How Far Would You Go?
Comparing Urban and Spatial Access in 10 Global Cities

Sahil Loomba and Matthew Garrod
Department of Mathematics, Imperial College London
(Dated: May 25, 2019)

Cities permit people to access a diverse range of venues and attractions with relative ease. A range of factors–spatial, behavioural and categorical–can determine how successfully they do so. This work develops principled measures of urban and spatial access around these factors using ideas from statistics, networks and topology. These measures are validated by using data on trip check-ins from the location-based social network Foursquare, on 10 cities spread across the globe. Correlation analysis reveals that people travel further for venue diversity, and for venue types they have a high affinity for. Consequently, there is a trade-off between local venue diversity and global venue popularity. This model-driven data analysis paves the way for future work on understanding how to quantify the diversity of resources that individuals have access to within a city, and help city planners to provide good urban access.

I. INTRODUCTION

More than half of the world's population now live in cities, and this number is only set to rise in years to come. Construction of integrated transport, introduction of novel technologies, faster and farther movement of diverse people within and between countries, mixing of cultures, rising complexity of socio-economic interactions, will all push us towards an era of global planning culture. This realisation provides an interesting challenge to urban planners, of how to best plan a city that works for its people in this complex and dynamic urban environment. Particularly, in terms of providing better access to a diverse set of facilities–such as food, health, education, entertainment, and mobility, amongst others–that improve peoples overall standards and satisfaction of living.

In this work we aim to define and assess the accessibility of 10 different cities spread across the globe. We hypothesise a city to be more *accessible* if it is possible for people to have access to a diverse set of public venues, that satisfy different dimensions of societal existence. Although this is an intuitive hypothesis, we seek to prove or disprove it, provide measurable statistics that can diagnose the accessibility of cities, and assist city-planners to develop cities that people want to live in.

The rich Foursquare dataset allows us to obtain three different classes of information:

Behavioural: This pertains to the trips that users choose to make. For example, users may make trips between venues of different categories, say between "restaurants" and "metro stations".

Categorical: This pertains to the distribution of different venues across the city. We may be interested in both the pure count and spatial distribution of different categories and also how individuals move between these different classes of venue.

Spatial: Data which contains information such as the trip distance or the distribution of venue categories within certain regions.

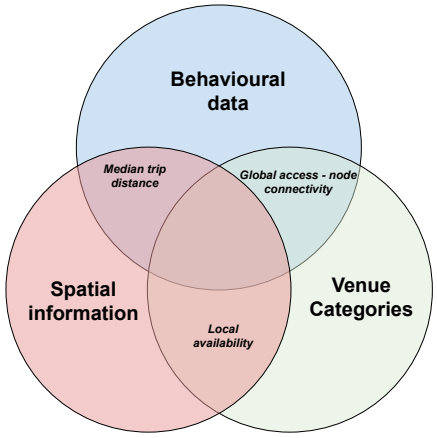


FIG. 1: Schematic illustrating the different classes of data which three statistics: (i) median trip distance, (ii) local availability $a_i(R)$, and (iii) global access - node connectivity $\hat{\beta}_0$, belong to. Each statistic covers two of the different data classes.

Previous studies of data from Foursquare suggest that human mobility patterns show some level of universality across different cities [1]. Consequently, testing for correlations between behavioural, categorical and spatial information should provide insight into both the accessibility of different venue categories and of people's preferences for different venue categories. The Foursquare dataset has rich temporal information on venue check-ins across 10 cities, and the category of each venue. We exclude temporal analysis from the present scope of our study, and make use of venue coordinates, check-in and category information only. A detailed data description is provided in the appendix (V A).

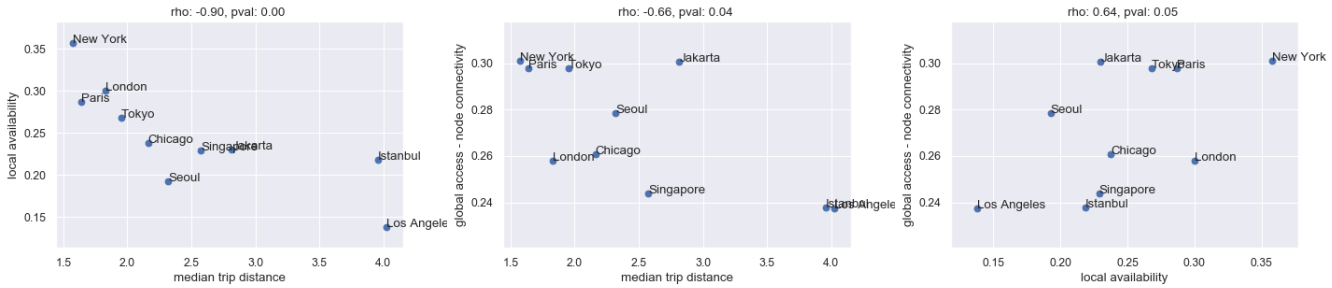


FIG. 2: Scatter Plots relating (a) median trip distance to local availability $a_i(R)$, (b) median trip distance to global access connectivity $\hat{\beta}_0$, and (c) $a_i(R)$ with $\hat{\beta}_0$.

II. SPATIAL ACCESS STATISTICS

We consider the following spatially dependent measures of the accessibility of each city.

1. The mean, median and maximum trip distances obtained by taking a sample of 5000 trips from each dataset.

2. Let $N_{\text{Tot}}(i, R)$ be the number of different venue categories within a radius R of venue i. Let $N_{\text{Vis}}(i, R)$ be the number of venue categories, within a radius R, that users visit. Let k be the number of different venue categories in the city of interest. We define the *local availability* of venue i to be:

$$a_i(R) = \frac{N_{\text{Tot}}(i, R)}{k}. \tag{1}$$

This statistic is purely mechanistic or spatial, and captures the availability of different venue categories within the vicinity of venue i. Now, incorporating trips undertaken by people, we consider two more statistics: $\mathcal{A}_i(R) = \frac{N_{\text{Vis}}(i,R)}{N_{\text{Tot}}(i,R)}$ and $\mathcal{A}_i^G(R) = \frac{N_{\text{Vis}}(i,R)}{k}$, which we will refer to as the local and global *spatial access* respectively. Note that $\mathcal{A}_i^G(R) = \mathcal{A}_i(R)a_i(R)$. These represent the degree to which users *make use of* the range of different venue categories in their immediate vicinity, versus in the entire city.

The first of these measures combines spatial and behavioural information, while the second, $a_i(R)$ contains spatial and categorical information but no behavioural information. Figure 1 shows the overlap of the median trip distance and $a_i(R)$ in terms of the information they include. Also shown is the overlap with the $\hat{\beta}_0$ statistic defined in section III.

The full motivation and derivation of these measures is described in the appendix (see section V B). In the following analysis we will use values of the above statistics estimated by sampling 100 random venues from each city. We will fix R = 1km as this is representative of the typical distance that an individual might walk from a particular venue without resorting to public transport. A more detailed investigation of the effect of varying R on the metrics discussed above is beyond the scope of this preliminary study.

III. URBAN ACCESS STATISTICS

A spatial-statistics view of venue interaction networks can see graphs between nodes embedded in a Euclidean 2D space, corresponding to the physical space of venue interactions. An alternative view is a category-based view, wherein n venues are "generated" from k categories following a certain multinomial distribution $\boldsymbol{\pi} \in \{0, 1\}^k$. Then, interaction edges A_{ij} are generated between any two venues from categories i and j, from a block or "affinity"matrix Ψ such that $a_{ij} \sim \text{Bernoulli}(\Psi_{ij})$. Succinctly, if $z_i \in [0, 1]^k$ be the assignment vector where $z_{ij} = 1$ iff venue i is in venue-category j, and $Z \in [0, 1]^{n\times k}$ be the assignment matrix for n venues, then

$$\begin{aligned} z_i &\sim \text{Multinomial}(\boldsymbol{\pi}) \\ A_{ij} &\sim \text{Bernoulli}(Z\Psi Z^T) \end{aligned} \tag{2}$$

This resembles the framework of stochastic block models, wherein categories are usually latent variables to be inferred [2], although here we treat them as the given observed venue-categories. It is easy to see the generated network depends entirely on the distribution of venue-categories $\boldsymbol{\pi}$ ("category information"), and the inter-category affinities encoded by Ψ ("behavioural information"). This purely categorical view might initially seem at odds with the idea of spatial networks, where a certain assumption of spatial homophily seems applicable. But we wanted to see if other forms of purely categorical homophily and access can explain away spatial homophily and access. In effect, we wanted to establish metrics based purely on $\boldsymbol{\pi}, \Psi$ that can capture the spatial access statistics defined above.

A categorical view can have advantages in terms of better city planning: it's easier to control for the distribution

of categories than it is to control the distances between them. To see this better, let us define these parameters for our model. Let C_{ij} be the (possibly asymmetric) trip-count matrix which encodes the number of trips from a venue of category i to j. (And sum of trips from a particular category i be $C_i = \sum_j C_{ij}$.) Now trips between two categories can be high for two reasons: either there are too many venues of those two categories, or people indeed preferentially travel between venues of these two categories:

$$\begin{aligned} C_{ij} &\propto \pi_i \pi_j \Psi_{ij} \\ C_{ij} &= \mathrm{tr}(C^T C)\pi_i \pi_j \Psi_{ij} \\ \implies \Psi_{ij} &= \frac{C_{ij}}{\mathrm{tr}(C^T C)\pi_i \pi_j} \end{aligned} \tag{3}$$

Note that we use the sum of elements of the count matrix $\mathrm{tr}(C^T C)$ as a normalizing constant, to account for differences in usage of Foursquare across different cities. In this manner, we have decoupled the reasons for observing a certain trip-count C_{ij} into two orthogonal parameters. For city planners, this provides a way to provide good urban access (loosely defined here in terms of encouraging more trips): increase the number of venues of categories where people have high affinity.

Correspondingly, we define two types of metrics for measuring and comprehending urban access of a given city:

- **Local Access:** This refers to *statistical moments* of affinity between pairs of venue categories.

  1. Variance in Absolute Affinity: a smaller value implies similar affinities between many venue categories, that is, people "use-up" the available venue categories.

$$\Phi_\sigma = \left\{ \sum_i \left(\frac{C_i}{\mathrm{tr}(C^T C)} \right)^2 \frac{1}{\pi_i} \right\} - 1$$

  2. Mean of Relative Affinity: a smaller value implies people have more affinity to travel between venues of different categories than of the same, thus encouraging category mixing.

$$\widetilde{\Phi}_\mu = -\sum_{i,j} \pi_i \pi_j \log \frac{C_{ij} C_{ji}}{C_{ii} C_{jj}}$$

- **Global Access:** This refers to *topological measures* of "reachability" amongst all venue categories. They are global because of an ordering they impose on pair of communities, by the extent of pairwise affinity between them.

  1. Connectivity: $\hat{\beta}_0$ is area under the Betti-0 curve; a smaller value implies venue categories which people have a high affinity to travel between are more abundant in the city. **That is, smaller value implies higher connectivity.**

  2. Edge Density: $\hat{\beta}_1$ is area under the Betti-1 curve; a larger value implies venue categories which people have a high affinity to travel between are (significantly) more abundant in the city. **That is, larger value implies higher edge density.**

For a full description, derivation, motivation and relationship between these statistics, see the appendix (V C).

IV. COMPARING SPATIAL & URBAN ACCESS

To see if there exists a relationship between spatial and urban access, the 4 urban access statistics $\Phi_\sigma, \widetilde{\Phi}_\mu, \hat{\beta}_0, \hat{\beta}_1$ were computed for all 10 cities in the dataset, and their Spearman correlations to the 8 spatial access statistics were estimated. To check if a purely fundamental statistic could correlate with spatial access, spatial access measures were correlated with 2 more candidate urban statistics: (a) number of categories (k) and (b) dispersion of distribution of venue categories ($\sum_i \pi_i^2$)–a purely category-distribution based urban access statistic. We plot the full correlation matrix in the appendix section V D in figure 7, but below we describe the key result of this analysis, as elucidated in figure 3, 2.

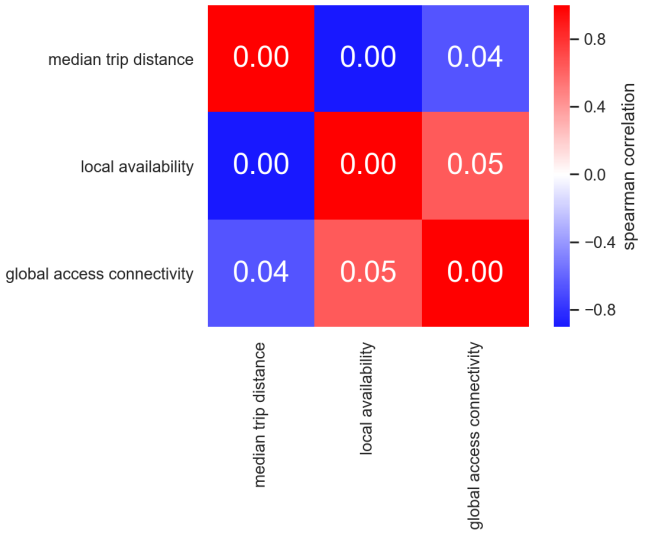


FIG. 3: Spearman Correlation Coefficient between select Spatial and Urban access statistics. Considering p-value≤ 0.05 as significant, median trip distance correlates with local availability $a_i(R)$, and global access–node connectivity $\hat{\beta}_0$.

- *Median trip distance significantly negatively correlates with local availability $a_i(R)$:* this captures the intersection of human behaviour and city space, that of spatial homophily. People are willing to travel longer distances when enough venue categories are not locally available.

To make this more concrete, consider the two cities with the largest and smallest median trip distances respectively: Los Angeles and New York. New York was found to have a large local availability, meaning that a large fraction of the possible categories ($\approx 35\%$) can be found within 1km of a randomly chosen venue, whereas this number is closer to ($\approx 15\%$) for Los Angeles. This is intuitive, as users are less likely to have to travel larger distances if a large range of different venue categories are locally accessible.

- *Median trip distance significantly negatively correlate with global access connectivity $\hat{\beta}_0$:* this captures the intersection of human behaviour and venue categories. Cities which offer more venues of categories that people have a high affinity for are the ones where people are willing (or made) to travel longer distances. One could intuit this is simply because commuting categories such as "metro stations" would enjoy a high affinity, and they would be further apart in bigger cities thus causing this correlation. However, when we plot the connection-events corresponding to "metro stations" for New York and Los Angeles (see figure 6), which have the largest and smallest areas under Betti-0, and the smallest and largest median trip distance, we do not see those events alone causing the significant decline in the curve for either of the cities. **Indeed, other non-commuting venue categories are contributing to this effect.**

- *Local availability $a_i(R)$ significantly positively correlates with global access connectivity $\hat{\beta}_0$:* this captures the intersection of city space and venue categories. Cities which offer more venues of categories that people have a high affinity for can only do so at the expense of making some venues less locally available. For example, Los Angeles has a large $\hat{\beta}_0$, meaning that it provides its residents with more venues of those categories that they have higher affinity to. However, the average $a_i(R)$ is relatively small meaning that the area surrounding any particular venue is less likely to have a diverse range of venue categories. This seems to suggest a **fundamental trade-off in urban planning**: between keeping venues diverse and keeping popular venue categories plentiful. Looking at figure 2, we surmise that London best provides the sweet-spot of local venue diversity and global venue popularity.

V. DISCUSSION

The aim of this study was to explore how the accessibility of different venue categories within different cities interplays with users' behavioural preferences. We have shown that metrics based on behavioural, spatial and categorical information are inherently related. People prefer venue diversity, and are willing to travel longer distances for it. Cities, on the other hand, have to play the balance of keeping city regions locally diverse, while providing more venues of categories that people have a high affinity for.

Methodological advantages. Comparing the statistics defined in sections V B and V C gives us complimentary ways for urban planners to understand the interactions between different venue categories at both a local and a global level. Both the venue-category graph (see section V B 2) and the inter-category affinity matrix, Ψ, (section V C) are amenable to treatment using the vast array of tools from modern network analysis [3]. For example, performing community detection on the weighted venue interaction graph defined in section V B 3 allows us to explore groups of venues which perform the same "function" in a city.

Shortcomings. The correlations observed in Figure 7 were observed on the basis of 10 data points. The 10 different cities were spread across different continents so we might expect some level of robustness to variability caused by cultural and geographical factors. Nonetheless, to be more confident about our conclusions it would be necessary to: (a) consider a much larger database of cities, or (b) compute these statistics at a higher spatial resolution (e.g at the level of neighbourhoods within cities).

When studying both the urban and spatial access statistics we have yet to contrast our findings with a suitable null model. For example, in the urban access statistic we might make a comparison with the data we would obtain if users travelled purely at random between categories while keeping the same number of overall trips. Furthermore, in the case of spatial access we might need to explore further the heterogeneity in distribution of venue categories to provide a more detailed interpretation of our results.

In our construction of the venue-category graph in V B we only consider a relatively small fraction of the total number of venues. Consequently, the graphs obtained may not give us a substantial picture of the full richness of venue category interactions. In addition, we have considered only a single value of radius R in the construction of the venue-category graph. Preliminary simulations suggest that the results are relatively stable to different choices of R, however, a full investigation of this effect is beyond the scope of this work.

Future directions. It will be of interest to compare our analysis with existing analysis of urban mobility using tools from network science [4], and to see if current principles of city planning can be rediscovered, or new ones propounded, using our work. Another intriguing prospect would be to connect cities' urban access with people's social access and health. Studying how one shapes the other will truly help us design cities of the future.

ACKNOWLEDGEMENTS

Thanks to Nick Jones and Antonia Godoy-Lorite for comments and discussions.

[1] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo, PloS one **7**, e37027 (2012).
[2] B. Karrer and M. E. Newman, Physical review E **83**, 016107 (2011).
[3] M. Newman, *Networks: an introduction* (Oxford university press, 2010).
[4] O. J. Sagarra Pascual, Non-binary maximum entropy network ensembles and their application to the study of urban mobility (PhD thesis, Universitat de Barcelona) (2016).
[5] G. Carlsson, Bulletin of the American Mathematical Society **46**, 255 (2009).

APPENDIX

A. Data Description

We use data about Foursquare check-ins for 10 different cities. The data used consists of two parts:

1. **Venue information.** For each venue in the table we have: venue id, name, latitude, longitude and a category.

2. **Movement information.** For each journey we have: id, id_2, Data, Period (MORNING, MIDDAY, AFTERNOON, NIGHT, OVERNIGHT), Number of check-ins (this is the number of check-in pairs).

Merging the venue and movement information allows us to study the distribution of trips made between different venue categories. In this analysis we will ignore the period and number of check-ins between venue pairs as we are focused on the overall of diversity of different venues that people are visiting. Some statistics about the data are shown in table I.

B. Quantifying Spatial Access

1. *Typical Trip Distances*

A 'basic' metric which captures the spatial access in a given city is the typical distance of trips made in the dataset. We estimate the distribution of trips by taking a sample of 5000 trips from the venue in each case. Estimates of the median trip distance based on this subsample are shown in Table I. In the analysis below we will also consider the maximum distance between trips estimated from the same sample. This gives us a proxy for the typical size of a city.

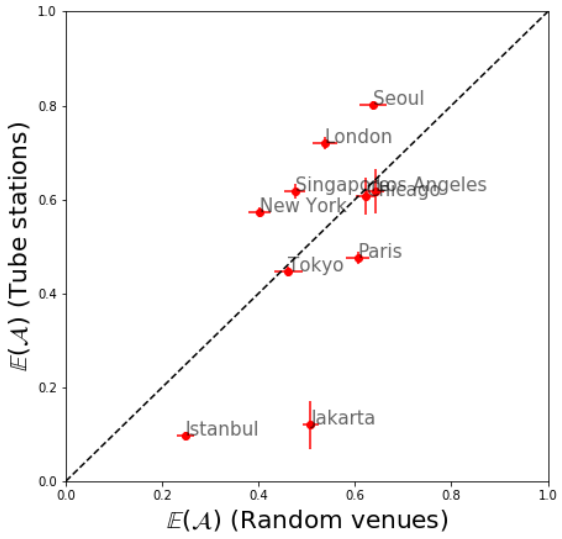


FIG. 4: Plot showing average value of $\mathcal{A}_i(R)$ for 100 random venues vs. all the metro stations in each of the 10 cities. The error bars represent the standard error on the mean. The Pearson correlation coefficient is 0.63 with a p-value of 0.049. The line y=x is shown as a guide for the eye. Shown for the case where $R = 1$km

2. *Assessing Local Diversity of Category Types*

Let $\mathcal{V}$ be the set of venues and $\mathcal{C}$ be the set of categories. We are interested in the diversity of venue categories *locally* accessible to users across different cities. Consequently, for a given venue i we will consider the set of venue categories which lie within R km of i. Let $N_{\rm Tot}(i, R)$ be the total number of venue categories within distance R from i.

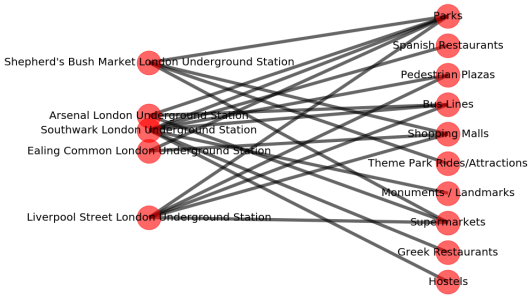
Let $\mathcal{C}(i, R)$ be the set of venue categories within distance R of venue i. We will define a category $K \in \mathcal{C}(i, R)$ as being *visited* if at least one trip originating from i terminates at a venue of category, K. From this we can define $N_{\rm Vis}(i, R) \leq N_{\rm Tot}(i, R)$ as the number of categories visited from i[1].

We can form a bipartite graph consisting of venues and categories. Form an edge between $i \in \mathcal{V}$ and $K \in \mathcal{C}$ if:

[1] In this analysis we only consider the journeys originating at the venue i rather than those starting at venues within its vicinity.

City	Number of venues	Number of categories	Number of metro stations	Number of trips	Median trip distance (km)
Chicago	13904	501	28	7775376	2.26
Paris	13588	464	295	7574139	1.63
Singapore	23324	521	81	7723757	2.68
Istanbul	113752	670	104	7372799	4.09
Tokyo	57810	592	366	7798240	1.87
New York	32971	603	282	7805871	1.48
London	22689	530	241	7650994	1.84
Los Angeles	15868	513	17	7721731	3.84
Jakarta	21813	469	4	7801368	2.75
Seoul	15545	433	283	7768926	2.41

TABLE I: Information about the number of venues and categories for each of the 10 cities in the dataset. Median trip distances are estimated from a random sample of 5000 trips from each city.



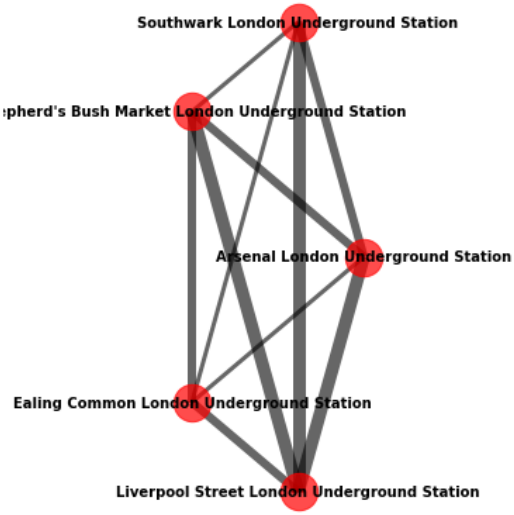


FIG. 5: Bipartite venue category graph for 5 underground stations in London and the corresponding weighted venue-venue graph.

1. $K \in \mathcal{C}(i, R)$

2. and K is visited at least once from i[2].

We denote the set of visited categories by: $\mathcal{C}_V(i, R) \subseteq \mathcal{C}(i, R)$. Figure 5 illustrates an example of the venue category graph for 5 underground stations in London.

In this bipartite graph, the degree of a venue i, $\kappa_i = N_{\text{Vis}}(i, R)$. Define the *local spatial access* of a venue, i, to be:

$$\mathcal{A}_i(R) = \frac{N_{\text{Vis}}(i, R)}{N_{\text{Tot}}(i, R)}. \tag{4}$$

This measure informs us about the number of different categories that individuals choose to visit out of the number of possible categories. It therefore, servers as a proxy for the accessibility and the quality of the venues within the vicinity of i.

If $\mathcal{A}_i(R) = 1$ then individuals visit all the possible categories out of the available categories, whereas if $\mathcal{A}_i(R) \approx 0$ then users are likely to visit few different categories.

We compute the distribution of $\mathcal{A}_i(R)$ values across the 10 cities for:

1. The venues in the category 'metro station'. We note that some cities, such as Jakarta, only have a relatively small number of metro stations included in the dataset. Consequently, the comparison of the accessibility of metro stations to that of other cities may not be fair.

2. A sample of 100 randomly selected venues.

[2] A more robust way of associating venues with categories might be obtained by ensuring that a set threshold number of trips occur between the specified venue and category pair within the threshold distance. However, we do not consider this in our initial analysis

Figure 4 shows the mean values of $\mathcal{A}_i(R)$ for random venues plotted against the value for metro stations for the 10 cities. The correlation between the two indicates that the accessibility of metro stations in a given city has a similar mean distribution to that of venues in general.

3. Reduction of the Bipartite Graph

We can project the bipartite graph considered in the preceding section onto the set of venues $\mathcal{V}$ in order to obtain a new weighted graph. This graph can be represented by a weighted adjacency matrix, W, where the edge weights:

$$W_{ij} = |\mathcal{C}_V(i, R) \cap \mathcal{C}_V(j, R)|, \tag{5}$$

represent the number of common categories which users have chosen to visit from the venues i and j (within a radius R).

We consider the following graph statistics:

- **Mean weighted degree.** The weighted degree of a node represents the total number of shared categories with other nodes. If the projected graph has a large mean weighted degree then venues have i) a large number of visited categories which are ii) shared with many other venues.

- **Average weighted clustering coefficient.** We can interpret this metric as follows: If venue A shares many venues with B and C then venues B and C are also likely to share a large number of venues.

Note that since we have fixed R at a specific value that these statistics tell us about the local similarity between randomly chosen venues.

C. Quantifying Urban Access

1. Local Access

This refers to a measure of affinity between pairs of venue-categories, $\Psi_{ij}, \forall i, j \in \{1, 2, \dots k\}$, which can be averaged over to obtain overall pairwise-affinities enjoyed by a given category: $\phi_i, \forall i \in \{1, 2, \dots k\}$. This can be further aggregated to obtain the mean and variance in the overall affinity enjoyed by the venue-categories of a given city: Φ_μ, Φ_σ. Mathematically, there are various ways to define these ideas concretely. We use the simplest and most interpretable formulations to write in matrix notation:

$$\begin{aligned}
\boldsymbol{\phi} &= \Psi\boldsymbol{\pi} \\
\Phi_\mu &= \mathbb{E}_{\boldsymbol{\pi}}[\boldsymbol{\phi}] = \boldsymbol{\pi}^T\boldsymbol{\phi} = \boldsymbol{\pi}^T\Psi\boldsymbol{\pi} = 1 \text{ (from eq 3)} \\
\Phi_\sigma &= \mathbb{E}_{\boldsymbol{\pi}}[(\boldsymbol{\phi} - \Phi_\mu)^T(\boldsymbol{\phi} - \Phi_\mu)] \\
&= \boldsymbol{\pi}^T \text{diag}((\boldsymbol{\phi} - \Phi_\mu)(\boldsymbol{\phi} - \Phi_\mu)^T) \\
&= \boldsymbol{\pi}^T \text{diag}((\Psi\boldsymbol{\pi} - \boldsymbol{\pi}^T\Psi\boldsymbol{\pi})(\Psi\boldsymbol{\pi} - \boldsymbol{\pi}^T\Psi\boldsymbol{\pi})^T) \\
&= \left\{ \sum_i \left(\frac{C_i}{\text{tr}(C^TC)} \right)^2 \frac{1}{\pi_i} \right\} - 1
\end{aligned} \tag{6}$$

Due to the normalisation we have performed, $\Phi_\mu = 1$ for all cities. But if we hadn't, it would have captured total number of trips, an intuitive measure of urban access. Φ_σ on the other hand captures variance in trip-counts across categories. For a null model where all k categories are equally distributed and have the same trip counts, it would amount to exactly 0. These local access measures capture notions of **absolute** affinity. Another interesting affinity to look at would be **relative**, with regards to affinity of a category to itself. This would directly capture the idea of categorical homophily. One simple way of doing this is to take a negative-log-ratio of the affinity matrix Ψ with respect to its diagonal elements. Correspondingly, let us define a new relative affinity matrix $\widetilde{\Psi}$ such that $\widetilde{\Psi}_{ij} = -\log\left(\frac{\Psi_{ij}}{\Psi_{ii}}\right)$. Clearly, if people take trips between venues of same categories with the same affinity as to other categories ("ambiphily"), this value is 0, positive if they prefer trips between venues of same category ("homophily") and negative otherwise ("heterophily"). Definitions for relative local access, $\widetilde{\Phi}_\mu$, $\widetilde{\Phi}_\sigma$, analogously follow. Specifically, we highlight:

$$\begin{aligned}
\widetilde{\Phi}_\mu &= -\sum_{(i,j)\in S} \pi_i\pi_j \log \frac{\Psi_{ij}\Psi_{ji}}{\Psi_{ii}\Psi_{jj}} \\
&= -\sum_{(i,j)\in S} \pi_i\pi_j \log \frac{C_{ij}C_{ji}}{C_{ii}C_{jj}}
\end{aligned} \tag{7}$$

where set S is set of all category-pairs i, j such that $i < j$. (Note that a 0 trip-count between any two venue categories will lead to an improper value of this statistic. We get around this problem by finding a good approximation $\widehat{\Psi}$ of the Ψ matrix not containing any 0 entries, whose statistic can be computed instead. This is further described in section V C 3.) It is easy to see that in the null model where every category has same number of trips to another category as it does to itself (agnostic "category mixing"), this value is exactly 0, irrespective of distribution of categories itself. A negative (positive) value encourages more (less) mixing. Say $\log \frac{C_{ij}C_{ji}}{C_{ii}C_{jj}}$ is some constant 2κ, then $\widetilde{\Phi}_\mu = -\kappa\left(1 - \sum_i \pi_i^2\right)$. The second factor is a positive number that measures dispersion in the distribution of venue categories, with a small value indicating few categories in a disproportionate balance, and a large value indicating many categories

in proportionate balance. For "heterophilous" networks, $\kappa > 0$ which leads to good urban access, further bettered by many equiproportioned categories. For "homophilous networks" $\kappa < 0$ which leads to poor urban access, further worsened by many equiproportioned categories.

2. Global Access

This refers to overall measures of "reachability" amongst all categories in a network. While most trips can be considered to occur in isolation, people typically pass through multiple venues, spread across venue categories, in a typical trip across the city. Thus, a more global measure of access should capture the ease of performing multiple and diverse trips in succession. As before, there can be multiple mathematical frameworks to capture this idea. We define our access statistic inspired from topological data analysis [5], in a manner that's easily interpretable.

Since we have defined a probabilistic model, Ψ_{ij} can be treated as an activation threshold over which edges are added between venues of category i and j. Correspondingly, we can define a Vietoris-Rips complex for the venue-interaction network wherein the distance between venues of two communities is given by $\delta_{ij} = -(\Psi_{ij}+\Psi_{ji})$. (Although distances cannot be negative, we make use of the negative merely to demonstrate that categories with higher pairwise affinities between them are closer in this metric space than those with lower affinities.) Now given some activation threshold $\epsilon \in [\min(\delta_{ij}), \max(\delta_{ij})]$, we can define a filtration of simplicial complexes to obtain a "barcode" for the family of venue-interaction networks generated as a sample of the probabilistic model. We only consider simplicial 1-complex, that is, 0-d (nodes) and 1-d (edges) complexes. More precisely, given a network with n venues, this barcode is completely defined by n, $\boldsymbol{\pi}$ and Ψ. While Ψ determines the threshold ϵ, n and $\boldsymbol{\pi}$ determine the number of edges added. Correspondingly, we can obtain the Betti curves for 0-d and 1-d homologies, wherein the Betti numbers are plotted against ϵ. Betti-0 at given ϵ, $\beta_0(\epsilon)$, corresponds to number of connected components $k(\epsilon)$. Since we consider only upto 1-d complexes, Betti-1 at given ϵ, $\beta_1(\epsilon)$, corresponds to $n-k(\epsilon)+m(\epsilon)$, where $m(\epsilon)$ is the number of edges added up to a certain threshold. Since we do not wish for n itself to affect the analysis, we can normalize β_0 by n and β_1 by n^2 and assume $n \to \infty$. Under this asymptotic assumption, β_1 simply corresponds to the edge density, and β_0 to the density of connected components. Note that since Ψ_{ij}s follow a log-normal distribution for this dataset, we logarithmically transform them before defining the distance, that is $\delta_{ij} = -\log(\Psi_{ij}\Psi_{ji})$. Also, to make things comparable between cities, we can normalise ϵ to lie in $[0, 1]$. We plot these curves for the ten cities in Figure 2. The β_0 curves start at 1, where every node is its own connected component, and then monotonically decrease upto 0 where we have 1 large connected component. Similarly, the β_1 curves start at 0, when no edges are added, and monotonically increase upto 1 where we have a fully connected network. Correspondingly, we can make use of area under these "normalised" Betti curves as a statistic. We refer to area under the Betti-0 curve ($\hat{\beta}_0$) as **global access - connectivity** and that under Betti-1 curve ($\hat{\beta}_1$) as **global access - edge density**. It is easy to see that a smaller (larger) $\hat{\beta}_0$ ($\hat{\beta}_1$) signifies that venue-categories which people have a high affinity to travel between exist in a larger proportion in the city, a directly interpretable measure of urban access.

The reason we refer to these as global access measures is because the filtration process imposes a global ordering G on pairs of venue-categories, from highest to lowest affinity. In this global ordering, consider consecutive category pairs (p, q) and (r, s) such that $\delta_{pq} < \delta_{rs}$. Consider the (normalised) Betti-1 curve $\beta_1(\epsilon)$ which essentially is the edge density, then between $\epsilon = \delta_{pq}$ and $\epsilon = \delta_{rs}$, the curve gains a height of $\pi_r\pi_s$ over a step size proportional to $\delta_{rs} - \delta_{pq}$. That is, $\beta_1(\delta_{rs}) = \beta_1(\delta_{pq}) + \pi_r\pi_s$. Applying the trapezoid rule, this leads to an increase in area under the Betti-1 curve proportional to $\left(\beta_1(\delta_{pq}) + \frac{\pi_r\pi_s}{2}\right) \log \frac{\Psi_{pq}\Psi_{qp}}{\Psi_{rs}\Psi_{sr}}$. Comparing it to equation 7, we note that ($\hat{\beta}_1$) resembles the mean relative local access in its form, except it sums over a global ordering of category pairs, therefore acting as a "global" access. We omit the expressions for $\hat{\beta}_0$ here for brevity, but it leads to a similar expression composed of sums of π_i, π_j instead of products, with an additional condition that bridging i, j must reduce the number of connected components.

3. Inferring Latent Categories

Consider the category affinity matrix $\Psi \in \mathbb{R}_{\geq 0}^{k\times k}$ and category distribution vector $\boldsymbol{\pi} \in \{0, 1\}^k$, where these k categories are observed, such as "pubs", "metro stations", "parks", etc. However, one can imagine only a few m number of latent binary categories to represent each of these k observed categories in a one-hot encoding of size m. Furthermore, since one-hot encoding implicitly assumes independence of latent categories, one can independently compute affinities of observed categories as product of affinities of latent categories.

To illustrate with an example, consider $k = 8$, and let $m = \text{ceil}(\log_2(k)) = 3$. That is, each of the categories can be represented as bitstrings $000, 001, \ldots 111$ such that $\widehat{\Psi}(b_1^p b_2^p b_3^p, b_1^q b_2^q b_3^q) \propto \Psi_{b_1^p b_1^q}\Psi_{b_2^p b_2^q}\Psi_{b_3^p b_3^q}$ and $\widehat{\pi}(b_1^p b_2^p b_3^p) = \pi_{b_1^p}\pi_{b_2^p}\pi_{b_3^p}$. In general, we make use of the Kronecker product to write in matrix/vector notation $\widehat{\Psi} = \kappa \bigotimes_{p=1}^{m} \Psi^p$ and $\widehat{\boldsymbol{\pi}} = \bigotimes_{p=1}^{m} \boldsymbol{\pi}^p$ where $\Psi^p \in \mathbb{R}^{2\times 2}$ and $\boldsymbol{\pi}^p \in \{0, 1\}^2$, $\forall p$. The problem of inference here is to obtain $\widehat{\Psi}$ as close to Ψ as possible.

We propose a simple and interpretable way to infer Ψ^is via eigendecomposition. Assuming Ψ to be a symmetric matrix, let $Q \in \mathbb{R}^{k\times k}$ be the set of orthogonal eigen-

vectors of Ψ. While the first eigenvector corresponds to the overall "connectivity", second eigenvector onwards we obtain successively orthogonal basis of the matrix which can correspond to the latent categories. Let us pick $Q^\dagger \in \mathbb{R}^{k\times m}$ as the set of top m eigenvectors after the first one, and project Ψ to obtain $\Psi^\dagger = \Psi Q^\dagger$. Now, the pth position in the latent bitstring of observed community i is given simply by $\Psi^\dagger_{ip} > 0$.

Once all k communities are partitioned into disjoint sets $0_p, 1_p$ corresponding to pth latent binary category, $\boldsymbol{\pi}^p$ and Ψ^p can be inferred by summing the appropriate entries in the original vector/matrix. Precisely,

$$\begin{aligned}
\pi^p_0 &= \sum_{i\in 0_p} \pi_i \\
\pi^p_1 &= \sum_{i\in 1_p} \pi_i \\
\Psi^p_{00} &= \frac{\sum_{i,j\in 0_p} C_{ij}}{\mathrm{tr}(C^T C)\pi^p_0\pi^p_0} \\
\Psi^p_{01} &= \frac{\sum_{i\in 0_p, j\in 1_p} C_{ij}}{\mathrm{tr}(C^T C)\pi^p_0\pi^p_1}
\end{aligned} \tag{8}$$

and Ψ^p_{11}, Ψ^p_{10} analogously follow. The whole reason for performing this latent category inference is to obtain a good approximation to Ψ that doesn't contain any 0 entries. Since we have transformed Ψ into an approximate Kronecker product of various 2×2 matrices that are almost surely not going to contain 0 entries, this objective is certainly solved. Moreover, since in estimating relative local access statistic we take the ratio of entries of Ψ, we need not infer the constant κ which gets cancelled out. We can simply pick an m (a free parameter which can be chosen through domain knowledge or using the eigengap heuristic), infer $\widehat{\Psi}$ and find the statistic from it instead. (However, if one really wanted to infer κ, then an extra constraint can be used, such as for $\widehat{\Psi}$ to have the same largest eigenvalue as that of Ψ.)

D. Full Correlations

Figure 7 shows the matrix of correlations between the different spatial and urban access statistics. We observe the following:

- $\mathcal{A}_i(R)$ *significantly negatively correlates with* Φ_σ*:* (Figure 2a) cities which offer high local venue diversity, are the ones where people have similar affinities to different venue categories. This indicates people have a preference for diversity, or else we would have noted a positive correlation between the two, or no correlation if they were agnostic.
- $\mathcal{A}_i(R)$ *significantly negatively correlates with* k: cities which offer high local venue diversity are the ones with fewer number of categories. This is possibly because people's usage of new venue categories scales sub-linearly with the number of categories added.
- *The dispersion of the distribution of venue categories alone does not correlate significantly with any spatial measures.* This suggests that the distribution of venues across categories alone does not determine overall spatial access. We expect this to be the case given the reported universality in mobility patterns between different cities [1].

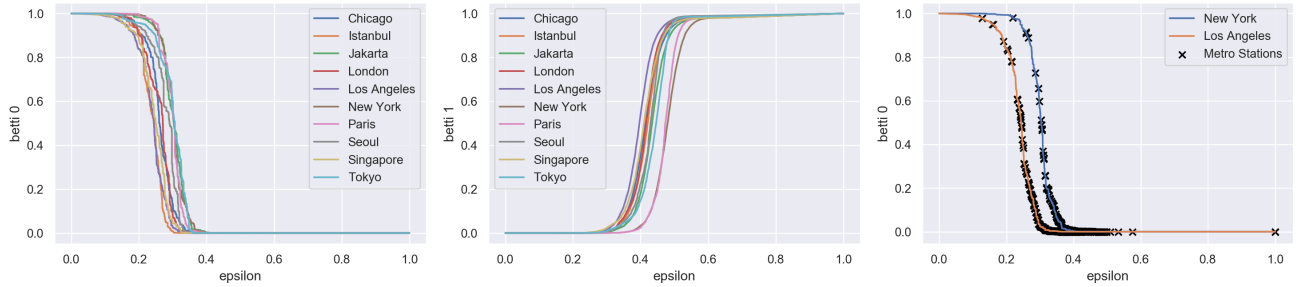


FIG. 6: Normalised Betti curves for (a,b) the ten cities, and (c) Betti-0 curves for New York and Los Angeles with "Metro Station" events marked.

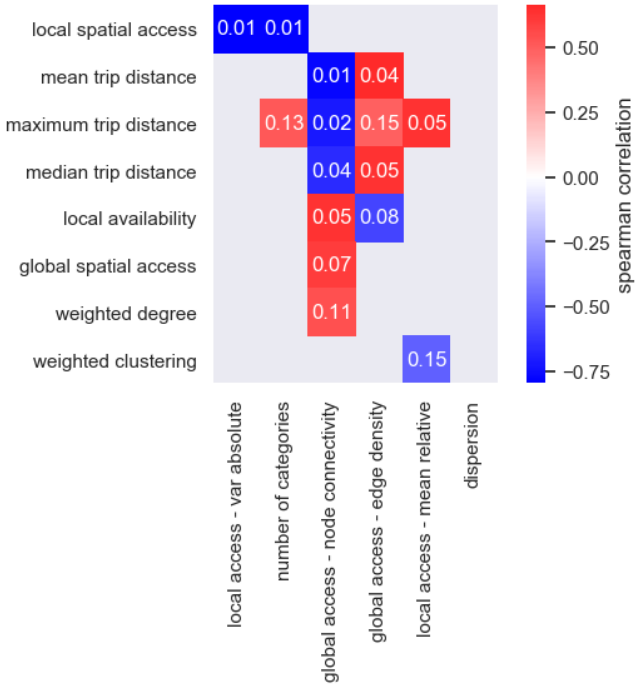


FIG. 7: Spearman Correlation Coefficient between various Spatial and Urban access statistics. Correlations with p-value>0.15 have been masked. Considering p-value≤ 0.05 as significant, Φ_σ correlates with $\mathcal{A}_i(R)$, and global access statistics $\hat{\beta}_0, \hat{\beta}_1$ correlate with trip distances.