College Students' Perceptions of the Data Economy

Anindya Roy

&

Aaron M. Kessler

Abstract: This work presents results of a survey focused on understanding college students' perceptions of the data economy in the US. Results from 113 students across two years of sampling indicate that most students have trepidations around data being collected, but less concern around educational data. Students' responses exhibited limited understanding of laws and technical processes that govern and facilitate such data collection. Surprisingly, despite the pandemic moving most students online between year one and year two of the sampling, little variation was observed from year to year. Potential implications of these, and other, results and next steps for the survey are highlighted.

Introduction

Regardless of an individual's intent, anyone using an internet enabled device (computer, smart phone, smart home technologies) is participating in the data economy (Cyphers & Gebhart, 2019). "A data economy is a global digital ecosystem in which data is gathered, organized, and exchanged by a network of vendors for the purpose of deriving value for the accumulated information...The gathered data is then passed to individuals or firms which typically take a fee" (Wikipedia, 2019). The data economy is often associated with "free" applications installed on our phone or advertisement firms purchasing our search information in service of better targeting clients, services, and goods. While it is true these groups play a major role, many other sectors are continuing to expand their use and activity in data collection beyond the ever-expanding data economy ecosystem. A hallmark of the data economy is data collection that involves some form of monetization. Although many educational institutions and third-party services contracted by those institutions collect and utilize student data through similar processes, they may not be classified as being part of the data economy since they are not engaging in specific monetization activities.

At the same time, some educational platforms (i.e. learning management systems, web-based tutoring systems, collaboration tools) and the institutions who use them are active participants in the data economy (Young, 2019). Recent shifts to hybrid, online, and remote instruction as a response to the COVID-19 pandemic have only increased many Higher Education institutions' engagement with, and reliance on, such systems. Yet, even before these shifts many platform companies and instructional developers (faculty and instructional design professionals broadly) had been exploring best ways to use student data for instructional and administrative purposes (Young, 2019). Higher education institutions have even used behavioral data (e.g. class attendance, dining hall usage, and writing samples), often compiled across platforms, to make decisions regarding social and instructional interventions (Gardner, 2019; Harwell, 2019). The effect of personal data being used for either monetization or "instructional" purposes seems to have resulted in a blurring of lines that often leave users uncomfortable. For many users it is unclear where one process ends and another begins. Given this blurring, it seems important that higher education institutions engage students in the conversations related to data collection and use, instead of making blanket top down decisions.

Outside education, the data economy is almost impossible to avoid (Hill, 2019) and despite the seemingly accepted use of personal data as part of the data economy associated with higher education, little is understood about students and users' (i.e. Faculty, TA's, Staff) perceptions of personal data being captured and leveraged in any number of ways. The survey at the center of this project was developed with the goal of interrogating students' perceptions of the data economy and some of the most common (in both media and in technical circles) ideas associated with the topic. This survey set out to answer four broad questions: 1) What do students report to know about how their data is used within the data economy, specifically data produced as a direct result of Higher Education activity? 2) What, if any, concerns and benefits do students perceive are the outputs of engaging with the data economy? 3) How familiar are they with the laws and regulations governing the collection and the use of personal data, both educational and non-educational? 4) What is their understanding of the technical processes and the infrastructure that allow the data economy to operate?

Method

Instrument

The survey content was iteratively developed to include the exploration of important terms and topics grounded in previous work (Gardner, 2019; Harwell, 2019). Inclusion of topics and terms in the survey was constrained by the need to balance collecting enough student responses to adequately understand their perceptions of the data economy while limiting the survey length to between 5 and 10 minutes. In the end, the final survey (Appendix A) contained 10 questions (2 Likert style, 4 multiple choice style, three open-response questions, and one Yes/No Grid). The survey asked the level of comfort (LoC) the respondents had regarding their personal data being collected by for-profit entities, as well as about educational data being accessed by their educational institution and other third-party vendors. Further, students' familiarity with terms related to data collection mechanism, their interest in the topic, and the advantages and the disadvantages associated with organizations using collected data were included.

In order to streamline the survey, a 5x5 grid was created to capture students' willingness to share 5 different types of data with campus/affiliate entities in 5 different roles. The question (Q9) was crafted in a way that allowed for minimizing the priming effect by introducing two versions, where about half the respondents received the "OK to share" prompt, and the rest received the "not OK to share" prompt in the question. These responses are combined (described later) to determine the percentage of net-OK responses.

Data Sources

The final survey was anonymously administered in early 2020 (n=48) and 2021 (n=87) to students participating in a student-organized web development course and competition that runs every year during the break between the Fall and the Spring terms of a highly selective research university in the US northeast. Most participants were in the first two years of their undergraduate program in Computer Science, and had a strong interest and/or experience in computer programming. In accordance with the guiding IRB protocol, the survey was available only when a potential respondent certified that they are over 18 years of age and taking the survey voluntarily.

The survey was presented as part of a lecture which included reading and discussions of articles broadly related to the data economy and higher education data usage. To prevent these discussions from biasing the respondents, the survey was administered at the beginning of the session, and the participants had no access to the materials discussed in the class except the definition of data economy from Wikipedia. Forty-eight students completed the survey out of approximately 75 people present during the 1st year of the session and 87 out of >100 in year two. The lecture in 2021 was presented virtually via zoom.

Along with the responses, we also had access to the time it took a respondent to complete the survey. We determined informally that if this survey was completed in under 60 seconds, it did not receive adequate attention from the respondent, and we did not include those responses in our results. With this new filter, we had 46 eligible responses from 2020, and 67 responses from 2021. In Fig.1 we show the histograms of the time-to-completion for all responses. The plot has an upper cutoff of 10 minutes, but such responses are included in our analysis.

Figure 1: Histogram of time taken to complete the survey.

Data Analysis

Quantitative Analysis. The LoC question (Q1) was a 5-point Likert scale response that was collapsed to 3 categories (comfortable, neutral, and uncomfortable). For most questions in the survey the total counts were recorded for each answer category and reported as percentages. The exceptions were the questions on advantages (Q5) and disadvantages (Q6) of the data economy and the LoC in sharing education data with stakeholders (Q9). To reduce any priming effect for Q9, we randomly assigned respondents into 2 groups -- who received the question prompts as "OK", or "not OK". We report the difference as "net OK", converted as the percentage of all eligible respondents.

Qualitative Analysis. An emergent codebook development process was used to develop categories of responses for the open-response questions 5 and 6 (Atheide *et al.*, 2008). In the first year of the project, each open-response question entry was classified by each of the two researchers. After the initial round of coding, the two researchers discussed each classification while deciding on code definitions. Based on these discussions final codes were assigned to each open-ended response. Each response was classified using as many codes as appropriate based on the definitions. This resulted in survey responses with multiple assigned codes. The codes for responses related to the benefits of the data economy were classified into 5 categories, while the final codes for responses related to downside resulted in 8 categories. Using the same codebook from year one, the year two responses to questions 5 and 6 were again collaboratively coded by the two researchers.

Results

Year to Year Comparison

Given the survey was administered across two years, we first selectively investigated the difference between the 2020 and 2021 samples. We analyzed the LoC data by pooling the data into 2 categories: uncomfortable (somewhat and very uncomfortable), and non-uncomfortable, which includes neutral, somewhat comfortable and very comfortable levels. We have enough responses to perform a chi square test of significance by pooling the responses from 2020 and 2021 as described above.

Data Type	Uncomf. '20	Comf. '20	Uncomf. '21	Comf. '21	<i>p-value</i>
Educational	10	36	16	49	0.901
Health	25	21	39	26	0.690
Financial	29	17	48	17	0.314
Location	26	20	41	24	0.618
Communication	31	15	55	10	0.056
Search history	30	16	47	18	0.556
Voice	26	20	48	17	0.089
Visual	28	18	33	32	0.390
Aggregate	26	20	44	21	0.317

Table 1: Total number of respondents who were uncomfortable (Uncomf.) and non-uncomfortable (Comf.) from 2020 ('20) and 2021 ('2021), and the corresponding *p-value* from chi-square test.

We see that "Voice data" and "Communication data" show a statistically significant rise in the level of discomfort at $p=0.1$ significance level. However, we are performing multiple tests

(9 questions), and a more stringent threshold of statistical significance is more appropriate (e.g., by applying Bonferroni correction $0.1/9 \approx 0.01$) for the analysis. With this modification, we see that there is no statistical significance in the year-to-year change in the level of discomfort.

We also found no statistically significant difference on the questions related to adequacy of laws, potential for discrimination, and the level of familiarity with the technical terms. This shows a surprising stability in respondents' opinions overall. As such, the data we report moving forward in this section is pooled across years.

Pooled results

The pooled results of this study begin to inform how higher education students perceive data collection processes associated with the data economy. When asked about their LoC with their personal data being collected by for-profit entities, the respondents expressed most comfort regarding their educational data by far, and least comfort about the communication metadata (Figure 2). While only about 23% of the respondents were somewhat or highly uncomfortable with their personal educational data being collected, about 55%--77% of the respondents expressed that they are uncomfortable in all other personal data categories.

Figure 2: Percentage of respondents who are uncomfortable (blue, bottom segment), neutral (red, mid segment), and comfortable (green, top segment) with sharing specific data types with for-profit entities.

When asked specifically about their on-campus educational data being accessed by campus entities and other third-party services, the respondents were more open to the instructors and the institute officials/registrar's office having access to their grades and attendance records, as opposed to third-party service providers. In Figure 3 we share a

heatmap of the percentage of students who responded with being “ok” with the data-collection process. Within each square we report the net percentage of “ok” in relation to respondents who reported either ok or not ok, after all neutral and no responses were withheld from the analysis. Students strongly dislike the idea of their wifi access records (and to a lesser extent, campus facilities access records) being shared with anyone, unlike other data types which show a range of acceptance based on the roles of those who can access their data. This points to a possible distinction in how students view their educational records (e.g., grades, attendance etc.) from their activities on campus, which can be determined via wifi/facilities access records.

Figure 3: Net percentage of students responding “OK” to sharing data types (x-axis) with different campus entities and third-party educational services (y-axis).

In Figure 4 the “Adv. only” (n=6) are the number of respondents who responded to the survey prompt on advantages of engaging in the data economy (Q5) but did not provide a response related to disadvantages (Q6). The “Disadv. only” (n=7) represents the number of respondents who answered the disadvantage prompt (Q6) but left the advantage question (Q5) blank. The intersection of the circles represents the number of respondents (n=72) who responded to both questions.

Figure 4: Venn diagram of students who mentioned the advantages and disadvantages in their comments. The intersection represents those who mentioned both.

As for reported advantages of the data economy (Q5), Figure 5 shows the percentage of coded responses for each of the 5 emergent categories observed in the data. The most common are associated with allowing for advancement in society through data-driven research and enabling personalized products. Note, the third most common response category supporting free services and products is usually the one most publicized in popular media, yet was noticeably low in our sample.

Figure 5: Advantage response categories coded for Q5.

The respondents expressed concerns (Q6) about the data economy that were categorized into a larger number of groups (Figure 6). The most common response in this category, also likely the simplest category, was loss of privacy. The next three most common responses were potential for abuse/stalking/doxing and being manipulated, and an increased chance of data theft and data security issues.

Figure 6: Disadvantage response categories coded for Q6.

Considering these responses in the context of other questions in our survey it is no surprise that almost 60% of the respondents feel that they are vulnerable to discrimination because of what their personal data might reveal about them. Only about 7% of respondents feel that the existing laws in the U.S. are adequate to protect individuals from the downsides of the data economy, while 8 times as many respondents find the existing laws inadequate. About 2 in 5 respondents claimed that they read something related to data-economy issues within the past month, while 1 in 5 said that they are not interested in this issue.

When it comes to the underlying technologies of data collection, it appears that the data-harvesting and economization processes are not very clear to the respondents. For example, when asked if they are familiar with the data collection-related technical terms such as “browser fingerprinting” and “tracking pixels”, only about 50% or fewer respondents said that they were familiar or somewhat familiar, revealing the opacity of the technology to the end users (Figure 7). This is one of the most surprising results, as our respondents represent a group of tech-savvy undergraduates. The highest level of “very familiar” was only 11% for the “third-party data brokers” category from 2020 responses suggesting a clear lack of understanding of the technologies that underpins the data economy.

Figure 7: Level of familiarity with technical terms associated with data-harvesting processes. These terms are: browser fingerprinting (browfing), transport-layer security (tls), third-party data brokers (tpdb), and tracking pixels (trapix).

Having analyzed responses at the individual-question level, we decided to explore how the responses cluster together between different questions. Specifically, we investigated two categories:

(1) The correlation in responses among different data categories for the LoC (Q1) question. In other words, a strong correlation between two data categories indicates that the respondents generally have a similar LoC for those categories. In Figure 8 we present a correlogram containing the Spearman correlation of the 5-level Likert scale data on 9 data categories. We see, for example, the responses on search-history data and the communication data are well correlated ($r > 0.5$), while the correlation between education data and communication data is much weaker (~ 0.2). These trends may indicate how individuals “bucket” the importance of privacy for different types of personal data, and this insight can be used to inform the future versions of the survey.

(2) We investigated if the students with different opinions on 4 specific questions have different levels of comfort as expressed in Q1. These 4 questions are: the adequacy of the privacy laws, potential for discrimination for the vulnerable population, data-harvesting technology related to third-party data brokers, and the respondents’ reading habits on the subject. To analyze this, we assigned an average LoC score for each respondent based on their Q1 answer, by converting their Likert scale responses to numerical scores, and summing up the numerical scores for all 9 data types for each individual. The Likert scale responses were first collapsed to 3 levels: comfortable, neutral, and uncomfortable categories, followed by numerical conversion to 1, 0 and -1, respectively.

Figure 8: Correlogram of individual respondents' level of comfort with different data types.

For each of these questions we divided the respondents into 2 pools, and performed a T-test for the average LoC score of the responses in these 2 pools. The two pools for the adequacy of data privacy laws divided the respondents into “Yes” and “I don’t have an opinion” group vs those who responded “No”. For the vulnerability question, the groups were divided by the answer “Yes”, vs. “No” and “May be” taken together. For the reading habits, the groups were divided by their topical reading within the last week or the last month, vs. those who read on the topic less recently, or are not interested. Finally, on the question on the familiarity of the term “third-party data brokers”, the groups were divided by those who are “Very familiar” and “Somewhat familiar” vs. those who are “Not familiar”.

Except in the first question on the adequacy of the laws, we did not find any statistically significant difference at $\alpha = 0.05$ level on the average LoC score. The average LoC score differed significantly among those who responded that the existing laws are inadequate to protect individuals, vs. those who believe the laws to be adequate or have no opinion on the matter. Those who responded that the laws are inadequate have a higher level of discomfort than the other group with a *p-value* of 0.013 (determined from one-sided t-test), and an associated moderate effect size of 0.45 (*Cohen’s d* measure).

Conclusion

We conducted year one of the survey before most higher-ed institutions began operating remotely because of COVID-19. We recognize that for many institutions such a radical transformation in the mode of instructional delivery was only possible due to the data-driven economy, where many extensively used services exist on a “freemium / pay-with-your-personal-data” model (e.g., Google drive, Zoom free calls etc.).

As the reliance on “free” or “integrated” instructional systems became even more utilized over the course of the pandemic we thought it important to implement year two of data collection remotely in order to see if students' perceptions or interest changed as their on-campus lives shifted to an online learning world. To our surprise, no difference is seen in the responses from year to year. With no significant difference between responses in years one and two, we pooled the data to learn that respondents feel more comfortable with education data being collected as compared to other forms of “sensitive” data. Further, students are most comfortable with instructors and the registrar's office having access to grades, attendance, and video watching habits, while being almost completely against data being shared with third parties or having wifi access information shared with anyone.

The participants, a technically savvy group participating in a programming and web development course, had limited knowledge of the technical mechanisms that underpin how the data economy functions. The fact that those with strong technical knowledge in their age and peer group do not understand the underlying processes of data economy emphasizes the need for education and awareness for the community at large. One might argue that it is a very concerning starting point for future explorations with a less technically savvy set of participants.

This work represents a first step in exploring the perception of higher education students about the way in which their data are collected, shared, and used. To be clear, we are not advocating for stopping all forms of data sharing. However, we do think there is a gap between how the user data are utilized, and the users' understanding and the comfort with the process. This phenomenon extends within our college campuses. Perhaps the first steps would involve an open conversation about the purpose and the processes of data collection. As a first study, this survey points out the need to explore these issues in greater detail, and to collaborate among different institutions to gain insights on the broader higher education population. If you are interested in using the instrument with your learners (either in programs, classes, or informal settings) please contact the authors.

References

- Altheide, D., Coyle, M., DeVriese, K., & Schneider, C. (2008). Emergent qualitative document analysis. *Handbook of emergent methods*, 127-151.
- Cyphers, B. & Gebhart, G. (2019). Behind the One-Way Mirror: A Deep Dive Into the Technology of Corporate Surveillance. A publication of the Electronic Frontier Foundation.

- Gardner, L. (2019). Students Under Surveillance?: Data Tracking Enters a Provocative New Phase. Retrieved from <https://www.chronicle.com/article/Students-Under-Surveillance-/247312>.
- Harwell, D. (2019). Colleges are turning students' phones into surveillance machines, tracking the locations of hundreds of thousands. Retrieved from <https://www.washingtonpost.com/technology/2019/12/24/colleges-are-turning-students-phones-into-surveillance-machines-tracking-locations-hundreds-thousands/>.
- Hill, K. (2019). I Cut the Big Five Tech Giants From My Life. It Was Hell. Retrieved from <https://gizmodo.com/i-cut-the-big-five-tech-giants-from-my-life-it-was-hel-1831304194>.
- Young, J. (2019). How Tech Companies are Selling Colleges on Mass Data Collection. Retrieved from <https://www.edsurge.com/news/2019-10-18-how-tech-companies-are-selling-colleges-on-mass-data-collection>.

Appendix A. Survey Instrument

Q1. Please indicate how comfortable you are about your personal data being collected by for-profit entities?

- Educational data
- Health data Financial data Location data
- Communication metadata (e.g., phonecall records)
- Internet search history
- Voice data collected by services (e.g., Alexa, Google Assistant)?
- Visual data as collected by CCTVs in public places
- Aggregate data connecting one or more of the above categories

Choose one for each case above: Very comfortable, Somewhat comfortable, Neutral, Somewhat uncomfortable, Very uncomfortable

Q2. Please list the kind of personal data not mentioned above that you are OK with being collected by for-profit entities:

Open-response textbox

Q3. Do you believe the existing data protection laws are adequate to protect individuals' rights in the data economy?

Choose one of: Yes, No, I don't have an opinion, I have more to say on this (with space to type text)

Q4. Do you believe your personal data could make you potentially vulnerable to discrimination?

Choose one of: Yes, No, May be, I have more to say on this (with space to type text)

Q5. As an individual, what do you believe to be the benefits of the data economy?

Open-response textbox

Q6. As an individual, what do you believe to be the downsides of the data economy?

Open-response textbox

Q7. Indicate how familiar you are with the following terms in connection to the technicalities of the data-harvesting process:

- Browser fingerprinting
- Transport Layer Security (TLS)
- Third-party data brokers
- Tracking pixels

Choose one for each case above: Not familiar, Somewhat familiar, Very familiar

Q8. When was the last time you read something related to data economy?

Choose one for each case above: Within the last week, In the last month, In last 6 months, I'm not interested in this topic

Q9A (Q9B). Please select which of the personal educational data you are OK (not OK) with each of the following entities collecting:

- Your instructors
- Registrar's office
- Sports coach
- Other institute admins
- Third-party service providers (e.g., Piazza)

Choose as many apply for each case above: Course grades, Attendance records, Educational video watching records, Campus wifi access records, Campus facilities (e.g., dining, gym) access records

Q10. When you sign up for a service which may collect your personal data, to you, the data-privacy implications are:

Choose one of: Clear, Somewhat clear, Not clear