1.204 Final Project
11 December 2012
J. Cressica Brazier

Modelling exploration and preferential attachment properties in individual human trajectories

using the methods presented in

Song, Chaoming, Tal Koren, Pu Wang, and Albert-László Barabási. 2010.
"Modelling the Scaling Properties of Human Mobility."

1. Introduction

I took the final project as an opportunity to understand the next step of 'trajectory modeling' after the Homework 3 analysis that employed the continuous time random walk (CTRW), by following the methods presented by Song et. al. (2010) in the paper "Modelling the Scaling Properties of Human Mobility" ("Song"). Song proposes an extension of the CTRW model to include the human mobility tendencies of exploration and preferential return. The authors resolve the three 'scaling anomalies' of the asymptotic behavior of humans (see Section 2), although they still do not incorporate short-term 'periodic modulations' such as diurnal travel patterns and correlations in the order in which locations are visited (Song 2010, 5).

This project is also a starting point for learning how to work with mobile phone data, which has great potential as a proxy for traditional mobility datasets. Mobility models based on call data records (CDR) have been used to describe and approximate differences in mobility based on urban form properties across different cities (Kang et al. 2011) as well as complex and patterns of use within cities (Toole et al. 2012) that are otherwise difficult to document without trip journals. Therefore, in this project, I am working towards an exploration of what kinds of information are contained within mobile phone records, by making a preliminary analysis of the relationships between trajectory measures and local urban form characteristics. In the future, I hope to incorporate these measures into research on comprehensive energy use monitoring across both transport (jump distances) and building operations (pause times at different locations).

The core of the individual mobility model proposed by Song contains only two parameters to describe the probability of visiting a new location, as well as one function to describe preferential return. But to consistently describe the empirical data, the modeler must also parameterize the displacement and waiting time distributions for that dataset. Furthermore, to verify that the different properties of a dataset, such as the CDR data used in this project, are consistent with the interrelationships of these parameters as proposed by Song, several more trajectory measures should be calculated. This process results in four main steps, which structure this final report:

- Determine the rate of exploration of new locations, the individual probability of visiting new locations, and the probability of returning to a previously-visited location (section 3a)
- Parameterize the probability density functions for displacement and waiting time (section 3b)
- Determine the relationship between the number of distinct locations visited and the time duration, as well as between the frequency of visiting a location and its order of visitation, to confirm the relationship between these parameters are consistent, as proposed by Song (section 3c)
- Generate simulated trajectories from the parameterized model, and compare the properties of these trajectories to the empirical data (section 4)

2. Methods and Data

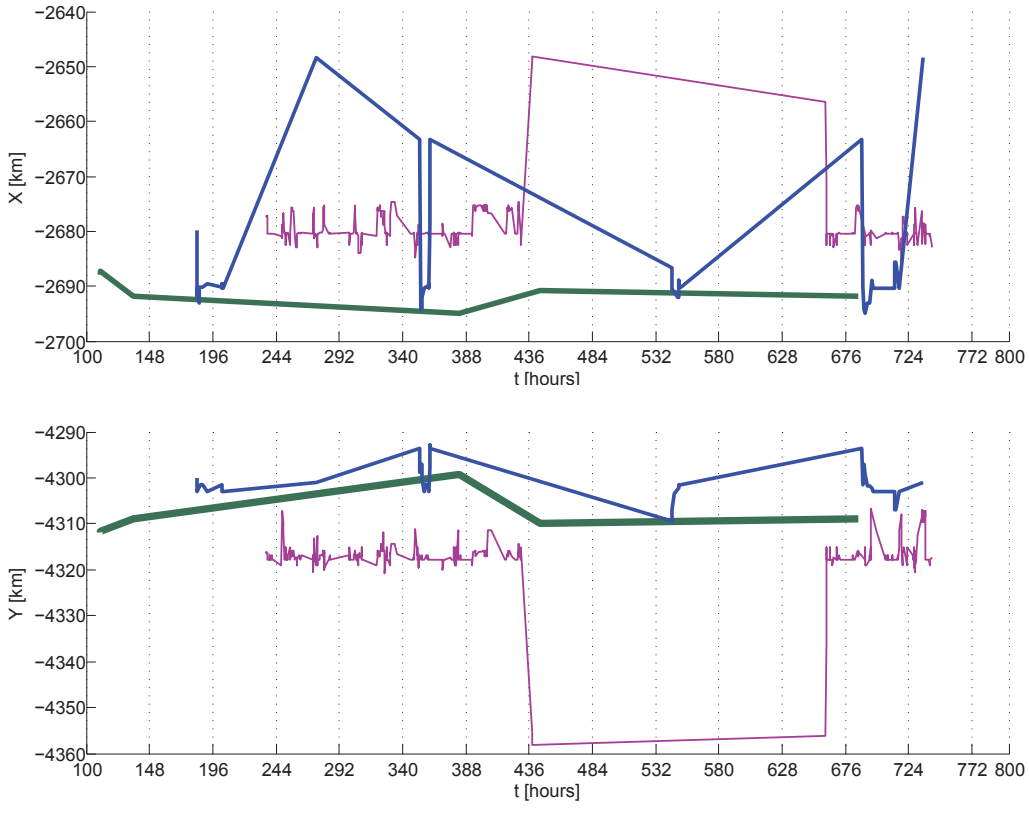
Song uses two mobility datasets, mobile phone trajectories and location-based service records, to empirically demonstrate the properties of the exploration and preferential return model. The mobile phone dataset contains the "time-resolved trajectories of three million anonymized mobile-phone users" (Song 2010, 1) for an entire year.

For this project, I will use the San Francisco mobile phone dataset containing 3126 users' trajectories (3107 after removing users with S = 1), a difference in sample size that is an important consideration for the meaning of the comparison with the article's results. The trajectories span one month, instead of one year. I read the data file into Matlab as a matrix with the following columns for each user, after removing the semicolon delimiters:

User ID	**Number of events**			
4082000002	18			
Time [seconds]	**X [km]**	**Y [km]**	**Tower ID**	**Other indices**
398224	-2687.8144	-4311.8173	3041	1 2
398318	-2687.2603	-4311.6812	3119	1 2

...

The graphs below illustrate the expected heterogeneity in the CDR trajectors among users, with some users having frequent activity, while others contain little usable data:



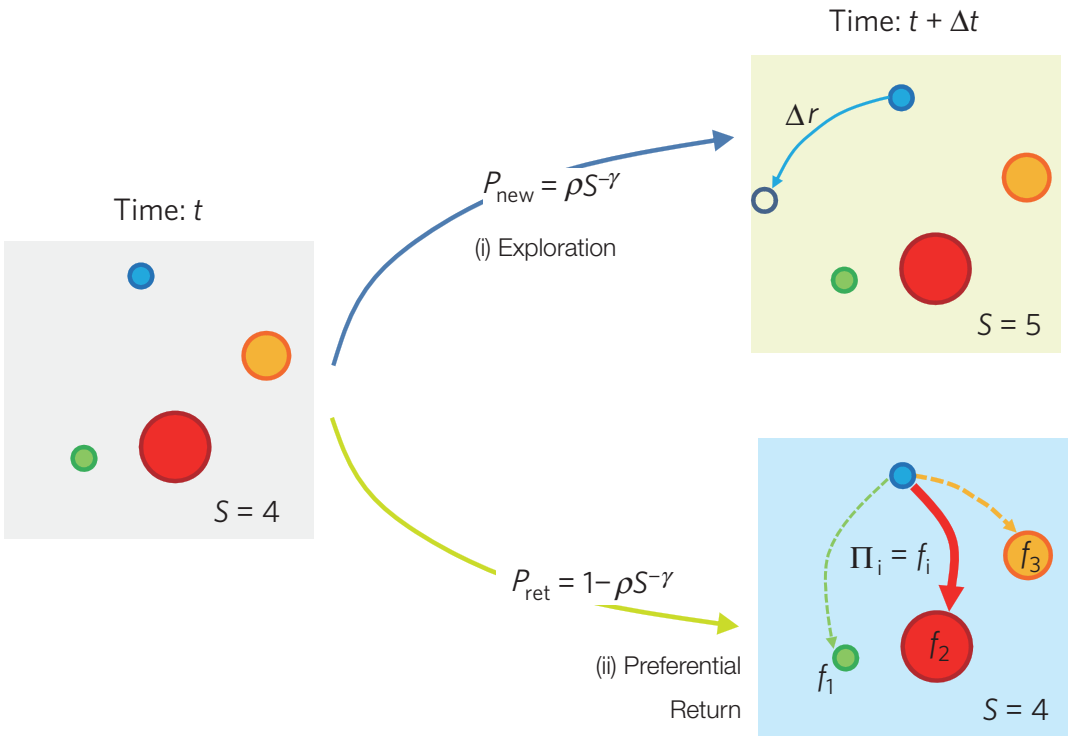
With these datasets, Song illustrates three critical differences, or 'scaling anomalies', between the human trajectories and CTRW models:

1. Humans visit fewer distinct locations over time, compared with the approximations of CTRW and Levy flights.
2. The visitation frequency distribution of humans tends towards zero, instead of becoming asymptotically constant for the least-visited locations, as it does in CTRW and Levy flights.
3. The mean square displacement (MSD) of humans grows more slowly than the CTRW prediction of logarithmic growth, so ultraslow diffusion should be incorporated into the model.

The authors then propose a new CTRW formulation that accounts for exploration limits (new locations are added more slowly over time) and preferential return (visitation probability is linked to the historical frequency of visiting each location). The model consists of three steps (shown below in Figure 2 from Song):

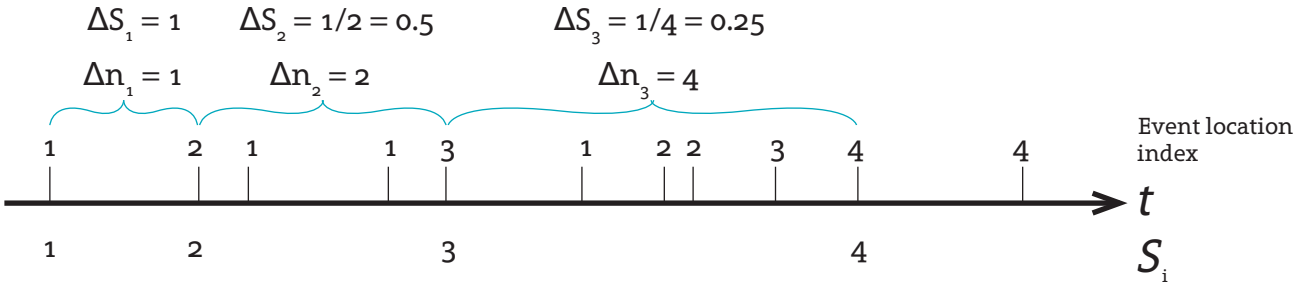
1. The individual pauses for an amount of time selected from the pause distribution P(Δt)
2. The individual may then move to a new location with the probability $P_{new} = \rho S^{-\gamma}$, or return to an already visited location with the complementary probability $P_{ret} = 1 - \rho S^{-\gamma}$
3. If the individual went exploring, s/he will move a distance selected from the jump distribution P(Δr). If the individual made a preferential return, s/he will return to a location with probability $\Pi_i = f_i$, where f_i is the frequency of each location's visits, and $f_k \sim k^{-\zeta}$.

The authors validate the model by generating trajectories and comparing the resulting radius of gyration distribution to those of the original mobile phone trajectories, showing they are in agreement.



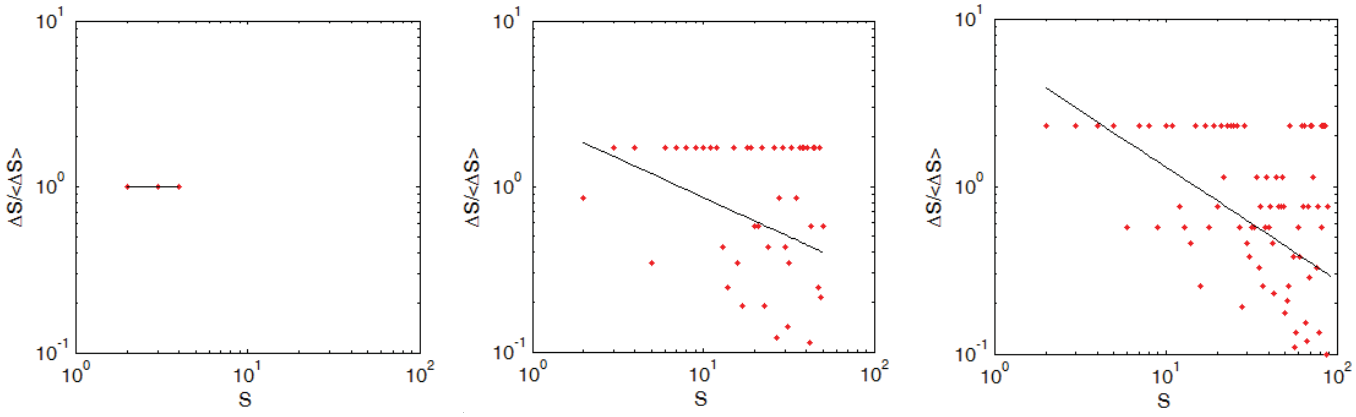
3. Model Parameterization

a. Measure rate of visiting new locations, ***ΔS vs. S***; estimate ***γ***; estimate ***ρ*** for each user

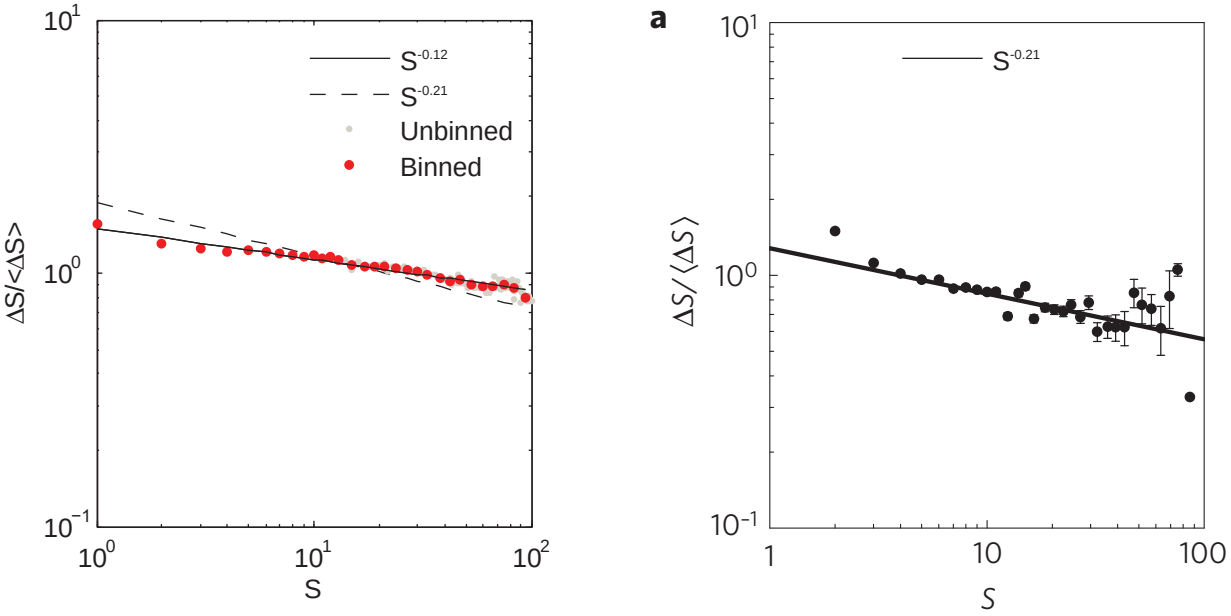


In the figure above, I diagram the procedure for determining the rate of visiting new locations versus returning to previously visited locations. Song initially sets up a vector in which visited locations are denoted with 0 and new locations receive the value 1, then averages all the visited locations corresponding to each S; instead, I directly count the number of visited locations up to and including the next new location, Δn_i. The change in the number of distinct locations corresponding to the previously visited S_i is then $\Delta S_i = 1 / \Delta n_i$.

Song then 'normalizes' ΔS for each user by dividing by its average value, <ΔS>, for 1 < S < 100. The rate of change for each individual, ΔS/<ΔS>, is therefore not dependent on that individual's ρ value, which cannot be determined until the global relationship $P_{new} \sim S^{-\gamma}$ has been established. Unfortunately, for users with S< 100, this average <ΔS> will be artificially high and will result in a lower γ value in the $\Delta S/\langle\Delta S\rangle \sim S^{-\gamma}$ relationship. Only 370 out of 3100 users reach 70 distinct locations within a month, and just 130 visit one hundred different cell towers. It is difficult to obtain a sample size that is not affected by the averaging problem and that does not have substantial noise in the data point distribution. I illustrate this problem in the three plots below, for a user with S ~= 5, 50, and 100, respectively. The 'slope' of the fitted line ranges from 0 < γ < 0.61, which demonstrates that users who don't span enough locations (S) have highly variable fitted line results. For a small sample size, an accurate value of γ cannot be determined.

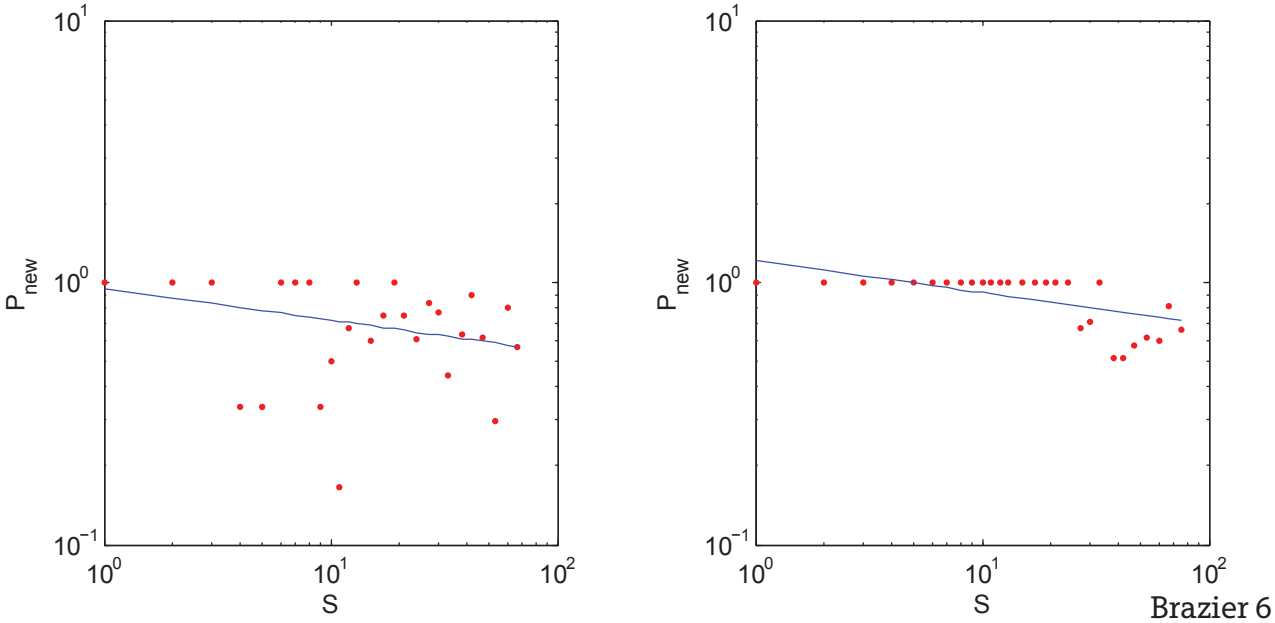


To move forward, I conducted a sensitivity analysis on the amount of users to include for determining the global $\Delta S/\langle \Delta S \rangle \sim S^{-\gamma}$, and the fitted parameter range was $0.098 < \gamma < 0.121$ for the range of users with $5 < S_{max} < 90$. I finally select users with $S_{max} > 50$, with a corresponding model parameter of $\gamma = 0.121$. The plot below compares these results to the corresponding Figure 3a (Song 2010, 4).

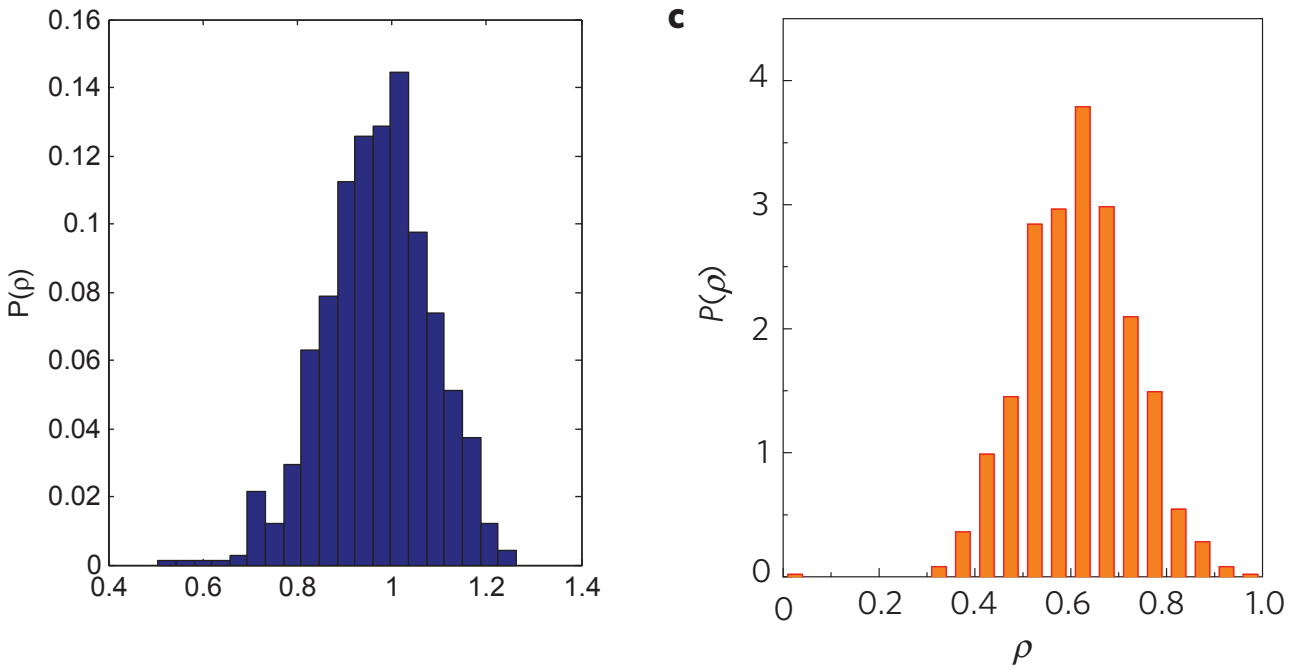


The Song data point mean is centered much lower than in my plot, which means that my methods require further review. In the interests of time, I did not reproduce Figure 3b, which shows the empirical and estimated relationship between the probability of return to a location versus the frequency of prior visitation, $\Pi = f$. Instead, I will directly use the stabilized frequencies f_i, which are calculated in Section 3c, as a probability distribution for the simulations.

I give two examples of determining ρ in the relationship $\Delta S = P_{new} = \rho S^{-0.12}$ for each user, below.

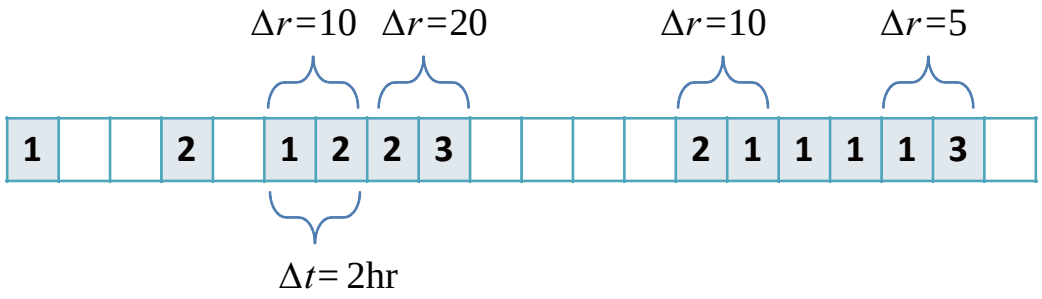


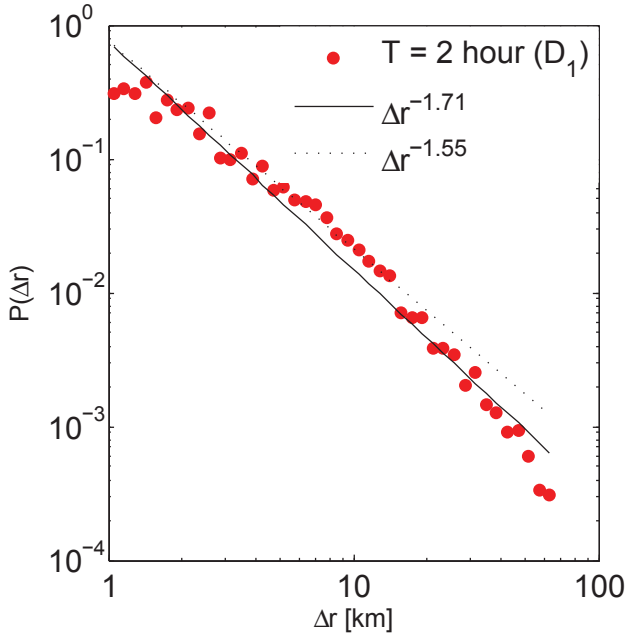
The following histogram for ρ over all users with $S_{max} > 50$ exhibits the same trend as the ρ distribution in Song's Figure 3c. The distribution is centered around ρ ≈ 0.97, compared with the average value of ρ ≈ 0.6 in Figure 3c. Again, this distribution represents a less complete dataset, in which users who have visited fewer locations by the end of one month also produce lower ρ than in Song's dataset.

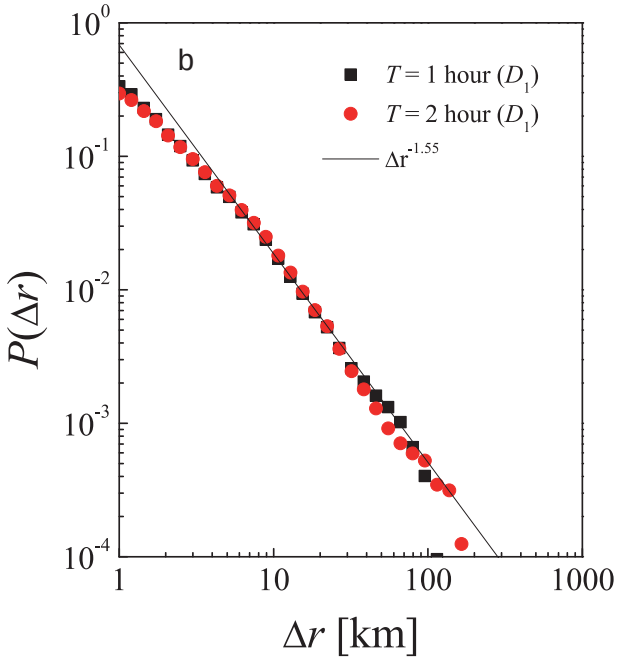


b. Measure jump distance and pause time distributions, ***P(Δr)*** and ***P(Δt)***; estimate ***α***, ***β***

To measure the distances and pauses, I must first filter the CDR to obtain the events that are separated by a fixed time interval (2 hours), in order to "correct the bias from the widely varying interevent times that characterize the calling pattern of each user" (Song 2010, SC4). Song shows that the remaining trajectory points represent a set of known distances traveled at known time intervals, which are not "driven by the statistics of call activity." I based the following diagram of this filtering process on Figure S3 in the Supplementary Information (Song 2010). The shaded boxes represent events for which the interevent time interval is 2 hours. I then fitted the probability density function to the exponential curve, $P(\Delta r) \sim \Delta r^{-1-\alpha}$, to estimate ***α***. The jump distance parameter ***α = 0.71*** appears to follow a more rapidly decaying distribution, compared with Song's estimate of ***α = 0.55.*** ***(see next page)***

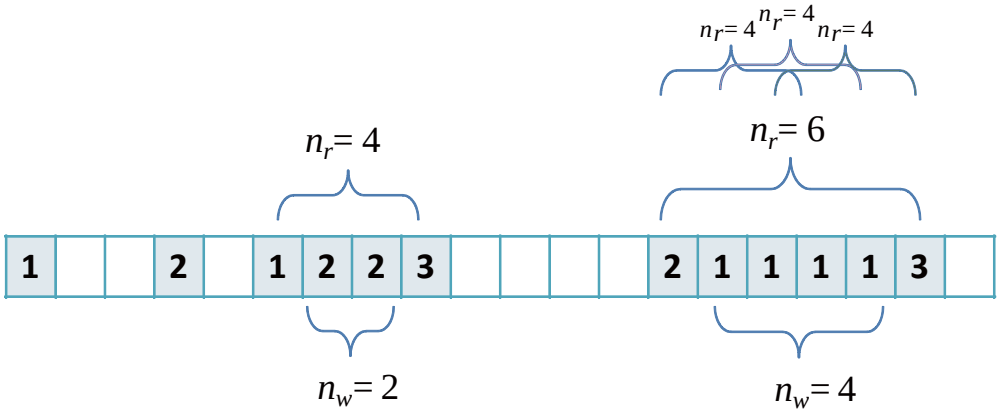






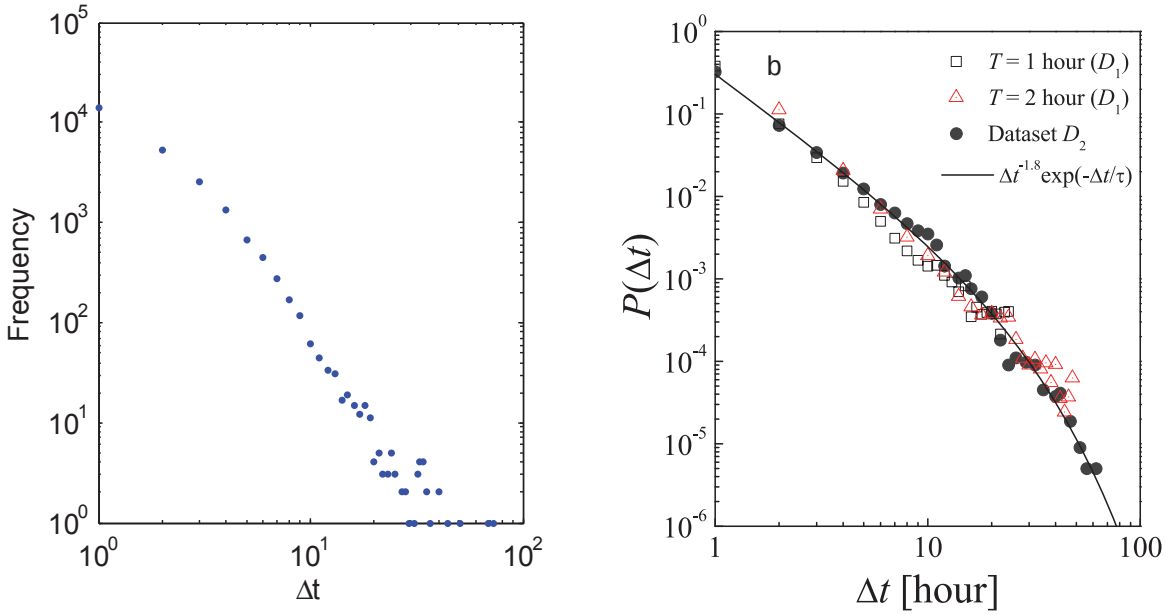
The estimation of pause times, Δt, is slightly more involved. In order to determine the probability of finding a pause time of length Δt, which needs to be filtered from the complete trajectory according to similar criteria as the displacements (that the location at a fixed time interval must be known for the events before and after), I must divide the number of those pause times by the probability of sampling an equally long interval in which events may occur at multiple locations, per the diagram and relationship below (adapted from Song 2010, Figure S3).

$$P_w(n_w) = P_{measure}(n_w)/P_{sample}(n_r). \qquad \text{(Equation S2)}$$



Moreover, the probability density function of the pause times follows an exponential distribution that contains the additional term of ***exp(-Δt/τ)***, where ***τ*** is the cutoff time of 17 hours in Song's case. The complete relationship is then $P(\Delta t) \sim \Delta t^{-1-\beta} \exp(-\Delta t/\tau)$.The distribution for this dataset is shown below, in comparison with the supplementary information Figure S4. The fitted parameter is ***β = 0.x***, versus Song's finding of ***β = 0.8.***

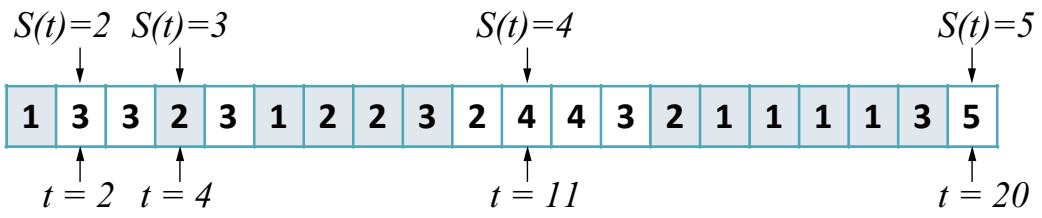
Shown below is the frequency of pauses of length n_w. I could not get the measurement of n_r to work in time, but the plot below is beginning to exhibit a similar distribution to Song's plod

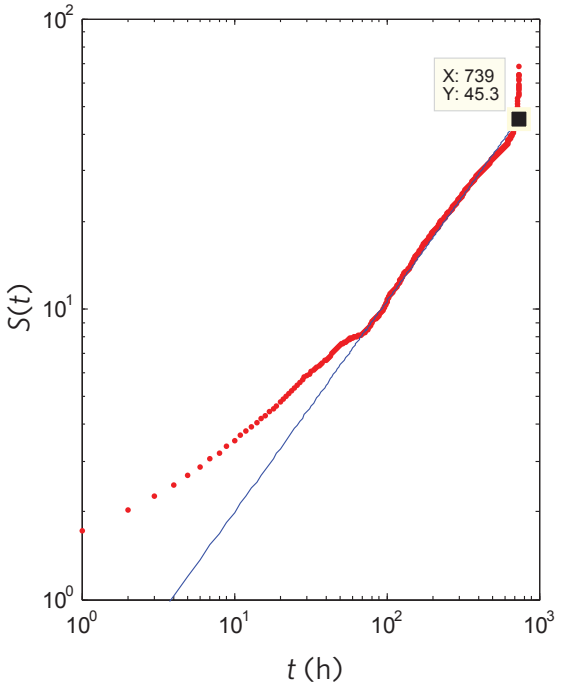


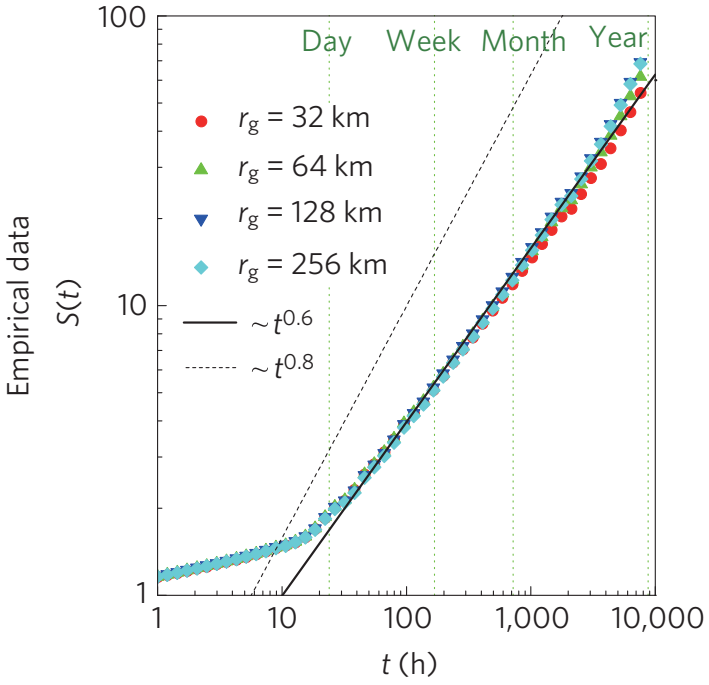
c. Confirm relationships $S(t) \sim t^{\mu}$ and $f_k \sim k^{-\zeta}$, estimate μ and ζ

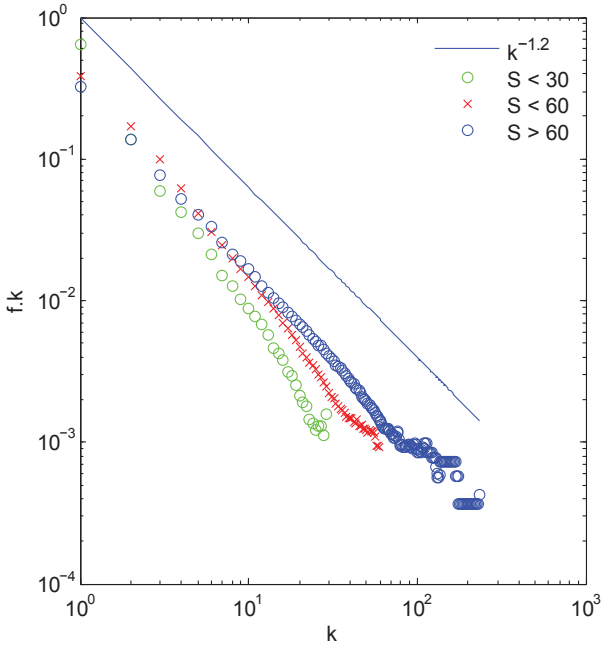
First I check that the number of new locations visited, S(t), has an exponential relationship with the time duration, $S(t) \sim t^{\mu}$. I generated a vector of the S_i value at a regular time interval of Δt = 2 hours, and then averaged these values for each S_i across all users. Note that this user set is not the same as the final set of users for estimating γ in Section 3a. By fitting the exponential curve, I obtained ***μ = 0.66***, which is fairly consistent with the value of 0.6 shown in Song's Figure 1a (see below).

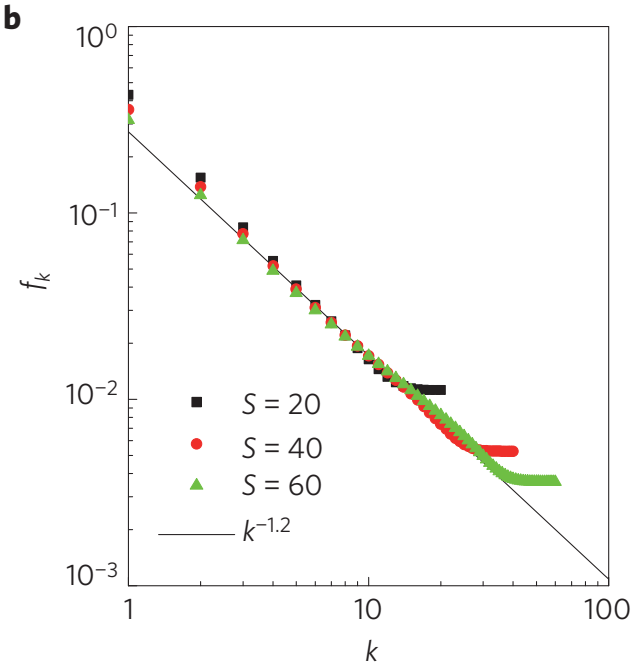
Next I generated the Zipf plot for the frequency of visitation of each location, $f^{-\zeta}$, versus the ordered rank of its first visitation. I did not have time to re-sample or bin the users within particular S regimes (i.e., $S_{max} \approx 30$, $S_{max} \approx 60$, etc.), but the plot below does show that the parameter ζ varies with S_{max}, which agrees with the finding of potentially different γ values for different regimes of S_{max} in Section 3a. The parameter value is similar to Song's, ζ ≈ 1.2. However, the article's Figure 1b plot also shows that the Song dataset may contain users for which S < 100, in which case my observations in Section 3a regarding the influence of users' different maximum S on the relationship $\Delta S/\langle \Delta S \rangle \sim S^{-\gamma}$ may be incorrect. The comparative plots are below.


S(t)=2
S(t)=3
S(t)=4
S(t)=5
1 3 3 2 3 1 2 2 3 2 4 4 3 2 1 1 1 1 3 5
t = 2
t = 4
t = 11
t = 20


X: 739
Y: 45.3
S(t)
t (h)


Empirical data
S(t)
Day
Week
Month
Year
r_g = 32 km
r_g = 64 km
r_g = 128 km
r_g = 256 km
~ $t^{0.6}$
~ $t^{0.8}$
t (h)


f.k
k
$k^{-1.2}$
S < 30
S < 60
S > 60


b
f_k
k
S = 20
S = 40
S = 60
$k^{-1.2}$

In the Supplementary information to the article, Song derives the relationships between the parameters (γ, β, μ, ζ) which follow equations (5) and (6) (Song 2010, 3), shown below.

The relationship between ζ and γ seems to dictate that γ must = 0.2 for ζ ≈ 1.2; however with γ ≈ 0.12 in my case, the steeper curves shown in the Zipf plot for $f^{-\zeta}$ do not suggest that ζ < 1.2. For the relationship μ = β / (1 + γ), my parameter estimates imply that ____ (could not estimate \beta in time). Therefore, there are some internal inconsistencies with the parameter estimate, which might be due to the CDR sample size.

$$\mu = \beta/(1+\gamma) \qquad \zeta = \begin{cases} 1+\gamma, & \gamma > 0 \\ 1-\rho, & \gamma = 0 \end{cases}$$

4. Model Results

If I had more time, I would set up the model generation function similar to the Levy flight function in Homework 3. The function would follow the procedure laid out in Section 2:

1. Accept inputs of total number of events to generate, model parameters (γ, α, β), and individual model parameters and return location probabilities (ρ, f).
2. Generate a pause time by selecting from the pause distribution P(Δt).
3. Generate a random number between 0 and 1 and compare this number to $P_{new} = \rho S^{-\gamma}$
4. If the number satisfies the probability of exploration, then select a distance from the jump distribution P(Δr) and a random direction for the move.
5. If the number instead falls in the range of the complementary probability of return to a previous location, $P_{ret} = 1 - \rho S^{-\gamma}$, select a return location from the probability distribution $\Pi_i = f_i$, where f_i is the frequency of each location's visits.
6. Repeat process for total number of events.

In the CDRs, there is not an explicit 'pause time' between each event at different locations, so I think the above model that inserts pause times before each move is acceptable but am not certain. I also intended to calculate the radius of gyration for each user, r_g, to compare the model results to the empirical data.

5. Concluding Remarks

I have been exposed to some important considerations when working with CDR data, as opposed to more continuous time and locational data. I have also surmised that the parameter values of the Song model can be quite variable, especially when the dataset is small. I will leave to future work the more rigorous analysis of

the relationship between mobility measures, such as the 'strength of exploration' ρ, and urban form variables that had initially prompted me to undertake this project. Following Yan Ji's thesis (2011), I would have liked to assign income and population predictors according to the user's home tower area. Then I would calculate certain destination distance predictors, 'distance from city center' and 'distance from local center', for each tower, as well as other urban form indicators that measure connectivity and accessibility. Based on a regression or clustering analysis, I would have liked to explore the usefulness of the parameters in Song's model compared with parameters proposed in other studies (cf. Kang 2011, Toole 2012).

References

Ji, Yan. 2011. "Understanding Human Mobility Patterns Through Mobile Phone Records: A Cross-Cultural Study". MIT.

Kang, C., X. Ma, D. Tong, and Y. Liu. 2011. "Intra-urban Human Mobility Patterns: An Urban Morphology Perspective." Physica A: Statistical Mechanics and Its Applications.

Song, Chaoming, Tal Koren, Pu Wang, and Albert-László Barabási. 2010. "Modelling the Scaling Properties of Human Mobility." Nature Physics 6 (10) (September 12): 818–823.

Toole, J. L., M. Ulm, M. C. González, and D. Bauer. 2012. "Inferring Land Use from Mobile Phone Activity." In Proceedings of the ACM SIGKDD International Workshop on Urban Computing, 1–8.