The price of choice:
models, paradoxes, and inference for 'mobility as a service'

Damon Wischik[1]

Abstract— **A city's transportation network is made up of subsystems, often under separate management, linked together through the choices made by users. This paper introduces a transport model which combines a discrete choice model of users, with a resource allocation model of a subsystems. This combined model gives a direct economic interpretation of tradeoffs in the system. For example, it tells us how much of a rideshare price is attributable to the cost of running the platform and how much is profit-making. The model can also be used to predict knock-on effects in the style of Braess's paradox, where an improvement in one part of the network might induce problems in other parts because of selfish choices made by users and by subsystems.**

I. INTRODUCTION

Selfish routing is a classical mathematical model of how self-interested users might route traffic through a congested network. It was introduced by Wardrop [1] as a model for assigning road traffic, for the purpose of computing the cost-effectiveness of road improvement schemes. Wardrop's model is still in widespread use in highway agencies for exactly this purpose, implemented directly in software packages such as SATURN [2] and PTV Visum [3], and indirectly in every agent-based traffic simulator that incorporates user choice.

Urban planning today is facing the challenges of 'Mobility as a Service' (MaaS). This term refers loosely to a collection of new technologies such as autonomous vehicles and electric scooters; to new platforms such as Uber and Zipcar and CitiBike; to the shift away from personally-owned cars; and to integrated multi-modal ticketing including apps such as Whim. Urban transport planning has typically been organized into a highways agency, a bus agency, etc., but now there is a need to reorient away from per-mode management and towards "How well can people get where they want to, over whatever mode they choose?", bearing in mind also environmental and health targets.

To model the MaaS world, we can add an extra class of decision maker to Wardrop's formulation. He described two classes: individual drivers within the model who choose their routes selfishly; and a network planner outwith the model who wants the best overall average benefit across all drivers. We propose three classes: individual users within the model who choose their routes/modes selfishly; a subsystem operator within the model who manages a mode or platform selfishly; and a network planner outwith the model as before. Some subsystems are privately run companies, such as Uber, in which case it is inherently reasonable to treat them as maximizing their own profits. Some subsystems are notionally under control of a single authority, such as London streets and London buses, but are typically run as separate agencies with their own targets and procedures.

The key contribution of this paper is to reformulate discrete choice modelling so as to integrate it with models of optimal resource allocation. In descriptive statistics, a standard model for user choice is logistic regression. In econometrics, a standard model is that each choice has an associated utility that is random, and users pick the choice with the highest utility; for a well-chosen random distribution this is equivalent to logistic regression. This paper introduces a novel model in which users choose a level of resource consumption in order to maximize their utility.

The combined model of discrete route choice and constrained resource allocation can then be used to analyse MaaS problems. For example, in a rideshare model, it explains how the rideshare operator's objective "find the cheapest flow of empty vehicles that will rebalance the network" is externalized as prices charged to users, affecting their mode choice. Another example: it can help a network planner to anticipate the effects of interventions. For example, if I close a lane here, what will the knock-on effects be throughout the city, taking into account the choices made by individual drivers and also the induced changes in the rideshare network?

Outline. Section III gives an example of a Braess-style paradox, which illustrates the problem of unintended knock-on effects. Section VII shows how our combined discrete choice + resource allocation model can give answers.

Sections IV–VI build up the model. We first see the core technique for converting an individual user's choice into a utility optimization. We next apply the technique to the problem of mode choice in rideshare, and then to the problem of estimating demand given aggregated data.

The goal of MaaS modelling at the policy level is to design mechanisms in the form of information flows and incentives: for example, how can I design a tax system so that rideshare operators are incentivised to route around pollution hotspots? Sections VII and VIII conclude by discussing the tools we need to build, to help cities manage rideshare platforms and other MaaS scenarios.

A note on terminology: this paper uses the term 'rideshare' broadly to refer to e-hail platforms like Uber and Lyft,

*This work is supported by a grant from the Toyota Mobility Foundation. It has benefitted from discussions with Sid Banerjee of Cornell, and Neil Walton of Manchester University.

[1]Damon Wischik is with the Computer Science and Technology department at the University of Cambridge, and with the Alan Turing Institute in London

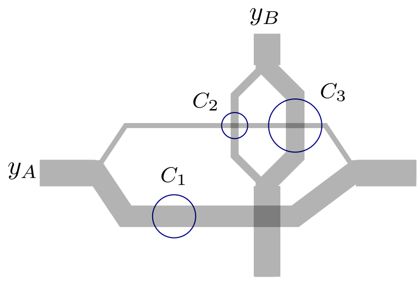


Fig. 1: *A multipath resource allocation problem, with two multipath flows that use three resources. The flow (y_A, y_B) is admissible if and only if $y_A \leq C_1 + \min(C_2, C_3)$, $y_B \leq C_2 + C_3$, and $2y_A + y_B \leq 2C_1 + C_2 + C_3$. How should flow rates be allocated subject to these constraints?*

whereas some of the literature uses it only for car pooling.

II. RELATED WORK

Wardrop's equilibrium assignment model has led to many strands of work. For this paper, the two most relevant are discrete choice modelling and network resource allocation.

Ever since McFadden's work on discrete choice and random utility modelling, including an analysis of whether Bay Area residents would use BART, discrete choice models have been widespread in the transportation literature and planning industry. Most studies that involves passenger choice use this framework. For recent examples, see [4], [5]. This paper starts with random utility modelling, and reformulates it as a resource allocation problem.

Multipath resource allocation have attracted much study in communication networks. Generalized cut constraints as illustrated in Figure 1 were studied in [7]. A distributed model for rate allocation was introduced in [8], which sets up an overall system optimization problem "maximize the utility of all flows minus a congestion loss function at each resource", and describe a distributed algorithm for solving it. Algorithms for distributed resource allocation based on [8] are now used for multipath Internet congestion control [10]. Generalized cut constraints describe what rebalancing of flows is permissible, and an analysis of the system optimization problem can show what knock-on rebalancing of flows will actually occur if, for example, one of the capacities is altered [9]. This present paper also describes a system optimization problem for resource allocation, but based on a discrete choice model for utility rather than a utility model for rate control.

Recent work proposes that socially optimal flows in a multimodal transport network might be achieved simply by building an app that suggests socially appropriate routes, rather than accommodating free user choice [6].

There is a growing literature on rideshare and carpool modelling, e.g. [11], generally emphasizing the matching mechanism between passengers and drivers, and how a rideshare operator might design this optimally. This paper differs in that it distinguishes between the network planner and the rideshare operator, treating them both simultaneously as decision makers, and presenting a joint optimization model that combines passenger choice with rideshare optimization.

In transport modelling, planners typically use the "four step model". The first two steps are forecasting demand, the third step is forecasting mode choice, and the fourth step is assigning it to routes using Wardrop's equilibrium model. There is a powerful and flexible model which combines the first three steps, based on statistical physics [12]. In this model, we consider a population of users who are assigned randomly to origins and destinations; then we look for the most likely configuration conditional on observed data. This corresponds to maximizing the entropy of the user distribution subject to constraints. For example, let T_{ij} be the number going from i to j; and suppose the constraints are (i) the total demand from node i is O_i, (ii) the total demand for node j is D_j, (iii) the total travel budget $\sum_{ij} T_{ij} c_{ij}$ is capped. Then the entropy-maximizing distribution is $T_{ij} = \alpha_i \beta_j O_i D_j e^{-\beta c_{ij}}$. The model in this paper yields similar demand matrices, but arising from a Wardrop-style discrete choice problem combined with resource allocation, rather than from statistical physics. This means that (i) it integrates the fourth step of the four-step model, and (ii) it lends itself directly to pricing calculations based on the tradeoff between transport supply and demand.

III. A BRAESS-STYLE PARADOX

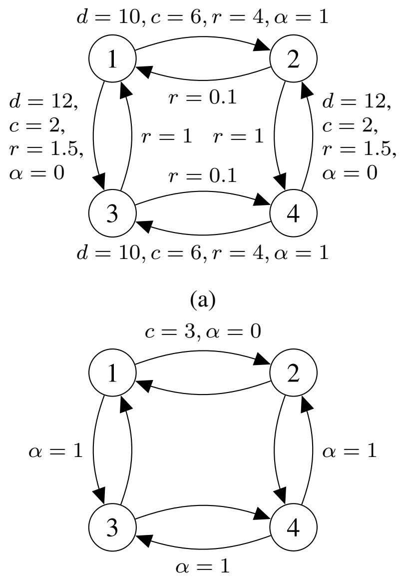
Consider the following model for a transport system with bus and rideshare, shown in Figure 2. There is a fixed demand $d_{ij} \geq 0$ on the route from node i to j, which can be served by either bus or rideshare. The edges in the graph denote complete routes from origin to destination, so there is no need to consider multihop paths. The cost of a bus ride is $c_{ij} > 0$, and the cost of rideshare is $\mu_i r_{ij}$ where μ_i is the surge multiplier at i and $r_{ij} > 0$ is the underlying cost for driving a vehicle between those nodes. Let $\alpha_{ij} d_{ij}$ be the volume of users who choose rideshare from i to j, leaving $(1 - \alpha_{ij}) d_{ij}$ to take the bus.

User model. Users make their decision based only on price: $\alpha_{ij} = 1$ if $\mu_i r_{ij} < c_{ij}$, and $\alpha_{ij} = 0$ if $\mu_i r_{ij} > c_{ij}$. In the case $\mu_i r_{ij} = c_{ij}$, users are indifferent, and we consider α_{ij} to be under the control of the rideshare operator—it might exert this control by jittering its price just under or just over c_{ij}.

Subsystem model. The rideshare operator sets the surge prices μ_s. In addition to the vehicles carrying passengers, it rebalances its fleet by running empty vehicles from i to j at rate $\beta_{ij} \geq 0$. The total flow must form a circulation, i.e.

$$\sum_i \Big(\alpha_{ij} d_{ij} + \beta_{ij} \Big) = \sum_k \Big(\alpha_{jk} d_{jk} + \beta_{jk} \Big) \quad \forall j.$$

Total revenue is $\sum_{i,j} \alpha_{ij} \mu_i r_{ij}$, and total cost is $\sum_{i,j} (\alpha_{ij} + \beta_{ij}) r_{ij}$. The rideshare operator's objective is to maximize revenue minus cost, by choosing λ and β (and α when users are indifferent).



(b)

Fig. 2: *A transport network with bus costs £c, rideshare base costs £r, and total demand d, of which αd takes rideshare. The bus cost c_{12} is reduced from £6 in (a) to £3 in (b), and this single change has a knock-on effect on α throughout the network.*

It is straightforward to compute the user and subsystem optimum using brute force. The rideshare operator will set surge prices so that at each node i there is some node j^* such that $\mu_i r_{ij^*} = c_{ij^*}$; otherwise it could increase profit by increasing μ_i. So the model can be solved by enumerating all possible values for each surge multiplier, solving a linear program to compute α and β, and picking the one that yields the highest profit for the rideshare operator.

A paradox. Consider the four-node network shown in Figure 2(a). The optimal route assignment and surge multipliers are as shown. Suppose that the bus fare c_{12} is reduced from £6 to £3: the new solution is shown in Figure 2(b). One might expect that reducing bus fares should have the effect of increasing bus passenger numbers, but the net effect is in fact the opposite: total bus passenger numbers are *reduced* from 24 to 10. Rideshare profits are also reduced from £38 to £21.93.

Braess described a paradox of selfish routing, a simple road network in which *adding* a new road has the effect of *worsening* average travel times, even though total demand is unchanged. His paradox is compelling because of sign reversal. A healthy modelling skepticism should lead to doubts about quantifying the price of anarchy ("maybe the benefit of building a road will be a bit less than predicted because of selfish routing; but all models are wrong, and at least we're doing some good"). But by demonstrating a sign reversal, Braess's example shows that even well-intentioned interventions can have detrimental outcomes.

In our network in Figure 2 likewise, intervening to reduce bus fares has the opposite effect to what one one would see when considering the bus network in isolation. The issue is that a small change in one part of the bus network has an impact on the rideshare subsystem, and the rideshare subsystem has a control mechanism (fleet rebalancing β) that transmits this impact to the rest of the network. Braess's paradox showed us that selfish routing by individual users can lead to sign reversal, so it's no surprise that selfish control by an agent with many more degrees of freedom can do the same.

IV. DISCRETE CHOICE MODEL

We now introduce a building block: a discrete choice model for a single user, which we'll use in the next two sections for building full system models.

Consider a user with a choice between two alternatives, route 1 and route 2, each with its own cost. A standard descriptive model is the logistic regression

$$\mathbb{P}(\text{choose } 1) = \frac{e^{\kappa - \gamma_1 p_1 + \gamma_2 p_2}}{1 + e^{\kappa - \gamma_1 p_1 + \gamma_2 p_2}} \tag{1}$$

where p_1 and p_2 are the prices on each of the routes and κ, γ_1 and γ_2 are parameters that can be fit from data. The exponent could include a variety of factors, but for present purposes we will be focusing on prices in a rideshare network, so we treat all other factors as lumped together into κ.

Another standard formulation is the random utility model. Let the utility of route i be a random variable,

$$U_i \sim \kappa_i - \gamma_i p_i + E_i, \quad E_i \sim \mathrm{Gumbel}() \tag{2}$$

where $\mathrm{Gumbel}()$ refers to a Gumbel distribution with cumulative distribution function $F(x) = \exp(-e^{-x})$, and suppose that the user picks the route with the higher utility. This leads again to (1), with $\kappa = \kappa_1 - \kappa_2$.

Now consider a user who has utilities given by (2), but who is coerced into making a different choice: suppose the user picks route $\arg\max_i (U_i + \theta_i)$, where θ_i are constants describing the amount of coercion. Only the difference $\theta_1 - \theta_2$ actually matters. Let $\alpha = \alpha(\theta_1 - \theta_2)$ be the coerced probability of choosing route 1. The expected utility will also depend on the amount of coercion; it is

$$\mathbb{E}\big(U_1 1_{U_1+\theta_1 > U_2+\theta_2} + U_2 1_{U_1+\theta_1 < U_2+\theta_2}\big).$$

After some algebra, and rewriting in terms of α, we find that the expected utility is

$$\alpha(\kappa_1 - \gamma_1 p_1) + (1-\alpha)(\kappa_2 - \gamma_2 p_2) + H(\alpha, 1-\alpha)$$

where H is the entropy of a $\mathrm{Bernoulli}(\alpha)$ random variable. Call this $u(\alpha \,|\, p_1, p_2)$. The user's choice can thus be written as an optimization problem:

$$\text{maximize} \quad u(\alpha \,|\, p_1, p_2) \quad \text{over } \alpha \in [0, 1]$$

and the solution coincides exactly with (1).

Why think of coercion in this way, as tweaking the thresholds for comparing random utilities? A simpler brute-force coercion is to simply say "The system dictates that the user take route 1 with probability α". This leads to an expected utility of

$$\alpha(\kappa_1 - \gamma_1 p_1) + (1-\alpha)(\kappa_2 - \gamma_2 p_2). \tag{3}$$

However, this type of coercion is not consonant with the notion of user choice expressed by (1) nor by (2)—there is no obvious way to make (1) arise through maximizing (3).

V. RIDESHARE PRICING

We now consider a joint model of user choice and resource allocation, a relaxed version of the model in Section III. Let demand d, public transit fare c, and rideshare base cost r be as before.

User model. Users make their decision based only on price: the fraction of users taking rideshare from i to j is

$$\alpha_{ij}(p_{ij}) = \frac{e^{\bar{\kappa}_{ij} - \gamma p_{ij} + \gamma' c_{ij}}}{1 + e^{\bar{\kappa}_{ij} - \gamma p_{ij} + \gamma' c_{ij}}}. \tag{4}$$

Subsystem model. The rideshare operator chooses rebalancing rates β. The total cost of all vehicle movements is

$$R(\alpha, \beta) = \sum_{i,j} (d_{ij}\alpha_{ij} + \beta_{ij}) r_{ij}$$

and the rebalancing rates solve

$$\begin{aligned} &\text{minimize } R(\alpha,\beta) \text{ over } \beta \geq 0 \\ &\text{such that } d\alpha + \beta \text{ is a circulation.} \end{aligned}$$

In addition, the rideshare operator sets a price p_{ij} on each edge, so as to maximimize the total revenue

$$\Big(\sum_{i,j} \alpha_{ij} d_{ij} p_{ij}\Big) - R(\alpha, \beta)$$

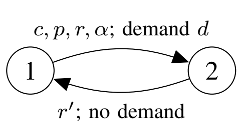
taking account of the impact of price on α and thence on β.

System optimization problem. Consider the following problem: maximize

$$\begin{aligned} \sum_{i,j} \Big\{ d_{ij} \Big[&\alpha_{ij}(\tilde{\kappa}_{ij} - \gamma m_{ij}) + (1-\alpha_{ij})(\kappa'_{ij} - \gamma' c_{ij}) \\ &+ H\big(\alpha_{ij}, 1-\alpha_{ij}\big) - \frac{\alpha_{ij}}{1-\alpha_{ij}} \Big] \\ &+ \beta_{ij}(m_{ij} - r_{ij}) \Big\} \end{aligned}$$

over $\alpha_{ij} \in [0,1]$, $\beta_{ij} \geq 0$, and $m_{ij} \in \mathbb{R}$, such that $d\alpha + \beta$ is a circulation.

It can be shown that any solution to the system problem solves the user and subsystem models, and vice versa (with $\bar{\kappa} = \kappa - \kappa'$). The proof is intricate but uninteresting and there isn't enough space here. Instead, to illuminate, consider a single link:



There must be a rebalancing flow of size $d\alpha$ from 2 to 1, so the rideshare subsystem sets p to maximize

$$d\alpha_p p - d\alpha_p (r + r').$$

In this equation we're writing α_p to emphasize that users adapt to p by choosing α. The maximum is at $\alpha_p + \alpha'_p(p - r - r') = 0$. By differentiating (4) we find $\alpha'_p = -\gamma\alpha_p(1-\alpha_p)$. Substituting this back in, the rideshare subsystem's choice of p solves

$$p = r + r' + \frac{1}{\gamma(1-\alpha_p)}. \tag{5}$$

Now, given p, we know from section IV that α maximizes

$$\alpha(\kappa - \gamma p) + (1-\alpha)(\kappa' - \gamma' c) + H(\alpha, 1-\alpha). \tag{6}$$

When both the subsystem optimization and the user optimization are jointly solved, then (5) is satisfied and (6) is maximized simultaneously, thus α maximizes

$$\begin{aligned} \alpha\big(\kappa - \gamma(r + r')\big) &+ (1-\alpha)(\kappa' - \gamma' c) \\ &+ H(\alpha, 1-\alpha) - \frac{\alpha}{1-\alpha}. \end{aligned}$$

This has the same form as the system optimization problem. The proof of the general case follows this strategy of endogenizing the prices, but starting from the Lagrangian of the system optimization problem so that the circulation constraints are accounted for via dual variables.

Interpretation. In the system optimization problem, m plays the role of an 'accounting cost'. If there's an excess of vehicles coming into i and a shortage at j then it's beneficial to send empty vehicles from i to j, and in a socialist world passengers would pay less than r_{ij} to travel on that link. This will be reflected by $m_{ij} < r_{ij}$.

The $\alpha/(1-\alpha)$ term reflects profit-taking by the rideshare operator. On links where α is close to 1, the term pushes α lower than it would otherwise be. The rideshare operator is extracting profit, and it does this by setting prices higher than they need to be, which pushes some users away. In links with α small, there is less opportunity to extract profit because public transit is appealling.

In this version of the system problem we have allowed prices to be varied freely per link. If the rideshare operator is constrained to set prices according to a surge multiplier, that constraint can simply be added to the system problem.

VI. VARIABLE DEMAND MODEL

In resource allocation problems it can be useful to treat demand as arising endogenously—as a tradeoff between users who want a resource versus the cost of that resource. The discrete choice model from Section IV does a poor job of this. It's possible to add a 'null option', call it route 0, and interpret it as "don't take either route"; but this model can only generate at most one trip per user. Informally, it's a binomial model and we'd prefer a Poisson.

Here is a version of the random utility framework that accommodates variable demand. Suppose a user can make a number of trips, and each trip can be on one of two routes. Let u_1 and u_2 be constants reflecting the underlying utility of each route, and let u_0 be a constant corresponding to not taking a trip. Suppose the user uses the following procedure for deciding how many trips to take and on which routes:

1) Generate $U_i \sim u_i + \mathrm{Gumbel}()$
2) If U_0 is the largest then stop
3) Otherwise, take a trip on route i where U_i is the larger, and gain utility $U_i - U_0$ by doing so
4) Go back to step 1

Under this procedure, the expected number of trips on route i is $e^{u_i - u_0}$.

We can turn this trip-generating procedure into a resource allocation problem, using a similar technique to Section IV. Suppose the user is coerced into making different choices by using $U_i + \theta_i$ for comparisons in steps 2 and 3, but that the utilty gain is still $U_i - U_0$. (Why is the utility gain $U_i - U_0$ in step 3 rather than just U_i? So that if the user is coerced into taking lots of trips, e.g. by making θ_1 very large, then there is an overall reduction in utility.)

Reparameterizing in terms of the total number of trips y and the fraction α_i made on route i, the expected utility is

$$y\Big[\alpha_1(u_1 - u_0) + \alpha_2(u_2 - u_0) + H(\alpha_1, \alpha_2)\Big] + (1+y)\, H\Big(\frac{1}{1+y}, \frac{y}{1+y}\Big). \quad (7)$$

We make no claim that this utility function is intrinsically true. It's simply a well-behaved function which is consonant with the fixed-demand model from Section IV. For example, the probabilty that a trip is on route 1 given that there is a trip is $e^{u_1}/(e^{u_1} + e^{u_2})$, which fits with the logistic regression model (1).

Estimating demand from data. Suppose we have measurements of total traffic on certain links in a transport network. A natural way to estimate the origin-destination matrix is to maximize a net utility function made up of terms like (7), subject to the constraint implied by the observed data.

The optimization problem "maximize total user utility subject to link constraints" is very similar to the bandwidth allocation problem studied in [8] in the context of communications networks. There are simple, fast, distributed algorithms for finding the optimum bandwidth allocation, and we speculate that such algorithms can be adapted to give fast algorithms for demand estimation.

A retail demand model. An interesting choice of utility is $u_i = \log W_i - 2 \log d_i$, where W_i is the 'attractiveness' of destination i, for example the floorspace in a shopping center, and d_i is the distance to it. If $u_i \ll u_0$ for each i then the utility-maximizing allocation is

$$\alpha_i y \approx \kappa \frac{W_i}{d_i^2} \quad (8)$$

which is a standard demand function in transport modelling [12].

Suppose now that users are constrained by a limited capacity transportation network. We can solve for the resulting allocation subject to the capacity constraint, using the demand estimation method described above. If the network were uncongested then the solution would still be (8). If some roads are congested then some demands will be lower than (8), and the dual variables for the congested roads will measure the economic opportunity cost of congestion—the amount of spending that is foregone because users can't get to the shops.

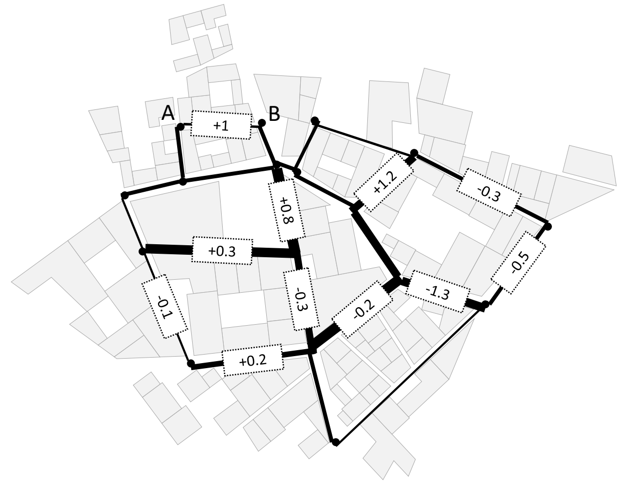


Fig. 3: *A toy network with bus and rideshare. Line thickness indicates number of bus passengers using each street. When fares are lowered on $A \to B$, the effect is transmitted throughout the rest of the network via the rideshare subsystem, and bus passenger numbers can go up or down.*

VII. Seeing Knock-on Effects

The tactical goal of MaaS modeling is to anticipate the effects of interventions, bearing in mind that changes in one part of the network have a knock-on effect via user choice and via subsystem optimization.

Consider a transport system with bus and rideshare as in Section V. We saw that the equilibrium outcome is the solution to an optimization problem. We can use this to compute knock-on effects as follows. Suppose for example we want to examine the effect of changing the public transport cost c_{AB} on some route $A \to B$.

1) Write out the Lagrangian.
2) Take total derivatives of all the variables and dual variables with respect to c_{AB}.
3) Solve the resulting system of equations.

This approach has been described before in the context of communications networks [9]. Figure 3 illustrates how the output might be shown to city planners.

All models are wrong. What's the point of this Lagrangian-based sensitivity analysis, when one could just run simulations to test the impact of changes? The reason is parameter-fitting. We posed the question "What's the impact of changes *to the current state of affairs*?", and to test this in simulation we'd need to tune the simulation parameters so as to reproduce the current state of affairs. This is intractable, for any simulator sufficiently detailed to capture the richness of what can be seen in big city datasets. It's daft if not futile to spend effort reproducing what's already there in the data—yet this is what a simulation-based approach requires.

The Lagrangian-based sensitivity analysis works differently. It starts with the data as is, and then it estimates *deltas* to the present state of affairs using a model. It's impossible to go beyond the data and answer counterfactual questions without a model, so we certainly need either a simulator or a mathematical model. Lagrangian-based sensitivity analysis is the best of both data and modelling worlds.

For helping with policy design, we should provide tools that *augment* a dataset. All models are wrong, so any final answer from a mathematical model which says "this policy is better than that" is not to be trusted. But if the tool takes in a dataset as rich as we can give it, and gives detailed output, then the policy maker can explore in fine detail the likely consequences of possible actions, paying closer attention to places that are known trouble spots, disregarding places where the data or model is sketchy.

A challenging direction for future research is to find ways in which a simulator can be used in lieu of a mathematical optimization model. Different cities face different problems and have different control levers; and there are more programmers who can implement a simulator than mathematicians who can formulate an appropriate optimization problem. Is it possible to use a simulator for a similar calculation to the Lagrangian sensitivity analysis—to estimate deltas, not values? There is a similar thrust at the moment in Probabilistic Programming Languages, which allow a programmer to specify a model by programming a simulator, and then use a general-purpose software tool to solve Bayesian inference problems.

VIII. INFERENCE CHALLENGES

The following problem was suggested to us by operators for Transport for London. The challenge is not to model it—it is embarrassingly easy to invent a model—but rather to find evidence from scant data of whether or not the problem actually occurs.

Mobility problem. Some streets in the city center become congested. This is detected in real time in the central control room (e.g. from closed-loop detectors at intersections that monitor vehicles passing, or from traffic cameras, or from bluetooth sensors that measure transit times). Highways agency operators act to relieve congestion, by changing the signal timing at upstream intersections. This causes delays and some diversions, including to buses. Passengers learn over time that buses are unreliable, so they switch to personal cars. This worsens city center congestion.

Available data. The city can be expected to keep detailed records of infrastructure, e.g. traffic flow rates or the presence of queues at traffic intersections, though the measurement types vary from city to city. Data on passenger decision making is harder to come by: it's reasonable to assume that the city has an estimate of total passenger numbers per hour, bus route, origin bus stop, and destination bus stop. It's not reasonable to assume we have access to data about the journeys made by individually identifiable passengers: even when such data is collected, privacy restrictions mean it can only be used for direct line-of-business processing, rather than speculative data science investigations.

MaaS challenge. This is a problem with two agencies—the highways agency and the bus transit agency—linked together by choices made by users. The challenge is to make it quick and easy to assess possible interactions between the agencies, given data that's readily available. If such a preliminary assessment raises concerns, it might be followed by careful expensive model building including e.g. passenger surveys, then simulating remedies.

There are actually two challenges for data inference. The first is inferring what *is happening*, and the second is inferring what *might happen*. For the first challenge, a typical task is "Estimate the origin-destination demand matrix, given aggregate measurements on certain links", and we have already in Section VI how to solve it. For the second challenge, a typical task is "Estimate the γ parameters in the discrete choice model (1), in order to model how users will react when the price changes."

We propose that a pragmatic approach for the second challenge and for many related problems is to generate a synthetic population of individuals consistent with available aggregate data. In the MaaS problem with buses on congested highways, take the daily passenger counts on each route / origin / destination as marginal data, and generate a synthetic population in which each individual has a trip diary that spans days. This population can then be analysed using whatever statistical analyses are appropriate, in this case a discrete choice model to test whether 'experienced more traffic intervention episodes' is correlated with 'less likely to take bus'. The procedure should be repeated for multiple synthetic populations, to assess the robustness of the answer. If there is a widely accepted probabilistic model for user behaviour it should be used (i.e. the synthetic populations should be drawn from a conditional probability distribution); otherwise the synthetic populations might be generated adversarially, to find maximum and minimum values of the target statistic. We call this general approach *reconstruction*.

This approach, of generating synthetic populations of individuals given marginal observations, is meant as a widely applicable first step for many analyses. On one hand it seems naive—"just invent the data you don't have". On the other hand it seems like nothing more than existing practice, from the classic origin-destination inference problem (generate a population of users each of whom occupies a path in a network, given marginal data about origins and destinations) to much more recent work on passenger trip reconstruction (generate a population of train passengers, given their tap-in and tap-out data, [13]).

We believe it's useful to give a label to this general approach. We hope by doing so to draw attention to a common data science pattern in mobility modelling. Within this limited application domain, as opposed to multiple imputation in statistics in general, it might be possible to develop a suite of reconstruction methods that work together and that can be offered to city planners as building blocks within a MaaS modelling toolkit. The concrete task of efficiently reconstructing the population, given marginal datasets and target statistics, is likely to pose interesting algorithmic questions.

REFERENCES

[1] J. G. Wardrop, "Some theoretical aspects of road traffic research," *Proc. Inst. Civil Engineers*, vol. 1, no. 3, pp. 325–362, 1952.

[2] D. V. Vliet *et al.*, "SATURN," https://saturnsoftware2.co.uk.

[3] "PTV Visum," http://vision-traffic.ptvgroup.com/en-us/products/ptv-visum.

[4] L. Ding and N. Zhang, "A travel mode choice model using individual grouping based on cluster analysis," *Procedia Engineering*, vol. 137, 2016.

[5] S. Zhu and D. Levinson, "Do people use the shortest path? An empirical test of Wardrop's first principle," *PLoS ONE*, 2015.

[6] C. Samal, L. Zheng, F. Sun, L. J. Ratliff, and A. Dubey, "Towards a Socially Optimal Multi-Modal Routing Platform," *ArXiv e-prints*, 2018.

[7] C. N. Laws, "Resource pooling in queueing networks with dynamic routing," *Adv. Appl. Probability*, vol. 24, no. 3, 1992.

[8] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control for communication networks: Shadow prices, proportional fairness and stability," *The Journal of the Operational Research Society*, vol. 49, no. 3, pp. 237–252, 1998.

[9] D. Wischik, M. Handley, and C. Raiciu, "Control of multipath tcp and optimization of multipath routing in the internet," in *Network control and optimization*, ser. Lecture notes in computer science, R. N. nez Queija and J. Resing, Eds. Springer, 2009, vol. 5894.

[10] D. Wischik, C. Raiciu, A. Greenhalgh, and M. Handley, "Design, implementation and evaluation of congestion control for multipath tcp," in *Proc. NSDI*, 2011.

[11] S. Banerjee, C. Riquelme, and R. Johari, "Pricing in ride-share platforms: a queueing-theoretic approach," *SSRN*, 2015.

[12] A. Wilson, "Entropy in urban and regional modelling: Retrospect and prospect," *Geographical Analysis*, vol. 42, no. 4, pp. 364–394, 2010.

[13] M. Ibrahimi, D. J. Wischik, C. D. Pluntke, B. S. Prabhakar, and D. Merugu, "Transportation system reconstruction," U.S. Patent US20 150 177 011A1, 12 20, 2013.