

UNIVERSITY OF CALIFORNIA
Santa Barbara

A Temporal Approach to Defining Place Types
based on User-Contributed Geosocial Content

A dissertation submitted in partial satisfaction
of the requirements for the degree of Doctor of Philosophy in
Geography

by

Grant Donald McKenzie

Committee in Charge:

Professor Krzysztof Janowicz, Co-Chair

Professor Martin Raubal, Co-Chair

Professor Konstadinos Goulias

Professor Andrew Flanagin

March 2015

The dissertation of
Grant Donald McKenzie is approved:

Professor Konstadinos Goulias

Professor Andrew Flanagin

Professor Martin Raubal, Committee Co-Chairperson

Professor Krzysztof Janowicz, Committee Co-Chairperson

March 2015

A Temporal Approach to Defining Place Types
based on User-Contributed Geosocial Content

Copyright © 2015

by

Grant Donald McKenzie

To my parents and sister for their love and support

Acknowledgements

As I close the *graduate student* chapter of my life and finally look up from my laptop, it becomes shockingly apparent how restrictive a single page of acknowledgments is.

I would like to thank the members of my committee for their input and guidance over the years. Though our physical paths diverged early in the doctoral process, Martin did a extraordinary job of continuing to advise, support and provide input on many aspects of my degree, proving that a good advisor is not bounded by physical space. Working with Jano, I found not only an advisor with exceptional drive, but a mentor who was able to inspire creativity even in the darkest depths of the dissertation process. I am still not entirely sure how he was able to do it, but Kostas helped keep to the entire Ph.D. process in perspective while providing invaluable input. I am very thankful for Andrew's perspective on my research. The value of feedback from a researcher outside of one's primary field, especially one as sharp as Andrew cannot be overstated.

Being able to work with intelligent and thoughtful colleagues such as those in the STKO lab is truly what has made the last few years of research enjoyable. I am incredibly grateful for all of their help and look forward to many long days and late nights of project work with them in the future. I must also acknowledge some great friends and colleagues I made along the way: Ben, Kate, Mike, Carlos, Burke, as well as the countless graduate and undergraduate students, postdocs and researchers in the Department of Geography. A heartfelt thank you to Sarah who made the entire process bearable, for putting up with me through the doldrums and the gale force winds, words cannot express my gratitude.

Last, and most importantly, I would like to thank my parents and sister for their unconditional love and support. They truly provided the foundation on which all of this has been possible. I could honestly never ask for a more caring and supportive family than the one I have.

Curriculum Vitæ

Grant Donald McKenzie

EDUCATION

- 2015 Doctor of Philosophy in Geography with an Emphasis in Technology and Society, University of California, Santa Barbara, United States of America (Expected)
- 2008 Master of Applied Science in Geomatic Engineering, The University of Melbourne, Australia
- 2004 Advanced Diploma in Geographic Information Systems, The British Columbia Institute of Technology, Canada
- 2002 Bachelors of Arts in Geography, The University of British Columbia, Canada

PROFESSIONAL EXPERIENCE

- 2011 – 2015 Research Assistant, University of California, Santa Barbara
- 2010 – 2013 Teaching Assistant, University of California, Santa Barbara
- 2011 Visiting Researcher, Institute of Cartography and Geoinformation, ETH Zürich
- 2009 – 2015 Geospatial Technologist, Spatial Development International
- 2008 – 2010 Geospatial Technologist, Private Consulting

2008 – 2009

Geospatial Software Developer, CH2M Hill

PUBLICATIONS

Refereed Journal Articles, Conference Proceedings & Book Chapters

Adams, B., **McKenzie, G.**, Gahegan, M. (Accepted) Frankenplace: Interactive thematic mapping for ad hoc exploratory search. *The 24th International World Wide Web Conference (WWW 2015)*. (May 18-22, Florence, Italy)

McKenzie, G., Klippel, A., (In Press) The Interaction of Landmarks and Map Alignment in You-Are-Here Maps. *The Cartographic Journal*. Maney Press.

McKenzie, G., Janowicz, K., Gao, S., Yang, J., .Hu, Y., (In Press) POI Pulse: A Multi-Granular, Semantic Signatures-Based Information Observatory for the Interactive Visualization of Big Geosocial Data. *Cartographica: The International Journal for Geographic Information and Geovisualization*. University of Toronto Press.

Savage, S. Forbes, A., Toxtli, C., **McKenzie, G.**, Desai, S. M., Hollerer, T. (2014) Visual Analysis of Targeted Audiences. *The 11th International Conference on the Design of Cooperative Systems* (May 27-30, Nice, France)

McKenzie, G., Janowicz, K., Adams, B. (2014) A Weighted Multi-Attribute Method for Matching User-Generated Points of Interest. *Cartography and Geographic Information Science*. 41(2) pp 125-137 Taylor & Francis

McKenzie, G., Janowicz, K., Adams, B. (2013) Weighted Multi-Attribute Matching of User-Generated Points of Interest. *Proceedings of The 21st ACM SIGSPATIAL In-*

ternational Conference on Advances in Geographic Information Systems (Short Paper).
pp 440-443 (November 5-8, Orlando, FL)

Gau, S., Hu, Y., Janowicz, K., **McKenzie, G.**, (2013) A Spatiotemporal Scientometrics Framework for Exploring the Citation Impact of Publications and Scientists. *Proceedings of The 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp 204-213 (November 5-8, Orlando, FL)

Hu, Y., Janowicz, K., **McKenzie, G.**, Sengupta, K., Hitzler, P. (2013) A Linked Data-driven Semantically-enabled Journal Portal for Scientometrics. *The 12th International Semantic Web Conference (ISWC '13)* pp 114-129 (October 21-25, Sydney, Australia)

McKenzie, G., Adams, B., Janowicz, K. (2013) A Thematic Approach to User Similarity Built on Geosocial Check-ins. *In Proceedings of the 16th AGILE Conference on Geographic Information Science (AGILE '13)* pp 39-54 (May 14-17, Leuven, Belgium)

Adams, B. and **McKenzie, G.** (2012) Inferring Thematic Places from Spatially Referenced Natural Language Descriptions. In: D. Sui, S. Elwood, and M. Goodchild (Eds.), *Crowdsourcing Geographic Knowledge*, pp 201-221. Springer

Refereed Extended Abstracts & Demonstration Papers

McKenzie, G. (Accepted) The pretense of residential privacy in geosocial networking data. *International Conference on Location-Based Social Media Data*. (March 13-14, Athens, GA)

Gong, L., Gao, S., **McKenzie, G.** (Accepted) POI Type Matching based on Culturally Different Datasets. *International Conference on Location-Based Social Media Data*. (March 13-14, Athens, GA)

McKenzie, G., Janowicz, K. (2014) Coerced Geographic Information: The not-so-voluntary side of user-generated geo-content. *Extended Abstracts of the 8th International Conference on Geographic Information Science (GIScience '14)*, Vienna, Austria.

Gao, S., Yang, J-A., Yan, B., Hu, Y., Janowicz, K., **McKenzie, G.** (2014) Detecting Origin-Destination Mobility Flows From Geotagged Tweets in Greater Los Angeles Area. *Extended Abstracts of the 8th International Conference on Geographic Information Science (GIScience '14)*, Vienna, Austria.

McKenzie, G., Janowicz, K. (2014) Activities in a New City: Itinerary Recommendation Based on User Similarity. *Spatial Knowledge and Information (SKI) - Canada* (Feb 7-9, Banff, AB)

McKenzie, G., Janowicz, K., Hu, Y., Sengupta, K., Hitzler, P. (2013) Linked Scientometrics: Designing Interactive Scientometrics with Linked Data and Semantic Web Reasoning. *The 12th International Semantic Web Conference (ISWC '13)* (Demonstration Paper) Sydney, Australia

Adams, B. and **McKenzie, G.** (2012) Frankenplace: An Application for Similarity-Based Place Search. *The 6th International AAAI Conference on Weblogs and Social Media (ICWSM)* (Demonstration Paper) (June 4-7, Dublin, Ireland) AAAI Press

McKenzie, G. and Raubal, M. (2012) Ground-truthing spatial activities through on-line social networking data. *Extended Abstracts of the 7th International Conference on Geographic Information Science (GIScience '12)* Columbus, OH.

Hu, Y., Li, W., Janowicz, K., Deutsch, K., **McKenzie, G.**, Goulias, K. (2012) Using spatial-temporal signatures to infer human activities from personal trajectories on location-enabled mobile devices. *Extended Abstracts of the 7th International Conference on Geographic Information Science (GIScience '12)*, Columbus, OH.

McKenzie, G., Raubal, M. (2011) Adding social constraints to location based services. *Extended Abstracts of the 8th Symposium on Location-Based Services*. (November 21-23, Vienna, Austria)

Refereed Workshop Proceedings

Janowicz, K., Adams, B., **McKenzie, G.**, Kauppinen, T., (2014) Towards Geographic Information Observatories. *Proceedings of Workshop on Geographic Information Observatories at the Eight International Conference on Geographic Information Science (GIO @ GIScience '14)*, (September 23, Vienna, Austria)

Hu, Y., **McKenzie, G.**, Yang, J., Gao, S., Abdalla, A., Janowicz, K., (2014) A Linked-Data-Driven Web Portal for Learning Analytics: Data Enrichment, Interactive Visualization, and Knowledge Discovery. *LAK Data Challenge Workshop at The 4th International Conference on Learning Analytics and Knowledge (LAK 2014)*, (March 24, Indianapolis, IN)

Gao, S., Janowicz, K., **McKenzie, G.**, Li, L., (2013) Towards Platial Joins and Buffers in Place-Based GIS. *The First International Workshop on Computational Models of Place (CoMP)* at the *The 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, (November 5-8, Orlando, FL)

McKenzie, G., Barrett, T., Hegarty, M., Goodchild, M., Thompson, W. (2013) Assessing the Effectiveness of Visualizations for Accurate Judgements of Geospatial Uncertainty. *Visually-Supported Reasoning with Uncertainty workshop at The 11th International Conference on Spatial Information Theory (COSIT '13)* (September 2, Scarborough, UK)

McKenzie, G., Deutsch, K., Raubal, M. (2013) What, When and Where: The real-world activities that contribute to online social networking posts. *Action and Interaction in Volunteered Geographic Information (ACTIVITY) Workshop at The 16th AGILE Conference on Geographic Information Science (AGILE '13)* (May 14-17, Leuven, Belgium)

McKenzie, G. (2011) Gamification and Location-based Services. *Cognitive Engineering for Mobile GIS Workshop at the 15th International Conference on Spatial Information Theory (COSIT '11)* (September 12-16, Belfast, ME)

Non-refereed Publications & Magazines

Deutsch, K., **McKenzie, G.**, Hu, Y., Janowicz, J., Li, W., and Goulias, K., (2012) Examining the Use of Smartphones and Data Quality in Data Collection for Travel

Behavior Research. *Proceedings of the International Association of Travel Behavior Research*, Toronto, ON.

McKenzie, G. (2011) Determining Social Constraints in Time Geography through Online Social Networking. *The 15th International Conference on Spatial Information Theory 2011 (COSIT '11)* - Doctoral Colloquium. (September 12-16, Belfast, ME)

Deutsch, K., **McKenzie, G.**, Stevenson, D., Dara-Abrams, D., and Goulias, K. (2011) Integrating GPS and Smart Phone Technologies for Behavioral Data Collection. *Eurogeo: The meeting of the Association of European Geographers*. (June 2-4, Athens, Greece)

Smith, E., **McKenzie, G.**, Duckham, M. (2008) The only game in town. *Position Magazine* 33 pp. 51-54

Master's Thesis

McKenzie, G. (2007) Exploring the Influence of Landmark Presence and Map Alignment on Wayfinding Performance Using Virtual Environment Modeling.

AWARDS

2013 Excellence in Research Award - Department of Geography, UCSB

2012 Jack & Laura Dangermond Fellowship for promising research in GIScience

2012 Excellence in Teaching Award - Department of Geography, UCSB

2008 Google Doctoral Colloquium Award for promising research
2007 J H Mirams Memorial Research Scholarship

Abstract

A Temporal Approach to Defining Place Types based on User-Contributed Geosocial Content

Grant Donald McKenzie

Place is one of the foundational concepts on which the field of Geography has been built. Traditionally, GIScience research into place has been approached from a spatial perspective. While space is an integral feature of place, it represents only a single dimension (or a combination of three dimensions to be exact), in the complex, multidimensional concept that is *place*. Though existing research has shown that both spatial and thematic dimensions are valuable, time has historically been under-utilized in its ability to describe and define places and their types. The recent availability and access to user-generated geosocial content has allowed for a much deeper investigation of the temporal dimension of place. Multi-resolution temporal signatures are constructed based on these data permitting both place instances and place types to be compared through a robust set of (dis)similarity measures. The primary contribution of this work lies in demonstrating how places are defined through a better understanding of temporal user behavior. Furthermore, the results of this research present the argument that the temporal dimension is *the most indicative* placial dimension for classifying places by type.

Contents

Acknowledgements	v
Curriculum Vitæ	vi
Abstract	xiv
1 Introduction	1
1.1 Context & Motivation	1
1.2 The Dimensionality of Place	4
1.2.1 The Temporal Dimension	6
1.2.2 Semantic Signatures	7
1.2.3 Place Types	9
1.3 Research Contribution	10
1.4 Methods	14
1.5 Outline	17
2 What, When and Where: The real-world activities that contribute to online social networking posts	20
2.1 Introduction	22
2.2 Related Work	25
2.2.1 Activity Categorization	25
2.2.2 Online Social Networking	26
2.2.3 Activity Prediction	26
2.3 Methods	27
2.3.1 Data Collection	28
2.3.2 Activity Categorization	30
2.3.3 Facebook Posts	33
2.3.4 Analysis	33
2.4 Results & Discussion	36
2.4.1 What	36
2.4.2 Where	37

2.4.3	When	38
2.5	Limitations & Next Steps	39
2.6	Conclusions	40
3	A Thematic Approach to User Similarity Built on Geosocial Check-ins	42
3.1	Introduction	44
3.2	Related Work	47
3.3	Methodology	49
3.3.1	Data Source	50
3.3.2	Data Collection	51
3.3.3	Themes	52
3.3.4	Variability vs. Commonality Weighting	55
3.3.5	Comparing Users	58
3.4	Results & Discussion	59
3.4.1	Validating the model	61
3.5	Limitations	66
3.6	Conclusion and Future Work	67
4	Where is <i>also</i> about time: A location-distortion model to improve reverse geocoding using behavior-driven temporal signatures	69
4.1	Introduction and Motivation	53
4.2	Research Contribution and Running Example	55
4.3	Temporal Signatures-based Location-distortion Model	58
4.3.1	Distortion Models	58
4.3.2	Activity Categories	60
4.3.3	Geosocial Check-ins	60
4.3.4	Constructing Temporal Semantic Signatures	60
4.3.5	Indicativeness of Temporal Bands	61
4.4	Distortion Functions and Weights	62
4.4.1	Spatiotemporal Distortion Functions	62
4.4.2	Weights	58
4.5	Evaluation and Discussion	65
4.6	Geosocially Distorting the User’s Location	68
4.7	Related Work	69
4.8	Conclusions & Future Work	69
4.1	Introduction and Motivation	72
4.2	Research Contribution and Example Scenario	78
4.3	Temporal Signatures-based Location-distortion Model	84

4.3.1	Distortion Models	84
4.3.2	Activity Categories	87
4.3.3	Geosocial Check-ins	88
4.3.4	Constructing Temporal Semantic Signatures	89
4.3.5	Indicativeness of Temporal Bands	92
4.4	Distortion Functions and Weights	93
4.4.1	Spatiotemporal Distortion Functions	94
4.4.2	Weights	96
4.5	Evaluation and Discussion	101
4.6	The Next Step: Geosocially Distorting the User’s location	107
4.7	Related Work	109
4.8	Conclusions & Future Work	112
5	POI Pulse: A Multi-Granular, Semantic Signatures-Based Information Observatory for the Interactive Visualization of Big Geosocial Data	114
5.1	Introduction and Motivation	117
5.2	A Data-Driven and Theory-Informed POI Taxonomy	121
5.2.1	Multi-dimensional Characterization of POI Types	123
5.2.2	Data Cleaning	128
5.2.3	Information Gain	129
5.2.4	Interactive Classification	132
5.3	Interaction and Visualization – Rasters vs. Vectors	136
5.3.1	The Tipping Point	138
5.3.2	Technology	140
5.3.3	Pre-loading Map Tiles	142
5.4	Default Behavior vs. Real-time Bursts	143
5.4.1	Default Behavior	144
5.4.2	Real-time Burst Mode	147
5.5	Conclusions and Future Work	152
6	How Where Is When? On the Regional Variability and Resolution of Geosocial Temporal Signatures Mined from Point Of Interest Check-ins	154
6.1	Introduction	157
6.2	Research Contribution	163
6.3	Raw Data and Temporal Signatures	166
6.4	Regional Variation	169
6.4.1	Significance of Placial Variations	170
6.4.2	Variability Between Categories	173

6.4.3	Concordance Between Dissimilarity Measures	177
6.4.4	Hierarchy Homogeneity	178
6.5	Cross-Cultural Comparison: Shanghai, CN	181
6.5.1	Chinese Check-in Dataset	182
6.5.2	POI Type Similarity Comparisons	183
6.6	Exemplary Investigation of Temporal Signature Differences	185
6.7	Related Work	189
6.8	Conclusions and Future Work	191
7	Conclusions	194
7.1	Discussion	195
7.1.1	Theoretical Contribution	196
7.1.2	Practical Implications	197
7.2	Limitations	203
7.2.1	Data	203
7.2.2	Methods	205
7.2.3	Conceptual	206
7.3	Future Research	207
7.3.1	Point of interest Matching, Conflation & Alignment	207
7.3.2	Geoprivacy	209
7.3.3	Real-world Activities	210

List of Figures

1.1	Linear representation of the temporal signature for Mexican Restaurant. The signature is constructed out of 168 hourly bands.	15
1.2	Temporal Signature for Mexican Restaurant aggregated by (a) Day of the Week and (b) Hour of the Day.	16
1.3	Circular representation of the temporal signature for Mexican Restaurant. The signature is constructed out of 168 hourly bands.	17
3.1	Similarity of User A, B & C to Focal User (randomly selected topics)	60
3.2	Similarity of User A, B & C to Focal User (topics with highest entropy)	61
3.3	Similarity of User A, B & C to Focal User (topics with lowest entropy)	62
3.4	Map fragment showing graduated symbols for the 30 nearest venues	63
4.1	Point-feature distance between a resort and the lounge located inside of it (screenshot from Google Maps).	75
4.2	Left: Different services list different locations for The French Press Cafe in Santa Barbara, CA. Google Maps (G), OpenStreetMap (O), Foursquare (F), Yelp (Y), Bing Maps (B). Right: Uncertainty in POI location and user location. French Press Cafe (1) and Los Arroyos Mexican Restaurant (14). The red pin marks the user's most probable location. Note that the circles of uncertainty are not drawn to scale; in actuality they would appear larger.	77
4.3	Coordinates from user's device (red pin) and nearby POI (blue markers).	82
4.4	Four possible distortion models and examples of their realization for shifting POI locations based on the temporal probability of their types (e.g., restaurant); exaggerated.	86
4.5	Daily temporal signatures for four POI categories.	90
4.6	Hourly temporal signatures for four POI categories.	91
4.7	Mean Reciprocal Rank for Four Equation and associated weight values compared to Distance-only.	99
4.8	Nearby POI locations (dark blue markers) adjusted by temporal probability at 10AM on Monday. Original POI locations visible as light blue markers. Three example locations (A, B, M) are shown in red, indicating pushed further away and green, indicating pulled closer to the assumed user location	101

4.9	Nearby POI locations adjusted by temporal probability at 11PM on Saturday. Original POI locations visible as light blue markers. Three example locations (A, B, M) are shown in red, indicating pushed further away and green, indicating pulled closer to the assumed user location.	102
4.10	Example visualization of a user's actual location (faded red pin), adjusted location (bright red pin), location uncertainty (large blue circle), Foursquare POI (blue markers) and Twitter and Instagram activity markers.	108
5.1	The pre-generated video (a) and the interactive POI Pulse system (b).	118
5.2	The weekly temporal bands for selected POI types by hour.	125
5.3	Words that make up LDA topics scaled by their relative probability	126
5.4	Ripleys K for 10 types (The y-axis value represents the difference between the observed value of K-measure at a given distance and the expected value under the CSR simulation process.)	129
5.5	Fragment of a Multi-Dimensional Scaling plot showing the upper-level classes as colors; Kruskal stress: 0.258.	136
5.6	Raster vs. Vector tile loading times (in ms).	139
5.7	Donut-pie charts showing OS category probabilities for two different POI.	147
5.8	Real-time Check-in counts for Santa Monica at 7pm on a Wednesday	149
5.9	Tweets in Santa Monica (red circles with semi-transparent white fill)	151
6.1	A fragment of the Museum type from Schema.org	159
6.2	Stacked bar plots showing amount of check-ins to each parent class as a percentage of overall check-ins. Check-ins have been split in to regions.	167
6.3	Check-in data represented as (a) a binned temporal signature by hour of the day and (b) a smoothed temporal signature. Both temporal signatures show averaged check-in behavior over 24 hours (a typical Tuesday) at a Mexican Restaurant.	168
6.4	Circular histograms depicting temporal signatures for Theme Park (a,b,c) and Drugstore (d,e,f).	171
6.5	Original Foursquare POI hierarchy supertypes by prevalence of the 100 most similar subtypes and the 100 most dissimilar subtypes.	180
6.6	Bottom-up signature-bases POI hierarchy supertypes by prevalence of the 100 most similar subtypes and the 100 most dissimilar subtypes.	181
6.7	Temporal Signatures for the POI type Theme Park in Los Angeles, New York City and Chicago, United States and Shanghai, China.	187

6.8	Temporal Signatures for the POI type Football Stadium in three cities in the United States. Note that data from Shanghai China is not shown here as no matching POI type was found.	187
6.9	Temporal Signatures for the POI type Drug Store / Pharmacy in Los Angeles, New York City and Chicago, United States and Shanghai, China.	188

List of Tables

2.1	Example activities categorized by What	31
2.2	Example activities categorized by Where	32
2.3	Activities categorized by What	35
2.4	Activities categorized by Where	36
3.1	Example Tips	51
3.2	Sample topics derived from Tip text represented as word clouds, where larger words are higher probability words for the topic.	53
3.3	Placement of actual venue based on similarity to Hypothetical Venue	65
4.1	POI Categories shown on Figure 3 with distance to device location and temporal probabilities on Monday 10 AM and Saturday 11 PM.	83
4.2	The 10 overall most indicative hours according to their information gain and the 10 least indicative hours.	93
4.3	Maximum Mean Reciprocal Rank (MRR), Maximum Sum of the Reciprocal Rank (Max SRR), normalized Discounted Cumulative Gain (nDCG), Number of POI ranked in the first position and associated weight for each Equation.	100
4.4	Comparing the results of the Distance Only method to our method which includes temporal signatures.	104
4.5	Example of Foursquare Search API query results ordered by distance and limited to 10. Known check-in location in bold face.	105
5.1	The 7 overall most diagnostic bands according to their information gain, the most diagnostic thematic and spatial bands, and the least diagnostic band.	131
5.2	F-score, Precision, and Recall for upper-level classes after the 2nd run.	134
5.3	Confusion Matrix after final class predictions.	134
6.1	Percentage of POI types that are statistically different between regions as determined by the Watson’s non-parametric two sample U 2 test of homogeneity. The results for three significance values (0.01, 0.05, 0.1) are reported.	172
6.2	Top five and bottom five dissimilar POI types based on normalized difference in Gini coefficient and split by region pairs.	174
6.3	Top five and bottom five dissimilar POI types based on normalized Jensen-Shannon Distance and split by region pairs.	175
6.4	Top five and bottom five dissimilar POI types based on normalized Earth Mover’s Distance and split by region pairs.	177

6.5	Kendall's coefficients of concordance W for pairs of regions and combinations of dissimilarity measures ($p < 0.01$ in all cases).	178
6.6	Ten highly dissimilar POI types and ten highly similar POI types selected from the U.S. Foursquare dataset. The Earth Mover's Distance was calculated between each Foursquare POI type its Chinese Jiebang counterpart. The values were normalized between the most dissimilar and most similar POI type	184

Chapter 1

Introduction

1.1 Context & Motivation

When Martin Cooper conceived of the idea for the first handheld mobile phone in the early 1970s (Cooper et al. 1975), he could not have imagined the immensity of the social impact it would have on society. I doubt that he could have anticipated how this ground-breaking idea would give rise to a world increasingly reliant on mobile technology. Through this invention emerged many of the fundamental technological concepts we hold dear: Short Message Service (SMS), camera-phones, mobile gaming and any of the 1.3 million mobile platform applications available on *Google Play* today. Perhaps one of the last possible notions in Cooper's mind was that people forty years in the future would rely on a tiny chip inside their portable mobile devices to trilaterate their geographic position based on temporal data sent at nanosecond accuracy from satellites orbiting the earth.

This same handheld device would then use a combination of wireless local area and cellular technologies to “look-up” the closest restaurant from a digital gazetteer of over 60 million Points of Interest based on these geographic coordinates, the entire task completing within seconds and the final result being the *purposeful sharing* of this information with millions of strangers around the world. While Cooper most likely has never met the young entrepreneur Dennis Crowley, Cooper’s work was most certainly crucial to establishing a platform that does exactly that.

In the late 1990s, Crowley and colleagues redefined a concept that, until then, was used solely in the transportation domain, *checking in*, to describe the social process of “self-reporting one’s position.” While the idea of sharing one’s location was not new, Crowley’s belief in the concept of geosocial *check-ins* led to the founding of numerous companies, most notably *Foursquare*. Foursquare succeeded as a company because it was established at just the right time. Location-aware technology was just beginning to make its way onto mobile devices and online social networking services, such as Facebook, were very much in the public eye. The idea of not only being able to share photos and textual updates with friends, but also share your location struck a chord with many technologically aware individuals and lead to the evolution of *Location-based Social Networking (LBSN)* or what is now being referred to as *Geosocial Networking (GeoSN)*.

From a research perspective, the ubiquitous adoption of *GeoSN* technology offers a new and exciting look at human activity and placial¹ behavior. At last count, 55 million users (Foursquare 2015a) contributed over 5 billion (Foursquare 2015c) check-ins to over 60 million Points of Interest around the world (Foursquare 2015b). These check-ins represent actual visiting behavior of real people to real Points of Interest.² Though the motivation and bias of this check-in behavior must be addressed, this is the first time in history that researchers have had access to this type of information at such a high level of spatial and temporal resolution and at such a large quantity. In many ways, *GeoSN* data is a ripple in the tidal wave of data and analytics that gave rise to a new paradigm of science, namely the *Fourth Paradigm*.

In early 2007, Jim Gray gave a presentation in which he proposed that a new paradigm of science has arisen, one of data-intensive scientific discovery. Throughout this presentation as well as in a book on the subject (Hey et al. 2009) the idea is presented that we now live in a world that is creating massive amounts of data containing a plethora of information related to everything from the eating habits of insects to political ideologies and even *geosocial* check-ins. Recent advances in computational capabilities have brought us to a tipping point in research. In

¹The term *placial* in this case parallels the term *spatial* but specifically for places rather than spaces.

²With acknowledgement of the fact that this is user-generated content and that POI, check-ins and users may be manufactured.

addition to these massive datasets, we now have the analytic tools and the computer processing power to mine, analyze and study the informational content that is presented through these data. The research presented in this dissertation is founded in this new paradigm of science in which we now have the ability to study the temporal aspects of placial human behavior.

It is important to note that the primary focus of this work is on the relationship between place types and the human temporal behavior that defines them, *not* on the value of *geosocial networking data*. While the data employed for much of this work is novel, it is critical that the reader understand that the methods and findings of this research are not specific to geosocial networking data but that these data form the basis from which to start the discussion.

1.2 The Dimensionality of Place

“Place is not only a fact to be explained in the broader frame of space, but it is also a reality to be clarified and understood from the perspectives of the people who have given it meaning.” (Tuan 1979)

Place is a difficult concept to define. Places are not items of substance in the physical sense but rather psychological constructs that mean different things to different people. Extensive research from a wide variety of disciplines has gone into defining and understanding place (Relph 1976, Tuan 1977a,b, Shamai 1991, Cresswell 2013) but to the uninitiated, the terms *space* and *place* are often used

interchangeably. In fact *Dictionary.com* defines *place* as “a particular portion of space, whether of definite or indefinite extent.” While these two concepts may be synonymous to some, a considerable amount of research has been dedicated to the dependencies between *Place* and *Space*. In the field of geographic information science, research into place has almost exclusively been approached from a spatial perspective (Winter et al. 2009). In other words, GIScientists have historically asked the question *What does space have to say about place?* While this is an interesting question, it has been researched and discussed extensively in the geospatial science literature (Agnew 2011, Goodchild et al. 2000, Kwan et al. 2003).

While space is a fundamentally defining feature of place, it is only a single dimension, or a combination of three dimensions to be exact. Other dimensions of place have been shown to have significant influence on our understanding of places and place types. For example, recent research has explored the role of *thematic space* on defining the places in our environment. A thematic dimension of place can be constructed out of user-generated unstructured natural language text with the purpose of showing similarity between places (Adams & McKenzie 2013) and the indicativeness of terms and phrases (Adams & Janowicz 2012, Adams & McKenzie 2012). Further work has explored thematic space for defining neighborhood boundaries (Cranshaw & Yano 2010, Joseph et al. 2012, Ferrari et al. 2011) and location-based recommendation models (Hu & Ester 2013, Kurashima

et al. 2010). The research presented in Chapter 3 touches on the relationship between the thematic dimension and temporal dimensions of place, taking a thematic approach to modeling user similarity through the use of geosocial check-in data.

1.2.1 The Temporal Dimension

Though existing research has shown that both space and theme are important dimensions of place, research into the ability of the temporal dimension to describe place has been lacking. Work has been done in areas such as Time Geography (Pred 1984, Miller 1991, Raper 2000), but the recent availability of user-generated online geo-content, geosocial check-ins for example, has allowed for a much deeper investigation of the temporal dimension of place. *Time* is a powerful source of information contributing not only to a better understanding of the activities in which people partake, but also the places at which these activities occur. The mobility patterns of human beings are shown to be quite repetitive (Song et al. 2010, Lu et al. 2013); We are creatures of habit. Our daily routine, the activities we do and the locations at which we do them, are quite predictable. Moreover, the times at which we conduct activities are predictable as well (Lin et al. 2012, Noulas et al. 2013). The temporal aspects of our placial visiting behavior are, in

many ways, more indicative of the activities being conducted than the geospatial locations at which they occur.

To show the importance of time let us take, for example, the following scenario. An individual is standing inside of a building in downtown Los Angeles, CA. The building contains three different types of establishments, *a nightclub*, *a bakery* and *an Italian restaurant*. In which of these establishments is the individual most likely to be? Without any information other than the geographic location of the individual, the probability of her being at any one of these establishments is equal and not at all helpful in determining her placial location. As humans our initial reaction to this question is to ask a follow up question, namely, *What time is it?* The value of knowing the time that an individual is conducting an activity should not be underestimated. Given the previous example, it is clear to the reader that *nightclubs* tend to be frequented late at night, *bakeries* in the morning and *Italian restaurants* at lunch or dinner time. An overly simplistic example, I admit, but it very clearly shows the value of the temporal dimension and the need for placial-temporal research.

1.2.2 Semantic Signatures

The amount of solar radiation reflectance of any given material on the surface of the earth varies depending on wavelength. While certain features on the earth

show similar reflectance values at specific wavelengths, research in the domain of remote sensing has shown that combining these values across different wavelengths produces unique *spectral signatures* that can be used to categorize features on the surface of the earth (Silva et al. 1971, Biehl & Stoner 1985, Hunt 1977). Analogous to this idea is that of *semantic signatures*: a unique set of semantic *bands* that can be used to describe and categorize human-defined geographic features, namely *places* in our environment (Mülligann et al. 2011, Janowicz 2012a, Adams & Janowicz 2012). In the same way that individual wavelengths are grouped into spectral bands (e.g., visible or infrared bands), so too can semantic wavelengths (e.g., spatial or temporal bands) be grouped. It is only through the combination of these bands that we can differentiate certain surfacial features in our environment. For example, in order to spectrally separate coniferous and deciduous vegetation, data from both the visible and infrared spectrums are required. Similarly, differentiating place types requires a combination of bands. For example, *Fire Stations* and *Post Offices* show similar spatial distribution patterns in a city and it is only through the inclusion of thematic and temporal bands that we can begin to differentiate them.

1.2.3 Place Types

Types and categories are psychological constructs and *types of places* are no different in this way. As humans, we have an innate desire to classify the world around us (Plato et al. 1995, Evangeliou 1988). It helps us to better organize and therein, understand the environment and our place in it. Simply stating that a certain place is a *Park* clearly brings to mind all the common (and arguably prototypical) attributes of a *Park* that went into classifying the place as such (e.g., open space, fresh air, sports, etc.). Individuals rely on these attributes to classify places into different types.

A *classical* or *Aristotelian* view on categorization focuses purely on the innate properties of an object (or place in this case) that are similar between objects. Concepts are isomorphic with respect to the properties and interrelationships of what we might call the *real* world (McCloskey & Glucksberg 1978). In contrast, a *behaviorist* approach to categorization would state that the classification of places into types should be based purely on individuals' views and behaviors towards places (Sellars 1963) and that the properties of the places themselves should have no influence on defining place types. Clearly, this work does not adhere strictly to either of these approaches (or any of the other theoretical categorization approaches), but rather sees value in merging different aspects of each view to establish a more general method for determining place types. Recent work in ontology

engineering mirrors this argument through merging of top-down and bottom-up approaches to categorization (Janowicz 2012a). Further discussion on this can be found in Chapter 5.

Throughout this work, places will be discussed in terms of their *type*. While the temporal dimension of place at an *instance* level is of interest, there tends to be very little difference between the temporal signatures of individual places within a type (e.g., *La Super-Rica Taqueria* vs. *Lilly's Tacos*). In contrast, the temporal signatures of *Mexican Restaurants* and *Eastern European Restaurants* are shown to be statistically significant. Furthermore, discussing places at an instance level only allows for statements to be made about the individual places themselves, rather than the *type* of place as a whole. One of the more important contributions of this research is the temporal inferences that can be made about a specific place based purely on knowledge of the place *type*.

1.3 Research Contribution

The primary contribution of this research lies in demonstrating how places are defined through a better understanding of temporal activity behavior. The times that we choose to visit certain places are uniquely informative. Through a diverse set of geocomputational models and data mining techniques this work indicates

how temporal data can be employed to differentiate between *types* of places in our environment. Furthermore, the results of this research present the argument that the temporal dimension is *the most indicative* placial dimension for classifying places by type.

Research Questions

Each of the individual chapters in this dissertation explore the relationship between *Place* and *Time* from a different perspective. Each chapter asks and addresses its own distinct set of research questions. In this section, the common thread or *leitmotiv* is introduced along with the overarching research questions that stem from this thread. As this dissertation is formed as a structured cumulation of published works, understanding how the different chapters fit together is of the utmost importance.

Understanding the relationship between the real-world activities in which individuals partake and the online actions of these individuals is an important first step in this work. Before it can be stated that online check-in data is representative of actual human activities, it must first be shown that there is indeed a correlation between real-world activities and digital ones. This leads to the following research question.

R1 *What is the strength of the relationship between real-world and online activities? Specifically, what types of real-world activities are men-*

tioned online, how often are they mentioned and what is the temporal relationship between an online social post and the real-world event?

Given knowledge of this relationship, it is necessary to look at the places where these activities occur as well as the people who conduct these activities. By bringing in the thematic dimension of place it is possible to explore the nuanced differences between places as well as the (dis)similarities between the place-goers themselves. The findings of this research suggest that the time at which people choose to conduct activities has a crucial role in assessing the similarity between places and people. This fits into the *leitmotiv* as it shows that *time* is very much linked with place type and the visiting behavior of individuals. The strength of this link and the influence of time are the focus of this research question.

R2 *When determining the similarity of individuals based on the descriptive language of the places that they visit, does the time at which an individual visits a place significantly influence the accuracy of such a model?*

Keeping in mind that the *spatial dimension* is a highly defining dimension of place, what can the temporal dimension add to this? While the previous research question focused on the intersection between thematic (natural language descriptions) and temporal dimensions, this question targets the interaction of the *spatial* and temporal dimensions.

R3 *Can the inclusion of a temporal component enhance existing spatial-only geolocation methods? Charged with the task of reverse geocoding*

geographic coordinates, can the temporal descriptive aspect of a place be merged with the spatial layout of a region in order to increase placial accuracy?

The previous two research questions focused on the relationship between two of these dimensions, namely *Theme & Time* and *Space & Time*. The next step along the common thread of this research is to explore the value-add of each dimension in defining place types. It has been shown that each of these dimensions is of great importance in our understanding of place type, but how do they compare?

R4 *How important is time in defining the difference between places of interest? Typically the dimension of space is taken to be the most indicative of place types and recent research has shown that theme plays an important role. What role does the temporal dimension play in defining place types and how does it compare to the indicativeness of spatial and thematic attributes?*

Last, provided that regionally-aggregated temporal check-in behavior has been shown to be valuable in defining place types, the regional specificity should be examined.

R5 *Does temporal check-in behavior vary by region? In other words, in order to be useful for determining place types, should regionally specific temporal signatures be constructed? Furthermore, do some POI types vary by region while others do not?*

1.4 Methods

Many methods were used in collecting and analyzing data as part of the research presented in this work. This section outlines a few of the overarching methods that were employed.

Given the importance of the temporal dimension, how do we model temporal behavior and show that there are clear differences in temporal behavior at different places? The striking increase in user-generated geo-content available online has provided researchers with access to a rich corpus of geosocial check-in data. While limited by the data-silo nature of many of the location-based social networking application providers, *Application Programming Interfaces (APIs)* allow modest access to much of the content contributed to these platforms.

In this work, one of the primary ways in which temporal behavior is investigated, is through the collection, analysis and reformation of geosocial check-in data into *Temporal Signatures*. Drawing on previous research in this area (Ye, Janowicz, Mülligann & Lee 2011, Mülligann et al. 2011, Noulas et al. 2011), *temporal signatures* fit within the concept of *semantic signatures*. Temporal signatures were constructed from hourly check-in data from a variety of regions and categories. The use of temporal signatures is ubiquitous throughout the work presented in

this dissertation and necessary for understanding how places differ in the temporal aspects of their activities.

Rather than simply describing the data, it is often more useful to visually present this information. For example, these data can be visualized either as a linear plot of 168 bands (Figure 1.1) or split into *Daily* and *Hourly* patterns (Figure 1.2). Given the inherent cyclical nature of time, visualizing the temporal signatures linearly may not be the most appropriate method. An important distinction here is that a linear approach makes the assumption that Sunday and Saturday are at opposite ends of the temporal spectrum while in reality they are adjacent to one another. Taking this into consideration, it may be more suitable to represent a temporal signature as a circular histogram (Figure 6.4).

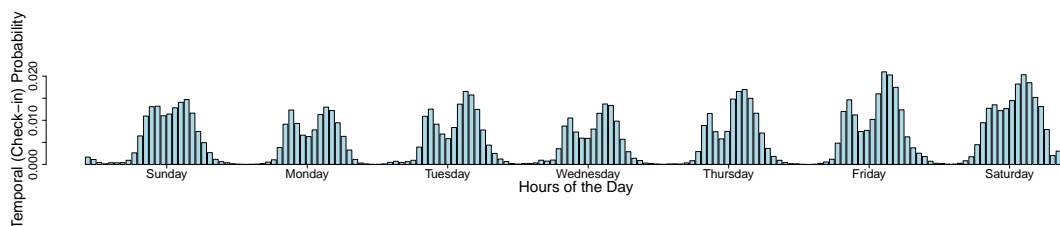


Figure 1.1: Linear representation of the temporal signature for *Mexican Restaurant*. The signature is constructed out of 168 hourly bands.

In addition to the construction of temporal signatures, numerous other methods were used in the analysis, visualization and presentation of results. *Shannon (Information) Entropy* and the *Information Gain* associated with the the addi-

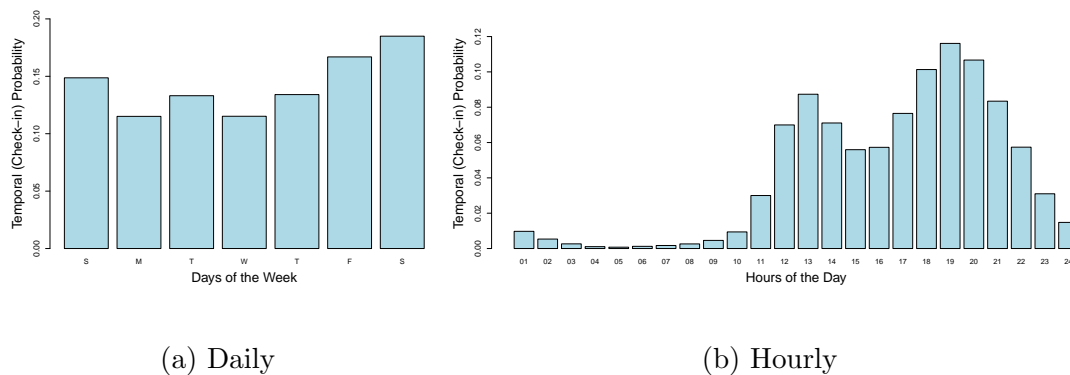


Figure 1.2: Temporal Signature for *Mexican Restaurant* aggregated by (a) *Day of the Week* and (b) *Hour of the Day*.

tion or removal of temporal/semantic bands were employed extensively in this research. As the temporal variability and uncertainty of place types is one of the primary focuses of this work, these approaches proved to be invaluable in both quantitatively presenting the differences between places, but also permitting the similarity of place types to be assessed.

(Dis)similarity measures such as *the Jensen-Shannon Divergence*, *Earth Mover's Distance* and *Watson's U^2 Test of Homogeneity*, were exploited in this research to show the differences and similarities between place types and the individuals that visit those places. Results of these similarity analyses were evaluated through various rank statistical methods and validated through concordance and correlation statistics. Further details on the methods are presented in each Chapter.

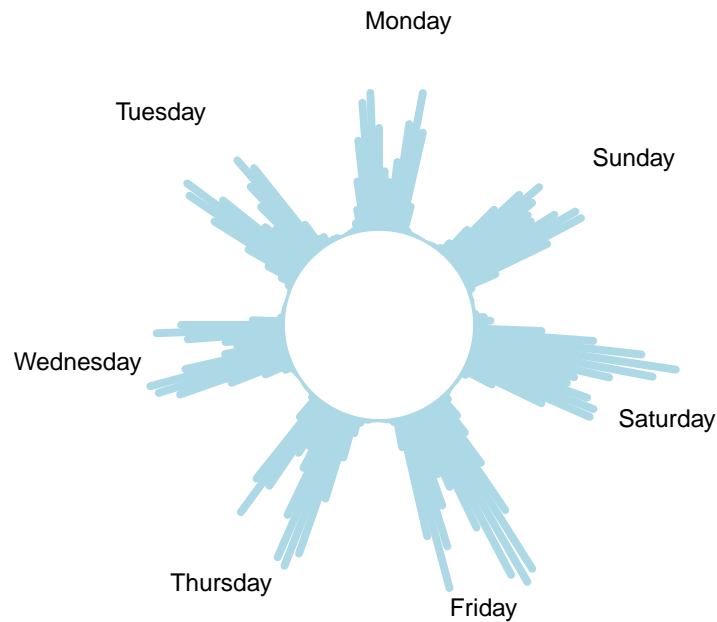


Figure 1.3: Circular representation of the temporal signature for *Mexican Restaurant*.

The signature is constructed out of 168 hourly bands.

1.5 Outline

The remainder of this dissertation is summarized and outlined in the paragraphs below.

Prior to exploring the power of user-generated content, Chapter 2 first explores the relationship between online contributions and the real-world activities that they represent. What types of activity, location and temporal information do people contribute online and how do these contributions relate to real-world activ-

ities? This chapter shows that there is a strong correlation between the activities that people say they will do and the activities the individuals actually complete. Additionally, it is shown that the time the activity is contributed online and the time the activity is completed are also related.

In Chapter 3, the role of user-contributed unstructured textual data (tips and reviews) is examined as a means for calculating similarity between users based on the places that they visit. While it is shown that textual data analyzed through a topic modeling approach is able to access the nuanced differences between different places, it is also shown that time plays a significant role in constructing a user-similarity model. The time at which a user completes an activity at a given place is fundamental in the construction of an accurate user-similarity model.

Next, Chapter 4 investigates the relationship between time and space through explicitly merging spatial and temporal dimensions. Framed by the real-world problem of geolocation, this work shows that it is possible to augment existing spatial-only reverse geocoding through the inclusion of a temporal component. Through a model that (theoretically) distorts space by a factor of time, this work shows that the ability to correctly identify a place significantly improves through the inclusion of temporal bands. Multiple methods for combining space and time are proposed and evaluated.

Focusing on the greater Los Angeles area as a sample region, Chapter 5 uncovers the multidimensionality of place in order to describe and categorize Points of Interest. Various classification and machine learning techniques are used with semantic bands of space, time and theme to develop an interactive mapping application that depicts the pulse of Los Angeles. In combining these three dimensions of place, it is shown that the temporal dimensions (temporal signatures) rank much higher in terms of information gain, indicating that time plays a much more important role in defining the pulse of the city than do other dimensions of place.

The previous chapters have shown that the dimension of time has a significant role to play in the idea of Place. Up until this point the temporal dimension has been depicted as a single set of temporal signatures used to describe place types regardless of their spatial location. In Chapter 6, the regional variability of place types is investigated through their temporal signatures. It is shown that some place types are aspatial, (regionally invariant) while others vary quite significantly with a change in region. In other words, temporally, not all regions are created equal.

Finally, Chapter 7 presents a general discussion on the findings of this work. Conclusions are presented as well as a section outlining the limitations of this research. Last, the *Future Work* section presents possible directions for future research and further studies.

Chapter 2

What, When and Where: The real-world activities that contribute to online social networking posts

In this chapter, the complex relationship between online social networking contributions and the real-world activities that they represent is investigated. The research introduced here investigated the types of real-world activities that are reported online and the spatio-temporal aspects of these posts. The results show that certain unique activity types are more likely to be shared through online sources while more common activities are more likely to go unmentioned. In addition, a temporal link between real-world activities and their related online contributions is discussed. Given that future chapters rely considerably on data contributed by users of online geosocial networking applications, this chapter is essential in that it sets the foundation on which future chapters are based.

Chapter 2. *What, When and Where: The real-world activities that contribute to online social networking posts*

Peer Reviewed Publication	
Title	What, When and Where: The real-world activities that contribute to online social networking posts
Authors	Grant McKenzie ¹ , Kathleen Deutsch ¹ , Martin Raubal ²
Institutions	¹ Department of Geography, The University of California, Santa Barbara, ² Institute of Cartography and Geoinformation, ETH Zürich
Venue	Proceedings of the International Workshop on Action and Interaction in Volunteered Geographic Information (ACTIVITY) at the 16th Association of Geographic Information Laboratories in Europe (AGILE) Conference
Location	Leuven, Belgium
Editors	Peter Mooney, Karl Rehr
Submit Date	March 29, 2013
Accepted Date	April 21, 2013
Presentation Date	May 14, 2013
Copyright	Grant McKenzie, Kathleen Deutsch, Martin Raubal

Abstract

Understanding the relationship between online social networking posts and real-world activities is key for many areas of research including activity prediction and recommendation engines. This paper presents the results of an exploratory study conducted with the purpose of extracting the types of an activity that are reported in an online post. The spatio-temporal components of activities are explored as well as the categories of the activities being conducted. Results suggest that activities that occur with less frequency are more likely to contribute to online action than those that are more routine.

2.1 Introduction

As Online Social Networking (OSN) applications grow in influence and user base, a multitude of questions have arisen focusing on the relationship between our real world and the online virtual one. The OSN application *Facebook* recently reported an average of over 552 million daily active users ([Facebook 2012](#)) with almost as many interacting with the application via a mobile device. These statistics indicate that OSNs have become fully integrated into our everyday lives, though questions remain as to the extent of this integration. Given the ubiquity of OSN applications and our desire to increase social worth, the propensity to

perform an activity in the real world and then broadcast the accomplishment of that activity through an OSN must be investigated. A better understanding of the relationship between our real-world accomplishments and our online social lives will have significant impacts in research areas ranging from activity behavior to location prediction systems.

A better understanding of the relationship between online activities and non-online activities should also direct the studying of the social structures of daily life. Social factors have been shown to have a significant impact on the activities we perform in the real world ([Ahas et al. 2007](#), [Elliott & Urry 2010](#), [Páez & Scott 2007](#)), but the question remains as to how that translates to the virtual world. Current location-based social networks (LBSN) such as *Foursquare*, *Yelp* and even *Facebook* allow users to broadcast their real-world locations along with status updates and photos related to the activities they are conducting at these locations, and with whom. As the users of these applications grow in number, it becomes more apparent that the desire to share one's activities is at least partially socially motivated. Discussions related to social capital and social worth emerge this notion of activity broadcasting ([Pultar et al. 2010](#)). What effect does publicly announcing your activities have on your social relationships? The connection between the types of activities that one perceives to increase social worth and the types of activities that actually do is undoubtedly an area of future

research. Taking one step at a time, our research offers a first step in exploring this connection and the activity types.

This paper presents a feasibility study that explores the *types, locations* and *time frames* of real-world activities and the likelihood that these activities are reported on an online social network. Using activity surveys and online activity tracking, this research offers an insightful view into the types of daily interactions and events that result in an online announcement. This exploratory study categorizes real-world activities based on established research guidelines and statistically determines which categories of activities are most prone to producing an OSN interaction. Given the spatio-temporal nature of activities (an activity must occur in space and time), we predict that the location and time of an activity play important roles in the types of activities that are reported.

The remainder of this paper is organized as follows. Section 2.2 introduces related work on the relationship between real and online activities. In Section 2.3 we describe the data collection and statistical methods used. Section 2.4 presents the results of our feasibility study while Section 2.5 and 2.6 discuss a number of limitations and offer conclusions and next steps for this area of research.

2.2 Related Work

2.2.1 Activity Categorization

Chapin defines activities as “classified acts or behavior of persons or households which, used as building blocks, permit us to study the living patterns or life ways of socially cohesive segments of society” (Chapin 1974). These acts of behavior may be categorized into a number of different classes. Researchers investigating time use and activity patterns have produced a plethora of classification schemes (Axhausen 2008, Parkka et al. 2006) each focused on a specific field. In addition, guidelines and suggestions have been developed to aid in classification of activities in endeavors such as activity based travel demand modeling (NCHRP 2008). Activities may be classified based on anything from their frequency, duration and sequence to social interaction (Golledge 1997). Additionally, the location at which an activity takes place may also form the framework on which activities are categorized.

Chapin developed a classification scheme for aggregating activities into two levels. He categorized activities based on a glossary of approximately 230 activity codes (Chapin 1974). Szalai et al. (1972) built on Chapin’s classification and produced a simpler system that they presented in their multinational time-budget study in 1972. These categories ranged from work to cultural events to trans-

portation and travel. It is upon much of this work that the categories defined later are based.

2.2.2 Online Social Networking

In the area of online social networking, it has been shown that interaction through online social networking applications such as *Facebook* have a tendency to increase social capital (Ellison et al. 2007). Not surprisingly, measures and types of social capital change with the type of online social interaction. Research in the area of social networks has also explored the role of “friendships” both on- and offline. While there is definite overlap in “friends” that one has on an online social networking application and their real-world social network, the research suggests that online social networks play a role in strengthening different aspects of offline friendships (Subrahmanyam et al. 2008). This research takes the next step in understanding the activities that motivate these online interactions.

2.2.3 Activity Prediction

One of the primary motivations for conducting this research is to recognize the relationship between online posts and real-world activities. An understanding of this relationship can offer significant insight into activity prediction. A number of studies have explored the usefulness of social network data in predicting future

activities (Chang & Sun 2011, Cheng et al. 2011) while others have investigated the effectiveness of travel trajectories in determining a user’s hometown or place of residence (Liben-Nowell et al. 2005, Backstrom et al. 2010).

Though most of these studies involve the exploration and use of online communities, surprisingly few have ground-truthed the data with real-world travel or activity data. While it may be possible to predict an individual’s location based on her previous location history, was this person at any of the locations previously reported by the social networking application?

Previous studies have explored the idea of activity-based ground-truthing given real-world social survey information (Ahas & Mark 2005), but to our knowledge, very little, if any research has investigated the position of online social networking data in determining an individual’s real-world location (McKenzie & Raubal 2012).

2.3 Methods

This section presents the methods used for data collection and analysis.

2.3.1 Data Collection

Participants

In order to assess the relationship between Facebook posts and real-world activities, a representative sample of online social network users was required. A total of 30 participants (15 female) between the ages of 20 and 45 (mean = 28.6, stdev= 5.2) were asked to participate in the research in return for a \$20 USD gift card. These participants were sampled from Facebook using a *snowball* sampling method (Biernacki & Waldorf 1981). Social acquaintances of the principal researcher were contacted initially with the offer to participate in the research study. Additional participants were recruited through word-of-mouth interest from the initial contacts. The sole requirement for participation was that each participant have a history (the two weeks before the study) of posting to Facebook a minimum of once a day on average. This ensured a reasonable amount of data for analysis and removed bias due to lack of participation.

Collection Methods

The study requested participation over a continuous 3-week period by completion of two components. First, study participants were asked to record their activities (to the nearest hour) through a daily activity diary available online. Participants were asked to record the start time, end time, location and description

of the activity performed. Both the location and description fields were presented as free-text boxes, allowing participants to use their own words when describing the location and activity. The instructions asked participants to be as detailed as possible when recording this information since the final goal of the location field was the ability to geocode.

Second, participants were required to install an application developed by the research team, allowing access to basic Facebook profile information¹ and social activities. The application gathered profile information on the specific participant (originally provided to Facebook by the user) as well as posts made by the user on her own wall (typically only visible to friends). By granting access to this application, online social network data was downloaded for each participant over the same 3-week period in which they were completing the daily activity diary. During the course of data collection, access to 2 of the participant's Facebook accounts was interrupted resulting in incomplete datasets, reducing the number of participants to 28. The study took place over a three-month period from October to December 2011.

¹Profile information consisted of name, gender, birth date, location, email, education, hometown and username. It is important to note that Facebook considers email as the only required field.

2.3.2 Activity Categorization

From the 28 participants in the study, 3,198 activities were recorded to the daily activity diaries (mean=114, stdev=37.6). As discussed previously, the data consisted of activity and location descriptions entered as free-text by participants. In order to include these data in a statistical model, it was necessary to categorize each of the 3,198 activities recorded by the 28 participants. The two levels of classes into which each activity was grouped consisted of the type of activity (What) and the location of the activity (Where). These classes are discussed in detail in the next sections.

Categorization of activities was achieved through manual processing. The activities were anonymized and 3 researchers independently categorized each activity in both the type and location classes. The principle researcher then reviewed the results in order to ensure consistency in coding. When a disagreement of activity categorization was discovered, conflicts among classification were resolved. The purpose of this multiple-categorization was to remove as much bias as possible from the categorization process.

Class: *What*

The *What* class conveys the type of activity the participant was performing. It is important to note that an activity can be (and often was) classified in multiple

categories. Table 2.1 lists the general categories established based on the data contributed by users. Example activities are shown as well.

It should be noted that the difference between local and distance travel is primarily within a city (commute) and between cities respectively.

Category	Example activity
Eating	Eating Dinner @ Home
Drinking (non-alcoholic beverages)	Yumm Coffee @ Starbucks
Drinking (alcoholic beverages)	Drinking beer @ OHares
Fitness / Active	Bike Ride @ SB Bike Path
Watch TV / Play Video Game / Surf the Web	Watching Football @ Buddys house
Local Transportation	Bus to work
Distance Transportation	Catching my flight to Toronto @ YVR Airport
Shopping	Christmas shopping @ Brentwood mall
Errands	Dentist Appointment
School	Classes, working @ UCSB
Work	Sound Design @ Work
Self Maintenance	Shower @ Home
Sporting Event	Go Canucks @ Rogers Arena
Cultural Entertainment (Concert, Museum, Play, etc.)	Sounds of Vienna Concert @ Kursalon, Vienna
Movie Theater	Mission Impossible @ South Edmonton Theaters
Vacation	Relaxing Vaca @ Victoria
Party / BBQ	Christmas Party @ Work
Other	Visiting @ Friends house

Table 2.1: Example activities categorized by *What*

Class: *Where*

The location at which each activity took place was categorized within the *Where* class. Unlike the *What* class, activities were categorized into a single location type (an activity could not take place at multiple locations). From a data-entry perspective, participants were asked to enter the name of a single location for each activity they performed. The example given was a spatial point described as an intersection of two streets. Actual entries from participants ranged from residential address to city level precision. For this reason, it was decided to simply group the *Where* activity tags into very general categories. Table 2.2 shows the categories as well as an example activity.

Category	Example activity
Home	Dinner with Family @ Home
Work/School	Working @ Westjest campus, Calgary
Transportation/Trip	Bus Home @ Waterfront Station
Other	Grocery Shopping @ West 4th & Vine

Table 2.2: Example activities categorized by *Where*

Class: *When*

The temporal component of activity posts is also worth exploring. The online activity diary required that participants enter both a start and an end time to their activities. The diary allowed participants to specify times to the nearest 30

minutes. Given the temporal bounds of these real-world activities, we are able to compare them to the online posts to which they are related.

2.3.3 Facebook Posts

For the purposes of this research, we define *Facebook Post* as a contribution of digital content to a user's social "wall." To refine it further, our research only analyzed textual input in the form of a status update made by one participant on her own wall. The total number of status updates for all users was 352 over the three-week period (mean=12.6, stdev=9.9). Of the 3,198 activities entered through the activity diary by the 28 participants, 75 of the activities were linked to one or more online post. The linking of these activities to posts was again achieved by manually matching an individual's online contribution to their real-world activity.

2.3.4 Analysis

The purpose of this research is to determine what types of activities at which locations are most likely to result in a post on the online social network Facebook. In order to achieve this goal, the two classes above were analyzed independently based on descriptive statistics.

Analysis of *What*

In exploring the original categories, we combined a number of the categories that were often tagged together. These categories roughly follow the *Time Budget Classification of Activities* framework developed by Szalai et al. (1972) and built upon Chapin and Logan’s activity classes (Chapin & Logan 1969). Merging activity types was also a result of the frequency of matched-tags based on the participant-contributed descriptions.

Eating and drinking (non alcohol related activities) were grouped together as were common activities often done in sequence at home (sleeping, watching television, showering, etc.). *Cultural entertainment* was also combined ranging from visits to a museum to concerts and attending sporting events. Lastly, *shopping and running errands* were combined as the two are often done together, or could be synonymous. Table 2.3 lists the number of occurrences categorized from the self-reported activity diaries as well as the number of Facebook posts in which an activity of that category was reported.

Analysis of *Where*

While the categories related to *What* type of activity were aggregated, the *Where* categories were not. This was primarily due to the vagueness of locations reported in the self-reported activity diaries. As was the case with the categorical

Category	Count	FB Post	Percentage reported on FB
Distance Travel	60	11	18.33
Vacation	30	3	10.00
Entertainment (Concerts, Museum, Theater, Sporting Events)	41	3	7.32
Drinking alcohol	96	5	5.21
Party / BBQ	70	2	2.86
Fitness	161	4	2.48
Shopping & Errands	453	9	1.99
Sleeping, watching TV, video games, browsing internet, self maintenance	969	17	1.75
Eating & Drinking (non-alcohol)	1017	17	1.67
Local Travel	240	4	1.67
School & Work	851	14	1.65

Table 2.3: Activities categorized by *What*

What data, each location category resulted in a number of online posts. Table 2.4 displays these data in a format mirroring the *Where* categories.

Analysis of *When*

Given the temporal bounds of a real-world activity, as provided by a participant, we are able to evaluate the relationship to the related online post in terms of time. As mentioned in previous sections, of the 3,198 daily activities, 86 Facebook posts were judged to be directly related to participants' real-world activities. It

Category	Count	FB Post	Percentage reported on FB
Transportation/Trip	298	14	4.70
Other	1235	38	3.08
Home	1022	16	1.57
Work/School	642	7	1.09

Table 2.4: Activities categorized by *Where*

is important to note that these 86 posts include duplicate real-world events as a number of participants posted more than once regarding a specific activity.

In exploring the relationship between real-world activities and online Facebook posts from a temporal perspective, we ask, how close to an activity does a post occur in relation to the start of an activity? The data shows that on average a post occurs approximately 9.31 hours before the start of an activity, with a median of 4.78 and standard deviation of 32.89 hours.

2.4 Results & Discussion

2.4.1 What

Both *Distance Travel* and *Vacation* show the highest percentage of posts related to real-world activities. Activities categorized as *Drinking Alcohol* and *Cultural Entertainment* presented lower percentages indicating that they only slightly influence the likelihood of an OSN user posting online. Not surprisingly, the more

mundane activities such as *Sleeping, watching TV and Self Maintenance* are the least influential in contributing to an online post, along with *Shopping & Errands*. More surprising is the fact that activities related to *BBQs & Party* had little to no sway on online posts. This could be largely due to the small sample size upon which this model was built.

The categories that demonstrated the highest influence on Facebook posts both related to events that occur less frequently than most other categorized occurrences. It follows that OSN users feel some increased sense of social value from traveling, perhaps as it presents a divergence from their average routine. Travel as it relates to vacation (and work for that matter) symbolizes financial stability, recreational activities and freedom from our daily routine that most cultures desire. This is slightly mirrored in the activities related to *Cultural Entertainment* such as concerts and sporting events.

2.4.2 Where

Again, we explore these data based simply on the number of online posts stemming from each location category. In looking at table 4, we see that the vast majority of activities is categorized as either *Home or Other* with both *Transportation* and *Work* producing less. Given the sheer number of activities categorized as *Other* (occurring outside of home, work or travel), it is not surprising to see it

resulting in the most number of online posts. However, it is interesting to note that the percentage of influence is much higher for *Other* and *Transportation* than *Home* or *Work*. Again, as with the *What* categorization, this increased influence on online contributions reflects a divergence from the routine activities that most likely occur at home and work. The *Transportation* category echoes the results from the *What* categorization showing that *Distance Travel* has a large influence on posts. This fits with our preconceived notions that activities done at work, school and home are not as interesting as those completed at other locations and therefore less worthy of being broadcast to our social circle.

2.4.3 When

The analysis of the temporal relationship between posts and activities indicates that on average, posts occur 9.31 hours before the start of an activity. In total, 79% of posts were written before the activity, ranging from approximately 9 days prior to the activity, to the exact start time of an activity (remember we are dealing with 30 minute resolution). In fact, the vast majority (91%) of posts in our sample set are within 24 hours of the start of an activity. These results suggest that access to an individual's online activity may offer insight into that individual's future real-world activities. These results imply significant value to research in the areas of activity prediction and recommendation engines.

2.5 Limitations & Next Steps

The method and results presented in this paper have a number of limitations. Given that this research presents an exploratory study, the number of both participants and Facebook posts made by those participants is small. This small sample size should be taken into account when interpreting the results. For example, the category *Party & BBQ* shows 70 occurrences and only 2 posts tagged to this category. The first step in expanding this to a full study would be to increase the number of participants as well as the duration of the study. Moving from a participant pool of 30 individuals to 300 (for example) would allow for further statistical exploration using existing as well as new and more robust methods. Given this increased sample size, next steps would involve evaluating variable correlations and employing a binary choice model (Lee 1979) to explore the influence of the different categories. It is expected that an increased number of participants would act to strengthen the results presented in this paper.

The self-reported activity survey should also be enhanced to include some level of participant tracking (e.g., GPS enabled mobile phones) along with strongly typed category choices for activities and locations. The free-text entry method undertaken in this research required considerable manual classification that could be avoided with standardized drop-down lists, or multiple-choice with an open

ended “other” category. A more extensive, formalized list of categories combined with the increased number of participant activities would offer more insight into types of activities broadcast through an OSN. Alternatively, natural language processing approaches could potentially be employed to extract location information from the updates themselves. Machine learning techniques and geographic information retrieval methods could both be of considerable value to this area of research.

Additionally, a wider range of social networking applications will provide more breadth to the study and enable the generalization of results on a larger scale. Facebook is by far the most ubiquitous online social network today and the perfect source for a preliminary study such as this, but future research in this area can make use of the abundance of online applications surfacing every day.

2.6 Conclusions

This paper presented methods and results from an exploratory study investigating the relationship between daily activity schedules and online social networking posts. This first step showed that it is possible to conduct a study that explores the interaction between the real world and the virtual one. While this is an introductory stride in the much larger research agenda of ground-truthing

online social networking data, the methods produced encouraging results. This study suggests that activities that occur with less frequency are more likely to contribute to online action than more routine activities. These results also intimate that users place high social value on activities that diverge from the norm such as vacations and travel. In summary, this exploratory study offers encouraging results for understanding the relationship between the real and the online social world.

Chapter 3

A Thematic Approach to User Similarity Built on Geosocial Check-ins

In this chapter, a model for assessing user similarity based on placial visiting behavior is constructed. Through the use of geosocial *check-in* trajectories, the similarity between individuals is measured based on the topics that are extracted from unstructured textual reviews contributed online to the places that they visit. The model can be adjusted to establish similarity based on *common* or *variable* topics. The temporal dimension is featured in this work through the order and time of places within a person's daily trajectory. Ultimately, the time at which someone chooses to visit a place is crucial to the effectiveness of a user similarity model based on placial behavior.

Peer Reviewed Publication	
Title	A Thematic Approach to User Similarity Built on Geosocial Check-ins
Authors	Grant McKenzie ¹ , Benjamin Adams ² , Krzysztof Janowicz ¹
Institutions	¹ Department of Geography, The University of California, Santa Barbara, ² Centre for eResearch, The University of Auckland
Venue	Geographic Information Science at the Heart of Europe - Proceedings of the 16th Association of Geographic Information Laboratories in Europe (AGILE) Conference
Location	Leuven, Belgium
Editors	Danny Vandenbroucke, Bndicte Bucher, Joep Cromptvoets
Publisher	Springer
Pages	39-53
Submit Date	November 15, 2012
Accepted Date	December 29, 2012
Presentation Date	May 14, 2013
Copyright	Reprinted with permission from Springer Publishing

Abstract

Computing user similarity is key for personalized location-based recommender systems and geographic information retrieval. So far, most existing work has focused on structured or semi-structured data to establish such measures. In this work, we propose topic modeling to exploit sparse, unstructured data, e.g., tips and reviews, as an additional feature to compute user similarity. Our model employs diagnosticity weighting based on the entropy of topics in order to assess the role of commonalities and variabilities between similar users. Finally, we offer a validation technique and results using data from the location-based social network Foursquare.

3.1 Introduction

Online social networking (OSN) offers new sources of rich geosocial data that can be exploited to improve geographic information retrieval and recommender systems. OSN platforms such as *Foursquare*, *Twitter*, and *Facebook* have taken advantage of the popularity of GPS-enabled mobile devices, allowing users to geotag their contributions, thus adding spatiotemporal context to their social interactions.

This increase in social networking through portable devices has resulted in a shift from location-static updates to location-dynamic interactions, freeing online communication from the clutches of the desktop and immersing it in our mobile lives. Social network users post updates on the go from anywhere in the world, be it from a restaurant, mountain top, or airplane. These data are having a profound impact in the study areas of human mobility behavior, recommendation engines, and location-based similarity measurements.

The abundance of data published through online sources provides an exceptional foundation from which to investigate user similarity. To many users of these OSNs, the benefits of allowing access to this personal information is worth the cost of privacy. From a research perspective, these data offer an unprecedented opportunity to observe human behavior and design new methods for exploring the similarity between individuals. Studying similarity is important for several reasons. First, it can be used to suggest new contacts and thus, enrich the social network of a user. Second, as similar users are more likely to share similar interests, user similarities play a key role in recommender systems (Matyas & Schlieder 2009) and geographic information retrieval (Jones & Purves 2008). For instance, the *Last.fm* music platform offers social networking functions by which users can explore their *musical compatibility* with others and listen to their personalized radio stations. Third, and of most importance for our work, the information

available about users, their locations, and activities is still sparse. User similarities can be exploited to predict *types* of activities and places preferred by a user based on those of users with similar preferences.

So far, most work on user similarity has mainly focused on structured, e.g., geographic coordinates, or semi-structured, e.g., tags and place categories, data. Unfortunately, these data are often unable to uncover nuanced differences and similarities. For instance, two users may frequently visit places tagged as *bar* and rated with a *Yelp* price range of \$\$\$. However, unstructured, textual descriptions reveal that only one of these users constantly visits places that offer pub quizzes. In this paper we suggest exploring location-based social networking (LBSN) data to enhance current user similarity measures by focusing on unstructured data, namely *tips* provided by users. This approach explicitly focuses on the non-spatial components of user-contributed data, utilizing *topic modeling* together with *diagnosticity weights* determined by the entropy of different topics. The temporal properties of a user's trajectory are also included when calculating user similarity. Our initial results show that the similarity between individuals is not uniform throughout the day. Thus, instead of generalizing similarity simply to the user level, we propose a method for assessing similarity on an activity-by-activity basis, exploiting the temporal as well as the spatial attributes of a user's trajectory.

The remainder of the paper is organized as follows. In Section 3.2, we discuss related work on user similarity and location-based social networks. Section 3.3 focuses on data mining and the methods used for defining user similarity. In Section 3.4, we present results based on actual user data. Section 3.5 discusses a few of the limitations we faced in conducting this research and Section 3.6 presents our conclusions and points out directions for future work.

3.2 Related Work

Assessing user similarity has become an important topic in information retrieval and recommender systems over the past few years. The motivations for developing user similarity measures range considerably, from recommendation systems (Guy et al. 2009, Horozov et al. 2006) and dating sites (Hitsch et al. 2010) to location and activity prediction (Lima & Musolesi 2012, Noulas et al. 2012).

A number of recent studies have focused on measuring user similarity through trajectory comparison (Lee et al. 2007, Li et al. 2008, Ying et al. 2010). Lee et al. (2007), explore a geometric approach to trajectory similarity by exploiting three types of distance measures in order to group trajectories. While their *Partition-and-Group* framework is unique, it is limited to the geospatial realm, overlooking the types of activities and social information related to the activity locations.

Similarly, [Li et al. \(2008\)](#) focused on the spatial components of user trajectories. Their method employs hierarchical trajectory sequence matching to determine similar users. Making use of GPS tracks, Li et al. extract *stay points* at which a user's activity is determined based on the affordances of a specific location.

While the above methods measure user similarity based on geospatial aspects of user trajectories, we argue that an understanding of the semantics of an activity space are essential. [Ye, Shou, Lee, Yin & Janowicz \(2011\)](#) investigate the concept of semantic annotations for venue categorization. In developing a semantic signature for a categorized place based on *check-in* behavior, similar, uncategorized places could be discovered. This concept of semantic signatures may also be applied to assessing user similarity through semantic trajectories. In this vein, [Ying et al. \(2010\)](#) measured semantic similarity between user trajectories in order to developed a *friend recommendation system*. This work focuses on the type of activities completed by each user and the sequence in which these activities take place. Akin to the *stay point* work presented by [Li et al. \(2008\)](#), the authors focus on *stay cells* and obtaining a semantic understanding of the types of activities conducted within the cells. From there, a semantic trajectory is formed and patterns are assessed and compared between users.

Activity prediction research can also benefit from exploring user similarity. Based on check-in data gathered through *Foursquare*, [Noulas et al. \(2012\)](#) exploit

factors such as transition between types of places, mobility flows between venues and spatial-temporal characteristics of user check-in patterns to build a supervised model for predicting a user’s next check-in. This method, while exploring previous check-ins across users, does not assess similarity between users in predicting future locations, an aspect that our research suggests is beneficial. Traditional work in collaborative filtering (e.g., Amazon recommendations) has also focused on measuring user similarity, but typically concentrates on ”structured” data such as numerical (star) ratings (Linden et al. 2003, Herlocker et al. 2004).

Recently, Lee & Chung (2011) presented a method for determining user similarity based on LBSN data. While the authors also made use of check-in information, they concentrated on the hierarchy location categories supplied by *Foursquare* in conjunction with the frequency of check-ins to determine a measure of similarity. By comparison, our approach is novel in that it makes use of an abundance of unstructured descriptive text (tips) provided by visitors of specific venues rather than a single categorical value.

3.3 Methodology

In this section, we describe the data collection, topic extraction, and methodology used for developing our user similarity measures.

3.3.1 Data Source

The location-based social networking platform, *Foursquare*, was used as our primary source of modeling data based on the sheer number of crowdsourced venues as well as its ubiquity as a location-based application. As the application defines it, a venue is a user-contributed “physical location, such as a place of business or personal residence.”¹ and as of publication, *Foursquare* boasts over 9 million venues in the continental United States alone. This platform allows users to *check in* to a specific venue, sharing their location with anyone they have authorized as well as other OSNs such as *Facebook* or *Twitter*. Built with a gamification strategy, users are rewarded for checking in to locations with badges, in-game points, and discounts from advertisers. This game-play encourages users to revisit the application, compete against their friends and contribute *check-ins*, *photos* and *tips*.

Venue Tips

An additional feature of *Foursquare*, is the ability for a user to contribute text-based *tips* to a venue. *Tips* consist of user input on a specific venue and can range from a restaurant review to a hiking recommendation. Lacking any official descriptive text for venues on *Foursquare*, these unstructured tips describe

¹<https://foursquare.com/>

and define the venue and location. As with most crowdsourced data, the length, content, and number of tips vary significantly throughout the *Foursquare* venue data set. Of the 9 million *Foursquare* venues available in the continental United States, approximately 22.8% included at least one tip. Taking only venues that have had more than ten unique user check-ins, this value jumps to 54.0%. Of the venues to which our sample population checked in, 77.0% include at least one tip with the mean length of a single tip being 74 characters (stdev = 49.3). Table 3.1 shows a few examples of tips left at different venues.

Order your tacos with flour tortilla and use their amazing green salsa!
Free wifi & power outlets outside work. Let's support and make sure they'll be there a long time
I just bought some leather chairs and I love them, great quality furniture

Table 3.1: Example tips

3.3.2 Data Collection

Publicly geotagged *Foursquare* check-ins were accessed via the *Twitter API* for 6000 users over a period of 128 days. Check-ins to venues with less than ten tips were removed as well as users with an overall check-in count less than 16. This resulted in a dataset totaling 24,788 check-ins over 11,915 venues for 797 users (mean of 31.1 check-ins per user). From a geosocial perspective, we define

an individual’s activity identity as an amalgamation of the venues to which she checks in.

3.3.3 Themes

In this work, we use a Latent Dirichlet Allocation (LDA) topic model to extract a finite number of descriptive themes (topics) from the user-generated tips assigned to venues in our Foursquare dataset. While numerous topic models are discussed in the literature, LDA is a state-of-the art generative probabilistic topic model that can be used to infer the latent topics in a large textual corpus in an unsupervised manner (Blei et al. 2003). A topic is a multinomial distribution over terms, where the distribution describes the probabilities that a topic will generate a specific word. LDA models each document as a mixture of these topics based on a Dirichlet distribution. Several mature implementations of LDA with improvements exist; for this work we employ the implementation in the MALLET toolkit (McCallum 2002).

A topic model is run across all *Foursquare* venues in the continental United States containing ten or more *tips* (approximately 125,265 venues). Tips are grouped by unique venue ID and all stop-words, symbols, and punctuation are removed as well as the 30 most common words.





High Entropy	Low Entropy
 <p>A word cloud for a high entropy topic. The most prominent words are 'great', 'love', 'friendly', 'care', 'staff', 'hospital', 'doctor', 'awesome', 'nurses', 'doctors', 'office', 'time', 'nice', 'dogs', 'amazing', 'health', 'dr', 'place', and 'waitress'.</p>	 <p>A word cloud for a low entropy topic. The most prominent words are 'great', 'service', 'food', 'good', 'time', 'wait', 'horrible', 'server', 'place', 'don', 'slow', 'order', 'back', 'terrible', 'worst', 'waitress', 'awesome', and 'staff'.</p>
 <p>A word cloud for a high entropy topic. The most prominent words are 'love', 'amazing', 'great', 'hair', 'salon', 'color', 'massage', 'place', 'make', 'years', 'nails', 'hairstylist', 'job', 'haircut', 'rocks', 'time', 'vc', 'make', and 'stylist'.</p>	 <p>A word cloud for a low entropy topic. The most prominent words are 'great', 'good', 'place', 'service', 'friendly', 'amazing', 'love', 'food', 'staff', 'people', 'town', 'prices', 'atmosphere', 'awesome', 'super', 'family', 'excellent', and 'nice'.</p>

Table 3.2: Sample topics derived from Tip text represented as word clouds, where larger words are higher probability words for the topic.

Venue Themes

Using this model we are able to express each venue as a mixture of a given number of topics. The model was tested with 40 topics at 2000 iterations. Future work could involve running similarity models with a varied number of topics. A few of the topics are concerned with a specific type of food, while others are focused on tourism and even baseball. Table 3.2 shows four examples of topics, based on top terms, extracted using LDA.

Temporal Themes

The daily trajectories for each of the 797 users in our dataset are grouped by user and aggregated to a single day. Given the limited number of check-ins, aggregating user activities to a single day was deemed appropriate. Over the 128 days of data collection, this produced a sparse average of 31.1 check-ins per user. This would not be sufficient for any prediction and additionally highlights the need to select similar users as proxies. Selecting one user as our base-line or *focal user*, each check-in in her trajectory is buffered by 1.5 hours. This so-called 3 hour *time window* is used as the temporal bounds from which all additional users' activities are collected. From there we calculate the topic signature for all users within this same time window. This produces an aggregate venue topic distribution for every user over a 3-hour time window around each of the *focal user's* check-ins; 1.5 hour before and 1.5 hour after the check-in. Given these distinctive topic signatures, it is feasible to compare users temporally, across these topics in order to produce a user similarity measure.

A topic *signature* is computed for each of the collections via Equation 3.1 where T_i is one topic in the collective topic distribution, n is the number of venues in the collection, $\#V_j$ is the number of times the same venue appears in the collection and $t_i^{V_j}$ is a single topic probability of Venue j . It is important to note that this method takes the frequency of check-ins to a unique venue into consideration.

This ensures that multiple check-ins to a single location do not over-influence the topic distribution.

$$T_i = \sum_{j=1}^n (\log_{10} \#V_j + 1) t_i^{V_j} \quad (3.1)$$

3.3.4 Variability vs. Commonality Weighting

This approach to calculating the topic signature for a collection of venues puts an equal amount of emphasis on all topics. This is not ideal when measuring the similarity between signatures as some topics are more prevalent across all venues than others. In order to augment the similarity model, we compute the entropy for each topic across all venues. In Table 3.2, two of the word clouds are examples of topics showing high entropy while the other two represent topics with low entropy.

Let t_i be the weight of topic t for venue i . A new discrete variable is defined for topics over venues by normalizing each t_i to t'_i by setting $t'_i = \frac{t_i}{\sum_{j=1}^n t_j}$, where n is the number of venues. The topic's entropy over all venues, E_T , is defined in Equation 3.2.

$$E_T = - \sum_{j=1}^n t'_j \log_2 t'_j. \quad (3.2)$$

Given this set of entropy values, a method for incorporating them as weights in a user similarity model must be assessed. This leads to questioning the role of topic prevalence in constructing a model for assessing user similarity. The approaches we present in the following subsections are influenced by literature in

the cognitive sciences that examined the role of context (or framing) in human similarity assessments. Tversky (1977) found that when two objects are compared for similarity, the set of objects from which the two objects are selected has the effect of making some properties more or less salient in the similarity judgment. The properties that are more salient are termed to be more ‘diagnostic’. Tversky argued that two factors contribute to the *diagnosticity* of a property. The first is *variability*, which finds that the properties that vary across the elements of the context set are used more to determine the similarity (or dissimilarity) of two objects. The second factor *commonality*, is the opposite, that properties that are shared by most elements of the context set are the important properties, because they help explain what is important in the domain of discourse.

Although this context effect is well-studied in the cognitive sciences most computer science similarity measurements are without context in this sense. A notable exception is the *Matching-Distance Similarity Measure* (MDSM), created to compare similarity of spatial entity classes (Rodriguez & Egenhofer 2004). MDSM defines commonality and variability metrics for feature-based classes. In the following sections we adopt these notions to the venue topic signatures.

Variability

One approach postulates that though the commonality topics remain critical in defining the venue (or user), they are less valuable in determining the similarities between two users. For example, if all venues in a dataset are high in a topic related to coffee, this topic does little in determining which two users are most similar. It is the less ubiquitous topics which are more *diagnostic* in the similarity model. Based on the literature on similarity (Tversky 1977), we call this type of diagnosticity, the *variability* weight.

In order to add weight to these more diagnostic topics, we build our similarity model based on a subset of ten topics with the highest entropy. Given the reduction in the number of topics, the collective topic distribution must then be normalized ($n=10$) to sum to 1 in order to compare distributions.

Commonality

It may be argued that the inverse effect of variability, *commonality* is more applicable. A *commonality* weight implies that more prevalent topics should be more influential in measuring user similarity. In essence, the more coffee shops one visits, the more similar they are to other coffee shop visitors.

The influence of entropy on topics using this commonality method involves taking the top ten topics with the lowest entropy and building our similarity

model based purely on those topics. Again, the collective topic distribution is normalized in order to sum to 1.

3.3.5 Comparing Users

Since each aggregate venue signatures consist of a distribution over an equal number of topics, a divergence metric may be used to measure the similarity between our *focal user* and all other users at any given activity. Using the *Jensen-Shannon divergence (JSD)* (Equation 3.3), we compute a dissimilarity metric between each user's topic distribution and the *focal user's* respective topic signature. $U1$ and $U2$ represent the topic signatures for User 1 and User 2 respectively, $M = \frac{1}{2}(U1 + U2)$ and $KLD(U1 \parallel M)$ and $KLD(U2 \parallel M)$ are *Kullback-Leibler divergences* as shown in Equation 3.4.

$$JSD(U1 \parallel U2) = \frac{1}{2}KLD(U1 \parallel M) + \frac{1}{2}KLD(U2 \parallel M) \quad (3.3)$$

$$KLD(P \parallel Q) = \sum_i P(i) \log_2 \frac{P(i)}{Q(i)} \quad (3.4)$$

The *JSD metric* is calculated by taking the square root of the value resulting from the equation. Given the inclusion of the logarithm base 2, the resulting metric is bound between 0 and 1 with 0 indicating that the two users' topic signatures are identical and 1 representing complete dissimilarity.

3.4 Results & Discussion

Selecting a *focal user* at random from the 797 users, we first run the basic *JSD* dissimilarity model without including an entropy weight. In order to keep the number of topics uniform across all models, a set of ten topics are randomly selected for comparison. Figure 3.1 shows the dissimilarity metrics at activity level resolution for 3 individuals compared to the *focal user*. As one can see, *User A's* similarity to the *focal user* generally decreases as the day progresses, with late evening proving to be the most similar time of day, *User B* is similar around lunchtime and quite dissimilar in the morning. Lastly, *User C* mirrors the average for most of the day with a small bump in the morning and a sharp peak of similarity at around 16:30.

In comparison, Figure 3.2 shows the effect of including the entropy measure with the purpose of emphasizing more diagnostic topics within the venue distributions. The same three users are compared to our *focal user*, but this time the venue distribution is composed of topics high in variability. The most visible outcome of the variability weight inclusion is an increase in range of similarity measures across users. Each of the three users is completely dissimilar to our *focal user* at some point during the day and the average dissimilarity across all users has increased.

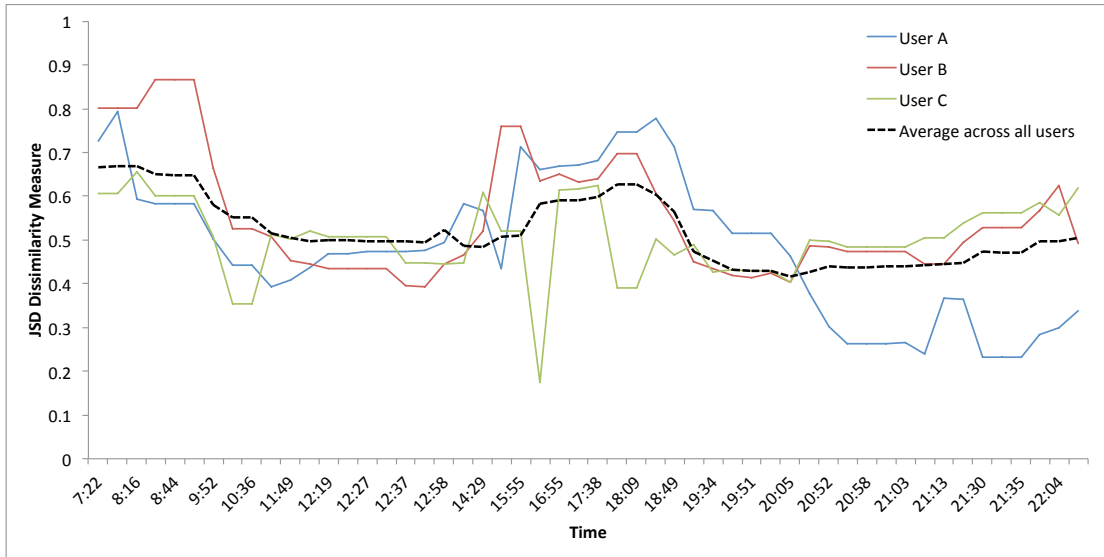


Figure 3.1: Similarity of User A, B & C to Focal User (randomly selected topics)

Interestingly enough, each of the sampled users increased their similarity to the *focal user* at least once throughout the day. Given that these topics offer the largest variability within the dataset, it is not surprising that a measure of similarity between users based purely on these topics will decrease overall in comparison to the non-entropy selection. This variability model will return specific peaks of similarity between users given that it is emphasizing the topics not as common across all venues. *User A* and *User B* show dramatic increases in similarity in the morning, with *User C* peaking around dinnertime. As this figure makes apparent, the change in user similarity is not uniform across all activities or users, it is dependent on the prevalence of a given topic (or combination of topics) within the aggregated distribution of an activity venue.

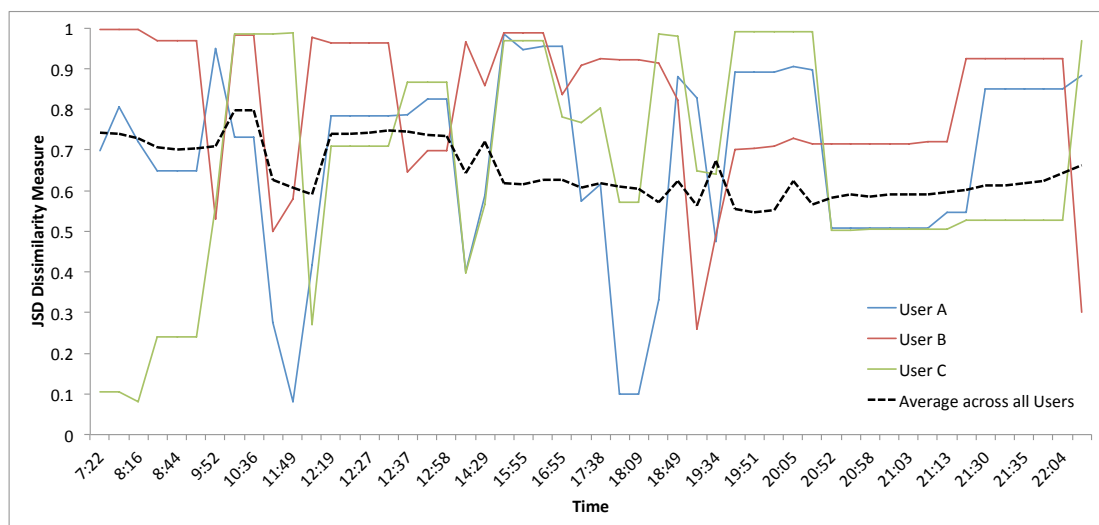


Figure 3.2: Similarity of User A, B & C to Focal User (topics with highest entropy)

The *commonality* model offers a very different perspective. Figure 3.3 shows that on average, the similarity between all users and the *focal user* increased. While some semblance of the random-topics figure still exists, the users appear more uniform in their similarity to our *focal user*.

3.4.1 Validating the model

This section presents the methods used to validate the similarity model as well as the results of the validation. Both of the entropy-based similarity models are evaluated along with the non-entropy model. The methods below are applied to each model.

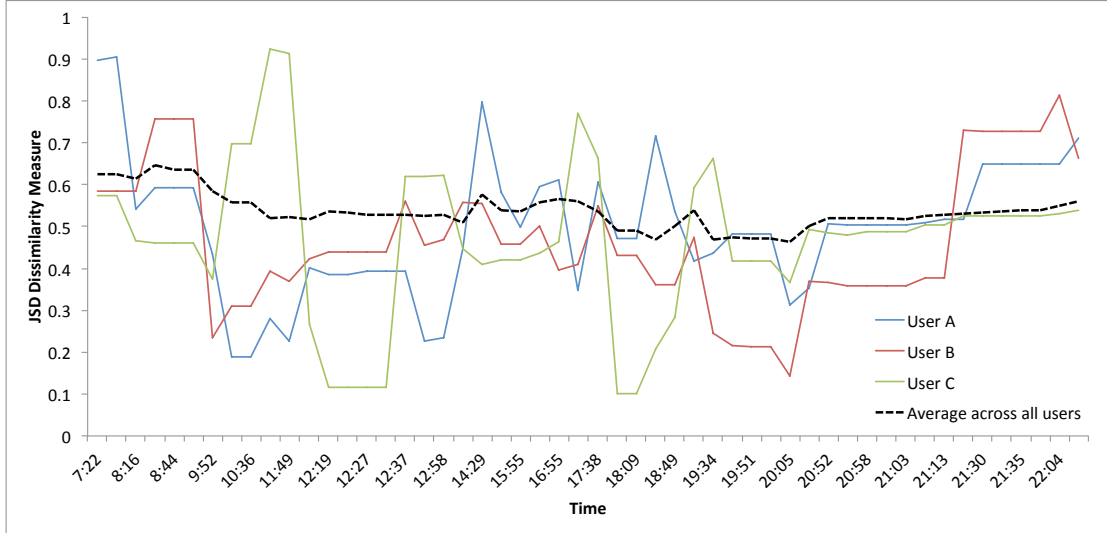


Figure 3.3: Similarity of User A, B & C to Focal User (topics with lowest entropy)

To start, the topic distributions for the top- k most similar users for each check-in are combined using Equation 3.5. The influence of each user on the combined topic distribution (HV) is calculated by multiplying the topic by the similarity value sim where $sim = 1 - dissimilarity$. This ensures that more similar users have a larger impact on the overall topic signature.

$$HV_{T_i} = \sum_{j=1}^n ((sim_j) * T_i^j) / \sum_{i=1}^m T_i \quad (3.5)$$

The resulting topic distribution represents a *hypothetical venue (HV)* that is the most similar to the *focal user's* check-in location as possible based on the model. In order to evaluate this *hypothetical venue*, we extract the 29 nearest (physically) venues (along with their topic distributions) for each of the *focal*

user's check-ins. This collection of venues, along with the actual check-in venue, form the test set from which the similarity model is assessed.

The 30 sample venues are ranked in order of similarity to the *hypothetical venue* and the position of the real check-in venue within this ranked set is recorded. Figure 3.4 shows an example with graduated symbol markers representing the dissimilarity of each venue (large dark color = low dissimilarity). In this example, the top 5 most similar venues are labeled with the actual check-in venue resulting in 1 (the most similar venue to the *hypothetical venue*). This process is run across

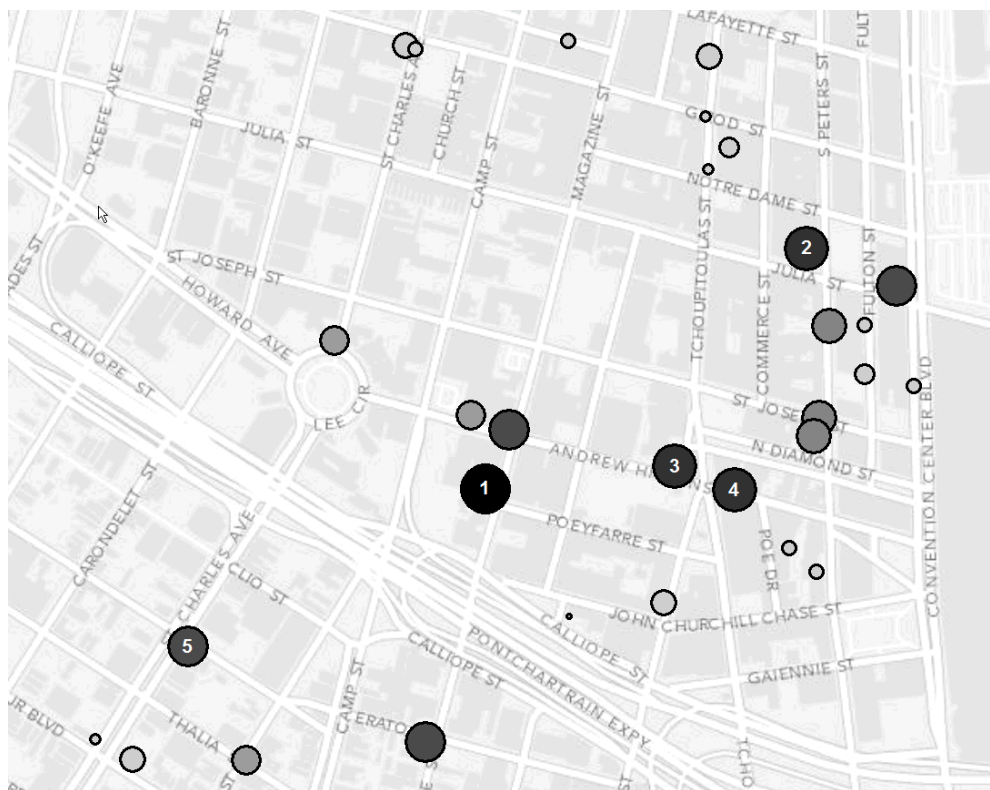


Figure 3.4: Map fragment showing graduated symbols for the 30 nearest venues

all check-ins for all users with the three levels of weighting. Table 3.3 shows an ordered-position table based on 3188 sampled check-ins over the 797 users in our dataset (4 randomly sampled check-ins per user). Both the 40 Topic model and the 30 Topic model are present in this table, showing the results for the *Variability*, *Commonality* and *No weight* models.

The *Commonality* weighted model produced the best results with over 77% of the *hypothetical venues* contributing to a correct estimation of the actual venue. In fact, the *Commonality* weighted model placed the actual check-in venue within the first 3 most similar venues 95% of the time. By comparison, the *Variability* weighted model was significantly less accurate, correctly estimating the actual check-in venue 45% of the time. While this performance is not as strong as the *commonality* weighted model, it is to be expected as the purpose of exploiting the *variability* topics within the topic distribution is to find the nuanced differences between venues rather than the overall commonality between them. Lastly, the results of the *non-weighted* , randomly-sampled topic model are presented. As a base-line, we see that even without the inclusion of entropy weighting, this similarity model produces excellent results with 65% of actual venues being correctly estimated. In all cases, these results suggest that the model performs quite well in estimating an actual check-in based purely on the check-ins of similar users.

Placement	Commonality (%)	Variability (%)	Random (%)
1	77.02	45.04	65.85
2	14.16	17.17	18.05
3	4.30	9.38	7.22
4	1.98	5.65	4.02
5	1.16	2.86	2.23
6	0.53	2.17	1.04
7	0.28	1.88	0.56
8	0.16	1.22	0.25
9	0.09	1.16	0.25
10	0.03	0.97	0.16
11	0.03	0.50	0.03
12	0.06	0.88	0.06
13	0.00	0.53	0.03
14	0.03	0.35	0.00
15	0.03	0.16	0.00
16	0.00	0.22	0.06
17	0.00	0.97	0.03
18	0.00	0.63	0.00
19	0.00	0.44	0.03
20	0.00	0.50	0.00
21	0.00	0.09	0.03
22	0.06	0.31	0.03
23	0.00	0.19	0.00
24	0.03	0.22	0.00
25	0.00	0.53	0.03
26	0.00	0.82	0.03
27	0.00	1.10	0.00
28	0.00	1.29	0.00
29	0.03	1.69	0.00
30	0.00	1.07	0.00

Table 3.3: Placement of actual venue based on similarity to *Hypothetical Venue*

3.5 Limitations

While the methods presented in this paper offer a promising approach to assessing user similarity through unstructured data, there are a number of limitations. Since the topic models are built on crowdsourced data (*tips*) from users of the application, the standard bias and errors of crowdsourcing are present. There is no way to ensure that a user submitting a tip has ever been to the venue or is offering a truthful tip. Additionally, since all tips for a single venue are combined in order to run the LDA model, those tips with more content have a large impact on the overall generation of topics. While there has been an increase in the number of people using LBSN applications, it should be noted that one's *Foursquare* check-in history does not account for every single activity that the user conducts throughout her day; the average user does not *check in* to every venue that she visits. It is more likely that a user checks in to locations that are unique or different from those to which she normally checks in. To some users, one venue might offer more social capital (Pultar et al. 2010) than another (e.g., nightclub vs. hospital) and user's opinions range on what is *unique*. However, the limitations discussed here also hold for most other methods designed based on volunteered geographic information and are a research challenge.

3.6 Conclusion and Future Work

The work presented in this paper offers an overview of an innovative approach to assessing user similarity across sparse, unstructured geosocial check-ins. In this paper, we explicitly extract the non-spatial components from the spatial data by focusing purely on the textual descriptions of locations. Given the amorphous nature of online social networking data, topic modeling has allowed us to extract themes from crowdsourced social data. These themes are merged across venues to produce a unique signature that defines an individual's geosocial activities at any given point in time. Through exploration of *variability* and *commonality* measures, based on the entropy calculated across these themes, we have shown two opposing methods for evaluating user similarity through publicly available *check-in* data. A model based on *Commonality* within the data produces the best results when estimating real check-ins from a set of nearby locations. The *Variability* within the venue topics allows us to explore the nuanced similarities between users and the venues they frequent. In all, these methods demonstrate value in their ability to enhance existing user similarity models.

Future work in this area will flow in a number of directions. With an increase in the amount of user check-ins, the data will allow for further temporal factoring to reflect day of the week and month. It is expected that a user's activity patterns

are not limited to hours within a day, but also reflects days of the week. The addition of temporal components will further enhance the ability of the model to discover similar users. Exploring the factors that contribute to this measure of user similarity will be a next step in this area of research as well. Analysis involving the correlation between location types and similarity measurements should be examined as well as outside factors that may contribute to similarity between users (e.g., demographic data, climate, etc).

Additional sources of unstructured geosocial content will be explored with the goal of enhancing the extraction of topics for venues. An incredible amount of unstructured geo-tagged content is available online and the addition of this data to our model will dramatically increase its accuracy. Lastly, while the sparsity of the data and the results gathered from such data is a novelty of this research, more precise activity information for a population of individuals (through a GPS enabled mobile device for example) will be tested order to assess the robustness of the model.

Chapter 4

Where is *also* about time: A location-distortion model to improve reverse geocoding using behavior-driven temporal signatures

This chapter takes the current spatial distance-only approach to reverse geocoding as a baseline and presents a novel method for enhancing this approach through the inclusion of temporal behavior data. Reverse geocoding (or place search) is a task that researchers and developers continually face when attempting to determine a person's *placial* location based on their geospatial coordinates. Current ranked spatial proximity methods perform this task with a moderate level of accuracy. In this chapter it is shown that spatial distance can be distorted by a factor of temporal check-in probability to produce a reverse geocoding method that significantly outperforms the baseline. This chapter fits in to the overall theme of this dissertation through its combining of the temporal dimension with the spatial di-

Chapter 4. *Where is also about time: A location-distortion model to improve reverse geocoding using behavior-driven temporal signatures*

mension to show one of the ways in which temporally-based place type definitions can augment an existing and currently spatial-only geolocation task.

Peer Reviewed Publication	
Title	Where is <i>also</i> about time: A location-distortion model to improve reverse geocoding using behavior-driven temporal signature
Authors	Grant McKenzie, Krzysztof Janowicz
Venue	Computers, Environment and Urban Systems
Editors	J.C. Thill
Publisher	Elsevier
Pages	Under Revision
Submit Date	November 19, 2014
Accepted Date	Under Revision
Publication Date	Under Revision
Copyright	Under Revision

Abstract

While geocoding returns coordinates for a full or partial address, the converse process of reverse geocoding maps coordinates to a set of candidate place identifiers such as addresses or toponyms. For example, numerous Web APIs map geographic point coordinates, e.g., from a user’s smartphone, to an ordered set of nearby Places Of Interest (POI). Typically, these services return the k nearest POI within a certain radius and use distance to order the results. Reverse geocoding is a crucial task for many applications and research questions as it translates between spatial and platial views on geographic location. What makes this process difficult is the uncertainty of the queried location and of the point features used to represent places. Even if both could be determined with a high level of accuracy, it would still be unclear how to map a smartphone’s GPS fix to one of many possible places in a multi-story building or a shopping mall, for example. In this work, we break up the dependency on space alone by introducing time as a second variable for reverse geocoding. We mine the geo-social behavior of users of online location-based social networks to extract temporal semantic signatures. In analogy to the notion of scale distortion in cartography, we present a model that uses these signatures to distort the location of POI relative to the query location and time, thereby reordering the set of potentially matching places. We demonstrate the

strengths of our method by evaluating it against a purely spatial baseline by determining the mean reciprocal rank and the normalized discounted cumulative gain.

4.1 Introduction and Motivation

Translating back and forth between spatial and placial representations of location is a crucial task underlying many research questions, applications, and systems. Geocoding, for instance, is the process of assigning corresponding geographic coordinates to other types of structured geographic identifiers such as addresses. The converse process, called reverse geocoding, assigns place identifiers, such as toponyms, to geographic coordinates. More specifically, it maps a geometry in the sense of OGC's Simple Feature model to an ordered set of candidate place identifiers. Typically, the Euclidean distance between the query coordinates and the point-feature representation of the candidate places is used to establish a ranking. To successfully match a user's location to a visited place, new geosocial approaches also consider popularity, e.g., how many users checked-in or wrote reviews about a place. Additionally, many (reverse) geocoding systems consider place hierarchies and granularity.

The following queries nicely illustrate the difference between a spatial and placial perspective as well as the arbitrariness of relying on point coordinates for the query and the candidate places alone. While not a reverse geocoder in the strict sense, the Flickr *flickr.places.findByLatLon* API call (Flickr 2014) returns place IDs given a lat/lon coordinate and accuracy value. This allows users to find photos for particular places. The API *rounds up* to the nearest place type, i.e., it returns a city ID for street-level coordinates rather than returning a street or building. Latitudes and longitudes are truncate to three decimal points. In each case, the query coordinates represent the same fix at the Griffith Observatory in Los Angeles. However, the query is run with different accuracy levels where 16 corresponds to the street level, 11 to the city level, and 7 to the county level. The respective responses from the Flickr API are as follows.

```
<places latitude="34.118341" longitude="-118.300458" accuracy="16" total="1">
<place place_id="HqDLYDJTUb8XihYDg" woeid="23511984" latitude="34.125"
longitude="-118.306" [...] place_type="neighbourhood" place_type_id="22"
timezone="America/Los_Angeles" name="Hollywood United, Los Angeles, CA, US,
United States" woe_name="Hollywood United" />
</places>

<places latitude="34.118341" longitude="-118.300458" accuracy="11" total="1">
[...] latitude="34.146" longitude="-118.248" [...]
place_type="locality" place_type_id="7" name="Glendale, California,
United States" [...] /> [...]

<places latitude="34.118341" longitude="-118.300458" accuracy="6" total="1">
[...] place_type="county" place_type_id="9" [...] name="Los Angeles County,
California, United States" [...] /> [...]
```

The fact that even small differences in spatial accuracy may have strong impacts, e.g., on routing choices, has been demonstrated in the literature before (Bowling & Shortridge 2010). What makes the example above interesting is the place hi-

erarchy. Hollywood is a district of Los Angeles, while Glendale is a city in Los Angeles County. From a human-centered *placial* perspective, one would assume the queries to return Hollywood (in fact, it should be the Los Feliz neighborhood), Los Angeles, and finally Los Angeles County. Instead the neighboring city of Glendale is returned for the city-level accuracy query, thereby breaking the expected hierarchical composition of places. From a computation-centric *spatial* perspective Glendale is selected simply because its centroid representation is closer to the query location than the centroid of Los Angeles.

The arbitrariness and imprecision of point-feature representations as well as the effect of missing topological relations also strikes on the level of small-scale features such as Places Of Interest (POI).¹ Figure 4.1 illustrates a common issue. First, the resort marker (A) is placed at the entrance to the parking lot. While this may be acceptable, other POI databases place it at the center of the building which is nearly 150m away. Second, the lounge is *inside* the resort but its marker (B) is shown over 100m away from the resorts marker. As most reverse geocoders rely on distance alone, such differences will lead to substantially different and often misleading results, e.g., when proposing a user's check-in location.

As the omnipresence of location-enabled mobile devices increases, more robust, accurate, context-aware, and data-rich geolocation services are required. Today,

¹Frequently also referred to as *Points Of Interest*.

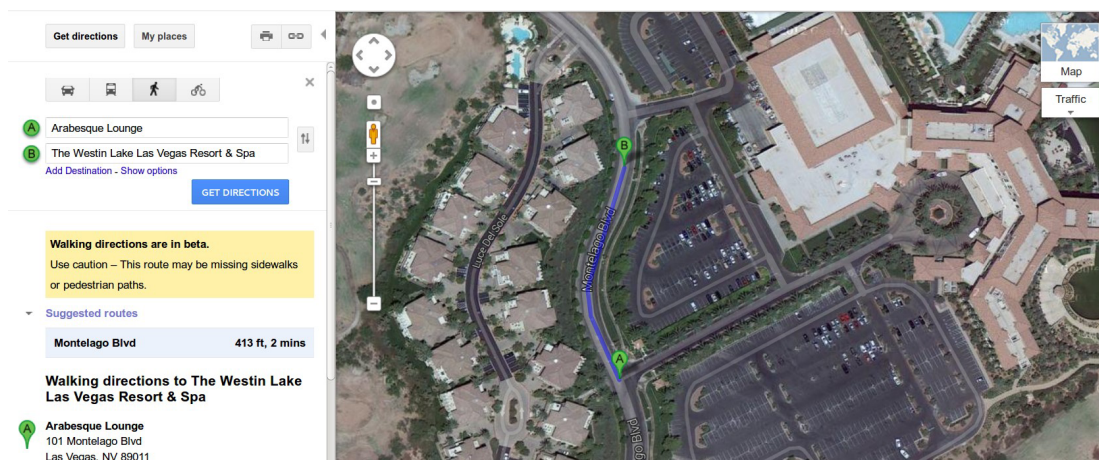


Figure 4.1: Point-feature distance between a resort and the lounge located inside of it (screenshot from Google Maps).

the ability to link spatial coordinates to an actual place has become essential in many aspects of our everyday lives including navigation applications, place recommendation, location-based advertising, and critical infrastructure. It is interesting to note that the challenge is not one of more accurate GNSS and Wi-Fi-based positioning systems (WPS) alone. The information that a person checked-in or is present at a place is semantically richer than the spatial data alone. To give a concrete example, the fact that a person is standing in front of a *food truck* is substantially different from the fact that a person checked-in to the *food truck* and is likely to order something. Spatial information is richer than just spatial proximity.

Commercial companies such as *Google* as well as open source platforms like *GeoNames* have made names for themselves offering application programming

interfaces (APIs) and web services that allow both developers and consumers to query gazetteers and POI databases using geographic coordinates as input. With the increase in user-generated geo-content, new services such as *Foursquare* and *Yelp* have emerged allowing anyone with a location-enabled mobile device to contribute or update the location of an entity in a crowd-sourced system. It is important to note that while these systems involve the contribution of geo-content from individual users, there is still some discussion as to whether or not they fit in to the category of *Volunteered Geographic Information* (Harvey 2014, McKenzie & Janowicz 2014). Previous work on POI matching has shown that the median distance of a single POI between different geolocation service providers is 62.8 meters apart and can reach up to several hundreds meters under extreme circumstances (e.g., for a golf course) (McKenzie et al. 2014). Figure 4.2 illustrates this fact by showing the position of markers from five major services. While this offset may not be a substantial issue in rural areas due to their low POI density, it will cause substantial problems for geolocation services (e.g., check-in services) in a high-density urban areas.

The task of determining the place an individual is visiting based on coordinates gathered from their mobile device becomes more difficult given the uncertainty associated with each POI in the dataset. That is, selecting the nearest POI to a user's location becomes an artifact of the arbitrary point-coordinate representation

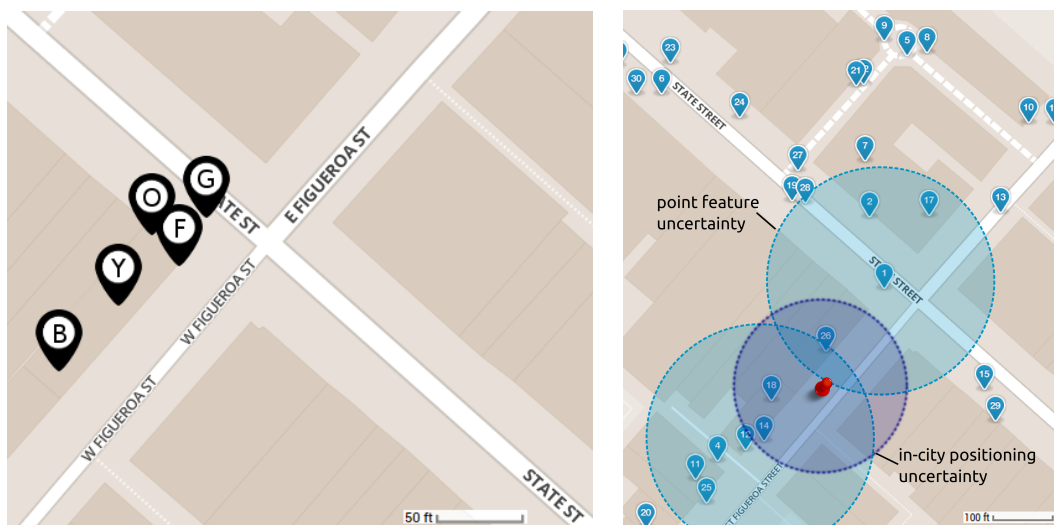


Figure 4.2: **Left:** Different services list different locations for *The French Press Café* in Santa Barbara, CA. Google Maps (G), OpenStreetMap (O), Foursquare (F), Yelp (Y), Bing Maps (B). **Right:** Uncertainty in POI location and user location. French Press Café (1) and Los Arroyos Mexican Restaurant (14). The red pin marks the user’s most probable location. Note that the circles of uncertainty are not drawn to scale; in actuality they would appear larger.

of nearby POI. Leaving the actual POI locations aside, another facet of uncertainty plagues traditional geolocation services, namely the positional accuracy of a location-enabled device. While most devices make use of a range of positioning technologies (e.g., GNSS, WPS, Cellular Network), each of these technologies has its own issues related to accuracy, imparting a level of uncertainty on any device location. Therein lies one of the problem facing traditional geolocation services such as reverse geocoders. Given all of this uncertainty, how can a geolocation service be expected to accurately predict a POI given geographic coordinates? An example of this challenge is shown in Figure 4.2. A number of POI are shown on

the map along with their associated uncertainty. Additionally, the red pin shows the most probable location of a mobile device and its two-dimensional depiction of uncertainty.

4.2 Research Contribution and Example Scenario

Clearly, relying on geographic coordinates alone to infer a place based on a user's mobile device position is not sufficient. However, there are other contextual clues that can be taken into account. Time is one such clue and in contrast to many other contextual information it is readily available with every position fix. Current reverse geocoding services solely exploit geographic location while in reality human behavior dictates that approximately the same location in geographic space can serve a variety of purposes at different times of the day or days of the week. The motivation for visiting a specific city block on a Tuesday morning is considerably different than visiting that same block on a Saturday night. While the geographic coordinates determined by one's location-enabled mobile device may be temporally-agnostic, the *probability* of conducting an activity at a nearby place is not.

In fact, place categories are implicitly defined by time. For instance, the likelihood of being at the *Department of Motor Vehicles* on a Sunday at 1 AM is

negligibly low. Not only is this likelihood driven by socio-institutional constraints (Raubal et al. 2004), but also by observable human-placial behavior patterns. Existing research in this area has shown that categories of places (e.g., Hospital, Restaurant, Bar) can be uniquely identified by the temporal patterns of their visitors (Ye, Janowicz, Mülligann & Lee 2011, McKenzie et al. In Press, Noulas et al. 2011). In this work, we make the case for *time* being an additional readily available clue for reverse geocoding and geosocial check-ins in specific. We demonstrate that given a time-stamp of a mobile device location fix, these unique *temporal signatures* (McKenzie et al. In Press) can be incorporated with existing distance-only based methods to substantially enhance the accuracy of place estimations.

The research contributions of this work are as follows:

- In analogy to the notion of scale distortion in cartography, we present a model that uses temporal signatures to distort the location of POI relative to the query location and time, thereby reordering the set of potentially matching places. Using the check-in frequency of a POI category at a specific time, geographic space is distorted by a factor of the temporal probability. Places that show a high check-in frequency at the provided time are shifted closer to the queried geographic coordinates of the user while those with low probabilities are pushed further away. Intuitively, given a user’s location fix

at 10pm, a nearby cinema is preferred over a closer bakery as the temporal signature of the place type *Bakery* indicates that people rarely visit bakeries during the night.

- We explore and report on multiple models for this temporal distortion analogy including linear, non-linear, symmetric and non-symmetric functions. Our study indicates that a non-linear, non-symmetric rational function produces the best results.
- We demonstrate the strengths of our method by evaluating it against a purely spatial baseline (used by most currently available services) by determining the Mean Reciprocal Rank and the normalized Discounted Cumulative Gain.² Our enhanced method increases the estimated accuracy of an individual’s location Mean Reciprocal Rank from 0.359 to **0.453** and the normalized Discounted Cumulative Gain from 0.583 to **0.711**. Additionally, we demonstrate that our model can also be used to improve the prediction accuracy of geosocial systems such as *Foursquare* which is noteworthy given their detailed ground-truth data.
- Many potential contextual clues are available to improve the quality of location services. Examples include weather information, mode of transporta-

²These statistical rank approaches will be further explained in Section 4.2.

tion, previously visited location, user preferences, and so forth. Many of them, however, are not available outside of commercial data silos, are difficult to mine, require different index schemes, or substantially increase the complexity of (pre-)computing candidate places. While time is readily available with every position fix and we provide signatures for each hour of the week, some use cases require pre-computed results. By computing information gain, we show that the temporal signatures vary greatly with respect to their indicativeness. Consequently, a few selected time-frames can already improve place estimation.

- Finally, we present an outlook on user-location distortion models. Our current work uses *default behavior* to compute the temporal probability of POI categories for different times. People (and places), however, do not always follow such established patterns. For instance, there might be an event at a location that would be closed otherwise. By enriching the default mode with a *dynamic real-time* model, we can adjust for such circumstances. We discuss the role of *Instagram* photos and *Tweets* to determine trending areas in real-time. We propose an inverse-distance weighed method to alter the user's query location, pulling it closer to areas of high online-social networking popularity.

Stepping back from the research contributions for a moment, let us explore a real-world scenario depicting the problem. This scenario will act as running example throughout the paper. Figure 4.3 shows a query location (red pin) and a number of nearby POI. A standard distance-based approach would simply calculate the distance between each POI and the query location and return a ranked set of distances allowing the user to make the assumption that she is currently at the closest POI. In referencing the temporal signatures for the different POI types, we find a *visit probability value* for each category of POI at any given hour of the day on any day of the week.

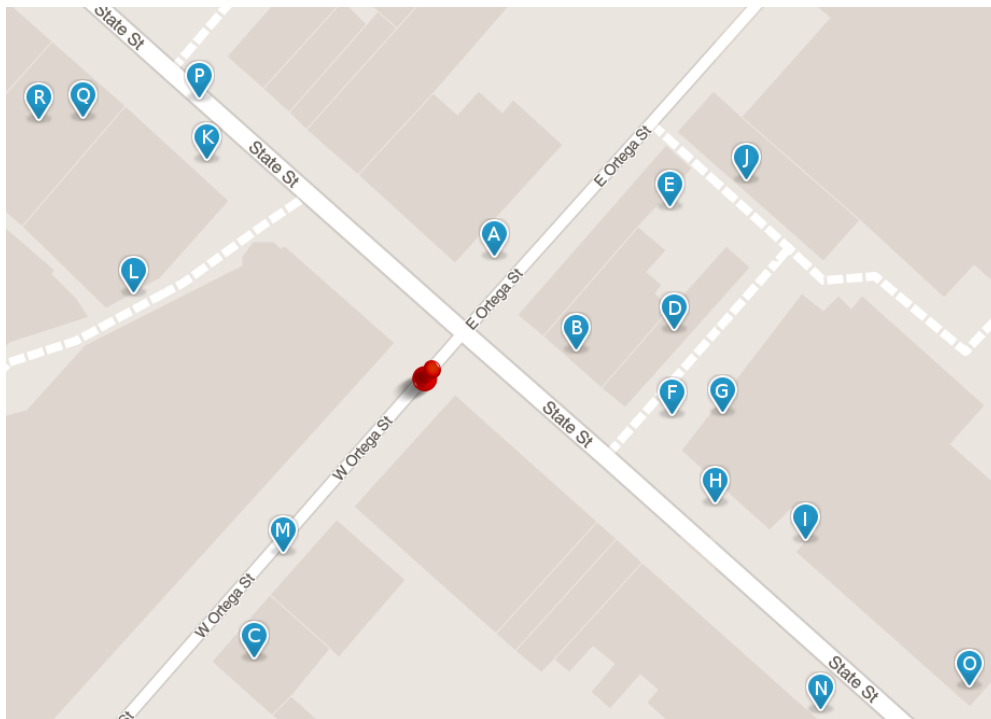


Figure 4.3: Coordinates from user’s device (red pin) and nearby POI (blue markers).

Table 4.1 shows the categories associated with each POI in Figure 4.3, the geographic distance to the query location, as well as the temporal probabilities for those POI types at both 10 AM on Monday and 11 PM on Saturday. As one can see, the popularity of nearby POI change significantly between the two times. Rather than assuming that there is an equal likelihood of a user visiting a POI, irrespective of time, it follows that temporal probability should be included in determining the most likely place.

Marker	Category	Distance (m)	Monday 10AM (10^{-3})	Saturday 11PM (10^{-3})
A	Bakery	39.2	6.28	4.08
B	Nightclub	41.4	0.26	44.16
C	Nightclub	69.9	0.26	44.16
D	American Restaurant	62.7	1.61	9.50
E	Bakery	73.7	6.28	4.08
F	Fast Food	65.0	4.80	5.78
G	Apparel Store	85.8	2.51	1.09
H	Ice Cream Shop	82.6	0.84	15.88
I	Movie Theater	94.2	1.44	11.00
J	Pub	88.9	0.53	22.66
K	Cosmetics Shop	60.9	3.87	1.57
L	Diner	70.0	5.49	7.56
M	Italian Restaurant	45.7	1.42	7.96
N	Furniture Store	114.9	4.79	5.01
O	Grocery Store	147.8	4.53	1.38
P	BBQ Joint	82.3	0.43	9.35
Q	Burrito Place	88.1	0.54	3.16
R	Italian Restaurant	93.6	1.42	7.96

Table 4.1: POI Categories shown on Figure 4.3 with distance to device location and temporal probabilities (sum of probabilities across all categories sums to 1) on Monday 10 AM and Saturday 11 PM.

The remainder of the paper is structured as follows. In Section 4.3 we introduce our temporal signatures-based location-distortion model, the extracted temporal signatures, and the used data. Next, Section 4.4 discusses the tested functions and their weights. In Section 4.5 we evaluate our proposed method. We present an outlook on dealing with real-time information in Section 4.6. In Section 4.7, we contrast our work to related research and discuss relevant findings. Finally, Section 4.8 offers conclusions and directions for future work.

4.3 Temporal Signatures-based Location-distortion Model

In this Section, we discuss the distortion models, the temporal signatures they use, and the data from which they were derived.

4.3.1 Distortion Models

The majority of current geolocation services take a position fix as input and return a set of ascending distance-ranked POI based on the geographic coordinates of those POI. Given a robust set of category-defining temporal probabilities gathered from location-based social networking check-ins, this paper offers a model for increasing the accuracy of the distance-based approach through the inclusion of

a temporal component. Different types of POI show fluctuations in visit probabilities throughout the day. Based on check-in behavior, these fluctuations reflect increases and decreases in POI type popularity. We leverage these probabilities to enhance distance-based geolocation approaches. To do so, we propose an analogy to scale distortion in cartography and distort space by a factor of the temporal probability. That is, we pull or push POI in the users vicinity depending on their type’s visiting likelihood during a particular time of the day.³ We realize our models by exploring four types of functions (Figure 4.4) for altering the geographic distance between the query location and each POI by a weighted temporal probability.

These four models represent different approaches to combining distance and time. The *linear* approach *symmetrically* adjusts the distance by pushing POI with low check-in probabilities away from the query location at a linear rate equivalent to the amount that high-probability venues are pulled towards the query location. Alternatively, one could model a changing, i.e., *non-linear*, push/pull rate that changes with the probability. While still symmetrical in its design, the assumption underlying this model is that highly likely or unlikely places should be pulled or pushed at a different rate while values close to the mean should approx-

³It is worth noting that all analogies are partial. We mathematically model the relative impact of distance and time to alter the POI ranking returned to the user but do not actually modify the underlying base geo-data.

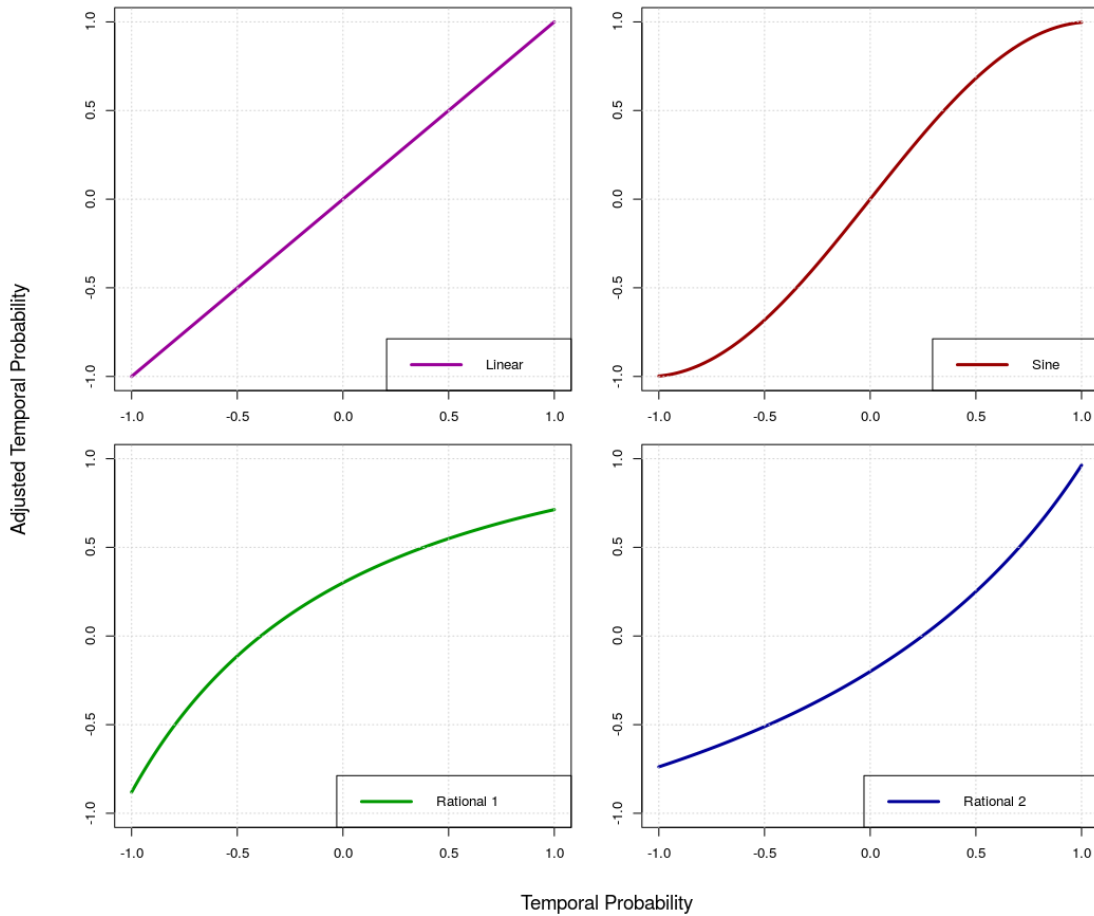


Figure 4.4: Four possible distortion models and examples of their realization for shifting POI locations based on the temporal probability of their types (e.g., restaurant); *exaggerated*.

imate a linear behavior. Here we employ a particular interval of *sine* functions for the symmetric non-linear model. We also explore *non-linear, non-symmetrical* options. Rational function 1 shown in Figure 4.4 depicts an example of one such option. In this case, as the probability of a user checking in to a POI increases the amount by which the distance decreases diminishes. Correspondingly, as the

temporal probability decreases, the amount by which the distance increases grows. In essence, those POI with low check-in likelihoods are punished at a higher rate than those with high probabilities are rewarded. The inverse of this function is also presented in Figure 4.4, *Rational 2*, decreasing the influence on geographic distance as temporal probability values move to the left while exponentially increasing the influence on distance as values move to the right. Intuitively, the rationale behind both non-linear, non-symmetrical models is to study whether pushes and pulls should be performed at different rates. Each of these four models is unique in its approach to the data. We compare them, their realizations, and their parameterization in later sections.

4.3.2 Activity Categories

When a new POI is contributed through the Foursquare mobile application, the creator is able to assign a category tag by selecting from a pre-defined hierarchical set of activity categories. Originally generated by user-contributed tags, governors of the Foursquare application refined the list on multiple occasions, eventually restricting category assignments to just those provided via the application. While the set does occasionally undergo minor adjustments, at time of writing, this category set consists of 421 unique categories divided between three hierarchical levels (Foursquare 2014a). Contributors to the application are asked to assign

at least one category to any venue they generate, though this is not enforced (Foursquare 2014b). A sample⁴ of 15,731,452 POI from across the United States showed that 86.19% of venues were assigned one categorical value, 0.07%, 2 or more and 13.74% had no category.

4.3.3 Geosocial Check-ins

Geosocial *check-in* data were collected via the Foursquare API with the purpose of constructing temporal signatures for specific venue categories. A total of 908,031 randomly selected Foursquare venues⁵ were accessed via the application API, divided amongst 421 categories, with a goal of accessing 240 venues per category. Unfortunately given a the uniqueness of a number of categories (e.g., Molecular Gastronomy Restaurant) it was difficult to achieve this number of POI for each category. Once the venues were chosen, check-in data were accessed every hour for four months starting October 2013. Each request for check-in information returned a value of *HereNow* which indicates the total number of users checked-in to the specific venue at any given time. Provided the number of venues listed above, a total of 3,640,893 check-ins were temporally analyzed. To account for regional variations, the data was collected from Los Angeles, New York City, Chicago, and New Orleans.

⁴Accessed through the public-facing API

⁵*Venue* in this case is the Foursquare-specific term for Point of Interest

It is worth noting that the Foursquare data is biased towards a particular user population, places, and place types. For instance, the typical Foursquare user is a 30-year-old American male and more likely to check-in at a trendy nightclub than an airport. We mitigate this problem by aggregating the data to the type level, i.e., over millions of check-ins, even though some places and place types receive less check-ins, nightclub still peak during weekend nights, while airports have a more uniform high-entropy visiting probability throughout the day and week with dips in the late night/early morning. More importantly, however, our work is concerned with studying the role of time for reverse geocoding and the different distortion models, not the particular geosocial dataset. Other data sources, e.g., from large-scale transportation surveys, could be used as well. Unfortunately, to the best of our knowledge, no alternative data sources with a similar spatial, temporal, and thematic resolution exist. Finally, the majority of geolocation services target a similar audience to Foursquare. We will revisit the Foursquare bias in the evaluation (Section 4.5).

4.3.4 Constructing Temporal Semantic Signatures

Provided 720 (30 days x 24 hours) *HereNow* values for every POI in the venue set, the values were aggregated by category, hour, and day of the week. The resulting 168 values for each category span every hour of a week. Normalizing

this data by the total number of check-ins for each category shows the check-ins per hour as a percentage of the total week.

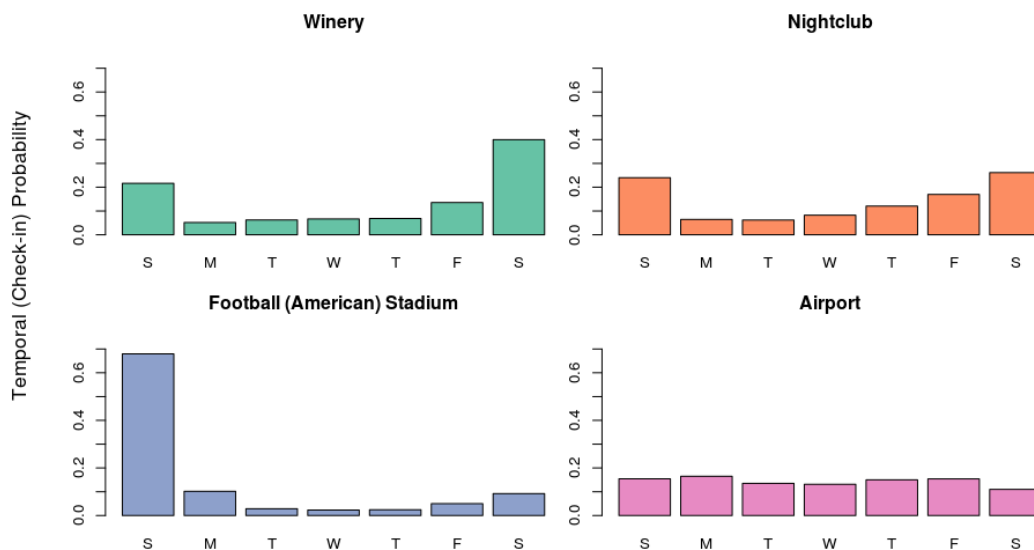


Figure 4.5: Daily temporal signatures for four POI categories.

While these check-in data are limited to a four month time-span, the high resolution allows for temporal signatures to be constructed for each category. In visualizing the temporal distribution of the check-ins grouped by category, one can decipher novel temporal patterns for each category in the set. These are called temporal *bands* and *signatures* in analogy to spectral signatures in remote sensing and follow a semantics-driven *social sensing* approach proposed in previous work (Janowicz 2012b). Figures 4.5 and 4.6 show daily and hourly temporal bands (respectively) for four POI categories that jointly form signatures to uniquely

identify categories via the spatiotemporal behavior of users of location-based social networks.

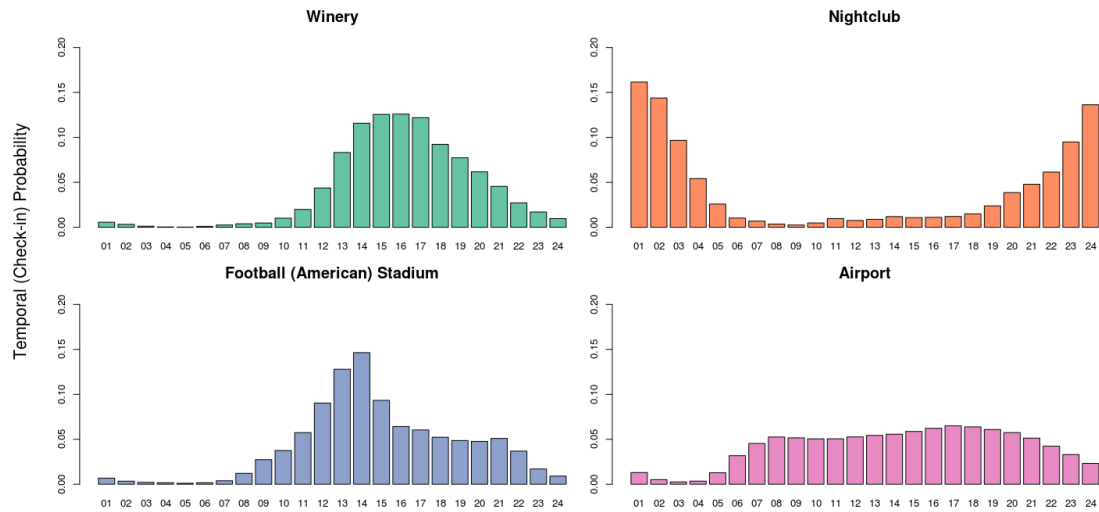


Figure 4.6: Hourly temporal signatures for four POI categories.

Modeling the daily check-in bands separately from the hourly check-in bands exposes some interesting nuances in the data. Both *Wineries* and *Nightclubs* are social and entertainment venues that serve alcohol, and show very similar temporal check-in patterns over a week time period with peaks on the weekend. In contrast, the hourly temporal bands show a very different pattern. These data show Winery visits peaking in the mid-afternoon while nightclub check-ins peak late at night (very early morning). This presents an excellent example of why varying temporal scales are necessary for constructing robust temporal signatures. Figures 4.5 and 4.6 also depict a contrast between activities in which time plays a defining role,

e.g., American football games on Sunday afternoons, and those where temporal aspects are less indicative of a POI type, e.g., Airports.

4.3.5 Indicativeness of Temporal Bands

This leads to the interesting question of which hours and days are most indicative and whether it is possible to compress the bands instead of storing all potentially relevant 168 values per POI type. To investigate this question, we look at the signatures from a classification perspective and consider each band as a discretized feature (attribute) of a class-labeled set of training tuples. Here, we use the entropy-based *information gain* as indicativeness measure. Equation 4.1 shows the computation of Shannon’s information entropy for a distribution D , where p_i is the probability of band i and Equation 4.2 computes the information gain ($\Delta(b_t)$) for a temporal band with $\frac{|D_j|}{|D|}$ being the weight of the j th partition of the training set according to this band. Table 4.2 shows the 10 most indicative hours as well as the 10 least indicative hours. Intuitively, the typical lunchtime hours (11am-12pm), close of business hours (4-5pm), and dinner/nightlife hours (10-11pm) are most indicative of a POI type, as is the distinction between work-days and weekends. In contrast, the early morning hours, e.g., Monday 5am, are significantly less-indicative. Consequently, visiting probabilities at these times

Band	Hour	Info. Gain	Band	Hour	Info. Gain
143	Friday 11pm	0.772	101	Thursday 3am	0.112
59	Monday 11am	0.750	150	Saturday 6 am	0.097
107	Thursday 11am	0.744	124	Friday 4am	0.093
60	Monday 12pm	0.725	26	Monday 2am	0.082
35	Sunday 11am	0.712	27	Monday 3am	0.079
161	Saturday 5pm	0.695	125	Friday 5am	0.063
88	Wednesday 4pm	0.693	28	Monday 4am	0.052
167	Saturday 11pm	0.69	100	Thursday 4am	0.046
142	Friday 10pm	0.688	149	Saturday 5am	0.045
131	Friday 11am	0.687	29	Monday 5am	0.034

Table 4.2: The 10 overall most indicative hours according to their information gain and the 10 least indicative hours.

will not differ substantially between POI type and thus can be pruned without impacting the signatures to save storage or optimize indexing.

$$H(D) = - \sum_{i=1}^n p_i \log_2(p_i) \quad (4.1)$$

$$\Delta(b_t) = H(D) - \sum_{j=v}^n \frac{|D_j|}{|D|} \times H(D_j) \quad (4.2)$$

4.4 Distortion Functions and Weights

In this section we discuss the concrete distortion functions that realize the models presented above as well as the parametrization of these functions.

4.4.1 Spatiotemporal Distortion Functions

In order to combine the temporal signatures with the existing spatial distance-based ranking, we introduce a new ranking distance attribute (d_t) for each POI. This attribute is a distortion of the existing geospatial distance (between the POI and the query coordinates) by a factor of the temporal probability. To determine the value of this new distance attribute, two variables need exploration; the function by which time and distance are combined and the ratio of influence (weight) that both distance and time should have on the new attribute.

A number of methods for combining space and time were explored, the most significant of which are shown in Figure 4.4. As described in the previous sections, initially a *linear* method was employed. First, \tilde{t}' is defined as the actual temporal probability of the POI, t' , subtracted from the mean temporal value, t'_m (Equation 4.3) then a weighted combination of the normalized distance and normalized temporal probability is taken (Equation [linear-type](#)) where w is the assigned weight and d' is normalized geographic distance. This method adjusts the geographical coordinates of the POI by increasing or decreasing the distance between each POI and the query location linearly and symmetrically by the POI's normalized temporal probability subtracted from the mean of the normalized temporal probabilities for all POI in the result set.

$$\tilde{t}' = t'_m - t' \quad \text{Where } t' \in [-1, 1] \quad (4.3)$$

$$d_t = d' \cdot w + \tilde{t}' \cdot (1 - w) \quad (\text{linear-type})$$

While effective, this *linear* distortion approach is restrictive in that it pushes and pulls all POI at the same rate, regardless of their distance from the temporal mean. Another approach is to use a non-linear function, e.g., a *sine* function. It approximates the linear approach as \tilde{t}' approaches zero, but will decrease in magnitude as the values move away from zero. An example of the distortive effects is shown in Figure 4.4. We compute d_t as follows (Equation [sine-type](#)).

$$d_t = d' - \sin(\tilde{t}') \cdot w \quad (\text{sine-type})$$

While appropriate for the data, the sine function still assumes that POI on either side of the temporal mean should be distorted symmetrically. Non-symmetric models are explored as well, to decrease the adjustment of the temporal probability on the positive side of the mean at a greater rate than those values on the negative side of the mean (for instance). We model this by employing a weight-adjusted rational function (Equation [rational-type 1](#)) We also study the inverse effect by using Equation [rational-type 2](#). Relaxing the symmetry requirement is a logical approach to distorting geospatial distance as those POI that are less

probable (of being visited at the given time of the day/week) should arguably be pushed further away from the query location at a higher rate than those being pulled closer.

$$d_t = d' - \left(1 - \frac{w}{\tilde{t}' + w}\right) \quad (\text{rational-type 1})$$

$$d_t = d' - \left(\frac{w}{-\tilde{t}' + w} - 1\right) \quad (\text{rational-type 2})$$

4.4.2 Weights

In the next step we determined the most suitable weight ratio between the normalized distance and the normalized temporal probabilities by using a set of geosocial check-in test data.

Using the *Twitter Streaming API* (Twitter 2014), 3,500 geolocated *Foursquare check-ins* were sampled from within the Greater Los Angeles region between 01:00 on November 1st and 23:59 on November 20th, 2013. The geographic coordinates as well as the category of the POI in which the Twitter user checked in were accessed. The number of check-ins (and the associated POI) were reduced to 2,800 to ensure that only those POI that showed at least 15 other POI within a 100 meter radius were included in the sample. This restriction ensured that the

results were not biased due to a lack of available POI from which the model could make a selection.

The geographic coordinates of these 2,800 check-ins/POI were employed as the base *user* locations from which the geolocation model would be built. In order to mimic the accuracy of a GPS enabled mobile device and arbitrariness in point-feature placement, an location-uncertainty component was introduced. New test locations were drawn from a normal distribution with a mean of 30 meters and standard deviation of 10 meters from the POI's real geographic location. The directional (angular) offset was randomly assigned for each set of coordinates. These coordinate values were taken as individual *user locations* which then formed the basis on which the geolocation model could be trained. As discussed in the introduction this is a conservative estimate of the involved uncertainties.

Provided these test user locations, a baseline test was developed. Each of the 2,800 test locations were queried against a comprehensive set of 15,729 POI and all POI within a 100 meter radius of each queried test user location were returned and ranked by geographic distance from shortest to longest. The ranked position of the POI known to be the user's true check-in location was recorded for each scenario and the *Mean Reciprocal Rank (MRR)* was then calculated for the overall test results. MRR, shown in Equation 4.4, is a statistical measure for evaluating the results of a ranked set of N (Number of POI in this case) responses.

$$MRR = \frac{1}{|N|} \sum_{i=1}^{|N|} \frac{1}{rank_i} \quad (4.4)$$

Using the *distance-only* MRR as a baseline, we tested which combination of weight and function maximized the MRR value, i.e., we quantified the relative importance of time for reverse geocoding as well as the particular distortion model that would yield the best results. Four other sets of MRR values were calculated based on the combination of temporal probability with geographic distance using each of the four functions depicted earlier (Figure 4.4). Each model was tested multiple times with a weight value increasing from zero at increments of 0.1. Figure 4.7 shows that all of the weighted functions out-perform the distance-only method at some point.

To validate this finding and ensure that the selected functions and weights are not merely an artifact of using MRR as the measure, additional rank comparison measures were computed. A sum of the reciprocal rank (SRR) method was explored as well as counting the number of correctly identified POI (rank position 1). Finally, the popular normalized Discounted Cumulative Gain (Equation 4.5) measure was computed for each of the functions where *DCG* is defined by Equation 4.6 and *POIcount* is the number of POI identified at the specified *i*th ranked position. *IDCG* is the *ideal* discounted cumulative gain which in this case is 2,800 given that an ideal result would correctly identify all POI in the

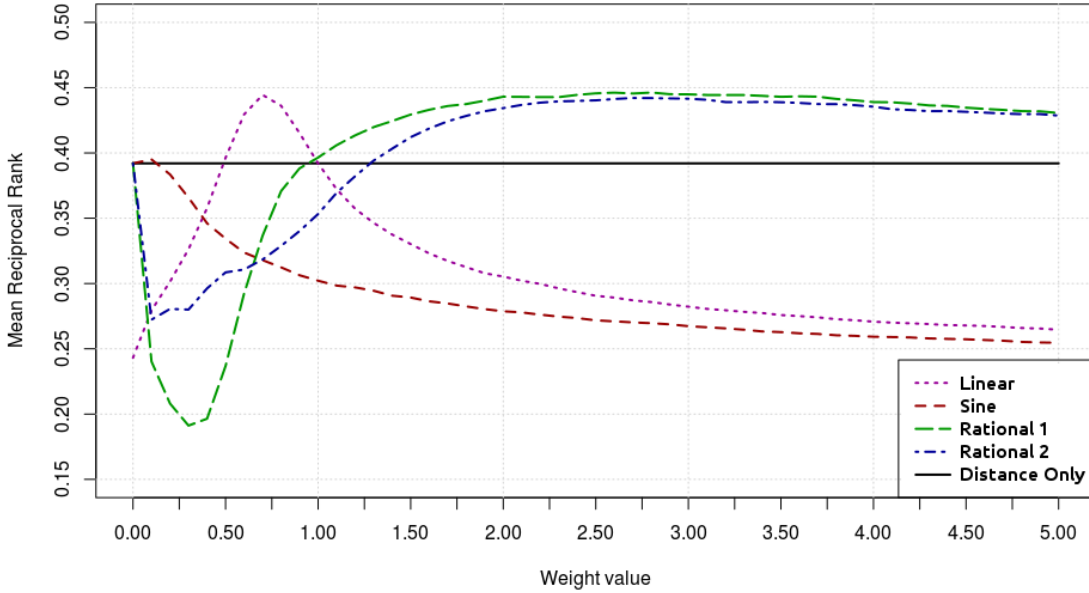


Figure 4.7: Mean Reciprocal Rank for Four Equation and associated weight values compared to Distance-only.

first ranked position. The maximum MMR values, SSR, nDCG and first ranked position count along with their associated weights for each function are shown in Table 4.3 indicating that the Rational 1 based model produces the best overall results with a weight of 2.8.

$$nDCG = \frac{DCG}{IDCG} \quad (4.5)$$

$$DCG = POIcount_1 + \sum_{i=2}^N \frac{POIcount_i}{\log_2(i)} \quad (4.6)$$

Function	Max MRR	Max SRR	nDCG	First Pos.	Weight
Distance Only	0.392	1095	0.621	485	NA
Linear	0.444	1245	0.665	661	0.7
Sine	0.395	1154	0.642	539	0.1
Rational 1	0.446	1250	0.669	665	2.8
Rational 2	0.442	1239	0.662	657	2.7

Table 4.3: Maximum Mean Reciprocal Rank (MRR), Maximum Sum of the Reciprocal Rank (Max SRR), normalized Discounted Cumulative Gain (nDCG), Number of POI ranked in the first position and associated weight for each Equation.

Taking this result, we revisited our running example introduced in Section 4.2 and distorted the query location and the POI locations by shifting them closer or further away. Figure 4.8 depicts this adjustment given a query time of 10 AM on Monday morning. The original distance from the query location to each POI is shown in the table and the original locations are shown as faded markers on the map. The new distorted distances are listed in the table as well as shown on the map via the bright blue markers. By comparison, Figure 4.9 shows the same process for 11 PM on Saturday night. Note that in the original distance-only scenario (see Figure 4.3), the distance to the Bakery (A), the Nightclub (B), and the Italian Restaurant (M) are similar where the distorted cases lists very different distances with the Bakery (A) being nearest in Figure 4.8 and the Nightclub (B) being closest in Figure 4.9. The marker colors of these POI switch from red to green and vice versa between the two figures indicating a pull (green) or push

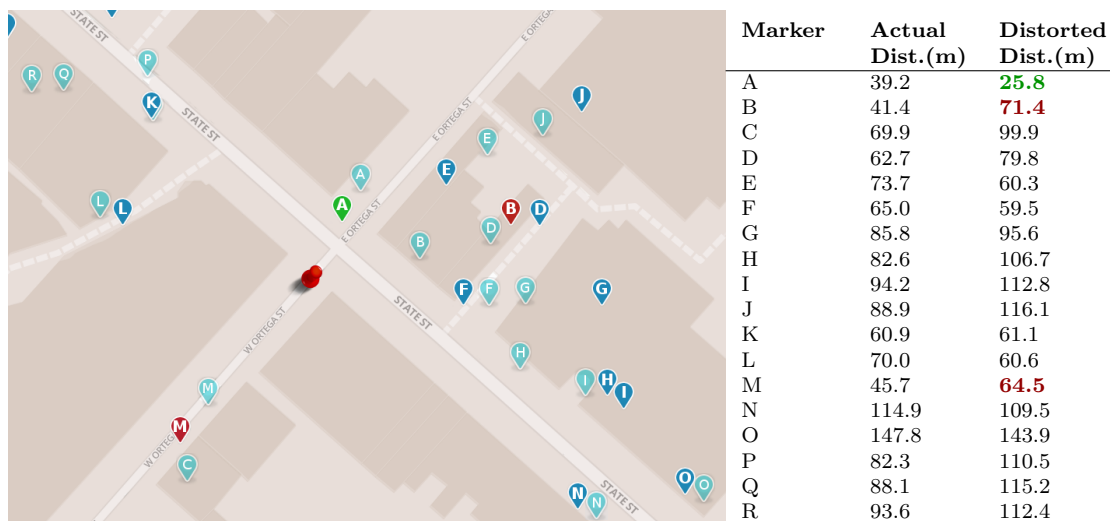


Figure 4.8: Nearby POI locations (dark blue markers) adjusted by temporal probability at 10AM on Monday. Original POI locations visible as light blue markers. Three example locations (A, B, M) are shown in red, indicating pushed further away and green, indicating pulled closer to the assumed user location.

(red) from the query location. Additionally, note that the Italian Restaurant (M) remains red between both figures indicating that it is not a very probable location at either time.

4.5 Evaluation and Discussion

In order to test the validity of the temporally weighted geolocation approach, we designed an experiment with geosocial user data that tests the selected non-linear non-symmetric model with a weight of 2.8 against a distance-only based approach for a new test set of known locations and check-ins.

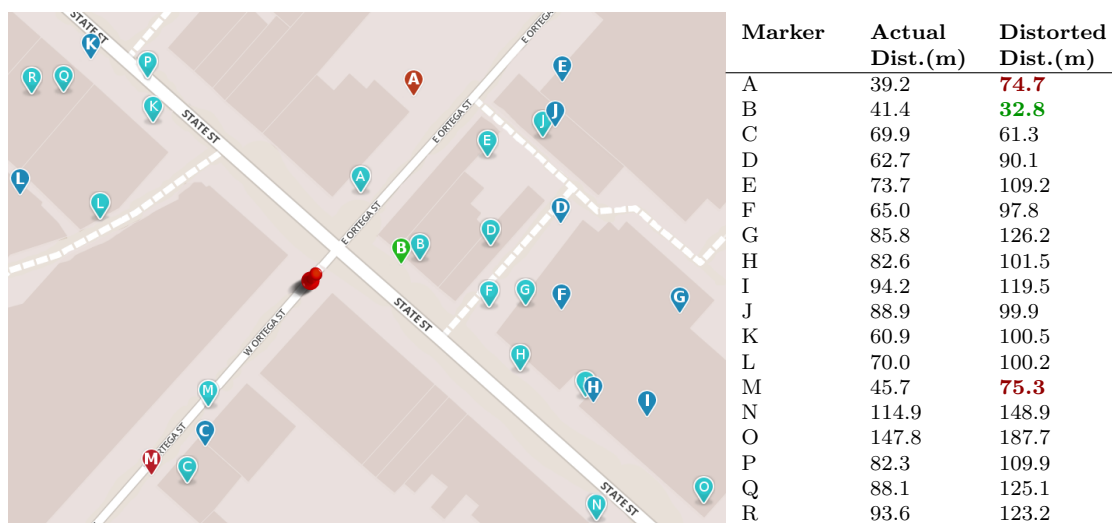


Figure 4.9: Nearby POI locations adjusted by temporal probability at 11PM on Saturday. Original POI locations visible as light blue markers. Three example locations (A, B, M) are shown in red, indicating pushed further away and green, indicating pulled closer to the assumed user location.

Specifying the Greater Los Angeles region as the boundary, the *Twitter Streaming API* was used to collect tweets that shared a Foursquare check-in. When a user of the Foursquare application decides to check-in to a place, they are given the option of sharing this data on their Twitter Feed. While Foursquare check-in data itself is not publicly available, the majority of Twitter feeds are. Using this method, 1,663 unique check-ins were accessed over a 24 hour period.

Immediately on receipt of the check-in data, the geographic coordinates of the POI were randomized using the method described in Section 4.4.2 to reflect standard GPS inaccuracy and a new set of geographic coordinates were established for the user. These *user* coordinates were queried against the Foursquare Venues

API (with the *Intent* parameter set to *Browse*⁶) and a set of 30 nearby POI were returned containing the distance from the query coordinates, *HereNow* (number of Foursquare users currently checked in to the POI), and *TotalCheckins* (total number of all-time check-ins to a specific POI).

Additionally, a separate query was made to the Foursquare Venues API with the *Intent* parameter set to *Checkin*. According to the Foursquare documentation *Browse* takes a distance-only approach to querying the gazetteer returning a set of nearby POI ordered by distance from query location, shortest to longest. The *Checkin* approach is not full explained in the documentation and simply states that the returned set of POI are ordered based on where a typical user is likely to check-in to at the provided latitude and longitude at the current moment in time. This option is most likely based on the company's internal popularity counts. In addition to the *Intent* parameter, each query was executed with additional parameters that specified a radius of 100 meters and minimum of 20 and maximum of 30 nearby POI. This limited bias due to a lack of nearby places.

Provided the set of nearby POI returned for each of the 1,663 queried *user* locations, the *distance-only* method can be compared against our new *temporal signatures* enhanced method. Since the actual POI to which the user checked in is known, it is possible to calculate a number of different measures for each approach.

⁶Foursquare offers four methods for querying their gazetteer: browse, checkin, global and match.

Table 4.4 presents the difference between these two methods across MRR, SRR, nDCG and First positions measures. The table shows that the inclusion of the temporal signatures model with a weight of 2.8, substantially outperforms the *distance only* method over all measures. In fact, the mean reciprocal rank (MRR) values rise from 0.359 to 0.453, an increase of **26.34%** and the nDCG values increase by **21.96%**.

Method	MRR	SRR	nDCG	First Pos.
Distance Only	0.359	443.8	0.583	211
Temporally Adjusted	0.453	793.5	0.711	423

Table 4.4: Comparing the results of the Distance Only method to our method which includes temporal signatures.

Ranking the POI based purely on *TotalCheckins* produces a MRR of 0.678. Such a large discrepancy in numbers between *distance-only* and *TotalCheckins* method is an important reminder of how biased the Foursquare data and its users are, i.e., a very high percentage of the total user base predictably visits a small number of establishments. While *TotalCheckins* works well for an application such as Foursquare, the majority of geolocation services do not rely on a closed community and explicit check-ins from their users, but have to estimate the place based on space (and time) alone. Interestingly, adding our temporal distortion method to *TotalCheckins* can further improve Foursquare’s results. If we use the *TotalCheckins* values in lieu of geographic distance, first normalizing the values

and then subtracting them from 1. This resulted in an MRR value of 0.692 a 2.1% increase over *TotalCheckins* alone.

POI ID	Distance(m)	TotCheckins	HereNow
4bba348c53649c746bc248fb	16	1398	1
4d14fbb981cea35d9e80d7ec	16	705	0
4a52bc1cf964a520f7b11fe3	22	479	0
4af22b13f964a5204be621e3	24	877	0
4acbf6abf964a52077c820e3	29	900	0
51301edfe4b01507da6114f2	37	675	1
516327d7e4b063c6e8320956	41	8	0
4a12b3baf964a5208e771fe3	43	3282	0
4e01174b1f6ef39c29422260	45	2560	0
4cd19cf9f6378cfa8e8abcd6	45	59	0

Table 4.5: Example of Foursquare Search API query results ordered by distance and limited to 10. Known check-in location in bold face.

The *HereNow* approach (ranking POI by the number of users currently checked in) to determining a user’s placial location is self fulfilling. Note that this validation model is based on real check-in data and ranking a set of nearby POI based on the number of users currently checked in will always involve a high degree of bias. The correct POI will always have at least one current check-in. Examining Table 4.5, the influence of this bias becomes immediately apparent. The vast majority of POI do not show a single current check-in with a limited few listing 1. Were this example scenario to be run multiple times, one would expect the known POI to be correctly identified half of the time and the POI ranked 6th in the list (also showing a *HereNow* value of 1) to be identified half of the time.

This *tie*, so to speak, can be broken through the inclusion of temporal signatures. Again, replacing the d' variable with the normalized *HereNow* value subtracted from 1, Equation [rational-type 1](#) is applied resulting in a **3.1%** increase and MRR measure of 0.872.

Lastly, Foursquare's closed check-in method is examined. It must be reiterated that while the Foursquare method does produce a very high MRR value (0.733) it relies on data not available to most geolocation services and involves a significant amount of user bias which is likely exploited by this method ([Shaw et al. 2013](#)). Though its performance is strong, it may be further enhanced through our temporally-enhanced method. In this case, the nearby POI returned from the search query are assigned a rank value based on their order within the set. This ranked value is normalized and assigned to the d' variable in the [rational-type 1](#) equation. The resulting MRR of 0.747 is **2%** higher than the proprietary-only approach showing that even a calibrated, in-house built method can be improved upon through the inclusion of temporal signatures.

4.6 The Next Step: Geosocially Distorting the User's location

The previous sections outline a method for distorting the geographic location of POI based on the temporal probability of an individual visiting these POI as determined by their type. In this section we outline, instead, a model that focuses on distorting the geographic location of the *user's device location* based on the presence of geosocial activity nearby. The geosocial activity referred to in this case pertains to online activities such as *geotagged tweets* and *geotagged Instagram photographs* that do not include placial tags but are tagged with geographic coordinates. Since these posts cannot be directly assigned to POI, they cannot influence the amount and direction by which a POI location is distorted. Instead, these activities impact the ability to geolocate an individual through distorting the actual query coordinates themselves.

Figure 4.10 presents an example scenario. The blue markers on the map indicate the location of POI, similar to figures shown previously. Instagram (camera icon) and Twitter (*t* icon) markers are shown on the map as well. These geosocial activities are collected over a one hour time period. In looking at this map, it is apparent that an event is occurring at the plaza (green region) given the high number of tweet and photo activity in the past hour. Combining this informa-

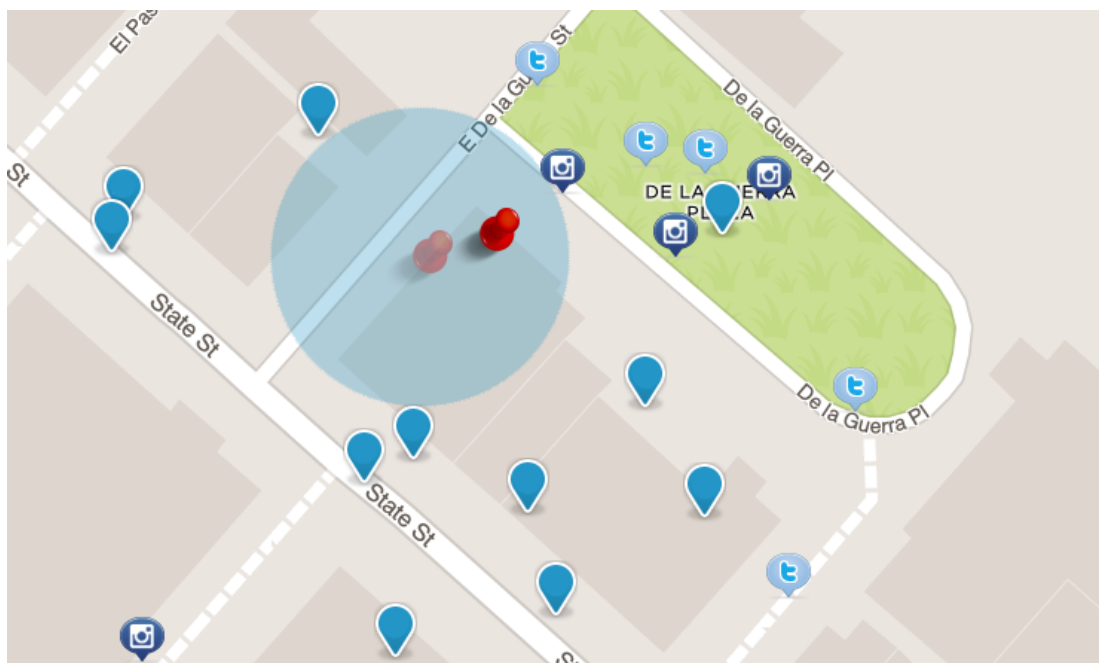


Figure 4.10: Example visualization of a user’s actual location (faded red pin), adjusted location (bright red pin), location uncertainty (large blue circle), Foursquare POI (blue markers) and Twitter and Instagram activity markers.

tion with the knowledge that the user’s query location is subject to uncertainty, adjusting the query location closer to the plaza is a reasonable proposal.

Using vector addition, a new vector is calculated from which distortion direction is ascertained. The *amount* by which the query location is adjusted is based on two factors. An inverse distance weight is calculated for each geo-social activity, assigning a greater weight to nearby activities than those occurring further away. Note that actual content of the tweet or Instagram caption is irrelevant in this approach. While some individual’s may prefer one source of social content over another, for our purposes, all geo-social activities are of equal value, their

influence on the query coordinates are based solely on distance and direction. The second factor influencing the distortion is the global weight value with which the combination of these activities influence the query location. This global weight is the focus of future research and will involve additional training in order to establish an optimal value.

It is worth noting that we do not assume that the presence of a Tweet or Instagram photo in a specific region indicate that this is the only area where an activity is occurring. This approach takes the presence of geosocial data as an additional and readily viable variable that can be employed to better geolocate an individual based on places that are currently *trending*. This method makes the assumption that the presence of a tweet or photo represents an increase in probability that some activity is taking place at this location.

4.7 Related Work

Existing research on user and mobile device specific geolocation services can be split in to two roughly defined groups. One approach focuses on the technical aspects associated with determining one's location, increasing the accuracy of location-based technologies (Tawk et al. 2014, Fallah et al. 2013) as well as enhancing the efficiency of location services on mobile devices (Schilit et al. 2003,

Paek et al. 2010). For better or worse, these advances are reflected in a number of patents filed recently (Zeto III et al. 2013, Brewington et al. 2013). While useful, these approaches do not consider non-technical sources of geolocation information, but instead focus on reducing the uncertainty associated with a device’s geographic coordinates.

The second approach has arisen from place recommendation research. Many of these approaches take advantage of the rise in geosocial check-ins and posts to explore user-similarity, (McKenzie et al. 2013, Cheng et al. 2013, Xiao et al. 2010) as well as user’s home locations (Backstrom et al. 2010, Hecht et al. 2011). Additionally, recent work has begun to explore temporal patterns in user behavior through online social networking check-ins (Gao et al. 2013a, Cheng et al. 2011) as well as human mobility patterns through mobile device tracking (Yuan et al. 2012, Palmer et al. 2013). Shaw et al. (2013) explored the use of check-in data for enhancing venue search results in the Foursquare application. While the authors did investigate both the temporal and spatial components of check-ins, they did so without exploiting category types. Additionally, their methodology for merging spatial and temporal data is sparse and clearly does not consider distorting space by a function of time. Lastly, though their work does produce promising results, these results are specific to the Foursquare application and founded on a level

of data-access restricted to Foursquare employees and thus of limited use to the reverse geocoding community outside of the company.

From a temporal signatures perspective, early work by [Ye, Janowicz, Mülligann & Lee \(2011\)](#) extracted check-in behavior from the online location-based social network *Whrrl* to determine daily and hourly default temporal patterns for a number of *Whrrl* place types. [Yuan et al. \(2013\)](#) took this a step further using these temporal patterns to recommend points of interest based on the time of day. Furthermore, [Wu et al. \(2014\)](#) show how social media check-in data can be used for combining a movement-based approach with activity-based analysis in studying human mobility patterns. In exploring Flickr data, [Hauff \(2013\)](#) recently found that the popularity of venues plays an important role (orders of magnitude) in the accuracy of geotagged Flickr photos. Additionally, a large study on mobile phone usage by [Yuan et al. \(2012\)](#) found unique activity patterns based on age and gender indicating that temporal signatures may differ not only by POI category, but also by visitor demographics.

While much of this work has focused on extracting user behavior from social-sharing platforms, it has been used to estimate, predict or make recommendations on places an individual may have visited (past) or should/may visit (future). To the authors' knowledge, very little research has focused on using existing public, place-based check-in behavior to enhance existing technical approaches to geolo-

cation in real-time. Additionally, no published work can be found that distorts geographic distance by a factor of temporal probability.

4.8 Conclusions & Future Work

The striking increase in location-based mobile applications in recent years is driving the need for better and more accurate geolocation services to the forefront of geo-computational research. Compounded by the inaccuracies of user-generated geo-content, the need for geolocation methods built on more than mere Euclidean distance are a necessity. Online geosocial networking solutions now offer researchers the ability to monitor human activity behavior which supply the foundation for categorically unique check-in signatures. By incorporating these semantic signatures with existing distance-only based geolocation services, more accurate results can be ascertained.

In this paper we demonstrate a novel technique for incorporating temporal signatures with geographic distance by virtually distorting (pushing and pulling) the geographic coordinates of nearby Places of Interest. In order to achieve the highest accuracy, a non-linear, non-symmetric approach was employed significantly outperforming the distance-only based geolocation service. Additionally, this same method was used to enhance existing state-of-the-art check-in and proprietary

methods offered by top mobile applications on the market today. Finally, we outline a method for the enhancement of this method through the use of geotagged social content such as *tweets* and *Instagram photographs*.

Future work in this area will include the continued enhancement and fine-tuning of the existing temporal signature weight and function as well as the inclusion of geosocial activities outlined in Section 4.6. A limitation of this work is evident in the three month span of data collection. An increase in the temporal extent of the data will allow further research into seasonal effects, holidays and climate fluctuation to name a few. Additional work aims to investigate regional variance in categorical-temporal signatures (e.g., Nightclubs in New York vs. Nightclubs in Los Angeles) as well as the influence of local weather patterns and daylight effects. The enhancement of the existing dataset will server to increase the accuracy and robustness of the temporal signatures-based approach. Finally, an online service is in development that will allow interested parties to increase the accuracy of existing services in real-time and over large datasets.

Chapter 5

POI Pulse: A Multi-Granular, Semantic Signatures-Based Information Observatory for the Interactive Visualization of Big Geosocial Data

While the previous chapters primarily focus on a combination of two placial dimensions, namely *Thematic & Spatial* and *Temporal & Spatial*, this chapter employs all three dimensions of place in defining and differentiating place types. Using Points of Interest from the Greater Los Angeles Area, this work takes a *top down* and *bottom up* approach to defining place types based on their spatial, thematic and temporal signatures. Through an interactive web mapping application, the temporal and spatial dimension of the data are displayed, in essence showing the *pulse* of the city. An important finding to take away from this chapter (linking to the overarching theme of the dissertation) is that, of the three dimensions, time is the most indicative in defining place types.

Peer Reviewed Publication	
Title	POI Pulse: A Multi-Granular, Semantic Signatures-Based Information Observatory for the Interactive Visualization of Big Geosocial Data
Authors	Grant McKenzie ¹ , Krzysztof Janowicz ¹ , Song Gao ¹ , Jiue-Ann Yang ^{1,2} , Yingjie Hu ¹
Authors	¹ Department of Geography, The University of Santa Barbara, ² Department of Geography, San Diego State University
Journal	Cartographica: The International Journal for Geographic Information and Geovisualization
Editors	Monica Wachowicz, Emmanuel Stefanakis
Publisher	the University of Toronto Press
Pages	In Press
Submit Date	April 18, 2014
Accepted Date	May 25, 2014
Publication Date	In Press
Copyright	Reprinted with permission from the University of Toronto Press

Abstract

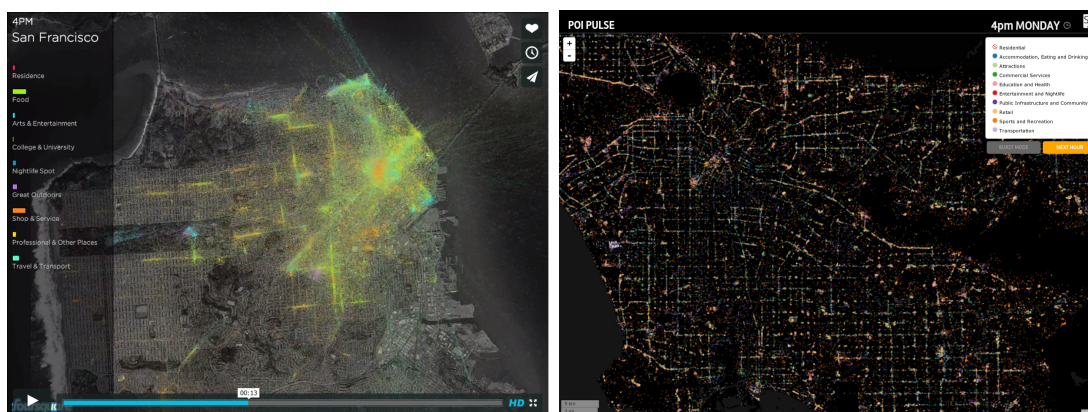
The volume, velocity, and variety at which data are now becoming available allow us to study urban environments based on human behavior at a spatial, temporal, and thematic granularity that was not achievable until now. Such data-driven approaches opens up additional, complementary perspectives on how urban systems function, especially if they are based on User-Generated Content (UGC). While the data sources, e.g., social media, introduce specific biases, they also open up new possibilities for scientists and the broader public. For instance, they provide answers to questions that previously could only be addressed by complex simulations or extensive human participant surveys. Unfortunately, many of the required datasets are locked in data silos that are only accessible via restricted APIs. Even if these data could be fully accessed, their naïve processing and visualization would surpass the abilities of modern computer architectures. Finally, the established place schemata used to study urban spaces differ substantially from UGC-based Point of Interest (POI) schemata. In this work, we present a multi-granular, data-driven, and theory-informed approach that addressed the key issues outlined above by introducing the *theoretical and technical* framework to interactively explore the pulse of a city based on social media.

5.1 Introduction and Motivation

Today's data universe offers access to a plethora of data at a spatial, temporal, and thematic resolution unthinkable just a few years ago. This data revolution is accompanied by the emerging 4th paradigm of science (Hey et al. 2009, Elwood et al. 2013) in which synthesis is the new analysis. Those changed realities cast off visions of *information observatories* (Tiropanis et al. 2013) in which complex systems, such as urban spaces¹, could be observed and better understood based on exploiting the variety, volume, and velocity of Big Data (MacEachren et al. 2011, Leetaru et al. 2013). Those, however, who tried to explore these new possibilities often encountered equally big challenges. First, major parts of Big Data still reside in closed proprietary silos with limited API access. Second, the metadata, e.g., provenance, and conceptual schemata required for any serious use by scholars are often not present, intransparent, or differ substantially to those established in science. Finally, the sheer volume and velocity makes interacting with or even just visualizing the data difficult to say the least.

For many of us, an information observatory for urban spaces in which user-generated *real-time* content reveals spatial, temporal, and thematic patterns and traits of human behavior, is a tempting idea as it aligns well with the Digital Earth

¹See <http://www.urbanobservatory.org/compare/index.html> for an early example.



(a) Screenshot from *Foursquare* video (b) *POI Pulse* interactive visualization

Figure 5.1: The pre-generated video (a) and the interactive *POI Pulse* system (b).

vision. Consequently, a posting² on Foursquare’s infographics blog in October 2013 raised a lot of attention. It linked to a series of videos showing the *pulse* of different cities such as San Francisco. The animations were entirely derived from mining massive amounts of user check-ins to the Foursquare Location-based Social Network and were aggregated to a single virtual day; see Figure 5.1a.

While the visualization itself is quite stunning, the Foursquare videos have several shortcomings: **(I)** The videos are not interactive, e.g., one cannot click at any of the check-in events or places to gain additional insights.³ **(II)** The videos are rendered based on a fixed geographic scale and focused on a particular part of the city. Thus, one cannot pan or zoom. **(III)** The millions of check-ins are

²<https://foursquare.com/infographics/pulse>

³Interested readers may try to find an explanation for the moving *Food* cluster in San Francisco at 4am; see <https://foursquare.com/infographics/pulse#san-francisco>.

aggregated to a single non-specific day, thus hiding well known patterns, e.g., weekdays versus weekends. **(IV)** Foursquare’s POI taxonomy consists of more than 400 POI types grouped into 9 top-level classes (see Figure 5.1a). While such generalized classes are necessary and useful, it is not clear how they were derived nor why certain POI types are categorized in specific ways. Furthermore, a binary class membership on such a coarse level will necessarily introduce arbitrary decisions and thus will significantly alter the observed temporal pulse of the city. For instance, *Cemeteries* are categorized under the *Great Outdoors* category. **(V)** Similar to other UGC, Foursquare contains data of widely varying quality. For instance, users often classify their own houses as *Castle* or check-in to features of types *Road*, *Trail*, or *Taxi*. While this is a consequence of UGC, it is important to clean the data before doing any serious analysis.

Inspired by Foursquare’s pulse videos and the theoretical and technical limitations of interacting and visualizing Big Data, we decided to address the aforementioned restrictions by designing a *POI Pulse* information observatory for Los Angeles;⁴ see Figure 5.1b. Naturally, as scientists we are more interested in those theoretical and technical aspects than the application as such, but we will use it as the joint **leitmotiv** that connects the following **research questions** which make up the scientific contribution of this work:

⁴Explore the portal at <http://www.poipulse.com>

$\mathcal{R}1$: Given the >400 POI type defined by Foursquare users, is it possible to derive an alternative top-level classification that is informed by existing and well established POI schemata (e.g., defined by Ordnance Survey) and still true to the original Foursquare data and user-behavior?

$\mathcal{R}2$: Most likely, the reason for showing a pre-rendered video is the fact that even the most modern Web browser using HTML5, CSS3, and effective JavaScript engines, cannot render the hundreds of thousands of POI as vectors thus making interaction cumbersome. Is it possible to use a scale-dependent, seamless combination of raster and vector tiles to render approximately 200,000 POI for Los Angeles, and still make the interface interactive and responsive? What is the tipping point from which vector tiles will be faster than raster tiles?

$\mathcal{R}3$: Given the legal API limits of closed data silos such as Foursquare, can we generalize check-ins, individual POI, and their attributes, e.g., tips, to a type-level *default behavior* that allows us to model the pulse of a city with minimal data requirement? Is it possible to seamlessly switch to a real-time, *burst* mode at zoom scales that do not exceed the daily API limits and thus also give access to real time data?

$\mathcal{R}4$: Can we improve on the Foursquare baseline by offering a pulse for all hours of the full week instead of a single day? Can we show binary upper-level

categories but seamlessly switch to a more nuanced view at a reduced zoom level to show a probabilistic category membership?

In the following, we present a multi-granular, data-driven, and theory-informed⁵ approach that addresses these research questions by introducing the *theoretical and technical* framework to interactively explore the pulse of a city based on social media.

5.2 A Data-Driven and Theory-Informed POI Taxonomy

In this section, we discuss how to derive a POI taxonomy by combining data-driven techniques with existing top-down classification schema. Many different POI vocabularies, taxonomies, and schemata have been defined in the past few years, e.g., schema.org, the Ordnance Survey POI classification system, the OpenStreetMap map features, OpenCyc, the Linked Geo Data ontology, the GeoNames ontology, or even WordNet, to name a few. Unfortunately, most of these are not suitable for our purpose. Sources such as WordNet are not specific enough, while platforms, such as OpenCyc, introduce distinctions (e.g., *man-made structure*) that are interesting from an ontological perspective but hinder the task at hand.

⁵I.e., including existing top-down schemata from the research literature and industry.

OpenStreetMap is notorious for its flat key-value pair classification and also introduces many feature types that are not POI specific. Similarly, the GeoNames feature classes are not suitable, since all POI types defined in Foursquare would end up in the same class (*S spot, building, farm*).

Schema.org is a data markup ontology jointly constructed by Google, Yahoo, Microsoft, Yandex, and the W3C. Intuitively, one would assume that such an ontology is most suitable to provide an upper-level abstraction for the >400 Foursquare types and should be able to replace the 9 top-level classes. One may also expect that schema.org was developed with datasets such as Foursquare, Yelp, etc, in mind. Surprisingly, however, that turned out not to be the case. For instance, schema.org distinguishes between *Places* and *Organization* as one of its top-level distinctions. While this is not wrong, the fact that Internet cafes are considered organizations but movie theaters are places is surprising.⁶ Due to many similar cases and ontological decisions taken by schema.org, it became clear that we needed another classification.

Eventually, we selected the Ordnance Survey (OS) POI classification system (v. 2.3) (OrdnanceSurvey 2014). In contrast to Web and data-driven resources, the OS classification is an administrative and UK-specific resource. The OS system

⁶Via: Thing >Organization >LocalBusiness >InternetCafe (see <http://schema.org/InternetCafe>) and Thing >Place >CivicStructure >MovieTheater (see <http://schema.org/MovieTheater>).

consists of 9 classes at the 1st level, 49 classes at the 2nd level, and 600 POI types at the 3rd level. We are only interested in the first level here (OS1). It consists of the following classes: **01** *Accommodation, eating and drinking*, **02** *Commercial Services*, **03** *Attractions*, **04** *Sport and entertainment*, **05** *Education and health*, **06** *Public infrastructure*, **07** *Manufacturing and production*, **09** *Retail*, and **10** *Transport*. We will use this classification as our top-down, theory-informed POI schema and in the following section describe how to use data-driven techniques to semi-automatically align the Foursquare types to this schema.⁷

5.2.1 Multi-dimensional Characterization of POI Types

The variety of big data presents new possibilities to understand POI from different perspectives. In previous work, we proposed the concept of *semantic signatures* to characterize a place using spatial, temporal, and thematic patterns (Janowicz 2012a). As an analogy to *spectral signatures* in remote sensing, *semantic signatures* differentiate types of places based on multiple *bands*. In this work, we employ *semantic signatures* and extract a number of descriptive dimensions from the Foursquare data to characterize POI by social sensing.

⁷To improve readability, we will refer to the Foursquare classes as *POI types* and to the OS1 classes as *upper-level classes*.

Temporal Bands

The temporal bands are derived from 3,640,893 check-ins to 938,031 venues from 421 Foursquare categories in Los Angeles, New York City, Chicago, and New Orleans. These check-ins have been collected for 4 months starting October 1st, 2013. Consequently, we cannot use them to understand seasonal effects but focus on the 168 hours of the week. The temporal resolution of the data is 2 hours, i.e., while we have hourly check-in times, the duration of check-ins is unknown and users are automatically checked out after 2 hours. In our work, we are neither interested in the particular venues, check-ins, nor users,⁸ but in studying the temporal **default behavior** of users towards **types** of POI. In other words, we are interested in the fact that bars are visited in the evenings and especially during weekends, while universities are mostly visited during the workdays between 7am-5pm. Figure 5.2, depicts 168 bands that jointly form the temporal signature for four POI types. The data represents probability values for check-ins to the given type (by hour bins), i.e., the 168 bands sum up to 1. Despite the large sample, we had to remove outliers as some of the POI types, e.g., *Molecular Gastronomy Restaurant*, have fewer venues than others. We used 4 standard deviations from the mean as cutoff. While we have not used these temporal bands before, we applied a coarser and more limited temporal signature to predict types

⁸Even more, due to API restrictions these data should not be stored for more than 24 hours.

for untagged POI successful (Ye, Shou, Lee, Yin & Janowicz 2011). Thus, we expect the temporal bands to play a major role in the derivation of the POI taxonomy.

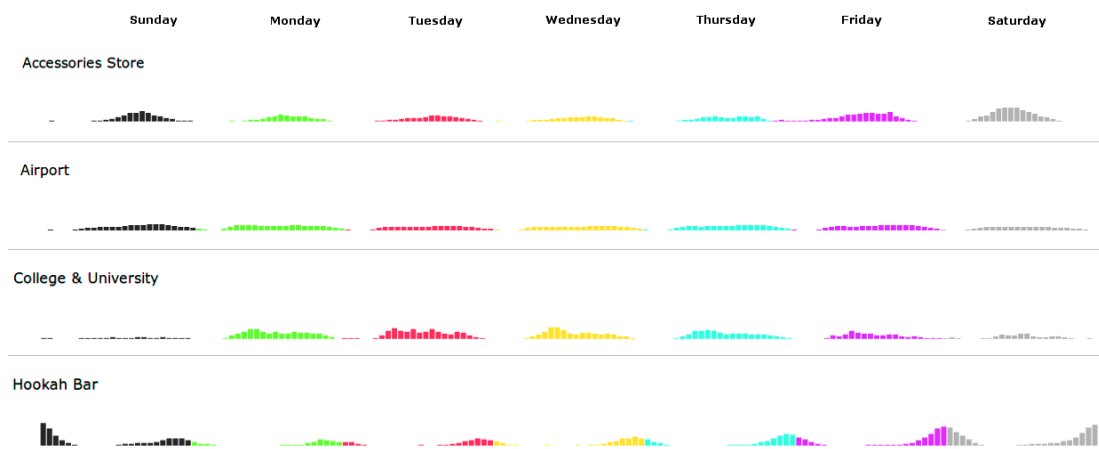


Figure 5.2: The weekly temporal bands for selected POI types by hour.

Thematic Bands

A representative subset of the venues (274,404) from the 421 Foursquare categories (POI types) have been used to derive another, yet very different set of bands that will jointly form the thematic signature; cf. (Tanasescu et al. 2013). We collected all user-contributed tips for those venues, stemmed all words, generated venue-specific documents out of them, grouped these documents by POI type, and then used Latent Dirichlet Allocation (LDA) (Blei et al. 2003)⁹ to ex-

⁹Due to the API restrictions, we are only storing the derived latent topics per POI type.

tract topics. LDA is an unsupervised, generative probabilistic model used to infer latent topics in a textual corpus. We trained LDA by treating the tips associated with all venues of a given type as single documents. LDA uses a *bag-of-words* approach to uncover topics that are represented as multinomial distributions over words. Each topic is composed of multiple words and their relative importance for this topic. Figure 5.3 uses word clouds to visualize the top 18 words in three topics by scaling them according to their probability. It is important to note that each stemmed word extracted from the tips appears in each topic with a different probability. LDA topics do not necessarily correspond to themes typically formed by humans.



Figure 5.3: Words that make up LDA topics scaled by their relative probability.

Topic 53, for example, is interesting as it prominently contains Spanish terms, while topic 26 is related to terms about markets, flowers, plants, and so forth. We are **not** interested in the specific terms but only their indicativeness, i.e., how diagnostic they are in predicting the type of place. For instance, topic 53 is more

likely to appear in relation to POI *types* such as *Mexican Restaurant* than within tips contributed to *Yoga Studios*.

Spatial Bands

The spatial distribution patterns of POI types in urban areas differ. To achieve a more holistic signature of the POI, we also introduce 14 spatial bands. The first set of bands is derived from the average of nearest-neighbor distances (ANND) among all POI types. The values have been normalized to [0-1] such that larger value indicates dispersion while the smaller value represents clustering. The next set of bands are derived from Ripley's K which offers the potential for detecting both different types and scales of spatial patterns. The K measure computes the average number of neighboring venues (of the same type) associated with each POI within a given distance and then compares them to the expected value under completely spatial randomness. We chose 10 distance thresholds and calculated the corresponding Ripley's K measures as 10 spatial bands for all POI types. Figure 5.4 shows that the K measure helps to evaluate how the spatial clustering or dispersion pattern of each POI type changes when the neighborhood distance changes. For instance, The values of ANND *Police Station* (0.721) and *Night Club* (0.702) are very close, while their spatial clustering patterns are different at multi-distance bands (scales). Both ANND and Ripley's K measures only

consider the distance or the number of neighboring venues but ignore the POI type information for spatial point pattern analysis. In urban areas many POI types (such as nightclubs and bars) often clustered together. The different types of spatial mixture patterns should also be taken in to consideration. To address this issue, we introduce a third family of bands called the *J Measure*. The *J Measure* involves generating a Delaunay triangulation between all POI of the same type, counting the number of other distinct POI types within each triangle and dividing it by the total number of POI types. We computed the mean, median, and standard deviations for the *J Measure* for all POI types. For instance, the mean *J Measure* for *Police Station* (0.257) is larger than that of *Night Club* (0.176), which indicates a larger POI type diversity between adjacent police stations.

5.2.2 Data Cleaning

As a next step we cleaned the dataset by removing all POI types that either refer to clearly linear features or were overly generic. Examples include types such as *Road* and *Trail*, and non-descriptive types such as *Building* or *City*. We also removed types that are a pure artifact of UGC and that we know have no instances in the Greater Los Angeles, e.g., *Volcano*. Similarly, we removed the type *Castle*, assuming that the 77 POI within the dataset are from user's that took the "my home is my castle" motto too literally. Finally, we removed clearly

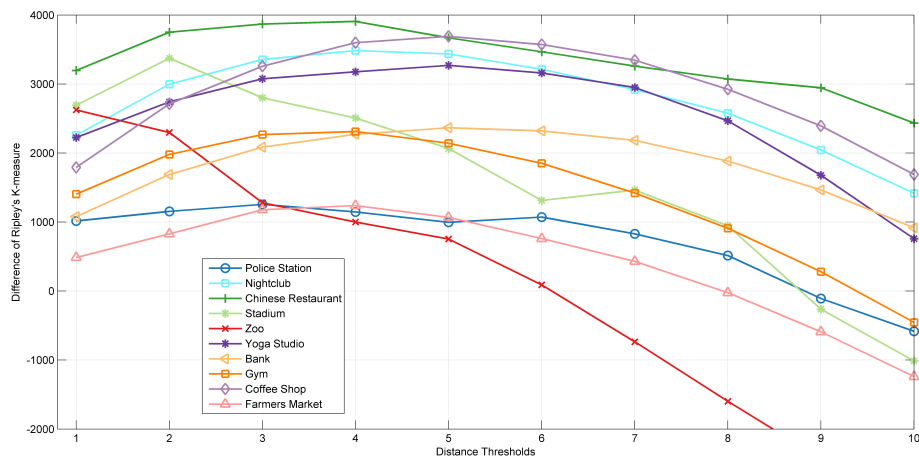


Figure 5.4: Ripley’s K for 10 types (The y-axis value represents the difference between the observed value of K-measure at a given distance and the expected value under the CSR simulation process.)

non-stationary POI such as *Plane* and *Taxi*, while leaving the *Food Truck* type in the dataset. This reduced the number of Foursquare POI types from 421 to 387 and the LA POI dataset from 178,814 POI to 164,902.¹⁰

5.2.3 Information Gain

Given the 246 different bands that jointly form our semantic signatures, it is interesting to discover which of these bands are most diagnostic in terms of their ability to estimate the membership of a particular POI type with respect to an upper-level class. This is for two reasons: First, it allows us to reduce the high-dimensional space by excluding dimensions/features that do not contribute

¹⁰Our original dataset contained 191,998 POI but this included uncategorized POI as well.

(much) to the classification; Second, it provides intuition about the expected bias, and, thus, the limits of data-driven classification. For instance, if most of the thematic bands were not diagnostic, it would be difficult to tell apart the *Airport* type from the *Emergency Room* type as the two share similar temporal signature, i.e., people visit them during all hours and days of the week. Hence, both types would more likely be classified as belonging to the same class, e.g., *Education and Health*, while they should belong to two distinct classes namely *Transport* and *Education and Health*.

Information gain is a measure of the expected decrease in entropy (Russell & Norvig 2010). It provides an assessment of the contribution of a particular feature (i.e., a specific band) for predicting the dependent variable, i.e., the upper-level class. To compute the information gain of the 246 bands, we jointly agreed on a set of POI types considered as clear matches for the respective OS1 classes. For instance, *German Restaurant* was manually classified as being a subtype of *Accommodation, Eating and Drinking*. Next, the information gain for all (discretized) bands was computed using this training set and the median and arithmetic mean scores were determined. Assuming that simpler models can better capture the underlying structures (Russell & Norvig 2010), all bands with information gains scores below the mean were removed, leaving 159 bands that were considered diagnostic.

Band	Information Gain	Band	Information Gain
temp143	0.772	temp161	0.695
temp59	0.750	temp88	0.693
temp107	0.744	theme39	0.519
temp60	0.725	spatial4	0.234
temp35	0.712	temp29	0.034

Table 5.1: The 7 overall most diagnostic bands according to their information gain, the most diagnostic thematic and spatial bands, and the least diagnostic band.

Table 5.1 shows some of the results. It is interesting to note that all top bands are temporal. In fact, the first non-temporal band (*theme39*) is ranked 56th. This thematic band is graphically represented in Figure 5.3a. The first spatial bands (*spatial4*) is ranked 134th. Examining the top temporal bands shows that the typical lunchtime hours (11am-12pm), close of business hours (4-5pm), and dinner/nightlife hours (10-11pm) are most relevant, as is the distinction between workdays and weekends. Band *temp143*, for instance, corresponds to Friday 11pm while the least diagnostic band (*temp29*) corresponds to Monday 5am. Consequently, while all 159 bands will contribute to the classification, we can expect the classifier to have more difficulties in learning the membership for classes such as *Public Infrastructure* that consist of POI types with widely varying temporal bands, e.g., *Police Station* versus *City Hall*. This will result in lower precision and recall values for such upper-level classes and will be discussed in the following sections. One could, of course, consider and extract additional bands.

However, this is out of scope for the paper at hand and significantly restricted by the availability of attribute data from typical POI data sources.

5.2.4 Interactive Classification

The creation of bands and their reduction via information gain set the stage for classifying the POI types from Foursquare using the Ordnance Survey level 1 classes. To do so we used a combination of machine learning-based classification and manual corrections in two different runs.

First, we selected the previously generated training set of POI types and trained a Support Vector Machine (SVM) (Cortes & Vapnik 1995) with a polynomial kernel. Next, we **predicted** the OS1 classes of all POI types using the same training set. We check all cases where the assigned and the predicted classes varied and decided manually which class to use. Interesting examples where we changed our initial decision include *Bagel Shop* that we initially classified as *Retail* while they are rather a breakfast place (thus, *Accommodation, eating and drinking*) in the US. Similar cases included *Brewery*, *Nail Salon*, and other POI type that could be categorized as belonging to different classes. Another good example are all college buildings. For instance, should *College Football Field* be categorized as *Education and Health*, *Sports and Entertainment*, or *Attractions*? From the point of view of a social check-in application such as Foursquare, the

number of users that view a football field as an attraction is orders of magnitude above the actual players (for which the football fields should belong to the sports class). We will address this multi-class nature of many POI types from a visual perspective in section 4.

Finally, we trained the SVM with the new training set and subsequently with all POI types. We computed the recall and precision for this run and manually inspected all mismatching class predictions. This led to some interesting findings about the bias in the Foursquare data, its crowd-sourcing nature in contrast to the administrative OS level 1 classes, as well as socio-political differences between the US-based type data and the UK-based schema. For instance, according to the OS classification *Recycling Facility* should be categorized as *Public Infrastructure* while they are *Commercial Services* in the US. Other interesting cases included *Public Art* that SVM successfully categorized as *Attraction*, or *Tailor Shop* that was predicted to belong to *Retail* (but could also have been a *Commercial Service*).

After inspecting the predicted class membership probabilities for each single type, we realized that based on the nature of the Foursquare POI types as well as the previously mentioned bias in our 159 most diagnostic bands, we would need to change some of the OS level 1 classes. We decided to remove the *Manufacturing and production* class as it only has three subtypes in our dataset, renamed *Public infrastructure* to *Public Infrastructure and Community* to also include religious

Target Class	F1	Precision	Recall
Accommodation, eating and drinking	0.8343	0.869	0.8022
Attractions	0.6479	0.561	0.7667
Commercial Services	0.5882	0.6667	0.5263
Education and health	0.7792	0.7692	0.7895
Entertainment and Nightlife	0.8235	1	0.7
Public Infrastructure and Community	0.5946	0.6471	0.55
Retail	0.8611	0.8267	0.8986
Sports and Recreation	0.7568	0.6774	0.8571
Transport	0.6957	0.7273	0.6667

Table 5.2: F-score, Precision, and Recall for upper-level classes after the 2nd run.

places, and finally split *Sport and entertainment* into two distinct classes: *Sport and Recreation* as well as *Entertainment and Nightlife*. As the temporal bands were found to be the most diagnostic features in our dataset and as we wanted to show the pulse of a city by hours and days, a joint class for POI types such as *Basketball stadium*, *Martial Arts Dojo*, and *Strip Club* was not feasible.

Target Class	A.E.D.	Attr.	Comm.	Edu.	Entert.	Public	Retail	Sports	Trans.	#
Accommodation, Eat, Drink	73	6	0	1	0	0	6	5	0	91
Attractions	0	23	0	0	0	0	4	3	0	30
Comm. Services	0	4	20	5	0	3	1	2	3	38
Education, Health	0	0	3	30	0	3	0	2	0	38
Entertainment, Nightlife	10	0	0	1	28	0	0	1	0	40
Public Infrastructure, Community	0	2	1	1	0	11	1	4	0	20
Retail	0	3	1	0	0	0	62	3	0	69
Sports, Recreation	1	1	3	1	0	0	1	42	0	49
Transport	0	2	2	0	0	0	0	0	8	12
#	84	41	30	39	28	17	75	62	11	387

Table 5.3: Confusion Matrix after final class predictions.

The second run consisted of a new training set based on the new upper-level classes and the lessons learned from the first run. We trained a SVM and predicted class membership for the training set as well as all other POI. The F-score, precision, and recall for this 2nd run are listed in Table 5.2. While the results for the new *Entertainment and Nightlife* class or the OS1 class *Accommodation, eating and drinking* are *very high*, other classes are more difficult to predict. This is largely due to the heterogeneity within such classes as well as the fact that some POI types cannot be distinguished based on the temporal, thematic, and spatial signatures. The class *Public Infrastructure and Community* offers good examples of this, and thus, has a relatively low F-score. The class includes POI types such as *Police Station*, *City Hall*, and *Mosque*, that vary substantially with respect to all bands. Finally, some POI types would require very different bands for their successful classification, e.g., sentiment analysis could be used to better distinguish police from fire stations. Table 5.3 shows a confusion matrix to give an overview of the varying classification success.

Figure 5.5 shows a fragment of a Multi-Dimensional Scaling plot. Each node corresponds to a types, while colors indicate class membership. The lines represent the top 20 % most similar pairs, while the node sizes indicate Kruskal stress. Classes such as *Accommodation, eating and drinking* (blue) and *Entertainment and Nightlife* (yellow) form densely connected clusters while other classes, e.g.,

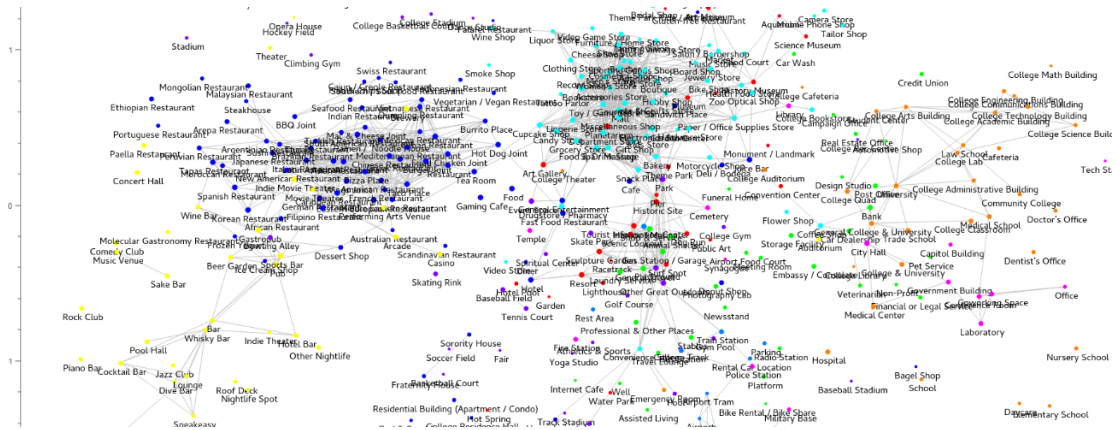


Figure 5.5: Fragment of a Multi-Dimensional Scaling plot showing the upper-level classes as colors; Kruskal stress: 0.258.

Public Infrastructure and Community (pink) are less coherent. This essentially confirms our findings visually.

Summing up, we derived a new upper-level classification schema based on an existing top-down schema as well as a data-driven way in which we let the data speak for itself to inform (and in most cases decide on) our final classification.

5.3 Interaction and Visualization – Rasters vs. Vectors

Once the upper-level classes were established and the POI types were successfully classified, the focus shifted to the challenge of visually rendering the over 165,000 individual POI. One of the primary issues that continues to plague web

mapping and cartography, is the speed at which data can be displayed. Recent advances in browser technology have allowed for dramatic changes in the way data can be processed and visualized. From a web mapping perspective, this increased reliance on browsers allows for improved interaction with data, reducing the need for continual client-server requests. Given the large amount of data and interactivity needs of this project, optimization of the web mapping component is essential. This section presents an efficient method for visualizing this big geosocial data.

Since a true city pulse requires equal contribution from all POI, an early decision was made not to cluster or reduce the POI when constructing the visualization. This created a challenge in determining an efficient means for visualizing approximately 165,000 points through a web browser. The state-of-the-art for many years has been to serve a collection of static image tiles pre-rendered by a mapping toolkit. The structure of these tiles typically follows a simple coordinate system. Each tiles has a Z (map scale) coordinate and an X and Y coordinate that describe its position within a square grid. For every Z-level increase, the number of tiles required increases by a factor of four, leading to extremely large tilecaches, depending on the number of zoom levels and extent of the area of interest. For reasons of practicality, most mapping applications restrict the number of tiled zoom levels to 20. Zoom level 0 represents the entire world in a single tile while level 19 projects the Earth at a map scale of 1:1,000. The power of the

image tiling scheme is that the size of the image file transmitted to the client is minimally influenced by the size of the data.

Recent W3C standards, such as *HTML5*¹¹ and *Scalable Vector Graphics (SVG)*¹², combined with powerful modern web browsers continue to push the boundaries of what can be done in web cartography. While image tiles allow cartographers to solely *display* content via the web, *Vector Tiles* offer users the enhanced ability to interact directly with the content. Vector tiles take a similar approach as image tiles in that they divide data in to smaller sizes in order to enable faster loading times leveraging modern browser parallelization and asynchronous data requests. Unfortunately, the enhancement of offering direct data interaction also increases the burden on the client side as data rendering is now being executed locally. Thus, it is important to study the *tipping point* at which a web mapping framework should switch between raster and vector tiles.

5.3.1 The Tipping Point

Ideally, vector tile representations of POI should be displayed at all map scales allowing for maximum interaction with the data. An experiment was run on three different networks in which both Vector and Raster representations of the POI dataset were loaded. Each tile format was loaded 200 times at each of the zoom

¹¹<http://www.w3.org/TR/html5/>

¹²<http://www.w3.org/TR/SVG11/>

levels between 10 and 16. The loading times (in milliseconds) were recorded and averaged and the results are shown in Figure 5.6. As one can see, the loading time required to display all POI (Zoom level 10) in vector format is simply not practical. With each increase in zoom level, the transfer/rendering time for the vector tiles decreases. Only those tiles that intersect with the view-port are transferred to the browser and rendered. This means that fewer and fewer points are displayed, reducing the amount of data to be transmit and rendered.

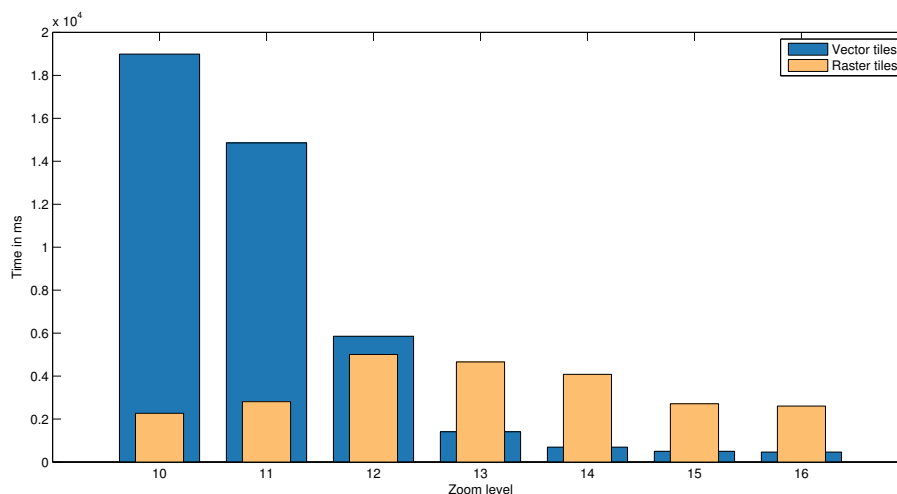


Figure 5.6: Raster vs. Vector tile loading times (in ms).

Comparatively, the loading times for image tiles (PNG-8) increase between zoom levels 10 and 12 and steadily decrease thereafter. This initial increase can be explained by the number of rendered tiles. If a given view-port requires 32 tiles to cover the entire map display, for example, only 10 of these tiles will contain

data at zoom level 10. This means that 22 tiles do not need to be transferred from the server or rendered by the client. As zoom levels approach 12, the number of tiles that render content increases until all 32 visible tiles contain some POI. The reduction in loading times between zoom levels 12 and 16 can then be explained by the amount of data rendered in each PNG. As the map scale increases, the total number of POI visible in the view-port decreases, indicating that the average number of POI rendered in each PNG approaches zero.

Given the significant disparity in loading times, the POI can clearly not be rendered solely in vector format. With the purpose of reducing loading time and maximizing user-data interaction, the switch in tile formats should occur between zoom levels 12 and 13. While loading time is a key contributor to this decision, the total number of POI is also relevant. Interacting with POI data at a map scale smaller than 1:70,000 is simply not possible as the ability to select an individual POI at this scale is arduous.

5.3.2 Technology

Adopting this idea of switching between vector and raster representation, the POI Pulse application is implemented. This platform is built using a novel combination of web technologies. Based in HTML5 and Javascript, the *Leaflet* v0.7 Javascript framework is employed for map display and interaction. The *Data-*

Driven Documents (D3) (Bostock et al. 2011) v3.4 library is used for manipulating and rendering data in Scalable Vector Graphics format through Javascript. On the back-end, *TileStache* v1.49 extracts the data from a PostGreSQL 9.2/PostGIS 2.0 spatially enable database, and organizes the POI in the file structure required by Leaflet. *TileMill* and *CartoCSS* are employed for cartographic styling exported as XML. *MapNik* v2.2 reads these style documents and renders image tiles while *TopoJSON* (Bostock & Davies 2013) v1.4.2 vector tiles are generated through *TileStache* and rendered on-the-fly with D3 and cascading style sheets (CSS). On page load, the map shows all POI in white with an opacity value used to indicate popularity for each hour of the week. For zoom levels 11 and 12, diverging colors, selected through the *ColorBrewer* application (Harrower & Brewer 2003), were assigned to the ten upper-level classes. In order to allow visibility control for each class, ten separate image tile caches were created.

Respectively, vector tiles are not pre-rendered and thus do not require numerous tile caches. Since vector tiles contain the raw geographic location and attributes, these data can be rendered on the client. Toggling the visibility of classes within the vector tiles is handled by iterating through the vectors and changing the visibility parameter for the appropriate SVG element. Restricting the zoom levels at which these tiles are rendered means that a limited number of POI require real-time rendering, but this feature requires the generation of a large

number of vector tiles. Remember that the number of tiles required at each zoom level increases by a factor of four as the zoom level increases. Excluding empty tiles, this means that level 13 requires 704 vector tiles while level 14 requires 2,666.

5.3.3 Pre-loading Map Tiles

Once the image and vector tiles have been generated, implementing them in the user experience becomes the next challenge. The temporal nature of these data require that a new set of tiles be displayed to the viewer with each click of the hour-advancement button; see Figure 5.1b. Regardless of the number of tiles or tile size, the process of adding tiles to the map always takes the web mapping framework a split second to organize the tiles and display them. The most common technique in viewing time series data is to procedurally remove one set of tiles from the map and add another set. In theory, this makes sense, but in practice, this process produces a moment where neither the previous tiles nor the new tiles are visible on the map. This creates what psychologists refer to as a *mask* between experiment tasks, removing any link between the previous image and the next. Unfortunately this has a negative impact on the application's user experience. Since the changes in activity are quite small from one hour to the next, this masking effect overpowers the visual effect of the minute changes necessary to understand the urban pulse. In order to circumvent this issue, we pre-render

tiles on the client and set the opacity value of zero. Initially an array of four hours of data are loaded on to the map with only the first hour being visible. As the user clicks the button to advance through time, the appropriate tiles are made visible while the previous hour's tiles are removed from the map. The process of changing the visibility of layer is computationally minimal compared to the task of adding the tiles. When the number of map tile layers preceding the currently visible layer reaches two, the next four hours of tiles are invisibly added to the map. This process ensures that a seamless flow of visual information is presented to the user.

5.4 Default Behavior vs. Real-time Bursts

This section presents two contrasting views of the POI-driven city pulse. First, the default behavior view aims at representing the constant and steady changes in activities conducted in the city. Temporally, the city goes through changes in activity dominance. This implies that specific activities, and the POI (types) at which these activities take place vary in intensity through out the day/week. This leads one to describe *Coffee Shops* as being mostly visited during the morning

while *Nightclubs* are most active at night. We call this the Foursquare population's *Default Behavior* (towards POI).¹³

While humans are often described as creatures of habit (and the temporal bands support this), on an individual level, our behavior is often quite spontaneous and unpredictable. Analysis and visualization of these phenomena cannot be explored by focusing at the POI ecosystem as a whole, but rather at a large scale or neighborhood level. It is virtually impossible to view an overview of a city and attempt to understand the individual activities and behaviors of every inhabitant. Existing research by [Cranshaw et al. \(2012\)](#), explored this phenomena of UGC-driven neighborhoods previously, but in a very different way. The authors show that a city can be split into subregions based on the social media contents generated by its residents. Our work takes a very different approach focusing at real-time information presented in subregions rather than individual neighborhoods. For this reason, we developed the *Social Burst View*.

5.4.1 Default Behavior

From a systems architecture standpoint, the default behavior is accessible by zooming and panning through all map zoom levels. Visual exploration of default temporal behavior and spatial patterns is also encouraged by panning through

¹³We are well aware of the fact that Foursquare is a biased data sources and thus our POI Pulse is biased (but this is not the focus of this paper).

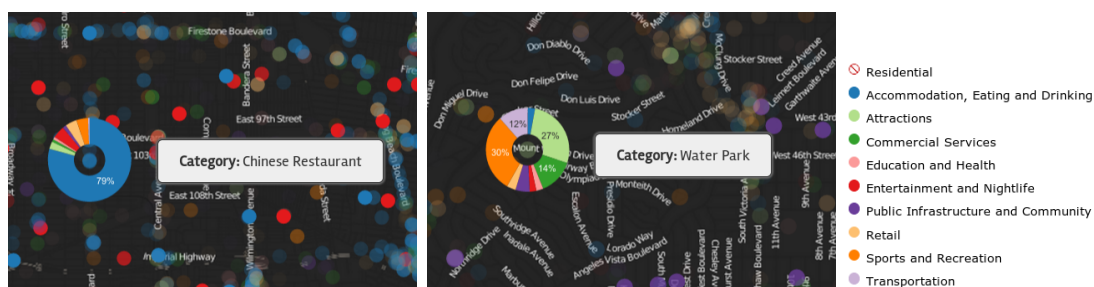
time (clicking the *Next Hour* button) or jumping to a selected time (clicking the clock button and selecting from a drop down list of hours and days). This shows how the POI-driven urban system changes over time. At the initial map scale, a single color value is used to represent all user-contributed POI in the Greater Los Angeles area. Advancing through time while visualizing the POI in this way provides the user with a better understanding of the flow of the city as a whole. This view is essential to understanding which regions of the city are dynamic and the overall variability in activity level for the entire region. Increasing the map scale by one zoom-level, the user is presented with new upper-level classes. Again, panning through both space and time, the viewer gains a better understanding of the distinction between classes. As the opacity value of each POI marker changes, the user is made aware that the level of activity is changing both between and within classes. For example, the class of *Entertainment and Nightlife* is very prominent at 12am on Sunday while it is completely overshadowed by categories such as *Commercial Services* on Monday at 9am.

Zooming in further, the data format switches from image to vector tiles. While the color scheme, marker size, and opacity do not change between zoom levels, the capabilities of the vector data format allow for much greater user interaction. Hovering one's mouse over any POI between zoom levels 13 and 16 results in the Foursquare POI type name appearing beside the POI as well as a *donut-pie chart*

surrounding the marker. The donut-pie chart is a technique employed to visually explain the OS class probabilities determined for each POI type in section 2.4 thus going beyond binary classification.

The value of being able to interact with the map through mouse events, for example, is that one can visually explore the probability distribution of classes for each individual POI. The standard marker visualization forces each POI to be assigned a single color representing a single class, but in actuality, the POI may exhibit high probabilities in more than one class and the primary marker color could be ascertained by a very small margin. When the user hovers over a POI, the donut-pie chart is displayed, demonstrating the multi-class characteristics of the venue. Each portion of the donut represents a category that contributes to this venue, and the color of each portion reflects the class. The size of each portion is defined by the percentage of this contribution based on the learned SVM model.

Figure 5.7 shows mouse-over interaction with two different POI. The central marker in Figure 5.7a is styled blue indicating that the primary OS class for this POI is *Accommodation, Eating and Drinking*. The accompanying donut-pie chart clearly shows that the highest probable classification for this POI **type** is indeed *Accommodation, Eating and Drinking* followed by small fractions of *Entertainment and Nightlife* and so forth. Comparatively, Figure 5.7b shows the prominent class for *Water Park* to be *Sports and Recreation* which makes sense given that



(a) Chinese Restaurant

(b) Water Park

Figure 5.7: Donut-pie charts showing OS category probabilities for two different POI.

users of the geosocial application are likely to visit a water park to engage in physical activities and recreation. The second highest class, as shown by the donut-pie chart is *Attraction* and it is the second highest probability by only a few percentage points. It is this ability to explore the discrepancies between classes, to dig into the underlying data, that is the true power of such an application.

5.4.2 Real-time Burst Mode

Understanding the pulse of a city involves not only looking at the city as a whole, but exploring the individual subregions or neighborhoods. What are people talking about in this part of the city? What places are popular right now? These are questions that should be asked, not at a city-wides scale, but rather at a local scale where the contents can be understood. Recent work by Purves et al. (2011) explored this notion of describing place based on data contributed

to geosocial applications such as Flickr and Geograph. In addition, the *LIVE Singapore!* project (Kloeckl et al. 2012) allows individuals to access a range of real-time information from a variety of sources as well as contribute back to the system. While it does not include default temporal behavior as a foundation, it does offer real-time access to an assortment of city sensors.

Microblogging applications such as *Twitter* offer users the ability to geo-tag their content before publishing it. By accessing the streaming API,¹⁴ these tweets can be added to the map immediately after they are published, providing the user with (near) real-time information on what is happening in a certain region. Additionally, Foursquare provides current check-in counts for any venue in their dataset through their rate-limited API.¹⁵ This information is valuable in that it shows the true popularity of both Foursquare as a service, and the POI at which its users choose to check-in. Clicking the *Burst Mode* button immediately changes the temporal state of the map to the current hour of the day and week and begins to show real-time tweets and check-in counts based on the map view-port. Adding the real-time behavior on top of the default behavior is only possible at zoom levels that ensure that the necessary queries do not exceed the rate-limit of the used APIs.

¹⁴<https://dev.twitter.com/docs/streaming-apis>

¹⁵<https://developer.foursquare.com/docs>

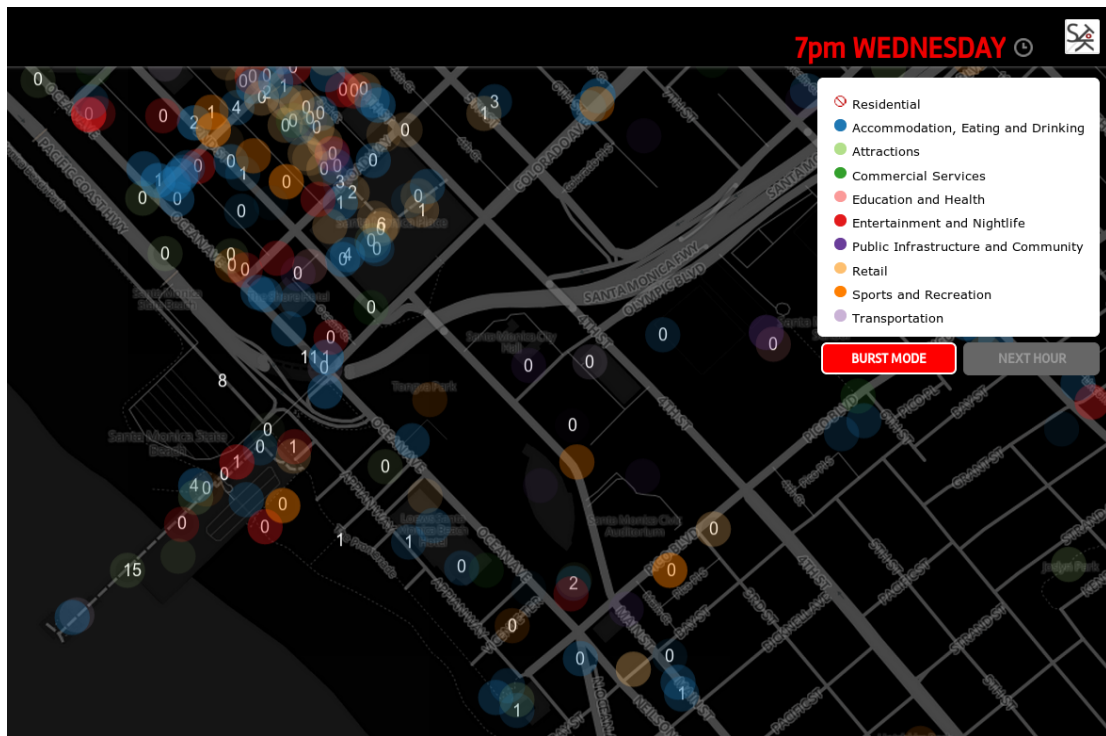


Figure 5.8: Real-time Check-in counts for Santa Monica at 7pm on a Wednesday

Real-time tweets

The Twitter streaming API offers users the ability to filter public streaming tweets by a specific geographic region. A listener service provides bounding box coordinates of the study area and all tweets geotagged within the region are inserted into a PostGIS database table. Individual tweets older than one minute are purged from the database complying with Twitter's terms of service.¹⁶ Though Twitter claims to restrict all tweets accessed through the filter streaming API to 1% of the real-time data, the influence of adding a geographic filter is not fully

¹⁶<https://dev.twitter.com/terms/api-terms>

known. The average rate of tweets over a 24 hour period filtered to within the Greater Los Angeles area is approximately 113 per minute.

On the client/browser side, an asynchronous JavaScript (AJAX) request is made every 1000ms to a server side handler. The JavaScript request provides the view-port extent of the browser in geographic coordinates in order to restrict the returned tweets to only those within the user's view extent. In addition, only those tweets published within the last 2 minutes are requested. Upon return, the tweets are added to the map via a *D3* vector layer which produces an animation that mimics water droplets (Figure 5.9). The animation lasts for 1000ms while another request is made to the server.

Given the sheer number of tweets, it is not technically prudent nor cognitively reasonable to display tweets on a small scale map. Recognizing this, users are given the option to view live tweets only within specific regions. The factor that determines the size and zoom scale of these regions is the number of POI within the view-port. A threshold of 1000 POI inside a view-port is the value at which users are given the option to view live tweets. Statistically, POI density is a good indication of neighborhood popularity, as the original POI were generated through crowd-sourcing means. It is important to note that this threshold is set independent of zoom level. As Figure 5.9 indicates, in some cases (Santa Monica Pier for example) the map scale will need to be quite large in order to fit less than

1000 POI in a view-port. Alternatively, parts of South-East Los Angeles reveal a lower POI density and therefore do not necessitate as large a map scale in order to visualize tweets.

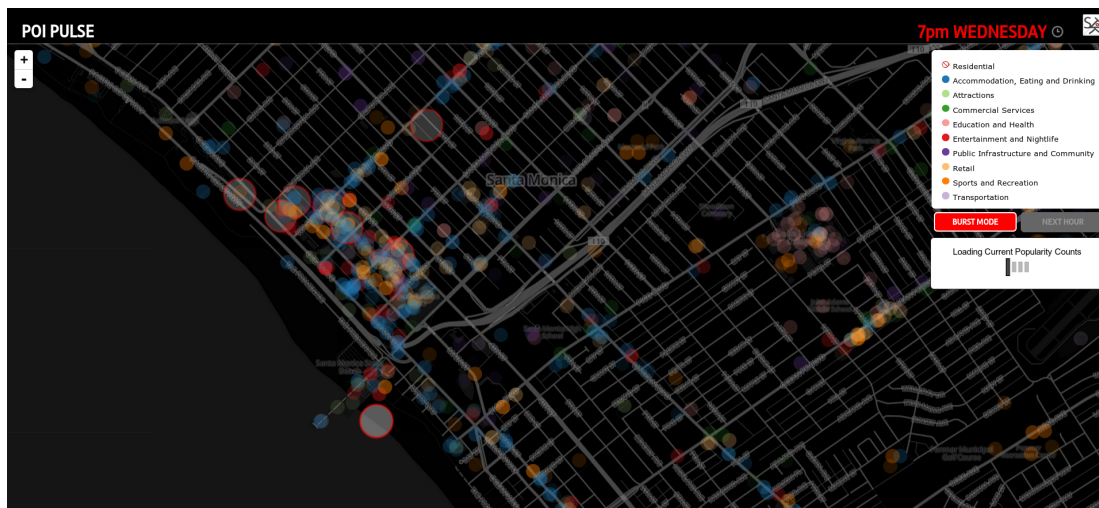


Figure 5.9: Tweets in Santa Monica (red circles with semi-transparent white fill)

Real-time POI popularity

While live geotagged-tweets offer insight into region-specific themes, the real-time geosocial popularity of POI in the region can also be determined. The Foursquare API permits requests to specific venues in order to determine the number of Foursquare users currently checked-in. Given API rate limits¹⁷, a 500 POI view-port restriction ensures that any request made to a region returns only valid responses. From an system architecture perspective, when a user clicks the

¹⁷API limitations require that a user request data through OAuth. Requests are limited to 500 per hour.

Burst Mode button, an AJAX request is made to the API which includes the geographic extent of the map view-port. The response from this request is then returned to the browser and check-in count values are added to the map for 500 POI, overlaid on top of the existing POI markers.

5.5 Conclusions and Future Work

Inspired by Foursquare’s city pulse videos, 5 major shortcomings were identified that must be addressed to make the POI pulse useful from a scientific perspective and to contribute to the vision of information observatories for urban systems. Based on those shortcomings, we derived 4 theoretical and technical research questions that have to be successfully addressed to implement an improved urban pulse. In this work we addressed those questions by a combination of data-driven and theory-informed techniques to arrive at a semantics signatures-based POI taxonomy. We investigated how to seamlessly switch between multi-buffered image and vector tiles to implement a responsive Web portal that can handle about 165,000 POI, thus pushing the envelope of state of the art Web cartography. We studied the *tipping point* between those cached image and vector tiles, and finally proposed a method to seamlessly switch between a default mode of human behav-

ior derived from empirical probabilities and streaming real-time geosocial data.

We implemented the POI Pulse system as showcase for our proposed solution.

In the future, we would like to add more services to the real-time burst mode. We are especially interested in a combination of platial, e.g., current check-ins, and spatial, e.g., Instagram pictures, data. As a proof-of-concept, we have discussed how to classify the human-generated content POI types to an administrative POI classification schema based on semantic signatures. However, we have only done so for level 1 and plan to add the second OS level in the future. Additionally, integration of POI and attribute data from alternative sources ([McKenzie et al. 2014](#)) would increase the variety and robustness of the proposed classification model. We also plan to add more bands, e.g., based on the place personalities proposed by [Tanasescu et al. \(2013\)](#). Next steps will also involve other methods of delineating class types, e.g., using a combination of color ramps to visually represent combinations of class probabilities and uncertainty.

Chapter 6

How Where Is When? On the Regional Variability and Resolution of Geosocial Temporal Signatures Mined from Point Of Interest Check-ins

The previous chapters have shown that the temporal behavior of individuals is crucial to defining the types of places that they visit. Until this point this work has taken an *aspatial* approach to constructing temporal signatures by aggregating check-ins from all across the United States. In this chapter, the regional variability and resolution of temporal signatures are the subject of investigation. Temporal check-in patterns from three cities across the U.S. as well as one city in China are compared. The results show that there is regional variability at it pertains to some place types and a lack of regional variability in others. This work takes

Chapter 6. How Where Is When? On the Regional Variability and Resolution of Geosocial Temporal Signatures Mined from Point Of Interest Check-ins

the next logical step in discussing the role of time in determining place types by investigating the placial indicativeness of temporal signatures.

Peer Reviewed Publication	
Title	How Where Is When? On the Regional Variability and Resolution of Geosocial Temporal Signatures Mined from Point Of Interest Check-ins
Authors	Grant McKenzie ¹ , Krzysztof Janowicz ¹ , Song Gao ¹ , Li Gong ²
Institutions	¹ Department of Geography, The University of California, Santa Barbara, ² Institute of RS and GIS, Peking University
Venue	Computers, Environment and Urban Systems
Editors	TBA
Publisher	Elsevier
Pages	Under Review
Submit Date	March 10, 2015
Accepted Date	Under Review
Publication Date	Under Review
Copyright	Under Review

Abstract

The temporal characteristics of human behavior towards Points of Interest (POI) differ significantly with place *type*. Intuitively, we are more likely to visit a restaurant during typical lunch and dinner times than at midnight. Aggregating geosocial check-ins of millions of users to the place type level leads to powerful *temporal signatures*. In previous work these signatures have been used to estimate the place being visited based purely on the check-in time, to label uncategorized places based on their individual signature's similarity to a type signature, and to mine POI categories and their hierarchical structure from the bottom-up. However, not all hours of the day and days of the week are equally *indicative* of the place type, i.e., the information gain between temporal bands that jointly form a place type signature differs. To give a concrete example, places can be more easily categorized into weekend and weekday places than into Monday and Tuesday places. Nonetheless, the *regional variability* of temporal signatures has not been studied so far. Intuitively, one would assume that certain types of places are more prone to regional differences with respect to the temporal check-in behavior than others. This variability will impact the predictive power of the signatures and reduce the number of POI types that can be distinguished. In this work, we address the regional variability hypothesis by trying to prove that all place types

are born equal with respect to their temporal signatures, i.e., temporal check-in behavior is *aspatial*. We reject this hypothesis by comparing the inter-signature similarity of 321 place types in three major cities in the USA (Los Angeles, New York, and Chicago). Next, we identify a common core of least varying place types and compare it against signatures extracted from the city of Shanghai, China for cross-culture comparison. Finally, we discuss the impact of our findings on POI categorization and the reliability of temporal signatures for check-in behavior in general.

6.1 Introduction

Points of Interest (POI)¹ are inextricably linked to modern (mobile) search, recommender systems, location-based social networks, transportation studies, navigation and tourism systems, urban planning, predictive geo-analytics such as crime forecasting, and so forth. In terms of their computational representation, POI can be described and categorized in many different ways. Typical approaches are either based on features or functionality. The former describe POI based on attributes/properties such as price range, Wi-Fi availability, wheelchair

¹We use the term *Point of Interest* here to keep in line with related work in research and industry and because these places are typically represented by point geometries. On the long term and due to the increase in richer geometric representations, *Place of Interest* seems to be the more appropriate name.

access, ambience, noise level, room size, customer satisfaction, and so forth. Leaving pre-defined types such as *Restaurant*, *Hotel*, or *National park*, aside, POI can be grouped into ad-hoc categories (Barsalou 1983) based on their common features such as “expensive places” or “attractions that offer wheelchair access.” A functionality-centric view describes and categorizes POI based on what they *afford*, e.g., dining, travel, trade, or shelter (Jordan et al. 1998, Winter & Freksa 2014). While both approaches can be combined to account for their distinct strengths and weaknesses, they are typically realized in a schema-first manner in which features or functionalities are defined top-down and then populated with data (Glushko 2014). An example of such a schema is shown in Figure 6.1 which depicts properties defined for *museum* as well as the higher-level types from which these properties were inherited.

Alternatively, and assuming that meaning emerges from social structure (Gärdenfors 1993), POI types can be described and categorized by aggregating how people behave towards places, e.g., when they visit them, what they say/write about them, and so forth. In addition to top-down schemata, such an approach reveals meaningful patterns suitable for a bottom-up, observations-first characterization of POI (types). To give a few concrete examples, certain types of places are visited mostly during the weekends, while others are visited primarily during the workweek. Similarly, some types have their visitation peaks during the evenings

Thing > Place > CivicStructure > Museum

A museum.

Property	Expected Type	Description
Properties from CivicStructure		
openingHours	Duration	The opening hours for a business. Opening hours can be specified as a weekly time range, starting with days, then times per day. Multiple days can be listed with commas ',' separating each day. Day or time ranges are specified using a hyphen '-'. - Days are specified using the following two-letter combinations: Mo, Tu, We, Th, Fr, Sa, Su. - Times are specified using 24:00 time. For example, 3pm is specified as 15:00. - Here is an example: <code><time itemprop="openingHours" datetime="Tu,Th 16:00-20:00">Tuesdays and Thursdays 4-8pm</time></code> . - If a business is open 7 days a week, then it can be specified as <code><time itemprop="openingHours" datetime="Mo-Su">Monday through Sunday, all day</time></code> .
Properties from Place		
address	PostalAddress	Physical address of the item.
aggregateRating	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.
containedIn	Place	The basic containment relation between places.
event	Event	Upcoming or past event associated with this place, organization, or action. Supersedes events .
faxNumber	Text	The fax number.
geo	GeoCoordinates or GeoShape	The geo coordinates of the place.

Figure 6.1: A fragment of the *Museum* type from Schema.org.

while others peak during typical business hours from 9am-5pm. Even the lack of such distinct peaks is indicative (e.g. of major airports). Textual descriptions and other sources of observations can be used accordingly. For instance, mining latent topics from social media such as textual user reviews of places from Los Angeles reveals very characteristic Spanish-language topics (McKenzie et al. In Press).

As an analogy to spectral signatures and bands in remote sensing, we have proposed *semantic signatures* that support the categorization of POI based on a multitude of spatial, temporal, and thematic bands (Janowicz 2012a). Simply put, in the domain of remote sensing, geographic entities on the surface of the Earth are classified via their unique reflection and absorption patterns in different wavelengths of electromagnetic energy called *bands* (Schowengerdt 2006). In some cases a particular band is sufficiently indicative to distinguish entity types (e.g., paved concrete from bare red brick), while in other cases a combination of multiple bands is required to form a unique spectral signature (e.g., deciduous and conifer trees cannot be distinguished via the visible light band alone).

Temporal semantic signatures and bands are of particular interest as they are relatively easy to mine and at the same time are strongly indicative for a variety of POI types (Shaw et al. 2013, Ye, Janowicz, Mülligann & Lee 2011). Consequently, they have been successfully used for the labeling of uncategorized places, for data cleansing and deduplication, for the construction of bottom-up POI hierarchies, for geolocation tasks such as estimating which place a user visited based on GPS fixes, and further tasks that benefit from this kind of social sensing. Recognizing the role of time has also lead to new fields of study such as time-aware POI recommendation (Yuan et al. 2013). Some POI types require additional (non-

temporal) bands for their more fine-grained classifications (McKenzie et al. In Press). However, we will exclusively focus on temporal signatures in this work.

Interestingly, not all hours of the day and days of the week are equally indicative for the classification of POI types, i.e., the information gain of temporal bands differs. Intuitively, places can be more easily categorized into evening and morning place types (e.g., *bars* versus *bakeries*) than into early morning and late afternoon places. To further exploit the analogy to spectral signatures, it is interesting to note that the *resolution* of temporal bands is characterized and bounded by human behavior. While hourly, daily, and seasonal bands have predictive power, second or minute-based bands do not (at least not for POI). This leads to the question of whether temporal signatures also have a *placial resolution*. Note that we use the term *placial* (or *regional*) instead of *spatial* here as the variation is expected to be non-linear². For example, San Diego, CA and Tijuana, Mexico are neighboring cities, yet we expect them to vary more with regards to the temporal signatures than San Diego, CA and San Francisco, CA which are over 700km apart. Conversely, *aspatial* implies *placial* (regional) invariance.

Clearly, as temporal signatures are mined from human behavior, certain POI types will be affected by cultural differences. For instance, the peak dinner time for restaurants in Italy is around 8pm while it is approximately 6pm in the United

²In the literature, ‘*placial*’ and ‘*platial*’ are used in parallel, we prefer *placial* here as it more readily points to the term’s origin, namely ‘place-based’.

States. We may even expect differences between the West and East Coasts of the U.S. In contrast, meaningful differences between the neighboring cities of New York, NY and Newark, NJ are less likely. Understanding such regional variations, their resolution, and magnitude, is important as they will effect the indicativeness of the signatures and thus their contribution to the aforementioned tasks; cf. (Gao et al. 2013a). In other words and referring back to the wordplay in the title, we will ask how much the *where*, i.e., regional-effects, impacts the *when*, i.e., the time people tend to visit certain types of venues. We will put this placial resolution research question to the test by hypothesizing that all place types are born equal with respect to their temporal signatures, i.e., that the temporal check-in behavior is *aspatial*.

The remainder of the paper is structured as follows. Section 6.2 outlines our research contributions. Section 6.3 introduces the data and the temporal signatures mined from these data. Next, in Section 6.4, methods, results and discussion on placial variation are presented. Section 6.5 discusses a small subset of the results in further detail, while Section 6.6 compares these results to another dataset from Shanghai, China. We discuss related work in Section 6.7 and finish with a discussion of the overall results and the conclusions in Section 6.8.

6.2 Research Contribution

The *regional variability hypothesis* can be illustrated using the following intuitive example. Given a user location derived from a positioning fix of a mobile device and a set of Points Of Interest in the vicinity of this fix; can we match the user's *spatial* location (lat/long coordinate) to a *placial* location (venue)? In other words, can we estimate which place a user visited, e.g., the Hollywood Palladium, based on the spatial location, e.g., the GPS fix 34.0981,-118.3249. Intuitively, the probability of *checking-in* at a particular place is inversely proportional to the distance between the spatial footprint of the POI and the user's location fix. As argued previously, check-in times can be aggregated to *type-indicative* temporal signatures. Now, given the example above, if the GPS fix was recorded at 8am, the user is more likely to be at the nearby Waffle eatery than the spatially closer Hollywood Palladium since the check-in probability for a concert venue is negligibly low in the morning. In contrast, the same fix recorded on Friday at 7pm most likely indicates a visit to the Palladium.

In fact, performing such an experiment with real data from over 2,800 check-ins in Los Angeles, CA shows that incorporating temporal signatures (aggregated from multiple cities) improves the geolocation estimation as measured by the Mean Reciprocal Rank (MRR) from 0.359 to 0.453, i.e., by about **26%**. Simplifying,

the check-in probability for a given place depends on the distance of its spatial footprint to the user's location as well as the temporal check-in likelihood at this *type* of place. Now, if the temporal signatures for POI types would be *aspatial*, i.e., there would be no regional variability, then the geolocation estimation quality would not differ based on the origin of those signatures. Temporal signatures derives from New York or Chicago check-in data would lead to the same increase in MRR over the distance-only baseline as signatures derived from Los Angeles (or signatures aggregated over multiple cities). However, performing such an experiment shows that using New York signatures for the geolocation estimation in Los Angeles leads to an MRR of **0.425**. This is lower than the performance of the aggregated signatures (**0.453**) but higher than the distance-only method commonly used to date (**0.359**). Consequently, while the New York signatures still outperform the baseline, there must be a *regional* variation in the check-in behavior. Alternatively, we can hypothesize that the signature differences are explainable by *random* variations.

This raises several interesting research questions; three will be explored in the following sections:

RQ1 Are POI types regionally invariant and the observed differences described above due to random fluctuations? We will try to reject this null hypothesis using the circular Watson's Two-Sample test.

RQ2 Are regional signature variations equally strong across all POI types, i.e., are there types that are affected more or less by such variations? Furthermore, given a POI type hierarchy, do certain supertypes form around more or less varying subtypes? We will address these questions by comparing the inter-signature similarity of POI types from three major cities in the USA (Los Angeles, New York, and Chicago). To ensure that these similarities are not merely artifacts of the used measure, we will use the Gini Coefficient, Jensen-Shannon Divergence, and Earth Mover’s Distance, and study the concordance of the resulting similarities by computing Kendall’s W.

RQ3 Given a common core of *least* varying POI types determined by their signatures from major US cities, how do these signatures hold up when compared against data from a very different cultural region, e.g., against signatures extracted for Shanghai, China? To approach this research question, we will select POI types that can be *aligned* between the U.S. and Chinese POI schemata and then divide them into two groups, those that vary clearly within the U.S. and those that do not. Next, we will use Earth Mover’s Distance to test whether these groups remain *stable* when using the Chinese signatures, i.e., whether POI types in the regionally invariant group remain in this group and vice versa.

6.3 Raw Data and Temporal Signatures

Check-in information was accessed hourly via the public *Foursquare* API to collect a total of 3,640,893 check-ins to 938,031 venues from 421 POI types across three regions: Los Angeles (LA), New York City (NY), Chicago (CHI), and New Orleans (NOLA).³ The Foursquare POI type schema groups these 421 POI types in to 9 top-level classes. To gain a better understanding of the data, Figure 6.2 shows the percentage of user check-ins divide by those 9 classes and split by region. *Travel & Transport* is by far the most prominent POI class followed by *Arts & Entertainment* which is more pronounced in Los Angeles than in either New York City or Chicago. In contrast, both New York City and Chicago show a higher percentage of check-ins at *Outdoors & Recreation* POI types.

For the purposes of this research, these check-in data were accessed during the fall/winter of 2013. The goal was to access check-ins to 60 venues in each city from each of the 421 POI types.⁴ Given the limited number of venues of some POI types in the selected cities (e.g., Belarusian Restaurant), this was not always possible. The hourly check-in data were aggregated by POI type, region, hour, and day of the week. Given 24 hours over 7 days, this resulted in 168 hourly *bands* used to construct a temporal signature normalized and aggregated

³Region boundaries are based on the *2010 Census Urban Areas* boundaries.

⁴<https://developer.foursquare.com/categorytree>

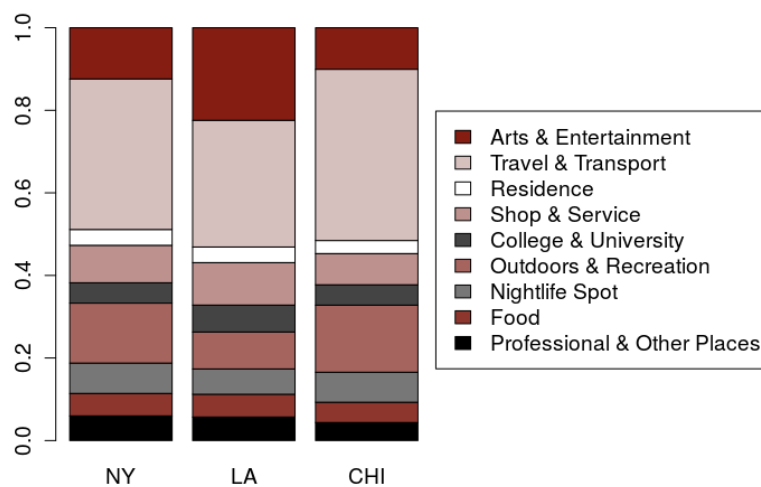


Figure 6.2: Stacked bar plots showing amount of check-ins to each parent class as a percentage of overall check-ins. Check-ins have been split in to regions.

to a single week. In order to ensure the robustness of the temporal signatures, any type whose venues appeared less than 30 times in a given region was removed from analysis. This reduced the number of POI types from 421 to 321. Additionally, the New Orleans dataset was dropped from analysis due to the limited availability of certain types which would have considerably restricted the categories available for comparison. The remaining 321 POI types in the three regions, Los Angeles, New York City and Chicago form the basis of the analysis to be discussed in the remained of the paper. Lastly, the signatures were cleaned by removing data errors and outliers. Note that due to the usage restrictions of the Foursquare API, no individual check-ins or venues were stored for this research but merely type-level aggregates.

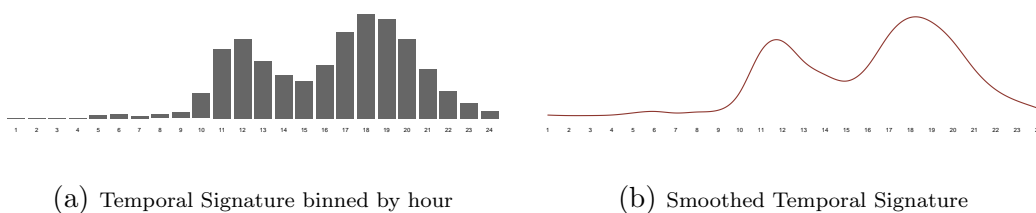


Figure 6.3: Check-in data represented as (a) a binned temporal signature by hour of the day and (b) a smoothed temporal signature. Both temporal signatures show averaged check-in behavior over 24 hours (a typical Tuesday) at a Mexican Restaurant.

Before using these signatures, it is important to understand their *temporal resolution* (Jensen & Cowen 1999), in this case, the smallest change in quantity (i.e., check-ins counts) that can be observed via a sensor (i.e., the check-in *Apps* and *API*). Reporting and using data below such a resolution may lead to erroneous results. For this reason, it is common practice to round data to their significant digits. While some Location-based Social Networking (LBSN) APIs return the check-in timestamp, others return check-in counts per venue and have to be scanned repeatedly at an interval that corresponds to the temporal resolution. More importantly, timestamps do not represent the time a user entered a place. For instance, a user would most likely enter a coffee shop, ordered an espresso, sit down, and *then* use his/her smartphone to check-in. Other systems, however, may check-in users automatically; see (Malmi et al. 2012) for the resulting differences between manual and automatic check-ins. Additionally, most LBSN platforms do not provide an option for *checking out* of a place and therefore, many ser-

vices will typically check their users out automatically after a certain time, e.g., 2 hours. Consequently, reporting temporal signatures on the level of minutes (even for large aggregates of data) or trying to draw conclusions from check-out times invites misunderstanding. This is of particular importance for the research at hand as we will compare signatures aggregated via Foursquare with those from *Jiepan*, a leading Chinese LBSN services (*Jiepan*) whose APIs return different temporal resolutions of data.

Consequently, we report the data at an hour-resolution as depicted in Figure 6.3a. If appropriate and necessary, the signatures can be smoothed via a kernel function; see Figure 6.3b.

6.4 Regional Variation

In this section, a number of methods for analyzing regional variations between POI types are presented. First, the question of whether or not types are *aspatial* is examined followed by an analysis of how much individual POI types vary regionally. Finally, POI hierarchies are examined in terms of their temporal signature homogeneity. We will define the terms and introduce the used measures in the corresponding subsection.

6.4.1 Significance of Placial Variations

Before we can explore the regional differences between particular POI type, we have to exclude the possibility that the temporal signature variations are merely a sampling artifact or produced through random fluctuations (**RQ1**). In order to do so, we start with the hypothesis that all types of POI are regionally equal, in other words they are placially invariant. Using *Watson's* Non-parametric Two Sample U^2 Test Of Homogeneity (Watson 1961, Zar 1976) we can test this hypothesis. The *Watson's* U^2 test (Equation 6.1) starts with the assumption that all samples (temporal signatures) are drawn from the same population (region). The variable N is the sum of the number of values in each sample (n_1, n_2) and d_k is the difference between the two cumulative signatures.

$$U^2 = \frac{n_1 n_2}{N^2} \left[\sum d_k^2 - \frac{(\sum d_k)^2}{N} \right] \quad (6.1)$$

The test also assumes that the temporal signatures are circular in nature (e.g, Monday is equally as close to Sunday in temporal distance as Sunday is to Saturday). Figure 6.4 visually depicts circular representations of temporal signatures for the POI types of *Theme Park* and *Drugstore*. Clearly, temporal signatures for *Theme Park* tend to vary stronger with place than those for *Drugstore*.

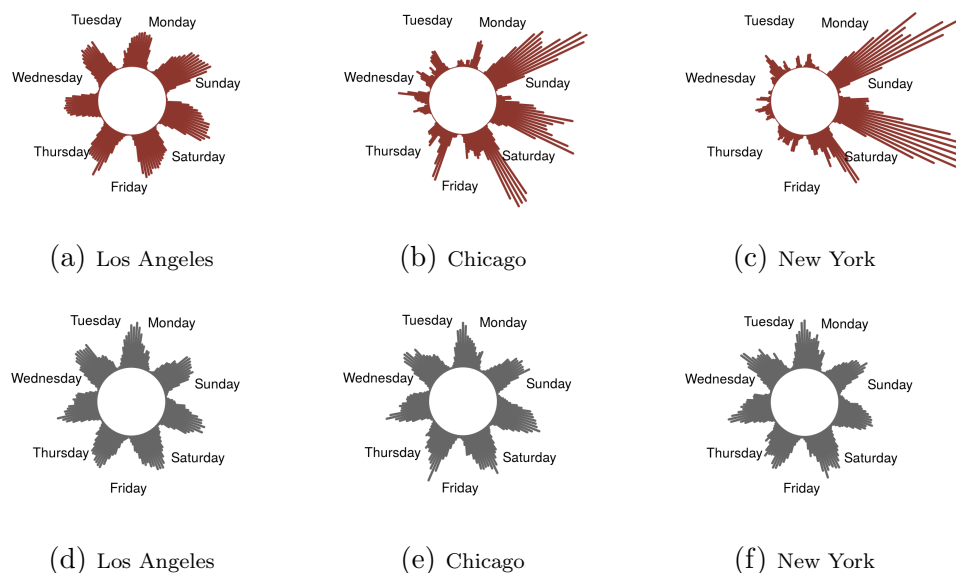


Figure 6.4: Circular histograms depicting temporal signatures for *Theme Park* (a,b,c) and *Drugstore* (d,e,f).

Altering the significance level⁵, the categorical circular distributions of 168 temporal bands are compared between each pair of regions (e.g., NY & LA). The results shown in Table 6.1 present the percentage of POI types that are significantly different between regions based on the provided significance level. For example, a significance value of 0.05 shows that in **52%** of the cases, the hypothesis is rejected for the pair of LA & NY meaning that 167 out of 321 POI types differ significantly between the two regions. Similarly, LA & CHI and NY & CHI pairs reject the hypothesis for **50%** and **48%** of POI types respectively. Provided this information, arguments can be made that (1) measurable and meaningful re-

⁵*alpha parameter* in the `watson.two.test` function in R.

gional variability does exist between POI types and (2) some types are regionally dependent while others are not.

	0.01	0.05	0.1
NY & CHI	33%	48%	57%
LA & CHI	37%	50%	59%
LA & NY	36%	52%	63%

Table 6.1: Percentage of POI types that are statistically different between regions as determined by the *Watson's* non-parametric two sample U^2 test of homogeneity. The results for three significance values (0.01, 0.05, 0.1) are reported.

These results confirm our intuition and reject the null hypothesis (**RQ1**). On the one hand, temporal signatures for POI types or check-in times in general have been successfully used in the literature (Ye, Janowicz, Mülligann & Lee 2011, Yuan et al. 2013, Shaw et al. 2013, Gao et al. 2013a, McKenzie et al. In Press, Yuan et al. 2014) because they are stable and generalizable over individual samples. On the other hand, even when applying the very conservative 0.01 alpha level, at least 106 POI types differ significantly between regions. Thus, understanding and quantifying these differences opens up new ways to substantially improve POI recommendation, classification, and so forth.

The question remains as to *which* POI types are placially variant and by what amount? This will be answered in the following subsections.

6.4.2 Variability Between Categories

In considering **RQ2** it is necessary to explore how temporal signatures of different POI types change based on region. In order to determine the amount by which some POI types are regionally dependent, we analyzed the variability using three dissimilarity measures.

Difference in Gini Coefficients

The *Gini coefficient* is a measure of the inequality of a given distribution. Originally intended to represent the income distribution of a country's residents (Gini 1912), a distribution of P is said to be equal (all values are the same) if $G(P)$ results in 0 and be completely unequal should $G(P)$ be 1. As shown in Equation 6.2, this coefficient provides a rough value used to describe any given distribution. In comparing two distributions, the *Gini coefficient* of one distribution can be subtracted from the other (which we refer to as the *difference in Gini Coefficients* or DGC) to give a broad indication of the (dis)similarity of two distributions.

$$G(P) = \frac{\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|}{2n^2\mu} \quad (6.2)$$

Table 6.2 lists the five most dissimilar POI types as well as the five most similar types based on the difference in Gini coefficients. The types are split based on region pairs. The value shown in parenthesis beside each type is the

difference in Gini coefficient value normalized by the most dissimilar type (Theme Park) and the most similar type (American Restaurant). Normalization allows for comparison between POI type as well as between dissimilarity measures (cf. Table 6.3 and Table 6.4).

NY & LA	NY & CHI	LA & CHI
Dissimilar POI Types		
Theme Park (0.844)	Recycling Facility (0.825)	Theme Park (1)
Real Estate Office (0.739)	Resort (0.797)	Resort (0.802)
E. European Restaurant (0.68)	Farm (0.703)	Baseball Stad. (0.74)
Recycling Facility (0.586)	Historic Site (0.702)	Donut Shop (0.711)
Farm (0.582)	Basketball Stad. (0.686)	Garden Cntr (0.704)
Similar POI Types		
Drugstore / Pharmacy (0.004)	Furniture / Home Store (0.003)	Monument / Landmark (0.005)
Gym (0.001)	Harbor / Marina (0.002)	Men's Store (0.003)
Community College (0.001)	Yoga Studio (0.002)	Gym (0.002)
Pet Store (0.001)	Laboratory (0.001)	Community College (0.001)
Art Museum (0.001)	Wings Joint (0.001)	American Restaurant (0.000)

Table 6.2: Top five and bottom five dissimilar POI types based on normalized *difference in Gini coefficient* and split by region pairs.

Jensen-Shannon Distance

While informative, the *difference in Gini coefficient* approach primarily focuses on the minima and maxima of a distribution. The *Jensen-Shannon Divergence (JSD)* is a method for measuring dissimilarity between two probability distributions (P, Q) (Lin 1991). In this case, comparison between distributions is done through a one-to-one bin approach. The *distance metric* is calculated by taking the square root of the value resulting from the divergence and is bounded between 0 (identical distributions) and 1 (complete dissimilarity). The *JSD* cal-

ulation is shown in Equation 6.3 where $M = \frac{1}{2}(P + Q)$ and KLD represents the *Kullback-Leibler Divergence* specified in Equation 6.4. While useful as a dissimilarity metric, *JSD's* one-to-one bin comparison does not take in to account neighboring bins.

$$JSD(P \parallel Q) = \frac{1}{2}KLD(P \parallel M) + \frac{1}{2}KLD(Q \parallel M) \quad (6.3)$$

$$KLD(P \parallel Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)} \quad (6.4)$$

Table 6.3 shows to the top five and bottom five dissimilar POI types split by region pair. As we saw with the *difference in Gini coefficient* approach (Table 6.2), the most dissimilar POI types are often *Theme Parks* or *Stadiums*. Interestingly, the most similar POI types are shown to be a variety of *Stores* (e.g., Grocery Store).

NY & LA	NY & CHI	CHI & LA
Dissimilar POI Types		
Football Stadium (1.000)	Theme Park (0.863)	Football Stadium (0.843)
Baseball Stadium (0.687)	Recycling Facility (0.677)	Theme Park (0.835)
Theme Park (0.603)	Food Truck (0.651)	Recycling Facility (0.733)
Basketball Stadium (0.594)	Funeral Home (0.627)	Skate Park (0.710)
Campground (0.584)	Basketball Stadium (0.586)	Food Truck (0.707)
Similar POI Types		
Electronics Store (0.021)	Grocery Store (0.000)	University (0.021)
Furniture / Home Store (0.039)	Residential Building (0.037)	Electronics Store (0.035)
Hospital (0.035)	Home (private) (0.023)	Hardware Store (0.030)
Grocery Store (0.032)	Department Store (0.021)	Drugstore / Pharmacy (0.024)
Department Store (0.031)	Mall (0.018)	Gym (0.022)

Table 6.3: Top five and bottom five dissimilar POI types based on normalized Jensen-Shannon Distance and split by region pairs.

Earth Mover's Distance

Given *JSD*'s reliance on a one-to-one bin comparison, the *Earth Mover's Distance* (*EMD*) is utilized as well. Originally introduced by the computer vision community (Rubner et al. 1998, 2000), *EMD* compares each bin in a distribution (*P*) to all bins in a second distribution (*Q*) assigning a cost value based on bin distance. Simply put, *EMD* is the minimum amount of work it takes to convert one distribution into the other, where $F_{i,j}$ is a flow matrix (amount of earth to move between bins) and $C_{i,j}$ is the cost matrix representing the cost of moving the flow. The total cost is then shown in Equation 6.5.

$$EMD(P, Q) = \sum_{i=1}^n \sum_{j=1}^n F_{i,j} C_{i,j} \quad (6.5)$$

As with both *DGC* and *JSD*, calculating the *EMD* across all types for all pairs of regions allows us to rank POI types by their regional similarity with high values indicating high dissimilarity. Table 6.4 lists the five most and five least dissimilar types split by region. The normalized *EMD* values are shown in parenthesis next to the type name. Similarities between the regional pairs are apparent in both the highly dissimilar and similar (shaded gray) groups with *Theme Parks* and *Stadiums* again, showing to be the most dissimilar POI type and *Stores* and *Residences* proving to be the most similar.

NY & LA	NY & CHI	LA & CHI
Dissimilar POI Types		
Theme Park (0.789)	Theme Park (0.710)	Theme Park (1.000)
Football Stad. (0.686)	Resort (0.614)	Resort (0.600)
Real Estate Office (0.471)	Basketball Stad. (0.549)	Baseball Stad. (0.575)
East Euro Restaurant (0.416)	Winery (0.506)	Garden Center (0.442)
Farm (0.402)	Recycling Facility (0.453)	Donut Shop (0.423)
Similar POI Types		
College Residence Hall (0.013)	Home (0.005)	Monument / Landmark (0.015)
Shoe Store (0.011)	Hardware Store (0.005)	University (0.009)
Military Base (0.010)	Doctor's Office 0.003	Drugstore/ Pharmacy (0.006)
Convenience Store (0.001)	Comm. College (0.002)	Home (0.004)
Drugstore/ Pharmacy (0.000)	Airport Gate (0.000)	Convenience Store (0.002)

Table 6.4: Top five and bottom five dissimilar POI types based on normalized Earth Mover's Distance and split by region pairs.

Summing up, with respect to **RQ2**, these three dissimilarity measures show that there are clear differences between POI types. Some, e.g., *Theme Park*, show a strong regional variability, while others, e.g., *Convenience Store*, do not.

6.4.3 Concordance Between Dissimilarity Measures

While these three statistical dissimilarity measures yield individual results for inter-signature comparison, the real value of these measures is shown in their agreement. Here *Kendall's coefficient of concordance* is employed (Kendall & Smith 1939). Each of the three regions is compared to each other region using *Earth Mover's Distance*, *Jensen-Shannon distance* and *difference in Gini coefficient*. These produce a single dissimilarity value from each region pair for each POI type. Kendall's W is then used to calculate the measure of concordance between each dissimilarity measure across all POI types.

Measures	NY & LA	LA & CHI	NY & CHI
EMD & JSD	0.80	0.82	0.78
EMD & GINI	0.91	0.91	0.88
GINI & JSD	0.74	0.73	0.74

Table 6.5: Kendall’s coefficients of concordance W for pairs of regions and combinations of dissimilarity measures ($p < 0.01$ in all cases).

A *Kendall’s* W value of 1 indicates complete concordance where a value of 0 represents no concordance at all. As shown in Table 6.5, all W values are greater than random with the values for *EMD* & *GINI* producing the highest coefficient of concordance followed closely by *EMD* & *JSD* and *GINI* & *JSD*. This indicates a high level of agreement between dissimilarity measures, thus excluding the possibility that the observed similarities are merely artifacts of choosing a specific measure. We will focus on EMD for the remaining analysis.

6.4.4 Hierarchy Homogeneity

Typically, POI types are not flat but form a hierarchy consisting of one or more root types followed by multiple type-levels. Figure 6.1 shows such a hierarchy from schema.org with *Thing* as the root type. The subsumption relation is transitive, i.e., as *Place* is a supertype of *CivicStructure* and *CivicStructure* is a supertype of *Museum*, *Museum* is also a subtype of *Place*. Such hierarchies are not only important means for knowledge engineering but also key for various information retrieval techniques such as query expansion.

The second part of **RQ2** poses the interesting question of whether supertypes, e.g., *Retail*, in a POI hierarchy are homogeneous with respect to the temporal signature variability of their subtypes, e.g., *Hardware store*. To address this question, we grouped the top 100 most and top 100 least varying subtypes and then compared their distribution with respect to the supertypes. Intuitively, homogeneous supertypes should mainly contain subtypes from one group but not from both.

Figure 6.5 depicts the results of our analysis for the supertypes provided by Foursquare. By necessity, hierarchies introduce some arbitrariness by highlighting certain perspectives and hiding others. The Foursquare POI hierarchy is an interesting case as its supertypes seem like mixed bags, e.g., grouping *Cemeteries* under the *Outdoors & Recreation* root type and even introducing a *Professional & Other Places* “catch-all” type. While some POI types such as *Nightlife Spot* and *Travel & Transport* are homogeneous, the majority do not show a clear trend. This confirms our intuition. In fact, this very problem has been addressed before, combining spatial, thematic, and temporal signatures to construct a more appropriate POI type hierarchy for Foursquare from the bottom-up (McKenzie et al. In Press). We can now use this hierarchy to compare it to the original Foursquare categorization. Intuitively, the bottom-up version should be more homogeneous, i.e., supertypes predominantly contain either similar or dissimilar subtypes (with

regards to their temporal signatures between U.S. cities). Figure 6.6 confirms this assumption, the *Accommodation*, *Eating & Drinking* and *Attractions* types being particularly clear examples. It is interesting to note that in both hierarchies the transportation-centric types contain mostly similar POI types, while the service types consist of subtypes too diverse to show a clear picture.

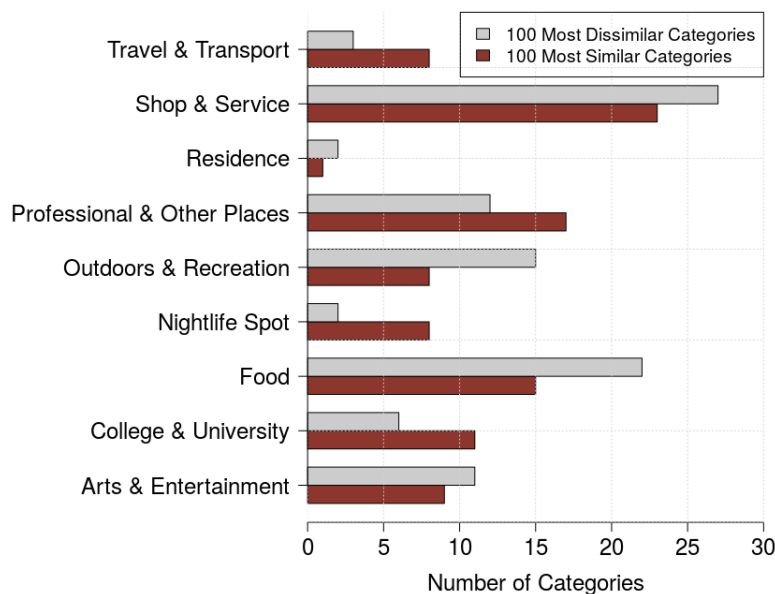


Figure 6.5: Original Foursquare POI hierarchy supertypes by prevalence of the 100 most similar subtypes and the 100 most dissimilar subtypes.

Summing up, to answer the second part of **RQ2**, POI hierarchies are not generally homogeneous with respect to the regional variability of the temporal signatures of their types. Nonetheless, some supertypes show clear patterns even across different POI hierarchies.

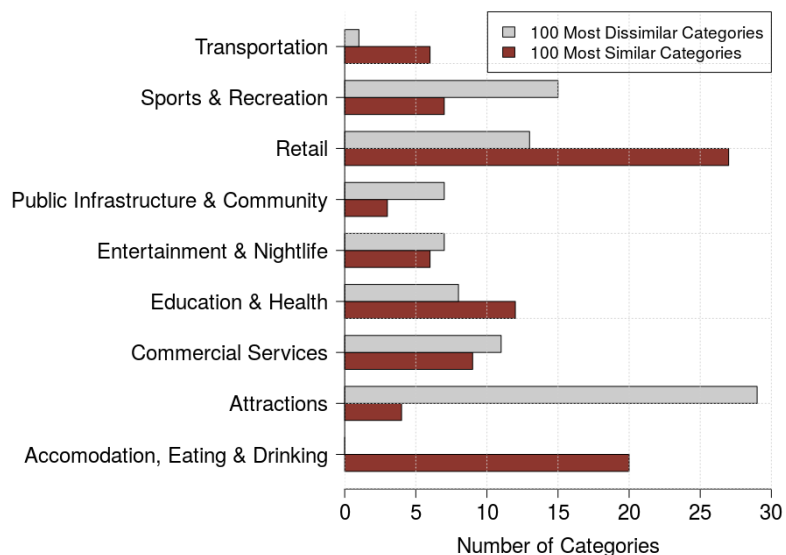


Figure 6.6: Bottom-up signature-bases POI hierarchy supertypes by prevalence of the 100 most similar subtypes and the 100 most dissimilar subtypes.

6.5 Cross-Cultural Comparison: Shanghai, CN

The next step in examining regional variation in POI types is to investigate how cultural differences influence placial variation in temporal signatures. With respect to research question **RQ3**, it is appropriate to examine the temporal signatures of POI in a city outside of the United States. In filling this role we chose to compare U.S. temporal signatures based on Foursquare data with temporal signatures constructed from *Jiepang*⁶ check-ins, which is one of the largest location-based social networking platforms in China.

⁶<http://jiepang.com>

6.5.1 Chinese Check-in Dataset

The *Jiebang* dataset on which this section is based contains more than 20 million location-based social check-in records from a one year period starting in September of 2011. All check-in data is from the Shanghai region of China and has been used to explore inter-urban mobility in previous work (Liu et al. 2014, Lian et al. 2014). Approximately 75,000 venues from 156 POI types grouped within 8 root-level types were extracted from user check-ins. Note that the predefined classification system of the Chinese check-in dataset is different from the type schema used in Foursquare. For example, the *(American) football stadium* type is popular in the United States while it does not exist in the *Jiebang* POI hierarchy. Furthermore, *Doctor's Office*, *Hospital* and *Medical Center* types from Foursquare are merged into a single Chinese LBSN POI type. Interestingly, and confirming our results from Section 4, the POI types that show clear regional differences within the U.S. are among those that are most difficult to align to the Chinese dataset, e.g., *Theme Park*, various types of sports facilities, and *Donut Shop*. In contrast, it was easier to find corresponding *Jiebang* POI types for the top *regionally invariant* types. As a sample comparison, 10 highly regionally invariant and 10 highly varying POI types were selected and manually matched between both datasets (see Table 6.6).

6.5.2 POI Type Similarity Comparisons

The 20 sample POI types were selected based on dissimilarity analysis within the three U.S. cities presented in the previous sections. Similar to the methodology discussed in Section 6.4.2, we applied the Earth Mover's Distance to calculate the dissimilarity of these POI types between the averaged temporal signatures for U.S. cities and the city of Shanghai.

Table 6.6 lists the 20 most and least dissimilar POI types along with the normalized EMD values for within the United States and between the United States and Shanghai. Please note that the *Mean EMD Within U.S.* is calculated by taking the mean of the EMDs reported from each regional pair while the *Shanghai vs. U.S. Mean EMD* is calculated as the EMD between the *regionally averaged* U.S. temporal signature and the Jieyang temporal signature. While the average EMD of temporal signatures between *the Shanghai vs. U.S. Mean* for all POI types is higher than that of *Within U.S.*, the magnitude difference between highly dissimilar POI types and highly similar POI types remains the same across cultures. In other words, POI types that are highly variable in the U.S. are also highly variable in Shanghai China (means of **0.62** & **0.68** respectively) while the most stable POI types remain stable across cultures (means of **0.07** and **0.23** respectively). The *Spearman* correlation coefficient of these sets of normalized EMD values is **0.64** (*Pearson* = 0.70) indicating above average similarities between the two.

POI Type	Shanghai vs. U.S. Mean n EMD	Mean n EMD Within U.S.
Dissimilar POI Types		
Theme Park	0.89	1.00
Farm	0.89	0.82
Historic Site	0.43	0.69
Zoo	0.42	0.59
Cemetery	1.00	0.58
Gaming Cafe	0.63	0.58
Pool Hall	0.25	0.54
Burger Joint	0.89	0.53
Gas Station/Garage	0.42	0.46
Public Art	0.98	0.36
Similar POI Types		
Toy/Game Store	0.00	0.15
Furniture/Home Store	0.45	0.13
College Library	0.05	0.13
Shoe Store	0.28	0.11
Mall	0.19	0.10
Grocery Store	0.09	0.04
Hotel	0.37	0.01
University	0.35	0.01
Home (private)	0.10	0.00
Drugstore / Pharmacy	0.23	0.00

Table 6.6: Ten highly dissimilar POI types and ten highly similar POI types selected from the U.S. Foursquare dataset. The Earth Mover’s Distance was calculated between each Foursquare POI type its Chinese *Jiepan* counterpart. The values were normalized between the most dissimilar and most similar POI type.

In response to **RQ3**, this section shows that the most regionally invariant types in the U.S. show reasonable stability when compared to Shanghai, China, but that highly variable types within the U.S. are also high variable in the Chinese

dataset. This is a very valuable insight as it indicates that some POI type may be represented by signatures with potentially global coverage.

6.6 Exemplary Investigation of Temporal Signature Differences

The analysis presented in the previous sections shows that POI types do in fact vary regionally with some showing significant changes between the regions of Los Angeles, New York City, Chicago, and Shanghai, and others displaying no significant difference in their temporal signatures. In this section, we discuss a few select examples of these POI types with the purpose of illustrating why regional variability exists for some types but not others.

The POI type that shows the highest level of dissimilarity across all pairs of regions and all dissimilarity measures is *Theme Park*. While this POI type may not immediately come to mind when thinking about regional differences, the reason is apparent when one examines the discretized temporal signatures (168 hourly bands of the week) shown in Figure 6.7. Check-in probabilities remain quite constant throughout the week for Los Angeles, while weekend peaks are much more pronounced for both New York and Chicago. These differences in temporal signatures can be explained through a better understanding of the regions themselves.

While a number of Theme parks exist in the Greater Los Angeles area, the predominant amusement park in the area is *Disneyland Resort*. In 2013, the park hosted approximately 16.2 million guests making it the third most visited park in the world that year (TEA 2014). Given the “holiday destination” nature of *Disneyland Resort*, it is not surprising that the temporal signature displays very little difference between weekend and weekdays. Moreover, a strong argument can be made for the impact of weather on theme park visits. As stated in Section 3, data collection took place through the seasons of Fall and Winter. Weather variability in Southern California is minimal relative to the seasonal variability experienced in both Chicago and New York. In actuality, many Theme parks in New York and Chicago close completely for the winter months (November - March) and a limited few remain open on the weekends for special events. Interestingly, check-in data from Shanghai shows similar weekend behavior but additionally we see a tendency toward a peak in the morning during the weekdays. This indicates the need for seasonal temporal signature bands, which we plan to address in the future.

Based on the variability analysis done in Section 4, *Football Stadium* are shown to be another POI type high in dissimilarity between regions. Since professional American football is traditionally played on Sundays, one might expect temporal signatures to be quite similar between regions in the United States. Upon further examination we find a number of different factors contributing to this dissimilarity

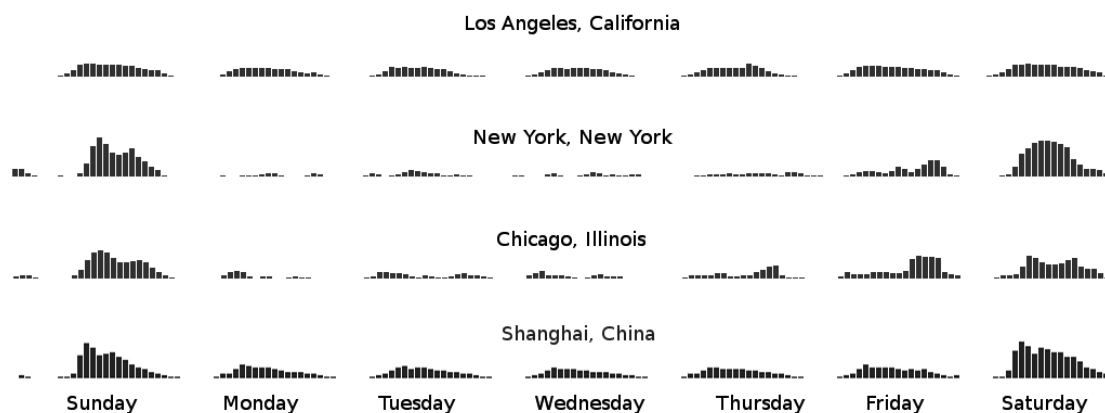


Figure 6.7: Temporal Signatures for the POI type *Theme Park* in Los Angeles, New York City and Chicago, United States and Shanghai, China.

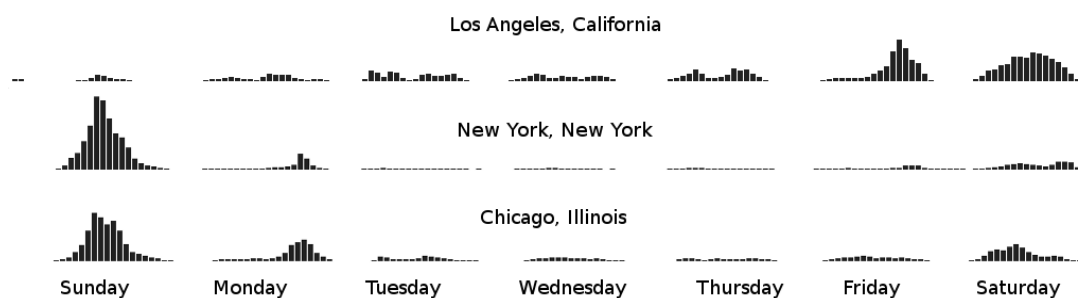


Figure 6.8: Temporal Signatures for the POI type *Football Stadium* in three cities in the United States. Note that data from Shanghai China is not shown here as no matching POI type was found.

ranking. First, while *professional* football is played on Sundays, College football is often played on Saturdays and High School football is typically played on Friday nights. It is important to know that Los Angeles does not have a professional football team which means that the peak one would expect on Sunday afternoon (which is seen in Chicago and New York City) is not found in the Los Angeles

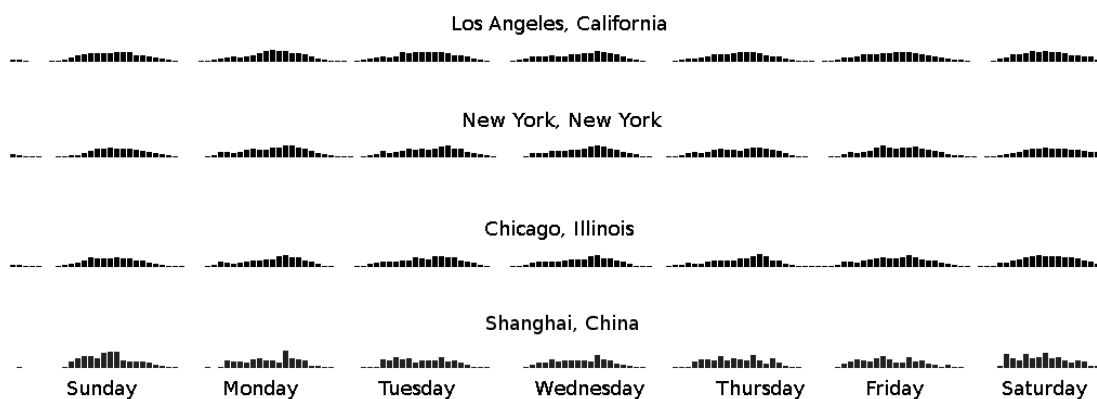


Figure 6.9: Temporal Signatures for the POI type *Drug Store / Pharmacy* in Los Angeles, New York City and Chicago, United States and Shanghai, China.

signature (Figure 6.8). Instead, we see the influence of both College and High School Football with peaks on Friday night and Saturday afternoon. Furthermore, football stadiums as with other types of stadiums, routinely host events other than just football matches. Major music concerts, trade fairs, and other sporting events often take place in large football stadiums which would also contribute to the regional difference in temporal signatures.

Lastly, we will look at the category of *Drug Store / Pharmacy* which presents the highest regional similarity across all POI types. From a conceptual perspective, one often thinks of a drug store or pharmacy as being an *atemporal* type of category. One would be hard-pressed to list the *typical* times of day that an individual would choose to visit a drug store as drug stores offer a wide range of products. Found on many street corners in the United States and China, drug

stores are often the “closest” place to pick up anything from sunscreen to birthday cards. Figure 6.9 shows the *atemporal* nature of drug stores with check-in values shown for most of the daylight hours and less check-ins late at night when many drug stores may be closed. Furthermore, Figure 6.9 shows the lack of regional variability in the temporal signatures between Los Angeles, New York City, Chicago and Shanghai.

In summary, while statistical methods applied to temporal signatures show that there are regional differences in POI types, a better understanding of the data behind these variations can be gained through a detailed examination of a subset of POI types.

6.7 Related Work

Previous studies have explored the role of LBSN data in analyzing human behaviors and urban dynamics. For instance, [Cheng et al. \(2011\)](#) found that users follow the “Lévy Flight” mobility pattern and adopt periodic behaviors in check-ins, which were bounded with their social ties as well as geographic and economic constraints. [Wu et al. \(2014\)](#) further analyzed the temporal transition probability of different activities (e.g., working, dining and entertainment) using social media check-in data. In work by [Noulas et al. \(2011\)](#), the authors reveal different

temporal rhythms in the top 10 most popular Foursquare categories (e.g., home, café, highway, and bar) between weekdays and weekends. The distinct temporal bands of POI types can be useful for data cleaning, place recommendation and decision making in LBSN (Ye, Janowicz, Mülligann & Lee 2011). From the urban informatics perspective, the POI data mined from user-generated content provides a fresh and updated view on the *city-in-use* versus the *city-in-plan*. Thus, it can help study neighborhood variations and monitor land-use changes (Quercia & Saez 2014). Cranshaw et al. (2012) spatially clustered POI as urban neighborhoods and studied how multiple factors shape urban dynamics. Recently, McKenzie et al. (In Press) introduced a multi-granular, semantic signatures-based approach for the interactive visualization of the city pulse using millions of POI data in the Greater Los Angeles area. A data-driven and theory-informed POI classification approach has also been introduced in this work focusing on the multi-dimensional (spatial, temporal and thematic) characteristics of POI types. Although there is a large volume of literature studying POI location recommendation based on users' historical check-in records and spatio-temporal patterns (Bao et al. 2012, Wang et al. 2013, Zhang & Chow 2013, Gao et al. 2013b, Yuan et al. 2013), to the best of our knowledge, no existing research has addressed the *placial perspective* and the role of regional variability in categorically defined temporal signatures.

6.8 Conclusions and Future Work

In this work we have discussed the regional variability and resolution of temporal signatures for Points of Interest. To study the variability, we assumed that POI type signatures are regionally invariant and hypothesized that the observed differences are merely random fluctuation. We rejected this hypothesis using Watson's Two-Sample test. Consequently, there are measurable and meaningful regional differences between POI types. This is an important finding as temporal signatures are a valuable social sensing methodology for various tasks including data cleansing, geolocation, POI recommendation, and categorization.

Next, we discussed the magnitude and the distribution of these differences within the U.S. by comparing major cities. To ensure that the comparison is not driven by the choice of similarity measure, we tested three measures and determined the concordance between them. The results confirm that the regional temporal signature variation is not homogeneous across POI types. A POI type that does not show regional variations when comparing New York to Los Angeles data, is also likely to show no substantial variation when comparing any of these cities to Chicago. Interestingly, the picture is more difficult for types that display strong regional variability. These types differ in unique ways, so to speak, independent of the compared cities. Finally, we compared U.S.-based signatures to

those from Shanghai, China to test whether types that show less variance would also remain stable when compared to data from a very different culture. Our first results indicate that this is the case.

Summing up, temporal signatures built from social media data, user check-ins in this case, are *not aspatial*. They vary to a degree where methods and applications would benefit from region-specific signatures. However, this does not mean that one would have to generate and store a multitude of local signatures. First, as the geolocation example in Section 2 demonstrates, *aggregated* signatures are very powerful and second, not all types vary to a degree that would justify the additional overhead. The suitable placial resolution for regionally varying signatures depends on the concrete application needs and the expected benefits. Defining country-wide signatures may be an appropriate resolution for some tasks but not for others. However, using the same signatures world-wide will only prove useful for a certain subset of relatively invariant types. Once again, this highlights the local nature of information and the role of space and place in studying Physical-Cyber-Social relations in general (Sheth et al. 2013). Our findings are important as today's research applies temporal POI and check-in data uniformly across space.

Future work in this area will involve expanding the dataset to include additional regions from major cities around the world. We will also explore the

difference between rural and urban settings as well as the influence of weather and seasonal effects on certain types of POI. Along the same lines, we focused on regional differences here while demographic differences may also be key drivers.

Finally, the work at hand is part of a long-term project ([Janowicz 2012a](#)) to publish an openly available library of semantic signatures with the hope that it will be equally as transformative as *spectral signature libraries* have been to the field of remote sensing. Signatures are difficult and time consuming to mine; the research community will benefit from having common access to well described and documented spatial, temporal, and thematic signatures for Points Of Interest and other features.

Chapter 7

Conclusions

In this dissertation, temporal human behavior is shown to be a valuable dimension for defining place types. In fact the temporal check-in patterns of individuals are more indicative than what has traditionally been seen as the most important dimension of place, namely *space*. The methods presented in previous chapters show that the activities that individuals conduct at places in their environment are highly temporally descriptive. In order to complete this research, data was accessed through voluntary activity surveys as well as user-contributed placial data collected through online geosocial networking platforms (e.g., Foursquare, Twitter, Yelp). The results of the research presented in this dissertation were evaluated through the construction of varied temporal probability models and rank statistical methods. The findings are chapter specific and are summarized within the respective chapters. This chapter will synthesize the findings and present the overall theoretical contribution of this work as well as the practical implications.

Limitations of this work are also presented along with directions for future areas of research.

7.1 Discussion

The increased availability of user-generated geo-content represents new opportunities for human behavior, activity and place-based geoinformatics research. Through the analysis of these rich sources of information we are able to better quantify the dimensionality of place. While *place* is in many ways still very much an abstract concept, analysis of data contributed by millions of visitors to these places offers new and unique insights into understanding and defining place. The research presented in this dissertation explores the concept of place, and more specifically *place types* through a multidimensional lens. Spatial, Thematic and Temporal components of place are three such dimensions and form the basis for this placial exploration. Of particular interest is the role that *time* plays in defining place types. Each Chapter of this work took a different approach to determining the value of the human behavior-driven temporal dimension, all stemming from the common thread that time is essential to understanding place.

7.1.1 Theoretical Contribution

Previous work on the topic of *Place*, from a geographic information science standpoint, has traditionally approached place from a *spatial* perspective (Agnew 2011, Goodchild et al. 2000, Kwan et al. 2003). This is not surprisingly, given that the study of Geography as a discipline is founded on the notion that *spatial* location matters. It must be realized though, that *space* is only a single dimension in the multifaceted definition of *place*. While other research has demonstrated the utility of the *thematic* dimension (Adams et al. 2015, Cranshaw & Yano 2010), research into the role of the *temporal* dimension on understanding places has been lacking. The research presented in this dissertation demonstrates the value of *time* in defining the place types. The statements below outline the theoretical contributions of this work, filling a gap that currently exists in place-based research and in doing so, advancing the field of geographic information science.

- While there is a correlation between online contributions and real-world activities, unique activities (those that exist outside one's daily routine) are more likely to be the subject of online social networking contributions. This indicates that inferences about the real world can be made via data collected from online actions (e.g., geosocial check-ins).

- People can be assessed as more similar or less similar to one another based on the places they choose to visit. While the place types that they visit are important, so is the time at which the places are visited.
- While multiple dimensions contribute to differentiating place types, the temporal component of visiting behavior is the most indicative dimension contributing to the delineation of place types.
- Combining the spatial components of place with human visiting patterns (temporal data) enhances the accuracy of positively identifying a place within a region.
- Temporal activity behavior varies with place type and region. In other words, human behavior as it pertains to frequenting places, differs between regions.

7.1.2 Practical Implications

One of the questions that is often asked of this research is *Why is a better understanding of place necessary?* and in that same vein *Why should we care about the temporal component of place?* In answering these questions, let us take a scenario from the domain of *Urban Planning*. In contributing to the design and plan of a city, it is important to not only understand the spatial dispersion

of a city's inhabitants, but also understand the ways in which these individuals interact with places in their environment. Since it has been shown that *time* is an important dimension and defining contributor to place, it follows that this temporal dimension would play a significant role in contributing to the development of a city. This temporal component of place allows urban planners to better understand the activities behavior of individuals as well as how places are connected via the temporal visiting behavior of people. In doing so, it allows for the city to be designed not only spatially, but also temporally, accounting for interaction between people through the appropriate application of land-use and zoning laws.

Aside from the theoretical contributions of this research, there are clear practical implications of this work in a wide variety of domains. The remainder of this section outlines many of the ways in which place and the temporal dimension of place are important.

Recommender Systems & Activity Prediction

The results of this research, as hinted at in Chapter 3, can greatly benefit the field of recommender systems. Current work in this area primarily makes use of top-down schemata for defining place types. For example, some systems make the suggestions that since you visited a *bar* previously, you would most likely like to visit a *bar* in the future. Few of these systems actually pull apart the dimension

of the individual bars that you visited to see what it was about those places that enticed you to attend them. Rather than taking an authoritative view on place types, this research investigates place types from the bottom up, mining the data of users that actually went to places and exploring the user behavior at these places (both thematically and temporally). This step contributes to both a better match between the instance of the place and the place type assigned to it, as well as a recommender system that makes use of these nuanced differences between places. This work clearly has implications on three types of recommender systems:

- **Local Place** recommender systems take data based on places that you have been previously and recommends new places for you to try. The results of this research would allow users of current recommender systems to adjust the variability parameters of the model to suggest more or less common types of places based on where they have gone previously.
- **Itinerary** recommender systems are responsible for designing an out-of-town trip itinerary based on one's interests. Such a recommender system would make use of this user-similarity work to find users in the region in which one plans to visit (proxies) that have similar activity interests as the focal user and then recommend places to her based on the places the proxy user visits.

- **Friend/Partner** recommender systems currently focus on the textual descriptions of the activities that people like to do as well as personal interest, etc. As has been shown throughout this and previous work (Scellato et al. 2011, Cho et al. 2011, Silva et al. 2014), the places that people choose to visit say a lot about them as individuals. Current friendship/partner recommendation services could use the work presented in the user-similarity model to propose possible friendships and relationships based on this information.

The flip side of *activity recommendation* is *activity prediction*. A good recommender system should be able to recommend a place to you that you would like to visit. Using this same approach, an activity prediction model would be able to predict the places you would likely go based on the data and place-type information.

Reverse Geocoding / Place Search

As shown in Chapter 4, the results of this research have significant implications for current state-of-the-art reverse geocoding services. Currently, top *place searching* services take a distance-only approach to determining one's placial location based on their spatial coordinates. This work shows that through the inclusion of temporal check-in data, specifically temporal signatures, the accuracy of existing place search approaches can be increased by over 26%. Ongoing work in this area

is currently focused on developing an enhanced reverse geocoding service that makes use of these findings.

Data-driven classification and hierarchy construction

Current efforts to categorize and organize place types into hierarchies are typically driven by top-down processes. Vocabularies such as *Schema.org*, *The Ordnance Survey* and even *Foursquare's* internal *venue tree* are all constructed by groups of people discussing how these types should be organized. The data-driven aspect of the research presented in this dissertation approaches the construction of place-type hierarchies from the bottom-up. Instead of organizing the data based on what we think are place-type delineations, this research proposes that we instead look at the actual data and investigate how humans interact with places. Categorization and schematization should be driven by real placial behavior. This is not to say that the two approaches are mutually exclusive, but rather that this work suggests that both options offer value and should be involved in the generation of place-type hierarchies.

Point of interest matching and conflation

Much of the appeal of so called *Big* data lies in the variety of data that is available to the average consumer. The recent increase in Point of Interest data

fits into this discussion through the emergence of numerous applications that provide POI data and platforms that assist in the user-generation of POI. The value of all of these platforms is that there is a wide range of data accessible for POI, everything from *user check-ins* to the *ambience* or even the presence of *Wi-Fi* are offered through these platforms. Access to rich content about a specific place through these various services is useful. The difficulty is that while many of these services offer attribute information for the same instance of a place, matching and conflating these places are difficult tasks. Initially one might think that given the geographic location of these POI, it would be possible to match them purely based on geographic coordinates. In actuality, these platform-specific POI have been shown to be over 62.8 meters apart on average (McKenzie et al. 2014). This is where the multidimensionality of place comes in. Both the thematic attributes of the space (e.g., textual reviews) as well as the temporal patterns (e.g., check-in behavior) can be included in the spatial dimension to offer a robust, multi-attribute method for matching points of interest.

7.2 Limitations

This research is not without its limitations. It is important to expose the issues and limitations that were present in this work as well as the steps that were taken to mitigate or minimize their impact.

7.2.1 Data

Bias

The majority of the data collected for this research was collected from online geosocial networking platforms. It has been widely recognized that use of these platforms is restricted in its demographics. The average *Foursquare* user is a single white young adult male with an annual income between \$25,000 and \$50,000. The reality of conducting research with this type of data is that any inferences can only be made for the population that the data represents. That being said, geosocial networking services such as Foursquare do represent over 50 million people worldwide. While restrictive in its demographic representation, analysis of the data does offer significant insights into the behavior of the individuals producing the data. For example, this data can be used in urban planning to see how people interact with the city and to determine roughly the burden being placed on the transportation system.

The bias towards certain types of POI is also recognized. The *social saliency* of a place should not be underestimated. Visits to place types such as *Hospitals* or *Jails* are most likely under-represented in check-in data as there is very little, if any, social capital to be gained from checking in to these places. Alternatively, a visit to a trendy *Nightclub* or *Ski Resort* is much more likely to be reflected in a check-in due to the increase in social value associated with “having it known” that you were there.

Much of the data on which this research is founded comes from a single data provider, namely Foursquare. While the argument can be made that in fact the data comes from over 50 million data providers, the fact remains that one service is responsible for validating, cleaning and other unknown practices on the data before serving it back to the public. The unfortunate reality is that at the time of conducting this research, Foursquare is the only provider to offer a large and robust enough dataset on which to conduct this research. As geospatial technology gains traction in social networking applications, more and more of this data will be made available, increasing the variety and richness of data for constructing temporal signatures.

Access to Data

Virtually all of the data used in conducting this research was ascertained via a public facing application program interface (API). Most large data providers offer restricted access to a limited amount of their data through an API. The purpose is to allow users of these API to gain a sample of the data on which to build third party applications which will in turn increase the revenue of the original data provider. While smart from a business standpoint, it severely limits one's ability to access data for research purposes. Given access to more data over a larger timespan, research into the influence of seasons and weather on check-ins would be feasible as well as large-scale studies on regional check-in variation. Unfortunately private *data silos* have become commonplace in today's user-generated data market and this has negatively impacted the ability to do large scale, data-driven behavioral research.

7.2.2 Methods

Dealing with Circular Data

One of the features that makes working with temporal data so unique is its circularity. In aggregating data to a certain temporal resolution, the day of the week or hour of the day for example, acknowledging this circularity is impor-

tant. It is imperative that the methods used to analyze circular data reflect this importance. For example, a similarity method that takes a linear approach to determining similarity would might make the erroneous assumption that Sunday and Saturday are at opposite ends of a linear scale, where in reality Sunday is as close to Saturday as Friday is to Saturday. This becomes important when using methods that assign a *transportation cost* for comparing two distributions (e.g. Earth Mover's Distance).

Real-world Activities vs. Online Posts

In Chapter 2 a preliminary study was shown that asked users to complete a diary of their daily activities. At the same time a *Facebook* application monitored their online posts. While this proved to be a good first step, asking individuals to track their own activities could prove to be problematic. A mobile application that automatically determines an individual's location as well as estimates their activities would be more suitable for this type of study.

7.2.3 Conceptual

A conceptual limitation of this work is found in the fact that for much of this research, place must be represented on a cartographic display. Since the concept of place is either unique for each individual or shared socially, justifying a boundary

on a map is difficult. In much of this research, places are represented as *Points of Interest* on a map. It should be noted that the deceptions of places as points is not a statement of their geometry, but rather a necessity of geocomputational research.

In addition to the visual and geometric depictions of place, the need to define the dimensionality of place is a limitation of placial research. The number of possible dimensions which places occupy is restricted to three in this work. This is not to say that there are not additional dimensions of place, but rather that for the purposes of this dissertation, working with primarily one dimension but discussion three was sufficient.

7.3 Future Research

The different paths by which future research on this topic may extend are plentiful. While not an exhaustive list, a number of these are outlined below.

7.3.1 Point of interest Matching, Conflation & Alignment

As stated in Section 7.1.2, additional effort can be made towards point of interest matching and conflation. This research has been done at a very interesting time for user-generated geo-content. Numerous data providers exist all offering

their own unique set of points of interest. Many of these places are referencing the same real-world instance of a place, but offer different details and attributes about the place. The difficulty is determining that the place in set A is the same as the place in set B. Additional work needs to be done in this area to not only match POI between providers, but also conflate these POI. The research community as a whole will benefit greatly from a robust set of POI in which multiple data platforms have contributed attribute information. Inclusion of information from multiple datasets will also reduce the bias inherent in information provided by a single application.

In addition to this, alignments need to be made between place type vocabularies. Currently Google, Microsoft and others use the *Schema.org* vocabulary for structuring the world while Facebook and Foursquare, for example, use their own internal vocabularies. In order to know that *Middle-Eastern European Restaurant* in one vocabulary is the same as *Persian Food Establishment* in another requires some form of alignment between the two vocabularies. The work presented in Chapter 5 presents a first step towards aligning the numerous vocabularies available today.

7.3.2 Geoprivacy

One of the top concerns of both producers and consumers of online user-generated content is the privacy of the data. As more and more of our lives are based online this concern for privacy is justified. The work presented in this dissertation touches on a number of the ways that an individual's publicly shared data can be used to expose their personal activity behavior. While this is not the purpose of this work, I can clearly lead in that direction. The unfortunate reality of many of the geosocial applications in the market today is that the benefits of the services that they offer are often gained at the cost of one's private location information (Duckham & Kulik 2005, Vicente et al. 2011, Kwan et al. 2004). For example, searching for a good restaurant nearby means sharing your current geolocation with services such as Yelp or Google, services that most likely already have additional personal information about you, such as your personal interests, communications, etc. While one concern in all of this is the amount of private location information we are willing to share with commercial companies, another is concern over how aware contributors are of what can be done with this data. An informed application user choosing to share this information is one thing, but an uninformed, naive user coerced into sharing this information is something else entirely. Much of these concerns have been discussed in a recent publication

(McKenzie & Janowicz 2014), though this is an area of user-generated geo-content that requires a lot more research.

7.3.3 Real-world Activities

In Chapter 2, the interaction between online social networking posts and the real-world activities they represent is discussed. This work presents a first step in a much needed exploration of the relationship between these two *worlds*. Specifically, the correlation between online *geosocial* content and real-world temporal activity space should be investigated in greater detail. As the use of geosocial applications become ubiquitous, the availability of data will continue to entice researchers into using these online sources to make inferences about real-world activity behavior. As this research states, it is important to make sure the relationship between online contributions and real-world activities is well understood in order to substantiate these inferences.

Bibliography

- Adams, B. & Janowicz, K. (2012), On the geo-indicativeness of non-georeferenced text., *in* 'ICWSM', pp. 375–378.
- Adams, B. & McKenzie, G. (2012), Frankenplace: An application for similarity-based place search., *in* 'ICWSM'.
- Adams, B. & McKenzie, G. (2013), Inferring thematic places from spatially referenced natural language descriptions, *in* 'Crowdsourcing Geographic Knowledge', Springer, pp. 201–221.
- Adams, B., McKenzie, G. & Gahegan, M. (2015), Frankenplace: Interactive thematic mapping for ad hoc exploratory search, *in* 'Proceedings of the 24th International World Wide Web Conference (WWW'15)'.
- Agnew, J. (2011), 'Space and place', *The SAGE handbook of geographical knowledge* pp. 316–330.
- Ahas, R., Aasa, A., Silm, S., Aunap, R., Kalle, H. & Mark, Ü. (2007), 'Mobile positioning in space–time behaviour studies: Social positioning method experiments in estonia', *Cartography and Geographic Information Science* **34**(4), 259–273.
- Ahas, R. & Mark, Ü. (2005), 'Location based services new challenges for planning and public administration?', *Futures* **37**(6), 547–561.
- Axhausen, K. W. (2008), 'Social networks, mobility biographies, and travel: survey challenges', *Environment and planning. B, Planning & design* **35**(6), 981.
- Backstrom, L., Sun, E. & Marlow, C. (2010), Find me if you can: improving geographical prediction with social and spatial proximity, *in* 'Proceedings of the 19th international conference on World wide web', ACM, pp. 61–70.
- Bao, J., Zheng, Y. & Mokbel, M. F. (2012), Location-based and preference-aware recommendation using sparse geo-social networking data, *in* 'Proceedings of

BIBLIOGRAPHY

- the 20th International Conference on Advances in Geographic Information Systems', ACM, pp. 199–208.
- Barsalou, L. W. (1983), 'Ad hoc categories', *Memory & cognition* **11**(3), 211–227.
- Biehl, L. L. & Stoner, E. (1985), 'Reflectance properties of soils', *Adv. Agron* **38**, 1–44.
- Biernacki, P. & Waldorf, D. (1981), 'Snowball sampling: Problems and techniques of chain referral sampling', *Sociological methods & research* **10**(2), 141–163.
- Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *Journal of Machine Learning Research* **3**, 993–1022.
- Bostock, M. & Davies, J. (2013), 'Code as cartography', *The Cartographic Journal* **50**(2), 129–135.
- Bostock, M., Ogievetsky, V. & Heer, J. (2011), 'D³ data-driven documents', *IEEE Transactions on Visualization and Computer Graphics* **17**(12), 2301–2309.
- Bowling, E. & Shortridge, A. (2010), A dynamic web-based data model for representing geographic points with uncertain locations, in 'Spatial Accuracy Symposium 2010', pp. 1–4.
- Brewington, B. E., Brown, B. G., Guggemos, J. A., Hawkins, D. & Stout, B. (2013), 'Augmentation of place ranking using 3d model activity in an area'. US Patent 8,533,187.
- Chang, J. & Sun, E. (2011), Location 3: How users share and respond to location-based data on social networking sites, in 'Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media', AAAI, pp. 74–80.
- Chapin, F. S. (1974), *Human activity patterns in the city: Things people do in time and in space*, Wiley New York.
- Chapin, F. S. & Logan, T. (1969), Patterns of time and space use, in 'Contributions to the Future Forum on New Resources in an Urban Age'.
- Cheng, H., Arefin, M. S., Chen, Z. & Morimoto, Y. (2013), 'Place recommendation based on users check-in history for location-based services', *International Journal of Networking and Computing* **3**(2), 228–243.
- Cheng, Z., Caverlee, J., Lee, K. & Sui, D. Z. (2011), 'Exploring millions of footprints in location sharing services.', *ICWSM 2011*, 81–88.

BIBLIOGRAPHY

- Cho, E., Myers, S. A. & Leskovec, J. (2011), Friendship and mobility: user movement in location-based social networks, *in* 'Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 1082–1090.
- Cooper, M., Dronsuth, R. W., Leitich, A. J., Lynk, J. C. N., Mikulski, J. J., Mitchell, J. F., Richardson, R. A. & Sangster, J. H. (1975), 'Radio telephone system'. US Patent 3,906,166.
- Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Machine learning* **20**(3), 273–297.
- Cranshaw, J., Schwartz, R., Hong, J. I. & Sadeh, N. M. (2012), The livelihoods project: Utilizing social media to understand the dynamics of a city., *in* 'The 6th International Conference on Weblogs and Social Media', AAAI.
- Cranshaw, J. & Yano, T. (2010), Seeing a home away from the home: Distilling proto-neighborhoods from incidental data with latent topic modeling, *in* 'CSSWC Workshop at NIPS', Vol. 10.
- Cresswell, T. (2013), *Place: A short introduction*, John Wiley & Sons.
- Duckham, M. & Kulik, L. (2005), A formal model of obfuscation and negotiation for location privacy, *in* 'Pervasive computing', Springer, pp. 152–170.
- Elliott, A. & Urry, J. (2010), *Mobile lives*, Routledge.
- Ellison, N. B., Steinfield, C. & Lampe, C. (2007), 'The benefits of facebook friends: social capital and college students use of online social network sites', *Journal of Computer-Mediated Communication* **12**(4), 1143–1168.
- Elwood, S., Goodchild, M. F. & Sui, D. (2013), Prospects for VGI research and the emerging fourth paradigm, *in* 'Crowdsourcing Geographic Knowledge', Springer, pp. 361–375.
- Evangelidou, C. (1988), *Aristotle's Categories and Porphyry*, Vol. 48, Brill.
- Facebook (2012), 'Facebook statistics', <http://newsroom.fb.com/content/default.aspx?NewsAreaId=22>. Accessed May 2, 2012.
- Fallah, N., Apostolopoulos, I., Bekris, K. & Folmer, E. (2013), 'Indoor human navigation systems: A survey', *Interacting with Computers* **25**(1), 21–33.

BIBLIOGRAPHY

- Ferrari, L., Rosi, A., Mamei, M. & Zambonelli, F. (2011), Extracting urban patterns from location-based social networks, *in* ‘Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks’, ACM, pp. 9–16.
- Flickr (2014), ‘Flickr developer documentation’, <http://www.flickr.com/services/api/flickr.places.findByLatLon.htm>.
- Foursquare (2014a), ‘Foursquare category hierarchy’.
URL: <https://developer.foursquare.com/categorytree>
- Foursquare (2014b), ‘What is the style guide for adding and editing places?’, <http://support.foursquare.com/hc/en-us/articles/201064960-What-is-the-style-guide-for-adding-and-editing-places>.
- Foursquare (2015a), ‘About foursquare’, <https://foursquare.com/about/>. Accessed: 01/10/2015.
- Foursquare (2015b), ‘The foursquare blog - december 2013’, <http://engineering.foursquare.com/2014/01/03/the-mathematics-of-gamification/>. Posted: 01/03/2014.
- Foursquare (2015c), ‘Foursquare’s new big data initiative is going to help it thrive, even as the check-in withers’, <http://blog.foursquare.com/post/70494343901/ending-the-year-on-a-great-note-and-with-a-huge>. Posted: 12/19/2013.
- Gao, H., Tang, J., Hu, X. & Liu, H. (2013a), Exploring temporal effects for location recommendation on location-based social networks, *in* ‘Proceedings of the 7th ACM conference on Recommender systems’, ACM, pp. 93–100.
- Gao, H., Tang, J., Hu, X. & Liu, H. (2013b), Exploring temporal effects for location recommendation on location-based social networks, *in* ‘Proceedings of the 7th ACM conference on Recommender systems’, ACM, pp. 93–100.
- Gärdenfors, P. (1993), ‘The emergence of meaning’, *Linguistics and Philosophy* **16**(3), 285–309.
- Gini, C. (1912), ‘Variabilità e mutabilità’, *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi* **1**.
- Glushko, R. J. (2014), *The Discipline of Organizing*, O’Reilly Media, Inc.

BIBLIOGRAPHY

- Golledge, R. G. (1997), *Spatial behavior: A geographic perspective*, Guilford Press.
- Goodchild, M. F., Anselin, L., Appelbaum, R. P. & Harthorn, B. H. (2000), 'Toward spatially integrated social science', *International Regional Science Review* **23**(2), 139–159.
- Guy, I., Ronen, I. & Wilcox, E. (2009), Do you know? recommending people to invite into your social network, *in* 'Proceedings of the 14th international conference on Intelligent user interfaces', pp. 77–86.
- Harrower, M. & Brewer, C. A. (2003), 'Colorbrewer.org: An online tool for selecting colour schemes for maps', *The Cartographic Journal* **40**(1), 27–37.
- Harvey, F. (2014), We know where you are. and were more and more sure what that means, *in* 'Emerging Pervasive Information and Communication Technologies (PICT)', Springer, pp. 71–87.
- Hauff, C. (2013), A study on the accuracy of flickr's geotag data, *in* 'Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 1037–1040.
- Hecht, B., Hong, L., Suh, B. & Chi, E. H. (2011), Tweets from justin bieber's heart: the dynamics of the location field in user profiles, *in* 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', ACM, pp. 237–246.
- Herlocker, J., Konstan, J., Terveen, L. & Riedl, J. (2004), 'Evaluating collaborative filtering recommender systems', *ACM Transactions on Information Systems (TOIS)* **22**(1), 5–53.
- Hey, A. J., Tansley, S., Tolle, K. M. et al. (2009), *The fourth paradigm: data-intensive scientific discovery*, Microsoft Research Redmond, WA.
- Hitsch, G. J., Hortaçsu, A. & Ariely, D. (2010), 'Matching and sorting in online dating', *The American Economic Review* **100**(1), 130–163.
- Horozov, T., Narasimhan, N. & Vasudevan, V. (2006), 'Using location for personalized poi recommendations in mobile environments', *SAINT* p. 124129.
- Hu, B. & Ester, M. (2013), Spatial topic modeling in online social media for location recommendation, *in* 'Proceedings of the 7th ACM conference on Recommender systems', ACM, pp. 25–32.

BIBLIOGRAPHY

- Hunt, G. R. (1977), 'Spectral signatures of particulate minerals in the visible and near infrared', *Geophysics* **42**(3), 501–513.
- Janowicz, K. (2012a), 'Observation-driven geo-ontology engineering', *Transactions in GIS* **16**(3), 351–374.
- Janowicz, K. (2012b), 'Observation-driven geo-ontology engineering', *Transactions in GIS* **16**(3), 351–374.
- Jensen, J. R. & Cowen, D. C. (1999), 'Remote sensing of urban/suburban infrastructure and socio-economic attributes', *Photogrammetric engineering and remote sensing* **65**, 611–622.
- Jones, C. & Purves, R. (2008), 'Geographical information retrieval', *International Journal of Geographical Information Science* **22**(3), 219–228.
- Jordan, T., Raubal, M., Gartrell, B. & Egenhofer, M. (1998), An affordance-based model of place in gis, in '8th Int. Symposium on Spatial Data Handling, SDH', Vol. 98, pp. 98–109.
- Joseph, K., Tan, C. H. & Carley, K. M. (2012), Beyond local, categories and friends: clustering foursquare users with latent topics, in 'Proceedings of the 2012 ACM Conference on Ubiquitous Computing', ACM, pp. 919–926.
- Kendall, M. G. & Smith, B. B. (1939), 'The problem of m rankings', *The annals of mathematical statistics* **10**(3), 275–287.
- Kloeckl, K., Senn, O. & Ratti, C. (2012), 'Enabling the real-time city: Live singapore!', *Journal of Urban Technology* **19**(2), 89–112.
- Kurashima, T., Iwata, T., Irie, G. & Fujimura, K. (2010), Travel route recommendation using geotags in photo sharing sites, in 'Proceedings of the 19th ACM international conference on Information and knowledge management', ACM, pp. 579–588.
- Kwan, M.-P., Casas, I. & Schmitz, B. C. (2004), 'Protection of geoprivacy and accuracy of spatial information: how effective are geographical masks?', *Cartographica: The International Journal for Geographic Information and Geovisualization* **39**(2), 15–28.
- Kwan, M.-P., Janelle, D. G. & Goodchild, M. F. (2003), 'Accessibility in space and time: A theme in spatially integrated social science', *Journal of Geographical Systems* **5**(1), 1–3.

BIBLIOGRAPHY

- Lee, J.-g., Han, J. & Whang, K.-Y. (2007), Trajectory Clustering : A Partition-and-Group Framework , *in* ‘International Conference on Management of Data’, pp. 593–604.
- Lee, L.-F. (1979), ‘Identification and estimation in binary choice models with limited (censored) dependent variables’, *Econometrica: Journal of the Econometric Society* pp. 977–996.
- Lee, M. & Chung, C. (2011), ‘A user similarity calculation based on the location for social network services’, *DASFAA* pp. 38–52.
- Leetaru, K., Wang, S., Cao, G., Padmanabhan, A. & Shook, E. (2013), ‘Mapping the global twitter heartbeat: The geography of twitter’, *First Monday* **18**(5).
- Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W. & Ma, W.-Y. (2008), ‘Mining user similarity based on location history’, *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems - GIS '08* p. 34.
- Lian, D., Zhu, Y., Xie, X. & Chen, E. (2014), Analyzing location predictability on location-based social networks, *in* ‘Advances in Knowledge Discovery and Data Mining’, Springer, pp. 102–113.
- Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P. & Tomkins, A. (2005), ‘Geographic routing in social networks’, *Proceedings of the National Academy of Sciences of the United States of America* **102**(33), 11623–11628.
- Lima, A. & Musolesi, M. (2012), Spatial dissemination metrics for location-based social networks, *in* ‘UbiComp 2012’.
- Lin, J. (1991), ‘Divergence measures based on the shannon entropy’, *Information Theory, IEEE Transactions on* **37**(1), 145–151.
- Lin, M., Hsu, W.-J. & Lee, Z. Q. (2012), Predictability of individuals’ mobility with high-resolution positioning data, *in* ‘Proceedings of the 2012 ACM Conference on Ubiquitous Computing’, ACM, pp. 381–390.
- Linden, G., Smith, B. & York, J. (2003), ‘Amazon. com recommendations: Item-to-item collaborative filtering’, *Internet Computing, IEEE* **7**(1), 76–80.
- Liu, Y., Sui, Z., Kang, C. & Gao, Y. (2014), ‘Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data’, *PloS one* **9**(1), e86026.

BIBLIOGRAPHY

- Lu, X., Wetter, E., Bharti, N., Tatem, A. J. & Bengtsson, L. (2013), 'Approaching the limit of predictability in human mobility', *Scientific reports* **3**.
- MacEachren, A. M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X. & Blanford, J. (2011), Senseplace2: Geotwitter analytics support for situational awareness, *in* 'Visual Analytics Science and Technology (VAST 2011)', IEEE, pp. 181–190.
- Malmi, E., Do, T. M. T. & Gatica-Perez, D. (2012), Checking in or checked in: comparing large-scale manual and automatic location disclosure patterns, *in* 'Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia', ACM, p. 26.
- Matyas, C. & Schlieder, C. (2009), 'A spatial user similarity measure for geographic recommender systems', *GeoSpatial Semantics* pp. 122–139.
- McCallum, A. K. (2002), Mallet: A machine learning for language toolkit. <http://mallet.cs.umass.edu>.
- McCloskey, M. E. & Glucksberg, S. (1978), 'Natural categories: Well defined or fuzzy sets?', *Memory & Cognition* **6**(4), 462–472.
- McKenzie, G., Adams, B. & Janowicz, K. (2013), A thematic approach to user similarity built on geosocial check-ins, *in* 'Geographic Information Science at the Heart of Europe', Springer, pp. 39–53.
- McKenzie, G. & Janowicz, K. (2014), Coerced geographic information: The not-so-voluntary side of user-generated geo-content, *in* 'Extended Abstracts of the Eighth International Conference on Geographic Information Science'.
- McKenzie, G., Janowicz, K. & Adams, B. (2014), 'A weighted multi-attribute method for matching user-generated points of interest', *Cartography and Geographic Information Science* **41**(2), 125–137.
- McKenzie, G., Janowicz, K., Gao, S., Yang, J.-A. & Hu, Y. (In Press), 'Poi pulse: A multi-granular, semantic signatures-based approach for the interactive visualization of big geosocial data', *Cartographica: The International Journal for Geographic Information and Geovisualization* .
- McKenzie, G. & Raubal, M. (2012), Ground-truthing spatial activities through online social networking data, *in* 'Extended Abstracts of the Seventh International Conference on Geographic Information Science (GIScience 2012), Columbus, OH'.

BIBLIOGRAPHY

- Miller, H. J. (1991), 'Modelling accessibility using space-time prism concepts within geographical information systems', *International Journal of Geographical Information System* **5**(3), 287–301.
- 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', Springer, pp. 350–370.
- NCHRP (2008), Report 571 standardized procedures for personal travel surveys, Technical report, National Cooperative Highway Research Program.
- Noulas, A., Mascolo, C. & Frias-Martinez, E. (2013), Exploiting foursquare and cellular data to infer user activity in urban environments, *in* 'Mobile Data Management (MDM), 2013 IEEE 14th International Conference on', Vol. 1, IEEE, pp. 167–176.
- Noulas, A., Scellato, S., Lathia, N. & Mascolo, C. (2012), International conference on data mining, *in* 'Mining User Mobility Features for Next Place Prediction in Location-based Services'.
- Noulas, A., Scellato, S., Mascolo, C. & Pontil, M. (2011), 'An empirical study of geographic user activity patterns in foursquare.', *ICWSM* **11**, 70–573.
- OrdnanceSurvey (2014), 'Points of interest', <http://www.ordnancesurvey.co.uk/business-and-government/products/points-of-interest.html>.
- Paek, J., Kim, J. & Govindan, R. (2010), Energy-efficient rate-adaptive gps-based positioning for smartphones, *in* 'Proceedings of the 8th international conference on Mobile systems, applications, and services', ACM, pp. 299–314.
- Páez, A. & Scott, D. M. (2007), 'Social influence on travel behavior: a simulation example of the decision to telecommute', *Environment and Planning A* **39**(3), 647.
- Palmer, J. R., Espenshade, T. J., Bartumeus, F., Chung, C. Y., Ozgencil, N. E. & Li, K. (2013), 'New approaches to human mobility: Using mobile phones for demographic research', *Demography* **50**(3), 1105–1128.
- Parkka, J., Ermes, M., Korpipaa, P., Mantyjarvi, J., Peltola, J. & Korhonen, I. (2006), 'Activity classification using realistic data from wearable sensors', *Information Technology in Biomedicine, IEEE Transactions on* **10**(1), 119–128.
- Plato, Annas, J. & Waterfield, R. (1995), *Plato: The Statesman*, Cambridge Texts in the History of Political Thought, Cambridge University Press.

BIBLIOGRAPHY

- 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.
- Pultar, E., Winter, S. & Raubal, M. (2010), 'Location-based social network capital', *GIScience, Extended Abstracts* .
- Purves, R., Edwardes, A. & Wood, J. (2011), 'Describing place through user generated content', *First Monday* **16**(9).
- Quercia, D. & Saez, D. (2014), 'Mining urban deprivation from foursquare: Implicit crowdsourcing of city land use', *Pervasive Computing, IEEE* **13**(2), 30–36.
- Raper, J. (2000), *Multidimensional geographic information science*, CRC Press.
- Raubal, M., Miller, H. J. & Bridwell, S. (2004), 'User-centred time geography for location-based services', *Geografiska Annaler: Series B, Human Geography* **86**(4), 245–265.
- Relph, E. (1976), *Place and placelessness*, Vol. 67, Pion London.
- Rodriguez, M. A. & 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.
- Rubner, Y., Tomasi, C. & Guibas, L. J. (1998), A metric for distributions with applications to image databases, *in* 'Computer Vision, 1998. Sixth International Conference on', IEEE, pp. 59–66.
- Rubner, Y., Tomasi, C. & Guibas, L. J. (2000), 'The earth mover's distance as a metric for image retrieval', *International Journal of Computer Vision* **40**(2), 99–121.
- Russell, S. J. & Norvig, P. (2010), *Artificial intelligence: a modern approach*, Prentice hall Upper Saddle River, NJ.
- Scellato, S., Noulas, A. & Mascolo, C. (2011), Exploiting place features in link prediction on location-based social networks, *in* 'Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 1046–1054.
- Schilit, B. N., LaMarca, A., Borriello, G., Griswold, W. G., McDonald, D., Lazowska, E., Balachandran, A., Hong, J. & Iverson, V. (2003), Challenge: Ubiquitous location-aware computing and the place lab initiative, *in* 'Proceedings

BIBLIOGRAPHY

- of the 1st ACM international workshop on Wireless mobile applications and services on WLAN hotspots', ACM, pp. 29–35.
- Schowengerdt, R. A. (2006), *Remote sensing: models and methods for image processing*, Academic press; 3 edition.
- Sellars, W. (1963), 'Philosophy and the scientific image of man', *Science, perception and reality* **2**, 35–78.
- Shamai, S. (1991), 'Sense of place: An empirical measurement', *Geoforum* **22**(3), 347–358.
- Shaw, B., Shea, J., Sinha, S. & Hogue, A. (2013), Learning to rank for spatiotemporal search, in 'Proceedings of the sixth ACM international conference on Web search and data mining', ACM, pp. 717–726.
- Sheth, A., Anantharam, P. & Henson, C. (2013), 'Physical-cyber-social computing: An early 21st century approach', *Intelligent Systems, IEEE* **28**(1), 78–82.
- Silva, L., Hoffer, R. & Cipra, J. (1971), 'Extended wavelength field spectroradiometry', *Proc. 7th Int. Symp. on Remote Sensing of Environment* **2**, 1509–1518.
- Silva, T. H., de Melo, P. O. V., Almeida, J., Musolesi, M. & Loureiro, A. (2014), You are what you eat (and drink): Identifying cultural boundaries by analyzing food & drink habits in foursquare, in 'International Conference on Web and Social Media', AAAI.
- Song, C., Qu, Z., Blumm, N. & Barabási, A.-L. (2010), 'Limits of predictability in human mobility', *Science* **327**(5968), 1018–1021.
- Subrahmanyam, K., Reich, S. M., Waechter, N. & Espinoza, G. (2008), 'Online and offline social networks: Use of social networking sites by emerging adults', *Journal of Applied Developmental Psychology* **29**(6), 420–433.
- Szalai, A. et al. (1972), 'The use of time: Daily activities of urban and suburban populations in twelve countries.', *The use of time: daily activities of urban and suburban populations in twelve countries.* .
- Tanasescu, V., Jones, C. B., Colombo, G., Chorley, M. J., Allen, S. M. & Whitaker, R. M. (2013), The personality of venues: Places and the five-factors ('big five') model of personality, in 'Computing for Geospatial Research and Application (COM. Geo), 2013 Fourth International Conference on', IEEE, pp. 76–81.

BIBLIOGRAPHY

- Tawk, Y., Tomé, P., Botteron, C., Stebler, Y. & Farine, P.-A. (2014), 'Implementation and performance of a gps/ins tightly coupled assisted pll architecture using mems inertial sensors', *Sensors* **14**(2), 3768–3796.
- TEA (2014), 'Tea/aecom 2013 global attractions report', http://www.aecom.com/deployedfiles/Internet/Capabilities/Economics/_documents/ThemeMuseumIndex_2013.pdf". Retrieved June 6, 2014.
- Tiropanis, T., Hall, W., Shadbolt, N., De Roure, D., Contractor, N. & Hendler, J. (2013), 'The web science observatory', *IEEE Intelligent Systems* **28**(2), 100–104.
- Tuan, Y.-F. (1977a), 'Sense and place', *Schoff, Gretchen Holstein, and Yi-Fu Tuan. Two Essays on a Sense of Place* pp. 1–13.
- Tuan, Y.-F. (1977b), *Space and place: The perspective of experience*, U of Minnesota Press.
- Tuan, Y.-F. (1979), *Space and place: humanistic perspective*, Springer.
- Tversky, A. (1977), 'Features of similarity', *Psychological Review* **84**(4), 327–352.
- Twitter (2014), 'Twitter developer documentation', <https://dev.twitter.com>.
- Vicente, C. R., Freni, D., Bettini, C. & Jensen, C. S. (2011), 'Location-related privacy in geo-social networks', *Internet Computing, IEEE* **15**(3), 20–27.
- Wang, H., Terrovitis, M. & Mamoulis, N. (2013), Location recommendation in location-based social networks using user check-in data, *in* 'Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems', ACM, pp. 364–373.
- Watson, G. S. (1961), 'Goodness-of-fit tests on a circle', *Biometrika* pp. 109–114.
- Winter, S. & Freksa, C. (2014), 'Approaching the notion of place by contrast', *Journal of Spatial Information Science* **2014**(5), 31–50.
- Winter, S., Kuhn, W. & Krüger, A. (2009), *Spatial Cognition and Computation: Computational Models of Place (Special Issue)*, Vol. 9, Taylor & Francis.
- Wu, L., Zhi, Y., Sui, Z. & Liu, Y. (2014), 'Intra-urban human mobility and activity transition: Evidence from social media check-in data', *PloS one* **9**(5), e97010.

BIBLIOGRAPHY

- Xiao, X., Zheng, Y., Luo, Q. & Xie, X. (2010), Finding similar users using category-based location history, *in* ‘Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems’, ACM, pp. 442–445.
- Ye, M., Janowicz, K., Mülligam, C. & Lee, W.-C. (2011), What you are is when you are: the temporal dimension of feature types in location-based social networks, *in* ‘Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems’, ACM, pp. 102–111.
- Ye, M., Shou, D., Lee, W.-C., Yin, P. & Janowicz, K. (2011), On the semantic annotation of places in location-based social networks, *in* ‘Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining’, ACM, pp. 520–528.
- Ying, J. J.-C., Lu, E. H.-C., Lee, W.-C., Weng, T.-C. & Tseng, V. S. (2010), ‘Mining user similarity from semantic trajectories’, *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks - LBSN '10* pp. 19–26.
- Yuan, Q., Cong, G., Ma, Z., Sun, A. & Thalmann, N. M. (2013), Time-aware point-of-interest recommendation, *in* ‘Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval’, ACM, pp. 363–372.
- Yuan, Q., Cong, G. & Sun, A. (2014), Graph-based point-of-interest recommendation with geographical and temporal influences, *in* ‘Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management’, ACM, pp. 659–668.
- Yuan, Y., Raubal, M. & Liu, Y. (2012), ‘Correlating mobile phone usage and travel behavior—a case study of harbin, china’, *Computers, Environment and Urban Systems* **36**(2), 118–130.
- Zar, J. H. (1976), ‘Watsons nonparametric two-sample test’, *Behavior Research Methods* **8**(6), 513–513.
- Zeto III, M. J., Rippetoe, D., Shaw, D., Mercer, A. R., Gaxiola Jr, G., Williams, R. T. & Johansson, E. A. O. (2013), ‘System and methods for delivering targeted marketing content to mobile device users based on geolocation’. US Patent App. 13/911,956.

BIBLIOGRAPHY

Zhang, J.-D. & Chow, C.-Y. (2013), igslr: personalized geo-social location recommendation: a kernel density estimation approach, *in* 'Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems', ACM, pp. 324–333.