Understanding the Perception and Awareness of Education Technologies' Privacy and Security Issues

Rakibul Hasan
Arizona State University
Tempe, Arizona, USA
rakibul.hasan@asu.edu

ABSTRACT

The rapid growth of education technologies (EdTech), accompanied by large-scale data collection with obscure policies for use and sharing, poses numerous privacy and security risks to users. While policymakers and the research community struggle to catch up with EdTech's fast development, adequate risk awareness among the users and other stakeholders can go a long way in ensuring proper use and initiating movements against invasive technologies. Past studies surveyed users to understand their risk perceptions, but they predominantly focused on students, who may have little understanding of potential privacy risks (e.g., young students) or control over EdTech's deployment. Moreover, while past studies repeatedly identified a general concern about surveillance and data collection, studies examining users' perceptions of and experiences with specific EdTech apps remain scarce.

Toward a broader and more nuanced understanding of risk perception, experience, and awareness involving a larger population, this paper reports findings based on large-scale online datasets: 9M Twitter posts and 0.5M Reddit posts related to EdTech. We implemented a novel pipeline to automatically identify posts related to EdTech's privacy and security issues; overall, we detected a low level of awareness about these issues among the examined population. Through manual annotation of more than 7,000 tweets and qualitative analysis of 186 Reddit posts, we surfaced concerns people expressed about *specific* technologies.

Moreover, we identified several EdTech stakeholder groups (e.g., educators) from online profiles to examine their perspectives. Most educators cared for students' privacy—affirming past studies; but importantly, we also found educators to be concerned about their own privacy. In contrast, some educators regarded students' concerns for privacy as unwarranted and advocated for EdTech's *increased* use. We discovered power asymmetry and tensions between stakeholders (e.g., instructors and administrators) regarding EdTech's deployment and use, which may exacerbate the risks for students. Finally, we surfaced threats to the privacy the people around EdTech users (e.g., family members), as many apps conduct multi-device tracking and home network scanning. We reflect on these findings and make recommendations for future research in this fast-growing domain.

Proceedings on Privacy Enhancing Technologies 2023(4), 269–286

https://doi.org/10.56553/popets-2023-0110

KEYWORDS

education technology, privacy, surveillance, user perception

1 INTRODUCTION

Education technologies (EdTech) are being integrated into every step of educational processes [129]. The Covid-19 pandemic imposed an abrupt move to virtual settings, and EdTech applications (apps) were often deployed bypassing proper vetting processes [26, 44, 60], and without adequate IT infrastructures and training resources for educators and IT personnel [11, 20, 67, 89, 118] to properly configure, maintain, and use those tools. EdTech's rapidly growing market (estimated to exceed $404B by 2025 [63]) hints at their continued and expanded use in the post-pandemic world [44, 75, 92].

EdTech apps, like most other digital tools, mine massive amounts of user data [1, 46], and analyze them for profiling, evaluation, and comparison [58, 129]. Moreover, educational institutes are increasingly offloading data collection, storage, and maintenance responsibilities to for-profit entities [44, 45] that may not be subject to data protection laws like FERPA (Family Educational Rights and Privacy Act) [11, 44]. Consequently, the marketplace for EdTech data is growing [87, 105, 129], and researchers and privacy activists are expressing increasing concerns regarding the security, privacy, and safety risks from EdTech [79, 82, 87, 129].

Are the stakeholders—students, educators, parents, school officials, and developers—aware of and concerned about these risks? Public awareness and resistance to surveillance technologies can prevent their normalization. For example, parental opposition to inBloom—a multi-state program to collect data from educational institutes and share them with private companies—prevented it from materializing [40].

Past research investigated risk perception and awareness by EdTech users [10, 19, 68, 70–72, 83, 115, 120, 124, 126]; most of them, however, focused only on students; a few included other stakeholders [70, 71, 83, 115]. Furthermore, findings varied between populations—ranging from students being unaware of EdTech's use [19], to being aware but unconcerned about data collection [124], to demanding transparency and greater control over their data [116, 126]. Such inconsistencies might have resulted from framing effects, cultural variations, and biases [19, 49, 54, 124]. Finally, while prior research repeatedly discovered general anxiety around data collection, they rarely spotted specific privacy concerns from certain tools or experiences of privacy violations, presumably, because surveying large populations or iterating over a comprehensive list of potential risks could be prohibitively expensive.

This paper aims toward a broader and simultaneously nuanced understanding of people's perceptions of and experiences using

EdTech. We complement prior survey-based research by analyzing large-scale data from two popular platforms: Twitter and Reddit. These platforms offer diverse perspectives on current socio-technical issues and have been used in numerous studies on privacy and security (e.g., [7, 64, 65, 78, 98, 101]). From Twitter and relevant subreddits (e.g., */r/teachers*, */r/students*, and */r/edtechhelp*), we collected 11M tweets and 0.5M Reddit posts that contained EdTech-related keywords; these posts were made between January 2008 and February 2022. We analyzed these data to identify posts about the privacy and security of EdTech, as well as mapped the post authors to EdTech stakeholders (e.g., educators) to understand risk perception and awareness of different groups.

Trend analyses revealed that tweets about EdTech in general and tweets specific to EdTech's privacy/security issues peaked after the Covid-19 lockdown, but the latter had a much smaller scale than the former. This result suggests that the examined population may lack risk awareness despite news on data breach [35, 95, 131] and abuse [24, 42, 59, 103] are on the rise and frequently posted on Twitter. Although small in relative terms, we identified more than 7,000 tweets discussing EdTech's privacy and security issues. Manual annotation of the tweets revealed that people expressed generic concerns about surveillance and profiling as well as specific risks they *perceived* or *experienced* while using certain technology (e.g., sharing video during online classes). Critically, more tweets expressed specific concerns than generic concerns; the former also reached a larger audience (through retweets and likes) than the latter, especially, after the Covid-19 lockdown. We broke down the above findings across stakeholder groups: we found that, among other results, academics engaged only moderately in discussing EdTech's potential risks and creating public awareness.

Reddit posts, compared to tweets, are more descriptive, and their qualitative analyses offered nuanced understandings of the issues at hand. The key discoveries include educators' concerns regarding self-privacy being risked by EdTech, and tensions among educators and school administrators regarding EdTech's deployment and use. Additionally, we surfaced *novel* threats: not only students but also their family members may experience privacy violations as EdTech enter personal spaces (e.g. home) and conduct multi-device tracking through stealthy behaviors; such threats may impact marginalized populations disproportionately due to their greater reliance on school-issued or shared devices [55]. Finally, we found that some educators—who maintain that students' aversion to using EdTech due to privacy risks contradicts their use of social media and that EdTech's benefits outweigh their risks—advocated for EdTech's *increased* deployment and use.

In sum, this paper provides a broader view of EdTech's privacy/security issues as *perceived* and *experienced* by a large population, over a 12-years timeline. Instead of solicitation, we relied on voluntarily provided data that are less likely to be biased; we expect the findings to portray an accurate picture of people's concerns and awareness. Based on these findings, we offer guidelines for future research in this rapidly emerging space.

2 LITERATURE REVIEW

2.1 Data mining by EdTech

As students interact with technologies for educational purposes, their activities are logged, mined, and analyzed for patterns. For example, web-based learning management systems collect every click, page transition, and contextual information (e.g., date and time) [58, 89]. Remote tutoring and proctoring tools continuously collect audio and video data; and mobile-based apps monitor students' activities in and out of school campuses and in contexts unrelated to educational activities [77, 89, 129].

2.2 Privacy and security risks from EdTech

EdTech's massive-scale data collection is accompanied by an increasing number of data breaches at educational institutes [79], secondary uses of students' data [82, 87, 129], and an expansion of data marketplaces [4, 43, 107]. Harms from data breaches are explicit: leaking identity and other sensitive attributes of students have led to tax fraud, unlawful extortion of money, and even the death of a student [79]. Behavioral data mining may appear less dangerous, but profiling and identifying information can be inferred from such data [37, 58]. Moreover, constant monitoring of students' activities may have gradual, but more insidious effects of destroying independent and creative thinking capabilities [129]. Integrating applications from multiple vendors and third-party developers through APIs (Application Programming Interface) [14, 41, 51] not only subject more people to the existing risks (by scaling up data volume and parties that can access the collected data) but also create new risks since triangulating cross-platform data facilitates re-identification (even if the original data were deidentified) and profiling [130]).

2.3 User awareness and risk perception

Researchers have investigated the level of awareness and concerns about EdTech's security and privacy issues by surveying students; unfortunately, some findings were inconsistent across samples. Jones et al. reported that students lack awareness of both data collection for learning analytics and their purposes [72]. Phu et al. found students to be aware of their interactions with e-learning portals being recorded but carefree [124]. Contrarily, many researchers reported that students felt threatened by EdTech and wanted more control over their data and a consenting mechanism before data collection and analyses [116]. Students' risk perception also differed based on the amount and the type of data being collected as well as the purposes for data collection [10, 120, 126]. Very few studies investigated other stakeholders such as educators [70, 71] and developers [115]; both groups generally were supportive of better privacy for students but fell short of properly understanding or carrying out their responsibilities to achieve that goal [71]. A study based on large-scale, unsolicited opinions on EdTech from the public could fill the gaps in the literature and resolve conflicts in findings.

2.4 Preventive measures

Researchers have proposed guidelines for trustworthy and privacy-respecting implementations of EdTech [39, 102]. Unfortunately,

contributions from the security and privacy research community to realize those guidelines and subsequently audit them have been scarce, except for a few studies [13, 18, 52, 53, 58, 85]. Policymakers' efforts to protect consumers' rights and interests have also been severely lacking [11, 44]. A comprehensive and nuanced understanding of the level of public awareness, perception, and experiences may help propel research efforts to better inform the users and the public, establish policies to regulate data collection and usage, and innovate technical means to mitigate the risks.

2.5 Social sensing for privacy and security research.

Online platforms (e.g., Twitter and Reddit) have become major venues for discussing socio-technical issues. Twitter provides an open platform for researchers, academics, and practitioners to discuss various aspects of research and teaching [57], and Reddit offers role-specific forums for different EdTech stakeholders. These platforms were instrumental in acquiring threats intelligence and understanding public attitudes toward security/privacy issues at scale (e.g., [64, 78, 98]). Furthermore, non-security experts also participate online in such discussions [114], and critically, large-scale public discussions have led to desirable outcomes such as enhancing product security [62] and banning invasive remote proctoring apps at educational institutes [23].

3 METHODS OF TWITTER DATASET COLLECTION AND ANALYSIS

In this paper, we take a quantitative approach and analyze large-scale data from Twitter and Reddit (Section 5) to identify privacy and security concerns regarding EdTech. This approach has been adopted in numerous prior research investigating such concerns in different domains (e.g., [56, 64, 78, 90, 93, 98, 125]). While interview-based studies allow one to identify rich and nuanced content, collecting data from a large sample is not usually possible in those settings. More importantly, we view interview-based and observational studies are complementary to each other, rather than competitors: the latter can be used to confirm and extend the findings from the former employing a larger sample. Since many prior studies interviewed or surveyed EdTech users and stakeholders to identify concerns, we focus on whether those findings apply to a larger sample including diverse stakeholder groups.

This section details Twitter data collection and analysis methods; we combined methodological steps from past research (e.g., [56, 90, 93, 125]) with our own to create a novel iterative process of data analyses. Results from automated analyses were manually reviewed and validated by the authors as well as independent annotators to establish reliability.

We used Tweepy [27] to query Twitter's public API (Application Programming Interface) for academic research [123] which allows searching for tweets posted *anytime* since Twitter's inception. We compiled a list of hashtags, account names, and keywords (described below) that were used as search terms. The collected tweets included id, text, number of retweets and likes, and authors' profile information. We collected original tweets, as well as quotes and replies, but omitted retweets as they do not add information. After collection, tweet content and profile descriptions were pre-processed for topic clustering and sentiment analyses. We describe these steps below.

3.1 Collecting search terms

Both Twitter handles and hashtags associated with EdTech apps were used to search tweets. First, we listed apps from several authoritative sources: the most popular 100 EdTech apps on Android and iOS marketplaces, apps recommended by *Common Sense Education*,[1] top 100 apps ranked by SimilarWeb [6] that has been used in numerous prior research (e.g., [36, 119]), and other popular online sources (see Appendix A.1). After removing duplicates, apps without an associated Twitter profile (e.g., Kiddopia), and apps that were not directly used for educational purposes (e.g., *Here Comes the Bus*, which is used for school bus scheduling), we were left with 45 apps and corresponding Twitter handles (full list in Appendix A.2). Hashtags were collected from sources listed in Appendix A.1, as well as from recommendations made by Twitter when we searched with already collected tags and Twitter handles; we stopped searching after reaching convergence. In total, we collected 54 unique hashtags (e.g., #edtech, #eddata, and #remoteteaching), see the full list in Appendix A.2. To further expand the set of search terms, the collected hashtags and handles were concatenated with the following keywords: surveillance, surveil, track, risk, vulnerable, threat, security, privacy, student, secure, insecure, private, dataveillance, invasion, and invasive, totaling 1,365 search terms.

Next, we collected tweets that contained any of the hashtags, handles, or search phrases. Note that we collected tweets *containing* the handles, rather than posted by the associated Twitter account; the former includes tweets from consumers or the public and expresses opinions about the apps or their developers, while the latter includes tweets that are posted by companies and usually intended for advertisements. In total, we collected 11,315,305 unique tweets posted from 33,27,801 unique accounts. Only English tweets (N = 9,745,446) were analyzed in this paper.

3.2 Clustering Twitter users

To understand who discusses EdTech on Twitter, we mapped users to stakeholder groups (e.g., educators) by clustering profile descriptions, as detailed below.

Pre-processing and language detection of user profiles. First, texts describing profiles were pre-processed to remove numbers, special characters, and stopwords [33]. The remaining words were lemmatized to obtain their base or dictionary form [109]. To determine the profile language, the pre-processed texts were fed into three popular language detection models: Langid [30], Google's Compact Language Detector [29], and a fasttext pre-trained model [73]. The language was determined based on majority voting. We identified 1,698,096 profiles in English, 142,066 in Spanish, 91,536 in French, and 42,593 in German; 1,235,386 profiles remained undetermined and the rest were in other languages. Only English profiles were used in subsequent analyses. We could not determine the language of 1,235,386 ($\approx$ 11%) profiles, either because the profile descriptions were empty or contained only emojis or URLs; these

[1] Common Sense Education is part of *Common Sense Media* [3] which is a leading organization in reviewing and rating technologies aimed at children

profiles were discarded from further analyses. We note that empty profile is not uncommon on Twitter; prior research has found that more than 20% of the profiles can have empty descriptions [96, 111].

Clustering pre-processed profiles. The pre-processed profiles were embedded in a 100-dimensional vector space where each profile was represented by a point and 'similar' profiles can be grouped to form clusters. For the embedding, we trained fasttext [16] following the skip-gram model, which places texts that are *semantically* (rather than syntactically) similar next to each other, and thus similar profiles get embedded in nearby points [17]. We adapted Sia et al.'s implementation of a vector clustering algorithm to cluster the profiles [113].

Since clustering is an unsupervised algorithm, the appropriate number of clusters needs to be determined experimentally. We used the *coherence score*, one of the most popular metrics [106], to determine the number of clusters. First, the clustering algorithm was executed multiple times with different cluster numbers (as a parameter): 10, 20, 30, 40, 50, 80, 100, 120, and 150. Each time, the individual cluster's coherence score was computed by averaging the pairwise cosine similarity among the 10 most important words in that cluster (i.e., words representing the cluster center [113]). We settled on 40 clusters since this structure achieved the highest average coherence score of 0.69 (range: 0–1) across the specified number of clusters [106]. Following cluster identification, each profile was assigned to the cluster with which it had the highest cosine similarity.

3.3 Identifying tweets expressing concerns

The sheer volume of online data, with only a tiny fraction of it being related to privacy and security (e.g., .12% [90] to .5% [93] for app reviews) makes manual analyses infeasible. Thus, following prior works [56, 90, 93, 125], we combined automated and manual steps to identify and categorize relevant tweets. Figure 1 depicts the analysis pipeline.

First, tweets were identified using keyword search (see below) and then clustered using topic modeling. Next, we performed topic modeling to distinguish tweets related to EdTech from other topics. Finally, we conducted sentiment analysis and discarded tweets that conveyed *positive* or *neutral* sentiment as they were unlikely to express concerns. The remaining tweets were manually annotated. The following sections detail these steps.

Keyword search. Following prior work [90, 93], we identified tweets that contained keywords related to privacy and security. The following keywords were compiled from literature review [56], news articles, and our observation of tweets: *privacy*, *security*, *student*, *surveillance*, *surveil*, *dataveillance*, *vulnerable*, *vulnerability*, *infosec*, *abuse data*, *misuse data*, *aggregation*, *secondary use*, *intrude*, *intrusion*, *appropriation*, *appropriate*, and *specific purpose*.

Tweet topic clustering. Tweets were pre-processed by tokenizing (using an open-source library [117]), replacing emoticons with corresponding words [32], converting contractions to full forms [31], removing special characters and stopwords, and lemmatizing the remaining words. Next, tweets were converted into vectors using fasttext [16] and then clustered [113]. As before, we experimented with different cluster numbers: 5 8 10 15 20 25 30, and identified 10 clusters based on the coherence score (0.71).

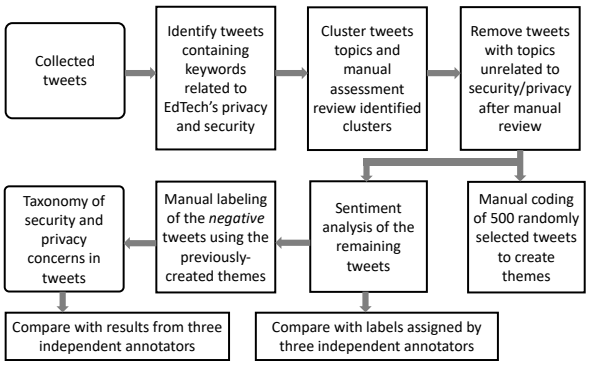

Figure 1: Steps to create a taxonomy of privacy and security concerns about EdTech from the tweet dataset, and evaluate it with randomly sampled data labeled by independent annotators.

Manual assessment of the clusters. To evaluate the cluster quality, we reviewed 10 representative words in each cluster. As Table 1 shows, clusters 3, 8, and 10 were difficult to interpret from the words alone without more context. Hence, we manually reviewed a random sample of 50 tweets from each cluster, a total of 500 (all tweets in clusters 7, 9, and 10 were included as they contain less than 50 tweets).

This review step served two additional purposes: it helped identify clusters irrelevant to our research and served as a basis for creating a taxonomy of concerns that we later used to label all tweets (Section 3.4). Tweets in clusters 5, 7, 8, and 9 were removed from further analyses; almost all tweets in cluster 5 were related to security/privacy risks in Bring Your Own Device paradigm (#BYOD) in employment contexts, tweets in cluster 7 were related to two incidents involving Facebook [38, 91], and clusters 8 and 9 contained tweets on random topics.

Sentiment analysis. We trained a supervised machine learning model for sentiment analysis using two datasets: the open dataset [88] containing 6.3 million tweets labeled as *positive*, *negative*, or *neutral* by the *AWS Comprehend API* [8], and the SemEval dataset [104] containing 50,000 tweets labeled similarly by crowd workers. The combined dataset was divided into training (80%) and evaluation (20%) splits. We trained a fasttext (multi-class) classification model on the training split. To improve the classification accuracy, we used bi-grams from pre-processed tweet texts (e.g., to correctly interpret phrases like "not happy"). The model yielded 91.3% classification accuracy on the test split which is higher than a state-of-the-art model trained only on the SemEval dataset [25] (we experimented with a popular lexical-based model named Vader [66], but it performed worse in the test split.)

The trained model was used to infer the sentiment of tweets in our dataset. Only tweets predicted as *negative* were used in subsequent analyses as positive and neutral tweets were unlikely to express concerns or risks.

Cluster	#Tweets	Keywords
1	12,735	edtech, cybersecurity, dfir, skillsgap, nlproc, ingramtechsol, systemsthinke, stcenturyskill, thecybersecurityplace, cybint
2	1,205	datum, protect, sotu, schoolprivacyzone, paigekowalski, datadrivesresult, rachelbarrer, takectrl, replukemesser, benjaminbherold
3	6,447	username, share, talk, story, daily, event, awesome, tune, david, advisor
4	702	school, lifeleader, schoolchildren, safeschool, studentvoice, chooseprivacy, schooldistrict, coetail, powerschool, indyschool
5	37,108	security, byod, mobile, device, enterprisesecurity, byodsecurity, fiercemobileit, securityrisk, dellsecurity, symantec
6	104,699	privacy, student, news, washington, priva, myname, privac, stud, studentright, priv
7	48	facebook, internship, flaw, austrian, lose, harvard, expose, highlight, messenger, cancel
8	699	conclude, cpdp, myriad, minefield, txcto, asugsv, sotn, bettertogether, nuisance, rundown
9	3	neighborhood, georgetown, auburn, nazi, chattanooga, tulsa, knife, sixth, uconn, franklin
10	12	gauge, pseudonym, tyranny, culprit, sarcasm, ludicrous, nefarious, pedophile, disingenuous, fwiw

Table 1: The number of tweets and most representative 10 words in each cluster.

3.4 Manual annotation of tweets expressing concerns.

This section details our process of categorizing the tweets identified in the previous step.

Creating categories. During our review of 500 tweets (Section 3.3), we also categorized their content type. For example, some tweets expressed general privacy concerns and fear of data abuse, while others mentioned specific concerns (e.g., visual data privacy during a virtual meeting) or risks from a specific technology. Many tweets discussed privacy/security incidents (such as a data breach) and (the lack of) laws or policies to regulate EdTech. Table 2 lists these categories.

Applying the categories. In this step, we applied the categories to all tweets expressing concerns and simultaneously validated this taxonomy of concerns. But, instead of crowdsourcing, we decided to label the full dataset ourselves based on the following observations.

While labeling the first 500 tweets, we noted that many tweets could not be fully interpreted from the text alone as they may contain only URLs, news headlines, or replies to another tweet, and required visiting Twitter to gather contextual information. Moreover, many tweets referred to specific EdTech providers by Twitter handle contained technical terms (e.g., bias in AI-based remote proctoring apps), or referred to specific people (e.g., a privacy researcher or activist) that are likely to be unfamiliar to crowd workers. Furthermore, replies to other tweets required us to read the whole tweet thread to properly categorize them. Finally, many tweets were duplicates of others (e.g., a news article shared by many); we had to scan all tweets multiple times to spot such duplicates. Thus, crowdsourcing the labeling task in our case could produce unreliable results.

Note that the categories were created and applied based on the presence of certain entities (e.g., Twitter handles, links to news articles, and technical phrases referring to privacy or security vulnerabilities), and thus unlikely to be influenced by subjective biases. Four tweets did not fit any of the categories; hence, we created a new category to label those: *tech support* (Table 2). All other tweets were annotated by the previously created categories, demonstrating the comprehensiveness of the taxonomy. To establish further reliability on the results, we compared our annotations with annotations created by independent annotators on random tweet subsets (Section 4.5).

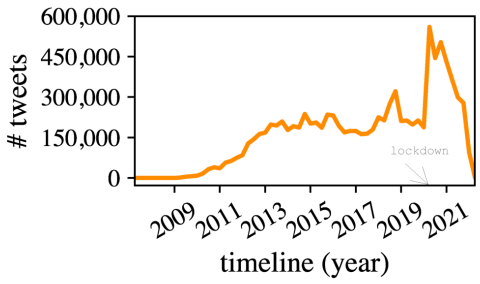

Figure 2: The trend in English tweets related to EdTech.

4 FINDINGS

We collected 11,315,305 tweets from 3,327,801 users; 9,745,446 (86%) were in English (from 1,698,096 users). Figure 2 plots the tweet trend: EdTech-related discussions peaked around mid-2020 once educational institutes went remote. The following sections report findings based only on English tweets and the users who posted them.

4.1 User groups discussing EdTech on Twitter

As Section 3.2 explained, we identified 40 profile clusters, manually reviewed them, and merged similar clusters, resulting in seven larger groups that resembled EdTech stakeholders (Table 3). We note the semantic similarities among the words representing each group; the non-words were hashtags or Twitter handles for relevant entities (e.g., universities or research organizations, we confirmed

Label	Explanation
Privacy generic	Tweets expressing concerns around EdTech's surveillance, data collection, and misuse.
Privacy specific	Tweets referring to privacy/security concerns or experiences specific to a context or a technology.
Privacy incidents	Tweets discussing privacy-violating incidents such as data breaches.
Policy	Tweets discussing the applicability or (in)adequacy of laws and policies to protect EdTech consumers.
Article	Tweets sharing a news or journal article related to EdTech. The most common article topics included security and privacy challenges posed by EdTech and the current status of laws and policies for consumers' protection.
Other student privacy	Tweets discussing students' privacy issues or concerns but not in the context of EdTech.
Other privacy	Tweets that discussed generic security or privacy concerns and issues with other technologies.
Info branding	Tweets advertising online courses, tutorials, or other resources that may help mitigate EdTech's privacy/security issues. These tweets were usually posted by the creator of the advertised items.
Tech branding	Tweets advertising software products created to prevent security/privacy risks in the context of using EdTech.
Exchanging knowledge	Tweets posting questions (e.g., how to configure some app), discussion prompts, or tips related to EdTech.
EdTech generic	Tweets about generic questions, discussions, or complaints regarding education apps.
EdTech support	Tweets advocating technologies' deployment or usage in education.
Other	Tweets that were unrelated to education technologies or security/privacy threats from technology.

Table 2: Tweet categories and their explanations.

this by manual searches) and thus represented contextually related information.[2]

Surprisingly, we did not find any cluster representing students. The *Academic techie* group contained "student" among the top 10 words; we manually reviewed 100 profiles randomly selected from this group, but only one of them was a student profile; the rest included the word "student" in contexts other than group identity (e.g., "love my students"). Similar searches based on keywords that could be present in students' profiles (e.g., "learner" and "researcher") yielded only 0–2% of profiles belonging to students. Thus, we did not create a separate group for students.

4.1.1 The number of tweets across user groups. *Academic techie* posted the highest number of EdTech-related tweets (1,791,852), followed by *Personal brand* (1,144,087). *Influencer*, *Academic*, *News*, *Personal*, and *Techie* posted 948,885, 723,787, 213,018, 161,053, and 141,259 tweets, respectively. Based on the keywords describing these groups, people who build software and analytical systems (possibly including EdTech) and advocate technologies' integration in pedagogical processes seem to be the most vocal on Twitter regarding EdTech. Relatively fewer tweets were from advertising accounts (likely maintained by online influencers [57, 108]); this result is unsurprising since we excluded tweets from EdTech providers' official profiles.

4.2 Tweets containing privacy or security-related keywords.

We identified 216,667 tweets, posted from 90,579 unique accounts, that contained at least one keyword related to privacy or security. Twitter's terms of service [122] prohibit the use of tweets posted from accounts that were later deleted. We identified 53,009 tweets that were posted by such accounts (N=26,772), i.e., the accounts do not exist anymore; thus, we discarded those tweets. Among the remaining 163,658 tweets, 108,666 tweets were posted by the user groups identified above. Table 4 shows the number of tweets containing privacy/security-related keywords along with the ratio of such tweets to the total number of tweets from each group. In the following sections, we report how many of these tweets actually expressed concerns related to EdTech's privacy and security issues, break it down across user groups, and identify EdTech apps that were most frequently mentioned in those tweets.

4.3 Tweets related to EdTech's privacy and security.

4.3.1 Identifying and labeling tweets expressing concerns. Sentiment analysis identified 7,309 tweets with *negative*, 148,082 with *neutral*, and 8,267 with *positive* sentiment. We manually categorized the concerns expressed by the *negative* tweets.[3] As Table 5 shows, 4,281 tweets were assigned to different concern categories. The rest 3,028 tweets were duplicates of other tweets. For example, variants of the news regarding the Electronic Frontier Foundation accusing Google of tracking students [103], and students being suspended for refusing to wear a school-issued RFID tracker [80] appeared in our dataset more than 400 times. Notably, more tweets mentioned specific risks or expressed concerns regarding certain technology than tweets expressing general concerns, and four tweets advocated EdTech's *increased* use. Also note that only 424 (< 6%) tweets were false positives (i.e., unrelated to EdTech's privacy/security), demonstrating the reliability of our automated selection method.

4.3.2 Most important keywords and entities that were associated with the identified concerns. Table 6 shows the most important 25 keywords and entities (based on *Term Frequency Inverse Document Frequency* or TF-IDF score [34]) that occurred in tweets expressing either generic or specific privacy concerns, or common to both. Note that many tweets in the latter group referred to specific privacy violations or invasive behaviors by certain technologies, and mentioned (i.e. contained Twitter handles) 129 unique EdTech providers

[2] Appendix A.3 lists all 40 clusters separately.

[3] As discussed in Section 3.3, we did not include tweets with positive or neutral sentiment since they are less likely to include any concerns about privacy issues or complaints about related experiences.

Group name and description	Keywords representing the group	#Users
Academic: profiles emphasizing educator roles in research or teaching institutes.	school, principal, research, researcher, researching, lunduniversity, researched, keeleuniversity, reseach, stfxuniversity, universitysa, simaduniversity, teacher, learner, paraeducator, director, professor, assistant, associate.	195,762
Academic techie: educator profiles advocating EdTechs' adoption in pedagogical processes or showcasing related certifications.	education, technology, learn, tech, passionate, student, techology, msuepet, specialist, certified, trainer, integrationist, nearpodcertified, classcraft, certifiedmie, edtechie, educator, edugator.	133,749
Techie: profiles describing computer science, engineering, and related professions.	software, engineer, developer, programmer, developper, coldfusion, softwares, getpostman, klarna, couchbase, data, scientist, datum, analytics, analytic, visualization, analysts, datavisualization, boozallen, linkeddata.	54,057
Personal brand: profiles claiming leading positions or expertise in tech or business fields.	founder, entrepreneur, entreprise, dealmaker, wantrepreneur, angellist, capitalfactory, businessperson, lead, professional, provide, development, training, focus, expert, develop, create, inspire, empower, mission, strive.	164,168
Influencer: profiles maintained by celebrities, influencers, or entrepreneurs and may be used for product endorsements.	online, website, blogspot, prospectus, service, company, solution, multinational, enabling, follow, twitter, account, official, instagram, facebook, business, social, digital, market, strategist	191,884
News: profiles describing journalist or editorial professions.	editor, journalist, fivethirtyeight, primetime, journaliste, wlrn, journalistin, geekwire, journos, thismorne, news, information, event, bring, late, source, daily, issue, cover, relate.	43,134
Personal: profiles highlighting personal relationships.	husband, father, proud, christian, amazing, awesome, marry, married, wonderful, lucky, wife, mother, friend, daughter, sister, mama, mommy, grandmother, momma, grandma.	68,667

Table 3: Profile groups after clustering and merging, 20 most important keywords, and number of Profile in each group.

User group	#Tweets	% total	% group
Academic	8,197	7.5%	1.13%
Academic techie	19,170	17.5%	1.07%
Influencer	44,097	40.5%	4.65%
News	12,314	11.3%	5.78%
Personal brand	20,846	19.2%	1.82%
Techie	4,042	3.7%	2.86%
Personal	2,254	2.1%	1.4%

Table 4: The number of tweets related to privacy and security across user groups (and the percentage of such tweets relative to the total number of tweets and tweets posted by the respective group).

Label	Count	Label	Count
Other student privacy	864	Privacy incident	190
Privacy specific	725	Info brand	131
Other privacy	574	Tech branding	86
Article	494	EdTech generic	76
Other	424	Knowledge exchange	45
Privacy generic	406	EdTech support	4
Policy	262		

Table 5: Distribution of tweets related to EdTech's privacy and security concerns.

(top five: goguardian, chegg, examsoft, edpuzzle, and pearsonvue). In contrast, tweets expressing generic concerns mentioned only 32 unique EdTech providers. Most tweets with *mentions* complained about privacy threats experienced while using the mentioned application.

Above results not only confirm past studies [9, 71] (reporting concerns about invasive data collection and vague privacy policies) but also extended those studies by identifying EdTech apps and providers that users perceived as the most threatening.

4.3.3 Which stakeholder group expressed what concerns? Table 7 breaks down the number of tweets expressing concerns by each stakeholder group. Interestingly, the group with the highest enthusiasm for EdTech, *Academic techie*, also expressed the most concerns. Another surprising finding is that Twitter profiles for self- or business advertisements posted reasonably large proportions of such tweets. To examine any group-specific peculiarities in discussion topics, we report the top 10 words (according to TF-IDF score) from the tweets in Table 7. Most notably, the *Personal* user group went beyond discussing privacy risks from EdTech's surveillance to users' expectations and EdTech providers' accountability.

4.4 Trend and reach of tweets expressing concerns.

To understand the discussion trends on EdTech's privacy issues, incidents, and regulatory policies, we plotted the average number

Category	Most frequent keywords
Common	username, privacy, student, edtech, datum, surveillance, school, issue, teacher, concern, app, kid, security, bad, online
Generic	eddata, parent, education, protect, child, tool, company, time, learn, information
Specific	software, invasion, violation, google, exam, policy, zoom, violate, proctorio, feel

Table 6: Most important keywords and entities in tweets that expressed (generic or specific) privacy concerns.

Label	Author group	Count	Most important keywords
privacy-generic	Academic	33	privacy, student, edtech, surveillance, datum, law, time, platform, issue, security.
	Academic techie	86	privacy, student, edtech, datum, school, surveillance, issue, app, teacher, kid.
	Influencer	29	privacy, student, datum, school, edtech, byod, surveillance, onlinesafety, security, fear.
	News	7	student, mooc, loss, major, drawback, blow, surveillance, people, rep, appropriately.
	Personal brand	25	privacy, student, datum, edtech, eddata, teacher, concern, school, issue, sxswedu.
	Techie	1	(omitted due to the low number of tweets.)
	Personal	6	student, expectation, accountability, understand, secure, threaten, pjnet, keepyourpromise, intrusive, stopcommoncore.
privacy-specific	Academic	80	student, privacy, issue, datum, surveillance, concern, school, violation, teacher, tool.
	Academic techie	125	privacy, student, edtech, school, issue, security, google, surveillance, app, datum.
	Influencer	44	privacy, student, school, facebook, issue, security, concern, kid, reliance, danger.
	Personal brand	34	privacy, student, edtech, datum, surveillance, breach, security, school, issue, software.
	News	16	student, privacy, surveillance, issue, online, username, exam, unfair, global, spark, software.
	Techie	21	security, privacy, exam, student, delete, business, fail, information, account, company.
	Personal	13	student, violate, edtech, zoom, time, assumption, guilt, right, danger, invasion.

Table 7: Tweets expressing privacy concerns and top 10 keywords in those tweets across user groups.

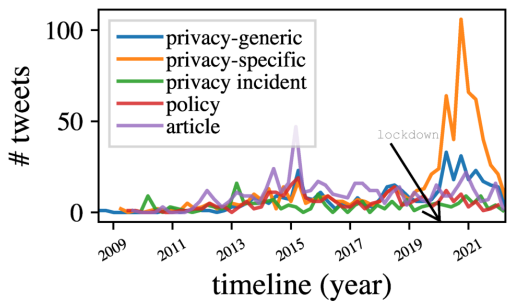

Figure 3: The trend of tweets related to EdTech's privacy/security issues.

of tweets on those topics posted every three months (Figure 3). *Importantly, tweets mentioning specific risks started to increase in number after March 2020 (as the lockdown started due to Covid-19), and peaked at the beginning of 2021, when most educational institutes had already gone fully remote.* Tweets expressing generic concerns followed similar patterns but on a smaller scale. Discussions around regulating EdTech and laws to protect students' privacy, and tweets mentioning privacy incidents experienced occasional small bumps, but seem to be uncorrelated to the recent surge in EdTech's deployment. The number of tweets sharing news articles slightly correlated with the recent surge but had larger spikes in the past.

Next, we looked at whether such discussions were limited to a small group of people. The tweets were posted from 3,577 unique accounts. Only 21 accounts posted more than 10 tweets, and eight accounts posted more than 20 tweets. The maximum number of tweets from a single account was 76, and 75% of the accounts posted only one tweet. Thus, user participation was fairly broad and a few accounts did not dominate discussions.

Finally, we investigated whether such discussions drew attention from the audience (and potentially raised awareness among the public). On average tweets expressing concerns received 4.83 likes and 2.02 retweets—two ways content gets propagated to a new audience on Twitter.[4] Figure 4 shows the number of retweets and likes received by tweets of different categories aggregated over three months. Attention to tweets that expressed specific concerns increased starting from 2020 and peaked near the end of 2021. Surprisingly, generic concerns, while tweeted more frequently after the pandemic began than before (Figure 3), did not receive much attention from the public.[5]

[4]For comparison, for the full set of English EdTech-related tweets, the average number of likes and retweets were 2.19 and 0.84, respectively. For general tweets, retweets usually follow an exponential decay [69] where only 4% of tweets get more than 50 retweets; in previous work, we found that average retweets can range from 3.5–9 based on the user [57]

[5]Due to indexing and ranking, results from APIs do not always match with manual searches (see https://twittercommunity.com/t/tweets-missing-from-twitter-api-v2-recent-search/179031. Thus we might have missed some highly popular tweets.

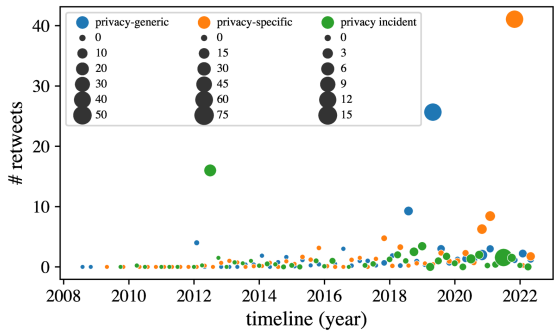

Figure 4: Visibility of tweets expressing concerns and incidents related to EdTech. The size of the dots corresponds to the average number of *likes*.

4.5 Comparing results with data labeled by independent annotators.

To establish further reliability of our findings, we hired three undergraduate students (majoring in computer science) as independent annotators. They were paid at a rate of $14.5 per hour following our institutional guidelines.

4.5.1 Labeling privacy/security-related tweets. The first task was to identify privacy/security-related tweets from a random sample of 3,000 English tweets. The annotators were briefed about the task and shown some positive and negative examples. They then independently labeled the tweets with high consistency (Krippendorff's $\alpha = 0.93$),[6] conflicts were resolved by majority voting.

The annotators identified 47 (1.6%) tweets related to EdTech's privacy/security issues, while our semi-automated analysis identified 2.4% (3,857 out of 163,658) such tweets. Importantly, 43 of the 47 tweets (91.2%) overlapped with the tweets we identified. The seeming paradox of annotators identifying positive samples at a lower rate than the semi-automated system could be a result of too many false positives in the latter case or annotators' unfamiliarity with technical or domain-specific hashtags and keywords, or both. We test this in the next study.

4.5.2 Labeling samples identified by the automated approach. To investigate the precision of our automated analysis mechanism, we conducted another study to estimate the number of false positives it generated. We randomly sampled 1,000 tweets from the set of 3,857 tweets identified as *positive* by this method. As before, we hired three undergraduate students (different than the previous study) to annotate these tweets. We found that, for 912 tweets, at least one annotator agreed with the categorization (i.e., labeled them as *positive* samples), and all three annotators agreed with our classification for 784 tweets. At least two annotators agreed with our classification for 852 tweets; Thus, a majority voting-based scheme indicates 1.48% false positives in our dataset. We inspected the tweets for which annotators differed with us in classification; While we found no strong pattern, tweets including special keywords, hashtags, or Twitter handles, and tweets about the security of BYOD paradigm in school, tweets including URLs to news articles that are not fully comprehensible from the title (such as the collaboration between Facebook and Udacity to train models in privacy-preserving ways and how deploying WiFi 6 protocol in schools may protect privacy) were most prevalent. We note that in general identification rate is less than 1% when crowd workers labeled such data [56, 90, 93, 125]. In our study, we employed undergraduate students majoring in computer science, who presumably are more knowledgeable in the subject matter than crowd workers, and indeed achieved a higher detection rate (1.6% as reported above). Extrapolating this observation, we infer that the false positives in this study also resulted from a lack of domain knowledge and technical terms relevant to the context, since we used such terms in the keyword-based search phase to identify tweets expressing privacy concerns. Additionally, false positives can be filtered out with manual reviews and preferred to false negatives; Thus, our semi-automated analysis pipeline proves to be a valuable one in this domain.

4.5.3 Categorizing concerns expressed in tweets. We randomly selected 100 tweets from each of the following categories: *privacy generic*, *privacy specific*, *privacy incidents*, *policy*, and *article* (Table 2), totaling 500 tweets. The annotators, after familiarizing themselves with the categories and seeing two examples from each category, independently labeled these tweets. Conflicts were resolved through discussion and majority voting. For 462 tweets (92.4%), the annotators agreed with our labels, again demonstrating the high reliability of our results. Looking at the tweets with mismatched labeling, we found that most of them fell into *specific privacy* and *policy* categories and contained keywords or hashtags that are not widely used in general discussions.

4.5.4 Profile groups. The clustering of Twitter profiles was similarly validated. The same annotators independently assigned one of the seven groups (Table 3) to 700 randomly selected profiles (100 per group). Their grouping matched with ours for 639 (91.3%) profiles, demonstrating the correctness of our cluster identification and assignment methods.

4.6 Summary of findings

Our findings reveal the trend of EdTech-related discussions over 12 years. Comparing Figure 2 with Figure 3 reveals that both generic and privacy/security-related discussions followed similar temporal patterns—they peaked after the Covid-19 lockdown—but the latter was discussed at a much smaller scale than the former, indicating a low level of risk awareness among the public despite popular outlets are publishing news on data breach [35, 95, 131] and abuse [24, 42, 59, 103] at an increasing rate.

Manual categorization of tweets revealed that *specific* concerns were expressed more frequently than *generic* concerns, especially after the lockdown. Importantly, we uncovered concerns with specific technologies, such findings are generally difficult to obtain from surveys and interview-based studies due to resource constraints .

As Table 6 shows, surveillance, data collection, and privacy risks to students were the most frequently discussed topics in tweets

[6] We used Krippendorff's alpha [81] as a measure of inter-rater reliability, as it is more suitable for textual content analyses and more generalized than other measures (such as Fleiss' kappa and Cohen's kappa) in terms of the number of coders and categories.

expressing both generic and specific concerns; tweets in the latter category also mentioned risks experienced in a given context or from a specific technology. Frequent keywords were common across user groups (Table 7). Only the *Personal* group that likely included parents and other guardians touched on concepts such as students' expectations and privacy rights, and EdTech providers' accountability.

Interestingly, while *Zoom* appeared among the most frequently mentioned words, we did not find a single mention of *Zoom bombing* [127], which was a hot topic in 2020 on Twitter and other platforms [86]. A manual search of related keywords (e.g., *Zoom bombing*, *Zoom_bombing*, and *Zoom_bomb*) on Twitter revealed that tweets containing those terms, as well as Twitter accounts (such as zoom_bombing) that were, presumably, being used to share Zoom meeting IDs for bombing attacks [86], were removed. While excluding such tweets may affect findings related to discussions on Zoom security, we note that our focus was identifying a broader set of concerns, rather than a few topics that went viral.

Mapping the Twitter profiles into potential EdTech stakeholders, we found limited engagement from academics in discussing EdTech's privacy and security. Techies, who are likely to possess the most technical knowledge, touched on those topics in most of their tweets (Table 4), but rarely expressed concerns about the potential risks (Table 7). Interestingly, the most EdTech enthusiastic group, *Academic techies*, expressed the most concerns related to EdTech (Table 7).

5 METHODS OF REDDIT DATASET COLLECTION AND ANALYSIS

5.1 Subreddit selection and data collection

To identify subreddits related to our research, we conducted web searches and consulted Reddit posts listing education-related subreddits [28]. Additional subreddits were identified from Reddit's recommendations. In total, we identified eight subreddits related to education: */r/education*, */r/edtechhelp*, */r/higheredsysadmin*, */r/students*, */r/professors*, */r/k12sysadmin*, */r/teaching*, and */r/teachers*. Using the PMAW [97] wrapper around the Pushshift API [12], we collected all posts and comments that were submitted on those subreddits anytime before February 2022.

5.2 Identifying EdTech-related posts and their manual labeling

Similar to the Twitter dataset, we searched for EdTech-related terms (i.e., names of EdTech providers, tags, and phrases) in the title and body of the posts; this step yielded 3,792 posts containing at least one keyword. These posts were then filtered using the same set of privacy and security-specific keywords as for tweets (Section 3.3). Only 186 posts contained one or more of those keywords in their title or body. The sentiment analysis step was omitted as the number of posts for manual analyses was much smaller than the Twitter dataset. These 186 posts were manually coded using the labels created in section 3.4. Additionally, since discussions on Reddit were more detailed and nuanced than tweets, we performed qualitative analyses on this dataset (Section 6.3).

Label	# Posts in subreddits
Privacy specific	Teachers: 18, k12sysadmin: 10, Professors: 9
Exchanging tips	Teachers: 10, k12sysadmin: 9, Professors: 3
Privacy generic	Professors: 1, Teachers: 2, k12sysadmin: 1
EdTech support	Professors: 2, Teachers: 3
EdTech generic	Professors: 1, Teachers: 1
Other privacy	k12sysadmin: 3
Privacy incident	Professors: 1
Tech branding	Professors: 1, k12sysadmin: 1

Table 8: The number of posts in each concern category.

6 FINDINGS FROM THE REDDIT DATASET

In total, we collected 545,502 Reddit posts from the seven subreddits (*/r/education*: 334581, */r/teachers*: 165573, */r/professors*: 23666, */r/k12sysadmin*: 12659, */r/student*: 8906, */r/higheredsysadmin*: 97, */r/edtechhelp*: 20).

6.1 Statistics of posts related to EdTech's privacy and security

Initially, we identified 3,792 posts that contained EdTech-related terms. Among these posts, at least one keyword related to security/privacy was found in 186 posts on the following subreddits: */r/teachers*: 81, */r/professors*: 45, */r/k12sysadmin*: 43, */r/education*: 10, */r/student*: 2, */r/higheredsysadmin*: 0, and */r/edtechhelp*: 0.

6.2 Distribution of types of concerns expressed

We manually reviewed and assigned 83 posts under the concern categories. Table 8 breaks down this number into categories and subreddits, and the following section provides qualitative insights from these posts. *Notably, similar to Twitter data, we found more posts mentioning specific threats than general concerns.* The most important 20 keywords were: zoom, students, school, privacy, google, online, goguardian, teacher, student, meeting, class, security, policy, kids, camera, feel, proctoru, access, teachers, issues, video, people, classroom, app, parents. This list greatly overlaps with the top words in tweets for both generic words (e.g., security, privacy, and student) as well as EdTech providers (e.g., Google, Proctoru, and Zoom). Unlike Twitter data, a relatively large number of posts shared detailed tips on properly configuring and using EdTech to avoid security, privacy, and legal issues; the following section provides qualitative insights from these analyses.

6.3 Qualitative findings

While fewer posts discussed EdTech on Reddit compared to Twitter, they were more detailed and thus provided richer information about people's attitudes and concerns. We discuss the major patterns we observed below from different stakeholders (e.g., educators and system administrators). The stakeholder groups were identified by manually reviewing the profile descriptions and the posts (e.g., "...in the class I teach..."), and based on membership in relevant subreddits (e.g., */r/teachers*, */r/professors*, and */r/sysadmin*).

Educators generally cared about students' privacy, but not everyone. Educators discussed EdTech's privacy implications for

their students. They expressed anxiety about improperly configuring remote teaching applications and wanted access to "premium" versions of the software that were considered to offer enhanced security and compliance with relevant policies (e.g., COPPA). Several instructors pointed out the lack of preparation and administrative support for remote teaching. An instructor was concerned about tracking students across devices when they used their home computers to remotely attend classes. Additionally, concerns about surveillance and censorship surfaced when using district-owned devices and federally monitored networks.

In contrast, five posts advocated *increased* use of technology in education and complained about the backlash against EdTech from students and their parents. They reasoned that EdTech's benefits outweigh their risks, and students' unwillingness to use EdTech due to privacy concerns contradicts their use of social media.

Educators also expressed worries about self-privacy. Eight posts from educators expressed concerns about their own privacy being violated. A faculty member of a higher education institute remarked that many other educators felt the "chilling effect" as EdTech monitor and log their activities, which can be reviewed by authorities. Two educators felt discomfort being recorded during remote teaching from their homes and other private spaces. Two posts discussed a recent incident where an instructor was publicly mocked after a recorded lecture was leaked on TikTok [5], a social media platform.

Tension between educators and administrations. One teacher mentioned that, on two occasions, their school administration asked them to use remote teaching applications in a way that may violate terms of service. The said teacher did not feel comfortable with the administration's assurance of taking the liability should any problem arise as the administration refused to provide a written statement.

Another senior faculty member revealed that they were given *full access* to *all* data collected by their institute's learning management system and such "backdoor" access was given to all high-level administrators without faculty members' knowledge. They also condemned the lack of institutional policy or guidance about such access privileges, as it was unclear who should take responsibility if problems arise.

Surfacing novel privacy threats. One educator pointed out that some apps may automatically install (browser) extensions even when students use personal devices to log into school accounts. Such stealthy behaviors not only enable device-wide tracking of students' activities but also risk the privacy of other people who share those devices (e.g., family members). A student posted their observation that GoGuardian [50] was filtering their internet activities even at home while using school-issued Chromebooks. One educator complained about other teachers' online sharing of photos taken during virtual meetings with students; such photos could facilitate online harassment.

Reddit: a medium to discuss security vulnerabilities and share tips to avoid them. Several posts in */r/teachers*, */r/professors*, and */r/k12sysadmin* discussed apps' vulnerabilities and shared tips to avoid them in detail. Two posts explained how Zoom bombing works and what precautions may help prevent it. Three posts cautioned fellow educators about using Zoom and laid out long lists of preparatory steps before deploying and using Zoom. Four posts compared two or more alternative applications in terms of security and privacy protection they offer. While such discussions may help navigate times of crisis, risks of wrong or misleading information remain [101]. Moreover, spammers or advertisers may hijack those discussion threads; future research could investigate these issues.

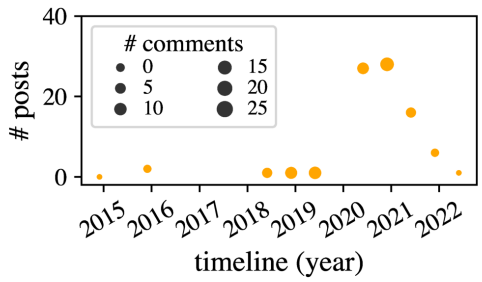

Figure 5: The trend of engagement with privacy and security-related posts on Reddit. The dot size correlates with the average number of comments or replies to those posts.

6.4 Trends and popularity of posts related to privacy concerns

Similar to Twitter, we wanted to understand the number and popularity of posts related to EdTech's privacy and security issues on Reddit. Figure 5 shows the average number of posts (aggregated across all categories) in every six months and the average number of comments on those posts (indicating other users' participation in those discussions). Similar to Twitter, the number of posts expressing concerns peaked during mid-2020. Unsurprisingly, those posts had a larger number of comments on average compared to pre-pandemic posts.

6.5 Summary of findings.

Overall, we found that students, instructors, and system administrators were concerned about EdTech's privacy and security implications, confirming prior research examining educators' attitudes toward EdTech [71]. Our analysis, however, also identified posts from educators arguing for EdTech's increased use; we identified similar content on Twitter as well (Section 4.3.1). Such findings demonstrate the suitability of studying unsolicited expressions to discover different opinions. Additionally, we uncovered instructors' concerns regarding self-privacy, new privacy threats from using school-issued devices or monitored networks, and EdTech's sneaky behaviors, as well as tensions between instructors and other school officials.

7 LIMITATIONS

Similar to any research based on online data, our findings may not fully reflect the public perception of EdTech, since the online samples, even though taken from two of the most popular platforms, may differ from the rest of the population in their experience with

EdTech and its reporting. The findings also depended on several automated analysis steps (e.g., keyword search, clustering, and sentiment analysis). While these steps were evaluated through manual reviews, we may still have missed some data since all relevant posts may not contain the keywords we searched with, and sentiment analysis is not 100% accurate. Therefore, our results set a lower bound on the privacy and security concerns expressed on Twitter and Reddit.

Another limitation of our method is that the stakeholder groups were identified based on profile information on Twitter, titles and descriptions of subreddits, and employment status reported on the posts. While we manually reviewed the identified groups, we cannot verify the post authors' real-world occupation. Thus, the results should be interpreted with caution.

A key EdTech stakeholder group, students, was underrepresented in our datasets. There are many possible reasons underlying their absence in the dataset. For example, both platforms have a minimum age requirement of 13 [100, 121] that excludes a large number of students, students who use these platforms may not reveal that they are students in their profile or may feel discomfort in expressing their views online where it could be monitored by school authorities [48], and so on. While our findings may not generalize to this population, they complement prior studies where the focus was mostly on students and other stakeholders were absent.

Finally, our data do not allow us to differentiate between EdTech's use in K-12 or higher educational institutes. These two settings may differ in the apps they use, the types and levels of their concerns, and the technical and financial resources they can leverage to tackle anticipated risks.

8 ETHICAL CONCERNS

Our data collection and analysis procedure was reviewed and approved by our institutional ethics board. We took additional measures to protect data subjects' privacy and safety since our datasets may contain information about members of vulnerable populations. These measures aimed at maximizing the potential benefits and minimizing harms, according to the principle of *beneficence* in the widely followed ethical standards detailed in the Belmont Report [15]. Before data collection, we reflected on the potential benefits of this research to society, and in particular to the student population, including minors and marginalized groups. Since our research aims to improve students' privacy and safety, we believe that the potential benefits justify undertaking this research.

To minimize harm, we followed suggested guidelines for the ethical use of online data for research [99, 128], avoided replicating tweets and Reddit posts in the paper, and carefully presented the findings to maintain platform users' anonymity. We followed the terms and conditions of the respective platforms in analyzing the data. The datasets are stored in a secured server and will be deleted post-publication. We will only retain post IDs for reproducibility and may share them on request with other research groups only after reaching an agreement over data protection standards [99].

9 DISCUSSIONS AND CONCLUSIONS

Overall, the results from the two platforms converged and corroborated previous research, establishing the robustness of our methods and the reliability of our findings. On both platforms, a tiny fraction of EdTech-related discussions touched privacy or security issues—this result agrees with prior studies looking into privacy concerns of mobile applications [56, 90, 93, 125]. The absence of student posts corroborates prior research that found students to be unaware of or unconcerned about EdTech's use or data collection. Such disregard, however, may also be attributed to enhanced trust towards educational institutes [9] (compared to e.g., private entities), and power asymmetry (e.g., between students and instructors or school administration) where students feel 'pressured' to adopt EdTech and resign their privacy [110] due to learned helplessness [61]. Similarly, a general acceptance of technologies in education due to their usability [82], lack of risk awareness [71], as well as organizational factors, e.g., fear of being monitored on online platforms by employers [48], may have dissuaded academics from critically engaging with EdTech's increasing use. Future research could investigate the roles of these factors in creating or exacerbating privacy/security risks in educational contexts.

While many findings supported prior research, we also highlight the novelty of our study. Unlike any prior studies (e.g., [9, 19, 68, 71, 72, 124]), we investigated the awareness and perception of two large online populations regarding EdTech; These platforms were used in understanding security threats [64, 78, 98] as well as other sociotechnical issues [57] as they offer a glimpse of public perceptions of those issues. Our longitudinal data spanning over 12 years provides insights into how perceptions and concerns about EdTech's privacy and security varied over time, in particular, before and after the Covid-19 pandemic.

Most studies investigating EdTech were conducted pre-pandemic and focused mostly on LMS (Learning Management Systems) [19, 68, 72, 124]. In contrast, we discovered privacy concerns regarding more than 130 EdTech apps that are diverse in nature and are used in various educational contexts. Future research can use our results to rank apps based on the number of concerns that were raised about them. We further identified novel threats and discussed their effects on marginalized student groups (see below), an important topic that has not been highlighted enough in the EdTech context.

Our focus on observational data avoided biasing opinions and allowed us to gather diverse perspectives from different stakeholders. Tsai et al. remarked that structural issues, trust, and power imbalance may exacerbate students' privacy risks [120]; we provide evidence of such issues (e.g., conflicts between educators and school administration). Additionally, the stakeholder group *Personal* that likely included parents and other guardians, underscored accountability for EdTech providers and their obligation to meet users' expectations of privacy, possibly hinting toward a shift in how EdTech is conceptualized, developed, and deployed. Analyzing unsolicited expressions also allowed us to discover attitudes toward EdTech and their privacy implications that are unlikely to be surfaced in interview/survey-based studies (e.g., due to social desirability bias). For example, multiple prior studies reported that educators value student privacy highly [70, 71], and select technologies to use in the classroom keeping students' privacy and safety in mind [71, 84]. In both of our datasets, however, we discovered a preference for EdTech because of its usefulness and a disregard toward students' concerns by a few educators. We hypothesize that such attitudes are more prevalent than what we found (since they

are unpopular and less likely to be posted online) and need to be considered by researchers studying and improving EdTech's privacy and security (e.g., to identify ways to properly communicate the risks and encourage to adopt mitigation strategies).

Discussions about EdTech's privacy peaked in 2021, raising an interesting question: did the pandemic create more risks or just surfaced the existing ones? We maintain that the pandemic exposed a broader population to the existing risks, and simultaneously, created new risks. For example, remote participation in classes and exams using tools that continuously record private spaces threatens the privacy of students as well as other coexisting people (e.g., family members and roommates). Additionally, using school-issued devices at home, or device sharing among family members resulted in new privacy threats for both students and their friends and family. Future studies could look into an in-depth understanding of privacy threats to people who are not EdTech users.

Technical research in this domain, such as detecting vulnerabilities or improper data collection and sharing practices of different apps, is just getting traction [21, 26]. Similar research in other domains generally selected apps to analyze based on popularity; our findings offer another selection criteria: based on the (perceived or experienced) risks. The overlap in risks and apps found in both datasets suggests the generalizability of the results and their potential to guide future research in this direction.

Some of the identified threats, such as multi-device tracking and home network monitoring, may disproportionately affect marginalized and at-risk populations. Students from historically disadvantaged communities rely more on school-issued or shared devices, and consequently, they and their families may face increased surveillance [55]. Inferring demographic information from activity logs [58] may out gender-nonconforming students, and social media scanning by EdTech apps can create legal issues if students discuss pregnancy or abortion on those platforms [47, 74, 76, 94]. We recommend future research in this area pay special attention to vulnerable populations.

An important avenue for future research is to explore ways to raise risk awareness among students—the most vulnerable stakeholders, and educators—who prescribe EdTech's use in their classes. In addition to trust and learned helplessness, the fine-grained nature of data (e.g., clicks on web pages) can contribute to underestimating the risks: collecting such data may seem harmless to those who may not realize that private information, including identity, demographic attributes, personality traits, and behavioral patterns, can be inferred by aggregating such data [22, 58, 79, 112, 132]). Future studies could examine whether communicating concrete risks—e.g., instead of stating what data will be collected, explaining what sensitive information that data will reveal—affects risk perception and awareness.

EdTech's safe use also depends on other stakeholder groups who are involved in EdTech's procurement, setting institutional policies regarding data collection and ownership, or responsible for their development or maintenance. A lack of or conflicting understanding of potential risks among them may negatively affect end users' privacy and safety. We suggest future research to investigate the organizational procedures for EdTech's procurement, whether and how privacy and security issues are prioritized while creating data-sharing guidelines and negotiating the ownership and use of data with EdTech vendors, and ensuring accountability of the responsible parties if data is misused.

Finally, we make two recommendations about data collection and analyses. First, we noticed that tweets tend to focus on specific events (such as data breaches and data abuse by EdTech vendors) and activism (e.g., the movement against *Common Core* [2]), whereas Reddit posts focus on nuanced discussions and personal experiences. Thus, future research focusing on a qualitative understanding of interpersonal or organizational factors may benefit more by analyzing data from discussion forums like Reddit than micro-blogging platforms like Twitter. Second, regarding manual labeling of data, we suggest employing either experts or crowd workers with domain knowledge for annotation; the average crowd workers may mislabel data due to unfamiliarity with domain-specific entities or contexts.

In conclusion, we believe that this study will motivate and guide future research exploring ways to enhance risk awareness among stakeholders and invent techniques to mitigate the identified risks.

ACKNOWLEDGMENTS

We thank the students and crowd annotators for their participation in this research. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] 2021. The Rise of Education Surveillance. *Bloomberg.com* (Nov. 2021). https://www.bloomberg.com/news/newsletters/2021-11-03/the-rise-of-education-surveillance
[2] 2022. Common Core. https://en.wikipedia.org/w/index.php?title=Common_Core&oldid=1122626774 Page Version ID: 1122626774.
[3] 2022. Common Sense Media. https://en.wikipedia.org/w/index.php?title=Common_Sense_Media&oldid=1113145620#cite_note-20 Page Version ID: 1113145620.
[4] 2022. Remote Learning Software Tracked Kids' Data to Sell to Brokers. https://gizmodo.com/remote-learning-data-brokers-privacy-1848975202
[5] 2022. TikTok - Make Your Day. https://www.tiktok.com/en
[6] 2022. Website Traffic - Check and Analyze Any Website. https://www.similarweb.com/
[7] Fernando Alves, Aurélien Bettini, Pedro M Ferreira, and Alysson Bessani. 2021. Processing tweets for cybersecurity threat awareness. *Information Systems* 95 (2021), 101586. https://doi.org/10.1016/j.is.2020.101586
[8] Amazon. 2022. Amazon comprehend API. https://docs.aws.amazon.com/comprehend/latest/dg/API_Reference.html
[9] Kimberly E Arnold and Niall Sclater. 2017. Student perceptions of their privacy in leaning analytics applications. In *Proceedings of the seventh international learning analytics & knowledge conference (LAK '17)*. Association for Computing Machinery, New York, NY, USA, 66–69. https://doi.org/10.1145/3027385.3027392
[10] David G Balash, Dongkun Kim, Darika Shaibekova, Rahel A Fainchtein, Micah Sherr, and Adam J Aviv. 2021. Examining the examiners: Students privacy and security perceptions of online proctoring services. In *Seventeenth symposium on usable privacy and security (SOUPS 2021)*. USENIX Association, 633–652. https://www.usenix.org/conference/soups2021/presentation/balash
[11] Lindsey Barrett. 2020. Governance of student data. (2020). Publisher: UNICEF.
[12] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *CoRR* abs/2001.08435 (2020). https://arxiv.org/abs/2001.08435 arXiv: 2001.08435 tex.bibsource: dblp computer science bibliography, https://dblp.org tex.biburl: https://dblp.org/rec/journals/corr/abs-2001-08435.bib tex.timestamp: Fri, 24 Jan 2020 15:00:57 +0100.
[13] Peter Bautista and Paul Salvador Inventado. 2021. Protecting student privacy with synthetic data from generative adversarial networks. In *International conference on artificial intelligence in education*. 66–70.
[14] Lois Beckett. 2019. Under digital surveillance: how American schools spy on millions of kids. https://www.theguardian.com/world/2019/oct/22/school-student-surveillance-bark-gaggle
[15] United States. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1978. *The Belmont report: ethical principles*

and guidelines for the protection of human subjects of research. Vol. 2. Department of Health, Education, and Welfare, National Commission for the ….

[16] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* (2016).

[17] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* (2016).

[18] Nigel Bosch, R Crues, Najmuddin Shaik, and Luc Paquette. 2020. " Hello,[REDACTED]": Protecting student privacy in analyses of online discussion forums. *Grantee Submission* (2020). Publisher: ERIC.

[19] Mohammad Botnevik, Stianand Khalil and Wasson Barbara. 2020. Student awareness and privacy perception of learning analytics in higher education. In *Addressing global challenges and quality education*. Springer International Publishing, Cham, 374–379.

[20] Monica Bulger. 2016. Personalized learning: The conversations we're not having. *Data and Society* 22, 1 (2016), 1–29.

[21] Ben Burgess, Avi Ginsberg, Edward W. Felten, and Shaanan Cohney. 2022. Watching the watchers: bias and vulnerability in remote proctoring software. In *31st USENIX security symposium (USENIX security 22)*. USENIX Association, Boston, MA, 571–588. https://www.usenix.org/conference/usenixsecurity22/presentation/burgess

[22] Guanliang Chen, Dan Davis, Jun Lin, Claudia Hauff, and Geert-Jan Houben. 2016. Beyond the MOOC platform: Gaining insights about learners from the social web. In *Proceedings of the 8th ACM conference on web science (WebSci '16)*. Association for Computing Machinery, New York, NY, USA, 15–24. https://doi.org/10.1145/2908131.2908145

[23] Monica Chin. 2020. An ed-tech specialist spoke out about remote testing software — and now he's being sued. https://www.theverge.com/2020/10/22/21526792/proctorio-online-test-proctoring-lawsuit-universities-students-coronavirus

[24] Monica Chin. 2020. Privacy group files complaint against five online test-proctoring services. https://www.theverge.com/2020/12/9/22166023/epic-proctorio-examity-privacy-online-testing-school-lawsuit-proctoring

[25] Mathieu Cliche. 2017. BB_twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs. *arXiv preprint arXiv:1704.06125* (2017).

[26] Shaanan Cohney, Ross Teixeira, Anne Kohlbrenner, Arvind Narayanan, Mihir Kshirsagar, Yan Shvartzshnaider, and Madelyn Sanfilippo. 2020. Virtual classrooms and real harms. *CoRR* abs/2012.05867 (2020). https://arxiv.org/abs/2012.05867

[27] Contributors. 2022. Tweepy: An easy-to-use Python library for accessing the Twitter API. https://www.tweepy.org/

[28] Reddit contributors. 2012. Teaching & education subreddits. https://www.reddit.com/r/redditlists/comments/hrccp/teaching_education_subreddits/

[29] Various contributors. 2021. Compact language detector v3 (CLD3). https://github.com/google/cld3

[30] Various contributors. 2021. Langid. https://github.com/saffsd/langid.py

[31] Wikipedia contributors. 2022. Contraction (grammar). https://en.wikipedia.org/wiki/Contraction_(grammar)

[32] Wikipedia contributors. 2022. List of emoticons. https://en.wikipedia.org/wiki/List_of_emoticons

[33] Wikipedia contributors. 2022. Stop word. https://en.wikipedia.org/wiki/Stop_word

[34] Wikipedia contributors. 2022. TF-IDF. https://en.wikipedia.org/wiki/Tf\T1\textendashidf

[35] Sam Cook. 2021. US schools leaked 28.6 million records in 1,851 data breaches since 2005. https://www.comparitech.com/blog/vpn-privacy/us-schools-data-breaches/

[36] James C. Cooper and John M. Yun. 2021. Antitrust & Privacy: It's Complicated. https://doi.org/10.2139/ssrn.3871873

[37] Ana-Maria Creţu, Federico Monti, Stefano Marrone, Xiaowen Dong, Michael Bronstein, and Yves-Alexandre de Montjoye. 2022. Interaction data are identifiable even across long periods of time. *Nature Communications* 13, 1 (2022), 313. https://doi.org/10.1038/s41467-021-27714-6

[38] LYDIA DISHMAN. 2015. Facebook revokes internship after student exposes messenger flaw. https://www.fastcompany.com/3049810/facebook-revokes-internship-after-student-exposes-messenger-flaw

[39] Hendrik Drachsler and Wolfgang Greller. 2016. Privacy and analytics: It's a DELICATE issue a checklist for trusted learning analytics. In *Proceedings of the sixth international conference on learning analytics & knowledge (LAK '16)*. Association for Computing Machinery, New York, NY, USA, 89–98. https://doi.org/10.1145/2883851.2883893

[40] Lisa Fleisher Elizabeth Dwoskin. 2014. Parental opposition fells inBloom education-software firm. https://www.wsj.com/articles/SB10001424052702304049904579516111954826916

[41] Facebook. 2021. Facebook for education. (2021). https://education.fb.com

[42] Todd Feathers. 2020. Privacy group asks for investigation into software that spies on students. https://www.vice.com/en/article/4adajq/privacy-group-asks-for-investigation-into-software-that-spies-on-students

[43] Todd Feathers. 2022. This private equity firm is amassing companies that collect data on america's children. https://themarkup.org/machine-learning/2022/01/11/this-private-equity-firm-is-amassing-companies-that-collect-data-on-americas-children

[44] Barbara Fedders. 2019. The constant and expanding classroom: Surveillance in K-12 public schools. *North Carolina Law Review* 97, 6 (2019).

[45] Tobias Fiebig, Seda Gürses, Carlos H Gañán, Erna Kotkamp, Fernando Kuipers, Martina Lindorfer, Menghua Prisse, and Taritha Sari. 2021. Heads in the clouds: Measuring the implications of universities migrating to public clouds. *arXiv preprint arXiv:2104.09462* (2021).

[46] Gennie Gebhart Karen Gullo Frida Alim Nate Cardozo and Amul Kalia. 2017. Spying on students: School-issued devices and student privacy. https://www.eff.org/wp/school-issued-devices-and-student-privacy

[47] GENNIE GEBHART. 2017. Spying on Students: School-Issued Devices and Student Privacy. https://www.eff.org/wp/school-issued-devices-and-student-privacy

[48] Luke Gentile. 2022. Texas school district to monitor employees' social media accounts. https://www.yahoo.com/video/texas-school-district-monitor-employees-194500153.html

[49] Nina Gerber, Benjamin Reinheimer, and Melanie Volkamer. 2019. Investigating people's privacy risk perception. *Proc. Priv. Enhancing Technol.* 2019, 3 (2019), 267–288.

[50] GoGuardian. 2012. GoGuardian. https://www.goguardian.com/

[51] Kalervo Gulson, Carlo Perrotta, Ben Williamson, and Kevin Witzenberger. 2021. Should we be worried about google classroom?: The pedagogy of platforms in education. *Journal of Professional Learning* 14 (2021), 46–50. https://search.informit.org/doi/10.3316/informit.484619281890842 Number of pages: 5 Place: Surry Hills, NSW Publisher: Australian Education Union New South Wales Teachers Federation Branch.

[52] Song Guo, Deze Zeng, and Shifu Dong. 2020. Pedagogical data analysis via federated learning toward education 4.0. *American Journal of Education and Information Technology* 4, 2 (2020), 56. Publisher: Science Publishing Group.

[53] Mehmet Emre Gursoy, Ali Inan, Mehmet Ercan Nergiz, and Yucel Saygin. 2017. Privacy-preserving learning analytics: Challenges and techniques. *IEEE Transactions on Learning Technologies* 10, 1 (2017), 68–81. https://doi.org/10.1109/TLT.2016.2607747

[54] Tzipora Halevi, Nasir Memon, James Lewis, Ponnurangam Kumaraguru, Sumit Arora, Nikita Dagar, Fadi Aloul, and Jay Chen. 2016. Cultural and psychological factors in cyber-security. In *Proceedings of the 18th international conference on information integration and web-based applications and services (iiWAS '16)*. Association for Computing Machinery, New York, NY, USA, 318–324. https://doi.org/10.1145/3011141.3011165

[55] DeVan L. Hankerson, Cody Venzke, Elizabeth Laird, Hugh Grant-Chapman, and Dhanaraj Thakur. 2021. *Student Privacy Implications of School-Issued Devices and Student Activity Monitoring Software*. Technical Report. https://cdt.org/insights/report-online-and-observed-student-privacy-implications-of-school-issued-devices-and-student-activity-monitoring-software/

[56] Hamza Harkous, Sai Teja Peddinti, Rishabh Khandelwal, Animesh Srivastava, and Nina Taft. 2022. Hark: A deep learning system for navigating privacy feedback at scale. In *2022 IEEE symposium on security and privacy (SP)*. https://www.computer.org/csdl/proceedings-article/sp/2022/131600b562/1CIO8duve24

[57] Rakibul Hasan, Cristobal Cheyre, Yong-Yeol Ahn, Roberto Hoyle, and Apu Kapadia. 2022. The impact of viral posts on visibility and behavior: A longitudinal study of scientists on twitter. In *Proceedings of the international AAAI conference on web and social media (ICWSM '22), to appear*.

[58] Rakibul Hasan and Mario Fritz. 2022. Understanding utility and privacy of demographic data in education technology by causal analysis and adversarial-censoring. *Proceedings on Privacy Enhancing Technologies* 2022, 2 (2022), 245–262. https://doi.org/doi:10.2478/popets-2022-0044

[59] Alex Hern. 2014. Google faces lawsuit over email scanning and student data. https://www.theguardian.com/technology/2014/mar/19/google-lawsuit-email-scanning-student-data-apps-education

[60] Benjamin Herold. 2017. The case(s) against personalized learning. https://www.edweek.org/technology/the-cases-against-personalized-learning/2017/11

[61] Donald S Hiroto and Martin E Seligman. 1975. Generality of learned helplessness in man. *Journal of personality and social psychology* 31, 2 (1975), 311. Publisher: American Psychological Association.

[62] Rae Hodge. 2020. Zoom security issues: Zoom buys security company, aims for end-to-end encryption. https://www.cnet.com/news/privacy/zoom-security-issues-zoom-buys-security-company-aims-for-end-to-end-encryption

[63] HolonIQ. 2021. 10 charts to explain the Global Education Technology Market.

[64] Sameera Horawalavithana, Abhishek Bhattacharjee, Renhao Liu, Nazim Choudhury, Lawrence O. Hall, and Adriana Iamnitchi. 2019. Mentions of security vulnerabilities on reddit, twitter and GitHub. In *IEEE/WIC/ACM international conference on web intelligence (WI '19)*. Association for Computing Machinery, New York, NY, USA, 200–207. https://doi.org/10.1145/3350546.3352519

[65] Sameera Horawalavithana, Abhishek Bhattacharjee, Renhao Liu, Nazim Choudhury, Lawrence O. Hall, and Adriana Iamnitchi. 2019. Mentions of Security Vulnerabilities on Reddit, Twitter and GitHub. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI '19)*. Association for Computing Machinery, New York, NY, USA, 200–207. https://doi.org/10.1145/3350546.3352519

[66] C Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 216–225. https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[67] Dirk Ifenthaler. 2017. Are higher education institutions prepared for learning analytics? *TechTrends : for leaders in education & training* 61, 4 (2017), 366–371. Publisher: Springer.

[68] Dirk Ifenthaler and Clara Schumacher. 2016. Student perceptions of privacy principles for learning analytics. *Educational Technology Research and Development* 64, 5 (Oct. 2016), 923–938. https://doi.org/10.1007/s11423-016-9477-y

[69] Maximilian Jenders, Gjergji Kasneci, and Felix Naumann. 2013. Analyzing and predicting viral tweets. In *Proceedings of the 22Nd international conference on world wide web (WWW '13 companion)*. ACM, New York, NY, USA, 657–664. https://doi.org/10.1145/2487788.2488017

[70] Kyle Jones, Amy VanScoy, Kawanna Bright, and Alison Harding. 2021. Do they even care? Measuring instructor value of student privacy in the context of learning analytics. (2021).

[71] Kyle M. L. Jones, Alison Harding Amy VanScoy, Kawanna Bright, and Sanika Vedak. 2022. A measurement of faculty views on the meaning and value of student privacy. *Journal of Computing in Higher Education* (2022).

[72] Kyle M L Jones, Andrew Asher, Abigail Goben, Michael R Perry, Dorothea Salo, Kristin A Briney, and M Brooke Robertshaw. 2020. "We're being tracked at all times": Student perspectives of their privacy in relation to learning analytics in higher education. *Journal of the Association for Information Science and Technology* 71, 9 (2020), 1044–1059. https://doi.org/10.1002/asi.24358

[73] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016).

[74] Jeremy Kahn. 2022. After Roe, fears mount about A.I.'s ability to identify those seeking abortions. https://fortune.com/2022/06/28/after-roe-v-wade-fear-of-a-i-surveillance-abortion/

[75] Katherine Mangan. 2021. The surveilled student. Publication title: The chronicle of higher education.

[76] MARK KEIERLEBER. 2021. The spy tech that followed kids home for remote learning—and won't leave. https://www.fastcompany.com/90677509/the-spy-tech-that-followed-kids-home-for-remote-learning-and-wont-leave

[77] Heather Kelly. 2019. School apps track students from classroom to bathroom, and parents are struggling to keep up. https://www.washingtonpost.com/technology/2019/10/29/school-apps-track-students-classroom-bathroom-parents-are-struggling-keep-up/

[78] Rupinder Paul Khandpur, Taoran Ji, Steve Jan, Gang Wang, Chang-Tien Lu, and Naren Ramakrishnan. 2017. Crowdsourcing cybersecurity: Cyber attack detection using social media. In *Proceedings of the 2017 ACM on conference on information and knowledge management (CIKM '17)*. Association for Computing Machinery, New York, NY, USA, 1049–1057. https://doi.org/10.1145/3132847.3132866

[79] Mark Klose, Vasvi Desai, Yang Song, and Edward Gehringer. 2020. EDM and privacy: Ethics and legalities of data collection, usage, and storage. *International Educational Data Mining Society* (2020). Publisher: ERIC.

[80] DAVID KRAVETS. 2013. Student suspended for refusing to wear RFID chip returns to school. https://www.wired.com/2013/08/student-rfid-chip-flap/

[81] Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

[82] Daniel G Krutka, Ryan M Smits, and Troy A Willhelm. 2021. Don't be evil: Should we use google in schools? *TechTrends : for leaders in education & training* (2021). https://doi.org/10.1007/s11528-021-00599-4

[83] Priya C Kumar, Marshini Chetty, Tamara L Clegg, and Jessica Vitak. 2019. Privacy and security considerations for digital technology use in elementary schools. In *Proceedings of the 2019 CHI conference on human factors in computing systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300537

[84] Priya C. Kumar, Marshini Chetty, Tamara L. Clegg, and Jessica Vitak. 2019. Privacy and Security Considerations For Digital Technology Use in Elementary Schools. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300537

[85] Charles Lang, Charlotte Woo, and Jeanne Sinclair. 2020. Quantifying data sensitivity: Precise demonstration of care when building student prediction models. In *Proceedings of the tenth international conference on learning analytics & knowledge (LAK '20)*. Association for Computing Machinery, New York, NY, USA, 655–664. https://doi.org/10.1145/3375462.3375506 Number of pages: 10 Place: Frankfurt, Germany.

[86] Chen Ling, Utkucan Balc, Jeremy Blackburn, and Gianluca Stringhini. 2021. A first look at zoombombing. In *2021 IEEE symposium on security and privacy (SP)*. 1452–1467.

[87] Asim Majeed, Said Baadel, and Anwar Ul Haq. 2017. Global triumph or exploitation of security and privacy concerns in e-learning systems. In *International conference on global security, safety, and sustainability*. 351–363.

[88] Charles Malafosse. 2019. FastText sentiment analysis for tweets: A straightforward guide. https://github.com/charlesmalafosse/FastText-sentiment-analysis-for-tweets

[89] Roxana Marachi and Lawrence Quill. 2020. The case of Canvas: Longitudinal datafication through learning management systems. *Teaching in Higher Education* 25, 4 (2020), 418–434. https://doi.org/10.1080/13562517.2020.1739641 Publisher: Routledge.

[90] Debjyoti Mukherjee, Alireza Ahmadi, Maryam Vahdat Pour, and Joel Reardon. 2020. An empirical study on user reviews targeting mobile apps' security & privacy. *arXiv preprint arXiv:2010.06371* (2020).

[91] Shadia Nasralla and Angelika Gruber. 2015. Austrian student's lawsuit vs Facebook bogged down in procedure. https://www.reuters.com/article/us-facebook-austria-lawsuit/austrian-students-lawsuit-vs-facebook-bogged-down-in-procedure-idUSKBN0N019420150409

[92] Natasha Singer. 2021. Learning apps have boomed in the pandemic. Now comes the real test. https://www.nytimes.com/2021/03/17/technology/learning-apps-students.html Publication title: The new york times.

[93] Duc Cuong Nguyen, Erik Derr, Michael Backes, and Sven Bugiel. 2019. Short text, large effect: Measuring the impact of user reviews on android app security & privacy. In *2019 IEEE symposium on security and privacy (SP)*. 555–569. https://doi.org/10.1109/SP.2019.00012

[94] Barbara Ortutay. 2022. Why some fear that big tech data could become a tool for abortion surveillance. https://www.pbs.org/newshour/economy/why-some-fear-that-big-tech-data-could-become-a-tool-for-abortion-surveillance

[95] Carly Page. 2021. Pearson to pay $1M fine for misleading investors about 2018 data breach. https://techcrunch.com/2021/08/16/pearson-to-pay-1m-fine-for-misleading-investors-about-2018-data-breach

[96] Arjunil Pathak, Navid Madani, and Kenneth Joseph. 2021. A Method to Analyze Multiple Social Identities in Twitter Bios. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–35. https://doi.org/10.1145/3479502

[97] Matthew Podolak. 2022. PMAW: Pushshift multithread API wrapper. https://github.com/mattpodolak/pmaw

[98] Shrestha Prasha, Arun Sathanur, Maharjan Suraj, Saldanha Emily, Arendt Dustin, and Volkova Svitlana. 2020. Multiple social platforms reveal actionable signals for software vulnerability awareness: A study of GitHub, Twitter and Reddit. *PLOS ONE* 15, 3 (April 2020), 1–28. https://doi.org/10.1371/journal.pone.0230250 Publisher: Public Library of Science.

[99] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society* 7, 2 (April 2021), 205630512110190. https://doi.org/10.1177/20563051211019004

[100] Reddit. 2021. Reddit user agreement. https://www.redditinc.com/policies/user-agreement#p_37

[101] Elissa M Redmiles, Noel Warford, Amritha Jayanti, Aravind Koneru, Sean Kross, Miraida Morales, Rock Stevens, and Michelle L Mazurek. 2020. A comprehensive quality evaluation of security and privacy advice on the web. In *29th USENIX security symposium (USENIX security 20)*. USENIX Association, 89–108. https://www.usenix.org/conference/usenixsecurity20/presentation/redmiles

[102] Joel R Reidenberg and Florian Schaub. 2018. Achieving big data privacy in education. *Theory and Research in Education* 16, 3 (2018), 263–279. https://doi.org/10.1177/1477878518805308

[103] Adi Robertson. 2015. EFF complaint says Google broke privacy pledge by tracking students. https://www.theverge.com/2015/12/1/9832210/eff-google-student-privacy-pledge-ftc-complaint

[104] Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 502–518. https://doi.org/10.18653/v1/S17-2088

[105] N Cameron Russell, Joel R Reidenberg, Elizabeth Martin, and Thomas B Norton. 2018. Transparency and the marketplace for student data. *Va. JL & Tech.* 22 (2018), 107. Publisher: HeinOnline.

[106] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on web search and data mining (WSDM '15)*. Association for Computing Machinery, New York, NY, USA, 399–408. https://doi.org/10.1145/2684822.2685324

[107] D Samuelson. 2017. Student data being sold to the highest bidder as privacy rights have gone down the drain. https://www.naturalnews.com/2017-05-12-student-data-being-sold-to-the-highest-bidder-as-privacy-rights-have-gone-down-the-drain.html

[108] Andrea Schlüschen. 2016. Celebrity endorsement in social media. In *Encyclopedia of E-commerce development, implementation, and management*. IGI Global, 1940–1956.
[109] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Vol. 39. Cambridge University Press Cambridge.
[110] John S. Seberger, Marissel Llavore, Nicholas Nye Wyant, Irina Shklovski, and Sameer Patil. 2021. Empowering resignation: There's an app for that. In *Proceedings of the 2021 CHI conference on human factors in computing systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445293 Number of pages: 18 Place: Yokohama, Japan tex.articleno: 552.
[111] Konstantinos Semertzidis, Evaggelia Pitoura, and Panayiotis Tsaparas. 2013. How people describe themselves on Twitter. In *Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks*. ACM, New York New York, 25–30. https://doi.org/10.1145/2484702.2484708
[112] Suranga Seneviratne, Aruna Seneviratne, Prasant Mohapatra, and Anirban Mahanti. 2015. Your installed apps reveal your gender and more! *SIGMOBILE Mob. Comput. Commun. Rev.* 18, 3 (Jan. 2015), 55–61. https://doi.org/10.1145/2721896.2721908 Number of pages: 7 Place: New York, NY, USA Publisher: Association for Computing Machinery tex.issue_date: July 2014.
[113] Suzanna Sia, Ayush Dalmia, and Sabrina J Mielke. 2020. Tired of topic models? Clusters of pretrained word embeddings make for fast and good topics too!. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*. Association for Computational Linguistics, Online, 1728–1736. https://doi.org/10.18653/v1/2020.emnlp-main.135
[114] Mohit Singhal, Nihal Kumarswamy, Shreyasi Kinhekar, and Shirin Nilizadeh. 2021. The prevalence of cybersecurity misinformation on social media: Case studies on phishing reports and zoom's threats. https://doi.org/10.48550/ARXIV.2110.12296
[115] Kaiwen Sun, Christopher Brooks, Abraham H Mhaidli, Florian Schaub, and Sonakshi Watel. 2018. Taking student data for granted? A multi-stakeholder privacy analysis of a learning analytics system. In *EDM 2018 workshop on policy and educational data mining*.
[116] Kaiwen Sun, Abraham H Mhaidli, Sonakshi Watel, Christopher A Brooks, and Florian Schaub. 2019. It's my data! Tensions among stakeholders of a learning analytics dashboard. In *Proceedings of the 2019 CHI conference on human factors in computing systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300824
[117] Jared Suttles. 2011. TweetTokenize. https://github.com/jaredks/tweetokenize/
[118] Nadia Tamez-Robledo. 2021. What do teachers know about student privacy? Not enough, researchers say. https://www.edsurge.com/news/2021-10-08-what-do-teachers-know-about-student-privacy-not-enough-researchers-say
[119] Victor V. Timchenko, Sergey Y. Trapitsin, and Zoya V. Apevalova. 2020. Educational Technology Market Analysis. In *2020 International Conference Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS)*. 612–617. https://doi.org/10.1109/ITQMIS51053.2020.9322982
[120] Yi-Shan Tsai, Alexander Whitelock-Wainwright, and Dragan Gašević. 2020. The privacy paradox and its implications for learning analytics. In *Proceedings of the tenth international conference on learning analytics & knowledge (LAK '20)*. Association for Computing Machinery, New York, NY, USA, 230–239. https://doi.org/10.1145/3375462.3375536
[121] Twitter. 2021. About parental consent on Twitter. https://help.twitter.com/en/using-twitter/parental-consent
[122] Twitter. 2021. Twitter terms of service. https://twitter.com/en/tos
[123] Twitter. 2022. Twitter API: Academic research access. https://developer.twitter.com/en/products/twitter-api/academic-research
[124] Phu Vu, Megan Adkins, and Shelby Henderson. 2019. Aware, but don't really care: Student perspectives on privacy and data collection in online courses. *Journal of Open, Flexible and Distance Learning* 23, 2 (2019), 42–51. https://search.informit.org/doi/10.3316/informit.980808045440057 Publisher: Flexible Learning Association of New Zealand.
[125] Ge Wang, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2021. Protection or punishment? Relating the design space of parental control apps and perceptions about them to support parenting for online safety. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021). https://doi.org/10.1145/3476084 Number of pages: 26 Place: New York, NY, USA Publisher: Association for Computing Machinery tex.articleno: 343 tex.issue_date: October 2021.
[126] Deborah West, Ann Luzeckyj, Bill Searle, Danny Toohey, Jessica Vanderlelie, and Kevin R Bell. 2020. Perspectives from the stakeholder: Students' views regarding learning analytics and data collection. *Australasian Journal of Educational Technology* 36, 6 (Dec. 2020), 72–88. https://doi.org/10.14742/ajet.5957
[127] Wikipedia contributors. 2022. Zoombombing. https://en.wikipedia.org/wiki/Zoombombing
[128] Matthew L Williams, Pete Burnap, and Luke Sloan. 2017. Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. *Sociology* 51, 6 (Dec. 2017), 1149–1168. https://doi.org/10.1177/0038038517708140
[129] Ben Williamson, Sian Bayne, and Suellen Shay. 2020. The datafication of teaching in Higher Education: critical issues and perspectives. *Teaching in Higher Education* 25, 4 (2020), 351–365. https://doi.org/10.1080/13562517.2020.1748811 Publisher: Routledge.
[130] Elad Yacobson, Orly Fuhrman, Sara Hershkovitz, and Giora Alexandron. 2019. De-identification is not enough to guarantee student privacy: De-anonymizing personal information from basic logs. In *Companion proceedings 10th international conference on learning analytics & knowledge (LAK20)*.
[131] NBC New York. 2022. Data of more than 800,000 NYC public school students compromised in data hack. https://www.nbcnewyork.com/news/local/data-of-more-than-800000-nyc-public-school-students-compromised-in-data-hack/3616939/
[132] Sha Zhao, Shijian Li, Julian Ramos, Zhiling Luo, Ziwen Jiang, Anind K. Dey, and Gang Pan. 2019. User profiling from their use of smartphone applications: A survey. *Pervasive and Mobile Computing* 59 (2019), 101052. https://doi.org/10.1016/j.pmcj.2019.101052

A APPENDICES

A.1 Online resources

In addition to the top 100 EdTech apps on iOS and Android marketplaces, the following sources were used to collect additional app names.

(1) https://www.commonsense.org/education/selections-for-learning
(2) http://www.iu19.org/top-10-education-hashtags-to-follow-on-twitter
(3) https://www.teachthought.com/twitter-hashtags-for-teacher
(4) https://www.teachthought.com/technology/top-education-hashtags
(5) https://www.techlearning.com/how-to/recommended-hashtags-to-discover-top-edtech-tools-and-apps
(6) https://www.filamentgames.com/blog/top-10-hashtags-for-amplifying-your-edtech-and-game-based-learning-tweets
(7) https://builtin.com/edtech/edtech-companies
(8) https://www.oceanfrogs.com/list-of-edtech-companies-in-the-usa
(9) https://spdload.com/blog/top-education-startups
(10) https://www.softwaresuggest.com/blog/best-online-exam-proctoring-software

A.2 Search terms

Hashtags

#crisisteaching, #edtech, #eddata, #k12online, #teachingtools, #distancelearning, #Udacity, #edtechchat, #edadmin, #OnlineEd, #mlearning, #googleedu, #edapp, #edtechinHE, #flipclass, #edreformtribe , #LMS, #remoteteaching, #schoology, #onlinelearning, #remotelearning, #e-learning, #elearning, #openeducation, #edreformtribe, #ipadchat, #studentdataprivacy, #BYOD, #blendchat, #globaled, #edreform, #schoolreform, #homeschooling, #EdTechX, #ERL, #iPaded, #GsuiteEdu, #digped, #GoogleEdu, #blendedlearning, #disted, #MovingEduForward, #Schoology, #digitalequity, #moodle, #Edtools, #mooc, #EdApps, #virtuallearning, #schoolreform, #dropouts, #flatclass

Twitter handles

@moodle, @Schoology, @udacity, @PearDeck, @Schoolytics, @Verificient, @MercerMettl, @LearningAtScale, @ExamSoft, @examity, @AIProctor, @Educ_Technology, @edXOnline, @mergeedu, @quizizz, @speedexam, @EdTech_K12, @BookCreatorApp, @MicrosoftEDU, @ProctortrackPro, @edpuzzle, @ProctorU, @MerlynMind, @Newsela, @udemy, @edmodo, @Flipgrid, @PSIServicesLLC, @coursera, @pearsonvue, @blackboard, @Seesaw, @proctorio, @conductexam, @CanvasLMS, @merittracsvc, @edtechdigest,

@MyPowerSchool, @Examus1, @ClassDojo, @DigitalPromise, @TeachFX, @GoogleForEdu, @in_examonline, @ProctorExam

A.3 User clusters

The following Table 9 lists all 40 user clusters identified based on profile descriptions, which clusters were grouped together according to their resemblance to potential EdTech stakeholders, top keywords representing the cluster, and the cluster consistency score.

Group	Cluster keywords	Score
Academic	research, researcher, researching, lunduniversity, researched, keeleuniversity, reseach, stfxuniversity, universitysa, simaduniversity	0.740
	teacher, learner, math, lifelong, multiage, lpselma, paraeducator, nkcsd, classkick, lbpsb	0.638
	university, director, professor, assistant, associate, universityofky, professors, drakeuniversity, educause, rowanuniversity	0.602
	school, principal, hillcrest, elmhurst, oakdale, annandale, haddonfield, prekindergarten, olmsted, beechwood	0.700
Academic techie	education, student, technology, learn, tech, passionate, teach, teaching, techology, msuepet	0.585
	educator, specialist, certified, trainer, integrationist, nearpodcertified, classcraft, certifiedmie, edtechie, edugator	0.649
Techie	software engineer developer programmer developper coldfusion softwares getpostman klarna couchbase	0.692
	data scientist datum analytics analytic visualization analysts datavisualization boozallen linkeddata	0.676
Personal brand	founder, entrepreneur, entreprise, dealmaker, wantrepreneur, angellist, entrepreneure, exited, capitalfactory, businessperson	0.652
	lead, professional, provide, development, experience, training, focus, expert, develop, include	0.574
	create, inspire, empower, mission, difference, strive, bringing, collectively, striving, invigorate	0.589
Influencer	follow, twitter, account, official, instagram, page, update, office, feed, facebook	0.612
	online, website, start, check, visit, click, standout, blogspot, prospectus, peruse	0.528
	service, company, solution, solutions, operation, multinational, enabling, managment, caliber, allocation	0.618
	business, social, digital, marketing, medium, media, strategy, market, strategist, programmatic,	0.622
Business	editor, journalist, fivethirtyeight, primetime, journaliste, wlrn, journalistin, geekwire, journos, thismorne,	0.661
	news, information, event, bring, late, source, daily, issue, cover, relate	0.554
Personal	husband, father, proud, christian, amazing, awesome, marry, married, wonderful, lucky	0.668
	wife, mother, friend, daughter, sister, mama, mommy, grandmother, momma, grandma	0.743
Ungrouped	calhoun, augustana, frankfort, muskegon, canandaigua, susquehanna, chattahoochee, valparaiso, wyandotte, flagstaff	0.699
	illegitimate, flyover, covfefe, andrewyang, ronpaul, inalienable, twill, gerrymandering, moreincommon, stronghold	0.556
	lover, enthusiast, nerd, geek, avid, traveler, addict, fanatic, foodie, adventurer	
	suppose, heartless, supposedly, tagline, annoyance, undecided, vomit, effervescent, unwilling, suess	0.708
	feminist, lgbtqs, sexnotgender, progressivism, pākehā, downtrodden, letlenilead, sisepuede, nohumanisillegal, saynotoracism	0.637
	sport, football, soccer, dodgeball, fantasyfootball, racquetball, basketballer, kickball, lpga, sportsball	0.679
	health, care, interprofessional, medicate, phlebotomist, delirium, uhsft, tracheostomy, multimorbidity, greatormondst	0.613
	yvonne, frazier, mcbride, wilcox, michaela, mcclure, macleod, marjorie, cuthbert, davey	0.694
	writer, author, book, write, creator, kidslisten, authored, hardcover, chicagoreader, truestory	0.592
	environmental, environ, groundwater, mangrove, tectonic, enviroment, disturbance, builtenvironment, vegetation, deforestation,	0.630
	view, tweet, opinion, personal, endorsement, tweets, opinions, post, express, retweet	0.704
	hourglass, hokage, concoction, clockwork, groundskeeper, lamborghini, cauldron, parakeet, groundhog, marzipan	0.566
	fiction, novelist, fairytale, historicalfiction, unpublished, novella, histfic, thrilling, labyrinth, shortstorie	0.701
	design, artist, designer, graphic, photographic, artistry, sketching, photographe, photographs, creating	0.585
	community, support, dedicate, commit, advocating, coordinating, communion, wraparound, commence, nfps	0.511
	love, life, world, live, people, family, time, share, enjoy, passion	0.599
	game, geekdom, stuffs, warframe, nintendoswitch, gunpla, chiptune, lulz, gumroad, ghibli	0.599
	music, musician, soundtrack, baritone, crossover, radiohead, singersongwriter, rocknroll, drumming, dubstep	0.585,
	uzbekistan, libyan, caracas, latvian, kyrgyzstan, daffodil, whoeurope, internacional, tashkent, yokohama	0.593
	londonderry, northwich, skipton, highbury, llandudno, pontypridd, northernireland, basildon, birkenhead, westmidlands	0.716
	food, coffee, wine, cooking, cheesecake, avocado, cheeseburger, guacamole, pretzel, chocolatier	0.619

Table 9: All 40 profile clusters, one per line. The horizontal lines indicate which clusters were combined based on the similarity of the most important words representing them. The middle and right columns show the top 10 words defining the clusters and their consistency scores, respectively.