小红书 | DataFun.

小红书社区反作弊探索与实践

费栋 @小红书

目录 CONTENT

小红书 | DataFun.

01

社区反作弊的意义

作弊的定义和行业风险

作弊：通过非正常手段滥用产品功能，以牟取利益的行为

不同行业作弊风险并不固定，需要结合产品形态和业务模式发现和定义

社区反作弊的意义

- 平台声誉：内容和数据不真实都会降低公众对于平台的认可度
- 品牌方：商业分析结论偏差，导致投放效果不如预期，品牌方低估流量价值

- C端用户：内容生态不健康不真诚和流量数据不真实，影响体验，长期会带来用户流失
- 作者：作弊者数据虚高对其他作者不公平，甚至可能导致“劣币”驱逐“良币”

- 数据导向：数据不准确带来分析、决策的偏差

- 面向监管：诈骗等问题存在监管风险
- 机器资源：大量作弊行为可能导致服务堵塞，影响正常用户使用

小红书 | DataFun.

02

社区黑灰产生态

作弊背后的产业链：分工明确

作弊投入 — 核心物料

手机号

猫池　接码平台

IP

代理IP　秒波IP

设备

模拟器改机　云控手机

作弊投入 — 技术实现

做号

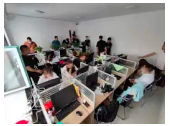
注册账号　养号

自动脚本和工具

作弊变现 — 运营

刷量引流服务

诈骗团伙

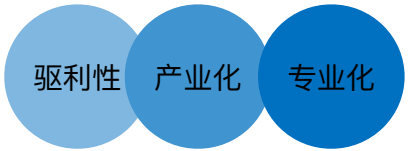


作弊手法迭代：从自动化工具逐渐演变为真人众包，作弊成本增加，识别难度变大

脱机类接口作弊

虚拟机、群控

众包



小红书 | DataFun.

03

作弊防控策略

作弊防控思路

目标

降低作弊行为占比

杜绝作弊行为

思路

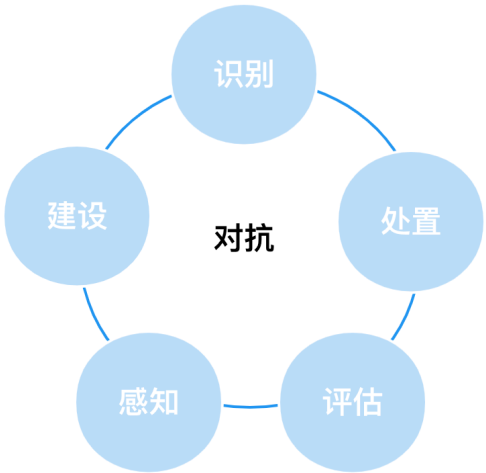
提高作弊成本，压缩作弊的获利空间，以降低作弊动机

关键链路

被动识别 → 主动防御

构建风险感知能力
控制核心资源
- 账号、设备
- 准入壁垒+存量清理

作弊防控思路



风险感知

更快发现风险，变被动救火为主动防御
分为情报运营、黑产卧底、红蓝对抗等模块

能力建设

建立面向对抗的快速响应能力

- 端+云联防
- 快速接入可灵活配置的风控系统
- 跨场景协同使用的风险画像

风险识别

提高识别准召

- 扩充数据：设备特征+账号特征+行为特征
- 增加维度：行为序列和团伙挖掘

风险处置

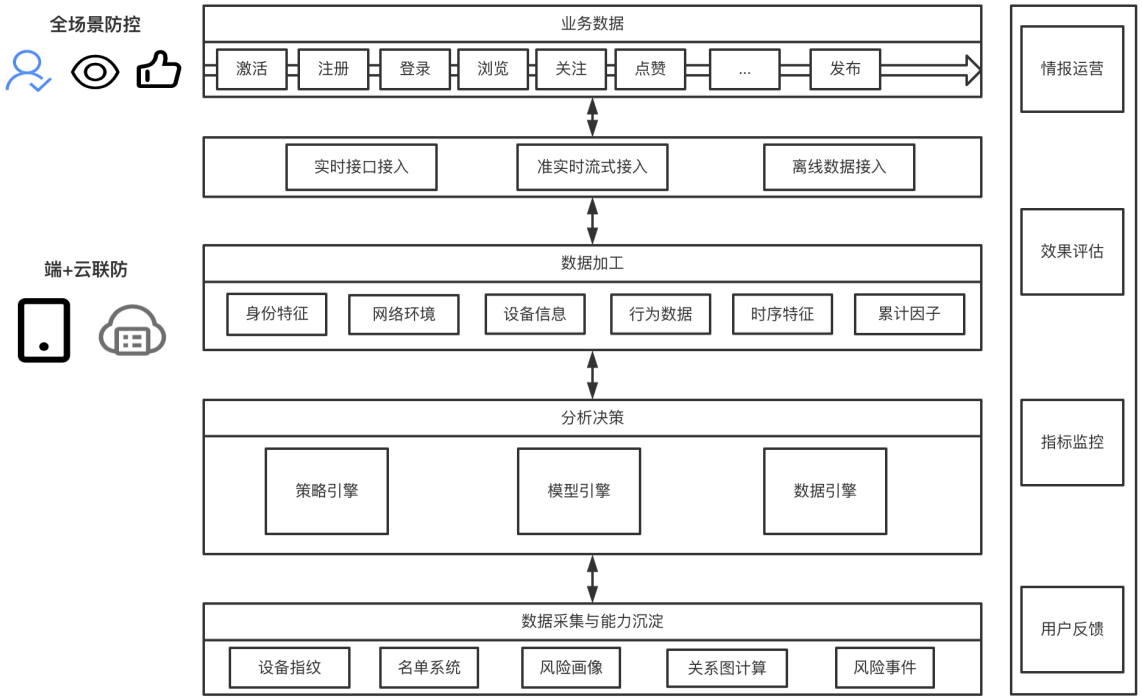
选择更有效的方式来降低对抗成本

- 打击对象：账号、设备、笔记、作者等
- 打击手段：请求拦截、人机校验、限流等

效果评估

通过指标评估风险水位，如作弊漏过滤、作弊服务价格、账号价格等

实现方案 — 风控体系：支持快速接入分析、灵活配置与能力迁移



小红书 | DataFun.

04

社区反作弊实践

数据刷量反作弊实践 — 风险治理

		治理方案	治理效果
影响	• 博主的虚假粉丝 • 笔记的虚假点赞、收藏、分享、阅读等	清理作弊行为所得	• 绕过及对抗强 • 尝试作弊无边际成本
实现链路	• 买量者购买刷量服务，或在众包平台发单 • 刷量者通过提供作弊服务套利	针对刷量账号做治理	• 账号成本变高，刷量服务价格上涨 • 刷量者尝试作弊有成本
作弊动机	• 买量者通过刷假数据提高自我流量“价值”，从而实现商业化流量变现 • 刷假数据也是商业合作中完成效果交付的手段	按作弊程度作流量分发降权与商业权益限制	• 买量者作弊意愿降低

从 治理【风险影响】 转变为 治理【实现链路】与【作弊动机】，作弊意愿降低，作弊量级下降显著

数据刷量反作弊实践 — 风险识别

阶段一

单点

基于行为主体的特征判断

基本假设

作弊主体有明确的特征异常

识别方法

- 限速策略
- 参数校验
- 环境异常
- 设备伪造、改机等识别
- 基于统计特征的监督学习

优点：解释性强

缺点：容易绕过

阶段二

群体

基于一组行为主体的特征

基本假设

作弊团伙存在明显的特征相似性

识别方法

- 无监督聚类模型
- 频繁项级挖掘

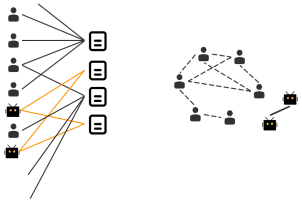
优点：专家知识依赖性较低

缺点：通过特征伪造和养号、真人作弊可绕过

阶段三

群体

基于一组主体的关系



基本假设

- 作弊用户倾向于给正常用户不喜爱的笔记/作者互动
- 作弊行为通常具有团伙性质

识别方法

- 实体关联图分割
- 高密子图挖掘
- 社群发现模型
- 标签传播模型

优点：不易被绕过

缺点：如果提高作弊成本，每个作弊账号只做少量的行为，可以绕过

非常感谢您的观看

小红书 | DataFun.