Proceedings of the 2021 Winter Simulation Conference
S. Kim, K. Smith, Z. Zheng, S. Masoud, B. Fang, C. Szabo, and M. Loper, eds.

DO PEOPLE FAVOR PERSONAL DATA MARKETS IN A SURVEILLANCE SOCIETY?

Ranjan Pal
Charles Light
Yifan Dong
Mingyan Liu

Electrical and Computer Engineering
University of Michigan
Ann Arbor, MI, 48109, USA

Yixuan Wang

Computer Science Carnegie
Mellon University
Pittsburgh, PA, 15213, USA

Pradipta Ghosh

Facebook
Menlo Park, CA, 94025, USA

Harshith Nagubandi

Electrical and Computer Engineering
University of California San Diego
La Jolla, CA, 92093, USA

Leana Golubchik

Computer Science
University of Southern California|
Los Angeles, CA, 90089, USA

Bodhibrata Nag

Operations Management
Indian Institute of Management Calcutta
Joka, WB, 700104, India

Swades De

Electrical Engineering
Indian Institute of Technology Delhi
Hauz Khas, DL, 110016, India

ABSTRACT

We investigate and rationalize how individuals in a surveilled *developing* economic society value a human-centric data economy (HCDE). We first design and conduct a **non-online** pilot field experiment on approximately 22500 human subjects across India from 2014-2019, and collect data reflecting the impact of monetary incentives on these subjects to voluntarily trade their (personal) data in the digital surveillance age. Consequently, we study how various degrees of incentive influence subject preferences - both, when they are, or are not well-informed about the commercial malpractices their personal data might be subjected to. We analyze and rationalize **two main observations** in general for the Indian population: (i) despite being warned of the commercial malpractices associated with their personal data in the mobile and IoT age, *they prefer to trade data for incentives*, and (ii) the willingness of individuals to trade personal data is statistically heavy-tailed, and *hints at following a weak power-law*.

978-1-6654-3311-2/21/$31.00 ©2021 IEEE

1 INTRODUCTION

We're in a digital economy where data acting as the new oil is more (fundamentally) valuable than ever, and the modern key to smooth functionality of everything from the government to local companies. Today, an unprecedented amount of analyzable information on humans, things, and nature opens up vast opportunities for accelerated insights, innovation, and economic growth (Jones and Tonetti 2019). According to a UN Financial Trade Quarterly (FTQ) report of 2019 (Milletler 2019), the five largest data firms in the world: Apple, Amazon, Facebook, Google and Microsoft - are actors in the digital data economy with a combined market value of nearly USD 4 trillion that represents approximately 20% of market capitalization in the USA. However, interestingly enough, the people - whose raw data is driving the fourth industrial revolution - play a rather passive role in the modern digital economy as they are often left out of the value chain that transforms their raw data into huge monetary benefits (Wu 2017; Zuboff 2019). In addition, this digital economy brings an added disadvantage to common people in the form of privacy risks. Most visibly, the Cambridge Analytica scandal and the influence of the 2016 US elections demonstrated that individuals are increasingly at privacy risk to become manipulated through Big Data aggregating and analyzing firms. More generally, as valuable human experiences-bearing personal data account for a large part of Big Data in the mobile and IoT age, they subsequently gave rise to a new form of exploitative capitalism - "surveillance capitalism" (Zuboff 2019), in which raw and free information on society individuals is systematically and often shamelessly, and unfairly analyzed using powerful AI tools to sell predictions on their behaviors to targeted advertising agencies.

1.1 Practical Justifications Behind the Need of a Human-Centric Data Economy

Since the inception of the Warren and Brandeis Privacy Act in 1890, until a few recent efforts (Laudon 1996; Varian 2009; Acquisti et al. 2015; Pal and Crowcroft 2019; Laoutaris 2019; Posner and Weyl 2018; Lanier 2014; Pal et al. 2020; Schwartz 2003; Taylor et al. 2016; Stigler 1978; Samuelson 2000) have argued in favor of an alternative human-centric data economy (HCDE) in which people be paid/compensated whenever their data will be used for revenue-generating products and services. More specifically, the authors mention that (i) paying for data puts economic pressure on online services to apply *data minimization* principles, i.e., to collect and process only the minimum amount of data necessary for their operation, that are mentioned in the General Data Protection Regulation (GDPR) - thereby mitigating privacy risks; (ii) paying people for their data will contribute to the (universal) guaranteed minimum income principle in neo-classical economics and be an alternative to labor-based compensation in the future in which most work will be done by machines - a survey conducted recently (Posner and Weyl 2018) estimated that if fair remuneration algorithms are set in place, (a) a family of four could earn upto USD 20,000 per year from their data - this amount estimated as a sum of the direct monetary impact created by their data and the indirect socio-economic influence (externality) such data generates, and (b) investment and innovation in technology will be boosted; (iii) business models and machine learning algorithms have zero value without human data, and thus should be taxed for collecting the latter - a viewpoint shared by industry leaders such as Bill Gates, Mark Zuckerberg, and Elon Musk, (iv) online services market is not a zero-sum game - increasing the profits of people by paying for their data does not have to harm the profits of online services, and might actually result in higher quality and quantity of data being shared by individuals in many application scenarios.

Many would agree that an HCDE is unconventional and counter-intuitive for a privacy-conscious society. However, human behavior sometimes showcases non-intuitive privacy actions, even without the current presence of a regulated economy. The well known *privacy paradox* introduced by Susan Barnes (Barnes and Paradox 2006), and advocated for many scenarios by (Taylor et al. 2016; Trepte and Reinecke 2011; Taddei and Contena 2013; Gerber et al. 2018), have clearly shown a discrepancy between online privacy concerns and privacy behaviors, i.e., even when users/individuals have substantial concerns with regard to their online privacy, they engage in self-disclosing behaviors that do not adequately reflect their

concerns. We emphasize here that in the modern smart phone age, these behaviors are often supplemented by psychology-driven actions (e.g., binary opt-in policies set by apps, individuals preferring free apps over paid apps, a significant population of users being neutral to cookie downloads) by individuals and data extractors alike that increase the privacy risk for online individuals (Pal and Crowcroft 2019). In a recent review, the authors (Taylor et al. 2016) use three themes to connect insights from social and behavioral sciences towards showcasing why privacy risks are inevitable (hence defending a rationale for HCDEs to exist): (i) peoples' uncertainty about the consequences of privacy-related behaviors and their own preferences over those consequences, (ii) the context-dependence of people's concern, or lack thereof, about privacy, and (iii) the degree to which privacy concerns are malleable - manipulable by commercial and governmental oversight. The aforementioned inevitability has also been proved through economic theory models (with perfectly rational actors) that attribute privacy risks to statistical correlations between user's data (Pal et al. 2020; Pal et al. 2020; Pal et al. 2020) - these correlations being even more likely for the realistic boundedly rational actor settings.

1.2 Research Motivation and Contributions

Research Motivation - Though in principle it may be clear that an HCDE might ensure some socio-economic parity amidst the inevitable privacy risks in a data-driven society, apart from the multiple other benefits it promises, one big question still remains to be answered: *are humans in favor of such an economy in the first place and willing to trade their personal data (a proxy for privacy) for incentives?*. An 'yes' answer to this question from a significant fraction of population with respect to various demographics is a *bare necessity* (and by no means sufficient) for society to witness an HCDE in the (near) future. After all, it is not new to the world that a certain technology or an idea can promise important benefits, but its adoption never takes off in society (e.g. the case of the secure DNSSEC protocol) for a myriad of reasons. *In this paper, we aim to get an answer to this question for human subjects in the Indian subcontinent*, as a necessary 'test-the-water' exercise towards a futuristic HCDE society.

Research Contributions - In view of (a) pre-conceived notions based on existing literature, and (b) personal social experience in the Indian society, we are inclined towards experimentally verifying the following hypothesis for the Indian population, **H**: *individuals in society prefer trading their personal data for incentives, despite being made aware of how such data could be unfairly misused in online data markets.* To this end, we *designed and conducted randomized trial experiments, i.e., RCTs (Kendall 2003), on 22.5K individuals in India from 2014-2019* (see Section 2). Specifically, our research contributions are pivoted on the following three research questions (RQs), answers to which highlight different facets of the hypothesis for the Indian population.

- **RQ1**: Do individuals in modern smart societies *without any substantial knowledge* on the mal-practices regarding the commercial use of their digital personal data, favor a human-centric data economy (HCDE), and be voluntarily willing to trade their personal data (hence in a sense a proxy for privacy) for incentives?
- **RQ2**: In the event that human subjects in such economies, malpractice aware or unaware, *are made (re)aware*, i.e., warned, of the commercial malpractices associated with their personal data in the mobile and IoT age, and are given suggestions to better their privacy hygiene, are they still willing to trade their (personal) data for or without incentives? Comparing the answers from RQ1 and RQ2 would showcase the degree of impact monetary incentives have on human privacy preferences to trade data in such economies, and strengthen/weaken the hypothesis.
- **RQ3**: Is there an underlying formal statistical pattern in such economies regarding the human willingness to voluntarily trade their personal data, *with and without* incentives (both, in the presence, and in the absence of them being made aware of the commercial malpractices associated with their data)? The answer to RQ3 would help us decipher if and how the socio-economic effects of population diversity affect individuals' decisions in (de)embracing HCDEs.

1.3 Overview of Results and Their Significance

Results - Through our main experimental results (see Section 3), we show both, the expected, as well as the non-obvious. Specifically, we observe that in general, *higher incentives favorably shift individuals' preference towards trading their data (the expected) irrespective of their prior awareness of commercial malpractices their data is subject to (a non-obvious observation) - thereby possibly adding weight to the privacy paradox, and (b) personal data trading preferences for an Indian population are statistically heavy tailed and hints at following (but not always) a weak power-law (also a non-obvious observation) independent of incentives.* We detail the behavioral and economic rationale behind our results in Section 3.
Broader Significance - Economic inequality in the GDP-rich India (that boasts of a top five GDP nation globally) is quite blatant and according to some sources (e.g., *New World Health*) claims the second position as the most unequal country globally, with millionaires controlling 54 per cent of its wealth but average individual income being only USD 5 per month. With data science technologies seeping into, and enhancing every aspect of our lives, the country's economy is no exception, and this science can be used in a host of initiatives, with the aim of a more 'equal' distribution of wealth along with increasing GDP. Adding to this is the fact that 'predatory' and economically unfair data collection has never been so obvious as it has been since the early 2000s, as documented by (Zuboff 2019). Hence, one could only expect 'parity'-enforcing human-centric data economies to be considerably more successful in the modern digital age. However, apart from *obvious concerns* on the technology (Odlyzko 2003; Carrascal, Riederer, Erramilli, Cherubini, and de Oliveira 2013) and the broader legal dimensions (including privacy) that *necessarily* needs to be addressed to practically implement HCDEs successfully - there is the all important factor: *societal willingness* - without which HCDEs cannot take off in many societies (e.g., with democracy governments such as in India), even if technology and legal concerns are alleviated. *The broader significance of our research questions lie in inferring through statistics and AI whether societal interest towards embracing HCDEs is a vox populi in India-like economies - the positive affirmation of which will pave the way for a push in alleviating challenges in (privacy-enhancing) technology, law, and regulatory policies (see Section 4 for more details) to realize HCDEs.*

1.4 Novelty and Relation to Prior Work

Different forms of RQ1 without any specific distinction between malpractice-aware and non-aware subjects has been addressed before in (Benndorf and Normann 2018; Ackerman et al. 1999; Kokolakis 2017; Grossklags and Acquisti 2007; Gerber et al. 2018). *All these efforts have come to the conclusion that monetary incentives do positively influence humans, albeit via different degrees, to trade personal data.* However, the **novelty** in our proposed research, compared to these works, with respect to RQ1 is in (a) the larger scale and much increased (geographic and literacy) diversity of human subjects we wish to experiment with (see Section 2 for more details), compared to those mentioned in these stated works, (b) the experimental study done on individuals in an 'orthogonal' (Indian) socio-economic structure, compared to those in the literature, and (c) getting an app-specific view of answers to RQ1 - simply because, an individual today has access to different types of smart apps, and for each of them, her preference to embrace HCDEs might be different due to the difference in (personal) information collected by the apps (see Section 2 for details). More specifically, even though an individual might not be aware of privacy malpractices, she might have varied degrees of sensitivity to different information types. RQ2 and RQ3 clearly remain unanswered till date. The role of RQ2 in addressing the *overarching question* as to whether incentives catalyze humans to trade their privacy is pivotal. The power of incentives can be best judged against a worst case (not considered by existing research efforts) - which in our work is ensured through an RCT intervention program that precisely discusses the privacy pitfalls of data trading with individuals not aware of them, before recording their willingness to trade personal data. Earlier efforts have synonymized addressing RQ1 with addressing the overarching question. It is not necessarily so, particularly in societies where cyber-security literacy among Internet users is poor (as in India). For such subjects, a positive answer

to RQ1 does not necessarily imply a positive answer to RQ2. Through addressing RQ3, to the best of our knowledge, we propose to be the first to unearth the mathematical underpinnings behind the structure of underlying social patterns that influence personal data monetization mindsets (in a certain geography).

2 OUTLINE OF EXPERIMENT DESIGN

We *concisely* detail *(in the past tense)* the design of RCT experiments conducted in Indian cities to assess preference of individuals to personal data trading in the presence of monetary incentives.

Preview - We conducted **non-online** RCTs (see (Ranjan Pal 2021) for a rationalized blueprint of conducting RCTs) in 10 Indian cities on approximately 22.5K (22348) [52% men and 48% women in the age range 18-68] human subjects spread across five zones (east, west, central, north, south) with the following population count distribution (4237, 3986, 3463, 6327, 4335) from the summer of 2014 till the summer of 2019, and each subject having access to a smart phone. According to (Turner, A 2020; Kemp 2019; Center 2019), approximately 50% of the population in the world (around 500 million of them in India own a smart phone where the average user (aged between 16-64) spends approximately least 3 hours≠ 5.5 hours for an average Indian) on average, out of which 90% of the time is spent browsing mobile apps. On the other hand, the average world population of individuals spends around 7 hours of time per day on average in any online activity, approximately twice that of the time spent on mobiles alone (Kemp 2019). *Thus, it is fair to assume that analyzing the preferences of mobile app users will provide a strong characterization of their privacy preferences in general.*. The primary reasons for us to go non-online were (a) to mitigate the digital footprint created due to online survey responses being stored on third party servers, and being amenable to privacy risks, and (b) grab more attention of survey participants in a multi-cultural, multilingual country like India through a physically-conducted RCT intervention (the privacy awareness program/tutorial), compared to an online one where the participants would have less scope to interact with the intervention providing team. The number of subjects surveyed with consent (see (Ranjan Pal 2021) for the consent form) was approximately 60% of the approved capacity of 35K surveyors, constrained on voluntary participation. The PIs approached multiple regional institutional review boards (IRBs) in the different country zones prior to conducting RCT experiments, to get their approval. Each of these IRBs ceded authority to the IRB at the Indian Institute of Management Calcutta (IIMC).

Why RCTs for LMICs? - In our setting, we are interested in observing the outcome of whether individuals would prefer to embrace an HCDE (or otherwise) when provided with monetary incentives. However, the (positive) power of incentives can only be judged best when individuals prefer to embrace HCDEs, despite being made explicitly aware of the privacy risks that might accompany an HCDE. This awareness exercise is the intervention program of the RCT. While, people in more cyber-literate societies might be well aware of the potential privacy pitfalls of HCDEs, the ones in lesser cyber-literate societies (e.g., low and medium income demographics) need to be made explicitly aware before we can infer whether monetary incentives indeed tilt the preference of individuals to voluntarily embrace HCDEs, despite the knowledge of privacy risks accompanying them.

Secure and Private Data Management - The RCT survey form capturing integer-scaled (e.g., from 0 to 5) individual preferences to embrace HCDEs was anonymous and only required meta information like gender, profession, geographical zone, and whether one uses social network accounts to access apps. We *did not* require surveyors to fill in details such as their phone, ZIP code, email - attributes that are often privacy sensitive and can contribute to privacy breaches. Data relevant to our project will be made publicly available for sharing upon research publication. The project team will make use of a secure digital repository at the Indian Institute of Management Calcutta for permanent and HIPPA-compliant safe storage and archiving of project data. This data will be stored on a network file storage system with suitable permissions, and regularly backed up.

Subject Recruitment Specifics - The purpose (studying human preferences on personal data monetization) of the experiment, and the basic criteria for subject recruitment was advertised (without any cold-calling on our part), periodically every six months, through the following multiple channels: (i) emails from participating

Indian research institutions, i.e., IIMs and IIT, sent out to their members (staff, students, and faculty) within, and to partner educational institutions, (ii) the strongly industry-networked placement cell within these institutions being used to send ads to companies so as to target a corporate audience, (iii) participation flyers being circulated by staff, students, and faculty of such institutions in their living neighborhoods to target subjects from households, (iv) participation flyers being circulated by these household members to representatives of surrounding small local businesses (e.g., maids, shopkeepers, transport owners), (v) personal connections of research institute members in the higher administration to few celebrities (e.g., politicians, artists, business men/women, movie actors). This was made possible due to the multiple institutional social events where such celebrities were invited as chief guest in the past.

Every six months, we collected a list of participants who signed up for being part of our RCT experiments. Apart from the celebrities, all other subjects were selected via *consecutive sampling* (selecting everyone for the RCT experiments who fitted the basic criteria for the normal population), as per recommendations made by (Kendall 2003) for non-rarified populations. University research subjects were surveyed inside the universities, corporates were surveyed within their respective organizations, household members were individually surveyed in their homes or together with many neighboring household representatives in one home. This is a common culture in India to gather at one place especially if you are in a community setting., and small local business (SLB) representatives were surveyed in their business locations. The celebrities were surveyed in their offices. In order to ensure a suitable number of subjects necessary for effective RCTs, we used the standard Fragility Index (FI) measure (of statistical power) (Walsh et al. 2014) to estimate the threshold number of post-screened subjects. Each surveyed subject was given a *Paytm* gift cards worth INR 50 (equivalent to USD 0.75 roughly - an amount that can buy a lunch-time meal in many places in India, as remuneration. This was important to increase response rates and reduce drop-out rates (Maniaci and Rogge 2014).

The RCT Intervention Task - We conducted a non-online privacy awareness training program of about 45 minutes for intervention groups and post that ask the surveyors in both groups - the intervention group and the non-intervention, i.e., regular group, to complete the survey form (see (Ranjan Pal 2021)). The most salient discussion features of the training program for the intervention group included (a) what types of personal information are gathered by various apps for the purpose of targeted advertising, (b) the ones that cause high risk of privacy breaches, and more importantly (c) the degree of privacy hygiene one should adopt in the smartphone age. In this regard, we discussed the privacy paradox with real-life examples, where people on one side care about privacy, but on the other side voluntarily give up their data in the need to access certain service provided by apps. We then provided the audience with a big picture of how and why search engines, social networking sites, and e-commerce firms are making money (and how much) out of our personal data, without many of us being consciously aware of this fact. We further provided examples of popular data breach scandals such as the Target scandal and the Cambridge-Analytica scandal, and their negative impact on society. Finally, we discussed the popular channels in our day-to-day mobile lives via which privacy breach risks are increased, that included us talking about (with examples) socially-engineered emails, messages, ad-clicks, and spam phone calls. Following this, we raised awareness about proper ways of mobile and desktop web browsing to minimize leakage of personal information from apps to downstream data buyers.

Intervention-Centric Subject Sampling and Bias Mitigation - We adopted a *stratified sampling* (as per recommendations in (Kendall 2003)) in order to prevent the contraindication of the intervention program to be tested - thereby negating confounding variable effects. We sorted post-screened subjects based on their Westin scales (Westin 2003) (acting as RCT pivots), with placing x% of privacy risk-averse subjects into one pool prior to running RCTs, and then randomly allocating (via the *blinding* technique to mitigate bias (Kendall 2003)) half of this pool to the intervention program led *survey exercise*, and the other half directly to the survey exercise. The remaining $(100-x)$% of the volunteering subjects were initially perceived by us to be vulnerable to the privacy paradox. Thus, we separated this set of subjects into the *privacy pragmatist* and *privacy unconcerned* categories, and separately ran intervention programs on half of the

randomly sampled subjects in each of these categories. As a result of applying the blinding step, we made sure that neither the subjects, nor anyone performing subsequent measurements and data collection were aware of the random intervention group assignment, ensuring a setup that does not favor unfair attention to any group. In order to mitigate the chance of falling into the popular *Simpson's paradox* (Appleton et al. 1996; Perera 2006; Reintjes et al. 2000; Wagner 1982) in RCTs, we resorted to a proper qualitative social understanding of common parameters that most likely influenced individual trading preferences (in our case the Westin scale) and a proportional randomized allocation of such parameters among groups. However, it is widely accepted that the extent to which Simpson's paradox is likely to occur in experimental research, including RCTs, is difficult to determine (Ameringer et al. 2009), simply because there is no guarantee that every potential confounding variable will be known.

3 STATISTICAL RESULTS AND A BEHAVIORAL-ECONOMIC ANALYSIS

We observe and analyze patterns on individual subject willingness to trade their privacy with commercially motivated entities, for various incentive structures. More specifically we study **(i)** the proportion of individuals for a given app category that belong to various 'willingness to trade privacy' (WTP) buckets, i.e., bucket 0 to bucket 10, for different incentive structures and training experiences - validity of the study substantiated by the KS and Fragility Index tests, and **(ii)** the complementary CDF (CCDF) plot that helps to characterize mathematical trends (if any) behind the proportions obtained in (i). In the interest of space, we represent and analyze the experimental outcomes for three app categories, one each from the lower *(house and home)*, middle *(medical)*, and upper parts *(social communication)* of the app popularity curve (see Fig.3 in (Ranjan Pal 2021)). The results for the chosen app categories represent, w.l.o.g., all the other app categories in the corresponding curve parts (see (Ranjan Pal 2021) for plots in other app categories).
Observation #1 - With respect to **(i)**, for all the app categories studied, we observe in general (Figures 1-3 (a)-(d)) that (a) WTP values are high for around at least 45% of the subject population (buckets 10, 9, and 8) *irrespective* of both, the amount of monetary incentive, and whether the subjects undergo a training program or not, and (b) the willingness to trade personal data quite evidently decreases with training.
Statistical Inference for Observation #1 - Despite the observation made above through (b), using the Kolmogorov and Smirnov (KS) hypothesis test, *we reject the null hypothesis* (i.e., statistically significant p-values in RCTs less than 0.05) that *given incentives, the privacy training program has a major influence on an individual's willingness to trade personal data.* More specifically, a privacy training program on average reduces an individual's (even if he is privacy literate) WTP values only approximately by less 10% - when compared to the setting where such an individual does not participate in the awareness program (also evident via 'extreme' incentive scenario plots 1-3 (i)-(j)). The Fragility Index (FI) for all the app categories stands lowest at 15 in the RCTs, indicating the KS test results to be fairly robust, i.e., less sensitive to consumer preference changes, since high minimum values of FI complement low *p*-values.
Behavioral Economic Rationale behind Observation #1 - Observation #1 is well rationalized for a country like India - a developing low-medium economy country (LMIC), despite privacy being a right currently upheld by the Indian constitution - an exception being situations (e.g., fake news spread for social and communal harm) where upholding privacy for every individual might go against bringing anti-social elements to justice (Section 79 of IT Act). This is because of multiple reasons working together in tandem: (a) the basic awareness regarding good privacy-hygiene is lacking for a significant Indian population of digitally equipped but illiterate users - the inferences here are based on the population in major Indian cities, where the digital illiteracy factor is far greater than in sub-urban areas of the subcontinent, though much of the population in such areas are equipped with smart phones, (b) a significant section of the GDP-rich, highly inequitable LMIC population preferring to monetize their personal data in return for incentives that might increase their daily average income (even an amount as small as 0.75 USD is worth a meal in many parts of India), (c) a part of the same population being under the perception (due to data commercialization inevitability and its inherent economic unfairness) of accruing high *opportunity costs* (Angner 2012) of not being part of an HCDE, (d) a consensus of resentment in certain sections of the privacy-sensitive public

on the unfairness of existing data commercialization providing a basis for a *behavioral anchoring bias* (Angner 2012) that makes these individuals prefer embracing HCDEs when compared to staying true to their 'private' nature and shying away from them, (e) a sense of *confirmation bias* (Angner 2012) prevailing among parts of the privacy-literate LMIC population that an evidence of privacy enhancing technologies (PETs) being increasingly used by personal data collectors does not rule out the inevitable existence of unfair information asymmetry driven personal data commercialization, and (f) the strong close-knit socio-cultural fabric of India that enables voluntary personal data release by individuals on mobile social community platforms to garner social importance points (e.g., through Facebook likes). *Though our data analysis is solely on an Indian population, rationale (a)-(c) are likely to hold for populations in quite a few countries around the globe with similar socio-economic and political structures.*

Observation #2 - We observe another interesting counter-intuitive phenomenon throughout Figures 1-3: for the cases when incentives are provided, the number of high WTP value subjects with no training is *lesser* than that with training.

Behavioral Economic Rationale behind Observation #2 - Observation #2 can be explained in view of the *anchoring effect* and the *availability heuristic* (Angner 2012). High WTP zones usually represent individuals who are relatively more profit-minded than privacy-sensitive. The training program confirms their viewpoint on the gross economic unfairness of the current data economy. In the first place, they are willing to incur the least opportunity costs in an LMIC economy. Second, the training program induces an anchoring effect on parts of this population of high WTP individuals wherein their desires to make money out of their personal data increases post any 'resentment' feelings regarding the current state of gross economic unfairness in the data economy that we portrayed in the training program. Finally, it is a plausible behavioral reason that some individuals without training might *perceive* being more risk-averse to exhibit the high WTP values compared to the case when they are properly educated about how their data is commercialized - exhibiting a form of *social desirability bias*.

Observation #3 - Individual preferences to trade personal data are *seem to be heavy-tailed* (statistical validation pending) in general from CCDF plots of WTP preferences, irrespective of (the amount) of) monetary compensation.

Statistical (In)Validation of Observation #3 - To verify our intuition on the potential heavy-tailedness of WTP preferences, we conduct the following rigorous sequential statistical procedure: (I) find the best fitting power-law distribution to the upper tail - the choice of using a power law is WLOG and one could choose an arbitrary heavy-tailed (HT) distribution to start with, (II) evaluate its statistical plausibility using the standard χ^2 goodness-of-fit test, (III) compare the plausibility to alternative distributions fitted to the same part of the upper tail using the standard Vuong normalized likelihood-ratio test (see (Ranjan Pal 2021) for rationale on why they give better performance; also see (Ranjan Pal 2021) for formally analyzing the statistical procedure.) (Vuong 1989) - in our work we consider Log-normal (HT), Weibull (HT), and Exponential distributions (non-HT) as alternatives to the power-law HT distribution, and (IV) choose the distribution that has the maximum plausibility. Post application of this statistical methodology, we observe that human preferences to trade personal data (irrespective of receiving privacy training and/or monetary incentives), i.e., WTP values follow either (a) a weak power law (WPL) if we club buckets 9 and 10 together (given it might be difficulty for individuals to behaviorally perceive the difference between buckets 9 and 10), or (b) a log-normal distribution otherwise. *A distribution is termed as a WPL if for at least 50% of the application categories, a power-law distribution cannot be rejected (p-value* ≥ 0.1*), and the power-law region consists of at least 50 individuals (Broido and Clauset 2019)*. We emphasize here that though for each application type, we get a different *p*-value, all of them are above the threshold value of 0.1 - the condition that says that the hypothesis of WTP preferences being a power-law distribution cannot be rejected. *Additional fundamental details on distribution fitting, alternative distributions, and likelihood-ratio tests relevant to the analysis methodology is provided in (Ranjan Pal 2021)* .

Behavioral Rationale behind Observation #3 - The evidence of heavy-tailed weak power law relationships with respect to human personal data trading preferences with and without incentives clearly suggests that

the former are correlated via a social phenomenon (induced by behavioral economic rationale) catalyzed by a highly inequitable (and generally privacy illiterate) LMIC economy. More specifically, most of the surveyed population in this economy socially share the feeling to increment their average daily income by even a dollar by trading their personal information. In addition, privacy literate but non-sensitive individuals clearly do not want to miss out on opportunities to trade their personal data when they share a common belief that data commercialization is inevitable - more so in the wake of recent data scandals such as the UIDAI database breach in India, and Cambridge Analytica worldwide. Only a relatively much smaller portion of the surveyed population (of highly privacy sensitive and income oblivious individuals) are averse to personal data trading.

4 SUMMARY AND BROADER IMPACTS

As paper summary, we experimentally studied the impact of monetary incentives on the personal data trading preference of ≊2.5K human subjects in India having access to mobile apps. More specifically, influenced by our pre-conceived hypothesis that monetary incentives will influence people to give away their personal data irrespective of their prior knowledge of privacy and its related hygiene, we wished to test the validity of the hypothesis by first 'warning' society about the pitfalls of data leakage and then observing changes in their preferences to trade their personal data. To this end, we conducted large-scale non-online randomized controlled trial (RCT) experiments on a variety of human subjects in India from 2014-2019, and those using the most popular categories of apps. The RCT included an intervention step where we provided educational awareness on privacy-hygiene to a selected group of subjects, post which they were provided a survey form to elucidate their personal data trading preferences as a function of varying amount of monetary incentives. The remaining group were not subject to any educational intervention but participated in the survey. We observed the increase/decrease trend of human preferences to personal data trading for all the popular app categories for these two groups. Not surprisingly, for highly inequitable and low-medium economies like India, we observed more than 50% of the (generally privacy-illiterate) subject population (irrespective of whether they underwent the training program or otherwise) increased their preference to trade data with increases in remuneration, for all the app categories - clearly establishing (via statistical significance tests) that money (incentives) overpowers education-laden privacy sensitivities in the current age and possibly putting more weight on the privacy paradox. However, surprisingly, we observed that the statistical distribution of preferences (with or without incentives) is heavy-tailed and hints at being weakly scale-free in most app cases. This hints at the mathematical nature of underlying social patterns that influence information trading mindsets, to be similar to those arising in social networks - indicating a possibility that individuals in social networks might influence (induced by behavioral economic rationale) one another on personal data trading preferences, being catalyzed by a highly inequitable (and generally privacy illiterate) LMIC economy. *As part of future work*, we wish to investigate using ML algorithms, the effect of individual demographic factors on a person's willingness to trade personal data. The broader impact of our proposed research will have policy, socio-economic, and technology implications. On the policy front, our research outcomes will hint at creating/modifying/rethinking existing data protection regulations/policies (e.g., parallel to the GDPR and CCPA in the western world) in India-like socio-political economies that can better tie human preferences about commercial data use benefits and associated privacy fears.

On the socio-economic dimension, our research will hint at a possible demographic blueprint on how commercial and non-profit data collectors in the subcontinent, and over a wide application space, should target consumers and manage their data. Moreover, with data science technologies seeping into, and enhancing every aspect of our lives, a country's economy is no exception, and this science can be used in a host of initiatives, with the aim of a more 'equal' distribution of wealth along with increasing GDP for highly inequitable but GDP-rich economies like India. Finally, on the technical front, were HCDEs to become a reality in the (near) future motivated by a social push, would call for the design of (a) demographic and application specific privacy preserving mechanisms and subsequent economic contracts that result in

WTP for Various Categories. (9+10 in (g)(h): the combination of the population with score 9 and 10)

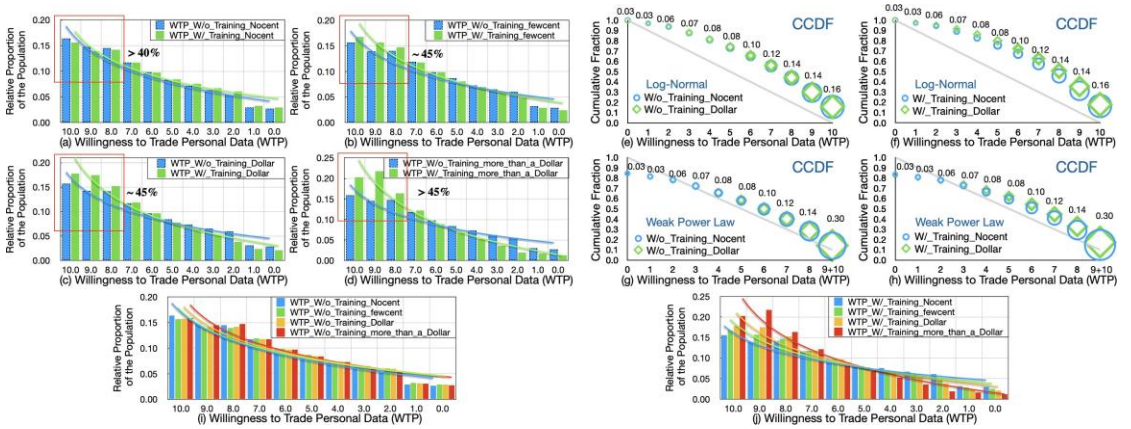

Figure 1: WTP for *House and Home* Category

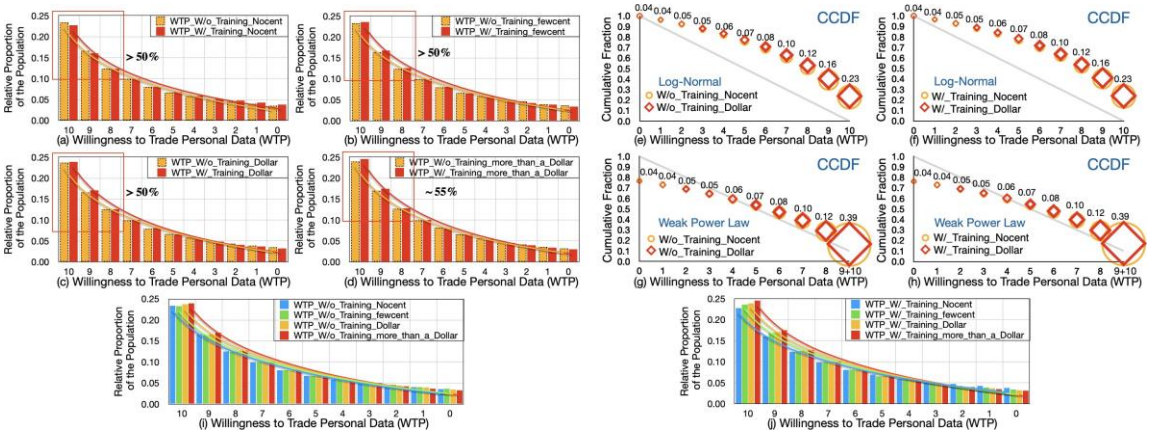

Figure 2: WTP for *Medical* Category

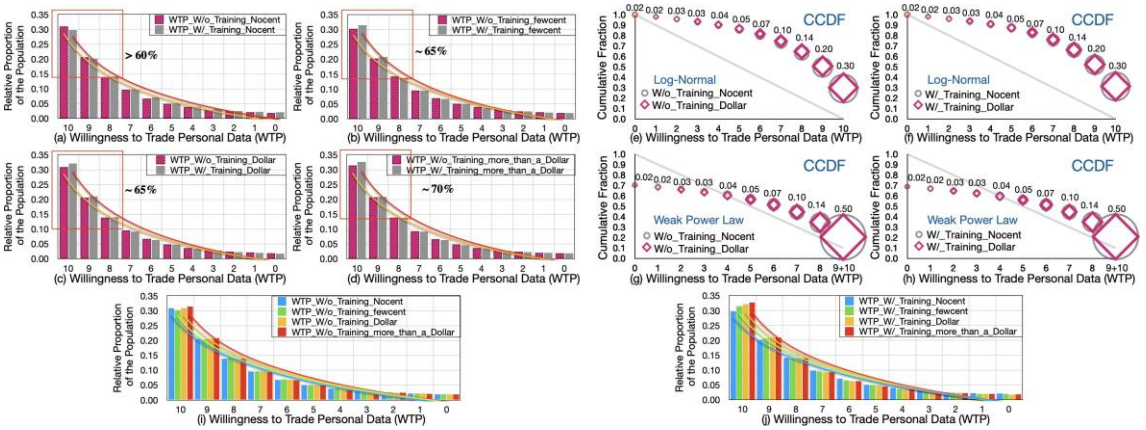

Figure 3: WTP for *Social Communication* Category

suitable privacy-utility tradeoffs, and (b) less computationally intensive security and privacy enhancing (cryptographic) protocols to enforce digital contracts with fair remuneration between the data sellers and the buyers.

ACKNOWLEDGMENTS

All the authors contributed in equal capacity for this work. The authors would also like to acknowledge Jon Crowcroft (University of Cambridge), Sagar Joglekar (Nokia Bell Labs), Ramesh Johari, and Johan Ugander (Stanford University) for their constructive comments on our experimental setup. Finally, the authors would like to acknowledge all support staff members associated with our experiments. This work is partly supported by (a) the NSF under grants CNS-1616575, CNS-1939006, CNS-1816887 and (b) the Army Research Office under grant ARO W911NF1810208.

REFERENCES

Ackerman, M. S., L. F. Cranor, and J. Reagle. 1999. "Privacy in e-commerce: examining user scenarios and privacy preferences". In *Proceedings of the 1st ACM conference on Electronic commerce*, 1–8.

Acquisti, A., L. Brandimarte, and G. Loewenstein. 2015. "Privacy and human behavior in the age of information". *Science* 347(6221):509–514.

Ameringer, S., R. C. Serlin, and S. Ward. 2009. "Simpson's paradox and experimental research". *Nursing research* 58(2):123.

Angner, E. 2012. *A course in behavioral economics*. Macmillan International Higher Education.

Appleton, D. R., J. M. French, and M. P. Vanderpump. 1996. "Ignoring a covariate: An example of Simpson's paradox". *The American Statistician* 50(4):340–341.

Barnes, S. B., and A. P. Paradox. 2006. "Social networking in the United States". *Peer-Reviewed Journal on the Internet*.

Benndorf, V., and H.-T. Normann. 2018. "The willingness to sell personal data". *The Scandinavian Journal of Economics* 120(4):1260–1278.

Broido, A. D., and A. Clauset. 2019. "Scale-free networks are rare". *Nature communications* 10(1):1–10.

Carrascal, J. P., C. Riederer, V. Erramilli, M. Cherubini, and R. de Oliveira. 2013. "Your browsing behavior for a big mac: Economics of personal information online". In *Proceedings of the 22nd international conference on World Wide Web*, 189–200.

Center, P. R. 2019. "Mobile fact sheet". *Internet & Technology.*.

Gerber, N., P. Gerber, and M. Volkamer. 2018. "Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior". *Computers & Security* 77:226–261.

Grossklags, J., and A. Acquisti. 2007. "When 25 Cents is Too Much: An Experiment on Willingness-To-Sell and Willingness-To-Protect Personal Information.". In *WEIS*.

Jones, C. I., and C. Tonetti. 2019. "Nonrivalry and the Economics of Data". Technical report, National Bureau of Economic Research.

Kemp, S. 2019. "Digital trends 2019: Every single stat you need to know about the internet". *The Next Web*.

Kendall, J. 2003. "Designing a research project: randomised controlled trials and their principles". *Emergency medicine journal: EMJ* 20(2):164.

Kokolakis, S. 2017. "Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon". *Computers & security* 64:122–134.

Lanier, J. 2014. *Who owns the future?* Simon and Schuster.

Laoutaris, N. 2019. "Why Online Services Should Pay You for Your Data? The Arguments for a Human-Centric Data Economy". *IEEE Internet Computing* 23(5):29–35.

Laudon, K. C. 1996, September. "Markets and Privacy". *Commun. ACM* 39(9):92–104.

Maniaci, M. R., and R. D. Rogge. 2014. "Caring about carelessness: Participant inattention and its effects on research". *Journal of Research in Personality* 48:61–83.

Milletler, B. 2019. "Data Economy: Radical Transformation or Dystopia?". *Frontier Technology Quarterly*.

Odlyzko, A. 2003. "The case against micropayments". In *International Conference on Financial Cryptography*, 77–83. Springer.

Pal, R., and J. Crowcroft. 2019. "Privacy trading in the surveillance capitalism age viewpoints on'privacy-preserving'societal value creation". *ACM SIGCOMM Computer Communication Review* 49(3):26–31.

Pal, R., J. Crowcroft, Y. Wang, Y. Li, S. De, S. Tarkoma, M. Liu, B. Nag, A. Kumar, and P. Hui. 2020. "Preference-Based Privacy Markets". *IEEE Access* 8:146006–146026.

Pal, R., J. Li, J. Crowcroft, Y. Li, M. Liu, and N. Sastry. 2020. "Privacy Risk is a Function of Information Type: Learnings for the Surveillance Capitalism Age". *IEEE Transactions on Network and Service Management.*

Pal, R., Y. Wang, J. Li, M. Liu, J. Crowcroft, Y. Li, and S. Tarkoma. 2020. "Data Trading with Competitive Social Platforms: Outcomes are Mostly Privacy Welfare Damaging". *IEEE Transactions on Network and Service Management*.

Perera, R. 2006. "Statistics and death from meningococcal disease in children". *British Medical Journal* 332(7553):1297–1298.

Posner, E. A., and E. G. Weyl. 2018. *Radical markets: Uprooting capitalism and democracy for a just society*. Princeton University Press.

Ranjan Pal, Yixuan Wang, Rohan Sequeira, Mingyan Liu, Pradipta Ghosh, Leana Golubchik, Bodhibrata Nag, Tathagata Bandyopadhyay 2021. "Appendix of How do Humans Value Data Privacy in Smart Cities?". Available at https://drive.google.com/file/d/1f6AHAfpPiCBFWp_Y-zq9RdZdfYOK628E/view?usp=sharing.

Reintjes, R., A. de Boer, W. van Pelt, and J. Mintjes-de Groot. 2000. "Simpson's paradox: an example from hospital epidemiology". *Epidemiology* 11(1):81–83.

Samuelson, P. 2000. "Privacy as intellectual property?". *Stanford law review*:1125–1173.

Schwartz, P. M. 2003. "Property, privacy, and personal data". *Harv. L. Rev.* 117:2056.

Stigler, G. 1978. "An Introduction to Privacy in Economics and Politics". *Journal of Legal Studies* 9(4).

Taddei, S., and B. Contena. 2013. "Privacy, trust and control: Which relationships with online self-disclosure?". *Computers in Human Behavior* 29(3):821–826.

Taylor, C., A. Acquisti, and L. Wagman. 2016. "The economics of privacy". *Journal of Economic Literature* 54(2):442–92.

Trepte, S., and L. Reinecke. 2011. *Privacy online: Perspectives on privacy and self-disclosure in the social web*. Springer Science & Business Media.

Turner, A 2020. "How many smartphones are in the world".

Varian, H. R. 2009. "Economic aspects of personal privacy". In *Internet policy and economics*, 101–109. Springer.

Vuong, Q. H. 1989. "Likelihood ratio tests for model selection and non-nested hypotheses". *Econometrica: Journal of the Econometric Society*:307–333.

Wagner, C. H. 1982. "Simpson's paradox in real life". *The American Statistician* 36(1):46–48.

Walsh, M., S. K. Srinathan, D. F. McAuley, M. Mrkobrada, O. Levine, C. Ribic, A. O. Molnar, N. D. Dattani, A. Burke, G. Guyatt et al. 2014. "The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index". *Journal of clinical epidemiology* 67(6):622–628.

Westin, A. F. 2003. "Social and political dimensions of privacy". *Journal of social issues* 59(2):431–453.

Wu, T. 2017. *The attention merchants: The epic scramble to get inside our heads*. Vintage.

Zuboff, S. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power: Barack Obama's Books of 2019*. Profile Books.

AUTHOR BIOGRAPHIES

RANJAN PAL is a faculty member of ECE at University of Michigan Ann Arbor. His primary research interest is in engineering robust cyber-security and information privacy solutions using decision and the applied mathematical sciences. His email address is palr@umich.edu.

YIXUAN WANG is a graduate student in the School of Computer Science at Carnegie Mellon University. Her research interests lie in information privacy. Her email address is yixuanwa@andrew.cmu.edu.

CHARLES LIGHT is a graduate student in ECE at the University of Michigan Ann Arbor. His research interests lie in cyber-security and its economics. His email address is cxlight@umich.edu.

YIFAN DONG is an undergraduate student in ECE at the University of Michigan Ann Arbor. His research interests lie in information privacy. His email address is spikedyf@umich.edu.

PRADIPTA GHOSH is a research engineer at Facebook, USA. His research interests lie in mobile and wireless networks. His email address is pradiptg@usc.edu.

MINGYAN LIU is an entrepreneur and the Peter and Evelyn and Fuss Chair Professor of ECE at University of Michigan. Among her diverse research interests include cyber-security and information privacy. Her email address is mingyan@umich.edu.

HARSHITH NAGUBANDI is a graduate student in ECE at the University of California San Diego. His research interests lie in communication networks. His email address is hnagubandi@ucsd.edu.

LEANA GOLUBCHIK is the Stephen and Eta Varra Professor of Computer Science at the University of Southern California. Her diverse research interests include cyber-security and information privacy. Her email address is leana@usc.edu.

BODHIBRATA NAG is a professor of Operations Management at the Indian Institute of Management Calcutta. Among his diverse research interests include cyber-security and information privacy management. His email address is bnag@iimcal.ac.in.

SWADES DE is a professor of Electrical Engineering at the Indian Institute of Management Calcutta. Among his diverse research interests include cyber-security and information privacy management. His email address is swadesd@ee.iitd.ac.in.