Understanding metropolitan patterns of daily encounters

Supporting Information Appendix

Lijun Sun, Kay W. Axhausen, Der-Horng Lee and Xianfeng Huang

Contents

I Data

Trip records were collected from Singapore's smart-card-based fare collection system, covering more than 96% of public transit trips. The system collects data for both bus and MRT (subway) modes. We employ bus – not MRT (Mass Rapid Transit, railway based) – trip records in this study, since it is difficult to identify close proximity interactions on the large MRT trains. For buses, once a smart card holder boards a vehicle (tapping-in), the system will generate a temporary transaction record; after he/she leaves the vehicle (tapping-out), a complete record will be stored with detailed information on the trip. A comprehensive summary on utilization of smart card data is presented in Ref. [1].

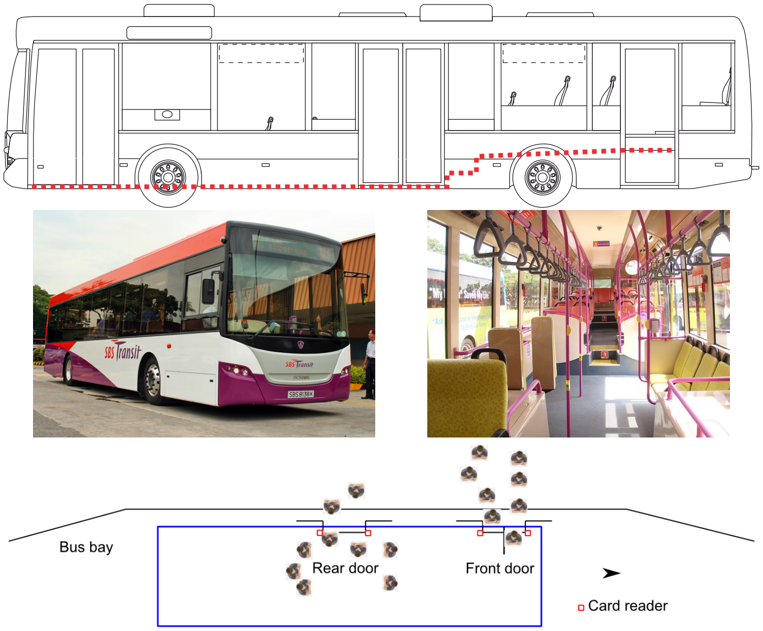

Figure S1: Layout of the Scania K230 single decker lower floor bus and location of smart the card readers

Fig. S1 shows design and layout of most common bus type in Singapore – Scania K230UB Euro V bus. Passengers can only board at front doors and are advised to alight at the rear door. Two smart card readers are employed on both sides of each door.

In general, a full bus trip may contain more than one stage with transfers from one route/vehicle to another; stage records are generated separately in the smart card system. Since our goal is to identify in-vehicle encounters and the people one may encounter in vehicles will differ from stage to stage, we use the term trip to represent stage in this document (see Ref. [2]). After processing the raw data, we obtained the trip records used in this study. Fields and their contents are provided in Tab. S1.

This study was performed on trip records for the week from 11, April, 2011 (Monday) to 17, April, 2011 (Sunday). The dataset contains 21,814,448 bus trip transaction records from 2,895,750 individual smart card holders.

Table S1: Fields and contents of trip record dataset

Field	Description
Trip ID	A unique number for each transit trip
Card ID	A unique coded number for each smart card (anonymized)
Passenger Type	The attribute of cardholder (Adult, Senior citizen and Child)
Service Number	Bus route service number (e.g. 96)
Direction	Direction of the bus route (0 and 1)
Bus Registration No.	A unique registration number for each vehicle (e.g. '0999')
Boarding Stop ID	A unique number for boarding stop (e.g. 40009)
Alighting Stop ID	A unique number for alighting stop (e.g. 40009)
Ride Date	Date of a trip (e.g. '2011-04-11')
Ride Start Time	Start (tapping-in) time of a trip (e.g. 08:00:00)
Ride End Time	End (tapping-out) time of a trip (e.g. 08:00:00)
Ride Distance	Distance of the trip (e.g. 12.0 km)

II Social demographic dependency

The use of transit services in general – and buses in particular – is differentiated along lines of ethnicity, gender, age and income. As a result, similar to residential segregation - metropolitan areas are organized with respect to ethnicity and income – daily transit use might exhibit social segregation patterns as well.

To address dependency and segregation of bus use on social demographic attributes, we incorporated two additional datasets in the analyses:

(1) **Census of population and Population Profile:** This dataset provides a statistical analysis of Singapore's changing population profile. The report presents trends and changes in geographic distribution and transport characteristics of Singapore's resident population [3].

(2) **HITS, 2008:** The Household Interview Travel Survey (HITS) is conducted every four to five years by the Land Transport Authority (LTA) of Singapore. HITS (2008) is the latest released survey.

II.1 Singapore population

These analyses are based on the population trends dataset. The total population of Singapore was 5.31 million in June, 2012.

Tab. S2 shows the share of transport modes for working residents older than 15. We see that public transit covers more than 40% of the overall working transport modes. Overall, about 50.5% of Singaporean residents (excluding foreigners) go to work using only public transit (excluding private bus/van, e.g. company bus).

Table S2: Key indicators of the resident population on transport by ethnicity (2010) (adapted from Ref. [3])

Transport mode (%)	Total	Chinese	Malays	Indians	Others
Resident working persons by usual transport mode to work (aged above 15)	100.0	100.0	100.0	100.0	100.0
Public bus only	19.3	18.3	24.1	22.3	18.2
Metro/Metro with bus	31.2	30.6	30.2	34.4	39.9
Car only	24.8	27.8	12.7	17.6	17.6
Private chartered bus/van only	3.6	3.5	4.2	4.4	3.2
Motocycle/scooter only	3.8	2.4	13.1	4.6	2.0
Others or no transport required	17.3	17.4	15.7	16.7	19.1

II.2 Social profile of bus users

With the increasing quantity and range of urban mobility, public transit is becoming more and more important as a mode of individual mobility. However, similar to other social activities, transit use pattern is also highly determined by individual social profile, such as gender, age and income. To have a comprehensive understanding of bus users' social profiling, we explored bus use dependency based on the following social attributes: gender, age and income.

II.2.1 Dependence on gender and age

We employed HITS (2008) dataset to explore social segregation of bus use on gender, age, and income. In Singapore, nation wide HITS is conducted every four to five years by the Land Transport Authority (LTA), to provide insight into residents' traveling and commuting behavior for transportation planners and policy makers. The 2008 survey covers about one percent of all Singapore households, with each household member answering detailed questions about their trips and social attributes.

The survey also aims at understanding challenges and policy issues facing Singapore's land transport use. Such surveys benefit research as well, by providing constructive insights on residents' commuting behaviors and their evolving traveling patterns: particularly transit use patterns distributed over individual social attributes.

Tab. S3 and the blue bars in Fig. S2a show gender and age distribution over all 35122 individuals from 10641 households in HITS (2008). Looking at all individuals, sex composition and age structure in HITS show similar patterns to results over the whole population. Based on this nation wide survey, we then explore dependence of transit use on gender and age. The last two columns of Tab. S3 and the red bars in Fig. S2a show the proportion of bus users over each sample group.

From the symmetrical pattern of Singapore bus users' proportion over gender (see Fig. S2), apparently bus use does not differ significantly by gender; both males and females show similar proportion of bus user

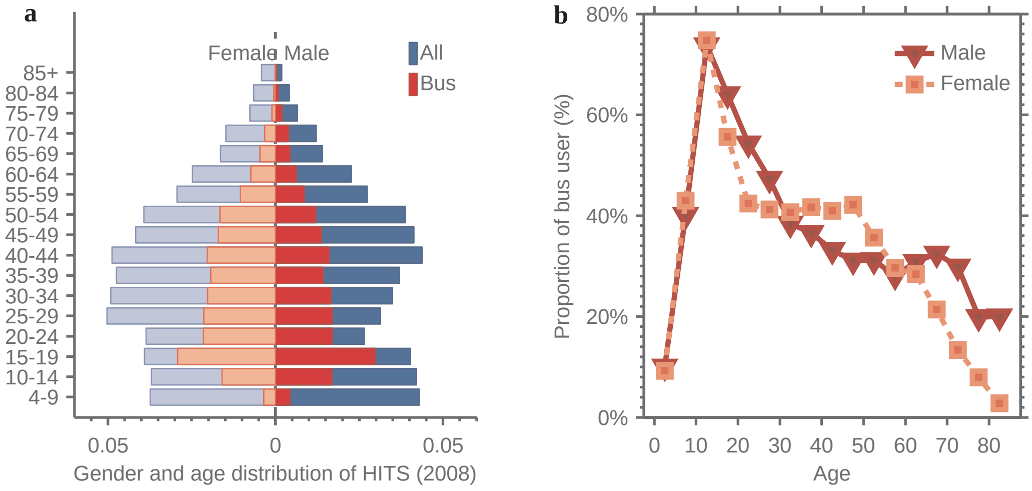

Figure S2: Gender and age distribution from HITS (2008). **(a)**, Gender and age distribution in HITS (2008). The blue bars indicate all individuals in the survey, while the red bars show only the individuals who used public transit services on the surveyed day. **(b)**, Detailed proportion of public transit users over age, for both genders.

Table S3: Gender and age distribution from HITS (2008) (all v.s. bus users)

Age	HITS all		HITS bus		Age	HITS all		HITS bus	
	Male	Female	Male	Female		Male	Female	Male	Female
4-9	1509	1312	152	121	50-54	1362	1378	423	581
10-14	1477	1300	592	559	55-59	963	1033	300	368
15-19	1415	1371	1044	1025	60-64	795	871	223	258
20-24	933	1356	598	754	65-69	493	574	152	163
25-29	1101	1766	598	749	70-74	428	520	139	111
30-34	1228	1727	581	712	75-79	231	269	69	36
35-39	1300	1666	499	677	80-84	146	227	29	18
40-44	1536	1713	562	714	85+	65	144	13	4
45-49	1450	1463	481	599	Total	11949	13674	5107	5910

across all age groups.

However, from the proportion shown in Fig. S2b, we can see that bus use shows clear differentiated pattern by age. Children under 9 and senior citizens above 70 seldom take buses. There are more bus users among the young (15∼25) than among the adults over 30. With the age increase from 30 to 70, we also observed a decreasing proportion of bus use.

II.2.2 Dependence on income and housing

To characterize income dependency, we first show income distribution over all individuals in HITS (2008). As shown in Fig. S3, over all individuals who agreed to report their monthly incomes in HITS (2008), more

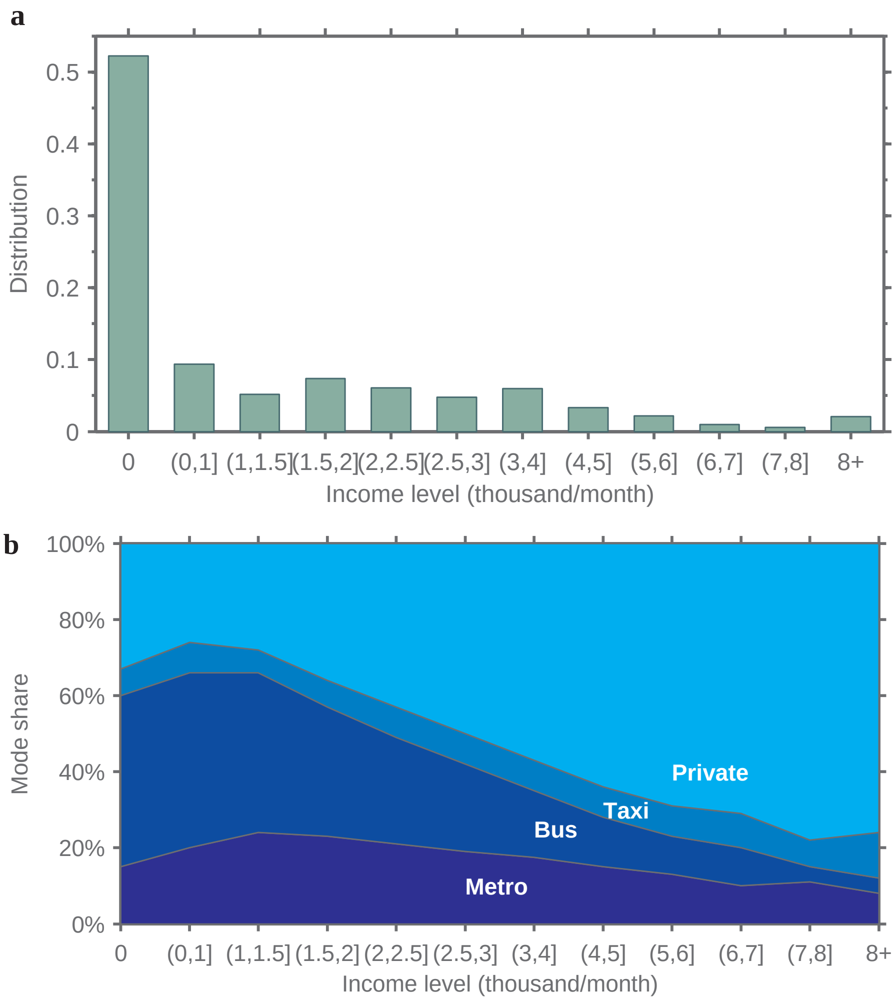

Figure S3: Dependence of bus use on income levels. **(a)**, Distribution of income levels from HITS (2008) (Unit:SGD). **(b)**, Transport model share across different income levels from HITS (2008) (adapted from Ref. [4]) (Unit:SGD).

than 50% have no income (mainly children or retired).

To better understand how income affects mode choice, Fig. S3b provides the mode share breakdown by private transport, taxi, bus and metro against income levels. We see that the share of metro and taxi remained stable over various income levels. However, the mode share for buses was highest for the low-income group (Singapore Dollar – SGD 0∼2000). With increase in income, mode share for buses begins to taper off. Taking both income distribution and mode share into consideration, bus trips actually covered about 37% of total urban mobility in the 2008 survey. Census of population 2010 also indicates that transport mode share varies with housing type (Tab. S4). About 62∼69% of resident working persons staying in HDB 1-3 room flats commuted by public transit, while the proportion is 48∼56% for people living in HDB 4 room or larger flats. Further analyses of transport mode by education (attending school) and travel time by different transport modes can be found in Ref. [3].

Taken together, observations suggest that transit use in Singapore varies clearly with age and income, while trends over ethnicity and gender are not clear. However, considering that the number of active indi-

Table S4: Key indicators of the resident population on transport by housing (2010) (adapted from Ref. [3])

Transport mode (%)	HDB 1/2 room flats	HDB 3 room flats	HDB 4 room flats	HDB 5 and executive	Condo and private flats	Landed properties
Resident working persons by usual transport mode to work (aged above 15)	100.0	100.0	100.0	100.0	100.0	100
Public bus only	39.9	28.7	21.5	15.3	10.3	8.6
Metro only	10.8	12.2	12.5	12.4	8.9	5.5
Metro with bus	18.2	19.9	19.8	18	10.9	9.4
Metro with other	0.6	1.3	1.9	2.7	2.9	2.2
Car only	2.2	9.3	15.6	29.3	50.6	59.6
Private chartered bus/van only	2.3	4.1	4.7	3.6	1.5	1.2
Motocycle/scooter only	4.4	5.1	5.2	3.3	0.8	0.6
Others	8.8	9.5	10.9	9.7	8.6	7.8
No transport required	12.9	9.9	7.8	5.9	5.6	5.1

viduals over one week is 2.9 million (covering 55% of the residents) and taking the observed patterns in Fig. 1C and 1D in the main paper into account, we conclude that public transit is the most important element of daily commuting trips (go to work/school) for Singaporeans.

III Characterizing collective transit use

To explore the pattern of collective transit activities and the pattern of transit usage at city-scale, we first characterized the following properties for all 21,814,448 trips:

(1) Trip start time

(2) Trip duration l (in-vehicle travel time)

III.1 Trip start time

Fig. S4a shows the number of bus boardings per hour during the week. Prominent morning and evening peaks can be readily identified with departure rates greater than 300,000 passengers/hour, i.e. the daily collective commuting flows in the city. Between the two peaks on weekdays, the buses are also highly utilized with boardings higher than 100,000 passengers/hour. In addition, we find quite similar daily demand patterns from Monday to Friday, revealing the regularity of collective transit use for both regular travelers and other passengers, in both peak and non-peak time. Without the heavy commuting flows of weekdays,

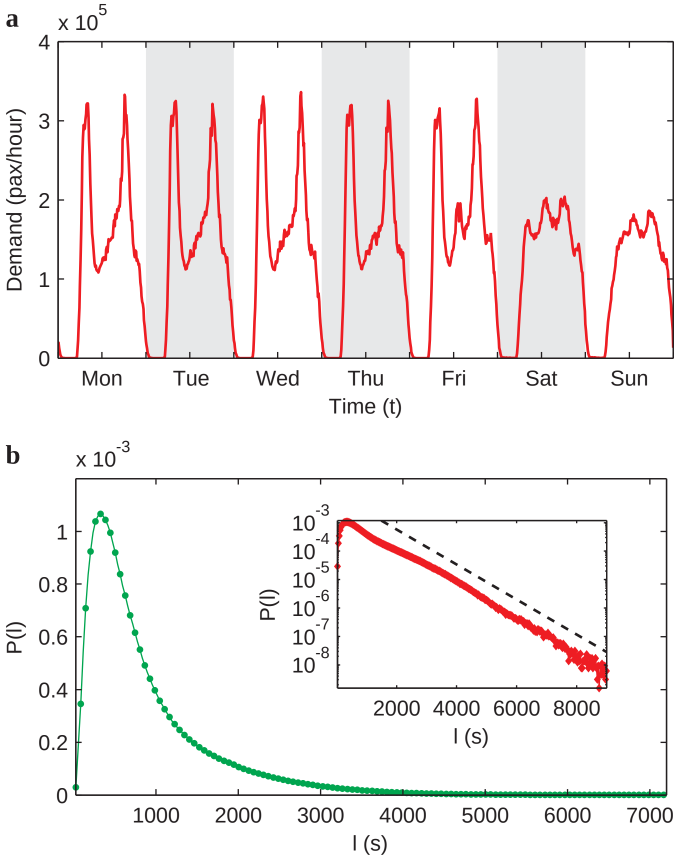

Figure S4: Collective transit activities. **(a)**, Departure rate of bus trips during the week. The demand is shown in number of boardings per hour. **(b)**, The probability density function $P(l)$ of trip durations. The inset plots $P(l)$ in semi-log scale (◆).

demand patterns are more uniform on weekends. Collective movements and travel behaviors across the world are reviewed in Ref. [5] for comparison.

III.2 Trip duration l (travel time)

Fig. S4b shows the distribution $P(l)$ of trip duration during the week, defined as the time interval between boarding (tapping-in) and alighting (tapping-out) activities. Considering that city transit route lengths are always upper bounded, distribution of travel time is also bounded with an upper limit of 9041s (2.5 hours). In addition, as the inset of Fig. S4b shows, we found that the tail of $P(l)$ can be well characterized by an exponential, when $t \geq 600s$:

$$P(l) \sim e^{-\frac{l}{\lambda_l}} \tag{S1}$$

with $\lambda_l \approx 704s$.

As a bus system is a feeder mode in a city with a subway system, durations of bus trips are relatively short. On average, durations of bus trips are $867.9 \pm 763.9s$ ($14.5 \pm 12.7min$, mean ± standard deviation).

IV Characterizing individual transit use

To explore individual transit use patterns, we analyzed the following attributes for all 2,895,750 individual smart card holders.

(1) Trip frequency f

(2) Number of encountered people n

(3) Time interval between consecutive bus trips τ

IV.1 Trip frequency f

The frequency of using bus services is the main indicator to characterize individual transit usage patterns. We measured frequency f of taking buses during the week for all individuals. The distribution $P(f)$ is shown in Fig. S5a. Given the fact that transit use is constrained by various personal and physical factors, $P(f)$ is also upper bounded (upper limit: 116 $week^{-1}$, the most active user) and exponentially distributed, as the inset of Fig. S5a shows:

$$P(f) \sim e^{-\frac{f}{\lambda_f}} \tag{S2}$$

with $\lambda_f \approx 5.9\ week^{-1}$.

On average, stage frequency during the week is $7.5 \pm 6.6\ week^{-1}$ (mean ± standard deviation).

IV.2 Number of encountered people n

Fig. S5b shows the distribution $P(n)$ of number of encountered individuals during the week. Travelers are grouped according to their bus trip frequencies f. Clearly, the more one travels, the more contacts he/she tends to have. The standard deviation of contacts also increases with f.

IV.3 Inter-event interval τ

People's communication activities via digital networks such as mobile phone calls, e-mails and online messages follow a non-Poissonian inter-event interval τ between consecutive communication activities,

$$P(\tau) \sim \tau^{-\beta} e^{-\frac{\tau}{\tau_c}} \tag{S3}$$

with exponent β and cutoff value τ_c.

This distribution $P(\tau)$ characterizes the bursty nature of human communication activities, showing that inter-event intervals τ are smoothly distributed as a truncated power-law [6, 7]. However, intuitively, transit activities are typically more regular and rhythmic than communications, in particular for commuters with daily work/home trips.

To explore transit use dynamics, we selected 2,507,783 individuals who take buses more than once during the week (86.6% of all bus users), obtaining 18,985,800 inter-event intervals of consecutive bus trips. The distribution $P(\tau)$ of intervals is shown in Fig. S5c. As can be seen, unlike the communication activities

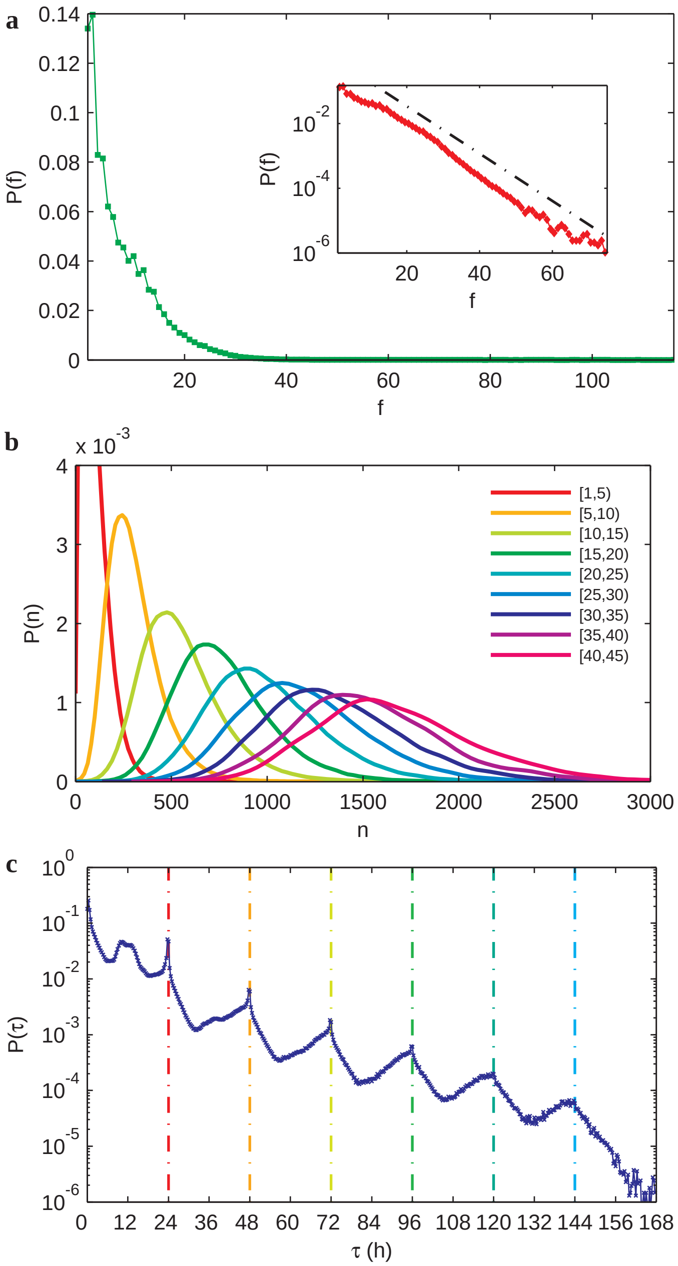

Figure S5: Individual transit use. (**a**), Distribution $P(f)$ of stage frequency by users during the week. The inset displays probability density in semi-log scale (♦). (**b**), Probability density function $P(n)$ of number of encountered people during the week, grouped by stage frequency f. (**c**), Distribution $P(\tau)$ of time intervals between consecutive bus trips, across overall transit user population during the week.

mediated by digital networks, $P(\tau)$ on bus activities has clear temporal patterns, neither exponentially nor power-law distributed. In most cases, individuals take their next trips within 18 hours after previous trips. Over this 18 hours, two peaks can be clearly identified:

- 0–2h, a prominent peak describing intervals between unlinked trips: including transfers to create a linked trip between two activities, as well as activity chains with short duration activities at the first stop;
- 8.5–14.5h, two close flat peaks, indicating intervals between daily home/work commuting trips;

When the interval is longer than 18 hours, the distribution shows repeated prominent peaks at $24h$, $48h$, $72h$, $96h$, $120h$ and $144h$. In contrast with the smooth asymptotic distribution of inter-event intervals obtained from digital communication networks, these prominent peaks capture the tendency of individuals to take buses at daily intervals. Note that a similar tendency with peaks of people returning to locations they previously visited is reported in Ref. [8] as well, showing that people's transit use displays significant regularity.

V Instantaneous in-vehicle encounter network

With the precise boarding/alighting times and the unique bus registration number, trip records provide us with the opportunity to construct a city-scale, time-resolved, in-vehicle encounter network. Given that the size of a bus is limited, encounters can be also considered as close physical proximity, important for understanding the spreading of airborne viruses and word-of-mouth information. Once an individual boards the vehicle, edges with all other on-board passengers are created, representing encounters; and when he/she leaves the vehicle, the edges with on-board passengers are destroyed. Thus, as a consequence of the time-continuous boarding/alighting activities, a temporal contact network is created. Unlike static networks such as road networks and internet networks, in temporal networks driven by activities, the edges are not continuously active over time, resulting in network evolution in temporal scale. Hence, the temporal resolution of edge-state (active or not) also determines the resolution of network evolution. Fig. S6a plots an instantaneous encounter network composed by individuals from 15 vehicles, at 9:00 a.m. on 11, April, 2011.

In contrast with temporal communication networks composed of a set of stars and lines instantaneously, the encounter network is more like a collaboration network on scientific papers or movies, since all individuals from one collaboration are fully connected [9]. However, as authors and actors can participate in more than one paper and movie at the same time, while one cannot appear in two vehicles instantaneously, this contact network is composed of many distinct components, each of which represents a fully connected contact network on one vehicle, as shown in Fig. S6a.

As opposed to collaboration networks (where the edges of a component are generally created and destroyed when the collaboration is begun or ended, as a consequence of individuals' boarding and alighting activities), the encounter networks change over time with the creation and destruction of edges. To model this temporal contact network, we defined two attributes to the edges: (1) creation time t_{cr}, and (2) contact duration t_d. Therefore, the edges of the contact network are denoted as $e_{t_{cr}}^{t_d}$, and the destruction time of

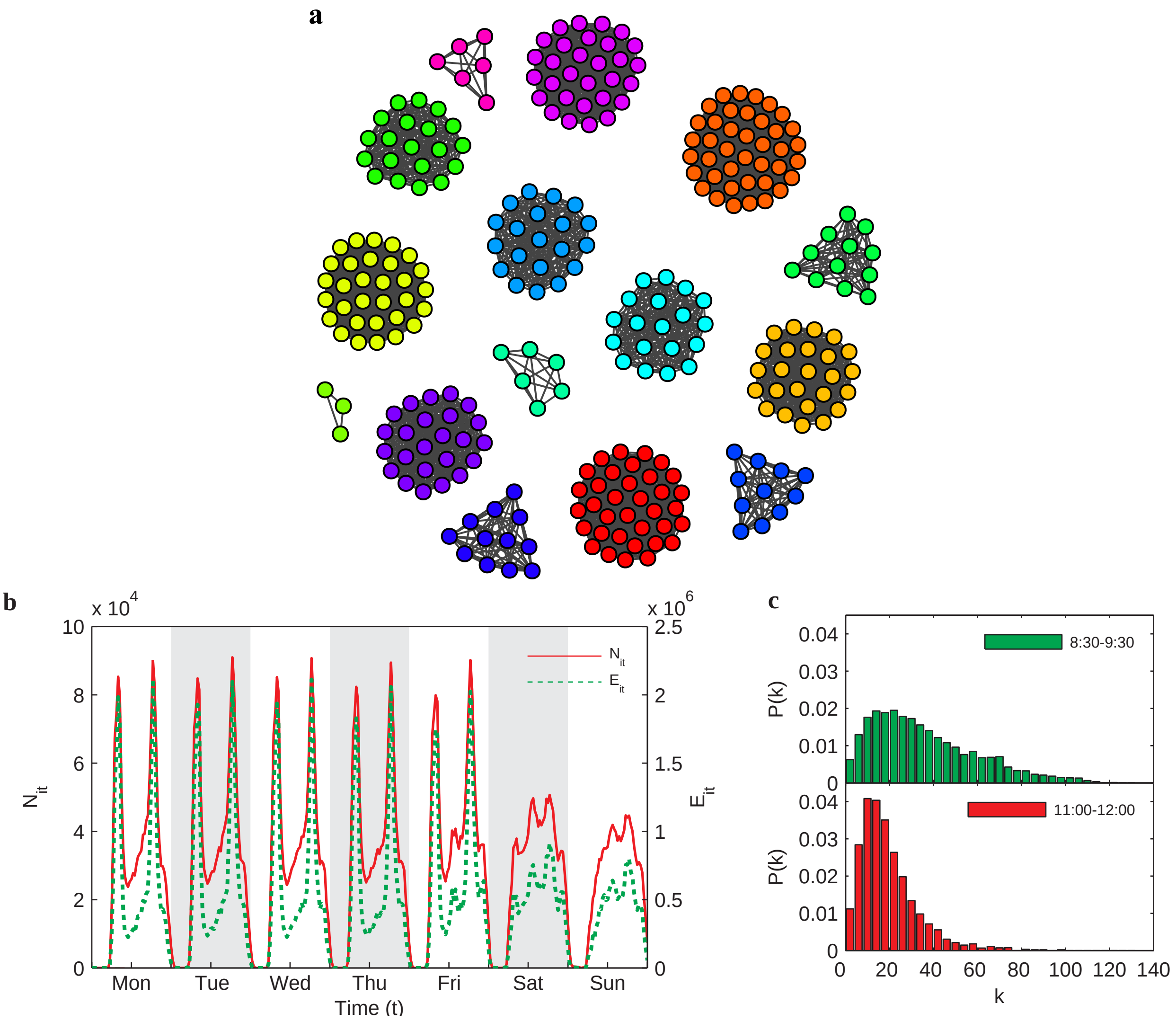

Figure S6: Instantaneous encounter network. (**a**), Contact network at 9:00 a.m. on 11, April, 2011. Here, we take only a snapshot of the encounter network, with 15 vehicles out of 2519, shown in different colors (only one half of on-board passengers from each vehicle are shown). (**b**), Number of vertices and number of edges of instantaneous network $G(t)$ during the week. Red solid curve shows number of vertices; green dashed line shows number of edges. (Data is in half hour intervals.) (**c**), Degree distribution $P(k)$ of instantaneous networks during peak (8:30–9:30) and non-peak hours (11:00–12:00). For each time slot, we selected 5 instantaneous networks (up: t=8:30, 8:45, 9:00, 9:15, 9:30) (down: t=11:00, 11:15, 11:30, 11:45, 12:00).

$e_{t_{cr}}^{t_d}$ is $t_{cr}+t_d$. Consequently, at any time t, the instantaneous contact network $G(t)$ can be defined as the composition of edges which are created before t and destroyed at or after t:

$$G(t) = \bigcup_{t_{cr}<t,\, t_{cr}+t_d \geq t} e_{t_{cr}}^{t_d} \tag{S4}$$

Fig. S6b shows the instantaneous number of vertices and number of edges respectively. In fact, for the instantaneous contact network $G(t)$, if individuals are assumed to be uniformly distributed to m vehicles, the number of edges may be written as:

$$E_{it} = \frac{m}{2} \times \left(\frac{N_{it}}{m} \langle k \rangle_{it} \right) = \frac{m}{2} \left(\frac{N_{it}}{m} \left(\frac{N_{it}}{m} - 1 \right) \right) \approx \frac{N_{it}^2}{2m} \tag{S5}$$

where N_{it} is number of vertices, and $\langle k \rangle_{it} = \left(\frac{N_{it}}{m} - 1 \right)$ is average degree of $G(t)$.

Therefore, with transit demand variation during the day, degree distribution $P(k)$ of instantaneous network $G(t)$ also changes, as shown in Fig. S6c. Obviously, degree distributions are markedly different in peak and non-peak hours.

VI Weekly-aggregated encounter network

To further explore encounter patterns, we extend the time dimension to an interval (t_1, t_2). Similar to instantaneous network $G(t)$, the aggregated encounter network $G(t_1, t_2)$ in interval (t_1, t_2) can be formulated by selecting edges created before t_1 and destroyed at or after t_2:

$$G(t_1, t_2) = \bigcup_{t_{cr}<t_1,\, t_{cr}+t_d \geq t_2} e_{t_{cr}}^{t_d} \tag{S6}$$

We set Monday's 0:00 a.m. as $t_1 = 0$ and the end of Sunday as $t_2 = 7 \times 24 = 168h$; thus, the weekly aggregated contact network can be denoted as $G(0, 168h)$. Considering that two individuals can encounter each other more than once during the week, it is possible that more than one time-labeled edge is created between two vertices in this network.

To explore statistical properties of the weekly-aggregated encounter network, we measure the following representative properties:

(1) Degree distribution $P(k)$

(2) Contact duration t_d

(3) Encounter frequency over the week f_e

(4) Total encounter duration d

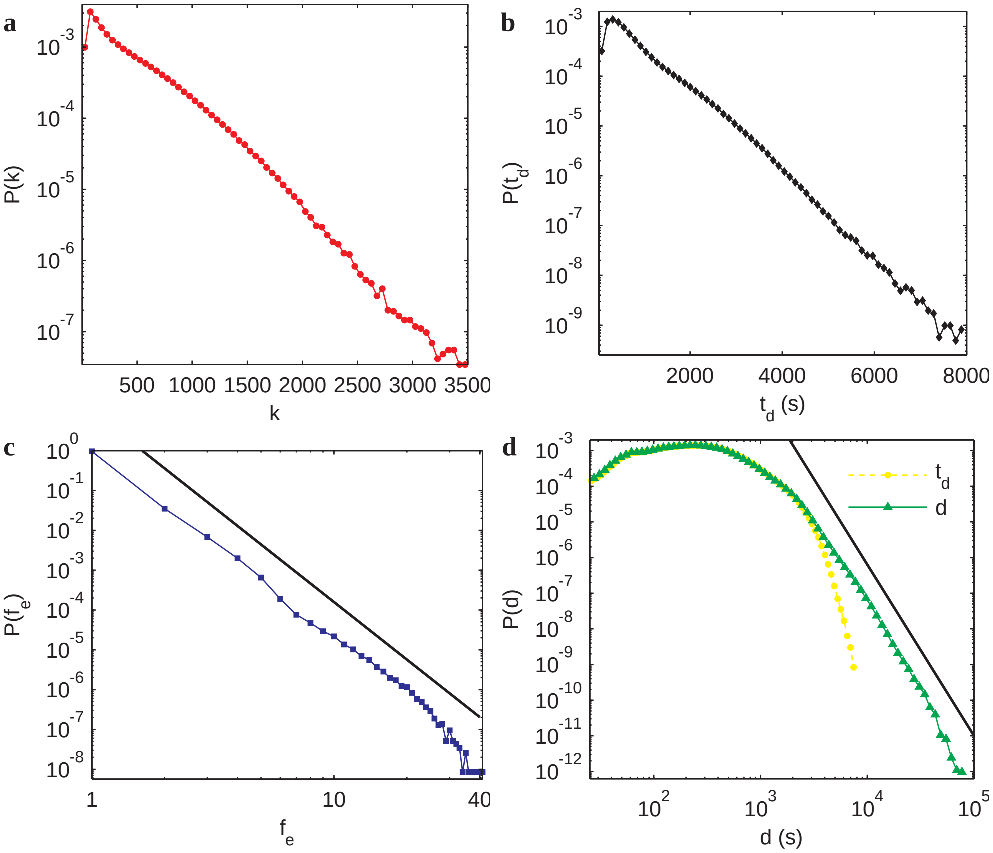

Figure S7: Statistical properties of weekly-aggregated in-vehicle encounter network. **(a)**, Distribution $P(k)$ of degrees. **(b)**, Probability density function $P(t_d)$ of encounter durations. This distribution is also plotted in log-log scale, as the yellow dashed line (●) shown in **(d)**. **(c)**, The distribution $P(f_e)$ of encounter frequencies by individual pairs over the week. As a guide, the black line shows the slope with $\beta = 4.8$ for comparison. **(d)**, Probability density function $P(d)$ (▲) of total encounter duration of pairs over the week. As a guide, the solid line indicates slope with $\beta = 4.8$.

VI.1 Degree distribution $P(k)$

In the aggregated network, the degree k for an individual stands for the number of people he/she has encountered during the week. Previous studies have shown the scale-free nature of static networks such as actor collaboration networks, internet and power grids, suggesting that the degree of these networks follows a heavy-tailed power-law distribution:

$$P(k) \sim k^{-\gamma} \tag{S7}$$

with exponent γ.

However, as shown in Fig. S7a, we find that $P(k)$ of the weekly-aggregated network shows a stable long exponential tail for $k \geq 50$: $P(k) \sim e^{-\frac{k}{\lambda_k}}$, with $\lambda_k \approx 286$. Unlike the scale-free property of some networks [10], the exponentially bounded tail suggests that most individuals have only a limited number of contacts during the week. Note that the exponential decaying of $P(k)$ is also reported in other human contact networks in Ref. [11, 12], suggesting similar patterns may exist in most, if not all, human physical contact networks. In fact, on one hand, the physical constraints on humans might be one reason for the short-tailed distribution given that one cannot take buses infinitely within one week. On the other hand, no preferential attachment is exhibited since all individuals are essentially random strangers to each other across the population. The average degree of the weekly-aggregated network is $\langle k \rangle = 342 \pm 333$.

VI.2 Contact duration t_d

In the weekly-aggregated network, 494,323,272 edges are created in total over the week. To explore the property of physical proximity, we extract the contact duration t_d of each edge. In Fig. S7b, we plot the distribution $P(t_d)$ of contact durations, which is one main indicator to measure connection strength between two individuals.

We observe that $P(t_d)$ can be also characterized by an exponential distribution $P(t_d) \sim e^{-\frac{t_d}{\lambda_{t_d}}}$ when $t_d \geq 600s$, with $\lambda_{t_d} \approx 498.4s$. The exponential tail is strong evidence that human transit use is highly constrained. It should be noted as well that t_d represents the length of overlap in two individuals' travel time and the shape of $P(t_d)$ is quite similar to the distribution $P(l)$ of travel time stated in Sec. III.2, together indicating the homogeneity between collective behaviors and individual behaviors.

VI.3 Encounter frequency f_e

For each individual pair, frequency of their encounters is a first indicator to quantify the strength of their social relationship. If individuals are assumed to have no correlation in their transit use with others, it might be rare that one encounters the same person repeatedly; however, frequent encounters are observed in the dataset. To measure the probability of repeated encounters, we calculated the frequency of encounters $f_e(i,j)$ for each individual pair (i,j). Fig. S7c shows the probability density distribution $P(f_e)$ for all pairs during the week. Strikingly, we found that the frequency of encounters can be characterized by a power-law distribution:

$$P(f_e) \sim f_e^{-\beta} \tag{S8}$$

with exponent $\beta \approx -4.8$, suggesting that while most people rarely encounter the same person repeatedly, there are still certain individuals who do encounter each other frequently during the week.

Hence, recurring encounters can represent substantial collective regularity of pairs of individuals regarding their transit use. This kind of social relation is the so-called "familiar stranger" [13, 14], the people whom we may encounter repeatedly in daily life. Nevertheless, since more than 95% of the 494,323,272 encounters happened only once during the week, people hardly noticed the existence of the huge contact network with their "familiar strangers".

VI.4 Total encounter duration d

Beyond encounter frequency, we propose a second indicator to measure strengths of connections between two individuals – total encounter duration d. For paired individuals who have encountered each other over the studied week, their total contact duration is defined as the sum of durations of all encounters:

$$d(i,j) = \sum_{k=1}^{f_e(i,j)} t_{d,k}(i,j) \tag{S9}$$

where $f_e(i,j)$ is the number of encounters between two individuals (i,j) during the week and $t_{d,k}(i,j)$ is the duration of their kth encounter.

Fig. S7d shows the distribution $P(d)$ of total encounter durations over all pairs. As can be seen, the probability density $P(d)$ generally overlaps with the distribution $P(t_d)$ shown in Sec. VI.2 when $d \leq 1200s$, as these individuals encounter each other only once in the studied period. However, the two distributions are separated into the long-tailed $P(d)$ and the short-tailed $P(t_d)$ part. The difference between $P(d)$ and $P(t_d)$ is due to the repeated encounters. When $d \geq 1200s$, the $P(d)$ can be well characterized by a power-law:

$$P(d) \sim d^{-\beta} \tag{S10}$$

with exponent $\beta \approx 4.8$.

Therefore, with the contribution from repeated encounters, the total encounter duration has overtaken the exponentially bounded t_d, resulting in the occurrence of heavy ties between two individuals in the aggregated network.

VII Measuring individuals

To explore how individual behavior influences his/her encounter patterns, we propose two attributes to capture the possibility of encounters and the regularity of transit use, respectively:

(1) Personal weight w_i

(2) Absolute trip difference m_i

VII.1 Personal weight w_i and rescaled encounter likelihood r_i

To quantify encounter likelihood individually, we first measure the total number of encounters for each individual, given as the sum of encounter frequency f_e with each encountered person:

$$u_i \equiv \sum_{j \in N(i)} f_e(i,j) \tag{S11}$$

where $N(i)$ is the set of encountered people (neighbors) of individual i.

However, since all the encounters are treated equally, u_i fails to capture the likelihood to encounter someone repeatedly. To measure the individual's chance of repeated encounters, we propose the personal weight w_i, given by the number of recurring encounters.

$$w_i \equiv \sum_{j \in N(i)} \left(f_e(i,j) - 1\right) \tag{S12}$$

hence, to individual i, the number of "familiar stranger" is $k_i = \sum_{j \in N(i), f_e(i,j)>1} 1$ (see Fig. S8 for u_i, w_i and k_i).

Given that the possibility of recurring encounters is also influenced by travel time, we rescale w_i to encountering rate to avoid the bias:

$$r_i = \frac{w_i}{T_i} \tag{S13}$$

where $T_i = \sum_j l_j$ is the total travel time (l_j is the duration of jth trip). Hence, r_i is a more appropriate attribute to capture a person's encounter likelihood.

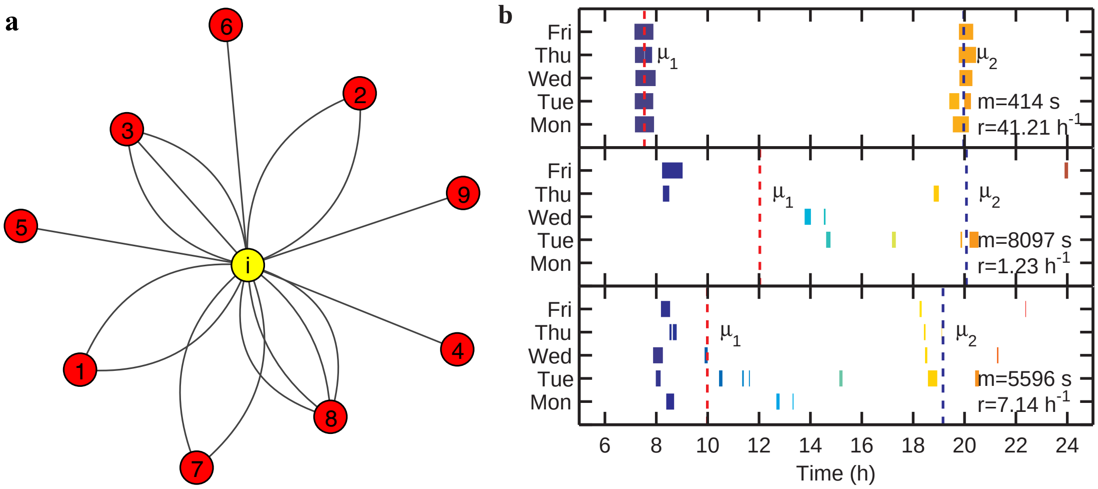

Figure S8: Individual transit use. **(a)**, Definitions of w_i and u_i. For individual i, his/her neighbour set is denoted as $N(i) = \{1,2,3,4,5,6,7,8,9\}$, of which individual i encounters $\{1,2,3,7,8\}$ more than once, such that $k_i = 5$; thus, the calculated personal weight and number of total encounters are $w_i = (2-1)+(2-1)+(3-1)+(1-1)+(1-1)+(1-1)+(2-1)+(4-1)+(1-1) = 8$ and $u_i = 2+2+3+1+1+1+2+4+1 = 17$ respectively. **(b)**, Transit usage plots for 3 individuals. Each bus trip is shown as a colored solid line, from t_{start} to t_{end}.

VII.2 Absolute trip difference m_i

The trip set of individual i is TR_i. To find out whether trip tr_c is regular or irregular, we first measure its similarity to the remaining trips, given by:

$$q\left(tr_c\right) \equiv \sum_{j \neq c, \left|t_j - t_c\right| \leq 45\,\mathrm{min}} l_j \tag{S14}$$

where $t_j = \frac{t_{start}+t_{end}}{2}$ is the mean of start and end times of the $jth\,(k = 1, 2, \cdots, n)$ trip $- tr_j$, t_c is mean time of current trip tr_c and l_j is the duration of the jth trip. Hence, $q(tr_c)$ measure the total duration of trips within $45min$ from tr_c.

Therefore, for trip $tr_c \in TR_i$, its outlyingness is defined as:

$$o_c \equiv \frac{q\left(tr_c\right)}{\sum\limits_j q\left(tr_j\right)} \tag{S15}$$

such that a higher o_c indicates tr_c is similar to other trips whereas a lower o_c indicates that the corresponding trip is irregular. Then, we define an irregular trip if $o_c < M_i - D_i$, where M_i and D_i are the mean value and standard deviation of $\{o_c\}$.

To avoid over-removing, we set a threshold so that the total duration of removed trips should be less than 10% of the duration of all trips. If the ratio is higher than 10%, the irregular trips will not be removed from TR_i given their importance.

After this, we apply the general K-means method to all the individuals [15], with the distance calculated as sum of absolute differences.

$$m_i = \begin{cases} \frac{\sum\limits_{k=1}^{2} \sum\limits_{t_j \in S_k} \left|t_j - \mu_k\right|}{n} & \text{if } \left|\mu_1 - \mu_2\right| \leq 2h \\ \frac{\sum\limits_j \left|t_j - \frac{\sum t_j}{n}\right|}{n} & \text{otherwise} \end{cases} \tag{S16}$$

where μ_k is the mean of $\{t_j | t_j \in S_k\}$.

The general algorithm of calculating m_i is shown as Algorithm S1:

In Fig. S8b, we show the transit use of 3 individuals. The top inset shows the transit use of a regular transit user. Generally, this individual took buses at the same time from Monday to Friday, resulting in a very low $m_i = 414s \approx 6.9min$. Owing to the very regular transit use, the encounter rate of this individual is $r_i = 41.21h^{-1}$, which is remarkably high. In the middle inset, we plot transit use of an irregular user. Clearly, the times of this individual taking buses are random over the week, with a high $m_i = 8097s \approx 135min$ and a low $r_i = 1.23h^{-1}$. The transit use of the third user is a hybrid of regular commuting and random trips. After measuring the total duration of random trips, they are not removed, given their importance. Thanks to the regular commuting trips, the encountering rate $r_i = 7.14h^{-1}$ is not too low. However, the calculated $m_i = 5596s \approx 93min$ is high, owing to the additional random trips.

VII.3 Distributions $P(r)$ and $P(m)$

To explore how individual regularity influences his/her encounter likelihood, we chose all the individuals who encounter at least one "familiar stranger" during the week, obtaining a population of 1,626,040 people.

Algorithm S1 Calculating m_i with irregular trips removed

```
for $i = 1$ to $N$ do
    $TR_i \leftarrow \{tr_1, \cdots, tr_n\}$ trip set of individual $i$
    $\{q(tr_c) | tr_c \in TR_i\} \leftarrow$ Calculate $q(tr_c)$ for each trip
    for $c = 1$ to $n$ do
        $o_c \leftarrow \frac{q(tr_c)}{\sum_k q(tr_k)}$
    end for
    $M_i \leftarrow mean\ \{o_c | c = 1, \cdots, n\}$
    $D_i \leftarrow standard\ deviation\ \{o_c | c = 1, \cdots, n\}$
    for $c = 1$ to $n$ do
        if $o_c < M_i - D_i$ then
            $TR_i \leftarrow TR_i \backslash \{tr_c\}$
        end if
    end for
    $(\{u\}, \{S\}) \leftarrow$ K-means($TR_i$)
    $m_i \leftarrow$ Calculate absolute trip difference.
end for
```

Thus, following the previous computation, we obtain parameters r_i and m_i for the selected population and measure the distribution of $P(r)$ and $P(m)$ separately. Fig. S9a shows the distribution of $P(r)$ over the individuals with $r_i \geq 1$. We find $P(r)$ is well characterized by a truncated power-law, indicating the rescaled encounter likelihood varies strongly across the population, with most people having a small r. Next, we plot the probability density function $P(m)$ of absolute trip differences (Fig. S9b), finding $P(m)$ is characterized by a narrow peak followed by a fat tail. Given the definition of m, small m indicates regular behaviors while high m suggests random traveling. Thus, the distribution results from a combination of regular commuters and random travelers.

To explore whether r_i is rooted in one's transit use patterns, we then measure the joint distribution $P(r,m)$ over the selected population (Fig. S9c). Interestingly, with the increase of r, we observe a convergence to $m \approx 10min$. We find that for the people with small m, the distribution $P(r)$ is wide, suggesting that it is not clear that individuals with small m encounter more "familiar strangers" given other factors such as trip timing and frequency. However, it is clear that people with larger r tend to behave more regularly. The dotted line indicates the average trend of $m < m_{0.95}$ ($m_{0.95}$ is the 95^{th} percentile of m), showing that the average decreases with r. Taken together, the distribution shows that one's high encounter likelihood is rooted in one's regular daily behavior.

VIII Evolution of aggregated network $G(0,t)$

To analyze the temporal contact network, we derive static networks to capture both temporal and topological properties (Fig. S10). To explore the evolution and dynamics creating $G(0, 168h)$, we reconstruct the

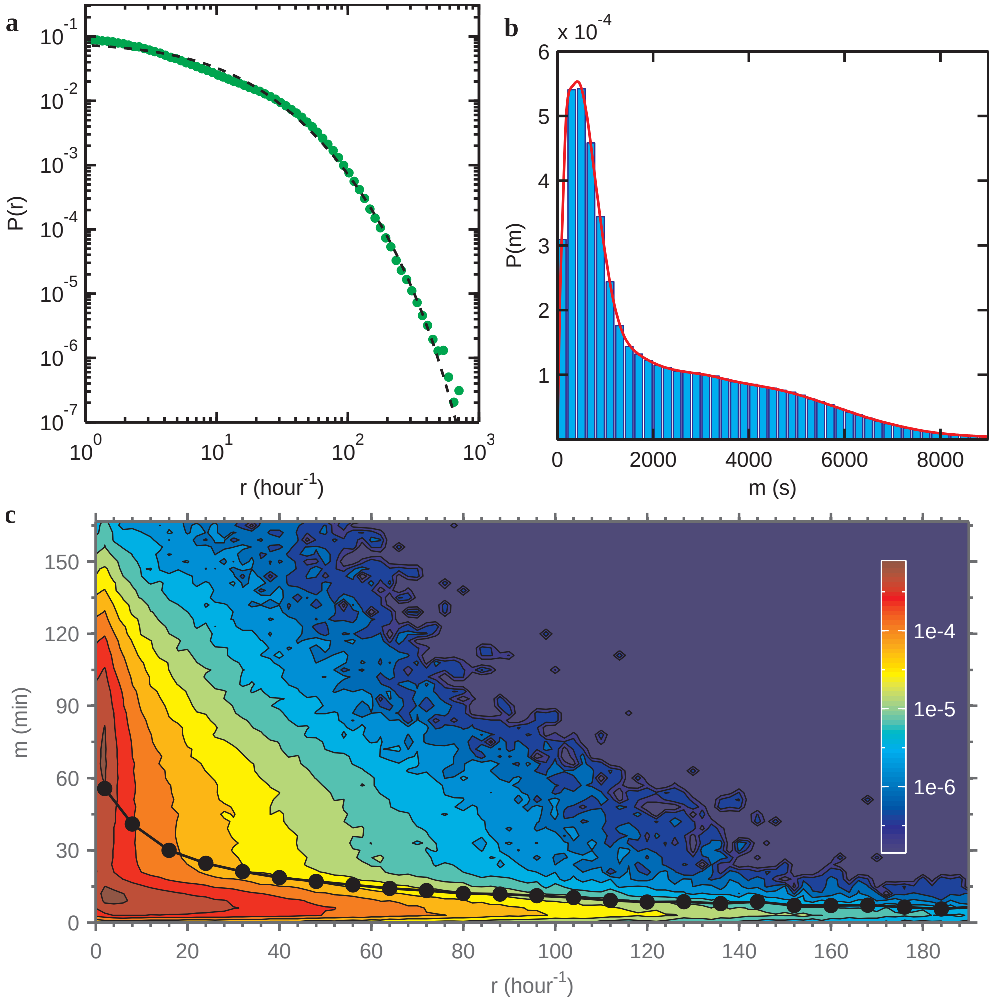

Figure S9: Probability distributions of rescaled encounter likelihood r and absolute trip difference m. **(a)**, Probability density function of $r_i(r_i \geq 1)$ in log-log scale. The dashed line indicates a truncated power law $P(r) \sim (r+r_0)^{-\beta}\exp\left(-\frac{r}{\alpha}\right)$, with $r_0 = 22.8$, $\beta = 2.9$ and $\alpha = 116$. **(b)**, Distribution $P(m)$ over the selected population. **(c)**, Joint probability density $P(r,m)$. Solid dotted line indicates the trend of mean value of $m < m_{0.95}$ against r.

network in temporal scale, obtaining the evolution of the aggregated network $G(0,t)$:

(1) Number of vertices N_t

(2) Degree distribution $P(k)$

(3) Average degree $\langle k \rangle_t$

(4) Clustering coefficient c_t

To distinguish regular commuters from random travelers, we divide the population into two groups: regular commuters with personal weight $w_i \geq 10$, and the rest. Then we create randomly drawn populations of 500,000 individuals, with different fractions of regular commuters from $\rho = 0\%$ to $\rho = 100\%$.

VIII.1 Number of vertices N_t

We first measure growth of the aggregated network $G(0,t)$ in terms of number of vertices N_t until time t. Fig. S11a shows the fraction of vertices — $\frac{N_t}{N}$, where $N = 500,000$ is the size of the final population.

We observe that, when the population is composed of regular commuters, the number of individuals N_t exhibits a rapid increase at the beginning of the week, containing about 90% of the population after Monday, followed by a slower growth to saturation. However, when we reduce the fraction of regular commuters, growth rate is reduced. In fact, this pattern results from the difference of inter-event interval τ of different types of transit users. Regular commuters tend to take buses every day, while random passengers appear randomly in the transit system during the week. Thus, if the population is composed of random passengers who take buses once and uniformly distributed during the week, we expect to observe a linear increase.

VIII.2 Degree distribution $P(k)$

Fig. S11b shows the degree distributions $P(k|\rho)$. As can be seen, $P(k|\rho = 0)$ can be well characterized by an exponential distribution, while the share of low k is decreasing gradually in $P(k|\rho)$ with the increase of ρ from 0% to 100%. In fact, we know that most of the population are regular commuters, so they tend to encounter more people, leading to the increase of $P(k|\rho = 1)$ from $k = 0$ to $k = 200$. Despite the variation, we find that all $P(k|\rho)$ can be well-fitted with exponentially decaying tails, which is in accordance with the degree distribution over the whole population (Fig. S7a).

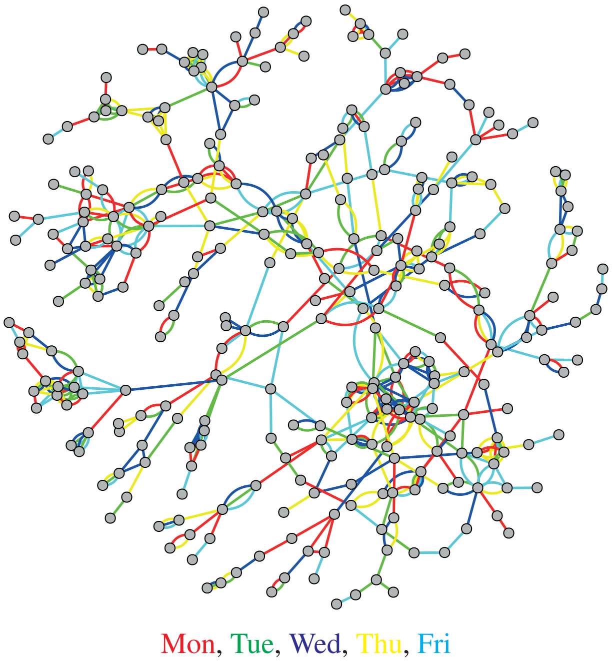

Figure S10: The growth of a sample encounter network over weekdays. The colors of edges indicate their creation by day of week. Please note that some pairs encounter each other more than once.

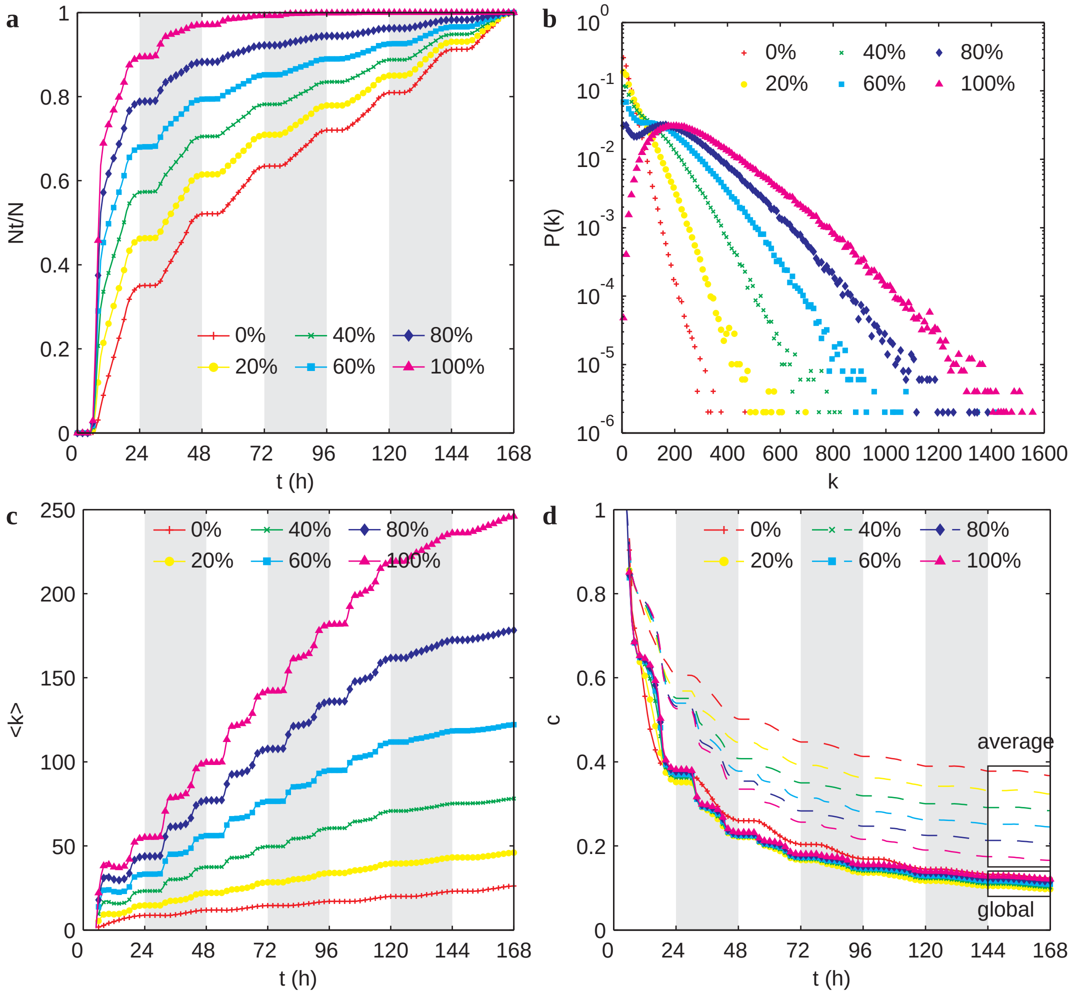

Figure S11: Evolution of aggregated encounter networks created over populations with different fractions of regular commuter. (We show only 6 of the 11 populations created for simplicity.) **(a)**, Fraction of presented users. **(b)**, Degree distribution $P(k)$. **(c)**, Average degree $\langle k \rangle$. **(d)**, Global clustering coefficient (solid lines) and average clustering coefficient (dashed lines).

VIII.3 Average degree $\langle k \rangle_t$

The average degree of aggregated network $G(0,t)$ at time t is given by:

$$\langle k \rangle_t = \frac{2E_t}{N_t} \tag{S17}$$

where E_t is the number of edges at t.

As stated before, the average degree of instantaneous networks can be estimated by $\langle k \rangle_{it} \approx \frac{N_{it}}{m}$. However, the cumulative degree k_t for one individual in an aggregated network is given by the sum of encounters created over $(0,t)$. In other words, k_t is also influenced by the frequency of taking buses. If we assume the number of encounters in each trip to be a constant n, the average degree may be calculated as:

$$\langle k \rangle_t = n\bar{f}_t \tag{S18}$$

where $\bar{f}_t$ is the average number of bus trips over $(0,t)$, which is linear increasing with t. Thus, we may find $\langle t \rangle$ is also linear increasing with t without saturation.

As shown in Fig. S11c, we observe steady linear increases of $\langle k \rangle_t$ with t. Hence, unlike social networks with stable relationships, even if we further extend the aggregation time, no saturation will be observed, given the partially random characteristics of encounters in this contact network.

Evolution of the average number of encountered people $\langle e \rangle$ is presented in the manuscript, suggesting that stronger social connections are emerging from the random daily encounters.

VIII.4 Clustering coefficient c_t

Two clustering coefficients are measured here according to the definition in Ref. [16].

The global clustering coefficient or transitivity of a network is defined as:

$$c_1 = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of vertices}} \tag{S19}$$

and the average clustering coefficient of a network is defined as:

$$c_2 = \frac{1}{n} \sum_i c_{li} \tag{S20}$$

where $c_{li} = \dfrac{\text{number of triangles connected to vertex } i}{\text{number of triples centered on vertex } i}$ is the local clustering coefficient of vertex i.

Therefore, both c_1 and c_2 lie in the range $0 \le c_t \le 1$. Fig. S11d shows the evolution of both global clustering coefficient and average clustering coefficient. Interestingly, we find the global clustering coefficients for different combinations of regular and irregular users display a strong convergence to $c_1 = 0.1095 \pm 0.0093$ (mean± standard deviation) after $168h$, whereas the average clustering coefficients exhibit a more dispersed pattern with $c_2 = 0.2649 \pm 0.0660$.

In fact, as stated in [16], c_2 tends to weight the contributions of vertices with low-degree more heavily and users with lower degree tend to have higher local clustering coefficients due to less transit frequency since the encounters of one individual are bounded in one or two vehicles. With the increase of ρ, the average frequency of transit trips and average degree are increasing as well, resulting in c_2 decreases from 0.3668 to 0.1656.

As opposed to c_2, c_1 actually measures the mean probability $P(connected)$ that two vertices are connected if they are connected to same third vertex. In terms of this encounter network, c_1 represents the probability of two individuals having encountered each other, if they both have encountered the same third individual.

Thus, the strong convergence of c_1 over the population with different ρ captures a general trend of $P(connected)$, although the population composition is quite different.

At the individual level, the local clustering coefficient c_{li} has been widely used to describe individuals in sociology. As reviewed in [16], many researchers found that $c_{li} \sim k_i^{-1}$ holds approximately true for certain models of scale-free networks, such as it implies $\ln(c_{li}) \sim -\ln(k_i)$. However, the aggregated network here is different from the previously studied scale-free networks, given their formation mechanism and the evolution of c_{li} with k_i. To further explore the evolution of clustering coefficient individually, we measure

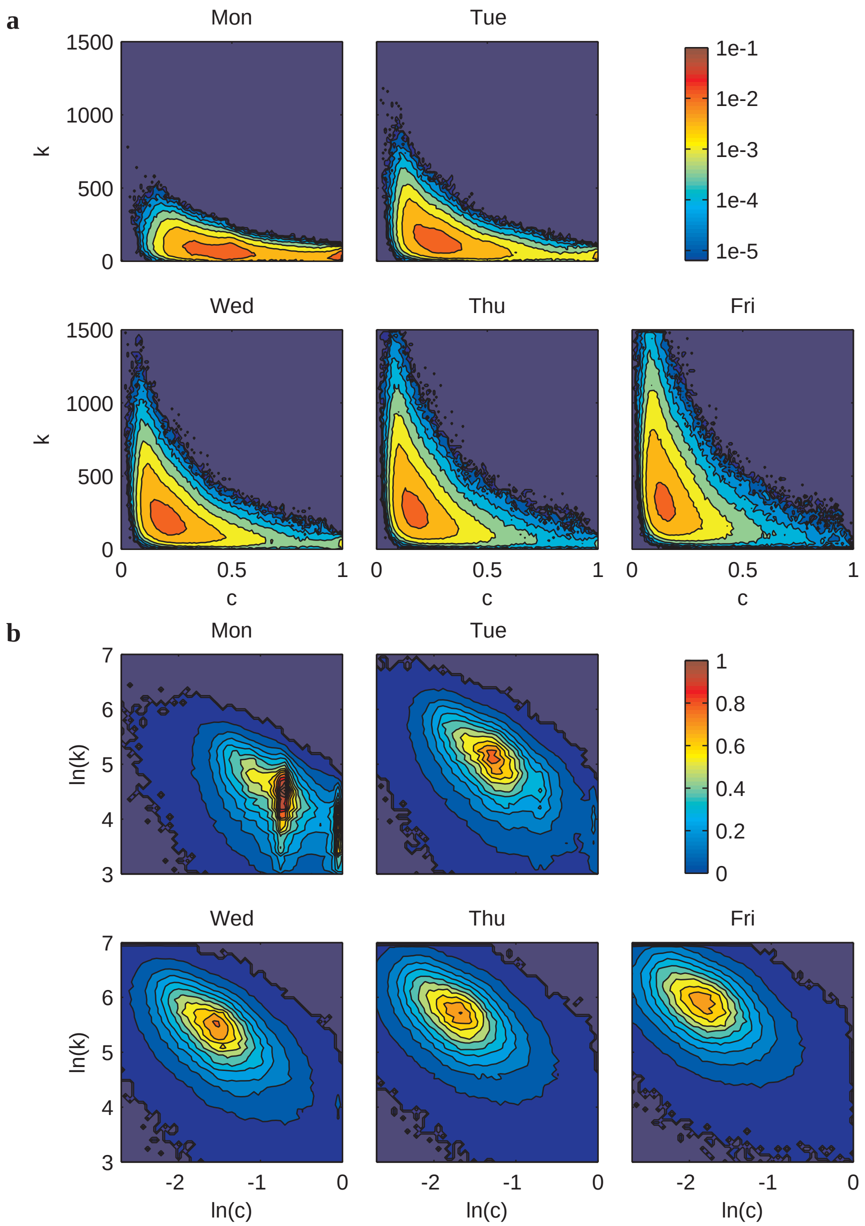

Figure S12: Joint probability distribution $P(k,c)$ of degree and local clustering coefficient at the end of each weekday. The parameters k and c are obtained from 783,247 individual with person weight $w_i \geq 10$. **(a)**, Evolution of $P(k,c)$. **(b)**, Evolution of $P(\ln(k),\ln(c))$. The correlation coefficient r of $\ln(k_i)$ and $\ln(c_{li})$ from Tuesday to Friday are $r = \{-0.61,-0.60,-0.56,-0.52\}$. The standard deviation of $\ln(k_i)$ and $\ln(c_{li})$ are $\sigma_{k_i} = \{0.69,0.66,0.62,0.59\}$ and $\sigma_{c_{li}} = \{0.47,0.48,0.48,0.48\}$, respectively.

the joint distribution $P(k_i, c_{li})$ over the population with $w_i \geq 10$ (k_i here means number of encountered people). The daily evolution of $P(k_i, c_{li})$ during the week is shown in Fig. S12a. From the movement of the central cluster, it can be seen that c_{li} decreases with the increase of k_i. When looking into the variation of $(\ln(k), \ln(c))$, strikingly, we found the distribution $P(\ln(k), \ln(c))$ presents a clear symmetrical pattern as shown in Fig. S12b after Monday. We also observe that the central cluster of $P(\ln(k), \ln(c))$ is moving with time, showing the general pattern that clustering coefficient c decreases with the increase of degree k.

References

[1] Pelletier MP, Trépanier M, Morency C (2011) Smart card data use in public transit: A literature review. *Transp Res Part C Emerg Technol* 19:557–568.

[2] Axhausen KW (2007) Definition of movement and activity for transport modelling. *Handbook of Transport Modelling, 2nd Edition* eds Hensher DA, Button KJ (Elsevier, Oxford).

[3] Singapore Department of Statistics (2011) *Census of population 2010: Statistical release 3, Geographic distribution and transport* (Department of Statistics, Ministry of Trade & Industry, Singapore).

[4] Cheong C, Toh R (2010) Household interview surveys from 1997 to 2008 – A decade of changing travel behaviours. *LTA JOURNEYS* 4:52–61.

[5] Schönfelder S, Axhausen KW (2010) *Urban rhythms and travel behaviour: Spatial and temporal phenomena of daily travel* (Ashgate, Farnham, UK).

[6] Barabási AL (2005) The origin of bursts and heavy tails in human dynamics. *Nature* 435:207–211.

[7] Vázquez A (2005) Exact results for the Barabási model of human dynamics. *Phys Rev Lett* 95(24):248701.

[8] González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. *Nature* 453:779–782.

[9] Perra N, Gonçalves B, Pastor-Satorras R, Vespignani A (2012) Activity driven modeling of time varying networks. *Sci Rep* 2:469.

[10] Barabási AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509–512.

[11] Isella L, et al. (2011) What's in a crowd? analysis of face-to-face behavioral networks. *J of Theor Biol* 271(1):166–180.

[12] Zhang Y, Wang L, Zhang YQ, Li X (2012) Towards a temporal network analysis of interactive wifi users. *Europhys Lett* 98(6):68002.

[13] Milgram S (1974) The frozen world of the familiar stranger. *Psychol Today* 17:70–80.

[14] Grannis R (2009) *From the ground up: Translating geography into community through neighbor networks* (Princeton University Press, New Jersey).

[15] Hartigan JA (1975) *Clustering algorithms* (John Wiley & Sons, Inc., New York).

[16] Newman MEJ (2003) The structure and function of complex networks. *SIAM Rev Soc Ind Appl Math* 45(2):167–256.