arXiv:2601.04134v1 [cs.SI] 7 Jan 2026

名人信息减少网络仇恨并限制其传播*

Eaman Jahani† Blas Kolic 马里兰大学 和麻省理工学院 IDE 马德里卡洛斯三世大学 Manuel Tonneau Hause Lin Daniel Barkoczi 牛津大学 麻省理工学院 南丹麦大学 Edwin Ikhuoria Victor Orozco Samuel P. Fraiberger 密德萨斯大学 世界银行 世界银行和纽约大学 †

摘要

网络仇恨传播迅速，但人们对预防性、可扩展的策略能否遏制网络仇恨知之甚少。我们开展了迄今规模最大的仇恨言论预防随机对照试验：在尼日利亚的 X 平台上进行为期 20 周、针对族群仇恨的消息宣传活动。此前曾参与仇恨言论活动的 73,136 名用户被随机分配接收尼日利亚名人发送的亲社会视频信息。宣传活动使处理期间的仇恨内容减少了 2.5% 至 5.5%，且约 75% 的降幅在随后四个月内持续保持。扩大触达用户受众的比例，降低了处理组和未处理组用户对该用户仇恨帖子的传播放大，使曝光度最高的账号所获仇恨内容转发减少了 50% 以上。可扩展的信息宣传活动无需删除内容，也能限制网络仇恨。

关键词： 随机对照试验；现场实验；社交媒体；仇恨言论；亲社会信息传播；网络效应；信息扩散；内容审核

*我们受益于在牛津互联网研究所社交媒体研究未来研讨会、IC2S2 23、Meta的CSS研讨会以及世界银行DECDI研讨会上获得的反馈。本研究是世界银行发展影响研究计划的一部分。**作者贡献：** E.J.和S.F.牵头、指导并监督了该项目。H.L.负责管理广告宣传活动。D.B.在H.L.的协助下收集了数据。M.T.计算了仇恨分数。B.K.和E.J.在M.T.、D.B.和H.L.的协助下整理了数据。B.K.在M.T.的协助下进行了描述性分析。E.J.在H.L.的协助下完成了其他所有数据分析。E.J.和S.F.牵头撰写了论文，B.K.、M.T.、H.L.和V.O.提供了协助。所有作者均参与了研究设计。

†致谁 通讯可联系：电子邮件：eaman@umd.edu；sfraiberger@worldbank.or g

引言

社交媒体平台重塑了信息的传播方式，使社会运动、抗议活动和公民参与得以迅速协调（Bond et al., 2012; Tufekci, 2017; Enikolopov et al., 2020）。然而，同样由注意力驱动的系统也会放大仇恨内容（Vosoughi et al., 2018; Bail et al., 2018; Finkel et al., 2020; Guess et al., 2020; Milli et al., 2025）。全球调查显示，约三分之二的互联网用户表示曾在网上遭遇仇恨言论，尤其是在社交媒体上（UNESCO, 2023）。既有证据表明，网络仇恨与线下伤害有关，包括缅甸的族群迫害、德国针对难民的袭击，以及西班牙针对移民和 LGBT 群体的仇恨犯罪（M¨uller and Schwarz, 2021; Whitten-Woodring et al., 2020; Ar- cila Calder´on et al., 2024）。

对此，平台主要依赖内容审核，删除或限制违反规则的帖子，以减少用户接触网络仇恨的机会（Gillespie，2018）。尽管审核可以减少仇恨内容，并提高对平台规则的合规程度（Chandrasekharan 等，2017；Kumarswamy 等，2023；Yildirim 等，2023；Horta Ribeiro 等，2023），但它也面临着重要局限。有害内容的传播速度可能快于审核速度；不同用户和语言受到的执法力度并不一致；而扩大人工审核的规模成本高昂（Vosoughi 等，2018；Pfeffer 等，2023；Giansiracusa，2021；Goldstein 等，2023；Tonneau 等，2025a、b）。自动化系统可能会错误分类内容，有时会漏掉不那么明显的有害内容，或删除无害帖子，从而引发对言论自由的担忧（Allen 等，2024；Lee 等，2024；Kaplan，2025）。这些局限凸显了对可扩展方法的需求：在不删除内容的前提下减少网络仇恨。

一种经过检验的非审查式方法是直接反仇恨言论，即通过公开回复仇恨帖子来遏制敌意。既有研究表明，这类回复能够减少处理组用户发表的仇恨言论（Munger, 2017; Siegel and Badaan, 2020; Hangartner et al., 2021; B¨ar et al., 2024; Gennaro et al., 2025b）。然而，直接反仇恨言论本质上属于被动应对，必须等仇恨内容出现并被检测到后才能采取行动，而这时内容往往已经传播开来。此外，这类干预难以扩大规模，因为它们依赖于对海量内容进行实时检测，而这种能力目前主要掌握在平台手中。

一种互补的方法是预防性反仇恨言论，旨在仇恨内容传播之前，触达有可能发布或分享此类内容的用户。通过广告投放定向信息，提供了一种主动且可扩展的渠道；研究表明，这种方法有望减少错误信息传播并提高疫苗接种率（Lin et al., 2024; Larsen et al., 2022）。如果信息中出现名人，其效果可能会进一步增强，因为名人能够提高关注度、参与度和可信度（Alatas et al., 2023）。然而，目前还没有大规模研究检验预防性信息传递能否减少网络仇恨，或产生更广泛的网络效应。

我们报告迄今规模最大的现场实验，检验减少网络仇恨的预防策略。在尼日利亚 X（前身为推特）平台开展的一项预注册随机对照试验中，我们向 73,136 名此前曾参与仇恨内容互动的用户推送了一段 42 秒的亲社会视频，视频中有尼日利亚名人出镜。鉴于社交媒体使用广泛且族群间紧张关系长期存在，尼日利亚是一个具有代表性的试验案例（Ezeibe，2021 年）。我们评估了该策略对仇恨内容生产和分享的影响、消息推送结束后的持续性，以及触达用户受众中的较大比例是否会在网络中产生间接效应。结果显示，针对特定用户推送名人消息可以减少网络仇恨并遏制其传播，而无需删除内容。

结果

我们分析了 2023年7月31日之前发布的超过 17 亿条公开帖子，并使用在尼日利亚文本上微调的语言模型检测仇恨内容（Tonneau 等，2024）。高分帖子通常表达了明显的敌意（“我对北方人的仇恨与日俱增”）或非人化（“比夫拉人是动物”）。我们确定了 73,136 名用户，他们在 2022年1月1日 至 2023年7月31日期间至少两次参与了针对尼日利亚群体的仇恨内容（见方法）。我们使用基于图的算法将这些用户划分为若干簇，该算法会将此期间频繁相互提及的账号归为一组（Ugander 等，2013）。随后，我们按相同比例将各簇随机分配至处理组和对照组，处理组用户有资格通过 X 广告平台接收预防性短视频（图 1；见方法）。在为期 140 天的处理期间，

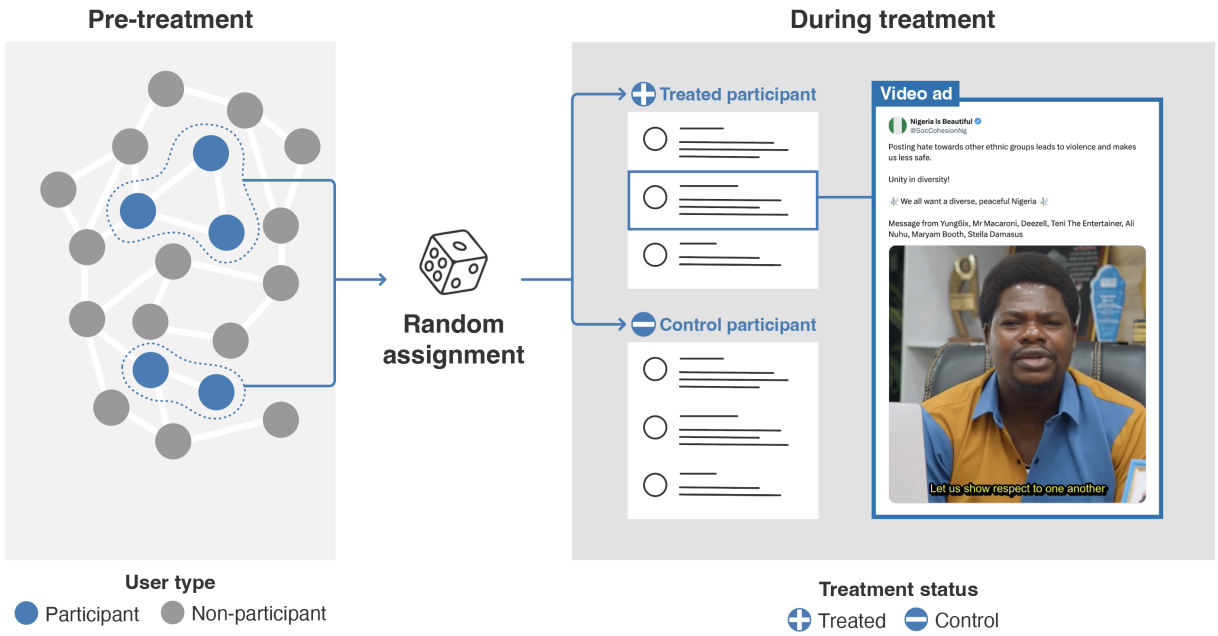


图1：**实验样本与干预实施**。左侧面板显示，73,136 个参与过仇恨言论活动的抽样 X 账号以蓝色表示，其他账号以灰色表示。用户根据处理前提及关系被分入不同簇，各簇随机分配至处理组或对照组。右侧面板显示，2023年11月29日至2024年4月17日期间，被分配至处理组的用户（加号）有资格通过 X 广告平台在信息流中直接收到尼日利亚名人的亲社会视频广告，而对照组用户（减号）则不会收到。

平台向 17,650 名用户投放了广告，占处理组的 45.8%。广告展示次数超过 90 万次，并带来了大量参与度，包括点赞和转发（见补充信息第 S1 节）。

我们首先评估这项宣传活动是否减少了仇恨内容的创作和分享。我们预先注册的主要结果指标是每位用户发布的仇恨帖子数量，分别统计所有帖子、原创帖子和转发，并涵盖三个时期：处理前（2023年1月1日至7月31日）、处理期间（2023年11月29日至2024年4月17日）和处理后（2024年5月1日至8月31日）。我们通过比较处理组用户和对照组用户来估计意向治疗（ITT）效应，并调整处理前水平，同时考虑集群随机设计（见方法）。宣传活动期间，处理组用户减少了仇恨内容的创作和分享（图2）。总体而言，仇恨帖子数量相较对照组下降了2.5%（$p = 0.032$），其中原创仇恨

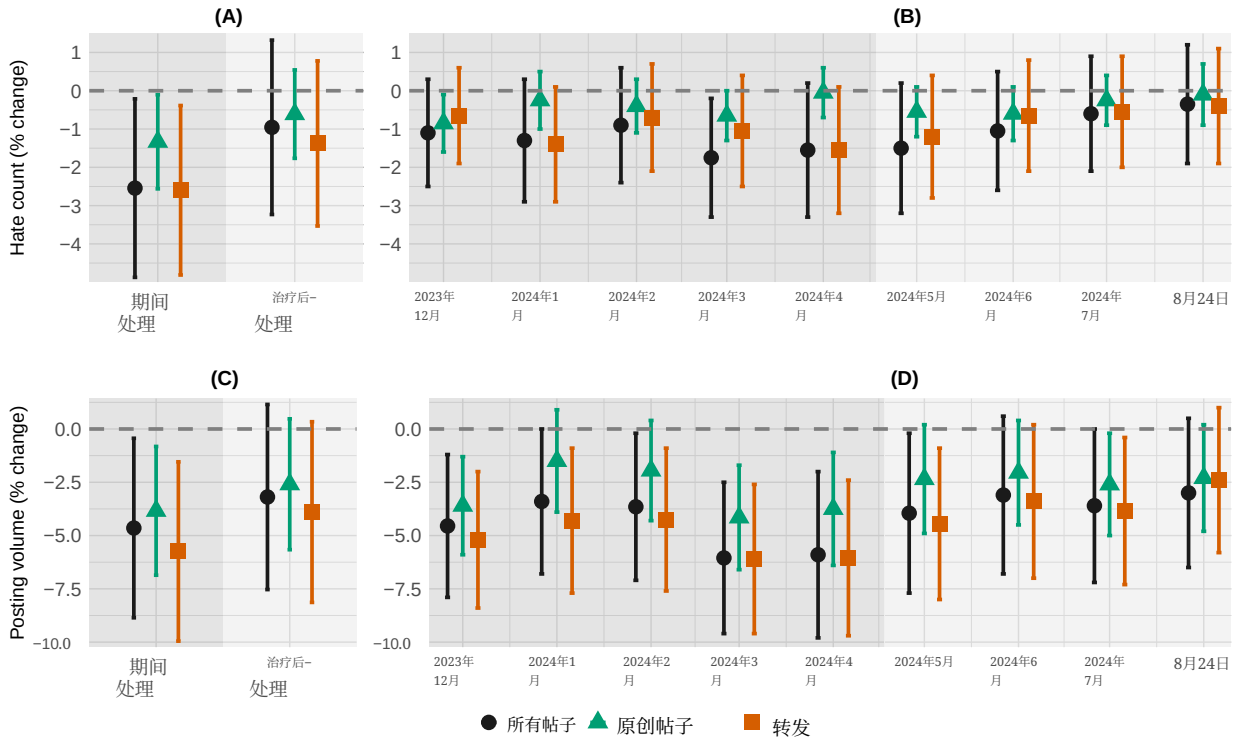


图2：**对仇恨帖子和发帖量的直接效应**。A–B面板报告了宣传活动干预期间及处理后时期（A）和按月份划分（B）的仇恨帖子的意向处理百分比变化。C–D面板显示了发帖量的百分比变化。结果分别按所有帖子（黑色）、原创帖子（绿色）和转发（红色）绘制。误差条表示95%置信区间。处理组用户在宣传活动期间减少了仇恨发帖和总体活动，其中很大一部分降幅在处理后仍持续存在。

帖子减少了1.3%（$p = 0.033$），仇恨内容转发减少了2.6%（$p = 0.021$；图 2A-B）。由于处理组用户中只有45.8%至少收到过一则广告，意向治疗效应低估了暴露的影响。我们将处理分配作为广告接收的工具变量，估计实际接受干预者效应为5.5%，这与既有证据一致：数字信息宣传活动通常会使行为改变1%至5%（Athey et al., 2023；Gordon et al., 2022；Lin et al., 2024）。值得注意的是，处理期间观察到的效果中，约75%（$p < 10^{-6}$）在处理后时期仍然持续（见补充信息第 S3 节）。结果对于其他仇恨定义和处理前调整方式均保持稳健（见补充信息第 S4 节）。

为进一步刻画处理效应，我们考察了未预注册的结果变量，首先是发帖量。宣传活动使治疗期间的发帖总量减少了4.6%（$p = 0.03$），

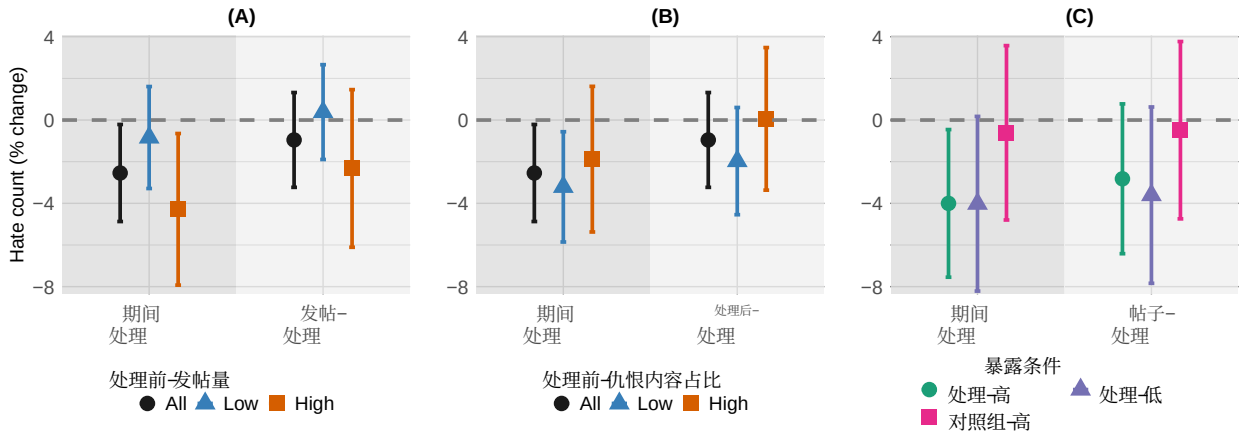


图 3：**仇恨帖子上的异质性处理效应**。处理效应因处理前发帖量 (A)、处理前仇恨内容占比 (B) 和暴露条件 (C) 而异。发帖量和仇恨内容占比的低、高组依据中位数分组确定。暴露综合了自身的分组情况与被分配到处理组的同伴占比（见方法）。误差线表示 95%置信区间。处理效应集中在更活跃、仇恨倾向较低的用户中，且没有证据表明参与者中存在间接效应。

处理后，这一效应有88%仍然持续（$p < 10^{-6}$），表明处理组用户的参与度更广泛地下降了（图2C-D）。活跃度的下降幅度超过了仇恨内容的减少幅度，但两者差值并不显著（$p > 0.32$）。我们还观察到处理组用户所放大传播的账号发生了变化。尽管处理组用户的转发总量减少了5.7%（$p = 0.007$），但他们将注意力转向了更具影响力的来源：他们转发的账号平均关注者多3.4%（$p = 0.045$），帖子也多2.8%（$p = 0.051$），与对照组用户转发的账号相比。这一模式表明，尽管转发总量有所下降，干预仍使注意力转向了更具影响力的账号（见补充信息第S5节）。

处理效应因处理前行为而异。对于处理前发帖量高于中位数的用户，治疗期间仇恨帖子减少了4.3%（$p = 0.021$）；而对于低于中位数的用户，处理效应不显著（$p = 0.50$；图3A）。处理效应也因处理前仇恨程度而异：对于处理前仇恨内容占比低于中位数的用户，仇恨帖子减少了3.2%（$p = 0.017$）；而对于高于中位数的用户，处理效应不显著（$p = 0.29$；图3B）。总体而言，该宣传活动对更活跃且处理前发布较少仇恨内容的用户更有效。使用发帖总量作为结果变量时，也出现了类似的异质性模式

（见补充信息第S6节）。

随后，我们检验参与者社交网络中的预注册间接效应。作为原始随机化设计的一部分，我们在每个簇内独立随机翻转一部分个体分配，从而产生了接受处理的同伴局部暴露程度的外生变异，并改变了接受处理的邻居所占比例（见方法）。由此形成四种暴露条件，依据用户自身的分配以及其邻居中被分配接受处理者所占的比例来界定：处理组—高暴露、处理组—低暴露、对照组—高暴露和对照组—低暴露。检测间接效应需要参与者之间持续互动，但他们转发的账号中仅有3.39%属于实验样本，而且被转发账号集合每月约有78%的变化，因此几乎没有足够的稳定性产生间接影响（见补充信息第S7.6节）。与这一限制相符，我们未发现对照组—高暴露与对照组—低暴露之间存在差值（$p = 0.77$），也未发现处理组—高暴露和处理组—低暴露相对于对照组—低暴露存在差值（$p = 0.99$；见图3C和补充信息第S7节）。

随后，我们考察了用户互动频繁的情境，在这种情境下，间接效应的影响范围更大。具体而言，我们检验了参与者选择转发的内容发生变化时，对因此受到暴露的非参与者产生的间接效应（图 4）。按照预注册方案，我们聚焦于 400 名上游用户，这些用户被定义为处理前其内容经常被参与者转发的非参与者（参见“方法”）。由于转发是主要的放大渠道，处理组用户减少转发可能降低上游用户的可见度，并限制仇恨内容的下游传播。我们以每位上游用户处理前受众中被分配至处理组的比例来量化间接暴露，并估计该比例每增加一个百分点，对处理前转发仇恨内容的用户数量和仇恨内容转发量减少的影响（图 5）。

间接暴露增加，使参与者中处理前转发仇恨内容的用户减少了 5.4%（$p = 0.013$），非参与者中减少了 4.5%（$p = 0.025$）；宣传活动结束后，降幅仍大致相同（图 5A-B；补充信息第 S1 节）。这些增幅主要集中在发帖量低于中位数的上游用户中，而发帖量较高的上游用户所受影响并不显著。

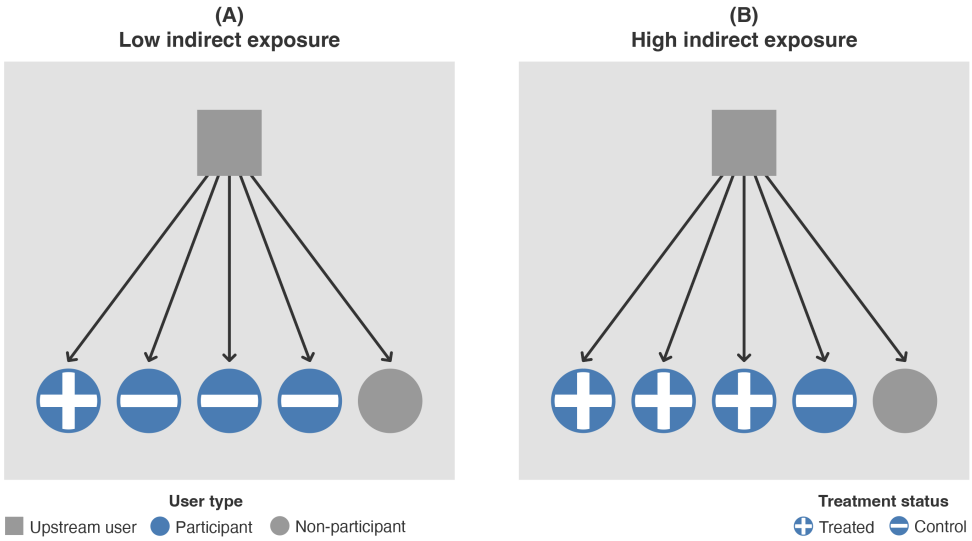


图4：**上游用户与间接暴露**。上游用户是指在处理前，其帖子经常被参与者转发的账号。(A) 某上游用户的转发者中，处理组参与者较少。(B) 某上游用户的转发者中，处理组参与者占比更高，意味着其受到干预的间接暴露程度更大。灰色方块代表上游用户；蓝色圆圈代表参与者；加号和减号分别表示处理状态和对照状态；灰色圆圈代表非参与者；箭头代表处理前转发链接。

对发帖量较高的上游用户而言，影响并不显著。仇恨内容转发也呈现出类似模式（图5C-D）：处理组受众占比每增加一个百分点，参与者的仇恨内容转发就减少 2.9%（$p = 0.019$），非参与者则减少 2.1%（$p = 0.075$）。将这一边际效应应用于观察到的暴露水平，意味着上游用户的仇恨内容转发平均减少约 53%。这些模式表明，处理组用户传播放大的减弱也传导至从未收到消息的非参与者，从而限制了处理组人群以外的下游传播。结果对其他设定具有稳健性（见补充信息 S10节）。

讨论

我们在此展示，尼日利亚名人发布的亲社会短视频大规模宣传活动，减少了 X 上仇恨内容的创作和分享。在族群间紧张关系长期存在的背景下，这项干预持续减少了仇恨帖子和转发

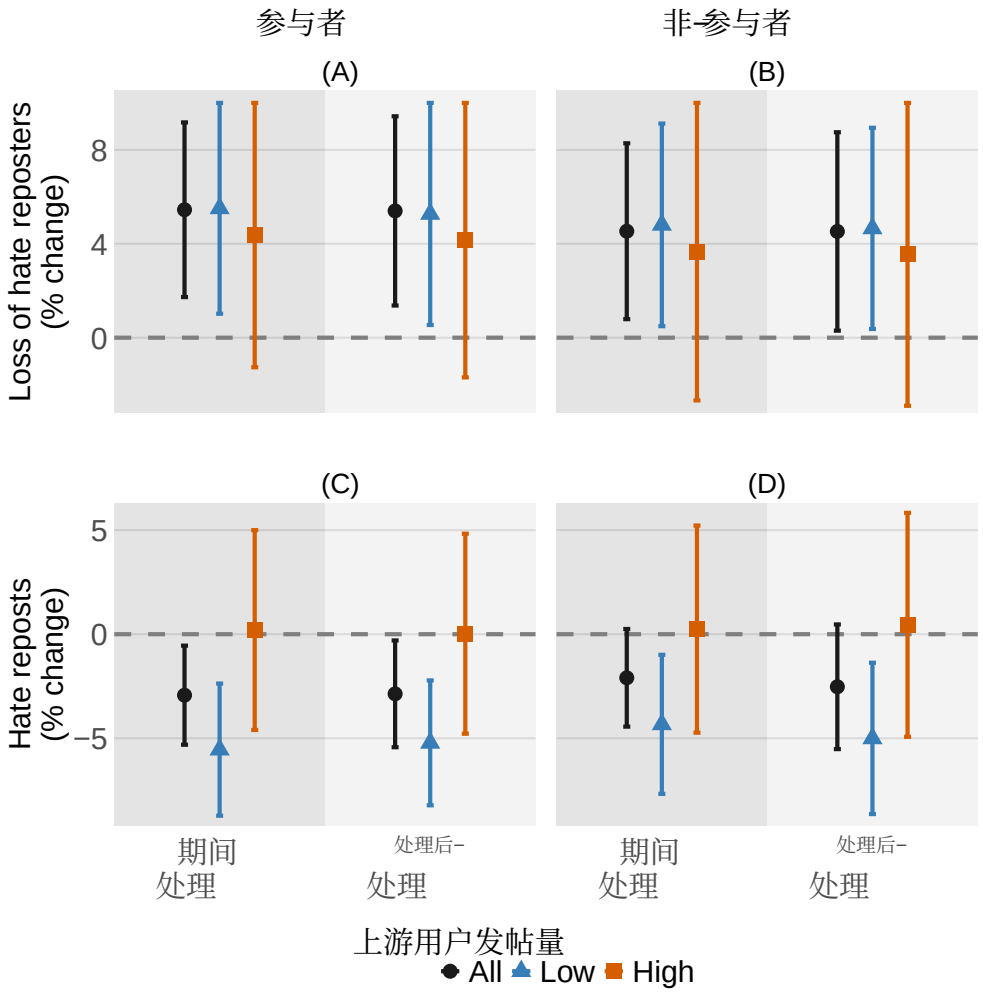


图5：**对上游用户的间接效应**。上游用户是指在处理前，其帖子经常被参与者转发的非参与者。各面板报告了上游用户处理前受众中处理组受众占比每增加一个百分点所对应的百分比变化。各行表示结果变量类型（处理前仇恨内容转发者流失；仇恨内容转发）。各列表示受众类型（参与者；非参与者）。颜色表示上游用户处理前的发帖量（全部、低、高）。处理组受众占比越高，仇恨内容转发者流失越多，仇恨内容转发越少，即使在非参与者中也是如此。误差条表示95%置信区间。

这种效果在宣传活动结束后仍然持续。其影响还扩展到了从未收到消息的用户，证明了主动且不审查的干预是可行的。

此前关于直接反仇恨言论的实地实验报告称，仇恨帖子有所减少，但这类方法需要实时检测，且难以大规模推广。相比之下，我们基于定向信息传递的预防策略能在用户发布有害内容之前触达他们。我们的效果幅度

与近期在政治、健康和错误信息情境下开展的数字信息传递实验报告的结果相当（Athey et al., 2023；Gordon et al., 2022；Lin et al., 2024）。这表明，主动信息传递能够有效遏制仇恨，同时也适用于大规模部署。

参与者随时间转发的账号之间重合度有限且不稳定，因此处理很少有机会通过本地连接扩散。这种模式可能反映了社交媒体平台在本地影响力方面的结构性局限：平台日益成为广播网络，提升大型账户的影响力，而非促进同伴间的反复互动（Chen, 2023）。尽管如此，当用户受众中有相当一部分接触到干预时，该用户仇恨内容的传播放大有所下降。这一发现对于仇恨内容产出量较高、且对直接处理反应较弱的用户尤其重要，因为降低其受众的参与度可以间接限制其触达范围（Gennaro et al., 2025a）。

我们在非参与者中观察到的间接效应，不太可能是转发链被简单中断所致。如果处理组用户打断了转发链，我们预计非参与者在上游用户转发受众中的相对排名会出现系统性变化，但我们并未观察到这种模式（见补充信息第 S11 节）。相反，证据与平台内部的排名响应相符：参与度降低会削弱较小上游用户的可见度。

这些结果凸显了主动信息传递策略作为内容审核补充手段的潜力。预防性宣传活动可以通过现有的广告基础设施，以较低的边际成本开展；相比之下，人工审核和执行系统则需要持续投入大量资源（见补充信息第 S12 节）。未来的研究应比较不同的信息传递者类型和信息框架（Munger, 2021；Hangartner et al., 2021；Siegel and Badaan, 2020），并探索自适应个性化或人工智能驱动的个性化（B¨ar, 2024；Costello et al., 2024）。随着平台缩减审核力度，以及算法系统继续放大有害内容（Corsi, 2024；Milli et al., 2025），预防性反仇恨言论提供了一种可扩展的非审查式方法，能够减少网络危害并促进更安全的数字环境。

材料与方法

预注册。 本研究在数据分析之前进行了预注册。参与者的实验设计、处理分配和主要结果已在 https://osf.io/q735x进行预注册。针对上游用户间接效应的分析计划已在 https://osf.io/rpygk进行预注册。

样本选择。 2021年1月1日至2022年12月31日期间，我们使用 X API 识别个人资料位置标注为尼日利亚的用户。我们递归地将这些用户在帖子中提及的所有尼日利亚账户纳入样本，并重复这一过程，直到不再发现新的用户。最终识别出超过 280 万名用户和逾 17 亿条帖子，涵盖了该平台在尼日利亚大量的活动。我们使用在尼日利亚社交媒体文本上微调的大语言模型 NaijaXLM-T 检测仇恨内容（Tonneau 等，2024）。该模型为每条帖子赋予 0 到 1 之间的连续仇恨得分，并对目标群体进行分类，包括族群、宗教和性别类别。潜在参与者是这些用户中的 X 账号：在 2022年1月1日至2023年7月31日期间，这些账号至少发布、转发或点赞过两条被判定为针对尼日利亚族群的仇恨内容的帖子。我们排除了无法访问的账户，以及少数社交关系高度密集的用户，以便估计间接效应；最终纳入 80,154 名实验参与者（见补充信息第 S1 节和 S1.8）。其中，7,018 个账号在处理前变为不活跃，或在处理期间及之后无法访问，因此分析样本包含 73,136 名用户。处理组与对照组之间的样本流失无显著差异（见补充信息第 S4.4 节）。

处理分配。如预注册所述，我们根据 2022年1月1日至2023年7月31日的用户提及构建了互动网络，并使用 3-网聚类算法将参与者划分为 4,253 个簇（Ugander 等，2013）。处理采用两阶段图簇随机化。首先，以相同概率将各簇独立分配至处理组或对照组。其次，独立地以 0.18 的概率翻转每位参与者的分配

，以改变簇内暴露程度（Eckles 等人，2017）。根据处理前活跃度，将参与者分为十个规模相等的组，并上传至 X 广告平台，以便我们考察不同活跃度下的广告投放模式。仅向处理组用户投放由尼日利亚名人出演的亲社会视频广告。为防止组间污染，广告在发生任何互动后都会自动删除并重新发布，确保互动不会使对照组用户看到广告（参见补充信息第 S1.7 节）。

结果测量指标。 根据预注册方案，主要结果是参与者在2023年1月1日至2024年8月31日期间发布的仇恨帖子。我们使用 NaijaXLM-T 测量仇恨内容，该模型为每条帖子分配一个介于 0 和 1 之间的连续仇恨得分。对于每位用户，我们统计仇恨分数高于 0.5 的帖子，并分别构建所有帖子、原创帖子和转发的测量指标。我们对结果进行对数转换，并在各时期内取平均值以进行估计。关于采用其他仇恨阈值和模型设定的稳健性检验，详见补充信息第 S4 节。

参与者直接效应的估计。 在假设参与者群体内部不存在间接效应的前提下，我们通过比较处理组与对照组参与者的结果来估计处理效应。平均处理效应采用差分估计量进行估计；该估计量利用重新缩放的处理前基线调整每位用户的结果，以提高精度，同时保持无偏性。我们通过根据分配程序所产生的联合处理概率，解析推导估计量的方差，从而在推断中纳入图聚类随机化的影响。形式化定义和方差推导见补充信息第 S2 节。

对参与者间接效应的估计。为了估计对参与者的间接效应，我们根据每位用户自身的处理分配，以及其干预前提及关系中指向处理组参与者的比例，将用户分配到四种暴露条件之一。暴露采用q比例规则$q = 0.7$来界定：当用户干预前的提及关系中至少70%指向

处理组用户时，将其归为高暴露；当至少70%指向对照组用户时，则归为低暴露；邻居构成比例介于这两个阈值之间的用户不归入任何一种暴露条件（Ugander 等，2013）。由此得到四种条件：处理组—高暴露、处理组—低暴露、对照组—高暴露和对照组—低暴露。我们通过蒙特卡洛模拟分配程序，获得图聚类随机化下的暴露倾向，并据此构造逆倾向权重。随后，采用与H´ajek估计量等价的加权回归，估计各暴露条件下的均值及其对比，并纳入处理前发帖量以提高精度。方差采用 Gao 和 Ding（2025）的网络稳健HAC估计量计算。完整的方法细节见补充信息第S7节。

上游用户选择。 我们确定了参与者在宣传活动前四个月（2023年8月1日至2023年11月29日）频繁转发的账号，并根据这些账号在处理前时期获得的转发中来自参与者的占比进行排名。我们排除了2023年宣传活动开始前发帖超过15,000条的用户。剩余账号中排名前400的账号被选为具有较高潜在暴露度的上游用户，这些账号获得的转发至少有37%来自参与者。对于这些用户，我们收集了2023年5月1日至2024年8月31日期间的所有原创帖子。

对上游用户间接效应的估计。我们在尖锐零假设下使用随机化推断来估计上游用户的间接效应，即其结果变量不受间接暴露变化的影响。对于每位上游用户i，暴露T_i 定义为其处理前转发者中被分配至处理组的比例。我们的检验统计量是回归中T_i 的系数：我们将每个结果变量对暴露进行回归，并控制处理前发帖量。随后，我们模拟 10,000 次重新打乱的分配，以复现原始实验设计，重新计算暴露，并重新估计相同的回归模型，从而生成尖锐零假设下检验统计量的分布。随机化推断p值将观测到的统计量与该分布进行比较，而边际效应的置信区间

则通过对该程序求逆获得，表示暴露增加一个百分点的效应。更多方法细节请参见补充信息第 S8 节。

致谢

本研究获得英国外交、联邦及发展事务部（FCDO）的资助，以及世界银行研究支持预算的支持。此外，纽约大学高性能计算提供了资源、服务和人员专业支持。本文所表达的研究结果、解释和结论均为作者观点，不一定代表国际复兴开发银行/世界银行及其附属组织、世界银行执行董事或其所代表政府的观点。S.F. 和 V.O. 是世界银行员工。研究期间，E.J.、B.K.、M.T.、H.L. 和 D.B. 曾担任世界银行顾问。E.J. 获得美国国家科学基金会（NSF）第 1745640 号资助。H.L. 获得加拿大社会科学与人文研究委员会三机构资助（第 192324 号）。本研究经健康媒体实验室机构审查委员会审查并批准（机构审查委员会编号：2310）。

参考文献

V. Alatas、A. G. Chandrasekhar、M. Mobius、B. A. Olken 和 C. Paladines。名人代言是否重要？在印度尼西亚开展的一项推广疫苗接种的推特实验。《经济学杂志》，134(659):913–933，2023年11月。ISSN 0013-0133。doi: 10.1093/ej/uead102。URL https://doi.org/10.1093/ej/uead102。J. Allen、D. J. Watts 和 D. G. Rand。量化错误信息和疫苗怀疑内容在 Facebook 上的影响。《科学》，384(6699):eadk3451，2024年。doi: 10.1126/science. adk3451。URL https://www.science.org/doi/abs/10.1126/science.adk3451。C. Arcila Calder´on、P. S´anchez Holgado、J. G´omez、M. Barbosa、H. Qi、A. Matilla、P. Amado、A. Guzm´an、D. L´opez-Mat´ıas 和 T. Fern´andez-Villazala。从网络仇恨言论到线下仇恨犯罪：煽动性语言在预测针对移民和 LGBT 群体的暴力行为中的作用。《人文与社会科学通讯》，11(1):1–14，2024年。P. M. Aronow 和 J. A. Middleton。随机实验中平均处理效应的一类无偏估计量。《因果推断杂志》，1(1):135–154，2013年。P. M. Aronow 和 C. Samii。在一般干扰情况下估计平均因果效应，并应用于一项社交网络实验。《应用统计年鉴》，11(4):1912–1947，2017年。S. Athey、K. Grabarz、M. Luca 和 N. Wernerfelt。大规模数字公共卫生干预：社交媒体广告对新冠疫苗相关信念和结果的影响。《美国国家科学院院刊》，120(5):e2208110120，2023年。C. A. Bail、L. P. Argyle、T. W. Brown、J. P. Bumpus、H. Chen、M. F. Hunzaker、J. Lee、M. Mann、F. Merhout 和 A. Volfovsky。在社交媒体上接触相反观点可能加剧政治极化。《美国国家科学院院刊》，115(37):9216–9221，2018年。D. B¨ar、A. Maarouf 和 S. Feuerriegel。生成式人工智能可能会适得其反，无法有效进行反仇恨言论。*arXiv* 预印本 *arXiv:2411.14986*，2024年。

R. M. Bond、C. J. Fariss、J. J. Jones、A. D. Kramer、C. Marlow、J. E. Settle 和 J. H. Fowler。一项涉及 6100 万人的社会影响与政治动员实验。《自然》，489(7415): 295–298，2012年。E. Chandrasekharan、U. Pavalanathan、A. Srinivasan、A. Glynn、J. Eisenstein 和 E. Gilbert。你不能留在这里：通过仇恨言论审视 Reddit 2015 年封禁措施的有效性。《*ACM*人机交互汇刊》，1(CSCW):1–22，2017年。B. X. Chen。社交媒体的未来。《纽约时报》，2023年4月。URL https://www.nytimes.com/2023/04/19/technology/personaltech/tiktok-twitter-facebook-social.html。访问日期：2025-11-22。G. Corsi。评估推特对低可信度内容的算法放大：一项观察性研究。《*EPJ*数据科学》，13(1):1–15，2024年。T. H. Costello、G. Pennycook 和 D. G. Rand。通过与人工智能对话，持久减少阴谋论信念。《科学》，385(6714):eadq1814，2024年。D. Eckles、B. Karrer 和 J. Ugander。网络实验的设计与分析：减少干扰造成的偏差。《因果推断杂志》，5(1)，2017年。R. Enikolopov、A. Makarin 和 M. Petrova。社交媒体与抗议参与：来自俄罗斯的证据。《计量经济学》，88(4):1479–1514，2020年。C. Ezeibe。尼日利亚的仇恨言论与选举暴力。《亚非研究杂志》，56 (4):919–935，2021年。E. J. Finkel、C. A. Bail、M. Cikara、P. H. Ditto、S. Iyengar、S. Klar、L. Mason、M. C. McGrath、B. Nyhan、D. G. Rand、L. J. Skitka、J. A. Tucker、J. J. Van Bavel、C. S. Wang 和 J. N. Druck-man。美国的政治宗派主义。《科学》，370(6516):533–536，2020年。M. Gao 和 P. Ding。网络实验中的因果推断：基于回归的分析与基于设计的性质，2025年。URL https://arxiv.org/abs/2309.07476。G. Gennaro、L. Bronner、L. Derksen、M. Kubli、A. Kotarcic、S. Kurer、P. Grech、K. Donnay、F. Gi-lardi 和 D. Hangartner。仇恨言论的分布及其对内容审核的影响。《政治科学研究与方法》，第 1–9 页，2025a。doi: 10.1017/psrm.2025.10063。G. Gennaro、L. Derksen、A. Abdelrahman、E. Broggini、M. A. Green、V. A. Haerter、E. Heer、I. Heidler、F. Kauer、H.-N. Kim 等。鼓励用户换位思考少数群体处境的反仇恨言论，能够减少社交媒体上的仇恨言论及其传播放大。《科学报告》，15(1):22018，2025b。N. Giansiracusa。Facebook 用欺骗性的数学手段掩盖其仇恨言论问题。《连线》，2021年。T. Gillespie。《互联网守护者：平台、内容审核以及塑造社交媒体的隐秘决策》。耶鲁大学出版社，2018年。I. Goldstein、L. Edelson、M.-K. Nguyen、O. Goga、D. McCoy 和 T. Lauinger。理解内容审核的（无）效性：以美国国会大厦骚乱背景下的 Facebook 为例。*arXiv* 预印本 *arXiv:2301.02737*，2023年。B. R. Gordon、R. Moakler 和 F. Zettelmeyer。够接近吗？对广告衡量非实验方法的大规模探索。《营销科学》，42(4):768–793，2022年。A. M. Guess、B. Nyhan 和 J. Reifler。在2016年美国大选期间接触不可信网站。《自然·人类行为》，4(5):472–480，2020年。D. Hangartner、G. Gennaro、S. Alasiri、N. Bahrich、A. Bornhoft、J. Boucher、B. B. Demirci、L. Derksen、A. Hall、M. Jochum 等。基于共情的反仇恨言论可在一项社交媒体现场实验中减少种族主义仇恨言论。《美国国家科学院院刊》，118(50):e2116310118，2021年。

M. Horta Ribeiro、J. Cheng 和 R. West。自动化内容审核能提高用户遵守社区准则的程度。载于*2023*年*ACM*万维网大会论文集，第2666–2676页，2023年。J. Kaplan。更多言论，少些错误。https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/，2025年1月。Meta新闻室，最初发表于2025年1月7日。B. Kolic、F. Aguirre-L´opez、S. Hern´andez-Williams 和 G. Garduˇno-Hern´andez。从信息茧房到回音室：运用二阶邻居方法量化推特气候讨论中的极化现象。《复杂网络期刊》，13(4)：cnaf020，2025年。N. Kumarswamy、M. Singhal 和 S. Nilizadeh。更严格的内容审核对 Parler 用户话语的影响。*arXiv*预印本 *arXiv:2310.08844*，2023年。B. J. Larsen、T. J. Ryan、S. Greene、M. J. Hetherington、R. Maxwell 和 S. Tadelis。借唐纳德·特朗普为新冠疫苗背书之机，为公共卫生注入强心针：一项大规模广告实验。*arXiv*预印本 *arXiv:2203.02625*，2022年。C. Lee、K. Gligori´c、P. R. Kalluri、M. Harrington、E. Durmus、K. L. Sanchez、N. San、D. Tse、X. Zhao、M. G. Hamedani 等。分享遭遇种族主义经历的人在网上受到人类和机器的噤声，但准则重构干预展现出希望。《美国国家科学院院刊》，121(38)：e2322764121，2024年。H. Lin、H. Garro、N. Wernerfelt、J. Shore、A. Hughes、D. Deisenroth、N. Barr、A. Berinsky、D. Eck-les、G. Pennycook 等。利用数字准确性提示广告大规模减少错误信息传播。2024年。H. Lin、G. Czarnek、B. Lewis、J. P. White、A. J. Berinsky、T. Costello、G. Pennycook 和 D. G. Rand。利用人类与人工智能对话说服选民。《自然》，第1–8页，2025年。S. Milli、M. Carroll、Y. Wang、S. Pandey、S. Zhao 和 A. D. Dragan。社交媒体上的参与度、用户满意度与煽动分裂内容的传播放大。《*PNAS Nexus*》，4(3)：pgaf062，2025年3月。ISSN 2752-6542。doi：10.1093/pnasnexus/pgaf062。URL https://doi.org/10.1093/pnasnexus/pgaf062。K. M¨uller 和 C. Schwarz。煽风点火：社交媒体与仇恨犯罪。《欧洲经济学会杂志》，19(4)：2131–2167，2021年。K. Munger。推文干预对发推者的影响：通过实验减少种族主义骚扰。《政治行为》，39：629–649，2017年。K. Munger。别@我：通过实验减少推特上的党派间失礼行为。《实验政治学期刊》，8(2)：102–116，2021年。J. Pfeffer、D. Matter 和 A. Sargsyan。推文半衰期。载于国际*AAAI*网络与社交媒体会议论文集，第17卷，第1163–1167页，2023年。P. R. Rosenbaum。随机实验中单位间的干扰。《美国统计学会期刊》，102(477)：191–200，2007年。A. A. Siegel 和 V. Badaan。#no2sectarianism：减少网上宗派仇恨言论的实验方法。《美国政治学评论》，114(3)：837–855，2020年。M. Tonneau、P. Quinta De Castro、K. Lasri、I. Farouq、L. Subramanian、V. Orozco-Olvera 和 S. Fraiberger。NaijaHate：使用代表性数据评估尼日利亚推特上的仇恨言论检测。载于 L.-W. Ku、A. Martins 和 V. Srikumar 编，计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），第9020–9040页，泰国曼谷，2024年8月。计算语言学协会。doi：10.18653/v1/2024.acl-long.488。URL https://aclanthology.org/2024.acl-long.488/。

M. Tonneau、D. Liu、N. Malhotra、S. A. Hale、S. Fraiberger、V. Orozco-Olvera 和 P. R¨ottger。HateDay：来自全球仇恨言论数据集的洞见，该数据集代表推特上一天的情况。载于 W. Che、J. Nabende、E. Shutova 和 M. T. Pilehvar 编，计算语言学协会第*63*届年会论文集（第*1*卷：长篇论文），第2297–2321页，奥地利维也纳，2025年7月。计算语言学协会。ISBN 979-8-89176-251-0。doi：10.18653/v1/2025.acl-long.115。URL https://aclanthology.org/2025.acl-long.115/。M. Tonneau、D. Liu、R. McGrady、K. Zheng、R. Schroeder、E. Zuckerman 和 S. Hale。社交媒体平台内容审核人员配置中的语言差异，2025年。URL https://doi.org/10.31235/osf.io/amfws_v1。Z. Tufekci。《推特与催泪瓦斯：网络化抗议的力量与脆弱性》。耶鲁大学出版社，2017年。J. Ugander、B. Karrer、L. Backstrom 和 J. Kleinberg。图聚类随机化：网络对多个宇宙的接触。载于第*19*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第329–337页，2013年。联合国教科文组织。《网络虚假信息与仇恨言论影响调查》。技术报告，联合国教科文组织/益普索，2023年9月。URL https://www.ipsos.com/sites/default/files/ct/news/documents/2023-11/unesco-ipsos-online-disinformation-hate-speech.pdf。调查于2023年8月至9月在16个国家开展，受访者为具有代表性的互联网使用人群，年龄为 18+。S. Vosoughi、D. Roy 和 S. Aral。网络上真假新闻的传播。《科学》，359(6380)：1146–1151，2018年。J. Whitten-Woodring、M. S. Kleinberg、A. Thawnghmung 和 M. T. Thitsar。若不懂如何使用，Facebook便如毒药：缅甸的Facebook、民主与人权。《国际新闻与政治期刊》，25(3)：407–425，2020年。doi：10.1177/1940161220919666。URL https://doi.org/10.1177/1940161220919666。M. M. Yildirim、J. Nagler、R. Bonneau 和 J. A. Tucker。比停用更进一步：停用警告如何减少推特上的仇恨言论。《政治学展望》，21(2)：651–663，2023年。

补充信息

目录

S1 实验设计 **19**
S1.1 仇恨言论检测.19 S1.2 参与者选择.19 S1.3 参与者之间的互动网络.19 S1.4 网络聚类.19 S1.5 集群随机化.20 S1.6 亲社会信息干预.20 S1.7 干预交付.20 S1.8 描述性统计.21

S2 直接效应 估计 22
S2.1 概念方法.22 S2.2 倾向概率.23 S2.3 点估计.24 S2.4 图聚类随机化下的推断.25 S2.5 保守方差估计.27 S2.6 随机化推断.29

S3 直接效应的持续性 29
S3.1 概念方法.30 S3.2 结果.30

S4 直接效应的稳健性 31
S4.1 替代处理前时期.31 S4.2 仇恨的替代测量.32 S4.3 负向对照检验.34 S4.4 参与者流失.35

S5 转发用户构成 **37**

S6 直接效应的异质性 **38**
S6.1 按处理前活动划分的异质性.38 S6.2 按处理前仇恨内容占比划分的异质性.39

S7 参与者之间的间接效应 39
S7.1 估计目标与暴露映射.40 S7.2 加权q比例暴露模型.40 S7.3 暴露对比.41 S7.4 倾向得分.41

S7.5 通过加权回归进行估计.42 S7.6 被转发用户的更替.43

S8 对上游用户的间接影响 **44**
S8.1 上游用户选择.44 S8.2 暴露定义.44 S8.3 随机化推断的检验统计量.45 S8.4 因果估计对象.45 S8.5 随机化推断.45

S9 间接效应的持续性 **46**
S9.1 概念方法.46 S9.2 结果.49

S10 间接效应的稳健性 **49**
S10.1 替代仇恨度量.49 S10.2 替代样本选择.51 S10.3 替代暴露定义.51 S10.4 替代处理前时期.53 S10.5 负向对照检验.54

S11 检验转发链中断 **55**

S12 成本效益 **57**

S1 实验设计

S1.1 仇恨言论检测

仇恨内容通过 NaijaXLM-T 识别。这是一种在尼日利亚社交媒体文本上训练的大语言模型，相关介绍见 Tonneau 等（2024）。该模型为每条帖子赋予 0 到 1 之间的连续仇恨得分，并可在 `https://huggingface.co/worldbank/naija-xlm-twitter-base-hate`公开获取。尽管该分类器能够检测多种形式的仇恨言论，我们的干预聚焦于族群仇恨。为识别目标群体，我们进一步对同一模型进行微调，以预测帖子是否针对 Tonneau 等（2024）所定义的八个尼日利亚社群之一。在2022年1月1日至2023年7月31日的检测期间，用户若发布、转发或点赞至少两条仇恨分数高于 0.9 且族群目标得分高于 0.5 的帖子，即符合纳入实验的条件。对随机抽取的 50 名符合条件用户进行人工审核后发现，约 80% 的用户明显参与过族群仇恨行为。

S1.2 参与者选择

将这些资格标准应用于完整数据集后，筛选出 95,140 名在2022年1月1日至2023年7月31日期间参与种族仇恨活动的用户。我们排除了 5,346 个无法访问的账户，并移除了同期发出提及数量排名前 2.5% 的用户，以及收到提及数量排名前 2.5% 的用户，以便估计间接效应；此举剔除了 3,956 名活跃度极高的用户。在其余用户中，80,154 名在处理前时期与另一名符合资格的用户至少存在一条转发或提及关系，构成实验总体；另有 5,684 名孤立用户被排除。在这些参与者中，2,369 名在2023年8月1日至11月29日处理开始前变为不活跃；此外，还有 4,649 名在随访期间停用账号、被暂停使用或将账号设为私密，最终分析样本为 73,136 名参与者。

S1.3 参与者之间的互动网络

我们使用2022年1月1日至2023年7月31日期间所有带有“@”符号的参与者互动，构建了一个加权有向网络。这些互动包括提及和转发；为简便起见，我们将它们统称为提及。每当用户a提及用户b时，我们就在两者之间添加一条从a指向b的有向边，边权重为提及总次数。为减少不频繁联系带来的噪声，我们仅保留每个用户权重最高的十条出边和十条入边。最终得到的剪枝网络包含521,961条边，保留了主要的影响渠道，并作为聚类和处理分配的基础。

S1.4 网络聚类

由于用户可能受到其他用户处理状态的影响，我们在簇层面而非个体层面随机分配处理，以控制干扰（Ugander 等，2013）。我们根据处理前的提及网络构建簇，假设间接效应主要通过直接关系发挥作用。遵循 Ugander 等（2013）提出的 3-net 聚类程序，我们迭代选择簇质心，移除每个质心两跳网络范围内的全部用户，并将剩余用户分配给最近的质心。我们唯一的

与原始程序的不同之处在于平局的处理：当用户到多个质心的距离相等时，我们依据互动强度进行分配，使联系紧密的用户归入同一簇。这种方法反映了在线互动的重尾结构，并提高了估计间接效应时的精确度。

S1.5 集群随机化

处理分配遵循基于图聚类随机化的两阶段设计（Ugan- der 等人，2013；Eckles 等人，2017）。第一阶段中，各个簇以相同概率独立分配至处理组或对照组，因此同一簇内的所有用户获得相同的分配。仅采用簇级随机化无法分别识别直接效应和间接效应，因为部分用户及其所有邻居都会获得相同的分配。为确保所有用户处于任一暴露条件下的概率均非零，我们以 0.18 的概率独立翻转每位用户的分配，这一做法通常称为打孔（Eckles 等人，2017）。第二阶段确保部分用户周围有许多接受处理的邻居，而另一些用户周围的此类邻居较少，从而能够在局部干扰假设下估计直接效应和间接效应。各簇之间的分配相互独立，由此产生的集群内处理相关性约为 0.41，为估计间接效应提供了足够的变异性和统计功效。

S1.6 亲社会信息传播干预

处理内容是一段简短视频，视频中，尼日利亚名人传达了一则亲社会信息。这些名人来自不同的族群和性别背景：Teni，女性约鲁巴族歌手兼词曲创作者；Maryam Booth，女性豪萨族演员，活跃于卡尼伍德和诺莱坞；Mr Macaroni，男性约鲁巴族喜剧演员兼内容创作者；Stella Damasus，女性伊博族演员兼制片人，活跃于诺莱坞；Deezel，男性豪萨族嘻哈艺人兼电影制片人；Ali Nuhu，男性豪萨族演员，活跃于卡尼伍德和诺莱坞，同时也是卡尼伍德的制片人和导演，以及尼日利亚电影合作公司的董事总经理；Yung6ix，埃多人男性说唱歌手兼嘻哈艺人。他们传达的信息如下：

“尼日利亚的同胞们，大家好。我是*Teni*、*Maryam Booth*、*Mr Macaroni*、*Stella Damasus*、*Deezel*、*Ali Nuhu*、*Yung6ix*，也就是飞船上的*Swaggalomo*。请让我们停止发布针对其他族群的仇恨言论。发布针对其他族群的仇恨言论是不对的。发布针对其他族群的仇恨言论会助长暴力，让我们所有人都更加不安全。让我们彼此尊重。请互相尊重。我们都希望尼日利亚和平，为了你、我和我们的孩子。和平与爱。”

该视频以推广帖文的形式出现在处理组参与者的信息流中，如图 S1 所示。视频共有七个版本，传达的信息相同，但名人出场顺序不同。视频的一个版本可通过 `https://drive.google.com/file/d/1AFNOHCL0A6Xfljs2hCL_0Oo3KWeesCLT/view?usp=sharing` 查看。

S1.7 干预交付

所有处理组用户均有资格通过 X 广告平台接收视频消息，但能否投放取决于平台算法。共有 17,650 名用户，占处理组的 45.8%

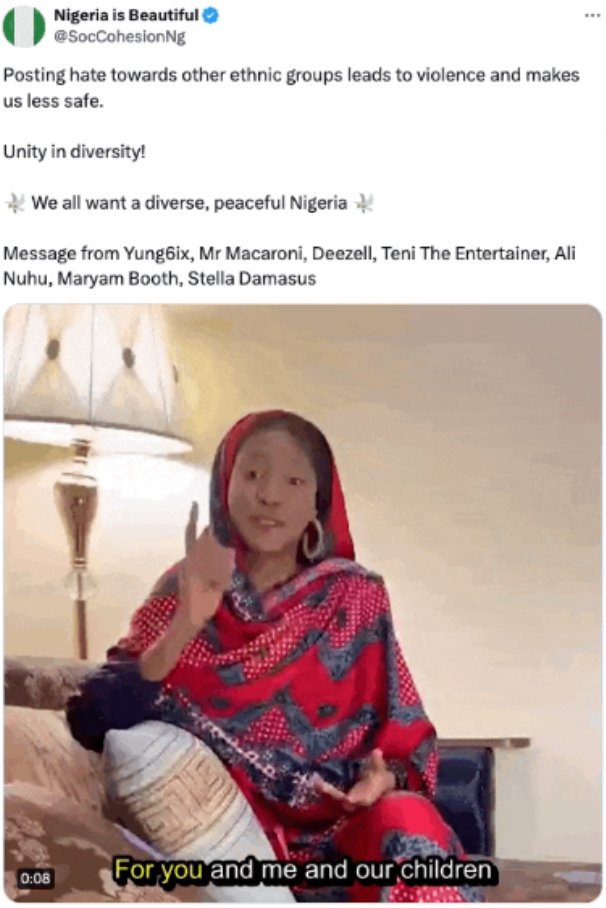


图 S1：出现在处理组用户信息流中的广告截图

用户至少看到了一条广告。在所有投放的广告中，此次宣传活动共产生 916,278 次曝光和 44,872 次互动，包括 17,795 次链接点击、 5,679 个点赞和 1,313 次转发。与对照组相比，处理组用户也更常在提及相关名人的同时使用广告相关词语。与“仇恨”和“暴力”等关键词同时出现的提及增加了 0.08%（$p = 0.00047$），此类提及在帖子总数中的占比也显著上升（$p = 0.03$），这表明广告吸引了用户关注。为防止通过社交互动造成非预期暴露，用户点赞或转发广告后，该广告会立即从其信息流中撤下，之后也只可能再次向该用户显示。这一方案不会影响其他用户的处理分配或资格。由于广告在每次互动后都会撤下，这些数据可能低估了标准投放条件下的互动情况。

S1.8 描述性统计

处理前时期的基线活动、仇恨内容和网络互动见表 S1。参与者的发帖行为差异很大，大多数活动是转发，而非原创帖子。仇恨帖子（仇恨分数 > 0.5）平均只占总体活动的一小部分，但其分布高度偏斜。与其他参与者的互动仅占总互动的一小部分，表明受众之间的直接

表 S1：处理前时期 73,136 位参与者的描述性统计。统计数据按用户计算。仇恨帖子定义为仇恨分数大于 0.5 的帖子。分享以百分比表示。

变量	Mean	SD	中位数 四分位距（25–75%）	
月度帖子				
帖子总数	417	910	133	[36, 412]
原创帖子	100	198	32	[6.6, 108]
转发	318	853	52	[7.6, 253]
月度仇恨内容（分数 >0.5）				
仇恨帖子	3.9	11	0.86	[0.14, 3.29]
原创仇恨帖子	0.88	3.6	0.14	[0.00, 0.57]
仇恨内容转发	3.10	9.8	0.43	[0.00, 2.14]
仇恨内容分享率（%）	1.5	3.1	0.5	[0.1, 1.4]
网络互动				
关注者	873	2855	273	[87, 826]
参与者转发率（%）	4.0	7.9	2.2	[0.5, 4.7]

重叠有限。处理组和对照组之间的协变量平衡情况见表 S2。全部变量的均值差异都很小，且均无统计学显著性（全部 p-值> 0.2），表明随机化使两组在基线时具有可比性。

S2 直接效应估计

S2.1 概念方法

本节介绍用于量化干预对参与者直接效应的估计目标、估计量和推断程序。我们假设不存在干扰，因此每位用户的结果变量仅取决于其自身的处理状态。在这一假设下，可采用标准的个体层面估计量。令 $Y'_i(t)$ 表示用户 i 在处理状态 $t \in \{0,1\}$ 下的结果变量，并按月份或研究期间进行汇总。由于大多数结果变量是计数，我们分析对数变换结果：

$$Y_i(t) = \log\big(Y'_i(t) + 1\big).$$

我们的主要估计目标是平均处理效应（ATE）：

$$\tau \;=\; \frac{1}{N}\sum_{i=1}^{N}\big(Y_i(1) - Y_i(0)\big).$$

我们在图2中报告了多个对数尺度结果变量的平均处理效应估计值。由于这些估计使用Horvitz–Thompson逆倾向加权估计量，我们首先描述处理

表 S2：处理组和对照组之间的处理前协变量平衡。除仇恨内容占比和参与者转发占比以百分比表示外，所有均值和差值均以对数尺度表示。双侧p值检验均值相等，并考虑了随机化设计。仇恨帖子定义为仇恨分数大于 0.5 的帖子。分享以百分比表示。

变量	均值（处理组）	均值（对照组）	差值	p值
月度帖子				
帖子总数	4.73	4.76	-0.00366	0.816
原创帖子	3.34	3.35	-0.00111	0.916
转发	3.83	3.86	-0.00289	0.828
月度仇恨内容（分数 >0.5）				
仇恨帖子	0.940	0.921	0.00190	0.613
原创仇恨帖子	0.348	0.341	0.000696	0.566
仇恨内容转发	0.728	0.710	0.00175	0.591
仇恨内容分享率（%）	1.76	1.69	0.0686	0.472
网络互动				
关注者	5.53	5.57	-0.0391	0.823
参与者转发率（%）	4.16	4.21	-0.0446	0.757

随机化设计所隐含的倾向概率。

S2.2 倾向概率

令$p_i(t) = \Pr(Z_i = t)$表示单位i 接受处理状态$t \in \{0,1\}$的概率，并令$p_{ij}(t_1,t_2)$表示单位i和j分别接受处理状态t_1和t_2的联合概率。后文推导的方差表达式需要用到联合倾向概率。在 S1.5节所述的带打孔的图聚类随机化中，所有单位的边际倾向概率相同，且不同簇之间的分配相互独立。

命题1 根据*S1.5*节所述的设计，每个单位的边际倾向概率为

$$p_i(1) = p_t + p_{hp} - 2p_t p_{hp}, \qquad p_i(0) = 1 - p_t - p_{hp} + 2p_t p_{hp},$$

其中，p_t 表示一个簇被分配到处理的概率，p^{hp} 表示个体分配被翻转的概率。对于不同的单位 $i \neq j$，令 $c(i)$ 表示单位 i 所属的簇。联合倾向概率为：

$$p_{ij}(t_1,t_2)=\begin{cases} p_i(t_1)p_j(t_2), & c(i)\neq c(j),\\ p_t+p_{hp}^2-2p_tp_{hp}, & c(i)=c(j),\ t_1=t_2=1,\\ 1+2p_tp_{hp}+p_{hp}^2-2p_{hp}-p_t, & c(i)=c(j),\ t_1=t_2=0,\\ p_{hp}(1-p_{hp}), & c(i)=c(j),\ t_1\neq t_2. \end{cases}$$

对于 $i=j$，我们可以得到边际倾向概率：$p_{ii}(t,t)=p_i(t)$，以及 $p_{ii}(t_1,t_2)=0$，当 $t_1\neq t_2$ 时。

在我们的实验中，簇级处理概率为$p_t=0.5$，打孔概率为$p_{hp}=0.18$。代入这些值可得 $p_i(1)=p_i(0)=0.5$，对于同一簇中的不同单位：

$$p_{ij}(1,1)=p_{ij}(0,0)=0.352,\qquad p_{ij}(1,0)=p_{ij}(0,1)=0.148.$$

这些倾向概率用于逆概率加权以及后续的方差计算。

S2.3 点估计

在无干扰假设以及 S1.5节所述的随机化下，处理状态t 下平均潜在结果的 Horvitz–Thompson（HT）估计量为

$$\widehat{\mu}(t)=\frac{1}{N}\sum_{i:Z_i=t}\frac{Y_i^{\text{during}}}{p_i(t)}.$$

由于用户活动存在高度异质性，我们使用差分估计量提高精度（Aronow and Middleton, 2013）。令Ypre $_i$表示处理前结果，并乘以常数 α重新缩放，使得 αY^{pre}的均值与Y期间的均值相匹配。定义调整后的结果变量：

$$\Delta_i^{\text{during}}=Y_i^{\text{during}}-\alpha Y_i^{\text{pre}}.$$

ATE 的 HT差值估计量为

$$\widehat{\tau}_{\text{diff}}=\widehat{\mu}_{\text{diff}}(1)-\widehat{\mu}_{\text{diff}}(0), \tag{1}$$

其中

$$\widehat{\mu}_{\text{diff}}(t)=\frac{1}{N}\sum_{i:Z_i=t}\frac{\Delta_i^{\text{during}}}{p_i(t)}.$$

由于处理前时期结果与处理相互独立，该估计量在保持无偏的同时，实现了显著更低的方差（Aronow and Middleton，2013）。正文和补充信息中报告的所有直接效应估计值均基于 $\widehat{\tau}_{\text{diff}}$。

S2.4 图聚类随机化下的推断

有效推断需要考虑簇内分配的相关性。 $\widehat{\tau}_{\text{diff}}$ 的方差由图聚类随机化下 Horvitz–Thompson均值估计量的方差和协方差得出，所用的边际倾向概率和联合倾向概率$p(t)$及$p(t_1,t_2)$如上所推导。为简洁起见，所有量均指差分结果。我们有以下定理，用以刻画 $\widehat{\mu}(t)$ 的方差和协方差。

定理1（图聚类随机化下的方差和协方差）采用打孔的图聚类随机化时，*Horvitz–Thompson* 估计量 $\widehat{\mu}(t)$的方差和协方差为

$$\operatorname{Var}\left[\widehat{\mu}(t)\right] = \frac{1}{N^2}\left[\left(\frac{p(t,t)}{p(t)^2} - 1\right)\sum_{c\in\mathscr{C}} S_c(t,t) + \left(\frac{p(t) - p(t,t)}{p(t)^2}\right) Q(t,t)\right] \tag{2}$$

对于 $t \in \{0,1\}$，以及对于 $t_1 \neq t_2$

$$\operatorname{Cov}\left[\widehat{\mu}(t_1),\widehat{\mu}(t_2)\right] = \frac{1}{N^2}\left[\left(\frac{p(t_1,t_2)}{p(t_1)p(t_2)} - 1\right)\sum_{c\in\mathscr{C}} S_c(t_1,t_2) - \frac{p(t_1,t_2)}{p(t_1)p(t_2)}\, Q(t_1,t_2)\right] \tag{3}$$

其中，$\mathscr{C}$表示簇的集合，且

$$S_c(t_1,t_2) = \sum_{i\in c}\sum_{j\in c} Y_i(t_1)\, Y_j(t_2)$$

$$Q(t_1,t_2) = \sum_{i=1}^{N} Y_i(t_1)\, Y_i(t_2)$$

证明。我们依据 Aronow and Samii（2017）关于 Horvitz–Thompson 估计量的一般方差和协方差公式。将记号简化后， $\widehat{\mu}(t)$的方差可写为（Aronow and Samii，2017）

$$\operatorname{Var}\left[\widehat{\mu}(t)\right] = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{p_{ij}(t,t)}{p_i(t)p_j(t)} - 1\right) Y_i(t)\, Y_j(t) \tag{4}$$

其中，$p_{ij}(t,t) = \Pr(Z_i = t, Z_j = t)$ 是单位 i和j 的联合倾向概率。在打孔设计下，所有单位具有相同的边际倾向概率 $p_i(t) = p(t)$；同一簇（不同簇）中的所有配对 $i \neq j$ 具有相同的联合倾向概率 $p_{ij}(t,t)$。我们利用这一点，将式（4）中的双重求和拆分为簇内配对和对角分量。首先，注意对于属于

不同 簇的任意配对（i_j），$p_ij(t,t)$ p(t)，因此括号中的因子为

$$\frac{p_{ij}(t,t)}{p_i(t)p_j(t)} - 1 = \frac{p(t)^2}{p(t)^2} - 1 = 0,$$

这些项会相互抵消。因此，只有同一簇内的配对会产生贡献。对于同一簇c中的$i \neq j$，有$p_{ij}(t,t) = p(t,t)$，所以

$$\frac{p_{ij}(t,t)}{p_i(t)p_j(t)} - 1 = \frac{p(t,t)}{p(t)^2} - 1$$

对于这类配对，该值为常数。对于$i = j$，有$p_{ii}(t,t) = p(t)$，因此

$$\frac{p_{ii}(t,t)}{p_i(t)p_i(t)} - 1 = \frac{p(t)}{p(t)^2} - 1 = \frac{1}{p(t)} - 1$$

利用这些观察结果，我们将式（4）改写为

$$\mathrm{Var}\left[\widehat{\mu}(t)\right] = \frac{1}{N^2}\left[\left(\frac{p(t,t)}{p(t)^2} - 1\right)\sum_{c\in\mathscr{C}}\sum_{\substack{i\in c\\ j\in c\\ i\neq j}} Y_i(t)\,Y_j(t) + \left(\frac{1}{p(t)} - 1\right)\sum_{i=1}^{N}\left(Y_i(t)\right)^2\right]$$

根据定义，簇内非对角和为

$$\sum_{\substack{c\in\mathscr{C}, i\in c, j\in c\\ i\neq j}} Y_i(t_1)\,Y_j(t_2) = \sum_{c\in\mathscr{C}} S_c(t_1,t_2) - Q(t_1,t_2) \tag{5}$$

代回后得到

$$\begin{aligned}\mathrm{Var}\left[\widehat{\mu}(t)\right] &= \frac{1}{N^2}\left[\left(\frac{p(t,t)}{p(t)^2} - 1\right)\left(\sum_{c\in\mathscr{C}} S_c(t,t) - Q(t,t)\right) + \left(\frac{1}{p(t)} - 1\right)Q(t,t)\right]\\ &= \frac{1}{N^2}\left[\left(\frac{p(t,t)}{p(t)^2} - 1\right)\sum_{c\in\mathscr{C}} S_c(t,t) + \left(\frac{p(t) - p(t,t)}{p(t)^2}\right)Q(t,t)\right]\end{aligned}$$

由此得到（2）。对于 $\widehat{\mu}(t_1)$ 和 $\widehat{\mu}(t_2)$ 在 $t_1 \neq t_2$, 之间的协方差，Aronow and Samii（2017）给出

$$\mathrm{Cov}\left[\widehat{\mu}(t_1),\widehat{\mu}(t_2)\right] = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{p_{ij}(t_1,t_2)}{p_i(t_1)p_j(t_2)} - 1\right)Y_i(t_1)\,Y_j(t_2) \tag{6}$$

在我们的设计下，$p_{ii}(t_1,t_2) = 0$ 对于 $t_1 \neq t_2,$，因此对角项变为

$$\frac{p_{ii}(t_1,t_2)}{p_i(t_1)p_i(t_2)} - 1 = -1$$

对于同一簇中的 $i \neq j$，我们有一个共同值 $p_{ij}(t_1,t_2) = p(t_1,t_2)$。对于不同簇中的 $i \neq j$，$p_{ij}(t_1,t_2) = p(t_1)p(t_2)$，因此这些项和之前一样消去。将求和拆分为簇内分量和对角分量，可得

$$\mathrm{Cov}\left[\widehat{\mu}(t_1),\widehat{\mu}(t_2)\right] = \frac{1}{N^2}\left[\left(\frac{p(t_1,t_2)}{p(t_1)p(t_2)} - 1\right)\sum_{c\in\mathscr{C},\, i\in c,\, j\in c \atop i\neq j} Y_i(t_1)\,Y_j(t_2) - \sum_{i=1}^{N} Y_i(t_1)\,Y_i(t_2)\right]$$

利用上述分解 5，我们可以将协方差改写为

$$\begin{aligned}\mathrm{Cov}\left[\widehat{\mu}(t_1),\widehat{\mu}(t_2)\right] &= \frac{1}{N^2}\left[\left(\frac{p(t_1,t_2)}{p(t_1)p(t_2)} - 1\right)\left(\sum_{c\in\mathscr{C}} S_c(t_1,t_2) - Q(t_1,t_2)\right) - Q(t_1,t_2)\right] \\ &= \frac{1}{N^2}\left[\left(\frac{p(t_1,t_2)}{p(t_1)p(t_2)} - 1\right)\sum_{c\in\mathscr{C}} S_c(t_1,t_2) - \frac{p(t_1,t_2)}{p(t_1)p(t_2)}Q(t_1,t_2)\right]\end{aligned}$$

■

已知估计的平均潜在结果 2 和 3 的方差与协方差，我们可以轻松计算估计的平均处理效应（$\widehat{\tau}$）的方差：

$$\mathrm{Var}\left[\widehat{\tau}\right] = \mathrm{Var}\left[\widehat{\mu}(1)\right] + \mathrm{Var}\left[\widehat{\mu}(0)\right] - 2\,\mathrm{Cov}\left[\widehat{\mu}(1),\widehat{\mu}(0)\right] \tag{7}$$

S2.5 保守方差估计

接下来，我们为 Horvitz–Thompson估计量 $\widehat{\mu}(t)$ 构造方差和协方差估计量。我们遵循 Aronow and Samii（2017）的做法，并将其结果应用于我们的设计：所有单位具有相同的边际倾向概率 $p_i(t) = p(t)$，且同一簇中的单位具有相同的联合倾向概率 $p_{ij}(t_1,t_2) = p(t_1,t_2)$。

S2.5.1 方差估计量

Aronow 和 Samii （2017）针对 $\widehat{\mu}(t)$ 的方差提出了以下基于设计的估计量：

$$\widehat{\mathrm{Var}}\left[\widehat{\mu}(t)\right] = \frac{1}{N^2}\sum_{i:Z_i=t}\sum_{j:Z_j=t}\left(\frac{1}{p_i(t)p_j(t)} - \frac{1}{p_{ij}(t,t)}\right)Y_iY_j$$

根据我们的设计，对于所有i 和$p_{ij}(t,t) = p(t,t)$，同一簇中的$p(t) = i$与$p \neq j$具有相同关系；而对于不同簇中的单元，$p_{ij}(t,t) = p(t)^2$。使用与其中相同的簇级记号

定理 2 中，该估计量可简化为

$$\widehat{\mathrm{Var}}\big[\widehat{\mu}(t)\big] = \frac{1}{N^2}\left[\left(\frac{1}{p(t)^2} - \frac{1}{p(t,t)}\right)\sum_{c\in\mathscr{C}}\widehat{S}_c(t,t) + \left(\frac{1}{p(t,t)} - \frac{1}{p(t)}\right)\widehat{Q}(t,t)\right] \tag{8}$$

其中

$$\widehat{S}_c(t_1,t_2) = \sum_{\substack{i:i\in c\\ Z_i=t_1}}\sum_{\substack{j:j\in c\\ Z_j=t_2}} Y_iY_j$$
$$\widehat{Q}(t) = \sum_{i:Z_i=t} Y_i^2$$

Aronow 和 Samii（2017）表明，当所有联合倾向概率 $p_{ij}(t,t)$ 对 $i \neq j$ 都严格为正时，该方差估计量是无偏的：

$$\mathbb{E}\big[\widehat{\mathrm{Var}}\big[\widehat{\mu}(t)\big]\big] = \mathrm{Var}\big[\widehat{\mu}(t)\big]$$

在打孔设计下，此条件成立，因为每一对单位同时接受处理t的概率均为正。

S2.5.2 协方差估计量

对于 $\widehat{\mu}(t_1)$ 与 $\widehat{\mu}(t_2)$ 之间的协方差，其中 $t_1 \neq t_2$, ，Aronow and Samii（2017）提出了以下估计量：

$$\begin{aligned}\widehat{\mathrm{Cov}}\big[\widehat{\mu}(t_1),\widehat{\mu}(t_2)\big] = \frac{1}{N^2}\Big[&\sum_{i:Z_i=t_1}\sum_{\substack{j:Z_j=t_2\\ j\neq i}}\left(\frac{1}{p_i(t_1)p_j(t_2)} - \frac{1}{p_{ij}(t_1,t_2)}\right)Y_iY_j\\ &- \sum_{i:Z_i=t_1}\frac{Y_i^2}{2p_i(t_1)} - \sum_{i:Z_i=t_2}\frac{Y_i^2}{2p_i(t_2)}\Big]\end{aligned} \tag{9}$$

在我们的设计下，对于$i \neq j$属于同一簇的情况，$p_i(t) = p(t)$，且$p_{ij}(t_1,t_2) = p(t_1,t_2)$；对于属于不同簇的单位，$p_{ij}(t_1,t_2) = p(t_1)p(t_2)$。与前面一样，只有簇内配对会对第一项作出贡献。使用簇级记号，该估计量可简化为

$$\begin{aligned}\widehat{\mathrm{Cov}}\big[\widehat{\mu}(t_1),\widehat{\mu}(t_2)\big] = \frac{1}{N^2}\Big[&\left(\frac{1}{p(t_1)p(t_2)} - \frac{1}{p(t_1,t_2)}\right)\sum_{c\in\mathscr{C}}\widehat{S}_c(t_1,t_2)\\ &- \frac{1}{2p(t_1)}\widehat{Q}(t_1,t_1) - \frac{1}{2p(t_2)}\widehat{Q}(t_2,t_2)\Big]\end{aligned} \tag{10}$$

正如 Aronow 和 Samii（2017）所示，只要 $p_{ij}(t_1,t_2) > 0$对于$i \neq j$成立，且结果变量Y_i为非负数，该协方差估计量就会向下有偏。这一条件在我们的随机化设计中得到满足，因此

$$\mathbb{E}\big[\widehat{\mathrm{Cov}}\big[\widehat{\mu}(t_1),\widehat{\mu}(t_2)\big]\big] \leq \mathrm{Cov}\big[\widehat{\mu}(t_1),\widehat{\mu}(t_2)\big]$$

表 S3：对各结果变量中治疗期间效应在处理后时期持续比例的估计与推断。

结果变量	$\widehat{\beta}$	标准误（簇稳健）	p值
仇恨帖子的对数	0.751	0.00376	$< 10^{-6}$
原创仇恨帖子的对数	0.642	0.00430	$< 10^{-6}$
仇恨内容转发日志	0.752	0.00422	$< 10^{-6}$
帖子总数的对数	0.876	0.00273	$< 10^{-6}$

S2.5.3 平均处理效应的保守方差

结合上述方差估计量和协方差估计量，可得到 ATE 方差的代入估计量

$$\widehat{\mathrm{Var}}\left[\widehat{\tau}_{\mathrm{diff}}\right] = \widehat{\mathrm{Var}}\left[\widehat{\mu}(1)\right] + \widehat{\mathrm{Var}}\left[\widehat{\mu}(0)\right] - 2\widehat{\mathrm{Cov}}\left[\widehat{\mu}(1), \widehat{\mu}(0)\right] \tag{11}$$

由于对于每个 t，$\widehat{\mathrm{Var}}\left[\widehat{\mu}(t)\right]$ 都是无偏的，而 $\widehat{\mathrm{Cov}}\left[\widehat{\mu}(1)\ \widehat{\mu}(0)\right]$ 低估了真实协方差，因此式（11）中的估计量是保守的：

$$\mathbb{E}\left[\widehat{\mathrm{Var}}\left[\widehat{\tau}_{\mathrm{diff}}\right]\right] \geq \mathrm{Var}\left[\widehat{\tau}_{\mathrm{diff}}\right]$$

补充信息和正文中报告的所有直接效应置信区间，均基于这一保守方差估计量。

S2.6 随机化推断

作为补充的稳健性检验，我们还报告了通过对精确图聚类设计下的随机化推断检验进行反演而得到的月度结果的95%置信区间。对于每个月份，我们通过回归Y_i^m对处理进行回归，并将分箱后的$Y^{\mathrm{pre}}{}_i$作为协变量，计算处理系数；从打孔设计中模拟10,000种分配，并将未在5%显著性水平上被拒绝的效应集合构建为置信区间。

S3 直接效应的持续性

我们研究的一个关键问题是，在不再投放广告后，干预期间观察到的处理效应在多大程度上仍能持续。我们并不通过直接比较处理后时期的处理组和对照组来衡量持续性，而是遵循 Lin 等人（2025）提出的方法，估计处理期间处理效应中在处理后时期仍然保留的比例。该方法将持续性建模为即时变化与延迟变化之间的比例关系；与在随访时比较处理组相比，研究表明，这种方法能够提供显著更高的统计功效和更强的可解释性（Lin 等人，2025）。下一节将简要介绍这一方法，并呈现持续性参数估计的结果。

S3.1 概念方法

为对持续性进行形式化描述，我们首先分别为处理期间和处理后时期指定结果模型。令$Y_{pre\,i}$,$Y^{during}{}_i$,$Y_{post\,i}$表示用户i在处理前、处理期间和处理后时期的结果，并令$Z_i \in \{0,1\}$表示处理分配。在处理期间，我们将结果建模为

$$Y_i^{\text{during}} = \alpha^{\text{during}} + \gamma^{\text{during}} Y_i^{\text{pre}} + \tau Z_i + \varepsilon_i^{\text{during}}, \tag{12}$$

其中，τ是处理期间的效应。纳入$Y_{pre\,i}$可调整后续结果中的任何线性趋势。同样，我们将处理后时期的结果建模为

$$Y_i^{\text{post}} = \alpha^{\text{post}} + \gamma^{\text{post}} Y_i^{\text{pre}} + \beta\,\tau Z_i + \varepsilon_i^{\text{post}}.$$

系数β表示处理期间效应中τ在处理后时期持续存在的比例，也是我们关注的估计目标。$\beta = 1$表示效应完全持续，而$\beta = 0$表示效应完全衰减。与处理期间模型一样，项$\gamma^{\text{post}}Y$pre $_i$用于捕捉与持续性无关的线性趋势。

尽管这些结构模型有助于阐明β的解释，但直接根据处理指标估计持续性在统计上可能效率较低，尤其是在处理后效应较小时。相反，我们可以通过一个将处理后结果与处理期间结果联系起来的单一回归模型来估计β：

$$Y_i^{\text{post}} = \alpha' + \gamma' Y_i^{\text{pre}} + \beta\, Y_i^{\text{during}} + \varepsilon_i'. \tag{13}$$

然而，由于结果变量的方差较大，这一设定的检验效能可能不足。按照我们分析中采用的相同策略以及 Lin 等人（2025）的方法，我们改为考察相对于处理前时期的变化量：

$$\begin{aligned}\Delta_i^{\text{during}} &= Y_i^{\text{during}} - Y_i^{\text{pre}} \\ \Delta_i^{\text{post}} \quad &= Y_i^{\text{post}} - Y_i^{\text{pre}}.\end{aligned}$$

随后，我们使用以下方法估计持续性参数

$$\Delta_i^{\text{post}} = \alpha' + \gamma' Y_i^{\text{pre}} + \beta\,\Delta_i^{\text{during}} + \varepsilon_i'. \tag{14}$$

Lin 等人（2025）的模拟表明，该估计量能够一致地恢复真实的持续性参数，并且与直接根据处理分配估计处理后的处理效应相比，具有显著更高的统计功效。

S3.2 结果

我们针对正文中报告的各项结果变量分别估计了公式（14）。我们没有对处理前结果使用单一线性项，而是将其取值范围划分为 40 个大小相等的区间组，并在模型中将其作为虚拟变量纳入。这种做法能更好地考虑潜在的

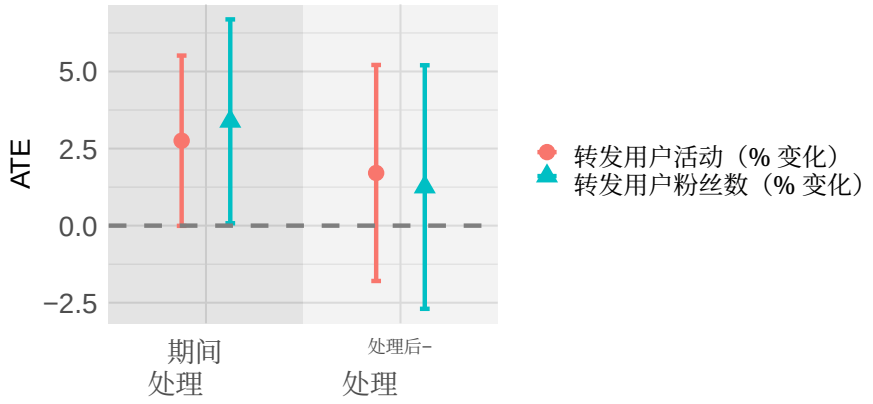


图 S2：被转发账户的构成。被转发账户的平均粉丝数对数和平均状态数对数的估计平均处理效应。误差条对应于95%的置信区间。

趋势非线性。由于我们的实验采用聚类随机化设计分配处理，我们使用了聚类稳健标准误，并按随机化聚类进行聚类，这些随机化聚类由 3-net算法构建。表 S3 报告了估计的持续性系数 β，以及聚类稳健标准误和p值。

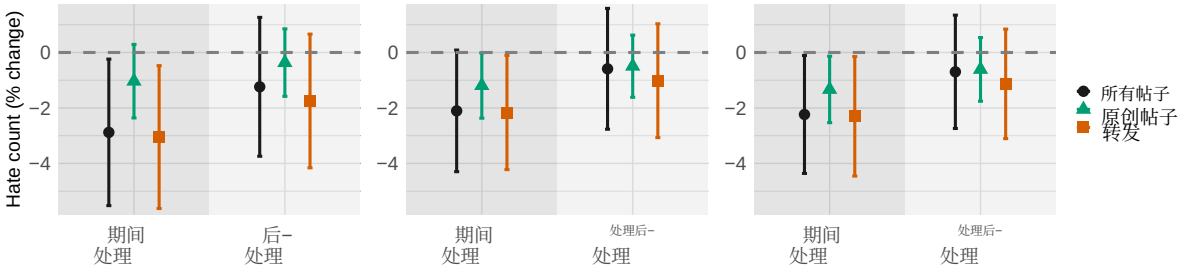
S4 直接效应的稳健性

为评估主要研究发现（图2）的稳健性，我们开展了多项稳健性检验。这些检验验证了估计的处理效应不会受到以下因素影响：(i) 处理前调整时期的选择，或 (ii) 使用仇恨的替代测量。除验证研究设计外，我们还在预期不会产生任何效应的期间进行了负向对照检验。每项稳健性检验均采用正文所述的估计程序，并在实验设计下使用相同的模型设定和推断方法。

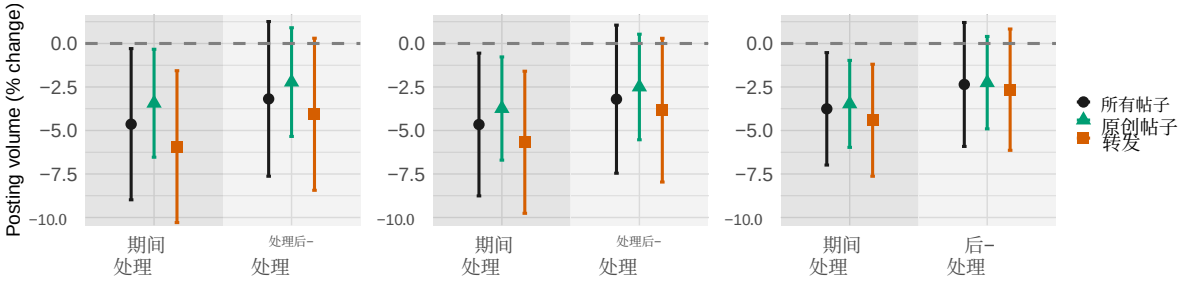
S4.1 替代处理前时期

为确保我们的结果并非由处理前时期（2023年1月1日至2023年7月31日）的选择所驱动，我们使用三个替代基线期重新估计了所有处理效应：(i) 2023年1月1日至2023年9月30日，(ii) 2023年1月1日至2023年5月31日，以及 (iii) 2023年5月1日至2023年8月31日。如 S4.3 节所述，处理组用户在暴露后删除处理前两个月发布的帖子的比例高于对照组用户。这表明，2023年10月至11月，以及在较小程度上2023年9月，可能包含部分处理泄漏，因此被排除在处理前窗口之外。

对于每种设定，我们都重复使用正文中的差分估计量，估计六种对数变换结果的平均处理效应：仇恨帖子总数、原创仇恨帖子、转发的仇恨帖子、帖子总数、原创帖子和转发。图 S3 展示了这些处理前定义下各主要结果变量的估计系数。表 S4 报告了每种结果变量在各处理前窗口下的平均处理效应及其 p 值，包括正文中采用的处理前窗口。



（a）仇恨相关结果



（b）发帖量结果

图 S3：不同替代处理前时期下处理效应的稳健性：2023年1月1日至5月31日（左）；2023年1月1日至9月30日（中）；2023年5月1日至8月31日（右）。误差条表示95%置信区间。

点估计在所有处理前设定下均一致，且统计显著性仍达到$p < 0.05$，只有一种情况例外。这些结果表明，观察到的仇恨内容减少并非特定处理前调整选择造成的假象。

S4.2 仇恨的替代测量

正文分析将仇恨定义为一个二元指标：对所有帖子求和，当 NaijaXLM-T 赋予的仇恨分数超过 0.5 的阈值时，该指标取值为 1。为评估结果对这一界定的敏感性，我们使用其他二元阈值、连续仇恨得分以及基于关键词的测量方法，重复进行估计。

二元阈值。 我们对正文分析进行了调整，采用模型赋予的仇恨分数的不同二元阈值（0.1、0.75 或 0.9）。图 S4 报告了各阈值对应的平均处理效应。在所有阈值下，处理效应均为负值，且大小相近，这表明我们的发现并非由临界值的选择所致。

连续分数。作为将仇恨分数二值化的替代方案，我们直接对各帖子经模型赋予的仇恨分数求和，并采用与主要分析相同的对数转换来构建结果变量。该指标对预测仇恨强度较高的帖子赋予更大权重，同时保留低于任意固定阈值的信息。图 S5 报告了由此得到的时期层面的平均处理效应，以及按月计算的基于随机化推断的估计，与正文中的图 2A 和

表 S4：不同处理前窗口下各结果变量的平均处理效应。

结果变量	2023年1月至9月	2023年1月至5月	2023年5月至8月	2023年1月至7月（正文）
仇恨总量	−2.1%（$p = 0.06$）	−2.9%（$p = 0.032$）	−2.2%（$p = 0.039$）	−2.5%（$p = 0.032$）
原始仇恨	−1.2%（$p = 0.046$）	−1.0%（$p = 0.126$）	−1.3%（$p = 0.029$）	−1.3%（$p = 0.033$）
转发的仇恨内容	−2.2%（$p = 0.039$）	−3.1%（$p = 0.020$）	−2.3%（$p = 0.036$）	−2.6%（$p = 0.021$）
帖子总数	−4.7%（$p = 0.026$）	−4.6%（$p = 0.036$）	−3.8%（$p = 0.022$）	−4.6%（$p = 0.030$）
原创帖子	−3.7%（$p = 0.013$）	−3.4%（$p = 0.030$）	−3.5%（$p = 0.006$）	−3.8%（$p = 0.013$）
转发	−5.7%（$p = 0.006$）	−5.9%（$p = 0.008$）	−4.4%（$p = 0.007$）	−5.7%（段落 $= 0.007$）

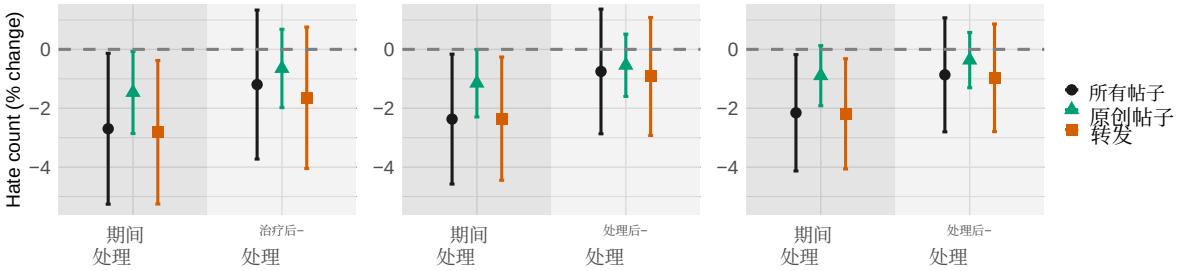


图 S4：使用替代结果变量检验处理效应的稳健性。该结果变量根据仇恨分数的二元阈值（左图为 0.1，中图为 0.75，右图为 0.9）定义为仇恨帖子的计数。

B 类似。估计效应的方向和大小均与使用二元仇恨指标得到的结果一致。

基于关键词的计数。 作为额外的稳健性检验，我们根据仇恨相关关键词是否出现，构建了一项仇恨的替代指标。为筛选这些关键词，我们利用 Tonneau 等（2024）的标注数据集，计算每个词在仇恨帖子和非仇恨帖子中的出现频率。随后，我们计算这些频率的比值，以衡量某个词出现在仇恨帖子中而非仇恨帖子中的可能性高出多少。根据这一排序，我们选出排名前 K 的仇恨相关关键词，其中 $K \in \{10, 50, 100, 200\}$，并以至少包含一个此类关键词的帖子计数来衡量仇恨。图 S6 显示，根据这些定义，处理效应仍为负且显著，结果一致。点估计大于基于阈值的二元指标。

总体而言，采用二元、基于分数和基于关键词的指标进行的稳健性分析，证实了处理效应在不同仇恨度量下具有稳定性。表 S5 报告了所有替代定义下的平均处理效应及其 p 值。

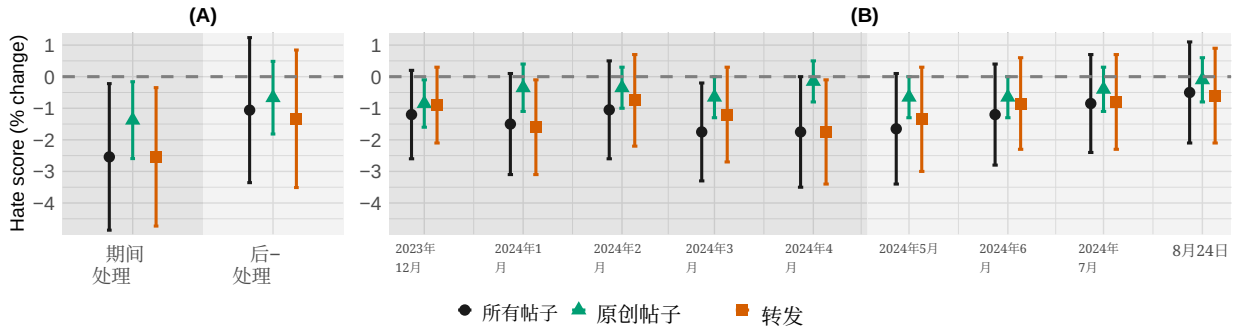


图 S5：以仇恨分数为结果变量时处理效应的稳健性：处理期间及处理后（A），以及按月份划分（B）。

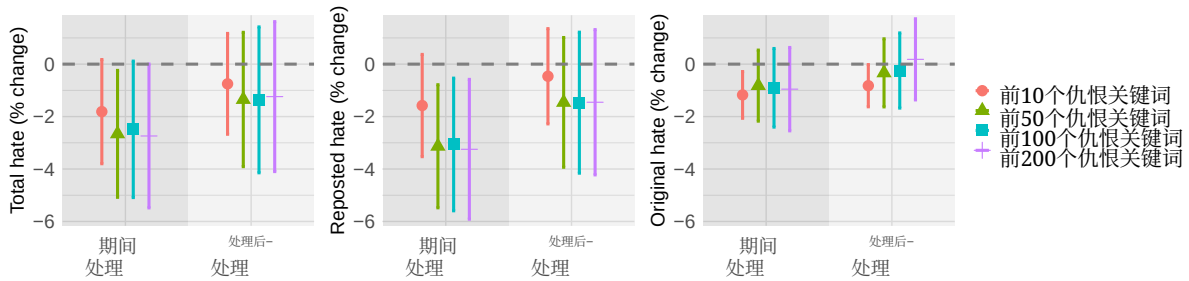


图 S6：使用替代结果变量（定义为包含与仇恨相关的前列关键词的帖子数量）时处理效应的稳健性。各面板分别显示仇恨帖子总量（左）、转发的仇恨帖子（中）和原创仇恨帖子（右）的结果。

S4.3 负向对照检验

为进一步验证结果，我们进行了负向对照分析，以评估宣传活动开始前处理组与对照组之间的差异。如果未发现系统性的处理前差异，我们对估计结果的信心将进一步增强，因为这表明干预期间观察到的效应反映了广告的真实因果影响，而非既有趋势。

我们将宣传活动前紧邻的几个月称为安慰剂期。估计程序与正文中的主分析相同：使用正文所述的同一差分估计量，根据处理前基线对安慰剂期内测量的结果进行调整。在这项检验中，基线调整期为2023年1月1日至6月1日，因此2023年6月1日至11月28日为安慰剂期。

如果实验设计有效，我们预计安慰剂期内的效应为零或微不足道，干预期间的处理效应会更强，而处理后月份的效应则会减弱。图 S7 展示了二元仇恨计数、仇恨分数和发帖量的月度处理效应，证实了这一模式。安慰剂期内估计的月度处理效应至少在最初四个月与零在统计上无显著差异，这表明观察到的处理效应并非源于两组之间既有的行为差异。不过，在处理前最后两个月，部分效应略微转为负值。我们将后期下降解读为处理组用户的选择性内容删除：这些用户在接受处理后，回溯性地删除了此前发布的一些帖子，这些帖子发布于

表 S5：在不同仇恨定义下，治疗期间对数结果变量的平均处理效应（以百分比变化表示）及其 p 值。

定义	仇恨总量	原始仇恨	转发的仇恨内容
分数 ≥ 0.1	$-2.70,\ p = 0.039$	$-1.46,\ p = 0.040$	$-2.82,\ p = 0.024$
分数 ≥ 0.75	$-2.37,\ p = 0.035$	$-1.15,\ p = 0.049$	$-2.35,\ p = 0.028$
分数 ≥ 0.9	$-2.15,\ p = 0.033$	$-0.89,\ p = 0.086$	$-2.19,\ p = 0.022$
原始分数	$-2.54,\ p = 0.032$	$-1.38,\ p = 0.027$	$-2.54,\ p = 0.023$
前10个关键词	$-1.81,\ p = 0.074$	$-1.17,\ p = 0.010$	$-1.58,\ p = 0.112$
前50个关键词	$-2.66,\ p = 0.031$	$-0.82,\ p = 0.236$	$-3.13,\ p = 0.009$
前100个关键词	$-2.49,\ p = 0.061$	$-0.90,\ p = 0.238$	$-3.06,\ p = 0.018$
前200个关键词	$-2.74,\ p = 0.050$	$-0.96,\ p = 0.239$	$-3.25,\ p = 0.017$

近期。此前关于直接反仇恨言论的社交媒体实验也记录了类似的删除行为，见 Hangartner 等人（2021）。

S4.4 参与者流失

在本节中，我们评估这项干预是否导致不同组别的样本流失。对于每个结果变量，我们使用公式11中定义的同一 Horvitz–Thompson 差值估计量和保守方差估计量来估计平均处理效应。所有指标的处理组和对照组流失率均无显著差异。

S4.4.1 处理前不活跃

我们首先考察在处理推出前变为不活跃、并在整个实验期间一直保持不活跃的用户。对于每位用户，我们定义一个指示变量，用以表示其在处理前六个月内是否不活跃。使用公式11中的 HT 差值估计量和保守方差估计量，我们发现处理组与对照组的不活跃率估计差值为 -0.00197（$p = 0.346$），表明对照组的不活跃率略高。该差异在统计学上并不显著，说明处理组之间的处理前不活跃程度没有可检测到的不平衡。

S4.4.2 处理后无法访问

接下来，我们评估不同处理组中用户因删除、设为私有或暂停使用而无法访问的比率是否不同。我们将相同的估计框架应用于处理后无法访问指示变量，估计得到的平均处理效应为 -0.000724 ($p = 0.769$)。这一差异在统计上并不显著，表明该干预并未提高用户在处理后变得无法访问的可能性。

S4.4.3 综合流失

最后，我们考虑一种综合指标，用以反映用户是否（i）在处理前变为不活跃，或（ii）在处理后变得无法联系。采用相同的估计量，该指标的平均处理效应为

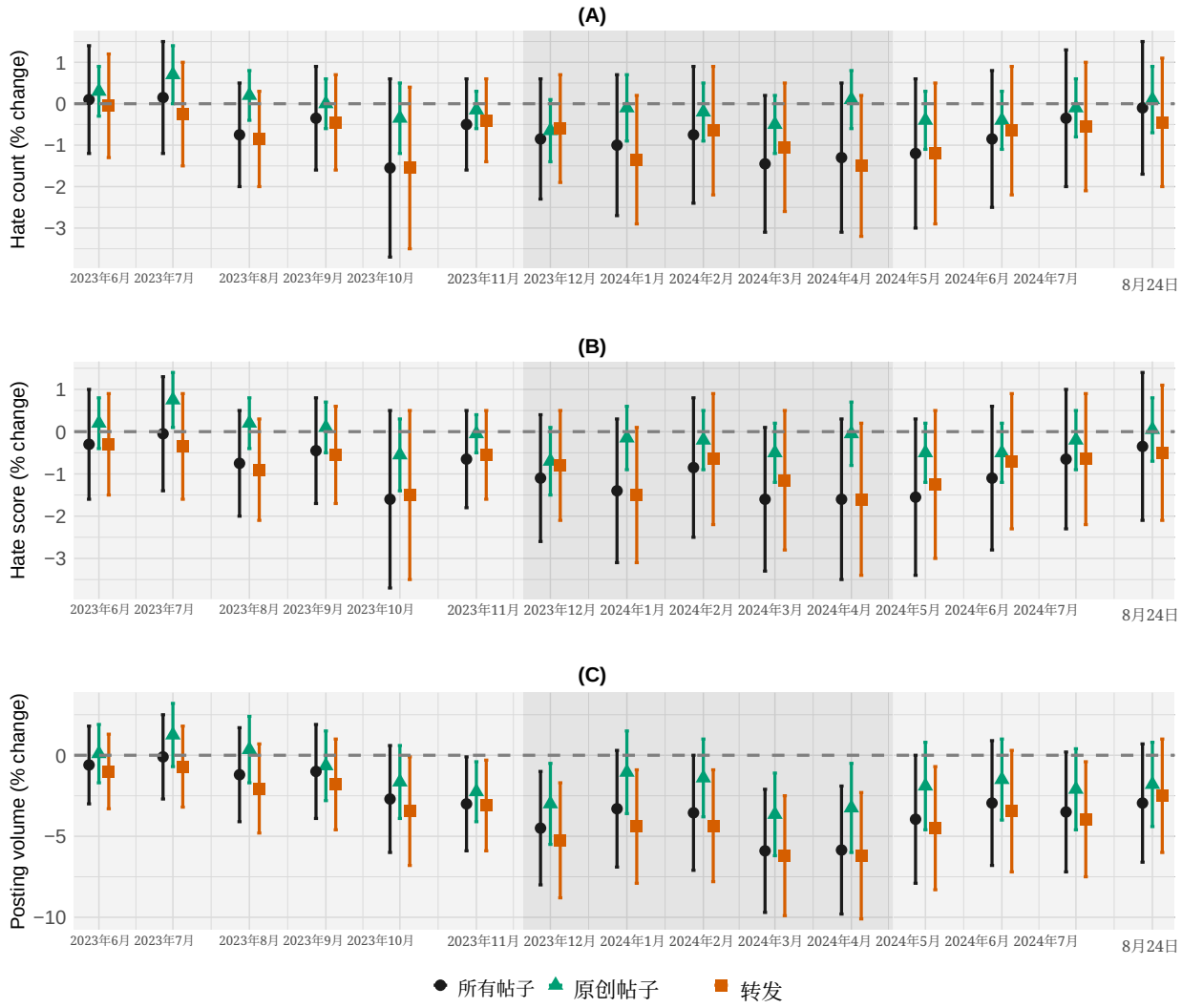


图 S7：基于相对于处理前基线估计的月度处理效应进行的负向对照检验。处理效应针对与正文相同的结果变量（二元仇恨计数和发帖量）以及连续仇恨得分进行估计。第一个灰色区域表示安慰剂期，第二个和第三个灰色区域分别表示处理期间和处理后月份。误差条表示 95% 随机化推断置信区间。

综合流失指标为 -0.00269 ($p = 0.458$)。同样，对照组的流失率略高，但处理组与对照组之间的差值在统计上并不显著。

总体而言，干预并未导致差异性流失，这表明选择性退出不太可能使估计的处理效应产生偏差。

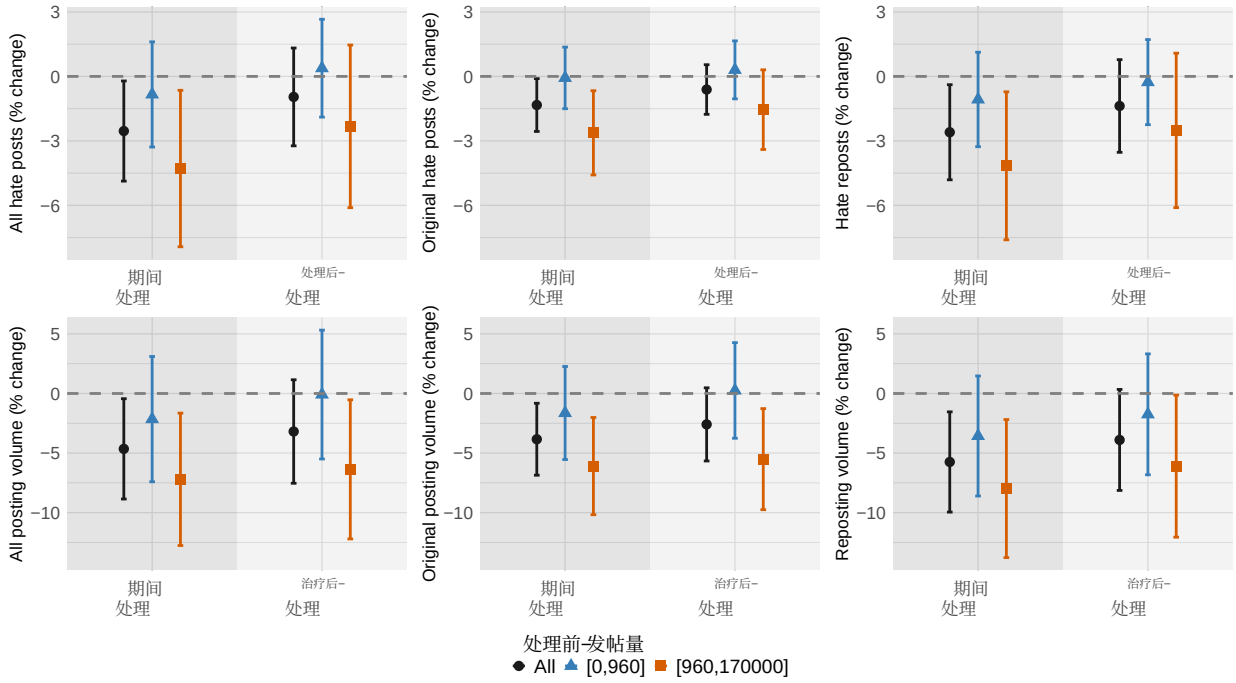


图 S8：按处理前活动水平划分的仇恨相关结果（上排）和发帖量（下排）的处理效应异质性。每个面板报告了高活动和低活动用户（高于／低于中位数）在不同结果变量上的估计效应。点估计以百分比变化表示，条形表示 95%置信区间。

S5 转发用户构成

如图2所示，与对照组用户相比，处理组用户总体上的发帖活动和转发活动较低。在本节中，我们考察在发生转发的条件下，处理组用户转发的账号的构成是否不同于对照组用户转发的账号。

为描述被转发账号的构成，我们记录每次转发所涉及账号的特征，重点关注（i）关注者数量和（ii）状态数量，后者定义为该账号自创建以来发布的帖子总数。由于这两个变量均高度偏斜，我们对其取对数。对于每位用户，我们计算其在处理前、处理期间和处理后所转发账号的平均粉丝数对数和平均状态数对数，并对该用户的所有转发取平均值。这些用户层面的平均值作为结果变量。随后，我们采用与直接效应分析相同的估计和推断程序，估计干预对各结果变量的平均处理效应（ATE）。

图 S2 报告了估计的处理效应。尽管转发量有所减少，但处理组用户在转发时往往会转发来自更受欢迎、发帖活动也更活跃的账号的内容。具体而言，处理组用户转发的账号，其关注者平均约多 3.4%，账号存续期间的发帖活动也高出 2.8%。这些结果表明，处理使转发行为转向高知名度账户，而非一概降低参与度。

表 S6：处理期间，按处理前活动划分的仇恨帖子和发帖量结果的平均处理效应（ATE）、标准误（SE）和 p 值。估计值以对数单位表示，并以百分比变化呈现。每个单元格第一行列出 ATE，第二行列出（SE，p 值）。

结果变量	All	Low	High
仇恨总量	-2.54 （1.19，$p = 0.032$）	-0.84 （1.24，$p = 0.50$）	-4.29 （1.86，$p = 0.021$）
原始仇恨	-1.33 (0.627, $p = 0.033$)	-0.070 (0.731, 段落 = 0.92)	-2.62 (1.00，$p = 0.009$)
转发的仇恨内容	-2.60 (1.13, 段落 = 0.021)	-1.07 (1.12, 段落 = 0.34)	-4.16 (1.76, 段落 = 0.018)
帖子总数	-4.65 (2.15, 段落 = 0.030)	-2.15 (2.68，$p = 0.42$)	-7.20 (2.83，$p = 0.011$)
原创帖子	-3.84 (1.54, 段落 = 0.013)	-1.64 (1.99, $p = 0.41$)	-6.09 (2.08, 段落 = 0.003)
转发	-5.75 (2.15, $p = 0.007$)	-3.57 （2.57，$p = 0.16$）	-7.98 (2.95，$p = 0.007$)

S6 直接效应的异质性

正文中，我们记录了直接处理效应在仇恨帖子方面的异质性。本节将异质性分析扩展至正文分析的其余结果变量：发帖总量、原创帖子、转发、原创仇恨帖子和仇恨内容转发。具体而言，我们考察处理效应如何随干预前用户行为的两个关键维度而变化：（i）总体发帖量；（ii）处理前仇恨内容占比。我们根据每个异质性变量的中位数，将用户分为高于和低于中位数的两组。各子组内的处理效应采用正文和 S2 节所述的相同框架进行估计。本分析评估发帖更活跃的用户，或参与仇恨内容相对较少的用户，是否会在所有结果测量指标上对宣传活动作出不同反应。

S6.1 处理前活动的异质性

图 S8 和表 S6 展示了按处理前发帖活动分组的用户的处理效应。在所有结果变量上，处理效应的方向均与正文报告的结果一致：活跃度高于中位数的用户，其仇恨内容和发帖总量均出现更大且具有统计显著性的下降；而不太活跃用户的效应较小，且不具有统计显著性。这些模式表明，干预对高参与度用户更有效，因为他们接触广告的机会更多，也更有可能在网络行为中产生可测量的变化。尽管直接比较时，高活跃度组与低活跃度组之间的差异并无统计显著性，但点估计始终显示，活跃用户的反应更强。这进一步支持了这样的解释：参与强度会增强用户对亲社会信息传播的响应。

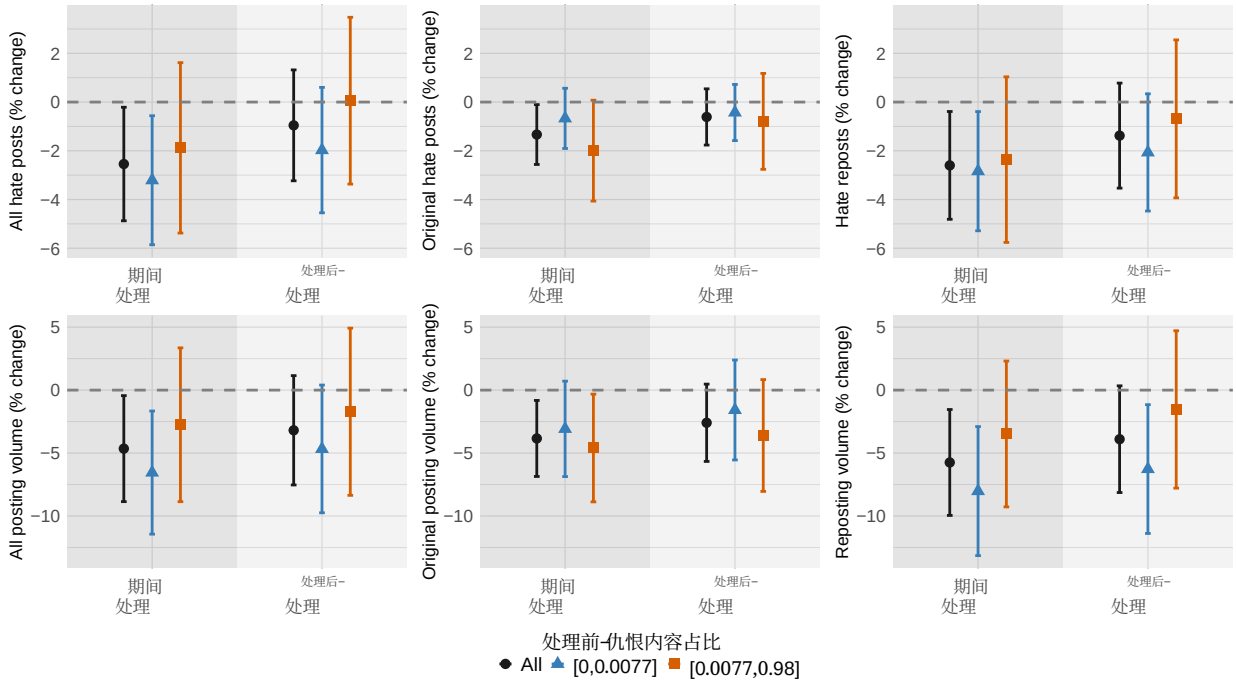


图 S9：按处理前仇恨内容占比分组的仇恨相关结果（上排）和发帖量（下排）的处理效应异质性。每个面板报告处理前活动中仇恨内容占比高于和低于中位数的用户的估计效应。点估计以百分比变化表示，条形对应95%置信区间。

S6.2 按处理前仇恨内容占比分组的处理效应异质性

图 S9 和表 S7 报告了按处理前仇恨内容占比分组用户时的类似结果。效应再次与正文所述结果相呼应：对于基线活动中仇恨内容占比较低的用户，宣传活动更有效地减少了仇恨内容。对于仇恨内容占比高于中位数的用户，效应较弱，通常不具有统计显著性，这与这些用户对规范性信息传播的响应较低相一致。然而，较少倾向于仇恨内容的用户，其仇恨性发帖行为和总体发帖行为均出现了可测量的减少。

S7 参与者中的间接效应

本节介绍我们如何定义和估计参与者中的间接效应（图3C）。我们首先介绍因果估计量和暴露映射的作用，随后说明用于操作化暴露条件的加权q比例暴露模型，将处理组低暴露、对照组高暴露和处理组高暴露效应定义为正式的估计目标，解释如何获得暴露倾向，最后概述基于 Gao 和 Ding（2025年）加权回归方法的估计流程。本文始终假设存在局部干扰，即用户的结果变量可能取决于其自身的分配以及其直接邻居的分配，但不取决于相隔超过一跳的单元。

表 S7：处理期间所有结果变量按处理前仇恨内容占比分组的平均处理效应（ATE）、标准误（SE）和p值。估计值以对数单位表示，并转换为百分比变化。每个单元格第一行列出 ATE，第二行列出（SE，p值）。

结果变量	**All**	**Low**	**High**
仇恨总量	-2.54 (1.19，$p = 0.032$)	-3.21 (1.35, $p = 0.017$)	-1.88 (1.78, 段落 $= 0.29$)
原始仇恨	-1.33 (0.627, $p = 0.033$)	-0.67 (0.630, $p = 0.29$)	-1.99 (1.06, 段落 $= 0.059$)
转发的仇恨内容	-2.60 (1.13, $p = 0.021$)	-2.84 (1.25, $p = 0.023$)	-2.36 (1.73, 段落 $= 0.17$)
帖子总数	-4.65 (2.15，$p = 0.030$)	-6.55 (2.49, $p = 0.009$)	-2.75 (3.12，$p = 0.37$)
原创帖子	-3.84 (1.54, $p = 0.013$)	-3.08 （1.93，段落 $= 0.11$）	-4.60 (2.18, $p = 0.035$)
转发	-5.75 (2.15，$p = 0.007$)	-8.02 (2.61，$p = 0.002$)	-3.48 (2.96，$p = 0.24$)

S7.1 估计目标与暴露映射

令$Z = (Z1, \ldots, Z_N)$表示随机处理分配向量，令$Y_i(z)$表示参与者i在分配向量z下的潜在结果。目标估计目标比较这样两个假想情境：所有用户都接受处理，$z = (1,1,\ldots,1)$，以及所有用户都未接受处理，$z = (0,0,\ldots,0)$，

$$\tau = \frac{1}{N}\sum_{i=1}^{N}\left[Y_i(1,1,\ldots,1) - Y_i(0,0,\ldots,0)\right].$$

映射 $z \mapsto Y_i(z)$ 是高维的，无法直接估计，因为不可能同时观测两个世界。为了解决这一问题，我们通常使用暴露映射，即函数$T_i = T(i,Z,A)$，将完整的分配向量和网络 A 映射为每个用户 i的低维暴露类别t 。在映射设定正确的情况下，完整处理向量 z 下的潜在结果$Y_i(t)$ 可简化为，这意味着只有暴露类别会影响用户的结果变量。这一简化使我们能够将因果估计目标定义为不同暴露类别之间平均潜在结果的差值，而不是完整分配向量之间的差值。更重要的是，如果暴露映射函数相对简单，例如在局部干扰假设下，则单个随机化方案中即可观测到全部暴露条件，从而实现识别。

S7.2 加权q分数暴露模型

为划分暴露条件，我们采用加权q比例暴露模型，其中$q = 0.7$；该模型将 Ugander 等（2013）的q比例框架推广到加权网络。对于每个用户i，令W_i 表示互动网络中i的所有邻居的边权重之和，如 S1.3 节所述。令$W_{i,T}$ 表示指向接受处理的邻居的总权重。假设用户i

当至少 70% 的邻居（按其边权重加权）接受处理时，该用户被视为间接受到处理暴露，即$W_{i,T}/W_i \geq 0.7$。根据这一规则，可将用户划分为四种暴露条件之一。

- **低暴露对照：**用户未接受处理，且至少 70% 的邻居也未接受处理。
- **高暴露对照：**用户未接受处理，但至少 70% 的邻居接受了处理。
- **低暴露处理：**用户接受了处理，但至少 70% 的邻居未接受处理。
- **高暴露处理：**用户接受了处理，且至少 70% 的邻居也接受了处理。

这一加权的 q 比例规则近似表达了这样一种观点：结果取决于有影响力的邻居中是否有足够大比例的人接受处理，而不要求所有邻居都接受处理。

S7.3 暴露对比

令 $\mu(t) = \frac{1}{N}\sum_{i=1}^{n} Y_i(t)$ 表示暴露条件为 t 时的平均潜在结果。处理组低暴露、对照组高暴露和处理组高暴露效应，定义为这些平均潜在结果之间的对比。处理组低暴露（TL）效应比较的是，当邻居实际上未接受处理时，接受处理的用户与未接受处理的用户，

$$\tau_{\mathrm{TL}} = \mu(\text{treated-low exposure}) - \mu(\text{control-low exposure}) \tag{15}$$

对照组高暴露（CH）效应比较的是，邻居接受处理的未处理用户与邻居未接受处理的未处理用户，

$$\tau_{\mathrm{CH}} = \mu(\text{control-high exposure}) - \mu(\text{control-low exposure}) \tag{16}$$

处理组高暴露（TH）效应比较的是，邻居接受处理的处理组用户与邻居未接受处理的未处理用户，

$$\tau_{\mathrm{TH}} = \mu(\text{treated-high exposure}) - \mu(\text{control-low exposure}) \tag{17}$$

由于暴露取决于个体及其邻居是否接受处理，打孔式聚类随机化所产生的邻居随机变化对于识别全部三个估计目标是必要的。

S7.4 倾向得分

对于每位用户i 和暴露类别t，令 $\pi_i(t) = \Pr(T_i = t)$ 表示在打孔随机化设计下，用户i落入暴露t的概率。这些概率已知

来自设计，但由于两阶段分配和加权q分数暴露条件，无法通过解析方法计算。因此，我们通过运行 50,000 次随机化程序的蒙特卡罗模拟来估计 $\pi_i(t)$。每次模拟中，我们在簇级重新抽取处理，以 0.18 的概率独立翻转每项分配，重新计算所有用户的暴露类别，并记录每位用户达到各暴露类别的频率。估计的倾向概率是$T_i = t$的模拟经验比例。我们还推导了一种基于动态规划的解析方法，类似于 Ugander 等（2013）针对 q 分数暴露模型所述的方法。不过，由于打孔步骤和加权暴露规则，这种动态规划更为复杂，因此此处不予介绍。尽管如此，我们确认，通过蒙特卡罗模拟估计的倾向概率与解析方法得到的结果非常接近。估计得到的 $\pi_i(t)$ 可确定用于估计的逆倾向权重$w_i = 1/\pi_i(T_i)$。

S7.5 通过加权回归进行估计

为估计方程 15、16 和 17 中每种暴露类别对应的目标估计目标 $\tau(t)$，我们采用 Gao 和 Ding（2025）的加权回归估计量，并以逆倾向概率作为权重；在我们的设定中，该方法等同于 H´ajek 估计量。令 $z_i(t)$ 表示用户 i 是否在暴露t下被观测到的指示变量。加权最小二乘通过对Y_i 关于完整的暴露指示变量集合进行回归来实现，权重为$w_i = 1/\pi_i(T_i)$。

在我们的实现中，该回归采用前文所述的相同双重差分结果变量，并将每种暴露类别的虚拟变量作为外生变量纳入模型，同时以分为 40 个等规模区间的处理前发帖量作为协变量，以提高估计精度。根据这一估计量，与暴露t 对应的拟合系数等于 H´ajek 逆概率加权均值，

$$\hat{\mu}(t) = \frac{\sum_i z_i(t) w_i Y_i}{\sum_i z_i(t) w_i}.$$

这种估计方法有两个优点。首先，它等同于我们在预注册中指定的 H´ajek 估计量，同时提供了一个统一的、基于回归的框架，可按需纳入协变量。其次，正如 Gao 和 Ding（2025）所示，该估计量在存在干扰的情况下仍保有良好的基于设计的性质，且其方差可通过网络稳健 HAC 方差估计量进行一致估计。

根据每种暴露条件下估计的平均潜在结果，通过对相关 $\hat{}\mu(t)$ 值取简单差值，得到对应于高强度处理组、高强度对照组和处理组低暴露效应的对比。我们使用加权回归中的同一网络稳健 HAC 协方差矩阵估计量计算这些对比的标准误，并假设暴露均值的对比是回归系数的线性组合。

正文中的图 3 展示了我们的结果。具体而言，我们发现，在此暴露模型下，间接效应（τ_{CH}）在统计上并不显著。采用另一种暴露阈值选择方式，即$q = 0.8$ 比例时，结果在定性上仍保持不变。

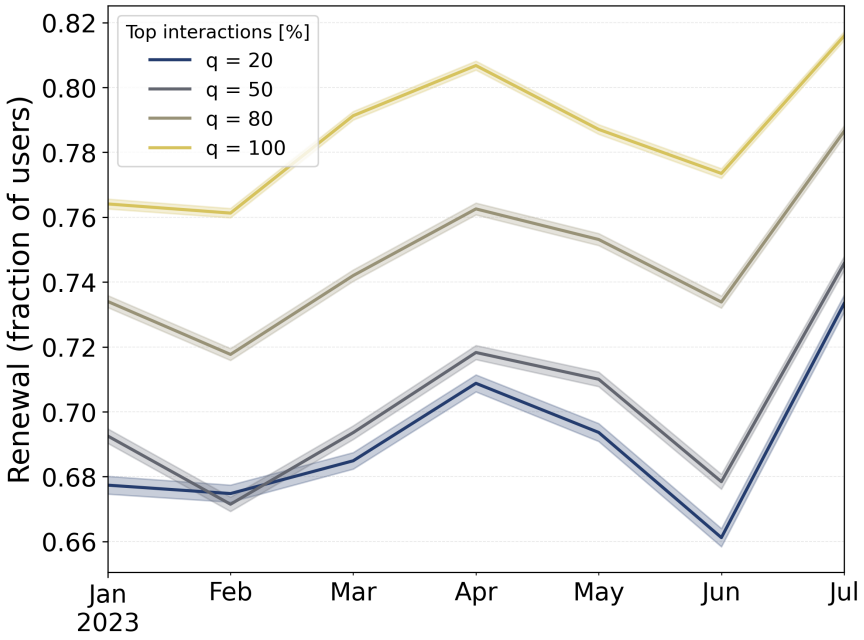


图 S10：参与者转发账号的月度更替情况。更替率是指在月份t 被转发、但在月份$t+1$未被转发的账号比例；不同条件水平依据互动最频繁的账号划分（见图例）。阴影区域表示 95% 置信区间。

S7.6 被转发用户的更替情况

由于分析涵盖了从2023年1月1日至2024年8月31日的较长时期，因此可以合理地预期，互动模式会随着研究的进行而演变。因此，我们探究参与者转发的账号集合随时间变化的速度。对于每位参与者u和每个月份t，令 $\mathcal{S}_u(t)$表示u 在t月转发的账号集合。我们将更新率定义为

$$r_u(t) = 1 - \frac{|\mathcal{S}_u(t) \cap \mathcal{S}_u(t+1)|}{|\mathcal{S}_u(t)|}, \tag{18}$$

它反映了月份t 用户u转发的账号中，在月份$t+1$未再次出现的比例。为了检验参与者是否有一组相对稳定的核心转发账号，我们还计算了以用户u 互动最频繁的账号中排名前q 百分位为条件的更替率。也就是说，我们依据月份t互动次数排名前q 的账号集合，对 $\mathcal{S}_u(t)$ 进行条件化。

图 S10 展示了不同条件水平下参与者的平均更新率。无条件限制时，平均月度更新率为 0.78，这表明被转发账户中有很大一部分会从一个月份变更到下一个月份。即使将关注范围限制在互动最频繁的账户中，更新率仍然很高：聚焦于互动量排名前 20% 的账户时，平均更新率为 0.68。这些结果表明，参与者转发的账户集合会随时间发生显著变化，这意味着在单个时间点构建的互动网络只能提供短暂的快照。这些更新率的大小与 Kolic 等人（2025年）记录的用户高周更新率高度一致。

此外，参与者转发的账户中，只有 3.39% 的账户本身也是参与者，这表明大多数相关互动路径都位于实验样本之外。高周转率与参与者彼此分享对方内容的低比例相结合，为我们未能检测到参与者中的间接效应提供了一个合理解释。

S8 对上游用户的间接效应

S8.1 上游用户的选择

为识别候选上游用户，我们收集了80,154名参与者在2023年8月1日至11月28日的处理前时期转发的全部仇恨内容。从这些内容中，我们提取出原始作者，并排除了参与者。随后，对于剩余的作者，我们确定了在同一时期转发过其任何原创内容的全部用户。为避免纳入参与者仅占其总参与度中极小比例的极高活跃账户，我们将分析范围限定为处理前时期发帖数少于15,000条的用户。对于每位剩余用户，我们计算了转发占比，其定义为该用户收到的全部处理前转发中归因于参与者的转发所占的比例。形式上，设 $\mathscr{O}_i$ 为上游用户i 在处理前创建的原创帖子的集合；对于每条原创帖子$k \in \mathscr{O}_i$，设 $\mathscr{R}_{ik}$ 表示转发该帖子的全部用户集合。转发占比为

$$F_i = \frac{\sum_{k\in\mathscr{O}_i}\sum_{u\in\mathscr{R}_{ik}} \mathbf{1}\{u \in \text{participants}\}}{\sum_{k\in\mathscr{O}_i} |\mathscr{R}_{ik}|} \tag{19}$$

我们将 F_i 值最高的 400 名上游用户定义为主要分析集。由于F_i 反映了用户受众中可能暴露于实验的最大比例，因此也代表该用户可能受到处理的最大暴露程度，我们将在下一节对此进行分析。

S8.2 暴露定义

对于每位上游用户 i，我们参照转发占比 F_i，以类似方式定义其对实验的暴露，但现在将每位参与者转发者的处理状态纳入考量。换言之，暴露是处理组参与者所带来的处理前转发占比：

$$T_i = \frac{\sum_{k\in\mathscr{O}_i}\sum_{u\in\mathscr{R}_{ik}} \mathbf{1}\{u \in \text{participants}\}\mathbf{1}\{u \text{ treated}\}}{\sum_{k\in\mathscr{O}_i}\sum_{u\in\mathscr{R}_{ik}} \mathbf{1}\{u \in \text{participants}\}}. \tag{20}$$

因此，T_i 表示上游用户处理前参与度中归因于被分配至处理组的用户的比例。T_i 值越高，用户i的潜在间接暴露程度越高，因为其处理前受众中处理组参与者所占比例更大。注意，对于任何上游用户i，$0 \le T_i \le F_i \le 1$ 。

S8.3 随机化推断的检验统计量

对于每个结果变量，我们使用基于回归的检验统计量，估计暴露T_i 与上游用户结果变量之间的边际关联。为处理直接效应并提高统计功效，只要存在有效的处理前测量值，我们就采用差分估计量。令$Y^{\text{during}}{}_i$ 表示处理期间的结果（处理后期间的估计方法类似），$Y^{\text{pre}}{}_i$ 表示相应的处理前数值。对于有定义此类处理前基线的结果（例如参与者或非参与者的仇恨内容转发日志），我们构造调整后的结果

$$\Delta_i^{\text{during}} = Y_i^{\text{during}} - Y_i^{\text{pre}} \tag{21}$$

并使用以下公式估计检验统计量

$$\Delta_i^{\text{during}} = \alpha + \theta T_i + \gamma_{s(i)} + \varepsilon_i, \tag{22}$$

其中，$\gamma_{s(i)}$ 表示上游用户处理前总发帖量各分组的固定效应。此协变量调整通过剔除用户之间与处理暴露无关的水平差异，提高了估计精度。

对于缺乏有意义的处理前基线的结果变量，例如流失转发者数量（该数量是相对于处理前受众定义的），我们使用原始结果 $Y^{\text{during}}{}_i$ 并估计：

$$Y_i^{\text{during}} = \alpha + \theta T_i + \gamma_{s(i)} + \varepsilon_i \tag{23}$$

$\hat{}$在所有情况下，我们都将系数 θ 用作随机化推断的检验统计量。它本身并非因果估计量，而是在随机化程序下对其分布进行评估的统计量。

S8.4 因果估计对象

我们假设采用线性暴露—响应模型

$$Y_i(T_i) = Y_i(0) + \tau T_i, \tag{24}$$

其中，τ 表示暴露增加一个单位时结果变量的边际变化（即所有处理前的转发者都接受处理）。由于T_i 的取值范围为 0 到 1，因此结果按T_i增加一个百分点进行报告。尖锐零假设 H_0： $\tau = \tau_0$ 完全规定了所有潜在结果。

S8.5 随机化推断

为检验尖锐零假设 H_0：$\tau = \tau_0$ 对任意给定的 τ_0 值（包括表示不存在间接效应的 $\tau_0 = 0$），并为 τ 构造置信区间，我们按如下方式实施随机化推断（RI）：

1. 通过打孔重复进行两阶段分配 10,000 次，并根据方程（20），针对每次模拟分配 b 和每位上游用户 i，重新计算 $T^{(b)}{}_i$。

2. 对每个分配 b，计算在零假设结果模型成立时本应观测到的调整后的结果 $\widetilde{\Delta}^{(b)}$。对于每个 τ_0，计算

$$\widetilde{\Delta}_i^{(b)}(\tau_0) = \Delta_i - \tau_0 T_i + \tau_0 T_i^{(b)}$$

其中，T_i 是实际处理分配下上游用户 i 的暴露，而 $T^{(b)}{}_i$是新分配 b下模拟得到的暴露。

3. 对每个 b，以 $\widetilde{\Delta}^{(b)}(\tau_0)$ 为因变量、以 $T^{(b)}{}_i$ 为自变量进行回归，并纳入相同的活动分箱固定效应 $\gamma_{s(i)}$，然后记录斜率估计值 $\widehat{\theta}^{(b)}(\tau_0)$。

4. $\{\widehat{\theta}^{(b)}(\tau_0)\}_{b=1}^{10,000}$的经验分布近似于尖锐零假设 H_0下检验统计量的分布。

$\widehat{\ }$RI 的 p 值是观测到的 θ在零假设分布 $\{\widehat{\theta}^{(b)}\}_{b=1}^{10,000}$下的双侧尾部概率（Rosenbaum，2007）。通过在一组 τ_0取值上检验 H_0：$\tau = \tau_0$并对检验结果求逆（即确定接受域），我们得到因果参数 τ的置信区间。这种方法通过在集群随机化和打孔所诱导的依赖结构下生成零假设分布，确保了精确推断。此外，该方法明确考虑了潜在的间接效应，除尖锐零假设和线性暴露-反应结果模型外，不需要任何建模假设。图 5 和补充信息中报告的所有 RI 结果均遵循这一程序。

S9 间接效应的持续性

与直接效应一样，我们将持续性参数定义为治疗期间上游用户所受间接效应中持续到处理后时期的比例。由于间接效应分析依赖随机化推断（RI）程序，我们将 Lin 等人（2025）的方法调整到 RI 框架中，并通过反演 RI 检验估计持续性参数的置信区间。下文假设，对于每个上游用户 i，我们都能观察到其在干预期间及干预后的结果变量，分别记为 Y期间 $_i$和 $Y^{\text{post}}{}_i$，例如仇恨内容转发日志，或处理前转发者的对数流失量。

S9.1 概念方法

如 S8 节所述，令 T_i 表示上游用户 i 的暴露程度（在采用打孔的两阶段设计中，i 的受众中接受处理的比例）。我们假设采用线性暴露—响应模型，其中宣传活动期间的间接效应为 τ，而处理后的间接效应是处理期间效应的 $\beta \in [0,1]$：

$$Y_i^{\text{during}}(T_i) = Y_i^{\text{during}}(0) + \tau T_i$$
$$Y_i^{\text{post}}(T_i, \beta) = Y_i^{\text{post}}(0) + \beta \tau T_i$$

其中，Y期间 $_i(0)$ 和 $Y^{\text{post}}{}_i(0)$ 分别表示上游用户 i 在期间和处理后处于零暴露状态下的潜在结果。持续性参数 β 表示处理期间效应在宣传活动结束后仍持续的比例。

S9.1.1 关于 τ 的条件设定。

针对 β 的随机化推断（RI）需要针对每一次处理分配的模拟重新随机化，构造相应的暴露 T_i 以及两个时期中隐含的潜在结果。在上述模型下，这些潜在结果取决于处理期间间接效应的大小 τ。如果不指定 τ，我们就无法在其他分配方案下插补反事实结果 $Y^{\text{during}}{}_i(T_i)$ 和Y处理后 $_i(T_i,\beta)$，因而也无法生成持续性统计量的随机化推断分布。

因此，我们在一系列合理的 τ 取值上进行随机化推断。对于每个候选 τ，我们（i）针对一系列 β_0 取值，在每种模拟分配下插补处理期间和处理后的结果；（ii）反演由此得到的随机化推断检验，以获得在给定该 τ 条件下 β 的置信区间。随后，对于每个假设的 τ，我们报告估计的 β、其随机化推断置信区间，以及针对无持续性原假设的随机化推断p值。

S9.1.2 处理前调整与差分结果。

与全文其他部分一样，我们通过考察相对于处理前基线的变化量来提高精度。令$Y^{\text{pre}}{}_i$ 表示上游用户i的处理前时期结果，并定义

$$\Delta_i^{\text{during}} = Y_i^{\text{during}} - Y_i^{\text{pre}}$$
$$\Delta_i^{\text{post}} = Y_i^{\text{post}} - Y_i^{\text{pre}}$$

S9.1.3 随机化推断检验统计量

随机化推断程序中使用的检验统计量，是以下回归模型中 $\Delta^{\text{during}}{}_i$ 的系数：

$$\Delta_i^{\text{post}} = \alpha + \theta\,\Delta_i^{\text{during}} + \gamma_{s(i)} + \varepsilon_i \tag{25}$$

其中，$\gamma_{s(i)}$ 是间接效应分析中使用的相同处理前活动分箱固定效应。我们可以模拟随机化程序：对于每次随机化，在尖锐零假设下根据 β_0 的取值推算结果，并估计检验统计量。这样便能构造 θ 检验统计量的尖锐零假设分布，并将其与观测值进行比较。

S9.1.4 随机化推断

我们针对尖锐零假设H_0 实施 RI： $\beta = \beta_0$ 以候选处理效应为条件 τ。该实现仿照第 S8 节中针对间接效应的 RI 程序：对于每次模拟重新随机化，我们 (i) 重新计算暴露；(ii) 根据零假设结果模型，通过将观察到的结果从实际暴露T_i“平移”至模拟暴露$T_{(b)\,i}$，构造零假设下本应观察到的结果。对于固定的一对 (τ,β_0)，按如下步骤进行：

表 S8：按假定的治疗期间效应划分的间接效应基于随机化推断的持续性估计 τ。对于每个结果变量和 τ，我们报告 95% 随机化推断置信区间以及 p 值，用于检验 $H_0: \beta = 0$。随机化推断基于 $0 \le \beta \le 1$ 进行，因此上限为 1。

假定的 τ	结果变量	95% 置信区间	p 值
面板 A：仇恨内容转发日志			
-0.01	仇恨内容转发日志（参与者）	(0, 1)	0.55
-0.01	仇恨内容转发日志（非参与者）	(0, 1)	0.29
-0.02	仇恨内容转发日志（参与者）	(0, 1)	0.11
-0.02	仇恨内容转发日志（非参与者）	(0.09, 1)	0.039
-0.03	仇恨内容转发日志（参与者）	(0.29, 1)	0.0097
-0.03	仇恨内容转发日志（非参与者）	(0.39, 1)	0.0061
-0.04	仇恨内容转发日志（参与者）	(0.43, 1)	0.0007
-0.04	仇恨内容转发日志（非参与者）	(0.5, 1)	0.0003
-0.05	仇恨内容转发日志（参与者）	(0.5, 0.99)	$< 10^{-6}$
-0.05	仇恨内容转发日志（非参与者）	(0.55, 1)	$< 10^{-6}$
B面板：日志中仇恨内容转发者流失			
0.03	日志中仇恨内容转发者流失（参与者）	(0, 1)	0.75
0.03	日志中仇恨内容转发者流失（非参与者）	(0, 1)	0.35
0.04	日志中仇恨内容转发者流失（参与者）	(0, 1)	0.11
0.04	日志中仇恨内容转发者流失（非参与者）	(0.91, 1)	0.008
0.05	日志中仇恨内容转发者流失（参与者）	(0.92, 1)	0.0065
0.05	日志中仇恨内容转发者流失（非参与者）	(0.96, 1)	0.0001
0.06	日志中仇恨内容转发者流失（参与者）	(0.97, 1)	0.0001
0.06	日志中仇恨内容转发者流失（非参与者）	(0.98, 1)	$< 10^{-6}$
0.07	参与者的日志中仇恨内容转发者流失	(0.98, 1)	$< 10^{-6}$
0.07	非参与者的日志中仇恨内容转发者流失	(0.99, 1)	$< 10^{-6}$

1. 通过打孔重复原始两阶段分配$B = 10{,}000$ 次。对于每次模拟分配b，通过方程 (20) 计算上游用户暴露$T^{(b)}{}_i$。

2. 对于每个b，构造在 τ和 β_0 成立时本应观察到的调整后的差分结果。具体而言，定义

$$\widetilde{\Delta}_i^{\text{during},(b)}(\tau) = \Delta_i^{\text{during}} - \tau T_i + \tau T_i^{(b)}$$
$$\widetilde{\Delta}_i^{\text{post},(b)}(\tau, \beta_0) = \Delta_i^{\text{post}} - \beta_0 \tau T_i + \beta_0 \tau T_i^{(b)}$$

其中，T_i 是上游用户i在实际分配下的暴露，而$T^{(b)}{}_i$ 是上游

用户i在模拟分配b下的暴露。

3. 对每个b，估计持续性回归

$$\widetilde{\Delta}_i^{\text{post},(b)}(\tau,\beta_0) = \alpha^{(b)} + \widehat{\theta}^{(b)}(\tau,\beta_0)\widetilde{\Delta}_i^{\text{during},(b)}(\tau) + \gamma_{s(i)} + \varepsilon_i^{(b)}$$

其中包括间接效应分析中使用的相同活动分箱固定效应$\gamma_{s(i)}$，并将斜率 $\widehat{\theta}_{(b)}(\tau,\ \beta_0)$记录为随机化推断检验统计量。

4. $\{\widehat{\theta}_{(b)}(\tau,\ \beta_0)\}_{b=1}^{B}$的经验分布近似于在$H_0:\beta=\beta_0$ 成立且以 τ为条件时检验统计量的零假设分布。

5. 对$H_0:\ \beta=\beta_0$ 进行随机化p值检验，其值为

$$p(\beta_0) = \frac{1}{B}\sum_{b=1}^{B}\mathbf{1}\{|\hat{\theta}^{(b)}(\tau,\beta_0)| \geq |\hat{\theta}(\tau,\beta_0)|\}$$

该值将观测到的检验统计量 $\hat{}\theta(\tau,\ \beta_0)$与其零假设分布进行比较。

6. 对一组候选值 β_0 重复步骤1–4，并收集所有未在 α 显著性水平（例如，$\alpha=0.05$）下被拒绝的 $H_0:\ \beta=\beta_0$ 对应的值，即可得到以间接效应 τ为条件的持续性参数 β 的（$1-\alpha$）置信区间。

S9.2 结果

我们采用上述随机化推断程序，在一系列合理的 τ 取值范围内，估计每种上游用户结果对应的持续性参数 β，正文中报告了这些结果。表 S8 报告了通过反演随机化推断检验得到的95%置信区间，以及与无持续性尖锐零假设相关的 p 值。在所有结果变量以及几乎所有 τ 取值下，我们发现间接效应在处理后时期仍具有显著持续性，这表明干预结束后，下游放大效应的降低仍在持续。

S10 间接效应的稳健性

本节考察我们对上游用户间接效应的估计在一系列替代测量和设计选择下的稳健性。全文采用与 S8 节相同的两阶段随机化和打孔方法，以及相同的基于随机化推断的推断程序，并使用相同的活动分箱固定效应$\gamma_{s(i)}$。我们报告的效应表示为上游用户处理组受众占比每增加一个百分点所对应的百分比变化，与正文中的图5一致。

S10.1 仇恨度量的替代方案

正文使用仇恨分数阈值 0.5，呈现上游用户原创内容的转发。我们通过以下方式评估结果的稳健性：(i) 使用其他二元阈值；(ii) 使用连续的、按分数加权的结果变量。

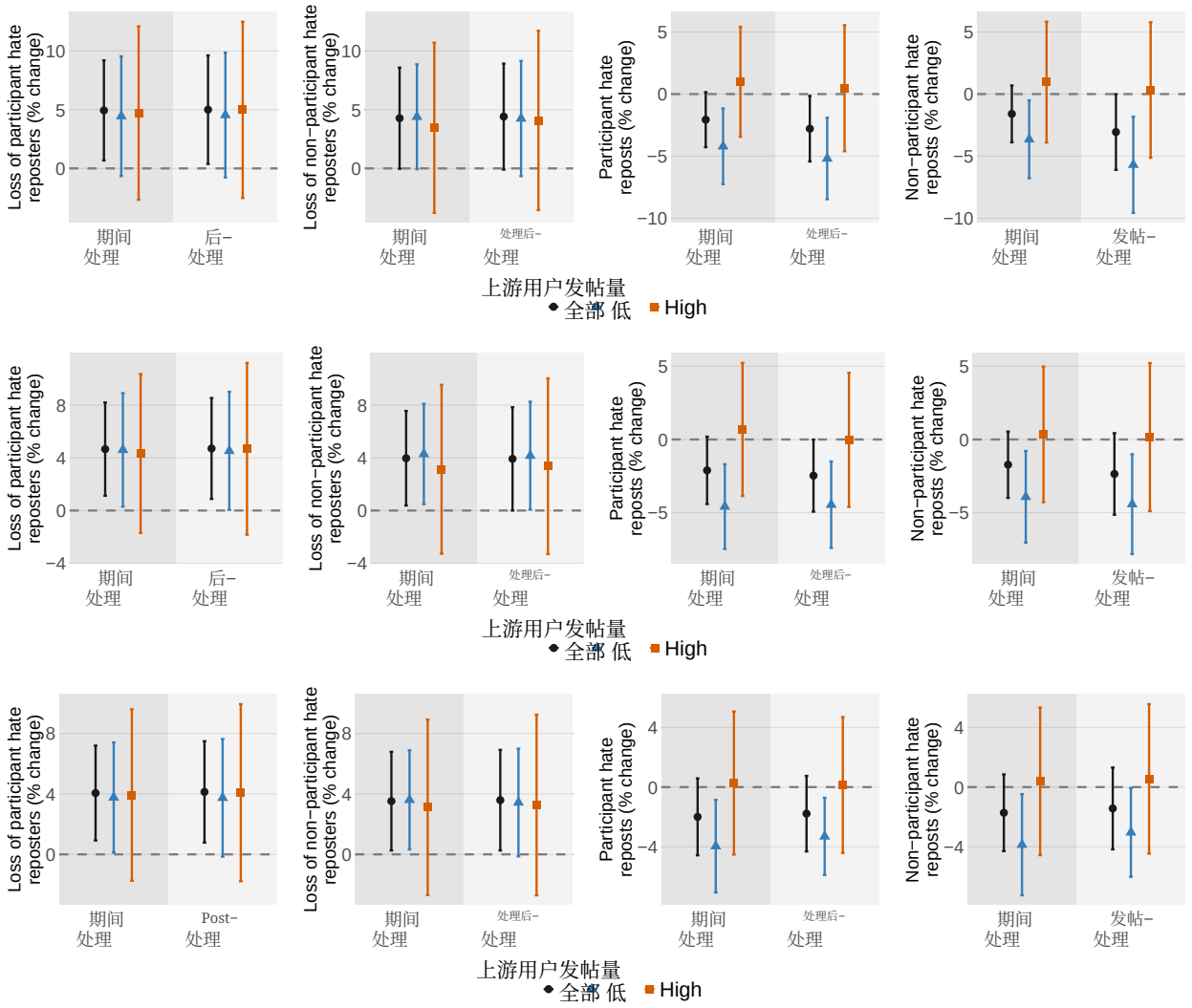
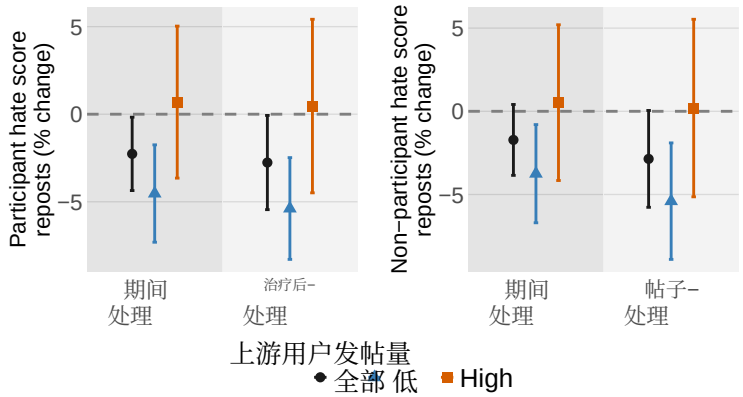


图 S11：在替代二元仇恨阈值 0.1（上）、0.75（中）和 0.9（下）下，对上游用户的间接效应。各面板报告了上游用户处理前受众中接受处理的比例每增加一个百分点所对应的百分比变化。结果变量的定义与正文中的图5相同。误差条表示 95% RI 置信区间。

二元阈值。 我们采用不同的二元阈值重新界定上游用户的仇恨内容。具体而言，对于上游用户发布的每条原创帖子，如果模型的仇恨分数高于阈值 $c \in \{0.1, 0.75, 0.90\}$，则将其仇恨指标设为 1，否则设为 0。对于每个阈值，我们重新计算原创仇恨帖子收到的仇恨内容转发，并针对四项关注的结果，重复进行相同的基于 RI 的间接效应估计（图 S11）。在不同阈值下，估计的间接效应在方向和幅度上仍与正文中的结果相近。

连续仇恨分数结果。 我们还用连续的分数加权指标替代二元仇恨转发计数。具体而言，我们不再只统计分数高于某一阈值的帖子的转发，而是将上游用户收到的所有被转发原创帖子的仇恨分数相加



图S12：使用连续仇恨分数结果的间接效应。各面板报告了处理组受众占比每增加一个百分点所对应的百分比变化，其中结果变量为每位上游用户收到的转发内容的仇恨分数总和。误差条表示 95% 随机化推断置信区间。

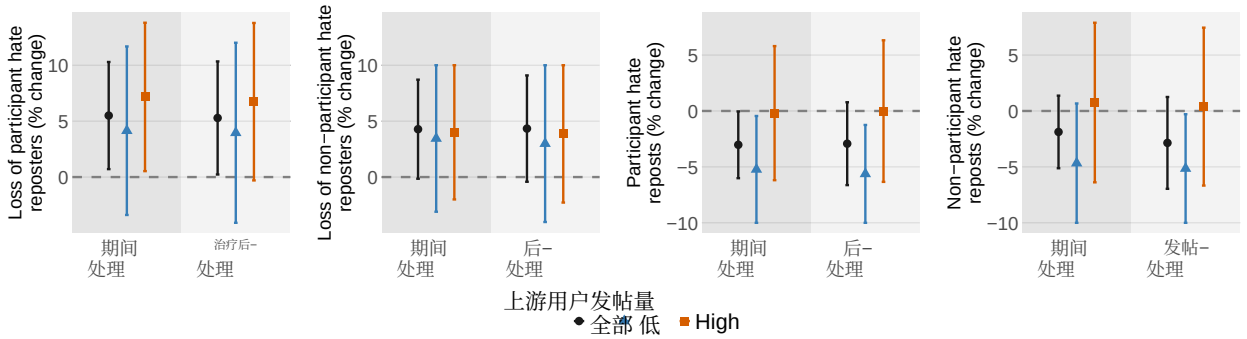
，从而得到一种结果变量：它会提高仇恨程度更高的转发的权重，同时保留任何固定临界值以下的信息。图S12报告了使用这一分数加权结果得出的参与者和非参与者的基于随机化推断的估计。对仇恨内容转发的估计效应显示，参与者的转发减少了 2.26%（$p = 0.037$），非参与者减少了 1.72%（$p = 0.098$），与主要估计结果一致。

S10.2 替代样本选择

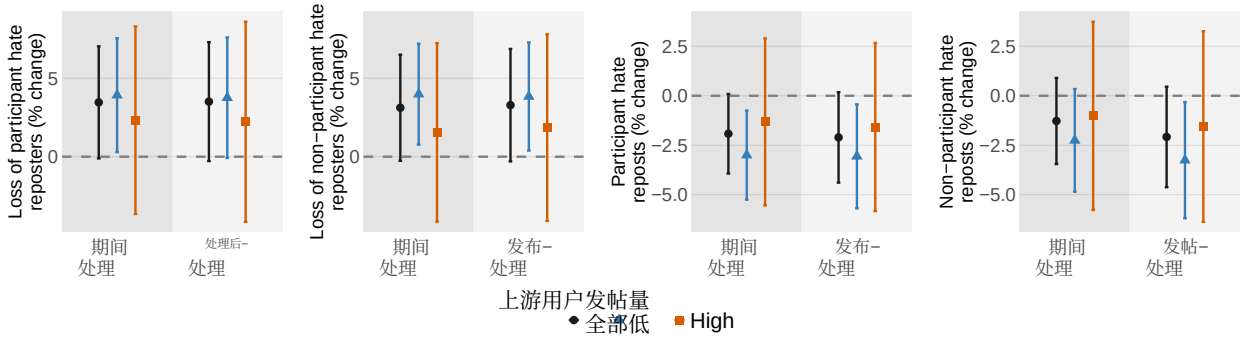
正文中呈现的预注册分析聚焦于按参与者潜在暴露度排名前400的上游用户。为评估结果对这一选择的稳健性，我们使用替代样本重复分析：根据公式19定义的潜在暴露度指标，选取排名前200和前600的上游用户。图S13展示了相应的估计结果。这些结果与正文图5中的发现高度一致。例如，在排名前200的用户中，参与者和非参与者转发的仇恨内容分别减少了3%（$p = 0.047$）和1.9%（$p = 0.25$）；处理组受众占比每增加一个百分点，参与者和非参与者中处理前转发仇恨内容的用户流失率则分别上升5.5%（$p = 0.034$）和4.3%（$p = 0.055$）。

S10.3 替代暴露定义

正文中，按照我们的预注册方案，上游用户是根据其对参与者的潜在暴露程度选出的，该程度以其处理前转发中可归因于参与者的比例来衡量。一种自然的替代方法是根据原创内容的仇恨分数为每条转发加权，使用户与仇恨程度更高的内容互动时，对测量暴露的贡献更大。因此，我们修改公式19中的原始潜在暴露度量，并定义一种



(a) 按曝光量排名前200的上游用户



(b) 按曝光量排名前600的上游用户

图S13：对具有最高潜在处理暴露量的前200名（上图）和前600名（下图）上游用户的间接效应。各面板报告上游用户处理前受众中接受处理的份额每增加一个百分点所对应的百分比变化。结果类型见图5。误差条表示 95% RI 置信区间。

替代性的仇恨加权度量：

$$F_i = \frac{\sum_{k\in\mathcal{O}_i}\sum_{u\in\mathcal{R}_{ik}}\mathbf{1}\{u\in\text{RCT}\}\,h_{ik}}{\sum_{k\in\mathcal{O}_i}\sum_{u\in\mathcal{R}_{ik}}h_{ik}} \tag{26}$$

其中，h_{ik} 是上游用户i发布的原创帖子k 的仇恨分数，$\mathcal{O}_i$ 是i在处理前发布的原创帖子的集合，$\mathcal{R}_{ik}$ 是处理前时期转发了帖子k 的用户集合。类似地，我们修改方程20中的实际暴露，定义实验参与者转发活动中归因于接受处理的 用户的仇恨加权份额

$$T_i = \frac{\sum_{k\in\mathcal{O}_i}\sum_{u\in\mathcal{R}_{ik}}\mathbf{1}\{u\in\text{RCT}\}\mathbf{1}\{u\text{ treated}\}\,h_{ik}}{\sum_{k\in\mathcal{O}_i}\sum_{u\in\mathcal{R}_{ik}}\mathbf{1}\{u\in\text{RCT}\}\,h_{ik}} \tag{27}$$

使用这一替代潜在暴露F_i，我们选取了排名前400的上游用户，并完全按照 S8 节的方法，重新进行了基于随机化推断的间接效应分析，同时将暴露度指标替换为方程 27 中仇恨加权的定义。图 S14 展示了采用这一替代暴露定义时的结果。根据这一指标，处理组受众每增加一个百分点，

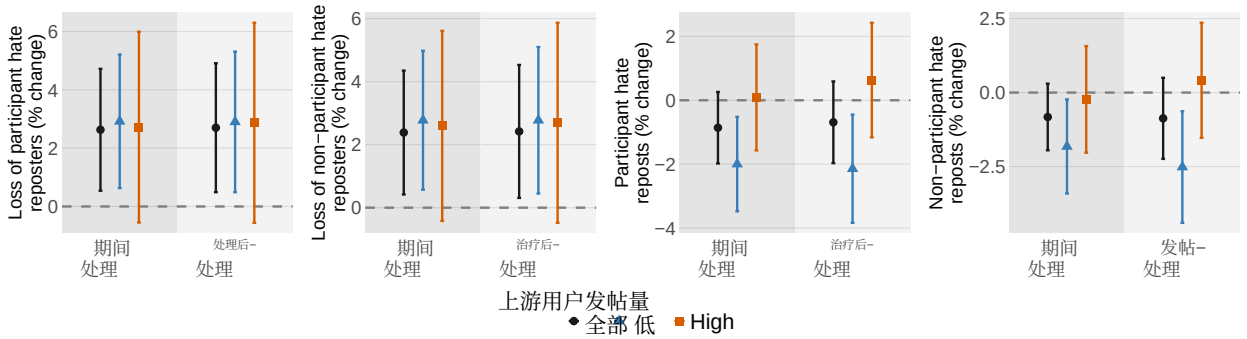


图 S14：使用仇恨加权暴露定义的间接效应。上游用户根据方程（26）中的替代指标选取排名前 400 名的用户，暴露于处理的定义见方程（27）。误差条表示 95% 随机化推断置信区间。

表 S9：间接效应对处理前时期选择的稳健性。各项结果报告处理期间结果变量的 95% 随机化推断置信区间；括号内报告针对零效应尖锐零假设的双侧随机化推断p值。

	仇恨内容转发者流失		仇恨内容转发	
处理前窗口	参与者	非参与者	参与者	非参与者
2023年6月1日至2023年11月28日	$[-1.0, 11.1]$% (段落 $= 0.078$)	$[-2.0, 10.0]$% (段落 $= 0.12$)	$[-5.7, -0.28]$% （段落 $= 0.037$）	$[-5.0, 0.54]$% (段落 $= 0.097$)
2023-07-01 至 2023年11月28日	$[-0.64, 10]$% (段落 $= 0.065$)	$[-1.1, 9.3]$% (段落 $= 0.095$)	$[-6.5, -0.55]$% (段落 $= 0.023$)	$[-6.1, 0.03]$% (段落 $= 0.051$)
2023-08-01 至 2023年11月28日 (正文)	$[1.6, 9.2]$% (段落 $= 0.013$)	$[0.68, 8.3]$% (段落 $= 0.025$)	$[-5.4, -0.54]$% ($p = 0.019$)	$[-4.4, 0.24]$% (段落 $= 0.075$)

暴露度指标显示，参与者的仇恨内容转发减少了 0.86% ($p = 0.14$)，非参与者减少了 0.83% ($p = 0.16$)。参与者中处理前转发仇恨内容的用户流失率增加了 2.7% ($p = 0.014$)，非参与者中则增加了 2.4% ($p = 0.019$)。与正文一样，上游发帖量带来的异质性依然存在：影响主要集中在低活跃用户中，而高活跃用户的下降在统计上不显著。

S10.4 替代处理前时期

由于受众流动性很大（如 S7.6 节所示），若要预测处理后转发行为，处理前指标应尽可能接近干预时间。因此，我们的主要设定采用 2023年8月1日至11月28日的处理前窗口来构建干预前调整。在本节中，我们使用两个时间跨度更长的替代处理前窗口重新估计间接效应：(i) 2023年7月1日至2023年11月28日，以及 (ii) 2023年6月1日至2023年11月28日。表 S9 报告了正文中研究的四项上游结果的基于随机化推断的估计：前

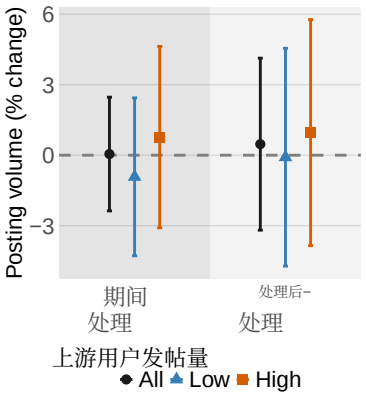


图 S15：使用上游用户发帖总量进行的负向对照检验。该图报告了处理组受众占比每增加一个百分点所对应的边际百分比变化。异质性是相对于处理前发帖量而言的。误差条表示 95% 随机化推断置信区间。

期参与者和非参与者仇恨内容转发量的变化，以及处理期间参与者和非参与者中处理前转发者的流失。每个单元格均报告基于随机化推断的 95% 随机化推断置信区间，以及针对零效应尖锐零假设的双侧随机化推断p值。最后一行对应于正文中的设定。在不同的替代处理前窗口下，估计的间接效应在大小和方向上相近。

S10.5 负向对照检验

最后，我们实施了两项安慰剂检验；对于这些检验，我们预计干预不会产生影响。首先，我们重复了 RI 分析，将上游用户的总体发帖活动作为安慰剂结果。由于活动广告只针对实验参与者，从未直接触达上游用户，因此我们预计他们的发帖量不会发生变化。图 S15 报告了该安慰剂结果的基于随机化推断的估计。与预期一致，估计效应较小，在统计上与零无显著差异（$p = 0.75$）。

作为第二项安慰剂检验，我们估计了完全发生在干预之前的时期内的间接效应。我们采用了上述相同的基于随机化推断的差异估计程序，并按照两种备选时期设定计算安慰剂效应：

1. 一种安慰剂设定：处理前调整期为 2023-01-01 至 2023- 04-01，处理期间安慰剂期为 2023-04-01 至 2023-08-01，处理后安慰剂期为 2023-08-01 至 2023-11-01。2. 另一种安慰剂设定：处理前调整期为 2023-04-01 至 2023- 07-01，处理期间安慰剂期为 2023-07-01 至 2023-11-01。在这种情况下，可用时间跨度不足以单独设定处理后安慰剂期。

由于这些期间均尚未实施处理，我们预计在两种设定中，上游暴露与结果变量之间都不存在系统性关系。图 S16 报告了这些安慰剂期基于随机化推断的估计。与预期一致，估计效应

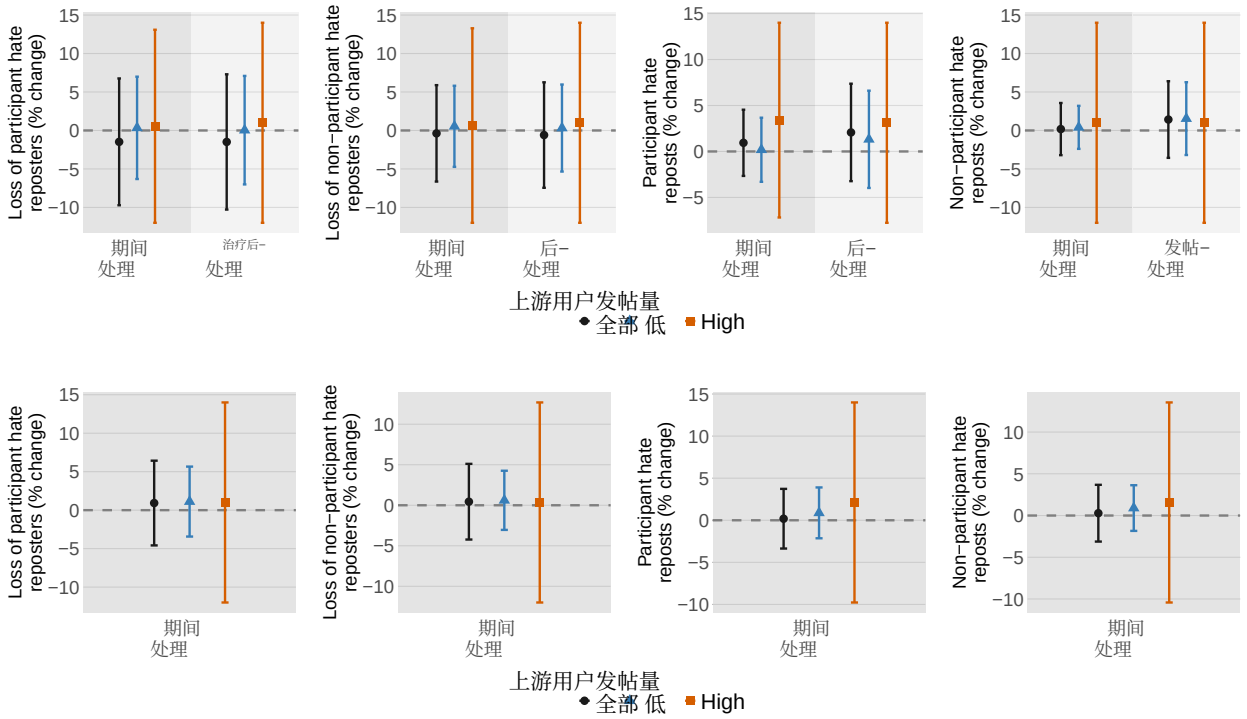


图 S16：使用干预前时期进行安慰剂检验。各面板报告的是基于实验启动前各时期计算的基于 RI 的间接效应估计值。上方面板对应设定 1，处理前时期为 2023-01-01 至 2023-04-01，涵盖处理期间和处理后时期的安慰剂期。下方面板对应设定 2，处理前时期为 2023-04-01 至 2023-07-01，仅涵盖处理期间的安慰剂期。误差条表示 95% RI 置信区间。

在参与者结果和非参与者结果中，无论是仇恨内容转发还是仇恨内容转发者流失，都接近于零，并且在统计上与零无显著差异。表 S10 报告了相应的点估计和 RI p 值。

S11 转发链中断检验

正文记录了对上游用户产生的间接效应，并延伸至从未被宣传活动针对的非参与者。对于这种非参与者效应，我们考察了两种相互竞争的机制。第一，这种效应可能机械地产生于转发级联的中断：如果处理组参与者减少转发，一些下游非参与者可能仅仅因为无法通过通常的转发链接触到上游帖子，而始终看不到这些内容。第二，处理组参与者可能以一种令平台将其解读为上游用户内容受欢迎程度下降的方式减少转发，从而降低这些内容的展示频率，进而使总体转发量下降，包括非参与者中的转发。由于第二种机制涉及平台内部机制，在无法访问排序与分发系统的情况下，很难直接检验。本文考察了第一种机制的影响，并未发现与转发链中断相一致的证据，这提高了平台响应解释的可信度

表 S10：使用干预前时期进行的基于随机化推断的安慰剂估计。为便于比较，最后一行列出了正文中使用正确时期设定得出的估计值。条目报告处理安慰剂时期内结果变量的 95 % 随机化推断置信区间；括号内报告针对零效应尖锐零假设的双侧随机化推断 p 值。

	仇恨内容转发者流失		仇恨内容转发	
处理前窗口	参与者	非参与者	参与者	非参与者
设定 *1:*				
2023-01-01 至 2023-04-01	$[-9.72, 6.76]$% (页 $= 1$)	$[-6.64, 5.89]$% (页 $= 0.80$)	$[-2.66, 4.54]$% (页 $= 0.68$)	$[-3.21, 3.57]$% (段落 $= 0.99$)
设定 *2*:				
2023-04-01 至 2023-07-01	$[-4.58, 6.43]$% (段落 $= 0.56$)	$[-4.23, 5.11]$% (段落 $= 0.73$)	$[-3.36, 3.73]$% (段落 $= 0.89$)	$[-3.12, 3.68]$% (段落 $= 0.75$)
正文规格:				
2023-08-01 至 2023年11月28日	$[1.6, 9.2]$% (段落 $= 0.013$)	$[0.68, 8.3]$% (段落 $= 0.025$)	$[-5.4, -0.54]$% ($p = 0.019$)	$[-4.4, 0.24]$% ($p = 0.075$)

平台响应解释的合理性。

转发链中断检验基于以下直觉：如果处理组参与者正在中断转发链，那么一些原本会在暴露于参与者转发后进行转发的非参与者，将不再这样做。在这种情况下，仍然进行转发的非参与者转发者的构成，会更多地转向那些独立进行转发的用户（例如，直接看到原始上游用户的帖子），而不是通过暴露于参与者转发来进行转发。因此，暴露于处理程度更高的上游用户，其参与者转发者与非参与者转发者的相对排序应发生变化，非参与者在参与者之后进行转发的频率会降低。

我们通过为每位上游用户构建一项指标，将这一想法具体化，该指标衡量参与者转发之后发生非参与者转发的频率。固定一位上游用户i，并令 $\mathscr{O}_i$ 表示i发布的原创帖子集合。对于每条原创帖子$k \in \mathscr{O}_i$，令 $\mathscr{P}_{ik}$ 表示转发k的参与者用户集合， $\mathscr{N}_{ik}$ 表示转发k的非参与者用户集合，并令t_{iku} 表示用户u转发k的时间。我们将排序结果定义为

$$R_i = \frac{\sum_{k \in \mathscr{O}_i} \sum_{u \in \mathscr{N}_{ik}} \sum_{v \in \mathscr{P}_{ik}} \mathbf{1}\{t_{ikv} < t_{iku}\}}{\sum_{k \in \mathscr{O}_i} |\mathscr{N}_{ik}||\mathscr{P}_{ik}|}$$

该指标具有简单的概率解释：对于随机选取的原创帖子k，它表示随机选取的参与者转发发生在同一帖子随机选取的非参与者转发之前的概率。接近一的值表明，非参与者往往在参与者之后转发（这更符合转发级联通过参与者传播的情况）；而接近零的值表明，非参与者往往在参与者之前转发。

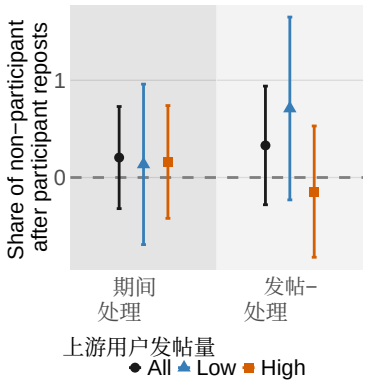


图 S17：对参与者转发和非参与者转发相对排序的间接效应。该图报告了处理组受众占比每增加一个百分点所对应的边际百分比变化。异质性是相对于处理前发帖量而言的。误差条表示 95% RI 置信区间。

我们将R_i 计算为上游用户层面的结果变量，并使用第 S8 节所述的、结合处理前差异调整的相同随机化推断方法，检验更高的暴露是否会改变参与者与非参与者转发者之间的相对顺序。图 S17 显示，在处理期间或处理后时期均不存在统计显著的效应（处理期间，$[-0.32, 0.73]$ 的 95% 置信区间，$p = 0.54$；处理后时期，$[-0.28, 0.94]$ 的 95% 置信区间，$p = 0.34$）。这些无效结果与转发链中断的解释不一致，反而支持这样一种解释：参与者参与度降低会触发平台响应，从而降低上游用户的可见度。

S12 成本效益

本节对我们的干预与传统平台内容审核的成本效益进行了说明性比较。其目的并非精确估算内容审核成本（因为这些成本无法通过公开信息获知），而是根据平台内容审核工作的近似规模及已有记录的有效性，为该干预提供一个基准。

我们根据公开报道的数据估算全球内容审核成本：X 在 2023 年初雇用了约 2,500 名人工审核员。假设每小时工资为 10 美元、每天工作 8 小时，并且每天持续提供审核服务，则全球内容审核预算约为每月 600 万美元。尽管投入了这些资源，已有证据表明，平台仅移除了有限比例的仇恨内容。具体而言，Frances Haugen 在 2021 年披露，Facebook 移除了约 5% 的仇恨言论。

由于我们的实证研究场景是尼日利亚，我们将全球审核成本按尼日利亚平台用户所占比例缩放，以此构建一个基准。尼日利亚约有 300 万用户，全球用户估计约为 3.5 亿，因此尼日利亚约占全球用户的 0.8%。在审核资源与用户数量成比例增长这一简化假设下，推算出的尼日利亚审核预算约为每月 50,000 美元。尽管这一分配未必精确，但在缺乏国家层面的内容审核数据的情况下，它提供了一个透明的基线，

这是因为缺乏国家层面的内容审核数据。

我们的干预每月成本约为 7,500 美元，并与仇恨内容的发布和转发减少约 2.5% 相关。将每月项目成本除以估算的仇恨减少比例后可知，与传统内容审核相比，该干预每降低一个百分点的成本约低 71%。这一比较未考虑作用机制上的重要差异——事后删除与事前行为改变——但表明，定向亲社会信息传递或可成为传统内容审核的一种相对经济高效的补充。