Stephen Scarano
November 2nd, 2020

An Investigation into the Impact of Data Commerce and Distribution on the Public Good

Summary

This paper seeks to originate an in-depth understanding of how the data market has compromised the best interests of the average citizen, if at all. Before analyzing such a tension, it is necessary to establish a broad legal and economic context on the history of information exchange. Analysis of data-commerce in respect to consumer interest must not limit itself to one company's efficacy, and therefore this paper's synthesis notes the impact of the brokerage industry as a whole. Specifically, research has focussed on the interaction between these organizations and their sale of individual's identities and data; in large part, ignoring the manner of their collaboration with clients.

Finding – The Business of Data Brokerage has Ballooned Significantly Since the 1990s

Data brokerage, the practice of collecting and selling consumer information to businesses, has a long-established market history in the United States of America. However, following the invention of the internet, collection became significantly less case by case and decreasingly localized, resulting in the mass aggregation of consumer information [4]. The highly profitable and loosely-regulated industry is now worth an estimated 300 billion dollars, and the Arkansas-based company Acxiom maintains and analyzes a database of more than 144 million households, averaging 1,500 attributes per entry [6]. The scope of the market exceeds intuition. A study published in the Association of Computing Machinery finds that roughly 90% of US-based Facebook accounts are linked to broker databases [5]. The New York Times reports that Clearview AI has consolidated upwards of 3 billion facial images, and that its CEO Mr. Ton-That has sold that information to both local police departments and the Federal Bureau of Investigations [7]. As of 2016, the faces of nearly half of all American adults are in police databases [8].

Finding – Data De-Identification Measures Are Weak to Outside Aggregation and Synthesis

Data anonymization remains the primary standard through which companies ethically sell information to brokers, but many remain skeptical as to the degree of that confidentiality. While information may be anonymized in a vacuum, cross-referencing data with publicly available sets can lead to reidentification. A case study as early as 2008 from the University of Texas at Austin follows the deanonymization of a Netflix database. The company publicly released ratings information with omitted identifying characteristics for a cash prize contest; subsequently, researchers demonstrated that through a combination of public data sets, they could re-identify Netflix's ratings and, more importantly, spotlight potentially private, sensitive information. For example, the study identified one user as highly rating *Jesus of Nazarus*, *The Gospel of John*, *Power and Terror: Noam Chomsky in Our Times*, *Fahrenheit 9/11*, *Bent*, and *Queer as Folk,* the scope of which provide insight into the individual's religious beliefs, political leanings, and sexual orientation [1].

Later studies have applied similar techniques more broadly as more data sets become publicly available, providing statistical models of re-identification from incomplete databases [2]. A 2015 research paper from the Conference on Online Social Networks demonstrates additional techniques to identify information pertaining to an individual's child or children. Through both Facebook and anonymized broking data, researchers profiled children by matching their parents to public voter registration information [3]. Statisticians and researchers illustrate the increasingly fragile oversights in standard data de-identification processes, advocating for greater holistic protections rather than the removal of identifying fields.

Finding – Information is not always accurate, and individuals face a difficult task in removing themselves

In contrast to debate regarding privacy infringement, studies pay little attention to the accuracy of the collected data. A research analysis published through the Association of Computing Machinery audited several dominant data brokerage companies, finding that 40% of stored attributes were described by participants as "not at all accurate" across companies [5]. Additionally cryptic—Acxiom, Epsilon, and similar services prove either uncooperative or intensely vague regarding the sources of such information. When interrogated by the United States Congress in 2012, both organizations refused to identify what organizations or resources provided their intelligence [6]. The Fair Credit Reporting Act (FCRA) requires entities to remove present misinformation, but mandates little incentive to update data without active consumer complaint [4]. If a customer does decide to opt out, the consumer must in most cases submit additional information and mail a check of 5 US dollars. Some may provide no clear path forward at all [5].

Conclusion

In the face of a largely unregulated industry, a considerable solution to the listed and serious concerns takes the form of systemic enforcement of citizen interest. We may look toward the European Union's data legislation. By law, the E.U. considers privacy a fundamental right, limiting the freedom of brokerage companies and banning the collection of personal data for an undetermined purpose [9]. The above findings also suggest the implementation of a revised FCRA which mandates that the specified collection organizations keep information accessible and updated to citizens. Enforced accessibility grants citizens visibility of their digital representation and does not rely on erroneously de-identified data sets. While a less ambitious change, required updates ensure the universal application of government-expungement, mandating the removal of inaccurate or unrepresentative information. Under such law, the government would bar the buying and selling of information judicially stricken from a person's record [4]. Data commerce as it stands poses threat to consumer privacy, safety, and opportunity, yet the state has taken little action in curtailing the bloated industry at the heart of the problem. Proper regulation must not only elucidate the subject of its trade, but empower individuals to challenge it.

Works Cited

[1] Arvind Narayanan. Vitaly Shmatikov. Robust De-anonymization of Large Sparse Datasets. *IEEE Symposium on Security and Privacy* **1** (2008).
DOI: https://doi.ieeecomputersociety.org/10.1109/SP.2008.33

[2] Rocher, L., Hendrickx, J.M. & de Montjoye, Y. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun* **10,** 3069 (2019). DOI: https://doi.org/10.1038/s41467-019-10933-3

[3] Ratan Dey. Tehila Minkus. Yuan Ding. Keith W. Ross. The City Privacy Attack: Combining Social Media and Public Records for Detailed Profiles of Adults and Children. *COSN '15* (2015). DOI: https://doi.org/10.1145/2817946.2817957

[4] Logan Wayne. The Data Broker Threat: Proposing Federal Legislation to Protect Post-Expungement Privacy. J. *Crim. L. & Criminology* **102,** 253 (2012). https://scholarlycommons.law.northwestern.edu/jclc/vol102/iss1/8

[5] Alan Mislove. Elissa Redmiles. Giridhari Venkatadri. Krishna Gummadi. Michelle Mazurek. Oana Goga. Piotr Sapiezynski. Auditing Offline Data Brokers via Facebook's Advertising Platform. WWW '19 (2019). 1920-1930. DOI: https://doi-org.silk.library.umass.edu/10.1145/3308558.3313666

[6] Leanne Roderick. Discipline and Power in the Digital Age: The Case of the US Consumer Data Broker Industry. *Critical Sociology* **40**, 5 (2014). DOI https://doi.org/10.1177/0896920513501350

[7] Kashmir Hill. The Secretive Company That Might End Privacy as We Know It. 2020. Retrieved October, 2020 from https://nyti.ms/3eihF1E

[8] Alvaro Bedoya. Clare Garvie. Jonathan Frankle. The Perpetual Line-Up: Unregulated Police Face Recognition in America. Georgetown Law Center on Privacy & Technology, (2016). DOI: https://www.perpetuallineup.org/

[9] Chih-Liang Yeh. Pursuing consumer empowerment in the age of big data: A comprehensive regulatory framework for data brokers. J. Science Direct **42,** 4 (2018). 282-292. DOI: https://doi.org/10.1016/j.telpol.2017.12.001