随机扰动，包括稀疏化扰动、随机扰动和随机交换. 其中，稀疏扰动方法[62]对图中的每条链路执行一次伯努利实验，根据实验结果进行链路的删除或保持，从而最终在尽可能保持图结构信息的同时对敏感链路推理进行干扰；随机扰动方法以某一概率删除图中的链路，然后随机选择同样数量不相连的链路进行添加，保持图中的总链路数不变；随机交换方法从图中随机选择两条链路进行删除，并添加交叉连接各自端节点的新链路到图中，重复多次[63]. 另外，研究发现敏

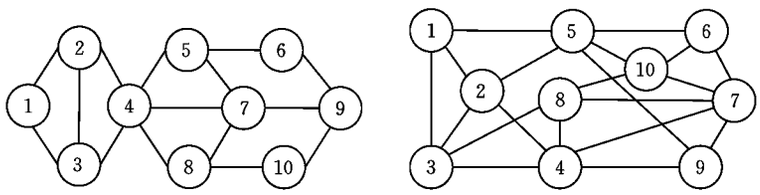
感链路推理虽然与全局拓扑相关,但局部结构往往发挥着更大作用.因此,研究人员提出了基于随机行走的结构扰动方法[64],同时通过使随机扰动保持在步长为 $t$ 的范围内,在实现隐私保护的同时尽可能保持了数据效用.另外,为了在实现隐私保护的同时保持数据效用,文献[65]提出了谱特征保持的结构扰动方法.

基于数据扰动的图数据隐私保护方法复杂度低,但其隐私保护水平受限于结构扰动程度.同时,此类方法中扰动链路的选择具有随机性,未考虑图数据固有的结构模式,被推理重构的可能性较高.数据扰动方法无法提供量化的隐私保护水平,匿名图仍然存在隐私泄露的风险.

(2) $k$ 匿名

$k$ 匿名保护是指通过数据操作使得任意一项信息所属的相等集内数据项的数量不小于 $k$,即对于每一项信息,数据集中都存在其它 $k-1$ 项与其无法区分的信息.$k$ 值越大,目标个体身份被攻击者识别的概率就越小,抵御攻击的能力就越强.

$k$ 度匿名是最早的图数据 $k$ 匿名保护方法.如果一个图满足 $k$ 度匿名,则表明图中任何一个节点至少与其它 $k-1$ 个节点具有相同的度.如图 5(a)是一个没有进行度匿名的原始图,度序列为[2,3,3,5,3,2,4,3,3,2],给定度序列和相应的节点 ID 序列.由于只有节点 4 具有度 5,而只有节点 7 具有度 4,所以任何人都可以重新识别出节点 4 和节点 7.通过 3-度匿名将图 5(a)修改为图 5(b),度序列为[3,4,4,5,5,3,5,4,3,4],这时图中任何一个节点都至少有其它 2 个节点与其度数相同,从而重新识别节点 4 或节点 7 的概率变为 1/3.




(a)原始图 (b)3度匿名图

图 5 $k$ 度匿名

文献[66]研究图数据中基于节点度的攻击问题,提出了 $k$-DA($k$-Degree Anonymity)方法.$k$-DA 首先在图中节点度序列的基础上通过贪婪策略建立一个新的 $k$ 匿名度序列,然后根据这个新的序列构建匿名化的图,其中每个节点都有至少 $k-1$ 个跟它有着相同度的节点.文献[67]提出一种基于节点平均度的 $k$ 度匿名隐私保护方案.此方案首先利用基于平均度的贪心算法对节点进行划分,使得同一分组中节点的度都修改为平均度,从而生成 $k$ 度匿名序列.然后利用优先保留重要边的图结构修改方法对图进行修改,从而实现图的 $k$ 度匿名.

由于攻击者能够利用节点的局部结构特征作为背景知识,$k$ 度匿名的隐私保护能力较差.为了进一步提升隐私保护水平,基于子图的 $k$ 匿名方法被提出.文献[68]将朴素匿名处理后的图划分为包含 $k$ 个块的多个分组,并在块与块之间进行匹配对齐.然后,对于没有对齐的结构进行链路复制,从而使每个分组中具有 $k$ 个相同的结构.虽然存在的 $k$ 匿名方法能够保护敏感节点信息不被重识别攻击,但无法保护节点之间的敏感链路信息.为此,文献[69]提出了基于 $k$-Isomorphism 的算法以保证敏感链路的每个相关节点都有 $k$ 个候选节点,从而使攻击者无法识别具体的敏感链路.

此外,研究领域针对复杂图数据匿名问题也进行了研究,文献[70]提出了面向加权图的 $k$ 匿名方法.考虑到链路方向也会造成敏感信息的泄露,文献[71]提出了面向有向图的 $k$ 度匿名方法.为避免使用不同时刻的图数据进行关联攻击,文献[72]提出一种基于逆向更新的动态图数据隐私保护方法.总体上,$k$ 匿名方法能够提供确定的隐私保护水平,对敏感信息的保护能力强,但此类方法执行效率低,不适合用于大规模图数据中.

(3)聚类方法

为了抵御基于结构和属性的隐私攻击,研究领域提出了基于聚类的匿名方法,该方法基于图结构和属性对图进行聚类划分,然后根据划分结果将图表示为超级节点(Super-Nodes)和超级边(Super-Edges).基于聚类方法的图匿名过程如图 6 所示,首先对图中的节点基于距离度量聚类为不同的簇,使得不同簇的节点之间距离尽可能的大,同一簇中节点之间的距离尽可能的小,如中间的子图所示.然后,在此基础上将同一簇的节点融合为超级节点,将簇与簇之间的边融合为超级边,从而实现对敏感信息的隐藏.

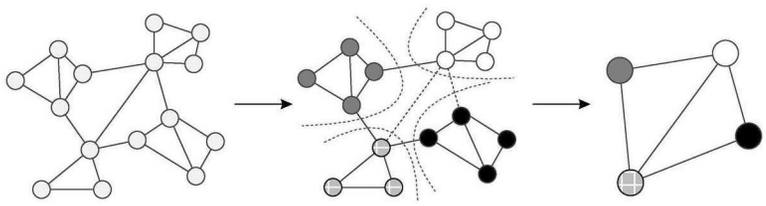


图 6　基于聚类匿名的图数据隐私保护

为了实现图的聚类匿名，文献[73]首先定义信息损失函数，然后基于此函数贪婪式的对没有类别的节点进行划分，最终基于划分结果进行泛化和匿名. 文献[74]提出基于随机块模型的聚类方法. 为了保证隐私保护水平，此方法将图结构分割成多个区块，要求每个区块至少有 $k$ 个节点. 经过匿名化算法后输出的是一个包含超级节点和超级边的泛化图，泛化图保留着原图的关键属性，对任何查询都满足 $k$ 匿名. 文献[75]综合结构信息和属性信息来计算节点之间的距离，然后基于 $k$-means 算法进行聚类，最终在降低信息损失的同时实现聚类匿名. 文献[76]定义了节点间的结构相似度和属性相似度，然后利用贪心方法实现属性图的聚类划分和匿名处理.

聚类匿名方法将图中所有结点聚类融合成若干超级节点，其中每个超级节点至少包含 $k$ 个结点，从而隐私保护能力强，具有广泛的适用性，可以防止多种类型的隐私攻击. 然而，聚类匿名方法中的融合操作需要合并节点和边，对数据效用的影响较大.

(4) 差分隐私

作为一种严格的和可证明的隐私定义，差分隐私近年来受到了极大关注并被广泛研究[77-78]. 差分隐私的主要思想是任何随机算法得到的输出难以区分仅相差一条记录的“相邻”数据集，从而实现隐私保护. 设有随机算法 $M$ 及其可能输出构成的子集 $S_M$，对于相邻数据集 $D$ 和 $D'$，如果满足式(6)，

$$P_r[M(D)\in S_M]\leq \exp(\varepsilon)\cdot P_r[M(D')\in S_M] \tag{6}$$

则称算法 $M$ 提供了 $\varepsilon$-差分隐私保护. 其中，$\varepsilon$ 为隐私保护的预算，exp()为指数函数，$P_r$ 为概率. 一般 $\varepsilon$ 取很小的值. 如果 $\varepsilon$ 等于 0，则表明在相邻数据集 $D$ 和 $D'$ 上算法 $M$ 输出两个概率分布完全相同的结果，从而隐私保护水平最高. 差分隐私匿名的基本过程如图 7 所示，潜在的恶意攻击者查询相邻数据集 $D$ 和 $D'$. 为了保护敏感数据记录，差分隐私方法对获得的精确查询结果添加噪声，使得在相邻数据集上尽可能输出相同的结果. 实质上，差分隐私就是要保证数据集中存在或不存在某一数据项时，对最终发布的查询结果几乎没有影响，最常用的噪声产生方式是采用 Laplace 机制.

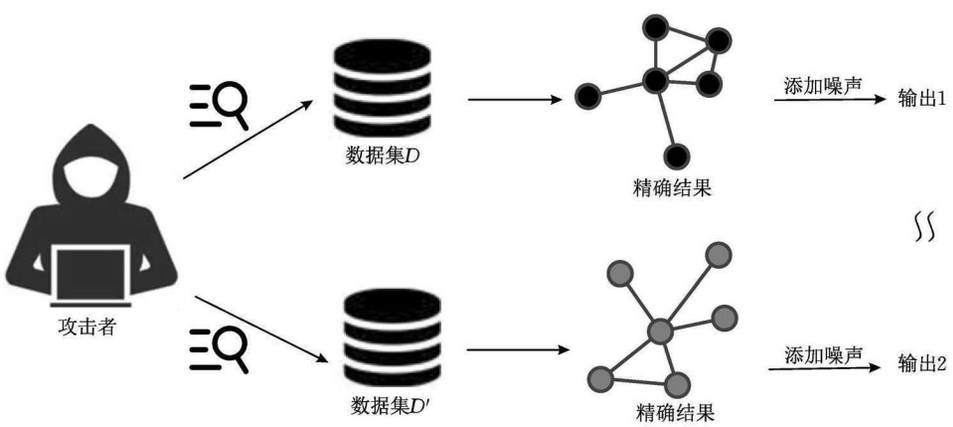



图 7　差分隐私保护示意图

为了进行图数据的隐私保护，文献[79]提出了基于子图的差分隐私保护方法. 该方法先将图划分为多个子图，然后在每个子图中利用四叉树进行区域划分，并在树的叶子节点添加满足差分隐私保护的噪声. 文献[80]提出了社团结构保持的图数据差分隐私模型. 该模型在社团检测的基础上，对社团内部的链路添加 Laplace 噪声，然后基于社团内节点的度信息生成社团之间的链路. 文献[81]提出了基于 dK 序列的图数据差分隐私保护方法. 该方法首先以子图为单位将图表示为关于节点度的统计数据，即 dK 序列，然后提出基于 Laplace 函数的扰动算法 dK-PA. 该算法产生噪声并注入生成的 dK 序

列中使其满足 ε-差分隐私. 另外,文献[82]发现通过估计节点之间的连接概率而不是直接考虑观察到的边可以大大降低差分隐私添加的噪声规模,为此提出利用层次随机图进行结构推理,并采用马尔可夫链蒙特卡罗方法采样可能的层次随机图以实现差分隐私. 差分隐私基于数学理论为隐私保护进行了严格的定义并提供了量化评估方法,但相关算法普遍复杂度较高、面临大规模数据隐私保护的挑战.

表 1 是对图数据隐私保护方法的一个小结. 数据隐私保护问题近二十年来得到了研究领域较为充分的探索,图数据隐私保护也得到了较多的关注,其中数据扰动方法简单易实现,但发生隐私泄露的风险较高. $k$ 匿名、聚类方法以及差分隐私方法能够提供更高水平的隐私保护,但聚类方法对数据效用的影响较大,$k$ 匿名和差分隐私方法能够提供确定水平的隐私保护,但它们复杂度较高. 总体上,由于图数据隐私保护场景的复杂性、背景知识的不可控性以及要求保持数据效用等原因,仍然还有待提出效率更高、适用性更好的图数据隐私保护方法.

**表 1 图数据隐私保护方法小结**

| 类型 | 操作 | 优点 | 缺点 | 算法 | 描述 |
|---|---|---|---|---|---|
| 数据扰动 | 添加、删除边 | 复杂度低,易实现 | 隐私保护水平受限于结构扰动程度,无法提供量化的隐私保护水平,被攻击的风险较高 | 稀疏扰动方法[62] | 对每条边独立执行伯努利实验 |
| | | | | 随机交换方法[63] | 随机删除两条边,并添加交叉连接各自端节点的新边 |
| | | | | 随机行走方法[64] | 通过 $t$ 步随机行走添加边 |
| | | | | 谱特征随机扰动[65] | 根据特征值的变化选择添加、删除和交换的边 |
| $k$ 匿名 | 添加、删除边和节点 | 能提供确定水平的隐私保护 | 执行效率低,不适用于大规模图数据 | $k$-DA[66] | 基于贪婪策略生成 $k$ 匿名度序列 |
| | | | | 平均度 $k$ 度匿名[67] | 贪心算法及优先保留重要边的结构修改 |
| | | | | $k$-automorphism[68] | 在图划分的基础上进行结构对齐和边复制 |
| | | | | $k$-isomorphism[69] | 保证敏感链路的每个相关节点都有 $k$ 个候选节点 |
| | | | | $k$-weighted[70] | 面向加权图的 $k$ 匿名方法 |
| | | | | $k$-directed[71] | 面向有向图的 $k$ 度匿名方法 |
| | | | | $k$-dynamic[72] | 基于逆向更新的动态图 $k$ 匿名方法 |
| 聚类方法 | 融合边和节点 | 保护能力强,适用性广,可防多种攻击 | 对数据效用的影响大 | 贪婪划分[73] | 结合损失函数和贪婪策略进行图划分 |
| | | | | 随机块聚类方法[74] | 基于随机块模型的图结构分割与泛化 |
| | | | | $k$-means 聚类方法[75] | 结合机构、属性距离和 $k$-means 算法的聚类划分 |
| | | | | 贪心聚类方法[76] | 结合结构、属性相似度和贪心法的聚类划分 |
| 差分隐私 | 添加边和节点 | 具有理论基础,能提供确定水平的隐私保护 | 算法复杂度高、面临大规模数据隐私保护的挑战 | 子图差分隐私[79] | 基于四叉树的区域划分和噪声添加 |
| | | | | 社团差分隐私[80] | 在社团检测的基础上添加 Laplace 噪声 |
| | | | | dK 序列差分隐私[81] | 基于 Laplace 函数产生噪声对生成的 dK 序列进行扰动 |
| | | | | 连接概率差分隐私[82] | 基于马尔科夫链蒙特卡罗方法采用层次随机图 |

## 4 面向图数据安全的攻击与防御方法

图数据安全关注图模型训练数据的窃取与泄露风险. 特别是对于提供开放服务的图学习系统,恶意攻击者可能基于系统的输出结果窃取模型的训练数据或者确定训练数据集中的成员信息,从而暴露敏感信息、侵害服务提供商的合法权益.

### 4.1 图数据窃取攻击的基本概念

在模型训练过程中,优化求解算法通过最优化参数取值以拟合数据的分布特征,充分拟合使得模型记录了训练数据的细粒度信息,从而基于模型输出可以获得训练数据的特征线索. 在此基础上,根据目标模型对于训练数据的拟合程度,攻击者能够通过访问目标模型窃取训练数据的相关信息,此过程称为模型逆向攻击(Model Inversion Attack)和成员推理攻击(Membership Inference Attack). 具体地:

(1) 模型逆向攻击. 模型逆向攻击的目标在于基于模型输出、背景知识等信息尽可能地获得模型的原始训练数据. 例如,对于训练好的、对外提供智能服务的图节点分类模型,攻击者基于节点特征、节点标记等背景知识以及模型的输出结果可推理重构原始图数据,使推理获得的图数据尽可能地与原始图数据相一致[83],具体如图 8 所示.

(2) 成员推理攻击. 对于“某一输入记录”和“目标模型(Target Model)”,成员推理攻击的任务是构建“攻击模型(Attack Model)”从而能够区分目标模型在训练数据集和非训练数据集上的不同输出,以便确定此输入记录是否包含在目标模型的训练数据集中[84-85]. 因此,成员推理攻击的实质是一个分类问题. 为了构建攻击模型,需要创建多个“影子模型(Shadow Model)”以模仿目标模型的行为,影子模

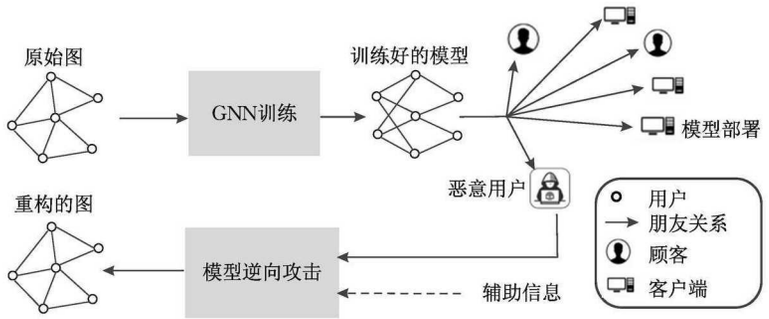



图 8　图模型逆向攻击示意图[83]

型中各个记录是否属于其训练数据集是已知的，基于影子模型的“输入记录”和“该记录是否属于影子模型训练数据集”训练攻击模型，从而使攻击模型具有分类能力.

对于任何恶意攻击，良好的攻击效果都依赖于对目标图模型的充分了解. 只有尽可能地获得目标图模型的类型、架构、参数以及求解方法等信息，才能设计出有效的攻击方法. 根据攻击者对目标模型了解程度的不同，可以将攻击分为不同类型.

(1) 白盒攻击. 在白盒攻击(White-box Attack)中，攻击者拥有关于目标模型的全部信息，包括模型类型、模型架构、模型参数等. 在此情况下，攻击者完全根据目标模型的真实特征设计攻击方法，较容易实现有效攻击. 然而，在实际应用中往往难以获得关于目标模型的详细信息，从而此类攻击方法可行性较差.

(2) 灰盒攻击. 相对于白盒攻击中攻击者拥有目标图模型的全部信息，在灰盒攻击(Gray-box Attack)中攻击者只有关于目标模型的部分信息，可能是模型架构、模型参数或者训练数据中的某一类. 设计有效的灰盒攻击方法的难度较大，但灰盒攻击可行性更强、更加危险.

(3) 黑盒攻击. 在黑盒攻击(Black-box Attack)中，攻击者无法获得关于模型训练数据和模型内部特征的任何信息，只能通过构造输入数据获得模型的响应输出. 由于不依赖于目标模型的具体信息，黑盒攻击方法适用于任何模型，具有很强的危险性.

### 4.2　图模型逆向与成员推理攻击方法

模型逆向攻击表明攻击者可以通过模型结果反向推理模型的训练数据. 为了评估模型训练数据的安全风险、提升对智能服务的安全保障水平，近年来模型逆向攻击受到研究人员的广泛关注[86-87]. 然而，由于图数据拓扑结构和特征属性的复杂性，对图模型进行逆向攻击更具挑战性，当前图学习领域的模型逆向攻击研究仍处于萌芽阶段.

在图模型逆向攻击方面，文献[89]针对图节点分类任务提出了 GraphMI 方法. 给定训练好的 GNN 模型和节点标记、属性等背景知识，GraphMI 试图逆向重构训练数据集中节点之间的链路. 为此，考虑标记一致性、特征光滑性以及结构稀疏性等因素，GraphMI 中的投影梯度下降模块通过最小化节点预测标记与真实标记之间的距离求解接近原始图结构的邻接矩阵，其目标函数定义如式(7)所示，

$$\arg\min_{\boldsymbol{A}\in\{0,1\}^{N\times N}} \mathcal{L}_{\text{attack}}=\mathcal{L}_{\text{GNN}}+tr(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{L}}\boldsymbol{X})+\beta\|\boldsymbol{A}\|_F \tag{7}$$

然后，基于图自编码模型(Graph Auto-Encoder)编码节点特征、邻居向量等信息获得节点的嵌入表示，最终通过解码推理图的拓扑结构.

在图模型成员推理攻击方面，给定目标模型 $\mathcal{M}$ 及其训练图数据 $G_t=(V_t,E_t)$，文献[88]基于节点 $v$ 及其邻居结构推理节点 $v$ 是否属于 $V_t$，该方法包含影子模型训练、攻击模型训练和成员关系推理三个阶段. 首先，该方法基于与目标模型训练数据具有相同分布的输入数据及其在目标模型上的预测结果构造影子模型. 然后，基于输入数据是否属于影子模型的训练集合及其在影子模型上的预测结果训练二分类攻击模型. 最后，基于某一节点在目标模型上的预测结果，通过攻击模型判断此节点是否属于目标模型的训练数据. 为了进一步分析节点特征和子图结构对成员关系的影响，基于与文献[88]类似框架，文献[89]分别以节点特征、邻居子图作为背景知识，并基于此背景知识和影子模型构建成员、非成员数据集，从而利用此数据集训练二分类攻击模型，其流程如图 9 所示.

与以上节点成员关系推理不同，文献[90]以目标数据集的节点属性、部分图结构以及辅助数据集作为背景知识，基于目标模型输出进行链路成员关系推理攻击. 此方法假设如果两个节点共享更多相似属性和目标图模型的预测结果，那么它们更有可能被连接，为此针对不同背景知识提出了无监督攻

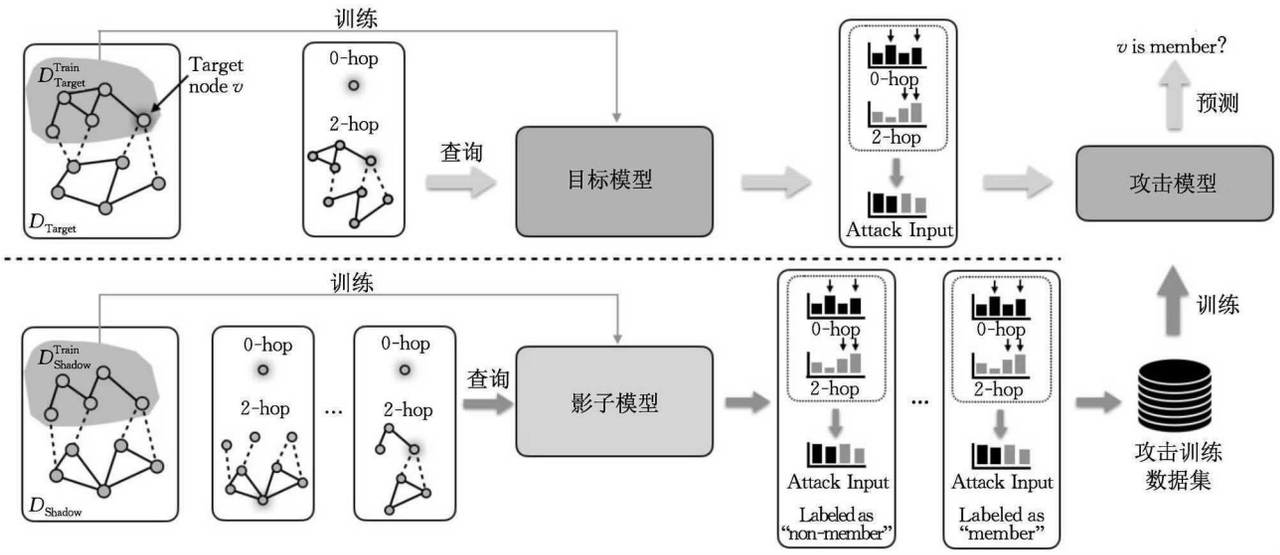



图 9 节点成员关系推理攻击[83]

击、二元分类器、迁移攻击等方法.

### 4.3 图模型逆向与成员推理攻击防御方法

关于模型逆向与成员推理攻击的防御,在传统机器学习领域常采用差分隐私、随机失活(Dropout)和模型集成(Model Stacking)等方法进行实现.为了防御通过图模型逆向攻击窃取模型的原始训练数据,文献[77]针对模型的每轮训练过程,探索了在裁剪梯度上增加高斯噪声的差分隐私(Differential Privacy)防御机制的有效性.另外,为了防御成员推理攻击,文献[83]提出了在目标模型的训练数据集中随机添加链路(Random Edge Addition)、目标模型仅输出预测标签(Label-Only Output)两种防御方法,在多个图神经网络模型上的测试结果表明这两种防御方法能够有效降低成员推理攻击的准确率,但它们会损失目标模型的实用性.

表 2 是对图模型逆向与成员推理攻击及其防御方法的一个小结.总体上,当前关于图模型逆向攻击的研究工作仍比较少,还没有形成较为系统的研究成果.存在的少数工作虽然对图模型脆弱性进行了分析,但还缺乏深入研究,对影响模型逆向攻击成功率的因素还缺乏认识.在图模型逆向攻击的防御方面,当前仅有的研究只是传统机器学习领域防御方法的迁移应用,实际效果仍有待更多的研究验证.同时,由于对影响模型逆向攻击的底层机制缺乏深入认识,图模型逆向攻击防御方法的研究缺乏充分的理论指导.

**表 2 图模型逆向与成员推理攻击及其防御方法小结**

| 类型 | 算法 | 威胁 | 目标模型 | 背景知识 | 优点 | 缺点 | 描述 |
|---|---|---|---|---|---|---|---|
| 图模型逆向攻击 | GraphMI[83] | 白盒 | GCN,GAT,GraphSAGE | 节点特征及其类别标签 | 首次给出了图数据抽取攻击框架 | 对模型和背景知识的强假设,与实际不符 | 基于投影梯度优化求解方法和GAE模型推理训练图数据,发现影响力越大的边越容易被重构 |
| 节点成员推理攻击 | MIAttack[88] | 黑盒 | GCN,GAT,SGC,GraphSAGE | 未知目标模型训练数据 | 首次研究黑盒条件下节点成员推理攻击 | 基于实验比较,缺少理论分析 | 基于影子模型训练、攻击模型构建和成员关系推理三阶段框架,揭示节点类别数量、过拟合及模型架构对模型脆弱性的影响 |
| | NodeMI[89] | 黑盒 | GIN,GAT,GraphSAGE | 影子数据集、影子模型与节点拓扑 | 提供了清晰的攻击框架 | 基于实验比较,缺少理论分析 | 基于三阶段框架,探索节点自身特征、节点局部结构对成员关系推理的影响 |
| 链路成员推理攻击 | LinkSA[90] | 黑盒 | GCN | 分8种情况分别讨论 | 较为全面的分析讨论 | 缺少清晰的统一框架 | 按背景知识不同提出无监督、二元分类器、迁移攻击方法 |
| 逆向攻击防御方法 | Differential Privacy[83] | 白盒 | GCN | 目标模型及其求解算法 | 探索了基于训练过程的防御 | 效果不理想 | 增加高斯噪声的差分隐私方法 |
| | Random Edge Addition、Label-Only Output[89] | 白盒 | GraphSAGE | 目标模型及其训练数据 | 探索数据和模型角度进行防御 | 影响目标模型的实用性 | 在模型训练阶段随机添加链路、使目标模型仅输出预测标签 |

## 5　面向图模型隐私的攻击与防御方法

在智能服务系统中尽管智能服务接口是开放的，但其后台学习模型是私有和保密的.然而，研究表明攻击者可以通过开放的服务接口 API(Application Programming Interface)获取模型的参数、架构、功能等信息，从而造成模型机密信息的泄露.研究领域将以上无需训练数据和模型背景知识，通过开放 API 获得模型架构、参数的过程定义为模型萃取攻击(Model Extraction Attack).此攻击通过构建查询输入并获得相应的模型输出形成输入、输出数据集.这样的数据集能够反映目标模型(Target Model)的内在特征.在理想情况下，只要构建了充分覆盖目标模型输入空间的样本数据并获得相应的输出结果，就可以复制目标模型.

根据目标模型的复杂度不同，模型萃取攻击可以分为“简单模型精确提取”以及“复杂模型近似提取”两大类.对于决策树、逻辑回归、支持向量机等简单模型，其数学定义清晰，模型萃取攻击只需要获得充分的输入、输出对，即可通过方程求解的形式推理模型参数和超参取值，获得较准确的目标模型[91-92].相对的，对于深度学习、强化学习、集成学习等复杂模型，无法通过明确的数学公式进行形式化定义，模型萃取攻击需要通过输入、输出数据集间接刻画目标模型的内在特征，然后基于此数据集训练替代模型(Substitute Model)以近似目标模型的功能(Functionalities)[93-94].模型萃取攻击的基本架构如图 10 所示.

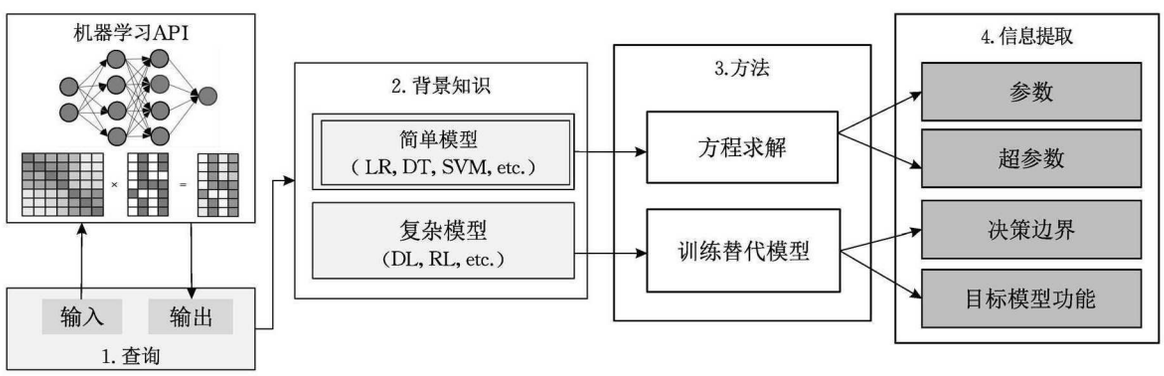



图 10　智能模型萃取攻击基本架构

### 5.1　图模型萃取攻击的主要方法

在传统机器学习领域，研究人员已经进行了模型萃取攻击研究.文献[91]分别基于置信度和类别标签提出了针对经典机器学习模型的萃取攻击方法.文献[92]实现了对逻辑回归、支持向量机等经典模型超参的萃取攻击.文献[95]提出了基于学习的高效查询算法从而使模型萃取攻击适用于大规模系统.

随着图模型的广泛应用，图模型萃取攻击问题逐渐受到研究人员的关注.针对基于 GNN 的节点分类模型，文献[96]提出一个基于学习的图模型萃取攻击方法.该方法根据节点类别采样目标模型的训练数据以获得子图集合，然后对得到的子图结构进行扰动，并基于目标模型对子图中的节点进行类别标记.最终，基于输入子图及其对应的节点类别标记训练替代模型从而实现模型的萃取.然而，此方法假设已知目标模型的训练数据并基于此构建查询样本，此假设与现实情况差异较大.与文献[96]类似，文献[97]对基于 GNN 的节点分类模型进行了萃取攻击，其框架如图 11 所示.该方法随机选择节点并进行控制，并基于这些节点向目标模型发出查询请求.由于选择的节点可能不相互连接，因此需要构建合成节点以关联采样节点形成训练图.然后，基于训练图以及节点类别标记构建具有相似功能的替代模型.结果表明替代模型的准确率与选择的攻击节点的数量密切相关.该工作与文献[96]都是基于目标模型的训练数据构建查询子图，其它的训练数据的构建方法及其效果仍有待研究.

为了探究背景知识对模型萃取攻击的影响，文献[98]考虑目标模型训练数据的结构信息、节点属性和辅助数据三种因素，根据攻击者具有完全、部分和无法获得相关信息的不同情况讨论图模型萃取攻击问题，分别提出了相应的攻击方法.这些方法基于背景知识查询目标模型以构建替代模型的训练数据，对于缺失信息通过可用数据进行构造补全.

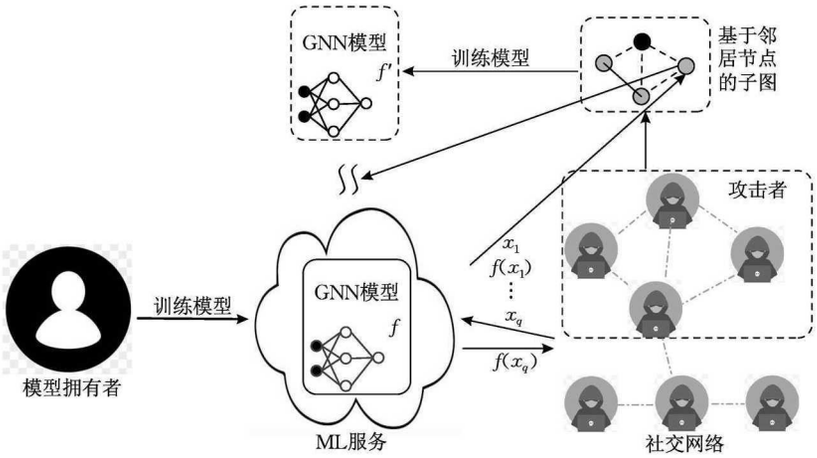



图 11 图模型萃取攻击示意图[92]

以上研究工作都是关于直推式图学习模型的，文献[99]关注归纳式图神经网络的萃取攻击问题，作者通过将攻击者的背景知识归纳为查询图和目标模型响应两个维度，系统性的定义了图模型萃取攻击的问题体系，并形成了六种攻击场景. 在此基础上，作者定义了适用于所有攻击场景的通用图模型萃取攻击框架，该框架主要包含两个部分：第一部分用于当查询图的结构信息不可用时学习离散图的结构；第二部分通过对节点特征和目标模型响应进行联合学习来建立替代模型.

### 5.2 图模型萃取攻击的防御方法

为了防御模型萃取攻击，传统机器学习领域提出了限制模型输出详细结果，或者对置信度的精度进行截断的保护策略. 同时，还提出了利用差分隐私方法对目标模型的输出结果进行保护，通过降低替代模型训练数据的准确性实现对模型萃取攻击的防御. 另外，在传统机器学习领域，研究人员提出了基于密码学的模型隐私保护方法、基于模型水印的模型隐私保护方法以及基于 Dropout 等策略防止过拟合的模型隐私保护方法. 然而，关于图模型的萃取攻击问题，当前研究人员还未提出针对性的防御方法，传统的通用性防御方法对图模型萃取攻击的有效性也仍有待验证和分析.

表 3 是对图模型萃取攻击方法的一个小结. 总体而言，与经典机器学习模型相比较，图模型的萃取攻击更具挑战性. 首先，由于图数据的非独立同分布(Non-IID)特征，相关学习模型主要为随机分块、图神经网络等复杂模型，从而攻击目标以训练具有相似功能的替代模型为主. 其次，由于图数据的关联性、稀疏性、低秩性等特征，如何构建少量具有代表性的查询样本从而获得关于整体图数据的学习模型的特征成为巨大挑战. 另外，由于图数据的关联性以及开放服务接口的异常检测机制，构建的查询样本要合法高效且相互兼容. 最后，当前研究工作主要是针对节点分类模型，对图分类、链路预测等任务模型的安全风险缺乏关注.

**表 3 图模型萃取攻击方法小结**

| 类型 | 算法 | 威胁 | 目标模型 | 背景知识 | 优点 | 缺点 | 描述 |
|---|---|---|---|---|---|---|---|
| 图模型萃取攻击 | GNNExtract[96] | 黑盒 | GCN | 目标模型训练数据及节点特征分布 | 基于子图训练替代模型，简单易实现 | 假设已知目标模型训练数据，与现实情况差异大 | 基于目标模型构建多个输入子图及其对应的节点类别标记以训练替代模型 |
| | ConMEA[97] | 黑盒 | GCN | 目标模型训练数据中受控制的攻击节点 | 考虑查询的合法性、隐匿性 | 已知训练图不符合现实情况；合成节点可能影响攻击效果 | 结合受控节点的查询结果和邻居连接构造训练数据 |
| | TaxMEA[98] | 黑盒 | GCN | 节点属性、图结构以及辅助子图 | 结合现实情况充分考虑不同背景知识对萃取攻击的影响 | 已知训练图不符合现实情况；无法提供统一攻击方法 | 基于背景知识查询目标模型构建训练数据，对缺失信息进行构造补全 |
| | InductiveGNN[99] | 黑盒 | GraphSAGE，GAT，GIN | 查询图的节点特征、图结构信息 | 适用于多种目标模型 | 只适用于节点水平的计算任务 | 根据背景知识和模型结果进行讨论，构建统一攻击框架 |

## 6　面向图模型安全的攻击与防御方法

在传统机器学习领域，模型算法主要应用于企业内部，训练数据和模型输入都是安全可控的. 然而，越来越开放的智能应用使模型算法面临不可信、甚至恶意对抗的计算环境. 例如，攻击者可以通过改变训练数据以形成不准确、有缺陷的模型算法. 攻击者可以蓄意构造输入数据从而欺骗目标模型. 因此，各种各样的对抗攻击(Adversarial Attacks)行为使得图模型面临严重的安全风险.

### 6.1　图模型对抗攻击框架

大量研究工作通过给良性样本添加特定噪声证明了深度学习的脆弱性[100-101]. 对于图学习，攻击者也通过修改图拓扑结构成功误导图模型产生了错误结果[25]. 图模型对抗攻击通过在模型训练阶段、模型测试阶段操纵图结构和节点特征影响模型性能. 根据攻击能力、扰动类型和攻击目标的不同，对抗攻击可以分为不同类型，整体框架如图 12 所示.

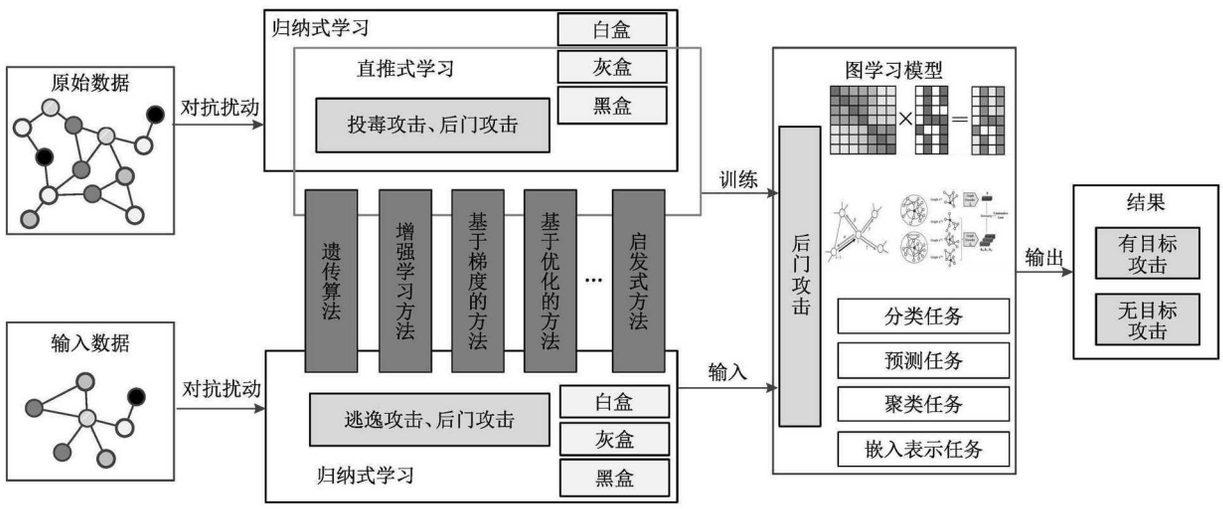



图 12　图模型的对抗攻击框架

(1) 攻击能力

攻击者可能在图模型的训练阶段或者测试阶段进行恶意攻击，具体依赖于实际场景中攻击者进行对抗扰动的能力. 研究领域将能够在模型数据采集阶段对训练数据进行对抗扰动的攻击称为“投毒攻击”，将能够在模型测试阶段对输入数据进行对抗扰动的攻击称为“逃逸攻击”，将能够在学习系统的全阶段注入和激活触发器的攻击称为“后门攻击”. 具体如下：

① 投毒攻击. 基于获得的原始数据资源，恶意攻击者根据目标模型的可用信息以及具体攻击目标，设计对抗扰动、对原始图数据的特征属性和结构拓扑进行修改投毒，使学习到的图模型具有低准确性或潜在漏洞，如图 12 中的上半部分所示. 关于投毒攻击，攻击者可以在已知、未知目标模型的情况下都可以进行数据扰动，然后基于投毒后的训练数据进行模型更新，从而导致模型性能下降.

② 逃逸攻击. 针对训练得到的图模型，恶意攻击者根据目标模型的可用信息设计对抗扰动、对模型的输入数据进行修改，使得目标模型对于修改后的输入数据无法给出正确的响应输出，从而逃脱目标智能系统的正常作用和功能，如图 12 中的下半部分所示.

③ 后门攻击. 在数据收集、模型训练、模型部署和更新阶段，攻击者可以针对训练数据和局部模型进行后门攻击，如图 12 各模块所示. 遭受后门攻击的图模型当遇到攻击者预先准备的触发器时后门被激活，从而产生攻击者预设的结果，否则按原模型进行预测，以此达到隐藏后门的目的.

(2) 扰动类型

为了进行图模型的对抗攻击，当前恶意攻击者主要通过对图数据产生不易察觉的对抗扰动来实现. 根据扰动信息类型的不同，扰动操作可以分为结构扰动和特征扰动，具体如下：

① 结构扰动. 攻击者根据影响模型训练或者逃脱模型检测的需要，在满足扰动成本限制的条件下，对图数据中的链路进行添加、删除以及重写操作，或者进行节点的添加和删除.

② 特征扰动. 攻击者根据对抗扰动的需要对图数据中的节点特征、链路特征进行修改.

(3) 攻击目标

根据攻击者通过对抗扰动导致模型产生特定行

为或者使模型整体性能受到影响的不同，图模型的对抗攻击可以分为有目标攻击(Targeted Attack)和无目标攻击(Untargeted Attack).

①有目标攻击. 针对特定的一组节点和链路进行投毒、逃逸和后门攻击使其产生特定的预测结果，例如将节点分类为特定类别.

②无目标攻击. 通过恶意扰动导致图模型输出错误结果，但不限定错误结果的范围. 例如，通过投毒攻击降低图数据中未标记节点的整体分类准确率.

### 6.2 图模型对抗攻击方法

(1) 面向图预测任务的对抗攻击

图预测任务是指通过已知的图拓扑结构和特征信息预测尚未存在链路的节点之间产生连接的可能性，在商品推荐、知识计算等领域具有广泛的应用前景. 文献[102]首次研究了图链路预测任务的投毒攻击问题，提出了面向图自编码 GAE(Graph Auto-Encoder)模型的迭代式梯度对抗攻击方法 IGA(Iterative Gradient Attack). 该方法基于交叉熵构建关于目标链路的损失函数，如式(8)所示，

$$\widetilde{L}=-wY_t\ln(\widetilde{A}_t)-(1-Y_t)\ln(1-\widetilde{A}_t) \tag{8}$$

其中 $Y_t\in\{0,1\}$ 是目标链路 $E_t$ 的真实标签，$\widetilde{A}_t$ 是 GAE 输出的目标链路 $E_t$ 的存在概率. 然后，计算损失函数 $\widetilde{L}$ 关于链路 $A_{i,j}$ 的偏导，如式(9)所示，

$$g_{i,j}=\partial\widetilde{L}/\partial A_{i,j} \tag{9}$$

此偏导取值反映了链路对于损失函数的影响程度. 对抗攻击方法 IGA 从对称化处理后的偏导中选择梯度取值最大的链路进行添加和删除，从而实现对抗攻击. 该方法计算简单、易于实现，但只适用于梯度求解类算法且要求目标模型已知.

结构相似性度量是链路预测的代表性方法之一，文献[103]将基于相似性的链路预测的投毒攻击描述为一个优化问题. 为了进行问题求解，作者提出了贪婪和启发式策略，从而攻击者通过链路扰动降低基于相似性的链路预测方法对目标链路集合的相似性取值. 该工作为基于相似性的链路预测投毒攻击提供了理论分析框架，但所提方法仍有待充分的验证评价. 针对结构相似性链路预测方法 RA(Resource Allocation)，文献[104]提出了随机的、启发式的、基于遗传进化算法的投毒扰动方法，通过实验比较证明了基于遗传进化算法的对抗扰动方法具有更好的效果. 然而，由于遗传进化算法循环执行选择、交叉、变异操作，基于遗传进化算法的投毒扰动方法收敛速度慢、搜索能力差.

为了探索链路预测方法面对投毒攻击的脆弱性，文献[40]提出了基于深度集成编码的投毒攻击方法 DEC(Deep Ensemble Coding). DEC 方法给出了链路预测投毒攻击的基础框架，结合原始图和攻击方法产生投毒图数据，从而使链路预测算法在其上产生比在原始图上更低的准确率，如图 13 所示. DEC 方法假设图的结构模式是链路预测算法的基础，因此链路预测的投毒攻击可以通过扰动链路、改变图的结构模式来实现. 为了发现对图结构模式具有重要影响的链路，DEC 提出了图结构增强方法，其基于低秩稀疏理论进行网络建模，如式(10)所示，

$$\min_{\boldsymbol{Z}}\lambda\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}\|_F^2+\|\boldsymbol{Z}\|_F^2 \tag{10}$$

通过模型求解获得最优稀疏矩阵 $\boldsymbol{Z}^*$，并通过式(11)进行拓扑重构，

$$\boldsymbol{X}^*=\boldsymbol{X}\boldsymbol{Z}^* \tag{11}$$

其中 $\boldsymbol{X}$ 是邻接矩阵. 通过迭代执行以上模型进行图结构增强，然后考虑全局和局部结构特征选择对图模式具有重要影响的链路进行添加和删除. 此方法不依赖目标链路预测模型，适用范围广，具有解析解，计算效率高，但启发式的重要链路选择机制无法保证获得最优的扰动链路.

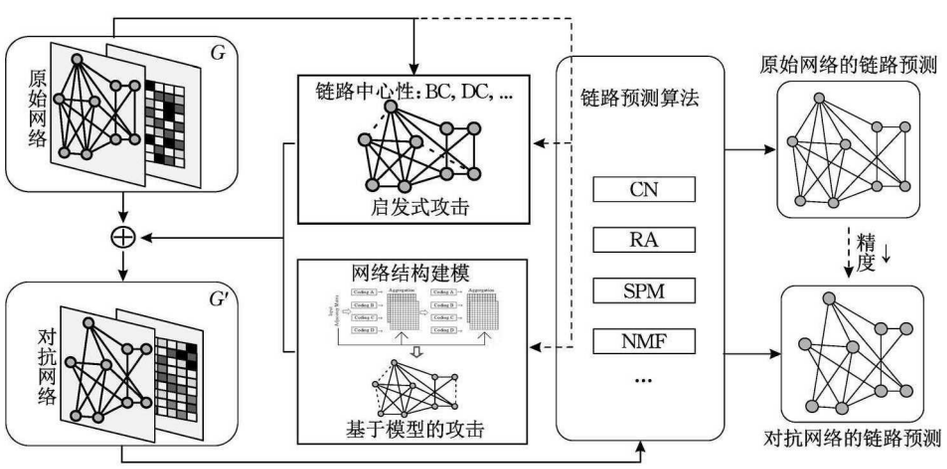



图 13 链路预测对抗攻击框架图[40]

针对动态图中的链路预测问题，文献[105]以归纳式的链路预测模型为目标，提出了基于强化学习的黑盒逃逸攻击方法. 该方法将图中要扰动的链路建模为强化学习要执行的动作，并通过求解强化学习中的策略函数来获得最终解. 同时，强化学习以链路预测模型预测精度的损失量为各个动作对应的奖励. 最终，通过迭代式的优化求解获得有效的扰动链路. 然而，该方法复杂度较高，无法适用于大规模图数据. 另外，针对基于邻接子图分类的图神经网络链路预测模型 SEAL[106]，文献[107]针对白盒逃逸攻击问题提出了基于贪心策略的启发式结构扰动方法. 对于输入图数据中的节点对，该方法在每一步添加、删除链路使损失函数取值最大，最终获得导致 SEAL 模型产生错误预测的子图. 该方法逻辑简单、易于实现，但是每一步扰动链路的选择都需要执行目标预测模型，复杂度高.

(2) 面向图聚类任务的对抗攻击

图聚类任务，也称为社团检测，其根据拓扑结构将图分为不同的节点和链路集合，使集合内部节点之间的连接尽可能的紧密、集合之间节点间的连接尽可能的稀疏. 文献[108]首次研究了图聚类模型的投毒攻击问题，提出了目标噪声注入(Targeted Noise Injection)和小社团(Small Community)两种攻击方法. 其中，目标噪声注入方法基于给定的噪声水平添加节点和链路以镜像图中已存在的结构，从而图划分算法无法区分真实节点和噪声节点. 小社团方法在完全图的基础上启发式的选择节点和链路进行删除. 这两种方法简单易实现，但对数据效用影响较大，且缺乏充分的理论支持.

针对社团检测的投毒攻击问题，文献[109]提出了基于遗传算法的攻击方法 Q-Attack. 此方法通过选择、交叉、变异等操作生成候选扰动链路，同时基于社团的模块度定义适应函数(Fitness Function)进行扰动链路的选择，如式(12)所示，

$$f=2*e^{-Q} \tag{12}$$

其中 $Q$ 为社团的模块度，表示节点和链路的聚集水平. 通过此机制，遗传算法选择降低模块度、增加适应度取值的链路作为扰动链路. 在 Q-Attack 的基础上，文献[110]研究了宏观、介观和微观三个尺度的社团检测投毒攻击问题，提出了基于遗传算法的演化扰动攻击 EPA(Evolutionary Perturbation Attack)方法. 以上两种方法有理论保证，具有较好的攻击效果，但遗传算法收敛速度慢，无明确的终止准则.

为了保护目标节点免受社团检测算法的检测，文献[111]提出了黑盒条件下的社团检测投毒攻击方法 CD-ATTACK. 该方法定义了包含社团检测替代模型和对抗攻击生成器的迭代式 Actor-Critic 框架，其中社团检测替代模型基于 GNN 进行节点嵌入表示，并根据嵌入表示的相似性划分社团. 对抗攻击生成器提出基于隐变量模型(Latent Variable Model)的对抗图生成机制，如式(13)所示，

$$P(\hat{A}|\boldsymbol{A},\boldsymbol{X})=fq_{\varphi}(\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{X})p_{\theta}(\hat{A}|\boldsymbol{A},\boldsymbol{X},\boldsymbol{Z})\mathrm{d}\boldsymbol{Z} \tag{13}$$

其中 $\hat{A}$ 为生成图，$\boldsymbol{A}$ 为邻接矩阵，$\boldsymbol{X}$ 为节点特征矩阵，$q_{\varphi}(\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{X})$ 为编码部分，$\varphi$ 为编码参数集合，$p_{\theta}(\hat{A}|\boldsymbol{A},\boldsymbol{X},\boldsymbol{Z})$ 为图产生部分，$\theta$ 为产生图的参数集合. CD-ATTACK 方法为社团检测投毒攻击问题提供了完整的模型框架，但此方法的复杂度和效率有待进一步的分析和验证.

(3) 面向图节点分类任务的对抗攻击

图节点分类任务，即图中未标记节点的分类问题，目标是基于给定的属性图及其部分节点的类别标签推理其余节点的类别. 面向基于 GCN 的半监督节点分类模型，文献[26]研究了节点分类的投毒、逃逸攻击问题，并提出攻击方法 Nettack. 该方法基于分类模型分别定义了关于图结构和节点特征的损失函数 $s_{\mathrm{struct}}$ 和 $s_{\mathrm{feat}}$. 然后，根据不可感知对抗扰动(Unnoticeable Perturbations)的要求，寻找使图节点统计特征变化量小于门限值的结构和特征扰动候选集合 $C_{\mathrm{struct}}$ 和 $C_{\mathrm{feat}}$. 最终，通过循环迭代发现使损失函数产生最大变化的最优结构扰动和特征扰动. Nettack 方法首次提出了图节点分类模型对抗攻击问题，但其扰动对象的发现与选择缺乏理论保障.

除了以上基于 GCN 的半监督节点分类模型的对抗攻击外，文献[112]分别基于强化学习、梯度更新和遗传算法提出了对目标节点两跳范围图结构进行扰动的逃逸攻击方法 RL-S2V、GradArgmax 和 GeneticAlg. 该工作比较系统全面，对当前图节点分类模型对抗攻击领域代表性的技术方法进行了实践，但是对各方法的性能、特性缺乏充分的比较分析. 已有工作都是针对基于 GNN 的节点分类模型，协同分类(Collective Classification)模型的对抗攻击缺乏关注. 为此，文献[113]提出了线性信念传播 LinLBP 分类模型的逃逸攻击方法，在白盒、灰盒条件下，通过最优化建模图链路的对抗扰动从而改变目标节点的类别，如式(14)所示，

$$\begin{aligned}&\min_{\boldsymbol{B}}\sum_{u,v\in V,u<v}B_{uv}C_{uv},\\&\text{s.t. }FNR=1,\\&\quad B_{uv}\in\{0,1\},\ \text{for } u,v\in V,\\&\quad \sum_{v}B_{uv}\leq K,\ \text{for } v\in V\end{aligned} \tag{14}$$

其中 $B_{uv}$ 为链路是否扰动的指示变量，$C_{uv}$ 为扰动成

本，$K$ 为允许的扰动上界. 该方法首次关注协同分类模型，但要求目标模型和数据的相关先验知识.

为了降低节点分类模型的整体性能，文献[114]提出了基于元学习(Meta-learning)的节点分类投毒攻击方法 Mettack. 此方法将图结构矩阵 $\boldsymbol{G}$ 作为超参数，通过计算攻击者训练之后的损失函数关于 $\boldsymbol{G}$ 的梯度以生成对抗扰动，如式(15)所示，

$$\begin{aligned}&\nabla_{\boldsymbol{G}}^{\text{meta}}:=\nabla_{\boldsymbol{G}}\ell_{\text{atk}}(f_{\theta^*}(\boldsymbol{G}))\\&\text{s.t. }\theta^*=opt_{\theta}(\ell_{\text{train}}(f_{\theta}(\boldsymbol{G})))\end{aligned}\tag{15}$$

其中 $opt(\cdot)$ 为可微的优化求解过程，$\ell_{\text{train}}$ 为训练时的损失函数，$\ell_{\text{atk}}$ 为攻击者试图优化的损失函数，$f_{\theta}$ 为分类模型. Mettack 对抗攻击方法对先验信息要求低、有理论保证，但复杂度较高，且只适用于基于损失函数优化求解的模型. 此外，基于链路添加和删除的扰动会明显改变图数据，为了构建不可察觉的扰动，文献[115]提出了重连操作(Rewiring Operation)并设计了基于强化学习的逃逸攻击方法. 考虑实际应用中进行链路扰动的困难性，文献[116]研究了基于节点注入(Node Injection)的图节点分类投毒攻击问题，提出了基于强化学习的非目标投毒攻击方法.

针对后门攻击问题，文献[117]基于节点特征构建触发器，给定原始图 $\boldsymbol{G}$ 以及图中节点 $v$ 的特征向量 $\boldsymbol{x}$，改变此特征向量中的特征子集设计特征触发器使得该节点被分类为目标类别 $y_t$. 通过对原始图数据部分节点设置触发器形成带后门的训练数据集，在此基础上进行模型训练获得后门 GNN 模型 $\Phi$，以及基于原始图数据进行模型训练获得 GNN 模型 $\Phi_o$，从而后门攻击可以定义为

$$\begin{cases}\Phi(v,x_t;\boldsymbol{G})=y_t\\\Phi(v,\boldsymbol{x};\boldsymbol{G})=\Phi_o(v,\boldsymbol{x};\boldsymbol{G})\end{cases}\tag{16}$$

其中 $\Phi(v,x_t;\boldsymbol{G})$ 表示在具有触发器 $x_t$ 的图数据 $\boldsymbol{G}$ 上执行后门 GNN 模型、触发器被激活时的预测，$\Phi(v,\boldsymbol{x};\boldsymbol{G})$ 表示后门 GNN 模型在触发器未被激活时的预测，$\Phi_o(v,\boldsymbol{x};\boldsymbol{G})$ 表示原始 GNN 模型的预测. 该工作从后门攻击的解释性出发，探讨触发器注入位置选择问题以及触发器注入位置对节点分类性能的影响. 作者在预训练中微调图学习框架提出后门攻击方法，利用解释性方法 GraphLIME 在节点特征向量中设置触发器，从而在重新训练之后获得具有后门的图学习模型. 文献[118]面向节点分类任务利用图特征生成基于子图的触发器，其能够根据输入图进行自适应变化，且与下游任务无关. 该方法通过双层优化对图模型和触发器参数进行优化求解. 文献[119]以单个节点作为触发器，当目标节点与触发器相连时激活后门，并提出基于线性图卷积和基于梯度的两种后门攻击方法.

(4) 面向图嵌入任务的对抗攻击

图嵌入(Graph Embedding)旨在将图中的节点表示成低维的向量形式，使其可以在向量空间中具有表示和推理能力. 文献[47]研究了无监督图嵌入方法的鲁棒性问题，以 DeepWalk 作为随机行走图嵌入方法的代表设计了投毒攻击的目标函数，最终基于特征值扰动理论给出了投毒扰动方法. 该工作为图嵌入模型脆弱性的理论性分析提供了方案，但其与目标模型紧密相关，适用性不强. 针对图嵌入方法 DeepWalk，文献[120]提出基于遗传算法的投毒攻击方法 EDA 以尽可能改变表示空间中节点之间的欧氏距离. EDA 方法具有良好的理论支撑，但每次迭代都需要对可能的扰动链路集合进行交叉、变异和评价，导致其收敛速度慢、计算代价高. 以图嵌入方法 DeepWalk 和 LINE 为目标，文献[121]分析两种方法的求解过程并构建损失函数，然后提出基于投影梯度下降法的投毒、逃逸攻击方法，其加权邻接矩阵的更新公式如式(17)所示，

$$\boldsymbol{A}^{t+1}=Proj(\boldsymbol{A}^t-s_t\cdot\nabla_{\boldsymbol{A}}L)\tag{17}$$

其中 $Proj(\cdot)$ 为投影函数，$L$ 为损失函数，$s_t$ 为步长大小. 通过多次迭代更新获得加权邻接矩阵 $\boldsymbol{A}^{opt}$，然后基于其各项的取值进行链路的扰动. 该方法逻辑简单、易于实现，但可行性与目标模型紧密相关.

针对图嵌入方法的逃逸攻击问题，文献[122]提出了快速梯度攻击方法 FGA(Fast Gradient Attack). 该方法基于 GCN 模型的梯度信息构建对抗网络生成器，其目标损失函数如式(18)所示，

$$L_t=-\sum_{k=1}^{|F|}Y_{tk}\ln(Y'_{tk}(\boldsymbol{A}))\tag{18}$$

其中 $Y_{tk}$ 为节点的真实标签，$Y'_{tk}$ 是模型输出的预测结果. FGA 计算目标损失函数关于链路的偏导信息，如式(19)所示，

$$g_{ij}=\partial L_t/\partial A_{ij}\tag{19}$$

此信息反映了链路对损失函数的影响程度，从而基于此进行图结构的对抗扰动 FGA 方法逻辑清晰、易于实现，但只适用于基于梯度更新的目标模型.

另外，现有的图嵌入对抗攻击方法主要为白盒攻击，依赖目标模型的架构、参数等信息，对真实模型的威胁有限. 为此，文献[123]将图嵌入表示建模为具有过滤器的图信号处理过程，基于图过滤器和特征矩阵度量嵌入表示的质量，最终通过攻击给定模型对应的图过滤器提出黑盒攻击方法 GF-Attack.

为了方便进行图模型对抗攻击的比较分析，本文对代表性的攻击算法及其特征进行了归纳总结，具体如表 4 所示. 从表中可以看出，当前对抗攻击算

**表 4　图模型对抗攻击方法小结**

| 任务 | 算法 | 攻击类型 | 威胁 | 目标类型 | 优点 | 缺点 | 描述 |
|---|---|---|---|---|---|---|---|
| 图预测对抗攻击 | IGA[102] | 投毒攻击 | 白盒 | 非目标 | 计算简单、易于实现 | 适用于梯度求解类算法且要求已知目标模型 | 基于梯度更新方法的链路扰动 |
| | SimAttack[103] | 投毒攻击 | 白盒 | 目标 | 为基于相似性的链路预测对抗攻击提供了理论分析框架 | 所提方法有待充分的验证评价 | 基于贪婪方法和启发式的链路扰动 |
| | RAAttack[104] | 投毒攻击 | 白盒 | 非目标 | 比较不同类型的攻击方法 | 遗传算法收敛速度慢、搜索能力差 | 基于随机、启发式、遗传算法的链路扰动 |
| | DEC[40] | 投毒攻击 | 黑盒 | 非目标 | 不依赖目标链路预测模型，适用范围广，计算效率高 | 无法保证获得最优的扰动链路 | 基于低秩稀疏理论的结构模式增强和启发式扰动链路选择 |
| | RLAttack[105] | 逃逸攻击 | 黑盒 | 非目标 | 利用优化求解可获得有效扰动链路 | 复杂度较高，无法适用于大规模图 | 通过强化学习获得实施链路扰动的最优行为 |
| | SealAttack[107] | 逃逸攻击 | 白盒 | 非目标 | 逻辑简单、易于实现 | 复杂度较高，无法适用于大规模图 | 结合损失函数和贪婪策略的链路扰动 |
| 图聚类对抗攻击 | CluAttack[108] | 投毒攻击 | 黑盒 | 非目标 | 计算简单、易于实现 | 对数据效用影响较大，且缺乏充分的理论支持 | 通过启发式方法扰动节点和链路 |
| | Q-Attack[109] | 投毒攻击 | 黑盒 | 非目标 | 有理论保证和好的攻击效果，适用于不同的社团检测算法 | 收敛速度慢，局部搜索能力差，无明确终止准则 | 基于遗传选择进行扰动链路的生成和选择 |
| | EPA[110] | 投毒攻击 | 黑盒 | 非目标 | 影响图、社团和节点三个层面社团检测算法的划分结果 | 收敛速度慢，局部搜索能力差，无明确终止准则 | 基于遗传选择进行扰动链路的生成和选择 |
| | CD-Attack[111] | 投毒攻击 | 黑盒 | 非目标 | 为社团检测对抗攻击问题提供了完整的模型框架 | 复杂度和效率有待分析和验证 | 基于社团检测替代模型和对抗攻击生成器迭代式更新生成对抗图 |
| 图节点分类对抗攻击 | Nettack[26] | 逃逸攻击、投毒攻击 | 白盒 | 非目标 | 首次提出了图节点分类模型对抗攻击问题 | 扰动对象发现与选择缺乏理论保障 | 基于损失函数和对统计特征的影响启发式的发现扰动链路 |
| | RL-S2V、GradArgmax、GeneticAlg[112] | 逃逸攻击 | 白盒、黑盒 | 非目标 | 对图节点分类模型对抗攻击领域代表性的技术方法进行了实践 | 对各方法的性能、特性缺乏充分的比较分析 | 基于强化学习、梯度更新、遗传算法生成对抗图 |
| | LinLBP[113] | 逃逸攻击 | 白盒、灰盒 | 目标 | 首次关注协同分类模型 | 要求目标模型和数据的相关先验知识 | 基于最优化求解的扰动链路生成 |
| | Mettack[114] | 投毒攻击 | 灰盒 | 非目标 | 对先验信息要求低，有理论保证 | 复杂度高，只适用于基于损失函数优化求解的模型 | 基于元学习产生扰动图 |
| | ReWatt[115] | 逃逸攻击 | 黑盒 | 非目标 | 基于重写的扰动操作不易被察觉 | 复杂度高，不适用于大规模网络 | 基于强化学习发现扰动链路 |
| | NIPA[116] | 投毒攻击 | 黑盒 | 非目标 | 节点扰动符合实际场景 | 复杂度高，不适用于大规模网络 | 基于强化学习注入扰动节点 |
| | EBA[117] | 后门攻击 | 黑盒 | 目标 | 基于解释性的后门攻击 | 触发器固定不变，仅适用于节点分类 | 基于代表性节点特征信息构造特征触发器 |
| | GTA[118] | 后门攻击 | 黑盒 | 目标 | 触发器生成具有自适应性 | 模型求解复杂 | 对后门图模型和触发器生成模型的参数进行轮流优化 |
| | NB[119] | 后门攻击 | 白盒、黑盒 | 目标 | 仅需注入单个节点的触发器生成策略 | 针对 GCN 模型 | 提出线性图卷积后门和基于梯度下降的后门 |
| 图嵌入对抗攻击 | PGDAttack[121] | 逃逸攻击、投毒攻击 | 白盒 | 非目标 | 逻辑简单、易于实现 | 可行性与目标模型紧密相关 | 基于迭代更新和投影梯度下降进行链路扰动 |
| | FGA[122] | 逃逸投毒 | 白盒 | 非目标 | 逻辑清晰、易于实现 | 只适用于基于梯度更新的目标模型 | 通过损失函数关于链路的偏导生成扰动链路 |
| | GF-Attack[123] | 投毒攻击 | 黑盒 | 非目标 | 不依赖目标模型信息 | 参数取值缺乏指导原则 | 结合损失函数和循环搜索生成扰动链路 |

法主要以启发式方法、梯度更新、遗传算法、强化学习、优化求解为主，存在复杂度较高、收敛速度慢等问题，当前对相关方法还缺乏充分的比较分析. 同时，当前各种方法中的对抗扰动程度以经验为主，缺乏对成功实现攻击时最优扰动规模的研究. 此外，对抗攻击要求添加不易察觉的扰动，如何根据图数据特征提出扰动衡量标准也是一个重要问题.

### 6.3 图模型对抗攻击防御方法

(1) 基于预处理的防御方法

预处理是指在模型训练之前对图数据中潜藏的对抗扰动进行检测和清除. 文献[41]使用相似度度量 Jaccard 对图中节点之间存在链路的可能性进行判别，发现具有高度不同特征的节点对，通过移除这些链路达到防止对抗攻击的目的. 文献[42]研究对抗攻击方法 Nettack 发现其仅仅影响邻接矩阵奇异值分解中的高阶奇异值，从而提出利用主要奇异值的低秩近似消除对抗扰动的方法，主要思想如图 14 所示，图中最左边表示具有扰动的有毒图，图中黑色的直线表示加入的扰动，通过 SVD 分解后，具有扰动边的秩比正常边的低，如图左边第二个子图所示，以此通过获取图中高秩部分的数据恢复出原始图，如图中右边两个子图所示. 为了抵御对抗扰动，文献[43]提出了基于图生成模型、链路预测和异常检测的预处理方法.

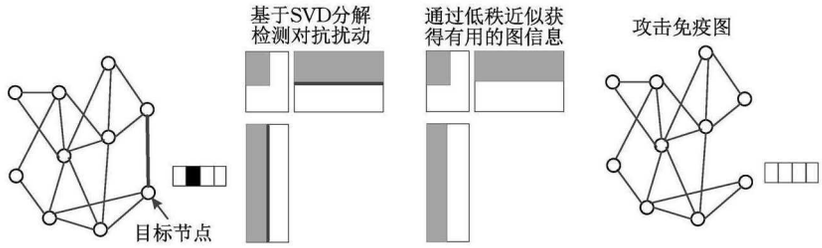



图 14 基于低秩近似的重构方法

(2) 基于对抗训练的防御方法

对抗训练方法在训练数据集中加入对抗样本，从而通过更加丰富的训练数据提升模型的鲁棒性. 文献[124]给出了基于模型解释性的对抗训练方法，该方法基于误差平方和寻找在局部范围内近似目标模型 $f$ 的简单可解释模型 $h$，然后基于此模型产生对抗样本进行对抗训练. 考虑图数据中结构关联性的影响，文献[125]面向半监督的节点分类模型提出了基于虚拟对抗训练的防御方法 GATV. 该方法利用最大-最小框架，在每轮迭代中先生成对抗样本以尽可能地最大化关联节点之间的差异、破坏光滑性，然后最小化目标函数、鼓励关联节点之间的光滑性进而完成模型训练. 类似的，文献[126]结合最大-最小框架和基于优化的对抗样本生成方法，给出了依次更新对抗扰动和模型参数的对抗训练算法.

(3) 基于鲁棒性模型设计的防御方法

鲁棒性模型设计是指针对对抗扰动的特征构建能够消除其对预测结果产生负面影响的学习机制. 基于注意力机制，文献[127]利用惩罚性聚合提出了鲁棒性图模型 PA-GNN，其通过为对抗扰动分配较低的注意力系数以限制对抗链路对预测结果的负面影响. 文献[128]提出基于注意力机制的图模型 RGCN，其在卷积层利用高斯分布进行节点表示，通过高方差刻画对抗扰动的不确定性，从而定义抑制扰动节点的聚合机制. 文献[129]提出适用于任何 GNN 模型的防御方法 GNNGuard，其利用同质性理论进行可疑链路识别，并基于此设计权重计算方法和消息传递机制.

(4) 基于攻击检测的防御方法

攻击检测方法是在模型测试阶段利用对抗扰动与正常数据的区别进行防御. 文献[43]提出基于链路预测、图生成机制以及异常检测方法对输入数据进行攻击检测，以实现图数据中潜在的恶意链路. 文献[130]认为对抗扰动会增加目标模型关于正常节点和扰动节点概率分布的差异性，从而提出基于假设检测的最大平均偏差(Maximum Mean Discrepancy)的攻击检测方法. 文献[131]认为异常节点的属性具有较差的预测能力，从而提出基于图感知标准过滤包含异常节点的子集的攻击检测方法 GraphSAC. 另外，文献[132]探讨和分析了近期提出的图模型攻击算法对图深度学习模型结构的影响，并提出通过 KL 散度度量节点及其邻居概率差异性的检测方法.

(5) 基于认证的防御方法

基于认证的防御主要基于随机平滑的方法，通过添加随机噪声来测试一定条件下分类器的鲁棒性. 在图学习领域，现阶段仅文献[133]提出了基于认证的图神经网络后门攻击防御方法，其采用随机子采样的方法构建平滑分类器，并对图中的标签进

行预测，采用投票的方式将预测多的标签作为最终的正确标签，以此防止后门攻击.

随着图模型安全风险的日益增加，对抗攻击的防御问题得到越来越多的关注.除了以上工作，研究领域还提出了基于学习的数据恢复[134]以及鲁棒性认证(Robustness Certification)[135-136]等防御方法.总体上，预处理方法基于对图数据模式规律的认知识别对抗扰动.该类方法简单易实现，但只有假设的结构模式与实际数据一致时才具有良好效果.同时，预处理方法独立于模型的训练过程，可能会错误地删除正常数据.对抗训练是当前应用最广泛的对抗攻击防御方法，其关键在于如何高效地生成高质量对抗样本，不足之处在于无法防御训练数据中不存在的对抗样本.鲁棒性模型设计主要基于图数据的低秩、稀疏、特征光滑等特征以及符合对抗样本特性的数学机制进行模型构建.攻击检测方法在获得输入数据之后、目标模型运行之前执行，要求低复杂度，当前主要以统计方法为主.基于认证的方法主要用于防御后门攻击.

## 7　开放资源

为了促进图学习隐私与安全问题的研究，研究人员贡献了众多开放资源.为了整合图数据隐私保护的相关成果，Ji 等人[137]开发了面向图数据匿名的统一开源平台 SecGraph，其集成了 11 种图数据匿名算法、19 种数据效用量化指标和 15 种基于结构的去匿名攻击方法.基于 SecGraph 平台，数据所有者能够利用最先进的匿名技术保护数据、测量数据效用、评估数据面对去匿名攻击时的隐私风险.

为了进行智能模型算法的安全性研究，Li 等人[138]开发了对抗攻击与防御平台 DeepRobust.该平台基于 PyTorch 实现，包含图学习领域的 12 种攻击算法和 6 种防御算法.PyTorch 基于具体功能将相关方法分为多个子包.其中，目标攻击子包(Targeted-attack Sub-package)包括目标攻击的基类和目标攻击算法，全局攻击子包(Global-attack Sub-package)包含全局攻击基类和全局攻击算法，防御子包(Defense Sub-package)包括 GCN 模型和其它防御图对抗攻击的方法，数据子包(Data Sub-package)提供了 Cora、Citeseer、Polblogs 等 10 余项公共基准图数据集.基于 DeepRobust 平台，研究人员可以高效的对所提出的攻击防御方法的有效性和图模型的鲁棒性进行分析评价.

除了集成化平台，Github 等开源社区上也具有相关攻击与防御方法的实现，具体如表 5 所示.

**表 5　图模型对抗攻击与防御开放源码**

| 来源 | 框架 | Github 库 |
|---|---|---|
| Zügner et al.[26] | TensorFlow | https://github.com/danielzuegner/nettack |
| Jin et al.[32] | PyTorch | https://github.com/DSE-MSU/DeepRobust/tree/master/deeprobust/graph |
| Li et al.[111] | Tensorflow | https://github.com/halimiqi/CD-ATTACK |
| Dai et al.[112] | PyTorch | https://github.com/Hanjun-Dai/graph_adversarial_attack |
| Zügner et al.[114] | TensorFlow | https://github.com/danielzuegner/gnn-meta-attack |
| Bojchevski et al.[47] | TensorFlow | https://github.com/abojchevski/node_embedding_attack |
| Chen et al.[122] | PyTorch | https://github.com/DSE-MSU/DeepRobust |
| Chang et al.[123] | Tensorflow | https://github.com/SwiftieH/GFAttack |
| Wu et al.[41] | PyTorch | https://github.com/DSE-MSU/DeepRobust |
| Entezari et al.[42] | PyTorch | https://github.com/DSE-MSU/DeepRobust |
| Feng et al.[125] | TensorFlow | https://github.com/fulifeng/GraphAT |
| Xu et al.[126] | TensorFlow | https://github.com/KaidiXu/GCN_ADV_Train |
| Tang et al.[127] | Tensorflow | https://github.com/tangxianfeng/PA-GNN |
| Zhu et al.[128] | PyTorch | https://github.com/KaidiXu/GCN_ADV_Train |
| Zang et al.[130] | PyTorch | https://github.com/chisam0217/Graph-Universal-Attack |
| Jin et al.[134] | PyTorch | https://github.com/ChandlerBang/Pro-GNN |
| Ling et al.[139] | PyTorch | https://github.com/ryderling/DEEPSEC |

## 8　研究难点与未来挑战

面对开放、复杂的现实环境，图学习遇到严重的隐私与安全威胁，直接影响图智能系统的实际应用.近年来，国内外研究人员对图学习的隐私与安全问题开展了研究工作，其中国外代表性研究团队包括密歇根州立大学数据科学与工程实验室、慕尼黑工业大学的数据分析与机器学习团队、爱荷华州立大学电子与计算机工程系团队、伊利诺伊大学香槟分

校安全学习实验室以及佐治亚理工学院机器学习中心团队等，国内代表性团队包括清华大学统计人工智能和学习团队、浙江大学网络系统安全与隐私实验室、浙江工业大学人工智能安全实验室以及重庆邮电大学本人所在团队.国内外相关研究情况表明，当前图学习隐私与安全的研究工作主要基于启发式方法、强化学习和遗传算法等通用框架以及其它领域模型方法的迁移应用，针对图数据的基础性、理论性研究较少，总体上还处于初期阶段.另外，现有研究工作主要集中在图神经网络以及节点分类任务的对抗攻击方面，对随机分块模型、随机行走、低秩稀疏等其它图学习模型以及链路预测、子图匹配、节点重要性排序等其它任务的关注较少，也对图模型逆向攻击、图模型抽取攻击的研究不充分.同时，当前的研究工作主要是在假设已知目标模型和训练数据的白盒条件下进行的，对结合图学习实际应用场景的灰盒、黑盒隐私与安全问题研究较少.少数黑盒条件下攻击与防御研究工作主要基于替代模型以及对抗样本在模型之间的迁移性假设，然而研究领域对对抗样本迁移性的存在条件和底层原理仍缺乏充分认识.考虑到当前图学习隐私与安全问题存在的以上难点，未来研究工作可从以下方面展开：

(1)隐私保护的图学习机制研究.当前图数据隐私保护主要以启发式方法为主，往往通过具体目标问题的分析进行匿名方法的定义，隐私保护水平没有保障，结合图数据特性的基础性、理论性研究有待加强.例如，基于图生成模型对图数据隐私风险进行量化研究，考虑同态加密的图数据加密处理研究以及设计密文图数据挖掘方法，或者基于安全多方计算和联邦学习进行隐私保护的图学习框架研究[140-141].因此，隐私保护的图学习框架及模型方法研究是未来工作的重要内容.

(2)图数据特征实证分析与鲁棒性建模研究.当前攻击与防御研究缺乏对图数据内在特征的认识和考虑.真实图数据往往是异质的、复杂关联的，现有对图数据的认识主要来源于传统复杂网络、社交网络的相关成果.当前对图数据特征的研究仍不成熟，对图数据结构模式与图模型底层机理的关系缺乏理解.除了少数模型设计工作考虑了图数据特征[142-143]，在图模型安全方面无论是对抗样本生成还是攻击检测都缺乏对图数据模式的充分利用.因此，图数据特征分析以及融合图数据特征的鲁棒性模型设计是图模型安全问题未来重要研究方向.

(3)图模型攻击防御机制的解释性研究.当前虽然研究领域针对图学习模型提出了许多攻击和防御方法并验证了它们的有效性，但关于图模型安全可信的基础性和解释性研究不足、对图模型的底层机理缺乏充分认识和理解，例如数据特征、模型架构以及模型求解方法等因素与模型脆弱性的关系及其影响机理、投毒攻击和逃逸攻击影响模型训练和推理的具体过程、对抗样本在替代模型和目标模型之间具有迁移性的根本原因等.与图像、文本等内容不同，图数据不具有直接的语义信息，相关的结构扰动需要结合图数据特征进行理解.因此，在后续工作中亟需结合图数据挖掘对图学习模型及其攻击防御方法进行解释性研究，从而利用形成的理论基础和基本原则指导鲁棒性图模型的构建.

(4)图学习隐私与安全研究体系的构建.虽然研究领域对图学习模型隐私和安全问题进行了初步研究，但至今仍缺乏完整的研究体系.图学习模型攻击与防御问题的定义仍然是多样性的，没有严格统一的理论框架、实验平台、数据集合以及评价指标，对图模型的鲁棒性缺乏度量方法，对攻击防御机制的可行性以及可能的最大防御效果缺乏评估方法，对图数据对抗扰动的不可感知性缺乏准确的定义.因此，除了具体的图模型算法研究外，研究体系的构建、度量方法的设计、评价指标的定义、实验平台以及数据集合的整理等都是未来的重点研究工作.

(5)面向实际应用的图模型攻击、防御方法研究.图学习模型广泛应用于欺诈检测、知识计算、模式识别、舆情分析等领域，除了文献[144-145]进行知识图谱的投毒攻击研究之外，当前对图学习模型具体应用问题的安全风险与防御机制的研究缺乏关注和深入分析.现有研究工作对模型方法的实际可行性考虑不足.例如，基于强化学习、遗传算法的对抗攻击方法都需要多次迭代更新，每次迭代都需要计算目标函数的取值，效率受限；基于梯度的对抗攻击方法需要计算目标函数关于图中节点和链路的梯度信息，无法适用于大规模图数据.另外，现有的攻击与防御研究往往假设对目标模型和数据都具有充分的先验知识，从而所提出的技术方法难以适用于背景知识受限的现实应用场景.综合以上情况，面向实际应用问题的图模型攻击、防御方法研究具有重要的现实意义.

## 9 总 结

当前，万物智能互联的应用场景使人工智能面临严重的安全威胁，直接影响智能系统的实际应用.

要促进人工智能产业的蓬勃发展，就必须解决好复杂条件下数据与模型算法的安全防护问题. 图学习作为人工智能的重要领域之一，由于图数据以及图学习模型的泛在性，进行图学习的隐私与安全问题研究具有十分重要的意义. 然而，当前研究领域对图学习隐私与安全的研究存在理论框架和评价体系缺乏、对图模型攻击与防御机理认识不足等问题. 为此，本文对近年来图学习隐私与安全的相关研究工作进行梳理归纳，从数据隐私、数据安全、模型隐私和模型安全四个方面分析讨论了存在的研究成果，理清了主要方法及其优势与不足. 同时，介绍了相关的开放资源和实验平台. 最后，指出当前研究面临的主要困难并展望了图学习隐私与安全问题未来可能的研究方向.

## 参 考 文 献


[1] Qiao S J, Han N, Gao Y J, et al. A fast parallel community discovery model on complex networks through approximate optimization. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(9): 1638-1651

[2] Washio T, Motoda H. State of the art of graph-based data mining. ACM SIGKDD Explorations Newsletter, 2003, 5(1): 59-68

[3] Yan X F, Han J W. gSpan: Graph-based substructure pattern mining//Proceedings of the 2002 IEEE International Conference on Data Mining. Maebashi, Japan, 2002: 721-724

[4] Lacasa L, Luque B, Ballesteros F, et al. From time series to complex networks: The visibility graph. Proceedings of the National Academy of Sciences, 2008, 105(13): 4972-4975

[5] Qian Zheng-Ping, Zhou Jing-Ren. The application and challenge of graph computing in Alibaba. Communications of the CCF, 2018, 7(14): 36-41(in Chinese)
(钱正平，周靖人. 图计算在阿里巴巴中的应用与挑战. 计算机学会通讯，2018，7(14)：36-41)

[6] Ji S X, Pan S R, Cambria E, et al. A survey on knowledge graphs: Representation, acquisition, and applications. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(2): 494-514

[7] Zhang Q S, Song X, Shao X W, et al. Object discovery: Soft attributed graph mining. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(3): 532-545

[8] Qiao S J, Han N, Zhou J L, et al. SocialMix: A familiarity-based and preference-aware location suggestion approach. Engineering Applications of Artificial Intelligence, 2018, 68: 192-204

[9] Gong N, Liu B. Attribute inference attacks in online social networks. ACM Transactions on Privacy and Security, 2018, 21(1): 1-30

[10] Wang P F, Guo J F, Lan Y Y, et al. Your cart tells you: Inferring demographic attributes from purchase data//Proceedings of the 9th ACM International Conference on Web Search and Data Mining. San Francisco, USA, 2016: 173-182

[11] Profiling T S U. Unified and discriminative influence model for inferring home locations//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 1023-1031

[12] Mahmud J, Nichols J, Drews C. Home location identification of Twitter users. ACM Transactions on Intelligent Systems and Technology, 2014, 5(3): 1-21

[13] Feng Deng-Guo, Zhang Min, Li Hao. Big data security and privacy protection. Chinese Journal of Computers, 2014, 37(1): 246-258(in Chinese)
(冯登国，张敏，李昊. 大数据安全与隐私保护. 计算机学报，2014，37(1)：246-258)

[14] Fang Bin-Xing, Jia Yan, Li Ai-Ping, et al. Privacy preservation in big data: A survey. Big Data Research, 2016, 2(1): 2016, 2(1): 1-18(in Chinese)
(方滨兴，贾焰，李爱平等. 大数据隐私保护技术综述. 大数据，2016，2(1)：2016，2(1)：1-18)

[15] Abawajy J H, Ninggal M I H, Herawan T. Privacy preserving social network data publication. IEEE Communications Surveys & Tutorials, 2016, 18(3): 1974-1997

[16] Li H X, Chen Q R, Zhu H J, et al. Privacy leakage via de-anonymization and aggregation in heterogeneous social networks. IEEE Transactions on Dependable and Secure Computing, 2020, 17(2): 350-362

[17] Wu T, Guo Y X, Chen L T, et al. Integrated structure investigation in complex networks by label propagation. Physica A: Statistical Mechanics and Its Applications, 2016, 448: 68-80

[18] Tong H H, Faloutsos C, Pan Y. Fast random walk with restart and its applications//Proceedings of the 6th International Conference on Data Mining. Washington, USA, 2006: 613-622

[19] Newman M E J. Network structure from rich but noisy data. Nature Physics, 2018, 14(6): 542-545

[20] Lee C, Wilkinson D J. A review of stochastic block models and extensions for graph clustering. Applied Network Science, 2019, 4(1): 1-50

[21] Pech R, Hao D, Pan LM, et al. Link prediction via matrix completion. Europhysics Letters, 2017, 117(3): 38002

[22] Xian X P, Wu T, Qiao S J, et al. NetSRE: Link predictability measuring and regulating. Knowledge-Based Systems, 2020, 196: 105800

[23] Xu Bing-Bing, Cen Ke-Ting, Huang Jun-Jie, et al. A survey on graph convolutional neural network. Chinese Journal of Computers, 2020, 43(5): 755-780(in Chinese)
(徐冰冰，岑科廷，黄俊杰等. 图卷积神经网络综述. 计算机学报，2020，43(5)：755-780)

[24] Xia F, Sun K, Yu S, et al. Graph learning: A survey. IEEE Transactions on Artificial Intelligence, 2021, 2(2): 109-127

[25] Rong Y, Xu T, Huang J, et al. Deep graph learning: Foundations, advances and applications//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. California, USA, 2020: 3555-3556

[26] Zügner D, Akbarnejad A, Günnemann S. Adversarial attacks on neural networks for graph data//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. London, UK, 2018: 2847-2856

[27] Zügner D, Borchert O, Akbarnejad A, et al. Adversarial attacks on graph neural networks: Perturbations and their patterns. ACM Transactions on Knowledge Discovery from Data (TKDD), 2020, 14(5): 1-31

[28] Wu B, Yang X, Pan S, et al. Adapting membership inference attacks to GNN for graph classification: Approaches and implications//Proceedings of the IEEE International Conference on Data Mining (ICDM). Auckland, New Zealand, 2021: 1421-1426

[29] Ji S L, Mittal P, Beyah R. Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey. IEEE Communications Surveys and Tutorials, 2016, 19(2): 1305-1326

[30] Casas-Roma J, Herrera-Joancomartí J, Torra V. A survey of graph-modification techniques for privacy-preserving on networks. Artificial Intelligence Review, 2017, 47(3): 341-366

[31] Sun L C, Dou Y T, Yang C, et al. Adversarial attack and defense on graph data: A survey. arXiv preprint arXiv:1812.10528, 2018

[32] Jin W, Li Y X, Wang Y Q, et al. Adversarial attacks and defenses on graphs. ACM SIGKDD Explorations Newsletter, 2020, 22(2): 19-34

[33] Chen L, Li J T, Peng J Y, et al. A survey of adversarial learning on graphs. arXiv preprint arXiv:2003.05730, 2020

[34] He Y Z, Meng G Z, Chen K, et al. Towards privacy and security of deep learning systems: A survey. IEEE Transactions on Software Engineering, 2020, 48(5): 1743-1770

[35] Ji Shou-Ling, Du Tian-Yu, Li Jin-Feng, et al. Security and privacy of machine learning models: A survey. Journal of Software, 2021, 32(1): 41-67(in Chinese)
(纪守领, 杜天宇, 李进锋等. 机器学习模型安全与隐私研究综述. 软件学报, 2021, 32(1): 41-67)

[36] Xu H, Ma Y, Liu H C, et al. Adversarial attacks and defenses in images, graphs and text: A review. International Journal of Automation and Computing, 2020, 17(2): 151-178

[37] Mo Xiao-Mei, Shen Hao, Yu Ding-Guo. Analysis of global news flow patterns based on complex networks. Journal of Southwest University (Natural Science), 2020, 42(12): 15-24 (in Chinese)
(莫小梅, 沈浩, 俞定国. 基于复杂网络的全球新闻流动模式分析. 西南大学学报(自然科学版), 2020, 42(12): 15-24)

[38] Bullmore E, Sporns O. Complex brain networks: Graph theoretical analysis of structural and functional systems. Nature Reviews Neuroscience, 2009, 10(3): 186-198

[39] Jia J Y, Gong N Z. AttriGuard: A practical defense against attribute inference attacks via adversarial machine learning//Proceedings of the 27th USENIX Security Symposium. Baltimore MD, USA, 2018: 513-529

[40] Xian X P, Wu T, Qiao S J, et al. Deep ensemble coding: Adversarial attacks against structure prediction models. Neurocomputing, 2021, 437: 168-185

[41] Wu H J, Wang C, Tyshetskiy Y, et al. Adversarial examples on graph data: Deep insights into attack and defense. arXiv preprint arXiv:1903.01610, 2019

[42] Entezari N, Al-Sayouri S A, Darvishzadeh A, et al. All you need is low (rank) defending against adversarial attacks on graphs//Proceedings of the 13th International Conference on Web Search and Data Mining. Houston, USA, 2020: 169-177

[43] Xu X J, Yu Y, Li B, et al. Characterizing malicious edges targeting on graph neural networks//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019: 1-13

[44] Newman M. The structure and function of complex networks. SIAM Review, 2003, 45(2): 167-256

[45] Costa L D F, Rodrigues F A, Travieso G, et al. Characterization of complex networks: A survey of measurements. Advances in Physics, 2007, 56(1): 167-242

[46] Lv L Y, Pan L M, Zhou T, et al. Toward link predictability of complex networks. Proceedings of the National Academy of Sciences, 2015, 112(8): 2325-2330

[47] Bojchevski A, Günnemann S. Adversarial attacks on node embeddings via graph poisoning//Proceedings of the World Wide Web. San Francisco, USA, 2019: 695-704

[48] Xuan Q, Shan Y, Wang J, et al. Adversarial attacks to scale-free networks: Testing the robustness of physical criteria. arXiv preprint arXiv:2002.01249, 2020

[49] Zhou X, Liang X, Du X, et al. Structure based user identification across social networks. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(6): 1178-1191

[50] Korayem M, Crandall D. De-anonymizing users across heterogeneous social computing platforms//Proceedings of the 7th International AAAI Conference on Weblogs and Social Media. Boston, USA, 2013: 689-692

[51] Shu K, Wang S H, Tang J L, et al. User identity linkage across online social networks: A review. ACM SIGKDD Explorations Newsletter, 2017, 18(2): 5-17

[52] Nilizadeh S, Kapadia A, Ahn Y. Community-enhanced de-anonymization of online social networks//Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. Scottsdale, USA, 2014: 537-548

[53] Lee W H, Liu C C, Ji S L, et al. Blind de-anonymization attacks using social networks//Proceedings of the 2017 on Workshop on Privacy in the Electronic Society. Dallas, USA, 2017: 1-4

[54] Gong N, Liu B. You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors//Proceedings of the 25th USENIX Security Symposium. Austin, USA, 2016: 979-995

[55] Wu L T, Ying X W, Wu X T. Reconstruction from randomized graph via low rank approximation//Proceedings of the 2010

SIAM International Conference on Data Mining. Austin, USA, 2010: 60-71

[56] Zhang Y, Humbert M, Surma B, et al. CTRL+Z: Recovering anonymized social graphs. arXiv preprint arXiv:1711.05441, 2017

[57] Xian X P, Wu T, Liu Y B, et al. Towards link inference attack against network structure perturbation. Knowledge-Based Systems, 2021, 218: 106674

[58] Narayanan A, Shmatikov V. De-anonymizing social networks //Proceedings of the 30th IEEE Symposium on Security and Privacy. Berkeley, USA, 2009: 173-187

[59] Backstrom L, Dwork C, Kleinberg J. Wherefore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography//Proceedings of the 16th International Conference on World Wide Web. Alberta, Canada, 2007: 181-190

[60] Zheleva E, Getoor L. To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles//Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain, 2009: 531-540

[61] Xian X P, Wu T, Qiao S J, et al. Multi-view low-rank coding-based network data de-anonymization. IEEE Access, 2020, 8: 94575-94593

[62] Bonchi F, Gionis A, Tassa T. Identity obfuscation in graphs through the information theoretic lens. Information Sciences, 2014, 275: 232-256

[63] Hay M, Miklau G, Jensen D, et al. Anonymizing social networks. Computer Science Department Faculty Publication Series, 2007, 180: 1-18

[64] Mittal P, Papamanthou C, Song D. Preserving link privacy in social network based systems. arXiv preprint arXiv:1208.6189, 2012

[65] Ying X W, Wu X T. Randomizing social networks: A spectrum preserving approach//Proceedings of the 2008 SIAM International Conference on Data Mining. Georgia, USA, 2008: 739-750

[66] Liu K, Terzi E. Towards identity anonymization on graphs//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. Vancouver, Canada, 2008: 93-106

[67] Xu Jia-Yu, Zhang Hong-Yan, Xu Li, et al. $k$-degree anonymous privacy protection scheme based on average degree of node in social networks. Computer Systems & Applications, 2021, 30(12): 308-316(in Chinese)
(许佳钰，章红艳，许力等. 社会网络中基于节点平均度的 $k$ 度匿名隐私保护方案. 计算机系统应用，2021，30(12)：308-316)

[68] Zou L, Chen L, Özsu M T. $K$-automorphism: A general framework for privacy preserving network publication//Proceedings of the VLDB Endowment. Lyon, France, 2009: 946-957

[69] Cheng J, Fu A W, Liu J. $K$-isomorphism: Privacy preserving network publication against structural attacks//Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. Indianapolis, USA, 2010: 459-470

[70] Liu C G, Liu I H, Yao W S, et al. $K$-anonymity against neighborhood attacks in weighted social networks. Security and Communication Networks, 2015, 8(18): 3864-3882

[71] Casas-Roma J, Salas J, Malliaros F D, et al. $K$-degree anonymity on directed networks. Knowledge and Information Systems, 2019, 61(3): 1743-1768

[72] Dong Xiang-Xiang, Gao Ang, Liang Ying, et al. Method of privacy preserving in dynamic social network data publication. Journal of Frontiers of Computer Science and Technology, 2019, 13(9): 1441-1458(in Chinese)
(董祥祥，高昂，梁英等. 动态社会网络数据发布隐私保护方法. 计算机科学与探索，2019，13(9)：1441-1458)

[73] Campan T, Truta T. A clustering approach for data and structural anonymity//Proceedings of the 2nd ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD. Las Vegas, USA, 2008: 8-18

[74] Hay M, Miklau G, Jensen D, et al. Resisting structural re-identification in anonymized social networks//Proceedings of the VLDB Endowment. Auckland, New Zealand, 2008: 102-114

[75] Zhou Yi-Hua, Zhang Bing, Yang Yu-Guang, et al. Cluster-based social network privacy protection method. Computer Science, 2019, 46(10): 154-160(in Chinese)
(周艺华，张冰，杨宇光等. 基于聚类的社交网络隐私保护方法. 计算机科学，2019，46(10)：154-160)

[76] Jiang Huo-Wen, Zhan Qing-Hua, Liu Wen-Juan, et al. Clustering anonymity approach for privacy preservation of graph data-publishing. Journal of Software, 2017, 28(9): 2323-2333(in Chinese)
(姜火文，占清华，刘文娟等. 图数据发布隐私保护的聚类匿名方法. 软件学报，2017，28(9)：2323-2333)

[77] Xu S, Su S, Xiong L, et al. Differentially private frequent subgraph mining//Proceedings of the IEEE 32nd International Conference on Data Engineering (ICDE). Helsinki, Finland, 2016: 229-240

[78] Jorgensen Z, Yu T, Cormode G. Publishing attributed social graphs with formal privacy guarantees//Proceedings of the 2016 International Conference on Management of Data. San Francisco, USA, 2016: 107-122

[79] Wang Jun-Li, Liu Xian-Hui, Guan Min. Differential privacy protection based generation model of social network publication graph. Journal of Tongji University (Natural Science), 2017, 45(8): 1227-1232(in Chinese)
(王俊丽，柳先辉，管敏. 基于差分隐私保护的社交网络发布图生成模型. 同济大学学报(自然科学版)，2017，45(8)：1227-1232)

[80] Liu P, Xu Y X, Jiang Q, et al. Local differential privacy for social network publishing. Neurocomputing, 2020, 391: 273-279

[81] Sala A, Zhao X H, Wilson C, et al. Sharing graphs using differentially private graph models//Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference. Berlin, Germany, 2011: 81-98

[82] Xiao Q, Chen R, Tan K L. Differentially private network data release via structural inference//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2014: 911-920

[83] Zhang Z X, Liu Q, Huang Z Y, et al. GraphMI: Extracting private graph data from graph neural networks//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21). Montreal, Canada, 2021: 3749-3755

[84] Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models//Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP). San Jose, USA, 2017: 3-18

[85] Ateniese G, Mancini L V, Spognardi A, et al. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. International Journal of Security and Networks, 2015, 10(3): 137-150

[86] He Ying-Zhe, Hu Xing-Bo, He Jin-Wen, et al. Privacy and security issues in machine learning systems: A survey. Journal of Computer Research and Development, 2019, 56(10): 2049-2070(in Chinese)
(何英哲, 胡兴波, 何锦雯等. 机器学习系统的隐私和安全问题综述. 计算机研究与发展, 2019, 56(10): 2049-2070)

[87] Truex S, Liu L, Gursoy M E, et al. Demystifying membership inference attacks in machine learning as a service. IEEE Transactions on Services Computing, 2021, 14(6): 2073-2089

[88] Olatunji I E, Nejdl W, Khosla M. Membership inference attack on graph neural networks. arXiv preprint arXiv:2101.06570, 2021

[89] He X L, Wen R, Wu Y X, et al. Node-level membership inference attacks against graph neural networks. arXiv preprint arXiv:2102.05429, 2021

[90] He X L, Jia J Y, Backes M, et al. Stealing links from graph neural networks//Proceedings of the 30th USENIX Security Symposium. USA, 2021: 2669-2686

[91] Tramèr F, Zhang Fan, Juels A, et al. Stealing machine learning models via prediction APIs//Proceedings of the 25th USENIX Security Symposium. Austin, USA, 2016: 601-618

[92] Wang B H, Gong N Z. Stealing hyperparameters in machine learning//Proceedings of the 2018 IEEE Symposium on Security and Privacy (SP). San Francisco, USA, 2018: 36-52

[93] Correia-Silva J R, Berriel R F, Badue C, et al. Copycat CNN: Stealing knowledge by persuading confession with random non-labeled data//Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN). Rio de Janeiro, Brazil, 2018: 1-8

[94] Orekondy T, Schiele B, Fritz M. Knockoff nets: Stealing functionality of black-box models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4954-4963

[95] Jagielski M, Carlini N, Berthelot D, et al. High accuracy and high fidelity extraction of neural networks//Proceedings of the 29th USENIX Security Symposium. Boston, USA, 2020: 1345-1362

[96] DeFazio D, Ramesh A. Adversarial model extraction on graph neural networks. arXiv preprint arXiv:1912.07721, 2019

[97] Wu B, Pan S R, Yuan X L. Towards extracting graph neural network models via prediction queries. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(18): 15925-15926

[98] Wu B, Yang X W, Pan S R, et al. Model extraction attacks on graph neural networks: Taxonomy and realization. arXiv preprint arXiv:2010.12751, 2020

[99] Shen Y, He X, Han Y, et al. Model stealing attacks against inductive graph neural networks//Proceedings of the IEEE Symposium on Security and Privacy. San Francisco, USA, 2022: 1175-1192

[100] Zhang Si-Si, Zuo Xin, Liu Jian-Wei. The problem of the adversarial examples in peep learning. Chinese Journal of Computers, 2019, 42(8): 1886-1904(in Chinese)
(张思思, 左信, 刘建伟. 深度学习中的对抗样本问题. 计算机学报, 2019, 42(8): 1886-1904)

[101] Ren K, Zheng T H, Qin Z, et al. Adversarial attacks and defenses in deep learning. Engineering, 2020, 6(3): 346-360

[102] Chen J Y, Shi Z Q, Wu Y Y, et al. Link prediction adversarial attack. arXiv preprint arXiv:1810.01110, 2018

[103] Zhou K, Michalak T P, Rahwan T, et al. Attacking similarity-based link prediction in social networks. arXiv preprint arXiv:1809.08368, 2018

[104] Yu S Q, Zhao M H, Fu C B, et al. Target defense against link-prediction-based attacks via evolutionary perturbations. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(2): 754-767

[105] Fan H X, Wang B H, Zhou P, et al. Reinforcement learning-based black-box evasion attacks to link prediction in dynamic graphs. arXiv preprint arXiv:2009.00163, 2020

[106] Zhang M H, Chen Y X. Link prediction based on graph neural networks. Advances in Neural Information Processing Systems, 2018, 31: 5165-5175

[107] Lin W Y, Ji S X, Li B C. Adversarial attacks on link prediction algorithms based on graph neural networks//Proceedings of the 15th ACM Asia Conference on Computer and Communications Security. Taipei, China, 2020: 370-380

[108] Chen Y Z, Nadji Y, Kountouras A, et al. Practical attacks against graph-based clustering//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. Dallas, USA, 2017: 1125-1142

[109] Chen J Y, Chen L H, Chen Y X, et al. GA-based Q-attack on community detection. IEEE Transactions on Computational Social Systems, 2019, 6(3): 491-503

[110] Chen J Y, Chen Y X, Chen L H, et al. Multiscale evolutionary perturbation attack on community detection. IEEE

Transactions on Computational Social Systems, 2020, 8(1): 62-75

[111] Li J, Zhang H L, Han Z C, et al. Adversarial attack on community detection by hiding individuals//Proceedings of the Web Conference. Taipei, China, 2020: 917-927

[112] Dai H J, Li H, Tian T, et al. Adversarial attack on graph structured data//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 1115-1124

[113] Wang B H, Gong N Z. Attacking graph-based classification via manipulating the graph structure//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. London, UK, 2019: 2023-2040

[114] Zügner D, Günnemann S. Adversarial attacks on graph neural networks via meta learning. arXiv preprint arXiv: 1902.08412, 2019

[115] Ma Y, Wang S H, Derr T, et al. Attacking graph convolutional networks via rewiring. arXiv preprint arXiv: 1906.03750, 2019

[116] Sun Y W, Wang S H, Tang X F, et al. Non-target-specific node injection attacks on graph neural networks: A hierarchical reinforcement learning approach//Proceedings of the International World Wide Web Conference. Taipei, China, 2020: 673-683

[117] Xu J, Xue M, Picek S. Explainability-based backdoor attacks against graph neural networks//Proceedings of the 3rd ACM Workshop on Wireless Security and Machine Learning. Abu Dhabi, United Arab Emirates, 2021: 31-36

[118] Xi Z, Pang R, Ji S, et al. Graph backdoor//Proceedings of the 30th USENIX Security Symposium (USENIX Security 21). Online, 2021: 1523-1540

[119] Chen L, Peng Q, Li J, et al. Neighboring backdoor attacks on graph convolutional network. arXiv preprint arXiv:2201.06202, 2022

[120] Yu S Q, Zheng J, Chen J Y, et al. Unsupervised Euclidean distance attack on network embedding//Proceedings of the 2020 IEEE 5th International Conference on Data Science in Cyberspace. Hong Kong, China, 2020: 71-77

[121] Sun M J, Tang J, Li H C, et al. Data poisoning attack against unsupervised node embedding methods. arXiv preprint arXiv:1810.12881, 2018

[122] Chen J Y, Wu Y Y, Xu X H, et al. Fast gradient attack on network embedding. arXiv preprint arXiv: 1809.02797, 2018

[123] Chang H, Rong Y, Xu T Y, et al. A restricted black-box adversarial framework towards attacking graph embedding models//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 3389-3396

[124] Liu N, Yang H, Hu X. Adversarial detection with model interpretation//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK, 2018: 1803-1811

[125] Feng F L, He X N, Tang J, et al. Graph adversarial training: Dynamically regularizing based on graph structure. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(6): 2493-2504

[126] Xu K D, Chen H G, Liu S J, et al. Topology attack and defense for graph neural networks: An optimization perspective//Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19). Macao, China, 2019: 3961-3967

[127] Tang X F, Li Y D, Sun Y W, et al. Transferring robustness for graph neural network against poisoning attacks//Proceedings of the 13th International Conference on Web Search and Data Mining. Houston, USA, 2020: 600-608

[128] Zhu D Y, Zhang Z W, Cui P, et al. Robust graph convolutional networks against adversarial attacks//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Anchorage, USA, 2019: 1399-1407

[129] Zhang X, Zitnik M. GNNGuard: Defending graph neural networks against adversarial attacks. arXiv preprint arXiv: 2006.08149, 2020

[130] Zhang Y X, Regol F, Pal S, et al. Detection and defense of topological adversarial attacks on graphs//Proceedings of the International Conference on Artificial Intelligence and Statistics. 2021: 2989-2997

[131] Ioannidis V N, Berberidis D, Giannakis G B. GraphSAC: Detecting anomalies in large-scale graphs. arXiv preprint arXiv:1910.09589, 2019

[132] Zhang Y X, Khan S, Coates M. Comparing and detecting adversarial attacks for graph deep learning//Proceedings of the Representation Learning on Graphs and Manifolds Workshop, International Conference on Learning Representations. New Orleans, USA, 2019: 1-7

[133] Zhang Z, Jia J, Wang B, et al. Backdoor attacks to graph neural networks//Proceedings of the 26th ACM Symposium on Access Control Models and Technologies. Spain, 2021: 15-26

[134] Jin W, Ma Y, Liu X R, et al. Graph structure learning for robust graph neural networks//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. California, USA, 2020: 66-74

[135] Wang B H, Jia J Y, Cao X Y, et al. Certified robustness of graph neural networks against adversarial structural perturbation. arXiv preprint arXiv:2008.10715, 2020

[136] Zügner D, Günnemann S. Certifiable robustness and robust training for graph convolutional networks//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Anchorage, USA, 2019: 246-256

[137] Ji S L, Li W Q, Mittal P, et al. SecGraph: A uniform and open-source evaluation system for graph data anonymization and de-anonymization//Proceedings of the 24th USENIX Security Symposium. Washington, USA, 2015: 303-318

[138] Li Y X, Jin W, Xu H, et al. DeepRobust: A PyTorch library for adversarial attacks and defenses. arXiv preprint arXiv:2005.06149, 2020

[139] Ling X, Ji S, Zou J, et al. DEEPSEC: A uniform platform for security analysis of deep learning model//Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP). San Francisco, USA, 2019: 673-690

[140] Wang B H, Li A, Li H, et al. GraphFL: A federated learning framework for semi-supervised node classification on graphs. arXiv preprint arXiv:2012.04187, 2020

[141] Wang B H, Guo J Y, Li A, et al. Privacy-preserving representation learning on graphs: A mutual information perspective. arXiv preprint arXiv:2107.01475, 2021

[142] Zhu J, Yan Y J, Zhao L X, et al. Beyond homophily in graph neural networks: Current limitations and effective designs. arXiv preprint arXiv:2006.11468, 2020

[143] Sankar A, Liu Y Z, Yu J, et al. Graph neural networks for friend ranking in large-scale social platforms//Proceedings of the Web Conference 2021. Ljubljana, Slovenia, 2021: 2535-2546

[144] Zhang H T, Zheng T H, Gao J, et al. Data poisoning attack against knowledge graph embedding. arXiv preprint arXiv:1904.12052, 2019

[145] Banerjee P, Chu L Y, Zhang Y, et al. Stealthy targeted data poisoning attack on knowledge graphs//Proceedings of the IEEE 37th International Conference on Data Engineering (ICDE). Chania, Greece, 2021: 2069-2074

**XIAN Xing-Ping**, Ph. D., lecturer. Her research interests include graph data mining, intelligent security and data privacy protection.

**WU Tao**, Ph. D., associate professor, doctoral supervisor. His research interests include intelligent security, privacy preservation and knowledge computing.

**QIAO Shao-Jie**, Ph. D., professor. His main research interests include big data and social network analysis.

**WU Yu**, Ph. D., professor, Ph. D. supervisor. Her research interests include network intelligence, network behavior analysis and data visualization.

**LIU Yan-Bing**, Ph. D., professor, Ph. D. supervisor. His main research interests include Internet of vehicles and medical image processing.

**Background**

Graph provides sufficient data resources for scientific research and commercial applications, which can be utilized for discovering the underlying knowledge and patterns of real-world systems and promoting the development of intelligent society. Accordingly, graph machine learning models have been widely used in social networks, knowledge graphs, e-commerce and so on. However, theoretical researches and practical experiments demonstrate that current technologies about graph machine learning are not yet mature and have high privacy and security risks. Thus, because of the realistic demand of artificial intelligence security and the extensive influence of graph machine learning models, the research on the privacy and security of graph learning has become an important issue in this field.

In recent years, the security threats to machine learning systems and the privacy preservation of graphs have been widely researched, but the security of graph learning model has only recently begun to catch the attention. At the same time, the privacy and security issues in intelligent computing systems are often inseparable. However, so far, there is no any comprehensive analysis on the research progress of graph learning privacy and security. To this end, this article reviews the recent works on the privacy and security of graph learning, in which the research progress of graph data privacy, graph data security, graph model privacy and graph model security is summarized systematically and the main achievements and shortcomings are discussed. Based on the analysis of the existing methods, the main challenges and the future research directions are outlined.

This work is supported by the National Natural Science Foundation of China "Research on Security and Dependability of Graph Machine Learning (62106030), Research on Network Reconstruction and Regulation Algorithms and Its Application in Privacy Preservation (61802039)" and Chongqing Municipal Natural Science Foundation (Postdoctoral Fund) (cstc2021jcyj-bsh0176, cstc2020jcyj-msxmX0804), and Partially Supported by the National Key R&D Program of China (2018YFB0904900, 2018YFB0904905), the National Natural Science Foundation of China (61772098), the Sichuan Province Science and Technology Planning Project (2021JDJQ0021), and the Chengdu Technology Innovation R&D project (2021-YF05-00491-SN).