arXiv:2303.03445v1 [cs.IR] 6 Mar 2023

审计方法如何影响我们对 YouTube 推荐系统的理解

Sarmad Chandio、Daniyal Pirwani Dar、Rishab Nithyanand

{sarmad-chandio, rishab-nithyanand}@uiowa.edu
mdar@cs.stonybrook.edu

摘要

对社交媒体网站进行审计所生成的数据，构成了我们理解算法内容推荐系统所呈现偏见的基础。随着世界各地的立法者开始考虑监管驱动在线平台的算法系统，确保这些推断出的偏见准确无误至关重要。然而，正如本文将展示的那样，由于与从特定平台收集数据的审计相关的配置参数十分复杂，要做到这一点并不容易。

我们以 YouTube 为重点，说明要通过审计推断 YouTube 推荐系统的特性，其方法学挑战比人们预想的更大。要获得科学上有效的结果，需要考虑许多方法层面的决策，而每项决策都会带来成本。例如，审计人员在收集算法推荐内容时，是否应使用（获取成本高昂的）已登录 YouTube 账号，以获得更准确的推断？我们探究了这一决策及许多其他决策的影响，并对影响 YouTube 推荐内容的方法选择作出了一些令人惊讶的发现。综合来看，我们的研究提出了一些审计配置折中方案，YouTube 审计人员和研究人员可以借此在不牺牲推断准确性的前提下，从经济和计算两方面降低审计开销。我们还指出了若干会显著影响测量所得推断准确性的配置参数，审计时应予以仔细考虑。

1 引言

审计内容推荐系统正变得愈发重要。随着社交媒体平台及其采用的算法持续对我们的社会政治现实产生越来越大的影响，准确地对其进行审计，出于多种原因已成为一项日益重要的任务。毕竟，这些审计往往聚焦于算法推荐系统，在制定针对在线平台和算法的有效监管措施（whi 2022），以及加深对算法在政治极化（Barberá 2020）和错误信息传播（Hussein、Juneja 和 Mitra 2020）等方面所起作用的认识上，发挥着重要作用，以及其他社会行为。例如，以 YouTube 平台为研究对象，先前的研究揭示了算法推荐系统的若干令人担忧之处，例如容易形成过滤气泡（Tomlein 等人 2021）、推荐不适龄内容（Papadamou 等人 2019）、错误信息（Tomlein 等人 2021；Hussein、Juneja 和 Mitra 2020），甚至极端主义内容（Ribeiro 等人 2020年；Papadamou 等人 2020）。然而，这些研究的结论似乎常常相互矛盾——例如，先前研究表明，YouTube 推荐系统会产生主流化效应（即优先推荐热门内容而非小众内容）（Ledwich 和 Zaitsev 2019），同时也显示它倾向于推荐小众和极端主义内容（Ribeiro 等人 2020年）。在算法审计研究的结果常常相互矛盾的情况下，制定有效的监管措施并深入理解算法对社会的影响颇具挑战。本文（*1*）指出，审计方法是造成此类结果相互矛盾的一个原因；（*2*）提出了减少此类矛盾的方案。

从概念上讲，设计推荐算法审计很简单。由于被审计算法的不透明性，研究人员依赖所谓的“傀儡账号”审计方法（Sandvig 等人）。在这种方法中，审计可以概括为一个简单的三步流程。

创建傀儡账号。创建旨在冒充真实人类用户的傀儡账号或用户画像。其目标是使用自动化工具（通常是网络爬虫），向底层推荐系统提供一组交互方式，使其能够据此学习傀儡账号的某些特征。在 YouTube 的情境下，这可能包括让傀儡账号加载一组视频（称为训练集），为推荐算法学习用户行为和偏好提供基础。

测量推荐树。在此步骤中，傀儡账号执行能够生成推荐内容的种子交互。该组推荐内容构成推荐树的第一层。通常，还会递归地与这些推荐内容本身进行交互，以形成更深层的推荐树。应用于 YouTube 时，此步骤包括向傀儡账号提供一个种子视频，并收集由此产生的所有推荐内容。随后加载与每条

版权所有 © 2021，美国人工智能促进协会（www.aaai.org）。保留所有权利。

推荐内容相关的视频本身，以填充推荐树。

假设检验。最后，给定与具有不同特征的傀儡账号相关联的一个或多个推荐树，我们会检验关于底层推荐算法的假设，并据此作出推断。

在实践中，算法审计颇具挑战性，也可能迫使研究人员作出方法论妥协。先前描述的流程乍看简单，但其中每个步骤都涉及若干常被忽视的关键决策。例如，在进行爬取以构建傀儡账号时，研究人员需要决定将哪些视频纳入傀儡账号训练集、训练集中应包含多少视频，以及选择哪个视频作为种子，等等。每个视频可能对收集到的推荐内容产生何种影响，存在不确定性，这使得这些决策颇具挑战。更复杂的是，即使对上述问题采用严谨可靠的判断依据，坚持方法论严谨性仍会带来高昂的资金和计算成本。

高昂成本导致的方法论妥协。 包括 YouTube 在内的在线平台，会让研究人员难以自动创建开展有效审计所需数量的账号——它们会向试图创建账号的网页自动化工具提供验证码（或直接封锁这些工具），而且通常要求每个账号都绑定经过验证的电话号码。规避这些障碍的成本可能高得令人难以承受，迫使研究人员做出妥协，从而影响推论的有效性。例如，研究人员可能会简单地为每个傀儡账号配备一个独立的浏览器（Cookie），以绕过为每个傀儡账号获取可验证电话号码所带来的困难和高昂成本。然而，这种规避做法往往是基于一种希望：审计所得推论的准确性不会因此受到损害——也就是说，他们假定 YouTube 对待已登录用户的推荐方式，与对待浏览器中存有 YouTube Cookie 的未登录用户相同。

高昂计算成本带来的方法论妥协。此外，当爬虫遇到大量时长一小时（或更长）的视频时，视频爬取会消耗大量计算资源和时间。再加上统计上可靠的假设检验需要收集大量数据，单次审计可能需要耗费数千小时的机器运行时间。这带来了另一个两难问题：为了得出有效结论，是否应承担完整观看每个视频所需的高昂计算成本？是否应遍历推荐树的所有路径？虽然只抽取树中的部分路径进行采样、且不将视频播放至完成度的替代做法更易实行，但目前尚不清楚这些做法是否会影响爬取过程中随后收集到的推荐内容。

简而言之，目前还没有针对 *YouTube* 等平台上的傀儡账号式审计的最佳实践或指南。本文将特别聚焦于 YouTube，并尝试通过回答以下研究问题来填补这一空白。

研究问题1：傀儡账号训练集、推荐种子与推荐之间有什么关系？推荐树？（§3）我们首先研究训练集和种子对其生成的推荐树的影响。具体而言，我们进行了一项实验，使用两种截然不同的种子和训练集的所有组合来训练四组傀儡账号。随后，我们分析这些傀儡账号生成的推荐树，以了解推荐内容如何随着种子和训练集的变化而改变。

研究问题2。审计过程中降低现金成本会产生什么影响？（§4）我们研究 YouTube 审计人员最常采用的一种节省成本措施所带来的后果：不为每个傀儡账号使用真实的 YouTube 账户，而是依赖浏览器 Cookie 将傀儡账号的身份泄露给 YouTube。我们通过比较四组傀儡账号生成的推荐树来开展这项分析，这些账号组反映了审计研究中常见的做法。除了维护 YouTube 账户“状态”的方式不同之外，这些傀儡账号组在其他方面完全相同。

研究问题3。审计过程中降低计算成本会产生什么影响？（§5）最后，我们探讨与降低计算成本相关的权衡所带来的后果。我们尤其关注每个傀儡账号训练视频所花费的时间，以及推荐树探索的深度和广度所产生的影响。为此，我们训练多组傀儡账号，让它们以不同的播放完成度“观看”视频，并测量其收集到的推荐树之间的差异。随后，我们研究本研究收集的所有推荐树中抽取的节点特征，以确定这些特征是否会因节点在树中的位置而异。

2 方法论

我们总共进行了十一个实验（参见 表1），在这些实验中，我们改变了特定的审计参数。我们利用这些审计中收集到的推荐树，来确定参数变化的影响。本节将介绍我们选择的审计配置和分析方法论。

2.1 配置参数

傀儡账号训练集 (T_{niche}, T_{main})。在全部 11 项实验中，我们首先使用 T_{niche} 或 T_{main}中包含的视频来训练傀儡账号。此前的一项研究（Papadamou 等人（2021））表明，只需 22 个视频就足以个性化 YouTube 视频推荐，而我们决定为每个训练集使用 32 个视频[1] 进行训练。小众训练集（T_{niche}）。 T_{niche} 中的视频经过人工筛选，旨在体现边缘化（例如阴谋论）且相对不受欢迎（观看次数较少）的内容。该训练集中的视频选自多个边缘化的 subreddit，例如*r/climateskeptics* 和 *r/theworldisflat*等。该训练集中的视频支持与相应 subreddit 主题相关的立场（例如支持地平说）。 T_{niche} 中的视频平均获得了 2.5 万次观看。

[1]每个训练集中的视频完整列表可在以下网址查看：
https://osf.io/3j5u8/?view_仅=c2e21b99dbc3470e86bc9e904b39e6d3

主流训练集（T_{main}）。T_{main} 中的每个视频都经过筛选，以涵盖与对应小众视频相同的主题，但这些视频来自 YouTube 搜索该主题的结果（例如“揭穿地平说”）。搜索结果中最受欢迎的视频（按观看次数计）被添加到T_{main} 中。T_{main} 中的视频代表“主流”观点，并支持这些主题中广受认可的世界观（例如，地球不是平的）。T_{main} 中的视频平均获得了 590 万次观看。

这种训练集构建方法为推荐算法提供了截然不同的两种输入，从而可以衡量训练集对推荐内容的任何效应。

推荐树种子(s_{niche} 和 s_{main}**)**。种子视频是收集推荐树的起点（即推荐树的根节点）。与训练集类似，我们使用了两个种子中的一个（s_{niche}和 s_{main}），并根据直觉选择它们，认为它们会对推荐树产生截然不同的影响。实验中使用的 s_{niche} 视频是一段关于非法移民主题的边缘政治视频，观看次数为 7.1K；而 s_{main} 视频则是一段非常热门的主流视频，聚焦于“Slapgate”（sla 2022），观看次数超过 3.8M。选择种子视频的主题时，我们有意确保：(1) 它们与T_{main} 或 T_{niche} 中的任何视频都不重叠，以便区分训练集与种子各自造成的效应；(2) 它们彼此之间也不重叠，以最大限度地扩大其推荐树之间任何可衡量的差异。如果观察到由 s_{niche}和 s_{main} 生成的推荐树之间不存在差异，则表明种子对观察到的推荐内容影响甚微。

账户状态 (A_{full}、 A_{cookies} 和 A_{clear}**)**。为了更好地了解在不同的 YouTube 账户假设下进行审计时，推荐算法是否会以不同方式运行，我们采用四种不同的账户假设来收集推荐树。(1) A_{full} 表示爬虫在开始训练和收集推荐内容之前，已登录新创建的 YouTube 账户进行审计。这代表了理想情况，即每个傀儡账号都有自己的全新 YouTube 账户。(2) A_{cookies} 表示爬虫未登录，但在整个爬取过程中始终保留浏览器中的 YouTube Cookie。这代表了审计文献中最常见的爬取方式。(3)A_{clear} 表示爬虫在登录状态下进行爬取，并在同一账户用于另一次爬取之前清除其观看历史。这种方法旨在让不同傀儡账号复用账户。

观看时长 ($W_{100\text{pc}}, W_{50\text{pc}}, W_{25\text{pc}}, W_{10\text{pc}}$**)**。训练用傀儡账号可能会耗费大量计算资源，因为训练集中通常包含时长较长的视频。为了了解是否需要完整观看训练集中的视频，我们从四个爬虫收集推荐树。这些爬虫的配置完全相同，唯一的区别在于：它们在切换到下一个视频之前，会将每个训练视频播放到不同的完成度。$W_{100\text{pc}}$、$W_{50\text{pc}}$、$W_{25\text{pc}}$和 $W_{10\text{pc}}$ 分别将视频播放至100%、50%、25%和10%的完成度。

交互方式 (I_{get}，I_{click}**)**。由于可靠性方面的困难，要让爬虫实际点击超链接是一项颇具挑战性的任务。一个常用的替代方法是解析 DOM 来获取链接，再让浏览器加载目标链接。遗憾的是，没有实际点击这一行为也会被常见的机器人检测工具视为特征，并可能导致服务器端采取差异化处理（Singh等，2017年；Khattak等，2016年；Ahmad等，2020年；Jueckstock等，2021年）。我们开展了一项实验，以了解点击推荐视频是否会影响后续的推荐内容。I_{click} 表示一种审计方式：在爬取推荐树的过程中，每个爬虫都会实际点击视频以加载视频。I_{get} 表示另一种审计方式：每个爬虫只从 DOM 获取视频的 URL，并指示浏览器加载该 URL。

探索广度 (P_{left}、 P_{right}**)**。YouTube 的推荐内容是动态加载的，用户向下滚动页面时，推荐选项通常还会继续出现。这会增加推荐树每一层的宽度。在我们的试点测试中，我们发现每个视频的推荐内容数量至少为 40 条（许多情况下要多得多）。我们分析推荐树爬取过程中出现在推荐列表顶部的视频（即树中由 P_{left} 表示的最左侧路径），以及出现在推荐列表底部的视频（即树中由 P_{right}表示的最右侧路径）。

探索深度 ($D_{\text{top}}, D_{\text{bottom}}$**)**。最后，我们考察了对推荐树的测量特征进行深度爬取的必要性。为此，我们分析了仅加载种子视频后观察到的所有视频的特征（即推荐树中由 D_{top} 表示的 1st 层级），并将其与在树的 10th 层级观察到的所有视频的特征（即我们 采集的、由 D_{bottom}表示的推荐树底部）进行比较。

2.2 数据采集

最大限度减少潜在混杂变量的影响。推荐树会受到大量变量的影响，其中一些处于研究人员的控制之下（例如我们的配置参数），而另一些则不受控制。在本研究中，我们通过以下方法，尽最大努力减少这些潜在影响。

考虑搜索索引的更新。由于 YouTube 上不断有大量新内容产生，搜索索引和推荐候选列表也在持续变化。因此，在相隔很长时间的两个时间段内收集推荐的爬取结果，可能无法进行比较，因为可供推荐的选项会有很大差异。我们通过同步每项实验中的爬取来减轻这类影响：对于每个使用一种配置来收集推荐树的爬虫，都有另一个同步运行的爬虫使用备用配置来收集用于比较的推荐树。这种同步在节点级别进行——也就是说，我们确保每棵树在其各自的推荐树中到达完全相同的节点位置，且与对应的另一棵树相差不超过

问题	参数	配置	#树	#视频
RQ1	训练集	T_{main}, T_{niche}	16	32K
	种子视频	s_{main}, s_{niche}	16	32K
RQ2	账号	A_{full}, A_{cookies}	8	14K
		A_{full}, A_{clear}	8	13K
RQ3	观看时长	$W_{100\text{pc}}$, $W_{50\text{pc}}$	8	15K
		$W_{50\text{pc}}$, $W_{25\text{pc}}$	8	15K
		$W_{25\text{pc}}$, $W_{10\text{pc}}$	8	15K
	互动	I_{get}, I_{click}	8	16K
	广度	P_{left}, P_{right}	all*	69K
	深度	D_{top}, D_{bottom}	all*	35K

*分析这些参数时使用了所有树中的数据。

表1：**实验**。“参数”列表示每项实验中被修改的参数，“配置”列表示实验中为该参数指定的值。“# 树”列表示为分析采集的推荐树总数，“# 视频”列表示在这些树中观察到的推荐内容总数。

几秒钟。因此，使用同一参数的不同备用配置收集的树具有可比性。

考虑分布式基础设施和地理位置效应。正如先前研究所示（Hannak 等，2013），网络服务器可能分布在广阔区域，不同地点或数据中心的服务器可能具有不一致的搜索索引，或提供地理位置特定推荐。为减轻这些因素对所采集树的影响，我们所有的数据采集实验都在同一地点进行，并为 YouTube 使用静态域名系统条目，以确保所有内容请求以及与平台的互动至少由同一区域内的网络服务器处理。

纳入 *A/B* 测试因素。众所周知，平台在测试新功能或算法更新时，会对用户进行 A-B 测试（Facebook，2022）。为尽力减轻此类测试的影响，我们针对研究中测试的每个参数，收集至少八次完全相同且同步的爬取数据。

收集推荐树。傀儡账号经过训练并获得种子视频后，我们便开始探索推荐树。遗憾的是，由于对每棵待收集的树所遍历的每条路径，以及每种待测试配置，都需要一个傀儡账号，因此无法完整探索推荐树。这是因为此前观看过的视频会影响后续推荐，所以一个傀儡账号只能沿推荐树进行单向（向下）遍历。此外，我们会为每个视频至少收集 40 条推荐内容。因此，深度为 n 的推荐树至少有 40^n 条从根节点到叶节点的路径，每条路径都需要一个独立的傀儡账号。在遍历树的过程中，我们探索了五条独特的路径——最左侧路径（由每个节点的第一条推荐内容组成）、最右侧路径（由每个节点的最后一条推荐内容

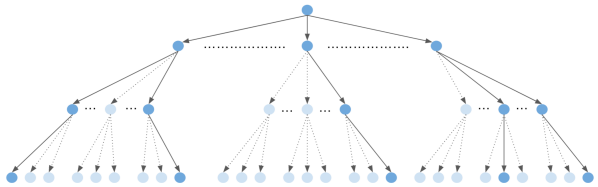

图 1：基于 5 个傀儡账号生成的推荐树。扁平箭头显示每个傀儡账号各自唯一的推荐路径。从种子开始，每个节点代表由其父节点推荐的一个视频。

组成），以及从中间预先选出的三条路径（按齐夫权重抽样，以体现对推荐列表中排名靠前视频的偏好）。我们同时探索每条路径，并为每条路径分配一个独立但训练、配置和种子设置均相同的傀儡账号，将探索深度设为 10，并记录沿途的所有推荐内容。我们将这些路径和观察结果拼接起来，获得完整推荐树的一个子集，并据此开展分析。对于每种参数配置，我们至少收集四棵这样的树，同时确保与配置交替的审计保持同步。图 1 展示了这样一棵树的示例。在图中，每条路径代表构成傀儡账号的视频集合，节点代表视频，任意两个节点（$parent$、 $child$）之间的有向边表示，在与 $parent$直接互动后，推荐了 $child$ 。这棵树的根节点代表用于生成第一组推荐内容的种子视频。

2.3 推荐树特征

在我们的分析中，我们重点研究推荐树中观察到的视频的受欢迎程度、频道多样性和主题。我们选择这些参数，是因为平台审计通常会关注这些指标（或其变体），以识别推荐算法造成的回音室效应、兔子洞效应或主流化效应。

推荐内容的受欢迎程度。我们使用视频观看量作为代理指标来衡量推荐内容的受欢迎程度，从而了解算法倾向于推荐小众内容还是主流内容。我们记录每个节点所观察到的推荐视频观看量分布。如果一棵推荐树在每个节点上的视频大多受欢迎程度较低，则表明相关傀儡账号配置倾向于推荐小众内容。反之，如果一棵树在每个节点上的视频大多受欢迎程度较高，则表明相关傀儡账号配置倾向于推荐主流内容。如果两种配置生成的树在组内和组间差异方面存在显著差异，则表明其中一种配置比另一种更倾向于推荐主流（受欢迎）内容。**推荐内容的频道多样性**。每个关于某个主题的视频都体现了上传该视频的频道的观点。因此，我们将树中频道的多样性作为推荐内容所提供观点范围的代理指标。我们通过记录以下内容来衡量这一点：

每个节点上推荐内容中各频道的熵。若一棵推荐树在每个节点上的推荐内容熵都较高，则表明推荐多样性较高，并意味着不存在兔子洞效应。如果两种配置生成的树在组内和组间差异方面存在显著差异，则表明其中一种配置比另一种更倾向于呈现多样性较低的推荐内容。

推荐内容的语义相似度。我们提取树中观察到的每个视频所关联的标题和文本描述。随后，我们对这些文本执行标准的自然语言处理预处理操作（即分词、移除停用词和URL、移除在数据集中出现频率超过 50% 的词元，以及词形还原）。我们将树中每个节点观察到的所有视频的这些处理后文本合并，并将其用作该节点推荐内容中观察到的主题的表征。为了对任意两个节点进行语义比较，我们使用 SpaCy 文档相似度方法（SpaCy 2022年），该方法采用词袋方法，计算文本中所有词向量平均值之间的余弦相似度。尽管此方法存在局限，无法捕捉内容的极性（例如，反疫苗文本与驳斥反疫苗文本的论述会有很高的相似度），但我们的人工验证发现，该方法能够捕捉文本中主题的相似性。我们还测试了其他用于衡量相似度的语义方法（潜在语义索引（Rosario 2000年））和词汇方法（潜在狄利克雷分配（Blei、Ng 和 Jordan 2003年）），但发现这些方法在捕捉数据集的主题相似度方面表现不佳。我们进行了一项试点研究，随机抽取由 SpaCy 的 docsim、LDA 和 LSA 判定为高度相似的视频对，并人工评估其文本以验证相似性。根据这项研究，我们认定 SpaCy 的 docsim 效果最佳。由两种配置生成的树，其组内差异和组间差异若存在显著差异，则表明其中一种配置产生的推荐主题存在可测量的差别。

2.4 审计配置的比较

我们的每项实验都会产生两组推荐树——每组对应一种待测试的审计参数配置（例如， A_{full} 和 A_{cookies}）。每组中的树彼此同步收集。对于这两组推荐树，我们会沿着§2.3所述的三个维度，计算组间差异（例如， A_{full} 与其同步的 A_{cookies} 树之间的差异）和组内差异（例如，两棵同步的 A_{full} （或A_{cookies}）树之间的差异）。下文将介绍这一过程。

推荐树节点的记录特征。令 n_{ij} 表示位于推荐树路径 P_i 上、深度为 j 的已遍历节点（即已观看的视频），并令 r_{ijk} 表示在n_{ij}观察到的 k^{th} 推荐。在每个节点 n_{ij} ，我们记录：(1) 受欢迎程度标量值 $pop(n_{ij}) = \mu(views(r_{ij,1}) \dots views(r_{ij,40}))$，表示该节点所有已观察到的推荐视频的观看次数；(2) 频道熵标量值$div(n_{ij}) = entropy(channel(r_{ij,1}) \dots, channel(r_{ij,40}))$ 表示该节点推荐视频中频道的多样性；以及 (3) 文档向量 $doc(n_{ij}) = docvec(desc(r_{ij,1}) \dots, desc(r_{ij,40}))$，它表示与该节点所有推荐视频的视频描述相关联的文档向量。

比较推荐树的特征。给定两棵推荐树 T 和 T'，我们以节点位置为单位计算其特征差异——也就是说，我们针对 T 和 T'中的每个节点位置，分别计算流行度向量、频道熵和文档向量的差异。计算方式如下：

$$\delta_{pop}(T, T') = mean([\forall i, \forall j : pop(n_{ij}) - pop(n'_{ij})])$$

$$\delta_{div}(T, T') = mean([\forall i, \forall j : div(n_{ij}) - div(n'_{ij})])$$

$$\delta_{sem}(T, T') = mean([\forall i, \forall j : docsim(docvec(n_{ij}) \; docvec(n'_{ij}))])$$

这些值有效地捕捉了 T 与 T'之间节点到节点差异的均值。由于我们研究中收集的所有树都遍历了推荐树中的同一组路径，因此可以进行这种节点到节点的比较。在树比较中保持这种节点位置依赖性非常重要，因为它能够处理节点在推荐树中的位置可能导致的特征差异。例如，将来自 T 、深度为=1处的顶层推荐与来自 T' 、深度为=10 处的 40th 推荐进行比较，可能会将由推荐排名变化引起的树特征差异错误地归因于审计配置变化的影响。

计算组内差异和组间差异。给定两个审计配置 $\mathcal{C}$ 和 $\mathcal{C}'$，它们分别生成树集合 $\mathcal{T}$ 和 $\mathcal{T}'$，我们计算：(1)组内差异，即在 $\mathcal{T}$ 和 $\mathcal{T}'$内部的树之间观测到的特征差异的分布；以及(2)组间差异，即在 $\mathcal{T}$ 和 $\mathcal{T}'$之间的树之间观测到的差异的分布。这些差异表示为：

$$\Delta_x^{within}(\mathcal{T}) = [\forall (T_i, T_j) \in (\mathcal{T} \times \mathcal{T}) : \delta_x(T_i, T_j)]$$

$$\Delta_x^{across}(\mathcal{T}, \mathcal{T}') = [\forall (T_i, T_j) \in (\mathcal{T} \times \mathcal{T}') : \delta_x(T_i, T_j)]$$

$$\forall x \in \{pop, div, sem\}$$

在采用相同审计配置生成的所有树上计算出的组内差异，使我们能够建立一个基线，用于衡量由审计人员无法控制的因素（例如，概率性推荐算法、A/B 测试等）造成的特征变化。组间差异则展示了审计配置以及 外部因素变化所造成的差异。

量化审计参数配置的影响。给定与配置（$\mathcal{C}$、 $\mathcal{C}'$）相关的分布 Δ_x^{within} 和 Δ_x^{across}，我们使用包含 100 万个样本的自助法（DiCiccio 和 Efron，1996；Efron，1987），为组内差异和组间差异的均值计算 95%置信区间。我们还使用这些自助法样本，为效应量计算 95%置信区间——即自助法样本中组内差异与组间差异之间的差值。令 $[CI_{lower}, CI_{upper}]$ 为效应量的 N%置信区间。

参数		视频受欢迎度（观看次数，以百万计）				频道多样性（熵，以比特计）			内容语义（相似度分数）			
固定	变化	μ_{views}	效应（95%置信区间）	效应（99%置信区间）	μ_{effect}	μ_{entropy}	效应（95%置信区间）	效应（99%置信区间）	μ_{effect}	效应（95%置信区间）	效应（99%置信区间）	μ_{effect}
s_{main}	T_{main} T_{niche}	7.15 4.94	**[0.34, 1.33]**	**[0.19, 1.49]**	0.84	3.63 3.49	[-0.16, 0.17]	[-0.21, 0.22]	0.00	[-0.02, -0.00]	[-0.02, 0.00]	-0.01
s_{niche}	T_{main} $T_{\text{生态位}}$	4.32 1.80	**[1.46, 2.19]**	**[1.35, 2.30]**	1.82	3.38 3.26	[-0.27, 0.19]	[-0.34, 0.26]	-0.04	**[-0.04, -0.02]**	**[-0.05, -0.01]**	-0.03
T_{main}	s_{main} s_{niche}	10.71 7.78	**[0.73, 2.31]**	**[0.50, 2.56]**	1.51	3.17 2.97	[-0.14, 0.14]	[-0.18, 0.18]	0.00	**[-0.04, -0.02]**	**[-0.04, -0.02]**	-0.03
T_{niche}	s_{main} $s_{\text{生态位}}$	4.93 1.72	**[2.68, 3.05]**	**[2.62, 3.11]**	2.87	4.02 3.44	**[0.12, 0.45]**	**[0.07, 0.51]**	0.28	**[-0.05, -0.02]**	**[-0.05, -0.01]**	-0.03

表2：改变训练集（前两行）和随机种子（后两行）所造成的影响。各列分别表示每组在某一特征上观察到的节点平均值、测得效应量的95%和99%置信区间（即组内差异与组间差异之差；参见 §2.4），以及平均效应量。粗体数值表示在相应置信水平下具有统计显著性的效应量。

当且仅当 ($CI_{lower} \leq CI_{upper} < 0$) 或 ($CI_{upper} \geq CI_{lower} > 0$) 时，我们认为该效应在此置信水平下具有统计显著性——即，当 *N%* 的自助法样本中观测到的效应量具有相同极性时。在我们的工作中，我们报告效应量的 95% 和 99%置信区间。我们还将平均效应量报告为自助法样本中观测到的所有效应量的均值。

3 训练集和随机种子

实验设置。 我们的目标是测量训练集和随机种子对审计生成的推荐树特征的影响。为此，我们从四种不同的审计配置中收集了32棵推荐树：使用 T_{main} 和 s_{main}、 T_{main} 和 s_{niche}、 T_{niche}和 s_{main}，以及 T_{niche} 和 s_{niche}进行的审计各生成八棵树。我们将每组树各分成两组，每组四棵，分别记为($\mathcal{T}_{\text{main, main}}$,$\mathcal{T}'_{\text{main, main}}$)、($\mathcal{T}_{\text{main, niche}}$,$\mathcal{T}'_{\text{main, niche}}$)、($\mathcal{T}_{\text{niche, main}}$,$\mathcal{T}'_{\text{niche, main}}$)和($\mathcal{T}_{\text{niche, niche}}$,$\mathcal{T}'_{\text{niche, niche}}$)。为便于准确进行组内和组间比较，这些树是在同步条件下收集的（参见 §2.2）（参见§2.4）。将每组八棵树拆分为两组、每组四棵，可以避免重复使用树来检验多个假设。

衡量训练集变化的影响。为了揭示审计中使用的训练集对推荐树特征的影响，我们计算了两项分析中所得组内差异、组间差异和效应量（参见§2.4）的均值以及 95% 和 99% 置信区间：(1) 比较 $\mathcal{T}_{\text{main,main}}$ 与 $\mathcal{T}_{\text{niche,main}}$ ——即使用相同的主流种子，同时改变训练集；(2) 比较 $\mathcal{T}_{\text{main,niche}}$ 与 $\mathcal{T}_{\text{niche,niche}}$ ——即使用相同的生态位种子，同时改变训练集。

衡量种子变化的影响。我们对以下分析重复采用相同的方法：(1) 比较 $\mathcal{T}'_{\text{main,main}}$ 与 $\mathcal{T}'_{\text{main,niche}}$ ——即在使用侧重争议性话题的主流训练集时改变种子；(2) 比较 $\mathcal{T}'_{\text{生态位,主流}}$ 与 $\mathcal{T}'_{\text{生态位,生态位}}$ ——即在保持边缘且涉及争议性话题的训练集不变时改变种子。

结果。 我们的结果汇总于表2。总体而言，我们发现，改变训练集或种子的特征总是会影响审计中观察到的视频受欢迎程度。然而，频道多样性和语义并非如此。具体而言，我们的分析得出了以下见解。

推荐内容中似乎存在强有力的“近因偏差”证据。关注表2的最后两行可以看到，将种子从小众视频改为主流视频所产生的效应几乎总是具有统计显著性且幅度较大；唯一的例外是使用 T_{main}进行训练时记录频道多样性。对推荐视频的受欢迎程度和熵产生的（显著）效应，也高于改变训练集（前两行）所观察到的效应。改变种子所产生的最显著效应体现在“受欢迎程度”这一维度：将种子视频从小众视频切换为主流视频后，经过 T_{main} 和 T_{niche}训练的视频推荐内容，其观看次数平均分别增加了151万和287万。我们仅发现推荐视频的语义发生了边际性（但具有显著性）的变化——也就是说，将种子从主流视频切换为小众视频后，推荐内容的语义相似度会降低1%至5%。这表明，无论使用何种训练集，种子的选择都可能大幅改变推荐树的特征以及审计推断。将这一发现外推可知，最新视频将对未来的推荐内容产生超出比例的影响。

频道多样性并不总是取决于训练集和随机种子。我们的分析表明，选择不同的训练集和随机种子，频道多样性基本不会受到影响。只有一种例外：在对某个 T_{niche} 训练集进行审计时，改变随机种子会产生影响（参见表2第四行）。我们看到，从 s_{main} 切换到 s_{niche}，会使每个节点的频道多样性平均降低0.28（以比特为单位的熵）。虽然这一发现似乎印证了有关算法具有兔子洞效应倾向的说法，但需要注意的是，只有在审计过程中仅与边缘内容互动时（训练集中仅包含边缘内容，且使用相应随机种子），才会出现这种下降。鉴于只要发生任何其他互动，这种效应就会消失，因此，这一发现可能是因为讨论该生态位内容的创作者数量较少。

要点。 综合来看，这些结果为理解 YouTube 的推荐系统及其

参数	视频受欢迎度（观看次数，以百万计）				频道多样性（熵，以比特计）				内容语义（相似度分数）			
	μ_{views}	效应（95%置信区间）	效应（99%置信区间）	μ_{effect}	μ_{entropy}	效应（95%置信区间）	效应（99%置信区间）	μ_{effect}	效应（95%置信区间）	效应（99%置信区间）	μ_{effect}
A_{full} A_{Cookie}	9.20 7.72	[-0.90, 0.99]	[-1.20, 1.29]	0.05	3.36 3.57	[-0.11, 0.31]	[-0.18, 0.36]	-0.01	[-0.02, -0.00]	[-0.03, 0.00]	-0.01
A_{clear} A_{full}	12.34 8.47	**[1.82, 3.46]**	**[1.54, 3.70]**	2.65	3.55 2.86	[-0.03, 0.53]	[-0.12, 0.62]	0.26	**[-0.05, -0.01]**	**[-0.05, -0.01]**	**-0.03**

表 3：更改登录状态（第 1 行）和清除观看历史（第 2 行）所造成的影响。各列分别表示每组在某项特征上的节点平均值、测得效应量的 95% 和 99% 置信区间（即组内差异与组间差异之差；参见 §2.4），以及平均效应量。粗体数值表示在相应置信水平下具有统计显著性的效应量。

审计提供了不同的观点。研究人员不仅需要特别关注训练和播种，还必须明白，他们对推荐视频的测量结果，在很大程度上取决于傀儡账号已经遍历过的最近节点。具体而言，近因偏差似乎会导致单个视频压过大量先前视频的影响，从而影响审计得出的最终推断。总体而言，我们建议审计推断（例如，是否存在主流化效应）应以以下因素为条件：（1）训练集和种子的具体特征；（2）从推荐树中选择节点所采用的具体策略。

4 节省成本的配置

实验设置。在本节中，我们重点了解常用的傀儡账号管理策略对由其生成的推荐树的影响。

衡量基于 *Cookie* 的傀儡账号的有效性。为了了解基于 Cookie 的傀儡账号与真实账号之间的差异，我们分别为 $\mathcal{T}_{\text{full}}$ 和 $\mathcal{T}_{\text{cookies}}$ 收集了四棵推荐树。这两组的所有参数配置均保持一致，唯一的区别是 $\mathcal{T}_{\text{full}}$ 使用了已登录的个人资料，而 $\mathcal{T}_{\text{cookies}}$ 未登录，但保留了 YouTube Cookie。$\mathcal{T}_{\text{full}}$ 和 $\mathcal{T}_{\text{cookies}}$均使用了（$T_{\text{main}}$， s_{main}）训练集和种子。

衡量清除账号历史记录的有效性。为了验证清除账号历史记录是否确实能够清除观看历史记录效应（即使删除观看历史记录后，用户仍会获得类似的推荐内容），我们分别为 $\mathcal{T}'_{\text{full}}$ 和 $\mathcal{T}_{\text{clear}}$ 收集了四棵推荐树。$\mathcal{T}'_{\text{full}}$ 和 $\mathcal{T}_{\text{clear}}$ 均使用了已登录的个人资料，并使用了（T_{main}， s_{main}）训练集和种子。不过，在基于种子 s_{main}收集推荐内容之前，$\mathcal{T}_{\text{clear}}$ 的观看历史记录已被删除。

为了深入了解不同账号管理策略所产生的可测量效应，我们计算了组内差异、组间差异和效应量对应的均值、95%和99%置信区间。

结果。结果汇总于表 3。我们的分析得出了两个明确结论。

审计并不需要为每个傀儡账号创建新的账号。首先，我们关注了在使用已登录 YouTube 账户（$\mathcal{T}_{\text{full}}$）的傀儡账号与仅保留其浏览器 Cookie（$\mathcal{T}_{\text{cookies}}$）的傀儡账号之间切换所产生的影响。我们发现，两者的推荐内容在任何测量特征上均不存在显著差异。这带来了显著的节省成本机会，因为傀儡账号可以与浏览器实例关联，而无需克服自动化账号创建和电话号码验证所带来的障碍。

通过清除历史记录实现账号复用的可能性 与配置和同步方式完全相同的傀儡账号相比，$\mathcal{T}_{\text{full}}$ 傀儡账号的受欢迎程度和内容语义存在显著差异。这表明，清除历史记录后，$\mathcal{T}_{\text{clear}}$ 可能已清除训练阶段形成的受欢迎程度语境和主题语境，而这些语境仍由 $\mathcal{T}_{\text{full}}$ 保留。简而言之，清除账号历史记录或许能让账号在大规模研究中复用，尤其适用于测量受欢迎程度和内容语义的研究（例如，量化主流化和兔子洞效应的审计）。不过，我们并未声称清除观看历史等同于获得一个全新账号（全新账号意味着 Google 在后端未存储该个人资料的任何数据，而我们并未对此进行检查）。

要点。这些发现为审计人员节省账号创建和维护所涉及的巨额成本提供了机会。我们表明，能够保留 YouTube Cookie 的浏览器与 YouTube 账户同样有效。此外，账户重复使用（清除历史记录后）也是审计人员研究平台受欢迎程度和内容语义的可行选择。

5 计算折衷

实验设置。本节分析为节省计算资源而可能作出的三种折衷所产生的影响：（1）在推荐树中仅观看每个视频预先确定的一部分；（2）使用 Selenium 的 driver.get(URL) 方法，而非自动执行对推荐视频的用户点击；以及（3）进行低深度和窄广度审计。

衡量视频观看时长的影响。为回答审计是否需要将视频“观看”至播放完成度这一问题，我们收集并分析了四棵推荐树，其中审计将所有视频“观看”至播放完成度（$\mathcal{T}_{\text{w=100}}$）；八棵推荐树，其中审计仅将视频“观看”至其总时长的 50%（$\mathcal{T}_{\text{w=50}}, \mathcal{T}'_{\text{w=50}}$）；以及四棵推荐树，其中审计仅将视频“观看”至其总时长的 25%（$\mathcal{T}_{\text{w=25}}$）。两组审计均使用了（$T_{\text{main}}, s_{\text{main}}$）训练集和种子。

衡量互动机制的影响。我们收集了四棵推荐树，其中审计实际定位并点击了推荐

参数	视频受欢迎度（观看次数，以百万计）				频道多样性（熵，以比特计）				内容语义（相似度分数）			
	μ_{views}	效应（95%置信区间）	效应（99%置信区间）	μ_{effect}	μ_{entropy}	效应（95%置信区间）	效应（99%置信区间）	μ_{effect}	效应（95%置信区间）	效应（99%置信区间）	μ_{effect}
$W_{100\text{pc}}$ $W_{50\text{pc}}$	7.69 7.84	[-0.86, 0.28]	[-1.04, 0.46]	-0.29	3.74 3.51	[-0.23, 0.22]	[-0.30, 0.29]	0.00	[-0.03, -0.00]	[-0.03, 0.00]	-0.01
$W_{50\text{pc}}$ $W_{25\text{pc}}$	12.13 9.55	[-3.52, 1.53]	[-4.28, 2.34]	-1.03	3.21 3.60	[-0.41, 0.15]	[-0.50, 0.25]	-0.13	[-0.03, 0.00]	[-0.03, 0.01]	-0.01
$W_{25\text{pc}}$ $W_{10\text{pc}}$	14.11 13.59	[-2.10, 0.43]	[-2.51, 0.83]	-0.85	3.61 3.47	[-0.56, 0.12]	[-0.67, 0.24]	-0.22	[-0.03, 0.00]	[-0.04, 0.01]	-0.01
I_{click} I_{get}	7.62 6.93	[-0.59, 0.64]	[-0.79, 0.82]	0.02	3.79 3.88	[-0.20, 0.11]	[-0.24, 0.16]	-0.04	[-0.02, 0.00]	[-0.02, 0.00]	-0.01
P_{left} $P_{\text{右}}$	8.33 7.47	[-0.65, 0.98]	[-0.91, 1.25]	0.16	3.72 3.33	[-0.03, 0.20]	[-0.07, 0.24]	0.08	**[-0.03, -0.02]**	**[-0.03, -0.02]**	-0.02
D_{top} D_{bottom}	13.73 5.97	**[5.04, 6.67]**	**[4.78, 6.92]**	5.86	4.59 3.12	**[1.05, 1.24]**	**[1.02, 1.26]**	1.14	**[-0.02, -0.01]**	**[-0.03, -0.01]**	-0.02

表 4：因改变视频观看时长（第 1–3 行）、交互机制（第 4 行）、推荐选择策略（第 5 行）和爬取深度（最后一行）而产生的影响。各列分别表示在每组中观察到的特定特征的节点平均值、测得效应量的 95% 和 99% 置信区间，以及平均效应量。粗体数值表示在相应置信水平下具有统计显著性的效应量。

视频链接 ($\mathcal{T}_{\text{click}}$)，以及四棵推荐树，其中审计仅识别推荐视频的 URL，并使用 driver.get(URL) 命令获取视频 ($\mathcal{T}_{\text{get}}$)。两组审计使用了 (T_{main}, s_{main}) 训练集和种子。

衡量爬取广度和深度的影响。我们分析了本研究中收集的全部 96 棵推荐树中最左侧和最右侧路径的特征（$\mathcal{T}_{\text{left}}$和 $\mathcal{T}_{\text{right}}$）。这些路径分别对应于在每个视频下仅点击顶部和底部推荐内容所获得的路径。我们还分析了本研究中获得的全部 96 棵树在深度 1 和 10 处观察到的推荐内容特征（$\mathcal{T}_{\text{top}}$ 和 $\mathcal{T}_{\text{bottom}}$）。

与之前一样，在每项分析中，我们都会计算均值、95% 和 99% 置信区间，以及组内差异、组间差异和效应量。

结果。结果如表 4 所示。值得注意的是，除爬取深度不同的配置外，我们所做的更改均未导致测得的推荐特征出现具有统计显著性的差异。这对审计人员有几项重要启示。

视频不必播放至结束。在我们所有改变视频观看时长比例的审计配置中，推荐视频特征的变化与审计配置的观看比例之间均不存在统计关系。这一发现令人意外，表明观看视频 10% 与观看 100% 对后续推荐内容的影响程度并无差异。进一步调查后，我们发现，有证据表明 YouTube 只要求观看 30 秒，就会将视频计为一次“观看”（Parsons 2017年；Funk 2020年）。根据这些先前的发现，我们推测，这一 30 秒的观看阈值也用于判定视频是否应影响后续推荐内容。由于推荐树中的视频时长远超 300 秒（许多视频长达 20 至 60 分钟），即使只观看视频的 10%，也会被计为一次“观看”。这一发现表明，视频不必观看到特定的播放完成度比例，而是只需达到固定的观看时长阈值，这为未来的审计人员提供了一种有望节省计算成本（且不影响准确性）的途径。

无需自动点击推荐视频。我们的分析显示，在 $\mathcal{T}_{\text{get}}$ 和 $\mathcal{T}_{\text{click}}$中观察到的推荐树特征之间，任何差异都不具有统计显著性。这表明，无需使用浏览器自动化工具（例如 Selenium WebDriver 的动作链（Seleniu m 2022））来显式点击视频链接。在不影响审计推断准确性的前提下，这使审计人员可以用一种简单可靠的方法取代原先的方法：后者需要较高的编程开销、计算成本高且不可靠，而新方法只需编程控制浏览器，从 DOM 中获取指定 URL 即可访问后续推荐内容。

爬取深度会影响推荐特征。 我们对爬取深度影响的分析显示，推荐树的所有特征均具有统计显著性。具体而言，我们注意到，推荐树顶部的节点通常比树底部的推荐内容更受欢迎、更多样化，且语义相似度更低。这一发现再次展示了强烈的近因偏差影响推荐内容的可能性。有趣的是，我们并未发现推荐排名最高和最低的视频之间存在具有统计显著性的差异——这表明审计人员需要特别关注其爬取的深度。

要点。 我们的分析为研究人员带来了两项显著的计算成本节省。具体而言，我们发现无需将视频播放至结束，并且点击视频与仅“获取”视频所对应的 URL 会产生相同的结果，这降低了审计所需的计算和工程开销。此外，我们的工作还凸显出，推荐树的不同深度可能产生不同的推荐特征。为考虑这些效应，任何基于审计得出的推论都必须以所使用树的深度为条件。

6 相关研究

YouTube 推荐系统的审计。 本文的灵感来自近期大量涌现的 YouTube 审计研究，而这些研究往往得出了相互矛盾的结果。例如，Lutz 等人（Lutz 等人 2021）提供了不存在兔子洞效应的证据，同时证明了各种政治意识形态中存在主流化效应。其他研究（Ledwich 和 Zaitsev 2019；Munger 和 Phillips 2022； Hosseinmardi 等人 2021； Makhortykh 和 Urman 2020）也对 YouTube 存在兔子洞效应这一观点提出了质疑，并显示出推荐内容将用户引向主流和中立内容的证据。与这些发现相反，Haroon 等人（Haroon 等人 2022）提供了证据，表明 YouTube 通过“接下来播放”和首页推荐，将用户引向日益偏颇和激进的政治内容。这些发现与另一批研究（Bryant 2020；Ribeiro 等人 2020 年；Tomlein 等人 2021；Kirdemir 和 Agarwal 2021；Papadamou 等人 2019、2021）相互补充；后者认为，YouTube 推荐在政治、科学和健康相关领域加剧了极化。不同于此前的研究，我们的目标并不是支持或削弱关于 YouTube 倾向于影响极化的特定理论。相反，我们旨在揭示造成这些差异的可能原因，并提供指导方针，以避免审计研究社群内部出现此类混淆和矛盾。最近，Hussein、Juneja 和 Mitra（Hussein、Juneja 和 Mitra 2020）展示了用户资料的人口统计特征如何改变 YouTube 的推荐。在一项聚焦于 YouTube 去货币化算法的研究中，Dunna 等人（Dunna 等人 2022）发现了推荐算法与去货币化算法存在关联的证据。此外，Google 还发表了大量描述 YouTube 所使用推荐算法的出版物。这些研究提出，用户资料、观看历史、视频观看时长和点击率可作为其内容排序算法的特征（Zhao 等人 2019；Tang 等人 2019；Fu 等人 2016；Covington、Adams 和 Sargin 2016；Zhao 等人 2015；Brodersen、Scellato 和 Wattenhofer 2012；Davidson 等人 2010）。这些描述为我们选择审计参数提供了依据。

提升基于爬虫的研究的可靠性。互联网测量领域也曾开展过与我们类似的工作。这些工作主要致力于促进万维网测量和隐私领域中更可靠、可复现的研究。Yadav 等人（Yadav 和 Goyal，2015年）研究了一组开源网络爬虫，并展示了每种爬虫适用于哪些不同的使用场景。近期，Ahmed 等人（Ahmad等，2020年）展示了不同爬虫对测量和安全研究推论的影响。类似地，Zeber 等人（Zeber 等人，2020年）和 Jueckstock 等人（Jueckstock等，2021年）也展示了爬虫及其配置的选择如何损害实验的可重复性。我们的工作在这些研究基础上进一步识别了平台特定的审计挑战。

7 结论

局限性。从根本上说，我们的工作是一项尽力而为的研究，旨在了解不同审计方法决策对从 YouTube 收集的推荐内容的影响。YouTube 是视频托管平台中流媒体播放量最高的平台之一（Duò 2023）。尽管我们的审计范围仅限于 YouTube，但它有助于为其他视频平台上的审计研究铺路。因此，我们的研究并非没有局限性。首先，我们自身在计算资源和经费方面都有限，因此不得不决定要探索哪些爬取参数。这影响了我们进行以下工作的能力：（1）探索每棵推荐树中的更多路径；（2）开展超过八次同步树探索；以及（3）以更大的深度探索推荐树。为缓解这些局限可能导致的错误推论，我们仅进行同类节点和位置依赖比较，并确保研究中测得的任何差异都考虑到推荐算法普遍具有概率性这一特点，方法是测量组间差异并将其与组内差异进行比较。其次，从我们外部的观察角度来看，存在一些无法控制的潜在效应，而我们的工作实际上是在测量一个黑箱系统。我们尽力识别其中一些因素（例如 A/B 测试、数据中心位置、测量位置等），并尝试逐一加以控制。不过，仍有可能存在未计入的效应影响我们的结果。最后，我们承认，所选训练集和种子视频可能不足以观察到互动对推荐系统产生的所有影响。尽管如此，我们仍提供了有参考价值的数据点，供面临大量相互矛盾结果的研究群体考量。

结论。 本研究展示了审计配置对其生成的推荐树特征的影响。具体而言，我们表明，尽管训练集确实会对推荐内容产生统计学影响，但 YouTube 推荐中的“近期偏见”会显著减弱这些影响（§3）。因此，在审计中选择观看视频时需要格外谨慎。更重要的是，必须披露这些决策，并且必须 以这些决策为条件进行审计推断。我们对不同类型审计配置文件的分析（§4）表明，与在整个审计期间仅保留 YouTube Cookie 相比，获取干净 YouTube 账号这一成本高昂的任务并不会带来显著不同的结果。此外，我们的研究结果还表明，可以利用 YouTube 提供的“清除历史记录”功能复用账号。最后，我们对审计中的各种计算折衷所做的分析（§5）表明，无需观看视频的特定比例，该视频就能影响后续推荐（相反，预设阈值似乎已足够）；审计人员无需执行诸如编程模拟光标点击视频之类颇具挑战性的自动化任务；爬取深度会影响推荐树的特征（因此，报告审计推断时应将其作为条件）。

参考文献

2022年。白宫科技平台问责听证会纪要。白宫简报室。URLhttps://www.whitehouse.gov/briefing-room/statements-releases/2022/09/08/readout-of-white-house-listening-session-on-tech-platform-accountability/。

2022年。威尔·史密斯—克里斯·洛克掌掴事件——维基百科。维基百科。URLhttps://en.wikipedia.org/wiki/Will_Smith%E2%80%93Chris_Rock_slapping_事件。

Ahmad, S. S.; Dar, M. D.; Zaffar, M. F.; Vallina-Rodriguez, N.; and Nithyanand, R. 2020年。表象错觉还是顿悟？爬虫如何影响我们对万维网的理解。WWW '20。美国纽约州纽约市：计算机协会。

Barberá, P.2020年。社交媒体、回音室与政治极化。社交媒体与民主：领域现状与改革前景34。

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003年。潜在狄利克雷分配。《机器学习研究杂志》。

Brodersen, A.; Scellato, S.; and Wattenhofer, M. 2012. YouTube 全球化：视频的地理流行度。载于第 *21* 届万维网国际会议论文集。URL http://doi.acm.org/10.1145/2187836.2187870。

Bryant, L. V. 2020年。YouTube 算法与另类右翼过滤气泡。《开放信息科学》4(1)：85–90。

Covington, P.; Adams, J.; and Sargin, E. 2016. 用于 YouTube 推荐的深度神经网络。RecSys '16。计算机协会。ISBN 9781450340359。

Davidson, J.; Liebald, B.; Liu, J.; Nandy, P.; Van Vleet, T.; Gargi, U.; Gupta, S.; He, Y.; Lambert, M.; Livingston, B.; and Sampath, D. 2010. YouTube 视频推荐系统。RecSys '10。计算机协会。

DiCiccio, T. J.；和 Efron, B. 1996. 自助法置信区间。《统计科学》11(3): 189–228.

Dunna, A.；Keith, K.；Zuckerman, E.；Vallina-Rodriguez, N.；O'Connor, B.；和 Nithyanand, R. 2022. 关注幕后算法：提高 YouTube 去货币化算法的透明度。载于 *2022* 年 *ACM* 计算机支持的协同工作会议论文集（*ACM CSCW 2022*）。URL https://sparta.cs.uiowa.edu/people/rishab/papers/cscw-2022-demon.pdf。

Duò, M. 2023。2023 年值得考虑的 10 种视频托管解决方案（免费与付费）。URL https://kinsta.com/blog/video-hosting/#:~:文本=YouTube，YouTube%20主页&文本=YouTube%20是%20最大的视频，月%2C%20没有什么%20能与之媲美。Efron，B. 1987。更优的自助法置信区间。美国统计学会期刊82(397): 171–185。

Facebook。2022年。Facebook A/B 测试。URLhttps://www.facebook.com/business/help/1738164643098669?id=445653312788501。Fu，B.；Chi，E.；Cao，P.；Yang，R.；和 Singh，S. 2016。视频观看时长与评论情感：来自 YouTube 的经验。Funk，M. 2020。YouTube 如何计算观看次数？这比你想象的要复杂！Tubics。URLhttps://www.tubics.com/blog/what-counts-as-a-view-on-youtube。Hannak，A.；Sapiezynski，P.；Molavi Kakhki，A.；Krishnamurthy，B.；Lazer，D.；Mislove，A.；和 Wilson，C. 2013。网络搜索个性化测量。WWW'13。计算机协会。doi：10.1145/2488388.2488435。URLhttps://doi.org/10.1145/2488388.2488435。

Haroon, M.；Chhabra, A.；Liu, X.；Mohapatra, P.；Shafiq, Z.；和 Wojcieszak, M. 2022. YouTube，极端化推手？审计并缓解 YouTube 推荐中的意识形态偏见。*arXiv* 预印本 *arXiv:2203.10666* 。

Hosseinmardi, H.; Ghasemian, A.; Clauset, A.; Mobius, M.; Rothschild, D. M.; 和 Watts, D. J. 2021. 考察 YouTube 上激进内容的消费情况。美国国家科学院院刊118(32): e2101967118.

Hussein, E.; Juneja, P.; 和 Mitra, T. 2020. 衡量视频搜索平台中的错误信息：一项针对 YouTube 的审计研究。*ACM* 人机交互汇刊4(CSCW1): 1–27.

Jueckstock, J.; Sarker, S.; Snyder, P.; Beggs, A.; Papadopou- los, P.; Varvello, M.; Livshits, B.; 和 Kapravelos, A. 2021. 迈向切实且可复现的万维网爬取测量。在 *2021* 年万维网大会论文集，WWW '21。美国计算机协会。

Khattak, S.; Fifield, D.; Afroz, S.; Javed, M.; Sundaresan, S.; McCoy, D.; Paxson, V.; 和 Murdoch, S. J. 2016. 你看到的和我看到的一样吗？匿名用户的差别化待遇。发表于*NDSS*。

Kirdemir, B.; 和 Agarwal, N. 2021. 探索 YouTube 视频推荐网络中的偏见与信息茧房。发表于复杂网络及其应用国际会议，166–177。Springer。

算法极端主义：考察 YouTube 的激进化兔子洞。*CoRR*abs/1912.11211。URL http://arxiv.org/abs/1912.11211。

Lutz, M.; Gadaginmath, S.; Vairavan, N.; 和 Mui, P. 2021. 考察 YouTube 搜索与推荐算法中的政治偏见。发表于 *2021 IEEE* 计算智能研讨会系列（*SSCI*），1–7。IEEE。

Makhortykh, M.；和 Urman, A. 2020年。《伟大的随机化器：使用虚拟代理审计 YouTube 推荐算法对意识形态色彩浓厚的新闻内容分发的效应》。《*AoIR* 互联网研究精选论文集》。

Munger, K.；和 Phillips, J. 2022年。右翼 YouTube：供需视角。《style id='1'>《国际新闻/政治学杂志》27(1)：186–219。

Papadamou, K.; Papasavva, A.; Zannettou,S.; Blackburn, J.; Kourtellis, N.; Leontiadis, I.; Stringhini, G.; and Sirivianos, M. 2019. 受扰的 YouTube 儿童版：刻画并检测 YouTube 上令人不安的内容。*CoRR*abs/1901.07046. URL http://arxiv.org/abs/1901.07046。

Papadamou, K.; Zannettou,S.;Blackburn,J.; Cristofaro, E. D.; Stringhini, G.; and Sirivianos, M. 2020年。理解 YouTube 上的非自愿独身者社群。*CoRR*abs/2001.08293. URLhttps://arxiv.org/abs/2001.08293。

Papadamou, K.; Zannettou, S.; Blackburn, J.; Cristofaro, E. D.; Stringhini, G.; and Sirivianos, M. 2021。“这只是一种流感”：评估观看历史对 YouTube 伪科学视频推荐的效应。

Parsons,J. 2017年。观看 YouTube 视频多长时间才算一次观看？GrowTraffic。URL https://growtraffic.com/blog/2017/08/youtube-video-counts-view。

Ribeiro, M. H.; Ottoni, R.; West, R.; Almeida, V. A. F.；以及 Meira, W. 2020年。审计 YouTube 上的激进化路径。载于*2020* 年公平性、问责制与透明度会议论文集。美国计算机协会。doi:10. 1145/3351095.3372879。URLhttps://doi.org/10.1145/3351095.3372879。

Ribeiro, M. H.; Ottoni, R.; West, R.; Almeida, V. A. F.; 和 Meira, W. 2020年。审计 YouTube 上的激进化路径。载于 *2020*年公平、问责与透明度会议论文集。美国纽约州纽约市：美国计算机协会。Rosario, B. 2000年。潜在语义索引：概述。技术报告 *INFOSYS* 。Sandvig, C.;Hamilton, K.; Karahalios,K.; 和 Langbort, C. ???? 审计算法：检测互联网平台歧视的研究方法。Selenium。2022年。Selenium 操作链。
U
R
L
h
t
t
p
s://
w
w
w.
s
e
l
e
n
i
u
m.
d
e
v/
s
e
l
enium/docs/api/py/webdriver/selenium.webdriver.common.action
chains.html。_

Singh, R.; Nithyanand, R.; Afroz, S.; Pearce, P.; Tschantz, M. C.; Gill, P.; 和 Paxson, V. 2017年。Tor 出口屏蔽的性质与动态特征。载于 第 *26* 届 *USENIX* 安全研讨会（*USENIX Security 17*）。加拿大不列颠哥伦比亚省温哥华：USENIX 协会。

SpaCy。2022年。SpaCy 语言学特征。URL https://spacy.io/usage/linguistic-features。

Tang, J.; Belletti, F.; Jain, S.; Chen, M.; Beutel, A.; Xu, C.; 以及 H. Chi, E. 2019。面向长程依赖用户序列的神经混合推荐器。WWW ’19。美国纽约州纽约市：美国计算机协会。URLhttps://doi.org/10.1145/3308558.3313650。

Tomlein, M.; Pecher, B.; Simko, J.; Srba, I.; Móro, R.; Stefancova, E.; Kompan, M.; Hrckova, A.; Podrouzek, J.; 和 Bieliková, M. 2021. YouTube 错误信息过滤气泡审计：气泡破裂与近期行为变化。载于 *RecSys ’21*：第十五届 *ACM* 推荐系统会议，阿姆斯特丹，荷兰，*2021*年*9*月*27*日—*2021*年*10*月*1*日。doi:10. 1145/3460231.3474241。URLhttps://doi.org/10.1145/3460231.3474241。

Yadav, M.; 和 Goyal, N. 2015. 开源爬虫比较：综述。

Zeber, D.; Bird, S.; Oliveira, C.; Rudametkin, W.; Segall, I.; Wollsén, F.; 和 Lopatka, M. 2020年。自动化网页爬取作为人工浏览替代方案的代表性。载于 *2020*年万维网大会论文集，WWW ’20。美国纽约州纽约市：计算机协会。URLhttps://doi.org/10.1145/3366423.3380104。

Zhao, Z.; Cheng, Z.; Hong,L.; and Chi, E. H.2015. 通过行为分解改进用户主题兴趣档案。WWW ’15。

Zhao, Z.; Hong, L.; Wei, L.; Chen,J.; Nath, A.; Andrews, S.; Kumthekar, A.; Sathiamoorthy, M.; Yi, X.; and Chi, E. 2019. 推荐接下来观看的视频：多任务排序系统。RecSys ’19。美国纽约州纽约市：美国计算机协会。doi: 10.1145/3298689.3346997。URLhttps://doi.org/10.1145/3298689.3346997。