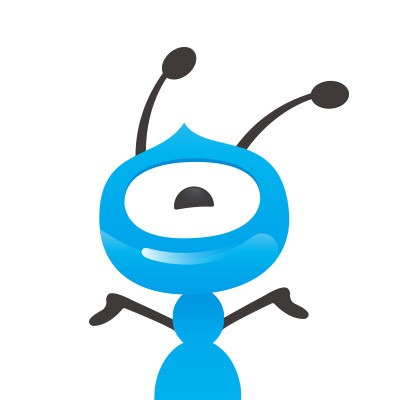
基于异构图神经网络的恶意账户识别方法

[极验](javascript:void(0);) 前天

[文章转载自公众号 蚂蚁金服科技 ， 作者 蚂蚁金服科技](https://mp.weixin.qq.com/s/oWNMP9ul0ZmQN7sarJ4C6w##)

ACM CIKM 2018 全称是 The 27th ACM International Conference on Information and Knowledge Management，会议于2018年10月22日-26日在意大利都灵省举行。

CIMK 是国际计算机学会（ACM）举办的信息检索、知识管理和数据库领域的重要学术会议。本次大会目的在于明确未来知识与信息系统发展将面临的挑战和问题，并通过征集和评估应用性和理论性强的高质量研究成果以确定未来的研究方向。

本篇文章分享了蚂蚁金服在CIKM 2018上的一篇论文《Heterogeneous Graph Neural Networks for Malicious Account Detection》，作者包括刘子奇、陈超超、杨新星、周俊、李小龙、宋乐。

在线账户伴随着（移动）互联网的诞生而产生，在金融服务领域，这种通过批量、低成本注册的恶意账户的存在是十分危险的。本文介绍了一种新的基于异构图的、面向恶意账户识别的图神经网络方法GEM ，这也是这也是世界上已知的第一个面向恶意账户检测的图神经网络方法。

**概述**

本文的主要目的是介绍一种新的基于异构图的、面向恶意账户识别的图神经网络方法（GEM, Graph Embeddings for Malicious accounts）[1]。该方法是支付宝为保障体系内账户安全，降低恶意账户带来的资损，保障支付体系健康，在人工智能领域所做的努力和尝试。这也是世界上已知的第一个面向恶意账户检测的图神经网络方法。

下面，我们将为大家讲解如下内容：

1.  恶意账户是什么？具有什么特点？

2.  为什么图神经网络方法能够高效识别恶意账户？

3.  我们的 GEM 方法如何工作？

在微信后台回复**【恶意账户检测】**获取论文资源~

**恶意账户**

什么是恶意账户

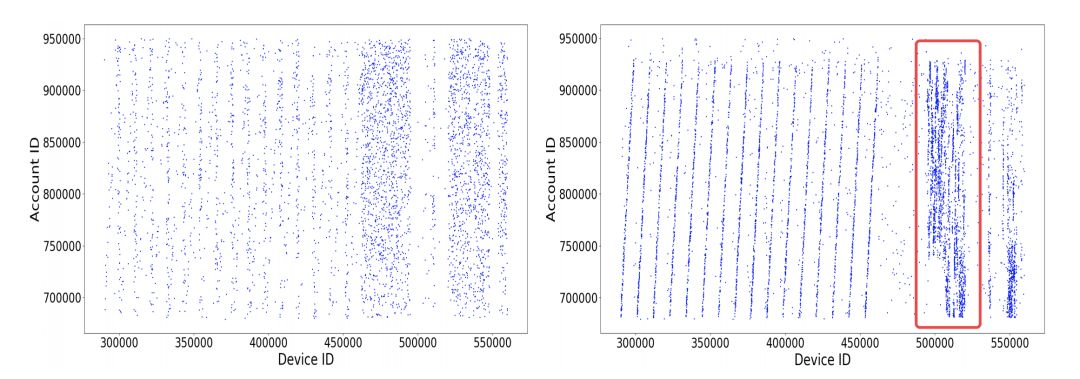
随着（移动）互联网诞生，催生了多种形式的在线服务，在线账户随之产生。比如：Gmail 提供的邮件服务，微博/Twitter 提供的短消息分享服务，支付宝提供的支付服务等。通过注册大量 Gmail 邮箱账户，恶意用户就可能迅速、大量地扩散垃圾广告等信息。微博账户等也可能催生僵尸账户达到某种非法营销、传播目的。在金融服务领域，这种恶意账户的存在就更加危险，比如注册大量新账户达到薅羊毛、洗钱、欺诈等目的。

用一句话总结：恶意账户具有强烈获取利益倾向和团伙性质，往往是通过批量、低成本注册的账户。

恶意账户的特点

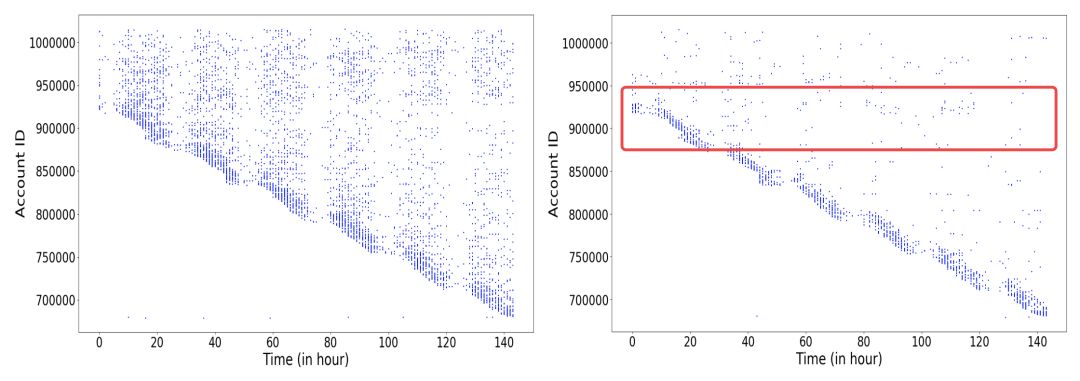
本文我们对黑产账户数据进行分析，并总结如下特点：

1.设备聚集性：见下图所示。两张图分别展示了用户（纵轴）过去是否在设备（横轴）有过登录行为。蓝色点代表该用户过去有在某设备上登录过。其中，左图显示的是正常账户特征，右图显示的是恶意账户特征。从图中可以看出，左图  pattern 较为均匀（regular），即便不同类型的设备（媒介）上，其和账户连接的 pattern 可能密度不太一样。右图则完全不同，我们可以在黑产账户上看到极为有规律和稠密的 pattern。这说明，黑产账户更倾向于在设备（媒介）的连通上有着高聚集性。



设备聚集性：黑产在设备上的聚集效应

2.时间聚集性：见下图所示。两张图分别展示了账户（纵轴）在时间（横轴）上的行为序列。其中一个蓝点代表该用户在某时间点上有登录行为。左图仍然为正常账户，而右图则是黑产账户。因为我们取的都是新注册账户，所以在注册时间点前无行为。从左图可以看出，正常账户在注册之后，每天会有均匀的登录 pattern，右图中的黑产账户则只在某个时间段内集中达成某种行为，这种 pattern 我们称之为时间聚集性。

时间聚集性：黑产在时间上的聚集效应

小结：这两种特征是黑产账户所固有形成的。即，这些黑产受利益所驱动而无法绕开这些模式（只要能准确捕获黑产账户之间共享的设备信息，这里的设备不限于某一个手机、某一个IP地址，可以认为是一种媒介）。我们针对这些数据特点设计了基于图的神经网络算法识别黑产账户。

图神经网络识别黑产

一个直观的方法是联通子图方法。我们先构建账户-设备二部图，由于设备聚集性，我们可以计算每个联通子图的节点数目，每个账户的危害程度取决于该节点所在联通子图的节点数目。该函数本质上可以用图神经网络抽象。

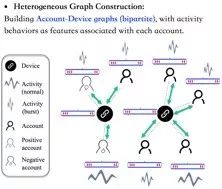
该方法可以准确识别那些设备聚集度特别高的黑产账户。但是对于设备聚集程度一般或较低的账户，很难做出准确区分。

GEM 方法如何工作

基于我们前面对设备聚集性和时间聚集性的分析，我们将上面的用于刻画联通子图的图神经网络方法进行扩展：

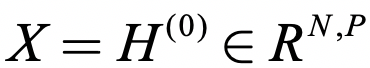
第一、我们构建异构图，包括账户类节点，以及多种类型的设备信息，如：电话、MAC、IMSI 以及其他 ID。

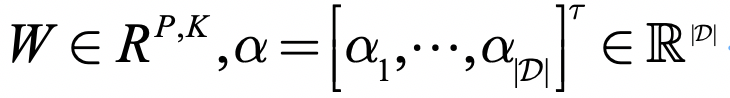
第二、我们为每个账户加入时间上的行为特征 X ∈ RN,P。其中每行 Xi 表示节点（账户或设备）在时间上的行为特征。我们希望构建的神经网络模型可以学习到通过设备聚集在一起的账户在行为特征上的模式，从而更准确的做出判别。



我们的算法如下：

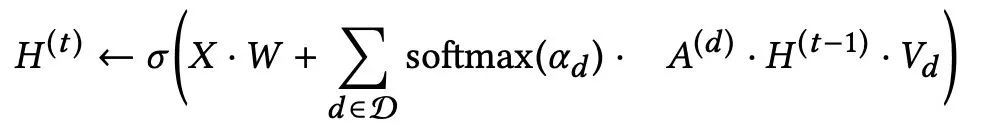
input：https://mmbiz.qpic.cn/mmbiz_png/8VOiack4f29dQSIZx08sfzkru6hK4S3UE4R1gVtlVqduVhX6fly4tDw3VS9tYksbX1PKDBU7aAkyAcazEOiaydfg/640?wx_fmt=png&wxfrom=5&wx_lazy=1&wx_co=1 account-divide adjacency matrices https://mmbiz.qpic.cn/mmbiz_png/8VOiack4f29dQSIZx08sfzkru6hK4S3UE5nyHiaKq1vCnia2T2Eok5htK1icpSwWqzZ0ibR9ez5Ju72PejV5IOALEAQ/640?wx_fmt=png&wxfrom=5&wx_lazy=1&wx_co=1.

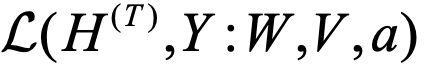
Account activities,depth of propagations *T，labels*

output：Parameters ，https://mmbiz.qpic.cn/mmbiz_png/8VOiack4f29dQSIZx08sfzkru6hK4S3UETjuqvEuGj9zg0RJVxn1lE3Aic3ArENEnEUURUribqRE7FCF3p92fjMcA/640?wx_fmt=png&wxfrom=5&wx_lazy=1&wx_co=1

While not converged **do**

https://mmbiz.qpic.cn/mmbiz_png/8VOiack4f29dQSIZx08sfzkru6hK4S3UEICrSLYibPeaquFEzXR0seGrOm9l6bGaB4ts9G3TJOLx3jeNDDM04iacQ/640?wx_fmt=png&wxfrom=5&wx_lazy=1&wx_co=1



Backpropagation based on  

**结果**

我们使用了连续 4 周的数据，比较了 GEM 和其他有竞争力的方法在这些数据上的 AUC 和 F1-score 上的表现。

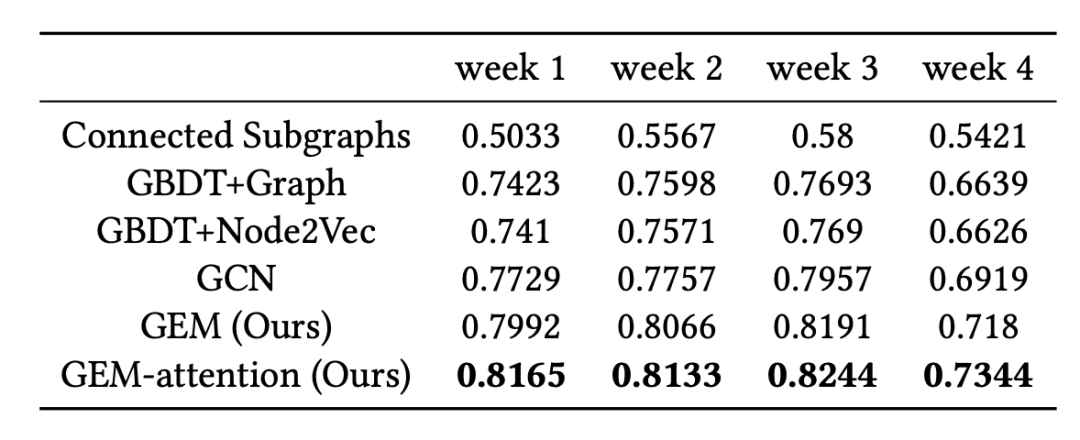


表1F1-score

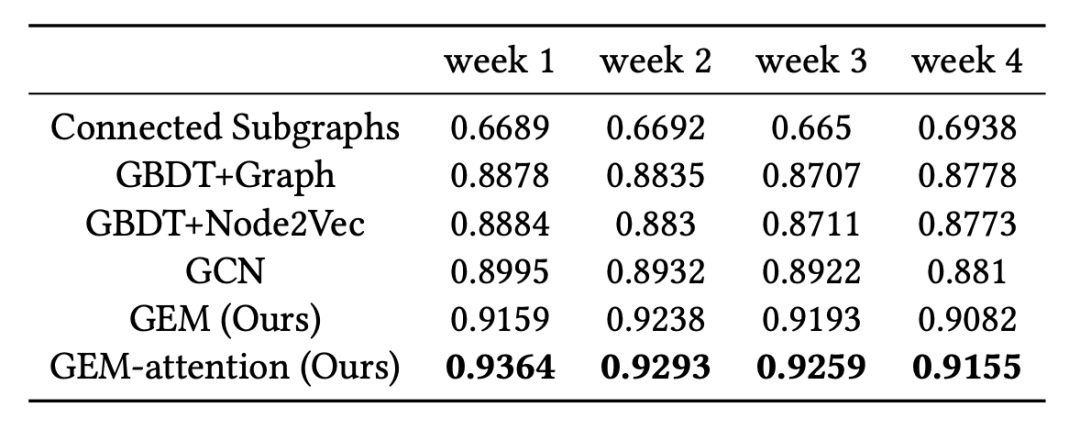


表2AUC

下面，我们比较了这些方法在 Precision-Recall 曲线上的表现。从图中可以看出，GEM 可以在召回的头部到尾部，都保持相对一致的高准确度。这使得我们的方法可以在避免打扰正常账户的同时，有效打击黑产账户。

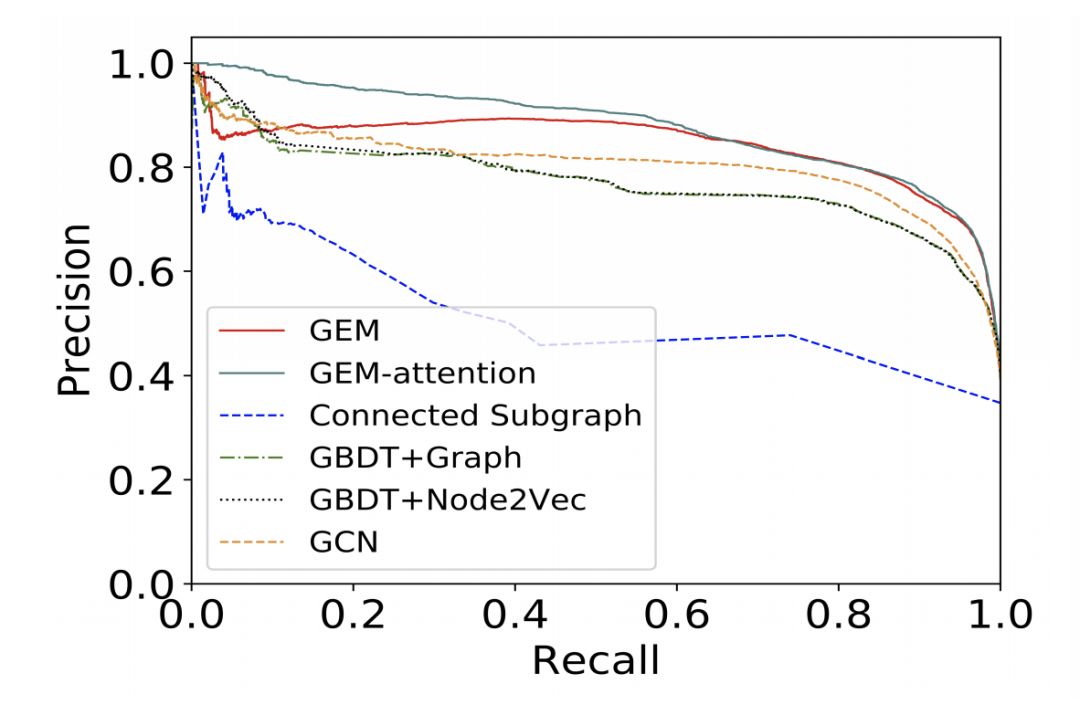
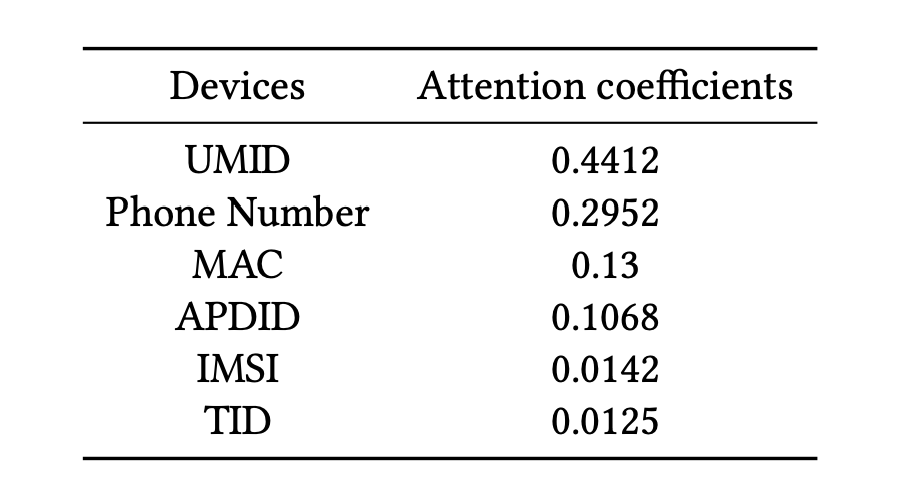


图1Precision-Recall曲线

最后，我们分析了我们的算法自动识别的异构图中不同类型账户的有效性。这些分析可以帮助我们更加有效理解哪类设备在当前有高概率会被利用，以及随时间变化，黑产策略的调整等。



参考文献

[1] Ziqi Liu, Chaochao Chen, Xinxing Yang, Jun Zhou, Xiaolong Li, Le Song. Heterogeneous Graph Neural Networks for Malicious Account Detection. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Turin 2108.