International Journal of Geo-Information

MDPI

*Article*

# A Thematic Similarity Network Approach for Analysis of Places Using Volunteered Geographic Information

**Xiaoyi Yuan [1,*], Andrew Crooks [1,2] and Andreas Züfle [2]**

[1] Department of Computational and Data Sciences, George Mason University, 4400 University Drive, MS 6B2, Fairfax, VA 22030, USA; acrooks2@gmu.edu

[2] Department of Geography and Geoinformation Science, George Mason University, 4400 University Drive, MS 6C3, Fairfax, VA 22030, USA; azufle@gmu.edu

* Correspondence: xyuan5@gmu.edu; Tel.: +1-703-993-9298

Received: 7 May 2020; Accepted: 6 June 2020; Published: 10 June 2020


**Abstract:** The research presented in this paper proposes a thematic network approach to explore rich relationships between places. We connect places in networks through their thematic similarities by applying topic modeling to the textual volunteered geographic information (VGI) pertaining to the places. The network approach enhances previous research involving place clustering using geo-textual information, which often simplifies relationships between places to be either in-cluster or out-of-cluster. To demonstrate our approach, we use as a case study in Manhattan (New York) that compares networks constructed from three different geo-textural data sources—TripAdvisor attraction reviews, TripAdvisor restaurant reviews, and Twitter data. The results showcase how the thematic similarity network approach enables us to conduct clustering analysis as well as node-to-node and node-to-cluster analysis, which is fruitful for understanding how places are connected through individuals' experiences. Furthermore, by enriching the networks with geodemographic information as node attributes, we discovered that some low-income communities in Manhattan have distinctive restaurant cultures. Even though geolocated tweets are not always related to place they are posted from, our case study demonstrates that topic modeling is an efficient method to filter out the place-irrelevant tweets and therefore refining how of places can be studied.

**Keywords:** geo-textual data; volunteered geographic information; crowdsourcing; similarity network analysis; topic modeling

## 1. Introduction

Place and space are among the most fundamental concepts of geography [1,2]. Space is often considered to be points of locations represented by coordinates. Place, on the other hand, is an "experience-based dynamic construct" [3]. Compared to space, the concept of place emphasizes on the meaning-making process that is complex, dynamic, and individualistic [4]. In this paper, we study how different places are semantically similar, based on textual topics that appear in Volunteered Geographic Information (VGI) in these places. Our goal is to create a thematic similarity network that connects places of similar topics regardless of their physical distance. By applying a network clustering algorithm, we find groups of semantically similar places and analyze their topics and spatial autocorrelation qualitatively and quantitatively.

Analyzing and theorizing about places from a variety of perspectives, has a long history in geographical analysis—from social area analysis [5–8] to more recent geodemographic analysis that derives collective behaviors and characteristics from demographic data of geographic regions [9,10]. In the past decade, studies on place have taken advantage of a new data source, that of VGI [11]. VGI comes in many different forms, from that of maps created by users to text from Wikipedia which

has a geographic component (e.g., place names) [12]. The research presented in this paper uses the textual form of VGI, specifically crowdsourced reviews from TripAdvisor and geolocated Twitter data (see Section 3). Such platforms provide large amounts of textual data with either explicit or implicit geographic information contributed by users [13,14]. Leveraging this unstructured geographical information found in such texts allows us to comprehend the complexity of places at scale [15–17].

Generally speaking, the most common method used by prior research to analyze geo-textual data is to structure the unstructured texts into themes (i.e., topics) through topic models (e.g., [18]). This is then often followed by applying clustering algorithms (e.g., k-means) to expose the underlying patterns of sentiments, experiences, or activities captured in the text (e.g., [12,19–24]). When places are clustered for further analysis, however, those in the same cluster are assumed to be carrying similar characteristics. Relationships between places are reduced to being either in-cluster or out-of-cluster. However, we would argue that the connectedness and relationships of places, in reality, are more complex. For instance, when connecting places in a network, places at the edge of their own clusters still have relatively weak out-of-cluster connections. The network approach presented in this paper, recognizes them as places with both in-cluster and out-of-cluster connections. Thus, this approach does not limit us to only perform network-level place clustering, but also to discover unique places based on their positions in the networks. To highlight this we use a case study to demonstrate the approach in the context of Manhattan, New York. In the remainder of the paper, we will first discuss related research pertaining to topic modeling and thematic similarity network analysis (Section 2). This is followed by introducing the data (Section 3) and the methodology (Section 4) that we applied to our case study. The results are then presented in Section 5 and finally, the implications and conclusions of our research are presented in Section 6.

## 2. Related Work

The approach proposed in this paper involves two major steps, topic modeling using geo-textual data and thematic similarity network analysis. In what follows we review related work with respect to these steps. For step one, topic modeling is a widely used language model for understanding large amounts of unstructured textual data. Previous research has adopted generic topic modeling algorithms (e.g., [18]) to ones that incorporate geographic information (e.g., [25,26]). Using these geographical topic models, studies have been able to derive the topics from travel blogs and Flickr tags to specific geographical units, such as states [25,26]. Other work has analyzed the relationships between topics and countries from online news articles and blogs [27], or generated activity patterns from check-in data [23]. In addition, topic modeling has been used to recommend travel destinations using travel blogs [28,29], create location related question-answering systems using Twitter and blogs [30], and predict the future distribution of topics [31]. Despite research on innovating geographic topic models, many researchers often chose to use generic topic models (such as Latent Dirichlet Allocation, LDA [18]) to analyze geo-textual data. In such instances, the geolocational information in the text does not contribute to the results of the topic models but is used only after applying the model. For example, Adams et al. [20] explored the temporal themes related to places using travel blogs and applied a similarity score between places based on the topics. Jenkins et al. [32] compared themes of geographic areas from Twitter and Wikipedia whereas Xu et al. [33] model the topics of restaurants of a city.

In addition, previous work has defined "place" at various levels of aggregation—countries, cities, neighborhoods, buildings, but such aggregations artificially split geographical areas. For example, at the neighborhood level, Cranshaw et al. [34] detected boundaries of neighborhoods using check-in data and Foursquare venue descriptions in order to show that crowdsourced and official neighborhood definitions differed. At a more aggregated level, Preoţiuc-Pietro [35] viewed cities as collections of Foursquare venues and clustered cities hierarchically using venue descriptions to show that similarities between cities can be captured through crowdsourced data. Since Foursquare venue data also provides venue categories, Noulas et al. [36] clustered both geographic areas (in terms of $625 \times 625$ square meters) and users based on their visit history in order to enhance recommendation systems for different users. In another work,

Crooks et al. [12] proposed a multi-level (individual building, streets, and neighborhoods level) approach for discovering social functions through mining place topics. Clustering at different aggregations also allows us to find places where people share similar experiences [20,37] along with places with similar functions [12]. When applying clustering, however, the relationship between places becomes binary, being similar or not similar, and thus the relationships between places in *different clusters* are often ignored.

Turing to work pertaining to thematic similarity network analysis (i.e., the second step of our approach), previous studies have analyzed place similarities but rarely used a network approach in the context of geo-textual data. For example, Janowicz et al. [38] used semantic similarity for developing geographic information retrieval applications. While Yan et al. [39] trained word embeddings for place types that was then used for exploring similarity and relatedness between point-of-interests types. In terms of using a similarity network-based approach, Quercini and Samet [40] created graph-based similarity measures to address spatial relatedness of a concept to a location using Wikipedia articles. In another work, Hu et al. [41] placed cities into networks based on their semantic relatedness (i.e., number of news articles which contain the co-occurrences of the two cities). Similarity networks, however, have seen much wider applications in domains outside of geography, ranging from analyzing protein sequences and structures [42], genome data [43] to that of hospital patients [44,45]. Methodologically, such studies have demonstrated that one of the most important analysis for similarity networks is clustering (i.e., community detection), which captures groups of nodes that are most similar to each other. Although place clustering does not require connecting places in networks, one of the advantages of conducting network-based clustering is that it enables for downstream node level analysis in relation to clusters. For example, Valavanis et al. [42] discovered structural similarities of protein folds and classes in the downstream analysis after carrying out network clustering. Similarly, in the case study presented throughout the rest of this paper, we will apply clustering to the similarity network as well as identifying special nodes (i.e., places) based on their positions in the network.

## 3. Data

### *3.1. Data Collection*

To apply our methodology (see Section 4) to showcase how a network approach can be used to study place, data was needed to be collected. In this study we used two geo-textual data sources: TripAdvisor and Twitter. The rationale for choosing these data sources are two-fold. First, they are open source that have been widely used by previous research (as discussed in Section 2). Therefore, future studies could use these data sources to extend the research presented in this paper. Secondly, most prior research using geo-textual data often choose only one data source. In the research presented in this paper, we aim to provide a thematic similarity network approach which can compare multiple geo-textual data sources. The TripAdvisor data was collected in September 2019 which included reviews for attractions and restaurants in New York City. For each attraction and restaurant, the addresses, neighborhood, and reviews were retrieved. An example of this is shown in Figure 1, in which we highlight content that was used in our analysis (i.e., locational information and reviews). With respect to Twitter, we were only interested in tweets that had a precise geographical coordinate. The Twitter data that was collected from 1 January 2015 to 31 December 2015 with a bounding box of latitude ranging from 40.481867 to 40.9325 and longitude between −74.2721 and −73.626201, which includes the New York City.

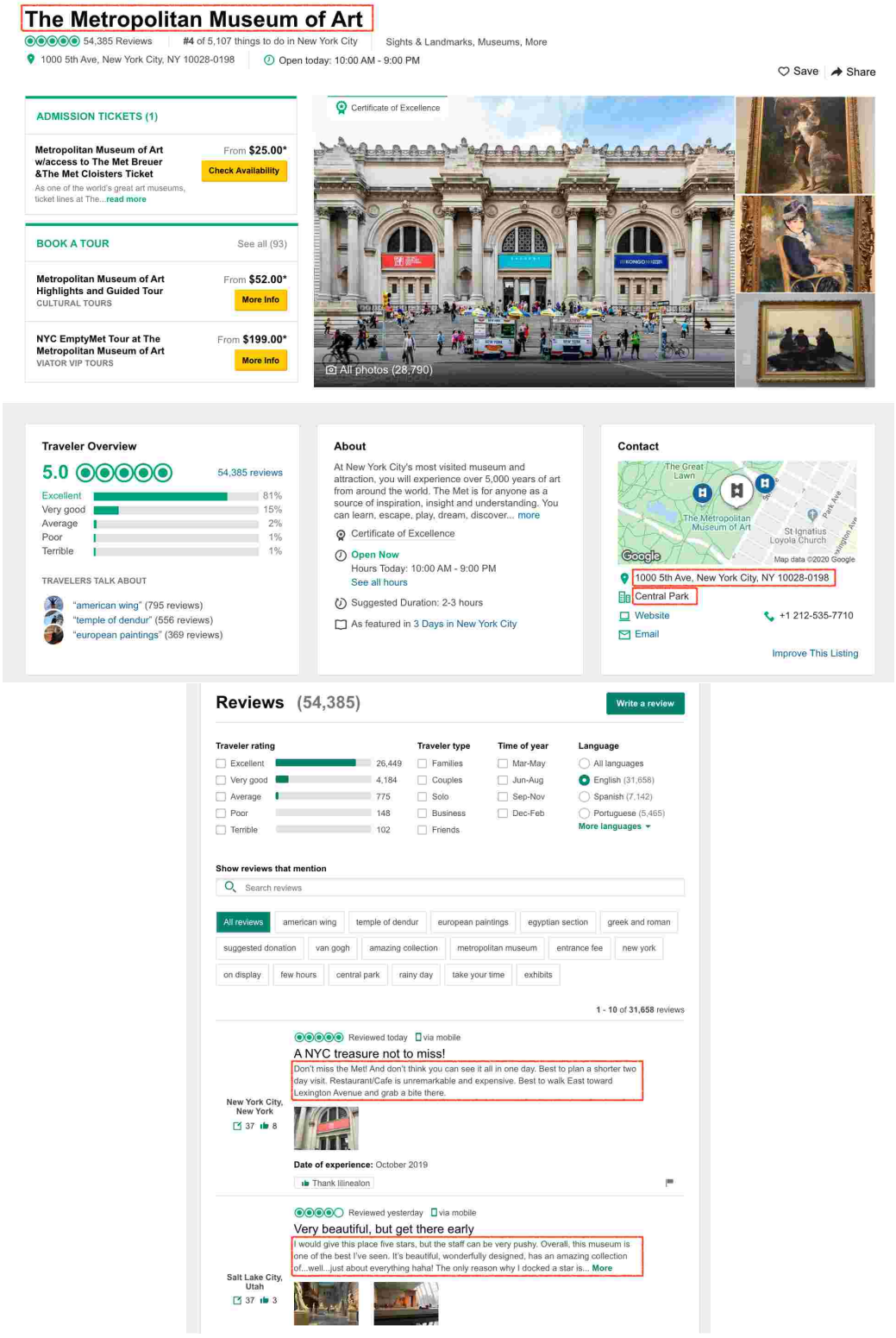



**Figure 1.** An example of a TripAdvisor page and the highlights are the information scraped from the page.

### *3.2. Data Pre-Processing and Aggregation*

As the locational information from TripAdvisor attractions and TripAdvisor restaurants was in the format of addresses, the first step of data pre-processing was to geocode TripAdvisor data addresses using Google Maps Geocoding application programming interface (API) [46]. After all the data was

geocoded, the second step was to define what a place is. As was discussed in Section 2, previous studies have treated places at various levels of aggregation. For this case study, a place is a census tract defined by the United States Census [47]. Although the aggregation may result in the modifiable areal unit problem that statistical summaries of the aggregated area are influenced by the shape and size of the area [48], the reason of using census tract in this research was to incorporate the census demographic data into the analysis. The rationale for this was to be able to explore the connection between the patterns found in crowdsourced reviews (or tweets) and the underlying geodemographics of an area. Furthermore, by using census data, while not only demonstrating how our case study allows for a novel approach to studying places through thematic similarity networks, but it also allows for others to use it in different areas within the United States or in other countries where census data is available (e.g., as in the United Kingdom). Int should be noted however, if readers are not interested in comparing the geo-textual data to census data, our approach could be applied to other levels of aggregation such as grids, road segments etc. (see [12]).

After aggregating data to the census tract level, the final step was to select only tracts that appeared in all three datasets to make them comparable. It was found that most of the attractions within New York City from TripAdvisor were located in Manhattan, and thus for the analysis in this paper, the tracts only in Manhattan were analyzed. Table 1 shows the number of restaurants/attractions, reviews/tweets, and census tracts after restricting the study to Manhattan. Furthermore, the texts (i.e., reviews and tweets) were filtered to only be those that were in English. Although special characters such as "@", emojis, and stop words may contribute to the meaning of the text [49], we do not consider them in this work as is common in text pre-processing (e.g., [50]). Next all words were converted into lower case to treat all words with the same text the same. Finally, in order to reduce the number of vocabularies (e.g., words with the same meaning such as walking, walk, walked), a stemmer (i.e., Porter Stemmer) was applied and only the stems of the words were retained [51].

**Table 1.** Statistical summaries of data sizes.

| Dataset | Count | Number of Reviews | Number of Tracts |
|---|---|---|---|
| TripAdvisor attractions | 956 (attractions) | 446,747 | 210 |
| TripAdvisor restaurants | 7946 (restaurants) | 865,055 | 210 |
| Twitter | 268,224 (users) | 2,009,498 | 210 |

## 4. Methodology

In this section, we will first introduce how the topic model was trained on each dataset and how the thematic similarity network is constructed based on the similarities between the derived topics (Section 4.1). Figure 2 illustrates the workflow from data input to thematic similarity network output including data collection and prepossessing which was described in Section 3. After the thematic similarity networks are constructed, we carried out the network community detection on these networks (Section 4.2) and node level network analysis (Section 4.3). Finally, the algorithm and implementation is presented in Section 4.4.

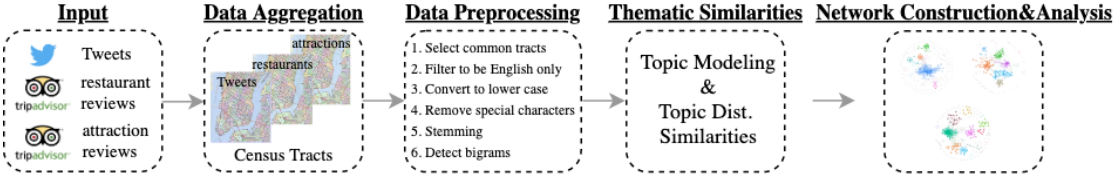



**Figure 2.** Workflow from data input to the construction of the thematic similarity network and analysis (i.e., community detection and unique nodes discovery).

### 4.1. Topic Modeling and Thematic Similarity Networks

As noted in Section 2, The first major step towards gaining a meaning from large collections of text is topic modeling. Topic modeling is a type of statistical model that discovers latent topics in documents. For example, when writing a review, the reviewer might not be thinking of specific topics, but topic modeling assumes that there are underlying topics, which are known as latent topics. One of the most widely used topic models is Latent Dirichlet Allocation (LDA) [18], which is a generative probabilistic model that treats each document as a distribution of latent topics and each topic as a distribution of words. A document can be a news article, a review or a social media post. In this research, each document is comprised of all the texts for each census tract from one dataset. For example, for the TripAdvisor restaurant dataset, tract "36061000700" had 57 restaurants, which had total of 1951 reviews. The 1951 reviews were treated as one document in the LDA model.

One LDA model is trained for each of the three datasets. To train LDA models, we need to tune the hyper-parameters using the datasets to obtain models that have the best performance. A LDA model can predict the words in each topic and the topics are in each document. Therefore the hyper-parameters that need to be tuned include the a-priori probability vector $\alpha$ that maps each topic to a probability, and the a-priori probability vector $\beta$ that maps each word to a probability. Moreover, the total number of topics ($K$) need to be learned from the data as well. The model was implemented using gensim LDA library, in which alpha and beta was set as "auto" so that both hyper-parameters can be learned from the data [52]. To ensure a better model interpretability, we detected the bi-grams in the texts, after which the corpus became a mixture of uni-grams and bi-grams. In addition, the vocabularies in the corpus was truncated since otherwise many of the most frequent words that bears less concrete meanings in the context of our tasks, such as "I" and "is", would become the top words of the topics. However, the threshold of the word frequency ($top\_n$) to be truncated is a parameter of the model that needs to be tuned during training as well. To tune $K$ and $top\_n$, experiments on each dataset were carried out. Since there is no ground truth for topic models, the common model evaluation metrics are perplexity and coherence [53]. However, the experiments showed that optimizing coherence or perplexity scores in all three datasets did not generate models with better topic interpretability (code https://bitbucket.org/xiaoyiyuan/network_vgi/src/master/script/topic_model_results.ipynb?viewer=nbviewer). As a result, we adopted interpretability and manual observations as the metrics for evaluating the topic model quality that can be found. For instance, when $k$ is too high, the model produces topics that have many common words, meaning that new topics are not contributing to generating new knowledge about the data. When $top\_n$ is too high, more documents have only one or two topics, which makes it hard to comprehend topic meanings. Table 2 shows the parameters that produces the most interpretable model for each dataset. Each of the experiments and the results with various values for the hyper-parameters can be found in the shared source code.

**Table 2.** Parameters of the trained LDA models on the datasets.

| Dataset | $K$ | $top\_n$ |
|---|---|---|
| TripAdvisor attractions | 30 | 100 |
| TripAdvisor restaurants | 40 | 500 |
| Twitter | 70 | 700 |

When the LDA model is trained, each document is represented by a distribution of topics. The square root of Jensen–Shannon divergence is a commonly used metric of measuring distance between discrete distributions. The Jensen–Shannon distance between two (topic) probability distributions $P$ and $Q$ is defined formally as:

$$JSD(P||Q) = \sqrt{\frac{D_{KL}(P||M) + D_{KL}(Q||M)}{2}}, \text{ where } M = \frac{1}{2}(P + Q).$$

The Jensen–Shannon divergence is symmetrical (i.e., $JSD(P||Q) = JSD(Q||P)$). As a result, the edges of the similarity networks are not directed but weighted and the weights are the similarity scores. Using the same method, three similarity networks were constructed from the three datasets. Since there is always a similarity score between each pair of tracts, there is always an edge between them in the networks as well, making the networks fully connected.

### 4.2. Community Detection

Discovering communities of a fully connected network requires network sparsification [54]. Network sparsification reduces number of edges while preserving structural and statistical properties of interest. Thus, the principle is to reduce the network size by retaining only the important edges and in a similarity networks, the edge weight (i.e., the similarity score) is an indicator of edge importance [55]. The cut-off value for edge weights to sparsify the networks depends on the data and the clustering algorithm. In this work, we used the Girvan–Newman algorithm [56] to conduct network community detection (i.e., clustering) on the sparsified networks. The Girvan–Newman algorithm is a hierarchical method of detecting communities in complex networks, which can also be applied to weighted networks [56]. Within each step of the Girvan–Newman community detection algorithm, it progressively removes edges with highest edge betweenness centrality (i.e., edges with highest number of shortest path passing through them) and recalculates edge betweenness after each iteration of removal. By removing these high betweenness edges, the communities are separated from each other and consequently, the underlying community structure of the network is revealed. In a weighted network, the Girvan–Newman algorithm calculates the edge betweenness as described above, ignoring the edge weights. Then it divides the edge betweenness by the weight of the corresponding edge. As with unweighted networks, the algorithm then removes edges of highest betweenness. The result of the algorithm is a dendrogram which repeats the steps until no edges can be removed or the most ideal communities have achieved (i.e., the highest modularity of clusters). Therefore, we need to cross validate two parameters, the edge weight cut-off threshold for sparsification and the number of iterations for the Girvan–Newman algorithm. The process of sparsifying fully connected network to community detection is shown in a stylized network in Figure 3.

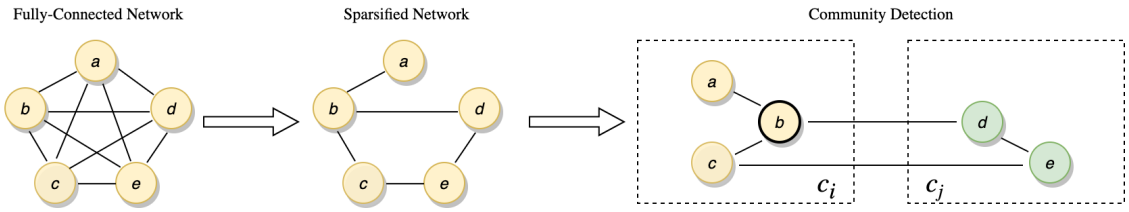
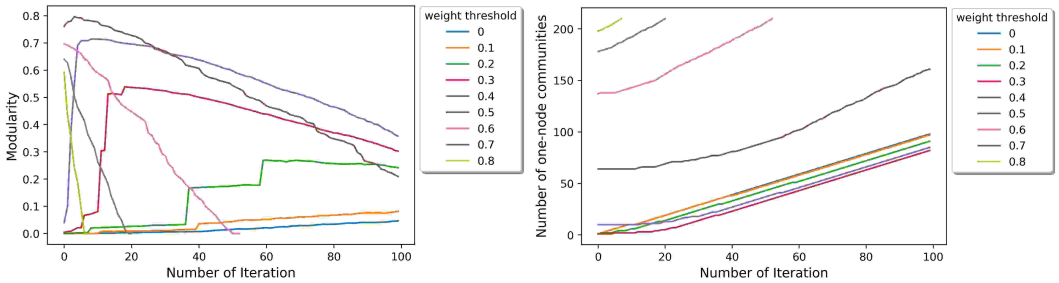


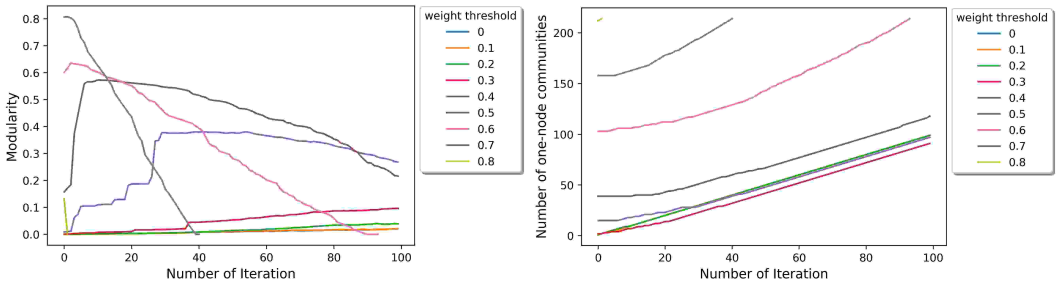

**Figure 3.** A stylized network demonstrating the process of community detection from a fully connected similarity network.

A common evaluation metric for network community detection quality is modularity, which is a measure of the strength of division of a network into communities [57]. High modularity means that within each community detected, there are dense connections within the community and sparse connections between nodes in different communities. However, using modularity as the sole metric for the Girvan–Newman algorithm is not sufficient for our task—Figure 4 shows that when modularity is at its highest value, the network has become too scattered with a large number of one-node communities, which hinders the downstream analysis of interactions between communities and relationships between nodes and communities. To mitigate this problem, we selected the set of parameters that has the highest modularity without generating many one-node communities. For instance, the highest modularity for TripAdvisor attraction network (Figure 4a, left, brown) is 0.8 at iteration 8 but it produced more than 50 one-node communities (Figure 4a, right, brown). However, choosing a model with a slightly lower modularity value, e.g., 0.7 (Figure 4a, left, purple) significantly reduced the

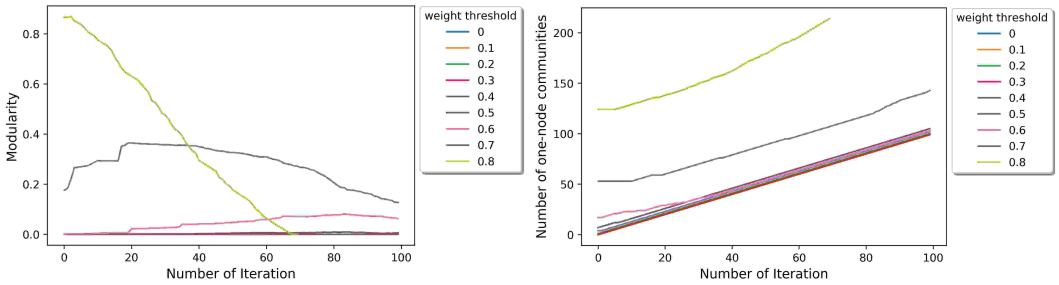
number of one-node communities (Figure 4a, right, purple). Using the same heuristics, the best parameters for each network are presented in Table 3.




(**a**) TripAdvisor attractions thematic similarity network.




(**b**) TripAdvisor restaurants thematic similarity network.




(**c**) Twitter thematic similarity network.

**Figure 4.** Cross validation results for community detection in three networks, modularity (**Left**) and number of one-node communities (**Right**).

**Table 3.** Parameters used for community detection in the three networks.

| Network | Iteration | Weight Threshold | Modularity |
|---|---|---|---|
| TripAdvisor attractions | 8 | 0.4 | 0.714 |
| TripAdvisor restaurants | 10 | 0.5 | 0.573 |
| Twitter | 19 | 0.7 | 0.365 |

### *4.3. Discovering Unique Nodes*

Since the edge weights represent similarity, a node (i.e., a place) with a low-degree centrality value means that it bears low similarity to other nodes in the network. It is straightforward, therefore, to discover the ends of the uniqueness spectrum—on the one end, the highly unique places are nodes with a low-degree centrality while at the other end high degree centrality nodes are the least unique places. Other than these two extremes, there are nodes that act as bridges between communities that carry their unique characteristics, i.e., the community boundary nodes. The concept of community boundaries is rarely applied in similarity network analysis but is often used in social

network analysis. In social networks, the community boundaries are the people who convey outside information to those in the community with no out-of-community connections [58,59]. We adopted and modified the definition of community boundaries from the social network analysis by Guerra et al. [58]. This modified definition of a community boundary is that of a node $v$ that is a boundary node of community $C_i$ for community $C_j$ when:

1. node $v \in C_i$ has at least one edge connecting to community $C_j$ and
2. all the neighborhoods of $v$ have no edge connecting to community $C_j$.

The community boundaries are identified in the three sparsified networks instead of the original full-connected ones because identifying boundary nodes relies on the community structures detected in the sparsified network. Finally, the last step in Figure 3 illustrates a stylized network with community boundaries. For community $C_i$, node $b$ and node $c$ both have edges connecting to the outside community $C_j$. Node $b$ qualifies as a boundary node for community $C_i$ because it has a neighbor (node $a$) with *no* edges connecting to community $C_j$. Since node $c$ does not have a neighbor meeting this requirement, node $c$ does not count as a community $C_i$'s boundary node to $C_j$. In social networks, the definition of boundary nodes guarantees that node $b$ is the only node that brings outside information to node $a$. In the context of place similarity networks, a node with no connection with the outside community (i.e., node $a$) indicates it has characteristics that are unique to its own community and the boundary nodes (e.g., node $b$) are the ones that connect the uniqueness of the communities.

### *4.4. Algorithm and Implementation*

To summarize what has been discussed above with respect to our methodology, the pseudo-code for it is described in Algorithm 1. The algorithm takes one data source (e.g., Twitter corpus or TripAdvisor) as an input. The loop in Lines 1–6 constructs topic models from the input texts and Lines 7–11 calculate similarities between each document of the input. Line 12 constructs a thematic similarity network (which was explained in Section 4.1). Finally, Lines 12–14 detect communities in the network (Section 4.2) and the loop from Lines 15–22 discover boundary nodes (Section 4.3). The complete Python code and information pertaining to the software versions is available at https://bitbucket.org/xiaoyiyuan/network_vgi.

**Algorithm 1:** Network Construction and Community and Boundary Node Detection

**Input:** Corpus split by their geolocated census tracts $D = d_1, d_2, ..., d_n$

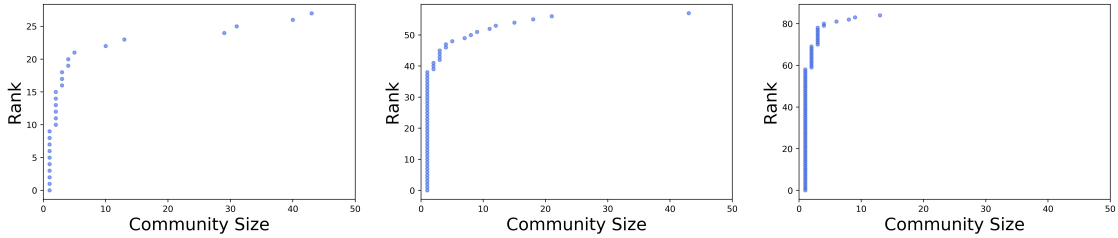
**foreach** *d in D* **do**
  $pairs \leftarrow [\,]$
  $pair\_similarities \leftarrow [\,]$
  `/* TM maps topic ID t and words w (from document D) to a probability */`
  $TM(t, w) \leftarrow topic\_model(D)$
  $d.topics = topic\_model(d)$
**end**
**foreach** $d \neq d'$ *where* $1 \leq d, d' \leq n$ **do**
  $pairs$.insert($[d, d']$)
  `/* Jensen-Shannon Distance */`
  $distance \leftarrow$ JSDistance($d.topics, d'.topics$)
  $pair\_similarities$.insert($distance$)
**end**
$G = (D, D \times D, pair\_similarities)$
`/* Sparsify the graph by pruning edges with low similarity */`
$G \leftarrow$ sparsify($G$)
`/* Girvan-Newman Community Detection */`
$C \leftarrow$ girvan_newman($G$) `/*` $C = c_1, ..., c_{|C|}$ `*/`
$boundary\_nodes \leftarrow [\,]$
**foreach** $d_i, d_j$ *in* $D$ **do**
  `/*` $d_i$ `and` $d_j$ `are from different communities and are connected */`
  $condition1 = d_i \in c_i \wedge d_j \in c_j \wedge d_i \neq d_j \wedge d_i.has\_edge(d_j)$
  $condition2 = d_i neighbors.has\_no\_edge(c_j)$
  **if** $condition1 \wedge condition2$ **then**
    $boundary\_nodes$.insert($d_i$)
  **end**
**end**
**return** $G, C, boundary\_nodes$

## 5. Results

Building upon our methodology, in this section, we will present the results for the thematic similarity network analysis of places in Manhattan, New York. Specifically, Section 5.1.1 maps and visualizes the major network communities and their topics and presents results from the spatial autocorrelation of these communities using Moran's I measure of spatial autocorrelation(Section 5.1.2). In Section 5.2, we enrich the network nodes with geodemographic data and finally in Section 5.3 we identify and analyze nodes by their degrees of uniqueness.

### *5.1. Major Network Communities and Their Topics*

In this section, we evaluate the clusters found using our proposed community detection approach described in Section 4.2. For this purpose, we first visualize and qualitatively analyze the community clusters. Then to ensure that the communities that we found are clustered significantly, we test each community for spatial autocorrelation using Moran's I. The sizes of the communities are show in Figure 5. Even though we lowered the number of one-node communities in the community detection (as discussed in Section 4.2), the distributions of the community sizes still appear to be long-tailed. For the sake of clear visualization, only the major communities (i.e., communities with a size equal or larger than 5 nodes) from the community detection are presented for each network. The topic modeling results for all communities are https://bitbucket.org/xiaoyiyuan/network_vgi/src/master/script/topic_model_results.ipynb?viewer=nbviewer.



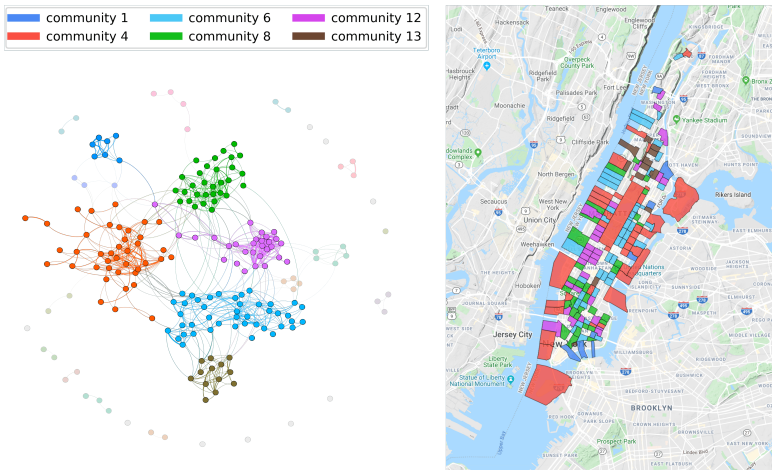

(**a**) TripAdvisor attractions (**b**) TripAdvisor restaurants (**c**) Twitter

**Figure 5.** The sizes of communities from the community detection results of the three networks.

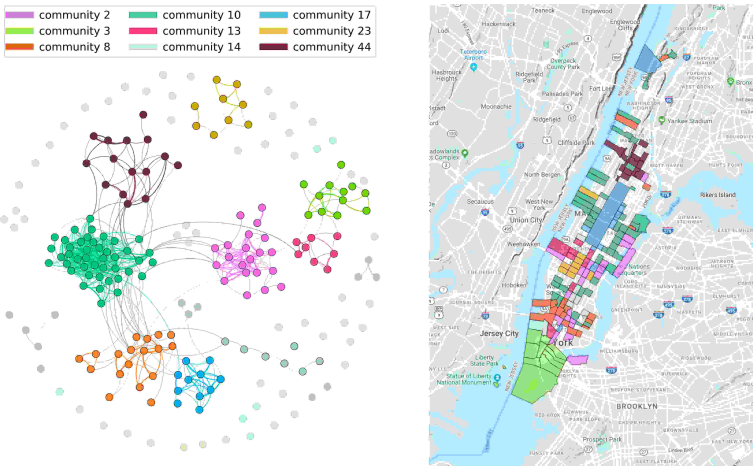
#### 5.1.1. Network Visualization and Mapping

Figures 6a and 7a shows the visualizations of networks, maps, and topics from the community detection results of the TripAdvisor attractions thematic similarity network. Major communities are highlighted, and tracts of the major communities are mapped in the same colors. In Figure 7a, dominant topics (i.e., topics with coefficients equal or higher than 0.1) of the major communities are shown. Based on the words in the topics, communities can be characterized into categories such as church tracts in Harlem (Community 13), restaurant tracts that includes two famous restaurant areas in Chelsea and Chinatown (Community 8), bridge tracts (Community 1), and theater tracts (Community 12) while other communities have more hybrid characteristics (i.e., topics). Observing the combination of the topics and their locations on the map (Figure 6a), some communities have tracts which are visually close to each other and their topics reflect the main characteristics of the attractions in these geographic regions. For instance, topics of Community 12 are about "broadway", "theater", "concert", and "venu" (venue) and most of these tracts are clustered around the Broadway theater district. Furthermore, as shown on the map (Figure 6a), not all communities are not clustered perfectly in a geographic region and some of the tracts of a community are in the same region. For example, even though most of the tracts of Community 12 are located Midtown, the rest of the tracts are scattered around the Downtown area. The reason is that the topics of Community 12 include not only Broadway but also more broadly "concert", "game", and "venu" (venue) (Figure 7a). A similar example is that of Community 13 that has a dominant topic with keywords "harlem", "church", and "theater" and most of the tracts of Community 13 are located in Harlem and tracts that are not in Harlem have church related attractions such as The New York Mosque in Midtown Manhattan and Mariners' Temple Baptist Church in Downtown Manhattan. Such findings indicate that the network communities are reflections of people's similar experiences of various attractions as they are mined from a large amount of crowdsourced reviews from individuals.

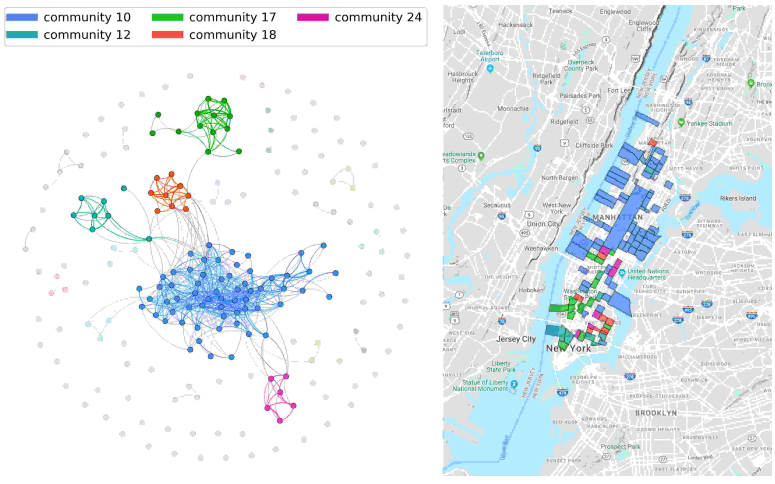
For the restaurant thematic similarity network, communities show higher level of spatial proximity (Figures 6b and 7b). One of the most prominent of such is that of Community 3, which is shown in Figure 6b clustered in Downtown Manhattan. Primary topics of community 3 (Figure 7b) are "pub" and "eatali" (Eataly food market), and "financi district" (financial district). In addition, tracts of Community 8 have close geographic proximity as well. This is evident from the map of Figure 7b, where most of the tracts in Community 8 are located between Downtown and Midtown Manhattan. Community 8 has Topics 14 and 32 featuring word stems such as "greenwich_villag" (Greenwich Village), "west_villag" (West Village), "japanes" (Japanese), and "bagel". Similarly, Community 17 has Topic 36 that can be interpreted as Central Park related even though it has the common Topic 32 that shows up across many other communities (e.g., Community 2, 8, 10, 14, and 17). Interestingly, communities from TripAdvisor attractions network have counterparts from the restaurants network communities. For example, Community 13 from attraction network and Community 44 from restaurant network are about Harlem, which can be seen from the geographic clusters on the map and their dominant topics. A similar finding is for the theater district, which appears in both Community 12 of attraction network and Community 23 of the restaurant network. This suggests that people's dining experiences can be intertwined with the characteristics of the surrounding attractions or vice versa.




(**a**) TripAdvisor attractions

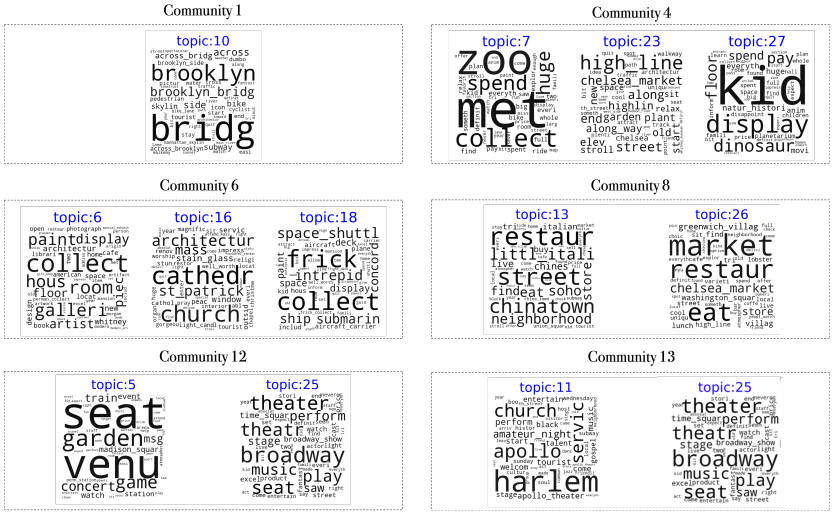



(**b**) TripAdvisor restaurants

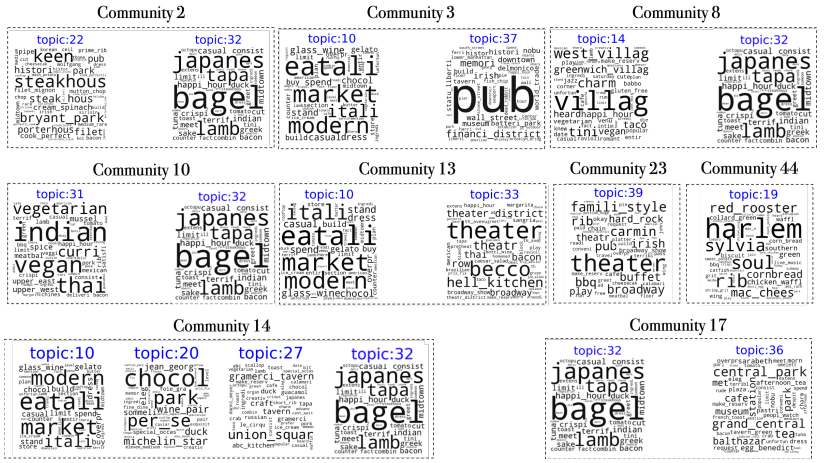



(**c**) Twitter

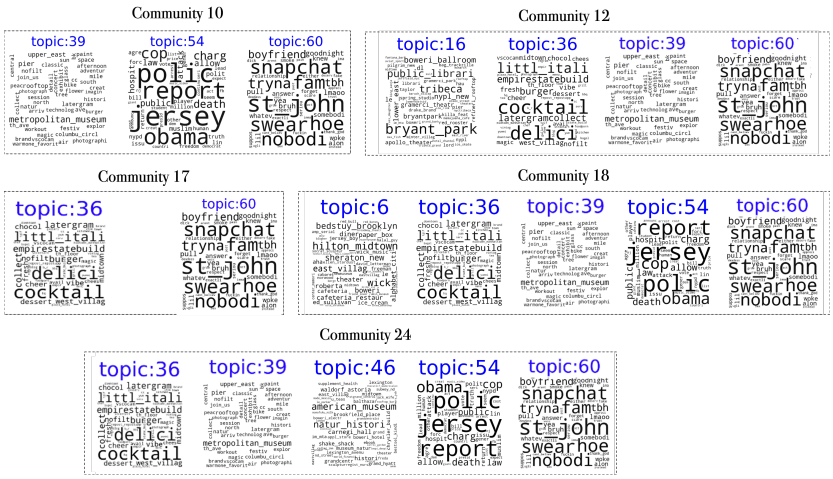
**Figure 6.** Network visualization of all communities from the thematic similarity networks using Gephi [60] Fruchterman–Reingold layout with major communities highlighted. Only the major communities are shown on the map for the sake of clarity. Major communities in Network visualization and mapping for each network are colored the same and thus the legend applies for both.




(**a**) TripAdvisor attractions




(**b**) TripAdvisor restaurants




(**c**) Twitter

**Figure 7.** Dominant topics of all major communities in each thematic similarity network. Dominant topics are topics with coefficients equal or higher than 0.1.

Turning to the results of the Twitter thematic similarity network, one of the most noticeable pattern is that of Community 10 (i.e., the blue community in Figure 6c) which dominates this network. Unlike the communities in Trip Advisor attractions and restaurants where there is a more even distribution of community sizes. Furthermore, communities from the Twitter dataset have more diverse topics than that of restaurant and attraction networks from TripAdvisor (Figure 7c). This could partly due to the distinction between Twitter and TripAdvisor as data sources for studying places. In that TripAdvisor reviews are directly about places but this is not necessarily the case for Twitter, which is a more generic social media platform where users can contribute a whole variety of topics [13]. Therefore, some topics (e.g., Topic 54 and Topic 60 in Figure 7c) from Twitter are not about places but relate to news or social and political discussions. This indicates that although geolocated tweets can be used to study people's perceptions and experiences about places, it needs to be used with awareness that the texts may need to be filtered. The results in Figure 7c shows that it is viable to use topic modeling to filter out the non-related topics (e.g., Topic 54 which relates to police reporting and New Jersey). The reason could be that tweets pertaining to social discussions often use different vocabularies than texts directly about places. Since topic modeling is a bag-of-words approach, the model is sensitive to vocabularies and thus can "tell them apart" as separate topics.

#### 5.1.2. Quantitative Test for Spatial Autocorrelation of Communities

In the previous section (Section 5.1.1), we discussed network communities and whether the communities have geographically proximate tracts. In this section, we will present the results of Moran's I measure of spatial autocorrelation to quantify geographical proximity of the major communities in Table 4 and the Moran's I results for all communities are found in Table A1 in the appendix. Moran's I is a measure for spatial autocorrelation that is often applied on continuous data. To measure each community's autocorrelation level, we therefore encoded tracts of a specific community as 1 and all the other tracts as 0. We defined neighborhood using Queen's contiguity, i.e., any polygons (i.e., tracts) that shares a point-length border are neighbors.

Among the three networks, all the major communities from the TripAdvisor restaurant network have statistical significance in their spatial autocorrelation results (Table 4). Communities 3, 8, 10, 23, and 44 of the restaurant network have spatial autocorrelation at the 99% confidence interval, which are generally about pubs, west village, vegetarian Indian, theater, and Harlem respectively. These results inform us that these geographic clusters in Manhattan have their own restaurant culture, manifested by the topical summaries from TripAdvisor reviews. Other major communities from the restaurant network (i.e., Communities 2, 13, 14, and 17) also show statistical significance (at a confidence interval of over 95%) in their spatial autocorrelation results. These communities have relatively lower scores from Moran's I test, which can be observed on the map as they are more spread out over Manhattan. For the attraction network, Communities 4, 6, 12, and 13 have spatial autocorrelation and have topics pertaining to zoo/High Line/kid, cathedral/gallery, Broadway/seat/venue, and Harlem/Broadway, which can summarized from word clouds in Figure 7a. Therefore, attractions from these communities are more of a mixture of many different topics, which also explained the reason of the Moran's I for the attraction network being relatively lower than that from the restaurant network. Similarly, for the Twitter thematic network, besides the biggest community (i.e., Community 10), the others have low Moran's I values which suggests that the topics discussed on Twitter are less correlated with their geographic locations.

**Table 4.** Moran's I spatial autocorrelation of major communities in each network.

| Network | Community ID | Moran's I | *p* Value |
|---|---|---|---|
| TripAdvisor attraction | 1 | −0.022740 | 0.413 |
| | 4 | 0.120743 | 0.014 * |
| | 6 | 0.142676 | 0.006 ** |
| | 8 | 0.002285 | 0.410 |
| | 12 | 0.107151 | 0.014 * |
| | 13 | 0.106869 | 0.034 * |
| TripAdvisor restaurants | 2 | 0.120561 | 0.017 * |
| | 3 | 0.682913 | 0.001 *** |
| | 8 | 0.223211 | 0.001 *** |
| | 10 | 0.174987 | 0.001 *** |
| | 13 | 0.112362 | 0.026 * |
| | 14 | 0.167896 | 0.010 ** |
| | 17 | 0.255313 | 0.002 ** |
| | 23 | 0.345028 | 0.001 *** |
| | 44 | 0.572042 | 0.001 *** |
| Twitter | 10 | 0.328338 | 0.001 *** |
| | 12 | 0.056188 | 0.085 |
| | 17 | 0.118175 | 0.020 * |
| | 18 | 0.194717 | 0.007 ** |
| | 23 | 0.019150 | 0.258 |

* Significant at $p \leq 0.05$; ** Significant at $p \leq 0.01$; *** Significant at $p \leq 0.001$.

### *5.2. Enriching Network Communities with Geodemographic Attributes*

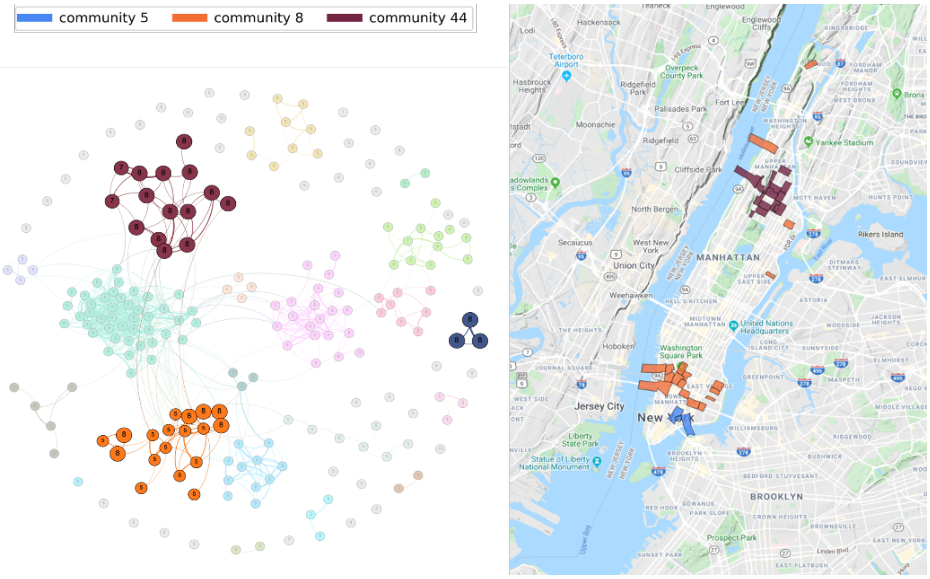
One advantage of examining places as the Census tracts is to combine Census demographic data with the results from the derived networks. If we were to use the demographic data from the US Census such as the American Community Survey (ACS), a tract can be described by multiple variables (e.g., total population, mean household income, education attainment, and marital status). An alternative is that proposed by Spielman and Singleton [61] who took the ACS data and clustered it to generate a single variable description known as a geodemographic classification (e.g., "Hispanic and Kids" and "Wealthy Nuclear Families"). We enriched our node attributes with this geodemographic classification at the tract level to explore the relationship between the network communities and their demographics.

Based on the results from Spielman and Singleton [61] shown in Table 5, most of the tracts that we study are classified as wealthy and the column "Percentage" shows percentages of tracts in that demographic classification. We use it as baseline to compare the percentage of each demographic classification for network communities. For instance, if a community has more than 22.90% of Demographic Type 8, based on Table 5, we define that community to have a high proportion of low-income residents. Using this baseline, we discovered that even though Manhattan tracts are mostly rich, and the majority of low-income tracts reside in a few network communities, Communities 5, 8, and 44. From the topics of these communities, two of them are in Chinatown and Harlem, presented in Figure 8). This suggests that these low-income areas have a distinctive restaurant culture. When applying the same method to the communities from TripAdvisor attractions and Twitter thematic networks, we do not find communities with high percentages of demographic types. This implies that discussions on Twitter and TripAdvisor attractions in Manhattan do not have patterns that correspond to the characteristics of its residents.

**Table 5.** Geodemographic distributions of tracts (i.e., network nodes).

| | Type | Type Description | Percentage |
|---|---|---|---|
| High Income | 2 | "Wealthy Nuclear Families" | 1.87% |
| | 5 | "Wealthy, urban without Kids" | 68.22% |
| | 7 | "Wealthy Old Caucasian" | 2.80% |
| Low Income | 8 | "Low income, mix of minorities" | 22.90% |
| Others | 10 | "Residential Institutions" | 1.40% |
| | 3 | "Middle Income, Single Family Home" | 0.47% |

Note: the geodemographic type and type descriptions are from research by Spielman and Singleton [61]. The descriptions are abbreviated to give the reader a sense of the classification schema and only the types found within our study area are shown. The percentages are based on all tracts across all network communities in our study area and thus are used as baseline for defining if network community is predominantly high income or low income.




(**a**)


(**b**)

**Figure 8.** Low-income communities highlighted and the label nodes represent the geodemographic type. (**a**) Network visualization of all communities and mapping of major communities (colored the same as Figure 6b). The node label represents their demographic classification. (**b**) Word cloud of topics in major communities. Topics of low-income communities are in visualized (**b**).

### *5.3. Identifying Nodes with Degrees of Uniqueness*

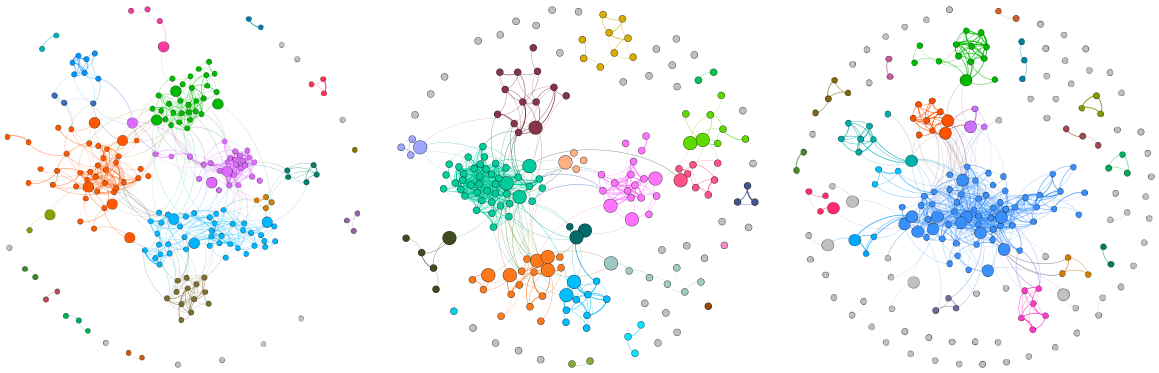
Besides network-level analysis, node level analysis allows us to identify important or interesting places. As discussed in Section 4.3, nodes with the lowest weighted centrality are the most unique ones and vice versa. In this Section, we will first examine the central nodes (top 5 highest weighted centrality nodes) and the outliers (top 5 lowest weighted centrality nodes), followed by exploring the community boundary nodes in the networks (i.e., nodes that act as bridges between communities that carry their unique characteristics).

Table 6 shows the topics for the central and outlier nodes in the network of TripAdvisor restaurants. Observing the number of topics for the two kinds of nodes, central nodes tend to have more diverse topics than the outliers. The topics of the outlier nodes show that these are the tracts with attractions that are unique to Manhattan, including "Skylin" (Skyline), "rockefel_center" (Rockefeller Center), "time_squar" (Time Square), "grand_central" (Grand Central Station), "Statu" (Statue of Liberty), and "elli" (Ellis Island). Since they are unique and distinctive, the outlier nodes have very low weighted degree centralities. This pattern of low-degree centrality nodes with distinctive topics also applies to the thematic similarity network from Twitter and TripAdvisor restaurants. On the contrary, the central nodes have a combination of common topics that enable them to have connections with many other nodes.

**Table 6.** Topics of central nodes and outlier nodes in thematic similarity network of **TripAdvisor attractions**.

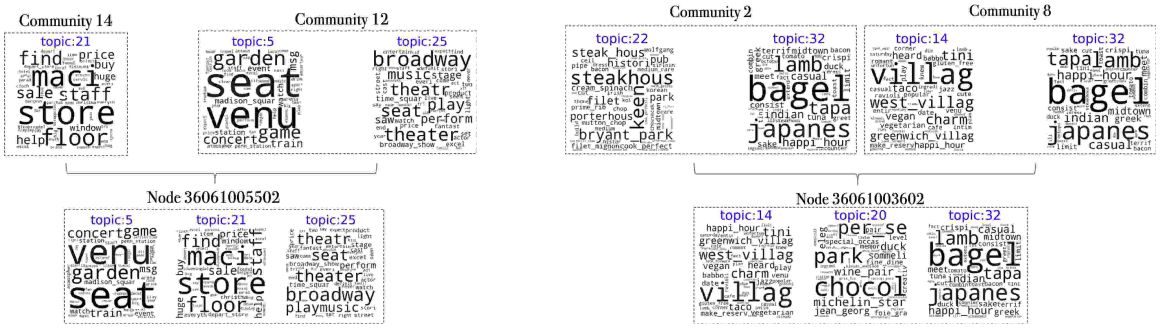
| Central Nodes | | | | | Outlier Nodes | |
|---|---|---|---|---|---|---|
| | topic:6 | topic:16 | topic:25 | topic:29 | | topic:2 |
| 36061012000 | display roomhous collect floor artist galleri paint | architectur st_patrick stain_glass cathedr church mass | seat theatr play theater perform broadway stage music | fountain play dog jewish side locat | 36061013700 | staten_island skylin island brooklyn_bridg cruis seen boat brooklyn top_rock manhattan_skylin |
| | topic:6 | topic:13 | topic:25 | topic:29 | | topic:24 |
| 36061005000 | floor room galleri paint display collect piec hous architectur | chinatown neighborhood little_itali restaur eat street store find soho | theater play broadway seat theatr music | jewish play dog side fountain locat | 36061010400 | rockefel_center spectacular year christma sunset queue book rockett includ top rock radio_citi light staff moma fantas floor |
| | topic:6 | | topic:29 | | | topic:12 |
| 36061005400 | galleri new floor room paint display collect architectur hous piec artist | | locat dog play side fountain batteri_park | | 36061007600 | queue attract definit pass lift stay floor staff fantast everyth th_floor say tourist tri expect light icon time squar |
| | topic:25 | topic:27 | topic:29 | | | topic:14 |
| 36061016700 | theater broadway theatr perform play music seat time_squar | pay spend kid display floor dinosaur | jewish side locat play dog fountain | | 36061009200 | station termin restaur grand_central architectur train train_station |
| | topic:5 | topic:21 | topic:25 | | | topic:22 |
| 36061005502 | madison_squar venu game concert train seat event garden msg station | store find staff sale price buy maci floor help huge window | music broadway perform theatr seat play stage theater | | 36061000100 | pedest crown immigr island elli island queue secur_book inclus inform advanc statu boat liberti |

Figure 9 shows the positions of community boundary nodes in the three networks, which to be expected are often at the edges of the communities. Identifying nodes with these special positions facilitates us to identify places with hybrid characteristics from both communities. Community boundary nodes connect the uniqueness between communities. To demonstrate the ways that the topics from community boundary nodes include topics from the communities, here we will show two examples from the TripAdvisor networks. First, Figure 10a shows an example of the topics and characteristics of a community boundary node and how the boundary node has the topics from two communities (i.e., Communities 14 and 12). However, when two communities have overlapping topics, the topics of the community boundary node are not always the perfect combination of topics from the two communities as shown in Figure 10b.



(**a**) TripAdvisor attraction (**b**) TripAdvisor restaurant (**c**) Twitter

**Figure 9.** Visualization of the networks and nodes where large node size represents boundary nodes. Communities are colored the same as Figure 6.




(**a**) (**b**)

**Figure 10.** Two examples of communities with boundary nodes and their respective topics. (**a**) An example from TripAdvisor attractions. (**b**) An example from TripAdvisor restaurants.

## 6. Conclusions

A place is a geographic location that has individuals' experiences and meaning-making processes from it. As a place has different meanings to different individuals, it is difficult to summarize the collective sense of place (e.g., [4,13,16]). As more and more textual data that describe people's experiences of places are now available online via social media etc., it is possible to study places by crowdsourcing the online textual data from place reviews and geolocated social media. Furthermore, through the use of network science and natural language processing of crowdsourced experiences collected from individuals we can also study the connections between places which reveals not only richer information about the place itself and how places relate to each other.

To this end, in this paper we used TripAdvisor reviews and geolocated Twitter data to understand the characteristics and the connectedness of places. The complex relationships between places were modeled via thematic (i.e., topical) similarity networks. While previous research using crowdsourced data allowed for the clustering of places (e.g., [32]), they do not explore the connections between places and clusters. The network approach developed in this paper enables us to perform clustering (i.e., network community detection) to discover the network and node level patterns of places. More specifically, the contributions of this research are as follows. First, similar to previous research with place clustering, the network approach of places also allow the performance of clustering on places using a network community detection algorithm. The case study in Section 5.1 show that community detected from the thematic similarity network from restaurant reviews tend to have higher Moran's I value (i.e., geographical proximity) than that of attraction reviews and tweets. It suggests that certain geographical clusters correspond to certain restaurant culture in Manhattan (e.g., the bar

and pub area at Downtown). Second, by using the network approach (as discussed in Section 4), we can discover places of interest by exploiting the positions of the places in the network. In the case study shown in the paper, the places of interest are places of different levels of uniqueness (Section 5.3). Third, from the TripAdvisor restaurant network results, we found that even though most of the study area in Manhattan is high income, the low-income communities have a distinctive restaurant culture that the high-income areas do not have (Section 5.2). Fourth, by comparing different datasets (i.e., Trip Advisor restaurants and attractions reviews and Twitter), we show implications of using such data for studying places (Section 5.1). TripAdvisor review data represents experiences and perceptions people have directly about places, whereas geolocated Twitter data does not necessarily reflect places. However, as our case study shows, by using topic modeling one can overcome this challenge and filter out place-irrelevant topics, which do not require the time consuming hand labeling process as supervised learning (as shown in Section 5.1.1).

While this study has shown how places are connected through individuals' experiences and adds to the growing area of geographic data science [62], there are several limitations to this research. First, although the clustering algorithm used in this study (i.e., the Girvan–Newman algorithm [56]), produces deterministic results, it might not be an ideal choice when the networks become larger, say when expanding this research to larger areas. Therefore, researchers who expand this research might want to consider less computationally inexpensive algorithms such as Louvain community detection algorithm [63] which use modularity optimization and has been shown to be scalable [64]. Second, in this study, we define places as census tracts and further analysis is required to test whether some of the results still stand when places are defined otherwise (e.g., zip codes, city blocks etc.). Nonetheless, using census tracts in this research had the advantage of combining textual VGI data with Census data for further analysis (as shown in Section 5.2). Turning to future work, other centrality measures (e.g., betweenness centrality, eigenvector centrality) could be explored to discover places of interest other than degree centrality and boundary nodes. Additionally, topic models could be trained by merging data from three datasets so that the topics are comparable across networks. As this work does not take tokens such as emojis into consideration, future work could explore topic models by incorporating them (e.g., [65,66]). The network can also constructed differently with edges representing similarities measured by methods other than topic similarity such as similarity based on users' visit history, which has often been used in collaborative recommender systems [67]. Even with these limitations and potential areas of further work, the research presented in this paper demonstrates a novel approach of studying places and their connections by combining textual VGI with network analysis.

**Author Contributions:** All authors contributed to the design of the methodology and the writing of the paper. Xiaoyi Yuan collected the data and carried out the analysis. All the authors contributed to the preparation of the manuscript and approved the final version to be published. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

**Table A1.** Moran's I for all communities.

| Network | Community ID | Moran's I | *p* Value |
|---|---|---|---|
| TripAdvisor Attractions | 1 | −0.022740 | 0.413 |
| | 2 | −0.002260 | 0.105 |
| | 3 | 0.000641 | 0.054 |
| | 4 | 0.120743 | 0.014 * |
| | 5 | 0.162247 | 0.010 ** |
| | 6 | 0.142676 | 0.006 ** |
| | 7 | −0.020767 | 0.035 * |
| | 8 | 0.002285 | 0.410 |
| | 9 | −0.003477 | 0.131 |
| | 10 | 0.057185 | 0.097 |
| | 11 | 0.141883 | 0.015 * |
| | 12 | 0.107151 | 0.014 * |
| | 13 | 0.106869 | 0.034 * |
| | 14 | 0.053348 | 0.111 |
| | 15 | −0.005520 | 0.326 |
| | 16 | −0.005640 | 0.280 |
| | 17 | −0.006208 | 0.157 |
| | 18 | −0.005532 | 0.345 |
| | 19 | −0.015263 | 0.430 |
| | 20 | −0.000060 | 0.073 |
| | 21 | −0.007165 | 0.067 |
| | 22 | −0.005024 | 0.492 |
| | 23 | −0.004145 | 0.208 |
| | 24 | −0.006130 | 0.140 |
| | 25 | −0.005303 | 0.389 |
| | 26 | −0.009513 | 0.381 |
| | 27 | 0.171691 | 0.009 ** |
| | 28 | 0.338116 | 0.002 ** |
| TripAdvisor Restaurants | 1 | 0.004695 | 0.052 |
| | 2 | 0.120561 | 0.017 * |
| | 3 | 0.682913 | 0.001 *** |
| | 4 | −0.001215 | 0.045 * |
| | 5 | 0.459894 | 0.002 ** |
| | 6 | −0.003185 | 0.100 |
| | 7 | −0.009306 | 0.014 * |
| | 8 | 0.223211 | 0.001 *** |
| | 9 | −0.005161 | 0.416 |
| | 10 | 0.174987 | 0.001 *** |
| | 11 | −0.006029 | 0.140 |
| | 12 | −0.025863 | 0.045 * |
| | 13 | 0.112362 | 0.026 * |
| | 14 | 0.167896 | 0.010 ** |
| | 15 | 0.038441 | 0.170 |
| | 16 | −0.004510 | 0.365 |
| | 17 | 0.255313 | 0.002 ** |
| | 18 | −0.003777 | 0.186 |
| | 19 | 0.116421 | 0.020 * |
| | 20 | −0.005524 | 0.266 |
| | 21 | −0.004135 | 0.232 |
| | 22 | −0.006081 | 0.139 |
| | 23 | 0.345028 | 0.001 *** |
| | 24 | 0.085735 | 0.044 * |
| | 25 | −0.009043 | 0.358 |
| | 26 | −0.004745 | 0.425 |

**Table A1.** *Cont.*

| Network | Community ID | Moran's I | *p* Value |
|---|---|---|---|
| TripAdvisor Restaurants | 27 | 0.102172 | 0.030 * |
| | 28 | 0.076950 | 0.046 * |
| | 29 | −0.006169 | 0.111 |
| | 30 | −0.005366 | 0.317 |
| | 31 | −0.004373 | 0.352 |
| | 32 | −0.004645 | 0.415 |
| | 33 | −0.011160 | 0.188 |
| | 34 | −0.004428 | 0.350 |
| | 35 | −0.005358 | 0.343 |
| | 36 | −0.004536 | 0.124 |
| | 37 | −0.008849 | 0.015 * |
| | 38 | −0.004531 | 0.391 |
| | 39 | −0.005106 | 0.442 |
| | 40 | −0.004121 | 0.251 |
| | 41 | −0.004531 | 0.390 |
| | 42 | −0.004252 | 0.282 |
| | 43 | −0.005566 | 0.301 |
| | 44 | 0.572042 | 0.001 *** |
| | 45 | −0.004745 | 0.444 |
| | 46 | −0.004745 | 0.429 |
| | 47 | −0.005653 | 0.277 |
| | 48 | −0.005114 | 0.443 |
| | 49 | −0.005894 | 0.195 |
| | 50 | −0.002693 | 0.085 |
| | 51 | 0.004695 | 0.038 * |
| | 52 | −0.004088 | 0.232 |
| | 53 | −0.005155 | 0.420 |
| | 54 | −0.002693 | 0.075 |
| | 55 | −0.007618 | 0.024 * |
| | 56 | 0.004695 | 0.053 |
| | 57 | 0.004695 | 0.043 * |
| | 58 | 0.004695 | 0.042 * |
| Twitter | 1 | 0.004695 | 0.046 * |
| | 2 | −0.003338 | 0.086 |
| | 3 | −0.006387 | 0.161 |
| | 4 | 0.004695 | 0.046 * |
| | 5 | −0.001215 | 0.048 * |
| | 6 | 0.004695 | 0.046 * |
| | 7 | −0.010278 | 0.404 |
| | 8 | −0.002693 | 0.095 |
| | 9 | −0.004061 | 0.221 |
| | 10 | 0.328338 | 0.001 *** |
| | 11 | −0.012091 | 0.006 ** |
| | 12 | 0.056188 | 0.085 |
| | 13 | −0.010706 | 0.293 |
| | 14 | −0.010190 | 0.414 |
| | 15 | −0.003404 | 0.113 |
| | 16 | −0.007665 | 0.031 * |
| | 17 | 0.118175 | 0.016 * |
| | 18 | 0.194717 | 0.013 * |
| | 19 | −0.004334 | 0.347 |
| | 20 | −0.005818 | 0.224 |
| | 21 | −0.015831 | 0.304 |
| | 22 | −0.006734 | 0.076 |
| | 23 | −0.009470 | 0.010 * |
| | 24 | 0.019150 | 0.246 |
| | 25 | 0.196924 | 0.005 ** |
| | 26 | 0.077334 | 0.063 |
| | 27 | −0.003777 | 0.169 |
| | 28 | −0.009687 | 0.451 |

* Significant at $p \leq 0.05$; ** Significant at $p \leq 0.01$; *** Significant at $p \leq 0.001$.

## References


1. Goodchild, M.F. Formalizing Place in Geographic Information Systems. In *Communities, Neighborhoods, and Health: Expanding the Boundaries of Place*; Burton, L.M., Matthews, S.A., Leung, M., Kemp, S.P., Takeuchi, D.T., Eds.; Social Disparities in Health and Health Care; Springer: New York, NY, USA, 2011; pp. 21–33. [CrossRef]
2. Tuan, Y.F. *Space and Place: The Perspective of Experience*; U of Minnesota Press: Minneapolis, MN, USA, 1977.
3. Agnew, J. Space and Place. *SAGE Handb. Geogr. Knowl.* **2011**, *23*, 316–330.
4. Cresswell, T. *Place: An Introduction*; John Wiley & Sons: Chichester, UK, 2014. Available online: http://xxx.lanl.gov/abs/sdzhBQAAQBAJ (accessed on 3 December 2014).
5. Shevky, E.; Bell, W. *Social Area Analysis; Theory, Illustrative Application and Computational Procedures*; Stanford University Press: Palo Alto, CA, USA, 1955.
6. Anderson, T.R.; Egeland, J.A. Spatial Aspects of Social Area Analysis. *Am. Sociol. Rev.* **1961**, *26*, 392–398. [CrossRef]
7. Spielman, S.E.; Thill, J.C. Social Area Analysis, Data Mining, and GIS. *Comput. Environ. Urban Syst.* **2008**, *32*, 110–122. [CrossRef]
8. Spicker, P. Charles Booth: The Examination of Poverty. *Soc. Policy Adm.* **1990**, *24*, 21–38. [CrossRef]
9. Webber, R. Papers: Designing Geodemographic Classifications to Meet Contemporary Business Needs. *Interact. Mark.* **2004**, *5*, 219–237. [CrossRef]
10. Singleton, A.D.; Longley, P.A. Creating Open Source Geodemographics: Refining a National Classification of Census Output Areas for Applications in Higher Education. *Pap. Reg. Sci.* **2009**, *88*, 643–666. [CrossRef]
11. Goodchild, M.F. Citizens as Sensors: The World of Volunteered Geography. *GeoJournal* **2007**, *69*, 211–221. [CrossRef]
12. Crooks, A.; Pfoser, D.; Jenkins, A.; Croitoru, A.; Stefanidis, A.; Smith, D.; Karagiorgou, S.; Efentakis, A.; Lamprianidis, G. Crowdsourcing Urban Form and Function. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 720–741. [CrossRef]
13. Stefanidis, A.; Crooks, A.; Radzikowski, J. Harvesting Ambient Geospatial Information from Social Media Feeds. *GeoJournal* **2013**, *78*, 319–338. [CrossRef]
14. Sui, D.; Elwood, S.; Goodchild, M. *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*; Springer Science & Business Media: Dordrecht, Netherlands, 2012; Available online: http://xxx.lanl.gov/abs/SSbHUpSk2MsC (accessed on 15 February 2020).
15. MacEachren, A.M. Leveraging Big (Geo) Data with (Geo) Visual Analytics: Place as the Next Frontier. In *Spatial Data Handling in Big Data Era: Select Papers from the 17th IGU Spatial Data Handling Symposium*; Zhou, C., Su, F., Harvey, F., Xu, J., Eds.; Advances in Geographic Information Science; Springer: Singapore, 2017; pp. 139–155. [CrossRef]
16. Hu, Y. 1.07—Geospatial Semantics. In *Comprehensive Geographic Information Systems*; Huang, B., Ed.; Elsevier: Amsterdam, The Netherlands, 2018; pp. 80–94. [CrossRef]
17. Yuan, X.; Crooks, A. Assessing the Placeness of Locations through User-Contributed Content. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 15–23. [CrossRef]
18. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
19. Ballatore, A.; Adams, B. Extracting Place Emotions from Travel Blogs. *Proc. AGILE* **2015**, *2015*, 1–5.
20. Adams, B.; McKenzie, G. Inferring Thematic Places from Spatially Referenced Natural Language Descriptions. In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*; Sui, D., Elwood, S., Goodchild, M., Eds.; Springer: Dordrecht, The Netherlands, 2013; pp. 201–221. [CrossRef]
21. Mai, G.; Janowicz, K.; Prasad, S.; Yan, B. Visualizing the Semantic Similarity of Geographic Features. In Proceedings of the AGILE Conference, Lund, Sweden, 12–15 June 2018; pp. 12–15.
22. Hu, Y.; McKenzie, G.; Janowicz, K.; Gao, S. Mining Human-Place Interaction Patterns from Location-Based Social Networks to Enrich Place Categorization Systems. In Proceedings of the Workshop on Cognitive Engineering for Spatial Information Processes at COSIT 2015, Santa Fe, NM, USA, 12 October 2015.
23. Hasan, S.; Ukkusuri, S.V. Urban Activity Pattern Classification Using Topic Models from Online Geo-Location Data. *Transp. Res. Part C Emerg. Technol.* **2014**, *44*, 363–381. [CrossRef]

24. Gao, S.; Janowicz, K.; Couclelis, H. Extracting Urban Functional Regions from Points of Interest and Human Activities on Location-Based Social Networks. *Trans. GIS* **2017**, *21*, 446–467. [CrossRef]
25. Yin, Z.; Cao, L.; Han, J.; Zhai, C.; Huang, T. Geographical Topic Discovery and Comparison. In *Proceedings of the 20th International Conference on World Wide Web*; Association for Computing Machinery: New York, NY, USA, 2011; pp. 247–256. [CrossRef]
26. Mei, Q.; Liu, C.; Su, H.; Zhai, C. A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs. In *Proceedings of the 15th International Conference on World Wide Web*; Association for Computing Machinery: New York, NY, USA, 2006; pp. 533–542. [CrossRef]
27. Wang, C.; Wang, J.; Xie, X.; Ma, W.Y. Mining Geographic Knowledge Using Location Aware Topic Model. In *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*; Association for Computing Machinery: New York, NY, USA, 2007; pp. 65–70. [CrossRef]
28. Hao, Q.; Cai, R.; Wang, C.; Xiao, R.; Yang, J.M.; Pang, Y.; Zhang, L. Equip Tourists with Knowledge Mined from Travelogues. In *Proceedings of the 19th International Conference on World Wide Web*; Association for Computing Machinery: New York, NY, USA, 2010; pp. 401–410. [CrossRef]
29. Hu, B.; Ester, M. Spatial Topic Modeling in Online Social Media for Location Recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems*; Association for Computing Machinery: New York, NY, USA, 2013; pp. 25–32. [CrossRef]
30. Yuan, Q.; Cong, G.; Ma, Z.; Sun, A.; Thalmann, N.M. Who, Where, When and What: Discover Spatio-Temporal Topics for Twitter Users. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2013; pp. 605–613. [CrossRef]
31. Schmid, K.A.; Züfle, A.; Pfoser, D.; Crooks, A.; Croitoru, A.; Stefanidis, A. Predicting the evolution of narratives in social media. In *International Symposium on Spatial and Temporal Databases*; Springer: Cham, Switzerland, 2017; pp. 388–392.
32. Jenkins, A.; Croitoru, A.; Crooks, A.T.; Stefanidis, A. Crowdsourcing a Collective Sense of Place. *PLoS ONE* **2016**, *11*, e0152932. [CrossRef] [PubMed]
33. Teng, X.; Yang, J.; Kim, J.S.; Trajcevski, G.; Züfle, A.; Nascimento, M.A. Fine-Grained Diversification of Proximity Constrained Queries on Road Networks. In *Proceedings of the 16th International Symposium on Spatial and Temporal Databases*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 51–60.
34. Cranshaw, J.; Schwartz, R.; Hong, J.; Sadeh, N. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In *Sixth International AAAI Conference on Weblogs and Social Media*; The AAAI Press: Palo Alto, CA, USA, 2012.
35. Preoţiuc-Pietro, D.; Cranshaw, J.; Yano, T. Exploring Venue-Based City-to-City Similarity Measures. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*; Association for Computing Machinery: New York, NY, USA, 2013; pp. 1–4. [CrossRef]
36. Noulas, A.; Scellato, S.; Mascolo, C.; Pontil, M. Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-Based Social Networks. In *Fifth International AAAI Conference on Weblogs and Social Media*; The AAAI Press: Palo Alto, CA, USA, 2011.
37. Adams, B.; Raubal, M. Identifying Salient Topics for Personalized Place Similarity. *Res. Locate* **2014**, *14*, 1–12.
38. Janowicz, K.; Raubal, M.; Kuhn, W. The Semantics of Similarity in Geographic Information Retrieval. *J. Spat. Inf. Sci.* **2011**, *2011*, 29–57. [CrossRef]
39. Yan, B.; Janowicz, K.; Mai, G.; Gao, S. From ITDL to Place2Vec: Reasoning About Place Type Similarity and Relatedness by Learning Embeddings From Augmented Spatial Contexts. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*; Association for Computing Machinery: New York, NY, USA, 2017; pp. 1–10. [CrossRef]
40. Quercini, G.; Samet, H. Uncovering the Spatial Relatedness in Wikipedia. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*; Association for Computing Machinery: New York, NY, USA, 2014; pp. 153–162. [CrossRef]
41. Hu, Y.; Ye, X.; Shaw, S.L. Extracting and Analyzing Semantic Relatedness between Cities Using News Articles. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 2427–2451. [CrossRef]
42. Valavanis, I.; Spyrou, G.; Nikita, K. A Similarity Network Approach for the Analysis and Comparison of Protein Sequence/Structure Sets. *J. Biomed. Inform.* **2010**, *43*, 257–267. [CrossRef] [PubMed]

43. Wang, B.; Mezlini, A.M.; Demir, F.; Fiume, M.; Tu, Z.; Brudno, M.; Haibe-Kains, B.; Goldenberg, A. Similarity Network Fusion for Aggregating Data Types on a Genomic Scale. *Nat. Methods* **2014**, *11*, 333–337. [CrossRef]
44. Brown, S.A. Patient Similarity: Emerging Concepts in Systems and Precision Medicine. *Front. Physiol.* **2016**, *7*. [CrossRef]
45. Pai, S.; Bader, G.D. Patient Similarity Networks for Precision Medicine. *J. Mol. Biol.* **2018**, *430*, 2924–2938. [CrossRef] [PubMed]
46. Google Geocoding API. Available online: https://developers.google.com/maps/documentation/geocoding/start (accessed on 3 February 2020).
47. US Census. Available online: https://www2.census.gov/geo/pdfs/education/CensusTracts.pdf (accessed on 3 February 2020).
48. Openshaw, S. The Modifiable Areal Unit Problem. *Quant. Geogr. Br. View* **1981**, 60–69.
49. Kouloumpis, E.; Wilson, T.; Moore, J. Twitter Sentiment Analysis: The Good the Bad and the Omg! In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*; The AAAI Press: Menlo Park, CA, USA, 2011.
50. Boyd-Graber, J.; Mimno, D.; Newman, D. Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements. *Handb. Mixed Membsh. Model. Their Appl.* **2014**, *226–250*, 225255.
51. Schofield, A.; Mimno, D. Comparing Apples to Apple: The Effects of Stemmers on Topic Models. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 287–300. [CrossRef]
52. Řehůřek, R.; Sojka, P. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*; ELRA: Valletta, Malta, 2010; pp. 45–50. Available online: http://is.muni.cz/publication/884893/en (accessed on 15 February, 2020).
53. Stevens, K.; Kegelmeyer, P.; Andrzejewski, D.; Buttler, D. Exploring Topic Coherence over Many Models and Many Topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2012; pp. 952–961.
54. Serrano, M.A.; Boguná, M.; Vespignani, A. Extracting the Multiscale Backbone of Complex Weighted Networks. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 6483–6488. [CrossRef] [PubMed]
55. Lindner, G.; Staudt, C.L.; Hamann, M.; Meyerhenke, H.; Wagner, D. Structure-Preserving Sparsification of Social Networks. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Paris, France, 25–28 August 2015; pp. 448–454. [CrossRef]
56. Newman, M.E.J. Analysis of Weighted Networks. *Phys. Rev. E* **2004**, *70*, 056131. [CrossRef] [PubMed]
57. Newman, M.E.J. Modularity and Community Structure in Networks. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 8577–8582. [CrossRef] [PubMed]
58. Guerra, P.C.; Meira, W., Jr.; Cardie, C.; Kleinberg, R. A Measure of Polarization on Social Media Networks Based on Community Boundaries. In *Seventh International AAAI Conference on Weblogs and Social Media*; The AAAI Press: Menlo Park, CA, USA, 2013.
59. Schaeffer, S.E. Graph Clustering. *Comput. Sci. Rev.* **2007**, *1*, 27–64. [CrossRef]
60. Bastian, M.; Heymann, S.; Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. In *Third International AAAI Conference on Weblogs and Social Media*; The AAAI Press: Menlo Park, CA, USA, 2009.
61. Spielman, S.E.; Singleton, A. Studying Neighborhoods Using Uncertain Data from the American Community Survey: A Contextual Approach. *Ann. Assoc. Am. Geogr.* **2015**, *105*, 1003–1025. [CrossRef]
62. Singleton, A.; Arribas-Bel, D. Geographic Data Science. *Geogr. Anal.* **2019**. [CrossRef]
63. Blondel, V.D.; Guillaume, J.L.; Lambiotte, R.; Lefebvre, E. Fast Unfolding of Communities in Large Networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. [CrossRef]
64. Que, X.; Checconi, F.; Petrini, F.; Gunnels, J.A. Scalable Community Detection with the Louvain Algorithm. In Proceedings of the 2015 IEEE International Parallel and Distributed Processing Symposium, Hyderabad, India, 25–29 May 2015; pp. 28–37. [CrossRef]
65. Swartz, M.; Crooks, A. Comparison of Emoji Use in Names, Profiles, and Tweets. In Proceedings of the 2020 IEEE 14th International Conference on Semantic Computing (ICSC), San Diego, CA, USA, 3–5 February 2020; pp. 375–380.

66. Swartz, M.; Crooks, A.T.; Kennedy, W. Emoji and Keyword Cues for Diversity in Social Media. In Proceedings of the 11th International Conference on Social Media and Society, Online, 22 July 2020.
67. Almazro, D.; Shahatah, G.; Albdulkarim, L.; Kherees, M.; Martinez, R.; Nzoukou, W. A Survey Paper on Recommender Systems. 2010. Available online: http://xxx.lanl.gov/abs/1006.5278 (accessed on 15 February 2020).