Quantifying the sensing power of vehicle fleets

Kevin P. O'Keeffe[a,1], Amin Anjomshoaa[a], Steven H. Strogatz[b], Paolo Santi[a,c], and Carlo Ratti[a]

[a]Senseable City Lab, Massachusetts Institute of Technology, Cambridge, MA 02139; [b]Department of Mathematics, Cornell University, Ithaca, NY 14853; and [c]Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche, 56124 Pisa, Italy

Edited by Michael F. Goodchild, University of California, Santa Barbara, CA, and approved May 14, 2019 (received for review December 19, 2018)

Sensors can measure air quality, traffic congestion, and other aspects of urban environments. The fine-grained diagnostic information they provide could help urban managers to monitor a city's health. Recently, a "drive-by" paradigm has been proposed in which sensors are deployed on third-party vehicles, enabling wide coverage at low cost. Research on drive-by sensing has mostly focused on sensor engineering, but a key question remains unexplored: How many vehicles would be required to adequately scan a city? Here, we address this question by analyzing the sensing power of a taxi fleet. Taxis, being numerous in cities, are natural hosts for the sensors. Using a ball-in-bin model in tandem with a simple model of taxi movements, we analytically determine the fraction of a city's street network sensed by a fleet of taxis during a day. Our results agree with taxi data obtained from nine major cities and reveal that a remarkably small number of taxis can scan a large number of streets. This finding appears to be universal, indicating its applicability to cities beyond those analyzed here. Moreover, because taxis' motion combines randomness and regularity (passengers' destinations being random, but the routes to them being deterministic), the spreading properties of taxi fleets are unusual; in stark contrast to random walks, the stationary densities of our taxi model obey Zipf's law, consistent with empirical taxi data. Our results have direct utility for town councilors, smart-city designers, and other urban decision makers.

mobile sensing | urban monitoring | urban sustainability | city science

Monitoring urban environments is a challenging task; pollution, infrastructural strain, and other quantities of interest vary widely over many scales—both spatial and temporal—requiring much effort to accurately measure. The field of urban sensing seeks to solve this problem (1–3). With the proliferation of powerful and affordable sensors, it promises a cost-efficient way to monitor urban phenomena at the required fine-scale spatiotemporal resolutions. Even so, traditional approaches to urban sensing, which fall into two main categories (Fig. 1), have limitations. At one extreme, airborne sensors such as satellites scan wide areas, but only during certain time windows. At the other extreme, stationary sensors collect data over long periods of time, but with limited spatial range. Recently, however, "drive-by sensing" has emerged as a new sensing strategy, which offers coverage good in both space and time (4–7). Here, sensors are mounted on "crowd-sourced" urban vehicles, such as cars, taxis, buses, or trucks. This piggy-backing allows the sensors to scan the wide areas traversed by their hosts, allowing the spatiotemporal profile of a city to be explored with great ease and accuracy.

Research on drive-by sensing has so far been technological, focused on the engineering difficulties of the sensors (8), managing the dynamic network they comprise (9, 10), and parsing the data they collect (11–13). Yet, the efficiency of the crowd-sourced aspect of drive-by sensing, on which the viability of the approach rests, has not been analyzed—how many vehicles are required to accurately monitor a city's environment? The answer to this question hinges on the mobility patterns of the host fleet; wide coverage requires the vehicles to densely explore a city's spatiotemporal "volume." We call the extent to which a vehicle fleet achieves this their sensing power. In what follows, we present a case study of the sensing power of taxi fleets. We choose to study taxis as sensor hosts because they are pervasive in cities and because datasets characterizing their mobility patterns are publicly available.

Consider a fleet of sensor-equipped vehicles $\mathcal{V}$ moving through a city, sampling a reference quantity X during a time period $\mathcal{T}$. We represent the city by a street network S, whose nodes represent possible passenger-pickup and -dropoff locations and whose edges represent street segments potentially scannable by the vehicle fleet during $\mathcal{T}$. We use the proviso "potentially scannable" since some segments are never traversed by taxis in our datasets and so are permanently out of reach of taxi-based sensing, as further discussed in *SI Appendix*. To model the taxis' movements, we introduce the taxi-drive process, a schematic of which is presented in Fig. 2 *A*–*C*. The model assumes that taxis travel to randomly chosen destinations via shortest paths, with ties between multiple shortest paths broken at random. Once a destination is reached, another destination is chosen, again at random, and the process repeats. To reflect heterogeneities in real passenger data, destinations in the taxi-drive process are not chosen uniformly at random. Instead, previously visited nodes are chosen preferentially: The probability $q_n(t)$ of selecting a node n is proportional to $1 + v_n^{\beta}(t)$, where $v_n(t)$ is the total number of times node n has been visited at during $[t_{start}, t)$ and β is an adjustable parameter that depends on the city. This "preferential return" mechanism is known to capture the statistical properties of human mobility (14) and, as we show, also captures those of taxis.

Results

To compare our model to data, we quantify the sensing power of a vehicle fleet as its covering fraction $\langle C \rangle$, defined as the average

Significance

Attaching sensors to crowd-sourced vehicles could provide a cheap and accurate way to monitor air pollution, road quality, and other aspects of a city's health. But in order for so-called drive-by sensing to be practically useful, the sensor-equipped vehicle fleet needs to have large "sensing power"—that is, it needs to cover a large fraction of a city's area during a given reference period. Here, we provide an analytic description of the sensing power of taxi fleets, which agrees with empirical data from nine major cities. Our results show taxis' sensing power is unexpectedly large—in Manhattan; just 10 random taxis cover one-third of street segments daily, which certifies that drive-by sensing can be readily implemented in the real world.

Author contributions: K.P.O., A.A., P.S., and C.R. designed research; K.P.O., A.A., S.H.S., P.S., and C.R. performed research; K.P.O. and A.A. analyzed data; and K.P.O., S.H.S., and P.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: kokeeffe@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1821667116/-/DCSupplemental.

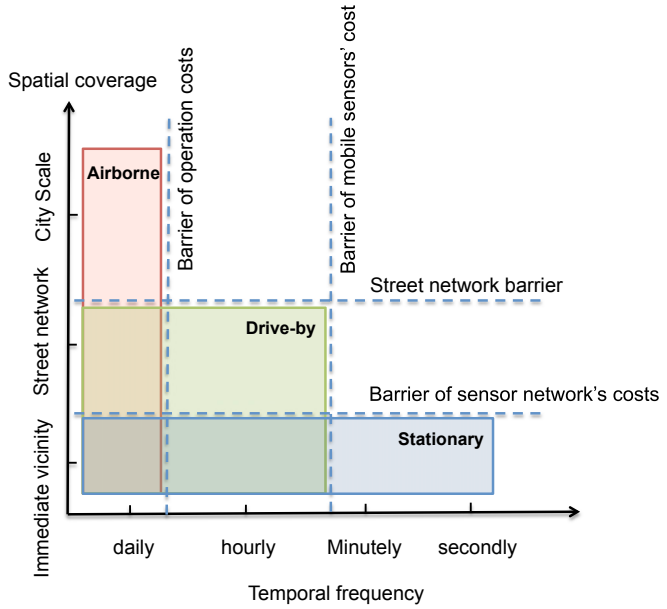

Fig. 1. Comparison of different sensing methods. Airborne sensors, such as satellites, provide good spatial coverage, but their temporal coverage is limited to the time interval when the sensors pass over the location being sensed. Conversely, stationary sensors collect data for long periods of time, but have limited spatial range. Drive-by sensing offers some advantages of both methods. By using host vehicles as "data mules," drive-by sensing offers a cheap, scalable, and sustainable way to accurately monitor cities in both space and time.

fraction of street segments in S that are "covered" or sensed by a taxi during time period $\mathcal{T}$, assuming that N_V vehicles are selected uniformly at random from the vehicle fleet $\mathcal{V}$. (In *SI Appendix*, we consider an alternate definition.)

We have computed $\langle C \rangle$ for 10 datasets from 9 cities: New York (confined to the borough of Manhattan), Chicago, Vienna, San Francisco, Singapore, Beijing, Changsha, Hangzhou, and Shanghai. (We used two independent datasets for Shanghai, one from 2014 and the other from 2015. For the 2015 dataset, we chose the subset of taxi trips starting and ending in the subcity "Yangpu" and hereafter consider it a separate city.) Each dataset consists of a set of taxi trips. The representation of these trips differs, however, by city and roughly falls into two categories. The Chinese cities comprise the first category, in which the global positioning system (GPS) coordinates of each taxi's trajectory were recorded, along with the identification (ID) number of the taxi. Knowing taxi IDs lets us calculate $\langle C \rangle$ explicitly as a function of the number of sensor-equipped vehicles N_V, as desired. Accordingly, we call these the "vehicle-level" datasets. For the remaining cities, however, trips were recorded without taxi IDs; in these cases, we know only how many trips were taken, not how many taxis were in operation for the duration of our datasets. (Although taxi IDs are available for Yangpu and New York City, for reasons discussed in *SI Appendix*, we exclude them from the vehicle-level datasets.) So for these "trip-level" datasets, we can only calculate the dependence of $\langle C \rangle$ on N_T, the number of trips, which serves as an indirect measure of the sensing power. Finally, since we represent cities by their street networks, and not as domains in continuous space, we map GPS coordinates to street segments using OpenStreetMap, so that trips are expressed by sequences of street segments $(S_1, S_2, \dots)$.

We find that, despite its simplicity, the taxi-drive process captures the statistical properties of real taxis' movements. Specifically, it produces realistic distributions of segment popularities p_i, the relative number of times (so that the p_i sum to 1) each street segment is sensed by the fleet $\mathcal{V}$ during $\mathcal{T}$ (in turn, these p_i allow us to calculate our main target, $\langle C \rangle$). Fig. 2*D* shows the empirical distribution of the p_i obtained from our New York dataset (brown histogram). Consistent with previous findings (15), the distribution is heavy-tailed and closely follows Zipf's law (this is also true of the other cities; *SI Appendix*, Fig. S2). The distribution predicted by the taxi-drive process (blue histogram) is consistent with the data. This good agreement is surprising. One might expect that the many factors absent from the taxi-drive process—variations in street-segment lengths and driving speeds, taxi–taxi interactions, human-routing decisions, and heterogeneities in passenger-pickup and -dropoff times and locations—would play a role in the statistical properties of real taxis. Yet our results show that, at the macroscopic level of segment popularity distributions, these complexities are unimportant. Moreover, the agreement of the model and the data are not trivial. Compare, for example, the predictions of a random-walk model (Fig. 2*E*). With their skewed unimodal distribution, the random-walk p_i fails to capture the qualitative behavior observed in the data.

Having obtained the segment popularities p_i, we can predict the sensing power $\langle C \rangle_{N_V}$ analytically by using a simple ball-in-bin model. We treat street segments as "bins" into which "balls" are placed when they are traversed by a sensor-equipped taxi. Using the segment popularities p_i as the bin probabilities (recall that the p_i are normalized and can thus be used as

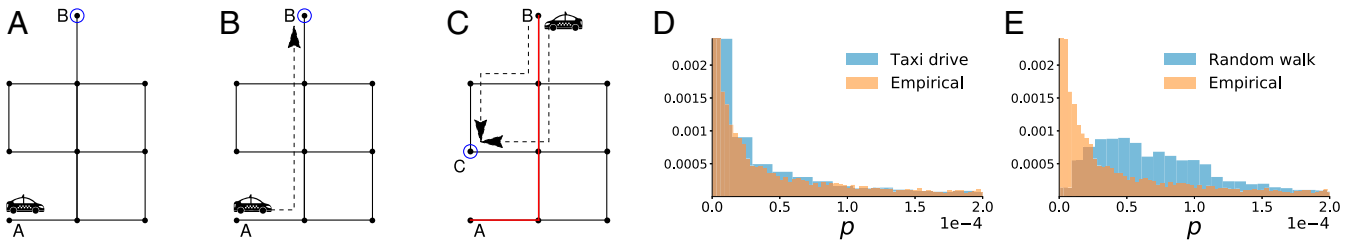

Fig. 2. Taxi-drive process. *A*–*C* show a schematic of the taxi-drive process. (*A*) A taxi picks up a passenger at node *A*. Then, a destination node *B* (blue circle) is randomly chosen. (*B*) The shortest path between *A* and *B* is taken (dashed arrow). No edges have yet been sensed. (*C*) After the edges connecting *A* and *B* have been traversed by the sensor-equipped taxi, they become "sensed," which we denote by coloring them red. Now, at *B*, the taxi proceeds to its next pickup at, say, *C*. There are two shortest paths connecting *B* and *C*, so one is chosen at random. This process then repeats. (*D*) Distribution of street-segment popularities *p* predicted by the taxi-drive process (blue histogram) agrees with empirical data from Manhattan (brown histogram). (*E*) By contrast, a random-walk model of taxi movement (i.e., a random walk performed on the street) incorrectly predicts a skewed, unimodal distribution of street-segment popularities, in qualitative disagreement with the data. For *D* and *E*, the (directed) Manhattan street network on which the taxi-drive and random-walk processes were run was obtained by using the Python package "osmnx." The taxi-drive parameter β was 1.5, and the process was run for $T = 10^7$ time steps, after which the distribution of p_i was approximately stationary.

probabilities), we derive (*Materials and Methods*) the approximate expression

$$\langle C\rangle_{N_V} \approx 1 - \frac{1}{N_S}\sum_{i=1}^{N_S}(1-p_i)^{\langle B\rangle * N_V}. \quad [1]$$

Here, $\langle B\rangle$ is the average distance (measured in segments) traveled by a taxi chosen randomly from $\mathcal{V}$ during $\mathcal{T}$, and N_S is the number of scannable street segments (*SI Appendix*). The trip-level expression $\langle C\rangle_{N_T}$ is the same as Eq. **1** with $\langle B\rangle$ replaced by $\langle L\rangle$, the average number of segments in a randomly selected trip (Materials and Methods, Eq. **10**).

Fig. 3 compares the analytic predictions for $\langle C\rangle$ against our data for a reference period of $\mathcal{T} = 1$ d (see *SI Appendix* for how the empirical C were calculated). We tested the prediction given by Eq. **1** in two ways: using p_i estimated from our datasets (thick line) and using p_i estimated from the stationary distribution of the taxi-drive process (dashed line). In both cases, theory agrees well with data, although the latter estimate is less accurate (which is to be expected, since it is derived from a model). Note that the $\langle C\rangle$ curves from different cities in Fig. 3 are strikingly similar. This similarity stems from the near-universal distributions of p_i (shown in *SI Appendix*, Fig. S1 and discussed in *SI Appendix*) and suggests that $\langle C\rangle$ might also be universal.

Fig. 4 tests for universality in the $\langle C\rangle$ curves. Using the vehicle-level data, we plot $\langle C\rangle$ vs. $N_V/\langle B\rangle$, which removes the city-dependent term $\langle B\rangle$ from Eq. **1** (*SI Appendix*, Figs. S4 and S5 shows how $\langle B\rangle$ and $\langle L\rangle$ are city-dependent. Note that we assume the p_i are universal, so we do not scale them out.) With no other adjustments, the resulting curves nearly coincide, as if collapsing on a single, universal curve. (The fidelity of the collapse, however, varies by day; *SI Appendix*.) In *SI Appendix*, Fig. S10, we perform the same rescaling for the trip-level data, which shows a poorer collapse. However, since these datasets are of lower quality than the vehicle-level data, less trust should be placed in them. Hence, given the good collapse of the vehicle-level data, we conclude that the sensing power of vehicle fleets, as encoded by $\langle C\rangle$, might be universal.

The fast saturation of the $\langle C\rangle$ curves tells us that taxi fleets have large, but limited, sensing power; popular street segments are easily covered, but unpopular segments, being visited so rarely, are progressively more difficult to reach. A law of diminishing returns is at play, which means that, while scanning an entire city is difficult, a significant fraction can be scanned with relative ease. In particular, as detailed in *SI Appendix*, ∼50% of vehicles are required to scan 80% of a city's scannable street segments, but 50% of segments are covered by just $N_V^* \sim 200 \sim 5\%$ of vehicles (*SI Appendix*, Table S3 reports similar numbers for the trip-level data). Most strikingly, as shown in *SI Appendix*, Fig. S11, one-third of the street segments in Manhattan are sensed by as few as 10 random taxis. In fact, because our estimates for B are lower bounds (*SI Appendix*), the above-quoted values for N_V^* are likely lower. These remarkably small values of N_V^* and N_T^* are encouraging findings and certify that drive-by sensing is readily feasible at the city scale, thus achieving the main goal of our work.

Discussion

Requiring that segments be scanned just once a day, as assumed in our analysis, could be too coarse a temporal resolution for some urban quantities which we desire to monitor. Air quality, for instance, has large temporal variations and would therefore require multiple readings dispersed evenly over time to be adequately sensed. To see if drive-by sensing can accommodate more demanding temporal requirements, we derived (*SI Appendix*) an expression for the adjusted sensing power $\langle C^*\rangle(N_T, N_w)$ of a vehicle fleet, defined as the average number of segments covered at least once in each of N_w equally sized subintervals of 1 d. Larger N_w thus corresponds to greater temporal resolution. In *SI Appendix*, Fig. S12, we plot $\langle C^*\rangle$ for $N_w = 3$, so that segments must be scanned at least once in the morning, afternoon, and evening, defined by the intervals $(12\text{AM}, 8\text{AM}), (8\text{AM}, 4\text{PM}), (4\text{PM}, 12\text{AM})$. We find that while reduced, the sensing power is still large; $N_V = 355 = 3\%$ of vehicle are needed to scan half of Manhattan's street segments, and $N_V = 100 < 1\%$ of vehicles scan one-third of segments. This

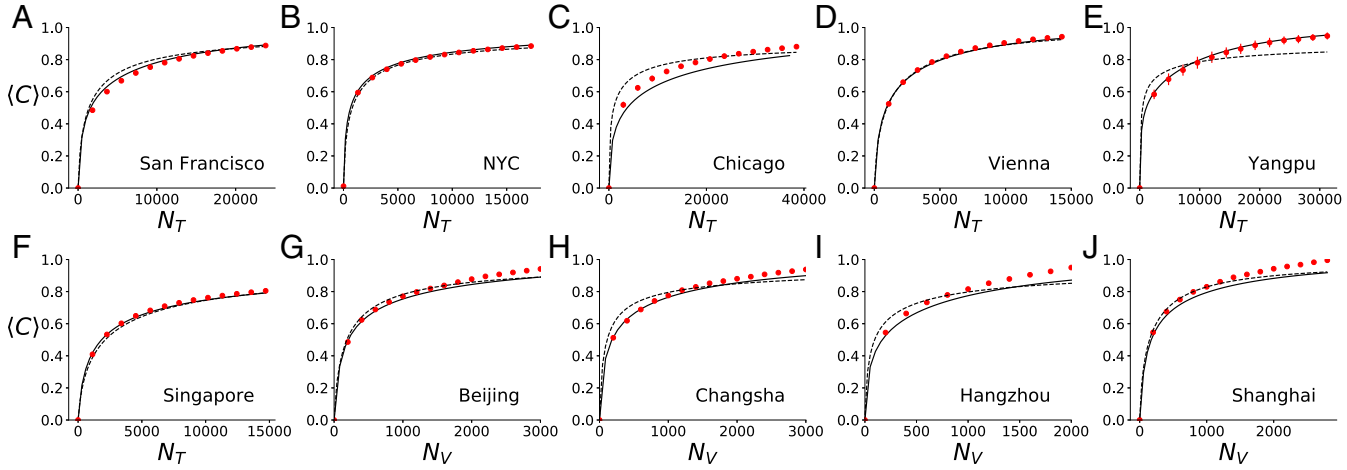

Fig. 3. Sensing power ⟨C⟩. Theoretical and empirical street-covering fractions $\langle C\rangle$ for all datasets are shown. *A–F* show the trip-level data, where the independent variable is the number of trips N_T, and *G–J* show the vehicle-level data, where the independent variable is the number of vehicles N_V. Thick and dashed curves show the analytic predictions for $\langle C\rangle$ by using p_i estimated from data and the taxi-drive process, respectively. Red dots show the empirical $\langle C\rangle$, whose calculation we describe in *SI Appendix*. Notice in *A–F* that the number of trips needed to scan half a city's street segments, N_T^*, is remarkably low—∼2,000 = 10%—and in *G–J*, $N_V^* \sim 5\%$. Exact figures for each N_T^*, N_V^* are given in *SI Appendix*. We list the city name, date, parameter β, and goodness-of-fit parameter for when empirical p_i are used (r^2) and taxi-drive p_i are used ($\tilde{r}^2$) for each city. (*A*) San Francisco, 05/24/08, $\beta = 0.25$, $r^2 = 0.99$, $\tilde{r}^2 = 0.99$. (*B*) New York City (NYC), 01/05/11, $\beta = 1.5$, $r^2 = 0.99$, $\tilde{r}^2 = 0.99$. (*C*) Chicago, 05/21/14, $\beta = 3.0$, $r^2 = 0.93$, $\tilde{r}^2 = 0.92$. (*D*) Vienna, 03/25/11, $\beta = 0.25$, $r^2 = 0.99$, $\tilde{r}^2 = 0.99$. (*E*) Yangpu, 04/02/15, $\beta = 2.75$, $r^2 = 0.99$, $\tilde{r}^2 = 0.71$. (*F*) Singapore, 02/16/11, $\beta = 1.0$, $r^2 = 0.99$, $\tilde{r}^2 = 0.99$. (*G*) Beijing, 03/01/14, $\beta = 1.0$, $r^2 = 0.95$, $\tilde{r}^2 = 0.95$. (*H*) Changsha, 03/01/14, $\beta = 1.75$, $r^2 = 0.95$, $\tilde{r}^2 = 0.94$. (*I*) Hangzhou, 04/21/15, $\beta = 1.25$, $r^2 = 0.90$, $\tilde{r}^2 = 0.90$. (*J*) Shanghai, 03/06/14, $\beta = 0.75$, $r^2 = 0.92$, $\tilde{r}^2 = 0.93$.

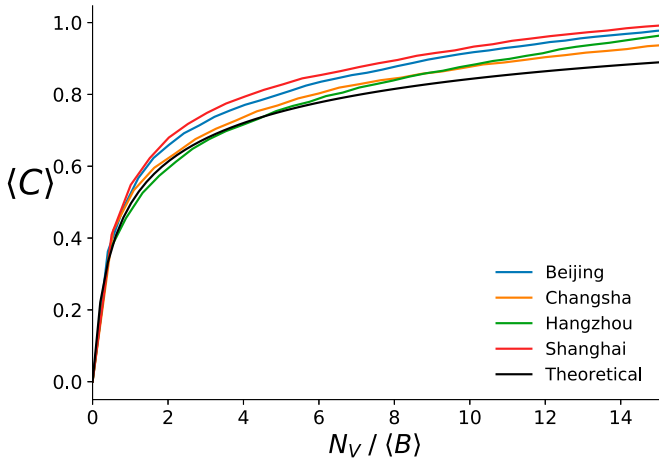

Fig. 4. Scaling collapse. Empirical street-covering fractions $\langle C \rangle$ vs. normalized number of sensor-equipped vehicles $N_V/\langle B \rangle$ from the four vehicle-level datasets. Remarkably, with no adjustable parameters, the curves for all four datasets fall close to the same curve, suggesting that, at a statistical level, taxis cover street networks in a universal fashion. For each dataset, the estimated values of $\langle C \rangle$ were found by drawing N_V vehicles at random and computing the covering fractions. This process was repeated 10 times. The variance in each realization was $O(10^{-3})$, so error bars were omitted. For the theoretical curve Eq. **1**, the p_i was estimated by using the taxi-drive process with $\beta = 1.0$ on the Beijing street network. The choice of Beijing was arbitrary, since, recall, the p_i from different cities are nearly universal.

demonstrates that drive-by sensing can still efficiently provide sensing at finer resolutions in time. *SI Appendix* further discusses this point.

Because taxis are concentrated in commercial and tourist areas, taxi-based drive-by sensing has an inherent spatial bias. This bias could have harmful consequences, such as underserving socioeconomically disadvantaged neighborhoods. A hybrid approach to sensing could overcome this pitfall. Sensor-equipped taxis could be used to scan popular areas of a city, while the remaining hard-to-reach areas could be scanned by vehicles dedicated exclusively to sensing (as opposed to third-party vehicles on which drive-by sensing "parasitically" relies). We discuss the spatial bias of drive-by sensing comprehensively in *SI Appendix*.

There are many ways to extend our results. To keep things simple, we characterized the sensing power of taxi fleets with respect to the simplest possible cover metric: the raw number of segments traversed by a taxi at least once, $C = \sum 1_{(M_i \geq 1)}$ (where, as defined in *Materials and Methods*, M_i is the number of times the i-th segment is sensed at the end of the reference period). A more general metric would be $C = \sum b_i 1_{(M_i \geq 1)}$, where b_i could represent the length of the segment or an effective sensing area. Extensions to the taxi-drive model would also be interesting. We have shown that the taxi drive produces realistic distributions of segment popularities p_i, but have said nothing about the spatial patterns of p_i—could the model be tweaked to capture these patterns? The techniques used in ref. 16 could be useful here. Finally, it would also be interesting to see if the sensing power could be inferred from the street network S alone—that is, without data on taxi-mobility patterns. First steps in this direction are taken in *SI Appendix*.

Taxis traveling in cities share some of the features of non-standard diffusive processes. Like Levy walks (17) or the run-and-tumble motion of bacteria (18), their movements are partly regular and partly random. As such, they produce stationary densities on street networks that obey Zipf's law, contrary to a standard random walk. Future work could examine if other aspects of taxis' spreading behavior are also unusual. Perhaps the hybrid motion exemplified by taxis offers advantages in graph exploration (19), foraging (20), and other classic applications of stochastic processes (21).

The work most closely related to drive-by sensing is on "vehicle-sensor networks" (22). Here, sensors capable of communicating with each other are fitted on vehicles, resulting in a dynamic network. The ability to share information enables more efficient, "cooperative" sensing, but has the drawback of large operational cost. Most studies of vehicle-sensor networks are therefore in silico (23). Since the sensors used in drive-by sensing do not communicate, drive-by sensors are significantly cheaper to implement than vehicle-sensor networks.

Vehicles other than taxis can be used for drive-by sensing. Candidates include private cars, trash trucks, or school buses. Since putting sensors on private cars might lead to privacy concerns, city-owned buses or trucks seem better choices for sensor hosts. The mobility patterns of school buses and trash trucks are, however, different from those of taxis; they follow fixed routes at fixed times, limiting their sensing power.

The diverse data supplied by drive-by sensing have broad utility. High-resolution air-quality readings can help combat pollution, while measurements of air temperature and humidity can help improve the calibration of meteorological models (24, 25) and are useful in the detection of gas leaks (26). Degraded road segments can be identified with accelerometer data, helping inform preventive repair (27, 28), while pedestrian-density data can be helpful in the modeling of crowd dynamics (29). Finally, information on parking-spot occupancy, Wi-Fi access points, and street-light infrastructure—all obtainable with modern sensors—will enable advanced city analytics as well as facilitate the development of new big-data and internet-of-things services and applications.

In short, drive-by sensing will empower urban leaders with rich streams of useful data. Our study reveals these to be obtainable with remarkably small numbers of sensors.

Materials and Methods

We derive an expression for the sensing power of a vehicle fleet. We quantify this by their covering fraction $\langle C \rangle_{N_V}$, the average fraction of street segments covered at least once when N_V vehicles move on the street network S, according to the taxi-drive process during a reference period $\mathcal{T}$. Given the nontrivial topology of S and the non-Markovian nature of the taxi-drive process, it is difficult to solve for $\langle C \rangle_{N_V}$ exactly. We can, however, derive a good approximation. It turns out that it is easier to first solve for the trip-level $\langle C \rangle_{N_T}$ metric—that is, when N_T, the number of trips, is the dependent variable, so we begin with this case; the vehicle-level expression $\langle C \rangle_{N_V}$ then follows naturally.

Imagine we have a population $\mathcal{P}$ of taxi trajectories. We define a taxi trajectory T_r as a sequence of street segments $T_r = (S_{i_1}, S_{i_2}, \ldots)$. The source of the population $\mathcal{P}$ of trajectories is unimportant for now; it could come from a taxi (or fleet of taxis) moving according to the taxi-drive process or from empirical data, as we later discuss. Given $\mathcal{P}$, our strategy to find $\langle C \rangle_{N_T}$ is to map to a "ball-in-bin process": We imagine street segments as bins into which balls are added, when they are traversed by a trajectory taken from $\mathcal{P}$. Note that, in contrast to the traditional ball-in-bin process, a random number of balls is added at each step, since taxis' trajectories have random length.

Trajectories with Unit Length. Let L be the random length of a trajectory. The special case of $L = 1$ is easily solved, because then drawing N_T trips at random from $\mathcal{P}$ is equivalent to placing N_T balls into N_S bins, where N_S is the number of segments, and each bin is selected with probability p_i. As indicated by the notation, we estimate the bin probability with the segment popularities discussed in *Results*; the segment popularities can be used as probabilities, since, recall, they are defined as the relative number of times each segment is traversed, and therefore sum to 1. Let $\vec{M} = (M_1, M_2, \ldots, M_{N_S})$, where M_i is the number of balls in the i-th bin. It is well known that the M_i are multinomial random variables,

$$\vec{M} \sim \text{Multi}(N_T, \vec{p}), \qquad [2]$$

where $\vec{p} = (p_1, p_2, \ldots p_{N_S})$. The (random) fraction of segments covered is

$$C = \frac{1}{N_S} \sum_{i=1}^{N_S} 1_{(M_i \geq 1)}, \quad [3]$$

where 1_A represents the indicator function of event A. The expectation of this quantity is

$$\langle C \rangle_{(N_T, L=1)} = \frac{1}{N_S} \sum_{i=1}^{N_S} \mathbb{P}_{N_T}(M_i \geq 1), \quad [4]$$

(note that we introduce L as a subscript for explanatory purposes). The number of balls in each bin is binomially distributed $M_i \sim Bi(N_T, p_i)$. The survival function of the binomial function is $\mathbb{P}(M_i \geq 1) = 1 - (1 - p_i)^{N_T}$. Substituting this into Eq. **4** gives the result

$$\langle C \rangle_{(N_T, L=1)} = 1 - \frac{1}{N_S} \sum_{i=1}^{N_S} (1 - p_i)^{N_T}. \quad [5]$$

Trajectories with Fixed Length. Trajectories of fixed (i.e., nonrandom) length $L > 1$ impose spatial correlations between the bin-ball counts M_i (recall that in the classic ball-and-bin problem, the M_i are already correlated, since their sum is constant and equal to the total number of balls added). This is because trajectories are contiguous in space; a trajectory that covers a given segment is more likely to cover neighboring segments. Given the nontrivial topology of the street network S, the correlations between bins are hard to characterize. To get around this, we make the strong assumption that for $N_T \gg 1$, the spatial correlations between bins are asymptotically zero. This assumption greatly simplifies our analysis. It lets us reimagine the ball-in-bin process so that adding a trajectory of length L is equivalent to adding L balls into (not necessarily contiguous) bins chosen randomly according to p_i. Then, selecting N_T trajectories of length L from $\mathcal{P}$ is equivalent to throwing $L{*}N_T$ balls into N_S bins $\langle C \rangle_{(N_T, L_{fixed})} = \langle C \rangle_{(N_T * L, L=1)}$. Hence, the expected coverage is a simple modification of Eq. **5**:

$$\langle C \rangle_{(N_T, L_{fixed})} = 1 - \frac{1}{N_S} \sum_{i=1}^{N_S} (1 - p_i)^{L * N_T}. \quad [6]$$

Assuming neighboring segments are spatially uncorrelated is a drastic simplification and effectively removes the spatial dimension from our model. Yet, surprisingly, as shown in Fig. 3, it leads to predictions that agree well with data.

Trajectories with Random Lengths. Generalizing to random L is straightforward. Let $S_{N_T} = \sum_{i=1}^{N_T} L_i$ be the number of segments covered by N_T trajectories. By the law of total expectation

$$\langle C \rangle_{(N_T, L)} = \sum_{n=0}^{\infty} \langle C \rangle_{(n, L_{fixed})} \mathbb{P}(S_{N_T} = n). \quad [7]$$

The first term in the summand is given by Eq. **6**. For the second term, we need to know how the trajectory lengths are distributed. In *SI Appendix*, Fig. S4, we show $L \sim Lognormal(\tilde{\mu}, \tilde{\sigma}^2)$ (note, L here measures the number of segments covered by a taxi, and not the distance of the trip; *SI Appendix*). It is known that a sum of lognormal random variables is itself approximately lognormal $S_{N_T} \sim Lognormal(\mu_S, \sigma_S^2)$, for some μ_S and σ_S. There are many different choices for μ_S, σ_S; for a review, see ref. 30. We follow the Fenton–Wilkinson method, in which $\sigma_S^2 = \ln\left(\frac{\exp \tilde{\sigma}^2 - 1}{N_T} + 1\right)$ and $\mu_S = \ln(N_T \exp(\tilde{\mu})) + (\tilde{\sigma}^2 - \sigma_S^2)/2$. Then,

$$\mathbb{P}(S_{N_T} = n) = \frac{1}{n\sigma_S\sqrt{2\pi}} e^{-\frac{(\ln n - \mu_S)^2}{2\sigma_S^2}}. \quad [8]$$

Substituting this into Eq. **7** gives

$$\langle C \rangle_{(N_T, L)} = \frac{1}{N_S n \sigma_S \sqrt{2\pi}} \sum_{n=0}^{\infty} \sum_{i=1}^{N_S} \left(1 - (1 - p_i)^n\right) e^{-\frac{(\ln n - \mu_S)^2}{2\sigma_S^2}}. \quad [9]$$

The above equation fully specifies the desired $\langle C \rangle_{(N_T, L)}$. It turns out, however, that the sum over n is dominated by its expectation, so we collapse it, replacing n by its expected value $\langle L \rangle * N_T$. This yields the much simpler expression $\langle C \rangle_{(N_T, L)} = \langle C \rangle_{(N_T * \langle L \rangle, L=1)}$, or

$$\langle C \rangle_{N_T} \approx 1 - \frac{1}{N_S} \sum_{i=1}^{N_S} (1 - p_i)^{\langle L \rangle * N_T}, \quad [10]$$

which appears in *Results*.

Extension to Vehicle Level. Translating our analysis to the level of vehicles is straightforward. Let B be the random number of segments that a random vehicle in $\mathcal{V}$ covers in the reference period $\mathcal{T}$ (in *SI Appendix*, Fig. S4, we show how B are distributed in our datasets). Then, we simply replace $\langle L \rangle$ with $\langle B \rangle$ in the expression for $\langle C \rangle_{N_T}$ to get $\langle C \rangle_{N_V}$,

$$\langle C \rangle_{N_V} \approx 1 - \frac{1}{N_S} \sum_{i=1}^{N_S} (1 - p_i)^{\langle B \rangle * N_V}. \quad [11]$$

Model Parameters. The parameters $\langle L \rangle$, $\langle B \rangle$ in Eq. **11** as easily estimated from our datasets (*SI Appendix*). The bin probabilities p_i are trickier. They have a clear definition in the ball-in-bin formalism, but in our model, the interpretation is not as clean; they represent the probability that a subunit of a trajectory taken at random from $\mathcal{P}$ covers the i-th segment S_i. As mentioned above, we estimate these with the segment popularities, which we calculate in two ways: (*i*) deriving them directly from our datasets; or (*ii*) from the taxi-drive process (recall that these methods led to similar distributions of p_i, as shown in Fig. 2*D*).

ACKNOWLEDGMENTS. We thank Allianz, Amsterdam Institute for Advanced Metropolitan Solutions, Brose, Cisco, Ericsson, Fraunhofer Institute, Liberty Mutual Institute, Kuwait–MIT Center for Natural Resources and the Environment, Shenzhen, Singapore–MIT Alliance for Research and Technology, UBER, Victoria State Government, Volkswagen Group America, and all of the members of the MIT Senseable City Laboratory Consortium for supporting this research. S.H.S. was supported by NSF Grants DMS-1513179 and CCF-1522054.

1. N. D. Lane, S. B. Eisenman, M. Musolesi, E. Miluzzo, A. T. Campbell, "Urban sensing systems: Opportunistic or participatory?" in *Proceedings of the 9th Workshop on Mobile Computing Systems and Applications* (ACM, New York, 2008), pp. 11–16.
2. D. Cuff, M. Hansen, J. Kang, Urban sensing: Out of the woods. *Commun. ACM* **51**, 24–33 (2008).
3. T. Rashed, C. Jürgens, *Remote Sensing of Urban and Suburban Areas* (Springer Science & Business Media, New York, 2010), vol. **10**.
4. U. Lee, M. Gerla, A survey of urban vehicular sensing platforms. *Comput. Networks* **54**, 527–544 (2010).
5. B. Hull *et al.*, "Cartel: A distributed mobile sensor computing system" in *Proceedings of the 4th International Conference on Embedded Networked Sensor Systems* (ACM, New York, 2006), pp. 125–138.
6. P. Mohan, V. N. Padmanabhan, R. Ramjee, "Nericell: Rich monitoring of road and traffic conditions using mobile smartphones" in *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems* (ACM, New York, 2008), pp. 323–336.
7. A. Anjomshoaa *et al.*, City scanner: Building and scheduling a mobile sensing platform for smart city services. *IEEE Internet Things J.* **5**, 4567–4579 (2018).
8. J. H. Ahnn, M. Potkonjak, "VeSense: Energy-efficient vehicular sensing" in *2013 IEEE 77th Vehicular Technology Conference* (IEEE, Piscataway, NJ, 2013).
9. A. Skordylis, N. Trigoni, Efficient data propagation in traffic-monitoring vehicular networks. *IEEE Trans. Intell. Transp. Syst.* **12**, 680–694 (2011).
10. M. J. Piran, G. R. Murthy, G. P. Babu, E. Ahvar, "Total gps-free localization protocol for vehicular ad hoc and sensor networks (vasnet)" in *2011 Third International Conference on Computational Intelligence, Modelling and Simulation (CIMSiM)*, (IEEE, Piscataway, NJ, 2011) pp. 388–393.
11. I. Turcanu, P. Salvo, A. Baiocchi, F. Cuomo, An integrated vanet-based data dissemination and collection protocol for complex urban scenarios. *Ad Hoc Networks* **52**, 28–38 (2016).
12. R. Bridgelall, Precision bounds of pavement distress localization with connected vehicle sensors. *J. Infrastruct. Syst.* **21**, 04014045 (2014).
13. G. Alessandroni *et al.*, "Sensing road roughness via mobile devices: A study on speed influence" in *2015 9th International Symposium on Image and Signal Processing and Analysis (ISPA)* (IEEE, Piscataway, NJ, (2015).
14. C. Song, T. Koren, P. Wang, A. L. Barabási, Modelling the scaling properties of human mobility. *Nat. Phys.* **6**, 818–823 (2010).
15. P. Wang, T. Hunter, A. M. Bayen, K. Schechtner, M. C. González, Understanding road usage patterns in urban areas. *Sci. Rep.* **2**, 1001 (2012).
16. Y. Song, H. J. Miller, X. Zhou, D. Proffitt, Modeling visit probabilities within network-time prisms using m arkov techniques. *Geogr. Anal.* **48**, 18–42 (2016).

17. M. F. Shlesinger, J. Klafter, B. J. West, Levy walks with applications to turbulence and chaos. *Physica A Stat. Mech. Appl.* **140**, 212–218 (1986).
18. M. J. Schnitzer, Theory of continuum random walks and application to chemotaxis. *Phys. Rev. E* **48**, 2553–2568 (1993).
19. B. Tadić, "Exploring complex graphs by random walks" in *AIP Conference Proceedings* (AIP, Melville, NY, 2003), vol. **661**, pp. 24–27.
20. G. M. Viswanathan, M. G. Da Luz, E. P. Raposo, H. E. Stanley, *The Physics of Foraging: An Introduction to Random Searches and Biological Encounters* (Cambridge Univ Press, Cambridge, UK, 2011).
21. D. Ben-Avraham, S. Havlin, *Diffusion and Reactions in Fractals and Disordered Systems* (Cambridge Univ Press, Cambridge, UK,2000).
22. D. Van Le, C.K. Tham, Y. Zhu, "Quality of information (qoi)-aware cooperative sensing in vehicular sensor networks" in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)* (IEEE, Piscataway, NJ, 2017), pp. 369–374.
23. M. Gerla, J. T. Weng, E. Giordano, G. Pau, "Vehicular testbeds—validating models and protocols before large scale deployment" in *2012 International Conference on Computing, Networking and Communications (ICNC)* (IEEE, Piscataway, NJ, 2012), pp. 665–669.
24. M. I. Mead *et al.*, The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. *Atmos. Environ.* **70**, 186–203 (2013).
25. R. J. Katulski *et al.*, Mobile system for on-road measurements of air pollutants. *Rev. Sci. Instrum.* **81**, 045104 (2010).
26. P. S. Murvay, I. Silea, A survey on gas leak detection and localization techniques. *J. Loss Prev. Process Indust.* **25**, 966–973 (2012).
27. T. M. Nadeem, M. T. Loiacono, "Mobile sensing for road safety, traffic management, and road maintenance." US Patent 8,576,069 (2013).
28. M. Wang, R. Birken, S. S. Shamsabadi, "Framework and implementation of a continuous network-wide health monitoring system for roadways" in *Nondestructive Characterization for Composite Materials, Aerospace Engineering, Civil Infrastructure, and Homeland Security 2014* (International Society for Optics and Photonics, Bellingham, WA, 2014), vol. **9063**, p. 90630H.
29. M. B. Kjærgaard, M. Wirz, D. Roggen, G. Tröster, "Mobile sensing of pedestrian flocks in indoor environments using wifi signals" in *2012 IEEE International Conference on Pervasive Computing and Communications (PerCom)* (IEEE, Piscataway, NJ, 2012), pp. 95–102.
30. B. R. Cobb, R. Rumi, A. Salmerón, Approximating the distribution of a sum of log-normal random variables. *Stat. Comput.* **16**, 293–308 (2012).