Economic outcomes predicted by diversity in cities

Chong et al. *EPJ Data Science* (2020) 9:17
https://doi.org/10.1140/epjds/s13688-020-00234-x

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Economic outcomes predicted by diversity in cities

Shi Kai Chong[1†], Mohsen Bahrami[2,3*†], Hao Chen[4], Selim Balcisoy[5], Burcin Bozkaya[3,6] and Alex 'Sandy' Pentland[1]

*Correspondence:
bahrami@mit.edu
[2]MIT Connection Science, Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, USA
[3]School of Management, Sabanci University, Istanbul, Turkey
Full list of author information is available at the end of the article
†Equal contributors

Abstract

Much recent work has illuminated the growth, innovation, and prosperity of entire cities, but there is relatively less evidence concerning the growth and prosperity of individual neighborhoods. In this paper we show that diversity of amenities within a city neighborhood, computed from openly available points of interest on digital maps, accurately predicts human mobility ("flows") between city neighborhoods and that these flows accurately predict neighborhood economic productivity. Additionally, the diversity of consumption behaviour or the diversity of flows together with geographic centrality and population density accurately predicts neighborhood economic growth, even after controlling for standard factors such as population, etc. We develop our models using geo-located purchase data from Istanbul, and then validate the relationships using openly available data from Beijing and several U.S. cities. Our results suggest that the diversity of goods and services within a city neighborhood is the largest single factor driving both human mobility and economic growth.

Keywords: Economic growth; Urban economy; Diversity; Interaction; Information flow; Consumer city; Huff gravity model

1 Introduction

Cities are known as engines of industry and innovation, and economists largely take the view of cities as production centers [1–3]. Previous studies have illuminated the growth, innovation, and prosperity of entire cities, but the causal processes that produce these results are complex and the causality is unclear. Moreover, while substantial evidence is accumulating concerning the growth and prosperity of cities as a whole, there is less clarity about the factors and processes that determine the prosperity and attractiveness of individual neighborhoods or districts.

This paper provides a simple and practical method for forecasting local neighborhood prosperity that accounts for around half of the variance in economic growth as well as accurately predicting interaction patterns between neighborhoods (e.g., number of workers and shoppers from other areas of the city). The method has been validated on data from three continents, and sheds new light on the causal processes underlying the evolution of city neighborhoods.

Springer

We begin by utilizing open sourced map data and geo-tagged expenditure records from the banking sector, covering a significant portion of the population of Istanbul. Using this data we show that (1) volume of inflow of people into a neighborhood is strongly correlated with the diversity of amenities within the neighborhood as measured by the Shannon entropy of the store categories, (2) the volume of inflow of people into a neighborhood is strongly correlated with the productivity of the neighborhood, and (3) the diversity of amenities within a neighborhood is strongly correlated with the diversity of people entering the neighborhood.

Having developed a model that relates diversity of shopping categories, flows of people, and productivity, we then turn to the important problem of predicting the economic growth of neighborhoods. Here we show that (4) the diversity of shopping categories within a neighborhood accurately predicts the economic growth of the neighborhood during the following year. Moreover, the prediction of economic growth using diversity of shopping categories remains accurate even when we control for population density, housing price, and centrality of the neighborhood. Combining (3) and (4) we also find that the diversity of people entering the neighborhood also predicts economic growth, again even when these other variables are controlled.

Finally, we validate this growth prediction model in other cities on two other continents using publicly available data from the social networking and crowd sourcing websites Yelp, Dianping, and Meituan (Chinese group buy and crowd-sourced review sites, respectively). While it is well known that consumer consumption has a positive effect on economic growth [4–6], our work is novel in that it provides a simple quantitative method using openly available data that results in usefully accurate predictions and accounts for a large proportion of variance in growth rates.

Understanding the relationship between shopping diversity, the volume and diversity of flows of people, and economic productivity and growth can yield insight into a variety of important societal issues and lead to many practical applications such as urban planning. Policy makers and planners who understand the factors that contribute to economic growth in different areas of the city can allocate their resources efficiently in order to make cities grow better and help even out the distribution of wealth across neighborhoods.

While our results do not demonstrate causal processes, they suggest that the diversity of goods and services within a city neighborhood, and the diversity of people flowing into a neighborhood, may be the largest single factor driving human mobility, productivity, and economic growth. These findings suggest that theories of economic growth that emphasize the spread of ideas and opportunities among diverse populations may have stronger causal effects than is generally believed.

2 Background

Immigration of human population from rural to urban areas has affected human lifestyle. Majority of human population now live in urban areas, while only 20% of them were living in cities during the 19th century [7]. This continuous immigration has led to increase of population density in urban areas. The high population density in cities is associated with high air pollution, ease of contagious disease spread, high crime rate, inequality, and segregation [8–10]. Despite all mentioned negative aspects, people continue to move to cities and the cities become denser. Studies show that the cities facilitate economic activities, support industries and production, use resources more efficient, offer better and cheaper transportation, and easier access to difference locations [8–15].

Urban scientists have been trying to define the characteristics of a so-called "great" city for decades. By the advent of new computational methods and availability of big datasets from various industries, researchers have been enabled to validate their hypotheses leveraging large-scale datasets. Jane Jacobs, a famous urban scientist, suggested a number of essential characteristics that promote quality of life in urban area [16]. The main characteristics she introduces are: mixed land use, diversity, and density. Later Sung et al. [17] validated Jacobs' proposed characteristics in the city of Seoul using various surveys. Recently, De Nadai et al. [18] use a large-scale mobile phone dataset and show that Jacobs' proposed conditions hold for Italian cities as well.

Jacobs in her book *The Death and Life of Great American Cities* [16] argues that the density and diversity of population and amenities are associated with level of interaction among citizens. Many research studies have shown that these interactions lead to information exchange, idea creation, innovation, productivity, and more opportunities for citizens [19–25]. Recent studies have shown that the population density and volume of human flow in urban areas are extremely predictive of economic productivity and wealth in those areas [4–6, 26–28].

Glaeser et al. [29] argue that the productivity is not the only attraction of cities, since in some cities the production cost is more than consumption. For example, in some cities rent prices increase faster than wages increase. The study results suggest that the desirability of urban amenities is one of the important reasons for population attraction to the area. This is in line with Jacobs' idea of the cities that offer more and diverse amenities, satisfy the human desire for consumption of a large variety of goods and services and provide a vital urban life for citizens. In this study, inspired by those previous researches, we aim to understand the dynamics of economic growth and prosperity of single neighborhoods in urban areas.

3 Materials and methods

3.1 Data

For this study, we utilize various datasets from different sources, including publicly available datasets, data published by governments and census bureaus, datasets from financial sectors, and data provided by a commercial digital map production company. Detailed descriptions about each dataset used in this study are available in Additional file 1. Moreover, for replication purposes, we made all required data and code available at: github.com/cshikai/Cities).

3.1.1 Points of interest data

The Internet prompted improvements in logistics and supply chain operation, and the omnipresence of online business models means that manufactured goods are national goods that are easily transported. However, cultural goods [30] such as restaurants and theaters are confined to a locality and are representative of the attractiveness of a region as commodities. Therefore, the scope of commodities studied should not be limited to consumer goods. In order to study various types of commodities we use datasets created by a commercial digital mapping company that contains Points of Interest (POIs) found in each neighborhood. The data is collected via data collection vehicles moving around in every country. Moreover, several external datasets provided by local organizations are utilized to enrich the data.

POIs are grouped into twelve types, namely: business centers, community service centers, financial institutes, educational institutes, entertainment places, shopping places, restaurants, hospitals, parks, travel destinations, auto services, and transportation hubs. Further information about this dataset including metadata, is provided in Additional file 1 section and the research GitHub page.

3.1.2 Consumption data

Three datasets are utilized to study urban environments in three different countries from different continents: Turkey, China, and the United States.

Istanbul, Turkey The first dataset is a set of geo-tagged credit card transactions in Istanbul that covers the expenditure of a sample of more than 62,000 customer over a period of 1 year, from July 2014 to June 2015. It contains more than 4.2 million transactions. The transaction records contain hashed customer IDs, transaction amounts, merchants' business categories, and their locations. Customer information dataset includes customers' demographic information such as their age, gender, marital status, job type, education level, income, and their home and work location. This dataset was donated by one of the largest banks in Turkey for this specific scientific study. The number of sampled customers per district shows a 0.817 correlation with the districts' population during the time frame data was collected. Moreover, there is a 0.6115 correlation between the average yearly household income and the average yearly expenditure per credit card customer in each district. Therefore, we contend that the sample is well balanced across the metropolitan of Istanbul and representative.

More details about this dataset are provided in the Additional file 1. In the provided data for replication purposes in the research GitHub page, transactions and demographic features are aggregated at districts level. Thus, there is no information that could potentially be used to identify individual customers. Basic statistics of the dataset are shown in Table 1.

Beijing, China Second is the data on consumer behaviour for people in Beijing, China, which is publicly sourced from the Chinese phone apps Meituan and Dianping. These phone apps are similar to Groupon and Yelp respectively, where users can purchase discount coupons and look up reviews for various amenities in the city. A total of 136,000 deals offered by 6500 food businesses for four months (November–December 2016, and April–May 2017) are considered. A total of two million customers are in this dataset. This dataset is all publicly available and downloaded from Meituan-Dianping website. Dong et al. [31] use the same Dianping dataset to accurately predict socioeconomic attributes of various neighborhoods in several Chinese cities including Beijing. The attributes are daytime and nighttime population, number of firms, and consumption level in those areas. Their study results show that the Dianping dataset is highly informative. No potential information in the Beijing data that can identity individual users.

Table 1 Credit card dataset summary

Data collection time frame	July 1, 2014 to June 30, 2015
Number of transactions	4,254,652
Number of unique customers	62,392
Number of unique merchants	75,448

United States Third is the publicly available Yelp Data Challenge dataset that we use to study the consumption patterns of individuals in the United States. This dataset includes records on check-ins, reviews, and ratings from 654,135 users of 26,149 businesses across 42 different discontinuous urban areas in the United States from 2009 to 2015. Dataset can be downloaded from: https://www.yelp.com/dataset/download and should be filtered by the dates (2009 to 2015) to result in the same data we use here.

3.2 Theoretical framework

3.2.1 A predictive model of human flows

Flow network Understanding the dynamics of economic growth in cities has been a very important research field in the context of urban studies. Experimental results suggest that interactions as a result of human flow to different regions in cities promote productivity. Examples of these theories include the theory of structural holes [19], weak ties [20], the effect of social interaction on economy flourish [21, 22], and the importance of information flow in the Research & Development [23, 24]. Motivated by these research studies [19–25, 32, 33], we propose a network model of human flow and leverage various datasets for this purpose.

Here we use a notion of directed network to show the flows across different regions of a city. In this directed network, region units are the nodes and an edge is established from region i to j if a resident of region i visits a distinct point of interest in region j. The points of interest can be of different types such as workplace, retail store, restaurant, or any of the other types mentioned in Sect. 3.1.1. Visits from other regions to region i are called in-degree centrality or inflow (IF_i), and visits made by residents of region i to other regions are called out-degree centrality or outflow (OF_i). For the Istanbul case, we consider each district of Istanbul as a node and obtain empirical values of in-flow and out-flow using the credit card transaction dataset. The network edge weight W_{ij} is defined as the total volume of flow from i to j. Let the set of individuals that reside in district i be S_i and the set of transactions by person k that occur in distinct places of district j be T_{kj} and $C_{T_{kj}}$ be the count of transactions made by person k in district j. Then the network edge weight is given by equation below:

$$W_{ij} = \sum_{k \in S_i} C_{T_{kj}}. \tag{1}$$

This approach results in a fully connected directed network with 36 nodes, and 1296 edges.

Huff gravity model Our aim is to understand and predict the flow of citizens between different districts using credit card and map data, and investigate the factors associated with the volume of flows. Based on the definition of flow given above, people visit different places in a region either in relation to their work or for other purposes such as shopping and visits with hedonistic motivations [34]. Motivated by notion of "consumer city" by Glaeser et al. [29], we contend that the number of amenities as well as their diversity could be a region's attractiveness for citizens. In statistical physics, Shannon entropy of a system is a measure of the number of permutations of the states of the system [35]. In the context of urban amenities, it measures the number of different ways to combine these amenities together. Therefore, it represents a metric of the variety of experiences

that city dwellers obtain from this mix of amenities. Moreover, ease of access to the region of interest is another important factor to attract more visitors [36]. Thus, we use the Huff gravity model [37] to model the flows between districts. Huff model is a technique in spatial analysis which is based on the principle that the probability that a consumer in district i will visit and purchase something in district j is a function of the relative distance d_{ij} and attractiveness of district j.

While new machine learning and AI methods have found their way to various spatial analyses such as flow modeling [38], the Huff model and its variations are widely used in numerous applications. In a previous study by Suhara et al. [39], Huff model performance has been validated by transactional large-scale dataset. The Huff gravity model can be used for both individual and aggregated flows, therefore, using the model at districts level does not violate its assumptions. In this model, the flows are governed as following: for an individual that originates from district i, each district j has a utility U_{ij} that is proportional to its attractiveness A_j and inversely proportional to the distance d_{ij} from the origin.

$$U_{ij} = \frac{A_j}{d_{ij}^{\gamma}}. \tag{2}$$

In particular, we used the product of POI count, Q and POI diversity, D (raised to appropriate powers) as the measure of attractiveness for each district. The diversity of POIs in district j is measured by the Shannon Entropy of the distribution of POI types $P_k^{(j)}$. While it is very rare to have only one type of POI in an urban area, such situation will result in zero entropy and lead to an attractiveness equal to zero. This problem can be handled by transforming entropy to (entropy + 1) or exp(entropy) in the model. These transformations won't change the rest of the algorithm.

$$A_j = (Q_j)^{\alpha}(D_j)^{\beta}, \tag{3}$$

$$D_j = \sum_k -P_k^{(j)} \log\left(P_k^{(j)}\right). \tag{4}$$

Instead of Euclidean distance, we use the travel time via public transportation as the distance d_{ij}. This is a more realistic measure of the ease of access. We use Google Maps API to get these travel times. Google Map Distance Matrix API allows users to obtain the average travel time between two distinct points by entering the origin and destination coordinates. We use the credit card transaction dataset to in turn determine the origin and destination points in each district. Origin for each district is calculated as the spatial average of residents' home locations and destination is the spatial average of merchant locations in those districts. We avoid using district centroids as origin or destination, because the population distribution is not spatially uniform over the districts. For example, in some districts lakes or hills cover large areas of central parts and make those areas uninhabitable. Travel time matrix is available at: github.com/cshikai/cities/blob/master/data/istanbul/effective_distance.csv.

Under this model, the proportion of trips originating from i that will have j as its destination, $\hat{f}_{ij}$, is proportional to its utility relative to other competing districts.

$$\hat{f}_{ij} = \frac{U_{ij}}{\sum_{j'} U_{ij'}}. \tag{5}$$

To the best of our knowledge, this study is the first to model the shopping area choice with the Huff model and the proposed novel area attractiveness measure with number and diversity of urban amenities.

Optimizing parameters Each Huff gravity model that describes the flows of district i is parameterized by α, β, and γ. In order to calibrate the parameters, we considered three different methods. The first two involve fitting a generalized linear model (GLM) with Poisson and Negative Binomial distributions on the count of flows [40]. The third method involves normalizing the flow counts by the total count of each district i to obtain the probability of movement to j given that the origin is i. Then the model is linearized and parameters are optimized via ordinary least squares regression according to the methodology listed in Huff & McCALLUM (2008) [41]. Our analyses show that the third method (via OLS) has the best result among three, therefore, we use that method for calibrating the parameters as described in what follows. Details about the performance and comparison metrics are provided in the Additional file 1. Following the third approach, we transform the model into a linear form in the parameters by applying the following transformation to P_{ij}:

$$\log\left(\frac{P_{ij}}{\tilde{P}_i}\right) = \alpha \log\left(\frac{Q_{ij}}{\tilde{Q}_i}\right) + \beta \log\left(\frac{D_{ij}}{\tilde{D}_i}\right) - \gamma \log\left(\frac{d_{ij}}{\tilde{d}_i}\right), \tag{6}$$

where $\tilde{P}_i$, $\tilde{Q}_i$, $\tilde{D}_i$, and $\tilde{d}_i$ are the geometric means of P_{ij}, Q_{ij}, D_{ij}, and d_{ij} respectively. Then we use ordinary least squares to estimate the parameters of the model.

3.2.2 Measuring economic productivity

We attempt to measure the economic productivity of the city of Istanbul via the Gross Domestic Product (GDP). Since official GDP records from Turkey are only published with granularity at a city level, we use the insurance sales data as a proxy for GDP at a finer district level. This dataset is provided by a Turkish Insurance Company that has been ranked one of the top three in sales volume based on different insurance categories. The dataset contains the company's total contract sales values in all categories in Istanbul districts level, during 2014, 2015, and 2016.

The proxy used is the sum of insurance sales in each district. Among all the insurance categories, we consider four categories, namely: Home, Workplace, Vehicle liability, and Vehicle Collision insurances. The reason why these insurance types were chosen is that all of them are based on physical assets like buildings or vehicles, which is a more direct measure of economic output. In order to test whether the chosen insurance types can be used as proxy for GDP, the district-level insurance sales in the chosen four categories for 80 cities in Turkey for the year 2014 was summed up and compared with published GDP of those 80 cities in the same year. The evaluation metric used was Pearson's correlation, which was found to be 0.99249 between two sets (more details are provided in Additional file 1). The result shows that insurance valuations published could serve as a proxy for GDP at the district level.

3.2.3 Inferring consumption diversity from data

We begin by obtaining the set of local businesses that exist within each regional unit j, $S^{(j)}$. These sets of local businesses are divided based on their categories k into subsets $S_k^{(j)}$.

The categories include types of restaurant, such as fast-food, Japanese, Mexican, etc., or types of facilities like parks, movie theatres, etc., based on the nature of the businesses. Additional information about the regional units and business categories are provided in the Additional file 1.

From credit card expenditures in Istanbul and the Meituan dataset in China, the consumption of a good or service is directly obtainable. For the Yelp and Dianping data sets the rate of consumption of a good or service i, C_i is estimated by adding up the total number of customers who have either checked in, rated, left a review, or posted a picture about the amenity. Thus, for each region j, we are able to obtain the proportion of consumption in a particular category k.

$$P_k^{(j)} = \frac{\sum_{i \in S_k^{(j)}} C_i}{\sum_k \sum_{i \in S_k^{(j)}} C_i}. \tag{7}$$

We also define a metric of diversity of urban amenities consumed by people in region j using the Shannon entropy [42].

$$D(POI_j) = \sum_k -P_k^{(j)} \log\left(P_k^{(j)}\right). \tag{8}$$

By obtaining different sets of amenities that were consumed by users across different time periods, we are able to investigate the relationship between attractiveness of local goods and economic growth across different years.

3.2.4 Diversity of the inflow of people

For each individual k, we have a vector X_k that describes the demographics of an individual. For the study of Istanbul, $X_k \in R^6$, where the vector X_k contains information on the *home district*, *work district*, *age category*, *gender*, *education level*, *and income deciles* of the individual.

We compute $q_x^{(j)}$, the proportion of inflows in district j associated with people with a particular demographic x. The demographic diversity of people entering district j is thus given by

$$D_j = \sum_{x \in X} -q_x^{(j)} \log\left(q_x^{(j)}\right). \tag{9}$$

3.2.5 Measuring economic growth

Istanbul We measure growth of Istanbul via changes in its GDP between 2014 and 2016.

$$G_{j,t} \approx \frac{GDP_{j,t+1} - GDP_{j,t}}{GDP_{j,t}}. \tag{10}$$

Beijing We measure the economic development in region j of Beijing at time t, $C_{j,t}$ by the total capital asset of secondary and tertiary sectors in each Beijing district. Capital accumulation is a key determinant of positive economic growth in many established economic models, and thus we take this metric as an indicator of future economic growth.

$$G_{j,t} \approx \frac{C_{j,t+1} - C_{j,t}}{C_{j,t}}. \tag{11}$$

United States Similar to Turkey, we measured the growth of each district via changes in its economic output. Again, since official GDP records are only published at a state level, we approximated it via the sum of personal incomes in each district j, which is obtained from the five-year United States census.

$$G_{j,t} \approx \frac{Total\ Personal\ Income_{j,t+1} - Total\ Personal\ Income_{j,t}}{Total\ Personal\ Income_{j,t}}. \tag{12}$$

3.2.6 Growth model

We propose the following model in order to measures the relationship between economic growth rate (G), and various factors in urban area:

$$G_{j,t} = \beta_0 + \beta_1 H_{j,t} + \beta_2 \rho_{j,t} + \beta_3 I_{j,t} + \beta_4 D_{j,t} + \epsilon_{j,t}, \tag{13}$$

where $G_{j,t}$, $H_{j,t}$, $\rho_{j,t}$, $I_{j,t}$ and $D_{j,t}$ are consumption diversity, population density, housing index, and eigenvalue centrality of the district j at time t respectively. The control variables are explained below.

Population density The population density $\rho_{j,t}$ is defined as the number of people per unit area (in 1000/km^2).

Housing index Rental prices act as a medium that regulates investment into the region. Increase in rental prices may discourage businesses to be physically set up in the region, or reduce the supply of labor as individuals are reluctant to move into the area. To account for the effects of property prices on growth, $I_{j,t}$ is included as a covariate, which is defined as the ratio of the current rental price (per unit area) to the cheapest rental in the first time period t_0.

$$I_{j,t} = \frac{Rent_{j,t}}{Rent_{j,t_0}}, \tag{14}$$

where $j^* = \arg\min_j Rent_{j,t_0}$

Housing index information is provided by the leading real estate listing company in Turkey for research purposes. Detailed information and data are available at Additional file 1 and our GitHub page.

Eigenvalue centrality in a geographical network The success of a district is also dependent on its location in the city. Generally speaking, the more central the location, the easier the access to the businesses, and therefore, the more likely it would attract customers. To control for the location factor, we first computed a geographical network of districts with the edges between districts i and j weighted by the reciprocal of the travel distance (in measured in minutes) between them. We then computed the eigenvalue centrality of each business in this geographical network [43].

4 Results

4.1 Flow and diversity

We model the interactions that happen across different regions in a city via a network model of human flow, where the nodes are the regional units (districts, neighborhoods,

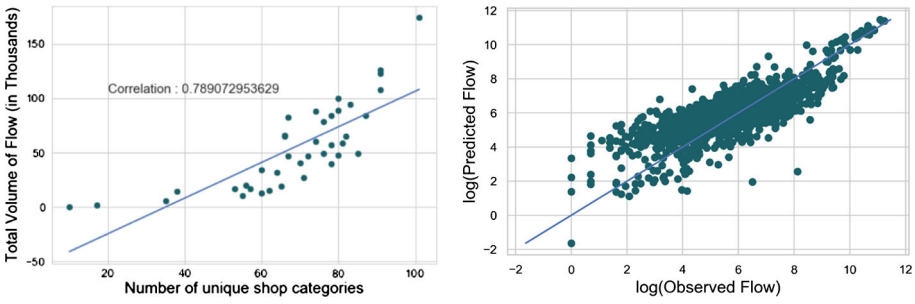

Figure 1 Left—Scatter plot of the number of unique shop categories versus total inflow volumes of each district in Istanbul. Right—Scatter plot of predicted flows vs observed flows

Table 2 Huff Model Parameters resulting from optimization via OLS method

	Coeff	Std Err	t	$P > \|t\|$	[0.025	0.975]
α	0.9982	0.089	11.214	0	0.824	1.173
β	3.4496	0.224	15.413	0	3.011	3.889
γ	−1.9589	0.055	−35.663	0	−2.067	−1.851

zip codes, etc.), and the edges represent the volume of flows between the nodes. We investigated how the quantity and diversity of amenities in those regions relate to the different volumes of flows across neighborhoods. Various points of interest in an urban environment were identified using a dataset produced by a commercial digital mapping company. This dataset includes a map of Istanbul with details such as different levels of administrative boundaries (e.g. districts and neighborhoods), and various categories of Points of Interest (POI), published quarterly from 2015 to the end of the first quarter of year 2016. Available POI types are already provided in the Methods section. The scatter plot in Fig. 1 (left) shows that there is a significant relationship between the total volume of inflows and diversity of commodities (measured via number of unique shop categories) in the area, with a correlation value of 0.789.

Other than the total volume of inflow, we are also interested in how the quantity and diversity of amenities factor into flows across each pair of districts. As such, we model this inter-district flow of citizens with a Huff gravity model [37] that uses the quantity (Q) and diversity (D) of amenities as the attractiveness measure and travel times (d) between the districts as proximity measure. Under this model, the probability of moving from region i to region j is proportional to the attractiveness of region j and inversely proportional to the proximity of two regions. Formally,

$$P(i \to j) \propto \frac{Q_j^{\alpha} D_j^{\beta}}{d_{ij}^{\gamma}}. \tag{15}$$

Using a global parameter value for the Huff gravity models across all districts, we optimized the value of the parameters α, β, and γ via OLS, which correspond to the scaling factor of Quantity, Diversity, and Distance in the model as shown in equation 15. Figure 1 (right) summarizes the performance of this model. The mean R^2 values of the models is 0.648, with corresponding parameters $\alpha = 0.9$, $\beta = 3.4$, and $\gamma = 2.1$.

The good fit of this model suggests that measurements of attractiveness and not just geographical distance is critical for understanding urban flows. Table 2 shows the Huff model

parameters resulting from optimization via OLS. We observe that the optimal model returns an α value of 0.9, which is less than β, at 3.4, indicating that diversity plays a larger role than the quantity of commodities in determining the attractiveness of a region. Attractiveness scales sub-linearly with quantity, while scaling super-linearly with the diversity. There is also super-linear scaling with the geographical accessibility. For robustness, we also fitted the flows across districts with generalized linear models (with Poisson and negative binomial distributions [40]), and obtained similar results, with POI sum scaling sub-linearly and POI Diversity scaling super-linearly. More information is provided in the Additional file 1.

We also fitted independent parameter values for each district, and the district-by-district optimized flow can be found in the Additional file 1. Despite the large increase in the degree of freedom of the parameters, the increase in performance, measured by the mean R^2 value, is marginal (0.623 to 0.658). This provides evidence for a robust, homogeneous rule that governs the human flow across all districts of Istanbul. This universality across districts is in contradiction with past observations that the rules governing flows vary strongly with physical locations [44].

4.2 Flow and productivity

Here, we investigate how the volume of human flow can account for the difference in economic productivity in different districts of a city. Existing literature suggests that productivity in urban environments is driven by idea exchange and innovation [22]. Accordingly, we expect that a large number of interactions to arise from large inflows of people, leading to more information flow, innovation, and more economic opportunities for people.

Figure 2 shows the relationships between the flows of each district within a one-year period and the corresponding economic productivity (details about the economic indicator are provided in the Method section). The Pearson correlation coefficients between the inflow, outflow, and total flow with the economic indicator are 0.843, 0.475, and 0.840 respectively, indicating a positive relationship between the flows and the districts' economic productivity. We test the significance of the effect of inflow and outflow on economic productivity via linear regressions. The results are shown in Table 3.

The results in Table 3 indicate that the inflow and outflow are significant variables in all models, with a positive relationship with productivity. The results after step-wise inclusion of *residential* population size and within-district flow as independent variables (Models 2, 3, and 4) alleviate the confounding effects of residential population density and flow of the residents within their own district. Moreover, the R^2 values of the regression models (at 0.742, 0.743, 0.751, 0751 for Models 1, 2, 3, and 4 respectively) show that inflows

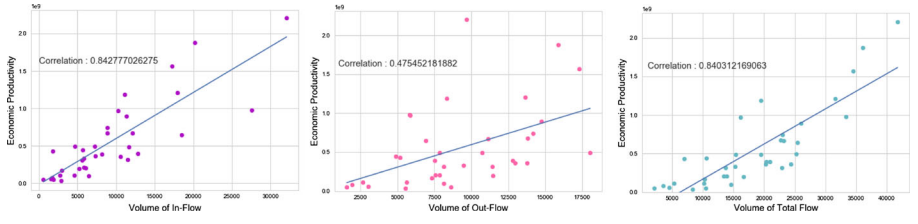

Figure 2 Scatter plot of total flow volumes (Left: Inflow, Middle: Outflow, Right: Total Flow) and the economic productivity of each district in Istanbul

Table 3 Regression coefficients for prediction of Economic Productivity using Human Flows

	Model 1	Model 2	Model 3	Model 4
Inflow	21.080***	21.712***	20.553***	20.512***
	(2.6715)	(3.7222)	(2.7105)	(3.899)
Outflow	14.752***	16.831*	20.064**	19.955*
	(5.1697)	(9.8939)	(7.1530)	(10.344)
Within – district F low		–4.132		0.253
		(16.675)		(17.202)
Population			–0.386	–0.397
			(0.36984)	(0.387)
Constant	–153,476.8	–164,618.5	–92,124.5	–91,219.5
	(107,900)	(118,342)	(121,931)	(138,224)
Observations	36	36	36	36
R^2	0.7426	0.7433	0.7517	0.7517
Adjusted R^2	0.7192	0.7193	0.7284	0.7286

Standard errors in parentheses. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

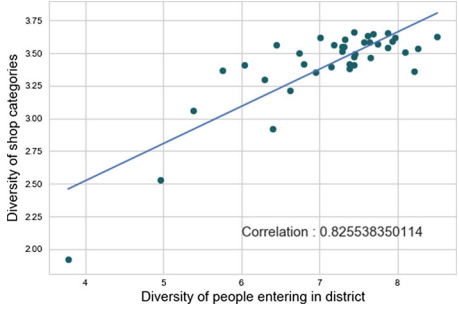

Figure 3 Demographic diversity of people entering each district vs diversity of amenities in those districts (Istanbul)

and outflows explain a large percentage of the variation in the economic productivity of the region. Our results support the hypothesis that the inflow of people and the resulting proximity and interaction with other districts' residents give rise to information diffusion. Additionally, the outflow of residents to other districts allows exploration of new knowledge, which could possibly then be exploited and transferred to the home district upon return. New information then leads to more economic opportunities in the area.

4.3 Consumption diversity and flow diversity

Interestingly, our study shows that there is a significant correlation between the demographic diversity of the inflow of people (computation details are already provided in Methods section) and the availability of commodities in the region. Figure 3 shows the scatter plot of the diversity of the people (demographics include characteristics such as age, gender, income levels, work, and home districts) entering each district of Istanbul versus the diversity of goods and services available (categorized using the merchant category codes-MCC) in these districts.

The two variables have a correlation of 0.825. According to Glaeser et al. [45], availability of diverse goods and services provide the means to attract people with different demographics and varying taste and preferences. Similarly, the inflow of diverse people provides the conditions for a wider variety of businesses to be set up in the region, allowing the local economy to flourish.

4.4 Diversity and economic growth

We investigate the above-mentioned relationship between the diversity in a neighborhood, quantified via the Shannon Entropy, and the economic growth. Specifically, diversity is given by the diversity of goods and services consumed in the neighborhood, as well as the diversity of the inflow of people into the area.

We start by investigating the relationship between diversity of consumption and the growth of the economy. Leveraging publicly available data from social networking and crowd sourcing websites, we are able to extend this part of the study from one city to 3 urban regions in three different countries: Turkey, China, and the United States. These urban regions are divided and studied on a sub-city granularity. Details of the economic indicators and consumption data are defined in the Methods section.

Figures 4(A), 5(A), and 6(A) show the scatter plots for regions in Istanbul, Beijing, and various urban areas in the United States. In all three cases, we see that the diversity of consumption exhibits significant statistical positive correlations with the economic growth in the following year at 0.71 (Istanbul, Fig. 4(A)), 0.54 (Beijing, Fig. 5(A)), and 0.52 (U.S., Fig. 6(A)), indicating that the diversity of consumption alone accounts for between one quarter and half of the variance in future economic growth.

Figures 4(B), 5(B), and 6(B) show the residual scatter plots after we account for the correlations with variables such as population density, housing price index, and the geographical centrality of the district within the city. Even after controlling for covariates that could potentially affect economic growth rates, we observe that the diversity still has significant partial correlations with economic growth at 0.72 (Istanbul, Fig. 4(B)), 0.41 (Beijing, Fig. 5(B)), and 0.57 (US, Fig. 6(B)).

We fit an ordinary least squares (OLS) regression model to all available variables. The relationship between economic growth, G and consumption diversity, H is modelled with

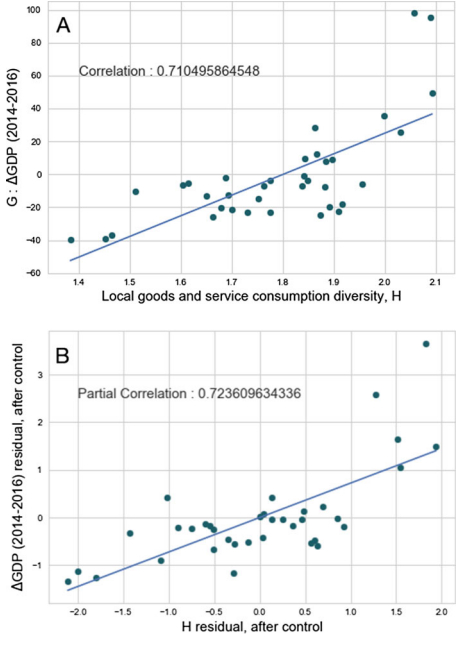

Figure 4 Scatter plot of growth vs diversity of goods consumed. Each data point corresponds to a district in Istanbul in a specific year. Top (**A**)—Full correlation. Bottom (**B**)—Partial correlation after controlling for other variables

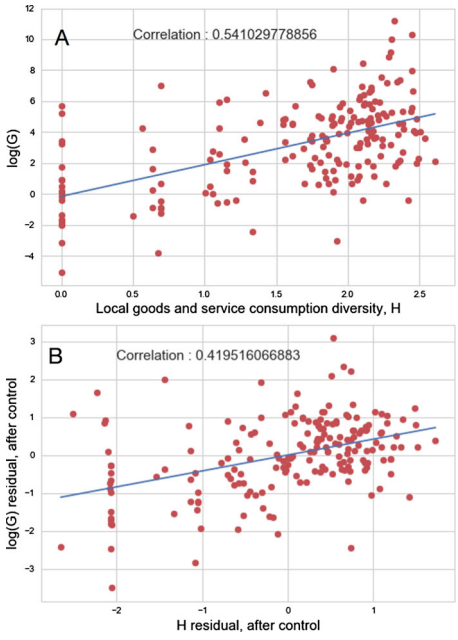

Figure 5 Scatter plot of growth vs diversity of goods consumed. Each data point corresponds to a district in Beijing in a specific year. Top (**A**)—Full correlation. Bottom (**B**)—Partial correlation after controlling for other variables

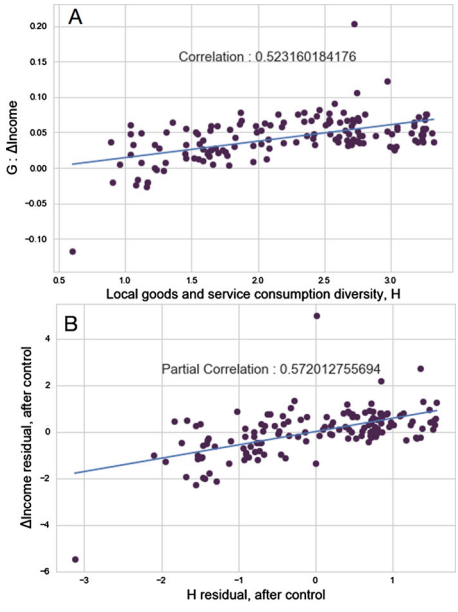

Figure 6 Scatter plot of growth vs diversity of goods consumed. Each data point corresponds to a census block in United States in a specific year. Top (**A**)—Full correlation. Bottom (**B**)—Partial correlation after controlling for other variables

the following equation:

$$G_{j,t} = \beta_0 + \beta_1 H_{j,t} + \beta_2 \rho_{j,t} + \beta_3 I_{j,t} + \beta_4 D_{j,t} + \epsilon_{j,t}, \tag{16}$$

where $G_{j,t}$, $H_{j,t}$, $\rho_{j,t}$, $I_{j,t}$ and $D_{j,t}$ are the growth rate, consumption diversity, population density, housing index, and eigenvalue centrality of the district j at time t, respectively.

The regression diagnostics are presented in Table 4. The results show that consumption diversity has a significant, positive effect on growth in all three cases, even after accounting for the effects of covariates such as population density, housing price index, etc. We ob-

Table 4 Regression coefficients for prediction of Economic Growth using Consumption Diversity

	Beijing	Istanbul	USA
Consumption Diversity, H	0.233***	0.707***	0.583***
	(0.0493)	(0.121)	(0.0706)
Population Density, ρ	0.474***	–0.389*	–0.213**
	(0.0644)	(0.464)	(0.0729)
Housing Index, I	0.164***	–0.113	0.0981
	(0.0469)	(0.147)	(0.0703)
Geographic Centrality, D	0.222***	0.245	0.270***
	(0.0635)	(0.209)	(0.0727)
Constant	–1.97E–09	4.11E–09	–2.19E–09
	(0.0425)	(0.117)	(0.0678)
Observations	187	36	145
R^2	0.671	0.574	0.357
Adjusted R^2	0.664	0.519	0.338

Standard errors in parentheses. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

serve that amongst all the variables, consumption diversity has the most consistent effect across all three regions; it is the only variable that has a consistent direction of dependence and is statistically significant.

We should note that the model is observed to have a poorer fit for the study conducted in the urban areas of the United States, with an R^2 of 0.357 (Table 4). This may be attributed to the geographically sparse, and possibly biased, data sample. The Yelp data set we used, merely contains data on a subset of census blocks that were geographically distributed across different states as opposed to the districts studied in China and Turkey, which were all contained within the same city and provided a complete picture of Beijing and Istanbul respectively. In addition, the census blocks are sampled across different time periods t. Nevertheless, the regression results in Table 4 show that the positive relationship between consumption diversity and economic growth is reasonably strong and statistically significant.

5 Discussion

Our results show that human flow between districts in a city can be effectively explained by the relative attractiveness due to the local amenities and ease of access to a district. Our proposed variation of Huff gravity model that takes into account the number and diversity of POIs in a region accounts for more than 85% of the variation in flows between districts. Additionally, our analysis shows that movement to districts of a city scales super-linearly with the diversity of amenities, in favor of the theory of intervening opportunities [46], suggesting that trip making is not explicitly dependent on physical distance but on the accessibility of resources satisfying the objective of the trip.

We also show that the differences in the volume of flow of people into different districts can predict the differences in economic productivity among various districts of Istanbul. We find that the inflow of people has a strong positive relationship with the economic productivity of the region. These results improve the findings of Bettencourt et al. [8, 11], Gomez-Lievano et al. [12] and Pan et al. [13] who report super-linear scaling of economic productivity in relation to population density at the aggregate level as well as studies by others [8, 19, 20, 22, 24, 47, 48] in that one can now quantitatively predict, with a linear model, the variations of economic output in different parts of a city with an R^2 value of 0.909. The effects of inflow and outflow of people remain significant predictors of eco-

nomic output even after accounting for indicators traditionally associated with urban scaling phenomena such as population. Our results are in favor of the concept that productivity in urban settlements is primarily driven by the strong interaction, idea generation and exchange within the urban network.

Finally, we show that the consumption diversity and the diversity of human flow in an urban area can be utilized as a signal about its future economic growth, as there is a positive relationship between neighborhood diversity and economic growth a year later, even after controlling for population density, housing price index, and the geographical centrality of the district within the city.

We contend that the following mechanism can be employed to explain these results and findings that the flow of diverse people catalyzes the production of yet more diverse goods and services, which attracts even more diverse and larger flows of people. Specialized shops and services not only require larger pools of the population but also diverse patrons with different tastes and preferences in order to reach viable levels of demands. These businesses are economically feasible only in diverse, bustling cities. This makes it likely for business owners to invest money in these urban areas, which can support a large variety of specialized businesses, such as niche restaurants that specialize in various cuisines. The result is the increase in the level of economic activity in these areas. Overall, this sets up a dynamic process in which entrepreneurs respond to the flow of people with injections of money and investments into new businesses. These new amenities provide the conditions, both in terms of their inherent utility and new work opportunities, to continuously attract a flow of people and allow the economy to expand.

While we only have direct consumption data in Istanbul, we extended our study to urban areas in two other continents, specifically Beijing, China and some disconnected census blocks in the U.S.A. In these regions, we obtained information on consumption diversity via publicly available data. Our analysis with the data from Dianping and Meituan platforms as well as the data from the Yelp Dataset Challenge suggest that our basic premise on diversity of amenities driving economic productivity is still well-supported, suggesting a robust relationship.

6 Conclusion

Cities are home to more than half of the world's population and based on the UN projections, cities will attract almost all of the growth in the human population over the next three decades [7]. While cities are known as engines of industry and production, various undesirable factors such as income inequality and socioeconomic segregation affect the citizens' well-being [8–10, 49]. On the other hand, cities, with their high population density, facilitate the consumption of a wide variety of commodities. Thus, understanding how urban environments impact citizens' behaviour in the cities is a research field of increasing importance. This study highlights computational techniques that leverage large-scale datasets to help the planners and policy makers better understand the effect of urban characteristics on economic productivity.

Past work by Glaser et al. [45] has shown that the quantity of amenities is positively correlated with the *population growth rate* of cities, suggesting that people move into cities because they value the diversity of consumption that urban environments provide. Here we extend this idea to the more granular level of individual neighborhoods and districts, and show how commodity diversity predicts both mobility and economic outcomes, by using data on the diversity of amenities in Istanbul and the flow of people within the city.

One limitation of our study is that the data is not a perfect representation of the flow of people around the cities. Geo-tagged credit card records do not completely capture the movements of people; only flows that involve economic transactions are captured. Therefore, we would expect higher recorded flows in regions with a higher number of shops. A better source of flow data could potentially be mobile devices equipped with the capability of pinging an individual's location every few minutes. Nevertheless, we mitigated this limitation by including a variable in the model that takes into account the number of POIs in the region. Even after this mitigation, the diversity of amenities well explained the flow of people.

Our findings have practical implications for the urban planners and officials of cities. In this new era of Big Data, readily available data sources offer many dynamic measures of a city's economic health from different perspectives [50], and our work demonstrates how electronic records of expenditures can be utilized to do so. City officials need not rely on annualized values of traditional economic indicators for planning purposes but are instead able to obtain up-to-date metrics of how well the city is doing. As a city's ability to remain attractive as a consumption nexus becomes increasingly associated with its success in retaining high-value-adding people and improving its economic well-being, decision makers should consider making their cities a more diverse and vibrant place not just for work, but a pleasant place to live and play in.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-020-00234-x.

Additional file 1. Supplementary information (PDF 8.2 MB)

Acknowledgements
We thank the MIT Trust Data Consortium and Prof. Guan at the Capital Institute of Science and Technology Development Strategy, Beijing for making this project possible. We also thank X. Dong, Y. Leng, and Y. Yuan for helping with the collection and cleaning of the Meituan/Dianping dataset. We are indebted to the companies (the ones we introduced in the manuscript and those who wish to remain anonymous) for providing the insurance, credit card, and housing data in Istanbul. We also thank the anonymous reviewers for their valuable comments and contributions towards improving our paper.

Funding
No funding was provided for this research.

Abbreviations
GDP, Gross Domestic Product; GLM, Generalized Linear Model; OLS, Ordinary Least Squares; POI, Points of Interest; MCC, Merchant Category Codes.

Availability of data and materials
The data and code used in the analysis are available at: https://github.com/cshikai/Cities

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
SKC collected data, performed research and analyzed data for consumption and economic growth; MB collected data, performed research and analyzed data for commodities and human flow; HC collected the economic data for Beijing; SB collected the data for Istanbul; and SKC, MB, AP and BB contributed to writing. All authors read and approved the final manuscript.

Author details
[1]The Media Lab, Massachusetts Institute of Technology, Cambridge, USA. [2]MIT Connection Science, Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, USA. [3]School of Management, Sabanci University, Istanbul, Turkey. [4]School of Economics and Resource Management, Beijing Normal University, Beijing, China. [5]Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey. [6]New College of Florida, Sarasota, USA.

Publisher's Note

Received: 26 November 2019 Accepted: 10 June 2020 Published online: 24 June 2020

References

1. Jacobs J (1969) The economy of cities. Vintage, New York
2. Jacobs J (1985) Cities and the wealth of nations: principles of economic life. Vintage, New York
3. Polèse M (2005) Cities and national economic growth: a reappraisal. Urban Stud 42(8):1429–1451
4. Glaeser EL, Gottlieb JD (2006) Urban resurgence and the consumer city. Urban Stud 43(8):1275–1299
5. Clark TN, Lloyd R, Wong KK, Jain P (2002) Amenities drive urban growth. J Urban Aff 24(5):493–515
6. Nichols Clark T (2003) Urban amenities: lakes, opera, and juice bars: do they drive development? In: The city as an entertainment machine. Emerald Group Pub., Bingley, pp 103–140
7. Crane P, Kinzig A (2005) Nature in the Metropolis. Science 308(5726):1225
8. Bettencourt LM, Lobo J, Helbing D, Kuhnert C, West GB (2007) Growth, innovation, scaling, and the pace of life in cities. Proc Natl Acad Sci 104(17):7301–7306
9. Clark WA (1986) Residential segregation in American cities: a review and interpretation. Popul Res Policy Rev 5(2):95–127
10. Trounstine J (2018) Segregation by design: local politics and inequality in American cities. Cambridge University Press, Cambridge
11. Bettencourt L, West G (2010) A unified theory of urban living. Nature 467(7318):912–913
12. Gomez-Lievano A, Patterson-Lomba O, Hausmann R (2017) Explaining the prevalence, scaling and variance of urban phenomena. Nat Hum Behav 1(1):0012
13. Pan W, Ghoshal G, Krumme C, Cebrian M, Pentland A (2013) Urban characteristics attributable to density-driven tie formation. Nat Commun 4:1961
14. Becker GS, Glaeser EL, Murphy KM (1999) Population and economic growth. Am Econ Rev 89(2):145–149
15. Glaeser EL (2011) Triumph of the city: how our greatest invention makes us richer, smarter, greener, healthier, and happier. Penguin, New York
16. Jacobs J (1961) The death and life of great American cities. Random House, New York
17. Sung H, Lee S, Cheon S (2015) Operationalizing jane Jacobs' urban design theory: empirical verification from the great city of Seoul, Korea. J Plan Educ Res 35(2):117–130
18. De Nadai M, Staiano J, Larcher R, Sebe N, Quercia D, Lepri B (2016) The death and life of great Italian cities: a mobile phone data perspective. In: Proceedings of the 25th International Conference on World Wide Web, pp 413–423
19. Burt RS (2009) Structural holes: the social structure of competition. Harvard University Press, Cambridge
20. Granovetter MS (1977) The strength of weak ties. In: Social networks. Academic Press, San Diego, pp 347–367
21. Wu L, Waber BN, Aral S, Brynjolfsson E, Pentland A (2008) Mining face-to-face interaction networks using sociometric badges: predicting productivity in an it configuration task. Available at SSRN. http://dx.doi.org/10.2139/ssrn.1130251
22. Granovetter M (2005) The impact of social structure on economic outcomes. J Econ Perspect 19(1):33–50
23. Allen TJ (1984) Managing the flow of technology: technology transfer and the dissemination of technological information within the R&D organization. MIT Press, Cambridge
24. Reagans R, Zuckerman EW (2001) Networks, diversity, and productivity: the social capital of corporate R&D teams. Organ Sci 12(4):502–517
25. Dong X, Suhara Y, Bozkaya B, Singh VK, Lepri B, Pentland AS (2018) Social bridges in urban purchase behavior. ACM Trans Intell Syst Technol 9(3):33
26. Di Clemente R, Luengo-Oroz M, Travizano M, Xu S, Vaitla B, González MC (2018) Sequences of purchases in credit card data reveal lifestyle in urban populations. Nat Commun 9(1):1–8
27. Yang S, Wang M, Wang W, Sun Y, Gao J, Zhang W, Zhang J (2017) Predicting commercial activeness over urban big data. Proc ACM Interact Mob Wearable Ubiquitous Technol 1(3):1–20
28. Dong L, Chen S, Cheng Y, Wu Z, Li C, Wu H (2017) Measuring economic activity in China with mobile big data. EPJ Data Sci 6(1):29
29. Glaeser EL (2001) Consumer city. J Econ Geogr 1(1):27–50
30. Kloosterman RC (2014) Cultural amenities: large and small, mainstream and niche—a conceptual framework for cultural planning in an age of austerity. Eur Plan Stud 22(12):2510–2525
31. Dong L, Ratti C, Zheng S (2019) Predicting neighborhoods' socioeconomic attributes using restaurant data. Proc Natl Acad Sci 116(31):15447–15452
32. Centola D, Macy M (2007) Complex contagions and the weakness of long ties. Am J Sociol 113(3):702–734
33. Zhou X, Hristova D, Noulas A, Mascolo C, Sklar M (2017) Cultural investment and urban socio-economic development: a geosocial network approach. R Soc Open Sci 4(9):170413
34. Arnold MJ, Reynolds KE (2003) Hedonic shopping motivations. J Retail 79(2):77–95
35. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423
36. Sim A, Yaliraki SN, Barahona M, Stumpf MP (2015) Great cities look small. J R Soc Interface 12(109):20150315
37. Huff DL (1963) A probabilistic analysis of shopping center trade areas. Land Econ 39(1):81–90
38. Yeghikyan G, Opolka FL, Nanni M, Lepri B, Lio P (2020) Learning mobility flows from urban features with spatial interaction models and neural networks. arXiv preprint. arXiv:2004.11924
39. Suhara Y, Bahrami M, Bozkaya B, Pentland A (2019) Validating gravity-based market share models using large-scale transactional data. arXiv preprint. arXiv:1902.03488
40. Beir´o MG, Bravo L, Caro D, Cattuto C, Ferres L, Graells-Garrido E (2018) Shopping mall attraction and social mixing at a city scale. EPJ Data Sci 7(1):28
41. Huff D, McCallum BM (2008) Calibrating the huff model using arcgis business analyst. ESRI White Paper
42. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423
43. Newman ME (2016) Mathematics of networks. In: The new Palgrave dictionary of economics, pp 1–8
44. Isaacman S, Becker R, Cáceres R, Kobourov S, Rowland J, Varshavsky A (2010) A tale of two cities. In: Proceedings of the eleventh workshop on mobile computing systems and applications 2010. ACM, New York, pp 19–24

45. Glaeser EL, Mare DC (2001) Cities and skills. J Labor Econ 19(2):316–342
46. Stouffer SA (1940) Intervening opportunities: a theory relating mobility and distance. Am Sociol Rev 5(6):845–867
47. Pan W, Aharony N, Pentland A (2011) Fortune monitor or fortune Teller: understanding the connection between interaction patterns and financial status. In: 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing. IEEE Press, New York, pp 200–207
48. Quigley JM (1998) Urban diversity and economic growth. J Econ Perspect 12(2):127–138
49. Chen Y, Rosenthal SS (2008) Local amenities and life-cycle migration: do people move for jobs or fun? J Urban Econ 64(3):519–537
50. Glaeser EL, Kominers SD, Luca M, Naik N (2018) Big data and big cities: the promises and limitations of improved measures of urban life. Econ Inq 56(1):114–137