OPEN ACCESS Freely available online

PLoS one

Collective Human Mobility Pattern from Taxi Trips in Urban Area

Chengbin Peng[1,2], Xiaogang Jin[2], Ka-Chun Wong[1], Meixia Shi[3], Pietro Liò[4]*

1 Mathematical and Computer Sciences and Engineering Division, King Abdullah University of Science and Technology, Jeddah, Kingdom of Saudi Arabia, **2** Institute of Artificial Intelligence, College of Computer Science, Zhejiang University, Hangzhou, China, **3** College of Environmental and Resource Sciences, Zhejiang University, Hangzhou, China, **4** Computer Laboratory, Cambridge University, Cambridge, United Kingdom

Abstract

We analyze the passengers' traffic pattern for 1.58 million taxi trips of Shanghai, China. By employing the non-negative matrix factorization and optimization methods, we find that, people travel on workdays mainly for three purposes: commuting between home and workplace, traveling from workplace to workplace, and others such as leisure activities. Therefore, traffic flow in one area or between any pair of locations can be approximated by a linear combination of three basis flows, corresponding to the three purposes respectively. We name the coefficients in the linear combination as traffic powers, each of which indicates the strength of each basis flow. The traffic powers on different days are typically different even for the same location, due to the uncertainty of the human motion. Therefore, we provide a probability distribution function for the relative deviation of the traffic power. This distribution function is in terms of a series of functions for normalized binomial distributions. It can be well explained by statistical theories and is verified by empirical data. These findings are applicable in predicting the road traffic, tracing the traffic pattern and diagnosing the traffic related abnormal events. These results can also be used to infer land uses of urban area quite parsimoniously.

Citation: Peng C, Jin X, Wong K-C, Shi M, Liò P (2012) Collective Human Mobility Pattern from Taxi Trips in Urban Area. PLoS ONE 7(4): e34487. doi:10.1371/journal.pone.0034487

Editor: Matjaz Perc, University of Maribor, Slovenia

Received January 17, 2012; **Accepted** February 28, 2012; **Published** April 18, 2012

Funding: CP was supported by Graduate Fellowship from King Abdullah University of Science and Technology. CP and XJ were supported by the National Science Foundation of China under Grant No. 61070069. PL was supported by the following project: RECOGNITION: Relevance and Cognition for Self-Awareness in a Content-Centric Internet (257756), which is funded by the European Commission within the 7th Framework Programme (FP7). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Pietro.Lio@cl.cam.ac.uk

Introduction

Urban traffic has drawn the attention of physicists since more than one decade ago. Generally, there has been two kinds of approaches for the traffic analysis. In microscopic models, some researchers represent vehicles as particles interacting with each other [1,2], while some others use the cellular automata framework [1,3,4]. Based on game theory, the impact of individuals' irregular behaviors on traffic system is also emphasized [5]. On the other hand, from the macroscopic perspective, the idea of fluid dynamics is introduced [1,6].

In recent years, a new and more fundamental approach for traffic analysis is emerging: human mobility, by drawing statistical inferences from the enormous empirical data [7–9]. Several reasons boost the research in this area.

Firstly, the knowledge of the mobility pattern is essential in traffic modeling [10,11] for simulation, forecasting [12,13] and control [11]. In addition, by measuring the traffic flow during some time interval to see whether or not it agrees with the verified estimation, the collective mobility analysis can serve as a tool for abnormality definition and detection [14,15]. Compared to computer vision based detection [16,17], collective mobility model based abnormality detection can be applied in a much larger scale of area, for example, the whole city.

Secondly, the mobility pattern and the consequential traffic flow can also interact with the land use. The characteristics of traveling strongly influence urban formation, evolving, and future planning [18–21], whereas the land use can also affect the urban traffic [22–24] and the human mobility [25].

Thirdly, the better understanding of human mobility can help to more easily control the spreading of contagious diseases by limiting the contact among individuals [26], since the transmission of infected people from one place to another is an important way to infect the susceptible ones, either in a small scale area [27,28] or from a worldwide viewpoint [29–31]. Similar theories hold for viruses contamination with malicious code among wireless communication devices [32,33].

Due to the high importance of human mobility research, and the availability of the large amount of empirical data as a consequence of the prevalence of wireless communication devices, researchers become more and more interested in the statistical features of human mobility pattern via real world data [34]. Ref. [7] and Ref. [9] suggest that human travels are reminiscent of Lévy Flights [35] according to the trajectories of bank notes and taxies respectively, while Ref. [36] reports some variances by the GPS information from volunteers. These differences are later recognized as a result of the periodic pattern of individual's traveling [8] and recently Ref. [37] discovers up to 93% of total time when individual locations are predictable in their data set, which contains trajectories of mobile phone users. For taxi trips, Ref. [38] studies the distribution of the travel distances and time.

Nevertheless, previous statistical inferences of human mobility mostly focus on individual level, while this article analyzes the citizens' collective dynamics in the urban area. In our research, based on the traveling purposes, we discovered three distinct basis patterns for collective traffic flow regardless of the location. In addition, a distribution is revealed that can characterize the fluctuation of the traffic flow at any time in each location. As mentioned above, these findings can be useful for urban planning, traffic estimation and anomalous detection. Further studies on interaction between different areas will provide a more detailed collective mobility model, and would additionally benefit the research on epidemic spreading in urban area.

Analysis

Data Description and Background Assumptions

In this research, the data [39] are collected from about two thousand taxies operating within the urban area of Shanghai, China. These data mainly focus on the central part of city, and the population in this part is about seven million according to the fifth national population census [40]. The information about when and where passengers were picked up and dropped off can be retrieved from the raw data, and every pair of picking and dropping information is defined as a taxi trip. The data set includes about 1.58 million taxi trips. The longitude and latitude location information in the data by GPS is converted to positions in a planar coordinate system, with the city landmark Oriental Pearl Tower as the origin. For the ease of analyzing and representing, the urban area is divided into squares, similar to a chessboard. The side lengths of each square is identically 200 meters. In our context, each location corresponds to one of these squares. More details can be found in Appendix S1.

Basis Traffic Flows: the Constancy

As we know, even a $200\mathrm{m}\times 200\mathrm{m}$ area in a city can possess land of several different types, for example, containing schools, shops and apartments at the same time. In this section, we will discuss how to categorize the taxi trips according to the traveling purposes, and then use these categories to infer the land use composition for each square.

First of all, we consider the taxi trip categorization. People setting out in the same location would possibly have different purposes: some may go to workplaces while some others may go for entertainment. Meanwhile, for trips belonging to the same category but in different locations, the collective pattern should be similar, regarding to the departure and arrival time in a large amount of data. For example, if the number of trips between residential area and workplaces (for commuting purpose) reaches the highest at 8:00 am (going to work) and 5:00 pm (getting off work), then the number of trips in this category in any place would peak almost at the same time, although the scale may be different.

In short, we can define a set of basis collective patterns, each of which corresponding to a trip category respectively. Then linear combinations of these patterns can describe the macro traveling pattern of each location. Finally, the coefficients in a linear combination can reflect the land uses of the location.

Directly from the taxi data, we can only calculate the macro patterns. Therefore, we should adopt appropriate inference methods to find the basis patterns and the coefficients for each location.

To represent our method more formally, we define (i,j) to index the square in ith row and jth column among all the squares divided within the city. If m is the number of rows and n is the number of columns for squares in the map, then $i\in[1,m]\cap\mathbb{Z}$, and $j\in[1,n]\cap\mathbb{Z}$. Let h be the number of time slots, normally 24 for one day. Therefore for location (i,j), the numbers of departure and arrival trips (macro pattern) along time each day can be represented by a $1\times h$ vector $\mathbf{S}_{i,j}$, which is easy to calculate. We can also define a set of $1\times h$ vectors containing normalized numbers of trips along time: $\mathbf{B}_1$, $\mathbf{B}_2$, $\mathbf{B}_3$, …, $\mathbf{B}_K$, each for one basis pattern that we seek for.

The macro pattern is a linear combination of basis patterns, so we have

$$\mathbf{S}_{i,j}=\mathbf{P}_{i,j}\begin{bmatrix}\mathbf{B}_1\\ \mathbf{B}_2\\ \mathbf{B}_3\\ \vdots\\ \mathbf{B}_K\end{bmatrix} \tag{1}$$

where $\mathbf{P}_{i,j}$ is a row vector containing K coefficients for the linear combination on the right-hand side.

By taking all the locations into account, it can also be written as

$$\begin{bmatrix}\mathbf{S}_{1,1}\\ \mathbf{S}_{1,2}\\ \vdots\\ \mathbf{S}_{1,n}\\ \mathbf{S}_{2,1}\\ \mathbf{S}_{2,2}\\ \vdots\\ \mathbf{S}_{2,n}\\ \vdots\\ \mathbf{S}_{m-1,n}\\ \mathbf{S}_{m,1}\\ \vdots\\ \mathbf{S}_{m,n}\end{bmatrix}=\begin{bmatrix}\mathbf{P}_{1,1}\\ \mathbf{P}_{1,2}\\ \vdots\\ \mathbf{P}_{1,n}\\ \mathbf{P}_{2,1}\\ \mathbf{P}_{2,2}\\ \vdots\\ \mathbf{P}_{2,n}\\ \vdots\\ \mathbf{P}_{m-1,n}\\ \mathbf{P}_{m,1}\\ \vdots\\ \mathbf{P}_{m,n}\end{bmatrix}\begin{bmatrix}\mathbf{B}_1\\ \mathbf{B}_2\\ \mathbf{B}_3\\ \vdots\\ \mathbf{B}_K\end{bmatrix} \tag{2}$$

and abbreviated as

$$\mathbf{S}=\mathbf{PB} \tag{3}$$

Because the two matrices on the right-hand side of Eq. (3) are unknown, there are many matrix decomposition methods that may apply. However, according to the physical meaning of $\mathbf{P}$ and $\mathbf{B}$, all the entries of these two matrices should be nonnegative. Therefore, we choose nonnegative matrix factorization (NMF) [41,42] for the decomposition.

In our context, it is a method to factorize a matrix $\mathbf{S}\in\mathbb{R}_+^{mn\times h}$ into two nonnegative factors $\mathbf{P}\in\mathbb{R}_+^{mn\times K}$ and $\mathbf{B}\in\mathbb{R}_+^{K\times h}$ approximately. By this approach, we can find the basis patterns (the row vectors of $\mathbf{B}$) and the parameter vectors (the row vectors of $\mathbf{P}$) simultaneously. As vector $\mathbf{P}_{i,j}$ (the $[(i-1)m+j]$th row of matrix $\mathbf{P}$) is only responsible for vector $\mathbf{S}_{i,j}$ (the $[(i-1)m+j]$th row of matrix $\mathbf{S}$), in fact, each element of $\mathbf{P}_{i,j}$ denotes the scale of traffic flow with respect to the corresponding category, in location (i,j). Hence, we also call these elements the traffic power because they reflect how strong the traffic flows of different categories are.

Now the only thing left is to determine K, the number of the basis patterns.

From the algorithmic perspective, we noticed that NMF starts with random initial conditions [41]. By experiments on the taxi data with many different random initial conditions, we find that only when K equals 3, the factorization results can be stable. This fact indicates that with parameter $K=3$, NMF can find out statistically significant characteristics for the data, and Fig. 1 demonstrates the resulted basis pattern $\mathbf{B}_1$, $\mathbf{B}_2$ and $\mathbf{B}_3$.

On the other hand, from the land-use and trip-category perspective, $K=3$ is a reasonable choice in categorizing trip purposes.

There are several land-use definitions related to the topic of mobility. For example, each place may be classified as a residential (home), working, shopping, or recreational location [27]. It may also be regarded as one of the following types: a residential area, a workplace, a commercial zone, a recreation area and educational facilities [43]. In Ref. [44], these types are simplified into workplace, home and shop. Specifically for the city of Shanghai based on GIS information, Ref. [45] refers to the land types including residence, industry, agriculture, roads, water, land for construction and other urban land. In our context, we can simplify the land-use definition to be: residences, workplaces and others. Here workplaces shall include any industrial and office workplaces as well as schools, and other places can include shopping and recreational facilities, hospitals, etc.

For trips, some scientists categorize these individual activities into several orientations: family, work, leisure and service-based movement [46]. Similarly, according to our land-use definition, we can use three purpose-based categories for the trips: commuting between home and workplace ($\mathbf{B}_1$), business traveling between two workplaces ($\mathbf{B}_2$), and trips from or to other places ($\mathbf{B}_3$). This representation is in accordance with the algorithmic result in Fig. 1. Take a typical workday as an example, based on our three categories, the major traffic flows in the city are supposed to be as follows: those from home to workplaces in the early morning (green line), from one workplace to another in the daytime (red line), from workplace to home or to other places at dusk (green line again), and those between other places and home in the night (blue line).

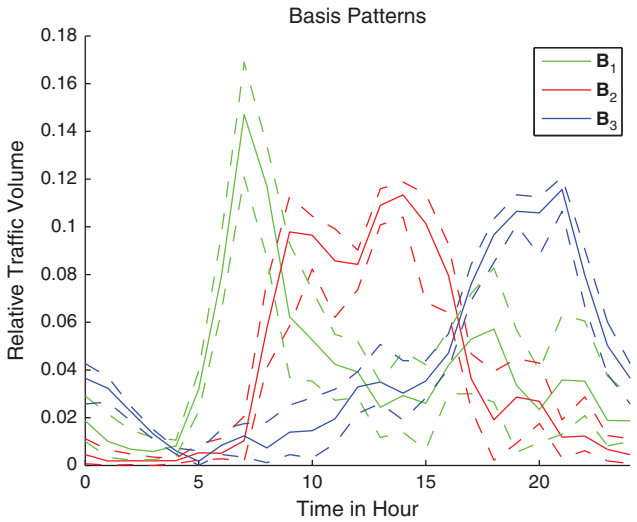

Figure 1. Basis Pattern B: Green is B1, Red is B2, and Blue is B3. Solid Lines Represent the Mean $\langle \mathrm{B} \rangle$, while Dashed Lines Represent the Positive and Negative Deviations Averaged on Different Days.
doi:10.1371/journal.pone.0034487.g001

Therefore, $K=3$ is an effective and reasonable choice.

In the following sections with $K=3$, for clarity, we will use **Bc**, **Bw** and **Bo** to replace $\mathbf{B}_1$, $\mathbf{B}_2$ and $\mathbf{B}_3$ respectively:

$$\mathbf{B}=\begin{bmatrix}\mathbf{B}_1\\ \mathbf{B}_2\\ \mathbf{B}_3\end{bmatrix} = \begin{bmatrix}\mathbf{Bc}\\ \mathbf{Bw}\\ \mathbf{Bo}\end{bmatrix} \tag{4}$$

We also use $Pc_{i,j}$, $Pw_{i,j}$ and $Po_{i,j}$ to represent the three entries in vector $\mathbf{P}_{i,j}$:

$$\mathbf{P}_{i,j}=\left[Pc_{i,j},Pw_{i,j},Po_{i,j}\right] \tag{5}$$

Appendix S2 describes the detailed implementation about applying NMF to this problem. The basis pattern on different days are averaged to $\langle \mathbf{B} \rangle$. Then, $\mathbf{P}_{i,j}$, the traffic power, can be recalculated based on $\langle \mathbf{B} \rangle$ for different day. If it variants in an acceptable interval day by day, the daily average of $\mathbf{P}_{i,j}$, represented by $\langle \mathbf{P}_{i,j} \rangle$, can indicate the land use of location (i,j). For example, if $\langle Pc_{i,j} \rangle$ is large, then the traffic flow corresponding to basis pattern $\langle \mathbf{Bc} \rangle$ is large, suggesting that location (i,j) serves mainly for residences or workplaces, while if $\langle Pw_{i,j} \rangle$ is the largest, we can be quite sure that this location is mainly for workplaces. In addition, if the variation of $\mathbf{P}_{i,j}$ on some day goes out of the acceptable interval, it indicates that something abnormal happens on that day. This feature can be helpful for anomaly detection on human activities in a large area. In the next section, we will analyze the variance of $\mathbf{P}_{i,j}$, to determine what is an acceptable interval.

Daily Traffic Power: the Variation

Typically in a city, the volume of the traffic flow is quite regular everyday [8]. However even for the same time in the same location but on different days, the volume is vulnerable to change within a certain range. This section is devoted to analyze how $\mathbf{P}_{i,j}$ fluctuates everyday. In this case, $\mathbf{P}$ is calculated from the average basis pattern $\langle \mathbf{B} \rangle$ according to Appendix S2.

We define a random variable α to represent the relative variance of the traffic power.

The empirical distribution function of α can be simply extracted from a collection of the following expressions in different locations on different days:

$$\frac{Pc_{i,j}}{\langle Pc_{i,j} \rangle}, \frac{Pw_{i,j}}{\langle Pw_{i,j} \rangle}, \frac{Pw_{i,j}}{\langle Pw_{i,j} \rangle} \tag{6}$$

where $\langle \cdot \rangle$ means the daily average, as we have used.

We also find the theoretical distribution function of α, which is more complex.

First, we try to find α only for the first category of trips in location (i,j). We define pn as the potential population that may affect the first-category traffic in this location, and r as the probability (ratio) that an individual in the population finally becomes part of that traffic flow. Then the number of such trips follows a binomial distribution:

$$P_{TN}(tn) = \binom{pn}{tn} r^{tn}(1-r)^{pn-tn} \tag{7}$$

where tn can be any non-negative integer less than pn. Because it is a binomial distribution, the corresponding CDF can be written in terms of the beta functions:

$$\begin{aligned} & D_{TN}(tn) \\ = \; & P_{TN}(TN \le tn) \\ = \; & 1 - P_{TN}(TN \ge tn+1) \\ = \; & 1 - I_r(tn+1, pn-tn) \end{aligned} \tag{8}$$

where $I_r(tn+1, pn-tn) = \frac{B(r : tn+1, pn-tn)}{B(tn+1, pn-tn)}$. $B(x : c_1, c_2)$ is the incomplete beta function as $B(x : c_1, c_2) = \int_0^x u^{c_1-1}(1-u)^{c_2-1}du$ and $B(c_1, c_2)$ is the beta function as $B(c_1, c_2) = \int_0^1 u^{c_1-1}(1-u)^{c_2-1}du$. Eq. (8) is strictly equal when tn is a positive integer, while for a real positive number of tn, we may use this approximation:

$$\begin{aligned} & D_{TN}(tn) \\ \approx \; & \frac{1}{2}\{[1 - I_r(tn+1, pn-tn)] \\ & + [1 - I_r((tn-1)+1, pn-(tn-1))]\} \end{aligned} \tag{9}$$

According to the definition, $\alpha = \frac{Pc_{i,j}}{\langle Pc_{i,j} \rangle} = \frac{TN}{\langle TN \rangle}$, where $\langle TN \rangle$ is equivalent to $pn \times r$ by the property of expectation of the binomial distribution, and can be treated as a constant for a given location. Therefore, the probability density function (PDF) of α is:

$$\begin{aligned} & P_\alpha(a) \\ = \; & P_{\alpha \times \langle TN \rangle}(a \times \langle TN \rangle) \\ = \; & P_{TN}(a \times \langle TN \rangle) \\ = \; & \binom{pn}{a \times \langle TN \rangle} r^{a \times \langle TN \rangle}(1-r)^{pn - a \times \langle TN \rangle} \end{aligned} \tag{10}$$

where a should satisfy the condition that $a \times \langle TN \rangle$ is a non-negative integer. The cumulative distribution function (CDF) is

$$\begin{aligned} & D_\alpha(a) \\ = \; & P_\alpha(\alpha \le a) \\ = \; & P_{TN}(TN \le \lfloor a \times \langle TN \rangle \rfloor) \\ = \; & D_{TN}(\lfloor a \times \langle TN \rangle \rfloor) \\ = \; & \sum_{k=0}^{\lfloor a \times \langle TN \rangle \rfloor} \binom{pn}{k} r^k (1-r)^{pn-k} \end{aligned} \tag{11}$$

where $\lfloor \cdot \rfloor$ where represents the floor function. We call this distribution the normalized binomial distribution of α. As listed in Appendix S3, the moment generation functions of α indicate that $\langle TN \rangle$ plays an essential role in the distribution. Numerical simulations also provide evidence that the distribution of α is strongly affected by $\langle TN \rangle$ (the product of pn and r), but is almost irrelevant to pn or r alone. Therefore, we can assign an constant integer N to pn.

Let $\mathbf{v}$ be a vector containing all the possible values of $\langle TN \rangle$. Then the PDF of α with $\langle TN \rangle = v_k$ can be written in this form

$$P_{\alpha, v_k}(a) = \binom{N}{a \times v_k} r^{a \times v_k}(1-r)^{N - a \times v_k} \tag{12}$$

and the CDF is

$$D_{\alpha, v_k}(a) = \sum_{k=0}^{\lfloor a \times v_k \rfloor} \binom{N}{k} r^k (1-r)^{N-k} \tag{13}$$

Finally, we discuss how to make α representative for variations of any traffic category in any location. We define a vector $\mathbf{s}$, in which each entry σ_k denotes the proportion of traffic flow corresponding to $\langle TN \rangle = v_k$. Then for a randomly selected traffic flow, when the average number of trips $\langle TN \rangle$ is not given, a general expression for the CDF of α is

$$D_\alpha(a) = \sum_k \sigma_k D_{\alpha, v_k}(a) \tag{14}$$

By beta approximation as in Eq. (9), it can be written into a continuous version

$$\begin{aligned} & D_\alpha(a) \\ = \; & \sum_k \sigma_k D_{\alpha, \langle TN \rangle}(a, v_k) \\ \approx \; & \sum_k \frac{1}{2}\sigma_k \{[1 - I_r(a \times v_k + 1, N - a \times v_k)] \\ & + [1 - I_r((a \times v_k - 1) + 1, N - (a \times v_k - 1))]\} \end{aligned} \tag{15}$$

Results

In this section, we demonstrate how our theoretical results are supported by the empirical investigation.

The general characteristics of our data set, such as the displacement distribution in Fig. 2 and the visiting frequency distribution in Fig. 3, are similar to others' [8,38]. The plot of daily

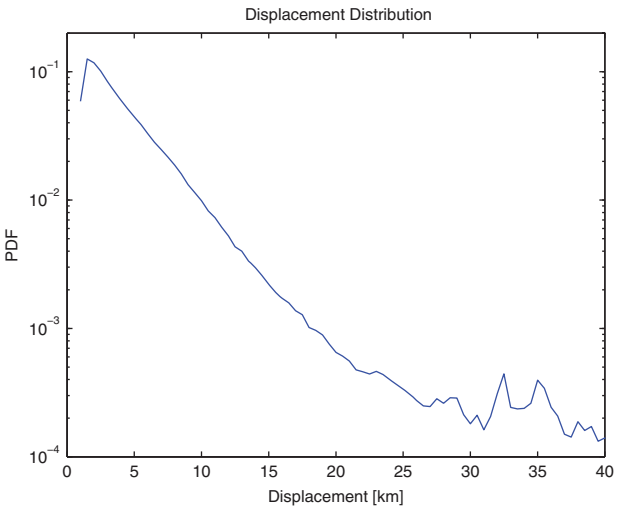

Figure 2. Traveling Distance Distribution.
doi:10.1371/journal.pone.0034487.g002

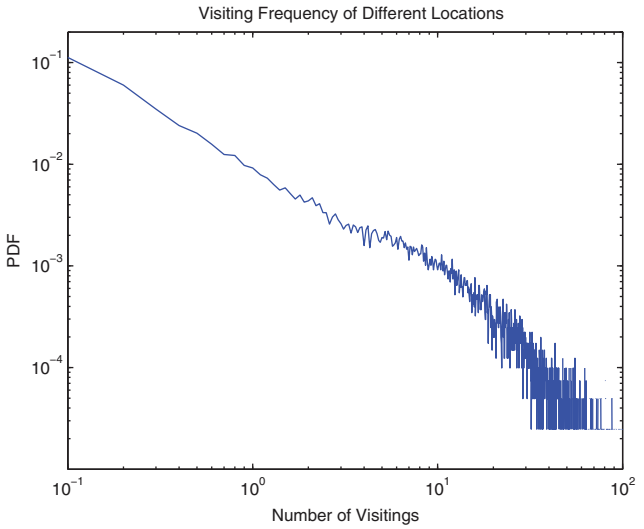

Figure 3. Visiting Frequency Distribution of Different Locations.
doi:10.1371/journal.pone.0034487.g003

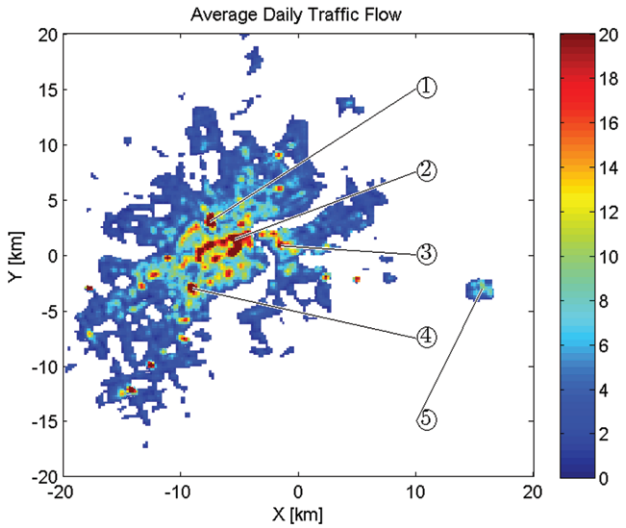

Figure 4. The Average Traffic Flow of Each Location, and the Tags Corresponding to Following Locations: ①Shanghai Railway Station; ②Nanjing Road & People's Square; ③Lujiazui Finance & Trade Zone; ④ Shanghai South Railway Station; ⑤ Pudong International Airport.
doi:10.1371/journal.pone.0034487.g004

traffic flow in Fig. 4 exhibits some hot areas by red, including the most flourishing commercial street Nanjing Road as the largest red block, Shanghai Railway Station, Shanghai South Railway Station, Lujiazui Finance & Trade Zone, etc. The largest isolated area in blue is the Pudong International Airport.

Without any intentional intervention, by NMF with random initial values, we find that the normalized basis pattern on workdays is generally quite similar (Fig. 1). Therefore, we can use the traffic power **P** to analyze the mean and the deviation of daily traffic.

In Fig. 5, the three components of $\mathbf{P}_{i,j}$ in every location is normalized and represented by yellow, red and blue respectively. For example, a location in yellow color means the traffic flow of the first category (**Bc**: commuting between home and workplace) is dominant there. Mixed colors in some places indicate a mixture of traffic flows of different categories. It is noticeable that in area where the traffic flow is large, the positive (Fig. 6(a)) and negative (Fig. 6(b)) deviation of the traffic power **P** is quite small. The distribution of this deviation can be represented accordingly by Fig. 7(a) and Fig. 7(b), which is fitted well with Eq. (15). This fitting result is quite different from the best fitted normal distribution by the central limit theory, which verifies Eq. (14) and Eq. (15) that α should be a collection of random variables following a set of distributions with different parameters. The proportion of traffic flow with $\langle TN \rangle = v_k$ is σ_k, as plotted in Fig. 8. Here we limit each σ_k to be no larger than twice of the empirical value. According to the result in Fig. 7, for the whole city, 80% of the deviations are within the range of $0.5 \sim 1.5$. Although the lengths of vectors **s** and **v** are identically 50 in our estimation, the number of active pairs ($\gg 0$) of σ_k and v_k is only about 10, and this number can be reduced if we only calculate for a small area given the sufficient amount of data. In short, we can see that Eq. (15) can be a reasonable approximation for the relative deviation of the daily traffic flow. Fig. 9(b) presents the components of **P** for the central part of the city in comparison with the urban planning map for Year 2004–2020 in Fig. 9(a). Generally, it can be seen that the residence area have a large volume of traffic with respect to **Bc** and **Bo**, corresponding to trips between home and workplaces and trips for other purposes, while in the workplace area especially for business, there are lots of flows corresponding to the second

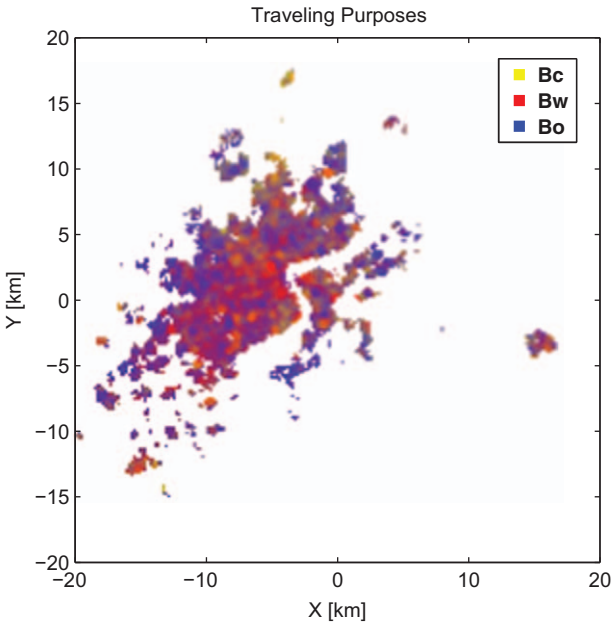

Figure 5. The Average Component Proportions of $\mathbf{P}_{i,j}$ in Each Location, Equivalent to the Categorical Proportion of the Traffic.
doi:10.1371/journal.pone.0034487.g005

category **Bw**, and in the remaining area, the third one **Bo** is quite significant. We should note that the urban planning map (2004–2020) is not an exact description for the land uses of Year 2007, and consequently, the patterns of the two figures may not agree well in some small areas. For example, the red patch around point $(-5,-5)$ in Fig. 9(a) is planned as an industrial land, namely,

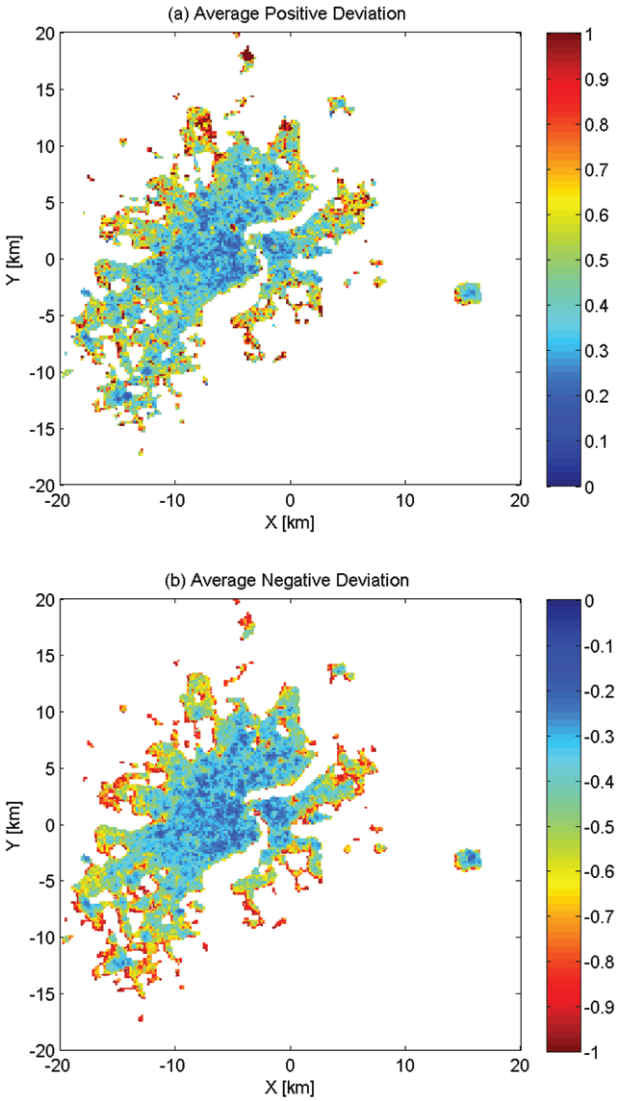

Figure 6. The Relative Deviation for Components of $P_{i,j}$ in Each Location: (a) the Average Positive Deviation; (b) The Average Negative Deviation.
doi:10.1371/journal.pone.0034487.g006

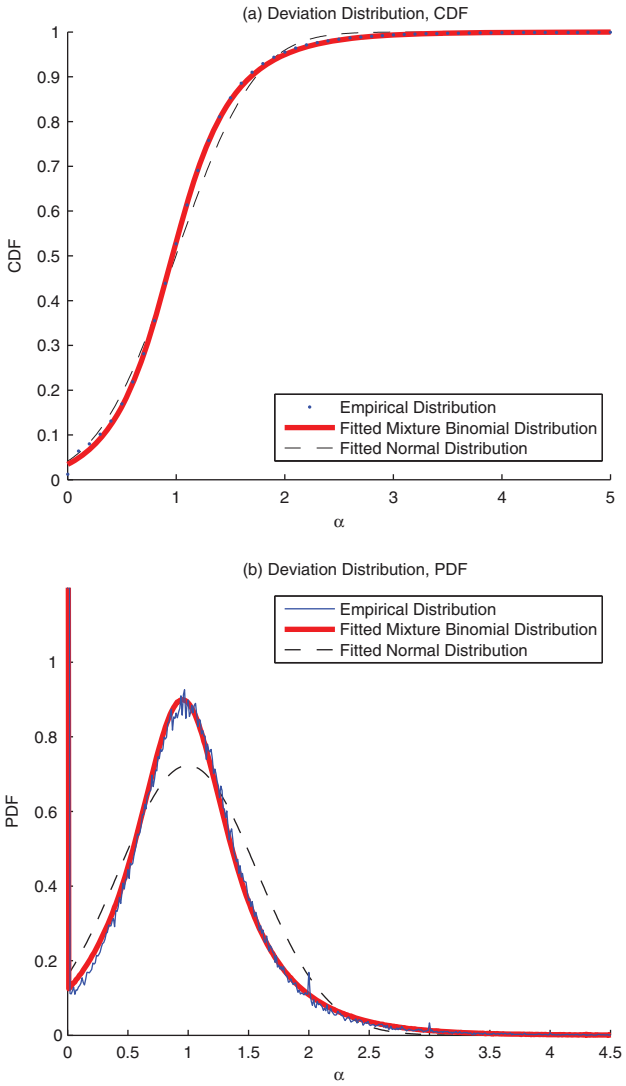

Figure 7. The Distribution of the Relative Deviation for Components of $P_{i,j}$: (a) CDF; (b) PDF.

workplace in our context, while in fact it was a construction site for Expo 2010 Shanghai China with very few taxi traffic in Year 2007. Yet it is still reasonable for a construction site to have the major taxi flows of type **Bo** as shown in Fig. 9(b) because in the evening workers would be very likely to go out for recreation, entertainments, etc.

In addition, we can see how the government planning [47] is affected by what it is now. For example, Nanjing Road and near by is the largest block with high traffic throughput, and traffic flows are constituted mainly by those of workplaces related (**Bw**) and other facilities related (**Bo**) categories. In the planning, it is designed to be a public activity center for administrative, business and shopping purposes. Lujiazui is another similar but smaller zone, which is planned mainly for business and shopping centers.

Discussion

In this research, we find that the traffic on workdays can be divided into three categories according to the different purposes:

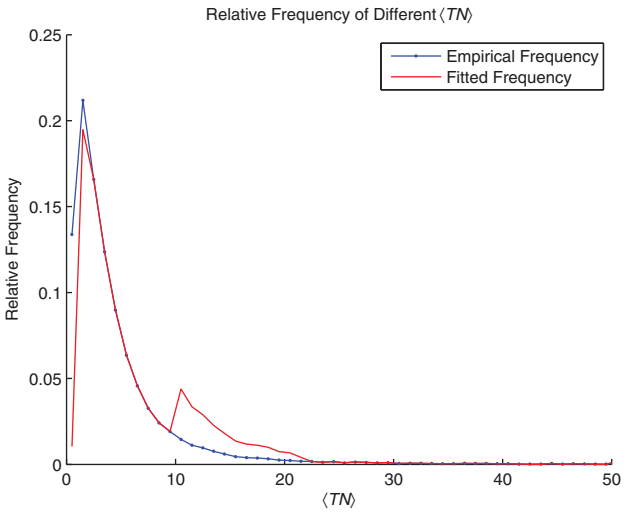

Figure 8. The Parameters for the Distribution.
doi:10.1371/journal.pone.0034487.g008

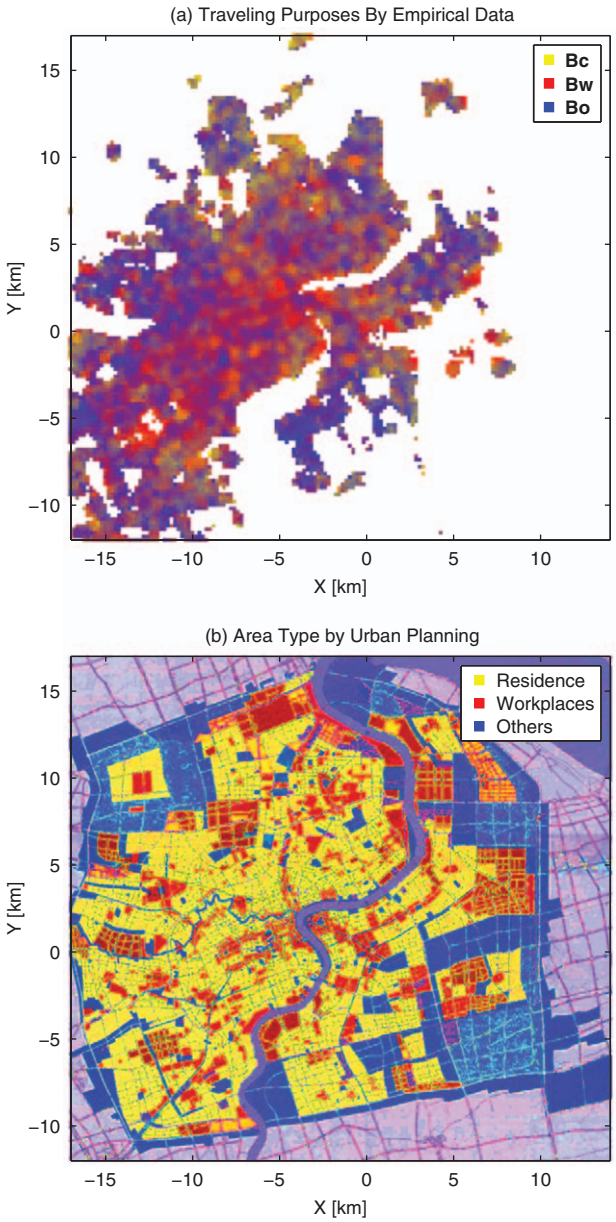

Figure 9. Comparing the Empirical Data to Urban Planning Map: (a) the Area Type from Urban Planning [47] for Central Part of the City; (b) the Average Categorical Proportion of Traffic for Central Part of the City.
doi:10.1371/journal.pone.0034487.g009

commuting between home and workplaces, traveling from workplace to workplace, and others such as leisure activities. Each of these categories has a highly distinguishable basis pattern: **Bc**, **Bw** or **Bo**. The relative daily deviation of the traffic flow in each category can be modeled as Eq. (14), which is a mixture of normalized binomial distributions, with a continuous approximation as Eq. (15).

This basis pattern theory is applicable to data sets containing the beginning and ending information of trips, such as the bicycle departure and arrival data [48], cell phone based mobility information [8], GPS based data, etc.

The first contribution of this research is, it provides a very economical approach to understand how the urban traffic at different locations are composed from the three categories. For instance, a large $Pc_{i,j}$ means there is a large portion of traffic between home and workplaces at location (i,j). This theory can also help to infer the land use composition by a quite easy, real-time, and automated way. For example, the evidence of a large $Pc_{i,j}$ everyday indicates location (i,j) is mainly for residential or working purpose, while a large $Pw_{i,j}$ can imply that it has lots of workplaces. A mixture of different land uses in a single location can be found by this method as well.

Second, based on the NMF approach, the time series of the total traffic at any location can be expressed as a linear combination of the basis patterns. Therefore, we can compress the traffic data of a large area into a very small data size, but still with a quite high resolution. Namely, we only need to store the global basis patterns, and for each location, we use a small vector for the traffic power to represent how strong each basis pattern is.

Third, we find that the distribution of the relative deviation is not a normal distribution, indicating that the random variable α is not identical from one place to another, or from time to time. The significance of Eq. (14) and Eq. (15) is, they provide an expression of how traffic fluctuates for various unknown positions and time intervals. This description of relative deviation can also be helpful to estimate the change of the traffic flow, which would be important in traffic predicting, controlling and urban planning.

Finally, with the deviation distribution, we can not only predict the change of traffic, but also diagnose the abnormality of the traffic: where, when, why, and how. The first two functions are obvious, while 'why' abnormal can be disclosed by the traffic power, and 'how' abnormal can be revealed by the probability of the deviation. For example, if some traffic flow is very abnormal one day, the probability density of the variance on that day should be very small.

Our analysis focusing on the traffic flows in different locations on different workdays. Our results can also be extend to the traffic on a road. The road traffic is a summation of the traffic passing this road from several sources and to several destinations. Therefore, the volume and the deviation of the road traffic flow can also be explained in our framework.

Supporting Information

Appendix S1 More on Data Description and Background Assumptions.
(PDF)

Appendix S2 Implementation Details about the Factorization.
(PDF)

Appendix S3 Moment Generation Function of α.
(PDF)

Acknowledgments

We would like to thank Wireless and Sensor networks Lab (WnSN, Shanghai Jiao Tong University, China) for providing the data source. We thank Dr. Min-You Wu, Yang Yang (Shanghai Jiao Tong University, China) for supports in data. We also thank Xianchuang Su, Dr. Yixiao Li, Dr. Yong Min and Chuanzi Chen (Zhejiang University, China), Dr. David Keyes and Dr. Xiangliang Zhang (King Abdullah University of Science and Technology, Saudi Arabia) for precious suggestions. For computer time, this research used the resources of the Supercomputing Laboratory at King Abdullah University of Science & Technology (KAUST) in Thuwal, Saudi Arabia.

Author Contributions

Conceived and designed the experiments: CP XJ PL. Performed the experiments: CP KW. Analyzed the data: CP XJ KW MS PL. Contributed reagents/materials/analysis tools: CP PL. Wrote the paper: CP XJ PL.

References

1. Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. Physics Reports 329: 199–329.
2. Nagel K (1996) Particle hopping models and traffic flow theory. Physical Review E 53: 4655.
3. Esser J, Schreckenberg M (1997) Microscopic simulation of urban traffic based on cellular automata. International Journal of Modern Physics C-Physics and Computer 8: 1025–1036.
4. Simon P, Nagel K (1998) A simplified cellular automaton model for city traffic. Arxiv preprint cond-mat/ 9801022.
5. Perc M (2007) Premature seizure of traffic flow due to the introduction of evolutionary games. New Journal of Physics 9: 3.
6. Helbing D (1995) Improved fluid-dynamic model for vehicular traffic. Physical Review E 51: 3164.
7. Brockmann D, Hufnagel L, Geisel T (2006) The scaling laws of human travel. Nature 439: 462–465.
8. González M, Hidalgo C, Barabási A (2008) Understanding individual human mobility patterns. Nature 453: 779–782.
9. Jiang B, Yin J, Zhao S (2009) Characterizing the human mobility pattern in a large street network. Physical Review E 80: 021136.
10. Leutzbach W (1987) Introduction to the theory of traffic flow. Springer Verlag.
11. Kerner B (2009) Introduction to modern traffic flow theory and control: the long road to threephase traffic theory. Springer Verlag.
12. Kitamura R, Chen C, Pendyala R, Narayanan R (2000) Micro-simulation of daily activity-travel patterns for travel demand forecasting. Transportation 27: 25–51.
13. Kuppam A, Pendyala R (2001) A structural equations analysis of commuters' activity and travel patterns. Transportation 28: 33–54.
14. Liao Z, Yang S, Liang J (2010) Detection of Abnormal Crowd Distribution. In: IEEE/ACM International Conference on Green Computing and Communications & IEEE/ACM International Conferenceon Cyber, Physical and Social Computing IEEE. pp 600–604.
15. Candia J, González M, Wang P, Schoenharl T, Madey G, et al. (2008) Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical 41: 224015.
16. Andrade E, Blunsden S, Fisher R (2006) Modelling crowd scenes for event detection. In: Proceedings of the 18th International Conference on Pattern Recognition IEEE, volume 1. pp 175–178.
17. Mehran R, Oyama A, Shah M (2009) Abnormal crowd behavior detection using social force model. In: IEEE Conference on Computer Vision and Pattern Recognition IEEE. pp 935–942.
18. Handy S (1996) Methodologies for exploring the link between urban form and travel behavior. Transportation Research Part D: Transport and Environment 1: 151–165.
19. Horner M, O'Kelly M (2001) Embedding economies of scale concepts for hub network design. Journal of Transport Geography 9: 255–265.
20. Dieleman F, Dijst M, Burghouwt G (2002) Urban form and travel behaviour: micro-level household attributes and residential context. Urban Studies 39: 507.
21. Waddell P (2002) Modeling urban development for land use, transportation, and environmental planning. Journal of the American Planning Association 68: 297–314.
22. Boarnet M, Crane R (2001) The influence of land use on travel behavior: specification and estimation strategies. Transportation Research Part A: Policy and Practice 35: 823–845.
23. Wegener M (2004) Overview of land use transport models. Handbook of transport geography and spatial systems 5: 127–146.
24. Handy S (2005) Smart growth and the transportation-land use connection: what does the research tell us? International Regional Science Review 28: 146.
25. Han X, Hao Q, Wang B, Zhou T (2011) Origin of the scaling law in human mobility: Hierarchy of traffic systems. Physical Review E 83: 036117.
26. Longini I Jr., Nizam A, Xu S, Ungchusak K, Hanshaoworakul W, et al. (2005) Containing pandemic influenza at the source. Science 309: 1083.
27. Eubank S, Guclu H, Kumar V, Marathe M, Srinivasan A, et al. (2004) Modelling disease outbreaks in realistic urban social networks. Nature 429: 180–184.
28. Easley D, Kleinberg J (2010) Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge University Press.
29. Anderson R, Fraser C, Ghani A, Donnelly C, Riley S, et al. (2004) Epidemiology, transmission dynamics and control of SARS: the 2002–2003 epidemic. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 359: 1091.
30. Hufnagel L, Brockmann D, Geisel T (2004) Forecast and control of epidemics in a globalized world. Proceedings of the National Academy of Sciences of the United States of America 101: 15124.
31. Riley S (2007) Large-scale spatial-transmission models of infectious disease. Science 316: 1298.
32. Kleinberg J (2007) The wireless epidemic. Nature 449: 287–288.
33. Hu H, Myers S, Colizza V, Vespignani A (2009) WiFi networks and malware epidemiology. Proceedings of the National Academy of Sciences 106: 1318.
34. Castellano C, Fortunato S, Loreto V (2009) Statistical physics of social dynamics. Reviews of modern physics 81: 591–646.
35. Shlesinger M, Zaslavsky G, Frisch U (1995) Lévy flights and related topics in physics. In: Lévy Flights and Related Topics in Physics: Proceedings of the International Workshop Held at Nice, France. volume 450.
36. Rhee I, Shin M, Hong S, Lee K, Chong S (2008) On the levy-walk nature of human mobility. In: INFOCOM 2008 The 27th Conference on Computer Communications. IEEE. IEEE. pp 924–932.
37. Song C, Qu Z, Blumm N, Barabasi A (2010) Limits of predictability in human mobility. Science 327: 1018.
38. Liang X, Zheng X, Lv W, Zhu T, Xu K (2012) The scaling of human mobility by taxis is exponential. Physica A: Statistical Mechanics and its Applications 391: 2135–2144.
39. Shanghai Jiao Tong University, China (2007) SUVnet-Trace data. Available: http://wirelesslab.sjtu.edu.cn. Accessed 2012 Mar 9.
40. Shanghai Population and Family Planning Commission, China (2001) From the fifth population census to evaluate the population condition for the sustainable development of Shanghai. Available: http://www.popinfo.gov.cn/yearbook/2001nj/zhuanwen/7-4.htm. Accessed 2012 Mar 9.
41. Lee D, Seung H (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401: 788–791.
42. Lin C (2007) Projected gradient methods for nonnegative matrix factorization. Neural computation 19: 2756–2779.
43. Hollick M, Krop T, Schmitt J, Huth H, Steinmetz R (2004) Modeling mobility and workload for wireless metropolitan area networks. Computer Communications 27: 751–761.
44. Ben-Akiva M, Bowman J, Ramming S, Walker J (1998) Behavioral realism in urban transportation planning models. Transportation Models in the Policy-Making Process: Uses, Misuses and Lessons for the Future. pp 4–6.
45. Zhang L, Wu J, Zhen Y, Shu J (2004) A GIS-based gradient analysis of urban landscape pattern of Shanghai metropolitan area, China. Landscape and Urban Planning 69: 1–16.
46. Onnela J, Saramäki J, Hyvönen J, Szabó G, Lazer D, et al. (2007) Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences 104: 7332.
47. Shanghai Municipal Bureau of Planning and Land Resources, China (2009) Shanghai urban planning: land-use planning. Available: http://www.china.com.cn/aboutchina/zhuanti/09dfgl/2009-09/08/content184882372.htm. Accessed 2012 Mar 9.
48. Kaltenbrunner A, Meza R, Grivolla J, Codina J, Banchs R (2008) Bicycle cycles and mobility patterns-Exploring and characterizing data from a community bicycle program. Arxiv preprint arXiv: 08104187.