A glimpse of big data: how social media can inform urban design

Yang Song, North Dakota State University

Jessica Fernandez, Clemson University

The last few decades have seen an explosion in the phenomenon known as ‘big data.’ A product of social media, smart phones, sensors, and the internet, large quantities of data have never been more accessible or more powerful. Big data technologies have been widely adopted in a variety of industries including finance, marketing, energy, telecommunications, agriculture, and real estate to make more informed decisions and to be able to predict trends. Many publicly available websites such as Yelp, TripAdvisor, and Realtor provide a plethora of information such as behavior and perceptions associated with places and communities. However, few studies have been published in the fields of planning and the built environment to examine the efficacy of big data use and implementation.

This paper presents a case study using the social media website TripAdvisor to inform the concept development of an urban design project in Chicago, Illinois. By extracting, processing, and analyzing large amounts of geocoded information in TripAdvisor, the patterns of urban activities that were previously inaccessible to designers become accessible and useful. The study also displays how implementing the research of data in TripAdvisor supports a better understanding of the relationship between people and places, and advances the processes of urban transformation. As a response, landscape architects and urban planners might realize the potential of a big data approach, influencing the future of design and research related to the built environment.

A glimpse of big data: how social media can inform urban design

Keywords**:** Social media, Urban design, TripAdvisor, Social activities.

Introduction

Data is being created, copied and transformed with unprecedented speed and scale (The Economist, 2010). The abundance of social media, smart phones, sensors, and the internet all sense, share and process data around the world. People are increasingly able to search, aggregate, and analyze data with the exponential increase of volume, velocity and variety of data available (Chandler, 2015). Big data technologies have been widely adopted in a variety of industries including finance, marketing, energy, telecommunications, agriculture, and real estate to make more informed decisions, as well as to be able to predict trends. Boyd and Crawford (2012) note that "[from] big data has emerged a system of knowledge that is already changing the objects of knowledge, while also having the power to inform how we understand human networks and community."

Urban design includes the design and shaping of cities and other urban spaces in a variety of scales. The scope includes not only the formation of place, but also considers infrastructure and amenities for neighborhoods, districts, and even entire cities. Making decisions on urban issues such as typology and density, pedestrian zones, greenspaces, aesthetics, accessibility, and sustainability requires a deep and comprehensive understanding of how the built environment influences public behavior and cultural activities. Therefore, analysis of societal needs and public perceptions becomes key to building a more equitable urban environment that is beneficial to a broad spectrum of people (Dobbins, 2009).

Nowadays, many publicly available online databases such as Yelp, TripAdvisor, and Realtor provide a plethora of information such as behavior and perceptions associated with places and communities. Usage of this data has the great potential to assist urban designers studying the preferences of city users. However, few studies have been published in the fields of planning and design of the built environment to examine the efficacy of big data use and implementation. This study presents the value of utilizing insights from TripAdvisor towards a more in-depth analysis on urban activities.

Data collection

This study has developed a database from TripAdvisor (www.tripadvisor.com). TripAdvisor is one of the most highly used online platforms related to tourism. Known worldwide, TripAdvisor also includes the most complete and current human attraction profiles. Therefore, attraction review information from this website is used as the source for data sampling in this study. The data focuses on the attractions in the TripAdvisor "things to do" category, excluding data on hotels, flights, and vacations rentals. For this study, 1604 attraction profiles from Chicago, Illinois cover eighteen categories including sights and landmarks, museums, tours, nature and parks, outdoor activities, concerts and shows, boat tours and water sports, zoos and aquariums, water and amusement parks, nightlife, food and drink, shopping, fun and games, transportation, traveler resources, spas and wellness, classes and Fjoworkshops, and events.

Specific attributes for each attraction profile were collected. An example of one segment of attraction data is listed in Table 1. Basic information such as attraction name, ranking by TripAdvisor for its particular type category, location information such as latitude and longitude, review information such as the number of reviews rated as Excellent, Very good, Average, Poor

and Terrible, and time information such as the year of the the first review posted were included. In the end, a database with all these information of 1604 attractions was created.

Id	Name	Ranking	Type	Latitude	Longitude	Excellent	Very good	Average	Poor	Terrible	Year
10	Grant Park	41	Nature & Parks	41.882789	-87.618923	261	186	48	12	5	7

Table 1. Data of one attraction in TripAdvisor

Analysis and Results

This study explores the potential of applying big data from TripAdvisor for urban design studies. Three different forms of analysis are presented in the following sections, including location and categorical information, attraction index and proximity, and time and growth.

Location and categorical information

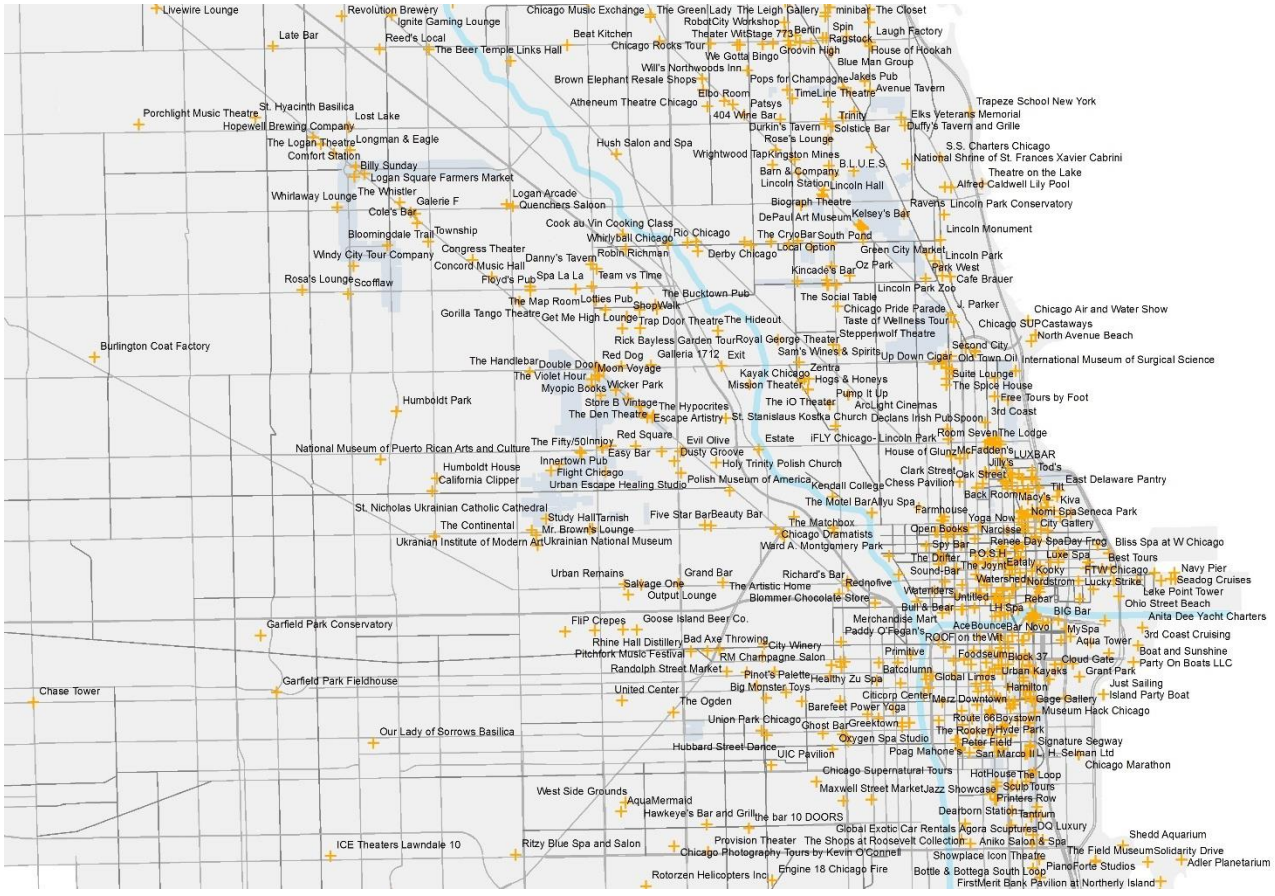

Figure 1: Locations of attractions in TripAdvisor

TripAdvisor data was imported into ArcGIS for analysis and visualization. In order to provide a more clear representation, only the Chicago, Illinois downtown and north Chicago were included in the analysis extent. The location (longititude and latitude) of each attraction was

obtianed by a geocoding service provided by Google. As Figure 1 shows, all of the attraction points were drawn and labeled by their names.

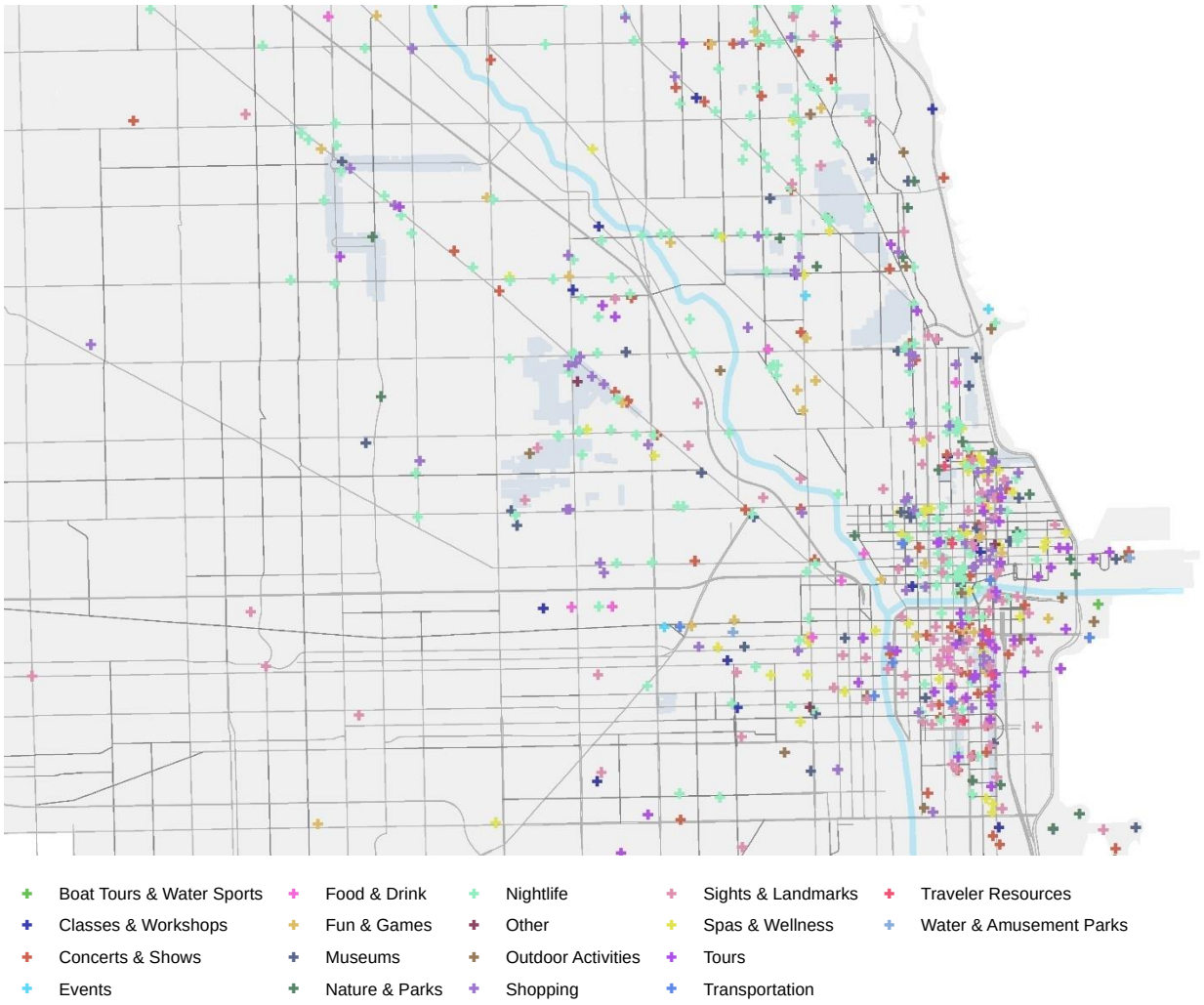

Figure 2: Categories of the attractions in TripAdvisor

The comprehensive classification system available on TripAdvisor made it possible to plot the categories of all of the attractions to show the distribution of urban services and amenities. Figure 2 presents eighteen categories to the extents of the scope of this study.

Attraction index and proximity

With hundreds of attractions included, it can be assumed that not every attraction has the same impact on the overall urban living experience. Some attractions have more reviews and higher ratings than others, have gained higher exposure, and could be considered more valued by the residents. The urban areas near these attractions are usually well developed, with many relevant and well-liked businesses. In Chicago, for example, Millennium Park shows to be the strongest hub of the waterfront area. The Chicago Riverwalk also appears to play an important role in promoting the development of the inner downtown commercial area.

To expand upon these observations, this study introduced an index to indicate the level of significance of each attraction. The index was calculated using a true Bayesian estimate ("Bayes estimator," 2017), a similar approach to how IMDB rates movie titles. The formula for calculating the attraction index is as follows:

weighted index (WI) = (v ÷ (v+m)) × R + (m ÷ (v+m)) × C

where: R = average rating for the attraction(excellent as 5, very good as 4, average as 3,poor as 2, terrible as 1); v = number of review for the attraction; m = minimum reviews as reasonable mount (currently 10); C = the mean rating across the whole dataset including attractions less than minimum reviews (currently 4.3032960467363619).

This index takes both the number of reviews and the rating of reviews into consideration, and systematically weights them into one value. It provides an appropriate solution to balancing highly rated attractions with a small number of reviews against attractions with a high number of reviews and lower ratings. Figure 3 shows the analysis of all attractions in the extent of the study. The size of the red dots represents the different levels of attraction index. Larger dots indicate a higher value on the attraction index, signifying higher review amounts and ratings.

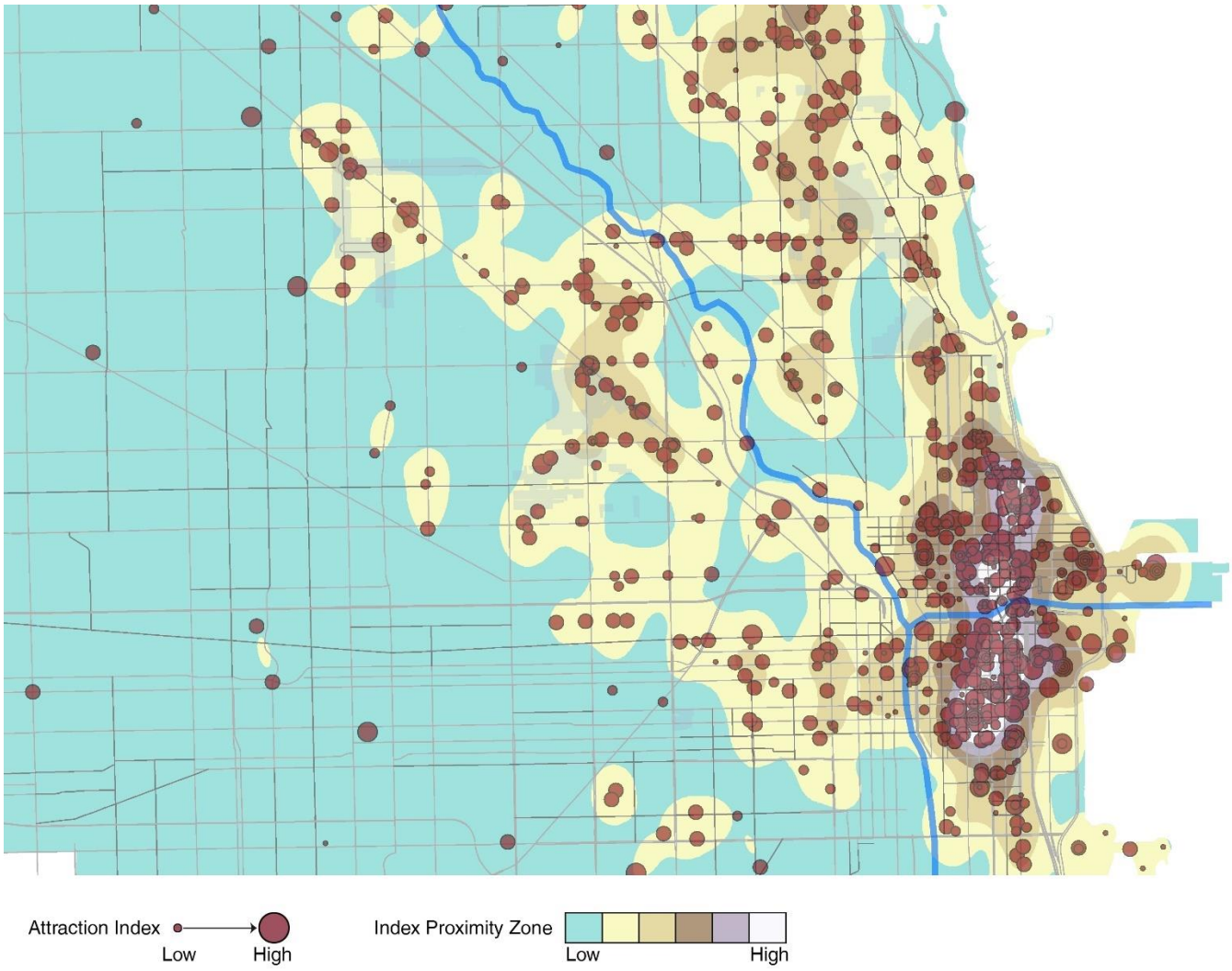

Figure 3 Attraction Index and Index Proximity Zone

Furthermore, through use of the kernel density function in ArcGIS, Index Proximity Zones were calculated and mapped. This map identifies areas that share similar access to the attraction resources. The calculation depends on the distance to the nearby attractions and their index values. For example, in Figure 3 the areas with light yellow color signify somewhat low access to the attraction resources, meaning that those areas either have attractions with low index value or have fewer attractions nearby.

According to Figure 3, the Chicago downtown area has large amount of attractions with high index values. The other significant clusters are located around Lincoln park (north) and Wicker park (northwest). Through this mapping and data analysis process, urban designers and planners can see the specific distributions and patterns occurring based on the documented impact of TripAdvisor attractions. Since the indexes typically represent thousands of online reviews, their reliability and validity is stronger than various typical approaches of urban designers for data collection and assessment including personal experience, sampling the opinions of a sub-section of a population, anecdotal observation, and online articles and blogs.

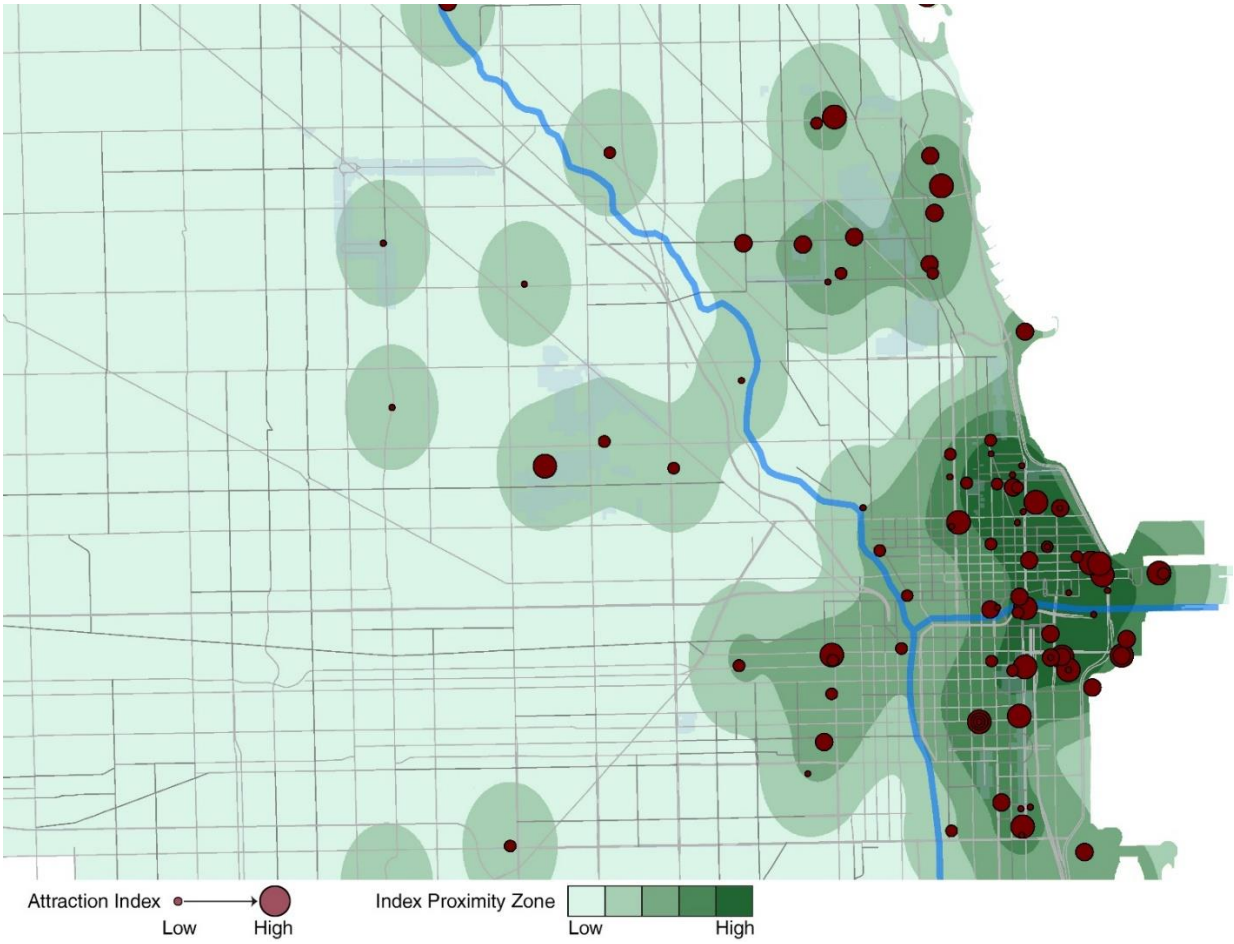

Figure 4: Attraction Index and Index Proximity Zone for the category of health and outdoor activities.

Moreover, the attraction index can be used to assess specific categories of attractions. For example, in Figure 4, dots specifically display attractions related to outdoor and health activities within the TripAdvisor categories of Outdoor Activities, Water and Amusement Parks, Nature and Parks, Boat Tours and Water Sports, and Spas and Wellness. The method provides specific insights beneficial to the planning of greenspace and development of outdoor amenities.

Time and Growth

Perhaps the most unique component of TripAdvisor is its documentation of time. Every review in each attraction has an associated time tag, providing great potential for assessment of places over time. This study explores one basic application of the growth of attractions over time. Because the time (year) of the first review of each attraction was collected in the database, this study can present when the attractions started on TripAdvisor. Figures 5-8 display how the index and proximity zones of attractions developed over the past sixteen years (2001- 2017).

Each of figures 5-8 indicate the extent and intensity of development during the specific indicated timeframe. The downtown district has the most significant cluster of attraction growth throughout the seventeen-year period assessed in this study. Aside from the downtown area, the Lincoln park region (north of downtown) started to popularize during the 2008-2012 timeframe. Wicker park region (north west of downtown) become the fastest area to grow in popular attractions in recent years (2013-2017). Overall, most attractions showed up along the Chicago river, with the west side of the river growing in desirable attractions at the fastest rate.

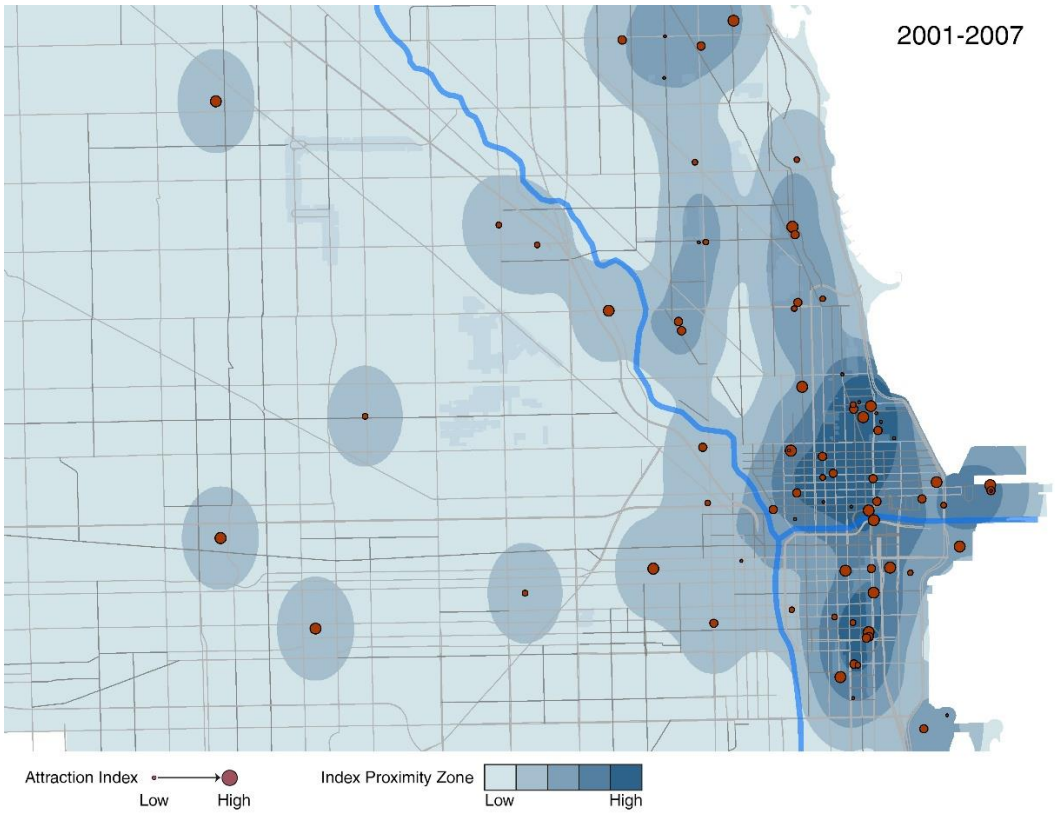

Figure 5: Attraction Index and Index Proximity Zone for the attractions during 2001-2007

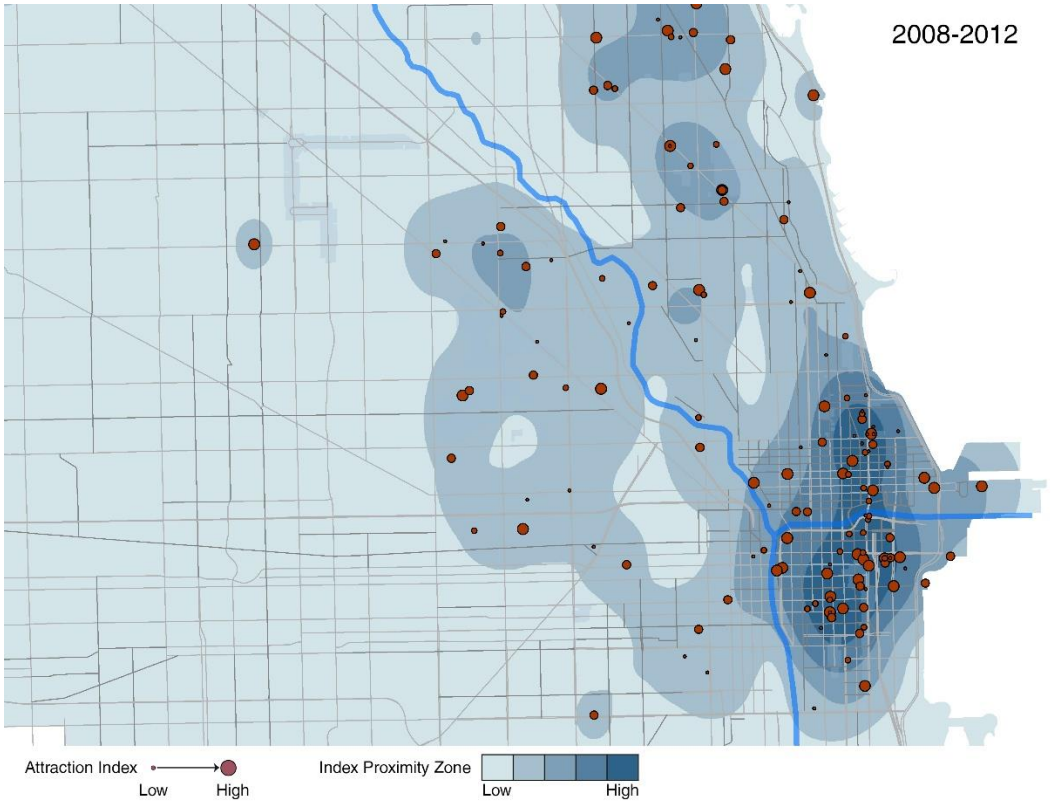

Figure 6: Attraction Index and Index Proximity Zone for the attractions during 2008-2012

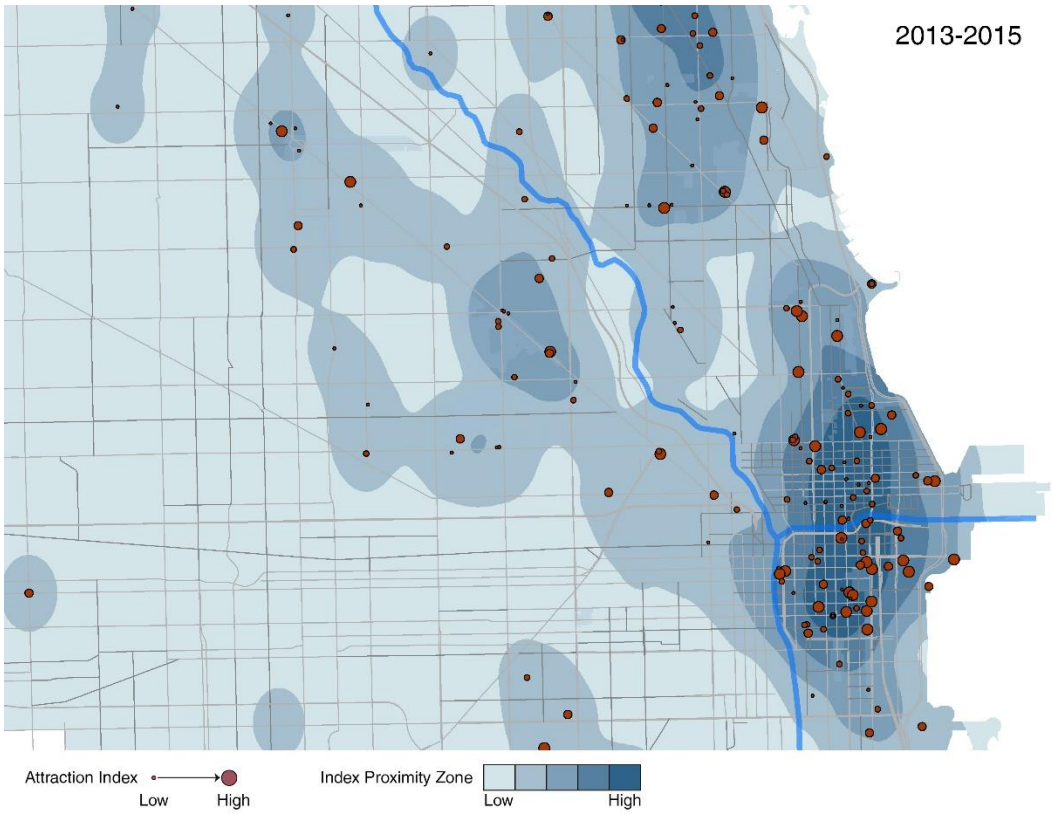

Figure 7: Attraction Index and Index Proximity Zone for the attractions during 2013-2015

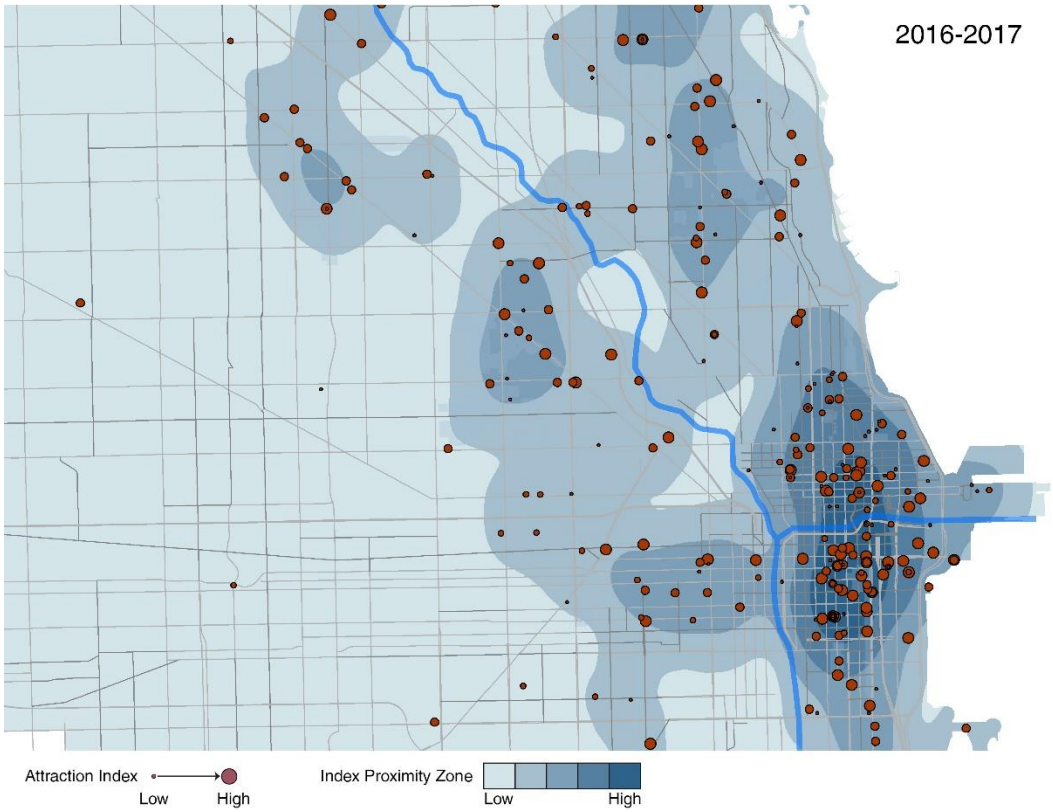

Figure 8: Attraction Index and Index Proximity Zone for the attractions during 2016-2017

Conclusion and Discussion

This study presents an approach to utilizing data from TripAdvisor, introducing three forms of analysis that might assist with the formation of more informed urban design decisions. First, the specific locations and categories of TripAdvisor attractions were drawn to display the distribution of the places where residents share their experiences online. Second, the attraction index and proximity zones were calculated to identify and classify the areas with significant urban activities. Third, a time and growth analysis documented how the study extent in Chicago prospered and transformed through time. By extracting, processing, and analyzing large amounts of geocoded information in TripAdvisor, this study identified the patterns of urban activities that were previously inaccessible to designers. The study also demonstrates how implementing big data in social media can support a better understanding of the relationship between people and places, and advances the processes of urban transformation.

However, future research on the application of big data specifically from TripAdvisor is needed. For instance, millions of review comments are associated with specific attractions and times. This data can provide a plethora of information to help understand the relationships between urban behaviors, public opinions, and urban design and management. Because urban design relates to a range of issues regarding public spaces, an improved awareness of people's opinions online can enable designers to create more responsive and holistic design solutions.

Bibliography:

Bayes estimator. (n.d.). In Wikipedia. Retrieved May 21, 2017, from https://en.wikipedia.org/wiki/Bayes_estimator#cite_note-7

'Data, Data Everywhere', The Economist, 15 February 2010. Available at: http://www.economist.com/node/15557443. Last accessed 26 February 2015.

Chandler, David. 2015. A World without Causation: Big Data and the Coming of Age of Posthumanism. Millennium-Journal of International Studies 43(3): 833–851.

Boyd, Danah, and Kate Crawford. 2012. CRITICAL QUESTIONS FOR BIG DATA Provocations for a cultural, technological, and scholarly phenomenon. Information Communication & Society 15(5): 662–679.

Boeing; et al. (2014). "LEED-ND and Livability Revisited". Berkeley Planning Journal. 27: 31–55. Retrieved 2015-04-15

Dobbins, Michael. 2009. Urban design and people. Hoboken, N.J.: Wiley