nature portfolio

https://doi.org/10.1038/s43588-021-00130-y

Supplementary information

Vector-based pedestrian navigation in cities

In the format provided by the authors and unedited

Vector-based Pedestrian Navigation in Cities
Supplementary Information

August 9, 2021

1 Street Networks

In Supplementary Figure 1, we present overview and zoomed-in vies of the street networks used in the study.

2 Walking Paths

2.1 Shortest Paths

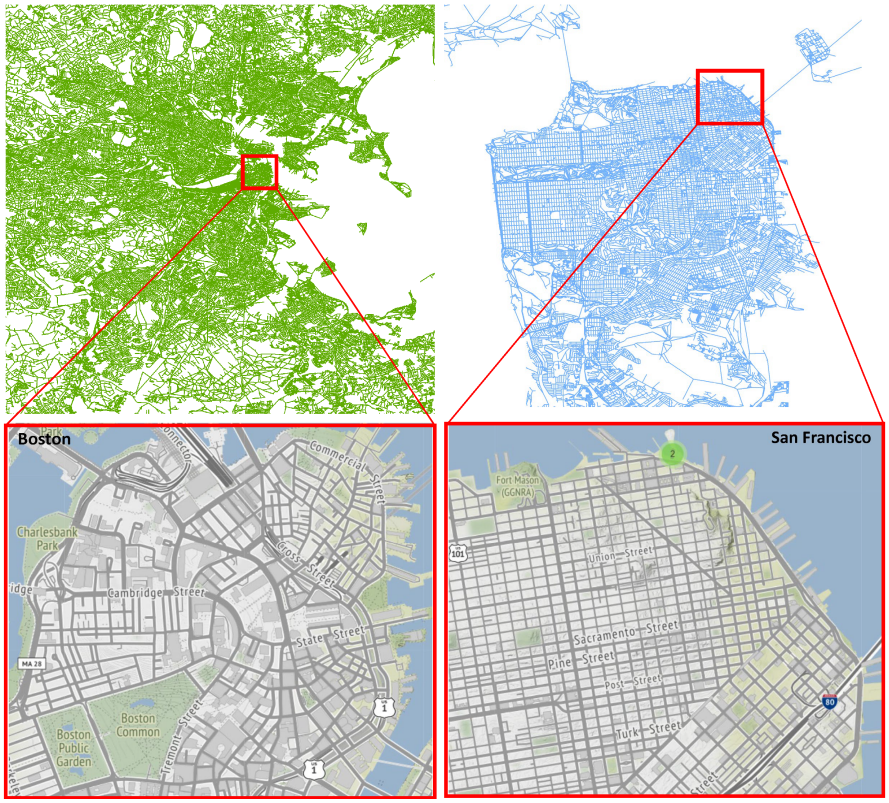
The summary statistics of the shortest paths in both Boston and San Francisco is shown in Supplementary Table 1.

2.2 Human Paths

The summary statistics of the human paths in both Boston and San Francisco is shown in Supplementary Table 1. Supplementary Figure 2 reports the trip velocity distribution extracted from data, in in Boston and San Francisco.

3 Likelihood Cross-Validation

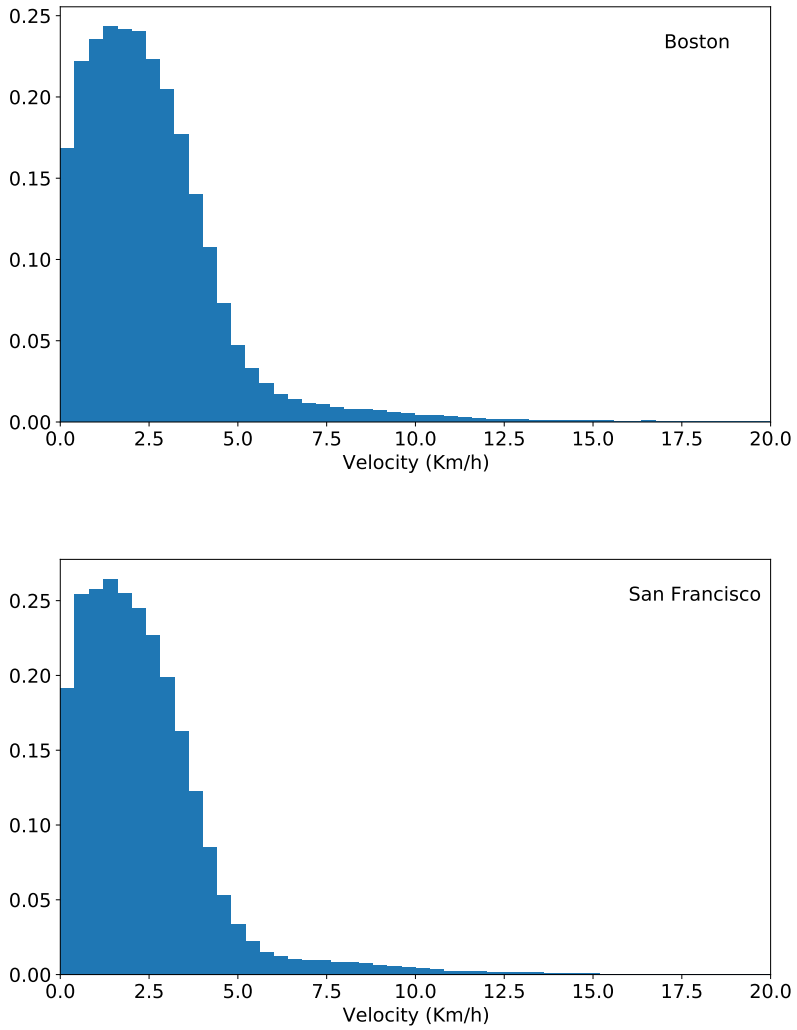
A certain fraction of paths will have zero sample probability in our simulation, even though, in a hypothetical analytical derivation, no path would have a zero probability. Such zero probabilities need to be replaced by a small non-zero threshold c, because taking a logarithm of a zero results in infinite likelihood. Therefore, we associate path probability c to each path with a sample probability smaller than c. To ensure the robustness of the results, we obtained qualitatively similar results with different values of $.000001 < c < .001$. Notably, given the uncontrolled set up of our data-generation procedure, this thresholding would be intrinsically necessary even using analytical probability derivation, since it is not possible to eliminate outliers in human paths. For example, an outlier



Supplementary Figure 1: Overviews (upper) and zoomed-in views (lower) of the street network used in the analysis in Boston (right) and in San Francisco (left).

Supplementary Table 1: Summary statistics of human paths and shortest distance paths after data cleaning.

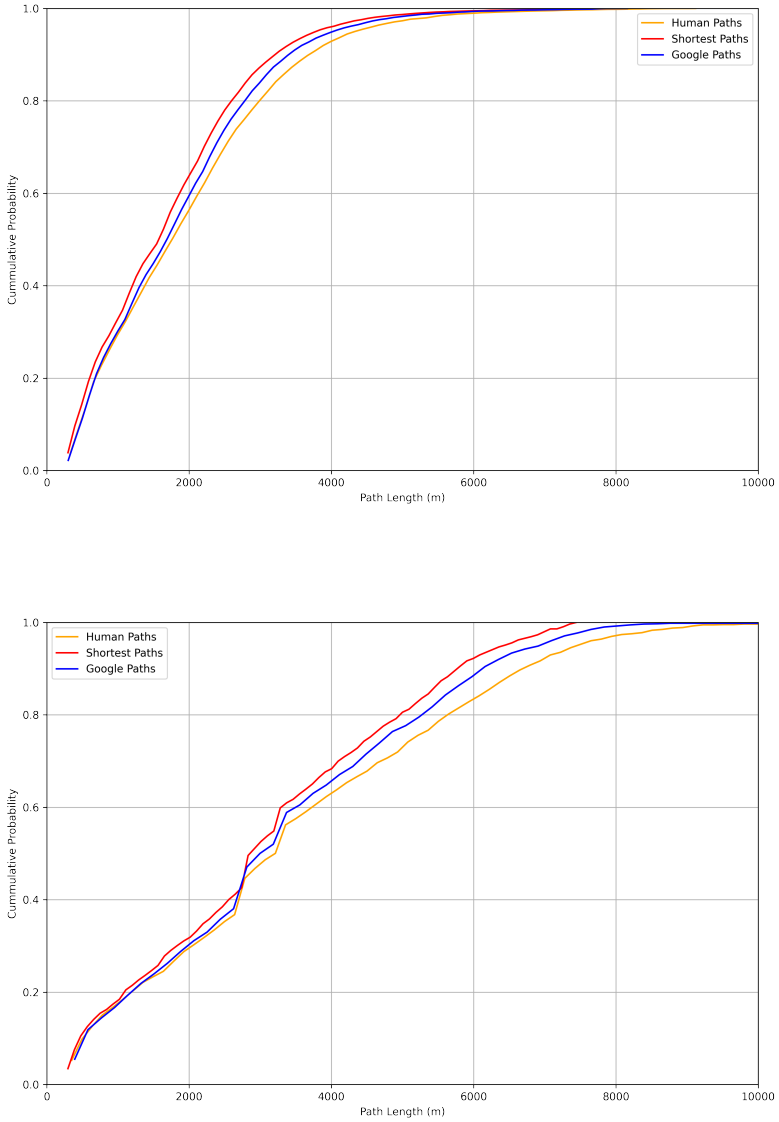
Walking Paths		**Boston**		**San Francisco**	
		Human	Shortest	Human	Shortest
Count		165,645	165,645	189,075	189,075
Length (m)	**mean**	856.0	758.0	868.1	781.7
	std	843.6	718.3	912.1	796.0
	min	200.0	200.0	200.0	200.0
	25%	372.9	345.9	363.7	341.9
	50%	596.4	536.9	583.7	535.3
	75%	1,019.7	899.0	1,017.4	914.9
	max	23,167.1	18,396.8	36,377.7	29,176.8



Supplementary Figure 2: Distribution of path velocity in Boston and San Francisco.

could be a pedestrian who is having a detour to meet a friend, or see a shop. Obviously, no model can predict such cases.

The contour plots in Supplementary Figure 4 show the dependence of the parameter σ on the OD separation and c. The results are qualitatively similar in both cities. When c is low, the optimal σ must be high. This is quite reasonable; in fact, to obtain non-zero probabilities for the outliers, the objective costs of



Supplementary Figure 3: Comparison among the cumulative densities of human, shortest, and Google path lengths in Boston (upper) and San Francisco (lower).

the paths must be strongly perturbed.
Surprisingly, we can see that σ decreases as the OD distance increases, which

Supplementary Table 2: Length and geometric comparison results among human paths, shortest paths, and Google paths in Boston.

Boston	**Path Length (m)**			**Jaccard Similarity**		**Hausdorf Distance**	
	Human	Shortest	Google	H vs. G	H vs. S	H vs. G	H vs. S
count	10,853	10,853	9,254	10,853	10,853	10,853	10,853
mean	1,950.9	1709.1	1,839.9	0.39	0.35	185.2	191.2
std	1,324.2	1152.4	1,209.8	0.31	0.30	182.2	185.1
min	214.1	202.6	202.6	0.00	0.00	0.0	0.0
25%	832.7	731.3	813.8	0.13	0.11	66.4	66.9
50%	1,764.7	1575.5	1,686.0	0.32	0.28	131.6	138.7
75%	2,724.3	2382.5	2,548.4	0.63	0.54	246.2	256.6
max	9,119.9	7868.9	8,159.6	1.00	1.00	1,560.8	2,080.7

Supplementary Table 3: Length and geometric comparison results among human paths, shortest paths, and Google paths in San Francisco.

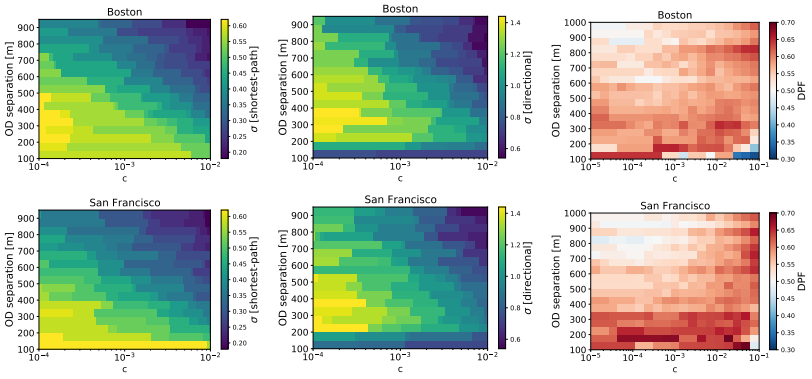
San Francisco	**Path Length (m)**			**Jaccard Similarity**		**Hausdorff Distance**	
	Human	Shortest	Google	H vs. G	H vs. S	H vs. G	H vs. S
count	1,719	1,719	1,492	1,719	1,719	1,719	1,719
mean	3,476.8	3,073.6	3,271.6	0.39	0.38	293.4	311.3
std	2,273.9	1,906.6	2,067.2	0.34	0.37	375.1	374.1
min	207.3	207.3	207.3	0.00	0.00	0.0	0.0
25%	1,683.2	1,508.1	1,611.6	0.09	0.06	42.1	35.9
50%	3,211.5	2,861.9	2,997.6	0.30	0.24	161.8	192.2
75%	5,147.9	4,534.2	4,747.5	0.63	0.67	418.8	432.1
max	11,659.7	7,441.9	15,092.9	1.00	1.00	3,591.5	2,427.8

may be related to the traveling budget-time (Supplementary Figure 5). Whereas pedestrians could treat the shorter paths in a more leisurely fashion, the longer paths are more likely premeditated, hence lower σ.

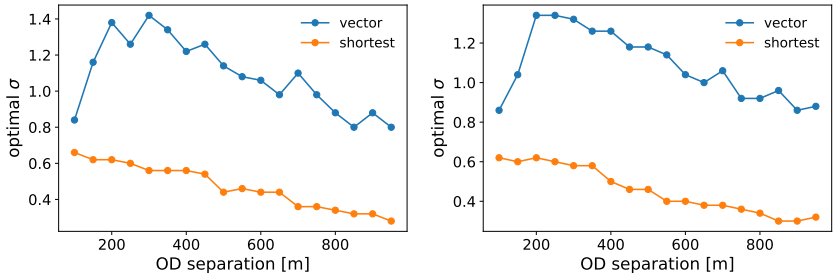
The rightmost panels in Supplementary Figure 4 show the DPF in leave-one-out cross-validation obtained for different OD separation and c. This analysis shows that DPF is weakly affected by c, and that $DPF > 0.5$ for a wide set of values.

4 Pseudo-code of Navigation Algorithms

A pseudo-code description of the navigation algorithms is reported below.



Supplementary Figure 4: Probability threshold calibration. Upper panels refers to Boston, lower to San Francisco. Left panels shows the optimal σ for the stochastic distance minimization model; central panels shows the optimal σ for vector-based navigation, right panel shows the DPF in leave-on-out cross validation. All values are obtained for different c and origin destination distance



Supplementary Figure 5: Optimal sigma for $c = .001$ as a function of the OD separation. Left plot refers to Boston; right plot refers to San Francisco.

5 Additional Analysis

5.1 Individual Performances

In this section we explore to what extent individuals display different performances in their ability of finding a shortest-path to destination. In particular, for each individual we aim to measure the fraction of times he/she correctly identifies the shortest path. Such fraction, computed for the specific individual, can then be compared to the average fraction computed across the entire population, and thus be considered as a metric of individual performance in finding shortest paths.

In order to have sufficient statistical accuracy, we restricted our analysis to the 616 individuals for which we have at least 50 paths recorded in the data set.

Algorithm 1 Calculate Stochastic shortest path

Data: Network, origin, destination, σ
forall $Edge \in Network.Edges$ **do**
  $Edge.Length := Sample(exp(\mathcal{N}(log(Edges.Length), \sigma^2)))$
end
path : = Dijkstra(Network, origin, destination)
return *path*

Algorithm 2 Calculate Vector-based path

Data: Network, origin, destination, σ
Let $dNetwork$ be a directed form of $Network$
forall $Edge \in dNetwork.Edges$ **do**
  $\alpha := \angle(\overrightarrow{Edge}, < \overrightarrow{Start(Edge), destination} >)$
  $Edge.Length := Sample(exp(\mathcal{N}(log(\alpha\ Edges.Length), \sigma^2)))$
end
path : = Dijkstra(dNetwork, origin, destination)
return *path*

At the aggregate level of the entire population of 616 individuals, we have that 33% of times the paths chosen equals the shortest path. However, at individual level we observe a large variation around this average value. In principle, the fraction of paths equal to the shortest path for each individual could be compared with the average value and we could perform a statistical test to check whether the observed deviation from a binomial distribution is significant. However, the observed deviation might be biased since not all the individuals have the same length distribution in their path sets. As shown in Supplementary Figure 6 (a), individuals who under-perform are characterized by a set of path with longer lengths than the control, and, of course, longer paths have lower probability to match exactly the shortest path. To account for this, we must control the bias introduced by differences in path length. This type of problem is typically addressed with matching set theory, that in a multi-dimensional setting would require the definition of a propensity score. However, being in our case the confounding variable only the length, the match can be addressed directly on this variable. Specifically, to asses if a certain individual has a higher (or lower) fraction of path equal to the shortest, its sample fraction cannot be compared directly with average fraction of 33% compute across the entire population; rather, we should select a tailored control set that match approximately the path length distribution of the tested individual. To do that, we binned the length of the path in steps of $50m$, and for each individual we sampled randomly a set of path that match the count of path on each bin of the individual under study. From such random sample, we obtain the control proportion of path equal to the shortest path. We repeated such a random sample 100,000 times. In Supplementary Figure 6 (c),(d),(e), we report the outcome of this process for three individuals who under-perform (c), perform the same (d), or

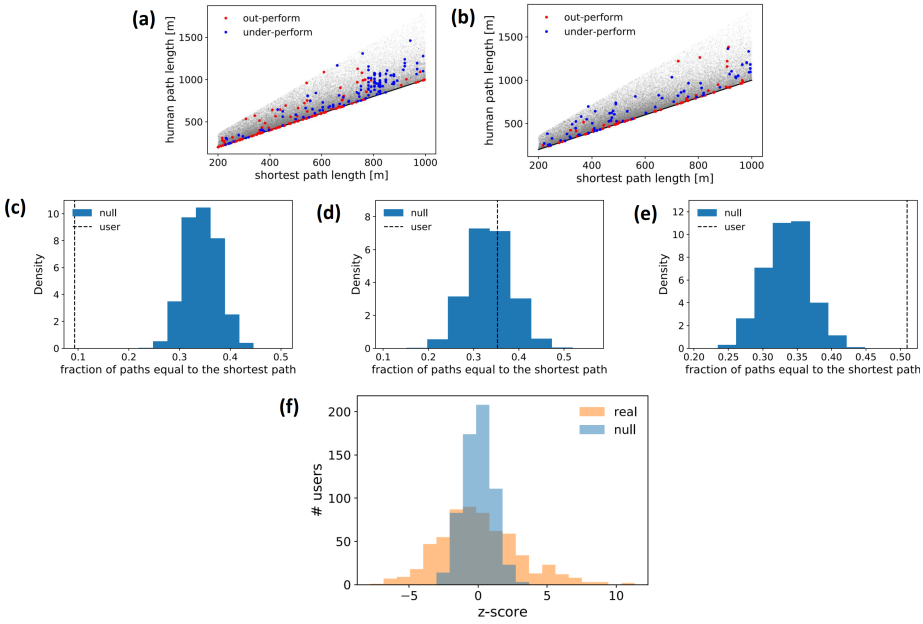
over-perform (e) with respect to the null distribution. The null distribution converges to a Normal being the sum of independent Bernoulli trials, therefore in Supplementary Figure 6 (f), we can use a z-score to highlight the presence of outliers. After this matching in Supplementary Figure 6 (b), we show two users, one that under-performs and another that out-performs and both show to be unbiased with respect to the length distribution.

5.2 First Segment Strategy

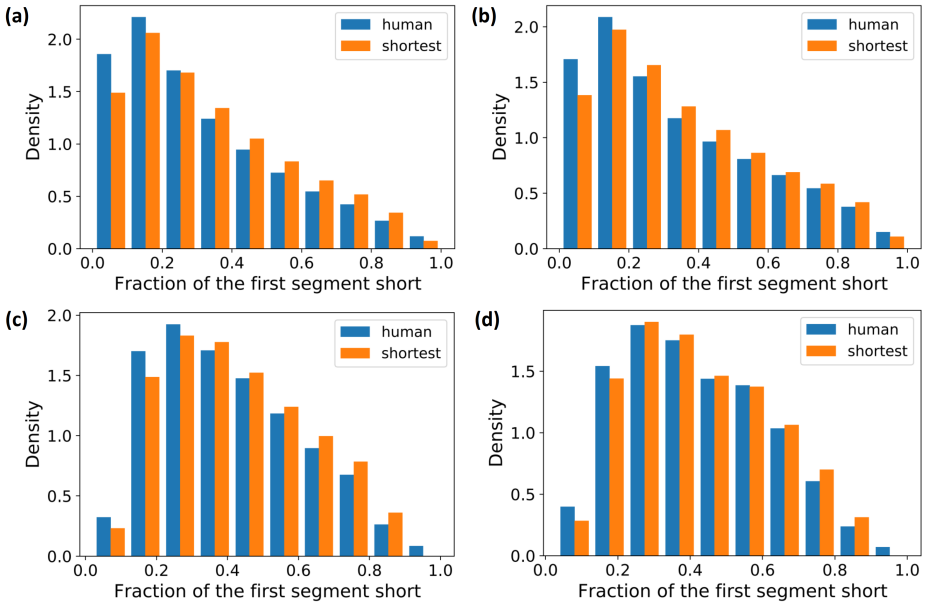
A possible cause of the observed asymmetry in human paths might be the tendency of selecting a relatively long straight segment at the beginning of the path, a tendency that has been observed in the literature and referred to as the Initial Straightest Segment (ISS) strategy. In this section, we perform a statistical test to assess this hypothesis.

In order to define the fist straight line sub-segment of each path, we simplified the trajectory with the DP algorithm with a cutoff of 30m. Then, we measured the path length for the first two points in of the DP-simplified representations. The average length covered by the first segment for humans in Boston is 219m while for the shortest path is 226m; for San Francisco both human and shortest path cover on average 256m. By looking at the fraction of the total path length covered by the first segment (Supplementary Figure 7), in both cities we have observed that this fraction is larger for shortest than that of human paths. To further confirm this observation, we also restricted our analysis to paths with an OD separation of at most 300m. In this case, the absolute length covered by the first segment is slightly longer for the human paths (133m in both Boston and San Francisco) than for the shortest paths (130m). However, the fraction of the total path length covered by the first segment still shows higher value for the shortest path, as depicted in the lower panel of Supplementary Figure 7.

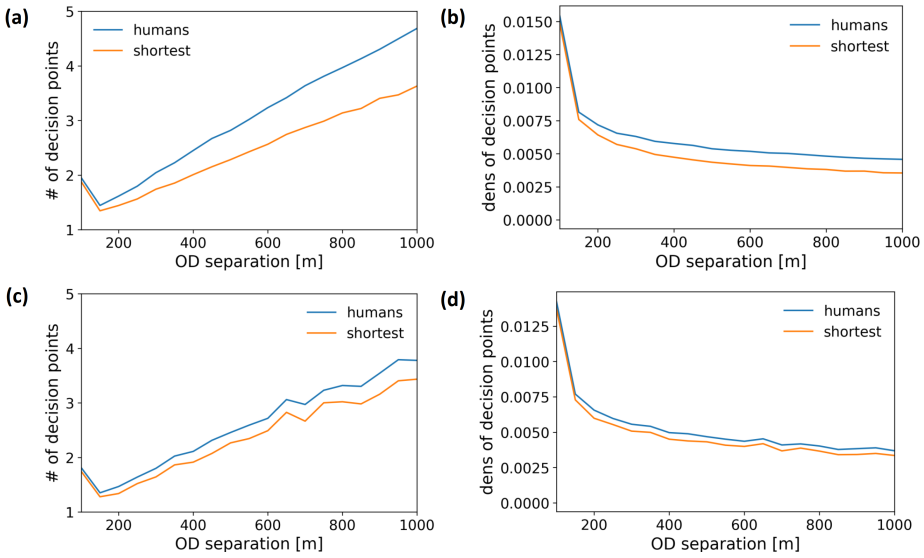
To account for that, for the path with OD separation smaller than 300m we modeled straight-first segment propensity by setting a cost equal to zero to all straight segment that depart from each origin and searching for the shortest path. To do that, we need to identify the collections of street segments without a significant angle starting from the origin. This problem can be addressed by considering the dual edge-edge network. In this representation, a node is a street segment, and two street segments are linked if they are connected by a street intersection. The link among two street intersections can be weighted by the angle among the two street. Therefore, in order to select the collection of all paths with a minimal or smooth angle from the origin, it is enough to cut all links (in the dual representation) with an angles higher than 20 degree and then find the set of street intersections that lie in a connected component that contains the origin node. However this approach fails in reproducing the slight increment in the first segment length observed on the human path: in fact, as result of this modeling the average total length covered by the first segment reaches 165m and 168m for Boston and San Francisco respectively, which are significantly longer than the observed ones.



Supplementary Figure 6: (a) each dots correspond to a path; red dots are the paths of an individual whose fraction of paths equal to the shortest path is significantly lower than the average fraction of 33%; blue dots represent an individual whose fraction of path equal to the shortest path is significantly higher than 33%. Figure (b) shows other two individuals who under-perform and out-perform after bias in path length distribution has been removed through matching. Figures (c),(d),(e) are three example of individuals with corresponding proportion of matching the shortest path (dotted line), compared with the expected proportion obtained from a control that match their length distribution. Figure(f) shows the z-score of each individual, showing that the distribution of z-scores of individuals in the data set is broader that the expected distribution of z-score (in blue) in case of no outliers.



Supplementary Figure 7: Distributions of the fraction of path length covered by the first segment in Boston (a) (c) and San Francisco (b) (d). The upper panels refers to all the paths, the lower panels are restricted to paths whose OD separation is at most 300m.



Supplementary Figure 8: (a) (c) Average number of decision points as a function of the OD separation, (b) (d) density of decision points as a function of the the OD separation. (a) and (b) refer to Boston; (c) and (d) refer to San Francisco.

5.3 Decision Points

In this section we explore the hypothesis that humans might have a tendency to minimize the number of decision points (road intersections) in their trajectory. To define such a number, we simplified each trajectory with DP algorithm with a 30m threshold. After that, the number of decision points is the total number of simplified segments minus one. We performed this analysis both on shortest paths and human paths for different OD separations. In upper panels of Supplementary Figure 8, the average absolute number of decision points per trajectory is systematically higher on the human trajectory in both cities. We further computed the density of decision points since human paths are on average longer than shortest paths. The density of decision points were calculated as the number of decision points divided by the path length. In both cities, we reported a higher density of decision points for the humans than their shortest counterparts. We can then conclude that minimizing the number of decision points in the trajectory is likely not a significant factor in pedestrian path formation.