学校代码： 10246
学　　号： 14212010034

硕 士 学 位 论 文

（专业学位）

大规模社交网络用户访问控制策略与信息分享行为分析

A Large Empirical Analysis of Access Control Settings and Information Sharing Behaviors in Online Social Networks

院　　系：　软件学院

专业学位类别（领域）：　软件工程

姓　　名：　倪敏悦

指 导 教 师：　韩伟力　副教授

完 成 日 期：　2016 年 10 月 15 日

指导小组成员名单

韩伟力 副教授

Dr. Jun Pang (University of Luxembourg)

目 录

摘要

近年来，社交网络在全球范围内获得了巨大的成功。Facebook、Twitter 和 Instagram 等一系列社交网络已经成为了人们日常生活中不可或缺的一部分。与 Facebook 相对封闭的好友圈不同，在 Twitter 和 Instagram 这一类更开放的社交网站上，我们希望了解，用户在社交网络中的信息保护有哪些行为表现？用户在信息分享时有怎样的心理？如何改进现有的访问控制机制？

本文针对 Twitter 和 Instagram 这两大社交网站上用户的访问控制策略与信息分享行为两方面进行了实证分析研究，并对这一类社交网站上现有的访问控制机制提出相应的改进建议。

- 从 2015 年 10 月 15 日至 2016 年 1 月 12 日，我们通过调用 Twitter 和 Instagram 官方提供的 API，采集了 155,387 位 Twitter 用户与 282,066 位 Instagram 用户的相关数据。在此期间，Twitter 的隐私用户比例上升 0.73%，Instagram 的隐私用户比例上升 4.84%。其中，女性、亚裔、年轻用户更关注网络空间中的个人隐私保护。部分用户会频繁地修改个人主页访问控制设置，5.21%的 Twitter 用户平均修改 2.29 次，19.95%的 Instagram 用户平均修改 3.40 次。而开启了访问控制设置的用户具有较低活跃度、清理好友以及发布更具有隐私性的内容的行为特点。同时，重大事件与重要节日会影响用户对个人主页访问控制设置的决策。
- 本文以 Instagram 平台上的热门标签#like4like 的相关内容作为对用户信息分享行为分析的切入点。在采集了 143,586 名用户发布的 1,770,643 篇相关帖子后，针对这些数据进行实证分析结果显示：从 2012 年第一季度至 2016 年第一季度，参与用户人数增长 600 倍，人均发帖数量翻了一番，并仍然保持着上升的趋势，而用户的人群分布显示具有女性主导和低龄化的特性。热门标签使得用户发布的内容的受欢迎程度不断上升，平均每篇帖子获得的点赞数量由最初的 37.00 上升至 49.41，其中大约 80%的点赞来自于陌生人。但是点赞用户获得回赞的比例非常低，为 6.11%，其中陌生人获得回赞的比例仅为 3.38%。表明了用户倾向于通过热门标签提升关注度，但并不乐意遵从#like4like 标签的本意与点赞用户互动。
- 我们使用机器学习领域中的相关模型算法，对用户的个人主页访问控制设置进行了训练与预测，预测得到的 ROC（Receiver Operating Characteristic）曲线下面积最优结果为 0.70。这一结果从一定程度上验证了自动预测用户访问控制设置的可行性，也为社交网络中用户访问控制设置机制的研究提供了全新的见解和思路。

在全面认识用户在社交网络中的隐私保护与信息分享行为之后，本文提出，Twitter 和 Instagram 现有的全局访问控制机制已经不能很好地满足用户的隐私保护与信息分享的双向需求。因此我们建议在 Twitter 和 Instagram 这一类社交网站上，企业应提供更细粒度的访问控制机制，避免用户频繁地更改访问控制设置，造成隐私数据的泄露。

关键词：社交网络，访问控制，数据挖掘，机器学习
中图分类号：TP309

Abstract

In recent years, along with the rapid development of the Internet, online social networks have gained a huge success, among which, Facebook, Twitter and Instagram have become an important part of human daily life. Different from Facebook, users on Twitter and Instagram have much wider social circles. Thus, the questions like how users protect their data on OSNs; what users' thoughts are while sharing information online; and how we could improve the current access control mechanism, are still under-studied.

In this paper, we mainly focus on the analysis of users' access control settings and information sharing behaviors on Twitter and Instagram.

- From October 15th, 2015 to January 12th, 2016, we collect data of 155,387 Twitter users and 282,066 Instagram users through official APIs. During these three months, we find that the percentage of private users on Twitter increased about 0.73%, and that on Instagram increased about 4.84%. Female, Asian and younger users would pay more attention to online information protection. Meanwhile, some of the users would change their access control settings frequently. On average, 5.21% of Twitter users have changed 2.29 times, and 19.95% of Instagram users have changed 3.40 times. What's more, cleaning social circles, posting with certain topics and important festivals and events are possible reasons for users to change their access control settings.
- For users' information sharing behaviors, we collect data of 143,586 users related to popular tags on Instagram named as #like4like. From the first quarter of 2012 to the first quarter of 2016, the total amount of participated users increased by 600 times, and the number of posts per user doubled. Female and younger users are dominant users who published the posts with the tag of #like4like. According to our analysis, the tag #like4like has helped attract more users' *like*, which is an action made by a stranger or a friend when he or she sees the post with #like4like. The average number of *like* for each post increased from 37.00 to 49.41, and 80% of *likes* were from strangers. However, only 6.11% of users who liked the posts got like back, and the rate of strangers who got like back was much lower, which was 3.38%. The results indicate that users tend to use the popular tag #like4like to help increase the popularity of their certain posts, but will not follow what

the tag means to have further interactions with those users who liked their posts.

- We leverage machine-learning techniques to conduct a prediction on whether users would enable their access control setting or not. Our prediction achieve a promising result in which the area under the Receiver Operating Characteristic Curve is 0.70, which indicates a user's access control setting can be predicted to some extent.

Based on our findings, we suggest that the current access control mechanism on Twitter and Instagram cannot meet users' requirements. On the contrary, it leads to users' frequent changes of their access control setting, and potential risk to users' sensitive data online. The companies should seek a way to make their access control mechanism more fine-grained.

Keywords： Online Social Networks, Access Control, Data Mining, Machine Learning

CLC Number: TP309

第一章 引言

1.1 研究背景

在近十年的互联网发展中，社交网络迅速崛起，并在全球范围内获得了巨大的成功。如今，社交网络早已成为了人们日常生活的一部分，他们通过这些平台与家人、朋友，或者陌生人交流、沟通和分享各自的生活。在众多社交网络平台中，Facebook、Twitter 和 Instagram[1]异军突起，成为了全球首屈一指的社交网络大型平台，每天吸引着千万名来自世界各地的用户。数据显示，截止至 2015 年 9 月，Facebook 每月的活跃用户人数达到 15 亿，其中超过 83.5%的用户来自除美国和加拿大之外其他的国家和地区，而 Twitter 和 Instagram 每月也各自拥有超过 3 亿和 4 亿的活跃用户[2]。每天，有超过 5 亿条信息通过 Twitter 发布，Instagram 上也会新增 6 千万张新的照片[3]。虽然都是热门的社交网络平台，但是用户使用这些网站的目的却不尽相同。在 Facebook 上，用户更多地与自己现实生活中的同学、朋友、家人建立好友关系，通过发布状态、博文、照片来记录并分享自己生活的点滴。而 Twitter 更类似于新闻传播平台，Instagram 则只能发布单张照片或一段视频。

为了对庞杂的网络信息进行整合、归类，并便于用户进行搜索，标签功能如今已被各个社交网站所采用。在 Instagram 上，用户通过给自己发布的内容添加相应的标签来提升自己照片的浏览量，吸引更多陌生人的关注。目前，已有少数研究者通过对特定标签相关内容的分析来揭示其所反映出的用户某一行为的特点。Flavio 等研究者们关注到了自拍（#Selfie）在世界各地的兴起，许多人通过发布自拍来获取更多的关注[20]。而 Yelena 等人则注意到人们对各国食物的高度关注，通过对 Instagram 网站上#foodporn 标签的研究来分析用户希望传播和分享的内容[28]。但是这两项研究都具有特定主题的针对性，不能有效地反映用户希望与更多其他用户发生互动的社交需求。此时，另一个极具特色的标签进入了我们的视线，#like4like。我们发现，#like4like 标签从 Instagram 上线初期出现，至今在 Instagram 网站上已有 2 亿多篇相关的帖子，成为排名第二的热门标签。这个标签意为 like for like，字面意思表达了用户在获得其他用户点赞之后将会回赞的意愿，表现了用户希望通过这个标签获取更多的点赞、结交更多朋友的想法。我们认为，这是研究用户在社交网站上信息分享时心理活动的一个很好的切入点。

然而，社交网络为用户的社交生活带来极大便利的同时，用户的个人隐私信

[1] Alexa 世界网站排名中，Facebook, Twitter 和 Instagram 分别位列第 3、第 10、第 15。
[2] 数据来源参考：http://newsroom.fb.com/company-info，访问于 2016 年 2 月。
[3] 数据来源参考：http://bit.ly/1Fij4er，访问于 2016 年 2 月。

息也面临着更大的安全隐患。已有研究发现，用户在社交网络上发表的每一条状态和每一次签到信息都能够或多或少地对外透露出用户的个人信息[18][25]。即使用户不公开展示个人信息栏，攻击者仍然能够通过用户发布的文字信息或者照片信息寻找蛛丝马迹，从而获取用户的隐私信息。而为了帮助用户，尤其是那些不具备信息安全专业知识的普通用户，各大社交网站都结合了各自网站的特点实施了不同粒度的访问控制机制。Facebook 提供了最细粒度的访问控制机制[4]。用户可以将好友分类，并且可以对每一条发布的内容设置不同的访问控制策略，控制阅读每一条内容的目标好友，比如只有家人能够访问家庭出游的旅行照片等。而 Twitter[5]和 Instagram[6]都提供的是全局的访问控制机制，一旦用户开启了个人主页的访问控制设置，除了该用户的个人头像和包括好友数量和发布内容的数量在内的基本数据信息，他们的所有内容都对陌生人不可见；并且陌生人必须只有在用户通过了他的好友申请之后才能进行关注。为了在现有基础上更进一步提升社交网络中的访问控制策略的使用效率与效果，近年来，学术界也对各类访问控制模型展开了大量的研究。许多研究人员将研究重点放在为社交网络访问控制机制建立新的适用模型，以及寻求能够精准定义社交网络策略的方法。Philip W. L. Fong 等人提出了一种新的二阶访问控制模型[10]，即假设用户 A 想要访问用户 B 的所有共享资源，那么用户 A 与用户 B 必须先建立起一定的社交关系。Barbara Carminati 等人提出了包含三方面的可以被应用于社交网络访问控制模型的规则[4]，包括用户之间的社交关系、社交网络中两个用户之间的距离，以及置信度。他们进一步提出了通过语义网技术来定义访问控制策[3]。Glenn Bruns 等人则提出将混合逻辑作为访问控制策略设置的语言，并进一步验证了这一语言的强大[2][9][11]。而在后人的研究中，混合逻辑也被不断地提及以及使用[6][7][14][15][17][22]。

尽管有一些研究者也已关注到对真实用户行为研究的重要性并展开了相应的分析，但是几乎所有的研究都仅针对 Facebook 平台，并且他们都仅通过数千名用户来研究用户的访问控制策略的使用[12][8][13][21]。数据量小而局限，不能很好地反映整体趋势。同时，正如我们之前所说的，用户对各个社交网站的使用目的不同，因此对于大部分普通用户而言，Facebook 上的好友圈反而具有更高的封闭性，分享的内容也相对更加私密；而 Twitter 和 Instagram 上的信息却会在陌生人中有更高的曝光度，用户本身也更容易与陌生人建立社交网络好友关系，好友圈关系更加开放与复杂。

[4] Facebook 的访问控制机制可参考：:https://www.facebook.com/help/325807937506242.
[5] Twitter 的访问控制机制可参考：https://support.twitter.com/articles/14016.
[6] Instagram 的访问控制机制可参考：https://www.instagram.com/developer/endpoints/.

1.2 研究内容及意义

社交网站上的访问控制机制已经成为了研究热点之一，但是，其中的大部分研究都是针对 Facebook 平台。而如第 1.1 节中所指出的，Twitter 与 Instagram 无论在用户的使用目的，还是网站的访问控制机制上，都完全不同于 Facebook。因此，本文首次专注于 Twitter 与 Instagram 两大社交网站，提出并通过实证分析回答以下三个问题。

- 用户在社交网站中的信息保护有哪些行为表现？

本文运用数据可视化以及自然语言处理领域中的相关技术，从静态与动态两个角度，对超过 15 万 Twitter 用户和将近 30 万 Instagram 用户访问控制设置的情况进行了数据分析与挖掘，掌握了真实用户在不同的访问控制设置下的行为特点，以及影响用户个人主页访问控制设置决策的潜在因素。

- 用户在信息分享时有怎样的心理？

全面了解用户的行为特点、明白用户的隐私保护与信息分享的双向需求，是设计真正契合用户需求的访问控制机制的必要条件。本文以 Instagram 网站上排名第二的标签#like4like 作为对用户信息分享行为分析的切入点，采集了超过 14 万 Instagram 用户发布的相关内容，运用数据可视化技术，结合社会学的相关理论，分析了用户在信息分享时的心理并得到相应结果。

- 如何改进现有的访问控制机制？

本文在以上两方面分析结果的基础上，针对现有的社交网络访问控制机制，向社交网站服务提供方提出了合理的建议。并应用机器学习领域中的逻辑回归模型、随机森林模型与梯度推进模型，量化评估了自动预测用户社交网站上个人主页访问控制设置的可行性。

与前人的研究相比，本文是首个对大规模真实用户在社交网络中行为表现的研究，并且量化地分析了用户信息保护与信息分享两种截然相反的行为特点，掌握了用户在社交网站上信息保护与共享的双向需求。另外，本文向社交网站服务提供方提出了有意义的建议，并首次提出了用户个人主页访问控制设置自动预测的方法，为社交网站访问控制机制的设计提供了全新的见解和思路。

1.3 论文的组织结构

本文的第二章将介绍本文所使用的相关技术，包括数据可视化技术、自然语言处理和机器学习分析模型等三方面的主要分析技术，以及在数据采集过程中运用到的相关技术。第三章具体介绍数据集的采集和预处理。第四章介绍了对社交网络用户访问控制设置情况的分析结果，并对自动预测用户访问控制设置进行了建模，对其可行性进行了量化评估。第五章阐述对社交网络用户信息分享时的心

理表现的分析结果。第六章对分析实验中的限制进行了讨论，并给出了针对现有访问控制机制的改进建议。最后，第七章总结全文，并对未来工作进行展望。

第二章 相关技术综述

2.1 数据可视化

数据可视化技术在大数据时代中的地位显得尤为重要与突出，而数据可视化也已经成为了一个异常热门的研究领域。数据可视化，实质就是将大量数据通过科学又美观的呈现方法，清楚并高效地转化为人类可读且有意义有价值的信息的过程。因此，数据可视化是一项集合了统计学、图形学、艺术美学等多个跨学科领域的现代科学。在本文中，我们主要使用了盒图与词云这两个目前已经比较成熟且流行的数据可视化方法，来对我们的研究数据进行可视化的展现，帮助我们得到有价值的结果。

2.1.1 盒图

盒图（Boxplot）由美国统计学家 John W. Tukey 发明，是一种通过标识出数据集合中四分位数的方式来展现数据离散程度的数据可视化方法。

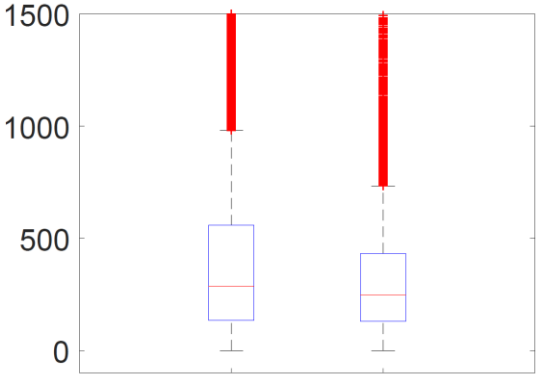

图 2.1 盒图 Boxplot

图 2.1 展示了盒图的一种样式。盒图主要由五个数据点构成，它们分别是：最小观测值（minimum）、下四分位数（1st quarter， Q1）、中位数（medium）、上四分位数（3rd quarter， Q3）、以及最大观测值（maximum）。其中，下四分位数、中位数与上四分位数组成一个带有分隔线、形状像盒子的长方形，因此将此类图表称为盒图。盒图的另外一个重要组成部分是盒主体到最小观测值与最大观测值之间的两条延伸线，这两条延伸线称为“胡须（whisker）”，所以盒图又称盒子-胡须图（box-and-whisker diagram）。

盒子与两条胡须已经能够表示数据集中的大部分数据，但是由于现实世界中获取到的数据往往存在着大量的离群数据（outlier），所以为了防止这些离群数据影响整体数据集的分布情况，导致整体特征偏移，盒图将所有的离群数据单独在图上绘出，有效避免了数据分布偏移的发生。同时，因为受离群数据的影响，

盒图的最小观测值与最大观测值可能不是整个数据集的最小值和最大值，而是在去除了所有离群数据点之后获得的最小值与最大值。

“胡须”的长度决定了盒图展示结果的有效性，理论而言，合适的“胡须”长度一般为四分位距的 1 至 1.5 倍。四分位距（IQR，Interquartile Range 的简称），又称四分差，即上四分位数和下四分位数之间的差值$\mathrm{IQR} = Q_3 - Q_1$，是描述统计学中的一种方法，可用来描述统计资料中各个变量的离散程度。在盒图中，盒子的长度就代表了一个四分位距。所以，以“胡须”长度为 1.5 倍四分位距为例：

1) 最小观测值为 $Q_1 - 1.5 * \mathrm{IQR}$。如果存在离群数据小于最小观测值，那么“胡须”的下端即为最小观测值，离群数据单独绘出；反之，“胡须”的下端就是最小观测值，也是整个数据集的最小值。
2) 最大观测值为 $Q_3 + 1.5 * \mathrm{IQR}$。如果存在离群数据大于最大观测值，那么“胡须”的上端为最大观测值，离群数据单独绘出；反之，“胡须”的上端就是最大观测值，也是整个数据集的最大值。

盒图能够有效地帮助识别数据集中的离群数据，并通过观察盒子的长度、分隔情况以及“胡须”的长度来判断数据集数据的离散程度和偏向，而盒图的一大优势在于可以通过两个数据集的盒图来比较两者的分布情况。由于受到曼•惠特尼 U 检验[29]的启发，我们在本文中使用盒图来对两个数据集分布情况及其均值的差异进行比较，得到有意义的统计比较结果。

2.1.2 词云

词云（word cloud），又称标签云（tag cloud），是一种针对文本信息的数据可视化方法。词云实际是一个带有权重信息的列表，而它的起源可以追溯到上个世纪末，加拿大小说家 Douglas Coupland 的小说 Microserfs 中最先出现了一个带有权重的英语关键词列表。词云真正成为一项数据可视化技术被广泛运用到实际研究中则是在 21 世纪初期互联网 2.0 时代到来之后，它的最初目的就是为了对大量互联网中产生的文本数据信息进行可视化。

词云大多由多个单词组成。每个单词的权重，即为该词在整个数据集中出现的频次，在词云中通过单词的大小来表示单词的权重。而字体颜色则可以表示单词的重要性。图 2.2 展示了词云的一种呈现结果。词云这一类数据可视化方法的最大优势在于，用户可以从词云中迅速定位到出现频次高或重要性高的单词。本文中，我们主要使用词云来展现用户在社交网络 Instagram 中标签的使用次数的情况。

图 2.2 词云 Wordcloud

2.2 自然语言处理

自然语言处理（Natural Language Processing，简称 NLP）是一个将计算机与人类语言之间的交互作为研究对象的，介于计算机科学、人工智能、以及计算语言学之间的分支领域。虽然关于自然语言处理的研究仍然处于一个比较初期的阶段，但是已经有不少的研究人员提出了一些高效的、并且可靠的自然语言处理模型和算法，使得计算机能够识别、理解并且处理人类语言。在本文中，我们利用自然语言处理领域中的隐含狄利克雷分布模型，来对用户在社交网络中发布的信息进行文本主题的有效提取与归类。

隐含狄利克雷分布（Latent Dirichlet Allocation，简称 LDA），由 David Blei、吴恩达等人在 2003 年提出[1]，是自然语言处理领域中的一种主题生成模型，它可以将每篇文档的主题按照概率分布的形式给出。

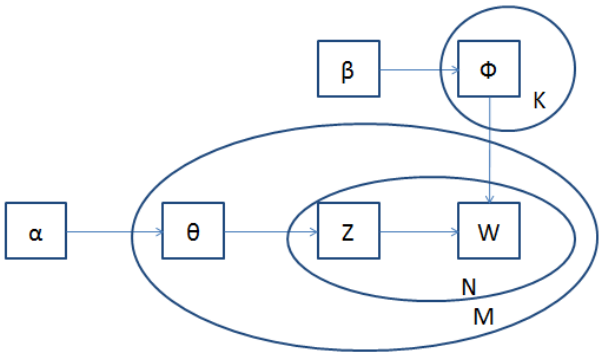

图 2.3 LDA 贝叶斯网络结构

LDA 模型是一种计算机自动挖掘输入文章中隐藏的话题的方法。它的基本原理是，对于一篇输入的文章，计算机不需要理解每一个单词之间的语义与逻辑关系，而只看作是一组单词构成的集合。这个集合可以包含多个主题，集合中的每一个单词都可以被归入其中一个主题。类似于 Beta 分布是二项式分布的共轭先验概率分布，LDA 是一种多项式分布的共轭先验概率分布。因此，正如图 2.3

中 LDA 贝叶斯网络结构所描述的，使用隐含狄利克雷模型生成文档主题可以分为以下 4 步：

1) 输入文档 X，从狄利克雷分布 α 中取样生成文档 X 的主题分布 θ_x；
2) 从主题的多项式分布 θ_x 中取样生成文档 X 中第 y 个词语的相关主题 $T_{x,y}$；
3) 从狄利克雷分布 β 中取样生成主题 $T_{x,y}$ 的单词分布 $\phi_{Tx,y}$；
4) 从单词的多项式分布 $\phi_{Tx,y}$ 中采样最终生成单词 $W_{x,y}$。

由此可以得到整个模型中的所有可见变量以及隐藏变量的联合分布为：

$$p(W_x, T_x, \theta_x, \phi|\alpha, \beta) = \prod_{y=1}^{N} p(\theta_x|\alpha)p(T_{x,y}|\theta_x)p(\phi|\beta)p(W_{x,y}|\theta_{T_{x,y}}) \quad (1)$$

最终，一篇文档的单词分布的最大似然估计可以通过公式（1）中的 θ_x 以及 ϕ 进行积分和对 T_x 进行求和得到：

$$p(W_x|\alpha, \beta) = \int_{\theta_x} \int_{\phi} \sum_{T_x} p(W_x, T_x, \theta_x, \phi|\alpha, \beta) \quad (2)$$

之后可以根据$p(W_x|\alpha, \beta)$的最大似然估计，再进一步通过吉布斯采样等方法估计出模型中的参数。

2.3 机器学习分析模型

机器学习是计算机科学的分支领域之一，它结合了模式识别与人工智能中的计算学习理论。1959 年时，Arthur Samuel 将机器学习定义为“是一个为了赋予计算机在没有编码的前提下自主学习能力的研究领域”[7]。机器学习领域的主要研究目标是探究基于数据的学习与预测算法。本文中，我们使用了 3 个目前机器学习领域中较为成熟的模型来对社交网络用户访问控制设置情况进行预测，这三个模型分别是逻辑回归模型，随机森林模型和梯度推进模型。

2.3.1 逻辑回归模型

1958 年，英国统计学家 David Cox 提出了逻辑回归模型（Logistic Regression），之后一直被广泛应用于二元预测的问题解决中。逻辑回归模型将一系列预测变量（又称独立变量）作为输入，输出二元问题中二元结果的预测概率。因此，逻辑回归模型可以用来评价输出值与输入值之间的关联度。

逻辑回归模型因其简单高效的特点，是目前在机器学习领域中被广泛运用的分析模型，逻辑回归模型的实质就是在线性回归模型的基础之上，再应用一层 Sigmoid 函数，又称逻辑函数。

$$g(z) = \frac{1}{1 + e^{-z}} \quad (3)$$

[7] Phil Simon 在 2013 年发表的书籍《Too Big to Ignore》中提及。

对于线性边界的情况，边界形式为：

$$\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n = \sum_{i=1}^{n} \theta_i x_i = \theta^T x \quad (4)$$

最终获得构造函数 h 为：

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} \quad (5)$$

在二分逻辑回归问题中，函数$h_\theta(x)$的值的含义为结果取 1 的概率，所以对于某一特定输入，得到预测结果类别 1 的概率为$h_\theta(x)$，类别 0 的概率为 1-$h_\theta(x)$。

2.3.2 随机森林模型

在机器学习领域中，另一种被广泛用来解决回归以及分类问题的模型是随机森林模型。随机森林模型不同于决策树模型的一点在于，它在训练的过程中会生成多棵不同的决策树，而非唯一的决策树。因此，随机森林分析模型最终的输出针对两类不同的问题也会有所区别。对于分类问题，随机森林中的每一棵决策树会先给出各自的分类判断结果，之后对所有的结果进行统计，将最多的结果作为最终的输出值；而在回归问题中，随机森林的输出则会是所有决策树输出结果的平均值。

决策树模型在机器学习领域中已经是一个比较成熟且应用极其广泛的模型算法，在输入不同变换形式的特征值的条件下，决策树的训练仍然能够保持其鲁棒性来排除不相干因素的干扰，并且生成稳定的决策树模型。但是，决策树模型的不足在于他的准确率相对较低。而另一点值得注意的是，因为在训练的过程中，单一的决策树会为了顾及数据集中数据表现出的各个方面的特点而学习许多不规则的模式并加入决策树的分支中，从而导致生成的决策树深度过大。在这种情况下会导致生成的决策树有大方差、低偏差的特点，此时生成的决策树其实是对训练集数据的过拟合。针对单一决策树模型的这一劣势，随机森林模型很好地解决了这个问题。随机森林模型通过学习训练集数据的不同方面来生成具有针对性的不同的决策树模型，以此来降低数据的方差。这样做虽然会在一定程度上增加数据的偏差，但是却可以大大提升随机森林模型的性能，并有效解决单一决策树导致的过拟合问题。

所以在随机森林中，每一棵决策树的训练规则是：

1) 假设训练集中的数据样本总共有 N 个，然后通过重复多次抽样来获得 n 个样本作为生成决策树的训练集；
2) 如果有 M 个输入变量，每个节点都将随机选择 m（m<M）个特定的变量，并以此 m 个变量来确定最佳的分裂点。在整个决策树生成的过程中，m 保持不变；
3) 任其最大限度地生长，不对任何一棵决策树进行剪枝；

4) 最后，对所有决策树的输出结果进行加总获得最终的输出结果，即新的预测数据。

2.3.3 梯度推进模型

梯度推进模型也是机器学习领域中一种被用来解决回归和分类问题的比较多见的模型，它会生成一个类似于决策树的预测模型。梯度推进算法，和其他推进算法相类似，拥有推进的不同阶段，又称为迭代。但是梯度推进算法不同的地方在于，它在迭代的时候选择梯度下降的方向来保证最后的结果最优。

模型的损失函数可以被用来描述模型的可信程度，假设模型没有过拟合，那么损失函数越大，就会导致模型的错误率越高。所以如果建立的模型能够让损失函数持续下降，那么模型就能够在这一过程中得到不断地改进，换言之，即使得模型的错误率不断下降。而最好的方式就是让损失函数在其梯度方向上下降。算法 1 展示了梯度推进算法的具体推进流程。在整个算法推进的过程中，我们一共训练 F_0 到 F_m 总共 m 个基学习器，而后沿着梯度下降的方向不断更新ρ_m和a_m来对模型的函数直接进行更新，从而利用参数可加性进一步推广到整个函数空间。

算法 1 梯度推进算法

$F_0(x) = \arg min_\rho \sum_{i=1}^{N} L(y_i, \rho)$

for m = 1 to M do

$\tilde{y}_i = -[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}]_{F(x)=F_{m-1}(x)}, i = 1, N$

$a_m = \arg min_{a,\beta} \sum_{i=1}^{N} [\tilde{y}_i - \beta h(x_i : a)]^2$

$\rho_m = \arg min_\rho \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \rho h(x_i : a_m))$

$F_m(x) = F_{m-1}(x) + \rho_m h(x : a_m)$

end for

end

2.4 数据采集与处理技术

由于本文的实验对象为 Twitter 与 Instagram 两大社交网站上真实用户的行为数据，而目前没有任何可直接获得数据的渠道。因此，本文中所使用的全部数据都是通过调用社交网站服务提供方官方提供的数据访问接口，以及其他数个不同服务提供的接口，来采集并处理后得到的。在本节中，我们将简单介绍本文中所运用到的主要数据访问接口。

2.4.1 社交网站用户数据采集

Twitter 和 Instagram 都各自提供了遵循 OAuth 协议的官方数据访问接口来对

外部用户提供有限的数据资源。

在 Twitter 上，我们通过调用 Streaming API[8]来采集用户的地理位置信息并进行筛选获得目标种子用户。Streaming API 向开发者提供 Twitter 在全球范围内的流数据。并且开发者可根据实际情况调用不同的流接口来实现针对不同用例的数据采集。Streaming API 共有三种不同的流接口，分别是公开流、用户流和站点流，适用场景如表 2.1 所示。

表 2.1 Twitter Streaming API 的三种不同流接口

	适用场景
公开流	可获取 Twitter 网站上所有的公开数据流。适用于针对某些用户或话题的数据采集，以及数据挖掘。
用户流	适用于获取单一用户的数据流，可返回特定用户的所有数据。
站点流	针对获取多个用户数据流的场景。

除了 Streaming API，Twitter 还提供了另一类 REST API[9]以供开发者调用来获取相应的数据。REST API 提供了一系列的 GET 方法，开发者可以通过调用不同的方法，获取指定的数据。以 GET statuses/user_timeline 为例，开发者可通过此方法输入特定用户的数字帐号或屏幕名称获取该用户的已发表的帖子数据，例如，用户 Alice 的 Twitter 数字账号为 13579，那么通过传入参数 user_id=13579，开发者可获得以 JSON 文件格式返回的 Alice 的最新 3200 篇帖子数据。

Instagram 也提供了类似的 REST API[10]，其中包括一系列 GET 方法，开发者亦可通过调用不同的方法获取以 JSON 文件格式返回的指定数据。

2.4.2 人脸识别

由于各个社交网站上访问数据接口的权限限制，我们无法直接获取到用户的任何个人属性的数据，于是我们转而使用了一项开源的人脸识别服务 Face++[11]来获取用户个人属性信息。

Face++™ 是北京旷视科技有限公司旗下的新型云端视觉服务平台，旨在提供一整套世界领先的视觉技术服务。Face++利用数据挖掘与深度学习等尖端技术，提供包括人脸检测、人脸识别以及面部分析在内的三项核心视觉服务。我们应用 Face++提供的 DETECT API[12]，将包含用户脸部形象的图像作为输入，进行个人属性信息的转换，转换返回 JSON 格式的数据文件，其中包括用户性别、种族（亚裔、非裔、白人）、年龄以及各项数据的置信度。在返回的 JSON 数据文件中，

[8] Twitter Streaming API，可参考：https://dev.twitter.com/streaming/overview.
[9] Twitter REST API，可参考：https://dev.twitter.com/rest/public.
[10] Instagram REST API，可参考：https://www.instagram.com/developer/endpoints.
[11] Face++，可参考：http://www.faceplusplus.com/.
[12] Face++人脸识别 DETECT API，可参考：http://www.faceplusplus.com/detection_detect/。

还包括了人脸五官分布、笑脸识别、图像基本信息等一系列其他的数据，对人脸的识别非常全面到位。

Face++是目前广泛应用的、比较可靠的人脸识别工具，这项工具已经赢得了不少国际赛事的奖项，也被许多其他的科学研究者所应用来获取用户的个人属性信息[19][20]。

2.4.3 地理位置信息转换

Google 地图向开发者提供了地理编码与反向地理编码的服务。地理编码是将人类可读的文本地址转换为经纬度地理坐标的过程，而反向地理编码则是将地理坐标转换为文本地址的过程。本文中通过调用 Google 地图提供的 Geocoding API 实现将经纬度地理坐标转换为具体国家、地区等文本信息的反向地理编码的数据处理。

开发者可通过 HTTP 或 HTTPS 访问 Google 地图的 Geocoding API，当转换数据包含敏感用户信息时，应当通过 HTTPS 访问该服务。要实现反向地理编码，应提供以下两种传入参数中的任意一种：经纬度地理坐标 latlng，或者地点 id。在完成了数据转化处理之后，文本数据将以 JSON 文件格式被返回，开发者可进一步处理获取所需的地址类型，例如仅保留国家或地区信息，或者具体到街道地址的信息。

第三章 数据准备

3.1 社交网站用户访问控制设置数据准备

对于社交网站用户访问控制设置情况的分析，我们将目标用户设定为美国纽约地区的用户。虽然纽约用户不是一个在全球范围内的随机采样样本，但是 Ratan Dey 等人的研究结果表明，纽约用户具有高多样性与代表性[8]，所以对于纽约用户的分析能够从一定程度上反应全球用户的趋势。

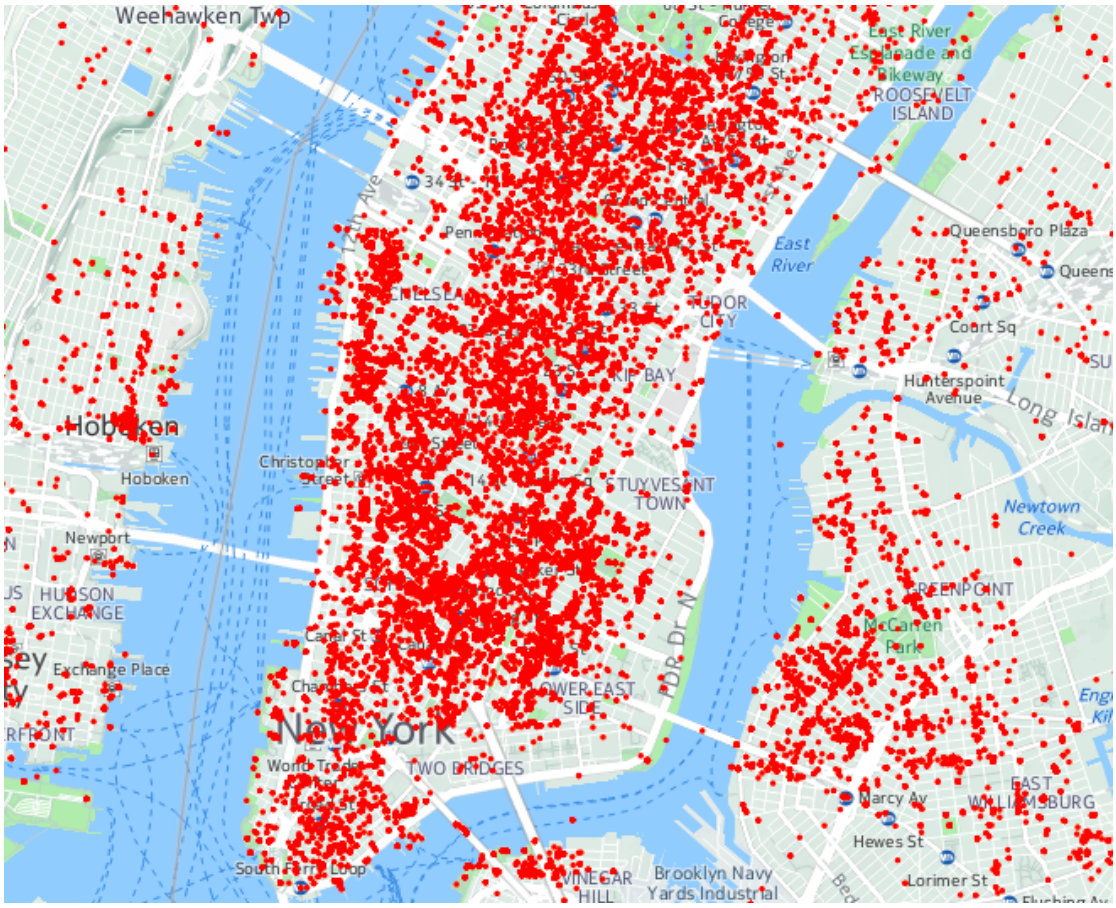

图 3.1 Instagram 纽约用户签到分布图

由于我们无法通过官方提供的 API 获取到 Twitter 和 Instagram 上每一个用户的所在地信息，我们使用用户分享在社交网站上的签到数据来对用户人群进行所在地定位。目前，随着移动设备例如智能手机、平板电脑等的大规模普及，越来越多的用户会使用移动客户端来登录社交网站，有数据显示，80%的 Twitter 活跃用户来自于移动端[6]。为了适应这一趋势，主流的社交网站都在各自的移动端程序中添加了利用手机 GPS 定位传感器获得的地理位置进行地理位置分享的功能，Twitter 和 Instagram 也不例外。因此，我们运用 Twitter 的 Streaming API 和 Instagram 的 REST API 来分别采集两大社交网站上用户的地理位置信息，并筛选

出纽约地理坐标范围内的所有用户。为了尽可能确保最终作为实验对象的所有用户都是纽约市的常住居民，我们仅保留拥有 10 条以上在纽约范围内签到信息的用户。图 3.1 展示了 Instagram 上纽约样本用户签到的地理位置情况。

在获得了种子目标用户之后，从 2015 年 10 月 15 日至 2016 年 1 月 12 日连续 3 个月的时间内，我们通过调用 Twitter 和 Instagram 各自提供的 REST API，对用户每天的访问控制设置数据以及其他相关数据信息进行采集，包括关注用户数量、粉丝数量、发布内容数量、点赞数量等。同时，为了确保我们的目标用户都是普通用户以及实验结果不受公众效应的影响，我们过滤了所有明星名人以及传销公众账号的信息。在此，我们根据整体的数据分布，将拥有 2000 名以上粉丝的帐号定义为明星名人帐号，而将关注用户数量比粉丝数量多超过 1000 人的帐号定义为传销公众账号，并将这些账号全部删除。

最终，我们使用 2015 年 11 月 12 日的数据作为静态分析的实验对象[13]，共有 175,202 名 Twitter 用户以及 292,406 名 Instagram 用户。而将那些在我们采集数据周期中始终活跃的用户作为动态分析的实验对象，共有 155,387 位 Twitter 用户，282,066 位 Instagram 用户[14]。

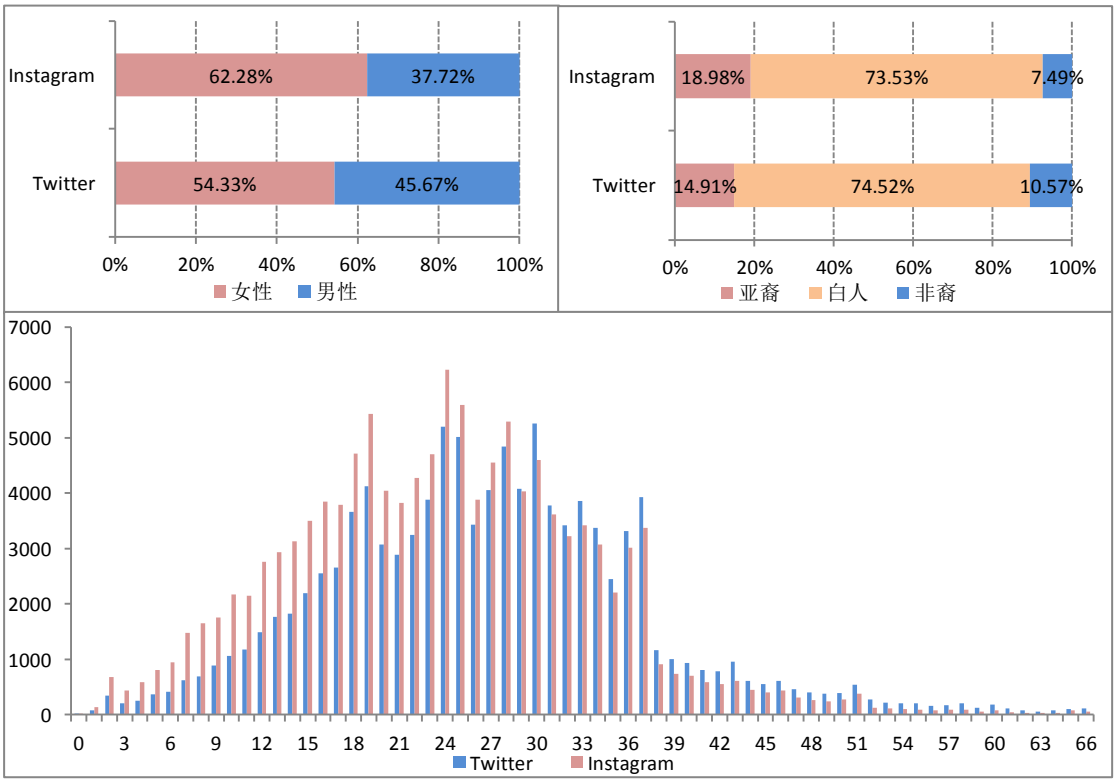

图 3.2 Twitter 和 Instagram 用户个人属性分布图

由于 Twitter 与 Instagram 都未提供能够获取用户基本属性信息的 API，而用

[13] 根据我们对其他日期用户数据的分析结果，对于单日用户的行为表现与趋势都比较类似，所以我们使用 2015 年 11 月 12 日的数据作为静态分析的代表。

[14] 动态分析的数据集用户数量略少于静态分析的数据集导致的原因比较复杂，但主要的可能原因有用户在我们采集数据周期内注销了帐号，或者变为非活跃用户导致我们无法采集到该用户每天的数据信息。

户的个人属性信息是我们的分析中至关重要的一项数据依据，因此我们通过对用户头像进行人脸识别来获取用户个人属性信息。在此，我们运用开源人脸识别服务 Face++获取用户个人属性信息，获取的数据包括用户的性别、种族以及具体的年龄信息。图 3.2 展示了 Twitter 和 Instagram 用户个人属性数据的分布情况，分布符合各网站公布出的用户整体分布数据。例如整体而言，社交网络更受女性用户的青睐，而 Twitter 比 Instagram 能吸引更多的男性用户。Instagram 相较于 Twitter 拥有更多的年轻用户，所以整体年龄分布更靠前。

值得注意的一点是，当我们使用 Instagram 提供的 API 来获取隐私用户的数据时，由于受到官方设置的权限限制，而无法获取到隐私用户的任何数据信息，因此我们的实验分析也会受到一些影响。在表 3.1 中我们罗列了针对 Twitter 和 Instagram 两个社交网站，我们的实验开展情况。

表 3.1 Twitter 和 Instagram 网站上实验实施情况

	静态分析			动态分析			预测分析
	个人属性	线上行为	线下行为	个人属性	线上行为	重大事件	
Twitter	√	√	√	√	√	√	√
Instagram			√	√		√	

3.2 社交网站用户信息分享行为数据准备

社交网站上用户信息分享行为的数据采集方法与社交网站访问控制设置的数据采集方法略有不同，我们没有局限在纽约城市范围内的用户数据，而是通过随机采样，在全球范围内进行了种子用户的采集。

首先，我们需要确定在 Instagram 上目前仍然有效的用户数字账号区间，借鉴了其他研究人员类似的取样方法[5][20]。由于在 Instagram 上，所有用户的数字账号拥有着先后顺序，而 2,000,000,000 之后的有效账户就趋于 0，因此我们在前 2,000,000,000 个数字账号中随机采样 40,000,000 个帐号，之后再通过调用 Instagram 的 REST API 找出所有仍然在使用的活跃用户帐号。经过过滤后，我们共采集到 11,982,242 个数字账号。在此基础之上，我们进一步通过调用 Instagram 的 REST API 来收集所有用户发布的所有帖子内容，数据包括发布时间、使用的标签、照片的滤镜、获得的点赞数量以及评论数量等。如果用户发布的帖子内容中包含了地理位置共享信息，即地理经纬度的坐标，我们也会一同获取。在之后的实验过程中，我们通过调用 Google 提供的 Geocoding API 将所有的地理坐标数据转换为具体的国家、地区等可读数据。在采集了所有这些数据之后，我们最后进行过滤，保留所有包含了#like4like 标签的帖子数据。最终，我们获取到 143,586 个有效的公开用户的信息，并获取这些用户从 2010 年 12 月至 2016 年 3

月发布过的所有#like4like 相关内容，共 1,770,643 篇帖子。与访问控制设置数据采集相同的是，我们仍然通过调用 Face++的 DETECT API 来对用户的头像照片进行人脸识别后获得用户的个人属性数据。

其次，为了更好地探究用户与陌生人之间的互动行为，我们进一步采集了#like4like 相关的动态数据。在 2016 年 5 月期间，我们使用 Instagram 的 REST API 随机采集了 11 天的动态数据，数据采集结果如表 3.2 所示。

表 3.2 用户社交行为实施程度数据集

数据采集日期	#like4like/#l4l 帖子数量
2016-05-06	23
2016-05-07	18
2016-05-12	26
2016-05-13	62
2016-05-14	38
2016-05-17	36
2016-05-18	31
2016-05-19	39
2016-05-20	7
2016-05-24	26
2016-05-25	25
总共	**331**

由于 Instagram 的官方 API 不提供时间信息，我们无法直接获取用户的点赞与回赞时间，我们通过追踪的方法来获取这一部分的时间数据。假设用户 Alice 发布了一片带有#like4like 标签的帖子。当用户 Ben 为 Alice 发布的帖子点赞之后，我们会记录下这一时刻，并立即触发另一个进程去获取后续的动态数据。我们首先遍历用户 Ben 个人主页上所有已发布的帖子，查看 Alice 在此之前是否为 Ben 点赞，若有则记录下点赞的帖子和时间，这一部分数据将作为 Alice 与 Ben 之间好友关系判断的依据。然后我们持续追踪 Alice 与 Ben 的互动状态，查看 Alice 是否给 Ben 回赞，直至发现 Alice 的回赞并记录下 Alice 回赞的帖子内容与时间，以及这些帖子在 Ben 个人主页上的序号与包含的其他标签。

第四章　社交网站用户访问控制设置分析

4.1 静态分析

社交网站用户访问控制设置使用情况的静态分析以 2015 年 11 月 12 日的数据作为研究对象，得到了隐私用户与公开用户的异同性的分析结果。首先，我们给出以下定义：开启了个人主页访问控制设置的用户称为*隐私用户*；而未开启访问控制设置，即完全公开个人主页的用户，我们定义为*公开用户*。

4.1.1 整体统计数据

如表 4.1 所示， 2015 年 11 月 12 日，分别有 5.22%的 Twitter 用户和 11.92%的 Instagram 用户开启了个人主页的访问控制设置，将自己设置为隐私用户。而 Instagram 用户比 Twitter 用户更注重个人隐私保护的原因可能是用户使用这两个社交网站的目的不同：用户更倾向于从 Twitter 上获取新闻资讯，因此较少地分享与个人隐私相关的内容；而 Instagram 则不同，作为一个单纯的照片分享平台，用户的大量个人信息都有可能通过分享在网站上的照片而被陌生人获取，带来巨大的信息安全隐患。所以，Instagram 上隐私用户的比例远高于 Twitter。

表 4.1 社交网站用户统计数据

	用户分类	整体	使用真人头像	未使用真人头像
Twitter	隐私用户	9,145（5.22%）	6,066（5.65%）	3,097（4.54%）
	公开用户	166,057（94.78%）	101,347（94.35%）	64,710（95.46%）
Instagram	隐私用户	34,844（11.92%）	-	-
	公开用户	257,562（88.08%）	-	-

目前 Twitter 和 Instagram 也都没有提供任何关于各自网站上隐私用户比例的官方数据。Cha 等人统计了 Twitter 网站上隐私用户的比例大约为 7%[5]，这与我们的数据比较接近，而细微的差距可能由取样方法的不同导致。Cha 等人采用的是在所有 Twitter 用户中随机取样的方法，而我们分析的目标用户则仅局限于纽约市的用户。对于 Instagram 网站上的隐私用户比例尚未有可作参考的研究数据。但是由于我们的研究重点在于分析现实世界中，用户是如何部署各自的访问控制设置的，所以在此对整体情况的数据仅作参考数值。

4.1.2 用户的个人属性

我们将通过 Face++软件分析用户个人头像所得到的性别、种族与年龄作为用户的个人属性，结合他们的访问控制设置情况进行分析并发现：对于不同性别、

不同种族、不同年龄阶段的用户，他们具有各自不同的访问控制设置偏好。其中，女性、年轻人以及亚裔用户表现出更强的对线上个人隐私的保护意愿。

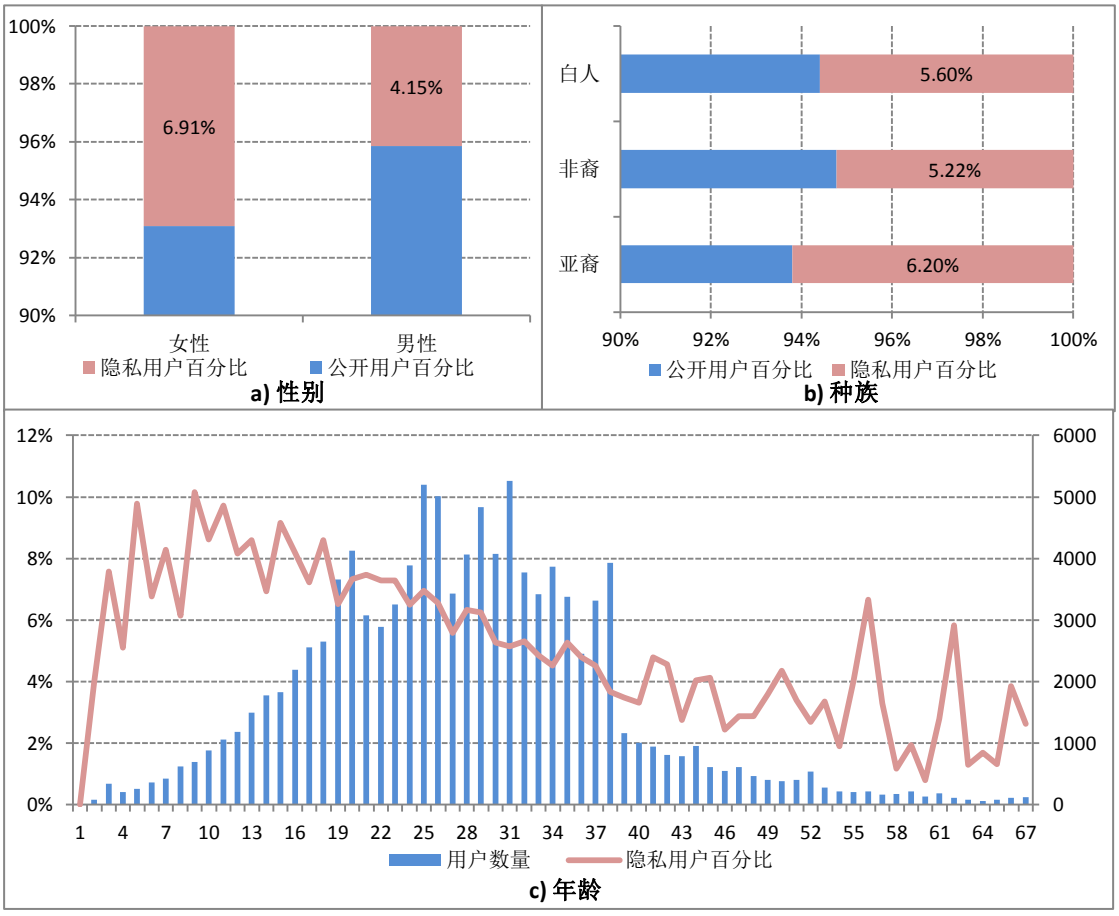

图 4.1 Twitter 用户个人属性分布情况

性别 通过分别对女性用户与男性用户的隐私用户数量统计后我们发现，6.91%的女性用户将自己设置为隐私用户，阻止了陌生人对于其个人主页的访问，而仅有 4.15%的男性用户开启了自己的隐私设置（如图 4.1a）所示）。女性比男性更注重保护自己的个人隐私。

种族 从图 4.1b）中我们不难发现，在纽约用户中，亚裔用户拥有最高比例的隐私用户（6.20%），远高于白人与非裔用户的隐私用户比例，分别为 5.60%和 5.22%。我们认为导致这一现象的可能原因之一是文化差异：亚洲人，尤其以中国、韩国、日本等国家为例，长久以来都接受着传统儒家文化的熏陶而以含蓄内敛为主要文化特点，因此，即使生活在西方国家，以华裔、韩裔、日裔为首的亚裔用户仍然保留着传统思想，而比其他种族的用户更关注个人隐私的保护。

年龄 对于所有年龄在 10 岁以上的 Twitter 用户，他们的隐私用户比例随着年龄的增长而不断下降（如图 4.1c）所示）。而这一趋势在年龄段处于 20-40 岁之间的用户中格外明显，也就是说，年轻用户比年长用户更在意个人网上信息的安全

问题。我们还发现，10 岁以下的用户也保持着较高的隐私用户比例。由于在现实生活中，10 岁以下的儿童对于外界的认知度有限，但无论是对社交网站的认识与使用，还是对个人隐私保护的关注度，都表现出在作为分析对象的用户人群中，这一群年龄在 10 岁以下的人群不是真正的儿童，而更可能是希望通过设置非真实头像来保护个人隐私的成年人。

所以，我们进一步对这一群 10 岁以下的人群进行了性别、种族与个人访问控制设置情况的具体分析。通过图 4.2 与图 4.1 的对比，不难发现，无论男性还是女性，在 10 岁以下的用户群体中都有更高的隐私用户比例，女性隐私用户占比上升 1.33 个百分点达到 8.24%，而男性则上升了 1.84 个百分点达到 5.99%。而对于这组人群中不同种族的用户来说，三个种族的隐私用户比例也都同时有了大幅度上升。其中，亚裔用户的隐私用户比例上升了 2.1 个百分点达到 8.30%，依然是三个种族人群中隐私比例最高的人群。这两个结果与之前得到的女性与亚裔用户更关注隐私保护问题的结论相一致，且更重要的是，他们验证了 10 岁以下的“儿童”其实是使用了儿童头像、且非常在意个人隐私保护的用户群体。

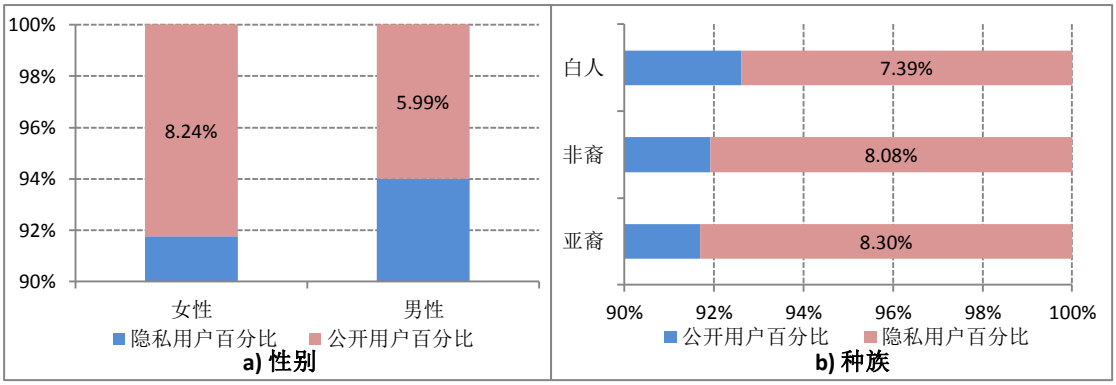

图 4.2 10 岁以下 Twitter 用户个人属性分布情况

个人头像 由于我们是通过使用 Face++工具对用户个人头像进行分析后得到个人属性信息的，所以如果用户没有使用人物头像，而是使用了例如宠物、风景等其他照片作为个人头像，那么我们将无法获取到他们的个人属性数据。Twitter 用户中，使用了人物头像与未使用人物头像的人群比例统计如表 4.1 所示。Twitter 网站上，使用了人物头像的用户共 107,413 人，其中 6,066 名用户为隐私用户，占比约 5.65%。而未使用人物头像的用户共 67,789 人，其中隐私用户有 3,079 人，占比 4.54%。由此可见，在 Twitter 这一类网站上，未使用人物头像的用户对个人信息泄漏的警惕性要略低于使用了人物照片作为个人头像的用户。我们认为可能的原因是，Twitter 作为一个新闻传播平台，使用非真实照片作为个人头像已经能够给用户带来更多的安全感，从一定程度上保护了他们的个人隐私。

4.1.3 用户的线上活动

我们通过四个数值来对用户线上活动的两方面进行量化，它们分别是：1）用户线上的活跃度，通过用户发布的帖子总数与用户点赞的帖子总数来量化；2）用户线上的社交圈大小，通过用户当前的粉丝总人数与被关注者总人数来量化。

由于隐私用户与公开用户的人数数量级相差巨大，我们通过盒图来展示这两类人群在用户线上活跃度与用户线上社交圈大小这两个方面的数据分布并进行比较。在盒图中，我们标识出了整体分布的最小值、下四分位数、中位数、上四分位数，通过这些标志线来展示并比较隐私用户与公开用户的整体分布情况。

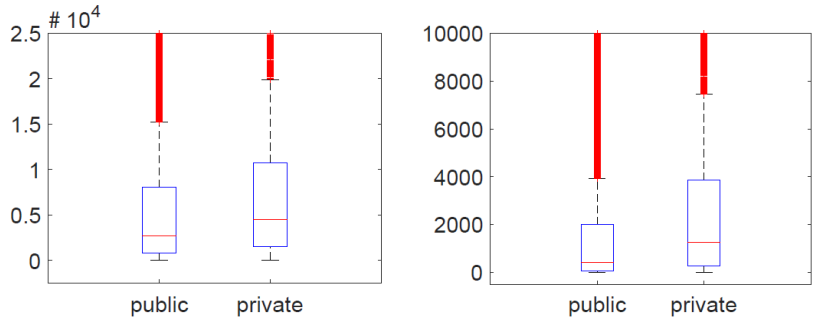

图 4.3 公开用户与隐私用户的发布帖子总数分布（左）及点赞帖子总数分布（右）

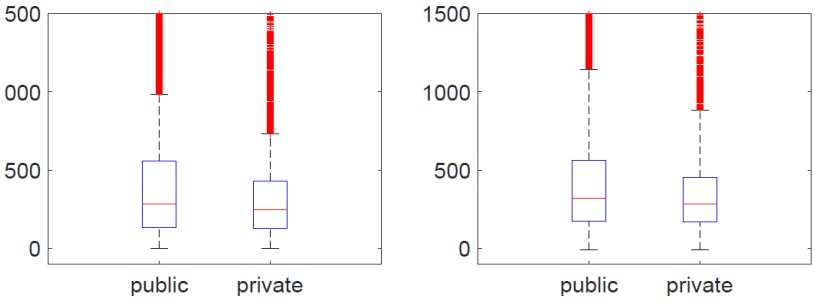

图 4.4 公开用户与隐私用户的粉丝人数分布（左）及关注人数分布（右）

从图 4.3 中我们可以看到，与公开用户比较，隐私用户发布了更多的帖子并给更多的帖子点赞。隐私用户平均每人总共发布了 8,864.07 篇帖子，点赞了 3,380.43 篇帖子，而公开用户平均每人总共发布了 7,550.96 篇帖子，点赞了 2,277.58 篇帖子。可能导致这一现象的原因有两种：

1) 隐私用户由于设置了个人主页的访问控制，将陌生人屏蔽在外，所以大量潜在的攻击者不能查看到隐私用户发布的所有内容，这使得隐私用户的安全感增强，所以他们非常放心地在网络上表达自己的所思所想，因此他们的活跃度高于公开用户。
2) 隐私用户是经验更加丰富的网民，与其他用户相比，他们更加关注个人网上隐私信息的保护，也对网络空间中个人信息泄露所能带来的威胁有更深刻的认识。也因为隐私用户拥有更长的网龄，所以他们积累了更多的发布和点赞的帖子。

在第 4.2.3 章节中，我们将通过对隐私用户与公开用户进一步的动态分析来具体阐述最终的结果。

与网络活跃度的差异情况刚好相反的是，我们发现，隐私用户的网络社交圈要远小于公开用户，换言之，公开用户拥有更多的粉丝与被关注者（如图 4.4 所示）。公开用户平均拥有 423.26 位粉丝，平均关注了 431.73 位用户，而隐私用户平均拥有 329.37 位粉丝，平均关注了 355.17 位用户。在第 1.2 章节中我们介绍了 Twitter 与 Instagram 提供的访问控制机制是当用户开启了自己的个人主页访问控制设置之后，所有新添加的粉丝必须向用户发出申请，在用户同意通过粉丝申请之后才能成为粉丝。从巨大的平均粉丝数量差值可以看出，这一操作能够有效帮助用户过滤并隔离陌生人。而另一方面，有趣的是，我们发现隐私用户关注的用户数量也远少于公开用户，这意味着隐私用户不仅会筛选过滤不必要的粉丝，他们在关注其他用户的时候也是有所挑选的，以此来避免个人网上的社交圈因过大而产生潜在的信息安全隐患。

4.1.4 用户的线下活动

从 Twitter 和 Instagram 网站上分别获取到的用户签到数据能够有效地反映出纽约用户的线下生活。我们获取到的签到数据包含了用户的签到地点与签到时间，所以接下来我们将从这两方面入手对用户的线下生活进行分析。

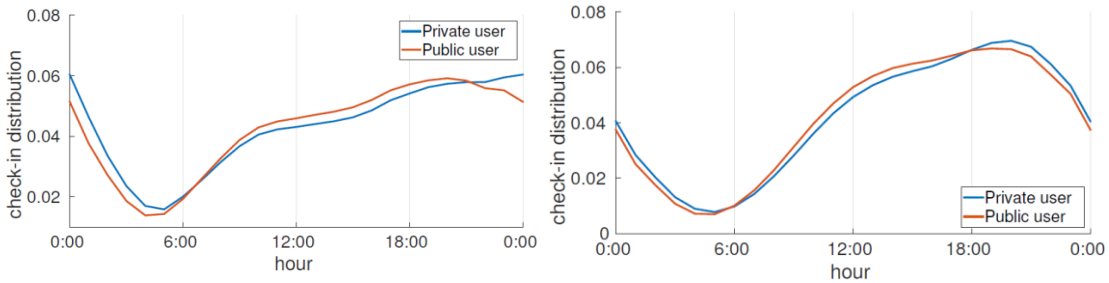

图 4.5 隐私用户与公开用户在 Twitter（左）和 Instagram（右）上签到时间的分布情况

签到时间 图 4.5 展示了 2015 年 11 月 12 日当天 24 小时内，Twitter 用户和 Instagram 用户在各个时间段签到的人数分布。虽然 Twitter 和 Instagram 两个网站上用户的签到时间分布曲线有些许的不同，但是两个网站上隐私用户与公开用户签到时间的差异是一致的：与公开用户相比，隐私用户在夜间表现得更加活跃。由于 2015 年 11 月 12 日是周四，属工作日，白天是上班时间，那么夜间时间就成为了大部分线下社交活动发生的时间段。由此我们推断，在线下的社交活动中，隐私用户比公开用户参与度更高。

签到地点 由于 Twitter 与 Instagram 两个网站不同的接口设计，我们只能获取到 Instagram 用户的具体签到地点的分类，而不能够对 Twitter 进行分析。在此，我

们对于签到地点的分类基于 Foursquare[15]产生，Foursquare 是一个提供基于地点的内容分享的社交网络，因此它提供了合理的树形结构的地点分类。我们在此获取了其中第一层的地点分类，将所有的签到地点共分为了 9 个大类，分别是：娱乐（Entertainment）、大学（University）、食品（Food）、夜生活（Nightlife）、室外运动（Outdoor）、住宅区（Residence）、商店（Store）、交通枢纽（Transportation）以及专门性场所（Professional）。

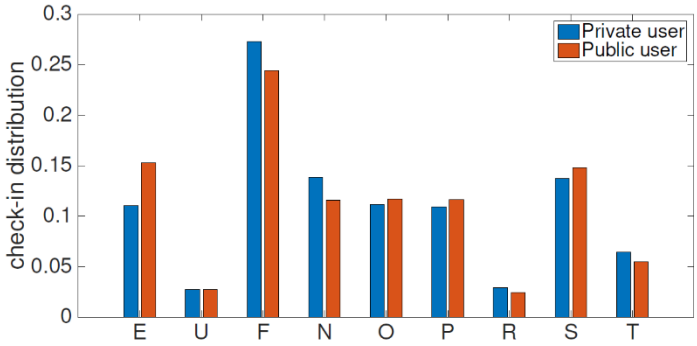

图 4.6 Instagram 上隐私用户与公开用户签到地点的分布情况

图 4.6 展示了隐私用户与公开用户在 2015 年 11 月 12 日当天 24 小时内在各个时间段签到地点的分布情况。从图中可以看出隐私用户在食品类场地与夜生活场地的签到明显多于公开用户。由于现实生活中，绝大多数的社交活动都会发生在类似于饭店、酒吧等一系列的食品类场地与夜生活场地，结合用户签到时间的表现，我们进一步断定在现实生活的社交活动中，隐私用户比公开用户表现得更加活跃，并且隐私用户因为有着更高的安全感（不会被陌生人看到自己分享的内容，本人的社交圈也更干净），他们会在社交网站上比较频繁地分享自己的线下社交生活。

4.2 动态分析

通过静态分析，我们了解了隐私用户与公开用户各自的特点，接下来，我们希望通过动态分析来探究导致用户改变他们的访问控制设置的可能因素。

在此，动态分析指的是基于某一个连续时间段内获取到的数据，进行用户每天访问控制设置情况动态变化的分析。我们通过分析影响用户决策的三方面因素入手，一是包括用户的性别、种族、年龄等个人属性在内的个人属性因素，二是用户在社交网站上外在表现的行为变化，三是与外界事件相关的外在因素。我们希望通过这三方面因素的大数据分析来全面地了解用户的真实想法。

[15] Foursquare，可参考：https://foursquare.com/。

4.2.1 整体趋势

首先，我们统计了用户在抓取数据的将近 3 个月的时间区间内整体的变化趋势与变化频率。我们发现，绝大多数的用户对于个人主页访问控制的设置决策是多变的，也就是说他们会频繁地修改个人主页的访问控制设置。然而，在目前越来越多的人意识到网络数据安全问题的大环境下，整体而言，隐私用户的比例仍然在不断上升，越来越多的人通过将自己设置为隐私用户来保护社交网络中个人信息的安全。

表 4.2 用户访问控制设置变化情况的统计数据

变化次数	Twitter 用户	百分比	Instagram 用户	百分比
1	3,458	45.56%	17,390	30.91%
2	2,494	32.86%	15,252	27.11%
3	473	6.23%	4,397	7.82%
4	506	6.67%	5,179	9.21%
5	171	2.25%	2,121	3.77%
6	153	2.02%	2,597	4.62%
7	83	1.09%	1,220	2.17%
8	59	0.78%	1,470	2.61%
9	44	0.58%	849	1.51%
10	39	0.51%	955	1.70%
11	20	0.26%	552	0.98%
12	23	0.30%	721	1.28%
13	11	0.14%	455	0.81%
14	12	0.16%	493	0.88%
15	5	0.07%	267	0.47%
>15	28	0.51%	553	4.16%
总数	7,590		56,261	

根据我们的分析数据显示，在 2015 年 10 月 14 日，Twitter 与 Instagram 这两个主流社交网站上的用户中各自有 4.89%和 9.36%的隐私用户。而截止至 2016 年 1 月 12 日，两大社交网站上的隐私用户比例分别上升至 5.62%和 14.20%。由此可见，在我们抓取数据的 3 个月时间内，两个网站上的隐私用户占比都有所提升，尤其是 Instagram，隐私用户占比提升了接近 5 个百分点。而研究人员在另一主流社交网站 Facebook 上也得到了类似的结果[8][21]。这一数据很好地证明了人们对于在社交网站上个人信息的保护意识正在逐步提升。

表 4.2 列出了用户在 Twitter 和 Instagram 两个社交网站上个人主页访问控制

设置变化情况的统计数据。从 2015 年 10 月 14 日至 2016 年 1 月 12 日，Twitter 网站上，共有 7,590 位用户至少修改了一次访问控制设置，占比达到 5.21%。而这一比例在 Instagram 上则更高，共有 56,261 名纽约用户改变过个人主页的访问控制设置，变化人群比例高达 19.95%。

表 4.2 中还具体列出了 Twitter 和 Instagram 两个社交网站上用户改变个人主页访问控制设置的次数的统计情况。在这两个网站上，绝大多数改变了个人主页访问控制设置的用户都表现出了设置决策的不确定性。我们称所有修改过个人主页访问控制设置的用户为*变化人群*。Twitter 网站上有 54.44%的变化人群多次改变访问控制设置；而在 Instagram 上，则有将近 70%的变化人群在 3 个月时间内曾多次改变自己的设置，更有 4.16%的 Instagram 变化用户人群（553 人）改变次数多于 15 次，这个数字意味着他们平均一周就会改变一次设置。

这些数字反映了在目前社交网站访问控制策略设计尚不是非常成熟的阶段，用户对于自己在社交网站上访问控制设置的决定具有高度的不确定性。他们会经常地修改自己的设置，有许多可能的原因，在接下来的实验中，我们就希望通过一系列的科学分析来探求这其中的蛛丝马迹，了解用户的真实想法。

4.2.2 用户的个人属性

通过统计分析，我们发现，女性用户与年轻用户会更加频繁地改变个人主页访问控制设置，并且他们中有更多的人会由公开用户变为隐私用户。而在针对不同种族用户的分析中我们发现，相较而言，白人对个人主页访问控制设置的决策最稳定，并且他们也更加开放，隐私用户比例的增长速率最为缓慢。接下来我们从变化频率与趋势两方面进行分析。

- **访问控制设置变化频次与个人属性的关联**

根据个人属性[16]，Twitter 和 Instagram 两个社交网站上用户个人主页访问控制设置的变化情况如表 4.3 所示。

从表 4.3 中可以看到，在我们数据采集的 3 个月期间，Twitter 用户中有 6.52%的女性用户和 3.66%的男性用户改变了他们的个人主页访问控制设置，Instagram 用户中有 19.20%的女性用户与 13.92%的男性用户也修改了个人主页访问控制设置。另外，这些变化的人群中，无论是 Twitter 还是 Instagram，女性用户对于个人主页访问控制设置的不确定性均高于男性用户。这一点在 Instagram 上尤其明显，女性用户平均修改 3.60 次，而男性用户平均修改 2.93 次。

除了性别之外，不同的种族人群之间也有着较大的差异。在 Twitter 上，亚裔用户拥有最高比例的变化人群，达到 6.1%，而非裔用户则平均改变访问控制

[16] 不同于静态分析，由于用户的个人主页访问控制设置改变，在我们分析研究的 3 个月期间内，有一部分用户由公开用户变成了隐私用户，所以我们可以在动态分析中对这一部分用户进行基于个人属性的分析。

设置的次数最多，为 2.4 次。而在 Instagram 上，20.63%的非裔用户修改过个人主页的访问控制设置，成为三个种族中比例最高的人群，但是亚裔用户的决策却最多变，平均每人修改了 3.84 次。有趣的是，白人无论在 Twitter 还是 Instagram 上，都表现出了最高的稳定性，不仅更少的人会修改原本的访问控制设置，而且在修改过后也是表现得最稳定的一群人。其中的原因可能不仅因为西方文化相较于亚洲文化更加得开放，也因为我们的研究目标为纽约，以白人为主的整个社会环境使得白人更有安全感。

表 4.3 Twitter 和 Instagram 网站上根据不同的个人属性，访问控制设置变化情况的统计数据

		整体	性别		种族		
			女	男	亚裔	非裔	白人
Twitter	变化用户人数比例（%）		6.52	3.66	6.11	5.16	5.04
	平均变化次数（次）	2.29	2.37	2.09	2.37	2.40	2.25
	变化 1 次的用户比例（%）	45.59	43.77	49.44	42.28	42.14	46.89
Instagram	变化用户人数比例（%）		19.20	13.92	19.33	20.63	16.32
	平均变化次数（次）	3.40	3.60	2.93	3.84	3.78	3.22
	变化 1 次的用户比例（%）	34.85	34.68	42.08	27.79	30.53	37.55
		整体	年龄段				
			0-10	11-30	31-45	>=46	
Twitter	变化用户人数比例（%）		7.72	5.95	3.29	2.41	
	平均变化次数（次）	2.29	2.52	2.28	2.23	2.07	
	变化 1 次的用户比例（%）	45.59	41.15	44.77	51.23	48.45	
Instagram	变化用户人数比例（%）		20.19	18.09	13.14	11.34	
	平均变化次数（次）	3.40	3.74	3.47	2.82	2.90	
	变化 1 次的用户比例（%）	34.85	32.26	33.26	43.63	46.71	

如表 4.3 所示，我们根据用户年龄将用户分为 4 类：0-10 岁，11-30 岁，31-45 岁以及 45 岁以上。在 Twitter 和 Instagram 两个社交网站上，都表现出了年轻人会更频繁地改变他们的个人主页访问控制设置的趋势。例如，在 Instagram 上，18.09%的 11-30 岁用户改变了他们的访问控制设置，并且平均每人改变了 3.74 次。而在 31-45 岁的人群中，这些数字就低了许多，只有 13.14%的用户改变了个人访问控制设置，平均的修改次数也下降到了 2.82 次。另外，10 岁以下的用户仍然保持着超高的活跃度，有最多的人改变了个人访问控制设置，并且平均改变次数也是最高的。这一点表明了那些使用了儿童的照片作为个人头像的用户比其他用户表现得更加矛盾，他们一方面非常在意网络空间中个人隐私的保护，另一方面也频繁地想要获得更多的关注。

- **变化趋势与个人属性的关联**

图 4.7 展示了不同性别、种族、年龄段的用户群个人主页访问控制设置的变化趋势。图 4.7a）显示无论是 Twitter 还是 Instagram，女性隐私用户的增长速率

都大于男性。而这一趋势在 Instagram 上更为显著，女性隐私用户增长了约 10%，而男性隐私用户增长了约 8%。从图 4.7b）中我们可以看到三个不同种族的人群在两个网站上有着不同的表现。其中差异最大的是非裔用户，在 Twitter 上他们拥有最小比例的隐私用户并且增长速率最慢，而在 Instagram 上，非裔却拥有最高比例的隐私用户，并且有着最快的增长速率。我们认为这可能是用户使用这两个网站的不同目的导致的。不同年龄段的隐私用户比例与增长速率在图 4.7c）中展示，不难发现，年轻用户始终有更高比例的隐私用户，并且增长速率也更快。

4.2.3 用户的线上活动

- **用户网上活跃度**

为了对用户网上活跃度进行有效的动态分析，我们将用户分为 3 组，分别是：三个月内始终保持私密状态的*隐私用户*；三个月内始终保持公开状态的*公开用户*；以及更改过设置状态的*变化用户*，并分别对他们的网上行为表现进行了相应的分析。与 4.2 节中相对应的，我们采用 4 个度量值来量化衡量各个用户网上的活跃度情况，这 4 个度量值分别是：1）每日发帖数量；2）每日点赞数量；3）每日新增粉丝数；4）每日新增关注用户人数。

图 4.8 展示了隐私用户与公开用户每日活跃度的动态变化情况。在 4.1.3 节中我们发现相较于公开用户，隐私用户在网上发布了更多的帖子，并给更多的帖子点了赞。然而，从图 4.8a）中我们有了有趣的发现，代表隐私用户的两条折线始终低于公开用户的两条折线。换言之，根据每天新增的发帖数目与点赞数目来看，除去极个别日期的特殊情况，隐私用户的活跃度始终低于公开用户。在 4.1.3 节中我们给出了两种猜测解释：一为隐私用户由于设置了有效的访问控制，他们更愿意在网上分享自己的感受与体验；二是隐私用户可能是那些拥有更长网龄的成熟网民，所以他们更清楚地明白互联网给个人信息带来的威胁。而在此我们的发现给予了第二种解释强有力的支撑，也就是说，隐私用户他们并不是因为开启了访问控制设置能给自己带来安全感而变得活跃，相反地，他们的活跃度始终低于公开用户，但是因为他们拥有更长的网龄，才会使得他们积累了更多的已发布和已点赞的帖子。

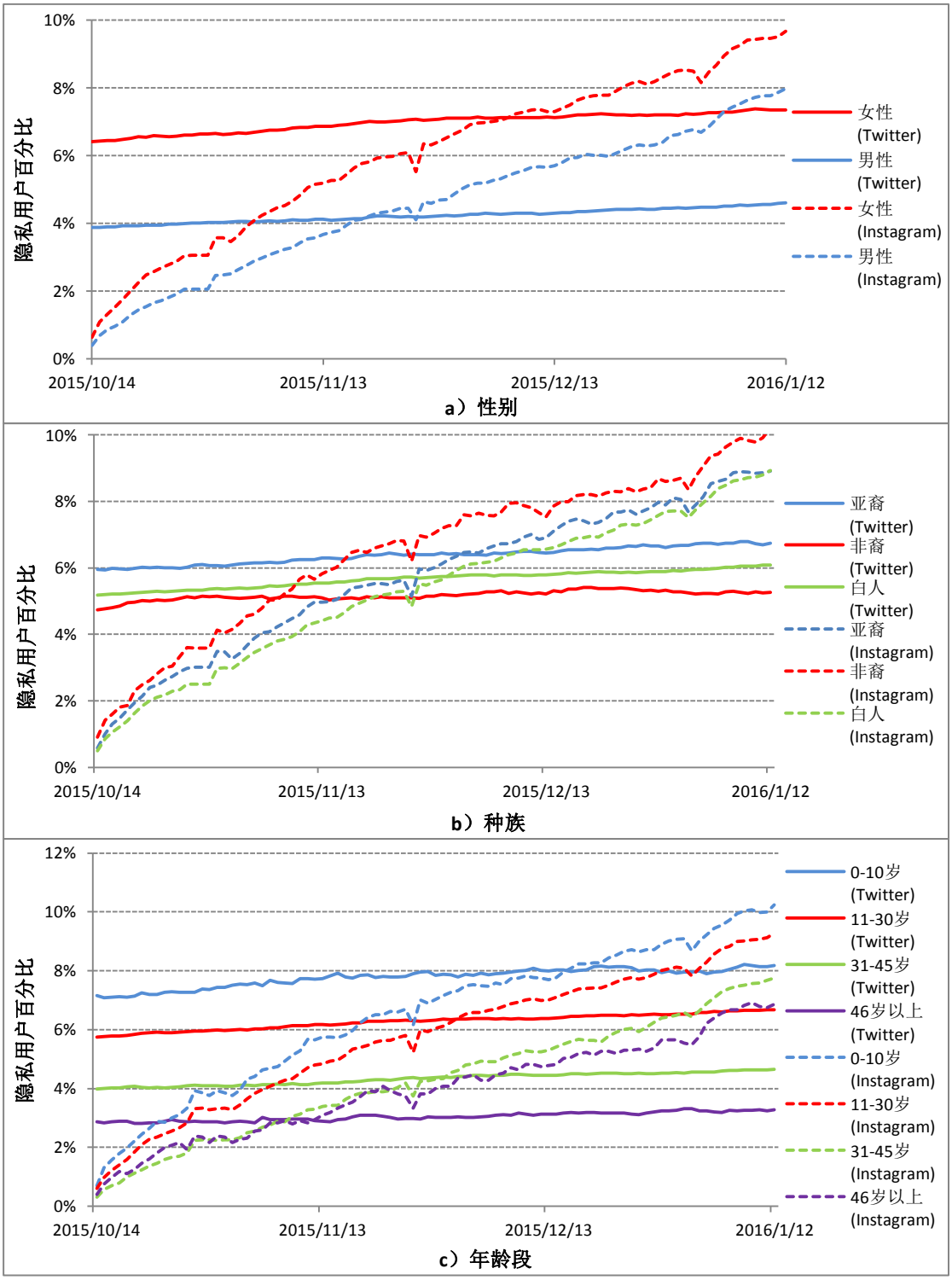

图 4.7 基于用户个人属性的隐私用户比例变化趋势图[17]

另外，从图 4.8b）中，我们还发现那些始终保持私密状态的隐私用户极少扩张个人的网上社交网络朋友圈。他们几乎不添加新的关注，并且每天平均每个隐私用户添加的新粉丝数为负增长。相反地，公开用户无论是每天的新关注用户数量，亦或每天的新增加粉丝数量都大于 0，说明公开用户整体仍然在不断地扩张

[17] 由于我们无法获取 Instagram 上隐私用户的个人信息，所以 Instagram 的初始隐私用户接近于 0。

个人的社交网络朋友圈。这明显的反差进一步支撑了 4.1.3 节中的结论，隐私用户对个人社交网络朋友圈的筛选比公开用户更严格、更谨慎。

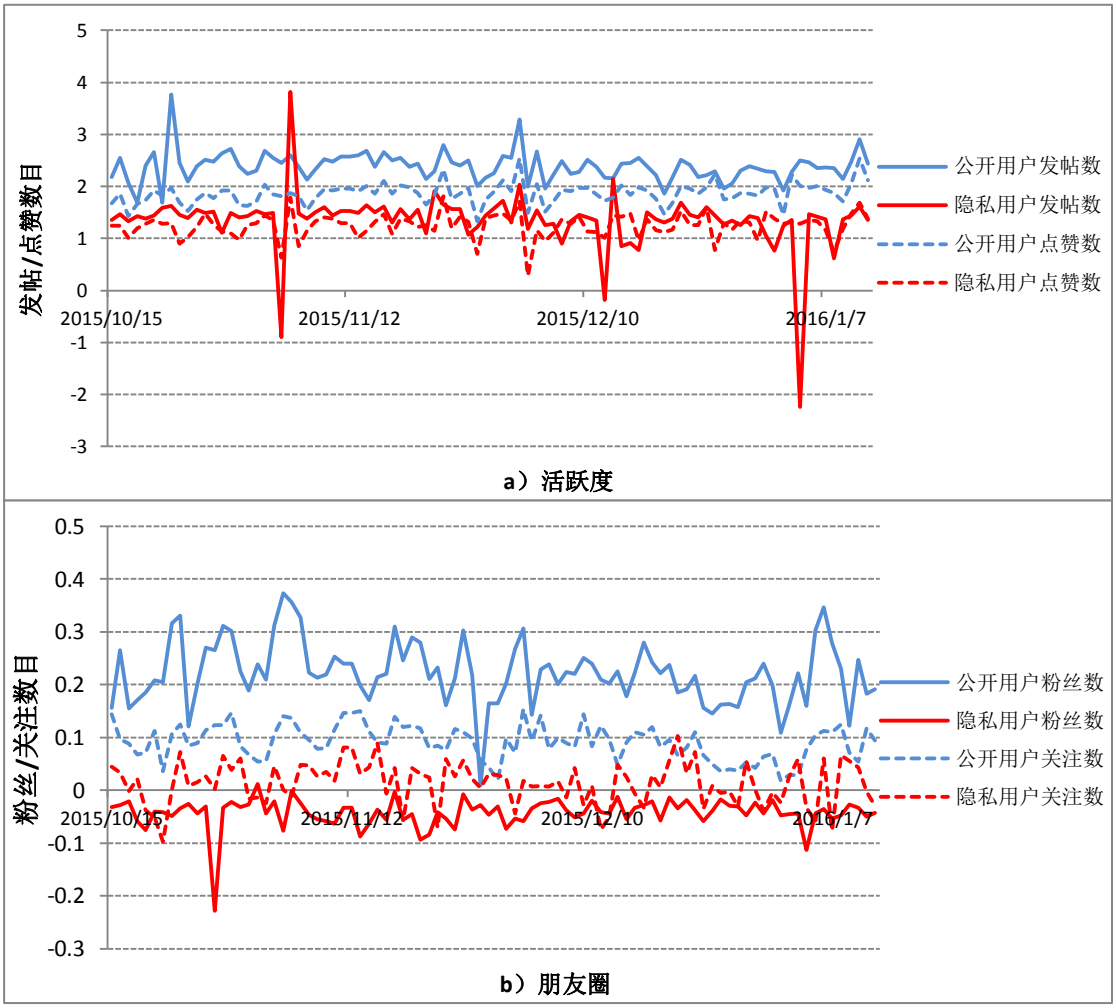

图 4.8 公开用户与隐私用户网上行为动态分析

表 4.4 变化状态用户网上活动数目统计情况

	处于私密状态时	处于公开状态时
发布帖子数目	3.12	5.64
点赞帖子数目	4.35	4.67
新增粉丝数量	-0.15	0.25
新增关注数量	-0.04	0.14

对于变化用户，我们采用了另一种分析方法。基于之前提出的 4 个度量值，我们对变化用户在私密和公开两个不同状态下的网上行为表现分别进行了计算，得到如表 4.4 所展示的数据结果。有趣的是，即便是同一批用户，他们在处于私密状态时与处于公开状态时的网上行为表现也有着较大的差异。用户处于私密状态时的活跃度明显低于处于公开状态时。日均发帖数量，处于私密状态时为 3.12，而处于公开状态时为 5.64。而当用户处于私密状态时，日均新增粉丝数量与日均

新增关注数量均为负增长。这一现象意味着，用户改变个人主页的访问控制设置是具有目的性。他们极有可能在私密状态时清理个人的朋友圈。

- **用户谈论话题**

我们通过分析变化用户在修改个人主页访问控制设置前后在网上的发帖内容来识别用户谈论的话题，发帖内容指的是 Twitter 上的一条状态或是 Instagram 上伴随照片发布的语句。

表 4.5 用户改变个人主页访问控制设置前后一天讨论的话题

Twitter	**由公开用户转变为隐私用户**		**由隐私用户转变为公开用户**	
	改变前	**改变后**	**改变前**	**改变后**
话题 1	happy years new	woman get never	just one time	can party still
话题 2	music nothing three	family made truth	person every crying	one just kids
话题 3	nice looking needs	like boys text	bitch whole one	team really win
Instagram	**由公开用户转变为隐私用户**		**由隐私用户转变为公开用户**	
	改变前	**改变后**	**改变前**	**改变后**
话题 1	follow keep coming	thankful already missing	good morning feeling	god person remember
话题 2	go let strong	feels puppy wake	come show true	inspiration goodnight sleep
话题 3	can’t wait next	get also link	family friends lit	art music yesterday

我们利用自然语言处理领域中的 LDA 文档主题生成模型算法来探测用户发布的帖子的话题。通过抓取用户在变化个人主页访问控制设置前后各一天的发帖内容后，分别组成该用户变化访问控制设置前后的两个文档，而后过滤掉其中所有的标点符号与停止词。之后，我们将所有的文档进行整理，再过滤掉其中出现在少于 20 个文档或多于 70%的文档中的词[24]，最后构成我们对用户话题分析的

数据集合。虽然我们无法获取隐私用户的任何信息，但是由于用户的设置状态是动态的，那么一旦隐私用户更改个人设置公开自己的信息，我们就可以获取到他在私密状态时发布的状态。

我们分别分析了三个月内由公开用户转变为隐私用户和由隐私用户转变为公开用户在变化访问控制设置前后各一天的发帖内容，在表 4.5 中罗列了 Twitter 和 Instagram 上用户改变个人主页访问控制设置前后，排名前三的话题关键词。不难发现，当用户开启了访问控制设置处于私密状态时，用户发布的内容话题也更加私密。例如表 4.5 中列出的 Twitter 公开用户在新年期间谈论的热门话题“happy, years, new”，Instagram 公开用户谈论的“follow, keep, coming”等一系列大众化的关键词。而在用户状态为私密的时候，话题关键词则出现了“family”、“missing”、“thankful”等带有个人感情色彩的字眼。

4.2.4 重大事件/重要节日

在第 4.2.1 节中提到，越来越多的用户开始关注到社交网络上个人信息的安全问题，所以无论在 Twitter 或是 Instagram 网站上，隐私用户的比例都在不断上升。因此，每天新增的由公开状态转变为私密状态的用户人数应多于由私密状态转变为公开状态的用户人数。然而，当我们计算了每天这两类人群的数量差之后，我们发现：在某些全球性的重大事件或重要节日发生的当下，会有更多的人选择公开个人主页，有更强烈的意愿与外界沟通交流。

在我们抓取数据的三个月期间，有三个美国当地重要的节日，分别是感恩节（2015 年 11 月 26 日）、圣诞节（2015 年 12 月 25 日）和元旦（2016 年 1 月 1 日）。而在这三个节日当天，无论是 Twitter 还是 Instagram，都有更多的用户改变自己的访问控制设置为公开，也就是说，在这些日子里，人们变得更加开放了，暂时放下了对个人敏感信息的顾虑，会开放自己的个人主页来结交新的朋友，并与大家分享自己的过节方式。

而另一方面，我们也发现了一些重大事件也会刺激人们变得更加开放。例如，在巴黎恐怖袭击事件（2015 年 11 月 13 日）和加州枪击案（2015 年 12 月 3 日）发生之后，都有更多的隐私用户选择公开个人主页。我们猜测这一现象的原因可能是用户希望看到更多的关于这些事件的进展，也更愿意在网络上表达个人意见。巴黎恐怖袭击发生之后，在欧洲地区曾发起了公开 Twitter 个人主页接纳陌生人的公益活动，呼吁大家公开 Twitter 个人主页。

不仅是重要节日与重大的国际性事件，一些本地的活动也会影响用户对个人主页访问控制设置的决策。在 2015 年 11 月 1 日，Twitter 和 Instagram 上都出现了许多隐私用户改变为公开用户的现象。经过调查后发现，2015 年 11 月 1 日是 New York Mets 与 Kansas City Royals 之间的棒球联赛冠军赛。虽然纽约队在那一

场比赛中落败，但是仍然有大量的纽约用户由隐私用户转变为公开用户。

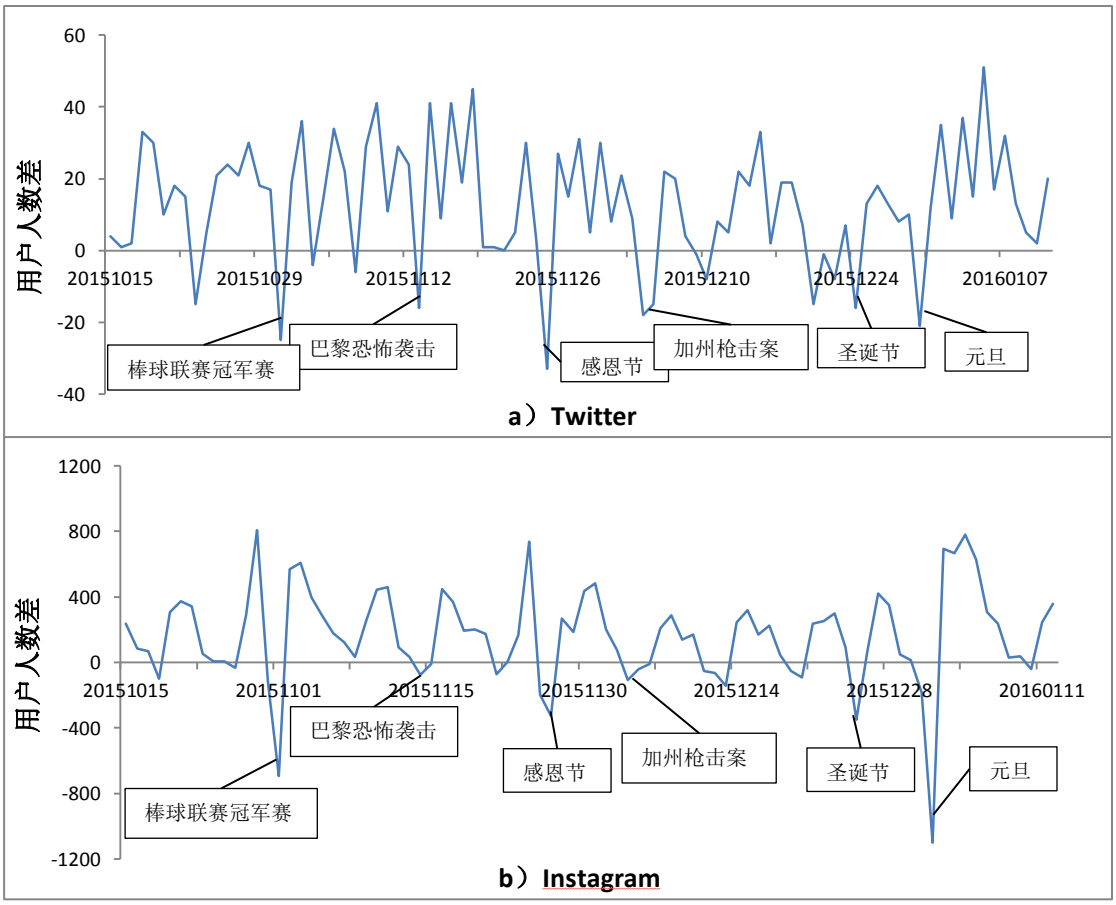

图 4.9 新增公开用户人数与新增隐私用户人数差

4.3 访问控制设置预测

在全面分析了真实用户个人主页访问控制的设置情况之后，我们希望验证是否能够通过构建模型的方法来对用户的访问控制设置进行自动预测。如果我们能够通过用户在社交网站上的一系列行为数据，以高正确率来准确地预测真实用户的访问控制设置，那么无论对于学术界还是工业界，都将开启社交网络个人访问控制机制研究的新篇章。未来，类似 Twitter 和 Instagram 的一系列社交网站都能够帮助用户自动调整他们的个人主页访问控制设置，不仅能够帮助用户们从复杂的访问控制设置的理解中解放出来，还能够帮助用户更有效地保护个人隐私，切实避免由于用户个人疏忽而产生的信息泄漏隐患；而政府部门也可以主导开发网络空间中面向所有用户的隐私信息保护建议模型，在可能的信息泄漏发生的第一时间向用户发出警示。

目前，我们的预测仅仅基于用户的静态信息，通过用户的个人属性和网上活动的静态数据构建相应模型后，对他们的个人主页访问控制设置是否开启进行预测。我们将对用户个人主页访问控制设置的预测问题规范为一个 0 或 1 的二分问

题，如果用户开启了个人主页访问控制设置，我们将其置为 1，反之，我们将其置为 0。并构建机器学习领域中的三个常用模型来解决这一预测分类问题。我们将输入的数据集分为两类来对用户个人主页访问控制设置进行预测，分别是：数据集 1 中，我们只导入用户线上活动的静态数据进行学习与预测，线上活动的静态数据包括用户发布的帖子总数与点赞的内容总数，以及他们的粉丝数与被关注用户数量；在数据集 2 中，我们不仅导入用户线上活动的静态数据，同时也将性别、种族、年龄段等用户个人属性数据同时导入，对于性别与种族两类文本数据，我们将文本转化为数字，以便于模型的处理与学习。在构建了以上 2 种不同的输入数据集之后，我们使用机器学习领域中的 3 个不同的分类模型对数据集进行训练和预测，它们分别是逻辑回归分析模型（logistic regression）、随机森林模型（random forest）、梯度推进模型（gradient boosting）。我们使用了 ROC 曲线[18]与 AUC[19]来对预测结果进行量化评定。

表 4.6 访问控制设置预测结果

	数据集 1	数据集 2
逻辑回归分析模型	0.59	0.62
随机森林分析模型	0.61	0.64
梯度推进分析模型	0.69	0.70

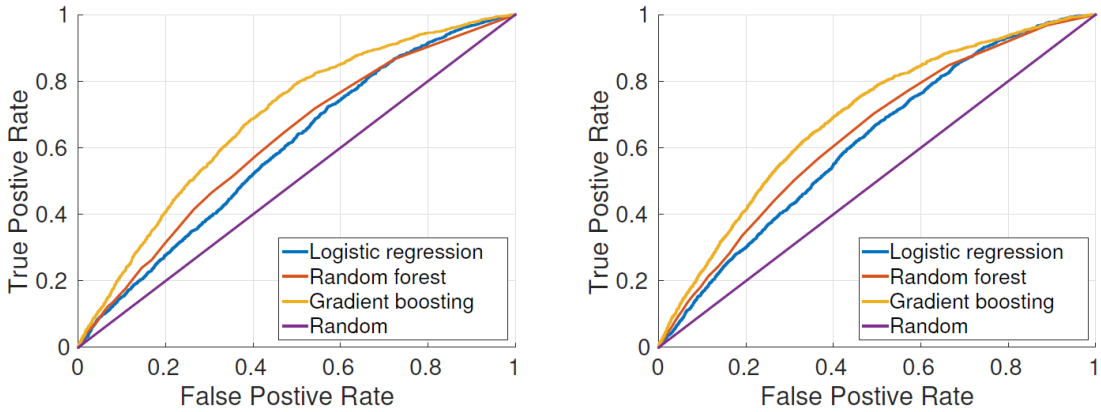

图 4.10 数据集 1 与 2 在经过不同分类训练与预测值后，得到的 ROC 曲线

表 4.6 列出了两个模型分别利用 3 个不同的机器学习分类模型对数据进行训练与预测的 AUC 结果。可以看到，我们获得的 AUC 最优结果为 0.70，是在导入用户线上活动数据和用户个人属性数据，再建立梯度推进分析模型来进行训练与预测后得到的结果，这一结果从一定程度上验证了用户个人主页的访问控制设置是可以被训练与自动预测的。同时，我们看到数据集 2 的预测结果都优于数据集 1，也就是说，用户的个人属性数据可以进一步提高预测用户的个人主页访问

[18] ROC 曲线，Receiver Operating Characteristic Curve，又称感受性曲线。ROC 曲线是以真阳性率为纵坐标，假阳性率为横坐标，在特定条件下采用不同的判断标准得出的不同结果画出的曲线图。
[19] AUC，Area Under ROC Curve，ROC 曲线下方面积，是一种用来度量分类模型好坏的标准。

控制设置的准确性。图 4.10 进一步展示了两个模型在经过不同机器学习分类算法训练并预测之后的 ROC 曲线。

在此需要指出的是，由于我们无法获取到 Instagram 上隐私用户的个人属性与线上活动数据，所以我们的预测仅限于 Twitter 用户的个人主页访问控制设置预测。但是，由于之前的所有分析都显示 Twitter 用户与 Instagram 用户拥有着相类似的表现，我们认为，Twitter 网站上获取到的预测结果可以推及到 Instagram 用户，换言之，真实用户在以 Twitter 和 Instagram 为代表的社交网站上的访问控制设置可以被自动预测。

4.4 本章结论

本章主要阐述了用户在 Twitter 和 Instagram 两大社交网站上设置访问控制策略的行为特点，从静态与动态两个角度对用户个人主页访问控制设置情况以及变化情况进行了探讨。我们发现：

- 用户对社交网络中个人隐私的保护意识正在不断加强。三个月期间，Twitter 隐私用户比例上升 0.40%，Instagram 隐私用户比例上升 4.84%。其中，女性、亚裔、年轻人更注重个人信息的保护。
- 用户会频繁地修改他们的个人主页访问控制设置。三个月期间，5.21% 的 Twitter 用户与 19.95%的 Instagram 用户修改过个人主页访问控制设置，平均修改 2.29 次与 3.40 次。其中，女性与年轻人修改更为频繁。
- 隐私用户具有低活跃度、清理好友圈、发布更具有私密性的帖子等一系列行为特点，重要节日与重大事件会影响用户对个人主页访问控制设置的决策。

基于以上的发现，我们提出可以通过机器学习训练并自动预测用户的访问控制设置，并使用机器学习模型验证了自动预测用户个人主页访问控制设置的可行性，预测最优结果 ROC 曲线下面积达到 0.70。

第五章　社交网站用户信息分享行为分析

5.1 时序动态分析

#like4like 是 Instagram 网站上排行第二的最受欢迎标签。因此，我们以这一标签的相关内容作为切入点，基于我们获得的数据，从参与度、获得的关注度（获得的点赞数量与评论数量）、发布者的个人属性（性别、年龄与种族）、以及与其他用户的互动情况等方面，展开对用户在社交网站上信息分享时心理的研究。

5.1.1 用户参与度

图 5.1 展示了使用#like4like 标签的帖子数量与参与发帖的用户人数的增长速率，其中 x 轴表示以季度为单位的时间轴，主 y 轴表示发布了#like4like 内容的数目与参与用户的增长速率（以 2012 年第一季度为基准），次 y 轴表示平均每位参与用户发布的帖子数量的增长速率。若某一季度的增长率为 1，则表示该季度发布的帖子数量或者参与用户人数与基准季度，即 2012 年第一季度持平；若某季度的增长率为 10，则表示该季度发布的帖子数量或者参与用户人数相较于基准季度增长了 10 倍。

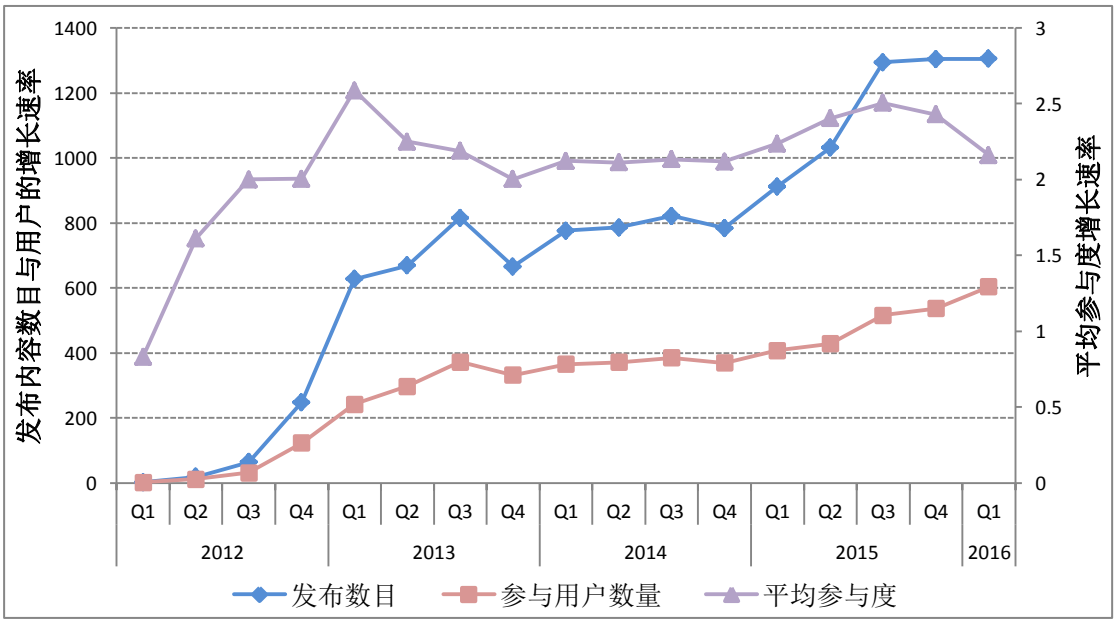

图 5.1 #like4like/#l4l 内容的发布情况

由图 5.1 可以看到在 2013 年的第一季度，#like4like 标签的发帖数量与参与用户人数出现了第一次骤增，相较于前一季度共 36,242 篇帖子以及 5,569 位参与用户，该季度共发布帖子 91,655 篇，吸引了 10,920 位用户的参与。而后在 2015 年第三季度，#like4like 标签的使用出现第二次激增，该季度发布的帖子数目相较于前一季度增加了近 4 万条，用户人数增长超过 20%。我们进一步计算平均参

与度后发现，平均参与度在 2013 年第一季度时达到最高值，为平均每人发布 8.39 条内容；之后有所下降，但基本稳定在平均每人每季度发布 7.21 篇帖子。由此可以看出，#like4like 标签从 2010 年 12 月第一次出现至今，始终保持着长盛不衰的趋势，吸引着越来越多的用户参与。换言之，越来越多的用户乐意于使用这一类热门标签来与别人分享自己发布的内容。

5.1.2 获得的关注度

图 5.2 展示了从 2012 年第一季度至 2016 年第一季度，添加#like4like 标签的帖子受欢迎程度的变化情况。蓝色折线段表示平均每条内容获得的点赞数量，虽然在 2011 年至 2013 年第一季度的两年时间里，平均每条内容获得赞的数量有所波动，但是在之后发布内容数量与参与用户人数都稳步增长的同时，我们看到平均每篇帖子获得的点赞数量也在稳步上升。2013 年第一季度平均每篇帖子仅获得 21.16 个点赞，而截止至 2016 年第一季度该数值上升至 49.41。然而与平均获得点赞数量不同的是，平均每篇帖子获得的评论数量并没有出现明显的起伏，而是始终徘徊在 2 条上下。

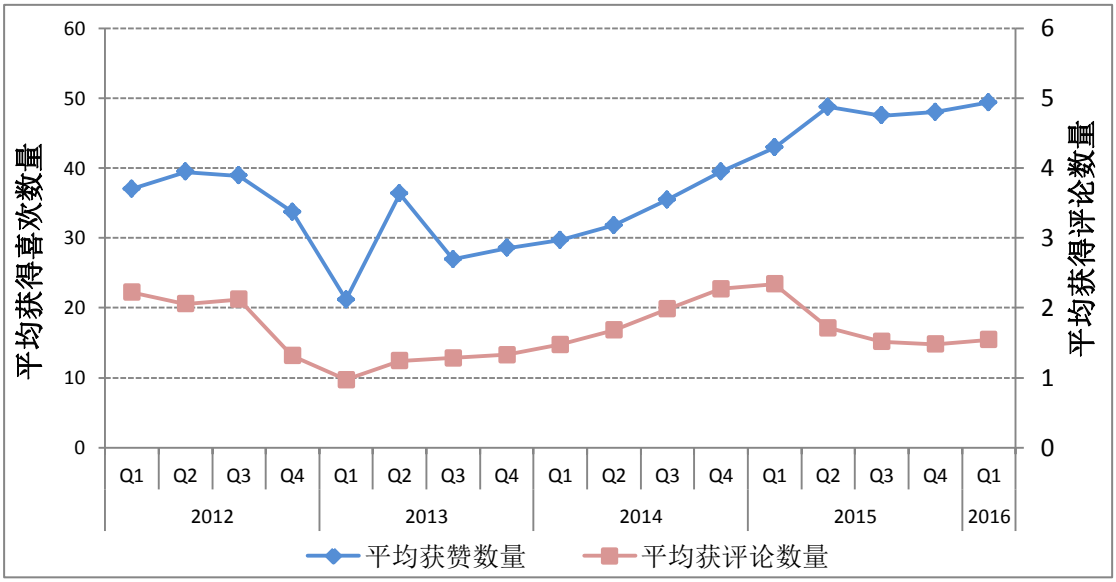

图 5.2 #like4like/#l4l 内容获得的喜欢与评论的情况

Moira Burke 等人将用户在社交网络上的互动交流行为分为三类，分别是点赞式交流（one-click communication）、广播式交流（broadcast communication）以及创作式交流（composed communication）[30]，这三种行为在用户交互程度上循序渐进。点赞行为属于点赞式交流，是非常机械性地动作，并没有实质的交流内容；而评论则属于创作式交流，也就是说评论更需要用户之间发自内心的交流。而我们在此获得的点赞数量与评论数量各自不同的发展趋势，恰好印证了这一观点，并进一步说明了#like4like 这一类标签对于普通用户而言，尽管可以提升用户帖子的曝光度，但是用户之间并不会产生发自内心的交流沟通，大多都只是停

留在非常浅层的关注层面。

5.1.3 用户的个人属性

用户个人属性分布情况 通过用户性别、种族以及年龄等一系列个人属性的数据，我们得以进一步发现发布带有#like4like 标签帖子的用户特点。图 5.3 分别展示了：a）女性用户比例随时间的变化；b）亚裔用户比例随时间的变化；c）各年龄段用户比例随时间的变化。

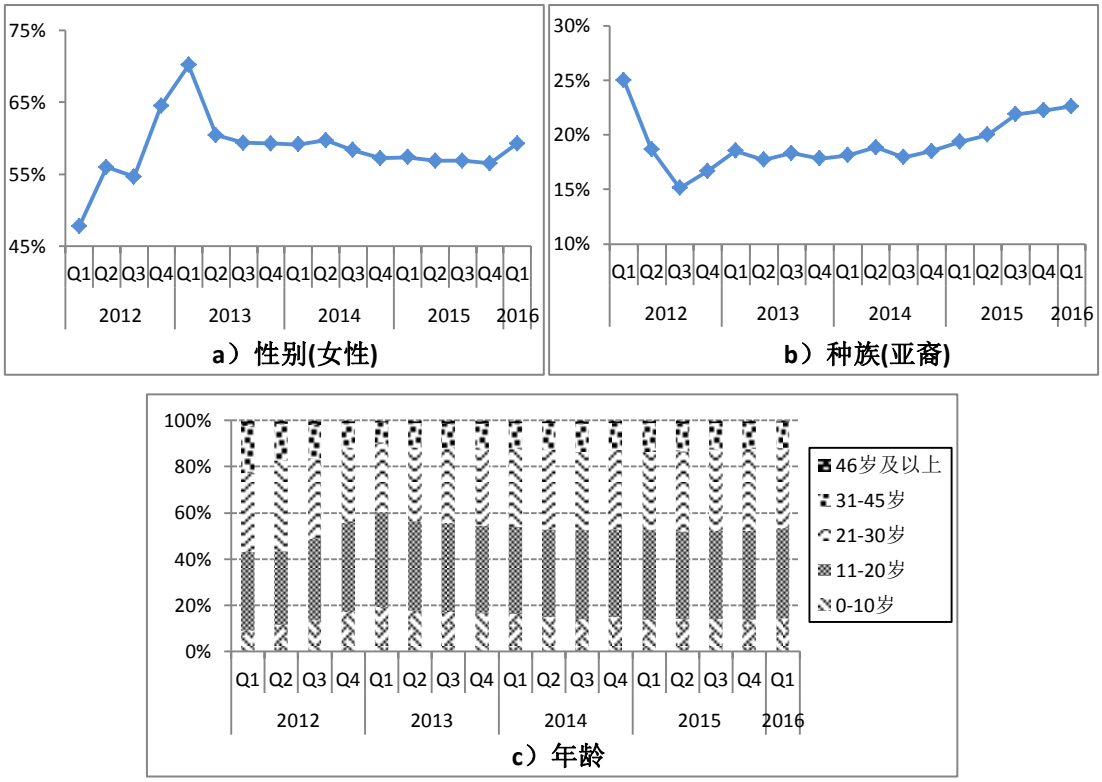

图 5.3 发布#like4like/#l4l 内容用户个人属性的分布情况

根据多项数据[20,21,22,23,24]表明，Instagram 网站上全球女性用户的比例超过 60%。从 5.3a）中可以看出虽然女性仍然占据着主导地位，2013 年第一季度时女性用户的比例高达 70%。但是随着时间的推移，女性比例略有下降，意味着有更多的男性加入到了发布#like4like 内容的行列中来。根据我们获取到的数据，如 5.3b）所示，发布#like4like 内容的用户中亚裔用户的比例正在不断上升，由 2012 年第三季度的 15.12%上升至 22.65%。相应地，白人用户比例逐年下降，而非裔用户比例始终保持在 5%左右，未有明显变化。5.3c）中显示了发布#like4like 内容的用户年龄分布情况。由图中可以看出，年龄段在 20 岁以下的用户比例由最初的

[20] https://www.quora.com/What-is-the-male-female-ratio-on-Instagram，访问于 2016 年 5 月
[21] http://wersm.com/men-vs-women-on-instagram/，访问于 2016 年 5 月
[22] http://www.pewresearch.org/fact-tank/2015/08/28/men-catch-up-with-women-on-overall-social-media-use/，访问于 2016 年 5 月
[23] http://www.pewresearch.org/fact-tank/2013/09/12/its-a-womans-social-media-world/，访问于 2016 年 5 月
[24] https://www.brandwatch.com/2015/01/men-vs-women-active-social-media/，访问于 2016 年 5 月

43%上升并稳定在 53%。其中，11 至 20 岁的用户占 39%。而 31 岁以上的用户比例则在不断下降，由 22%下降至 12%。由此可见，年轻用户更倾向于发布带有#like4like 标签的内容来吸引更多陌生人的关注。

交互同质性 交互同质性是由 Miller McPherson 等人提出的一种社会学假设，可以用来解释用户在社交网站上与其他人发生互动的过程中的趋同性[27]。而在我们接下来的分析中，我们将来验证这种社交网络中的同质性是否也能在我们所研究的用户点赞的交互行为中得到肯定的结果，换言之，我们希望验证是否用户会收到更多的来自同性别、同种族、同年龄段的用户的点赞。

我们根据同质性的定义，对用户的性别与年龄，分别给出了以下算术定义。

性别同质性：$H_{Gender} = \frac{F_{male}+F_{female}}{F_{interactions}}$　　（6）

其中，$F_{interactions}$ 表示用户互动的总数，F_{male} 与 F_{female} 分别表示男性与男性之间的互动数量和女性与女性之间的互动数量。因此，*H_Gender* 衡量了在所有的用户交互中，同性别交互的占比，取值区间为[0, 1]，若 *H_Gender* 等于 1 则代表了所有的用户点赞行为都发生在同性别之间。

种族同质性：$H\_Race = \frac{F_{asia}+F_{africa}+F_{white}}{F_{interactions}}$　　（7）

与性别同质性相类似，$F_{interactions}$ 表示用户交互的总数，而 F_{asia}、F_{africa} 与 F_{white} 则分别代表亚裔、非裔与白人三个种族各自内部发生的交互数量。因此与 H_Gender 类似，H_Race 衡量了同种族交互的占比，取值区间为[0, 1]，H_Race 为 1 表示所有的用户点赞行为都发生在同种族人之间。

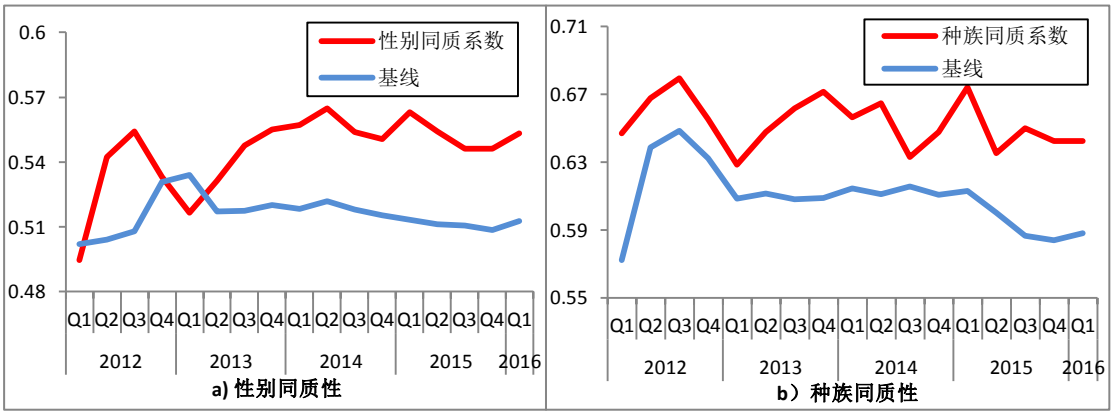

图 5.4 性别同质性与种族同质性

从图 5.4 中得出结论，社交网络上的点赞确实会更多地发生在相同性别和相同种族的用户之间。根据我们的数据采集情况，截止至 2016 年第一季度，性别同质系数逐渐趋于平稳，大约为 0.55，而性别同质系数的基线大约稳定在 0.51。这一同质性在各个种族人群中表现得更为明显。种族同质系数大约在 0.65 左右徘徊，而基线在 2015 年第三季度之后下降至 0.59 以下。

年龄同质性：$H\_Age = RMSE(A_i, A_j)^{-1}$　　（8）

与前两者有略微不同的是，由于用户的年龄是完全离散的数值，我们在此只能通过计算整体的点赞用户与发帖用户之间的年龄差距来衡量年龄同质性。RMSE 指的是均方根误差，在此我们使用均方根误差而不使用绝对差是为了增加当点赞用户与发帖用户年龄差较大时对整体均值的影响。若 $H\_Age = 1$，即 $RMSE(A_i, A_j) = 1$，也就是说整体点赞用户与发帖用户之间的年龄均方根误差为 1 岁。所以 H_Age 的取值区间为(0, +∞)，数值越大，意味着年龄差越小，具有越高的年龄同质性。

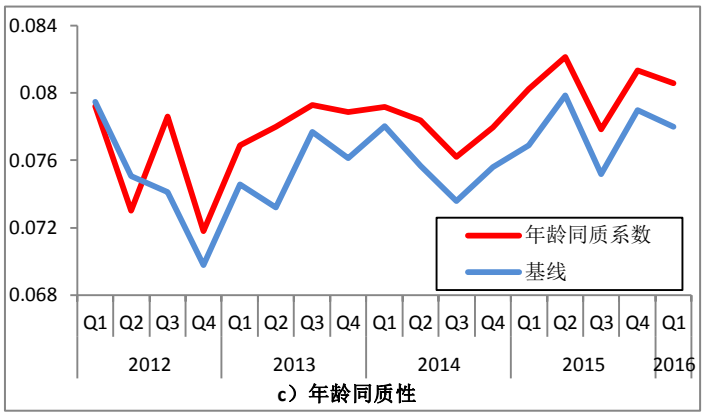

图 5.5 年龄同质性

图 5.5 展示了年龄同质性的分析结果。排除前期阶段数据的不稳定表现，年龄同质系数始终略微高于基线。2016 年第一季度，年龄同质系数约为 0.081，即发帖用户与点赞用户的年龄均方根误差大约为 12.35 岁；而该季度基线的数值为 0.078，表示发帖用户与点赞用户的年龄均方根误差大约为 12.82 岁。但是，相较于性别和种族而言，年龄同质系数与基线的差异非常小，导致这一结果可能的原因是用户的年龄是一个相对分散的离散数值，而在所有使用 Instagram 的用户中有约 90%的用户处于 20-40 岁的年龄段内，这意味着任意两个用户之间的年龄差有限。尽管我们已经通过使用均方根误差的方法将年龄差进行了放大，但是受到总量的影响，不会对年龄同质性系数造成更大的影响。所以我们认为在此，年龄同质性的表现不明显。

5.2 关注度与原贴内容的关联分析

在这一章节中，我们将对用户发表的带有#like4like 标签的帖子内容进行分析，从标签类型、用户类型以及国家文化三个角度来探究并验证哪一类帖子内容更容易受到更多人的喜爱。

5.2.1 与标签类型的关系

首先，我们对标签进行分类，标签的分类参考了 Yelena Mejova 等研究者的

实验结果[28]，稍作调整后，我们将标签共分为以下 7 类：

- 情感类：表达情感、个人意见等标签；
- 健康类：生活方式、饮食方式等标签；
- 社交类：事件、活动、节日等于社交生活有关的标签；
- 地点类：地理位置、城市名、国家名等标签；
- 食品类：特定食物名称、饮料名称等标签；
- 时间类：描述年、月、日、时刻等一类的标签；
- 其他：其他不属于上述 6 类标签的标签。

表 5.1 进一步列出了标签的分类情况。

图 5.6 和图 5.7 分别展示了不同标签分类的帖子受欢迎程度的情况，即他们平均获得的点赞数量与评论数量。在此，我们使用标签来为发布的内容分类，分类方法为若该篇帖子包含了至少 1 个该标签类别中的标签，则这条内容属于该类发布内容。比如，某一条发布内容使用了标签#like4like #friend #happy #summer，那么这条发布内容将被归入 3 个标签类别，分别是情感类（#like4like，#happy）、社交类（#friend）和时间类（#summer）。

表 5.1 标签分类情况

分类	数量	举例
情感类	115	like4like, l4l, happy, love 等
健康类	136	nutrition, organic, sport, vegetarian 等
社交类	134	friend, igers, holiday, instagood 等
地点类	810	jakarta, Italy, home, japanese 等
食品类	657	banana, beef, beer, breakfast 等
时间类	52	summer, monday, midnight, weekend 等
其他	176	adventure, heaven, homemade, tao 等

如图 5.6 所示，我们发现，健康类的发布内容获得了最多的点赞，平均获得 52.81 个点赞（中位数为 36）。地点类的发布内容获得的点赞仅次于健康类，它们平均获得 44.95 个赞（中位数为 30）。而不同于点赞的数量在不同类别的帖子之间有明显的区别，评论数量在不同类别的内容之间几乎相同，如图 5.7 所示。正如第 5.1.2 节中提到的，评论相对于点赞，具有更强的社交性，所以我们认为可能的原因是即使使用热门标签能够有效提升帖子的曝光度，发自内心的语言交流仍然极少发生，并不会因此而有任何的影响。但是我们仍然惊喜地发现，从整体分布而言，健康类的帖子可能获得更多的评论，平均每篇帖子获得 2.15 条评论，中位数为 1。

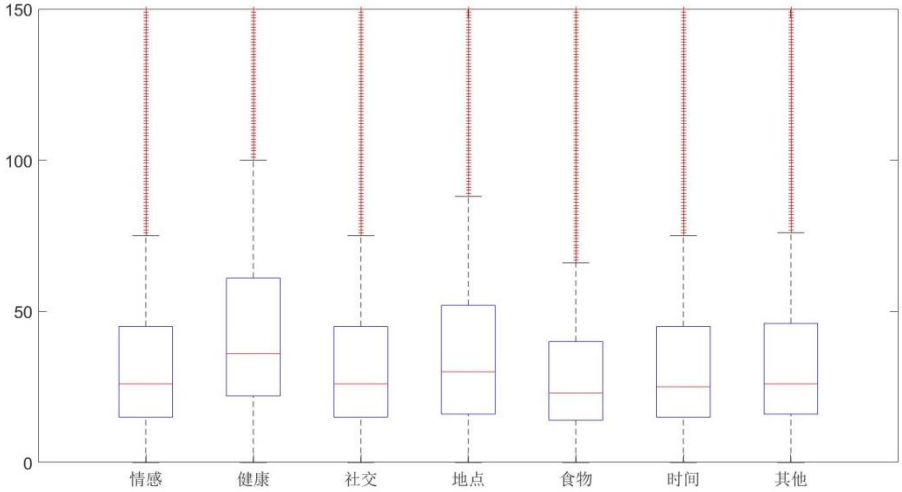

图 5.6 各类别内容平均获得的点赞数量的分布盒图

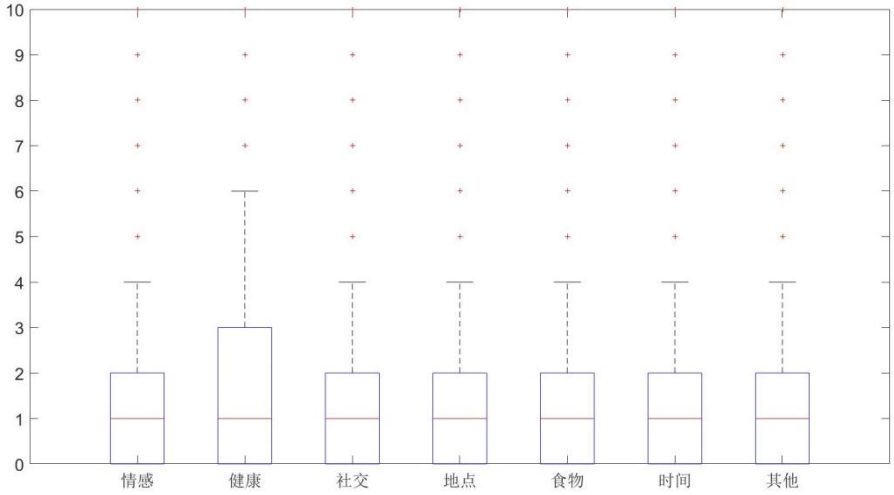

图 5.7 各类别内容平均获得的评论数量的分布盒图

5.2.2 与用户类型的关系

我们根据用户发布的#like4like 的次数与内容，将所有用户分为 3 类：

- 单次用户：在 Instagram 上仅发布过 1 次带有#like4like 标签的帖子的用户，共 50,311 人；
- 旅行用户：在多个国家发布过带有#like4like 标签的内容的用户，共 2,635 人。我们获取了用户发布的所有带有签到地理坐标信息信息的#like4like 标签的内容，并进行了地理坐标的转换，获取用户签到地点的国家信息。
- 所有用户：我们获取的所有发布过带有#like4like 标签的帖子的用户，共 143,586 人。

表 5.2 列出了用户分类的具体信息。我们发现，虽然旅行用户人数非常少，仅有 2,635 人，占总人数的 1.84%，但是他们异常活跃，共发布了 137,533 条带

有#like4lik 标签的帖子，占总发布内容数量的 7.77%。而他们的平均发帖量更是高达平均每人发布 52.19 条，相较于整体均值每人 12.33 条高出将近 40 条。而在帖子的受欢迎程度方面，旅行用户获得的关注也是最多的，平均每条内容获得 53.07 个赞，1.94 条评论。相较而言，单次用户获得的关注度就远不及旅行用户，平均每条内容获得 35.01 个赞，低于整体均值。

表 5.2 不同类别用户发布内容的情况

	用户数量	发布内容数量	平均发帖	平均获赞	平均评论
单次用户	50,311(35.04%)	50,311(2.84%)	1	35.01	1.87
旅行用户	2,635(1.84%)	137,533(7.77%)	52.19	53.07	1.94
所有用户	143,586	1,770,643	12.33	39.29	1.61

为什么同样发布了带有#like4like 标签的内容，单次用户与旅行用户发布的帖子的受欢迎程度，尤其是获得的点赞的数量会有如此巨大的差异？我们希望通过对照片的文本信息进行分析得到相应的结果。由于大部分用户会在发布一条内容的时候使用多个不同的标签，而这些标签往往可以从侧面反映出他们发布的图片内容，所以，我们接下来希望通过对这三类用户使用的标签进行分析。

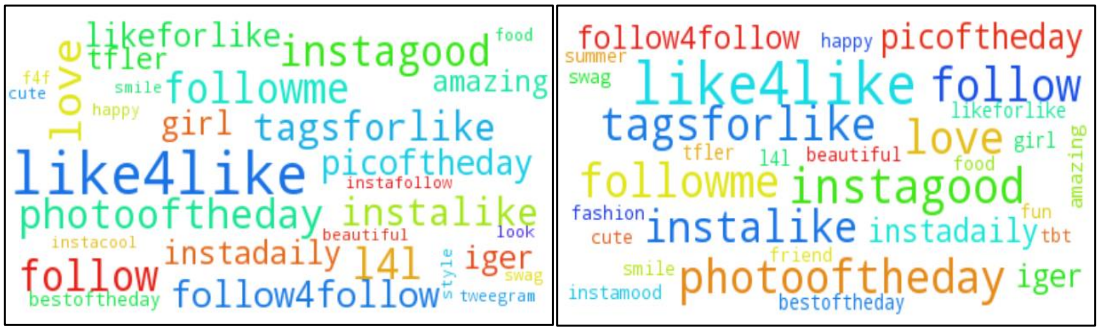

图 5.8 单次用户（左）与旅行用户（右）使用次数排名前 30 的热门标签词云

表 5.3 三类用户各类标签使用情况

分类	情感	健康	社交	位置	食品	时间	其他
单次用户	41.83%	0.91%	24.84%	2.35%	3.16%	2.70%	24.21%
旅行用户	37.85%	0.94%	25.97%	3.69%	3.41%	2.80%	25.34%

通过对 3 类用户使用的标签的分析，结合表 5.3 与图 5.8，单次用户较多地使用了情感类标签，比如排名前 30 的标签当中绝大部分都是类似#like4like、#love、#instalike、#followme 等一系列表达希望别人给自己点赞愿望的标签，他们希望通过发布这些内容吸引更多的关注；而旅行用户则使用了更多的例如#friend、#fun、#smile 等表达正面情绪的标签。另外，我们从表 5.3 中发现，旅行用户相较于单次用户，使用了更多的地点类标签，这极有可能是因为旅行用户喜欢在自己的旅途中分享一些带有地域性标志的内容；旅行用户还使用了更多的健康类与食品类的标签。根据这些数据我们推断，旅行用户会频繁地分享自己在旅行中的点点滴滴，可能是特色景点，也可能是特色食物等等，而这些有特色的内容极有

可能是他们发布的内容更受欢迎的主要原因之一。

5.2.3 与国家文化的关系

通过获取用户的所有签到信息，我们对用户的国籍进行确认，我们将在同一个国家有 10 次以上签到信息且仅有一个国家签到信息的用户认为是该国公民，表 5.4 列出了排名前五的国家用户的发帖信息。

从表 5.4 中可以看到，意大利拥有最多的发布了#like4like 内容的用户，而印尼和巴西是最活跃的两个国家。但是有趣的是，印尼虽然拥有最多的活跃用户，他们发布的帖子的受欢迎程度却是最低的，无论是获得的点赞数量，亦或是获得的评论数量，都远远低于排名前五的其他四个国家。所以，我们进一步对这五个国家用户发布内容的文本信息进行分析。

表 5.4 发帖量排名前 5 的国家用户数据

排名	国家	发帖用户	人口基数	发帖比例	平均获赞	平均评论
1	意大利	4,365	12,209	35.75%	43.04	1.51
2	美国	3,927	12,486	31.45%	48.77	2.53
3	印尼	3,571	7,489	47.68%	31.95	1.42
4	巴西	2,875	6,222	46.21%	42.09	1.81
5	英国	1,014	5,997	16.91%	41.01	1.59

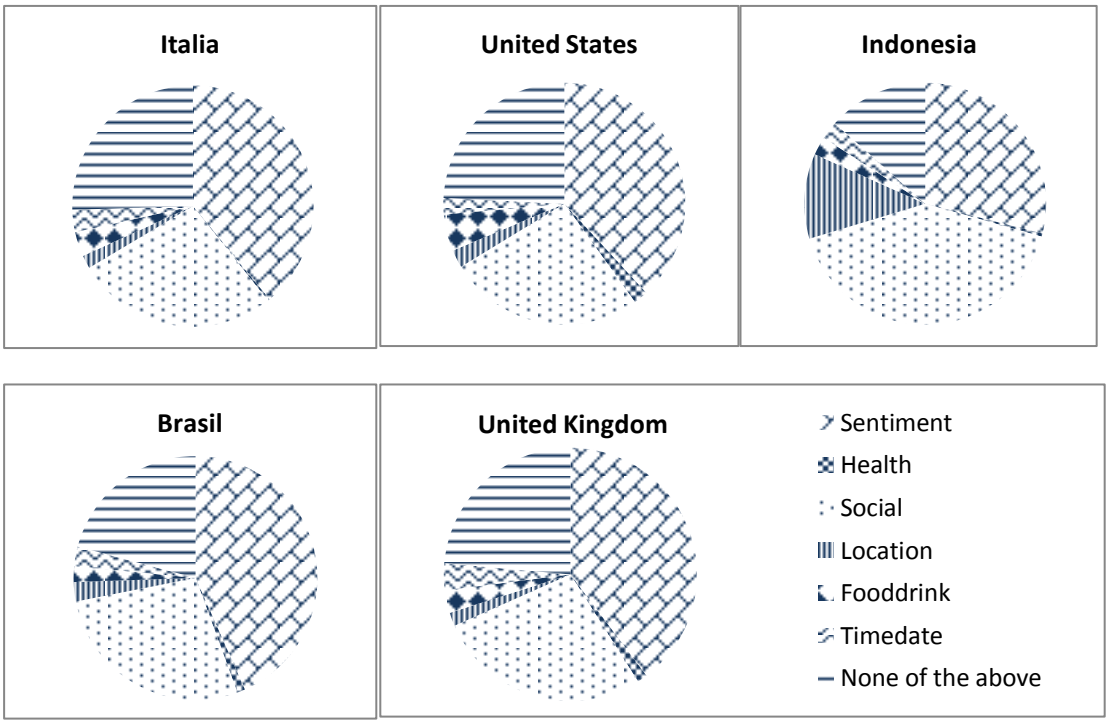

图 5.9 发帖量排名前五的国家各类标签的使用情况

图 5.10 印尼用户使用最多的 30 个热门标签的词云

如图 5.9，各个国家使用的标签类别有较为明显的区别。美国相较于其他 4 个国家而言，使用了更多的健康类以及食品类的标签，英国次之。而令人奇怪的是，印度尼西亚的用户使用了最多的社交类标签，使用比例远远高于其他所有 4 个国家。我们使用词云对印尼用户使用的标签进行了可视化，如图 5.10 所示。我们发现，印尼用户非常喜爱使用与网上购物相关的标签，如#olshopindo、#trustedolshop、#onlineshop 等，这一类标签在其他国家都非常少见；而如#bajulucu、#bajumurah、#jualanmurah 等标签则是直接指出了网上购物具体的网络平台。我们进入 Instagram 网站查询了相关的内容后发现，使用这一类标签的发布的内容，大多与网上购物有关，因此我们推断，印度尼西亚用户的高活跃度原因在于他们使用#like4like 这一类热门标签希望借此来对网购产品进行推广与营销。许多研究者在对用户对于互联网营销广告的态度进行研究后发现，用户都不喜爱这一方面的广告内容，更少与广告发布者进行对话[32][33]，因此，如表 5.4 中的结果显示，印度尼西亚用户虽然拥有着高活跃度，却因为他们发布的大多数是广告而最不受欢迎。

5.3 用户互动行为分析

2016 年 5 月期间，我们随机采集了 11 天共 331 篇带有#like4like 标签的新帖子，并对发帖用户与点赞用户的行为进行了追踪。在本节中，我们将基于这些数据，对用户的信息分享之后的社交互动行为进行分析。

5.3.1 点赞与回赞的发生时间

点赞的发生时间如图 5.11 所示，55.32%的点赞发生在带有#like4like 或#l4l 标签的内容发布后一个小时内，而 94.01%的点赞发生于内容发布后的 24 小时之内。仅有大约 6%的点赞发生在内容发表后的 24 小时之后。

回赞的发生时间 除去用户没有回赞的情况，发布#like4like 内容的用户回赞其他用户的反应速度非常快。如图 5.12 所示，所有的回赞都发生在用户收到点赞之后的 15 分钟之内，其中接近于 90%的回赞发生在收到点赞的 5 分钟之内（回赞

了 242 位点赞用户主页上共 556 条发布的内容）。

基于上述点赞时间与回赞时间两方面的数据结果，可以看出，对于某一条特定的发布内容，其热度在发布后 1 小时就会出现大幅度减弱，而在超过 24 小时之后就只有非常少数的关注了。但是用户的回赞的反应速度相对而言更加迅速，所有的回赞都发生在收到点赞后的 15 分钟之内。

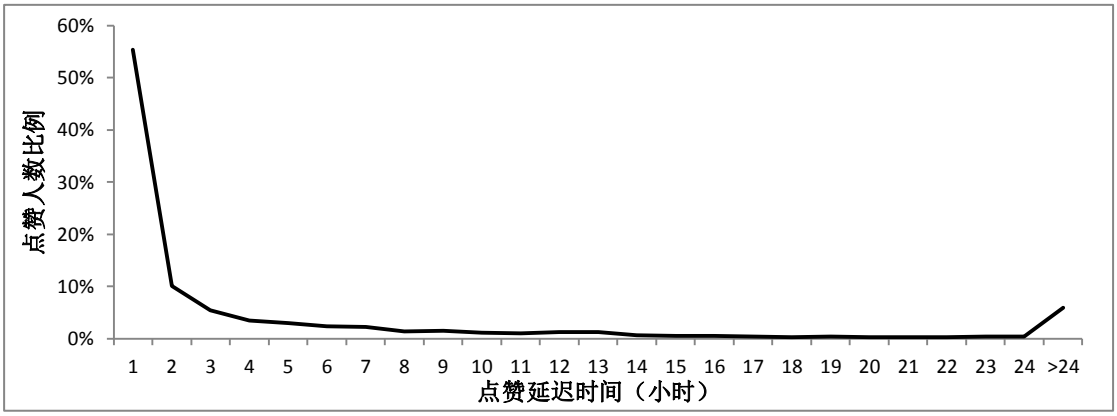

图 5.11 用户点赞的反应速度

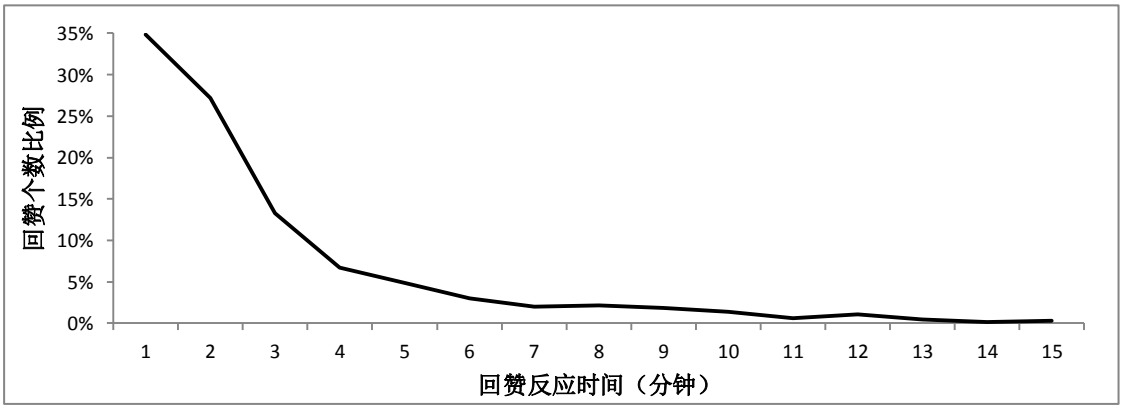

图 5.12 用户回赞的反应速度

5.3.2 获得回赞的用户关系

表 5.5 用户的回赞情况分布

回赞数量	0	1	2-5	>5
用户人数	175	57	66	34
比例	52.87%	17.22%	19.94%	10.27%

表 5.5 列出了总共 331 位发帖用户的回赞情况。从整体比例上来看，仅有 6.11%的点赞用户得到了来自原发帖用户的回赞，比例相当低。在所有发帖用户中，52.87%的用户没有回赞任何一名为其点赞的用户；而仅有 10.27%的用户回赞了 5 名以上的点赞用户。然而，这 331 篇帖子平均每篇获得 57 个赞。这一结果表明了，使用#like4like 这一类标签的用户并没有表现出强烈的社交意愿，不会真正地落实“like for like”的承诺，回复点赞的用户。而他们使用这些标签的用意

只是希望收获更多的关注，与更多的人分享自己的信息。

我们希望通过进一步分析发帖用户与点赞用户之间的关系来了解用户回赞行为的针对性。我们将用户在 Instagram 上的社交关系定义为 2 种，好友及陌生人。在此，由于受到数据获取权限及各方面因素等的限制，我们无法直接获得任意两个用户之间是否为好友的信息，所以我们选择了一种间接的方式来判断这一用户关系。我们通过查看两个用户之间是否发生过互动来判断两个用户是否为好友或者是已经认识的人。换言之，如果两个用户在我们采集数据的时间点之前没有互相点赞等一系列的交互行为，那么我们就认为这两个用户为陌生人。

表 5.6 好友或陌生人与发帖用户发生互动的比例

	点赞用户中好友的比例	好友获得回赞的比例	陌生人获得回赞的比例
整体	20.17%	15.03%	3.38%
有回赞行为的用户		33.62%	7.27%

如#like4like 的热门标签确实能够为发帖用户带来更多的陌生人的关注，大约 80%的点赞来自于陌生人。这一结果并不意外，#like4like 早已成为了 Instagram 上热度排名第二的热门标签，自然会吸引许多人对这一标签的关注，而获得更多来自陌生人的关注也是用户使用这一类标签的初衷。表 5.6 中列出了两种社交关系的用户获得回赞的比例。尽管整体而言获得回赞的用户比例非常低，只有 15.03%的好友会在点赞之后获得反馈，然而陌生人获得回赞的比例更低，仅为 3.38%。而如果仅计算有回赞行为的发帖用户时，这一差距变得更为悬殊，好友获得回赞的比例为 33.62%，而陌生人仍然只有 7.27%。由此，我们发现，用户并不是对所有用户都一视同仁的，用户更可能回复好友的点赞，而极少对陌生人的点赞做出反应。根据 Mark S. Granovetter 所提出的观点，社交网络中用户的关系可分为强关系（strong ties）与弱关系（weak ties）两种[26]，强关系比较多的是在现实生活中也互相熟悉的好友关系，而弱关系则更多依托于社交网络平台，关系的双方在现实生活中更可能是陌生人或者不熟悉的人。即使点赞本身是社交网络中强度最弱的一种互动模式，用户也不会因为获得了点赞而去回复弱关系用户。而在此之前，也有相关研究证明了弱关系用户互动发生的小概率。例如 Amanda Lenhart 和 Mary Madden 的调查发现，接近 80%的用户在社交网站上收到了来自陌生人的邀请之后会直接删除或不予理会[25]。通过对发帖用户与点赞用户之间的关系的分析，我们得到初步结论，#like4like 等一系列类似的热门标签并不会从根本上改变用户对于陌生人与自己社交的态度，他们更多的只想通过使用热门标签来吸

[25] 数据参考 Amanda Lenhart 和 Mary Madden 于 2007 年 4 月 18 日发表于 Pew Research Center 网站上的文章 Friendship, Strangers and Safety in Online Social Networks。

引更多人的关注，与更多人分享自己发布的内容，而并没有表现出与点赞用户有进一步交流互动的积极性，且与陌生人发生互动的可能性远低于好友。

5.3.3 回赞内容

最后我们希望通过用户回赞的内容来辨别用户回赞的目的。如图 5.13 所示，78.99%的用户在收到用户的点赞之后给相应的用户回复了一个赞，10.03%的用户回复了两个赞。另外，我们统计了所有用户回赞的第一篇帖子的序号，发现在所有回复的点赞中，79.71%的回赞从用户主页上的第一篇帖子开始，而 8.55%的回赞从用户主页上的第二篇帖子开始，仅有 4.13%的回赞发生在用户主页上第 10 篇帖子之后。而我们进一步分析了用户回赞的帖子的文本信息后发现，超过 50%的帖子不包含任何文本内容，其余的帖子标签也具有很大的随机性。由此证明，大部分用户的回赞并不介意帖子的内容，而只是为了回赞而回赞。

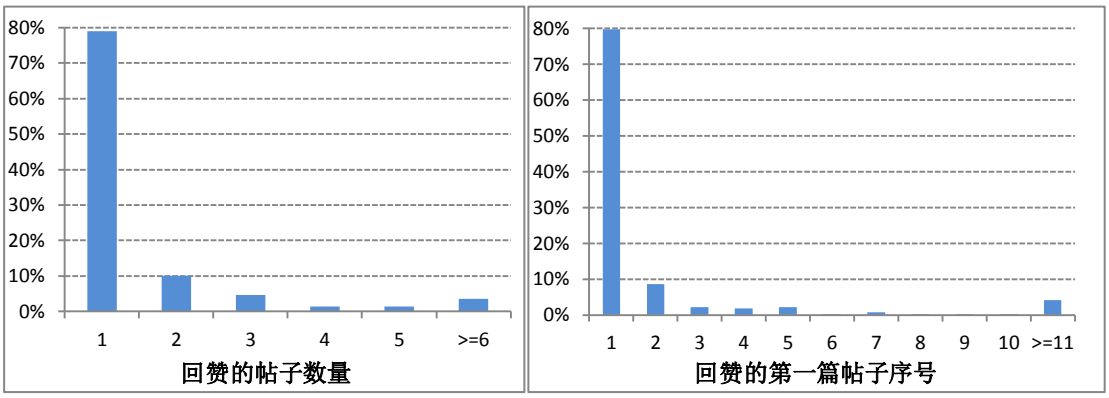

图 5.13 回赞内容统计情况。左侧为所有用户回赞的帖子数量的人数统计分布情况；右侧为所有用户回赞的第一篇帖子序号的人数统计分布情况。

5.4 本章结论

本章以#like4like 标签的相关内容作为切入点，对用户在 Instagram 平台上的信息分享时的心理展开了实证分析研究。我们发现：

- 用户更乐意使用热门标签来提升自己帖子的曝光度。从 2012 年第一季度至 2016 年第一季度，参与用户人数增长 600 倍，发布帖子的总数增长超过 1300 倍，人均发帖数量翻了一番，并仍然保持着上涨的趋势。
- 热门标签有效帮助用户提升帖子的曝光度。2012 年第一季度发布的所有帖子平均每篇获得 37.00 次点赞，而 2016 年第一季度时该数值上升至 49.41，且其中 80%的点赞来自于陌生人。但是评论的数量无明显波动。
- 用户之间的互动具有同质性。用户更可能获得同性别、同种族的用户的点赞。健康类内容最受欢迎，平均每篇帖子获得点赞 52.81 次；地点类内容次之，平均每篇帖子获得点赞 44.95 次。

- 用户没有表现出与点赞用户之间互动的积极性。根据回赞的比例来看，仅有 6.11%的点赞用户得到回赞，陌生人得到回赞的比例更低，仅为3.38%。

用户使用#like4like 这一类标签会为他们带来更高的曝光率，更多的点赞，但是并不会获得更多发自内心的交流。而用户本人并未遵从标签本意，他们并没有表现出与点赞用户有进一步互动的积极性。

第六章 讨论

6.1 限制

在我们对社交网络真实用户行为的分析研究中，由于受到数据获取权限等各方面因素的影响，我们的数据集存在着一定的限制，从而也会对我们的分析方法产生一定的影响。

- 数据不完整。Twitter 和 Instagram 作为世界领先水平的两大社交网络平台，他们对于各自平台上用户个人信息的保护是非常到位的。并且，目前也没有发生过严重的信息泄漏事件。所以，我们的数据采集只能通过一定的取样方法，调用这两大平台官方提供的 API 接口，获得部分合法的数据，这也就导致了部分重要数据我们无法获得。例如，在 Instagram 网站上，一旦用户设置为隐私用户，不仅陌生人无法浏览他们的个人主页，我们也无法通过任何渠道获取这些隐私用户的任何数据信息。
- 研究对象的局限性。由于我们的采样方法，我们收集的数据集中用户具有局限性。在对用户网上信息分享心理的分析中，我们通过在全球范围内随机取样获得种子用户；而在对用户访问控制设置情况的分析中，我们将研究对象限定为纽约城市范围内的用户。尽管我们的数据集足够大，例如我们在对用户访问控制设置情况的分析中采集了约 15 万 Twitter 用户和超过 28 万的 Instagram 用户作为研究对象，但是仍然会有一些因地域、民族等各方面客观原因而导致的取样不全面。但是在我们对取样结果进行检测后，我们发现与 Twitter 和 Instagram 网站官方提供的一些用户数据保持一致，所以我们的研究结果仍然是有意义和代表性的。
- 研究对象社交圈信息的不足。无论在对用户信息分享心理的分析中，还是对用户访问控制设置情况的分析中，我们都希望可以获得更多的用户社交圈内好友的信息。例如在第 4.2.3 节中，我们发现用户在将个人主页的访问控制设置由公开修改为私密之后，他们会删除个别原本在社交圈内的好友，但是我们无法分析获取被删除的“好友”的具体信息。导致这一结果的主要原因是 Twitter 和 Instagram 提供的 API 都有这方面的限制。我们无法获取 Twitter 用户社交圈内好友的信息[26]，而 Instagram 也仅能提供某一用户非常少量的粉丝信息和被关注者的信息[27]。

[26] 可参考 Twitter API 说明：https://dev.twitter.com/rest/public/rate-limiting
[27] 可参考 Instagram API 说明：https://www.instagram.com/developer/endpoints/relationships

6.2 建议

我们通过对真实用户在社交网络中的行为表现的分析，既了解了用户在Twitter和Instagram上对个人隐私的保护行为，也了解了他们在使用热门标签信息分享时的心理。基于这些分析，我们给出以下两条针对当前以Twitter和Instagram为代表的一系列社交网络实行的访问控制机制的建议：

- 在Twitter、Instagram等一类社交网站上，访问控制设置可以被自动预测。根据我们的实验分析，在对真实用户在社交网络上的个人属性与线上活动等静态数据进行训练之后，访问控制设置预测的ROC曲线下面积达到0.70。而根据我们进一步的动态分析发现，隐私用户与公开用户在动态的行为表现中也有非常明显的区别，因此，我们认为在如Twitter和Instagram这一类的社交网络中，可以用自动预测的方式替代用户必须手动设置的机制，或者可以将预测的结果作为推荐设置提示用户，这样可以有效防止用户因个人疏漏而导致的信息泄露。
- Twitter和Instagram现有的全局访问控制机制已经不能很好地满足用户的双向需求，应实行更细粒度的访问控制机制。通过对用户信息分享心理的分析，我们发现大部分用户的需求是矛盾的，而全局访问控制设置无法兼顾。用户需要在个人信息受到保护的同时，又能够通过发布某些内容来吸引更多来自陌生人的关注。那么，这就导致了在现行的全局访问控制机制模式下，用户必须频繁地修改个人设置，以满足自己的社交需求。然而，一旦用户为了某条内容而关闭访问控制设置，那么他的所有个人信息都将被暴露在陌生人面前，也就无法起到保护隐私的作用。因此，我们建议，在Instagram这一类社交网站上实行更细粒度的访问控制机制，例如可以在启动访问控制设置的同时，由用户选择公开发布某些内容；或者当用户公开个人主页的同时可以私密地发布单篇内容。

第七章 总结和展望

7.1 总结

现代社会，社交网站已经成为了人们日常生活中不可或缺的一部分。而随着大数据时代的到来，任何一个普通用户每天都会在社交网站上产生大量的数据。透过社交网络这个大平台，研究人员能够通过分析数据的方方面面，来获得有意义的信息。通过研究分析，我们回答了在第一章中提出的三个问题。

- 用户在社交网站中的信息保护有哪些行为表现？

本文采集了真实用户在社交网站上的公开数据，应用了数据可视化、自然语言处理等领域的相关技术，从静态表现和动态变化两个方面，对用户的访问控制设置情况进行分析与挖掘。从 2015 年 10 月 15 日至 2016 年 1 月 12 日，Twitter 隐私用户比例上升 0.73%，Instagram 隐私用户比例上升 4.84%。其中，女性、亚裔以及年轻用户更关注社交网络中个人信息的保护。部分用户会频繁地修改个人主页访问控制设置，5.21%的 Twitter 用户平均修改 2.29 次，19.95%的 Instagram 用户平均修改 3.40 次。并且，隐私用户具有较低活跃度、清理好友、发布更具有隐私性内容等行为特点。同时，重大事件与重要节日会影响用户对个人主页访问控制设置的决策。

- 用户在信息分享时有怎样的心理？

我们以 Instagram 平台上排行第二的热门标签#like4like 作为切入点，应用了数据可视化领域中的先进技术，结合社会学中相关理论，分析了用户在社交网站上与好友、陌生人互动的行为表现。我们发现，用户变得更乐意于使用热门标签吸引更多陌生人的关注，从 2012 年第一季度至 2016 年第一季度，参与发布 #like4like 相关内容的用户上涨了 600 倍，人均发布内容翻了一番。平均获得点赞的数量也由 37.00 上升至 49.41，其中约 80%的点赞来自陌生人。但是他们并没有表现出与点赞用户产生进一步交流的积极性。整体的回赞比例为 6.11%，其中超过半数的用户没有回复任何一名用户的点赞，而陌生人得到回赞的比例更低，仅为 3.38%。由此可见，用户仅希望利用热门标签收获关注度，而并不乐意因此产生进一步的社交互动的行为。

- 如何改进现有的访问控制机制？

本文提出了一种全新的社交网站访问控制机制的设计思路，通过运用机器学习的相关技术对用户的访问控制设置进行自动预测。我们运用了机器学习领域中的逻辑回归、随机森林和梯度推进三种不同的分析模型对用户的访问控制设置进行训练与预测，将用户的个人属性数据和静态网上活动数据作为输入数据集，得到的 ROC 曲线下面积最优结果达到 0.70，这一结果从一定程度上验证了自动预

测用户在社交网络中的访问控制设置的可行性。

同时，在结合了上述两方面的分析结果之后，我们发现类似于 Twitter 和 Instagram 这一类社交网站上现有的全局访问控制机制已不能很好地满足用户的需求。用户希望在保护个人隐私的同时，能够通过发布某些特定的内容来获得更广泛的关注。所以应当使用更细粒度的访问控制机制来取代现有的全局访问控制机制，以此来更好地迎合用户使用社交网络时兼顾社交与隐私保护的双重需求。

7.2 展望

社交网络中用户的行为分析还有许多可以进一步完善的方面。首先，可以进一步提高数据集本身的质量。目前我们通过对用户头像照片进行人脸识别获取用户的个人属性信息，并依赖于此展开了一系列的分析。考虑到部分用户会使用非真人照片作为个人头像的可能性，之后我们会通过对用户主页上多张自拍照片进行人脸识别来获取用户的个人属性数据，以此来提高用户个人属性数据的准确性，使得后续的分析结论更可靠。

其次，受曼·惠特尼 U 检验的启发，本文使用盒图规避了对不同数据集均值比较的显著性检验。在后续的研究中，我们希望通过使用更有效、更符合本文研究数据集要求的统计学检验方法，来对均值比较的显著性做出量化的判断。

再次，针对用户社交网络访问控制设置的预测，目前我们仅仅将用户的静态数据，即用户的个人属性以及线上活动的静态数据，作为构建预测模型的输入。但是，通过分析我们也发现，用户会频繁地修改个人访问控制的设置，并且一些外界因素以及用户的动态都会影响用户的访问控制设置决策。因此，在之后的研究中我们会添加动态数据和外界客观因素数据作为训练模型的输入，来提高用户访问控制设置预测的准确率。

最后，本文主要研究的是以 Twitter 和 Instagram 为首的国外社交网站。而今，国内也涌现出了一大批类似的社交网站，例如新浪微博、腾讯微博等平台。未来，我们会将研究延伸到国内的社交网站，并对国内社交网站的访问控制机制也展开深入的研究。

参考文献

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[2] G. Bruns, P. W. L. Fong, I. Siahaan, and M. Huth. Relationship-based access control: its expression and enforcement through hybrid logic[A]. In Proc. 2nd ACM Conference on Data and Application Security and Privacy[C]. ACM, 2012: 117-124.

[3] B. Carminati, E. Ferrari, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. A semantic web based framework for social network access control[A]. In Proc. 14th ACM Symposium on Access Control Models and Technologies[C]. ACM, 2009: 177-186.

[4] B. Carminati, E. Ferrari, and A. Perego. Rule-based access control for social networks[A]. In Proc. IFIP WG 2.12 and 2.14 Semantic Web Workshop, volume 4278 of LNCS[C]. Springer, 2006: 1734-1744.

[5] M. Cha, H. Haddadi, and F. B. K. P. Gummadi. Measuring user influence in Twitter: The million follower fallacy[A]. In Proc. 4th AAAI Conference on Weblogs and Social Media[C]. The AAAI Press, 2010: 10-17.

[6] M. Cramer, J. Pang, and Y. Zhang. A logical approach to restricting access in online social networks[A]. In Proc. 20th ACM Symposium on Access Control Models and Technologies[C]. ACM, 2015: 75-86.

[7] J. Crampton and J. Sellwood. Path conditions and principal matching: a new approach to access control[A]. In Proc. 19th ACM Symposium on Access Control Models and Technologies[C], pages 187-198. ACM, 2014.

[8] R. Dey, Z. Jelveh, and K. Ross. Facebook users have become much more private: A large-scale study[A]. In Proc. 2012 IEEE International Conference on Pervasive Computing and Communications Workshops[C]. IEEE, 2012: 346-352.

[9] P. W. L. Fong. Preventing sybil attacks by privilege attenuation: a design principle for social network systems[A]. In Proc. 32nd IEEE Symposium on Security and Privacy[C]. IEEE CS, 2011: 263-278.

[10] P. W. L. Fong, M. M. Anwar, and Z. Zhao. A privacy preservation model for Facebook-style social network systems[A]. In Proc. 14th European Symposium on Research in Computer Security, volume 5789 of LNCS[C]. Springer, 2009: 303-320.

[11] P. W. L. Fong and I. Siahaan. Relationship-based access control policies and their policy languages[A]. In Proc. 16th ACM Symposium on Access Control Models and

Technologies[C]. ACM, 2011: 51-60.

[12] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing Facebook privacy settings: user expectations vs. reality[A]. In Proc. 2011 ACM SIGCOMM conference on Internet measurement conference[C]. ACM, 2011: 61-70.

[13] M. Mondal, Y. Liu, B. Viswanath, K. P. Gummadi, and A. Mislove. Understanding and specifying social access control lists[A]. In Proc. 10th Symposium on Usable Privacy and Security[C]. USENIX Association, 2012: 271-283.

[14] J. Pang and Y. Zhang. A new access control scheme for Facebook-style social networks[A]. In Proc. 9th Conference on Availability, Reliability and Security[C]. IEEE CS, 2014: 1-10.

[15] J. Pang and Y. Zhang. Cryptographic protocols for enforcing relationship-based access control policies[A]. In Proc. 39th Annual IEEE Computers, Software & Applications Conference[C]. IEEE CS, 2015: 484-493.

[16] J. Pang and Y. Zhang. Location prediction: communities speak louder than friends[A]. In Proc. 3rd ACM on Conference on Online Social Networks[C]. ACM, 2015: 161-171.

[17] J. Pang and Y. Zhang. A new access control scheme for Facebook-style social networks[J]. Computers & Security, 2015, 54:44-59.

[18] M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health[A]. In Proc. 5th AAAI Conference on Weblogs and Social Media[C]. The AAAI Press, 2011: 265-272.

[19] M. Redi, D. Quercia, L. Graham, and S. Gosling. Like partying? your face says it all. Predicting the ambiance of places with profile pictures[A]. In Proc. 9th AAAI Conference on Weblogs and Social Media[C]. The AAAI Press, 2015: 347-356.

[20] F. Souza, D. de Las Casas, V. Flores, S. Youn, M. Cha, D. Quercia, and V. Almeida. Dawn of the selfie era: The whos, wheres, and hows of selfies on Instagram[A]. In Proc. 3rd ACM on Conference on Online Social Networks[C]. ACM, 2015: 221-231.

[21] F. Stutzman, R. Gross, and A. Acquisti. Silent listeners: The evolution of privacy and disclosure on Facebook[J]. Journal of Privacy and Confidentiality, 2013, 4(2):2.

[22] E. Tarameshloo, P. W. L. Fong, and P. Mohassel. On protection in federated social computing systems[A]. In Proc. 4th ACM Conference on Data and Application Security and Privacy[C]. ACM, 2014: 75-86.

[23] J. W. Tukey. Exploratory Data Analysis[M]. Pearson, 1977.

[24] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing

Twitter and traditional media using topic models[A]. In Proc. 33rd European Conference on IR Research, volume 6611 of LNCS[C]. Springer, 2011: 338-349.

[25] Y. Zhong, N. J. Yuan, W. Zhong, F. Zhang, and X. Xie. You are where you go: Inferring demographic attributes from location check-ins[A]. In Proc. 8th ACM International Conference on Web Search and Data Mining[C]. ACM, 2015: 295-304.

[26] Mark S. Granovetter. The Strength of Weak Ties[J]. American Journal of Sociology, 1973, 78(6): 1360-1380.

[27] Miller McPherson, Lynn Smith-Lovin, and James M Cook. BIRDS OF A FEATHER: Homophily in Social Networks[J]. Annual Review of Sociology, 2001, 27(1): 415–444.

[28] Yelena Mejova, Sofiane Abbar, and Hamed Haddadi. Fetishizing Food in Digital Age: #foodporn Around World[A]. In Proc. 10th International AAAI Conference on Web and Social Media[C]. The AAAI Press, 2016: 250-258.

[29] Henry B.Mann, and Donald R Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other[J]. Annals of Mathematical Statistics, 1947, 18(1): 50-60.

[30] Moira Burke and Robert E. Kraut. 2014. Growing closer on facebook: changes in tie strength through social network site use[A]. In Proc. of the SIGCHI Conference on Human Factors in Computing Systems[C]. ACM, 2014: 4187-4196.

[31] Moira Burke, Cameron Marlow, and Thomas Lento. 2010. Social network activity and social well-being[A]. In Proc. of the SIGCHI Conference on Human Factors in Computing Systems[C]. ACM, 2010: 1909-1912.

[32] Cho Chang-Hoan, and Cheon Hongsik John. Why do people avoid advertising on the Internet?[J]. Journal of Advertising, 2004, 33(4): 89-97.

[33] Katherine A. MacKinnon. User Generated Content vs. Advertising: Do Consumers Trust the Word of Others Over Advertisers?[J]. The Elon Journal of Undergraduate Research in Communications, 2012, 3(1): 14-22.

在读期间发表论文

[1] **Minyue Ni**, Yang Zhang, Weili Han and Jun Pang. An Empirical Study on User Access Control in Online Social Networks. In 21st ACM Symposium on Access Control Models and Technologies (ACM SACMAT 16), 2016. (CCF C 类会议)

[2] Weili Han, Zhigong Li, **Minyue Ni**, Guofei Gu and Wenyuan Xu. Shadow Attack based on Password Reuses: A Quantitative Empirical View. IEEE Transactions on Dependable and Secure Computing, 2016. (CCF A 类期刊)

[3] Zeqing Guo, Weili Han, Liangxing Liu, Wenyuan Xu, Ruiqi Bu and **Minyue Ni**. SPA: Inviting Your Friends to Help Set Android Apps. In 20th ACM Symposium on Access Control Models and Technologies (ACM SACMAT 15) , 2015: 221 - 231. (CCF C 类会议)

致 谢

毕业论文的完成，代表着自己的学生生涯即将画上句号。在此，我首先想要感谢我的研究生导师韩伟力老师。在我攻读硕士学位的两年半时间里，韩老师给予我的谆谆教诲让我受益匪浅。在我撰写毕业论文的过程中，韩老师严谨的科学态度和精益求精的工作作风，督促着我不断进步。韩老师严肃认真、孜孜不倦、实事求是的科研精神，也值得我一直努力学习。此外，韩老师除了在学术研究上对我进行指点，在为人处世和生活中也给予我点播、帮助和关心。这些点滴将会使我受益终生。

其次，我要感谢卢森堡大学的庞军老师和张阳同学，以及复旦大学信息安全实验室的所有同学。我与庞老师、张阳同学的合作非常顺利以及愉快，他们也给予了我莫大的帮助与支持。实验室的同学们在科研工作上都十分刻苦和专注，在我有难处的时候给了我许多的帮助、启发和动力。在这样一个有爱、融洽的团队中度过了两年半研究生的学习和工作生涯，是一件非常幸运的事情，我的心中充满感恩。

当然，我的成长同样也离不开父母的培养与心血。父母与我一路同甘共苦，在我低落沮丧时听我倾诉，在我取得进步时为我自豪与骄傲。感谢他们无私的付出，也感谢他们无论何时都会站在我的身后鼓励我，尊重我的选择，支持我的决定，做我最坚强的后盾。

学生生涯的结束意味着与许多同窗即将分别，我们一同成长、一起迷茫、一起对未来充满希望。因为他们，校园生活更多了欢声笑语。

现在的心情是感慨万千，依然清晰地记得当年刚跨入复旦大学校园的我，怀着一颗忐忑与不确定的心来到了这里，开始了软件工程专业的学习。求学生涯的结束，对我的人生而言，是一段全新旅途的起点。这是一段精彩的人生经历，我将永远珍藏在心。

复旦大学
学位论文独创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。论文中除特别标注的内容外，不包含任何其他个人或机构已经发表或撰写过的研究成果。对本研究做出重要贡献的个人和集体，均已在论文中作了明确的声明并表示了谢意。本声明的法律结果由本人承担。

作者签名：_____ 日期：_____

复旦大学
学位论文使用授权声明

本人完全了解复旦大学有关收藏和利用博士、硕士学位论文的规定，即：学校有权收藏、使用并向国家有关部门或机构送交论文的印刷本和电子版本；允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。涉密学位论文在解密后遵守此规定。

作者签名：_____ 导师签名：_____ 日期：_____