A

Cohesion and community: the effects of space, place and time on foursquare check-ins in New York City

KENNETH JOSEPH, Carnegie Mellon University
KATHLEEN M. CARLEY, Carnegie Mellon University
JASON I. HONG, Carnegie Mellon University

In previous work, we used Latent Dirichlet Allocation (LDA), a technique commonly applied to document collections to discover latent themes, to cluster foursquare users in New York City. We found that although the feature set used was agnostic of geo-spatial location, time, users' friends on social networking sites and venue function, qualitative evidence existed that groups of people of different types (e.g. tourists), communities (e.g. users tightly clustered in space) and interests (e.g. people who enjoy athletics) could be uncovered. In the present work, we use the same feature set and a similar methodology, but we extend these efforts in seeking a more quantitative understanding of why groups of users frequent certain venues. Specifically, we develop metrics to test the cohesiveness in time, space and function of sets of venues uncovered by LDA that are checked in to by similar users. We find that nearly all venue sets the model uncovers are more cohesive than we would expect by chance along one or more of these metrics, supporting previous work in a variety of domains. In addition, we discover a significant negative correlation between the spread of venue sets in space and function, thus suggesting a "neighborhood" effect observed in other recent work. Finally, we show that the model captures distinct "micro-cultures" within the city and discuss how we can understand these based on the notion of self-representation and by leveraging latent connections between users. These findings are intended to both support and inform social science in the way that location-based services can help to understand community and human behavior in the urban environment.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Sociology; H.1.2 [**User/Machine Systems**]: Human Information Processing

General Terms: Urban computing, topic modeling, location-based service

Additional Key Words and Phrases: foursquare, topic modeling, community structure, urban analytics

ACM Reference Format:
Joseph, K., Carley, K.M., Hong, J.I. 2013. Cohesion and community: the effects of space, place and time on foursquare check-ins in New York City ACM Trans. Intell. Sys. and Tech. V, N, Article A (January 2013), 22 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

This work was supported in part by the Office of Naval Research (ONR) through a MURI N00014081186 on adversarial reasoning. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research or the U.S. government.
Author's addresses: K. Joseph and K.M. Carley and J.I. Hong, School of Computer Science, Carnegie Mellon University

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

Over the past decade, location-sharing services like foursquare[1] and Facebook Places[2] and the increased number of mobile phones have produced massive quantities of information on the movements, actions and social structure of large human systems. Such data is exponentially more granular, more accurate and larger in scale than location data collected in the past via more traditional means, such as surveys. The availability of such data has allowed researchers to better understand how large collections of humans behave and interact as they move through space, providing verifications and extensions of traditional social science concepts in this realm. For example, recent work has confirmed at the scale of an entire nation that social structure is still largely clustered by distance [Onnela et al. 2007], and that people within cities tend to go about their day within relatively distinct regions that do not always conform to government-induced boundaries [Cranshaw et al. 2012].

In the present work, we focus on garnering a better understanding of how different factors combine to affect the places (or venues) people go within a city. In particular, we consider how human tendencies to stay within small geographic spaces [Brockmann et al. 2006], to move with predictable circadian rhythms [Cho et al. 2011] and to go to places of similar function [Lindqvist et al. 2011] play important but distinct roles in the specific venues that people go in New York City. While we will discuss recent work focusing on similar questions, our approach is unique from many in that it asks how these different factors affect the specific venues people go and not just their position in space. We thus seek to better understand questions of cohesion amongst places frequented by the same people - in other words, what is it about two places that causes people to attend them both regularly?

Our ability to focus on the specific place a person goes to, as opposed to simply their position in space, is due to our use of public "check-ins" of users of foursquare. Foursquare is a socially-driven location sharing application [Lindqvist et al. 2011], where users can check-in to different places (e.g. the Starbucks on 10th Street) and have these check-ins be shared with friends both on foursquare and other social networking sites. Because of the richness of the data generated, check-ins on foursquare have been the focus of much recent work in a similar vein to the questions explored here (e.g. [Noulas et al. 2011b; Cheng et al. 2011; Cranshaw et al. 2012; Lindqvist et al. 2011; Ferrari et al. 2011; Joseph et al. 2012; Tang et al. 2010; Noulas et al. 2011b]).

Using the foursquare check-ins of over 10,000 users in New York City from a larger dataset collected by the authors of [Cranshaw et al. 2012], we group people using Latent Dirichlet Allocation (LDA), a method for unsupervised clustering. This clustering is done with only the unique identifier of each place - thus, the place's position in space, temporal distribution and function are not utilized in the feature set. Clusters of users are discovered by associating them with a set of hidden "topics", learned by the model based on the shared check-ins of users. Each topic can be thought of as a set of characteristic venues LDA generates that are frequented by similar users.

Given these collections of venues, obtained only by considering the frequency with which users check-in to them, we ask two questions to better understand how places frequented by similar people are cohesive in time, location and function. The first question is whether or not these places are more cohesive along these dimensions than one would expect by chance. In order to determine this, we create a randomized set of "fake" topics from the same venues, which provides a null model for comparison. We find that in many cases, the topics LDA generates are significantly cohesive in one or

[1] http://www.foursquare.com
[2] https://www.facebook.com/about/location

more of time, space and function at a significance level of $\alpha<.05$. Interestingly, however, we find that it is rarely the case that these venues are more cohesive in *all* of these ways, suggesting that human behavior is not always affected by each of these factors unconditionally. The second question we ask is how these different factors are correlated - for example, when venues are closer in space, do they have similar temporal structures? We find here a negative correlation between the diversity of venue functions within a topic and the geo-spatial spread of these venues, suggesting that within the city, regions confined to smaller spaces frequented by the same people often satisfy a more diverse set of needs. This finding validates recent work on the subject of neighborhoods by Cranshaw et al. 2012. In addition to these two questions, we complete exploratory work on how the network of topics our model generates exposes isolated communities of users and opportunities to drive integration between otherwise disparate populations within a city.

This paper is an extension of a previous work by similar authors [Joseph et al. 2012], where LDA is explored as a mechanism for understanding different types of people, community structures and interest groups within the city. The present work extends previous efforts in several ways. First, we use a more quantitative approach to determining the number of topics appropriate for LDA and use a different implementation which allows for a more accurate result. Second, we extend our review of the related work to include several recent papers also considering LDA on foursquare data, giving a better idea of the novelty of our work. Third, we consider quantitatively the cohesion within topics in time, space and function, as opposed to a strictly qualitative approach in the previous work. Finally, we here consider a network of topics via shared users and consider its implication on future work on understanding culture and community within the urban environment.

The rest of the paper is structured as follows. In Section 2, we detail the dataset studied, research studying the typical usage patterns of foursquare and a more detailed explanation of LDA. In Section 3, we describe previous work on human behavior using location data. In Section 4, we then describe in detail the approach taken - how we select the number of topics for LDA, and how we understand cohesion within a topic in time, space and function. In Section 5, we consider the results of the two main research questions described above. In Section 6, we consider the network of topics and what it suggests for interesting avenues for future work. We finish with some concluding remarks in Section 7.

2. BACKGROUND

2.1. Data and Cleaning

The data we use is a set of approximately 360,000 check-ins posted to Twitter from users of foursquare located in New York City, and is part of a larger dataset given to us by the authors of [Cranshaw et al. 2012]. The collection period was spread over two distinct segments, but in total the data comprises approximately 18 months of foursquare check-ins. In the dataset used, a check-in provides a unique user ID from Twitter, the time-stamp of the check-in, an optional user description (e.g. “the coolest place ever!”) and the ID of the venue at which the check-in occurred. Using this venue’s ID, the original data collectors also obtain the venue’s name, geo-location, and “category” information by querying the foursquare API[1]. These categories are drawn from a set of hierarchical names given by foursquare - examples include “Food::Burger Joint”, “Food::Bakery” and “Travel Spots::Boat or Ferry”, where the “::” operator separates levels of the hierarchy.

[1]https://developer.foursquare.com/index

Check-ins to Twitter are a specific and likely biased subsample of all foursquare check-ins, as users are not required to share all check-ins with their Twitter followers. Lindqvist et al. 2011 found that only 18% of the users surveyed allowed their check-ins to be posted to Twitter, though Cramer et al. 2011 found, less than a year later, that 68% of users had their foursquare accounts linked to Twitter. However, 63% of those studied by Cramer et al. 2011 had not shared their last check-in on Twitter, for reasons most often associated with either the potential of a check-in to annoy followers or only wanting to push "interesting" check-ins to this more public sphere.

This notion of pushing only interesting check-ins to Twitter extends to check-ins in general, and has been studied by several researchers under the term self-representation. Self-representation argues that people only check-in to places that represent them in a manner they desired to be viewed, and thus may or may not truly represent their actual interests and the places they actually frequent. For example, users surveyed by Lindqvist et al. 2011 tended not to want to check-in to places they perceived to be uninteresting (e.g. work) or embarrassing (e.g. fast food). The effects of self-representation have also been observed by Tang et al. 2010, who draw on social psychology literature to discuss the theoretical groundings for this effect.

Such findings affect our understanding of foursquare data in two important ways. First, they present an interesting and complicating factor to applying the notion of homophily [McPherson et al. 2001], which states that people can often be grouped together based on shared characteristics, to foursquare data in that it is difficult to determine real versus projected interests. Similarly, as we will see, when attempting to understand community as expressed by venues frequented by similar users, it is difficult to tell whether the user is actually a member of the community, or whether they merely want to be perceived as a member. As we will argue, however, both of these represent interesting social phenomenon regarding community and culture in the city.

2.2. Latent Dirichlet Allocation (LDA)

LDA is a member of a larger family of Bayesian frameworks referred to as "topic models". It was first introduced by Blei et al. 2003 as a latent space model that can be used to better understand text corpora by representing a large collection of documents in a more compact set of hidden topics. In a typical usage of LDA, a text document is represented as a set of words, where each word is assumed to belong to one or more hidden topics. Thus, each document can be described by considering how heavily the words within it relate to the various topics, and each topic can be described by the words most heavily associated with it.

In order to model user check-in behaviors with LDA, we use the analogy of a document to represent a user, and each word to represent a specific venue that a user checked in to. For example, a user who has checked in to Yankee Stadium twice and the local Pizza shop four times could be represented by the vector $< ...0, 0, 2, 0, 4, 0... >$, where zeros and ellipses are intend to show that in LDA, all possible venues (i.e. any venue attended by any user in the dataset) must be represented for each user. Thus, a better understanding of features that can cluster users in lower dimensional space is desirable, and something that our work suggests is readily possible. Note that we represent each venue as being unique from all others- this means, for instance, that the Starbucks on 5th Street will be different than the Starbucks on 10th Street.

Given this model, there are two different but analogous interpretations of the output generated by LDA in the present work. From an unsupervised clustering perspective, the topics generated are discovered by the frequency with which venues are attended by the same users, who themselves attend similar sets of places. Thus, clustering is done on the users, who are represented by the model to be a distribution over all topics (but are almost always highly associated with a small subset of them). Similarly,

topics are defined by the venues that these users have in common - they are a distribution over all venues, but will almost always be represented by a small set of venues representing a set of places frequently checked in to by similar users.

In the form of the generative process often used to describe LDA [Blei et al. 2003], topics represent latent factors that drive a user's check-in behaviors. That is, we assume that each check-in made is driven by a latent factor, which in turn determines their likelihood of checking in at a particular venue. This model, while perhaps more difficult to interpret, fits with models of human thought along the lines of both discrete choice [McFadden 1980] and schema-based [Rumelhart 1978] frameworks for cognition.

However, while such a fit provides precedence for the use of LDA, it is important to understand that the model comes with several other assumptions, many of which have been relaxed in later topic models. In particular, our use of LDA induces the assumption that the order of check-ins is irrelevant. Though, as we will see, check-ins close in time have a tendency to be close together in space, we are interested in latent factors driving general check-in behavior, and thus we feel that a unigram model captures this concept effectively. Second, LDA presumes no strict correlations between the different factors. Later models, including the Correlated Topic Model (CTM) [Blei and Lafferty 2007] and more recently hierarchical PAM [Li et al. 2012] relax this assumption. However, we find that tests using the CTM faired significantly worse on the data we use. It is possible that the reason such a model does not fair well is the poor distributional characteristics of the data (in terms of numbers of check-ins per user), but a better understanding of this is important in future work.

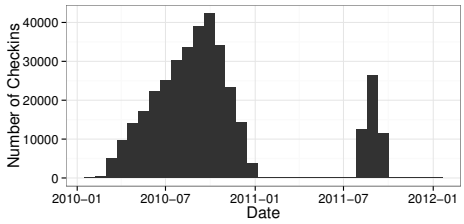


Fig. 1. Histogram of the number of check-ins over the duration of our data set

Finally, LDA does not allow us to model the dynamics of topics in the data, found to be a limitation of topic models applied to this data in previous work [Joseph et al. 2012]. Though we will give evidence that a dynamic model is appropriate for the task we are interested in, the data we use is segmented into two periods of data collection several months apart (see the histogram in Figure 1). Such a distribution would affect our usage of relevant dynamic topic models, in particular the work on latent periodic topic analysis [Yin et al. 2011]. We intend to explore this interesting avenue on different data in the future.

3. RELATED WORK

In working to understand human behavior with location-based data, many researchers have focused specifically on modeling and quantifying movement from location to location. Early work modeled human movement as a Lévy flight model [Brockmann et al. 2006], a model also found to well-approximate the distance between two successive check-ins by the same user on foursquare [Cheng et al. 2011]. Similarly, [Noulas et al. 2011b] study distances users travel between successive check-ins, noting (as we would

expect from a distribution modeled as a Lévy flight) that nearly 80% of the total check-ins for a user occur within 10 kilometers of the previous one. Along these same lines, work in [Noulas et al. 2011a] finds it is nearly impossible to tell how far two people will travel on the range of [0-100] meters, but that movement can be predicted with high accuracy in different cities by taking into account the density of venues in those cities. Such studies point to the obvious effect of space on where users check-in - given these findings, we would expect that venues similar users attend should be closer in space than we would expect by chance.

While such models give insight into the problem at hand, other work has focused on how temporal characteristics of a person's behavior tend to regulate their movement through space. In one of the earlier works studying large-scale human behavior through location-based data, Gonzlez et al. 2008 observe that the Lévy Flight model of human mobility can be explained to a large extent by human circadian rhythms, by which people tend to frequent the same few places with striking periodicity. Such periodicity is observed in later works, including that by Becker et al. 2011, who find that people who worked in a suburban American town could be distinguished from people who "partied" there by the times at which they were active in different sections of the town.

Specific to foursquare data, Bauer et al. 2012 use a novel, spatio-temporal topic model to understand the temporal and geographic regularities of different words in the textual content of tweets that included a foursquare check-ins. They observe that regularities in time and space of different words uncover the dynamics of certain regions in New York City, such as areas of work and areas of tourism. Ferrari et al. 2011 and Kling and Pozdnoukhov 2012 also use LDA to understand the temporal and geo-spatial dynamics of different cities (including New York City), finding clear distinctions in temporal signatures between different topics. These efforts provide core evidence that topic models can uncover collections of venues with specific temporal signatures. However, these works give little evidence as to how cohesive these signatures are *within* the different topics- is it a single venue within the topic driving the given pattern, or do all venues have the same temporal signatures? Given work suggesting periodicity in human behaviors, we would expect that most topics would be highly heterogeneous in their temporal signatures, as people tend to frequent the same places at different times of the day and week. Interestingly, we observe that there exist both cohesive and highly incohesive topics along the temporal dimension in Section 5.

In addition to temporal effects on human movement, there has been significant interest in how a person's social connections affect his or her movement. One might be tempted to assume that a person's check-ins can be better understood if they are conditioned on the check-ins of friends. Indeed, the location of a user can often be predicted based on the location of their friends on social networking sites (see [Sadilek et al. 2012; Scellato et al. 2011] for recent examples). Furthermore, recent work has shown that neighborhoods implied by census boundaries can be inferred from social graphs [Hipp et al. 2012], an obvious indicator of the relationship between location and social spheres. However, evidence from [Cho et al. 2011] suggests that while location prediction (e.g. a latitude-longitude point) can be done with reasonable accuracy, predicting the specific place a user will go based on where their friends on location-based services go is not as straightforward. Cho et al. 2011 observe that people who are friends on Brightkite and Gowalla have a check-in in common less than ten percent of the time. Thus, due in part to this evidence and in part to our inability to infer social ties from the data we use, cohesion in social structure is not considered in the present work.

In addition to temporal and social effects, the function of a place has been seen to affect human movement as well. Lindqvist et al. 2011 revealed a bimodal distribution of check-ins at more private locations, such as home and work. While most users never

checked in to places with these categories, those who did tended to do it frequently - one to two times per day. Similarly, interviews by Cramer et al. 2011 suggest a reinforcement effect- places that are observed to be popular received increased interest. Thus, popularity, which is often related to function, may be a factor in driving similar people to similar venues.

In noting these effects are all intricately intertwined, many have considered how various combinations of place function, temporal signature, social effects and proximity in space affect user movement. Cho et al. 2011 uncover correlations between human geographic movement, temporal dynamics and the social structure of the population they study. They create a spatio-temporal model, including two different locations functioning as a "home" and a "work" to model human movement within the city. Jiang et al. 2012 complete principal component analysis on large-scale survey data and then use K-means clustering to understand the spatio-temporal distributions of different activities in different locations in the Chicago area. Yuan et al. 2012 use a topic model on transitions of taxi-cabs within Beijing, where distributions of topics are conditioned on a dirichlet multinomial regression (DMR) on points of interest within different regions of the city. The model is used to understand areas within Beijing with different functional characteristics.

Cranshaw et al. 2012 use spectral clustering to understand how foursquare data gives insight into the dynamics of neighborhood boundaries in urban areas. Venues are linked in a graph based on their distance from each other, and the weight of the link then becomes a function of the number of users who visited both locations. This work is similar in the question it asks to work by similar authors, which utilizes LDA to better understand neighborhoods within regions in space [Cranshaw and Yano 2010]. In both [Cranshaw et al. 2012; Cranshaw and Yano 2010], the authors observe that regions frequented by similar users within a tight geographical space tended to have a diverse set of functions-often people within these regions found places to go to do what they wanted to do, and thus found little reason to travel outside of their neighborhood. We consider how this correlation manifests in our data as well, which uses a similar model but removes the constraint of linking only venues within a given distance.

Given the work above, it is clear that our methodology is well-grounded in the literature in two obvious ways. First, topic modeling, and even LDA specifically, has been heavily utilized to reduce dimensionality and to better understand human behavior using location-based data. We differentiate our work from the others along these lines in three ways. First, the manner in which we define the input to LDA is distinct from any previous feature set we are aware of - in particular, we use LDA chiefly as a mechanism for unsupervised clustering, whereas many of the previous works used it specifically in its original, text-based conceptualization. Second, unlike nearly all of the previous works, we employ an empirically rigorous methodology to select the number of topics appropriate for the task, as discussed in the following section. Finally, we consider cohesion within the topics generated by LDA, as opposed to only comparing across topics. We find that certain aggregate statistics are better than others in certain cases-for example, some topics simply cannot be explained by their temporal regularities, and thus may be better described by, for instance, their function. The second way in which our methodology is well-grounded in the literature is that the effects of time, geo-spatial location and function on the places a person go have been heavily studied. While this is the case, no study we are aware of has focused specifically on quantifying the extent to which venues frequented by similar users are intertwined in time, space and function as we do here

4. METHODS

4.1. Model

In order to ready the dataset we obtain for use with LDA, we remove users who checked-in at less than 5 unique venue and venues with less than 10 total check-ins. We repeated pruning iteratively until all such venues and users were removed. This approach of pruning data points is common in document modeling, as in imperfect documents there tend to be spelling mistakes and vocabulary unique to specific documents which are thus not of interest in understanding general trends. We begin with a total of 448,156 check-ins, 36,388 users and 44,312 venues. After pruning, the dataset we inputted to the model consisted 364,633 check-ins, 10,652 users and 34,812 venues. Note that, due to the power-law distribution of check-ins noted in [Noulas et al. 2011b] (and confirmed in our data), we still keep more than half of the check-ins while the number of venues and users decreases significantly.

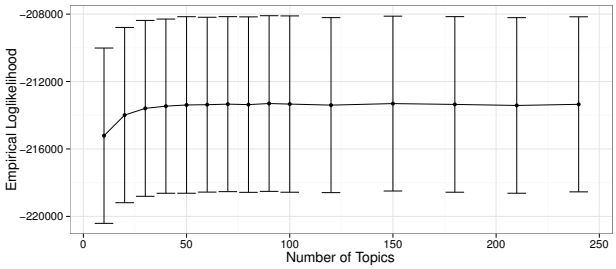


Fig. 2. Empirical Loglikelihood (y-axis) of the model at different numbers of topics (x-axis). Error bars are 95% CI

One pitfall of LDA not discussed above is the need to fix the number of topics, k, when deciding on a model. In order to determine the optimal k for the data we study, we compute the average empirical log-likelihood of the model at various k using the "left-to-right" algorithm described in [Wallach et al. 2009b] on left-out data using ten-fold cross validation. We test the range of values for k = $10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 130, 160, 190, 210, 240$. Figure 2 shows the results of our cross validation, where error bars represent 95% confidence intervals around the mean log-likelihood. In Figure 2, the higher the mean value of empirical log-likelihood, the better the model with the given k is at predicting the location of users in the held-out portion of the data. We find that there exists an obvious increase in the mean empirical log-likelihood up until fifty topics, at which point the means are nearly identical. Given the lack of an obvious number of topics to choose, we select the k with the absolute maximum mean empirical likelihood, which is 90 topics. Wallach et al. 2009a state that with an asymmetric prior on α (which we use here), it is safe to choose a reasonably large number of topics, justifying our choice of 90 over smaller numbers of topics. Furthermore, we find, qualitatively, that with numbers higher than 90, the topics we observe to be interesting still exist. We therefore choose a number that maximizes our empirical exploration of k while also allowing us to minimize difficulties in interpreting the model.

Final results were calculated using a parallelized version of LDA implemented in MALLET[3]. It is important to note that, as mentioned, the model we use computes

[3]http://mallet.cs.umass.edu

the posterior using an asymmetric prior, which has a tendency to move more popular "words" (in our case, venues) into the same topics [Wallach et al. 2009a]. Though we utilize such a method because it increases the empirical log-likelihood of the model, it also influences many of the topics LDA discovers to be composed of sets of venues with a small number of check-ins. In order to remove the dependency of our results on such uninteresting topics, we do not consider topics having less than 1000 check-ins total across the top 15 venues used to represent it (as described in the following section). Thus, in our analysis, we consider the results from only 48 of the original 90 topics.

4.2. Cohesion Metrics

In most uses of LDA (e.g. [Blei et al. 2003]), the top N words (venues) most closely associated with each topic are selected as being representative of the topic, and the rest are ignored. We choose fifteen venues a priori to represent each topic, and do not change the value throughout the analysis. Post-hoc testing suggests that our findings are reasonably robust to both including more and removing venues from this count.

In order to understand how cohesive the different topics LDA discovers are in time, space and function, we can now simplify this task to developing metrics to understand the cohesiveness of these fifteen venues. We focus first on function, for which we leverage the hierarchical categories associated with each venue in our dataset. In order to analyze cohesiveness in place function, we consider metrics for both a high-level function, derived from the top-level category of a venue, and a low-level function, derived from the full categorical description of each venue. Thus, for example, Madison Square Garden, which has the category "Arts & Entertainment::Stadium", would have a high-level function of "Arts & Entertainment" and a low-level function of "Arts & Entertainment::Stadium". Because each place has a single high-level and low-level function, a natural way to analyze the discrete distribution resulting from these values for each topic is information entropy. Information entropy, defined as $-\sum_{i=1}^{|Functions|} p(function_i) * log(p(function_i))$, measures the uniformity of this discrete distribution. In the equation, $|Functions|$ represents the number of possible functions at the high-level (approximately 10) or low-level (approximately 300), depending on the metric being calculated.

Both time and geo-spatial location define continuous spaces within which information entropy is not found to be the most natural definition of cohesion. Instead, for each, we define each venue with some vector, determine an appropriate distance metric for these vectors, and then compute the average pairwise distance between all fifteen places. For geo-spatial cohesion, the representation of each venue is its latitude and longitude. The distance metric utilized is the "Manhattan distance", defined as $|latitude_i - latitude_j| + |longitude_i - longitude_j|$, where i and j are two venues within a topic. While we use the Manhattan distance due to the fact that most venues are located on the island of Manhattan, we find that there is little difference in the rank-order of the topics when using more intricate distance measures.

A vector representation of the temporal distribution of a given place is not nearly as straightforward. Given the uneven distribution of our dataset over longer periods of time, we choose to define the temporal distribution of a venue as the number of check-ins that occurred during each hour of each day within the period of a week. Thus, each venue is represented as vector of 168 values, one for every hour of every day of the week. The value at each position in the vector is the number of times that any user checked in to that venue at that hour/day combination over the entire data set. This vector, when normalized, becomes a probability distribution for the likelihood of a user checking in at a location at any given hour of any given day over the period of one week.

Similar binning approaches were used by Ye et al. 2011, who defined a temporal metric using a simple kernel smoothing method on binned data to define the dissimilarity in temporal profiles between two distributions. Instead, we use a simplification of the Kolmogorov-Smirnov test statistic [Massey 1951] as our distance metric. In order to do so, we first compute the Empirical Cumulative Distribution Function (ECDF) for each venue. In order to describe the ECDF, we use an example where we consider the likelihood of a user checking in to a given location on any day of the week. This gives us vector of length seven, for example, $< .1, .1, .1, .1, .1, .2, .3 >$. The ECDF of a discrete probability distribution, defined as $\frac{1}{n}\sum_{i=1}^{n} I(X_i < t)$, where t is the day of the week, is simply the mass of the distribution occurring on or before a given day. Thus, the ECDF for the given vector would be$< .1, .2, .3, .4, .5, .7, 1 >$.

The distance metric we use simply takes the supremum (analogous to the maximum) difference between the ECDFs of two places at any index in the vector. For example, if we had two vectors, $< .2, .6, 1 >$ and $< .1, .9, 1 >$, the outcome of our distance metric would be equivalent to $max(|.2-.1|, |.6-.9|, |1-1|)$ and thus equal to .3. By considering the differences between the ECDFs of our temporal representation of different venues in this way, we find that our metric does a reasonable job of accounting for slight differences in the temporal structure between two different venues without the parametric requirements of the metric proposed in [Ye et al. 2011]. However, our method does suffer, as we will see, from differences in the magnitudes of check-ins across different factors- this is also a difficulty in applying more standard time-series clustering approaches, such as iSAX [Camerra et al. 2010]. Future work to improve the given metric is thus an avenue we hope to approach.

4.3. Obtaining random topics

While we would expect that our topics are going to be cohesive in time, space and/or function, it is not necessarily the case that they will be more cohesive than a random set of venues within the city across all factors. For example, Cranshaw et al. 2012 suggest that topics tightly clustered in geo-space may not necessarily be particularly cohesive in function. Additionally, work such as [Cho et al. 2011] and [Noulas et al. 2011a] suggest that humans travel with relative freedom in reasonably-sized geo-spatial regions. Thus, in order to understand whether or not the topics we find are truly more cohesive than one would expect by chance, it is necessary to create some form of a null model for comparison.

We create a random sample of 1000 topics which we can use to test the significance of cohesion in our "real" topics (i.e. those generated by LDA). To generate the random topics, we pull 15 venues uniformly from the venues representing each topic discovered by LDA. With 48 topics (recall we only consider topics with greater than 1000 check-ins total), this means we uniformly sample 15 venues out of a possible 720 one thousand times to generate our random topics. We can utilize these random topics to obtain an understanding of the significance of the cohesion of the real topics along each of the different metrics. That is, if we select a significance level of α<.05 (where α refers to the level at which we reject the null hypothesis and not the prior in LDA), we can consider a real topics to be significantly cohesive on a given metric if the value of that metric for that topic shows it is more cohesive than the random topic at the 5th percentile of the randomized topic set. In all cases below, we utilize a significance level of α<.05, given its typical usage in the social science literature. Where we refer to significance or to likelihood greater than chance, we are thus referring to the process described here.

5. RESULTS

Table I shows the percentage of topics that are more cohesive than we would expect by chance for each metric, along with the percentage of topics not significantly cohe-

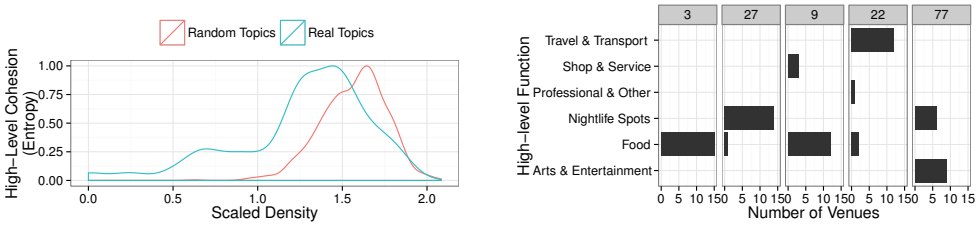
Table I. Percentage of topics that were more cohesive than expected by chance for each metric, as well as percentage of topics not significantly cohesive on any metric

Time	48%
High-Level Function	33%
Low-Level Function	44%
Distance	69%
None	13%

sive on any of our metrics. As expected, many of the topics were cohesive in distance- nearly 70% of the topics were more cohesive than we would expect by chance. In contrast, we were reasonably surprised to find that nearly half of our topics- 48%- were significantly cohesive in their temporal patterns. Given the findings of [Cho et al. 2011] in the periodicities of user movement, we would suspect most topics would have had highly variable patterns in time. Finally, nearly 13% of our topics (6 out of 48) were not significantly cohesive along any of the metrics we used. As we will discuss, such topics tended to represent areas outside of Manhattan, suggesting that while the were not cohesive as compared to the random topics generated here, it is likely the case that across a larger sample of possible venues, cohesion would have been significant.

Given that 87% of our topics were more cohesive than we would expect by chance along these metrics, this high-level view of the results suggests that reducing the dimensionality of the feature set used to cluster users is an interesting avenue for future work. However, beyond these high-level findings, our metrics present interesting insights into the roles of function, time and geo-spatial location in the topics that resulted from our use of LDA on the venue IDs. In the sections below, we give more detail and insight into the levels of cohesion across each of the metrics independently, and then explore correlations across metrics.

5.1. Cohesion in High-level and Low-level Function



(a) (b)

Fig. 3. a) A gaussian kernel density estimate for the entropy distribution of the real and randomized topics for high-level function, and b) The high-level function distribution for the five topics lowest in entropy. The topic number is indicated in the grey bar at the top of each plot

Figure 3a shows the distribution of entropy scores for the high-level function metric for both the real and randomized topics. As is clear, the distribution of cohesion for the real topics was skewed towards higher levels of cohesion (lower entropy). However, Figure 3a also suggests that there were few topics that were highly cohesive. Figure 3b shows the high-level function distribution for the five topics having the lowest entropy in high-level function. As is clear, our model uncovers topics which are distinct in their

functionality - topic 27 is a collection of nightlife spots, while topic 22 was a selection of venues frequented by those traveling by plane and staying at hotels in the city.

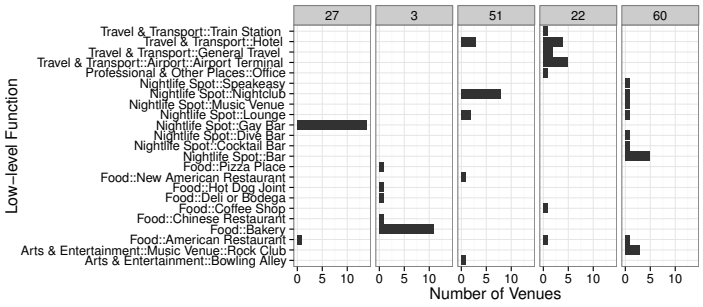


Fig. 4. Distribution of low-level function (place categories) for the five topics with the smallest low-level function entropy. The topic number is indicated in the grey bar at the top of each plot.

As one would imagine, though, it is often necessary to use other measures to differentiate topics beyond their distribution of high-level functions. For example, both topic 3 and topic 9 heavily revolve around food. We can differentiate between these two quite easily by looking at their low-level function distributions. Here, we find that topic 3 is a collection of restaurants (mostly bakeries). One can see this is the case in Figure 4, which shows the distribution of low-level function for the five topics lowest in entropy. In contrast, topic 9 is a collection of mostly Chinese, Korean and Japanese restaurants. Interestingly, while topic 9 is significantly cohesive in high-level function, it is not so at the lower level. Such a finding points to the practical usage of a hierarchical measure of function, such as the one employed by foursquare, in that we can understand cohesiveness at varying levels of complexity, shaping interesting collections that in a flat topic structure may have been missed.

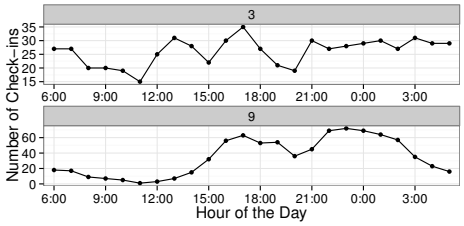


Fig. 5. Number of check-ins per hour of the day, aggregated over the entire dataset, for topics 3 and 9.

Another way we could have differentiated between topic 3 and topic 9 was via their temporal signatures - as Figure 5 shows, topic 3 appears to have been frequented during lunch and dinner times, whereas topic 9 appears to have been frequented more at dinner and late at night. However, in making such a comparison, we ignore the fact that the temporal cohesion of topic 3 is not significantly different than random, and thus that the aggregate statistic presented in Figure 5 does not necessarily represent the temporal structure of each (or any) of the venues individually. While aggregate claims are still certainly of value, the study of these two topics suggests that there are often ways to describe collections of venues frequented by similar users in ways that

are significantly cohesive. Here, in the case of topic 3, it is low-level function, and in the case of topic 9, time. Where possible, we thus suggest that topic characteristics that are cohesive should be used to describe, compare and contrast topics, as our notion of cohesion provides a quantitative basis for using these characteristics to understand information about the individual components of topics as well as the aggregate.

Regardless of these claims, it is clear that lower-level functionality gives us a much more nuanced view of the topics found to be homogeneous in function by the model. In particular, by far the most homogeneous collection of venues in low-level function was topic 27, a collection of almost entirely "gay bars" in two of the more prevalent homosexual areas in New York City. A similar topic, discovered in our previous work in both New York City and San Francisco foursquare data [Joseph et al. 2012], suggests that members of some underlying gay community in these cities tended to frequent these same locations. We explore in greater detail the extent to which we can define the users associated with this topic to be members of a sub-community within New York City in Section 6.

5.2. Geo-spatial Cohesion

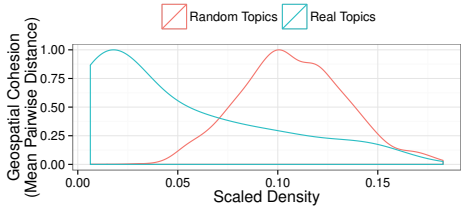
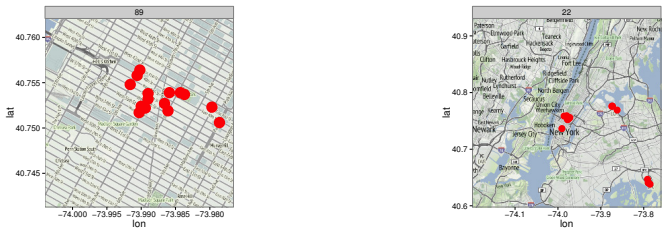


Fig. 6. A Gaussian kernel density estimate for the mean pairwise geo-spatial distance distribution of the real and randomized topics

Figure 6 displays the distribution of geo-spatial distance cohesion for the real and randomized topic sets and gives further evidence that distance cohesion is far higher in the topics discovered by LDA than one would expect by chance. As such, it is quite clear that our model finds that users tend to travel within reasonably small spaces of the possible distances they could travel. Figure 7a shows the topic with the highest spatial cohesion (topic 89) is concentrated most heavily in the area between 8th Avenue and 5th avenue, between 34th and 42nd street. This area can be loosely defined as the Garment District of New York, known to be one of the centers of the fashion industry.

While Figure 6 is validated by (and validates) many previous studies suggesting users tend to stay within small geographic spaces, it is interesting to note that there are cases in which topics contained venues that were much more spread geographically than we would expect by chance. For example, Figure 7b shows the second most spread topic, topic 22, where collections of points at the two prominent New York airports cause high spatial dispersion. Having discussed topic 22 in Section 5.1, we are aware that it appears to define a collection of people associated with the purpose of visiting New York. This finding, paralleled by previous work in [Joseph et al. 2012], suggests that when defining factions of people within a city, it is necessary to consider context beyond their movements in space to the actual venues they are moving to. These venues give important information, like the consideration of circadian rhythms gave in work by Gonzlez et al. 2008 on individual mobility patterns, that help to better understand the root cause of different individual mobility patterns within the city.



(a) (b)

Fig. 7. a) The spatial distribution of the topic with the lowest mean pairwise geo-spatial distance, and b) the topic with the second highest value on this metric. Each red dot indicates one of the top 15 venues associated with that topic

Thus, while limitations of foursquare data exist in the form of a sort of "filter" through which we get to see the location of a user, the observations made in this section give clear evidence that the context provided within foursquare data gives a richer understanding of place than data on location alone, as is often the case in location data from mobile phone calls and SMS. In contrast, data from mobile phones is often provided with few limitations on the actual location of a person - Dimmick et al. 2011 find in particular that SMS is regularly used by people at all times of the day and in nearly all locations. Thus, our results suggest that future work in understanding movement within the city would do well to combine data with context, like foursquare check-ins, which more concrete knowledge of physical location, such as data from mobile phones.

5.3. Temporal Cohesion

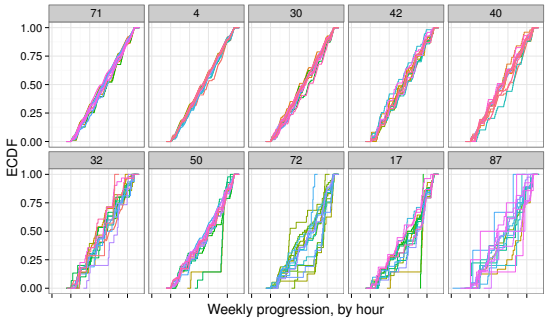


Fig. 8. Plots of the temporal patterns for the five factors highest (top row) and lowest (bottom row) in our temporal metric. On the x-axis, we plot time over the course of one week. The y-axis represents the ECDF, as discussed above. Different topics are labeled by the grey bar above each plot- within each plot, each colored line represents a different venue

Figure 8 shows the ECDFs for each venue for the topics with the top five most cohesive temporal distributions (top row) and the five with the least cohesive temporal distributions (bottom row). Within a specific topic, each colored line represents a different venue, and the x-axis of each plot covers the span of a single week. The visible differences between the top and bottom rows of graphs suggest qualitative evidence that the metric we provide to understand distinctions in temporal differences of venues within each topic is reliable. However, it is important to note that there appears to be a heavy reliance of the metric on the number of check-ins per venue. While this makes

some sense, given that the ECDF of a venue becomes smoother with greater numbers of check-ins, we find that the metric does still present interesting distinctions between topics with similar amounts of check-ins.

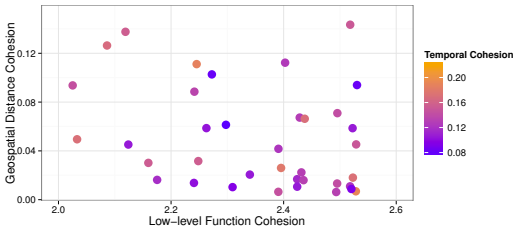
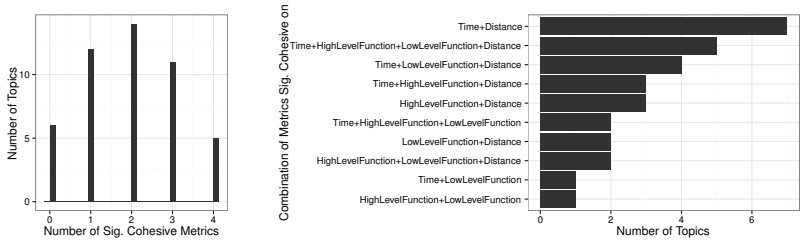


Fig. 9. A plot of the relation between low-level function (x-axis), geo-spatial cohesion (y-axis) and temporal cohesion (point color, as indicated by the scale on the right)

An exploration of the topics highest in temporal cohesion reveals that they tended to be more highly correlated in function and more highly dispersed in distance than the topics lowest in temporal cohesion. Figure 9 plots each topic as a point, where the y-axis is geo-spatial distance cohesion, the x-axis is low-level function cohesion, and the color of the point represents the temporal cohesion. Though we do see a slight tendency for venues lower in temporal cohesion (more blue in color) to have higher levels of low-level function entropy, a linear regression reveals that geo-spatial distance cohesion nor low-level function cohesion are significant predictors of temporal cohesion. Thus, we find that our data gives some qualitative evidence for but does not support the general hypothesis that topics that are highly temporally cohesive will be cohesive in function, nor that those cohesive in distance will be less cohesive in time. We explore this question in more detail in the following section.

5.4. Correlations



(a) (b)

Fig. 10. a) A histogram of the number of topics that were more cohesive than expected by chance on 0,1,2,3 or 4 metrics, and b) The number of topics that were cohesive on different combinations of metrics

Figure 10a shows the distribution of the number of metrics along which the topics were significantly more cohesive than we would expect by chance. As the plot suggests, when a venue was significantly cohesive in at least one metric, there was a 67.5% chance that it would be cohesive in multiple metrics. Such a finding supports long-standing claims of the intricate correlations between time, space and function that have been suggested in previous works in this area (e.g. [Cho et al. 2011; Cranshaw et al. 2012]).

Figure 10b shows the number of topics that were significantly cohesive across the given subset of metrics. We see that the most likely combination of cohesive factors are time and distance. This finding qualitatively supports the claims of Yuan et al. 2012, where different sections of the city are shown to be utilized at different times of the day. However, it at the same time serves to confound our qualitative claims in the previous section, and our belief that neighborhoods, which are also highly cohesive in space, would be less cohesive in time than one would expect by chance.

Table II. Spearman correlations between the different metrics, with significance determined by permutation test with N=1000

	High-level Function	Low-level Function	Geo-spatial Distance
Low-level Function	0.48[a]		
Geo-spatial Distance	0.19	-0.28[b]	
Temporal Distribution	0.08	0.05	0.18

[a] p <.0001
[b] p <.1

While Figure 10b indicates several correlations between metrics may exist, it is of course necessary to test such correlations empirically. We use Spearman rank correlation and compute significance usinga permutation test with 1000 iterations. Table II shows, unsurprisingly, that high-level function and low-level function are highly positively correlated- as entropy in one increases, entropy in the other increases as well. Interestingly, however, we find that there is a significant, negative correlation between cohesiveness in place and cohesiveness in low-level functionality. This finding enriches the claims made by Cranshaw et al. 2012 by showing that even with no distance-based constraints on the model, we discover topics tight in space that seem to serve a diverse set of functions.

We thus find an interesting differentiation between functional region and neighborhood, both of which are expected to be cohesive in space. Where functional regions were shown in [Yuan et al. 2012] to be of a specific function and be frequented at distinct times of day, our work and suggestions from Cranshaw et al. 2012 suggest that neighborhoods sohuld be diverse in function and, we would hypothesize, diverse in the temporal make-up of venues as well.In showing that many topics cohesive in time are cohesive in space, and also giving a significant negative correlation between space and function, our model thus suggests that both of these types of areas cohesive in space may exist in our model. Though we do not pursue this question in detail, this distinction between neighborhood and functional region is an interesting avenue to pursue in future work.

5.5. Topics Lacking Significant Cohesion

Several of the topics that lack cohesion on any metric appear to be lacking in such due to the venues within them defining neighborhoods in the towns and small cities surrounding the City of New York. Figure 11, which displays the geo-spatial distribution of four of the six topics that were not significantly cohesive along any metric, shows that, while noisy, these topics center reasonably well around regions not in Manhattan that appear to define cohesive areas in space. Given our finding in Section 5.4, we would expect that neighborhoods have venues less cohesive in function. Thus, we can explain the lack of cohesion in function across these places by noting that these topics seem to represent, for the most part, communities, which are likely to have low levels of functional cohesion. In turn, due to the fact that amenities are not in as tightly packed an area as they are within the confines of New York City, it would make sense

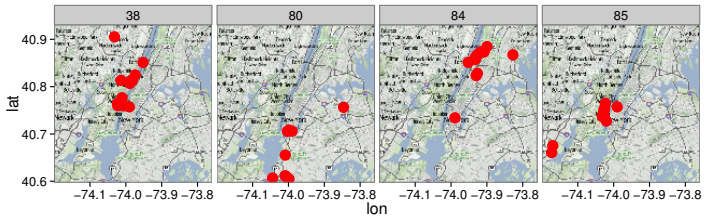


Fig. 11. The geo-spatial distribution of four of the six topics that were not significantly cohesion along any of the metrics used

that people must travel farther to reach them, hence the larger geo-spatial spread than we would expect by chance.

This finding provides two interesting conclusions with respect to the usefulness of our approach and the general usefulness of a data-driven approach to understanding community. First, the fact that our model does not concretely hinge on distances between places allows it to understand how community structure may extend beyond a priori beliefs of the distances across which a community exists. Second, this finding gives interesting evidence of the ability of location-based data to capture the distinction between suburban and urban life, in that those in the suburbs must be willing and able to travel larger distances. As such, one use of this data in the future might be to distinguish suburban or urban areas which are particularly well or poorly planned based on the expected radius of topics in the two settings.

6. DISCUSSION

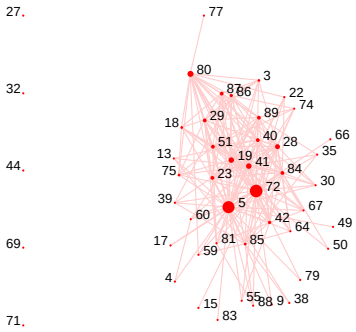


Fig. 12. A network of the latent topics, where links were created between topics when users were associated with both. The topic number is next to the node representing the topic - the size of a node is a function of the number of incoming links, and only links with weight greater than three are visualized.

Figure 12 shows the topics utilized in our analysis organized into a network diagram, visually laid out using multi-dimensional scaling [Borg and Groenen 2005]. Topics are represented by dots and associated topic numbers - the size of the dot represents the number of links to that topic. Recalling that in LDA a user is defined by a probability

distribution detailing their association with one or more topics, a link was created between two topics when the same user had more than 15% of their probability mass associated with each. Thus, for example, if user 3 was associated with topics 1,2,3 and 4 with a magnitude of .25 each, links would be formed between each pair of topics. We select 15% a priori, due to our belief that no user would likely be associated with more than 6 different topics, but find that this selection does not heavily influence our results. The weight of a link between two topics is simply the number of users that shared the two topics.

Figure 12 shows all links with weight greater than three. The figure provides a series of interesting qualitative conclusions which serve both to further the claims made in Section 5 and to generate interesting questions for future study - while there are several points of interest, we focus, due to space constraints, only on two here. First, one can readily see that topic 27 (the topic heavily associated with "gay bars") is isolated in the network. In addition, topic 77, which contained venues associated with New York's "hipster" crowd, had only a single link[4].

Figure 4b and Figure 12 thus combine to suggest that users in the "gay bar" and "hipster" clusters almost exclusively associated themselves with venues aligned only with these notions of self[5]. A natural question then becomes what to make of the fact that users affected by these two topics are segregated from almost all other users in their check-in patterns. We here suggest that this finding, interpreted through the lens of self-representation, implies that these two topics represent distinct and important "micro-cultures" within New York City.

In order to make this argument, we reference the work of Cooper and Denner 1998, who, quoting Ethier and Deaux 1994, argue "[an individual's desire to express a cultural identity] depends on the competing needs for inclusiveness and uniqueness". Thus, our claim of these two topics representing distinct cultures within New York becomes an argument of the extent to which we would expect these two cultures to derive a stronger notion of "uniqueness" that other communities our model discovered. As is widely assumed, one of the general goals of the "hipster" movement was (and is) to create and portray a *unique identity* [Alfrey 2012]. Similarly, according to the notions put forth by Affect Control Theory, homosexuals, as a discriminated minority community, are more likely to identify strongly with their distinct and minority culture than many other social groups [Smith-Lovin and Douglas 1992].

Via the claim that these two communities perceive themselves to be a part of a unique identity, and that foursquare allows users to present themselves as the person they want to be, we argue that our model seems to observe distinct subsections of culture within the city of New York. Perhaps most interestingly, these populations are not necessarily confined to certain sections of the city - the locations they exist in span various neighborhoods (or, by visual inspection, more appropriately "livehoods"). Though the empirical confirmation of these claims is outside the scope of this work, it is clear that a greater understanding of culture and community within the city using location-based data, and how such an understanding relates to long-standing social theory, is thus a very interesting avenue of future work.

The observations above indicate that particular sub-communities within the city seem to mingle in limited ways with other types of people within the city. This presents

[4]In order to verify beyond personal knowledge the extent to which topic 77 qualified as a "hipster" topic, we use Yelp (http://www.yelp.com), a crowd-sourced review website, to observe the number of venues which had received reviews having the word "hipster" in them. We find that ten out of the fifteen venues had reviews referring specifically to the venue being home to a hipster crowd.

[5]Note that we select these two topics because they are identifiable from the venues within these two topics - other communities may have been discovered by our model, however, we were unable to interpret such communities given the data we had

a case for the large levels of segregation and modularity known to exist across social groups (e.g. [Hipp et al. 2012]). One interesting point that can be considered by looking at the network of topics above, however, is where one might be able to find different types of people mingling in the city, and thus where opportunities might exist to lower barriers of segregation in the larger city community. Figure 12 shows that two topics, topic 5 and topic 72, appear to be most often associated with users who were affected by other topics as well. As one may have guessed, one of the topics (topic 5) is heavily associated with sport stadiums -in particular, Yankee Stadium, Arthur Ashe Stadium, IZOD Center and Citi Field (home to the New York Yankees, the US Open, the New Jersey Nets and Devils and the New York Mets, respectively). In many ways, this fits our intuition - athletic events tend to yield large, diverse crowds, suggesting the chance for intermingling of populations, and occur with relative infrequency, suggesting that users would tend to spend most of their time affected by other check-in factors. Similarly, topic 72 is represented by several places on Coney Island, well-known for its amusement park (and hot dogs) and the Bronx Zoo. While these popular places fit our intuition as to where people of diverse interests may meet, other venues strongly associated within these topics, specifically certain restaurants and bars in Manhattan, provide less obvious places at which city planners might encourage interaction between people who might otherwise share little common ground.

7. CONCLUSION

The study of urban environments using large quantities of location-based data presents an unprecedented opportunity to understand how people move, behave and interact within the city. In the present work, we further research in this area by testing the hypothesis that locations within New York City frequented by similar patrons are cohesive in space, time and function. We also consider the correlation between these factors, and how we can utilize the clusters resulting from our model to better understand community within an urban environment. Our main contributions can be summarized as follows:

— 87% of venue topics our model discovers, on a feature set agnostic of anything but unique venue IDs, are more cohesive than chance in time, space, and/or function.
— A negative correlation, significant at $p<.1$, exists between the cohesion of venue functionality and cohesion in distance, suggesting a "neighborhood" effect in the topics we study
— Analysis of the network of topics, as connected by similar users, suggests the existence of isolated gay and "hipster" subcommunities within New York City, but also locations at which city planners may encourage interactions between people of mixed interests

While we believe our contributions significant to the fields of both urban computing and sociology, there are certain limitations to the work we present which should be addressed in future work. First, though we show that our temporal metric generates output that fits with general intuitions, it is not immediately clear that the metric performs well where it must deal with venues having few check-ins, as it is highly correlated with the number of check-ins within a topic. Thus, future work would benefit from a more powerful non-parametric approach to comparing venue temporal signatures with small amounts of data.

Second, although we justify the model selected, LDA is one of the simplest topic models. While we find that a more complex model, the Correlated Topic Model [Blei and Lafferty 2007], does not perform as well as LDA on the data at hand, work done here may benefit from the utilization of a more complex framework. In particular, hierarchical models, such as Hierarchical LDA [Blei et al. 2004] or Hierarchical PAM [Li

et al. 2012], may help to better understand meso-level associations between collections of people and venues within the city. DMR-based topic models [Mimno and McCallum 2012], like the one used in [Yuan et al. 2012], may be used to incorporate into the model cohesion in time, space and function, as opposed to testing for cohesion along these dimensions in a feature set agnostic of these values. Such a model is particularly interesting in that Dirichlet Multinomial Regression can be considered a direct extension of McFadden's discrete choice model [Guimaraes and Lindrooth 2005]. Thus, such a model would also present an interesting tool to use in seeding multi-agent simulations of the movement of people within a city. Finally, a dynamic model, applied on a more continuous dataset than the one utilized here, may be able to differentiate between "bursty", periodic and continuous topics [Yin et al. 2011].

Beyond addressing these limitations, plenty of work still exists in understanding community and culture utilizing location-based data from services such as foursquare. For instance, given the association of hashtags with community on Twitter [Yang et al. 2012], it may be interesting to understand the extend to which community in place is similar to community online. Regardless of the avenue of research pursued, it is evident that the influx of "big data", particularly with respect to location-based data, have generated a significant number of new opportunities to understand human behavior in the urban environment. While we urge computational researchers to draw on the interesting and relevant work from centuries of study on the urban environment via more traditional data, we believe that the works we have cited here, and the efforts we have provided, are only the beginning of how new data sources can be used to improve what we know about the places so many of us live.

ACKNOWLEDGMENTS

We would like to thank Chun How Tan for his work on the previous version of this article. In addition, we would like to acknowledge the assistance, in no particular order, of Justin Cranshaw, Ju-sung Lee, Michael Martin, Geoffrey Morgan and Jonathan Morgan in their assistance in formalizing the ideas which were presented in this work.

REFERENCES

ALFREY, L. 2012. The search for authenticity: how hipsters transformed from a local subculture to a global consumption collective.

BAUER, S., NOULAS, A., SAGHDHA, D. O., CLARK, S., AND MASCOLO, C. 2012. Talking places: Modelling and analysing linguistic content in foursquare.

BECKER, R., CACERES, R., HANSON, K., LOH, J., URBANEK, S., VARSHAVSKY, A., AND VOLINSKY, C. 2011. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing 10,* 4, 18 –26.

BLEI, D. M., GRIFFITHS, T. L., JORDAN, M. I., AND TENENBAUM, J. B. 2004. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*. MIT Press, 2003.

BLEI, D. M. AND LAFFERTY, J. D. 2007. A correlated topic model of science. *The Annals of Applied Statistics*, 17–35.

BLEI, D. M., NG, A. Y., AND JORDAN, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res. 3*, 993–1022.

BORG, I. AND GROENEN, P. J. F. 2005. *Modern multidimensional scaling: Theory and applications*. Springer.

BROCKMANN, D., HUFNAGEL, L., AND GEISEL, T. 2006. The scaling laws of human travel. *Nature 439,* 7075, 462–465.

CAMERRA, A., PALPANAS, T., SHIEH, J., AND KEOGH, E. 2010. iSAX 2.0: Indexing and mining one billion time series. In *2010 IEEE 10th International Conference on Data Mining (ICDM)*. 58 –67.

CHENG, Z., CAVERLEE, J., LEE, K., AND SUI, D. Z. 2011. Exploring millions of footprints in location sharing services. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. ICWSM '11. AAAI, 81–88.

CHO, E., MYERS, S. A., AND 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*. KDD '11. ACM, New York, NY, USA, 10821090.

COOPER, C. R. AND DENNER, J. 1998. Theories linking culture and psychology: Universal and community-specific processes. *Annual Review of Psychology 49,* 1, 559–584.

CRAMER, H., ROST, M., AND HOLMQUIST, L. E. 2011. Performing a check-in: emerging practices, norms and 'conflicts' in location-sharing using foursquare. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*. MobileHCI '11. ACM, New York, NY, USA, 5766.

CRANSHAW, J., SCHWARTZ, R., HONG, J. I., AND SADEH, N. 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. ICWSM '12. AAAI.

CRANSHAW, J. AND 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 2010*. NIPS '10. AAAI.

DIMMICK, J., FEASTER, J. C., AND RAMIREZ, A. 2011. The niches of interpersonal media: Relationships in time and space. *New Media & Society 13,* 8, 1265–1282.

ETHIER, K. A. AND DEAUX, K. 1994. Negotiating social identity when contexts change: Maintaining identification and responding to threat. *Journal of Personality and Social Psychology 67,* 2, 243.

FERRARI, L., ROSI, A., MAMEI, M., AND 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*. LBSN '11. ACM, New York, NY, USA, 916.

GONZLEZ, M. C., HIDALGO, C. A., AND BARABSI, A.-L. 2008. Understanding individual human mobility patterns. *Nature 453,* 7196, 779–782.

GUIMARAES, P. AND LINDROOTH, R. 2005. Dirichlet-multinomial regression. Econometrics 0509001, EconWPA. Sept.

HIPP, J. R., FARIS, R. W., AND BOESSEN, A. 2012. Measuring neighborhood: Constructing network neighborhoods. *Social Networks 34,* 1, 128 – 140.

JIANG, S., FERREIRA,JR., J., AND GONZALEZ, M. C. 2012. Discovering urban spatial-temporal structure from human activity patterns. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*. UrbComp '12. ACM, New York, NY, USA, 95102.

JOSEPH, K., TAN, C. H., AND CARLEY, K. M. 2012. Beyond "Local", "Categories" and "Friends": clustering foursquare users with latent "Topics". In *Proceedings of the 4th International Workshop on Location-Based Social Networks*. LBSN'12. Pittsburgh, PA.

KLING, F. AND POZDNOUKHOV, A. 2012. When a city tells a story: Urban topic analysis.

LI, W., BLEI, D., AND MCCALLUM, A. 2012. Nonparametric bayes pachinko allocation. *arXiv:1206.5270*.

LINDQVIST, J., CRANSHAW, J., WIESE, J., HONG, J., AND ZIMMERMAN, J. 2011. I'm the mayor of my house: examining why people use foursquare - a social-driven location sharing application. In *Proceedings of the 2011 annual conference on Human factors in computing systems*. CHI '11. ACM, New York, NY, USA, 24092418.

MASSEY, FRANK J., J. 1951. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association 46,* 253, pp. 68–78.

MCFADDEN, D. 1980. Econometric models for probabilistic choice among products. *Journal of Business*, 1329.

MCPHERSON, M., LOVIN, L., AND COOK, J. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 1, 415–444.

MIMNO, D. AND MCCALLUM, A. 2012. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *arXiv:1206.3278*.

NOULAS, A., SCELLATO, S., LAMBIOTTE, R., PONTIL, M., AND MASCOLO, C. 2011a. A tale of many cities: universal patterns in human urban mobility. *ArXiv e-prints*.

NOULAS, A., SCELLATO, S., MASCOLO, C., AND PONTIL, M. 2011b. An empirical study of geographic user activity patterns in foursquare. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. ICWSM '11. AAAI, 570573.

ONNELA, J.-P., SARAMKI, J., HYVNEN, J., SZAB, G., LAZER, D., KASKI, K., KERTSZ, J., AND BARABSI, A.-L. 2007. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences 104,* 18, 7332–7336.

RUMELHART, D. 1978. *Schemata: The building blocks of cognition*. Center for Human Information Processing, University of California, San Diego.

SADILEK, A., KAUTZ, H., AND BIGHAM, J. P. 2012. Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining*. WSDM '12. ACM, New York, NY, USA, 723732.

SCELLATO, S., NOULAS, A., AND 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*. KDD '11. ACM, New York, NY, USA, 10461054.

SMITH-LOVIN, L. AND DOUGLAS, W. 1992. An affect control analysis of two religious subcultures. *Social perspectives on emotion 1*, 217–47.

TANG, K. P., LIN, J., HONG, J. I., SIEWIOREK, D. P., AND SADEH, N. 2010. Rethinking location sharing: exploring the implications of social-driven vs. purpose-driven location sharing. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. Ubicomp '10. ACM, New York, NY, USA, 8594.

WALLACH, H., MIMNO, D., AND MCCALLUM, A. 2009a. Rethinking LDA: why priors matter. *Advances in Neural Information Processing Systems 22*, 1973–1981.

WALLACH, H. M., MURRAY, I., SALAKHUTDINOV, R., AND MIMNO, D. 2009b. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. ACM, New York, NY, USA, 11051112.

YANG, L., SUN, T., ZHANG, M., AND MEI, Q. 2012. We know what @you #tag: does the dual role affect hashtag adoption? In *Proceedings of the 21st international conference on World Wide Web*. WWW '12. ACM, New York, NY, USA, 261270.

YE, M., JANOWICZ, K., MLLIGANN, C., AND 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*. GIS '11. ACM, New York, NY, USA, 102–111.

YIN, Z., CAO, L., HAN, J., ZHAI, C., AND HUANG, T. 2011. LPTA: a probabilistic model for latent periodic topic analysis. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. 904 –913.

YUAN, J., ZHENG, Y., AND XIE, X. 2012. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '12. ACM, New York, NY, USA, 186194.

Received October 2012; revised ; accepted