基于文本信息挖掘的电商黑灰产自动化监测及评估

1. 电商黑灰产现状

电商相关的网络黑灰产业务链可分为上游、中游、下游。

（1）上游：黑产基础服务，包括提供手机验证码、图片验证码、空包单号等服务

很多网站都会通过使用手机验证码来避免垃圾注册，或者图片验证码来识别机器行为。买家通过手机验证码平台获得手机号并填入所需注册验证的网站，平台给买家返回收到的验证码，从而通过网站的验证，一个手机验证码收费从0.1元到3元不等；黑产通过图片验证码平台组织打码工人人肉识别图片验证码，便可轻松绕过图片验证码的防控，每个仅需0.001元至0.25元。

如：

手机验证码平台：迅码 <http://www.xunma.net/index.html>

空包单号网站：空包100 <http://www.kongbao100.com/jiage1.asp>

（2）中游：账号生产销售，包括批量注册、维护、销售账号等

账号批量生产一般是通过软件进行垃圾注册或黑客通过拖库撞库洗号等手段获得账号数据，最后通过账号分销平台进行销售。

如：

淘宝小号批量注册机：<http://www.cncrk.com/downinfo/194697.html>

淘宝小号自助交易平台：<http://www.money668.com/>

（3）下游：交易活动平台，包括发布任务、刷单套现、店铺售卖转让等服务。其中刷单是支柱性黑产业务。

如：

淘刷刷 <http://www.taoshuashua.net/>

Yes！淘 <http://www.yestao.net/>

鸟人网 <http://www.birdhelper.com/>

有些网站同时包含了上游、中游、下游三个方面的业务，是综合性的平台。

以上这些网站是搜空包、刷单等关键词搜索引擎排在前几位的。虽然互联网+环境下刷单现象已经非常严重，但目前已知的电商活动刷单平台一共几十个，大量的其他黑灰产业务更加是无从考察，因此对电商行业的黑灰产网站进行大规模自动化检测，对黑灰产市场进行全面的监控和评估，具有很大的现实意义。了解电商商黑灰产的分布、运作和盈利情况，也有利于电商平台加大对相关作弊的检测和治理力度；另外电商黑灰产与其他行业的黑灰产业务链具有相似性，例如金融、社交、生活、游戏等行业，后续的研究方向可以从电商黑灰产扩大为网络黑灰产。

1. 主要问题
2. 对电商黑灰产网站进行全面检测

电商黑灰产网站规模巨大，初步预计有上千个，有的网站为避免被关闭在后期会更改网址[[[1]](#endnote-1)]或故意降低搜索引擎对其pagerank的计算，而有的网站故意包含很多黑灰产关键词以提高其排名。需要a)从已有的互联网海量舆情文本数据中挖掘黑灰产相关的网站名称或URL，b)并进一步爬取网站内容、结构、流量等特征，判别其是否为电商黑灰产网站。（漫爬，粗略）

1. 对电商黑灰产网站进行精确分类，并自动化监测和评估其规模

a)对电商黑灰产网站内容进行文本分析，根据已有的黑词库对网站生成标签，从而对网站类型进行归类，如刷单网站、物流单号网站、账号分销网站、综合性平台等等；（漫爬，精细，文本分析）

b)对不同类型的网站内容进行语义分析，提取相应关键信息。例如对于验证码平台、单号平台等，可以提取每个验证码服务、运单号的价格；对于刷单平台，提取卖家数量（对应卖家账号）、刷手数量（对应买家账号）、刷单任务数、任务佣金、商品名称链接、商品价格、刷单步骤等；（提取，语义分析）

c)根据提取的关键信息，评估网站的规模、盈利情况；建立商品价格-任务佣金回归模型，对刷单成本、刷单检测率等进行分析；（统计分析）

d）基于电商平台数据进行验证：黑灰产网站检测的准确率、买家、卖家交易情况等。

e)（后续：对不同网站生成指纹，对网站迁移进行持续监控）。

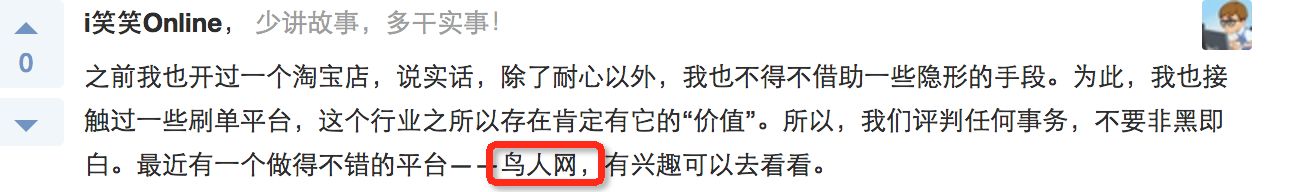
1. 对不同电商平台的黑灰产运作流程、刷单热度、发展趋势进行对比分析

（这个小方向对数据量的要求不高，近期可以先做，以便熟悉问题）

根据同一卖家在不同电商平台店铺的对应关系（已有淘宝-京东对应店铺名单10000组），爬取不同黑灰产网站中相关店铺的作弊情况并进行统计分析，结合新闻事件热度（一鸣），微博热度（晨阳），对淘宝、京东平台的作弊情况进行对比分析。

1. 研究方案
2. 从网络舆情文本数据中挖掘黑灰产相关的网站名称或URL

如：





1. URL解析对应到网站名称或网站地址，过滤掉白名单网站，得到待检测的网站
2. 对（2）中得到的所有网站进行粗略的漫爬，根据网站内容、结构、流量等特征，建立黑灰产网站检测模型
3. 对（3）中检测出的黑灰产网站进行精细漫爬，使用文本分析的方法结合已有黑词库对网站进行分类
4. 对（4）中不同类型网站的内容提取关键信息，如：对于刷单平台，提取任务帖中的佣金、商品链接、商品价格、店铺名称、电商平台等



对于空单网，提取物流价格等



1. 根据（5）提取的关键信息，对网站规模、盈利、刷单成本、刷单检测率等进行统计分析；对不同电商平台的作弊情况进行对比分析。
2. 验证
3. 研究进展
4. 从阿里目前提供的10000条舆情文本数据中提取了约2000条URL。其中有些是黑灰产网站，如：

' http://www.zhenshua.com/ '(交流论坛)

'http://t.cn/RbRKaOo'（流量平台）

'http://www.milioo.com'（刷单平台）

'http://www.88dan.com/?uid=393'（空单服务）

'http://www.100estore.com'（店铺转卖）

还有一些是淘宝、考拉海购、京东等的商品链接，还有微博的链接，类似于这样的：

http://t.cn/R6VhICj???

http://t.cn/R6cXgHe??

http://t.cn/8Fv70Jc

http://t.cn/R6cxLlk??

1. 对提取的URL对应的网站进行了内容的爬取，但是只选取了一两个作为start\_urls，还没有进行全量的爬取。

1. []以下五个网站为2015年www的文章《》中分析的五个淘宝刷单平台，目前已经全部关闭。调研过程中发现大量此种失效网址，相关网站是否迁移到其他网址还不能确定。

   <http://www2.88sxy.com/>

   <http://www.ntyjy.com/>

   <http://www.shuazuanshuaxinyu.com/>

   <http://www.kus.cc/Index.html>.

   <http://www.shuakewang.com/> [↑](#endnote-ref-1)