NEIGHBORHOOD SELF-IDENTITY AND POINT OF INTEREST IDENTIFICATION ON AIRBNB

by

Peter Thomas
A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Master of Science
Geoinformatics and Geospatial Intelligence

Committee:

_____________________________________ Dr. Arie Croitoru, Thesis Director

_____________________________________ Dr. Anthony Stefanidis, Committee Member

_____________________________________ Dr. Andrew Crooks, Committee Member

_____________________________________ Dr. Dieter Pfoser, Department Chairperson

_____________________________________ Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science

_____________________________________ Dr. Peggy Agouris, Dean, College of Science

Date: ______________________________ Spring Semester 2018
George Mason University
Fairfax, VA

Neighborhood Self-Identity and Point of Interest Identification on Airbnb

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

by

Peter Thomas
Graduate Certificate
George Mason University, 2017
Bachelor of Arts
The College of William and Mary, 2007

Director: Arie Croitoru, Associate Professor
George Mason University

Spring Semester 2018
George Mason University
Fairfax, VA

DEDICATION

This is dedicated to Priscilla and Pandora, the other two Ps in my pod.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. C for his guidance and insights, my committee, and the George Mason GGS department faculty for inspiring my research path. Finally, thank you to Priscilla, my family, and my friends for their patience with my constant excuse of “sorry, I have to work on my thesis.”

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

Density-Based Spatial Clustering of Applications with Noise	DBSCAN
InsideAirbnb.com	Inside Airbnb
Natural Language Toolkit	NLTK
Points of Interest	POIs
Support Vector Machines	SVM
Term Frequency–Inverse Document Frequency	TF-IDF
Volunteered Geographic Information	VGI

ABSTRACT

NEIGHBORHOOD SELF-IDENTITY AND POINT OF INTEREST IDENTIFICATION ON AIRBNB

Peter Thomas, M.S.

George Mason University, 2018

Thesis Director: Dr. Arie Croitoru

Room-sharing marketplace Airbnb is disrupting the short-term rental market and leisure travel industry by providing a platform to connect accommodation producers (Hosts) and consumers (Guests). Airbnb is growing rapidly and has more than 3 million Listings worldwide. Airbnb Listings are a rich and under-researched corpus of Volunteered Geographic Information (VGI). A Listing's neighborhood description, written by the Host, contains a wealth of information on local attractions such as parks, restaurants, nearby landmarks, and neighborhoods. This thesis uses the neighborhood descriptions in geolocated Airbnb Listings to delineate neighborhood boundaries and discover and geolocate unique Points of Interest (POIs) in New York City. This study constructs neighborhood maps through DBSCAN convex hull creation and hex assignment. Results show that the context of a neighborhood name changes based on how early it occurs in the neighborhood overview field: the first sentence is the Listing's location, and subsequent sentences are references to other nearby neighborhoods. Network analysis demonstrates that Listings reference nearby neighborhoods frequently and distant neighborhoods infrequently. The DBSCAN clustering algorithm is applied to effectively identify which frequent ngrams are highly spatially clustered and likely to represent a unique POI. This work is a novel application of crowdsourced neighborhood and POI identification techniques to a new VGI dataset.

1. INTRODUCTION

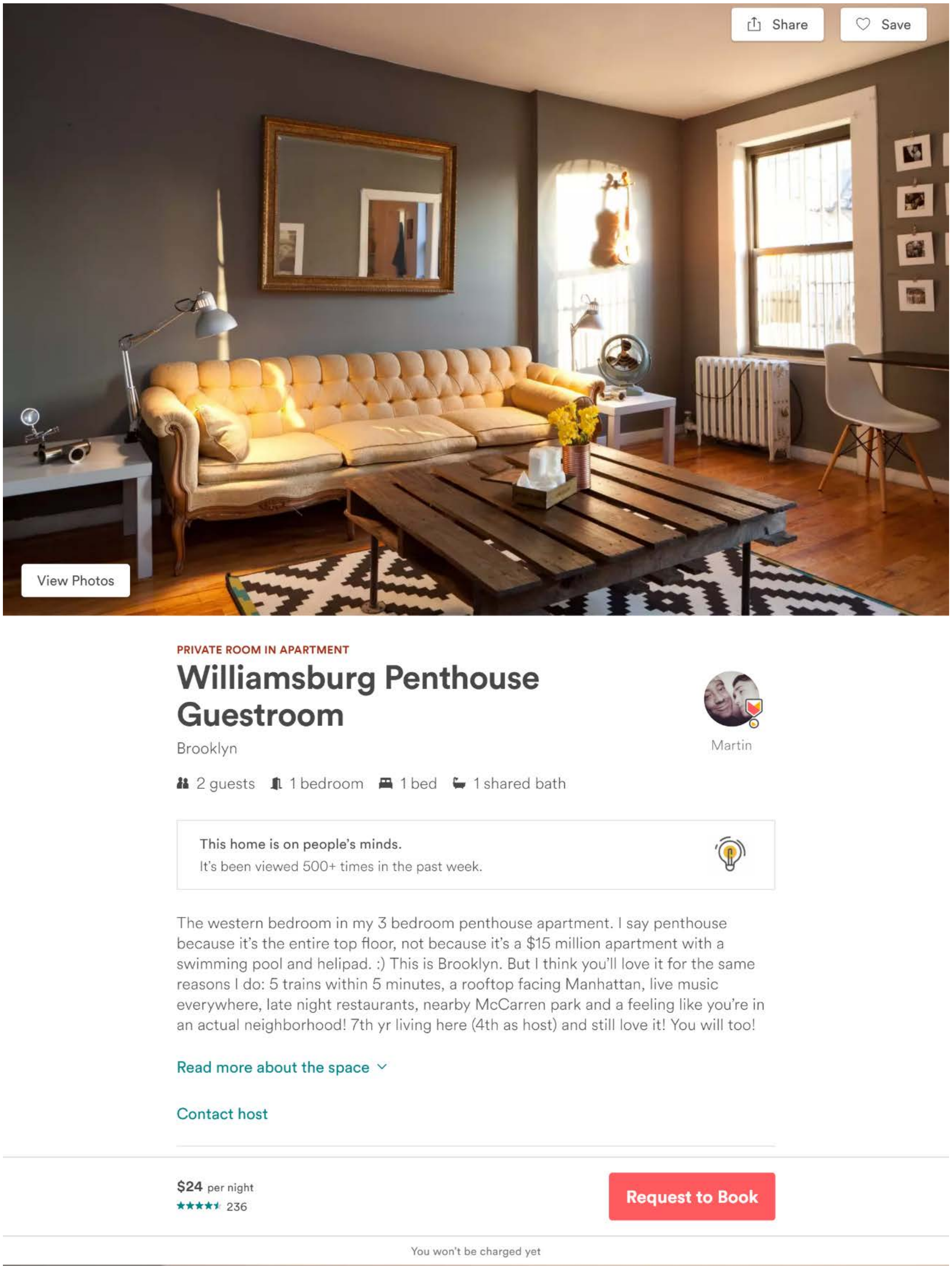
The rise of collaborative consumption—the "sharing economy"—is a defining story of the 2010s. The sharing economy allows users to offer and consume peer-to-peer a variety of goods and services, such as transportation and housing, and disrupts existing business models. The sharing economy has also created a wealth of new research opportunities and data for study. This thesis uses publicly-available Airbnb datasets to identify neighborhood boundaries and discover Points of Interest (POIs) in New York City.

1.1 <u>About Airbnb</u>

Room-sharing marketplace Airbnb provides a user-friendly platform to match available accommodation with short term renters. Airbnb is the second-most valuable company in the sharing economy, second only to ridesharing service Uber. Airbnb has experienced rapid growth from humble beginnings in 2008 and currently boasts a private valuation of $31 billion (Zaleski, 2017) and a global presence, with more than 4.5 million Listings worldwide in more than 191 countries (Airbnb, n.d.). Airbnb is easily accessible as a website and as a mobile application and has been widely adopted by independent travelers. The extent of Airbnb's disruptive impact on housing markets, the hotel industry, and benefits to the local economy are the subject of fierce ongoing debate, and many municipalities are enacting or considering legislation to regulate unlicensed room sharing. Although Airbnb’s growth is slowing due to market saturation, concerns with privacy and safety, and regulatory uncertainty, use is still increasing: 25% of travelers used Airbnb in 2017, and 80% were aware of Airbnb (Scaggs, 2017).

1.2 Airbnb Listings

Airbnb Listings are a large and under-researched corpus of Volunteered Geographic Information (VGI). This paper makes extensive use of the Airbnb terms Listing, Host, and Guest. A Listing is a post advertising either a room or an entire apartment or house and is written by the property owner (“Host”) to attract renters (“Guests”). Hosts can own multiple Listings, and it is not uncommon for a Host to post different rooms in a house or apartment as separate Listings. Figure 1 displays part of a representative Airbnb Listing. Listings include pictures, information on amenities, sleeping arrangements, house rules, cancellation policy, reviews, and neighborhood.



PRIVATE ROOM IN APARTMENT

Williamsburg Penthouse Guestroom

Brooklyn

2 guests 1 bedroom 1 bed 1 shared bath

This home is on people's minds.
It's been viewed 500+ times in the past week.

The western bedroom in my 3 bedroom penthouse apartment. I say penthouse because it's the entire top floor, not because it's a $15 million apartment with a swimming pool and helipad. :) This is Brooklyn. But I think you'll love it for the same reasons I do: 5 trains within 5 minutes, a rooftop facing Manhattan, live music everywhere, late night restaurants, nearby McCarren park and a feeling like you're in an actual neighborhood! 7th yr living here (4th as host) and still love it! You will too!

$24 per night
★★★★★ 236

You won't be charged yet

Figure 1 Airbnb Listing (Martin, n.d.).

The Neighborhood Overview, written by the Host, provides a wealth of information on local attractions that from the Host’s point of view would be of interest to prospective Guests, such as parks, restaurants, nearby landmarks, and neighborhoods. Figure 2 illustrates the 50 most frequent words in the Neighborhood Overview field for October 2017 Listings in New York City, scaled by frequency. The most frequent words are attractions, placenames, and distance measures.

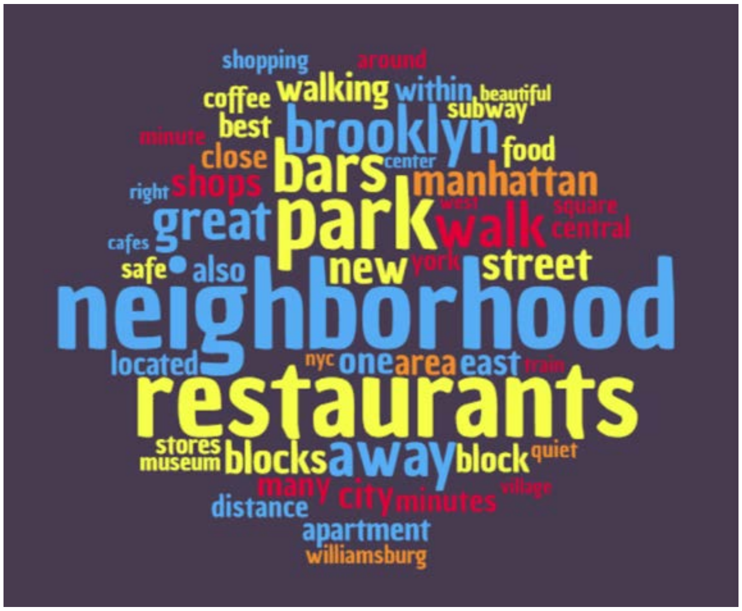


Figure 2 Word cloud of Neighborhood Overview words scaled by frequency.

Because Listings are geolocated, the Neighborhood Overviews can be aggregated and analyzed to delineate neighborhood boundaries and locate POIs. Naturally, Host bias will be evident as hosts are trying to advertise the positive aspects of their neighborhood and attract Guests. Hosts may reference landmarks and attractions that may not be close by, or misrepresent a Listing as located in a more fashionable adjacent neighborhood.

1.3 Airbnb Location Anonymization

A disadvantage of working with Airbnb data is that Listing geocoordinates are not precise. Listing addresses are not publicly available and are only shared with Guests after making a booking. Inside Airbnb data include the Airbnb-provided geocoordinates for each Listing; however, Airbnb anonymizes Listing geocoordinates to protect host privacy, and the coordinates may be 0-150 meters from the actual location. When there are multiple Listings within the same building, each Listing is displayed with a different location offset.

The author stayed in an Airbnb in New York City in January 2018 and was able to find related Listings in the October 2017 Inside Airbnb data and independently verify the Listing's geolocation anonymization. The Listing was advertised as a private room in an apartment, but the apartment had a total of eight rooms available for rent. Figure 3 shows the actual location of the Listing (the apartment building in blue) and the Airbnb geolocations for the eight Listings (in red). In this example, Flushing Avenue divides the Airbnb neighborhoods of Williamsburg and Bushwick. The Listing is in Williamsburg, and despite immediate adjacency to the neighborhood boundary, all points are located in Williamsburg, suggesting that Airbnb's method of introducing error keeps all points within the Listing's actual geolocated neighborhood.

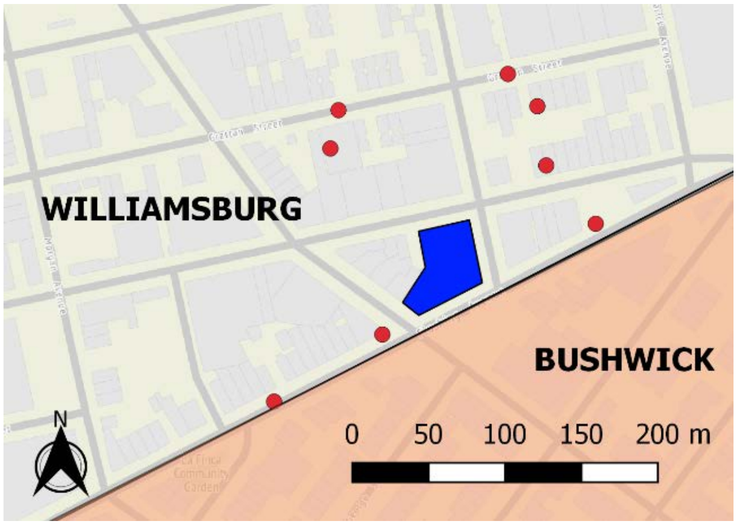


Figure 3 Airbnb Listing locations (red) for multiple rooms listed in a single building (blue).

This thesis uses Airbnb Listing data to investigate neighborhood identity and locate POIs. As more people live in urban areas, it is increasingly important to understand how citizens perceive and relate to their neighborhoods. Likewise, as the sharing economy grows, it is increasingly important to understand how users frame and portray their neighborhoods to others. Airbnb data is ideal for researching the intersection between the sharing economy and modern concepts of the neighborhood.

2. RELATED WORK

The main body of academic research on Airbnb studies Airbnb’s impact on housing, the hotel industry, and sharing economy social issues such as discrimination, bias, reciprocity, trust, reputation, and social exchange. High-profile research on Airbnb’s impact on housing prices is also commissioned and published by the public sector, such as a 2015 report produced as part of an investigation by the New York Attorney General (Delgado-Medrano & Lyon, 2016).

To the best of our knowledge, this is the first study to apply neighborhood or POI identification techniques to Airbnb data. This thesis combines elements of previous research on VGI, neighborhood identity, and POIs.

2.1 Volunteered Geographic Information (VGI)

VGI provides a large volume of information at low cost, but with unknown data quality. How credible is VGI, and can it be trusted? Information previously was deemed credible by gatekeepers who could manage and maintain information, but such an approach is not feasible in an environment of information abundance, where data is provided by untrained end-users. Traditional geospatial data quality elements such as positional accuracy, attribute accuracy, logical consistency, completeness, lineage, temporal accuracy, and semantic accuracy (Guptill & Morrison, 2013) can be difficult to assess for VGI. However, comparisons of VGI such as OpenStreetMap to authoritative data have found that data quality is very good (Haklay, 2010), and although there are problems of data heterogeneity (Girres & Touya, 2010), VGI can be more current than authoritative data (Fonte et al, 2015). A detailed overview on VGI quality evaluation efforts is given by Antoniou and Skopeliti (2015).

Airbnb data faces similar data quality challenges of other forms of VGI. For Airbnb, many attributes are self-reported by the hosts, and precise positional accuracy is obfuscated by the previously discussed displacement of 0-150m (Section 1.3). Airbnb data quality is also influenced by Host motivations. Not only are hosts unlikely to be trained in providing accurate data, it is in their interest to provide only positive information in order to attract more Guests and earn more revenue.

VGI quality assurance can be done through crowd-sourcing, social, and geographic approaches (Goodchild & Li, 2012), and these methods are applied to this study. First, the crowd-sourcing approach requires repeated observation of neighborhood characteristics and POI references, using independent, consistent reports (in this case, multiple Listings) to converge on a consensus definition. Second, the geographic approach involves comparing crowdsourced findings with broad geographic knowledge. For example, Airbnb Listings should exhibit spatial dependence: purported facts should be consistent with known geographic facts. This study compares Airbnb-derived findings with known locations of neighborhoods and POIs. Finally, Airbnb uses a social quality assurance approach to self-regulate hosts. Users of user-volunteered information are sophisticated in their assessment of credibility and are perceptive to the number of and content of user reviews (Metzger et al, 2010). Hosts receive scores on a variety of metrics, including accuracy. Listings with low scores are less likely to attract guests, so hosts are motivated to provide accurate information.

Social quality assurance is further enhanced because Listing author identity is clear: descriptions are provided by the Hosts, who are required to maintain a profile with a photo and personal information. Credibility in this case is based on believability, trustworthiness, and expertise. This credibility-as-perception is subjective and key to understanding VGI (Flanagin & Metzger, 2008). The sharing economy is built on this trust: trust that a service provider will provide a certain service in their area of expertise, and trust that the customer will use the service provider’s resources responsibly. The Airbnb experience is claimed as better than alternatives because hosts are in touch with the community, and guests use their Host’s knowledge to experience the location with local insight and “live like a local” (“Largest Airbnb Campaign to Date,” 2016).

Airbnb’s terms of service also require that Hosts maintain accurate Listing descriptions (Airbnb, 2017). It is unknown how actively Airbnb monitors content and removes Listings which violate their terms of service. Airbnb should be motivated by the bottom line and therefore expected to encourage Listing accuracy to improve customer satisfaction: customers are less likely to use Airbnb if they perceive Listing descriptions to be inaccurate.

Given these crowd-sourcing, social, and geographic quality assurance methods, we believe Airbnb Listing data is appropriate for use in identifying the locations of neighborhoods and POIs.

2.2 Neighborhood Identity

Neighborhood definitions are subjective: neighborhoods mean different things to different people (Haeberle, 1988). The consensus definition within the literature is that a “neighborhood” is a contiguous geographic unit of limited size, with relatively homogeneous characteristics, and a symbolic significance to residents (Weiss et al, 2007). For the purposes of this work, a neighborhood is considered to be a contiguous area, referred to with a consistent name by Airbnb Hosts, at a smaller scale than New York City’s five boroughs (Manhattan, Brooklyn, Queens, the Bronx, and Staten Island).

The significance and nature of neighborhoods are discussed extensively by Galster (2001) and Kearns and Parkinson (2001). Neighborhoods are important because they foster social networks and connection to a place, and residents are happier when they feel connected to their neighborhood and to the people who live in their neighborhood (Leyden et al, 2011). The society dependent on a location forms a symbolic community which evolves and provides continuity over long periods of time (Hunter, 1974).

The boundaries of neighborhoods are often subjective and malleable (Weiss et al, 2007). Urban planners and policy stakeholders desire well-defined neighborhood boundaries, but the reality is that neighborhood boundaries are often fuzzy and organic in areas without natural boundaries (Chaskin, 1997). Neighborhoods are an example of a vague cognitive region, and do not have a crisp boundary (Montello et al, 2003). Earlier work to investigate neighborhood boundaries often relied on field work techniques such as resident interviews (Guest & Lee, 1984) and map-drawing exercises (Coulton et al, 2001), which are time consuming to collect and difficult to perform at scale. Newer studies in this area make use of social media and VGI.

Although the body of literature on neighborhood identity does not address Airbnb neighborhoods, there are several papers which use social media sources such as Foursquare, Twitter, or Flickr to investigate neighborhood identity and define neighborhood boundaries, and these techniques can be applied to Airbnb data. Notable applications of Foursquare data include the “Hoodsquare” neighborhood detection algorithm, which uses spatial clustering methods to identify neighborhoods (Zhang et al, 2013), and Cranshaw and Yano’s neighborhood identification through Latent Topic Modeling analysis of check-in location categories (2010). Twitter applications include network analysis of geotagged Tweets using the Infomap algorithm to model the overlapping and nested nature of neighborhoods in New York City (Poorthuis, 2017), and a proposed methodology to automatically generate neighborhood guides (Tasse et al, 2016). Other applications include identifying city cores and neighborhoods using kernel density estimation and Flickr data (Hollenstein & Purves, 2010), and the identification of areas of interest in Shanghai using Panoramio images and Qieke check-ins (Liu et al, 2012).

To the best of our knowledge, no currently published studies use Airbnb data to delineate neighborhoods. Previous studies which touch on Airbnb neighborhoods include machine learning analysis of Airbnb Listing text and images to predict the Listing’s neighborhood and price (Tang & Sangani, 2016), and an Airbnb-NYU collaboration on longitudinal use trends and profitability at the neighborhood level, which found that from

2011-2016 more Listings dispersed to outlying areas of New York City (Coles et al, 2017).

At a broader level of research, there are numerous clustering methods to identify the spatial footprints of imprecise regions such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN), kernel density estimation, and machine learning with support vector machines (SVM). A detailed study on clustering methods is given by Xu and Tian (2015). These methods have been combined with social media data to delineate the boundaries of European countries using geotagged photos (Grothe & Schaab, 2009), locate ambiguous placenames such as the Scottish Highlands using web scrapes (Jones et al, 2008), and investigate the boundaries of “northern” and “southern” California (Gao et al, 2017). This thesis uses similar methodology to Gao et al (2017), including grouping data by hexagonal grids and counting occurrences to derive membership value, and using DBSCAN point clustering and convex hull creation to create polygon maps. Gao et al (2017) concluded that attitude is a more important component to spatial identity than exact physical location: an area can be further north than San Francisco and be deemed less “northern California”.

2.3 Point of Interest Identification

A Point of Interest (POI) is a geographic feature that occupies a specific point. POIs are important because they represent places of significance or utility to a user, and POIs form the foundational unit for location-based applications. Although the term POI has a broad range of uses, in the context of this thesis a POI refers to a named point of some significance, such as a landmark, as opposed to an unnamed POI category such as restaurants, bars, or coffee shops.

POI identification is typically accomplished using similar methods to those for identifying regions: via spatial clustering algorithms such as k-means, mean shift, or DBSCAN, through spectral clustering (graph theory), or machine learning. POI discovery has been applied extensively to Flickr and other geotagged photo data (Vasardani et al, 2013). Approaches include scalable mean-shift clustering and SVM at global scale (Crandall et al, 2009), self-tuning spectral clustering without need to specify parameters (Yang & Gong, 2011, 2015), grid-cell based geolocation (O'Hare & Murdock, 2013), and analysis of camera orientation (Lacerda et al, 2012).

DBSCAN has been used without modification to identify POIs in Australia (Lee et al, 2013), or adapted as with the P-DBSCAN variant, which incorporates adaptive density as well as a density threshold to correct for multiple contributions by a single user (Kisilevich et al, 2010).

Other crowdsourced methods include using Wikipedia data to train a model to automatically discover and localize POIs in Foursquare and Gowalla data (Rae et al, 2012), and combining ngram analysis and Term Frequency–Inverse Document Frequency (TF-IDF) value of names within a cluster (Mummidi & Krumm, 2008). This thesis will combine ngram analysis and DBSCAN techniques on New York City Airbnb data to identify neighborhood boundaries and discover POIs.

3. DATA

This thesis uses data from InsideAirbnb.com (Inside Airbnb), a public archive of scraped Airbnb data, and provided under a fair use claim. Inside Airbnb is the largest publicly available Airbnb dataset and is used in several notable studies in the literature. Inside Airbnb is run by New York City-based community activist Murray Cox to provide transparency to Airbnb's operations and promote a data-driven public discussion and analysis of Airbnb's impact on communities.

Inside Airbnb has monthly data scrapes of all data publicly available for New York City dating back to March 2015 and provides smaller datasets from 43 other cities worldwide. Publicly available data in an Airbnb Listing are any data visible to a prospective Guest browsing Airbnb and does not include any additional information provided after a booking is made, such as addresses or arrival instructions. The data are extensive; 96 fields of publicly available attributes are collected.

This thesis uses the October 2017 scrape for New York City, which includes 44,317 geolocated records (Inside Airbnb, 2017). Figure 4 shows a heat map of the Listings.[1] Listings are densely grouped in Mid- and Lower Manhattan and Williamsburg, and sparse in outer neighborhoods which are less popular with travelers.

[1] Maps in this thesis are displayed in EPSG 2263 NAD83 / New York Long Island (ftUS), which is the projection used by New York City agencies which produce GIS products. The base layer is the Stamen Terrain Openlayers plugin.

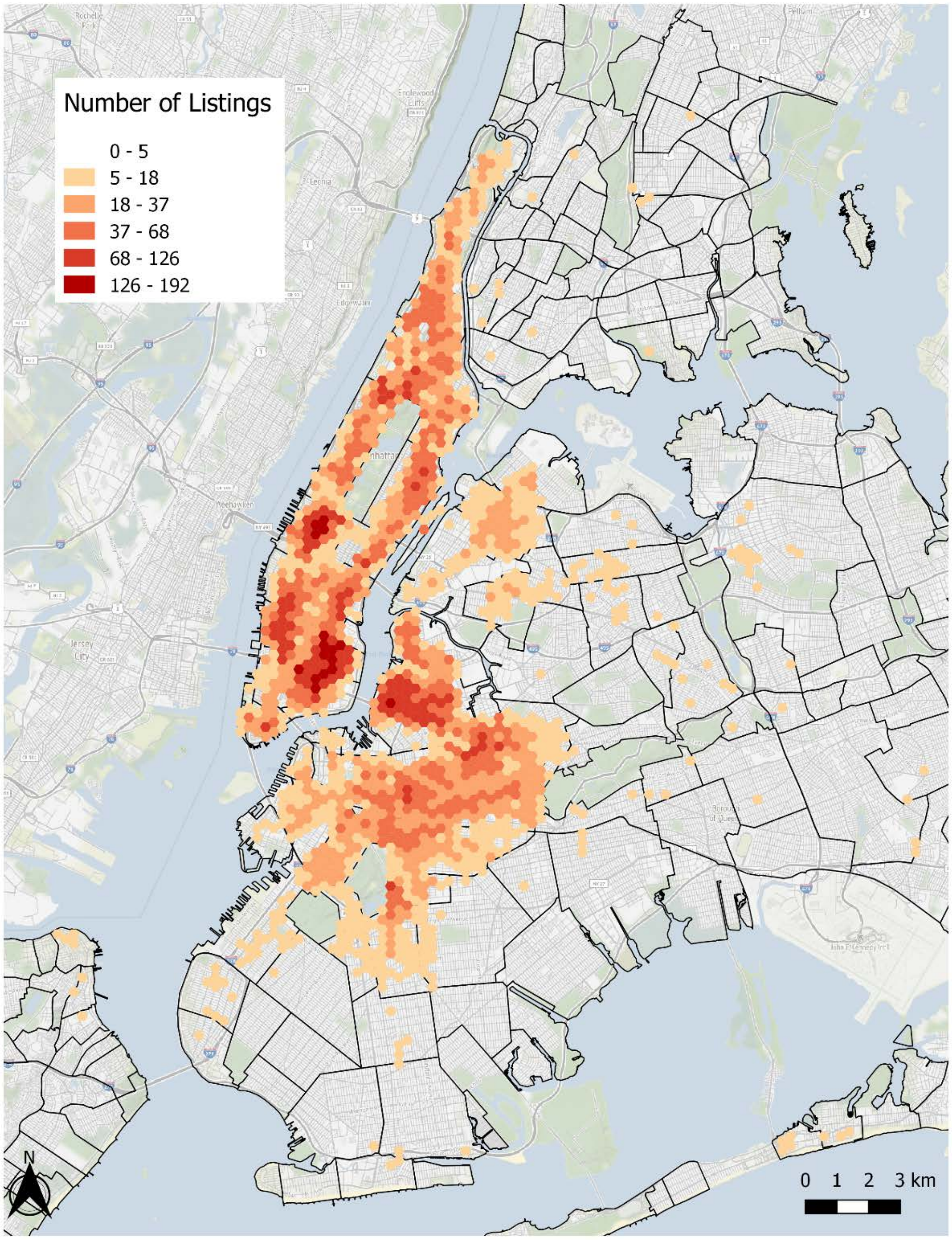


Figure 4 Heatmap of New York City October 2017 Airbnb Listing locations.

4. NEIGHBORHOOD IDENTIFICATION

The Inside Airbnb data include three neighborhood fields: the neighborhood as identified by Airbnb, and two fields assigned by Inside Airbnb post-collection via spatial join: the actual New York City neighborhood boundaries, and the group level neighborhood (the five boroughs: Manhattan, Brooklyn, Queens, the Bronx, and Staten Island). As there is no official reference source for New York City neighborhood boundaries (NYC Department of City Planning, 2018), this study uses the Pediacities neighborhood boundaries produced by NYC-native GIS developers (BetaNYC, 2015).

Airbnb's own neighborhood definitions are not accurate (Inside Airbnb, n.d.). The Airbnb neighborhood field is assigned based on the Listing address using Airbnb's own neighborhood definitions. Regardless of the neighborhood assigned, only the Borough-level name is named in the Listing. Hosts cannot change the displayed neighborhood, and the neighborhood name is not visible to the Host when they are prompted to write the Neighborhood Overview field. Neighborhood-based search is available to end users but is not a prominent feature of Airbnb's search interface.

Instead of using neighborhood fields, this thesis extracts references to known neighborhood names in the Neighborhood Overview field, which is written by the Host. 25,149 (58%) of the scraped New York City Listings include the Neighborhood

Overview field. Table 1 shows an example of Neighborhood Overview data, with identified neighborhood strings in bold.

Table 1 Example Neighborhood Overview data

"**Williamsburg** is one of the best neighborhoods in the world. Whether you want to try any type of cuisine (Thai, Vietnamese, BBQ, Japanese, American Nouveau, French to name a few) or grab a drink at a local bar (there's so many within walking distance of the apartment) or check out McCarren Park there's a limitless amount of things to do. You might not even go into Manhattan at all!"
"Fantastic **Williamsburg** neighborhood right off the L train. Close to tons of bars/restaurants (Okonomi/YUJI Ramen, Haymaker's, Campbell Cheese). A short stroll to the **Williamsburg** waterfront or **Greenpoint**. Explore Manhattan/Brooklyn from this peaceful base."
"**Greenpoint** is a thriving artistic community with cafes, boutiques, great restaurants, and fun bars all within walking distance. Also close to **Williamsburg** which has world-class restaurants, hip nightlife, indie cinemas and galleries galore!"
"Hands down TVs best part of **Williamsburg**. Tons of bars/restaurants and just a short walk to Bedford (without the nightmare of being on top of it)."
"Close to Manhattan, **Williamsburg** is NYC's trendiest neighborhood. Surrounded by hipster bars and local restaurants, this apartment is right in the heart of **Williamsburg**--steps away from shopping, live music, good eats, and public transportation."

4.1 Preprocessing

The python Natural Language Toolkit (NLTK) package was used to sanitize the Neighborhood Overview field (Bird et al, 2009). Sanitization included tokenization, lower casing, and removal of English stopwords and punctuation, while preserving sentence structure. A stemmed version of the sanitized field was also created using the Porter stemmer. The combined Neighborhood Overview corpus consisted of 1,654,176 tokens—789,530 after stopwords were removed—and a total of 79,408 sentences.

4.2 Sentence Structure Analysis

Sentence structure was preserved because analysis of the data indicated that the neighborhood name context changes in multi-sentence Listings. When a Host names a neighborhood in the first sentence of a Listing it is referring to the Listing's location, whereas mentions only in subsequent sentences are typically references to other neighborhoods which are nearby, or popular destinations with convenient transportation options. The examples in Table 1 illustrate common examples of neighborhood placement in sentence order.

Figure 5, Figure 6, Figure 7, and Figure 8 map examples of the difference in sentences. 75% of Listings mentioning Williamsburg in the first sentence of the Neighborhood Overview are located within the yellow boundary of Williamsburg (Figure 5), while only 2.5% of Listings mentioning Williamsburg in only second or subsequent sentences are inside the yellow boundary (Figure 6). Bushwick, the second largest neighborhood by Listing count, also clearly displays this difference between first sentence (68.1%) and subsequent sentence mentions (7.1%) (Figure 7 and Figure 8).

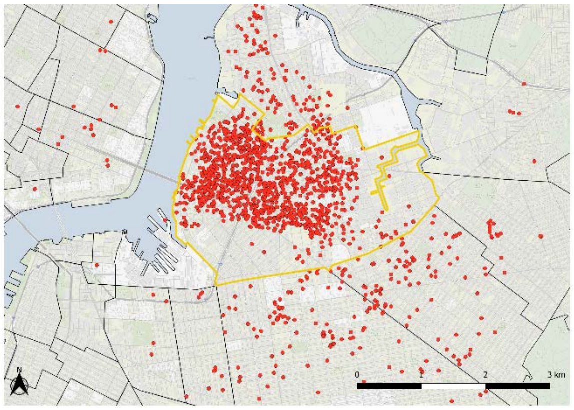


Figure 5 Listings with Williamsburg in the first sentence of the Neighborhood Overview.

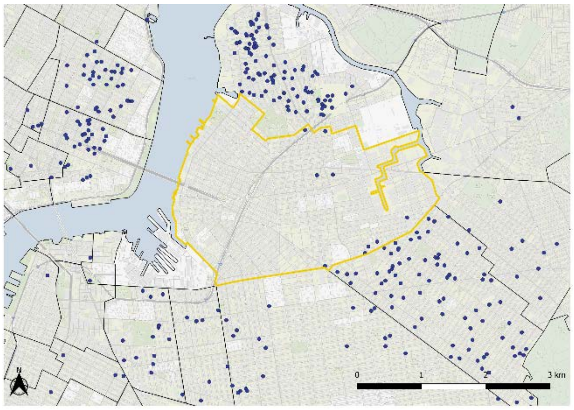


Figure 6 Listings with Williamsburg in only the second and subsequent sentences of the Neighborhood Overview.

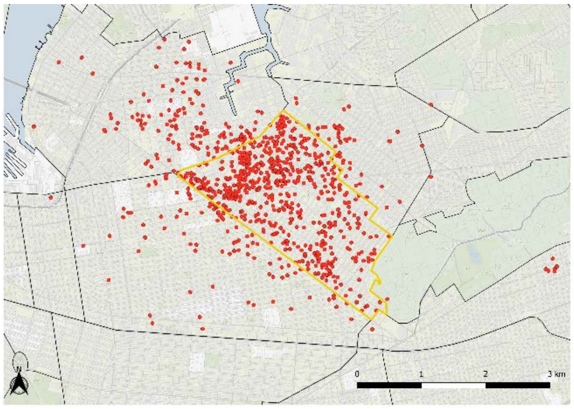


Figure 7 Listings with Bushwick in the first sentence of the Neighborhood Overview.

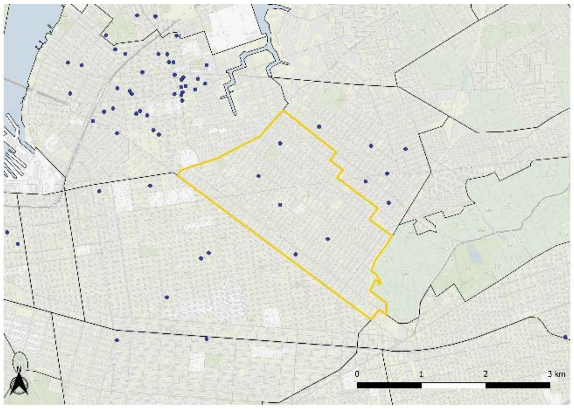


Figure 8 Listings with Bushwick in only the second and subsequent sentences of the Neighborhood Overview.

This analysis was extended to other neighborhoods to determine if the pattern is consistent. Table 2 displays the 25 neighborhoods with the most identified Listings. For each neighborhood, the Table includes the number of Listings mentioning it in the first sentence and the percentage of those Listings which are within the actual neighborhood boundary. The average for all neighborhoods with 10 or more total points is 58.7%, with a standard deviation of 20.7%. The table also shows the number of Listings mentioning the neighborhood in subsequent sentences (i.e. not the first sentence) and the percentage inside the actual boundary. The subsequent sentence Listings are only rarely present within the actual boundary: the average for all neighborhoods with 10 or more total points is 4.3%, with a standard deviation of 3.9%. Figure 9 and Figure 10 show the histograms or percentages for neighborhoods with 10 or more total points.

Several of the neighborhoods have a small percentage of first sentence Listings within the boundary, which indicates that the actual boundary does not necessarily align with the popular perception of these neighborhood locations. Midtown is commonly used in a broader context: it is considered to be the entire central lengthwise section of Manhattan rather than the demarcated neighborhood. Many of the Park Slope Listings are in the adjacent South Slope neighborhood, which can also be considered part of Park Slope. Soho is a popular neighborhood, and the Chinatown and Little Italy neighborhood polygons have small footprints. This thesis therefore identifies neighborhoods using only Listings which refer to the neighborhood in the first sentence of the Neighborhood Overview. Listings which mention multiple neighborhoods in the first sentence are included in all of those neighborhoods.

Table 2 Neighborhood boundary analysis

Rank	Neighborhood	First Sentence Listings	% in Boundary	Subsequent Sentence Listings	% in Boundary
1	Williamsburg	1561	75.0%	361	2.5%
2	Bushwick	903	68.1%	70	7.1%
3	Harlem	863	68.9%	36	16.7%
4	East Village	625	80.8%	160	4.4%
5	Soho	389	31.1%	343	1.7%
6	Midtown	467	30.8%	197	2.0%
7	Lower East Side	420	50.2%	176	4.5%
8	West Village	339	61.1%	179	2.2%
9	Park Slope	426	40.8%	90	0.0%
10	Greenpoint	395	86.8%	72	9.7%
11	Chelsea	391	81.1%	55	5.5%
12	Astoria	418	61.7%	23	4.3%
13	Chinatown	235	28.5%	178	8.4%
14	Upper East Side	354	84.2%	31	3.2%
15	Upper West Side	331	87.3%	13	7.7%
16	Fort Greene	245	57.1%	91	2.2%
17	Crown Heights	307	86.0%	28	7.1%
18	Clinton Hill	272	48.9%	29	0.0%
19	Little Italy	153	14.4%	132	5.3%
20	Nolita	141	47.5%	76	2.6%
21	Greenwich Village	135	54.1%	77	0.0%
22	Prospect Heights	171	53.2%	29	0.0%
23	Gramercy	129	43.4%	46	2.2%
24	Tribeca	103	41.7%	61	0.0%
25	Long Island City	99	73.7%	42	7.1%

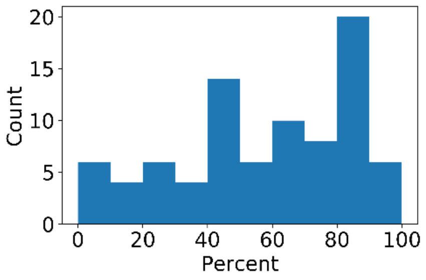


Figure 9 Histogram of percent of first sentence Listings within the neighborhood boundary.

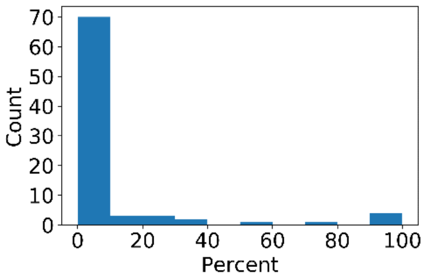


Figure 10 Histogram of percent of subsequent sentence Listings within the neighborhood boundary

4.3 Network Analysis

Listing sentence structure was further explored using network analysis. Figure 11 shows an excerpt of a directed network graph of 23,354 edges constructed from neighborhood mentions in the Listings. Graph nodes are grouped using the Gephi Geolayout plugin (they are geolocated to the centroid of the actual neighborhood polygon) and color-coded by borough. Nodes are scaled by in-degree: neighborhoods which are the target of more edges are larger. The graph illustrates that popular neighborhoods such as Williamsburg, Soho, Bushwick, the East Village, and Midtown have the largest in-degree.

Each edge is weighted for the number of Listings which reference the source neighborhood in the first sentence and the target neighborhood in a second or subsequent sentence. Edges are read in a clockwise direction: for example, Greenpoint Listings frequently mention Williamsburg, but Williamsburg Listings reference Greenpoint far less frequently. Self-edges are not included. The graph illustrates that Listings most frequently reference other nearby neighborhoods, and there are few strong connections between distant neighborhoods.

The full graph has an average weighted degree of 59.3, average path length of 2.9, and a network diameter of 9. The average clustering coefficient 0.373 (probability that two neighbors of a randomly selected node are themselves neighbors) indicates that the Graph is not highly clustered.

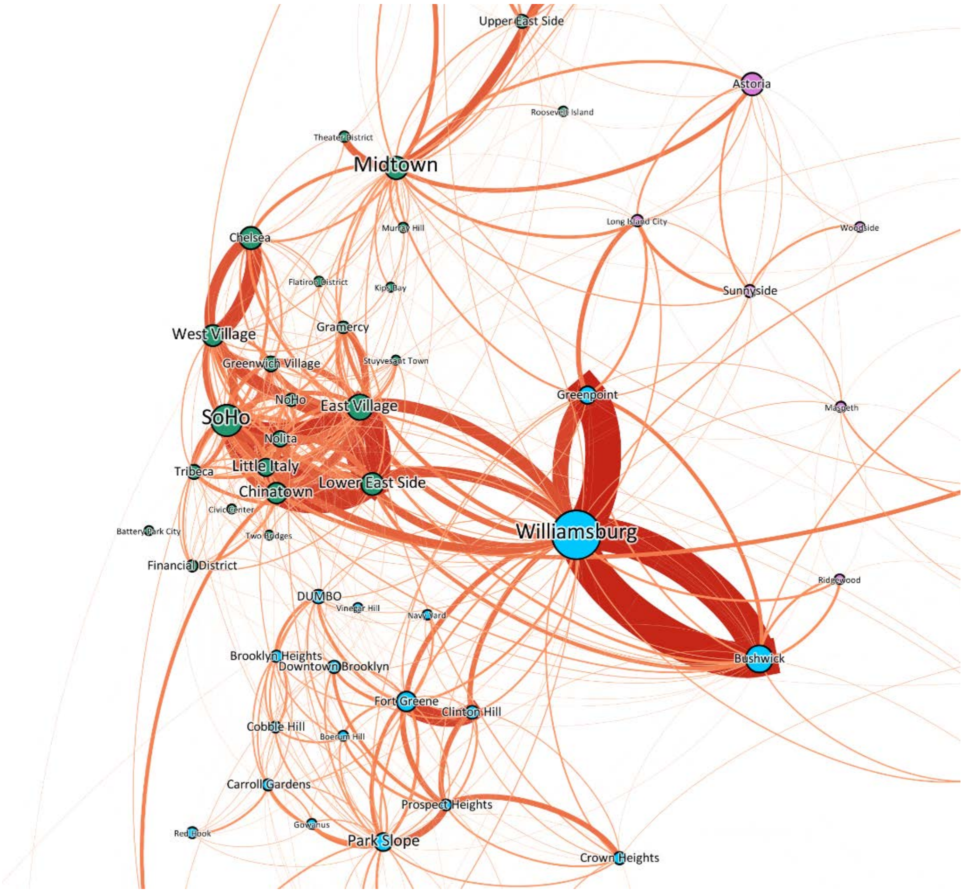


Figure 11 Network graph of neighborhood cross-mentions.

Table 3 lists the top 25 Neighborhoods by total degree. Soho, Little Italy, Fort Greene, and Tribeca have a high indegree-outdegree ratio, indicating that these popular neighborhoods receive disproportionately more mentions from other neighborhoods. The Upper West Side, Prospect Heights, and the Upper East Side have low indegree-outdegree ratios, indicating that these neighborhoods are more likely to discuss other neighborhoods. Figure 12 and Figure 13 show histograms of the neighborhood indegrees and outdegrees, which demonstrate power-law distribution.

Table 3 Neighborhood network analysis

Rank	Neighborhood	Indegree	Outdegree	Indegree / Outdegree	Eigenvector Centrality	Betweenness Centrality
1	Williamsburg	1705	1978	0.86	0.69	664.99
2	Harlem	1020	1207	0.85	0.26	119.94
3	Bushwick	803	1116	0.72	0.33	108.67
4	East Village	670	1052	0.64	0.54	241.23
5	SoHo	936	533	1.76	0.65	142.34
6	Lower East Side	527	786	0.67	0.49	32.24
7	Midtown	626	565	1.11	0.66	424.05
8	Chelsea	527	597	0.88	0.42	46.40
9	Astoria	512	608	0.84	0.15	24.05
10	Park Slope	392	550	0.71	0.45	177.15
11	West Village	518	410	1.26	0.52	65.69
12	Greenpoint	320	539	0.59	0.28	36.45
13	Chinatown	457	364	1.26	0.51	250.06
14	Fort Greene	463	356	1.30	0.36	47.04
15	Upper East Side	188	389	0.48	0.31	19.61
16	Clinton Hill	207	352	0.59	0.26	6.71
17	Little Italy	339	216	1.57	0.49	271.51
18	Crown Heights	183	348	0.53	0.11	19.11
19	Nolita	248	278	0.89	0.38	7.41
20	Upper West Side	121	362	0.33	0.16	16.33
21	Greenwich Vill.	225	227	0.99	0.48	24.37
22	Prospect Heights	101	260	0.39	0.27	32.12
23	Tribeca	203	158	1.28	0.44	33.52
24	Gramercy	136	211	0.64	0.29	21.76
25	Flatbush	160	132	1.21	0.14	25.03

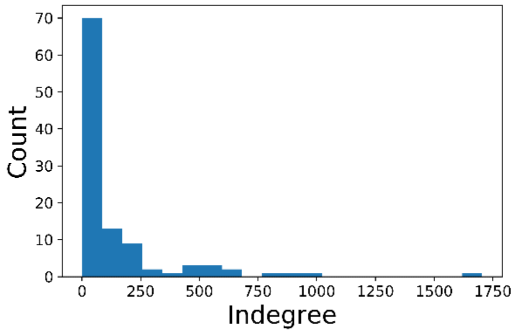


Figure 12 Histogram of neighborhood indegrees.

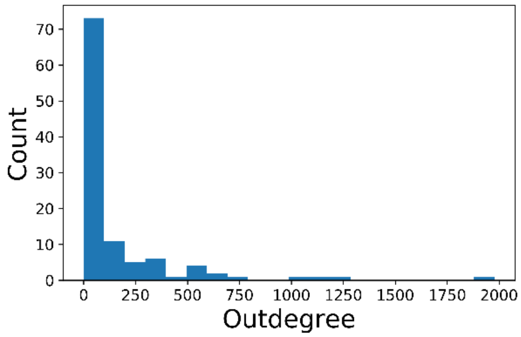


Figure 13 Histogram of neighborhood outdegrees.

Eigenvector centrality indicates that Williamsburg, Midtown, and Soho are the most strongly connected to other high degree neighborhoods. Although Harlem has a high indegree, it has a low Eigenvector centrality because it is not strongly connected to other high-scoring neighborhoods. Peripheral neighborhoods Astoria, Crown Heights, and Flatbush likewise have low Eigenvector centrality.

Betweenness centrality indicates how many shortest paths pass through the neighborhood. Williamsburg and Midtown have high betweenness centrality and are key nodes, while small central neighborhoods such as Clinton Hill and Nolita have low betweenness centrality.

4.4 Clustering Analysis

A clustering algorithm was used to remove noise points and identify neighborhood boundaries. For each neighborhood, all Listings mentioning the neighborhood in the first sentence of the Neighborhood Overview were selected. The Scikit-learn DBSCAN algorithm was applied to the selected Listings with Min_pts = 5 and epsilon = 500m to produce clusters for the neighborhood. Each Listing was assigned to a cluster if there were 4 or more other Listings within 500 meters. Epsilon of 500 meters was initially chosen to approximate the median neighborhood size of 0.35 square miles reported by survey respondents in Coulton et al (2013). A circular neighborhood with an area of 0.35 square miles (0.91 square kilometers) would have a radius of 537 meters.

Figure 14 shows the results of the DBSCAN algorithm on Listings referring to Williamsburg in the first sentence of the Neighborhood Overview. DBSCAN identified two clusters: a large main cluster overlapping the actual boundary of Williamsburg, and a small cluster in lower Manhattan. While most of the main cluster Listings are within the actual boundary, the cluster bleeds into other neighborhoods to the north (Astoria) and south (Bushwick and Bedford-Stuyvesant).

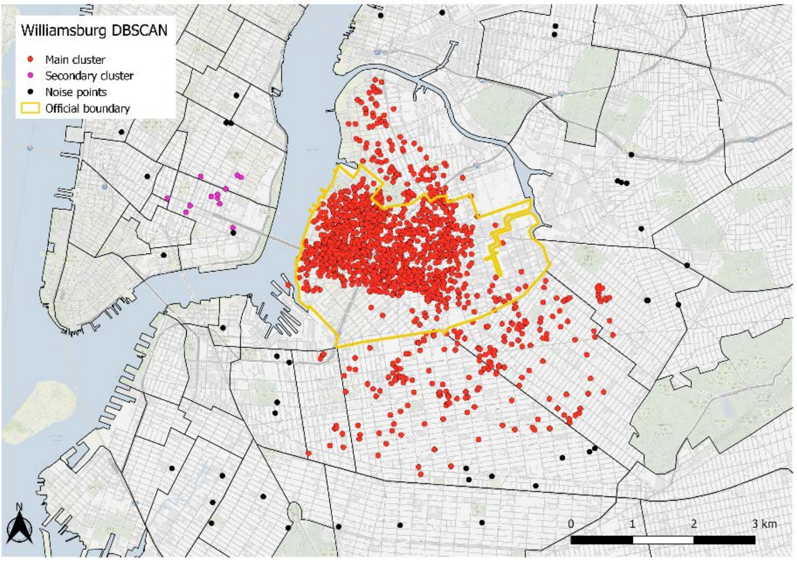


Figure 14 Williamsburg initial DBSCAN results.

The small secondary cluster is in Manhattan, and the Listings are in fact referring to the Manhattan terminus of the Williamsburg Bridge. The initial parameters also classified too many outlying Listings as part of the main cluster rather than noise points. These results indicated that the DBSCAN parameters were too lenient, so DBSCAN was rerun with Min_pts = 10 and epsilon = 250m to create more tightly defined neighborhoods. These new parameters were chosen to eliminate observed secondary clusters.

Each cluster was then converted to a polygon using the QGIS Convex Hull operation, which constructs a polygon containing all of the cluster Listings. Figure 15 shows the results of the Convex Hull operation for all neighborhood clusters. Polygons are overlapped in Manhattan and Brooklyn where many neighborhoods closely coexist. However, the clustering process did not identify many neighborhoods in outlying regions where data are sparse and the minimum points threshold isn’t met. Note that the popular Bedford-Stuyvesant neighborhood is not identified: residents more commonly refer to it using the nickname Bed-Stuy, which was not included in the neighborhood list for preprocessing.

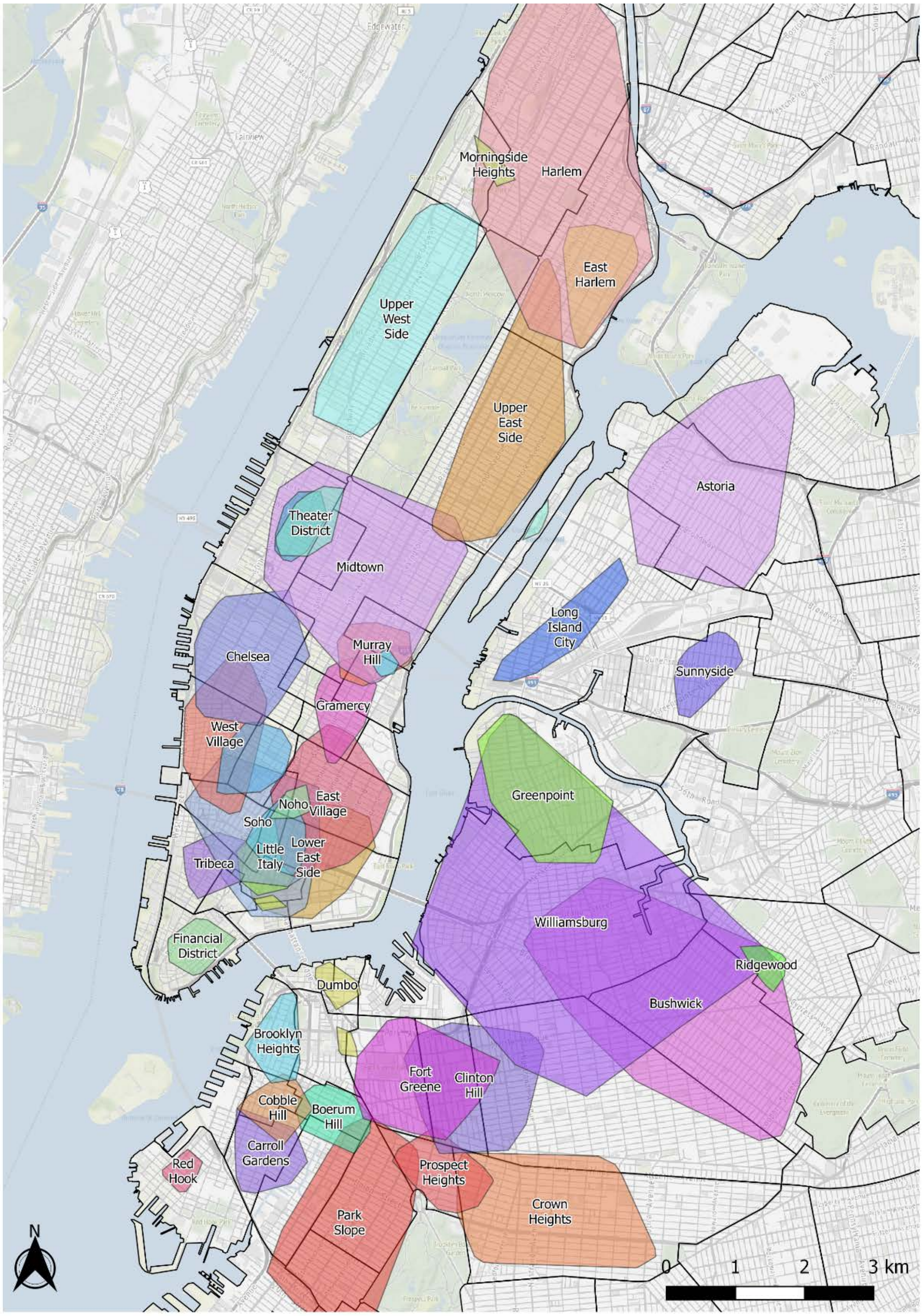


Figure 15 Identified neighborhood polygons.

To quantify the similarity of the actual and identified polygons, a spatial intersect operation was performed to find the area of overlap of each polygon pair. The intersect area was then compared to both the area of the actual polygon and the area of the identified polygon to compute overlap percentages. Table 4 shows the computed values for the 25 largest neighborhoods. For example, the identified polygon for Park Slope fully encompasses the actual polygon. The intersect area is therefore 100% of the actual polygon, but only 42% of the larger and more expansive identified polygon. Although the average identified neighborhood polygon size is 2.1 square kilometers, the median size is 0.87 square kilometers, which is very close to the median neighborhood size of 0.91 square kilometers reported by Coulton et al (2013).

Figure 16 and Figure 17 are histograms of the overlap percentages of the actual and identified polygons with the intersect polygons. Both distributions are bimodal. Local maxima for actual polygon overlaps are below 10% and above 80%, and local maxima for identified polygons are at 50% and above 90%.

Table 4 Neighborhood polygon comparisons

Rank	Neighborhood	Intersect Area (km^2)	Actual Area (km^2)	Intersect /Actual	Identified Area (km^2)	Intersect /Identified
1	Williamsburg	7.77	7.86	99%	15.38	51%
2	Harlem	5.36	5.99	90%	9.58	56%
3	Bushwick	4.46	4.74	94%	9.12	49%
4	Midtown	3.10	3.56	87%	6.14	50%
5	Astoria	3.33	4.72	71%	5.33	63%
6	Park Slope	1.76	1.77	100%	4.22	42%
7	Upper E. Side	3.41	4.54	75%	4.08	84%
8	Crown Heights	3.69	5.99	62%	4.03	92%
9	Upper W. Side	3.58	5.05	71%	3.60	99%
10	Clinton Hill	1.32	1.43	93%	2.95	45%
11	SoHo	0.80	0.91	87%	2.68	30%
12	East Village	1.48	1.83	81%	2.64	56%
13	Greenpoint	2.42	4.44	54%	2.61	93%
14	Fort Greene	1.35	1.63	83%	2.54	53%
15	Chelsea	2.00	2.74	73%	2.48	80%
16	Lower E Side	1.12	1.85	61%	2.33	48%
17	Wshngtn Hts	2.08	4.49	46%	2.08	100%
18	West Village	1.11	1.33	83%	1.96	57%
19	Chinatown	0.50	0.50	100%	1.38	37%
20	East Harlem	1.25	3.88	32%	1.25	100%
21	Long Isl. City	1.09	8.28	13%	1.09	100%
22	Prospect Hts	0.66	0.81	82%	1.05	63%
23	Carroll Gdns	0.71	0.93	77%	0.94	75%
24	Gramercy	0.39	0.59	67%	0.89	44%
25	Greenwich Vill.	0.60	0.75	80%	0.85	71%

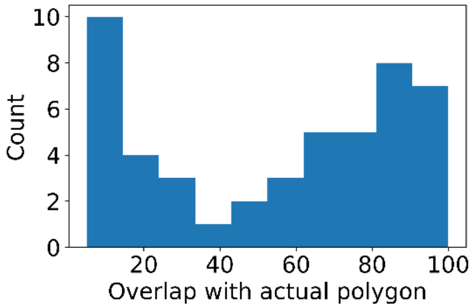


Figure 16 Histogram of percent overlap of the intersect polygon with the actual polygon.

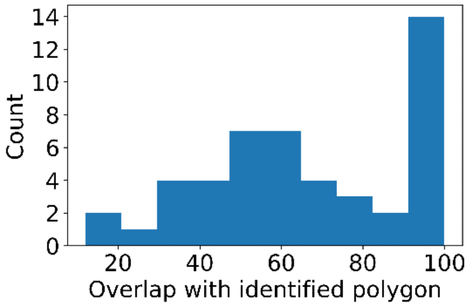


Figure 17 Histogram of percent overlap of the intersect polygon with the identified polygon

Figure 18 plots the percentage overlap of all 48 identified neighborhood polygons. Each point is a neighborhood, scaled by the size of the actual polygon. The figure is divided into four labeled quadrants to aid interpretation. Most of the identified polygons are either Smaller (wholly within) or of a Similar size to the actual polygons. Few identified polygons were either Larger (encompassing the actual area) or wholly Dissimilar (low overlap percentages). Scaling reveals that the Smaller identified polygons are often within large actual polygons, and Larger identified polygons often encompass small actual polygons.

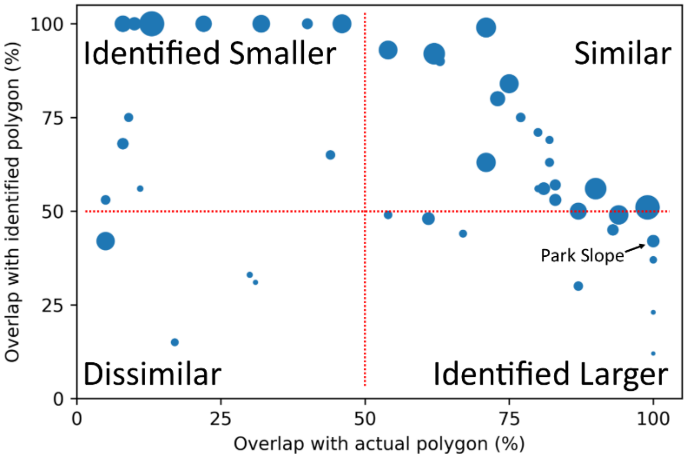


Figure 18 Intersect overlap comparison: identified vs. actual polygon.

Figure 19 presents examples for each of the quadrants. The darker section of each polygon is the intersection area of the identified polygon and the actual polygon. Brooklyn Heights, Cobble Hill, Boerum Hill, Fort Greene, and Carroll Gardens are all classified as Similar. Dumbo is Dissimilar; the identified polygon is mostly within the Vinegar Hill neighborhood. Red Hook is Smaller, and as previously discussed, Park Slope is Larger.

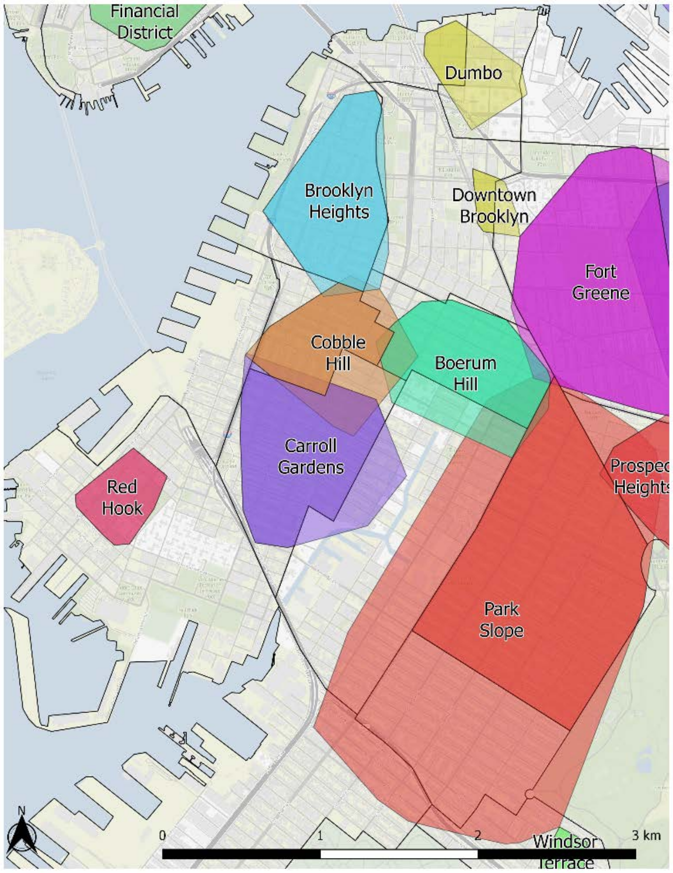


Figure 19 Intersect quadrant examples.

4.5 Hex Assignment

Defining neighborhoods as overlapping polygons means that it is possible for a location to be in multiple neighborhoods, which is often not desirable for urban planning purposes. An alternative method of neighborhood construction is by hex assignment. Hex assignment uses a winner-take-all approach to assign each area of the map to a neighborhood and avoid overlaps, but at some loss of spatial resolution.

To create the hex map in Figure 20, a grid of hexagons with side lengths of 175 meters was created overlaying the map extent. This hexagon size (0.06 square kilometers) was chosen after visual inspection of hexes of 0.53 square kilometers and 2 square kilometers. The larger hex sizes did not allow sufficient resolution to distinguish smaller neighborhoods.

A spatial join was performed on the Listings and hexes to assign each Listing to a hex. A python script was used to count the first sentence neighborhood mentions for all Listings within each hex and assign the hex to the neighborhood name with the highest sum. Hex color was then scaled at count thresholds of 5, 2, and 1 to better show neighborhood density.

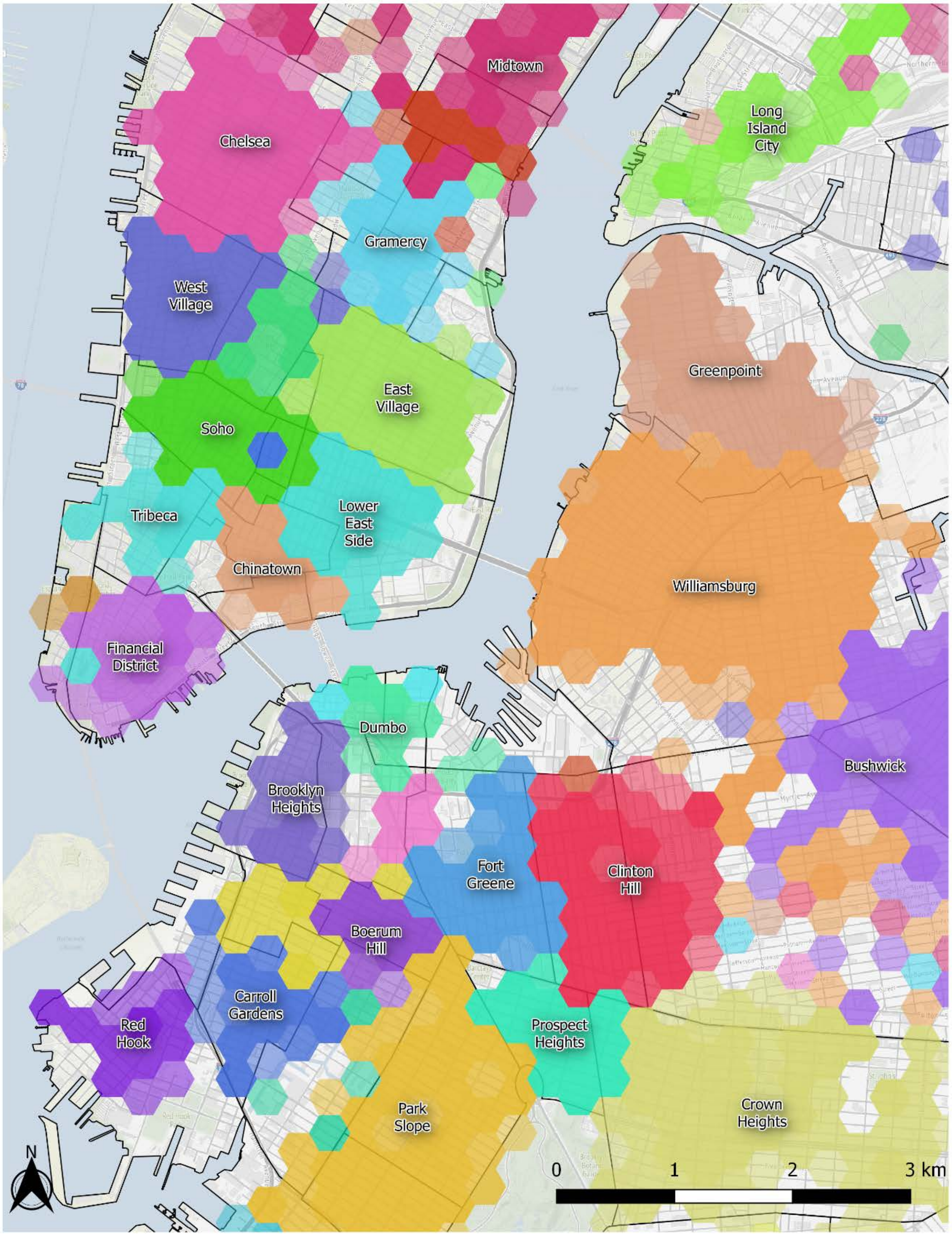


Figure 20 Neighborhood hex assignment.

5. POINTS OF INTEREST DISCOVERY

The Neighborhood Overview field of an Airbnb Listing contains frequent references to POIs such as landmarks and parks (examples given previously in Table 1). By combining ngram analysis and clustering methods, unknown POIs in the dataset can be discovered and geolocated.

5.1 Preprocessing

First, additional data sanitization was performed on the dataset to remove the previously analyzed neighborhood names. Chunking using the NLTK ngrams function on the stemmed, tokenized Neighborhood Overview field identified 710,520 unique bigrams and 632,504 unique trigrams.

Table 5 and Table 6 show the top 10 stemmed bigrams and trigrams. Distance measures ("block away") and landmarks ("central park") are evident.

Table 5 Stemmed bigrams

Rank	Bigram	Count
1	block away	3687
2	walk distanc	3606
3	new york	3244
4	minut walk	3043
5	restaur bar	3031
6	central park	2578
7	coffe shop	2218
8	bar restaur	2120
9	prospect park	1940
10	groceri store	1844

Table 6 Stemmed trigrams

Rank	Trigram	Count
1	within walk distanc	1357
2	new york citi	1184
3	lower east side	798
4	5 minut walk	578
5	10 minut walk	570
6	brooklyn botan garden	502
7	one block away	500
8	upper east side	471
9	great restaur bar	443
10	2 block away	419

5.2 Discovery

The DBSCAN clustering algorithm was used to identify which of the common ngrams are highly spatially clustered and likely to be a unique POI. The Scikit-learn DBSCAN algorithm was automated to run on the Listings for each of the top 1,000 bigrams and top 1,000 trigrams, with Min_pts = 10 and eps = 500m. If the DBSCAN results produced a non-noise cluster containing greater than 75% of the points, the ngram was identified as a POI. With this method, unique POIs can be distinguished from common ngrams such as “coffee shops”. Figure 21 is a histogram of the calculated percentage of points in the largest non-noise cluster for each ngram. Most of the ngrams tested did not contain non-noise clusters. The parameter of 75% was chosen based on observation of an increase in counts at and above 75%: these bins contain 10.75% of the data (prior to removal of neighborhood ngrams).

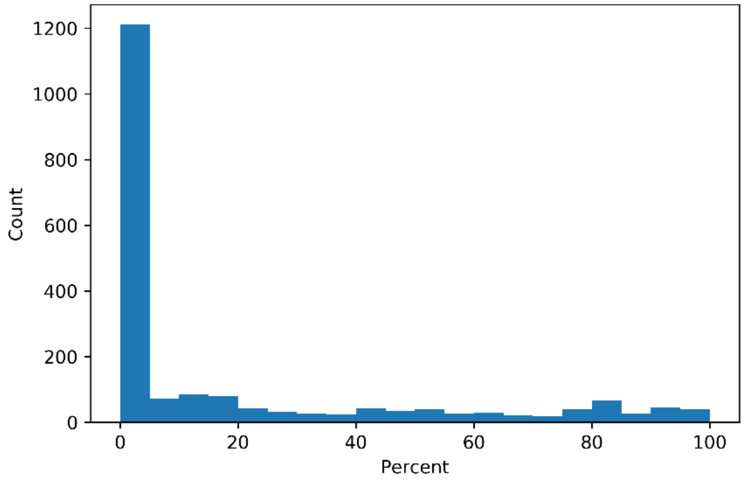


Figure 21 Histogram of percent of points in largest non-noise cluster.

86 of top 1000 bigrams and 103 of the top 1000 trigrams were identified as POI candidates. Figure 22 shows a selection of identified POI candidate clusters. Colombia University, Central Park, Barclay Center, and Prospect Park are correctly identified, as is the route of the “L” subway through Brooklyn. Of the 189 identified POI candidates, 78 correctly represented POIs, 33 were variants of known neighborhoods which weren’t screened (“central Harlem”, “east Williamsburg”), and 9 represented linear areas of interest such as the High Line Park. 69 of the candidates were non-POI descriptive text (“street art”, “vibrant night life”) or could not be matched with a known POI (“square park”). The 78 matched POIs included 32 parks, 19 cultural sites (theaters, museums, and churches), 11 landmarks, 7 restaurants, and 6 academic institutions.

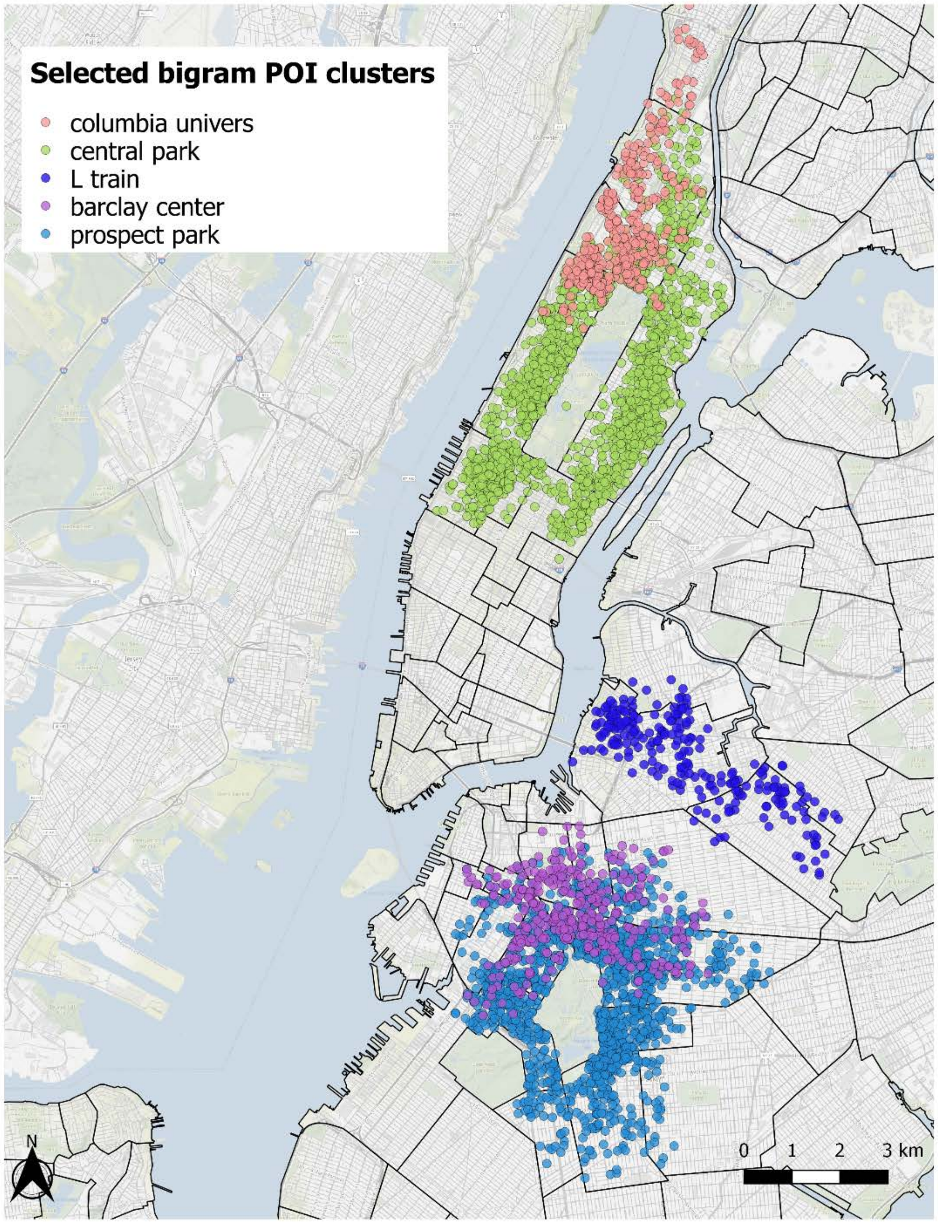


Figure 22 Examples of ngram DBSCAN clustering POI results.

5.3 Geolocation

The mean center of each POI cluster was mapped and compared to the known location of the corresponding landmark from NYC OpenData's lists of Points of Interest and Areas of Interest, which are compiled from across multiple New York City agencies

(NYC OpenData, n.d.). Figure 23 shows the identified POI locations in red, connected by a line to the corresponding actual POI location in green. Identified POI locations were on average 438 meters away from the actual POI location, with a standard deviation of 321 meters. The Empire State Building was the most accurately located POI at a distance of 24 meters, and Brooklyn Bridge Park was the most inaccurately located POI at 1278 meters.

POI locations were less accurate on the edge of residential areas, such as coastal features or features on the boundaries of large parks. In such cases the contributing Listings are to one side of the POI, and the cluster mean center is pulled in that direction. For example, nine out of twenty identified POI locations for parks were located within the park boundary, such as Central Park, but none of the seven coastal parks' identified locations were within the park boundary.

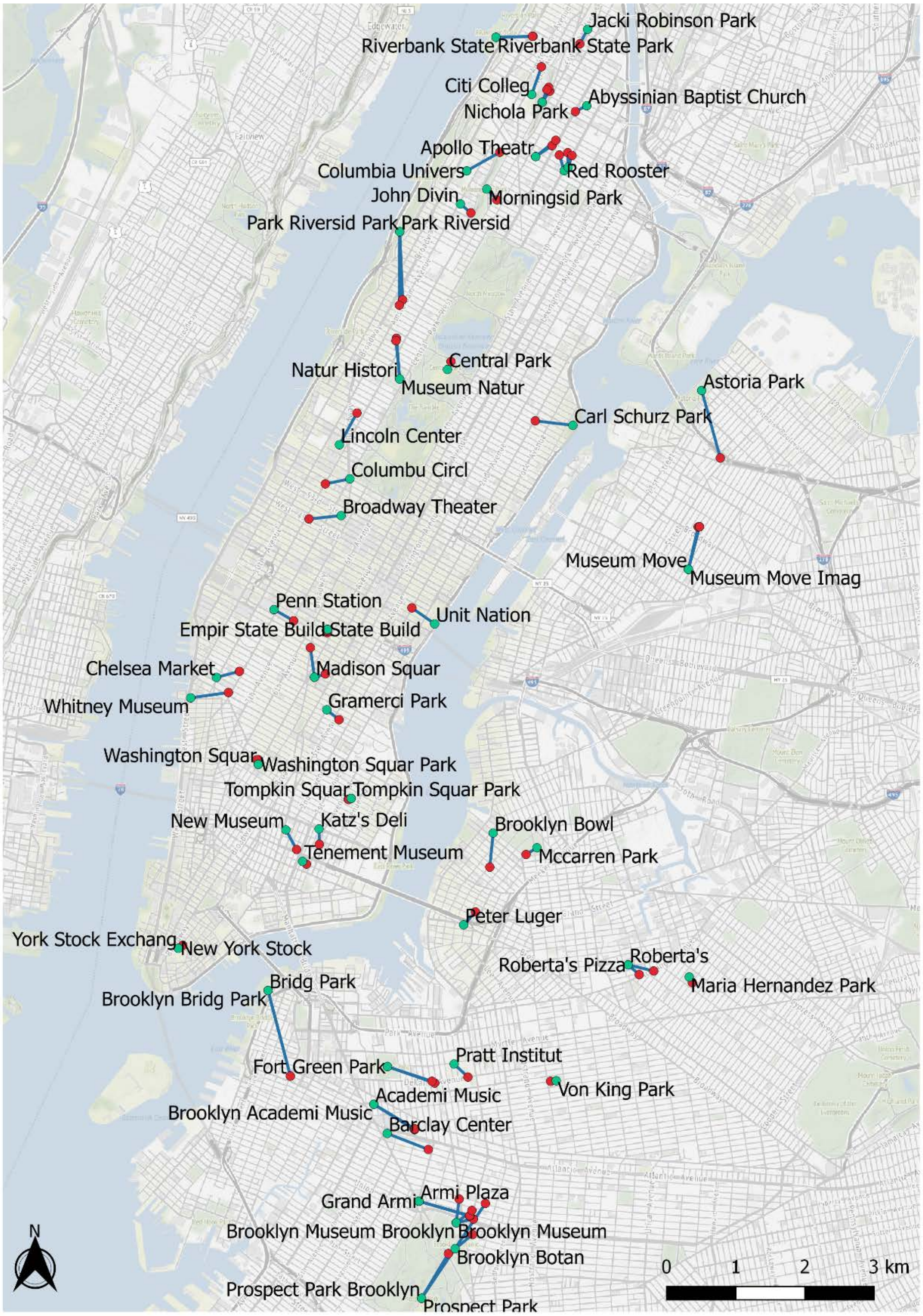


Figure 23 Identified POIs (red) connected to the actual POI locations (green).

6. DISCUSSION

Airbnb data provides several useful avenues for investigating residents' perceptions of their city and the nature of neighborhoods. Neighborhoods are challenging to delineate: there is rarely consensus on neighborhood boundaries, and they slowly evolve over time. Airbnb Listings afford researchers a new source to capture resident conceptions of neighborhood locations at scale. This study shows that Airbnb data can be used to both construct reasonably accurate neighborhood maps and geolocate POIs where Listings are sufficiently dense.

Care must be taken to correctly identify a Listing's neighborhood. Sentence structure analysis is significant because it allows a Listing's neighborhood to be assigned based only on text in the neighborhood overview. Interpreting the context of a neighborhood name is critical to determining whether the Host is referring to their own neighborhood or a different one. This knowledge is then applied in combination with clustering methods to produce neighborhoods polygons.

The produced polygons overlap each other, which illustrates a key challenge in the study of neighborhoods: boundaries between neighborhoods are often fuzzy. Neighborhoods do not follow geospatial topological rules, and a single location may belong to multiple adjacent or nested neighborhoods. Map construction by hex assignment resolves problems with overlap but loses some data fidelity.

Network analysis of the ties between neighborhoods confirms Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). Airbnb Listings refer to near neighborhoods more frequently than distant neighborhoods. Network analysis also provide insight into neighborhood popularity: neighborhoods with low indegree/outdegree ratios mention other neighborhoods more frequently than they are mentioned.

Airbnb hosts make frequent reference to POIs in the neighborhood overview field. POIs often represent locations which are perceived to be attractive to potential Guests; not necessarily POIs to the Host, but what the Host would consider to be POIs to a Guest. Ngram and clustering analysis can be used to identify POIs, however, less than half of the identified POI candidates could be matched to real-world POIs. Additional analysis would be needed to distinguish and delineate linear features or areas of interest. Although this study used a list of NYC neighborhoods, it would also be possible to use the POI discovery methods to identify possible neighborhood names for polygon construction.

Airbnb's continued growth in both new and established markets provides an expanding corpus of VGI for analysis. However, there are limitations to relying on Airbnb data. Besides the aforementioned data quality aspects (Section 2.1), Airbnb is not representative data: Hosts are not a representative sample of the population at large, and the text they write is motivated by financial gain. Additionally, Listings are only present in residential areas of neighborhoods. While POIs can be identified in commercial or industrial districts, they can only be accurately geolocated when surrounded by residential areas with Listings.

7. CONCLUSIONS

New York City Airbnb Listings are a rich source of VGI. By applying text analysis and area and point location techniques to the Neighborhood Overview field, Airbnb data can be used to successfully delineate neighborhoods and discover and geolocate major POIs. Effectiveness is dependent on Listing density; this methodology does not perform well in regions of data scarcity.

Neighborhood name placement in sentence structure allows distinguishing between a Listing’s own neighborhood and other nearby neighborhoods. Network analysis shows that Listings reference nearby neighborhoods frequently and distant neighborhoods infrequently. Both DBSCAN convex hull creation and hex assignment are effective methods to construct neighborhood maps, and DBSCAN combined with ngram analysis can effectively discover and geolocate unknown POIs.

The Airbnb Neighborhood Overview field data provide many additional opportunities for study. Text analysis could be expanded to characterize each neighborhood in terms of attractions and descriptors. For example, which neighborhoods or areas are described as walkable, which are quiet, and which are lively?

Many of the most frequent ngrams are distance measures quantified in terms of travel time (minutes) or distance (blocks). By extracting the direct object, it should be possible to find the actual distance between the Listing and a POI and quantify host consensus definitions of proximity and distance. POI analysis could also be expanded to examine the relationship of POI ngram placement within the Neighborhood Overview and determine if a POI's placement within the Listing's sentence structure carries the same significance as a neighborhood's placement.

The size of the identified neighborhood polygons could also be studied to determine if there is a relationship between neighborhood size and population density or Listing density. Manhattan is more densely populated than Brooklyn and the other outlying boroughs, and it is expected that neighborhoods in Manhattan would be smaller.

Neighborhood analysis could also be combined with additional demographic data to explore Host demographics. Given that Hosts have property with additional living space, are Hosts more affluent than the population at large, or more entrepreneurial? Is there a relationship between the demographics of a neighborhood and the density of Listings?

Longitudinal analysis of Airbnb data could be used to study the evolution and changing boundaries of neighborhoods over time, and potentially yield insights into the process of gentrification. The level of adoption of Airbnb may be indicative or predictive of gentrification and may also change in response to external factors such as regulation.

Airbnb has a global presence; although this study used New York City as an example, these POI and neighborhood identification methods could be applied to other cities, and differences between cities or countries could be analyzed and compared. The methods used in this study could easily be adapted for application to any similar source of VGI where users are asked to describe their neighborhoods.

REFERENCES

Airbnb. (2017). Terms of Service. Retrieved 2018 from https://bit.ly/2HAFktT

Airbnb. (n.d.). Fast Facts. Retrieved 2018 from https://bit.ly/2GAZ3Z5

Antoniou, V., & Skopeliti, A. (2015). Measures and indicators of VGI quality: An overview. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2*, 345.

BetaNYC. (2015). Pediacities NYC Neighborhoods. Retrieved 2018 from https://bit.ly/2JO8IgP

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Chaskin, R. J. (1997). Perspectives on neighborhood and community: a review of the literature. *Social Service Review*, *71*(4), 521-547.

Coles, P. A., Egesdal, M., Ellen, I. G., Li, X., & Sundararajan, A. (2017). Airbnb usage across New York City neighborhoods: geographic patterns and regulatory implications.

Coulton, C. J., Jennings, M. Z., & Chan, T. (2013). How big is my neighborhood? Individual and contextual effects on perceptions of neighborhood scale. *American Journal of Community Psychology*, *51*(1-2), 140-150.

Coulton, C. J., Korbin, J., Chan, T., & Su, M. (2001). Mapping residents' perceptions of neighborhood boundaries: a methodological note. *American Journal of Community Psychology*, *29*(2), 371-383.

Crandall, D. J., Backstrom, L., Huttenlocher, D., & Kleinberg, J. (2009). Mapping the world's photos. *Proceedings of the 18th International World Wide Web Conference* (pp. 761-770). Association for Computing Machinery.

Cranshaw, J., & Yano, T. (2010). Seeing a home away from the home: Distilling proto-neighborhoods from incidental data with latent topic modeling. *Neural*

Information Processing Systems Conference (Vol. 10). The Neural Information Processing Systems Foundation.

Delgado-Medrano, H.M., & Lyon, K. (2016). Short changing New York City: the impact of Airbnb on New York City's housing market. Prepared by BHJ Advisorys LLC. for Housing Conservation Coordinators Inc. and MFY Legal Services Inc.

Flanagin, A. J., & Metzger, M. J. (2008). The credibility of volunteered geographic information. *GeoJournal*, *72*(3-4), 137-148.

Fonte, C. C., Bastin, L., Foody, G., Kellenberger, T., Kerle, N., Mooney, P., ... & See, L. (2015). VGI quality control. *ISPRS Geospatial Week 2015*, 317-324.

Galster, G. (2001). On the nature of neighbourhood. *Urban Studies*, 38(12), 2111-2124.

Gao, S., Janowicz, K., Montello, D. R., Hu, Y., Yang, J. A., McKenzie, G., ... & Yan, B. (2017). A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science*, *31*(6), 1245-1271.

Girres, J. F., & Touya, G. (2010). Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS*, *14*(4), 435-459.

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*, 110-120.

Grothe, C., & Schaab, J. (2009). Automated footprint generation from geotags with kernel density estimation and support vector machines. *Spatial Cognition & Computation*, *9*(3), 195-211.

Guest, A. M., & Lee, B. A. (1984). How urbanites define their neighborhoods. *Population and Environment*, *7*(1), 32-56.

Guptill, S. C., & Morrison, J. L. (Eds.). (2013). *Elements of spatial data quality*. Elsevier.

Haeberle, S. H. (1988). People or place: Variations in community leaders' subjective definitions of neighborhood. *Urban Affairs Quarterly*, *23*(4), 616-634.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, *37*(4), 682-703.

Hollenstein, L., & Purves, R. (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, *2010*(1), 21-48.

Hunter, A. D. (1974). Symbolic communities: The persistence and change of Chicago's local communities. Chicago, Illinois: University of Chicago Press.

Inside Airbnb. (n.d.). Disclaimers. Retrieved 2017 from https://bit.ly/2HAYmQM

Inside Airbnb. (2017). Detailed Listings data for New York City. Retrieved 2017 from https://bit.ly/1NI9pGi

Jones, C. B., Purves, R. S., Clough, P. D., & Joho, H. (2008). Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, *22*(10), 1045-1065.

Kearns, A., & Parkinson, M. (2001). The significance of neighbourhood. *Urban Studies*, *38*(12), 2103-2110.

Kisilevich, S., Mansmann, F., & Keim, D. (2010). P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application* (p. 38). Association for Computing Machinery.

Lacerda, Y. A., Feitosa, R. G. F., Esmeraldo, G. Á. R. M., Baptista, C. D. S., & Marinho, L. B. (2012). Compass clustering: A new clustering method for detection of points of interest using personal collections of georeferenced and oriented photographs. *Proceedings of the 18th Brazilian Symposium on Multimedia and the Web* (pp. 281-288). Association for Computing Machinery.

Largest Airbnb Campaign to Date Tells Travellers to Live Like a Local. (2016). Little Black Book. Retrieved 2018 from https://bit.ly/2vdnB9n

Leyden, K. M., Goldberg, A., & Michelbach, P. (2011). Understanding the pursuit of happiness in ten major cities. *Urban Affairs Review*, *47*(6), 861-888.

Lee, I., Cai, G., & Lee, K. (2013). Mining points-of-interest association rules from geo-tagged photos. *Proceedings of the 2013 46th Hawaii International Conference on System Sciences*, (pp. 1580-1588). IEEE Computer Society.

Liu, J., Huang, Z., Chen, L., Shen, H. T., & Yan, Z. (2012). Discovering areas of interest with geo-tagged images and check-ins. *Proceedings of the 20th ACM*

International Conference on Multimedia (pp. 589-598). Association for Computing Machinery.

Martin. (n.d.). Williamsburg Penthouse Guestroom. Retrieved 2018 from https://bit.ly/2JNxSwm

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, *60*(3), 413-439.

Montello, D. R., Goodchild, M. F., Gottsegen, J., & Fohl, P. (2003). Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation*, *3*(2-3), 185-204.

Mummidi, L. N., & Krumm, J. (2008). Discovering points of interest from users' map annotations. *GeoJournal*, *72*(3-4), 215-227.

NYC Department of City Planning (2018). Neighborhood Tabulation Areas. Retrieved 2018 from https://on.nyc.gov/2GYXGaM

NYC OpenData. (n.d.). Points of Interest. Retrieved 2018 from https://bit.ly/2HkBh7p

O'Hare, N., & Murdock, V. (2013). Modeling locations with social media. *Information Retrieval*, *16*(1), 30-62.

Poorthuis, A. (2017). How to draw a neighborhood? The potential of big data, regionalization, and community detection for understanding the heterogeneous nature of urban neighborhoods. *Geographical Analysis*.

Rae, A., Murdock, V., Popescu, A., & Bouchard, H. (2012). Mining the web for points of interest. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 711-720). Association for Computing Machinery.

Scaggs, A. (2017). Morgan Stanley says we're reaching peak Airbnb. The Financial Times. Retrieved 2018 from https://on.ft.com/2iKBOkD

Tang, E., & Sangani, K. (2016). Neighborhood and price prediction for San Francisco Airbnb listings. Unpublished manuscript.

Tasse, D., Chou, J. T., & Hong, J. I. (2016). Generating Neighborhood Guides from Social Media. *Proceedings of the Tenth International AAAI Conference on Web and Social Media. CityLab: Technical Report WS-16-16.* Association for the Advancement of Artificial Intelligence.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*(sup1), 234-240.

Vasardani, M., Winter, S., & Richter, K. F. (2013). Locating place names from place descriptions. *International Journal of Geographical Information Science*, *27*(12), 2509-2532.

Weiss, L., Ompad, D., Galea, S., & Vlahov, D. (2007). Defining neighborhood boundaries for urban health research. *American journal of preventive medicine*, *32*(6), S154-S159.

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2), 165-193.

Yang, Y., & Gong, Z. (2011). Identifying points of interest by self-tuning clustering. *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 883-892). Association for Computing Machinery.

Yang, Y., & Gong, Z. (2015). Identifying points of interest using heterogeneous features. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5*(4), 68.

Zaleski, O. (2017). Airbnb Is Said to Double Revenue to $1 Billion Last Quarter. Bloomberg News. Retrieved 2018 from https://bloom.bg/2iVUmhO

Zhang, A. X., Noulas, A., Scellato, S., & Mascolo, C. (2013). Hoodsquare: Modeling and recommending neighborhoods in location-based social networks. *Proceedings of the 2013 International Conference on Social Computing (SocialCom),* (pp. 69-74). IEEE Computer Society.

BIOGRAPHY

Peter Thomas received his Bachelor of Arts in Global Studies: Russian and Post-Soviet Studies and Psychology from The College of William & Mary in 2007. He has worked as an international event planner, business developer, webmaster, travel blogger, and surveyor, and is currently a Software Engineer at BigBear Inc. Peter received his Graduate Certificate in Geospatial Intelligence from George Mason University in 2017 and plans to receive his Master of Science in Geoinformatics and Geospatial Intelligence from George Mason University in May 2018.