Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Mingtao Xu Date

Agglomeration Density and Business-Customer Matching

By

Mingtao Xu
Master of Science

Business Administration

_____
Giacomo Negro, Ph.D.
Co-Advisor

_____
Anand Swaminathan, Ph.D.
Co-Advisor

_____
L. G. Thomas, Ph.D.
Committee Member

Accepted:

_____
Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

_____
Date

Agglomeration Density and Business-Customer Matching

By

Mingtao Xu
M.S. Georgia Institute of Technology, 2013

Advisor: Giacomo Negro, Ph.D.
Advisor: Anand Swaminathan, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
In partial fulfillment of the requirements for the degree of

Master of Science in Business Administration

2015

Abstract

Agglomeration Density and Business-Customer Matching

By

Mingtao Xu

Studies in agglomeration have shown that agglomeration can benefit businesses by enhancing their productivity. In this paper, I examine the relationship between agglomeration density and customers' evaluation of businesses and propose a mechanism that represents a non-productivity benefit of agglomeration. In an urban setting, I argue that when businesses offer differentiated products and services and customers have different preferences, agglomeration of businesses provides customers with more diverse choices, enables them to try different products and services and find the ones that match their preferences the best. The outcome is that customers evaluate businesses more positively. This study uses Yelp data on chain restaurants, users, and reviews to test these propositions.

Agglomeration Density and Business-Customer Matching

By

Mingtao Xu
M.S. Georgia Institute of Technology, 2013

Advisor: Giacomo Negro, Ph.D.
Advisor: Anand Swaminathan, Ph.D.

A thesis submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
In partial fulfillment of the requirements for the degree of

Master of Science in Business Administration

2015

Table of Contents

List of Tables

List of Figures

Introduction

In this paper, I propose that agglomeration can work as an automatic matching mechanism that helps businesses find the right customers and promotes customers' evaluations of businesses. The paper intends to bridge related studies in agglomeration, organizations, and urban economics, and contributes to the literature by empirically examining the non-productivity benefits brought about by the improvement in matching between businesses and customers. Using Yelp data on chain restaurants, the paper aims to answer the following questions: 1) Is a higher agglomeration density associated with more positive customer ratings of businesses? 2) If it is, can the phenomenon be explained by better matching between businesses and customers?

The first question is grounded in a large literature on agglomeration economies. Scholars in economic geography have shown that agglomeration enhances productivity, heightens demand, and is associated with higher product variety (Combes & Gobillon, 2015; Fischer & Harrington, 1996; Hanson, 2001). On the production side, in a seminal work on agglomeration, Marshall (1920) points out that agglomeration can enhance productivity through three mechanisms: input sharing, knowledge spillover, and labor market pooling. Under the umbrella of transport cost reduction, the three mechanisms respectively represent reduced costs for moving goods, ideas, and people (Ellison, Glaeser, & Kerr, 2010).

Agglomeration benefits firms by lowering their costs for obtaining input goods. Marshall (1920) argued that firms that share inputs collocate close to suppliers to save transport costs, thus allowing collocated firms to acquire inputs at a lower price than isolated firms. Also, if inputs are sensitive to transport cost, which means the market for input goods is localized, downstream firms can create a larger market for input goods by collocating. In this manner, input suppliers can enjoy economies of scale that further lower firms' cost for obtaining input goods (Rosenthal & Strange, 2004). Second, agglomeration benefits organizational learning by enhancing the level of knowledge spillover among firms. Geographical proximity brings more chances for human interaction, which is essential to knowledge exchange. More frequent knowledge exchange

promotes learning, and enhanced learning within an agglomeration promotes firms' productivity. This mechanism of localized learning has been supported by empirical studies that suggest that patents of the same metropolitan statistical area (MSA) are more likely to be cited (Jaffe, Trajtenberg, & Henderson, 1993). Finally, producers in denser areas can reap productivity gains brought about by workers of higher quality (Costa & Kahn, 2000) and workers whose skills match the job requirements better. In a labor market where workers have heterogeneous skills and firms have different skill requirements, agglomeration helps workers find firms with the best skill match and leads to higher per worker productivity (Amiti & Pissarides, 2005; Wheeler, 2001). Wheeler (2001) quotes Glaeser (1994) to describe the scenario: "In a one-company town, individuals who are imperfectly matched to that company have nowhere else to go."

In addition to the productivity benefits, scholars have also studied agglomeration from the consumption side. Urban amenities provide a natural context for much research as they are both production centers and consumption centers (Glaeser, Kolko, & Saiz, 2001). Consumers in large cities can find products, services, and amenities that are not available in small cities (Glaeser et al., 2001), and that are more tailored by producers to match consumers' preferences (George & Waldfogel, 2003). Also, agglomeration of businesses reduces consumers' search cost and allow consumers to compare more businesses. From businesses' perspective, agglomeration heightens demand for their products and services by increasing their visibility to customers and increasing customers' subsequent purchases (Chung & Kalnins, 2001).

Scholars in management and industrial organization study agglomeration density from a competition perspective. The earliest work is Hotelling (1929) that suggests firms will choose to collocate if price competition does not exist. Later works in both organization theory and economics suggest that competition intensifies as organization density increases and organizations differentiate to avoid competition (Baum & Haveman, 1997; d'Aspremont, Gabszewicz, & Thisse, 1979; Hannan & Freeman, 1977). One important dimension of differentiation is quality. A recent study on online ratings finds that agglomeration improves

product quality. After removing reviewer effects, Gottlieb and Shkolnik (2014) found that coffee and tea places in locations with higher competitor density receive higher customer ratings (which indicates those coffee and tea places are of higher quality). The authors proposed that competition and knowledge spillover were potential mechanisms.

Higher customer ratings can reflect higher levels of quality; they also reflect customers' subjective evaluations of businesses. I would like to build on previous studies and examine whether a better matching between businesses and customers that resulted from greater diversity could be an explanation of the rating premium. This brings us to the second question of business-customer matching.

A substantial part of research on matching focuses on firm-worker matching in the labor market. Studies have found evidence on the positive association between diversity and goodness of matching. Greater firm diversity along the vertical productivity dimension improves matching in a way that high-productivity firms pay higher wages to court high-quality workers. And this improved quality matching amplifies wage inequality (Sorensen & Sorenson, 2007; Wheeler, 2001). On the other hand, greater firm diversity along the horizontal skill requirement dimension improves matching in that workers' different skills can be matched well with firms' different requirements. It reduces inequality since a worker's skill that is not valued by one firm may be valuable to another firm. (Sorensen & Sorenson, 2007). Studies have also found evidence on the positive association between preference matching and customer ratings. Kovács and Sharkey (2014) studied the effect of winning a prestigious book award on book ratings and reported that an award brought the book readers who were not predisposed to the book. Those readers read the book not because there is a strong fit in tastes but because reading the award-winning book signals their social identity, and those reviewers contribute to negative reviews of the book.

There exists indirect evidence that agglomeration density improves preference matching. Couture (2015) used Google restaurant data and travel time data to measure customers' "gains from density". Given a larger restaurant density, consumers should travel a shorter distance to

find a restaurant. However, after controlling for congestion, the author found that consumers' travel time did not reduce that much. This finding implies that in denser areas, although consumers drive past many restaurants on the way, they choose to dine in restaurants that are more distant but match their preferences better. In denser areas, people trade time for a better match. An extension of the argument could be that a restaurant in a denser area has a better fit with its customers compared to one in a less dense area.

The article proceeds as follows. First, I develop theoretical hypotheses regarding how density affects business-customer matching. Second, I describe my empirical strategy. Third, I give details of the dataset and measurements of constructs. Fourth, I present my models and report results. Fifth, I discuss my findings and conclusion. Finally, I lay out the limitations of this study and consider future directions.

Theory and Hypotheses

Density and Ratings

In this section, I discuss the relationship between agglomeration density and customer ratings in an urban business setting. Ratings of urban businesses such as restaurants, bars, and barber shops reflect levels of quality of their products and services as well as customers' perception. On average, the higher the business's quality, the higher the ratings it should receive from its customers.[1] Thus, factors that improve the quality should also bring better evaluations to the businesses. Agglomeration research suggests that quality can benefit from proximate firms as a consequence of knowledge spillover or competition intensity. Knowledge spillover occurs when the focal firm can observe and learn from proximate firms and improve its quality. Competition

[1] If we consider customers' perceptions, an argument can be made that the better a business meets customers' expectations, the higher the ratings it should receive from its customers. Higher quality leads to higher ratings only when the quality meets or exceeds customers' expectations. Anderson (1973) contains a review on psychological theories explaining the relationship between expectations and perceived product performance.

also leads to higher quality. First, it exerts pressure on businesses and forces them to improve their quality to avoid price competition that may substantially hurt profit. Second, from an ecological perspective, localized competition increases the failure rate of businesses (Baum & Mezias, 1992). The same business that can survive in a less dense area may not be able to survive in a denser area where there exists more fierce competition. The result is an overall higher quality of urban businesses in denser areas.

From the customer side, there is another mechanism that may lead to better evaluations. Better customer evaluations can come from a different set of customers. Customers who choose a business in a denser area differ from customers in a less dense area in that they have a better fit with characteristics of the business. Given a cluster of differentiated businesses instead of a single one, customers choose to go to a place, rather than "having nowhere else to go". Because of the greater variety of options, customers in areas with high business agglomeration density are more likely to find one that matches their preferences better. For example, one is more likely to find a dress that fits her perfectly in a large shopping mall with dozens of clothing stores than in an isolated clothing store. Fit can come from preferences in various aspects of the product and of the business, from price, texture, style, designer, brand reputation, to even layout, brightness and music of the store. A better fit enhances customers' perceived quality of products and services and increases customers' likelihood of giving better evaluations.

As a result, we can predict a positive density-rating relationship among urban businesses created by improved business-customer matching.

Proposition 1: For a given quality level, higher agglomeration density is associated with better customer evaluations.

Hypothesis 1: For a given quality level, higher restaurant density is associated with higher restaurant ratings.

I argue that the mechanism described above consists of two processes: first, higher density leads to greater diversity among businesses; second, greater diversity leads to improved

matching between businesses and customers. As a consequence of the improved matching, businesses receive higher customer ratings. The discussion below address these processes.

Diversity and Categories

Consumers differ in their preferences, and businesses differ in their offerings. With more options, consumers are more likely to find businesses that match their preferences better. Key to this logic is business diversity, and diversity comes from differentiation.

As density increases, resource competition among organizations becomes more intense, resulting in higher failure rates or lower levels of performances (Carroll & Hannan, 1989). Organizations differentiate to relieve the pressure that comes from competition (Blau, 1970; Hannan & Freeman, 1977; Lincoln, 1979; Swaminathan & Delacroix, 1991). Studies have shown that organizations differentiate vertically and horizontally. In their study of metropolitan areas in the U.S., Berry and Waldfogel (2010) examined quality differentiation (i.e., vertical differentiation) in the restaurant industry and found evidence that the quality range increases with market size. Baum and Haveman (1997) studied hotels in Manhattan and found that while hoteliers collocate to enjoy agglomeration benefits, they differentiate in size to avoid localized competition. In his modeling paper, Kuksov (2004) demonstrated that when searching costs are reduced (which is one of the consequences of agglomeration), firms choose to differentiate their products to alleviate price competition.

Proposition 2: Higher agglomeration density is associated with greater business diversity.

In my investigation of the restaurant industry, instead of quality or size, I examine differentiation along cuisine categories. I focus on the diversity in cuisine categories because cuisine categories are one of the most distinguishable characteristics of restaurants and are one of the most important aspects when customers evaluate their fit with restaurants. When diners make their restaurant choices, a common question is “What/where do you want to eat?” The answer is most likely to reflect one’s cuisine category preferences, such as “Spanish”, “French”, “Sushi”,

etc. It is true that the answer can also be "somewhere not too crowded," "somewhere close," "somewhere cozy," or "somewhere cheap". But compared to cuisine categories, some of these characteristics can hardly serve as objective foundations of restaurant categorization since they are subject to changes in time (such as crowdedness) or in customers (such as proximity and coziness). Other characteristics such as costliness can be used to categorize restaurants, but customers' preferences on them are subject to change. Thus, cuisine categories serve as an ideal foundation for studying the matching between restaurant categories and customers' preferences. The table below lists all Yelp restaurant categories in the U.S.

[Insert Table 1 about here]

Restaurants differentiate in cuisine categories as a response to competitive pressure. If an area already has a high restaurant density, an entrant is more likely to choose a different cuisine category to differentiate itself from existing restaurants. An incumbent who has no differentiation advantage is more likely to fail or move away from this area to avoid homogeneous competitors. As a result, restaurants that remain in a dense cluster will exhibit a high degree of diversity. I argue that cuisine categories in denser areas are more diverse. Hypothesis 2 tests Proposition 2 in this empirical setting and is stated as follows:

Hypothesis 2: Higher restaurant density is associated with greater restaurant cuisine category diversity.

Matching and Fit

A result of diverse offerings in an area is a better matching between businesses and customers. From businesses' perspective, a population of businesses with greater diversity offers a finer partition of the market (consumers), which enables each business to have a better fit with its customers. The more diverse the neighboring businesses are, the more people neighboring businesses will attract who have poorer fit with the focal restaurant. Figure 1 compares situations where there are one and two options available to consumers. Changing from one option to two, consumers of triangle ABC, who prefer r_2 over r_1, switch to r_2, while consumers of ABDE still

choose r_1. This process continues as the number of proximate competing businesses increases. In the end, only consumers who prefer the focal business the most are left and become its customers. The result of this group of highly matched customers is higher ratings for r_1. More detailed illustration of how diversity affects matching is given in Appendix 1.

[Insert Figure 1 about here]

Given the positive association between density and diversity, I argue that agglomeration density indirectly affects business-custom matching. Agglomeration density does not directly affect matching. Consider a hypothetical scenario that businesses and their products and services in an area are perfectly substitutable and completely homogeneous, then consumers' preferences are the same for each option, and business-customer matching will not be improved. This suggests that diversity is a mediator of density's effect on business-customer matching.

Proposition 3: Business diversity mediates the effect of agglomeration density on business-customer matching.

In the context of restaurants, if consumers choose a restaurant out of all other nearby surrounding restaurants with diverse offerings, they must have strong preference for the focal restaurant. Otherwise, consumers will simply go to dine elsewhere. On average, with greater diversity, the market share of each restaurant will decrease but each restaurant will be better matched with its market segment (customers). I hypothesize:

Hypothesis 3: Higher diversity is associated with better matching between restaurants and customers.

If Hypothesis 2 and Hypothesis 3 hold, Hypothesis 4 below will also hold:

Hypothesis 4: On average, restaurants in denser areas have better fit with their customers than restaurants in less dense areas.

Associations hypothesized in H4, H2 and H3 tests Proposition 3 and examine diversity as a mediator of the effect of density on restaurant-customer matching.

Evaluations and Ratings

In this section, I discuss how better matching affects customers' evaluations. Products in categories that we like more give us more satisfaction. A higher level of satisfaction enhances a customer's perceived quality of an organization (Gotlieb, Grewal, & Brown, 1994). The outcome of better perceived quality is a higher likelihood that a customer will give better evaluations. Kovács and Sharkey (2014) showed that a poor fit between readers' tastes and books may contribute to negative book reviews. In the restaurant context, a poor match between customers' preferences and restaurants' cuisine categories may lead to unsatisfactory dining experiences, resulting in low levels of satisfaction and low levels of perceived quality. In the Yelp case, a better or worse evaluations are translated to higher or lower star rating associated with a review. Restaurants with a group of customers whose preferences are well-matched with the restaurants' offerings are likely to collect more higher-rating reviews and have higher average ratings.

Hypothesis 5: Better matching between restaurants and customers is associated with higher restaurant ratings.

With Hypothesis 5, we have a complete chain of associations from restaurant density to restaurant ratings. These associations suggest the indirect effect of diversity on ratings via the mediator of matching. However, it is possible that there exists a direct path from diversity to ratings, which means that for the same group of customers, being exposed to more options is associated with higher ratings. Psychological studies by Reibstein, Youngblood, and Fromkin (1975) and Gotlieb et al. (1994) suggest that more options give customers a feeling that they have more freedom and more control over their choices, which lead to higher satisfaction and potentially higher ratings. I will also examine this direct relationship in the models.

In all, there are two mediators through which urban business agglomeration density affects customer evaluations. First, diversity mediates the effect of density on matching improvement. Second, better matching mediates the effect of diversity on ratings. The diagram

below summarizes the hypothesized relationships and major previous studies that this paper is indebted to.

[Insert Figure 2 about here]

Variety-Seeking and Omnivore

I have argued that with high diversity in surrounding restaurants, a focal restaurant's customers will have a strong preference for categories that the restaurant is in. They may either frequently visit restaurants in these categories or always give high ratings for restaurants in these categories. In addition, diversity may attract consumers who are more omnivorous. Psychological studies have shown that consumers have a tendency to seek variety (McAlister & Pessemier, 1982). In his study of the aesthetics of elites, Peterson (1992) use the term omnivore to describe the people who have appreciation of all distinct art categories. In the context of this paper, I define omnivore as people who appreciate different cuisine categories. With more choices at hand, omnivores can change their choices to gain a level of stimulation from novelty. Thus, it is possible that omnivores tend to go to areas with diverse options since the variety in those areas allows them to try different categories. This variety-seeking argument predicts that customers of restaurants in denser areas are more omnivorous, which means they do not have strong preferences for the categories of the restaurant they choose. This prediction is opposite to the prediction of improved matching. Empirically, customers' omnivorousness should be controlled for.

Empirical Strategy

Empirically, I use a dataset on 4,200 chain restaurants to test my hypotheses. One reason for using chain restaurants is to control for variations in quality. Controlling for quality is necessary since a natural alternative explanation to higher ratings of restaurants in denser areas is that those restaurants in denser areas have higher levels of quality. Scholars have used different

variables to measure quality; the variables include hygiene inspection scores (Jin & Leslie, 2003), reviewer-adjusted user-generated ratings (Gottlieb & Shkolnik, 2014), and expert-generated ratings (Berry & Waldfogel, 2010). My strategy is to use chain restaurants to eliminate the effect of product quality. After controlling for between-chain rating differences, within-chain rating differences should be isolated from effects of quality. Another reason for using chain restaurants is that their online ratings suffer less from fraud. In general, fraudulent reviews make up 16% of total Yelp reviews and give more extreme ratings (Luca & Zervas, 2015). However, fraud is less of a problem for chain businesses because their reputations have been firmly established by marketing and advertising so that they don't have much to gain from fraudulent reviews (Luca, 2011; Luca & Zervas, 2015; Mayzlin, Dover, & Chevalier, 2014).

Data

Yelp Data

Yelp.com initiated a Yelp Data Challenge[2] and published a dataset consisting of 1,569,264 reviews from 366,715 users, and 61,184 businesses in 10 cities. Along with each review, there is an integer star rating for the business from one to five. The 61,184 businesses are in ten cities: four cities overseas (Edinburgh, UK; Karlsruhe, Germany; Montreal and Waterloo, Canada) and six American metropolitan areas (Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, and Madison). Among the 61,184 businesses, 21,892 are restaurants. 60,785 out of 61,184 businesses and 21,799 out of 21,892 restaurants have reviews. The 21,799 restaurants have 986,672 reviews. Figure 3 shows distributions of review ratings. Figure 4 shows distributions of restaurants' and all businesses' average ratings. Noticing that the distribution for restaurants is less skewed towards the right side than the distribution for other businesses.

[Insert Figure 3 about here]

[2] More details are presented at: http://www.yelp.com/dataset_challenge.

[Insert Figure 4 about here]

Due to the availability of demographic data, I only study the six U.S. metropolitan areas. The six U.S metropolitan areas have 52,795 establishments.[3] Among them, 17,665 contain a "Restaurants" category label and I identify them as restaurants. Establishments such as cafes or food shops without a "Restaurants" label are excluded. These 17,665 restaurants are reviewed by 255,868 reviewers and have 940,548 reviews. Among these 17,665 restaurants in the U.S., I select chain restaurants with 10 or more locations. This chain restaurant dataset contains 4,275 businesses of 120 chains and their 84 thousand reviews.[4] Among the 4,275, 4,200 are left with complete demographic information and restaurant categories.[5]

Table 2 summarizes Yelp data by city:

[Insert Table 2 about here]

Raw data on businesses includes businesses' unique encrypted business IDs, names, stars (rounded to half-stars), review counts, categories, and locational variables (latitudes, longitudes, city, and state). Raw data on users include users' IDs, review counts, average stars, elite years, and dates started yelping. Raw data on reviews contains reviewed businesses' ID, reviewing users' ID, stars, texts, and dates. Samples of a business, a user, and a review entry is given below:

[Insert Table 3 about here]

[3] There are in total 61,184 businesses. Among them, 60,785 have reviews and 52,842 are located in the US. Out of 52,842, 52,795 businesses can be matched to ZCTA code and have ACS data. This gives us a dataset with 52,795 establishments with business and demographic data.

[4] Names of restaurants have inconsistencies. For example, "McDonald's" also appears as "McDonalds" or "Mcdonalds". The name "Bojangles'" has ten variations. These are all recognized as different names. The author hand corrected these names.

[5] Restaurants only have a single "Restaurants" label but lack restaurant categories information are excluded.

Demographic Data

I collect demographic data at Zip Code Tabulation Area (ZCTA) [6] level from 2009-2013 5-year American Community Survey (ACS) estimates. Reviews in the dataset were written in a period ranging from December, 2004 to January, 2015. The 2009-2013 period covers the five years with most reviews in the dataset.[7] ZCTA-level variables include: median household income (MHHI)[8], population density, the percentage of Whites, the percentage of Hispanics, male to female sex ratio, the percentage of population with a graduate or professional degree, with bachelor's degree, and with high school degree. The 4,200 chain restaurants locate in 229 ZCTAs.

Variables and Measurements

Categories

I use category labels of Yelp restaurants as my category system. The set *C* contains all the *N* restaurant categories (e.g. Mexican, Korean, Ethiopian), $C = \{c_1, c_2, \dots, c_N\}$. For each restaurant r, categories that it belongs to are denoted as $C_r = \{c_{r_1}, \dots, c_{r_N}\}$, $r_N \leq N$, and $C_r \subseteq C$. In most cases, C_r has no more than four elements. Since my observations are restaurants, the category of "Restaurants" is the common category for all of them. I remove the common "Restaurants" label and only count subcategories of "Restaurants". As an example, Mon Ami Gabi, the most reviewed restaurant in Las Vegas, has a category set consisting of three elements: "French", "Steakhouses", and "Breakfast & Brunch" after removing the "Restaurants" category.

---

[6] The Census uses Zip Code Tabulation Area (ZCTA) for zoning, which is not exactly the same as the USPS Zip Code area. More information can be found at: https://www.census.gov/geo/reference/zctas.html.
[7] There are only a few reviews at the end of 2004 or the beginning of 2015. The more recent 2010-2014 ACS 5-year estimates are still not available, and the planned release date is December 10, 2015.
[8] Among 52,849 U.S. businesses, 52,832 of them have corresponding ACS data. 1,135 of the 33,120 ZCTAs have no MHHI data. Among the rest, 13 have MHHIs labeled as "smaller than 2500" and 10 have MHHIs labeled as "greater than 250000", they are manually imputed as 2,499 and 250,001 respectively in my data. But none of the chain restaurants in the dataset locates in those extremely wealthy or poor ZCTAs, thus my income data is not censored.

The data also contain some 2nd-tier categories, such as Calabrian, Sardinian, and Tuscan under the 1st-tier category of Italian, as shown in Table 3. I flatten the hierarchical structure and keep both 1st-tier and 2nd-tier categories. Admittedly, some categorizations are tricky. For instance, "Sushi Bars" is not under "Japanese" and is a 1st-tier category; "Tapas Bars" and "Tapas/Small Plates" are two distinct 1st-tier categories.

Variables on Restaurants

Ratings

I calculate business star ratings from stars of each review.[9] The star rating for a restaurant is calculated as the arithmetic average of all the star ratings in the reviews of that restaurant. The equation is:

$$avg_r = \frac{1}{I_r}\sum_{i=1}^{I_r} s_{ir}$$

where s_{ir} is individual *i*'s star rating for restaurant *r*, and I_r is the total number of reviews of restaurant *r*.

A more sophisticated way to understand the ratings is through the following equation, which is also used in Gottlieb and Shkolnik (2014):

$$s_{ir} = \lambda_i + \gamma_r + \tau_{ir}$$

where λ_i is the fixed-effect attributed to reviewers, γ_r is the fixed-effect attributed to restaurants. The residual τ_{ir} can be interpreted as differences in different reviewers' ratings of the same restaurant, or differences in the same reviewer' ratings of different restaurants Either way, this residual τ_{ir} should be associated with individual preference p_{iC^j} ($C_r = C^j$). The benefit of using this τ_{ir} lies in its ability to separate producer-side and customer-side effects on ratings. Relating the two measures of ratings, I have $avg_r = \overline{s_{ir}} = \gamma_r + \overline{\lambda_\iota + \tau_{\iota r}}$. The first part γ_r captures the

[9] I don't use stars Yelp gives to each business because those stars are rounded to the nearest half-star. In their calculation, if a restaurant has an average rating of 3.74 stars, Yelp will display 3.5 stars, but if a restaurant has 3.76, Yelp will display 4 stars (see Anderson and Magruder (2012)).

fixed-effect of restaurant quality; the second part $\overline{\lambda_\iota + \tau_{\iota r}}$ captures the effect of customers. If we control for the fixed-effect of restaurant and isolate its effect on avg_r, from the effect of reviewers, then avg_r should give us pure information about the effect of a restaurant's customers on ratings. By using chain restaurants and adding chain-specific dummies, I control for within-chain restaurant quality difference, and avg_r serves as an ideal measure of customers' subjective evaluation of businesses.

Popularity

Besides star ratings, popularity is another important characteristic of a restaurant. I use the number of reviews a restaurant received to measure its popularity, denoted as rc_r. Assume that the propensity to write a review is the same across customers of restaurants, although only a small proportion of customers write reviews, I may still contend that a higher review number indicates more customers.

Chains

Restaurants in the same chain have very similar (if not identical) menus and share lots of resources, bus each chain has its idiosyncratic characteristics. In addition to chain-specific dummies, I add two chain-level that are the number of restaurants in each chain in the dataset (*chaincou*) and the mean star ratings of each chain (*chainmean)*. Figure 5 ranks the 120 chains in the dataset by average ratings. Figure 6 ranks the 120 chains in the dataset by their numbers of restaurants. Figure 7 gives the distribution of average ratings of the 120 chains.

[Insert Figure 5, Figure 6, and Figure 7 about here]

Variables on Neighboring Restaurants

Density

Most organizational studies define density as the number of organizations in a given domain, the density here is the number of business establishments in a given geographical area.

For each restaurant *r*, density is measured by five variables, which are numbers of restaurants (including the focal restaurant itself) within areas of radiuses 0.2, 0.5, 1, 2, and 5 mile(s). I use locations of all restaurants in the Yelp dataset to calculate densities. The five density variables for a restaurant *r* are denoted as: $count02_r$, $count05_r$, $count1_r$, $count2_r$, and $count5_r$. Figure 8 show distributions of these five variables for the 4,200 restaurants in the dataset.

[Insert Figure 8 about here]

An issue about these density measures is that they are only proxies of density and are not equal to the true neighboring environment of a restaurant, because it is not guaranteed that all of those restaurants in the dataset operate at the same time. Some restaurants may have closed permanently and some others may just opened recently. The true restaurant density changes over time. My measures aims to describe the environment around the location of the focal restaurant generally, not to portrait the dynamic evolution process of agglomeration density. Measuring density as a longitudinal variable can be a major future development.

Diversity

I use measure diversity in two ways. First, I count the number of unique categories represented in an area. For a focal restaurant, diversity is calculated as numbers of represented categories within distances of 0.2, 0.5, 1, 2, and 5 miles. As an example, if within an area with a radius of 1 mile, there is a German and American restaurant and an American and Mexican restaurant, then the diversity index will be 3. Let C_{rm_l} be the category set of one of the N_{rm} restaurants that are within an *m*-mile distance from the focal restaurant *r*, $\bigcup_{l=1}^{N_{rm}}(C_{rm_l})$ denotes the union set of categories of all these N_{rm} restaurants. The diversity within this *m*-mile-radius area around restaurant *r* can be written as the cardinality of the set:

$$Card_{rm} = \left| \bigcup_{l=1}^{N_{rm}} C_{rm_l} \right|$$

where $m \in \{0.2, 0.5, 1, 2, 5\}$, $l \in \{1, 2, \dots, N_{rm}\}$. Similar to the density measurement, for each restaurant *r*, this measure of diversity has five elements, I denote them as: $rdiv02_r$, $rdiv05_r$, $rdiv1_r$, $rdiv2_r$, and $rdiv5_r$.

Second, based on categories represented in each restaurant's surrounding area, for each area, I calculate the Simpson Index [10], which is a widely adopted measure of diversity. For the N_{rm} restaurants that are within an *m*-mile distance from restaurant *r*, I count the total of categories $CN_{rm} = \sum_{N_{rm}} |C_{rm_l}|$. Among them, CN_{rm,c_k} is the number of restaurants that belong to category c_k, then the "share" of category c_k is: $s_{c_k} = \frac{CN_{rmc_k}}{CN_{rm}}$. Simpson Index follows as one minus the sum of squares of these "shares":

$$si_{rm} = 1 - \sum_{r=1}^{R_i} {s_{c_k}}^2$$

The larger the Simpson Index, the more diverse a restaurant's surrounding environment is. I calculate Simpson Indices at five radius levels for each restaurant: $si02_r$, $si05_r$, $si1_r$, $si2_r$, and $si5_r$.

Popularity of Neighboring Restaurants

I calculate average review counts of restaurants within radiuses of 0.2, 0.5, 1, 2, and 5 miles to measure the general popularity of those areas. Notations are $rc02_r$, $rc05_r$, $rc1_r$, $rc2_r$, and $rc5_r$.

Quality of Neighboring Restaurants

My arguments are around preference matching, which I argue is a result of diversity, but not knowledge spillover. The quality of neighboring restaurants matters for the knowledge spillover mechanism, as it may enhance the efficacy of learning. I include the quality of neighbors in the analysis to separate the effect of density (the number of neighbors) from

[10] The same diversity index is more often called Herfindahl–Hirschman Index in economics to use firms' market shares to describe the concentration of an industry. In Ecology, the Simpson Index (Simpson, 1948) is used to measure group diversity when individuals are in different categories.

spillover (the quality of neighbors). For the radiuses of 0.2, 0.5, 1, 2, and 5 miles, average star ratings of neighboring restaurants are calculated and used as proxies for quality of neighboring restaurants. Their notations are $avg02_r$, $avg05_r$, $avg1_r$, $avg2_r$, and $avg5_r$.

Preference and Fit

Individual Preference

By saying "preference", I mean the extent to which an individual likes a category; by using "fit", I measure how well is a business matched with its customers' preferences. It is hard to measure customers' preferences on environment and service, as the categories of those factors are not available in data. Thus in operationalization, I assume that customer's ratings reflect their preferences on cuisine types and I only test the matching along cuisine categories. Restaurant categories are given by Yelp, and I will calculate preference scores for each customer. Notice that only reviews are observed, I use this revealed preferences as a proxy of true consumers' preferences, so the measurements rely on the assumptions that a customer's tendency to write a review for different categories are the same.

I adopt four different measures of customer preferences. First, I use the original method by Kovács and Sharkey (2014), which is the total stars of all the restaurant a reviewer reviewed that are in that category, divided by the total number of restaurants one reviewed in that category. Given an individual *i*'s star rating to restaurant *r*, denoted as s_{ir}[11], for all the categories $c_k \in C$, whether or not a category is associated with restaurant *r*, I write down *i*'s preferences to each c_k from s_{ir} as:

$$p_{irc_k} = \begin{cases} s_{ir}, & if\ c_k \in C_r \\ 0, & if\ c_k \notin C_r \end{cases}$$

If we aggregate the individual's star ratings for all restaurants in category c_j and take the average, we get the individual's preference score of category c_k:

[11] In the dataset, since Yelp gives five star options, $s_{ir} \in \{1, 2, 3, 4, 5\}$.

$$p1_{ic_k} = \frac{1}{R_{ic_k}} \sum_{r=1}^{R_i} p_{irc_k}$$

where R_{ic_k} is the number of restaurants *i* reviewed that belong to category c_k.

Second, a customer's preference for a category is calculated as the total stars of all the restaurant one reviewed that are in that category, divided by the total number of restaurants he/she reviewed. This preference measure is a modified version of the measure by Kovács and Sharkey (2014). The definition is as follows:

$$p2_{ic_k} = \frac{1}{R_i} \sum_{r=1}^{R_i} p_{irc_k}$$

where R_i is the number of restaurants *i* reviewed.

As an example, a reviewer X reviewed restaurant A, B, C, and D. X gave A, a Thai restaurant, a 4-star rating; B, a Japanese and Korean restaurant, a 3-star rating; C, a Korean restaurant, a 4-star rating, and D, another Korean restaurant, a 5-star rating. Then X's preference for Thai is 4/4=1.00, for Japanese is 3/4=0.75, for Korean is (3+4+5)/4=3.00. Note that I impute 0 for all the categories that a restaurant does not belong to. Taking the Thai restaurant as an example, X's preference for that restaurant is 4/4=(4+0+0+0)/4=1.00. This method takes the frequency of visiting into consideration. The logic behind is that more visits to restaurants in a category show one's preference for that category. Otherwise, if we simply take the average of ratings, the single one review to Thai restaurant A will produce a 4-star preference for the Thai category, which makes X have the same preference score for Thai category and Korean category. The fact that three out of the four reviews are about Korean restaurants is not reflected in that calculation.

Third, I calculate the ratio of R_{ic_k} and R_i. This proportion tells us how many of an individual *i*'s reviewed restaurants are in the category c_k. The higher the proportion, the more frequent *i* chooses category c_k. The definition is as follows:

$$p3_{ic_k} = \frac{R_{ic_k}}{R_i}$$

By definition, there is the following relationship among these three measures:

$$p3_{ic_k} = \frac{p2_{ic_k}}{p1_{ic_k}}$$

The three measures above rely on another two assumptions: 1) the more one likes a cuisine category, the higher ratings one gives to restaurants in that category ($p1_{ic_k}$and $p2_{ic_k}$); 2) the more one likes a cuisine category, the more often one chooses a restaurant in that category ($p3_{ic_k}$). The three measures allows inter-personal comparison of preferences. For example, we can infer from these measures that an Individual 1 prefers French category more than another Individual 2 if (1) Individual 1 gives an average of 4.5 stars to French restaurants and another Individual 2 gives 3.4 to French restaurants, or (2) 80% of A's reviews are about French restaurants and only 40% of B's reviews are about French restaurants.

However, one can argue that such comparisons are incorrect. When consumers are making decisions of which category to choose, the comparison is mostly subjective. One's preferences on cuisine categories are compared with one's own preferences for other categories. Each consumer has his/her own subjective ranking of cuisine categories based on his/her own preferences. Using this idea, numerical star ratings are more like realizations of one's ordinal ranking of categories other than cardinal numbers.

The fourth measure is constructed based on the above idea. For each individual *i*, I put down $p1_{ic_k}$, which is his/her average ratings for each category. I then rank categories by $p1_{ic_k}$ from high to low. In this way, I have each individual's subjective ranking of categories. Denote a category c_k's rank in individual *i*'s ranking as κ_{ic_k}, and individual *i*'s total number of reviewed categories as K_i, I standardize each individual's ranking scale and compute *i*'s preference for c_k as a percentile:

$$p4_{ic_k} = 1 - \frac{\kappa_{ic_k}}{\mathrm{K}_i}$$

where $p_{ic_k}^4 \in (0,1]$. The higher the $p_{ic_k}^4$, the more *i* prefers category c_k.

Establishment Level Fit

Consider a category combination $C^j = \left\{c_1^j, c_2^j, \dots, c_{N_j}^j\right\} \subseteq C$, individual *i*'s preference for the category combination (C^j) is the average of his/her preferences for each of the categories contained in C^j.

$$p_{iC^j} = \frac{1}{N_j}\sum_{N_j} p_{ic_k}, \forall c_k \in C^j$$

where N_j is the number of categories C^j contains.

For a restaurant *r* of categories C^j, the restaurant's fit score with its customers is calculated as the average of all its reviewers' individual preferences for that category combination p_{iC^j}.

$$Fit_r = \frac{1}{I_r}\sum_{i=1}^{I_r} p_{iC^j}$$

where I_r is the total number of reviews of restaurant *r* [12]. I have four Fit_r's ($Fit1_r$, $Fit2_r$, $Fit3_r$, and $Fit4_r$) for each restaurant, corresponding to four measures of customers' preferences. Figure 9 gives the distribution of these veriables for chain restaurants.

[Insert Figure 9 about here]

Note that for a restaurant *r* , if $C_r = C^j$, s_{ir} may not equal to p_{iC^j}. Individual preference p_{iC^j} is defined by category combinations, not by restaurants. A restaurant's categories link the restaurant with individual preferences. An individual's preference scores of restaurants with exactly same categories are the same.

[12] Notice that I_r is the total number of reviews, not reviewers, this is because some users update their reviews of the same restaurant. I treat multiple reviews from the same person on a restaurant as different reviews. In operationalization, I_r is the total number of reviews, and I aggregate p_{rC^j} by reviews, not reviewers.

As an example, X's preferences for restaurant A, B, C, and D are 1.00, 1.875, 3.00, and 3.00 respectively. If another individual Y also reviewed A and B and the preference scores are 2.00 and 3.125 respectively. Assume that X and Y are the only two reviewers of A, B, C, and D, then if we combine the preferences scores of X and Y, I can get the fit scores of A, B, C, and D with their reviewers are 1.50, 2.50, 3.00 and 3.00 respectively.

There is a tricky issue regarding $Fit3_r$, because this measure three also reflect customers' variety seeking behavior. While the preference matching increases a restaurant's fit with its customers and positively affect $Fit3_r$, customers variety seeking behavior diversify their choices and negatively affect $Fit3_r$. The total effect on $Fit3_r$ remains uncertain.

Differences in Customers

I compute several variables from review data to capture differences in customers of different restaurants, including differences in geographical pattern and categorical pattern of their reviews.

Visitors

It is possible that visitors (tourists or business travelers) have different criteria or expectations about restaurants and different willingness to travel than local residents. To control for this effect, I calculate a user-city index to imply whether a city is an individual's city of residence or not.

Yelp data gives the number of reviews of each user. However, for many users, a large proportion of their reviews are out of this dataset on the six U.S. MSAs. So for an individual *i*, denote his/her total number of reviews as V_i and numbers of reviews in the metro area of Phoenix, Las Vegas, Pittsburgh, Charlotte, Madison, and Urbana-Champaign as $V_{i1}, V_{i2}, V_{i3}, V_{i4}, V_{i5}$ and V_{i6} respectively. The indices of individual *i* in MSA *j* are calculated as:

$$v_{ij} = \frac{V_{ij}}{V_i}$$

where $j \in \{1, 2, 3, 4, 5, 6\}$. I have six v_{ij}'s for each reviewer. $\sum_j V_{ij} / V_i$ is often smaller than one since a part of a reviewer's reviews are not in these six MSAs.

A larger index means a larger proportion of one's ratings are in the focal metro area, which makes one less likely to be a visitor (more likely to be a resident) of that metro area.[13]As an example, if Individual 1 wrote one review in City A and nine reviews in City B, and these are all her reviews, then Individual 1's visitor indices for City A and City B are 0.1 and 0.9 respectively. A v_{ij} is attached to each review based on the reviewer and the location of the restaurant. Then aggregating these v_{ij}'s to the establishment level, I denote this variable as vi_r. The higher the vi_r, the more "local" the customers of restaurant *r*.

Omnivorousness

Some customers only go to restaurants of one or a few categories whereas some others are more omnivorous and like to try different categories. To capture this heterogeneity among customers, I calculate the Simpson Index (usi_i) for each individual and count the number of categories one reviewed (ucc_i). A higher Simpson Index (usi_i) means an individual writes reviews more evenly in different restaurant categories. The higher the Simpson Index or the larger the number of reviewed categories, the more omnivorous the reviewer is. I aggregate these two variables to the establishment level and take the average, then add variables of $avgucc_r$, its log $laucc_r$, and $avgusi_r$ to each restaurant. The higher the three variables, the more omnivorous the customers of a restaurant are. $avgucc_r$ and $avgusi_r$ are calculated as:

[13] Since the answer to the question of whether a city is the city of residence of a reviewer is by nature binary. I can also construct a binary variable based on v_{ij}. A threshold $\bar{v}$, such as 0.5, can be used to distinguish the home city and visiting cities. All ratings in cities that don't meet the threshold will be labeled as visitors' ratings. When $\max_j(v_{ij}) < \bar{v}$, this means none of the six MSAs is the home city of *i*, and all of individual *i*'s ratings will be labeled as visitors' ratings. The visitor effect variable is a 0-1 dummy that distinguishes ratings of visitors from those of residents:

$$v_{ij} = \begin{cases} 1, & if\ v_{ij} < \bar{v} \\ 0, & if\ v_{ij} \geq \bar{v} \end{cases}$$

$$avgusi_r = \frac{1}{I_r}\sum_{i=1}^{I_r} usi_i$$

$$avgucc_r = \frac{1}{I_r}\sum_{i=1}^{I_r} ucc_i$$

where I_r is the total number of reviews of restaurant *r*.

Experience

Reviewers have different levels of experience and familiarity with Yelp. I assume that the more restaurant reviews the reviewer writes, the more experienced the reviewer is. Denote an individual *i*'s number of restaurant reviews written as urc_r_i, then I calculate the average number of restaurant reviews that reviewers of a restaurant have written and denote it as $uavgrc\_r_r$. The larger the $uavgrc\_r_r$., the more experienced the customers are.

$$avgurc\_r_r = \frac{1}{I_r}\sum_{i=1}^{I_r} urc\_r_i$$

where I_r is the total number of reviews of restaurant *r*

Demographic Variables

At the ZCTA (Zip Code Tabulation Area) level, demographic variables of median household income (MHHI), population density, the proportion of Whites, Hispanics, sex ratio, and percentages of population with graduate or professional degree, bachelor's degree, and high school degree are directly from 2009-2013 ACS, matched to each restaurant by ZCTA code using ArcGIS. Additional controls include chain-specific and state-specific fixed-effects.

Table 4 below lists all variables of the chain restaurant dataset and their descriptive statistics.

[Insert Table 4 about here]

Figure 10 gives the correlation matrix of variables. Variables with a radius of 1 mile are shown on the figure. From the figure we can easily recognize correlations between density measures of *count1, lcount1*, and among diversity measures of *si1, rdiv1*, and *lrdiv1*.

[Insert Figure 10 about here]

Methods

The hypothesized theoretical model in Figure 2 has a two-step two-mediator structure where the antecedent is density, the outcome is ratings, the first-step mediators is diversity, and the second-step mediator is business-customer matching. H2, H3, and H5 hypothesize direct effects of density on diversity, diversity on matching, and matching on ratings respectively. H1 and H4 hypothesize indirect effects of density on ratings, and density on matching. In terms of models, due to the simultaneity and complexity of the two-step mediation process, the major tool I use is Structural Equation Modelling (SEM) methodology.

Recall that a greater diversity in the surrounding area can possibly attract a group of more omnivorous customers, and the omnivorousness of customers can be associated with higher ratings. To control for this possibility, I add the construct of customers' omnivorousness to the structural equations.

The structural equation path analysis system contains five main constructs: agglomeration density, business diversity, business-customer matching, customers' onmivorousness, and businesses' ratings. The system consists of the following five equations:

(1) $density = \Lambda_1^d X_d + \zeta_1$;

(2) $diversity = \beta_2^e density + \Lambda_2^d X_d + \Lambda_2^c X_c + \zeta_2$;

(3) $omnivorousness = \beta_3^d diversity + \beta_3^e density + \Lambda_3^d X_d + \Lambda_3^c X_c + \zeta_3$;

(4) $matching = \beta_4^o\ omnivorousness + \beta_4^d\ diversity + \beta_4^e density + \Lambda_4^d X_d + \Lambda_4^c X_c + \zeta_4$;

(5) $ratings = \beta_5^m matching + \beta_5^o omnivorousness + \beta_5^d\ diversity + \beta_5^e density$

$$+\Lambda_5^d X_d + \Lambda_5^c X_c + \zeta_5.$$

where X_d is a geographical variable vector containing demographic variables and state fixed-effects; X_c is a control vector containing customers' differences variables, neighboring restaurant variables, and chain fixed-effects; ζs are errors.

Theoretically, I don't claim that customers' omnivorousness affects the preference matching between businesses and customers. But recall that the four measures of matching involve either category-specific ratings or/and category-specific visiting frequency, which are expected to be associated with customers' omnivorousness. Because of this association, I add omnivorousness to the right-hand-side of the fourth regression. Among the five constructs, a total of ten direct paths are estimated.

Results

This section reports results from the SEM analyses. Figure 12.a-12.d present non-standardized estimates for all the four matching measures at the radius of one mile[14], using $si1_r$ as business diversity measure, and $avgusi_r$ as customers' omnivorousness measure[15]. Table 5 and Table 6 present non-standardized coefficients and standard errors at five radius levels using $Fit3_r$ and $Fit4_r$ as the matching measure.[16]

[Insert Figure 12 about here]

[Insert Table 5-8 about here]

[14] I also estimated structural equations at other radius levels, but constructs show best correlations when radius is one mile.

[15] Recall that both variables are Simpson Indices, $si1_r$ measures the business diversity in an area and $avgusi_r$ measures customers' "diversity" in their reviewed categories.

[16] I also estimate structural models using Fit1 and Fit2 at other radius levels (0.2, 0.5, 2, and 5 miles), those models show similar results and are not shown in Figure 13. I am glad to offer full statistical analyses results upon request.

Hypothesis 2 is supported by highly significant paths from density to diversity[17] in Figure 12, suggesting that the business diversity is positively associated with agglomeration density. In areas with a high business density, businesses have more differentiated offerings, enabling customers to choose from a wider range of options. Similarly significant associations are found in models using other levels of radiuses in Table 5-8.

However, these models only offer limited support for Hypothesis 3. In Figure 12 and most models in Table 5-8, paths from diversity to goodness of matching are insignificant, meaning a more diverse surrounding environment is not associated with a better business-customer matching. However, among the twenty models, Model 6 in Table 6 and Model 11 and 12 in Table 7 show significantly positive direct paths from diversity to matching. This finding suggests that within an area of 5-minute walking distance, the higher the restaurant density, the better matching between restaurants and their customers. The radiuses of these three models are 0.2 miles and 0.5 miles, and the matching measures of these three models are $Fit2_r$ and $Fit3_r$. Given these facts, the findings seem to suggest that the matching improvement may only occur at very small geographical areas. In the restaurant context, the number of options matters only if the options are within consumers' walking distances; and consumers choose the best fit among restaurants within a short walking distance. Due to the limited support on the positive association between diversity and goodness of matching, Hypothesis 4 also lacks support when the radius is greater than 0.5 miles.

Hypothesis 5 is supported by significant positive associations between $Fit1_r$, $Fit2_r$, and $Fit4_r$ and restaurants' ratings. But, results on $Fit1_r$ and $Fit2_r$ are not surprising since these two measures use customers' category-specific ratings as proxies of their preferences, thus they have the rating part built-in. The positive association between $Fit4_r$ and ratings shows that customers who rank highly of categories that the restaurant is in give higher ratings to the restaurant. The

[17] I use R package of piecewiseSEM to estimate these structural models, see Rosseel (2012) and Lefcheck (2015) for manuals.

result on $Fit4_r$ provides more credible support for Hypothesis 5. A significant negative association is found between $Fit3_r$ and ratings, which deviates from prediction. This negative association indicates that the less concentrated the customers are on the restaurant's categories, the higher ratings the restaurant tends to receive. Similar to what have been discussed before on $Fit3_r$, it is unclear whether this negative association undermines evidence for Hypothesis 5 because the smaller concentration can reflect either a poor business-customer matching or customers' omnivorousness.

The insignificance direct path from density to ratings suggests the potential effect of density on ratings is indirect. About the path of this indirect effect, although I argued about the improved business-customer matching mechanism, it is not well supported by data of larger geographical areas. Models support that the rating premium is associated with customers' omnivorousness. In this way, Hypothesis 1 is still supported, but the path appears to be different from my original hypotheses.

Discussion

Although the analyses only lend limited support for H3 and H4, an unexpected finding is that omnivorousness plays a role in these processes, and this section is devoted to further discussion of this finding. Restaurants in denser areas do seem to attract more omnivores and restaurants' ratings are positively associated with their customers' omnivorousness. Figure 11.a shows a revised framework with the additional construct of customers' omnivorousness.

[Insert Figure 11 about here]

One argument that can undermine the hypothesized preference matching mechanism is that many customers do not have strict ordered preferences for cuisine categories. When Peterson (1992) introduced the term "omnivore" to the study of cultural consumption, he reported in the survey for musical tastes that many respondents were omnivores and were unable to pick one

category as their favorite. A similar situation may exist in the restaurant context. If it is true, continuous measures of customers' preferences on categories (such as the four measures in this paper) become inappropriate, especially for customers who are omnivores. Transforming these continuous consumers' preference measures to some discrete categorical measures may be a direction that is worth further examination.

Back to the discussion of omnivorousness, although signs of direct effects of customers' omnivorousness on restaurant ratings are inconsistent in Figure 12 (positive in Model 1-5 and Model 16-20 in Table 5 and Table 8; but negative in Model 6-15 in Table 6 and Table 7), the total effects are significantly positive[18]. A plausible alternative framework that deserves our attention is one that replaces preference matching with customers' omnivorousness as the second-step mediator of the relationship between density and ratings. Figure 11.b shows this revised theoretical framework. In this framework, the association of diversity and customers' omnivorousness has been supported by structural equations in Figure 12. In explaining the association between more omnivorous customers and higher ratings, reasons can be that omnivores enjoy changes or they are less critical. On the one hand, the novelty of new cuisine categories brings omnivores extra satisfaction; on the other hand, their relatively less experience in one category makes them less capable of judging critically.

Additional Studies

In this section, I run additional studies to check the robustness of the findings on (1) the positive association between agglomeration density in a small area and business-customer matching, and (2) the role of omnivorousness in these processes.

To check the robuseness of findings on the relationship between diversity and matching at smaller geographical areas I change the diversity measure to the number of distinct restaurant

[18] Estimated total effects of omnivorousness on ratings in the four SEMs are: 1.327*1.115+(-0.438)=1.042, (-2.242)*1.037+3.366=1.041, (-1.053)*(-0.352)+0.671=1.042, and 0.437*3.093+(-0.311)= 1.041.

categories in the surrounding area and repeat the analysis in Figure 12 at radius levels of 0.2 miles and 0.5 miles. Model 21-28 in Table 9 present the results. To test Hypothesis 3, the path that we are interested in is the one from *lrdiv* (which is the diversity measure) to *Fit* (which measures the goodness of matching). At the radius level of 0.2 miles, three of the four models show positive significant association; at 0.5 miles level, however, none of the models is significant. Results of these additional studies further suggest that the business-customer matching mechanism becomes significant when businesses are close enough such that they don't differ in terms of consumers' transportation cost.

[Insert Table 9 about here]

To check the robustness of findings on omnivorousness, I remove the matching variable and estimate structural equation models using different measures of omnivorousness and diversity. The revised system only consists of four equations:

(1) $density = \Lambda_1^d X_d + \zeta_1$;

(2) $diversity = \beta_2^e density + \Lambda_2^d X_d + \Lambda_2^c X_c + \zeta_2$;

(3) $omnivorousness = \beta_3^d diversity + \beta_3^e density + \Lambda_3^d X_d + \Lambda_3^c X_c + \zeta_3$;

(4) $ratings = \beta_4^o omnivorousness + \beta_4^d\ diversity + \beta_4^e density + \Lambda_4^d X_d + \Lambda_4^c X_c + \zeta_4$.

Non-standardized results are presented in Figure 13.

[Insert Figure 13 about here]

In Figure 13.a, I still use $si1_r$ as diversity measure and $avgusi_r$ as omnivorousness measure. In Figure 13.c and 13.d, I changed the measure of diversity to numbers of unique categories represented in the surrounding 1-mile radius area; in Figure 13.b and 13.d, I changed the measure of customers' omnivorousness to the average numbers of restaurant categories customers reviewed. All four SEM systems show significant positive associations between surrounding areas' diversity and customers' omnivorousness, and customers' omnivorousness and restaurants' ratings. The indirect path from diversity to ratings has been supported. I

discussed the possibility that more diverse options "give customers a feeling that they have more freedom and more control over their choices." However, models of Figure 13 don't find the direct association between diversity and ratings significant as none of the four direct paths are significant.

Also, I estimate structural models using variables in Figure 13.a, but at different radius levels. Paths of interest are presented in Table 10, in which *lcount*, *si*, *avgusi* are used as measures of agglomeration density, diversity, and customers' omnivorousness respectively. Model 31 in Table 10 is the exact model as in Figure 13.a. Significant paths at all radius levels lend additional support for the theoretical model using omnivorousness as a second-step mediator.

[Insert Table 10 about here]

Concluding Remarks

In this paper, I extend our understanding of agglomeration effects and study how agglomeration density affects the matching between businesses and customers, and customers' evaluations of businesses. Evidence supports that agglomeration density is positively associated with business diversity and customers' evaluations. The hypotheses that greater business diversity matches businesses with customers who prefer focal businesses more and who evaluate businesses more positively is only significant at the 0.2-mile radius level. This finding implies that consumers are sensitive to transportation cost, which should be strictly controlled for to isolate and show the improved business-customer matching. In addition, at all radius levels, I find consistent evidence that greater business diversity matches businesses with a group of more omnivorous customers. More general, this indicates that heterogeneous customer groups have different evaluation patterns. For a business establishment, the omnivorousness of its customers is positively associated with the ratings it receives from them.

In terms of methodology, this study contributes to literature by developing and comparing different measures of consumers' revealed preferences. In this study, the four

measures of customers' revealed preferences do not produce consistent statistical inferences. I use consumers' numerical ($Fit1_r$) and ordinal ($Fit4_r$) evaluations, frequencies of visiting ($Fit3_r$), and the combination of numerical evaluations and frequencies ($Fit2_r$) to infer consumers' preferences for different categories.

The study of ratings also has practical importance for businesses. Higher ratings have been shown to lead to higher revenue (Luca, 2011), this study provides evidence that locating in denser areas can benefit businesses by attracting more omnivorous customers who are also more generous in giving ratings.

Studying consumers' omnivorousness in small and large cities can be another extension. Peterson (1992) found that the proportion of omnivorous consumers is higher in higher occupational status groups. Being omnivorous is high-status individuals' way of labeling their difference. From a geographical perspective, it is interesting to examine whether the proportion of omnivorous consumers is higher in larger cities than in smaller cities. Do large cities cultivate their citizens to be more omnivorous? If consumers in large cities are more omnivorous, does this mean consumers in large cities provide more opportunities for niche businesses, although businesses may face more competition in large cities.

Limitations

In this paper, I assume that more options in general enables consumers to find their best fit and leads to more satisfaction, however, there are processes that make more options result in poorer perceptions. Studies have shown that more options may make people feel less satisfied (Iyengar, Wells, & Schwartz, 2006; Schwartz, 2004; Shah & Wolford, 2007). Combining effects of advantages and disadvantages of having more options may produce an inverted-U-shape function (Shah & Wolford, 2007). Following this logic, when density is too high, businesses' ratings may be lower as a consequence of the suffering that customers have to bear in making the

hard decision among many options. As a result, the relationship between density and ratings can also have an inverted-U-shape.

I concede that in this study, each individual's choice set is inferred by business density other than directly measured, but the issue of affective costs in decision-making is not likely to be dominant in my empirical context. Psychological studies that emphasize the affective cost of evaluating too many options during the decision-making process. However, in the setting of this paper, choosing a restaurant is not a decision as stressful as choosing job offers (Iyengar et al., 2006). Understandably, choosing a career path or making life choices can be of great significance and of great benefit/loss to an individual, and the decision-making deserves large cognitive resources. However, choices such as where to eat can hardly be that important, and the consideration process should be much less costly and shorter (otherwise, the cost of hunger rises significantly). Customers are not likely to experience an affective cost higher than the benefit of a better match. Thus, having more restaurant choices and finding a restaurant that is a better match should still bring greater satisfaction to customers.

Another issue that this study does not address is service quality. Although using chain restaurants is able to control for variations in service quality to some extent, I have not included a service quality variable. The key issue is whether service quality is correlated with density. If because of competition or other reasons, the service quality of chain restaurants in denser areas is better, then it serves as an alternative explanation to the one suggested in this study. But significant customer traffic in denser areas may also affect service quality negatively. If overall, the service quality in denser areas is lower, then arguments in this paper will be strengthened. A content analysis of each review could help to identify this potential confounding factor.

At last, data used in this study include restaurants in six American cities. Generalizing and testing the arguments in other industries such as retailing and other geographical areas can be fruitful.

References

Amiti, M., & Pissarides, C. A. 2005. Trade and industrial location with heterogeneous labor. ***Journal of International Economics***, 67(2): 392-412.

Anderson, M., & Magruder, J. 2012. Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database. ***Economic Journal***, 122(563): 957-989.

Anderson, R. E. 1973. Consumer Dissatisfaction: The Effect of Disconfirmed Expectancy on Perceived Product Performance. ***Journal of Marketing Research***, 10(1): 38-44.

Baron, R. M., & Kenny, D. A. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. ***Journal of Personality and Social Psychology***, 51(6): 1173-1182.

Baum, J. A. C., & Haveman, H. A. 1997. Love Thy Neighbor? Differentiation and Agglomeration in the Manhattan Hotel Industry, 1898-1990. ***Administrative Science Quarterly***, 42(2): 304-338.

Baum, J. A. C., & Mezias, S. J. 1992. Localized Competition and Organizational Failure in the Manhattan Hotel Industry, 1898-1990. ***Administrative Science Quarterly***, 37(4): 580-604.

Berry, S., & Waldfogel, J. 2010. Product Quality and Market Size. ***The Journal of Industrial Economics***, 58(1): 1-31.

Blau, P. M. 1970. A Formal Theory of Differentiation in Organizations. ***American Sociological Review***, 35(2): 201-218.

Carroll, G. R., & Hannan, M. T. 1989. Density Dependence in the Evolution of Populations of Newspaper Organizations. ***American Sociological Review***, 54(4): 524-541.

Chung, W., & Kalnins, A. 2001. Agglomeration effects and performance: a test of the Texas lodging industry. ***Strategic Management Journal***, 22(10): 969-988.

Combes, P.-P., & Gobillon, L. 2015. Chapter 5 - The Empirics of Agglomeration Economies. In J. V. H. Gilles Duranton, & C. S. William (Eds.), ***Handbook of Regional and Urban Economics***, Vol. Volume 5: 247-348: Elsevier.

Costa, D. L., & Kahn, M. E. 2000. Power Couples: Changes in the Locational Choice of the College Educated, 1940–1990. ***The Quarterly Journal of Economics***, 115(4): 1287-1315.

Couture, V. 2015. Valuing the Consumption Benefits of Urban Density.

d'Aspremont, C., Gabszewicz, J. J., & Thisse, J. F. 1979. On Hotelling's "Stability in Competition". ***Econometrica***, 47(5): 1145-1150.

Ellison, G., Glaeser, E. L., & Kerr, W. R. 2010. What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns. ***The American Economic Review***, 100(3): 1195-1213.

Fischer, J. H., & Harrington, J. E., Jr. 1996. Product Variety and Firm Agglomeration. ***The RAND Journal of Economics***, 27(2): 281-309.

George, L., & Waldfogel, J. 2003. Who Affects Whom in Daily Newspaper Markets? ***Journal of Political Economy***, 111(4): 765-784.

Glaeser, E. L. 1994. ***Economic growth and urban density: A review essay***: Hoover Institution on War, Revolution, and Peace, Domestic Studies Program.

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

Gotlieb, J. B., Grewal, D., & Brown, S. W. 1994. Consumer satisfaction and perceived quality: Complementary or divergent constructs? ***Journal of Applied Psychology***, 79(6): 875-885.

Gottlieb, J. D., & Shkolnik, D. 2014. Agglomeration and Quality.

Hannan, M. T., & Freeman, J. 1977. The Population Ecology of Organizations. ***American Journal of Sociology***, 82(5): 929-964.

Hanson, G. H. 2001. Scale economies and the geographic concentration of industry. ***Journal of Economic Geography***, 1(3): 255-276.

Hayes, A. F., Preacher, K. J., & Myers, T. A. 2011. Mediation and the estimation of indirect effects in political communication research. ***Sourcebook for political communication research: Methods, measures, and analytical techniques***: 434-465.

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

Iyengar, S. S., Wells, R. E., & Schwartz, B. 2006. Doing Better but Feeling Worse: Looking for the "Best" Job Undermines Satisfaction. ***Psychological Science***, 17(2): 143-150.

Jaffe, A. B., Trajtenberg, M., & Henderson, R. 1993. Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. ***The Quarterly Journal of Economics***, 108(3): 577-598.

Jin, G. Z., & Leslie, P. 2003. The effect of information on product quality: Evidence from restaurant hygiene grade cards. ***Quarterly Journal of Economics***, 118(2): 409-451.

Kovács, B., & Sharkey, A. J. 2014. The Paradox of Publicity: How Awards Can Negatively Affect the Evaluation of Quality. ***Administrative Science Quarterly***, 59(1): 1-33.

Kuksov, D. 2004. Buyer Search Costs and Endogenous Product Design. ***Marketing Science***, 23(4): 490-499.

Lefcheck, J. S. 2015. piecewiseSEM: Piecewise structural equation modeling in R for ecology, evolution, and systematics. ***arXiv preprint arXiv:1509.01845***.

Lincoln, J. R. 1979. Organizational Differentiation in Urban Communities: A Study in Organizational Ecology. ***Social Forces***, 57(3): 915-930.

Luca, M. 2011. Reviews, reputation, and revenue: The case of Yelp. com. ***Harvard Business School NOM Unit Working Paper***(12-016).

Luca, M., & Zervas, G. 2015. Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. ***Harvard Business School NOM Unit Working Paper***(14-006).

Marshall, A. 1920. Principles of economics: an introductory volume.

Mayzlin, D., Dover, Y., & Chevalier, J. 2014. Promotional Reviews: An Empirical Investigation of Online Review Manipulation. ***American Economic Review***, 104(8): 2421-2455.

McAlister, L., & Pessemier, E. 1982. Variety Seeking Behavior: An Interdisciplinary Review. ***Journal of Consumer Research***, 9(3): 311-322.

Peterson, R. A. 1992. Understanding audience segmentation: From elite and mass to omnivore and univore. ***Poetics***, 21(4): 243-258.

Reibstein, D. J., Youngblood, S. A., & Fromkin, H. L. 1975. Number of choices and perceived decision freedom as a determinant of satisfaction and consumer behavior. ***Journal of Applied Psychology***, 60(4): 434-437.

Rosenthal, S. S., & Strange, W. C. 2004. Chapter 49 Evidence on the nature and sources of agglomeration economies. In J. V. Henderson, & T. Jacques-François (Eds.), ***Handbook of Regional and Urban Economics***, Vol. Volume 4: 2119-2171: Elsevier.

Rosseel, Y. 2012. lavaan: An R package for structural equation modeling. ***Journal of Statistical Software***, 48(2): 1-36.

Schwartz, B. 2004. ***The paradox of choice: Why more is less***.

Shah, A. M., & Wolford, G. 2007. Buying Behavior as a Function of Parametric Variation of Number of Choices. ***Psychological Science***, 18(5): 369-370.

SIMPSON, E. 1949. Measurement of Diversity. ***Nature***, 163: 688-688.

Sorensen, J. B., & Sorenson, O. 2007. Corporate Demography and Income Inequality. ***American Sociological Review***, 72(5): 766-783.

Swaminathan, A., & Delacroix, J. 1991. Differentiation Within An Organizational Population: Additional Evidence from The Wine Industry. ***Academy of Management Journal***, 34(3): 679-692.

Wheeler, Christopher H. 2001. Search, Sorting, and Urban Agglomeration. ***Journal of Labor Economics***, 19(4): 879-899.

Appendix 1. Logics of Business-Customer Matching

In this section, I use a simple model (Figure 1) to illustrate how consumers with different preferences are matched with horizontally differentiated businesses and how diversity in consumers' options improves business-customer matching.

Assume that there are only two restaurants, r_1 and r_2, in a closed town. The two restaurants are horizontally differentiated and their categories are C_{r_1} and C_{r_2} respectively. For simplification, I assume that both C_{r_1} and C_{r_2} only have one category element and they are different. A customer i's preferences for C_{r_1} and C_{r_2} are $p_{iC_{r_1}}$ and $p_{iC_{r_2}}$ respectively. I assume that there is no transportation cost, prices of two restaurants are the same, and that consumers only differ in preferences for categories and restaurants only differ in categories. Categories of the two restaurants are fixed, and there are not adaptation, learning, or advertising that may change the goodness of preference matching between restaurants and customers.

A consumer's utility of dining at home is $\underline{U_i}$, and utilities of dining at r_1 and r_2 are $U_i\left(p_{iC_{r_1}}\right)$and $U_i\left(p_{iC_{r_2}}\right)$. U_i is an increasing function of preference p_i, meaning that a stronger preference always brings more satisfaction to a consumer. V_i is the inverse function of $U_i(p_i)$. Define $\underline{p_i} = V_i(\underline{U_i})$, which is the lowest preference level that a consumer will choose to dine out instead of at home. When $max\left[U_i\left(p_{iC_{r_1}}\right), U_i\left(p_{iC_{r_2}}\right)\right] > \underline{U_i}$ or equivalently, $max\left[p_{iC_{r_1}}, p_{iC_{r_2}}\right] > \underline{p_i}$, the resident will choose to dine out, and he/she will choose the restaurant with a higher preference level.

Figure 1 shows $p_{C_{r_1}}, p_{C_{r_2}}, \underline{p}$, and the distribution of residents' choices between r_1 and r_2.

The upper panel shows the scenario where there are two restaurants, and the lower panel shows the scenario where there is only r_1. $F_{r_1}^2$ and $F_{r_1}^1$ are the fit scores r_1 has with its customers when there are two and one restaurants respectively.

Assume that consumers are uniformly distributed[19] in the two-dimensional Hotelling square space $(p_{C_{r_1}}, p_{C_{r_2}})$, and $p_{C_{r_1}}, p_{C_{r_2}} \in (0, \overline{p})$, recall that a restaurant's fit score with its customers is defined as the mean of customers' preferences on the restaurant's category, then we have $F_{r_1}^2 = \sqrt{\frac{\overline{p}^2+\underline{p}^2}{2}} > \frac{\overline{p}+\underline{p}}{2} = F_{r_1}^1$. Intuitively, this is because consumers in triangle ABC "have to" choose r_1 since there is no r_2, and their preferences on r_1 are not as strong as those of residents in trapezoid ABDE. This model shows that restaurants in areas where consumers have more options have a smaller share of consumers but have better fit with their customers' preferences.

[19] In general, if there is only one restaurant r_1, its fit score with its customers is
$F_{r_1}^1 = E\left(p_{C_{r_1}} \middle| p_{C_{r_1}} > \underline{p}\right)$ when it is the only restaurant, and
$F_{r_1}^2 = E\left(p_{C_{r_1}} \middle| p_{C_{r_1}} > \underline{p} \ and \ p_{C_{r_1}} > p_{C_{r_2}}\right)$ when there are two restaurants.
When there are *n* restaurants, then r_1's fit scores with its customers will be:
$F_{r_1}^n = E\left(p_{C_{r_1}} \middle| p_{C_{r_1}} > \underline{p} \ and \ p_{C_{r_1}} > \max_{i \neq 1} p_{C_{r_i}}\right)$, which means customers has to choose to dine out and prefer r_1 over any other restaurants.

Appendix 2. Testing Mediation Effects

Baron and Kenny (1986) proposed a four-step approach for testing single-level mediation. As shown Figure A.1, X is the antecedent, Y is the outcome, and M is the mediator. In Figure A.1, if *a*, *b*, and *c* are all significant, but after controlling for *a* and *b*, *c'* becomes insignificant, then we may conclude that M totally mediate the effect of X on Y.

[Insert Figure A.1 about here]

The theoretical model in this paper follows a two-step two-mediator structure, as shown in Figure A.2. In addition to the direct path, among the three possible indirect paths from density to ratings, the one this paper focuses on is the path that progresses through diversity and then fit. Using Structural Equation Modelling (SEM), my hypotheses predict $\boldsymbol{a_1, a_3, \text{ and } b_2}$ to be significant. I argue that diversity totally mediates the effect of density on fit and don't expect a_2 to be significant. Notice that a_2 in the diagram below means the direct effect of density on fit, which is not equal to Hypothesis 4 in my argument, which follows an indirect effect logic.

I also argue that fit totally mediates the effect of diversity on ratings so that b_1 will be insignificant. However, I do expect the direct path *c'* to be significant. Recall that in addition to the consumer side explanation, the producer side explanation suggests that density enhances quality and subsequent ratings. This path will be captured by the coefficient of path *c'*. If we exclude all producer-side effects of density on ratings, path *c'* should be insignificant. Here, notice again that path *c'* in the figure indicates the effects of density on fit that are exogenous to this system, which is not H1 in my argument, which hypothesize an endogenous $a_1 \rightarrow a_3 \rightarrow b_2$ indirect path.

[Insert Figure A.2 about here]

Appendix 3. Review-level Analyses

Hypothesis 5 states at the establishment level that the level of preference the customers have should be positively associated with the ratings they give. It will be interesting to test a similar hypothesis by looking at each single review:

Hypothesis 5b: At the review level, the likelihood of a higher star rating is positively associated with the preference the reviewer has for the categories that the reviewed restaurant is in.

To test the hypothesis, I use $p3_{iC_k}$ as the measure of preference since it is the only measure whose calculation does not involve ratings. Based on the definition of $p3_{iC_k}$, Hypothesis 5b can be rewritten as:

Hypothesis 5b': At the review level, the likelihood of a higher star rating is positively associated with the proportion of reviews a reviewer wrote in categories that the reviewed restaurant belongs to.

I ran two single-equation ordered logit models for a preliminary test of Hypothesis 5b. Model 1 uses all the 983 thousand reviews of restaurants in the ten U.S. and international cities; and Model 2 does the same analysis on the 84 thousand reviews of the 4,200 chain restaurants. Table A.1 summarizes variables in these two models.

[Insert Table A.1 about here]

Maximum Likelihood estimates are reported in Table A.2.

[Insert Table A.2 about here]

Hypothesis 5b is supported; the independent variable of interest $catfreq$ ($p3_{iC_k}$) has positive coefficients in both models. This suggests that customers who have stronger preferences for categories that the reviewed restaurant belongs to are more likely to give higher star ratings. However, usi, which measures customers' omnivorousness, has negative coefficients in both

models. The result means that customers who are more omnivorous are more likely to give lower ratings, which is different what we found in establishment level analyses.

Control variables that are included are the average rating of a restaurant, the average rating of a reviewer, the total reviews of a restaurant, the total reviews of a reviewer, and the status of the reviewer. Coefficients of the average rating of a restaurant and a reviewer are positive, as expected. Coefficients of the total reviews of a restaurant and a reviewer are significant but at neglectable scale. The Yelp "Elites" tends to be more critical in giving ratings. The results strengthen the finding of the positive association between customers' preference and ratings, although further research needs to be done (perhaps by including more control variables) to examine in the inconsistent result on omnivorousness.

Table 1. Yelp Restaurant categories in the U.S.

Afghan	Chech	Indonesian	Polish
African	Cheesesteaks	Irish	Portuguese
Senegalese	Chicken Shop	Italian	Poutineries
South African	Chicken Wings	*Calabrian*	Russian
American (Old)	Chinese	*Sardinian*	Salad
American (New)	*Cantonese*	*Tuscan*	Sandwiches
Arabian	*Dim Sum*	Japanese	Scandinavian
Argentine	*Shanghainese*	*Conveyor Belt Sushi*	Scottish
Armenian	*Szechuan*	*Kushikatsu*	Seafood
Asian Fusion	Creperies	*Ramen*	Singaporean
Australian	Cuban	*Teppanyaki*	Slovakian
Austrian	Czech	Korean	Soul Food
Bangladeshi	Delis	Kosher	Soup
Barbeque	Diners	Laotian	Southern
Basque	Ethiopian	Latin American	Spanish
Beer Hall	Fast Food	*Colombian*	Sri Lankan
Belgian	Filipino	*Salvadoran*	Steakhouses
Brasseries	Fish & Chips	*Venezuelan*	Supper Clubs
Brazilian	Fondue	Live/Raw Food	Sushi Bars
Breakfast & Brunch	Food Court	Malaysian	Taiwanese
British	Food Stands	Meatballs	Tapas Bars
Buffets	French	Mideterranean	Tapas/Small Plates
Burgers	Gastropubs	*Falafel*	Tex-Mex
Burmese	German	Mexican	Thai
Cafes	Gluten-Free	Middle Eastern	Turkish
Cafeteria	Greek	*Egyptian*	Ukrainian
Cajun/Creole	Halal	*Lebanese*	Uzbek
Cambodian	Hawaiian	Modern European	Vegan
Caribbean	Himalayan/Nepalese	Mongolian	Vegetarian
Dominican	Hot Dogs	Moroccan	Vietnamese
Haitian	Hot Pot	Pakistani	**Total: 133**
Puerto Rican	Hungarian	Persian/Iranian	
Trinidadian	Iberian	Peruvian	
Catalan	Indian	Pizza	

Note: Categories in *Italian* are subcategories under the category in front of them. Category names are from: https://www.yelp.com/developers/documentation/v2/all_category_list

Table 2. Yelp data entries by metropolitan area

MSA	Population (2014 estimate)	Population Rank	Number of Restaurants	Number of Reviewers	Number of Reviews	Date of First Review
Phoenix, AZ	4.5 million	12	6,899	92,770	355,763	2005-02-01
Charlotte, NC-SC	2.4 million	22	1,835	19,676	61,238	2004-12-19
Pittsburgh, PA	2.4 million	23	1,156	14,465	42,213	2005-05-03
Las Vegas, NV	2.1 million	30	4,293	129,012	382,838	2004-10-26
Madison, WI	0.64 million	86	856	9,340	29,576	2005-03-03
Champaign-Urbana	0.24 million	191	223	3,028	7,719	2004-10-12

Note: Data from Yelp (http://www.yelp.com/dataset_challenge)

Table 3. Data samples

Sample raw data on a business establishment:

Business ID	Name	Stars	Review Count	Categories
1vK7gWQ_b5ehAyOidOsYtg	Aiello's Pizza	4	115	['Italian', 'Restaurants']
Longitude	Latitude	City	State	
-79.9232555	40.4332453	Pittsburgh	PA	

Sample raw data on a user:

User ID	Review Count	Average Stars	Elite	Yelping since
--4fX3LBeXoE88gDTK6TKQ	23	4.31	2012, 2013	2012-03

Sample raw data on a review:

Business ID	User ID	Stars	Text	Review Date
bcBMAa0UQpNLFvvdZ4dxtQ	--0HEXd4W6bJI8k7E0RxTA	5	Great food!	7/13/2013

Note: Data from Yelp (http://www.yelp.com/dataset_challenge)

Table 4. Summary Statistics

Usage	Statistic	N	Mean	St. Dev.	Min	Max	Meaning
Rating measure	avg	4,200	3.037	0.786	1	5	Average star rating
Popularity	lrc	4,200	2.492	1.006	1.099	6.33	Log of review count of a restaurant
Neighbors' popularity	lrc02	4,200	3.414	0.836	1.099	6.238	Log of average review counts of neighboring restaurants within 0.2 miles
	lrc05	4,200	3.56	0.75	1.099	5.924	Log of average review counts of neighboring restaurants within 0.5 miles
	lrc1	4,200	3.665	0.672	1.099	5.687	Log of average review counts of neighboring restaurants within 1 mile
	lrc2	4,200	3.77	0.601	1.386	5.447	Log of average review counts of neighboring restaurants within 2 miles
	lrc5	4,200	3.881	0.525	1.386	5.03	Log of average review counts of neighboring restaurants within 5 miles
Density measure	count02	4,200	12.841	13.898	1	150	Number of neighboring restaurants within 0.2 miles
	count05	4,200	29.592	35.271	1	242	Number of neighboring restaurants within 0.5 miles
	count1	4,200	59.723	66.344	1	533	Number of neighboring restaurants within 1 mile
	count2	4,200	163.17	167.24	1	1,081	Number of neighboring restaurants within 2 miles
	count5	4,200	726.26	569.03	1	2,454	Number of neighboring restaurants within 5 miles
	lcount02	4,200	2.178	0.879	0	5.011	Log of count02
	lcount05	4,200	2.935	0.967	0	5.489	Log of count05
	lcount1	4,200	3.664	0.936	0	6.279	Log of count1
	lcount2	4,200	4.697	0.925	0	6.986	Log of count2
	lcount5	4,200	6.258	0.905	0	7.805	Log of count5
Neighbors' quality measure	avg02	4,200	3.278	0.377	1	4.668	Average star rating of neighboring restaurants within 0.2 miles
	avg05	4,200	3.334	0.285	1	4.533	Average star rating of neighboring restaurants within 0.5 miles
	avg1	4,200	3.372	0.208	1.6	4.102	Average star rating of neighboring restaurants within 1 mile
	avg2	4,200	3.401	0.144	2.6	4.143	Average star rating of neighboring restaurants within 2 miles
	avg5	4,200	3.416	0.097	2.985	3.964	Average star rating of neighboring restaurants within 5 miles
Diversity measure	si02	4,200	0.848	0.154	0	0.969	Simpson diversity index of neighboring restaurants within 0.2 miles
	si05	4,200	0.9	0.102	0	0.968	Simpson diversity index of neighboring restaurants within 0.5 miles
	si1	4,200	0.93	0.059	0	0.973	Simpson diversity index of neighboring restaurants within 1 mile
	si2	4,200	0.948	0.029	0	0.972	Simpson diversity index of neighboring restaurants within 2 miles
	si5	4,200	0.958	0.018	0	0.971	Simpson diversity index of neighboring restaurants within 5 miles
	rdiv02	4,200	11.452	7.065	1	47	Number of unique categories of neighboring restaurants within 0.2 miles
	rdiv05	4,200	18.015	10.082	1	60	Number of unique categories of neighboring restaurants within 0.5 miles
	rdiv1	4,200	25.443	12.32	1	67	Number of unique categories of neighboring restaurants within 1 mile
	rdiv2	4,200	38.259	15.582	1	79	Number of unique categories of neighboring restaurants within 2 miles
	rdiv5	4,200	60.275	17.845	1	91	Number of unique categories of neighboring restaurants within 5 miles

Table 4. Summary Statistics (Continued)

Usage	Statistic	N	Mean	St. Dev.	Min	Max	Meaning
Diversity measure (continued)	lrdiv02	4,200	2.222	0.719	0	3.85	Log of rdiv02
	lrdiv05	4,200	2.702	0.679	0	4.094	Log of rdiv05
	lrdiv1	4,200	3.097	0.575	0	4.205	Log of rdiv1
	lrdiv2	4,200	3.547	0.476	0	4.369	Log of rdiv2
	lrdiv5	4,200	4.043	0.365	0	4.511	Log of rdiv5
Chain controls	chmean	4,200	3.035	0.457	1.928	4.178	Mean average star ratings of a chain
	chaincou	4,200	86.495	85.408	10	295	Mean review count of a chain
	lchaincou	4,200	4.003	0.974	2.303	5.687	Log of chaincou
Demographic controls	whishr	4,200	0.671	0.148	0.138	0.964	Proportion of Whites, ZCTA level
	hisshr	4,200	0.224	0.162	0.01	0.822	Proportion of Hispanics, ZCTA level
	sexr	4,200	0.996	0.113	0.711	3.08	Male population/Female population, ZCTA level
	lpopden	4,200	7.946	0.968	2.494	9.694	Log of population density, ZCTA level
	lmhhi	4,200	10.868	0.35	9.879	11.743	Log of median household income, ZCTA level
	HS	4,200	23.921	7.057	5.1	42.1	Proportion of population with high school degree, ZCTA level
	Bach	4,200	20.141	8.784	2.5	48.4	Proportion of population with Bachelor's degree, ZCTA level
	Gra	4,200	11.234	6.934	0	45	Proportion of population with graduate or professional degree, ZCTA level
Fit measure	fit1	4,200	3.189	0.618	1	5	Fit score of a restaurant with its customers, measure 1
	fit2	4,200	0.797	0.413	0.028	4.5	Fit score of a restaurant with its customers, measure 2
	fit3	4,200	0.302	0.130	0.025	1	Fit score of a restaurant with its customers, measure 3
	fit4	4,200	0.412	0.132	0.000	0.889	Fit score of a restaurant with its customers, measure 4
Customers' characteristics	avgurc_r	4,200	81.197	71.94	1	734	Average number of restaurant reviews written by a customer
	avgusi	4,200	0.791	0.104	0	0.955	Average Simpson Index of reviewers of a restaurant
	avgucc	4,200	17.655	6.956	1	55.5	Average number of reviewed categories for reviewers of a restaurant
	laucc	4,200	2.783	0.451	0	4.016	Log of avgucc
	vi	4,200	0.646	0.153	0.025	1	Average visitor's index of customers

Note: (1) Demographic data is from American Community Survey (2008-2013).
(2) Location-related variables are calculated from latitudes and longitudes offered by Yelp.
(3) Business, customer, review data are from Yelp.

Table 5. Structural Equation Path Analysis results using Fit1 as the matching measure at all radius levels

Path	*Radius (mile(s))*				
	0.2	0.5	1	2	5
	(1)	(2)	(3)	(4)	(5)
†H2: lcount(r) -> si(r)	0.116***	0.070***	0.040***	0.023***	0.011***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
lcount(r) -> avgusi	-0.005**	-0.001	0.001	0.000	0.010**
	(0.003)	(0.002)	(0.003)	(0.003)	(0.004)
lcount(r) -> Fit1	0.006	0.018	0.016	0.007	-0.003
	(0.014)	(0.013)	(0.015)	(0.019)	(0.024)
lcount(r) -> avg	0.003	0.000	0.008	0.012	0.021
	(0.009)	(0.008)	(0.009)	(0.012)	(0.015)
†si(r) -> avgusi	0.050***	0.049***	0.093***	0.198***	0.120
	(0.014)	(0.019)	(0.032)	(0.070)	(0.095)
†H3: si(r) -> Fit1	-0.117	-0.133	-0.044	-0.523	-0.259
	(0.073)	(0.106)	(0.180)	(0.397)	(0.544)
si(r) -> avg	-0.046	-0.039	-0.111	0.042	0.005
	(0.047)	(0.068)	(0.113)	(0.246)	(0.337)
avgusi -> Fit1	1.206***	1.227***	1.327***	1.355***	1.364***
	(0.084)	(0.087)	(0.088)	(0.089)	(0.089)
†avgusi -> avg	-0.433***	-0.452***	-0.438***	-0.443***	-0.446***
	(0.056)	(0.056)	(0.057)	(0.057)	(0.057)
†H5: Fit1 -> avg	1.074***	1.101***	1.115***	1.121***	1.124***
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
Demographic Controls	Yes	Yes	Yes	Yes	Yes
Neighbor Quality Controls	Yes	Yes	Yes	Yes	Yes
Customers' Differences Controls	Yes	Yes	Yes	Yes	Yes
State Effects	Yes	Yes	Yes	Yes	Yes
Chain Effects	Yes	Yes	Yes	Yes	Yes
Observations	4,200	4,200	4,200	4,200	4,200

Note: *p<0.1; **p<0.05; ***p<0.01

(1) Structures of all models are the same as the model in Figure 12.a, with different radius levels. Model (3) is the exact model in Figure 12.a.
(2) Among the five main constructs, a total of 10 direct paths are shown.
(3) Fit1 uses each customer's ratings for categories as the foundation of preference calculation.
(4) a -> b indicates a direct path from variable a to variable b.
(5) † paths of interest.

Table 6. Structural Equation Path Analysis results using Fit2 as the matching measure at all radius levels

Path	*Radius (mile(s))*				
	0.2	0.5	1	2	5
	(6)	(7)	(8)	(9)	(10)
†H2: lcount(r) -> si(r)	0.116***	0.070***	0.040***	0.023***	0.011***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
lcount(r) -> avgusi	-0.005**	-0.001	0.001	0.000	0.010**
	(0.003)	(0.002)	(0.003)	(0.003)	(0.004)
lcount(r) -> Fit2	0.001	0.014	0.026	0.026	0.015
	(0.009)	(0.008)	(0.009)	(0.012)	(0.015)
lcount(r) -> avg	0.008	0.006	-0.001	-0.008	0.001
	(0.015)	(0.015)	(0.016)	(0.021)	(0.026)
†si(r) -> avgusi	0.050***	0.049***	0.093***	0.198***	0.120
	(0.014)	(0.019)	(0.032)	(0.070)	(0.095)
†H3: si(r) -> Fit2	0.084*	0.092	0.120	-0.321	-0.265
	(0.047)	(0.067)	(0.112)	(0.245)	(0.335)
si(r) -> avg	-0.249***	-0.276**	-0.285	-0.208	-0.006
	(0.081)	(0.118)	(0.199)	(0.439)	(0.602)
avgusi -> Fit2	-2.288***	-2.280***	-2.242***	-2.226***	-2.222***
	(0.054)	(0.054)	(0.055)	(0.055)	(0.055)
†avgusi -> avg	2.971***	3.149***	3.366***	3.416***	3.438***
	(0.111)	(0.115)	(0.116)	(0.117)	(0.117)
†H5: Fit2 -> avg	0.921***	0.987***	1.037***	1.052***	1.058***
	(0.027)	(0.028)	(0.028)	(0.028)	(0.028)
Demographic Controls	Yes	Yes	Yes	Yes	Yes
Neighbor Quality Controls	Yes	Yes	Yes	Yes	Yes
Customers' Differences Controls	Yes	Yes	Yes	Yes	Yes
State Effects	Yes	Yes	Yes	Yes	Yes
Chain Effects	Yes	Yes	Yes	Yes	Yes
Observations	4,200	4,200	4,200	4,200	4,200

Note: *p<0.1; **p<0.05; ***p<0.01

(1) Structures of all models are the same as the model in Figure 12.b, with different radius levels. Model (8) is the exact model in Figure 12.b.
(2) Among the five main constructs, a total of 10 direct paths are shown.
(3) Fit2 uses each customer's ratings for categories weighted by frequencies of visiting as the foundation of calculation.
(4) a -> b indicates a direct path from variable a to variable b.
(5) † paths of interest.

Table 7. Structural Equation Path Analysis results using Fit3 as the matching measure at all radius levels

Path	*Radius (mile(s))*				
	0.2	0.5	1	2	5
	(11)	(12)	(13)	(14)	(15)
†H2: lcount(r) -> si(r)	0.116***	0.070***	0.040***	0.023***	0.011***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
lcount(r) -> avgusi	-0.005**	-0.001	0.001	0.000	0.010**
	(0.003)	(0.002)	(0.003)	(0.003)	(0.004)
lcount(r) -> Fit3	-0.002	0.002	0.004*	0.007***	0.008*
	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)
lcount(r) -> avg	0.009	0.021	0.027	0.022	0.021
	(0.017)	(0.017)	(0.019)	(0.025)	(0.031)
†si(r) -> avgusi	0.050***	0.049***	0.093***	0.198***	0.120
	(0.014)	(0.019)	(0.032)	(0.070)	(0.095)
†H3: si(r) -> Fit3	0.034***	0.024*	0.030	-0.047	-0.061
	(0.010)	(0.014)	(0.024)	(0.052)	(0.071)
si(r) -> avg	-0.162*	-0.176	-0.150	-0.562	-0.310
	(0.092)	(0.135)	(0.230)	(0.509)	(0.698)
avgusi -> Fit3	-1.053***	-1.053***	-1.053***	-1.052***	-1.053***
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
†avgusi -> avg	0.566***	0.512***	0.671***	0.682***	0.671***
	(0.181)	(0.190)	(0.195)	(0.197)	(0.198)
†H5: Fit3 -> avg	-0.283**	-0.368**	-0.352**	-0.374**	-0.395***
	(0.140)	(0.147)	(0.151)	(0.153)	(0.154)
Demographic Controls	Yes	Yes	Yes	Yes	Yes
Neighbor Quality Controls	Yes	Yes	Yes	Yes	Yes
Customers' Differences Controls	Yes	Yes	Yes	Yes	Yes
State Effects	Yes	Yes	Yes	Yes	Yes
Chain Effects	Yes	Yes	Yes	Yes	Yes
Observations	4,200	4,200	4,200	4,200	4,200

Note: *p<0.1; **p<0.05; ***p<0.01

(1) Structures of all models are the same as the model in Figure 12.c, with different radius levels. Model (13) is the exact model in Figure 12.c.
(2) Among the five main constructs, a total of 10 direct paths are shown.
(3) Fit3 uses each customer's frequencies of visiting as the foundation of preference calculation.
(4) a -> b indicates a direct path from variable a to variable b.
(5) † paths of interest.

Table 8. Structural Equation Path Analysis results using Fit4 as the matching measure at all radius levels

Path	*Radius (mile(s))*				
	0.2	0.5	1	2	5
	(16)	(17)	(18)	(19)	(20)
†H2:lcount(r) ->si(r)	0.116***	0.070***	0.040***	0.023***	0.011***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
lcount(r) ->avgusi	-0.005**	-0.001	0.001	0.000	0.010**
	(0.003)	(0.002)	(0.003)	(0.003)	(0.004)
lcount(r) -> Fit4	-0.002	0.000	0.003	0.004	-0.002
	(0.003)	(0.003)	(0.003)	(0.004)	(0.005)
lcount(r) ->avg	0.015	0.019	0.017	0.007	0.024
	(0.015)	(0.015)	(0.016)	(0.021)	(0.026)
†si(r) ->avgusi	0.050***	0.049***	0.093***	0.198***	0.120
	(0.014)	(0.019)	(0.032)	(0.070)	(0.095)
†H3:si(r) -> Fit4	0.011	-0.001	-0.008	-0.144	0.032
	(0.016)	(0.023)	(0.038)	(0.084)	(0.114)
si(r) ->avg	-0.203**	-0.180	-0.136	-0.090	-0.388
	(0.080)	(0.117)	(0.198)	(0.435)	(0.597)
avgusi-> Fit4	0.422***	0.426***	0.437***	0.442***	0.441***
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
†avgusi->avg	-0.312***	-0.364***	-0.311***	-0.315***	-0.311***
	(0.097)	(0.101)	(0.103)	(0.104)	(0.105)
†H5: Fit4 ->avg	2.786***	2.967***	3.093***	3.151***	3.168***
	(0.078)	(0.080)	(0.081)	(0.082)	(0.082)
Demographic Controls	Yes	Yes	Yes	Yes	Yes
Neighbor Quality Controls	Yes	Yes	Yes	Yes	Yes
Customers' Differences Controls	Yes	Yes	Yes	Yes	Yes
State Effects	Yes	Yes	Yes	Yes	Yes
Chain Effects	Yes	Yes	Yes	Yes	Yes
Observations	4,200	4,200	4,200	4,200	4,200

Note: *p<0.1; **p<0.05; ***p<0.01

(1) Structures of all models are the same as the model in Figure 12.d, with different radius levels. Model (18) is the exact model in Figure 12.d.
(2) Among the five main constructs, a total of 10 direct paths are shown.
(3) Fit4 uses each customer's ranking of cuisine categories as the foundation of calculation.
(4) a -> b indicates a direct path from variable a to variable b.
(5) † paths of interest.

Table 9. Structural Equation Path Analysis results using lrdiv as the diversity measure at radius levels of 0.2 and 0.5 miles

Path	*Radius (miles)*							
	0.2				*0.5*			
	Fit1	Fit2	Fit3	Fit4	Fit1	Fit2	Fit3	Fit4
	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)
†H2:lcount(r) ->lrdiv(r)	0.717***	0.717***	0.717***	0.717***	0.628***	0.628***	0.628***	0.628***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
lcount(r) ->avgusi	-0.018***	-0.018***	-0.018***	-0.018***	-0.015***	-0.015***	-0.015***	-0.015***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
lcount(r) -> Fit	0.008	-0.018	-0.011***	-0.009*	0.035	0.01	-0.003	0.001
	(0.025)	(0.016)	(0.003)	(0.005)	(0.026)	(0.016)	(0.003)	(0.006)
lcount(r) ->avg	0.021	0.046*	0.026	0.056**	0.002	0.029	0.039	0.037
	(0.016)	(0.027)	(0.031)	(0.027)	(0.016)	(0.028)	(0.033)	(0.028)
lrdiv(r) ->avgusi	0.026***	0.026***	0.026***	0.026***	0.028***	0.028***	0.028***	0.028***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)
†H3:lrdiv(r) -> Fit	-0.022	0.04**	0.019***	0.012*	-0.04	0.016	0.01	-0.001
	(0.031)	(0.020)	(0.004)	(0.007)	(0.037)	(0.023)	(0.005)	(0.008)
lrdiv(r) ->avg	-0.032	-0.093***	-0.051	-0.09***	-0.008	-0.068*	-0.048	-0.05
	(0.020)	(0.034)	(0.039)	(0.034)	(0.023)	(0.041)	(0.047)	(0.040)
avgusi-> Fit	1.203***	-2.289***	-1.054***	0.421***	1.229***	-2.28***	-1.054***	0.426***
	(0.084)	(0.054)	(0.012)	(0.019)	(0.087)	(0.055)	(0.012)	(0.019)
avgusi->avg	-0.43***	2.971***	0.564***	-0.31***	-0.452***	3.149***	0.514***	-0.362***
	(0.056)	(0.111)	(0.182)	(0.097)	(0.057)	(0.115)	(0.190)	(0.101)
†H5: Fit ->avg	1.075***	0.921***	-0.283**	2.79***	1.101***	0.986***	-0.368**	2.967***
	(0.010)	(0.027)	(0.140)	(0.078)	(0.010)	(0.028)	(0.147)	(0.080)
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Neighbor Quality Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Customers' Differences Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Chain Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,200	4,200	4,200	4,200	4,200	4,200	4,200	4,200

Note: *p<0.1; **p<0.05; ***p<0.01

(1) Structures of all models are the same as the model in Figure 12.
(2) Among the five main constructs, a total of 10 direct paths are shown.
(3) a -> b indicates a direct path from variable a to variable b.
(4) † paths of interest.

Table 10. Structural Equation Path Analysis results without business-customer matching construct at all radius levels

Path	*Radius (mile(s))*				
	0.2	0.5	1	2	5
	(29)	(30)	(31)	(32)	(33)
†H2: lcount(r) -> si(r)	0.118***	0.070***	0.040***	0.023***	0.011***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
lcount(r) -> avgusi	-0.005**	-0.001	0.001	0.000	-0.010**
	(0.003)	(0.002)	(0.003)	(0.003)	(0.004)
lcount(r) -> avg	0.009	0.020	0.026	0.019	0.017
	(0.017)	(0.017)	(0.019)	(0.025)	(0.031)
† si(r) -> avgusi	0.050***	0.049***	0.093***	0.198***	0.120
	(0.014)	(0.019)	(0.032)	(0.070)	(0.095)
si(r) -> avg	-0.0171*	-0.185	-0.160	-0.545	-0.286
	(0.092)	(0.135)	(0.230)	(0.509)	(0.699)
† avgusi -> avg	0.863***	0.899***	1.041***	1.076***	1.087***
	(0.105)	(0.110)	(0.113)	(0.114)	(0.115)
Demographic Controls	Yes	Yes	Yes	Yes	Yes
Neighbor Quality Controls	Yes	Yes	Yes	Yes	Yes
Customers' Differences Controls	Yes	Yes	Yes	Yes	Yes
State Effects	Yes	Yes	Yes	Yes	Yes
Chain Effects	Yes	Yes	Yes	Yes	Yes
Observations	4,200	4,200	4,200	4,200	4,200

Note: *p<0.1; **p<0.05; ***p<0.01

(1) Structures of all models are the same as the model in Figure 13.a, with different radius levels. Model (31) is the exact model in Figure 13.a.
(2) Among the four main constructs, a total of 6 direct paths are shown.
(3) a -> b indicates a direct path from variable a to variable b.
(4) † paths of interest.

Table A.1. Review-level summary statistics

Usage	Statistic	N	Mean	St. Dev.	Min	Max	Meaning
Dependent variable	stars	982,554	3.722	1.263	1	5	Star rating of this review
Restaurant and customer controls	avguser	982,554	3.743	0.774	1.000	5.000	Average star rating given by the reviewer
	avgbus	982,554	3.722	0.526	1.000	5.000	Average star rating of the restaurant
	userrc	982,554	130.046	261.063	0	4,573	Number of reviews of the reviewer
	rstrrc	982,554	349.097	625.264	3	4,578	Number of reviews of the restaurant
Customers' characteristics	usi	982,554	0.743	0.265	0.000	0.969	The reviewer's Simpson Index
	catfreq	982,554	0.347	0.321	0.001	1.000	How many of the reviewer's reviews are in this restaurant' categories ($p3_{ic_k}$)
	elite	982,554	0.271	0.444	0	1	A 0-1 dummy, equals 1 if the review is an "Elite"

Table A.1.a. Summary Statistics for reviews of all restaurants

Usage	Statistic	N	Mean	St. Dev.	Min	Max	Meaning
Dependent variable	stars	84,109	3.313	1.384	1	5	Star rating of this review
Restaurant and customer controls	avguser	84,109	3.582	0.798	1.000	5.000	Average star rating given by the reviewer
	avgbus	84,109	3.313	0.627	1.000	5.000	Average star rating of the restaurant
	userrc	84,109	141.365	282.987	1	4,534	Number of reviews of the reviewer
	rstrrc	84,109	73.184	87.187	3	561	Number of reviews of the restaurant
Customers' characteristics	usi	84,109	0.795	0.219	0.000	0.968	The reviewer's Simpson Index
	catfreq	84,109	0.295	0.290	0.001	1.000	How many of the reviewer's reviews are in this restaurant' categories ($p3_{ic_k}$)
	elite	84,109	0.266	0.442	0	1	A 0-1 dummy, equals 1 if the review is an "Elite"

Table A.1.b. Summary Statistics for reviews of chain restaurants

Table A.2. Review-level Ordered Logit Model results

	Dependent variable:	
	stars	
	(26)	(27)
†catfreq	0.209***	0.473***
	(0.014)	(0.049)
avgbus	1.392***	1.320***
	(0.004)	(0.012)
avguser	1.899***	1.820***
	(0.003)	(0.012)
rstrrc	-0.00001***	-0.0002**
	(0.00000)	(0.0001)
usi	-0.205***	-0.182***
	(0.017)	(0.065)
userrc	0.00003***	0.0001***
	(0.00001)	(0.00002)
elite	-0.157***	-0.048***
	(0.005)	(0.017)
Observations	982,554	84,109
Log Likelihood	-1,144,653.000	-106,392.300
Note:	*p<0.1; **p<0.05; ***p<0.01 † variable of interest	

Figure 1. Matching between businesses and customers

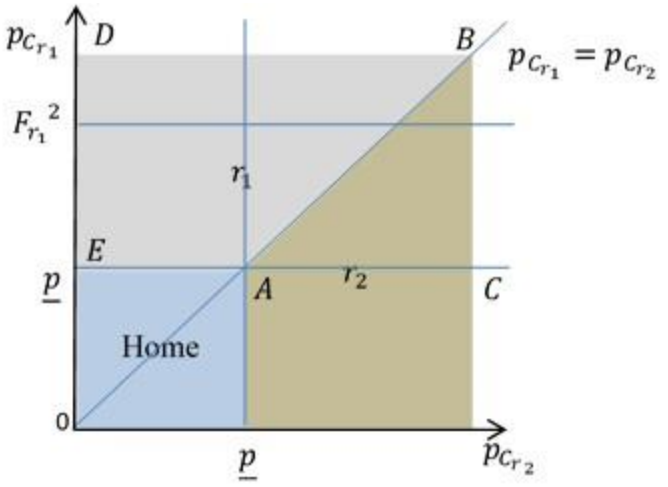


Figure 1.a. Consumers have three options (r_1, r_2, or neither of them).

Note: $p_{C_{r1}}$ and $p_{C_{r2}}$ are customers' preference for categories of r_1 and r_2. When there are two options, consumers whose preferences for either restaurant are greater than $\underline{p}$ choose to dine out. Among them, consumers who prefer r_1 than r_2 choose r_1. *ABDE* are consumers who choose r_1, and their average level of preference (fit level) for r_1 is $F_{r_1}^2$.

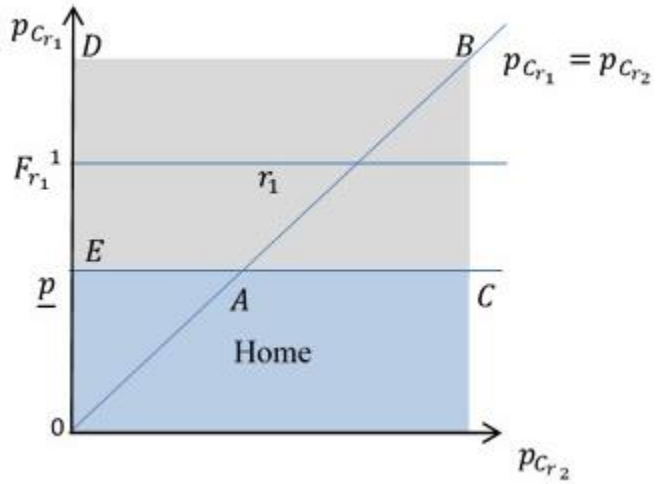
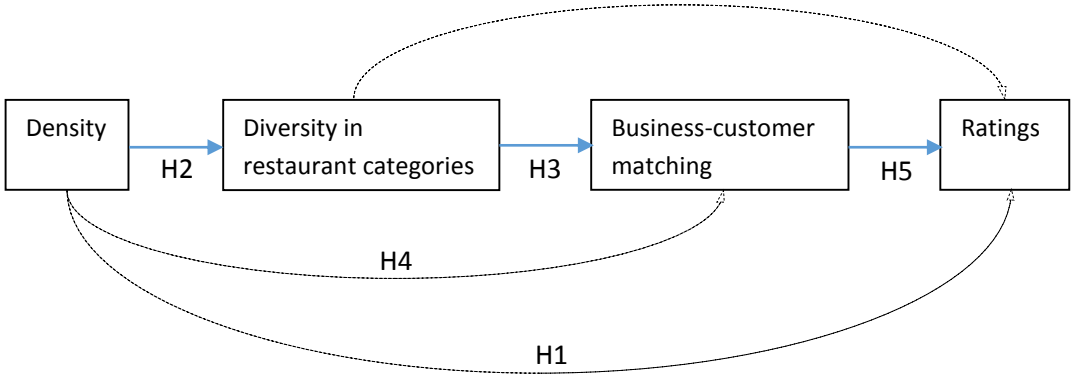


Figure 1.b. Consumers have two options (r_1 or nothing).

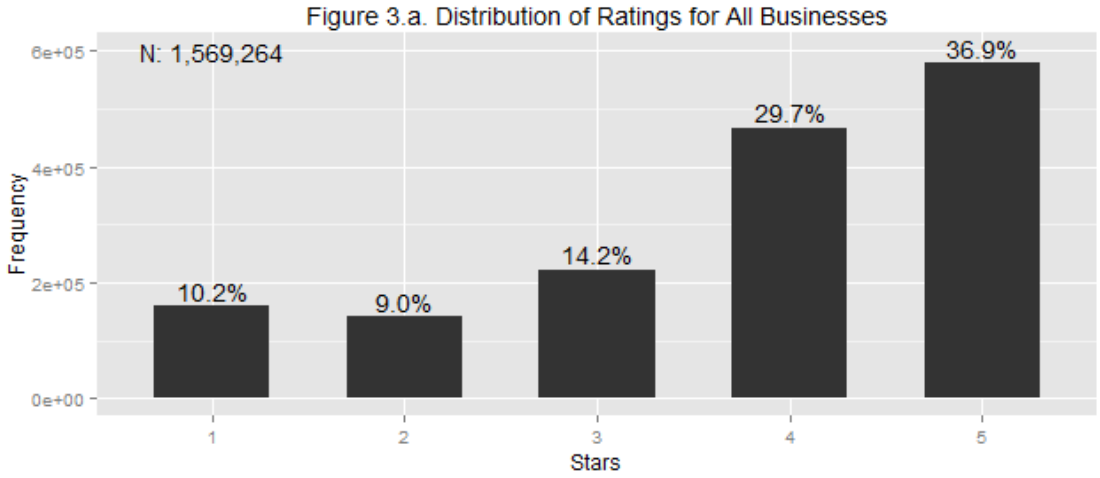
Note: When there is only one option, consumers whose preferences for $r_1$are greater than $\underline{p}$ dine out, and they have to choose r_1. *CBDE* are consumers who choose r_1. Compared with ***Figure 1.a.,*** consumers of *ABC*, who would have chosen r_2, have to become r_1's customers. The average level of preference (fit level) for r_1 when it is only one option is $F_{r_1}^1$, which is lower than the average fit level when there are two options $F_{r_1}^2$ if assuming that customers are uniformly distributed. This means when consumers have more options, businesses are better matched with its customers.

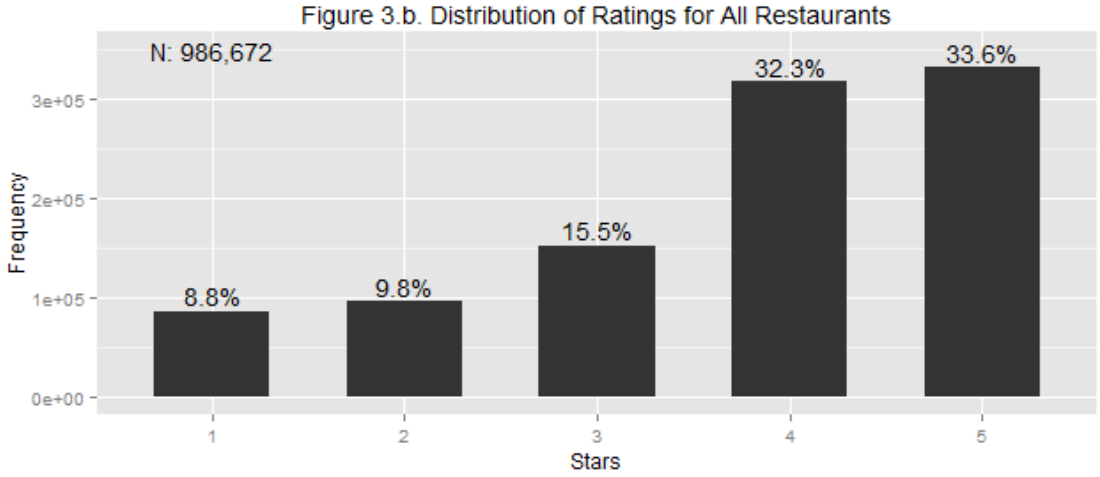
Figure 2. A two-step two-mediator theoretical framework



Note: Solid straight arrows indicate direct effects, dashed curved arrows indicate indirect effects. Hypotheses in the paper are shown.

Figure 3. Distributions of star ratings for all businesses, all restaurants, and chain restaurants





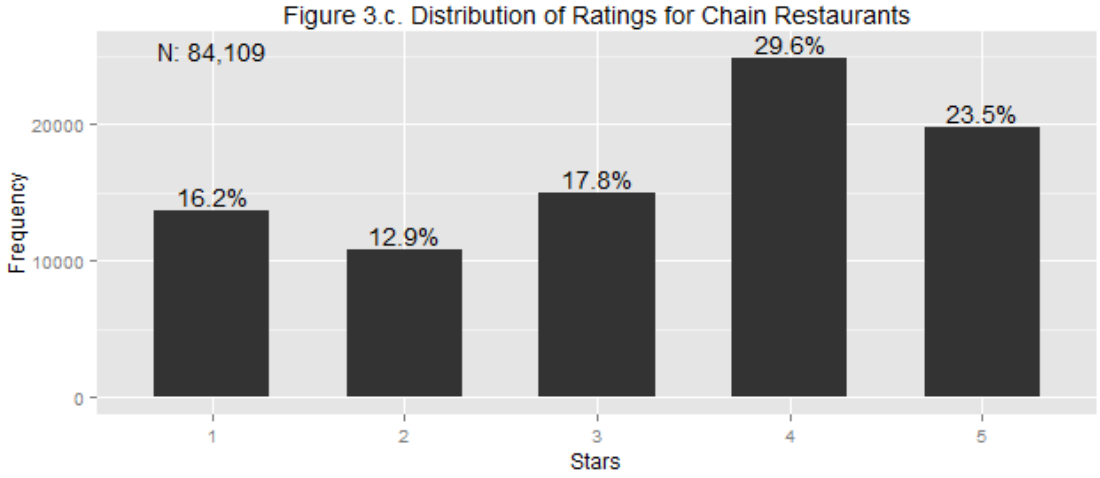
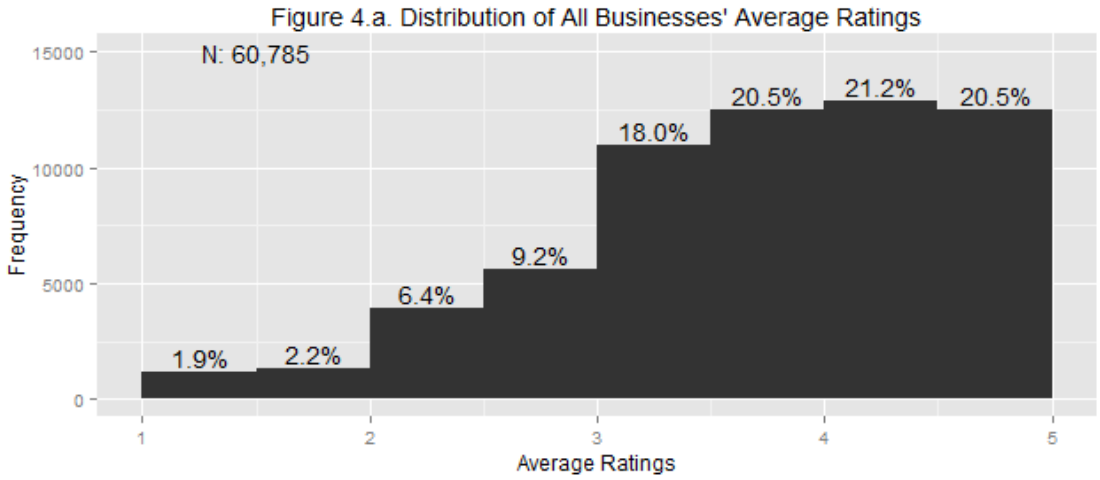
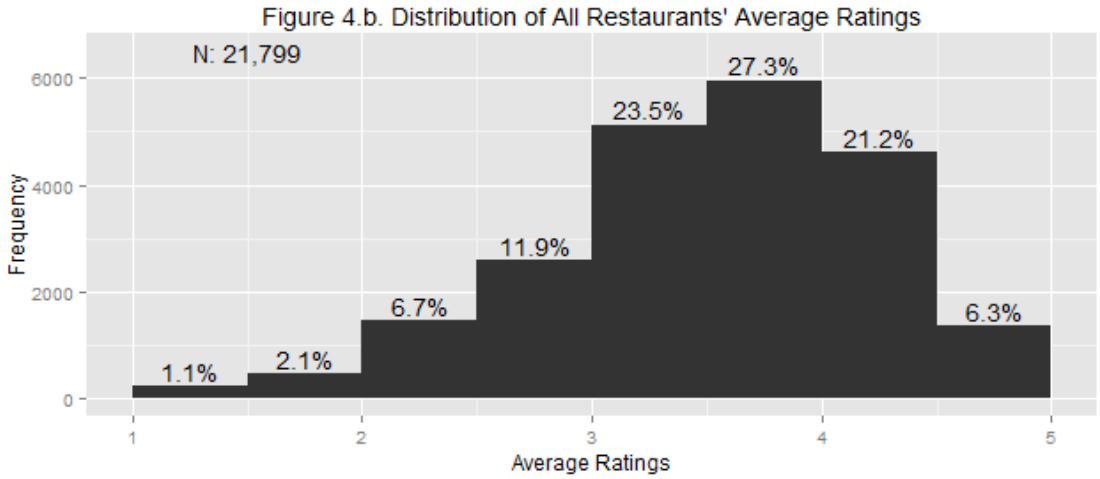


Figure 4. Distribution of restaurants' and other businesses' average ratings





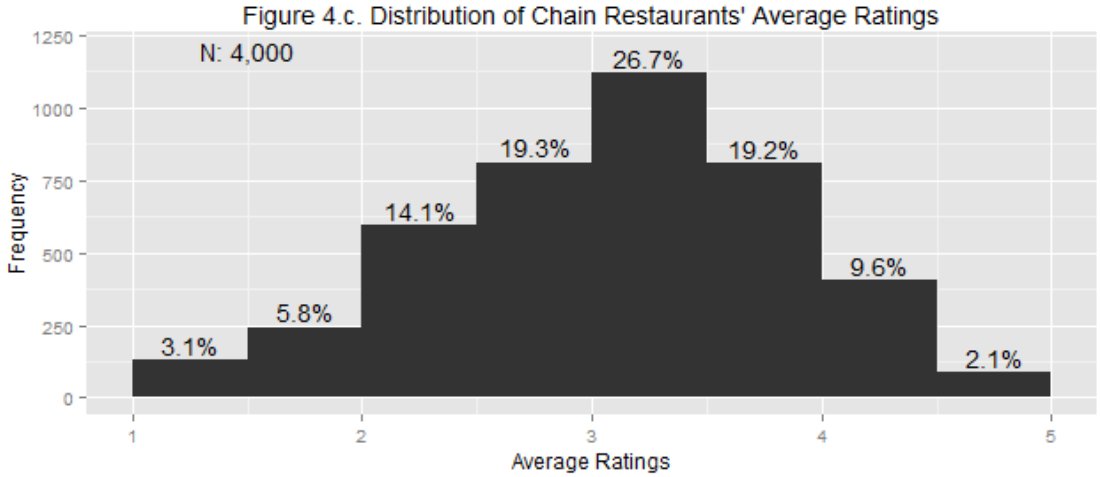


Figure 5. Average ratings of the 120 chains

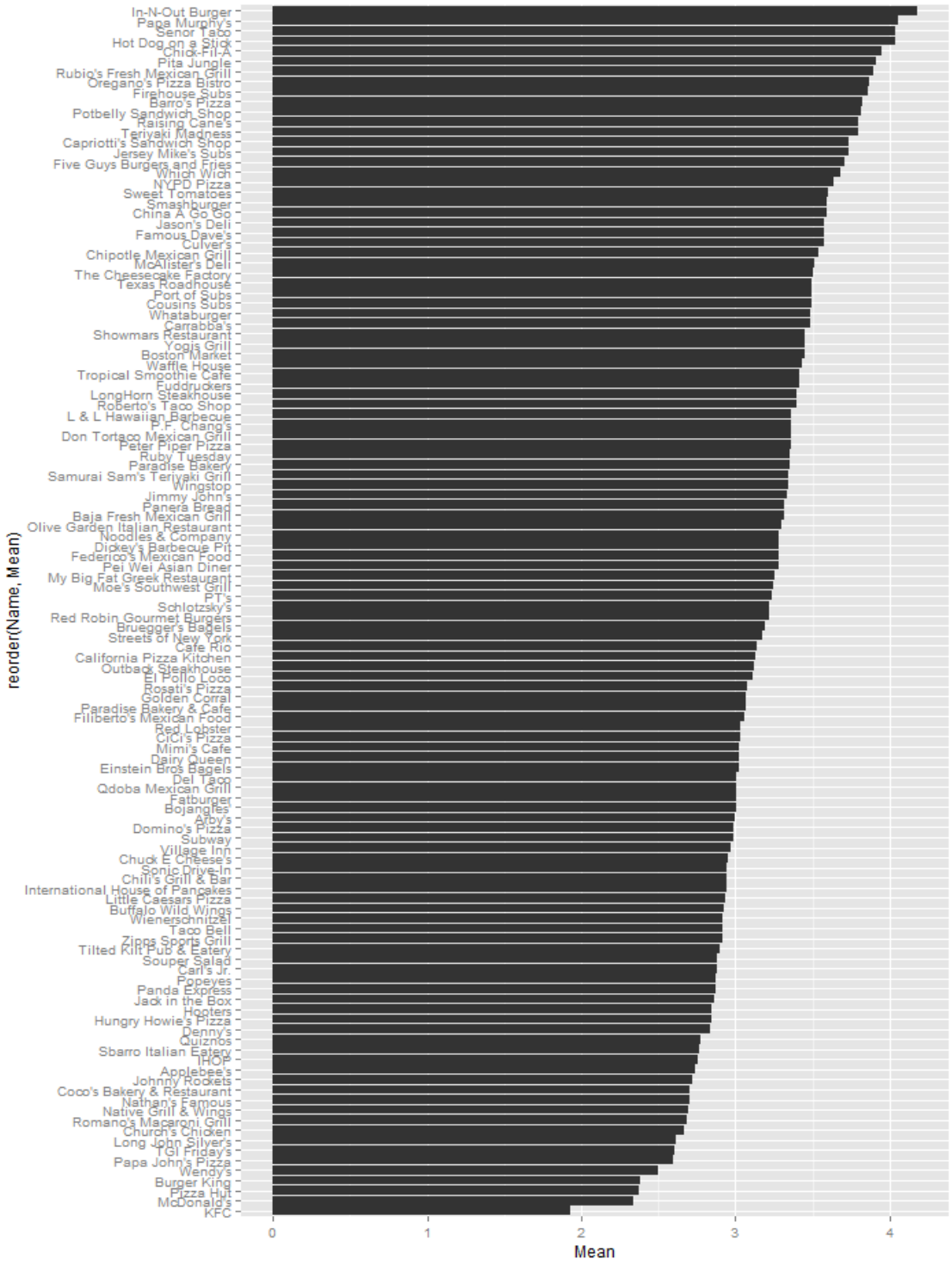


Figure 6. Number of restaurants of the 120 chains

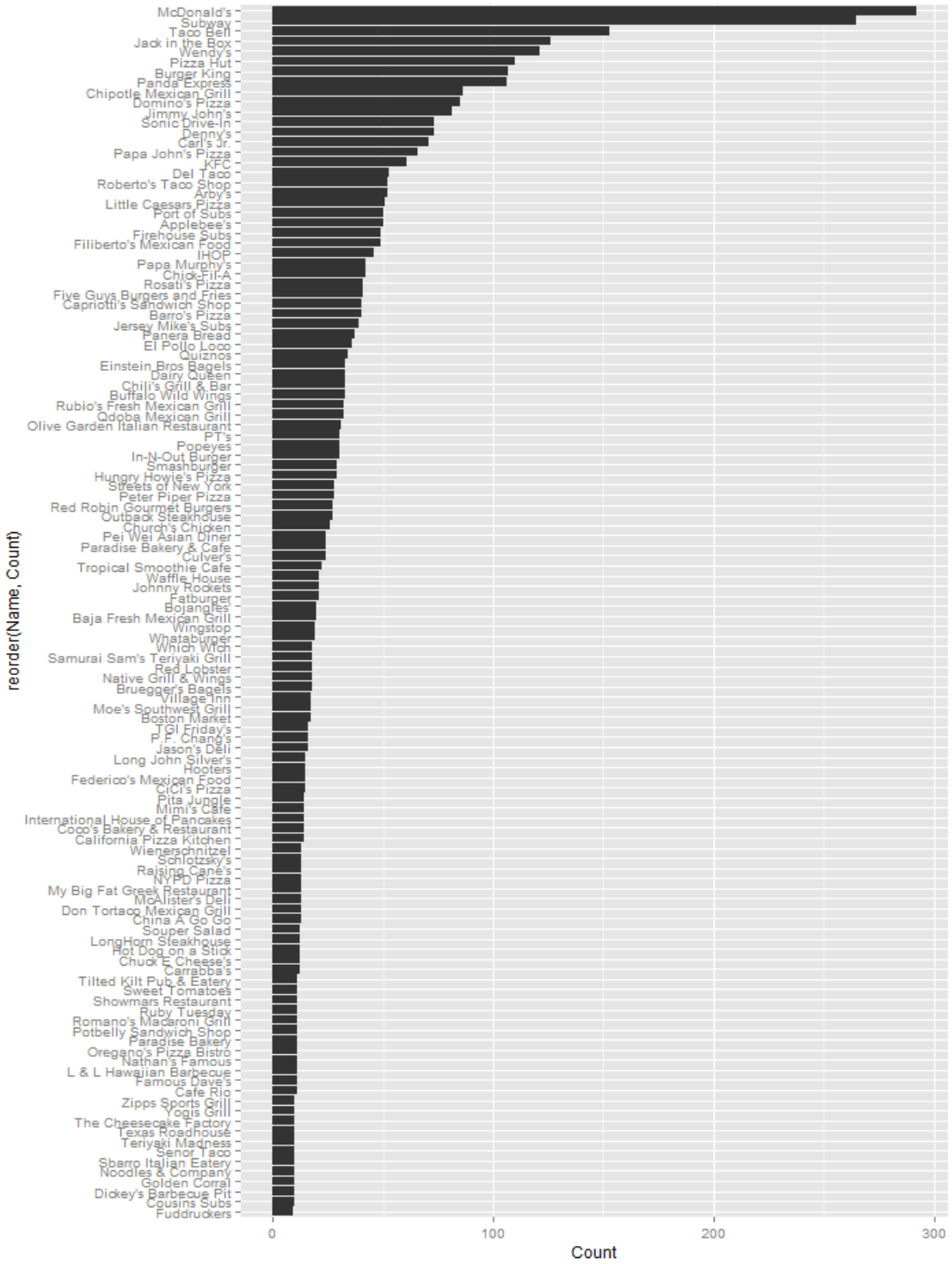


Figure 7. Distribution of mean ratings of the 120 chains.

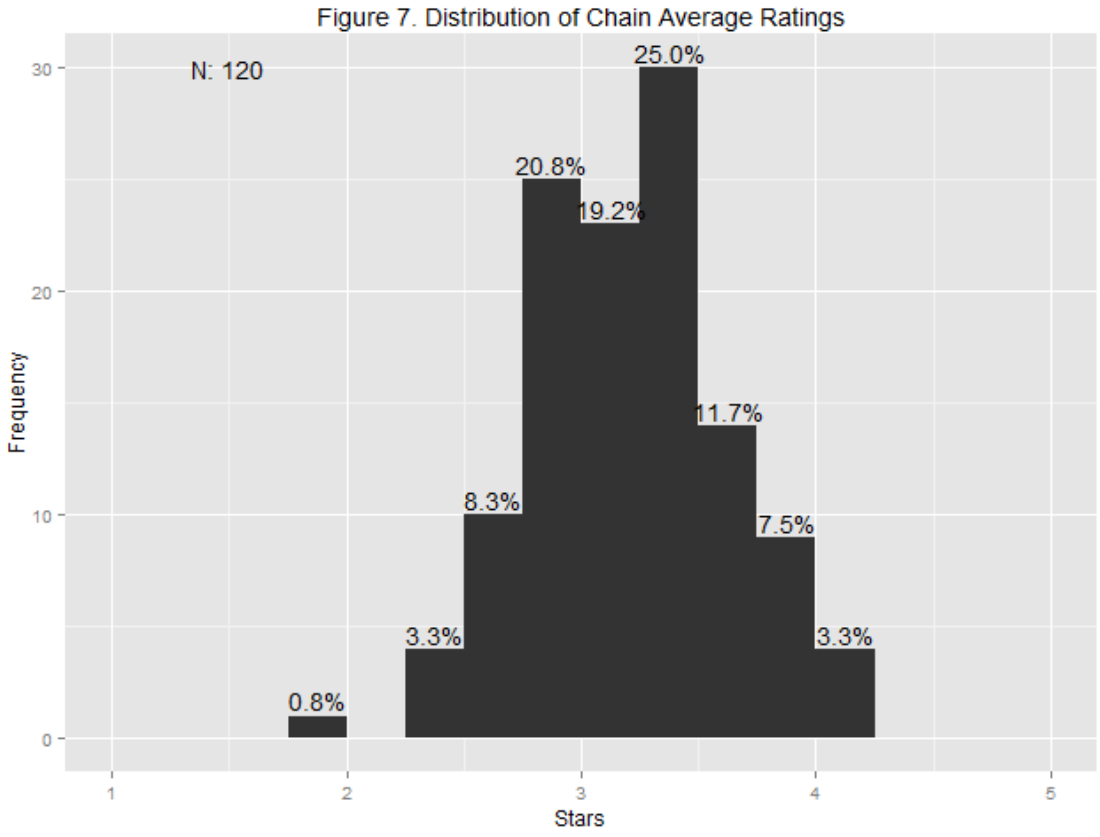
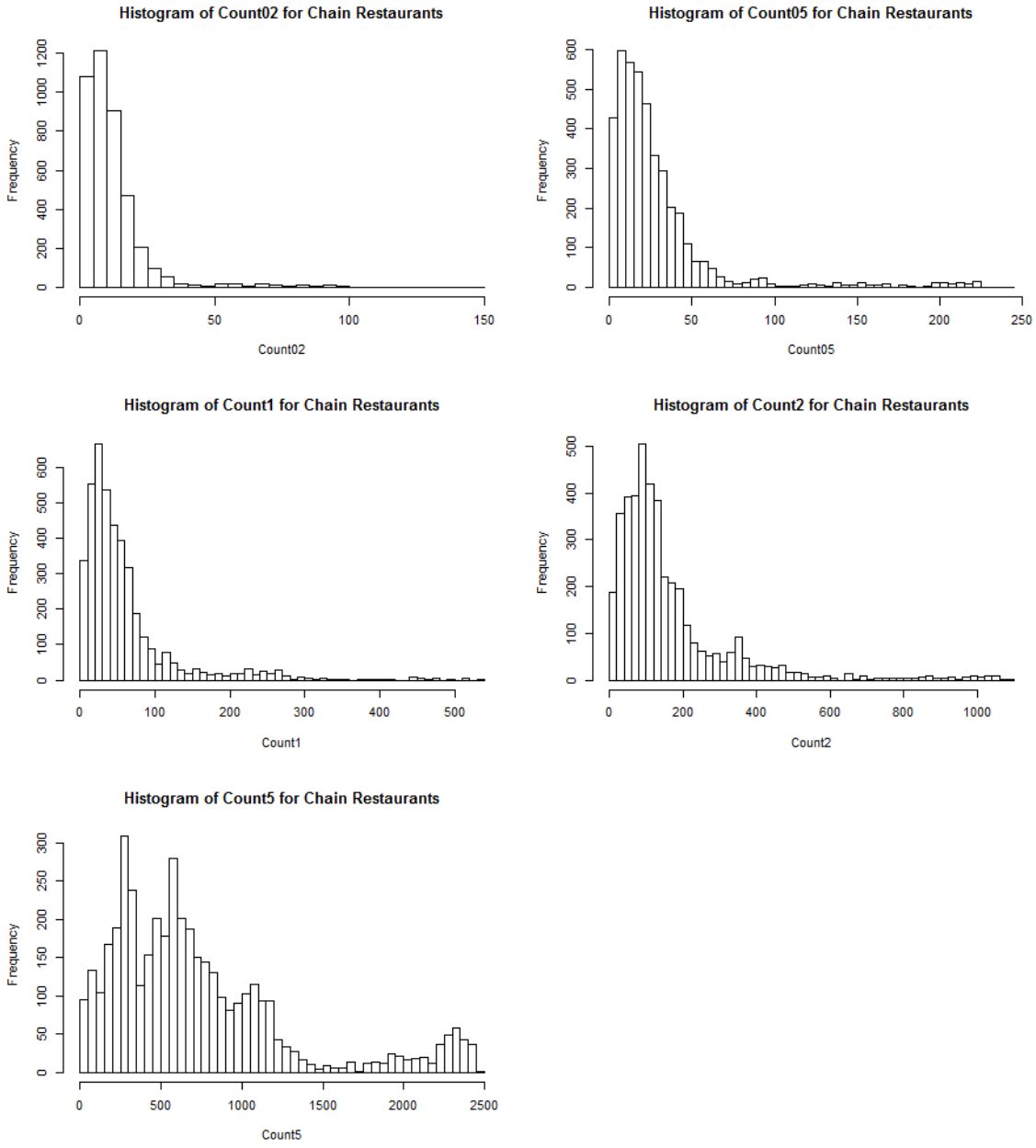
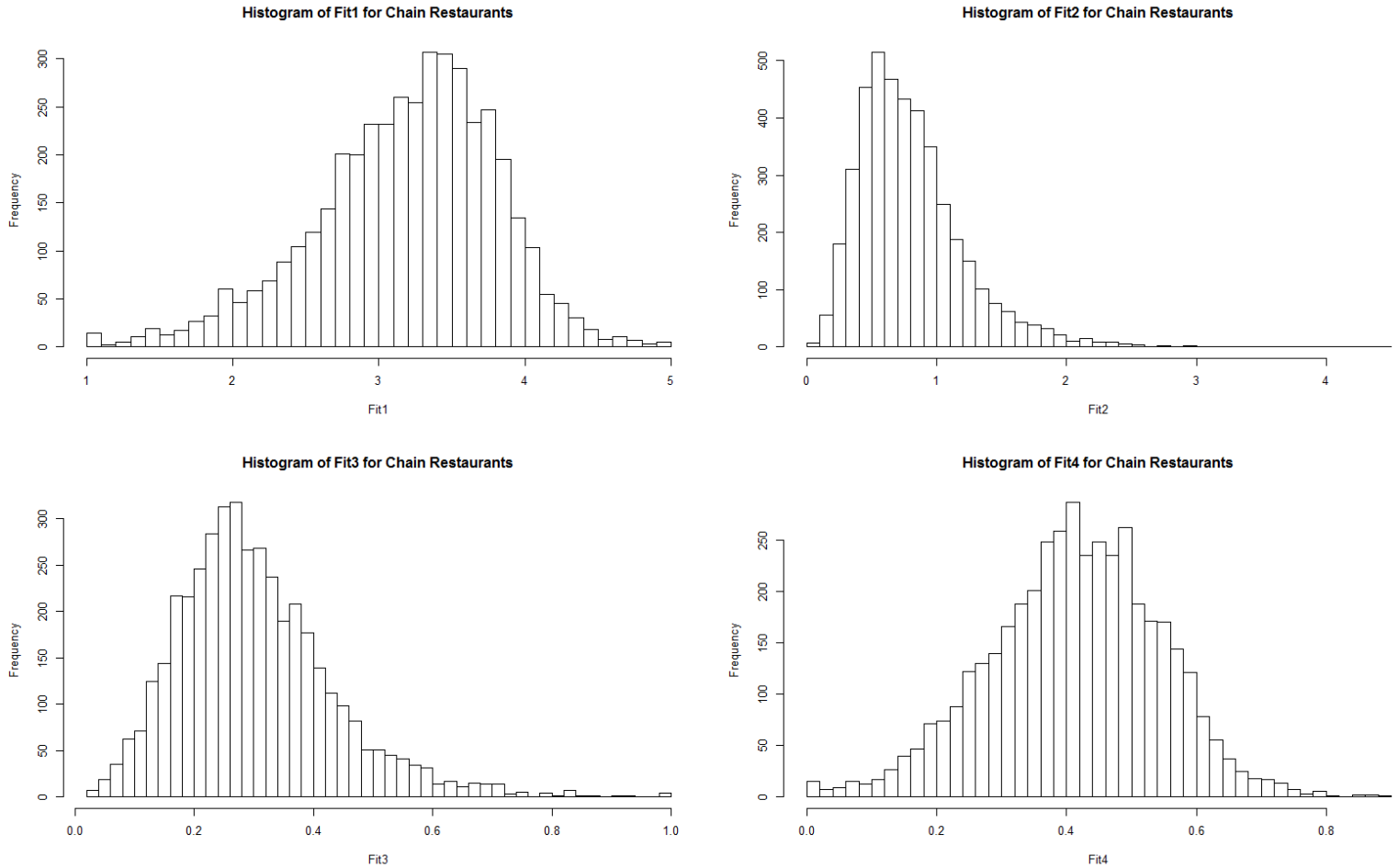


Figure 8. Distributions of density measures for chain restaurants



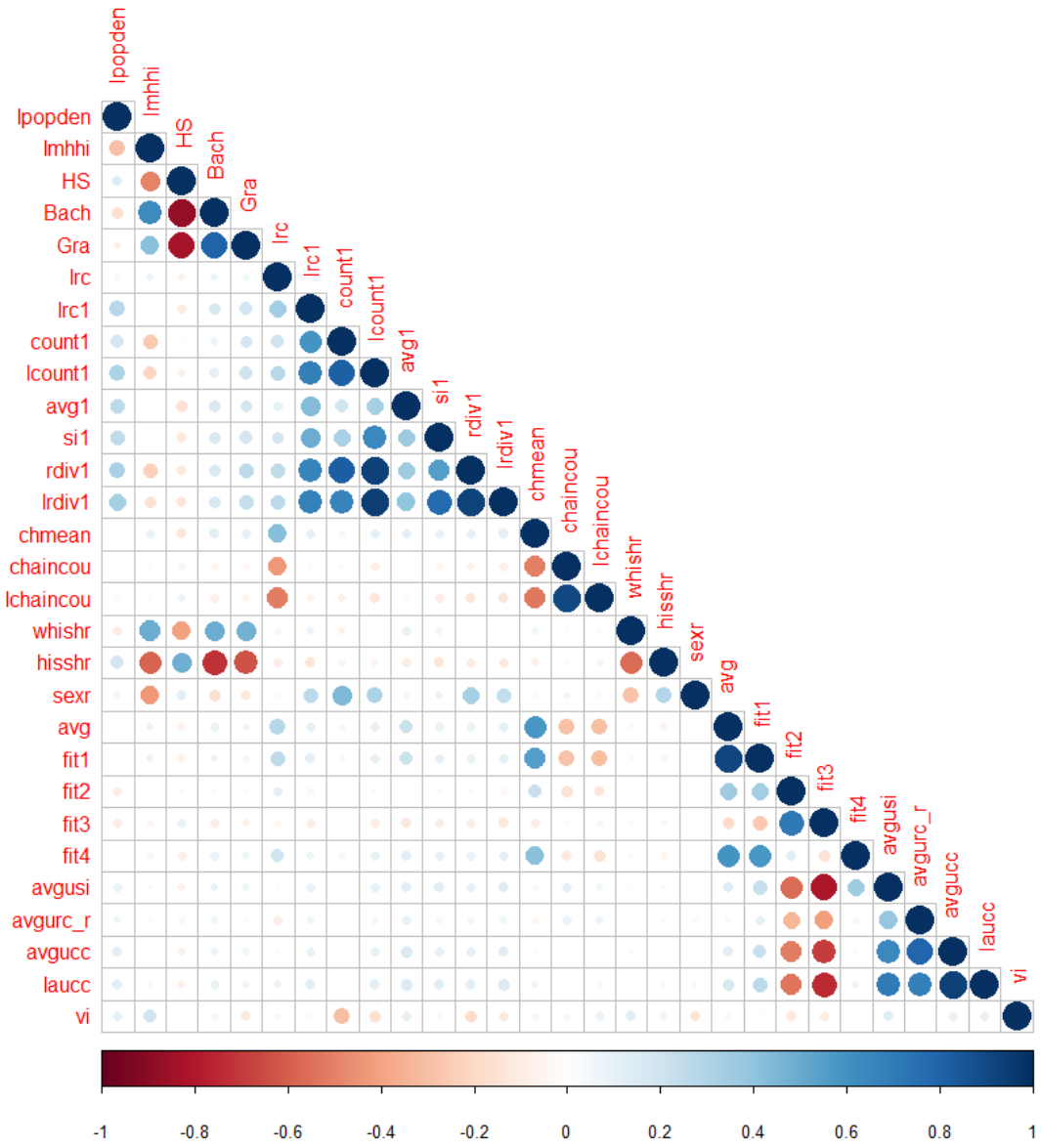
Note: N=4,200. Count02, Count05, Count1, Count1, and Count5 are the number of restaurants (not only chain restaurants) within a radius of 0.2, 0.5, 1, 2, and 5 mile(s) of a focal chain restaurant.

Figure 9. Distributions of Fit1, Fit2, Fit3, and Fit4 for chain restaurants



Note: N=4,200. Fit1, Fit2, Fit3, and Fit4 are the four measures for business-customer matching. The higher the Fit score, the better a restaurant is matched with its customers.

Figure 10. Correlation Matrix of variables with radius=1 mile



Note: The darker the color, the higher the correlation.

Figure 11. Revised frameworks with the addition of the omnivorousness construct

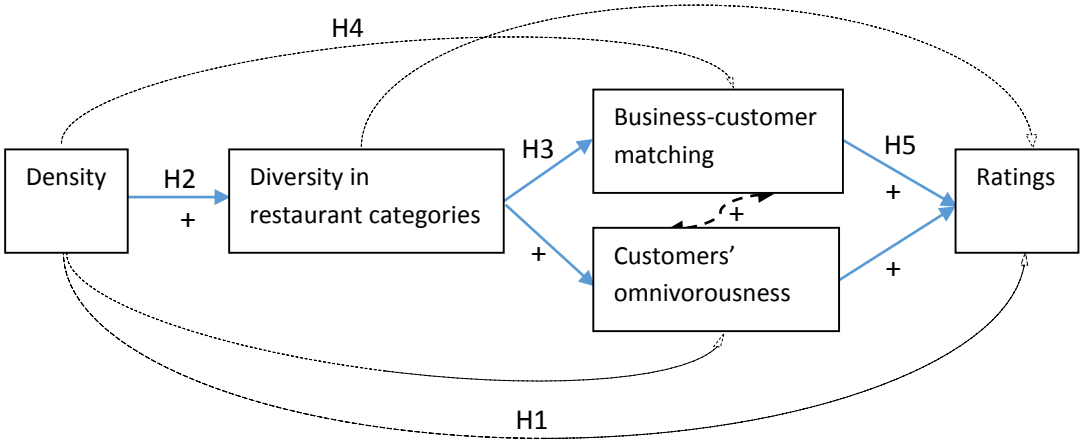


Figure 11.a. The Business-customer matching construct is still kept.

Note: Solid straight arrows indicate direct effects, dashed curved arrows indicate indirect effects. Plus signs indicate positive associations that have been verified from the analyses. Hypotheses in the paper are shown.

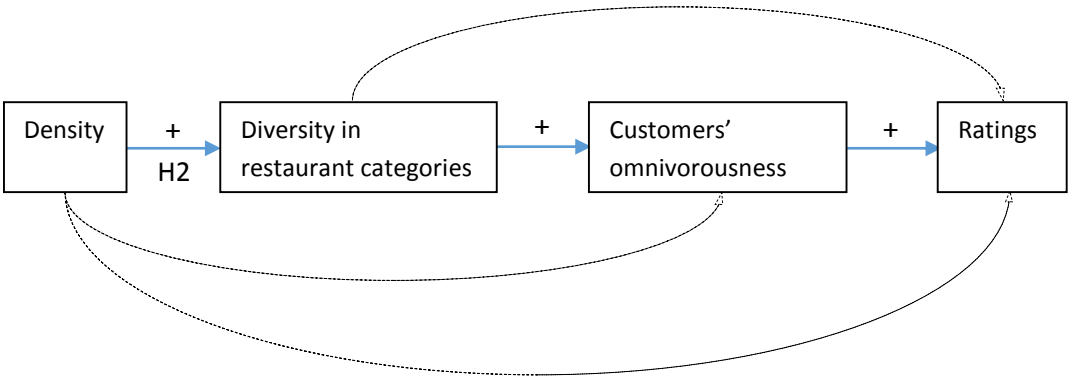


Figure 11.b. The Business-customer matching construct is removed.

Note: Solid straight arrows indicate direct effects, dashed curved arrows indicate indirect effects. Plus signs indicate positive associations that have been verified from the analyses. Hypotheses in the paper are shown.

Figure 12. Structural Equation Modeling results when radius = 1 mile

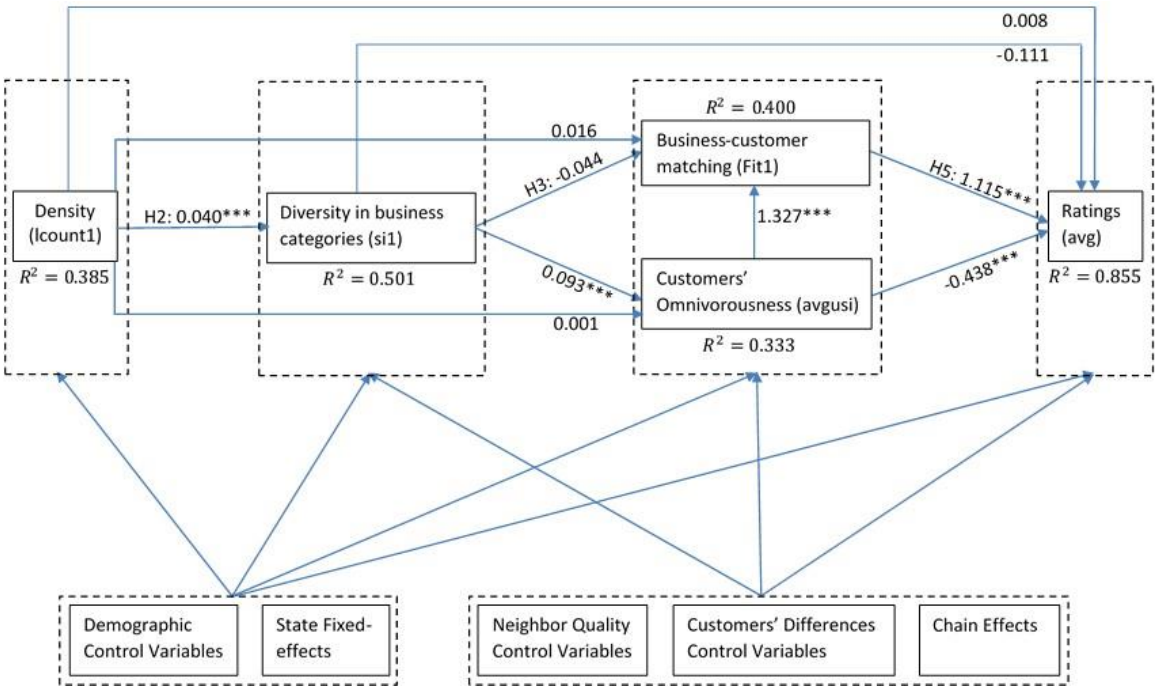


Figure 12.a. r=1 mile, surrounding restaurants' diversity measure: si1, matching measure: Fit1.

Note: *p<0.1, **p<0.05, ***p<0.01. All arrows indicate direct effects. Among the five main constructs, a total of 10 direct paths are shown. Boxes at the top are variables of interest; boxes at the bottom indicate control variables.

Figure 12. Structural Equation Modeling results when radius = 1 mile (Continued)

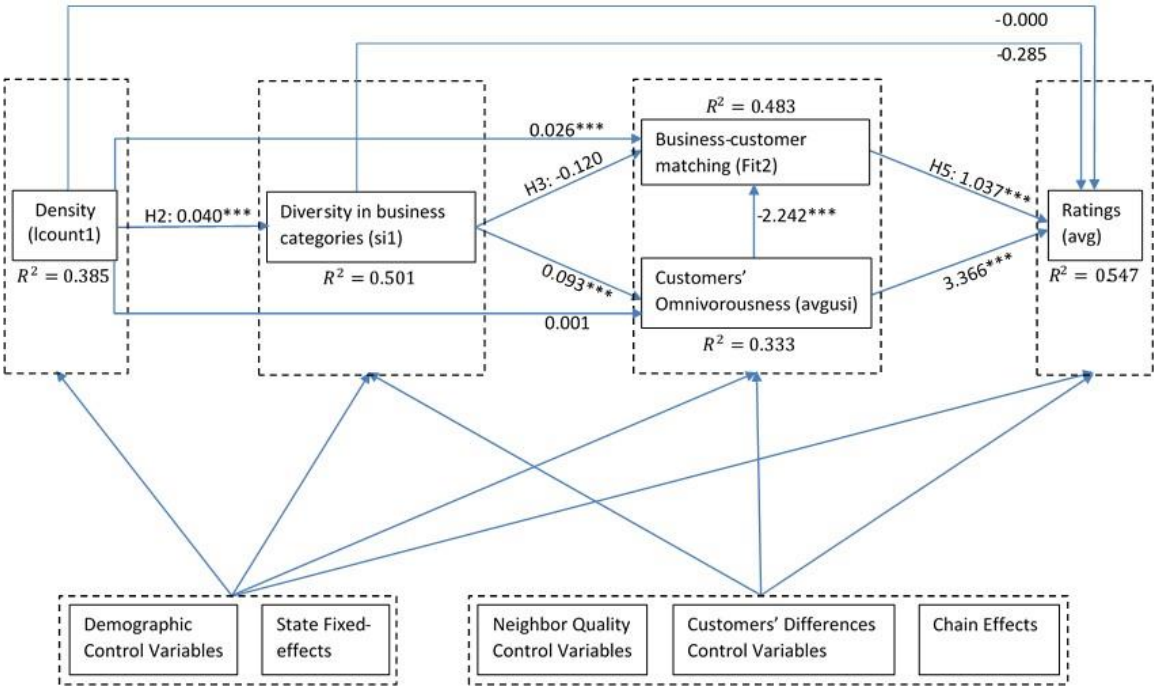


Figure 12.b. r=1 mile, surrounding restaurants' diversity measure: si1, matching measure: Fit2.

Note: *p<0.1, **p<0.05, ***p<0.01. All arrows indicate direct effects. Among the five main constructs, a total of 10 direct paths are shown. Boxes at the top are variables of interest; boxes at the bottom indicate control variables.

Figure 12. Structural Equation Modeling results when radius = 1 mile (Continued)

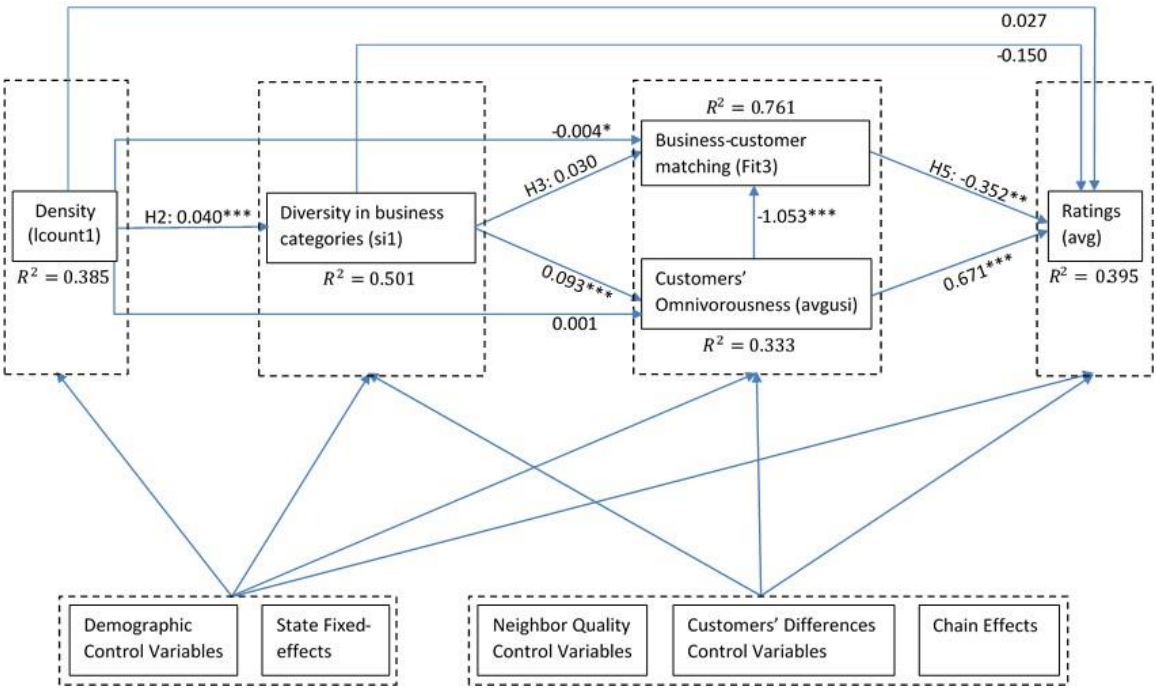


Figure 12.c. r=1 mile, surrounding restaurants' diversity measure: si1, matching measure: Fit3.

Note: *p<0.1, **p<0.05, ***p<0.01. All arrows indicate direct effects. Among the five main constructs, a total of 10 direct paths are shown. Boxes at the top are variables of interest; boxes at the bottom indicate control variables.

Figure 12. Structural Equation Modeling results when radius = 1 mile (Continued)

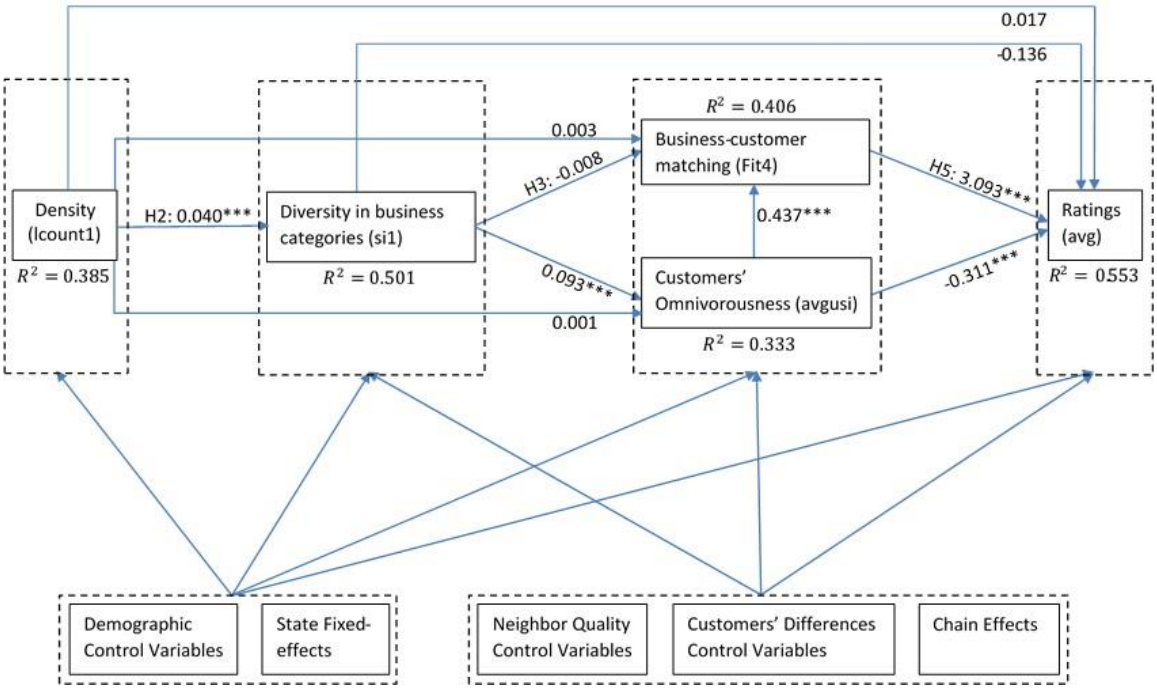


Figure 12.d. r=1 mile, surrounding restaurants' diversity measure: si1, matching measure: Fit4.

Note: *p<0.1, **p<0.05, ***p<0.01. All arrows indicate direct effects. Among the five main constructs, a total of 10 direct paths are shown. Boxes at the top are variables of interest; boxes at the bottom indicate control variables.

Figure 13. Structural Equation Modeling results without business-customer matching construct when radius = 1 mile

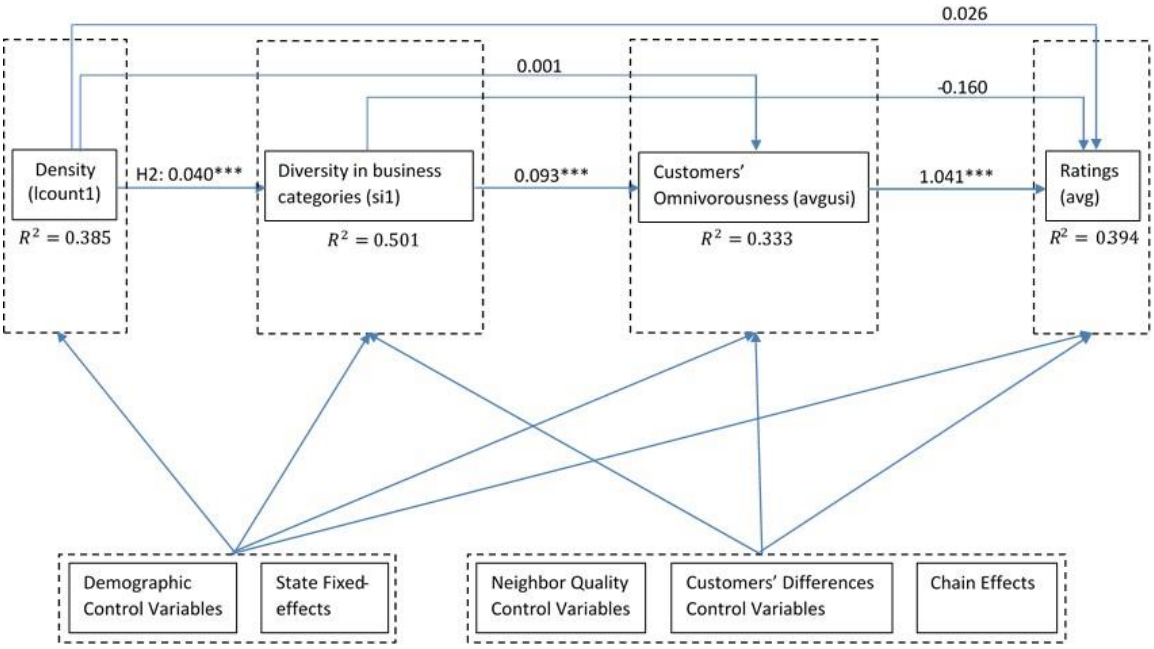


Figure 13.a. r=1 mile, surrounding restaurants' diversity measure: si1, customers' omnivorousness measure: avgusi.

Note: *p<0.1, **p<0.05, ***p<0.01. All arrows indicate direct effects. Among the four main constructs, a total of 6 direct paths are shown. Boxes at the top are variables of interest; boxes at the bottom indicate control variables.

Figure 13. Structural Equation Modeling results without business-customer matching construct when radius = 1 mile (Continued)

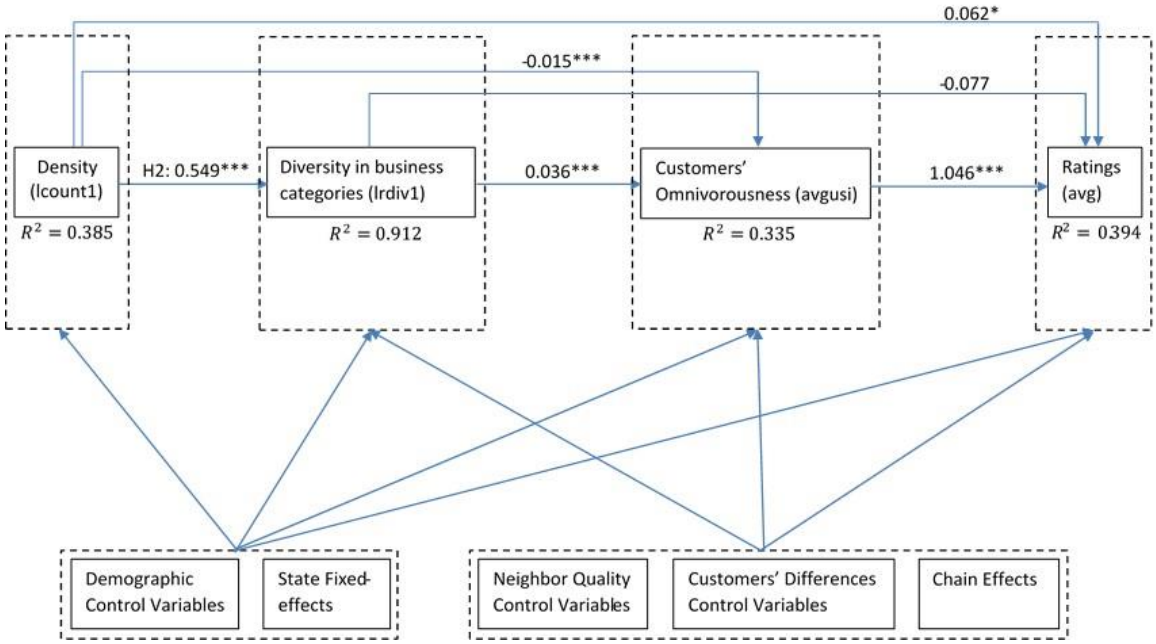


Figure 13.b. r=1 mile, surrounding restaurants' diversity measure: lrdiv1, customers' omnivorousness measure: avgusi.

Note: *p<0.1, **p<0.05, ***p<0.01. All arrows indicate direct effects. Among the four main constructs, a total of 6 direct paths are shown. Boxes at the top are variables of interest; boxes at the bottom indicate control variables.

Figure 13. Structural Equation Modeling results without business-customer matching construct when radius = 1 mile (Continued)

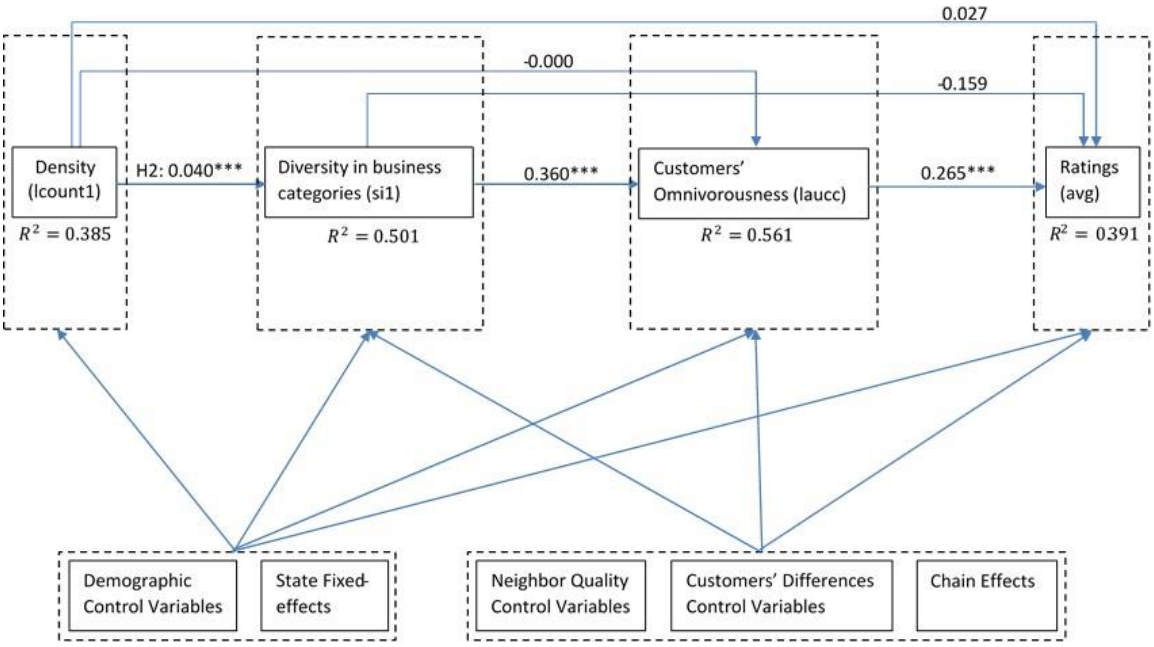


Figure 13.c r=1 mile, surrounding restaurants' diversity measure: si1, customers' omnivorousness measure: laucc.

Note: *p<0.1, **p<0.05, ***p<0.01. All arrows indicate direct effects. Among the four main constructs, a total of 6 direct paths are shown. Boxes at the top are variables of interest; boxes at the bottom indicate control variables.

Figure 13. Structural Equation Modeling results without business-customer matching construct when radius = 1 mile (Continued)

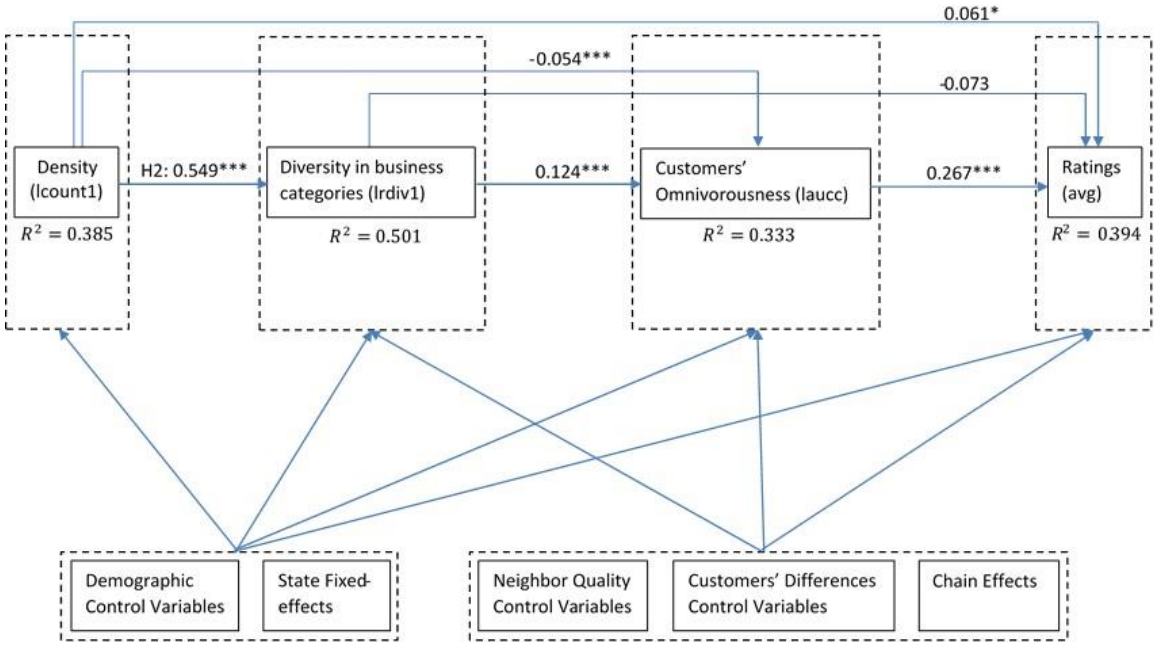
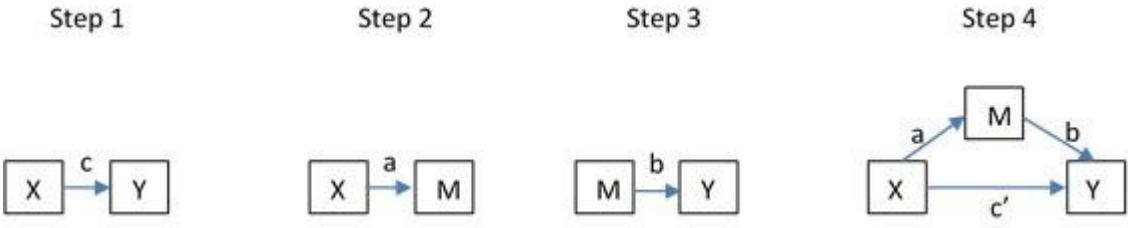


Figure 13.d r=1 mile, surrounding restaurants' diversity measure: lrdiv1, customers' omnivorousness measure: laucc.

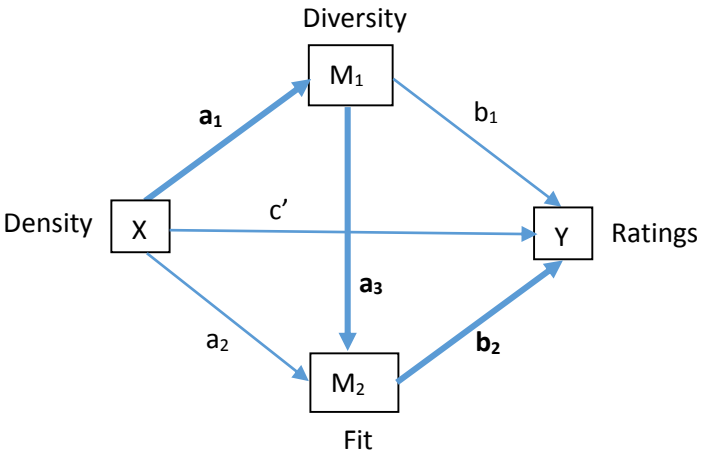
Note: *p<0.1, **p<0.05, ***p<0.01. All arrows indicate direct effects. Among the four main constructs, a total of 6 direct paths are shown. Boxes at the top are variables of interest; boxes at the bottom indicate control variables.

Figure A.1. Four-step testing for a simple mediation model



Note: All arrows indicate direct effects. To verify M as a full mediator, c, a, and b should be significant in the first three steps, then c' should become insignificant after adding M to the model. Otherwise M is a partial mediator if c' is also significant.

Figure A.2. A two-step two-mediator structure



Note: X is the treatment, Y is the effect, M1 and M2 are two mediators. All arrows indicate direct paths, arrows in bold are direct paths that are hypothesized to exist in the paper. a1, a3, and b2 correspond to H2, H3, and H5 respectively. See Hayes, Preacher, and Myers (2011) for a similar graph and a more general discussion of the two-mediator structure.