Datum: Managing Data Purchasing and Data Placement in a Geo-Distributed Data Market

Xiaoqi Ren, *Student Member, IEEE*, Palma London, *Student Member, IEEE*, Juba Ziani, *Student Member, IEEE*, and Adam Wierman, *Member, IEEE*

***Abstract*— This paper studies two design tasks faced by a geo-distributed cloud data market: which data to purchase (data purchasing) and where to place/replicate the data for delivery (data placement). We show that the joint problem of data purchasing and data placement within a cloud data market can be viewed as a facility location problem and is thus NP-hard. However, we give a provably optimal algorithm for the case of a data market made up of a single data center and then generalize the structure from the single data center setting in order to develop a near-optimal, polynomial-time algorithm for a geo-distributed data market. The resulting design, Datum, decomposes the joint purchasing and placement problem into two subproblems, one for data purchasing and one for data placement, using a transformation of the underlying bandwidth costs. We show, via a case study, that Datum is near optimal (within 1.6%) in practical settings.**

***Index Terms*— Data market, geo-distributed analytics.**

I. Introduction

TEN years ago computing infrastructure was a *commodity* – the key bottleneck for new tech startups was the cost of acquiring and scaling computational power as they grew. Now, computing power and memory are *services* that can be cheaply subscribed to and scaled as needed via cloud providers like Amazon EC2, Microsoft Azure, etc.

We are beginning the same transition with respect to *data*. Data is broadly being gathered, bought, processed and sold in various marketplaces. However, it is still a commodity, often obtained through offline negotiations between providers and companies [1]. Thus, acquiring data is one of the key bottlenecks for new tech startups nowadays.

This is beginning to change with the emergence of *cloud data markets*, which offer a single, logically centralized point for selling, buying and processing data. Multiple data markets have recently emerged in the cloud, e.g., Qlik Datamarket [2], Factual [3], InfoChimps [4], Xignite [5], IUPHAR [6], etc. These marketplaces enable data providers to sell and upload data and clients to request data from multiple providers (often for a fee) through a unified query interface. Also, major cloud service providers such as Google [7], Microsoft [8] and Amazon [9] all host various public datasets covering geospacial, environmental, scientific and online services as an extra benefit for their cloud clients. Current data markets provide a variety of basic services: (i) *aggregation* of data from multiple sources, (ii) *cleaning* of data to ensure quality across sources, (iii) *ease of use*, through a unified API, and (iv) *low-latency delivery* through a geographically distributed content distribution network. As these market places mature they are increasingly adding other services as well. Besides providing raw data to clients, it is an inevitable trend for data markets to carry out value-added services built upon the data, such as analytics and machine learning APIs.

Given the recent emergence of data markets, there are widely differing designs in the marketplace today, especially with respect to pricing. For example, the Qlik Datamarket [2] sets prices with a subscription model that allows a maximum number of queries (API calls) per month and limits the size of records that can be returned for a single query. Other data markets, e.g., Google BigQuery [7] and Infochimps [4], allow payments per query or per data set. In nearly all cases, the data provider and the data market operator each then get a share of the fees paid by the clients, though how this share is arrived at can differ dramatically across data markets. The task of pricing is made even more challenging when one considers that clients may be interested in data with differing levels of precision/quality and privacy may be a concern.

Not surprisingly, the design of pricing (both on the client side and the data provider side) has received significant attention in recent years, including pricing of per-query access [10], [11] and pricing of private data [12], [13].

In contrast, the focus of this paper is *not* on the design of pricing strategies for data markets. Instead, **we focus on the engineering side of the design of a data market**, which has been ignored to this point. Supposing that prices are given, there are important challenges that remain for the operation of a data market. Specifically, two crucial challenges relate to *data purchasing* and *data placement*.[1]

Manuscript received April 13, 2017; revised October 14, 2017 and February 2, 2018; accepted February 3, 2018; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor M. Mellia. Date of publication March 19, 2018; date of current version April 16, 2018. This work was supported in part by the National Science Foundation under Grant 1254169, Grant 1518941, Grant 1331343, and Grant 1637598, in part by the National Science Foundation Graduate Fellowship, and in part by the Resnick Sustainability Institute Fellowship. *(Corresponding author: Xiaoqi Ren.)*

The authors are with the Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: xren@caltech.edu; plondon@caltech.edu; jziani@caltech.edu; adamw@caltech.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TNET.2018.2811374

[1]Here we assume either the "raw" data is delivered or the computational overhead to process the data is negligible compared to the purchasing cost and placement cost. This is the case for the most existing data markets [2], [7]–[9], and also is consistent with [14] and [15]. We leave joint optimization of computation, data purchasing, and data placement as future work.

A. Data Purchasing

Given prices and contracts offered by data providers, which providers should a data market purchase from to satisfy a set of client queries with minimal cost?

B. Data Placement

How should purchased data be stored and replicated throughout a geo-distributed data market in order to minimize bandwidth and latency costs? And which clients should be served from which replicas given the locations and data requirements of the clients?

Clearly, these two challenges are highly related: data placement decisions depend on which data is purchased from where, so the bandwidth and latency costs incurred because of data placement must be balanced against the purchasing costs. Concretely, less expensive data that results in larger bandwidth and latency costs is not desirable.

The goal of this paper is to present a design for a geo-distributed data market that jointly minimizes data purchasing and data placement costs.

The combination of data purchasing and data placement decisions makes the task of operating a geo-distributed data market more complex than the task of operating a geo-distributed data analytics system [14]–[18]. Geo-analytics systems minimize the cost (in terms of latency and bandwidth) of moving the data needed to answer client queries, replacing the traditional operation mode where data from multiple data centers was moved to a central data center for processing queries. However, crucially, such systems do not consider the cost of obtaining the data (including purchasing and transferring) from data providers.

Thus, the design of a geo-distributed data market necessitates integrating data purchasing decisions into a geo-distributed data analytics system. To that end, our design builds on the model used in [15] by adding data providers that offer a menu of data quality levels for differing fees. The data placement/replication problem in [15] is already an integer linear program (ILP), and so it is no surprise that the addition of data providers makes the task of jointly optimizing data purchasing and data placement NP-hard (see Theorem 1).

Consequently, we focus on identifying structure in the problem that can allow for a practical and near-optimal system design. To that end, we show that the task of jointly optimizing data purchasing and data placement is equivalent to the uncapacitated facility location problem (UFLP) [19]. However, while constant-factor polynomial running time approximation algorithms are known for the *metric* uncapacitated facility location problem (see [20]–[22]), our problem is a *non-metric* facility location problem, and the best known polynomial running time algorithms achieve a $O(\log C)$ approximation via the greedy algorithm in [23] or the randomized rounding algorithm in [24], where C is the number of clients. Note that without any additional information on the costs, this approximation ratio is the smallest achievable for the non-metric uncapacitated facility location unless NP has slightly superpolynomial time algorithms [25]. While this is the best theoretical guarantee possible in the worst-case, some promising heuristics have been proposed for the non-metric case, e.g., [26]–[31].

Though the task of jointly optimizing data purchasing and data placement is computationally hard in the worst case, in practical settings there is structure that can be exploited. In particular, we provide an algorithm with polynomial running time that gives an exact solution in the case of a data market with a single data center (Section IV-A). Then, using this structure, we generalize to the case of a geo-distributed data cloud and provide an algorithm, named Datum (Section IV-B) that is near optimal in practical settings.

Datum first optimizes data purchasing as if the data market was made up of a single data center (given carefully designed "transformed" costs) and then, given the data purchasing decisions, optimizes data placement/replication. The "transformed" costs are designed to allow an architectural decomposition of the joint problem into subproblems that manage data purchasing (external operations of the data market) and data placement (internal operations of the data market). This decomposition is of crucial operational importance because it means that internal placement and routing decisions can proceed without factoring in data purchasing costs, mimicking operational structures of geo-distributed analytics systems today.

We provide a case study in Section V which highlights that Datum is near-optimal (within 1.6%) in practical settings. Further, the performance of Datum improves upon approaches that neglect data purchasing decisions by $> 45\%$.

To summarize, this paper makes the following main contributions:

1. We initiate the study of jointly optimizing data purchasing and data placement decisions in geo-distributed data markets.
2. We prove that the task of jointly optimizing data purchasing and data placement decisions is NP-hard and can be equivalently viewed as a facility location problem.
3. We provide an exact algorithm with polynomial running time for the case of a data market with a single data center.
4. We provide an algorithm, Datum, for jointly optimizing data purchasing and data placement in a geo-distributed data market that is within 1.6% of optimal in practical settings and improves by $> 45\%$ over designs that neglect data purchasing costs. Importantly, Datum decomposes into subproblems that manage data purchasing and data placement decisions separately.

II. Opportunities and Challenges

Data is now a traded *commodity*. It is being bought and sold every day, but most of these transactions still happen offline through direct negotiations for bulk purchases. This is beginning to change with the emergence of cloud data markets such as Qlik Datamarket [2], Factual [3], InfoChimps [4], Xignite [5]. As cloud data markets become more prominent, data will become a *service* that can be acquired and scaled seamlessly, on demand, similarly to computing resources available today in the cloud.

A. The Potential of Data Markets

The emergence of cloud data markets has the potential to be a significant disruptor for the tech industry, and beyond. Today, since computing resources can be easily obtained and scaled through cloud services, data acquisition has become the bottleneck for new tech startups.

For example, consider an emerging potential competitor for Yelp. The biggest development challenge is not algorithmic or computational. Instead, it is obtaining and managing high quality data at scale. The existence of a data market with detailed local information about restaurants, attractions, etc., would eliminate this bottleneck entirely. In fact, data markets such as Factual [3] are emerging to target exactly this need.

Another example highlighted in [11] and [32] is language translation. Emerging data markets such as Infochimps sell access to data on word translation, word frequency, etc. across languages. This access is a crucial tool for easing the transition tech startups face when moving into different cultural markets.

A final example considers computer vision. When tech startups need to develop computer vision tools in house, a significant bottleneck (in terms of time and cost) is obtaining labeled images with which to train new algorithms. Emerging data markets have the potential to eliminate this bottleneck too. For example, the emerging Visipedia project [33] (while free for now) provides an example of the potential of such a data market.

Thus, like in the case of cloud computing, ease of access and scaling, combined with the cost efficiency that comes with size, implies that cloud data markets have the potential to eliminate one of the major bottlenecks for tech startups today – data acquisition.

B. Operational Challenges for Data Markets

While data pricing within cloud data markets has received increasing attention, the engineering of the system itself has been ignored. The engineering of such a geo-distributed "data cloud" is complex. In particular, the system must jointly make both data purchasing decisions and data placement, replication and delivery decisions, as described in the introduction.

Even considered independently, the task of optimizing data placement/replication within a geo-distributed data analytics system is challenging. Such systems aim to allow queries on databases that are stored across data centers, as opposed to traditional databases that are stored within a single data center. Examples include Google Spanner [34], Mesa [35], JetStream [36], Geode [15], and Iridium [18]. The aim in designing a geo-distributed data analytics system is to distribute the computation needed to answer queries across data centers; thus avoiding the need to transfer all the data to a single data center to respond to queries. This distribution of computation is crucial for minimizing bandwidth and latency costs, but leads to considerable engineering challenges, e.g., handling replication constraints due for fault tolerance and regulatory constraints on data placement due to data privacy. See [15] and [18] for a longer discussion of these challenges and for examples illustrating the benefit of distributed query computation in geo-distributed data analytics systems.

Importantly, all previous work on geo-distributed analytics systems assumes that the system already owns the data. Thus, on top of the complexity in geo-distributed analytics systems, a geo-distributed cloud data market must balance the cost of data purchasing with the impact on data placement/replication costs as well as the decisions for data delivery. For example, if clients who are interested in some data are located close to data center A, while the data provider is located close to data center B (far from data center A), it may be worth it to place that data in data center A rather than data center B. In practice, the problem is more complex since clients are usually geographically distributed rather than centralized and one client may require data from several different data providers.

Additional complexity is created by versioning the data, i.e., the fact that clients have differing quality requirements for the data requested. For example, if some clients are interested in high quality data and others are interested in low quality data, then it may be worth it to provide high quality level data to some clients that only need low quality data (thus incurring a higher price) because of the savings in bandwidth and replication costs that result from being able to serve multiple clients with the same data.

C. Related Work

Our work focuses on the joint design of data purchasing and data placement in a geo-distributed cloud data market. As such, it is related to recent work on data pricing, content distribution networks, and geo-distributed data analytics systems. Further, the algorithmic problem at the core of our design is the facility location problem, and so our work builds on that literature. We discuss related work in these areas.

1) Data Pricing: The design of data markets has attracted increasing interest in recent years, especially in the database community, see [37] for an overview. The current literature mainly focuses on query-based pricing mechanism designs [10], [11], [13] and seldom considers the operating cost of the market service providers (i.e., the data cloud). There is also a growing body of work related to data pricing with differentiated qualities [12], [13], [38], often motivated by privacy.

Similarly, in [39] and [40] a data pricing scheme is proposed for XML trees in which data prices vary with data quality. Quality is measured in different ways, in the form of accuracy [39], or completeness [40]. In these works, data providers offer prices, but clients may propose a price less than that of the data provider, after which the quality of the data will be decreased. Their framework is based on uniform sampling of rooted subtrees in weighted XML documents. Additionally, Muschalle *et al.* [41] and Stahl and Vossen [42], [43] consider issues related to pricing strategies and data quality, and allow customers to suggest prices rather than fixed prices. They consider the *Name Your Own Price* mechanism [42], which is also used in [40]. We however allow data providers to set the price, and also allow for clients to view available quality levels and prices before making a query.

This work relates to data pricing on the data provider side and is orthogonal to our discussion in this paper. We assume

that prices are known, and are instead concerned with data purchasing and placement choices.

2) Geo-Distributed Data Analytics Systems: As cloud servers are increasingly located in geo-distributed systems, analysis and optimization of data stored in geographically distributed data centers has received increasing attention [14], [15], [18], [44], [45]. Bandwidth constraints [14], [15] as well as latency [18] are the two main challenges for system design, and a number of system designs have been proposed, e.g., see Section II-B for more discussion. Our work builds on the model of geo-distributed data analytics systems in [15] and [18], but is distinct from this literature because none of the work on geo-distributed data analytics systems considers the costs associated with purchasing data.

3) Content Distribution Networks: Content Distribution Networks (CDNs) [46]–[49] were originally introduced to improve quality of Internet services to meet the challenge of rapid growth in the intensity of the content. CDNs replicate content and data from source servers to many other servers that are located closer to the end customers; therefore making it possible to process requests for content locally, saving bandwidth and reducing latency [50]–[59]. As such, the algorithms governing CDNs are similar in spirit to those underlying data markets. However, the novelty of data cloud design stems from the fact that it is not enough to replicate content on a network of data centers so as to minimize the cost of delivering the data from sources to customers, data clouds must also consider the fact that there is not free access to the data – it must be *purchased* from providers. This cost significantly changes the algorithmic challenge and leads us to study a framework to *simultaneously* decide (i) what data to *purchase* from which providers and (ii) where to *place* the data on the network and how to route the data to the clients in order to minimize its operating costs.

4) Algorithms for Facility Location: Our data cloud cost minimization problem can be viewed as a variant of the uncapacitated facility location problem. Though such problems have been widely studied, most of the results, especially algorithms with constant approximation ratios, require the assumption of metric cost parameters [20]–[22], which is not the case in our problem. In contrast, for the non-metric facility location problem the best known algorithm is a greedy algorithm proposed in [19]. Beyond this algorithm, a variety of heuristics for solving UPFL have been proposed. However, (i) the usual UPFL heuristics do not take into account the added structure of our problem, and (ii) the new heuristic that we propose, Datum, shows that you can in fact simplify the intermediary's problem by separating purchasing and placement decisions and still provide near-optimal solutions in practical settings. The usual UPFL heuristics do not exhibit such a decomposition. Datum may also be valuable more broadly for facility location problems.

III. A Geo-Distributed Data Cloud

This paper presents a design for a geo-distributed cloud data market, which we refer to as a "data cloud." This data cloud serves as an intermediary between *data providers*, which gather data and offer it for sale, and *clients*, which interact with the data cloud through queries for particular subsets/qualities of data. More concretely, the data cloud purchases data from multiple data providers, aggregates it, cleans it, stores it (across multiple geographically distributed data centers), and delivers it (with low-latency) to clients in response to queries, while aiming at minimizing the *operational cost* constituted of both bandwidth and data purchasing costs.

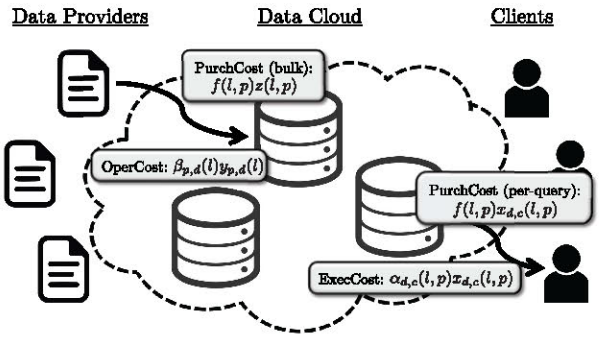


Fig. 1. An overview of the interaction between data providers, the data cloud, and clients. The dotted line encircling the data centers (DC) represents the geo-distributed data cloud. Data providers and clients interact only with the cloud. Data provider p sends data of quality $q(l,p)$ to data center d, and the corresponding operation cost is $\beta_{p,d}(l)y_{p,d}(l)$. Similarly, data center d sends data of quality $q(l,p)$ to client c, and the corresponding execution cost is $\alpha_{d,c}(l,p)x_{d,c}(l,p)$. In bulk data contracting, the corresponding purchasing cost is $f(l,p)z(l,p)$. In per-query data contracting, the corresponding purchasing cost is $f(l,p)x_{d,c}(l,p)$.

Our design builds on and extends the contributions of recent papers – specifically [15] and [18] – that have focused on building geo-distributed data analytic systems but assume the data is already owned by the system and focus solely on the interaction between a data cloud and its clients. Unfortunately, as we highlight in Section IV, the inclusion of data providers means that the data cloud's goal of cost minimization can be viewed as a non-metric uncapacitated facility location problem, which is NP-hard. For ease of exposition, we keep our model simple. We note that however, it can be easily adapted to include various pricing mechanisms on the data provider side, different query structures and execution plans in the data cloud, as well as different types of contracts and payment methods between data cloud and clients.

For reference, Figure 1 provides an overview of the interaction between these three parties as well as some basic notations.

A. Modeling Data Providers

The interaction between the data cloud and data providers is a key distinction between the setting we consider and previous work on geo-distributed data analytics systems such as [15] and [18]. We assume that each data provider offers distinct data to the data cloud, and that the data cloud is a price-taker, i.e., cannot impact the prices offered by data providers. Thus, we can summarize the interaction of a data provider with the data cloud through an exogenous menu of data qualities and corresponding prices.

We interpret the quality of data as a general concept that can be instantiated in multiple ways. For categorical data, quality

may represent the resolution of the information provided, e.g., for geographical attributes the resolution may be {street address, zip code, city, county, state}. For numerical data, quality could take many forms, e.g., the numerical precision, the statistical precision (e.g., the confidence of an estimator), or the level of noise added to the data.[2]

Concretely, we consider a setting where there are P data providers selling different data, *provider* $p \in \mathcal{P} = \{1, 2, \ldots, \mathcal{P}\}$.[3] Each data provider offers a set of quality levels, indexed by *level* $l \in \mathcal{L} = \{1, 2, \ldots, L_p\}$, where L_p is the number of levels that data provider p offers. We use $q(l, p)$ to denote the data quality level l, offered by data provider p. Similarly, we use $f(l, p)$ to denote the fee charged by data provider p for data of quality level l. Importantly, the prices vary across providers p since different providers have different procurement costs for different qualities and different data.

The data purchasing contract between data providers and data cloud may have a variety of different types. For example, a data cloud may pay a data provider based on usage, i.e., per query, or a data cloud may buy the data in bulk in advance. In this paper, we discuss both per-query data contracting and bulk data contracting. See Section III-C.1 for details.

B. Modeling Clients

Clients interact with the data cloud through queries, which may require data (with varying quality levels) from multiple data providers.

Concretely, we consider a setting where there are C clients, *client* $c \in \mathcal{C} = \{1, 2, \ldots, C\}$. A client c sends a *query* to the data center, requesting particular data from multiple data providers.[4] Denote the set of data providers required by the request from client query c by $G(c)$. The client query also specifies a minimum desired quality level, $w_c(p)$, for each data provider p it requests, i.e., $\forall p \in G(c)$. We assume that the client is satisfied with data at a quality level higher than or equal to the level requested.

More general models of queries are possible, e.g., by including a DAG modeling the structure of the query and query execution planning (see [15] for details). For ease of exposition, we do not include such detailed structure here, but it can be added at the expense of more complicated notation.

Depending on the situation, the client may or may not be expected to pay the data cloud for access. If the clients are internal to the company running the data cloud, client payments are unnecessary. However, in many situations the client is expected to pay the data cloud for access to the data. There are many different types of payment structures that could be considered. Broadly, these fall into two categories: (i) subscription-based (e.g., Azure DataMarket [61]) or (ii) per-query-based (e.g. Infochimps [4]).

In this paper, we do not focus on (or model) the design of payment structure between the clients and the data cloud. Instead, we focus on the operational task of minimizing the cost of the data cloud operation (i.e., bandwidth and data purchasing costs). This focus is motivated by the fact that minimizing the operation costs improves the profit of the data cloud regardless of how clients are charged. Interested readers can find analyses of the design of client pricing strategies in [10], [11], and [13].

C. Modeling a Geo-Distributed Data Cloud

The role of the data cloud in this marketplace is as an aggregator and intermediary. We model the data cloud as a geographically distributed cloud consisting of D data centers, *data center* $d \in \mathcal{D} = \{1, 2, \ldots, D\}$. Each data center aggregates data from geographically separate local data providers, and data from data providers may be (and often is) replicated across multiple data centers within the data cloud.

Note that, even for the same data with the same quality, data transfer from the data providers to the data cloud is not a one time event due to the need of the data providers to update the data over time. Here we consider a model where clients are enterprises and clients and data cloud sign on contracts in advance to decide on which data and what data quality to be transferred to the clients. Thus we target the modeling and optimization of data cloud within a fixed time horizon, given the assumption that queries from clients are known beforehand or can be predicted accurately. This assumption is consistent with [15] and [18] and reports from other organizations [62], [63]. Online versions of the problem are also of interest, but are not the focus of this paper.

1) Modeling Costs: Our goal is to provide a design that minimizes the operational costs of a data cloud. These costs include both data purchasing and bandwidth costs. In order to describe these costs, we use the following notation, which is summarized in Figure 1.[5]

- $x_{d,c}(l, p) \in \{0, 1\}$: $x_{d,c}(l, p) = 1$ if and only if data of quality $q(l, p)$, originating from data provider p, is transferred from data center d to client c.
- $\alpha_{d,c}(l, p)$: cost (including bandwidth and/or latency) to transfer data of quality $q(l, p)$, originating from data provider p, from data center d to client c
- $y_{p,d}(l) \in \{0, 1\}$: $y_{p,d}(l) = 1$ if and only if data of quality $q(l, p)$ is transferred from data provider p to data center d.
- $\beta_{p,d}(l)$: cost (including bandwidth and/or latency) to transfer data of quality $q(l, p)$ from data provider p to data center d.
- $z(l, p) \in \{0, 1\}$: $z(l, p) = 1$ if and only if data of quality $q(l, p)$, originating from data provider p, is transferred to the data cloud.

[2]A common suggestion for guaranteeing privacy is to add Laplace noise to data provided to data markets, see e.g., [13], [60]

[3]We distinguish data providers based on data, i.e., one data provider sells multiple data is treated as multiple data providers.

[4]We distinguish clients based on queries, i.e., one client sends multiple queries is treated as multiple clients.

[5]Throughout, subscript indices refer to data transfer "from, to" a location, and parenthesized indices refer to data characteristics (e.g., quality, from which data provider).

$f(l,p)$: purchasing cost of data with quality $q(l,p)$, originating from data provider p.

Given the above notations, the costs of the data cloud can be broken into three categories:

1) The *operation cost* due to transferring data of all quality levels from data providers to data centers is

$$\text{OperCost} = \sum_{p=1}^{P} \sum_{l=1}^{L_p} \sum_{d=1}^{D} \beta_{p,d}(l) y_{p,d}(l). \tag{1}$$

2) The *execution cost* due to transferring data of all quality levels from data centers to clients is

$$\text{ExecCost} = \sum_{c=1}^{C} \sum_{p \in G(c)} \sum_{l=1}^{L_p} \sum_{d=1}^{D} \alpha_{d,c}(l,p) x_{d,c}(l,p). \tag{2}$$

3) The *purchasing cost* (PurchCost) due to buying data from the data provider could result from a variety of differing contract styles. In this paper we consider two extreme options: *per-query* and *bulk* data contracting. These are the most commonly adopted strategies for data purchasing today.

In *per-query* data contracting, the data provider charges the data cloud a fixed rate for each query that uses the data provided by the data provider. So, if the same data is used for two different queries, then the data cloud pays the data provider twice. Given a per-query fee $f(l,p)$ for data $q(l,p)$, the total purchasing cost is

$$\text{PurchCost(query)} = \sum_{c=1}^{C} \sum_{p \in G(c)} \sum_{l=1}^{L_p} \sum_{d=1}^{D} f(l,p) x_{d,c}(l,p). \tag{3}$$

In *bulk* data contracting, the data cloud purchases the data in bulk and then can distribute it without owing future payments to the data provider. Given a one-time fee $f(l,p)$ for data $q(l,p)$, the total purchasing cost is

$$\text{PurchCost(bulk)} = \sum_{p=1}^{P} \sum_{l=1}^{L_p} f(l,p) z(l,p). \tag{4}$$

To keep the presentation of the paper simple, we focus on the per-query data contracting model throughout the body of the paper and discuss the bulk data contracting model (which is simpler) in Appendix D.

2) Cost Optimization: Given the cost models described above, we can now represent the goal of the data cloud via the following integer linear program (ILP), where OperCost, ExecCost, and PurchCost are as described in equations (1), (2) and (3), respectively.

$$\min_{x,y} \ \text{OperCost} + \text{ExecCost} + \text{PurchCost} \tag{5}$$

$$\text{subject to } x_{d,c}(l,p) \le y_{p,d}(l) \quad \forall c,p,l,d \tag{5a}$$

$$\sum_{l=1}^{L_p} \sum_{d=1}^{D} x_{d,c}(l,p) = 1, \quad \forall c, p \in G(c) \tag{5b}$$

$$\sum_{l=1}^{L_p} \sum_{d=1}^{D} x_{d,c}(l,p) q(l,p) \ge w_c(p), \quad \forall c, p \in G(c) \tag{5c}$$

$$x_{d,c}(l,p) \ge 0, \quad \forall c,p,l,d \tag{5d}$$

$$y_{p,d}(l) \ge 0, \quad \forall p,l,d \tag{5e}$$

$$x_{d,c}(l,p), y_{p,d}(l) \in \{0,1\}, \quad \forall c,p,l,d \tag{5f}$$

The constraints in this formulation warrant some discussion. Constraint (5a) states that any data transferred to some client must already have been transferred from its data provider to the data cloud.[6] Constraint (5b) ensures that each client must get the data it requested, and constraint (5c) ensures that the minimum quality requirement of each client must be satisfied. The remaining constraints state that the decision variables are binary and nonnegative.

An important observation about the formulation above is that data purchasing/placement decisions are decoupled across data providers, i.e., the data purchasing/placement decision for data from one data provider does not impact the data purchasing/placement decision for any other data providers. Thus, we frequently drop the index p.

Note that there is a variety of practical issues that we have not incorporated into the formulation in (5) in order to minimize notational complexity, but which can be included without affecting the results described in the following. A first example is that a minimal level of data replication is often desired for fault tolerance and disaster recovery reasons. This can be added to (5) by additionally considering constraints of the form $\sum_{d=1}^{D} y_{p,d}(l) \ge kz(l,p)$, where k denotes the minimum required number of copies. Similarly, privacy concerns often lead to regulatory constraints on data movement. As a result, regulatory restrictions may prohibit some data from being copied to certain data centers, thus constraining data placement and replication. This can be included by adding constraints of the form $y_{p,d}(l) = 0$ to (5) where p and d denote the corresponding data provider and data center, respectively. Finally, in some cases it is desirable to enforce SLA constraints on the latency of delivery to clients. Such constraints can be added by including constraints of the form $\sum_{p \in G(c)} \sum_{l=1}^{L_p} \sum_{d=1}^{D} \alpha_{d,c}(l,p) x_{d,c}(l,p) \le r_c$, where r_c denotes the SLA requirement of client c.

We refer the reader to [14], [15], and [18] for more discussions of these additional practical constraints. Each paper includes a subset of these factors in the design of geo-distributed data analytics systems, but does not model data purchasing decisions.

[6] For bulk data contracting model, one more constraint $y_{p,d}(l) \le z(l,p), \ \forall c,l,p,d$ is required. This constraint states that any data placed in the data cloud must be purchased by the data cloud.

IV. Optimal Data Purchasing & Data Placement

Given the model of a geo-distributed data cloud described in the previous section, the design task is now to provide an algorithm for computing the optimal data purchasing and data placement/replication decisions, i.e., to solve data cloud cost minimization problem in (5). Unfortunately, this cost minimization problem is an ILP, which are computationally difficult in general.[7]

A classic NP-hard ILP is the uncapacitated facility location problem (UFLP) [19]. In the uncapacitated facility location problem, there is a set of I clients and J potential facilities. Facility $j \in J$ costs f_j to open and can serve clients $i \in I$ with cost $c_{i,j}$. The task is to determine the set of facilities that serves the clients with minimal cost.

Our first result, stated below, highlights that cost minimization for a geo-distributed data cloud can be reduced to the uncapacitated facility location problem, and vice-versa. Thus, the task of operating a data cloud can then be viewed as a facility location problem, where opening a facility parallels purchasing a specific quality level from a data provider and placing it in a particular data center in the data cloud.

Theorem 1: The cost minimization problem for a geo-distributed data cloud given in (5) *is NP-hard.*

The proof of Theorem 1 (given in Appendix A of the online supplementary material) provides a reduction both to and from the uncapacitated facility location problem. Importantly, the proof of Theorem 1 serves a dual purpose: it both characterizes the hardness of the data cloud cost minimization problem and highlights that algorithms for the facility location problem can be applied in this context. Given the large literature on facility location, this is important.

More specifically, the reduction leading to Theorem 1 highlights that the data cloud optimization problem is equivalent to the *non-metric* uncapacitated facility location problem – every instance of any of the two problems can be written as an instance of the other. While constant-factor polynomial running time approximation algorithms are given for the *metric* uncapacitated facility location problem in [20]–[22], in the more general *non-metric* case the best known polynomial running time algorithm achieves a $\log(C)$-approximation via a greedy algorithm with polynomial running time, where C is the number of clients [23]. This is the best worst-case guarantee possible (unless NP has slightly superpolynomial time algorithms, as proven in [25]); however some promising heuristics have been proposed for the non-metric case, e.g., [26]–[31].

Nevertheless, even though our problem can, in general, be viewed as the non-metric uncapacitated facility location, it does have a structure in real-world situations that we can exploit to develop practical algorithms.

In particular, in this section we begin with the case of a data cloud made up of a single data center. We show that, in this case, there is a structure that allows us to design an algorithm with polynomial running time that gives an exact solution (Section IV-A). Then, we move to the case of a data cloud made up of geo-distributed data centers and highlight how to build on the algorithm for the single data center case to provide an algorithm, Datum, for the general case (Section IV-B). Importantly, Datum allows decomposition of the management of data purchasing (operations outside of the data cloud) and data placement (operations inside the data cloud). This feature of Datum is crucial in practice because it means that the algorithm allows a data cloud to manage internal operations without factoring in data purchasing costs, mimicking operations today. While we do not provide analytic guarantees for Datum (as expected given the reduction to/from the non-metric facility location problem), we show that the heuristic performs well in practical settings using a case study in Section V.

A. An Exact Solution for a Single Data Center

We begin our analysis by focusing on the case of a single data center, which interacts with multiple data providers and multiple clients. The key observation is that, if the execution costs associated with transferring different quality levels of the same data are the same, i.e., $\forall l, \alpha_c(l) = \alpha_c$, then the execution cost becomes a constant which is independent of the data purchasing and data placement decisions as shown in (6).

$$\text{ExecCost} = \sum_{c=1}^{C}\sum_{l=1}^{L} \alpha_c x_c(l) = \sum_{c=1}^{C} \alpha_c \left(\sum_{l=1}^{L} x_c(l) \right) = \sum_{c=1}^{C} \alpha_c \tag{6}$$

The assumption that the execution costs are the same across quality levels is natural in many cases. For example, if quality levels correspond to the level of noise added to numerical data, then the size of the data sets will be the same. We adopt this assumption in Section IV-A and extend to the general case in Section IV-B.

This assumption allows the elimination of the execution cost term from the objective. Additionally, we can simplify notation by removing the index d for the data center. Thus, in per-query data contracting, the data cloud optimization problem can be simplified to (7). (We discuss the case of bulk data contracting in Appendix D.)

$$\text{minimize} \sum_{l=1}^{L} \beta(l) y(l) + \sum_{c=1}^{C}\sum_{l=1}^{L} f(l) x_c(l) \tag{7}$$

$$\begin{aligned} \text{subject to } & x_c(l) \le y(l), \quad \forall c, l \\ & \sum_{l=w_c}^{L} x_c(l) = 1, \quad \forall c \\ & x_c(l) \ge 0, \quad \forall c, l \\ & y(l) \ge 0, \quad \forall l \\ & x_c(l), y(l) \in \{0,1\}, \quad \forall c, l \end{aligned} \tag{7a}$$

Note that constraint (7a) is a contraction of (5b) and (5c), and simply means that any client c must be given exactly one

[7]Note that previous work on geo-distributed data analytics where data providers and data purchasing were not considered already leads to an ILP with limited structure. For example, Vulimiri *et al.* [15] suggest only heuristic algorithms with no analytic guarantees.

quality level above w_c, the minimum required quality level.[8] The remaining constraints follow directly from (5) by dropping d since we only consider one data center case in (7). While this problem is still an ILP, in this case there is a structure that can be exploited to provide a polynomial time algorithm that can find an exact solution. In particular, in Appendix B we prove that the solution to (7) can be found by solving the linear program (LP) given in (8).

$$\begin{aligned}
&\text{minimize } \sum_{l=1}^{L}\beta(l)y(l)+\sum_{i=1}^{L}\sum_{l=i}^{L}S_i f(l)\chi_i(l)\\
&\text{subject to}\\
&\quad \chi_i(l)\le y(l), \quad \forall i,l\\
&\quad \sum_{l=i}^{L}\chi_i(l)=1, \quad \forall i\\
&\quad \chi_i(l)\ge 0, \quad \forall i,l\\
&\quad y(l)\ge 0, \quad \forall l
\end{aligned} \tag{8}$$

In (8), S_i is the number of clients who require a minimum quality level of i, and $\chi_i(l) = 1$ represents clients with minimum required quality level i purchase at quality level l.

Note that this LP is not directly obtained by relaxing the integer constraints in (7), but is obtained from relaxing the integer constraints in a reformulation of (7) described in Appendix B. The theorem below provides a tractable, exact algorithm for cost minimization in a data cloud made up of a single data center. (A proof is given in Appendix B of the online supplementary material).

Theorem 2: There exists a binary optimal solution to the linear relaxiation program in (8) *which is an optimal solution of the integer program in* (7) *and can be found in polynomial time.*

In summary, the following gives a polynomial time algorithm which yields the optimal solution of (7).

Step 1: Rewrite (7) in the form given by (4).

Step 2: Solve the linear relaxation of (4), i.e., (8). If it gives an integral solution, this solution is an optimal solution of (7), and the algorithm finishes. Otherwise, denote the fractional solution of the previous step by $\{\chi^r(l), y^r(l)\}$ and continue to the next step.

Step 3: Find $m_i \in \{i,\ldots,n\}$ such that $\sum_{l=i}^{m_i-1} y^r(l) < 1$, and $\sum_{l=i}^{m_i} y^r(l) \ge 1$. (See Appendix B for the existence of $\{m_i\}$.) And express $\{\chi_i(l)\}$ as a function of $\{y(l)\}$ based on (6). Substitute the expressions of $\{\chi_i(l)\}$ with $\{y(l)\}$ in (8) to obtain an instance of (7). Solve the linear programming problem (7) and find an optimal solution that is also an extreme point of (7).[9] This yields a binary optimal solution of (7). Use transformation (6) to get a binary optimal solution of (8), which can be reformulated as an optimal solution of (7) from the definition of $\{\chi_i(l)\}$.

[8]While the two constraints are equivalent for an ILP, they lead to different feasible sets when considering its LP-relaxation; in particular, facility location algorithms based on LP-relaxations such as randomized rounding algorithms need to use the contracted version of the constraints to preserve the $O(\log C)$-approximation ratio for non-metric facility location. It is equivalent to the reformulation given in Appendix A and does not introduce infinite costs that may lead to numerical errors.

[9]This step can be finished in polynomial time [64].

B. The Design of Datum

Unlike the data cloud cost minimization problem for a single data center, the general data cloud cost minimization is NP-hard. In this section, we build on the exact algorithm for cost minimization in a data cloud made up of a single data center (Section IV-A) to provide an algorithm, Datum, for cost minimization in a geo-distributed data cloud.

The idea underlying Datum is to, first, optimize data purchasing decisions as if the data market was made up of a single data center (given carefully designed "transformed" costs), which can be done tractably as a result of Theorem 2. Then, second, Datum optimizes data placement/replication decisions given the data purchasing decisions.

Before presenting Datum, we need to reformulate the general cost minimization ILP in (5). Recall that (5) is separable across providers, thus we can consider independent optimizations for each provider, and drop the index p throughout. Second, we denote the set of all possible subsets of data centers, e.g., $\{\{d_1\},\{d_2\},\ldots,\{d_1,d_2\},\{d_1,d_3\},\ldots\}$ by V.[10] Further, define $\beta_v(l) = \sum_{d\in v}\beta_d(l)$, and $\alpha_{v,c}(l) = \min_{d\in v}\{\alpha_{d,c}(l)\}$, where $v \in V$. Given this change, we define $y_v(l) = 1$ if and only if data with quality level l is placed in (and only in) data centers $d \in v$ and $x_{v,c}(l) = 1$ if and only if data with quality level l is transferred to client c from some data center $d \in v$. These reformulations allow us to convert (5) to (9) as following.

$$\begin{aligned}
&\text{minimize } \sum_{l=1}^{L}\sum_{v=1}^{V}\beta_v(l)y_v(l)+\sum_{c=1}^{C}\sum_{l=1}^{L}\sum_{v=1}^{V}\alpha_{v,c}(l)x_{v,c}(l)\\
&\qquad +\sum_{c=1}^{C}\sum_{l=1}^{L}\sum_{v=1}^{V}f(l)x_{v,c}(l) && (9)\\
&\text{subject to } x_{v,c}(l)\le y_v(l), \quad \forall c,l && (9a)\\
&\qquad \sum_{l=w_c}^{L}\sum_{v=1}^{V}x_{v,c}(l)=1, \quad \forall c && (9b)\\
&\qquad \sum_{v=1}^{V}y_v(l)\le 1, \quad \forall l && (9c)\\
&\qquad \sum_{v=1}^{V}x_{v,c}(l)\le 1, \quad \forall c,l && (9d)\\
&\qquad x_{v,c}(l)\ge 0, \quad \forall v,c,l && (9e)\\
&\qquad y_v(l)\ge 0, \quad \forall v,l && (9f)\\
&\qquad x_{v,c}(l), y_v(l)\in\{0,1\}, \quad \forall v,c,l && (9g)
\end{aligned}$$

Compared to (5), the main difference is that (9) has two extra constraints (9c) and (9d). Constraint (9c) ensures that data can only be placed in at most one subset of data centers across V. And constraint (9d) follows from constraint (9b). Using this reformulation Datum can now be explained in two steps.

[10]Note that, in practice, the number of data centers is usually small, e.g., $10-20$ world-wide. Further, to avoid exponential explosion of V, the subsets included in V can be limited to only have a constant number of data centers, where the constant is determined by the maximal number of replicas to be stored.

Step 1: Solve (IV-B) while treating the geo-distributed data cloud as a single data center. Specifically, define $Y(l) = \sum_{v=1}^{V} y_v(l)$ and $X_c(l) = \sum_{v=1}^{V} x_{v,c}(l)$. Note that, $Y(l)$ and $X_c(l)$ are $0-1$ variables from Constaint (9c) and (9d). Further, ignore the middle term in the objective, i.e., the ExecCost. Finally, for each quality level l, consider a "transformed" cost $\beta^*(l)$. We discuss how to define $\beta^*(l)$ below. This leaves the "single data center" problem (IV-B). Crucially, this formulation can be solved optimally in polynomial time using the results for the case of a data cloud made up of a single data center (Section IV-A).

$$\begin{aligned}
\text{minimize } & \sum_{l=1}^{L} \beta^*(l) Y(l) + \sum_{c=1}^{C} \sum_{l=1}^{L} f(l) X_c(l) \\
\text{subject to } & X_c(l) \le Y(l), \quad \forall c, l \\
& \sum_{l=w_c}^{L} X_c(l) = 1, \quad \forall c \\
& X_c(l) \ge 0, \quad \forall c, l \\
& Y(l) \ge 0, \quad \forall l \\
& X_c(l), Y(l) \in \{0,1\}, \quad \forall c, l
\end{aligned} \tag{10}$$

The remaining issue is to define $\beta^*(l)$. Note that the reason for using transformed costs $\beta^*(l)$ instead of $\beta_v(l)$ is that the optimal $\beta_v(l)$ cannot be known precisely without also optimizing the data placement. Thus, in defining $\beta^*(l)$ we need to anticipate the execution costs that result from data placement and replication given the purchase of data with quality level l. This anticipation then allows a decomposition of data purchasing and data placement decisions. Note that the only inaccuracy in the heuristic comes from the mismatch between $\beta^*(l)$ and $\min\{\beta_v(l) + \sum_{c \in C^*(l)} \alpha_{v,c}(l)\}$ where $C^*(l)$ is the set of customers who buy at quality level l in an optimal solution – if these match for the minimizer of (5) then the heuristic is exact. Indeed, in order to minimize the cost of locating quality levels to data centers, and allocating clients to data centers and quality levels, the set of data centers v where an optimal solution chooses to put quality level l has to minimize the cost of data transfer in the set v and allocating all clients who get data at quality level l, i.e. $C^*(l)$, to this set of data centers v.

Many choices are possible for the transformed costs $\beta^*(l)$. A conservative choice is $\beta^*(l) = \min_v \beta_v(l)$, which results in a solution (with Step 2) whose OperCost + PurchCost is a lower bound to the corresponding costs in the optimal solution of (5).[11] However, it is natural to think that more aggressive estimates may be valuable. To evaluate this, we have performed experiments in the setting of the case study (see Section V) using the following parametric form $\beta^*(l) = \min_v\{\beta_v(l) + \mu_1 \sum_{l' \le l} \sum_{w_c = l'} \alpha_{v,c}(l') e^{-\mu_2 (l - l')}\}$, where μ_1 and μ_2 are parameters. This form generalizes the conservative choice by providing a weighting of $\alpha_{v,c}(l')$ based on the "distance" of the quality deviation between l' and the target quality level l. The idea behind this is that a client is more likely to be served data with quality level close to the requested minimum quality level of the client. Here we use the exponential decay term $e^{-\mu_2(l-l')}$ to capture the possibility of serving the data with quality level l to a client with minimum quality level $l' \le l$. Interestingly, in the setting of our case study, the best design is $\mu_1 = \mu_2 = 0$, i.e., the conservative estimate $\beta^*(l) = \min_v \beta_v(l)$, and so we adopt this $\beta^*(l)$ in Datum.

Step 2: At the completion of Step 1 the solution (X, Y) to (IV-B) determines which quality levels should be purchased and which quality level should be delivered to each client. What remains is to determine data placement and data replication levels. To accomplish this, we substitute (X, Y) into (9), which yields (11).

$$\begin{aligned}
\text{minimize } & \sum_{l=1}^{L} \sum_{v=1}^{V} \beta_v(l) y_v(l) + \sum_{c=1}^{C} \sum_{l=1}^{L} \sum_{v=1}^{V} \alpha_{v,c}(l) x_{v,c}(l) \\
& + \sum_{c=1}^{C} \sum_{l=1}^{L} \sum_{v=1}^{V} f(l) x_{v,c}(l) && (11) \\
\text{subject to } & x_{v,c}(l) \le y_v(l), \ \forall c, l && (11a) \\
& \sum_{l=w_c}^{L} \sum_{v=1}^{V} x_{v,c}(l) = 1, \ \forall c && (11b) \\
& \sum_{v=1}^{V} y_v(l) = Y(l) && (11c) \\
& \sum_{v=1}^{V} x_{v,c}(l) = X_v(l) && (11d) \\
& x_{v,c}(l) \ge 0, \quad \forall v, c, l && (11e) \\
& y_v(l) \ge 0, \quad \forall v, l && (11f) \\
& x_{v,c}(l), y_v(l) \in \{0,1\}, \quad \forall v, c, l && (11g)
\end{aligned}$$

The key observation is that this is no longer a computationally hard ILP. In fact, the inclusion of (X, Y) means that it can be solved in closed form. Specifically, let $C(l)$ denote the set of clients that purchase data with quality level l, i.e., $C(l) = \{c : X_c(l) = 1\}$. Then (12) gives the optimal solution of (11). (A proof is given in Appendix C of the online supplementary material.)

$$y_v(l) = \begin{cases} 1, & \text{if } Y(l) = 1 \text{ and} \\ & v = \arg\min\{\beta_v(l) + \sum_{c \in C(l)} \alpha_{v,c}(l)\}, \\ 0, & \text{otherwise.} \end{cases} \tag{12a}$$

$$x_{v,c}(l) = \begin{cases} y_v(l), & \text{if } c \in C(l), \\ 0, & \text{otherwise.} \end{cases} \tag{12b}$$

V. Case Study

We now illustrate the performance of Datum using a case study of a geo-distributed data cloud running in North America. While the setting we use is synthetic, we attempt to faithfully model realistic geography for data centers in the data cloud, data providers, and clients. Our focus is on quantifying the overall cost (including data purchasing and bandwidth/latency costs) of Datum compared to two existing

[11] However the ExecCost cannot be bounded, thus we cannot obtain a bound for the total cost. The proof of this is simple and is not included in the paper due to space limit.

designs for geo-distributed data analytics systems and the optimal. To summarize, the highlights of our analysis are

1. Datum provides consistently lower cost ($> 45\%$ lower) than existing designs for geo-distributed data analytics systems.
2. Datum achieves near optimal total cost (within 1.6%) of optimal.
3. Datum achieves reduction in total cost by significantly lowering purchasing costs without sacrificing bandwidth/latency costs, which stay typically within 20-25% of the minimal bandwidth/latency costs necessary for delivery of the data to clients.

A. Experimental Setup

The following outlines the setting in which we demonstrate the empirical performance of Datum.

1) Geo-Distributed Data Cloud: We consider a geographically distributed data cloud with 10 data centers located in California, Washington, Oregon, Illinois, Georgia, Virginia, Texas, Florida, North Carolina, and South Carolina. The locations of the data centers in our experiments mimic those in [65] and include the locations of Google's data centers in the United States.

2) Clients: Client locations are picked randomly among US cities, weighted proportionally to city populations. We consider 100 clients. Each client requests data from a subset of data providers, chosen *i.i.d.* from a Uniform distribution. Unless otherwise specified, the average number of providers per client request is $P/2$. There are 8 quality levels. The quality level requested from each chosen provider follows a Zipf distribution with mean $L_p/2$ and shape parameter 30. P and L_p are defined as in Section III-A and Section III-B. We choose a Zipf distribution motivated by the fact that popularity typically follows a heavy-tailed distribution [66]. Results are averaged over 20 random instances. We observe that the results of the 20 instances for the same plot are very close (within 5%), and thus do not show the confidence intervals on the plots.

3) Data Providers: We consider 20 data providers. We place data providers in the second and third largest cities within a state containing a data center. This ensures that the data providers are near by, but not right on top of, data center and client locations.

4) Operation and Execution Costs: To set operation and execution costs, we compute the geographical distances between data centers, clients and providers. The operation and execution costs are proportional to the geographical distances, such that the costs are effectively one dollar per gigameter. This captures both the form of bandwidth costs adopted in [14] and the form of latency costs adopted in [18].

5) Data Purchasing Costs: The per-query purchasing costs are drawn *i.i.d.* from a Pareto distribution with mean 10 and shape parameter 2 unless otherwise specified. We choose a Pareto distribution motivated by the fact that incomes and prices often follow heavy-tailed distributions [66]. Results were averaged over 20 random instances. To study the sensitivity of Datum to the relative size of purchasing and bandwidth costs, we vary the ratio of them between $(0.01, 100)$.

6) Baselines: We compare the performance of Datum to the following baselines. leftmargin = *

- *OptCost* computes the optimal solution to the data cloud cost minimization problem by solving the integer linear programming (5). Note that this requires solving an NP-hard problem, and so is not feasible in practice. We include it in order to benchmark the performance of Datum.
- *OptBand* computes the optimal solution to the bandwidth cost minimization problem. It is obtained by minimizing only the operation cost and execution cost in the objective of (5). Bandwidth cost minimization is commonly considered as a primary goal for cost minimization in geo-distributed data analytics systems [15]. Due to computational complexity, heuristics are usually applied to minimize the bandwidth cost. Here, instead of implementing a heuristic algorithms, we optimistically use OptBand in order to lower bound the achievable performance. Note that this also requires solving an NP-hard problem and thus is not feasible in practice.
- *NearestDC* is a greedy heuristic for the total cost minimization problem that is often applied in practice. It serves the clients exactly what they ask for by purchasing the data and storing it at the data center closest to the data provider.

B. Experimental Results

1) Quantifying Cost Reductions From Datum*:* Figure 2(a) illustrates the costs savings Datum provides. Across levels of query complexity (number of providers involved), Datum consistently provides $> 45\%$ savings over OptBand and $> 51\%$ savings compared to NearestDC. Further, Datum is within 1.6% of the optimal cost in all these cases. The improvement of Datum compared to OptBand comes as a result of optimizing purchasing decisions at the expense of increased bandwidth. Importantly, Figure 2(b) shows that the extra bandwidth cost incurred is small, $20 - 25\%$. Thus, joint optimization of data purchasing and data placement decisions leads to significant reductions in total cost without adversely impacting bandwidth costs.

2) The Form of Client Queries: To understand the sensitivity of the cost reductions provided by Datum, we next consider the impact of parameters related to client queries. Figure 2 shows that the complexity of queries has little impact on the cost reductions of Datum. Figure 3 studies two other parameters: the heaviness of the tail of the per-query purchasing fee and the number of quality levels offered.

Across all settings, Datum is within 1.6% of optimal; however both of these parameters have a considerable impact on the cost savings Datum provides over our baselines. In particular, the lighter the tail of the prices of different quality levels is, the less improvement can be achieved. This is a result of more concentration of prices across quality levels leaving less room for optimization. Similarly, fewer quality levels provides less opportunity to optimize data purchasing decisions. At the extreme, with only quality level available, the opportunity to optimization data purchasing goes away and OptBand and OptCost are equivalent.

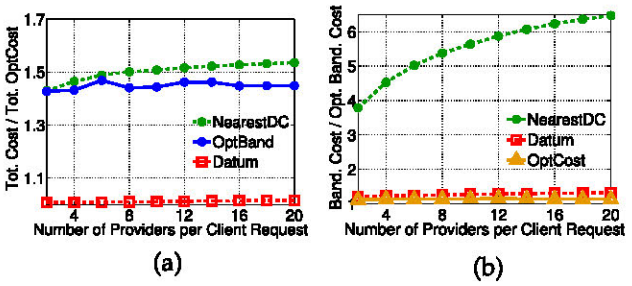


Fig. 2. Illustration of the near-optimality of Datum as a function of the complexity of client requests (i.e., the average number of providers data must be procured from in order to complete a client request) (a) Total cost. (b) Bandwidth cost.

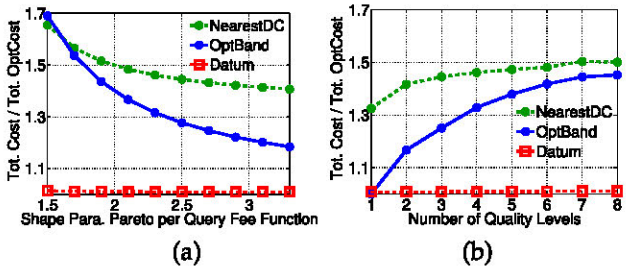


Fig. 3. Illustration of Datum's sensitivity to query parameters. (a) varies the heaviness of the tail in the distribution of purchasing fees. (b) varies the number of quality levels available. Note that Figure 2 sets the shape parameter of the Pareto governing purchasing fees to 2 and includes 8 quality levels.

3) Data Purchasing Vs. Bandwidth Costs: The most important determinant of the magnitude of Datum's cost savings is the relative importance of data purchasing costs. In one extreme, if data is free, then the data purchasing decisions disappear and the problem is simply to do data placement in a manner that minimizes bandwidth costs. In the other extreme, if data purchasing costs dominate then data placement is unimportant. In Figure 4 we only compare total costs among OptCost, OptBand, and Datum. NearestDC is far worse (more than 5 times worse than OptCost in some cases) and thus is dropped from the plots. Figure 4(a) studies the impact of the relative size of data purchasing and bandwidth costs. When the x-axis is 0, the data purchasing and bandwidth costs of the data center are balanced. Positive values mean that bandwidth costs dominate and negative values mean that data purchasing costs dominate. As expected, Datum's cost savings are most dramatic in regimes where data purchasing costs dominate. Cost savings can be 54% in extreme settings. Data purchasing costs are expected to dominate in the future – for some systems this is already true today. However, it is worth noting that, in settings where bandwidth costs dominates, Datum can deviate from the optimal cost by $10-20\%$ in extreme circumstances, and can be outperformed by the MinBand benchmark. Of course, Datum is not designed for such settings given its prioritization of the minimization of data purchasing costs.

4) Internal Vs. External Costs: An important aspect of the design of Datum is the decomposition of data purchasing decisions from data placement decisions. This provides a

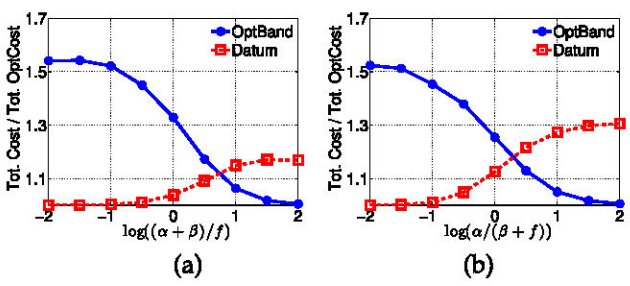


Fig. 4. Illustration of the impact of bandwidth and purchasing fees on Datum's performance. NearestDC is excluded because its costs are off-scale. (a) varies the ratio of bandwidth costs (summarized by $\alpha+\beta$) to purchasing costs (summarized by f). (b) varies the ratio of costs internal to the data cloud (α) to costs external to the data cloud ($\beta+f$). Note that in Figure 2 the ratios are set to $\log(\frac{\alpha+\beta}{f})=-0.5$ and $\log(\frac{\alpha}{\beta+f})=-1$.

separation between the internal and external operations (and costs) of Datum. Given this separation, it is important to evaluate the sensitivity of Datum's design to the relative size of internal and external costs.

Since Datum prioritizes the optimization of external costs (optimizing them in Step 1, see Section IV-B), it is natural to expect that Datum performs best when these costs dominate. This is indeed the case, as illustrated in Figure 4(b). Like in Figure 4(a), when the x-axis is 0, the internal and external costs are balanced. Positive values indicate the internal costs dominate and negative values indicate the external costs dominate. In settings where external costs dominate Datum can provide 50% cost savings and be within a few percent of the optimal. However, in cases when internal costs dominate Datum can deviate from the optimal cost by $10-30\%$ in extreme circumstances, and can be outperformed by the MinBand benchmark. Note that, as data purchasing costs grow in importance, external costs will dominate, and so we can expect that Datum will provide near optimal performance in practical settings.

5) Scalability: Computing the optimal solution to the ILP as our benchmark is a NP-hard problem and can quickly become computationally intractable as the problem size grows. Thus, we limit the number of clients to 100 in our evaluation. In contrast, Datum scales well as the problem size grows, since it only requires solving a linear program with size that scales linearly with the problem size.

VI. Concluding Remarks

This work sits at the intersection of two recent trends: the emergence of online data marketplaces and the emergence of geo-distributed data analytics systems. Both have received significant attention in recent years across academia and industry, changing the way data is bought and sold and changing how companies like Facebook run queries across geo-distributed databases [14], [15]. In this paper we study the engineering challenges that come when online data marketplaces are run on top of a geo-distributed data analytics infrastructure. Such cloud data markets have the potential to be a significant disruptor (as we highlight in Section II). However, there are many unanswered economic and engineering questions

about their design. While there has been significant prior work on economic questions [10], [11], [13], [32], [37], [67], the engineering questions have received much less attention.

In this paper, we have presented the design of a geo-distributed cloud data market: Datum. Datum jointly optimizes data purchasing decisions with data placement decisions in order to minimize the overall cost. While the overall cost minimization problem is NP-hard (via a reduction to/from the facility location problem), Datum provides near-optimal performance (within 1.6% of optimal) in realistic settings via a polynomial-time algorithm that is provably optimal in the case of a data cloud running on a single data center. Additionally, Datum provides $> 45\%$ improvement over current design proposals for geo-distributed data analytics systems. Datum works by decomposing the total cost minimization problem into subproblems that allow optimization of data purchasing and data placement separately, which provides a practical route for implementation in real systems. Further, Datum provides a unified solution across systems using per-query pricing or bulk pricing, systems with data replication constraints and/or regulatory constraints on data placement, and systems with SLA constraints on delivery.

This paper is meant to initiate the study of data purchasing and data placement for data markets; thus, there many directions are left for future exploration. For example, Datum assumes clients are single entities with fixed locations and which data they need is known a priori as a result of pre-signed contracts with data market providers. In practice, clients can also be geo-distributed large companies with different data requirements in different locations. Exploring the optimal data purchasing and data placement with a mixed types of clients is interesting and challenging. Further, in Datum, computing resources used to process/clean raw data are assumed to be negligible. As data markets provide services on top of the raw data, e.g., analytics or learning services, the amount of computational power needed will grow and joint optimization of computational power and data purchasing and placement will become a crucial challenge.

REFERENCES

[1] (2015). *Study Identifies Common Pain Points in Big Data Projects*. [Online]. Available: http://www.bigdataexchange.com/tag/list-of-third-party-data-providers/

[2] *Qlik*. [Online]. Available: http://www.qlik.com/us/products/qlik-data-market

[3] (2015). *Factual*. [Online]. Available: https://www.factual.com/

[4] (2015). *Infochimps*. [Online]. Available: http://www.infochimps.com/

[5] (2015). *Xignite*. [Online]. Available: http://www.xignite.com/

[6] (2015). *The IUPHAR/BPS Guide to Pharmacology*. [Online]. Available: http://www.guidetopharmacology.org/

[7] *Google BigQuery Public Datasets*. [Online]. Available: https://cloud.google.com/bigquery/public-data/

[8] *Azure Public Datasets*. [Online]. Available: https://docs.microsoft.com/en-us/azure/sql-database/sql-database-public-data-sets

[9] *AWS Public Datasets*. [Online]. Available: https://aws.amazon.com/public-datasets/

[10] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-based data pricing," in *Proc. 31st Symp. Principles Database Syst.*, 2012, pp. 167–178.

[11] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Toward practical query pricing with QueryMarket," in *Proc. SIGMOD*, 2013, pp. 613–624.

[12] L. K. Fleischer and Y.-H. Lyu, "Approximately optimal auctions for selling privacy when costs are correlated with data," in *Proc. 13th ACM Conf. Electron. Commerce*, 2012, pp. 568–585.

[13] C. Li, D. Y. Li, G. Miklau, and D. Suciu, "A theory of pricing private data," *ACM Trans. Database Syst.*, vol. 60, no. 12, pp. 79–86, 2014.

[14] A. Vulimiri, C. Curino, B. Godfrey, K. Karanasos, and G. Varghese, "WANalytics: Analytics for a geo-distributed data-intensive world," in *Proc. CIDR*, 2015, pp. 1–9.

[15] A. Vulimiri, C. Curino, P. B. Godfrey, J. Padhye, and G. Varghese, "Global analytics in the face of bandwidth and regulatory constraints," in *Proc. NSDI*, 2015, pp. 323–336.

[16] K. Hsieh *et al.*, "Gaia: Geo-distributed machine learning approaching LAN speeds," in *Proc. NSDI*, 2017, pp. 1–19.

[17] R. Viswanathan, G. Ananthanarayanan, and A. Akella, "CLARINET: WAN-aware optimization for analytics queries," in *Proc. OSDI*, 2016, pp. 1–16.

[18] Q. Pu *et al.*, "Low latency geo-distributed data analytics," in *Proc. SIGCOMM*, 2015, pp. 421–434.

[19] J. Krarup and P. M. Pruzan, "The simple plant location problem: Survey and synthesis," *Eur. J. Oper. Res.*, vol. 12, no. 1, pp. 36–81, Jan. 1983.

[20] M. Charikar, S. Guha, É. Tardos, and D. B. Shmoys, "A constant-factor approximation algorithm for the k-median problem (extended abstract)," in *Proc. STOC*, 1999, pp. 1–10.

[21] S. Guha and S. Khuller, "Greedy strikes back: Improved facility location algorithms," *J. Algorithms*, vol. 31, no. 1, pp. 228–248, 1999.

[22] K. Jain and V. V. Vazirani, "Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and Lagrangian relaxation," *J. ACM*, vol. 48, no. 2, pp. 274–296, Mar. 2001.

[23] D. S. Hochbaum, "Heuristics for the fixed cost median problem," *Math. Program.*, vol. 22, no. 1, pp. 148–162, Dec. 1982.

[24] V. V. Vazirani, *Approximation Algorithms*. Berlin, Germany: Springer, 2001.

[25] U. Feige, "A threshold of ln n for approximating set cover," *J. ACM*, vol. 45, no. 4, pp. 634–652, Jul. 1998.

[26] D. Erlenkotter, "A dual-based procedure for uncapacitated facility location," *Oper. Res.*, vol. 26, no. 6, pp. 992–1009, 1978.

[27] J. E. Beasley, "Lagrangean heuristics for location problems," *Eur. J. Oper. Res.*, vol. 65, no. 3, pp. 383–399, Mar. 1993.

[28] K. S. Al-Sultan and M. A. Al-Fawzan, "A tabu search approach to the uncapacitated facility location problem," *Ann. Oper. Res.*, vol. 86, pp. 91–103, Jan. 1999.

[29] M. Körkel, "On the exact solution of large-scale simple plant location problems," *Eur. J. Oper. Res.*, vol. 39, no. 2, pp. 157–173, Mar. 1989.

[30] D. Tuzun and L. I. Burke, "A two-phase tabu search approach to the location routing problem," *Eur. J. Oper. Res.*, vol. 116, no. 1, pp. 87–99, Jul. 1999.

[31] D. Ghosh, "Neighborhood search heuristics for the uncapacitated facility location problem," *Eur. J. Oper. Res.*, vol. 150, no. 1, pp. 150–162, Oct. 2003.

[32] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "QueryMarket demonstration: Pricing for online data markets," in *Proc. VLDB Endowment*, 2012, pp. 1–4.

[33] (2015). *Visipedia Project*. [Online]. Available: http://www.vision.caltech.edu/visipedia/

[34] J. C. Corbett *et al.*, "Spanner: Google's globally-distributed database," *ACM Trans. Comput. Syst.*, vol. 31, no. 3, Aug. 2013, Art. no. 8.

[35] A. Gupta *et al.*, "Mesa: Geo-replicated, near real-time, scalable data warehousing," in *Proc. VLDB Endowment*, 2014, pp. 1259–1270.

[36] A. Rabkin, M. Arye, S. Sen, V. S. Pai, and M. J. Freedman, "Aggregation and degradation in JetStream: Streaming analytics in the wide area," in *Proc. NSDI*, 2014, pp. 275–288.

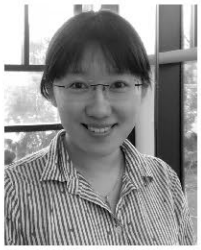
[37] M. Balazinska, B. Howe, and D. Suciu, "Data markets in the cloud: An opportunity for the database community," in *Proc. VLDB Endowment*, 2011, pp. 1–4.

[38] R. Cummings, K. Ligett, A. Roth, Z. S. Wu, and J. Ziani, "Accuracy for sale: Aggregating data with a variance constraint," in *Proc. ITCS*, 2015, pp. 317–324.

[39] R. Tang, H. Wu, Z. Bao, S. Bressan, and P. Valduriez, "The price is right," in *Database and Expert Systems Applications. DEXA* (Lecture Notes in Business Information Processing), vol. 8056, H. Decker, L. Lhotská, S. Link, J. Basl, and A. M. Tjoa, Eds. Berlin, Germany: Springer, 2013.

[40] R. Tang, A. Amarilli, P. Senellart, and S. Bressan, "Get a sample for a discount," in *Database and Expert Systems Applications. DEXA* (Lecture Notes in Computer Science), vol. 8644, H. Decker, L. Lhotská, S. Link, M. Spies, and R. R. Wagner, Eds. Cham, Switzerland: Springer, 2014.

[41] A. Muschalle, F. Stahl, A. Löser, and G. Vossen, "Pricing approaches for data markets," in *Enabling Real-Time Business Intelligence. BIRTE* (Lecture Notes in Business Information Processing), vol. 154, M. Castellanos, U. Dayal, and E. A. Rundensteiner, Eds. Berlin, Germany: Springer, 2012.
[42] R. Stahl and G. Vossen, "Data quality scores for pricing on data marketplaces," in *Intelligent Information and Database Systems. ACIIDS* (Lecture Notes in Computer Science), vol. 9621, N. T. Nguyen, B. Trawiński, H. Fujita, and T. P. Hong, Eds. Berlin, Germany: Springer, 2016.
[43] R. Stahl and G. Vossen, "Name your own price on data marketplaces," *Informatica*, vol. 28, no. 1, pp. 155–180, 2017.
[44] C.-C. Hung, L. Golubchik, and M. Yu, "Scheduling jobs across geo-distributed datacenters," in *Proc. 6th ACM Symp. Cloud Comput.*, 2015, pp. 111–124.
[45] H. Zhang *et al.*, "Live video analytics at scale with approximation and delay-tolerance," in *Proc. 14th USENIX Symp. Netw. Syst. Design Implement. (NSDI)*, Boston, MA, USA, 2017, pp. 377–392. [Online]. Available: https://www.usenix.org/conference/nsdi17/technical-sessions/presentation/zhang
[46] A. Vakali and G. Pallis, "Content delivery networks: status and trends," *IEEE Internet Comput.*, vol. 7, no. 6, pp. 68–74, Nov. 2003, doi: 10.1109/MIC.2003.1250586.
[47] G. Peng. (2004). "CDN: Content distribution network." [Online]. Available: https://arxiv.org/abs/cs/0411069
[48] G. Pallis and A. Vakali, "Insight and perspectives for content delivery networks," *Commun. ACM*, vol. 49, no. 1, pp. 101–106, Jan. 2006, doi: 10.1145/1107458.1107462.
[49] A.-M. K. Pathan and R. Buyya, "A taxonomy and survey of content delivery networks," Grid Comput. Distrib. Syst. Lab., Univ. Melbourne, Parkville, VIC, Australia, Tech. Rep. 4, 2007.
[50] J. D. Guyton and M. F. Schwartz, "Locating nearby copies of replicated Internet servers," in *Proc. Conf. Appl., Technol., Archit., Protocols Comput. Commun. (SIGCOMM)*, New York, NY, USA, 1995, pp. 288–298, doi: 10.1145/217382.217463.
[51] Z.-M. Fei, S. Bhattacharjee, E. W. Zegura, and M. H. Ammar, "A novel server selection technique for improving the response time of a replicated service," in *Proc. 17th Annu. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, vol. 2. Mar./Apr. 1998, pp. 783–791.
[52] S. Jamin *et al.*, "On the placement of Internet instrumentation," in *Proc. IEEE Conf. Comput. Commun. 19th Annu. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, vol. 1. Mar. 2000, pp. 295–304, doi: 10.1109/INFCOM.2000.832199.
[53] P. Krishnan, D. Raz, and Y. Shavitt, "The cache location problem," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 568–582, Oct. 2000, doi: 10.1109/90.879344.
[54] M. Gritter and D. R. Cheriton, "An architecture for content routing support in the Internet," in *Proc. 3rd Conf. USENIX Symp. Internet Technol. Syst. (USITS)*, vol. 3. Berkeley, CA, USA, Mar. 2001, p. 4. [Online]. Available: http://dl.acm.org/citation.cfm?id=1251440.1251444
[55] L. Qiu, V. N. Padmanabhan, and G. M. Voelker, "On the placement of Web server replicas," in *Proc. 20th Annu. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, vol. 3. Apr. 2001, pp. 1587–1596.
[56] J. Kangasharju, J. Roberts, and K. W. Ross, "Object replication strategies in content distribution networks," *Comput. Commun.*, vol. 25, no. 4, pp. 376–383, Mar. 2002, doi: 10.1016/S0140-3664(01)00409-1.
[57] C. Gkantsidis and P. R. Rodriguez, "Network coding for large scale content distribution," in *Proc. IEEE 24th Annu. Joint Conf. IEEE Comput. Commun. Soc.*, vol. 4. Mar. 2005, pp. 2235–2245, doi: 10.1109/INFCOM.2005.1498511.
[58] A. Flavel *et al.*, "FastRoute: A scalable load-aware anycast routing architecture for modern CDNs," in *Proc. 12th USENIX Symp. Netw. Syst. Design Implement. (NSDI)*, Oakland, CA, USA, 2015, pp. 381–394. [Online]. Available: https://www.usenix.org/conference/nsdi15/technical-sessions/presentation/flavel
[59] F. Chen, R. K. Sitaraman, and M. Torres, "End-user mapping: Next generation request routing for content delivery," in *Proc. ACM Conf. Special Interest Group Data Commun. (SIGCOMM)*, New York, NY, USA, 2015, pp. 167–181, doi: 10.1145/2785956.2787500.
[60] C. Dwork, "Differential privacy," in *Encyclopedia of Cryptography and Security*, New York, NY, USA: Springer, 2011, pp. 338–340.
[61] (2015). *Microsoft Azure*. [Online]. Available: https://azure.microsoft.com/en-us/
[62] J. Wiener and N. Bronson. (2014). *Facebook's Top Open Data Problems*. [Online]. Available: https://research.facebook.com/blog/1522692927972019/facebook-s-top-open-data-problems/
[63] G. Lee, J. Lin, C. Liu, A. Lorek, and D. Ryaboy, "The unified logging infrastructure for data analytics at Twitter," in *Proc. VLDB Endowment*, 2012, pp. 1771–1780.
[64] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*, Vol. 6. Belmont, MA, USA: Athena Scientific, 1997.
[65] (2012). *Google Data Center FAQ*. [Online]. Available: http://www.datacenterknowledge.com/archives/2012/05/15/google-data-center-faq/
[66] M. E. J. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemp. Phys.*, vol. 46, no. 5, pp. 323–351, 2005.
[67] M. Balazinska, B. Howe, P. Koutris, D. Suciu, and P. Upadhyaya, "A discussion on pricing relational data," in *Search of Elegance in the Theory and Practice of Computation*, Berlin, Germany: Springer, 2013, pp. 167–173.



Xiaoqi Ren received the B.S. degree in automation from Tsinghua University, China. She is currently pursuing the Ph.D. degree with the Department of Computer Science, California Institute of Technology. Her research focuses on optimization of today's large-scale data centers, including online scheduling, energy usage and sustainability, and new market mechanism designs (electricity market and data market).

Palma London received the double B.S. degree in electrical engineering and math from the University of Washington. She is currently pursuing the Ph.D. degree with the Department of Computing and Mathematical Sciences, California Institute of Technology. Her research interests are generally in convex optimization, distributed algorithms, and machine learning.

Juba Ziani received the B.S. degree from Supelec, France, in 2011, and the M.S. degree in operations research from Columbia University in 2012. He is currently pursuing the Ph.D. degree in computer science with the California Institute of Technology. His primary focus is to study the new challenges posed by the generation of larger and larger amounts of data. In particular, he is interested in mechanism design for exchanging and drawing useful conclusions from data, in the privacy and fairness concerns that come from utilizing private data, and in the effect that data and signaling have on optimal mechanism design.

Adam Wierman is currently a Professor with the Department of Computing and Mathematical Sciences, California Institute of Technology. He has coauthored papers that received best paper awards from the ACM SIGMETRICS, the IEEE INFOCOM, IFIP Performance, the IEEE Green Computing Conference, the IEEE Power and Energy Society General Meeting, and the ACM GREENMETRICS. His research interests center around resource allocation and scheduling decisions in computer systems and services. He received the 2011 ACM SIGMETRICS Rising Star Award and the 2014 IEEE Communications Society William R. Bennett Prize.