Homophily and Sorting Within Neighborhoods

Gregorio Caetano and Vikram Maheshri*

June 29, 2015

Abstract

Homophily, or the tendency of similar people to associate with each other, plays an important role in the formation of peer groups. One result of this phenomenon is that policy interventions that involve assigning subjects to different environments such as neighborhoods, schools or workplaces may fail to change the types of peers to which they are exposed. Using a new large data set from Foursquare, a popular mobile app that documents the activity of millions of users, we show that gender and age homophily is widespread and highly local: at least half of the sorting across tens of thousands of venues in eight major US cities happens within census blocks. This generates a wedge between the levels of neighborhood diversity and venue diversity to which people are exposed. We show that this wedge is primarily mediated by the local supply of venues. Neighborhoods with a greater variety of venues tend to attract more diverse visitors, but they sort more intensely across venues within the neighborhood. The resulting reduction in diversity in venues may lead to narrower social interactions. *JEL Codes: R10, R23, R33.*

1 Introduction

Similar people have long been observed to associate with each other, a phenomenon known as homophily (McPherson et al. (2001)). This self-selection into groups, or sorting, manifests itself in decisions as varied as selecting a neighborhood, visiting a restaurant or even joining a conversation. Such sorting can occur because similar individuals are attracted to the same environments or because similar individuals are attracted to the same types of people. Irrespective of the reason, this phenomenon directly affects the formation of individuals' peer groups.

*University of Rochester and University of Houston. We are extremely grateful to Foursquare, Inc. for generous access to their restricted, anonymized data, and especially to Michael Li and Blake Shaw for helpful discussions about the data. We thank Alexei Alexandrov, Dionissi Aliprantis, Dani Arribas-Bel, Donald Davis, Joan Monras, Richard Murphy, Romans Pancs, Stephen Ross, and various seminar and conference participants for helpful discussions. We also thank Riley Hadden and Hao Teng for excellent research assistance. All errors are our own. The appendix is available at http://www.gregoriocaetano.net/resources/Research/Homophily_Neighborhoods_Appendix.pdf.

Although many policy interventions have sought to expose individuals to different peers by manipulating their assignment to neighborhoods, schools, workplaces, and other environments, in reality, individuals do not interact uniformly with each other in these settings. While some individuals become friends and acquaintances, others may never meet. Homophily implies that sorting will lead to a less diverse peer group for an individual than the one encouraged by their assignment, which may substantially attenuate the intended effect of the intervention. Indeed, the formation of endogenous peer groups has been shown to offset the effects of assignment in small-scale randomized controlled trials (Carrell et al. (2013)) and has been suggested as an explanation for the surprisingly small neighborhood effects found in some larger randomized controlled trials such as the "Moving to Opportunity" program (Weinberg (2007)). Unfortunately, little is known about the pervasiveness of homophily within neighborhoods.

In this paper, we present large-scale evidence of gender and age homophily in the day-to-day activities of millions of individuals, and we show that their sorting into endogenous peer groups is mediated by specific features of the urban environment. We exploit a large and unique user-generated data set from a prominent location-based social network, Foursquare, to measure how individuals in eight major US cities - Atlanta, Chicago, Dallas, Los Angeles, Philadelphia, New York City, San Francisco and Washington, DC - sort into tens of thousands commercial and recreational venues such as restaurants, shops, parks and museums. We find that sorting happens quite locally: 80-90 percent of sorting by gender occurs between venues within census tracts, and, remarkably, over half of it occurs between venues within census blocks. (For some geographic perspective, Manhattan contains over 2,500 census blocks, and Chicago contains over 6,000 of them.)

Individuals sort into successively smaller areas from census tracts down to venues. We find that similar people tend to visit the same census tracts, the same census block groups within tracts, the same census blocks within block groups, and the same venues within blocks. Thus, we find evidence of homophily in all of the choices that we can observe in our dataset. By extension, our results likely understate the role of homophily in mitigating exposure to diverse peer groups since individuals likely sort further within venues at levels that we do not observe. For example, they may sort to different tables within restaurants, to different departments within stores or they may sort to the same venue at different times of day.

Local sorting across venues leads to a wedge between the general level of diversity to which

people are exposed in a neighborhood and the specific level of diversity to which they are exposed in a venue. We show that this wedge is primarily determined by the variety of venues that are on offer. Using a simple model of venue sorting in the spirit of Hotelling (1929), we illustrate how venue variety affects the levels of diversity observed both at the neighborhood level and at the venue level and show that the directions of both of these effects are theoretically ambiguous. We estimate these effects with three different identification strategies that rely on variation in the variety of venues in adjacent neighborhoods, the entry and exit of venues over time, and plausibly exogenous variation in zoning laws that all yield similar results. Greater venue variety attracts a more diverse set of individuals to a neighborhood, but once there, they sort more intensely across venues, thereby reducing the amount of diversity to which they are exposed. This finding underscores the difficulties involved in strengthening the social interactions that form the basis of thriving urban environments, as the endogeous sorting of individuals at very local levels may undermine well meaning place-based policies.

Social interactions have been proposed and identified as determinants of social learning and personality development (Blau (1964)), the evolution of social norms (Ostrom (2000)), the racial wage gap (Ananat et al. (2013)), technological adoption (Conley and Udry (2001)), job referrals (Bayer et al. (2008)), and the prevalence of crime (Glaeser et al. (1996)). A considerable theoretical and empirical literature has analyzed sorting into broader geographic areas such as cities, neighborhoods and school attendance areas (Roback (1982); Bayer et al. (2007)) with the implicit intent to gain insights into people's relevant interactions. In particular, exposure to female peers has been found to affect student achievement at the primary (Hoxby (2000); Whitmore (2005)), secondary (Lavy and Schlosser (2011); Hill (2015)), and post graduate (Arcidiacono and Nicholson (2005)) levels, bullying in schools (Espelage and Holt (2001)), substance abuse (Andrews et al. (2002)), corporate governance and performance (e.g., Brown et al. (2002); Adams and Ferreira (2009)), the expression of political beliefs (Huckfeldt (1995)), and the level of intimacy in social networks (Verbrugge (1977)).

All of our empirical findings - widespread and highly localized sorting, homophily in all location choices, and causal evidence that venue variety generates a wedge between neighborhood and venue diversity - are robust across all cities in our sample, from dense, older cities such as New York City to sprawling, new cities such as Dallas. Moreover, our results in full are qualitatively similar

when we conduct a parallel analysis of sorting by age. The robust patterns uncovered in this study may speak to the external validity of our results. In addition, we provide a detailed analysis of potential measurement error in our data, including a Monte Carlo study, to ensure the robustness of our results. This methodological contribution may be relevant to the growing number of empirical analyses of user-generated "big" data sets.

The remainder of the paper is organized as follows. In Section 2, we show that the endogenous formation of peer groups can substantially affect the interpretation of estimates of peer effects, even in randomized controlled trials. In Section 3, we describe our data set, and in Section 4, we show widespread evidence of homophily in location choices and sorting even in narrow areas. In Section 5, we explore the causes of this phenomenon with a simple model of sorting within neighborhoods, and we show empirically that the variety of venues on offer in neighborhoods impacts both the levels of diversity in venues and in neighborhoods, but in opposite directions. In Section 6, we study the role of measurement error in the analysis of user-generated data such as ours and show that our results are robust. A more detailed treatment of potential measurement error in our data is provided in the supplementary appendix. In Section 7, we briefly discuss the results of our parallel analysis of age sorting; the complete results are provided in the supplementary appendix. We conclude with a discussion of our findings in Section 8.

2 Homophily and the Endogenous Formation of Peer Groups

Although many policy interventions seek to expose individuals to different peers by manipulating their assignment to neighborhoods, schools, workplaces, and other environments, homophily will endogenously generate a less diverse peer group than the one implied by assignment. This can substantially attenuate the intended effects of the intervention (Weinberg (2007)). We illustrate this point with a simple example in the language of the potential outcomes framework.

Assume that there are two neighborhoods that vary in their gender composition: one has a high proportion of females, and the other has a low proportion of females. Every individual i in a neighborhood can further sort into a peer group of their choice, which is, for simplicity either one with a high proportion of females or one with a low proportion of females. In each neighborhood, it is always possible to find peer groups of both types. Researchers randomly assign a group of men from other neighborhoods to one of these two neighborhoods in order to understand the impact of

exposure to gender diversity. The assignment of man i can be denoted as $A_i \in \{0, 1\}$ where $A_i = 1$ if he is assigned to the neighborhood with a high proportion of women, and $A_i = 0$ if he is assigned to the neighborhood with a low proportion of women instead. In this context, researchers typically observe the assignment A_i and the outcome Y_i of each of those men, which allows them to estimate the "Intention-to-Treat" effect

$$ITT = \mathrm{E}\left[Y|A = 1\right] - \mathrm{E}\left[Y|A = 0\right] \tag{1}$$

Within a neighborhood, men join peer groups that can have either high gender diversity ($T_i = 1$) or low gender diversity ($T_i = 0$). The function $T_i\left(A_i\right)$ can be used to classify these men into one of four groups: compliers ($T_i\left(A_i\right) = A_i$), always takers ($T_i\left(A_i\right) = 1$), never takers ($T_i\left(A_i\right) = 0$) and defiers ($T_i\left(A_i\right) = 1 - A_i$). Compliers sort to a peer group that is predominantly composed of the majority gender in the neighborhood. Always takers sort to a mostly female peer group irrespective of the neighborhood to which they are assigned, and similarly, never takers sort to a mostly male peer group irrespective of the neighborhood to which they are assigned. Finally, defiers sort to a peer group that is predominantly composed of the minority gender in the neighborhood. Because similar individuals tend to associate with each other (i.e., because of homophily), men will tend to be never takers.[1] The function $T_i\left(A_i\right)$ allows us to write their outcomes as an implicit function of their assignment

$$Y_{iT_i} = T_i Y_{i1} + \left(1 - T_i\right) Y_{i0} \tag{2}$$

because the outcome is affected by exposure to peers as opposed to neighborhood assignment.

Assuming defiers do not exist, we can rewrite equation (1) as

$$ITT = \alpha_{\mathbb{C}}\left(\mathrm{E}\left[Y_{i1}|i \in \mathbb{C}\right] - \mathrm{E}\left[Y_{i0}|i \in \mathbb{C}\right]\right) \tag{3}$$

where $\mathbb{C}$ is the set of compliers, and α_C is its share of the population. Typically in this context researchers do not observe α_C, as they do not observe data in the level at which peer effects happen; they observe Y and A and directly estimate ITT via equation (1). But because of homophily, estimates of neighborhood effects will tend to be substantially smaller than the actual peer group

[1] If women were experimental subjects, homophily would lead most of them to be always takers.

effects since subjects would tend not to comply with their assignment (i.e, low $\alpha_{\mathbb{C}}$).

Thus, even if exposure to diversity in peers truly mattered, researchers might fail to find such an effect when randomly assigning individuals to different groups because the endogenous sorting of individuals into peer groups will leave them less exposed to diversity than intended. This endogenous formation of peer groups could explain the difficulty that researchers have faced in finding evidence of exposing poor individuals to more affluent neighborhoods in various large experiments such as "Moving to Opportunity" (Kling et al. (2007)) and the Metropolitan Toronto Housing Program (Oreopoulos (2003)). At the same time, it is consistent with the findings that exposure to roommates with high academic and social ability has a variety of effects on students (Sacerdote (2001)), and that workplace productivity increases with peer effects in two person teams (Mas and Moretti (2009)). After all, the assignment to small groups leaves little room for further endogenous sorting. These seemingly contradictory results highlight the need to understand the extent to which homophily operates within assignment groups.

3 Data

In order to analyze the localized sorting of individuals sorting, we require comprehensive, disaggregated data of their whereabouts across a large number of locations within small neighborhoods, which is difficult to observe directly. We circumvent this issue with novel, proprietary data from Foursquare, Inc., creators of the eponymous mobile app and social network that allows users to document their precise whereabouts electronically. Upon arriving at a venue, Foursquare identifies the venue by GPS on a user's mobile phone, and they can electronically "check in". We use information on the demographic composition of Foursquare users in each venue to construct a proxy for the actual demographic composition of all individuals (i.e., Foursquare and non-Foursquare users) in the venue. Although this raises important concerns of sample selection, we develop an empirical approach with these concerns specifically in mind. To that end, we provide a detailed treatment of potential sources of measurement error in our data and its implications for all of our empirical results in Section 6 and in the appendix. Ultimately, we argue that our particular empirical approach allows us to extract a meaningful signal about the sorting of all individuals across venues from the novel dataset provided.

Ours is the first study to use this large and highly detailed database of venue visitors to study

diversity within neighborhoods.[2] Foursquare is a particularly suitable data source for our analysis because it is a prominent location-based social network that boasts a large number of users (over 50 million worldwide as of March 2015) who are particularly active (users have made over 6 billion cumulative check-ins as of March 2015), which makes for a highly detailed catalog of activity.

Our data set contains information on all Foursquare activity in venues in eight major US cities: Atlanta, Chicago, Dallas, Los Angeles, New York City, Philadelphia, San Francisco and Washington, DC. Our specific sample regions are defined as the counties in which these cities are primarily located.[3] For each of the 76,377 venues that are tracked in these cities, Foursquare has directly provided to us in fully anonymized form the number of daily check-ins by male and female users from August 1, 2012 to July 31, 2013. This data is aggregated at the venue level, hence we cannot observe any characteristics of individual Foursquare users, nor can we track a particular individual's activity. We restrict our sample to venues that experienced at least 10 check-ins during the sample period to improve our measurements of the gender compositions of venues.[4] In total, these venues experienced 49.6 million check-ins during the sample period with the average venue in our sample experiencing 649 check-ins. Each venue in our data set is also geo-coded by latitude and longitude, which allows us to link venues to unique census tracts, block groups and blocks using neighborhood definitions from the 2010 Census.

In Table 1, we present the summary statistics for our sample. We summarize the sample by city and by venue classification. Not surprisingly, larger cities such as New York and Los Angeles have more venues and check-ins. Males tend to check in slightly more than females on average, but there is substantial, robust variation in the gender composition of venues in all cities. For each city in our sample, check-ins across venues are approximately distributed lognormally. It is immediate that there is more variation in the average gender composition of venues across categories than across cities: it ranges from 52% females in *Shops and Services* to 40% female in *Hotels*. The variation in gender composition within categories also ranges more widely than the variation in

[2] A small but growing number of studies (e.g., Arribas-Bel and Bakens (2014)) have begun to use Foursquare data obtained indirectly via the Foursquare API (application programming interface). Foursquare data obtained via the API unfortunately does not disaggregate check-ins along any demographic dimension.

[3] The counties are Fulton (Atlanta), Cook (Chicago), Los Angeles, New York, Philadelphia, and San Francisco respectively. We treat the entire District of Columbia as the "county" for Washington. Most of the cities in our sample are entirely contained in their corresponding county with the notable exception that New York County only contains the borough of Manhattan.

[4] We show empirically that this restriction does not bias our results (see appendix).

gender composition within cities. *Food* and *Shops and Services* are the most prevalent and heavily trafficked venue categories in our sample.

Table 1: Summary Statistics

City	**Venues**	**Check-ins**	μ	σ	$p_{75} - p_{25}$	**Tracts**	**B. Groups**	**Blocks**
Atlanta	4,115	2.84	0.46	0.17	0.19	180	361	1,307
Chicago	13,665	8.11	0.49	0.16	0.19	1,100	2,235	6,237
Dallas	5,065	2.40	0.45	0.16	0.19	421	774	1,986
Los Angeles	23,108	10.2	0.46	0.15	0.18	1,902	3,584	9,182
New York City	16,203	16.2	0.49	0.17	0.19	282	945	2,501
Philadelphia	3,933	2.10	0.47	0.16	0.19	301	568	1,757
San Francisco	6,601	4.78	0.42	0.15	0.16	182	440	1,898
Washington, DC	3,687	2.98	0.43	0.16	0.17	152	272	1,069
Category							**Unique Subcategories**	
Food	31,398	16.6	0.45	0.13	0.17		65	
Shops/Services	20,903	9.97	0.52	0.21	0.28		66	
Bars	6,441	6.52	0.44	0.12	0.13		20	
Outdoors	4,795	4.62	0.44	0.16	0.21		22	
Cafes	4,483	3.88	0.47	0.14	0.18		3	
Entertainment	4,189	4.08	0.46	0.13	0.15		29	
Hotels	1,798	2.24	0.40	0.11	0.13		5	
Gyms	1,625	1.41	0.49	0.23	0.34		12	
Spiritual	745	0.29	0.48	0.17	0.23		3	

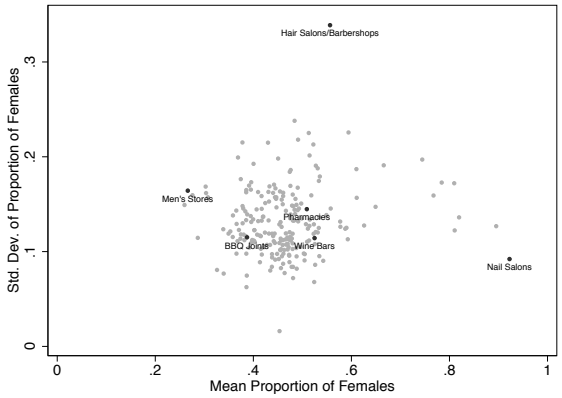
Notes: Check-ins reported in millions. μ and σ refers to the mean and standard deviation of the proportion of females in venues, and p_{25} and p_{75} refer to the 25^{th} and 75^{th} percentiles of the proportion of females in venues.

The 9 categories of venues that we observe are further classified into 225 narrow subcategories, and Foursquare users very actively check into even surprising types of venues such as *Banks*, *Cemeteries*, *Pharmacies, Synagogues*, and *Dog Runs*.[5] In Figure 1, we present a scatter plot of the mean and standard deviation of the gender composition of venues for each subcategory throughout our entire sample. Generally speaking, this summary of the gender composition of venues in subcategories looks intuitive and reasonable, and we highlight a few selected subcategories to convey some of the richness of the data. *Men's Stores*, not surprisingly, cater to mostly men, and this is fairly consistent across stores; conversely, *Nail Salons* cater to mostly women across all stores. *Hair Salons/Barbershops* cater to a mixed customer base in the aggregate; however the high standard

[5]Detailed summary statistics disaggregated by subcategory can be found in the appendix.

deviation of the gender composition of these venues suggests that they may serve very different clientele – either predominantly male or predominantly female. In contrast, *Wine Bars*, which exhibit a similarly mixed clientele in the aggregate seem to also exhibit this mixed gender composition at the venue level as well. These subcategorical scatter plots can be constructed separately for each city; although there are some small shifts in locations of the subcategories, their relative positions tend to be stable across cities.

Figure 1: Proportion of Females in Venues by Subcategory



Note: This scatter plot pools venues from all cities in the sample. Each dot represents all of the venues within a subcategory.

Because we observe daily check-ins at each venue, we can assess whether there are any dynamic trends in our data over the sample period. As shown in the first panel of Figure 2, there is substantial day-of-week variation in check-ins since venues are more highly frequented on weekends, but the gender composition of check-ins is nearly constant. This suggests that we can aggregate the data at least to the weekly level to analyze gender diversity. We do so and check for aggregate weekly trends in our data in the second panel of Figure 2. The results are strikingly similar: there is no systematic weekly variation in check-in frequency and no discernible seasonality or aggregate trend. More importantly, the gender composition of check-ins is roughly constant throughout the sample period. This suggests that we can aggregate the data set to the annual level to analyze gender diversity without loss of generality.

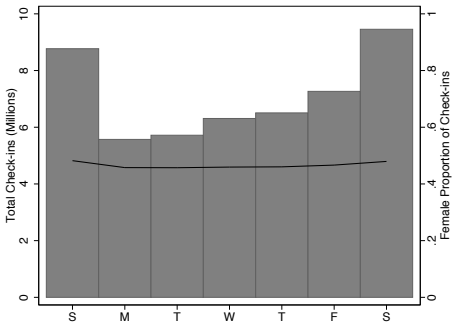
To further support this choice of aggregation, we check whether the gender compositions of indi-

vidual venues follow a trend over time. For each venue, we compute the net number of week-on-week increases (increases minus decreases) in the proportion of female check-ins over the sample period and we plot them against the total number of changes in the proportion of female check-ins in Figure 2.[6] Larger dots represent more venues in the sample, and the shaded region is defined to include 95% of all venues. It is immediate that most venues experience roughly as many relative increases in female popularity as relative decreases in female popularity. Because the gender composition of a venue has the tendency to vary around a fixed value, it is appropriate to interpret longitudinal variation in check-ins as measurement error, which we minimize by aggregating our data to the annual level to focus on the more salient cross-sectional variation in our data.[7]

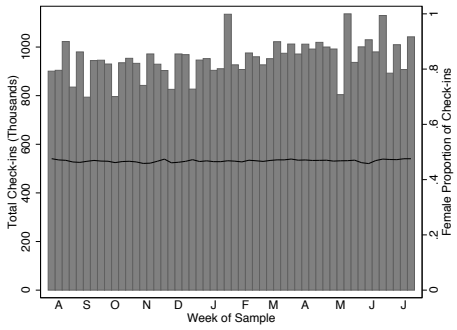
[6] A venue is defined to experience a week-on-week increase (decrease) in the female share if its female share increases (decreases) by a threshold of at least one percentage point, and the total number of changes in the proportion of female check-ins is equal to the sum of increases and decreases. We replicated panel (c) of Figure 2 with alternative thresholds of 5, 10 and 15 percentage points and obtained qualitatively similar results.

[7] As a robustness check, we replicated all main results of the paper by month-of-year and by day-of-week and found similar results, which are presented in the appendix.

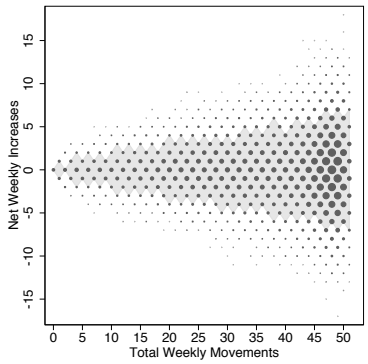
Figure 2: Check-ins and Gender Composition Over Time



(a) Day-of-Week Variation



(b) Week-of-Year Variation



(c) Within Venue Variation

Notes: (b): The 53rd week of the sample is omitted because it only contains a single day. (c): In this scatter plot of venues in our data, larger dots correspond to a greater numbers of venues. A venue experiences a weekly increase (decrease) in gender composition if the proportion of female check-ins rises (falls) by at least one percentage point.

4 Measuring Sorting Within Neighborhoods

A group of venues is more highly sorted by gender if their gender compositions differ greatly from one another. Fully unsorted venues all feature the same proportion of female customers and exhibit maximal gender diversity. The sorting of men and women across venues leads the gender compositions of the venues to differ from one another. In the extreme case, if some venues serve only females and the others serve only males, then the venues are fully sorted and exhibit no gender diversity. One important measure of sorting is the Theil (1967) index, which has been widely used in studies of segregation (e.g., Reardon and Firebaugh (2002); Chetty et al. (2014)).[8] Formally, if s_{jk}

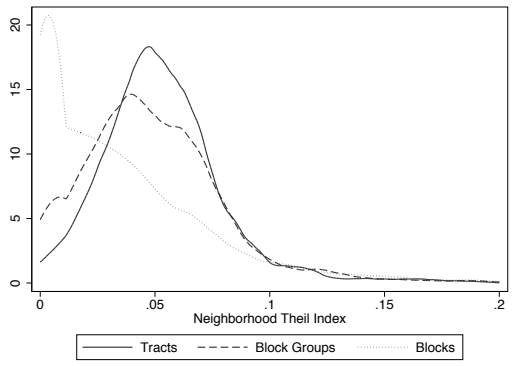
[8]Weitzman (1992) proposes a general, recursively defined measure of diversity that satisfies numerous attractive mathematical, economic and conceptual properties. In certain contexts, he shows it to be equivalent to the widely

is the share of females at venue j located in neighborhood k, then the Theil index of neighborhood k is given by

$$T_k = \frac{1}{n_k} \sum_{j \in k} \left(\frac{s_{jk}}{\bar{s}_k} \cdot \log \frac{s_{jk}}{\bar{s}_k} \right) \tag{4}$$

where n_k is the number of venues in the neighborhood and $\bar{s}_k$ is the simple average of s_{jk} across all venues in the neighborhood.[9] If the neighborhood is fully integrated (i.e., no sorting, maximal diversity), then all of its venues will have the same gender composition as the neighborhood overall, and $T_k = 0$. Neighborhoods with less diverse venues have larger values of T_k.[10] In practice, k can correspond to the entirety of a city (c), a census tract (t), a census block group (g) or a census block (b), so T_k represents the extent to which individuals sort within k.

Figure 3: Densities of Theil Indices for Various Neighborhood Definitions



Notes: All densities are estimated using a bandwidth of 0.005 and an Epanechnikov kernel. For clarity, we present the density only for values of the domain less than 0.2; fewer than 1% of neighborhoods of any type have a Theil Index in excess of 0.2. Theil Indices are pooled across neighborhoods in all cities.

We compute the Theil index for all tracts, block groups and blocks in the cities in our sample and present the densities of these indices in Figure 3.[11] The bulk of the density of T_t lies away

used Shannon index (Shannon and Weaver (1963)), which measures the amount of "true diversity" or the effective number of different types of "objects". In our application, objects correspond to venues by demographic composition, and the Shannon index reduces to the Theil index up to an additive constant.

[9] Our results are virtually unchanged if we denote s_{jk} as the share of men in venue j in neighborhood k. The same is true for our age analysis.

[10] The maximum value that the Theil index can take is $\log n_k$, which varies with the density of venues in a neighborhood. Where applicable, our results using the Atkinson (1970) index (the Theil index divided by $\log n_k$, thus normalized to values between 0 and 1) instead of the Theil index are all qualitatively equivalent. As we explain below, we use the Theil index instead of the Atkinson index in our analysis because of its decomposability properties.

[11] We calculated bootstrapped standard errors with 500 repetitions for the means of T_t, T_g and T_b for each city separately. All means are statistically significantly different from zero at the 99% level.

from zero, which reveals that individuals sort within tracts. Similarly, the bulk of the density of T_g lies away from zero, which reveals that individuals also sort within block groups. The density of T_b is close to zero for approximately 10% of the sample, so roughly 90% of blocks in these cities are further sorted by gender in venues. Mathematically, $T_b \leq T_g \leq T_t$ for all $b \in g \in t$. Because these three densities roughly coincide for higher values of the Theil index, all of the sorting in highly sorted tracts and block groups occurs within their constituent blocks as opposed to across them.

The Theil index possesses the attractive property of being additively decomposable.[12] This property allows for the Theil index of an entire city to be decomposed into a weighted average of the Theil indices of all of its constituent neighborhoods, which encapsulates sorting within neighborhoods, plus the extent of sorting across neighborhoods in the city. Formally, we can decompose venue sorting in city c into within- and across- tract components as

$$T_c = \underbrace{\sum_{t \in c} \alpha_t \cdot T_t}_{\text{sorting within tracts}} + \underbrace{\sum_{t \in c} \alpha_t \cdot \log \frac{\bar{s}_t}{\bar{s}_c}}_{\text{sorting across tracts}} \tag{5}$$

where the weights $\alpha_t = \frac{n_t s_t}{n_c s_c}$ correspond to the contribution of each tract to overall venue diversity in c (s_k represents the share of females across all venues in neighborhood k). T_c can be similarly decomposed to the block group or block levels. The key benefit of this simple decomposition is that we can analyze sorting across venues independently of sorting across neighborhoods. In Table 2, we present the proportion of the sorting across venues in cities that happens only across venues within neighborhoods, i.e., the contribution of the second term of equation (5). It is immediate that the majority of gender sorting across venues in cities occurs within neighborhoods.

[12] Although the Theil index is not the only such measure that is additively decomposable, it is the only one that is homogenous of degree zero (Bourguignon (1979)), which makes it invariant to rescaling. This is important in our application because males may be more or less likely to check in on Foursquare than females; hence in order to maintain the external validity of our estimates we should make only relative comparisons of segregation. In addition, as Shorrocks (1980) points out, other commonly used measures of segregation, diversity, exposure or inequality with other attractive properties which are based on the Herfindahl index (e.g., the index of segregation introduced in Ellison and Glaeser (1997)) or the Gini coefficient, are not additively decomposable, so they are less useful and appropriate in our context.

Table 2: Venue Sorting Within Neighborhoods

	Proportion of city-wide sorting due to sorting within:		
	Tracts	Block Groups	Blocks
Atlanta	0.89	0.83	0.59
Chicago	0.82	0.74	0.47
Dallas	0.79	0.71	0.48
Los Angeles	0.83	0.74	0.50
New York City	0.92	0.88	0.78
Philadelphia	0.85	0.78	0.50
San Francisco	0.83	0.78	0.57
Washington, DC	0.88	0.84	0.61

Note: Bootstrapped standard errors for all entries in all cities are less than 0.005 and are omitted for clarity.

To better interpret the measures in Table 2, we can benchmark the observed gender compositions of venues against the gender compositions of venues that would be hypothetically observed if there was no sorting within neighborhoods.[13] This exercise reveals how much additional sorting we are able to measure because we can observe sorting across venues within neighborhoods as opposed to only sorting across neighborhoods (as in the vast majority of studies). By observing sorting at the more disaggregated venue level, we are able to detect 2-4 times more sorting than in data aggregated at the block level, and 4-12 times more sorting than in data aggregated at the tract level.[14] For Manhattan, these numbers are on the higher end: we are able to detect 4 (12) times more sorting than we would have with data aggregated at the block (tract) level.

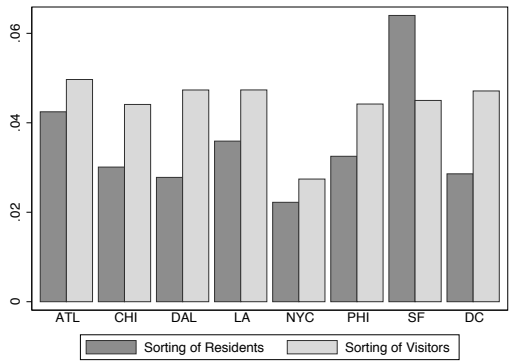
Neighborhood Residents vs. Visitors

Typically researchers can observe only certain broad sorting choices that individuals make, such as their neighborhoods of residence. Because we also observe the choices of which neighborhoods people visit, it is useful to compare the relative amounts of homophily observed in residents and visitors of the same neighborhood.

[13] We also conduct a falsificiation exercise in which individuals are not allowed to sort within blocks to validate this benchmarking exercise and ensure that our results are not simply artifacts of sampling error. The details and results of this exercise are provided in the appendix.

[14] To obtain these figures, we take the reciprocal of the proportion of observed venue sorting due to sorting within neighborhoods (e.g., $(1 - 0.89)^{-1} = 9.09$ for Tracts in Atlanta).

Figure 4: Sorting of Residents vs. Sorting of Visitors



Note: "Sorting of Residents" is calculated as the Theil index of the gender composition of block residents from the 2010 Census. For comparability, "Sorting of Visitors" is calculated as the Theil index of the gender composition of check-ins in blocks. Bootstrapped standard errors for all estimates are below 0.005 and are omitted for clarity.

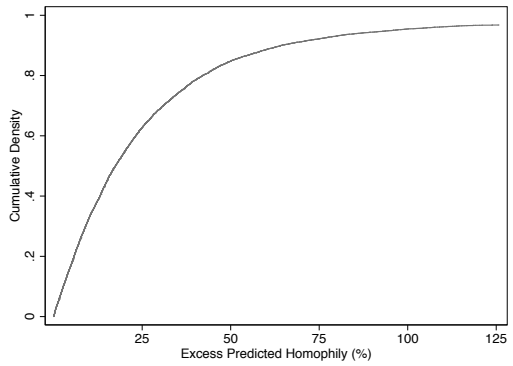
In Figure 4, we compare the sorting of residents across blocks with the sorting of visitors across blocks for each city in our sample. The former is calculated as the Theil index of the gender composition of block residents for each city from the 2010 Census. The latter is calculated as the Theil index of the gender composition of block visitors for each city from our data, which is equivalent to the second term in a decomposition of T_c into sorting within- and across- blocks according to equation (5). We find that for all cities except one, there is less residential sorting across blocks than visitor sorting across blocks.[15] This suggests that if anything, studies that rely on residential data understate the amount of sorting and lack of exposure to diversity over and above the numbers discussed previously.

We summarize the findings in this section in the explicit language of exposure to similar peers (homophily). For instance, if we were to observe residential data, then we would estimate that the average woman in neighborhood k would randomly encounter another woman with probability equal to the female share of residents in that neighborhood. However, if we were to observe data disaggregated to the venue level, then we could better estimate that the average woman in k would randomly encounter another woman with probability $\sum_{j \in k} \frac{f_{jk}}{f_k} \cdot s_{jk}$ where f_{jk} is the number of women visiting venue j, and f_k is the total number of women visiting neighborhood k. It is straightforward then

[15] One city in our sample, San Francisco, exhibits greater residential sorting by gender than visitor sorting by gender. We speculate that this is due to San Francisco's sizable gay population, which concentrates residentially in certain neighborhoods whose venues are visited by a very gender diverse population. Indeed, San Francisco, like all other cities in the sample, exhibits less residential sorting by age than visitor sorting by age (see appendix).

to calculate how much more likely we would estimate that a woman would encounter a woman in venue level data than in residential data, and we present the empirical cumulative distribution of this difference (which corresponds to the extent to which residential data fails to capture homophily that is observable in venue data) in Figure 5. To be conservative, we define neighborhoods as census blocks, which are the smallest geographic units that are commonly used for residential data by researchers. In roughly 40 percent of the neighborhoods, we would underestimate exposure to peers of the same gender by at least 25 percent (or about one standard deviation of the probability that a woman would encounter another woman in a venue) if we used residential data instead of venue data, and in roughly 20 percent of neighborhoods, we would underestimate exposure to peers of the same gender by at least 50 percent.

Figure 5: Excess Predicted Homophily in Venue Data



Note: In this figure, we present the empirical cumulative distribution of how much more likely we would predict that a woman would encounter another woman in a census block using venue level data than if we used residential data.

5 Diversity and Venue Variety

5.1 A Simple Model of Sorting Across Venues

Thus far, we have established that the amount of sorting within neighborhoods is substantial, and this sorting creates a wedge between the levels of gender diversity that are observed in venues and in neighborhoods. In order to explore the causes of this wedge, it is useful to consider a model of sorting across venues. Our goal is not to fully characterize sorting within and across neighborhoods. Instead, we seek only to explore the role of venue offerings in explaining diversity in neighborhoods and in venues. To that end, we present a simple and stylized model based on Hotelling (1929).

On the supply side, we model a neighborhood k as a collection of venues indexed by j, each of

which possess a single particular characteristic x_j that lies on the unit interval and differentiates the venues. This characteristic can be thought of as the venue's particular type of offering. The spatial distribution of venues on the unit interval corresponds to the variety of venues in the neighborhood. For example, a Mexican restaurant and a Chinese restaurant lie closer to each other on the interval than, say, a Mexican restaurant and a shoe store. More generally, when venues are more spread out, they collectively represent a greater variety of venue offerings, which we denote as V_k. To simplify our analysis as much as possible, we consider the simplest setting that could feature sorting across venues: a single neighborhood with two venues where supply is fixed. In such a neighborhood, $V_k = |x_1 - x_2|$.[16]

On the demand side, we assume that there is a mass of individuals, each of whom are indexed by i and possess the following utility function over venues

$$U_i(x) = u - (\delta_i - x)^2 \tag{6}$$

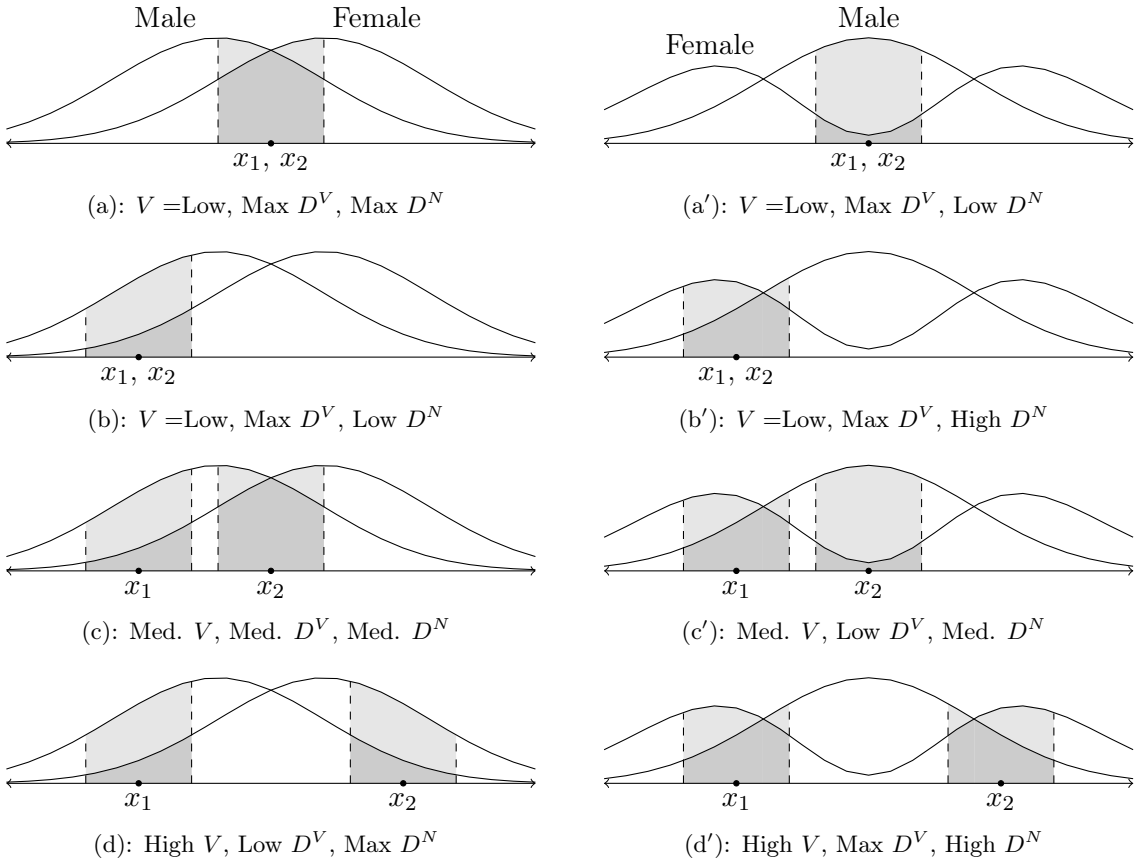
where δ_i is their ideal point and $u < 1$ is a constant. Once again, we consider the simplest specification of demand that could feature sorting across venues: individuals belong to one of two equal sized groups of potential venue-goers, say, males and females. The δ_i are drawn from different distributions depending on the group to which i belongs. Each individual is assumed to choose at most one venue that maximizes their utility provided $U_i > 0$.[17] If more than one venue offers an individual maximal positive utility, then ties are broken randomly.

We can combine the supply and demand sides to define equilibrium venue diversity and neighborhood diversity. Venue diversity in a neighborhood is measured by the negative Theil Index of the gender composition of venues, i.e., $D_k^V = -T_k$, since higher levels of T_k correspond to less diversity. The overall amount of diversity in neighborhood k can be understood as how representative the gender composition of *actual* venue-goers in the neighborhood is of the gender composition of *potential* venue-goers. Because the groups are of equal size, the latter is equal to $\frac{1}{2}$, so we can define neighborhood diversity as $D_k^N = -\left|\frac{f_1+f_2}{f_1+f_2+m_1+m_2} - \frac{1}{2}\right|$ where f_j and m_j are the number of female

[16] In general, x_j could also refer in part to the physical locations of venues. In such a formulation, connected subsets of the unit interval could correspond to physical neighborhoods, and we could study sorting across neighborhoods as well. For simplicity, we abstract away from this formulation because our empirical analysis exploits only very local variation in venue variety.

[17] Individuals for whom $U_i \leq 0$ for all x can be understood as choosing an outside option, which could reflect visiting another neighborhood.

Figure 6: Venue Variety and Diversity



(a): V =Low, Max D^V, Max D^N

(a′): V =Low, Max D^V, Low D^N

(b): V =Low, Max D^V, Low D^N

(b′): V =Low, Max D^V, High D^N

(c): Med. V, Med. D^V, Med. D^N

(c′): Med. V, Low D^V, Med. D^N

(d): High V, Low D^V, Max D^N

(d′): High V, Max D^V, High D^N

and male visitors to venue j respectively.

We use this simple model to illustrate the relationship between neighborhood venue offerings and diversity in a series of diagrams. In Figure 6, we consider four different neighborhoods in order of increasing venue variety (a) - (d). Each of these neighborhoods has a counterpart (a′) - (d′) that is identical except for the demand that it faces. In each of the neighborhoods in the first column, men's ideal points tend to be lower than women's ideal points. In neighborhood (a), there is no venue variety, as $x_1 = x_2$. As a result, there is no sorting, so the venues exhibit maximal diversity. Also, since the venues attract equal numbers of men and women, there is maximal neighborhood diversity as well. In neighborhood (b), $x_1 = x_2$ as before, so there is still no venue variety or sorting, and hence maximal venue diversity. However, the overall level of diversity of this neighborhood is low since venue-goers are unrepresentative of the population at large. In neighborhood (c), $x_1 \neq x_2$, so this neighborhood has a moderate level of venue variety, which is accompanied by a moderate

amount of sorting (and hence some venue diversity). As a result, this neighborhood has a moderate overall level of diversity relative to neighborhoods (a) and (b). Finally, neighborhood (d) features a high level of venue variety, which is accompanied by a high level of sorting, and hence low venue diversity. However, because the two venues cater to symmetric groups of consumers, an equal number of men and women go to one of the venues, and hence the neighborhood has maximal overall diversity. The four analogous neighborhoods in the second column, (a′) - (d′) face different demands. Men's and women's average ideal points are now both located at $\frac{1}{2}$, however women tend to prefer venues with low and high x_j's, whereas men tend to prefer venues with moderate x_j's.

We can draw three important conclusions from this stylized analysis. First, sorting is made possible only by venue variety; it is trivial to note that there will be no sorting across venues with identical x_j's (and minimal sorting across venues with very similar x_j's). Accordingly, venue variety is an attractive candidate for a determinant of the wedge between venue diversity and neighborhood diversity that we have established in Section 4. Second, the relationship between venue variety and venue diversity is theoretically ambiguous. In the neighborhoods $a - d$, venue variety and venue diversity are inversely related to each other, but in neighborhoods $a' - d'$, this relationship no longer holds. Third, the relationship between venue variety and overall neighborhood diversity is also theoretically ambiguous. In neighborhoods $a - d$, venue variety and neighborhood diversity are directly related to each other, but in neighborhoods $a' - d'$, this relationship no longer holds. The latter two implications suggest that we must empirically determine the relationships between venue variety and venue and neighborhood diversity in order to determine the extent to which venue variety creates this wedge.

5.2 A Proxy for Venue Variety

In order to generalize the model and take it to the data, we must first measure venue variety. Intuitively, venue variety should be lower in neighborhoods with more substitutable venues whose characteristics and offerings are more similar. One important characteristic of a venue is its location. All else constant, venues located far from each other should be less substitutable. In addition, the offerings of a venue can be proxied for by its categorization in our data.

Because the subcategories of venues are so narrowly defined, we can interpret them as proxies

for x_j provided that we compare venues only in narrow geographic areas (i.e., the same location).[18] Thus, we can recast the first conclusion drawn above in terms of something that is measurable with our data: The proportion of sorting within a neighborhood that is due to sorting across subcategories should be high if neighborhoods are narrowly defined. In Table 3 we present the proportion of sorting within neighborhood that occurs across venue types for each neighborhood definition. Our findings are consistent with the model. The bulk of sorting within neighborhoods occurs across subcategories; for instance, about 90% of sorting within census blocks occurs across subcategories. However, much less sorting within entire cities occurs across subcategories. This suggests that location is a better proxy for x_j when comparing venues that are located far from each other.

Table 3: Proportion of Within-Neighborhood Sorting by Gender Due to Sorting Across Subcategories:

	City	Tracts	Block Groups	Blocks
Atlanta	0.26	0.78	0.83	0.91
Chicago	0.26	0.84	0.89	0.94
Dallas	0.27	0.82	0.86	0.92
Los Angeles	0.20	0.83	0.87	0.92
New York City	0.31	0.70	0.81	0.90
Philadelphia	0.22	0.81	0.85	0.94
San Francisco	0.28	0.76	0.82	0.92
Washington, DC	0.26	0.75	0.82	0.91

Note: Subcategories (225) are defined in the appendix. Bootstrapped standard errors for all entries are less than 0.005 and are omitted for clarity.

[18] Venue amenities can be either exogenous or endogenous. In general, endogenous amenities are those that are a function of individuals' choices, such as the demographic makeup of a venue's clientele. As discussed by Schelling (1969), Caetano and Maheshri (2014) and others, if endogenous amenities were important determinants of sorting, then they would generically imply a dynamic process of tipping whereby venues would be systematically increasing or decreasing in their female share over time. However, Figure 2 shows no evidence of such tipping. We find similar results for age (see appendix). Taken together, these results suggest that these endogenous amenities are relatively unimportant to sorting in our context.

5.3 Identifying Causal Effects of Venue Variety on Venue and Neighborhood Diversity

The stylized model described above reaches ambiguous conclusions about the effects of venue variety on venue and neighborhood diversity. As a result, we attempt to identify these causal effects empirically. Consider two small, nearby neighborhoods that are otherwise similar except for their levels of venue variety. For instance, one neighborhood may feature only restaurants, whereas another neighborhood may feature both restaurants and shops (compare to neighborhoods (a) and (c) in Figure 6). Given their small sizes and proximity, it is reasonable to consider their locations and the demands that they face to be approximately the same. It follows that the observed differences in the amounts of diversity in venues within each neighborhood and the amounts of overall diversity in each neighborhood can be reasonably attributed to the difference in their venue variety. We implement an identification strategy that makes this comparison.

Following the model, the amount of local diversity in venues in block b can be measured by the negative Theil Index, $D_b^V = -T_b$, and the overall amount of neighborhood diversity in b can be measured by how representative the distribution of visitors in b are of the distribution of visitors in the whole city. As a generalization of the model, if f_{jb} and m_{jb} represent the total number of females and males in venue j in block b, and $s_b = \frac{\sum_{j \in b} f_{jb}}{\sum_{j \in b}(f_{jb}+m_{jb})}$, then we can define

$$D_b^N = -\left|s_b - s_c\right| \tag{7}$$

to be the overall amount of diversity in block b in city c. Finally, because we compare only small neighborhoods that are close to each other, we can take advantage of the classification of venues in our data to generalize the model above and define the venue variety of b, V_b, as either the number of unique categories or subcategories of venues that are on offer in that block.

We estimate the following regression equation:

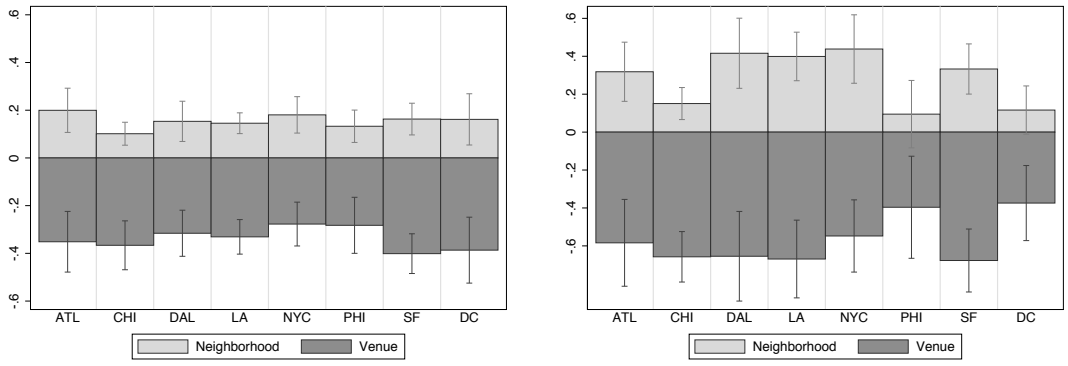
$$D_b^V = \beta^V V_b + \alpha_g^V + X_b\gamma^V + R_b\lambda^V + \epsilon_b^V \tag{8}$$

$$D_b^N = \beta^N V_b + \alpha_g^N + X_b\gamma^N + R_b\lambda^N + \epsilon_b^N \tag{9}$$

in which the α_g are fixed effects at the block group level, where $b \in g$, and X_b represents a set of block control variables that includes the total number of venues and amount of checkin activity in b, R_b represents a set of residential control variables that includes the total number and female share of residents in b, and ϵ_b^V represents an error term.[19] β^V and β^N are the coefficients of interest. To aid interpretation, we normalize all variables by their standard deviations, so β^V corresponds to the effect of a one standard deviation increase in venue variety on venue diversity (in units of its standard deviation), and β^N corresponds to the effect of a one standard deviation increase in venue variety on neighborhood diversity (in units of its standard deviation).

In Figure 7, we present estimates of β^V (darker bars) and β^N (lighter bars) along with their corresponding 95% confidence interval for each city separately, and for V_b defined as either the number of unique categories or subcategories. We systematically find that $\hat{\beta}^V < 0$ and $\hat{\beta}^N > 0$. This implies that any increase in neighborhood diversity due to a change in venue variety will be accompanied by more intense sorting between venues within the neighborhood, thereby reducing the exposure to diversity at the venue level. Indeed, a one standard deviation increase in venue variety will lead to roughly a 0.2 standard deviation increase in neighborhood diversity, and roughly a 0.4 standard deviation decrease in venue diversity.

Figure 7: $\hat{\beta}^V$ and $\hat{\beta}^N$ By City



(a) V_b = Num. of Categories (b) V_b = Num. of Subcategories

Notes: The dark bars represent estimates of $\hat{\beta}^V$ from equation (8), and the light bars represent estimates of $\hat{\beta}^N$ from equation (9). 95% confidence intervals are also shown from robust standard errors clustered at the block group level. The number of observations for each of the 16 regressions is equal to the number of census blocks in each city (see Table 1), and the R^2 of each regression varies from 0.33 to 0.50.

[19]The residential control variables are obtained from the 2010 Census Summary File 1 (SF1).

Can We Interpret $\hat{\beta}^V$ and $\hat{\beta}^N$ as Causal?

The causal parameters β^V and β^N are identified under the assumptions $\text{Cov}\left(\epsilon_b^V, V_b | \alpha_g^V, X_b, R_b\right) = 0$ and $\text{Cov}\left(\epsilon_b^N, V_b | \alpha_g^N, X_b, R_b\right) = 0$. Because we conduct our analysis at the block level, we explicitly consider small neighborhoods, and the inclusion of block group fixed effects α_g ensures that we only compare neighborhoods that are located near each other, which holds constant all determinants of the demand that vary at the block group level. Still, certain neighborhood amenities that are correlated to venue variety might attract different groups of people to different nearby blocks, so we control for X_b to ensure that we compare blocks that have similar numbers of venues and levels of foot traffic, and we control for R_b to ensure the number and type of residents is similar across these blocks.

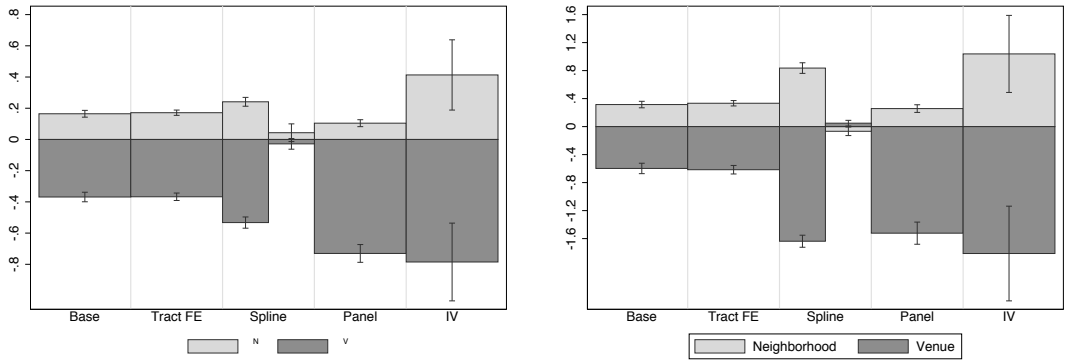
The remaining concern is that some unobserved neighborhood amenities that cannot be controlled for by even these variables may be correlated to venue variety. For instance, one might worry about simultaneity bias: different venues may decide to locate themselves in neighborhoods that attract diverse visitors, i.e. demand for venues causes supply of venues. The fact that neighborhoods are both small and close to each other in our context helps allay such concerns as this would only be problematic if venues had control over and preference for a particular nearby block to enter. Since entering a particular block requires a commercial vacancy and the blocks are similar in venue density, foot traffic, and location, it is difficult to think of a reason why this might be the case.[20]

Nevertheless, we provide four robustness checks that address these and other concerns. The results of these four robustness checks are shown in Figure 8, where we compare the baseline estimates of β^V and β^N from equations (8) and (9) pooled over all eight cities with estimates from four alternative specifications.[21] In the first set of bars, we define venue variety as the number of distinct categories in a neighborhood, and in the second set of bars, we define venue variety as the number of distinct subcategories in a neighborhood.

[20] We perform an F-test for whether all coefficients of R_b are equal to zero for each city and find that the differences in the numbers and gender compositions of nearby block residents do not help predict differences in either D_b^V or D_b^N in any city. This suggests that residents of the city do not consider nearby blocks to be in meaningfully different locations (since, at a minimum, the residents of those blocks do not).

[21] We also conducted these robustness checks for each city and obtained similar results, which are reported in the appendix. We were unable to obtain IV estimates disaggregated by city due to their lack of precision.

Figure 8: $\hat{\beta}^V$ and $\hat{\beta}^N$: Alternative Identification Strategies



(a) V_b = Num. of Categories (b) V_b = Num. of Subcategories

Notes: The dark shaded bars represent $\hat{\beta}^V$, and the light shaded bars represent $\hat{\beta}^N$. 95% confidence intervals are also shown from robust standard errors clustered at the block group level. The first bars correspond to baseline estimates from equations (8) and (9). The second bars replace the block group fixed effects in the baseline estimates with tract fixed effects. The third set of bars correspond to estimates of the parameters specified as a linear b-spline with a knot at 3 subcategories. The fourth bars correspond to estimates from equations (10) and (11) where the dataset is disaggregated to a monthly panel, and the block group fixed effects are replaced with block fixed effects. The fifth bars correspond to 2SLS estimates of the baseline regressions with zoning instruments.

For our first robustness check, we reestimate equations (8) and (9) with tract fixed effects instead of block group fixed effects. Tracts typically encompass two or more block groups, so these fixed effects no longer control for amenities varying across block groups within tracts which may confound our estimates. The results (denoted as "Tract FE") are virtually unchanged, which suggests that after controlling for X_b and R_b, amenities and local demand varying across block groups within tract are uncorrelated to V_b. It is difficult to conceive of unobservables that are correlated to V_b, that vary across blocks within block groups but do not vary across block groups within tracts.[22]

Second, we reestimate equations (8) and (9) using linear b-splines in V_b, which allows us to estimate separate marginal effects of venue variety on diversity for neighborhoods with three or fewer subcategories and for neighborhoods with four or more subcategories. If $\hat{\beta}^V$ and $\hat{\beta}^N$ are causal estimates, then they will likely decline in magnitude as we compare nearby blocks with higher levels of venue variety.[23] In contrast, if these estimates reflect confounding factors that are

[22]For instance, simultaneity could only be a concern if venues had more control or preference over their choice of which block within a block group to enter relative to their choice of which block group within a tract to enter.

[23]Extending the intuition of the model presented above, in a neighborhood with greater number of venues with

present irrespective of the level of V_b, then we should find that these effects do not decline for higher V_b. Indeed, we find that nearly all of these effects (denoted as "Spline") operate at low levels of venue variety in all cities in our sample.[24]

Third, we exploit the longitudinal variation in our data to estimate β^V and β^N using an alternative identification strategy. We can respecify equations (8) and (9) as

$$D_{bt}^V = \beta^V V_{bt} + \alpha_b^V + \alpha_{ct}^V + X_{bt}\gamma^V + \epsilon_{bt}^V \quad (10)$$

$$D_{bt}^N = \beta^N V_{bt} + \alpha_b^N + \alpha_{ct}^N + X_{bt}\gamma^N + \epsilon_{bt}^N \quad (11)$$

respectively. The key difference is that all of our main explanatory variables and controls now vary by month. By doing so, we can identify β^V and β^N using only variation in neighborhood venue variety that arises due to the entry and exit of venues over time. We implement this identification strategy by including block fixed effects (α_b^V and α_b^N) that additionally control for all unobserved determinants of diversity that vary across blocks within block groups that were not controlled for in equations (8) and (9). The fixed effects α_{ct}^V and α_{ct}^N control for city level amenities that may vary by month in order to absorb any seasonality that varies across cities. Our results (denoted as "Panel") suggest that our baseline estimates of β^V are conservative, which is consistent with our sensitivity analysis in Section 6.

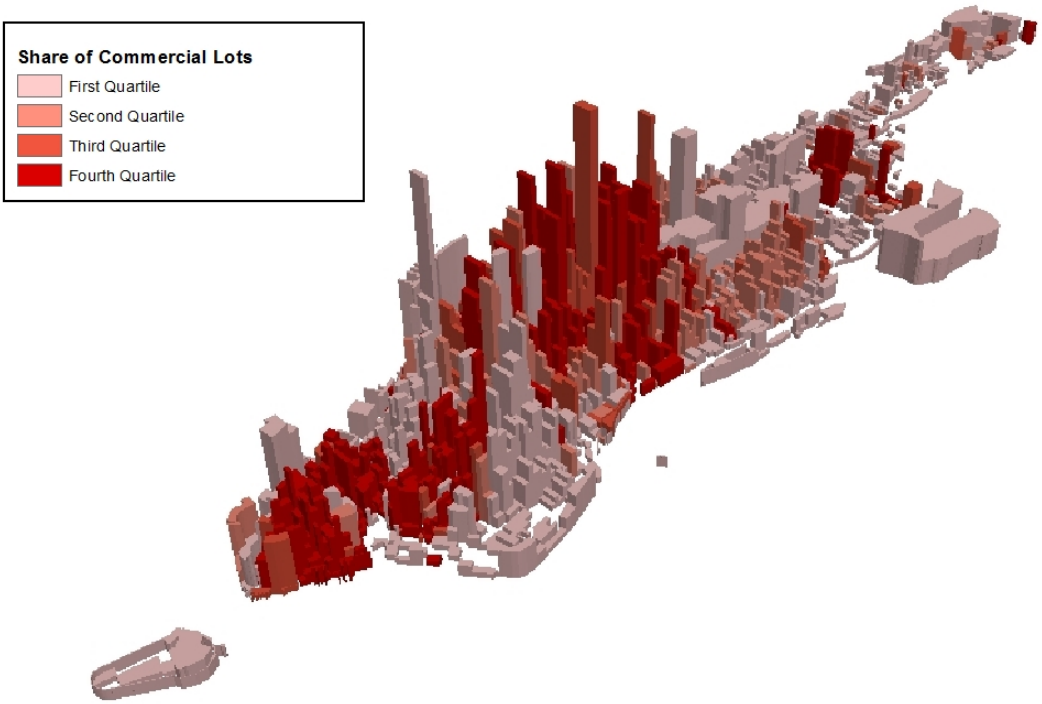
Finally, we estimate β^V and β^N using a third, distinct identification strategy that uses zoning laws as instrumental variables (IVs) for venue variety. By doing so, we only use identifying variation in the variety of venues that stems from the supply side of venues. This IV approach deals with remaining simultaneity concerns and all remaining confounders that are uncorrelated to zoning laws such as certain kinds of measurement error. Specifically, we use the share of lots in the block that are zoned to residential, commercial and mixed uses as instruments; hence, we effectively compare diversity in adjacent blocks that are zoned differently and have different levels of venue variety but a similar number of venues, overall traffic and numbers and types of residents.[25]

distinct x_j's, more of the support will be covered by venue-goers. As a result, a marginal increase in venue variety will have a smaller effect on both D_b^V and D_b^N since the additional venue will draw increasingly from individuals who were otherwise planning to go to another venue.

[24] These qualitative results do not change when we place the knot at $2, ..., 5$ subcategories.

[25] We obtained lot level data on zoning for each city from their respective planning offices. Lots can be zoned for

Figure 9: Commercial Zoning and Venue Variety (First Stage)



Notes: Each bar represents a census block in Manhattan. The height of each bar corresponds to V_b, the number of unique venue subcategories in b. Darker bars represent blocks with a greater proportion of commercially zoned lots.

Differences in zoning laws are found to generate differences in the variety of venues in census blocks. In Figure 9, we spatially illustrate the relationship between commercial zoning (categorized in quartiles for visual clarity) and venue variety (number of unique venue subcategories) in Manhattan census blocks, which is clearly positive. More formally, a joint F-test of the significance of the three instruments for the number of unique venue categories and unique venue subcategories yields $F_{3,5779} = 25.00\ (0.00)$ and $F_{3,5779} = 12.27\ (0.00)$, respectively, where the p-values shown in parenthesis are much smaller than 0.01.

Our estimates (denoted as "IV") are, if anything, larger in magnitude than all OLS estimates, which suggests that the OLS estimates may be attenuated by measurement error. As a result, our

other uses than the three that we use for IVs such as manufacturing and parks.

findings that $\beta^V < 0$ and $\beta^N > 0$ should be understood to be conservative. This interpretation is consistent with our treatment of measurement error described in Section 6 and the appendix.

Taken altogether, we conclude that the observed wedge between venue diversity and overall neighborhood diversity is largely attributable to sorting induced by the variety of venues on offer in neighborhoods. Greater venue variety leads to higher overall diversity in the neighborhood but lower diversity in venues within the neighborhood.

6 Robustness

Although the advent of user-generated datasets offers much promise, there are reasonable potential concerns regarding measurement error. Our primary concern is that the proportion of females that we observe in a venue may be systematically different from the proportion of females that actually visit the venue. We argue that this likely does not confound our analysis, and, in any case, we show empirically that our results are qualitatively robust to the extent that it does. Indeed, in the presence of such measurement error, our results should actually be understood as conservative estimates of the amount of sorting within neighborhoods and the effects of this sorting on neighborhood and venue diversity.

To fix ideas, let $\tilde{f}_{jk}$ and $\tilde{m}_{jk}$ represent the actual numbers of females and males who visit venue j in neighborhood k. We can write the relationships between the observed and actual variables as

$$f_{jk} = \gamma^f_{jk} \cdot \tilde{f}_{jk} \tag{12}$$

$$m_{jk} = \gamma^m_{jk} \cdot \tilde{m}_{jk} \tag{13}$$

where the γ_{jk} parameters represent gender and venue specific check-in rates. All observed variables previously defined in terms of f_{jk} and m_{jk} have an actual, unobserved counterpart denoted with a tilde.

When mismeasurement is not gender specific, i.e., $\gamma^f_{jk} = \gamma^m_{jk}$, the female shares of check-ins at venues are unchanged, so all of our results are unaffected. This is a particularly nice feature, as it ensures our results are robust to any basic form of measurement error due to the fact that not all venue customers use the Foursquare app. Moreover, if mismeasurement is gender specific, but the mismeasurement in the female share of venues is only neighborhood specific (i.e., $s_{jk} = \gamma^s_k \cdot \tilde{s}_{jk}$), then

our estimates of neighborhood Theil indices and their geographic decompositions are unchanged. This ensures that our results are robust to neighborhood specific sources of measurement error such as those correlated to unobserved neighborhood amenities.

In general, measurement error may be not only gender and neighborhood specific but also venue specific. We check the sensitivity of our main results to a general form of measurement error by conducting a Monte Carlo simulation. Without loss of generality, we define $\omega_{jk} = \frac{\gamma_{jk}^m}{\gamma_{jk}^f}$ to be the relative oversampling of males in venue j. For each iteration l, we randomly draw ω_{kj}^l for each venue from a uniform distribution $[\underline{\omega}, \overline{\omega}]$. We then calculate the "true" values of $\tilde{s}_{jk}^l$, $\tilde{T}_k^l$ for that iteration. Using these "true" values, we can simulate the main results of the paper, and the variation of the results across iterations allows us to construct confidence intervals. Although ω_{jk}^l is randomly drawn, it is positively correlated to $\tilde{s}_{jk}^l$ by construction.[26]

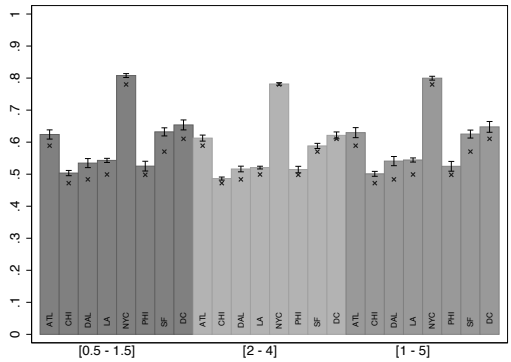
We conduct the Monte Carlo simulation under three separate parametrizations to capture qualitatively different types of measurement errors. In the first parametrization, we set $\underline{\omega} = 0.5, \overline{\omega} = 1.5$, which allows males to check in up to 50% less or more frequently than females, though they check in at the same rate on average. In the second parametrization, we set $\underline{\omega} = 2, \overline{\omega} = 4$. This increases the measurement error in two ways: it assumes that on average males check in three times more than females do, and it allows for greater dispersion of γ_{jk} across venues. In the third parametrization, we set $\underline{\omega} = 1, \overline{\omega} = 5$ which further worsens measurement error by allowing for even greater dispersion of γ_{jk} across venues.[27]

We report the Monte Carlo ($N = 500$) results for each of the three main estimates of the paper in Figure 10. Each panel contains 24 bars, which represent the three different sets of parameters for each city in our sample. For each set of parameters, the bars represent the average estimate of that result across all 500 iterations. We also show 95% confidence intervals for these estimates along with the previously presented value of that result under the assumption of no measurement error denoted with an "x".

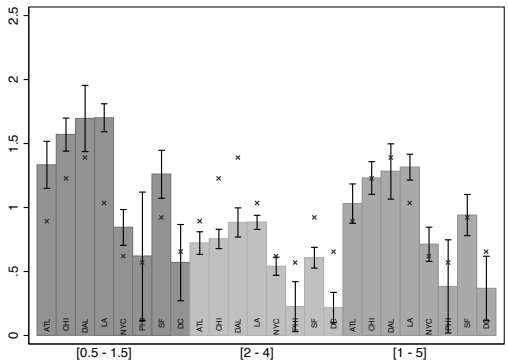
[26] We also performed alternative Monte Carlo simulations where we allowed ω_{jk}^l to be positively (or negatively) correlated to s_{jk} instead and obtained qualitatively similar results.

[27] We also performed analogous Monte Carlo simulations assuming females check in more rather than less frequently than males on average and found analogous results.

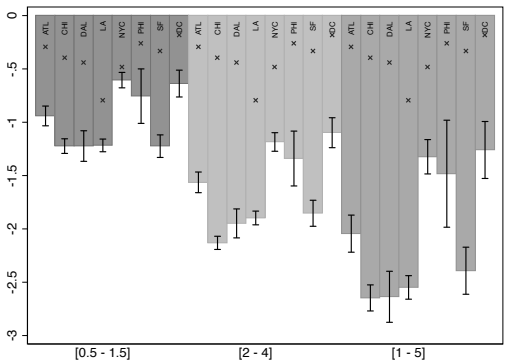
Figure 10: Robustness: Monte Carlo Results



(a) Proportion of City Sorting due to Within Census Blocks Sorting



(b) $\hat{\beta}^N$



(c) $\hat{\beta}^V$

Notes: Each panel presents Monte Carlo results for three different set of parameters $[\underline{\omega}, \overline{\omega}]$, which represent the interval of the uniform distribution from which γ_{jk} is drawn: $[0.5, 1.5]$, $[2, 4]$ and $[1, 5]$. The bars represent the estimates of the Monte Carlo with 95% confidence intervals, and "x" represents the estimates under the assumption of no measurement error, which are reported in the paper.

In the first panel of Figure 10, it is clear that our estimate of the fraction of the city sorting that occurs within census blocks is robust to various amounts of measurement error; if anything we underestimate the amount of sorting that occurs locally.[28] Even though the actual estimates under the assumption of no measurement error may fall outside of the confidence interval, they are qualitatively the same. A large proportion of sorting happens within blocks under all reasonable assumptions on measurement error. In second and third panels, we show how our regression results

[28] Because the measurment error that we introduce in the Monte Carlo simulation is correlated to the female share of venues, the across-neighborhood component of city sorting tends to be magnified more than the within-neighborhood component (see equation (5)).

are affected by different kinds of measurement errors. If anything, measurement error leads to attenuation bias, mainly in $\hat{\beta}^V$. This is consistent with the results of our panel and IV identification strategies and suggests that our conclusion that $\beta^V < 0$ and $\beta^N > 0$ may be conservative. Overall, these simulations suggest that our results are generally robust to measurement error. Even though erroneously assuming away measurement error might lead us to estimate parameters that would fall outside of the true confidence intervals in some cases, our qualitative conclusions should not be affected even by very extreme forms of measurement error.

There are other specific concerns that we address in depth in the appendix. For example, one might worry that relying on user-generated data could result in a selected sample of venues that might bias our results. One might also worry that our results are artifacts of sampling error. We find persuasive evidence our main results are qualitatively robust to all reasonable forms of measurement error.

7 Sorting Within Neighborhoods by Age

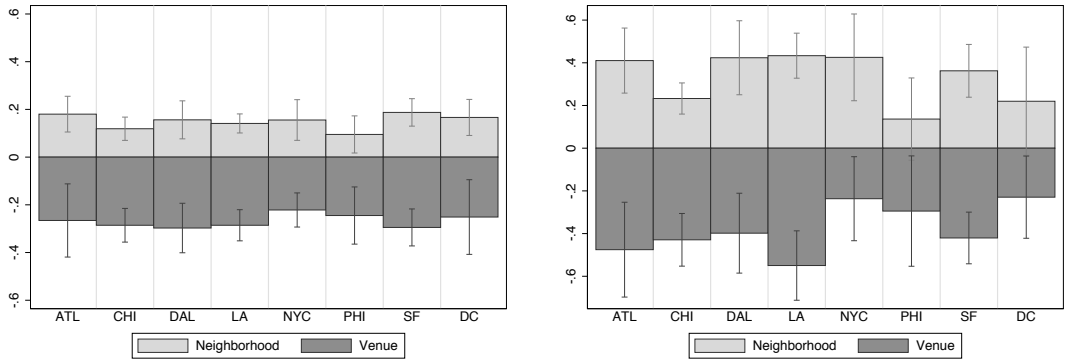
Foursquare collects demographic data of its users along one additional dimension: age. For each venue in our sample, we can observe the daily numbers of check-ins from users under 35 years of age and from users 35 years of age or older. With this information, we replicate our entire analysis, substituting for the proportion of females the proportion of youth. Our results are broadly similar to our results on gender sorting. Although we find roughly half as much sorting by age as we do sorting by gender, it occurs highly locally as shown in Table 4: about half of all age sorting in cities occurs within census blocks. As shown in Figure 11, the effects of venue variety on venue age diversity and neighborhood age diversity are both qualitatively and quantitatively similar to the respective effects on gender diversity in all cities. A full reporting of all results from this replication, including an analogous sensitivity analysis, is provided in the appendix.

Table 4: Venue Sorting Within Neighborhoods By Age

	Proportion of city-wide sorting due to sorting within:		
	Tracts	Block Groups	Blocks
Atlanta	0.75	0.68	0.45
Chicago	0.73	0.63	0.36
Dallas	0.76	0.68	0.45
Los Angeles	0.75	0.67	0.43
New York City	0.87	0.81	0.70
Philadelphia	0.68	0.61	0.34
San Francisco	0.83	0.76	0.53
Washington, DC	0.80	0.74	0.47

Note: Bootstrapped standard errors for all Theil indices in all cities are less than 0.005 and are omitted for clarity.

Figure 11: $\hat{\beta}^V$ and $\hat{\beta}^N$ For Age By City



(a) V_b = Num. of Categories (b) V_b = Num. of Subcategories

Notes: The dark bars represent estimates of $\hat{\beta}^V$ from equation (8), and the light bars represent estimates of $\hat{\beta}^N$ from equation (9). 95% confidence intervals are also shown from robust standard errors clustered at the block group level. The number of observations for each of the 16 regressions is equal to the number of Census Blocks in each city (See Table (1)), and the R^2 of each regression varies from 0.23 to 0.52.

We believe these complementary results on age sorting are informative for three reasons. First, they reveal that homophily is pervasive at all location choices along multiple uncorrelated demographic dimensions. Although we can only conclude with confidence that individuals sort into venues according to these patterns by gender and by age, our results are at a minimum suggestive that individuals might also sort similarly along other demographic dimensions. Second, they

establish that sorting by age is a widespread phenomenon, which is meaningful in its own right. Although peer effects with respect to age have not been widely studied, the systematically different beliefs and prior experiences that people of different ages may have suggest that homophily along this dimension might play a key role in the shaping of political preferences and the development of human capital by social interactions. Third, they serve as an additional indirect robustness check of our main results on gender sorting. Because the relative oversampling of youth is more acute than the relative oversampling of men, the fact that these our results are so robust and similar to our gender results offers additional evidence that our empirical approach does not suffer from measurement error.

8 Conclusion

The endogenous sorting of individuals into peer groups shapes the social environment in which we live. Homophily leads similar people to seek less diversity, which tends to mitigate the effects of well-meaning interventions to expose people to diversity when they are implemented at coarser levels than peer groups form. Using novel, user-generated data from Foursquare, a popular mobile app, we analyze how individuals sort into neighborhoods and further into venues in eight major US cities. We find that individuals sort by gender and by age across venues that are extremely close to each other and at a similar intensity in a variety of different city types, from the long established, dense, urban cores of New York City and Philadelphia to newer and more diffuse urban areas such as Los Angeles, Dallas and Atlanta. This lends some universality to the widespread, homophilic, endogenous peer group formation we observe.

Our findings echo the central themes of Jacobs (1961): individuals endogenously respond to the urban landscape around them, and it is the diversity of this landscape that gives rise to social interactions. However, they also invite a reassessment of whether mixed-use development in neighborhoods coupled with demographic density, which Jacobs and others have championed, are necessary – or even sufficient – ingredients for diversity to emerge. While we find that the resulting variety in the types of venues will lead to more overall diversity in neighborhoods, we also find that it will lead to *less* diversity at the venue level as similar individuals are able to more intensely sort into the same venues. Hence, strengthening the social interactions that form the basis for thriving communities may be a more complicated task for policymakers to achieve than previously thought.

Our results corroborate the idea advanced by Glaeser et al. (2001) and more recently shown by Couture (2014) that the variety of venues on offer is a primary amenity to urban consumers. They also relate to the recent debate on how well cities can offer exposure to a diversity of opinions that might be crucial for the formation of accurate and pro-social beliefs. If similar people tend to hold similar views, then homophily might impact the diversity of opinions to which they are exposed. On the one hand, Sunstein (2009) suggests that physical interaction inside venues as well as outside in neighborhoods might be an important source of exposure to diverse views.[29] On the other hand, Gentzkow and Shapiro (2011) find that news media (both online and offline) offer more exposure to diverse opinions than neighbors, co-workers and family members do. Our findings help reconcile these two positions: physical interaction may well be a crucial source of exposure to diverse opinions, but the substantial endogenous sorting of people within cities and neighborhoods may mitigate the amount of diversity to which people are actually exposed, reducing the impact of such interaction on the formation of their beliefs.

More generally, the formation of peer groups is a deeply personal choice. Although it is certainly affected by where people live, study and work, people make many smaller decisions on a daily basis that can shape their social environments in profound ways. These might revolve around a seemingly insignificant action such as frequenting a specific venue, meeting a future partner, or joining a conversation that turns out to be memorable and impactful. While the informal and personal nature of these decisions makes them difficult to observe in standard data sets, the proliferation of user-generated "big" data sets has the potential to offer researchers a window into this rich source of socialization. We view this work as an early step along that path.

References

Adams, R. B., Ferreira, D., 2009. Women in the boardroom and their impact on governance and performance. Journal of financial economics 94 (2), 291–309.

Ananat, E., Fu, S., Ross, S. L., 2013. Race-specific agglomeration economies: Social distance and the black-white wage gap. Tech. rep., National Bureau of Economic Research.

[29] "The diverse people who walk the streets and use the parks are likely to hear speakers' arguments; they might also learn about the nature and intensity of views held by their fellow citizens. (...) When you go to work or visit a park (...) it is possible that you will have a range of unexpected encounters" (p. 30).

Andrews, J. A., Tildesley, E., Hops, H., Li, F., 2002. The influence of peers on young adult substance use. Health psychology 21 (4), 349.

Arcidiacono, P., Nicholson, S., 2005. Peer effects in medical school. Journal of public Economics 89 (2), 327–350.

Arribas-Bel, D., Bakens, J., 2014. "the magic's in the recipe": Urban diversity and popular amenities. mimeo.

Atkinson, A. B., 1970. On the measurement of inequality. Journal of economic theory 2 (3), 244–263.

Bayer, P., Ferreira, F., McMillan, R., 2007. A unified framework for measuring preferences for schools and neighborhoods. Journal of Political Economy 115 (4), 588–638.

Bayer, P., Ross, S. L., Topa, G., 2008. Place of work and place of residence: Informal hiring networks and labor market outcomes. Journal of Political Economy 116 (6), 1150–1196.

Blau, P. M., 1964. Exchange and power in social life. Transaction Publishers.

Bourguignon, F., 1979. Decomposable income inequality measures. Econometrica 47 (4), 901–920.

Brown, D. A., Brown, D. L., Anastasopoulos, V., 2002. Women on boards: Not just the right thing... but the" bright" thing. Conference Board of Canada.

Caetano, G., Maheshri, V., 2014. School segregation and the identification of tipping behavior. Mimeo.

Carrell, S. E., Sacerdote, B. I., West, J. E., 2013. From natural variation to optimal policy? the importance of endogenous peer group formation. Econometrica 81 (3), 855–882.

Chetty, R., Hendren, N., Kline, P., Saez, E., 2014. Where is the land of opportunity? the geography of intergenerational mobility in the united states. Tech. rep., National Bureau of Economic Research.

Conley, T., Udry, C., 2001. Social learning through networks: The adoption of new agricultural technologies in ghana. American Journal of Agricultural Economics, 668–673.

Couture, V., 2014. Valuing the consumption benefits of urban diversity. mimeo.

Ellison, G., Glaeser, E. L., 1997. Geographic concentration in us manufacturing industries: A dartboard approach. Journal of Political Economy 105 (5), 889–927.

Espelage, D. L., Holt, M. K., 2001. Bullying and victimization during early adolescence: Peer influences and psychosocial correlates. Journal of Emotional Abuse 2 (2-3), 123–142.

Gentzkow, M., Shapiro, J. M., 2011. Ideological segregation online and offline. The Quarterly Journal of Economics 126 (4), 1799–1839.

Glaeser, E., Sacerdote, B., Scheinkman, J., 1996. Crime and social interactions. The Quarterly Journal of Economics, 507–548.

Glaeser, E. L., Kolko, J., Saiz, A., 2001. Consumer city. Journal of economic geography 1 (1), 27–50.

Hill, A. J., 2015. The girl next door: The effect of opposite gender friends on high school achievement. American Economic Journal: Applied Economics 7, 147–77.

Hotelling, H., 1929. Stability in competition. The Economic Journal 39 (153), 41–57.

Hoxby, C., 2000. Peer effects in the classroom: Learning from gender and race variation. Tech. rep., National Bureau of Economic Research.

Huckfeldt, R. R., 1995. Citizens, politics and social communication: Information and influence in an election campaign. Cambridge University Press.

Jacobs, J., 1961. The death and life of great American cities. Random House LLC.

Kling, J. R., Liebman, J. B., Katz, L. F., 2007. Experimental analysis of neighborhood effects. Econometrica 75 (1), 83–119.

Lavy, V., Schlosser, A., 2011. Mechanisms and impacts of gender peer effects at school. American Economic Journal: Applied Economics 3 (2), 1–33.

Mas, A., Moretti, E., 2009. Peers at work. The American economic review 99 (1), 112–145.

McPherson, M., Smith-Lovin, L., Cook, J. M., 2001. Birds of a feather: Homophily in social networks. Annual review of sociology, 415–444.

Oreopoulos, P., 2003. The long-run consequences of living in a poor neighborhood. The quarterly journal of economics, 1533–1575.

Ostrom, E., 2000. Collective action and the evolution of social norms. The Journal of Economic Perspectives, 137–158.

Reardon, S. F., Firebaugh, G., 2002. Measures of multigroup segregation. Sociological methodology 32 (1), 33–67.

Roback, J., 1982. Wages, rents, and the quality of life. The Journal of Political Economy, 1257–1278.

Sacerdote, B., 2001. Peer effects with random assignment: Results for dartmouth roommates. Quarterly Journal of Economics 116 (2), 681–704.

Schelling, T. C., 1969. Models of Segregation. The American Economic Review 59 (2), 488–493.

Shannon, C. E., Weaver, W., 1963. Mathematical theory of communication. University Illinois Press.

Shorrocks, A. F., 1980. The class of additively decomposable inequality measures. Econometrica: Journal of the Econometric Society, 613–625.

Sunstein, C. R., 2009. Republic. com 2.0. Princeton University Press.

Theil, H., 1967. Economics and information theory. North-Holland.

Verbrugge, L. M., 1977. The structure of adult friendship choices. Social forces 56 (2), 576–597.

Weinberg, B. A., 2007. Social interactions with endogenous associations. Tech. rep., National Bureau of Economic Research.

Weitzman, M. L., 1992. On diversity. The Quarterly Journal of Economics 107 (2), 363–405.

Whitmore, D., 2005. Resource and peer impacts on girls' academic achievement: Evidence from a randomized experiment. American Economic Review, 199–203.