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

NICHOLAS VINCENT, Northwestern University, USA
BRENT HECHT, Northwestern University, USA

Tech users currently have limited ability to act on concerns regarding the negative societal impacts of large tech companies. However, recent work suggests that users can exert leverage using their role in the generation of valuable data, for instance by withholding their data contributions to intelligent technologies. We propose and evaluate a new means to exert this type of leverage against tech companies: "conscious data contribution" (CDC). Users who participate in CDC exert leverage against a target tech company by contributing data to technologies operated by a competitor of that company. Using simulations, we find that CDC could be highly effective at reducing the gap in intelligent technologies performance between an incumbent and their competitors. In some cases, just 20% of users contributing data they have produced to a small competitor could help that competitor get 80% of the way towards the original company's best-case performance. We discuss the implications of CDC for policymakers, tech designers, and researchers.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**

KEYWORDS

conscious data contribution; data leverage; data as labor

ACM Reference Format:

Nicholas Vincent and Brent Hecht. 2021. Can "Conscious Data Contribution" Help Users to Exert "Data Leverage" Against Technology Companies? In *PACM on Human Computer Interaction*, Vol. 5, No. CSCW1, Article 103, April 2021. ACM, New York, NY, USA. 25 pages. https://doi.org/10.1145/3449177

1 INTRODUCTION

There is growing concern about the serious negative societal impacts of intelligent technologies operated by large technology companies. However, existing power dynamics between the public and tech companies limit the public's ability to change tech company behavior with regards to privacy erosion, harms to democracy, economic inequality and other issues [19,23]. 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 [20,45,48].

This was work was funded by NSF grants 1815507 and 1707296.

Authors email addresses: nickvincent@u.northwestern.edu, bhecht@northwestern.edu

https://doi.org/10.1145/3449177

In this paper, we explore how the public might 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" [1,45]—that is existentially important to data-driven technologies (e.g. [55–57]). This means that users can take advantage of their role as data generators to gain leverage against data-dependent firms [57].

Recent research indicates that fertile ground exists for data leverage: in one survey published at CSCW, 30% of U.S.-based respondents reported they have already stopped or changed their technology use as a form of protest against tech companies, while in another survey 33% of U.S.-based respondents reported they believe tech companies have a negative effect on the country [11,36].

Prior work identified one form of data leverage: **data strikes** [55]. In a data strike, a group of users who wishes 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 this research found through simulations that data strikes might be effective, data strikes must contend with the diminishing returns of data to machine learning (ML) performance [22]. Indeed, the ability to generalize from limited data is one reason machine learning is so powerful. 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 in some cases.

In this paper, we propose and evaluate an alternative means for users to exert data leverage against tech companies: **conscious data contribution** (or "CDC"). A group of users who wishes to protest a tech company using CDC contributes its data to a competing institution (e.g. another tech company) that has 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. CDC takes advantage of the fact that data is "nonrival"—many firms can use the same data [25], so deleting data or quitting an existing technology is not a requirement for CDC. A group of people could help to support a new competitor in the market using CDC, without the need to completely quit using existing technologies.

In theory, CDC has two desirable characteristics compared to data strikes. First, CDC is more realistic within short-term time frames (as some data strikes will require support from regulators), which is important given the growing demand for immediate changes to the power dynamics between users and tech companies. In terms of legal support for CDC, regulators in various jurisdictions, e.g. the European Union, are increasingly advancing legislation that protects "data portability", the right for users to receive and re-use their personal data [34]. Tech companies are also supporting data portability features, e.g. Google's "Takeout" feature that allows users to export their data [21]. As data portability laws and features become more common, CDC should become even easier to practice.

Second, while small data strikes must fight an uphill battle against the diminishing returns of data, CDC does not face diminishing returns until participation is high. While small data strikes may have a minor impact on a large company's technologies, 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. In other words, CDC can more easily operate in the "vertical" region of the performance vs. dataset size curve instead of the "horizontal" region of this curve.

The goal of this paper is to begin to understand how CDC might work in practice. To do so, we simulated CDC applied to four widely studied and business-relevant ML tasks: two recommendation tasks and two classification tasks. For context, we also consider data strikes—

combined with CDC and on their own—in which users delete their data from an offending company. In total, we simulated three different data leverage scenarios: *CDC only*, *data strike only,* and *CDC with data strike*. To measure the data leverage achieved in each scenario with different *participation rates*, we compared 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). To enable comparisons across ML tasks with different evaluation metrics, dataset sizes, and data formats, we defined **Data Leverage Power** (DLP), a metric that facilitates cross-task comparison and the comparison of CDC to data strikes. In our analyses, we compare performance using both DLP and traditional ML metrics, which provide a task-specific perspective.

Our findings suggest that CDC with relatively small participation rates can effectively reduce the gap between a data-rich incumbent and its small competitor. If just 20% of users participate in CDC, the small competitor can get at least 80% of the way towards best-case performance for all our ML tasks. In certain situations, participation by 5% of users is enough to boost the small competitor's ML performance to 50% of best-case performance improvement, and 20% of users can get the small competitor 90% of the way.

Our results suggest that CDC may be more powerful than data strikes for many real-world contexts and could provide new opportunities for changing existing power dynamics between tech companies and their users. While we must be cautious in comparing the effects of CDC and data strikes because they operate differently (i.e. helping a competitor vs. directly hurting a company), we see that CDC is effective even when data strikes are impossible. More generally, our simulation experiments highlight how methods from machine learning research can be used to study and change power dynamics between tech companies and the public.

2 RELATED WORK

2.1 Data Leverage and Consumer Leverage

Our interest in the CDC concept was heavily motivated by literature that has studied the use of data leverage to give the public more power in its relations with tech companies (and broadly, any organizations that use data-dependent technologies) [57]. Very recent work has proposed a framework of "data leverage" that consists of three "data levers": data strikes, conscious data contribution, and data poisoning. Data strikes involve data deletion/withholding, conscious data contribution involves data generation/sharing , and data poisoning involves data manipulation aimed at harming data-driven technologies [57]. In this paper, we focus on comparing data strikes and conscious data contribution with simulation experiments. In other words, we focus specifically on two branches of the broader data leverage framework.

Previously, Vincent et al. studied simulated "data strikes" in the recommender system context and found that data strikes of moderate size (e.g., 30-50% of users) could be impactful. Weyl and Lanier [33] have proposed that cooperative entities might guide collective action like data strikes. While the data strikes concept has focused on withholding or deletion of data, Brunton and Nissembaum [4] have written about *obfuscation-based protest* – feeding intelligent technologies junk data to reduce their predictive power, using tools like AdNauseum [24]. Finally, Kulynych, Overdorf, and colleagues [32] laid out a broad framework for "Protective Optimization Technologies" (POTs) – technologies that attack optimization systems like intelligent technologies in order to reduce externalities caused by such systems. This framework is inclusive of any tactics that contest a technology, including the data levers we explore in this paper (data strikes and CDC) or "data poisoning"—the third "data lever" in Vincent et al.'s data leverage framework [57]—which draws on obfuscation, POTs, and machine learning literature on adversarial data.

2.2 The Relationship between Data Leverage and Consumer Leverage

While interest in data leverage is relatively new, a large body of work has studied forms of consumer leverage. The definition of CDC we use here was influenced by existing types of consumer leverage. In particular, the practice of "political consumerism", in which the public uses its consumer purchasing power as a political tool, is a precedent for CDC [40]. Political consumerism includes both boycotts and "buycotts": buying products or services to support a specific company. Buycotts, in contrast to boycotts, represent a "positive approach" to consumer action, as they reward, rather than punishing, a company [15]. Drawing on the dichotomies of data vs. consumer leverage and positive reward-based approaches vs. negative punishment-based approaches, CDC can be seen as a data leverage version of the positive, reward-focused buycotting approach.

Both political consumerism and interest in protest against tech are prevalent, suggesting there may be a large market for CDC. Recent work suggests over 50% of U.S.-based survey respondents having engaged in boycotts or buycotts in 2017 [13]. In another survey, 30% of U.S.-based respondents reported stopping or changing their use of technology companies in particular [36], and some work has begun to design tools for technology-assisted political consumerism [35]. The results of the survey work, in particular, suggest that large-scale CDC is well within the realm of near-term feasibility: if CDC is made straightforward, it seems there are many people who will be interested.

2.3 Learning Curves and Diminishing Returns from Data

Research on the relationship between ML performance and dataset size provides important motivation for conscious data contribution. Many scholars have empirically studied "learning curves" – the relationships between training dataset size and ML performance – for a variety of models. When looking at a specific task (e.g., image classification), learning curves can be characterized as exhibiting diminishing returns. At some point, the relationship between performance and data size becomes "flat" [2,14]. Diminishing returns have been observed in a variety of contexts, for instance statistical machine translation [29], deep learning [5,22,49], logistic regression [44], decision trees [9], and matrix factorization recommender models [55].

The techniques used to study learning curves can be adapted to study data leverage. A typical procedure to generate a learning curve would entail randomly sampling a fraction of training data, retraining a model with this sample of data, and measuring performance of the new trained model. One might repeat this procedure for some fixed number of iterations to obtain the average performance for a random sample of a certain size. By doing this for a variety of fractions, we obtain a curve of performance vs. dataset size. To study data leverage, we are interested in the curve that relates performance to the fraction of users who contribute data (though in some cases we must use fraction of data as a proxy for fraction of users). At a high level, we obtain this curve using the same procedure for computing a learning curve that was used by Perlich et al. [44] and Cho et al. [5]. However, as we will describe below in Methods, there are some additional implementation details that distinguish data leverage simulations from learning curve experiments and help to increase the ecological validity of our simulations.

2.4 Data Leverage and Online Collective Action

The data leverage scenarios we simulated in this paper are instances of online collective action, a topic with a rich literature in social computing. This literature considers online collective action in a broad variety of contexts. These contexts include leadership as a collective activity in Wikipedia

[62], collective action for crowdworkers [51], the use of Twitter bots to organize civic volunteers [52], and the development of tools for "end-to-end" collective action [59]. It is likely that strategies that work well for these contexts may apply directly to data leverage scenarios, in particular by facilitating the organization and participation necessary to achieve large-scale participation.

For instance, leaning on findings from CSCW scholarship about leadership in the Wikipedia community, CDC groups might encourage members who are not formal leaders to take leadership actions using strategies from Wikipedia [61,62]. More generally, CDC stands to benefit from most research on successful peer production (e.g. [17,26]), as peer production participants (especially Wikipedia editors) are already critical sources of data labor that fuels AI systems [37,56], and share core similarities with CDC participants. Concretely, groups engaging in CDC can emulate peer production strategies, and in some cases participating in peer production could be a form of CDC (e.g. contributing labeled images to Wikimedia Commons with the goal of helping start-ups train computer vision models). Another area of CSCW research that is highly relevant to CDC is crowdwork. Although there are major differences between crowdwork and peer production, crowdworkers also provide crucial data labor [27] and have led successful collective action movements in the past [51].

Similarly, looking to research on bots and support tools in social computing, CDC participants might re-use or adapt existing bots, social platforms, and browser plug-ins to promote CDC engagement [35,52,53,59]. Tools to support political consumption, such as browser extensions that help people boycott websites, might be enhanced with CDC features [35].

Social computing researchers have called for more work that addresses "Computer Supported Collective Action" (CSCA) [53]. Generally, any given data leverage scenario (i.e. CDC and/or a data strike by some group) can be seen as an instance of CSCA [53]. This means those seeking to use data leverage must face the many challenges associated with collective action, e.g. challenges with leadership, communication, and planning. Conversely, models of success in CSCA can provide templates for successful CDC.

3 METHODS

We conducted a series of experiments to compare the ML performance of two simulated companies when users exert data leverage using CDC and/or data strikes. For our simulations, we assume the following scenario. There exists a large, data-rich incumbent company – called "Large Co." – that starts with a full dataset. Some users of Large Co.'s data-driven technologies are interested in protesting Large Co. because of its values or actions. To do so, they want to support a 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 simultaneously deleting it from Large Co.'s dataset (*CDC with data strike*) as well as variations in which deletion is impossible (*CDC only*). For additional context, we also considered variations in which users engage only in a data strike (*data strike only*). In all our simulations, data strikers delete all their data contributions and then models are retrained.

To begin our experiments, we first had to identify specific ML tasks to study and a corresponding ML approach to implement for each task. We selected four tasks that have both attracted substantial research attention and have clear industry applications: two recommendation tasks (using movie ratings and Pinterest interaction records) and two classification tasks (images and text). For each task, we sought out a top-performing ML approach with a publicly available implementation. Below, we further detail our simulation assumptions and the specific tasks, datasets, and ML approaches from prior work that we used.

At a high level, our experiments follow procedures similar to those used in learning curve research that has studied the relationship between ML performance and training dataset size [22]. For each task, we repeatedly retrained the corresponding model with samples of the benchmark training set corresponding to different CDC and/or data strike participation rates (e.g. 1%, 5%, etc.), and evaluated model performance.

Specifically, our data leverage simulations have three major differences from traditional learning curve simulations. First, in simulating data leverage scenarios, for datasets with user identifiers, we drew a sample of users to participate in CDC and/or a data strike instead of drawing a sample of data points. This approach simulates what would happen in a CDC scenario, in which data is added or removed on a user-by-user basis. For our classification datasets, which lack user identifiers, we randomly sampled data points, as in learning curve research. This is an inherent limitation, as many influential classification datasets lack user identifiers for privacy reasons.

Second, when simulating CDC, we are primarily interested in ML performance as evaluated from the perspective of each company. To get this *company perspective* evaluation, we use a test set drawn from each company's data sample. For instance, if Small Co. receives data from 10% of users, Small Co. must create its own test set using data from these 10% of users. However, as a secondary measurement, we can also hold out a separate, fixed test set that is hidden from each company. This *fixed holdout* perspective allows us to measure a performance while taking into account people who are accessing the technology as a brand new user (or anonymously, i.e. a user who receives recommendations in "Private Browsing" mode). This *fixed holdout* test set can also be seen as a more objective external measurement of model performance, as the *fixed holdout* set is the same across every simulation and is unaffected by which users or observations are available to a particular company. In a more practical sense, this *company perspective* vs. *fixed holdout* comparison is most relevant to personalization tasks, where performance might differ drastically for new (or anonymous) users.

Third, as mentioned above, our data leverage simulations allow us to consider the case in which one company gains data while another company loses data, i.e., in a data strike. Below, we will discuss how we addressed the challenges in comparing these scenarios (i.e. how do we define a comparison metric that allows us to compare the effectiveness of giving Small Co. data vs. deleting data from Large Co.?)

3.1 Simulation Details

Following past learning curve studies, we considered a range of participation rates. Specifically, we conducted simulations in which a group of users engages in one of the following scenarios: *CDC only*, *data strike only*, or *CDC with data strike*. We considered participation rates of 0.01 (1% of users or data), 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5. We leveraged a natural symmetry of our experiments to estimate effects for larger groups: the situation in which Small Co. has some fraction s of all data is equivalent to the situation in which Large Co. has lost 1 - s of all data. For instance, if Small Co. gains 10% of all users or data, the expected ML performance is the same as when Large Co. has 90% of its users or data deleted. Thus, without running additional experiments, we can also measure ML performance for participation rates of 0.6, 0.7, 0.8, 0.9, 0.95, 0.99. As we will discuss further below, we also consider baseline results (i.e. when Small Co. has very little data) and best-case results (i.e. when Large Co. has all possible training data) in our definition of Data Leverage Power.

The end result of our simulations is to generate a curve showing the effectiveness of data leverage scenarios at different participation rates. While our simulated scenario focuses on just two companies, this curve could be used to evaluate data leverage exerted against many companies. If several companies simultaneously benefit from CDC, we can use each beneficiary company's participation rate to get corresponding DLP. For instance, we might use this curve to compare the performance of one company that received CDC with 10% participation, another company that received CDC with 20% participation, and a third company that was the target of a data strike by 30% of its users.

In each simulation, Small Co. and Large Co. gain access to a scenario-specific dataset, split their respective datasets into a train set and company-specific test datasets, train their models, and evaluate their trained models on two test sets: (1) their *company perspective* test set (which is unique to each scenario) and (2) a *fixed holdout* test set. As mentioned above, the *fixed holdout* test set provides an objective evaluation of performance (it does not depend on the company-specific test set) and is particularly relevant to personalization tasks. We further detail the purpose of considering these two different test sets below.

For each training run, we use the same, fixed hyperparameters that achieved reported results in prior work instead of running hyperparameter search for each simulation. While this substantially reduces the computational cost of our experiments, this means that for any given scenario, our results do not necessarily represent the best possible performance for a given participation rate. For instance, if Small Co. only has 10% data to work with, it's possible Small Co. could use different hyperparameters (or other techniques, such as data augmentation, different training algorithms, etc.). to boost performance. However, by using fixed hyperparameters, we reduce the computational cost and make it easier to explore several tasks, use multiple sampling seeds, and replicate our findings.

More formally, our simulations followed the same procedure for each fractional group size s:

1. Identify the data that will be given to Small Co. via CDC and removed from Large Co. via data strikes. To identify this data, we randomly sampled s (the participation rate) of all users and took the data attributed to the sampled users. For tasks in which units of data are not attributed to specific users (image and text classification), we sampled s of all units of data.
2. Train and evaluate a model using Small Co.'s dataset. This gives us Small Co.'s performance in the *CDC only* scenario.
3. Train and evaluate a model using Large Co.'s dataset. This gives us Large Co.'s performance in both the *CDC with data strike* and *data strike only* scenarios.
4. Compare Small Co.'s performance and Large Co.'s performance to worst-case and best-case performance values. As we will describe below, we formalized this comparison by defining "Data Leverage Power", a measurement that we used (alongside traditional ML metrics) to compare data leverage across different ML tasks. At a high level, DLP measures how close CDC gets Small Co. to Large Co.'s performance, or how close a data strike moves Large Co.'s performance to low-data baseline performance. We expand on DLP's precise definition and the motivation underlying the measurement below.

In order to measure average performances for each participation rate, we repeated this procedure using five different seeds for sampling s users/data and calculated the average performance across these iterations. Thus, for each ML task we studied, we retrained a model 70 times (7 group sizes, 5 sampling seeds, and 2 different companies), which was made easier by our choice to avoid costly hyperparameter searches.

3.2 Tasks, Datasets and Machine Learning Implementations

As mentioned above, we first identified ML tasks of interest to ML researchers in industry and academia, and then sought out a publicly available implementation of a high-performing ML approach for each task. To make experimentation feasible, we also needed to identify implementations that would be possible to retrain many times. We identified four public implementations of successful approaches to well-studied, industry relevant ML tasks: Rendle et al.'s [47] implementation of recommender systems that use star ratings, Dacrema et al.'s [8] implementation of recommender systems that use binary interaction data, an image classifier from the Stanford DAWNBench competition [6], and a text classifier from Google Jigsaw's Toxic Comment Classification challenge hosted on Kaggle [63].

Below, we provide additional detail about the four ML tasks and the specific datasets and models from prior work that we used. In each case, we followed prior work closely in our implementations so that the best-case performance achieved by Large Co. when it has access to a "full dataset" (i.e. 100% of the dataset used in prior work) is comparable to the published results. To make our simulations ecologically valid, our goal was for Large Co.'s best-case, full dataset performance to be comparable to reported results in prior work, while keeping computational costs down. To this end, we made some small changes aimed at reducing the cost of experiments while minimally impacting performance. We used software from prior work where possible and make the code for our experiments available.[1]

The first two tasks we studied were recommendation tasks that involve training "recommender systems" to predict the rating a user will give an item and predicting whether a user will interact with an item. Recommender systems are enormously important to a variety of industries, have garnered huge attention in the computing literature, and are immensely profitable [18,39,60]. The second two tasks we considered were classification tasks: classifying images (with ten possible image classes) and classifying text as toxic or non-toxic. These tasks come from two large ML research areas (computer vision and natural language processing) and are representative of classification systems that are applicable to a huge variety of industries (i.e. identifying the class of an image or piece of text is broadly useful).

For the rating prediction task, we used Rendle's [46] factorization machine approach with the extremely influential MovieLens 10-M dataset [18], as Rendle et al. [47] demonstrated that this approach outperformed a variety of competing approaches in terms of Root Mean Squared Error (RMSE), a metric used in past recommender system competitions. Specifically, we used Bayesian Matrix Factorization with size 32 embeddings and 50 sampling steps, which substantially lowers training costs and slightly lowers performance compared to the most expensive configuration Rendle et al. used (size 500 embeddings and 500 sampling steps).

Dacrema et al. rigorously compared simple, yet well-calibrated baseline techniques to complex neural techniques for the interaction prediction task. We focus on the Pinterest recommendation dataset from Dacrema et al.'s work, originally from Geng et al. [16]. Dacrema et al. showed that a simple item-based k-nearest neighbor performs extremely well—better than neural techniques—for this dataset. We use Dacrema et al.'s implementation of a k-nearest neighbor recommender system. In terms of evaluation, Dacrema et al. used Hit Rate@5 (alongside other metrics that led to

[1] https://github.com/nickmvincent/cdc

similar findings). Hit Rate is defined by holding out a single item that each user interacted with and then using the model to rank the held-out item and 99 random items the user did not interact with. If the item that the user actually interacted with is in the top five ranked results returned by the model, this is considered a success, and Hit Rate@5 is the total fraction of users for which the model was successful. This evaluation procedure maps well to top-n recommendation features common on many platforms (e.g. "Top Videos for You").

For a classification task from computer vision, we consider the CIFAR-10 dataset, a popular benchmark dataset for image classification [31]. The Stanford DAWNBench challenge [6] includes a leaderboard that documents image classification approaches that achieve high accuracy using minimal training time. From this leaderboard, we used Page's [42] ResNet approach and the well-studied CIFAR-10 dataset [31] (which includes images belonging to ten different classes). For this task, we evaluated our models using accuracy: for this task, accuracy is defined as fraction of test images that are successfully classified.

From natural language processing, we consider the case of toxic comment classification, using labeled Wikipedia discussion comments from Google Jigsaw's ML challenge hosted on Kaggle [63]. We used a TF-IDF logistic regression approach from the Kaggle leaderboard that achieves performance comparable to top models. While the dataset has labels for six different overlapping categories of toxicity, we focused only on binary classification (toxic vs. non-toxic) such that we train only a single model for each simulation. We made this task binary by treating a comment with any toxicity-related label (toxic, severely toxic, obscene, threatening, insulting, hateful) as generally toxic. As in the Kaggle competition, we evaluate the model using area under the receiver operator curve (which we refer to as AUC, for area under the curve). The binary classification performance of the ML approach is very close to the average performance across the six categories (see code repo for more details[2]).

Our datasets also cover a range of sizes and data types: the ML-10M dataset has 10M explicit ratings (1-5 stars) from 72k users. The Pinterest dataset has 1.5M interactions from 55k users. CIFAR-10 has 50k train images and a fixed set of 10k test images. The Toxic Comments dataset has 160k comments. This presents a major challenge in comparing the results of our data leverage simulations: each task is typically evaluated in a different manner (i.e. RMSE vs. Hit Rate vs. Accuracy vs. AUC), our datasets are of different sizes, and the format of data varies substantially (i.e. a movie rating is different than an image or piece of text). For instance, we might perform several experiments that show that a contribution of x star ratings can improve recommender RMSE by y, whereas a contribution of x images can improve image classification accuracy by z. However, comparing changes in RMSE to changes in image classification accuracy is not straightforward. Below, we describe how we defined Data Leverage Power in order to address this challenge.

3.3 Measuring the Effectiveness of CDC with Data Leverage Power

Different ML tasks require different evaluation techniques and our datasets have different sizes and different data formats. Here, we describe how we compared the effectiveness of CDC across different tasks.

To measure effectiveness, we introduce a task-agnostic measurement: **Data Leverage Power** (DLP). The goal of defining DLP is to create a single metric that captures the effectiveness of data leverage across ML tasks and across different data leverage scenarios (e.g., strike vs. CDC). DLP is defined in a scenario-specific manner such that it tracks Small Co.'s ability to catch up with Large Co. in ML performance in CDC scenarios, but also tracks the ability for a data strike to lower

Large Co.'s performance in the *data strike only* scenario (in which Small Co.'s performance does not change).

For each participation rate, DLP takes into account four measurements: baseline performance (performance with a very low data approach, for instance a "random guess" approach for classification or "recommend most popular items" approach for recommendation), the full-data best-case performance, Small Co.'s average performance, and Large Co.'s average performance. Below, we refer to these four measurements, respectively, as *baseline*, *best*, *small*, and *large*. While measuring full-data best-case performance is straightforward, selecting a baseline is less so. After walking through exactly how we defined DLP below, we describe how we identified baseline performance for each task.

For our two scenarios that involve CDC (*CDC only* and *CDC with data strike*), DLP is defined as the ratio of Small Co.'s average performance improvement over baseline to Large Co.'s average performance improvement over baseline for a given participation rate. In other words, we compare how much better Small Co.'s performance improves on the baseline to how much Large Co.'s performance improves on the baseline. Mathematically, for our scenarios that involve CDC, DLP defined as is:

$$\frac{small - baseline}{large - baseline}$$

In *CDC only* scenarios, Large Co.'s performance never changes (no data strike occurs) and is therefore fixed at best-case performance, while Small Co.'s performance increases with participation rate. In the *CDC with data strike* scenarios, larger participation rates lower Large Co.'s performance while increasing Small Co.'s performance.

As an example, imagine that for some model, best-case performance is 1.0 accuracy and worst-case is 0.5. With full data, Large Co. achieves the best-case 1.0 accuracy and thus has an improvement over worst-case of 0.5. For CDC by 10% of users, Small Co.'s accuracy (averaged across iterations) is 0.7, an improvement over worst-case of 0.2. If this is accompanied by a data strike and this data strike causes Large Co.'s performance to drop to 0.9, Large Co. now has an improvement over worst-case of 0.4. In this case, the DLP for *CDC only* is 0.2 / 0.5 = 0.4 and the DLP for *CDC with data strike* is 0.2 / 0.4 = 0.5. By repeating the entire process for every group size, we obtain a full plot of DLP vs. participation rate. For each task, we set "baseline" performance as corresponding to the worst performance from all our experiments, which occurs when either company has as little data as possible (in our experiments, 1% of users/data). For ML-10M, this occurs when Small Co. has 1% of users and is evaluated on the *fixed holdout* test set. In this case, the model effectively guesses the mean rating for almost all predictions. For CIFAR-10, the worst-case performance also occurs when Small Co. has 1% of the data, and is about 10%, equivalent to randomly guessing one of ten classes. For Toxic Comments, the lowest performance occurs when Small Co. has a 1% sample of data (and is about 0.9 AUC). The Pinterest task is a special case, as our approach cannot make predictions for unseen users. To get a baseline for Pinterest, we followed Dacrema et al. and used the performance achieved when using a simple "recommend most popular items" approach with full data. As we will discuss below, we also "recommend most popular" to calculate *fixed holdout* performance (because in a *fixed holdout* scenario, the recommender will face unseen users).

In *data strike only* scenarios, however, Small Co.'s performance is fixed at worst-case performance and therefore Small Co.'s improvement over baseline is zero. This means the numerator of the ratio we used above is always zero in *data strike only* scenarios. Therefore, for

these scenarios, we calculate how much Large Co.'s performance has fallen from best-case performance and find the ratio of Large Co.'s performance loss to gap between best-case and baseline. This "no-CDC" version of DLP is still comparable to CDC version, as it measures the delta between Small Co. and Large Co.'s performance. Mathematically, DLP for *data strike only* scenarios is:

$$\frac{large - best}{baseline - best}$$

The DLP approach to comparing data leverage simulations accounts for the fact that datasets are of different sizes and are comprised of different, hard-to-compare data types (e.g. a single image is different from a single user-item interaction). By focusing on DLP and participation rate (instead of e.g., number of users, number of observations, number of gigabytes of data, etc.), we can make comparisons across ML tasks, e.g., how does CDC by 30% of ML-10M users compare to CDC by 30% of CIFAR-10 users?

In interpreting our results, we calculate the participation rates needed to achieve a certain DLP for each ML task. For instance, we ask "How many users does it take to get 80% of the way to Large Co.'s performance?" We consider a variety of reasonable round-number DLP thresholds, because acceptable performance levels will vary by user and ML task (identifying acceptable performance levels is an important area of future work). For instance, a user who is motivated to support Small Co. (e.g. because they feel strongly about protesting Large Co.'s values or actions and feel strongly about supporting Small Co.'s values or actions) might accept much worse performance from Small Co.'s technologies than a user who does feel as strongly about the companies' value or actions. For instance, even for a DLP of 0.5 – e.g. Small Co.'s recommender system gets just 50% of the way to best-case performance – a user who strongly supports Small Co. might not mind needing to scroll further through their recommendation lists.

It is important to note that while DLP was critical to our ability to compare different data leverage scenarios, is important to consider task-specific factors such as performance thresholds and the real-world value of improved performance, i.e., if performance changes from x to y, what are downstream effects on revenue, user retention, etc.? To increase the interpretability of our piecewise definition of DLP and address the second challenge, we also report the raw performance values (traditional metrics) that accompany a given DLP value. In doing so, we retain the benefits of DLP (easy comparisons across tasks) while still allowing those familiar with a particular task to understand task-specific effects of a data leverage campaign (i.e., if a DLP campaign moves performance over or under an important task-specific performance threshold). As we will highlight again throughout our Discussion, a similar approach that treats DLP and traditional metrics as complementary will also be useful for studying data leverage in practice.

Additionally, comparing a data strike to CDC is comparing an action that harms ML performance to an action that helps ML performance. In order to address this challenge, we defined DLP as piecewise, with a separate definition for CDC scenarios and *data strike only* scenarios, such that harming Large Co. and helping Small Co. both represent increased balance in power between Large Co. and Small Co. In other words, our piecewise definition is motivated by the assumption that CDC and data strikes represent two ways of achieving the same goal.

3.4 Company Perspective vs. Fixed Holdout Evaluation

As mentioned above, when simulating CDC, we evaluate our models using scenario-specific *company perspective* test sets (i.e. test sets drawn from the data that Small Co. or Large Co. have).

However, as a secondary measure, we can also look at evaluation metrics from a *fixed holdout set* that neither company can access, which summarizes ML performance while taking into account the experience of new users and anonymous users. In other words, we consider both a "subjective" test set and an "objective" test set (in the sense that the "objective" *fixed holdout* set is unaffected by the specifics of each data leverage scenario, while the "subjective" *company perspective* test set is affected). For ML-10M and Toxic Comments, we used the same approach used in Rendle et al.'s review and sampled a random 10% of data to create a *fixed holdout* set. For the Pinterest and CIFAR-10 datasets, we used the fixed holdout sets used in the prior work [8,31] that inspired our modeling approach.

The distinction between *fixed holdout* and *company perspective* test sets is most important for personalization tasks (e.g., recommendation). For these contexts, if a company has no data about a particular user (e.g., because that person is a brand-new user, is accessing a service anonymously, or is engaging in obfuscation), that user necessarily receives non-personalized worst-case performance. For the ML-10M case, the approach we used can only, at best, predict that anonymous users will give every item the mean rating of all items. For the Pinterest case, the approach we used cannot produce recommendations for unseen users, so we uniformly assigned these users a Hit Rate@5 contribution of 0.1668, corresponding to Hit Rate@5 documented by Dacrema et al. when using a non-personalized "recommend most popular items to everyone" approach.

The real-word scenario that *fixed holdout* evaluation maps to is the one in which users receive recommendations from both Small Co. and Large Co. but use one or both of the services as a new or anonymous user (or use obfuscation to make themselves effectively anonymous). For instance, if only 10% of users contribute data to Small Co., but every single user chooses to use Small Co.'s recommender system, it will necessarily perform poorly for the 90% of users for whom the model cannot provide personalized results.

Critically, if a company has strong *company perspective* performance, but poor *fixed holdout* performance, this means their technologies will be very effective for current users, but they may have trouble expanding their userbase. Our experiments involve random sampling, so each simulated company perspective test set is drawn from the same distribution as the fixed holdout set, with the main difference arising when a user appears in the *fixed holdout* set but not a particular company's test set. In practice, CDC and data strikes may be practiced by homogenous groups, and so the distinction between company perspective and fixed holdout may become even more important. In presenting our results, we focus first on *company perspective* evaluation, and then discuss the implications of looking at results using a *fixed holdout se*t. Comparing different "test set perspectives" will be an important component of future data leverage research.

4 RESULTS

Below, we present the results from our DLP simulations. We begin by focusing on our *CDC only* results. Next, we examine the additional effect of adding a data strike to CDC and examine our results for *CDC with data strike* and *data strike only* scenarios.

As mentioned above, our primary focus is on ML performance measured with *company perspective* evaluation. At the end of this section, we present our secondary measurement, *fixed holdout* evaluation, and describe how this secondary measurement informs us about interactions between data leverage and personalization systems.

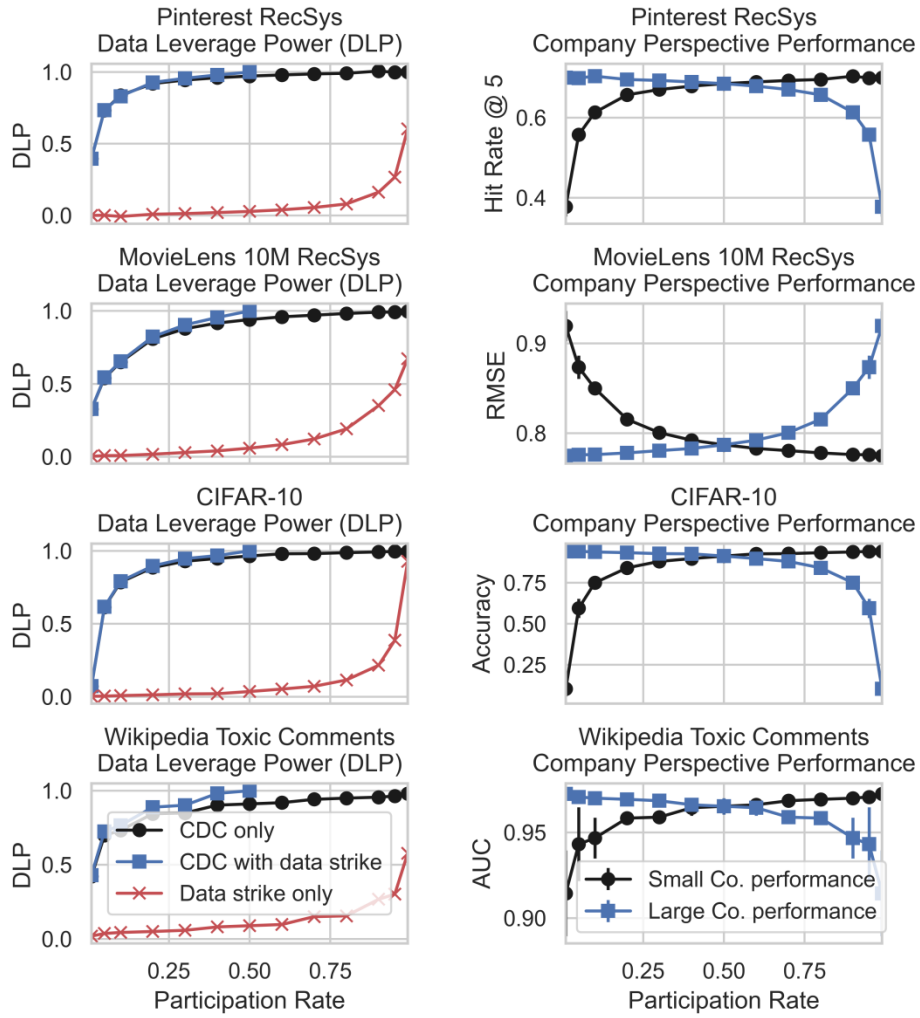

Fig. 1. The left column shows DLP plotted against participation rate for CDC only, CDC with data strike, and data strike only scenarios. Each row shows a different ML task. The right column shows the task-specific performance measurement we used to calculate DLP. Vertical bars show standard deviation for task-specific results.

4.1 *CDC Only* Scenarios using *Company Perspective* Evaluation

Our full set of *company-perspective* experimental results are shown in Figure 1. The left column of Figure 1 shows Data Leverage Power (DLP), our measurement described above that allows us to compare results across ML tasks. A higher DLP value means that a data leverage action was more effective, in terms of boosting Small Co. and/or reducing Large Co.'s performance. Within the left column, *CDC only* results are shown in black. To show how DLP relates to task-specific performance, the right column shows the various task-specific evaluation measurements that we

used to calculate DLP for each scenario: Hit Rate, Root Mean Squared Error (RMSE), Accuracy, and Area under the Receiver-Operator Curve (AUC). Here, there are only two colors: black shows Small Co.'s performance improving as CDC participation increases while blue shows Large Co.'s performance decreasing as data strike participation increases (we discuss these data strike results below). As described above, using baseline performance, best-case performance, and these two performance curves, we can compute DLP values (left column) for all three of our scenarios.

Examining the black curves in the left-hand column of Figure 1, it appears that CDC can be highly effective at allowing a small company to drastically reduce the performance gap between itself and a large competitor, as across our four tasks we see a CDC participation rate of at minimum 10% and at most 20% is needed to get Small Co.'s performance 80% of the way to best-case (i.e., the black curve reaches a DLP of 0.8). In general, both the scenario-specific evaluation curves (right column) and resulting DLP curves (left column) display diminishing returns of data. However, comparing the rows in Figure 1, we see that the effectiveness of CDC at reducing the performance gap between companies is not identical across tasks. The curves "level off" at different rates. Like learning curves, DLP curves are influenced by a variety of factors such as the algorithm used and the size of the full dataset. DLP provides us with a consistent way to make comparisons.

To systematically compare the black DLP curves from each row shown in the left column of Figure 1 is, we asked, "what group size of CDC is needed to achieve a given DLP"? Drawing from the results shown in Figure 1, Table 1 shows the CDC group size needed to achieve DLP thresholds of 0.5, 0.6, 0.7, 0.8, and 0.9.

Table 1. Shows the CDC participation rate needed to reach DLP thresholds of 0.5, 0.6, 0.7, 0.8, and 0.9 for each ML task we studied. For instance, the bottom left cell shows that 5% CDC participation is needed to reach a DLP of 0.5 for the Toxic Comments task.

		DLP Thresholds				
		0.5	**0.6**	**0.7**	**0.8**	**0.9**
ML Task	**Pinterest**	5%	5%	5%	10%	20%
	ML-10M	5%	10%	20%	20%	40%
	CIFAR-10	5%	5%	10%	20%	30%
	Toxic Comments	5%	5%	10%	20%	40%

Looking at Table 1, we see that CDC was especially effective for the Pinterest recommendation task, e.g., for the Pinterest task, only 20% of users need to engage in CDC to achieve 90% of best-case performance. For the other tasks, larger CDC participation (i.e. 30-40%) was required to achieve this same 90% DLP threshold.

Even for the cases that require greater participation to achieve a given DLP threshold, the results in Table 1 suggest CDC could be quite impactful. For instance, in the ML-10M case, 20% of users can achieve a DLP of 0.8, which may be good enough for some potential users to begin using Small Co.'s ML model. Furthermore, the Small Co. performance results are not upper bounds on performance for a given participation rate; it is possible that Small Co.'s data scientists could get even better performance by using a model that performs better for small datasets, by adjusting hyperparameters, by using data augmentation, etc.

Looking at the traditional performance metrics in the right column, we see underlying values from which DLP was calculated. To obtain DLP of 0.8 with in CDC only scenarios, Small Co. must achieve substantial improvements over the baseline. For instance, at participation rate of 20%, Small Co. achieves a Pinterest Hit Rate of 0.66, MovieLens RMSE of 0.82, CIFAR-10 accuracy of 0.84, and a Toxic Comments AUROC of 0.96. These are raw performance metrics that may very well be acceptable for Small Co.'s use. For instance, in their paper, Rendle et al. documented progress on ML-10M: an RMSE of 0.82 is equivalent to the state of the art in 2008, which may be adequate for CDC participants. Comparing the left and right columns of Figure 1, we see that converting DLP to raw performance is relatively straightforward for *CDC only*, because Large Co.'s performance never changes (as there is no data strike) and the DLP curve therefore has the same shape as the raw performance curve.

4.2 *CDC With Data Strike* and *Data Strike Only* using Company Perspective Evaluation

Next, we look at how DLP changes when users engage in CDC (to Small Co.) and simultaneously engage in a data strike (by deleting data from Large Co.). Returning to the left column of Figure 1, we focus on the blue squares (*CDC with data strike*) as well as the difference between the blue squares and black circles (*CDC only*).

Our primary finding here is that a data strike can add a small effect on top of CDC, but only when participation rates rise above around 20%. Even for a participation rate of 30%, adding a data strike adds at most 0.05 DLP. This means that for a group of 30% of users, the ability to delete data lowers Large Co.'s performance and thus "closes the performance gap" by at most an additional 5%. Of course, the logistical and potential legal challenge of adding the data strike on top of CDC may incur large costs relative to the gain.

Returning to the raw performance metrics in the right column, we can see how the additional "boost" from incorporating data strikes corresponds to reduced performance for Large Co. Looking at Large Co.'s performance curve in blue, we unsurprisingly see the same characteristic diminishing return curves: small data strikes only have minor impact on Large Co., explaining the relatively small DLP boost from adding data strikes to CDC.

Even at a participation rate of 0.5 (i.e. a group of 50% of all users included in the benchmark dataset), the benefits of adding a data strike are still modest, peaking at 9% for the Toxic Comments case. At the same 50% group size, deletion had only a 3% benefit for the Pinterest case. We cannot measure the benefits for groups larger than 0.5, because at this point Large Co. now has less data than Small Co. and we effectively begin traversing the DLP curve backwards (i.e., Small Co. has now become the "larger" company in terms of dataset size).

Finally, looking at the *data strike only DLP* curves (red x's in the left column), we see that data strikes alone have very little DLP relative to CDC approaches. This illustrates the challenge of diminishing returns of data. If a learning curve is flat (or almost flat) for a wide range of participation rates, a data strike alone must break out of this flat region to begin exerting substantial data leverage. As we will see immediately below, an interesting exception to this trend is when we consider anonymous users: users who engage in a data strike against Large Co. but continue to use Large Co.'s technology will receive non-personalized results, which can hurt Large Co.'s overall performance substantially.

To summarize, we see that data strikes have much less potential to exert data leverage than CDC at participation rates less than 50%. Of course, CDC requires the existence of a competitor, so in many cases people may only be able to data strike. Specifically, in monopoly contexts, data

strikes may be the only option available to users. We expand on how supporting CDC might support greater competition, as both a tool against monopoly but also in the service of innovation.

4.3 CDC Effectiveness using *Fixed Holdout* Evaluation

As discussed above, it is informative to examine the effects of CDC when ML models are evaluated using a fixed holdout test set that includes users who may not appear in the *company perspective* test set. Figure 2 shows the same measurements as Figure 1 but uses *fixed holdout* performance instead of *company perspective* performance.

We observed that for non-personalized CIFAR-10 and Toxic Comment cases, *fixed holdout* test set evaluation gave very similar results to *company perspective* evaluation. This is expected: the *company perspective* test sets were randomly drawn from the same source as the *fixed holdout* test sets in our experiments. The only notable difference was higher standard deviation (vertical bars visible in the last row of Figure 1) in performance for small amounts of data, because the *company perspective* test sets are smaller for smaller group sizes.

For ML-10M and Pinterest recommendation tasks (i.e. tasks that involve personalized results), *fixed holdout* test set performance is different from *company perspective* performance. Rather than a diminishing returns curve, ML performance (and the resulting DLP) is linearly dependent on participation rate. Specifically, as each company loses users, their recommendation performance linearly approaches non-personalized baseline performance.

As described above, this result corresponds to a situation in which some users access a recommender system with a new account (i.e. because they are new users, or perhaps because they deleted their account as part of a data leverage campaign), forcing the system to output non-personalized results. Specifically, imagine that 30% of users engage in CDC to support Small Co. Some other group of users access Small Co.'s recommender system with new accounts. These users who did not participate in CDC but continue to use Small Co.'s model anonymously will not see the benefits of other people's data contributions until they provide their own data. This result highlights that personalized ML systems require users to provide their own data before they can see the benefits of CDC.

The evaluation of recommender systems using a *fixed holdout set* shown in Figure 2 also highlights a case in which deletion can play a large role in reducing the performance gap between two personalization technologies. Put plainly, if a user insists on getting non-personalized results from one company, deleting their data from a competing company is effective at reducing the performance gap between the two: both companies will be forced to provide the user non-personalized results.

Overall, these results suggest that while the ability to delete data can enhance the ability of CDC by groups to increase competition between ML technologies, it will only make a large difference for relatively large campaigns or for cases in which people use personalized ML technologies with new accounts. Deletion raises ethical concerns as well: there are cases in which hurting some "Large Co." may be seen as anti-social, e.g. for classification models that are well known to assist physicians in achieving better health outcomes, but there are also cases where hurting Large Co. companies may be seen as pro-social, e.g. lowering the utility of disadvantage-reinforcing credit scoring systems. We expand on these concerns below.

5 DISCUSSION

In this section, we discuss the implications of our experimental results and the limitations of our study that might be addressed by future work.

As we highlighted in the Introduction, the study of power dynamics between users and tech companies using the combined lens of collective action and machine learning is relatively new. We were able to contribute to this emerging research area by leveraging simulation methods, with assumptions grounded in ongoing discussions about data leverage and ML. We have seen that CDC represents a promising and feasible means by which the public might gain additional leverage.

The potential impact of CDC could be amplified through a number of avenues that span design, policy, and research. In order to make CDC more broadly available, there are major opportunities to design new technologies to make CDC easier, as well as opportunities to institute policy that supports CDC. We discuss specific examples below.

5.1 Implications for Design and Policy

We observed that CDC can be effective in reducing the performance gap between two competitors. This finding suggests that constituencies interested in creating more competition between AI services—as an intervention aimed to prevent monopolies or as a means to increase innovation—may wish to further investigate CDC itself (e.g. conducting similar experiments to those described here) and explore avenues for making CDC easier at the grassroots level. Policymakers and advocates might push for data portability regulations, an area of growing discussion [10,50]. Specifically, policies that make it easier for people to obtain data they generate and transfer that data to another company could make CDC easier. For instance, the EU's GDPR has a "right to data portability" – other jurisdictions might emulate or extend this idea [64].

Regulatory support will be particularly important for making CDC possible for datasets with complex formats. For instance, for datasets that are captured with proprietary sensors (e.g., wearable tech personal health datasets), CDC will likely remain impossible without regulation that compels companies to make data transferrable in common (and machine-readable) format. The creation of such regulation will benefit from interdisciplinary research incorporating machine learning, law, and HCI/CSCW.

In the near-term, the technologies and companies to which CDC is applicable may be limited by the legal rights around data. In other words, while tools (e.g., scraping software from researchers) and voluntary choices by companies (e.g., Google's Takeout service) may make CDC possible today, there are certain contexts in which CDC needs legal support. In general, laws that give users more agency over the data they helped to generate will amplify the power of data leverage, whereas laws that make it harder to have agency over data will do the opposite. Data leverage will especially benefit from laws that are designed with a focus on the social and collective nature of data creation, as opposed to a framework of individual data ownership.

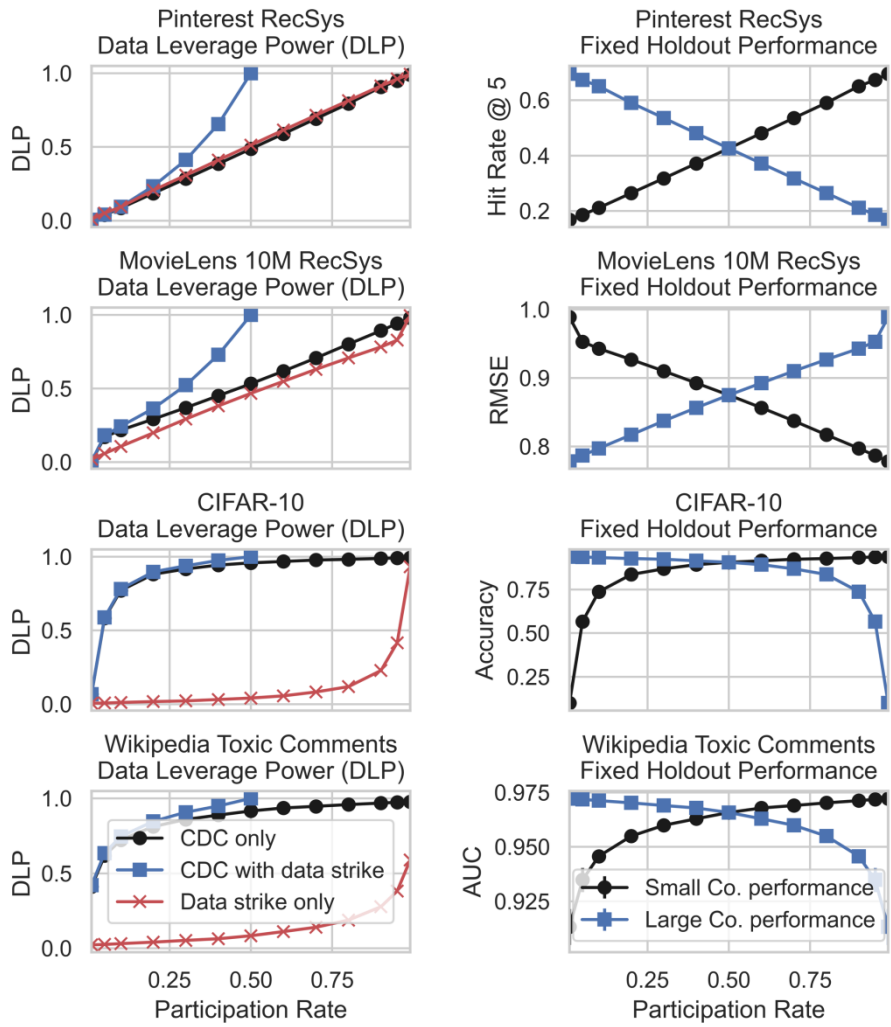

Fig. 2. DLP results using *Fixed Holdout* Evaluation. Like Fig. 1, the left column shows DLP plotted against participation, each row shows a different ML task, and the right column shows the task-specific performance measurement we used to calculate DLP.

Software can complement regulation or support CDC in the absence of regulation. Designers might create tools that make CDC easier, such as tools that help users collect their data from web platforms in a format that is easy to share with other companies or organizations. These tools might look similar to software used to collect data for studies in social computing and CSCW (e.g. software like data scrapers and scripts that use APIs to obtain data from sources like Twitter, Wikipedia, Reddit, etc. [41]). There are also opportunities for technology designers to create tools that help organize and make visible the impact of CDC. These tools could take inspiration from technologies that support collective action (e.g. Zhang et al.'s "WeDo" [59] and Li et al.'s "Out of

Site" [35]), and would aid in scaffolding the CDC process and communicating the impact of CDC to participants. Indeed, this direction would benefit especially from CSCW research around coordinating and motivating online groups (including work that has focused on peer production, e.g. [26,30,61,62]). An additional synergy between the CDC concept and CSCW research is that CDC is, by its nature, online. Any techniques, tools, and strategies developed by CSCW researchers for online collective action (activism, peer production, etc.) might additionally be used to support groups who wish to engage in CDC. In the same vein, new findings from CSCW that support collective action can likely be applied to CDC in a straightforward manner.

5.2 Ecological validity of simulations

The conceptualization of CDC we simulated is something that users can realistically engage in today, although data strikes may not be feasible in some contexts (e.g. if regulators do not enforce user requests for data deletion). Thus, our results that correspond to scenarios in which CDC participants also delete their data can be seen as measuring how much more effective CDC might be *in a world in which large-scale data deletion is possible* (and companies are forced to retrain their models regularly, protecting against "data laundering" through model parameters). In a very recent case, the Federal Trade Commission forced a privacy-violating company to delete both facial recognition data and the model trained using that data; this is a promising precedent for future data strikes [38].

An additional ecological validity concern is that expensive-to-train models, e.g. models that cost up to $245k to train [43], may be completely inaccessible to non-incumbent companies. Our experiments required repeated retraining of models, forcing us to select models that are fast to train. However, small challenger companies may also face financial constraints around model training, so by focusing first on fast-to-train models, we naturally select for approaches that small companies might realistically use. Furthermore, the datasets we investigated may be of similar size to some, but not all, models in production at tech companies. Future work that seeks to make expensive state-of-the-art model training cheaper could further widen the pool of models that are "CDC viable".

It is worth noting that some truly enormous models, such as OpenAI's recent GPT-3 language model, may be simply too expensive to ever study with the simulation approach we used here [3]. For these models, it may be possible to investigate the efficacy of CDC and data strikes through alternative approaches, such as "influence" estimation techniques from the machine learning literature [28,58].

Finally, an important ecological validity concern is that there may be cases in which a "Small Co." and "Large Co." in the real world compete without offering comparable ML services, because the absence or presence of an ML service is a selling point (e.g. a privacy-focused Small Co. that eschews recommendation entirely). For such cases, the methodological approach of CDC simulation will not be as helpful as an analysis that focuses on other factors that govern business success.

5.3 Identifying "Acceptable" Performance of ML Models

One approach we used to compare ML tasks was to look at the CDC participation rate needed to reach reasonable round-number performance thresholds. In practice, there may be thresholds specific to certain contexts that require looking at traditional ML evaluation metrics, e.g., perhaps for certain users of a music recommender system, there is a certain Hit Rate at which they will

stop using the recommender. By considering traditional metrics, it is possible to take such thresholds into account.

Public research that identifies these thresholds, and more generally the relationship between ML performance and downstream consequences (e.g., how performance impacts users leaving a platform, how performance maps to profit), will make it easier to plan CDC (following, for instance, Song et al.'s work on degraded search performance [54]). If CDC organizers know that DLP of 0.8 (or, returning to our results from ML-10M, an RMSE of 0.82) is "good enough" for most people, they can use simulations like ours to estimate what participation rate they need to achieve this DLP.

Studying the relationship between performance and downstream consequences will likely require proprietary data (especially in the case of performance's relationship with revenue). Furthermore, these relationships likely differ across ML tasks. Nonetheless, any findings in this area could be invaluable for informing CDC.

Work on recommender systems has already shown that offline evaluation metrics do not fully predict user perception of recommendations [39]: while offline evaluation metrics like Hit Rate are defined to correlate directly with revenue generation (i.e. each hit corresponds to revenue), they do not necessarily correlate very well with user satisfaction. For the purposes of getting new users to join a competing platform, this means getting only part of the way towards "best-case" performance may be more than adequate to give users a satisfying experience.

5.4 Negative Impacts of Data Leverage

For ML technologies that are clearly identifiable as societally beneficial, exerting data leverage via data strikes could be harmful. Conversely, for technologies that are societally harmful, exerting data leverage via CDC could be harmful. Of course, the classification of beneficial vs. harmful technologies requires ongoing discussion, but research has already begun to identify some cases. Notably, Kulynych, Overdorf and colleagues' work on Protective Optimization Technologies provides an overview of various harmful ML instantiations, including discriminatory facial recognition and credit scoring that create unjust economic feedback loops [32].

Our results suggest that CDC should be preferred over data strikes for exerting data leverage against ML operators of societally beneficial technologies. Some ML models directly impact health and safety outcomes, e.g., ML models that assist doctors or operate vehicles. For such cases, using data deletion to reduce the performance of a large incumbent organization (e.g., a major hospital or transit company) could induce substantial societal harms (e.g., missed diagnoses or vehicle crashes).

For cases in which a ML model is considered societally harmful, e.g. discriminatory facial recognition or unjust credit scoring [32], CDC does not represent a viable option. However, data strikes still face an uphill battle to break out of the flat region of diminishing returns curves. Our results suggest protest against such harmful ML models may be best accomplished through working towards regulation, perhaps in conjunction with data strikes or other types of POTs. Some jurisdictions, like San Francisco, have already moved towards such regulation, by banning the use of facial recognition by police [7]. Other jurisdictions might follow this example and directly regulate societally harmful technologies for which consumer leverage and data leverage are not well suited to address.

An additional consideration is that a small CDC beneficiary operating a less-than-ideal version of a particular technology could also induce harms. To address this possibility, a set of acceptable performance thresholds could be determined by an external body knowledgeable about task-

specific metrics (e.g., a government mandated minimum precision and recall for a medical model). CDC users seeking to a support a new health start-up would likely need to contribute enough data for the start-up to meet external standards before their models could be put into use.

A final consideration is that evidence from economics literature suggests that broader data sharing may be more desirable from a consumer welfare perspective, as more companies would be able to provide high quality technologies [25]. In the extreme, if a huge number of firms are subject to data strikes, AI technologies would broadly suffer. In the opposing extreme, if a huge number of firms had access to data from every person in the world, AI technologies would be very accurate, but privacy would be grossly violated at a global scale.

5.5 Data Sharing by Corporations

We have so far framed our research by considering scenarios in which users collectively engage in CDC as a way of exerting data leverage against companies. However, firms can also share data that is broadly useful to other organizations. For instance, a company interested in social responsibility might release a labeled dataset that is useful for societally beneficial ML technologies as part of a conscious corporate social responsibility program. We'd expect such "corporate CDC" to have similar effectiveness to the CDC we studied here: if a large company releases 10% of its "toxic comment detection" data, this may be enough data for other organizations to get 70% of the way towards best case performance on this task. Indeed, we see some steps in this direction through open data initiatives [65]. Looking forward, government programs could even financially incentivize such open data initiatives as part of an effort to address concerns about the market power of tech companies.

5.6 Limitations and Future Work

In general, while our simulations covered a variety of tasks and we took steps to maximize the ecological validity of our simulations, there are many opportunities to extend our data leverage simulations. We used only one high-performing model and hyperparameter setup for each simulation. This means our results do not represent an upper bound for the effectiveness of CDC, as Small Co.'s data scientists would likely seek alternate models or hyperparameters that perform better for small datasets. Furthermore, there are numerous ML tasks that could be studied using simulation, perhaps with the ultimate goal of creating a "catalog" of CDC effectiveness. Such a catalog would be useful to CDC organizers, but also to policymakers interested in incentivizing corporate CDC and promoting competition around ML.

It is worth nothing that while our DLP definition is critical for allowing us to compare ML tasks with different evaluation procedures and different data sizes, a weakness of the DLP definition is that it requires careful selection of a baseline and careful consideration of what constitutes a "full dataset". Selecting an extremely weak baseline could make DLP appear exaggerated. Selecting too small or too large of a "full dataset" size could miss important parts of the DLP curve. We addressed these challenges by carefully choosing comparable baselines for each task (such that the baseline corresponds to a "low-data" or "no-data" approach that Small Co. might use when it has access to very little data) and by taking our datasets from prior work.

One way future simulations can zoom in on specific tasks would be to perform simulations that take into account the costs and rewards of successes and errors. Researchers might estimate the cost of false positives and false negatives, estimate the reward associated with successful classification, and calculate the expected total cost or reward for each organization associated with

a given data leverage action (e.g. the cost to Large Co. and the benefit to Small Co.). This would move towards simulating the downstream consequences of CDC and data strikes.

Another direction for future data leverage research involves more directly modeling factors other than data leverage participation rate that facilitate success for businesses and technologies. In this paper, we did not address the intricacies of markets, consumer preferences, the ecosystems in which tech companies operate, or the collective action processes required to organize data leverage campaigns. Each of these factors will be important for future work that advances understanding of data leverage.

Beyond simulation, other directions for advancing this research area might involve in-the-wild experiments and observational studies of users exerting data leverage. This research might be conducted, at least in part, by organizations directly affected by data leverage, e.g. companies who are the targets of protest or the beneficiaries of CDC. These organizations will likely have access to unique data on the effectiveness of data leverage. In particular, as mentioned above, data that maps the effects of user-generated data to downstream consequences like business outcomes will be particularly valuable.

Finally, our experiments looked only at one ML task at a time. Future work should consider the interplay between datasets and ML tasks. For instance, how does CDC interact with ML pipelines and datasets that feed into multiple ML systems? Answering this question will be important for understanding the full effects of CDC and data leverage.

6 CONCLUSION

In this work, we proposed and evaluated **conscious data contribution**, a tactic the public might use to exert **data leverage** against tech companies to encourage them to change their behavior around key issues of interest. CDC entails users making data contributions aimed at reducing the performance gap between a large incumbent ML operator that users wish to protest and a small competitor that users wish to support. Using simulations, we measured the effectiveness of CDC in a variety of ML contexts using both a new metric called "data leverage power" and traditional ML metrics. Our results suggest that CDC represents a viable way to reduce the ML performance gap between a large incumbent and small competitor. We also observe that data deletion can enhance the effects of CDC, but the overall impact of data deletion is small compared to CDC. Overall, these results provide early information that inform the growing data leverage and provide guidance for constituencies interested in CDC.

7 ACKNOWLEDGMENTS

This was work was funded by NSF grants 1815507 and 1707296. We are grateful for discussions with Hanlin Li, Nicole Tilly, and Stevie Chancellor in developing the ideas in this paper. We additionally received very helpful feedback from colleagues at Grouplens at the University of Minnesota, CollabLab at Northwestern, and the Community Data Science Collective.

REFERENCES

[1] Imanol Arrieta Ibarra, 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 (2018).

[2] Michael Bloodgood and Chris Callison-Burch. 2010. Bucking the trend: Large-scale cost-focused active learning for statistical machine translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 854–864.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom

Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. *Language Models are Few-Shot Learners.*

[4] Finn Brunton and Helen Nissenbaum. 2015. *Obfuscation: A user's guide for privacy and protest.* Mit Press.

[5] Junghwan Cho, Kyewook Lee, Ellie Shin, Garry Choy, and Synho Do. 2015. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *arXiv preprint arXiv:1511.06348* (2015).

[6] Cody Coleman, Deepak Narayanan, Daniel Kang, Tian Zhao, Jian Zhang, Luigi Nardi, Peter Bailis, Kunle Olukotun, Chris Ré, and Matei Zaharia. 2017. Dawnbench: An end-to-end deep learning benchmark and competition. In *NIPS*.

[7] Kate Conger, Richard Fausset, and Serge F. Kovaleski. 2019. San Francisco Bans Facial Recognition Technology. *N.Y. Times* (May 2019). Retrieved from https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html

[8] Maurizio Ferrari Dacrema, 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*, ACM, 101–109.

[9] Thomas G Dietterich and Eun Bae Kong. 1995. *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms.* Technical report, Department of Computer Science, Oregon State University.

[10] Cory Doctorow. 2019. Regulating Big Tech makes them stronger, so they need competition instead. *The Economist.* Retrieved from https://www.economist.com/open-future/2019/06/06/regulating-big-tech-makes-them-stronger-so-they-need-competition-instead

[11] Carroll Doherty and Jocelyn Kiley. 2020. *Americans have become much less positive about tech companies' impact on the U.S.* Retrieved from https://www.pewresearch.org/fact-tank/2019/07/29/americans-have-become-much-less-positive-about-tech-companies-impact-on-the-u-s

[12] Michael D Ekstrand, F Maxwell Harper, Martijn C Willemsen, and Joseph A Konstan. 2014. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender systems*, ACM, 161–168.

[13] Kyle Endres and Costas Panagopoulos. 2017. Boycotts, buycotts, and political consumerism in America. *Research & Politics* 4, 4 (2017), 2053168017738632.

[14] Rosa L Figueroa, Qing Zeng-Treitler, Sasikiran Kandula, and Long H Ngo. 2012. Predicting sample size required for classification performance. *BMC medical informatics and decision making* 12, 1 (2012), 8.

[15] Monroe Friedman. 1996. A positive approach to organized consumer action: The "buycott" as an alternative to the boycott. *Journal of Consumer Policy* 19, 4 (1996), 439–451.

[16] Xue Geng, Hanwang Zhang, Jingwen Bian, and Tat-Seng Chua. 2015. Learning image and user features for recommendation in social networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 4274–4282.

[17] Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In *Proceedings of the 7th international symposium on wikis and open collaboration*, ACM, 163–172.

[18] F Maxwell Harper and Joseph A Konstan. 2016. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2016), 19.

[19] Brent Hecht, 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* (2018).

[20] Astead W. Herndon. 2019. Elizabeth Warren Proposes Breaking Up Tech Giants Like Amazon and Facebook. *The New York Times.* Retrieved April 4, 2019 from https://www.nytimes.com/2019/03/08/us/politics/elizabeth-warren-amazon.html

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

[22] Joel Hestness, 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* (2017).

[23] Vivian Ho. 2019. Tech monopoly? Facebook, Google and Amazon face increased scrutiny. *the Guardian* (June 2019). Retrieved from https://www.theguardian.com/technology/2019/jun/03/tech-monopoly-congress-increases-antitrust-scrutiny-on-facebook-google-amazon

[24] Daniel C Howe and Helen Nissenbaum. 2017. Engineering Privacy and Protest: A Case Study of AdNauseam. In *IWPE@ SP*, 57–64.

[25] Charles I Jones and Christopher Tonetti. 2019. *Nonrivalry and the Economics of Data.* National Bureau of Economic Research.

[26] Aniket Kittur and Robert E Kraut. 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, ACM, 37–46.

[27] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, ACM, 1301–1318.

[28] Pang Wei W Koh, Kai-Siang Ang, Hubert Teo, and Percy S Liang. 2019. On the accuracy of influence functions for measuring group effects. In *Advances in Neural Information Processing Systems*, 5255–5265.

[29] Prasanth Kolachina, Nicola Cancedda, Marc Dymetman, and Sriram Venkatapathy. 2012. Prediction of learning curves in machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, Association for Computational Linguistics, 22–30.
[30] Robert E Kraut, Paul Resnick, Sara Kiesler, Moira Burke, Yan Chen, Niki Kittur, Joseph Konstan, Yuqing Ren, and John Riedl. 2012. *Building successful online communities: Evidence-based social design*. Mit Press.
[31] Alex Krizhevsky, Geoffrey Hinton, and others. 2009. Learning multiple layers of features from tiny images. (2009).
[32] Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. 2020. POTs: protective optimization technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 177–188.
[33] Jaron Lanier and E Glen Weyl. 2018. A Blueprint for a Better Digital Society. *Harvard Business Review* (2018).
[34] Colin Lecher. 2018. California just passed one of the toughest data privacy laws in the country. *The Verge*. Retrieved from https://www.theverge.com/2018/6/28/17509720/california-consumer-privacy-act-legislation-law-vote
[35] Hanlin Li, Bodhi Alarcon, Sara M. Espinosa, and Brent Hecht. 2018. Out of Site: Empowering a New Approach to Online Boycotts. *Proceedings of the 2018 Computer-Supported Cooperative Work and Social Computing (CSCW'2018 / PACM)* (2018).
[36] Hanlin Li, 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 (2019), 87.
[37] Yilun Lin, Bowen Yu, Andrew Hall, and Brent Hecht. 2017. Problematizing and addressing the article-as-concept assumption in wikipedia. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2052–2067.
[38] Kim Lyons. 2021. FTC settles with photo storage app that pivoted to facial recognition. *The Verge* (January 2021). Retrieved from https://www.theverge.com/2021/1/11/22225171/ftc-facial-recognition-ever-settled-paravision-privacy-photos
[39] Ian Mackenzie, Chris Meyer, and Steve Noble. 2013. How retailers can keep up with consumers. *Mckinsey.com*. Retrieved from https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers
[40] Benjamin J Newman and Brandon L Bartels. 2011. Politics at the checkout line: Explaining political consumerism in the United States. *Political Research Quarterly* 64, 4 (2011), 803–817.
[41] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2, (2019), 13.
[42] David Page. 2018. How to Train Your ResNet. Retrieved from https://myrtle.ai/how-to-train-your-resnet
[43] Tony Peng. 2019. The Staggering Cost of Training SOTA AI Models. *Synced*. Retrieved from https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models
[44] Claudia Perlich, Foster Provost, and Jeffrey S Simonoff. 2003. Tree induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research* 4, Jun (2003), 211–255.
[45] Eric A Posner and E Glen Weyl. 2018. *Radical Markets: Uprooting Capitalism and Democracy for a Just Society*. Princeton University Press.
[46] Steffen Rendle. 2012. Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3, 3 (2012), 57.
[47] Steffen Rendle, Li Zhang, and Yehuda Koren. 2019. On the Difficulty of Evaluating Baselines: A Study on Recommender Systems. *arXiv preprint arXiv:1905.01395* (2019).
[48] Kenneth Rogoff. 2019. Big tech has too much monopoly power – it's right to take it on. *the Guardian* (April 2019). Retrieved from https://www.theguardian.com/technology/2019/apr/02/big-tech-monopoly-power-elizabeth-warren-technology
[49] Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. 2019. A Constructive Prediction of the Generalization Error Across Scales. *arXiv preprint arXiv:1909.12673* (2019).
[50] Gus Rossi and Charlotte Slaiman. 2019. Interoperability = Privacy + Competition. *Public Knowledge* (October 2019). Retrieved from https://www.publicknowledge.org/blog/interoperability-privacy-competition
[51] Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. 2015. We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), ACM, New York, NY, USA, 1621–1630. DOI:https://doi.org/10.1145/2702123.2702508
[52] Saiph Savage, Andres Monroy-Hernandez, and Tobias Höllerer. 2016. Botivist: Calling volunteers to action using online bots. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 813–822.
[53] Aaron Shaw, Haoqi Zhang, Andrés Monroy-Hernández, Sean Munson, Benjamin Mako Hill, Elizabeth Gerber, Peter Kinnaird, and Patrick Minder. 2014. Computer supported collective action. *interactions* 21, 2 (2014), 74–77.
[54] Yang Song, Xiaolin Shi, and Xin Fu. 2013. Evaluating and Predicting User Engagement Change with Degraded Search Relevance. In *Proceedings of the 22Nd International Conference on World Wide Web* (WWW '13), ACM, New York, NY, USA, 1213–1224. DOI:https://doi.org/10.1145/2488388.2488494
[55] Nicholas Vincent, 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*.
[56] Nicholas Vincent, Isaac Johnson, Patrick Sheehan, and Brent Hecht. 2019. Measuring the Importance of User-Generated Content to Search Engines. In *Proceedings of AAAI ICWSM 2019*.

[57] Nicholas Vincent, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent Hecht. 2021. Data Leverage: A Framework for Empowering the Public in its Relationship with Technology Companies. In *ACM FAccT 2021 (formerly FAT*)*.
[58] Nicholas Vincent, Yichun Li, Renee Zha, and Brent Hecht. 2019. Mapping the Potential and Pitfalls of "Data Dividends" as a Means of Sharing the Profits of Artificial Intelligence. *arXiv preprint arXiv:1912.00757* (2019).
[59] Haoqi Zhang, Andrés Monroy-Hernández, Aaron Shaw, Sean A Munson, Elizabeth Gerber, Benjamin Mako Hill, Peter Kinnaird, Shelly D Farnham, and Patrick Minder. 2014. WeDo: End-To-End Computer Supported Collective Action. In *Eighth International AAAI Conference on Weblogs and Social Media*.
[60] Renjie Zhou, Samamon Khemmarat, and Lixin Gao. 2010. The impact of YouTube recommendation system on video views. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, ACM, 404–410.
[61] Haiyi Zhu, Robert E Kraut, Yi-Chia Wang, and Aniket Kittur. 2011. Identifying shared leadership in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 3431–3434.
[62] Haiyi Zhu, Robert Kraut, and Aniket Kittur. 2012. Effectiveness of shared leadership in online communities. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, ACM, 407–416.
[63] 2020. Toxic Comment Classification Challenge. Retrieved from https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
[64] 2020. Art. 20 GDPR – Right to data portability. *General Data Protection Regulation (GDPR)*. Retrieved from https://gdpr-info.eu/art-20-gdpr
[65] 2020. Open Data Initiative. Retrieved from https://www.microsoft.com/en-us/open-data-initiative

Received June 2020; revised October 2020; accepted December 2020.