Simplicity in complex networks: a path-based centrality

Leonardo Bellocchi[1,*] and Nikolas Geroliminis[1]

[1]Urban Transport System Laboratory, École Polytechnique Fédérale de Lausanne, Switzerland

[*]Corresponding author: leonardo.bellocchi@epfl.ch

Keywords: Network analysis; path choice; complex networks; drivers' behaviour; link betweenness.

Abstract

Understanding travel behaviour in transportation system is key challenge to calibrate and simulate the usage of urban mobility networks. We define in this work a path-based centrality based on the simplicity of the path between two locations in road network. Analyzing a huge dataset of GPS points of more than 20'000 vehicles and 170'000 trips, we reconstructed the real trajectories and estimate the degree of simplicity of each of them. Interesting insights of drivers' behaviour came from the comparison with the shortest and the simplest path. This allowed us to categorize trips according with their complexity and extract general behavioural relation among drivers. Finally, we measured the effect of considering simplicity as path-choice factor influences the distribution of road usage and the link betweenness.

Introduction

In an urban network, the shortest path for a driver it is not always the most convenient and probable choice. Other factors rather than time and length are implied for the travellers' path choice. Habits, risks, as well as the *simplicity* have important influence in drivers' behavior. Learning how drivers chose their path it is of fundamental help for traffic management. In fact, knowing the most probably used road by transportation users can lead to a more efficient management of the urban network and prevent traffic jams and slowdowns. In sociology literature, it is well know the concept of Dunbar Number (see for example [6]), how the *limit of information* for certain type of social relation a person can effectively hold. This finding has inspired the results in [5] where the authors compute the information of the all shortest path algorithm as a measure of simplicity of spatial networks. Another definition of simplicity for complex networks has been done in the work of Costa et al. [4], where the authors identify simplicity with the regularity of the network and some homogeneity among local clustering coefficients and node centralities. The above-mentioned works have in common to describe the network in a static way and at node level, looking at the topological structure of the *reseau* and on its connection properties. In our work, we designed an algorithm to quantify the information that a driver has to retain during a trip counting the number of *changes* at the intersection. With *change* we intent when a driver make a decision during its trip, for example turning

right or left not following the natural extension of the road. For this scope, we defined a path-based centrality that measures the level of complexity of each trip and, at the big scale, identifies the more information-demanding road. A change happens every time the path deviates from its natural extension because of their relative angle deviation or because different functional road type. By studying the characteristics of real trajectories and comparing their degrees of complexity and length, it allows us to distinguish behavioural patterns among drivers and average coefficient of shortness and simplicity.

1 The simplest path algorithm

Let $\mathcal{G} = G(\mathcal{N}, \mathcal{L})$ be a graph representing a network with $\mathcal{N}$ nodes and $\mathcal{L}$ links. We consider a path in $\mathcal{G}$ a collection of links $\{\ell_1, \ell_2, \ldots, \ell_r\} \in \mathcal{L}^r$ with ℓ_i adjacent to ℓ_{i+1} $\forall i = 1, \ldots, r-1$. Given a minimal angle perception threshold Δ, we consider that a driver makes a change going from ℓ_i to ℓ_{i+1} if and only if *all* these conditions are satisfied:

(a) There is more than 1 link belonging to the same road type of ℓ_i;

(b) The angle between ℓ_i and ℓ_{i+1} is not the minimum;

(c) The angle between ℓ_i and ℓ_{i+1} is greater than Δ.

In our application, we fixed the threshold $\Delta = \pi/6$. This algorithm is based on the road perception that the driver has during her/his trip. Whenever the driver makes a decision, i.e. needs *information*, we count it as a change. Here, we propose two different ways to quantify this *change*: Boolean (1 or 0) or with a weight given by the $arctan()$ between the two consecutive road vectors of her/his path. We show in the results as these two methods brings similar conclusions for what concerning the classification of observed path and drivers. An illustrative example of this algorithm is shown in Fig. 1. We define *the simplest path* between two points (O, D) on the map, the path which has the minimal number of changes and, in case of multiple solutions, the shortest one among them.

2 Results

2.1 Simplicity and shortness in paths.

For each trips in our dataset we individuated its origin (O) and destination (D) in the urban map of Shenzhen. Then, we calculated the trip length of the observed trajectory ($RE(O, D)$) and the number of changes that the driver effectuated during his trip (O, D) with our algorithm. For the same pair of origin-destination, we computed the shortest path ($SH(O, D)$) and the simplest path ($SI(O, D)$) and, again, calculated the trip length and the number of changes. We visualize the results in Fig. 2. We can notice how the real paths use in average the double in term of number of changes and in trip length with respect to the simplest and the shortest path respectively. In order to analyze our dataset of observed trajectories, we defined the following parameters for

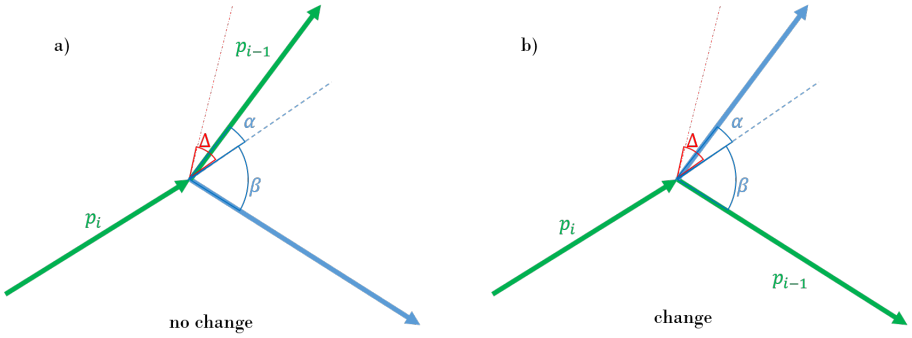

Figure 1: **Example of the counting changes algorithm.** On the left, the path in green has no change at the intersection $p_i - p_{i+i}$ because the angle α is the minimum between the other choice (with $\beta > \alpha$) and is under the perception threshold Δ. On the other hand, on the right, we count the passage between p_i and p_{i+1} as a change, considering p_{i+1} not the "natural extension" of the road p_i.

each Origin-Destination (OD) and path p:

$$\Delta^{OD}_{ch}(p) = \frac{ch(p(O,D)) - ch(SI(O,D))}{ch(SI(OD))} \qquad \text{with} \quad ch(SI(O,D)) \neq 0$$

$$\Delta^{OD}_{len}(p) = \frac{len(p(O,D)) - len(SH(O,D))}{len(SH(O,D))} \qquad \text{with} \quad len(SH(O,D)) \neq 0$$

with $ch(\cdot)$ and $len(\cdot)$ the functions that compute the number of change and the trip length respectively. In Fig. 3 we report the scatter plot of $\Delta^{OD}_{ch}(RE)$ (3. a) and $\Delta^{OD}_{len}(RE)$ (3.c) for each OD of our dataset and their averages $\bar{\Delta}_{ch}(RE)$ and $\bar{\Delta}_{len}(RE)$ (resp. 3.b and 3.d). We can notice that, however, the average $\bar{\Delta}_{ch}(RE)$ (≈ 0.55 for the observed and ≈ 0.2 for the SH) and $\bar{\Delta}_{len}(RE)$ (≈ 0.25 for RE and 0.13 for the SI) are both very stable during the whole day and that their distributions in panel *a*) and *c*) present some correlation with the traffic congestion. This means that a percentage of drivers and of trip are influenced by traffic not only for the increase of travel time but also in drivers' behaviour and path choice, with a substantial increment of complexity, detour and trip length.

2.2 Trip categorization and driving behavioural priorities

In Fig. 4 we compare the Δ^{OD}_{ch} of the observed path RE with which of the shortest path SH and the Δ^{OD}_{len} of the RE with which of the simplest path. Therefore, we define 5 categories of trips:

(N): trips OD with $\Delta^{OD}_{ch}(RE) > \Delta^{OD}_{ch}(SH)$ and $\Delta^{OD}_{len}(RE) > \Delta^{OD}_{len}(SI)$;

(L): trips OD with $\Delta^{OD}_{ch}(RE) > \Delta^{OD}_{ch}(SH)$ and $\Delta^{OD}_{len}(RE) < \Delta^{OD}_{len}(SI)$;

(C): trips OD with $\Delta^{OD}_{ch}(RE) \leq \Delta^{OD}_{ch}(SH)$ and $\Delta^{OD}_{len}(RE) > \Delta^{OD}_{len}(SI)$;

(LC): trips OD with $\Delta^{OD}_{ch}(RE) \leq \Delta^{OD}_{ch}(SH)$ and $\Delta^{OD}_{len}(RE) < \Delta^{OD}_{len}(SI)$;

(E): trips OD with $\Delta^{OD}_{ch}(RE) = \Delta^{OD}_{ch}(SH)$ and $\Delta^{OD}_{len}(RE) = \Delta^{OD}_{len}(SI)$.

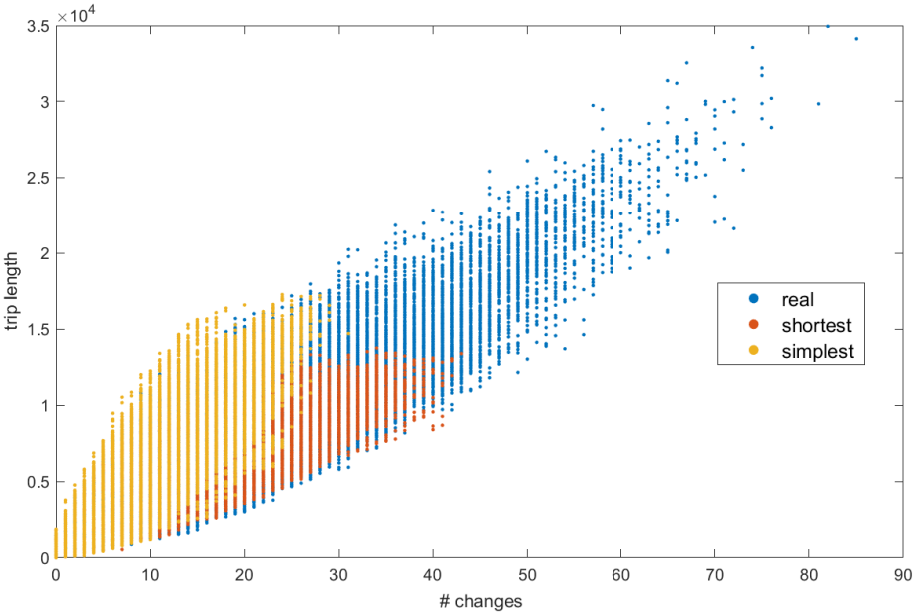

Figure 2: **Relation between total trip length and number of changes for each trip with the same Origin-Destination table.** We reported the results of the shortest paths (red) and the simplest path (yellow) compered to the observed trajectory (blue).

These 5 categories distinguish the different travel behaviour factors of the drivers. In particular, trips in cat(N) does not minimize neither the number of change nor the trip length. In cat(L) and the trip length is shorter than which of the simplest path while in cat(C) the paths are simpler than the corresponding shortest paths. The paths that belong to cat(LC) are the best compromise between shortness and simplicity, being those paths shorter than the corresponding simplest path and simpler than the corresponding shortest path. Finally in cat(E) there are the paths that coincide with the shortest and the simplest. The percentage of each of these categories are reported in the pie chart in Fig. 5. We reported 3 examples of trips belonging to categories (LC), (L) and (C) in Fig. 6.

We denote the fraction of links that two paths p_1 and p_2 with the same Origin and Destination have in common with the *overlapping function* $\mathcal{O}(p_1, p_2)$. Therefore, if $\mathcal{O}(p_1, p_2) = 1$ the two paths are the same and if $\mathcal{O}(p_1, p_2) = 0$ they do not have any links in common. In Fig. 7 we show the amount of OD pairs of the simplest (yellow) and the shortest (red) path that share at least a certain fraction of links with the observed path. The shortest path seems to have an higher overlapping score than the simplest path. We also plotted in yellow dashed line the overlap score if we use the simplest path algorithm weighted with the angle of each turn and it results lower than the other two cases.

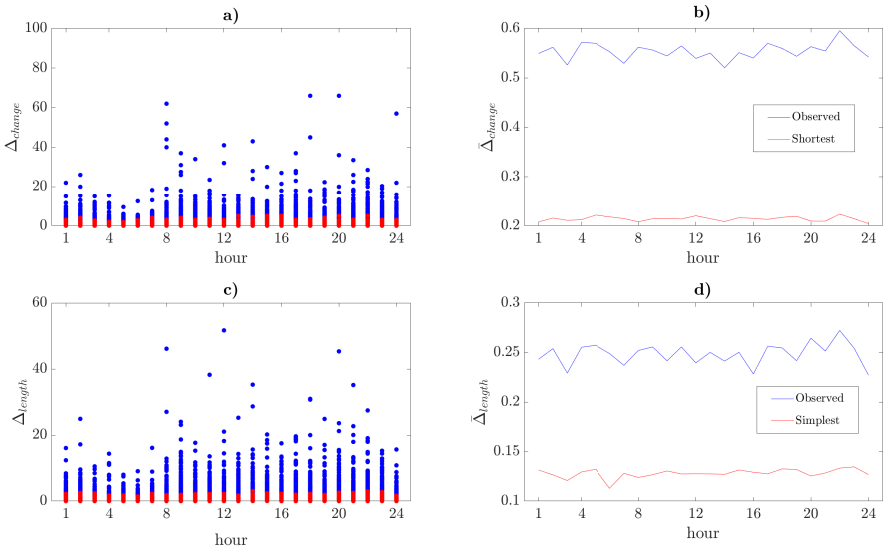

Figure 3: **Distribution (left) and average (right) of the *change ratio* Δ_{change} and *length ratio* Δ_{length}.**

2.3 Betweenness centralities for shortest, simplest and real path.

Once we compute the trajectory for real, shortest and simplest path, we are also able to estimate the usage of each road. The classical way to estimate the centrality of link in a graph is to calculate with an all shortest path algorithm the betweenness [1] of each node (or link) in the graph. As in [3], it corresponds to the weighed betweenness for link ℓ according to the three different ways to draw a path between points in a map, $\mathcal{B}_{RE}(\ell)$, $\mathcal{B}_{SH}(\ell)$, $\mathcal{B}_{SI}(\ell)$. The results that we show in Fig. 8 tell that the observed drivers' paths prefer to pass by the external arteria (on the left) while the betweenness $\mathcal{B}_{SI}(\ell)$ following the simplest path algorithm highlight the central straight arteria.

Another interesting result comes from the analysis of the distribution of the betweenness in the three scenarios (shortest-simplest-real). For this aim, we calculated the Gini coefficient of the distribution of the betweenness values at each time step (one every 6 mins) during all day (Fig. 9). Gini coefficient ([7]) is a index of inequality and the more is close to one the more the values $\{x_i\}$ are unequal distributed. In formulas,

$$G = \frac{E}{2M}, \qquad \text{with} \quad E = \frac{1}{n^2}\sum_i\sum_j |x_i - x_j| \qquad \text{and } M = \frac{1}{n}\sum_i x_i.$$

A Gini coefficient equal to 0 means that all the links have the same value while it is equal to 1 when only one element has the total value. We remark that the difference in time comes only for the different Origin-Destination table. The peak that we register during the night is due to the scarcity of trajectories in our dataset that imply a not full homogeneous coverage of the network enhancing inequality in betweenness distribution. We notice how the distribution of $\mathcal{B}_{RE}$

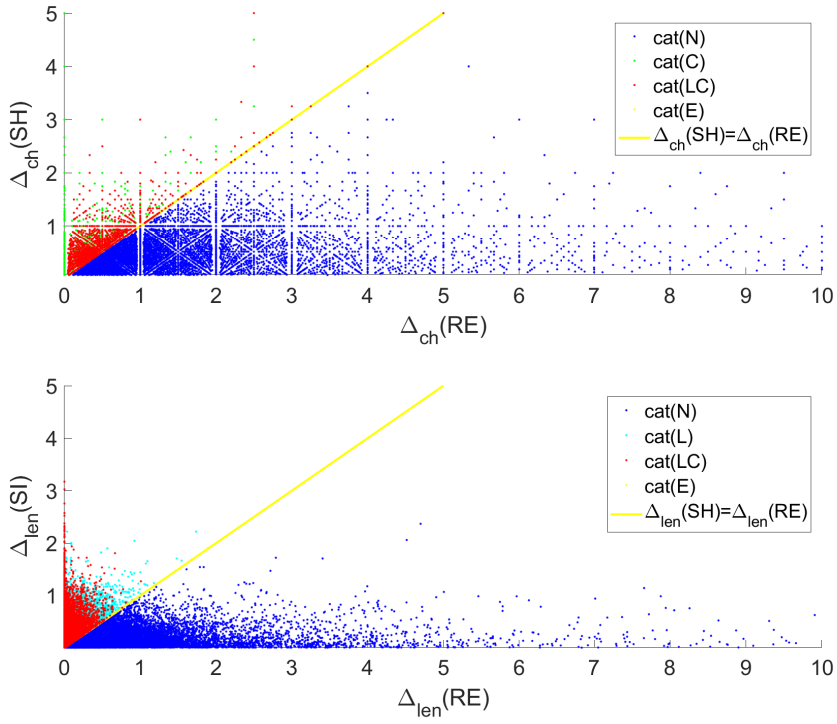

Figure 4: **Comparison of the degree of simplicity between Real and Shortest path (top) and the shortness between Real and Simplest paths (bottom).** In different colors are represented the 4 categories that we explain in the text. In particular, in red the real trajectories simpler than the shortest path and shorter than the simplest path while in blue the trip of cat(N) longer and with more changes than the shortest and the simplest. In the top panel in green the real trip with a simpler path than the shortest but longer than the simplest path and in cyan, in the bottom panel, the trip shorter than the simplest but with more change than the shortest path. The trip in cat(E) are in the diagonal (yellow) in both panels.

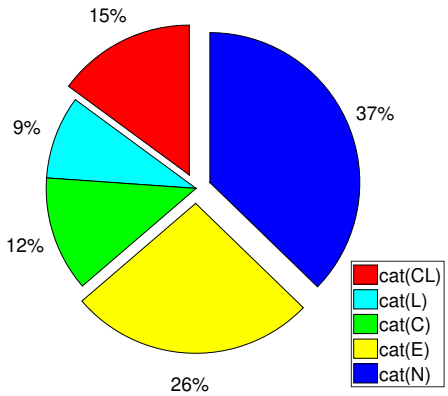

Figure 5: **Pie chart of the categories of trips.**

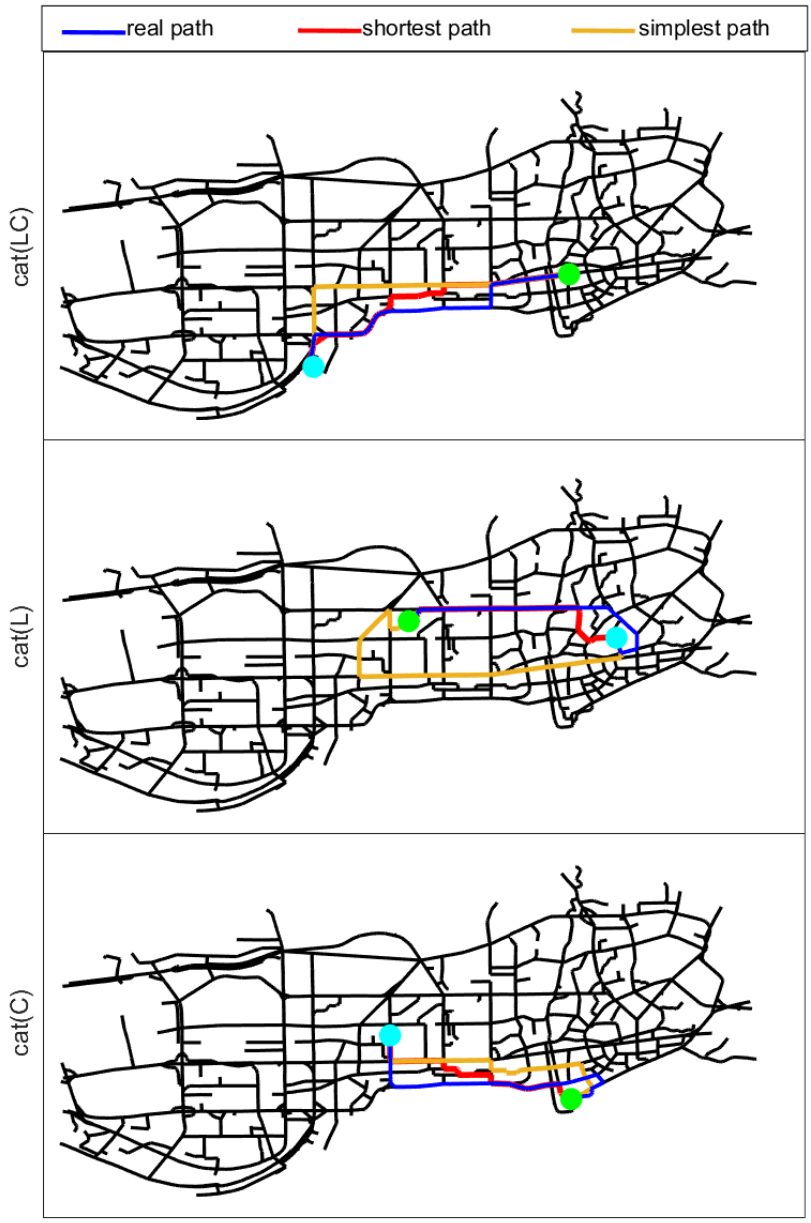

Figure 6: **Example of trip of categories LC (top), L (middle) and C (bottom).** In blue we draw the observed trajectory and we compare it with the shortest path (red line) and simplest path (dark yellow line). We can distinguish where the real user chose a path shorter than the simplest path (middle) or simpler than the shortest one (bottom) or both (top).

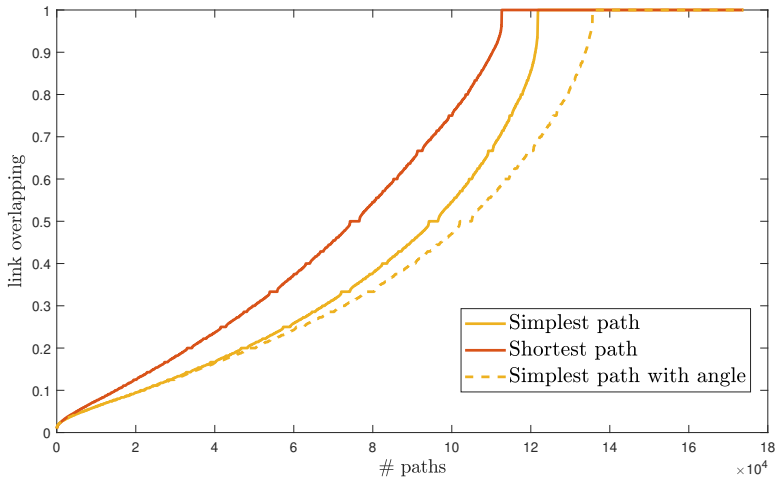

Figure 7: **Statistics of the link overlapping between the shortest path (red) and the simplest path (yellow) with the observed path for each trip.**

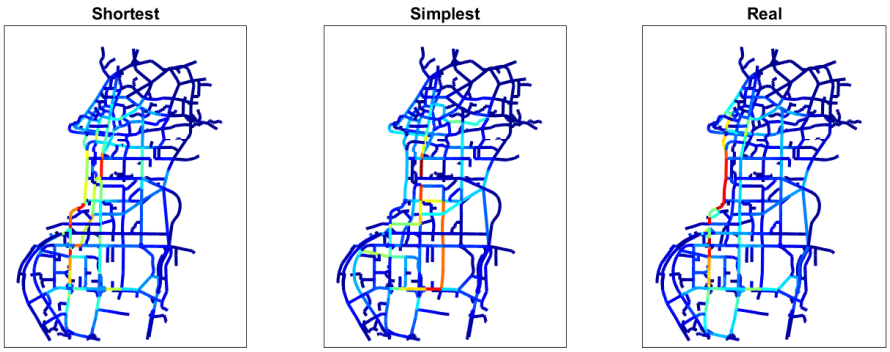

Figure 8: **Comparison of the tree different way to compute betweenness: a)** $\mathcal{B}_{RE}(\ell)$ **, b)** $\mathcal{B}_{SI}(\ell)$**, c)** $\mathcal{B}_{SH}(\ell)$**.**

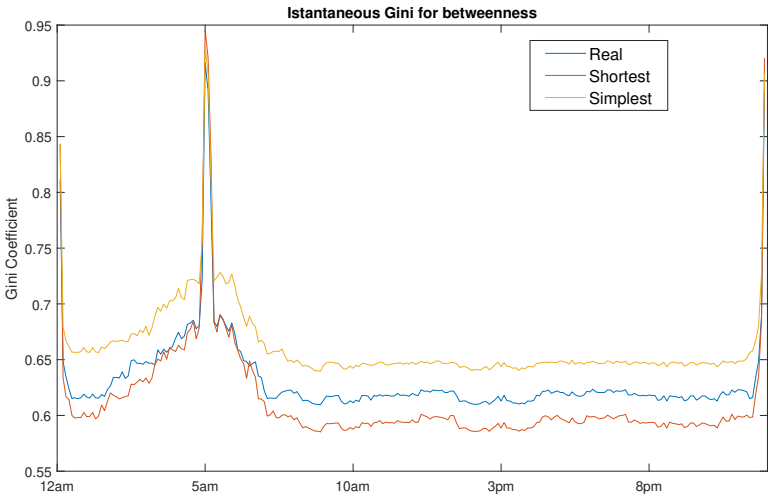

Figure 9: **Instantaneous Gini coefficient of the distribution of the betweeness centrality for the tree different path types.**

maintains its Gini coefficient always between the shortest (minimum) and simplest path. This plot shows and quantifies that while the shortest path method is the most *adaptable*, using different road and exploiting all possible alternatives, the simplest path method tends to pass through the same roads ignoring secondary roads. The observed drivers tend to use not always the shortest path exploring part of the city network that they do not know but they try to remain in the most know road with the consequences of a distribution of road usage between the two other proposed methods.

Conclusions

In this work we analyzed a large dataset of real trajectories reconstructed *ad hoc* with a efficient map matching algorithm. We designed a robust algorithm to count the number of decisions that a path in a road network implies. From this path-based measure of simplicity, we deduced the influence that the factor of simplicity has on real path choice. The strongest results come when we compare the real path with the simplest and the shortest path. In this way, we are able to classify each observed path with the ratio between trip length and number of changes. Different behaviours and priorities for drivers bring to different road choice and consequentially different spatial pattern for traffic. Based on a huge real demand table, we reconstructed the map of the city of Shenzhen with the cumulative link betweenness calculated with the three types of path strategies. The link usage might have application for traffic management and control and also to study the impact of the perception and sensibility of drivers with the particular transportation network design.

We also calculated the average simplicity and shortness of the observed path and this may have application on traffic model, microsimulator calibration and road management.

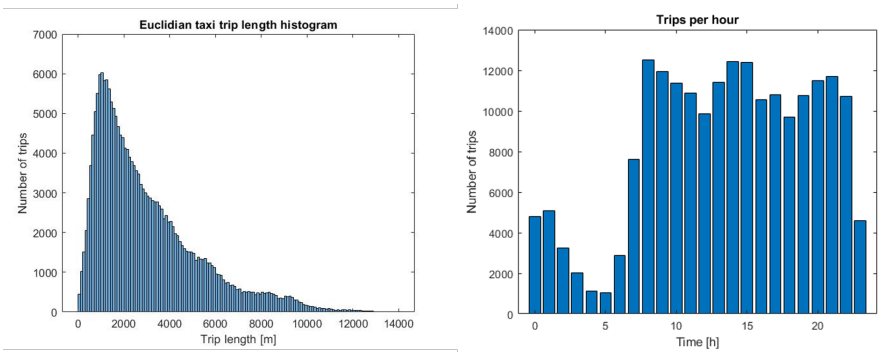

Figure 10: **Statistics about the used demand table for Shenzhen downtown 7th September 2011.** The total number of trips is more than 170,000 with 13,000 during the peak hours. The distribution of the trip lengths has a positive skew and the average is about 3 km and a maximum of 13km.

Finally, with our method we categorize the different kind of trips into 5 categories according to their comparison with the corresponding shortest path and simplest path. In particular, we distinguish drivers that care more about simplicity (trip simpler than the shortest path) from who care more about trip length (shorter that the simplest path).

3 Methods and data description

3.1 Data Analysis on real trajectories

Based on our dataset of more than 170k independent trips traced by about 20 millions of GPS points in Shenzhen downtown, we reconstructed each trajectory matching the GPS points into our simplified map. The frequency of GPS traces is around 30 seconds and whenever the signal was temporarily loss for a trajectory we used a shortest path algorithm to deduce the missing part. Given the high density of information we can assure that our method might affect just slightly the veracity of our results. In Fig. 10 we report the distribution of the trip lengths (on the left panel) and the distribution of trips par hour (on the rigth).

References

[1] Freeman, Linton C. *A set of measures of centrality based on betweenness.* Sociometry (1977): 35-41.

[2] Barthelemy, Marc. *Betweenness centrality in large complex networks.* The European physical journal B 38.2 (2004): 163-168.

[3] Crucitti, Paolo, Vito Latora, and Sergio Porta. *Centrality measures in spatial networks of urban streets.* Physical Review E 73.3 (2006): 036125.

[4] L. da F. Costa and F. A. Rodrigues Seeking for simplicity in complex networks. Europhysics Letters 85 4 (2009).

[5] Viana, Matheus P., et al. *The simplicity of planar networks.* Scientific reports 3 (2013): 3495.

[6] Hill, Russell A., and Robin IM Dunbar. *Social network size in humans.* Human nature 14.1 (2003): 53-72.

[7] C. Gini, Measurement of inequality of income, in: Economic Journal 31 (1921), 22-43.