Spatial and Social Frictions in the City: Evidence from Yelp*

Donald R. Davis[†]
Columbia and NBER

Jonathan I. Dingel[‡]
Chicago Booth

Joan Monras[§]
Sciences Po

Eduardo Morales[¶]
Princeton and NBER

October 2015
Preliminary and Incomplete

Abstract

We employ user-generated data from the website Yelp.com to estimate how spatial and social frictions combine to shape consumption choices within cities. Travel time, from both home and work, plays a first-order role in consumption choices. Users are two to four times more likely to visit a venue that is half as far away. Social frictions also play a large role. Individuals are less likely to visit venues in places demographically different from their own neighborhood. A one-standard-deviation increase in demographic distance is equivalent to a 21% increase in travel time in terms of reduced visits. Higher crime rates reduce visit probabilities. Women are about 50% more responsive than men to local robbery rates.

*We thank Treb Allen, David Atkin, Victor Couture, Ingrid Gould Ellen, Mogens Fosgerau, Marcal Garolera, Joshua Gottlieb, Jessie Handbury, Albert Saiz, and seminar audiences at the NBER Summer Institute URB meeting, Urban Economics Association, NBER Fall ITI meeting, Princeton, NYU Stern, Duke ERID spatial-equilibrium conference, University of Barcelona - IEB, Yale, the Université du Québec à Montréal, Toronto, and UCLA for helpful comments. We thank Bowen Bao, Luis Costa, David Henriquez, Charlene Lee, Ludwig Suarez, and especially Ben Eckersley, Hadi Elzayn, and Benjamin Lee for research assistance. Thanks to the New York Police Department, and especially Gabriel Paez, for sharing geo-coded crime data. Dingel thanks the Kathryn and Grant Swick Faculty Research Fund at the University of Chicago Booth School of Business for supporting this work. Monras thanks the Banque de France Sciences Po partnership and LIEPP for financial support. Part of this work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program LIEPP (ANR-11-LABX-0091, ANR-11-IDEX-0005-02).

[†]drdavis@columbia.edu
[‡]jdingel@chicagobooth.edu
[§]joan.monras@sciencespo.fr
[¶]ecmorale@princeton.edu

1 Introduction

Cities make us more productive, but they also provide attractive consumption opportunities (Glaeser, Kolko, and Saiz, 2001). Consumption in cities often requires travel, so the value of these consumption opportunities depends on the geography of the city (Couture, 2014). Home and work are our primary locations within the city. And so home, work, and the commute between the two are the primary bases from which our consumption is launched.

Consumers' decisions in cities are influenced by both the products offered and other features of the urban context. We use the term "friction" to refer to determinants of demand beyond products' characteristics and prices. These frictions, which can be understood as components of a broader set of product characteristics, depend on both the location of the product and the identity of the consumer. *Spatial* frictions are costs of traversing the city that influence consumer decisions by making consumption farther away less attractive. These frictions play a central role in theories of spatial competition, dating to Hotelling (1929).

Spatial frictions fragment the city, making it less valuable to consumers. But frictions are not only spatial, they are also *social*. Social frictions are demographic or socioeconomic characteristics of locations that influence individuals' decisions to consume there. If aversion to differences in ethnicity, race, or income reduces our willingness to take advantage of consumption opportunities, then the value of the city is diminished. Concerns over perceived safety, which often relate to such differences, may also shape consumer decisions. These social frictions need not affect all members of society similarly; they could vary by ethnicity, race, or gender.

In this paper, we explore empirically how spatial and social frictions govern our use of the city. Our work characterizes the consumption decisions of individuals living in New York City who use Yelp.com, a website where users review local businesses. The data we collected from Yelp users' reviews identify their home and work locations and businesses patronized. We combine this information with data on places within New York City. The resulting dataset is novel in that it captures four dimensions necessary to studying consumption within the city. It identifies the restaurants patronized by consumers, characteristics of restaurants both chosen and unchosen, mode-specific travel times from users' home and work locations to these restaurants, and measures of social frictions within New York City.

We characterize consumer preferences by estimating a discrete-choice model of restaurant-visit decisions. Our estimates show that both spatial and social frictions influence the geography of consumption. First, we quantify the role of spatial frictions for consumption within the city. We estimate consumers' aversion to incurring longer travel times, controlling for product characteristics like restaurants' ratings and prices. Our model specification accounts for the facts that consumption may originate at home, work, or the commute between them and that both automobile and public transit are possible modes of transport. Across origin-mode pairs, halving the minutes of travel time to a venue would imply that the user would be two to nearly four times more likely to visit the venue from that origin by that mode. These are important parameters for quantitative assessments of consumption in the city, such as Allen, Arkolakis, and Li (2015). Further, our estimates imply that models of urban consumption are misspecified if they omit visits originating from the workplace.

Second, we quantify the role of demographic differences in consumers' decisions and find that this social friction plays an important role. Ceteris paribus, a user would be 27% more

likely to visit a venue in a census tract that is one standard deviation more demographically similar to her home tract. Moreover, these frictions are not symmetric. While users on average are less likely to visit a census tract that has demographics different from those of their own residence, the negative effect of demographic differences is more than twice as large when the destination tract is plurality black. Finally, beyond tract-level demographic differences, the racial/ethnic identity of individual users predicts their consumption behavior. Our estimates demonstrate homophily with regard to the user's race/ethnicity, implying that socioeconomic shifts such as gentrification have heterogeneous consequences for residents as a function of their identities. The finding that demographic differences deter visits implies that economic interactions are even more segregated than would be predicted by the combination of residential segregation and spatial frictions.

Third, we find a significant gender differential in the incidence of crime. Users are less likely to visit venues in places where more robberies occur, and women are significantly less likely than men to visit venues in neighborhoods with more robberies. These effects are modest in magnitude given the low levels of robberies in contemporary New York City, but they imply that the substantial decline in crime in New York City over the last twenty-five years was particularly advantageous to females, since females' use of the city is more responsive to crime rates.

Our findings relate to several strands of literature. First, we study the geography of consumption within the city. A recent literature has documented cross-city variation in the tradable goods available for consumption (Handbury and Weinstein, 2011; Handbury, 2012), and geographic variation in non-tradables has been posited to shape the relative attractiveness of cities (Glaeser, Kolko, and Saiz, 2001). This dimension of economic life has grown increasingly important in recent decades.[1] Prior studies of the geography of consumption within the city include Couture (2014), who infers the benefits of variety due to urban density from the time individuals spend traveling from their homes to restaurants, Houde (2012), who demonstrates that a demand model incorporating potential consumers' commuting paths better matches observed sales than a model estimated under the assumption that all consumer trips originate at their home locations, Katz (2007), who studies the impact that driving time from home has on the choice of grocery stores by consumers, and Eizenberg, Lach, and Yiftach (2015), who estimate within-city travel costs for consumers using data on neighborhood-level expenditure shares for Jerusalem supermarkets. Relative to this prior work, we exploit data describing individuals' home and work locations, their demographics, and characteristics of the venues they patronize. This allows us to estimate the effect of spatial frictions on consumer decisions while accounting for product characteristics. Furthermore, while prior literature has focused on spatial frictions, we emphasize the role of social frictions in shaping consumption within cities.

Our paper is also related to a large literature on social and economic fragmentation related to demographic differences. Much of that literature has documented ethnic and racial fragmentation in terms of residential segregation (Cutler, Glaeser, and Vigdor, 1999;

[1]US households' share of food spending devoted to food prepared away from home grew from less than 26% in 1970 to more than 43% in 2012. While the number of daily commuting trips has stayed relatively steady for decades, trips for social/recreational purposes have steadily grown (Commuting in America III, 2006).

Echenique and Fryer, 2007; Bayer and McMillan, 2008).[2] Many US cities are *de facto* quite segregated. Figure 1 depicts the population of New York City using colored dots to represent people of four different demographic groups – whites (green), blacks (blue), Hispanics (orange), and Asians (red).[3] There is a clear pattern of residential segregation. Some of the mechanisms posited to explain residential segregation would also predict segregation in individuals' consumption patterns. For example, between countries, lower bilateral trust reduces trade and investment (Guiso, Sapienza, and Zingales, 2009), and closer cultural ties can facilitate economic exchange (Rauch, 2001). A number of studies have documented retail-shopping frictions related to racial/ethnic composition (Lee, 2000; Ayres, 2001; Antecol and Cobb-Clark, 2008; Schreer, Smith, and Thomas, 2009). This paper quantifies how demographic differences shape which places people visit for consumption purposes. We use information on the demographic composition of residents surrounding venues and users' residential locations, as well as the ethnicity or race of the user, to demonstrate that ethnic or racial dissimilarity reduces economic interactions within New York City after accounting for geography, venue characteristics, and income differences.

Third, while we know that crime affects local population size (Cullen and Levitt, 1999) and local housing prices (Gibbons, 2004; Linden and Rockoff, 2008; Pope, 2008), there is little evidence on the response of consumers to spatial variation in crime rates.[4] Crime rates, fear of crime, and residential segregation are interrelated phenomena (O'Flaherty and Sethi, 2007, 2010). Specifically, the fraction of a neighborhood's population that is black is significantly associated with its residents' perception of a crime problem, even after controlling for neighborhood crime rates (Quillian and Pager, 2001), and white survey respondents more strongly associate racial composition with perceived risk (Quillian and Pager, 2010). Our data on consumer choices allow us to separately study the influences of ethnic and racial composition and criminal activity on the decisions of non-residents to patronize various neighborhoods. Our estimates demonstrate a cost of crime beyond its immediate effects on victims and residents in terms of diminishing the consumption value of the city for the broader population.

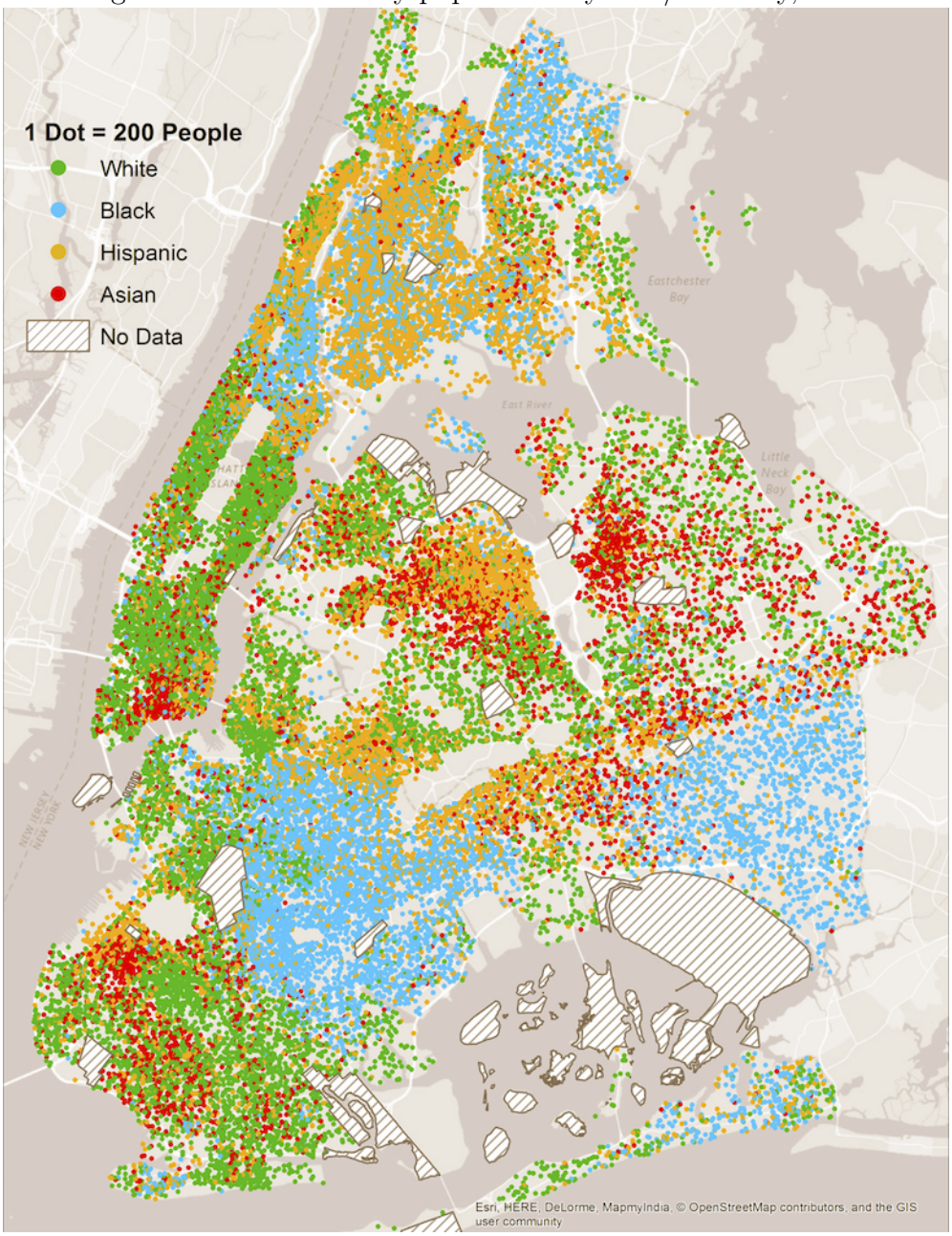
Finally, our results are also related to previous research on the interaction between gender and fear of crime (Doran and Burgess, 2011). A large majority of this literature is based on surveys such as the General Social Survey and the National Crime Survey (Ferraro, 1996). These studies uniformly find that fear of crime is higher among women (Pain (1991, p.416)). The surveys also indicate that a very large fraction of women adjust their behavior in regard to these fears, avoiding risky areas in ways that affect their social and leisure activities (Pain, 1997; Lorenc, Clayton, Neary, Whitehead, Petticrew, Thomson, Cummins, Sowden, and Renton, 2012). Our study complements these survey studies by using evidence coming

[2]Echenique and Fryer (2007) have also studied how race affects students' friendship networks within schools.

[3]This map was inspired by a *New York Times* project, "Mapping America: Every City, Every Block."

[4]There are consumer technologies devoted to this purpose. GPS devicemakers have developed technologies to incorporate crime statistics and demographic information when giving users navigational directions. Microsoft patented incorporating crime statistics into route calculation so as to direct the user "through neighborhoods with violent crime statistics below a certain threshold". A Manhattan-based smartphone app called "SketchFactor" crowdsourced opinions about the "sketchiness" of neighborhoods. The app came under criticism for being racist and is now defunct (*New Yorker*, July 29, 2015).

Figure 1: New York City population by race/ethnicity, 2010



NOTES: This figure depicts the residential NYC population in terms of four demographic categories that cover 97% of the population. Each dot represents 200 people. Tract-level population data from the 2010 Census of Population.

from actual consumption patterns of individuals. Additionally, a major advantage of our data is that we have information describing the venues under consideration, allowing us to control for these other determinants of demand when inferring how male and female users respond. We find that female users are particularly sensitive to spatial variation in crime rates.

The remainder of this paper proceeds as follows. Section 2 introduces the data and section 3 our empirical methodology. Section 4 reports our estimates of the roles of spatial and social frictions in the city.

2 Data

We combine data from individual Yelp users with information on New York City to describe how these individuals use the city.

2.1 Yelp data

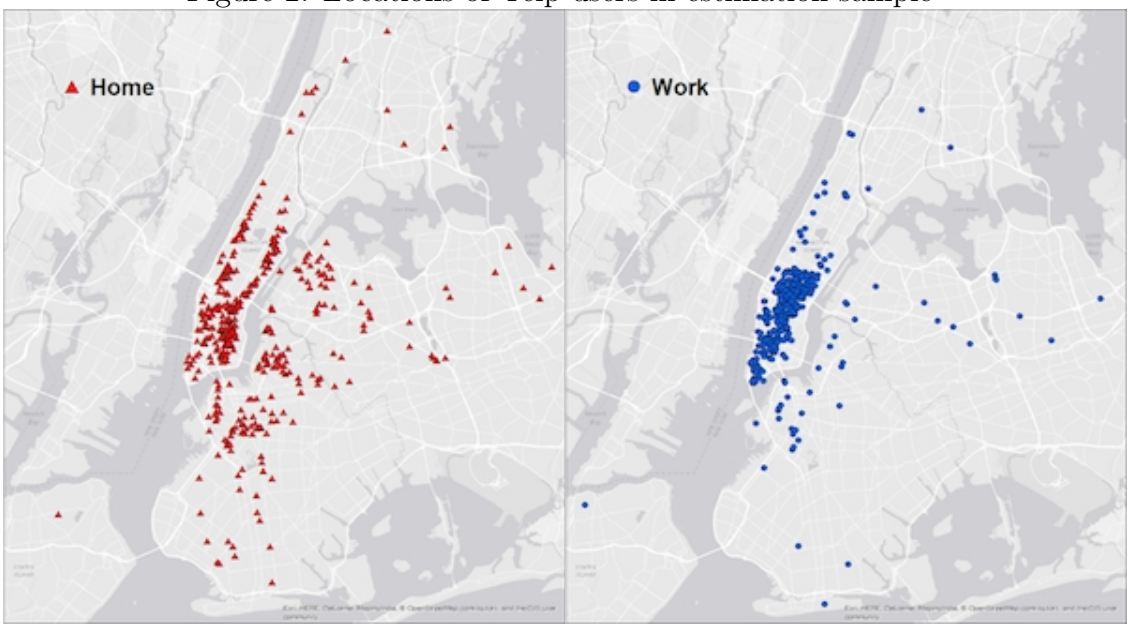
Yelp.com is a website where users review local businesses, primarily restaurants and retail stores (Yelp, 2013). The website describes a venue in terms of its address, hours of operation, average rating, user reviews, and a wide variety of other characteristics. Yelp's coverage of restaurants is close to comprehensive (see appendix section A.4). The website is relevant for the general population of restaurant consumers, as discontinuities in Yelp ratings have been shown to have substantive effects on restaurants' revenues (Luca, 2011) and reservation availability (Anderson and Magruder, 2012).

In addition to assigning a rating of one to five stars, users are encouraged to write reviews describing their personal experience with a business. These reviews vary greatly both in tone and length, ranging from a few sentences to many paragraphs. Crucial for our purposes is that users often disclose information in their reviews about their residential and work locations.[5] This provides us a novel description of the links between the consumption choices of individuals and a rich set of location- and venue-specific characteristics.

We examined Yelp users' reviews to identify their home and work locations. In mid-2011, we gathered data from the Yelp website for users who had reviewed venues in New York City. As described in detail in appendix A.3, we identified users' residential and work locations based on the text of reviews that contained at least one of 26 key phrases related to location, such as "close to me," "block away," and "my apartment." We classified whether the venue under review was proximate to the user's home and/or work locations. Using the set of venues associated with users' residential and work locations, we estimated the residential and work locations using the average of the relevant venues' latitude-longitude coordinates. Restricting our sample to users who do not reveal a change in residence or workplace within New York City and whose home and work locations can be assigned to census tracts with no missing covariates yielded an estimation sample containing 406 users who wrote 16,573

[5]Another example of using the information disclosed in reviews is searching Yelp to detect outbreaks of food poisoning unreported to NYC health authorities (Harrison, Jorder, Stern, Stavinsky, Reddy, Hanson, Waechter, Lowe, Gravano, and Balter, 2014).

Figure 2: Locations of Yelp users in estimation sample



NOTES: This figure depicts the distribution of home and work locations of the 406 users in our estimation sample.

reviews.[6]

Figure 2 depicts the home and work locations of the users in our estimation sample. Consistent with broader patterns, our identified users have a high concentration of employment in Manhattan below 59th Street and a more dispersed residential pattern.

Yelp users may also post information about themselves to their profile, such as a photo. In some estimation exercises, we use a user's apparent gender and ethnicity or race, inferred from their user photo, to examine how consumer behavior depends on the interaction of user characteristics and venue characteristics.[7] A previous study compared such observational measures of ethnicity and race inferred from photos with administrative data and found a high degree of accuracy in partitioning subjects into three groups: Asian, black, and white or Hispanic (Mayer and Puller, 2008).

Table 1 reports summary statistics for these users' gender, demographic, and locational information. Females constitute more than 60% of the users in our estimation sample.[8] Very few users were identified as blacks. Asians are overrepresented relative to their share of the general population, as they constitute about one-third of our identifiable users but only about 12% of New York City. The users in our estimation sample tend to live in census

[6]If we omit work locations and use only home locations, there are more than 1800 users with nearly 40,000 reviews meeting these criteria. See appendix C.2 for a discussion of how our results are altered if we fail to incorporate the workplace and commuting origins.

[7]While users may choose "male" or "female" for their gender on their Yelp profile, this information is not publicly displayed on their profile. All our empirical results describing differences between men and women in fact describe differences between users who were classified based on their gender presentation in their profile photo. Some profile photos do not present a gender (e.g. cartoon graphics, photos of animals).

[8]Users with profile photos for which we could not classify the gender have both male and female dummy variables equal to zero.

Table 1: User summary statistics

Variable	Mean	Std. Dev.
Number of restaurant reviews in estimation sample	40.95	42
User appears female in profile photo	0.61	0.49
User appears male in profile photo	0.34	0.47
User appears white or Hispanic in profile photo	0.44	0.5
User appears Asian in profile photo	0.26	0.44
User appears black in profile photo	0.03	0.16
User race/ethnicity indeterminate in profile photo	0.27	0.45
Median household income of home census tract (thousands dollars)	77.07	33.19
Share of home census tract population that is age 21-39	0.43	0.11

NOTES: This table describes the 406 Yelp users in our estimation sample. Census tract income data from 2007-2011 American Community Survey and demographic data from 2010 Census of Population.

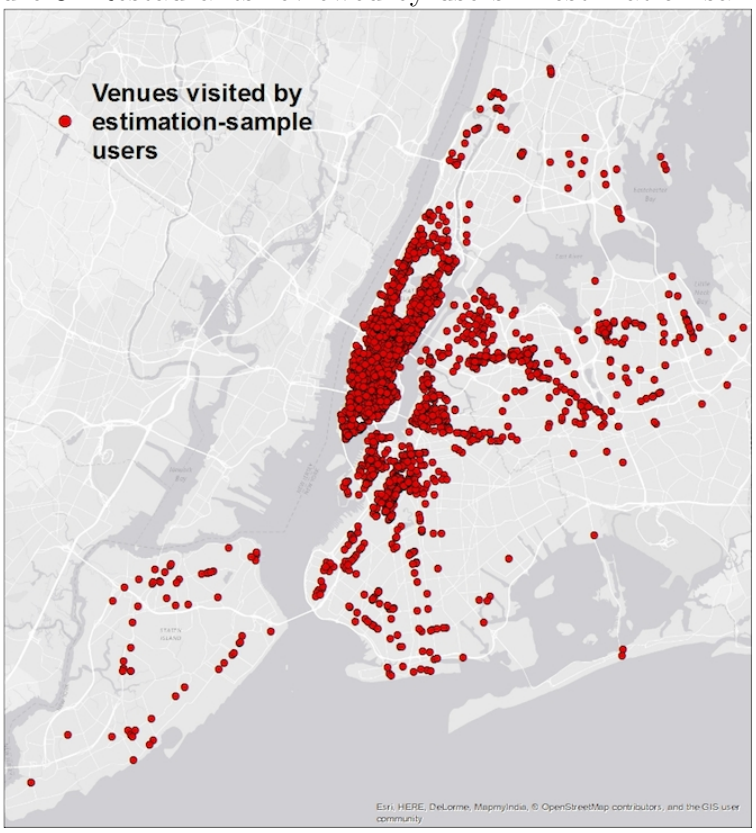
tracts with median incomes typical of Manhattan but higher than typical of New York City as a whole. The average of tract median household income in our estimation sample is near $77,000, while the city-wide mean is about $56,000. The users in our estimation sample tend to live in census tracts with a share of the population between the ages of 21 and 39 (43%) that is higher than both Manhattan (37%) and New York City as a whole (30%). These patterns are consistent with statements that Yelp's global user base tends to be younger, higher-income, and more educated than the population as a whole (Yelp, 2013).

The 406 users in our estimation sample posted more than 16,000 reviews of NYC restaurants between 1 January 2005 and 14 June 2011. Figure 3 displays all the restaurants reviewed by the users in our estimation sample. Unsurprisingly, these venues are heavily concentrated in lower Manhattan, but users in our estimation sample have reviewed venues in many parts of New York City.

Table 2 summarizes the distribution of reviews in terms of venues' prices, ratings, and boroughs. The first two columns describe the estimation sample in terms of the number and share of reviews. The third column reports these frequencies for all reviews of NYC Yelp restaurants in our data, constituting more than 700,000 reviews. Comparing the second and third columns of Table 2 shows that users in our estimation sample exhibit review frequencies similar to that of the broader Yelp population. The frequencies of reviews across restaurants' prices and ratings are very similar in columns two and three. This is consistent with the hypothesis that users whose residences and workplaces we located based on the text of their reviews are similar to the broader population of Yelp users in their restaurant-going behavior. Users in our estimation sample review Manhattan venues slightly more than the population of Yelp as a whole.

Figure 4 maps these data for two individuals in our estimation sample. In each panel, red dots denote Yelp venues reviewed by this user. The "H" denotes the average coordinates of those venues identified as residential locations in the text of this user's reviews. The "W" denotes the similarly defined work location. The user in the left panel lives and works in midtown Manhattan. The other user works in midtown Manhattan and resides in a south-eastern Manhattan development called Stuyvesant Town. At a glance, the maps suggests

Figure 3: Restaurants reviewed by users in estimation sample



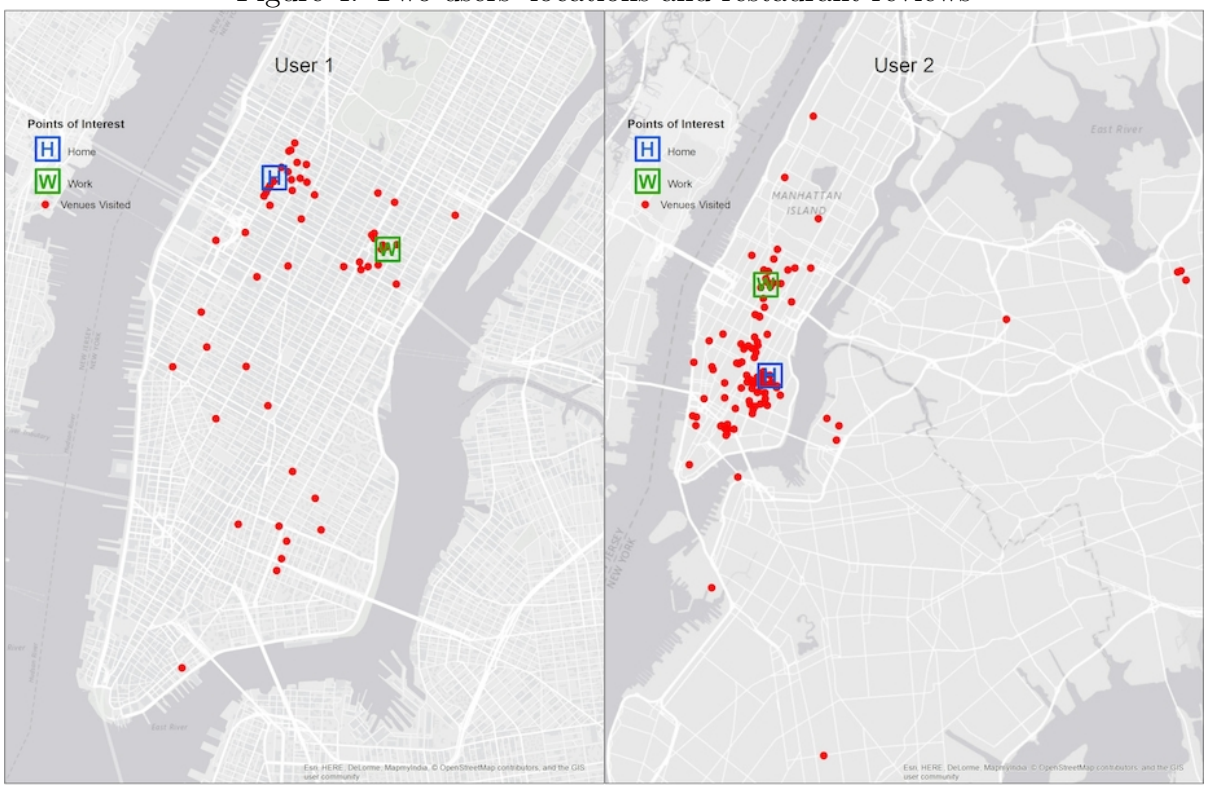
NOTES: This map depicts the locations of 4993 Yelp restaurant venues reviewed by users in our estimation sample. Each dot represents a venue.

Table 2: Venue review summary statistics

	Estimation sample		Share of
Restaurant characteristic	Observations	Share of reviews	all Yelp reviews
Price of $	3875	.233	.228
Price of $$	9345	.562	.566
Price of $$$	2745	.165	.161
Price of $$$$	659	.040	.045
Rating of 1 stars	9	.001	.001
Rating of 1.5 stars	32	.002	.002
Rating of 2 stars	120	.007	.011
Rating of 2.5 stars	590	.035	.038
Rating of 3 stars	2321	.140	.140
Rating of 3.5 stars	6202	.373	.359
Rating of 4 stars	6499	.391	.390
Rating of 4.5 stars	834	.050	.056
Rating of 5 stars	17	.001	.003
Located in Manhattan	13646	.821	.749
Located in Brooklyn	1862	.112	.169
Located in Queens	1003	.060	.069
Located in Bronx	52	.003	.008
Located in Staten Island	61	.004	.005

NOTES: This table summarizes the distribution of reviews across different venue characteristics in both our estimation sample (columns 1 and 2) and all Yelp users as a whole (column 3). Comparing columns 2 and 3 shows that the review behavior of users in our estimation sample is similar to that of Yelp users as a whole.

Figure 4: Two users' locations and restaurant reviews



NOTES: These two maps display two users' home and work locations and Yelp restaurant venues reviewed.

that both proximity and venue characteristics influence user behavior. Both users primarily review venues that are near their home or work locations. Both users visit more downtown venues than uptown venues, which may reflect differences in the quantity or quality of venues in these areas.

2.2 NYC crime, demographic, and transportation data

We combine the information from Yelp users' reviews with data describing locations' residential racial or ethnic composition, income levels, crime rates, and estimates of the time required to travel between locations.

Much of the demographic and income information is reported at the level of census tracts. Census tracts are geographic units defined by the US Census Bureau based on population. We use data from the 2010 Census of Population to describe each tract's residential racial/ethnic composition in terms of Asians, blacks, Hispanics, and whites.[9] These population counts are depicted in Figure 1. The data on median household incomes come from the 2007-2011 American Community Survey 5-Year Estimate. These tract-level characteristics

[9]To be precise, we divide the population into five racial/ethnic groups and use the population counts of non-Hispanic whites, non-Hispanic blacks, non-Hispanic Asians, and all Hispanics. The remainder, which includes Native Americans, Hawaiians, other races, and mixed-race categories, constitutes about three percent of the NYC population.

Table 3: Tract-level summary statistics

	Manhattan		New York City	
Variable	Mean	Std. Dev.	Mean	Std. Dev.
Population	5677	2993	3866	2115
Land area (square kilometers)	0.182	0.092	0.324	0.516
Median household income (thousands dollars)	76.339	44.344	56.292	27.152
Share of tract population that is Asian	0.117	0.129	0.125	0.154
Share of tract population that is black	0.142	0.195	0.245	0.297
Share of tract population that is Hispanic	0.232	0.228	0.265	0.222
Share of tract population that is white	0.484	0.302	0.335	0.31
Male share of census tract population	0.475	0.039	0.476	0.032
Share of census tract population that is age 21-39	0.374	0.117	0.302	0.084
Spectral segregation index for tract's plurality	0.171	0.354	0.914	2.394
Average annual robberies per resident in tract, 2007-2011	0.005	0.015	0.003	0.009
Tract is plurality Asian	0.032	0.177	0.082	0.274
Tract is plurality black	0.125	0.332	0.251	0.434
Tract is plurality Hispanic	0.215	0.412	0.246	0.431
Tract is plurality white	0.627	0.484	0.421	0.494
Number of observations	279		2110	

NOTES: This table describes 2010 NYC census tracts for which an estimate of median household income is available. Data on incomes from 2007-2011 American Community Survey, demographics from 2010 Census of Population, robberies from NYPD.

are summarized in Table 3.

To describe crime rates, we computed tract-level robbery statistics for 2007-2011 using confidential, geocoded incident-level reports provided to us by the New York Police Department.[10] We use robberies as our crime measure because these are the most common and relevant threat to individuals visiting a Yelp venue.[11]

To measure segregation, we calculate the Echenique and Fryer (2007) spectral segregation index (SSI) for the modal race/ethnicity in each census tract. This index measures the degree to which a census tract borders census tracts of the same demographic plurality, and the further degree to which those tracts themselves border tracts of the same plurality, *ad infinitum*.[12] This captures the idea that residents at the center of demographically homogeneous area are more segregated than those near the edge. For example, in Figure 1, the black census tracts at the center of the cluster of blue dots on the right edge of the map will have higher SSI values than those at the edge of the cluster.

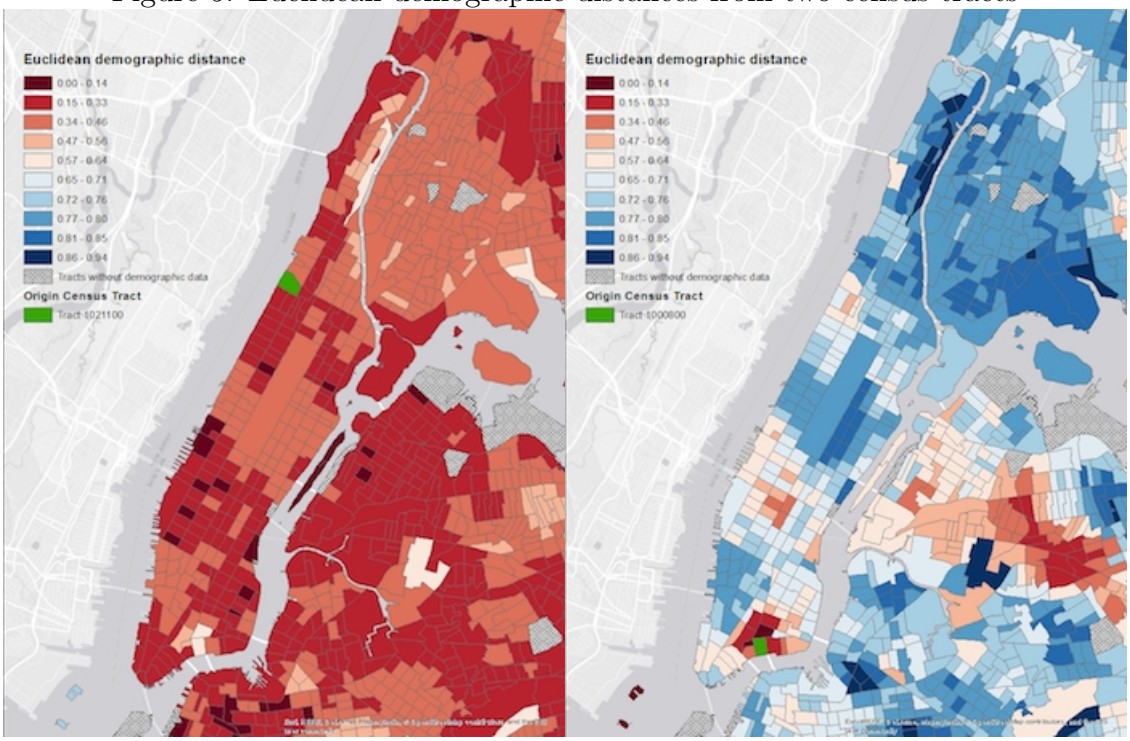
Table 4 presents summary statistics for pairs of census tracts. To measure racial/ethnic

[10]The 2007-2011 timespan matches the years covered by the American Community Survey data. Fewer than two percent of the Yelp reviews in our estimation sample were posted in 2005-2006.

[11]By contrast, burglaries are more relevant for residents than visitors; rapes, assaults and murders are certainly of real interest as potential fears, but their measures are variably polluted by the fact that the attacker may frequently or even predominantly be known to the victim, so these are potentially not clean measures of the threats that deters visits to venues.

[12]More formally, a census tract is a member of a network of tracts of the same demographic plurality that are connected to another member. On this connected component, the SSI is the largest eigenvalue of the irreducible submatrix of the fraction of neighboring tracts that are of the same plurality. See Echenique and Fryer (2007).

Figure 5: Euclidean demographic distances from two census tracts



NOTES: These maps depict Euclidean demographic distances from an origin tract to other tracts in NYC. In the left panel, the origin tract is in Morningside Heights; in the right panel, Manhattan's Chinatown. Demographic data from 2010 Census of Population.

demographic distances between two tracts, we calculate the Euclidean distance between the two tract's population shares for five demographic groups. Stacking these population shares in a five-element vector $shares_{tract}$, the "Euclidean demographic distance" between origin and destination tracts is

$$\|shares_{origin} - shares_{destination}\|/\sqrt{2},$$

where $\|\cdot\|$ indicates the L^2 norm. This measure can range from zero to one. Figure 5 illustrates the Euclidean demographic distance for two census tracts, one in Morningside Heights with a diverse population that is similar to most tracts and one in Manhattan Chinatown that is overwhelmingly Asian and thus quite demographically distant from most tracts.

To estimate travel times, we requested the public-transit and automobile travel times between the centroids of New York City census tracts from Google Maps.[13] In addition to direct travel to a restaurant from home or work, we consider the additional travel time to a Yelp venue that would be incurred by a user incorporating a visit to the venue as part of her commuting path rather than commuting directly between home and work. Denote the travel time from location x to location y by $time(x, y)$. For user i living in h_i and working in w_i, the travel time associated with visiting venue j from her commuting path p_i is computed

[13]Using tract-to-tract travel times keeps the computational burden under 10 million instances.

Table 4: Tract-pair summary statistics

Variable	Mean	Std. Dev.
Percentage absolute difference in median household income	0.506	0.355
Percentage difference in median household income	0	0.618
Euclidean demographic distance between tracts	0.455	0.226
Travel time by public transport in minutes	72.436	30.319
Travel time by automobile in minutes	24.937	10.589

NOTES: This table describes 4,452,012 pairs of 2010 NYC census tracts for which estimates of median household income and travel times are available. Data on incomes from 2007-2011 American Community Survey, demographics from 2010 Census of Population, and travel times from Google Maps.

Table 5: Female vs male users in estimation sample

Covariate	Female mean	Male mean	p-value for difference
Percent white, home	.573	.559	.547
Percent black, home	.052	.092	.007
Percent Asian, home	.162	.144	.233
Percent Hispanic, home	.182	.177	.751
Percent white, work	.630	.642	.528
Percent black, work	.052	.054	.821
Percent Asian, work	.182	.177	.742
Percent Hispanic, work	.107	.099	.429
Average annual robberies per resident, 2007-2011, home	.004	.003	.560
Average annual robberies per resident, 2007-2011, work	.016	.015	.944
Median household income (thousands dollars), home	7.87	7.55	.360
Median household income (thousands dollars), work	9.86	9.88	.963

NOTES: This table reports summary statistics for the home and work tracts of users in our estimation sample, distinguishing between male and female users. The last column reports the p-value from a t-test for differences in means between female and male users.

as

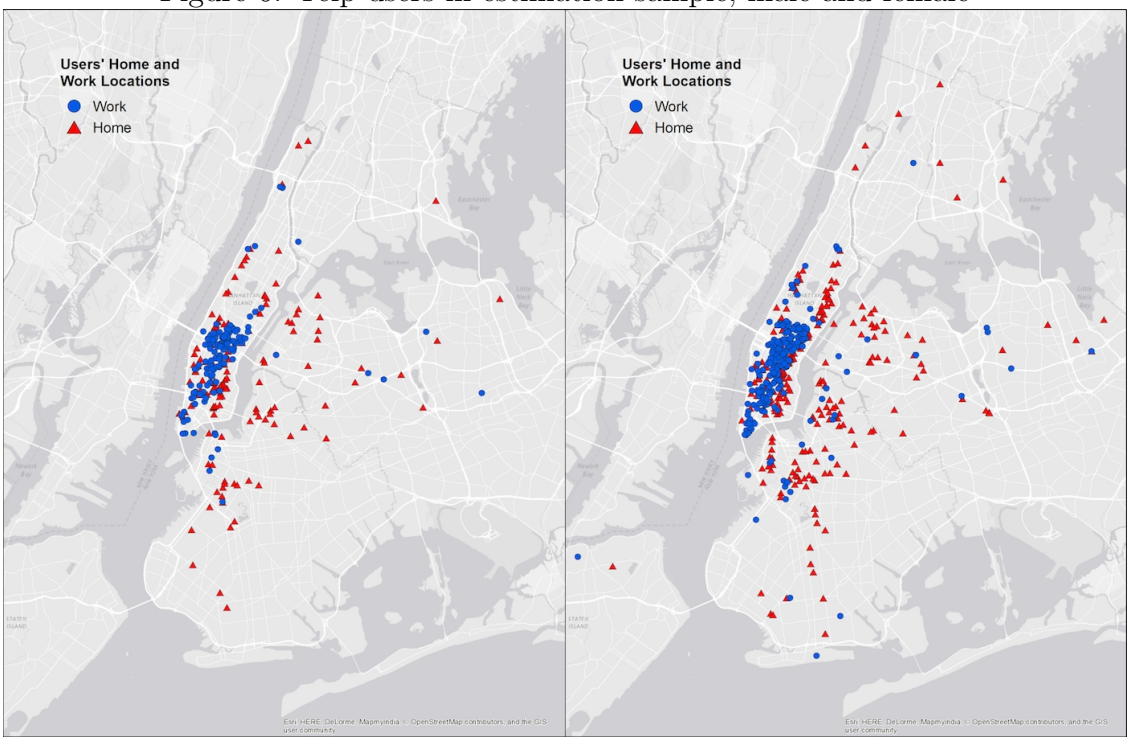
$$time(p_{i,}j) = \frac{1}{2}\max\left\{time(h_i, j) + time(w_i, j) - time(h_i, w_i), 0\right\}.$$

2.3 Gender balance

Since a number of our results describe how men and women use the city differently, in this section we describe the locations of male and females users in our estimation sample.

Figure 6 depicts the home and work locations of the users in our estimation sample akin to Figure 2, separating users by gender. Our estimation sample is geographically balanced across genders in most respects. Table 5 compares the average characteristics of female and male users' residential and workplace locations. For most characteristics, male and female users' locations don't exhibit statistically significant differences. However, male users tend to reside in census tracts that have a higher black population share.

Figure 6: Yelp users in estimation sample, male and female



NOTES: The left panel depicts the residential and workplace locations of Yelp users in our estimation sample who are identified as male; the right panel depicts females.

2.4 Observed behavior and frictions

Before introducing our behavioral model, we present moments from our data suggesting the spatial and social frictions that we investigate. It also illustrates the necessity of explicitly modeling consumer decisions in order to separately identify the roles of various frictions.

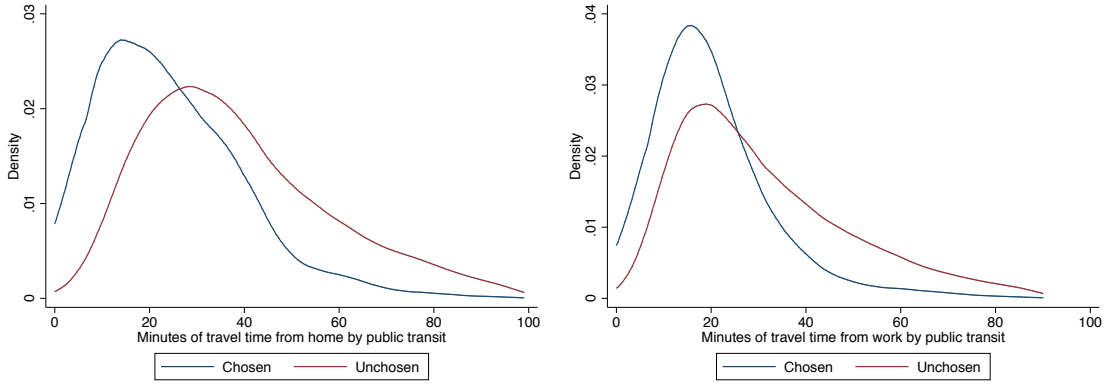
The data are consistent with the hypothesis that travel time plays a significant role in consumption decisions. The two users whose behavior was depicted in Figure 4 tended to consume near their home and work locations. This pattern is exhibited by the hundreds of users in our estimation sample. Figure 7 plots the density of travel times for all the users in our estimation sample, comparing the travel time to venues that were chosen by these users to a randomly selected sample of venues that were not chosen. Yelp users are far more likely to visit venues that are closer to their residential and workplace locations.[14]

The data are also consistent with a role for demographic differences in explaining users' consumption patterns. The left panel of Figure 8 plots the density of Euclidean demographic distances for our estimation sample, comparing venues chosen by these users to a randomly selected sample of unchosen venues. This plot shows that Yelp users are more likely to visit venues located in tracts with demographics more similar to those of their home tract.

The right panel of Figure 8 plots the analogous densities in terms of robberies per resident in the venue's census tract. Visually, the plot reveals little difference between the density of

[14]The fact that unchosen venues have shorter travel times to work than home likely reflects the fact that many venues are in Manhattan and most workplaces are in Manhattan.

Figure 7: Travel time and consumer choice



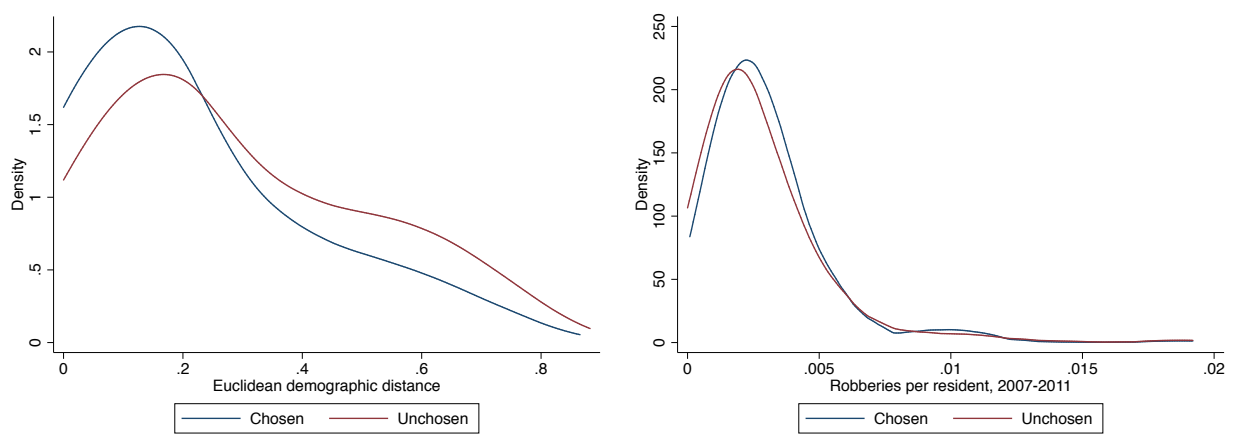
NOTES: These plots are kernel densities for two distributions of user-venue pairs: those venues chosen by users in our estimation sample and a random sample of venues not chosen by these users. The left panel plots the densities as functions of travel time from home by public transit; the right panel from work. Epanechnikov kernel with bandwidth of 3.

robberies per resident for venues chosen by the individuals in our sample and that density for a randomly selected sample of the venues that were not chosen. If anything, the venues chosen have slightly more robberies per resident. If higher crime rates deter potential consumers, their influence is masked by other sources of variation in consumer decisions.

While these density plots are consistent with the idea that both spatial and social frictions may influence consumers' decisions, they have limited capacity to identify the effect that these factors have on consumers' choices. Each density plot neglects the influence of both the other frictions and venue characteristics on consumption decisions. For example, given significant residential segregation in NYC (i.e. census tracts with similar demographics tend to be clustered together), measures of travel time are positively correlated with measures of demographic differences. Therefore, from Figures 7 and 8, it is not possible to quantify the relative influence of spatial and demographic distance. Similarly, robberies presumably occur in places that people choose to visit and, therefore, crime rates may be positively correlated with locational characteristics that attract consumers. If this is the case, the right panel in Figure 8 would understate any negative impact of robberies on consumer visits.

In order to address these concerns and quantify the impact of various frictions on consumers' decisions, in the next section we introduce a discrete-choice model that accounts for a large set of potential determinants of consumers' demand for restaurants.

Figure 8: Demographic differences, robberies, and consumer choice



NOTES: These figures plot kernel densities for two distributions of user-venue pairs: those venues chosen by users in our estimation sample and a random sample of venues not chosen by these users. The left panel plots the densities as functions of Euclidean demographic distances; the right robberies per resident. Epanechnikov kernel with bandwidths of 0.1 and 0.001, respectively. The right panel excludes seven census tracts with robberies per resident between 0.02 and 0.4.

3 Empirical approach

We estimate a standard discrete-choice model using data on Yelp users' choices to identify the parameters governing their demand function for restaurant venues. Individuals take repeated decisions about where to eat: they must choose whether to visit any venue and, if they do, which venue to visit. We index individuals by i, venues by j, and we index by t the occasions in which i needs to decide on whether to visit a venue. We denote the set of potential choices at period t as J_t. We denote the outside option of not visiting any venue as $j = 0$, and we assume that it belongs to every set J_t.

3.1 Demand Specification

When visiting a venue, individuals choose whether to visit it from home, work, or their commuting path, and choose whether to travel via public transit or car. We index pairs of origin locations and transportation modes by l and assume that a realized trip to a venue may be one of six types: from home via car ($l = hc$), from home via public transit ($l = hp$), from work via car ($l = wc$), from work via public transit ($l = wp$), from their commuting path via car ($l = pc$), or from their commuting path via public transit ($l = pp$). We denote the set of these six potential origin-mode pairs as $\mathcal{L} \equiv \{hc, hp, wc, wp, pc, pp\}$.

We adopt the standard random-utility representation of preferences and assume that the utility for individual i of visiting venue j in period t from origin-mode l may be written as

$$U_{ijlt} = \beta_l^1 X_{ijl}^1 + \beta^2 X_{ij}^2 + \nu_{ijlt} \tag{1}$$

where X_{ijl}^1 denotes covariates observable to the econometrician that vary by origin and transportation mode, while X_{ij}^2 denotes covariates that do not vary by origin-mode. Note

that we allow the coefficient on X^1_{ijl}, β^1_l, to vary flexibly with the pair of origin locations and transportation modes indexed by l. The variable ν is a scalar that is unobserved to the econometrician. We assume that the utility attached to the outside option of staying home is $U_{i0lt} = \nu_{i0lt}$.

In our empirical specification, X^1_{ijl} is the log minutes it would take individual i to travel to restaurant j using the transportation mode from the origin indexed by l. The fact that β^1_l is l-specific allows the disutility of travel time to depend on both whether the trip originates from home, work, or the commuting path, and whether the individual is traveling via public transit or automobile (private car or taxi). These disutilities might differ because the direct pecuniary cost of an additional minute of travel time might be different across modes of transportation (positive for the case of taxi, zero in the case of subway). Similarly, the disutility might be different because the opportunity cost of additional time spent traveling might be different when an individual is leaving from work and returning to work afterwards than when it is leaving from home. Visits from home may be likely to happen on weekends or evenings and, therefore, the opportunity cost of traveling farther is related to the marginal value of leisure time. Conversely, trips from work that are not linked to commuting are likely to happen on weekdays and in the middle of the workday and, therefore, the opportunity cost of traveling is related to the marginal value of work time. By allowing the coefficient on travel time to differ across the six different potential origin-mode pairs included in the set $\mathcal{L}$, we allow, among many other potential sources of heterogeneity in the value of time, for the marginal value of leisure time to differ from the marginal value of work time.

The vector X^2_{ij} includes a broad set of characteristics of the venue, indexed by j, and characteristics of the census tract in which the venue is located, which we denote k_j. The fact that β^2 is common across origin-mode pairs indexed by l means that we assume that, for example, a user's utility from going to a restaurant that has a high Yelp rating rather than a restaurant that has a low Yelp rating does not depend on the trip's origin nor whether the user employs public transportation or an automobile to reach the restaurant.

Although our data describes users' home and work locations, it does not identify the exact origin of each trip to a restaurant. We address this shortcoming by assuming that consumers jointly optimize over the restaurant they patronize and the origin from which they do so.[15]Thus, individuals optimally choose the restaurant-origin-mode combination jl that maximizes their utility. Accordingly, we define a variable d_{ijlt} that takes value 1 if individual i chooses to travel to venue j from origin-mode l at period t:

$$d_{ijlt} = \mathbb{1}\{U_{ijlt} \geq U_{ij'l't}; \forall\ j' \in J_t,\ l' \in \mathcal{L}\},$$

where $\mathbb{1}\{A\}$ is an indicator function taking value 1 if A is true. We also define a variable d_{ijt} that takes value 1 if individual i chooses alternative j at period t:

$$d_{ijt} = \sum_{l \in \mathcal{L}} d_{ijlt}.$$

[15]If we were to observe the origin of the trip and assume that it were exogenously determined, we could condition on this additional information when estimating. If we think that the origin of the trip is determined endogenously, then, even if we were to observe the origin of each trip, we would have to impose assumptions on individuals' joint optimization over restaurants and origins.

In order to estimate our parameters of interest, we assume that the vector of unobserved utilities for individual i at period t, $\nu_{it} = (\nu_{ijlt}; \forall\, j \in J_t,\, l \in \mathcal{L})$, is independent across individuals and time periods and has a joint extreme value type I cumulative distribution function:

$$F(\nu_{it}) = \exp\Big(-\sum_{j=1}^{J_t} \Big(\sum_{l \in \mathcal{L}} \exp(-\nu_{ijlt}) \Big) \Big). \tag{2}$$

This distribution yields a standard multinomial logit discrete-choice model. We make this assumption on the functional form because it allows us to handle the large number of venues in the choice set J_t. Assuming a logit error term means that the resulting choice probabilities exhibit the independence of irrelevant alternatives property. That is, the probability that an individual i at period t chooses venue j relative to the probability that she chooses venue j' does not depend on the characteristics of all the restaurants other than j and j' that are included in the choice set J_t. This implies that we can identify the vector of preference parameters β_l^1 and β^2 simply by comparing the frequency with which individuals choose restaurants among an arbitrary subset of all the restaurants from which they might potentially choose from. This property of the multinomial logit model was first described by McFadden (1978).

Given the distributional assumption in equation (2), the probability that individual i visits venue j from origin-mode l at period t is

$$P(d_{ijlt} = 1 | X_{it}; \beta) = \frac{\exp(V_{ijlt})}{\sum_{j' \in J_t} \Big(\sum_{l' \in \mathcal{L}} \exp(V_{ijl't}) \Big)},$$

with $X_{it} = \{X_{ijl}; \forall j \in J_t, l \in \mathcal{L}\}$, $X_{ijl} = (X_{ijl}^1, X_{ij}^2)$, and $\beta = (\{\beta_l^1; \forall l \in \mathcal{L}\}, \beta^2)$. The probability that individual i visits venue j at period t is simply the sum of these probabilities across all possible origin-mode pairs that individual i might use to visit venue j at t

$$P(d_{ijt} = 1 | X_{it}; \beta) = \sum_{l \in \mathcal{L}} P(d_{ijlt} = 1 | X_{it}, ; \beta) = \frac{\Big(\sum_{l \in \mathcal{L}} \exp(V_{ijlt}) \Big)}{\sum_{j' \in J_t} \Big(\sum_{l \in \mathcal{L}} \exp(V_{ijl't}) \Big)}. \tag{3}$$

3.2 Estimation

There are two reasons why we cannot directly use the choice probability in equation (3) to build a likelihood function that we may use to identify the parameter vector β. First, we observe only the reviews posted by Yelp users, not all visits to restaurants (i.e. we do not observe the value of d_{ijt} for every i, j, and t). Second, the cardinality of the choice set J_t is such that it is computationally infeasible to compute the denominator of the choice probability in equation (3). We explain in sequence how we solve these two problems.

The fact that we observe a sample of reviews posted on Yelp, rather than information on all visits from a random sample of the population of interest, implies that we need to impose some assumptions on how Yelp users write reviews. Specifically, we assume that (a) Yelp users do not write reviews for restaurants they have not visited; (b) Yelp users only write

reviews once per restaurant (independently of how many times they visit a restaurant); and (c) the probability that an individual writes a review is independent of *ex ante* restaurant characteristics (the characteristics that determine agents' dining choices; i.e. independent of X_{it}). Denoting by d^r_{ijt} a dummy variable taking value 1 when individual i writes a review on Yelp about a venue j visited at period t, assumptions (a) to (c) allows us to write the probability that we observe a review about j by i at t as

$$\begin{aligned} P(d^r_{ijt} = 1|X_{it};\beta) &= P(d^r_{ijt} = 1|d_{ijt} = 1, X_{it};\beta) \times P(d_{ijt} = 1|X_{it};\beta) \\ &= w_{it} \times \mathbb{1}\{j \neq 0, j \neq D^r_{it}\} \times P(d_{ijt} = 1|X_{it};\beta), \end{aligned} \tag{4}$$

The first equality relies on the assumption that individuals only write reviews about restaurants they actually visited. The second equality assumes that the probability that individual i writes a review about a choice j at period t is: (a) equal to zero when such choice was the outside option or belongs to the set of venues previously reviewed by individual i, denoted as D^r_{it}; (b) equal to an individual-time pair specific constant w_{it} for visits to restaurants not previously reviewed.

The assumption that review probabilities are independent of restaurant characteristics may seem implausible under some circumstances. First, one could claim that individuals are more likely to write reviews about dining experiences in which they were greatly surprised – either negatively or positively. However, this will not bias our estimates of the preference parameter vector β. The reason is that surprises are, by definition, independent of the variables that are in the information set of consumers when deciding which restaurant venue to patronize. Given that choices, by definition, can only be a function of variables in these information sets, it must be that all the restaurant characteristics included in the vector X^2_{ij} are included in such information sets and, therefore, independent of whatever variable caused the user's surprise.

Second, one could claim that users are more likely to write reviews of restaurants that have a small number of reviews, that do not already have a reputation well-known by most consumers, or that users want to signal they have patronized. Call this the "McDonald's" review pattern: conditional on having visited the corresponding restaurant, the probability that a Yelp user writes a review of a McDonald's is much lower than the probability that she writes a review of a non-chain restaurant. As Appendix B.1 shows, this behavior will only bias the estimates of the β coefficients on those characteristics that are both included in the vector X^2_{ij} and influence the review-writing probabilities of Yelp users. Specifically, as long as the probability that an individual writes a review of a restaurant does not depend on our measures of spatial or social frictions, conditional on the vector of restaurant characteristics X^2_{ij}, the estimates of the parameters characterizing consumers' responses to these frictions will not be biased.

The second reason why we cannot use the choice probability in equation (3) to build a likelihood function to identify the parameter vectors β is that the cardinality of the choice set J_t makes it computationally infeasible to construct the denominator of the choice probability in equation (3). McFadden (1978) and Train, McFadden, and Ben-Akiva (1987) show that, in a multinomial logit model, one can consistently estimate the vector of preference parameters β even if the researcher does not correctly specify the consumer's choice set J_t. In our case, while the probability of choosing a restaurant in equation (3) is generated by a multinomial

logit, the term $w_{it} \times \mathbb{1}\{j \neq 0, j \neq D^r_{it}\}$ implies that the probability of observing a review – i.e. equation (4) – is not exactly that generated by a multinomial logit. However, as we show here, one can use the logic in McFadden (1978) to obtain a likelihood function that does not depend on the actual choice set J_t and that will correctly identify the vector β.

For each individual i and time period t, we denote by J'_{it} the subset of restaurants in agent i's consideration set at period t that she has not previously reviewed (i.e $J'_{it} = J_t / \{D^r_{it} \cup \{j = 0\}\}$).[16] We define a set S_{it} that is a subset of the true consideration set minus the previously reviewed restaurants, J'_{it}. Following McFadden (1978) and Train, McFadden, and Ben-Akiva (1987), we construct S_{it} by including i's observed choice at period t plus a random subset of the other alternatives included in the set J'_{it}. Each of the choices included in S_{it} other than the actual choice of i at t are selected from J'_{it} with equal probability.[17]. As all elements of the set S_{it} other than the actual choice of i at t are selected randomly, the set S_{it} is a random variable. We denote by $\pi(S_{it}|d^r_{ijt} = 1)$ the probability of assigning the subset S_{it} to an individual i at period t who wrote a review about venue j. Our sampling scheme implies that

$$\pi(S_{it}|d^r_{ij} = 1) = \begin{cases} \kappa & \text{if } j \in S_{it}, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where κ is a constant such that $\kappa \in (0, 1)$. Therefore, the conditional probability of an individual i writing a review about venue j at period t, given a sample S_{it} randomly drawn by the econometrician, is:

$$\begin{aligned} P(d^r_{ijt} = 1|X_{it}, S_{it}; \beta) &= \frac{P(S_{it}|d^r_{ijt} = 1, X_{it})P(d^r_{ijt} = 1|X_{it}; \beta)}{\sum_{j' \in J_t} P(S_{it}|d^r_{ij't} = 1, X_{it})P(d^r_{ij't} = 1|X_{it}; \beta)}, \\ &= \frac{\pi(S_{it}|d^r_{ijt} = 1)P(d^r_{ijt} = 1|X_{it}; \beta)}{\sum_{j' \in J_t} \pi(S_{it}|d^r_{ij't} = 1)P(d^r_{ij't} = 1|X_{it}; \beta)}, \\ &= \frac{\pi(S_{it}|d^r_{ijt} = 1)P(d^r_{ijt} = 1|X_{it}; \beta)}{\sum_{j' \in S_{it}} \pi(S_{it}|d^r_{ij't} = 1)P(d^r_{ij't} = 1|X_{it}; \beta)}, \\ &= \frac{\kappa P(d^r_{ijt} = 1|X_{it}; \beta)}{\sum_{j' \in S_{it}} \kappa P(d^r_{ij't} = 1|X_{it}; \beta)}, \\ &= \frac{P(d^r_{ijt} = 1|X_{it}; \beta)}{\sum_{j' \in S_{it}} P(d^r_{ij't} = 1|X_{it}; \beta)}, \end{aligned} \tag{6}$$

as long as $j \in S_{it}$, and 0 otherwise. The first equality comes by applying Bayes' rule. The second equality accounts for the fact that our procedure to draw the samples of venues S_{it} does not depend on the restaurant characteristics, X_{it}, once we condition on the observed review of individual i at period t. Finally, the third, fourth and fifth equalities are implied

[16] In our empirical application, we assume that J_t is the set of all restaurants listed on Yelp, located in NYC, and for which information on characteristics such as the price and Yelp rating is available.

[17] This assignment mechanism satisfies the positive conditioning property (see McFadden 1978).

by equation (5). Combining equations (3), (4), and (6), we obtain that, for every $j \in S_{it}$

$$\begin{aligned} P(d^r_{ijt} = 1 | X_{it}, S_{it}; \beta) &= \frac{w_{it} \mathbb{1}\{j \neq 0, j \neq D^r_{it}\} \Big(\sum_{l \in \mathcal{L}} \exp(V_{ijlt}) \Big)}{\sum_{j' \in S_{it}} \Big\{ w_{it} \mathbb{1}\{j \neq 0, j \neq D^r_{it}\} \Big(\sum_{l \in \mathcal{L}} \exp(V_{ij'lt}) \Big) \Big\}} \\ &= \frac{\mathbb{1}\{j \neq 0, j \neq D^r_{it}\} \Big(\sum_{l \in \mathcal{L}} \exp(V_{ijlt}) \Big)}{\sum_{j' \in S_{it}} \Big\{ \mathbb{1}\{j \neq 0, j \neq D^r_{it}\} \Big(\sum_{l \in \mathcal{L}} \exp(V_{ij'lt}) \Big) \Big\}} \\ &= \frac{\Big(\sum_{l \in \mathcal{L}} \exp(V_{ijlt}) \Big)}{\sum_{j' \in S_{it}} \Big\{ \Big(\sum_{l \in \mathcal{L}} \exp(V_{ij'lt}) \Big) \Big\}} \end{aligned} \tag{7}$$

where the second equality divides by w_{it} –the probability that user i writes a review at t– in the numerator and denominator; and the third equality takes into account that $S_{it} \in J'_{it}$, and, therefore, $\mathbb{1}\{j \neq 0, j \neq D^r_{it}\} = 1$ for all elements of the set S_{it}. If we had not been careful to draw the random sets S_{it} from the subset of venues that have not been previously reviewed by each consumer i, then it would be possible that the indicator function $\mathbb{1}\{j \neq 0, j \neq D^r_{it}\}$ takes value 0 for one of the restaurants included in S_{it}. In this case, we would have to keep the term $\mathbb{1}\{j \neq 0, j \neq D^r_{it}\}$ in the denominator and keep track when constructing the probability $P(d^r_{ijt} = 1 | X_{it}, S_{it}; \beta)$ of which of the restaurants in S_{it} have previously been reviewed by each user. It is therefore only for computational simplicity that we draw the set S_{it} from the set of restaurants never previously reviewed by i, J'_{it}.

Using the last expression in equation (7), we define our log-likelihood function as

$$L = \sum_i \sum_t \sum_{j \in S_{it}} \mathbb{1}\{d^r_{ijt} = 1\} \ln \left(\frac{\sum_{l \in \mathcal{L}} \exp(V_{ijlt})}{\sum_{j' \in S_{it}} \Big\{ \sum_{l \in \mathcal{L}} \exp(V_{ij'lt}) \Big\}} \right).$$

As intended, this log-likelihood function is defined in terms of the probability that an individual i writes a review about restaurant j at period t and does not depend on the actual choice set J_t.

3.3 Identification Concerns

The estimation procedure described in sections 3.1 and 3.2 imposes two key identification assumptions: (1) absence of unobserved heterogeneity in individuals' valuations of observable characteristics, and (2) exogeneity of home and work locations. In this section, we discuss why we impose these assumptions and what they imply in our empirical context.

One limitation of the multinomial logit discrete-choice model is that it does not allow for unobserved heterogeneity across individuals in their preference parameters. For example, while in some specifications we estimate gender-specific coefficients on crime rates, we do not allow for within-gender heterogeneity in this coefficient. The standard approach in demand estimation to allow for heterogeneity in individuals' preferences for observed product characteristics is to assume that the parameters capturing those preferences follow a known distribution in the population of interest. The combination of this assumption with the

multinomial logit assumption on the distribution of the error terms ν_{ijlt} yields a mixed logit discrete choice model. In our specific setting, this is infeasible: unobserved heterogeneity in the parameter vector β would make the the choice-set construction results derived in McFadden (1978) inapplicable and therefore necessitate estimating a likelihood function using the actual choice set J_t, which is computationally infeasible in a city with tens of thousands of restaurants. A conceivable alternative would be to estimate a mixed logit model by arbitrarily assigning consumers choice sets that omit the majority of restaurants in New York City. Such choice-set reductions would almost invariably be wrong in a particular sense: mixed-logit parameter estimates can be very sensitive to the misspecification of the actual choice set that consumers take into account when deciding which product they prefer (Conlon and Mortimer, 2013). In other words, arbitrary decisions necessary to make the mixed-logit model computationally feasible would have a large impact on the resulting parameter estimates. Given this consideration, in sections 3.1 and 3.2 we have specified a model that may be consistently estimated while exploiting information contained in only a subset of the users' true choice sets.

While we do not allow for unobserved heterogeneity in preferences, we allow this preferences to vary with observed individual characteristics. Given the information available to us on each user's gender, race/ethnicity and home census tract median income, we will allow the preference parameters β to vary across groups of users by interacting these individual characteristics with both restaurant characteristics and our measures of spatial and social frictions. For example, we will allow users living in tracts of different income levels to value restaurants's prices and ratings differently, and users of different genders to differentially value census tracts' crime levels. As suggested by the testing procedure in Hausman and McFadden (1984), the parameter estimates we obtain will be robust to the particular choice sets used to estimate them only if this observed heterogeneity in preferences is sufficient to characterize users' preferences (so that the resulting model exhibits the independence of irrelevant alternatives property).[18] As we show in appendix C.3, our estimates vary very little across different randomly generated choice sets. Therefore, we infer that, in our particular application, it is unlikely that the independence of irrelevant alternatives assumption is driving our results.[19]

A second identifying assumption implicit in sections 3.1 and 3.2 is that individuals' home and work locations are exogenously determined. However, in practice, individuals choose where to live and work and, consequently, it is conceivable that the home and work locations of individuals in our sample might be endogenously determined as a function of restaurant characteristics. The endogenous location of home and work will not bias our estimates of the preference parameter β as long as the distribution of the vector of unobserved

[18]The formal testing procedure in Hausman and McFadden (1984) compares parameters estimated using the whole choice set to those estimated using a randomly selected subset. Since it is not computationally feasible to estimate our model with the whole choice set, we cannot implement this exact test. Instead, we compare estimates from models that differ only in their randomly selected choice sets.

[19]Katz (2007) and Pakes (2010) show that there is an alternative estimation approach that uses moment inequalities and that would allow both to handle potentially large unobserved choice sets and heterogeneity in the individuals' preferences for some observed restaurant characteristics. For the specific case of our empirical exercise, we discuss in Appendix B.3 the advantages and disadvantages of the moment inequality estimation approach relative to that described in sections 3.1 and 3.2.

restaurant characteristics affecting individuals' restaurant choices, $\{\nu_{ijlt}; l \in \mathcal{L}, j \in J_t\}$ for any individual i and period t is independent of the vector of characteristics determining the optimal selection of home and work location. One particular case that this assumption rules out is the possibility that the location of home or work is itself directly a function of $\{\nu_{ijlt}; l \in \mathcal{L}, j \in J_t\}$.[20]

Conversely, this assumption imposes no restriction on the restaurants characteristics that we explicitly account for in the vector X_{it} influencing users' home and work locations. Therefore, any locational endogeneity is less likely to bias our estimates of the preference parameters $\{(\beta_1^l, \beta_2); l \in \mathcal{L}\}$ the larger the vector of restaurant characteristics that we explicitly control for; i.e. the larger the set of characteristics affecting restaurant choice that we include in the vector X_{it} and that, therefore, need not be accounted for by the unobserved components $\nu_{it} = \{\nu_{ijlt}; l \in \mathcal{L}, j \in J_t\}$. The reason is that the fewer the variables that are accounted for by this unobserved component, the more likely it is that this composite is independent of the characteristics determining each individual's choice of home and work location. With this in mind, the empirical specifications we introduce in the next section incorporate a very large set of restaurant characteristics.

4 Estimation results

This section reports the results of estimating discrete-choice models of the form described in section 3 using the data introduced in section 2. We estimate different discrete-choice models that progressively expand the set of included covariates, X_{ijlt}, by introducing measures that describe spatial and social frictions.

In each specification, we include in X_{ij}^2 a number of venue and area characteristics that presumably influence consumer decisions but are not our frictions of interest. These are the venue's price (expressed in terms of four price categories) and average rating, these restaurant characteristics interacted with the user's home census tract's median household income, the log median household income of the tract in which the venue is located, the percentage difference and absolute percentage difference in median incomes between these two tracts, 28 area dummies, and 9 cuisine dummies.[21]

We include this rich set of covariates in order to address two potential concerns for obtaining consistent estimates of the frictions of interest: (1) omitted variable bias due to characteristics that might influence consumption decisions and be correlated with the spatial and social frictions of interest, and (2) selection bias coming from characteristics that might influence both consumption decisions and the endogenous location of home and work. For example, the restaurant characteristics and area dummies imply that our estimates of spatial frictions cannot be attributed to our sample of Yelp users coincidentally or endogenously living and working near attractive dining options.[22] Similarly, the covariates capturing income levels and differences imply that our estimates of social frictions cannot simply be attributed

[20] A detailed analysis of the effect of endogenous home and work location on our estimates is contained in Appendix B.2.

[21] We aggregate NYC community districts to partition New York City into 28 large areas; see appendix section A.1. We aggregate the cuisine categories reported in Yelp into 9 categories (American, Asian, European, Indian, Latin American, Middle Eastern, veggie, and unassigned).

[22] These controls address the colocation problem for common valuations of restaurant characteristics, such

to the fact that demographics and crimes covary with incomes across space. When we estimate gender-specific coefficients on our measures of spatial and social frictions, we also estimate gender-specific coefficients on the controls for the venue's price and rating, income in the tract in which the venue is located, and the tracts' income differences. In the interest of brevity, we do not report the coefficients on this large vector of controls.

In each specification in the main text, the choice set R_{it} contains 20 elements: the venue chosen by the individual at that time and 19 randomly drawn venues. All the specifications presented in the main text are estimated using a common choice set R_{it}, so that variation across columns and tables represents variation in results, not variation in the random subset of venues included in users' choice sets. We have also estimated these specifications repeatedly using different randomly sampled choice sets and choice sets with more elements (see appendix C.3). Consistent with the independence-of-irrelevant-alternatives assumption, the estimated coefficients are largely invariant to the particular randomly sampled subsets of the choice set.

A user making a trip to a Yelp venue may start that trip from her home location, work location, or during her commute, as described in section 3.[23] Similarly she decides on the mode of transport, traveling via mass transit or automobile. Thus, while R_{it} contains 20 elements, each user chooses one of 120 possible choices when there are three possible points of origin and two modes of transport.

4.1 Travel time

We first investigate the role of spatial frictions in terms of the number of minutes a consumer would have to travel in order to visit a restaurant. As discussed earlier, we consider three origins, home, work, and the commuting path between them, and two transport modes, car and public transit, so there are six potential origin-mode pairs. We first estimate a specification in which home is the only origin and then sequentially introduce the work and commuting origins.

Table 6 presents estimates for four multinomial-logit models. In the first column, we find large, negative coefficients on travel times from home. Users are much less likely to visit venues that are far from their home via public transit or automobile. In the second column, introducing the work origins yields comparable coefficients on travel times from home and negative coefficients on travel times from work. The coefficients on log minutes from work are roughly thirty percent larger than those on log minutes from home, consistent with the hypothesis that the opportunity cost of travel time from work is greater than that from home. In the third column, we introduce the travel time to the venue from the user's commuting path. The home and work coefficients are largely unchanged, and the commuting path travel times exhibit significant negative coefficients.

---

that our results cannot be attributed to users choosing to live near highly rated restaurants or areas that are nice places to dine. One potential concern would be that individuals with heterogeneous preferences locate near restaurants that they in particular find attractive. In Table 8, we address demographic-specific tastes for particular cuisine by introducing demographic-cuisine-category interactions.

[23] Knowing both the home and work locations of the Yelp users in our sample is key to our results. Appendix C.2 present estimates of specifications using only the home origin, and the results meaningfully differ.

Table 6: Travel time

	(1)	(2)	(3)	(4)
Log of travel time from home-public	-1.26^a	-1.23^a	-1.17^a	-1.19^a
	(.019)	(.023)	(.046)	(.071)
Log of travel time from home-public × female				.043
				(.094)
Log of travel time from home-car	-1.55^a	-1.37^a	-1.29^a	-1.40^a
	(.025)	(.020)	(.038)	(.067)
Log of travel time from home-car × female				$.146^c$
				(.082)
Log of travel time from work-public		-1.73^a	-1.65^a	-1.75^a
		(.060)	(.126)	(.209)
Log of travel time from work-public × female				.208
				(.259)
Log of travel time from work-car		-1.96^a	-1.92^a	-1.91^a
		(.043)	(.112)	(.156)
Log of travel time from work-car × female				.055
				(.212)
Log of travel time from commute-public			-1.09^a	-1.14^a
			(.030)	(.050)
Log of travel time from commute-public × female				.101
				(.064)
Log of travel time from commute-car			-1.32^a	-1.43^a
			(.034)	(.058)
Log of travel time from commute-car × female				$.153^b$
				(.072)
Number of origin-mode points	2	4	6	6
Number of venues in choice set	20	20	20	20
Log-Likelihood	-2.35	-2.30	-2.32	-2.31
Pseudo R-sq	.215	.229	.225	.227
Akaike Information Criterion	100	104	108	134
Number of trips	16573	16573	16573	15619
Number of individuals	406	406	406	385

NOTES: Each column reports an estimated multinomial logit model of individuals' decisions to visit a Yelp venue. Standard errors in parentheses. Statistical significance denoted by a (1%), b (5%), c (10%). Unreported controls include venue price and rating interacted with home tract income, log median household income in tract of venue, percentage difference and percentage absolute difference in income levels, 28 area dummies, and 9 cuisine dummies. In column 4, additional unreported controls are a female user dummy interacted with venue price and rating, log median household income in tract of venue, and percentage difference and percentage absolute difference in income levels.

In the fourth column, we estimate gender-specific coefficients on these travel-time covariates. The estimation sample is slightly smaller, as we restrict attention to users identified as male or female. In addition to estimating gender-specific coefficients on the spatial frictions, we allow the unreported covariates capturing restaurant characteristics and income levels to vary by gender. The estimates show that men and women respond to spatial frictions similarly; the coefficients on the log travel times are negative, highly significant, and similar to those in the third column. The interactions of travel times and the female dummy have positive coefficients, but they are small in magnitude compared to the main effects and not statistically significant in general.

Table 6 yields a very clear finding: spatial frictions matter. Travel time plays a first-order role in individuals' consumption choices within the city. Yelp users are less likely to visit venues that, in terms of mass-transit and automobile travel time, are more distant from their home and work locations, as well as the commuting path between these.[24] Since we control for restaurant and area characteristics, these findings cannot be attributed to our sample of Yelp users living close to attractive dining options. The finding that both individuals' home and work locations predict visit behavior is consistent with the finding in Houde (2012) that incorporating aggregate commuting flows improves market-share predictions. The fact that the coefficients on log travel time via automobile tend to be more negative than those on log travel time via public transit is consistent with the hypothesis that the marginal disutility of traveling via automobile is greater.[25] In New York City, public-transit fares are invariant to distance, while taxi fares are not. The gender-specific coefficients suggest, if anything, that females are slightly more inclined towards automobile transportation, relative to males, but these effects are small.

To illustrate the economic significance of travel time for consumer decision making, consider two hypothetical restaurants, identical in their characteristics except for the number of minutes away from the user. The first restaurant is 15 minutes from the user's workplace by car; the second restaurant is 30 minutes away. The estimated coefficients in column three imply that the user would be almost four times as likely to visit the more proximate venue from work by car ($2^{1.92} \approx 3.8$). Similarly, if the two restaurants were 15 and 30 minutes from the commuting path by public transit, the user would be more than twice as likely to visit the more proximate venue from his or her commute by public transit ($2^{1.09} \approx 2.1$).

Thus, we find that spatial frictions play a first-order role in urban consumption choices, consistent with a long tradition in theoretical models of spatial competition. Our estimated specifications recognize that consumption opportunities arise at both home and work, and we control for venue characteristics when inferring the disutility of additional travel time to a restaurant. Our quantification of these spatial frictions is an important input for models of consumption within the city (e.g. Allen, Arkolakis, and Li 2015). These magnitudes are also relevant for thinking about the welfare consequences of changes in the transportation network, such as new public-transit investments or private vehicle-sharing programs.

[24] We locate users based on reviews of Yelp venues (not necessarily restaurants) near their homes and workplaces. We have estimated the travel-time coefficients separately for users with few venues and more venues revealing this locational information and found that they are similar.

[25] The estimated specification assumes that there are no l-specific fixed costs in the utility function.

4.2 Demographic differences

Understanding the role and magnitude of social frictions in the city is at the heart of our effort. We investigate a series of questions. Is there a broad-based aversion to visiting venues in neighborhoods primarily populated by people other than those characteristic of one's home neighborhood? If yes, is this uniform across census tracts with the same demographic characteristics, or do users exhibit an even stronger relative aversion to going deep into neighborhoods populated by those different from residents near their homes? Are the areas populated by some races or ethnicities particularly attractive to all users? Are there specific bilateral patterns of attraction or aversion between racial and ethnic groups?

Table 7 presents estimates for four multinomial-logit models, one model per column. In addition to the unreported controls for restaurant and area characteristics included in all our specifications, each specification includes the travel-time covariates for the six origin-mode pairs introduced in the previous section. Since the coefficients on these travel-time covariates are stable across our specifications, we omit reporting these coefficients in Table 7 and subsequent tables where travel time is not the central focus.[26]

The first column of Table 7 introduces the Euclidean demographic distance measure and shows that users are less likely to visit venues located in census tracts with demographics different from those of their home census tract. Since our controls include travel times, this result cannot be attributed to residential segregation alone. Similarly, our controls include income differences, so this result cannot be attributed to differences in socioeconomic status. Finally, our controls include restaurant characteristics, area dummies, and the income level of tract k_j, such that this result cannot be attributed to the users in our estimation sample simply tending to reside in tracts with demographics similar to the demographics of tracts that host dining destinations all users find relatively attractive.[27]

The estimated coefficient on Euclidean demographic distance (EDD) implies an economically significant role for this social friction. Consider a user who contemplates visiting two venues that are identical except for their Euclidean demographic distances from the home census tract, which differ by one standard deviation.[28] Our estimates imply that the user would be 27% more likely to visit the venue in the more demographically similar census tract.[29]

We can also express the economic significance of demographic differences as a trade-off

[26]We have also found that our main results are robust to interacting log travel times with home tract median household income.

[27]By their nature, these controls cannot rule out the hypothesis that there are both unobserved restaurant characteristics and unobserved heterogeneous tastes for those characteristics that are correlated with demographic differences.

[28]For example, a user who lives in East Flatbush (tract 36047083800), which is 88% black, considers visiting a neighborhood near Atlantic Terminal (tract 36047017900) or a neighborhood in southwest Canarsie (tract 36047069601). The former is 59% black, 20% white, and 13% Hispanic; the latter is 36% black, 48% white, and 12% Hispanic. As a result, the Atlantic Terminal neighborhood has a Euclidean demographic distance of 0.25 from the user's home tract, while the Canarsie tract has an Euclidean demographic distance of 0.49. Thus, these two destinations' Euclidean demographic distances differ by about one standard deviation.

[29]Comparing two venues j and j' for which $X^1_{ij} = X^1_{ij'}$, $\frac{P(d_{ij'}=1|X_{it};\beta)}{P(d_{ij'}=1|X_{it};\beta)} = \exp(\beta^2(X^2_{ij} - X^2_{ij'}))$. Table 4 shows that the standard deviation of Euclidean demographic distance is 0.226, so the coefficient of -1.06 in column 1 of Table 7 implies that a venue that has EDD 0.226 lower than an otherwise-identical venue will be visited with 27% higher probability ($\exp(-1.06 \times -.226) \approx 1.27$).

Table 7: Demographic differences

	(1)	(2)	(3)	(4)
EDD between h_i and k_j	-1.06^a	-1.06^a	-1.27^a	-1.26^a
	(.072)	(.075)	(.114)	(.114)
SSI of k_j		$.091^a$	$.072^a$	$.052^b$
		(.022)	(.022)	(.024)
EDD × SSI		-.071	$-.135^b$	$-.227^c$
		(.056)	(.068)	(.116)
EDD × h_i is plurality Asian			.201	.332
			(.264)	(.265)
EDD × h_i is plurality black			-.064	-.368
			(.403)	(.419)
EDD × h_i is plurality Hispanic			$.726^b$	$.715^b$
			(.288)	(.289)
k_j is plurality Asian			$.312^a$	.148
			(.113)	(.125)
k_j is plurality black			.324	.342
			(.243)	(.310)
k_j is plurality Hispanic			$.317^a$	$.363^a$
			(.120)	(.122)
EDD × k_j is plurality Asian			.229	$.466^c$
			(.243)	(.276)
EDD × k_j is plurality black			-1.54^a	-1.12
			(.526)	(.778)
EDD × k_j is plurality Hispanic			$-.616^b$	$-.736^b$
			(.305)	(.312)
SSI × k_j is plurality Asian				$.429^a$
				(.090)
SSI × k_j is plurality black				-.663
				(.451)
SSI × k_j is plurality Hispanic				-.376
				(.286)
EDD × SSI × k_j is plurality Asian				$-.456^b$
				(.213)
EDD × SSI × k_j is plurality black				-.089
				(.906)
EDD × SSI × k_j is plurality Hispanic				.768
				(.482)
Number of origin-mode points	6	6	6	6
Number of venues in choice set	20	20	20	20
Log-Likelihood	-2.31	-2.31	-2.31	-2.30
Pseudo R-sq	.227	.227	.228	.229
Akaike Information Criterion	110	114	132	144
Number of trips	16573	16573	16573	16573
Number of individuals	406	406	406	406

NOTES: Each column reports an estimated multinomial logit model of individuals' decisions to visit a Yelp venue. Standard errors in parentheses. Statistical significance denoted by *a* (1%), *b* (5%), *c* (10%). "EDD" is Euclidean demographic distance; "SSI" is spectral segregation index. The unreported covariates are log travel times from six origin-mode pairs and the unreported controls in Table 6.

between demographic distance and time. To hold constant the utility of visiting a venue from home via mass transit, a venue one standard deviation more demographically distant would have to be about 21% closer in terms of travel time.[30]

Next, we investigate the possibility that demographic differences may matter more when the venue is located deep within a segregated area. To do so, we use a spectral segregation index (SSI) that describes a tract's demographic isolation in terms of its racial/ethnic plurality and interact it with the Euclidean demographic distance. The first column demonstrated that users are less likely to visit venues located in places with demographics different from those of their home location. Interacting the demographic distance with the demographic isolation of the venue tract serves to gauge the permeability of demographic boundaries. Holding fixed the demographics of the destination, are users less likely to visit venues in places with different demographics when the destination is surrounded by tracts of the same racial/ethnic plurality?

In the second column of Table 7, the interaction of Euclidean demographic distance and the spectral segregation index has a negative coefficient that is statistically insignificant and close to zero. Conditional on the demographic differences between the origin and destination tract, a destination tract near the edge of a racially or ethnically distinct area is as likely to be visited by a user as tract with the same demographic differences deep inside that area. Moreover, the positive coefficient on SSI implies that, on average and conditional on demographic differences, individuals find more demographically isolated tracts more attractive.[31]

The specifications in the first two columns of Table 7 treat all demographic differences and demographic isolation symmetrically. This may not be a very good description of the social frictions that influence consumer behavior. For example, a user who lives in an overwhelmingly white neighborhood may view visiting a restaurant in an overwhelmingly black neighborhood differently from a user who contemplates the reverse visit, even though the Euclidean demographic distance is the same for both trips by definition. Similarly, the experience of visiting a demographically isolated Asian neighborhood may be different from visiting a demographically isolated Hispanic neighborhood. We proceed to relax the symmetry assumption by estimating plurality-specific coefficients on the EDD and SSI terms, relative to the omitted categories of h_i and k_j having populations that are plurality white.

The third column of Table 7 reports plurality-specific coefficients on the Euclidean demographic distance. The common coefficient on Euclidean demographic distance remains large and highly significant. However, the estimates also show substantial heterogeneity in the effect of Euclidean demographic distance as a function of the home and destination tracts' demographics, rejecting the assumption of symmetry. Most notably, while users on average are significantly less likely to visit a census tract that has demographics considerably different from those of their own residence, the negative effect of Euclidean demographic distance is more than twice as large when the destination tract is plurality black. Hispanic

[30]To hold U_{ijlt} constant, a change of ΔX^2_{ij} would be offset by the change $\Delta X^1_{ijl} = -\beta^2 \Delta X^2_{ij}/\beta^1_l$. Since the unreported coefficient β^1_{hp} is -1.15 (quite similar to its reported value of -1.17 in the previous table), the change required to offset a one-standard-deviation increase in Euclidean demographic distance is $-1.06 \times .226/1.15 \approx -.21$.

[31]Note that both SSI and EDD are functions of the destination tract's demographic characteristics. In our estimation sample, EDD is positively correlated with SSI for destination tracts that are plurality Asian, black, and Hispanic, while it is (slightly) negatively correlated with SSI for plurality-white destination tracts.

tracts exhibit sharp asymmetries – the negative effect of Euclidean demographic distance is considerably weaker for users residing in plurality-Hispanic tracts and considerably stronger for restaurants located in plurality-Hispanic tracts. The differential effects of Euclidean demographic distance for Asian pluralities have point estimates of meaningful size compared to the average effect, but these are imprecisely estimated. Overall, this plurality-specific specification demonstrates substantial asymmetries in the social frictions associated with demographic differences. Since our control covariates include the two tracts' absolute and signed percentage difference in median household income, these results cannot simply be attributed to differences in income levels across demographic groups.

The fourth column of Table 7 reports plurality-specific coefficients on Euclidean demographic distances, spectral segregation indices, and their interaction. First, we again find heterogeneous effects of Euclidean demographic distances, consistent with the patterns found in the third column. The positive coefficient on plurality-Asian destinations substantially dampens the negative effect of Euclidean demographic distance on the probability of visiting restaurants in such tracts. Second, the estimates suggest that the relative attractiveness of segregated areas on average is driven by users' greater likelihood of visiting restaurants located in demographically isolated census tracts that are plurality Asian or, to a much lesser degree, white. This would be consistent with New York's Chinatowns being popular with Yelp users as a whole in a manner not predicted by their income levels and restaurant ratings. Though very imprecisely estimated, the coefficients on the black and Hispanic SSI terms are negative and an order of magnitude larger than the positive coefficient on SSI for plurality-white tracts. Third, we find weak evidence that users are less likely to visit demographically different tracts that are farther within segregated areas. The plurality-specific EDD-SSI coefficients are imprecisely estimated for black and Hispanic pluralities. The Asian-plurality EDD-SSI coefficient is notably negative, though this only dampens users' proclivity for highly segregated Asian areas.[32]

These social frictions might have a gendered component, for numerous potential reasons. In estimating, however, we have found little evidence of differential responses to these demographic traits across genders. Table C.1 in appendix C.1 reports gender-specific estimates of the coefficients in Table 7 and there is no reliable pattern in which women respond differently to demographic differences than men.

Because the demographic measures in Table 7 characterized tract-level population characteristics, these findings may capture at least two distinct phenomena. First, users of all races/ethnicities may prefer to spend time in places with demographics similar to those of the tract in which they reside, a preference for environmental similarity. Second, individuals may prefer to spend time in places populated by individuals similar to themselves, homophily. An example of the first would be if a white individual living in Harlem, a heavily black and Hispanic neighborhood, is more likely to visit black and Hispanic neighborhoods than a white individual living elsewhere. An example of the second would be if black and Hispanic individuals are more likely to visit black and Hispanic neighborhoods and Yelp users living in Harlem are more likely to be black or Hispanic.

We attempt to separate these possibilities by including individual demographic infor-

[32]For venues in tracts that are plurality Asian, the estimated coefficients imply that $\frac{\partial U_{ijl}}{\partial SSI_j} > 0$ up to the 95th percentile of EDD_j.

mation in specifications in Table 8. Before presenting the results, we note a few reasons for caution in interpreting the findings. The first is that the racial or ethnic identity of the individual is being inferred from profile photos on Yelp. Prior work comparing inferred demographics and administrative data suggests that this can be done reliably with respect to three groupings – Asian, black, and white or Hispanic (Mayer and Puller, 2008). There is therefore significant within-group heterogeneity, most notably in the white and Hispanic group. The second concern is that we only identified individual demographics for a subset of users, so the sample size is smaller than in previous tables. In particular, there are few black users in our estimation sample, so the estimated coefficients reflect the behavior of a small number of individuals. With these cautions in mind, the results in Table 8 nonetheless suggest a profound role for race and ethnicity in shaping how users use the city.

Column 1 in Table 8 characterizes demographic differences in terms of the share of individuals residing in the venue's census tract who do not belong to the same race or ethnicity as the user, pooling all demographic groups. The results suggest that, in general, users are more likely to visit venues in locations with populations more similar to their own identity. The estimates reveal that this role for homophily at the individual level is distinct from the tendency for users to visits restaurants in tracts with demographics similar to that of their home tract.

Columns 2 and 3 investigate whether this finding varies across users with different identities. The results suggest that users of all demographic groups are less likely to visit venues in census tracts whose residents have different racial and ethnic identities than their own. This pattern is particularly strong for black users and weakest for white or Hispanic users, in which case individual homophily may be zero or cannot be distinguished from environmental similarity as captured by the Euclidean demographic distance. These results may be a consequence of not distinguishing between whites and Hispanics when classifying users' profile photos. This finding suggests that consumption choices reflect not only a tendency towards environmental similarity, in which users are more likely to visit venues located in census tracts with demographics similar to those of their home tracts, but also homophily, in that users are more likely to consume in census tracts populated by individuals sharing their own ethnic or racial characteristics.

Columns 4 and 5 go further by investigating homophily category-by-category. For each user type, we interact the user identity with the residential population shares of Asians, blacks, Hispanics, and others in the venue's tract. This makes whites the omitted category. The positive and significant coefficient on the interaction of individual race/ethnicity with the population share of the same demographic group shows that users in our sample are more likely to visit venues located in census tracts with more residents of their own type. The evidence for homophily on the part of Asian and black users is clear, while our inability to distinguish between Hispanic and white users based on the profile photos likely muddles our attempt to discern homophily for these users. In column 5, we interact the cuisine dummies with individual demographics. Comparing columns 4 and 5 reveals that demographic-linked cuisine tastes explain a small fraction of the inferred homophily. While it is true that Asian users are more likely to eat at Asian restaurants, this does not explain the finding that Asian users are more likely to visit venues located in places with more Asian residents. Alongside this homophily, the consistently large, negative coefficient on Euclidean demographic distance demonstrates a strong role for environmental similarity.

Table 8: Individual demographics

	(1)	(2)	(3)	(4)	(5)
EDD between h_i and k_j	$-.935^a$		-1.13^a	-1.10^a	-1.10^a
	(.080)		(.084)	(.086)	(.086)
Percentage of k_j population of non-i race	$-.886^a$				
	(.060)				
i is Asian $\times$ share non-Asian in k_j		-1.11^a	-1.40^a		
		(.086)	(.090)		
i is black $\times$ share non-black in k_j		-2.36^a	-1.80^a		
		(.353)	(.372)		
i is whithisp $\times$ share non-whithisp in k_j		$-.577^a$	-.131		
		(.100)	(.107)		
i is Asian $\times$ share Asian in k_j				1.34^a	1.18^a
				(.098)	(.099)
i is Asian $\times$ share black in k_j				-.046	.030
				(.256)	(.257)
i is Asian $\times$ share Hispanic in k_j				-.309	-.242
				(.188)	(.189)
i is black $\times$ share Asian in k_j				-1.87^a	-1.90^a
				(.540)	(.548)
i is black $\times$ share black in k_j				1.14^a	$.983^b$
				(.414)	(.416)
i is black $\times$ share Hispanic in k_j				1.92^a	1.78^a
				(.370)	(.376)
i is whithisp $\times$ share Asian in k_j				$-.198^c$	-.005
				(.119)	(.120)
i is whithisp $\times$ share black in k_j				-.116	-.178
				(.241)	(.241)
i is whithisp $\times$ share Hispanic in k_j				$.480^a$	$.419^b$
				(.167)	(.168)
Number of origin-mode points	6	6	6	6	6
Number of venues in choice set	20	20	20	20	20
Cuisine-individual-demographic interactions					Yes
Log-Likelihood	-2.31	-2.32	-2.31	-2.31	-2.30
Pseudo R-sq	.226	.225	.227	.228	.231
Akaike Information Criterion	112	114	116	134	166
Number of trips	13257	13257	13257	13257	13257
Number of individuals	295	295	295	295	295

NOTES: Each column reports an estimated multinomial logit model of individuals' decisions to visit a Yelp venue. Standard errors in parentheses. Statistical significance denoted by a (1%), b (5%), c (10%). "EDD" is Euclidean demographic distance. The unreported covariates are log travel times from six origin-mode pairs and the unreported controls in Table 6. In columns 4 and 5, there are three more unreported covariates, the interactions of i's racial/ethnic category and the share "other" in k_j. In column 5, the cuisine dummies are also interacted with i's racial/ethnic category.

In Table 9, we investigate whether these social frictions vary with user gender. The columns of Table 9 are akin to the specifications appearing in columns 2 through 4 of Table 8 with gender-specific coefficients on both the unreported controls and the social frictions of interest. These estimates suggest that men and women respond similarly to Euclidean demographic distance, as reported above in the discussion of Table 7. We infer heterogeneity in the degrees of homophily across gender-race categories. In particular, the estimated coefficients suggest that Asian females exhibit weaker homophily than Asian males, while white or Hispanic females exhibit greater homophily than their male counterparts. Again, we have limited ability to interpret the white-or-Hispanic results given the substantial within-category heterogeneity.

The body of evidence presented in this section demonstrates that economic interaction, as measured by trips to restaurants, is *de facto* segregated within the city, over and above the fact that residential segregation itself stands as a barrier to integrated consumption patterns through the time cost of travel. The necessarily imperfect nature of our measures of these social frictions means that these findings could be consistent with more than one behavioral mechanism. For example, users may be more likely to visit restaurants located near the residences of friends and family members. If these relationships exhibit racial/ethnic homophily, then our estimates may reflect the influence of social networks rather than frictions involved in interacting with strangers. Similarly, if these social networks can be predicted on the demographic composition of users' home census tracts, conditional on incomes and travel times, the large effect of Euclidean demographic distances could reflect attraction to known persons rather than aversion to visiting areas populated by strangers with different demographics.

Regardless of the particular mechanisms underlying how demographic differences shape consumption in the city, our quantification of their influence demonstrates that this social friction plays an important role in consumer behavior. First, it is quantitatively large in terms of consumer decisions. Ceteris paribus, a user would be 27% more likely to visit a venue in a census tract that is one standard deviation more demographically similar to her home tract. Second, these frictions are not symmetric. We estimate that users are less likely to visit venues in demographically distant destinations when that tract's residents are plurality black. Third, the identity of the individual user matters for predicting this behavior. Our estimates suggesting homophily with regard to the user's race/ethnicity imply that quantitative accounts of consumption in the city should not rely on a representative agent's valuation of particular consumption opportunities.

Table 9: Gender-specific coefficients for individual demographics

	(1)	(2)	(3)	(4)
EDD between h_i and k_j		$-.954^a$		$-.959^a$
		(.133)		(.136)
EDD × female		-.240		-.170
		(.170)		(.174)
i is Asian × share non-Asian in k_j	-1.72^a	-1.88^a		
	(.133)	(.136)		
i is black × share non-black in k_j	-2.07^a	-2.16^a		
	(.771)	(.794)		
i is whithisp × share non-whithisp in k_j	$-.273^c$	.165		
	(.144)	(.157)		
i is Asian × share Asian in k_j			1.72^a	1.91^a
			(.144)	(.148)
i is Asian × share black in k_j			-.163	.126
			(.387)	(.398)
i is Asian × share Hispanic in k_j			-.225	-.092
			(.277)	(.283)
i is black × share Asian in k_j			-.742	-.583
			(.702)	(.710)
i is black × share black in k_j			1.66^b	1.85^b
			(.829)	(.852)
i is black × share Hispanic in k_j			1.55^b	1.79^b
			(.773)	(.785)
i is whithisp × share Asian in k_j			-.171	$.308^c$
			(.157)	(.171)
i is whithisp × share black in k_j			$-.983^a$	-.546
			(.346)	(.356)
i is whithisp × share Hispanic in k_j			$.434^c$	$.676^a$
			(.223)	(.229)
i is Asian × share non-Asian in k_j × female	$.898^a$	$.699^a$		
	(.162)	(.168)		
i is black × share non-black in k_j × female	-.454	.374		
	(.854)	(.883)		
i is whithisp × share non-whithisp in k_j × female	$-.521^a$	$-.543^a$		
	(.186)	(.200)		
i is Asian × share Asian in k_j × female			-1.03^a	$-.835^a$
			(.176)	(.182)
i is Asian × share black in k_j × female			-.244	-.203
			(.445)	(.457)
i is Asian × share Hispanic in k_j × female			-.414	-.311
			(.320)	(.328)
i is black × share Asian in k_j × female			-2.58^b	-2.57^b
			(1.06)	(1.07)
i is black × share black in k_j × female			-.238	-.979
			(.921)	(.948)
i is black × share Hispanic in k_j × female			.526	.181
			(.863)	(.879)
i is whithisp × share Asian in k_j × female			$-.893^a$	$-.923^a$
			(.207)	(.222)
i is whithisp × share black in k_j × female			$.709^c$	.657
			(.394)	(.407)
i is whithisp × share Hispanic in k_j × female			-.310	-.323
			(.267)	(.275)
Number of origin-mode points	6	6	6	6
Number of venues in choice set	20	20	20	20
Log-Likelihood	-2.31	-2.30	-2.30	-2.30
Pseudo R-sq	.227	.230	.229	.231
Akaike Information Criterion	146	150	182	186
Number of trips	13257	13257	13257	13257
Number of individuals	295	295	295	295

NOTES: Each column reports an estimated multinomial logit model of individuals' decisions to visit a Yelp venue. Standard errors in parentheses. Statistical significance denoted by a (1%), b (5%), c (10%). "EDD" is Euclidean demographic distance; "SSI" is spectral segregation index. The unreported covariates log travel times from six origin-mode pairs, the unreported controls in Table 6, and a female user dummy interacted with venue price and rating, log median household income in tract of venue, and percentage difference and percentage absolute difference in income levels.

4.3 Crime

Another social friction that receives much attention in urban settings and interacts with demographic differences is the threat of crime. To assess the influence of crime on consumption behavior in the city, we introduce the average annual robberies per resident in the census tract in which the restaurant is located as a choice characteristic.

Table 10 presents estimates for four multinomial-logit models. The specification in the first column adds this crime measure to the specification of column 2 in Table 7. The coefficient on robberies per resident is negative, large, and highly statistically significant. Users are less likely to visit venues in high-robbery places. The coefficients on the demographic covariates are largely unchanged. In the second column, we estimate plurality-specific coefficients on the demographic frictions and obtain a very similar effect of robberies per resident on the choice probability. The coefficient on Euclidean demographic distance for destinations that are plurality black is again very negative, demonstrating that this finding in Table 7 was not due to omitting spatial variation in robberies.

Thus, we find that robberies have an effect on the decisions of consumers to visit particular places. This social friction was not evident from simple variation presented in Figure 8 and demonstrates the value of estimating a behavioral model that allows us to separate spatial variation in crime rates from venue characteristics, spatial frictions, and demographic influences. Our estimates imply that a venue being located in a tract with one standard deviation higher robberies per resident is equivalent to that tract being about 3.2% farther away in terms of minutes of travel time from home by public transit.[33]

The effect of crime on users' decisions varies notably by gender. Columns 3 and 4 of Table 10 report specifications in which the coefficients on the spatial and social frictions (as well as unreported controls) are gender-specific. In these specifications, the negative coefficient on robberies per resident is about 50% larger for female users than male users. A meaningful portion of the finding that users are less likely to visit high-crime areas is that female users in particular are averse to visiting restaurants in places with high robbery rates. In columns 3 and 4, the estimates suggest that female users are less likely to visit restaurants located in demographically distant and isolated tracts. However, the standard errors on finer demographic measures, such as the interaction of EDD with spectral segregation or the black-plurality-specific EDD coefficient, are very large, such that we can no longer reject the null hypotheses of zero differential effect. The robust findings are that robberies reduce visits, female users are much more averse to robberies, Euclidean demographic distance reduces visits, and we are able to statistically distinguish the roles of crime and demographics.

As one means of illustrating the quantitative implications of crime on consumer behavior, we calculate the implied demand responses if robbery rates were to return to their peaks of the early 1990s. The left panel of Figure 9 depicts the time series of robberies in New York City, which declined more than 80% from 1990 to 2013. There was notable spatial variation in this decline, as depicted in the right panel of Figure 9, which plots precinct-level robberies in 2007-2011 as a fraction of the number of robberies in 1990. Manhattan experienced a sharper decline in robberies, in percentage terms, than the outer boroughs of the city.

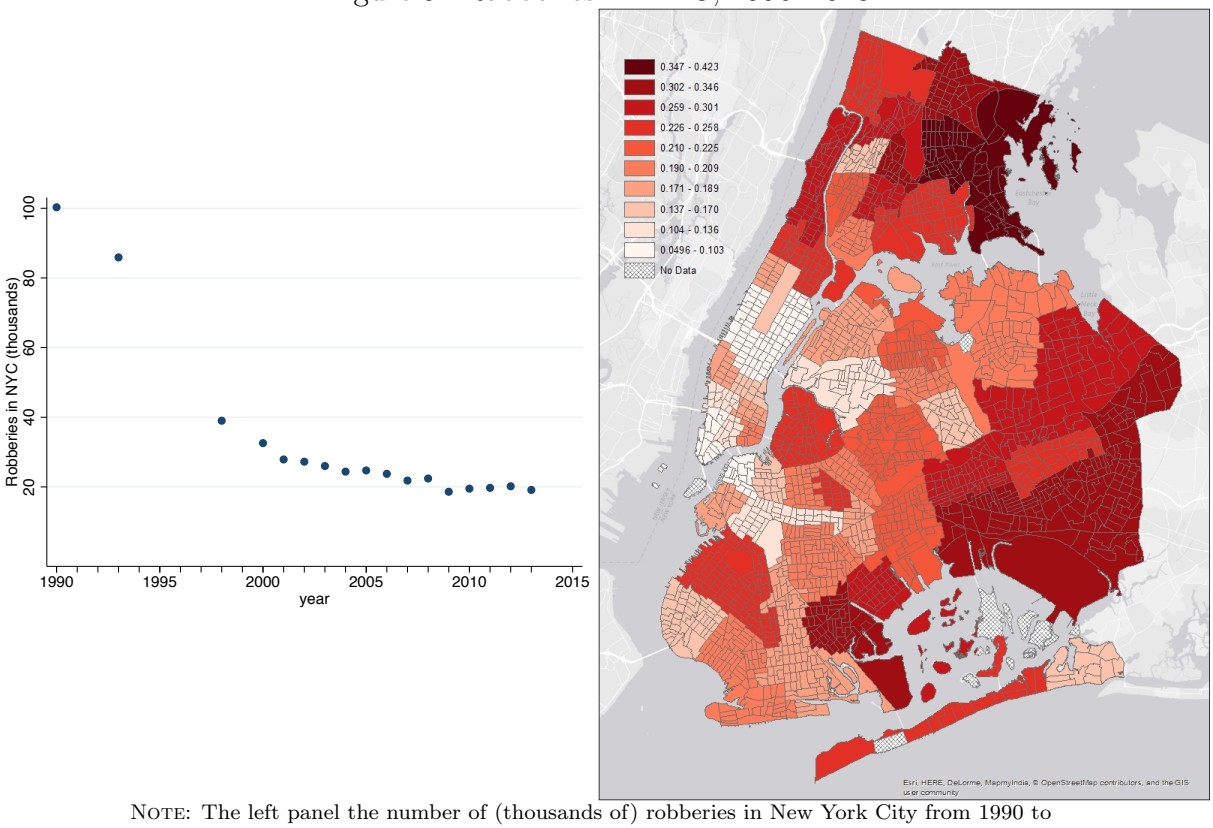
[33] Following the calculation in footnote 30 to hold U_{ijlt} constant and using the unreported coefficient $\beta^1_{hp} = -1.15$, the estimates in the second column of Table 10 imply that if if robberies per resident were 0.009 higher, this would be offset by a $-4.17 \times 0.009 / -1.15 = 0.032$ decrease in log minutes of travel time.

Table 10: Crime

	(1)	(2)	(3)	(4)
Average annual robberies per resident in k_j	-3.95^a	-4.19^a	-2.94^a	-3.10^a
	(.469)	(.473)	(.693)	(.694)
Average annual robberies per resident in k_j × female			-1.62^c	-1.73^c
			(.916)	(.912)
EDD between h_i and k_j	-1.04^a	-1.22^a	$-.985^a$	-1.22^a
	(.075)	(.115)	(.126)	(.158)
SSI of k_j	$.091^a$	$.053^b$	$.065^c$	-.005
	(.022)	(.024)	(.035)	(.046)
EDD × SSI	-.082	$-.229^b$	.040	.017
	(.057)	(.116)	(.091)	(.140)
EDD × female			-.111	-.041
			(.159)	(.163)
SSI × female			.033	.072
			(.036)	(.045)
EDD × SSI × female			$-.215^c$	$-.364^b$
			(.117)	(.146)
EDD × h_i is plurality Asian		.373		.402
		(.267)		(.273)
EDD × h_i is plurality black		-.473		-.583
		(.419)		(.440)
EDD × h_i is plurality Hispanic		$.682^b$		$.827^a$
		(.290)		(.300)
k_j is plurality Asian		$.299^b$		$.242^c$
		(.128)		(.132)
k_j is plurality black		.314		.040
		(.310)		(.330)
k_j is plurality Hispanic		$.380^a$		$.378^a$
		(.122)		(.126)
EDD × k_j is plurality Asian		.236		.346
		(.280)		(.289)
EDD × k_j is plurality black		-1.08		-.370
		(.777)		(.807)
EDD × k_j is plurality Hispanic		$-.770^b$		$-.749^b$
		(.313)		(.322)
SSI × k_j is plurality Asian		$.371^a$		$.434^a$
		(.092)		(.096)
SSI × k_j is plurality black		-.675		-.421
		(.451)		(.465)
SSI × k_j is plurality Hispanic		-.380		-.362
		(.287)		(.299)
EDD × SSI × k_j is plurality Asian		$-.396^c$		$-.523^b$
		(.216)		(.220)
EDD × SSI × k_j is plurality black		-.067		-.775
		(.903)		(.939)
EDD × SSI × k_j is plurality Hispanic		.769		.702
		(.483)		(.500)
Number of origin-mode points	6	6	6	6
Number of venues in choice set	20	20	20	20
Log-Likelihood	-2.31	-2.30	-2.30	-2.29
Pseudo R-sq	.228	.229	.231	.232
Akaike Information Criterion	116	146	150	180
Number of trips	16573	16573	15619	15619
Number of individuals	406	406	385	385

NOTES: Each column reports an estimated multinomial logit model of individuals' decisions to visit a Yelp venue. Standard errors in parentheses. Statistical significance denoted by *a* (1%), *b* (5%), *c* (10%). "EDD" is Euclidean demographic distance; "SSI" is spectral segregation index. The unreported covariates are log travel times from six origin-mode pairs, the unreported controls in Table 6. In columns 3 and 4, additional unreported controls are a female user dummy interacted with venue price and rating, log median household income in tract of venue, and percentage difference and percentage absolute difference in income levels.

Figure 9: Robberies in NYC, 1990–2013



NOTE: The left panel the number of (thousands of) robberies in New York City from 1990 to 2013. The NYPD only provides annual historical data for 2000 onwards. The map in the right panel depicts precinct-level robbery declines from 1990 to the 2007-2011 annual average. We merge precincts 33 and 34 for this purpose, since they were split apart in 1994. Tracts with zero population are reported as missing, since we assign tracts to the precinct in which the majority of the tract population resides.

Using the specification in column 4 of Table 10, in which robberies per resident reduce the probability of a user visiting a venue and this negative effect is particularly strong for female users, we calculate the probability that a user residing in Manhattan would visit each of the more than 10,000 Yelp restaurant venues in New York City under both current conditions and if robberies were at their 1990 levels.[34] This ceteris paribus exercise does not account for the ways that the monumental decline in crime rates both reflected and spurred New York City's changing demographic and economic circumstances over the past quarter-century.[35] However, it provides insights into how the consequences of crime for economic interactions are mediated by geography and gender.

To illustrate how spatial and social frictions interact and mediate changes in crime, consider the consequences, ceteris parbius, of a change in the covariates X_{ij}^2. Using $\mathcal{P}_{ij} \equiv P(d_{ijt} = 1|X_{it;}\beta)$ to denote the probability that a user visits a venue and $U_i \equiv \mathbb{E}_\nu \left(\max_{jl} \left(V_{ijl}\right)\right) = \ln\left(\sum_j \sum_l \exp\left(V_{ijl}\right)\right)$ to denote that user's expected utility, the consequences of a change dX_{ij}^2 are

$$d\mathcal{P}_{ij} = \beta^2 \mathcal{P}_{ij} \sum_{j' \neq j} \left\{\mathcal{P}_{ij'}(dX_{ij}^2 - dX_{ij'}^2)\right\}$$
$$dU_i = \beta^2 \sum_j \mathcal{P}_{ij} dX_{ij}^2$$

If spatial frictions were infinite so that individuals only spent time in their residential location, then the decline in robberies in one's own neighborhood would be a sufficient statistic for the welfare consequences of that change. On the other hand, if there were no spatial nor demographic frictions, then venues with larger declines in robberies would be more attractive for all New Yorkers, and demand would shift one-to-one with the change in robberies. In the world of finite frictions, the influence of these spatial and social frictions are summarized by the $\mathcal{P}_{ij}$ terms appearing in the probability-of-visit and utility expressions.

The implied utility gains from the decrease in robberies illustrate how spatial and social frictions mediate the consequences of changes in other covariates. First, there is meaningful spatial variation in the implied utility gains associated with the decrease in robberies per

[34] We compute these probabilities for venues for which we have a complete set of covariates. To calculate the visit probabilities and expected utility for a resident of a census tract, we also need to know where that individual works within New York City. Using commuting data from LEHD Origin-Destination Employment Statistics (LODES), we identify the five most common workplace tracts associated with a given residential tract. We then calculate the visit probabilities and expected utilities for five users of each gender with these residence-workplace combinations and take a weighted average over the five, using the commuting shares in the LODES data. To construct counterfactual outcomes, we calculate the same visit probabilities and expected utilities for this set of tracts with counterfactual values of the robberies per resident covariates in X_{ijl}. These counterfactual values come from 1990 population levels and tract-level robberies in 1990, which are inferred from tract-level robberies in 2007-2011 and precinct-level robbery declines from 1990 to 2007-2011. See appendix A.2. We perform these calculations for census tracts in Manhattan, omitting the tracts for which we lack an estimate of contemporary median household income or robberies in 1990.

[35] We illustrate the demand response to counterfactual crime levels while holding all other characteristics fixed, thereby refraining from making assumptions about the endogenous supply of restaurants. By contrast, Houde (2012) and Eizenberg, Lach, and Yiftach (2015) use their demand estimates to analyze a merger and local market power, respectively, and therefore make specific supply-side assumptions to compute counterfactual equilibria.

resident. For our hypothetical male residents, the average utility gain is 1.4% of initial utility, but the standard deviation across Manhattan tracts is 0.3 percentage points. These utility gains are only weakly predicted by the change in crime in the user's home tract.[36] Second, the utility gains associated with the decrease in robberies per resident are larger for women than men. Averaging across all Manhattan census tracts, the increase in utility for females associated with the lower crime rate is about 30% larger than the increase for males.[37] This suggests that the substantial decline in crime in New York City over the last twenty-five years was particularly advantageous to females, since females' use of the city is more responsive to crime rates.[38]

In summary, our estimates show a role for crime in consumer's consumption decisions in the city. These effects are modest in magnitude given the low levels of robberies in contemporary New York City, but they show a notably gendered dimension. Female users respond more strongly to higher robbery rates than male users. This finding, inferred from cross-sectional variation in crime, has consequences for how we interpret the very large decline in robberies over the last 25 years.

5 Conclusions

We use a novel data source to describe the consumption of non-tradables in New York City and exploit properties of the multinomial logit discrete-choice model to tractably identify how consumers value venues' and locations' characteristics. Our identification of Yelp users' home and work locations allows us to characterize how consumption in the city depends on travel times, demographic differences, crime rates, and user characteristics.

Our estimation results characterize how both spatial and social frictions influence consumption choices. Consistent with theories of spatial competition, spatial frictions play a large role in consumption within the city. Across origin-mode pairs, halving the minutes of travel time to a venue would imply that the user would be two to nearly four times more likely to visit the venue from that origin by that mode. Our estimates show that both home and work locations are relevant for predicting the restaurants patronized by users.

Social frictions matter as well. A venue in a location one standard deviation more demographically distant from a user's home is one-quarter less likely to be visited, ceteris paribus.

[36] Regressing the utility gain on the change in robberies per resident in the home tract yields an R^2 of 37%.

[37] This statement is about dU for females and males, not dU/U. Since we estimate gender-specific coefficients for restaurant and area characteristics, there are also differences in U by gender. Here we focus on the change associated with the counterfactual robbery rates, dU.

[38] The calculation in the text is a ceteris paribus experiment concerning the change in the number of robberies. Naturally a general equilibrium exercise can have different outcomes. At one extreme, the rise in local amenities could accrue entirely to local landlords if prices are free to move and housing supply is perfectly inelastic. At the other extreme, incumbent residents of a locality can receive the full amenity gain as a welfare increment if housing supply is perfectly elastic or prices are frozen. Where we fall between the extremes, of course, depends on the relevant elasticities of supply of housing and of residents to localities, as well as the particulars of housing price regulation. While housing supply is far from perfectly elastic in New York City, considerable new housing construction has taken place in recent decades and continues. Moreover, rent stabilization regulations have placed strong limits on the rise of many rents. Thus the real impact will fall somewhere in between the extremes. For a more detailed discussion, see appendix D.

This aversion to demographic differences exists irrespective of the demographics of the destination neighborhood. However, these effects appear to be weaker when the destination is an Asian neighborhood and approximately doubled when the destination is a black neighborhood. There is some evidence that users are positively attracted to demographically isolated Asian communities, such as New York's Chinatowns. But there is no significant evidence, conditional on demographic distance, that users are less likely to go deeply into isolated neighborhoods of other racial and ethnic groups.

These social frictions take the form of both a preference for environmental similarity and homophily at the level of the individual. While caution is needed in interpreting the estimates, users appear to be much more likely to visit venues in neighborhoods where a larger share of the residents share the user's individual ethnicity or race, even as tract-level demographic differences persist in predicting large differences in visits. There is no significant difference in the response of men and women to simple demographic differences between home and venue, although there are some differences with respect to individual characteristics, with homophily being greater for Asian women and weaker for white or Hispanic women. We have noted that significant residential segregation, combined with spatial frictions, implies that consumption will also be segregated. The results we report should therefore be interpreted as the incremental segregation of consumption due to social frictions based on race and ethnicity. Our work highlights asymmetries in the impact of these social frictions, an area worthy of further exploration.

We also examine the role of crime as a social friction deterring consumption. In spite of the dramatic 85 percent decline in robberies from the peak of the early 1990s, robberies per capita continue to have a measurable, if modest, negative effect on visits to venues. The magnitude of the effect is gendered, with women having 50 percent greater sensitivity to crime than men in selecting venues for visits. This implies that the gains from the decline in crime since the early 1990s were of particular benefit to women.

While our estimation approach exploits data on the decision to eat at restaurants across New York City, the consequences of frictions like travel time, demographic differences, and robbery rates apply to a much broader scope of life in the city. These would include both the broader scope of consumption of non-tradable services, from laundromats to retailers, and the vast array of non-market activities that cause residents to traverse the city. We illustrate the importance and interaction of spatial and social frictions by using our estimates to predict demand responses to a return to the peak crime levels of the early 1990s. Higher crime reduces a neighborhood's consumption value for all potential visitors, not just the residents living in high-crime neighborhoods, and has a greater effect on the consumption decisions of women. Our quantitative framework could also be used to evaluate particular infrastructure projects or transportation policies. Spatial and social frictions divide the city, but our work can thus also make clear the potential benefits of various efforts to reduce such frictions.

References

Allen, T., C. Arkolakis, and X. Li (2015): "Optimal City Structure," mimeo.

Anderson, M. L., and J. R. Magruder (2012): "Learning from the Crowd: Regression

Discontinuity Estimates of the Effects of an Online Review Database," *Economic Journal*, 122(563), 957–989.

Andrews, D. W. K., and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78(1), 119–157.

Antecol, H., and D. A. Cobb-Clark (2008): "Racial and ethnic discrimination in local consumer markets: Exploiting the army's procedures for matching personnel to duty locations," *Journal of Urban Economics*, 64(2), 496–509.

Ayres, I. (2001): *Pervasive Prejudice? Unconventional Evidence of Race and Gender Discrimination*. University of Chicago Press.

Bayer, P., and R. McMillan (2008): "Distinguishing Racial Preferences in the Housing Market: Theory and Evidence," in *Hedonic Methods in Housing Markets: Pricing Environmental Amenities and Segregation*, ed. by A. Baranzini, J. Ramirez, C. Schaerer, and P. Thalmann, pp. 225–244. Springer.

Conlon, C. T., and J. H. Mortimer (2013): "Demand Estimation under Incomplete Product Availability," *American Economic Journal: Microeconomics*, 5(4), 1–30.

Couture, V. (2014): "Valuing the Consumption Benefits of Urban Density," .

Cullen, J. B., and S. D. Levitt (1999): "Crime, Urban Flight, And The Consequences For Cities," *The Review of Economics and Statistics*, 81(2), 159–169.

Cutler, D. M., E. L. Glaeser, and J. L. Vigdor (1999): "The Rise and Decline of the American Ghetto," *Journal of Political Economy*, 107(3), 455–506.

Dickstein, M., and E. Morales (2015): "What do Exporters Know?," .

Doran, B., and M. Burgess (2011): *Putting Fear of Crime on the Map: Investigating Perceptions of Crime Using Geographic Information Systems*, Springer Series on Evidence-Based Crime Policy. Springer.

Echenique, F., and R. G. Fryer (2007): "A Measure of Segregation Based on Social Interactions," *The Quarterly Journal of Economics*, 122(2), 441–485.

Eizenberg, A. (2014): "Upstream Innovation and Product Variety in the U.S. Home PC Market," *The Review of Economic Studies*, 81(3), 1003–1045.

Eizenberg, A., S. Lach, and M. Yiftach (2015): "Retail Prices in a City: An Empirical Analysis," .

Ferraro, K. F. (1996): "Women's Fear of Victimization: Shadow of Sexual Assault?," *Social Forces*, 75(2), 667–690.

Gibbons, S. (2004): "The Costs of Urban Property Crime," *Economic Journal*, 114(499), F441–F463.

Glaeser, E. L., J. Kolko, and A. Saiz (2001): "Consumer city," *Journal of Economic Geography*, 1(1), 27–50.

Guiso, L., P. Sapienza, and L. Zingales (2009): "Cultural Biases in Economic Exchange?," *The Quarterly Journal of Economics*, 124(3), 1095–1131.

Handbury, J. (2012): "Are Poor Cities Cheap for Everyone? Non-Homotheticity and the Cost of Living Across U.S. Cities," Columbia University working paper.

Handbury, J., and D. E. Weinstein (2011): "Is New Economic Geography Right? Evidence from Price Data," NBER Working Papers 17067, National Bureau of Economic Research, Inc.

Harrison, C., M. Jorder, H. Stern, F. Stavinsky, V. Reddy, H. Hanson, H. Waechter, L. Lowe, L. Gravano, and S. Balter (2014): "Using online reviews by restaurant patrons to identify unreported cases of foodborne illness—New York City, 2012–2013," *Morbidity and Mortality Weekly Report*, 63(20), 441–445.

Hausman, J., and D. McFadden (1984): "Specification Tests for the Multinomial Logit Model," *Econometrica*, 52(5), 1219–40.

Holmes, T. J. (2011): "The Diffusion of Wal-Mart and Economies of Density," *Econometrica*, 79(1), 253–302.

Hotelling, H. (1929): "Stability in Competition," *The Economic Journal*, 39(153), pp. 41–57.

Houde, J.-F. (2012): "Spatial Differentiation and Vertical Mergers in Retail Markets for Gasoline," *American Economic Review*, 102(5), 2147–82.

Katz, M. (2007): "Supermarkets and Zoning Laws," Harvard PhD dissertation.

Lee, J. (2000): "The Salience of Race in Everyday Life: Black Customers' Shopping Experiences in Black and White Neighborhoods," *Work and Occupations*, 27, 353–376.

Linden, L., and J. E. Rockoff (2008): "Estimates of the Impact of Crime Risk on Property Values from Megan's Laws," *American Economic Review*, 98(3), 1103–27.

Logan, J. R., Z. Xu, and B. J. Stults (2014): "Interpolating U.S. Decennial Census Tract Data from as Early as 1970 to 2010: A Longitudinal Tract Database," *The Professional Geographer*, 66(3), 412–420.

Lorenc, T., S. Clayton, D. Neary, M. Whitehead, M. Petticrew, H. Thomson, S. Cummins, A. Sowden, and A. Renton (2012): "Crime, fear of crime, environment, and mental health and wellbeing: Mapping review of theories and causal pathways," *Health & place*, 18(4), 757–765.

Luca, M. (2011): "Reviews, Reputation, and Revenue: The Case of Yelp.com," Harvard Business School Working Paper No. 12-016.

Mayer, A., and S. L. Puller (2008): "The old boy (and girl) network: Social network formation on university campuses," *Journal of Public Economics*, 92(1-2), 329–347.

McFadden, D. (1978): "Modelling the Choice of Residential Location," in *Spatial Interaction Theory and Planning Models*, ed. by A. Karlqvist, L. Lundqvist, F. Snickars, and J. Weibull, pp. 75–96. North Holland, Amsterdam.

Morales, E., G. Sheu, and A. Zahler (2015): "Extended Gravity," .

O'Flaherty, B., and R. Sethi (2007): "Crime and segregation," *Journal of Economic Behavior & Organization*, 64(3-4), 391–405.

——— (2010): "The racial geography of street vice," *Journal of Urban Economics*, 67(3), 270–286.

Pain, R. (1991): "Space, sexual violence and social control: Integrating geographical and feminist analyses of women's fear of crime," *Progress in Human Geography*, 15(4), 415–431.

——— (1997): "Social Geographies of Women's Fear of Crime," *Transactions of the Institute of British Geographers*, 22(2), pp. 231–244.

Pakes, A. (2010): "Alternative Models for Moment Inequalities," *Econometrica*, 78(6), 1783–1822.

Pope, J. C. (2008): "Fear of crime and housing prices: Household reactions to sex offender registries," *Journal of Urban Economics*, 64(3), 601–614.

Quillian, L., and D. Pager (2001): "Black Neighbors, Higher Crime? The Role of Racial Stereotypes in Evaluations of Neighborhood Crime," *American Journal of Sociology*, 107(3), pp. 717–767.

——— (2010): "Estimating Risk: Stereotype Amplification and the Perceived Risk of Criminal Victimization," *Social Psychology Quarterly*, 73(1), 79–104.

Rauch, J. E. (2001): "Business and Social Networks in International Trade," *Journal of Economic Literature*, 39(4), 1177–1203.

Schreer, G. E., S. Smith, and K. Thomas (2009): ""Shopping While Black": Examining Racial Discrimination in a Retail Setting," *Journal of Applied Social Psychology*, 39(6), 1432–1444.

Train, K. E., D. L. McFadden, and M. Ben-Akiva (1987): "The Demand for Local Telephant Service: a Fully Discrete Model of Residential Calling Patterns and Service Choices," *The Rand Journal of Economics*, 18(1), 109–123.

Wollman, T. (2015): "Trucks Without Bailouts: Equilibrium Product Characteristics for Commercial Vehicles," mimeo.

Yelp (2013): "Yelp.com Welcomes 100 Million Unique Visitors in January 2013," .

A Data

A.1 NYC geographic and demographic data

Our data on census tracts' geographic areas and populations come from the 2010 Census of Population (Series G001 and P5). By 2010 Census definitions, there are 2168 tracts in New York City, of which 288 are in Manhattan.

The 2007-2011 American Community Survey 5-Year Estimate provides estimates of median household income (Series B19013) for 2110 and 279 of these tracts, for which summary statistics are provided in Table 3. Of the nine Manhattan tracts without median household income estimates, seven have a population below 25 persons, one is Inwood Hill Park (population 161), and one is Randall's Island (population 1648). More than 90% of the NYC tracts without median household income estimates have populations below 200 persons, the notable non-Manhattan exceptions being Bush Terminal (population 2105) and Rikers Island (inmate population of 11091).

Tract's historical demographic characteristics were obtained from the Longitudinal Tract Data Base, which maps prior Census years' population counts to the 2010 geographic definitions (Logan, Xu, and Stults, 2014).

We aggregated New York City's 59 community districts to define 28 areas, allowing us to use area dummies to control for unobservable characteristics when estimating. Each of Manhattan's 12 community districts constitute an area. We aggregated community districts to define 8 areas in Brooklyn (1, {2,6}, {3,8,9}, {4,5}, {7,11,12,13}, 10, {14,15}, {16,17,18}) and 6 in Queens ({1,2}, {3,4}, {5,6}, {7,8,11}, {9,10,14}, {12,13}). The boroughs of the Bronx and Staten Island each constitute one area. We assigned each census tract to one of these areas; tracts split across areas were assigned to the area with the largest share of tract land area.

A.2 NYC crime data

We computed our tract-level robbery statistics using confidential, geocoded incident-level reports provided to us by the New York Police Department. We aggregated robbery incidents to the census-tract level; we assigned each incident to a census tract based on a point-in-polygon matching strategy using ESRI's ArcMap software. We computed the average annual robberies over 2007-2011 for each census tract.

Historical crime rates are not available at a level of geographic detail below the NYPD precinct. Precinct-level crime statistics for 1990 are available on the NYPD webpage for each precinct. Precinct-level crime statistics for 2000 onwards are available from the NYPD website in an Excel file, Citywide Seven Major Felony Offenses by Precinct 2000-2014. We used a concordance between census blocks and NYPD precincts created by John Keefe to assign tracts to precincts based on the precinct in which the majority of the tract's population resides (92% of tracts have all component blocks in the same precinct). Our counterfactual 1990 robbery levels are inferred from tract-level robbery levels in 2007-2011 and precinct-level growth rates from 1990 to the 2007-2011 average.

A.3 Yelp users data

We collected Yelp user's locational information in two rounds, starting from 50,000 Yelp users who reviewed a venue in the five boroughs of New York City prior to 14 June 2011. The first round was an intensive examination of reviews written by Yelp users in a randomly selected 25% sample of these users. The second round was a more selective examination of the remaining 75% sample based on the the first round's lessons for successfully locating users. We restricted our estimation sample to users whose set of home locations was made up of venues all within 1.5 miles (2.414 kilometers) of each other. We imposed the same restriction on the set of work locations.

A.3.1 Yelp users data: First round

Between 1 January 2005 and 14 June 2011, users in the 25% sample wrote about 230,000 reviews of venues in New York and New Jersey. To identify residential and workplace locations, research assistants examined the text of reviews that contained at least one of 26 key phrases. Those key phrases are ten home-related phrases {I live, my apt, my apartment, my building, my neighborhood, my house, my place, my hood, my block, laundr}, seven work-related phrases {I work, coworker, colleague, lunch break, my office, my work, my job}, and nine phrases related to both {my local, delivery, block away, block from m, blocks from m, close to me, close to my, minutes from m, street from m}.

Parsing review texts, we flagged 16,425 reviews containing these phrases. Research assistants identified twenty-one percent of these flagged reviews as identifying a user's home location, and eleven percent of them as identifying a workplace. Reviews containing multiple home-related phrases identified a user's home location in 54% of cases; reviews with multiple work-related phrases yielded a work location 45% of the time.

This process identified about 1500 users with a residential location, 575 users with a workplace, and 450 users with both home and work locations. Thus, we found locational information for nearly one-fifth of the Yelp users we examined. The median user for which we obtained locational information had reviewed twenty venues in New York and New Jersey, while the median for which we obtained no information had reviewed five venues. Amongst users with more than ten reviews of NY/NJ venues, we obtained locational information for about 40%.

We identified individual who changed their residential and workplace locations via two means. First, research assistants reported any moves identified in the text of reviews containing the 26 key phrases above. Second, we reviewed the text of reviews containing at least of four key phrases: we moved, I moved, moving into, moving here.

This first round yielded 241 users who appear in our estimation sample.

A.3.2 Yelp users data: Second round

In the second round with the remaining 75% of users, we limited our examination to reviews that were likely to yield both home and work locations for a user. We investigated the text of 6,426 reviews of venues in New York City written by 569 users with at least one review containing two home-related phrases and at least one review containing two work-related phrases. In this round, we did not examine reviews in which the only key phrase

was "delivery". We used workers on Amazon's Mechnical Turk marketplace to classify the text. This work was performed in triplicate, and we only use observations with unanimous responses.

This process investigated 569 users and identified home locations for 173 users, work locations for 38 users, and both locations for 304 users. After imposing the other restrictions described in the main text, this second round yielded 165 users who appear in our estimation sample.

A.4 Yelp venue data

Yelp describes venues' locations by their street addresses. We translated these addresses to latitude-longitude coordinates and assigned venues to census tracts. We determined the latitude and longitude of each venue by a combination of methods. First, we matched the venue addresses to a point using the address locators provided by the New York City Department of Urban Planning and StreetMap North America. For venues with an incorrect ZIP code, we used the borough in the text of the venue's address. For venues not matched using these address locators, we used an alternative address located via an online geocoding service FindLatitudeAndLongitude. For the addresses that could not be matched using ESRI or the online service, we found the coordinates using GoogleMaps on a case-by-case basis. After determining venues' coordinates, we assigned each venue to a census tract based on a point-in-polygon matching strategy.

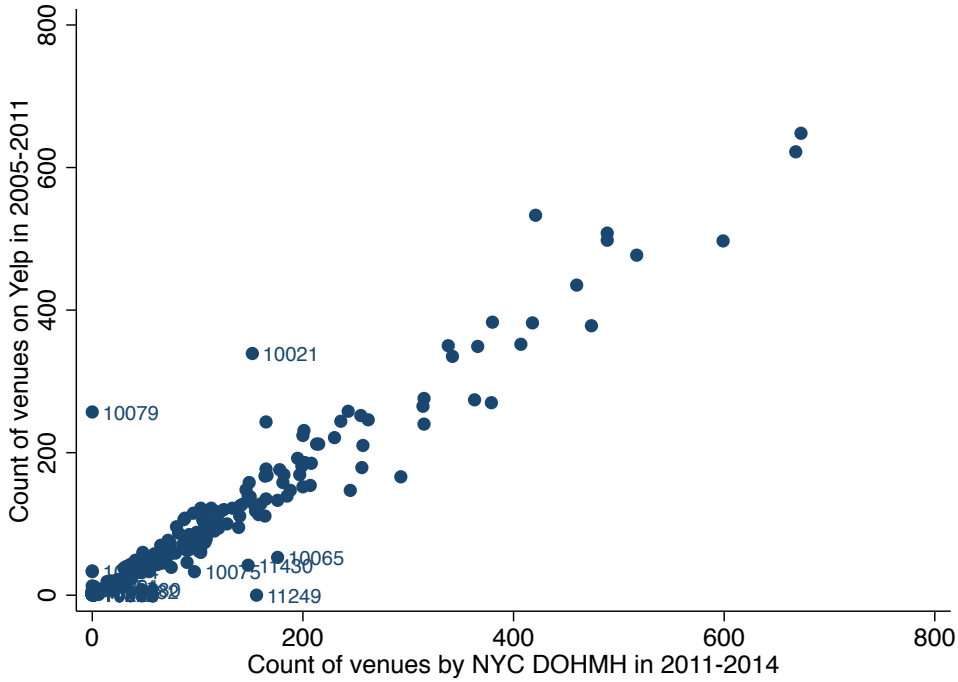
We created nine cuisine dummies by aggregating Yelp cuisine classifications into the following categories: African, American, Asian, European, Indian, Latin American, Middle Eastern, vegetarian. The omitted cuisine category is "unassigned" cuisine types, which includes venues with cuisine listed as "restaurant" on Yelp.

The Yelp venues included in our estimation sample as possible elements of users' choice sets meet three criteria. First, they had been reviewed at least once as of 2011. Second, they had both a rating and price listed on Yelp as of 2011. Third, they were located in a census tract for which a median household income estimate is available.

As one means of validating Yelp's venue coverage, we compare our count of Yelp restaurants by ZIP code to the number of establishments reported in health inspections data by the New York City Department of Health & Mental Hygiene. The DOHMH data report inspection results for 2011-2014, while our Yelp venue data, downloaded in 2011, covers trips to venues between 2005 and 2011. Despite this temporal mismatch, the two data sources report similar venue counts at the ZIP-code level, as shown in Figure A.1.[39]

[39]The outliers are often attributable to the temporal mismatch. The 10021 ZIP code was split into three components in 2007, creating 10065 and 10075 (Sam Roberts, "An Elite ZIP Code Becomes Harder to Crack", *New York Times*, 21 March 2007). A similar story explains 11211 and 11249 (Joe Coscarelli, "Williamsburg Hipsters Robbed of Prestigious 11211 Zip Code", *Village Voice*, 2 June 2011). 11430 is JFK Airport. The 10079 ZIP code does not exist; it appears to be a placeholder on Yelp.

Figure A.1: Venue counts by ZIP code, Yelp vs NYC DOHMH



B Econometrics

This appendix discusses how the assumptions made in section 3 with respect to review writing, exogenous locations and preference heterogeneity could be relaxed.

B.1 Correlated review writing

In this section, we discuss the potential bias in our estimates of the preference parameter β in the case in which the probability that a user reviews a restaurant depends on some characteristic of this restaurant that is observed by the econometrician and that affects individuals' dining choices. As indicated in the main text, the vector of covariates X_{ij}^2 accounts for both a broad set of restaurant characteristics and characteristics of the census tract in which it is located. Let's denote as X_{ij}^{2a} the subvector of X_{ij}^2 capturing restaurant characteristics and as X_{ij}^{2b} the subvector of X_{ij}^2 capturing characteristics of the census tract in which it is located. Analogously, we split the parameter vector β^2 into the two subvectors β^{2a} and β^{2b}.Assume that the probability that individual i reviews restaurant j upon visiting j at period t depends on some of the restaurant characteristics included in the vector X_{ij}^{2a} through the following function

$$P(d_{ijt}^r | d_{ijt} = 1, X_{it}) = w_{it} \mathbb{1}\{j \neq 0, j \neq D_{it}^r\} \exp(\gamma X_{ij}^{2a}). \tag{8}$$

The case in which $\gamma = 0$, this function is identical to that assumed in the main text an employed in equation (4). Using this expression, we can write the probability that we observe a review at period t written by user i about restaurant j as:

$$P(d^r_{ijt} = 1|X_{it};\beta) =$$
$$w_{it} \times \mathbb{1}\{j \neq 0, j \neq D^r_{it}\} \times \frac{\Big(\sum_{l\in\mathcal{L}} \exp(\beta^1_l X^1_{ijl} + (\beta^{2a} + \gamma)X^{2a}_{ij} + \beta^{2b} X^{2b}_{ij})\Big)}{\sum_{j'\in J_t}\Big(\sum_{l\in\mathcal{L}} \exp(\beta^1_l X^1_{ij'l'} + (\beta^{2a} + \gamma)X^{2a}_{ij'} + \beta^{2b} X^{2b}_{ij'})\Big)}, \quad (9)$$

As it is clear from this expression, we cannot separately identify the parameter vectors β^{2a} and γ. However, the estimates of the parameter vector β^1_l and β^{2b} are not affected by the dependency of the probability of writing a review on the vector of restaurant characteristics X^{2a}_{ij}. Furthermore, the probability in equation (9) is identical to that in equation (4) with the only exception that $\beta^2 = (\tilde{\beta}^{2a}, \beta^{2b})$ with $\tilde{\beta}^{2a} = \beta^{2a} + \gamma$ takes the place of the parameter vector β^2 in the main text. Therefore, following the same steps indicated in the main text, one can derive a likelihood function that identifies the parameter vector $(\beta^1, \tilde{\beta}^{2a}, \beta^{2b})$. This means that an expression for the probability of writing a restaurant review as that in equation (8) does not prevent us from obtaining consistent estimates of the preference parameters reflecting the impact of travel time and characteristics of the census tract in which the restaurant is located.

B.2 Endogenous home and work locations

The statistical model described in Section 3 implicitly assumes that individuals' home and work location are exogenously determined. However, in practice, individuals optimal choose where to live and work and, consequently, the home and work location of every individual observed in our sample might be endogenously determined as a function of the restaurant characteristics that they might consider visit. In this section, we discuss the assumptions that one would need to impose to guarantee that the endogenous selection of home and work does not bias the estimates of the preference parameters $\{(\beta^l_1, \beta_2), l \in \mathcal{L}\}$ computed following the estimation approach in Section 3.

Assume that, in some period 0, individuals choose their home and work locations by maximizing a utility function that is a weighted average of: (a) the expected utility that they will obtain in future periods from visiting restaurants from those locations (or from the commuting path); (b) characteristics of the home and work locations that have intrinsic value independently of their properties as sites from where to launch consumption.

In order to compute the expected utility of visiting restaurants when deciding on home and work location, we need to make an assumption on the content of agents' information sets at the time at which these decisions are taken. Using the notation in Section 3.1, let's assume that, at the time of deciding on where to live and work, every individual i knows the value of the vector $\{(X^1_{ijlt}, X^2_{ijt}); l \in \mathcal{L}, j \in J_t\}$ for any period $t > 0$ at which i will visit a restaurant. Conversely, assume that individuals know the distribution but ignore the realizations of the preference shocks $\{\nu_{ijlt}; l \in \mathcal{L}, j \in J_t\}$ for any period t posterior to that when the decision on residence and work is taken. Under this assumption, the expected utility for individual i

of visiting restaurants from a particular home and work location (h, w) is

$$V_{ihw} = \sum_{t \in \mathcal{T}} \log \Big(\sum_{j}^{J_t} \sum_{l \in \mathcal{L}} \exp(\beta_l^1 X_{ijlt} + \beta^2 X_{ijt}) \Big) + \gamma, \tag{10}$$

where γ is the Euler's constant, and, as a reminder, J_t is the set of all restaurants open for business at period t, and $\mathcal{L}$ is the set of possible origin-transportation mode pairs that an individual with home in location h and workplace in w may use to visit a restaurant, and $\mathcal{T}$ is the set of periods at which individual i will visit a restaurant.

For every possible pair of home and work locations (h, w) and individual i, let's denote the vector of characteristics defining their intrinsic value, independently of their properties as locations from where to launch consumption, as Z_{ihw}.

If we define as α the weight that individuals assign to the expected utility that they will obtain from visiting restaurants, we can write the utility for an individual i of living in location h and working in location w as

$$W_{ihw} = (1 - \alpha)\omega Z_{ihw} + \alpha V_{ihw},$$

where ω is a parameter vector of identical dimensions as Z_{ihw} that determines the impact of each of the characteristics in the utility for individual i of establishing her residence in location h and her workplace in location w. An individual i lives in location h^* and works in location z^* if

$$(h^*, w^*) = \arg \max_{w \in \mathcal{W}, h \in \mathcal{H}} \{(1 - \alpha)\omega W_{ihw} + \alpha V_{ihw}\}, \tag{11}$$

where $\mathcal{W}$ denotes the set of all possible work locations and $\mathcal{H}$ denotes the set of all possible home locations.

Extending the model of restaurant choice in Section 3 to account for the endogenous selection of home and work location, note that the relevant empirical probability would be the probability that an individual i chooses to visit restaurant j at period t *conditional on having chosen to live in location* h^* *and work in location* w^*:

$$P(d_{ijt} = 1 | X_{it}, h_i = h^*, w_i = w^*; \beta; \alpha) = \sum_{l \in \mathcal{L}} P(d_{ijlt} = 1 | X_{it}, h_i = h^*, w_i = w^*; \beta; \alpha),$$

and

$$P(d_{ijlt} = 1 | X_{it}, h_i = h^*, w_i = w^*; \beta; \alpha) =$$

$$\int_{\nu_{it}} \mathbb{1}\{\beta_1^l X_{ijlt}^1 + \beta_2 X_{ij}^2 + \nu_{ijlt} > \beta_1^l X_{ij'l't}^1 + \beta_2 X_{ij'}^2 + \nu_{ij'l't}; j' \in J_t, l \in \mathcal{L}\} f(\nu_{it} | X_{it}, h_i = h^*, w_i = w^*; \alpha, \beta) d\nu_{it},$$

where $\nu_{it} = \{\nu_{ijlt}; j \in J_t, l \in \mathcal{L}\}$ and $f(\nu_{it} | X_{it}, h_i = h^*, w_i = w^*; \alpha)$ denotes the density function of the vector ν_{it} conditional on the vector of observed characteristics determining the utility of restaurant visits, X_{it}, and conditional on the observed house and work locations h_i and w_i being the optimal choices of individual i. Using equation (11), we can rewrite this density function as:

$$f\big(\nu_{it} | X_{it}, (h^*, w^*) = \arg \max_{w \in \mathcal{W}, h \in \mathcal{H}} \{(1 - \alpha)\omega W_{ihw} + \alpha V_{ihw}\}\big).$$

This representation of the density function shows clearly that the conditioning set is a function of the vectors of observed covariates $X_i = \{X_{it}, t \in \mathcal{T}\}$ and $\{W_{ihw}, h \in \mathcal{H}, w \in \mathcal{W}\}$.[40] Therefore, we can recover the choice probability in equation (3) in the main text as long as we assume that the distribution of the vector of unobserved restaurant characteristics affecting individuals' restaurant choices, $\{\nu_{ijlt}; l \in \mathcal{L}, j \in J_t\}$ for any individual i and period t verifies two conditions: (a) it is independent of the vector of characteristics determining the optimal selection of home and work location, X_i and $\{W_{ihw}, h \in \mathcal{H}, w \in \mathcal{W}\}$; (b) it is distributed type I extreme value. The model in Section 3 already imposes that the distribution of the vector $\{\nu_{ijlt}; l \in \mathcal{L}, j \in J_t\}$ and distributed type I extreme value. Therefore, under the model for home and work location described above, allowing individuals to optimally determine their home and work will not bias the estimates described in Section 3 as long as we impose the additional restriction that $\{\nu_{ijlt}; l \in \mathcal{L}, j \in J_t\}$ is independent of the vector $\{W_{ihw}, w \in \mathcal{W}, h \in \mathcal{H}\}$ conditional on X_i.

B.3 Moment inequalities

Katz (2007) and Pakes (2010) show that there is an alternative estimation approach that uses moment inequalities and that would allow both to handle potentially large unobserved choice sets and heterogeneity in the individuals' preferences for some observed restaurant characteristics. In this approach, the utility for an individual i of visiting venue j in period t from origin-mode l may be written as:

$$U_{ijlt} = \beta_{li}^1 \mathbb{E}[X_{ijl}^1 | \mathcal{I}_{it}] + \beta_i^2 \mathbb{E}[X_{ij}^2 | \mathcal{I}_{it}], \tag{12}$$

with $\beta_{li}^1 = \beta_l^1 + \varepsilon_{li}^1$, $\beta_l^2 = \beta^2 + \varepsilon_i^2$, and $\mathcal{I}_{it}$ denoting the information set of individual i at the time of deciding which restaurant to visit at period t. Under the assumption that $\mathbb{E}_i[\varepsilon_{li}^1] = \mathbb{E}_i[\varepsilon_i^2] = 0$, where $\mathbb{E}_i[\cdot]$ denotes the expectation across individuals in the population of interest, Katz (2007) and Pakes (2010) show how to derive moment inequalities that will generate bounds for the average preference parameters β_l^1 and β^2. The behavioral model in equation (12) differs from that in equation (1) in that: (a) allows consumers to have imperfect information about the characteristics of the different restaurants at the time of deciding on which restaurant to visit; (b) allows individuals to differ in their preferences for the different observed restaurant characteristics included in the vector X_{it}; (c) assumes that there is no additional individual-restaurant-origin-mode specific characteristics that affects individual choices and is unobserved to the econometrician (i.e. assumes away the logit shock included in equation (1)).[41]

There are three reasons why we chose to estimate the model described in Section 3 instead of the behavioral model based on equation (12). First, the restaurant and locational characteristics included in the vector X_{ijt} are publicly available through Yelp.com, Google Maps, and SocialExplorer.com, so it is unlikely that individuals make large mistakes when

[40] Note that $\{V_{ihw}, h \in \mathcal{H}, w \in \mathcal{W}\}$ is itself a function of X_i.

[41] Dickstein and Morales (2015) show how to estimate a discrete choice model in which consumers may have imperfect information about the choice characteristics and their choices may be affected by individual-choice specific unobserved shocks. However, the estimator introduced in this paper is only applicable to binary choice problems and, therefore, cannot handle the large choice set that consumers face in our empirical application.

forecasting variables like the time that it takes to travel to a venue or the average price of each restaurant.[42] Second, while using moment inequalities to estimate and perform inference on bounds on a small set of parameters is computationally straightforward (e.g. Holmes, 2011; Eizenberg, 2014; Morales, Sheu, and Zahler, 2015; Dickstein and Morales, 2015; Wollman, 2015), doing so for the set of parameters that we estimate in some of our specifications (i.e. those accounting simultaneously for spatial and social frictions) is computationally unrealistic. [43] Third, even if we are controlling for a large set of observed restaurant characteristics, it is likely that there are still multiple unobservable factors (e.g. is the restaurant child-friendly? do the other people in my party like the restaurant? do I feel like eating at a French restaurant today?) that may vary across individual-restaurant-occasions triplets and that are likely to be important in determining individuals' restaurant picks. The model in Section 3 accounts for all these different factors through the unobserved preference shock ν_{ijlt}; conversely, the behavior model in equation (12) assumes these factors away.

C Robustness checks

C.1 Gender-specific coefficients

Table C.1 presents estimation results for specifications with gender-specific coefficients on demographic differences, analogous to the first three columns of Table 7. While the estimates suggest that female users are less likely to visit venues located in census tracts that are both demographically distant and demographically isolated, as captured by the triple interaction of Euclidean demographic distance, the spectral segregation index, and the female dummy, these coefficients are imprecisely estimated and not statistically significant. When we allow for asymmetries in the demographic differences, none of the estimated coefficients demonstrate differential responses across genders at the 5% level of statistical significance.

C.2 Number of origins

Section 4 presents estimation results for specifications with three origins: home, work, and commuting. This appendix section demonstrates that including multiple origins is important for some of our results. When we specify home as the only possible origin for the trip, we estimate qualitatively similar but quantitatively distinct coefficients on spatial and social frictions. Table C.2 demonstrates this by reporting specifications akin to those in Tables 7, 10, and C.1 estimated under the assumption that home is the only origin for visits to restaurants. In columns 1, 3, and 5, we use the sample estimation sample as in the main text. In columns 2, 4, and 6, we include all users whose home location meets our inclusion criteria. Not requiring workplace information more than quadruples the number of users and more than doubles the number of reviews in the sample. The additional observations

[42] While the NYPD only started making incident-level crime maps available on its website in December 2013, precinct-level crime statistics have been available on the NYPD website since 2003 and updated weekly. During our study period of 2007-2011, local newspapers like the *New York Times* produced incident-level maps based on felony reports.

[43] Applying the standard inference procedure to compute confidence sets (Andrews and Soares, 2010) for the large number of characteristics included in our exercise would be computationally prohibitive.

Table C.1: Gender-specific coefficients for demographic differences

	(1)	(2)	(3)
EDD between h_i and k_j	$-.942^a$	-1.00^a	-1.53^a
	(.112)	(.116)	(.200)
SSI of k_j		$.058^c$	.029
		(.034)	(.039)
EDD × SSI		.075	-.001
		(.083)	(.113)
EDD × female	$-.228^c$	-.117	.278
	(.131)	(.138)	(.235)
SSI × female		.049	.059
		(.035)	(.040)
EDD × SSI × female		$-.279^b$	-.218
		(.113)	(.141)
EDD × h_i is plurality Asian			-.187
			(.403)
EDD × h_i is plurality black			-.038
			(2.08)
EDD × h_i is plurality Hispanic			$.894^c$
			(.476)
EDD × h_i is plurality Asian × female			.793
			(.538)
EDD × h_i is plurality black × female			-.016
			(2.13)
EDD × h_i is plurality Hispanic × female			.029
			(.593)
k_j is plurality Asian			$.299^c$
			(.178)
k_j is plurality black			.462
			(.564)
k_j is plurality Hispanic			.101
			(.205)
EDD × k_j is plurality Asian			$.806^c$
			(.415)
EDD × k_j is plurality black			-2.01^c
			(1.18)
EDD × k_j is plurality Hispanic			.128
			(.526)
k_j is plurality Asian × female			-.103
			(.233)
k_j is plurality black × female			-.326
			(.631)
k_j is plurality Hispanic × female			.311
			(.253)
EDD × k_j is plurality Asian × female			-.661
			(.518)
EDD × k_j is plurality black × female			1.00
			(1.32)
EDD × k_j is plurality Hispanic × female			-1.04
			(.638)
Number of origin-mode points	6	6	6
Number of venues in choice set	20	20	20
Log-Likelihood	-2.31	-2.31	-2.30
Pseudo R-sq	.228	.228	.230
Akaike Information Criterion	112	120	156
Number of trips	15619	15619	15619
Number of individuals	385	385	385

NOTES: Each column reports an estimated multinomial logit model of individuals' decisions to visit a Yelp venue. Standard errors in parentheses. Statistical significance denoted by a (1%), b (5%), c (10%). "EDD" is Euclidean demographic distance; "SSI" is spectral segregation index. The unreported covariates are log travel times from six origin-mode pairs, venue price and rating interacted with home tract income, log median household income in tract of venue, percentage difference and percentage absolute difference in income levels, and 28 area dummies, 9 cuisine dummies and a female user dummy interacted with venue price and rating, log median household income in tract of venue, and percentage difference and percentage absolute difference in income levels.

tend to reduce the standard errors without substantially altering the point estimates of the coefficients. However, these home-only specifications yield less negative coefficients on Euclidean demographic distance and much greater female responsiveness to robberies per resident than those using all three origins. This suggests that incorporating information on both home and work locations is key to describing how people use the city.

C.3 Choice-set size

Table C.3 shows that our results are robust to resampling the choice set of size 20 (column 1) or constructing choice sets with more elements (columns 2-3). The estimated coefficients are stable across sampled choice sets, consistent with the independence-of-irrelevant-alternatives assumption.

Table C.2: Estimation results with one origin

	(1)	(2)	(3)	(4)	(5)	(6)
Average annual robberies per resident in k_j			-2.85^a	-3.14^a	-1.40^b	-1.22^b
			(.463)	(.321)	(.669)	(.580)
Average annual robberies per resident in k_j × female					-2.19^b	-2.38^a
					(.885)	(.756)
EDD between h_i and k_j	$-.720^a$	$-.712^a$	$-.682^a$	$-.680^a$	$-.823^a$	$-.883^a$
	(.116)	(.075)	(.116)	(.075)	(.159)	(.128)
SSI of k_j	$.082^a$	$.070^a$	$.070^a$	$.053^a$	-.026	-.038
	(.022)	(.015)	(.024)	(.016)	(.051)	(.043)
EDD × SSI	$-.200^a$	$-.146^a$	$-.357^a$	$-.244^a$	-.045	.093
	(.068)	(.042)	(.123)	(.071)	(.157)	(.109)
EDD × female					.159	.040
					(.164)	(.130)
SSI × female					$.114^b$	$.094^b$
					(.050)	(.043)
EDD × SSI × female					$-.397^a$	$-.393^a$
					(.153)	(.119)
EDD × h_i is plurality Asian	.249	.172	.411	$.362^b$	$.457^c$	.025
	(.262)	(.173)	(.265)	(.175)	(.271)	(.228)
EDD × h_i is plurality black	-.330	$-.665^a$	$-.768^c$	$-.945^a$	$-.968^b$	-1.34^a
	(.398)	(.250)	(.412)	(.261)	(.433)	(.382)
EDD × h_i is plurality Hispanic	.215	.070	.197	.057	.339	.105
	(.290)	(.175)	(.291)	(.176)	(.302)	(.252)
k_j is plurality Asian	$.264^b$	$.127^c$	$.260^b$	.125	.206	$-.223^c$
	(.113)	(.073)	(.128)	(.084)	(.132)	(.114)
k_j is plurality black	.137	$.297^b$	.187	$.338^c$	-.056	-.188
	(.243)	(.143)	(.313)	(.176)	(.335)	(.287)
k_j is plurality Hispanic	.199	$.289^a$	$.263^b$	$.320^a$	$.273^b$	$.307^a$
	(.122)	(.077)	(.124)	(.078)	(.127)	(.108)
EDD × k_j is plurality Asian	.104	$.470^a$	.060	$.407^b$	.153	1.15^a
	(.243)	(.157)	(.281)	(.183)	(.290)	(.242)
EDD × k_j is plurality black	-1.39^a	-1.85^a	-.899	-1.55^a	-.262	.075
	(.521)	(.310)	(.775)	(.439)	(.809)	(.672)
EDD × k_j is plurality Hispanic	$-.625^b$	$-.925^a$	$-.800^b$	$-.994^a$	$-.812^b$	$-.824^a$
	(.307)	(.192)	(.314)	(.197)	(.323)	(.263)
SSI × k_j is plurality Asian			$.307^a$	$.303^a$	$.381^a$	$.407^a$
			(.094)	(.064)	(.099)	(.090)
SSI × k_j is plurality black			$-.925^b$	$-.547^b$	-.722	$-.847^b$
			(.460)	(.267)	(.478)	(.412)
SSI × k_j is plurality Hispanic			-.297	-.064	-.254	-.148
			(.279)	(.102)	(.286)	(.201)
EDD × SSI × k_j is plurality Asian			-.192	-.225	-.352	$-.474^b$
			(.223)	(.142)	(.229)	(.188)
EDD × SSI × k_j is plurality black			.008	.058	-.633	-1.47^c
			(.914)	(.540)	(.953)	(.809)
EDD × SSI × k_j is plurality Hispanic			$.807^c$	.220	.687	.394
			(.471)	(.205)	(.481)	(.331)
Number of origin-mode points	2	2	2	2	2	2
Number of venues in choice set	20	20	20	20	20	20
Log-Likelihood	-2.34	-2.28	-2.34	-2.28	-2.33	-2.29
Pseudo R-sq	.216	.235	.217	.236	.219	.233
Akaike Information Criterion	124	124	138	138	164	164
Number of trips	16573	39807	16573	39807	15619	24515
Number of individuals	406	1810	406	1810	385	904

NOTES: Standard errors in parentheses. Statistical significance denoted by *a* (1%), *b* (5%), *c* (10%). The unreported covariates are log travel times from home-public and home-car and the unreported controls in Table 6. Columns 5 and 6 also include a female user dummy interacted with venue price and rating, median household income in tract of venue, and percentage difference and percentage absolute difference in income levels as unreported covariates.

Table C.3: Varying choice set sizes

	(1)	(2)	(3)
Log of travel time from home-public	-1.17^a	-1.14^a	-1.15^a
	(.072)	(.066)	(.065)
Log of travel time from home-public × female	.046	.029	.037
	(.095)	(.088)	(.087)
Log of travel time from home-car	-1.38^a	-1.35^a	-1.36^a
	(.068)	(.064)	(.063)
Log of travel time from home-car × female	$.148^c$	.123	$.131^c$
	(.083)	(.079)	(.078)
Log of travel time from work-public	-1.70^a	-1.72^a	-1.71^a
	(.198)	(.205)	(.199)
Log of travel time from work-public × female	.224	.248	.249
	(.242)	(.246)	(.238)
Log of travel time from work-car	-1.88^a	-1.87^a	-1.86^a
	(.159)	(.152)	(.149)
Log of travel time from work-car × female	.058	.077	.072
	(.214)	(.204)	(.199)
Log of travel time from commute-public	-1.12^a	-1.10^a	-1.11^a
	(.051)	(.048)	(.048)
Log of travel time from commute-public × female	.080	.094	.098
	(.066)	(.062)	(.062)
Log of travel time from commute-car	-1.37^a	-1.36^a	-1.38^a
	(.056)	(.055)	(.055)
Log of travel time from commute-car × female	.114	$.119^c$	$.136^c$
	(.072)	(.070)	(.069)
Average annual robberies per resident in k_j, 2007-2011	-3.24^a	-2.82^a	-2.82^a
	(.704)	(.686)	(.672)
Average annual robberies per resident in k_j, 2007-2011 × female	-1.08	-1.50^c	-1.42
	(.923)	(.901)	(.889)
EDD between h_i and k_j	$-.947^a$	$-.969^a$	-1.04^a
	(.127)	(.123)	(.122)
SSI of k_j	$.070^b$	$.066^c$	.057
	(.035)	(.034)	(.034)
EDD × SSI	.035	.044	.080
	(.096)	(.087)	(.090)
EDD × female	-.115	-.164	-.079
	(.160)	(.155)	(.153)
SSI × female	.040	.040	.048
	(.035)	(.035)	(.035)
EDD × SSI × female	$-.213^c$	$-.209^c$	$-.241^b$
	(.120)	(.113)	(.114)
Number of origin-mode points	6	6	6
Number of venues in choice set	20	40	60
Log-Likelihood	-2.30	-2.97	-3.36
Pseudo R-sq	.229	.194	.177
Akaike Information Criterion	150	151	152
Number of trips	15619	15619	15619
Number of individuals	385	385	385

NOTES: Each column reports an estimated multinomial logit model of individuals' decisions to visit a Yelp venue. Standard errors in parentheses. Statistical significance denoted by a (1%), b (5%), c (10%). "EDD" is Euclidean demographic distance; "SSI" is spectral segregation index. The unreported covariates are log travel times from six origin-mode pairs, the unreported controls in Table 6, and a female user dummy interacted with venue price and rating, log median household income in tract of venue, and percentage difference and percentage absolute difference in income levels.

D Incidence of the fall in crime

In section 4.3, we discuss the quantitative significance of users' aversion to restaurants in high-robbery areas in terms of the massive decline in New York City crime in the last twenty-five years. The numbers in the main text are the utility gains associated with a change in robbery rates, computed across locations and genders while holding residents, venues, and all other aspects of the city fixed. A natural question is whether the welfare gains we infer from this exercise would in fact be realized in a broader, general equilibrium in which spatial arbitrage occurs. If a location's improvement calls forth greater demand that raises local land prices, then some or all of the welfare gains we estimate might be enjoyed by landowners rather than incumbent residents. Naturally, the division of these gains depends on the elasticities of supply and demand.

To simplify, assume that the demand of residents to live in any census tract is continuous, so that the marginal resident is indifferent between her tract and another location. Assume a competitive housing market, so that a tract's equilibrium housing price is determined by the marginal resident's valuation of that tract. We assume that there are locations outside New York City unaffected by the changes we consider and that the marginal residents of some tracts are indifferent between their tracts and these outside locations.

First, consider an extreme case in which all census tracts receive a common positive valuation shock and housing supplies are perfectly inelastic. Housing prices must rise to exactly offset the valuation gain so that the marginal resident remains indifferent between her tract and another location. In this case, residents' welfares are unchanged and all gains accrue to the landowners.

Now consider the other extreme in which there is a common positive shock and housing supplies are perfectly elastic. In this case, tracts' housing prices are unchanged and the population of New York City increases as individuals move into the city until the marginal resident is again indifferent between a tract in NYC and some outside location. Those whose residence is unchanged experience gains equal to the valuation increase, while those who changed census tracts within New York will have experienced an even larger increase.

The estimated welfare gains are unlikely to have been entirely capitalized into housing prices. First, while the city's land supply is fixed, the housing supply is not. Over the last twenty-five years, substantial amounts of new housing were built in New York City. Second, institutional features of the city's housing market, such as rent stabilization, limit the extent to which local improvements pass through to landlords. Where rent ceilings are binding, the gains accrue to residents.

Evaluating the incidence of heterogeneous improvements, such as the decline in crime that is valued more by women than men, is slightly more complicated. Consider the perfectly inelastic case. The housing price will increase, but it will rise less than the increase in the location's value to women. The logic is simple. Absent a price increase, more women would enter. If the price increase fully offset the gain to the marginal female resident, some men would depart. Thus there are important welfare consequences to such changes even when housing is inelastically supplied. Inframarginal (incumbent) women gain the difference between the valuation increase and the price increase, while this is an upper bound on the gains enjoyed by new female entrants. Inframarginal (incumbent) men lose due to the uncompensated price increase, while this price rise is an upper bound on departing males'

losses. In both cases, the total effect depends on their next best alternative. If instead housing were perfectly elastically supplied in the relevant range, there would be no gains to landowners, full gains to inframarginal women, and no change for men.