STATISTICAL COMPLEX ANALYSIS OF TAXI MOBILITY IN SAN FRANCISCO

Oleguer Sagarra Pascual

PhD supervisor: Albert Diaz-Guilera

Departament de Física Fonamental, Universitat de Barcelona. 08028 Barcelona, Spain.

Abstract

The recent developments in technology of movement tracking devices such as Global Positioning (GPS), together with the increasing availability of consistent data bases, have lately given rise to the study of human mobility patterns in different environments. In this work a statistical characterization of real mobility GPS high-frequency data from taxis in San Francisco is performed. The different patterns taxi drivers and customers follow are shown through comparing behavior when cabs are empty or full and the information is presented using a weighted directed complex network metric, from which the author obtains some topological information such as correlations between nodes, assortativity and clustering. Some adapted measurements to weighted nets are presented together with some remarks to support the need for new tools for assortativity classification.

1 Introduction

The study of the patterns and dynamics of mobility has always been subject of great interest in several disciplines such as sociology [1], ecology [2], urban planning [3], traffic forecasting or virus/disease spread among others. However, until recently the quantity and quality of data available to perform such studies was *limited* on one hand due to technological reasons (the position tracking and storing devices were inaccurate, not very accessible and with low capacity) and in the other due to practical issues (most data was obtained based on surveys or manually collected from sighting records in the case of animals). Nowadays, the GPS has improved in such a way that many studies can be performed with large mobility, high frequency databases taking advantage as well from the advance in the processing capacity of our computers [4]. Many of these past studies have been centered in studying the *home range* or use of space of animals using modified diffusion equations [5] and based mainly in the theory of Lévy Flights [6], others explore the emerging scaling properties of human mobility [7] while some work also has been done on explaining the influence of the environment in changes on mobility patterns [8].

In the present work the author aims to present a new way on studying home-range mobility based on a GPS Taxi traces database in San Francisco, USA.

The idea is to firstly perform a complete study of the raw available data without taking into account the specific conditions from where it was taken to detect interesting features and then relate those features with the particular boundary conditions of the problem at hand. This way many different mobility environments can be compared and the results of the studies become more universal [9,10].

The document is structured as follows: Firstly the technical details of the data are presented, followed by a quick statistical overview of its main characteristics. The text then follows to analyze more specific statistical features of the individuals involved in the study to present some insight in the different behavioral patterns observed. With the gathered information the author presents a *complex network* representation of the data. Finally some preliminary conclusions are drawn and many ideas for further work are presented.

2 The Data

2.1 General information

The data used on this study was obtained from [11][1] and consists of high frequency GPS data (updated at a high pace of $< t_{i+1} - t_i >= 90$ s) from a set of 537 taxis of the same company collected on the interval May-June 2008. The data provides latitude and longitude coordinates of the taxi together with a time reference (UNIX time since epoch) and an indicator of it being occupied or not[2]. Other information about whether the taxis are operated by the same driver, if

[1] The data is published [12], where it has been kindly uploaded under public license by its authors.

[2] The dataset was filtered and 4 *anomalous* taxis excluded from the study. The parameters used to do so were the number of trips[3] per taxi ($n \geq 40$) to detect anomalous behavior. In the appendix to the document the reader may find additional information and figures on the procedures used and explained through the text. Refer to *Preparing the Data* for further details on the data preparation.

	Δt^e	Δt^f	Δr^e	Δr^f
i_c	200’	43’	15 km	16 km

Table 1: Starting values of the decreasing tails defined as $\{i_c | P(i_c) = 0.1, i = \Delta t, \Delta r\}$.

environmental information is shared between workers via radio contact as well as if the trips originate from a call/taxi station or spontaneous user decision is not available. We will consider the taxis as statistically independent (as done and justified in [11]) and consider their drivers as skilled enough to have a similar knowledge of the city [13] and usage of GPS due to the lack of further *a priori* information. No precision ranges were given with the data but the accuracy of the integrated GPS devices on the taxi for 2-D data such as these is typically of about 10 meters[4].

2.2 Features overview

From the overall of the data, we obtained a total of 5×10^5 empty and full trips out of around 11×10^6 GPS updates which conform a good statistical ensemble to work on. In figure 1 we present the complementary cumulative (CCDF) and probability (pdf) distributions of elapsed times Δt_i and distances Δr_i covered by the taxis during all the observation time. From both the CCDF and their pdf’s one sees several interesting features: Firstly we observe a noisy initial range $\mathcal{O} \sim [1, 10]$ (for meters and seconds) easily explained by the fact that these are not typical ranges of usage of a taxi, neither for distances nor for times. Aditionally we see slow decaying tails that extend to very large values, we also see that in the case of the distances the forms of the distributions seem to be quite similar[5]. Finally it is worthy to observe a protuberance in the middle of the tail at around the interval $[10 - 20]$ km for distances that may account for the trips to the airport (which is 15 km apart from the city approximately) and is a frequent destination as we will see in section §4.4. Also note that the distributions seem to be flat (flatter in the case of empty trips) with a decay in a fat tail behavior starting at values r_c, t_c that define a typical maximum distance (time) of taxi usage as shown in table 1 (most likely the typical time/distance spent in an *in-city* trip)[6].

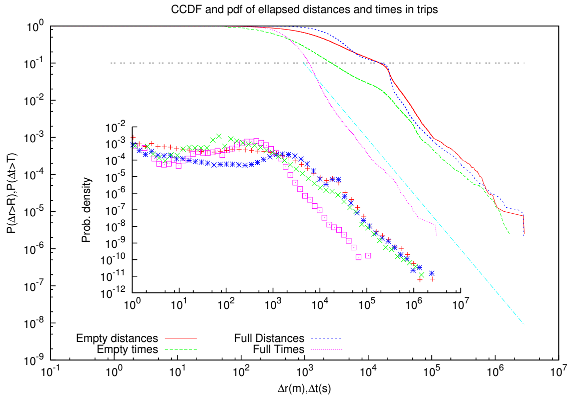


Figure 1: Probability density distribution and CCDF of elapsed times and distances for all the taxis with logarithmic binning. Note the slow decreasing tails. The light blue line is a guide for the eye with fitted exponent $\alpha = 2.030 \pm 0.002$ from the CCDF of full times, which was the only with statistically relevant results. For more details on the fits see the appendix.

3 Statistical Analysis

The aim of this section is to provide some insight in the individual behavior patterns for the individual taxis. As a first approach to do so we fitted the tails of the individual distributions of Δt, Δr for each taxi to detect strange patterns but all the data had power law tails with *similar* exponents[7]. Then we proceeded to study the efficiency of each taxi defined as,

$$v_e = \frac{\Delta r_e}{\Delta t_e} \quad v_f = \frac{\Delta r_f}{\Delta t_f}$$

$$\rho_r = \frac{\Delta r_f}{\Delta r_{total}} \quad \rho_t = \frac{\Delta t_f}{\Delta t_{total}} \quad \rho_v = \frac{v_e}{v_f}. \tag{1}$$

We show the results on table 2[8]. From the values obtained and the *peak* shaped distributions, we observe no substantial differences in the efficiency of the taxis, fact that provides an hypothesis for all of them showing similar behavior. Also it is noteworthy that in mean, half of the distance the taxis move, they do it occupied while they spend one third of the total time on duty. Interestingly enough, they all seem to have bigger typical speeds when full and this feature has a bigger spread, fact probably caused by the bias introduced by resting times (counted as in-time empty trips) on the data treatment and caused by different resting patterns of the drivers.

We also computed the center of masses of the movements for the different taxis $\vec{r}_{CM}$ (mean position over time updates) and compared the spread of the movements using the gyration radius defined as,

$$r_{gyr}^2 = \frac{1}{N} \sum_i^N |\vec{r}_i(t) - \vec{r}_{CM}|^2.$$

Where each values r_i represent a GPS location update. The results are shown in figure 2. We observe that the

[4] This information can be obtained from top retailer GPS tracking devices website.

[5] We shall study the correlation of long trips in section §3.

[6] Please note those are not the values at which the fits start, only where the decreasing trend does.

[7] Taking into account we are dealing with real data, with a certain amount of noise. Refer to appendix for further details.

[8] To see the associated distributions refer to the appendix.

	r	t	v
$< \rho >$	0.56 ± 0.04	0.31 ± 0.06	0.359 ± 0.102

Table 2: Efficiency mean values and associated standard deviation. $\rho_r > \rho_t$ is explained by the fact that drivers spend some waiting times looking for customers in taxi stops such as the airport. As for the relative speeds, one sees that taxis tend to move faster when full.

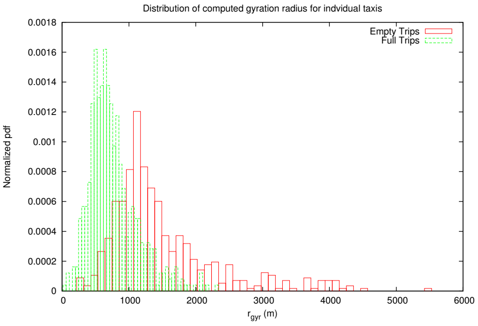


Figure 2: Gyration Radius distribution for empty and full situations.

spread in the reference system for each taxi is smaller whenever they are occupied ($< r^f_{gyr} >= 769 \pm 369$ m), situation that is explained by the fact that the destination in these cases is known an set by the customer, resulting in direct trips unlike the empty situation ($< r^e_{gyr} >= 1500 \pm 800$ m) where drivers tend to wander in search of customers.

Finally we focused our attention on the tails of the distribution (the long range trips) given the similar form that seems to exist between the tails of the empty/full distance distributions shown in figure 1. First we computed the distribution of maximum range trips and observed that a high overlap in the ranges of empty/full maximum distances exist, and that the long range full most lengthy trips were longer than the empty ones.

These hints directed us to study the correlation between the successive long trips (sequences of full/empty trips with a small intermediate time $\Delta\tau <$ 3 hours[9] and lengths $\Delta r > 20$ km) and successive short trips ($\Delta r < 5$ km and $\Delta\tau < 10$ minutes). We calculated for each taxi the Pearson correlation coefficient of the pairs of trips and their p-values[10] and averaged over all the taxis to obtain $C_{long} = 0.45 \pm 0.21$, $C_{short} = 0.11 \pm 0.006$. These differences in correlation of length of successive trips are explained by the fact that when taxi drivers drive away from the city, they do not have the legal right to pick up customers and so are forced to retrace their way back (thus having successive trips of *similar* length). The opposite situation (operating inside the city) does not generate correlation since this constraint no longer exists and successive trips can be considered independent.

The taxis are considered independent and do not seem to show significant differences in their efficiency parameters, spread around the city (r_{gyr}), mean position (CM). Moreover, the correlation existing in long trips indicate that the observed tails in the distributions of distances are caused mainly by the choice of destination of customers, and hence that the fat tails are explained by the heterogeneity of customers (behavior when full) and not searching strategies (empty taxis). This correlation also explains the similitude in the distance tails, whereas in the case of times the discrepancy is not explained[11]. Finally, the bumps present in the distributions seem to indicate the presence of (one or several) important destinations at an important distance from the city. Some of the overall facts presented here should be present and explained by the network built in section §4.

4 Complex Network Approach

In this second part a representation of the data using a complex network approach is shown. Firstly the building procedure of the network is quickly explained together with some general features of it listed and finally more involved calculations as well as some conclusions are drawn from it.

4.1 General Overview

The network is built from the trips present in the data, where nodes correspond to locations (starting or ending ride points) discretized[12] using a grid with bins of 100×100 m of surface that are connected with weighted edges that represent the number of trips linking each pair of locations. The net is directed (because it is in no way symmetric) and weighted (with the number of trips)[13] and *selfloops* as well as *isolated nodes* have been trimmed. The main features of the net are presented in table 3 and its distribution of strengths and degrees. We observe that the net representing empty trips is *denser*[14] than the full

[9] The time condition was added to avoid data discontinuities using non-related trips.

[10] Using the usual definition as found for instance in [14]. See appendix.

[11] The study of time intervals is inconclusive on one hand due to the definition of trip used and on the other because even though the constraint for long rides exists for distances, in the case of times this fact is much more difficult to study due to resting times not accounted for.

[12] Some noise coming from the discretization will be present on the results, but we cannot apply a non-constant metric on the grid without applying an *a priori* bias on the results, i.e. 1 km^2 may contain more information in the city center than in the outskirts of San Francisco but we are forced to make no assumptions on our analysis, so we adopted a constant grid (also for simplicity in our algorithm).

[13] Please refer to the appendix for a complete justification and the algorithm used to construct the net as well as for the distributions mentioned in the document and not showed.

[14] With a similar number of nodes, it has 80% less links.

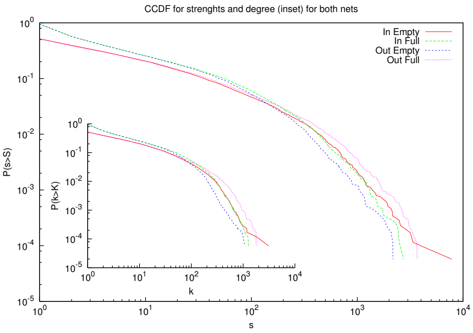
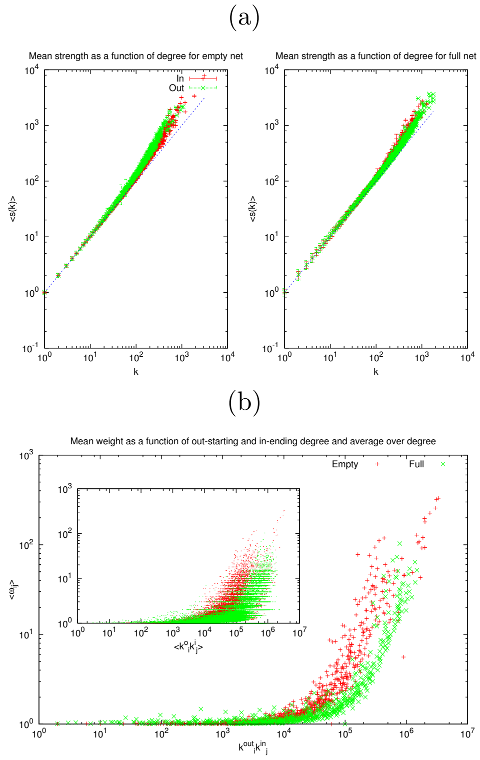


Figure 3: Strength and degree distribution (inset). We see the similarity between them and its fat tailed nature, both seem to have a functional form of $P(s), P(k) \sim k^{-\gamma} f(k/k_x)$ truncated power laws due to the finite size nature of the nets at study in accordance with [15].

one. Finally, the weight distribution decreases faster for the full net (indicating more relative importance of most used trips for the empty one). These features are explained by the fact that normally destinations for customers tend to be more diverse (home, work) and heterogeneous, hence more edges pointing to different places are present in the full net (greater strength out, greater density, more edges). In figure 3 we show the distribution of both strengths and degrees for the nets at study, we observe similar behavior for the complementary features (in /out) of both nets (empty/full) but this is an expected behavior emerging from the definition of trip mentioned earlier.


(a)

(b)

Figure 4: (a) Computed function $< s > (k)$. The straight lines serve as guide to the eye, we observe that the approximation $\omega_{ij} \equiv < \omega >$ is lost for high values of k. (b) Computed mean weight of edges as a function of starting and ending unweighted degrees, the inset is the raw data while the main plot corresponds to the mean over s. We observe again a general lack of correlations for several decades with a sudden increase for high values of $k_i^{(o)} k_j^{(i)}$.

4.2 Studying correlations

Once we have taken a look at the general features of the graphs, we want to further explore the correlations existing in the nets, to do so, we wish to study the relation between the strength and weights of edges emerging from nodes: Were no correlation present between them, we would obtain [15] $s_i =< \omega > k_i$ using the approximation $\omega_{ij} \equiv < \omega >$. In figure 4 we see the plot of this function and observe that whereas the approximation is valid for low values of k, the degree of nodes increases slower than its strength as values increase, thus showing that as nodes become more connected, they become more and more frequent (important). The second plot in figure 4 shows yet another interesting behavior, with the weight of edges being in general independent of the nodes they connect except for big values of $k_i^{(o)} k_j^{(i)}$, where we see the tendency of high connected nodes in the net being linked by frequent trips (heavy weighted edges) among themselves (specially in the case of the full net, where the pattern is clearer). Important places (*hot spots*) tend to be connected among themselves because people need to move between them (full net) and so are a frequent destination for taxi drivers in their search of customers (empty net)[15].

Finally we also computed the betweenness centrality of all the nodes in the net and averaged it over the nodes with the same values of strength as shown in figure 5 and found a power-law like behavior as done

	N	E	ρ	$< k^i >$	$< k^o >$	$< s^i >$	$< s^o >$	α_ω	$< \omega_{ij} >$
Empty	17465	278891	9.1×10^{-4}	15 ± 62	15 ± 43	22 ± 118	22 ± 79	2.125 ± 0.005	1.4
Full	17511	348166	1.1×10^{-4}	20 ± 60	20 ± 81	26 ± 98	26 ± 130	2.660 ± 0.003	1.1

Table 3: General features of the net. Refer to the appendix for the exact definitions of the parameters used together with the references for algorithms used in computing the values. Observe the conservation for both strengths and degrees, with different σ due to slight differences in their distribution tails. Finally α_ω refers to the fit of the weight distribution as a power law, the only consistent fit available (see appendix).

[15] Both nets are likely to be highly correlated and show similar patterns for most features, but an in-depth study of the correlations between them is out of the scope of this work.

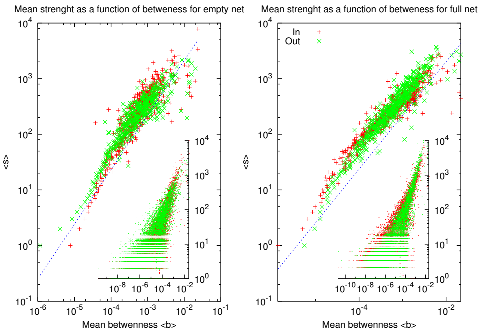


Figure 5: Mean strength as a function of betweenness $< s > (b)$, the inset corresponds to the raw data while the main plot is smoothed over repeated values of b.

in [15]. The betweenness definition in our networks give an idea of the routing habits of taxi drivers, since assuming the taxi drivers want to maximize their efficiency, they might want to cover all the important zones of the city concatenating occupied trips, and so high values of betweenness indicate important spots (places with high density of customers) that are well connected (many weighted short paths[16]) and thus distribute the transit over the different regions of the map.

From the features studied in this section, we can conclude the very important influence the top nodes of the net accumulate, on one hand due to the fact that the more connected they become in degree k, the more strength s they gather, and the more strength (traffic) they gain, the more central (b) they become, as well as the better connected with other important nodes $< \omega_{ij} > (k_i k_j)$ they get.

It needs to be noted that the analysis of correlations in this network is pretty simple and more advanced studies on weighted networks could have been performed [16,17,18], but the facts presented here allow us to indistinguishably use s or k as variables in our mean field approximation of the network, just simplifying our posterior assortativity analysis[17].

4.3 Assortativity: Pearson r coefficients and neighbor connectivity

To finish our study of the main characteristics of the net prior to its representation, we studied the assortativity of our networks. To do so we computed both the Pearson assortativity node strength coefficient r_s[18] and the mean neighbor degree $k_{nn}(k)$ to look for the linking nature of both nets. We need to be aware that since we are working with a directed network (which is highly non-reciprocal) we can compute several versions of the coefficient and neighbor degree (strength), all containing different information[19]. In table 4 we present the values obtained from the computation of the Pearson coefficient, and we observe some surprising differences. Whilst the full net seems to have an important disassortative character, the empty net gives non-conclusive results (with low values nevertheless) so from this preliminary analysis we need to conclude a general lack of assortativity in the empty network. To evaluate the effect superrich nodes may have on this statistic [22] due to the correlations existing between them seen in §4.2, we computed the evolution of the r_i coefficients recursively extracting the top weighted in and out degree nodes of the net as showed in figure 6, where we also show the relative size evolution of the biggest weakly and strongly connected components of the network.

	r_{in-in}	$r_{out-out}$	r_{out-in}	r_{in-out}	$< r >$
Empty s	-0.032847	0.017782	-0.054016	0.045915	-0.018 ± 0.043
Full s	-0.079386	-0.093057	-0.116725	-0.057891	-0.10 ± 0.03

Table 4: Pearson r coefficient for degrees in both networks. Please note that the minimum values of r are bounded [21].

We observe that as we exclude the top nodes, the networks' r coefficients evolve into positive values, fact that would indicate assortative nature, i.e. a tendency of nodes of similar (or increasing) degree to link. Moreover, the sudden change in the r coefficient indicates the strong tendency of superrich nodes to link with scattered destinations (very low connected nodes).

Accounting for the size of the biggest components, it is important to note that the nets are percolated with their giant component[20] occupying roughly the 98% of the network in both cases. As we exclude the very first top nodes, we see that the size of the giant component in the case of the empty net is more vulnerable, fact that would indicate a more assortative nature than the full case[21]. Moreover, we observe that the size of the strongly connected component is resilient to extraction of top nodes, fact that indicates a strong inter-connectivity for medium sized nodes (mostly points inside the city center).

To further clarify the assortativity of the net we present on figure 7 the mean neighbor degree of the networks at hand[22] in figure 7.

[16] Computed using Dijkstra's algorithm.

[17] In fact, all the figures shown have been computed for both s and k, refer to the appendix for the ones not appearing in the present document.

[18] Adapted to weighted nets from the definition in [19] and [20].

[19] *In-In* indicates *In* degree of node out-going neighbors related with *In* degree of parent node for instance. See the appendix for an explanation of the different meanings in the coefficients or [20].

[20] The weakly biggest component (or giant component) of our network is the subgraph of nodes that are connected between themselves via a path composed of directed or undirected links.

[21] As shown in [21] for different theoretical models. Albeit since we are dealing with real data, the pattern observed is much less extreme.

[22] We computed them at two stages, both with the complete nets and with the nets trimmed with 52 nodes, roughly 0.3% of

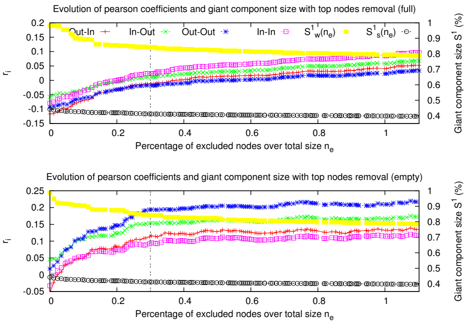
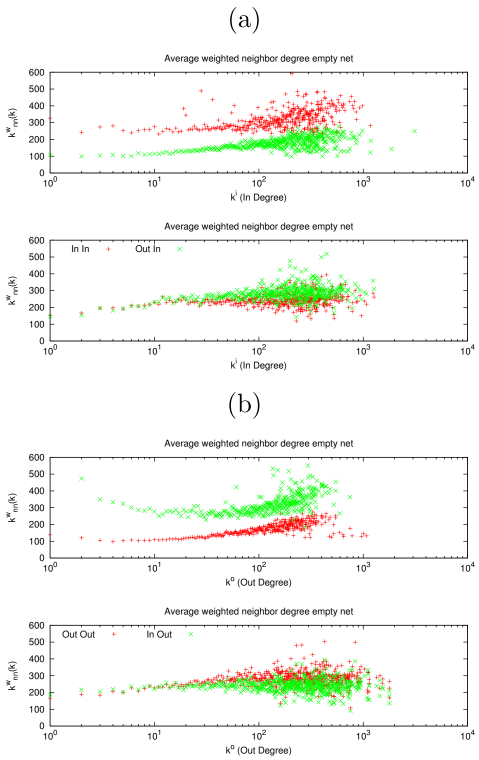


Figure 6: Evolution of r_i coefficients for both nets and of the size of the giant computing while top nodes are being removed, as a function of the percentage of nodes removed. We observe the disassortative negative values are quickly lost due to exclusion of main nodes and the assortative nature of the nets is strengthened as we exclude more and more nodes.

We start taking a look at the full net. It is important to note that in this case all the coefficients carry similar information, as important places are both a preferred destination and departure points for customers, and random spots in the streets (i.e. stoping a taxi *on the move*) or scattered locations such as workplaces behave the opposite way. We do not observe clear patterns, with nodes qualitatively seeming to show a faint assortative tendency.

The empty net seems to have a much more complicated pattern. Not all the pairs show the same information nor the same tendency. The clearer pattern is shown by the $In - Out$ pairs (U shape), where low out-degree nodes (occupied trips ending in a scattered locations) have a very strong tendency to link to high in-degree spots (taxi spots, *hot spots* for customers) that dramatically falls to a constant tendency of medium sized nodes ($k \sim \mathcal{O}(10)$) to increasingly link with similar nodes, ending in a pattern for busy spots tending to connect among themselves. A similar, less-extreme shape is observed for the $In - In$ pairs. Finally, both $Out - Out$ and $Out - In$ show a faint assortative tendency (in this case, drivers ending a trip do not want to visit places where they will not likely embark customers, hence they do not choose low out degree locations).

Our assortativity analysis is inconclusive due to the fact that the two indicators used appear to give different results due to the influence of the most important nodes in the net, evidence supported by the strong correlations detected for the supernodes of the net in previous sections §4.3. Moreover, the mix of linking behaviors (assortative and disassortative) present in the complex structural nature of the nets call for the need of a local based assortative analysis [[23]]. All this facts suggest that some additional tools need to be developed to study and better quantify the linking trends present in directed (weighted) networks.


(a)

(b)

Figure 7: Mean weighted degree of neighbor nodes in semi logarithmic scale. (a) In-In and Out-Out strength pairs. (b) In-Out and Out-In strength pairs. We observe an overall assortative-disassortative tendency for the full net while having an *s shaped* effect and a mixed pattern for the empty net.

4.4 Network Representations

For practical reasons we restrict our representation to the empty net[23], that shows the behavioral patterns of the taxi drivers. We have kept the giant component of the net with edges of $\omega_{ij} > 5$ (trimming out scattered, seldomly visited locations) and performed a modularity analysis on it [[24]] that successfully identified groups of communities. We observe in figure 8 the overall net with several clear clustering patterns: Taxi drivers tend to go to nearest local important nodes to look for customers (hence the successful clustering and the assortativity for medium-connected nodes), otherwise, they direct themselves (travelling longer distances) to the top nodes (disassortativity) that greatly influence our network and are well connected between themselves. In the figure we also see the top influence of the airport on the overall net, fact that explains the bumps found on the distributions in section §3. This

its total number. Besides changes for low values of the graphs, no substantial changes were observed possibly due to the qualitatively nature of the analysis. Refer to the appendix.

[23] More representations are available in the appendix.

sort of behavior is clearly marked by the knowledge of the system the drivers have, and is reminiscent of a truncated Lévy Flight searching strategy.

5 Preliminary conclusions and future work

We have firstly trimmed and studied individual taxi real data and found no sharp differences between the different taxi driver behavior patterns. Similar distribution shapes were obtained for the different taxis in times, distances and efficiency parameters, although no scaling was attained. A more pronounced dispersion was detected for empty taxis through the study of its radius of gyration r_{gyr} and a strong correlation for long empty and full trips caused by legal constraints was detected, that explain the similitude of the fat tails observed for the data of empty and full situations, governed by the heterogeneity of destinations chosen by customers[24]. To conclude the statistical analysis we also detected a frequent destination at long distance of the city, which would later correspond to a top node of the network representation performed.

In this sense we represented our data in two complex weighted directed networks and computed the main characteristics of the nets (power law distribution of weights, fat tailed in and out weighted degree distribution) as well as the patterns of the correlations existing between the nodes' betweenness, strength and degree. Finally an assortativity analysis was performed with contradictory results that showed the difficult linking nature of the net, the strong influence of superrich nodes on the statistics of the net (rising from the correlations detected and the shape of the distributions of weight and degree) and the need for local-based assortativity analysis tools.

To end this work some representations of the net of empty taxi trips are shown exhibiting a clustered nature for frequent movements attached to delimited geographical zones, indicating the tendency of drivers to maximize their occupancy efficiency by visiting their nearest busy spots after each run.

The author considers that the results of this study mainly open up three ways for further research: On one hand, the introduction of a weighted directed net forced the author to adapt some standard measures to these type of nets, and such measures should be refined and its efficiency further studied (specially in the case of assortative characterization, where we obtained many inconclusive results that call for a better defined local approach in its study[25]). On the other hand, if more Taxi mobility data were available on other cities, the properties studied here could be com-

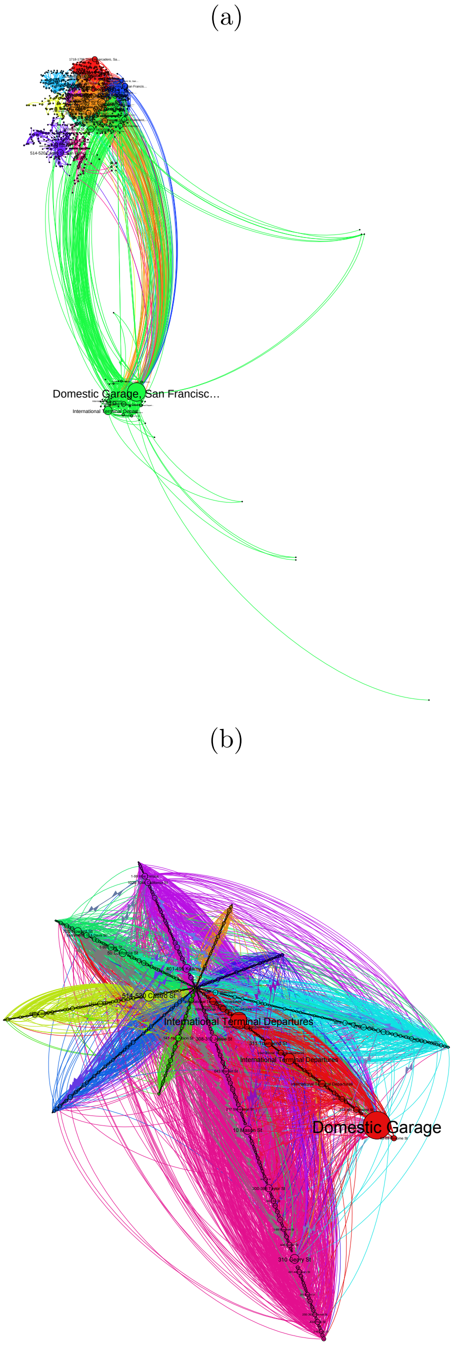

(a)
Domestic Garage, San Francisc...
(b)
International Terminal Departures
International Terminal Departures
Domestic Garage

Figure 8: Empty Net representations. (a) Geographic location of node with edges with $w_{ij} \geq 5$ with their size referred to their value of s^{in} and labelled by address. We observe a clear neighborhood clustering pattern with both airports (SFK and Oakland) far from the city. (b) Radial layout with the r coordinate representing decreasing values of mean weighted neighbor out degree k^w_{nn} and the nodesizes their incoming degree k^i. We observe that the clustering is broken by connections between top nodes and the general disassortativity, with big nodes present towards big radial values in each cluster. The node labelled *Domestic Garage* corresponds to the drivers resting area in the airport.

[24] Please note that some of the results obtained here coincide with very recent studies performed in mobility habits of population based on their personal car GPS data usage [4].

[25] Adapting the methodology in [25] to weighted networks for instance.

pared in the search for universal patterns (or particular divergences). Also a study of the nets' growth over time could be carried with additional data. Finally, the results of this analysis open the way for a model to be constructed on taxi behavior that via simulation validates the results obtained, and in the event of obtaining a realistic model optimization in the searching strategies could be devised.

To conclude, it needs to be stated that the methodology followed provides a good example of the power of complexity and data mining strategies to both study and represent without preliminary information a big set of data and obtain relevant information and patterns, as well as providing tools to compare different sets of similar data.

The author would like to thank its tutor for the bibliography and remarks provided as well as the developer community of the open-source package `NetworkX` for its useful programming tips and work in maintaining such a powerful tool in open access.

References

1. Horner, M. W. and O'Kelly, M. E. *Journal of Transport Geography* **9**(4), 255–265 (2001).

2. Horne, J. S., Garton, E. O., Krone, S. M., and Lewis, J. S. *Ecology* **88**(9), 2354–63 September (2007).

3. Eubank, S. *Nature* **429**, 180–184 (2004).

4. Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., and Trasarti, R. *The VLDB Journal* July (2011).

5. Benhamou, S. *PloS one* **6**(1), e14592 January (2011).

6. Raposo, E. P., Buldyrev, S. V., da Luz, M. G. E., Viswanathan, G. M., and Stanley, H. E. *J. Phys. A: Math. Theor.* **42**(43), 434003 October (2009).

7. Gonzalez, M., Hidalgo, C. a. C., Barabási, A. A.-L., and González, M. C. *Nature* **453**(7196), 779–782 June (2008).

8. Humphries, N. E., Queiroz, N., Dyer, J. R. M., Pade, N. G., Musyl, M. K., Schaefer, K. M., Fuller, D. W., Brunnschweiler, J. M., Doyle, T. K., Houghton, J. D. R., Hays, G. C., Jones, C. S., Noble, L. R., Wearmouth, V. J., Southall, E. J., and Sims, D. W. *Nature* **465**(09116), 1066–1069 June (2010).

9. Isella, L., Stehlé, J., Barrat, A., Cattuto, C., Pinton, J.-F., and Broeck, W. V. D. *J. Theor. Biol.* **271**, 166–180 (2010).

10. Isella, L., Romano, M., Barrat, A., Cattuto, C., Colizza, V., Van Den Broeck, W., Gesualdo, F., Pandolfi, E., Ravà, L., Rizzo, C., and Tozzi, A. E. *PLoS ONE* **6**(2), 10 (2011).

11. Piorkowski, M., Sarafijanovic-Djukic, N., Grossglauser, M., and Piorkowski M. Sarafijanovic-Djukic N., G. M. *2009 First International Communication Systems and Networks and Workshops* , 1–10 January (2009).

12. Kotz, D., Henderson, T., and Abyzov, I. (2004).

13. Girardin, F., Blat, J., and Fabien Girardin, J. B. *Pervasive Mob. Comput.* **6**(4), 424–434 August (2010).

14. Rodgers, J. L. and Nicewander, W. A. *American Statistician* **42**(1), 59–66 (1988).

15. Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. *PNAS* **101**(11), 3747–3752 March (2004).

16. Serrano, M. A., Boguna, M., and Pastor-Satorras, R. *Physical Review E - Statistical, Nonlinear and Soft Matter Physics* **74**(5 Pt 2), 055101 (2006).

17. Newman, M. E. J. *Phys. Rev. E* **70**(5), 9 (2004).

18. Barthelemy, M., Barrat, A., Pastor-Satorras, R., and Vespignani, A. *Physica A* **346**(1-2), 34–43 (2004).

19. Newman, M. *Phys. Rev. Lett.* **89**(20), 1–4 October (2002).

20. Foster, J. G., Foster, D. V., Grassberger, P., and Paczuski, M. *PNAS* **107**(24), 10815–20 June (2010).

21. Newman, M. *Phys. Rev. E* **67**(2), 1–13 February (2003).

22. Xu, X.-K., Zhang, J., Sun, J., and Small, M. *Phys. Rev. E* **80**(5), 1–7 November (2009).

23. Piraveenan, M., Prokopenko, M., and Zomaya, A. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* (October 2009) August (2010).

24. Newman, M. E. J. *PNAS* **103**(23), 8577–8582 (2006).

25. Serrano, M. A., Maguitman, A., Boguñá, M., Fortunato, S., and Vespignani, A. *ACM Trans Web* **1**(2) (2007).

STATISTICAL COMPLEX ANALYSIS OF TAXI MOBILITY IN SAN FRANCISCO: Additional information

Oleguer Sagarra Pascual

PhD supervisor: Albert Diaz-Guilera

Departament de Física Fonamental, Universitat de Barcelona. 08028 Barcelona, Spain.

Abstract

This documents presents the additional information that was not added to the main document but is needed to back up the facts there presented. It contains important information about the procedures followed, complementary plots, maps and figures as well as statistical methods and algorithms used.

1 Preparing the Data:

The data available was obtained from several `txt` files (one per each taxi) in a 4 column format, where the first and second columns referred to the latitude/longitude positioning of the taxi (in decimal degrees), the third to the taxi occupancy (0 for empty, 1 for full) and the last one showed the `UNIX` time since epoch.

Our definition of trip includes the group of successive data rows between a change on the taxi occupancy parameter (from empty to full or from full to empty). The time spent in each trip was computed from direct sum of the block of values contained in each occupancy (referred to the original starting time of the movement)[1]. Accounting for the distance, we used the `python` module `geopy` [1] that uses Vicenty's formulae to compute the geodesical distance[2] between two geodesical points with an accuracy of about 0.05% [2] and iterated for all the values in each *trip-block*. The metric used was euclidean due to the fact that the frequency of the data was high enough to avoid considering the use of other metrics such as the Manhattan (also called TaxiCab) metrics[3] and the ellipsoid used to do the conversion was the standard WGS-84 (as supplied in the data).

Finally, to avoid non-consistent taxis, we excluded 4 taxis with a low level of performed trips n ($n < n_{max} = 40$) or with a high level of zero-length percentage trips (in time or space)[4] $0.1 \leq \nu \leq 0.9$. The overall data accounted 464216 (464006) empty (full) trips from a total of 11257922 GPS updates.

	Empty	Full	Total
Trips	872 ± 285	872 ± 285	1744 ± 571
ν_t	$0.14 \pm 0.05\%$	0.02 ± 0.02	0.08 ± 0.03
ν_r	$0.14 \pm 0.05\%$	0.02 ± 0.02	0.08 ± 0.03

Table 1: Mean values and standard derivation of number of trips and zero-length percentages. We observe their values are equal within the error observed.

2 Statistical methods

Some comments are in order to account for the statistical features used. Regarding the binning of the data, all the cumulatives have been computed without binning and directly sorting the data (taking into account repeated results), while the histograms have been produced either with a normal binning or logarithmic in the case of strongly skewed distributions (presence of fat tails) and this circumstance is always stated on the text. The number of bins used is normally chosen with a thumb rule or via trial and error considering the value that smoothes the noise while capturing the main details of the distributions (but never taking bins that imply less than 20 data values on average). Please note all the tools used in such analysis have been the `python` standard `scipy` and `numpy` modules [4].

As for the fitting of the power laws, all fits (unless otherwise stated) have been performed using maximum likelihood methods described in [5] with the `python` package `plfit` [6].

Finally, all the mean values presented in the text are normally accompanied by an error equal to the standard derivation of the sample obtained from the second moment of the distribution (whenever its dis-

[1] Please note that the data available is gathered while taxis are on duty, and hence off-working times are not counted. This fact is not mentioned with the information provided with the data, but taking a look at the time intervals between trips one sees that just a very small minority are longer than *6* hours, which wouldn't be the case otherwise.

[2] Straight distance over the ellipsoid.

[3] As performed in [3] using the same dataset.

[4] I. e. number of 0 length trips over total, $\nu \equiv \frac{n_0}{n}$.

	$\Delta r_{min}, \Delta t_{min}$	$n(> \Delta r_{min}, \Delta t_{min})$	$< \alpha_r, t >$	L
Δt^e	3641	1409	3.11 ± 0.05	-12575.9
Δt^f	884	121705	2.030 ± 0.003	-1.06×10^6
Δr^e	24643.3	13093	3.35 ± 0.02	-139820
Δr^f	251444	107	2.7 ± 0.2	-1445.4

Table 2: Fit details of the overall distributions of time and distances.

tribution is *smooth enough*[5]) and the same holds for the presence of errorbars in the figures.

2.1 Total time and distance distribution fits

In table 1 we show the details of the fit of the total accumulated data together with the associated error and Likelihood measurement[6]. The only tail with a likely *pure power law* behavior is the one corresponding to full times and possibly the empty distances tail, while the two others resemble more a sort of *logbrownian* distribution. Anyhow, the number of decades is not consistent enough to say anything further than the fact that the tails are fat.

2.2 Time and distance individual distributions

In our hope to observe some differences among the different taxis, we made a preliminary step by trying to fit the tails of each of the distributions by a power law and we present the results in figure 1. Please note that those fits are just orientative and although some differences are observed among exponents, we see that most of the distributions seem to have similar values for the exponents using a consistent number of data for the fits (as seen in the inset figure). The tails start in a minimum range of $\Delta r_{min}, \Delta t_{min} \sim 500$ m,s up to $\Delta r_{min} \Delta t_{min} \sim 4000$ m,s for the shorter tails of distances (times). We observe as well that for the times those distributions are much more flattened, fact that indicates more dispersion (heterogeneity of the tails) and could explain the dispersion observed in the differences on velocity efficiency seen in §2.3.

To further explore differences between taxis, in figure 2 we show the pdf figures of all the taxis for both distances and times. We observe several interesting features, we observe dispersion more pronounced in the case of the times and an initial noisy part (which is not interesting since this is not the main usage of taxis). We also see two bumped zones, the initial one whose end marks the typical distance (time) of usage of a taxi in accordance with the results found in the main article and a second one that in the case of distances is most probably explained by the presence of the airport at an important distance of the city, that acts as a hub (as we will see further on). Although a scaling relation seemed feasible, a procedure based in the ideas presented in [7] with a parameter r_g[7] was tried to no satisfactory result.

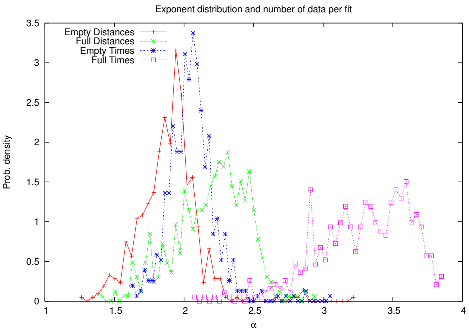


Figure 1: Distribution of fitted exponents for each taxi.

2.3 Taxi Efficiency in movement and time

We present in figure 3 the distributions from which the mean and standard deviation of taxi efficiency parameters presented in the main document have been derived, which support the hypothesis that mostly all the taxi drivers obtain similar productivity levels.

2.4 Correlation and Pearson p-values for long trips

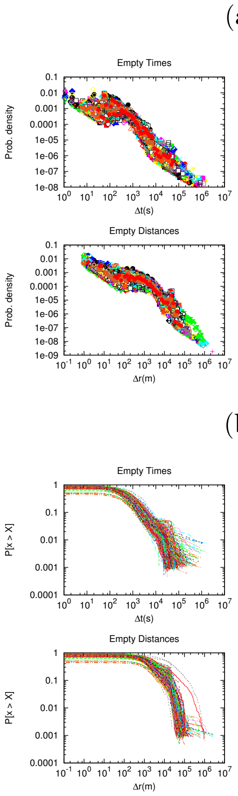
To check our hypothesis about correlation in long trips, an analysis of the extreme statistics of the data ensemble was performed and the maximum Δr and Δt range values for each taxi were gathered and counted as shown in figure 4. We observe a new hint about the behavior of the taxis, while their number of long trips seem to be quite similar, the waiting times / traveling times are radically different, fact that emerges from the definition of trip used (the resting *on duty* times are included in empty trips) together with the fact that while full, the taxis do have a preferred destination and hence move *quicker*. The second plot in figure 4 show the correlation pearson value and their accuracy presented as its *p-value* for successive occupied-unoccupied trip pairs. We observe big differences in the correlation behavior for long/short trips, fact that confirms the strong influence that long trips perform on the immediate search strategy for the drivers. The computation of the mean correlation values has been performed counting successive full/empty trips and excluding the ones with a significance level lesser than 5% for the full trips (no trimming for empty trips)[8].

[5] Which means with a concentrated enough *gaussian* form so that the second moment has any statistical sense as error.

[6] For details see [5].

[7] See *Statistical Analysis* in the main document.

[8] See the *p-value* distributions, that are concentrated for the full trips but not in the other case. It needs to be noted that we do not expect extremely significant correlation values, as



(a)

(b)

Figure 2: (a) Total histograms (normalized) with logarithmic binning of the distributions. We observe similar shapes and the usual bumped zone mentioned in the text, (b) CCDF from the raw data.

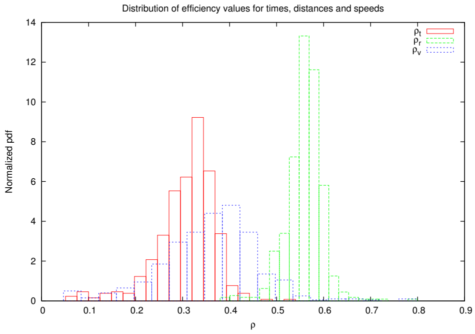
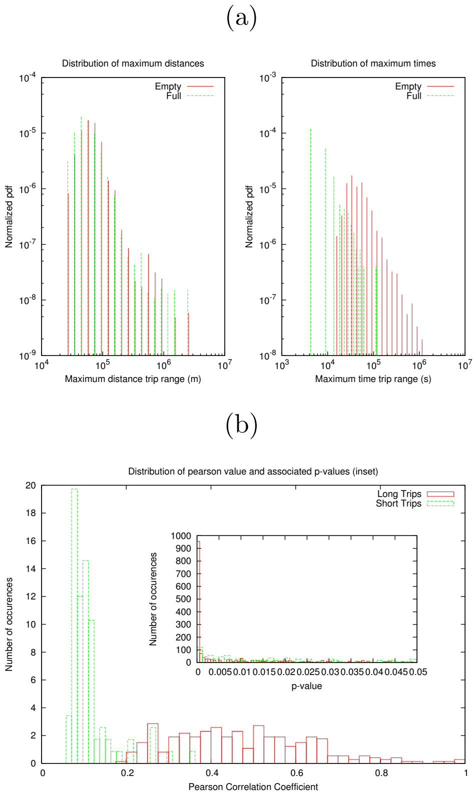


Figure 3: Efficiency parameters for the taxis. We do not observe a big variety nor heterogeneity in the data except a bigger spread on the velocity quotient. For the definitions of the variables used refer to main document.



(a)

(b)

Figure 4: (a) Distribution of maximum values for empty/full trips for Δr and Δt. We observe that while the distance distribution do overlap it is not the case for the time distributions, (b) Distributions of pearson correlation coefficients computed for long ($\Delta r > 20$km) and short ($\Delta r < 5$ km) trips and associated p-values (inset). We observe a pronounced difference between both, long trip correlation coefficients are more spread and have smaller *p-values* (more confidence) skewed to values at the right of the graph. Please note that a faint number of small *outsiders* with negative values have been omitted from the plot.

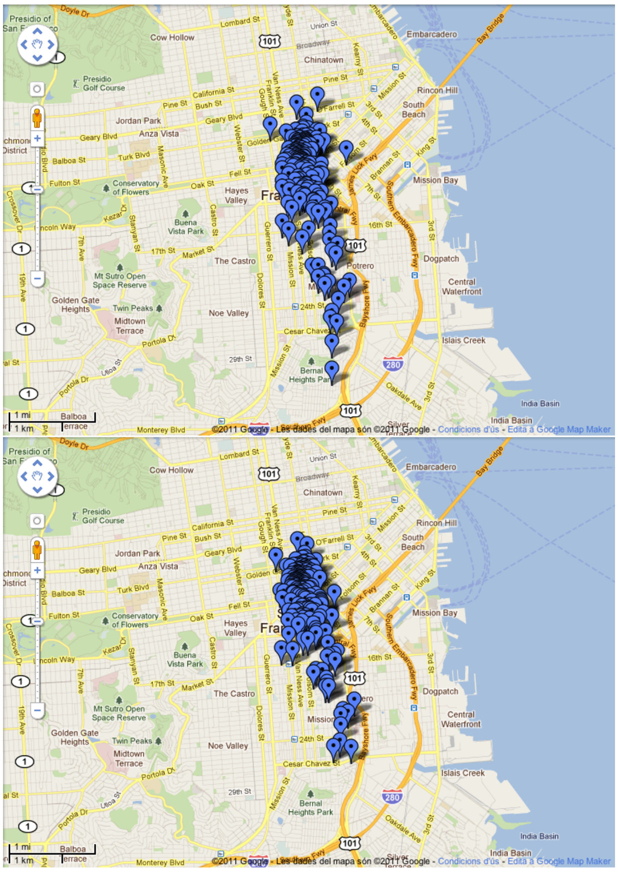


Figure 5: Map showing the locations of the mean position of the taxis when full (upper figure) and empty (lower). We see they group in the center of the city and that the spread is greater in the case of empty situation.

2.5 Gyration radius and center of masses

To end up the individual taxi data analysis, the center of masses of each taxi was computed and their mean squared spread over this value (r_{gyr}) (see main text). A *Google Map* created using the `python geopy` and `simpleklm` [8] is shown on figure 5 where the CM's are placed on the map, as expected, they all concentrate in the city center[9].

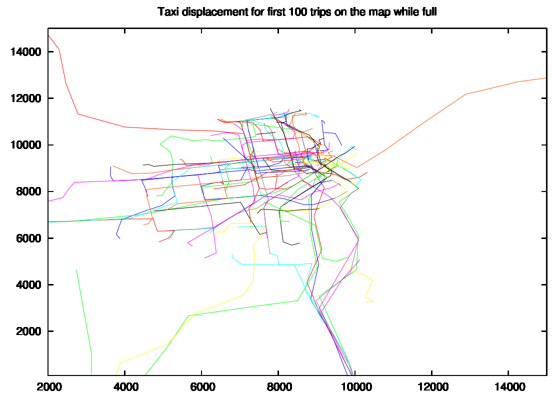


Figure 6: Example of taxi number 10's first 400 trajectories on the UTM 10S plane referred to our custom defined origin as directly translated from latitude/longitude positions.

3 Building the net

In this part of the appendix we explain in detail the procedure followed to construct the net and the calculations that support the choices made (and explained in the main document) for it to be directed and weighted.

We have used the `python` module `NetworkX` [9] to build the net and perform the numerical operations on it, as for what representation concerns, we used the open source `Gephi` suite [10]. Finally, for the coordinate transformation we used the module `pyproj` [11].

First of all the author wanted to refer the data to a bidimensional map of cartesian coordinates, to do so the *Universal Transverse Mercator coordinate system* (UTM) was used centered in zone map 10S (where the area of San Francisco is included). A convenient value of $(x_1^0, x_2^0) = (543442, 4173200)$ m was chosen as the origin of a grid of variable size (after some trials we have set the gridsize as squares of $\Delta x_1 \times \Delta x_2 = 100 \times 100$ m). Then a simple algorithm[10] was performed on both the empty and full data sets.

1. We create a both dictionaries `nodelist,edgelist`[11] and set `nodecount` $= 0$.

2. For every trip (group of pairs of latitude/longitude data) we translate the starting and ending positions to UTM (x_1, x_2) coordinates.

3. We compute $n_i \in \mathcal{Z} | n_i \leq x_i/\Delta x_i < n_i + 1$ for $i = 1, 2$ for initial (in) and ending points (end).

4. If $(n_1, n_2) \notin$ `nodelist` : `nodecount`$+= 1$ and `nodelist`[`nodecount`] $= (n_1, n_2)$ and

such values compute the likelihood of a linear relation between variables, and our goal is to check for any linear relationship in terms of orders of magnitude (i. e., after a long trip, taxi drivers will return to the city but not to the same exact location). In any case, for the long trips we roughly excluded 40% of the data.

[9] In fact, they concentrate in the area formed by the three most visited locations (Union Square, Embarcadero and SFK airport) as we shall see on our complex network approach.

[10] Please note that `python` is a *high level programming language* and hence it is better to perform calculations in a *Matlab* way rather than performing many loops for efficiency reasons.

[11] An array of pairs `key:value` where one can access the `value` anytime by querying `A[key]`. For more details see [12].

`edgeinitial` $=$ `nodecount`$_{\texttt{in}}$, `edgefinal` $=$ `nodecount`$_{\texttt{end}}$. If `edgefinal` $\neq$ `edgefinal`,

- If $(\texttt{edgeinitial}, \texttt{edgefinal}) \notin$ `edgelist`: `edgelist`$[\texttt{edgeinitial}, \texttt{edgefinal}] = 1$.
- Else: `edgelist`$[\texttt{edgeinitial}, \texttt{edgefinal}] += 1$.

5. Else: continue to next trip.
6. Until all trips have been computed. Then, we add directly the nodes in the dictionary to the net, adding as node attributes the latitude and longitude of each point obtained via inverse transformation of $x_i = n_i \times \Delta x_i + x_i^0$ from UTM to geodesical coordinates.
7. Finally we add the edges from the remaining dictionary, where their weight represents the number of trips between locations (non normalized).

Note that we create a net where isolated nodes may occur. We do not include it on the description of the algorithm but we computed and alternative dictionary with the selfloops. Also note that the algorithm needed to be modified to compute the second net, as we wanted the numeration of nodes to remain unchanged (in order to compare net nodes for both empty/full situations).

3.1 Isolated and selfloops

We trimmed the nets from isolates and selfloops for various reasons. First of all due to the fact that selfloops imply trajectories inside an area of 1 ha (short trips) which are not very representative of taxi usage, whereas isolated nodes are points of the grid from which no trips either leave ($k^o = 0$) or enter ($k^i = 0$), facts caused by either the discrete nature of the grid used or probably by bad GPS data. Secondly because many algorithms may fail used on an un-filtered net such as the one we are treating. An overview on the data of such points is found on table 3. Focusing a little bit further on the selfloops, we observe that the distribution of their weights is fat tailed as can be seen in figure 7. Finally, a map[12] of the overlapping isolates in both empty and full maps was computed but no significant relation between such points was detected[13].

	N_{iso}	$Overlap_{iso}$	E_s	$Overlap_s$	$< \omega_{ij} >$
Empty	217	4.6%	5143	63%	15 ± 41
Full	160	6.3%	4119	78%	3 ± 7

Table 3: Isolates N_{iso} and selfloops E_s information. Note that Overlap is defined as the percentage of isolated nodes(selfloops) appearing on both nets at the same time from the total of isolates N_i(selfloops E_s). The last column is the average weight of selfloops and its associated standard derivation values indicate a fat tail behavior (as seen in figure 3).

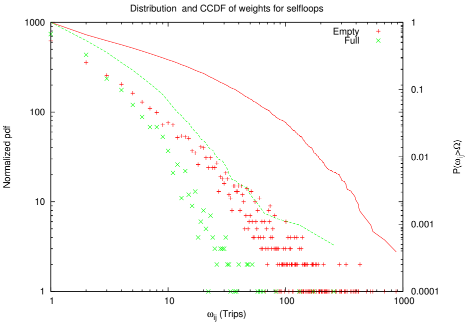


Figure 7: CCDF (axis at right) and distribution of trips (weights, axis at left side) in selfloops.

3.2 Weighted and directed net justification

We choose to represent our nets in a weighted and directed way. The reasons to use a weighted net are similar to those presented in for example[13], since accounting the effect or *load* travelling through the different edges can provide more accurate topological as well as dynamical information about the network at study. As for the reasons to represent the net in a directed way (thus increasing the level of difficulty of our study), since we deal with directed trips, it seems logical to adopt these strategy. Moreover, the statistics presented in table 4 show that since the net is not in any case symmetric, we are forced to adopt this approach.

	$< \lvert\omega_{ij} - \omega_{ji}\rvert >$	E_{sym}	ρ^{sym}	ρ^0	R
Empty	1.06 ± 5.1	23454	16%	45%	0.24
Full	1.03 ± 2.07	30701	17%	50%	0.26

Table 4: Mean difference of weight between edges that exist in both senses. E_{sym} is the number of pairs of edges that connect 2 nodes both ways and $\rho^{sym} \equiv \frac{2E_{sym}}{E_{total}}$ is the percentage over total number of nodes and $\rho^0 \equiv \frac{E(<\omega_{ij}-\omega_{ji}=0>)}{E_{sym}}$. R is the reciprocity of the net, see §4.1.

[12] One can consult via internet all the maps mentioned, to obtain the addresses refer to §6.

[13] In fact the top node concentrating most of the transit is the base of the taxi company from which the data was taken. Hence it seems correct to exclude those data from the study.

4 Studying the Net

4.1 General features

The general features presented in the table on the main document are defined as follows[14],

- Number of edges is E, number of nodes is N.
- Density: $\rho \equiv \frac{E}{N(N-1)}$
- Degree in (out) $k^{i,o}$: Number of edges entering (leaving) a node.
- Strength in (out) $s^{i,o}$: Total sum of weights of edges entering (leaving) a node[15]. Also referred as weighted in (out) degree.

$$s_i^{out} = \sum_{j=1}^{N} \omega_{ij} A_{ij} \qquad s_j^{in} = \sum_{i}^{N} \omega_{ij} A_{ij}$$

- Degree and strength assortativity : Measures the similarity of connections in the graph with respect to the node degree or strength. It is the pearson coefficient of the different degree (strength) pairs of nodes at each end of the edges of the net, applied to strengths by the author from the definition in [15] and [16] for pairs $In-In$, $Out-Out$, $In-Out$ and $Out-In$. For further details see §4.3. The original formula corresponds to,

$$r(\alpha,\beta) = \frac{E^{-1}\sum_i [(j_i^{\alpha} - <j^{\alpha}>)(k_i^{\alpha} - <k^{\alpha}>)]}{\sigma^{\alpha}\sigma^{\beta}}.$$

  Where the parameters α, β are the degree or strength (in or out) of the (source,destination) node pairs j, k.

- Betweenness centrality: Relative number of shortest paths that pass trough a node, thus giving an idea of the *importance* in traffic handling of such node. It is defined for node $v \in V$ as,

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}},$$

  and the number of shortest path for the weighted net is computed using a modified *Dijkstra's distance algorithm* as in [17]. Please note that using this algorithm the shortest path between two nodes is the one with the lowest sum of weights in the edges used to reach destination.

- Reciprocity R: Gives a measure of the similarity of edges connecting nodes in the two senses in a weighted manner. Adapted from [14] introducing weights by author. A value towards 1 indicates high (weighted) reciprocity, and the opposite towards 0.

$$R \equiv \frac{\sum \sqrt{\omega_{ij}\omega_{ji}A_{ij}A_{ji}}}{\sum A_{ij}\omega_{ij}}.$$

	S_w^1	ρ_w^1	S_w^2	ρ_w^2
Empty	17213	98.64%	4	0.023%
Full	17189	98.23%	3	0.017%
	S_s^1	ρ_s^1	S_s^2	ρ_s^2
Empty	7724	44.23%	1	0.005%
Full	7738	44.18%	2	0.011%

Table 5: Size of first and second biggest strongly and weakly connected components. We observe that the graph are well above the phase transition occurring in such nets [15] on their undirected versions.

- Size of weakly/strongly connected components $S_{w,s}^i$: The $i-th$ weakly (strongly) connected component is the $i-th$ biggest connected subgraph of nodes (by undirected edges for the weakly component, by directed edges for the strongly one) in the net, and $\rho_{w,s}^i \equiv= S_{w,s}^i/N$ its relative size. It is computed using the algorithm proposed in [18] for directed graphs. In table 5 we present the values obtained.
- Average Neighbor degree/strength: Mean degree of neighboring nodes for unweighted nets [19].

$$k_{nn,i} \equiv \frac{\sum_{j=1}^{N} a_{ij}k_j}{k_j} \qquad s_{nn,i} \equiv \frac{\sum_{j=1}^{N} a_{ij}s_j}{k_j}.$$

- Average Weighted In (Out) Neighbor degree/strength: Average number of degree for the neighbors of a node weighted by their out strength (relative preference of out-going connections).

$$k_{nn,i}^{w} \equiv \frac{1}{s_i}\sum_{j=1}^{N} a_{ij}\omega_{ij}k_j.$$

  Can be computed for the pairs $< x_{nn,}^{w} > (x)$: $(in-in), (out-out), (in-out), (out-in)$ $\quad x = k, s$ producing different information about assortativity of the network. Partially adapted for directed networks from [13].

We present on figure 8 the distribution of weights for both networks as well as the histogram (pdf) for the strengths and degrees of the net. Note that whereas these distributions seem to follow a power law, the distributions shown of degree and strength are very influenced by finite size effects or directly seem to be *logbrownian* shaped[16], this is why we do not present the results of the fits. The distribution of betweenness is also shown.

[14] The usual concepts of graph theory have been mainly extracted from [14].

[15] All the weighted quantities referred here are adapted from the undirected weighted versions in [13]

[16] In their CCDF plots shown in the main document.

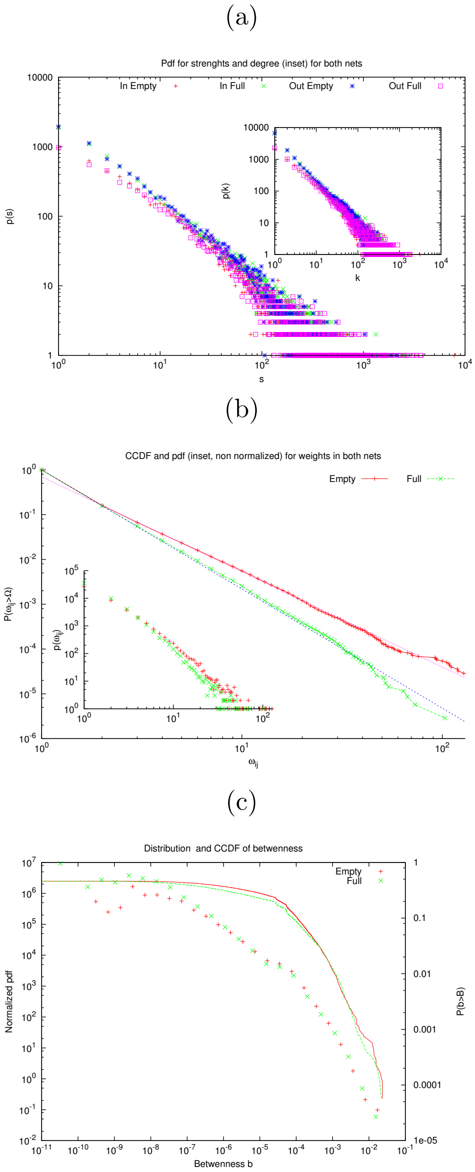


Figure 8: (a) Probability distribution (non normalized) for strength and degree (inset) for both nets. (b) Probability distribution (non normalized) and CCDF for the weights in both nets. The straight lines are fits to power law behavior with exponents $\alpha_e = 2.125 \pm 0.005$ and $\alpha_f = 2.660 \pm 0.003$. (c) Probability distribution (with logarithmic binning) and CCDF for weighted betweenness in both networks. We observe that the most central hubs accumulate up to a 2% of the traffic.

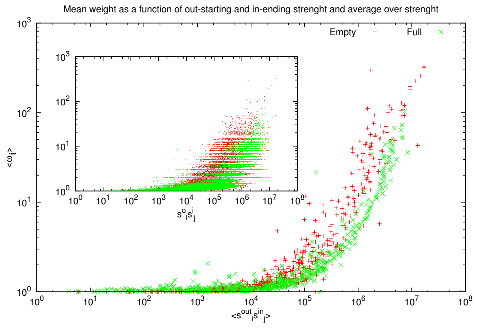


Figure 9: Mean weight of edges as product of starting ending out (in) weighted degree (strength).

		Empty	Full
In Degree	High	Taxi Stop, Driver resting area	Hot Spot (airport, train station...)
	Low	Low customer density spot	Particular location (house, work)
Out Degree	High	Driver resting area, Airport departures terminal	Hot Spot
	Low	Hot spot, Scattered Location, Taxi Stop	Particular location

Table 6: Examples of meaning of high and low degree spots.

4.2 Studying Correlations

To further study the correlations present in the network between weights and connections, we have computed the mean strength as a function of the betweenness of the nodes (showed in the main document), as well as the mean weight of edges as a function of the product of its starting and ending nodes' strength (figure 9) and degree (in main document). We confirm a roughly constant part accounting for lack of correlation for low values of k, s that is lost in the form of a potential function towards the values corresponding to most connected nodes.

We have repeated all the calculations shown in the main document for the degree of the nodes to see if there where any differences with the strength versions of the computation. Besides an increase in the noise, we did not see any differences, and this fact backs up our choice of applying a mean field approximation to strengths rather than degrees explained in the main document.

4.3 Assortativity

To compute the pearson coefficient for assortativity for the attributes mentioned (in/out degree and strength) we applied a usual r definition on a bidimensional array containing for each node's first attribute α^n its neighbors second attribute β^{nn}. We roughly present in figure 10 the meaning of each attribute pair assortativity measure used in the main document.

In table 6 we introduce examples of the meaning of high and low connected nodes could be for further clarification.

In table 7 we show the complete set of r values for both networks corresponding to k and s and in figure

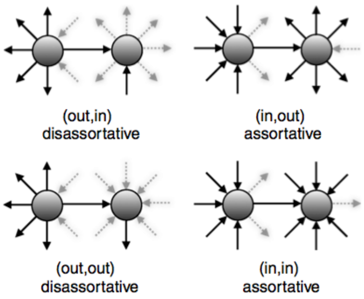


Figure 10: The four degree-degree correlations in directed networks. The fuzzy edges indicate that nodes can have any number of edges of this type, as they do not enter into the specific correlation. For each correlation we show an example typical of assortative or disassortative networks. Taken from [16].

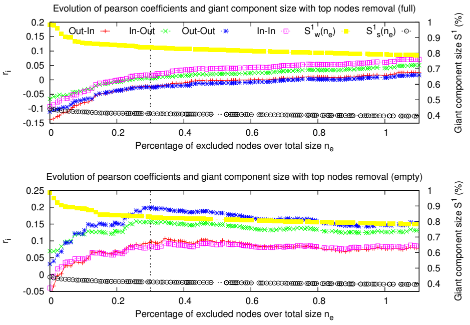


Figure 11: Evolution of the size (weakly and strongly connected) biggest component S^1 and r values with successive removal of top in and out degree nodes.

11 we show the same version of the plot appearing in the main document for the evolution of raw degrees of the net as we exclude top nodes. We observe very similar patterns as the ones explained there. To do so, we used a simple algorithm:

1. Copy original net.

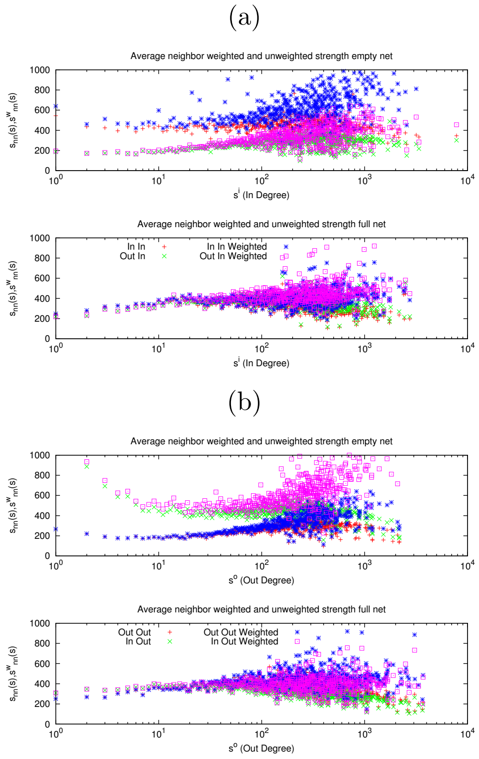
2. Obtain top nodes of the network for both in and out versions of desired attribute (strength or degree). In case of two nodes having the same rank, we place them randomly.

3. Recursively extract all the nodes in the list, grouped by their rank (if there are two top number 5 nodes, we extract them all at the same time).

4. Compute reduced r coefficients as defined in [20] for the nodes in the new net, but using the attributes from the old one. Please note that we do not recompute the attributes of the net (strength and degree), we simply eliminate the contribution of the top ones.

	r_{in-in}	$r_{out-out}$	r_{out-in}	r_{in-out}	$< r >$
Empty s	-0.032847	0.017782	-0.054016	0.045915	-0.018 ± 0.043
Empty k	-0.039805	0.031489	-0.067035	0.070131	0.02 ± 0.05
Full s	-0.079386	-0.093057	-0.116725	-0.057891	-0.10 ± 0.03
Full k	-0.092305	-0.110126	-0.138987	-0.066179	-0.090 ± 0.018

Table 7: Pearson r coefficient for strength and degree in both networks.


(a)

(b)

Figure 12: (a,b) Mean neighbor weighed and unweighted strength for both nets. We observe the very same shape as in the plots presented in the main document.

Finally we show the figures 13 representing the calculations performed when evaluating the assortativity of the net for both k and s with the complete net and the trimmed one. In fact we do not observe any differences in the shapes (which are the important factor of the qualitatively analysis performed on such graphs). Moreover, in figure 12 we observe the same shape as the graph presented with the raw degrees on the main document, hence confirming the general interchangeability of s and k in our study.

5 Plotting the net

We present some additional representations of both the full and empty networks not shown in the main document[17].

5.1 Empty Net

a) Giant component of trimmed net with condition for edges $\omega_{ij} \geq 5$ after a modularity analysis (modularity of 0.602) [21] representing nodes according to their geographical position (shown in main document). Node size is the incoming strength for nodes. Number of visible edges $E = 6503(2.33\%)$, nodes $N = 1168(6.98\%)$.

b) Radial figure of the previous plot with radial coordinate corresponding to decreasing values of incoming degree. We observe the proportionally studied for s,k.

c) Radial figure of plot a) with radial coordinate representing outgoing strength. We observe the overall assortative tendency for medium and big sized nodes (except the case of the airport, biggest dot that represents the resting area for taxi drivers in airport). Shown in main document.

5.2 Full Net

d) Plot under the same conditions as performed in a). Number of nodes $N = 1017(5.81\%)$ and edges $E = 5111(1.47\%)$. We observe that the modularity analysis no longer relates to delimited geographical zones.

e) Image showing the group nodes $\{g \in G | k^{in}(g) < 6 \cap k^{out}(g) < 6\}$ that contains $N_G = 11667(66.63\%)$ nodes and $E_G = 1298(0.38\%)$ edges. The size of the nodes represents their betweenness and the color their outgoing degree (black for 1, pale green for 5). We observe the almost inexistent tendency of low connected nodes to link among themselves since despite these nodes representing over 70% of the net, they just have links that represent a very small part of the total connections present in the network.

6 Maps

Finally please note that there is additional on-line material in `http://maps.google.com/maps/user?uid=218221286799316438623&hl=ca&ptab=2`, where the reader may find a collection of Google Maps featuring the positions of different important nodes of both networks[18].

The maps feature the top 25 nodes for each of the listed attributes to provide better information about top spots for taxis in San Francisco.

- Degree and strength (in and out).
- Hub and Authority number (computed with the HITS algorithm) [22].
- Betweenness.
- Isolated nodes.
- Selfloops.
- Center of masses for each taxi.

[17] The author recommends to visualize the attached images using a computer as their quality is good enough to be zoomed several times for a convenient exploration by the reader.

[18] Please note that Google disabled some functionality on their website, and thus we have attached to the present document the maps in `.kml` format to be opened with the GoogleEarth software.

References

1. Beck, B. (2006).
2. Hill, D. *ELT Journal* **xxm**(176) (1997).
3. Piorkowski, M., Sarafijanovic-Djukic, N., Grossglauser, M., and Piorkowski M. Sarafijanovic-Djukic N., G. M. *2009 First International Communication Systems and Networks and Workshops* , 1–10 January (2009).
4. Jones, E., Oliphant, T., Peterson, P., and Others. (2001).
5. Newman, M. E. J. *Contemp. Phys.* **46**(5), 323–351 September (2005).
6. Ginsburg, A. (2009).
7. Gonzalez, M., Hidalgo, C. a. C., Barabási, A. A.-L., and González, M. C. *Nature* **453**(7196), 779–782 June (2008).
8. Lancaster, K. (2011).
9. Hagberg, A. A., Schult, D. A., and Swart, P. J. In *Proceedings of the 7th Python in Science Conference SciPy2008,* Varoquaux, G., Vaught, T., and Millman, J., editors, volume 836, 11–15. Los Alamos National Laboratory (LANL), (2008).
10. Bastian, M., Heymann, S., and Jacomy, M. *American Journal of Sociology* , 361–362 (2009).
11. Whitaker, J. (2006).
12. Pilgrim, M. *Dive Into Python*, volume 203. Apress, (2004).
13. Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. *PNAS* **101**(11), 3747–3752 March (2004).
14. Newman, M. E. J. *Networks: An Introduction.* Oxford University Press, (2010).
15. Newman, M. *Phys. Rev. Lett.* **89**(20), 1–4 October (2002).
16. Foster, J. G., Foster, D. V., Grassberger, P., and Paczuski, M. *PNAS* **107**(24), 10815–20 June (2010).
17. Brandes, U. *Journal of Mathematical Sociology* **25**(2), 163–177 (2001).
18. Nuutila, E. and Soisalon-Soininen, E. *Information Processing Letters* **49**(1), 9–14 (1994).
19. Serrano, M. A., Maguitman, A., Boguñá, M., Fortunato, S., and Vespignani, A. *ACM Transactions on the Web* **1**(2), 10–es August (2007).
20. Xu, X.-K., Zhang, J., Sun, J., and Small, M. *Phys. Rev. E* **80**(5), 1–7 November (2009).
21. Newman, M. E. J. *PNAS* **103**(23), 8577–8582 (2006).
22. Brin, S. and Page, L. *Computer Networks and ISDN Systems* **30**(1-7), 107–117 April (1998).

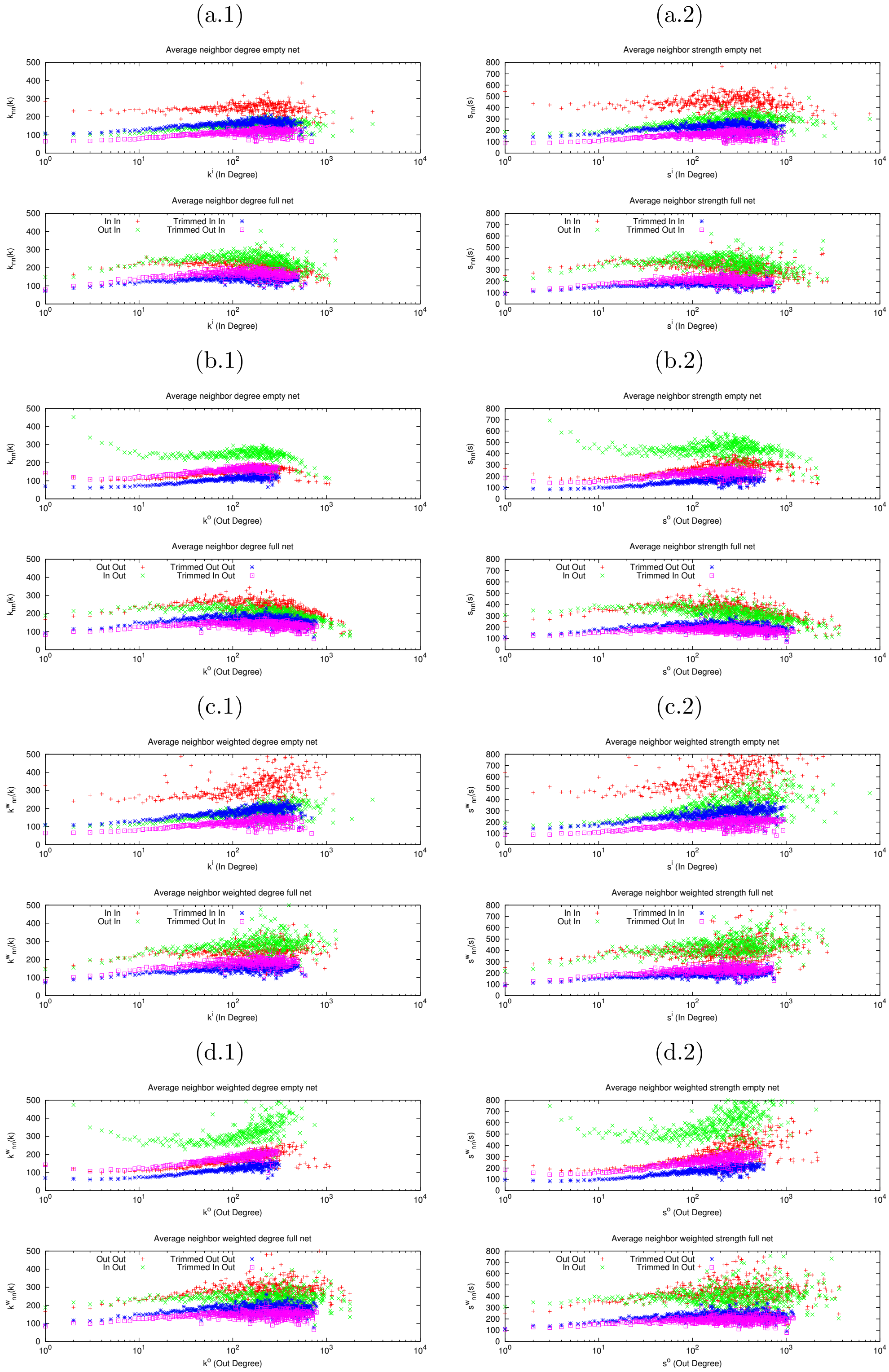
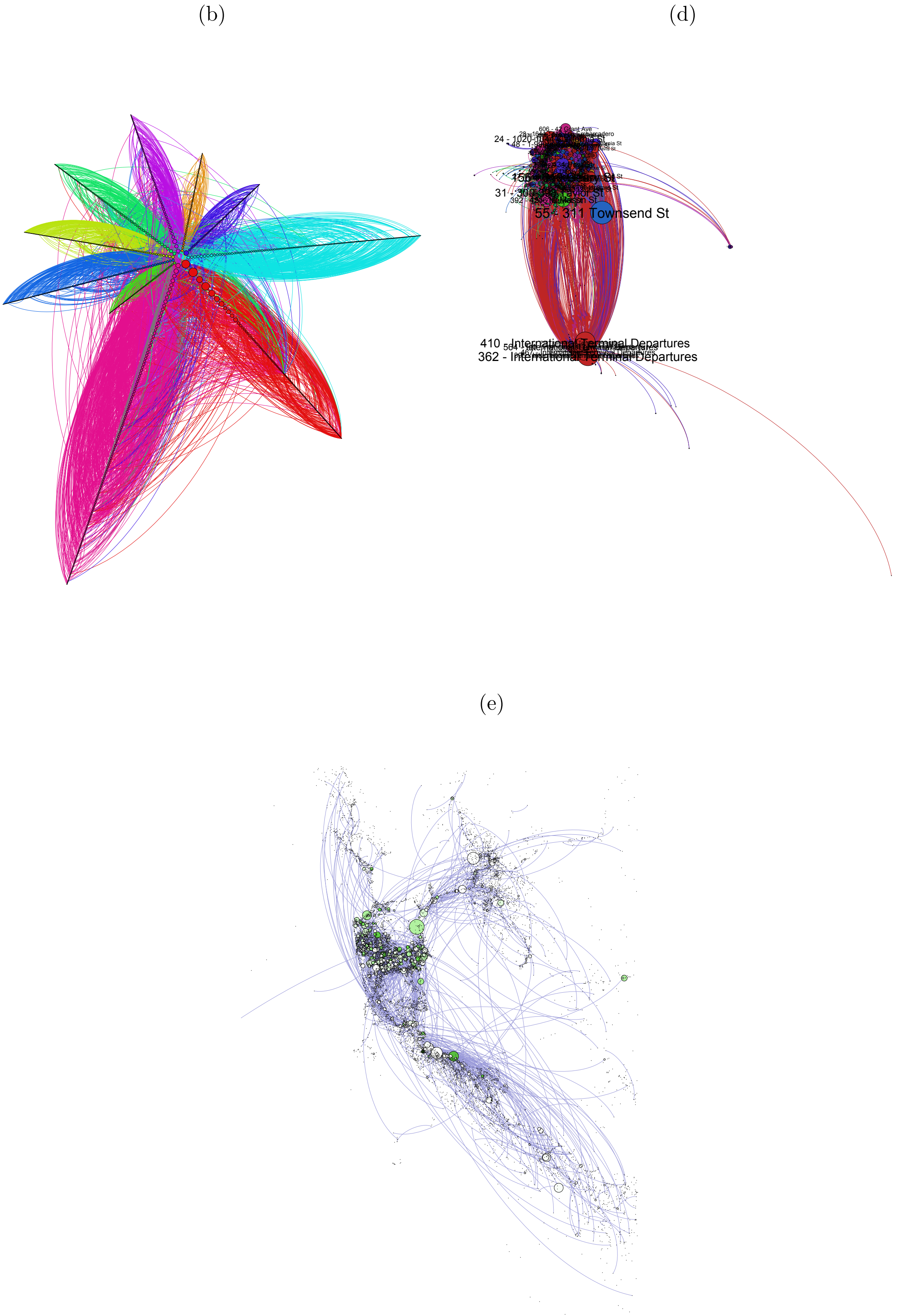


Figure 13: (a,b) Mean neighbor and strength degree (c,d) Mean weighted neighbor and strength degree, computed for the complete net and the trimmed net removing 0.3% of their top nodes. We observe some changes on the initial and final ranges of the shapes of the functions, corresponding to the influence of top nodes on small ones. We also see the more accurate description that the weighted average provides, enhancing the disassortativity for small nodes and linking tendency among top connected locations.


(b)

(d)

(e)

606 - 47 Grant Ave
24 - 1020 ...
31 - 300 ... Taylor St
55 - 311 Townsend St
410 - International Terminal Departures
362 - International Terminal Departures

Figure 14: Additional images of both nets. See main text for details.