Can "Conscious Data Contribution" Help Users to Exert "Data Leverage" Against Technology Companies?

NICHOLAS VINCENT AND BRENT HECHT, Northwestern University

1. INTRODUCTION

There is growing concern about the serious negative societal impacts of data-driven technologies operated by large technology companies (e.g. privacy erosion, harms to democracy, economic inequality). However, existing power dynamics between the public and tech companies limit the public's ability to change tech company behavior (Ho 2019; Hecht et al. 2018). For instance, attempts to exert leverage against tech companies through consumer protest, e.g. boycotts, must contend with the market power of large tech companies (Herndon 2019; Posner and Weyl 2018; Rogoff 2019).

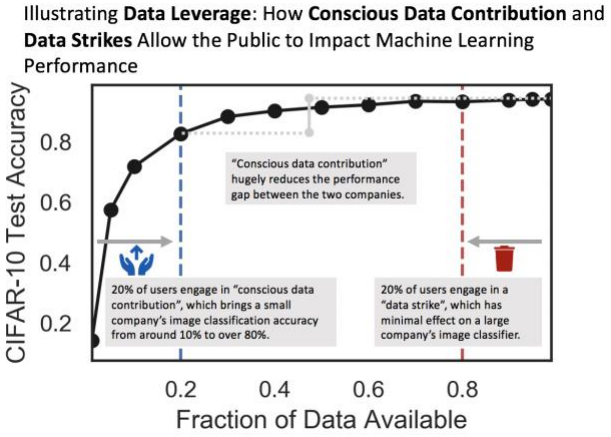

Fig. 1. Illustrates two forms of data leverage: conscious data contribution and data strikes.

In this abstract, we explore how the public might additionally exert **data leverage** against tech companies. Users (i.e. the public) play a critical role in the economic success of tech companies by providing training data—i.e. "data labor" (Arrieta Ibarra et al. 2018)—that is critical to the operation of data-driven technologies (Arrieta Ibarra et al. 2018; Posner and Weyl 2018; Vincent, Hecht, and Sen 2019; Vincent et al. 2019; McMahon, Johnson, and Hecht 2017). The literature studying this topic suggests that users can use this data labor role as a new form of leverage. Moreover, recent research indicates fertile ground exists for actioning this leverage: 30% of U.S.-based respondents reported they already stop or change their technology use as a form of protest against tech companies (Li et al. 2019).

In prior work, we identified one form of data leverage: **data strikes** (Vincent, Hecht, and Sen 2019). In a data strike, a group of users who wish to protest the values or actions of a tech company withholds and/or deletes their data contributions to reduce the performance of the company's data-driven technologies. While our prior work found through simulations that data strikes might be effective, data strikes must contend with the diminishing returns of data to machine learning (ML) performance (Hestness et al. 2017). This means that a small data strike will likely have a very small effect on other users. Additionally, a user who participates in a data strike hinders their own ability to benefit from personalization-based ML systems, which may make participation hard to sustain.

Here, we propose and evaluate an alternative means for users to exert data leverage against tech companies: **conscious data contribution** (henceforth CDC). In CDC, a group of users who wishes to protest a tech company contributes their data to a competing institution (e.g. another tech company) whose values or actions with which they agree more. They can additionally delete their data from the offending company's dataset, effectively combining a data strike and CDC. A group of users could even

stand up a new competitor in the market using CDC. CDC leverages the fact that data is "nonrival"—many firms can use the same data (Jones and Tonetti 2019).

In theory, CDC has two desirable characteristics compared to data strikes. First, CDC is more realistic within short-term time frames, which is important given the growing demand for immediate changes to the power dynamics between users and tech companies. In particular, CDC can utilize increasing support for data portability from regulators and tech companies (Herrman 2018). Second, while small data strikes must fight an uphill battle against the diminishing returns of data, CDC uses diminishing returns to its advantage. Even small contributions of data could hugely improve the performance of a CDC beneficiary's data-driven technologies, helping it to compete with the target of a protest.

To begin to understand how CDC might work in practice, we simulated CDC – both with and without data deletion from the offending company – for four widely-studied ML use cases: two recommendation tasks and two classification tasks. For context, we also simulated data strikes (i.e. data deletion only). We measured the effectiveness of *CDC*, *CDC-with-deletion,* and *data strikes* by defining a **Data Leverage Power** metric that allows us to compare the ML performance of a simulated large, data-rich incumbent company (the target of CDC) with that of a small competitor (the beneficiary of CDC). Our findings suggest that CDC with small participation rates can have large effects in terms of reducing the gap between a data-rich incumbent and its small competitor. If just 20% of users participate, the small competitor can get 70% of the way towards best-case performance for all our ML use cases. In certain situations, just 5% of users can get the small competitor 50% of the way to best-case performance, and 20% of users can get the small competitor 90% of the way. Furthermore, while we must be cautious in directly comparing the Data Leverage Power of CDC and data strikes because they operate differently (i.e. helping a competitor vs. directly hurting a company), we see that CDC is highly effective even when deletion of data from the offending company is impossible and may be more powerful than data strikes for many real-world contexts with small to medium participation rates.

2. EXPERIMENTS

We conducted a series of experiments to compare the ML performance of two simulated companies with access to different data sets. For each ML use case, we assume the following scenario: (1) There is a large, data-rich incumbent company (called "Large Co.") that starts with a full dataset. (2) Some users of Large Co.'s data-driven technologies are interested in protesting Large Co. because of its values or actions. (3) To do so, they want to support an existing small, data-poor competing company – "Small Co." that better aligns with their values. We considered variations in this scenario in which users can contribute data to Small Co.'s dataset while deleting it from Large Co.'s dataset (*CDC-with-deletion*) as well as variations in which deletion is impossible (*CDC-only*). For additional context, we also considered variations in which users only engage only in a data strike (*Deletion-only*).

Our experiments follow procedures similar to those used by learning curve research that has sought to understand the relationship between ML performance and training dataset size (Hestness et al. 2017). Our procedure involved identifying a highly accurate, commonly-used ML approach for each use case, repeatedly retraining the corresponding model with samples of the benchmark training set corresponding to different CDC and/or deletion participation rates (e.g. 1%, 5%, etc.), and evaluating model performance.

Data leverage simulations have two major differences from learning curve simulations. First, for recommendation datasets (which attribute data points to users), we randomly sample users (e.g. 1%, 5%, etc.) to engage in CDC and/or data deletion. For our classification datasets, we randomly sample data points directly, as in learning curve research. Second, when simulating CDC, we can evaluate each company's model using a test set split from that company's data sample (e.g. Small Co. receives a 20%

sample of data points, and creates a test set by splitting 10% of this sample) or we can create a test set that is hidden from each company, which allows us to measure how good a company's performance might be for users accessing the technology anonymously (e.g. a user who receives recommendations in "Private Browsing" mode). In this extended abstract, we focus on performance evaluated by each simulated company, which corresponds to how well each Co.'s system works for their current users.

In our experiments, we consider two recommender use cases and two classification use cases: recommenders that predict movie ratings (using the MovieLens 10-M dataset) (Rendle, Zhang, and Koren 2019; Harper and Konstan 2016), recommenders that predict Pinterest interactions (Dacrema, Cremonesi, and Jannach 2019), classification of the CIFAR-10 image dataset (Page 2018), and classification of the Wikipedia Toxic Comments dataset (Guocan 2018). Each ML use case uses a different metric, so to measure data leverage effectiveness, we introduce a context-agnostic measurement: Data Leverage Power (DLP). DLP considers ML performance relative to the gap between baseline performance (e.g. "random guess" or "recommend most popular" approaches) and "best-case" performance achieved with access to all data. For *CDC-only* and *CDC-with-deletion*, DLP tells us how far a CDC group gets Small Co.'s performance from baseline (DLP of 0) to best-case (a DLP of 1). For instance, a DLP of 0.5 means Small Co.'s performance is halfway from baseline to best-case. More formally, when users engage in CDC, DLP is defined as Small Co.'s performance improvement relative to baseline divided by the maximal gap between best-case and baseline. If users only engage in data deletion (data strike only), Small Co.'s performance is fixed, so instead DLP is defined as Large Co.'s performance loss relative to the gap between best-case (no strike) and baseline.

Looking at Fig. 2, we can see how effective CDC is across our ML use cases. Leveraging diminishing returns, CDC can be highly effective at allowing a small company to drastically reduce the performance gap between itself and a large competitor. We see that CDC by a small group (e.g. 10-20% of users) can get various models 80-90% of the way towards Large Co.'s performance. Furthermore, CDC can be very effective even if data strikes are not possible (the black *CDC-only* DLP results are very similar to the blue *CDC-with-deletion* results). Finally, we see that for small participation rates, CDC exerts much more DLP than data strikes alone (the red *Deletion-only* results).

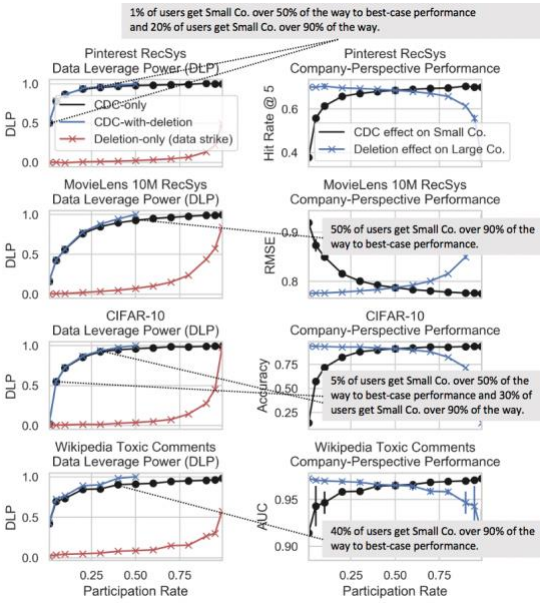

Fig. 2. The left column shows DLP, while the right column shows performance metrics for Small Co and Large Co. Error bars show standard deviation for five random groups per participation rate.

3. DISCUSSION

We observed that CDC can be highly effective in reducing the performance gap between two competitors. Constituencies interested in creating more competition between data-driven technologies may wish to further investigate CDC itself (e.g. conducting similar experiments to those described here) and explore avenues for making CDC easier for potential participants. Specifically, policymakers and advocates might push for data portability regulation and tools, an area of growing discussion (Rossi and Slaiman 2019; Doctorow 2019).

REFERENCES

Arrieta Ibarra, Imanol, Leonard Goff, Diego Jiménez Hernández, Jaron Lanier, and E Weyl. 2018. "Should We Treat Data as Labor? Moving Beyond 'Free.'" *American Economic Association Papers & Proceedings* 1 (1).

Dacrema, Maurizio Ferrari, Paolo Cremonesi, and Dietmar Jannach. 2019. "Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches." In *Proceedings of the 13th ACM Conference on Recommender Systems*, 101–109. ACM.

Doctorow, Cory. 2019. "Regulating Big Tech Makes Them Stronger, so They Need Competition Instead." *The Economist*, June 2019. https://www.economist.com/open-future/2019/06/06/regulating-big-tech-makes-them-stronger-so-they-need-competition-instead.

guocan. 2018. "Logistic Regression with Words and Char." Kaggle. February 2018. https://www.kaggle.com/guocan/logistic-regression-with-words-and-char-n-g-13417e.

Harper, F Maxwell, and Joseph A Konstan. 2016. "The Movielens Datasets: History and Context." *Acm Transactions on Interactive Intelligent Systems (Tiis)* 5 (4): 19.

Hecht, Brent, Lauren Wilcox, Jeffrey P Bigham, Johannes Schöning, Ehsan Hoque, Jason Ernst, Yonatan Bisk, Lana Yarosh, Bushra Amjam, and Cathy Wu. 2018. "It's Time to Do Something: Mitigating the Negative Impacts of Computing through a Change to the Peer Review Process." *ACM Future of Computing Blog*.

Herndon, Astead W. 2019. "Elizabeth Warren Proposes Breaking Up Tech Giants Like Amazon and Facebook." *The New York Times*, March 10, 2019, sec. U.S. https://www.nytimes.com/2019/03/08/us/politics/elizabeth-warren-amazon.html.

Herrman, John. 2018. "Google Knows Where You've Been, but Does It Know Who You Are?" *N.Y. Times*, September. https://www.nytimes.com/2018/09/12/magazine/google-maps-location-data-privacy.html.

Hestness, Joel, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. 2017. "Deep Learning Scaling Is Predictable, Empirically." *ArXiv Preprint ArXiv:1712.00409*.

Ho, Vivian. 2019. "Tech Monopoly? Facebook, Google and Amazon Face Increased Scrutiny." *The Guardian*, June. https://www.theguardian.com/technology/2019/jun/03/tech-monopoly-congress-increases-antitrust-scrutiny-on-facebook-google-amazon.

Jones, Charles I, and Christopher Tonetti. 2019. "Nonrivalry and the Economics of Data." National Bureau of Economic Research.

Li, Hanlin, Nicholas Vincent, Janice Tsai, Jofish Kaye, and Brent Hecht. 2019. "How Do People Change Their Technology Use in Protest?: Understanding 'Protest Users.'" *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): 87.

McMahon, Connor, Isaac L Johnson, and Brent Hecht. 2017. "The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies." In *ICWSM*, 142–151.

Page, David. 2018. "How to Train Your ResNet." September 2018. https://myrtle.ai/how-to-train-your-resnet.

Posner, Eric A, and E Glen Weyl. 2018. *Radical Markets: Uprooting Capitalism and Democracy for a Just Society*. Princeton University Press.

Rendle, Steffen, Li Zhang, and Yehuda Koren. 2019. "On the Difficulty of Evaluating Baselines: A Study on Recommender Systems." *ArXiv Preprint ArXiv:1905.01395*.

Rogoff, Kenneth. 2019. "Big Tech Has Too Much Monopoly Power – It's Right to Take It On." *The Guardian*, April. https://www.theguardian.com/technology/2019/apr/02/big-tech-monopoly-power-elizabeth-warren-technology.

Rossi, Gus, and Charlotte Slaiman. 2019. "Interoperability = Privacy + Competition." *Public Knowledge*, October. https://www.publicknowledge.org/blog/interoperability-privacy-competition.

Vincent, Nicholas, Brent Hecht, and Shilad Sen. 2019. "'Data Strikes': Evaluating the Effectiveness of New Forms of Collective Action Against Technology Platforms." In *Proceedings of The Web Conference 2019*.

Vincent, Nicholas, Isaac Johnson, Patrick Sheehan, and Brent Hecht. 2019. "Measuring the Importance of User-Generated Content to Search Engines." In *Proceedings of AAAI ICWSM 2019*.