



The 'Taxi Driver' Process and Passive Mobile Sensing on Networks

Citation

Giles, Michael. 2019. The 'Taxi Driver' Process and Passive Mobile Sensing on Networks. Bachelor's thesis, Harvard College.

Link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37364648>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#)

The ‘Taxi Driver’ Process and Passive Mobile Sensing on Networks

A thesis presented
to

Applied Mathematics

by

Michael Giles

in partial fulfillment of the honors requirements
for the degree of
Bachelor of Arts

Harvard College
Cambridge, Massachusetts
29 March 2019

Abstract

The modern world offers many examples and potential applications of mobile sensing agents passively traversing networks. However, the dynamics of these systems and their potential sensing power is relatively unexplored. In this paper, I examine a model designed to simulate this kind of process for urban sensing, extending it beyond urban street networks to general graphs. I calculate the number of visits, the mean first passage time, and the expected return time in the simplest case, and provide a bound and a scaling for the number of visits in more complex cases. Finally, I examine the spatial predictiveness of the model on a dataset of New York City taxi trips.

Acknowledgements

Many thanks to Dr. Kevin O’Keeffe for advising me and for sharing his research regarding the Taxi Driver process. Thanks to Prof. JP Onnela for his suggestions and for agreeing to help advise me. Thanks to Prof. Sean Eddy for encouraging me to work on this project when its future was in doubt, and for advising me throughout my time in the Applied Mathematics department. Finally, much gratitude to Prof. Andrew Murray and the entirety of the LS 50 professors and course staff for introducing me to applied mathematics and for opening a unique and enriching perspective onto it that I have appreciated ever since.

Contents

1	Introduction	4
1.1	Urban Sensing	4
1.2	Passive vs. Active Mobile Sensing	5
1.3	Random Search	5
1.4	Preferential Return	6
2	The ‘Taxi Driver’ Process	7
2.1	Definition	7
2.2	As a Random Walk on a Network of Routes	7
2.3	Node Visits	9
2.4	Mean First Passage	10
2.5	Recurrence Time	14
2.6	Exploring the dynamics for $\beta > 0$	15
2.6.1	Scaling of Expected Node Visits	15
2.6.2	Lower Bound on the Expected Value of Node Visits	17
2.6.3	Dominance of k Nodes	17
3	Spatial Prediction	19
4	Conclusion	22
4.1	Limitations	22
4.2	Additional Applications	22
4.3	Avenues for Further Inquiry	22
	References	23

1. Introduction

1.1 Urban Sensing

Cities are complex and dynamic environments, and the ability to design, govern, and inhabit them effectively depends on high-quality data about quantities of interest, such as pollution, vehicular traffic, and foot traffic. This makes urban data, as well as the sensors that collect it, vitally important. [12, 8]

Urban sensing is subject to a space-time trade-off in resolution. On one side are stationary sensors, such as traffic cameras, that offer good temporal resolution but are limited in space. On the other are aerial sensors such as satellites and plane flyovers, that offer excellent spatial resolution but occur infrequently [18]. These trade-offs are enforced by the cost of operating both methods – frequent flights are expensive, as is a comprehensive network of stationary sensors. In the middle of this trade-off lie mobile (or drive-by) sensors, which move around a city and continuously collect data. These might be vehicular [13] systems, or they may be handheld [15, 9] ones.

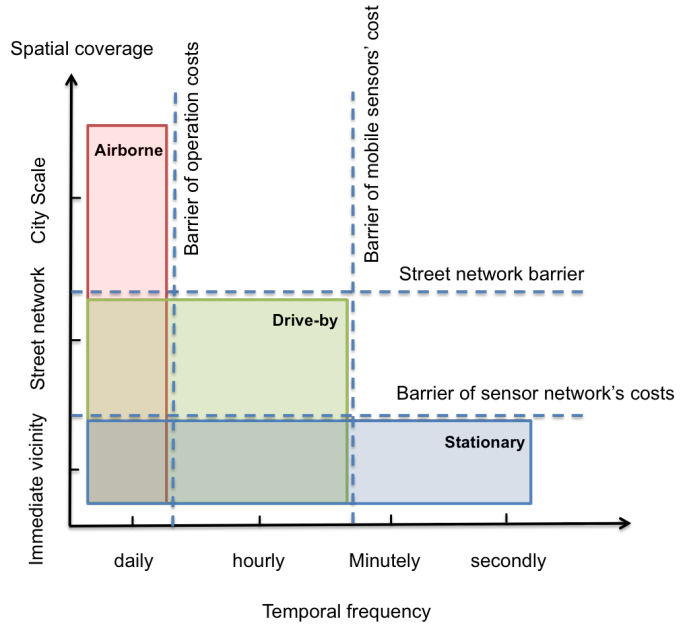


Figure 1.1: An illustration of the space-time trade-off for different sensing types (fig: [16])

1.2 Passive vs. Active Mobile Sensing

Mobile sensing can further be divided into active and passive paradigms. In *active* sensing, the mobile agent traverses the network for the express purpose of gathering data. A significant amount of work is being done on vehicular active sensing, both as the express purpose of a fleet of vehicles or as a component of a larger suite of capabilities [7, 2]. In contrast, *passive* sensors traverse the network for another purpose and data-gathering occurs incidentally. While it lacks some of the power and the scope of active sensing platforms, passive sensing is easier to adopt and less intrusive because it requires less dedicated engineering and infrastructure.

1.3 Random Search

Not all search protocols need to be random. Two fundamental kinds of graph search, depth-first search breadth-first search, are deterministic. In urban sensing, bus-lane monitoring approximates a kind of deterministic search, as the agents follow a predetermined path on a predetermined schedule [25]. In real systems, though, even this kind of patrol-based search is subject to randomness. Additionally, more explicitly random search is more robust to changes in the network, and can be more easily adapted to cover areas of higher demand.

The simplest random search protocol is a random walk, in which the agent diffuses around the network with equal probability of choosing any neighbor. In the case with weighted edges, the probability of traversing an edge may be inversely proportional to edge weight [14]. Other kinds of biases irrespective of edge weight might be imposed on the movement of the walker as well [4]. Other random search protocols are adaptations on random walks. In *intermittent random walks*, the agent proceeds to some point deterministically, and begins a random walk around that point. Intermittent random walks, when optimized correctly, can be superior to purely random walks in *blind search*, in which an agent is searching for a stationary target that it can only perceive when it is in near vicinity [17]. Another variety of random search is a *Lévy Flight*, or a random walk in which each step distance is drawn from a heavy-tailed Cauchy distribution [23]. Lévy flights have been compared to the way animals [20, 11], and have also been suggested as a way to model human mobility [19].

1.4 Preferential Return

The preferential return mechanism was developed to model patterns in human mobility [21]. Analogously to preferential attachment in network formation, they incorporate an exponent that makes it more likely for humans to return to locations they have already visited. This leads to an emergence of popular locations that is fit by a power law [10].

2. The ‘Taxi Driver’ Process

The ‘Taxi Driver’ Process is a random search process on graphs designed to simulate drive-by sensing via cab-mounted sensors [16]. It incorporates a preferential return mechanism and the variable step length given by Lévy walks and intermittent random walks. As such, it has interesting implications for modeling and understanding search on networks as well as human mobility.

2.1 Definition

Let $G = (V, E)$ be a connected (without self-loops) directed graph with nodes $V = \{1 \dots N\}$ and weight function W . Let R_{ij} be a set of routes from node i to node j , where $R_{ij}^{(k)}$ is the k^{th} element in this set and $1 \geq k \leq r_{ij}$, with r_{ij} being the number of routes from node i to j and M being the total number of routes. By the connectedness of G , the total number of sets R_{ij} must be $S = N(N - 1)$. In the urban sensing problem, R_{ij} is the set of geodesics between i and j and k is the number of geodesics between i and j . In general, this need not be true.

Beginning at a node i , an agent (the ‘taxi driver’) selects a destination node j (where $i \neq j$) according to the distribution

$$p_j(t) \sim v_j(t)^\beta$$

Where $v_j(t) > 0$ is the number of visits to node j for trip t , and β is a parameter fit to the network in question. The agent chooses a route r at random from R_{ij} and traverses it. Upon arriving at j , the agent repeats this process.

2.2 As a Random Walk on a Network of Routes

On G itself, this process is not Markovian, as the next node visited depends on the route currently being traversed. However, if the node popularities are constant it can be transformed into a process that is by considering it as a random walk on a network of routes. Enumerate all of the routes, and let $\mathbf{x}_t = (x_t^1, x_t^2, \dots, x_t^M)$ be a row vector giving the probability distribution across all of them. This distribution evolves according to the equation

$$\mathbf{x}_{t+1} = \mathbf{x}_t \mathbf{P}_t \quad (2.1)$$

To calculate $\mathbf{P}_t$, first find the probability that a given node is chosen as the next destination. Define the $M \times N$ matrix $\mathbf{C} = [C_{mn}] = [I(\text{node } n \text{ in route } m)]$ and the $N \times M$ matrix $\mathbf{V} = [V_{nm}] = \left[\frac{1}{r_{ij}} I(\text{node } n \text{ in route } m) \right]$, where r_{ij} is the number of routes running from the same origin and leading to the same destination as m . Then,

$$\mathbf{p}_t = \frac{1}{K} (tx_t V)^{\circ \beta} \mathbf{V} \quad (2.2)$$

Where K is a normalizing constant and $\circ \beta$ gives the element-wise (Hadamard) exponentiation, i.e.

$$\begin{aligned} \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}^{\circ \beta} &:= \begin{bmatrix} a_{11}^\beta & \dots & a_{1n}^\beta \\ \vdots & \ddots & \vdots \\ a_{n1}^\beta & \dots & a_{nn}^\beta \end{bmatrix} \\ \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}^{\circ 0} &:= \mathbf{I} \end{aligned}$$

The transition matrix is then

$$\mathbf{P}_t = \mathbf{K} \left[\begin{pmatrix} \vdots \\ p_{t\{1=0\}} \\ \vdots \end{pmatrix} \cdots \begin{pmatrix} \vdots \\ p_{t\{M=0\}} \\ \vdots \end{pmatrix} \right] \quad (2.3)$$

Where $\mathbf{K}$ is a diagonal matrix with normalizing constants for each row on the diagonal:

$$\mathbf{K} = \begin{bmatrix} 1/K_1 & \dots & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & 1/K_n \end{bmatrix}$$

and $p_{t\{i=0\}}$ is p_t with the i^{th} component set to zero. Then,

$$\mathbf{x}_{t+1} = \mathbf{x}_0 \mathbf{P}_0 \mathbf{P}_1 \cdots \mathbf{P}_t \quad (2.4)$$

Or, with $\mathbf{s}_0$ being the distribution of start states on nodes

$$\mathbf{x}_{t+1} = \mathbf{s}_0 \mathbf{C} \mathbf{P}_0 \mathbf{P}_1 \cdots \mathbf{P}_t \quad (2.5)$$

If we take the transition probabilities to be constant over all trips, then this reduces to a Markov chain whose behavior for large t approaches that of the stationary distribution $\boldsymbol{\pi}$, which is given by the left eigenvector of the transition matrix with eigenvalue 1 [6].

If $\beta = 0$, then all transition probabilities between nodes are equal. Imposing the constraint $R_{ij} = R_{ji} \forall i, j$, this further reduces to a random walk on an undirected network of routes. The stationary distribution on these routes is then

$$\boldsymbol{\pi} = (1/r_1 M, \dots, 1/r_S M) \quad (2.6)$$

and therefore for this case

$$\lim_{t \rightarrow \infty} \mathbf{x}_t = \frac{\mathbf{R} \mathbf{s}_0 \mathbf{C}}{M} \quad (2.7)$$

Where $\mathbf{R}$ is the $M \times M$ diagonal matrix with $r_1, \dots, r_S$ on the diagonal.

2.3 Node Visits

The most important measure of a mobile sensor is the number of nodes it has visited in a given time period. In their publication specifying the Taxi Driver process, the authors introduce a new measure for betweenness, which I will also use here. In this version of betweenness, the betweenness of a node i is given by the number of routes for which there exists an edge (v, i) divided by the total number of routes. In their paper, the authors demonstrated a connection between this betweenness and the mean traversals an intersection receives, which I will generalize and elaborate on here [16]

Define the $N \times M$ matrix $\mathbf{C} = [c_{ij}] = [I(\text{route } j \text{ contains node } i)]$ for the route between nodes m and n . Then, the number of visits each node has received at time t , $\mathbf{v}_t = (v_t^1, \dots, v_t^N)$, is given by

$$t \mathbf{x}_t \mathbf{C}$$

Let C_i be the i^{th} column of $\mathbf{C}$. Then the number of visits of i is

$$t \mathbf{x}_t C_i$$

If the transition probabilities are constant over all trips, then the visits approach

$$\mathbf{o}_t = t\boldsymbol{\pi}\mathbf{C}$$

for large t , and the number of visits at i is

$$t\boldsymbol{\pi}C_i \tag{2.8}$$

Approximation

For $\beta = 0$, we have $\boldsymbol{\pi} = (1/r_1M, \dots 1/r_M M)$ Assuming that $r_{ij} = r \ \forall i, j$, we have $\boldsymbol{\pi} = (1/rM, \dots 1/rM)$. Notice that the total for the i^{th} row of $\mathbf{V}$ is exactly $M\text{betweenness}(i)$. So, for large t

$$v_t^i \approx t \frac{\text{betweenness}(i)}{r} \tag{2.9}$$

The Taxi Driver process was simulated on six different graph types: **A** Line **B** Cycle **C** Circular Ladder **D** Lattice **E** Barabási-Albert [3] **F** Watts-Strogatz [24] A and B had $N = 100$ nodes, C had 50 nodes in each ring for a total of 100, D was 10×10 , E had $N = 100$ nodes and 4 new connections, and F had $N = 100$ nodes, 4 connections, and $p = 0.1$ of a new connection.

These graphs were chosen for their simplicity (A,B), their resemblance to road systems (B, C), and their importance as a common motif in emergent networks (E,F). The parameters were designed to make them as comparable as possible in both the number of nodes and the number of edges, if the graph type allowed.

Comment

The approximation was qualitatively very good, but it tended to overestimate the measured mean occupancy. This might have to do with the skewedness of visits on a per-node basis from trial to trial. The one exception to this was the lattice graph, which had much higher mean visits than predicted by the approximation. This likely had to do with the simplification on r – of all of the graph types, the lattice by far had the most variability in the number of equivalent paths between two nodes, making this simplification the least appropriate for this case.

2.4 Mean First Passage

The *mean first passage* ($\langle MFP_v \rangle$) for a node $v \in V$ is the amount of time it takes to first encounter v , averaged over all the nodes in V . $\langle MFP_v \rangle$ can be expressed

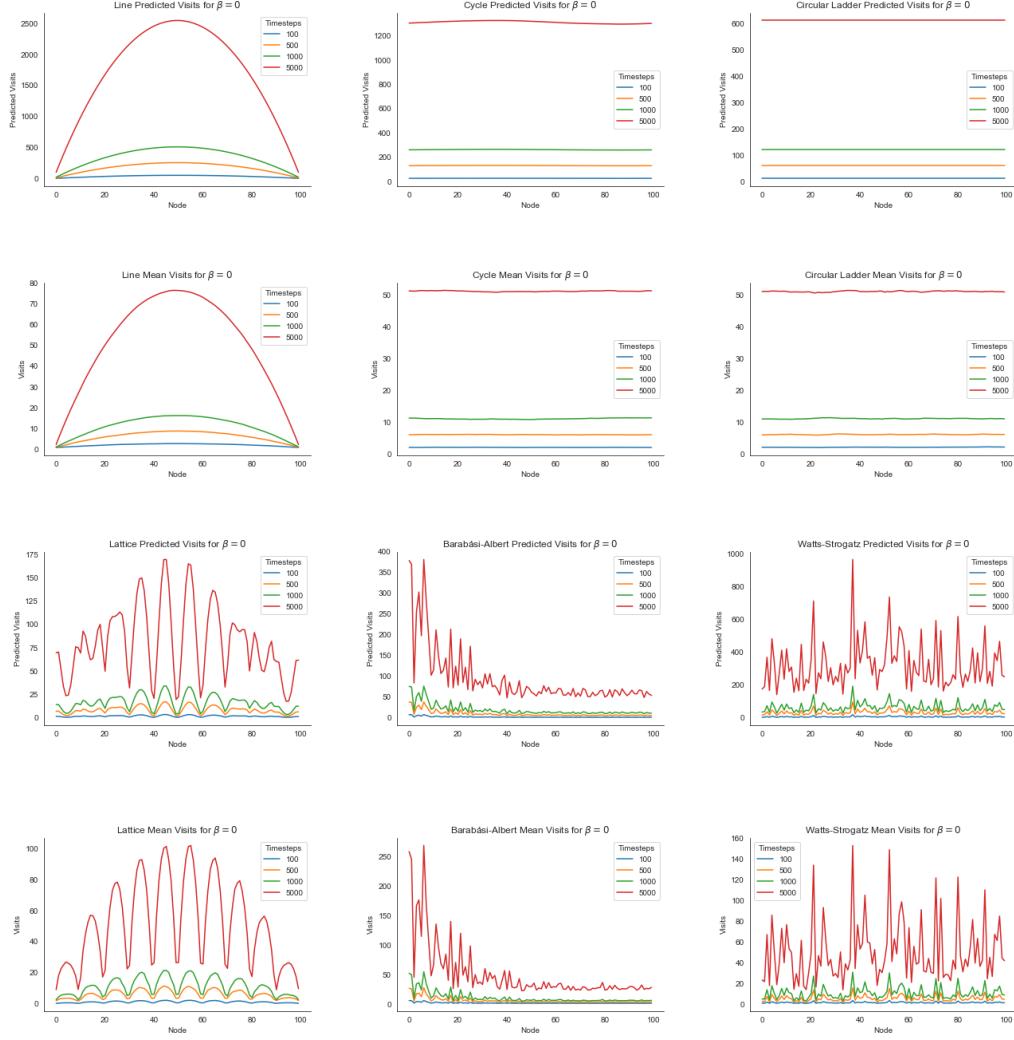


Figure 2.1: Simulation for node visits. 1000 iterations were performed per trial (with the exception of the lattice, which was done with 100 iterations), which lasted 1000 time points.

by adding an absorbing state F to the state space and connecting it to the subset of the routes containing v via b intermediate states $B = \{B_1, \dots, B_b\}$, where b is the number of routes in which v appears. Notice that b is exactly $M\text{betweenness}(v)$ times. The new transition probabilities are given by

$$\begin{aligned} P(X_{t+1} = A_i | X_t = R_i) &= 1, \quad \forall R_i : v \in R_i \\ P(X_{t+1} = F | X_t = B_i) &= 1, \quad \forall B_i \in B \\ P(X_{t+1} = F | X_t = F) &= 1 \end{aligned}$$

The number of trips until v is encountered can be calculated by defining a matrix $\mathbf{N}$

$$\mathbf{N} = \lim_{t \rightarrow \infty} \sum_{i=1}^t \prod_{j=1}^i \mathbf{Q}_j \quad (2.10)$$

Where $\mathbf{Q}_t$ is the $(M+2b) \times (M+2b)$ transition matrix of the transient states (i.e. F is excluded) at time t [6]. Notice that in the case where β is positive, this may not exist. If it does, the mean distance till absorption can be calculated by constructing the $N \times M+2b$ matrix $\mathbf{D} = [D_{nm}]$, where

$$D_{nm} = \begin{cases} L(R_m) & R_m \text{ begins with } n \\ 0 & \text{o.w.} \end{cases}$$

Where $L(R_m)$ is the sum of the edge weights along R_m until an edge ending in v is encountered. If R_m doesn't encounter v , this is merely the total length of the route.

Then, with $\mathbf{s}_0$ being the distribution on start states,

$$MFP_v = \mathbf{s}_0 \mathbf{D} \mathbf{N} \quad (2.11)$$

If the probabilities are constant [6],

$$\mathbf{N} = (\mathbf{I} - \mathbf{Q})^{-1}$$

Approximation

We take the formula for MFT_v^i , i.e. the mean first passage to v starting at i :

$$MFT_v^i = \frac{1}{N_M - 1} \left[\sum_{R_{ij} \in \mathcal{C}_i} \frac{L(R_{ij})}{r_{ij}} + MFT_v^j + \sum_{R_{uw} \in \mathcal{T}_i} \frac{L(R_{uw})}{r_{uw}} \right]$$

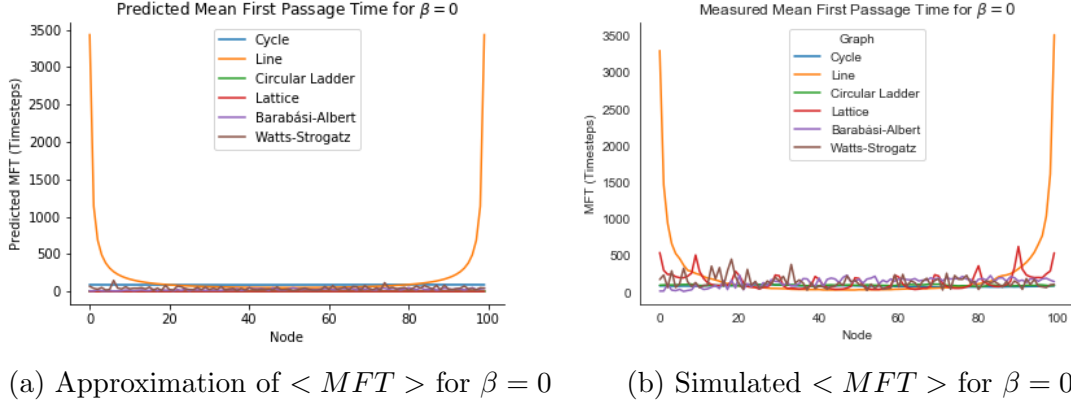


Figure 2.2: Approximate versus simulated $\langle MFT \rangle$ for the networks over 1000 iterations (100 for the lattice)

Where $\mathcal{T}_i$ and $\mathcal{C}_i$ are the sets of routes beginning with j that respectively do and do not encounter v . The cardinalities of these sets are $N_M - b_v$ and b_v , respectively

We make several simplifying assumptions: $MFT_v^i = MFT_v^j = M_v \forall i, j$, $L(R_{ij}) = l \forall R_{ij} \in \mathcal{C}_i$, $L(R_{uw}) = \frac{l}{2} \forall R_{uw} \in \mathcal{T}_i$, and $r_{ij} = r \forall i, j$. This gives us the equation

$$M_v = \frac{1}{N_M - 1} \left[(N_M - b_v) \left(\frac{l}{r} + M_v \right) + \frac{b_v l}{2r} \right]$$

With solution

$$M_v = -\frac{l \sum_{i,j} \frac{1}{r_{i,j}} (b_v - 2N_M)}{2(b_v - 1)(N_M - 1)}$$

Approximating b_v with $\text{betweenness}(v)N_M$, we have

$$\langle MFT_v \rangle \approx -\frac{(\text{betweenness}(v) - 2)lN_M \sum_{i,j} \frac{1}{r_{i,j}}}{2(N_M - 1)(\text{betweenness}(v)N_M - 1)} \quad (2.12)$$

2.2 compares the predicted MFT to the observed one. Of note is the high MFT for the endpoints of the line, as predicted by the model – these appear in only one shortest-path route, so as the number of nodes increase it takes longer and longer to visit them for the first time.

2.5 Recurrence Time

The *recurrence time* for a node i (t_R^i) gives the expected amount of elapsed time between visits to i . In the urban sensing application, this is useful because it gives a sense of the currency of information about a given intersection. For large t , the recurrence time for a given route j is [6].

$$\frac{1}{\pi_j}$$

To go from a stationary distribution on routes to one on nodes, multiply by the matrix $\mathbf{V}$. So, with V_v representing the v^{th} column of $\mathbf{V}$

$$t_R^i = \frac{1}{\boldsymbol{\pi} V_i} \quad (2.13)$$

Approximation

As again we have $\boldsymbol{\pi} = (1/r_1 M, \dots, 1/r_M M)$ for $\beta = 0$,

$$t_R^i = M \left(\frac{1}{\sum_j \frac{v_{ij}}{r_j}} \right)$$

Assuming $r_1 = \dots = r_M = r$, we have

$$t_R^i \approx \frac{r}{\text{betweenness}(i)} \quad (2.14)$$

Notice that this is in units of trips. To approximate clock time, multiply by the average path length $\bar{\ell}$.

Comment

For most of the graphs, the approximation above was reasonably accurate. However, for the lattice graph it was much less so. This is likely because the simplifications made in the approximation were not appropriate ones for the lattice. In particular, the distribution of path lengths for the lattice is extremely variable and not well-summarized by a single average path length. Because the estimate was multiplied by the average path length at the end, this had a large adverse effect on the quality of the approximation. Additionally, the distribution of r_{ij} is quite variable for the lattice, and approximating it by a single number may not have been appropriate.

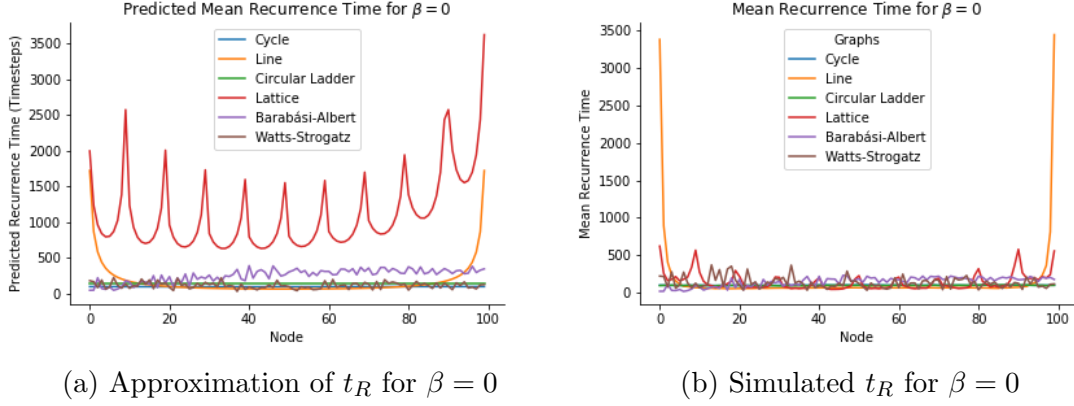


Figure 2.3: Approximate versus simulated t_R for the networks over 1000 iterations (100 for the lattice)

Nonetheless, the approximation even for the lattice was qualitatively quite good, and would likely be better if a more appropriate measure than average path length.

2.6 Exploring the dynamics for $\beta > 0$

2.6.1 Scaling of Expected Node Visits

Let $\mathbf{v}_t = (v_t^1, v_t^2 \dots v_t^N)$ be the visits each node has received after trip t . Then,

$$E[v_{t+1}^i | \mathbf{v}_t] = v_t^i + \sum_{j=0}^N \frac{(v_j^t)^\beta}{\|\mathbf{v}_t\|_\beta^\beta} I(i \in R_{ij}) \quad (2.15)$$

For $\beta > 1$, we can apply Hölder's Inequality:

$$E[v_{t+1}^i | \mathbf{v}_t] \geq v_t^i + \frac{\|\mathbf{v}_t\|_1^\beta (N\text{betweenness}(i))^{\frac{\beta}{\beta-1}}}{\|\mathbf{v}_t\|_\beta^{\beta-1}} \quad (2.16)$$

Applying Adam's Law and the linearity of expectation,

$$E[v_{t+1}^i - v_t^i] \geq (N\text{betweenness}(i))^{\frac{\beta}{\beta-1}} E \left[\frac{\|\mathbf{v}_t\|_1^\beta}{\|\mathbf{v}_t\|_\beta^{\beta-1}} \right] \quad (2.17)$$

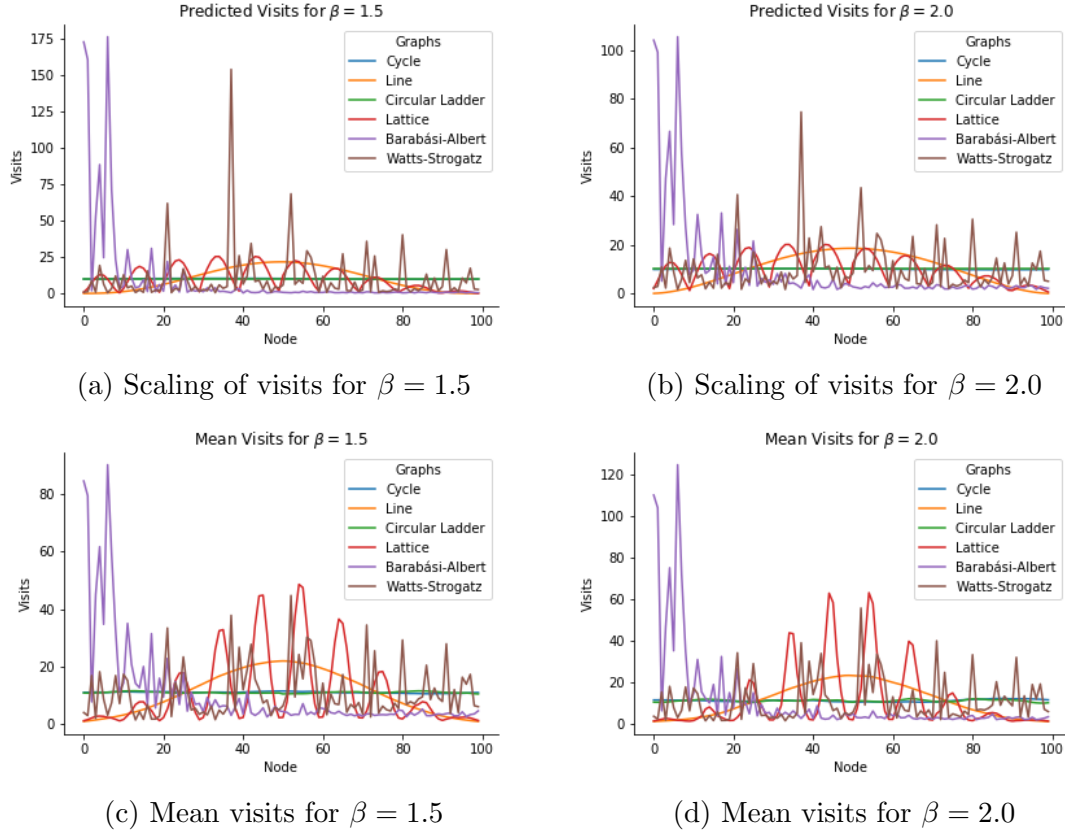


Figure 2.4: Visits predicted (top) and measured (bottom) for 1000 timesteps and $\beta = 1.5, 2.0$

Then, letting $K(t) = \sum_{j=0}^t E \left[\frac{\|\mathbf{v}_j\|_1^\beta}{\|\mathbf{v}_j\|_\beta^{\beta-1}} \right]$

$$E[v_{t+1}^i] \geq K(t)(N\text{betweenness}(i))^{\frac{\beta}{\beta-1}} \quad (2.18)$$

Thus demonstrating the scaling with $(N\text{betweenness}(i))^{\frac{\beta}{\beta-1}}$.

2.4 shows some disparities between the predicted and the actual behavior. The prediction overestimated the amount of visits for the Barabási-Albert graph, and over-sharpened peaks in the Watts-Strogatz graph that were not as prominent in the simulated trials. Additionally, the curve for the line graph is more pronounced in the simulation results. Perhaps most notably, it underestimated the visits for the lattice graph quite significantly.

2.6.2 Lower Bound on the Expected Value of Node Visits

From the above scaling, we can also calculate a lower bound on the expected value of visits for each node. By two further applications of Hölder's Inequality:

$$\begin{aligned} \frac{||\mathbf{v}_j||_1^\beta}{||\mathbf{v}_j||_\beta^{\beta-1}} &\geq \frac{\left(||\mathbf{v}_t||_\beta N^{\frac{-\beta}{\beta-1}}\right)^\beta}{||\mathbf{v}_t||_\beta^{\beta-1}} \\ &= ||\mathbf{v}_t||_\beta N^{\frac{-\beta^2}{\beta-1}} \\ &\geq ||\mathbf{v}_t||_1 N^{\frac{-(\beta+\beta^2)}{\beta-1}} \end{aligned}$$

So,

$$K(t) \geq N^{\frac{-(\beta+\beta^2)}{\beta-1}} \sum_{j=0}^t E[||\mathbf{v}_t||_1]$$

$E[||\mathbf{v}_t||_1] \leq tl_{min}$, where l_{min} is the minimum trip length

$$E[||\mathbf{v}_t||] = tE[\ell] = tl_{min}$$

Therefore

$$K(t) \geq N^{\frac{-(\beta+\beta^2)}{\beta-1}} \frac{t(t+1)}{2} l_{min} \quad (2.19)$$

Altogether,

$$E[v_{t+1}^i] \geq l_{min} \text{betweenness}(i)^{\frac{\beta}{\beta-1}} N^{\frac{-\beta^2}{\beta-1}} \frac{t(t+1)}{2} \quad (2.20)$$

In practice, this bound is far too permissive and the scaling from which it is derived is much more informative.

2.6.3 Dominance of k Nodes

Most of the previous analysis was done assuming the number of node visits was initially equal for every node in order to explore the effect of graph structure on the evolution of these probabilities. However, for $\beta > 0$ the Taxi Driver process demonstrates a great deal of sensitivity to an unequal distribution in visits. If the initial distribution of node visits is unequal, this may even be in spite of the structure of the graph. This is a useful feature of the model because in the urban sensing application, and others, locations are popular for many reasons outside the connectivity of

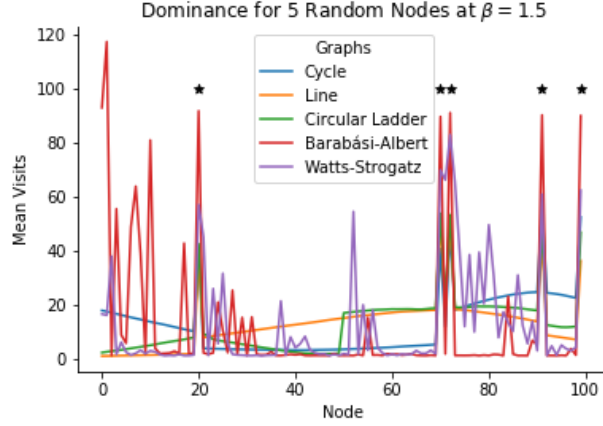


Figure 2.5: Example for $r = 10$ with a randomly-chosen subset of $k = 5$ nodes (marked with ★) for 1000 timesteps at $\beta = 1.5$

the graph. This stage of analysis is dedicated to exploring the effects of an unequal initial distribution of visits. To do this, we look at the difference in visits a required for a group of k nodes to be r times more popular than the remainder.

$$\frac{k(v+a)^\beta}{\|\mathbf{v}_t\|_\beta^\beta} \geq r \frac{(N-k)v^\beta}{\|\mathbf{v}_t\|_\beta^\beta}$$

$$a \geq v \left(\left(r \frac{(N-k)}{k} \right)^{1/\beta} - 1 \right)$$

With dimensionless parameter $\alpha = \frac{a}{v}$, this becomes

$$\alpha \geq \left(r \frac{(N-k)}{k} \right)^{1/\beta} - 1 \quad (2.21)$$

Discussion

Notice in 2.5 that the connectivity of the graph is still influential: though the chosen cohort of nodes are the most-visited in all other network types, in the Barabási-Albert graph they are not. This is because the most-visited node, 0, lies on the shortest path between the selected nodes. The popularity of these nodes therefore make 0 the most popular.

3. Spatial Prediction

It has previously been demonstrated that the Taxi Driver process is good at modeling the distribution of node visits for taxis on urban street networks [16]. However, while the overall distribution of visits matches that of the taxi traversals, it is unclear whether there is a street-to-street correspondence. In this section, I attempt to investigate this with a data set of Manhattan taxi trips. The dataset comprised a week of trips, 147800095 in total, from 1/18/11 to 1/25/11, and recorded the pick-up point, the drop-off point, and the intersections traversed in the duration of the journey [22]

The procedure discussed in the previous portion of the paper immediately ran into problems due to the size of the dataset involved. The taxi data set with which I was working traversed a section of New York containing 19266 intersections [5]. The dataset itself represented 3874 intersections, and given that many shortest-path routes might exist between two intersections, this became too much to handle. I began by restricting the set of intersections in two different ways. For the first, I chose the 500 most popular intersections by their frequency of occurrence as an origin or destination for a trip, accounting for 0.03271225646153308 of the visits in the dataset. For the second, I instead considered the 500 most-visited intersections, accounting for 0.4790159594760724 of the visits in the dataset. Even with this restriction, the dataset was prohibitively large: there were 11110739 all-pairs shortest paths between the intersections in the first subset and 10772297 between the intersections in the second. So, rather than construct and diagonalize this matrix, I instead modified my model.

I took

$$\mathbf{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_{500} \end{pmatrix}$$

to be the proportion of trips in which the most-popular intersections appeared, using the two methods above. Construct the 2500×500 matrix

$$\mathbf{R} = \begin{cases} \frac{1}{(\sum_j r_{ij})r_{ij}} & i \neq j \\ 0 & i = j \end{cases}$$

and the $N \times 2500$ matrix $\mathbf{M} = [M_{n,x(i,j)}]$ = number of times n appears in a route from i to j ,

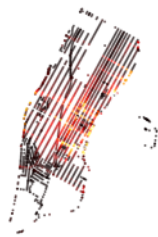
where x is a subsequence enumerating all pairs (i, j) . Then, the popularities of intersections represented by the subset is

$$\hat{\mathbf{p}} = \mathbf{MRp}$$

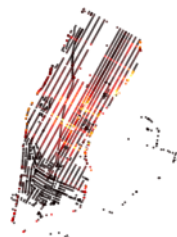
Then the predicted number of visits after t trips is $\hat{\mathbf{v}} = t\hat{\mathbf{p}}$

Discussion

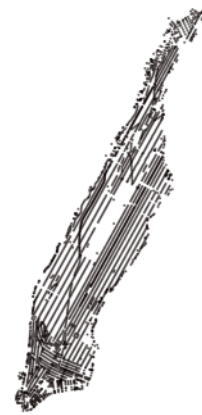
The best predictor of the actual proportion of visits was the sample of 100000 rides. The other two methods had a higher SSE likely because they tended to overemphasize the most popular routes as well as predict visits on nodes not in the dataset that happened to lie on a shortest path. Additionally, the emphasis on shortest paths eliminated most of the intersections in upper Manhattan from consideration, further raising the SSE. The increased accuracy of the sampling based off of all visits indicates that in reality the taxis are not taking the absolute shortest paths, but rather paths that are "trained" into the model by sampling from the most popular nodes. Further work is needed to determine what measure best predicts the popularity of a given path.



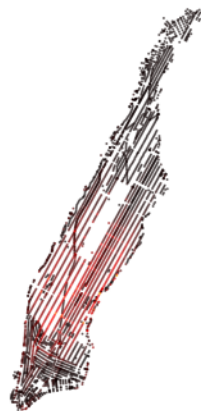
(a) Predicted visits by origin and destination (SSE: 0.3886)



(b) Predicted visits by most visits (SSE: $4.5834e-04$)



(c) Predicted visits from a random sample of 100000 rides (SSE: $1.5581e-05$)



(d) Measured visits

4. Conclusion

4.1 Limitations

One of the techniques that allowed us to calculate the quantities of interest for this process, and to approximate them for others, was by considering it to be a random walk on a network of routes. In doing this, we converted from clock time, where t was linked to the length of edges the agent traverses, to trip time, which is decoupled from this detail. To convert back to clock time, several times we multiplied by the average path length. This is a slight simplification, but it could pose some problems and impacted the accuracy of our predictions. A more precise approach might be to model the path length with a random variable $\ell(t)$ depending to the elapsed trip time. It would be an interesting future step as well to compare the distribution of this path length with the heavy-tailed Cauchy distribution used for Lévy walks.

4.2 Additional Applications

Outside of urban sensing, the Taxi Driver process might be useful in modeling passive sensing in other applications. One of these is indoor mapping, which is commercially important and academically interesting [27, 26]. In this application, the mobile agents are humans and the passive sensors might be smartphones, which can use wireless signals to map indoor spaces. The Taxi Driver process gives an idea for the number of agents and the amount of time required to map an indoor space completely.

A potential area of application outside of sensing is in workflow and process management. Workflows can be represented by graphs, with one process leading into another. If one process contains a flaw that causes some sort of global failure, the MFT and the recurrence time for the Taxi Driver on the workflow graph might give an idea of how long before that happens, as well as the expected amount of time in between global failures [1].

4.3 Avenues for Further Inquiry

This process was characterized with a single agent in isolation. Provided the agents do not interact, it is very straightforward to consider the process with more than

one agent after it has been characterized this way. However, there are potentially interesting situations for which this is not the case. If the cars belong to one fleet where the agents coordinate to reduce travel time, the dynamics of passive sensing for that fleet might be altered.

Another situation not captured by the current work is traffic and road link capacity. In addition to increasing travel time, traffic could have the effect of making the shortest-path route the fastest rather than the shortest one. Traffic also changes throughout the day, potentially necessitating re-routing and shortest-path sets that depend on time.

Bibliography

- [1] Wil van der Aalst. *Workflow management : models, methods, and systems*. eng. Cooperative information systems. Cambridge, Mass.: MIT Press, 2002. ISBN: 0262011891.
- [2] A. Anjomshoaa et al. “City Scanner: Building and Scheduling a Mobile Sensing Platform for Smart City Services”. In: *IEEE Internet of Things Journal* (2018), pp. 1–1. ISSN: 2327-4662. DOI: 10.1109/JIOT.2018.2839058.
- [3] Albert-László Barabási and Réka Albert. “Emergence of scaling in random networks”. In: *science* 286.5439 (1999), pp. 509–512.
- [4] R. Beraldi. “Biased Random Walks in Uniform Wireless Networks”. In: *IEEE Transactions on Mobile Computing* 8.4 (Apr. 2009), pp. 500–513. ISSN: 1536-1233. DOI: 10.1109/TMC.2008.151.
- [5] Geoff Boeing. “OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks”. In: *Computers Environment and Urban Systems* 65 (July 2017), pp. 126–139. DOI: 10.1016/j.compenvurbsys.2017.05.004.
- [6] Pierre Brémaud. *Markov chains : Gibbs fields, Monte Carlo simulation and queues*. eng. Texts in applied mathematics ; 31. New York: Springer, 1999. ISBN: 0387985093.
- [7] V Bychkovsky et al. “Data management in the CarTel mobile sensor computing system”. eng. In: *Proceedings of the 2006 ACM SIGMOD international conference on management of data*. SIGMOD ’06. ACM, 2006, pp. 730–732. ISBN: 1595934340.
- [8] Dana Cuff, Mark Hansen, and Jerry Kang. “Urban sensing: out of the woods”. eng. In: *Communications of the ACM* 51.3 (2008), pp. 24–33. ISSN: 1557-7317.

- [9] Prabal Dutta et al. “Common Sense: participatory urban sensing using a network of handheld air quality monitors”. eng. In: *Proceedings of the 7th ACM Conference on embedded networked sensor systems*. SenSys '09. ACM, 2009. ISBN: 9781605585192.
- [10] Marta C. González, César A. Hidalgo, and Albert-László Barabási. “Understanding individual human mobility patterns”. In: *Nature* 453.7196 (2008). ISSN: 0028-0836.
- [11] Nicolas E. Humphries and David W. Sims. “Optimal foraging strategies: Lévy walks balance searching and patch exploitation under a very broad range of conditions”. In: *Journal of Theoretical Biology* 358 (2014), pp. 179–193. ISSN: 0022-5193. DOI: <https://doi.org/10.1016/j.jtbi.2014.05.032>. URL: <http://www.sciencedirect.com/science/article/pii/S0022519314003051>.
- [12] Nicholas Lane et al. “Urban sensing systems: opportunistic or participatory?” eng. In: *Proceedings of the 9th workshop on mobile computing systems and applications*. HotMobile '08. ACM, 2008, pp. 11–16. ISBN: 9781605581187.
- [13] Uichin Lee and Mario Gerla. “A survey of urban vehicular sensing platforms”. eng. In: *Computer Networks* 54.4 (2010), pp. 527–544. ISSN: 1389-1286.
- [14] László Lovász. “Random Walks on Graphs: A Survey, Combinatorics, Paul Erdos is Eighty”. In: *Bolyai Soc. Math. Stud.* 2 (Jan. 1993).
- [15] Prashanth Mohan, Venkata N. Padmanabhan, and Ramachandran Ramjee. “Nericell: Rich Monitoring of Road and Traffic Conditions Using Mobile Smartphones”. In: *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems*. SenSys '08. Raleigh, NC, USA: ACM, 2008, pp. 323–336. ISBN: 978-1-59593-990-6. DOI: 10.1145/1460412.1460444. URL: <http://doi.acm.org/10.1145/1460412.1460444>.
- [16] Kevin O’Keeffe et al. “Urban sensing as a random search process”. In: *arXiv e-prints*, arXiv:1901.08678 (Jan. 2019), arXiv:1901.08678. arXiv: 1901.08678 [physics.soc-ph].
- [17] Gleb Oshanin et al. “Efficient search by optimized intermittent random walks”. In: *Journal of Physics A: Mathematical and Theoretical* 42.43 (Oct. 2009), p. 434008. DOI: 10.1088/1751-8113/42/43/434008. URL: <https://doi.org/10.1088/1751-8113/42/43/434008>.
- [18] *Remote Sensing of Urban and Suburban Areas*. eng. Vol. 10. Remote Sensing and Digital Image Processing. Dordrecht: Springer Netherlands, 2010. ISBN: 978-1-4020-4371-0.

- [19] I. Rhee et al. “On the Levy-Walk Nature of Human Mobility”. In: *IEEE/ACM Transactions on Networking* 19.3 (June 2011), pp. 630–643. ISSN: 1063-6692. DOI: 10.1109/TNET.2011.2120618.
- [20] David W. Sims et al. “Scaling laws of marine predator search behaviour”. English. In: *Nature* 451.7182 (Feb. 2008). Copyright - Copyright Nature Publishing Group Feb 28, 2008; Last updated - 2017-10-31; CODEN - NATUAS, pp. 1098–102. URL: <http://search.proquest.com.ezp-prod1.hul.harvard.edu/docview/204553175?accountid=11311>.
- [21] Chaoming Song et al. “Limits of Predictability in Human Mobility”. eng. In: *Science (Washington)* 327.5968 (2010), pp. 1018–1021. ISSN: 0036-8075. URL: <http://search.proquest.com/docview/21432829/>.
- [22] “New York City Taxi and Limousine Commission”. “*TLC Trip Record Data*”. 2011. URL: <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>.
- [23] G.M Viswanathan et al. “Lévy flights in random searches”. In: *Physica A: Statistical Mechanics and its Applications* 282.1 (2000), pp. 1–12. ISSN: 0378-4371. DOI: [https://doi.org/10.1016/S0378-4371\(00\)00071-6](https://doi.org/10.1016/S0378-4371(00)00071-6). URL: <http://www.sciencedirect.com/science/article/pii/S0378437100000716>.
- [24] Duncan J Watts and Steven H Strogatz. “Collective dynamics of ‘small-world’ networks”. In: *nature* 393.6684 (1998), p. 440.
- [25] X. Yu X. Li and K. He. “Towards Effective Bus Lane Monitoring Using Camera Sensors”. In: *Wireless Sensor Network* 3.5 (2011), pp. 174–182. DOI: 10.4236/wsn.2011.35020.
- [26] B. Zhou et al. “ALIMC: Activity Landmark-Based Indoor Mapping via Crowdsourcing”. In: *IEEE Transactions on Intelligent Transportation Systems* 16.5 (Oct. 2015), pp. 2774–2785. ISSN: 1524-9050. DOI: 10.1109/TITS.2015.2423326.
- [27] M. Zhou et al. “Personal Mobility Map Construction for Crowd-Sourced Wi-Fi Based Indoor Mapping”. In: *IEEE Communications Letters* 18.8 (Aug. 2014), pp. 1427–1430. ISSN: 1089-7798. DOI: 10.1109/LCOMM.2014.2333515.