截止2020年12月1日反作弊算法进展

反作弊算法内部项目代号：七剑下天山项目

代表性算法汇总：

模型	算法	业务出发点	技术亮点	算法效果	业务效果
社区消费行为作弊账号识别	GCN图神经网络	针对社区的消费行为，以及用户和笔记关系，用户和设备网络环境关系，来构造模型数据结构，识别社区的用户消费行为作弊	使用了图卷积神经网络模型，和aws ai lab深度合作，借助aws sagemaker同时打通了线上查询图数据获得结果同时进行模型的推理运算，在业界领先	目前，新老版本在社区消费场景点赞收藏关注评论分享场景下，占整体拦截的40%。以目前线上打击策略为评估集合，准确率95%，召回50%左右。额外增益量级较大（每天二三十万账号，几百万行为），现实中采用逐渐放开的策略化方案。	整体在社区几个消费场景下日均300万左右的拦截量。以新版本为例子，在点赞收藏场景下拦截量3.8w/日，其中增益量2.8w/日。新老两个版本在关注，点赞收藏，笔记发布环节使用，发现额外增益显著，尤其在真人和疑似真人作弊方面，以及纯作弊关联小号，和营销内容产生小号，以及最近两波黑产覆盖率在90%以上。
行为相似作弊用户的发现	kp聚类	作弊用户往往存在行为模式，聚集渠道，作弊对象等方面的相同和相似性	使用了新的k-prototype聚类算法，在原有的只采用连续特征的基础上，新增加了离线特征。同时基于连续和离散特征来进行聚类。	准确率96%，行为增益4万+/日	在点赞收藏环境发现被之前策略漏过，黑产成功的大量行为渠道相似用户的作弊者，
人审关联小号发现，黄牛图标签传播	标签传播	作弊的账号往往之前存在关联性，可以从已经发现确认的账号出发，关联出更大的团伙	基于最新的aws dgl库开发了传统和神经网络两个版本的标签传播算法，在计算时间和实现复杂度上大大减少。	人审关联小号的发现准确率在95%上下。黄牛图高风险段准确目前接近80%。	从人审标注的问题账户出发，成功关联出了团伙小号。在电商黄牛防控上，预期扩大对黄牛的防控
笔记评论正负面识别	情感识别	作弊或危害社区生态的内容例如笔记或是评论，往往存在正负面情感聚集性，需要解决基础情感识别	使用了最新的bert模型架构来进行迁移学习和训练，小红书内领先	积极情感召回　　84%，准确 87% ，中性情感召回 95%， 准确 88%，负面情感召回21%， 准确 91%	用于品牌投资项目，和一些内容分析。
笔记中出现品牌词的识别	品牌识别	以软广和刷量为目的的笔记和评论总是会有品牌词在其中，需要在保证准确和召回的基础上，识别出包含有品牌词的笔记	使用了最新的bert模型架构来进行迁移学习和训练，小红书内领先，生态安全第一个实时计算算法生产接口上线	品牌数据准确率 82.66%，覆盖率80.18%。商业主体准确率88.33%，覆盖率80.95% 。	业务方新版本软广观远监控使用，本周四商业笔记审核接口调用正式上线。
私信出发团伙发现，赞藏相似利益团伙发现	社区发现louvain算法	小红书天然的是一个大社区，而作弊群体往往是这个大社区内的小团伙，需要通过社区发现算法来进行识别，从团伙角度来识别和打击团伙	使用spark graphx进行开发，实现的优化目标函数最优值基本等于英伟达基于cuda的算法效果，体现了反作弊组算法的专业水平	赞藏场景下 大盘增益率：2% 准确率：92%	对于利用私信进行欺诈，营销，广告，勾搭的团伙进行了有效的识别发现。对于社区作弊对象存在的相似的利益团伙进行了识别发现。
黄牛图计算，粉丝关注图计算	实时图计算	黄牛团伙和粉丝批量关注都会天然的形成一个有别于正常人的网络结构，在线上实时或准实时进行计算识别	使用了最新的nebula图数据库，针对小红书超大规模图能够实时查询计算，业内领先	黄牛图，准确率接近100%，占整体拦截的85%。粉丝图新一版本比老版本召回扩大五倍左右，占整体拦截的20%。	在效率上实时，准实时可以识别相应的作弊，在效果上可以从团伙角度发现更多的黄牛和刷粉丝小号。

算法在业务中的结果也可以参考：

点赞收藏反作弊和粉丝反作弊，评估指标对应的关键节点

赞藏行为大盘

Follow检测打击策略指标变化统计