### 摘要&关键字

GitHub是影响力卓越的代码托管平台和社交平台，然而随着平台的发展，恶意推广用户也随之增加，严重危害了GitHub的生态，如何识别出恶意推广用户有显著的现实意义。本文基于真实的数据模拟用户行为构建了包含GitHub恶意推广行为的用户图数据集。在以往的GitHub异常检测任务中，只采用基于特征的方法，效果欠佳。本文基于图对比学习策略，创新性地提出了神经网络架构GCLProD，充分使用了用户之间的结构信息。为了验证方法的有效性，我们在构建的用户图数据集和公开的图异常数据集上均进行了实验，并与其他常见的分类方法进行了比较。实验结果表明，我们的方法能很好地检测GitHub恶意推广者，在公开数据集上的表现也优于各种先进的图节点异常检测网络架构。

GitHub is an influential code hosting and social platform. However, as the platform grows, malicious promoted users also increase, seriously harming GitHub’s ecology. It is of significant practical importance to identify malicious promoted users. In this paper, we build a user graph dataset containing GitHub malicious promotion behavior based on real data simulated user behavior. In previous GitHub anomaly detection tasks, only feature-based approaches were used, which were ineffective. In this paper, we innovatively propose a neural network architecture, GCLProD, based on graph comparison learning strategy, which fully uses the structural information among users. To verify the effectiveness of the method, we conducted experiments on both constructed user graph datasets and publicly available graph anomaly datasets and compared them with other common classification methods. The experimental results show that our method detects GitHub malicious promoters well and outperforms various advanced graph node anomaly detection network architectures on the public dataset.

GitHub · Graph Contrastive Learning · Graph Neural Network · Malicious Promoter Detection

### 1 Introduction

GitHub[ref1]是全球范围内流行的基于Git的代码托管平台和社交平台，是开源生态的重要依托点。截止2023年1月，GitHub平台内已经有超过1亿的开发人员和超过3.7亿的代码仓库。随着使用人数的增多，GitHub上的恶意行为也不免泛滥，本文研究的恶意行为专指不以正常的代码和产品质量去为仓库提供star、fork、issue等事件，而是在获取利益后为某些仓库有意提供，如dagster团队在其研究[ref2]中提到，目前已有诸多GitHub黑市提供这种推广服务。

这些行为严重危害了GitHub的开源生态，检测出这些恶意推广用户具有巨大的现实意义。目前已有一些检测GitHub恶意推广用户的研究[ref2,ref3]，为了检测出恶意推广用户，他们采用了“活动低频性”的简单启发式方法，或者是为GitHub用户构建一组特征进而使用无监督的聚类方法，或者是采取SVM这种有监督的方法。显然，目前的这些方法都是基于规则或是传统的基于特征的方法。

然而GitHub的所有用户构成了一张图网络，他们之间存在着诸多结构关系，用户通过在GitHub的follow关系或行为关系如共同star、共同fork等联系着彼此。本文采用之前工作[ref3]搜集原始恶意推广用户数据的方法，创建了一个仓库并从黑市为其购买恶意推广者，作为正样本群体；同时基于GitHub Rest API[ref4]和华东师范大学X-lab实验室开源的GitHub全域日志数据库[ref5]和一定的统计方法，构建了正常用户的负样本群体，最终将负样本和正样本按用户行为合并，模拟构建了包含恶意推广用户的GitHub用户图数据集。在此基础上，为了更好地检测出恶意推广用户，本文采用了基于图神经网络的深度学习方法，为了更充分地利用图结构信息，并减少对标注数据的需求，将无监督的对比学习的策略引入到该检测问题中，采样过程中参考了CoLA的方法，同时引入了图注意力机制，提出了新颖的基于对比学习的图神经网络架构GCLProD 。本文的主要贡献如下：

1. 采集真实的GitHub用户数据，创新性基于用户行为的模拟，构建了用户恶意推广用户检测的GitHub用户图数据集
2. 提出了新颖的基于对比学习的图神经网络GCLProD，据我们所知，我们是第一个使用图神经网络方法研究GitHub恶意推广检测的人
3. 在构建的GitHub用户图数据和公开的节点异常数据集上，GCLProD都取得了卓越的效果

### 2 Related Work

目前针对于GitHub恶意推广用户的相关研究，如[ref3]，在对恶意推广用户进行特征工程后，使用了包含逻辑回归、SVM在内的6种基于特征的机器学习方法进行分类；而在其他的GitHub异常检测研究[ref6,ref7]中，大多是对GitHub机器人的识别，即由代码生成进行自动化提交的假用户，在这些研究中，他们都极其依赖人工标注的标签和经过经验构建的特征。因此为了减少这些依赖，我们提出图对比学习框架GCLProD，充分挖掘用户之间的拓扑信息，着重于GitHub恶意推广用户图数据的结构性构建。

图数据上的恶意推广用户的识别，推广开来，即是图上异常节点的分类问题。从深度学习的角度来说，在以往图异常节点检测的研究中，大都只利用了节点自身的信息，如AMEN[ref8]对图上每个节点进行嵌入，基于其自我信息来检测异常；或者如DOMINANT [ref9] 基于GCN 进行自动编码，重构邻接矩阵和节点特征矩阵，它设置超参数将检测节点的异常程度定义为两个矩阵的重构误差和，虽然利用到了结构信息，但是也不针对于异常节点特殊的异常结构，而且基于自动编码的方法，在初始输入和最终重构时都需要输入全图数据，对内存的要求也难以接受。因此引入了对比学习的策略，以一定方式构建对比组合，并将全量数据拆分为这些组合，降低了训练时对内存的要求。虽然在语音、图像领域对比自监督学习已经取得了良好的效果[ref10]，但是在GitHub恶意推广用户检测的领域，这个方法还是首次进行研究。

### 3 Model Framework

在本节中，我们描述了 CSGNN 的总体框架，如图 3 所示。CSGNN 框架由四个部分组成：数据预处理和结构化、对比实例对采样、基于 GCN 的对比学习和污染评估。首先，对原始数据进行预处理和更好地结构化，以实现对具有复杂相关性的检测数据的特征挖掘。接下来，使用RWR的局部子图采样策略来获取正例和负例对。之后，我们使用基于 GCN 的对比学习模型来训练每批实例对。最后随机遍历所有的检测样本作为采样样本。检测样本的污染评估是通过评估采样样本与相邻样本组之间的一致性来完成的。

### 4 Experiments

### 5 Conclusion

文献引用顺序

1. GitHub Wikipedia
2. GitHub刷星博客
3. GitHub恶意推广问题
4. GitHub REST API
5. X-lab GitHub全域数据库
6. GitHub机器人检测Bot Detection in GitHub Repositories
7. GitHub issue pr机器人检测 A gˇ rˇound-trˇ utˇ h datˇ aseˇt and classification model
8. AMEN
9. DOMINANT
10. Representation Learning with Contrastive Predictive Coding 对比自监督的好处