PNAS
www.pnas.org

Supplementary Information for

Quantifying the sensing power of crowd-sourced vehicle fleets

Kevin P O'Keeffe, Amin Anjomshoaa, Steven Strogatz, Paolo Santi, Carlo Ratti

Kevin P O'Keeffe
E-mail: kokeeffe@mit.edu

This PDF file includes:

Supplementary text
Figs. S1 to S18
Tables S1 to S6
References for SI reference citations

Supporting Information Text

Data sets

We have 10 real-world data sets from 9 cities: New York City (confined to the buriough of Manhattan), Chicago, Vienna, San Francisco, Singapore, Beijing, Changsha, Hangszhou, and Shanghai. We had two independent data sets for Shanghai, independent in the sense they occurred on different years (2014 and 2015). For 2015 data set, we selected only those trips starting and ending in the subcity "Yangpu", and hereafter consider it a separate city. The data sets were collected from various sources. Those from Beijing, Changsha, and Hangszhou were provided by a third-party organization that collected driving data from taxi operation companies. The Shanghai data sets were provided by the "1st Shanghai Open Data Apps 2015" (an annual competition). The New York data set has been obtained from the New York Taxi and Limousine Commission for the year 2011 via a Freedom of Information Act request. The Vienna and Singapore data sets were provided to the MIT SENSEable City Lab by AIT and the Singapore government, respectively. The San Francisco and Chicago data sets were publicly available (9), (10). Note the NYC, Vienna, San Francisco, and Singapore data sets were the same as used in previous studies (1), (11).

The four data sets from Chinese cities were very large ($\sim$ GB worth of data per day). For computational convenience, we therefore subsampled these data sets – and not the data sets from the western cities – by selecting only those trips which occurred in a 20 km box surrounding the city center. Our choices for the city center – which of course are arbitrary since center centers do not have one precise location – have GPS coordinates $(39.9059631, 116.3912480)$, $(28.1979483, 112.9713300)$, $(30.2489634, 120.2052342)$, $(31.2253441, 121.4888922)$. We got these points from OpenStreetMap; they were the default locations returned when the city names were entered. A consequence of our subsampling is that we do not capture the polycentric structure of the Chinese cities, which may bias our results. To investigate this potential bias, we fitted $\langle C \rangle (N_T)$ against data from the full Shanghai data set in Supplementary Figure S17. As seen, the good agreement between our model and data persists, confirming that our subsampling method does not cause bias.

The temporal range of the data sets was not uniform. NYC was the most comprehensive, consisting of a year worth of taxi trips in Manhattan. The remaining data sets were for one week. The sizes of the street networks of each city were also different. We demonstrate this in Figures S6(a) and S7(a) by showing N_S, the number of scannable segments, for each city over the course of a week. As discussed in the main text, we use the qualifier 'scannable' in our definition of N_S because some segments are never traversed by taxis in our data sets and so are permanently out of reach of taxi-based sensing. Thus $N_S < N_{S,total}$, where $N_{S,total}$ is the total number of street segments in the street network. Given that city boundaries are ill-defined, a principled way to measure $N_{S,total}$ is difficult. Our strategy to approximate $N_{S,total}$ was to find the number of nodes in the smallest street network which contained as a subset all the segments scanned at least once in our data sets. Supplementary Table S1 shows results for $N_{S,total}$ and N_S using this approach.

Each data set consists of a set of taxis trips. The representation of these trips differs by data set. For the Chinese cities, a trip is the set of GPS coordinates of the taxi's position as its serves its passenger. Since in our model we represent cities by street networks, we convert the set of GPS coordinates to a trajectory T_r, defined in the main text as a sequence of street segments $T_r = (S_{i_1}, S_{i_2}, \dots)$. We matched the taxi trajectories to OpenStreetMap (driving networks) following the idea proposed in (12) which uses a Hidden Markov Model to find the most likely road path given a sequence of GPS points. The HMM algorithm overcomes the potential mistakes raised by nearest road matching, and is robust when GPS points are sparse.

For the remaining data sets, each trip i is represented by a GPS coordinate of pickup location O_i and dropoff location D_i (as well as the pickup times and dropoff times). As for the Chinese cities, we snap these GPS coordinates to the nearest street segments using OpenStreetMap. We do not however have details on the trajectory of each taxi – that is, on the intermediary path taken by the taxi when brining the passenger from O_i to D_i. So we need to approximate trajectories. We used two methods for this, one sophisticated, one simple. The sophisticated method was for the Manhattan data set. Here, as was done in (1), we generated 24 travel time matrices, one for each hour of the day. An element of the matrix (i, j) contains the travel time from intersection i to intersection j. Given these matrices, for a particular starting time of the trip, you pick the right matrix for travel time estimation, and compute the shortest time route between origin and destination; that gives an estimation of the trajectory taken for the trip. For the remaining cities, we used the simple method of finding the weighted shortest path between O_i and D_i (where segments were weighted by their length). As shown in the main text, in spite of the different representations of trajectories, the sensing properties of the taxi fleets from each city are very similar. This gives us confidence in the accuracy of our 'unsophisticated' method.

Lastly, for five of the nine cities – the Chinese cities plus NYC — taxi trips are recorded with the ID of the taxi which completed that trip. Hence for these 'vehicle-level' data sets we can calculate $\langle C \rangle_{N_V}$ – the sensing potential of a fleet as a function of the number of constituent vehicles N_V directly. For the remaining cities, it is unknown which taxis completed which trips. Hence for these 'trip-level' data sets, we can solve only for $\langle C \rangle_{N_T}$. Hence we hereafter divide our data sets into these two categories – 'vehicle-level' and 'trip-level' – and use these terms throughout the paper. For the sake of comparison, we decided to consider NYC and Yangpu part of the trip-level data sets. That way, three different representation of trajectories feature in the trip levels data sets, giving more confidence in the results produced from those data sets.

Supplementary Table 1 summarizes the properties of the data sets.

Estimation of parameters from data sets.

There are three parameters in our model: p_i, the segment popularities, B, the random distance (measured in segments) traveled by a taxi randomly selected from $\mathcal{V}$, and L, the random length of a taxis' trajectory; recall B is needed for the vehicle-level data for which $\langle C \rangle$ depends on the number of vehicles N_V, and L is needed for trip-level data for which $\langle C \rangle$ depends on the number of trips N_T. Supplementary Figures S4 and S5 show the distributions $\mathbb{P}(L)$ and $\mathbb{P}(B)$ for each city on a given day. L can be estimated from our data, but B cannot. This is because our data sets contain taxi *trips* only – a trip implying a passenger is on board – and do not include the distance traveled by taxis when they are empty. Hence, estimating B from our data sets constitutes a *lower* bound for the true B.

Coming back to Supplementary Figures S4 and S5, we see the distributions $\mathbb{P}(L)$ and $\mathbb{P}(B)$ are well fit by lognormals (shown as red curves in the figures). The lognormal fits well in all cases, with exceptions being Chicago, and to a lesser extent, San Francisco (which is contrast to the others appears to be monotonically decreasing). The distribution of taxi trajectory lengths have been studied before (17, 18). Notice however that these works measure a trajectory length in physical distance (i.e. kilometers), whereas our L measures distance in number of segments; that is, L simply counts the *number* of segments traversed by the taxi during the trip and is blind to the segment *lengths*. This definition of L was appropriate for our analysis for which we measured the sensing power in terms of numbers of segments covered $C(N_V) = (N_S)^{-1} \sum_j 1_{M_i > 1}$, and not the total length of road segments covered. Thus, our motivation for estimating L from data was that we needed it to compute the sensing power $\langle C \rangle (N_V)$, and not to study L in and of itself.

Supplementary Figures S6 and S7 shows the parameters of our model and other aspects of our data sets do not vary much on different days of the week (α characterizes the distribution of the segment popularities p_i as will discuss shortly). The low variations in these quantities are encouraging findings because they indicate the behavior of our model (which depends on these quantities) is general, and will not vary significantly on different days of the week.

Consistent with previous findings (13), the segment popularities p_i are long-tailed and appear to be universal, approximately following Zipf's law. To test for universality in p_i we fit each data set to the following heavy tailed distributions

$$\begin{aligned} P_{exponential}(x) &= \lambda e^{-\lambda(x - x_{min})} \\ P_{power\ law}(x) &= (\alpha - 1) x_{min}^{\alpha-1} x^{-\alpha} \\ P_{log\ normal}(x) &= x^{-1} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right) \\ P_{stretched\ exponential}(x) &= \beta \lambda x^{\beta-1} e^{-\lambda(x^{\beta} - x_{min}^{\beta})} \\ P_{truncated\ powerlaw}(x) &= \frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha, \lambda x_{min})} x^{-\alpha} e^{-\lambda x}. \end{aligned} \quad [1]$$

We performed the fitting using the python package 'powerlaw'. By default this package determines a minimum value p_{min} below which data are discarded. Since we want to model the full $\mathbb{P}(p)$ (and not just the tail), we set this equal to the minimum value in our data sets. Table 2 shows the results of the fittings. For each city either a truncated power law or stretched exponential was selected as the distribution of best fit. Thus, we only report the best-fit parameters for those two distributions (the parameters are defined by Eq. Eq. (1)). As detailed in documentation of 'powerlaw', parameters of best fit are found by maximum liklihood estimation. We estimated errors in these parameters by bootstrapping: new data sets $(p_i^*)_{i^*=1}^{N_S}$ were drawn uniformly at random from the original data set $(p_i)_{i=1}^{N_s}$ 1000 times, best fit parameters were found for each of these 1000 realizations, the standard deviation of which was taken as the standard error in each parameter. The 'goodness of fit' measure for each distribution is quantified by the KS (kolmogorov-smirnoff) parameter D, defined by

$$D = \max_x \left| CDF_{empirical}(x) - CDF_{theoretical}(x) \right| \quad [2]$$

where smaller D values indicate better fits, and where CDF denotes the cumulative density function. Finally, the likelihood-ratio test was used to compare the distribution of one fit to another. This has two parameters Λ, r. The sign of Λ tells which distribution is more likely to have generated the data (positive means the first, negative means the second), while the r-value gives a measure of the confidence in the value of Λ (the smaller, the more confident). We adopt the convention that $\Lambda > 0$ indicates the stretched exponential is preferred over the truncated power law (and $\Lambda < 0$ indicates the opposite).

As can be seen in Supplementary Table 2, the tests tell us $\mathbb{P}(p)$ of three of cities are best modeled by stretched exponentials, while the others are best modeled by truncated power laws. The values for r were all $< O(10^{-26})$ (and as small as $O(10^{-222})$), so we truncated all values to zero. There are some mild similarities in the best fit parameters, but no evidence of a convincing trend. Hence we conclude that while similar, the segment popularity distributions $\mathbb{P}(p)$ are not strictly universal.

Like $\mathbb{P}(L)$ and $\mathbb{P}(B)$, there is little daily variation in $\mathbb{P}(p)$. We demonstrate this in Supplementary Figure S6(b) and S7(b) where we show the maximum likelihood exponent α of the truncated power law fit measured day-by-day (for clarity, we do not display the β parameter of the stretched exponential, but they show the same trends).

Compare C_{model} and C_{data}. In the main text we compare our expression for $\langle C \rangle$ against data for a given reference period of a day. The empirical $\langle C \rangle$ were found by subsampling the data sets on a given day; random subsets were drawn from a day's worth of trips, and the average fraction of segments covered by those subsets was computed. As mentioned in the main text, we tested the analytic prediction in two ways: using p_i estimated by the stationary distributions of the taxi drive process (dashed line), and also directly from our data sets (thick line). In the latter case we calculated the distribution of p_i for each day of the week (excluding Sunday), then used those to calculate six separate $\langle C \rangle$, the average of which is shown. This way, both temporal fluctuations and the bias of using the same data sets to estimate p_i and the empirical $\langle C \rangle$ (which recall was calculated for a *single* day) was minimized. For both these cases, the parameter $\langle B \rangle$ was estimated from data sets.

Scaling Collapse

We first discuss the vehicle-level data. In the main text we derived

$$\langle C \rangle_{(N_V)} = 1 - \frac{1}{N_S} \sum_{i=1}^{N_S} (1 - p_i)^{\langle B \rangle * N_V}. \qquad [3]$$

which contains the parameters $p_i, \langle B \rangle$, and N_S. Since p_i and N_S specify the distribution of $\mathbb{P}(p)$, and since the distribution $\mathbb{P}(p)$ is approximately universal across cities (see Supplementary Figure S1), we only need to remove the parameter $\langle B \rangle$ from Eq. (3) to make it city independent. Thus, we plot $\langle C \rangle$ versus $N_V / \langle B \rangle$ which gives the city-independent quantity

$$\langle C \rangle_{(N_V / \langle B \rangle)} = 1 - \frac{1}{N_S} \sum_{i=1}^{N_S} (1 - p_i)^{N_V}. \qquad [4]$$

Supplementary Figure S9 shows the fidelity of the collapse varies by day of week. Hangzhou varies the most. This is not surprising, since as shown in Supplementary Figure S7, the Hangzhou data set has the highest temporal variation.

In Supplementary Figure S10 we apply the same procedure to the trip-level data, except now we plot $\langle C \rangle$ versus $N_T / \langle L \rangle$. There no universal scaling collapse, although there are some similarities between the data sets, Chicago, Yangpu, and San Francisco being nearly coincident. The lack of full universal behavior is perhaps due to the inferior quality of the trip-level data sets (recall the trip-level data are inferior because the trajectories are inferred for those data sets).

Sensing power figures

We here give explicit values for N_T^* and N_V^*, the numbers of trips and vehicles needed to cover half of the city's scannable street segments, i.e. the solutions to $\langle C \rangle(N_T^*) = 0.5$ and $\langle C \rangle(N_V^*) = 0.5$. We also report the numbers needed to cover 80%, which we define as N_T^{**} and N_V^{**}.

As previously discussed, while we consider Manhattan part of the trip-level data sets, taxi trips are recorded along with taxi IDs. This means we can find N_V^* for this data set (as opposed to only N_T^*). Supplementary Figure S11 shows $N_V^* = 30$ – just 30 random taxis cover half of the street segments. (Note in contrast to the rest of our work, the y-axis in Supplementary Figure S11 expresses the number of segments covered as a percentage of total number of segments $N_{S,total}$ and not the number of scannable street segments N_S.) Even more remarkably, over one third of the street segments N_S are scanned by just ten random taxis, and tell us the sensing power of New York taxis is very large.

Minimum street sampling problem

In the main text we quantified the sensing power of a vehicle fleet by their covering fraction $\langle C \rangle_{(N_V, m)}$, the average number of segments covered m times when N_V randomly selected vehicles were equipped with a sensor. Notice that in this definition the independent variable was the number of vehicles N_V. In some contexts, it might be advantageous to know the reverse scenario, in which the independent variable is C; that is, given a target coverage $\bar{C}$, to know how many vehicles are needed to ensure this target coverage is attained (with a given threshold probability guarantee $\bar{p}$). We call this the "minimum street sampling" problem. The minimum street sampling problem is similar in spirit to the classic "location set covering problems" from spatial optimization (14, 15) where the goal is to distribute 'facilities' on a network such that the network is optimally covered. By covered, we mean each node is within a certain distance of each facility. The difference between those works and ours is that our 'facilities' are non-stationary: the sensor-bearing taxis move around on the network.

(MINIMUM STREET SAMPLING): Given a street network S, an reference period $\mathcal{T}$, a minimum sampling requirement m for each street segment, and a collection $\mathcal{V}$ of vehicles moving in S during $\mathcal{T}$ where vehicle trajectories are taken from $\mathcal{P}$ according to a given probability distribution $\mathbf{P}$; what is the minimum number N_V^* of vehicles randomly selected from $\mathcal{V}$ such that $\mathbb{P}(C(N_V, m) \geq \bar{C}) \geq \bar{p}$, where $0 < \bar{C} \leq 1$ is the target street coverage and $\bar{p}$ is a target probabilistic sampling guarantee? The minimum street sampling problem is harder to solve that the 'sensing potential of a fleet' problem. This is because it requires the survival function of the multinomial distribution $\mathbb{P}_{N_T}(M_1 \geq m_1, M_2 \geq m_2 \ldots,)$, which to our knowledge has no known closed form. We here adapt a technique used in (8) to derive an excellent approximation to this survival function.

Approximation of survival function. The probability density function for the multinomial distribution is

$$\mathbb{P}_{N_B}(M_1 = m_1, M_2 = m_2, \ldots) = \frac{N_T!}{m_1! \ldots m_{N_S}!} \prod_{k}^{N_S} p_k^{m_k} \qquad [5]$$

where N_B is the number of balls which have been dropped, N_S is the number of bins, M_i is the random number of balls in bin i, and p_i is the probability of selecting bin i. We seek the survival function

$$\mathbb{P}_{N_B}(M_1 \geq m_1, M_2 \geq m_2, \ldots). \qquad [6]$$

The idea is to represent each M_i as an independent Poisson random variable, conditional on their sum being fixed (this is a well known identity between the Multinomial and Poisson distributions). First let A_i be the event $X_i \geq m_i$, where $X_i \sim Poi(sp_i)$, where s is a real number (we will explain its significance later). Using Bayes' Theorem, we express the survival function as

$$\mathbb{P}_{N_B}\Big(A_1, \dots, A_{N_S} | \sum_{i=1}^{N_S} X_i = N_B\Big) = \frac{\mathbb{P}(A_1 \dots, A_{N_S})}{\mathbb{P}(\sum_{i=1}^{N_S} X_i = N_B)} \mathbb{P}\Big(\sum_{i=1}^{N_S} X_i = N_B | A_1, \dots, A_{N_S}\Big). \qquad [7]$$

The numerator in the first term is easily found, since the events A_i are independent Poisson random variables. Recalling that if $X_i \sim Poi(\lambda_i)$ then $\mathbb{P}(X_i \geq m_i) = 1 - \Gamma(m_i, \lambda_i)/\Gamma(m_i)$, where $\Gamma(n, x) = \int_x^\infty t^{n-1} e^{-t} dt$ is the upper incomplete gamma function, we find

$$\mathbb{P}(A_1 \dots, A_{N_S}) = \prod_{i=1}^{N_S} \left(1 - \frac{\Gamma(m_i, sp_i)}{\Gamma(m_i)}\right). \qquad [8]$$

The denominator is also easy to find. Since $X_i \sim Poi(sp_i)$ and $\sum_i p_i = 1$, we see $\sum_i X_i \sim Poi(s)$ (sums of Poisson random variables are also Poisson distributed). Then

$$\mathbb{P}\Big(\sum_{i=1}^{N_S} X_i = N_B\Big) = \frac{s_B^N e^{-s}}{N_B!}. \qquad [9]$$

For the second term in Eq. (7), we note that conditioning on the joint event $A_1, A_2, \dots$ means the range of the summands are constrained to $[a_i, \infty]$. Hence the summands, which we call Y_i, are truncated Poisson random variables, which we denote by $Y_i \sim Poi_{[a_i, \infty]}(sp_i)$. We note that the mean of a truncated Poisson random variable is not the same as an untruncated one. In particular, if $W_i \sim Poi_{[a, \infty]}(\lambda)$, then

$$\mathbb{E}(W_i) = \lambda \frac{q_{a-1}}{q_a} \qquad [10]$$

$$Var(W_i) = \lambda^2 \frac{q_{a-2} q_a - q_{a-1}^2}{q_a^2} + \lambda \frac{q_{a-1}}{q_a} \qquad [11]$$

where

$$q_a = \begin{cases} 1 - \frac{\Gamma(a, \lambda)}{\Gamma(a)} & a \geq 1 \\ 0 & a < 1. \end{cases} \qquad [12]$$

Returning to the second term in Eq. (7), we find

$$\mathbb{P}\Big(\sum_{i=1}^{N_S} X_i = N_B | A_1, \dots, A_{N_S}\Big) = \mathbb{P}\Big(\sum_{i=1}^{N_S} X_i = N_B | A_1, \dots, A_{N_S}\Big) = \mathbb{P}\Big(\sum_{i=1}^{N_S} Poi_{[a_i, \infty]}(sp_i) = N_B\Big). \qquad [13]$$

We were unable to find an analytic form for the above sum. Instead, we used a first order normal approximation. This states that for a sequence of random variables $(W_i)_i$ with mean μ_i and variance σ_i^2,

$$\sum_{i=1}^{N_S} W_i \xrightarrow{d} N(s_\mu, s_\sigma) \qquad [14]$$

as $N_s \to \infty$, where

$$s_\mu = \sum_i \mu_i \qquad [15]$$

$$s_\sigma^2 = \sum_i \sigma_i^2. \qquad [16]$$

Then the term becomes

$$\mathbb{P}\Big(\sum_{i=1}^{N_S} X_i = N_B | A_1, \dots, A_{N_S}\Big) = \frac{1}{\sqrt{2\pi} s_\sigma} e^{-\frac{(N_B - s_\mu)^2}{2 s_\sigma^2}}. \qquad [17]$$

Pulling all this together gives

$$\mathbb{P}_{N_B}(M_1 \geq m_1, M_2 \geq m_2, \dots) \approx \frac{N_B!}{s^{N_T} e^{-s}} \frac{1}{\sqrt{2\pi} s_\sigma} e^{-\frac{(N_B - s_\mu)^2}{2 s_\sigma^2}} \prod_{i=1}^{N_S} \left(1 - \frac{\Gamma(m_i, sp_i)}{\Gamma(m_i)}\right). \qquad [18]$$

Now, the variable s is a free parameter. Determining the optimal s is an open problem. Following (8) we use $s = N_T$, which, when inserted into Eq. (18), along with Stirling's approximation $\frac{N_T!}{N_T^{N_T} e^{-N_T}} \approx \sqrt{2\pi N_T}$, yields our final expression

$$\mathbb{P}_{N_B}(M_1 \geq m_1, M_2 \geq m_2, \dots) \approx \sqrt{\frac{N_B}{s_\sigma^2}}\, e^{-\frac{(N_B - s_\mu)^2}{2 s_\sigma^2}} \prod_{i=1}^{N_S} \left(1 - \frac{\Gamma(m_i, N_B p_i)}{\Gamma(m_i)}\right). \tag{19}$$

To test the accuracy of the above approximation to the survival function we compared it to Monte Carlo estimates of this survival function. Supplementary Figure S3 shows the approximation is excellent.

Solve minimum street sampling. We leverage the survival function Eq. (19) to solve the minimum street sampling problem in the same way as we did to solve for C in the main text: we assume placing N_T trajectories of random length L into N_S bins is the same as placing $L * N_T$ balls into N_S bins,

$$\mathbb{P}_{(N_T, L)}(M_1 \geq m_1, \dots) = \sum_{n=0}^{\infty} \mathbb{P}_{(N_T, L=1)}(M_1 \geq m_1, \dots) \mathbb{P}(S_{N_T} = n) \tag{20}$$

where $\mathbb{P}_{(N_T, L=1)}(M_1 \geq m_1, \dots)$ is given by equation Eq. (19). As for the expression for C, this can be extended to the vehicle level by replacing L by B. Also as in the main text, this sum is well dominated by its average, leading to the simpler expression

$$\mathbb{P}_{(N_T, L)}(M_1 \geq m_1, \dots) = \mathbb{P}_{(\langle L \rangle * N_T, L=1)}(M_1 \geq m_1, \dots) \tag{21}$$

$$= \sqrt{\frac{\langle L \rangle N_T}{s_\sigma^2}}\, e^{-\frac{(N_B - s_\mu)^2}{2 s_\sigma^2}} \prod_{i=1}^{N_S} \left(1 - \frac{\Gamma(m_i, \langle L \rangle N_T p_i)}{\Gamma(m_i)}\right). \tag{22}$$

When full coverage $\bar{C} = 1$ is desired, equation Eq. (22) solves the minimum street sampling problem. However, when less than full coverage $C < 1$ is desired, we must marginalize over all combinations of $N_S * C$ segments above threshold. This is because in our formulation of the minimum street sampling problem we require just a bare fraction $\bar{C}$ of segments be covered, which is achievable by a large number of combinations of segments. Of course if *targeted* coverage were desired (i.e were specific street segments were desired to be senses with specific sensing requirements m), then Eq. (22) could be used. Staying within our current formulation however, an enumeration of all CN_S combinations of bins is required to marginalize $\mathbb{P}(M_1 \geq m_1, \dots)$. For large N_S enumerating these combinations is infeasible. To avoid this combintorial gallimaufry, we instead estimate $\mathbb{P}_{(\langle L \rangle * N_T, L=1)}(M_1 \geq m_1, \dots)$ by Monte Carlo; we draw samples of size $\langle L \rangle * N_T$ from a multinomial distribution 1000 times, and count the fraction of times at least $\bar{C}$ of the N_S bins are above the threshold m. This lets us estimate $\mathbb{P}(C > \bar{C})(N_T)$, from which we can read off the desired $N_T^*(\bar{P})$ solving the minimum street sampling problem.

Supplementary Figure S8 compares our predictions versus data for a target coverage of $\bar{C} = 0.5$. While the precise shapes of the theoretical and empirical curves do not agree, our model correctly captures the right range of variation: the $P(N_T)$ jumps to $P \approx 1$ at nearly the same N_T. In particular, the error $N_{T,model}(\bar{P} \approx 1)$ - $N_{T,data}(\bar{P} \approx 1)$ is ≈ 200. Expressed relative to the total number of trips, this is $\sim 10^{-4}$ for the NYC and Singapore data sets, and $\sim 10^{-2}$ for the other data sets which is good accuracy.

Sensing power at finer temporal resolutions

Here we extend our analysis of the sensing power to include temporal resolutions finer than $\mathcal{T} = 1$ day. We divide $\mathcal{T}$ into N_w windows of equal size, and define the adjusted sensing power $C^*(N_T, N_w)$ as the normalized number of segments that are covered at least once in each of the N_w windows. Let $M_i^\mu(N_T)$ be the number of times segment i is covered during window μ, when N_T trips have been randomly selected from $\mathcal{P}$, where $\mathcal{P}$ is the population of trips that occur in $\mathcal{T}$. As before, we derive $\langle C^* \rangle$ for the trip-level data first which is then easily generalized to the vehicle-level data. C^* is given by

$$C^*(N_T, N_w) = \frac{1}{N_S} \sum_{j=1}^{N_S} 1\Big(M_i^{\mu=1} \geq 1, M_i^{\mu=2} \geq 1, \dots, M_i^{\mu=N_w}\Big). \tag{23}$$

where 1 is the indicator variable. We approximate C^* by adapting our ball-in-bin analysis. Instead of adding indistinguishable balls into bins, we imagine balls come in N_w different colors (i.e. a different color for each time window N_w). We assume balls of different colors have the same probability of being chosen to be put into bins. Computing C^* then becomes equivalent to asking how many N_B balls need be drawn until $N_S * C^*$ bins have at least 1 ball of each color in them. Switching to numbers of trips N_T, which we recall is equivalent to adding a random number of balls (see methods section in the main text), the probability of this event is

$$\mathbb{P}_{N_T}(M_i^1 \geq, \bar{m}, \dots, M_i^{N_w} \geq 1) \tag{24}$$

We assume that the events $M_i^\mu \geq 1$ are independent

$$\mathbb{P}_{N_T}(M_i^1 \geq 1, \ldots, M_i^{N_w} \geq 1) = \prod_{\mu=1}^{N_w} \mathbb{P}_{N_T}(M_i^\mu \geq 1). \qquad [25]$$

Following our previous analysis we decompose the prodand

$$\mathbb{P}_{N_T}(M_i^\mu \geq 1) = \sum_{n}^{\infty} \mathbb{P}(N_T = n) \times \mathbb{P}_n(M_i^\mu \geq 1) \qquad [26]$$

which follows from the fact that trips have random length. Recall from the main text $\mathbb{P}(N_T = n) \sim Bin(N_T, p)$. Also from the main text, recall we assumed the sum in (26) is dominated by its average, so we collapse it and replace n by its average value $N_T \langle L \rangle$. Plugging this, along with the survival function for the binomial function, into the equation above gives

$$\mathbb{P}_{N_T}(M_i^1 \geq 1, \ldots, M_i^{N_w} \geq 1) = \prod_{\mu=1}^{N_w} 1 - (1 - p_i^\mu)^{\langle L \rangle N_B / N_w}. \qquad [27]$$

Finally, applying the expectation operator to Eq. (23), and plugging in the expression above gives

$$\langle C^* \rangle(N_T, N_w) = \frac{1}{N_S} \sum_{i=1}^{N_S} \mathbb{P}_{N_T}(M_i^1 \geq 1, \ldots, M_i^{N_w} \geq 1) \qquad [28]$$

$$= \frac{1}{N_S} \sum_{i=1}^{N_S} \prod_{\mu=1}^{N_w} \left[1 - (1 - p_i^\mu)^{\langle L \rangle \frac{N_T}{N_w}} \right] \qquad [29]$$

where p_i^μ are the segment popularities resulting from all trips which occur in window μ; we assume the size of the windows is large enough (or, equivalently, the number of windows N_w in low enough) so that the distribution of segment popularities p_i^μ is approximately stationary – an approximation that will get worse as N_w increases.

Supplementary Figure S12 shows $\langle C^* \rangle(N_T, N_w)$ for $N_w = 3$ and $N_w = 10$ along with the regular sensing power $\langle C \rangle$ for the NYC data set. There is reasonable agreement between data and theory for $N_w = 3$, but this agreement gets worse when the number of windows increases $N_w = 10$, as anticipated (in paragraph above). The figure also shows $\langle C^* \rangle$ decreases with increasing temporal resolution, as measured by increasing N_w. In particular, when $N_w = 3$, which can be thought of as requiring segments to be scanned in the morning, afternoon, and evening – defined by the intervals $(12AM, 8\text{AM}), (8\text{AM}, 4\text{PM}), (4\text{PM}, 12\text{AM})$ – we see $\langle C^* \rangle(\tilde{N_T}, N_w) = 0.33$ yields $\tilde{N_T} \approx 4000$. In terms of vehicles this translates to $\tilde{N_V} \approx 100$, compared to $N_V = 10$ for the regular sensing power – that is, 10 times more vehicles are needed to get the same coverage. Moreover, the equivalent ratio to cover half of the segments is approximately 12. Nevertheless, $N_V = 100$ to cover one-third of street segments still indicates a large sensing power. Table S5 summarizes our results.

Finally, we note the vehicle-level expression $\langle C^* \rangle(N_V, N_w)$ is found by replacing $N_T \to N_B$ and $\langle L \rangle \to \langle B \rangle$ in Eq. (29) as before.

Spatial bias of drive-by sensing

Taxi-based drive-by sensing has an inherent spatial bias because taxis do not spread out homogeneously over a city's areas; instead, taxis concentrate in the 'core' of a city: affluent, commercial, and touristic areas. This 'core-scanning' has both benefits and drawbacks. The benefit is that core-scanning might be useful for certain sensing goals. For simplicity, we have thus far assumed such goals were spatially uniform (mathematically expressed by $M_i \geq 1$ for each i). Certain urban quantities might however have *non-uniform* sensing requirements $M_i \geq m_i$ for some m_i, owing to the quantities non-trivial spatiotemporal character, as well as to the aims of the urban surveillant (certain areas might require greater monitoring than others). We conjecture for some urban quantities, greater scanning is required at the core, since the core has higher rates of pollution and increased infrastructural strain. For instance, if road quality were the urban metric being measured, it seems reasonable to assume that the scanning requirement m_i of a street segment would correlate with its usage (since the latter correlates with its depreciation rate). In the language of our model, $m_i \propto p_i$. Drive-by sensing is almost by definition the optimal choice of sensing strategy for this sensing requirement, since again by definition, it scans segments in proportion to their popularity.

The drawback of drive-by sensing's core-scanning is directly related to its preferential scan of the core; as such, it leaves unpopular, potentially socioeconomically disadvantaged, areas monitored at significantly lower resolutions. This is a serious concern, which could reinforce inequality and have other harmful consequences. In an effort to address the concern, we checked how segment population p_i correlate (spatially) with the median house-hold income w_i. These w_i were found from census data, obtained using SimplyAnalytics (16). Only data for the American cities NYC, Chicago, and San Francisco were available, so we restricted this part of our analyss to just these data sets. Census data was available for the same year as the taxi data for NYC (2011) and Chicago (2014), but not for San Francisco; here the taxi data was from 2008, but the closest census data was 2010, so we used those data. Supplementary Figure S13 shows a spatial plot of the w_i obtained from SimplyAnalytics and

Supplementary Figure S14 shows scatter plots of (w_i, p_i); note, because the spatial granularity of the w_i were lower than the p_i, multiple p_i's were mapped to a given w_i. The correlation between p_i and w_i for each city were mild, with Pearson coefficients $r_{NYC} = 0.31, r_{chicago} = 0.22, r_{sanfran} = -0.24$. Strangely, San Francisco showed a *negative* correlation, which we suspect means high income earners in San Francisco live outside of the city center. These low correlations are somewhat encouraging, insofar as they speak to a low sensing-bias to affluent neighbourhoods. However, to be sure of this low-sensing-bias a more detailed study is needed to corroborate our findings, which is a subject of future work.

We also visually inspected the drive-by sensing's spatial bias. Supplementary Figure S18 shows the spatial distribution of covered segments when $N_T = 10\%, 75\%$ for NYC, Chicago, and San Francisco where segments which have been covered are colored green, and those uncovered are colored red. As can be seen, city centers are indeed preferentially covered (although in Manhattan the cover is already quite homogeneous because it doesn't have a city center, per se).

Inferring sensing power from street network

Could the sensing power of a taxi fleet be inferred from the street network S alone? That is, without data characterizing the mobility pattern of the fleet M? In this note we explore this possibility. Beginning with the trip-level data, we see from the expression

$$\langle C \rangle_{N_T} = 1 - \frac{1}{N_S} \sum_i (1 - p_i)^{\langle L \rangle N_T} \qquad [30]$$

that if p_i and $\langle L \rangle$ could be inferred from S, then $\langle C \rangle$ could be inferred in turn. What network quantities could be used to infer p_i and $\langle L \rangle$ from S? We make the conjectures in Supplementary Table S6, that $p_i \approx b_i$, where b_i is the betweenness of node i, and that $\langle L \rangle = \langle l \rangle$, where $\langle l \rangle$ is the average path length of S. The $p_i = b_i$ approximation is an intuitive: if taxi origin and destinations are uniformly distributed, and taxis follow shortest paths, then by definition traffic densities on each edge (and therefore segment popularities) will correspond to edge betweenness. Of course, real taxis do not have uniformly distributed origin and destinations, and do not always follow shortest paths, so we do not expect $p_i = b_i$ to hold exactly. But as a first order approximation $p_i = b_i$ seems reasonable. The $\langle L \rangle = \langle l \rangle$ also follows from these assumptions; shortest paths implies $L = l$, and uniformly distributed origin and destinations imply every path in the graph is sampled, implying $\langle L \rangle = \langle l \rangle$.

Supplementary Figure S15 shows the probability density functions for p_i and b_i are reasonably similar. This similarity is not surprising, since, recall, for some of our simulated trajectories, shortest paths length routing was used. In the caption of this Figure we list $\langle L \rangle$ and $\langle l \rangle$ which are also similar. In Supplementary Figure S16 we plot the inferred sensing power $\langle \tilde{C} \rangle$, defined by Eq. (30) with $p_i \to b_i$ and $\langle L \rangle \to \langle l \rangle$, versus the normal sensing power for the trip-level data. The figure shows the inference is poor. Moreover, the bias between $\langle C \rangle$ and $\langle \tilde{C} \rangle$ is not uniform; for some data sets $\langle C \rangle > \langle \tilde{C} \rangle$ and for others $\langle C \rangle < \langle \tilde{C} \rangle$. We do not compute $\langle \tilde{C} \rangle$ for the vehicle-level data for two reasons. First, inferring the means distance traveled (in segments) by taxis $\langle B \rangle$ from street network quantities is difficult – how far a taxi travels in a day depends the driver's preferences, as well as the (varying) trip demand. Second, given the inference is so poor for the trip-level data, and given the trip-level data shows a better match between theory and data for the sensing power, we expect the vehicle-level inference will be poor too.

We conclude that an accurate inference of the sensing power from the topology of the street network is unlikely.

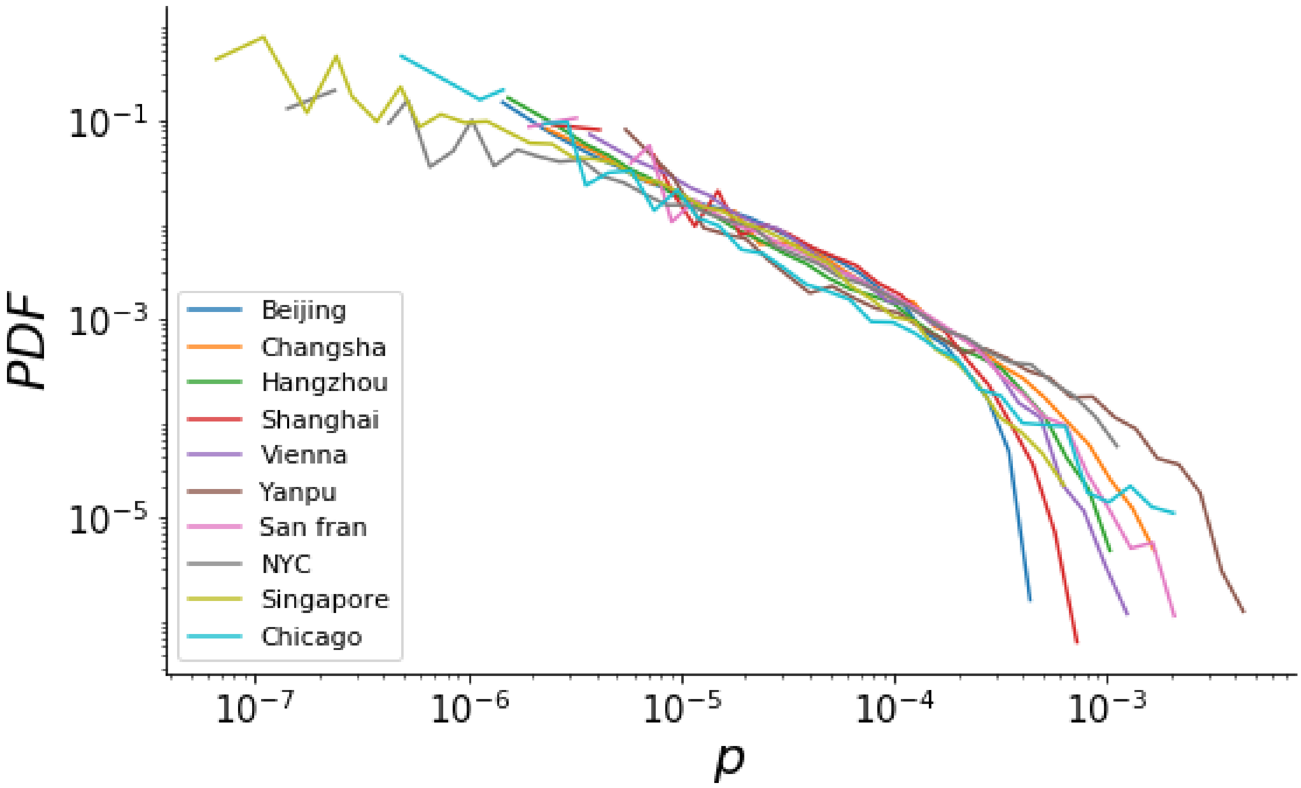

Fig. S1. Empirical segment popularities. Log-log plot of the distributions of segment popularities for each city, showing evidence of universality. See "Estimation of parameters from data sets" section.

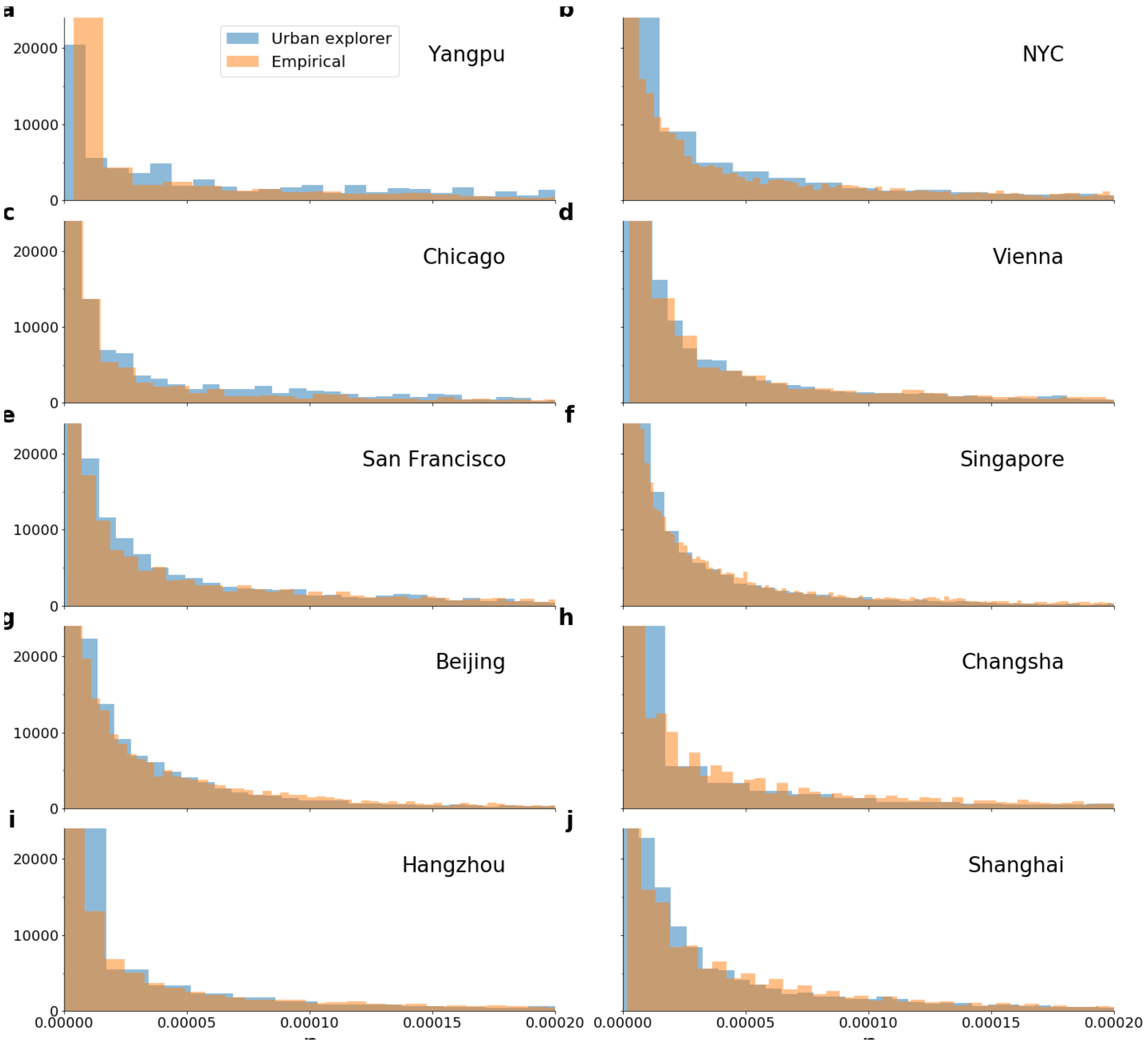

Fig. S2. Taxi drive segment popularities versus data. Segment popularities p_i derived from the taxi drive process (blue) and empirical data sets (orange) for all cities. Simulations were run for 10^7 timesteps after which the distribution of p_i were approximately stationary. To measure the similarity between the taxi-drive p_i and empirical p_i we use the Kolmogorov-smirnov statistic D, defined as $D = \max_x |CDF_1(x) - CDF_2(x)|$ where $CDF_i(x)$ is the empirical cumulative density function for the i-th data set. The D and bias parameter β (for the taxi-drive process) for each city were: (a) $(D, \beta) = (1.8, 2.75)$ (b) $(D, \beta) = (0.07, 1.5)$ (c) $(D, \beta) = (0.1, 3)$ (d) $(D, \beta) = (0.8, 0.25)$ (e) $(D, \beta) = (0.7, 0.25)$ (f) $(D, \beta) = (0.9, 1.0)$ (g) $(D, \beta) = (0.05, 1.0)$ (h) $(D, \beta) = (0.07, 1.75)$ (i) $(D, \beta) = (0.08, 1.25)$ (j)$(D, \beta) = (0.06, 0.75)$.

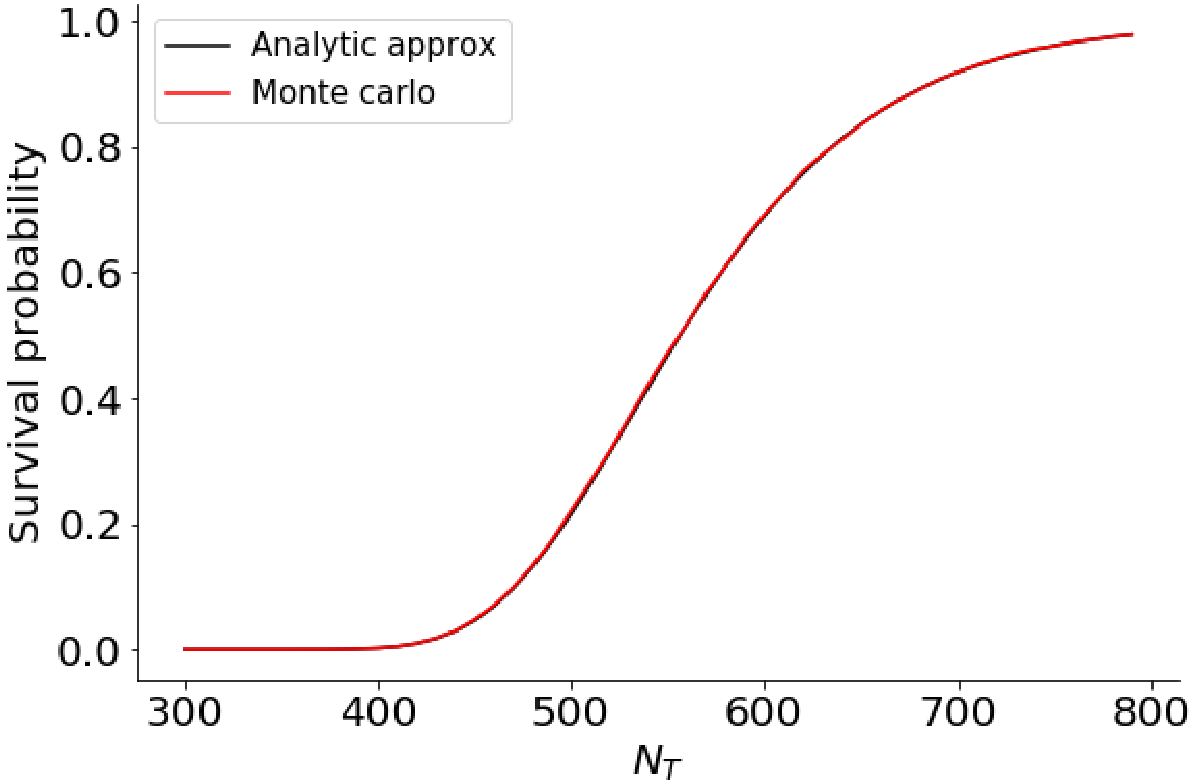

Fig. S3. Approximation of multinomial survival function. Survival probability for multinomial distribution estimated from Eq. (19), and via Monte Carlo. 10^5 trials were used in each Monte Carlo approximation. 50 bins were used, with $p_i = 1/50$. The survival probability is defined as $\mathbb{P}(M_1 > b, M_2 > b, \ldots)$. Here we took $b = 5$. Note the excellent agreement between theory and simulation (both curves lie on top of each other.)

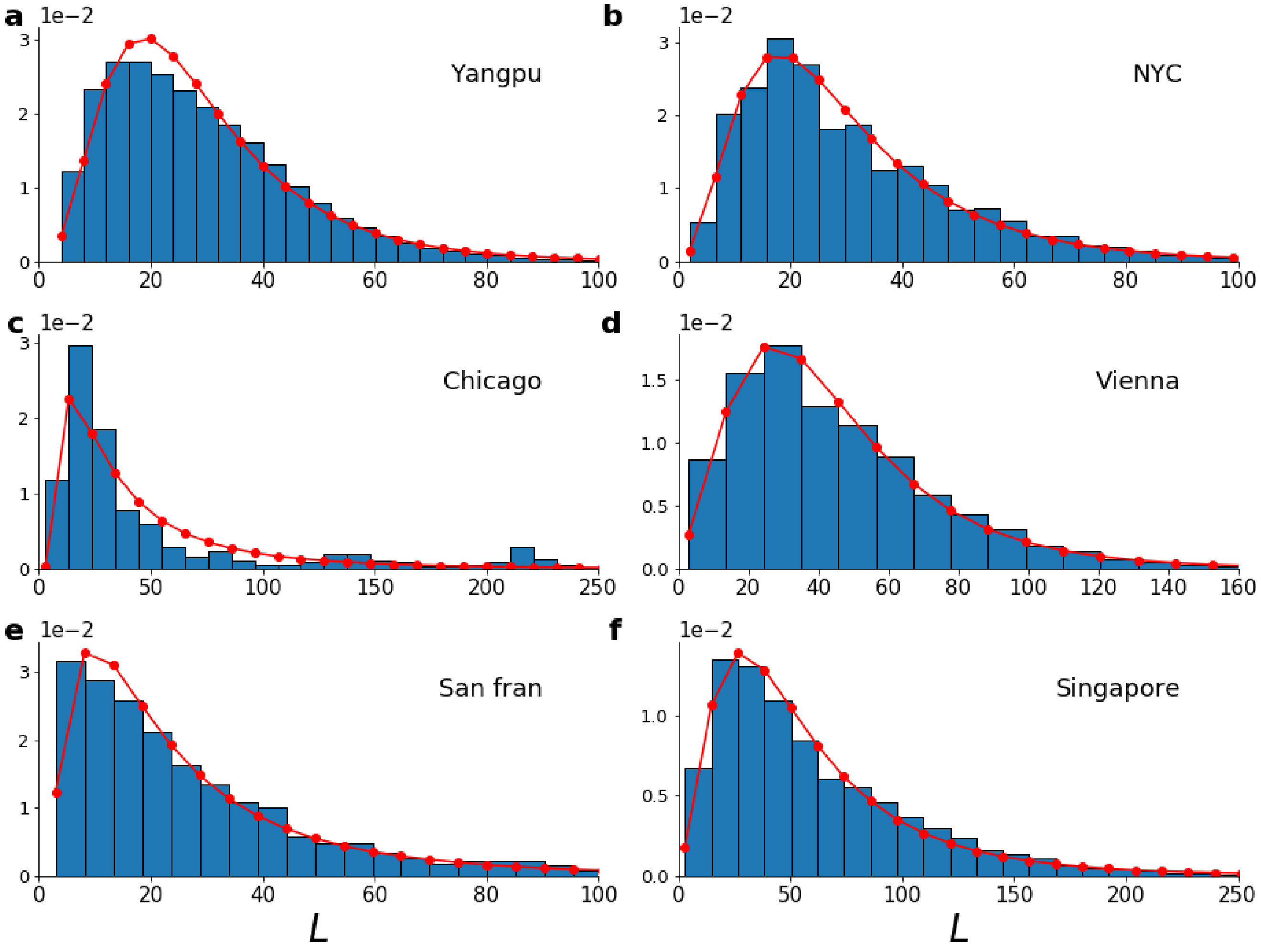

Fig. S4. Distributions of trajectories lengths for the trip-level data sets. Histograms of the trajectory lengths L during a given day for city. Red dotted lines show lognormal curves of best fit. We list the parameters of best fit μ, σ, the sample mean $\langle L \rangle$, and the day the data were taken from for each subplot. Notice Chicago appears to have two humps. Data taken from other days are qualitatively similar. It may be surprising that the L from different cities vary so much; one might expect the average distance traveled by taxis would be the same everywhere. Recall however that L measures the raw number of segments in a taxi trip, and that we assume each segment to have unit length. Hence, given that segments lengths might vary from city to city, we expect L to do so too. (a) Yangpu, 04/02/15, $(\mu, \sigma, \langle L \rangle) = (3.36, 0.52, 29.6)$ (b) NYC, 01/05/11, $(\mu, \sigma, \langle L \rangle) = (3.37, 0.57, 30.8)$ (c) Chicago, 05/21/14, $(\mu, \sigma, \langle L \rangle) = (3.36, 0.98, 51.02)$ (d) Vienna, 03/25/11, $(\mu, \sigma, \langle L \rangle) = (3.91, 0.51, 45.54)$ (e) San Fransisco, 05/24/08, $(\mu, \sigma, \langle L \rangle) = (2.99, 0.87, 28.90)$ (f) Singapore, 02/16/11, $(\mu, \sigma, \langle L \rangle) = (3.97, 0.68, 60.78)$.

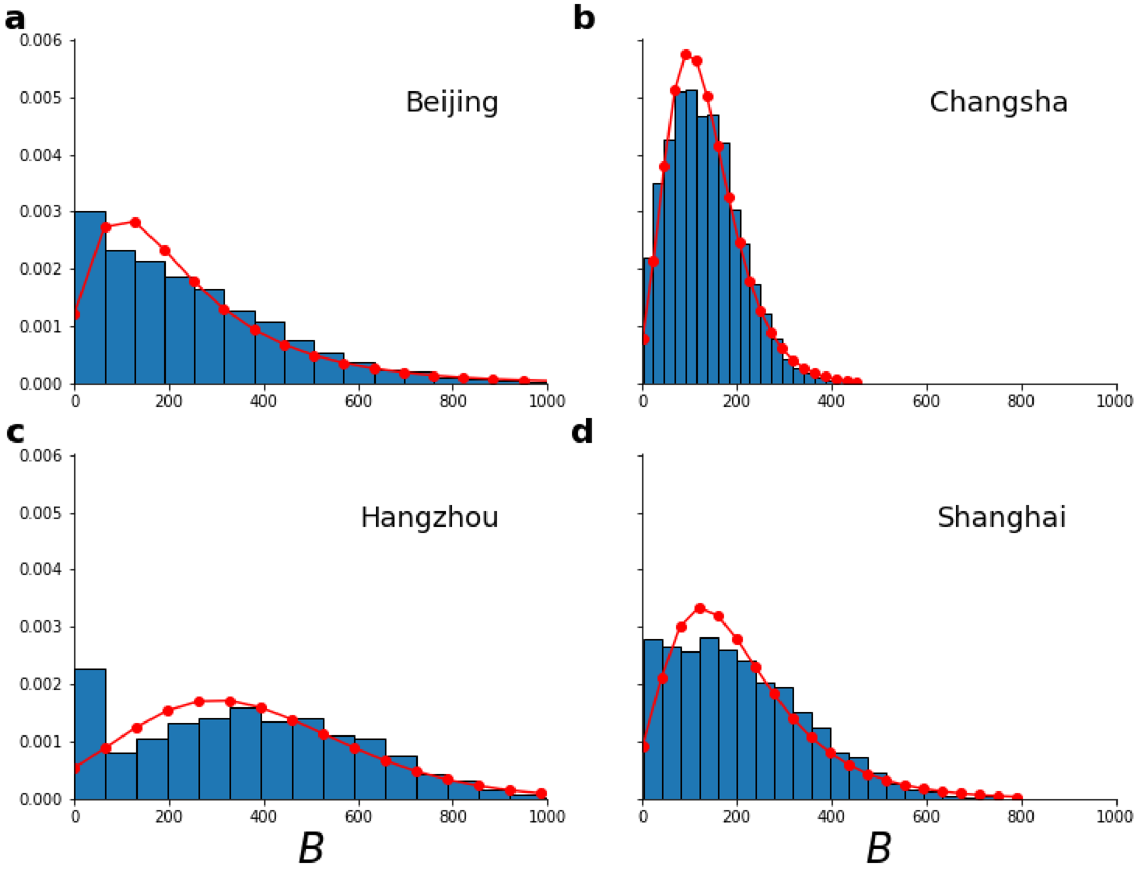

Fig. S5. Distributions of distance traveled by taxis for vehicle-level data sets. Histograms of B, the distance traveled (measured in number of segments) by a taxi in a day for each city. Red dotted lines show lognormal curves of best fit. We list the parameters of best fit μ, σ, the sample mean $\langle B \rangle$, and the day the data were taken from for each subplot (a) Beijing 03/02/13, $(\mu, \sigma, \langle B \rangle) = (5.56, 0.65, 245)$ (b) Changsha 03/02/14 $(\mu, \sigma, \langle B \rangle) = (5.46, 0.31, 131)$ (c) Hangzshou 04/22/14 $(\mu, \sigma, \langle B \rangle) = (5.13, 0.18, 366)$ (d) Shanghai 03/02/14 $(\mu, \sigma, \langle B \rangle) = (5.41, 0.35, 270)$.

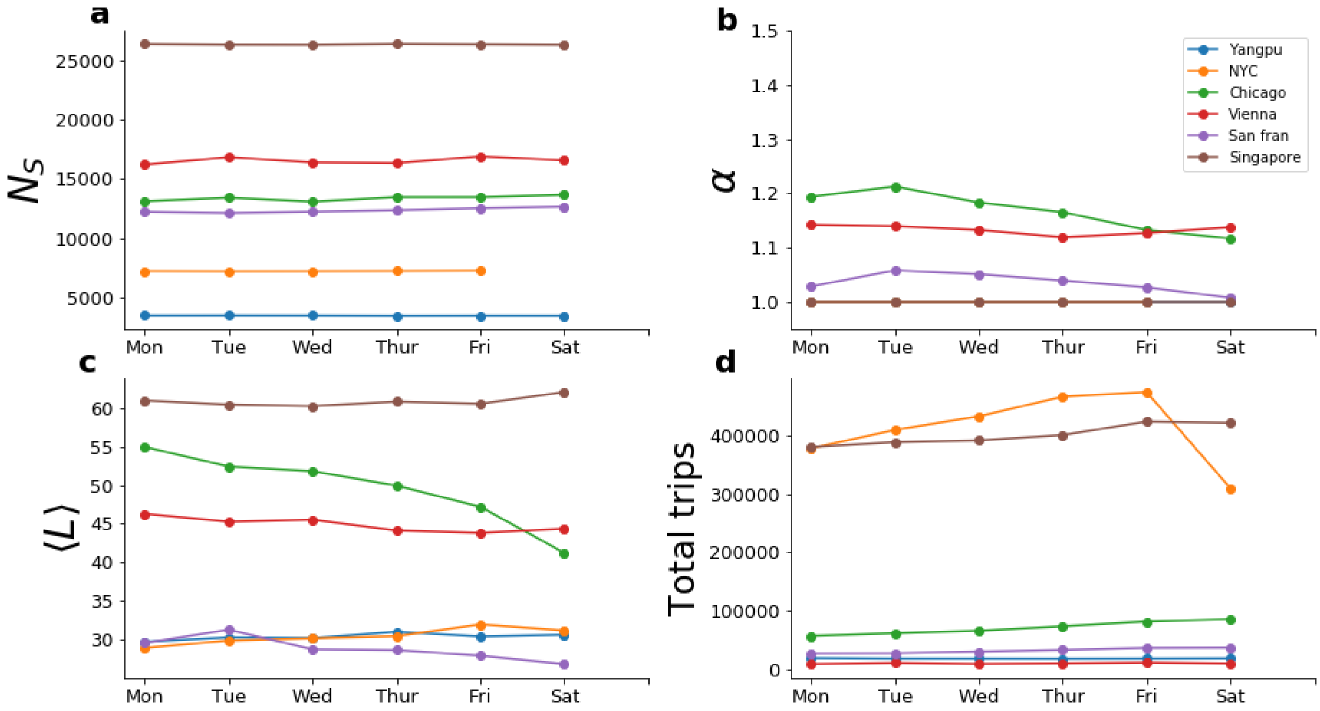

Fig. S6. Temporal fluctuations of trip-level data sets. Generally speaking, there is little daily variation in each quantity. (a) Number of scannable street segments (b) Best fit exponent in truncated power law α. (c) Average length of trajectory (d) Total number of trips.

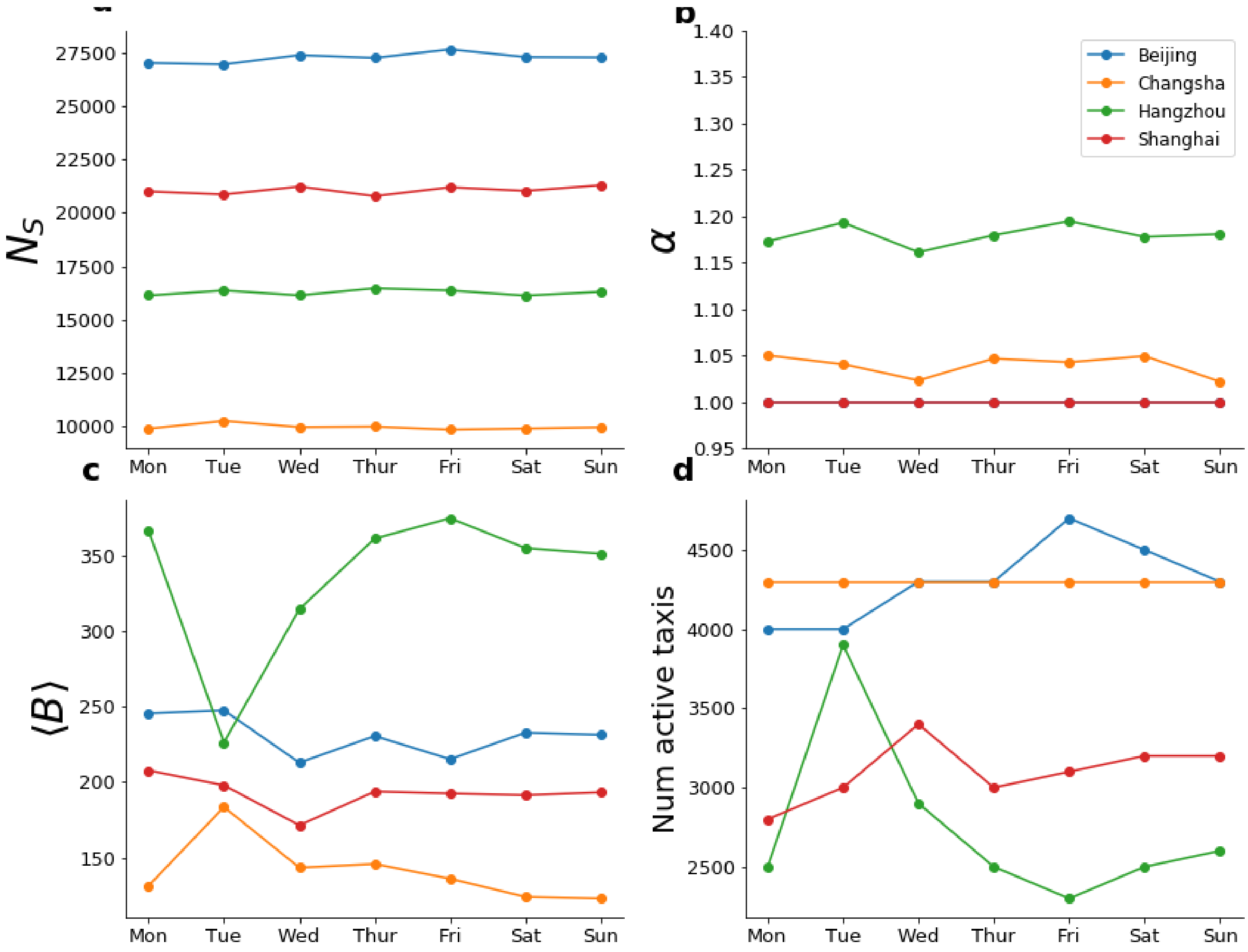

Fig. S7. Temporal fluctuations of vehicle-data. Generally speaking, there is little daily variation in each quantity. (a) Number of scannable street segments (b) Best fit exponent in truncated power law α. (c) Average daily distance traveled by a taxi (measured in number of segments) (d) Number of active taxis.

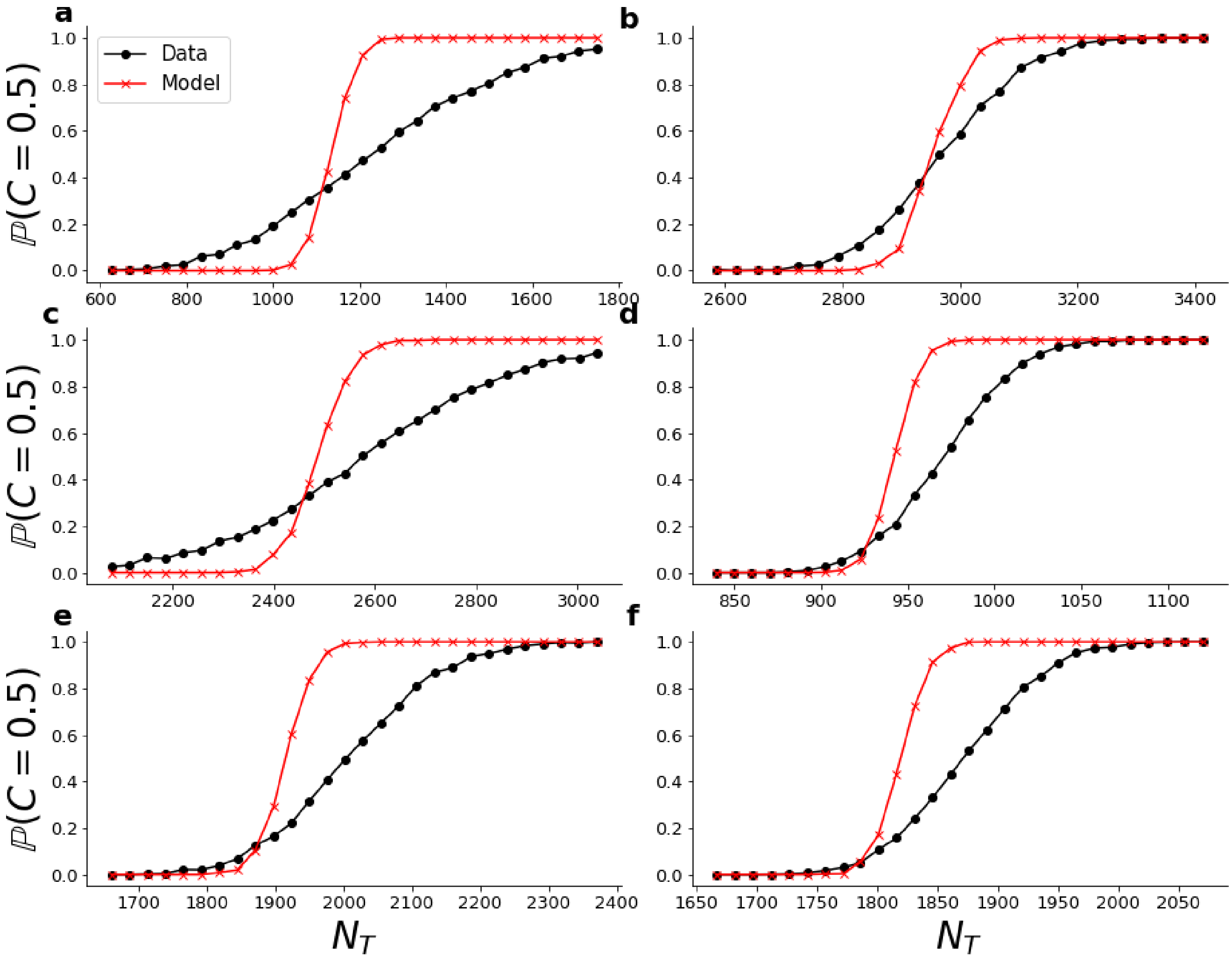

Fig. S8. Minimum street sampling problem. Analytic prediction versus trip-level data. The red curve shows theoretical results, while the black curve shows probabilities estimated from data. The parameters for each subplot were $\bar{C} = 0.5$, $m = 1$. The number of trials used in the Monte Carlo estimate of $\mathbb{P}(C)$ was 1000. (a) Yangpu on 04/02/15 (b) NYC on 01/05/11 (c) Chicago on 05/21/14 (d) Vienna on 03/25/11 (e) San Fransisco on 05/24/08 (f) Singapore on 02/16/11.

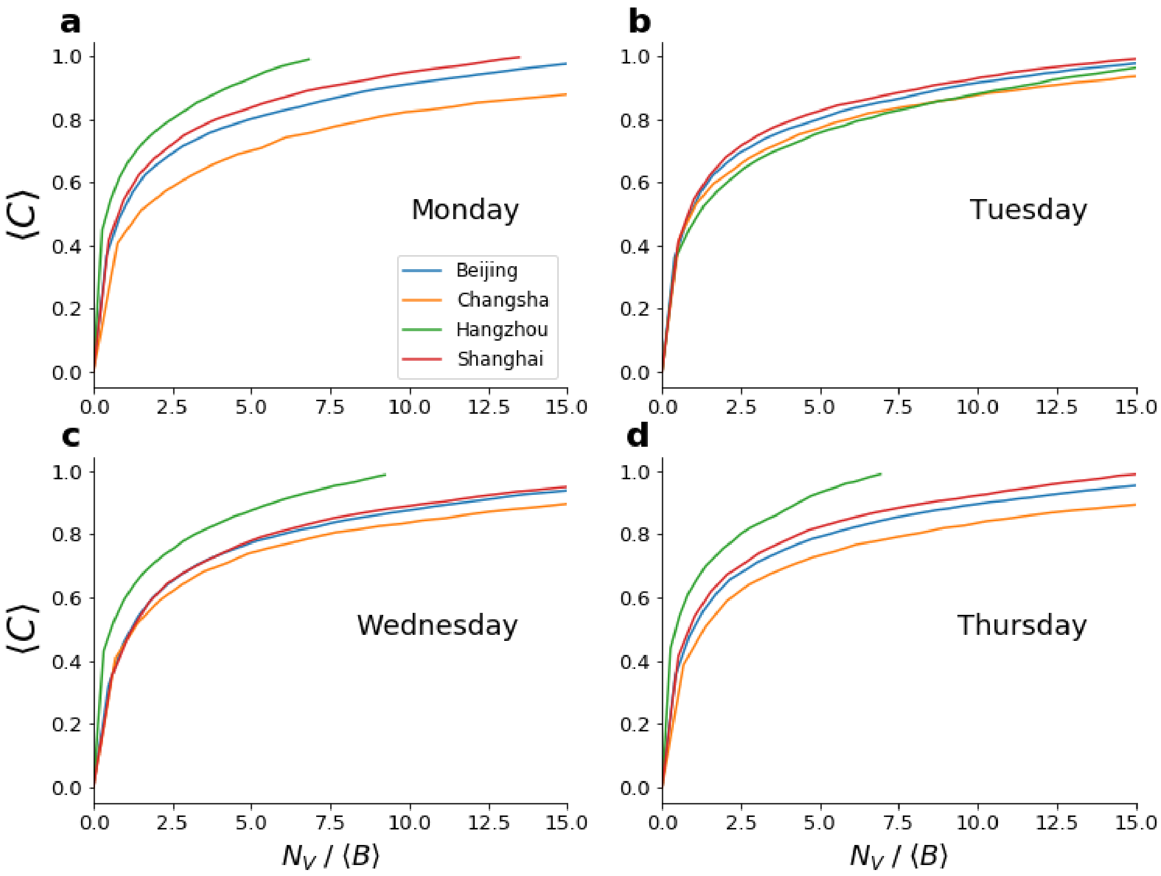

Fig. S9. Scaling collapse of vehicle-level data on different days. Counterpart of Figure 3 in the main text. As can be seen, a close approximation to a true scaling collapse is achieved only on Tuesday. Note the Hangzhou data set has strong variations. This is not surprising, since as shown in Supplementary Figure S7, this data set has strong temporal variations. In particular, $\langle B \rangle$ varies much more than the other data sets.

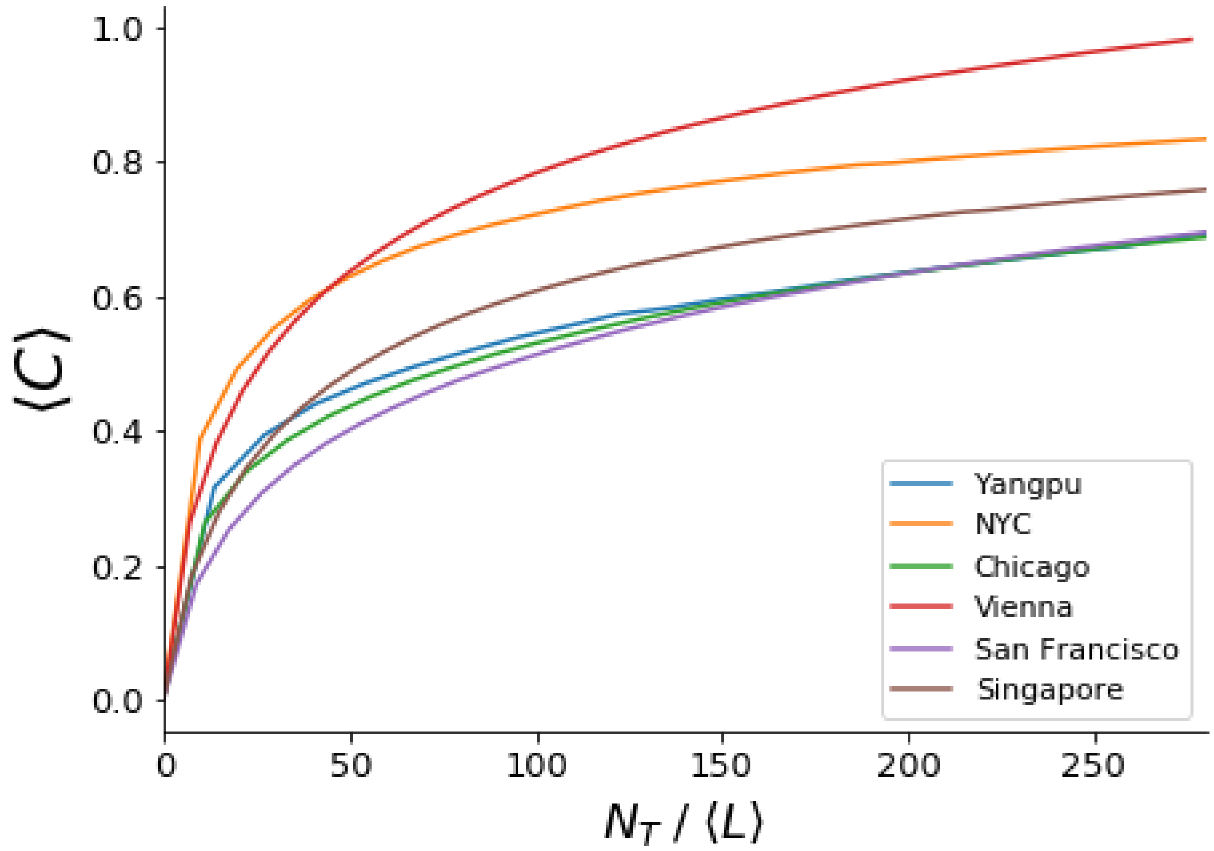

Fig. S10. Scaling collapse of trip-level data. In contrast to vehicle-level data sets – Supplementary Figure S9 – the trip-level data sets do not show universal behavior. There are however some trends. As can be seen the Chicago, San Francsico, and Yangpu data sets collapse to a common curve, where the other data sets do not. The data for each city are the same as those used in Figure 2 (main text). Trip data on different days show the same trends.

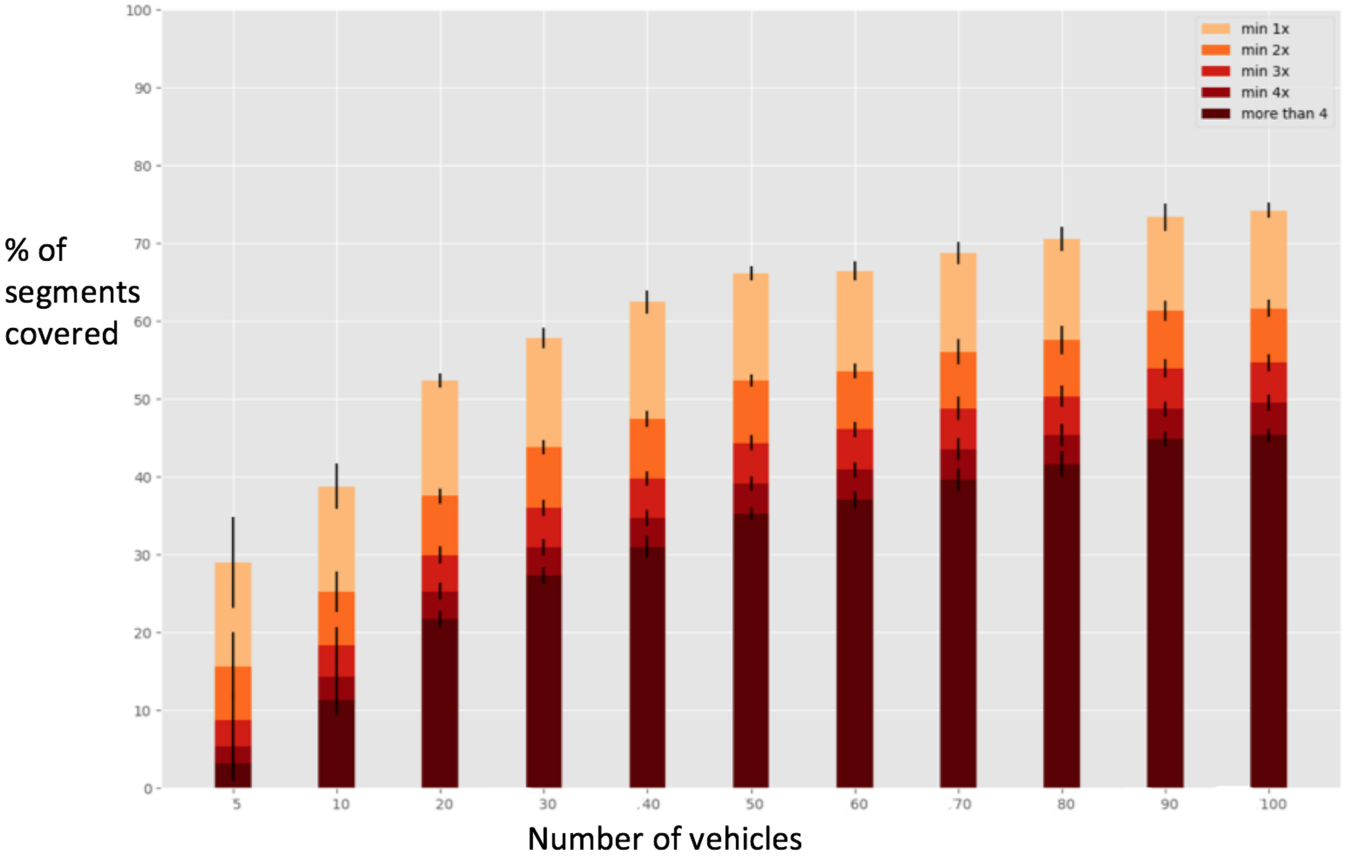

Fig. S11. Average segment coverage versus number of sensor-equipped taxis in Manhattan on 03/08/2011. Different colors show results for different scanning thresholds. That is, the % of segments covered at least m times, where $m = 1, 2, 3, 4$. Black lines show one standard deviation away from mean value. Notice that just 10 vehicles scan more than a third of scannable segments, while 30 scan more than half. See "Sensing power figures" subsection.

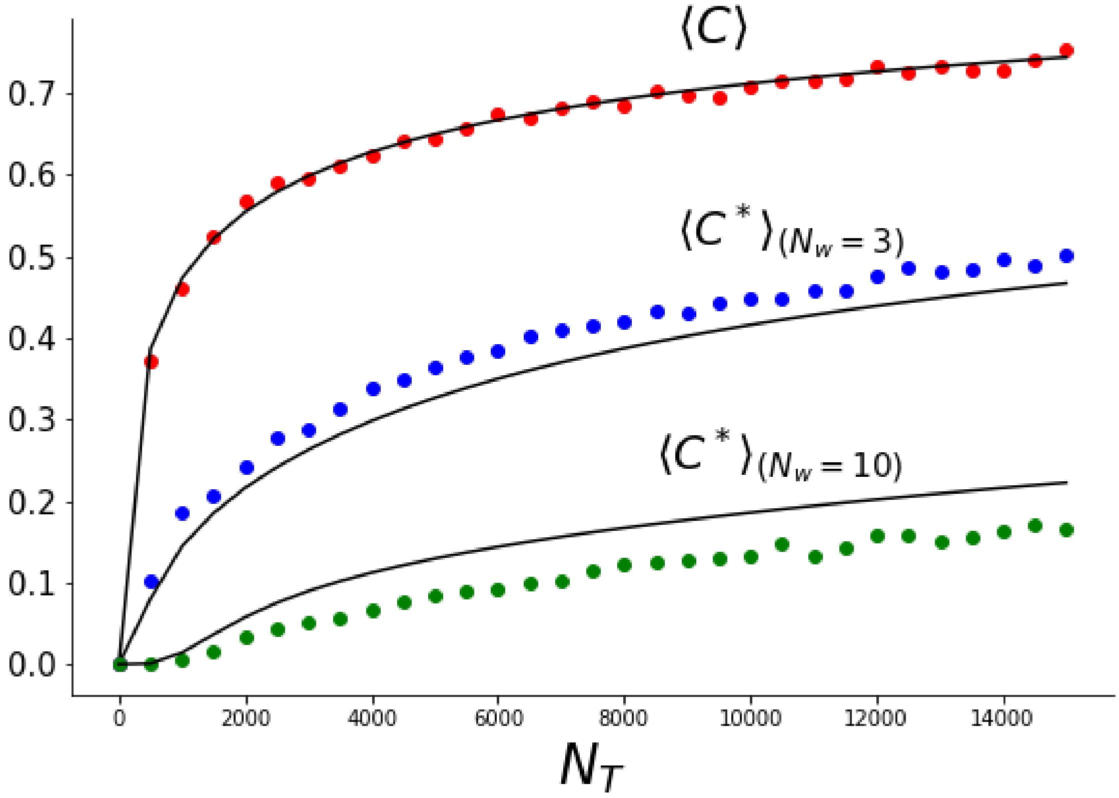

Fig. S12. Sensing power over finer temporal scale. The 'regular' sensing power $\langle C \rangle$ and adjusted sensing power $C^*(N_T, N_w)$ defined by $\mathrm{Eq.}$ (29) for $N_w = 3$ and $N_w = 10$. Note that when $N_w = 1$ $\langle C \rangle$ and $\langle C^* \rangle$ are equivalent. Thick black lines shows analytic predictions, while colored dots show empirical data. Each data point represents the average of 500 trials. As can be seen, requiring segments be sensed at a finer temporal resolution – that is, at least once in each of the N_w windows – significantly reduces their sensing power $\langle C^* \rangle < \langle C \rangle$. Note also the agreement between data and theory also gets worse for $N_w > 1$; see "Sensing power at finer temporal resolutions" section for a discussion on why this happens.

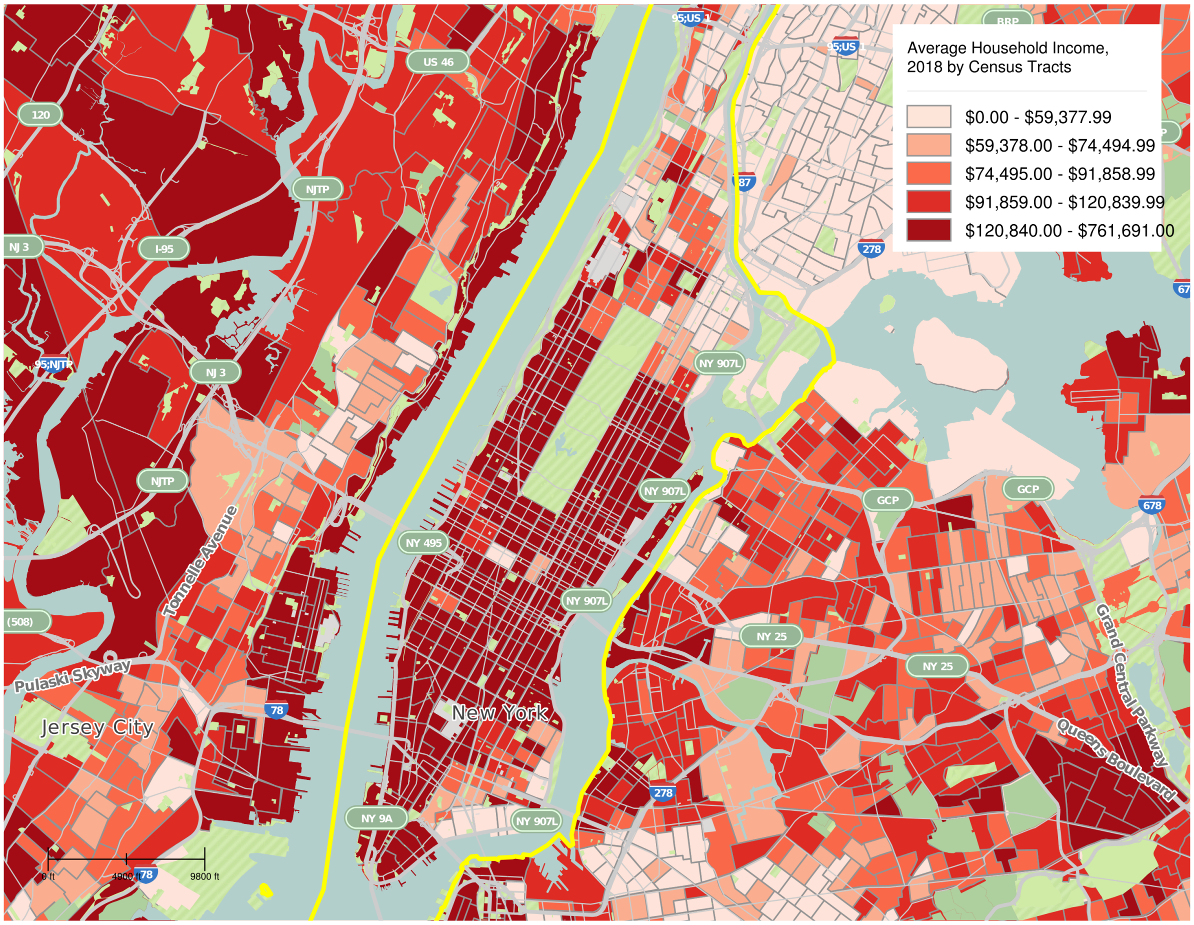

Fig. S13. Manhattan census data. Computed from SimplyAnalytic (16). See "Spatial bias of drive-by sensing" section

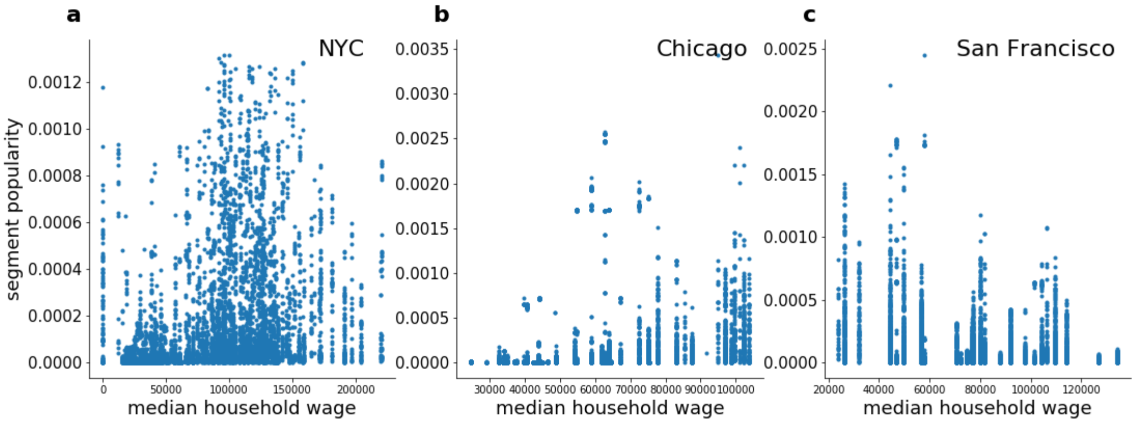

Fig. S14. Spatial correlation of segment popularity and median household income. For each subfigure we list the number of wages brackets N_{wages}, the number of scannable street segments N_S, and the pearson correlation coefficient. See "Spatial bias of drive-by sensing" section for how the data were collected. (a) NYC: $(N_{wages}, N_S, r) = (613, 6882, 0.31)$ (b) Chicago: $(N_{wages}, N_S, r) = (64, 10960, 0.22)$ (c) San Francisco: $(N_{wages}, N_S, r) = (27, 11573, -0.24)$.

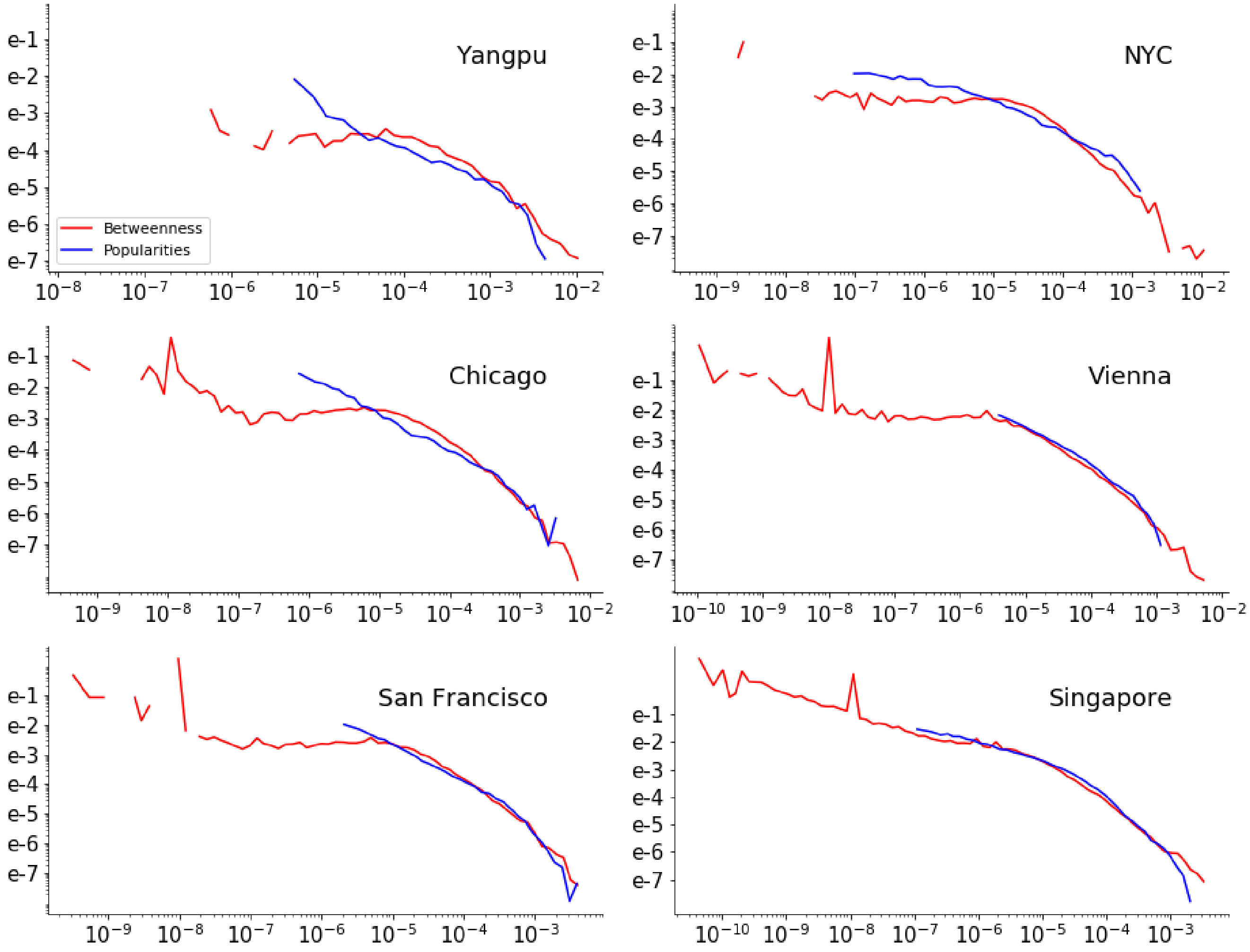

Fig. S15. Betweeness and segment popularities. Probability density functions of street network betweeness b and segment popularities p. The mean path length $\langle l \rangle$ of each graph, mean trip length $\langle L \rangle$, and Pearson correlation coefficient r, for each city are as follows. Note that the high correlation between p_i and b_i is to be expected, since our trajectory generation method relies on shortest-path routing. (a) Yangpu $(\langle l \rangle, \langle L \rangle, r) = (24.2, 29.6, 0.90)$ (b) NYC $(\langle l \rangle, \langle L \rangle) = (34.3, 30.8, 0.65)$ (c) Chicago $(\langle l \rangle, \langle L \rangle) = (51.0, 60.8, 0.95)$ (d) Vienna $(\langle l \rangle, \langle L \rangle, r) = (45.4, 60.9, 0.87)$ (e) San Fransisco $(\langle l \rangle, \langle L \rangle, r) = (28.88, 38.5, 0.92)$ (f) Singapore $(\langle l \rangle, \langle L \rangle, r) = (60.8, 73.8, 0.87)$.

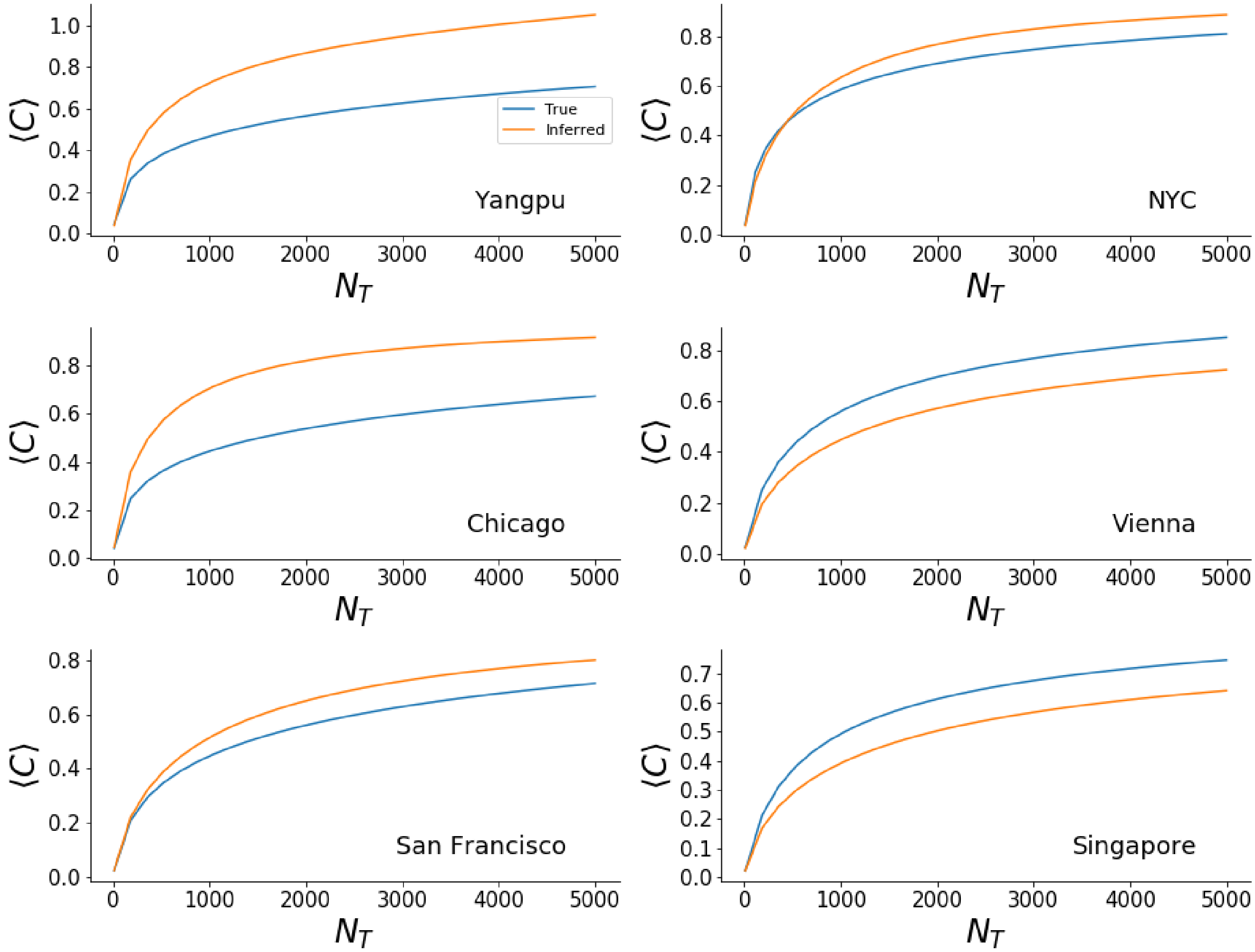

Fig. S16. Inferred sensing power. See "Inferring sensing power from street network" section.

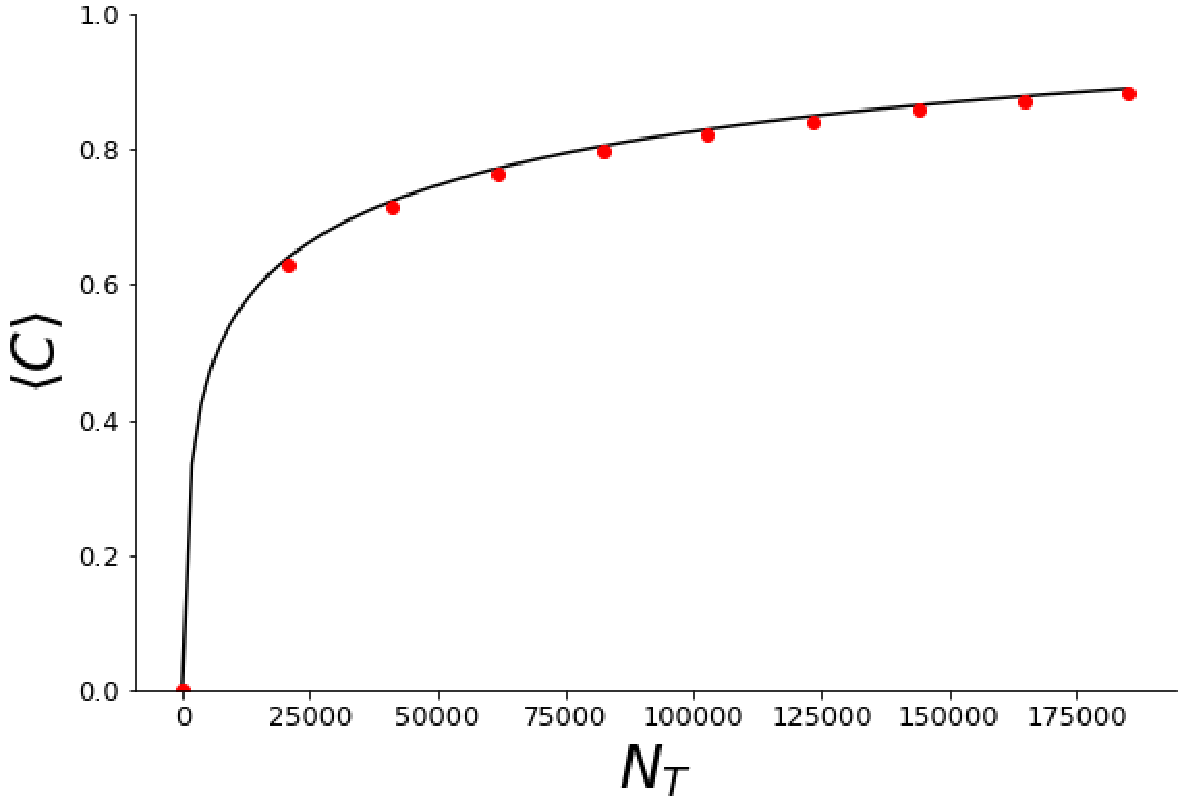

Fig. S17. Polycentric city. To check if the subsampling of the Chinese data sets negatively biases our results, we fit our model to data derived from the full Shanghai data set. As shown, our model still agrees well with data, thereby justifying our subsampling procedure; see "Datasets" section. Red dots show data, black thick line shows the model prediction.

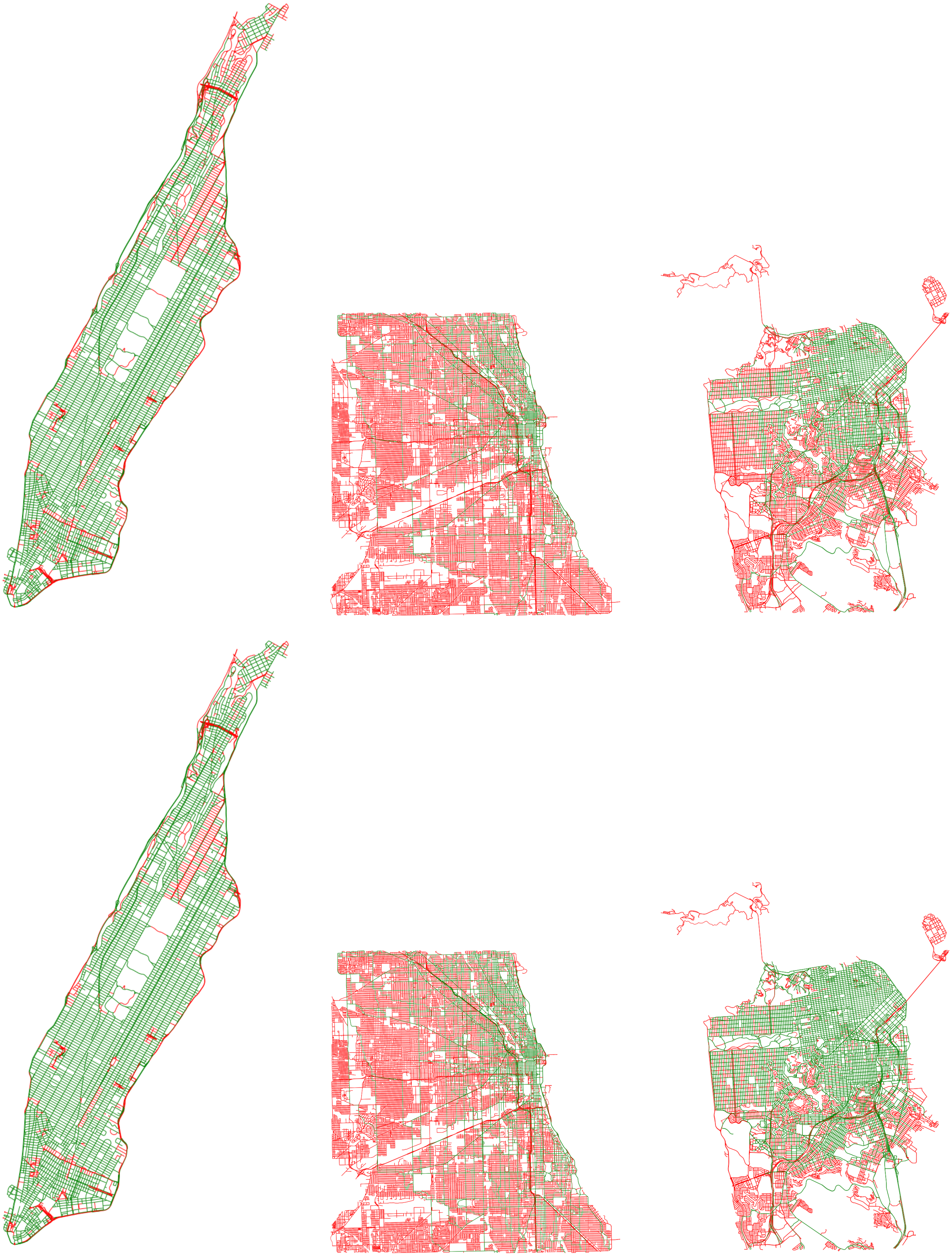

Fig. S18. Spatial coverage at at different saturation levels. NYC, Chicago, and San Francisco are shown in the first, second, and third columns. The top row shows the coverage when $N_T = 10\%$, and the bottom row shows the coverage when $N_T = 75\%$ of trips. Covered segments are colored green, while uncovered segment are colored red. As can be seen when N_T is small (top row), predominantly segments in the city center are covered. When N_T increases, areas outside of the city center become covered. See "Spatial bias of drive-by sensing" section.

City	Trajectories	Taxi Ids	Temporal range	$N_{S,total}$	N_S	$\frac{N_{S,total}}{N_S}$
Yangpu	Real (GPS)	Yes	1 Week: 04/01/15 – 04/04/15	2919	2657	0.94
NYC	Generated	Yes	1 Year: 12/31/10 – 12/31/11	7954	7265	0.91
Chicago	Generated	No	1 Week: 06/23/14 – 06/30/14	24054	12492	0.52
Vienna	Generated	No	1 Week: 03/07/11 – 10/07/11	24054	15775	0.66
San Francisco	Generated	No	1 Week: 05/21/08 – 05/28/08	15453	11708	0.76
Singapore	Generated	No	1 Week: 02/21/11 – 02/28/11	32362	25255	0.78
Beijing	Real (GPS)	Yes	1 Week: 03/01/14 – 03/07/14	54665	27024	0.49
Changsha	Real (GPS)	Yes	1 Week: 03/01/14 – 03/07/14	18067	9882	0.55
Hangszhou	Real (GPS)	Yes	1 Week: 04/21/15 – 04/28/15	39056	16125	0.41
Shanghai	Real (GPS)	Yes	1 Week: 03/01/14 – 03/07/14	49899	21002	0.49

Table S1. Properties of data sets.

Maximum liklihood parameters

	(λ, β, D) stretched exponential	(λ, α, D)trunc. power law	(Λ, r)
Yangpu	$\left((1.3 \pm 0.1) * 10^6, 0.266 \pm 0.003, 0.08\right)$	$(5830 \pm 10, 1.132 \pm 0.004, 0.07)$	$(-1327, 0)$
NYC	$\left((15 \pm 4) * 10^3, 0.499 \pm 0.005, 0.03\right)$	$(780 \pm 20, 1.00 \pm 10^{-6}, 0.25)$	$(1600, 0)$
Singapore	$\left((591 \pm 8) * 10^3, 0.499 \pm 0.004, 0.02\right)$	$(3400 \pm 200, 1 \pm (6 * 10^{-8}), 0.2)$	$(3282, 0)$
Chicago	$\left((3.4 \pm 0.9) * 10^6, 0.187 \pm 0.005, 0.04\right)$	$(650 \pm 30, 1.170 \pm 0.006, 0.03)$	$(-208, 0)$
San Francisco	$\left((47 \pm 7) * 10^5, 0.257 \pm 0.005, 0.03\right)$	$(1330 \pm 50, 1.156 \pm 0.008, 0.04)$	$(-218, 0)$
Vienna	$\left((41 \pm 0.5) * 10^5, 0.293 \pm 0.006, 0.04\right)$	$(2420 \pm 80, 1.196 \pm 0.008, 0.05)$	$(-278, 0)$
Beijing	$\left((1.2 \pm 0.4) * 10^5, 0.42 \pm 0.002, 0.06\right)$	$(5940 \pm 10, 1.00 \pm 10^{-6}, 0.08)$	$(824, 0)$
Changsha	$\left((7.5 \pm 0.2) * 10^5, 0.34 \pm 0.003, 0.04\right)$	$(1750 \pm 10, 1.02 \pm 0.02, 0.04)$	$(248, 0)$
Hangzhou	$\left((1.3 \pm 0.2) * 10^6, 0.23 \pm 0.003, 0.05\right)$	$(1770 \pm 20, 1.16 \pm 0.004, 0.04)$	$(560, 0)$
Shanghai	$\left((7.8 \pm 0.4) * 10^5, 0.43 \pm 0.004, 0.05\right)$	$(4970 \pm 10, 1.00 \pm 10^{-6}, 0.06)$	$(564, 0)$

Table S2. Maximum liklihood estimations of parameters.

City	N_T^*	N_T^{**}	$N_{T,total}$	$N_T^*/N_{T,total}$	$N_T^{**}/N_{T,total}$	Date
Yangpu	947	11716	17571	5 %	67 %	04/02/15
New York	1179	8007	466237	0.3 %	1.87 %	01/05/11
Chicago	2619	26110	67848	4 %	38 %	05/21/14
Vienna	1010	7552	10948	9 %	68 %	03/25/11
San fran	1923	13166	36089	5 %	36 %	05/24/08
Singapore	1782	14355	401879	0.44 %	4 %	02/16/11

Table S3. Coverage statistics. $N_{T,total}$ **refers to the total number of trips occurring on the specified day.**

City	N_V^*	N_V^{**}	$N_{V,total}$	$N_V^*/N_{V,total}$	$N_V^{**}/N_{V,total}$	Date
Beijing	211	1330	4000	5 %	33 %	03/01/14
Changsha	227	1516	4300	5 %	35 %	03/01/14
Hangzhou	132	1321	2500	5 %	52 %	04/21/15
Shanghai	148	1519	2800	5 %	54 %	03/01/14

Table S4. Coverage statistics. $N_{V,total}$ refers to the total number of taxis on the specified day.

Table S5. Comparison of regular and adjusted sensing power for Manhattan data set

	$\langle C \rangle = 0.33$	$\langle C^{*} \rangle (N_w = 3) = 0.33$	$\langle C \rangle = 0.5$	$\langle C^{*} \rangle (N_w = 3) = 0.5$
N_T	400	4000	1179	14750
N_V	10	100	30	355

Mobility Pattern M	Street network S
segment popularities p_i	edge betweenness b_i
mean length of trip $\langle L \rangle$	mean path length $\langle l \rangle$
mean distance traveled by taxi $\langle B \rangle$	N/A

Table S6. Conjectured relationships between quantities from taxis mobility patterns M and quantities from the street network S on which the taxis move.

References

1. Santi P, Resta G, Szell M, Sobolevsky S, Strogatz SH, Ratti C, "Quantifying the benefits of vehicle pooling with shareability networks", Proceedings of the National Academy of Sciences, Vol. 111, n. 37, pp 13290-13294.
2. Cavellin, Laure Deville, Scott Weichenthal, Ryan Tack, Martina S. Ragettli, Audrey Smargiassi, and Marianne Hatzopoulou. "Investigating the Use Of Portable Air Pollution Sensors to Capture the Spatial Variability Of Traffic-Related Air Pollution." Environmental Science & Technology 50.1 (2016): 313-20. Print.
3. Harvey, E. Therese, Susanne Kratzer, and Petra Philipson. "Satellite-based Water Quality Monitoring for Improved Spatial and Temporal Retrieval of Chlorophyll-a in Coastal Waters." Remote Sensing of Environment 158 (2015): 417-30. Print.
4. Mckercher, Grant R., Jennifer A. Salmond, and Jennifer K. Vanos. "Characteristics and Applications of Small, Portable Gaseous Air Pollution Monitors." Environmental Pollution 223 (2017): 102-10. Print.
5. Rosenfeld, Adar, Michael Dorman, Joel Schwartz, Victor Novack, Allan C. Just, and Itai Kloog. "Estimating Daily Minimum, Maximum, and Mean near Surface Air Temperature Using Hybrid Satellite Models across Israel." Environmental Research 159 (2017): 297-312. Print.
6. Vardoulakis, S., N. Gonzalezflesca, B. Fisher, and K. Pericleous. "Spatial Variability of Air Pollution in the Vicinity of a Permanent Monitoring Station in Central Paris." Atmospheric Environment 39.15 (2005): 2725-736. Print.
7. Jeff Alstott, Ed Bullmore, Dietmar Plenz. (2014). powerlaw: a Python package for analysis of heavy-tailed distributions. PLoS ONE 9(1): e85777
8. Bruce Levib. "A representation for multinomial cumulative distribution functions". The Annals of Statistics (1981) 1123–1126
9. CRAWDAD (Date of access: 01/04/2016). http://crawdad.org/dartmouth/campus/20090909 (2009).
10. https://catalog.data.gov/data set/taxi-trips.
11. Tachet, R., Sagarra, O., Santi, P., Resta, G., Szell, M., Strogatz, S. H., & Ratti, C. (2017). Scaling law of urban ride sharing. Scientific reports, 7, 42868.
12. Newson, P., & Krumm, J. (2009). Hidden Markov map matching through noise and sparseness. In Proceedings of the 17th ACM SIGSPATIAL. https://www.microsoft.com/en-us/research/publication/hidden-markov-map-matching-noise-sparseness/
13. Wang, Pu, et al. "Understanding road usage patterns in urban areas." Scientific reports 2 (2012): 1001.
14. Church, R., ReVelle, C. (1974, December). The maximal covering location problem. In Papers of the Regional Science Association (Vol. 32, No. 1, pp. 101-118). Springer-Verlag.
15. ReVelle, C., Toregas, C., Falkson, L. (1976). Applications of the location set-covering problem. Geographical analysis, 8(1), 65-76."
16. Silver, Breezy. "SimplyAnalytics." The Charleston Advisor 20.2 (2018): 51-56.
17. Zhao, Kai, et al. "Explaining the power-law distribution of human mobility through transportation modality decomposition." Scientific reports 5 (2015): 9136.
18. Liang, Xiao, et al. "The scaling of human mobility by taxis is exponential." Physica A: Statistical Mechanics and its Applications 391.5 (2012): 2135-2144.