基于微博粉丝页面爬虫所得数据的明星粉丝结构分析分析

一、背景介绍

当前的影视行业中，一群广受欢迎、粉丝众多的“流量明星”受到了人们的火热追捧。从 2011 年杨幂一夜走红，成为微博热搜常客，到近年来，蔡徐坤、迪丽热巴等微博粉丝过千万的流量明星不断涌现，“流量即王道”的思想逐渐被人们广泛接受。

明星是“注意力经济”的产物，粉丝越多商业价值越大。看似处于产业下游被收割地位的粉丝，已经开始主导着偶像的人气和商业吸引力。他们已经参与到偶像的职业规划中，用自己的力量为偶像争取好资源。而市场似乎也认可了这种做法，明星流量已成为众多企业的营销利器，现在广告商、制片方最看重的，不再是明星的业务水平，而是背后的粉丝力量。新浪微博，这个拥有 3.37 亿注册用户的平台，自然成为了企业产品重要的推广平台，企业邀请流量明星为自己的产品做广告推广，通过推广微博点赞数和转发数评价推广效果。

然而与此同时，“水军”、流量造假等现象也逐渐出现。2019 年 2 月，中央电视台以《惊人数据的秘密》为题，深入揭露了娱乐圈里一些艺人们流量数据造假的行为，将一些看似光芒万丈、流量数据惊人的“流量明星”打回原形。流量明星的实际影响力到底有多少？一条推广微博的转发点赞中有多少是真实粉丝？企业选择流量明星做广告到底是否是明智之举？

流量造假可以形成明星广受关注、人气很高的假象，使影视业逐渐走向“唯流量是从，忽视演员真实演技”的偏途，也为众多制片方、广告商带来投资困扰和经济损失。在粉丝经济时代，了解当下流量明星所能带来的真实流量以及分析真假粉丝特点，无论对制片方、广告商的投资决策还是大众理性认识流量明星所扮演的角色都具有重要意义。

二、数据来源与说明

三、描述性分析

为研究流量明星微博假粉的具体情况，我们选取了 6 大流量明星在 5 月中上旬发布且转发总量相似的六条微博，分别是：

迪丽热巴推广 ysl 香水（5 月 2 日，转发量 100 万+）（去重之后得到 5405 条）

王俊凯推广纯悦矿泉水（5 月 16 日，转发量 100 万+）（去重之后得到 4693 条）

王源推广中华牙膏（5 月 19 日，转发量 100 万+）（去重之后得到 4645 条）

易烊千玺推广依云矿泉水（5 月 16 日，转发量 100 万+）（去重之后得到 4106 条）

蔡徐坤公益形象大使（5 月 12 日，转发量 100 万+）（去重之后得到 7212 条）

朱一龙欧舒丹沐浴露（5 月 13 日，转发量 100 万+）（去重之后得到 6611 条）

为统一标准，我们爬取了这六条微博转发页面**前 10000 页**的转发数据，去掉一摸一样的重复转发后[1]，得到每人约 4000-7000 条样本数据。

每条样本数据包含的变量有：转发人微博 ID、转发人微博昵称、转发人微博关注人数、转发人微博粉丝数、转发人微博发博数、转发人微博性别、是否是微博会员、微博会员等级、微博账号等级、转发所用设备、转发所带评论、转发之后被评论、被转发、被点赞数量等。

1、总体数据把握

为对总体转发数据有一个整体把握，我们先选择了迪丽热巴和蔡徐坤的数据进行总体的描述性统计分析，发现他们的数据在微博等级、微博关注人数量、微博粉丝数量、微博发布数量的分布中都有特别突出的异常值。

为进一步寻找这些异常值出现的原因，我们在迪丽热巴和蔡徐坤的数据集中选取了微博等级为 4，微博关注人数小于 5 且微博粉丝数量小于 5 的若干条微博账号，访问他们的微博页面后，发现这些微博账号长期帮助迪丽热巴、蔡徐坤转发微博，同一条微博重复转发数量高达十多条，而且转发评论大多是表情符号，并无太大实际意义；转发之后的被点赞数、被评论数、再转发数几乎都为 0；最重要的是，这些微博 ID 名称大多是“用户+数字”的格式。于是，结合日常经验，我们有理由相信，这些账号是专门帮助特定明星转发微博的“假粉”，由他们产生的流量，可能并非真实粉丝。

---

[1] 这里指去掉一模一样的重复转发，暂时未去掉同一微博 ID 但不同转发内容的重复转发

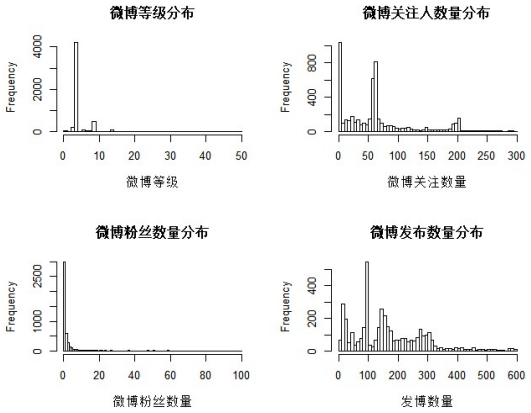


图 1：迪丽热巴总体数据分析

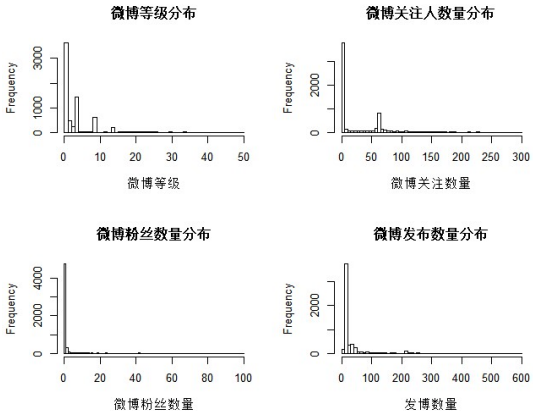


图 2：蔡徐坤总体数据分析

2、微博转发真粉假粉比例

于是，为了探究 6 位流量明星微博转发中假粉所占比例的情况，我们将数据集中**转发者微博账号的关注或者粉丝数少于等于 5、转发之后被点赞数评论数再转发数都为 0、微博会员等级为 0、微博账号等级小于等于 5 的数据，以及转发者的关注或者粉丝数大于等于 5 但昵称为“用户 XXXXXXXX”格式的样本**提取出来，作为假粉数据集，剩下的作为真粉数据集。

在计算了 6 位明星总体转发数量中的假粉比例后，我们发现：这六位明星的推广微博中有很大比例都是假粉在转发，平均假粉转发比例高达 76.12%，其中假粉转发比例最小的是朱一龙推广欧舒丹的沐浴露，只有 43.89%；而王俊凯推广纯悦矿泉水的假粉转发比例则醉倒，高达 91.62%。

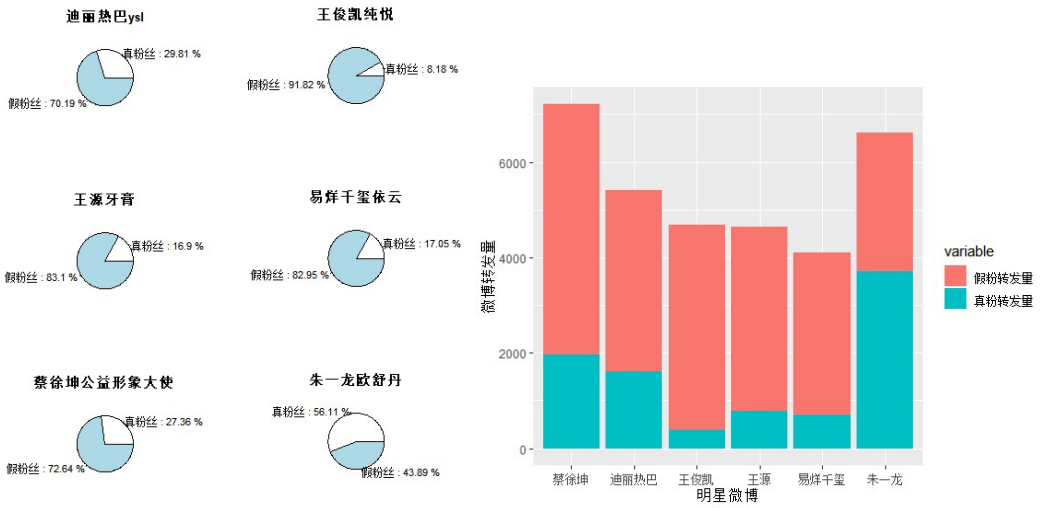


图 3：未去重时的真假粉转发数量比例

即使将真粉数据集和假粉数据集通过微博 ID 进行去重[2]，从探究总体转发中有多少是真假粉转发的到探究总体转发中有多少个真假粉丝后，我们发现，假粉比例的结果也相差不大。由此可看出，**明星推广产品的微博通过购买假粉来增加转发量的现象很普遍，而且这些的假粉占据了转发量的近 3/4，这些推广微博的转发中存在大量假流量，并不能反映一个明星的真实影响力。**

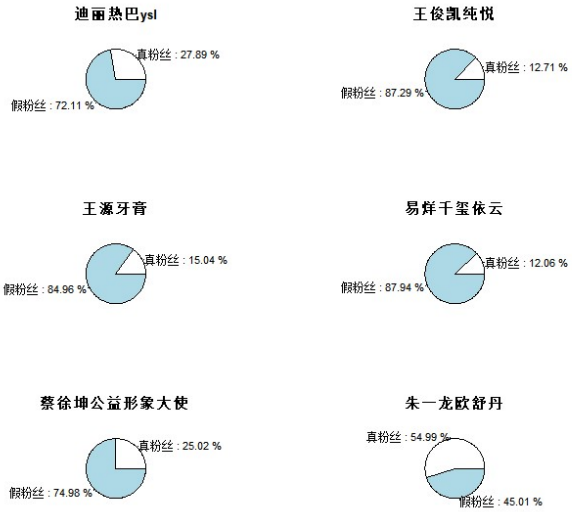


图 4：去重之后真假粉丝转发数量比例

[2] 去掉相同微博 ID 但是转发内容不同的重复转发微博

3、真假粉丝画像对比

为进一步刻画真假粉丝的画像，为商家提供识别真假粉丝的依据，我们对六个明星的真假粉丝数据集进行了分析，发现：真假粉丝群体在性别、转发设备、转发评论等方面都具有较为显著的差别。

3.1 性别

在真假粉丝画像的性别分析中，我们发现明星微博真粉丝和假粉丝性别比例存在较大差异。

真实粉丝群体中，蔡徐坤、王俊凯等男明星的粉丝都已女性为主，均高于 75%；而女星迪丽热巴则是女性略多于男性，占比 58.8%，这也符合我们追星群体以女性为主，男流量明星更吸引女性的日常认知。

而在**假粉丝群体**中，男性比例大幅增加，王俊凯、易烊千玺的男粉丝占比甚至高达 90%，女星迪丽热巴假粉中的男性比例也高于真实粉丝男性占比，这也进一步印证了我们的真假粉丝划分具有一定的合理性。

最后，我们将组合 TF-boys 中的三位影星王俊凯、王源、易烊千玺进行比较，发现三人虽属于同一组合，真实粉丝群体十分相近，但假粉类型却似乎并不相同，王源的假粉丝中有超过半数的女性粉丝，而另两位则都已男粉丝为主。

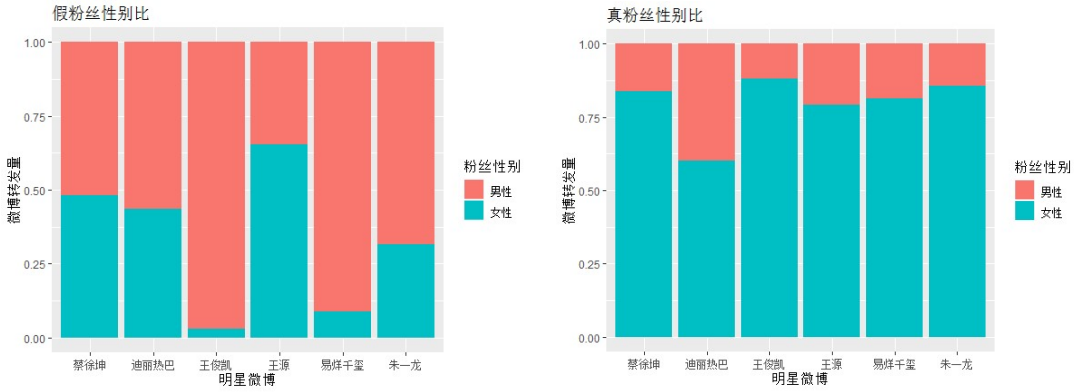


图 5：假粉丝性别比例（左）和真粉丝性别比例（右）

3.2 真假粉丝转发设备

通过对六位明星真假粉丝的转发设备分别作柱状图，可以看出不同明星的粉丝转发设备 top3 并不相同，同一位明星真粉丝与假粉丝群体的使用设备也均不相同。

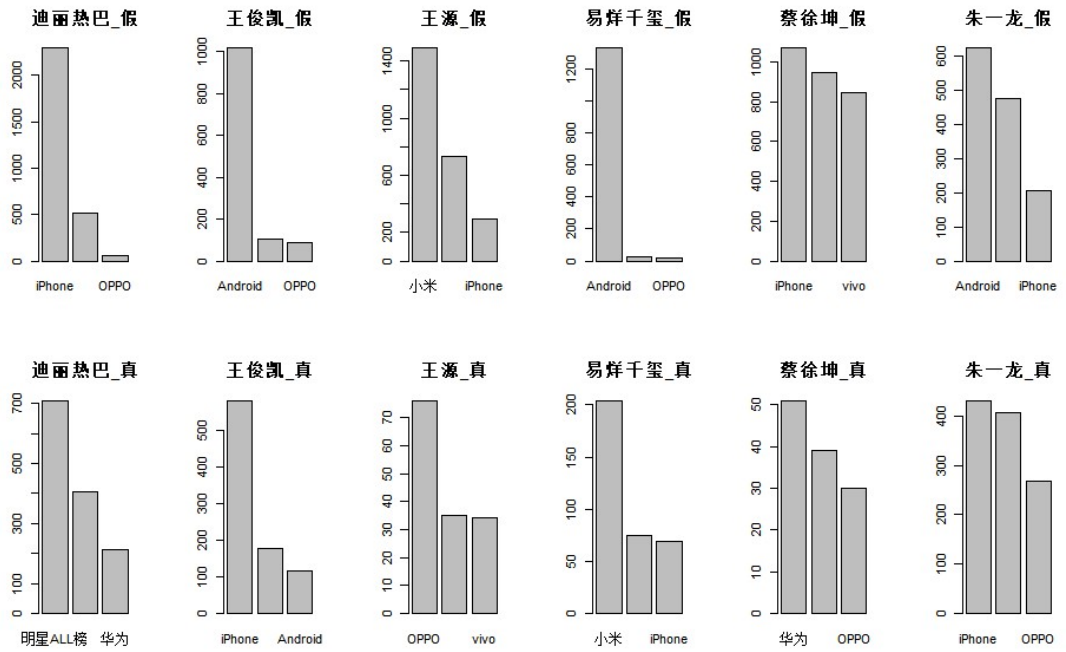


图 6：六位明星真假粉丝转发设备 top3

为进一步描述粉丝转发设备的特征，我们将这六位明星的真粉数据集合并在一起，假粉数据集合并在一起，重新做出了整体假粉和真粉的转发设备 top10。发现**假粉丝群体**使用安卓操作系统手机和 iPhone 为主，而 iPhone 则成为**真实粉丝**转发明星微博所使用的主要设备。假粉丝数据中显示的”Android”手机均为除华为等知名手机品牌外无法被识别具体品牌的安卓杂牌手机，这也反映了很大一部分假粉丝在转发时选择低成本转发设备的现象。

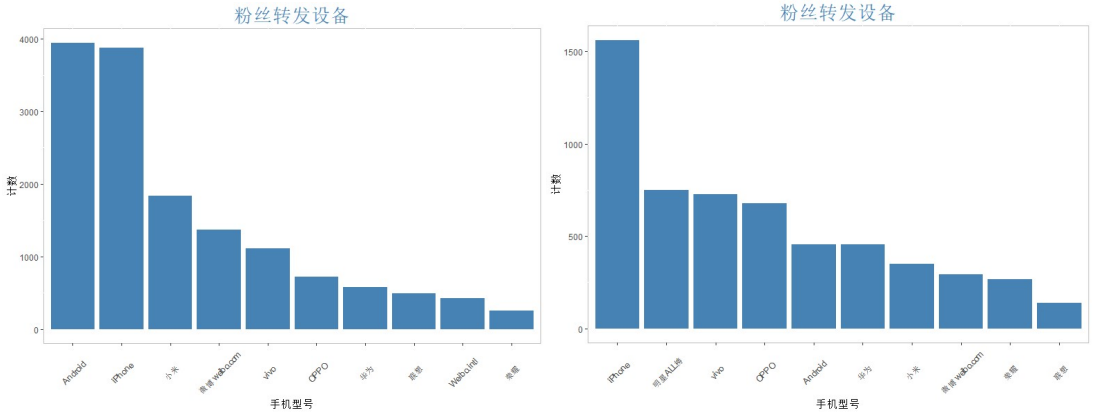


图 7：总体假粉丝（左）转和真粉丝（右）转发设备 top10

3.3 微博账号等级、发博数量、关注人数量、粉丝数量

利用上面 6 个明星合并在一起的假粉丝和真粉丝的总数据集，我们对假粉丝和真粉丝的微博账号等级、发博数量、关注人数量、粉丝数量对进行了统计，发现假粉微博账号等级平均 3，微博粉丝数量平均 1，微博关注人数量平均 31，而真粉丝的微博等级平均 12，微博粉丝数量平均 239，微博关注人数量 238，符合预期。

而后，我们画出了真粉丝和假粉丝的箱线图，在去除了 outlines 之后，可以看出，**真粉丝的数据分布较为分散，没有明显的偏向。**

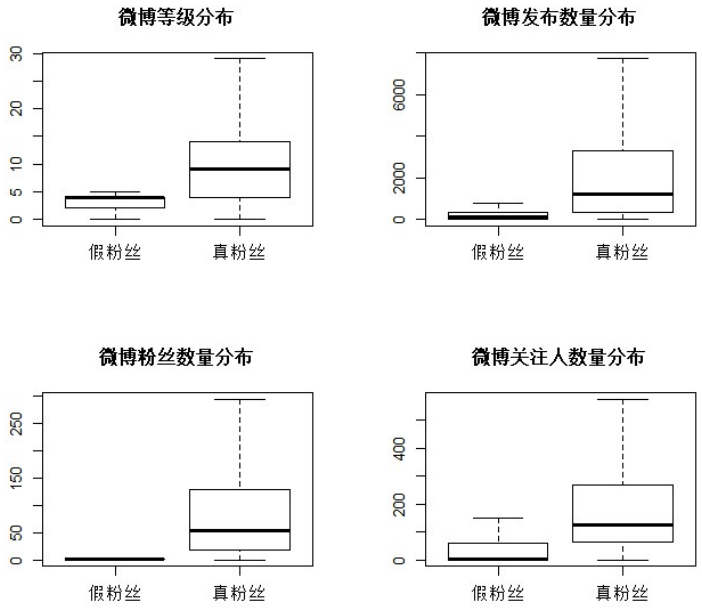


图 8：真假粉丝画像对比

3.4 转发评论、微博昵称、微博简介词云图

3.4.1 转发评论词云图

通过分别对明星真粉丝与假粉丝的转发评论作词云图，我们发现真粉丝与假粉丝在转发时的评论中都会提及偶像名字以及该微博主要推广对象，似乎并无明显区别；但如果仔细观察云图，并回顾原数据集，可以发现假粉丝转发评论多使用各种表情符号，如易烊千玺云图中的“羞嗒”，并常包括各种无关文字；真粉丝云图中，粉丝更关注偶像微博内容，且包含更多“超话”。

玺# #易烊千玺[超话]#炸燃舞台，万众瞩
@TFBOYS-易烊千玺

图 9：转发评论中的“超话”

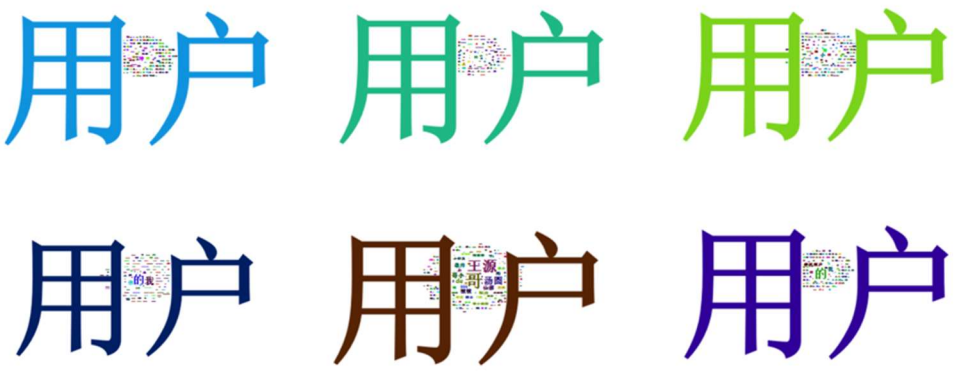
图 10：蔡徐坤真粉丝（左）与假粉丝（右）转发评论云图

图 11：易烊千玺真粉丝（左）与假粉丝（右）转发评论云图

3.4.2 微博昵称词云图

通过对粉丝微博昵称画词云图，可以发现真粉丝微博昵称是十分多样，很多真粉丝使用自己的偶像的名字或偶像的昵称作为自己微博昵称，除此之外，“火锅”“可爱”等各种生活词汇也多有出现。

图 12：真粉丝微博昵称词云图



假粉丝微博昵称的词云图中，“用户”二字占领绝对主导地位，而微博给予新用户的默认昵称正为“用户”后加一串数字的形式；于是我们尝试去掉“用户”二字再作云图，假粉丝昵称词云图中的确显示出大量数字串，我们以流量明星朱一龙的假粉丝为例进行展示。

图 13：假粉丝微博昵称词云图

图 14：去掉高频词“用户”的朱一龙假粉丝微博昵称词云图

3.4.3 微博简介词云图

在对 6 位明星的真假粉丝的微博简介做了词云图之后，我们发现，**真粉微博简介往往喜欢带自己偶像的姓名，而且会表现出对偶像强烈的喜欢的情感。**

图 15：真粉丝微博简介词云图

而**假粉丝绝大多数都没有微博简介**，比如我们爬取的易烊千玺和王俊凯的假粉样本，全部没有微博简介；而少有的几个微博简介都是些无关的字词，比如迪丽热巴和王源的假粉样本，都是些无关紧要的字词。

但是蔡徐坤和朱一龙假粉的微博简介引起了我们注意，乍一看他们的假粉的微博简介和真粉简介的词云图十分相似，而且朱一龙假粉简介中出现了“小号”这一高频词汇，让我们不得不猜想，难道这些假粉是高端定制类型的？竟然连微博简介也做的如此逼真？但是当我们回溯他们的假粉数据集才发现，其中 90%以上都是没有简介的正常假粉，但是其中约有**5%的他们的粉丝的追星小号因为满足我们的假粉要求被我们捕捉进假粉数据集**，所以他们的追星小号的微博简介在大多数空白的微博简介中被凸显出来，造成了和真粉相似的词云图，但是少数的误差并不影响我们对整体假粉的画像描述。

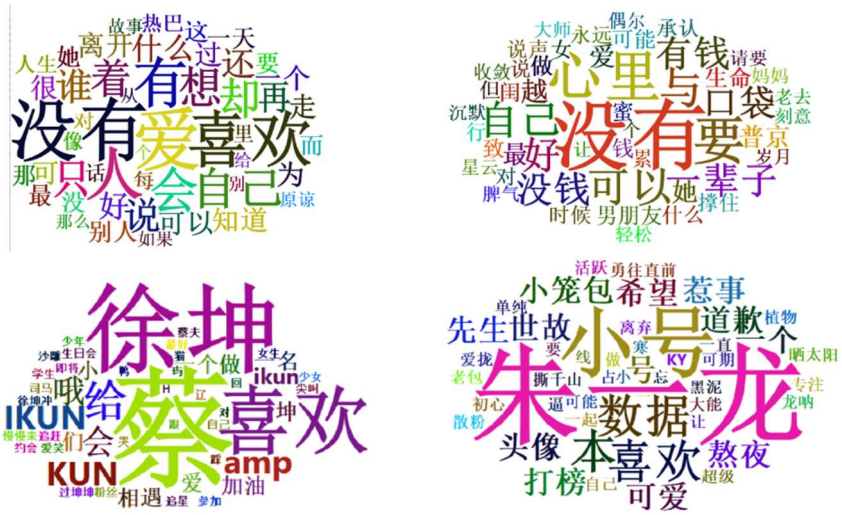


图 16：假粉丝微博简介词云图（分别是：迪丽热巴、王源、蔡徐坤、朱一龙）

3.5 真假粉丝重复转发倾向

之前我们提到，为了扩大偶像的影响力，假粉丝和真粉丝都有重复转发明星微博的行为。为了了解真假粉丝的重复转发倾向，我们利用未去重的真假粉丝转发总量和去重之后的真家粉丝转发数量计算出了 6 位明星真假粉丝的重复转发比例[3]。单独看每个人每条微博的重复转发倾向貌似没有显著性的规律，于是我们又另外爬取了这 6 位明星没人另外 5 条微博来研究他们每个人 6 条微博真假粉丝重复转发的平均比例。

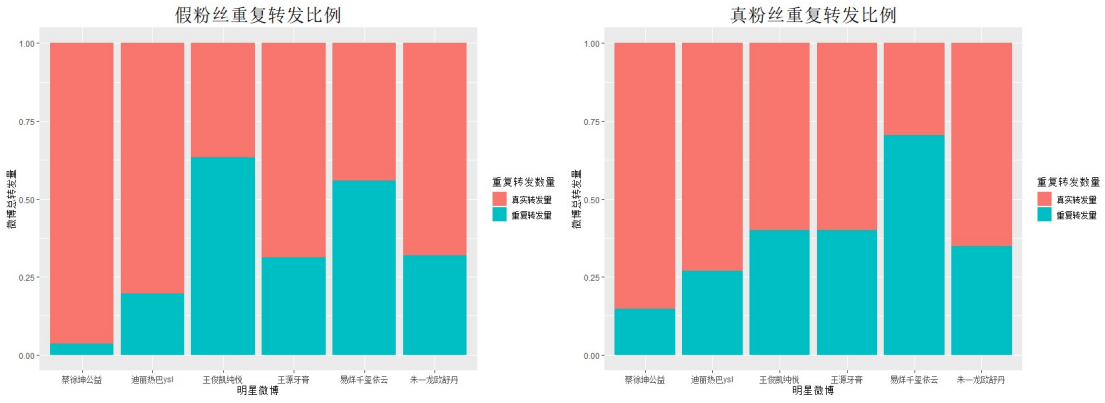


图 17：真假粉丝的重复转发比例

朱一龙	欧舒丹	迪丽热巴	ysl 香水	蔡徐坤	levis
	KFC		自拍		mv
	采访		东方卫视		公益
	电视剧		拉面		花衬衫
	戛纳		生日会		照片
	井然角色		杂志		新歌
王源	乐事	王俊凯	纯悦	易烊千玺	依云
	牙膏		碧浪		电影
	生活日常		Bolon		赛车
	唱作人作品		央视		时尚杂志
	新青年		生活日常		五一央视
	做饭		昆明		粤语

表 1：6 个明星的 6 条相关微博

[3] 重复转发分为一模一样评论内容的转发和不同评论内容的转发两种，由于我们爬取数据时将一模一样的转发清除掉了，所以这里的重复转发是指相同 ID 的不同评论内容的重复转发

在爬取了相似时间段和相似转发总量的 36 条微博之后，我们用同样的方法计算出了每个明星的真假粉丝重复转发比例，发现 6 位明星真粉丝的重复转发比例在 50%左右，即每四条转发中就有两条为粉丝的重复转发，而假粉丝的重复转发率依明星不同而不同。我们将六位明星的数据合并在一起后，发现真粉丝的重复转发比例为 52.1%，假粉丝在 36.4%，说明真粉丝比假粉丝有更高的重复转发倾向，为了帮自己的 idol 做数据，真粉丝还是很拼的。

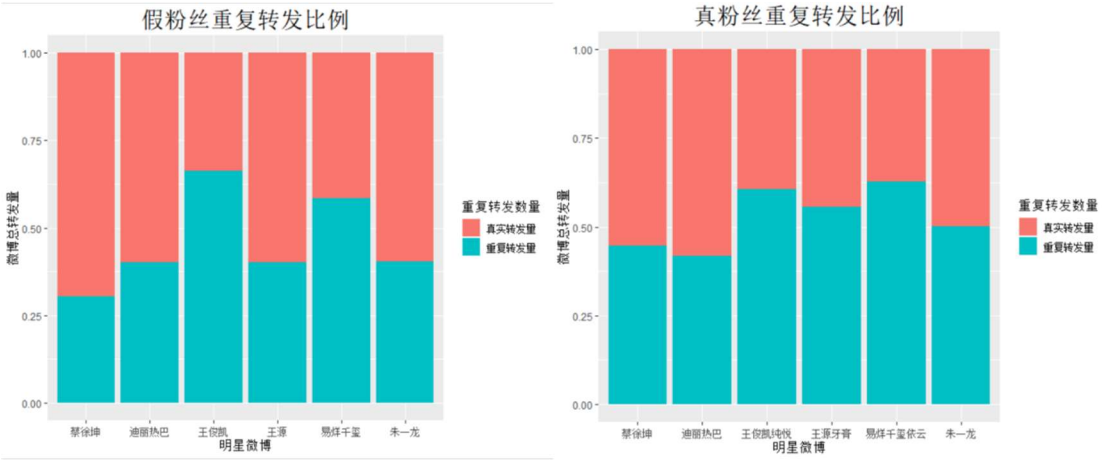


图 18：6 条微博平均真假粉丝重复转发比例

4、假粉 ID 重合度

在分析真粉假粉画像之后，自然而然，我们会思考，这些帮助明星转发不同微博的假粉会不会是同一个账号，于是我们分析了 6 位明星不同微博之间的真假粉 ID 的重合度。在对每位明星六条微博进行两两组合得到 15 组配对之后，我们计算了他们真粉假粉之间的 ID 重合度，发现真假粉重合度并无显著差异，当代流量明星存在部分持续关注其动态的忠实粉丝，也存在假粉丝帮同位明星不同微博多次刷流量的情况。

明星	15 组平均假粉 ID 重合度	15 组平均真粉 ID 重合度
朱一龙	9.51%	18.62%

迪丽热巴	6.81%	8.26%
蔡徐坤	14.75%	19.15%
王源	11.3%	8.26%
王俊凯	10.61%	9.57%
易烊千玺	3.12%	12.7%

5、总结

通过描述性分析发现，当代流量明星确实存在大量微博流量造假现象。真实粉丝和假粉丝在账号性别、转发设备、转发评论、微博昵称等多个方面都具有较显著差异。真实粉丝在微博账号上表现出更多的个性化，在转发偶像微博时展现出更好的针对性，且同样具有重复转发一条微博以帮助偶像增加人气的倾向；而假粉丝则具有更明确的刷量目的性，体现在微博账号昵称单一、转发评论常出现表情和无关字符等多个方面。