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 TCDE Awards Committee Chair

Johannes Gehrke
Microsoft, USA

Letter from the Special Issue Editor

Letter from Alexandra Meliou

Alexandra Meliou
University of Massachusetts, Amherst

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.

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.

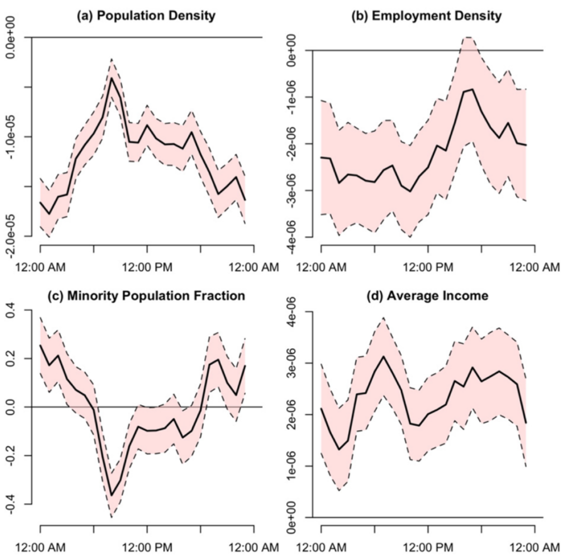


Figure 1: Coefficient estimates and 95% confidence interval of spatial error model for (a) population density, (b) employment density, (c)minority population fraction, and (d) average income [50].

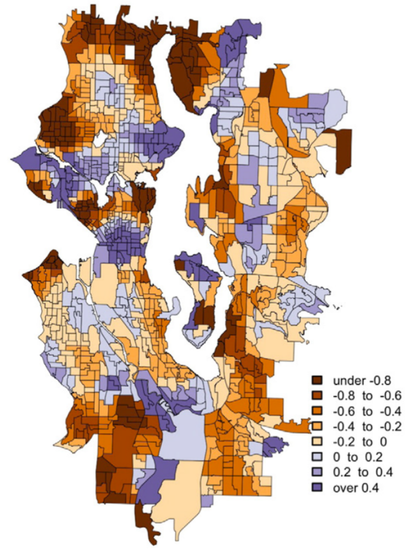


Figure 2: Coefficients for minority fraction from geographically weighted regression. Purple indicates a positive association between expected waiting time and minority fraction; gold indicates a negative association [50].

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.

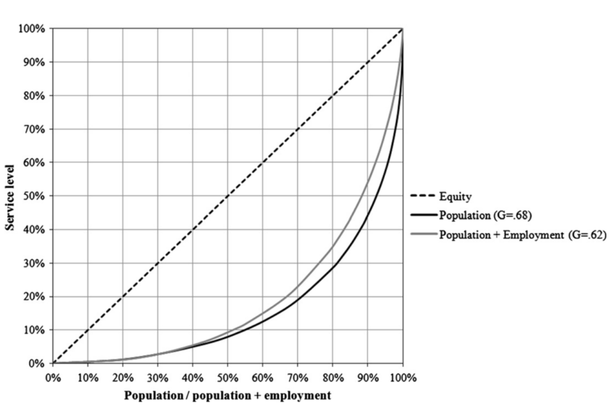


Figure 3: Use Lorenz curves to compare the equity of public transport service level to demand (population and employment) [28].

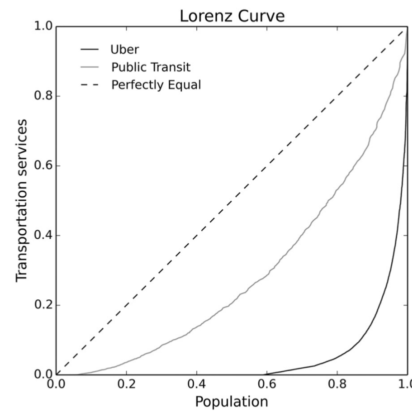


Figure 4: Use Lorenz curves to compare the equity of public transport and Uber service level to population [54].

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

Borough	Weekday		Weekend	
	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 1: 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 1). The results (see Table 1) 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:

$$Ratio = \frac{Number\ of\ minorities / Total\ number\ of\ minorities}{Number\ of\ non-minorities / Total\ number\ of\ non-minorities} \tag{1}$$

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, I 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.

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]:

$$\Pr(\mathbf{V}) = \prod_{X \in \mathbf{V}} \Pr(X|\mathbf{Pa}(X)) \tag{2}$$

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 $\mathtt{do}$ 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 $\mathbf{x} \in Dom(\mathbf{X})$:

$$P(O_{S\leftarrow 0}(\mathbf{U}) = 1|\mathbf{X} = \mathbf{x}, S = 1) = P(O_{S\leftarrow 1}(\mathbf{U}) = 1|\mathbf{X} = \mathbf{x}; S = 1) \tag{3}$$

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:

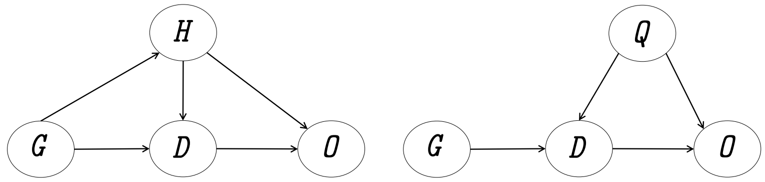
$$P(O = 1|do(\mathbf{P} = \mathbf{p})) = P(O = 1|do(\mathbf{P} = \mathbf{p}')) \tag{4}$$

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 4 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:

$$\Pr(O = o|do(S = 0), do(\mathbf{K} = \mathbf{k})) = \Pr(O = o|do(S = 1), do(\mathbf{K} = \mathbf{k})) \tag{5}$$



(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 $\Pr(A = 1|G = 1, D = \mathrm{A}) = 0.8$ but $\Pr(A = 1|G = 0, D = \mathrm{A}) = 0.2$). 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:

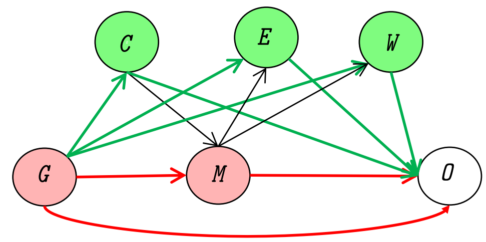
Attribute	*Res.*
MaritalStatus	*0.58*
Education	*0.13*
HoursPerwWeek	*0.04*
Age	*0.04*

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

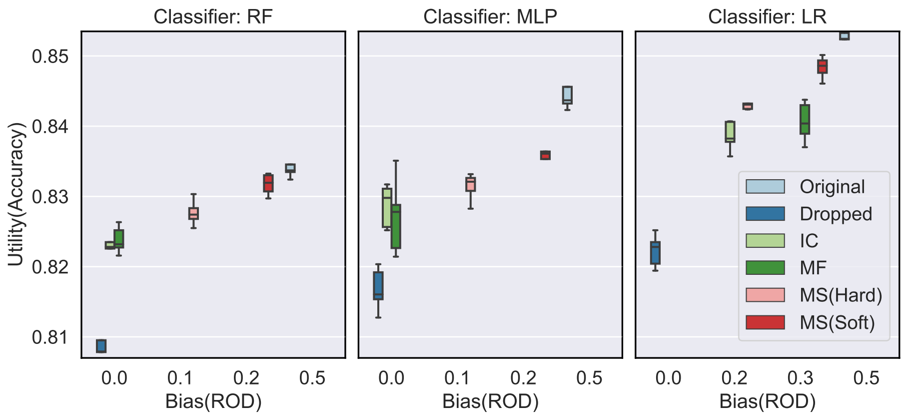


Figure 4: Performance of CAPUCHIN on Adult data.

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 Information-ssysteme"(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.

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

Data Engineering

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