Quantifying navigation complexity in transportation networks

Zhuojun Jiang [ID], Lei Dong*, Lun Wu and Yu Liu [ID]*

Institute of Remote Sensing and Geographical Information Systems, School of Earth and Space Sciences, Peking University, Beijing 100871, China
*To whom correspondence should be addressed: Email: arch.dongl@gmail.com; liuyu@urban.pku.edu.cn
Edited By: Javier Borge-Holthoefer

Abstract

The complexity of navigation in cities has increased with the expansion of urban areas, creating challenging transportation problems that drive many studies on the navigability of networks. However, due to the lack of individual mobility data, large-scale empirical analysis of the wayfinder's real-world navigation is rare. Here, using 225 million subway trips from three major cities in China, we quantify navigation difficulty from an information perspective. Our results reveal that (1) people conserve a small number of repeatedly used routes and (2) the navigation information in the subnetworks formed by those routes is much smaller than the theoretical value in the global network, suggesting that the decision cost for actual trips is significantly smaller than the theoretical upper limit found in previous studies. By modeling routing behaviors in growing networks, we show that while the global network becomes difficult to navigate, navigability can be improved in subnetworks. We further present a universal linear relationship between the empirical and theoretical search information, which allows the two metrics to predict each other. Our findings demonstrate how large-scale observations can quantify real-world navigation behaviors and aid in evaluating transportation planning.

Keywords: network navigability, search information, transportation networks, city science

---

Significance Statement:

Evaluating network navigability is crucial for understanding transportation efficiency. However, the real-world wayfinder navigation is understudied due to the lack of mobility data. Using large-scale subway trips from three mega-cities in China, we quantify navigation difficulty via information bits and find that the decision information in the subnetworks formed by repeatedly used routes is much smaller than the theoretical upper limit shown in previous studies. By modeling routing behaviors in growing networks, we show that navigability can be improved for actual trips even if the theoretical navigability of the network is deteriorating. We also reveal a universal linear relationship between the empirical and theoretical navigation information. These findings partially explain why people are not getting lost in complex transportation networks.

Introduction

With the expansion of urban transportation networks, the efficiency and navigability of cities have attracted increasing attention (1–5). In particular, the rapidly increasing number of edges (routes) in transportation networks has increased the complexity of navigation while making it easier for people to move around a city (6, 7). However, interestingly, people are not "lost" in complex transportation networks, which makes us wonder how individuals navigate during network evolution. The exploration of this question has important implications for understanding the correlation between the navigability of transportation networks and public travel behaviors.

Quantifying the routing costs in networks has offered novel insights into the navigation problems of transportation networks (8–11), brain networks (12–14), social networks (15), wireless networks (16), and many other disciplines (17–20). Among these studies, the information approach proposed by Rosvall et al. (9, 10) is an important starting point. By modeling navigation in a road net-
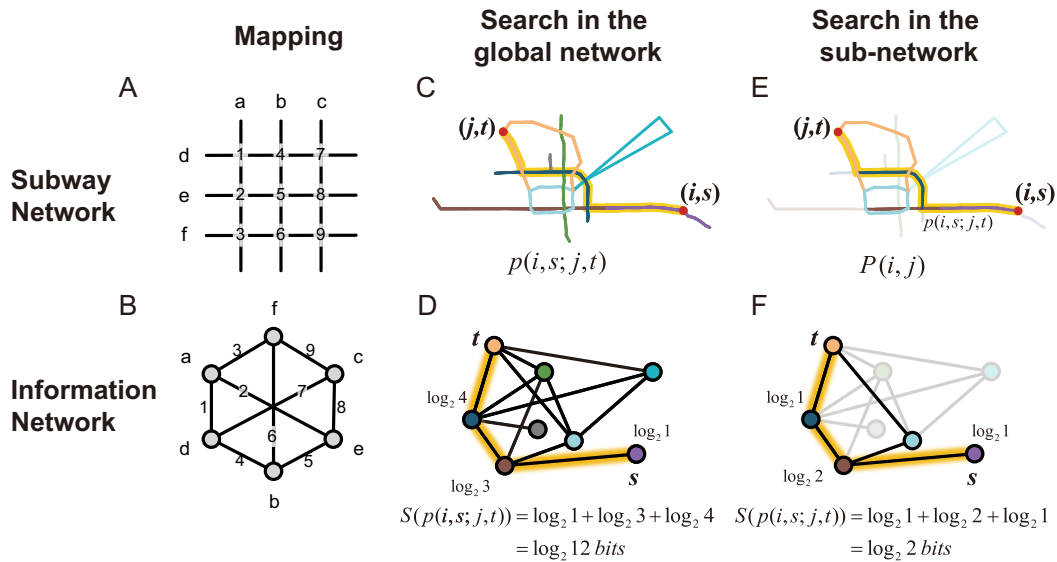
work as a signal transmission process, Rosvall et al. developed a "search information" metric to quantify navigation difficulty (9, 10). Unlike route choice models, which are widely used in transportation engineering (21–23), search information focuses more on measuring the complexity of the network [rather than modeling travelers' route choices and traffic assignment (24, 25)]. Specifically, the search information can be easily understood as the number of yes or no questions that a traveler has to answer when locating a route from a bird's-eye view of a map. However, previous studies on search information typically assumed that travelers have a global view of networks (8–11, 17), ignoring the critical fact that travelers generally use information from only part of the network during their actual navigation (26–28). More importantly, due to the lack of mobility data, previous studies often did not consider the actual traffic between network nodes, making it difficult to reflect the complexity of navigation in real networks.

Here, using 225 million subway ridership records from Beijing, Shanghai, and Shenzhen (three mega cities in China), we estimate

---

Fig. 1. Search information. (A) A simplified schematic of a subway network. (B) The corresponding information network of panel (A). Nodes represent subway lines (a–f), and edges represent transfer stations (1–9). (C) Route $p(i, s; j, t)$ (highlighted in yellow) from station i on line s to station j on line t. (D) Search in the global information network. One needs to answer $\log_2 k$ yes or no questions to locate the next line or $\log_2(k-1)$ excluding the line that it has come from. (E) Search in the subnetwork. We delimit the subnetwork of actual trips by the stations and lines within the set $P(i, j)$ of route choices from i to j. Subway lines without consideration are in faded colors. (F) The information network in solid colors is mapped from the subnetwork in panel (E). In contrast with the 3.6 ($\log_2 12$) yes or no questions based on global search in panel (D), travelers only need to make one decision during their search in the subnetwork ($\log_2 2$), i.e. to decide whether to travel on the dark blue or light blue line.

the route for each origin–destination (OD) record and obtain the subnetwork formed by route choices to represent traveler networks for route planning. By mapping the subway network into an information network, we calculate the search information (the measure of navigability) along the path in the information network (Fig. 1). According to ref. (9), the $\log_2 k$ bit is the minimal information necessary to locate the next move from k options, where k corresponds to the node degree in the information network [in Fig. 1(D), at each node except the origin node of the path, the incoming path is excluded from the k options, resulting in an information value of $\log_2(k-1)$ bits]. Notably, the degree k of the same node may be smaller in the subnetwork than in the global network, indicating that people are navigating with fewer line options (for transferring) in practice (Fig. 1).

We observe that most people only make a small number of route choices during a trip, which implies that people use a small subnetwork when planning their routes. This directly results in lower search costs during actual travel than in theoretical studies (8, 9, 11). By applying a simple route choice model to the historical subway networks of the three cities, we further investigate the impact of network growth on the decision information. We find that as the network grows, the navigation complexity of the subnetworks remains unchanged or even declines, while the theoretical global navigation complexity increases significantly. These differences suggest that the navigability of a network can be improved for actual travel behaviors, even if the global navigability is reduced during network growth. Moreover, we find a universal 3/4 linear relationship between empirical search information and theoretical search information, reflecting the hidden correlation between these two metrics in the complex network. This work bridges existing network navigability studies with traveler routing behaviors and has the potential to be used in evaluating urban transportation planning and understanding navigation in cities (29–32).

Results

The conserved number of routes

We estimate actual traveler routes using smart card data, subway networks, and a travel survey dataset. Here, we briefly introduce the data process; the detailed data description, the travel time estimation of a path, and the discrete choice model are provided in the "Materials and methods" section. First, for a given station pair, the top 10% and the bottom 10% of smart card records are trimmed according to the travel time distribution to exclude abnormal data. Second, for each record, those paths in the subway network whose temporal distances are within 10 min from the travel time of the record, are the candidate paths to be matched. Third, by adopting a discrete choice model considering travel time, distance, and the number of transfers, we estimate the probability of each candidate path being chosen. The parameters of the model are calibrated by the travel survey data (see the "Materials and methods" section). Finally, we match each record to the path with the largest probability of being chosen. To verify the matching results, we aggregate the flows by subway lines and calculate the correlation between the matched results and the official published number during the same period in Beijing (Supplementary Material Appendix, Fig. S1) (33). The high goodness of fit ($R^2 \approx 0.99$) indicates the effectiveness of our matching method (Fig. 2B).

The matched results show that, despite the large number of possible routes between a station pair, most people only use a few routes (Fig. 2C–E). For example, in 95% of the station pairs, people take no more than 4 routes in Beijing and Shanghai, and in Shenzhen, this ratio is 99%. Among those limited number of routes, we further find that more than 90% of people follow the simplest route to travel in the network (Supplementary Material Appendix, Fig. S2), i.e. the route with the fewest transfers (34). As more transfers usually mean higher time costs due to walking and waiting
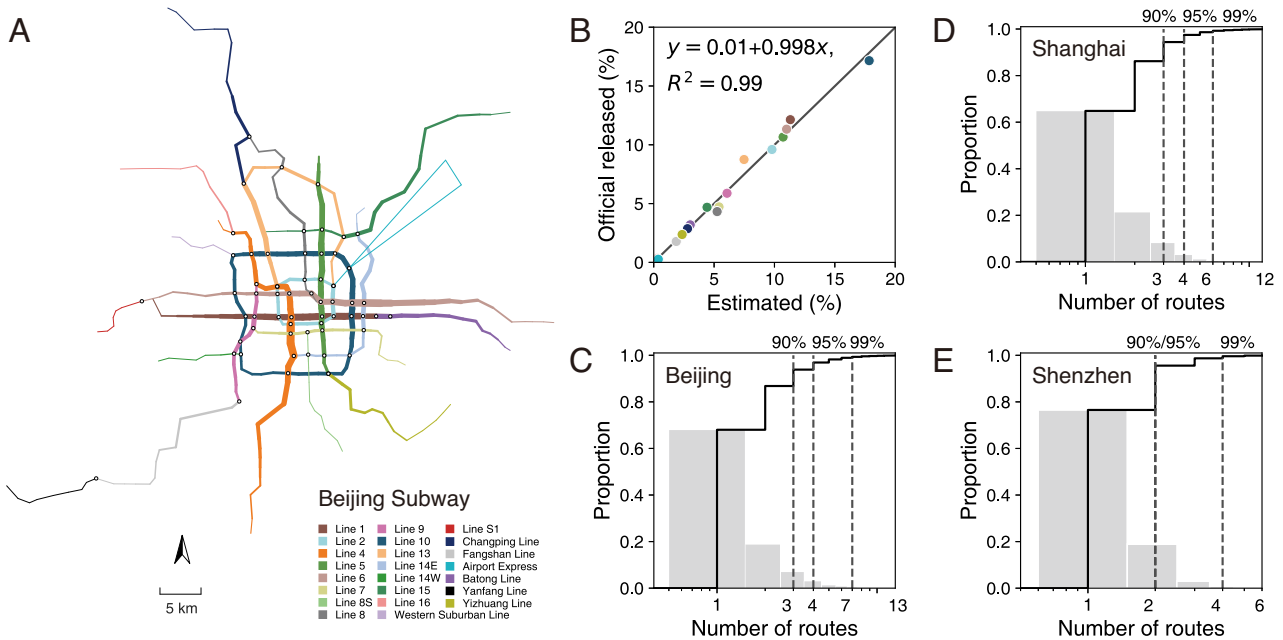
Fig. 2. Route matching based on smart card records of subway systems. (A) The estimated flow map of the Beijing subway network. (B) Correlation between the estimated flow and the official released flow on 16 lines in the Beijing subway. Each point represents a subway line. The flow is normalized by the summation value and indicates proportions in the plot. (C–E) The number of routes followed by travelers between stations in the three studied cities. The distributions are displayed in a histogram (gray) and a cumulative distribution (black), respectively. The dashed lines mark the number of routes followed by travelers between 90%/95%/99% of the station pairs.

for the next transfer (35), the simplest route can minimize transfer costs and tends to dominate people's choices.

Information measures

The adopted number of route choices observed in our data implies that actual travel occurs within subnetworks. It is possible that the information used for rider decisions consists of more than the subnetworks formed by the actual routes. However, since we can only observe traveled routes in our datasets, we assume that people's travel decisions are based on subnetworks formed by route choices in this paper. Here, we use the term "route" to denote people's route choices and the term "path" to denote the segments on the networks (the two terms are interchangeable).

First, we measure the navigation complexity in the information network via the information bits. Recall that a matched path $p(i, s; j, t)$ from station i on line s to station j on line t can be mapped to the information network (Fig. 1F), where each node represents a subway line and each edge represents the transfer station of the connected lines. According to ref. (11), the total amount of information S needed to locate $p(i, s; j, t)$ in the information network is

$$S(p(i, s; j, t)) = \log_2 k_s + \sum_{n \in p(i, s; j, t)} \log_2(k_n - 1), \quad (1)$$

where k_s is the degree of node s. k_n is the degree of node n on path $p(i, s; j, t)$ (except s and t), and one has to locate the next move from the $k_n - 1$ options. S measures the difficulty of locating a particular path, and a lower S means that the network has better navigability between the station pair. When there are multiple matched paths between i and j, we calculate the station-level empirical search information (ESI) by performing the flow-weighted average on all

the matched paths $P(i, j) = \{p(i, s; j, t)\}$ connecting i and j (see the "Materials and methods" section). The value of ESI indicates how difficult it is for travelers to find their way from i to j during the actual trip.

Unlike the definition of search in the subnetwork formed by actual trips, the theoretical search information (TSI) defines the information required to find the simplest path in the global information network (9, 11), i.e. the path with the fewest nodes from s to t. There may be multiple simplest paths between the stations; for simplicity, we use the fastest simplest path to define the station-level TSI according to ref. (11). The calculation for the TSI is the same as Eq. (1), but the node degrees k_s and k_n are calculated in the information network transformed from the whole network.

Figure 3(A) shows that the ESI values of most station pairs (> 81%) are smaller than the theoretical global values. In the Beijing subway, the mean value of ESI and the mean value of TSI are approximately 0.7 bits and 4.0 bits, respectively, indicating that the decision cost for travelers to determine their routes is much smaller than what they theoretically need to spend. Similar results can also be found in Shanghai and Shenzhen (Supplementary Material Appendix, Fig. S3). The main reason is that the subnetworks formed by route choices only include the lines and transfer stations that people have considered and adopted, and the number of these lines and transfer stations is far less than that of the whole network. This directly results in the node degrees of the subway lines in the subnetworks being smaller than those in the whole network when calculating the amount of information (Fig. 3(B) presents an example).

Although the TSI is greater than the ESI between most stations, we find that a small number of station pairs in the three cities have an ESI equal to a TSI (Fig. 3A). Most of these station pairs
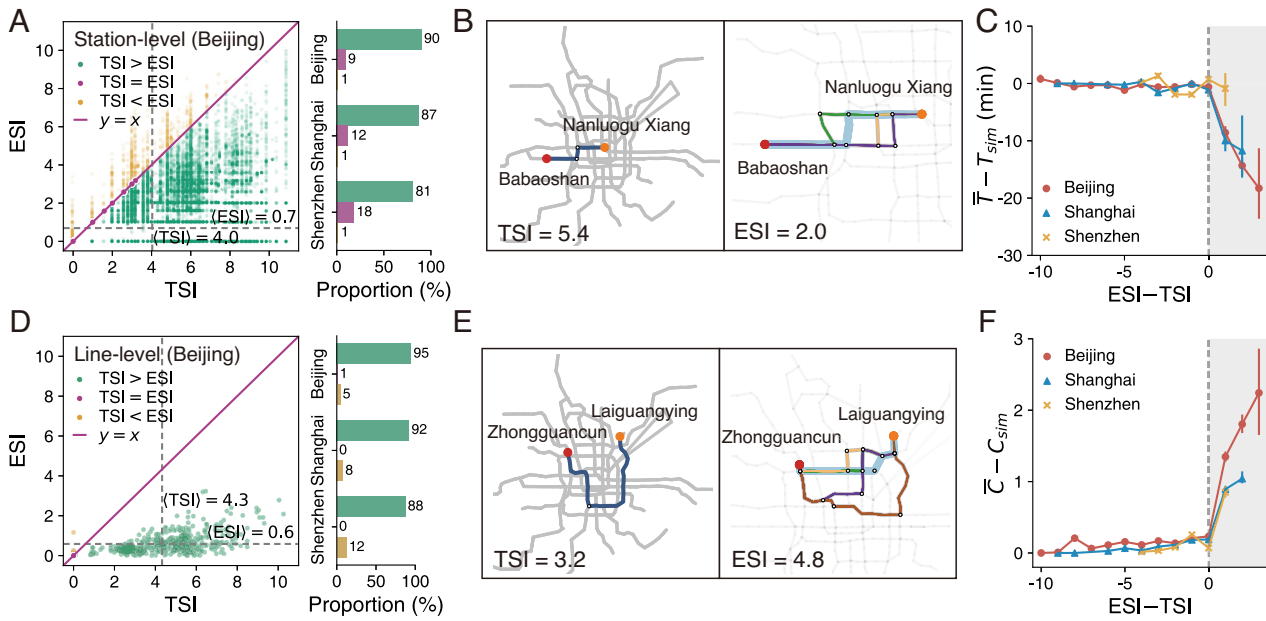
Fig. 3. Station-level and line-level search information. (A) The station-level ESI and TSI. In the left panel, each dot represents a station pair, and $\langle\cdot\rangle$ denotes the average over all pairs. In the right panel, the bars show the proportions of station pairs with TSI > ESI, TSI = ESI, and TSI < ESI. (B) An example of a station pair whose TSI > ESI. The subnetwork (right) formed by the four matched paths contains fewer lines and transfer stations than the global network (left), which makes ESI < TSI. (C) Time difference $\overline{T} - T_{sim}$ and information difference ESI−TSI at the station-level. For a station pair, $\overline{T}$ is the flow-weighted average of time for all the matched paths, and T_{sim} is the travel time of the simplest path. (D) The line-level ESI and TSI. Similar to panel (A). (E) An example of a station pair whose TSI < ESI. Between this station pair, people choose the paths with more transfers ($\overline{C} = 4.0$, $\overline{T} = 61.6$ min) in the subnetwork than the simplest path ($C_{sim} = 1$, $T_{sim} = 89.7$ min) in the global network, making the ESI exceed the TSI. (F) Transfer difference $\overline{C} - C_{sim}$ and information difference ESI−TSI at the station-level. The bars in panels (C, F) denote the 95% CIs.

are located on the same subway lines without any transfers, and hence both empirical and theoretical decision costs are zero.

A counter-intuitive finding is that there are still 1% station pairs with ESI > TSI (Fig. 3A and Supplementary Material Appendix, Fig. S3). This is mainly because people will trade higher navigation complexity for shorter times on some trips . Figure 3(C) shows that when ESI > TSI (i.e. the higher navigation complexity), the travel time between stations is shorter than the simplest path time ($\overline{T} - T_{sim} < 0$). Figure 3(E) shows a typical example. From Zhongguancun to Laiguangying, most people choose shorter paths with four transfers rather than the simplest path with one transfer in the whole network. Since the search information is accumulated at each transfer along the path, the increased number of transfers (i.e. $\overline{C} - C_{sim}$) makes the ESI more likely to exceed the TSI (Fig. 3F). Further analysis of station pairs with ESI > TSI also verifies our assumption (Supplementary Material Appendix, Fig. S4): The actual paths with more transfers are much shorter in travel time than the simplest path with fewer transfers.

In addition to the complexity of navigation between stations, another important dimension is the navigation between lines. In subway travel, the most important decision is how to best reach the destination through the choice of lines (rather than stations). Therefore, we average the amount of information on all the route choices that start on line s and end on line t to obtain the line-level search information S_l between the line pair s to t (see the "Materials and methods" section). In Fig. 3(D) and Supplementary Material Appendix, Fig. S5, we again observe that more than 87% of the line pairs have a lower ESI than TSI. The similar results between the station-level and line-level search information suggest that the small decision costs are universal across navigation dimensions in empirical navigability.

Evolution of navigation complexity

Subway systems in China have changed dramatically over the past two decades, giving us the opportunity to investigate the impact of network growth on navigation. To better understand the navigation complexity in historical networks without real-world trip data, we model navigation behaviors using the following steps: (1) generating a choice set, i.e. m alternative paths by the repeated shortest path algorithm (36) and (2) making the route decision based on the discrete choice models used in matching smart card records with network paths (see the "Materials and methods" section). In the second step, we further assume that the path with the highest probability (among the m paths) can represent the final route choice and use this path to measure the navigability in actual trips.

Specifying the size of choice set m is challenging for route choice modeling and it is usually based on laboratory settings due to the lack of empirical support (37, 38). Here, based on our findings (Fig. 2C–E), we use $m = 13$ in Beijing, $m = 12$ in Shanghai, and $m = 6$ in Shenzhen to characterize the choice sets and infer the subnetworks that people use for route decisions. To validate the robustness of the results, we also conduct tests on different values of m and obtain similar results (Supplementary Material Appendix, Figs. S6 and S7).

Figure 4(A) shows that both ESI and TSI increase over time as the network grows larger, but the growth rate of ESI is substantially lower than that of TSI. To obtain a closer look at the changes, we classify the trips into three groups based on the changes in value from the search information. The results show that the search information for the vast majority of trips is increasing in the global network (Fig. 4B). In Beijing, for example, more than 70% of the inter-line trips have an increased TSI. However, for search
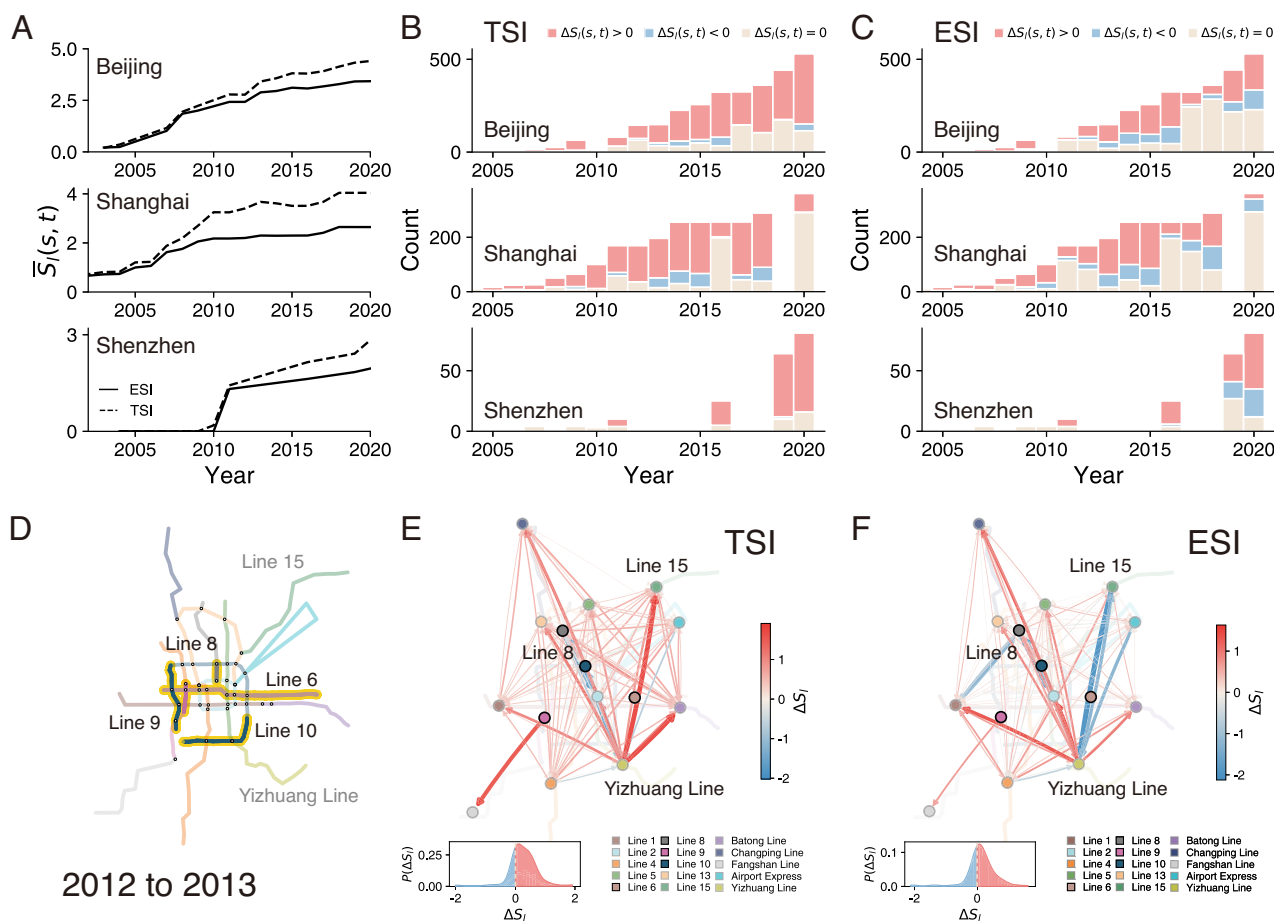
Fig. 4. Line-level search information. (A) The development of $\overline{S}_l(s,t)$ for ESI and TSI. (B) Line-level TSI is decomposed into three categories. Each bin represents the number of line pairs in the subway network of the year. The colors distinguish the line pairs whose TSI values increase (red), decrease (blue), or remain unchanged (pale) due to the network changes in that year. (C) Similar to panel (B), the ESI is decomposed into three categories. (D) The network changes in the Beijing subway from 2012 to 2013. The newly opened segments/lines are highlighted in yellow. (E, F) The impact of network changes on navigability. The inter-line trips are drawn in the arcs that point from the start lines to the end lines. The colors indicate the differences in search information between 2012 and 2013. The extension of Line 10 connecting the Yizhuang Line makes navigation from the Yizhuang Line to other northern lines, such as Line 15 and Line 8, much easier (F), while it increases the difficulty of global navigation for most trips (E).

information in the subnetworks, we observe that most trips are not affected or are even easier to navigate during network growth (Fig. 4C). Especially since 2017, more than half of the inter-line trips have an unchanged or reduced ESI each year in Beijing and Shanghai, which reflects a great improvement in the actual navigability. When looking at the specific trips (Fig. 4D–F), we also find that adding subway lines can facilitate navigation for travelers, even if the difficulty of the theoretical global navigation increases.

Figure 4 focuses on the average value of the search information, while the maximum value is also essential for a transportation network since the maximum value determines the navigation obstacle. Similar to Fig. 4(A), Fig. 5(A) shows that the maximum values of ESI and TSI increase as the network grows. In particular, the 8-bit upper limit of the TSI previously found in New York City (11) can also be verified in Shanghai (Fig. 5A). In addition to these specific values, we find a universal linear relationship with a slope of 3/4 between these two metrics despite the apparent complexity of network dynamics (Fig. 5B). This finding expresses a very tight quantitative constraint on navigation behaviors and allows the empirical search information to be predicted by global search information. Given the 8-bit upper limit of the TSI and the 3/4 relationship, we estimate that the upper limit of the empirical search is approximately 6 bits, which is confirmed in the three cities (Fig. 5A).

Discussion

In this work, we characterize traveler route choices from large-scale datasets and find that (1) people adopt a small number of repeatedly used routes and tend to prefer the simplest route; and (2) the decision cost for wayfinding in the subnetworks is much smaller than that in a theoretical global search. In analyzing the search information required for actual navigation, we use the subnetworks formed by route choices. This may miss some additional routing information, as people may use more alternative routes for decision-making, even if those routes are never traveled (in the observed data). These unobserved routes would make search information values larger than the empirical results in this paper but would still be lower than the global theoretical values.

We discover a 6-bit information limit from dynamic networks, which coincides with the bound of the human working memory capacity of 2–6 items in cognitive sciences (39–41). The 6-bit search information can also be linked to the number of connections in the network (which corresponds to the edges in the information network). According to ref. (11), the approximate relationship between search information S_l and the number of connections K_u of the information network is $S_l = \log_2 K_u$, when considering paths with $C = 2$ connections (note that this equation is derived assuming that the subway network is close to a regular
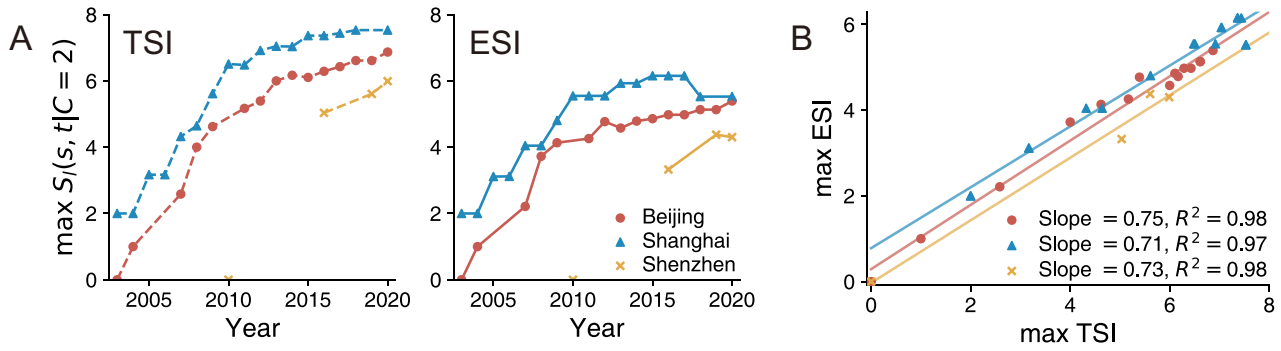
Fig. 5. The relationship between line-level TSI and ESI. (A) The development of maximum TSI and maximum ESI for locating the fastest simplest paths with $C = 2$ connections in the information network. (B) Maximum TSI versus maximum ESI. Each point represents the search information value of a year in panel (A). The maximum ESI is strongly correlated with the maximum TSI, and the slopes of the fitting lines are approximately 0.75.
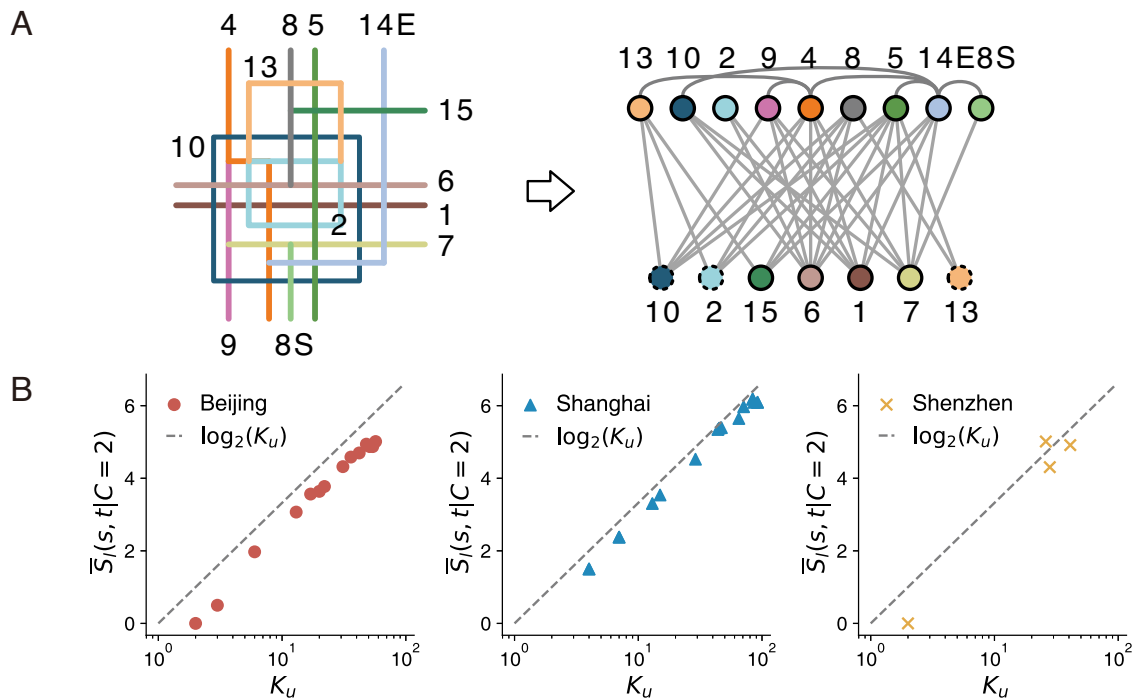
Fig. 6. Line-level search information and the number of network connections. (A) Beijing subway network. Beijing subway lines have a grid-like structure (left), and the information network is similar to the bipartite map (right). As loop lines (Lines 2/10/13) act both as horizontal and vertical lines, we express them as two points (solid and dashed) with the same color in the information network. (B) Considering a path with two connections, the average amount of line-level search information needed to locate the simplest path is close to $S_l = \log_2 K_u$, where K_u is the number of edges in the information network or the number of connections in the subway network (11).

lattice, as shown in Fig. 6(A); see (11) for details). We also observe similar relationships between $\overline{S}_l(s, t|C = 2)$ and the number of connections in the dynamic networks of the three cities (Fig. 6B). Therefore, the 6-bit limit implies that people are disturbed by 64 (2^6) connections in the subnetwork during the search for the most complex trips.

Notably, both our 6-bit and the previous theoretical 8-bit limits are obtained when only the routes with two transfers are considered (11), and in fact, many trips require more than two (some require even five) transfers to reach the destination. Therefore, the information limit summarized from the two transfer routes could be limited in representing the complexity of the network and defining the navigation complexity at different scales (or transfer numbers) can provide a more comprehensive evaluation of the network structure (42). Furthermore, identifying information limits may be critical, but reducing these travel complexities and designing a well-navigated transportation network are more profound.

Our work is helpful in evaluating transportation planning (43) and has implications for understanding human navigation in cities through large-scale datasets (44, 45). With the help of big data, quantifying real-world navigation behaviors may shed light on research in spatial cognition and psychology (32, 46–48), which is usually based on experimental data from volunteer participants (49–51). Our understanding of navigation complexity is still limited; especially due to the limitations of the data, we only study a subway network, which is a relatively simple public transportation network. Further research on multimodal networks (e.g. sub-

ways, buses, and other modes of transportation) and spatial cognition mechanisms will provide a deeper understanding of urban navigation (29–31).

Materials and methods
Data
The 225 million smart card records include 10.8 million users of the Beijing subway in 2019 May, 10.3 million users of the Shanghai subway in 2015 April, and 2.7 million in the Shenzhen subway in 2017 October (Supplementary Material Appendix, Table S1). Each record consists of a card identification, the entry and exit timestamps, and names of the entry and exit stations and lines. For each record, the interval between the entry and exit timestamps is calculated as the travel time of that trip.

The subway network data include 15 snapshots of the Beijing subway from 2003 to 2020, 18 snapshots of Shanghai from 2000 to 2020, and 8 snapshots of Shenzhen from 2004 to 2020 (Supplementary Material Appendix, Tables S2–S4). We obtain the stations and lines from the official online subway maps. The coordinates of the stations and the travel time between stations are collected from Baidu Maps.

To calibrate the discrete choice model and validate the estimated travel time adopted in this paper, we collected 272 subway trips with known routes and durations through questionnaires in the three studied cities (Supplementary Material Appendix, Figs. S8 and S9). Participants were asked to upload a screenshot of their subway ride record, including the entry and exit stations and times and the subway lines and transfer stations they took during the trip.

Travel time estimation and validation
In a subway network, nodes represent stations and edges represent subway lines connecting two consecutive stations. We set the in-vehicle time between a connected station pair as the temporal attribute of the edge and further assume that the travel time between the same adjacent stations between 2000 and 2020 is equal to the value in 2020 (not considering the effect of train speed increases).

To estimate the temporal distance of a specific path p_{ij} from stations i to j, we decompose the travel time into three parts: the in-vehicle time, transfer delay, and access/egress delay. We estimate each time separately: the in-vehicle time $T_{p_{ij}}^{veh}$ is obtained by the sum of the in-vehicle times on each edge; the transfer delay is assumed to be the product of the number of transfers $C_{p_{ij}}$ and the average transfer delay T_{trans}; and the access/egress delay is assumed to be a constant value T_{const}. The temporal distance of p_{ij} is

$$T_{p_{ij}}^{est} = T_{p_{ij}}^{veh} + C_{p_{ij}} \times T_{trans} + T_{const}, \tag{2}$$

where $T_{p_{ij}}^{veh}$ and $C_{p_{ij}}$ can be obtained from the subway network data. Although we assume that the average transfer delay T_{trans} and the access/egress delay T_{const} are constant at all stations, they are still difficult to observe. To solve this issue, we need to find some trips to be able to know their $T_{p_{ij}}^{est}$ and $C_{p_{ij}}$ and estimate T_{trans} and T_{const} as two regression coefficients. To do so, we assume that the trip with the least number of transfers among the k shortest paths is the most time-efficient trip, whose travel time is the shortest (arrival time − departure time) derived from the card data. This trip is also known as the Pareto-optimal trip (52). Here, we use $k = 3$, and we also test this on different k values(Supplementary Material Appendix, Fig. S10). $T_{p_{ij}}^{est}$ is estimated by the mean travel time

of the time-efficient trips between stations i and j. We then conduct the regression on the inputs of all station pairs to estimate the coefficients T_{trans} and T_{const} based on Eq. (2).

To validate the estimated travel time of the paths, we test Eq. (2) on the routes of the survey data. We calculate the expected time of each surveyed route in the network based on Eq. (2) and investigate the difference between this time and the actual travel time. Supplementary Material Appendix, Fig. S11 shows that in all three cities, the time difference peaks at approximately $\Delta T = 0$, and most trips (approximately 95.2%) are in the range of $|\Delta T| \leq 10$ min . The small gap verifies the effectiveness of the estimation [Eq. (2)].

Discrete choice model
To match the smart card records with routes on subway networks, we adopt the multinomial logit (MNL) model derived from utility theory, which is widely applied to traffic assignment (21, 23, 53).

In the MNL model, the probability of choosing the path p_{ij} from stations i to j in the choice set $P(i, j)$ is defined as

$$Prob_{p_{ij}} = \frac{e^{V_{p_{ij}}}}{\Sigma_{P(i,j)} e^{V_{p_{ij}}}}, \tag{3}$$

where $V_{p_{ij}}$ is the deterministic utility function of p_{ij} and is usually defined by a linear combination of factors affecting traveler route choices. We define the utility function by two explanatory variables as

$$V_{p_{ij}} = \beta_0 \times T_{p_{ij}}^{est} + \beta_1 \times PI_{p_{ij}}^{trans}, \tag{4}$$

where $T_{p_{ij}}^{est}$ is the travel time of p_{ij}. $PI_{p_{ij}}^{trans}$ is the cumulative transfer penalty index (23), specified as $(1 - e^{-C_{p_{ij}}})/d_{ij}$, where $C_{p_{ij}}$ is the number of transfers on p_{ij} and d_{ij} is the Euclidean distance between i and j. This penalty index assumes that the impedance of transfers to people's route choice increases in a nonlinear form as the number of transfers increases, and the effect of this cumulative impedance is inversely proportional to the distance traveled (23). β_0 and β_1 are the parameters measuring the effects of $T_{p_{ij}}^{est}$ and $PI_{p_{ij}}^{trans}$.

To calibrate the MNL models for different cities, we estimate the parameters by the maximum likelihood method using the survey data. According to the estimation, the parameters set for Beijing ($\beta_0 = -0.0063$, $\beta_1 = -30.99$), Shanghai ($\beta_0 = -0.0023$, $\beta_1 = -127.7$), and Shenzhen ($\beta_0 = -0.0031$, $\beta_1 = -113.2$) are statistically significant (Supplementary Material Appendix, Tables S5–S7).

To validate the discrete choice models, we first calculate the matching accuracy on the survey data. Based on the probability of being chosen given by the MNL model, we match the path with the highest probability in the choice set to that record and check whether this is the correct match. We finally obtain an overall correct rate of 92.1% across the three cities, with small differences between each city (Supplementary Material Appendix, Fig. S12). We then apply the models to the full set of smart card data and compare the aggregated flows on each line to the official published ridership numbers. The well-fitting regression also validates our models (Fig. 2B).

Search information
We measure the difficulty of navigating a subway network based on search information (9) and quantify the navigation complexity at the station/line level by aggregating the information needed to follow a path.

Station-level and line-level search information

The station-level search information is measured by averaging over all route choices $P(i, j) = \{p(i, s; j, t)\}$ from stations i to j. Since locating each path $p(i, s; j, t)$ in $P(i, j)$ may require different amounts of information, we adopt a flow-weighted method for the aggregation:

$$S_s(i, j) = \frac{\sum_{P(i,j)} S(p(i, s; j, t)) \cdot w(p(i, s; j, t))}{\sum_{P(i,j)} w(p(i, s; j, t))}. \quad (5)$$

For a given station pair i to j, ESI is calculated based on all the matched paths, and $w(p(i, s; j, t))$ is the weight value defined by the normalized flow on $p(i, s; j, t)$. For the TSI, the fastest simplest path is assumed to be the route choice between i and j according to ref. (11), and hence, only the single path is in $P(i, j)$, which simplifies Eq. (5) to $S_s(i, j) = S(p(i, s; j, t))$.

Similarly, we summarize the line-level search information by averaging over all route choices from line s to line t:

$$S_l(s, t) = \frac{\sum_{\{P(i,j)\}} S(p(i, s; j, t)) \cdot w(p(i, s; j, t))}{\sum_{\{P(i,j)\}} w(p(i, s; j, t))}. \quad (6)$$

For the ESI, $\{P(i, j)\}$ contains all the matched paths whose origin station i starts at line s and destination station j ends at line t. For the TSI, $\{P(i, j)\}$ contains all the fastest simplest paths between these stations.

Acknowledgments

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

Author Contributions

L.D., Y.L., and Z.J. designed the study; Z.J. processed and analyzed the data; Z.J., L.D, and Y.L. wrote the manuscript; and L.W. and Y.L. coordinated the study. All authors read, commented, and approved the final version of the manuscript.

Data Availability

The data and code used to reproduce the results are available at https://github.com/Jzjsnow/navi-complexity (54).

References

1. Gendreau M, Laporte G, Mesa JA. 1995. Locating rapid transit lines. J Adv Trans. 29(2):145–162.
2. Boguñá M, Krioukov D, Claffy KC. 2009. Navigability of complex networks. Nat Phys. 5(1):74–80.
3. De Domenico M, Solé-Ribalta A, Gómez S, Arenas A. 2014. Navigability of interconnected networks under random failures. Proc Natl Acad Sci. 111(23):8351–8356.
4. Dong L, Li R, Zhang J, Di Z. 2016. Population-weighted efficiency in transportation networks. Sci Rep. 6(1):26377.
5. Latora V, Marchiori M. 2001. Efficient behavior of small-world networks. Phys Rev Lett. 87(19):198701.
6. Çolak S, Lima A, González MC. 2016. Understanding congested travel in urban areas. Nat Commun. 7(1):10793.
7. Gallotti R, Barthelemy M. 2015. Anatomy and efficiency of urban multimodal mobility. Sci Rep. 4(1):6911.
8. Barberillo J, Saldaña J. 2011. Navigation in large subway networks: an informational approach. Phys A: Stat Mech Appl. 390(2):374–386.
9. Rosvall M, Trusina A, Minnhagen P, Sneppen K. 2005. Networks and cities: an information perspective. Phys Rev Lett. 94(2):028701.
10. Rosvall M, Grönlund A, Minnhagen P, Sneppen K. 2005. Searchability of networks. Phys Rev E. 72(4):046117.
11. Gallotti R, Porter MA, Barthelemy M. 2016. Lost in transportation: information measures and cognitive limits in multilayer navigation. Sci Adv. 2(2):e1500445.
12. Avena-Koenigsberger A, Misic B, Sporns O. 2018. Communication dynamics in complex brain networks. Nat Rev Neurosci. 19(1):17–33.
13. Amico E, Arenas A, Goñí J. 2019. Centralized and distributed cognitive task processing in the human connectome. Network Neurosci. 3(2):455–474.
14. Rajapandian M, Amico E, Abbas K, Ventresca M, Goñí J. 2020. Uncovering differential identifiability in network properties of human brain functional connectomes. Network Neurosci. 4(3):698–713.
15. Sneppen K, Trusina A, Rosvall M. 2005. Hide-and-seek on complex networks. Eur Lett. 69(5):853.
16. Boushaba M, Hafid A, Gendreau M. 2017. Node stability-based routing in wireless mesh networks. J Network Comp Appl. 93:1–12.
17. Zanin M, Buldú JM, Cano P, Boccaletti S. 2008. Disorder and decision cost in spatial networks. Chaos: An Int Jo Nonlin Sci. 18(2):023103.
18. Yin H, Clegg RG, Mondragón RJ. 2020. Simplification of networks by conserving path diversity and minimisation of the search information. Sci Rep. 10(1):19150.
19. Perotti JI, Billoni OV. 2012. Smart random walkers: the cost of knowing the path. Phys Rev E. 86(1):011120.
20. Cajueiro DO. 2009. Optimal navigation in complex networks. Phys Rev E. 79(4):046103.
21. Prashker JN, Bekhor S. 2004. Route choice models used in the stochastic user equilibrium problem: a review. Transp Rev. 24(4):437–463.
22. Prato CG. 2009. Route choice modeling: past, present and future research directions. J Choice Model. 2(1):65–100.
23. Kim I, Kim HC, Seo DJ, Kim JI. 2020. Calibration of a transit route choice model using revealed population data of smartcard in a multimodal transit network. Transportation. 47(5):2179–2202.
24. Liu Y, Bunker J, Ferreira L. 2010. Transit users' route-choice modelling in transit assignment: a review. Transp Rev. 30(6):753–769.
25. Raveau S, Guo Z, Muñoz JC, Wilson NHM. 2014. A behavioural comparison of route choice on metro networks: time, transfers, crowding, topology and socio-demographics. Transp Res A: Pol Pract. 66:185–195.

26. Muscoloni A, Cannistraci CV. 2019. Navigability evaluation of complex networks by greedy routing efficiency. Proc Natl Acad Sci. 116(5):1468–1469.

27. Lee SH, Holme P. 2012. Exploring maps with greedy navigators. Phys Rev Lett. 108(12):128701.

28. Gärling T, Axhausen KW. 2003. Introduction: habitual travel choice. Transportation. 30(1):1–11.

29. McKinlay R. 2016. Technology: use or lose our navigation skills. Nature. 531(7596):573–575.

30. Bongiorno C, *et al.* 2021. Vector-based pedestrian navigation in cities. Nat Comput Sci. 1(10):678–685.

31. Alessandretti L. 2021. A new computational model for human navigation. Nat Comput Sci. 1(10):642–643.

32. Coutrot A, *et al.* 2022. Entropy of city street networks linked to future spatial navigation ability. Nature. 604(7904): 104–110.

33. Beijing Subway. 2012. Official Weibo of Beijing Subway. https://weibo.com/bjsubway, last accessed: 27 June 2022.

34. Viana MP, Strano E, Bordin P, Barthelemy M. 2013. The simplicity of planar networks. Sci Rep 3(1):3495.

35. Van der Waard J. 1988. The relative importance of public transport trip time attributes in route choice. Presented at: PTRC Summer Annual Meeting, 16th, 1988, Bath, United Kingdom.

36. Yen JY. 1971. Finding the *k* shortest loopless paths in a network. Manage Sci. 17(11):712–716.

37. Bovy PHL. 2009. On modelling route choice sets in transportation networks: a synthesis. Transp Rev. 29(1):43–68.

38. Prato CG, Bekhor S. 2007. Modeling route choice behavior: how relevant is the composition of choice set? Transp Res Record: J Trans Res Board. 2003(1):64–73.

39. Cowan N. 2001. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behav Brain Sci. 24(1):87–114.

40. Cowan N. 2010. The magical mystery four: how is working memory capacity limited, and why? Curr Direct Psychol Sci. 19(1):51–57.

41. Halford GS, Cowan N, Andrews G. 2007. Separating cognitive capacity from knowledge: a new hypothesis. Trends Cogn Sci. 11(6):236–242.

42. Guimerà R, Díaz-Guilera A, Vega-Redondo F, Cabrales A, Arenas A. 2002. Optimal network topologies for local search with congestion. Phys Rev Lett. 89(24):248701.

43. Weckström C, Mladenović MN, Kujala R, Saramäki J. 2021. Navigability assessment of large-scale redesigns in nine public transport networks: open timetable data approach. Transp Res Part A Policy Pract. 147:212–229.

44. Manley EJ, Addison JD, Cheng T. 2015. Shortest path or anchor-based route choice: a large-scale empirical analysis of minicab routing in London. J Transp Geogr. 43:123–139.

45. Xu Y, González MC. 2017. Collective benefits in traffic during mega events via the use of information technologies. J R Soc Inter. 14(129):20161041.

46. Ramming MS. 2002. Network knowledge and route choice [PhD Thesis]. Massachusetts Institute of Technology. https://dspace.mit.edu/handle/1721.1/49797, last accessed: 1 August 2022.

47. Epstein RA, Patai EZ, Julian JB, Spiers HJ. 2017. The cognitive map in humans: spatial navigation and beyond. Nat Neurosci. 20(11):1504–1513.

48. Chrastil ER, Warren WH. 2014. From cognitive maps to cognitive graphs. PLoS One. 9(11):e112544.

49. Stangl M, Kanitscheider I, Riemer M, Fiete I, Wolbers T. 2020. Sources of path integration error in young and aging humans. Nat Commun. 11(1):2626.

50. Anggraini D, Glasauer S, Wunderlich K. 2018. Neural signatures of reinforcement learning correlate with strategy adoption during spatial navigation. Sci Rep. 8(1):10110.

51. Liao W, Kemloh Wagoum AU, Bode NWF. 2017. Route choice in pedestrians: determinants for initial choices and revising decisions. J R Soc Inter. 14(127):20160684.

52. Kujala R, Weckström C, Mladenović MN, Saramäki J. 2018. Travel times and transfers in public transport: comprehensive accessibility analysis based on Pareto-optimal journeys. Comput Environ Urban Syst. 67:41–54.

53. Jánošíková Ľ, Slavík J, Koháni M. 2014. Estimation of a route choice model for urban public transport using smart card data. Transp Plan Tech. 37(7):638–648.

54. Jiang Z, Dong L, Wu L, Liu Y. 2022. Supplementary data for "Quantifying navigation complexity in transportation networks." Zenodo. doi:10.5281/ZENODO.6828403.