Predicting traffic volumes and estimating the effects of shocks in massive transportation systems

Ricardo Silva[a,1], Soong Moon Kang[b], and Edoardo M. Airoldi[c]

[a]Department of Statistical Science and Centre for Computational Statistics and Machine Learning, University College London, London WC1E 6BT, United Kingdom; [b]Department of Management Science and Innovation, University College London, London WC1E 6BT, United Kingdom; and [c]Department of Statistics, Harvard University, Cambridge, MA 02138

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved March 20, 2015 (received for review July 8, 2014)

Public transportation systems are an essential component of major cities. The widespread use of smart cards for automated fare collection in these systems offers a unique opportunity to understand passenger behavior at a massive scale. In this study, we use network-wide data obtained from smart cards in the London transport system to predict future traffic volumes, and to estimate the effects of disruptions due to unplanned closures of stations or lines. Disruptions, or shocks, force passengers to make different decisions concerning which stations to enter or exit. We describe how these changes in passenger behavior lead to possible overcrowding and model how stations will be affected by given disruptions. This information can then be used to mitigate the effects of these shocks because transport authorities may prepare in advance alternative solutions such as additional buses near the most affected stations. We describe statistical methods that leverage the large amount of smart-card data collected under the natural state of the system, where no shocks take place, as variables that are indicative of behavior under disruptions. We find that features extracted from the natural regime data can be successfully exploited to describe different disruption regimes, and that our framework can be used as a general tool for any similar complex transportation system.

smart cities | transportation | regime change | complex systems

Well-designed transportation systems are a key element in the economic welfare of major cities. Design and planning of these systems requires a quantitative understanding of traffic patterns and relies on the ability to predict the effects of disruptions to such patterns, both planned and unplanned (1).

There is a long history of analytic and modeling approaches to the study of traffic patterns (2), for example using simulated scenarios in simple transportation systems (3), and analysis of real traffic data in complex systems, either focusing on a small samples (4) or using more aggregate data (5, 6). Here we take this approach to the next level by making use of smart-card data and incident logs to (*i*) predict traffic patterns and (*ii*) estimate the effect of unplanned disruptions on these patterns. We analyzed 70 d of smart-card transactions from the London transportation network, composed of ∼10 million unique IDs and 6 million transactions per day on average, resulting in one of the largest statistical analyses of transportation systems to date.

A related literature deals with various aspects of dynamics in complex networks and complex systems in general (7–9), using a variety of data sources, from emails (10) to the circulation of bank notes (11) to online experiments on Amazon Turk (12). More recently, a number of analyses have leveraged mobile phone data as proxies for mobility (4, 13–15).

However, smart-card technology allows us to obtain large samples of passenger location and movements without requiring noisy and potentially unreliable proxies such as mobile Global Positioning System traces (16), while also leveraging a more structured environment that imposes hard constraints on patterns of urban mobility (17). In particular, these constraints of the system allow us to identify a global model of passenger behavior under local line and station closures.

Transport for London Data

The London transportation system is composed of several connected subsystems. We focus on the Underground, Overground, and Docklands Light Rail (DLR), all of which are train services aimed at fast commuting within the Greater London area only. A map of the system is provided in Fig. S1.

Transport for London (TfL) provided us with smart-card readings covering 70 d, from February 2011 to February 2012. Smart-card readings comprise more than 80% of the total number of journeys (18). Each reading consists of a time stamp, a location code, and an event code. The location code uniquely identifies each of the 374 stations of the system that were active during the months covered by our data. The two events of our interest are generated when a passenger touches the smart-card reader at the entrance ("tap-in" event) or at the exit ("tap-out" event) of a station. Passenger IDs are anonymized and ignored in our analysis. We discarded all tap-in readings that are not matched to a tap-out, and vice-versa. Time resolution of the recorded time stamps is 1 min. Each day is composed of 1,200 min, starting at 5:00 AM until 1:00 AM of the next calendar day. Our analysis covers weekdays only. Weekdays are assumed to be exchangeable (see Fig. S2).

Overview of Analysis

We show that we can reliably predict passenger origin–destination (OD) traffic by combining around 140,000 nonparametric statistical models with hundreds of millions of smart-card data events. We also show that the same model provides features that explain behavior under a shock (or "disruption") to the system, defined as an unanticipated period during which a station or a line is (partially) closed down. The resulting model allows us to predict the outcome of disruptions and to evaluate stations by how prone to overcrowding they are given disruptions at peak time.

Significance

We propose a new approach to analyzing massive transportation systems that leverages traffic information about individual travelers. The goals of the analysis are to quantify the effects of shocks in the system, such as line and station closures, and to predict traffic volumes. We conduct an in-depth statistical analysis of the Transport for London railway traffic system. The proposed methodology is unique in the way that past disruptions are used to predict unseen scenarios, by relying on simple physical assumptions of passenger flow and a system-wide model for origin–destination movement. The method is scalable, more accurate than blackbox approaches, and generalizable to other complex transportation systems. It therefore offers important insights to inform policies on urban transportation.

Author contributions: R.S., S.M.K., and E.M.A. performed research and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1] To whom correspondence should be addressed. Email: ricardo@stats.ucl.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1412908112/-/DCSupplemental.

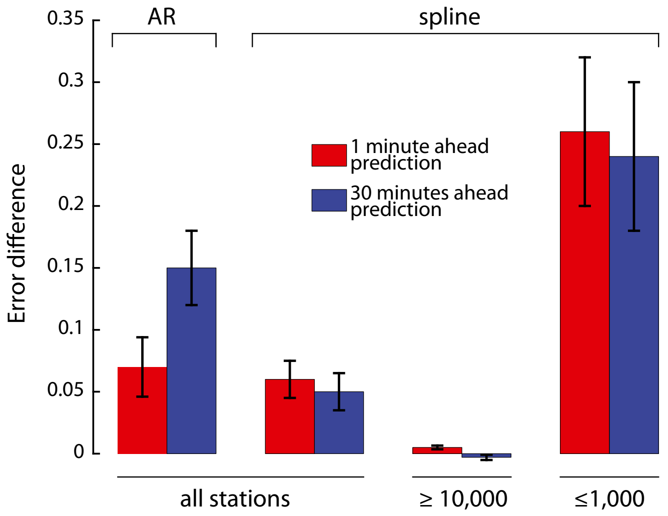

Fig. 1. RMSE difference per load between (*i*) the AR model and the tracking model and (*ii*) spline model and the tracking model. Fivefold cross-validation averages for 1-min- and 30-min-ahead predictions. Higher numbers mean an improvement given by the tracking model. Error bars show a 95% confidence interval (3 SEs).

Let N_{ijt} be the number of tap-out events at station S_j at time $t \in \{1,2,\ldots,1200\}$, caused by passengers who started their journey at station S_i (at possibly different starting times). Station S_i is the entering station, where a journey starts, and S_j is the exit station. We call N_{jt} the sum of $\{N_{ijt}\}$ over all possible entering stations, a quantity of interest for potential policies to deal with an excess number of passengers exiting through a particular destination.

Our approach can be divided into two steps. First, we develop a predictive model for N_{ijt} for all $374 \times 374 (\approx 140{,}000)$ possible pairs of stations at any minute of the day. This model represents the natural regime, where no planned or unplanned disruptions take place. Second, we create a model for N_{jt} under a disruption, knowing the type of disruption and the time period in which it occurs. Data on disruptions is provided by logs maintained by TfL, complementing the smart-card data. The model for the natural regime plays an important role here, because it is used to generate expected values of N_{ijt} according to what would have happened if no disruption had taken place. Such estimates of counterfactual variables are used as covariates (inputs) for the model for the factual outcomes, along with other structural features derived from the topology of the transportation network, where stations are vertices and edges connect stations that are directly physically linked by train tracks. A linear model provides a simple description of the relationship between topological structures, the natural regime, and the regime under disruption.

Intuitively, our disruption model is motivated by the following postulated relationship between N_{jt}^S, the number of exits from station S_j at time t under a disruption, and N_{jt}^0, the number under the natural regime:

$$N_{jt}^S = N_{jt}^0 - \mathrm{In}_{jt} + \mathrm{Out}_{jt}, \qquad [1]$$

where In_{jt} is the missing inflow, the number of passengers who cannot reach S_j because of the disruption but would have exited through S_j otherwise, and Out_{jt} is missing outflow, the number of passengers who cannot progress in their journeys in the usual way and will exit early at station S_j. Under a disruption, the variables in the right-hand side are unobservable, but their expectations can be estimated and used as covariates in a model of N_{jt}^S.

Modeling the Natural Regime: Results

We modeled $E[N_{ijt}|\mathrm{PAST}]$, the expected value of N_{ijt} given all past tap-in and tap-out events up to the given time in that particular day. This model was designed to predict three unknowns: (*i*) entering (tap-in) counts, (*ii*) the rate at which passengers remain inside the transportation system given these counts, and (*iii*) the rate at which passengers exit (tap-out) given the number of passengers inside the system and the length of their stay, according to origin. For each of these we used nonparametric regression models to account for the nonstationarity of the process over time (*Supporting Information*). We call our method the tracking model, because it keeps track of the number of passengers inside the network.

To assess the adequacy of this model, we performed a cross-validation procedure for predicting the overall aggregations $\{N_{jt}\}$ for all stations S_j. With our model, this is obtained simply by summing over the predicted N_{ijt} for each origin, for a fixed S_j. In *Supporting Information* and Figs. S3 and S4 we provide an illustration of predicting N_{jt} for the Oxford Circus station and also report a sensitivity analysis on how predictions change under different aggregations of origins and destinations.

The tracking model consists of tens of thousands of components, so there is a danger of overfitting. One way of assessing its adequacy is by comparing our predictions against blackbox models fitted directly to the aggregated data. We assessed a blackbox spline model regressing N_{jt} on the time index t. Notice that, for this model, $E[N_{jt}|\mathrm{PAST}] = E[N_{jt}]$. A second competing model is a standard linear autoregressive (AR) model, where each N_{jt} depends on $N_{j(t-30)}, N_{j(t-29)}, \ldots, N_{j(t-1)}$ (*Supporting Information*).

The cross-validation procedure is fivefold, implying 14 d (70 d/5) of test data for each fold. For the tracking model, we calculated the root mean squared error (RMSE) averaged over all stations, time points, and test days. We obtained an RMSE of 6.76 ± 0.08 tap-outs per minute for a 1-min-ahead forecast and 6.82 ± 0.09 for a 30-min-ahead forecast.

To aid the interpretability of the comparisons, we define the RMSE difference per load as the average difference between the RMSE of our model and a competitor, first calculated at a station level and then aggregated by taking a weighted average across

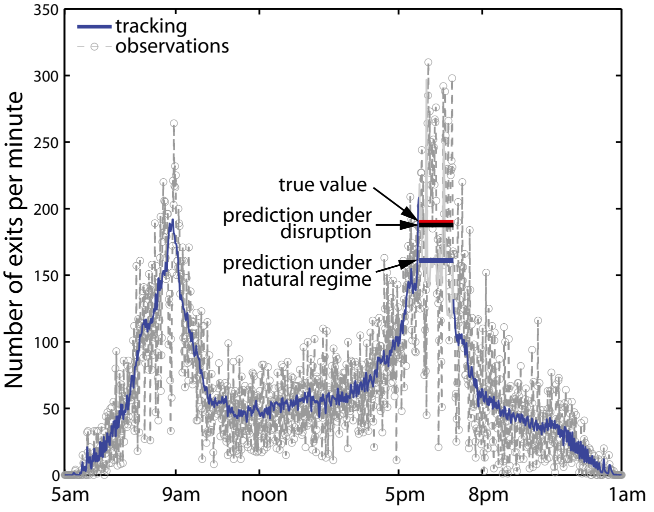

Fig. 2. Average number of exits per minute at Victoria LU station on Tuesday, January 17, 2012. The blue curve represents the 1-min-ahead prediction under the natural regime using the tracking model. Given a disruption from 6:00 PM to 7:00 PM between Victoria station and Brixton station in the Victoria line, the blue horizontal line indicates the average expected exit rate given by the tracking model under the natural regime, the red line the averaged observed exit count, and the black line the prediction given by the disruption model (Eq. **5**).

stations (weighted by the inverse of tap-out traffic volumes at that station). We discarded stations that have fewer than 10 tap-outs in the entire day.

We summarize the results of the fivefold cross-validation in Fig. 1. For instance, the RMSE per load against the AR model using all stations for a 1-min-ahead forecast is 0.07. This means that the difference of RMSEs between the AR and tracking methods has a magnitude that is ∼7% of the total traffic on average. We also assessed how predictions change when looking at subsets of the population. After discarding all stations with fewer than 10,000 exits per test day, the difference between our method and the time-independent spline method is essentially zero. For smaller stations (≤1,000 exits per test day), the difference is substantial. Thus, our model does not suffer from overfitting when compared against a blackbox model that estimates the aggregated counts directly, and it also improves the performance for the smaller stations.

Modeling the Effect of Shocks

We modeled the behavior of passengers under two types of disruption: bidirectional line segment closures and station closures. A line segment is a sequence of adjacent stations in one of the lines of the system (e.g., Piccadilly Line, see Table S1). Lines in the London system typically allow trains to go in two directions, and closures in a single direction have a weaker effect compared with closures in both directions so are of less interest when modeling larger changes. Here, stations are assumed not to close during a line segment closure, but because of the lack of trains, disrupted stations without any connection outside of the affected line segment will typically display a dramatic reduction in the number of tap-outs. During station closures trains will not stop, so passengers who planned to exit through that station will not be able to do so. Line segments are not closed during these events.

Outcome Variable. We assume that, for a given time interval $[t_1, t_F] = \{t_1, t_2, \ldots, t_F\}$ in which a disruption takes place, we have observed the behavior of the whole system up to time $t_1 - 1$. Our goal is to model the average expected tap-out count per minute, within the provided time interval, in each station of a given region of interest (ROI). A ROI is a subset of stations, selected independently of the data, in which a priori we expect to observe nontrivial changes in tap-out rate as a function of the topology of the network and type of disruption.

Although our model can predict the expected tap-out count at each minute individually, we modeled the average over $[t_1, t_F]$ because this quantity suffices to inform policy on station overcrowding and excess demand for alternative transportation. We assumed that the time interval is sufficiently short so that passenger behavior is not affected over time as a function of our covariates. As such, we define the outcome

$$\overline{N}^{S[j]}_{t_1:t_F} \equiv \sum_{t=t_1}^{t=t_F} N^{S}_{jt} \Big/ F, \qquad \textbf{[2]}$$

for each station S_j in the chosen ROI. Here N^{S}_{jt} is the number of tap-outs from station S_j at time t under disruption $\mathcal{S}$, excluding exits originated in S_j itself. Modeling this type of exit is straightforward and therefore we did not include it in the study. Fig. 2 provides an example of the prediction given by our model at Victoria Underground station.

Covariates for Line Segment Disruption. Consider the case where the disruption event $\mathcal{S}$ is the bidirectional disruption of line segment l along the sequence of stations $\mathcal{K}^l \equiv (S_{k(1)}, \ldots, S_{k(M)})$. Given this, we can define the set of covariates in the regression model for $\overline{N}^{S[j]}_{t_1:t_F}$. To distinguish between the natural regime and the regime under disruption $\mathcal{S}$, let N^0_{ijt} be the corresponding OD count at time t under the natural regime. Moreover, let $\mu^0_{ijt;t_1}$ be the expected value of N^0_{ijt} conditioned on observing all events of the day up to time $t_1 - 1$. Our set of covariates are functions of $\mu^0_{ijt;t_1}$.

Ideally, for each station $S_{k(n)} \in \mathcal{K}^l$, the disruption will be related to the amount of traffic for each OD pair (S_O, S_D) that passes either through the links $S_{k(n)} \to S_{k(n+1)}$ or $S_{k(n)} \to S_{k(n-1)}$ in the natural regime. However, only a fraction of the flow $S_O \to \cdots S_{k(n)} \to S_{k(n-1)} \to S_D$ might exit early at $S_{k(n)}$ if there are routes from the origin that do not necessarily use $S_{k(n)}$ or that might continue from $S_{k(n)}$ in a different line.

Given the target station $S_{k(n)}$, the expected missing outflow $\phi^{\mathrm{OUT}}(n)$ for $S_{k(n)}$ at time t is defined as

$$\phi^{\mathrm{OUT}}(n) \equiv \sum_{t=t_1}^{t=t_F} \phi^{\mathrm{OUT}}_{u}(n,t) \Big/ F, \qquad \textbf{[3]}$$

where

$$\phi^{\mathrm{OUT}}(n,t) \equiv \sum_{S_D \in \mathcal{K}_l \setminus S_{k(n)}} \; \sum_{S_O \neq S_D} \; \sum_{S_v \in \mathcal{N}_{\mathcal{K}_l}(n)} \pi^{\mathrm{OD}}_{k(n),v,l} \times \mu^0_{\mathrm{OD}t;t_1}.$$

In this equation, $\mathcal{N}_{\mathcal{K}_l}(n)$ are the neighboring stations to $S_{k(n)}$ in the set $\mathcal{K}_l$, and $\pi^{\mathrm{OD}}_{h,i,l}$ is the probability (under the natural regime) of passing first through S_h then S_i at line l during a journey from S_O to S_D (regardless of time). We restrict S_D to belong to $\mathcal{K}_l$, because these are the most difficult destinations to reach by an alternative route.

These probabilities are not directly identifiable from the smartcard data. The problem of estimating unobservable trajectories between two stations is a type of network tomography problem (19). However, TfL has survey data on passenger route choice, the Rolling Origin and Destination Survey (RODS) (20). Combined with prior information on likely routes using structural information of the network topology, we are able to produce Bayesian posterior expected values for $\pi^{\mathrm{OD}}_{h,i,l}$ (*Supporting Information*). The use of RODS data minimizes the need for more sophisticated network tomography models (21–24), for which no software is readily available for the scale of the problem we are operating at (to the best of our knowledge).

A potential difficulty with using the missing outflow as a covariate for our regression model for $\overline{N}^{S[k(n)]}_{t_1:t_F}$ is that, the more distant a destination is, the more likely a passenger will try a different route instead of tapping out early at $S_{k(n)}$. To control for this, we added as a second covariate $\phi^{\mathrm{DIST}}(n)$, the average physical distance (in kilometers) between $S_{k(n)}$ and each $S_{k(m)} \in \mathcal{K}^l, n \neq m$. This covariate is used in our model through a variety of nonlinear transformations (see Fig. S5 for an illustration).

A third covariate in this model is the missing inflow, the amount of traffic that would have exited through $S_{k(n)}$ but will not if the usual route would be through a vertex in the disrupted segment:

$$\phi^{\mathrm{IN}}(n,t) \equiv \sum_{S_O \neq S_{k(n)}} \; \sum_{S_v \in \mathcal{N}_{\mathcal{K}_l}(n)} \pi^{Ok_{(n)}}_{v,k(n),l} \times \mu^0_{Ok_{(n)}t;t_1},$$

with $\phi^{\mathrm{IN}}(n)$ defined analogously.

The fourth covariate is just the expected outcome under the natural state,

$$\phi^{\mathrm{NAT}}(n,t) \equiv \sum_{S_O \neq S_{k(n)}} \mu^0_{Ok_{(n)}t;t_1}$$

and, again, $\phi^{\mathrm{NAT}}(n)$ is defined analogously.

Finally, a fifth covariate, ϕ^{DELAY}, is a binary indicator of whether there were delays elsewhere happening in the same line during the disruption event. We extracted this covariate from the textual description of the disruption events according to TfL logs (*Supporting Information*).

Covariates for Station Disruption. Consider the disruption now being the closure of a single station $S_{\mathcal{K}}$, with no interruption of service except for the fact that no trains stop at $S_{\mathcal{K}}$. Our expectation is that, once $S_{\mathcal{K}}$ closes, passengers will have an increased probability of leaving at one of the stations S_h adjacent to it. We estimated the expected number of exits at each S_h with a regression model.

Define $\pi_h^{O\mathcal{K}}$ as the probability of passing through S_h on a journey that starts at S_O and ends at $S_{\mathcal{K}}$, regardless of which line is taken. We again used RODS data to estimate this quantity (*Supporting Information*). We define the expected missing outflow of S_h into $S_{\mathcal{K}}$ as

$$\phi^{\mathrm{OUT}}(h,t) \equiv \sum_{S_O \neq S_{\mathcal{K}}} \sum_{S_v \in \mathcal{N}_{\mathcal{K}_l}\left(S_{k(n)}\right)} \pi_h^{O\mathcal{K}} \times \mu^0_{O\mathcal{K}t;t_1} \qquad \textbf{[4]}$$

with $\phi^{\mathrm{OUT}}(h)$ defined analogously.

This covariate is meant to capture the excess tap-outs in S_h because of passengers leaving one station earlier than their intended destination $S_{\mathcal{K}}$. However, passengers might tap-out one station past $S_{\mathcal{K}}$, a role that can also be played by S_h with respect to other origins S_O. For this we define $\pi_h^{O\mathcal{K}'}$ as the probability of S_h being in the same line of the final leg of the journey between S_O and $S_{\mathcal{K}}$, but coming after $S_{\mathcal{K}}$. Covariate $\phi^{\mathrm{OUT}'}(h)$ is defined as in Eq. **4**, but using $\pi_h^{O\mathcal{K}'}$ instead.

We also define the covariate $\phi^{\mathrm{NAT}}(h)$, analogous to the case of line segment closure, and distance covariate $\phi^{\mathrm{DIST}}(h)$, the distance between S_h and $S_{\mathcal{K}}$ in kilometers.

Results

For the period of 70 d, we obtained the corresponding two-way line segment disruption events with 768 data points, and the station closure events with 191 data points (see Fig. S6 for raw data plots). Each data point corresponds to the outcome of a particular station at a particular disruption. The least-squares method was used to fit all models.

Table 1. Estimates of model for exit counts in affected line segments

	$\phi^{\mathrm{DELAY}} = 0$ (N = 344, R^2 = 0.93)		$\phi^{\mathrm{DELAY}} = 1$ (N = 424, R^2 = 0.92)	
Parameter	Estimate ± SE	P value	Estimate ± SE	P value
Intercept	−0.05 ± 0.33	0.88	0.07 ± 0.38	0.85
ϕ^{NAT}	1.16 ± 0.02	$<10^{-15}$	1.25 ± 0.02	$<10^{-15}$
ϕ^{IN}	−1.21 ± 0.11	$<10^{-15}$	−1.27 ± 0.07	$<10^{-15}$
ϕ^{OUT}	−0.05 ± 0.06	0.51	0.21 ± 0.06	$<10^{-3}$
$\phi^{\mathrm{OUT}}:\phi^{\mathrm{DIST}}$	0.10 ± 0.04	<0.01	−0.05 ± 0.02	0.05
$\phi^{\mathrm{OUT}}:\phi^{\mathrm{DIST2}}$	−0.001 ± 0.004	<0.01	0.004 ± 0.002	0.11

Disruptions of Line Segments. The ROIs for the line segment problems are stations within each affected segment $\mathcal{K} = \{S_{k(1)}, \ldots, S_{k(M)}\}$ having other connections outside $\mathcal{K}$. Stations without other connections have very few tap-outs or none because no trains can reach them. Stations elsewhere in the system show weaker effects that we did not consider in this study.

We define the model for expected outcomes as

$$E_x\left(\overline{N}^{\mathcal{S}[k_{(n)}]}_{t_1:t_F}\right) \equiv E\left[\overline{N}^{\mathcal{S}[k_{(n)}]}_{t_1:t_F}\middle|\mathsf{PAST}, \phi^{\mathrm{DELAY}} = x\right] \\ \equiv \beta_{0x} + \beta_{1x}\phi^{\mathrm{NAT}} + \beta_{2x}\phi^{\mathrm{IN}} + f_x\left(\phi^{\mathrm{DIST}}\right) \times \phi^{\mathrm{OUT}}, \qquad \textbf{[5]}$$

where the second-order polynomial

$$f_x\left(\phi^{\mathrm{DIST}}\right) \equiv \beta_{3x} + \beta_{4x}\phi^{\mathrm{DIST}} + \beta_{5x}\phi^{\mathrm{DIST}^2}$$

captures the impact of ϕ^{OUT} regulated by the average distance between $S_{k(n)}$ and the remaining elements of the ROI; $x \in \{0,1\}$, and we omit the indexing (n) for clarity.

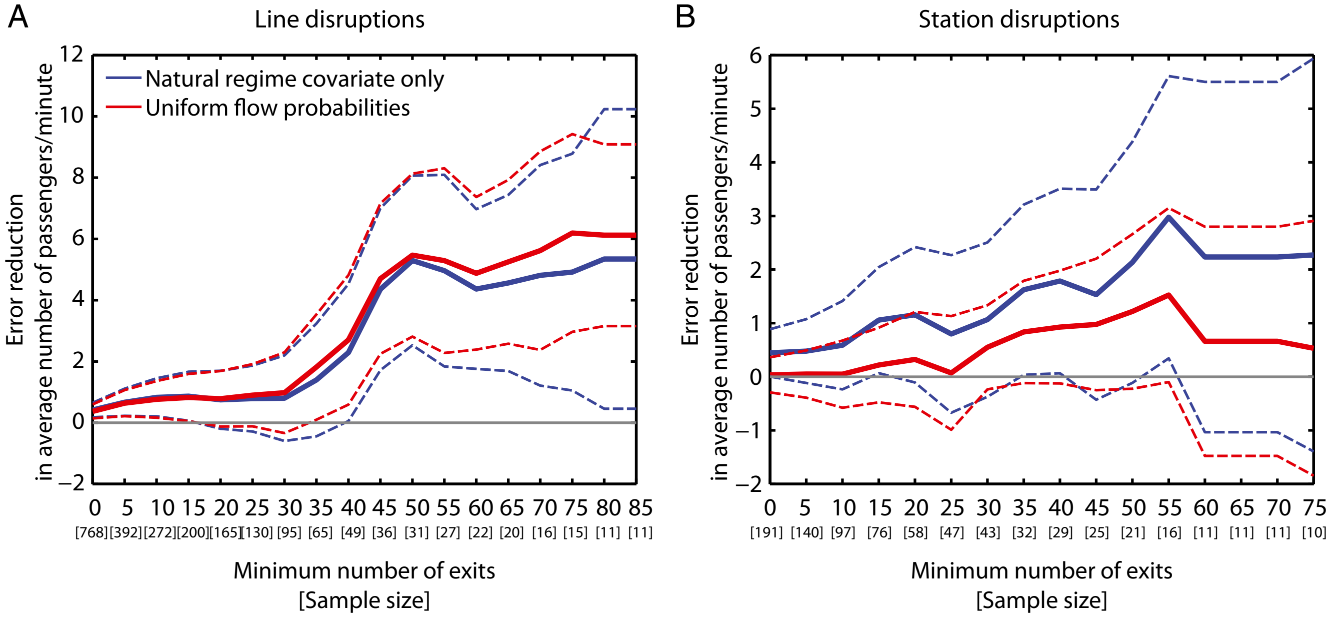

Fig. 3. Reduction in error provided by the model in Eq. **5** compared with baselines (*i*) natural regime covariate only and (*ii*) uniform flow probabilities. Dashed lines show the pointwise 95% confidence intervals on the error reduction. Each point considers test cases with minimum number of tap-outs per minute indicated in the horizontal axis. The number in brackets indicates the number of test cases. (*A*) Relative errors for line segment events. The absolute error of tracking model for the line segment disruption varies from 3.0 (all stations) to 12.2 (stations with 85 tap-outs per minute or more) persons per minute. (*B*) Relative errors for station events. The absolute error varies from 3.5 (all stations) to 10.5 (stations with 75 tap-outs per minute or more) persons per minute.

Table 2. Estimates of model for exit counts in affected neighboring stations

	Estimate ± SE (N = 191, R^2 = 0.95)	P value
Intercept	−0.07 ± 0.59	0.90
ϕ^{NAT}	1.07 ± 0.02	$<10^{-15}$
ϕ^{OUT}	0.59 ± 0.22	<0.01
$\phi^{\mathrm{OUT}}:\phi^{\mathrm{DIST}}$	−0.32 ± 0.20	0.11
$\phi^{\mathrm{OUT}}:\phi^{\mathrm{DIST2}}$	0.01 ± 0.02	0.32
$\phi^{\mathrm{OUT'}}$	0.89 ± 0.23	<0.01
$\phi^{\mathrm{OUT'}}:\phi^{\mathrm{DIST}}$	−0.86 ± 0.29	<0.01
$\phi^{\mathrm{OUT'}}:\phi^{\mathrm{DIST^2}}$	0.17 ± 0.07	0.02

Before fitting the model in Eq. **5**, we show models obtained without the distance covariate ϕ^{DIST},

$$E_0\left(\overline{N}^{S[k_{(n)}]}_{t_1:t_F}\right) = 1.15\phi^{\mathrm{NAT}} - 1.28\phi^{\mathrm{IN}} + 0.16\phi^{\mathrm{OUT}}, \qquad \textbf{[6]}$$

$$E_1\left(\overline{N}^{S[k_{(n)}]}_{t_1:t_F}\right) = 1.24\phi^{\mathrm{NAT}} - 1.23\phi^{\mathrm{IN}} + 0.09\phi^{\mathrm{OUT}}, \qquad \textbf{[7]}$$

because they are easier to interpret than Eq. **5** [standard errors of coefficients: 0.02, 0.11, and 0.02 for the no-delay case and 0.02, 0.07, and 0.02 for the delay case, respectively ($P < 10^{-7}$ each). Intercepts were removed ($P > 0.75$ each)]. This supports the postulated qualitative contributions of flows in Eq. **1**, where the signs match the postulated contribution of the respective flows and the magnitude of the ϕ^{NAT} component is not substantially different from unity. We conclude that, structurally, there is a significant contribution of missing inflows and outflows to the expected tap-out rate, which cannot be explained by a linear rescaling of the natural expected tap-out rate only. Most of the variability in the outcome can be explained by the natural regime and passenger flows ($R^2 > 0.9$).

As a matter of fact, the counterfactual flow ϕ^{NAT} was the covariate with the strongest contribution to the model: Fitting a model with this covariate only gives $E_0(\overline{N}^{S[k_{(n)}]}_{t_1:t_F}) = -0.29 + 1.10\phi^{\mathrm{NAT}}$ and $E_1(\overline{N}^{S[k_{(n)}]}_{t_1:t_F}) = -0.22 + \phi^{\mathrm{NAT}}$ (with $R^2 = 0.9$ and 0.88, respectively). Interestingly, this model seems to hide the impact of closures in the $\phi^{\mathrm{DELAY}} = 1$ case.

Table 1 presents the fitted models of Eq. **5**. The entries of $f_x(\phi^{\mathrm{DIST}}) \times \phi^{\mathrm{OUT}}$ can be interpreted as interaction terms in a linear model. The evidence suggests that the distance from affected stations to other affected stations matters in both cases. For the case with delays, discarding the nonsignificant quadratic term, the results agree with the intuition that as distance grows passengers may feel compelled to find alternative routes, and as such the missing outflow will be penalized. In the case without delays, the result seems contrary to intuition. We propose as an explanation that disruptions without delays are positively associated with line segments that offer fewer alternatives to reach their destinations. In fact, around 53% of the no-delay disruption events we observed included the end of the line (a feature which, on its turn, is associated with longer distances among stations and lack of alternative routes). In contrast, only 38% of the events with delays included the end of a line (*Supporting Information*).

We evaluated our framework by its predictive power using leave-one-out cross-validation (LOOCV). This consists of fitting a model with a training set containing all points but one, which is used for testing. For each fold, the error metric is the absolute difference between the predicted average number of tap-outs per minute against the true average in the test point.

We compare our performance against two baselines. The first is the model with ϕ^{NAT} as the only covariate, and the second a model where flow probabilities $\pi^{\mathrm{OD}}_{k(n),v,l}$ are defined to be constant (that is, they are removed from the definition in Eq. **3**). We focused on fitting models that aggregate both delayed and nondelayed events. To better compare models, we report the difference in the test error averaged over a decreasing subset of test points. Because the amount of tap-outs per station has a skewed distribution, a large number of small-traffic stations will mask the benefits achieved at larger stations. Results are shown in Fig. 3*A*. We report the difference in error between each baseline and our model, for each subset of the test folds considered. As we assess stations of larger traffic, the difference among our method and the baselines becomes more evident. The absolute error of our disruption model for the line segment case varies from 3.0 (all stations) to 12.2 (stations with 85 tap-outs per minute or more) persons per minute. See Tables S2–S5 and Fig. S7 for the absolute error in each class of station, prediction and error scatterplots, and for sensitivity analyses assessing variations of the model in Eq. **5**.

Disruptions of Single Stations. Our ROI for a station closure $S_{\mathcal{K}}$ consists of its neighbors S_h. The model for $\overline{N}^{S[h]}_{t_1:t_F}$, the average tap-count at each S_h, is

$$E\left[\overline{N}^{S[h]}_{t_1:t_F}\,\middle|\,\mathrm{PAST}\right] \equiv \beta_0 + \beta_1\phi^{\mathrm{NAT}} + f\left(\phi^{\mathrm{DIST}}\right) \times \phi^{\mathrm{OUT}} + f'\left(\phi^{\mathrm{DIST}}\right) \times \phi^{\mathrm{OUT'}}, \qquad \textbf{[8]}$$

where $f(\phi^{\mathrm{DIST}}) \equiv \beta_2 + \beta_3\phi^{\mathrm{DIST}} + \beta_4{\phi^{\mathrm{DIST}}}^2$ and $f'(\phi^{\mathrm{DIST}}) \equiv \beta_{2'} + \beta_{3'}\phi^{\mathrm{DIST}} + \beta_{4'}{\phi^{\mathrm{DIST}}}^2$. The fitted model is shown in Table 2.

We performed a LOOCV comparison against two baseline models (Fig. 3*B*) analogous to the line disruption case. The absolute error varies from 3.5 (all stations) to 10.5 (stations with 75 tap-outs per minute or more) persons per minute (see Table S3 for further details). Although there is no strong evidence our model outperforms the uniform flow model statistically (*Supporting Information*), and the improvement over the natural regime baseline is very small, the model is competitive while also revealing insights on passenger behavior. In particular, it suggests that passengers who tap-out at a station S_h immediately after $S_{\mathcal{K}}$ will do it less often as the distance between the two stations increases. This is a way of providing evidence of rational behavior of passengers, which can be used to validate whether announcements of station closures are being properly used by passengers—this

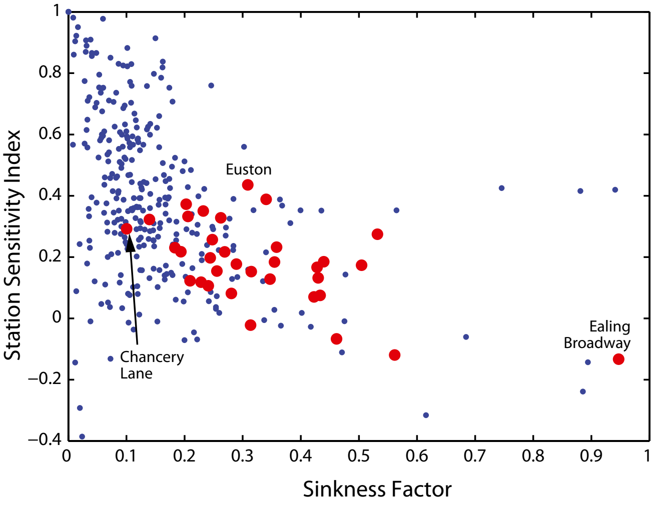

Fig. 4. Station sensitivity index versus sinkness factor for all stations. Red points represent the top 10% of stations as measured by number of tap-outs. Stations with the trivial sinkness of 1 were removed for better visualization.

might not be true if communication between staff and passengers is poor (i.e., if closures are announced only as the train passes through the closed station). This type of analysis can be applied to networks other than the London Underground as a validation of good communication between train drivers and passengers.

Station Sensitivity Index. Besides solving prediction tasks, the models described here allow for a structural understanding of the London transportation system. We provide, as an illustration of information extraction from the fitted models, a categorization of stations by how sensitive they are to closures at line segments containing them, information that is crucial when analyzing the vulnerable points of a transportation network. In particular, for any given station S, consider all sequences of four stations (S, S_1, S_2, S_3), all in the same line, which start at S and follow the physical adjacencies (if the line ends before four stations or if there is a bifurcation at a particular point, stop at the end or bifurcation instead). Consider the scaled expected change in exit numbers $(E[\overline{N}^{S}_{S,t_1:t_F}] - \phi^{\text{NAT}}(S))/\phi^{\text{NAT}}(S)$ as predicted by the model for endpoints without delays, where $t_1 : t_F$ is the peak period from 8:30 AM to 9:30 AM. The station sensitivity index for each S is defined as the maximum over the corresponding normalized expected changes. Notice that the index can be negative, meaning that a station is expected to have fewer passengers tapping out compared with the natural regime. This is the case when missing inflows outnumber other factors, which cannot be captured by the simpler models with only ϕ^{NAT} (*Supporting Information*).

The station sensivity index is the implicit result of several factors, including the degree by which station S is the final destination of passengers who reach at least S in their journey—a “sinkness” factor. The sinkness factor of a station S is given by the ratio N_S/M_S, defined as follows: for each OD pair (S_O, S_D) such that S lies in the shortest path between these two endpoints (as measured by the graph given by the union of all lines), add to N_S the total number of (S_O, S_D) journeys seen in our data, and add to M_S the total number of journeys between S_O and $S_{D'}$ where $S_{D'}$ lies between S and S_D in the shortest path $S_O \cdots \to S \to \cdots \to S_{D'} \to \cdots S_D$. Notice that the ratio N_S/M_S is large if S is the final destination point of a substantial fraction of journeys traversing it, and is equal to 1 if S is the end of a line. Fig. 4 shows a scatterplot between the station sensitivity index and the sinkness factor. The association is nonlinear and strong, summarized by a correlation coefficient of −0.44. In particular, the nonlinearity seems to be due to an interaction between station size with station sensitivity index and sinkness factor. We highlight the top 10% stations in Fig. 4, defined by their total volume of tap-outs in our data. In this case, the correlation coefficient is −0.60.

Discussion

We have shown that it is possible to predict traffic in a complex, real-world transportation network using a model consisting of tens of thousands of nonparametric statistical components. We have also shown how data from the London system provides overwhelming evidence for our hypothesis that traffic under disruption can be decomposed by contrasting it to a counterfactual output and flows that are split among over 100,000 OD pairs. This decomposition is validated by predictive performance under natural and disrupted regimes, and by structural insights that can be extracted from the model, of which we presented only a small sample of possibilities. The analysis presented, to the best of our knowledge, is the largest system-wide predictive study of a complex real urban railway network to date and integrates data from several sources, including smart-card data and passenger surveys.

In particular, our analysis introduces novel ideas on how to combine data from different regimes. Assumptions linking different regimes allow for estimating the effects of a particular shock using only observational data and natural experiments (25–27). Although our shocks are random and should not be strictly interpreted as nonrandom regime indicators, in the usual counterfactual sense (28), we believe that the work presented here provides an entirely novel way of modeling complex transportation networks. It explicitly makes use of modularity assumptions that allow structural claims from a relatively small set of unplanned shocks. Although we used the London transportation system as our case study, similar analyses can be undertaken in any transportation systems where smart-card data and disruption logs are available.

ACKNOWLEDGMENTS. We thank Transport for London for their kind support, including access to the data sources used in this work; Gareth Simmons, Samuel Livingstone, and Gail Leckie for editorial assistance; and the editor and two anonymous reviewers for comments that substantially improved the quality of our manuscript. This research was supported, in part, by National Science Foundation CAREER Award IIS-1149662 and Award IIS-1409177, by Office of Naval Research Young Investigator Program Award N00014-14-1-0485, and by an Alfred P. Sloan Research Fellowship to E.M.A.

1. Banavar JR, Maritan A, Rinaldo A (1999) Size and form in efficient transportation networks. *Nature* 399:130–132.
2. Boelter LMK, Branch MC (1960) Urban planning, transportation and system analysis. *Proc Natl Acad Sci USA* 46(6):824–831.
3. Carey M, Kwieciński A (1994) Stochastic approximation to the effects of headways on knock-on delays of trains. *Transportation Resarch B* 28B(4):251–267.
4. Eagle N, Pentland A, Lazer D (2009) Inferring friendship network structure by using mobile phone data. *Proc Natl Acad Sci USA* 106(36):15274–15278.
5. Guimerà R, Mossa S, Turtschi A, Amaral LAN (2004) The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proc Natl Acad Sci USA* 102(22):7794–7799.
6. Colizza V, Barrat A, Barthélemy M, Vespignani A (2005) The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc Natl Acad Sci USA* 103(7):2015–2020.
7. Newman M, Barabási A-L, Watts DJ, eds (2006) *The Structure and Dynamics of Networks* (Princeton Univ Press, Princeton).
8. Christakis NA, Fowler JH (2013) Social contagion theory: Examining dynamic social networks and human behavior. *Stat Med* 32(4):556–577.
9. Onnela J-P, et al. (2007) Structure and tie strengths in mobile communication networks. *Proc Natl Acad Sci USA* 104(18):7332–7336.
10. Dodds PS, Muhamad R, Watts DJ (2003) An experimental study of search in global social networks. *Science* 301(5634):827–829.
11. Brockmann D, Hufnagel L, Geisel T (2006) The scaling laws of human travel. *Nature* 439(7075):462–465.
12. Rand DG, Arbesman S, Christakis NA (2011) Dynamic social networks promote cooperation in experiments with humans. *Proc Natl Acad Sci USA* 108(48):19193–19198.
13. González MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. *Nature* 453(5):779–782.
14. Wang P, Gonzalez MC, Hidalgo CA, Barabási A-L (2009) Understanding the spreading patterns of mobile phone viruses. *Science* 324(5930):1071–1076.
15. Simini F, Gonzalez MC, Maritan A, Barabási A-L (2012) A universal model for mobility and migration patterns. *Nature* 484(7392):96–100.
16. Shirado H, Fu F, Fowler JH, Christakis NA (2013) Quality versus quantity of social ties in experimental cooperative networks. *Nat Commun* 4:2814.
17. Roth C, Kang SM, Batty M, Barthélemy M (2011) Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PLoS ONE* 6(1):e15923.
18. Transport for London (2012) TfL Factsheet. Available at https://www.tfl.gov.uk/cdn/static/cms/documents/tfl-factsheet.pdf. Accessed June 16, 2014.
19. Vardi Y (1996) Network tomography:Estimating source-destination traffic intensities from link data. *J Am Stat Assoc* 91:365–377.
20. Transport for London (2014) *Rolling Origin and Destination Survey: The Complete Guide, 2003*. Revised October 2010, March 2012, and January 2014 (London Underground Limited, UK).
21. Tebaldi C, West M (1998) Bayesian inference on network traffic using link count data. *J Am Stat Assoc* 93(442):557–573.
22. Cao J, Davis D, Van Der Viel S, Yu B (2000) Time-varying network tomography: router link data. *J Am Stat Assoc* 95:1063–1075.
23. Airoldi EM, Faloutsos C (2004) Recovering latent time-series from their observed sums: Network tomography with particle filters. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York), pp 30–39.
24. Airoldi EM, Blocker AW (2013) Estimating latent processes on a network from indirect measurements. *J Am Stat Assoc* 108(501):149–164.
25. Pearl J (2000) *Causality: Models, Reasoning and Inference* (Cambridge Univ Press, Cambridge, UK).
26. Imbens GW, Rubin DB (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge Univ Press, Cambridge, UK).
27. Dunning T (2012) *Natural Experiments in the Social Sciences* (Cambridge Univ Press, Cambridge, UK).
28. Morgan SL, Winship C (2014) *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (Cambridge Univ Press, Cambridge, UK).

Supporting Information

Silva et al. 10.1073/pnas.1412908112

SI Text

The Data. The public transportation system in Greater London consists mainly of an interconnected network of railways and buses. In 2011, public transportation accounted for about 19 billion passenger kilometers representing 43% of total journey stages, compared with 34% of journey stages by private transportation, 21% walking, and 2% cycling (1). The main railway networks comprise the London Underground network (also called "the Tube"), the Overground rail network, and the DLR.* Fig. S1 shows the map of these railway systems. The London Underground railway dates back to 1863, when it opened serving eight stations between Paddington and Farringdon. Since then, it has grown to 11 lines totaling 402 kilometers of extension serving 270 stations. The London Overground network covers 83 stations on six lines to provide connections between areas outside of Central London. The DLR is an automated light rail system with 45 stations in seven lines, which opened in 1987 to serve the redeveloped Docklands area of London. In 2011, the volume of passenger traffic was 9.5 billion passenger kilometers in Underground, 645 million passenger kilometers in Overground, and 456 million passenger kilometers in DLR (1). The local government body responsible for operation, management, and planning of transportation in London is TfL.

In 2003, TfL introduced an automated fare collection system called the "Oyster" card. This radiofrequency identification-based smart card is a form of electronic ticket used on public transportation system within the London fare zones. Users touch the card on an electronic reader when entering the transport system (a "tap-in") and leaving the system (a "tap-out") to deduct the fare. The Oyster card records the time and place of these transactions. The use of these smart cards is encouraged by the offering of substantially cheaper fares compared with those in cash. By June 2012, TfL issued more than 43 million Oyster cards and their use accounted for more than 80% of all public transport journeys in London (2).

For this study, we obtained from TfL the data containing each single journey taken using an Oyster card for a subset of days between February 14, 2011 and February 9, 2012, a total of 70 weekdays and 25 weekend days. Each Oyster card record (a "tap") contains the date of the transaction, anonymized user ID, and the event time (measured in minutes) and event code. Event codes represent transactions such as adding more credit to the card or cancelling previous transactions. In this study, we consider the events of entering or exiting an Underground, Overground, or DLR stations only. Each tap-out record includes the location of the first entry in the journey, and the corresponding entry time. For the time period covered by our data, the combined system of Underground, Overground, and DLR consisted of 374 stations.† From the raw data, we excluded cases where journeys were not completed, that is, tap-ins without a corresponding tap-out. This was done by considering only records of tap-out events, using the corresponding fields of time and location of first entry as the information about tap-ins. The missing tap-out events with a recorded tap-in can be safely treated as missing from a population that is not of interest to this study, because they are typically due to travelers who attempt to use the card in an invalid way (as in the case where there is no minimal sufficient credit while tapping-in) or due to TfL staff members. After these exclusions, we obtained 210,764,572 journeys by 10,687,141 unique users, with an average of 1.71 journeys per user per day, 1,756,756 unique IDs per day, and 3,010,922 journeys per day. For the analyses reported in the main text, we restrict our study to weekdays only, because the pattern of traffic between weekdays and weekends are different and the effects of disruptions on weekdays are more relevant. We discuss the differences between weekdays and weekend days in *SI Text*, *Assumptions About Weekdays*.

Lines of the London Tube and the Tube Graph. The lines for the Underground and Overground were segmented according to the official TfL classification, whereas the DLR was treated as a single line. Stations that lie at two or more lines were treated as single stations (for instance, King's Cross), with the exception of a few stations that have different identifier codes in the Oyster card system (for instance, Edgware Road at Hammersmith is distinguished from Edgware Road in the Circle Line). A full list of stations and lines is available upon request. We call the undirected graph formed by the conjunction of 374 stations and lines the "Tube graph," where each station is a vertex, with an edge between each pair of stations physically joined by tracks.

Assumptions About Weekdays. Weekdays are not strictly exchangeable. Standard rank tests (we used kstest2 from MATLAB) reject the null hypothesis of pairwise exchangeability at a 0.01 level for any two weekdays using a "bag of minutes" representation, where counts for every minute and every station are pooled together for each particular day of the week, $1{,}200 \times 374$ data points in each day. For instance, in the raw data there is an increase on the overall number of journeys in late Friday evenings compared with the rest of the week.

We nevertheless adopt the assumption of exchangeability of weekdays (and exclude weekends from our analysis) for simplicity. A quick visual inspection of the histograms of the bag of minutes representation reveals no evident visual features separating weekdays during the busiest times (9:00 AM–6:00 PM) and this assumption strongly simplifies the analysis. The difference between weekdays and weekends, however, is very strong. Fig. S2 depicts a summary of our data, illustrating what we claim to be three clusters of days: weekdays, Saturday, and Sunday.

The Model for the Natural Regime. Passengers enter and exit the system at different time points, where time is an integer from $t = 1$ (5:00 AM) to $t = 1{,}200$ (1:00 AM of the next day). Each value of t corresponds to a minute. Each day consists of $|V|^2 \times \mathcal{T}$ outcomes; $|V|$ is the number of stations and $\mathcal{T}$ is the number of time points, 374 and 1,200, respectively. If dependence across time is to be modeled at all, assumptions have to be chosen carefully on how to decouple such dependencies in a model that is easy to interpret and that can be fitted without much computational burden.

Patterns at different OD pairs are expected to show degrees of similarity, particularly in geographically close places. However, we should differentiate trend similarities from stochastic dependencies. Trend similarity states that model parameters for different stations or OD pairs over time should show some proximity according to suitable similarity metrics. Stochastic dependencies are probabilistic associations among random variables modeling different locations. If we postulate that (i) each passenger decides

[*] In addition to these main railway systems, there is a tram system (called Tramlink) of 28 km with 39 stops serving the south London boroughs of Croydon and Merton. The traffic generated by Tramlink in 2011 was only 148 million passenger kilometers. We do not include Tramlink in our study.

[†] The Blackfriars London Underground station and the DLR stations of Stratford, Bank, and Canning Town were mostly closed during this period.

independently when and where to leave from, and where to arrive at, and (*ii*) there are no physical constraints on the journey, then all OD counts will be stochastically independent given time of the day. The first assumption is approximately true when most passengers do not enter the system in groups and entrance times are not jointly affected by stochastic factors (such as delays on a bus bringing two passengers planning to enter the same station). The second assumption is false, taking into account that passengers cannot move independently within the system, and that variability on train arrivals will associate the times that different passengers take to arrive at their destinations. Moreover, given the entrance times of passengers, their exiting times are random but with a degree of predictability.

Our approach is to propagate stochastic dependencies and model exit rates through three sets of processes. The first set describes how many passengers enter a station to start a journey at particular times (entrance processes). The second set describes for how long passengers remain inside the system according to their origin and time of entrance (negotiation processes). Stochastic dependence over time is accounted by these two sets. The third set of processes contains $\mathcal{O}(|V|^2)$ models for OD counts conditioned on the state of the negotiation processes, but without any conditional stochastic dependence across time (exiting processes). We ignore trend similarity in our fitting process, that is, no regularization is used to penalize differences across models for different stations or OD pairs. This estimation procedure is justified by the large amount of data available and the computational cost of methods such as probabilistic matrix factorization (3). We describe the three sets of processes as follows.

The model for the entrance processes. Let L_{it} be the number of passengers entering station S_i at time t. We define $L_{it} \equiv 0$ for $t < 1$ and model the expected values of L_{it} for $t \geq 1$ given the entire past as

$$\mathsf{E}[L_{it}|\mathsf{PAST}, L_{it} > 0] = \left(\theta_{L_{it}} + \sum_{w=1}^{W} \beta_{L_{wi}} L_{i(t-w)}\right)_+, \qquad \textbf{[S1]}$$

where $(x)_+$ means $\max(x, 0)$, and

$$\mathcal{P}(L_{it} > 0|\mathsf{PAST}) \equiv \pi_{L_{it}}. \qquad \textbf{[S2]}$$

Parameter $\theta_{L_{it}}$ represents an unconditional time-dependent mean and explains most of the variation of the data. That is, given $\theta_{L_{it}}$, present behavior is mostly independent of past behavior. Assuming passengers arrive independently given a particular time of the day and location, there should be no dependence on the past at all. Our AR component captures weaker stochastic associations (e.g., people arriving in batches from buses before entering the station), but the AR coefficients $\{\beta_{L_{wi}}\}$ have little impact.

The model does not account for closed stations or unusual behavior. Standard autoregressive models (that is, with no $\theta_{L_{it}}$) are sensitive to station closures with some delay, as illustrated in *SI Text*, *Forecasting, the AR Model, and Sensitivity Analysis for the Natural State Predictive Performance*, but they provide no basis for prediction under a disrupted regime nor are they good models for the natural regime. Here, we are interested in modeling standard behavior, because we do not aim at predicting when external shocks happen (not possible with the data we have), but at deriving what happens when an external shock affects the system.

Parameters $\theta_{L_{it}}$, $t = 1, 2, \ldots, 1{,}200$ are fitted using cubic spline smoothing. We regress L_{it} on t, where $L_{it} > 0$. MATLAB function csaps [version 8.0.0.783 (R2012b)] is used with default parameters, which automatically selects the degree of smoothness. To avoid problems with extrapolation, we set $\theta_{L_{it}}$ to zero for $t \in [0, k_{bottom}] \cup [k_{upper}, 1200]$. Position k_{bottom}, typically taking place in the first few minutes of the day, is defined as the last time point before some tap-in happened in any day in our data. Position k_{upper} is the first time point, typically taking place in the last few minutes of the day, such that no tap-in happened afterward in any day of our data.

Parameters $\pi_{L_{it}}$, $1 \leq t \leq 1{,}200$, are fitted using decision trees by classifying Bernoulli variables $\{L_{it} > 0\}$ on t using R function rpart (version 4.1-8).

These estimates are then plugged into a constrained least-squares regression problem to fit the set $\{\beta_{L_{wi}}\}$ to the residuals $L_{it} - \hat{\theta}_{L_{it}}$, subject to positivity on the expected value of each training point. We use pcls, a function in the mgcv package (version 1.8-2) for nonparametric generalized additive models in R (4).

The fitted model is dominated by the parameters $\theta_{L_{it}}$, with fitted AR coefficients $\{\beta_{L_{wi}}\}$ having little impact, as expected.

The model for the negotiation processes. For each station S_i and time t, we have a (compressed) representation of the number of passengers who entered the system via vertex S_i and have not left the system by time $t - 1$. This representation is called a presence table. The presence table is the empirical distribution of such passengers by the amount of time they have been inside in the system. This empirical distribution is given in seven coarsened time brackets of [1,10] min, [11,20] min, …, [50,60] min, and more than 60 min. For each station S_i and time t, we have the vector $\mathbf{M}_{it} \equiv (M_{it}^1, \ldots, M_{it}^7)$ representing counts at these seven levels. For example, M_{it}^2 represents the number of passengers inside the system at time $t - 1$, who have started at station S_i, and who have entered the system during the interval $[t - 20, t - 11]$.

The temporal evolution of the entries of $\mathbf{M}_{it}$ is modeled through a cascade of nonparametric binomial regression models. We assume that, given time t, variation on M_{it}^k $(1 \leq k \leq 7)$ depends only on $M_{i,t-1}^k$ and $M_{i,t-1}^{k-1}$, being conditionally independent from the more distant past. The model for M_{it}^k for any given day is

$$M_{it}^k|\mathsf{PAST} \sim \mathsf{Binomial}\left(M_{i,t-1}^k + M_{i,t-1}^{k-1}, p_{it}^k\right), \qquad \textbf{[S3]}$$

where we define $M_{i,t-1}^0 \equiv L_{i,t-1}$. Parameter p_{it}^k is a different parameter for each station S_i and time of the day t. The model reflects the fact that whoever stays in the kth time bracket came from either the previous cohort of people in the same bracket k, or from bracket $k - 1$ (which for $k = 0$ corresponds to those who entered the system the minute before). For each station and bracket k, we fit the 1,200 $\{p_{it}^k\}$ parameters with the mgcv package for nonparametric binomial regression using t as the covariate. Although one should expect very weak dependence[‡] between entrance counts L_{it} and $L_{it'}$, we should expect much stronger (marginal) dependence within $\{M_{ik}^k\}$ across time, for the physical reasons explained above.

The model for the exiting processes. Let N_{ijt} be the number of passengers exiting (tapping-out) at station S_j at time t, having started at station S_i. Let R_{ijt} be the sum of the number of passengers in presence table $\mathbf{M}_{it}$, but only for the brackets within 10 min of the median commute time from S_i to S_j, the median estimate being the empirical median of the data. For instance, if the median time is 35 min, the corresponding brackets are {[21,30], [31,40], [41,50]} and $R_{ijt} = M_{it}^3 + M_{it}^4 + M_{it}^5$. The model for N_{ijt} is then

$$E\left[N_{ijt}|\mathsf{PAST}\right] = R_{ijt} \times q_{ijt}. \qquad \textbf{[S4]}$$

Regression is done separately for each of the $|V|^2$ models, by fitting N_{ijt}/R_{ijt} as a nonparametric function of t and using the least-squares cost function. MATLAB's csaps is used again.

Notice that, in principle, the data could allow for $N_{ijt} > R_{ijt}$, because measurement error or misuse of cards leads to tap-outs

[‡] Again, we are speaking of probabilistic dependence here: There is strong evidence that the means should vary smoothly over time. However, given the model, the L_{it} counts at different times should be essentially independent.

not being matched to some tap-ins. Nevertheless, in practice, this never happens, because R_{ijt} is usually far larger than N_{ijt}. R_{ijt} should therefore be considered a convenient summary of the presence table. Notice also that the dynamic models for the negotiation processes do not explicitly depend on N_{ijt}. The raw observed data for all exits is still used to fit the model for $\mathbf{M}_{it}$. $M^{k}_{i,t+1}$ is a function of the set $\{L_{it'}\}$ for $t' < t$ and the observable counts $\{N^{t'}_{idt}\}$, the number of passengers leaving the system who were inside the system for t' minutes, and who left the system at time t via exit point S_d. Even though $N^{t'}_{idt}$ is calculated to derive the presence tables to fit the negotiation process, we do not model $N^{t'}_{idt}$ directly, but only via the aggregates N_{idt} and M^{k}_{it}. Therefore, part of the observable information in the data are lost when compressing it into such models. We nevertheless believe this is a fine enough degree of modeling for our purposes, with $N^{t'}_{idt}$ playing no particular role in the sequel.

A model for N_{ijt} automatically gives a model for $N_{jt} \equiv N_{\cdot jt}$ (number of people exiting station S_j at time t) and $N_{i \cdot t}$ (number of people leaving at time t who started at station S_i). Parameters such as the presence table evolution p^{k}_{it} (Eq. **S3**) provide information about rate of evasion for passengers who started at S_i, and q_{ijt} (Eq. **S4**) can be used to compare destinations by how they absorb passengers from different origins. Some exploratory analysis can be done by clustering such curves, which we leave for future work.

In the next section, we will also use this model to test the assumption of "clumpiness" in the exiting process: given R_{ijt}, tap-out count N_{ijt} does not follow a binomial process with parameter q_{ijt}. The predictive coverage of N_{ijt} given R_{ijt} is not good using the binomial variance. An explanation is the fact that passengers arrive jointly in trains, so there will be a common source of variability because of this quantization effect on the time of departure. This effect is not negligible because we are looking at individual OD levels. Even if we aggregate OD pairs to predict the overall exit counts for a particular station, we must take this into account. For each destination S_j we also introduce the parameter ϕ_j, which does not vary over time. The parameter regulates the covariance of any pair of Bernoulli variables $\{X^{(m)}_{ijt}, X^{(n)}_{ijt}\}$, where $X^{(m)}_{ijt}$ is the binary indicator that a passenger m is leaving at S_j at time t, having started at S_i. Their correlation is given by $\phi_j \times q_{ijt} \times (1 - q_{ijt})$, with $\phi_j = 0$ in the binomial case. We estimate ϕ_j by a method of moments.

Effect of Traffic Stickiness. Another aspect of the model not captured by the blackbox models is that our model relates entrance behavior to exit behavior. This is reflected by the station-level parameter ϕ_j, which dictates the association level between individual exit events from the network. Our assumption that there is a sizeable level of dependence on how people leave stations can be explained by the fact that people are grouped within trains. One way of testing the hypothesis that $\phi_j > 0$ is by the predictive coverage of a model with our estimated $\{\phi_j\}$ parameters.

At any time point t, conditional on the presence table variables $\{M^{k}_{it}\}$, we generate predictive confidence intervals of three different magnitudes (90, 95, and 99%) and compare them against the intervals under the assumption $\{\phi_j = 0\}$. We generated aggregated confidence intervals for each N_{jt} by summing the means and variances of the predicted N_{ijt}, which are all independent in the model across stations and time once we conditioned on $\{M^{k}_{it}\}$. We then use a normal approximation to define the (90, 95, and 99%) predictive confidence intervals.

Let $X^{(m)}_{ijt}$ be the binary event of a particular passenger m leaving station S_j at time t, given the passenger is counted as being in the presence table summary R_{ijt} for N_{jt}. Referring to the model for N_{ijt}, we have

$$E\left[X^{(m)}_{ijt} | R_{ijt}\right] = q_{ijt},$$

where the association among passengers is given by

$$Corr\left(X^{(m)}_{ijt}, X^{(n)}_{ijt} | R_{ijt}\right) = \phi_j.$$

This implies

$$Var\left(N_{ijt} | R_{ijt}\right) = R_{ijt} q_{ijt} \left(1 - q_{ijt}\right)\left(1 + \left(R_{ijt} - 1\right)\phi_j\right). \qquad \textbf{[S5]}$$

To estimate ϕ_j, we first estimate q_{ijt} as before. Then, for a fixed S_j, we calculate the empirical variance of the corresponding Bernoulli trials, averaged over all days and time points, and solve the average of Eq. **S5** for ϕ_j. When the estimate is negative (which is possible, because q_{ijt} was estimated separately), we set ϕ_j to zero.

The predictive intervals obtained under the dependent model averaged over the five folds were $(0.87, 0.91, 0.95)$. For the (binomial) model with $\phi_j = 0$, the intervals were severely underdispersed, with a coverage of $(0.66, 0.73, 0.83)$ (all SE under 0.0001). This provides strong evidence for a need to include a dependence structure among the Bernoulli trials, which in our case has physical explanations.

Conditioning on the internal scale of the system (variables $\{M^{k}_{it}\}$) helps interpretability, because this conditioning allows one to separate variability owing to fluctuations in the entrance numbers at the origin from the degree of dependence between underlying Bernoulli trials of the exit events. Using the physical distance between each station and Oxford Circus as a surrogate to how frequently trains depart a station, we noticed a positive Spearman rank correlation of 0.32 between our estimates of $\{\phi_j\}$ and the physical distance, for our universe of 374 stations.

Fig. S3 shows the average 99% predictive confidence interval for a set of 14 test days independent of 56 d used for fitting the parameters.

Forecasting, the AR Model, and Sensitivity Analysis for the Natural State Predictive Performance. In the main text, we describe the use of a plain AR model for blackbox prediction of aggregated exit counts. The model is simply

$$E\left[N_{jt} | \mathsf{PAST}\right] = \beta_{\mathrm{AR}_0} + \sum_{w=1}^{30} \beta_{\mathrm{AR}_w} N_{j(t-w)}. \qquad \textbf{[S6]}$$

The model is analogous to the entrance process in *SI Text*, *The model for the entrance processes*, except that no smoothing parameter $\theta_{L_{it}}$ is used. The method of least-squares is used to fit this model.

For all models, including the plain AR model, step-ahead forecasts are done by propagating means. This ignores the truncation at zero from Eq. **S1** and similar equations, as positivity is nearly always satisfied and expected values then become linear functions of past expected values for all models. For instance, given tap-out counts observed up to time t_0, we forecast N_{jt} for $t = t_0 + 1, t_0 + 2, \ldots, t_0 + 30$ using the corresponding estimated AR model as follows:

```
i.  Let $C_w = N_{j(t_0+1-w)}$ for $w = 1, 2, \ldots, 30$
ii. For $i$ in $1, 2, \ldots, 30$
    iii. Let $\hat{N}_{j(t_0+i)} = \hat{\beta}_{\mathrm{AR}_0} + \sum_{w=1}^{30} \hat{\beta}_{\mathrm{AR}_w} C_w$
    iv.  For $w$ in $30, 29, \ldots, 2$, let $C_w = C_{w-1}$
         v. Let $C_1 = \hat{N}_{j(t_0+i)}$.
```

This is just an application of iterated expectations to Eq. **S6**.

The blackbox spline model used as a competitor is a regression function from time index t to expected outcome N_{jt},

$$E\left[N_{jt}\right]=f_j(t),$$

for some unknown function $f_j(\cdot)$, where $1\leq t\leq 1{,}200$. A different spline model is fit to each station. MATLAB function csaps for cubic spline fitting is used, as in *SI Text*, *The Model for the Natural Regime*.

To provide further evidence that our model for the natural regime is robust to overfitting, despite estimating every single OD pair traffic, we did further experiments at a coarser resolution of aggregation.

The task is to predict the aggregated exits for all stations in zones 1 and 2, the busiest zones in London, for traffic originated only in zones 3–9. A new blackbox spline model has to be fitted, because the one in the previous section was exclusively for the full aggregation. Our model, however, is exactly the same, but now aggregating different ODs.

With the same fivefold cross-validations setup, the average RMSE difference per load amounts to 0.001 (SE 0.002), providing more evidence that the OD model is robust.

Fig. S4 illustrates a comparison among the proposed tracking model, the blackbox spline model, and the AR model for Oxford Circus station on Monday, February 14, 2011. Oxford Circus is one of the Tube stations with the highest traffic.

The Probabilistic Flow Model. Although more sophisticated network tomography models are available for accurately estimating traffic volumes from and to pairs of stations (e.g., refs. 5–9), they are in general computationally infeasible at the scale of our massive and complex system, and to the best of our knowledge there is no available software applicable to this study. The approach we take highly simplifies computation of the estimators by relying on simple structural features of the network along a rich source of survey data measuring routes taken by actual passengers.

The RODS is a survey of passenger destinations and the routes chosen (10). For each respondent, his or her origin S_O and destination S_D are recorded, along with change points taken during the passenger's journey. We used the 2012 and 2013 surveys, in which 50,410 and 49,253 distinct routes were observed, respectively, with a total of 8,822,636 journeys.

We need an estimate of $\pi^{OD}_{h,i,l}$, the probability of passing first through S_h then S_i at line l, during a journey from S_O to S_D. In our context, given a RODS entry, the most important piece of route choice information for a S_O, S_D journey will be the last point of change.[§] From this, given a line closure event that takes place in the sequence $\mathcal{K}^l\equiv\{S_{k(1)},\ldots,S_{k(n-1)},S_{k(n)},S_{k(n+1)},\ldots,S_{k(M)}\}$ in line l, we obtain $\pi^{OD}_{k(n),v,l}$, for $S_v\in\{S_{k(n-1)},S_{k(n+1)}\}$. Let the shortest path in line l between two stations be defined as the shortest of all paths taken with respect to the subgraph of the Tube graph given by stations from line l only, with the respective edges. Given the last point of change S_X for a trip starting at some arbitrary S_O and ending at some $S_D\in\mathcal{K}^l$, we define

$$Aligned(X,k(n),v,D,l)=\begin{cases}1, & \text{if } S_X \text{ is in } l \text{ and the sequence}\\ & S_X-\cdots S_{k(n)}-S_v-\cdots S_D\\ & \text{is the shortest path in } l \text{ between } S_X \text{ and } S_D.\\ 0, & \text{otherwise.}\end{cases}\qquad\textbf{[S7]}$$

The above definition allows for the cases $S_X=S_{k(n)}$ and $S_v=S_D$.

The idea is to express $\pi^{OD}_{k(n),v,l}$ as a function of $\pi^{OD}_{(X,l)}$, defined as the probability of some S_X at l being the last point of change in a journey from S_O to S_D. The relationship is

$$\pi^{OD}_{k(n),v,l}=\sum_{S_X}\pi^{OD}_{(X,l)}\times Aligned(X,k(n),v,D,l).\qquad\textbf{[S8]}$$

That is, we sum over the probabilities of all possible last points of change for the (S_O,S_D) journey, but including only those S_X such that[¶] the final leg of the journey is $S_X\to\cdots S_{k(n)}\to S_v\to\cdots S_D$ in line l.

To estimate $\pi^{OD}_{(X,l)}$, we assign it a prior and use RODS data to calculate its posterior, using the posterior expected value as our estimate. Each RODS record specifies the last point of change for particular OD pairs, and the number of passengers who made that choice. Because RODS data do not specify the line of the last change point S_X, which might share multiple lines with S_D, we assume it is the line with the least number of hops between S_X and S_D.

The prior for $\pi^{OD}_{(X,l)}$ for all possible pairs of stations S_X and line l is a Dirichlet distribution on pairs (X,l) with hyperparameter entries $n\times\alpha(\pi^{OD}_{X,l})$, where $n=10$ is an effective sample size parameter. Each hyperparameter $\alpha(\pi^{OD}_{X,l})$ quantifies a prior choice for the corresponding probability of pair (X,l) with $\sum_{S_X}\sum_l\alpha(\pi^{OD}_{X,l})=1$, the sum going over all choices of station S_X and line l.

Hyperparameters $\alpha(\pi^{OD}_{X,l})$ are set as follows. Let c_1,c_2,c_3 be three auxiliary hyperparameters of our prior. Let $\mathcal{X}_L$ be the set of triplets (station, line, and cost) defined as follows:

i) If S_O shares a line l with S_D, add (S_O,l,d) to $\mathcal{X}_L$ where d is the distance in hops between S_O and S_D on l.

ii) If S_O does not share any line with S_D, but one can move from S_O to S_D with exactly one line change $l'\to l$ (where S_O is on l' but not l, and S_D is on l but not l'), add to $\mathcal{X}_L$ the triplet $(S_X,l,d+c_2)$ if the distance d for S_X (number of hops from S_O to S_X along l' plus hops from S_X to S_D along l) is the smallest among all stations on l. In case of ties, add all tied pairs.

iii) If moving from S_O to S_D requires at least two line changes, add $(S_X,l,d+c_3)$ to $\mathcal{X}_L$ if (*i*) S_X and S_D are on l, (*ii*) S_X minimizes the sum $d\equiv h_1+h_2$, where h_1 is the number of hops between S_X and S_D on l and h_2 is total number of hops between S_O and S_X in the Tube subgraph given by the union of all lines other than l.

iv) If S_X fails all three criteria above but it is present in some RODS entry as being the last point of change between S_O and S_D, add (S_X,l,∞) to $\mathcal{X}_L$ for all l containing both S_X and S_D.

With these criteria, we first set $\alpha(\pi^{OD}_{X,l})$ to zero for any (X,l) not in $\mathcal{X}_L$. Notice that the implied prior is partially empirical, because the fourth item above looks at RODS data. For all (X,l) that enters $\mathcal{X}_L$ via condition iv above, we set (for now, unnormalized) $\alpha(\pi^{OD}_{X,l})=1/374$.

For all (X,l) in $\mathcal{X}_L$ via i, ii, or iii above, we define

$$s_{Xl}=\max_{\mathcal{X}_L}(c)-c(X,l)+1,$$

where $c(X,l)$ is the corresponding cost entry of (S_X,l,c) in $\mathcal{X}_L$, and $\max_{\mathcal{X}_L}(c)$ is the maximum of all costs in set $\mathcal{X}_L$. We set hyperparameter $\alpha(\pi^{OD}_{(X,l)})$ to s_{Xl} if $\max(s_{Xl})-s_{Xl}\leq c_1$, or set it to baseline value $1/374$ otherwise. This thresholding adds an extra

[§] Notice the last point of change is S_O itself if no changes are made.

[¶] This is a simplifying assumption, because it discards the possibility of passengers' mistakes or irrational behavior such as passing through the destination of interest and then having to come back. However, we do not consider it worthwhile to assign positive probabilities to these events, because it considerably complicates the problem while having no obvious advantage.

penalty to (X, l) choices that are too far from the optimal choice given by $\max(s_{Xl})$. Finally, we normalize $\alpha(\pi^{\mathrm{OD}}_{X,l})$ so that it sums to 1. Hyperparameters (c_1, c_2, c_3) are set as 5,3,7 by trying different hyperparameter values and checking whether the resulting probabilities $\{\alpha(\pi^{\mathrm{OD}}_{X,l})\}$ were plausible according to the background knowledge of the authors.

For the station closure case, where we observe some station $S_{\mathcal{K}}$ closing but lines around it remaining open, we estimate $\pi^{O\mathcal{K}}_{h,\mathcal{K}}$ similarly, except that the sum in Eq. **S8** is now over all lines.

Notice that we could refine our notion of missing outflow by ignoring flows that go through some closed segments in $\mathcal{K}_l$. That is, we would redefine $\pi^{\mathrm{OD}}_{k(n),v,l}$ for events $S_O \to \cdots \to S_{k(n)} \to S_v \to \cdots S_D$ to be nonzero only where $S_{k(n)}$ is the first station in $\mathcal{K}_l$ in this route. The definition of $\phi^{\mathrm{OUT}}(n)$ (Eq. **3** of main text) includes trajectories that are not possible under disruption (that is, all those of the type $S_O \to \cdots \to S_{k(n\mp1)} \to S_{k(n)} \to S_{k(n\pm1)} \to \cdots S_D$). We chose not to account for this refinement to avoid complicating the definitions of inflow and outflow because otherwise we would need different definitions of $\pi^{\mathrm{OD}}_{k(n),v,l}$ for each case. In our sensitivity analyses in *SI Text*, *Sensitivity Analyses* we discuss models for stations only at the endpoints of a disruption, where this is not an issue.

Extracting Disruption Information from the TfL Logs. For the 70 d covered by our smart-card data we also obtained logs of recorded disruption events. Unplanned closures for the line segment events were selected as the ones labeled as "Part Suspended" or "Suspended" in the logs. For the station closure events, the only relevant label was "Closed." This resulted in 3,037 raw entries of line disruptions and 1,335 raw entries of station events. Each raw data point is characterized by the line of the disruption event, its two endpoints (if a line event), the starting and ending time, plus some extra textual information that records other relevant pieces of information, such as the cause of the event. Many events are represented by multiple entries. We merged entries if they had the same endpoints, happened in the same line, and the end time of an event was within 10 min from the start time of the next event. We excluded station events if the corresponding station participated in a line event at the same time. We also excluded Overground events, a service which on average is much less frequent than the Underground and DLR. After merging, we also excluded events less than 10 min long. This resulted in 180 line events, generating 786 data points corresponding to different stations in the ROI of each event. This also resulted in 96 station events, resulting in 191 data points corresponding to the neighbors of each affected station. Table S1 shows the distribution of line events broken down by line.

We want to filter and classify the entries in the available TfL logs in two ways: (*i*) Disruptions that take place in a single direction only are excluded, and (*ii*) events that have delays happening elsewhere in the line are marked as such.

To filter line segment closures that took place in only one direction, we searched for the presence of the substrings "bound", "w/b," and "clock," meant to detect the presence of the keywords "west-/east-/north-/south-bound" in the description, or "clockwise"/"counterclockwise" (used for the Circle Line of the Underground).

For classifying line segment closures as being accompanied or not by delays, we consider an event as a delay if the word "delay" was included in the textual description of the event. Many events were distinguished as "severe delays," but in our definition of the delay indicator we do not distinguish between severity levels.

A Note on Data Fitting. It should be noted that the data under the natural regime and the data under disruptions are not completely independent. We exclude a day of L_{jt}, $\mathbf{M}_{jt}$, and $\{N_{ijt}\}$ records when fitting the respective entrance, negotiation, and exiting processes if there is any disruption happening at S_j in the particular day. However, we do not exclude any records for the other processes. Recall that the negotiation processes for all stations are functions of all exit counts, and that there are weak but nonzero stochastic dependencies between time points within and time points outside disruptions. As a result, a minor degree of dependence between the natural regime and disruption data exists. However, we do not observe any impact of this dependence in our analysis. In particular, we repeated our analysis without excluding any records from the natural regime and observed no qualitative difference. To illustrate this, the two models obtained from fitting the line disruption model without distance covariates are now

$$E_0\left(\overline{N}^{\mathcal{S}[k_{(n)}]}_{t_1:t_F}\right) = 1.14\phi^{\mathrm{NAT}} - 1.25\phi^{\mathrm{IN}} + 0.16\phi^{\mathrm{OUT}}, \qquad \textbf{[S9]}$$

$$E_1\left(\overline{N}^{\mathcal{S}[k_{(n)}]}_{t_1:t_F}\right) = 1.24\phi^{\mathrm{NAT}} - 1.21\phi^{\mathrm{IN}} + 0.08\phi^{\mathrm{OUT}}. \qquad \textbf{[S10]}$$

SDs for the no-delay case are 0.02, 0.11, and 0.02. SDs for the delay case are 0.02, 0.07, and 0.02. Compared with Eqs. **6** and **7** in the main text, it is clear that no significant difference exists.

Visualization of Distance Functions. In Fig. S5*A* we show a visualization of the quadratic distance functions $f_0(\phi^{\mathrm{DIST}})$ and $f_1(\phi^{\mathrm{DIST}})$ as given by fitting the models for events without delays $(f_0(\cdot))$ and with delays $(f_1(\cdot))$ for line disruptions, as discussed in the main text. The functions are evaluated at observed ϕ^{DIST} points present in the data.

However, recall that the quadratic coefficient for $f_1(\cdot)$ had no strong statistical significance. To perform some sensitivity analysis on how relevant the quadratic term is for both functions, we added yet another nonlinear transformation of ϕ^{DIST} to generate the functions

$$g_x\left(\phi^{\mathrm{DIST}}\right) \equiv \gamma_{3x} + \gamma_{4x}\phi^{\mathrm{DIST}} + \gamma_{5x}\phi^{\mathrm{DIST2}} + \gamma_{6x}\log\left(\phi^{\mathrm{DIST}}\right).$$

The corresponding plot is shown in Fig. S5*B*. Whereas the shape for the no-delay curve has not been dramatically affected, the curve for the delay case confirms that the nonmononicity of $f_1(\cdot)$ is not strongly supported. Recall that we hypothesize that distance functions should be decreasing to penalize outflows, because passengers who are far from their destination are expected not to tap-out earlier, but to look for an alternative route inside the system. The estimated function for the events with delays conforms to this hypothesis. As explained in the main text, there are explanations of why this is not the case for the events without delays. For a final perspective, Fig. S5*C* shows the case when $f_0(\cdot)$ and $f_1(\cdot)$ are constrained to be linear. Once more, the overall conclusion is that the evidence for the no-delay case points to an increasing function, whereas the delay case points to a (more intuitive) decreasing function.

Visualization of Raw Data and Predictions. Fig. S6 provides scatterplots of the outcome variable and selected covariates under the two cases of full ROI or endpoints only. Fig. S6*A* may suggest that ϕ^{NAT} alone provides a good model for observed exit counts under disruption. Although the fit is good, including inflows and outflows in the model improves its predictive abilities compared with a model with ϕ^{NAT} only as shown in Fig. 3 (main text) and Fig. S7*A* and discussed in *SI Text*, *Sensitivity Analyses* below (in particular Table S4). Also, models with inflow and outflow covariates such as Eqs. **6** and **7** from the main text,

$$E_0\left(\overline{N}^{S[k_{(n)}]}_{t_1:t_F}\right) = 1.15\phi^{\mathrm{NAT}} - 1.28\phi^{\mathrm{IN}} + 0.16\phi^{\mathrm{OUT}},$$

$$E_1\left(\overline{N}^{S[k_{(n)}]}_{t_1:t_F}\right) = 1.24\phi^{\mathrm{NAT}} - 1.23\phi^{\mathrm{IN}} + 0.09\phi^{\mathrm{OUT}}$$

explain scenarios where the expected outcome might be less than the expected natural outcome ϕ^{NAT}, depending on the magnitude of ϕ^{IN} with respect to ϕ^{NAT} and ϕ^{OUT}. Without the ability of explaining decreases in expected outcome under disruption, a theory of disruption effects is incomplete. Comparing the two models above against the models with ϕ^{NAT} only,

$$E_0\left(\overline{N}^{S[k_{(n)}]}_{t_1:t_F}\right) = -0.29 + 1.10\phi^{\mathrm{NAT}},$$

$$E_1\left(\overline{N}^{S[k_{(n)}]}_{t_1:t_F}\right) = -0.22 + \phi^{\mathrm{NAT}},$$

it is clear that these do not account for cases where stations can have fewer than ϕ^{NAT} tap-outs under disruptions. Consider the case of Ealing Broadway station (Fig. 4 in main text; see also Fig. S1), which has fewer tap-outs if either its neighborhood in the Central line or its neighborhood in the District line closes down. A model with ϕ^{NAT} only cannot account for this. A theory of system-wide transportation behavior under shocks should allow for this context-sensitive variability, which we achieve using the fundamental concepts of inflows and outflows. Our framework of inflows and outflows interacting with counterfactual outcomes (Eq. **1**, main text) follows the philosophy of making a model simple, but no simpler than it should be, unlike the model with ϕ^{NAT} only.

Fig. S7 *A* and *B* compare the true outcomes under line disruption, for each of the affected stations (768 cases), against the leave-one-out predictions as given by the model combining delayed and nondelayed events. Fig. S7 *C* and *D* perform the analogous comparison for single-station disruption events.

Sensitivity Analyses. In *SI Text*, *Delays and endpoint models* we evaluate how delays interact with outcomes and under which conditions. A more in-depth look at predictive results is discussed in *SI Text*, *Predictive results*. The effect of distance measures and its evaluations is further explored in *SI Text*, *Distance model*. Finally, in *SI Text*, *Temporally fine-grained predictions* we comment on predictive results for the 1-min resolution setup.

Delays and endpoint models. Consider an alternative model for the line segment disruption problem, where the effect of delays is additive, as opposed to having two separate models for each state of the delay variable:

$$E\left[\overline{N}^{S[j]}_{t_1:t_F}\middle|\mathrm{PAST}\right] = \beta_0 + \beta_1\phi^{NAT} + \beta_2\phi^{IN} + \beta_3\phi^{OUT} + \beta_4\phi^{DELAY}.$$

The fit of this model is summarized in Table S2. It has a good R^2 fit compared with the individual models for $\phi^{DELAY} = 0$ and $\phi^{DELAY} = 1$, but the linear coefficient of the delay covariate does not significantly contribute to the model. However, consider the case where given a disrupted segment $\mathcal{K} = \{S_{k(1)}, \ldots, S_{k(M)}\}$, the ROI is composed only of the endpoints $\{S_{k(1)}, S_{k(M)}\}$ (again, excluding those that are not connected to any station outside $\mathcal{K}$—which will be the case for stations at the end of a line). A priori, this particular ROI suggests behavior that differs from the average station in $\mathcal{K}$, as they receive passengers from line l (as opposed to midpoints $S_{k(n)}$ in $\mathcal{K}$ with external connections; in that case, passengers would be planning to change lines at $S_{k(n)}$). We fit three other models for this subset of stations, again ignoring distance covariates to provide a set of models easier to interpret. These other three models are shown in Table S2. These models in general follow the theoretical structure of having positive outflow and negative inflow contributions, and no strong evidence for intercepts. There is evidence of different behavior between the models with and without delays—in particular, on the contribution of outflows, which is precisely the flow measure we believe to be most affected by delays. Differences in the contribution of outflows were also detected for the model regulated by average distance covariates and all stations, as discussed in the main text. At the same time, the additive contribution of the delay indicator is not significant (Table S2).

Predictive results. Fig. 1 in the main text shows predictive results for number of tap-outs per minute. To give a sense of scale for the RMSE using Underground stations as examples, the RMSE for the 30-min step-ahead problem is 24.9 ± 0.54 for Oxford Circus (average daily traffic per minute at the order of 60) and 22.35 ± 1.69 for King's Cross (average daily traffic per minute at the order of 50).

Table S3 shows predictive results corresponding to Fig. 3 in the main text. The table also addresses how the result changes under different delay conditions. Each of the four panels shows results for a different model: top left is for our model in Eq. **5** (main text) applied to all data, as in Fig. 3*A* (main text); top right and bottom left is our model in Eq. **5** (main text) applied to subsets of data classified according to ϕ^{DELAY}; bottom right is for our model in Eq. **8** (main text) applied to all data, and corresponds to the graph in Fig. 3*B* (main text).

The columns in Table S3 are as follows:

- Filter: indicates which test points are being used in the calculation of the respective statistics. A value of n for Filter means that only test points with an outcome variable of size n or more are used in the calculation of the remaining items in the respective row ($n = 0$ being the complete set of test folds);
- Sample: the number of test points that satisfy the filter condition;
- Error: average "absolute error," which for each test point is the absolute value of the difference between the test tap-out and the tap-out predicted by the respective model. This is averaged over the selected test points. For line disruptions, the model is the one in Eq. **5** of the main text, whereas for station disruptions the model is the one in Eq. **8** of the main text;
- Diff$_N$: difference between the absolute error of the model with ϕ^{NAT} as the only covariate, and the absolute error of our respective model, averaged over the selected test points;
- p_N: P value for the signed paired t test, null hypothesis $\mathrm{Diff}_N = 0$ against the alternative hypothesis $\mathrm{Diff}_N > 0$;
- Diff$_U$: difference between the absolute error of the model with uniform probabilities for the passenger flows, and the absolute error of our respective model, averaged over the selected test points;
- p_U: P value for the signed paired t test, null hypothesis $\mathrm{Diff}_U = 0$ against the alternative hypothesis $\mathrm{Diff}_U > 0$.

Concerning the results for station disruption (bottom right of Table S3), although there is no statistical difference with respect to the uniform probability case, overall our model shows a consistent advantage over this competitor. Flow probabilities seem to matter less in this problem. In particular, without distance covariates for simplicity the model for station exits[‖] is $1.10\phi^{\mathrm{NAT}} + 0.21\phi^{\mathrm{OUT}} + 0.25\phi^{\mathrm{OUT}'}$. That is, the contributions of ϕ^{OUT} and $\phi^{\mathrm{OUT}'}$ are approximately the same in this case.

[‖] All coefficients significant at a 0.01 level, SEs approximately 0.08 for the two outflow covariates.

Distance model. The main motivation for including distance covariates in our analysis is to provide some insight of the impact of outflows as distances to other stations in a disrupted segment change. However, although the contribution of the distance covariates to the model structure is strongly statistically significant, it provides no predictive gain. Table S4 illustrates this by comparing the predictions between our full model with distance covariates and the simpler model, which uses covariates ϕ^{NAT}, ϕ^{IN}, ϕ^{OUT} in the $\phi^{\mathrm{DELAY}} = 1$ case. The distance covariates only give a small but not statistically significant advantage at the cases with the larger stations.

To assess more complex uses of distance covariates, consider the case where, instead of averaging over distances with respect to all stations in the affected line, we weight each flow contribution by destination before aggregating them. In particular, we downweight $\pi^{\mathrm{OD}}_{k(n),v,l} \times \mu^{0}_{\mathrm{OD}t;t_1}$ by a function $g_u(dist(S_D, S_{k(n)}))$, where $g_u(\cdot)$ is some nonlinear transform of a normalized Euclidean distance $dist(\cdot, \cdot)$ between the stations.** Moreover, we once more avoid "self-exits" by summing over $S_O \neq S_D$ only. Therefore, for each function $g_u(\cdot)$, we define the covariate $\phi^{\mathrm{OUT}}_u(n)$ as

$$\phi^{\mathrm{OUT}}_u(n,t) \equiv \sum_{S_D \in \mathcal{K}_l \setminus S_{k(n)}} \sum_{S_O \neq S_D} \sum_{S_v \in \mathcal{N}_{\mathcal{K}_l}(n)} \frac{\pi^{\mathrm{OD}}_{k(n),v,l} \times \mu^{0}_{\mathrm{OD}t;t_1}}{g_u\left(dist\left(S_D, S_{k(n)}\right)\right)}$$

$$\phi^{\mathrm{OUT}}_u(n) \equiv \sum_{t=1}^{t=t_F} \phi^{\mathrm{OUT}}_u(n,t)/F. \qquad \textbf{[S11]}$$

We use a set of four different functions, $g_1(x) = 1$, $g_2(x) = \log(x)$, $g_3(x) = x$, and $g_4(x) = x^2$, to characterize four distinct covariates. As shown by results in Table S4 (columns Diff_2 and p_2), this extra complication did not provide a measurable payoff, whereas the model described in the main text provides an easier interpretation of the role of distance in our outflow measures.

Temporally fine-grained predictions. One final sensitivity analysis experiment covers the case where our model is not for the average exit count $\overline{N}^{\mathcal{S}_{k(n)}}_{t_1:t_F}$, but for each individual minute: $N^{\mathcal{S}_{k(n)}}_t$. As we mentioned in the main text, the model and the procedure for fitting it is directly applicable to any subset of $t_1 : t_F$, because it relies on the same counterfactual exit counts generated for the entire period. Treating each time point as a separate training point, we have a sample of 17,844 measurements for the case without delays and 21,953 for the case with delays. We fit models without the distance covariates, obtaining

$$\begin{aligned} E_0\left[N^{\mathcal{S}_{k(n)}}_t\right] &= 0.49 + 1.13\phi^{\mathrm{NAT}}_t(n,t) - 1.23\phi^{\mathrm{IN}}(n,t) \\ &\quad + 0.14\phi^{\mathrm{OUT}}(n,t), \\ E_1\left[N^{\mathcal{S}_{k(n)}}_t\right] &= -0.03 + 1.19\phi^{\mathrm{NAT}}_t(n,t) - 1.02\phi^{\mathrm{IN}}(n,t) \\ &\quad + 0.10\phi^{\mathrm{OUT}}(n,t), \end{aligned}$$

where all parameters are significant $(P < 10^{-15})$ except the intercept for the model with delays, assuming independence of the time points. Table S5 shows cross-validated predictive results for these two models using the same criteria as in the previous predictive evaluations. Cross-validation is performed by using as the test set the whole time series of one station at one disruption event. Errors are averaged over test folds, each fold averaged over time.

By comparing Table S5 to the results in Table S3, where errors varied between 3 and 11 persons per minute, it is evident that prediction at individual minutes is more difficult than averages over $t_1 : t_F$. We also improve the simple baseline based on $\phi^{\mathrm{NAT}}(n,t)$ only by a small margin (all stations, 0.3–1 in the delayed case, 0.2–3.7 in the no-delays case) because we lose predictive power at this resolution. We should, however, notice that for the purposes of transportation management and policy making (such as providing timely alternative transportation under disruption and long-term planning for expansions) the reliable average prediction over the disruption period is valuable information.

**We use latitude and longitude as coordinates. The normalization factor is the Euclidean distance between King's Cross and Heathrow Terminals 1–2–3, to make distances more interpretable.

1. Transport for London (2012) Travel in London, Report 5. Available at www.tfl.gov.uk/cdn/static/cms/documents/travel-in-london-report-5.pdf. Accessed May 26, 2013.
2. Transport for London (2012) Join in the celebrations across the capital this Summer with a limited edition Summer Oyster card. Available at https://www.tfl.gov.uk/info-for/media/press-releases/2012/june/join-in-the-celebrations-across-the-capital-this-summer-with-a-limited-edition-summer-oyster-card. Accessed May 26, 2013.
3. Mnih A, Salakhutdinov R (2008) Probabilistic matrix factorization. *Advances in Neural Information Processing Systems*, eds Platt JC, Koller D, Singer Y, Roweis ST (Curran Associates, Down, UK), Vol 20, pp 1257–1264.
4. Wood S (2006) *Generalized Additive Models: An Introduction with R* (Harvard Univ Press, Cambridge, MA).
5. Vardi Y (1996) Network tomography: Estimating source-destination traffic intensities from link data. *J Am Stat Assoc* 91:365–377.
6. Tebaldi C, West M (1998) Bayesian inference on network traffic using link count data. *J Am Stat Assoc* 93(442):557–573.
7. Cao J, Davis D, Van Der Viel S, Yu B (2000) Time-varying network tomography: router link data. *J Am Stat Assoc* 95:1063–1075.
8. Airoldi EM, Faloutsos C (2004) Recovering latent time-series from their observed sums: Network tomography with particle filters. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York), pp 30–39.
9. Airoldi EM, Blocker AW (2013) Estimating latent processes on a network from indirect measurements. *J Am Stat Assoc* 108(501):149–164.
10. Transport for London (2014) *Rolling Origin and Destination Survey: The Complete Guide, 2003*. Revised October 2010, March 2012, and January 2014 (London Underground Limited, UK).

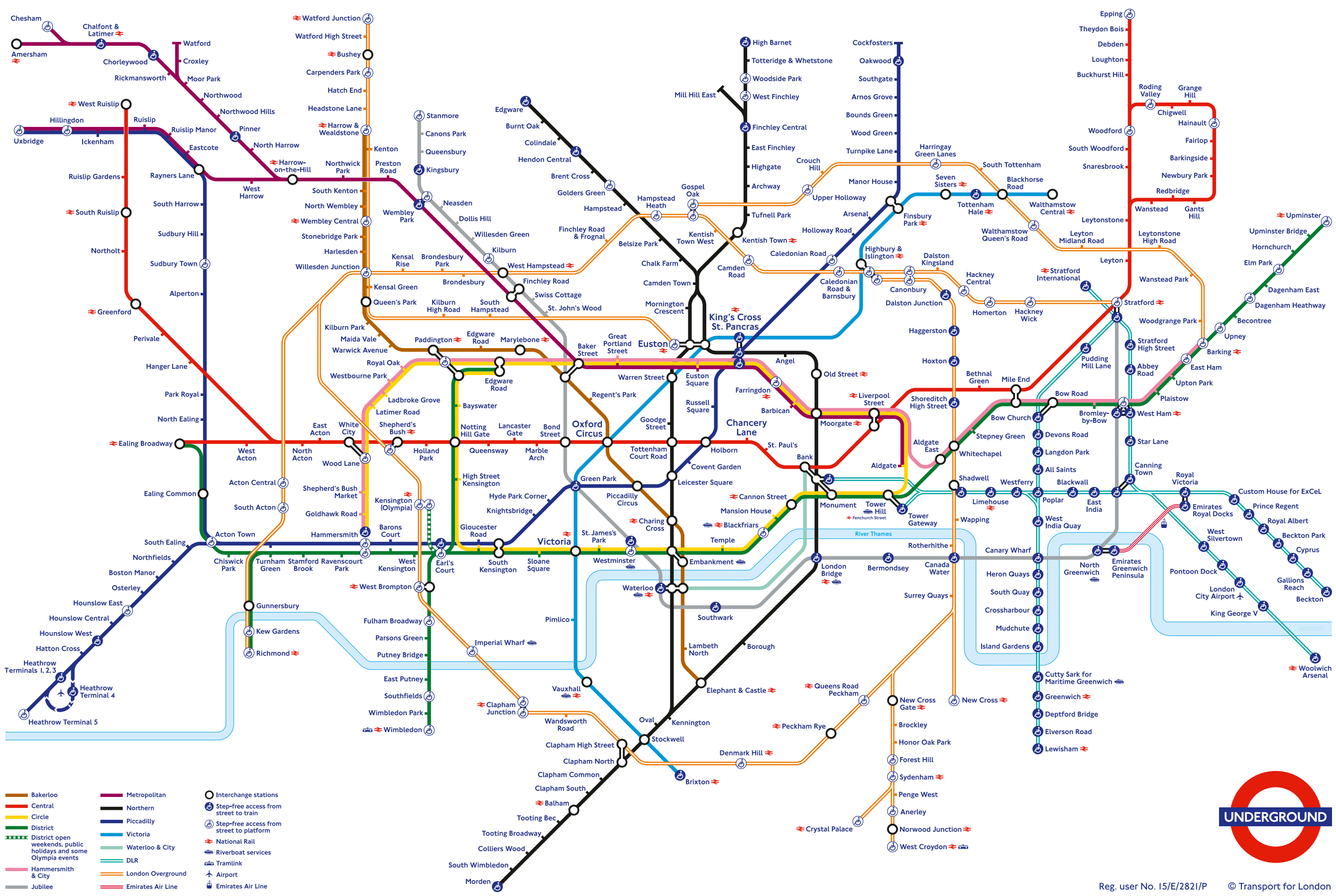

Fig. S1. Tube map including Underground, Overground, and DLR. Reproduced by kind permission of Transport for London, © TfL.

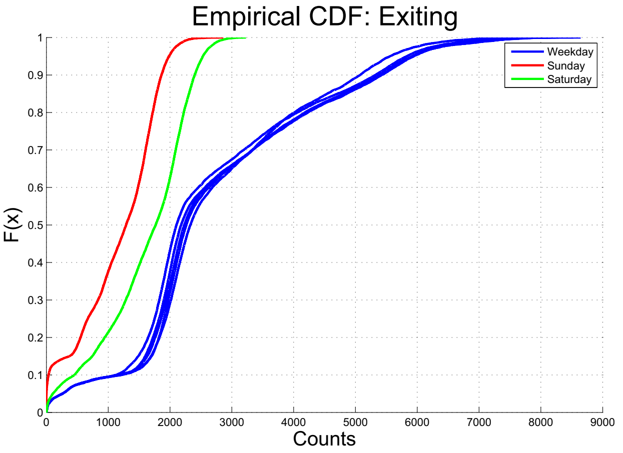

Fig. S2. Cumulative distribution function of exit counts aggregated per day for weekdays and weekends.

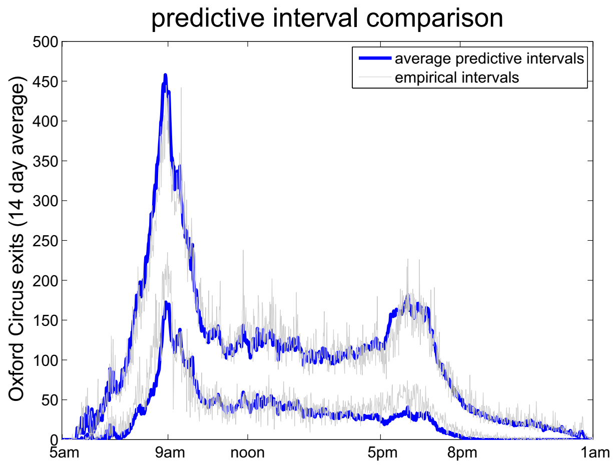

Fig. S3. Comparison between predictive 95% confidence interval and empirical intervals of exits from Oxford Circus station. Average predictive intervals are given by one-step-ahead predictions, which then are averaged over 14 test days. Empirical intervals are given as the empirical quantiles for the 5% symmetric tails.

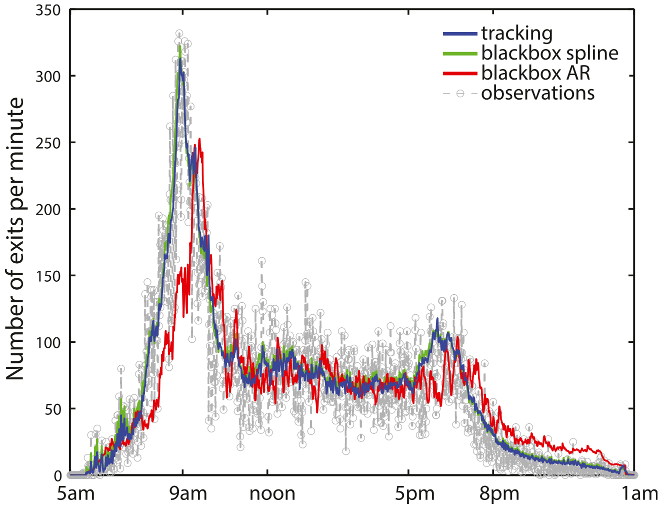

Fig. S4. Thirty-minutes-ahead prediction of the overall number of exits per minute at Oxford Circus station on Monday, February 14, 2011, given past observations of the day.

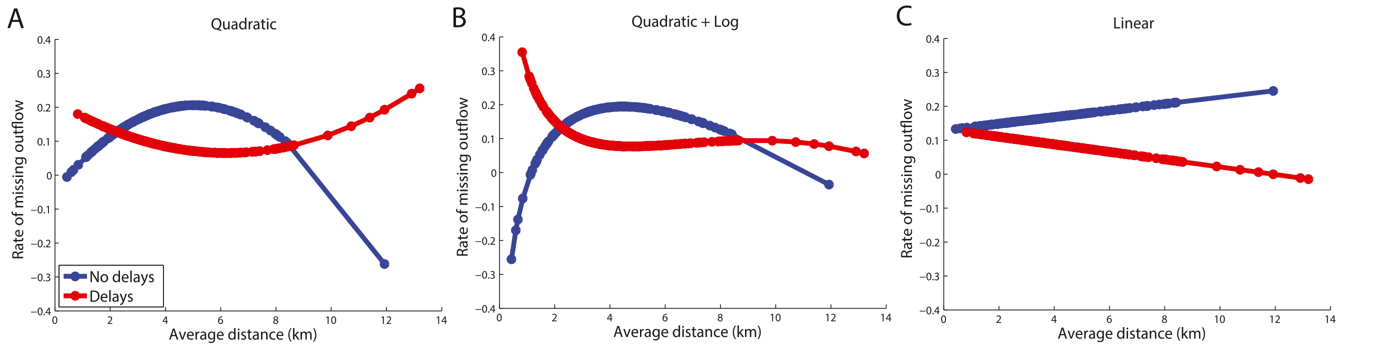

Fig. S5. Effect on ϕ^{DIST} (horizontal axis) on the weighting of ϕ^{OUT} under no delays and under delays using three variations of the distance function. (*A*) Effect for quadratic model. (*B*) Effect for model with quadratic and logarithmic transformations. (*C*) Effect for linear model.

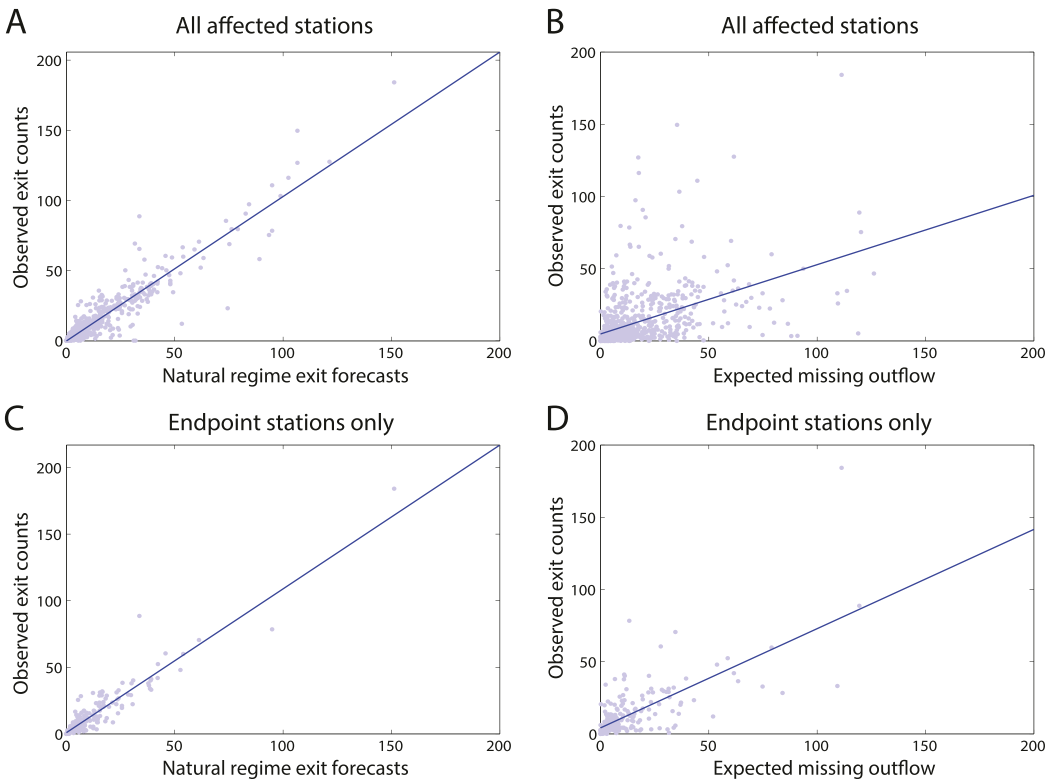

Fig. S6. Association between observed exit counts in stations affected by line disruptions, and some covariates used in prediction. (*A*) Association with respect to the predicted expectations of exits in overall region of interest under the natural regime. (*B*) Association with respect to the missing outflows ϕ^{OUT}. (*C*) Association with respect to natural regime at line segment endpoints only. (*D*) Association with respect to ϕ^{OUT} at line segment endpoints only. The lines in *A* and *C* are the fit given by least squares.

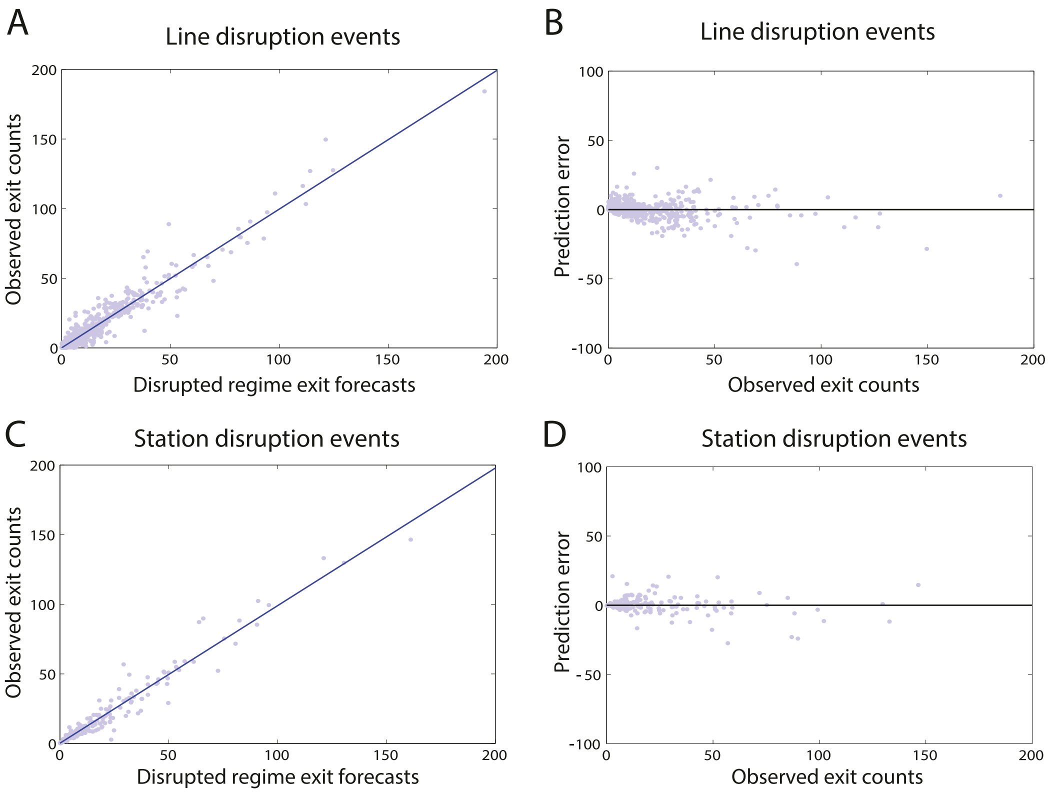

Fig. S7. Visual comparison of predicted and real outcomes for disruptions events. (*A*) All 768 cases of stations affected by line disruptions. (*B*) Corresponding residuals (difference between truth and predicted). (*C* and *D*) The analogous information for station disruption events. The lines in *A* and *C* are the fit given by least squares.

Table S1. Distribution of line events

Line name	No. of events
DLR	11
Bakerloo	12
Central	9
Circle	3
District	29
Hammersmith and City	6
Jubilee	25
Metropolitan	22
Northern	15
Piccadilly	26
Victoria	12
Waterloo and City	0

Table S2. Estimates of model for exit counts in affected line segments

	Linear delay effect (all stations) (N = 768, R^2 = 0.92)		Linear delay effect (endpoint stations only) (N = 204, R^2 = 0.91)	
Parameter	Estimate ± SE	P value	Estimate ± SE	P value
Intercept	0.05 ± 0.44	0.91	1.56 ± 0.87	0.07
ϕ^{NAT}	1.21 ± 0.01	$<10^{-15}$	1.02 ± 0.03	$<10^{-15}$
ϕ^{IN}	−1.22 ± 0.06	$<10^{-15}$	−0.76 ± 0.15	$<10^{-12}$
ϕ^{OUT}	0.10 ± 0.01	$<10^{-3}$	0.20 ± 0.03	$<10^{-6}$
ϕ^{DELAY}	−0.004 ± 0.09	0.96	−0.23 ± 0.16	0.15
	$\phi^{\mathrm{DELAY}} = 1$ (endpoint stations only) (N = 96, R^2 = 0.85)		$\phi^{\mathrm{DELAY}} = 0$ (endpoint stations only) (N = 108, R^2 = 0.91)	
Intercept	1.64 ± 0.90	0.07	0.57 ± 0.53	0.29
ϕ^{NAT}	0.99 ± 0.06	$<10^{-15}$	1.03 ± 0.03	$<10^{-15}$
ϕ^{IN}	−0.71 ± 0.21	0.001	−0.81 ± 0.15	$<10^{-12}$
ϕ^{OUT}	0.11 ± 0.03	0.001	0.20 ± 0.03	$<10^{-7}$

Table S3. Comparison of prediction errors of the full models against the model with ϕ^{NAT} only (Diff_N and p_N) and the model with flows given by uniform probabilities (Diff_U and p_U)

	Lines, all data						Lines, $\phi^{\mathrm{DELAY}} = 0$					
Filter	Sample	Error	Diff_N	p_N	Diff_U	p_U	Sample	Error	Diff_N	p_N	Diff_U	p_U
0	768	3.0	0.4	0.00	0.4	0.00	344	2.7	0.3	0.03	0.2	0.11
5	392	4.6	0.7	0.00	0.6	0.00	153	4.5	0.7	0.02	0.4	0.06
10	272	5.5	0.8	0.00	0.8	0.01	97	5.3	1.1	0.01	0.6	0.05
15	200	6.2	0.9	0.02	0.8	0.02	68	6.2	1.4	0.01	0.7	0.08
20	165	6.9	0.7	0.06	0.8	0.04	51	7.1	1.5	0.03	0.8	0.09
25	130	7.3	0.8	0.07	0.9	0.04	41	7.6	2.1	0.01	1.2	0.01
30	95	8.0	0.8	0.13	1.0	0.07	33	8.2	2.2	0.02	1.4	0.01
35	65	8.7	1.4	0.07	1.8	0.02	24	8.8	2.2	0.07	2.0	0.01
40	49	9.3	2.3	0.02	2.7	0.01	15	10.2	3.7	0.05	2.9	0.01
45	36	9.6	4.4	0.00	4.7	0.00	12	10.8	4.5	0.05	3.5	0.01
50	31	10.1	5.3	0.00	5.5	0.00	10	10.7	6.5	0.01	4.2	0.00
55	27	10.8	5.0	0.00	5.3	0.00	9	11.5	5.7	0.03	3.8	0.01
60	22	11.6	4.4	0.00	4.9	0.00	8	11.9	6.1	0.04	4.3	0.01
65	20	12.1	4.6	0.00	5.3	0.00	8	11.9	6.1	0.04	4.3	0.01
70	16	10.6	4.8	0.01	5.6	0.00	6	10.2	5.5	0.11	3.8	0.03
75	15	11.0	4.9	0.01	6.2	0.00	5	11.9	6.2	0.13	3.7	0.06
80	11	12.2	5.3	0.02	6.1	0.00	4	14.5	7.0	0.17	4.6	0.05
85	11	12.2	5.3	0.02	6.1	0.00	4	14.5	7.0	0.17	4.6	0.05
90	9	10.0	4.4	0.06	5.4	0.00	2	6.3	8.3	0.31	2.3	0.14
	Lines, $\phi^{\mathrm{DELAY}} = 1$						Stations, all data					
0	424	3.4	0.5	0.00	0.5	0.00	191	3.6	0.4	0.02	0.0	0.42
5	239	4.7	0.7	0.01	0.8	0.01	140	4.5	0.5	0.06	0.1	0.41
10	175	5.6	0.9	0.02	1.0	0.01	97	5.4	0.6	0.08	0.0	0.44
15	132	6.2	0.8	0.08	0.8	0.07	76	5.8	1.1	0.02	0.2	0.27
20	114	6.8	0.7	0.13	0.8	0.10	58	6.5	1.2	0.04	0.3	0.23
25	89	7.1	0.6	0.20	0.8	0.14	47	6.7	0.8	0.14	0.1	0.45
30	62	7.9	0.6	0.28	0.8	0.20	43	6.6	1.1	0.07	0.5	0.08
35	41	8.4	1.9	0.07	2.2	0.05	32	7.4	1.6	0.02	0.8	0.04
40	34	8.5	3.0	0.02	3.3	0.01	29	7.5	1.8	0.02	0.9	0.04
45	24	8.5	5.8	0.00	6.2	0.00	25	8.3	1.5	0.06	1.0	0.06
50	21	9.2	6.3	0.00	6.8	0.00	21	8.5	2.1	0.03	1.2	0.05
55	18	9.7	6.2	0.00	6.8	0.00	16	9.3	3.0	0.01	1.5	0.03
60	14	10.6	4.7	0.00	5.5	0.00	11	9.9	2.2	0.08	0.7	0.25
65	12	11.4	4.9	0.01	5.6	0.00	11	9.9	2.2	0.08	0.7	0.25
70	10	10.4	4.6	0.03	5.5	0.01	11	9.9	2.2	0.08	0.7	0.25
75	10	10.4	4.6	0.03	5.5	0.01	10	10.0	2.3	0.10	0.5	0.31
80	7	9.9	6.1	0.02	6.4	0.02	9	11.1	1.6	0.18	−0.1	0.55
85	7	9.9	6.1	0.02	6.4	0.02	9	11.1	1.6	0.18	−0.1	0.55
90	7	9.9	6.1	0.02	6.4	0.02	5	8.4	1.0	0.36	1.5	0.06

See text for details. In particular, absolute errors for our models are shown in the error column.

Table S4. Comparison of prediction errors of the full model (Eq. 5, main text) with distance covariates against the model without distance covariates, for $\phi^{\mathrm{DELAY}} = 1$

Filter	Sample	Error	Diff$_1$	p_1	Diff$_2$	p_2
0	424	3.4	−0.0	0.75	0.0	0.20
5	239	4.7	0.0	0.45	0.1	0.05
10	175	5.6	−0.0	0.55	0.1	0.25
15	132	6.2	−0.0	0.63	0.0	0.37
20	114	6.8	−0.0	0.56	0.1	0.32
25	89	7.1	−0.0	0.66	0.1	0.29
30	62	7.9	0.0	0.43	−0.1	0.85
35	41	8.4	0.0	0.45	−0.1	0.67
40	34	8.5	0.1	0.28	0.0	0.46
45	24	8.5	0.0	0.41	0.2	0.13
50	21	9.2	−0.1	0.64	0.1	0.29
55	18	9.7	−0.0	0.57	0.1	0.25
60	14	10.6	−0.2	0.78	0.0	0.49
65	12	11.4	−0.3	0.89	−0.0	0.55
70	10	10.4	−0.4	0.91	−0.2	0.81
75	10	10.4	−0.4	0.91	−0.2	0.81
80	7	9.9	−0.3	0.78	−0.1	0.65
85	7	9.9	−0.3	0.78	−0.1	0.65
90	7	9.9	−0.3	0.78	−0.1	0.65
95	6	11.2	−0.2	0.68	−0.0	0.52

Prediction error is defined by the absolute difference between the true number of tap-outs in an event of interest and the predicted number of tap-outs, averaged over test points in an LOOCV procedure. Column Diff$_1$ compares the difference in prediction error between the two models, positive numbers indicating an advantage for the full model. Column p_1 is the *P* value of a one-sided *t* test under the alternative hypothesis that our model (Eq. **5**, main text) is better than the competing model. Columns Diff$_2$ and p_2 are measured with respect to yet another way of using distance covariates, as explained in the text.

Table S5. Prediction errors of the models for all cases (with and without delay) for each individual minute

Filter	Error (no delay)	Error (delay)
0	6.8	8.1
5	11.6	11.8
10	14.1	13.9
15	16.0	15.4
20	17.4	16.3
25	18.5	17.0
30	20.3	18.9
35	22.8	21.6
40	26.8	23.0
45	28.6	26.0
50	30.3	27.3
55	32.1	28.9
60	32.2	30.4
65	32.2	32.4
70	35.0	34.3
75	36.6	34.3
80	37.7	31.3
85	37.7	31.3
90	43.3	31.3
95	43.3	31.6