开发者也需要保护：社交编码平台中的隐私视角与研究挑战

Nicol´as E. D´ıaz Ferreyra Abdessamad Imine Melina Vidoni Riccardo Scandariato软件安全研究所 洛林大学和 *Loria-* 计算机学院 软件安全研究所 汉堡工业大学 法国国家科学研究中心—*Inria* 南锡大区 澳大利亚国立大学 汉堡工业大学德国汉堡 法国南锡 澳大利亚堪培拉 德国汉堡 nicolas.diaz-ferreyra@tuhh.de abdessamad.imine@loria.fr melina.vidoni@anu.edu.au riccardo.scandariato@tuhh.de

arXiv:2303.01822v1 [cs.SE] 3 Mar 2023

摘要—**像 GitHub 这样的社交编码平台（SCPs）凭借其协作和版本控制功能，已成为现代软件工程的核心。与 Facebook 等主流在线社交网络（OSNs）一样，SCP 用户也面临隐私攻击和威胁，因为其个人资料和软件代码库中包含大量个人及项目相关数据。然而，与在线社交网络不同，现有文献尚未充分探讨或记录 SCP 用户的隐私顾虑与实践。在本研究中，我们介绍了一项在线调查（N=105）的初步结果，该调查探讨了开发者对 SCP 所带来隐私威胁的顾虑和看法。研究结果表明，尽管用户对社交和组织层面的隐私威胁表示担忧，但他们通常认为在这些平台上分享个人及项目相关信息是安全的。此外，相较于试图重新识别源代码贡献者的攻击，针对推断敏感属性的攻击被认为更有可能发生。基于这些研究发现，我们提出了一系列建议，以指导未来针对 SCP 隐私与身份管理的研究。**

索引词—**社交编码平台，隐私顾虑，可用安全，隐私工程**

一、引言

近年来，随着 GitHub 和 BitBucket 等社交编码平台（SCPs）的兴起，现代软件工程的格局发生了显著变化 [1, 2]。从软件爱好者到资深工程师，SCPs 汇聚了技术专长和背景各异的个人，共同参与公开协作式软件开发 [3]。此外，自 2000 年代末问世以来，SCPs 不仅吸引了新的参与者加入开源软件（OSS）生态系统，还通过提供免费托管和版本控制功能，并辅以社交功能，重新定义了 OSS 部署的协作性质 [4, 5]。

从本质上说，软件协作平台（SCP）具有许多主流在线社交网络（OSN）的特征，例如 Facebook 或 Twitter。比如，为了建立社交关系、开展协作并传播自己的项目，软件协作平台的用户通常会像图 1 所示那样创建个人资料。这些个人资料汇总了大量有关开发者人口统计信息及其在平台上的行为的信息，包括他们拥有、维护和参与贡献的项目和代码库。此外，针对某个用户可以获取的一些活动指标包括：(i) 他们创建的代码库数量，(ii) 他们报告的问题数量，(iii) 他们提交的拉取请求数量，以及 (iv) 他们撰写的提交数量 [6]。

以开发者为中心的SCP指标（及其汇总的信息）帮助研究人员刻画 GitHub社区，并揭示现实环境中常见的软件工程实践。然而，这类信息也可能被用于针对特定用户和社会技术群体发起定向网络攻击，例如针对特定性别或地区的个人，或具有特定技术背景的个人（如容易采用不安全编码实践的用户） [7]。此外，随着人工智能（AI）和机器学习（ML）不断发展，即使是保持较为私密、低调资料的用户，也可能因平台上可获取的大量数据而被重新识别。

研究动机： 在线社交网络（OSNs）语境下的隐私与身份管理已得到广泛研究。总体而言，现有文献 [8] 探讨的诸多问题包括个人信息的意外披露、匿名性以及数据滥用。同样，隐私研究者也高度关注如何从元数据中推断敏感属性，以及如何重新识别数据主体，并致力于为在线社交网络开发隐私增强技术（PETs） [9, 10]。尤其是塑造社会技术解决方案，以保护用户的身份和个人属性免遭重新识别、推断和可链接性攻击。

尽管软件协作平台与在线社交网络之间存在相似之处，但前者的隐私问题并未像后者的问题那样受到同等程度的关注。然而，软件协作平台上可获取的信息可能被用于实施社会工程攻击，进而危及公司和公共机构整体的安全。例如，2022年初，一群网络攻击者利用公开代码仓库中暴露的密钥，获取了三星190 Gb的数据 [11]。此外，GitGuardian最近的一份报告 [12] 显示，

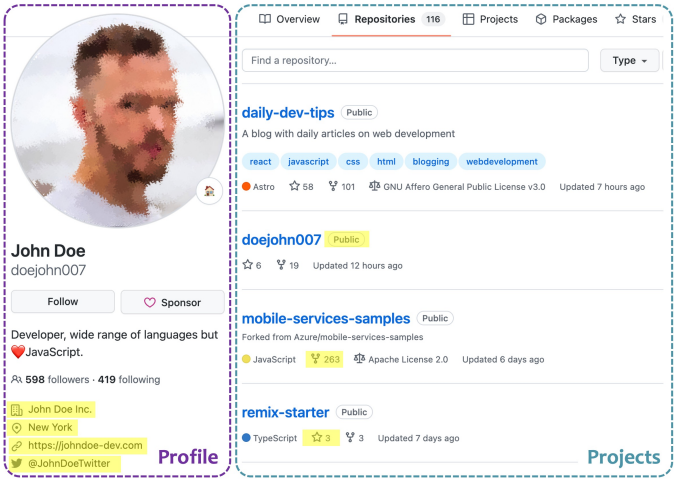

图 1. GitHub 上开发者页面的示例。

GitHub上公开暴露的机密信息数量自2020年以来增加了一倍多。尽管这一数字令人担忧，但并不应感到意外，因为开发者只需无意中将代码推送到公共代码仓库，就可能危及整个系统的安全 [13]。因此，有必要开展进一步研究，以刻画协作式编码环境中出现的隐私威胁与隐私问题。

贡献与研究问题： 在这项工作中，我们介绍并讨论了一项关于软件协作平台隐私顾虑的实证研究的初步结果。该研究采用在线调查（N=105），旨在回答以下研究问题（RQs）：

- **RQ1**： 软件协作平台的用户对自己的隐私有多担忧？为回答这一研究问题，我们根据现有文献中的隐私相关构念设计了一份问卷。具体而言，我们评估了开发者对其个人信息和项目相关信息的感知控制、对社会／组织威胁的顾虑，以及他们此前在软件协作平台上遭遇隐私侵犯的经历等。
- **RQ2**： 他们认为某些隐私威胁在软件协作平台上出现的可能性有多大？这一研究问题旨在确定软件协作平台用户是否低估了遭受隐私攻击的可能性。为此，我们请参与者评估两个假设性隐私攻击场景发生的可能性，场景分别描述了：（i）根据个人资料数据推断敏感属性；（ii）根据软件协作平台代码仓库中可获取的信息重新识别代码作者身份。

总而言之，我们的研究发现表明，社交隐私威胁和组织隐私威胁令社交编码平台（SCP）的用户感到担忧。尽管如此，许多参与者表示：(i) 在这些环境中发布个人及项目相关数据让他们感到安全；(ii) 他们觉得自己能够控制此类数据。我们将讨论这些见解所带来的影响，并进一步阐述一些未来研究的视角。

II. 背景与相关工作

A. 在线社交网络中的隐私研究

许多研究探讨了用户对服务提供商、信息访问以及在线社交网络中的个人数据共享的顾虑和态度。这些研究揭示了此类顾虑对人们的自我披露行为、访问控制偏好以及对在线隐私风险的整体认知所起的作用，并取得了重要研究发现 [8]。例如，已有研究反复表明，尽管在线社交网络用户对隐私高度关注，他们往往仍无法维护个人数据的情境完整性 [14]。这一现象通常与安全控制的可用性较低，以及平台在收集、处理和存储用户数据时缺乏透明度有关 [8]。近期的研究还关注到影响人们隐私决策过程的认知偏差 [15]。特别是，这些研究指出，在线社交网络用户倾向于用“经验法则”或启发式方法取代理性的风险判断，以降低网络安全选择的复杂性 [16]。因此，他们往往会采用不够理想的隐私配置，使自己的数据暴露给不可信受众或恶意用户。

先前的研究也试图刻画在线社交网络中潜在的隐私攻击。这类攻击通常利用用户在个人资料和帖子中披露的信息，为能够推断敏感属性和行为模式的机器学习模型提供数据[17]。例如，Gu 等人 [18] 开发了一种概率模型，利用以用户为中心的数据和社会关系数据来推断个人的家庭所在地。如果目标用户过去曾披露过某些位置数据（例如在酒吧或餐厅签到），该模型的准确率可超过 90%；在目标用户未披露过相关数据时，准确率约为 60%。同样，Comito [19]引入了一个监督学习框架，根据个人以往的移动行为预测其下一个位置。近期发表的研究还说明了深度神经网络（DNNs）在检测图像中的敏感信息方面的有效性，包括人物的年龄 [20]、性别 [20], 和种族 [21]。总体而言，基于 DNN 的隐私攻击凭借其在大型数据集 [17] 上的高性能和高效率，优于其他最先进的方法。

B. 社交编码平台中的隐私研究

与在线社交网络不同，针对软件协作平台的研究并未十分关注用户对隐私和数据保护的担忧。此外，目前的研究充其量可以被视为攻击模型，旨在揭示开发者的人口统计属性与这些平台上呈现的行为模式之间的关联。例如，Wu 等人[6] 观察到，GitHub用户生成的拉取请求和提交数量，与特定类别的个人资料图片（例如，人类、徽标或动物）以及人口统计特征相关。针对性别偏见的研究也促进了对软件协作平台中开发者行为的刻画。Imtiaz 等人的一项研究 [22] 表明，GitHub上的女性用户在沟通方面比男性更加克制，并且倾向于将工作集中在较少的项目和组织中。

同样，Newton 等人 [23] 发现，被认定为女性的GitHub贡献者在开源项目中的任职时间比男性更短。

针对代码归属的研究，在很大程度上也帮助描绘了软件协作平台中的隐私攻击全貌。研究尤其表明，公共软件仓库中的代码可被利用，通过反映个人编程风格的风格模式重新识别其贡献者。这些模式涵盖从注释和代码布局中的简单特征，到语法用法和控制流中的细微习惯。例如，Caliskan-Islam 等人 [24] 采用机器学习方法，利用从抽象语法树（AST）中提取的文体特征集，对 C/C++ 源代码作者进行去匿名化。此外，Alsulami 等人[25] 提出了一种用于自动提取此类特征的长短期记忆模型，而 Abuhamad 等人 [26]则提出了类似技术，旨在实现大规模、与语言无关且能抵御混淆的代码作者身份识别。后一种方法对语言敏感，能够识别使用四种编程语言（如 C、C++、Java 和 Python）的作者，即便作者使用了多种语言的组合（如 Java/C++、Python/C++）也不例外。

III. 方法

A. 调查设计

我们开展了一项在线调查，以回答第一节提出的研究问题。该调查主要分为两个部分：第一部分探讨软件协作平台中的隐私顾虑（RQ1），第二部分关注源于个人资料和代码库中可获取信息的隐私攻击（RQ2）。为 elicitation 隐私顾虑，我们参考了现有在线社交网络文献中的以下心理构念和量表：

- **对软件协作平台中披露的个人和项目相关信息的感知控制（CTRL）** [27]。
- **对软件协作平台中隐私风险的感知（RSK）** [27]。
- **既往内部隐私侵犯（PIIP）**，例如在软件协作平台中曾经历的隐私侵犯 [28]。
- **组织威胁隐私顾虑（PCOT）**，包括服务提供商和第三方对数据的收集和二次使用 [29]。
- **外部信息隐私顾虑（EIPC）**，涉及可能影响软件协作平台中他人（例如同事和合作者）隐私的信息披露实践 [28]。
- **社会威胁隐私顾虑（PCST）**，例如尴尬，以及其他软件协作平台用户可能实施的恶意行为 [29]。

这些量表由相应作者编制并验证：CTRL 和 RSK 由 Krasnova 等人验证[27]；EIPC 和 PIIP 由 Morlok 验证 [28]；PCOT 和 PCST 由 Krasnova 等人验证 [29]。为满足本研究的需要，我们对这些量表略作调整，将每个量表条目中的“在线社交网络”替换为“软件协作平台”。所有构念均采用 6 点李克特量表测量，范围从完全不同意到完全同意。我们通过以下方式评估其信度：

情景 1：您在软件协作平台上的活动（例如提交、注释和反应）可用于推断您的性别及其他社会人口统计信息。

情景 2：某人（例如一位同事）使用自己的账户将你编写的代码推送到公共代码仓库。随后，利用软件协作平台上可获取的信息（例如你编写的其他代码片段和个人资料信息），便可以追溯出这些代码的作者是你。

图 2. 软件协作平台中的假设性隐私攻击场景。

计算相应的克朗巴赫阿尔法系数，结果在所有情况下均高于 0.7。这表明每个量表中的条目具有较高的内部一致性，因为高于 0.7 的数值被认为是“可接受的”。

调查的第二部分由两个假设性的隐私威胁场景组成，参与者需要对其进行评估。如图 3 所示，这些场景代表了现有文献中记录的隐私攻击，而利用软件协作平台（第 II-B 节）中可获得的信息即可实施这些攻击。具体而言，(i) 发现开发者活动与其人口统计特征之间的敏感关联；以及 (ii) 通过编码风格模式重新识别其源代码贡献。我们特别询问了参与者，按照非常可能、可能、有些可能、有些不太可能、不太可能和非常不可能这几个选项，他们认为这些场景在现实生活中发生的可能性有多大。

B. 人口与抽样

参与者通过 Prolific[1] 招募，并根据其 (i) 自我报告的软件开发技术知识和 (ii) 自我报告的计算机编程技能进行预筛选。他们必须年满 18 岁才能参加研究；完成一项平均用时 15 分钟的调查后，可获得 2.00 英镑的报酬（通过 Prolific 支付）。作为标准质量控制措施，我们针对的是已经参与过至少 10 项其他研究且最低通过率为 98% [30] 的用户。调查中还设置了两道注意力检测题，以识别并剔除参与度不足的参与者的回答。最终，我们从 110 份完整提交中收集到 105 份有效回答。

C. 伦理考量

本研究遵循《赫尔辛基宣言》开展，并经德国实验经济研究协会伦理委员会批准。所有参与者在加入实验前均获知研究流程（包括数据隐私声明），并被要求提供知情同意。他们也可以随时退出，且其回答不会被记录。调查工具、同意书和研究结果均可在论文的**复现包**[2]中获取。

[1]https://prolific.co/ [2]https://doi.org/10.5281/zenodo.7692654

表 I 调查条目摘录。

构念	分析的构念条目	1	2	3	4	5	6
CTRL	我觉得自己能够控制个人和技术信息 我在软件协作平台上提供的信息	0	9	8	33	43	12
RSK	我觉得发布我的个人信息和项目相关信息是安全的 关于软件协作平台的信息	4	5	22	37	29	8
PIIP	我已经经历过 在软件协作平台上遭遇隐私侵犯	44	42	9	5	4	1
PCOT	我经常担心其他人可能会收集我公开的 在软件协作平台上可获取的信息	6	20	20	32	21	6
EIPC	我担心未经授权的人可能会访问我的代码 我在软件协作平台上分享的同事	9	17	23	37	14	5
PCST	我担心其他用户可能会利用他们 通过软件协作平台了解到的关于我的信息来占我便宜	14	24	23	26	16	2

注： 1=完全不同意， 2= 不同意， 3= 有些不同意， 4= 有些同意， 5= 同意， 6= 完全同意。

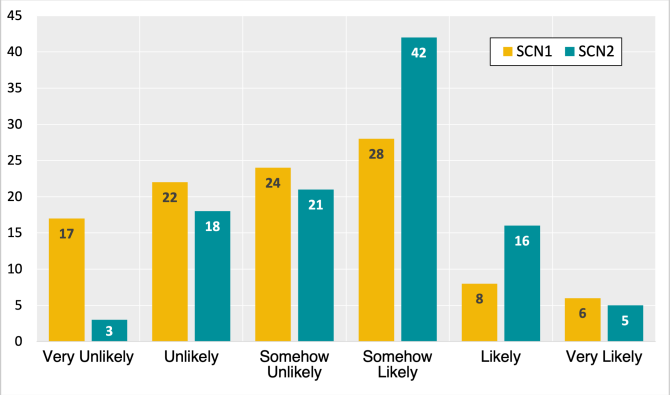

图 3. 隐私攻击场景评估。

IV. 结果与讨论

在 105 名研究参与者中，94 名为男性，10 名为女性，1 名为非二元性别。约 48.6% 的参与者表示，他们使用软件协作平台的经验为 2 至 5 年；14.3% 为 6 至 10 年，13.3% 少于 2 年，3.8% 超过 10 年。关于软件开发，35.2% 的参与者经验少于 2 年，另有 35.2% 为 2 至 5 年，16.2% 为 6 至 10 年，13.3% 超过 10 年。样本人口统计信息的完整概览见表 II。

调查第一部分所引出的每个构念（即 CTRL、RSK、PIIP、PCOT、EIPC 和 PCST）包含 3 至 7 条陈述或条目。我们为每个构念选取了一个具有代表性的条目，用于对软件协作平台中的开发者隐私担忧开展初步分析。尽管这种方法具有探索性，但符合本研究的目的，即提供一些初步见解，并推动对此问题开展进一步研究。论文的复现包中提供了每个量表条目所得结果的完整概览。

如表 I 所示，大多数参与者认同控制感陈述，认为自己能够控制在软件协作平台上披露的个人和技术信息。同样，约 71% 的参与者认同相应的安全感陈述（得分为 4–6），而 29% 的参与者表示，发布个人和项目相关信息时并不感到安全（得分为 1–3）。关于以往经验，只有 9.5% 的受访者曾遭受过隐私侵犯，其余受访者则倾向于不同意这一构念条目。另一方面，约 59% 的研究参与者表示担心第三方收集数据（PCOT），另有 56% 的参与者担心他人未授权访问自己分享的代码（EIPC）。最后，谈到社会威胁时，42% 的参与者倾向于认同隐私相关社会威胁陈述，即其他软件协作平台用户会利用公开信息。

图 3 展示了参与者对所提出隐私威胁场景的评估。我们观察到，60% 的参与者将情景 1（SCN1）置于量表的“不太可能”一侧（即 非常不可能、不可能 或 某种程度上不可能），而 40% 的参与者认为该情景在一定程度上“可能”（即 某种程度上可能、可能 或 非常可能）。令人意外的是，在情景 2（SCN2）中，这些数值发生了反转：60% 的参与者认为该情景在某种程度上“可能”，其余 40% 则认为“不可能”。分别分析每个可能性等级时，我们还可以发现 SCN1 与 SCN2 之间存在一些显著差异。尤其是，认为 SCN1 非常不可能 的参与者人数多于认为 SCN2 如此的人数，但选择可能 和某种程度上可能的人数则较少。

A. 对隐私研究的影响

(i) 隐私决策：我们的结果表明，尽管软件协作平台的用户通常觉得自己能够控制其信息，但他们也担忧以下隐私威胁：(i) 服务提供商和第三方的数据收集，(ii) 未经授权访问其他用户编写的代码，以及 (iii) 对公开信息的利用。此前针对在线社交网络的研究也报告了类似的背离现象，通常称为“隐私悖论” [31]。也就是说，用户的隐私顾虑与其实际隐私相关行为之间存在差异 [32]。然而，控制感通常有助于缓解这些顾虑，并最终促使用户披露更多信息 [32]。因此，应借助更多实证证据对这一观察结果进行进一步调查和讨论。

（ii）可用性与透明度：参与者对 PCOT、EIPC 和 PCST 条目的回答，也引发了人们对当前隐私增强机制可用性的疑问

表 I 调查自我报告的人口统计数据。

人口统计	范围	频数	%
性别	Male	94	89.5%
	女性	10	9.5%
	非二元性别	1	1.0%
教育	高中及以下	14	13.3%
	部分大学教育	19	18.1%
教育	本科（理学学士、文学学士）	48	45.7%
程度	研究生（理学硕士、博士）	24	22.9%
就业状况	学生	26	24.8%
	失业，正在找工作	10	9.5%
	失业，没有找工作	2	1.9%
	全职工作	13	12.4%
	兼职工作	54	51.4%
软件开发经验	<2 年	37	35.2%
	2-5年	37	35.2%
	6-10年	17	16.2%
	>10 年	14	13.3%
经验 使用软件协作平台	<2 年	35	13.3%
	2-5 年	51	48.6%
	6-10 年	15	14.3%
	>10 年	4	3.8%

（SCPs）中的隐私增强机制。人们对组织威胁的感知，往往源于平台在如何收集、处理、存储和共享（个人）信息方面缺乏透明度 [33]。对此，分析 SCP 用户在多大程度上能够合理理解常见的数据处理实践（例如，SCPs 和第三方的数据处理实践），以及他们如何将这些认知转化为更严格（或更宽松）的隐私配置（例如，软件代码库的可见性），将大有裨益。在某种程度上，软件仓库挖掘（MSR）社区的从业者也报告过类似的担忧 [34, 35]。尤其是，未经用户充分同意便从软件协作平台代码仓库中获取信息所带来的伦理影响。因此，关键不仅在于审视这些实践可能具有的侵扰性，还在于评估 SCPs 当前访问控制机制的局限与不足。

(iii) 多方隐私：用户对社交隐私和外部隐私的担忧，促使近期研究开始探讨在线社交网络中多方隐私冲突的解决方案 [36]。当用户分享的信息也可能损害他人的隐私时，便常会出现此类冲突，例如分享合照或在帖子中提及他人。总而言之，多方隐私从集体视角审视信息披露的规范与边界，尤其关注特定数据项的共同所有者之间相互冲突的隐私偏好 [37]。软件代码库中的秘密信息扩散可视为软件协作平台中此类冲突的一个实例。此外，基于拉取请求的工作流原则上有助于防止秘密信息共享，因为它们能促进对推送到特定代码库的代码进行监控和审查 [38]。不过，仍需系统地研究多方隐私冲突的具体来源、相关担忧及应对措施。

（iv）风险评估：以往研究表明，过去的经验在很大程度上会影响人们的隐私感知和行为 [39]。因此，数量相近的参与者倾向于不同意 PIIP 陈述、但同意 RSK 陈述（分别为 90% 和 70%），这并不令人意外。此外，参与者对所提出威胁场景的评估，也与他们对隐私控制和安全性的整体感知一致；许多人认为，SCN1 和 SCN2 极不可能、不太可能或有些不太可能发生。鉴于风险感知在隐私决策中的核心作用 [15], ，提高开发者对软件协作平台中出现的网络安全威胁的认识至关重要。同样重要的是，为他们提供方法和工具，帮助他们在评估潜在风险的同时，维护所分享信息的情境完整性。

B. 局限性与有效性威胁

本研究受到研究样本规模和构成方面局限性的影响。一方面，样本规模相对较小，因此我们无法将研究结果推广到整个软件协作平台社区。因此，从研究中获得的见解应视为初步结论，而非定论，并应促使该领域开展进一步研究。另一方面，通过众包平台招募的参与者可能会采取不诚实行为，例如谎称符合实验的资格标准，或提供随机答案。如第三节 B 部分所述，我们通过使用 Prolific 内置的资格筛选功能，招募能够积极参与研究的合适参与者，以尽量减少这些威胁。未来，我们计划通过设置 Danilova 等人提出的此类专业知识问题来改进这一环节 [40]。

V. 结论与未来工作

软件协作平台是了解开发者社会技术技能的丰富信息来源，并在很大程度上帮助刻画了当前软件工程的发展趋势。然而，尽管软件协作平台十分重要，人们却很少关注其用户的隐私相关行为，以及由不熟悉信息披露实践所引发的威胁。在本研究中，我们对此提供了一些实证见解，旨在为未来的研究铺平道路。总体而言，我们的结果表明，有必要进一步研究评估软件协作平台中现有隐私增强机制的可用性。也就是说，需要确定这些机制是否满足用户在透明度、访问控制和匿名性方面的个人与集体目标。

开发者对隐私威胁的了解和认知也需要更多研究，因为评估他们（可能存在的）意识不足，对于制定适当的网络安全培训计划至关重要。在线社交网络文献[8]已充分研究并记录了风险意识在采用隐私增强技术方面的作用。未来，我们计划深入探讨软件协作平台中风险意识与隐私相关行为之间的相互作用。尤其是研究开发者对潜在隐私威胁（例如秘密信息扩散或代码重新归属）的了解，以及他们对网络安全最佳实践（例如，参见 [38]）的采用情况。

参考文献

[1] T. Kinsman、M. Wessel、M. A. Gerosa 和 C. Treude，《软件开发者如何使用 GitHub Actions 自动化其工作流？》，载于 2021年IEEE/ACM第18届软件仓库挖掘国际会议（MSR），2021年，第420–431页。[2] T. Mens、M. Cataldo 和 D. Damian，《社交型开发者：软件开发的未来 [客座编辑导言]》，IEEE Software，第36卷，第1期，第11–14页，2019年。[3] J. E. Montandon、M. T. Valente 和 L. L. Silva，《挖掘 GitHub 用户的技术角色》，Information and Software Technology，第131卷，文章编号106485，2021年。[4] V. Cosentino、J. L. C. Izquierdo 和 J. Cabot，《使用 GitHub 进行软件开发的系统映射研究》，IEEE Access，第5卷，第7173–7192页，2017年。[5] M. Meli、M. R. McNiece 和 B. Reaves，《Git 能有多糟？公开 GitHub 代码仓库中的秘密信息泄露特征》，2019年网络与分布式系统安全研讨会论文集，2019年。[6] Y. Wu、Y. Zhang、T. Wang 和 H. Wang，《探索 GitHub 上开发者活动与个人资料图片之间的关系》，载于第11届亚太互联网软件研讨会论文集，2019年，第1–10页。[7] B. Lazarine、S. Samtani、M. Patton、H. Zhu、S. Ullman、B. Ampel 和 H. Chen，《识别科学网络基础设施中的存在漏洞的 GitHub 代码仓库和用户：一种无监督图嵌入方法》，载于2020年IEEE情报与安全信息学国际会议（ISI）。IEEE，2020年，第1–6页。[8] S. Oukemeni、H. Rif̂a-Pous 和 J. M. M. Puig，《微博客在线社交网络中的隐私分析：一项调查》，《ACM计算调查》（CSUR），第52卷，第3期，第1–36页，2019年。[9] J. Henriksen-Bulmer 和 S. Jeary，《重新识别攻击——系统性文献综述》，International Journal of Information Management，第36卷，第6期，第1184–1192页，2016年。[10] G. Beigi 和 H. Liu，《社交媒体隐私调查：识别、缓解与应用》，ACM数据科学汇刊，第1卷，第1期，第1–38页，2020年。[11] SecurityWeek。（2022年）《泄露的三星源代码中发现数千个密钥》。访问日期：2022年12月22日。[在线]。网址：https://www.securityweek.com/thousands-secret-keys-found-leaked-samsung-source-code [12] GitGuardian。（2022年）《2022年机密信息扩散现状》。访问日期：2022年12月22日。[在线]。网址：https://res.cloudinary.com/da8kiytlc/image/upload/v1646148528/GitGuardian StateOfSecretsSprawl2022.pdf [13] T. Segura。（2022年4月）《你的代码泄露了吗？》。访问日期：2022年12月22日。 [在线]。网址：https://www.infosecurity-magazine.com/blogs/has-your-code-leaked/ [14] H. Schwartz-Chassidim、O. Ayalon、T. Mendel、R. Hirschprung 和 E. Toch，《在社交网络上发布内容的选择性：隐私顾虑、社会资本与技术素养的作用》，Heliyon，第6卷，第2期，文章编号e03298，2020年。[15] S. Ostendorf 和 M. Brand，《在线隐私相关决策的理论概念化——引入三方自我披露决策模型》，Frontiers in Psychology，第13卷，2022年。[16] N. E. D´ıaz Ferreyra、S. Ostendorf、E. A¨ımeur、M. Heisel 和 M. Brand，《ENAGRAM：一款用于评估 Instagram 预防性助推的应用》，载于2022年欧洲可用安全研讨会（EuroUSEC ’22）论文集。ACM，2022年，第53–63页。[17] B. Liu、M. Ding、S. Shaham、W. Rahayu、F. Farokhi 和 Z. Lin，《机器学习遇上隐私：调查与展望》，《ACM计算调查》（CSUR），第54卷，第2期，第1–36页，2021年。[18] Y. Gu、Y. Yao、W. Liu 和 J. Song，《我们知道你在哪里：基于位置的社交网络中的家庭位置识别》，载于第25届计算机通信与网络国际会议（ICCCN ’16）。IEEE，2016年，第1–9页。[19] C. Comito，《NexT：基于位置的社交网络下一地点预测框架》，Knowledge-Based Systems，第204卷，文章编号106205，2020年，第。[20] A. Garain、B. Ray、P. K. Singh、A. Ahmadian、N. Senu 和 R. Sarkar，《GRA Net：一种基于面部图像进行年龄和性别分类的深度学习模型》，IEEE Access，第9卷，85689，2021年，85672页。– 出版商：IEEE。[21] T. Vo、T. Nguyen 和 C. T. Le，《使用深度卷积神经网络进行种族识别》，Symmetry，第10卷，文章编号564，2018年，第11期。[22] N. Imtiaz、J. Middleton、J. Chakraborty、N. Robson、G. Bai 和 E. Murphy-Hill，《探究性别偏见对 GitHub 的影响》，载于 *2019* 年 *IEEE/ACM*第*41*届国际软件工程会议（*ICSE*）。IEEE，2019年，第700–711页。[23] O. B. Newton 和 J. Song，《开源软件项目成员变动中的性别差异建模》，*arXiv*预印本 *arXiv:2206.08485*，2022年。[24] A. Caliskan-Islam、R. Harang、A. Liu、A. Narayanan、C. Voss、F. Yamaguchi 和 R. Greenstadt，《通过代码文体计量对程序员去匿名化》，载于第*24*届*USENIX*安全研讨会（*USENIX Security ’15*），2015年，第255–270页。[25] B. Alsulami、E. Dauber、R. Harang、S. Mancoridis 和 R. Greenstadt，《使用基于长短期记忆的网络进行源代码作者归属判定》，载于欧洲计算机安全研究研讨会。Springer，2017年，第65–82页。[26] M. Abuhamad、T. AbuHmed、A. Mohaisen 和 D. Nyang，《大规模且与语言无关的代码作者识别》，载于*2018*年*ACM SIGSAC*计算机与通信安全会议（*CCS ’18*）论文集，2018年，第101–114页。[27] H. Krasnova、S. Spiekermann、K. Koroleva 和 T. Hildebrand，《在线社交网络：我们为何披露信息》，《信息技术杂志》，第25卷，第2期，第109–125页，2010年。[28] T. Morlok，《分享并非关怀：外部隐私在社交网站用户信息披露行为中的作用》，载于第*20*届太平洋亚洲信息系统会议（*PACIS 2016*）论文集。AIESeL，2016年。[29] H. Krasnova、O. G¨unther、S. Spiekermann 和 K. Koroleva，《在线社交网络中的隐私顾虑与身份》，《信息社会中的身份》，第2卷，第1期，第39–63页，2009年12月。[30] J. Salminen、S.-g. Jung 和 B. J. Jansen，《在线用户研究建议》，载于国际人机交互会议。Springer，2021年，第127–146页。[31] S. Barth、M. D. de Jong、M. Junger、P. H. Hartel 和 J. C. Roppelt，《检验隐私悖论：具备技术知识、隐私意识和经济资源的用户的在线隐私与安全行为》，《远程信息处理与信息学》，第41卷，第55–69页，2019年。[32] B. Kim 和 D. Kim，《理解社交网络网站用户披露行为的关键前因：隐私悖论》，《可持续发展》，第12卷，第12期，第5163页，2020年。[33] D. Wilkinson、M. Namara、K. Patil、L. Guo、A. Manda 和 B. Knijnenburg，《追求透明度与控制：社交媒体广告解释的分类》，载于第*54*届夏威夷国际系统科学会议（*HICSS*）论文集，2021年，第763页。[34] N. E. Gold 和 J. Krinke，《软件代码库挖掘中的伦理问题》，《实证软件工程》，第27卷，第1期，第1–49页，2022年。[35] M. Vidoni 和 N. E. Diaz Ferreyra，《我该参与吗？为研究招募参与者而挖掘软件代码库的隐私风险》，载于第*1*届实证软件工程研究参与者招募国际研讨会（*RoPES’22*），2022年。[36] M. Humbert、B. Trubert 和 K. Huguenin，《相互依存隐私调查》，《*ACM*计算调查》（*CSUR*），第52卷，第6期，2019年10月。[37] J. M. Such 和 N. Criado，《社交媒体中的多方隐私》，《*ACM*通讯》，第61卷，第8期，第74–81页，2018年。[38] A. Krause、J. H. Klemmer、N. Huaman、D. Wermke、Y. Acar 和 S. Fahl，《海报：意外泄露秘密信息——防止和补救秘密信息泄露的策略》，载于第43届IEEE安全与隐私研讨会（IEEE S&P），2022年5月。[39] K. Renaud、M. Dupuis 和 R. Searle，《网络安全遗憾：我有过几次……但我并不后悔》，载于新安全范式研讨会（*NSPW ’22*），美国新罕布什尔州，2022年。[40] A. Danilova、A. Naiakshina、S. Horstmann 和 M. Smith，《你真的会编程吗？为程序员设计和评估在线调查筛选问题》，载于*2021*年*IEEE/ACM*第*43*届国际软件工程会议（*ICSE*）。IEEE，2021年，第537–548页。