Bulletin of the Technical Committee on

Data Engineering

September 2019 Vol. 42 No. 3

IEEE Computer Society

Letters

Opinions

Special Issue on Fairness, Diversity, and Transparency in Data Systems

2019 IEEE TCDE Awards

Conference and Journal Notices

The TC on Data Engineering

Membership in the TC on Data Engineering is open to all current members of the IEEE Computer Society who are interested in database systems. The TCDE web page is http://tab.computer.org/tcde/index.html.

The Data Engineering Bulletin

The Bulletin of the Technical Committee on Data Engineering is published quarterly and is distributed to all TC members. Its scope includes the design, implementation, modelling, theory and application of database systems and their technology.

Letters, conference information, and news should be sent to the Editor-in-Chief. Papers for each issue are solicited by and should be sent to the Associate Editor responsible for the issue.

Opinions expressed in contributions are those of the authors and do not necessarily reflect the positions of the TC on Data Engineering, the IEEE Computer Society, or the authors' organizations.

The Data Engineering Bulletin web site is at http://tab.computer.org/tcde/bull_about.html.

Letter from the Editor-in-Chief

Letter from Haixun

Haixun Wang
WeWork Corporation

Letter from the Special Issue Editor

The big data revolution and advancements in machine learning technologies have revolutionized decision making, advertising, medicine, and even election campaigns. Data-driven software now permeates virtually every aspect of human activity and has the ability to shape human behavior: it affects the products we view and purchase, the news articles we read, the social interactions we engage in, and, ultimately, the opinions we form. Yet, data is an imperfect medium, tainted by errors, omissions, and biases. As a result, discrimination shows up in many data-driven applications, such as advertisements, hotel bookings, image search, and vendor services. In this issue, we bring together an exciting collection of recent and ongoing work that focuses on the problems of fairness, diversity, and transparency in data-driven systems. This collection highlights the central role that the data management research community can play in detecting, informing, and mitigating the effects of bias, skew, and misuse of data, and aims to create bridges with work in related communities.

We start with "Nutritional Labels for Data and Models", by Stoyanovich and Howe. This paper argues for informational and warning labels for data, akeen to nutritional labels, that specify characteristics of data and how it should be consumed. These nutritional labels help humans determine the fitness of models and data, aiding the interpretability and transparency of decision-making processes.

The second paper, "Data Management for Causal Algorithmic Fairness", by Salimi, Howe, and Suciu, provides a brief overview of fairness definitions in the literature, and argues for the use of causal reasoning in defining and reasoning about fairness. The paper exposes a vision of the opportunities of applying data management techniques, such as integrity constraints, query rewriting, and database repair to enforcing fairness, detecting discrimination, and explaining bias.

In the third paper, "A Declarative Approach to Fairness in Relational Domains", Farnadi, Babaki, and Getoor focus on notions of fairness that capture the relational structure of a domain, and propose a general framework for relational fairness. Fairness-aware probabilistic soft logic includes a language for specifying discrimination patterns, and an algorithm for performing inference under fairness constraints.

The next paper, "Fairness in Practice: A Survey on Equity in Urban Mobility", by Yan and Howe, places its focus on practical societal implications of fairness in the domain of transportation. The paper presents the findings of equity studies in mobility systems, such as bike-sharing and ride-hailing systems, and reviews experimental methods and metrics.

Again motivated by the societal implications of fairness and diversity, Benabbou, Chakraborty, and Zick put their sights on the allocation of public resources. "Fairness and Diversity in Public Resource Allocation Problems" focuses on two real-world cases, the allocation of public housing in Singapore and public school admissions in Chicago, models them as constrained optimization problems, and analyzes the welfare loss in enforcing diversity.

We conclude with "Towards Responsible Data-driven Decision Making in Score-Based Systems", by Asudeh, Jagadish, and Stoyanovich. The paper focuses on designing fair and stable rankings, and discusses how these technologies can assess and enhance the coverage of training sets in machine learning tasks.

Thank you to all the authors for their insightful contributions, which bring into focus new and exciting challenges, and identify opportunities for data management research to contribute tools and solutions towards critical societal issues. Thank you also to Haixun Wang for his valuable assistance in putting together the issue. I hope you enjoy this collection.

Alexandra Meliou
University of Massachusetts, Amherst

Letter from the TCDE Awards Committee

The IEEE TCDE (Technical Committee of Data Engineering) has established several highly prestigious awards to encourage innovative long term contributions to the field of data engineering. It is our pleasure to present letters from the 2019 award winners in this issue.

Rising Star Award. The IEEE TCDE Rising Star Award is based on an individual's whole body of work in the first five years after the PhD. The award aims to promote current database researchers as they create their career. The 2019 IEEE TCDE Rising Star Award goes to Viktor Leis from the Technical University of Munich for *contributions to main-memory indexing and database architectures for NVM*.

Impact Award. The IEEE TCDE Impact Award recognizes database researchers whose research resulted in impact beyond the data engineering field, impact beyond research to industrial practice, and/or impact resulting in expansion of the data engineering field itself. The 2019 IEEE TCDE Impact Award goes to Christian Jensen from Aalborg University for *contributions to spatial, temporal, and spatio-temporal data management*.

Service Award. The IEEE TCDE Service Award recognizes an individual who has contributed significantly to ICDE, TCDE, and the data engineering community in general. The 2019 IEEE TCDE Service Award goes to David Lomet from Microsoft for *leadership as the Editor-in-Chief of the Data Engineering Bulletin for over 25 years*.

Congratulations again to the winners, and we hope you will enjoy reading their letters as much as we did.

The 2019 Awards Committee
Anastasia Ailamaki
Paolo Atzeni
Michael Carey
Xin Luna Dong
Johannes Gehrke (chair)
Sunita Sarawagi

Johannes Gehrke
Microsoft, USA

Value Creation from Massive Data in Transportation – The Case of Vehicle Routing

Christian S. Jensen
Aalborg University, Denmark

1 Introduction

Vehicular transportation will undergo profound change over the next decades, due to developments such as increasing mobility demands and increasingly autonomous driving. At the same time, rapidly increasing, massive volumes of data that capture the movements of vehicles are becoming available. In this setting, the current vehicle routing paradigm falls short, and we need new data-intensive paradigms. In a data-rich setting, travel costs such as travel time are modeled as time-varying distributions: at a single point in time, the time needed to traverse a road segment is given by a distribution. How can we best build, maintain, and use such distributions?

The travel cost of a route is obtained by convolving distributions that model the costs of the segments that make up the route. This process is expensive and yields inaccurate results when dependencies exist among the distributions. To avoid these problems, we need a path-centric paradigm, where costs are associated with arbitrary paths in a road network graph, not just with edges. This paradigm thrives on data: more data is expected to improve accuracy, but also efficiency. Next, massive trajectory data makes it possible to compute different travel costs in different contexts, e.g., for different drivers, by using different subsets of trajectories depending on the context. It is then no longer appropriate to assume that costs are available when routing starts; rather, we need an on-the-fly paradigm, where costs can be computed during routing. Key challenges include how to achieve efficiency and accuracy with sparse data. Finally, the above paradigms assume that the benefit, or cost, of a path is quantified. As an alternative, we envision a cost-oblivious paradigm, where the objective is to return routes that match the preferences of local, or expert, drivers without formalizing costs.

2 Background

Vehicular transportation is an inherent aspect of society and our lives: many people rely on vehicular transportation on a daily basis, we spend substantial time on transportation, and we are often forced to arrange our lives around traffic. As a reflection of this, society spends very substantial resources on enabling safe, reliable, clean, and inexpensive transportation. Due to a combination of interrelated developments, transportation will undergo profound changes in the years to come.

First, a range of key enabling technologies have reached levels of sophistication that make (semi-)autonomous vehicles possible. For example, Tesla cars already come with an autopilot that is a pre-cursor to autonomous driving, and virtually all major vehicle manufacturers are working to make autonomous cars. The state of affairs is similar to the one that applied to personal computing when Apple and Microsoft were created and the one that applied to the Internet when Google was founded. Second, the sharing economy trend is also gaining traction in relation to vehicular transportation, thus enabling better exploitation of under-utilized vehicles. For example, Uber enables transportation in private vehicles by private drivers. Online ridesharing services such as Lyft enable the sharing of trips. A large number of similar services exist across the globe. Next, other developments such as urbanization and the needs to combat air pollution and greenhouse gas emissions will also impact transportation. Many large cities are facing air quality problems, and the transportation sector is the second largest contributor to GHG emissions, trailing only the energy sector.

These increasingly pressing developments promise a perfect storm for transportation: While it is not clear exactly how this will play out, it is clear that transportation faces profound change. For example, Uber and similar services may eventually do away with under-paid drivers. When a person goes to a movie theater and cannot

find parking, the driver may instead let the car serve as a self-driving taxi, thus making money instead of paying money for parking while watching a movie.

We are also witnessing a digitalization trend that is unprecedented in the history of humanity: We are increasingly instrumenting societal and industrial processes with networked sensors. As a result, we are accumulating massive volumes of data that capture the states of processes and that may be used for enabling rational, data-driven processes and data-driven decision making. This also applies to transportation. Vehicles are increasingly online, via smartphones or built-in connectivity, and they are equipped with global navigation satellite system (GNSS) positioning capabilities, e.g., Galileo, GPS, and Glonass, via smartphones or in-vehicle navigation systems. As a result, rapidly increasing volumes of vehicle data are becoming available. This data includes vehicle trajectory data, i.e., sequences of GNSS records that record time and location. This new data source captures transportation at a level of detail never seen before.

With the diffusion of smartphones and in-vehicle navigation devices, routing is now available to a very large fraction of the population on Earth. Indeed, the availability of routing is now taken for granted, and routing is used widely. Further, the advances in autonomous and semi-autonomous vehicles make it a safe bet that more and more routing decisions will be taken by machines using some form of routing service, rather than by people. Thus, the importance of routing will increase over the coming years.

The foundation for traditional routing was built at a time where little data was available. We contend that given the above observations, new foundations are needed to enable routing capable of effectively exploiting available data to enable efficient and accurate, high-resolution routing services.

3 New Routing Paradigms

Traditional Routing The setting that underlies traditional routing services is one where a road network is modeled as a weighted graph and where the weight of an edge captures the cost of traversing the road segment modeled by the edge. In this setting, a graph with real-valued edge weights, capturing, e.g., travel distance, is given and some routing algorithm is applied to identify a route from a source to a destination with the minimum sum of edge weights. More advanced edge weights that capture travel time are also considered. While many different routing algorithms exist for such weighted road-network graphs, the prototypical algorithm is Dijkstra's algorithm [1]; hence, we call this Dijkstra's paradigm. This paradigm is well suited for settings were little travel data is available. Notably, by assigning weights to the atomic paths, i.e., individual graph edges, the paradigm makes the best possible use of available data. However, we contend that this simple edge-centric paradigm is obsolete and hinders progress in settings were travel costs are extracted from trajectories. Dijkstra's paradigm falls short when it comes to exploiting massive volumes of trajectory data for enabling more accurate and higher-resolution routing.

Given a (source, destination)-pair and a departure time, a typical routing service computes one or more paths from the source to the destination with the fastest travel time as of the departure time. "High resolution" implies that travel times in a road network are modeled (i) at a fine temporal granularity, as traffic changes continuously and affects travel time, and (ii) as distributions, as different drivers may have different travel times even when driving on the same path at the same time, and as traffic is inherently unpredictable. Further high resolution implies that routing takes into account the particular context, e.g., the driver, yielding personalized routing, or weather conditions [2, 3, 4].

We envision three new routing paradigms that are capable of exploiting massive trajectory data to enable more accurate and higher-resolution routing services.

Path-centric paradigm In this paradigm, costs are associated with arbitrary paths in a road network graph, rather than just with edges. This avoids unnecessary fragmentation of trajectories and automatically enables detailed capture of dependencies as well as turning and waiting times at intersections. This paradigm thrives

on data: the more trajectory data, the better the accuracy and resolution of the routing. Further, more data also promises more efficient routing, which is less intuitive. With this paradigm, the cost, e.g., travel time, of an arbitrary path is estimated from available costs of paths that intersect the path. Fewer costs have to be assembled than in the edge-centric paradigm. For example, with costs being probability distributions and a path containing 100 edges, convolution must be applied 99 times to assemble 100 distributions into one in Dijkstra's paradigm. With sufficient trajectory data, a path may be covered by a few long paths with costs in the path-centric paradigm. Thus, computing the path's cost will require only a few convolutions. Thus, this paradigm holds the potential to enable more efficient routing the more trajectory data that is available. In the extreme, computing the cost of an arbitrary path can be achieved by means of a lookup, with no need for convolution. Next, when using Dijkstra's algorithm, intuitively, when a search has reached a graph vertex, the lowest-cost path to reach that vertex is known and fixed; thus, all other paths for reaching the vertex can be disregarded, or pruned. In the new paradigm, the cost of reaching a vertex can change when the search proceeds from the vertex because a different set of path costs that reach into the past may be used. It may happen that the cost of the path used for reaching the vertex increases and that a lower-cost path now exists.

In the path centric-paradigm, the underlying data structure is no longer just a graph, as path weights need to be maintained, and the correctness of Dijkstra's algorithm is no longer guaranteed. In initial work [5, 6], we have taken first steps to define and explore some aspects of the path-centric paradigm. These studies confirm that the paradigm holds substantial promise and is "the right" paradigm when massive trajectory data is available.

On-the-fly paradigm Next, massive trajectory data makes it possible to compute different travel costs in different contexts, e.g., for different drivers, by using different subsets of trajectories depending on the context. In this setting, it is no longer appropriate to assume that precomputed costs are available when routing starts, which is the standard assumption. There are simply too many costs to compute and store, most of which will never be used. Instead, we need an on-the-fly paradigm, where costs can be computed during routing. When, during routing, we need to determine the cost distribution of an edge or a path, we need to retrieve the relevant parts of the available trajectories that contain useful cost information given the particular context considered. These parts are then used to form an accurate cost distribution. The retrieval task takes a path, the time-of-arrival at the path, and contextual information such as a user identifier and weather information as arguments. Then the task is to retrieve sub-trajectories that contain information relevant to these arguments. As a routing query should preferably take less than 100 milliseconds, it is very difficult to achieve the necessary efficiency, and indexing techniques are needed that go beyond existing techniques [7, 8, 9]. Another challenge is to determine which trajectories to actually use when computing the most accurate weight distributions. We have conducted preliminary studies focused on achieving better indexing [10] and understanding the accuracy problem [11, 12]. The studies indicate that the challenges are substantial.

Cost-oblivious paradigm The above paradigms rely on the same underlying assumption as does Dijkstra's paradigm: We use trajectory data for computing costs, and then we apply a routing algorithm to find lowest-cost paths. In essence, these paradigms only use trajectories for extracting costs such as travel time and GHG emissions [13]. However, trajectories contain much more information that could potentially be utilized for achieving better routing: Trajectories tell which routes drivers follow and seemingly prefer. This paradigm is behavioral in the sense that it aims to exploit this route-choice behavior. An earlier study [14] indicates that historical trajectories are better at predicting the route a driver will take from a source to a destination than is the route returned by a cost-based routing service. This study thus confirms that the cost-oblivious paradigm holds potential for enabling better routing. And again, this is a paradigm that is shaped to thrive on data: If enough data is available to cover all (source, destination)-pairs with trajectories, routing could be achieved by means of a lookup, with no need for a travel-cost based routing algorithm. We have already proposed a simple route-recommendation solution and have compared it with existing solutions [15]. These solutions do not contend well with sparse data. In addition,

we have proposed a first attempt at making better use of sparse data [16] for path recommendation within this paradigm.

Synergies It is important to observe that specific routing solutions can be composed of elements from Dijkstra's paradigm and all three new paradigms. For example, a predominantly on-the-fly solution may rely on pre-computed edge weights as a fall-back; and if insufficient data is available to a cost-oblivious solution, some limited form of routing may be applied. Beyond this, the fleshing out of the three paradigms relies on the same experimental infrastructure, encompassing computing capabilities, software pipelines, data, and methodologies.

4 Summary

In a world with more than 2.5 billion smartphone users and about 1 billion cars, and where routing decisions are increasingly being made by machines, the line of research outlined here has the potential for very large societal impact. It literally holds the potential to make a difference for on the order of a billion users. High-quality routing has significant benefits. It can make transportation more predictable, an important property of a transportation system that reduces the need to "leave early" and thus the time spent on transportation. In addition, it may increase the capacity of an existing infrastructure by making each trip more efficient, making room for more trips, and by incentivizing drivers to "spread out" their trips, e.g., by quantifying the time saved by traveling before or after rush hour. Routing also holds the potential to reduce the GHG emissions per trip [17, 18]. Finally, the above coverage of problems related to the use of massive trajectory data for value creation in transportation is by no means exhaustive.

Acknowledgments I would like to thank the many hard-working colleagues with whom I have worked and am working to make progress on the topics described here.

References

[1] E. W. Dijkstra. *A note on two problems in connexion with graphs*. Numer. Math., vol. 1, no. 1, pp. 269–271, 1959.

[2] J. Letchner, J. Krumm and E. Horvitz. *Trip Router with Individualized Preferences (TRIP): Incorporating Personalization into Route Planning*. In AAAI, 2006.

[3] B. Yang, C. Guo, Y. Ma and C. S. Jensen. *Toward personalized, context-aware routing*. VLDB J, vol. 24, no. 2, pp. 297–318, 2015.

[4] O. Andersen and K. Torp. *A Data Model for Determining Weather's Impact on Travel Time*. In DEXA, 2016.

[5] J. Dai, B. Yang, C. Guo, C. S. Jensen and J. Hu. *Path Cost Distribution Estimation Using Trajectory Data*. PVLDB, vol. 10, no. 3, pp. 85–96, 2016.

[6] B. Yang, J. Dai, C. Guo, C. S. Jensen and J. Hu. *PACE: a PAth-CEntric paradigm for stochastic path finding*. VLDB J, vol. 27, no. 2, pp. 153–178, 2018.

[7] B. B. Krogh, N. Pelekis, Y. Theodoridis and K. Torp. *Path-based queries on trajectory data*. In SIGSPATIAL GIS, 2014.

[8] B. B. Krogh, C. S. Jensen and K. Torp. *Efficient in-memory indexing of network-constrained trajectories*. In SIGSPATIAL GIS, 2016.

[9] S. Koide, Y. Tadokoro, C. Xiao and Y. Ishikawa. *CiNCT: Compression and retrieval of massive vehicular trajectories via relative movement labeling*. In ICDE, 2018.

[10] R. Waury, C. S. Jensen, S. Koide, Y. Ishikawa, and C. Xiao. *Indexing Trajectories for Travel-Time Histogram Retrieval*. In EDBT 2019.

[11] R. Waury, J. Hu, B. Yang and C. S. Jensen. *Assessing the Accuracy Benefits of On-the-Fly Trajectory Selection in Fine-Grained Travel-Time Estimation*. In MDM, 2017.

[12] R. Waury, C. S. Jensen and K. Torp. *Adaptive Travel-Time Estimation: A Case for Custom Predicate Selection*. In MDM, 2018.

[13] C. Guo, B. Yang, O. Andersen, C. S. Jensen and K. Torp. *EcoMark 2.0: empowering eco-routing with vehicular environmental models and actual vehicle fuel consumption data* Geoinformatica, vol. 19, no. 3, pp. 567–599, 2015.

[14] V. Ceikute and C. S. Jensen. *Routing Service Quality - Local Driver Behavior Versus Routing Services*. In MDM, 2013.

[15] V. Ceikute and C. S. Jensen. *Vehicle Routing with User-Generated Trajectory Data* In MDM, 2015.

[16] C. Guo, B. Yang, J. Hu and C. S. Jensen. *Learning to Route with Sparse Trajectory Sets*. In ICDE, 2018.

[17] O. Andersen, C. S. Jensen, K. Torp and B. Yang. *EcoTour: Reducing the Environmental Footprint of Vehicles Using Eco-routes*. In MDM, 2013.

[18] C. Guo, B. Yang, O. Andersen, C. S. Jensen and K. Torp. *EcoSky: Reducing vehicular environmental impact through eco-routing*. In ICDE, 2015.

Nutritional Labels for Data and Models *

Julia Stoyanovich
New York University
New York, NY, USA
stoyanovich@nyu.edu

Bill Howe
University of Washington
Seattle, WA, USA
billhowe@uw.edu

Abstract

An essential ingredient of successful machine-assisted decision-making, particularly in high-stakes decisions, is interpretability — allowing humans to understand, trust and, if necessary, contest, the computational process and its outcomes. These decision-making processes are typically complex: carried out in multiple steps, employing models with many hidden assumptions, and relying on datasets that are often used outside of the original context for which they were intended. In response, humans need to be able to determine the "fitness for use" of a given model or dataset, and to assess the methodology that was used to produce it.

To address this need, we propose to develop interpretability and transparency tools based on the concept of a nutritional label, drawing an analogy to the food industry, where simple, standard labels convey information about the ingredients and production processes. Nutritional labels are derived automatically or semi-automatically as part of the complex process that gave rise to the data or model they describe, embodying the paradigm of interpretability-by-design. In this paper we further motivate nutritional labels, describe our instantiation of this paradigm for algorithmic rankers, and give a vision for developing nutritional labels that are appropriate for different contexts and stakeholders.

1 Introduction

An essential ingredient of successful machine-assisted decision-making, particularly in high-stakes decisions, is interpretability — allowing humans to understand, trust and, if necessary, contest, the computational process and its outcomes. These decision-making processes are typically complex: carried out in multiple steps, employing models with many hidden assumptions, and relying on datasets that are often repurposed — used outside of the original context for which they were intended.[1] In response, humans need to be able to determine the "fitness for use" of a given model or dataset, and to assess the methodology that was used to produce it.

To address this need, we propose to develop interpretability and transparency tools based on the concept of a *nutritional label*, drawing an analogy to the food industry, where simple, standard labels convey information about the ingredients and production processes. Short of setting up a chemistry lab, the consumer would otherwise

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

*This work was supported in part by NSF Grants No. 1926250, 1916647, and 1740996.

[1] See Section 1.4 of Salganik's "Bit by Bit" [24] for a discussion of data repurposing in the Digital Age, which he aptly describes as "mixing readymades with custommades."

have no access to this information. Similarly, consumers of data products cannot be expected to reproduce the computational procedures just to understand fitness for their use. Nutritional labels, in contrast, are designed to support specific decisions by the consumer rather than completeness of information. A number of proposals for hand-designed nutritional labels for data, methods, or both have been suggested in the literature[9, 12, 17]; we advocate deriving such labels automatically or semi-automatically as a side effect of the computational process itself, embodying the paradigm of *interpretability-by-design*.

Interpretability means different things to different stakeholders, including individuals being affected by decisions, individuals making decisions with the help of machines, policy makers, regulators, auditors, vendors, data scientists who develop and deploy the systems, and members of the general public. Designers of nutritional labels must therefore consider *what* they are explaining, *to whom*, and *for what purpose*. In the remainder of this section, we will briefly describe two regulatory frameworks that mandate interpretability of data collection and processing to members of the general public, auditors, and regulators, where nutritional labels offer a compelling solution (Section 1.1). We then discuss interpretability requirements in data sharing, particularly when data is altered to protect privacy or mitigate bias (Section 1.2).

1.1 Regulatory Requirements for Interpretability

The European Union recently enacted a sweeping regulatory framework known as the General Data Protection Regulation, or the GDPR [30]. The regulation was adopted in April 2016, and became enforceable about two years later, on May 25, 2018. The GDPR aims to protect the rights and freedoms of natural persons with regard to how their personal data is processed, moved, and exchanged (Article 1). The GDPR is broad in scope, and applies to "the processing of personal data wholly or partly by automated means" (Article 2), both in the private sector and in the public sector. Personal data is broadly construed, and refers to any information relating to an identified or identifiable natural person, called the *data subject* (Article 4).

According to Article 4, lawful processing of data is predicated on the data subject's *informed consent*, stating whether their personal data can be used, and for what purpose (Articles 6, 7). Further, data subjects have *the right to be informed* about the collection and use of their data. [2] Providing insight to data subjects about the collection and use of their data requires technical methods that support interpretability.

Regulatory frameworks that mandate interpretability are also starting to emerge in the US. New York City was the first US municipality to pass a law (Local Law 49 of 2018) [32], requiring that a task force be put in place to survey the current use of "automated decision systems" (ADS) in city agencies. ADS are defined as "computerized implementations of algorithms, including those derived from machine learning or other data processing or artificial intelligence techniques, which are used to make or assist in making decisions." The task force is developing recommendations for enacting algorithmic transparency by the agencies, and will propose procedures for: (i) requesting and receiving an explanation of an algorithmic decision affecting an individual (Section 3 (b) of Local Law 49); (ii) interrogating ADS for bias and discrimination against members of legally protected groups, and addressing instances in which a person is harmed based on membership in such groups (Sections 3 (c) and (d)); (iii) and assessing how ADS function and are used, and archiving the systems together with the data they use (Sections 3 (e) and (f)).

Other government entities in the US are following suit. Vermont is convening an Artificial Intelligence Task Force to "... make recommendations on the responsible growth of Vermont's emerging technology markets, the use of artificial intelligence in State government, and State regulation of the artificial intelligence field." [33]. Idaho's legislature has passed a law that eliminates trade secret protections for algorithmic systems used in criminal justice [31]. In early April 2019, Senators Booker and Wyden introduced the Algorithmic Accountability Act of 2019 to the US Congress [6]. The Act, if passed, would use "automated decision systems impact assessment" to address and remedy harms caused by algorithmic systems to federally protected classes of people. The act

[2] https://gdpr-info.eu/issues/right-to-be-informed/

empowers the Federal Trade Commission to issue regulations requiring larger companies to conduct impact assessments of their algorithmic systems.

The use of nutritional labels in response to these and similar regulatory requirements can benefit a variety of stakeholders. The designer of a data-driven algorithmic method may use them to validate assumptions, check legal compliance, and tune parameters. Government agencies may exchange labels to coordinate service delivery, for example when working to address the opioid epidemic, where at least three sectors must coordinate: health care, criminal justice, and emergency housing, implying a global optimization problem to assign resources to patients effectively, fairly and transparently. The general public may review labels to hold agencies accountable to their commitment to equitable resource distribution.

1.2 Interpretability with Semi-synthetic Data

A central issue in machine-assisted decision-making is its reliance on historical data, which often embeds results of historical discrimination, also known as *structural bias*. As we have seen time and time again, models trained on data will appear to work well, but will silently and dangerously reinforce discrimination [1, 7, 13]. Worse yet, these models will legitimize the bias — "the computer said so." Nutritional labels for data and models are designed specifically to mitigate the harms implied by these scenarios, in contrast to the more general concept of "data about data."

Good datasets drive research: they inform new methods, focus attention on important problems, promote a culture of reproducibility, and facilitate communication across discipline boundaries. But research-ready datasets are scarce due to the high potential for misuse. Researchers, analysts, and practitioners therefore too often find themselves compelled to use the data they have on hand rather than the data they would (or should) like to use. For example, aggregate usage patterns of ride hailing services may overestimate demand in early-adopter (i.e., wealthy) neighborhoods, creating a feedback loop that reduces service in poorer neighborhoods, which in turn reduces usage. In this example, and in many others, there is a need to alter the input dataset to achieve specific properties in the output, while preserving all other relevant properties. We refer to such altered datasets as *semi-synthetic*.

Recent examples of methods that produce semi-synthetic data include database repair for causal fairness [25], database augmentation for coverage enhancement [4], and privacy-preserving and bias-correcting data release [21, 23]. A semi-synthetic datasets may be altered in different ways. Noise may be added to it to protect privacy, or statistical bias may be removed or deliberately introduced. When a dataset of this kind is released, its composition and the process by which it was derived must be made interpretable to a data scientist, helping determine fitness for use. For example, datasets repaired for racial bias are unsuitable for studying discrimination mitigation methods, while datasets with bias deliberately introduced are less appropriate for research unrelated to fairness. This gives another compelling use case for nutritional labels.

2 Nutritional Labels for Algorithmic Rankers

To make our discussion more concrete, we now describe Ranking Facts, a system that automatically derives nutritional labels for rankings [36]. Algorithmic decisions often result in scoring and ranking individuals — to determine credit worthiness, desirability for college admissions and employment, and compatibility as dating partners. Algorithmic rankers take a collection of items as input and produce a ranking – a sorted list of items – as output. The simplest kind of a ranker is a score-based ranker, which computes a score for each item independently, and then sorts the items on their scores. While automatic and seemingly objective, rankers can discriminate against individuals and protected groups [5], and exhibit low diversity at top ranks [27]. Furthermore, ranked results are often unstable — small changes in the input or in the ranking methodology may lead to drastic changes in the output, making the result uninformative and easy to manipulate [11]. Similar concerns apply in cases where

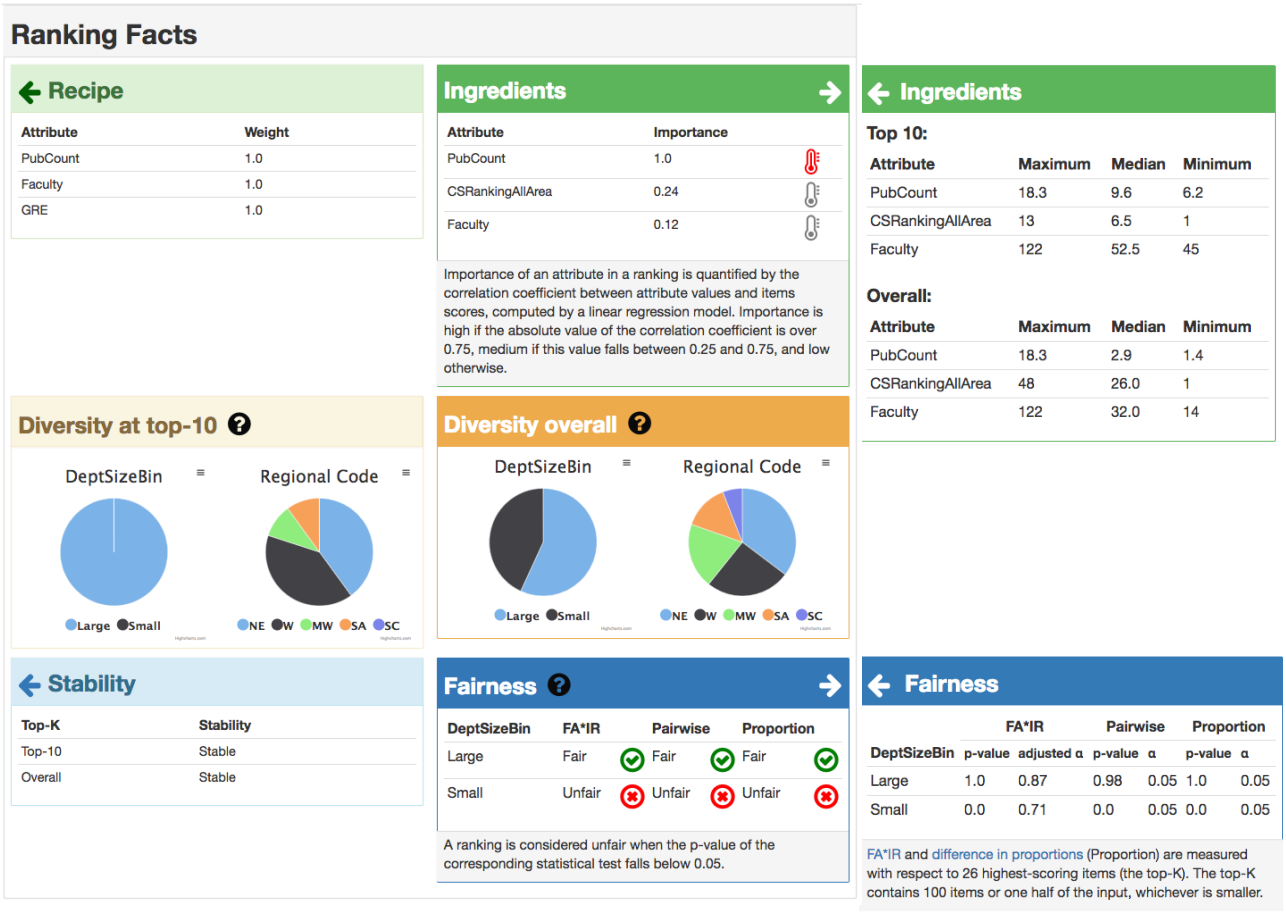


Figure 1: Ranking Facts for the CS departments dataset. The Ingredients widget (green) has been expanded to show the details of the attributes that strongly influence the ranking. The Fairness widget (blue) has been expanded to show the computation that produced the fair/unfair labels.

items other than individuals are ranked, including colleges, academic departments, and products.

In a recent work, we developed Ranking Facts, a nutritional label for rankings [36]. Ranking Facts is available as a Web-based tool[3], and its code is available in the open source [4]. Figure 1 presents Ranking Facts that explains a ranking of Computer Science departments. The data in this example was obtained from CS Rankings[5], augmented with attributes from the NRC dataset [6]. Ranking Facts is made up of a collection of visual widgets, each with an overview and a detailed view. Each widget addresses an essential aspect of transparency and interpretability, and is based on our recent technical work on fairness [3, 35], diversity [8, 27, 28, 34], and stability [2] in algorithmic rankers. We now describe each widget in some detail.

2.1 Recipe and Ingredients

These two widgets help to explain the ranking methodology. The Recipe widget succinctly describes the ranking algorithm. For example, for a linear scoring formula, each attribute would be listed together with its weight. The

[3] http://demo.dataresponsibly.com/rankingfacts/
[4] https://github.com/DataResponsibly/RankingFacts
[5] https://github.com/emeryberger/CSRankings
[6] http://www.nap.edu/rdp/

Ingredients widget lists attributes most material to the ranked outcome, in order of importance. For example, for a linear model, this list could present the attributes with the highest learned weights. Put another way, the explicit intentions of the designer of the scoring function about which attributes matter, and to what extent, are stated in the Recipe, while Ingredients may show attributes that are actually associated with high rank. Such associations can be derived with linear models or with other methods, such as rank-aware similarity in our prior work [27]. The detailed Recipe and Ingredients widgets list statistics of the attributes in the Recipe and in the Ingredients: minimum, maximum and median values at the top-10 and over-all.

2.2 Stability

The Stability widget explains whether the ranking methodology is robust on this particular dataset. An unstable ranking is one where slight changes to the data (e.g., due to uncertainty and noise), or to the methodology (e.g., by slightly adjusting the weights in a score-based ranker) could lead to a significant change in the output. This widget reports a stability score, as a single number that indicates the extent of the change required for the ranking to change. As with the widgets above, there is a detailed Stability widget to complement the overview widget.

An example is shown in Figure 2, where the stability of the ranking is quantified as the slope of the line that is fit to the score distribution, at the top-10 and over-all. A score distribution is unstable if scores of items in adjacent ranks are close to each other, and so a very small change in scores will lead to a change in the ranking. In this example the score distribution is considered unstable if the slope is 0.25 or lower. Alternatively, stability can be computed with respect to each scoring attribute, or it can be assessed using a model of uncertainty in the data. In these cases, stability quantifies the extent to which a ranked list will change as a result of small changes to the underlying data. A complementary notion of stability quantifies the magnitude of change as a result of small changes to the ranking model. We explored this notion in our recent work, briefly discussed below.

In [2] we develped methods for quantifying the stability of a score-based ranker with respect to a given dataset. Specifically, we considered rankers that specify non-negative weights, one for each item attribute, and compute the score as a weighted sum of attribute values. We focused on a notion of stability that quantifies whether the output ranking will change due to a small change in the attribute weights. This notion of stability is natural for consumers of a ranked list (i.e., those who use the ranking to prioritize items and make decisions), who should be able to assess the magnitude of the *region in the weight space* that produces the observed ranking. If this region is large, then the same ranked order would be obtained for many choices of weights, and the ranking is stable. But if this region is small, then we know that only a few weight choices can produce the observed ranking. This may suggest that the ranking was engineered or "cherry-picked" by the producer to obtain a specific outcome.

2.3 Fairness

The Fairness widget quantifies whether the ranked output exhibits statistical parity (one interpretation of fairness) with respect to one or more sensitive attributes, such as gender or race of individuals [35]. We denote one or several values of the sensitive attribute as a protected feature. For example, for the sensitive attribute gender, the assignment gender=F is a protected feature.

A variety of fairness measures have been proposed in the literature [38], with a primary focus on classification or risk assessment tasks. One typical fairness measure for classification compares the proportion of members of a protected group (e.g., female gender or minority race) who receive a positive outcome to their proportion in the overall population. For example, if the dataset contains an equal number of men and women, then among the individuals invited for a job interview, one half should be women. A measure of this kind can be adapted to rankings by quantifying the proportion of members of a protected group in some selected set of size k (treating the top-k as a set).

In [35], we were the first to propose a family of *fairness measures specifically for rankings*. Our measures are based on a generative process for rankings that meet a particular fairness criterion (fairness probability f) and

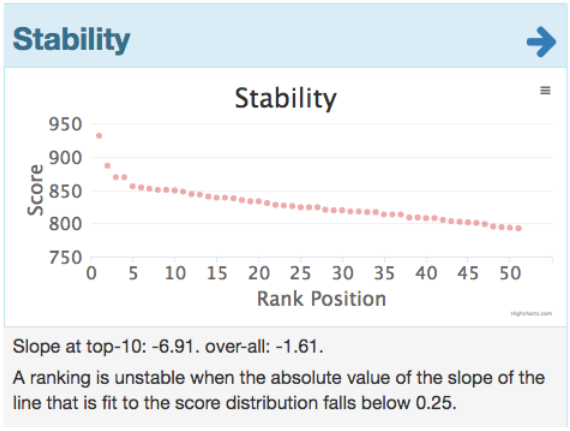


Figure 2: Stability: detailed widget.

are drawn from a dataset with a given proportion of members of a binary protected group (p). This method was subsequently used in FA*IR [37] to quantify fairness in every prefix of a top-k list. We also developed a pairwise measure that directly models the probability that a member of a protected group is preferred to a member of the non-protected group.

Let us now return to the Fairness widget in Figure 1. We select a binary version of the department size attribute DeptSizeBin from the CS departments dataset as the sensitive attribute, and treat the value and "small" as the protected feature. The summary view of the Fairness widget in our example presents the output of three fairness measures: FA*IR [37], proportion [38], and our own pairwise measure. All these measures are statistical tests, and whether a result is fair is determined by the computed p-value. The detailed Fairness widget provides additional information about the tests and explains the process.

2.4 Diversity

Fairness is related to diversity: ensuring that different kinds of objects are represented in the output of an algorithmic process [8]. Diversity has been considered in search and recommender systems, but in a narrow context, and was rarely applied to profiles of individuals. The Diversity widget shows diversity with respect to a set of demographic categories of individuals, or a set of categorical attributes of other kinds of items [8]. The widget displays the proportion of each category in the top-10 ranked list and over-all, and, like other widgets, is updated as the user selects different ranking methods or sets different weights. In our example in Figure 1, we quantify diversity with respect to department size and to the regional code of the university. By comparing the pie charts for top-10 and over-all, we observe that only large departments are present in the top-10.

This simple diversity measure that is currently included in Ranking Facts can be augmented by, or replaced with, other measures, including, for example, those we developed in our recent work [28, 34].

3 Learning Labels

The creation of nutritional labels is often cast as a design problem rather than a computational problem [9, 12]. Standard labels with broad applicability can amortize the cost of design, but the diversity of datasets, methods, and desirable properties for nutritional labels suggest a learning approach to help develop labels for a variety of situations. Since opaque automation is what motivated the need for labels in the first place, automating their creation may seem like a step backwards. But there are several benefits:

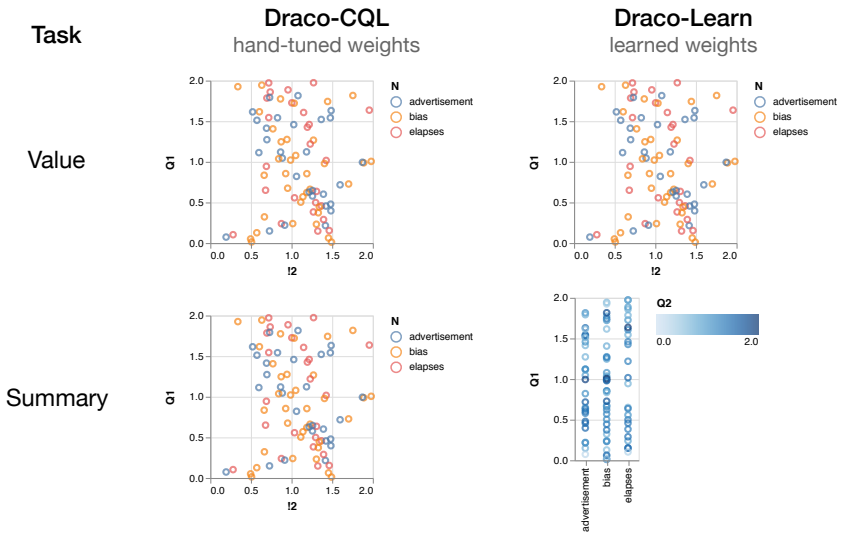


Figure 3: Draco can be used to re-implement existing visualization systems like CQL by hand-tuning weights (left) or be used to learn weights automatically from preference data (right). The visualizations selected can vary significantly, affording customization for specific applications. A similar approach can be used when generating nutritional labels for data and models.

- Coverage: *some* information provided in (nearly) *all* cases is preferable to *all* information provided in *some* cases, as there are many models and datasets being deployed.
- Correctness: Hand-designed labels imply human metadata attachment, but curation of metadata is essentially an unsolved problem. Computable labels reduce reliance on human curation efforts.
- Retroactivity: Some information can only be manually collected at the time of data collection (e.g., demographics of authors in a speech corpus to control for nationality bias). This opportunity is lost for existing datasets. However, inferring relevant properties based on the content of the data may be "better than nothing."

We now consider two approaches to the problem of learning labels, one based on the visualization recommendation literature, and one based on bin-packing optimization.

3.1 Learning as Visualization Recommendation

Moritz et al. proposed Draco [19], a formal model that represents visualizations as sets of logical facts, and represents design guidelines as a collection of hard and soft constraints over these facts, following an earlier proposal for the VizDeck system [14]. Draco enumerates the visualizations that do not violate the hard constraints and finds the most preferred visualizations according to the weights of the soft constraints. Formalized visualization descriptions are derived from the Vega-Lite grammar [26] extended with rules to encode expressiveness criteria [16], preference rules validated in perception experiments, and general visualization design best practices. Hard constraints *must* be satisfied (e.g., shape encodings cannot express quantitative values), whereas soft constraints express a preference (e.g., temporal values should use the x-axis by default). The weights associated with soft constraints can be learned from preference or utility data, when available (see example in Figure 3).

Draco implements the constraints using Answer Set Programming (ASP) semantics, and casts the problem of finding appropriate encodings as finding optimal answer sets [10]. Draco has been extended to optimize for constraints over multiple visualizations [22], and adapted for use in specialized domains.

Using Draco (or similar formalizations), the specialized constraints governing the construction of nutritional labels can be developed in the general framework of ASP, while borrowing the foundational constraints capturing

general visualization design principles. This approach can help reduce the cost of developing hundreds of application-specific labels by encoding common patterns, such as including descriptive statistics in all labels, or only showing fairness visualizations when bias is detected.

3.2 Learning as Optimization

Sun et al. proposed MithraLabel [29], focusing on generating task-specific labels for datasets to determine fitness for specific tasks. Considering the dataset as a collection of items over a set of attributes, each widget provides specific information (such as functional dependencies) about the whole dataset or some selected part of it. For example, if a data scientist is considering the use of a number-of-prior-arrests attribute to predict likelihood of recidivism, she should know that the number of prior arrests is highly correlated with the likelihood of re-offending, but it introduces bias as the number of prior arrests is higher for African Americans than for other races due to policing practices and segregation effects in poor neighborhoods. Widgets that might appear in the nutritional label for prior arrests include the count of missing values, correlation with the predicted attribute or a protected attribute, and the distribution of values.

4 Properties of a nutritional label

The database and cyberinfrastructure communities have been studying systems and standards for metadata, provenance, and transparency for decades [20, 18]. We are now seeing renewed interest in these topics due to the proliferation of data science applications that use data opportunistically. Several recent projects explore these concepts for data and algorithmic transparency, including the Dataset Nutrition Label project [12], Datasheets for Datasets [9], and Model Cards [17]. All these method rely on manually constructed annotations. In contrast, our goal is to *generate labels automatically or semi-automatically*.

To differentiate a nutritional label from more general forms of metadata, we articulate several properties:

- **Comprehensible**: The label is not a complete (and therefore overwhelming) history of every processing step applied to produce the result. This approach has its place and has been extensively studied in the literature on scientific workflows, but is unsuitable for the applications we target. The information on a nutritional label must be short, simple, and clear.

- **Consultative**: Nutritional labels should provide actionable information, rather than just descriptive metadata. For example, universities may invest in research to improve their ranking, or consumers may cancel unused credit card accounts to improve their credit score.

- **Comparable**: Nutritional labels enable comparisons between related products, implying a standard. The IEEE is developing a series of ethics standards, known as the IEEE P70xx series, as part of its Global Initiative on Ethics of Autonomous and Intelligent Systems.[7] These standards include "IEEE P7001: Transparency of Autonomous Systems" and "P7003: Algorithmic Bias Considerations" [15]. The work on nutritional labels is synergistic with these efforts.

- **Concrete**: The label must contain more than just general statements about the source of the data; such statements do not provide sufficient information to make technical decisions on whether or not to use the data.

Data and models are chained together into complex automated pipelines — computational systems "consume" datasets at least as often as people do, and therefore also require nutritional labels! We articulate additional properties in this context:

[7] `https://ethicsinaction.ieee.org/`

- **Computable**: Although primarily intended for human consumption, nutritional labels should be machine-readable to enable specific applications: data discovery, integration, automated warnings of potential misuse.
- **Composable**: Datasets are frequently integrated to construct training data; the nutritional labels must be similarly integratable. In some situations, the composed label is simple to construct: the union of sources. In other cases, the biases may interact in complex ways: a group may be sufficiently represented in each source dataset, but underrepresented in their join.
- **Concomitant**: The label should be carried with the dataset; systems should be designed to propagate labels through processing steps, modifying the label as appropriate, and implementing the paradigm of transparency by design.

5 Conclusions

In this paper we discussed work on transparency and interpretability for data and models based on the concept of a nutritional label. We presented Ranking Facts, a system that automatically derives nutritional labels for rankings, and outlined directions for ongoing research that casts the creation of nutritional labels as a computational problem, rather than as purely a design problem.

We advocate interpretability tools for a variety of datasets and models, for a broad class of application domains, and to accommodate the needs of a variety of stakeholders. These tools must be informed by an understanding of how humans perceive algorithms and the decisions they inform, including issues of trust and agency to challenge or accept an algorithm-informed decision. These tools aim to reduce bias and errors in deployed models by preventing the use of an inappropriate dataset or model at design time. Although the extent of data misuse is difficult to measure directly, we can design experiments to show how well nutritional labels inform usage decisions, and design the tools accordingly. More broadly, we see the review of human-curated and machine-computed metadata as a critical step for interpretability in data science, which can lead to lasting progress in the use of machine-assisted decision-making in society.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: Risk assessments in criminal sentencing. *ProPublica*, May 2016.

[2] Abolfazl Asudeh, H. V. Jagadish, Gerome Miklau, and Julia Stoyanovich. On obtaining stable rankings. *PVLDB*, 12(3):237–250, 2018.

[3] Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das. Designing fair ranking schemes. In *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019.*, pages 1259–1276, 2019.

[4] Abolfazl Asudeh, Zhongjun Jin, and H. V. Jagadish. Assessing and remedying coverage for a given dataset. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*, pages 554–565, 2019.

[5] Danielle K. Citron and Frank A. Pasquale. The scored society: Due process for automated predictions. *Washington Law Review*, 89, 2014.

[6] Cory Booker, Ron Wyden, Yvette Clarke. Algorithmic Accountability Act. `https://www.wyden.senate.gov/imo/media/doc/Algorithmic\%20Accountability\%20Act\%20of\%202019\%20Bill\%20Text.pdf`, 2019. [Online; accessed 3-May-2019].

[7] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, October 2018.

[8] Marina Drosou, HV Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. Diversity in Big Data: A review. *Big Data*, 5(2), 2017.

[9] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *CoRR*, abs/1803.09010, 2018.

[10] Martin Gebser. *Proof theory and algorithms for answer set programming*. PhD thesis, University of Potsdam, 2011.

[11] Malcolm Gladwell. The order of things. *The New Yorker*, February 14, 2011.

[12] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. *CoRR*, abs/1805.03677, 2018.

[13] David Ingold and Spencer Soper. Amazon doesn't consider the race of its customers. should it? *Bloomberg*, April 2016.

[14] Alicia Key, Bill Howe, Daniel Perry, and Cecilia R. Aragon. Vizdeck: self-organizing dashboards for visual analytics. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, pages 681–684, 2012.

[15] Ansgar R. Koene, Liz Dowthwaite, and Suchana Seth. IEEE p7003™ standard for algorithmic bias considerations: work in progress paper. In *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden, May 29, 2018*, pages 38–41, 2018.

[16] Jock D. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Trans. Graph.*, 5(2):110–141, 1986.

[17] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 220–229, 2019.

[18] Luc Moreau, Bertram Ludäscher, Ilkay Altintas, Roger S. Barga, Shawn Bowers, Steven P. Callahan, George Chin Jr., Ben Clifford, Shirley Cohen, Sarah Cohen Boulakia, Susan B. Davidson, Ewa Deelman, Luciano A. Digiampietri, Ian T. Foster, Juliana Freire, James Frew, Joe Futrelle, Tara Gibson, Yolanda Gil, Carole A. Goble, Jennifer Golbeck, Paul T. Groth, David A. Holland, Sheng Jiang, Jihie Kim, David Koop, Ales Krenek, Timothy M. McPhillips, Gaurang Mehta, Simon Miles, Dominic Metzger, Steve Munroe, Jim Myers, Beth Plale, Norbert Podhorszki, Varun Ratnakar, Emanuele Santos, Carlos Eduardo Scheidegger, Karen Schuchardt, Margo I. Seltzer, Yogesh L. Simmhan, Cláudio T. Silva, Peter Slaughter, Eric G. Stephan, Robert Stevens, Daniele Turi, Huy T. Vo, Michael Wilde, Jun Zhao, and Yong Zhao. Special issue: The first provenance challenge. *Concurrency and Computation: Practice and Experience*, 20(5):409–418, 2008.

[19] Dominik Moritz, Chenglong Wang, Gregory Nelson, Halden Lin, Adam M. Smith, Bill Howe, and Jeffrey Heer. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2019.

[20] Open provenance. `https://openprovenance.org`. [Online; accessed 14-August-2019].

[21] Haoyue Ping, Julia Stoyanovich, and Bill Howe. Datasynthesizer: Privacy-preserving synthetic datasets. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017*, pages 42:1–42:5, 2017.

[22] Zening Qu and Jessica Hullman. Keeping multiple views consistent: Constraints, validations, and exceptions in visualization authoring. *IEEE Trans. Vis. Comput. Graph.*, 24(1):468–477, 2018.

[23] Luke Rodriguez, Babak Salimi, Haoyue Ping, Julia Stoyanovich, and Bill Howe. MobilityMirror: Bias-adjusted transportation datasets. In *Big Social Data and Urban Computing - First Workshop, BiDU@VLDB 2018, Rio de Janeiro, Brazil, August 31, 2018, Revised Selected Papers*, pages 18–39, 2018.

[24] Matthew J. Salganik. *Bit By Bit: Social Research in the Digital Age*. Princeton University Press, 2019.

[25] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019.*, pages 793–810, 2019.

[26] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. Vega-lite: A grammar of interactive graphics. *IEEE Trans. Vis. Comput. Graph.*, 23(1):341–350, 2017.

[27] Julia Stoyanovich, Sihem Amer-Yahia, and Tova Milo. Making interval-based clustering rank-aware. In *EDBT 2011, 14th International Conference on Extending Database Technology, Uppsala, Sweden, March 21-24, 2011, Proceedings*, pages 437–448, 2011.

[28] Julia Stoyanovich, Ke Yang, and H. V. Jagadish. Online set selection with fairness and diversity constraints. In *Proceedings of the 21th International Conference on Extending Database Technology, EDBT 2018, Vienna, Austria, March 26-29, 2018.*, pages 241–252, 2018.

[29] Chenkai Sun, Abolfazl Asudeh, H. V. Jagadish, Bill Howe, and Julia Stoyanovich. MithraLabel: Flexible dataset nutritional labels for responsible data science. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, 2019.

[30] The European Union. Regulation (EU) 2016/679: General Data Protection Regulation (GDPR). `https://gdpr-info.eu/`, 2016. [Online; accessed 15-August-2019].

[31] The Idaho house of Representatives. House Bill No. 118. `https://legislature.vermont.gov/bill/status/2018/H.378`, 2019. [Online; accessed 15-August-2019].

[32] The New York City Council. Int. No. 1696-A: A Local Law in relation to automated decision systems used by agencies. `https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0`, 2017. [Online; accessed on 15-August-2019].

[33] Vermont General Assembly. An act relating to the creation of the Artificial Intelligence Task Force. `https://legislature.idaho.gov/wp-content/uploads/sessioninfo/2019/legislation/H0118A2.pdf`, 2018. [Online; accessed 15-August-2019].

[34] Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. Balanced ranking with diversity constraints. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 6035–6042, 2019.

[35] Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017*, pages 22:1–22:6, 2017.

[36] Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, H. V. Jagadish, and Gerome Miklau. A nutritional label for rankings. In *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*, pages 1773–1776, 2018.

[37] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo A. Baeza-Yates. FA*IR: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 1569–1578, 2017.

[38] Indre Zliobaite. Measuring discrimination in algorithmic decision making. *Data Min. Knowl. Discov.*, 31(4):1060–1089, 2017.

Data Management for Causal Algorithmic Fairness*

Babak Salimi*, Bill Howe†, Dan Suciu*
University of Washington
*{bsalimi,suciu}@cs.washington.edu, †billhowe@uw.edu

Abstract

Fairness is increasingly recognized as a critical component of machine learning systems. However, it is the underlying data on which these systems are trained that often reflects discrimination, suggesting a data management problem. In this paper, we first make a distinction between associational and causal definitions of fairness in the literature and argue that the concept of fairness requires causal reasoning. We then review existing works and identify future opportunities for applying data management techniques to causal algorithmic fairness.

1 Introduction

Fairness is increasingly recognized as a critical component of machine learning (ML) systems. These systems are now routinely used to make decisions that affect people's lives [11], with the aim of reducing costs, reducing errors, and improving objectivity. However, there is enormous potential for harm: The data on which we train algorithms reflects societal inequities and historical biases, and, as a consequence, the models trained on such data will therefore reinforce and legitimize discrimination and opacity. The goal of research on algorithmic fairness is to remove bias from machine learning algorithms.

We recently argued that the algorithmic fairness problem is fundamentally a data management problem [43]. The selection of sources, the transformations applied during pre-processing, and the assumptions made during training are all sensitive to bias that can exacerbate fairness effects. The goal of this paper is to discuss the application of data management techniques in algorithmic fairness. In Sec 2 we make a distinction between associational and causal definitions of fairness in the literature and argue that the concept of fairness requires causal reasoning to capture natural situations, and that the popular associational definitions in ML can produce misleading results. In Sec 3 we review existing work and identify future opportunities for applying data management techniques to ensure causally fair ML algorithms.

2 Fairness Definitions

Algorithmic fairness considers a set of variables $\mathbf{V}$ that include a set of *protected attributes* $\mathbf{S}$ and a *response variable* Y, and a classification algorithm $\mathcal{A} : Dom(\mathbf{X}) \rightarrow Dom(O)$, where $\mathbf{X} \subseteq \mathbf{V}$, and the result is denoted O

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

*This work is supported by the National Science Foundation under grants NSF III-1703281, NSF III-1614738, NSF AITF 1535565 and NSF award #1740996.

Fairness Metric	Description
Demographic Parity (DP) [7] a.k.a. Statistical Parity [12] or Benchmarking [44]	$S \perp\!\!\!\perp O$
Conditional Statistical Parity [10]	$S \perp\!\!\!\perp O \mid \mathbf{A}$
Equalized Odds (EO) [15] [2] a.k.a. Disparate Mistreatment [47]	$S \perp\!\!\!\perp O \mid Y$
Predictive Parity (PP)[9] [3] a.k.a. Outcome Test [44] or Test-fairness [9] or Calibration [9], or Matching Conditional Frequencies [15]	$S \perp\!\!\!\perp Y \mid O$

Figure 1: Common associational definitions of fairness.

and called *outcome*. To simplify the exposition, we assume a sensitive attribute $S \in \mathbf{S}$ that classifies the population into protected $S = 1$ and privileged $S = 0$, for example, female and male, or minority and non-minority (see [48] for a survey). The first task is to define formally when an algorithm $\mathcal{A}$ is fair w.r.t. the protected attribute S; such a definition is, as we shall see, not obvious. Fairness definitions can be classified as associational or causal, which we illustrate using the following running example (see [45] for a survey on fairness definitions).

Example 1: *In 1973, UC Berkeley was sued for discrimination against females in graduate school admissions. Admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance. However, it turned out that the observed correlation was due to the indirect effect of gender on admission results through applicant's choice of department. It was shown that females tended to apply to departments with lower overall acceptance rates [41]. When broken down by department, a slight bias toward female applicants was observed, a result that did not constitute evidence for gender-based discrimination. Extending this case, suppose college admissions decisions are made independently by each department and are based on a rich collection of information about the candidates, such as test scores, grades, resumes, statement of purpose, etc. These characteristics affect not only admission decisions, but also the department to which the candidate chooses to apply. The goal is to establish conditions that guarantee fairness of admission decisions.*

2.1 Associational Fairness

A simple and appealing approach to defining fairness is by correlating the sensitive attribute S and the outcome O. This leads to several possible definitions (Fig. 1). *Demographic Parity* (DP) [12] requires an algorithm to classify both protected and privileged groups with the same probability, i.e., $\Pr(O = 1|S = 1) = \Pr(O = 1|S = 0)$. However, doing so fails to correctly model our Example 1 since it requires equal probability for males and females to be admitted, and, as we saw, failure of DP cannot be considered evidence for gender-based discrimination. This motivates *Conditional Statistical Parity* (CSP) [10], which controls for a set of admissible factors $\mathbf{A}$, i.e., $\Pr(O = 1|S = 1, \mathbf{A} = \mathbf{a}) = \Pr(O = 1|S = 0, \mathbf{A} = \mathbf{a})$. The definition is satisfied if subjects in both protected and privileged groups have equal probability of being assigned to the positive class, controlling for a set of admissible variables. In the UC Berkeley case, CSP is approximately satisfied by assuming that department is an admissible variable.

Another popular measure used for predictive classification algorithms is *Equalized Odds* (EO), which requires both protected and privileged groups to have the same false positive (FP) rate, $\Pr(O = 1|S = 1, Y = 0) = \Pr(O = 1|S = 0, Y = 0)$, and the same false negative (FN) rate, $\Pr(O = 0|S = 1, Y = 1) = \Pr(O = 0|S = 0, Y = 1)$, or, equivalently, $(O \perp\!\!\!\perp S|Y)$. In our example, assuming a classifier is trained to predict if an applicant will be admitted, then the false positive rate is the fraction of rejected applicants for which the classifier predicted that they should be admitted, and similarly for the false negative rate: EO requires

that the rates of these false predictions be the same for male and female applicants. Finally, *Predictive Parity* (PP) requires that both protected and privileged groups have the same predicted positive value (PPV), $\Pr(Y = 1|O = i, S = 0) = \Pr(Y = 1|O = i, S = 1)$ for $i, = \{1, 0\}$ or, equivalently, $Y \perp\!\!\!\perp S|O$. In our example, this implies that the probability of an applicant that actually got admitted to be correctly classified as admitted and the probability of an applicant that actually got rejected to be incorrectly classified as accepted should both be the same for male and female applicants.

An Associational Debate. Much of the literature in algorithmic fairness is motivated by controversies over a widely used commercial risk assessment system for recidivism — COMPAS by Northpointe [18]. In 2016, a team of journalists from ProPublica constructed a dataset of more than 7000 individuals arrested in Broward County, Florida between 2013 and 2014 in order to analyze the efficacy of COMPAS. In addition, they collected data on arrests for these defendants through the end of March 2016. Their assessment suggested that COMPAS scores were biased against African-Americans based on the fact that the FP rate for African-Americans (44.9%) was twice that for Caucasians (23.5%). However, the FN rate for Caucasians (47.7%) was twice as large as for African-Americans (28.0%). In other words, COMPAS scores were shown to violate EO. In response to ProPublica, Northpointe showed COMPAS scores satisfy PP, i.e., the likelihood of recidivism among high-risk offenders is the same regardless of race.

This example illustrates that associational definitions are context-specific and can be mutually exclusive; they lack universality. Indeed, it has been shown that EO and PP are incompatible. In particular, Chouldechova [9] proves the following impossibility result. Suppose that prevalence of the two populations differs, $\Pr(Y = 1|S = 0) \neq \Pr(Y = 1|S = 1)$, for example, the true rate of recidivism differs for African-Americans and Caucasians; in this case, Equalized Odds and Predictive Parity cannot hold both simultaneously. Indeed, EO implies that $FP_i/(1 - FN_i)$ is the same for both populations $S = i$, $i = 0, 1$, while PP implies that $(1 - PPV_i)/PPV_i$ must be the same. Then, the identity

$$\frac{FP_i}{1 - FN_i} = \frac{\Pr(O = 1|S = i, Y = 0)}{\Pr(O = 1|S = i, Y = 1)} = \frac{\Pr(Y = 1|S = i)}{\Pr(Y = 0|S = i)} \frac{\Pr(Y = 0|O = 1, S = i)}{\Pr(Y = 1|O = 1, S = i)} = \frac{\Pr(Y = 1|S = i)}{\Pr(Y = 0|S = i)} \frac{1 - PPV_i}{PPV_i}$$

for $i = 0, 1$, implies $\Pr(Y = 1|S = 0) = \Pr(Y = 1|S = 1)$. We revisit the impossibility result in Sec 2.3.

2.2 Causal Fairness

The lack of universality and the impossibility result for fairness definitions based on associational definitions have motivated definitions based on causality [17, 16, 25, 37, 13]. The intuition is simple: fairness holds when there is no causal relationship from the protected attribute S to the outcome O. We start with a short background on causality.

Causal DAG. A *causal DAG* G over a set of variables $\mathbf{V}$ is a directed acyclic graph that models the functional interaction between variables in $\mathbf{V}$. Each node X represents a variable in $\mathbf{V}$ that is functionally determined by: (1) its parents $\mathbf{Pa}(X)$ in the DAG, and (2) some set of *exogenous* factors that need not appear in the DAG as long as they are mutually independent. This functional interpretation leads to the same decomposition of the joint probability distribution of $\mathbf{V}$ that characterizes Bayesian networks [27]:

$$\Pr(\mathbf{V}) = \prod_{X \in \mathbf{V}} \Pr(X|\mathbf{Pa}(X)) \tag{1}$$

d-Separation. A common inference question in a causal DAG is how to determine whether a CI $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})$ holds. A sufficient criterion is given by the notion of d-separation, a syntactic condition $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|_d \mathbf{Z})$ that can be checked directly on the graph (we refer the reader to [26] for details).

Counterfactuals and do Operator. A *counterfactual* is an intervention where we actively modify the state of a set of variables $\mathbf{X}$ in the real world to some value $\mathbf{X} = \mathbf{x}$ and observe the effect on some output Y. Pearl [27] described the do operator, which allows this effect to be computed on a causal DAG, denoted $\Pr(Y|do(\mathbf{X} = \mathbf{x}))$. To compute this value, we assume that X is determined by a constant function $\mathbf{X} = \mathbf{x}$ instead of a function provided by the causal DAG. This assumption corresponds to a modified graph with all edges into $\mathbf{X}$ removed, and values of the incoming variables are set to $\mathbf{x}$. For a simple example, consider three random variables $X, Y, Z \in \{0, 1\}$. We randomly flip a coin and set $Z = 0$ or $Z = 1$ with probability $1/2$; next, we set $X = Z$, and finally we set $Y = X$. The resulting causal DAG is $Z \to X \to Y$, whose equation is $\Pr(X, Y, Z) = \Pr(Z)\Pr(X|Z)\Pr(Y|X)$. The do operator lets us observe what happens in the system when we intervene by setting $X = 0$. The result is defined by removing the edge $Z \to X$, whose equation is $\Pr(Y = y, Z = z|do(X) = 0) = \Pr(Z = z)\Pr(Y = y|X = 0)$ (notice that $\Pr(X|Z)$ is missing), leading to the marginals $\Pr(Y = 0|do(X) = 0) = 1, \Pr(Y = 1|do(X) = 0) = 0$. It is important to know the casual DAG since the probability distribution is insufficient to compute the do operator; for example, if we reverse the arrows to $Y \to X \to Z$ (flip Y first, then set $X = Y$, then set $Z = X$), then $\Pr(Y = 0|do(X) = 0) = \Pr(Y = 1|do(X) = 0) = 1/2$ in other words, intervening on X has no effect on Y.

Counterfactual Fairness. Given a set of features $\mathbf{X}$, a protected attribute S, an outcome variable Y, and a set of unobserved exogenous background variables $\mathbf{U}$, Kusner et al. [17] defined a predictor O to be *counterfactually fair* if for any $\mathbf{x} \in Dom(\mathbf{X})$:

$$P(O_{S\leftarrow 0}(\mathbf{U}) = 1|\mathbf{X} = \mathbf{x}, S = 1) = P(O_{S\leftarrow 1}(\mathbf{U}) = 1|\mathbf{X} = \mathbf{x}; S = 1) \tag{2}$$

where $O_{S\leftarrow s}(\mathbf{U})$ means intervening on the protected attribute in an unspecified configuration of the exogenous factors. The definition is meant to capture the requirement that the protected attribute S should not be a cause of O at the individual level. However, this definition captures individual-level fairness only under certain strong assumptions (see [43]). Indeed, it is known in statistics that individual-level counterfactuals cannot be estimated from data [34, 35, 36].

Proxy Fairness. To avoid individual-level counterfactuals, a common approach is to study population-level counterfactuals or interventional distributions that capture the effect of interventions at population rather than individual level [28, 34, 35]. Kilbertus et al. [16] defined proxy fairness as follows:

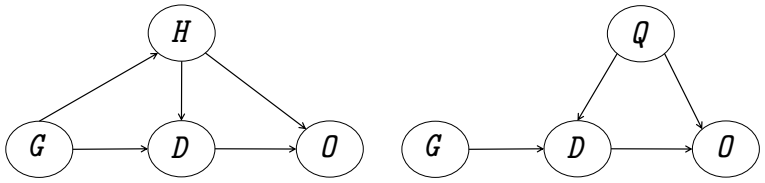
$$P(O = 1|do(\mathbf{P} = \mathbf{p})) = P(O = 1|do(\mathbf{P} = \mathbf{p}')) \tag{3}$$

for any $\mathbf{p}, \mathbf{p}' \in Dom(\mathbf{P})$, where $\mathbf{P}$ consists of proxies to a sensitive variable S (and might include S). Intuitively, a classifier satisfies proxy fairness in Eq 3 if the distribution of O under two interventional regimes in which $\mathbf{P}$ set to $\mathbf{p}$ and $\mathbf{p}'$ is the same. Thus, proxy fairness is not an individual-level notion. It has been shown that proxy fairness fails to capture group-level discrimination in general [43].

Path-Specific Fairness. These definitions are based on graph properties of the causal graph, *e.g.*, prohibiting specific paths from the sensitive attribute to the outcome [25, 22]; however, identifying path-specific causality from data requires very strong assumptions and is often impractical [4].

Interventional Fairness. To avoid issues with the aforementioned causal definitions, Salimi et al. [43] defined interventional fairness as follows: an algorithm $\mathcal{A} : Dom(\mathbf{X}) \to Dom(O)$ is $\mathbf{K}$-fair for a set of attributes $\mathbf{K} \subseteq \mathbf{V} - \{S, O\}$ w.r.t. a protected attribute S if, for any context $\mathbf{K} = \mathbf{k}$ and every outcome $O = o$, the following holds:

$$\Pr(O = o|do(S = 0), do(\mathbf{K} = \mathbf{k})) = \Pr(O = o|do(S = 1), do(\mathbf{K} = \mathbf{k})) \tag{4}$$



(a) College I (b) College II

College I	Dept. A		Dept. B		Total	
	Admitted	Applied	Admitted	Applied	Admitted	Applied
Male	16	20	16	80	32	100
Female	16	80	16	20	32	100

College II	Dept. A		Dept. B		Total	
	Admitted	Applied	Admitted	Applied	Admitted	Applied
Male	10	10	40	90	50	100
Female	40	50	10	50	50	100

Figure 2: Admission process representation in two colleges where associational fairness fail (see Ex.2).

An algorithm is called *interventionally fair* if it is $\mathbf{K}$-fair for every set $\mathbf{K}$. Unlike proxy fairness, this notion correctly captures group-level fairness because it ensures that S does not affect O in *any configuration* of the system obtained by fixing other variables at some arbitrary values. Unlike counterfactual fairness, it does not attempt to capture fairness at the individual level, and therefore it uses the standard definition of intervention (the do-operator). In practice, interventional fairness is too restrictive. For example, in the UC Berkeley case, admission decisions were not interventionally fair since gender affected the admission result via applicant's choice of department. To make it practical, Salimi et al. [43] defined a notion of fairness that relies on partitioning variables into *admissible* and *inadmissible*. The former are variables through which it is permissible for the protected attribute to influence the outcome. This partitioning expresses fairness social norms and values and comes from the users. In Example 1, the user would label department as admissible since it is considered a fair use in admissions decisions and would (implicitly) label all other variables as inadmissible, for example, hobby. Then, an algorithm is called *justifiably fair* if it is $\mathbf{K}$-fair w.r.t. all supersets $\mathbf{K} \supseteq \mathbf{A}$. We illustrate with an example.

Example 2: Fig 2 shows how fair or unfair situations may be hidden by coincidences but exposed through causal analysis. In both examples, the protected attribute is gender G, and the admissible attribute is department D. Suppose both departments in College I are admitting only on the basis of their applicants' hobbies. Clearly, the admission process is discriminatory in this college because department A admits 80% of its male applicants and 20% of the female applicants, while department B admits 20% of male and 80% of female applicants. On the other hand, the admission rate for the entire college is the same 32% for both male and female applicants, falsely suggesting that the college is fair. Suppose H is a proxy to G such that $H = G$ (G and H are the same); proxy fairness then classifies this example as fair: indeed, since Gender has no parents in the causal graph, intervention is the same as conditioning; hence, $\Pr(O = 1|do(G = \text{i})) = \Pr(O = 1|G = \text{i})$ for $i = 0, 1$. Of the previous methods, only conditional statistical parity correctly indicates discrimination. We illustrate how our definition correctly classifies this examples as unfair. Indeed, assuming the user labels the department D as admissible, $\{D\}$-fairness fails because $\Pr(O = 1|do(G = 1), do(D = \text{'A'})) = \sum_h \Pr(O = 1|G = 1, D = \text{'A'}, H = h)\Pr(H = h|G = 1) = \Pr(O = 1|G = 1, D = \text{'A'}) = 0.8$, and, similarly $\Pr(O = 1|do(G = 0), do(D = \text{'A'})) = 0.2$. Therefore, the admission process is not justifiably fair.

Now, consider the second table for College II, where both departments A and B admit only on the basis of student qualifications Q. A superficial examination of the data suggests that the admission is unfair: department A admits 80% of all females and 100% of all male applicants; department B admits 20% and 44.4%, respectively. Upon deeper examination of the causal DAG, we can see that the admission process is justifiably fair because the only path from Gender to Outcome goes through Department, which is an admissible attribute. To understand how the data could have resulted from this causal graph, suppose 50% of each gender have high qualifications and are admitted, while others are rejected, and that 50% of females apply to each department, but more qualified females apply to department A than to B (80% vs 20%). Further, suppose fewer males apply to department A, but all of them are qualified. The algorithm satisfies demographic parity and proxy fairness but fails to satisfy conditional statistical parity since $\Pr(A = 1|G = 1, D = \text{A}) = 0.8$ but $\Pr(A = 1|G = 0, D = \text{A}) = 0.2)$. Thus, conditioning on D falsely indicates discrimination in College II. One can check that the algorithm is justifiably fair, and thus our definition also correctly classifies this example; for example, $\{D\}$-fairness follows from $\Pr(O = 1|do(G = i), do(D = d)) = \sum_q \Pr(O = 1|G = i, D = d, Q = q))\Pr(Q = q|G = i) = \frac{1}{2}$. To summarize, unlike previous definitions of fairness, justifiable fairness correctly identifies College I as discriminatory and College II as fair.

2.3 Impossibility Theorem from the Causality Perspective

From the point of view of causal DAGs, EO requires that the training label Y d-separates the sensitive attribute S and the outcome of the classifier O. Intuitively, this implies that S can affect classification results only when the information comes through the training label Y. On the other hand, PP requires that the classifier outcome O d-separates the sensitive attribute S and the training labels Y. Intuitively, this implies S can affect the training labels only when the information comes thorough the outcome of classifier O. These interpretations clearly reveal the inconsistent nature of EO and PP. It is easy to show for strictly positive distributions that the CIs $(S \perp\!\!\!\perp O|Y)$ and $(S \perp\!\!\!\perp Y|O)$ imply $(S \perp\!\!\!\perp Y)$ or, equivalently, $\Pr(Y = 1|S = 0) = \Pr(Y = 1|S = 1)$ (see [43]). Indeed, from the causality perspective, EO and PP are neither sufficient nor necessary for fairness. In the causal DAG in Fig 3(b), suppose a classifier is trained on an applicant's qualifications Q to approximate admission committee decisions $\hat{O}$. It is clear that the classifier is not discriminative, yet it violates both EO and PP. The reader can verify that the causal DAG obtained by further adding an edge from Q to $\hat{O}$ (to account for the classifier outcome) does not imply the CIs $(G \perp\!\!\!\perp O|\hat{O})$ and $(G \perp\!\!\!\perp \hat{O}|O)$.

3 Data Management Techniques for Causal Fairness

3.1 Causal Fairness as Integrity Constraints

In causal DAGs, the missing arrow between two variables X and Y represents the assumption of no causal effect between them, which corresponds to the CI statement $(X \perp\!\!\!\perp Y|\mathbf{Z})$, where $\mathbf{Z}$ is a set of variables that d-separates X and Y. For example, the missing arrow between O and G in the causal DAG in Fig. 2(a) encodes the CI $(O \perp\!\!\!\perp G|H, D)$. On the other hand, the lack of certain arrows in the underling causal DAG is sufficient to satisfy different causal notions of fairness (cf. Sec 2.2). For instance, a sufficient condition for justifiable fairness in the causal DAG in Fig. 2(a) is the lack of the edge from H to O, which corresponds to the CI $(O \perp\!\!\!\perp G, H|D)$. Thus, fairness can be captured as a set of CI statements. Now to enforce fairness, instead of intervening on the causal DAG over which we have no control, we can intervene on data to enforce the corresponding CI statements.

Consequently, social causal fairness constraints can be seen as a set of integrity constraints in the form of CIs that must be preserved and enforced thorough the data science pipeline, from data gathering through the deployment of a machine learning model. The connection between CIs and well-studied integrity constraints in data management – such as Multi Valued Dependencies (MVDs) and Embedded Multi Valued Dependencies (EMVDs) [1] – opens the opportunity to leverage existing work in data management to detect and avoid bias in data.

SQL Query:
```
SELECT avg(Income)
FROM AdultData
GROUP BY Gender
```

Gender	SQL Query	Rewritten Query
Female	*0.11*	0.10
Male	0.30	0.11

Coarse-grained Explanation:

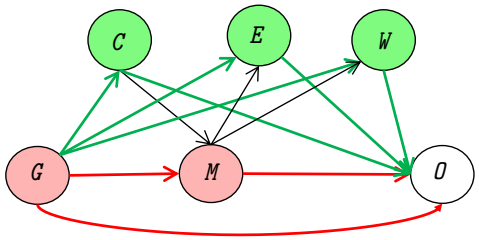
Attribute	*Res.*
MaritalStatus	*0.58*
Education	*0.13*
HoursPerwWeek	*0.04*
Age	*0.04*

Fine-grained Explanation:

Rank	MarialStatus	Gender	Income
1	Married	Male	1
2	Single	Female	0

Rank	Education	Gender	Income
1	Bachelors	Male	1
2	SomeCollage	Female	0

(a)



(b)

Figure 3: (a) HYPDB's report on the effect of gender on income (cf. Ex. 1). (b) A compact causal DAG with $O =$ income, $G =$ gender, $M =$ marital status, $C =$ age and nationality, $E =$ education and $W =$ work class, occupation and hours per week (cf. Ex. 3).

3.2 Query Rewriting

In data management, *query rewriting* refers to a set of techniques to automatically modify one query into another that satisfies certain desired properties. These techniques are used to rewrite queries with views [19], in chase and backchase for complex optimizations [29], and for many other applications. This section discusses query rewriting techniques for detecting and enforcing fairness.

3.2.1 Detecting Discrimination

As argued in Sec 2.2, detecting discrimination should rely on performing a hypothesis test on the causal effect of membership in minority $S = 1$ or privileged group $S = 0$ on an outcome of an algorithm O. The gold standard for such causal hypothesis testing is a *randomized experiment* (or an *A/B test*), called such because treatments are randomly assigned to subjects. In contrast, in the context of fairness, sensitive attributes are typically imputable; hence, randomization is not even conceivable. Therefore, such queries must be answered using *observational data*, defined as data recorded from the environment with no randomization or other controls. Although causal inference in observational data has been studied in statistics for decades, causal analysis is not supported in existing online analytical processing (OLAP) tools [41]. Indeed, today, most data analysts still reach for the simplest query that computes the average of O Group By S to answer such questions, which, as shown in Ex 1, can lead to incorrect conclusions. Salimi et al. [41] took the first step toward extending existing OLAP tools to support causal analysis. Specifically, they introduced the HYPDB system, which brings together techniques from data management and causal inference to automatically rewrite SQL group-by queries into complex causal queries that support decision making. We illustrate HYPDB by applying it to a fairness question (see [40] for additional examples).

Example 3: Using UCI adult Census data [20], several prior works in algorithmic fairness have reported gender discrimination based on the fact that 11% of women have high income compared to 30% of men, which suggests a huge disparity against women. To decide whether the observed strong correlation between gender and high income is due to discrimination, we need to understand its causes. To perform this analysis using HYPDB, one can start with the simple group-by query (Fig. 3(a)) that computes the average of Income (1 iff Income>

50k) Group By Gender, which indeed suggests a strong disparity with respect to females' income. While the group-by query tells us gender and high income are highly correlated, it does not tell us why. To answer this question, HYPDB automatically infers from data that gender can potentially influence income indirectly via MaritalStatus, Education, Occupation, etc. (the indirect causal paths from G to O in Fig. 3(b)). Then, HYPDB automatically rewrites the group-by query to quantify the direct and indirect effect of gender on income. Answers to the rewritten queries suggest that the direct effect of gender on income is not significant (the effect through the arrow from G to O in Fig. 3(b)). Hence, gender essentially influences income indirectly through mediating variables. To understand the nature of this influences, HYPDB provides the user with several explanations. These show that MaritalStatus accounts for most of the indirect influence, followed by Education. However, the top fine-grained explanations for MaritalStatus reveal surprising facts: there are more married males in the data than married females, and marriage has a strong positive association with high income. It turns out that the income attribute in US census data reports the adjusted gross income as indicated in the individual's tax forms; these depend on filing status (jointly and separately), could be household income. HYPDB explanations also show that males tend to have higher levels of education than females, and higher levels of education is associated with higher incomes. The explanations generated by HYPDB illuminate crucial factors for investigating gender discrimination.

Future Extensions. Incorporating the type of analyses supported by HYPDB into data-driven decision support systems is not only crucial for sound decision making in general, but it is also important for detecting, explaining and avoiding bias and discrimination in data and analytics. Further research is required on extending HYPDB to support more complex types of queries and data, such as multi-relational and unstructured.

3.2.2 Enforcing Fairness

Raw data often goes through a series of transformations to enhance the clarity and relevance of the signal used for a particular machine learning application [3]. Filter transformations are perhaps most common, in which a subset of training data is removed based on predicates. Even if the raw data is unbiased, filtering can introduce bias [3, 41]: It is known that causal DAGs are not closed under conditioning because CIs may not hold in some subset. Hence, filtering transformations can lead to violation of causal fairness integrity constraints. It is also known that conditioning on common effects can further introduce bias even when the sensitive attribute and training labels are marginally independent [26]. This motivates the study of *fairness-aware data transformations*, where the idea is to minimally rewrite the transformation query so certain fairness constraints are guaranteed to be satisfied in the result of the transformation. This problem is closely related to that of constraint-based data transformations studied in [3]. However, fairness constraints go beyond the types of constraints considered in [3] and are more challenging to address. Note that a solution to the aforementioned problem can be used to enforce fairness-constraints for raw data by applying a fair-transformation that selects all the data.

3.3 Database Repair

Given a set of integrity constraints Γ and a database instance D that is inconsistent with Γ, the problem of repairing D is to find an instance D' that is close to D and consistent with Γ. Repair of a database can be obtained by deletions and insertions of whole tuples as well as by updating attributes. The closeness between D and D' can be interpreted in many different ways, such as the minimal number of changes or the minimal set of changes under set inclusion (refer to [6] for a survey). The problem has been studied extensively in database theory for various classes of constraints. It is NP-hard even when D consists of a single relation and Γ consists of functional dependencies [21].

Given a training data D that consists of a training label Y, a set of admissible variables $\mathbf{A}$, and a set of inadmissible variables $\mathbf{I}$, Salimi et al [43] showed that a sufficient condition for a classifier to be justifiably fair is that the empirical distribution $\Pr$ over D satisfies the CI $(Y \perp\!\!\!\perp \mathbf{I}|\mathbf{A})$. Further, they introduced the CAPUCHIN system, which minimally repairs D by performing a sequence of database updates (viz., insertions and deletions

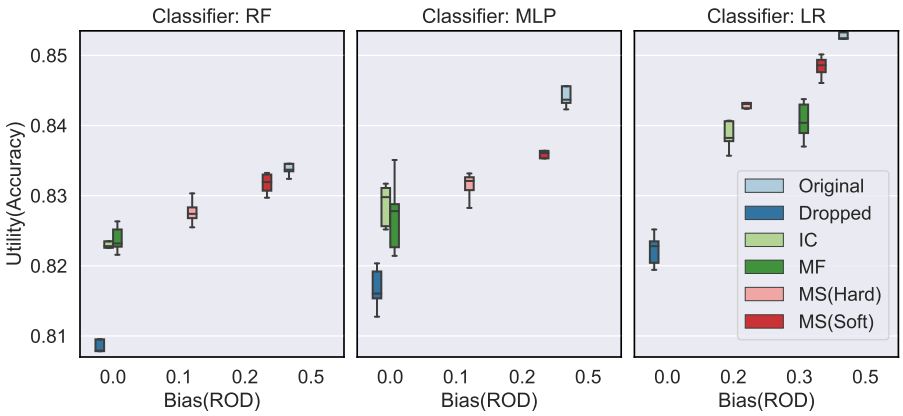


Figure 4: Performance of CAPUCHIN on Adult data.

of tuples) to obtain another training database D' that satisfies $(Y \perp\!\!\!\perp \mathbf{I}|\mathbf{A})$. Specifically, they reduced the problem to a minimal repair problem w.r.t. an MVD and developed a set of techniques, including reduction to the MaxSAT and Matrix Factorization, to address the corresponding optimization problem. We illustrate CAPUCHIN with an example.

Example 4: Suppose financial organisations use the Adult data described in Ex 1 to train an ML model to assist them in verifying the reliability of their customers. The use of raw data for training an ML model leads to a model that is discriminative against females simply because the model picks up existing bias in data, as described in Ex 3. To remove direct and indirect effects of gender on income (the red paths from G to Y in Fig. 4(b)) using the CAPUCHIN system, it is sufficient to enforce the CI $(O \perp\!\!\!\perp \mathbf{S}, \mathbf{M}|\mathbf{C}, \mathbf{E}, \mathbf{W})$ in data. Then, any model trained on the repaired data can be shown to be justifiably fair even on unseen test data under some mild assumptions [43]. To empirically assess the efficacy of the CAPUCHIN system, we repaired Adult data using the following CAPUCHINalgorithms: Matrix Factorization (MF), Independent Coupling (IC), and two versions of the MaxSAT approach: MS(Hard), which strictly enforces a CI, and MS(Soft), which approximately enforces a CI. Then, three classifiers – Linear Regression (LR), Multi-layer Perceptron (MLP), and Random Forest (RF) – were trained on both original and repaired training datasets using the set of variables $\mathbf{A} \cup \mathbf{N} \cup \mathbf{S}$. The classifier also trained on raw data using only $\mathbf{A}$, i.e., we dropped the sensitive and inadmissible variables. The utility and bias metrics for each repair method were measured using five-fold cross validation. Utility was measured by the classifiers' accuracy, and bias measured by the Ratio of Observational discrimination introduced in [43], which quantifies the effect of gender on outcome of the classifier by controlling for admissible variables (see [42] for details). Fig. 4 compares the utility and bias of CAPUCHIN repair methods on Adult data. As shown, all repair methods successfully reduced the ROD for all classifiers. The CAPUCHIN repair methods had an effect similar to dropping the sensitive and inadmissible variables completely, but they delivered much higher accuracy (because the CI was enforced approximately).

Future Extensions. The problem of repairing data w.r.t a set of CI constraints was studied in [43] for a single saturated CI constraint problem.[1] In the presence of multiple training labels and sensitive attributes, one needs to enforce multiple potentially interacting or inconsistent CIs; this is more challenging and requires further investigation. In addition, further research is required on developing approximate repair methods to be able to trade the fairness and accuracy of different ML applications.

[1] A CI statement is saturated if it contains all attributes.

3.4 Fairness-Aware Weak Supervision Methods

ML pipelines rely on massive labeled training sets. In most practical settings, such training datasets either do not exist or are very small. Constructing large labeled training datasets can be expensive, tedious, time-consuming or even impractical. This has motivated a line of work on developing techniques for addressing the data labeling bottleneck, referred to as *weak supervision methods*. The core idea is to programmatically label training data using, e.g., domain heuristics [31], crowdsourcing [32] and distant supervision [24]. In this context, the main challenges are handling noisy and unreliable sources that can potentially generate labels that are in conflict and highly correlated. State-of-the-art frameworks for weak supervision, such as Snorkel [30], handle these challenges by training label models that take advantage of conflicts between all different labeling sources to estimate their accuracy. The final training labels are obtained by combining the result of different labeling sources weighted by their estimated accuracy. While the focus of existing work is on collecting quality training labels to maximize the accuracy of ML models, the nuances of fairness cannot be captured by the exiting machinery to assess the reliability of the labeling sources. In particular, a new set of techniques is required to detect and explain whether certain labeling sources are biased and to combine their votes fairly.

3.5 Provenance for Explanation

Data provenance refers to the origin, lineage, and source of data. Various data provenance techniques have been proposed to assist researchers in understanding the origins of data [14]. Recently, data provenance techniques has been used to explain why integrity constraints fail [46]. These techniques are not immediately applicable to fairness integrity constraints, which are probabilistic. This motivates us to extend provenance to fairness or probabilistic integrity constraints in general. This extension is particularly crucial for reasoning about the fairness of training data collected from different sources by data integration and fusion, and it opens the opportunity to leverage existing techniques, such as provenance summarization [2], why-not provenance [8], and query-answers causality and responsibility [23, 38, 39, 5], explanations for database queries queries [33] to generate fine- and coarse-grained explanations for bias and discrimination.

4 Conclusions

This paper initiated a discussion on applying data management techniques in the embedding areas of algorithmic fairness in ML. We showed that fairness requires causal reasoning to capture natural situations, and that popular associational definitions in ML can produce incorrect or misleading results.

References

[1] Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

[2] Eleanor Ainy, Pierre Bourhis, Susan B Davidson, Daniel Deutch, and Tova Milo. Approximated summarization of data provenance. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 483–492. ACM, 2015.

[3] Dolan Antenucci and Michael Cafarella. Constraint-based explanation and repair of filter-based transformations. *Proceedings of the VLDB Endowment*, 11(9):947–960, 2018.

[4] Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 357–363, 2005.

[5] Leopoldo Bertossi and Babak Salimi. Causes for query answers from databases: Datalog abduction, view-updates, and integrity constraints. *International Journal of Approximate Reasoning*, 90:226–252, 2017.

[6] Leopoldo E. Bertossi. *Database Repairing and Consistent Query Answering*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2011.

[7] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*, pages 13–18. IEEE, 2009.

[8] Adriane Chapman and HV Jagadish. Why not? In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 523–534. ACM, 2009.

[9] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[10] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

[11] Rachel Courtland. Bias detectives: the researchers striving to make algorithms fair. *Nature*, 558, 2018.

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[13] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pages 498–510. ACM, 2017.

[14] Boris Glavic and Klaus Dittrich. Data provenance: A categorization of existing approaches. *Datenbanksysteme in Business, Technologie und Web (BTW 2007)–12. Fachtagung des GI-Fachbereichs" Datenbanken und Informationssysteme"(DBIS)*, 2007.

[15] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[16] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

[17] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.

[18] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9, 2016.

[19] Alon Y. Levy, Alberto O. Mendelzon, Yehoshua Sagiv, and Divesh Srivastava. Answering queries using views. In *Proceedings of the Fourteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, May 22-25, 1995, San Jose, California, USA*, pages 95–104, 1995.

[20] M. Lichman. Uci machine learning repository, 2013.

[21] Ester Livshits, Benny Kimelfeld, and Sudeepa Roy. Computing optimal repairs for functional dependencies. In *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, Houston, TX, USA, June 10-15, 2018*, pages 225–237, 2018.

[22] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *CoRR*, abs/1805.05859, 2018.

[23] Alexandra Meliou, Wolfgang Gatterbauer, Katherine F Moore, and Dan Suciu. The complexity of causality and responsibility for query answers and non-answers. *Proceedings of the VLDB Endowment*, 4(1):34–45, 2010.

[24] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

[25] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2018, page 1931. NIH Public Access, 2018.

[26] Judea Pearl. Causality: models, reasoning, and inference. *Econometric Theory*, 19(675-685):46, 2003.

[27] Judea Pearl. *Causality*. Cambridge university press, 2009.

[28] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

[29] Lucian Popa, Alin Deutsch, Arnaud Sahuguet, and Val Tannen. A chase too far? In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA.*, pages 273–284, 2000.

[30] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017.

[31] Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In *Advances in neural information processing systems*, pages 3567–3575, 2016.

[32] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322, 2010.

[33] Sudeepa Roy and Dan Suciu. A formal approach to finding explanations for database queries. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1579–1590. ACM, 2014.

[34] Donald B Rubin. *The Use of Matched Sampling and Regression Adjustment in Observational Studies*. Ph.D. Thesis, Department of Statistics, Harvard University, Cambridge, MA, 1970.

[35] Donald B Rubin. Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986.

[36] Donald B Rubin. Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, 103(484):1350–1353, 2008.

[37] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, pages 6414–6423, 2017.

[38] Babak Salimi and Leopoldo E. Bertossi. From causes for database queries to repairs and model-based diagnosis and back. In *ICDT*, pages 342–362, 2015.

[39] Babak Salimi, Leopoldo E Bertossi, Dan Suciu, and Guy Van den Broeck. Quantifying causal effects on query answering in databases. In *TaPP*, 2016.

[40] Babak Salimi, Corey Cole, Peter Li, Johannes Gehrke, and Dan Suciu. Hypdb: a demonstration of detecting, explaining and resolving bias in olap queries. *Proceedings of the VLDB Endowment*, 11(12):2062–2065, 2018.

[41] Babak Salimi, Johannes Gehrke, and Dan Suciu. Bias in olap queries: Detection, explanation, and removal. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1021–1035. ACM, 2018.

[42] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Capuchin: Causal database repair for algorithmic fairness. *CoRR*, abs/1902.08283, 2019.

[43] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pages 793–810. ACM, 2019.

[44] Camelia Simoiu, Sam Corbett-Davies, Sharad Goel, et al. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

[45] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

[46] Jane Xu, Waley Zhang, Abdussalam Alawini, and Val Tannen. Provenance analysis for missing answers and integrity repairs. *IEEE Data Eng. Bull.*, 41(1):39–50, 2018.

[47] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

[48] Indre Zliobaite. A survey on measuring indirect discrimination in machine learning. *CoRR*, abs/1511.00148, 2015.

A Declarative Approach to Fairness in Relational Domains

Golnoosh Farnadi[1,2], Behrouz Babaki[1], Lise Getoor[3]
[1]Polytechnique Montréal, [2] Mila, [3] UC Santa Cruz
farnadig@mila.quebec, behrouz.babaki@polymtl.ca, getoor@soe.ucsc.edu

Abstract

AI and machine learning tools are being used with increasing frequency for decision making in domains that affect peoples' lives such as employment, education, policing and financial qualifications. These uses raise concerns about biases of algorithmic discrimination and have motivated the development of fairness-aware machine learning. However, existing fairness approaches are based solely on attributes of individuals. In many cases, discrimination is much more complex, and taking into account the social, organizational, and other connections between individuals is important. We introduce new notions of fairness that are able to capture the relational structure in a domain. We use first-order logic to provide a flexible and expressive language for specifying complex relational patterns of discrimination. Furthermore, we extend an existing statistical relational learning framework, probabilistic soft logic (PSL), to incorporate our definition of relational fairness. We refer to this fairness-aware framework FairPSL. FairPSL makes use of the logical definitions of fairnesss but also supports a probabilistic interpretation. In particular, we show how to perform maximum a posteriori (MAP) inference by exploiting probabilistic dependencies within the domain while avoiding violations of fairness guarantees. Preliminary empirical evaluation shows that we are able to make both accurate and fair decisions.

1 Introduction

Over the past few years, AI and machine learning have become essential components in operations that drive the modern society, e.g., in financial, administrative, and educational spheres. *Discrimination* happens when qualities of individuals which are not relevant to the decision making process influence the decision. Delegating decision making to an automated process raises questions about discriminating against individuals with certain traits based on biases in the data. This is especially important when the decisions have the potential to impact the lives of individuals, for example, the decisions on granting loans, assigning credit, and employment.

Fairness is defined as the absence of discrimination in a decision making process. The goal of *fairness-aware* machine learning is to ensure that the decisions made by an algorithm do not discriminate against a population of individuals [14, 7, 16]. Fairness has been well studied in the social sciences and legal scholarship (for an in-depth review see [6]), and there is emerging work on fairness-aware ML within the AI and computer science communities. For example, fairness through awareness/Lipschitz property [11], individual fairness [27], statistical parity/group fairness [17], counterfactual fairness [19], demographic parity/disparate impact [14, 10], preference-based fairness [26], and equality of opportunity [16].

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

The existing work in fairness-aware machine learning is based on a definition of discrimination where a decision is influenced by an *attribute* of an individual. An attribute value upon which discrimination is based (such as gender, race, or religion) is called a *sensitive attribute*. The sensitive attribute defines a population of vulnerable individuals known as the *protected group*. A fair decision-making process treats the protected group the same as the *unprotected group*.

However, in many social contexts, discrimination is the result of complex interactions and can not be described solely in terms of attributes of an individual. For example, consider an imaginary scenario in an organization in which younger female workers who have older male supervisors have lower chances of promotion than their male counterparts.[1] This discrimination pattern involves two attributes of the individual (gender and age), a relationship with another individual (supervisor), and two attributes of the second individual. Addressing such complex cases poses two challenges. First, the concepts of discrimination and fairness need to be extended to capture not only attributes of individuals but also the relationships between them. Second, a process is required that ensures that fair decisions are made about individuals who are affected by such patterns. In this paper we address both of these challenges. We use first-order logic (FOL) to extend the notion of fairness to the relational setting. FOL is an expressive representation for relational problems which is also widely used for learning in relational domains. Moreover, we extend an existing framework for statistical relational learning [15] called probabilistic soft logic (PSL)[2] [5]. PSL combines logic and probability for learning and reasoning over uncertain relational domains. One of the most common reasoning tasks in PSL is called maximum a posteriori (MAP) inference, which is performed by finding the most probable truth values for unknowns over a set of given evidence. We develop a new MAP inference algorithm which is able to maximize the a posteriori values of unknown variables *subject to* fairness guarantees. An early version of this paper which this work builds upon and extends appeared in [13].

Our contributions are as follows: 1) we propose fairness-aware machine learning for the relational setting; 2) we extend PSL into a fairness-aware framework called FairPSL which can represent the logical definition of fairness; 3) we develop a new MAP inference algorithm which is able to maximize the posteriori values of unknown variables *subject to* fairness guarantees; 4) we empirically evaluate our proposed framework on synthetic data.

2 Motivation

Discrimination in social contexts have been studied in the field of social psychology [9, 8, 22]. There is a large literature on various aspects of relational bias in social contexts such as *in-group-out-group bias*, *gender bias*, and *ethnicity-based favoritism* that can result in discrimination. As an example, consider gender bias in the workplace that reflects stereotypically masculine criteria and male-based favoritism. Such gender bias typically places women in lower positions and negatively impacts their opportunities. Further, lack of women in leadership positions may affect the promotion of women and results in a glass ceiling that keeps women from rising beyond a certain level in the hierarchy. This scenario shows that considering protected attributes such as gender is not always sufficient to detect the source of bias and avoid discrimination, one also has to consider the relational information, in this case the organization hierarchy. Note that this can be generalized to any ingroup/outgroup scenario where the sensitive attribute could be race, religion, age, marital-status, etc.

The existing work on designing fair algorithms in machine learning exclusively focuses on *attribute-based fairness*, which is based on the following assumptions: First, there is an assumption that the individuals (sometimes referred to as units or entities) are independent and described by simple attribute vectors. Second, the group for which one wishes to ensure fairness (known as the *protected group*) is defined on the basis of some attribute values. Finally, there is a decision that is associated with each individual, and the goal is to ensure that members

[1] Of course, many other patterns may be possible: female bosses may promote female subordinates and discriminate against male workers, or male bosses may promote female employees. Our goal is to provide a general framework which is able to describe arbitrarily complex discrimination patterns.

[2] http://psl.linqs.org/

of the protected group are subject to a fair decision (we discuss different fairness measures in Section 4). We illustrate attribute-based fairness in the following example.

Example 1 (Loan Processing): A bank bases its decisions about granting a loan on attributes of the applicant. The goal is to decide whether to grant a loan to an applicant using a predictive model. The bank needs to ensure that the obey fair lending practices and ensure that gender, race, sexual orientation of applicants has no influence on the decision. In this scenario, the protected group is the historically disadvantaged applicants.

The current fairness-aware machine learning techniques are not capable of modeling relations and hence cannot be used to make the decision making model fair. However, in many decision making scenarios, especially in social and organizational settings, the domain is relational, and the protected group itself might be best represented using a relational definition. We illustrate this setting in the following scenario:

Example 2 (Performance Review): Consider an organization where decisions about the promotion of employees is based on two criteria: 1) an objective performance measure, and 2) the opinion of their direct and indirect managers above them. The opinions are inferred from the performance reviews which are collected periodically. Not every manager can submit a review for all its subordinates, this is especially the case for top-level managers who have a large number of subordinates. Hence, the opinions of managers are collectively inferred from the opinions of their sub-ordinates. However, some employees may be biased, and judge other employees unfavorably, by favoring employees who are similar to themselves (same gender, race, religion, etc.) over employees who are dissimilar. The organization needs to ensure that promotion of employees do not have any relational bias caused by in-group-out-group favoritism.

Example 2 describes a prediction problem over a database that consists of relations between employees. Such prediction tasks are best handled by techniques from the relational learning domain. To ensure fair prediction in such settings, we need to extend the notion of *attribute-based fairness* to *relational fairness*. Throughout this paper, we use the performance review problem as a running example for relational fairness.

3 Fairness Formalism

A representation that can describe different types of entities and different relationships between them is called relational. In this section, we use first-order logic to define relational fairness. We employ first-order logic as an expressive representation formalism which can represent objects and complex relationships between them. We start by defining an atom:

Definition 1 (Atom): An atom is an expression of the form $P(a_1, a_2, \ldots, a_n)$ where each argument $a_1, a_2, \ldots, a_n$ is either a constant or a variable. The finite set of all possible substitutions of a variable to a constant for a particular variable a is called its *domain* D_a. If all variables in $P(a_1, a_2, \ldots, a_n)$ are substituted by some constant from their respective domain, then we call the resulting atom a *ground atom*.

Example 3: In our loan processing problem (Example 1), we can represent applicants' attributes by atoms. For instance, atom $Female(v)$ indicates whether or not applicant v is female. Similarly, we can represent relations with atoms. In the performance review problem in Example 2 the atom $Manager(m, e)$ indicates whether or not employee m is a direct or indirect manager of employee e.

The relational setting provides the flexibility to express complex definitions with formulae.

Definition 2 (Formula): A formula is defined by induction: every atom is a formula. If α and β are formulae, then $\alpha \vee \beta$, $\alpha \wedge \beta$, $\neg\alpha$, $\alpha \rightarrow \beta$ are formulae. If x is a variable and α is a formula, then the quantified expressions of the form $\exists x\ \alpha$ and $\forall x\ \alpha$ are formulae.

To characterize groups of individuals based on a formula, we define the notion of *population*.

Definition 3 (Population): We denote formula F which has only one free variable v (i.e., other variables in F are quantified) by $F[v]$. The population defined by $F[v]$ is the set of substitutions of v for which $F[v]$ holds.

Example 4: Consider the formula $F[v] := \forall u$, *Manager*$(u, v) \rightarrow \neg$*SameGroup*(u, v). The population specified by this formula is the set of individuals all of whose managers belong to a group different from theirs.

The truth value of a formula is derived from the truth value of atoms that it comprises, according to the rules of logic. Each possible assignment of truth values to ground atoms is called an *interpretation*.

Definition 4 (Interpretation): An interpretation I is a mapping that associates a truth value $I(P)$ to each ground atom P. For Boolean truth values, I associates true to 1 and false to 0 truth values. For soft logic (see Definition 10) I maps each ground atom P to a truth value in interval $[0, 1]$.

In attribute-based fairness, it is assumed that there is a certain attribute of individuals, i.e, the sensitive attribute, that we do not want to affect a decision. Gender, race, religion and marital status are examples of sensitive attributes. Discrimination has been defined in social science studies as a treatment in favor or against a group of individuals given their sensitive attribute. This group of individuals is the protected group.

In a relational setting, both the sensitive attributes of an individual and their participation in various relations may have an undesired effect on the final decision. We characterize the protected group in a relational setting by means of a population. In practice, we are often interested in maintaining fairness for a specific population such as applicants, students, employees, etc. This population is then partitioned into the protected and unprotected groups. We define a *discriminative pattern* which is a pair of formulae to capture these groups: 1) $F_1[v]$: to specify the difference between the protected and unprotected groups and 2) $F_2[v]$: to specify the population over which we want to maintain fairness.

Definition 5 (Discriminative pattern): A discriminative pattern is a pair $DP[v] := (F_1[v], F_2[v])$, where $F_1[v]$ and $F_2[v]$ are formulae.

Example 5: The two formulae in the discrimination pattern $DP[v] := \big((\forall u$, *Manager*$(u, v) \rightarrow \neg$*SameGroup*$(u, v))$, *Employee*$(v)\big)$ specify two populations, namely all employees and those employees who belong to a group different from their managers.

Given the definition of the discriminative pattern, we have a rich language to define the scope of the protected and unprotected groups in a relational setting.

Definition 6 (Protected group): Given an interpretation I, the protected group is a population of the form:

$$PG := \{v : F_1[v] \wedge F_2[v]\}$$

which is defined as the set of all instances hold for variable v for which $F_1[v] \wedge F_2[v]$ is true under interpretation I, that is, $I(F_1[v] \wedge F_2[v]) = 1$. Similarly, the *unprotected group* is a population of the form:

$$UG := \{v : \neg F_1[v] \wedge F_2[v]\}$$

which is defined as the set of all instances hold for variable v for which $I(\neg F_1[v] \wedge F_2[v]) = 1$.

Example 6: The protected group of the discrimination pattern specified in Example 5 is $PG := \big\{v : \big(\forall u$, *Manager*$(u, v) \rightarrow \neg$*SameGroup*$(u, v)\big) \wedge$*Employee*$(v)\big\}$ and the unprotected group is $UG := \big\{v : \big(\exists u$, *Manager*$(u, v) \wedge$ *SameGroup*$(u, v)\big) \wedge$*Employee*$(v)\big\}$. This means our protected group is the set of employees belonging to a group different from their managers, and our unprotected group consists of other employees.

Discrimination is defined in terms of a treatment or decision that distinguishes between the protected and unprotected groups. Here we define the *decision* atom.

Definition 7 (Decision atom): A decision atom $d(v)$ is an atom containing exactly one variable v that specifies a decision affecting the protected group which is defined either by law or end-user.

Example 7: The decision atom $ToPromote(v)$ indicates whether or not v receives a promotion.

Note that the fairness formulation in this section is designed for the relational setting, however relational fairness subsumes the attribute-based fairness such that: a sensitive attribute is defined by an atom with one argument and $F_2[v]$ in discrimination pattern is *Applicant*(v). For example, discrimination pattern of our loan processing problem in Example 1 is of the form *DP* $:=$ (*Female*(v), *Applicant*(v)) that denotes female applicants as the protected group (i.e., $PG := \{v : \textit{Female}(v)\}$) and male applicants as the unprotected group (i.e., $UG := \{v : \neg \textit{Female}(v)\}$).

4 Fairness Measures

Over the past few years, many fairness measures have been introduced [24]. An important class of these measures are *group fairness* measures which quantify the inequality between different subgroups. Some of the most popular measures in this class include *equal opportunity*, *equalized odds*, and *demographic parity* [16]. In this paper we restrict our focus to the latter. In an attribute-value setting, demographic parity means that the decision should be independent of the protected attributes. Assume that binary variables A and C denote the decision and protected attributes, and the preferred value of A is one. Demographic parity requires that:

$$P(A = 1|C = 0) = P(A = 1|C = 1)$$

We will now generalize this measure to the relational setting using the notations defined in Section 3. Let a and c denote the counts of denial (i.e., negative decisions) for protected and unprotected groups, and n_1 and n_2 denote their sizes, respectively. Given the decision atom $d(v)$, discriminative pattern *DP*$(F_1[v], F_2[v])$, and interpretation I, these counts are computed by the following equations:

$$a \equiv \sum_{v \in D_v} I\big(\neg d(v) \wedge F_1[v] \wedge F_2[v]\big) \tag{5}$$

$$c \equiv \sum_{v \in D_v} I\big(\neg d(v) \wedge \neg F_1[v] \wedge F_2[v]\big) \tag{6}$$

$$n_1 \equiv \sum_{v \in D_v} I\big(F_1[v] \wedge F_2[v]\big) \tag{7}$$

$$n_2 \equiv \sum_{v \in D_v} I\big(\neg F_1[v] \wedge F_2[v]\big) \tag{8}$$

The proportions of denying for protected and unprotected groups are $p_1 = \frac{a}{n_1}$ and $p_2 = \frac{c}{n_2}$, respectively. There are a number of data-driven measures [20] which quantify demographic disparity and can be defined in terms of p_1 and p_2:

- **Risk difference**: $RD = p_1 - p_2$, also known as absolute risk reduction.
- **Risk Ratio**: $RR = \frac{p_1}{p_2}$, also known as relative risk.
- **Relative Chance**: $RC = \frac{1-p_1}{1-p_2}$ also, known as selection rate.

These measures have been used in the legal systems of European Union, UK, and US [1, 2, 3]. Notice that RR is the ratio of the proportion of benefit denial between the protected and unprotected groups, while RC is the ratio of the proportion of benefit granted. Finally, we introduce the notion of δ-fairness.

Definition 8 (δ-fairness): If a fairness measure for a decision making process falls within some δ-window, then the process is δ-*fair*. Given $0 \leq \delta \leq 1$, the δ-windows for measures RD/RR/RC are defined as:

$$\begin{aligned} -\delta \leq & RD \leq \delta \\ 1-\delta \leq & RR \leq 1+\delta \\ 1-\delta \leq & RC \leq 1+\delta \end{aligned}$$

To overcome the limitations of attribute-based fairness, we introduce a new statistical relational learning (SRL) framework [15] suitable for modelling fairness in relational domain. In the next section, we review probabilistic soft logic (PSL). We then extend PSL with the definition of relational fairness introduced above in Section 6. Our fairness-aware framework, "FairPSL", is the first SRL framework that performs fair inference.

5 Background: Probabilistic Soft Logic

In this section, we review the syntax and semantics of PSL, and in the next section we extend MAP inference in PSL with fairness constraints to define MAP inference in FairPSL.

PSL is a probabilistic programming language for defining hinge-loss Markov random fields [5]. Unlike other SRL frameworks whose atoms are Boolean, atoms in PSL can take continuous values in the interval $[0, 1]$. PSL is an expressive modeling language that can incorporate domain knowledge with first-order logical rules and has been used successfully in various domains, including bioinformatics [23], recommender systems [18], natural language processing [12], information retrieval [4], and social network analysis [25], among many others.

A PSL model is defined by a set of first-order logical rules called *PSL rules*.

Definition 9 (PSL rule): a PSL rule r is an expression of the form:

$$\lambda_r : T_1 \wedge T_2 \wedge \ldots \wedge T_w \rightarrow H_1 \vee H_2 \vee \ldots \vee H_l \tag{9}$$

where $T_1, T_2, \ldots, T_w, H_1, H_2, \ldots, H_l$ are atoms or negated atoms and $\lambda_r \in \mathbb{R}^+ \cup \infty$ is the weight of the rule r. We call $T_1 \wedge T_2 \wedge \ldots \wedge T_w$ the body of r (r_{body}), and $H_1 \vee H_2 \vee \ldots \vee H_l$ the head of r (r_{head}).

Since atoms in PSL take on continuous values in the unit interval $[0, 1]$, next we define soft logic to calculate the value of the PSL rules under an interpretation I.

Definition 10 (Soft logic): The ($\tilde{\wedge}$) and ($\tilde{\vee}$) and negation ($\tilde{\neg}$) are defined as follows. For $m, n \in [0, 1]$ we have: $m \tilde{\wedge} n = \max(m + n - 1, 0)$, $m \tilde{\vee} n = \min(m + n, 1)$ and $\tilde{\neg} m = 1 - m$. The $\tilde{}$ indicates the relaxation over Boolean values.

The probability of truth value assignments in PSL is determined by the rules' *distance to satisfaction*.

Definition 11 (The distance to satisfaction): The distance to satisfaction $d_r(I)$ of a rule r under an interpretation I is defined as:

$$d_r(I) = \max\{0, I(r_{body}) - I(r_{head})\} \tag{10}$$

$R1$	λ_1	: *IsQualified*$(e) \rightarrow$ *HighPerformance*(e)
$R2$	λ_1	: $\neg$*IsQualified*$(e) \rightarrow \neg$*HighPerformance*(e)
$R3$	∞	: *PositiveReview*$(e1, e2) \rightarrow$ *PositiveOpinion*$(e1, e2)$
$R4$	∞	: $\neg$*PositiveReview*$(e1, e2) \rightarrow \neg$*PositiveOpinion*$(e1, e2)$
$R5$	λ_1	: *PositiveOpinion*$(e1, e2) \wedge$ *Manager*$(m, e1) \rightarrow$ *PositiveOpinion*$(m, e2)$
$R6$	λ_1	: $\neg$*PositiveOpinion*$(e1, e2) \wedge$ *Manager*$(m, e1) \rightarrow \neg$*PositiveOpinion*$(m, e2)$
$R7$	λ_1	: *PositiveOpinion*$(m, e) \wedge$ *Manager*$(m, e) \rightarrow$ *IsQualified*(e)
$R8$	λ_1	: $\neg$*PositiveOpinion*$(m, e) \wedge$ *Manager*$(m, e) \rightarrow \neg$*IsQualified*(e)
$R9$	λ_1	: $\neg$*ToPromote*(e)
$R10$	∞	: *IsQualified*$(e) \rightarrow$ *ToPromote*(e)
$R11$	∞	: $\neg$*IsQualified*$(e) \rightarrow \neg$*ToPromote*(e)

Table 1: A simplified PSL model for the *Performance Reviewing* problem

By using Definition 10, one can show that the closer the interpretation of a grounded rule r is to 1, the smaller its distance to satisfaction. A PSL model induces a distribution over interpretations I. Let R be the set of all grounded rules, then the probability density function is:

$$f(I) = \frac{1}{Z} \exp[-\sum_{r \in R} \lambda_r (d_r(I))^p] \tag{11}$$

where λ_r is the weight of rule r, $Z = \int_I \exp[-\sum_{r \in R} \lambda_r (d_r(I))^p]$ is a normalization constant, and $p \in \{1, 2\}$ provides a choice of two different loss functions, $p = 1$ (i.e., linear), and $p = 2$ (i.e, quadratic). These probabilistic models are instances of hinge-loss Markov random fields (HL-MRF) [5]. The goal of maximum a posteriori (MAP) inference is to find the most probable truth assignments I_{MPE} of unknown ground atoms given the evidence which is defined by the interpretation I. Let X be all the evidence, i.e., X is the set of ground atoms such that $\forall x \in X, I(x)$ is known, and let Y be the set of ground atoms such that $\forall y \in Y, I(y)$ is unknown. Then we have

$$I_{MAP}(Y) = arg \max_{I(Y)} P(I(Y)|I(X)) \tag{12}$$

Maximizing the density function in Equation 11 is equivalent to minimizing the weighted sum of the distances to satisfaction of all rules in PSL.

Example 8: The simplified PSL model for the performance reviewing problem in Example2 is given in Table 1. The goal of MAP inference for this problem is to infer employees to promote. We simplified the model by assigning the same weight to all soft rules (i.e., $\lambda_i = 1$ where $i = \{1, 2, 5, 6, 7, 8, 9\}$). Below we explain the meaning of each rule in the model.

Rule $R1$ indicates that qualified employees have high performance and similarly rule $R2$ expresses that a negative qualification of employees is derived from their low performance. Rules $R5$ and $R6$ presents the propagation of opinion from bottom to top of the organizational hierarchy, i.e., managers have similar opinions towards employees given the opinions of their sub-ordinate managers. And rules $R7$ and $R8$ indicate that the positive/negative opinion of direct/indirect managers derive from the qualification of an employee. Rule $R9$ indicates the prior that not all employees get promoted. We also have four hard constraints (i.e., rules $R3$, $R4$, $R10$ and $R11$) where the weight of the rules are ∞. Rules $R3$ and $R4$ indicate that submitted positive/negative reviews should reflect positive/negative opinions. And two rules $R10$ and $R11$ show that a highly qualified employee should get promoted.

6 Fairness-aware PSL (FairPSL)

The standard MAP inference in PSL aims at finding values that maximize the conditional probability of unknowns. Once a decision is made according to these values, one can use the fairness measure to quantify the degree of discrimination. A simple way to incorporate fairness in MAP inference is to add the δ-fairness constraints to the corresponding optimization problem.

Consider risk difference, *RD*, where $RD \equiv \frac{\mathbf{a}}{n_1} - \frac{\mathbf{c}}{n_2}$. The δ-fairness constraint $-\delta \leq RD \leq \delta$ can be encoded as the following constraints:

$$n_2\mathbf{a} - n_1\mathbf{c} - n_1n_2\delta \leq 0 \tag{13}$$
$$n_2\mathbf{a} - n_1\mathbf{c} + n_1n_2\delta \geq 0 \tag{14}$$

Similarly, from $RR \equiv \frac{\mathbf{a}/n_1}{\mathbf{c}/n_2}$ and the δ-fairness constraint $1 - \delta \leq RR \leq 1 + \delta$ we obtain:

$$n_2\mathbf{a} - (1+\delta)n_1\mathbf{c} \leq 0 \tag{15}$$
$$n_2\mathbf{a} - (1-\delta)n_1\mathbf{c} \geq 0 \tag{16}$$

And finally, $RC \equiv \frac{1-\mathbf{a}/n_1}{1-\mathbf{c}/n_2}$ and the δ-fairness constraint $1 - \delta \leq RC \leq 1 + \delta$ gives:

$$-n_2\mathbf{a} + (1+\delta)n_1\mathbf{c} - \delta n_1n_2 \leq 0 \tag{17}$$
$$-n_2\mathbf{a} + (1-\delta)n_1\mathbf{c} + \delta n_1n_2 \geq 0 \tag{18}$$

A primary advantage of PSL over similar frameworks is that its MAP inference task reduces to a convex optimization problem which can be solved in polynomial time. To preserve this advantage, we need to ensure that the problem will remain convex after the addition of δ-fairness constraints.

Theorem 1: The following condition is sufficient for preserving the convexity of MAP inference problem after addition of δ-fairness constraints: The formulae $F_1[v]$ and $F_2[v]$ do not contain an atom $y \in Y$ and all atoms in $F_1[v]$ and $F_2[v]$ have values zero or one.

Proof: Since $I(F_1[v])$ and $I(F_2[v])$ do not depend on $I(Y)$, the values n_1 and n_2 are constants that can be computed in advance. Let us define the sets $D_v^a = \{v \in D_v : F_1[v] \wedge F_2[v] \text{ is true}\}$ and $D_v^c = \{v \in D_v : \neg F_1[v] \wedge F_2[v] \text{ is true}\}$. Since $F_1[v]$ and $F_2[v]$ can be only zero or one, we can rewrite the equations 5 and 6 as:

$$\mathbf{a} = \sum_{v \in D_v^a} I(\neg d(v)) = |D_v^a| - \sum_{v \in D_v^a} I(d(v))$$
$$\mathbf{c} = \sum_{v \in D_v^c} I(\neg d(v)) = |D_v^c| - \sum_{v \in D_v^c} I(d(v))$$

which indicates that $\mathbf{a}$ and $\mathbf{c}$ can be expressed as linear combinations of variables in the optimization problem. This means that constraints 13-18 are linear. Hence, addition of these constraints preserves the convexity of the optimization problem.

7 Experiments

We show the effectiveness of FairPSL by performing an empirical evaluation. We investigate two research questions in our experiments:

Q1 What is the effect of the fairness threshold δ on the fairness measures $RD/RC/RR$?

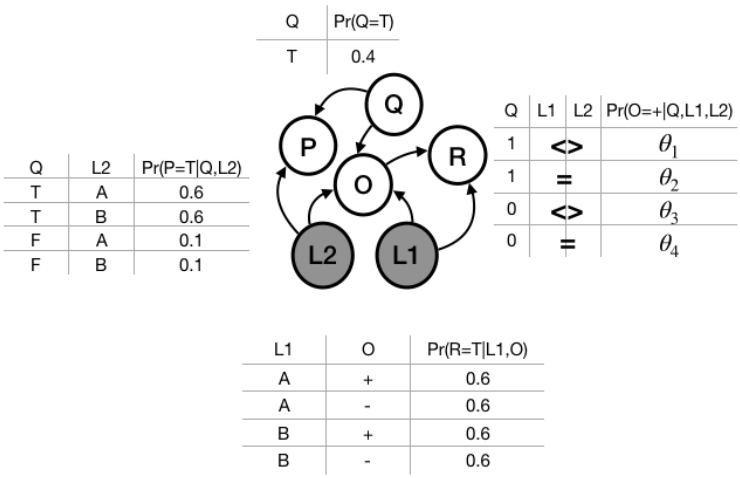
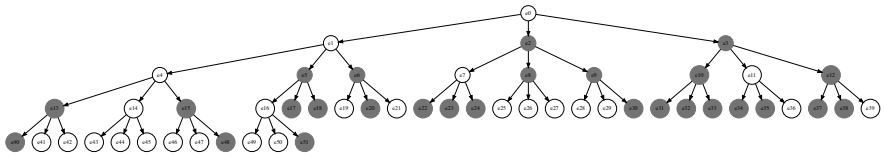


Figure 1: The model used for generating the datasets. There are four binary random variables, P, Q, O, and R. **P**: indicates whether or not the employee has high performance; **Q**: indicates whether or not an employee has high qualification; **O**: indicates whether or not the colleague submits the positive opinion towards the employee; **R**: indicates whether or not the colleague has a positive opinion towards the employee; **L1, L2**: indicates the label of the review provider and review receiver (observed).

Figure 2: An example of an organizational hierarchy with five levels and 50 employees with k=3. Each employee either has label A (shown with grey) or B (shown with white).



Q2 How is decision quality affected by imposing δ-fairness constraints?

Note that although we present the result for specific parameters of the framework in this section, we ran extensive analysis and the results we present are representative. We implemented the MAP inference routines of PSL and FairPSL in Python, using Gurobi-8.1[3] as the backend solver. The FairPSL code, code for the data generator and data is publicly available[4].

7.1 Data generation

We evaluate the FairPSL inference algorithm on synthetic datasets representing the performance review scenario (introduced in Example 2). The organization hierarchy is generated synthetically. The organization hierarchy generator is parameterized by two numbers: the number of employees in the organization (n) and the number of employees managed by each manager (k). Each employee is randomly assigned with a label *A* or *B*. An examples organization hierarchy with n=50 and k=3 is shown in Figure 2.

For each employee, we use the generative model of Figure 1 to draw assignments for all the random variables. We assume that only 40% of employees are qualified for promotion and regardless of their labels, employees submit only 60% of their opinions. In addition, due to various personal and environmental factors, only 60% of high quality employees perform well while 10% of low quality employees also perform well regardless of their labels. Note that these numbers are not specific and just chosen for the framework to serve as a representative setting and a proof of concept. The conditional probability table for the opinion variable O is parameterized by four values $(\theta_1, \theta_2, \theta_3, \theta_4)$ which together determine the degree of discrimination against the protected group. Since other parameters in the Bayesian network did not have a direct effect on the degree of discrimination, we fixed them to arbitrary values.

The results presented in this section are based on an organization hierarchy with 100 employees where $k = 5$. However, the results of the framework are not sensitive to the settings as we test the framework

[3] www.gurobi.com

[4] https://github.com/gfarnadi/FairPSL

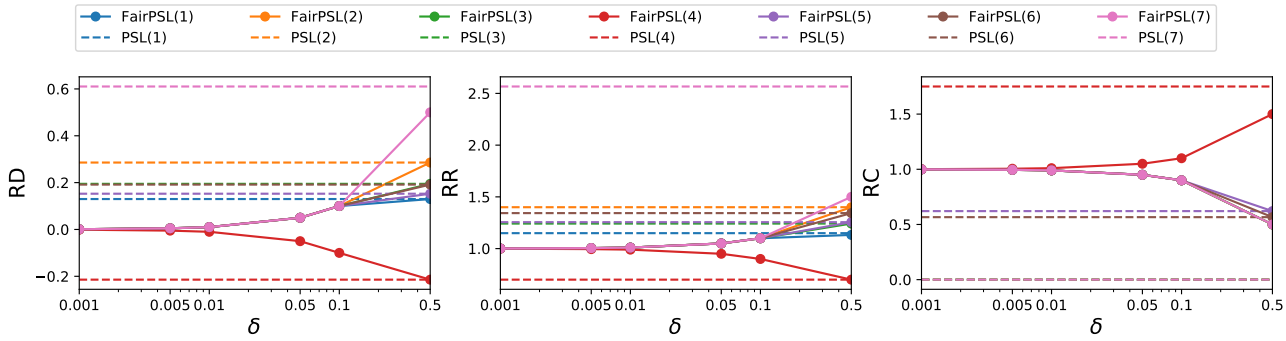


Figure 3: Fairness score of predictions obtained by MAP inference of PSL and FairPSL, according to the fairness measures *RD*, *RR*, and *RC*. The labels of datasets are mentioned with parenthesis next to the inference method. The FairPSL values of each measure are obtained by adding the δ-fairness constraint of that measure.

with various organization sizes ranging from 50 to 500 employees and various degree for k ranging from 3 to 10. We generated seven datasets given the organization hierarchy using different values for the θ parameters: $(0.0, 1.0, 0.0, 0.0)$, $(0.33, 1.0, 0.0, 0.0)$, $(0.66, 1.0, 0.0, 0.0)$, $(1.0, 1.0, 0.0, 0.0)$, $(1.0, 1.0, 0.0, 0.33)$, $(1.0, 1.0, 0.0, 0.66)$, $(1.0, 1.0, 0.0, 1.0)$.

In the first three settings the discrimination originates from negative opinions towards qualified outgroup employees. The first setup is an extreme case where the opinion towards outgroup employees is always negative. The discrimination in the last three settings originates from positive opinions towards unqualified ingroup employees. The last setup is an extreme case where the opinion towards ingroup employees is always positive. The fourth setup represent unbiased opinions where employees are treated similarly based on their qualification.

MAP Inference We use the model presented in Table 1 for MAP inference in PSL and FairPSL (recall that in FairPSL, the δ-fairness constraints corresponding to one of the fairness measures are also added to the model). The observed atoms are *Manager(m,e)*, *PositiveReview(e1,e2)* and labels of all employees. The truth values for all other atoms are obtained via MAP inference. We use the truth values obtained for the decision atoms *ToPromote(e)* to compute the fairness measures. We defined the discriminative pattern, and the protected and unprotected groups of this problem in Section 3.

7.2 Evaluation results

To answer **Q1**, we run the MAP inference algorithm of PSL and FairPSL on seven synthetic datasets. We run the MAP inference of FairPSL multiple times on each dataset: For each of the three fairness measures, we add the corresponding δ-fairness constraint with five thresholds $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$.

Figure 3 shows the fairness score of predictions in terms of the three fairness measures. As expected, tighter δ-fairness constraints lead to better scores. Note that the best possible score according to RD is 0, as it computes a difference. Since RR and RC compute ratios, the best possible score according to these measures is 1. In our experiments, with any of these measures, taking $\delta = 0.001$ pushes the score of predictions to its limit.

The δ-fairness constraints modify the optimization problem of MAP inference by reducing the feasible region to solutions that conform with fairness guarantees. Research question **Q2** is concerned with the effect of this reduction on the accuracy of predictions. Note that decision quality is the same as the accuracy of predictions. To answer this question, we compare the inferred values for the decision atoms *ToPromote(e)* against their actual values. These values are extracted from the known values of *IsQualified(e)* according to rules 11 and 12 in Table 1. Figure 4 shows the area under the curve of the receiver operating characteristic (AUC) of predicting the decision variable in three groups, namely the protected group, the unprotected group (i.e., promotion of the employees

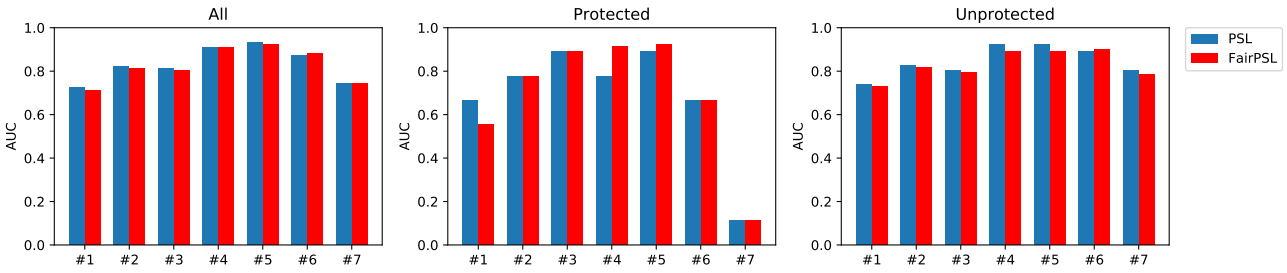


Figure 4: AUC score of predictions for truth values of unknown atoms *ToPromote(e)* using MAP inference of PSL and FairPSL with δ-fairness constraints RD with $\delta = 0.001$.

who have in-group managers), and all employees. By doing so, we make sure that our fairness constraints do not propagate bias towards either of the populations. Since the results of FairPSL with δ-fairness constraints RR and RC are very similar to the results of RD, we only report the latter here.

According to Figure 4, the results of both PSL and FairPSL in all seven datasets are close to each other. Note that although fairness may impose a cost in terms of overall accuracy, FairPSL often improves the accuracy of the protected class. Sometimes the overall predictions of FairPSL are even slightly better than PSL (e.g., dataset 6 and 7). As expected, the accuracy of the fourth setting where the opinions are unbiased are similar in both PSL and FairPSL. We observe that prediction of MAP inference for both FairPSL and PSL are similar, thus, in these settings at least, FairPSL guarantees fairness without hurting accuracy. Further investigation is required on the effect of the various ranges of discrimination (i.e., θ_1, θ_2, θ_3, θ_4) on the prediction results of FairPSL.

We also generate various types of organizations in which labels are not uniformly distributed, e.g., one population only occurs at the bottom levels of an organization. While we did not observe any differences in the behavior of our method with respect to accuracy and fairness measure, we found that the degree of discrimination is higher in such organizations. Further investigations on the structure of an organization on discrimination is an interesting direction for future research.

8 Conclusion and Future Directions

Many applications of AI and machine learning affect peoples' lives in important ways. While there is a growing body of work on fairness in AI and ML, it assumes an individualistic notion of fairness. In this paper, we have proposed a general framework for relational fairness which includes both a rich language for defining discrimination patterns and an efficient algorithm for performing inference subject to fairness constraints. We show our approach enforces fairness guarantees while preserving the accuracy of the predictions.

There are many avenues for expanding on this work. For example, here we assumed that the discriminative pattern is given, however an automatic mechanism to extract discriminatory situations hidden in a large amount of decision records is an important and required task. Discrimination discovery has been studied for attribute-based fairness [21]. An interesting next step is discrimination pattern discovery in relational data.

Acknowledgements

This work is supported by the National Science Foundation under Grant Numbers CCF-1740850 and IIS-1703331. Golnoosh Farnadi and Behrouz Babaki are supported by postdoctoral scholarships from IVADO through the Canada First Research Excellence Fund (CFREF) grant.

References

[1] European union legislation. (a) racial equality directive, 2000; (b) employment equality directive, 2000; (c) gender employment directive, 2006; (d) equal treatment directive (proposal), 2008.

[2] UK legislation. (a) sex discrimination act, 1975, (b) race relation act, 1976.

[3] United nations legislation. (a) universal declaration of human rights, 1948, (c) convention on the elimination of all forms of racial discrimination, 1966, (d) convention on the elimination of all forms of discrimination against women, 1979.

[4] Duhai Alshukaili, Alvaro A. A. Fernandes, and Norman W. Paton. Structuring linked data search results using probabilistic soft logic. In *International Semantic Web Conference (1)*, volume 9981 of *Lecture Notes in Computer Science*, pages 3–19, 2016.

[5] Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. *Journal of Machine Learning Research*, 18:109:1–109:67, 2017.

[6] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California Law Review*, 104:671, 2016.

[7] Danah Boyd, Karen Levy, and Alice Marwick. The networked nature of algorithmic discrimination. In *Data and discrimination: Collected essays*, pages 53–57. 2014.

[8] Marilynn B Brewer. In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological bulletin*, 86(2):307, 1979.

[9] Marilynn B Brewer. The social psychology of intergroup relations: Social categorization, ingroup bias, and outgroup prejudice. *Social Psychology: Handbook of Basic Principles*, 2007.

[10] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *CoRR*, abs/1703.00056, 2017.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *ITCS*, pages 214–226. ACM, 2012.

[12] Javid Ebrahimi, Dejing Dou, and Daniel Lowd. Weakly supervised tweet stance classification by relational bootstrapping. In *EMNLP*, pages 1012–1017. The Association for Computational Linguistics, 2016.

[13] Golnoosh Farnadi, Behrouz Babaki, and Lise Getoor. Fairness in relational domains. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 108–114. ACM, 2018.

[14] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pages 259–268. ACM, 2015.

[15] Lise Getoor and Ben Taskar. *Introduction to Statistical Relational Learning*. MIT press Cambridge, 2007.

[16] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NIPS*, pages 3315–3323, 2016.

[17] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *ICDMW*, pages 643–650. IEEE Computer Society, 2011.

[18] Pigi Kouki, Shobeir Fakhraei, James R. Foulds, Magdalini Eirinaki, and Lise Getoor. Hyper: A flexible and extensible probabilistic framework for hybrid recommender systems. In *RecSys*, pages 99–106. ACM, 2015.

[19] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *NIPS*, pages 4069–4079, 2017.

[20] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. A study of top-k measures for discrimination discovery. In *SAC*, pages 126–131. ACM, 2012.

[21] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. The discovery of discrimination. In *Discrimination and Privacy in the Information Society*, volume 3 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*, pages 91–108. Springer, 2013.

[22] Cecilia L Ridgeway and Shelley J Correll. Unpacking the gender system: A theoretical perspective on gender beliefs and social relations. *Gender & society*, 18(4):510–531, 2004.

[23] Dhanya Sridhar, Shobeir Fakhraei, and Lise Getoor. A probabilistic approach for collective similarity-based drug-drug interaction prediction. *Bioinformatics*, 32(20):3175–3182, 2016.

[24] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

[25] Robert West, Hristo S. Paskov, Jure Leskovec, and Christopher Potts. Exploiting social network structure for person-to-person sentiment analysis. *TACL*, 2:297–310, 2014.

[26] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, Krishna P. Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *NIPS*, pages 228–238, 2017.

[27] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML (3)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 325–333. JMLR.org, 2013.

Fairness in Practice: A Survey on Equity in Urban Mobility

An Yan, Bill Howe
University of Washington
{yanan15,billhowe}@uw.edu

1 Introduction

More than 54 percent of the world's population lives in urban areas [1]. Predicting dynamic urban activities such as energy consumption, air pollution, public safety, and traffic flows has become a fundamental task for improving the quality of human life. Urban mobility is closely intertwined with these problems, and is therefore a major determinant of quality of life [2], crucial to employment opportunities and access to resources such as education and health care [3].

Evidence suggests that residents of low-income and minority neighborhoods are concentrated away from economic opportunity and public resources [4]. Injustice of transportation services experienced by these residents further reinforces social exclusion as the availability and quality of transportation impact a person's access to opportunities [5, 6, 7, 8]. For example, one study revealed that living in neighborhoods with longer commute times is associated with lower employment rates of younger generations [9]. As a result, transportation equity issues have motivated government agencies to develop extensive multimodal transportation networks[6].

New mobility is about emerging transportation modes, including but not limited to car-sharing, bike-sharing, and ride-hailing or Transportation Network Companies (TNCs) [10]. New mobility services provide technology-based, on-demand, and affordable alternatives to traditional means. These services offer a chance to address persistent equity issues in transportation. However, new mobility services also bring new equity concerns. For example, people without internet service, smart phones, or credit cards are not able to get access to the services. Moreover, studies show that algorithms or human beings that distribute app-based mobility services may discriminate against people of color [11].

This paper reviews the methods and findings of mobility equity studies, with a focus on new mobility. The paper is structured as follows: Section 2 presents the background of transportation equity. Section 3 summarizes the main findings from current equity studies for mobility systems, with a brief discussion on future research. Section 4 reviews the commonly used methods for evaluating the equity of mobility service provision and usage and considers strengths and weaknesses. Section 5 discusses the relationship between the transportation equity community and the fairness in machine learning community. Section 6 concludes the paper.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

This work is supported by the National Science Foundation under grant NSF BIGDATA-1740996.

2 Equity in Mobility Systems: Background

Automated decision systems powered by machine learning and big data have been widely employed in many applications including credit scoring, criminal justice, online advertising, employment, etc [12, 13, 14, 15]. These systems have been hailed as efficient, objective, and accurate alternatives to human decision makers [16]. However, increasing evidence has shown that data-driven systems contain biases. For example, Google's image recognition system wrongly identified black users as gorillas [17]. Amazon's same-day delivery services excluded predominantly black neighborhoods in many cities to a varying degree [18].

Even if the algorithms themselves are well-intentioned, they can replicate and amplify human biases encoded in the data, thus resulting in unequal distribution of impacts across different demographic groups [12, 19, 20]. This effect is due to machine learning algorithms seeking to fit the training data as closely as possible to make accurate predictions. The process of learning also "accurately" captures historical signals of discrimination. In 2017, Caliskan et al. [21] found that an influential language corpus [22] generated by machine learning algorithms accurately reproduced historic biases. The corpus reflects societal stereotypes such as female names are more associated with family while male names are more associated with career. Not only do algorithms pick up discrimination in data, they also magnify them [23]. This effect is often due to the underrepresentation of minority groups in training data, leading to higher error rates for the minorities. One study [24] revealed that a widely-used predictive policing tool, PredPol, would reinforce the bias in the police-recorded, resulting in disproportionate policing of minority communities.

The heightened concerns about automated decision systems concentrate not only on discrimination, but also on a range of related issues, including transparency, privacy, and accountability [25]. These issues often intertwine and conflict with one another in practice. In the context of automatic decision systems, transparency is about the openness and understandability of data and models and accountability is about being responsible for the decisions [26]. Transparency is a critical prerequisite for accountability. In the absence of concrete evidence of intentional discrimination, it is difficult to hold an individual or organization accountable for biased decisions.

In practice, transparency for automatic decision systems is not easily achievable. Burrell [27] summarized three types of barriers to transparency: 1) intrinsic opacity, where some algorithms such as deep learning models are difficult to understand and interpret; 2) illiterate opacity, which says the general public may lack the expertise to understand the algorithms; and 3) intentional opacity, which is often resulted from intellectual property protection of the algorithm developers.

2.1 Definitions of transportation equity

Equity in the context of mobility has been studied independently since well before the recent interest in generalized fairness methods for machine learning. These efforts suggest that domain-specific and context-sensitive approaches should be incorporated into any fairness-aware ML system. Equity for mobility is the fair distribution of transportation costs and benefits, among current (and future) members of society [5].

There are mainly two perspectives from which to examine equity: horizontal equity and vertical equity. *Horizontal equity* (also called fairness and egalitarianism) is concerned with providing equal resources to individual or groups considered equal in ability and need, which means the public policies avoid favoring one individual or group over another. Horizontal equity suggests that those who pay more should receive superior services.

Vertical equity (also referred to as social justice, environmental justice, and social inclusion) is concerned with allocating resources to individuals or groups that differ in income, social class, mobility need, or ability. It advocates that public policies favor disadvantaged groups by providing discounts or special services, therefore compensating for overall inequities. One way to evaluate vertical equity is *equity of opportunity*, meaning that disadvantaged groups should have adequate access to transportation resources. Equity of opportunity is usually measured by access to services. In contrast, "*equity of outcome*" is usually measured by the actual usage of the

systems across groups. There is a general agreement about the goal of equity of opportunity, but less agreement about equity of outcome [5, 28, 29].

There are other ways to define transportation equity. Social equity indicates the differences between socioeconomic groups. Spatial equity refers to the differences in transport services among geographic regions [30]. These different definitions often overlap or conflict with each other. For example, horizontal equity requires the users to pay for what they get, whereas vertical equity prioritizes the needs of disadvantaged groups such as the low-income or ethnic minorities in the form of discounts [31].

2.2 Evaluation of mobility equity

Mobility equity research addresses a wide range of issues, including, for example, economic studies on how transportation is subsidized and taxed, and operational studies on how negative impacts of transportation systems are distributed among different groups [6]. Litman [5] proposed four variables to consider when performing any equity evaluation.

- Type of equity: horizontal equity or vertical equity
- Impact (costs and benefits) categories: public facilities and services, user costs and benefits (e.g., taxes and fares), service quality (e.g., public transportation service quality including frequency, speed, safety, reliability, comfort, etc.), external impacts (e.g., air pollution), economic impacts (e.g., access to employment), and regulation and enforcement (e.g., parking regulations)
- Measurement unit: per capita, per unit of travel (e.g., per trip), or per dollar.
- Categorization of people: demographics (e.g., age, household type, race), income class, ability, location, mode (e.g., pedestrians, public transit), industry (e.g., freight, public transit), and trip type (e.g., commutes)

This paper focuses on the equity of new mobility systems service provision and usage across different social-economic, demographic, or geographic groups.

3 Findings from Equity Research in Mobility Systems

We describe findings in the literature across 1) public transportation, and 2) new mobility services.

Public transportation Transportation equity has long been a major concern of governmental agencies, researchers, and the general public [5, 6, 7]. Despite the tremendous investment in transportation system development and progress in transportation equity research, there are still many long-standing equity issues resulted from unequal distributions of transport resources across different socioeconomic groups and spatial regions [32, 33, 8, 34]. A number of studies have found out that an uneven urban development has resulted in a lack of public transport supply for disadvantaged groups. For example, Vasconcellos [35] pointed out in Brazil, road systems are developed in a radial pattern. Low-income residents usually settled in fringes of the city with irregular pavements or hilly areas that are subject to landslides. The urban centers with good public services are mostly occupied by the high-income people. Similarly, Ricciardi [8] found that there is an unequal spatial distribution of public transport resources in two Australian cities. Their analysis showed that 70% of Perth's population shares one third of the public transit supply. Moreover, three socially disadvantaged groups — the elderly, low-income, and no-car households have less access to public transport services compared to the overall population. Some studies also showed that the economic burden and negative climate impacts of transportation systems is disproportionately imposed on disadvantaged people [33, 30, 36]. In recognizing these issues, many cities now have incorporated social equity into urban transportation planning. However, one study found that

social equity goals are often not translated into clear and actionable items and there is a lack of appropriate methods for assessing their achievements [37, 5]. Current literature on equity in public transport suggests that disadvantaged groups as a whole experience inequitable access to public transport services but suffer from significant negative impacts from the transportation systems.

3.1 New mobility

Bikeshare A number of researchers have studied equity in bikeshare systems. Several studies found that bikeshare stations were typically located in urban centers with high population density [38, 39], and there was a lack of stations in low-income areas. In an assessment of bikeshare systems in seven US cities, Ursaki et al. found significant differences in the race, education level, and income of population inside and outside bike share service areas in four cities [40]. Other studies also indicated that in North America, advantaged groups tend to have more access to docked bikeshare than disadvantaged groups [41]. Recently, free-floating (dockless) bike share systems have been introduced in several major cities in China and the United States [42, 43, 44]. Free-floating bikeshare systems may have different equity landscape from docked systems. There are no stations in the city, therefore there are no fixed service areas. In this way, access to bikes are transient and largely dependent on the placement of individual bikes, which is driven by user demand and companies' bike rebalancing strategies. As free-floating bikeshare systems are fairly new, the impact on equity are unclear. In examining access equity of dockless bikes in Seattle, Mooney et al. found out that more college-educated and higher-income residents have access to more bikes. They also found out that bike demand is highly correlated with rebalancing destinations [43], suggesting that the companies themselves are accountable for equity issues that arise.

Equal access to bikeshare does not imply equity of actual usage. Several studies found inequalities in the usage of bikeshare systems [45, 46]. For example, Daddio et al. [45] found a negative association between station-level usage with non-white population in Washington, D.C. The disparities in use partially stem from the inequalities of access, but there are many other factors that inhibit bikeshare use among disadvantaged groups. McNeil et al. found out that the biggest barrier to bikeshare is traffic safety, regardless of race or income [47]. Lower-income people of color have more concerns about costs of membership and more misconceptions about bikeshare than higher-income white people. Another study [48] found that credit card requirement, lack of computer access, annual subscription fee, and lack of bike lanes etc. are reported by low-income residents as barriers to bikeshare. Shaheen et al. [6] identified five types of barriers to use bikeshare including spatial, temporal, economic, physiological, and social barriers, and provided policy recommendations. Overall, current literature suggests that disparities exist in the access and use of bikeshare systems.

Ride-hailing Ride-hailing can potentially redefine car access, mitigating the mobility divide resulted from car ownership [49]. But the equity of ride-hailing services remains unclear. Several studies found that the service quality in terms of waiting times is not necessarily associated with the average income or minority fraction of pickup locations [50, 51]. A recent study [49] found that users in low-income neighborhoods actually use Lyft more frequently than users in high-income neighborhoods in Los Angeles. The findings of this study suggest that Lyft may provide automobile alternatives to neighborhoods with less access to cars. These findings contradict the conclusions from other studies, which suggest that TNCs provide poor services to low-income neighborhoods [52].

Another thread of research examined the discrimination in TNCs. Ge et al. [11] found out that TNC drivers discriminate against African American riders, resulting in longer waiting times and higher trip cancellation rate in Boston and Seattle. Similarly, Brown [49] found that black riders experienced four percent higher trip cancellation rates and longer waiting times than white riders in Los Angeles. Middleton [53] examined rider-to-ride discrimination in ridesharing. Results showed that white respondents in majority white counties are more likely to hold discriminatory attitudes towards riders of other races or class. A few studies investigated the relationship between TNCs and public transit. For example, Jin et al. [54] studied whether Uber contributes to the horizontal equity of transportation system. Their results implied that Uber has insignificant improvement

over transportation equity in New York City. In short, the extent to which ride-hailing forestall or exacerbate inequalities in transportation is not well understood.

4 Methods of Evaluating Transportation Equity

A variety of research methods including survey research [47], interviews and focus group [48], content analysis [37], correlational research [50, 51, 45, 46, 38], experimental research [11, 49], and equity metrics [28, 54, 39], have been employed to evaluate transportation equity. These methods differ in their focuses and features, but can be used together to complement each other. Statistical tests (that routinely used in correlational research) and equity metrics are two key techniques for discrimination discovery in both transportation and machine learning research. Experimental research allows the identification of causal relationships between variables. This section focuses on correlational research, experimental research, and equity metrics.

4.1 Correlational research

Correlational research aims to explore the relationship between two or more natural occurring variables. It determines which variables are related and how they are related (e.g., positive or negative) [55]. The two main steps involved in correlational research are measurement and data analysis. Researchers collect and measure variables from a variety of settings, but do not control over or manipulate them. Data analysis (e.g., statistical analyses, GIS methods, visualization) is applied to describe the relationships between variables. Correlational research does not establish a causal relationship between variables, but allows researchers to examine the associations among many variables at the same time.

Many equity studies employ correlational research to discover associations between transportation services provision (or usage) and sensitive attributes (i.e., percentage of minority in a neighborhood). Statistical methods (e.g., regression, t-test) are often used to discover statistical relationship. GIS methods (e.g., buffer analysis) are usually employed at the same time for generating variables for statistical tests, analyzing spatial distribution, and visualizing results. Three examples of correlational research are presented below.

Example 1: Quantifying the equity of bikeshare access in US cities Ursaki and Aultman-Hall [40] examined the access equity of docked bikeshare systems in seven US cities by comparing the socioeconomic characters of areas within and outside bikeshare service areas. A service area is defined as a 500m buffer around a bike station. The equitable situation for a city is that the characteristics (e.g., percent white) of population inside the service areas are not different from the population outside service areas.

The authors obtained docking station locations data from both the open data portals and the operators directly. Socioeconomic data including population density, race, education level, income, and age was obtained from ACS at census block group (CBG) level. Then the socioeconomic variables inside and outside service areas per CBG was calculated for each city. Student's t-tests were performed to assess statistical significance. Their results showed that the low-income, the elderly, and the minority have less access to bikeshare. For example, in Chicago, the percentage of African Americans inside and outside service areas is 18.7% and 41.9%, respectively.

This example examines seven cities in one study, providing a more holistic view of the equity of bikeshare access compared to studies that focus on only one city. Nevertheless, this study has several limitations. First, equity analysis is only conducted at city level. Although the socioeconomic variables inside and outside service areas were calculated at CBG level, the authors did not discuss the spatial heterogeneity within each city. Second, docking station placement is only one of the factors that influence access equity. This study did not consider other important factors such as the supply of bikes at each station over time. Lastly, the Student's t-tests may give misleading results in this study, because the spatial dependencies among neighboring CBG violate the independence assumption required by the test.

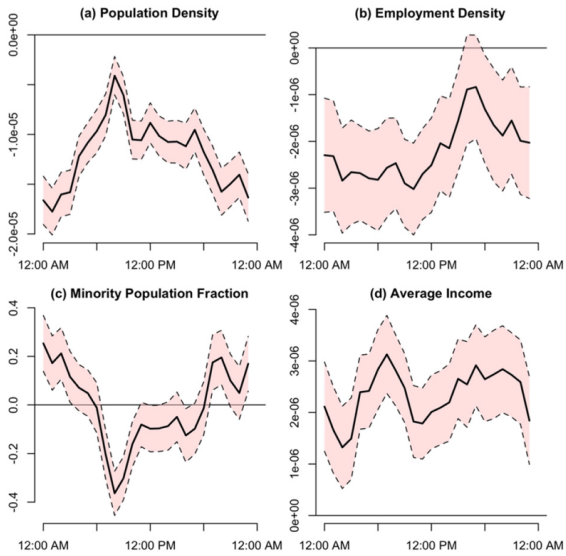


Figure 1: Coefficient estimates and 95% confidence interval of spatial error model for (a) population density, (b) employment density, (c)minority population fraction, and (d) average income [50].

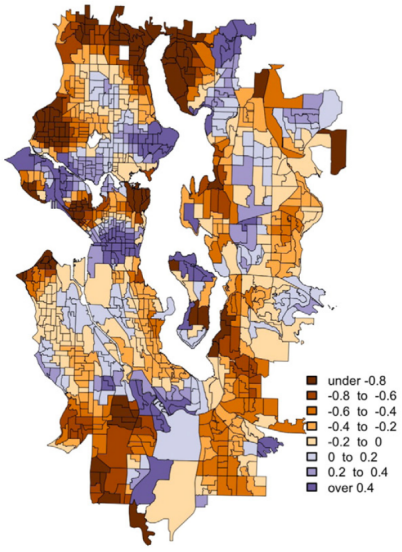


Figure 2: Coefficients for minority fraction from geographically weighted regression. Purple indicates a positive association between expected waiting time and minority fraction; gold indicates a negative association [50].

Example 2: Transportation network company wait times in Greater Seattle, and relationship to socioeconomic indicators Hughes and MacKenzie [50] investigated the relationships between wait times for UberX and socioeconomic indicators at census block group (CBG) level in Greater Seattle area. They obtained wait times by making UberX requests through Uber API using quasi-randomly selected locations throughout Greater Seattle. They collected about 1 million data points over a two-month period in 2015. Socioeconomic data including population density, employment density, average income, and minority population fraction was collected from the American Community Survey 5-year estimates (ACS).

They first fitted a regression model with mean waiting times in a CBG as dependent variable and socioeconomic attributes as independent variables. Using a Moran index test, they found significant spatial dependencies among waiting times in different CBG. Subsequently, they developed a spatial error model for each hour of the day to incorporate spatial effect into regression. Results showed that after adjusting the other covariates, higher population density and employment density were associated with shorter waiting time, but that the fraction of minorities in a block group did not significantly associated with waiting times, and that the relationship between these two variables varied between positive and negative throughout the day (Figure 1). In addition, higher average income is associated with longer wait times, suggesting that low-income areas enjoy better services. Geographically weighted regression (GWR) [56] was used to inform different impacts of each socioeconomic variable on different regions. GWR results showed that the relationship between the fraction of minority and wait times is mostly negative. They concluded that "white and wealthy" areas do not necessarily enjoy a better TNC service in terms of wait times.

The strength of this study is that it examined both spatial and temporal variations of the effects of different variables (e.g., minority fraction) on TNC waiting times. An interesting addition to this study is to include factors describing the urban form, such as road network into analysis into analysis. For example, the authors found out that higher income is associated with longer waiting time. It is possible that areas with dense road networks tend to experience shorter waiting times, and high-income individuals tend to live in areas with sparse road networks. If this is case, it implies that current urban infrastructure may contribute to the inequalities of new mobility

services. For the same reason, it is unclear if the relationships found in this study will generalize to other cities, of which urban forms (e.g., road network, crime rate, employment density, etc.) differ from Seattle.

Example 1 and Example 2 both examined the equity of service provision in terms of a single indicator (service area coverage and waiting times). These two examples sought to evaluate *equity of opportunity*. While waiting times and docking station locations are important, they do not fully imply the disparities in actual use. For example, individuals without a smartphone cannot use shared bikes even if the docking station is located close to them. The following example approaches this problem from another perspective, namely, focusing on evaluating the *equity of outcome*.

Example 3: Inequalities in usage of a public bicycle sharing Ogilvie and Goodman [38] explored the correlation between the usage of a bikeshare system in London and socioeconomic attributes. The dataset they use is the anonymized user registration data of a bikeshare system. They examined two dependent variables separately: mean number of trips made by a registered user per month (continuous) and whether a registered user has ever made any trip (binary). They constructed a series of independent variables from the registration data, including gender, place of residence, income deprivation (English Indices of Deprivation) at the level of the Lower Super Output Area (LSOA, a base unit of UK census data), non-White percentage of residential LSOA, distance from residence to nearest bike station, number of stations within 250m of residence, month of registration, etc.

The authors employed linear regression to examine the relationship between "mean number of trips per month" and independent variables, and logistic regression to examine the relationship between "ever made any trip" and independent variables. Spatial autocorrelation was accounted for using maximum likelihood estimation. Regression results showed that female users made fewer trips than males per month and users in more deprived areas are less likely to live close to a bike station. After adjusting for the distance from residential area to station, those in more deprived areas made more trips than those in the least deprived areas. They concluded that disparities exist in usage of the system across population, and the system has potentials to fulfill unmet meet if services expand to more deprived areas.

This study examined the equity of individual-level bikeshare usage. Although the authors found that female users tend to have fewer trips than male users, they cannot determine the cause of this observed relationship. It could be that females tend to have fewer bike trips at night or to regions with high crime rates due to safety concerns. After adjusting for crime rate or time of day, the association between bike usage and gender may change. This brings up another limitation of this research resulted from the use of automatically collected data from bikeshare system. The authors did not have control over the data collection process, so what they could study is also limited. Moreover, constrained by the data availability, the authors had to use area-level socioeconomic variables derived from the postcode of registration debit or credit card. It is thus unclear whether the conclusions would still hold if individual-level variables were available. The temporal scale of the data (7 months) limits the possibility to explore seasonal effects of bike usage.

Advantages and limitations of correlational research for equity studies By using correlational research, researchers can examine the relationships between transportation provision (or usage) and a wide range of variables collected from various sources. This is especially true when large amount of automatically collected data (e.g., smart card data, bikeshare trip database) is available. Correlational research is appropriate when researchers are unable to manipulate the variables due to practical or ethical reasons. For example, in equity studies, area-level variables such as average income of a CBG is not controllable.

One limitation of correlational research is that a significant correlation does not allow the researcher to determine a causal relationship, because there could be many factors that the researcher did not study but contribute to the correlation. And these factors could be independent of mobility service provision. Further inquiry is needed to corroborate the findings from a correlational study. Another limitation is that correlational research heavily depends on data availability and data quality, as discussed in Example 3.

4.2 Experimental Research

Experimental research enables researchers to identify causal relationship. In an experiment design, the researcher seeks to fully control the environment conditions so that variables of interest can be manipulated, while other variables are controlled (or randomized) across conditions. In this way, the effects of variables of interest can be tested by comparing between two or more conditions. Unlike correlational research, experimental research strictly controls for the impacts of variables not of interest, thus allowing the effects of variables of interest to be measured upon the outcome [55].

Example: Racial and gender discrimination in transportation network companies Ge et al. [11] studied the racial and gender discrimination in Transportation Network Companies (TNCs). They undertook two randomized control trails, hailing about 1500 rides in Seattle and Boston and recording service quality indicators. In the Seattle experiment, the treatment is race. Eight RAs (two African American females, two white females, two African American males, and two white males) were hired to request rides. Measures including estimated waiting times, acceptance time (time between trip request and acceptance), actual waiting times (time between acceptance and arrival), trip cancellation rate, trip duration, costs, and ratings were recorded by screenshots for each trip. To control for variables not of interest, the authors adopted a number of strategies. The RAs are undergraduate students, avoiding confounding factors such as age. They were given the identical smartphones using the same carriers, and received the same data collection instructions. The RAs were instructed to minimize their interactions with the driver, preventing the introduction of factors that influence ratings and travel time. Specific routes were developed to control for pick-up locations and travel duration. These routes were randomly assigned to RAs. RAs were also instructed to travel after evening rush hours from Mondays to Thursdays. Ordinary least squares regression (OLS) results showed that acceptance time is longer for African American riders than white riders for both UberX and Lyft.

In the Boston experiment, the authors adopted a within-group design. They hired eight RAs with a range of ethnic backgrounds summoning UberX or Lyft rides in Boston, each requesting rides under a "white-sounding" name and a "distinctively black name". This change in design aims to control the differences in data collection practices among RAs. In this case, the treatment is whether the rider has a black sounding name. Other aspects of experiment design are similar to those of the Seattle experiment. OLS results showed that riders with African American-sounding names experienced more frequent trip cancellations, and that African American males have higher cancellation rates than white males. Further analysis revealed that trip cancellations concentrated in pickup locations with low population density. They concluded that racial discrimination exists in TNC services in Seattle and Boston.

Advantages and limitations of experimental research for equity studies Experimental research allows for drawing causal conclusions. This is because experiments are conducted in controlled conditions and researchers can claim that the changes in outcomes are caused by the variable of interest.

Experimental research has notable limitations. First, the requirement that controlling all variables that might influence the outcomes is sometimes not realistic. This is especially true for experiments conducted in a natural environment. For example, in Ge et al.'s Seattle experiment [11], there are variances in the data collection practices (e.g., the time lag between taking screenshots and sending requests) among RAs. This influences the measurement of outcome variables. Second, compared to automatically collected data or survey data, experiments are often not able to produce large amount of data. Data collection in experiments is often expensive and labor-intensive. Finally, although experimental research can determine causal effects (e.g., racial discrimination exists in TNC services in Seattle), it cannot unveil the reasons why the outcome occurred (e.g., why TNC drivers discriminate against certain races). Further investigation through other research methods (e.g., interviews) is needed to understand the phenomenon.

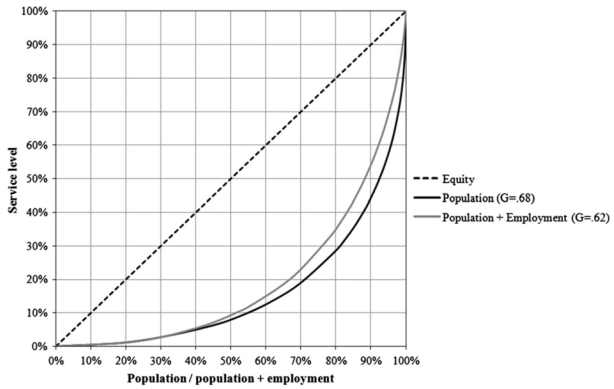


Figure 3: Use Lorenz curves to compare the equity of public transport service level to demand (population and employment) [28].

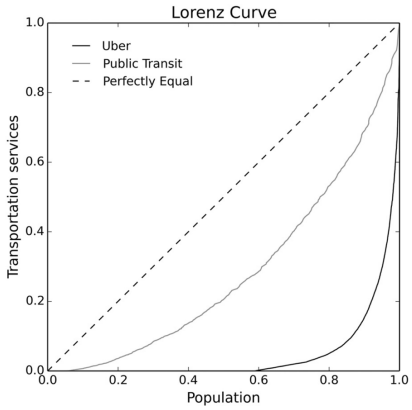


Figure 4: Use Lorenz curves to compare the equity of public transport and Uber service level to population [54].

4.3 Equity Metrics

Metrics that measure the distribution of some mobility system impacts (e.g., service level) have been widely adopted in transportation equity evaluation. Unlike statistical tests which focus on the discovery of discrimination or inequalities, metrics directly gauge the degree of equity by a single value. Equity metrics used in transportation equity research differ but overlap with those used in fairness in machine learning research. The similarities between them partially arise from the fact that both fields borrowed ideas from other domains such as social welfare and economics. For example, Gini coefficient (or Gini index), initially proposed to represent cumulative income and wealth distribution across a population, is one of the most popular equity metrics used in transportation to gauge the equity of transportation resource allocation [8]. However, Gini index has not yet received much attention in machine learning community [57]. Perhaps this is because fairness in machine learning research has primarily concentrated on classification problems that used in credit scoring, profiling of potential suspects, hiring, etc., for which other metrics are more appropriate. On the other hand, there are a few metrics, such as the "80% Rule" [58], were used in both communities [59]. The following examples introduce the use of Gini index and the "80% Rule" in transportation equity.

Example 1: Using Lorenz curves to assess public transport equity in Melbourne Delbosc and Currie [28] proposed to use Gini index as an equity metric of public transit service provision. A Lorenz curve is a graphical representation of Gini index. The figure below (see Figure 3) illustrates an example of a Lorenz curve representing the cumulative income across a population. The perfect equitable income distribution is plotted as the dashed line (line of equity) and an inequitable distribution of income is represented by the solid curve (Lorenz curve). A point on the solid curve can be interpreted as X percent (e.g., 70%) of population shares about Y percent (e.g., 25%) of the total income of the population. Gini index is the ratio of the area between the line of equity and the Lorenz curve (A), divided by the total area under the line of equity (A+B).

Delbosc and Currie applied a gini index to compare the equity of public transport service level to a proxy of demand (population and employment) in Melbourne. Service level of a census tract is expressed as a composite index taking into account bus, tram, and rail service areas and frequency. Using the service level index and the population of each census tract, the authors generated the first Lorenz curve (black solid curve) as shown in Figure 3. The Gini index is 0.68 for overall population in Melbourne. This can be interpreted as 70% of the population shares 19% of the public transport services. A second Lorenz curve (a grey solid curve) was calculated, taking into account the employment density. The Gini index for total population and employment is 0.62, appearing

	Weekday		Weekend	
Borough	**Public transit**	**Public transit + Uber**	**Public transit**	**Public transit + Uber**
Bronx	0.5803	0.5802	0.6079	0.6073
Brooklyn	0.5738	0.5725	0.5852	0.5783
Manhattan	0.6324	0.6276	0.6625	0.6352
Queens	0.7425	0.7424	0.7359	0.7332
Staten Island	0.4336	0.4335	0.4951	0.4949
Whole city	0.6653	0.6321	0.6429	0.6349

Table 2: Gini coefficient without and with Uber [54]

more equitable than the first curve. Nevertheless, these two curves suggest that inequalities exist in public transit, with only a small portion of the population enjoying the majority of transit services.

In this example, the Gini index serves as a measure of horizontal equity, that is, providing equal resources to those equal in need. The need for transportation supply of each census tract is approximated by population and employment density. So the perfect equitable distribution is that every unit of population and jobs shares the same transport resources. The need of special demographic groups (vertical equity) is not considered.

Example 2: Using Lorenz curves to access the equity of Uber and public transit in New York City Most recently, Jin et al. [54] employed Lorenz curves and Gini index to study the equity of Uber services in New York City (see Figure 4). They calculated service level for Uber and public transit using a similar approach as Delbosc and Currie [28]. Their results suggested that Uber is less equitable than public transit: 20% of population shares the 95% of Uber services.

They further compared Gini indexes of different boroughs for public transit and public transit + Uber (see Table 2). The results (see Table 2) shows that with Uber, the Gini index of the whole city reduced by about 0.03 on weekdays and by about 0.008 on weekends. This implies that Uber has insignificant impact on the transportation equity of New York City.

This study exemplifies how Gini index can be used to compare transportation equity across regions and across modes. This is possible because Gini index has several desirable features: it does not depend on the size of the population, the overall transit supply level, or the geographic units. For example, Gini index can be used to examine the equity of a neighborhood, a city, or a country. And it enables the comparison of equity between a city with high level of transit supply and one with low supply.

One limitation of Gini index is its heavy reliance on data. As Jin et al. noted, the main reason to choose New York City as study area is data availability. Beyond availability, all data sources have limitations (e.g., census data is not up-to-date) that would be calculated into Gini index. Another limitation lies in the way the service level is calculated. Studies that employed Gini indexes tend to use different methods to calculate service level [30]. It is unclear whether changing the service level indicator will significantly affect Gini index. These two limitations suggest that Gini indexes should be interpreted with caution.

Example 3: Evaluation of the equity of bikeshare system accessibility Meng [39] applied the "80% Rule" to evaluate access equity of a bikeshare system in Chicago. The "80% Rule" was advocated by the US Equal Employment Opportunity Commission to detect disparate impacts in employee selection procedures. The 80% Rule states that if the selection rate for minorities is less than 80% of the rate of non-minorities, the procedure is deemed to be discriminatory [58]. Similar to Ursaki and Aultman-Hall (see Example 1 of Section 4.1), the analysis is based on the locations of docking stations. The author created a 0.25-mile buffer around each station as service area, and calculated the demographic characteristics (i.e., race, gender, education, language proficiency, and income level) of population inside each service area. For each station, the equity metric based on the 80% Rule is calculated as follows:

$$Ratio = \frac{Number\ of\ minorities/Total\ number\ of\ minorities}{Number\ of\ non-minorities/Total\ number\ of\ non-minorities} \tag{19}$$

The results show that more than 33% of the stations have ratios below 0.8 for all demographic characteristics (except for gender) under examination.

There are several limitations of this study. First, instead of providing a city-level ratio, the authors computed station-level ratios and examined equity using the percentage of stations that violate the 80% Rule. This approach is problematic when the stations are not equally distributed. It is possible that a majority of the stations are all located in a small portion of the city and they tend to have similar ratios. Second, docking station placement cannot sufficiently represent access to bikeshare, as discussed in Example 1 of Section 4.1. Despite these weaknesses, this study serves as a typical example of using fairness metrics to evaluate vertical equity in new mobility systems.

Advantages and limitations of equity metrics Equity metric provides a single measure of equity, making it possible to track trends over time and conduct comparative studies between cities. It is easier for non-experts to interpret compared to statistical tests, therefore suited for conveying evaluation results to broader audience.

However, the reliability of metrics heavily depends on the quality of data sources. Moreover, different metrics often reflect competing goals. For example, Gini index measures horizontal equity, emphasizing individuals with equal ability or need gets equal resources. The 80% Rule shares a similar spirit of group fairness [59], which advocates for equal resource distribution across difference demographic groups. The choice of metrics may significantly affect evaluation results, so the use of multiple metrics is important.

4.4 Other methods

Apart from the three research methods described above, surveys, interviews, and focus groups have been used for transportation equity studies. These methods can be used to develop a deeper understanding of why inequalities exist based on the opinions, attitudes, and experiences from stakeholders of mobility systems. For example, McNeil, Nathan, et al. [47] conducted a survey of residents living in underserved neighborhoods with bikesahre stations. The findings revealed that minority respondents have more barriers, for example, costs of membership, to using shared bikes than non-minorities. This helps to explain why providing adequate spatial access to disadvantaged neighborhoods alone is not enough to address the disparities in actual use.

5 Transportation Equity and Fairness in Machine learning

In examining the fairness (equity) definitions from transportation equity community and fair machine learning (FairML) community, we observe that a natural mapping between them can be established, though further effort is needed to create a consistent mapping between concepts in one domain to the other. *Horizontal equity* echoes the spirit of *individual fairness* (similar people should be treated similarly). *Vertical equity* resembles *group fairness* (sensitive attributes should be independent from outcomes). This is true in cities where there is an uneven distribution of transport supply across different socioeconomic groups. Vertical equity encourages compensating for such inequalities by policies favoring disadvantaged groups. This aligns with group fairness that the level of transportation supply in a city should be the same across different groups. Vertical equity and group fairness are only "roughly" related because by definition, group fairness stresses "independence" between sensitive attributes and outcome, whereas vertical equity does not.

The most commonly used method for evaluating horizontal equity is Gini index. It has not attracted much attention in machine learning community. This may be partially due to the fact that not much attention has been paid to resource allocation problems in fair machine learning research. On the other hand, machine learning

community has developed a few metrics for individual fairness. Individual fairness requires that the "similarity" between a pair of individuals from two demographic groups respectively has to be defined. For example, in making hiring decisions, the algorithm has to possess perfect knowledge of how to compare the "qualification" of two individuals. This is often not realistic in practice and we have to come up with a suitable similarity metric that is best agreed upon among domain experts of a task. Theoretically, individual fairness can be used to evaluate horizontal equity. For example, in a simplified shared bike allocation problem, we use population and employment density as the demand for bikes. Then the differences in demand between two areas a and b can be expressed as $d(a, b)$ according to some similarity function d. Suppose we have an algorithm assigning bikes to areas, the number of bikes that area a and b will get is $f(a)$ and $f(b)$, respectively. A fair allocation satisfying individual fairness requires that for every two pairs of areas in the city: $D(f(a), f(b)) <= d(a, b)$, where D is another similarity function. The difficulty again, lies in the fact that we do not have perfect knowledge to determine the similarity in demand between two areas.

The majority of transportation equity research focuses on vertical equity. Likewise, more attention has been devoted to group fairness than individual fairness in machine learning community. Transportation equity heavily employs statistical tests for equity analysis, which is appropriate for discovering unfairness. Machine learning uses fairness metrics much more often, because metrics allow researchers to reduce achieving fairness goals to a much simpler problem: minimizing a value that represents unfairness. This is also valid in terms of algorithm design. In fact, some metrics, such as the 80% Rule, have been used in both communities. This connection may open great possibilities for bridging these two domains.

Fair machine learning community focuses almost exclusively on methods, whereas transportation equity concerns more about applications, policies, and interventions. Although fair machine learning approaches hold great promises in optimizing resource allocation in mobility settings, there is a long way to go to design, deploy, and evaluate a fairness-aware data-driven system as a real-world application. At the end of this paper, I hope to highlight the urgency of convergence of these two fields. Ultimately, researchers with knowledge in both fields, practitioners, policy-makers, and citizens should work together towards a common goal: a fair and effective transportation system for all citizens.

6 Conclusion

This paper summarized the findings and methods of equity studies in mobility systems, with a focus on new mobility systems. For new mobility services, it is generally agreed that disparities exist in the access and use of docked bikeshare system, but the equity implications of ride-hailing are still unclear. Further research is needed to understand how to deliver a more equitable new mobility system to serve the need of different groups. Many research methods have been employed in transportation equity studies. Different methods vary in their objectives, strengths and weaknesses. Correlational research can exploit a wide range of data sources and discover associations among many factors, but it cannot determine causal relationships. Equity metrics enable comparative studies among cities and assessment of changes over time, but their reliability is highly dependent on data. Experimental research can produce reliable findings, but is expensive and difficult to control all extraneous variables. The choice of research methods depends on research goals, and multiple methods can be used together to complement each other.

Given the similarities in objectives, concepts, and methods between transportation equity community and fairness in machine learning community, bridging these two domains together holds promise to enable multi-disciplinary breakthroughs.

References

[1] Chao Zhang, Keyang Zhang, Quan Yuan, Haoruo Peng, Yu Zheng, Tim Hanratty, Shaowen Wang, and Jiawei Han. Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In *Proceedings of the 26th International Conference on World Wide Web*, pages 361–370. International World Wide Web Conferences Steering Committee, 2017.

[2] Jason Shafrin, Jeff Sullivan, Dana P Goldman, and Thomas M Gill. The association between observed mobility and quality of life in the near elderly. *PloS one*, 12(8):e0182920, 2017.

[3] Ya-Fen Chan, Shou-En Lu, Bill Howe, Hendrik Tieben, Theresa Hoeft, and Jürgen Unützer. Screening and follow-up monitoring for substance use in primary care: an exploration of rural–urban variations. *Journal of general internal medicine*, 31(2):215–222, 2016.

[4] Qi Wang, Nolan Edward Phillips, Mario L Small, and Robert J Sampson. Urban mobility and neighborhood isolation in america's 50 largest cities. *Proceedings of the National Academy of Sciences*, 115(30):7735–7740, 2018.

[5] Todd Litman. *Evaluating transportation equity*. Victoria Transport Policy Institute, 2018.

[6] Susan Shaheen, Corwin Bell, Adam Cohen, and Balaji Yelchuru. Travel behavior: Shared mobility and transportation equity. Technical report, U.S. Department of Transportation, 2017.

[7] Chelsey Palmateer and David M Levinson. Justice, exclusion, and equity: An analysis of 48 us metropolitan areas. Technical report, University of Minnesota: Nexus Research Group, 2017.

[8] Anthony Michael Ricciardi, Jianhong Cecilia Xia, and Graham Currie. Exploring public transport equity between separate disadvantaged cohorts: a case study in perth, australia. *Journal of transport geography*, 43:111–122, 2015.

[9] Raj Chetty, Nathaniel Hendren, Frina Lin, Jeremy Majerovitz, and Benjamin Scuderi. Childhood environment and gender gaps in adulthood. *American Economic Review*, 106(5):282–88, 2016.

[10] Todd Goldman and Roger Gorham. Sustainable urban transport: Four innovative directions. *Technology in society*, 28(1-2):261–273, 2006.

[11] Yanbo Ge, Christopher R Knittel, Don MacKenzie, and Stephen Zoepf. Racial and gender discrimination in transportation network companies. Technical report, National Bureau of Economic Research, 2016.

[12] Indre Zliobaite. A survey on measuring indirect discrimination in machine learning. *CoRR*, abs/1511.00148, 2015.

[13] Kevin Petrasic, Benjamin Saul, James Greig, Matthew Bornfreund, and Katherine Lamberth. Algorithms and bias: What lenders need to know. *White & Case*, 2017.

[14] Claire Cain Miller. When algorithms discriminate. *The New York Times*, 9, 2015.

[15] Cynthia Rudin. Predictive policing using machine learning to detect patterns of crime. *Wired Magazine, August*, 2013.

[16] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NIPS Tutorial*, 2017.

[17] Jessica Guynn. Google photos labeled black people'gorillas'. *USA Today*, 1, 2015.

[18] David Ingold and Spencer Soper. Amazon doesn't consider the race of its customers. should it. *Bloomberg*, 2016.

[19] Bruno Lepri, Jacopo Staiano, David Sangokoya, Emmanuel Letouzé, and Nuria Oliver. The tyranny of data? the bright and dark sides of data-driven decision-making for social good. In *Transparent data mining for big and small data*, pages 3–24. Springer, 2017.

[20] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[21] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[22] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[23] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *CoRR*, abs/1810.08810, 2018.

[24] Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

[25] Meg Young, Luke Rodriguez, Emily Keller, Feiyang Sun, Boyang Sa, Jan Whittington, and Bill Howe. Beyond open vs. closed: Balancing individual privacy and public accountability in data sharing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 191–200. ACM, 2019.

[26] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 31(4):611–627, 2018.

[27] Jenna Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):2053951715622512, 2016.

[28] Alexa Delbosc and Graham Currie. Using lorenz curves to assess public transport equity. *Journal of Transport Geography*, 19(6):1252–1259, 2011.

[29] Karen Lucas, Bert Van Wee, and Kees Maat. A method to evaluate equitable accessibility: combining ethical theories and accessibility-based approaches. *Transportation*, 43(3):473–490, 2016.

[30] Tao Feng and Harry JP Timmermans. Trade-offs between mobility and equity maximization under environmental capacity constraints: A case study of an integrated multi-objective model. *Transportation Research Part C: Emerging Technologies*, 43:267–279, 2014.

[31] Nikolas Thomopoulos, Susan Grant-Muller, and MR Tight. Incorporating equity considerations in transport infrastructure evaluation: Current practice and a proposed methodology. *Evaluation and program planning*, 32(4):351–359, 2009.

[32] Steven Raphael and Lorien Rice. Car ownership, employment, and earnings. *Journal of Urban Economics*, 52(1):109–130, 2002.

[33] Robert D Bullard. Addressing urban transportation equity in the united states. *Fordham Urb. LJ*, 31:1183, 2003.

[34] Ahmed El-Geneidy, David Levinson, Ehab Diab, Genevieve Boisjoly, David Verbich, and Charis Loong. The cost of equity: Assessing transit accessibility and social disparity using total travel cost. *Transportation Research Part A: Policy and Practice*, 91:302–316, 2016.

[35] Eduardo Alcantara Vasconcellos. Urban transport policies in brazil: The creation of a discriminatory mobility system. *Journal of transport geography*, 67, 2018.

[36] Stuart Cohen and Clarrissa Cabansagan. A framework for equity in new mobility, 2017.

[37] Kevin Manaugh, Madhav G Badami, and Ahmed M El-Geneidy. Integrating social equity into urban transportation planning: A critical evaluation of equity objectives and measures in transportation plans in north america. *Transport policy*, 37:167–176, 2015.

[38] Flora Ogilvie and Anna Goodman. Inequalities in usage of a public bicycle sharing scheme: socio-demographic predictors of uptake and usage of the london (uk) cycle hire scheme. *Preventive medicine*, 55(1):40–45, 2012.

[39] Chao Meng. Evaluation of the equity of bikeshare system accessibility: A case study of chicago. Technical report, Georgia Institute of Technology, 2018.

[40] Julia Ursaki, Lisa Aultman-Hall, et al. Quantifying the equity of bikeshare access in us cities. Technical report, University of Vermont. Transportation Research Center, 2015.

[41] C Scott Smith, Jun-Seok Oh, and Cheyenne Lei. Exploring the equity dimensions of us bicycle sharing systems. Technical report, Western Michigan University. Transportation Research Center for Livable, 2015.

[42] Xuefeng Li, Yong Zhang, Li Sun, and Qiyang Liu. Free-floating bike sharing in jiangsu: Users' behaviors and influencing factors. *Energies*, 11(7):1664, 2018.

[43] Stephen J Mooney, Kate Hosford, Bill Howe, An Yan, Meghan Winters, Alon Bassok, and Jana A Hirsch. Freedom from the station: Spatial equity in access to dockless bike share. *Journal of Transport Geography*, 74:91–96, 2019.

[44] Chengcheng Xu, Junyi Ji, and Pan Liu. The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets. *Transportation research part C: emerging technologies*, 95:47–60, 2018.

[45] David William Daddio and N Mcdonald. Maximizing bicycle sharing: an empirical analysis of capital bikeshare usage. *University of North Carolina at Chapel Hill*, 8, 2012.

[46] R Alexander Rixey. Station-level forecasting of bikesharing ridership: station network effects in three us systems. *Transportation Research Record*, 2387(1):46–55, 2013.

[47] Nathan McNeil, Jennifer Dill, John MacArthur, Joseph Broach, and Steven Howland. Breaking barriers to bike share: Insights from residents of traditionally underserved neighborhoods. *Transportation Research and Education Center (TREC)*, 2017.

[48] SS Kretman, DC Johnson, and WP Smith. Bringing bike share to a low-income community: Lessons learned through community engagement, minneapolis, minnesota. *Preventing Chronic Disease*, 2011.

[49] Anne Elizabeth Brown. *Ridehail revolution: Ridehail travel and equity in Los Angeles*. PhD thesis, UCLA, 2018.

[50] Ryan Hughes and Don MacKenzie. Transportation network company wait times in greater seattle, and relationship to socioeconomic indicators. *Journal of Transport Geography*, 56:36–44, 2016.

[51] Mingshu Wang and Lan Mu. Spatial disparities of uber accessibility: An exploratory analysis in atlanta, usa. *Computers, Environment and Urban Systems*, 67:169–175, 2018.

[52] Jennifer Stark and Nicholas Diakopoulos. Uber seems to offer better service in areas with more white people. that raises some tough questions. *The Washington Post*, 2016.

[53] Scott Scott Russell Middleton. *Discrimination, regulation, and design in ridehailing*. PhD thesis, Massachusetts Institute of Technology, 2018.

[54] Scarlett T Jin, Hui Kong, and Daniel Z Sui. Uber, public transit, and urban transportation equity: A case study in new york city. *The Professional Geographer*, pages 1–16, 2019.

[55] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.

[56] A Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2003.

[57] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2239–2248. ACM, 2018.

[58] The U.S. EEOC. Uniform guidelines on employee selection procedures. *March 2, 1979*, 43:111–122, 1979.

[59] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

Fairness and Diversity in Public Resource Allocation Problems*

Nawal Benabbou[1], Mithun Chakraborty[2], Yair Zick[2]
[1]Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6 F-75005 Paris, France
nawal.benabbou@lip6.fr
[2]Department of Computer Science, National University of Singapore, Singapore
{mithun,zick}@comp.nus.edu.sg

Abstract

In this article, we address important extensions to the problem of allocating indivisible items to a population of agents: The agents are partitioned into disjoint groups on the basis of attributes (e.g., ethnicity) and we want the overall utility of the allocation to respect some notion of diversity *and/or* fairness *with respect to these groups. We study two specific incarnations of this general problem. First, we address a constrained optimization problem, inspired by diversity quotas in some real-world allocation problems, where the items are also partitioned into blocks and there is an upper bound on the number of items from each block that can be assigned to agents in each group. We theoretically analyze the price of diversity – a measure of the overall welfare loss due to these capacity constraints – and report experiments based on two real-world data sets (Singapore public housing and Chicago public school admissions) comparing this constrained optimization-based approach with a lottery mechanism with similar quotas. Next, instead of imposing hard constraints, we cast the problem as a variant of fair allocation of indivisible goods – we treat each group of agents as a single entity receiving a bundle of items whose valuation is the maximum total utility of matching agents in that group to items in that bundle; we present algorithms that achieve a standard relaxation of envy-freeness in conjunction with specific efficiency criteria.*

1 Introduction

Over the years, the Singapore government has adopted several social integration measures, to accommodate its multi-ethnic and multi-cultural population; one of these is the *Ethnic Integration Policy* (EIP) used by the Housing and Development Board (HDB) since 1989 [27] to determine housing allocations. This government body is charged with the construction of government subsidized public housing estates, and selling them to Singapore residents. The EIP sets upper bounds on the percentage of flats in every estate that can be owned by households of every major ethnic group: since 5 March 2010, every HDB housing block is required to hold no more than 87% Chinese, 25% Malay, and 15% Indian/Others [17, 14]. These ethnic quotas prevent the over-representation of any one group in an estate (resulting in de-facto segregation). HDB uses a lottery to allocate new developments: all

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

*Parts of this article are published in AAMAS 2018 [7] and accepted to IJCAI 2019 [8].

applicants who apply for a particular development pick their flats in a random order. Under the lottery mechanism, applicants selected later in the order may not get their top choices; in fact, they may be rejected because the quota for their ethnic group has been filled, *even if there are empty flats that they are willing to take*.

Example 1: Consider an estate with 100 flats; Amirah (who is ethnically Malay) receives a queue number of 30 — i.e., she is the 30th person to choose a flat; if, by random chance, 25 ethnic Malays precede her in the queue and pick before her, the ethnic quota for Malays (25%) will be filled and Amirah will no longer be allowed to select a flat in that estate. On the other hand, if Amirah is 110th in line but all families preceding her in the queue are ethnically Chinese, then at most 87 out of the 100 flats will be taken (since the Chinese have an 87% quota) and Amirah will have at least 13 flats to choose from. ♦

Over 72% of Singapore apartments are HDB flats [34], housing an estimated 81% of Singapore residents [18] as of 2018; thus, the HDB public housing market has a significant impact on the life and welfare – both individual and collective – of this island nation [13, 14, 20, 32, 37]. The HDB lottery mechanism, coupled with ethnicity constraints, adds another layer of complexity to what is, at its core, similar to the classic weighted bipartite matching or *assignment* problem [21, 26]: agents (buyer households) have idiosyncratic values/utilities for items (flats) while a central planner (HDB) wishes to allocate at most one item to each agent and vice versa, with the economic criterion of *overall utility/welfare/efficiency* in mind – but also with an added social goal of *promoting diversity*. Inspired by diversity-respecting allocation mechanisms — prevalent not only in Singapore public housing but also in other domains such as matching residents to hospitals in Japan, school admissions in many U.S. cities, and more [19, 36, 16] — we formally study the balance between maximizing allocative welfare and promoting allocative diversity. This underlying problem of allocation/assignment of goods distinguishes our contribution from the extensive literature on fairness and diversity in subset selection (see e.g., [35, 10] and references therein).

In Section 2, we analyze a (simplified) HDB housing market as an extension to the assignment problem where the sets of agents and items are partitioned into subsets called *types* and *blocks* respectively; there is a pre-specified upper bound on the number of agents of each type that can be assigned items in each block, called *type-block capacities*. We analyze the *price of diversity*, i.e., the fractional loss in the overall (optimal) welfare due to these capacity constraints, and relate it to a measure of *type disparity*; we also report simulations based on recent, real-world data sets, comparing our constrained optimization approach with a lottery mechanism with ethnicity quotas in terms of welfare.

In Section 3, we present an alternative approach towards the efficiency-diversity trade-off, drawing on and adding to the rich literature on the fair division of indivisible goods (see e.g., [28, 12, 22, 4, 3]). Here, each type is represented by a *super-agent* that is allocated a *bundle* of items; the super-agent's valuation of a bundle is given by the maximum utility assignment of items in the bundle to agents of that type. When agent-item utilities are binary, we provide a polynomial-time algorithm that computes an allocation with the maximum (utilitarian) social welfare while satisfying a popular fairness criterion: *envy-freeness up to one item* (EF1) [11]; for arbitrary real-valued utilities, we show experimentally that a heuristic extension of the classic *envy-graph algorithm* due to Lipton et al. [24] produces an EF1 allocation with low *waste* (a new inefficiency concept that we have introduced for this setting).

2 Diversity through hard capacity constraints

First, let us rigorously formulate the welfare maximization problem in a bipartite matching setting under type-block capacity constraints. Detailed proofs of all theoretical results in this section can be found in Benabbou et al. [7]. Throughout the paper, for $s \in \mathbb{N}$, we denote the set $\{1, 2, \ldots, s\}$ by $[s]$.

Definition 1 (ASSIGNTC): An instance of the Assignment with Type Constraints (ASSIGNTC) problem is given by: (i) a set N of n agents partitioned into k types $N_1, \ldots, N_k$, (ii) a set M of m items/goods partitioned

into l blocks $M_1, \ldots, M_l$, (iv) a utility $u(i,j) \in \mathbb{R}_+$ for each agent $i \in N$ and each item $j \in M$, and (iv) a capacity $\lambda_{pq} \in \mathbb{N}$ for all $(p,q) \in [k] \times [l]$; λ_{pq} is an upper bound on the number of agents of type N_p allowed in block M_q. W.l.o.g. we assume that for all p, q, $\lambda_{pq} \leq |M_q|$.

An assignment of items to agents, which we will also sometimes call an (N, M)-matching, can be represented by a $(0,1)$-matrix $X = (x_{ij})_{n \times m}$ where $x_{ij} = 1$ if and only if item j is assigned to agent i. Our objective is to maximize the utilitarian social welfare or USW, i.e., the sum of agent utilities: $u(X) \triangleq \sum_{i \in N} \sum_{j \in M} x_{ij} u(i,j)$. Clearly, this optimization problem can be formulated as an integer linear program (ILP) as in Figure 1. Here, the first set of inequalities captures our *type-block constraints* while the last three sets are the usual *matching constraints* jointly ensuring that each item (resp. agent) is assigned to at most one agent (resp. item). In general, the decision version of the ASSIGNTC problem is NP-complete (Benabbou et al. [7, Theorem 3.2]) but admits a polynomial-time $\frac{1}{2}$-approximation algorithm (Benabbou et al. [7, Theorem 3.4]); moreover, the problem can be solved in $\mathrm{polynomial}(n, m)$ time by a minimum-cost network flow-based algorithm (Benabbou et al. [7, Theorem 3.6]) if the utility matrix satisfies one of the following two conditions: (i) *type-uniformity* i.e., all agents of a type have identical utilities (for all $p \in [k]$ and for all $j \in M$, there exists $U_{pj} \in \mathbb{R}_+$ such that $u(i,j) = U_{pj}$ for all $i \in N_p$); (ii) *block-uniformity* i.e., all items in a block are clones of each other (for all $q \in [l]$ and for all $i \in N$, there exists $U_{iq} \in \mathbb{R}_+$ such that $u(i,j) = U_{iq}$ for all $j \in M_q$).

$$
\begin{aligned}
\max \sum_{i \in N} \sum_{j \in M} & x_{ij} u(i,j) & \\
s.t. \sum_{i \in N_p} \sum_{j \in M_q} & x_{ij} \leq \lambda_{pq} & \forall p \in [k], \forall q \in [l] \\
\sum_{j \in M} & x_{ij} \leq 1 & \forall i \in N \\
\sum_{i \in N} & x_{ij} \leq 1 & \forall j \in M \\
& x_{ij} \in \{0,1\} & \forall i \in N, \forall j \in M
\end{aligned}
$$

Figure 1: ILP formulation of ASSIGNTC.

We are mainly interested in how the imposition of type-block capacities impacts allocative efficiency. If we denote the set of all valid item assignments X by $\mathcal{X}$, and all assignments additionally satisfying our type-block constraints by $\mathcal{X}_\mathcal{C}$, then the unconstrained and unconstrained optimal social welfares for any given utility matrix $(u(i,j))_{n \times m}$ are given by $\mathrm{OPT}(u) \triangleq \max_{X \in \mathcal{X}} u(X)$ and $\mathrm{OPT}_C(u) \triangleq \max_{X \in \mathcal{X}_\mathcal{C}} u(X)$. Clearly, $\mathrm{OPT}_C(u) \leq \mathrm{OPT}(u)$ since $\mathcal{X}_\mathcal{C} \subseteq \mathcal{X}$. This leads to a natural measure of welfare loss for the ASSIGNTC problem; we call this the *Price of Diversity* (a similar definition appears in [1, 10]): $PoD(u) \triangleq \mathrm{OPT}(u)/\mathrm{OPT}_C(u)$. By definition, $PoD(u) \geq 1$, and its exact value depends on agent utilities, but we can bound it, regardless of the utility model, in terms of the *fractional type-block capacities* defined as $\alpha_{pq} \triangleq \lambda_{pq}/|M_q|$ for each $(p,q) \in [k] \times [l]$.

Theorem 2: For any ASSIGNTC instance, $PoD(u) \leq 1/\min_{(p,q) \in [k] \times [l]} \alpha_{pq}$.

The following example shows that there is an ASSIGNTC instance whose PoD reaches the upper bound in Theorem 2; in other words, this bound is tight.

Example 2: Suppose, $|N_{p_0}| \geq |M_{q_0}|$ for some type-block pair (p_0, q_0) in the set $\mathrm{argmin}_{(p,q) \in [k] \times [l]} \alpha_{pq}$ in an ASSIGNTC instance, and the utilities are given by $u(i,j) = 1$ if $i \in N_{p_0}$ and $j \in M_{q_0}$, $u(i,j) = 0$ otherwise. An optimal unconstrained assignment fully allocates the items in block M_{q_0} to agents in N_{p_0} for a total utility of $|M_{q_0}|$ whereas any optimal constrained assignment allocates exactly $\lambda_{p_0 q_0}$ items in M_{q_0} to agents in N_{p_0} for a total utility of $\lambda_{p_0 q_0}$. Hence, for this family of instances, $PoD(u) = |M_{q_0}|/\lambda_{p_0 q_0} = 1/\alpha_{p_0 q_0}$. ♦

In general, the bound in Theorem 2 is linear in the number of items i.e., the welfare loss due to hard diversity constraints can be significant in some instances (e.g., Example 2). However, type-block capacities are determined by a central planner in our model; a natural way of setting them is to fix the fractional capacities α_{pq} in advance, and then compute $\lambda_{pq} = \alpha_{pq} \times |M_q|$ when block sizes become available: by committing to a fixed minimum type-block quota α^* (i.e., $\alpha_{pq} \geq \alpha^*$ for all $(p,q) \in [k] \times [l]$), the planner can ensure a $PoD(u)$ of at most $1/\alpha^*$, regardless of the problem size and utility function. Higher values of α^* reduce the upper

bound on $PoD(u)$ but also increase the capacity of a block for every ethnicity, having an adverse effect on the diversity of block composition – α^* thus functions as a tunable tradeoff parameter between ethnic integration and worst-case welfare loss. In the Singapore housing allocation problem, the EIP fixes a universal percentage cap, slightly higher than the actual corresponding population proportion for every ethnicity. Plugging the current EIP percentages mentioned in Section 1 into the bound in Theorem 2, we get that the Singapore housing system has $PoD(u) \leq \frac{1}{0.87+0.25+0.15} \approx 6.67$. We mention that the price of diversity compares the *best* constrained allocation with the best unconstrained allocation; mechanisms deployed in practice do not necessarily try to find the optimal allocation. For example, the HDB mechanism uses a lottery to allocate flats, and thus may theoretically exhibit greater welfare loss than the bound set in Theorem 2.

Theorem 2 offers a worst-case tight bound on the price of diversity, making no assumptions on agent utilities, but Example 2 suggests that this upper bound is attained when social welfare is solely extracted from a single agent type and a single block. Intuitively, we can improve our bound if a less 'disparate' optimal assignment exists. To formalize this notion, we introduce a new parameter. For any optimal unconstrained assignment $X^* \in \mathcal{X}$, let $\beta_p(X^*)$ denote the ratio of the average utility of agents in N_p to the average utility of all agents under X^*. The *inter-type disparity parameter* $\beta(X^*)$ is defined as: $\beta(X^*) \triangleq \min_{p\in[k]} \beta_p(X^*) = \min_{p\in[k]} \frac{u_p(X^*)/|N_p|}{u(X^*)/n}$. Notice that $\beta(X^*)$ is in $(0, 1]$, can be computed in polynomial time and is fully independent of the type-block capacities (it uses X^*, an *unconstrained* optimal assignment). The closer $\beta(X^*)$ is to 1, the lower the disparity between average agents of different types under X^*.

Theorem 3: For any ASSIGNTC instance and any unconstrained optimal assignment $X^* \in \mathcal{X}$, we have $PoD(u) \leq \frac{1/\beta(X^*)}{\sum_{p\in[k]} \nu_p \min_{q\in[l]} \alpha_{pq}}$, where $\nu_p = \frac{|N_p|}{n}$ is the proportion of type $p \in [k]$ in the agent population.

For the Singapore public housing problem, if we use the ethnic proportions reported in the 2010 census report [33] i.e., $|N_1|/n = 0.741$ (Chinese), $|N_2|/n = 0.134$ (Malay), and $|N_3|/n = 0.125$ (Indian/Others) and the same block quotas α_{pq} as before, then in the case of no disparity (i.e., $\beta(X^*) = 1$), a simple calculation based on Theorem 3 shows that $PoD(u) \leq 1.43$ (approx.). Combining Theorems 2 and 3, if we plot the $PoD(u)$ against the disparity parameter $\beta(X^*)$ based on Singapore data, the point corresponding to any ASSIGNTC instance must lie in the shaded region of Figure 2.

2.1 Experimental Analysis

In this section, we present simulations of the ASSIGNTC problem using recent, publicly available datasets: Singapore demographic and housing allocation statistics, and the Chicago public school admission data. We compare the welfare of three assignment mechanisms: unconstrained optimal (maximizing welfare while ignoring the diversity constraints), constrained optimal (finding the optimal allocation under diversity constraints), and one-shot lottery-based (running a lottery with diversity constraints, as is the case for the HDB mechanism).[1] Conducting large-scale surveys to elicit agent preferences over items was beyond the scope of this work, so we simulated utilities based on reasonable models for both problems. We solved both the unconstrained and constrained social welfare maximizations using the Gurobi Optimizer. We refer the reader to `https://git.io/fNhhm` for full implementation details.

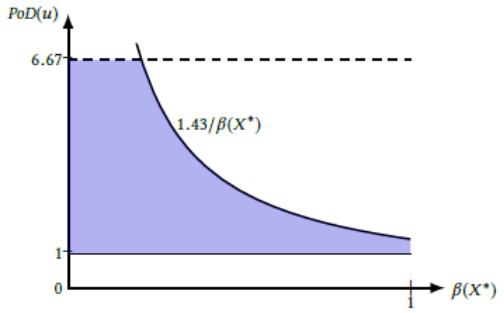
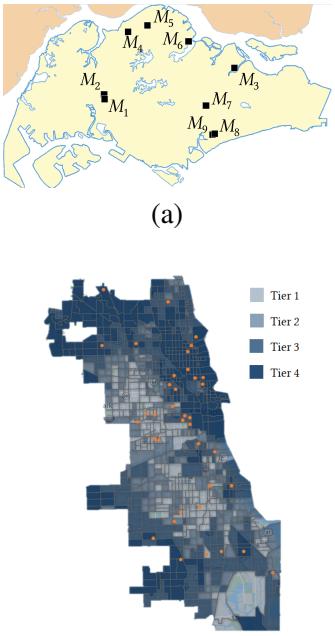


Figure 2: PoD vs disparity parameter for the HDB problem with current EIP quotas and ethnic proportions from Census 2010.

[1] We generated a uniformly random sequence over agents and assigned to each an item for which it has the highest utility among unassigned items in blocks for which that agent's ethnicity quota had not been filled yet. We abstract away other complications of actual lottery-based approaches used in our problem domains to focus on diversity constraints.

The Singapore Public Housing Allocation Problem. We collected data on the locations and numbers of flats of recent HDB housing development projects advertised over the second and third quarters of 2017.[2] Each development constitutes a block in our simulations, for a total of $m = 1350$ flats partitioned into $l = 9$ blocks $M_1, \ldots, M_9$ (Figure 3(a)), consisting of 128, 162, 156, 249, 108, 94, 104, 190, 159 flats respectively. There are pre-specified *categories* of flats, viz. 2-room flexi, 3-room, 4-room, and 5-room; our data set includes lower and upper bounds, $LB(t, q)$ and $UB(t, q)$ respectively, on the monthly cost (loan) for a flat of category t in block M_q for every t and q. We simulate 2 pools of n applicants whose ethnic composition follows the 2010 Singapore census report [33], as shown in Table 3. From the same census report, we collected the average salary $S(p)$ of each ethnicity group $p \in [k]$, given in Singapore dollars: $S(1) = S\$7,326$, $S(2) = S\$4,575$ and $S(3) = S\$7,664$. From publicly available data[3] on Singapore's Master Plan 2014,[4] we glean the locations of the geographic centers of the 55 planning areas that Singapore is divided into; we also obtained the population sizes of the three ethnicity groups under consideration in each planning area from the General Household Survey 2015 data available from the Department of Statistics, Singapore.[5] Our block capacities follow latest HDB block quotas [14]: $\alpha_{1q} = 0.87, \alpha_{2q} = 0.25, \alpha_{3q} = 0.15$ for every block M_q.



(a) (b)

Figure 3: (a) Singapore public housing block locations. (b) Tier statuses of Chicago census tracts and magnet school locations (orange dots).

We simulate 4 utility models; each has one parameter that does not come from the data: **(i)** ***Distance-based*** **(**$Dist(\sigma^2)$**)**: Each agent $i \in N$ has a preferred geographic location $\vec{a}_i \in \mathbb{R}^2$ (chosen uniformly at random within the physical landmass of Singapore) that she would like to live as close as possible to (say, the location of her parents' apartment, workplace, or preferred school). For every block M_q, we generate the utility of that agent i for apartment $j \in M_q$ by first drawing a sample from the normal distribution $\mathcal{N}(1/d(\vec{a}_i, loc(M_q)), \sigma^2)$, where $loc(M_q) \in \mathbb{R}^2$ is the geographical location of block M_q and $d(\cdot, \cdot)$ represents Euclidean distance, and then renormalizing to make the sum of utilities of each agent for all apartments in M equal to 1. **(ii)** ***Type-based*** **(**$Type(\sigma^2)$**)**: We assume that all agents of the same type (i.e., ethnic group) have the same preferred location (i.e., $\forall p \in [k], \forall i, i' \in N_p, \vec{a}_i = \vec{a}_{i'}$); the rest is similar to the above distance-based model. **(iii)** ***Project approval-based*** **(**$Project(\rho)$**)**: We construct, for each type, a categorical distribution over the 55 planning areas of Singapore, the probability of each area being proportional to the fraction of the sub-population of that type living in that area; for each agent i, we sample a preferred planning area from the above distribution corresponding to i's type; if a project M_q is within a radius ρ of the geographic center of agent i's preferred planning area, then i approves of the project i.e., $u(i, j) = 1\ \forall j \in M_q$, else i disapproves of the project i.e., $u(i, j) = 0\ \forall j \in M_q$. **(iv)** ***Price-based*** **(**$Price(\sigma^2)$**)**: Each agent $i \in N_p$ has a salary s_i that is generated according to the normal distribution $\mathcal{N}(S(p), \sigma^2)$. Each flat $j \in M_q$ of category t has a monthly cost p_j chosen uniformly in $[LB(t, q), UB(t, q)]$. The utility that agent i derives from flat j is then defined by $u(i, j) = 1/(p_j - \frac{s_i}{3})^2$, assuming that agent i is willing to pay one-third of her monthly salary on mortgage installments;[6] the rationale for the utility formula is that a high cost relative to the budget makes flats unaffordable, while a much lower cost indicates unsatisfactory quality, making the agent unhappy in both scenarios.

[2] http://www.hdb.gov.sg/cs/infoweb/residential/buying-a-flat/new/bto-sbf

[3] https://data.gov.sg/dataset/master-plan-2014-planning-area-boundary-web

[4] https://www.ura.gov.sg/Corporate/Planning/Master-Plan/

[5] https://www.singstat.gov.sg/publications/ghs/ghs2015content

[6] Inspired by the Singapore Central Provident Fund Board-endorsed "3-3-5 rule", as of 21 Sep 2017.

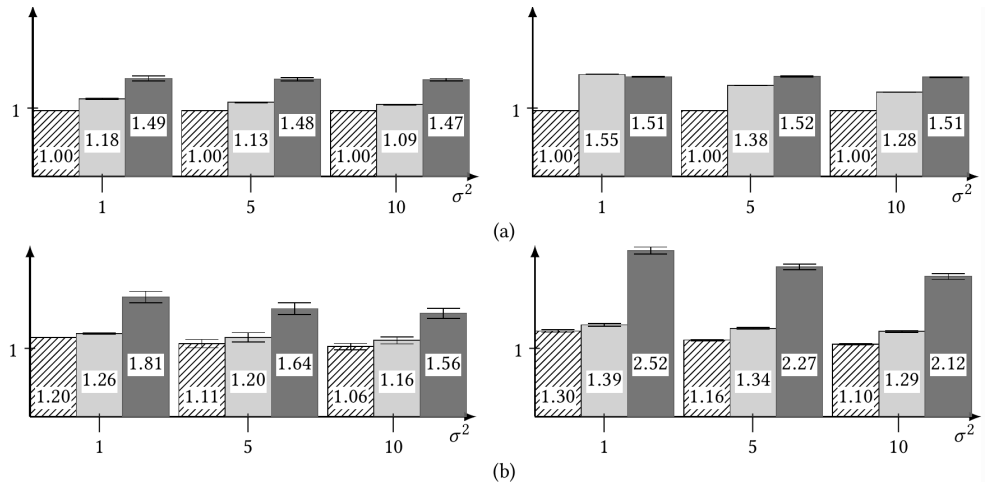


Figure 4: Averaged utility losses for (a) $Dist(\sigma^2)$ and (b) $Type(\sigma^2)$ with $n = m = 1350$ (left) and $n = 3000$, $m = 1350$ (right).

n	$\lvert N_1\rvert$	$\lvert N_2\rvert$	$\lvert N_3\rvert$
1350	1000	180	170
3000	2223	402	375

Table 3: #applicants (types 1, 2, and 3 are Chinese, Malay, Indian/Others respectively)

For each of our treatments (Figures 4–6), we plot the realized $PoD(u)$ (hatched bar), the theoretical upper bound on $PoD(u)$ as per Theorem 3 (dark gray bar), and the relative loss of the HDB lottery mechanism (i.e., the ratio of $\mathrm{OPT}(u)$ to the total utility of the assignment produced by the lottery mechanism) averaged over 100 agent permutations (light gray bar) against the values of the corresponding model parameters (σ^2 or ρ). In order to compare $Dist(\sigma^2)$ with $Type(\sigma^2)$, we vary both σ^2 in $\{1, 5, 10\}$ and n in $\{1350, 3000\}$; the results reported in Figures 4 are on average performance over 100 randomly generated instances. Our first observation is that, in all our experiments, $Dist(\sigma^2)$ exhibits virtually no utility loss due to the imposition of type-block constraints (see the hatched bars in Figures 4(a)). This is because utilities in $Dist(\sigma^2)$ are independent of ethnicity, resulting in a very low value for the inter-type disparity parameter β (indicated by the dark gray bars) — in fact, for any utility model where utilities are independent of ethnicity, the expected value of the disparity parameter is 1. For the $Type(\sigma^2)$ model-based utilities, the disparity parameter is somewhat higher (utilities do strongly depend on ethnicity), resulting in a higher $PoD(u)$ (see the hatched bars in Figures 4(b)). Despite making no attempt to optimize social welfare under type-block constraints, the HDB lottery mechanism does surprisingly well when the number of agents equals the number of apartments (see the light gray bars in the left part of Figure 4), extracting at least 84% of the optimal unconstrained welfare under the $Dist(\sigma^2)$ utility model, and at least 79% of the social welfare under the $Type(\sigma^2)$ model. However, the lottery-induced welfare is negatively impacted by the number of agents (see the light gray bars in the right part of Figure 4); for instance, it only extracts 65% of the optimal unconstrained welfare under $Dist(1)$ with $n = 3000$ and, in fact, the lottery-induced welfare loss for this treatment even exceeds the theoretical upper bound on the price of diversity.

For $Project(\rho)$, we use the fact that one degree of latitude or longitude at the location of Singapore corresponds to roughly 111 km to compute distances; we vary ρ in $\{5, 7.5, 10\}$ (in km). The results averaged over 100 runs are provided in Figure 5. In all instances, $PoD(u)$ is almost one and the lottery-induced welfare is also nearly as good, achieving at least 87% of the unconstrained optimum for 1350 agents and practically 100% for 3000 agents; the disparity parameter is also consistently close to its ideal value of 1, keeping the upper bound at around 1.45 regardless of the radius. Thus, this can be considered an example of a utility model for which the lottery mechanism virtually implements a constrained optimal allocation for a wide range of model parameters.

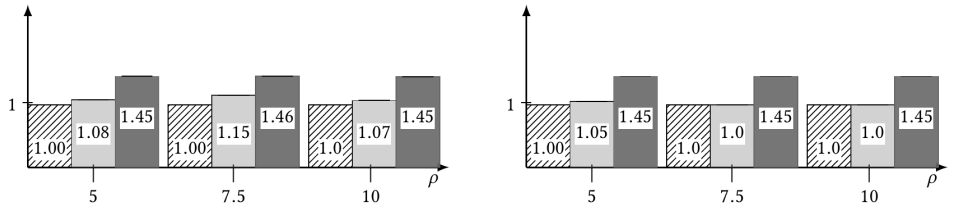


Figure 5: Averaged utility losses for $Project(\rho)$ with $n = m = 1350$ (left) and $n = 3000$, $m = 1350$ (right).

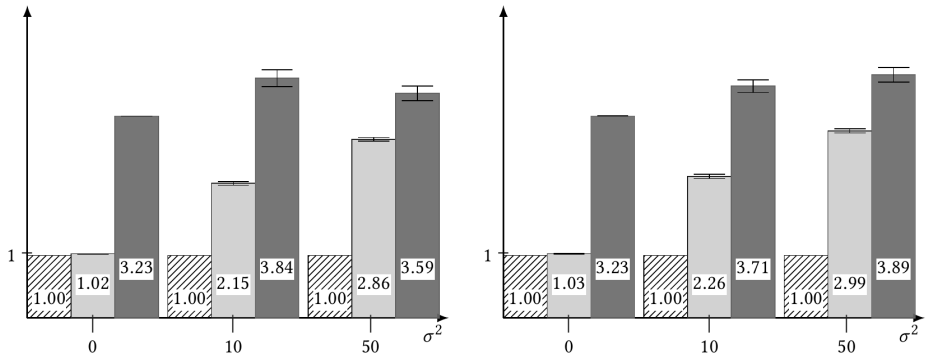


Figure 6: Averaged utility losses for $Price(\sigma^2)$ with $n = m = 1350$ (left) and $n = 3000$, $m = 1350$ (right).

Finally, for $Price(\sigma^2)$, we vary σ^2 in $\{0, 10, 50\}$; the results obtained by averaging over 100 runs are given in Figure 6. While the price of diversity is practically equal to one in all instances, the welfare loss observed with the lottery mechanism drastically increases with σ^2 (recall that agents from the same ethnicity group have identical preferences when $\sigma^2 = 0$): for instance, for 1350 agents, it extracts 98% of the optimal unconstrained welfare under $Price(0)$ while it only extracts 35% of this value under $Price(50)$. These numerical tests show that utility models exist for which the lottery mechanism may perform poorly compared to the optimal constrained allocation mechanism, even in allocation problems with a very low price of diversity.

Chicago Public School Admissions. Chicago Public Schools (CPS) is one of the largest school districts in the U.S.A.,[7] overseeing more than 600 schools of various types.[8] The application and selection processes for these schools involve a number of computerized lotteries, with a significant number of entry-level seats in magnet and selective enrollment schools being filled by lotteries based on a *tier system* based on the family *socio-economic status* (SES) as part of a social integration policy. The city computes a multi-factor, composite SES score for each of the census tracts that Chicago is divided into, and places each tract in one of four tiers based on its score in such a way that each tier contains (roughly) a quarter of school-aged children. The tier of a child is determined by their residential address. Of the seats in each school earmarked for a *citywide SES lottery* or *general lottery*, an equal number is allocated to each tier, and there is an upper limit on the number of schools that a child may apply to (see [30, 29]). We apply our setup to a simplified version of the CPS student-seat allocation scenario.

We collected data from the Chicago Public Schools website[9] on the locations of magnet schools in Chicago (which use a lottery mechanism to select students), as well as the total number of students enrolled in these schools in 2018, which we divided by 9 to obtain the approximate number of first-graders (there are nine grades

[7] http://www.cps.edu/About_CPS/At-a-glance/Pages/Stats_and_facts.aspx

[8] http://cpstiers.opencityapps.org/about.html

[9] http://cps.edu/Pages/home.aspx

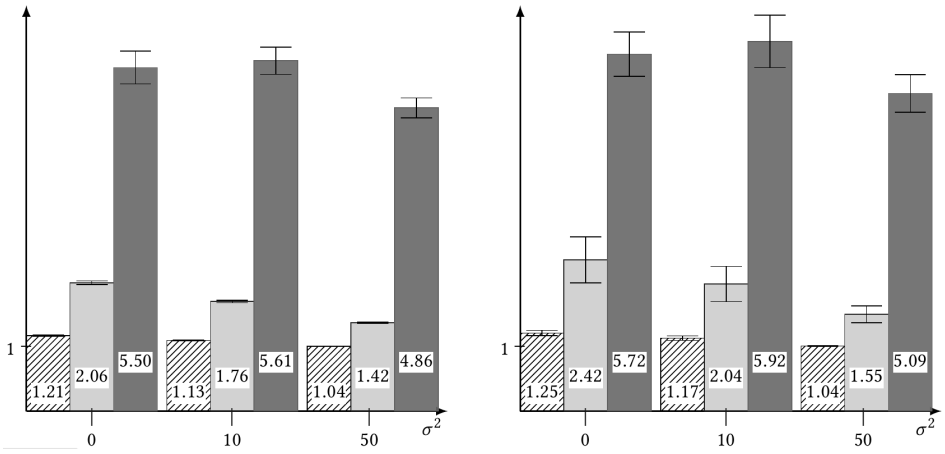


Figure 7: Averaged utility losses for $Dist(\sigma^2)$ with $n = m = 2261$ (left) and $n = 5000, m = 2261$ (right) in our Chicago-based simulations.

in total). This leads us to instances with $l = 37$ blocks (schools) and $m = 2261$ items (available seats) in total. In this school admission problem, students are partitioned into $k = 4$ types, viz. Tiers 1, 2, 3 and 4, depending on their residence (see Figure 3(b)[10]). In our experiments, we simulate 2 pools of n students where the type composition follows the real-world proportion, as shown in Table 4. Our type-block capacities are $\lambda_{pq} = 0.25|M_q|$ for every pair (p, q). For our student-seat utility simulations, we use the distance-based utility model $Dist(\sigma^2)$ we introduced in the housing domain, with the following important modifications: we choose the preferred location of a student uniformly at random from the collection of census tracts (polygons) belonging to her tier (see Figure 3(b)), where the position of every polygon is approximated by taking the averaged coordinates of its extreme points; we reset each student's utility to 0 for any school ranked 21st or lower in the preference ordering induced by her utilities (since students are allowed to apply to at most 20 schools), and then renormalize the utilities.

n	$\lvert N_1\rvert$	$\lvert N_2\rvert$	$\lvert N_3\rvert$	$\lvert N_4\rvert$
2261	613	622	533	493
5000	1355	1375	1180	1090

Table 4: #students (type p is Tier p for each $p \in [4]$.)

In our experiments, we vary σ^2 in $\{0, 10, 50\}$, and report 100-run averages of the same measurements as in the Singapore-based simulations (Figure 7). We observe that both the price of diversity (hatched) and the loss of the lottery mechanism (light gray bar) decrease as σ^2 increases, remaining well below the Theorem 3 bound (dark gray). However, the lottery mechanism loss is quite high in all instances and, just as in the Singapore case, gets worse for a higher number of students. Our experiments suggest that the lottery mechanism is better suited to problems with an equal number of agents and items.

3 Group Fairness in Allocation

Up to this point, we have only explored the welfare loss due to capacity constraints; however, allocative efficiency is only one facet of group fairness. In some settings, groups might receive an overall worse outcome from the allocation mechanism, as compared to others. This is known in the fairness literature as *envy*. In this section, we explore how envy-freeness (i.e., having no agent group envy another's allocation) affects the allocation outcome.

[10]Based on data from `http://cpstiers.opencityapps.org/` and `http://cps.edu/ScriptLibrary/Map-SchoolLocator/index.html`.

We work with a model similar to that in Section 2 with two differences: (i) the items M are not partitioned into blocks (or, equivalently, there is one block); (ii) we assume that for each agent $i \in N$ (resp. item $j \in M$), there is at least one item $j \in M$ (resp. agent $i \in N$) with $u(i,j) > 0$. We adopt an alternative view of diversity-respecting assignment as a task of allocating bundles (i.e., disjoint subsets of M) to *super-agents* (i.e., the types $N_1, \ldots, N_k$) in a manner that is both *fair* and *efficient* (see [8] for further details and complete proofs).

Each type p is represented by a *super-agent*. We define $v_p(S)$, the *valuation* of any super-agent $p \in [k]$ for any bundle of items $S \subseteq M$, as the maximum utilitarian social welfare of matching items in S to agents of type N_p; $v_p(\cdot)$ is a monotonic, submodular set function (see [8, Theorem 1]). Moreover, our model does not satisfy the additive bundle-valuation assumption or the public goods assumption prevalent in most prior work (see e.g., [12, 5, 2, 31] and references therein), necessitating novel solution techniques.

Definition 4 (Allocation): An *allocation* $\mathcal{A}$ is a collection of bundles $M_1^{\mathcal{A}}, \cdots, M_k^{\mathcal{A}}$, such that $M_1^{\mathcal{A}} \cup \ldots \cup M_k^{\mathcal{A}} \subseteq M$ and $M_p^{\mathcal{A}} \cap M_q^{\mathcal{A}} = \emptyset$ for all $p, q \in [k]$ with $p \neq q$, along with a maximum-USW matching between each type N_p and its allocated bundle $M_p^{\mathcal{A}}$ for all $p \in [k]$, thereby inducing a unique (N, M)-matching $X^{\mathcal{A}} = (x_{ij}^{\mathcal{A}})_{i \in N, j \in M}$.

We call $M_0^{\mathcal{A}} = M \setminus \cup_{p \in [k]} M_p^{\mathcal{A}}$ the set of *withheld items* under allocation $\mathcal{A}$. In general, withholding items means that an allocation violates, by definition, the *completeness* condition, a commonly used efficiency criterion. A type p's *marginal utility* for an item j is the difference in p's valuation of S with and without item j: $\Delta_p(S; j) \triangleq v_p(S \cup \{j\}) - v_p(S)$ if $j \notin S$; $\Delta_p(S; j) \triangleq v_p(S) - v_p(S \setminus \{j\})$ if $j \in S$. We say that an item $j \in M_p^{\mathcal{A}}$ is *unused* under $\mathcal{A}$ if it is either unassigned in the corresponding matching between N_p and $M_p^{\mathcal{A}}$ or is assigned to an agent $i \in N_p$ such that $u(i,j) = 0$. *Cleaning* is the procedure of revoking all unused items from an allocation and putting them in the withheld set. An item $j \in M$ is *wasted* by an allocation $\mathcal{A}$ if it is either withheld (i.e., $j \in M_0^{\mathcal{A}}$) or allocated to a type p and unused, although there is some other type q with $\Delta_q(M_q^{\mathcal{A}}; j) > 0$. A *non-wasteful* allocation has no wasted items. Non-wastefulness is a reasonable efficiency concept in this setting; in fact, it turns out to be a relaxation ([8, Lemma 1]) of a popular efficiency criterion, *Pareto optimality*: an allocation is Pareto optimal among types if the realized bundle-value of no type under this allocation can be strictly improved without strictly diminishing that of another.

We base our fairness criterion on the concept of *envy*: a type p envies a type q if $v_p(M_p^{\mathcal{A}}) < v_p(M_q^{\mathcal{A}})$; p envies q up to ν items, $\nu \in [|M_q^{\mathcal{A}}|]$, if there is a subset $C \subseteq M_q^{\mathcal{A}}$ of size $|C| = \nu$ such that $v_p(M_p^{\mathcal{A}}) \geq v_p(M_q^{\mathcal{A}} \setminus C)$ and, for every subset $C' \subseteq M_q^{\mathcal{A}}$ with $|C'| < \nu$, $v_p(M_p^{\mathcal{A}}) < v_p(M_q^{\mathcal{A}} \setminus C')$. Ideally, we want no type to envy another but such an allocation may not exist; a relaxation that always exists is the following:

Definition 5 (Envy-freeness up to one item [11]): Allocation $\mathcal{A}$ is *envy-free up to one item* (EF1) among types if for any two types $p, q \in [k]$, p either does not envy q or envies q up to one item i.e., there exists an item $j \in M_q^{\mathcal{A}}$ such that $v_p(M_p^{\mathcal{A}}) \geq v_p(M_q^{\mathcal{A}} \setminus \{j\})$.

We want our allocation to be not just EF1 among types (thereby respecting diversity) but also efficient in one of the ways discussed above. Our first result in this vein applies to the *binary utility model*: $u(i,j) \in \{0, 1\}$, $\forall i \in N, \forall j \in M$. This captures scenarios where each agent either approves or disapproves of an item but does not distinguish among its approved items. Moreover, in formal conversations with stakeholders, we have found that a binary utility model is likely consistent with how agents value items in many real-world situations, e.g., in housing markets, a potential buyer might want a flat of a particular category only (such as a 3BHK within a 5-km radius of her workplace), being indifferent among flats of the same category.

Theorem 6: For any problem instance with a binary utility model, Algorithm 1 computes in $poly(n, m)$ time an EF1 allocation that also maximizes the utilitarian social welfare of the induced (N, M)-matching.

It is easy to see that optimal utilitarian social welfare automatically implies Pareto optimality among types, and hence non-wastefulness. Thus, Algorithm 1 solves the fair and efficient allocation problem for binary utilities.

Algorithm 1: Maximum-size Matching with Envy-Induced Transfers

Compute a maximum-size matching of bipartite graph (N, M) such that there is an edge between i and j iff $u(i, j) = 1$, and clean the resulting allocation; designate the subset of items matched to agents in N_p as type p's allocated bundle $M_p^{\mathcal{A}}$ $\forall p \in [k]$.
/*Envy-Induced Transfers*/
while *there are two types* p, q *such that* p *envies* q *up to more than* 1 *item.* **do**
    Find item $j' \in M_q^{\mathcal{A}}$ such that $\Delta_p(M_p^{\mathcal{A}}; j') > 0$.
    $M_q^{\mathcal{A}} \leftarrow M_q^{\mathcal{A}} \backslash \{j'\}$; $M_p^{\mathcal{A}} \leftarrow M_p^{\mathcal{A}} \cup \{j'\}$.
    Compute a maximum-size (N_p, M_p)-matching.
end

For more general utilities in $\mathbb{R}_+$, an algorithm that guarantees a similarly fair and efficient allocation remains elusive. However, we note that it is possible to obtain a type-complete TEF1 allocation in polynomial time by a natural extension (called **L** hereafter) of the algorithm due to Lipton et al. [24]: iterate over the items $j \in M$, giving item j to a type, say p, that is currently not envied by any other type for its current bundle M_p; compute an optimal matching with the augmented bundle $M_p \cup \{j\}$; construct the *envy graph* where there is a directed edge from a type q to a type r whenever q envies r and eliminate any cycle in this graph by transferring the bundle of every type on this cycle to its predecessor on this cycle (to ensure that there is an unenvied type in each iteration), followed by re-matching within each such type. Although no item is withheld, it is possible for the final allocation to be wasteful: an item may be allocated to a type which has zero marginal utility for it or may *become* unassigned after a bundle is transferred between types.

One heuristic that could reduce waste is the following: in each iteration, find an item-type pair having the maximum marginal utility among all currently unenvied types and all unallocated items (breaking further ties uniformly at random, say), and allocate that item to that type. We call **L**, augmented with this heuristic, **H**.

		$\%Waste$	
Data set	#Items	**L**	**H**
UNEQUAL	50	13%	0
	100	39%	0
EQUAL	50	0%	0
	100	0.005%	0

To test how this marginal utility maximization heuristic performs in practice, we experimentally compared procedures **L** and **H** using the percentage of items wasted averaged over runs, denoted by $\%Waste$, as our performance metric. We simulated two data sets with $n = 100$ agents partitioned into $k = 3$ types: UNEQUAL (ethnic proportions following Singapore 2010 census [33]): $|N_1| = 74$, $|N_2| = 13$, $|N_3| = 13$; EQUAL: $|N_p| \approx n/k$ for all types $p \in [k]$. For each agent, we sampled m numbers uniformly at random from $[0, 1]$ and normalized them to generate utilities for all m items, with $m \in \{50, 100\}$. The results are shown in the adjoining table: the main observation is that **H** produces *zero* waste for *all* experiments. Thus, augmenting **L** with a simple heuristic produces a surprising improvement in performance over a wide range of problem parameters.

4 Discussion and future work

One of the extensions of the work presented here that we are currently pursuing is a rigorous analysis of the lottery mechanism with diversity quotas which we experimentally compared with our constrained optimization benchmark in Section 2.1. We are trying to assess whether certain lotteries are better than others in maintaining diverse but efficient outcomes in theory and in practice i.e., how the different parameters (the number of types, their respective percentage caps, sizes, and their utility structures) interact with the randomness of the draws to affect the welfare of the entire population as well as welfare-discrepancies among types.

One other major direction we are investigating is an extension of/alternative to Algorithm 1 for arbitrary

real-valued utilities. Several other possible approaches towards a tradeoff between fairness/diversity and efficiency are also worth exploring: diversity through the optimization of carefully constructed objective functions [23, 1]; extensions of non-envy-based fairness concepts (group-wise egalitarian welfare, maximin shares [3, 6], etc.) to our matching-based setting, and so on.

Acknowledgments

Chakraborty and Zick are supported by Singapore NRF Fellowship R-252-000-750-733, and Benabbou by ANR project 14-CE24-0007-01-Cocorico-CoDec; a major part of the work was done when Benabbou was a post-doctoral research fellow at National University of Singapore (2017-18), supported by Singapore NRF Fellowship R-252-000-750-733. The authors would like to thank Xuan-Vinh Ho, Jakub Sliwinski (supported by MOE grant R-252-000-6255-133), and Edith Elkind as co-authors of publications on which this article is based, and Ayumi Igarashi for insightful discussions. Thanks are also due to the anonymous reviewers of AAMAS 2018 and IJCAI 2019, and the attendees of COMSOC 2018 and FAMAS 2019 where parts of this work were presented.

References

[1] F. Ahmed, J. P. Dickerson, and M. Fuge. Diverse Weighted Bipartite b-Matching. *Proc. 26th IJCAI*, pp. 35–41, 2017.

[2] M. Aleksandrov and T. Walsh, Toby. Group envy freeness and group Pareto efficiency in fair division with indivisible items. *Künstliche Intelligenz*, pp. 57–72, 2018.

[3] S. Barman and S. K. Krishnamurthy. Approximation Algorithms for Maximin Fair Division. *Proc. 18th EC*, pp. 647–664, 2017.

[4] S. Barman, S. K. Krishnamurthy, and R. Vaish. Finding Fair and Efficient Allocations. *Proc. 19th EC*, pp. 557-574, 2018.

[5] S. Barman, S. K. Krishnamurthy, R. Vaish. Greedy algorithms for maximizing Nash social welfare. *Proc. 17th AAMAS*, pp. 7–13, 2018.

[6] S. Barman, A. Biswas, S. K. Krishnamurthy, and Y. Narahari. Groupwise maximin fair allocation of indivisible goods. *Proc. 32nd AAAI*, pp. 917–924, 2018.

[7] N. Benabbou, M. Chakraborty, X. V. Ho, J. Sliwinski, and Y. Zick. Diversity Constraints in Public Housing Allocation. *Proc. 17th AAMAS*, pp. 973–981, 2018.

[8] N. Benabbou, M. Chakraborty, E. Elkind, and Y. Zick. Fairness Towards Groups of Agents in the Allocation of Indivisible Items *Accepted to 28th IJCAI*, 2019.

[9] S. Bouveret, Y. Chevaleyre, and N. Maudet. Fair Allocation of Indivisible Goods. *Handbook of Computational Social Choice*, ed. F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia, ch. 12, pp. 284–310, Cambridge University Press, 2016.

[10] R. Bredereck, P. Faliszewski, A. Igarashi, M. Lackner, and P. Skowron. Multiwinner Elections with Diversity Constraints. *Proc. 32nd AAAI*, pp. 933–940, 2018.

[11] E. Budish The combinatorial assignment problem: Approximate competitive equilibrium from equal incomes. *Journal of Political Economy*, University of Chicago Press, Vol. 119, No. 6, pp. 1061–1103, 2011.

[12] I. Caragiannis, D. Kurokawa, H. Moulin, A. D. Procaccia, N. Shah, J. Wang. The unreasonable fairness of maximum Nash welfare. *Proc. 17th EC*, pp. 305–322, 2016.

[13] B. Chua. Race relations and public housing policy in Singapore. *Journal of Architectural and Planning Research*, pp. 343–354, 1991.

[14] Y. Deng, T. F. Sing, and C. Ren The story of Singapore's public housing: From a nation of home-seekers to a nation of homeowners. *The Future of Public Housing: Ongoing Trends in the East and the West*, ed. J. Chen, M. Stephens, and Y. Man, ch. 7, pp. 103–121, Springer, 2013. ISBN 978-3-642-41621-7.

[15] B. Fain, A. Goel, and K. Munagala. The core of the participatory budgeting problem. *Proc. 12th WINE*, pp. 384–399, 2016.

[16] D. Fragiadakis and P. Troyan. Improving matching under hard distributional constraints. *Theoretical Economics*, Wiley Online Library, Vol. 12, No. 2, pp. 863–908, 2017.

[17] Housing and Development Board, Singapore. Policy changes to support an inclusive and cohesive home [Press release]. 5 Mar 2010. http://www.nas.gov.sg/archivesonline/speeches/record-details/809e76bf-115d-11e3-83d5-0050568939ad.

[18] Housing and Development Board, Singapore. Annual Report 2016/2017: Key Statistics. 2017 http://www10.hdb.gov.sg/ebook/AR2018/key-statistics.html.

[19] Y. Kamada and F. Kojima. Efficient matching under distributional constraints: Theory and applications. *American Economic Review*, Vol. 105, No. 1, pp. 67–99, 2015.

[20] S. Y. Phang and K. Kim. Singapore's Housing Policies: 1960-2013. *Frontiers in Development Policy: Innovative Development Case Studies*, pp. 123–153, 2013.

[21] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics*, 2(1-2):83–97, 1955.

[22] D. Kurokawa, A. D. Procaccia, and J. Wang. When can the maximin share guarantee be guaranteed? *Proc. 30th AAAI*, pp. 523–529, 2016.

[23] J. Lang and P. K. Skowron Multi-Attribute Proportional Representation. *Proc. 30th AAAI*, pp. 530–536, 2016.

[24] R. J. Lipton, E. Markakis, E. Mossel, and A. Saberi. On approximately fair allocations of indivisible goods. *Proc. 5th EC*, pp. 125–131, 2004.

[25] L. Lovász and M. D. Plummer. Matching theory. *American Mathematical Society*, Vol. 367, 2009.

[26] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of SIAM*, Vol. 5, No. 1, pp. 32–38, 1957.

[27] Parliament of Singapore. Better racial mix in HDB housing estates. Parliament Debates: Official Report. 16 Feb 1989. Vol. 52, cols. 650-668.

[28] A. D. Procaccia and J. Wang. Fair enough: Guaranteeing approximate maximin shares. *Proc. 15th EC*, pp. 675–692, 2014.

[29] K. Quick. Chicago Public Schools: Ensuring Diversity in Selective Enrollment and Magnet Schools. *The Century Foundation.* https://tcf.org/content/report/chicago-public-schools, 14 Oct 2016.

[30] Chicago Public Schools Chicago Public Schools Policy Manual: Admissions Policy for Magnet, Selective Enrollment and Other Options For Knowledge Schools and Programs. Section 202.2. Board Report 17-0426-PO2. Available at http://policy.cps.edu/Policies.aspx, 26 Apr 2017.

[31] E. Segal-Halevi and W. Suksompong. Democratic Fair Allocation of Indivisible Goods. *Proc. 27th IJCAI*, pp. 482–488, 2018.

[32] L. L. Sim, S. M. Yu, and S. S. Han. Public housing and ethnic integration in Singapore. *Habitat International*, Vol. 27, No. 2, pp. 293–307, 2003.

[33] Department of Statistics, Singapore. Singapore 2010 Census: Key Indicators of the Resident Population. 2010.

[34] Department of Statistics, Singapore. Singapore in Figures. 2018.

[35] J. Stoyanovich, K. Yang, and H.V. Jagadish. Online set selection with fairness and diversity constraints. *Proc. 21st EDBT*, pp. 241–252, 2018.

[36] U.S. Department of Education, Office of Elementary and Secondary Education. Improving Outcomes for All Students: Strategies and Considerations to Increase Student Diversity. Washington, D.C. https://www.ed.gov/diversity-opportunity, 19 Jan 2017.

[37] M. Wong. Estimating the distortionary effects of ethnic quotas in Singapore using housing transactions. *Journal of Public Economics*, Vol. 115, pp. 131–145, 2014.

Towards Responsible Data-driven Decision Making in Score-Based Systems *

Abolfazl Asudeh[†], H. V. Jagadish[‡], Julia Stoyanovich[§]
[†]University of Illinois at Chicago, [‡]University of Michigan, [§]New York University
[†]asudeh@uic.edu, [‡]jag@umich.edu, [§]stoyanovich@nyu.edu

Abstract

Human decision makers often receive assistance from data-driven algorithmic systems that provide a score *for evaluating the quality of items such as products, services, or individuals. These scores can be obtained by combining different features either through a process learned by ML models, or using a weight vector designed by human experts, with their past experience and notions of what constitutes item quality. The scores can be used for different evaluation purposes such as ranking or classification.*

In this paper, we view the design of these scores through the lens of responsibility. We present technical methods (i) to assist human experts in designing fair and stable score-based rankings and (ii) to assess and (if needed) enhance the coverage of a training dataset for machine learning tasks such as classification.

1 Introduction

Big data technologies have affected every corner of human life and society. These technologies have made our lives unimaginably more shared, connected, convenient, and cost-effective. Using data-driven technologies gives us the ability to make wiser decisions, and can help make society safer, more equitable and just, and more prosperous. However, while having an enormous potential to help solve societal issues, *irresponsible* implementation of these technologies can not only fail to help, but may even make matters worse. Racial bias in predictive policing and data-driven judgeship, harming marginalized people and poor communities, and sexism in job recommendation systems are a few examples of such failures. In order to minimize societal harms of data-driven technologies, and to ensure that objectives such as fairness, equity, diversity, robustness, accountability, and transparency are satisfied, it is necessary to develop proper *tools, strategies, and metrics*.

Human decision makers often receive assistance from data-driven algorithmic systems that *provide a score for evaluating objects, including individuals*. The scores can be computed by combining different attributes either through a process learned by ML models (using some training data), or using a weight vector assigned by human experts. For example, a support vector machine learns the parameter values that define a linear separator in some regularized multi-dimensional feature space. Learning methods require that there be labeled data, and assume that there is some known ground truth.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

*This work was supported in part by NSF Grants No. 1741022 and 1926250.

In contrast, an expert-specified method does not require any labeled data, but may be ad-hoc. The scores are used either to evaluate an object independently (by only looking at its score) or in comparison with others.

Two major categories of methods that use scores to supporting decisions are (i) *Classification* and (ii) *Ranking*[1]. Classification is often used for predicting future outcomes based on historical data. Predicting recidivism and classifying individuals based on how likely they will commit a crime in the future is a societally important example of this kind. Ranking, on the other hand, is used for assignment by comparison on the existing data. For example, a company may rank its employees, and then reward high-ranked employees (with a raise or promotion) and fire low-ranked employees. College admissions is another example where the top-k applicants may be admitted by a college. Similarly, the international football association FIFA considers its rankings as "a reliable measure" for seeding the international tournaments such as the World Cup [3].

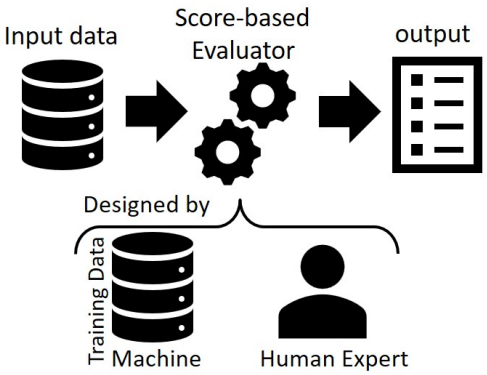


Figure 1: General architecture of score-based systems

Rankings are relative while labels in classification are absolute. That is, the rank of an individual depends on the others in the dataset, while class labels are assigned solely based on the score of an individual. The scoring mechanism for classification is usually learned by ML models. It can be a linear model such as regression and SVM, or a complex deep learning model. On the other hand, scoring mechanism for ranking is usually designed by human experts. US News university rankings, FIFA rankings, and CSRankings[2] are some of these examples.

Of course, the dichotomy above is not as clean as the preceding paragraph may suggest. Not all classification scoring is machine-learned. For example PSA (Public Safety Assessment) scores[3], used in data-driven judgeship, are human-designed. Similarly, ranking can be the basis for classification through the introduction of a cut-off rank, as in the case of college admissions.

Figure 1 shows the general architecture of score-based systems for data-driven decision making. The central component in these systems are the score-based evaluators that assign a score to each individual in the input data and generate the output by, for example, ranking or classifying the input. The output provides the evaluation of individuals that is used for decision making. Note that the scoring module can be designed by experts, or be learned by a machine, using some training data.

We, in our project Mithra, view *human experts* (for human-designed evaluators) and *training data* (for machine learned evaluators) as the keys to achieving responsibility in score-based systems. That is, for human-designed tasks such as ranking, we advocate designing assistive tools that help experts make sure their evaluators meet the objectives of fairness and stability. On the other hand, for machine learning tasks such as classification, we advocate assessing and repairing training data to make sure that, for example, the data is representative of minority groups [4], and models trained on that data do not reflect results of historical discrimination [5]. In the following, first in § 2, we explain our research for score-based ranking. Then in § 3, we provide our results for machine learning tasks such as classification by assessing and enhancing coverage for a (given) training dataset.

[1]Note that in some contexts ranking or classification is done without scoring. For instance, rank aggregation from partial ranked lists [1] or pairwise comparisons [2] is popular for group opinion collection. Our focus in this paper are evaluations (including ranking and classification) based on scores.

[2]csrankings.org

[3]www.psapretrial.org/about/factors

$\mathcal{D}$				f	f'
id	x_1	x_2	location	$\langle 1,1\rangle$	$\langle 1.11,.9\rangle$
t_1	0.63	0.71	Detroit	1.34	1.338
t_2	0.72	0.65	Chicago	1.37	1.384
t_3	0.58	0.78	Detroit	1.36	1.387
t_4	0.7	0.68	Chicago	1.38	1.389
t_5	0.53	0.82	Detroit	1.35	1.321
t_6	0.61	0.79	Chicago	1.4	1.388

Figure 2: Example 1-Data

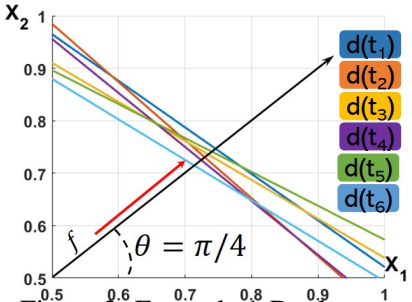


Figure 3: Example 1-Dual space

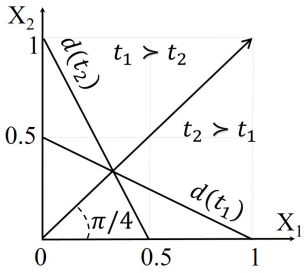


Figure 4: Ordering exchange between t_1:[1,2] and t_2:[2,1]

2 Responsible Ranking

Ranking of individuals is commonplace today, and is used, for example, for college admissions and employment. Score-based evaluators, designed by human experts, are commonly used for ranking, especially when there are multiple criteria to consider. The scores are usually computed by linearly combining (with non-negative weights) the relevant attributes of each individual from some dataset $\mathcal{D}$. Then, we sort the individuals in decreasing order of score and finally return either the full ranked list or its highest-scoring sub-set, the top-k.

Formally, we consider a dataset $\mathcal{D}$ to consist of n items, each with d scalar scoring attributes. In addition to the scoring attributes, the dataset may contain non-scoring attributes that are used for filtering, but they are not our concern here. Thus we represent an item $t \in \mathcal{D}$ as a d-length vector of scoring attributes, $\langle x_1, x_2, \ldots, x_d\rangle$. Without loss of generality, we assume that the scoring attributes have been appropriately transformed: normalized to non-negative values between 0 and 1, standardized to have equivalent variance, and adjusted so that larger values are preferred. A *scoring function* $f_{\vec{w}} : \mathbb{R}^d \to \mathbb{R}$, with weight vector $\vec{w} = \langle w_1, w_2, \ldots, w_d\rangle$, assigns the score $f_{\vec{w}}(t) = \Sigma_{j=1}^{d} w_j t[j]$ to any item $t \in \mathcal{D}$.

Linear scoring functions are straightforward to compute and easy to understand [6]. That is the reason they are popular for ranking, and for evaluation in general. However, it turns out that the rankings may highly depend on the design of these functions. To further explain this, let us consider the following toy example.

Example 1: *Consider a real estate agency with two offices in Chicago, IL and Detroit, MI. The owner assigns agents based on need (randomly) to the offices. By the end of the year, she wants to give a promotion to the "best" three agents. The criteria for choosing the agents are* x_1 : `sales` *and* x_2 : `customer satisfaction`. *Figure 2 shows the values in* $\mathcal{D}$*, after normalization. Considering the two criteria to be (roughly) equally important, the owner chooses the weights* $\vec{w} = \langle 1,1\rangle$ *for scoring. That is, the score of every agent is computed as* $f = x_1 + x_2$. *The 5th column in Figure 2 shows the scores, based on this function. According to function* f*, the top-3 agents are* t_6, t_4, *and* t_2, *with scores 1.4, 1.38, and 1.37, respectively. Note that, according to* f*, all top-3 agents are located in Chicago and no agent from Detroit is selected.*

The specific weights chosen have a huge impact on the score and hence rank for an item. In Example 1, the owner chose the weight vector $\vec{w} = \langle 1,1\rangle$, *in an ad-hoc manner*, without paying attention to the consequences. However, small changes in the weights could dramatically change the ranking. For example, the function f' with the weight vector $\vec{w'} = \langle 1.1,.9\rangle$ may be equally good for the owner and she may not even have a preference between $\vec{w}$ and $\vec{w}'$. Probably her choice of weights is only because $\vec{w}$ is more intuitive to human beings. The last column in Figure 2 shows the scores based on f', which produce the ranking $f' : \langle t_4, t_6, t_3, t_2, t_1, t_5\rangle$. Comparing it with the ranking generated by $f : \langle t_6, t_4, t_2, t_3, t_5, t_1\rangle$, one may notice that the rank of each and every individual has changed. More importantly, while according to f all promotions are given to the agents of the Chicago office, f' gives two promotions to Chicago and one to Detroit.

Many sports use ranking schemes. An example is the FIFA World Ranking of national soccer teams based on recent performance. FIFA uses these rankings as "a reliable measure for comparing national A-teams" [3].

Despite the trust placed by FIFA in these rankings, many critics have questioned their validity. University rankings is another example that is both prominent and often contested [7]: various entities, such as U.S. News and World Report, Times Higher Education, and QS, produce such rankings. Similarly, many funding agencies compute a score for a research proposal as a weighted sum of scores of its attributes. These rankings are, once again, impactful, yet heavily criticized. Similarly, in criminal justice, COMPAS [8] was originally intended to provide services and positive interventions, under resource constraints. That is, a score computed by COMPAS would then be used to rank individuals to prioritize access to services. Many other impactful examples can be mentioned, such as a company that evaluates its employees to promote some and let go some others, and a college admissions officer who decides to admit a small portion of the applicants.

Surprisingly, similar to Example 1, despite the enormous impact of score-based rankers, attribute weights are usually assigned in an ad-hoc manner, based only on intuitive reasoning and common-sense of the human designers. For instance, in the case of FIFA rankings, the scoring formula combines the past four years of performance of each team as $x_1 + 0.5x_2 + 0.3x_3 + 0.2x_4$, where x_i is the team's performance in the past i^{th} year. Of course, the designers tried to come up with a set of weights that make sense. For them $0.98x_1 + 0.51x_2 + 0.29x_3 + 0.192x_4$ would probably be equally acceptable, since the weight values are virtually identical: they choose the former formula simply because round numbers are more intuitive. This issue, in the context of university ranking, is further elaborated by Malcolm Gladwell in [7].

Assuming that the designers of rankings are willing to accept scoring functions similar to their initial functions, Mithra provides a toolbox and algorithms to help human experts practice responsible ranking. In the following, we start by providing some necessary background from computational geometry in § 2.2, followed by an explanation of fairness and stability in our framework in § 2.1.

2.1 Fairness and Stability Models

Decisions based on rankings may impact the lives of individuals and even influence societal policies. For this reason, it is essential to make the development and deployment of rankings transparent and otherwise principled. Also, since rankings highly depend on what weights are chosen in the scoring function, it is necessary to make sure that generated rankings are fair and robust.

2.1.1 Fairness

There is not a single universal definition of fairness. Impossibility theorems [9] have established that we cannot simultaneously achieve all types of fairness. Indeed, the appropriate definitions of fairness greatly depend on the context and on the perspective of the user. Sometimes, it may even be prescribed by law. As such, we consider a general definition of fairness in our work. Our focus is on societal (v.s. statistical) bias [10] and group fairness [11] (v.s. individual fairness [12]).

We consider some attributes, used for decision making (e.g. sales and customer satisfaction in Example 1), to be *non-sensitive*. Some other attributes, such as race and gender (and location in Example 1), we consider to be *sensitive*. We adopt the Boolean fairness model, in which a fairness oracle $\mathcal{O}$ takes as input an ordered list of items from $\mathcal{D}$, and determines whether the list satisfies a set of *fairness constraints*, defined over the sensitive attributes: $\mathcal{O} : \nabla_f(\mathcal{D}) \rightarrow \{\top, \bot\}$. A scoring function f that gives rise to a fair ordering over $\mathcal{D}$ is said to be *satisfactory*. For instance, in Example 1, assume that the owner knows that, because of some hidden factors, sales and customer satisfaction patterns are different in Chicago and Detroit. Hence, she considers the selection of the top-3 agents to be fair, if it assigns at least one of the promotions to each one of the offices. Note that according to this criterion, the ranking provided by $f = x_1 + x_2$ is not fair as it assigns all three promotions to the agents in Chicago. On the other hand the ranking generated by function $f' = 1.1x_1 + .9x_2$ assigns two of the promotions to Chicago and one to Detroit, and hence is considered to be fair.

2.1.2 Stability

We want a ranking to be *stable* with respect to changes in the weights used for scoring. Given a particular ranked list of items, one question a consumer will ask is: how robust is the ranking? If small changes in weights can change the ranked order, then there cannot be much confidence in the correctness of the ranking. We call a ranking of items *stable* if it is generated by a large portion of scoring functions in the neighborhood of the initial scoring function specified by the expert.

Every scoring function in a universe $\mathcal{U}^*$ of scoring functions induces a single ranking of the items. But each ranking is generated by many functions. For a dataset $\mathcal{D}$, let $\mathfrak{R}_\mathcal{D}$ be the set of rankings over the items in $\mathcal{D}$ that are generated by at least one scoring function $f \in \mathcal{U}^*$. Consider the set of scoring functions that generate a ranking $\mathfrak{r} \in \mathfrak{R}_\mathcal{D}$. Because this set of functions is continuous, we can think of it as a region in the space of all possible functions in $\mathcal{U}^*$. We use the region associated with a ranking to define the ranking's stability. The intuition is that a ranking is stable if it can be induced by a large set of functions. If the region of a ranking is large, then small changes in the weight vector are not likely to cross the boundary of a region and therefore the ranked order will not change. For every region R, let its volume, $vol(R)$, be the measure of its bulk. Given a ranking $\mathfrak{r} \in \mathfrak{R}_\mathcal{D}$, the stability of $\mathfrak{r}$ is the proportion of scoring functions in $\mathcal{U}^*$ that generate $\mathfrak{r}$. That is, stability is the ratio of the volume of the ranking region of $\mathfrak{r}$ to the volume of $\mathcal{U}^*$. Formally:

$$S_\mathcal{D}(\mathfrak{r}) = \frac{\text{vol}(R_\mathcal{D}(\mathfrak{r}))}{\text{vol}(\mathcal{U}^*)} \tag{20}$$

We emphasize that stability is a property of a ranking (not of a scoring function).

2.2 Geometric Interpretation

In the popular geometric model for studying data, each attribute is modeled as a dimension and items are interpreted as points in a multi-dimensional space. We transform this *Primal space* into a *dual space* [13].

We use the dual space in $\mathbb{R}^d$, where an item t is presented by a hyperplane $\mathsf{d}(t)$ given by the following equation of d variables $x_1 \dots x_d$:

$$\mathsf{d}(t):\ t[1] \times x_1 + \cdots + t[d] \times x_d = 1 \tag{21}$$

Continuing with Example 1, Figure 3 shows the items in the dual space. In $\mathbb{R}^2$, every item t is a 2-dimensional hyperplane (i.e. simply a line) given by $\mathsf{d}(t) : t[1]x_1 + t[2]x_2 = 1$. In the dual space, a scoring function $f_{\vec{w}}$ is represented as a ray starting from the origin and passing through the point $[w_1, w_2, ..., w_d]$. For example, the function f with the weight vector $\vec{w} = \langle 1, 1\rangle$ in Example 1 is drawn in Figure 3 as the origin-starting ray that passes through the point $[1, 1]$. Note that every scoring function (origin-starting ray) can be identified by $(d-1)$ angles $\langle \theta_1, \theta_2, \cdots, \theta_{d-1}\rangle$, each in the range $[0, \pi/2]$. Thus, given a function $f_{\vec{w}}$, its angle vector can be computed using the polar coordinates of w. For example, the function f in Figure 3 is identified by the angle $\theta = \pi/4$. There is a one-to-one mapping between these rays and the points on the surface of the origin-centered unit d-sphere (the unit hypersphere in $\mathbb{R}^d$), or to the surface of any origin-centered d-sphere. Thus, (the first quadrant of) the unit d-sphere represents the universe of functions $\mathcal{U}$.

Consider the intersection of a dual hyperplane $\mathsf{d}(t)$ with the ray of a function f. This intersection is in the form of $a \times \vec{w}$, because every point on the ray of f is a linear scaling of $\vec{w}$. Since this point is also on the hyperplane $\mathsf{d}(t)$, $t[1] \times a \times w_1 + \cdots + t[d] \times a \times w_d = 1$. Hence, $\sum t[j]w_j = 1/a$. This means that the dual hyperplane of any item with the score $f(t) = 1/a$ intersects the ray of f at point $a \times \vec{w}$. Following this, the ordering of the items based on a function f is determined by the ordering of the intersection of the hyperplanes with the vector of f. The closer an intersection is to the origin, the higher its rank. For example, in Figure 3, the intersection of the line t_6 with the ray of $f = x_1 + x_2$ is closest to the origin, and t_6 has the highest rank for f.

Consider the dual presentation of two items $t_1 : [1, 2]$ and $t_2 : [2, 1]$, shown in Figure 4, and a function that passes through this intersection. We name this function the *ordering exchange* between t_1 and t_2. That is because

this function partitions the space in two regions, where every function in the top-left region ranks t_1 higher than t_2 while every function in bottom-right ranks t_2 higher. In general, the ordering exchange between a pair of items t_i and t_j is identified by (the set of function on) the following origin-passing hyperplane:

$$\sum_{k=1}^{d}(t_i[k] - t_j[k])w_k = 0 \tag{22}$$

2.3 Designing Fair Ranking Schemes

We interpret fairness to mean that (a) disparate impact, which may arise as a result of historical discrimination, needs to be mitigated; and yet (b) disparate treatment cannot be exercised to mitigate disparate impact when the decision system is deployed. Disparate impact arises when a decision making system provides outputs that benefit (or hurt) a group of people sharing a value of a sensitive attribute more frequently than other groups of people. *Disparate treatment*, on the other hand, arises when a decision system provides different outputs for groups of people with the same (or similar) values of non-sensitive attributes but with different values of sensitive attributes. To avoid disparate treatment, it is desirable (and in many cases mandated by law) to not use information about an individual's membership in a protected group as part of decision-making. Following these, our goal is to build a system that helps human expert to design fair ranking schemes, in the sense that those *both* mitigate disparate impact (by ensuring that appropriate proportionality constraints are satisfied) *and* do not exercise disparate treatment (by not explicitly using information about an individual's membership in a protected group) during deployment. That is, a *single* evaluator will be used for all items in the dataset, irrespective of their membership in a protected group.

Our goal [4] is to build a system to assist a human designer of a scoring function in tuning attribute weights to achieve fairness. Formally, our *closest satisfactory function* problem is: Given a dataset $\mathcal{D}$ with n items over d scalar scoring attributes, a fairness oracle $\mathcal{O} : \nabla_f(\mathcal{D}) \rightarrow \{\top, \bot\}$, and a linear scoring function f with the weight vector $\vec{w} = \langle w_1, w_2, \cdots, w_d\rangle$, find the function f' with the weight vector $\vec{w'}$ such that $\mathcal{O}(\nabla_{f'}(\mathcal{D})) = \top$ and the angular distance between $\vec{w}$ and $\vec{w'}$ is minimized.

Since the tuning process does not occur too often, it may be acceptable for it to take some time. However, we know that humans are able to produce superior results when they get quick feedback in a design or analysis loop. Ideally, a designer of a ranking scheme would want the system to support her work through interactive response times. Our goal is to meet this need, to the extent possible, by providing a query answering system. From the system's viewpoint, the challenge is to propose similar weight vectors that satisfy the fairness constraints, in interactive time. To accomplish this, our solution operates with an offline phase and then an online phase. In the offline phase, we process the dataset, and develop data structures that will be useful in the online phase. In the online phase, the user specifies a *query* in the form of a scoring function f. If the ranking based on f does not meet the predefined fairness constraints, we suggest to the user an alternative scoring function that is both satisfactory and similar to f. The user may accept the suggested function, or she may decide to manually adjust the query and invoke our system once again.

The notion of ordering exchanges, explained in § 2.2 is a key in the preprocessing phase. Consider the set of ordering exchange hyperplanes between all pairs of items in $\mathcal{D}$. Similar to Figure 4, each hyperplane $h_{i,j}$ (the ordering exchange between t_i and t_j) partitions the function space $\mathcal{U}$ in two regions where in one region t_i outranks t_j while in the other t_j is ranked higher. The collection of these hyperplanes provide an arrangement [14] in the form of a dissection of the space into origin-starting connected d-cones with convex surfaces, we call *ranking regions*. All functions in a ranking region generate the same ranking while every ranking is generated by the functions of (at most) one ranking region. Hence, in the offline time it is enough to identify the satisfactory ranking regions whose rankings satisfy fairness constraints.

For 2D, we design a raw-sweeping algorithm. At a high level, using a min-heap for maintaining the ordering exchanges, the algorithm sweeps a ray from the x to y-axis. It first orders the items based on the x-axis and

gradually updates the ordering as it visits an ordering exchange along the way. As the algorithm moves from one ranking region to the other, it checks if the new ranking is fair, and if so, marks the region as satisfactory. Each satisfactory region in 2D is identified by two angles as its beginning and end. We construct a the sorted list of (the borders of) satisfactory regions in the offline phase. Given a query function f in online phase, we apply a binary search on the sorted list. If f falls in a satisfactory region, the algorithm returns f, otherwise it returns the closest satisfactory border to f.

Discovering the satisfactory regions in MD is challenging when there are more than two attributes. That is because the complexity of the arrangement of ordering exchanges is exponential in the number of attributes, d. Even given the satisfactory regions, answering user queries in interactive time is not possible. The reason is that we need to solve a non-linear programming problem for each satisfactory region, before answering each query. To address this issue, we propose an approximation algorithm for obtaining answers quickly, yet accurately. Our approach relies on first partitioning the function space, based on a user-controlled parameter N, into N equi-volume cells, where each cell is a hypercube of $(d-1)$-dimensions. During preprocessing, we assign a satisfactory function f'_c to every cell c such that, for every function f, the angle between f and f'_c is within a bounded threshold (based on the value of N) from f and its optimal answer. To do so, we first identify the cells that intersect with a satisfactory region, and assign the corresponding satisfactory function to each such cell. Then, we assign the cells that are outside of the satisfactory regions to the nearest discovered satisfactory function. In the online phase, given an unsatisfactory function f, we need to find the cell to which f belongs, and to return its satisfactory function. This can be done in $O(\log N)$ by performing binary searches on the partitioned space.

2.4 Obtaining Stable Rankings

Magnitude of the ranking regions that produces an observed ranking identify its stability. Stability is a natural concern for consumers of a ranked list. If a ranking is stable, then the same ranking would be obtained for many choices of weights. But if this region is small, then we know that only a few weight choices can produce the observed ranking. This may suggest that the ranking was engineered or "cherry-picked" by the producer to obtain a specific outcome. Human experts who produce scoring functions for generating the rankings desire to produce stable results. We argued in [15] that stability in a ranked output is an important aspect of algorithmic transparency, because it allows the producer to justify their ranking methodology, and to gain the trust of consumers. Of course, stability cannot be the only criterion in the choice of a scoring function: the result may be weights that seem "unreasonable" to the ranking producer. To support the producer, we allow them to specify a range of reasonable weights, or an *acceptable region* in the space of functions, and help the producer find stable rankings within this region.

We develop a framework [16] that can be used to assess the stability of a provided ranking and to obtain a stable ranking within the the acceptable region of scoring functions. We address the case where the user cares about the rank order of the entire set of items, and also the case where the user cares only about the top-k items. We focus on efficiently evaluating an operator we call GET-NEXT, which can be used to discover the stable rankings, ordered by their stability. Formally, for a dataset $\mathcal{D}$, a region of interest $\mathcal{U}^*$, and the top-$(h-1)$ stable rankings in $\mathcal{U}^*$, discovered by the previous GET-NEXT calls, our goal is to find the h-th stable ranking $\mathfrak{r} \in \mathfrak{R}$. Note that the GET-NEXT operator enables discovering the top-ℓ stable rankings, for any arbitrary ℓ. That simply can be done by calling the operator ℓ times. Our technical contribution for 2D is similar to the one for fair rankings. The 2D algorithm first discovers the ranking regions in $\mathcal{U}^*$ by sweeping a ray in it. The discovered rankings, along with their stabilities, are moved to a heap data structure. Then, every call of GET-NEXT returns the next stable ranking from heap.

For MD, we design a threshold-based algorithm that uses an arrangement [4] tree data structure, AKA cell tree [17], to partially construct the arrangement of ordering exchange hyperplanes. Specifically, given that our objective is to find stable rankings and that the user will likely be satisfied after observing *a few* rankings, rather than discovering all possible rankings, we target the discovery of only the next stable ranking and delay the

arrangement construction for other rankings. Arrangement construction is an iterative process that starts by partitioning the space into two half-spaces by adding the first hyperplane. The construction then iteratively adds the other hyperplanes; to add a new hyperplane, it first identifies the set of regions in the arrangement of previous hyperplanes with which the new hyperplane intersects, and then splits each such region into two new regions. The GET-NEXT operator, however, only breaks down the largest region at every iteration, delaying the construction of the arrangement in all other regions. Please refer to [16] for more details about the algorithm.

While being efficient in practice for medium-size settings, the algorithms based on arrangement construction are not scalable, as their worst-case complexities are cursed by the complexity of the arrangement. Next, we discuss function sampling as a powerful technique for aggregate estimation using Monte-carlo methods [18], as well as an effective technique for search-space exploration.

2.5 Function Sampling for Scalability

Uniform sampling from the scoring function space enables designing randomized algorithms for evaluating and designing score-based evaluators. In the following, we first discuss sampling from the complete function space and then propose an efficient unbiased sampling from a region of interest $\mathcal{U}^*$.

As explained in § 2.2, there is a 1-1 mapping between the universe of scoring functions and the points on the surface of (the first quadrant of) the unit d-sphere. That is, every point one the surface of the d-sphere correspond to a scoring function and vice versa. Hence, the problem of choosing functions uniformly at random from $\mathcal{U}$ is equivalent to choosing random points from the surface of a d-sphere. As also suggested in [19], we adopt a method for uniform sampling of the points on the surface of the unit d-sphere [20, 21]. Rather than sampling the angles, this method samples the weights using the *Normal distribution*, and normalizes them. This works because the normal distribution function has a constant probability on the surfaces of d-spheres with common centers [21]. Therefore, in order to generate a random function in $\mathcal{U}$, we set each weight as $w_i = |\mathcal{N}(0, 1)|$, where $\mathcal{N}(0, 1)$ draws a sample from the standard normal distribution.

In order to compute an aggregate (or conduct exploration) by perturbing in a region of interest $\mathcal{U}^*$ in the neighborhood of some function f, we need to only sample from the set of functions with the maximum angle around the ray of f. An acceptance-rejection method [22] can be used for this purpose. That is, to draw a sample, uniformly at random, from $\mathcal{U}$ and accept it, if it belongs to $\mathcal{U}^*$. The efficiency of this method, however, depends on the volume of $\mathcal{U}^*$, as if it is small, the algorithm will reject most of the generated samples. Alternatively, we propose a sampler that works based on the following observation: modeling $\mathcal{U}^*$ as the surface unit d-spherical cap, each Riemannian piece of the surface forms a $(d-1)$-sphere. Following this observation, our sampler first selects a Riemannian piece, randomly, proportional to its volume. Then it uses the Normal distribution to draw a sample from the surface of the Riemannian piece. Please refer to [16] for more details about the design of this sampler. We would like to emphasize that the proposed sampler has a *linear complexity* to the number of attributes d. It therefore provides a powerful tool for studying the score-based evaluators in higher dimensions.

We use function sampling for different purposes, including (i) evaluating the stability of a given ranking, (ii) designing a randomized algorithm for finding the stable rankings, and (iii) on-the-fly query processing for discovering fair functions. For (i) and (ii), in [16], we design Monte-carlo methods [18] that consume a set of N function samples for finding the rankings in $\mathcal{U}^*$ and computing their stabilities. For (iii), function sampling provides a heuristics for on-the-fly fair ranking scheme query processing in large-scale settings [4].

We used function sampling in MithraRanking [23], our web application[4], designed for responsible ranking design. After uploading a dataset, or choosing among available datasets, the application allows the user to specify the fairness constraints she wants to satisfy. For instance, in Figure 5, the user has added a constraint that the top-30% of the ranking should contain at most 30% with age more than 56 years old. Note that the interface gives the user the ability to add multiple fairness constraints. She then, as in Figure 6, specifies the weight vector

[4] http://mithra.eecs.umich.edu/demo/ranking/

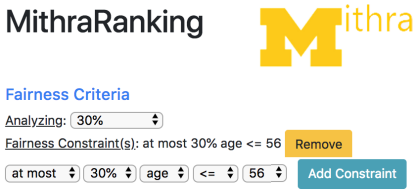


Figure 5: Specifying fairness constraints

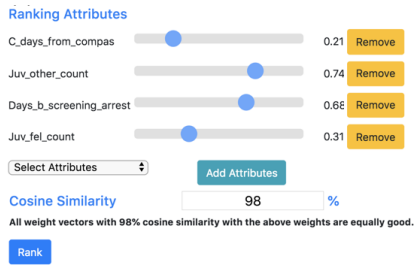


Figure 6: Specifying a weight vector

Figure 7: System results

of the initial scoring function and a region of interest around it, by specifying a cosine similarity. The system then ranks the data based on the specified function and checks if the ranking satisfies the fairness criteria. It also draws unbiased function samples from the region of interest to estimate the stability of the generated ranking. The system also uses the samples for finding the most stable rankings in the region of interest, the most similar fair function to the initial function, and a function (not necessarily the most similar) that generates a fair and stable ranking (Figure 7). The user can then accept any of those suggestions and change the ranking accordingly.

3 Coverage in Training Data

So far in this paper, we discussed responsible design of scoring functions by a human expert. Scoring models are also used for tasks such as classification and prediction Such scoring models can be complex and are often determined using machine learning techniques. An essential piece to the learning is a labeled training dataset. This dataset could be collected prospectively, such as through a survey or a scientific experiment. In such a case, a data scientist may be able to specify requirements such as representation and coverage. However, more often than not, analyses are done with data that has been acquired independently, possibly through a process on which the data scientist has limited, or no, control. This is often called "found data" in the data science context. It is generally understood that the training dataset must be representative of the distribution from which the actual test/production data will be drawn. More recently, it has been recognized that it is not enough for the training data to be representative: it must include enough examples from less popular "categories", if these categories are to be handled well by the trained system. Perhaps the best known story underlining the importance of this inclusion is the case of the "google gorilla" [24]. An early image recognition algorithm released by Google had not been trained on enough dark-skinned faces. When presented with an image of a dark African American, the algorithm labeled her as a "gorilla". While Google very quickly patched the software as soon as the story broke, the question is what it could have done beforehand to avoid such a mistake in the first place. The Google incident is not unique: there have been many other such incidents. For example, Nikon introduced a camera feature to detect whether humans in the image have their eyes open – to help avoid the all-too-common situation of the camera-subject blinking when the flash goes off resulting in an image with eyes closed. Paradoxically for a Japanese company, their training data did not include enough East Asians, so that the software classified many (naturally narrow) open Asian eyes as closed [25].

The problem becomes critical when the data is used for training models for data-driven algorithmic decision making. For example, consider a tool designed to help judges in sentencing criminals by predicting how likely an individual is to re-offend. Such a tool can provide insightful signals for the judge and have the potential to make society safer. On the other hand, a wrong signal can have devastating effects on individuals' lives. So it is important to make sure that the tool is trained on data that includes adequate representation of individuals similar to each person that will be scored by it. In [26], we study a real dataset of criminals used for building such a tool, published by Propublica [8]. We demonstrate that a model with an acceptable overall accuracy had an accuracy worse than random guess for Hispanic females, due to inadequate representation.

While Google's resolution to the gorilla incident was to "ban gorillas" [27], a better solution is to ensure that

the training data has enough entries in each category. Referring to the issue as "disparate predictive accuracy", [28] also highlights that the problem often is due to the insufficient or skewed sample sizes. If the only category of interest were race, as in (most of) the examples above, there are only a handful of categories and this problem is easy. However, in general, objects can have tens of attributes of interest, all of which could potentially be used to categorize the objects. For example, survey scientists use multiple demographic variables to characterize respondents, including race, sex, age, economic status, and geographic location. Whatever be the mode of data collection for the analysis task at hand, we must ensure that there are enough entries in the dataset for each object category. Drawing inspiration from the literature on diversity [29], we refer to this concept as *coverage*.

Lack of coverage in a dataset also opens up the room for adversarial attacks [30]. The goal in an adversarial attack is to generate examples that are misclassified by a trained model. Poorly covered regions in the training data provide the adversary with opportunities to create such examples. For instance, consider the gorilla incident again. Knowing that black people are under-represented in the dataset gives the adversary the information that the models trained using this dataset are not well-trained for this category. The adversary can use this information to generate examples that are misclassified by the model.

Our objective here is two-fold. First, we want to help the dataset users to assess the coverage, as a characterization, of a given dataset, in order to understand such vulnerabilities. For example, we use information about lack of coverage as a widget in the nutritional label [15] of a dataset, in our MithraLabel system[5][31]. Once the lack of coverage has been identified, next we would like to help data owners improve coverage by identifying the smallest number of additional data points needed to hit all the "large uncovered spaces".

Given multiple attributes, each with multiple possible values, we have a combinatorial number of possible *patterns*, as we call combinations of values for some or all attributes. Depending on the size and skew in the dataset, the coverage of the patterns will vary. Given a dataset, our first problem is to efficiently identify patterns that do not have sufficient coverage (the learned model may perform poorly in portions of the attribute space corresponding to these patterns of attribute values). It is straightforward to do this using space and time proportional to the total number of possible patterns. Often, the number of patterns with insufficient coverage may be far fewer. In [26], we develop techniques, inspired from set enumeration and association rule mining (*apriori*) [32], to make this determination efficient.

A more interesting question for the dataset owners is what they can do about lack of coverage. Given a list of patterns with insufficient coverage, they may try to fix these, for example by acquiring additional data. In the ideal case, they will be able to acquire enough additional data to get sufficient coverage for all patterns. However, acquiring data has costs, for data collection, integration, transformation, storage, etc. Given the combinatorial number of patterns, it may just not be feasible to cover all of them in practice. Therefore, we may seek to make sure that we have adequate coverage for at least any combination of ℓ attributes, where we call ℓ the *maximum covered level*. Alternatively, we could identify important pattern combinations by means of a *value count*, indicating how many combinations of attribute values match that pattern. Hence, our goal becomes to determine the patterns for the minimum number of items we must add to the dataset to reach a desired maximum covered level or to cover all patterns with at least a specified minimum value count.

We consider the low-dimensional categorical (sensitive) attributes $\mathcal{A} = \{A_1, A_2, ..., A_d\}$ such as `race`, `gender`, and `age` for studying coverage. Where attributes are continuous valued or of high cardinality, we consider using techniques such as (a) bucketization: putting similar values into the same bucket, or (b) considering the hierarchy of attributes in the data cube for reducing the cardinality of any one attribute. To capture lack of coverage in the dataset, we define a pattern P as a vector of size d, in which $P[i]$ is either X (meaning that its value is unspecified) or is a value of attribute A_i. We name the elements with value X as non-deterministic and the others as deterministic. We say an item t *matches* a pattern P (written as $M(t, P) = \top$), if for all i for which $P[i]$ is deterministic, $t[i]$ is equal to $P[i]$. For example, consider the pattern $P = X1X0$ on four binary attributes A_1 to A_4. It describes the value combinations that have the value 1 on A_2 and 0 on A_4. Hence, for example,

[5]http://mithra.eecs.umich.edu/demo/label/

$t_1 = [1, 1, 0, 0]$ matches P, while $t_3 = [1, 0, 1, 0]$ does not match it, because $P[2] = 1$ and $t_3[2] = 0$. Using the patterns to describe the space of value combinations, we define the coverage of a pattern P as the number of items in $\mathcal{D}$ that match it. We say a pattern P is dominated by another pattern P' if all value combinations matching it also match P'. Our (lack of coverage) identification problem is to discover *Maximal Uncovered Patterns* (MUPs), the set of uncovered patterns (patterns with coverage less than a threshold) that are not dominated by another uncovered pattern. This problem is #P-complete. At a high level, we define a directed acyclic graph (DAG) that captures the domination relation between the patterns and transform the problem into an enumeration on this graph while using the monotonicity property of coverage for pruning the search space.

We note that not all combinations of attribute values are of interest. Some may be extremely unlikely, or even infeasible. For example, we may find few people with attribute `age` as "teen" and attribute `education` as "graduate degree". A human expert, with sufficient domain knowledge, is required to be in the loop for (i) identifying the attributes of interest, over which coverage is studied, (ii) setting up a *validation oracle* that identifies the value combinations that are not realistic, and (iii) identifying the uncovered patterns and the granularity of patterns that should get resolved during the coverage enhancement.

Our coverage enhancement problem is: given a dataset $\mathcal{D}$, its set of material MUPs $\mathcal{M}_{\mathcal{D}}$, and a positive integer number λ, to determine the minimum set of additional tuples to collect such that, after the data collection, the maximum number of deterministic values in any MUP is at least λ. The problem, using a polynomial-time reduction from the vertex cover problem, turns out to be NP-complete. Since a single tuple could contribute to the coverage of multiple patterns, we shall show that this problem translates to a hitting set [33] instance. Using this transformation, we show that the greedy approach provides a logarithmic approximation-ratio for the problem. Given the exponential number of value combinations, the direct implementation of hitting set techniques can be very expensive. Hence, we also provide an efficient implementation of the greedy approach.

4 Final Remarks

In this article we explained our results towards responsible data-driven decision making in score-based systems. The scores, in these systems, are obtained by combining some features using either machine learning models or human-designed weight vectors. We provided our results for (i) assisting the experts to design fair and stable rankings, and (ii) assessing and enhancing coverage in a (given) training dataset for tasks such as classification.

So far, in (i) our focus has been on ranking, where the scores are used for comparing the items in a pool. Human-designed scores are also used for tasks such as classification. Extending our results for these tasks is part of our future work. Also, we would like to adopt the proposed techniques for linear machine learning models. The idea is to first train a machine learning model and then adjust the model to, for example, satisfy some fairness criteria. A similar idea can also be applied for designing ensemble methods for combining the outcome of multiple ML models. In (ii), we used a fixed threshold across different value combinations, representing "minor subgroups". We consider further investigations on identifying threshold value and minor subgroups for future work. We will also investigate other properties (in addition to coverage) for assessing and enhancing the fitness of training data for responsible data science tasks.

References

[1] A. Marcus, E. Wu, D. Karger, S. Madden, and R. Miller. Human-powered sorts and joins. *PVLDB*, 5(1):13–24, 2011.

[2] A. Asudeh, G. Zhang, N. Hassan, C. Li, and G. V. Zaruba. Crowdsourcing pareto-optimal object finding by pairwise comparisons. In *CIKM*, pages 753–762. ACM, 2015.

[3] FIFA. Fifa/coca-cola world ranking procedure. www.fifa.com/fifa-world-ranking/procedure/men.html, 2008.

[4] A. Asudeh, H. Jagadish, J. Stoyanovich, and G. Das. Designing fair ranking schemes. In *SIGMOD*. ACM, 2019.

[5] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu. Capuchin: Causal database repair for algorithmic fairness. In *SIGMOD*. ACM, 2019.

[6] A. Asudeh, N. Zhang, and G. Das. Query reranking as a service. *PVLDB*, 9(11):888–899, 2016.

[7] M. Gladwell. The order of things: What college rankings really tell us. *The New Yorker Magazine*, 2011.

[8] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: Risk assessments in criminal sentencing. *ProPublica*, 2016.

[9] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

[10] A. Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *FAT**, 2018.

[11] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *NIPS*, pages 3315–3323, 2016.

[12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *ITCS*, 2012.

[13] H. Edelsbrunner. *Algorithms in combinatorial geometry*, volume 10. Springer Science & Business Media, 2012.

[14] T. Jan. Redlining was banned 50 years ago. it's still hurting minorities today. Washington Post, 2018.

[15] K. Yang, J. Stoyanovich, A. Asudeh, B. Howe, H. Jagadish, and G. Miklau. A nutritional label for rankings. In *SIGMOD*, pages 1773–1776. ACM, 2018.

[16] A. Asudeh, H. Jagadish, G. Miklau, and J. Stoyanovich. On obtaining stable rankings. *PVLDB*, 12(3):237–250, 2018.

[17] B. Tang, K. Mouratidis, and M. L. Yiu. Determining the impact regions of competing options in preference space. In *SIGMOD*, 2017.

[18] C. P. Robert. *Monte carlo methods*. Wiley Online Library, 2004.

[19] A. Asudeh, A. Nazi, N. Zhang, G. Das, and H. Jagadish. RRR: Rank-regret representative. In *SIGMOD*. ACM, 2019.

[20] M. E. Muller. A note on a method for generating points uniformly on n-dimensional spheres. *CACM*, 2(4), 1959.

[21] G. Marsaglia et al. Choosing a point from the surface of a sphere. *The Annals of Math. Statistics*, 43(2), 1972.

[22] S. Lucidl and M. Piccioni. Random tunneling by means of acceptance-rejection sampling for global optimization. *Journal of optimization theory and applications*, 62(2):255–277, 1989.

[23] Y. Guan, A. Asudeh, P. Mayuram, H. Jagadish, J. Stoyanovich, G. Miklau, and G. Das. Mithraranking: A system for responsible ranking design. In *SIGMOD*, 2019.

[24] M. Mulshine. A major flaw in google's algorithm allegedly tagged two black people's faces with the word 'gorillas'. Business Insider, 2015.

[25] A. Rose. Are face-detection cameras racist? Time Business, 2010.

[26] A. Asudeh, Z. Jin, and H. Jagadish. Assessing and remedying coverage for a given dataset. *ICDE*, 2019.

[27] A. Hern. Google's solution to accidental algorithmic racism: ban gorillas. The Guardian, 2018.

[28] I. Chen, F. D. Johansson, and D. Sontag. Why is my classifier discriminatory? In *NeurIPS*, 2018.

[29] M. Drosou, H. Jagadish, E. Pitoura, and J. Stoyanovich. Diversity in big data: A review. *Big data*, 5(2), 2017.

[30] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *ECML PKDD*, 2013.

[31] C. Sun, A. Asudeh, H. V. Jagadish, B. Howe, and J. Stoyanovich. Mithralabel: Flexible dataset nutritional labels for responsible data science. In *CIKM*. ACM, 2019.

[32] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *VLDB*, 1994.

[33] V. V. Vazirani. *Approximation algorithms*. Springer Science & Business Media, 2013.

Letter from the Impact Award Winner

I was very happy and humbled to receive this year's TCDE Impact Award, with the citation "for contributions to spatial, temporal, and spatio-temporal data management." I would like to thank those who nominated me as well as the awards'd committee. Conducting research is very much a social, or collaborative, activity, and I have worked with many excellent colleagues on the three topics mentioned in the citation, and they deserve most of the credit for the results that I have contributed to achieving. I will mention some of them as I cover aspects of my research journey. I started out working on temporal databases and then later transitioned to working on spatial and spatio-temporal databases. To achieve some degree of brevity, I will offer an account of only some of the activities related to temporal data management. I thus start at the very beginning of my academic life.

The Early Years—Ph.D. Studies I received my M.Sc. degree in computer science from Aalborg University in 1988. At that time, the M.Sc. study had a formal duration of five and a half years and included two B.Sc. degrees (in my case, in Mathematics and Computer Science). The last half year was devoted to the M.Sc. thesis, but the mindset at the time was that you were not serious if you spent less than a year. Thus, having received the M.Sc. degree after six years of study, I received a scholarship to go and study for a Ph.D. for two and a half years anywhere in the world. All I needed to do was to write a thesis—the course requirements were already satisfied.

In early September 1988, I then arrived at Dulles Airport. My M.Sc. supervisor, Lars Mathiassen, now a professor at Georgia State University, had recommended that I study under the direction of Leo Mark, then a young faculty member at the University of Maryland. I still remember driving with Leo from Dulles to his house in the late evening with all the windows open in his (by Danish standards) huge and very American Chevy. An exciting journey had started.

A November 25, 1988 plan gave the following working title for my thesis: "A By-Relation Implemented Object Oriented Data Model Supporting Efficient Storage and Retrieval of Versions of Complex Objects in Engineering Applications." I started out looking at the versioning aspect, and this led to studies of support for transaction time, which I viewed as an ideal foundation for fine-grained version support. The eventual title of the thesis was "Towards the Realization of Transaction Time Database Systems," and I had become interested in temporal databases.

The Pursuit of Industrial Impact Having completed the Ph.D. studies and defended the thesis back in Denmark in January 1991, I packed up my car in Greenbelt, MD and drove cross-country to Tucson, AZ, where I was to work with the most visible temporal database researcher, Rick Snodgrass, then a young faculty member at the University of Arizona. I had received a faculty position at Aalborg University that allowed me to spend my first semester with Rick. Our interests matched very well, and we got off to a very good start. This turned into three more sabbaticals, in 1992, 1994, and 1999, where I also got the opportunity to work with Rick's students, Curtis Dyreson, Nick Kline, and Mike Soo.

The 1990s were exciting times in temporal databases. The field had witnessed a proliferation of temporal data models and query languages, almost to the point of each researcher having their own model and language. It was felt that this blocked industrial impact, and initiatives were taken to achieve a consensus temporal data model and query language. This resulted in the TSQL2 query language, which was designed by an 18-person committee led by Rick.

Pursuing the goal of achieving industrial impact, Rick subsequently was the main force behind attempts to standardize TSQL2. This turned out to be a difficult process, in part due to politics and a variety of interests, but we also made technical progress. Specifically, we learned that the TSQL2 design approach did not scale well: Adding support for some temporal functionality to SQL worked fine, but adding comprehensive support following the TSQL2 approach was not pretty. While SQL is not a pretty language in the first place in terms of design, the TSQL2 approach yielded a result that was uglier than we would have liked. Something different was

needed. As we were making these revelations, Michael Böhlen joined the University of Arizona as a postdoc. He had worked on an approach to language design that inspired the introduction of so-called statement modifiers into TSLQ2. The idea is that many temporal queries can be expressed intuitively and unambiguously as a single-state, non-temporal (and easy-to-formulate) SQL query that is then performed, as specified by a statement modifier, on all states of a temporal relation, after which the results are combined into a temporal relation. So a temporal query could then be formulated by a non-temporal query prefixed by some modifiers. A careful design based on this approach was introduced into standards proposals, and an "academic" version called ATSQL was also designed and documented in a TODS 2000 paper titled "Temporal Statement Modifiers."

In parallel with the above, I also worked on a range of other subjects in temporal databases, including database design, covering logical and conceptual temporal database design; data model and query language design aspects; support for the notion of "now" and for data aging; indexing; implementation of temporal algebra operators; query optimization; and architectures for implementing temporal query language support. I worked with five of my first six Ph.D. students on these topics: Kristian Torp, Heidi Gregersen, Dieter Pfoser, Janne Skyt, and Giedrius Slivinskas.

The Recent Years While spatial and spatio-temporal databases started to take over as my main activity around year 2000, I have continued to maintain an interest in temporal databases. Following his postdoc at Arizona, Mike joined the faculty at Aalborg University. He later moved to the Free University of Bozen-Bolzano and he is now back home in Switzerland, at the University of Zurich. I have been fortunate to be able to continue to work on temporal databases with Mike, Hans Gamper from Bolzano, and most recently Anton Dignös, as a Ph.D. student at Zurich and now as a faculty member at Bolzano. A key goal was to achieve an implementation of ATSQL. With other colleagues, we looked at many options, but it took until 2016, i.e., 16 years, before we had solid results. In particular, Anton's Ph.D. thesis and a TODS 2016 paper titled "Extending the Kernel of a Relational DBMS with Comprehensive Support for Sequenced Temporal Queries" show how to extend the kernel of PosgreSQL to enable efficient support for the functionality described in the TODS 2000 paper.

Impact and Lessons Looking back, one may ask what the impact of this work has been. Certainly, the literature suggests that the work has influenced other research in the field, but there has also been impact beyond academia. One highlight is that Teradata put temporal support into their system based on the statement modifier approach, which made them a pioneer in offering temporal support. This was done before ANSI/ISO standardization. Today, Teradata in addition supports the temporal tables and (limited) query language syntax in the standard. Another highlight is that the PostgreSQL implementation described in the TODS 2016 paper is available for anyone to use. A different line of impact is in the area of database design, where national statistics bureaus (e.g., Statistics Denmark) and archives (e.g., Danish National Archives) make use of temporal tables, including bi-temporal tables, when organizing their data. I have been contacted by, and have interacted with, several such entities. While the standards have adopted a language design approach that I think does not scale, and while there is a disconnect between SQL standardization and academia, I do believe that the standard is influenced by advances in temporal database research. For example, the standard supports bitemporal tables: We studied such tables in depth and even coined the term bitemporal.

Finally, I want to make a few points. First, research is often a social and collaborative effort. One should try to work with good colleagues (check!) and try to be a good colleague. Second, it can take decades to achieve societal impact, which is at odds with the increasing dependence on short externally funded projects in order to be able to perform research. Third, the disconnect between stardardization and academia is unfortunate from a societal perspective. Fourth, in research, one often does not quite know where one ends when starting.

Christian S. Jensen
Aalborg University, Denmark

Letter from the Service Award Winner

Icing on the Cake

I have had the honor and pleasure of serving for 25+ years and over 100 issues as the Editor-in-Chief (EIC) of the Data Engineering Bulletin, the very publication in which this letter is being published. I never dreamed, while pondering the Bulletin EIC offer from Rakesh Agrawal, then the TCDE chair in 1992, that I would make the Bulletin so significant a part of my career. To now get rewarded with the TCDE Service Award is truly "icing on the cake". I am thankful to the TCDE both for the opportunity to serve as Bulletin EIC and now for being honored for this service with this award.

The Bulletin has been such a large part of my technical career and my primary service activity until just recently, when I have become involved with Computer Society governance. And the beauty of how this all worked out is that the Bulletin has truly been a "labor of love". Where else can database professionals learn what is happening in a subarea of our field, brought together in a single issue, with contributions from research and industrial leaders.

In the database area, which changes so fast, the ability of the Bulletin to provide a special issue on a new topic is both unique and invaluable. The ability of Bulletin editors to bring leading technologists together to write articles for an issue is the "magic sauce" that makes the entire enterprise a success. Over the years, it has been my pleasure to work with so many of the gifted editors whose work you see in every issue published. I like to think that I also contributed to the success of the Bulletin– but my success was one level indirect. It was my success over the years of convincing distinguished members of the database community to serve as Bulletin editors. As one mark of this success, the editors I have appointed include seven Codd Award winners, all but one prior to their receiving the award. And I have no doubt there will be more winners in the future.

The Bulletin would not exist without articles written by so many distingushed members of our database community. Their willingness to contribute articles is a direct result of you, our readers, who so eagerly consume Bulletin articles. The result of this is a virtuous cycle: distinguished editors attract distinguished authors, who write articles that are read and cited by many members of our database community. So you, dear reader, have played an essential role in making this system work.

Over the years, the Bulletin has transformed from solely paper publication to a mixed paper-electronic publication to finally an entirely electronic publication. Over that time, my job at Digital Equipment Corp. (DEC) transformed into a job at Microsoft. My thanks to both employers, who so generously permitted me to spend time on the Bulletin for so many years, and who provided the initial web infrastructure that made the Bulletin available electronically.

Haixun Wang, my successor and current Bulletin EIC, now has three issues "under his belt". So the future of the Bulletin looks very promising. He has recently introduced an "opinion" section, and asked me to contribute an opinion piece in the first issue with the new section. This was my first non-letter Bulletin publication since 1987 (before I became EIC). I am hoping it is not the last as, like so many others in our community, I value the Bulletin as a channel for publishing my technical contributions.

And now, finally, I too have the pleasure of reading Bulletin articles– focusing on their technical content, rather than being concerned (and consumed) by formatting and editorial issues. I have already begun enjoying this post-EIC role, and look forward to this continuing. Thank you all for contributing to the success of the Bulletin and for making my involvement so personally gratifying.

David Lomet
Microsoft Research, USA

Letter from the Rising Star Award Winner

I am honored to have received the 2019 IEEE TCDE Early Career Award "for contributions to main-memory indexing and database architectures for NVM". Let me use the opportunity of this letter to describe three open, interrelated problems in this area that I consider both interesting and important.

Is the current dominance of LSM trees over B-tree justified?

For decades, virtually all database systems relied on B-trees for indexing (with hashing being a distant second). Most modern NoSQL, NewSQL, and cloud database systems, in contrast, primarily rely on Log-Structured Merge-trees (LSM) as their main data structure. B-trees and LSMs differ in terms of many different dimensions: in-place vs. out-of-place writes, eager writes vs. background merges, favoring reads vs. writes, etc. I therefore wonder: Have B-trees become obsolete? Are LSMs just a fad? Is it possible to design a data structure that combines the best properties of the two approaches?

Do we need a new class of database systems for flash arrays?

In the past 7 years, main memory capacities have stagnated. The first commercially-available version of byte-addressable non-volatile memory ("Intel Optane DC Persistent Memory") turned out to be as expensive as DRAM, but significantly slower. Flash, on the other hand, has become much cheaper during this time frame and is now $20\times$ cheaper than DRAM per byte. Furthermore, flash has become much faster, and it is now possible to directly attach a dozen or more devices to a single server, which results in a theoretical aggregated bandwidth close to DRAM. Neither traditional disk-based, nor modern in-memory or NVM-based database systems are capable of exploiting such extremely fast flash devices. This raises the question of whether a new system design is needed and how it would differ from existing approaches.

How to exploit hardware fluidity in the cloud?

When developing high-performance database systems, most of us implicitly assume that the hardware is fixed and optimize for a particular configuration. Given how most organizations procure hardware, this a reasonable approach. In the cloud, however, because it is easy to migrate to a different instance with potentially very different underlying properties, hardware should not be thought of as fixed. After all, users care about performance and cost, not about which kind of instances their service runs on. Therefore, cloud-native database systems could autonomously optimize the hardware configuration they run on. This requires an economical, literally cost-based approach that takes actual market prices into account.

Viktor Leis
Friedrich-Schiller-Universität Jena

It's FREE to join!

TCDE

tab.computer.org/tcde/

The Technical Committee on Data Engineering (TCDE) of the IEEE Computer Society is concerned with the role of data in the design, development, management and utilization of information systems.

- Data Management Systems and Modern Hardware/Software Platforms
- Data Models, Data Integration, Semantics and Data Quality
- Spatial, Temporal, Graph, Scientific, Statistical and Multimedia Databases
- Data Mining, Data Warehousing, and OLAP
- Big Data, Streams and Clouds
- Information Management, Distribution, Mobility, and the WWW
- Data Security, Privacy and Trust
- Performance, Experiments, and Analysis of Data Systems

The TCDE sponsors the International Conference on Data Engineering (ICDE). It publishes a quarterly newsletter, the Data Engineering Bulletin. If you are a member of the IEEE Computer Society, you may join the TCDE and receive copies of the Data Engineering Bulletin without cost. There are approximately 1000 members of the TCDE.

Join TCDE via Online or Fax

ONLINE: Follow the instructions on this page:

www.computer.org/portal/web/tandc/joinatc

FAX: Complete your details and fax this form to **+61-7-3365 3248**

Name _____

IEEE Member # _____

Mailing Address _____

_____

Country _____

Email _____

Phone _____

TCDE Mailing List

TCDE will occasionally email announcements, and other opportunities available for members. This mailing list will be used only for this purpose.

Membership Questions?

Xiaoyong Du
Key Laboratory of Data Engineering and Knowledge Engineering
Renmin University of China
Beijing 100872, China
duyong@ruc.edu.cn

TCDE Chair

Xiaofang Zhou
School of Information Technology and Electrical Engineering
The University of Queensland
Brisbane, QLD 4072, Australia
zxf@uq.edu.au

IEEE Computer Society
10662 Los Vaqueros Circle
Los Alamitos, CA 90720-1314

Non-profit Org.
U.S. Postage
PAID
Los Alamitos, CA
Permit 1398