DISCOVERING A COLLECTIVE SENSE OF PLACE THROUGH CROWD-GENERATED CONTENT

by

Andrew Jenkins
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Earth Systems and GeoInformation Sciences

Committee:

__________________________________ Dr. Anthony Stefanidis, Dissertation Director

__________________________________ Dr. Arie Croitoru, Committee Member

__________________________________ Dr. Andrew Crooks, Committee Member

__________________________________ Dr. Dieter Pfoser, Committee Member

__________________________________ Dr. Anthony Stefanidis, Department Chairperson

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

__________________________________ Dr. Peggy Agouris, Dean, College of Science

Date: ___________________________ Spring Semester 2016
George Mason University
Fairfax, VA

Discovering A Collective Sense of Place Through Crowd Generated Content

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Andrew Jenkins
Master of Science
George Mason University, 2012
Bachelor of Science
University of Maryland University College, 2009

Director: Anthony Stefanidis, Professor
Department of Earth Systems and GeoInformation Sciences

Spring Semester 2016
George Mason University
Fairfax, VA

DEDICATION

I would like to dedicate this work to my loving family. Without their unconditional love and support none of this would have been possible. In particular, my dearest wife and best friend, Jaime, who selflessly assumed more work and child duties so that I may have time and fewer distractions to complete this work.

And yes Braelyn, daddy can go outside and play now. ☺

ACKNOWLEDGEMENTS

First and foremost I wish to thank my advisor, Dr. Anthony Stefanidis. Thank you, Tony, for your patience, guidance, and intellectual mentoring throughout this process and all my graduate studies. And thank you for believing in my abilities and for your invaluable insights, ideas, and concepts that profoundly shaped this research.

My sincere thanks and gratitude also go to the other members of my committee who supported me throughout this process. I want to thank Dr. Arie Croitoru for his technical insights and the many intellectual discussions that made a significant impact on this work. I also want to thank Dr. Andrew Crooks for his overall constructive comments and suggestions towards place literature. And last but not least, I want to thank Dr. Pfoser for his candid remarks and insightful comments on this dissertation.

I would also like to thank the faculty and staff of the Department of Geography and GeoInformation Science at GMU for their various forms of support and teachings over the course of my graduate studies. In addition, I am honored to have been selected by the faculty as the 2016 Outstanding PhD Student awardee.

I would like to gratefully acknowledge the GMU Office of the Provost for their research grant and award so that I may complete this research. I would also like to recognize the GMU University Libraries for their financial support in publishing a portion of this work in an open access journal.

I would like to thank my various professional colleagues and friends who I've had the pleasure working with over the course of my career. In particular, I would like to thank my many friends at the U.S. Army Geospatial Center and DigitalGlobe Inc.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS

Application Programming Interface	API
Author-topic-community	ATC
Bi-gram Topic Model	BTM
Central Park	CP
Density-based spatial clustering of applications with noise	DBSCAN
Named Entity Recognition	NER
Natural Language Processing	NLP
New York City	NYC
Non-metric Multidimensional Scaling	NMDS
Normalized Google Distance	NGD
OpenStreetMap	OSM
Pointwise Mutual Information	PMI
Points of Interest	POI
Principal Component Analysis	PCA
Theatre District	TSD
Topical n-gram model	TNG
Twitter	TW
Self-organizing maps	SOM
Sense of Place	SOP
Singapore	SG
Spatial Temporal Co-Occurrence Patterns	STCOPS
Latent Dirichlet Allocation	LDA
Los Angeles	LA
London	LDN
Lower Manhattan	LM
Volunteered Geographic Information	VGI
Wikipedia Link Vector Model	WLVM

ABSTRACT

DISCOVERING A COLLECTIVE SENSE OF PLACE THROUGH CROWD GENERATED CONTENT

Andrew Jenkins, Ph.D

George Mason University, 2016

Dissertation Director: Dr. Anthony Stefanidis

Place is generally defined as location given meaning through human experience. The topic of place has been widely debated and studied throughout geography and the social sciences as a theoretical construct. However, the rise and availability of user-generated content now affords new opportunities to computationally analyze and quantify the social meaning of place, but the question still remains of how well such content can be mined to discover place or so-called platial knowledge. This research investigates the question by focusing on the shared meaning of place by generalizing people's collective sense of place. It is argued that taking a crowd-centric approach of collective and implicit sense of place meanings will lead to the discovery of emerging platial themes. Moreover, given the semantic-spatial-temporal characteristics of human activities within urban spaces, one can observe the emergence of unique themes that characterize different locations. In this dissertation, a novel quantitative approach is presented with statistical validation for deriving such platial themes from crowd-contributed content. This approach leverages unsupervised probabilistic topical n-gram modelling for dimensionality reduction, knowledge base labelling using semantic association, and spatial clustering with iterative distance analysis. Experimental results are presented from four different study areas that depict the emergence of unique places, thematic alignment across different data sources, and co-occurrence trends. The discovery and identification of locations that convey a collective sense of place contributes to the goal of observing how people transform a location to a place and shape its characteristics.

CHAPTER 1: INTRODUCTION

1.1 Problem Statement

This dissertation seeks to contribute an approach to discover and represent the collective meaning of place from human centric views that emerge in user-generated content from the places themselves. As place is generally defined as location given meaning through human experience. The topic of place has been widely debated and studied throughout geography and the social sciences as a theoretical construct. The rise and availability of user-generated content continues to afford new opportunities to computationally analyze and quantify the social meaning of place, but the question still remains of how well such content can be mined to discover place or so-called platial knowledge.

The view of place from a dynamic human centric approach has the advantage of offering new insights taken from the bottom-up (individuals or groups) instead of top-down layering of static information. The traditional view of place has largely been an explicit geometric representation (i.e. place-name gazetteers) derived from authoritative processes that assumes the meaning rarely changes. Although this view of place has its utility, the challenge continues as to the discovery and representation of social meanings beyond these traditional space perspectives. An unanswered question remains as to whether geo-located crowd-contributed sources contain so-called platial knowledge?

Places are routinely represented as explicit locations (i.e. footprints) in place-name gazetteers that use approaches that are committee centric, and often limited to administrative views of official place-names. Place-name gazetteers originating from different producers typically have similar schemas and category types in which feature matching and merging is relatively straightforward for comparison. However, extracting place meanings from one crowd-contributed source, let alone multiple, remains a challenge. Nevertheless, it is still unknown whether different sources containing platial information are semantically consistent thus contributing to a unique identity of place?

Current approaches to formulate and maintain place information are not equipped to handle evolving and bottom-up changes. For instance, take Central Park in New York City, a typical place-name gazetteer such as GeoNames.org or points of interest (POIs) provides a name, class or type, and locational footprint, but it does not express the collective social meaning at the time as to the crowd's perceived meaning of the park. This human centric view is nevertheless a dynamic characteristic of a place, as expressed by the crowd that include changes in semantics, temporal variations, and spatial representation (footprint).

A crowd-centric approach is warranted due to the problem of focusing solely on individual perspectives as to the meaning of place, which becomes highly subjective as interpretations and experiences can vary significantly. Crowds are used to derive implicit consensus as to the generally associated or lowest common denominator of place meaning. Given the notion of collective place meaning coupled with the vast amount and availability of crowd-contributed sources, the question still remains of how well such content can be mined to discover place or so-called platial knowledge. This research investigates the question by focusing on the shared meaning of place by generalizing people's collective sense of place. It is argued that taking a crowd-centric approach of collective and implicit sense of place meanings will lead to the discovery of emerging platial themes. Moreover, given the semantic-spatial-temporal characteristics of human activities within urban spaces, one can observe the emergence of unique themes that characterize different locations.

The scale of place, whether at the building, neighbourhood, or city level can have a significant impact on the uniqueness and discernibility of place as social data sources are globally connected. Since place is inherently multi-scalar, is there a decaying effect or gradient of place meaning in crowd-contributed sources as some scholars have found in various studies using surveys and human subjects? This scale effect has not been exhaustively explored in social media in general and, Twitter and Wikipedia, in particular, and open questions still remain. This work explores such problem of place meaning to understand the changes between the neighbourhood and city scales and particularities of place emergence to discover that zooming out to the city scale reveals more of the medium. Undoubtedly, the aggregation of sources at varying scales of analysis will produce different results as the areal units change, which prompts separate investigations at different scales.

As our world is becoming increasingly urbanized and dynamic, gaining an understanding of the building blocks of these urban environments is bringing forth the need for new approaches for sensing place. This dissertation contributes towards such goal, as it provides a new lens to observe platial content as it emerges from the people themselves, and allows us to do so at levels of spatial and temporal granularity that far exceed our past capabilities.

In order to discover emerging platial themes, a framework is developed to harvest implicit expressions of place using crowd-centric platial content by leveraging probabilistic topic modeling, semantic association, and platial clustering. The platial clusters are represented as one-to-one relationship of a single cluster mapped to a high-level category. Moreover, the platial representation of place seeks to address the dynamics of place through multiple relationships or connections to alternate perceptions of meaning beyond top-down classification. This problem is evident in certain types of user-generated content that are uncurated (i.e. Twitter) and not designed to reach consensus, but nevertheless offer real-time streams of content for dynamic spatiotemporal reasoning. The challenge of consensus is overcome by using existing sources of curated user-generated content (i.e. Wikipedia), which are designed and structured for the crowd to achieve consensus.

A significant challenge when considering platial thematics is the extraction of semantic content with both spatial and temporal attribution. To highlight the reasons, consider a simple scenario of a neighborhood that has undergone a renewal process and several buildings were torn down and new ones constructed that offer different affordances. Now consider two platial regions over the new construction that partially overlap spatially, have a temporal relation of before and after, and have a non-overlapping semantic relation of textual content. Thus, understanding these types of situations is essential for appropriately representing and eventually applying reasoning techniques.

1.2 Motivation

This research is motivated by the need to develop new methodologies and investigative approaches to study our urban places from the human perspective. As the study of places from this human-centric view will undoubtedly offer new insights into how we as humans assign perceive are environments and assign meanings. This democratized view of letting people assign meaning to places would have a profound affect on how urban planners and architectures design built spaces. In addition to human geographers and social scientists remotely sensing societal meanings from afar. And as Goodchild (2015) suggests the notion of a platial world is rich with hierarchies, associations and support for human perceptions and their impacts.

Only recently with emergence of massive amounts of geotagged content (Croitoru et al., 2014) has continuous monitoring, observing and studying human perceptions remotely been possible. I now have access to real-time streams of worldwide crowd-generated content, for example in the form of geotagged tweets and blog entries or flickr and instagram imagery. While it has been shown that such content can capture breaking events in the context of citizen journalism (see e.g. Goode, 2009; Kwak et al., 2010; Mathioudakis and Koudas, 2010; Bruns et al., 2012) an overall question and motivation for this research is how well such crowd-contributed content can be mined for platial knowledge. Some recent work has alluded to the fact that such platial knowledge discovery is possible from crowd-contributed sources, but such work has only begun to investigate this potential (Adams and Janowicz, 2012; Adams and McKenzie, 2013). There exists a current lack of combined methods to utilize all available semantic and spatiotemporal information contained within platial data sources (Steiger et al., 2015).

The spectrum of applications that may benefit from such an understanding of collective place meaning is broad, ranging from business (e.g. supporting location-based services) to security (e.g. detecting hotspots of unrest) and even health (e.g. studying health patterns not only as geometrical constructs, but also in association to underlying sociocultural data and attitudes). By moving from a geometrical view of the world around us to a platial view, we better support the quantitative study of the world's character, in addition to its layout.

One can consider a Geographer or, Social Scientist trying to determine whether a particular location is suitable for a place-based study on human behaviors or possibly agent-based modeling of social perceptions of place. Our current informational layers do little to represent such places as they change over time to satisfy more sophisticated inquiries. Thus researchers and analysts are forced to assimilate multiple sources of information and make inferences in order to develop a story of place. In essence, platial knowledge, built from collective crowd expressions from within spaces, should provide the essential components to construct bottom-up stories of place to help inform more rigorous investigation.

1.3 Research Hypothesis

Given the above problem statement and motivation, the following hypothesis is formulated.

Mining user-generated data leads to the characterization and localization of places.

According to the hypothesis, this research addresses the following three components:

- The algorithmic development of techniques for harvesting platial themes from crowd-contributed content based on topic modeling, knowledge base semantic reasoning, and spatial statistics.
- An investigation into the thematic alignment of platial information through a quantitative assessment of neighborhoods and cities from multi-source data.
- An assessment of how well such mined platial themes exhibit spatial and temporal alignment.

1.4 Research Assumptions

The proposed approach is based on the following key assumptions.

- A significant volume of geo-located social media in the English language is available for a continuous month time period.
- The linguistic style and character length of Twitter will cause erroneous results in some cases, to which the results will never be perfect.

- All information used from Wikipedia has an assumed state of assumed consensus derived from the crowd.
- The notion of place and perceived meanings are imprecise making a rigorous accuracy assessment difficult.
- The spatial representation of place, especially from implicit content, will be an imperfect approach as place boundaries are vague and incomplete.

1.5 Intended Audience

This work is applicable to anyone researching, studying, or engineering processes that involve an understanding of place and its collective representation in crowd-generated data. Specifically, Geographers and Social Scientists will find particular resonance with this research, as the study of place and human interactions with it; have been the foci for decades. The hope is this work will contribute to recent calls for platial technology in GIScience that are human-centric and more readily suited to understanding human behaviors (Goodchild 2015).

It is also intended that professionals working in consumer marketing will find interest in this work for tailoring ad-campaigns to specific places. The association between types of consumers with various place meanings could offer new verticals for mobile advertising revenue. Lastly, the representation of place has received much attention in the field of artificial intelligence, which has produced large volumes of work on spatial reasoning were researchers train robots to interpret the meaning of their environment (e.g. Rios-Martinez et al., 2015). The ability to capture collective meanings of place provides a new means for acquiring training features and abilities for context awareness.

1.6 Organization of Dissertation

The organization of this dissertation is presented here and structured so that each chapter provides observations and findings to evaluate the hypothesis. Chapter 2 provides a review of background literature and the current state of the art as it applies to each subject area. A non-exhaustive review of place, considering the breadth of place research, beginning with early geographical thinking through qualitative and quantitative turns, technologies, and crowdsourcing; ending with present day thought and the renewed interest with the rise of platial content.

Supporting work for crowdsourcing and consensus through the wisdom of the crowd are discussed as this forms the basis for the crowd-centric approach. Previous and current approaches to discovering platial information from crowdsourced content are reviewed with particular focus on spatio-temporal clustering, machine learning, and semantic methods. Current literature from the areas of semantic and spatio-temporal analysis is discussed with highlights on the importance of these approaches in discovering and quantifying platial information.

Chapter 3 outlines the proposed framework for discovering and thematically labeling platial knowledge from crowd produced content. Each section of this chapter presents the theoretical foundations for the processes used to harvest, statistical validation, and spatial alignment. Specifically, the chapter starts with a brief description and supporting claims for the chosen approach and selected methodology. The collection process is presented using Latent Dirichlet Allocation (LDA) topic modeling for dimensionally reduction with web-based semantic similarity and classification using Wikipedia. In addition, the spatial clustering with statistical significance checks is presented as the approach to detect clusters of semantic consensus from the crowd. The chapter concludes with two approaches are introduced for investigating the dynamic and evolving nature of platial knowledge through trend analysis and co-occurrence.

Chapter 4 presents the application of the proposed framework for harvesting collective and crowd-centric platial content from Twitter and Wikipedia in the following study areas: New York City, Los Angeles, London, and Singapore. A subsequent study is performed for three different neighborhood areas in New York City in order to contrast the differences in scale. Results are presented based on semantic and spatial alignment of platial clusters to known points of interest that have common meanings (i.e. school). A cross-source analysis is conducted between Twitter and Wikipedia to investigate the emergence of unique platial knowledge.

Chapter 5 extends on the findings from Chapter 4 by investigating the dynamic aspects of platial knowledge over time. The overall approach of topic modeling, semantic labeling, and finding significant hotspots is iterative applied at smaller temporal sizes by binning the data into one week intervals to investigate trends and polygon co-occurrences. Additional techniques are applied to the resulting platial hotspots to programmatically identify hotspots and derive area features.

Chapter 6 concludes the dissertation by providing a summary of the tasks to include accomplishments and major findings. The most significant contribution of this work is discussed with proposed extensions for future work, research directions, and speculated outlook on platial research and technology.

CHAPTER 2: BACKGROUND & PREVIOUS WORK

2.1 Place

The topic of place has been extensively studied throughout multiple disciplines, from geography (e.g. Relph, 1976; Tuan, 1977) and planning (e.g. Graham and Healey, 1999; Healey, 2004), to health informatics (e.g. Kearns and Joseph, 1993), sociology (e.g. Gieryn, 2000; Law and Urry, 2004), and psychology (e.g. Devine-Wright and Lyons, 1997; Gustafson, 2000; Devine-Wright and Clayton, 2010).[1] There exist volumes of literature concerning place as a social construct that interlaces geography and philosophy, but the definition most often used is Tuan's (1977) "…places are spatial locations that have been given meaning by human experience." This definition implies that places can only exist where humans have manifested or established a connection to a physical location or space through events. Relph (1976) described place, as an object comprised of three dimensions:

1. Observable activities that occur in relation to the place.
2. The meanings that are created by a person in that location, and;
3. The physical features that comprise the location's concrete or tangible attributes.

[1] A full discussion of place is beyond the scope of this dissertation. The purpose here in this chapter is to discuss place and how it relates to this dissertation

The combination of such characteristics of place, and in particular knowledge of people's expressions and experiences, whether expressed explicitly or implicitly, gives rise to a common meaning or means of identifying with a location used by the crowd. So as Relph (1976) described the basic building blocks of place from the individual perspective, he further articulated that the aggregation from multiple people forms consensus and mass images of place as seen here:

> *"Although one particular place may have quite different identities for different groups, there is nevertheless some common ground of agreement about the identity of that place. This is the consensus identity of a place, in effect its lowest common denominator." (Relph 1976)*

Accordingly, place consensus, or what Relph (1976) refers to as consensus identity of place, is formed through the reoccurrence of shared experiences (by groups) through repeated activities at a certain location (Brandenburg and Carroll 1995; Williams and Stewart 1998). These experiences assign a common meaning to locations, transforming them from geometric concepts (the three dimensional space itself) to experiential constructs (places that convey public perceptions). Ultimately, the concept of place is constantly evolving, as over time places are reconstituted with new meaning, reflecting for example urban dynamics (Batty et al., 1999), evolving sociocultural perceptions (Salesses et al., 2013), or significant events (Crooks et al, 2015).

Recently, the individual human centric expressions of place have been defined as platial information that is either expressed explicitly as place names or implicitly as semantic associations (Goodchild 2015). The platial perspective is one of human discourse defined by textual place names, linguistic descriptions, and the semantic relationships between places (Janowicz 2009; Goodchild and Li 2012; Gao et al. 2013). Quesnot and Roche (2015) distinguished between *explicit platial data* and *implicit platial data* in the era of geospatial Big Data, whereby, explicit platial data contains coordinates and place name mentions that correspond to social media check-ins. The researchers refer to implicit platial data as similarly having coordinates and check-ins at specific locations, but the mentions refer to places or activities not semantically associated with their current place.

Nevertheless, over time, place whether explicitly or implicitly characterized is reconstituted with new meaning, in the form of platial information, that in turn continuously reforms our sense of place making it a highly dynamic process. The notion of sense of place is formed through the reoccurrence of experiences through repeated activities at a location developed by a person or group (Brandenburg and Carroll 1995; Williams and Stewart 1998). In this work, 'sense of place' is used as a means to frame the relationship between crowds of people, place, and events/activities.

Whilst numerous definitions exist, sense of place is perhaps most simply considered as an overarching concept that subsumes other concepts describing relationships between human beings and spatial settings (Shamai 1991, Jorgensen and Stedman 2001). Silver and Grek-Martin (2015) argued that relationships between people and their environment have been conceptualized in many ways, drawing on distinct but related concepts such as "sense of place" (Tuan, 1974), "placelessness" (Relph, 1976), “insideness” (Relph, 1976), and “place attachment” (Altman and Low, 1992). In particular, the conceptualization of sense of place is investigated here despite its multitheorectical, complex, and contested positions. The literature is often chaotic and ‘sense of place’ may simultaneously appear ambiguous and distinct.

The study of place from a qualitative perspective has largely focused on both theoretical constructs and human subjects to understand behaviors towards assigning meanings to places. A sample of early work combined human subjectivity and cognitive maps as the primary tools for understanding place (Lynch 1960), but in a static representation. Pred (1984), proposed a theoretical framework that combined place and time-geography to account for the continuous reproduction of what the author called social and cultural forms of place. Ultimately, place remains complicated by the multitude of meanings individuals’ hold (Kridel 2010), which is evident in the breadth of field studies focusing on place (Lippard and Dawson 1997; Jordan et al. 1998; Zhao et al. 2011; Jones and Evans 2012).

Notably, Golledge (1992) elicited place boundaries for downtown Santa Barbra, CA from participants and compared the variations in boundaries that were equated to different meanings of place. Similar approaches have elicited knowledge from human subjects to understand the social meanings of places through the drawing of home range boundaries (Purcell 1997; Gustafson 2001). Hidalgo and Hernandez (2001) measured and computed place attachment within three spatial ranges (house, neighborhood, and city) and two dimensions (physical and social) by interviewing 177 people from different areas of Santa Cruz de Tenerife (Spain).

Aside from direct human subjects, and more recently, our highly networked urban environments have impelled the rate at which the meanings of places change through the connectedness of people, places, and devices in real-time (Crang and Graham 2007). Mobile computing has transformed our activities at spatial locations into connected activities thus making Tobler's (1970) first law of geography literally true – everything is connected to everything else. This has prompted significant efforts toward georeferencing place descriptions and processing spatial queries, such as using ontologies of place (Jones et al. 2001), qualitative spatial reasoning frameworks (Yao and Thill, 2006), fuzzy objects (Montello et al. 2003), probability models in combination with uncertainty (Guo et al. 2008; Liu et al. 2009), kernel-density estimation (Jones et al. 2008), description logics (Bernad et al. 2013), as well as knowledge discovery from data techniques for platial search (Adams and McKenzie 2012).

Recently, a review by Vasardani et al. (2013) suggested that a synthesis approach would provide improvements in locating place descriptions, and that new opportunities exist in identifying places from public media and volunteered sources by using Web-harvesting techniques. Similar to ontologies, the notion of geo-folksonomies, which are folksonomies of tagged places that have been used to encode relationships between humans and the places they label to (ElGindy and Abdelmoty 2014).

Significant attention has been devoted to methods for addressing the spatial vagueness of place using anchor theory (Galton and Hood, 2005), fuzzy models with vector data representation (McIntosh and Tuan, 2005), and supervaluation theory (Kulik, 2001). Other researchers have combined statistical machine learning with textual narratives of place to extract meaning using named entity recognition (NER). Freksa and Barkowsky (1996) categorized places into entities and formed relationships with nearby geographical objects to form a network structure to model place dependencies and changes in relatedness.

The scale and granularity of place representation has an undoubting affect on the perceived human activities, which has a dependency on the types of places and their functions. For instance, Richter (2009), although an indoor study, distinguished a functional, a social, and a physical hierarchy of places using a lattice of spatial granularity. Lewicka (2010) explored the relationship between place scale (apartment, neighborhood, city) and strength of place attachment with results showing the strongest tie at the neighborhood scale. Although this dissertation does not explore the emotional bond of people and place, the work of Lewicka (2010) does offer place scale insight at the neighborhood level that is also observed in this work.

The scale of sense of place is better viewed within social media as a spectrum of homogeneity and heterogeneity respectively from streets, neighborhoods, and city. Winter and Freksa (2015) determined that humans recognize and memorize places at different levels of granularity, and places at one level or scale of granularity can belong to places at coarser levels of granularity. This abstract concept runs in parallel with studying places at different scales. Still other scholars have noted, in the context of place, that relative position and structure in both large-scale spatial reasoning and small scale, will result in simpler or generalized relations that take the place of more precise representations (Cohn 1997).

2.2 Crowdsourcing Place

As we now have access to continuous streams of worldwide crowd-generated content, for example in the form of geotagged tweets and blog entries or Flickr and Instagram imagery. The proliferation of Web 2.0 technologies to mobile devices has allowed users to freely explore social spaces and places. The innovations of Web 2.0 technologies allow for instantaneous uploads of crowd-generated content while physically present at places and communicating cyber-socially through message posts, check-ins, social linking, photos, movies, and audio files through web all the time (Beck 2008). The massive volumes of crowd-generated content have created new challenges in the form of geospatial big data, but subsequently offer new opportunities for harvesting and synthesizing such information to extract knowledge (Stefanidis et al., 2011).

The use of Web 2.0 technologies has enabled the crowd to contribute vast quantities of spatial and place-based information explicitly in the form of volunteered geographic information (VGI) with sites like OpenStreetMap and Wikimapia, and implicitly with sites such as Foursquare and Twitter (Croitoru et al., 2014). Although important progress has been made in recent years in harvesting spatial and temporal data from social media, (Sui and Goodchild 2011) the quality and credibility of data for scientific research and decision-making still need further investigation. Croitoru et al., (2015) and Goodchild and Sui (2011) have articulated that new processes are needed for the fusion of GIS (i.e. mapping), social media, and social networks.

The importance of developing new approaches to mining and reasoning with platial knowledge is evident (Goodchild and Li, 2012; Steiger et al., 2015; Roche 2015).

Recent research concerning spatial semantics, location modeling, and cognitive models can be interpreted as initial work towards place research in Geographic Information Science (GIScience) (Winter et al., 2009). However, the types of data sources (i.e. geotagged media) available for which to mine platial knowledge present their own inherent challenges with respect to dimensionality, spatial and temporal coverage, linguistic style, etc. Thus the use of combined techniques from across multiple disciplines (i.e. geography, computer science, social science) to study platial knowledge has proven most effective to date. With the vast majority of social topics having a spatial footprint as described by Cano et al. (2011) provides us with unique opportunities to explore social activities at a single place.

Crowdsourcing the social meaning of places has been previously investigated using a variety of sources from location-based social networks, geotagged photos, and textual place descriptions to derive collective experiences. The closest with respect to this work presented here is the work of Steiger et al., (2015) that used a similar mixed approach to account for the spatial, temporal, and textual characteristics of crowd-generated content. In particular, the researchers extracted spatiotemporal and semantic features from Twitter using a combination of spatial statistics, semantic modeling, and a neural network to detect human activity patterns from perceived platial content. Cranshaw et al. (2013) demonstrated the construction of social neighborhoods through the clustering of collective human activities using Foursquare check-ins to form what the authors called Livehoods. The discovered clusters represented the perceived boundaries of neighborhoods that in some cases aligned with official boundaries and yet in others revealed entirely different areas influenced by demographics.

Liu et al. (2015) presented an approach to mining human representations of place and social activities from taxi trajectories and social media check-ins using a remote sensing framework to create social images. The images provide arguably land-use representations of platial knowledge and also highlight the fusion of multiple data sources to derive such information. Adams and McKenzie (2015) evaluated platial content in travel blogs using topic modeling and spatiotemporal features to determine the thematic similarity between different places. This work highlights yet another dimension and data source of platial content, provided in this case, through tourist narratives at both an individual and collective event levels.

Lee et al. (2014) showed that collective urban experiences through so-called crowd lifelogs could be extracted from sites such as Twitter, Foursquare, and Facebook Places. This work investigated the particular movements of crowds through urban places by measuring cognitive distance of place related terms to the crowd's physical clusters. However, this study did not include any semantic analysis on the implicit meaning of activities. Purves and Derungs (2015) combined crowd-contributed Flickr images with key terms, historical narratives of place, and kernel density analysis to investigate the dynamics of place over time by studying natural landmarks. Tammet et al. (2013) developed a process for mining popular sights or places using a platform called SightMap that linked POIs, Wikipedia places, and Panoramio photo densities. The researcher used text similarity by applying Levenshtein edit distance to match key words and categories co-located near places.

Keßler et al. (2009) described an agenda for a next generation gazetteer from user-contributed and vernacular geographic information with particular emphasis on harvesting implicit places. Goldber et al. (2009) used an agent-based semi-supervised learning approach with named entity recognition (NER) to extract place-name from webpages that were then geocoded using online white pages. Ivre and Machado (2011) developed an ontological approach to record semantic connections of places from webpages that they applied to news feeds. Recently, Gao et al. (2014) implemented a distributed Hadoop based architecture to harvest place from geotagged datasets in Flickr. Although Gao et al. (2014) used a bottom-up approach to construct place-based content from crowdsourced data; they did not incorporate any semantic or temporal attributes.

According to Elwood et al. (2013), volunteered geographic information (VGI) from the popular site OpenStreetMap (OSM) assumes the characteristics of human discourse (platial information) through user tags rather than scientific measurements and thus has strong links to the world of place. The concept of VGI has received much attention since its first introduction by Goodchild (2007). VGI has come to be defined as user-generated geographic information that is openly contributed conveying both space and place. The crowd-contributed concept seeks to harness the power of volunteers or the crowds to produce, assemble, and contribute geographic data in semi-structured formats that are curated by the crowd and remain relatively static in representation. VGI sources, such as OSM, have proved invaluable for humanitarian and disaster response situations like the

Haiti earthquake in 2010, in which large crowds of volunteer mappers from around the world contributed content that resulted in the most detailed GIS layers ever created within 24hrs after the initial event.

On the uncurated side of the spectrum, social media has grown to one of the central modes of implicit place communication while staying in touch with friends and family as well as sharing, coordinating, and commenting on socially relevant events and gatherings. The social context of multiple actors represented in a network allows for a more dynamic and event-based view of places, which ultimately convey patterns of life (Frank et al., 2013). Coupling this with the granularity of spatio-temporal information and social linking, social media data sources provide near real-time characteristics of place meaning. This is evident in studies showing the predictive mapping of human mobility between urban places (Noulas et al., 2012). More recent studies have investigated the characteristics of urban environments and perceptions of the places within it by using social media to construct sound maps (Aiello et al., 2016) and smell maps (Quercia et al., 2015).

2.3 Crowds, Consensus, and Crowd Generated Data

Crowdsourcing has emerged out of the need to combine, in the problem solving process, existing knowledge from several scientific and professional fields, with the “wisdom of the crowd” (Surowiecki, 2004). According to Surowiecki (2006), “under the right circumstances, groups are remarkably intelligent, and are often smarter than the smartest people among them”. The principal axiom of this approach is that no one knows everything; everyone has a specific expertise, and therefore, solutions can be reached by combining everybody’s knowledge and experience (Papadopoulou and Giaoutzi 2014). As people communicate more about a place, social consensus will create increased similarity between and within people's judgments of it (Davies 2009).

An early study on place consensus used local vernacular names for mapping, at least in countries like Great Britain (Harley 1971), and relied on attempts to identify and reflect local consensus: only so many names could be shown on a map or returned by a gazetteer. More recently, Haklay at el., (2010) concluded that given enough VGI contributors reviewing geographic data, Linus’ Law holds true and the accuracy and quality of the data increases. Linus’ Law originates from the software engineering field and more specifically states, “given enough people reviewing software code, issues will be quickly fixed or become shallow”. By juxtaposing Linus’s Law with the work of Davies (2009), it logically follows that given a large enough crowd of people conveying similar social meanings of place, the predominate consensus of place meaning should emerge. One such example of this concept is the work of Lamprianidis and Pfoser (2012), in which the authors constructed point-clouds from multiple Flickr users to form collective regions of colloquial meaning.

A central focus of this research is to use crowdsourced content. In particular, Wikipedia and Twitter are used extensively, however; there design principles and intended uses vary according to levels of consensus. Wikipedia first launched in 2001, with the intent of expediting the creation of encyclopedic articles for another site called

Nupedia that relied solely on expert contributors. However, the experts working on Nupedia took months to create only several articles. Wikipedia's founders, Jimmy Wales and Larry Sanger, had envisioned that contributors would quickly generate articles that experts could review and merge into Nupedia. Within the first year, contributors created more than 20,000 articles in 18 languages, and Wikipedia quickly became the flagship project.

Initially, a handful of administrators and editors were put in place to apply a loose level of control over the site, but ultimately the community of contributors provided governance through a peer production model. The Wikimedia Foundation was formed in 2003 by Wales to raise money for operating costs. By late 2005, the English version of Wikipedia along had more than 750,000 entries. Today, Wikipedia is the world's largest collectively edited source of encyclopedic knowledge (Krötzsch et al. 2007). It is the sixth most widely used website in the world with more than 4.7M English articles and over 35.3M total pages in 262 languages.

Wikipedia is the world's largest collectively edited source of encyclopedic knowledge and the largest reference website on the Internet (Krötzsch et al. 2007). The content is collaboratively written and updated by volunteers (Wikipedia 2007); it is extremely useful as a resource due to its size, variation, accuracy and quantity of hyper-links and meta-data (Kinzler 2005, Weaver *et al.* 2006). One of the strengths of Wikipedia is its continued rapid growth and ability to stay in sync with world events (Milne and Witten 2013). Kämpf et al. (2012) contend that a significant volume of Wikipedia contributions are motivated by news and mass media. The sheer number of articles, linked structure, and openness of Wikipedia have made it an ideal source for text classification (Wang et al. 2009), semantic relatedness (Iosif and Potamianos 2010), and knowledge engineering (Baumeister et al. 2011) research.

Chris Lüer (2006) identified disambiguation techniques used within Wikipedia as the mapping from a word to an article: disambiguation of polynyms is accomplished in Wikpedia by a combination of requiring every article to have a unique guessable name and explicit disambiguation pages. Disambiguation of synonyms is achieved through a network of redirect pages (Lüer 2006). The responsibility is then on a page author (and editors) to correctly link to the intended pages they reference.

Wikipedia is being used more and more in geographic information retrieval: for corpus generation, ontology generation (Buscaldi *et al.* 2006), gazetteer generation (Silva *et al.* 2004, Buscaldi *et al.* 2006), query expansion (Hauff *et al.* 2006) and ground truth generation (Lüer 2006). The category and template meta-data greatly simplifies article classification. Such projects as 'WikiProject: Geographic Coordinates' and 'PlaceOpedia' are currently underway allowing users to geographically tag Wikipedia (Wikipedia 2007, Steinberg 2007). Wikipedia's category structure has been used to create massive ontology and linked knowledge base, integrated with WordNet, and GeoNames (Hoffart et al., 2013)

As Wikipedia represents a paradigm of digital crowd curation, the popular social media site Twitter is considered just the opposite with its freeform and uncurated style. The popularized microblogging and online social networking service has become ubiquitous with its 140 characters of free text and specialized symbols. Twitter has been compared to news media and other real-time platforms because of its instantaneous streams of user reactions, opinions, and experiences related to events and places. Significant attention has been given to Twitter for events such as earthquakes (Earle *et al.* 2012; Crooks et al. 2013), political elections (Tumasjan et al. 2010), and terrorist activities (Oh et al. 2011). Other Twitter research has focused on sentiment analysis related to geographic distribution (Mitchell et al. 2013), event types (Thelwall et al. 2011), and topical categories (Tsytsarau and Palpanas 2012).

The common theme throughout much of Twitter research is the sheer volume of tweets needed to derive some meaningful results. Unlike Wikipedia, Twitter was not designed to form consensus or to harness the crowd to perform tasks. Wikipedia is an example of collective intelligence harnessed by a specific design pattern that allows the crowd to contribute, but relies on administrators to make final judgments on whether to delete a challenged article. In Wikipedia, we already have a consensus, and its content is an expression of said consensus guided by editing rules much like the content of authoritative encyclopedias (Medelyan et al. 2009). Articles in Wikipedia reflect an experiential definition of places, as conveyed by the crowd, whereby; we have one contribution made by many. In Twitter we have many contributions made by individuals without a controlled design pattern. Despite its shortcomings, Twitter provides benefits over Wikipedia with respect to human activities and the “what’s happening now” mantra.

Social media sources have inherent biases due to several factors. First, the design pattern and intended use of social media rarely account for notion of place or include it as core design principle. Most social media outlets, for instance Twitter, is meant to send short messages or communications about lifestyle, entertainment, or the "what's happening now" content in people's personal lives. Notions of place within social media are more of by-product of humans discussing their experiences at a particular location, which may or may not, relate to their surrounding environment. With this in mind, content analysis studies have been performed to assess the degree of bias in social media, particularly Twitter, and found that majority of the content is personal narratives of lifestyle and entertainment topics (Zhao et al. 2011).

To overcome the problem of biases in social media, one could model the signal-to-noise within the data to characterize what content has a relationship to a specific place. Additionally, although geographic coordinates make up a small percentage of social media content, ~2% (Burton et al. 2012), spatial clustering has shown the ability to reveal locations that describe similar human activities (i.e. political protests). In the specific case of Twitter, which has been compared to news media and other real-time platforms, instantaneous streams of user reactions, opinions, and experiences ultimately have some dependency or relatedness to place by virtue of a particular events or personal experience. Specific events with relations to place have been observed during earthquakes (Earle *et al.* 2012; Crooks et al. 2013), political elections (Tumasjan et al. 2010), and terrorist activities (Oh et al. 2011).

Such events ultimately have an affect on the volume and content of social media and thus skewing the information towards a particular topic. In this case, the event should be considered sporadic, although specific place information could emerge in greater fidelity, it would not represent normal regularities of a particular area and thus create a geographical unevenness (Graham et al. 2014). Additionally, social media is highly dependent on technological infrastructure and digital devices resulting in sparse representation of locations beyond the digital divide. These situations present problems for extracting place information, for which, other means and sources would have to be investigated. This is apparent from the work of Mislove et al. (2011) with their comparison of U.S. Census data and Twitter. U.S. population data correlates well with Twitter along the eastern and western seaboards, but diminishes in the Midwestern states due to low population densities.

2.4 Probabilistic Topic Modeling

Probabilistic topic models are a suite of algorithms whose aim is to discover the hidden thematic structure or topics in large volumes of documents (Blei 2011). This is a classical problem in text mining and information retrieval using term frequency and co-occurrence to form a bag-of-words structure. Latent Dirchlet Allocation (LDA) is probably the best-known approach in the field of topic modeling, which can be applied using either supervised or unsupervised models. It discovers latent structure in a collection of documents by representing each document as a mixture of latent topics, where a topic is itself represented as a distribution of words that tend to co-occur (Blei *et al*. 2003). In other words, the topics are essentially groups of correlated words and the correlation is revealed by word co-occurrence patterns in documents.

LDA, and its many variants, have shown significant utility when used with short messages like tweets, tips, or reviews (e.g. Farrahi et al. 2011, Pozdnoukhov et al. 2011, Ferrari et al. 2011, Hong et al. 2012). Topic models have shown advantage in activity modeling tasks due to their ability to effectively characterize discrete data represented by bags (i.e., histograms of discrete items) (Farrahi and Gatica-Perez 2011). Farrahi and Gatica-Perez used topic modeling to determine daily mobile user routines by designing a *bag of location sequences* based on cell tower locations and home (H), work (W), or other (or out) (O) location labels. Other researchers have attempted to predict Twitter user locations based on topics and known locations of those topics since most tweets have geographic location info (Hong et al., 2012).

Lim et al. (2013) developed a Twitter-Network topic model to jointly model the text and social network information. He at el. (2014) proposed a dynamic joint sentiment-topic model to detect and track current sentiment and key topic term. Li et al. (2014) developed a generative model approach they call author-topic-community (ATC) model for representing a corpus of linked documents. Interestingly, this approach works with social network and perceived hyperlink data in Wikipedia, although the authors did not try this, but nevertheless, this supervised approach jointly models author profile interests and linked community structure.

Wu et al. (2016) proposed the combination of topic modeling and pointwise mutual information (PMI) (Turney 2001) to profile the social network of user contributions for personalized recommendations. This work interestingly uses the resulting topics and PMI to find associations, but like majority of topic modeling approaches it only uses single terms as topics that disregard collocation of others to produce phrases. For instance, the phrase "yearly egg hunt at the white house" would return separate terms such as "egg", “hunt”, “white”, and “house” as opposed to phrases such as “egg hunt” and “white house” that have more semantic meaning.

Several variations of LDA have been proposed to address the unigram topic problems, which are referred to as n-gram based topic models. One of the first to explore word order or collocation using the LDA approach was Wallach (2006). This work extended Blei’s (2003) process by including n-gram statistics to produce bi-gram topic models. However, this approach only generates bi-grams and does not consider whether or not two terms should form a bi-gram or remain unigrams. Thus risking the grouping of terms together to form bi-grams that might not make sense. Griffiths et al. (2007) first proposed a method that added additional variables to not only include collocation to produce phrases, but also whether to produce a phrase or single word based on the dependencies or co-occurrence of the previous word.

Wang et al. (2007) further extended the work of Wallach (2006) and Griffiths et al. (2007) by proposing the topical n-gram model (TNG). Hence, claiming their approach is more generalizable by combining both works, but with the inclusion of nearby context in the bi-gram or unigram determination. For example, their method could differentiate the phrase “white house” within the context of a real estate description of any arbitrary house versus a reference to the president’s home within a single corpus. Thus producing separate unigrams and bigrams within the same context compared to the other approaches that would create a single bi-gram or unigram. To this end, the TNG approach was utilized in this research and is later discussed in Section 3.2.

2.5 Web-based Semantic Similarity and Crowd Generated Data

Semantic similarity measures are instrumental in the study of information retrieval (Srihari et al., 2000), ontologies (Sanchez et al., 2012), and the Semantic Web (Berners-Lee et al., 2001;Rettinger et al., 2012) to assess the association between two or more words or thematic categories. Ultimately, the semantic similarity between two terms or categories can change over time and across domains or geographic areas. The reasoning for the use of web-based or retrieval engine similarity measure is it best captures the current interests and dynamic volatility of the collective crowd (Li 2014; Franzoni and Milani 2012). Given that place meanings are highly fluid and event-based; a static knowledge base or lexical database would not have the most up-to-date meanings.

Measures of web-based semantic similarity abound for determining numerically how similar or related two words are. Resnik (1995) first proposed an information theory-based method for semantic similarity using a taxonomy hierarchy and frequency of occurrences of two concepts. More recently, page counts from search results have been used in conjunction with knowledge bases and search engines. Bollegala et al. (2007) presented several popular co-occurrence measures such as Jaccard (Jaccard 1912), Overlap (Simpson 1949), Dice (1945), and PMI (Turney 2001) in modified web-based form, denoted by the appending of “Web” to each name, to account for page counts presented in Table 1. Also included in Table 1 is the popular Normalized Google Distance (NGD) metric developed by Cilibrasi and Vitanyi (2004).

To briefly describe the metrics in Table 1, the notation $f(w_1)$ and $f(w_2)$ denotes the page or article counts, depending on the knowledge base or search engine, w_1 and w_2

denote the query terms or words. Bollegala et al., (2007) included c as a threshold value and N as the total number of articles or pages contained a search engine or knowledge base. For NGD, being a distance metric, the closer the resulting value is to zero, the more similar the terms. So calculating the score for the exact same terms would result in zero distance, compared to completely opposite terms resulting in one.

Table 1 Selected semantic similarity equations for use with web-based page count. Content adapted from Bollegala et al., (2007) and Cilibrasi and Vitanyi (2004).

WebJaccard(w_1,w_2)	$= \begin{cases} 0 & if\ f(w_1 \cap w_2) \leq c \\ \frac{f(w_1 \cap w_2)}{f(w_1)+f(w_2)-f(w_1 \cap w_2)} & otherwise \end{cases}$
WebOverlap(w_1,w_2)	$= \begin{cases} 0 & if\ f(w_1 \cap w_2) \leq c \\ \frac{f(w_1 \cap w_2)}{\min(f(w_1),f(w_2))} & otherwise \end{cases}$
WebDice(w_1,w_2)	$= \begin{cases} 0 & if\ f(w_1 \cap w_2) \leq c \\ \frac{2f(w_1 \cap w_2)}{f(w_1)+f(w_2)} & otherwise \end{cases}$
WebPMI(w_1,w_2)	$= \begin{cases} 0 & if\ f(w_1 \cap w_2) \leq c \\ \log_2\left(\frac{\frac{f(w_1 \cap w_2)}{N}}{\frac{f(w_1)}{N}\frac{f(w_2)}{N}}\right) & otherwise \end{cases}$
NGD(w_1,w_2)	$= \begin{cases} 0 & if\ f(w_1)=f(w_2) \\ 1 & if\ f(w_1) \neq f(w_2) \\ \frac{\max(\log_2 f(w_1),\log_2 f(w_2))-\log_2(f(w_1),f(w_2))}{\log_2 N-\min\ (\log_2 f(w_1),\log_2 f(w_2))} & otherwise \end{cases}$

In the context of Wikipedia, there has been numerous semantic relatedness and entity disambiguation approaches proposed throughout the literature. WikiRelate (Strube and Ponzetto 2006), one of the first proposed, uses Wikipedia's category structure to compute relatedness. The Wikipedia Link Vector Model (WLVM), proposed by Witten and Milne (2008), calculates semantic relatedness based on the number of hyperlinks to a particular article divided by the total number of articles. Newman et al. (2010) determined similarity by treating Wikipedia as a meta-document to score n-gram pairs using term co-occurrence. Niraula et al. (2013) used semantic similarity measures and LDA to assess topic quality and found that PMI (Turney 2001) method was best when compared to human judgments of topic coherence. They counted the frequency of the co-occurring words in a window of 10-word in Wikipedia corpus and 5 in case of Google 5-grams.

Genc et al. (2011) associated Twitter messages with their most similar Wikipedia pages and then used the network distances between the article pages as a proxy. Michelson and Macskassy (2010) showed that Wikipedia could be used to profile the topics of Twitter users into coherent categories. To briefly illustrate, Figure 1 shows their overall approach to discover the main topics of interest particular users. They start by extracting entities from Twitter using all capitalized and non-stop words, which they then send to Wikipedia to disambiguate the meanings and extract the categories. The last step involves a frequency rank metric calculated for each category per user. Although interesting, given millions of tweets, this approach would require a significant amount of computational time since the approach extracts all non-stop words. Plus, this approach suffers from a loss of meaning by only looking at single terms and not phrases such as bi-grams. Nevertheless, it provides a bases for the approach proposed in this dissertation.

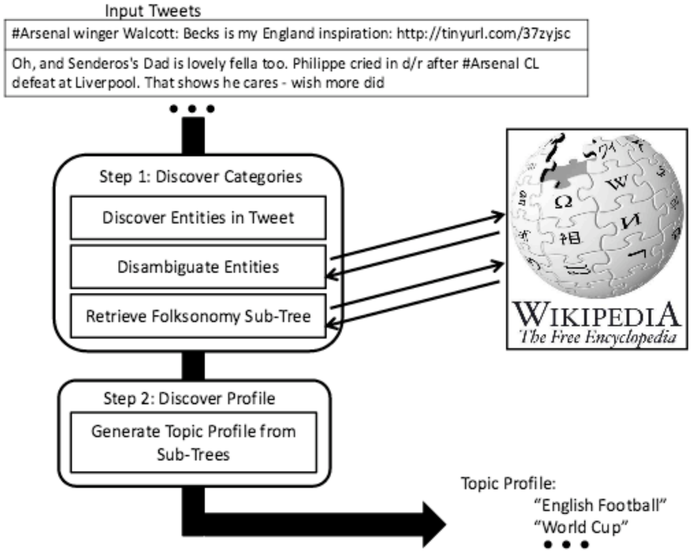

Input Tweets
#Arsenal winger Walcott: Becks is my England inspiration: http://tinyurl.com/37zyjsc
Oh, and Senderos's Dad is lovely fella too. Philippe cried in d/r after #Arsenal CL defeat at Liverpool. That shows he cares - wish more did
Step 1: Discover Categories
Discover Entities in Tweet
Disambiguate Entities
Retrieve Folksonomy Sub-Tree
WIKIPEDIA
The Free Encyclopedia
Step 2: Discover Profile
Generate Topic Profile from Sub-Trees
Topic Profile:
"English Football"
"World Cup"

Figure 1 Sample architecture (source Michelson and Macskassy 2010). Illustrative example that combines Twitter messages with Wikipedia to semantically link entities to higher-level categories.

2.6 Semantics, Topics, and Spatial Clusters

Semantic similarity measures have been widely studied and applied in GIScience (Rodriguez and Egenhofer 2004; Schwering and Raubal 2005; Li and Fonseca 2006; Ahlqvist and Shortridge 2006). Most of these measures are hybrid in a sense that they combine different approaches to similarity, such as features, regions in a multi-dimensional space, or network distances (Mülligann et al., 2011). However, most of the aforementioned studies assume structured data with predefined schemas. And unfortunately, when dealing with crowd-contributed content, structured attributes and schemas are not the norm. In the case of Twitter, the semantic content is presented in freeform text that has to be mined for entities and meaning. As previously discussed, LDA and its variants, have been used considerably to extract semantic content and perform dimensionality reduction to find key components or the most frequent terms.

Given the extraction of semantic terms or topics, and embedded geo-location information within the data source, point pattern analysis statistics can then be applied to determine semantic-spatial clusters. Adams and McKenzie (2015) evaluated platial content in travel blogs using LDA and spatiotemporal clustering to determine the thematic similarity between different places. This work shows another dimension and data source of platial content, provided in this case, through tourist narratives at both an individual and collective event levels.

Steiger et al. (2015) formed semantic and spatiotemporal clusters using Twitter data by applying LDA and two spatial statistics to discover frequencies of human activities, within places, that were then compared with census data for workplaces and residential areas. Figure 2 depicts the overall process the authors used to derive human activity clusters. However, there are significant differences from this approach to the research presented here in Chapter 3. The most notable being the use of Wikipedia as a knowledge base, the ability to select any high-level category or event (not just basic activity terms, “work”), and iterative distance analysis using Global Moran’s I to adjust for different geographies.

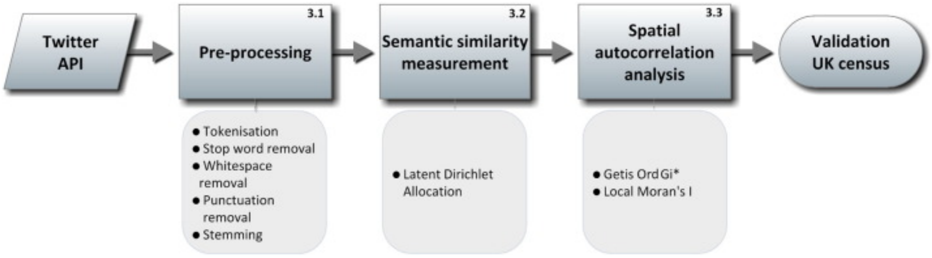


Figure 2 Sample process (source Steiger et al., 2015). Illustrative example that combines Twitter, LDA, and spatial statistics.

Fan and Stewart (2015) combined LDA, Wikipedia, spatial interpolation, and Getis-Ord Gi* (Getis and Ord, 1992) to detect the spatial patterns of wildfires. The researchers used the supervised version of LDA that was trained on the entire Wikipedia corpus using both geo-located articles about wildfires and geo-located articles for places such as towns, landmarks, etc. to characterize the semantics of wildfire related articles. Interestingly, they calculated a spatial probability surface by applying Kriging, or spatial interpolation, that combines wildfire related terms such as dry, heat, etc. with geo-located places with similar terms. They finally used Getis-Ord Gi* statistic to assess spatial alignment of individual terms with places, and noted some ambiguity, particularity with the term “mobile” and Mobile County in Alabama. To build the classifier, they used a high performance-computing environment given the sheer volume of data, which does not scale well for other applications.

Fuchs et al. (2013) investigated personal behavioural patterns in geo-located Twitter data by extracting knowledge about significant personal places and interests. The authors manually created topics and related terms, instead of using a machine learning approach, and then categorized the terms into higher-level categories. After assigning the tweets to a category, they then applied density based clustering using spatial proximity with an arbitrary distance threshold to data in hourly temporal slices. Once again, and in another paper by Steiger et al., (2015), they used LDA, Getis-Ord Gi*, and neural networks known as self-organizing maps (SOM), to extract spatiotemporal and semantic features from Twitter to detect human activity patterns from perceived platial content. Similar to their previous work on human activity patterns, the researchers restricted the categories to terms regarding daily routines.

To address the scale mixing of human activities found in crowd-sourced data, Westerholt et al., (2015) presented a modified scale-sensitive version of the Getis-Ord Gi* statistic that they combined with Latent Semantic Indexing (LSI). They used the resulting LSI values as inputs to their local spatial clusters analysis. To address the difference in scales, the researchers modified the spatial weights matrix by including scale-adjusted neighborhoods using a distance interval [d_{min}, d_{max}], as opposed to a fixed neighborhood distance. This approach has the advantage of detecting spatial clusters of different sizes within a single data source like Twitter. Although they provide a means to deal with varying scales, it still requires the distance values or interval to be produced much like the original Getis-Ord Gi* statistic.

In this chapter, a background literature review and current state of the art was presented for each subject area. A non-exhaustive review of place, considering the breadth of place research, beginning with early geographical thinking through qualitative and quantitative turns, technologies, and crowdsourcing; ending with present day thought and the renewed interest with the rise of platial content. Supporting work for crowdsourcing and consensus through the wisdom of the crowd are discussed as this forms the basis for the crowd-centric approach. Previous and current approaches to discovering platial information from crowdsourced content are reviewed with particular focus on spatio-temporal clustering, machine learning, and semantic methods. Current literature from the areas of semantic and spatio-temporal analysis is discussed with highlights on the importance of these approaches in discovering and quantifying platial information. The next chapter outlines the proposed framework for discovering and thematically labeling platial knowledge from crowd produced content. Each section of this chapter presents the theoretical foundations for the processes used to harvest, statistical validation, and spatial-temporal alignment.

CHAPTER 3: PROPOSED FRAMEWORK

3.1 Introduction

This chapter presents the proposed framework for extracting and classifying collective experiences from crowd contributions, spatial clustering with statistical validation, quantification of thematic alignment, and exploring thematic trends between platial content. A topic modeling approach using a variation of Latent Dirichlet Allocation (LDA), topical n-gram (TNG), is applied for dimensionally reduction of textual content. A web-based semantic labeling approach is applied to the resulting topical n-grams using a knowledgebase and a semantic similarity measure normalized by web article count. A spatial clustering approach using incremental spatial autocorrelation (Global Moran's I) and hotspot analysis (Getis Ord Gi*) with statistical significance checks is presented. Thematic alignment is proposed using proportional indexing and non-metric multidimensional scaling. And lastly, an approach to explore localized platial trends is proposed using polygon co-occurrence and the Mann-Kendall statistic.

3.2 The Approach

The proposed approach is designed with the goal of first extracting platial themes from such crowd-contributions and secondly to allow for additional investigation on the resulting themes. As such, the basic components are presented to extract semantic terms, thematically label the terms using a knowledge base, and spatially cluster. Once the clusters are formed and labeled, additional approaches are presented to investigate the semantic, spatial, and temporal alignment of the resulting platial themes.

The combination of these two approaches allows for a quantitative assessment of the data to evaluate the hypothesis. The overall framework is grounded in theoretical thought and with practical design considerations in order to discover emerging platial themes and investigate dynamics through experimentation. Figure 3 presents an overview of the proposed framework with the core components being data collection, semantic topic extraction, thematic labeling, and spatial clustering. The thematic, spatial, and temporal alignment analysis is performed on the resulting spatial clusters through experimentation to test the hypothesis.

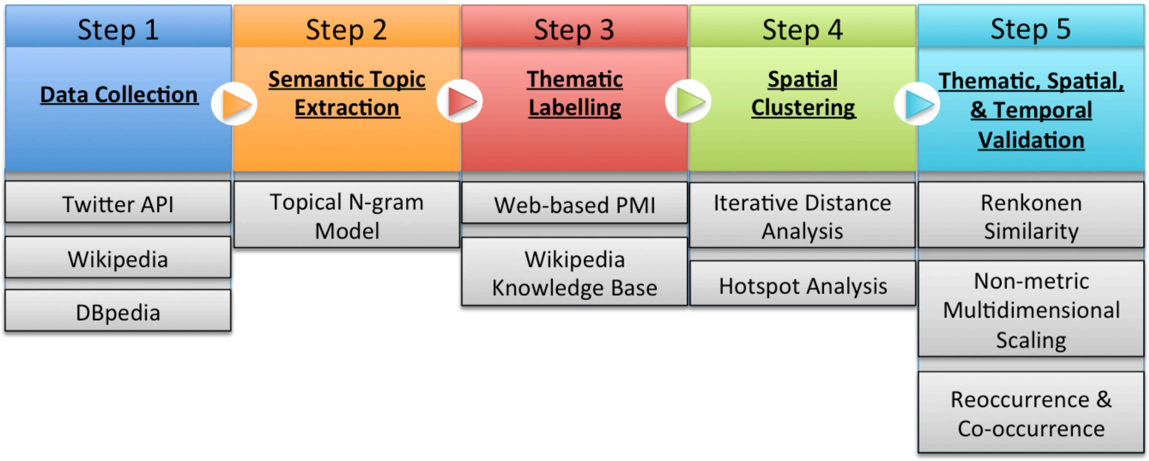


Figure 3 Overview of proposed framework.

3.3 Topical N-Gram Modeling

The specifics of the original LDA algorithm are presented to explain the process, flowed by the chosen topical n-gram approach and reasoning for its selection. Once again, and as discussed in Chapter 2, LDA is an unsupervised generative probabilistic model that discovers latent structure in a collection of documents by representing each document as a mixture of latent topics (unigrams), where a topic is itself represented as a distribution of words that tend to co-occur (Blei *et al.* 2003). LDA uses a bag-of-words assumption that does not preserve word order and only uses the frequency of co-occurring words in a document.

Blei (2003) described the use of LDA for dimensionality reduction for feature selection in supervised classifiers such as Support Vector Machine (SVM). Thus the application of LDA for dimensionality reduction is analogous to principal component analysis (PCA), albeit more complex, but nevertheless, useful for reducing large corpus of text to low-dimensional topic spaces. The algorithm converts the high-dimensional and noisy space of words and document allocations into a low dimensional topic and document allocations. This approach is applied to reduce the volume of unstructured crowd-contributed data into semantically coherent features.

LDA defines the following generative process for each document.

1. For each document, pick a topic from its distribution over topics.
2. Sample a word from the distribution over the words associated with the chosen topic.
3. The process is repeated for all the words in the document.

LDA is briefly presented using more formal notation from Newman et al. (2009). LDA models each of the D documents in a collection as a mixture over K latent topics, or topics not directly observed but rather inferred from observed words within a document, with each topic being a multinomial distribution over a vocabulary of W words. For document j, we first draw a mixing proportion θ_j from a Dirichlet with parameter α. For the i^{th} word in the document, a topic $z_{ij} = k$ is drawn with probability $\theta_{k|j}$. Word x_{ij} is then drawn from topic z_{ij} with x_{ij} taking on value w with probability $\phi_{w|zij}$. A Dirichlet prior with parameter β is placed on the word-topic distribution ϕ_k. Given the observed words $\mathbf{x} = \{ x_{ij} \}$, the task of Bayesian inference for LDA is to compute the posterior distribution over the latent topic assignments $\mathbf{z} = \{ z_{ij} \}$, the mixing proportions θ_j , and the topics ϕ_k.

Given the formal explanation of LDA, this generative approach seeks to learn the structure of words, by making assumptions of structure, and in this case a mixture of topics within a document, to determine the probability of a word belonging to one or more categories or topics based only on the data itself. This has the advantage over discriminative approaches, or at least in this application, that seek to determine a decision boundary between categories (Jordan 2002). The traditional LDA approach only looks at single words as part of a topic. This works well given a document with formal grammar, sentence structure, and higher word counts.

On the other hand, this approach does not produce coherent results on short text messages that typically refer to a single topic. In this case, a corpus of text (i.e. Twitter) can contain thousands of different topics for which the co-occurrence of words, regardless of the message context, can be grouped together to form topics. Thus the use of only unigrams, without longer phrases, can lead to disambiguation problems given polysemous words (Omar and Najib 2013). For example, Table 2 presents sample results showing generated topics using single words from newspaper articles and a collection of Twitter messages.

Table 2 Sample topics from original LDA model

Newspaper			Twitter
music	stock	theater	park
band	market	play	large
song	percent	production	night
rock	fund	show	noisy
album	investors	stage	quiet
jazz	companies	director	music
singer	trading	broadway	day

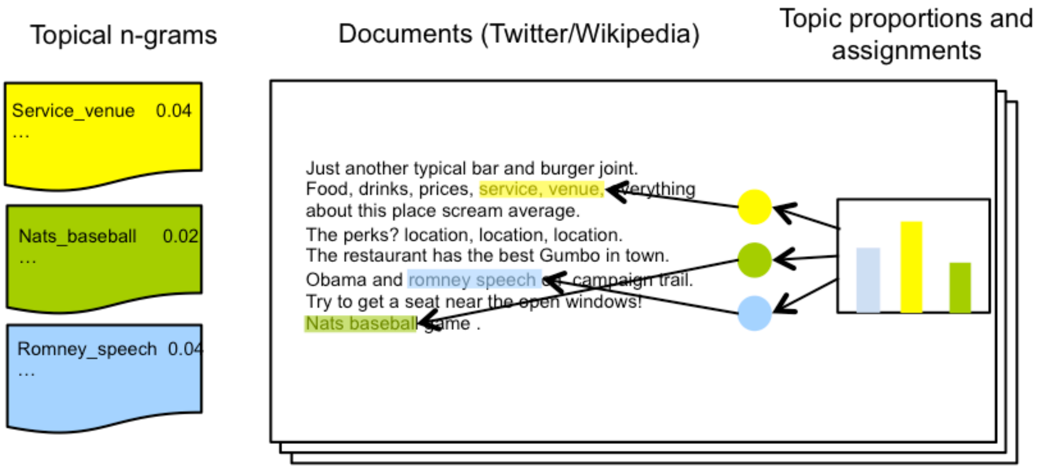
In Table 2, I observe more coherent topics from the newspaper articles with clear groupings between music, business, and theater. These topics are formed from structured text that has logical order and separation of topics into paragraphs or sections. The newspaper topics also present terms using a linguistic style that gives formal descriptions that convey context. Conversely, the Twitter topics are less formal and convey more experience and activities since the terms forming the topic have come from different tweets and thus contain a grouping of terms from various contexts. So the semantic meaning of the topic is somewhat difficult to interpret, much less assign the topical terms to a collective theme.

To overcome the described issues, the topical n-gram model (TNG) (Wang et al., 2007) was selected over the original LDA model, and previously discussed models in

Chapter 2, to provide a higher degree of topical term coherence (i.e. bi-gram) given the short 140-character limit of tweets. Thus allowing for the formation of bi-grams using collocation discovery so that terms contained within the learned topics express greater meaning individually from the context in which they are used. As previously discussed, TNG can discover unigrams, bi-grams, and longer phrases simultaneously, while allowing a single topical term to appear in multiple topics.

TNG is an extension to the original LDA model that captures the word-to-word or collocation dependencies directly in the generative process. So in essences, TNG infers both the per-topic word distribution using the standard LDA model and, for each word in the vocabulary, a distribution over the words that follow it (Wang et al. 2007). A bi-gram is formed through the co-occurrence frequency of two terms.

TNG forms longer phrases by appending to existing bi-grams by using the last word in the bi-gram in the same manner as bi-gram construction. Unsurprisingly, longer phrases produced by TNG using Twitter data were very incoherent given the shortness of the text. And notably, Ciaramite et al., (2015) remarked on the incoherence of TNG phrases using different data sources. For the purposes of this research only bi-grams were used. Figure 3 presents a graphical representation of the bi-gram process and sample topical n-gram outputs with color-coding.


Topical n-grams
Documents (Twitter/Wikipedia)
Topic proportions and assignments
Service_venue 0.04
…
Nats_baseball 0.02
…
Romney_speech 0.04
…
Just another typical bar and burger joint.
Food, drinks, prices, service, venue, everything
about this place scream average.
The perks? location, location, location.
The restaurant has the best Gumbo in town.
Obama and romney speech on campaign trail.
Try to get a seat near the open windows!
Nats baseball game .

Picture adapted from http://www.cs.princeton.edu/~blei/papers/Blei2012.pdf

Figure 4 Graphical representation of the topical n-gram modelling process.

Lastly, an important parameter in any unsupervised topic modeling process is setting the number of topics. According to Arun et al. (2010), finding the right number of latent topics in a given corpus has remained an open-ended question. However, while investigating Wikipedia articles, Arun et al. (2010) found a lower Kullback-Leibler divergence value when the number of topics ranged from 15 to 30 for a single article. For Twitter, TNG was executed on 24-hour time periods forming documents and running 1000 iterations of Gibbs sampling with the number of topics set at 700. This number of topics was determined qualitatively by assessing the specificity of topic outputs using multiple trials. Stopwords (i.e. *and*) and special symbols were removed during each trial using a standard list initially that was iteratively modified with common words specific to Twitter.

3.4 Web-Based PMI

PMI is a measure of association originating from information theory and statistics that calculates the probability of two events occurring given their individual and joint distributions (Turney 2001). The measure has been applied in the field linguistics, from which I use here, to find associations between words using an online knowledge base. Recchia and Jones (2009) compared PMI with Latent Semantic Analysis (LSA) using Wikipedia and Spearman rank correlation between human judgments of semantic similarity.

Recchia and Jones (2009) reported correlations ranging between 0.73 and 0.86, and concluded that the simple PMI metric with large volumes of data correlates more closely with human semantic similarities ratings than more complex models. The selection of PMI was based solely on empirical evidence and some exploratory analysis of different metrics as the contribution of this research is discovering platial themes and not an exhaust inquiry into the optimal metric.

The PMI calculation used is adapted from Matsuo et al., (2006) with the notation shown in Equation 1. The probability $p(w_1)$ is estimated by f_{w1}/N, where f_{w1} is the Wikipedia web count of topic term w_1 and N is total number of Wikipedia articles. The probability $p(w_2)$ is estimated by f_{w2}/N, where f_{w2} is the Wikipedia web count of theme w_2 and N is total number of Wikipedia articles. The probability of co-occurrence $p(w_1,w_2)$ is calculated by $f_{w1,w2}/N$, where $f_{w1,w2}$ is the Wikipedia web count of w_1 AND w_2 and N is the constant total number of Wikipedia articles.

Equation 1 web-based pointwise mutual information.

$$\mathrm{PMI}(w_1, w_2) = \log_2\left(\frac{\frac{p(w_1, w_2)}{N}}{\frac{p(w_1)}{N}\frac{p(w_2)}{N}}\right) \tag{1}$$

PMI values are inferred using the following conditions:

- If PMI(w_1, w_2) > 0, then topic term w_1 and category term w_2 are correlated with high probability within Wikipedia and the larger the value the higher the probability.
- If PMI(w_1, w_2) = 0, then topic term w_1 and topic term w_2 are isolated from each other within Wikipedia.

Given the resulting PMI values for each of the six categories per topic term w_1; the category with the largest PMI value is assigned to the single topic term. The thematic assignment of a category to a term is performed despite the level of correlation with one exception. The largest PMI value must be greater than zero for assignment, otherwise no assignment will occur. The rationale for this, and previously mentioned, is that the topic term and categorical term are isolated within Wikipedia. Given the linguistic style of Twitter containing grammatical errors, abbreviations, and short messages that are aggregated to form documents in the topic modeling process, erroneous and highly ambiguous terms arise frequently. Applying this condition filters noisy terms that Wikipedia cannot disambiguate.

Algorithm 1 below presents the approach for using topical n-gram terms from the crowd-contributed data sources with PMI. The code is initiated with a corpus of text that could be Twitter data divided into temporal bins or Wikipedia articles. High-level categories such as Sports, Entertainment, etc. are provided as inputs for calculating PMI. An empty list, *maxList*, is provided for storing the maximum PMI per topic and high-level category. The algorithm iterates through the documents and calculates topics for each. It then loops through the results topics and high-level categories using PMI and Wikipedia article counts to determine the category with the highest PMI value, which is then returned.

Algorithm 1 Thematically labelling topics using TNG and PMI.

```
Input:
T ← textDocuments()                         // Twitter or Wikipedia article.
L ← $l_1,l_2...l_n$                          // High-level categories (i.e. Sports).
maxList ← ∅                                 //Empty list for calculated returns.
for all d in T do
    tng ← TNG(d)                            //Calculate topical bi-grams.
    for all t in tng do
        topicLabels ← ∅
        for all l in L do
            p ← PMI(t,l)                    //Calculate PMI using Wikipedia.
            topicLabels.add({t,l,p})
        topic, label, pmi ← maxPMI(topicLabels)   //Calculate maximum PMI.
        maxList.add({topic, label, pmi})
return maxList
```

3.5 Incremental Distance Analysis and Local Spatial Clustering

Turning to the spatial aspect of the approach, this section builds on the previous, in that, the topics, labels, and PMI features generated are spatially clustered. In doing so, simply observing spatial or platial processes using a single distance value across multiple locations will produce undesirable results as geographies vary considerably (Lewicka, 2010). This is particularly evident in crowd-contributed content were events and activities can occur at varying scales (Westerholt et al., 2015). To account for the differences in scale and distance for which spatial clustering occurs regionally or at the city scale, incremental distance analysis using Global Moran's I (Odland, 1988; Paez and Scott, 2005) is used.

The spatial statistic tests for spatial autocorrelation based on both feature locations and feature values simultaneously using a spatial weights matrix. Moreover, this classic spatial statistic assesses whether patterns within a data source are clustered, dispersed, or random. The statistic outputs three measures that include Moran's *I* index, a z score, and p-value.

The resulting index value is assessed on a range from -1 to 1, where negative values indicate neighboring points have different values, values near zero indicate random neighboring values, and positive values indicate similar neighboring values. The process is an inferential statistic whereby the results are interpreted using the null hypothesis that states that the spatial pattern is random, compared to the alternate hypothesis that the underlying pattern is more clustered than expected. When the resulting p-value and z score are significant, and the null hypothesis is rejected, the direction of the z score is used to determine whether the pattern is dispersed or clustered.

The formula for calculating Global Moran's *I* is provided in Equation 2 (Odland, 1988). Where N is the total number of locations, x_i is value of the variable (i.e. PMI) at location i, x_j is the value at location $j, \bar{x}$ is the mean for the variable across all locations, and w_{ij} is the value of the spatial weighting function to assess location x_i and x_j based on proximity. To determine statistical significance, from a computed value for Moran's *I*, a standardized z score is obtained under the assumption of randomization (see Odland, 1988). A significant positive z score indicates that neighboring locations, in this case geo-located and thematically labeled tweets have similar PMI values than expected.

Equation 2 Global Moran's I statistic for spatial autocorrelation.

$$I \quad = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \qquad (2)$$

Once again, the Global Moran's I statistic is used to assess regional patterns using a spatial weights matrix based on proximity and a selected feature attribute. As such, testing for patterns as varying distances in crowd-contributed content becomes problematic, as the distance is typically unknown and differs depending on the location. To overcome this limitation, Global Moran's I is calculated incrementally over a specified distance interval to determine the most appropriate value that corresponds to the highest statistical significance using the reported z scores. The graph in Figure 4 provides an example of spatial autocorrelation incrementally run over varying distances to find the peak. Such analysis is performed on each of the high-level categories individually and at each study location to account for multi-scale activities.

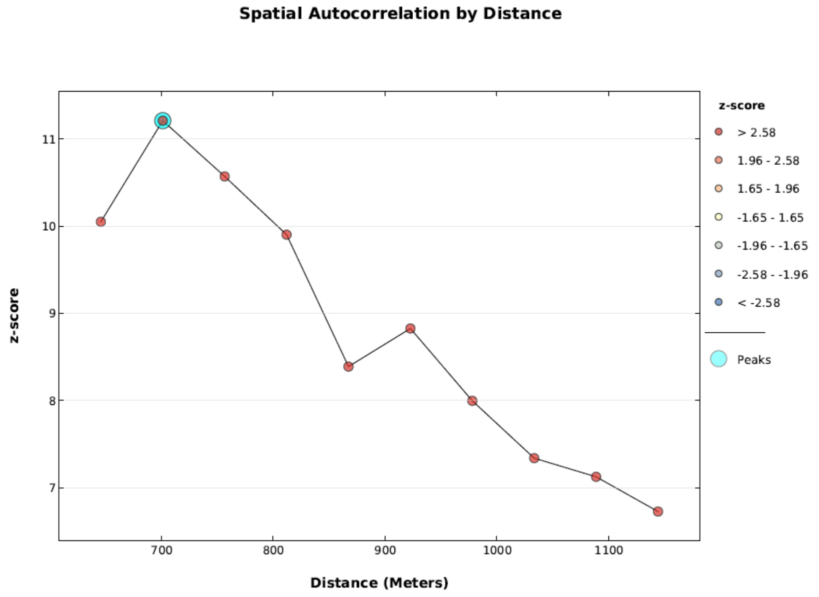


Figure 5 Incremental spatial autocorrelation by distance using Global Moran's I and thematically labelled Twitter data for Los Angeles, CA.

Once the analysis distance is determined, a local spatial autocorrelation statistic is needed to find areas that exhibit clustering of high values. Even more, and to assess the hypothesis of this work, the local statistic has to calculate not only high values of PMI per category, but high values surrounded by other high values to form a collective cluster. Given this, the Getis-Ord Gi* (Getis and Ord, 1992) statistic is used to discover local spatial clusters with high semantic values within the boundaries of each. Grieve (2012) showed the use of combining both Global Moran's I and Getis-Ord Gi* to study regional linguistic variations by looking at key words found in a corpus of letters to the editor. The research presented in this dissertation is closely aligned with the work of Grieve, as discussed in section 2.6, but differs in that incremental spatial autocorrelation is calculated here.

To describe the Getis-Ord Gi* statistic provided in Equation 3 (Getis and Ord, 1992) where N is the total number of locations, x_i is value of the variable (i.e. PMI) at location i, x_j is the value at location $j, \bar{x}$ is the mean for the variable across all locations, and w_{ij} is the value of the spatial weighting function to assess location x_i and x_j based on proximity. The local statistic produces z-scores and p-values for each feature by looking at neighbouring features given a weighting, in this case distance is used as the weight, to determine spatial clusters of high or low values. The statistic is calculated by taking the sum of PMI values combined with its neighbours within the set distance, and compared this to the expected local sum from all PMI values combined. When the difference between these two proportions is considerable and not by random chance, the single tweet is labelled as significant.

Equation 3 Getis-Ord Gi* statistic for local spatial autocorrelation.

$$G_i^* = \frac{\sum_j w_{ij} x_i - \bar{x} \sum_j w_{ij}}{\sqrt{\frac{\sum_j x_j^2}{N} - \bar{x}^2} \sqrt{\frac{N \sum_j w_{ij}^2 - \left(\sum_j w_{ij}\right)^2}{N-1}}} \tag{3}$$

Given the definitions of Global Moran's I and Getis-Ord Gi*, the two statistics are combined using Algorithm 2. The inputs include a spatial dataset that is thematically labeled and contains PMI values for each feature and minimum, maximum, and increment distances. The algorithm iterates through the range of distances and at each step calculates Global Moran’s I, in which the resulting distance and z score is recorded. Once complete, the maximum z score and corresponding distance are determined from all the values as in Figure 4.

The resulting distance is then used to compute the Getis-Ord Gi* statistic were the output is filtered to only return statistically significant points with z score greater than or equal to 1.96 representing high-values. In doing so, a feature with a high PMI similarity score is interesting, but may not be a statistically significant hot spot or a place with nearby features with similar place expressions. Moreover, a feature (i.e. single geo-located tweet) will have a high PMI similarity value and be surrounded by other features with high values as well.

Algorithm 2 Combining Incremental Global Moran’s I and Getis-Ord Gi*.

```
Input:
D ← getThematicSpatialData()               // Shapefile or PostGIS with PMI
I ← [$d_{min}$,$d_{max}$]                   // Min and max distance range.
Step ← number                               // Increment value.
maxDistList ← ∅                             // Empty set for results.
for all d in range(I[$d_{min}$], I[$d_{max}$],Step)
    w_Moran ← SpatialWeightMatrix(D[x,y],d)
    dist,z ← GlobalMoranI(w_Moran,D[pmi])      //Calculate Moran’s I.
    maxDistList.append({dist,z})
distMaxZscore ← max(maxDistList)                //Get distance with max zscore.
w_Getis ← SpatialWeightMatrix(D[x,y], distMaxZscore)
G ← GetisOrdGi*(w_Getis,D[pmi])                 //Calculate Getis-Ord Gi*.
return {zscore ∈ G | zscore ≥ 1.96}             //Only return points z ≥ 1.96.
```

3.6 Multi-Source Semantic Alignment

One approach to evaluate the hypothesis regarding emerging and unique platial themes is to test the similarity of crowd-contributed content from multiple sources is to assess thematic proportional alignment. Since platial themes are formed through human activities, and crowd-contributed content is formed through different means and platforms (i.e. Twitter and Wikipedia), it stands to reason that unique places should exhibit similarities regardless of the medium. As such, the Renkonen (1938) percentage similarity index, Equation 4, was selected to assess proportional alignment of high-level categories due to the minimal affect of sample size differences (Wolda 1981).

The Renkonen similarity measure is traditionally used in computing the proportional abundances of species between communities and habitats, which is analogously applied here between cities and neighbourhoods. The variables $p_{1,i}$ and $p_{2,i}$ in Equation 4 are determined by the total sum of entities assigned to a thematic category, which are then transformed into percentages. When comparing multi-sources over the same geographic locations, the minimum percentage from the same thematic category in both sources is determined and then the minimum sum from all the categories forms the final similarity value. The resulting percentage value is reported between 0 and 1.

Equation 4 Renkonen percentage similarity.

$$PS \quad = \textstyle\sum_{1}^{i} \min\left(p_{1,i}, p_{2,i}\right) \qquad (4)$$

To visualize collective uniqueness of places, non-metric multidimensional scaling (NMDS) is used by taking the Renkonen similarities as distances of dissimilarity using the rank order of the data (Kenkel and Orloci 1986). The combination of NMDS and Renkonen similarity has previously been used to investigate and visualize percentage similarities of habitat abundances (Janáč and Jurajda 2013) and categorical abundance (de Mattos et al. 2013). Both works cite this approach as having low sensibility to a limited number of high-level categories, which makes this appropriate for this framework. Given the Renkonen similarities values expressed between 0 and 1, I first converted these values to dissimilarities by subtracting each value by 1 to form our matrix.

NMDS seeks an ordination of the data in which the distances between all pairs of data points, in this case the places from Twitter and Wikipedia, are at the greatest distance apart, while maintaining rank-order agreement with their dissimilarities. Meaning data pairs with dissimilarity less than other data pairs will have a shorter distance apart in the visualization from other data pairs with greater dissimilarity values. NMDS uses an iterative process to decrease the stress of rank-order agreement between distances and dissimilarity values; as such, I explored various iteration values to minimize the stress to a near 0 value as described in (Kenkel and Orloci, 1986)

3.7 Spatiotemporal Alignment

This section presents an approach to investigate the research component of how well such mined platial themes exhibit spatiotemporal alignment. Since the human activities that form collective platial content are highly dynamic and place meanings evolve over time it’s only appropriate to investigate the emergence of such phenomena. Given that crowd-contributed content, in the form of Twitter data, affords both spatial and temporal attributes, that when combined with thematic labeling, has the potential to show both the emergence of new places and more established places that frequently emerge within the data. To assess if spatiotemporal alignment exists, a general framework is proposed that uses the thematically labeled Getis-Ord Gi* output, Density-based spatial clustering of applications with noise (DBSCAN) (Ester et al., 1996), and convex hulls to form polygon representations of place.

In regards to the geometric representation of place, and to reiterate from Chapter 2, the literature makes clear the difficulties in assigning any geometric shape to places of social meanings, as the boundaries are indeterminate and vague. The framework presented here is general so as to assess co-occurrence and alignment with the understanding that using even more elaborate approaches remain problematic. Numerous studies exist that suggest more elaborate methods when addressing the uncertainty and shapes of place, for example, using the point-radius method (Wieczorek et al. 2004), probabilistic methods (e.g., Liu et al. 2009), or fuzzy-set approaches (Zadeh 1975). Schlieder et al. (2001) developed a spatial footprint representation, in the form of connection graphs derived from a qualitative abstraction of tessellation polygons. Gao et al. (2014) introduced a fuzzy-set-based method to extract geographic footprints of polygonal places based on a distance-decay function between a geotagged photo and the centroid point of the cluster.

The framework presented here is similar to that of Pillai et al. (2012), who used polygon representations of solar flares and sun spots to find spatiotemporal co-occurrence patterns. The authors proposed an apriori algorithm approach to mine the frequent patterns, which is not used in this framework. However, the use of polygon intersection and overlap is adapted and later implemented as a database query.

To assess the spatiotemporal alignment using polygon representations the resulting Geits-Ord Gi* spatial hotspots have to be uniquely identified as clusters. Getis-Ord Gi* outputs statistically significant points within a neighborhood of high values, but the neighborhood is not identified as a single cluster, just individual points. To unique identify each cluster, DBSCAN (Ester et al., 1996) is proposed since the number of clusters is not known, but the distance is using iterative Global Moran's I. Different clustering algorithms abound, the density-based clustering approach, DBSCAN, was selected over k-means type approaches as they require the user to specify the number of clusters, which in this application is not known. Numerous other density-based clustering algorithms exist in the literature and in practice, such as OPTICS (Ankerst et al., 1999), DenClue (Hinneburg and Keim, 1998), and CLIQUE (Agrawal et al., 1998), which offer various advantages over DBSCAN, however, an investigation into this is reserved for future research.

DBSCAN locates regions of high density that are separated from one another by regions of low-density using points within a specified radius (Ester et al., 1996). This is accomplished given the two parameters, *epsilon* and *minPoints*. The algorithm starts by picking a random point in the dataset and determining if the total number of points surrounding the chosen point, within the distance of *epsilon*, is greater than *minPoints*. If the criteria is satisfied, the reachable points form a cluster, the cluster is then expanded recursively until no other points meet the criteria, in which case a new point that has not been visited is selected and the process is repeated. When a single point is selected that does not meet the criteria to form a cluster, the point is labeled as noise and become an outlier.

The DBSCAN process outputs clusters of points, in this case the Getis-Ord Gi* results, of dense regions that are converted into convex hulls to form polygons. Convex hull is a shape given by the convex closure of the set of points (Berg et al., 2000). It is the smallest convex polygon shape or bounding region that encloses a set of points. The input geometry to a convex hull operation can be points, lines, or polygons. More simplistically, the elastic band analogy is often used for visualization by imagining a band stretched around the outermost part of a set of points, and creates straight lines between the outermost features resulting in a convex hull. Figure 5 gives an example of a simple convex hull.

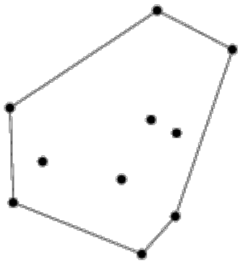


Figure 6 Sample convex hull. (Image from http://mathworld.wolfram.com/ConvexHull.html.)

The process for deriving polygons is presented in Algorithm 3 below. Once again, the statistically significant points from Algorithm 2 are taken as input. The *epsilon* value is specified by the distance with the highest z score using Global Moran's I incrementally. The number of *minPoints* is heuristically set to three to allow for the formation of polygon regions that are later investigated using overlap and intersect. The DBSCAN algorithm is then executed on the data and clusters are returned. A convex hull is then created for each cluster iteratively.

Algorithm 3 Forming polygons using DBSCAN and Convex Hull.

```
Input:
P ← getGetisOrdGiPnts()              // Thematic Getis-Ord Gi* points
Eps ← getIncrMoranIdist()            // Use max incremental distance value.
minPnts ← 3
clusters ← DBSCAN(P,Eps,minPnts)
polygons ← Ø
for all c in clusters do
    p ← ConvexHull(c)
    polygons.add(p)
return polygons
```

3.8 Summary

In summary, this chapter presented the proposed framework for discovering and thematically labeling platial knowledge from crowd produced content. Each section of this chapter presents the theoretical foundations for the processes used to harvest, statistical validation, and spatial alignment. Specifically, the chapter starts with a brief description and supporting claims for the chosen approach and selected methodology.

The collection process is presented using topical n-gram modeling for dimensionally reduction with web-based semantic similarity and classification using Wikipedia. In addition, the spatial clustering with statistical significance checks is presented as the approach to detect clusters of semantic consensus from the crowd. The chapter concludes with two approaches are introduced for investigating the dynamic and evolving nature of platial knowledge through trend analysis and co-occurrence. In the following Chapter 4, I present the application of the proposed framework for harvesting collective and crowd-centric platial content from Twitter and Wikipedia.

CHAPTER 4: A FRAMEWORK FOR ASSESSING SPATIAL AND THEMATIC ALIGNMENT

4.1 Introduction

This chapter presents a framework in which the approaches in Chapter 3 are applied to crowd-contributed data to address the research question of discovering and alignment of platial content. Specifically, this Chapter focuses on two distinct crowd-contributed sources: the consensus expressing crowd-curated Wikipedia content, and the uncurated content of individual tweets. While they are both expressions of crowd views they differ substantially in their purpose (Kaplan and Haenlein, 2010).

The former, i.e. Wikipedia entries, captures platial knowledge in the form of place entries, whose content is the outcome of a collaboratively-derived consensus on the main characteristics of such locations. Accordingly it can be viewed as an expression of the collective perception of such places. The latter, i.e. individual geotagged tweets made from such locations, simply express individuals' concerns, observations, or interests while they are there. More specifically, research question pursued here is how to extract platial content from such crowd contributions, capturing the sociocultural characteristics of a place, and to quantify the alignment of these two sources with respect to platial content.

4.2 Gathering Crowd Generated Content

The first steps in addressing the research questions is the collection of Wikipedia entries and semantic access statistics for various locations and Twitter streams originating from these locations as well. Four major cities are selected for the study areas, namely New York City (NYC), Los Angeles (LA), Singapore (SG), and London (LDN). These locations were chosen because of Twitter volume, population, and number of English speakers from different countries around the world. In order to highlight the research question at a finer geographical scale, three additional study areas are selected at the neighbourhood scale from NYC (Lower Manhattan, Central Park, and Theatre District). It should be noted here that in the context of this paper the term neighbourhood is used to refer to such sub-city level areas. Considering that the notion of place is inherently multiscalar, such a study at both scales is necessary, as cities encompass a broad, diverse geographic area of affordances for human activities compared to neighborhoods that our more homogenous (Lewicka 2010). Additionally, the selected neighborhood locations were chosen because each contains popular landmarks (i.e. Wall Street, Broadway, Central Park) that relate to one of the selected high-level categories, which contributes to the validation of the discovered platial clusters and cross-source consensus.

Each single Wikipedia entry was accessed programmatically using the article's webpage address. Geo-located Wikipedia articles were also investigated as an additional crowd-contributed data source, along with their associated online semantic access statistics that are spatially contained with each study area. The rational for inclusion of this data is that Twitter is highly dynamic and event driven, and a single Wikipedia article (e.g. NYC), although derived from collective consensus, is comparatively more static. Hence, by thematically classifying each geo-located article using its title as the topic and the number of online page accesses from the same month as the categorical proportions, I extract a more dynamic representation of place from Wikipedia.

DBPedia (Auer et al., 2007) which offers a more structured means to access Wikipedia content, was used to collect titles and geographic coordinates from geotagged articles using the methodology adapted from (Popescu and Grefenstette, 2010). The official boundaries of each study were used as before with Twitter to spatially filter the geotagged Wikipedia articles. A Wikipedia statistics site (Stats.grok.se, 2015) was used to gather the number of page accesses for each article title over the same time period as the Twitter data was collected (i.e. April 2014).

Geotagged tweets were collected from each of the study areas using Twitter's streaming Application Programming Interface (API) that returns a sample of 1% of all tweets. Previous studies have shown that such precisely geotagged tweets typically reflect a sample of approximately 2% (Burton et al., 2012) of the entire Twitter API traffic. Twitter data was gathered continuously for an entire month (April 2014), resulting in approximately 4.5 million tweets for NYC, 4.3 million for LA, 821,000 for SG, and 1.3 million for LDN. City boundaries for each location were downloaded from the local city governments to include the sub districts for NYC. The tweets were further filtered by English language and spatially using the geographic boundaries of each city and stored in a relational database.

An additional data input to this study is the selection of thematic categories used to semantically related topics to collective grouping. The following were selected politics, business, education, recreation, sports, and entertainment as high-level categories. These particular categories were chosen because they have been shown to be dominant in both Wikipedia and Twitter (Kittur et al., 2009; Zhao et al., 2011), with entertainment in particular having a disproportionately strong presence in user-generated media (Shao, 2009; Hargittai and Litt, 2011). Each of the above-listed high-level categories is an explicit category in Wikipedia and appears in over 50,000 articles. Previous studies have demonstrated content alignment with said categories for classifying trending topics (Lee et al., 2011) filtering tweets into new groups (Dilrukshi et al., 2013), and categorizing the professions of users (Wagner et al., 2013).

4.3 Mining Platial Content From Twitter and Wikipedia

The overall approach is presented in Figure 6, for which the Twitter processing is briefly described first. In order to extract and label topics from our data corpus I applied a variation of the probabilistic topic model LDA (Blei et al., 2003), known as the topical *n*-gram model (TNG) (Wang et al., 2007), treating 24-hour subsets of our data corpus as the equivalent of individual documents for the TNG analysis (see Algorithm 1). Each detected bi-gram is considered part of a *topic of discussion*. Discussion topics may be classified linguistically as belonging to multiple high-level categories by considering their *PMI* value (Turney, 2001). In the context of this research I thematically classify each topic as belonging to the high-level category that corresponds to its highest PMI value, which are estimated using the entire online Wikipedia corpus and comparing the frequencies of topic and category co-occurrences.

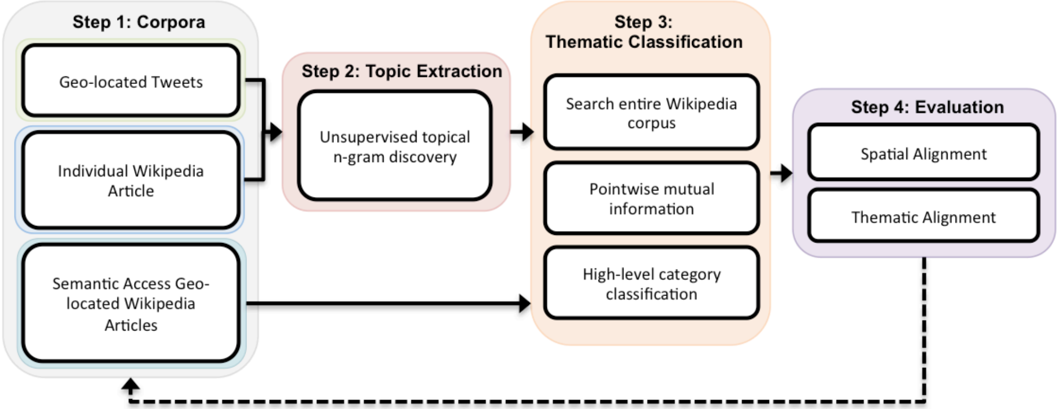

Step 1: Corpora
Geo-located Tweets
Individual Wikipedia Article
Semantic Access Geo-located Wikipedia Articles
Step 2: Topic Extraction
Unsupervised topical n-gram discovery
Step 3: Thematic Classification
Search entire Wikipedia corpus
Pointwise mutual information
High-level category classification
Step 4: Evaluation
Spatial Alignment
Thematic Alignment

Figure 7 Flowchart describing overall process used to discover platial alignment. Step 1: sequential ingestion of Twitter, and Wikipedia article representing each spatial location, and all geo-located Wikipedia article titles spatially and semantic access statistics contained within the same location. Step 2: unsupervised topic discovery using topical *n*-grams. Step 3: determine the semantic overlap of topical *n*-grams and defined categories using Wikipedia search counts and PMI. Step 4: determine the statistical significance of each platial category and compare proportional alignment between Twitter and Wikipedia.

For this particular study, bi-grams perform better than unigrams or phrases, when it comes to their thematic classification. This is consistent with studies addressing the performance of *n*-gram classification in Twitter, which also supported the notion that bi-grams outperform other alternatives (Pak and Paroubek, 2010). This is not surprising, as given the brief nature of tweets (at only 140 characters per tweet), unigrams are ambiguous, while meaningful longer phrases tend to be scarce. Based on the labelling of the bi-grams contained within them, individual tweets were classified under the 6 top-level categories as shown in Table 2.

Table 2 demonstrates the dominance of entertainment content in Twitter data. The results of this Table were obtained by setting a maximum value of 700 to the number of bi-gram topics that were extracted from our data corpus. In our studies we also compared the effect of varying the number of topics (i.e. from 100 to 700 in increments of 100) and found that the proportional allocations to high-level themes were consistent.

Table 3 Total count of tweets per high-level category containing a semantically related topical n-gram.

Cities	High-level Categories					
	Politics	Education	Entertainment	Sports	Business	Recreation
NYC	12,893	27,122	242,226	34,998	32,060	55,832
LA	11,207	28,258	281,556	32,412	42,264	57,390
LDN	3,697	5,765	52,412	6,227	10,613	21,711
SG	2,207	5,717	28,863	3,621	6,247	7,363

To recall from Arun *et al.* (2010), finding the right number of latent topics in a given corpus has remained an open-ended question. In which case, the numbers of topics were iteratively in Figure 7. The purpose of this approach was to determine at what point topic coherence might become so degraded that Wikipedia would have trouble disambiguating the terms and thus the categorical counts would exhibit a downward trend. It was observed that in each location the topic proportions per category followed an upward linear trend for each increment.

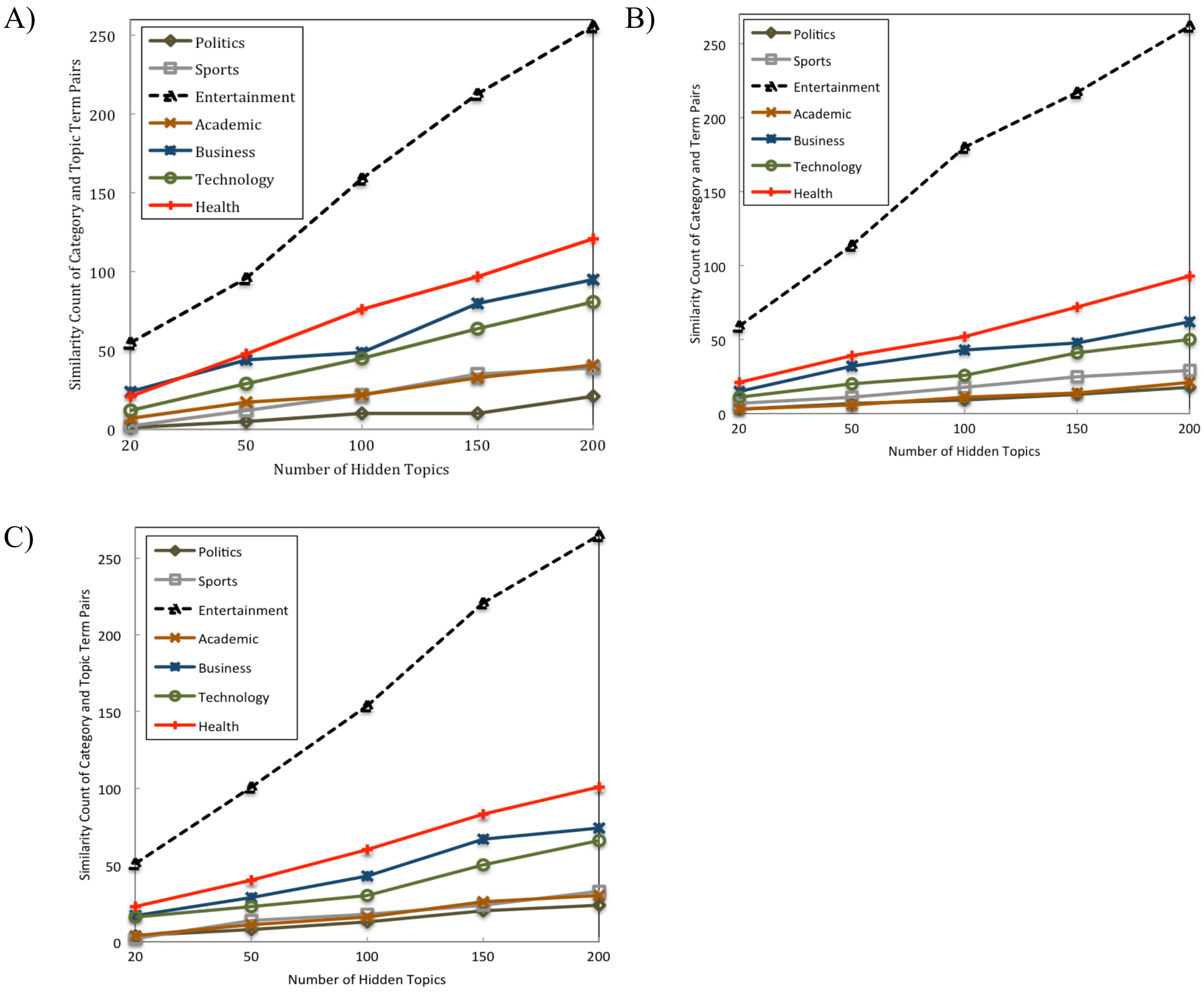


Figure 8 Graph of k-topic iterations with categorical similarity count. (A) Singapore (B) NYC (C) LA

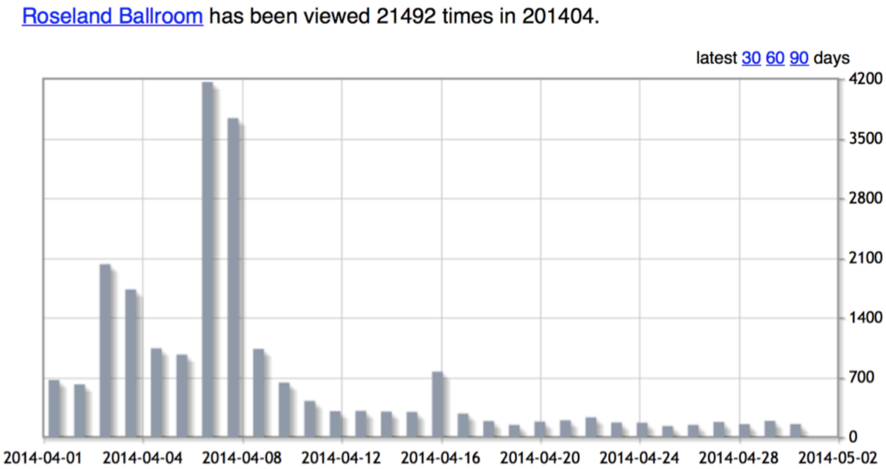
The discovered trend depicted a linear relationship in the semantic counts per category as the number of topics increase, but also a significant bias of entertainment related tweets within the data. The graphs were produced using additional thematic categories initially to get a sense of content, but those such as health and technology were determined to have excessive semantic overlap with other categories.

For single Wikipedia articles (Figure 6), I applied the same approach to analyzing individual article content, and more specifically the single entry for each city and neighbourhood that are subjects to our study. Each single Wikipedia entry was accessed programmatically using the article's webpage address. I extracted topical bi-grams from these entries, and allowed for up to 30 topics per document, with up to 10 bi-grams per topic, reflecting the smaller size of these documents, compared to our Twitter data corpus. The number of topics was based on the empirical evidence from (Arun et al., 2010), as 30 was shown to optimize the Kullback-Leibler (1951) divergence value. Bi-grams within each topic were thematically classified against each of the six high-level categories using the aforementioned PMI process.

I also investigate geo-located Wikipedia articles and their associated online semantic access statistics that are spatially contained with each study area. The rational for inclusion of this data is that Twitter is highly dynamic and event driven, and a single Wikipedia article (e.g. NYC), although derived from collective consensus, is comparatively more static. Hence, by thematically classifying each geo-located article using its title as the topic and the number of online page accesses from the same month as the categorical proportions, we extract a more dynamic representation of place from Wikipedia.

I used DBPedia (Auer et al., 2007), which offers a more structured means to access Wikipedia content, to collect titles and geographic coordinates from geotagged articles (Popescu and Grefenstette, 2010). The official boundaries of each study were used as before with Twitter to spatially filter the geotagged Wikipedia articles. I then used a Wikipedia statistics site (Stats.grok.se, 2015) to gather the number of page accesses for each article title over the same time period as the Twitter data was collected (i.e. April 2014).

As shown in Figure 6, the title of each article is processed using the same thematic classification approach as the topical n-grams. Instead of using topic counts to calculate category proportions (see Table 2), the number of page accesses was used. To illustrate, take for example the geo-tagged Wikipedia article entitled “Roseland Ballroom”, the article was thematically classified as entertainment and accessed 21,492 times in the month of April 2014. The process is repeated for all articles within the same boundaries and the total proportions for each category are calculated. Figure 8, provides the semantic page view statistics for Roseland Ballroom for the entire month of April 2014 with the total views.



Roseland Ballroom has been viewed 21492 times in 201404.

Figure 9 Wikipedia page view statistics for Roseland Ballroom. Image taken from Stats.grok.se.

I selected a variation of LDA, topical n-gram model (i.e. bi-grams), given the short 140-character limit of tweets to achieve higher topic coherence. LDA is an unsupervised generative probabilistic model that discovers latent structure in a collection of documents by representing each document as a mixture of latent topics (unigrams), where a topic is itself represented as a distribution of words that tend to co-occur (Blei et al., 2003). In general, topic modelling enables the conversion of high-dimensional and noisy spaces of words and document allocations into a low-dimensional topic and document allocations. Blei et al. (2003) described the use of LDA for dimensionality reduction for feature selection in supervised classifiers such as Support Vector Machine (SVM).

Individual bi-grams within each topic were programmatically sent to Wikipedia's search engine to capture the search count (total number of articles containing the terms). This process was repeated for each high-level category and finally the joint count of each bi-gram and high-level category. I normalized each search count over the total number of articles in Wikipedia, which at the time of this study was N = 4,502,037. A PMI value was calculated for every bi-gram for each of the six high-level categories.

The equation $PMI(w_1, w_2) = \log_2(\frac{\frac{p(w_1, w_2)}{N}}{\frac{p(w_1)}{N}\frac{p(w_2)}{N}})$ depicts the PMI calculation adapted from (Matsuo et al., 2006). The probability $p(w_1)$ is estimated by f_{w_1}/N, where is the Wikipedia web count of topic term w_1 and *N* is total number of Wikipedia articles. The probability $p(w_2)$is estimated by f_{w_2}/N, where f_{w_2} is the Wikipedia web count of theme w_2 and *N* is total number of Wikipedia articles. The probability of co-occurrence

$p(w1, w2)$ is calculated by $f_{w_1}f_{w_2}/N$, where $f_{w_1}f_{w_2}$ is the Wikipedia web count of w_1 and w_2, and *N* is the total number of Wikipedia articles.

In order to assess the PMI of the topical *n*-gram results related to the high-level categories I performed a manual assessment of a random subset of our Twitter data and found a nominal F_1 score of 0.70. The bi-grams that Wikipedia could not disambiguate or interpret, thus resulting in a search count of zero, were not included in the results. This accuracy is consistent with similar studies using PMI, albeit without Twitter (Iosif and Potamianos, 2010). Recchia and Jones (2009) compared PMI with Latent Semantic Analysis (LSA) using Wikipedia and Spearman rank correlation between human judgments of semantic similarity. These researchers reported correlations ranging between 0.73 and 0.86, and concluded that the simple PMI metric with large volumes of data correlates more closely with human semantic similarities ratings than more complex models.

I used Global Moran's I (Páez and Scott, 2005) to find an appropriate analysis distance per location by incrementally checking for spatial statistical significance at incremented distances using Global Moran's I to test for spatial autocorrelation and the distance with the highest significance value. The distance was recorded for each of the four cities and parameterized as the fixed distance threshold in the local Getis-Ord Gi* (Getis and Ord, 1992) along with the PMI values for each category. A feature with a high PMI similarity score is interesting, but may not be a statistically significant hot spot or a place with nearby features with similar place expressions. Moreover, a feature (i.e. single geotagged tweet) will have a high PMI similarity value and be surrounded by other features with high values as well.

I determined the Getis-Ord distance parameter by observing the spatial autocorrelation using Global Moran's I statistic at a series of distances. Using the geotagged tweets and their PMI values per high-level category, I found the prominent distance with a statistically significant z-score ($z > 1.96, P < 0.05$), and detected the following measures: LA 701.3m, NYC 619.2m, SG 1,334.9m, LDN 731.4m. I applied each distance and the PMI values as weights in computing Getis-Ord Gi* statistic to discover local spatial clusters with high semantic values within the boundaries of each city. I determined statistical significance at the $P < 0.05$ level and considered only z-scores greater than +1.96.

I selected the Renkonen (1938), percentage similarity index ($PS = \Sigma\, min(p_{1i}, p_{2i})$) to assess proportional alignment of high-level categories due to the minimal affect of sample size differences (Wolda, 1981). The similarity measure is traditionally used in measuring the proportional abundances of species between communities and habitats. I used the counts per high-level category to calculate percentages between each individual source at different locations (i.e. just Twitter or Wikipedia) and then a cross-source comparison (i.e. Twitter and Wikipedia) to determine a similarity score between 0 and 1 reported as a percentage.

I used non-metric multidimensional scaling (MDS) to visualize the Renkonen similarities as distances of dissimilarity using the rank order of the data (Kenkel and Orlóci, 1986). The combination of non-metric MDS and Renkonen similarity has previously been used to investigate and visualize percentage similarities of habitat abundances (Janáč and Diel, 2013) and categorical abundance (de Mattos et al., 2013). Both works cite this approach as having low sensibility to a limited number of high-level categories, which makes this appropriate for my work.

4.4 Evaluation and Results

In order to test the proposed approach for discovering platial content in crowd-contributed content it is necessary to present a formal assessment of the results. Within this section, the spatial alignment of the resulting thematic hotspots is investigated in detail with places of known affordances and understood meanings. Additionally, the proportional alignment of each thematic category and study area is assessed using the individual data source and across multiple sources and locations.

4.4.1 Assessing Spatial Alignment

In order to assess the degree to which Twitter content originating from various locations reflects the characteristics of these locations, I compare spatial clusters of thematic content in Twitter data to the corresponding physical neighbourhoods and their thematic characteristics as reflected in Wikipedia. Figure 10 shows the resulting high value clusters for two categories, namely entertainment (red) and recreation (blue) in a subset of Manhattan, NYC.

In this Figure I show the spatial distribution in finer resolution in order to better communicate the level of alignment between Twitter clusters and corresponding physical places: green clusters align at, or near, known areas of recreational affordances such as parks and squares, with the most notable cluster spatially aligned with Central Park. Similarly, the entertainment clusters align with established local entertainment hubs such as Times Square, Madison Square Garden, and the most notable cluster being at Broadway in the Theatre Sub District, just southwest of Central Park.

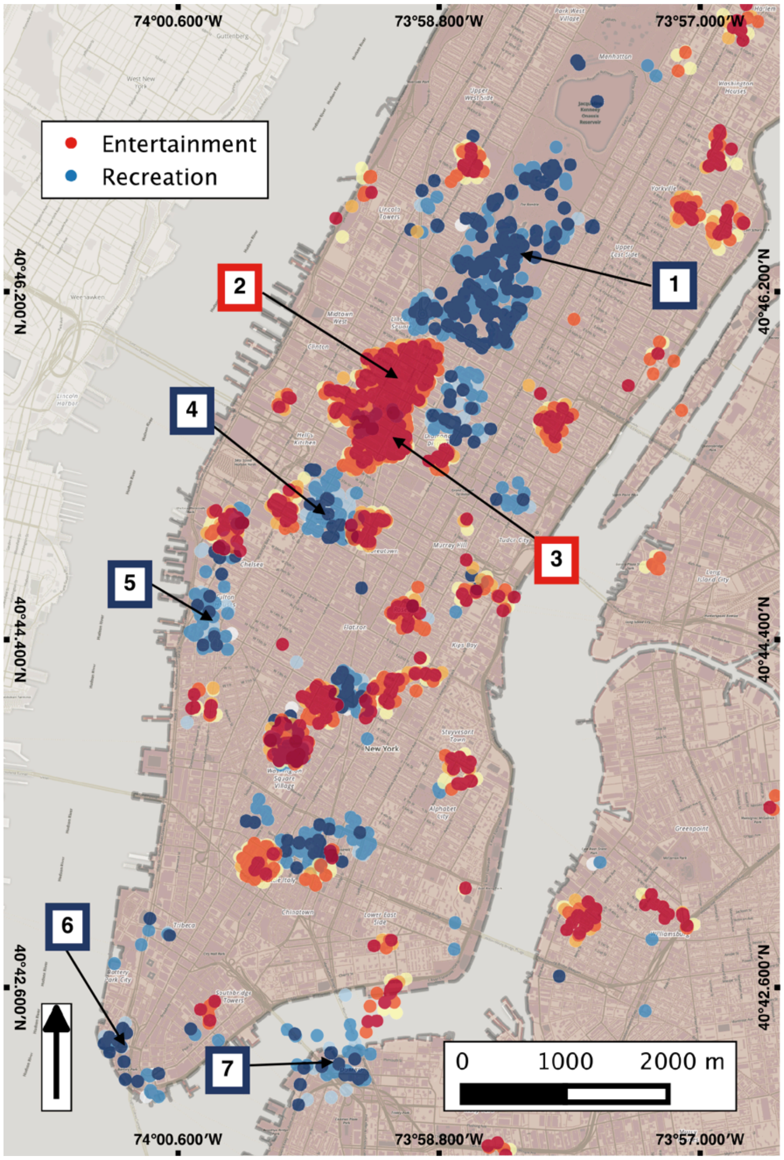


Figure 10 Statistically significant clusters of recreation and entertainment categories concentrated over Manhattan, NYC. For reference, the following sample places are labelled in the map with colour coded borders based on the legend features: (1) Central Park, (2) Broadway, (3) Times Square, (4) Madison Square Garden, (5) High Line Park, (6) Battery Park, and (7) Brooklyn Bridge Park.

Once these clusters are aggregated locally I have an even clearer view of this thematic alignment. These clusters are visualized in Figure 11, showing that the recreational clusters of Twitter content are overlapping not just with Central Park, but also with multiple parks throughout the city, including Battery Park, Brooklyn Bridge

Park, High Line Park, etc. This supports the argument that at this scale of analysis Twitter content is reflective of the local platial characteristics. In the same figure, the pie chart on its right-hand side shows the relative portion of Twitter content that is considered as *recreation* within our Twitter data corpus for this area (namely 13.7%, second to the dominant *entertainment* category).

Of course, as we see in Figure 11, I also receive few false positives (e.g. spots that are not aligned with parks in that figure). This is due to two reasons. First, there is some inherent fuzziness in the use of various terms, which may have multiple meanings in various contexts, yet are assigned their primary label in our clustering. For example, I often encounter the term *read* in tweets originating in the vicinity of airports, leading to the detection of local educational hotspots (as the term is primarily considered indicative of educational activities), even though in that case it is simply used by passengers who read books or periodicals while awaiting their plane. Second, the social media data may often indicate need rather than support for a certain type of activity. A small local square for example, may be serving as an informal recreational spot, even though formally it may not be recognized as such**.**

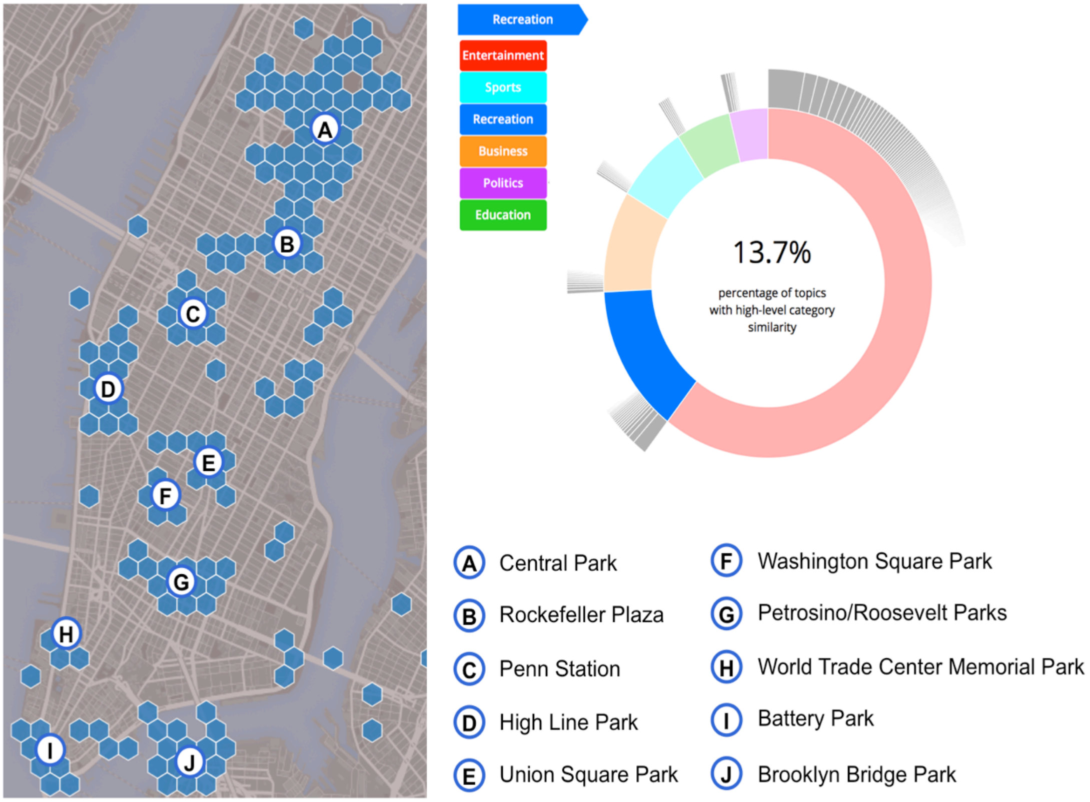


Figure 11 Spatially significant high value recreation clusters over NYC. The hexagon map depicts the statistically significant spatial clusters by point count coloured in blue. The results show a significant spatial alignment with prominent parks in NYC. The pie chart (right) shows that 13.7% of the total topics discovered from the entire NYC Twitter resulted in the highest semantic similarity with recreation out of the other five categories. Large clusters are annotated with location names to show spatial alignment.

While the above results show the overall alignment of thematic hotspots, I also compared the extent to which the crowdsourced spatial footprint of a thematic place aligns with its formally-defined extent. In Figure 12 I show the spatial extent of entertainment Twitter clusters (red dots) at the neighbourhood level for the Theatre Sub District, NYC. Beyond the spatial alignment, this Fig also illustrates how a place affordance extends beyond the formal boundary of that place due to human activity and/or expression: as the public moves to or from the Theatre District, it remains *immersed* in the thematic character of that place. As a result of this process, I have a collective reconfiguration of the platial boundary, shifting it westwards of its formal outline. Thus, social media content can capture places not just as the aggregate of their form and functions (Crooks et al. 2014), but also by revealing the transitional immersion into the themes that these functions enable.

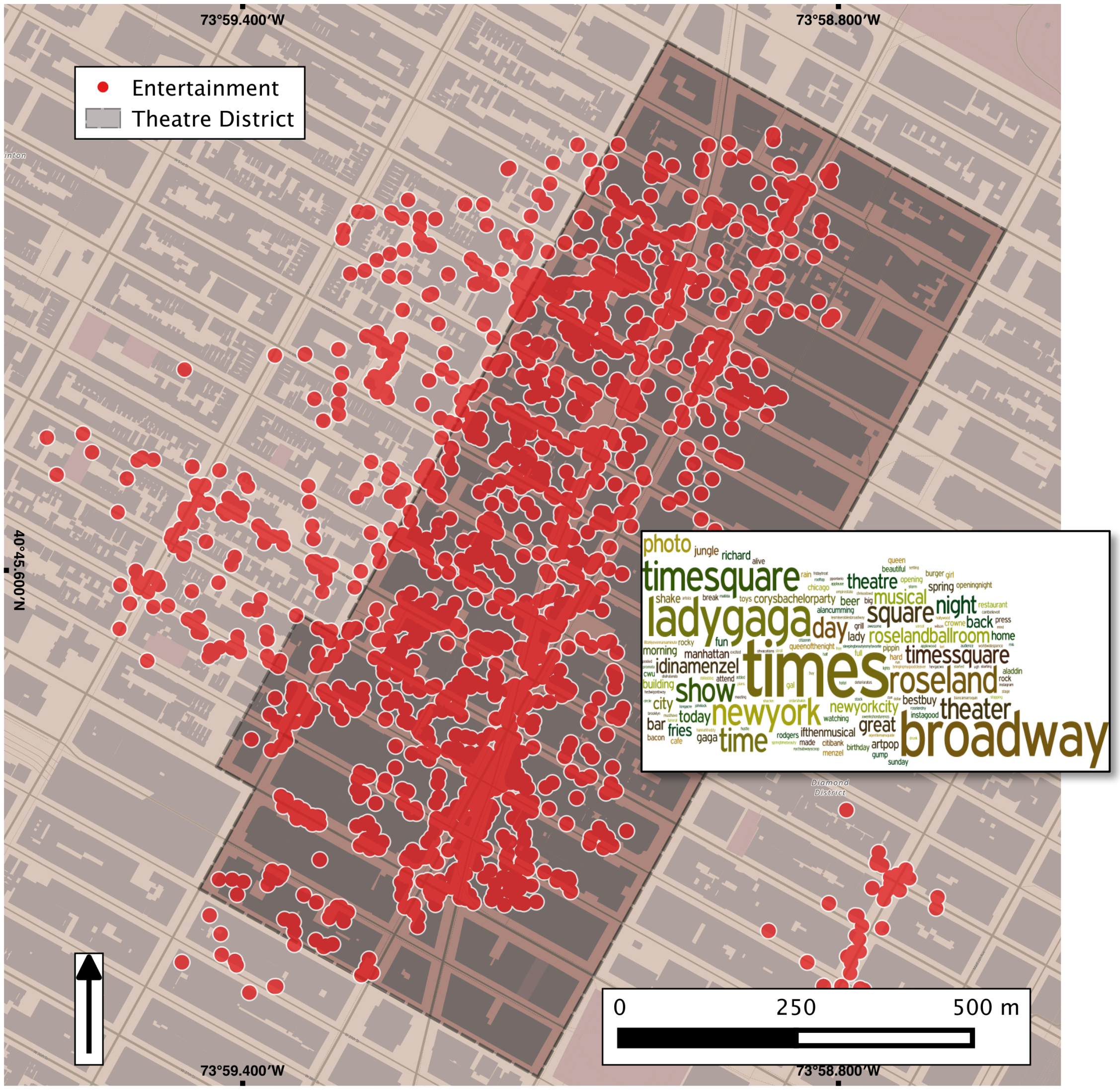


Figure 12 Depiction of spatially significant clusters and the Theatre Sub District, NYC. The spatial concentration of entertainment tweet clusters (red dots) contained within, and nearby, the Theatre Sub District (dark brown polygon with hash outline) show the alignment of platial expressions with local affordances (i.e. theatres around Broadway). I observe the dynamics of platial expressions with spaces adjacent to the official boundary being reconstituted with entertainment expressions. Additionally, the word cloud is shown to depict the topical terms semantically linked to entertainment.

I ascertained similar spatial alignment in SG, LA, and LDN for business, education, and sports categories using the aforementioned methods. Fig 13 provides overview maps containing the statistically significant clusters per category. Here I briefly highlight some specific physical features in each area that have common affordances of either shared activities or experiences, to further demonstrate spatial alignment. In SG sports clusters were detected in the western portion of the city near Jurong West Stadium and the Golazo Futsal Singapore.

Additional alignment was observed in the central part of the city near Clementi Stadium as well as to the north near Yishun Stadium. Majority of the education clusters were found near schools and universities as expected with notable locations as the United World College of South East Asia, Nanyang Polytechnic, and Temasek Polytechnic. Business clusters were detected in the Yishun district primarily in the central business district. Additional clusters are observed in the southern portion of the city near the Orchard Road location, which is a popular shopping area according to Wikipedia, and a large tourist attraction.

For LDN, I found overlapping thematic clusters of business, politics, and entertainment concentrated in the West End of London area. The location contains numerous government buildings, businesses, and theatres. I observed sports clusters near both Emirates and Olympic Stadiums, and also near restaurants and sports pubs. For recreation, our results indicated strong spatial alignment with parks and mostly notably LDN's largest, Richmond Park. In LA, significant clusters in the business category were observed near outdoor and indoor malls, shopping centres, and restaurants. However, I must note that I did not find a significant business cluster in LA's Financial District. Comparing our results against geographic data from the LA city government, I observed clustering in the education category near such example locations as the University of

Southern California, California State University Los Angeles, and California Lutheran University.

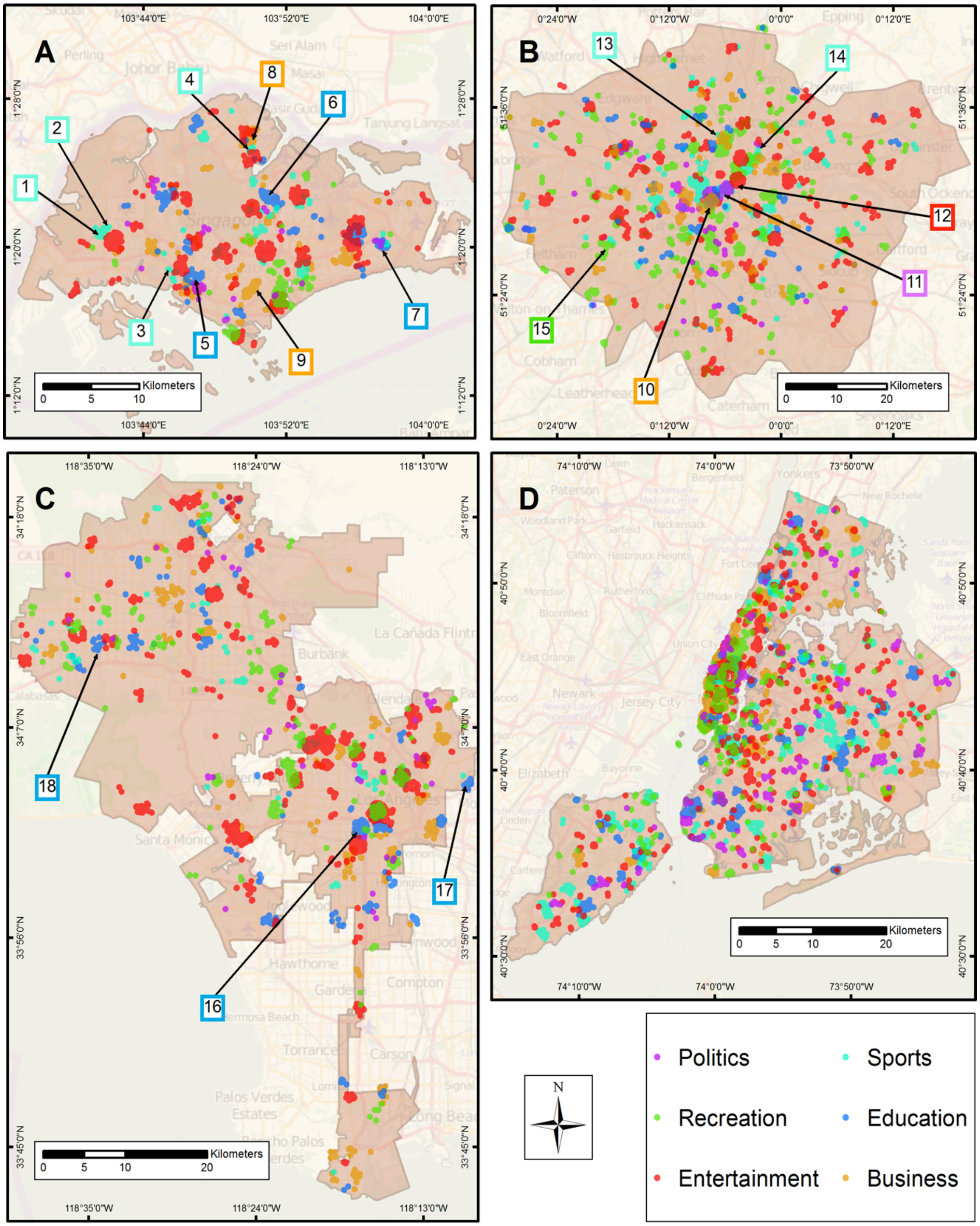


Figure 13 Maps depict significant hotspots for each of the high-level categories for (A) Singapore, (B) London, (C) Los Angeles, and (D) New York City. Additionally, the maps are annotated with sample locations were the hotspot aligns with a place with common meaning such as a school or stadium. The outlines of the numbered boxes correspond to the map legend colours. In map (A) Singapore, 1) Jurong West Stadium 2) Golazo Futsal Singapore 3) Clementi Stadium 4) Yishun Stadium 5) United World College of South East Asia 6) Nanyang Polytechnic 7) Temasek Polytechnic 8) Yishun District and 9) Orchard Road. In map (B) London, 10-12) West End of London 13) Emirates Stadium 14) Olympic Stadium and 15) Richmond Park. In map (C) Los Angeles, 16) University of Southern California 17) California State University Los Angeles and 18) California Lutheran University.

To further investigate the relationship between thematic hotspots and corresponding locations of interest, I analyzed the Euclidean distance between thematic hotspots relating to education and sports and corresponding facilities in LA using average nearest neighbour and proximity analysis. I first used average nearest neighbour analysis to understand the spatial distribution of facilities and then calculated the proximity of thematic hotspot points to these locations.

To briefly describe average nearest neighbour analysis, the process calculates the minimum distance between a point and its closest neighbour, and repeats the process for all points to derive a mean distance that is compared to an expected mean random nearest neighbour distance (Clark and Evans, 1954). This process produces the average nearest neighbour distance and gives a measure as to whether the points are spatially clustered or dispersed. I then subsequently looked at the average proximity of tweets to said facilities to get a relative indication of the distance between the two sources.

In doing so, I focused on LA due to the accessibility of a GIS database, and chose education and sports because these are themes that have facilities with understood meanings. I first calculated an average nearest neighbour distance of 1298.6m for official sports facilities using LA city government data (Cityplanning.lacity.org), and a nearest neighbour index of 0.43 and z-score of -16.67. With an index value less than 1 and a significant z-score, this result indicates the spatial distribution of sports facilities within the city are clustered. I then calculated the average proximity distance between all sports hotspot points to their nearest sports facility. I found that distance to have a mean value of 158.6 m, which is less then the average nearest neighbour distance and indicates an overall closeness in proximity between sports related tweets and facilities such as stadiums and athletic fields.

This also suggests that tweets conveying an athletic theme tend to originate in the vicinity of such facilities, conveying the platial nature of such locations. Accordingly, the mean distance of 158.6 m can be viewed as a measure of the extent of this particular theme's immersion potential. Similarly, I calculated an average nearest neighbour distance among all schools, colleges, and universities in LA and found it to be 956.7 m with a nearest neighbour index of 0.75 and z-score of -10.13.

As before, this result indicates the educational facilities are also spatially clustered within the city boundary. The corresponding average proximity between education-related tweets and the nearest educational facility had a mean value of 406.1 m. This pattern is consistent with the above-made observations regarding sports facilities and their relation to thematically-focused Twitter content.

4.4.2 Assessing Thematic Alignment

Above I discussed the spatial alignment of Twitter thematic clusters with corresponding facilities and neighbourhoods to show how thematic hotspots tend to coincide with relevant physical locations and their characteristics. I now turn to assessing the thematic alignment among the uncurated content of individual tweets and the consensus expressing crowd-curated Wikipedia platial content. In that sense, each neighbourhood can be seen as having a sociocultural signature, expressing the ensemble of themes that assign it a particular character.

Our objective is to compare such signatures as they emerge in Twitter traffic to the corresponding content as it is harvested from Wikipedia. I do so by studying the various percentages of different thematic categories as they are used to describe various neighbourhoods in Wikipedia and compare them to the corresponding percentages of Twitter traffic originating from these locations. In order to pursue this objective I present our results first at the neighbourhood level, and then aggregating up to the city level.

In order to highlight platial themes as they emerge at the *neighbourhood* level of analysis, I present results from NYC's Theatre District, Central Park, and the Lower Manhattan Financial District in Figure 14. Each figure is a composite view of the proportions from Twitter and both Wikipedia topics and semantic page accesses by title. The outer pie chart on each graphic represents Wikipedia topics, the middle shows Twitter topics, and the inner depicts Wikipedia page accesses. In the Figure I have summarized the degree to which various themes express the particular nature of each neighbourhood, capturing for example the more recreational nature of Central Park and the more entertainment-oriented character of the Theatre District.

For the pair-wise comparison of such data I use the Renkonen index of similarity, whereby 100% similarity would indicate perfect thematic alignment, whereas 0% would reflect no alignment. Using this metric, the average similarity among the Twitter signatures (as shown in Figure 14) of the three locations was found to be 7%. The three locations have significantly low proportional alignment and thus indicate a uniqueness of each location amongst collective Twitter content.

This is visually evident in Figure 14 with the Theatre Sub District comprised almost entirely of entertainment related content at 88%. The Theatre District contains numerous entertainment establishments, to include Broadway. Conversely, and as one would expect, 99% of the statistically significant content in Central Park is recreation related. As Central Park is one of the most well known parks in one of the largest cities in the world. For Lower Manhattan, I find business as the largest proportion at 78%, which is home to Wall Street and the most powerful financial district in the world. Ultimately, these findings highlight the differences in human experiences and activities at each location, and more importantly that unique expressions of place emerge from Twitter through uncurated processes.

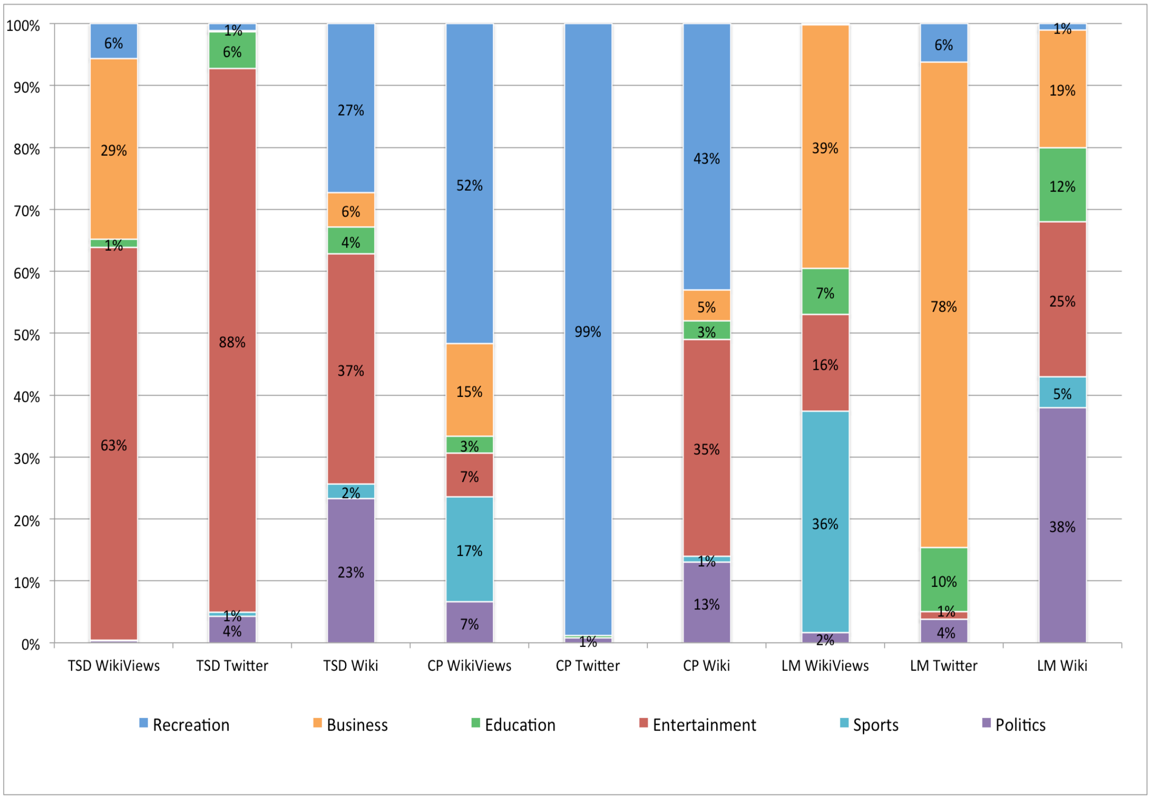


Figure 14 Percentages for each category, data source, and location in New York City neighbourhoods. The bar chart labels are presented in <Location> <Data Source> form. The locations are Theatre Sub District (TSD), Central Park (CP), Lower Manhattan District (LM). The data sources are Wikipedia Article Topics (Wiki), Spatial Twitter Topics (Twitter), and Wikipedia Semantic Accesses (WikiViews).

I assessed the single Wikipedia articles for each location (Figure 14). The average overall Renkonen similarity of this content for these three locations is 64%, which shows that single Wikipedia articles have higher proportional alignment and less uniqueness of place compared to Twitter. Central Park and the Theatre District resulted in the highest pairwise similarity of 87% compared to Lower Manhattan similarities of 49% (Central Park) and 57% (Theatre District).

Considering the nature of Wikipedia entries as a reflection of crowd consensus, these similarity values can be viewed as an indicative reflection of the thematic variability among these locations. Regarding individual neighbourhoods, in the Theatre Sub District, entertainment is clearly dominant (37%), followed by recreation (27%). For Central Park, I found near the largest division between recreation (43%) and recreation (35%). In the Lower Manhattan District, home of City Hall and the financial district, proportions differ significantly from the other two locations with politics leading (38%), followed by entertainment (25%), and business (19%).

Additionally, I evaluated the Wikipedia titles with page accesses using Renkonen similarity from the percentages in Figure 14 and assessed the average overall similarity for these three locations at 40%. Thus indicating less proportional alignment and more uniqueness of place between the neighbourhoods. Although the uniqueness is not as distinct as Twitter at 7%, I do see a reduction in alignment compared to the individual Wikipedia articles taken from the same data source. Looking at the most prominent proportions in each neighbourhood from Figure 14; the Theatre District is 64% entertainment, Central Park 51% recreation, and Lower Manhattan 39% business followed by 35% sports.

I compared the thematic content of Wikipedia (single articles and geo-located semantic access) and Twitter per location to assess proportional alignment by looking at Figure 15. Our findings show that for the Theatre District the highest similarity between Twitter and the geo-located article access at 66% followed by the single article and Twitter at 57%. The neighbourhood displays more pronounced particularities, namely its entertainment orientation. For Central Park, we have very high thematic alignment, compared to the other areas, between Twitter and Wikipedia (both single and geo-located page views) content (53% and 39% respectively).

Conversely, in the Lower Manhattan District, which is home to diverse affordances and therefore more multifaceted, the thematic alignment between Twitter and the neighbourhood single Wikipedia article was lower (36%). The comparison Twitter and geo-located page access resulted in higher alignment at 50%. Interestingly, our findings show the highest proportional alignment between two different neighbourhoods, Central Park and Theatre District, but using the single Wikipedia article for each area. This results points to homogeneity of Wikipedia, referring to its content and afforded activities, but also that these two neighbourhoods are spatially joined. Meaning the descriptions and expressions of place within Wikipedia for these two places share common semantics and lineages.

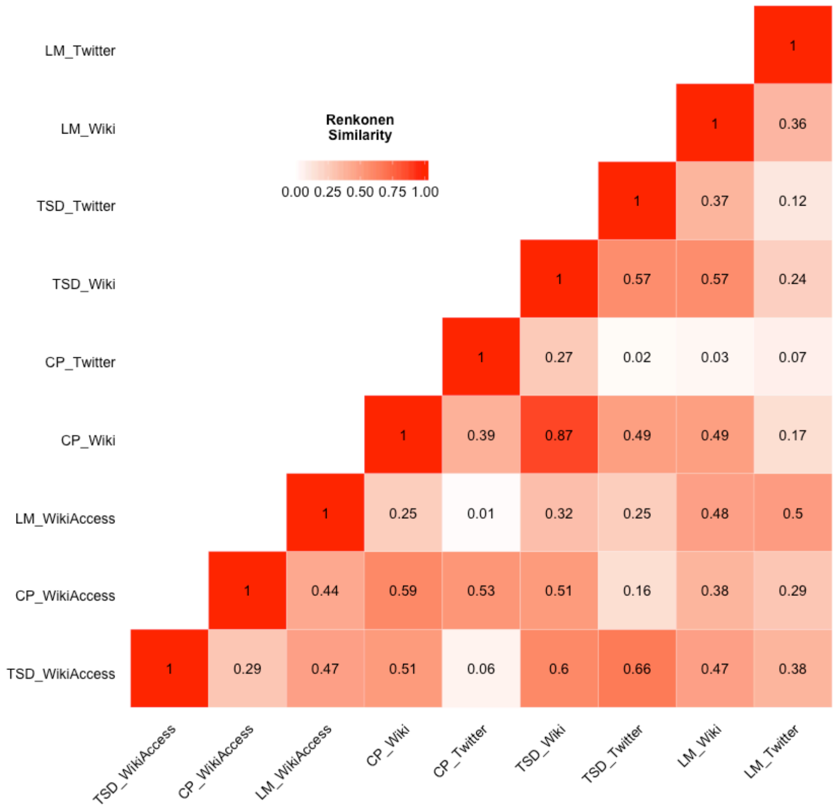


Figure 15 Renkonen similarity matrix for all three locations and data sources. The matrix schema shows a gradient between 1 (red) and 0 (white). The neighbourhoods are abbreviated as follows: Theatre District (TSD), Central Park (CP), and Lower Manhattan (LM). The data sources are differentiated as Twitter, single Wikipedia articles (Wiki), and geo-located Wikipedia article with semantic accesses (WikiAccess).

In Figure 16, I visualize the neighbourhoods using non-metric multidimensional scaling to better understand the proportional similarities. As previously discussed, the process of ordination uses the Renkonen dissimilarities derived from Figure 15. The distance between data points in Figure 16 is determined from the rank-ordering of dissimilarity values and distance values in ordination space for each pair of data points. From this I see two visual trends emerge from Figure 16. The first trend is the closeness of Twitter and Wikipedia semantic access proportions for each location, which indicates that at the neighbourhood scale a relative uniqueness of place, is revealed across these two different dynamic sources. Conversely, the second trend is the relative grouping of Wikipedia articles from different locations thus representing more homogeneity of the data source vice uniqueness of the places.

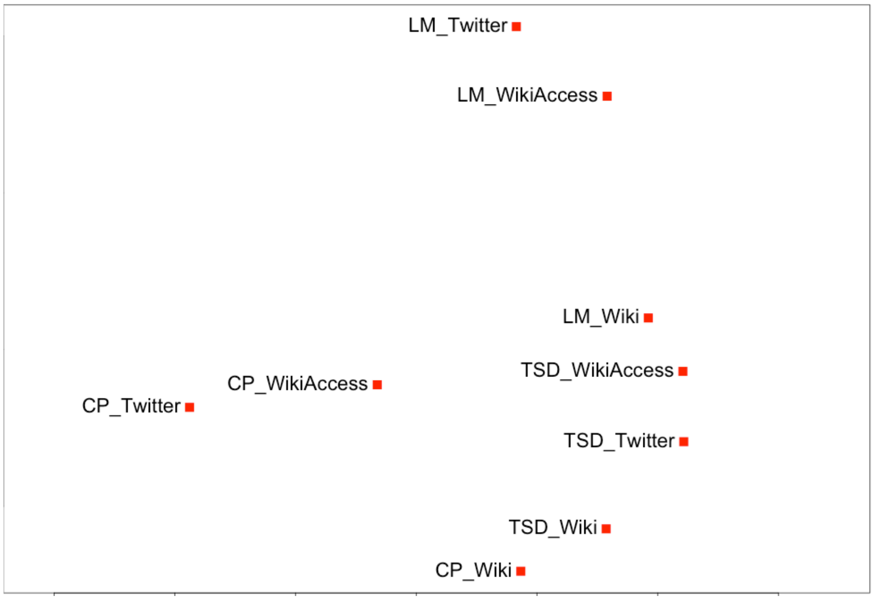


Figure 16 Non-metric multi-dimensional scaling of each neighbourhood and data source. The first and second dimensions of the ordination are represented with distances computed by rank-ordering between Renkonen dissimilarity values and distances in ordination space. The neighbourhoods are abbreviated as follows: Theatre District (TSD), Central Park (CP), and Lower Manhattan (LM). The data sources are differentiated as Twitter, single Wikipedia articles (Wiki), and geo-located Wikipedia article with semantic access (WikiAccess).

I performed the same analysis at the city level, aggregating all results within the city boundaries, for NYC, LA, SG, and LDN. Focusing on Twitter, percentage data in Figure 17 and Renkonen similarities in Figure 18, our results show similar proportional distributions across the four cities with an overall average similarity of 69%. When compared to Twitter at the neighbourhood scale this result indicates a loss of locational uniqueness. This further suggests that aggregating social media content at the city level of analysis has a smoothing effect on thematic content, removing the fine level platial characteristics that I observed at the neighbourhood level. Moreover, one could make the argument that at this level of analysis the platial character dissipates, and I observe Twitter at large rather than particulars.

Comparing the Wikipedia articles (Figures 17 and 18) produced an average similarity of 76% across the four cities. Our findings show that NYC and LA have the highest pairwise similarity (Figure 10) of 94% with entertainment as the dominant category. The results for Wikipedia semantic accesses are somewhat alike the Wikipedia articles with an average similarity of 80%, with NYC and LA having 93% proportional alignment. Like our neighbourhood findings, I see strong similarity between Wikipedia articles and semantic accesses for different areas, pointing to the homogeneity of the source as opposed to locational distinction or scale.

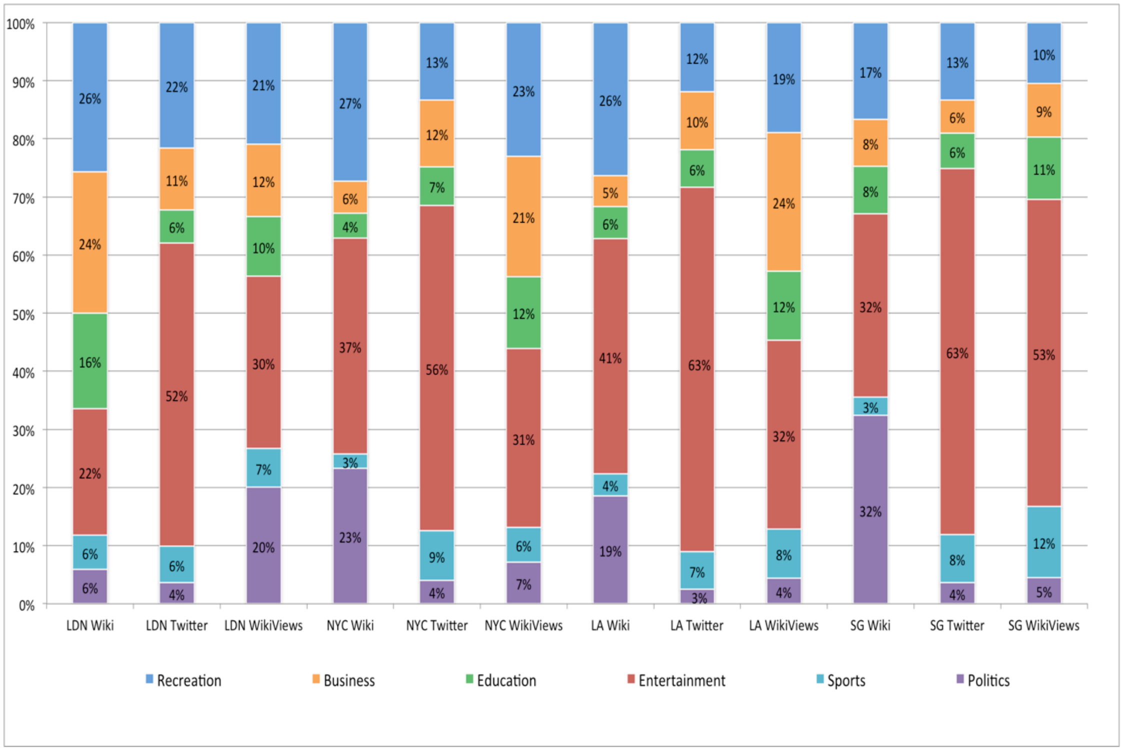
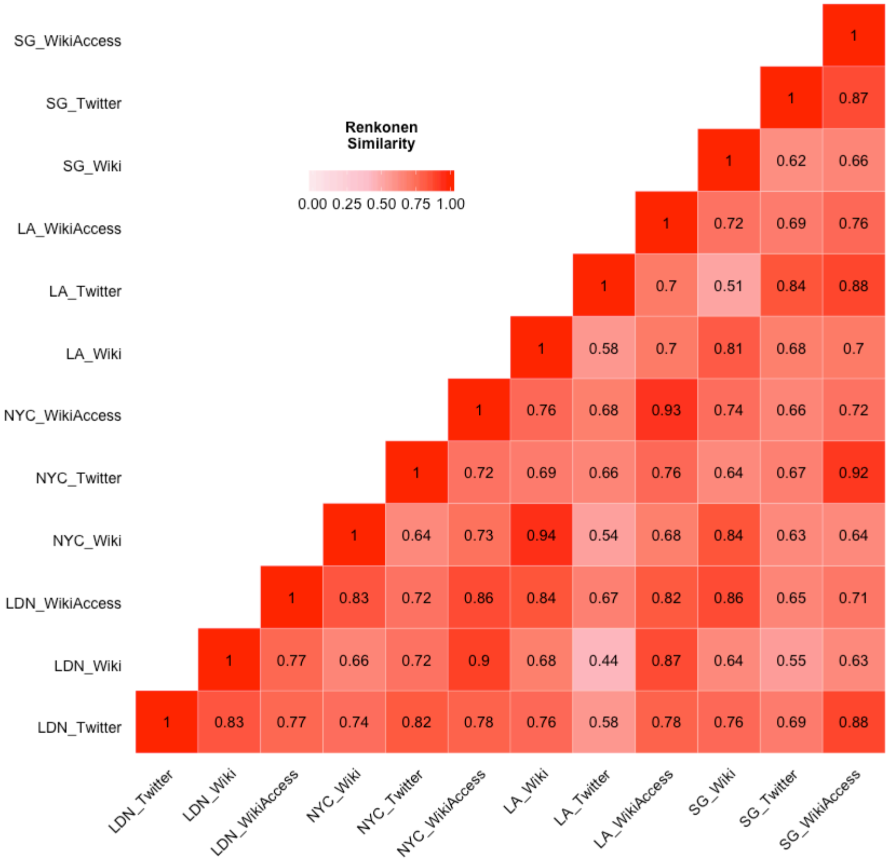


Figure 17 Percentages for each category and data source by city. The bar chart labels are presented in <Location> <Data Source> form. Locations include London (LDN), New York City (NYC), Los Angeles (LA), and Singapore (SG) and data sources are Wikipedia article topics (Wiki), spatial Twitter topics (Twitter), and geo-located Wikipedia semantic accesses (WikiViews).



	LDN_Twitter	LDN_Wiki	LDN_WikiAccess	NYC_Wiki	NYC_Twitter	NYC_WikiAccess	LA_Wiki	LA_Twitter	LA_WikiAccess	SG_Wiki	SG_Twitter	SG_WikiAccess
SG_WikiAccess												1
SG_Twitter											1	0.87
SG_Wiki										1	0.62	0.66
LA_WikiAccess									1	0.72	0.69	0.76
LA_Twitter								1	0.7	0.51	0.84	0.88
LA_Wiki							1	0.58	0.7	0.81	0.68	0.7
NYC_WikiAccess						1	0.76	0.68	0.93	0.74	0.66	0.72
NYC_Twitter					1	0.72	0.69	0.66	0.76	0.64	0.67	0.92
NYC_Wiki				1	0.64	0.73	0.94	0.54	0.68	0.84	0.63	0.64
LDN_WikiAccess			1	0.83	0.72	0.86	0.84	0.67	0.82	0.86	0.65	0.71
LDN_Wiki		1	0.77	0.66	0.72	0.9	0.68	0.44	0.87	0.64	0.55	0.63
LDN_Twitter	1	0.83	0.77	0.74	0.82	0.78	0.76	0.58	0.78	0.76	0.69	0.88

Figure 18 Renkonen similarity matrix for all three locations and data sources. The matrix schema shows a gradient between 1 (red) and 0 (white). The cities are abbreviated as follows: London (LDN), New York City (NYC), Los Angeles (LA), and Singapore (SG). The data sources are differentiated as Twitter, single Wikipedia articles (Wiki), and geo-located Wikipedia article with semantic accesses (WikiAccess).

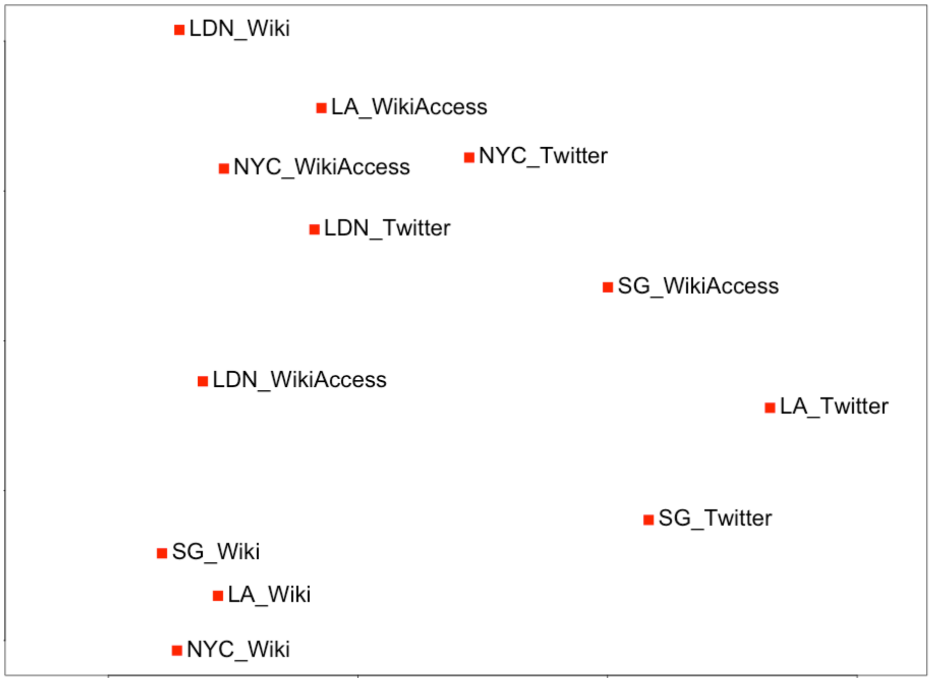


Figure 19 Non-metric multi-dimensional scaling of each city with data source. The first and second dimensions of the ordination are represented with distances computed using rank-ordering of Renkonen dissimilarities. The cities are abbreviated as follows: London (LDN), New York City (NYC), Los Angeles (LA), and Singapore (SG) The data sources are differentiated as Twitter, single Wikipedia articles (Wiki), and geo-located Wikipedia articles with semantic access (WikiAccess).

In Figure 19, I view the cities using non-metric multidimensional scaling as previously shown with the neighbourhoods to better understand the proportional similarities across the all data sources and cities. Once again, the Renkonen percentage similarities from Figure 18 are converted to dissimilarity values and visualized so that the data points are in rank-order agreement between distance and dissimilarity. As previously discussed, I explored various iteration values to minimize the stress or fit between distance and dissimilarity in the visualization. The apparent smoothing effect of scale and aggregation is further revealed at this level with the notable grouping of Wikipedia articles, minus LDN, and lack of clear groupings of Twitter and semantic accesses by location. The distances of dynamic content, Twitter and Wikipedia access, are tempered by the lack of a unique single category for any one location.

4.5 Summary

This chapter presented the application of the proposed framework for harvesting collective and crowd-centric platial content from Twitter and Wikipedia in the following study areas: New York City, Los Angeles, London, and Singapore. A subsequent study was performed for three different neighborhood areas in New York City in order to contrast the differences in scale and investigate aggregation of content for cross-source consensus. The results presented are based on semantic and spatial alignment of platial clusters to known points of interest that have common meanings (i.e. school). In particular, the high-level categories of recreation, entertainment, sports, and education demonstrated spatial alignment with locations with understood meaning such as parks, theatres, stadiums, and universities respectively.

A cross-source analysis was conducted between Twitter and Wikipedia to investigate the emergence of unique platial knowledge. The findings showed that the aggregation of content at different scales exhibits varying levels of uniqueness. It was shown that at the neighborhood level, in particular, the collective human activities in these places share common and semantically similar meanings. On the other hand, a collective sense of place or collective place meaning is not observed at the city scale because of the aggregation of dissimilar human activities.

CHAPTER 5: A FRAMEWORK FOR ASSESSING PLATIAL REOCCURRENCES AND CO-OCCURRENCES

5.1 Introduction

This Chapter presents an assessment of how well such mined platial themes exhibit spatiotemporal alignment by extending the work of Chapter 4. In the previous Chapter, Twitter and Wikipedia were investigated to discover socio-cultural signatures of place that emerge, and to what extent they align, from each crowd-contributed source from particular places. The work showed that place meaning is multiscalar, and that investigating places at the neighborhood scale conveyed greater uniqueness of place given a single month of data. However, collective places formed through social activities are highly dynamic and constantly change with events, weather, and temporal cycles. The focus in this Chapter is to better understand the spatio-temporal dynamics of these collective thematic activities through the reoccurrence and co-occurrence of such platial clusters.

Ultimately, place meaning is formed through the reoccurrence of shared experiences (by groups) through repeated activities at a certain location (Brandenburg and Carroll 1995; Williams and Stewart 1998). In the context of sense of place, looking at reoccurrences and co-occurrences allows for the emergence of what Granham et al. (2009) referred to as place dependences. Meaning, certain locales are predisposed to common physical or social affordances that allow individuals or groups to carry out specific activities. These places can have dual purposes such as a stadium for sports and entertainment venues that reoccur. Places also afford multiple activities that can happen at the same time such as parks for recreational biking or relaxing while reading a book, which conveys an implicit educational meaning for those escaping to the park. Nevertheless, the intent here is to investigate whether or not these activities (i.e. platial themes) exhibit a reoccurrence pattern over time for single or multiple thematic clusters and/or co-occurrence pattern whereby different thematic clusters intersect.

The investigation into reoccurrence and co-occurrence alignment is best described using a subset of Allen's (1983) temporal relations, specifically *precedes, preceded by, equals* and *meets* and the spatial predicate *intersects*. These predicates are appropriate given that the data is spatially clustered and, more specifically, temporally binned in such a way that only these types of relationships can exist. Similar approaches have been used in mining spatio-temporal co-occurrence patterns (STCOPS) to discover evolving regions (Pillai et al. 2012), co-location patterns (Shekhar and Huang, 2001), and spatiotemporal databases (Pfoser et al., 2011). Although the algorithms presented in the cited literature are beyond the scope of this research, they do offer considerable insights for robust methodologies that are useful in assessing reoccurrence and co-occurrence of thematic platial clusters.

A subset of the LA Twitter data presented in the previous Chapter is selected due to spatial alignment of high-level categories. Although spatial alignment was discovered in each of the cities presented in Chapter 3, the purpose here is to showcase another city in detail and LA accounted for the second largest number of clusters compared to NYC. However, the temporal period is extended to three months and reprocessed at both single month and weekly increments. The Twitter data is assembled concurrently for LA using the publicly available streaming API for the months of March (3.1M), April (4.4M), and May (3.5M) of 2014. The collection is once again filtered on tweets with geolocation and the English language. The tweets were further filtered spatially using the geographic boundaries from the LA city government website. Similar to the city and neighborhood comparison in Chapter 4, a finer temporal scale is additionally pursued to investigate the affect of dividing the LA data into 13-week intervals.

5.2 Discovering Platial Themes Over Time

Figure 20 provides an overview of the process similar to Figure 7, except step 4 is now a spatiotemporal evaluation of reoccurrence and co-occurrence alignment, unlike the spatial and thematic alignment investigated previously. To briefly describe the approach, the tweets were grouped into 24hr time periods to form documents using the entire three months of data and processed using topical n-gram modeling to extract bi-grams (see Algorithm 1). The extracted terms were then thematically labeled using the entire Wikipedia knowledge base and PMI that were finally stored in a relational database.

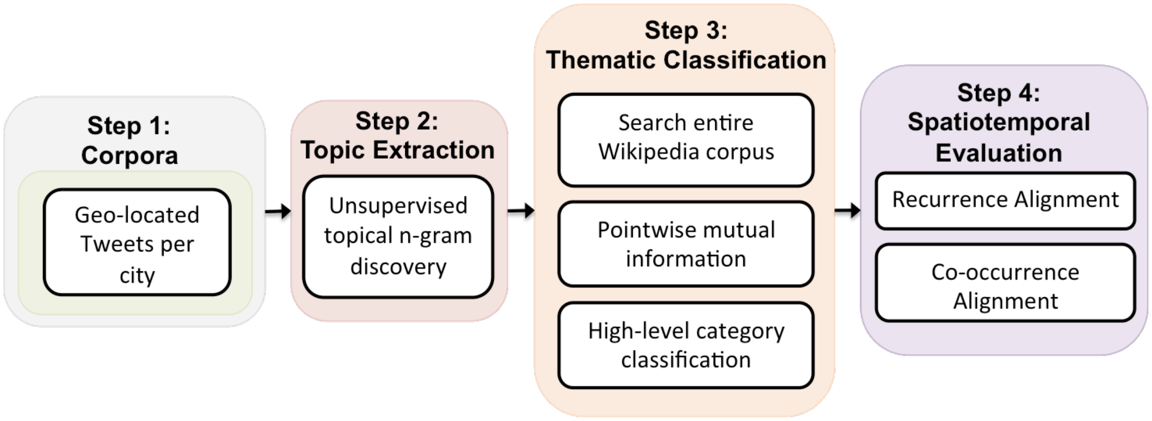

Step 1:
Corpora
Geo-located Tweets per city
Step 2:
Topic Extraction
Unsupervised topical n-gram discovery
Step 3:
Thematic Classification
Search entire Wikipedia corpus
Pointwise mutual information
High-level category classification
Step 4:
Spatiotemporal Evaluation
Recurrence Alignment
Co-occurrence Alignment

Figure 20 Flowchart describing overall process used to discover platial alignment. Step 1: sequential ingestion of Twitter, and Wikipedia article representing each spatial location, and all geo-located Wikipedia article titles spatially and semantic access statistics contained within the same location. Step 2: unsupervised topic discovery using topical *n*-grams. Step 3: determine the semantic overlap of topical *n*-grams and defined categories using Wikipedia search counts and PMI. Step 4: determine the statistical significance of each platial category and compare recurrence and co-occurrence alignment over three months for three cities and 13 weeks for one city.

A subset of the original six high-level categories used in Chapter 4, namely sports, recreation, and education, were chosen here due to the complexity and data volume incurred with additional temporal bins. Another factor for the selection of categories was the degree to which they spatially aligned with places of known affordances as seen in Chapter 4. Using places of known affordances, such as stadiums, schools, and parks offer a means of asserted ground truth for evaluation.

Given the thematically labeled tweets, the data is temporally divided into three month and 13 week intervals. Figure 22 shows the number of thematically labeled tweets per category for each of the months. Interestingly, the total collected number of tweets for April was 4.4M, almost a 1M higher than May and 1.5M over March, yet this time period has the least amount of labeled tweets. Conversely, the lowest number of tweets was collected in March, which has the highest thematic counts. It is perceived, although not known, that more events occurred in this month since Twitter is highly event-based (Croitoru et al., 2013).

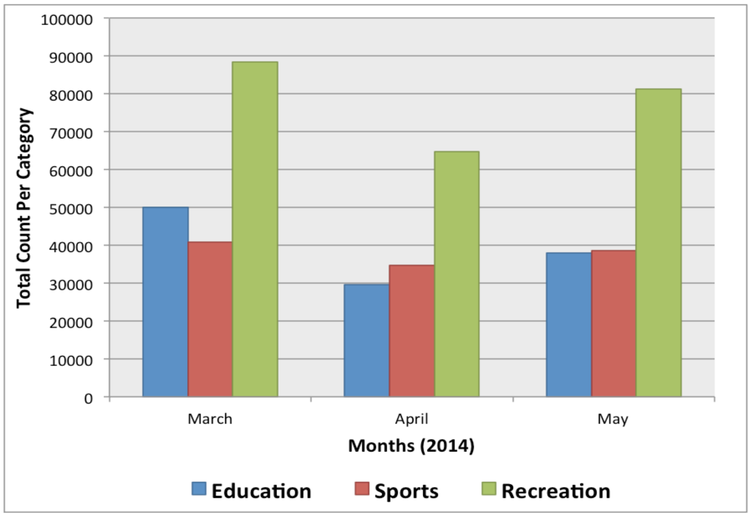


Figure 21 Bar chart depicts the total thematically labelled counts for LA by month.

Similarly, the per category counts are shown in Figure 23 for each of the 13 weeks and unsurprisingly follow downward trend in April. Another observation from both Figure 22 and 23 is the proportionally high number of thematically labelled recreational tweets. As seen in Chapter 4, with respect to the large portion of entertainment tweets, the content in Twitter is uncurated and unfortunately biased towards certain topics of discussion. With this in mind, content analysis studies have been performed to assess the degree of bias in social media, particularly Twitter, and found that majority of the content centers on personal narratives of lifestyle and entertainment topics (Zhao et al. 2011). Notions of place within social media are more of a by-product of humans discussing their experiences at a particular location, which may or may not, relate to their surrounding environment.

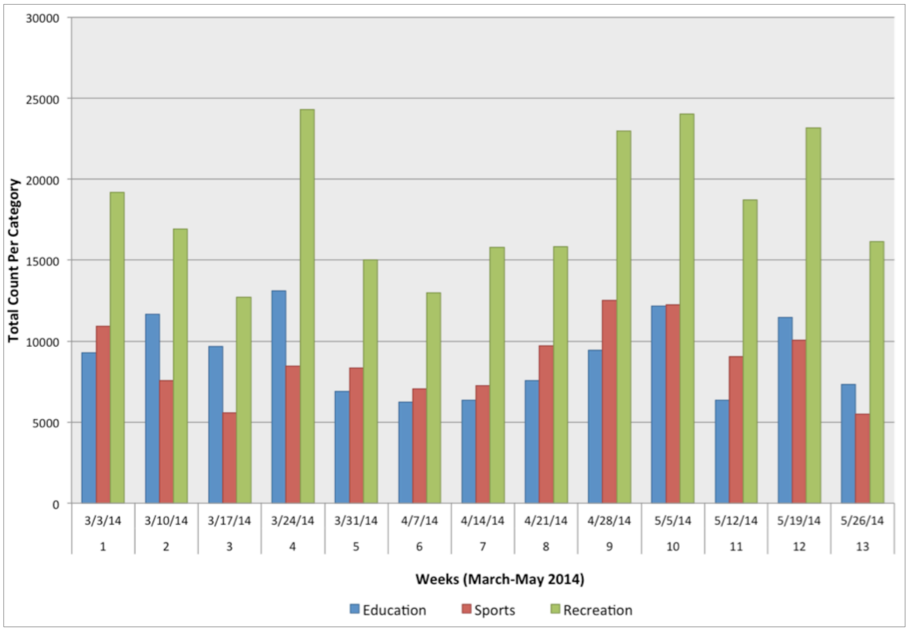


Figure 22 Bar chart depicts the total thematically labelled counts for LA for 13 weeks.

Using the labelled tweets per category, the next step is to find an appropriate analysis distance for each category and temporal bin (months and weeks). This is accomplished by incrementally checking for spatial statistical significance at incremented distances using Global Moran's I (Páez and Scott, 2005) to test for spatial autocorrelation and the distance with the highest significance value. The following distances were observed by iterating from 100m to 3000m using 50m increments for each month and by category in order of education, sports, and recreation respectively: March (840m, 723m, 976m), April (948m, 893m, 1052m), and May (786m, 834m, 872m).

The distances for the 13-weeks were additionally determined to find the max statistically significant z score. Figure 23 presents the distances plotted for the education category from 100m to 3000m (x-axis) for each of the 13-weeks (y-axis) and z scores (z-axis). Both sports and recreation were calculated using the same method (see Algorithm 3).

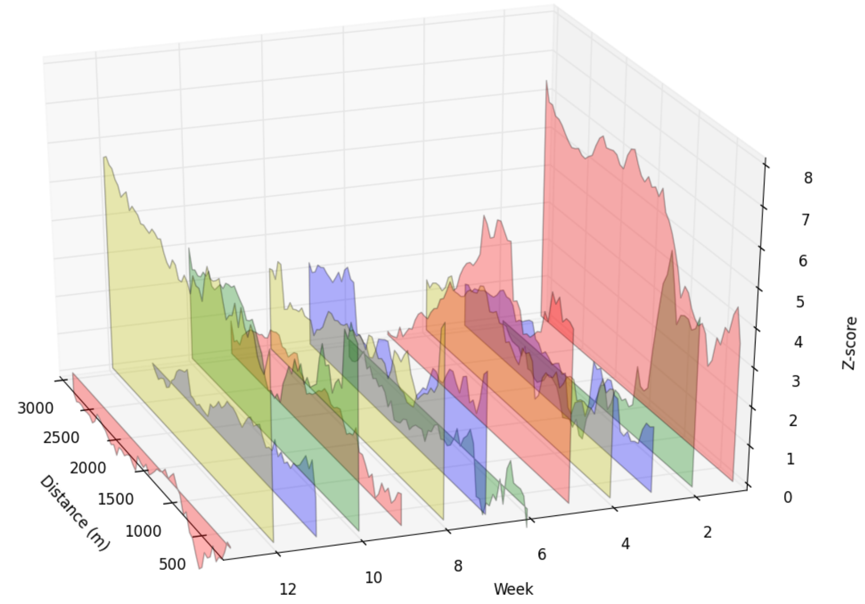


Figure 23 LA education incremental Global Moran's I for each of the 13 weeks.

Using the highest z score (see Algorithm 2), the distance was recorded for each of the time periods and parameterized as the fixed distance threshold in the local Getis-Ord Gi* (Getis and Ord, 1992) along with the PMI values for each category. The PMI values were applied as weights in computing Getis-Ord Gi* statistic to discover local spatial clusters with high semantic values within the boundaries of LA. The statistical significance at the $P < 0.05$ level and considered only z-scores greater than +1.96.

Figure 23 depicts week 13 below the +1.96 threshold, which either means the data was random or the data clustered outside of the min/max interval values. In either case, the highest z score was still used to determine the distance value, for which Getis-Ord Gi* would determine the final statistically significant clusters. It is understood that this is not the optimal approach, although several clusters were still discovered for week 13, future versions of Algorithm 2 will check for this condition and rerun at different intervals. As a feature with a high PMI similarity score is interesting, it may not be a statistically significant hot spot or a place with nearby features with similar place expressions. Moreover, a feature (i.e. single geotagged tweet) will have a high PMI similarity value and be surrounded by other features with high values as well.

The resulting spatial clusters from the 13-week periods per category were further clustered using DBSCAN and transformed into polygons by creating convex hulls. This step is performed because assessing the reoccurrence and co-occurrence of platial clusters as point data for 13-weeks is difficult to visually interpret. Although using a convex hull can misrepresent the area, majority of the clusters are small, given the time period, which reduces obvious region errors and ultimately place is vague and has indeterminate boundaries.

DBSCAN locates regions of high density that are separated from one another by regions of low-density using points within a specified radius (Ester et al., 1996). This is accomplished given the two parameters, *epsilon* and *minPoints*. The algorithm starts by picking a random point in the dataset and determining if the total number of points surrounding the chosen point, within the distance of *epsilon*, is greater than *minPoints*. DBSCAN is used because the resulting Getis-Ord Gi* clusters are not individually identified so that each can be turned into a polygon (see Algorithm 3). The parameter inputs for DBSCAN are *epilson* and *minPnts*. The same distance values used for Getis-Ord Gi* were assigned to epilson and the minPnts was heuristically set to three to minimally form triangles. The DBSCAN and convex hull process was not applied to the results from the monthly time period since an assessment using the just the point was feasible.

5.4 Evaluation and Results

To test the proposed approach for discovering the reoccurrence and co-occurrence of platial content in crowd-contributed content it is necessary to present a formal assessment of the results. Within this section, the spatiotemporal alignment of the resulting thematic clusters for each of the time periods is investigated in detail with places of known affordances and understood meanings. The section proceeds by assessing the data from LA on a monthly frequency followed by weekly.

5.4.1 Assessing Monthly Recurrences and Co-occurrences

In order to assess the degree to which Twitter content originating from various locations exhibits reoccurrence and/or co-occurrence patterns, I compare the spatial clusters by first looking at within theme reoccurrences followed by cross theme categories for co-occurrence. Since the within theme clusters are temporally binned into discrete one month increments, the following observations and assessment is made using the temporal relations of *meets* and *precedes*, or *preceded by* and spatial intersection. Additionally, prominent physical features such as stadiums, park, etc., that are overlapped by the spatial clusters are used as references.

Figure 24 shows the resulting high value clusters for each of the three categories, namely education (blue), recreation (green), and sports (red). The Figure is presented as an overview of clusters discovered at the city level for each of the three months. As noted earlier in Figure 21, the reduced number of thematically labelled tweets for the month of April obviously produced less spatial clusters. Although the collection of tweets for the months of March and May resulted in less totals compared to April, the number of observable unique clusters is greater for these months. From Figure 24 I also observe clustering in each of the categories in Downtown LA, which is explored next in greater detail for reoccurrences and co-occurrence.

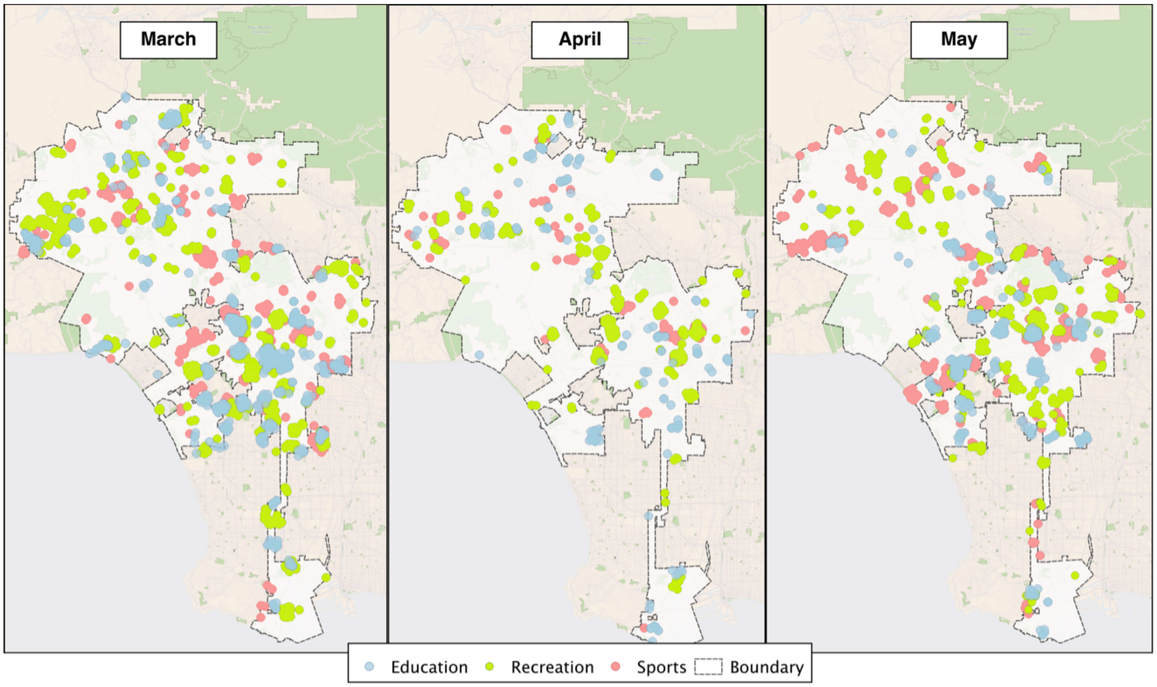


Figure 24 City level overview map showing thematic platial clusters for three different months.

Figure 25 provides a focused view of Downtown LA to better assess the reoccurrence and co-occurrences across each of the three months. I proceed by assessing the individual categories first for reoccurrence and then an aggregated view of all the categories for co-occurrences. Starting with the Downtown LA Sports map (top-left), I observe a reoccurrence of clusters (A) from March and May, but not April, located in residential areas within the vicinity of Koreatown and Mid City. A similar reoccurrence of clusters (C) appear once again between March and May, and to some degree April, near the numerous downtown restaurants and shopping areas. Surprisingly, reoccurrence is not observed for clusters B (Staples Center) and D (Dodger Stadium) appearing only in April, despite these places having known affordances for sports activities.

Looking at Downtown LA Recreation map (Figure 25 top-right), reoccurrence is once again observed in clusters (E) from March and May, but not April, located in residential areas within the vicinity of Koreatown and Mid City. A recurrence clusters (F) at the Staples Center are seen concurrently from March to April. Additionally, a single cluster (G) for the month of April emerged for recreational activities at Dodger Stadium, but did not include over months. Overall, smaller reoccurrence clusters that are not labelled were found in residential areas and indicate people discussing topics related to recreation possibly from watching TV or the Internet.

For the education results (Figure 25 bottom-left), reoccurrence clusters (H and I) appear for each of the three months at the annotated locations. The first cluster (H), is a residential neighbourhood consisting of several elementary schools, middle schools, and high schools all in close proximity, while the other cluster (I), is the University of Southern California (USC).

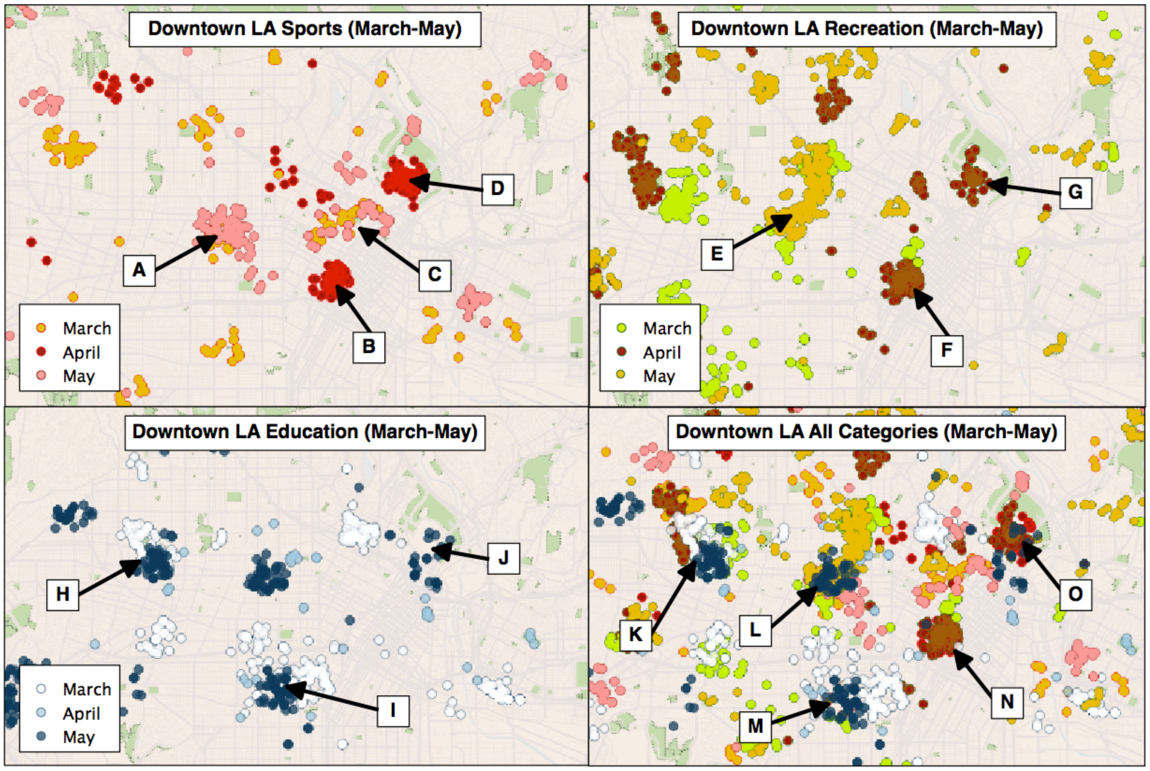

Downtown LA Sports (March-May)
A
B
C
D
March
April
May
Downtown LA Recreation (March-May)
E
F
G
March
April
May
Downtown LA Education (March-May)
H
I
J
March
April
May
Downtown LA All Categories (March-May)
K
L
M
N
O

Figure 25 Quad maps show clusters of sports, recreation, and education categories over Downtown LA for three consecutive months. The different maps are depicted as follows: Downtown LA Sports, Downtown LA Recreation, Downtown LA Education, and Downtown LA All Categories. A description of each map and letter callout is given within the body of text. Note: the legend for the Downtown LA All Categories map is not provided as it is consolidated map of the other three.

Given the reoccurrence assessment of the individual categories, I now turn to the co-occurrence or mixing of multiple categories that appear at the same location within the same month or over the entire three-month period. Looking the All Categories map (Figure 25 bottom-right), I observed the most prominent co-occurrences in clusters K and L emerging from the neighbourhood areas previously mentioned. The intersection of multiple thematic clusters over these areas indicate that people are discussing multiple topics, and suggests arguably greater personal uniqueness of place (meaning homes), whereby there is less of a crowd connection or attachment of place meaning, other than general homes, compared to public parks or stadiums. In this case, the proximity factor of people living closely within a neighbourhood, and in general, suggests similar experience of daily routines and activities causing semantically similar content to emerge from social places and not traditional physical places.

For the clusters M, overlapping USC, I observed co-occurrence between education (dominate category), recreation, and sports, which makes sense since USC is a major university with multiple sports teams and a large student body. Clusters N and O depict co-occurrence of sports, recreation, and limited education themes that appear within the month of May. Given the concentration of cluster in a relatively small area these are perceived as scheduled events taking place at the Staples Center and Dodger Stadium. The mixing of education clusters at location O either represents the repeated use of a bi-gram that has higher semantic similarity with education in Wikipedia or possibly ambiguous terms taken out of context and mislabelled.

5.4.2 Assessing Weekly Recurrences and Co-occurrences

I now turn to the assessment of recurrences and co-occurrence using a weekly interval. As previously discussed, the weekly clusters were further processed using DBSCAN and converted to convex hull polygons to more easily assess the intersection of 13 different time periods. This section provides a focused assessment of selected locations for the reoccurrences of individual thematic clusters and then proceeds with another assessment of Downtown LA. I start by looking at the map of Dodger Stadium in

Figure 26 for sports clusters. The physical location of the stadium is located in the center of the map, which also contains the highest concentration of intersecting and overlapping polygons.

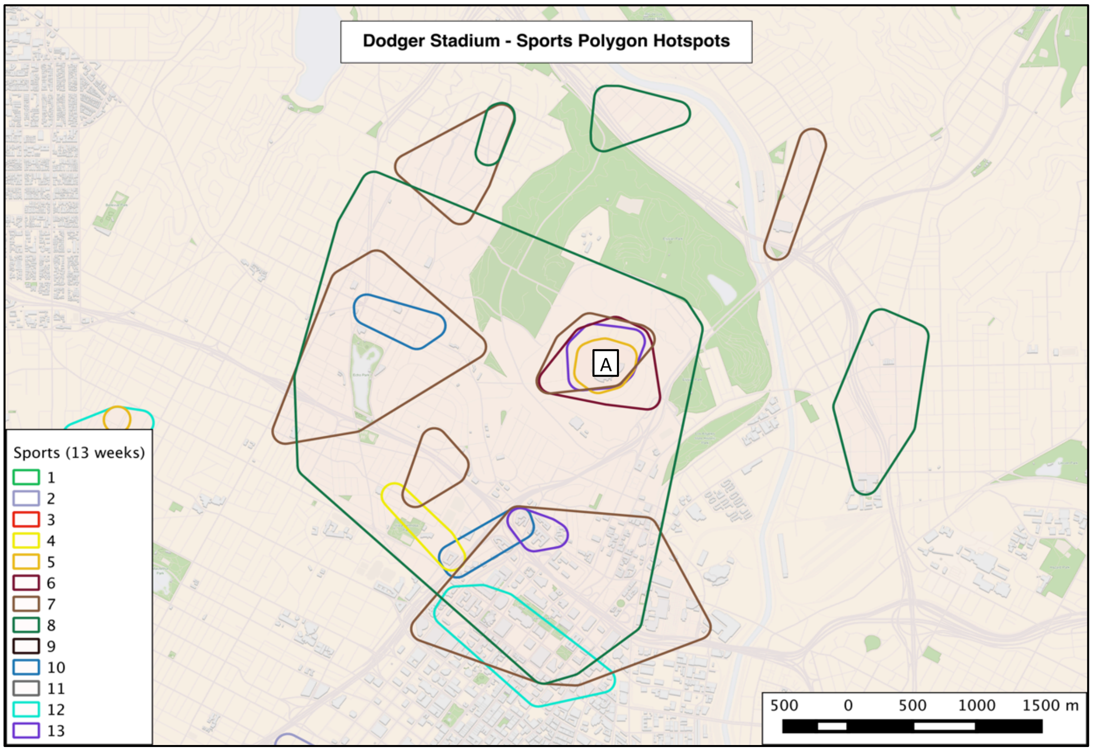


Figure 26 LA convex hulls representing sports hotspots near Dodger Stadium (A) using 13-week temporal bins.

Out of 13 weeks, in Figure 26, 5 different polygons reoccur over this location and ultimately correspond temporally with the major league baseball team's home and away schedule. The team played away games from the beginning of the season in March to the first part of April. Then they played several away games mid April and another set of home games at the end of April. This cycle continued for the month of May, which aligns with the reoccurrence of clusters seen in the map. The reoccurrence clusters observed away from the stadium align with both away and home games, indicating sports bars and restaurants that are frequented by fans regardless of where the game is actually played.

In the Figure 27 map, educational clusters are shown over USC. Compared to the dense reoccurrence clustering in the previous stadium map, Figure 26, the polygons here encompass larger areas suggesting the physical layout and form of the USC campus and nearby neighborhoods has an affect on the footprints. Of particular note here is the lack of clusters in the beginning of March when the school is in spring recess. The reoccurrence of clusters does not appear to follow cycle similar to that of the sporting event with regularly scheduled games. Of the 13 weeks, only 3 intersect the campus boundary, while the other near polygons could represent student housing or other nearby schools.

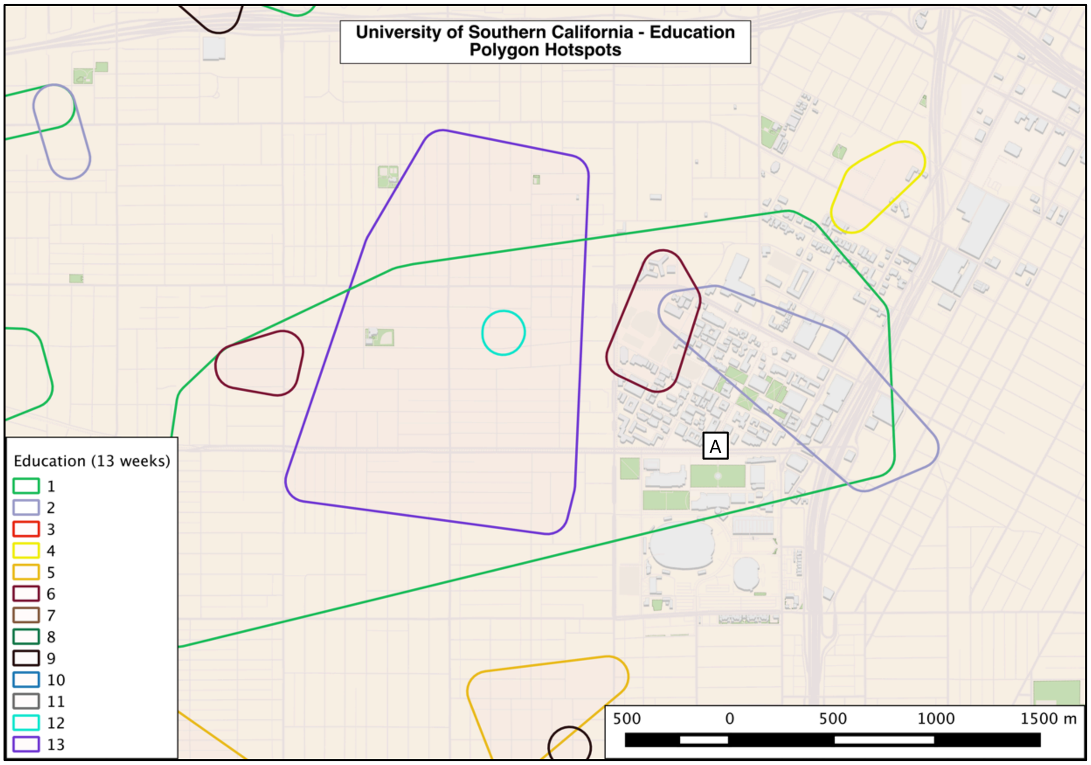


Figure 27 LA convex hulls representing education hotspots for University of Southern California (A) using 13-week temporal bins.

The map in Figure 28 depicts recreational clusters over Downtown LA with the following feature labeled: (A) Staples Center (B) downtown restaurants (C) Pershing Square. Surprisingly, only a single polygon is observed over the Staples Center for the 5th week. The highest number of intersection of cluster reoccurrences is seen overlapping the downtown restaurants (B). However, the reoccurrences are not continuous over the 13 weeks with only 5 different weeks that emerge at different times. Along with its numerous restaurants, multiple hotels and outdoor courtyards are located within this area to accommodate tourists visiting the city. Lastly, a three recreation clusters form over Pershing Square (C) representing weeks 2, 4, and 12. The park like square holds numerous outdoor art showings and venues.

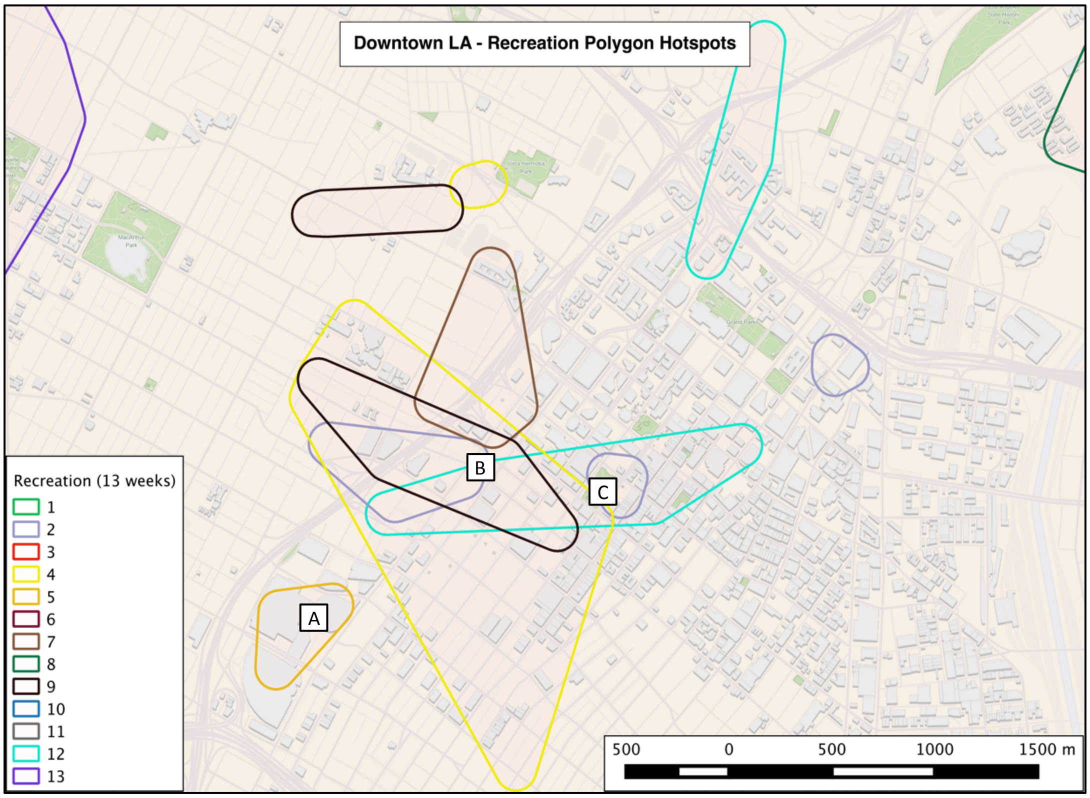


Figure 28 LA convex hulls representing recreation hotspots in Downtown LA using 13-week temporal bins; (A) LA Convention Center and Staples Center, (B) numerous restaurants, (C) Pershing Square. The map depicts locations of recurrence and the popular (A) LA Convention Center and Staples Center where only a single polygon emerged.

The map in Figure 29 provides a visual for assessing the combined co-occurrences for each of the three categories. In doing so, each map presents a single thematic category, but zoomed further out to include some of the surrounding areas. The repeated letters show selected locations were two or more categories intersect. Location A depicts the co-occurrence of both a sports and recreation polygons during week 4 near the downtown restaurants and hotels.

I observed that at location B, residential neighborhood, sports and recreation intersect at two different time periods, 2 and 5, over the same location. The co-occurrence of sports and recreation is once again discovered at location C during week 12. And lastly, an intersection during week 13 is observed for sports and education at the Dodger Stadium (C). The overall co-occurrence trend in Figure 29 is that of sports clusters, which appeared in each case being the most frequent. This co-occurrence trend is closely followed by recreational that coincided with sports, which is not overly surprising given the semantic similarity of the two categories and the frequency of their appearances together.

It's not difficult to imagine the relationship between these two categories as one could consider the watching of a sporting event to be a recreational activity. The appearance of sport and education was previously discussed as to the different meanings of words and their semantic interpretation without the full context of the message. Although only three categories were chosen for this assessment, they nevertheless revealed, to a degree, the semantic composition of places and their uniqueness that result in repeated emerges at particular locations.

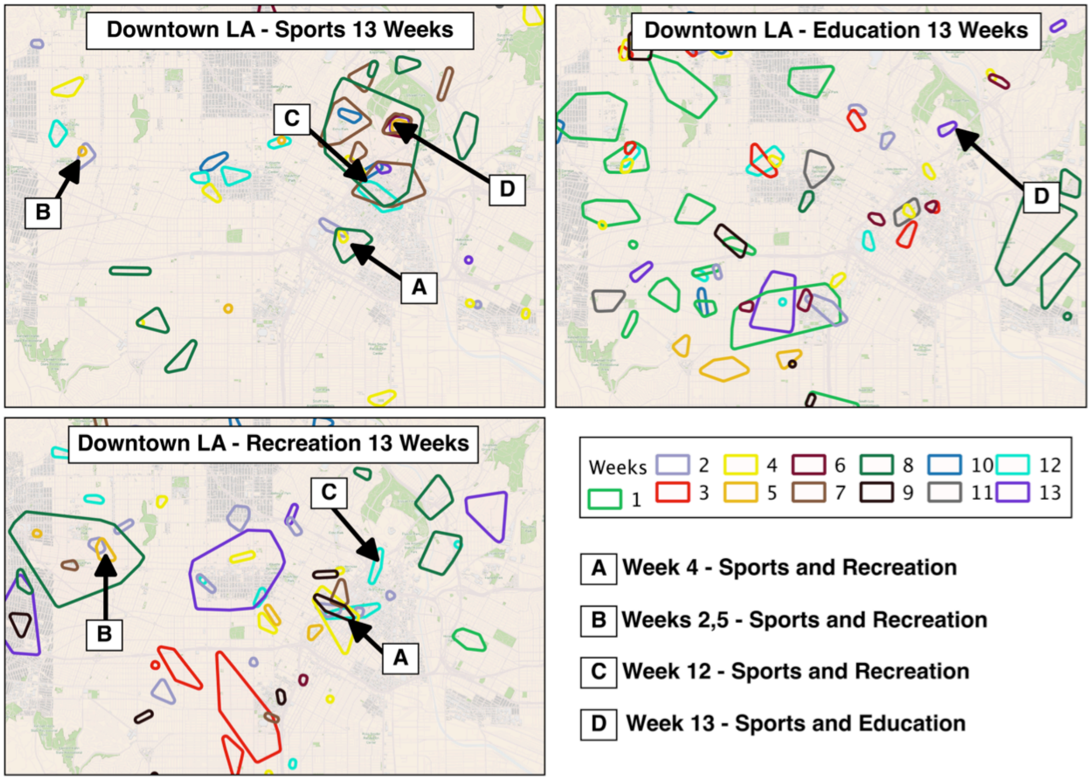

Downtown LA - Sports 13 Weeks
Downtown LA - Education 13 Weeks
Downtown LA - Recreation 13 Weeks
A
B
C
D
Weeks
1
2
3
4
5
6
7
8
9
10
11
12
13
A Week 4 - Sports and Recreation
B Weeks 2,5 - Sports and Recreation
C Week 12 - Sports and Recreation
D Week 13 - Sports and Education

Figure 29 LA convex hulls representing recreation, sports, and education hotspots in Downtown LA using 13-week temporal bins. The duplicate letters signify spatial and temporal overlap discovered within the maps.

CHAPTER 6: CONCLUSION AND FUTURE RESEARCH

6.1 Summary and Conclusion

It was hypothesized that taking a crowd-centric approach towards the mining of people's collective sense of place would lead to the discovery, extraction, and alignment of platial content. It was formally posited that mining user-generated data leads to the characterization and localization of places. To address the hypothesis, an algorithmic approach was developed and proposed that combined probabilistic topic modelling, semantic association, and spatial clustering to identify locations of collective sense of place. The proposed framework was applied to crowd-contributed content in the form of Twitter and Wikipedia to assess the semantic, spatial, and temporal alignment of the resulting platial clusters. The findings demonstrate how the meaning of place can be harvested through the analysis of crowd-generated content as thematic groupings.

In Chapter 4, a framework was presented to assess the spatial and thematic alignment of platial content produced the proposed approach. By contrasting such locations with the corresponding Wikipedia entries and semantic access, it demonstrated for the first time the thematic and spatial alignment between these two sources, supporting the argument that such content can be analyzed to reveal the shared meaning of place, as it emerges through human activities and perception. Of course, the approach also revealed few questionable clusters as to their semantic labelling and spatial alignment to places of unknown affordances. In these situations, it is difficult to fully assess whether the sensed platial content is truly characteristic of the physical place, as ground truth is hard to acquire and place meaning can be vague. For this reason, the assessments of alignment, whether semantic, spatial, or temporal, were conducted using locations of understood and general collective meanings. Moreover, the results demonstrated that the approach does address a significant portion of the signal-to-noise problem found in Twitter, whereby, people may be in a place physically, but not necessarily participating or contributing to the shared meaning of a place. The results indicate this to be the case in light of questionable clusters.

Another contributing factor of the demonstrated results was the selection of high-level categories that minimize semantic overlap without becoming too narrow, but still remain broad enough to relate many concepts in a meaningful way. The NYC results make this clear with the recreational and entertainment clusters that emerged with discernable spatial alignment with many parks and theatres. Moreover, the proportional alignment of semantic content between Twitter and Wikipedia demonstrated not only the uniqueness of places, specifically at the neighbourhood scale, but the importance of selecting high-level categories to differentiate platial content.

Ultimately, the scale of place has a significant impact on its discernibility within sources as shown in our findings of neighbourhoods and cities. At the neighbourhood scale the particularities of place emerge, whereas zooming out to the city scale reveals more of the medium such as Twitter and Wikipedia instead of a particular location as one would expect with aggregation. Undoubtedly, the aggregation of sources at varying scales of analysis will produce different results, which is why I separately investigated cities and neighbourhoods. This is also true of temporal scales demonstrated in Chapter 5. Whereby, the thematically labelled platial content was separated into monthly and weekly intervals that changed the neighbourhood densities of points to form different clusters. To deal with scale, the approach of incrementally checking for statistical significance at varying distances to determine the appropriate scale per location and thematic category. The results of apply incremental spatial autocorrelation at the weekly time period underscored the need for rigours distance analysis as the distances varied considerably. Thus highlighting the fact that Twitter and other crowd-contributed content are composed of events and other human activities at varying scales (stadium vs. park) all within a single data source.

Concerning Chapter 5, the results from assessing reoccurrence and co-occurrences of platial clusters demonstrated once again the uniqueness of particular locations and the mixing affect of others. By extending the proposed framework approach into the temporal dimension, this demonstrated the utility of such combined approaches to not only investigate the static representation of place meaning, but over time as well. Investigating the intersection of platial clusters and physical places of known affordances together, at varying time periods, showed that platial content remerges over time as activities are repeated and collective place meanings are implicitly reaffirmed.

As our world is becoming increasingly urbanized and dynamic, gaining an understanding of the building blocks of these urban environments is bringing forth the need for a new science of cities (Batty, 2013). The work that I presented here is contributing towards this goal, as it provides a new lens to observe platial content as it emerges from the people themselves, and allows us to do so at levels of spatial and temporal granularity that far exceed our past capabilities.

6.2 Future Research

The frameworks introduced in this dissertation provide a natural starting point for future platial information research. Ultimately, there are many directions that could be taken to improve upon this work from both an investigation into platial content within other crowd-contributed data sources and development of new quantitative methods. Ultimately, the platial world is rich with human perspectives and discourses based on places that the spatial world is not able to describe (Goodchild 2014). Although place research is nothing new, as seen in the literature review, the proliferations of Web 2.0 technologies to produce crowd-contributed content in real-time and advances in data discovery techniques offer a renewed focus whereby the early social constructs and theories of place are now observable at scales never before imaginable. Thus making the outlook on new platial research extremely promising.

In reflecting on the proposed approach, and looking to the future, there are several directions that become clear for mining platial content. One future research direction is the investigation of other topic modeling variations to improve not only topic coherence, but also using online processing methods vice the chosen batch approach. This would allow processing of platial content in real-time as the data is streaming directly from the APIs. In particular, the more recent developments of Cheng et al., (2014) use of a biterm topic model (BTM) to address the document level word co-occurrence pattern issues found in tweets. Their results show increased topic coherence over vanilla LDA by using global word co-occurrence patterns from an entire corpus of tweets vice dividing the corpus into smaller groups or documents that model terms locally. Additionally, they proposed online algorithms to store a small fraction of data on the fly for model update, which are much more efficient than the batch algorithm on large scale data set (Cheng et al., 2014)

In addition to future work in extracting meaningful phrases from text is alternate or improved methods for thematic labeling. The proposed approach used Wikipedia and PMI with selected high-level categories. There exist a plethora of other ways to calculate semantic similarity to include network-based approaches (Li et al., 2015). For which, semantic similarity is calculated based on the network distance between two different concepts.

The proposed approached performed a one-to-one mapping between bi-gram and high-level categories, which generalized the thematic clusters into a single class. Although this demonstrated the emergence of platial content, new research is needed to represent multiple themes simultaneous as seen from the co-occurrence results showing intersecting clusters that in some case could represent multiple activities within the context of a single event. In which case, spatio-temporal data mining approaches are needed to find frequent patterns that are difficult to visually synthesis given large volumes of highly linked data.

One obvious finding from this platial research is that combining locational, temporal, and textual context are paramount. However, there are sources of information that could provide context to improve the thematic labeling of platial clusters. One of those being location-based social networks, whereby the clustering of friends in network space could further validate or add confidence to observations the platial information. A future research endeavor could include the construction of an ensemble-based approach that combines observations from each dimension to boost the thematic classification results or perform soft voting. In either case, there exists a significant amount of research to be done on approaches to enhance the semantic meaning of platial content.

The use of incremental spatial autocorrelation at varying distances was proposed to address changes in scale due to events and general unevenness of the spatial distribution of tweets. Future research is proposed to investigate the combination of incremental spatial autocorrelation, local scale-sensitive indicators, and temporal space-time interactions tailored for crowd-contributed data. Westerholt et al., (2015) presented a modified scale-sensitive version of the Getis-Ord Gi* statistic to address the scale mixing of human activities found in crowd-sourced data.

To address the difference in scales, the researchers modified the spatial weights matrix by including scale-adjusted neighborhoods using a distance interval [$d_{min},$ d_{max}], as opposed to a fixed neighborhood distance. This approach has the advantage of detecting spatial clusters of different sizes within a single data source like Twitter. However, their approach does not address the generation of distance intervals or account for varying time periods.

Lastly, Twitter and Wikipedia were the only crowd-contributed data sources investigated in this study. The demonstrated results from these sources are promising, but at the same time beg the question of whether similar findings would emerge in other data sources and other languages. Ultimately, platial information is formed through textual descriptions that explicitly or implicitly describe a location, which broadly encompasses many crowd derived data sources. Some of these data sources include Foursquare, Flickr, Yelp, and Facebook Places, and sources predominately not the English such as Sina-Weibo from China and CKontakte (VK) from Russia to name a few. The use of knowledge base for determining semantic similarity between categories and phrases offers the added flexibility of being able to work with different languages. Although, a different language tokenizer would be necessary for using the topic modeling approach, however, must language tokenizer are freely available on the internet.

Lastly, this research contributed an algorithmic technique for harvesting platial themes from social media and Wikipedia, which had not been previously explored together in the context of collective meanings of place. The goal of validating results at multiple scales using cross-source consensus was accomplished to discover the uniqueness of place at the neighborhood scale and lack of uniqueness at the city scale. The accomplishments presented in this work provide a foundation for future research to further investigate the dynamics of collective place meanings and the inherent dynamics.

REFERENCES

Adams, B., & Janowicz, K. (2015). Thematic signatures for cleansing and enriching place-related linked data. *International Journal of Geographical Information Science*, *29*(4), 556-579.

Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998). *Automatic subspace clustering of high dimensional data for data mining applications* (Vol. 27, No. 2, pp. 94-105). ACM.

Ahlqvist, O., & Shortridge, A. (2006). *Characterizing land cover structure with semantic variograms* (pp. 401-415). Springer Berlin Heidelberg.

Aiello, L. M., Schifanella, R., Quercia, D., & Aletta, F. (2016). Chatty maps: constructing sound maps of urban areas from social media data. *Open Science*, *3*(3), 150690.

Alani, H., Jones, C. B., & Tudhope, D. (2001). Voronoi-based region approximation for geographical information retrieval with gazetteers. *International Journal of Geographical Information Science*, *15*(4), 287-306.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, *26*(11), 832-843.

de Andrade, F. G., de Souza Baptista, C., & Davis Jr, C. A. (2014). Improving geographic information retrieval in spatial data infrastructures. *GeoInformatica*, *18*(4), 793-818.

Andrea Rodriguez, M., & Egenhofer, M. J. (2004). Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. *International Journal of Geographical Information Science*, *18*(3), 229-256.

Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., & Thom, D. (2013). Thematic patterns in georeferenced tweets through space-time visual analytics. *Computing in Science and Engineering*, *15*(3), 72-82.

Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999, June). OPTICS: ordering points to identify the clustering structure. In *ACM Sigmod Record* (Vol. 28, No. 2, pp. 49-60). ACM.

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Advances in Knowledge Discovery and Data Mining* (pp. 391-402). Springer Berlin Heidelberg.

Batty, M. (2007). *Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals*. The MIT press.

Baumeister, J., Reutelshoefer, J., & Puppe, F. (2011). KnowWE: a Semantic Wiki for knowledge engineering. *Applied Intelligence*, *35*(3), 323-344.

Beard, K. (2012, November). A semantic web based gazetteer model for VGI. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information* (pp. 54-61). ACM.

Beaubouef, T., Ladner, R., & Petry, F. (2004). Rough set spatial data modeling for data mining. *International Journal of Intelligent Systems*, *19*(7), 567-584.

De Berg, M., Van Kreveld, M., Overmars, M., & Schwarzkopf, O. C. (2000). *Computational geometry* (pp. 1-17). Springer Berlin Heidelberg.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific american*, *284*(5), 28-37.

Bishr, M., & Janowicz, K. (2010, September). Can we trust information?-the case of volunteered geographic information. In *Towards Digital Earth Search Discover and Share Geospatial Data Workshop at Future Internet Symposium, volume* (Vol. 640).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

Brandenburg, A. M., & Carroll, M. S. (1995). Your place or mine?: The effect of place creation on environmental values and landscape meanings. *Society & Natural Resources*, *8*(5), 381-398.

Brindley, P., Goulding, J., & Wilson, M. L. (2014, November). A data driven approach to mapping urban neighbourhoods. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 437-440). ACM.

Burton, S. H., Tanner, K. W., Giraud-Carrier, C. G., West, J. H., & Barnes, M. D. (2012). " Right time, right place" health communication on Twitter: value and accuracy of location information. *Journal of medical Internet research*, *14*(6), e156.

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). Btm: Topic modeling over short texts. *Knowledge and Data Engineering, IEEE Transactions on*, *26*(12), 2928-2941.

Cilibrasi, R. L., & Vitanyi, P. (2007). The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, *19*(3), 370-383.

Cohn, A. G., Bennett, B., Gooday, J., & Gotts, N. M. (1997). Qualitative spatial representation and reasoning with the region connection calculus. *GeoInformatica*, *1*(3), 275-316.

Coleman, D. J., Georgiadou, Y., & Labonte, J. (2009). Volunteered geographic information: The nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research*, *4*(1), 332-358.

Crang, M., & Graham, S. (2007). Sentient cities ambient intelligence and the politics of urban space. *Information, Communication & Society*, *10*(6), 789-817.

Cranshaw, J., Schwartz, R., Hong, J. I., & Sadeh, N. (2012, June). The livehoods project: Utilizing social media to understand the dynamics of a city. In *International AAAI Conference on Weblogs and Social Media* (p. 58).

Croitoru, A., Wayant, N., Crooks, A., Radzikowski, J., & Stefanidis, A. (2015). Linking cyber and physical spaces through community detection and clustering in social media feeds. *Computers, Environment and Urban Systems*, *53*, 47-64.

Croitoru, Arie, Andrew Crooks, Jacek Radzikowski, R. Vatsavai, Anthony Stefanidis, and Nicole Wayant. (2014). Geoinformatics and Social Media: A New Big Data Challenge. In Big Data Techniques and Technologies in Geoinformatics (H. Karimi, editor), CRC Press, pp. 207-232.

Croitoru, A., Crooks, A., Radzikowski, J., & Stefanidis, A. (2013). Geosocial gauge: a system prototype for knowledge discovery from social media. *International Journal of Geographical Information Science*, *27*(12), 2483-2508.

Crooks, A., Pfoser, D., Jenkins, A., Croitoru, A., Stefanidis, A., Smith, D., ... & Lamprianidis, G. (2015). Crowdsourcing urban form and function. *International Journal of Geographical Information Science*, *29*(5), 720-741.

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). # Earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, *17*(1), 124-147.

Davies, C., Holt, I., Green, J., Harding, J., & Diamond, L. (2009). User needs and implications for modelling vague named places. *Spatial Cognition & Computation*, *9*(3), 174-194.

Davies, C. (2009). Are places concepts? familarity and expertise effects in neighborhood cognition. In *Spatial Information Theory* (pp. 36-50). Springer Berlin Heidelberg.

de Mattos, B., Bezerra, L., Aguiar, F., Moura-Neto, C., Zucco, C. A., & Cascon, P. (2013). Diet, activity patterns, microhabitat use and defensive strategies of Rhinella hoogmoedi Caramaschi & Pombal, 2006 from a humid forest in northeast Brazil. *The Herpetological Journal*, *23*(1), 29-37.

Madirolas, G., & de Polavieja, G. G. (2012). Wisdom of the confident: Using social interactions to eliminate the bias in wisdom of the crowds. *Collective intelligence*, *2015*.

Dey, A., & Prukayastha, B. S. (2013). Named Entity Recognition using Gazetteer Method and N-gram Technique for an Inflectional Language: A Hybrid Approach. *International Journal of Computer Applications*, *84*(9).

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297-302.

Duckham, M., Goodchild, M. F., & Worboys, M. (Eds.). (2004). *Foundations of geographic information science*. CRC Press.

Earle, P. S., Bowden, D. C., & Guy, M. (2012). Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, *54*(6).

ElGindy, E., & Abdelmoty, A. (2014). Enriching user profiles using geo-social place semantics in geo-folksonomies. *International Journal of Geographical Information Science*, *28*(7), 1439-1458.

Elwood, S. (2008). Geographic Information Science: new geovisualization technologies–emerging questions and linkages with GIScience research. *Progress in Human Geography*.

Elwood, S., Goodchild, M. F., & Sui, D. (2013). *Prospects for VGI research and the emerging fourth paradigm* (pp. 361-375). Springer Netherlands.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).

Fan, J., & Stewart, K. (2015). Detecting Spatial Patterns of Natural Hazards from the Wikipedia Knowledge Base. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *2*(4), 87.

Farrahi, K., & Gatica-Perez, D. (2011). Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(1), 3.

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

Frank, M. R., Mitchell, L., Dodds, P. S., & Danforth, C. M. (2013). Happiness and the patterns of life: A study of geolocated tweets. *Scientific reports*, *3*.

Franzoni, V., & Milani, A. (2012, December). PMING Distance: A collaborative semantic proximity measure. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on* (Vol. 2, pp. 442-449). IEEE.

Gelernter, J., Ganesh, G., Krishnakumar, H., & Zhang, W. (2013, November). Automatic gazetteer enrichment with user-geocoded data. In *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information* (pp. 87-94). ACM.

Genc, Y., Sakamoto, Y., & Nickerson, J. V. (2011). Discovering context: classifying tweets through a semantic transform based on wikipedia. In *Foundations of augmented cognition. Directing the future of adaptive systems* (pp. 484-492). Springer Berlin Heidelberg.

Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical analysis*, *24*(3), 189-206.

Ghosh, D., & Guha, R. (2013). What are we 'tweeting'about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science*, *40*(2), 90-102.

De Felice, G., Fogliaroni, P., & Wallgrün, J. O. (2011). A hybrid geometric-qualitative spatial reasoning system and its application in gis. In *Spatial Information Theory* (pp. 188-209). Springer Berlin Heidelberg.

Goldberg, D. W., Wilson, J. P., & Knoblock, C. A. (2009). Extracting geographic features from the internet to automatically build detailed regional gazetteers. *International Journal of Geographical Information Science*, *23*(1), 93-128.

Golledge, R. G. (1992). Place recognition and wayfinding: Making sense of space. *Geoforum*, *23*(2), 199-214.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211-221.

Goodchild, M. F. (2015). Space, place and health. *Annals of GIS*, *21*(2), 97-100.

Goodchild, M. (2009). NeoGeography and the nature of geographic expertise. *Journal of location based services*, *3*(2), 82-96.

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

Goodchild, M. F., & Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, *3*(3), 231-241.

Gracia, J., & Mena, E. (2008). Web-based measure of semantic relatedness. In *Web Information Systems Engineering-WISE 2008* (pp. 136-150). Springer Berlin Heidelberg.

Graham, M., Hogan, B., Straumann, R. K., & Medhat, A. (2014). Uneven geographies of user-generated information: patterns of increasing informational poverty. *Annals of the Association of American Geographers*, *104*(4), 746-764.

Grieve, J. (2012). Sociolinguistics: Quantitative Methods. *The encyclopedia of applied linguistics*.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, *114*(2), 211.

He, Y., Lin, C., Gao, W., & Wong, K. F. (2013). Dynamic joint sentiment-topic model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5*(1), 6.

Hidalgo, M. C., & Hernandez, B. (2001). Place attachment: Conceptual and empirical questions. *Journal of environmental psychology*, *21*(3), 273-281.

Hill, L. L., Frew, J., & Zheng, Q. Geographic Names. The implementation of a gazetteer in a georeferenced digital library. Digital Library. 5 (1), 1999.

Hinneburg, A., & Keim, D. A. (1998, August). An efficient approach to clustering in large multimedia databases with noise. In *KDD* (Vol. 98, pp. 58-65).

Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. (2013). YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, *194*, 28-61.

Iosif, E., & Potamianos, A. (2010). Unsupervised semantic similarity computation between terms using web documents. *Knowledge and Data Engineering, IEEE Transactions on*, *22*(11), 1637-1647.

Iosif, E., & Potamianos, A. (2015). Similarity computation using semantic networks created from web-harvested data. *Natural Language Engineering*, *21*(01), 49-79.

Machado, I. M. R., de Alencar, R. O., de Oliveira Campos Jr, R., & Davis Jr, C. A. (2011). An ontological gazetteer and its application for place name disambiguation in text. *Journal of the Brazilian Computer Society*, *17*(4), 267-279.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New phytologist*, *11*(2), 37-50.

Janáč, M., & Jurajda, P. (2013). Diel differences in 0+ fish samples: effect of river size and habitat. *River Research and Applications*, *29*(1), 90-98.

Jones, C. B., Purves, R., Ruas, A., Sanderson, M., Sester, M., Van Kreveld, M., & Weibel, R. (2002, August). Spatial information retrieval and geographical ontologies an overview of the SPIRIT project. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 387-388). ACM.

Jones, C. B., Alani, H., & Tudhope, D. (2001). Geographical information retrieval with ontologies of place. In *Spatial information theory* (pp. 322-335). Springer Berlin Heidelberg.

Jones, P., & Evans, J. (2012). Rescue geography: Place making, affect and regeneration. *Urban studies*, *49*(11), 2315-2330.

Jordan, A. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, *14*, 841.

Jordan, T., Raubal, M., Gartrell, B., & Egenhofer, M. (1998, July). An affordance-based model of place in GIS. In *8th Int. Symposium on Spatial Data Handling, SDH* (Vol. 98, pp. 98-109).

Kämpf, M., Tismer, S., Kantelhardt, J. W., & Muchnik, L. (2012). Fluctuations in Wikipedia access-rate and edit-event data. *Physica A: Statistical Mechanics and its Applications*, *391*(23), 6101-6111.

Keßler, C., Janowicz, K., & Bishr, M. (2009, November). An agenda for the next generation gazetteer: Geographic information contribution and retrieval. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in Geographic Information Systems* (pp. 91-100). ACM.

Kenkel, N. C., & Orlóci, L. (1986). Applying metric and nonmetric multidimensional scaling to ecological studies: some new results. *Ecology*, 919-928.

Kridel, C. (Ed.). (2010). *Encyclopedia of curriculum studies* (Vol. 1). Sage.

Kulik, L. (2001). A geometric theory of vague boundaries based on supervaluation. In *Spatial information theory* (pp. 44-59). Springer Berlin Heidelberg.

Kulik, L. (2001). A geometric theory of vague boundaries based on supervaluation. In *Spatial information theory* (pp. 44-59). Springer Berlin Heidelberg.

Lamprianidis, G., & Pfoser, D. (2012, November). Collaborative geospatial feature search. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems* (pp. 169-178). ACM.

Lee, R., Wakamiya, S., & Sumiya, K. (2015). Exploring geospatial cognition based on location-based social network sites. *World Wide Web*, *18*(4), 845-870.

Lewicka, M. (2010). What makes neighborhood different from home and city? Effects of place scale on place attachment. *Journal of environmental psychology*, *30*(1), 35-51.

Li, B., & Fonseca, F. (2006). TDD: A comprehensive model for qualitative spatial similarity assessment. *Spatial Cognition and Computation*, *6*(1), 31-62.

Li, C., Cheung, W. K., Ye, Y., Zhang, X., Chu, D., & Li, X. (2015). The Author-Topic-Community model for author interest profiling and community discovery. *Knowledge and Information Systems*, *44*(2), 359-383.

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, *40*(2), 61-77.

Li, P., Wang, H., Zhu, K. Q., Wang, Z., Hu, X., & Wu, X. (2015). A large probabilistic semantic network based approach to compute term similarity. *Knowledge and Data Engineering, IEEE Transactions on*, *27*(10), 2604-2617.

Li, Y. (2014). Semantic image similarity based on deep knowledge for effective image retrieval.

Lih, A. (2004). Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. *Nature*.

Lim, K. W., Chen, C., & Buntine, W. (2013, December). Twitter-network topic model: A full Bayesian treatment for social network and text modeling. In *NIPS2013 Topic Model workshop* (pp. 4-4).

Lippard, L. R. (1997). *The lure of the local: Senses of place in a multicentered society* (p. 9). New York: New Press.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., ... & Shi, L. (2015). Social sensing: a new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers*, *105*(3), 512-530.

Liu, Y., Guo, Q. H., Wieczorek, J., & Goodchild, M. F. (2009). Positioning localities based on spatial assertions. *International Journal of Geographical Information Science*, *23*(11), 1471-1501.

Lombard, M. (2014). Constructing ordinary places: Place-making in urban informal settlements in Mexico. *Progress in Planning*, *94*, 1-53.

Longan, M. W. (2002). Building a global sense of place: The community networking movement in the United States. *Urban Geography*, *23*(3), 213-236.

Haklay, M. (2010). How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordinance Survey datasets for London and the rest of England. *Environ Planning B*, *37*, 682-703.

Manguinhas, H., Martins, B., & Borbinha, J. (2008, November). A geo-temporal web gazetteer integrating data from multiple sources. In *Digital Information Management, 2008. ICDIM 2008. Third International Conference on* (pp. 146-153). IEEE.

Massey, D. (2010). *A global sense of place* (pp. pp-232). aughty. org.

McIntosh, J., & Yuan, M. (2005). Assessing similarity of geographic processes and events. *Transactions in GIS*, *9*(2), 223-245.

Medelyan, O., Milne, D., Legg, C., & Witten, I. H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, *67*(9), 716-754.

Michelson, M., & Macskassy, S. A. (2010, October). Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data* (pp. 73-80). ACM.

Milne, D., & Witten, I. H. (2013). An open-source toolkit for mining Wikipedia. *Artificial Intelligence*, *194*, 222-239.

Mislove, A., Lehmann, S., Ahn, Y. Y., Onnela, J. P., & Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. *ICWSM*, *11*, 5th.

MIT Technology Review. The Decline of Wikipedia (2013-10-22). http://www.technologyreview.com/featuredstory/520446/the-decline-of-wikipedia/

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*, *8*(5), e64417.

Moncla, L., Renteria-Agualimpia, W., Nogueras-Iso, J., & Gaio, M. (2014, November). Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 183-192). ACM.

Mountain, D., & Mcfarlane, A. (2007). Geographic information retrieval in a mobile environment: evaluating the needs of mobile individuals. *Journal of Information Science*.

Mülligann, C., Janowicz, K., Ye, M., & Lee, W. C. (2011). Analyzing the spatial-semantic interaction of points of interest in volunteered geographic information. In *Spatial information theory* (pp. 350-370). Springer Berlin Heidelberg.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010, June). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 100-108). Association for Computational Linguistics.

Newman, D., Asuncion, A., Smyth, P., & Welling, M. (2009). Distributed algorithms for topic models. *The Journal of Machine Learning Research*, *10*, 1801-1828.

Niraula, N., Banjade, R., Ştefănescu, D., & Rus, V. (2013). Experiments with semantic similarity measures based on lda and lsa. In *Statistical Language and Speech Processing* (pp. 188-199). Springer Berlin Heidelberg.

Noulas, A., Scellato, S., Lathia, N., & Mascolo, C. (2012, December). Mining user mobility features for next place prediction in location-based services. In *Data mining (ICDM), 2012 IEEE 12th international conference on* (pp. 1038-1043). IEEE.

Oh, O., Agrawal, M., & Rao, H. R. (2011). Information control and terrorism: Tracking the Mumbai terrorist attack through twitter. *Information Systems Frontiers*, *13*(1), 33-43.

Omar, A. H., & Salleh, M. N. M. (2013). Modeling Unstructured Document Using N-gram Consecutive and WordNet Dictionary. In *pie* (Vol. 77, p. 1).

Peuquet, D. J. (1994). It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of american Geographers*, *84*(3), 441-461.

Pfoser, D., Tao, K., Mouratidis, K, Nascimento, M., Mokbel, M., Shekhar, S., and Huang, Y., eds. (2011). *Advances in Spatial and Temporal Databases*. Vol. 6849. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg.

Pred, A. (1984). Place as historically contingent process: Structuration and the time-geography of becoming places. *Annals of the association of american geographers*, *74*(2), 279-297.

Purcell, M. (1997). Ruling Los Angeles: Neighborhood movements, urban regimes, and the production of space in southern California. *Urban Geography*, *18*(8), 684-704.

Purves, R. S., & Derungs, C. (2015). From Space to Place: Place-Based Explorations of Text. *International Journal of Humanities and Arts Computing*, *9*(1), 74-94.

Quesnot, T., & Roche, S. (2015, January). Platial or locational data? toward the characterization of social location sharing. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on* (pp. 1973-1982). IEEE.

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior research methods*, *41*(3), 647-656.

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior research methods*, *41*(3), 647-656.

Renkonen O. (1938).“*Statistisch-ökologische Untersuchungen über die terrestrische Käferwelt der finnischen Bruchmoore.”*, Doctoral dissertation, Societas zoologica-botanica Fennica Vanamo.

Rios-Martinez, J., Spalanzani, A., & Laugier, C. (2015). From proxemics theory to socially-aware navigation: A survey. *International Journal of Social Robotics*, *7*(2), 137-153.

Rettinger, A., Lösch, U., Tresp, V., d'Amato, C., & Fanizzi, N. (2012). Mining the semantic web. *Data Mining and Knowledge Discovery*, *24*(3), 613-662.

Sánchez, D., Batet, M., Isern, D., & Valls, A. (2012). Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications*, *39*(9), 7718-7728.

Scassa, T. (2013). Legal issues with volunteered geographic information. *The Canadian Geographer/Le Géographe canadien*, *57*(1), 1-10.

Schlieder, C., Vögele, T., & Visser, U. (2001). Qualitative spatial representation for information retrieval by gazetteers. In *Spatial Information Theory* (pp. 336-351). Springer Berlin Heidelberg.

Schönhofen, P. (2009). Identifying document topics using the Wikipedia category network. *Web Intelligence and Agent Systems: An International Journal*, *7*(2), 195-207.

Schwering, A., & Raubal, M. (2005). *Spatial relations for semantic similarity measurement* (pp. 259-269). Springer Berlin Heidelberg.

Shamai, S. (1991). Sense of place: An empirical measurement. *Geoforum*, *22*(3), 347-358.

Shekhar, S., & Huang, Y. (2001). Discovering spatial co-location patterns: A summary of results. In *Advances in Spatial and Temporal Databases* (pp. 236-256). Springer Berlin Heidelberg.

Silver, A., & Grek-Martin, J. (2015). “Now we understand what community really means”: Reconceptualizing the role of sense of place in the disaster recovery process. *Journal of Environmental Psychology*, *42*, 32-41.

Simpson, E. H. (1949). Measurement of diversity. *Nature*.

Spielman, S. E. (2014). Spatial collective intelligence? Credibility, accuracy, and volunteered geographic information. *Cartography and geographic information science*, *41*(2), 115-124.

Stefanidis, A., Crooks, A., & Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. *GeoJournal*, *78*(2), 319-338.

Steiger, E., Resch, B., & Zipf, A. (2015). Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographical Information Science*, 1-23.

Steiger, E., Westerholt, R., Resch, B., & Zipf, A. (2015). Twitter as an indicator for whereabouts of people? Correlating twitter with uk census data. *Computers, Environment and Urban Systems*, *54*, 255-265.

Strube, M., & Ponzetto, S. P. (2006, July). WikiRelate! Computing semantic relatedness using Wikipedia. In *AAAI* (Vol. 6, pp. 1419-1424).

Su, Q., Xiang, K., Wang, H., Sun, B., & Yu, S. (2006). Using pointwise mutual information to identify implicit features in customer reviews. In *Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead* (pp. 22-30). Springer Berlin Heidelberg.

Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business. *Economies, Societies and Nations*, *296*.

Tammet, T., Luberg, A., & Järv, P. (2013). *Sightsmap: crowd-sourced popularity of the world places* (pp. 314-325). Springer Berlin Heidelberg.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, *62*(2), 406-418.

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

Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, *24*(3), 478-514.

Tuan, Y. F. (1977). *Space and place: The perspective of experience*. U of Minnesota Press.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, 0894439310386557.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, 0894439310386557.

Boland, P. (2010). Sonic geography, place and race in the formation of local identity: Liverpool and Scousers. *Geografiska Annaler: Series B, Human Geography*, *92*(1), 1-22.

Vögele, T. J., & Stuckenschmidt, H. (2001). Enhancing Gazetteers with Qualitative Spatial Concepts. In *Proceedings of the Workshop on Hypermedia in Environmental Protection*.

Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., & Studer, R. (2006, May). Semantic wikipedia. In *Proceedings of the 15th international conference on World Wide Web* (pp. 585-594). ACM.

Wales, J. (2005). Jimmy Wales Talks Wikipedia. Retrieved April 24, 2016, from http://www.writingshow.com/articles/transcripts/2006/01012006.html

Wallach, H. M. (2006, June). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning* (pp. 977-984). ACM.

Wang, P., Hu, J., Zeng, H. J., & Chen, Z. (2009). Using Wikipedia knowledge to improve text classification. *Knowledge and Information Systems*, *19*(3), 265-281.

Westerholt, R., Resch, B., & Zipf, A. (2015). A local scale-sensitive indicator of spatial autocorrelation for assessing high-and low-value clusters in multiscale datasets. *International Journal of Geographical Information Science*, *29*(5), 868-887.

Williams, D. R., & Stewart, S. I. (1998). Sense of place: An elusive concept that is finding a home in ecosystem management. *Journal of forestry*, *96*(5), 18-23.

Winter, S., & Freksa, C. (2012). Approaching the notion of place by contrast. *Journal of Spatial Information Science*, *2012*(5), 31-50.

Winter, S., Kuhn, W., & Krüger, A. (2009). Guest editorial: Does place have a place in geographic information science?.

Witten, I., & Milne, D. (2008, July). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA* (pp. 25-30).

Wolda, H. (1981). Similarity indices, sample size and diversity. *Oecologia*, *50*(3), 296-302.

Wu, L., Wang, D., Guo, C., Zhang, J., & wen Chen, C. (2016, January). User Profiling by Combining Topic Modeling and Pointwise Mutual Information (TM-PMI). In *MultiMedia Modeling* (pp. 152-161). Springer International Publishing.

Zandbergen, P. A. (2009). Accuracy of iPhone locations: A comparison of assisted GPS, WiFi and cellular positioning. *Transactions in GIS*, *13*(s1), 5-25.

Zesch, T., & Gurevych, I. (2010). Wisdom of crowds versus wisdom of linguists–measuring the semantic relatedness of words. *Natural Language Engineering*, *16*(01), 25-59.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval* (pp. 338-349). Springer Berlin Heidelberg.

Zook, M., Graham, M., Shelton, T., & Gorman, S. (2010). Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake. *World Medical & Health Policy*, *2*(2), 7-33.

BIOGRAPHY

Andrew Jenkins was born in Winchester, Virginia in 1982, and graduated from Sherando High School in Stephens City, Virginia, in 2000. He was employed as a Senior Geospatial Analyst in the US Army on active duty from 2002-2009 with several combat deployments. While in the US Army, he received his Bachelor of Science in Computer and Information Science from University of Maryland University College in 2009. After leaving active duty, Andrew worked as a government researcher at the US Army Topographic Engineering Center, Ft. Belvoir, Virginia serving as a principal investigator on many applied research projects. Andrew received his Master of Science in GeoInformatics and Geospatial Intelligence and a Graduate Certificate in Geospatial Intelligence from George Mason University in 2012. He also served in the US Army reserve from 2009-2012 as a Senior Geospatial Analysis Instructor at the National Geospatial-Intelligence College at NGA. In the spring of 2013, he entered the Ph.D. program in Earth Systems and Geoinformation Sciences at George Mason University in Fairfax, Virginia. He was a graduate research assistant with the Center for Geospatial Intelligence working on place research and later received a grant to focus on his doctoral dissertation. In the fall of 2013, Andrew left the government and entered the private sector to work for DigitalGlobe Inc. where he is currently a Principal Data Scientist.