Check-ins in “Blau space”: Applying Blau’s macrosociological theory to foursquare check-ins from New York City

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

Peter Blau was one of the first to define a latent social space and utilize it to provide concrete hypotheses. Blau defines social structure via social “parameters” (constraints). Actors that are closer together (more homogenous) in this social parameter space are more likely to interact. One of Blau’s most important hypotheses resulting from this work was that the consolidation of parameters could lead to isolated social groups. For example, the consolidation of race and income might lead to segregation. In the present work, we use foursquare data from New York city to explore evidence of homogeneity along certain social parameters and consolidation that breeds social isolation in communities of locations checked-in to by similar users.

More specifically, we first test the extent to which communities detected via Latent Dirichlet Allocation are homogenous across a set of four social constraints - racial homophily, income homophily, personal interest homophily and physical space. Using a bootstrapping approach, we find that fourteen (of twenty) communities are statistically, and all but one qualitatively, homogenous along one of these social constraints, showing the relevance of Blau’s latent space model in venue communities determined via user check-in behavior. We then consider the extent to which communities with consolidated parameters, those homogenous on more than one parameter, represent socially isolated populations. We find communities homogenous on multiple parameters, including a homosexual community and a “hipster” community, that show support for Blau’s hypothesis that consolidation breeds social isolation. We consider these results in the context of mediated communication, in particular in the context of self representation on social media.

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 social network, macrosociology

Additional Key Words and Phrases: foursquare, community structure, urban analytics, latent social space

ACM Reference Format:
Joseph, K., Carley, K.M., Hong, J.I. 2013. Check-ins in “Blau space”: Applying Blau’s macrosociological theory to foursquare check-ins from 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

With the advent of GPS on mobile phones and the increasing popularity of location-sharing services such as foursquare[1] and Facebook Places[2], scientists have in recent years been privy to massive quantities of information on the movements, actions and social structure of large human systems. These data are substantially more granular, more accurate and larger in scale than location data collected in the past via more traditional means, such as surveys.

Though the literature around these data has focused on a variety of problems, two general concepts are pervasive throughout. The first is that human interaction and travel are bound by space [Brockmann et al. 2006; Cheng et al. 2011]. The second is that homophily, the principle that “a contact between similar people occurs at a higher rate than among dissimilar people” [McPherson et al. 2001, p.416], is a strong driver of both where humans go and who they interact with. Models emphasizing homophily by placing human movement in the context of the actions of others having similar interests [Noulas et al. 2012b] or user’s social connections [Sadilek et al. 2012] have been increasingly successful in, for example, predicting where a given person will go next.

However, while successful in predicting human movement, this recent work has also suggested that the constraints of space and homophily on sociality are messy and multi-faceted. [Sadilek et al. 2012] find spatial constraints alone are insufficient to predict a person’s location and [Cho et al. 2011] observe that social structure can explain only a small percentage of user behavior on a service similar to foursquare. What these studies have found, but failed to state in the context of sociological theory, is that homophily is a “cross-cutting phenomena” [McPherson et al. 2001, p.418]. By cross-cutting, we refer to the fact that human interaction and movements are driven heavily by homophily, but that this homophily exists in a high-dimensional, unobservable social space.

This social space has often been referred to as “Blau space” [McPherson and Ranger-Moore 1991] after Peter Blau, one of the first to formulate such a notion [Blau 1974; 1977a; 1977b]. In contrast to homophily as a driving factor in sociality, Blau framed homophily as one family of factors within a larger set of constraints on social behavior which are the parameters of this latent space. Social groups tend to form, Blau argued, in collections of individuals that are homogenous on one or more of these social constraints and thus groups that occupy a similar position in the latent social space. This formalization is convenient, as it allows us to set homophily and other mechanisms influencing social behavior (like spatial position) into a consistent conceptual framework.

In previous work [Joseph et al. 2012], we utilized Latent Dirichlet Allocation (LDA) [Blei et al. 2003] to cluster locations (venues) in foursquare data from New York City and the Bay Area[3]. We viewed the resulting clusters of venues as representations of social communities and provided a qualitative explanation of the results. Here, we concern ourselves only with data from New York City. New York is of particular interest due to the magnitude of the data we have for the city, its ethnic diversity and its high population density, all of which promote the possibility of interesting sociological phenomenon.

Our chief observations in this previous work were two-fold. First, we found that the communities we discovered appeared homogenous in at least one observable manner.

[1] http://www.foursquare.com

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

[3] foursquare is a location-based social network where users can “check-in” to different locations and broadcast these check-ins to their friends on social networking sites

These communities could thus, we believed, be rationalized within Blau's notion of a high-dimensional social space. Second, and more interestingly, we observed that certain communities appeared to be homogenous along multiple social dimensions (e.g. space and function). This finding also falls into Blau's work. Blau used his conceptualization of latent social space to argue that 'consolidated [social constraints] impede intergroup relations" [Blau 1977b, p.45]. In other words, Blau argued that in addition to aiding in the creation of social relations, homophily, when correlated along multiple social constraints within communities, can also serve to limit social interactions across groups. From this, we can infer that a community homogenous along multiple constraints may exist in a narrow domain of the latent social space and, consequently, exist as an isolated social group.

In the present work, we extend previous efforts in two ways[4]. We first consider more rigorously the quality of the communities that LDA provides for sociological study under different parameterizations. While many human social structures are dynamic, the macrosociological processes of interest to Blau were presumed to be relatively stable. We thus desire a model that provides a stable estimate of community over time. In order to find a parameterization of LDA that provides this, we first split the data into two equally-sized partitions. Across a wide parameter space, we run LDA on both datasets and calculate the Normalized Mutual Information (NMI) [Lancichinetti et al. 2009] of the resulting clusterings. We find that there exists certain parameterizations of LDA in which a level of stable community structure arises.

Using the best model parameterization, we then run LDA on the full data and provide a more quantitatively rigorous analysis of the resulting communities. Our focus is on expanding our two observations in the previous work within Blau's framework. The latent space we consider centers on four well-validated social constraints- racial homophily, income-based homophily, geographic constraints and personal interest homophily. These social constraints are not used during venue clustering and therefore can then be viewed as latent factors influencing the creation of the communities LDA uncovers. Assuming these constraints are pertinent to the unobserved social space and that the communities we observe follow Blau's conceptualization, the following hypothesis should hold:

H1: *The communities discovered will be more homogenous than one would expect by chance on at least one of the proposed social constraints*

Using this same framework, we can also test whether or not Blau's hypothesis that consolidation of parameters breeds social isolation extends to our observed communities. This leads to the following hypothesis:

H2: *Communities homogenous on more than one of the proposed social constraints will represent socially isolated groups*

In order to test our two hypotheses, we first develop statistics to measure the level of homogeneity in each community along each constraint. We then construct a null sampling distribution for each statistic on each community via the bootstrap method [Efron and Tibshirani 1993]. Our hypotheses can then be tested by determining whether or not our social constraints were more homogenous in the observed communities than one would expect given the null distribution.

Our results indicate partial support for both hypotheses, suggesting that Blau's theoretical model seems to hold in new data source and a relatively new conceptualization of the urban community and thus its importance in future work on understanding social integration via urban computing. With respect to H1, all but one of the twenty communities discovered qualitatively matches the expected outcome. Quantitatively,

[4] Code used for analysis and results of our parameter experiments are available at https://github.com/kennyjoseph/tist_article

six communities are not significantly different than expected by chance at $\alpha = .05$, indicating a variety of possible confounds. In particular, we discuss how self representation, the fact that users only check-in to places where they want to be seen, may have affected results. With respect to H2, we find three communities in which more than one constraint is significantly more homogenous than chance (at $\alpha = .05$). In two of the three cases, the communities discovered represent populations within the city that are often marginalized and segregated, while in the third, results are inconclusive.

In the following sections, we first briefly discuss how communities have been defined and studied within the city in recent work. We then provide a brief introduction to LDA and the location-based social networks (LBSNs) literature. Following this, we describe our methodology, in particular our method of parameter selection for LDA and details on the algorithms we use to bootstrap our analysis. We then present our results, a short discussion and end with some remarks on limitations and future work.

2. RELATED WORK

2.1. Blau and Urban Sociology

Blau's work has been applied to the urban setting chiefly with respect to the topic of racial segregation [Massey and Denton 1988], where his notion of social groups has been aligned with the neighborhood as defined by government boundaries. Decades of research have shown that these neighborhoods represent segments of a city's population that are homogenous beyond spatial positioning- for example, individuals in the same neighborhood often have similar socio-demographic characteristics and socio-economic status [Schelling 1971].

Recently, however, the notion of a spatially, culturally homogenous neighborhood defined by government boundaries has been questioned by scholars. [Hipp et al. 2012] construct neighborhoods bound by social ties in addition to geographic constraints, arguing that "the presence of social ties is a characteristic of neighborhoods" (p. 128). [Cranshaw et al. 2012] show that by clustering a geospatially constrained graph of venues checked-in to by foursquare users, one finds neighborhoods that rarely conform to government-induced boundaries. Finally, though not yet a reality, the increasing level to which we live "in media" [Deuze 2012] suggests an even further breakdown of the spatial component of a neighborhood may be eminent.

In the present work, we take the first approach we are aware of to assume a purely social, data-driven view of the community within the urban setting with the intention of analyzing the relationship between social constraints and social isolation. We assume that the only defining characteristic of a community is that it is a collection of places within a city frequented by the same people. Our focus is thus on how Blau's predictions extend to such communities.

2.2. Latent Dirichlet Allocation

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 could 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 modeled as a set of words, where each word is assumed to belong to one or more topics. Model output describes each document as a distribution over the topics and each topic is described as a distribution over all words. We can use the analogies of a document to represent a user and a word to represent a venue when modeling user check-in behaviors with LDA. In this way, a user who has checked in to Yankee Stadium twice and the local Pizza shop four times is modeled as the term vector $< ...0, 0, 2, 0, 4, 0... >$, where zeros and ellipses show that in LDA, all venues (words) in the data are used to define each user (document)[5].

While this word to document metaphor provides a nice connection to the original application area, from an unsupervised clustering perspective LDA is simply a methodology for co-clustering users and venues in a k dimensional space, where k is the number of clusters as set by the researcher. However, because we have taken it out of its initially intended purpose, using LDA on the present data comes with several assumptions. In particular, our use of LDA induces the assumption that the order of check-ins is irrelevant. Because we are interested in social constraints expected to influence the creation of long-standing communities, we feel that such a model is appropriate. Additionally, LDA presumes no strict correlations between the different latent topics. Later models, including the Correlated Topic Model (CTM) [Blei and Lafferty 2007] and more recently hierarchical PAM [Li et al. 2012] provide different assumptions that might serve to improve the clustering of our venues by explicitly assuming correlations exist between topics. In the present work, we chose LDA to allow ourselves the ease of interpretation it presents and because our previous results suggested that it provides clusters of sociological interest.

2.3. Location Based Social Networks

Formally, foursquare is a socially-driven location sharing application [Lindqvist et al. 2011]. Services that fit this description are often referred to as Location Based Social Networks [Zheng 2011]. On foursquare, 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 on other social networking sites. Here, we review qualitative research on why people check-in, statistical models of human mobility and sociality within the city using LBSN data and other works clustering venues to obtain representations of communities. We detail how such work helps to both understand and motivate our efforts.

2.3.1. Why Users Check-in. The data we, and indeed nearly all other works on foursquare use, are a subset of all foursquare check-ins posted to the public Twitter feed. Check-ins to Twitter are a specific and likely biased subsample of all foursquare check-ins, as users are not required to share check-ins with their Twitter followers. [Lindqvist et al. 2011] found that only 18% of the users they surveyed allowed check-ins to be posted to Twitter, though [Cramer et al. 2011] found, less than a year later, that 68% of users surveyed had 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 suggests that people only check-in to places that depict them in a manner they desired to be viewed. Consequently, these venues may not truly portray users’ actual interests nor 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 restaurants). The effects of self representation have also been observed by [Tang et al. 2010], who draw on social psychology literature to discuss the cognitive groundings for this effect.

[5] 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.

Self representation affects our understanding of foursquare data in two important ways. First, it presents an interesting and complicating factor in applying the notion of homophily because 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 users are actually members of a community or whether they merely want to be perceived as a member.

2.3.2. Statistical Modeling of LBSN data. The most pervasive use of LBSN data has been to study how human movement and interaction are constrained in space. 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 [Noulas et al. 2011]. More recently, it has been suggested that models of human movement are better explained by taking into account the population density of a given area in addition to spatial constraints [González et al. 2008; Cheng et al. 2011]. These concepts will be important when we develop our constrained samples for the null distributions of our statistics.

Specific to foursquare data, [Bauer et al. 2012] use a 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-in. 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] 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 previous uses of LDA, though applying different assumptions and asking different questions, provide some precedence for its use here. However, an important point these works do not address is the extent to which venues communities uncovered by LDA are homogenous along the dimensions studied. While aggregate data is interesting when compared across clusters, we focus on within-cluster variance and the applications that it presents.

Though questions of spatial and spatio-temporal patterns dominate the literature, recent work has also begun to focus on sociality, with an emphasis on the effects of homophily. Early work using location information to model homophily between individuals included efforts by [Li et al. 2008]. Since this initial work, it has been shown that the location of a user can be predicted to some extent based on the location of their friends on social networking sites [Sadilek et al. 2012; Scellato et al. 2011]. [Cho et al. 2011] uncover correlations between human geographic movement and the social structure of the population they study, also considering the temporal constraints on where people move throughout the day. In addition to works providing insight into human behavior, research in clustering users has also led to methods for providing recommendations to users based on methodologies similar to those used here [Bao et al. 2012].

Additional work has begun to move beyond spatio-social data as well, including, for example, textual information from user posts [Vasconcelos et al. 2012]. In a related vein, [Noulas et al. 2012b] show that categorical features of venues on foursquare are a reasonable predictor of human movement. Indeed, their work, which compares how a variety of constraint-based predictors can be used to predict human movement is highly relevant to our work here by showing that spatial, categorical and social factors all play a role in human movement. Though it is well known that homophily induces correlations in human behavior, this work, as well as the others discussed here, makes clear that these tendencies are observable in LBSN data.

2.3.3. Modeling Collections of Venues. Several recent works have utilized foursquare to infer communities via the clustering of venues. [Cranshaw et al. 2012] use spec-

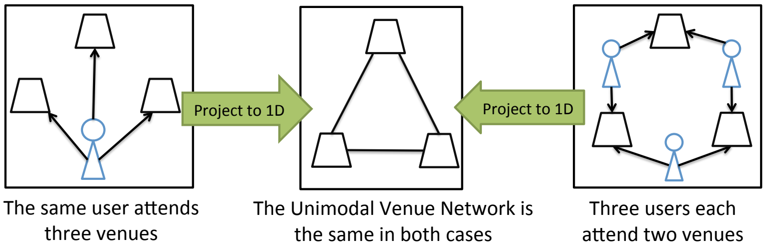
Project to 1D
Project to 1D
The same user attends three venues
The Unimodal Venue Network is the same in both cases
Three users each attend two venues

Fig. 1. An example of how information is lost when transforming a bimodal network of users to venues (where links are check-ins) to a unimodal network of venues to venues (where links are some function of shared user base)

tral clustering to understand how foursquare data gives insight into the dynamics of neighborhood boundaries in urban areas. [Lawlor et al. 2012] cluster venues using the OSLOM algorithm, a new methodology for finding statistically significant subsets of the data. The authors find that “different cities have varying degrees of stratification”, but that “[f]urther research is required to understand if cultural, racial or socio-economic segregation can be the reason for this” (p. 6).

Both of these works use a unimodal representation of the network of venues, created by computing some measure of the shared user base between each pair of venues connected in the underlying network representation. While a practical methodology when considering a sparse, geospatially regulated venue network, it is well known that the projection of a bimodal network (here, of users and venues, with links representing check-ins) onto a unimodal network (of venues to venues) loses a significant amount of information [Barber 2007].

Consider, for example, Figure 1, where we see on the left three venues in the bimodal user-venue network that have been checked in to by the same user. In the figure, the blue figures represent people and the trapezoids represent venues. On the right, we see three venues which are only connected in a “six-path” [Opsahl and Panzarasa 2009]; that is, the three venues are connected by three users, each of whom attended two of the same venues. Though these two bimodal networks depict different ways in which the three venues are connected, the unimodal projection of these two bipartite networks using cosine similarity results in an identical network of venues. In the present work, we consequently find it desirable to use a methodology that clusters the bipartite network explicitly in order to retain the information in the underlying data.

A variety of algorithms exist to cluster bipartite networks [Fortunato 2010]. However, the data we use presents a variety of restrictions that lead us to the utilization of a Bayesian latent variable model, namely LDA. First, many of the bipartite clustering algorithms place each element into a single community - in the case of foursquare venues, it is more likely that certain places, like the movie theatre or the grocery store, may be central to multiple communities. Second, algorithms which allow nodes to exist in multiple clusters generally focus on the concept of a “bi-clique” - the natural extension of a graph clique into bipartite networks. Through exploration of bi-cliques in our data using the BCFinder application [Lehmann et al. 2008], we find a large number of very small bi-cliques that provide little information of relevance to our hypotheses. This is due to the existence of venues with extremely high degree (i.e. stadiums). Using an asymmetric prior over the topics, LDA is able to overcome these uneven distributions over the network [Wallach et al. 2009a].

Finally, we do not expect all venues to be strongly associated with a community - some venues, particularly those with lower levels of check-ins, may simply be unimportant to any existing social community. We therefore desire a clustering algorithm that is able to filter out such uninteresting nodes from the model. While OSLOM, the method used by [Lawlor et al. 2012], fits this conceptualization, it has not yet been extended to bipartite networks. Though this is an interesting avenue of future work, LDA provides a second, equally applicable clustering framework that fits the requirements specified above.

3. METHODOLOGY

In this section, we first describe how we obtain check-in data from foursquare through Twitter, information on venue location and category from foursquare and information on the racial and income distributions that we will use for each venue from the US census. We then describe how we use LDA to cluster venues into communities and how we test for the best parameterization of LDA. Following this, we detail four statistics, one for each social constraint utilized, that define the extent to which each community is homogenous with respect to that constraint. Finally, we detail an algorithm used to bootstrap random clusters, which we can use to see how much more homogenous each cluster our model finds is on each constraint than we would expect by chance.

3.1. Data Collection

The data we use is a set of approximately 450,000 check-ins posted to Twitter from around 35,000 users of foursquare located in New York City. These data are part of a larger dataset given to us by the authors of [Cranshaw et al. 2012], formed by pulling foursquare check-ins visible to the public Twitter stream. No attempt was made to uncover social connections between users checking in. In total, the data comprises approximately 18 months of foursquare check-ins.

Each check-in provides a unique user ID from Twitter, the time-stamp of the check-in and the ID of the venue at which the check-in occurred. Using this venue’s ID, the original data collectors also obtained the venue’s name, geo-location, and “category” information by querying the foursquare API[6]. 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.

In addition to this information provided by the original data providers, we also collect racial, income and population size information for each venue. These values, described in more detail below, represent an estimate of these three quantities for the residents living in the same census block as the venue. This information is drawn from the 2011 American Community Survey (ACS) 5-year sample[7]. The ACS is a yearly collection of census-type data that provides information down to the level of a census

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

[7] http://www.census.gov/acs/www/

Table I. Parameters and tested values for model specification for LDA

k	20,40,80
$MinUser$	1,4
$MinVenue$	1,4,9
a	0.6,20,50,100
$MinWeight$	0.005,0.01,0.1
$BinarizeData$	Yes, No

block (between five and ten city blocks within Manhattan, and roughly a square mile in less populated areas of our dataset). We note that although nearly all venues in the communities we study (79%) are in their own census block, there do exist venues that share the same block with another venue and thus not all are unique. Further limitations of this data are discussed in Section 6.

In sum, for each point, we collect information on four factors representing social constraints along which communities may be homogenous - location and categorical information about each venue and racial and income details of residents of the census block in which the venue exists. We also collect the population size of the block each venue is located within, which we will use when creating our bootstrapped samples. In order to avoid biases from missing data, we ignore venues that do not have data on all five of these points. Few venues fit this description, and those that did tended to be venues missing categorical information, which were often residences or other places checked-in to by few users anyway. In our final dataset (including the additional cleaning steps described below), we study approximately 13,000 venues.

3.2. Clustering with LDA

Authors utilizing LDA often test multiple parameterizations of the model and use results from the parameterization that scores best on some outcome metric, typically a function of the likelihood of left out data. In particular, because the number of topics, k, must be fixed by the researching in LDA, it is common to test for different numbers of topics. However, using the methodology out of its original context provides additional parameters that must be considered when selecting a model. Table I gives the parameter space we consider in the present work. In addition to testing different values of k, we also consider five other tunable parameters.

First, many works utilizing LDA for text analysis pre-process the data by removing terms that occur rarely and documents that have few words. We experiment with a variety of cut-offs for the minimum number of check-ins per user, $MinUser$, and minimum number of checkins per venue, $MinVenue$, required of a user or venue to be included in the analysis.

Second, we leverage a version of LDA that computes the posterior using an asymmetric prior on the topic distributions[8]. This form of the model has a tendency to move more popular "words" (venues) into the same clusters [Wallach et al. 2009a], allowing us to implicitly account for the large variations in popularity of the different venues. The parameter a is used to define the unevenness in the prior distribution over the latent variables. Though a is optimized during estimation of the model, we check to make sure that starting conditions do not have an obvious effect on the new domain.

Third, because LDA is often used to provide a surface-level understanding of data, the discretization of the posterior distribution of "words" (venues) to "topics" (communities) is generally made in an ad-hoc fashion. That is, the top N words for each topic are typically said to be representative of the topic regardless of their actual weight in the posterior. As an alternative, we vary a cutoff value, $MinWeight$, at which a venue

[8] The MALLET toolkit was used to run all LDAs

is determined to be associated with a given community. All venues with a probability greater than $MinWeight$ in the posterior of any cluster are considered to be representative of that venue community, and all venues below this value are considered to be noise and thus un-representative.

Finally, our previous work suggested that binarization of the user-venue vectors provided a much better qualitative interpretation of the data. That is, we found more interesting clusters by representing each user as a binary vector over all venues, where a venue is set to 1 if the user ever checked-in to that venue and a 0 otherwise, rather than including the additional information provided when users check-in to the same venue multiple times. We test the effectiveness of this decision quantitatively on our ability to obtain stable clusters in the present work.

The outcome metric we use to determine the best model parameterization is based on the stability of the resulting clusters across two independent subsets of the data. For each parameterization, we split the dataset into two temporally contiguous, equally-sized samples of check-ins and run LDA on both[9]. We then compare the resulting clusters using a variant of the Normalized Mutual Information (NMI) metric for overlapping communities, developed by [Lancichinetti et al. 2009]. NMI is a measure of the extent to which two clusterings of two possibly unequal sets of data are the same. An NMI of 1 is given when the two clusterings are exactly the same, a value of 0 is given when the two clusterings are entirely different.

Intuitively, we would expect that NMI would increase as all parameters except for k increase, as these parameters set thresholds at which we remove information that can be considered noise in some way. While we find this to generally be the case, several of the parameterizations produced communities which were too small to be of practical interest. We decide to ignore any parameterizations that provided venue communities with a mean size less than 10 (i.e. runs where on average, each community was represented by less than ten venues). The selection of ten venues as a cutoff was made based on a qualitative exploration of the resulting data, and may be considered an additional parameter to be tested in future iterations of this work.

3.3. Measuring Social Constraint Within Clusters

In this subsection, we introduce the statistics used to measure the level of homogeneity of each community on each social constraint. To define these statistics, we first describe a pair-wise distance metric for each constraint. A statistic to measure the homogeneity of a particular community on a particular constraint is then determined by calculating the mean pairwise distance for all venues in that community on the distance metric for that constraint. Below, we give the mechanism used to determine pairwise distances between two values for each of the four constraints.

3.3.1. Spatial Constraint. The Euclidean distance between locations of the venues expressed as latitude/longitude points is used.

3.3.2. Interest Constraint. Recall that the category schema used by foursquare is hierarchical. Pairwise distance is calculated via a slight variation on depth-first search of the hierarchy of interests. Two venues are maximally similar in interest when they match exactly and are both of depth 2 or 3 (very few venues were described at all three levels of the hierarchy, hence we treat these the same). If this is the case, the venues receive a distance score of 1. Otherwise, the distance is specified by $4 - L$, where L is the first level of the tree at which the two categorical representations differ. Thus, 'A::B' and

[9] We choose to split the data into two samples, as opposed to a higher number of samples, because splitting a higher number of times would have allotted too few check-ins to each sample.

'A::C' would have a distance of 3, 'A' and 'B::C::D' a distance of 4, and so on[10]. The value L has a maximum of 2 and a minimum of 0.

3.3.3. Income-based Constraint. To represent the income level of a census block, the median household income level is used. Each venue's income value is thus the median household income of the census block in which it is located. The pairwise distance between two venues becomes simply the absolute difference between the two values.

3.3.4. Racial Constraint. For each census block, we obtain best estimates of the number of people described by six different racial categories. The racial categories used are "White alone", "Black or African American alone", "Asian alone", "Some other race alone", "Two or more races" and "Hispanic or Latino". For each venue, we thus have a six dimensional vector. The vector for each venue is normalized to sum to one, and then Euclidean distance is used to define the distance between two venues.

3.4. Bootstrapping Null Distributions

Our interest in understanding how latent factors affect social connections is an area of much work in the field of social network analysis. Exponential Random Graph Modeling (ERGM) [Robins et al. 2007], the Multiple-Regression Quadratic Assignment Procedure [Dekker et al. 2007] and latent space modeling [Krivitsky et al. 2009] have all been used to explore the extent to which latent effects influence social networks. However, these models generally apply to unimodal data and are known to have trouble with the quantity of data considered here [Snijders 2011]. Additionally, these models are used to determine which latent effects impact sociality across an entire system. In contrast, our focus is on how latent factors impact the creation of particular venue communities within a larger system. In order to do so, we therefore desire an understanding of the homogeneity of the communities LDA uncovers on these latent factors relative to what we would expect by chance.

Because we already expect communities our model uncovers to be homogenous (by H1), it does not make sense to compare them with each other. Consequently, we must obtain randomized collections of venues to compare against, and we therefore employ a bootstrapping approach [Efron and Tibshirani 1993]. Bootstrapping involves two steps. First, we resample the set of venues in our analysis with replacement to generate communities formed by chance. Second, we use this set of randomized communities to create a null distribution for each statistic computed on each community. This distribution can then be used to test whether or not a statistic measured on a community discovered by LDA is significantly different than what is expected by chance via a simple hypothesis test.

The broad null hypothesis under study is thus that along a given social constraint, a community uncovered via LDA on the user to venue check-in network will be no more homogenous than a community formed by chance. However, prior work suggests the need for a more discriminatory statement to provide conclusions of sociological interest. We thus formulate two different constrained null hypotheses, one for the geospatial constraint and one for the other three constraints. We use these null hypotheses to determine how the bootstrapped samples of randomized communities are created for each constraint.

The null hypothesis we use for the geospatial statistic is the following: "the given community is no more homogenous in space than a randomized community of the same size containing venues *within census blocks of similar population size*." Regulat-

[10] Note then that we "score" each level as a distance of 1, not choosing to decrease this value at lower levels of the hierarchy. This decision was made because we could not justify an appropriate scale to define the shrinkage in distance. However, this may be an interesting avenue for future work.

ALGORITHM 1: Bootstrapping process for Racial, Income and Interest homogeneity

```
Input: Set of clusters, $C$, and $NumDesiredSamples$
Output: Bootstrapped samples for each cluster in $C$ for a given statistic
for each cluster, $c$ in $C$ do
    Obtain $kde$, the Kernel Density Estimator for the population sizes associated with venues in $c$;
    Compute $D$, a vector of the distance of each venue in $c$ from the geographic center of $c$;
    Sort $D$ into 5 equal sized bins, $B_0, B_1, B_2, B_3, B_4$;
    $clusterSamples$ = [];
    while $|clusterSamples| < NumDesiredSamples$ do
        $sample$ = [];
        Get random venue, $v_0$ according to the distribution over all venues defined by $kde$;
        Obtain $N_0$, all venues within $max(D)$ of $v_0$;
        for $i$ from 0 to 4 do
            define $N_{B_i}$, the set of points in $N_0$ within $distance\ range$ $[B_{i,0}, B_{i,|B_i|-1}]$ of $v_0$ ;
            if $|N_{B_i}| \geq |B_i|$ then
                Add $|B_i|$ random points from $N_{B_i}$ to $sample$;
            end
            else
                // This sample does not fit the constrained null hypothesis
                $sample = []$;
                $break$;
            end
        end
        // If we have a valid sample
        if $|sample|$ == $|c|$ then
            $clusterSamples$.append($sample$);
        end
    end
    // We have enough samples
    yield $clusterSamples$;
end
```

ing the distribution of population size controls for the "urban-ness" of a cluster, which is already well known to affect travel patterns of individuals [González et al. 2008; Cheng et al. 2011]. Under this null hypothesis, a random community created via our bootstrapping procedure should have a distribution of population sizes that is approximately the same as the community found by LDA we wish to test. To ensure this, we first obtain a kernel density estimation[11] of the population sizes for the venues in the community of interest. We can then generate a bootstrap sample of communities by repeatedly sampling random collections of venues that have approximately the same distribution of population size.

While controlling for population is an adequate condition to test for geospatial homogeneity, the other three social constraints we consider are highly correlated not only with the extent to which they exist in an urban setting, but also with space itself. For example, racial distribution in a community located in a small geospatial area is almost always going to be more homogenous than a community dispersed over a larger spatial region, even when controlling for population size. Consequently, a more appropriate null hypothesis for these social constraint statistics is the following: "The given community is no more homogenous along constraint X than a randomized community of the same size, within census blocks of similar population sizes, *in a similarly ho-*

[11] The bandwidth of the kernel is estimated using Scott's rule [Scott 2009].

mogenous spatial region". To obtain a bootstrapped sample of communities from this null distribution for each statistic for each community found by LDA, we utilize Algorithm 1.

For each community, we first compute the location of it's geographic center and calculate the distance of each venue in the community to this center. We store these distances in the set D, where D_i is the ith furthest venue from the community's center. We then cut D into five bins, B_0, B_1, B_2, B_3, B_4 each having an equal (up to a difference of one) number of venues. We chose to use five bins as opposed to a smaller number because it assured that all bins would have at least one venue for all communities we obtained by LDA. We chose five as opposed to a larger number of bins because we found it gave a reasonable spatial distribution and adding bins would increase computational costs. After this procedure, B_0 contains the smallest $|D|/5$ distances, B_1 contains the second set of $|D|/5$ smallest distances, and so on. Each bin is defined by a *distance range*, $[B_{i,0}, B_{i,|B|-1}]$. The term $B_{i,0}$ defines the first element in bin B_i and thus, by construction from the ordered set D, the minimum distance in B_i. The term $B_{i,|B|-1}$ refers to the final element in B_i and, consequently, the element with the maximum distance from the community's center in that bin.

We can use the number of venues per bin and the distance ranges of the bins to control the spatial spread of the bootstrapped samples by drawing randomized communities that can be placed into bins of equal size and equal distance ranges from a given center point. This process is straightforward. While we do not have enough samples, we select a random venue, v_0, to serve as the geographic center of a new sample. This venue, v_0, is sampled according to its probability in kde, the kernel density estimator of the actual cluster's population size distribution. Using a KD tree [Bentley 1975], we obtain N_0, the set of all venues within distance $max(D)$ of v_0. We can then sample from N_0 a collection of venues that fit into the same binned distribution as those from our community of interest.

That is, for each bin B_i, we take the subset of venues in N_0, defined as N_{B_i}, that are within the distance range of B_i to v_0. If there are fewer than $|B_i|$ venues that fit this description, the sample will not fit the null hypothesis and is discarded. The process then begins again with an empty sample at bin B_0. If, however, $|N_{B_i}| \geq |B_i|$, we draw from N_{B_i} uniformly, pulling $|B_i|$ of these venues and adding them into our sample. We then continue to the next bin. If there are enough venues in N_0 in each of the five distance ranges, the sample fits the null hypothesis and is added to our set of samples. We continue this process until we have reached the desired number of randomized community samples.

The outcome of Algorithm 1 is, for each community found via LDA, a set of $NumDesiredSamples$ random venue collections each having approximately the same spatial and population distribution as the cluster itself. We run Algorithm 1 for each of the three constraints, thus generating unique sample sets for each statistic, for each venue community. From these samples, we can define the null distributions we will use in our analysis by calculating our statistic on the randomized communities in the set of bootstrapped samples.

4. RESULTS

In this section, we first discuss results of our parameter selection experiments carried out to find the best model under consideration. We then use the resulting model to explore our two hypotheses.

4.1. Optimal LDA Parameterization

To find the optimal parameterization of LDA, we explore the parameter space defined in Table I. We define optimal as the parameterization having the maximum NMI across

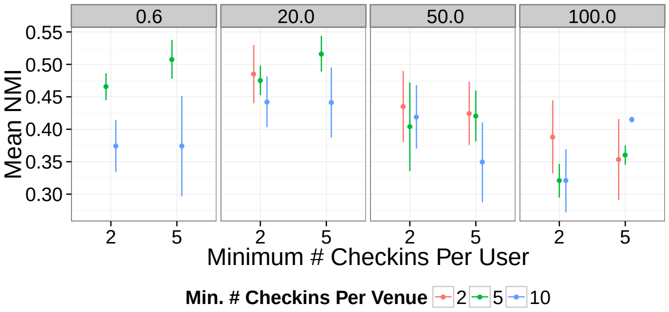

Fig. 2. The mean NMI across three runs. The x-axis represents the minimum number of check-ins needed to include a user in the clustering; color represents the minimum number of check-ins needed to include a venue. The grey bars above the plots are different values for a. Confidence intervals represent a single standard deviation from the mean estimate in both directions.

clusters obtained from the two independent subsets of the data. As LDA is a stochastic process, the NMI for a given parameterization was determined by averaging across three replications.

Figure 2 shows results from a subset of the parameter space in order to provide a clearer interpretation of the results. More specifically, we only show results for $k = 20$ which proved, as expected, to be the best setting. We also only show results with binary data, where the first partition gave, on average, 2.5 times more information about communities in the second time period than in the multinomial case. We suspect that the cause of this difference is that users tended to have multiple check-ins to venues that were of interest to few others, such as home and work [Lindqvist et al. 2011]. Finally, results are only shown with a $MinWeight$ of .005, as all other $MinWeight$ values provided clusters where the mean number of venues per community does not fit our specified criterion of 10.

From Figure 2, we see that our outcome metric faired best when a was set to either 0.6 or 20 and when $MinVenue = 5$ ($MinUser$ appeared to have a minimal effect). Within these parameterizations, a maximum NMI of .58 was reached. This suggests that regardless of model parameters, the communities produced do differ between the two time periods. This is to be expected - many communities are dynamic entities within cities [Cranshaw et al. 2012]. However, such a high value for the NMI of the two sets of clusters suggests that while a portion of the communities discovered are dynamic, many communities are stable over time with optimal model parameters.

We re-ran the top three parameterizations according to mean NMI on the full dataset ten times in order to generate the final set of communities we use to test H1 and H2. We select as our final set of communities the output of the run maximizing the empirical log-likelihood of left out data using the "left-to-right" algorithm described in [Wallach et al. 2009b] with 10% held out data. The final model came from one of the ten runs that had the same parameterization as the run with the highest mean NMI in Figure 2 ($k = 20$, $MinWeight = .005$, $Binarize = Yes$, $a = .6$, $MinUser = 5$, $MinVenue = 5$). Note that we use both stable and dynamic communities in the following subsection, where we explore results for our two hypotheses.

4.2. Main Findings

Figure 3 show the spatial positions of each venue in the twenty communities discovered. Black dots represent venues, red dots the geospatial center of the community, and

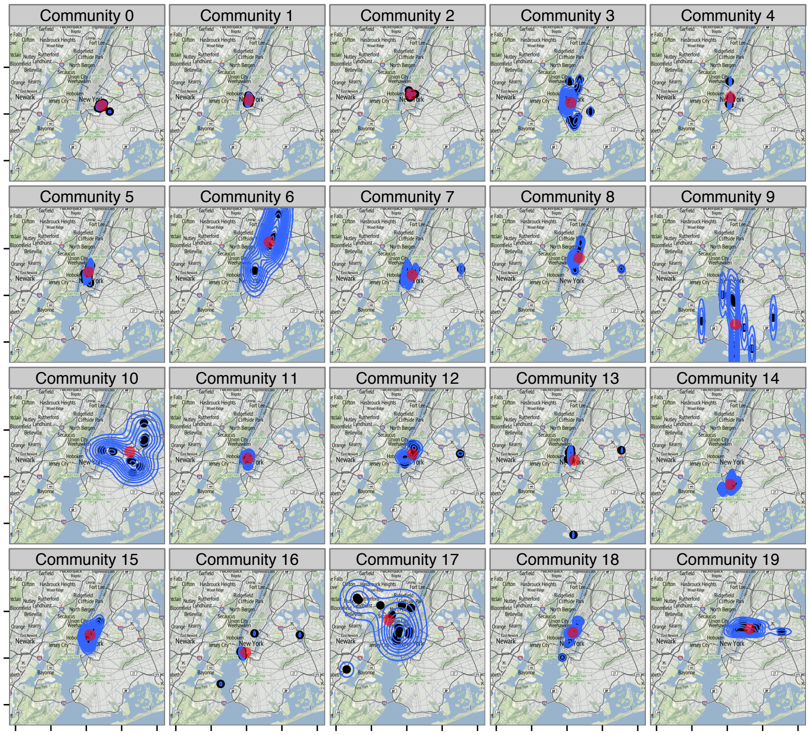

Fig. 3. The spatial distribution of the venue communities. Each black dot represents a single venue within the community. The red dot represents the geographic center. In blue is a kernel density estimation of the spatial distribution of venues to ease the effects of over plotting. The width of each map is approximately thirty miles.

blue a gaussian kernel density estimate to ease the effects of over-plotting. Communities homogenous in space exist almost entirely in Manhattan, the city center (Communities 1-5,11,12,18 and 19) or Brooklyn, the most highly populated borough (0 and 14).

4.2.1. ***H1:*** *Social constraints and cluster formation.* Our first hypothesis, H1, was that at least one the four social constraints studied would be significantly more homogenous than expected by chance in each community. Figure 4 shows results for each constraint in a different subplot. Within each, the statistic for each community is given as a red dot. The black lines represent the 95% confidence interval (CI) of the bootstrapped sample for each statistic, computed using a percentile interval [Wasserman 2003, p. 36] with $NumDesiredSamples$ set to 1,000. Where the red point falls outside of the bootstrapped CI, a standard hypothesis test would conclude that there is significant evidence to reject the corresponding null hypothesis at $\alpha = .05$. If the point falls to the left of the range, the community was significantly *more* homogenous, to the right, significantly *less* homogenous.

As is clear, geospatial proximity is estimated to be the most important constraint on community formation. Only seven communities are not significantly more homogenous in space than expected by chance, even when controlling for population. While three of these seven have homogeneity statistics near the end of the 95% bootstrapped interval, the fact that not all twenty are significantly different from null can be grounded

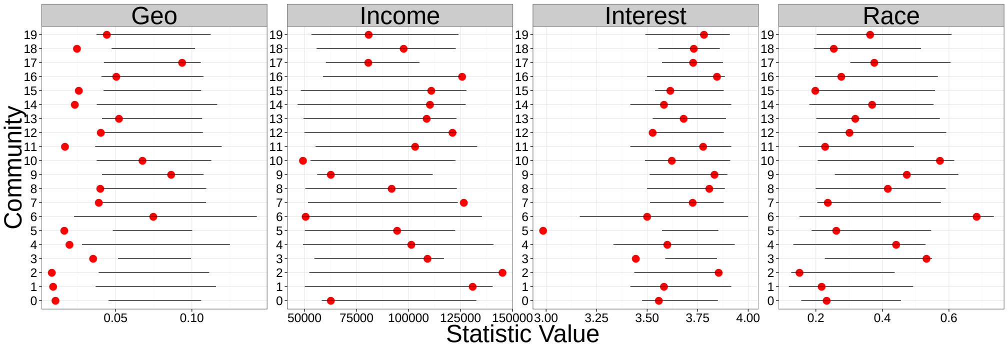

Fig. 4. The four panels represent the outcome for each community for each of the four social constraints tested. The red dot represents the measured value of the statistic on the given constraint for the given community. The black line is the 95% bootstrapped confidence interval of the null distribution of the statistic.

in previous work. [Noulas et al. 2012a] show that while spatial constraints play an important role in human movement, at small spatial scales the effect is significantly lower. Our findings differ from this previous work, however, in that we find this with respect to the spatial distribution of social communities as opposed to human mobility.

Importantly, we also consider the fact that communities not significantly homogenous in space should be homogenous on a different social constraint. To this end, we observe that Community 10 is significantly more homogenous than the null model with respect to income. While we will discuss in our results associated with H2 a case where spatial and income constraints are both significantly more homogenous than chance, Community 10 presents a rather surprising finding. The venues within this community are as spread in space as one would expect by chance (controlling for population size), but sit in areas of homogenously low income levels. We explore this in more detail in Section 5.

Community 10 is, however, the only one that exhibits quantitative support for H1 out of the seven communities not homogenous spatially. This may be due to the subset of constraints we consider in the present work or to various other biases, discussed in Section 6. We do note, though, that Community 6 and Community 9 have interest homogeneity statistics well below the mean bootstrapped estimate. In addition to the three communities that qualitatively appear to differ from the null hypothesis of the geospatial statistic, qualitative evidence thus suggests that there may be only one cluster not aligned with H1.

4.2.2. ***H2:*** *Consolidated social constraints.* H2 concerns the extent to which communities significantly homogenous on more than one social constraint are socially isolated. We found three communities such communities in our data. In Community 5 and Community 3, both spatial and interest homogeneity were significant. In Community 0, spatial and income homogeneity were both significant.

Table II shows the venues in Community 5. Twelve of the twenty-one venues in this community are “Gay Bars”. Similar venue communities were observed in our previous qualitative work in both the New York City and Bay Area data sets. To the extent that the gay community is well known to be an isolated minority culture [Smith-Lovin and Douglas 1992; McCabe et al. 2010], Community 5 presents direct support for H2. Community 0, homogenous on the income and spatial dimensions, also supports our hypothesis. It contains venues that are all located within Williamsburg, a neighborhood of Brooklyn widely regarded as one of the most “hipster” neighborhoods in the

Table II. Name and Categories of Venues in with Community 5

Place Name	PlaceCategory	Place Name	Place Category
Duplex	Nightlife Spot::Gay Bar	Boiler Room	Nightlife Spot::Gay Bar
Stonewall Inn	Nightlife Spot::Gay Bar	Splash Bar	Nightlife Spot::Gay Bar
XES Lounge	Nightlife Spot::Gay Bar	Therapy NYC	Nightlife Spot::Gay Bar
Posh	Nightlife Spot::Gay Bar	GYM Sportsbar	Nightlife Spot::Gay Bar
Barrage	Nightlife Spot::Gay Bar	Ritz Bar and Lounge	Nightlife Spot::Gay Bar
Boxers NYC Sportsbar	Nightlife Spot::Gay Bar	Industry Bar	Nightlife Spot::Gay Bar
Elmo Restaurant and Lounge	Food::American Restaurant	AMC Empire 25	Movie Theater::Multiplex
High Line Park	Great Outdoors::Park	New York Penn Station	Travel & Transport::Train Station
Blockheads Burritos	Food::Burrito Place	Vynl	Food::Breakfast Spot
Chelsea Clearview Cinemas	Movie Theater::Multiplex	Bamboo 52	Food::Sushi Restaurant

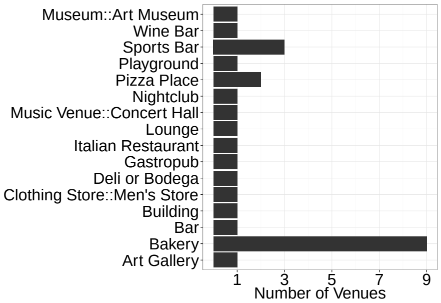

Fig. 5. A histogram of the categories for the venues in Community 3 (the top level of the hierarchy is excluded).

United States[12]. One of the general goals of the “hipster” movement was (and is) to create and portray a *unique identity* [Alfrey 2012]. The hipster community thus aligns with Blau’s hypothesis in that hipsters generally seek their own culture and therefore tend segregate themselves from the rest of the city.

While these two clusters provide support for H2, it is important to consider the implications of our data source on this finding, in particular the issue of self representation. 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; Veelen et al. 2013]. Hipsters, in an attempt to depict a unique identity, may be similarly affected [Cooper and Denner 1998; Ethier and Deaux 1994]. This suggests that users associating themselves with the homosexual or hipster cultures in New York City are more likely to check in to places representative of these cultures than members of more mainstream communities. While this does not discredit the support for H2, it does suggest that communities surrounding minority cultures in our model (and other models using foursquare data) may be idealized versions of underlying social structures, representing only the most salient aspects of that community.

In contrast to the previous two cases, where the underlying social group is inferable, Community 3 contains a high level of noise. Figure 5 provides the histogram of the categories represented by the venues in Community 3. A closer inspection of the numerous bakeries reveals that seven of nine are located in the Chinatown neighborhood of Manhattan. Homogeneity of these specific venues in this particular location suggests the tentative hypothesis that the community represents individuals of a particular na-

[12] Forbes magazine recently referred to Williamsburg as the “unofficial East Coast birthplace of hipsterism”; http://www.forbes.com/sites/morganbrennan/2012/09/20/americas-hippest-hipster-neighborhoods/

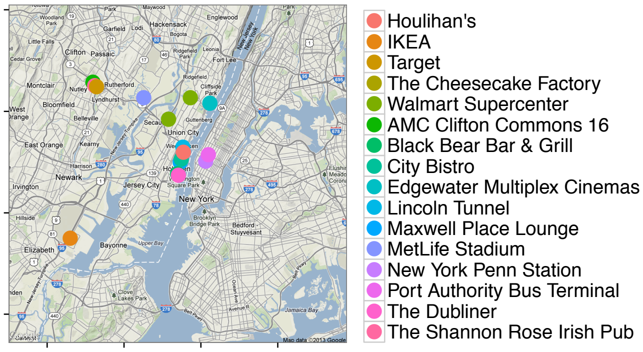

Fig. 6. Venues associated with Community 17

tional culture (with a devotion to specific delicacies), which may or may not be socially isolated. Alternatively, in line with previous results, this community is simply representative of those that desire a self representation online of this particular form of Chinese culture. Against intuition, however, racial homogeneity is significantly *higher* than we would expect from chance, as these bakeries are the only venues in the cluster to exist within an area with a high population of residents of Asian descent. While the hypothesized underlying community would support H2, it thus appears that clustering methodologies explicitly leveraging social constraints (as opposed to testing for their effects) may be necessary to uncover it.

5. DISCUSSION

We provide here a short discussion of two specific results from our study of H1 that are of qualitative interest. The first result is the existence of a community that was homogenous on income but not on space in the New York data. One-tailed paired t-tests of the median household income distribution of Community 10 against all other communities were completed with $\alpha = .05$[13]. Results show that Community 10 had a significantly lower median household income across its venues than 15 of the 19 others discovered. There is thus strong evidence that these venues represents a community where people in lower income levels go out of their way to places in other low income areas. The second finding of interest is Community 17, which does not even qualitatively match the predictions of H1. Figure 6 shows a close-up view of the venues in this community. From the plot, a reasonable qualitative explanation is that people within the urban setting must go out to less populated areas in the surrounding suburbs to access certain venues which require more space to function, like IKEA and Walmart.

The common theme in these two findings is the apparent effect of non-social factors. In particular, it is quite possible that these communities are the result of what we might term “goods effects”, where product constraints rather than social constraints mediate movement. In the case of Community 17, goods that cannot be acquired within a user’s traditional radius of movement must be obtained from locations outside the city center in an array of directions. In the case of Community 10, a similar argument applies, although the restriction is not because of the location of goods but rather because of their cost.

[13] Because these were planned comparisons, we did not to control for multiple comparisons, see discussion in [Seltman 2012]

6. CONCLUSION

In the present work, we use foursquare data to consider the extent to which social constraints affect the formation of communities within the urban environment (H1) and how consolidation of these constraints may indicate the existence of isolated social groups within the city (H2). We observe significant qualitative and quantitative support for H1. Specifically, we find that all but one of the communities uncovered presents qualitative support for the hypothesis and 70% of the communities found provide quantitative support. In addition, we find that even when controlling for population size, geospatial constraints are estimated to be the most influential on social community formation. We did, however, uncover an interesting case where spatial homogeneity did not appear to affect community formation, while income homogeneity did. We also observe reasonable support for H2. Specifically, two of the three communities we discovered, which we aligned to homosexual and hipster cultures respectively, are well known to exist in relative social isolation from the rest of the city. Results for the third community were inconclusive, as there was some, but not enough, evidence of a culturally distinct community.

These inconclusive results are indicative of the exploratory and sociological nature of our study, and we note that results may be affected by several biases. First and foremost, null hypotheses may be under or over-constrained, leading to confidence intervals that are too small or too large. Second, LDA is used outside of its intended domain, and is furthermore a relatively simple Bayesian framework. While this allows for increased interpretability, our model may very well have detected some clusters of venues that are poor representations of social communities. Third, our use of population-level statistics, while perhaps the best available data for our purpose, may be subject to scrutiny in their application to judging homogeneity of social constraints [Butts et al. 2012]. For example, as census data represents residents of a particular location, it may not be representative of the actual population distribution at certain times of the day (e.g. business districts). Finally, the data used does not allow us to explicitly distinguish between certain socio-psychological processes of interest, namely self representation versus social segregation.

Beyond addressing these limitations, further work might explore additional means of testing our hypotheses beyond the traditional approach of hypothesis testing that is utilized here, in particular approaches formalizing novel prediction problems that address our questions. Along these lines, interesting avenues of future work might include using Dirichlet Multinomial Regression (DMR) based topic models [Mimno and McCallum 2012; Yuan et al. 2012] to explicitly incorporate the social constraints studied here as opposed to testing for their latent effects. Such a model is particularly interesting in that DMR can be considered a direct extension of McFadden's discrete choice model [Guimaraes and Lindrooth 2005]. Thus, it could be used to seed homophily-based multi-agent simulations [Carley et al. 2009] of the movement of people within a city, as agents could have an empirically-grounded cognitive model of movement decisions. This work would complement recent efforts to include geospatial information into such models [Moon and Carley 2007; Carley et al. 2011].

Regardless of the avenue of research pursued, however, it is evident that the influx of "big data", particularly with respect to location-based data, has created new opportunities to understand human behavior in the urban environment. In particular, we have barely scratched the surface of what is possible with respect to the study of macrosociological processes. While the costs associated with these opportunities, such as the use of observational, mediated data, are not to be regarded lightly, we believe that the works we have relied upon 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 of Alexander Loewi, Michael Martin, Ju-sung Lee, Geoffrey Morgan and Jonathan Morgan in their assistance in formalizing the ideas which were presented in this work. Last but certainly not least, we gratefully acknowledge Justin Cranshaw, who provided both guidance and the data utilized here and our anonymous reviewers for their insightful responses to previous versions of this article.

REFERENCES

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

BAO, J., ZHENG, Y., AND MOKBEL, M. F. 2012. Location-based and preference-aware recommendation using sparse geo-social networking data. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. 199–208.

BARBER, M. J. 2007. Modularity and community detection in bipartite networks. *Physical Review E 76,* 6, 066102.

BAUER, S., NOULAS, A., SEAGHDHA, D., CLARK, S., AND MASCOLO, C. 2012. Talking places: Modelling and analysing linguistic content in foursquare. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. 348–357.

BENTLEY, J. L. 1975. Multidimensional binary search trees used for associative searching. *Communications of the ACM 18,* 9, 509–517.

BLAU, P. 1977a. *Inequality and heterogeneity: A primitive theory of social structure*. New York: Free Press.

BLAU, P. 1977b. A macrosociological theory of social structure. *American journal of sociology*, 26–54.

BLAU, P. M. 1974. Presidential address: Parameters of social structure. *American Sociological Review 39,* 5, 615–635.

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.

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

BUTTS, C. T., ACTON, R. M., HIPP, J. R., AND NAGLE, N. N. 2012. Geographical variability and network structure. *Social Networks 34,* 1, 82–100.

CARLEY, K. M., MALLOY, E., AND ALTMAN, N. 2011. Multi-agent modeling of biological and chemical threats. In *Infectious Disease Informatics and Biosurveillance*. Springer, 361380.

CARLEY, K. M., MARTIN, M. K., AND HIRSHMAN, B. R. 2009. The etiology of social change. *Topics in Cognitive Science 1,* 4, 621–650.

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, 1082–1090.

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, Stockholm, Sweden, 57–66.

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.

DEKKER, D., KRACKHARDT, D., AND SNIJDERS, T. A. B. 2007. Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. *Psychometrika 72,* 4, 563–581.

DEUZE, M. 2012. *Media Life* 1 Ed. Polity, Cambridge, MA.

EFRON, B. AND TIBSHIRANI, R. 1993. *An introduction to the bootstrap*. Vol. 57. Chapman & Hall/CRC.

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, 9–16.

FORTUNATO, S. 2010. Community detection in graphs. *Physics Reports 486,* 3, 75–174.

GONZÁLEZ, 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.

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. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. 482–485.

KRIVITSKY, P. N., HANDCOCK, M. S., RAFTERY, A. E., AND HOFF, P. D. 2009. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social networks 31,* 3, 204–213.

LANCICHINETTI, A., FORTUNATO, S., AND KERTSZ, J. 2009. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics 11,* 3, 033015.

LAWLOR, A., COFFEY, C., MCGRATH, R., AND POZDNOUKHOV, A. 2012. Stratification structure of urban habitats. In *PERVASIVE 2012*.

LEHMANN, S., SCHWARTZ, M., AND HANSEN, L. K. 2008. Biclique communities. *Physical Review E 78,* 1, 016108.

LI, Q., ZHENG, Y., XIE, X., CHEN, Y., LIU, W., AND MA, W.-Y. 2008. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. GIS '08. ACM, New York, NY, USA, 34:1–34:10.

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, Vancouver, BC, Canada, 2409–2418.

MASSEY, D. S. AND DENTON, N. A. 1988. The dimensions of residential segregation. *Social Forces 67,* 2, 281–315.

MCCABE, S. E., BOSTWICK, W. B., HUGHES, T. L., WEST, B. T., AND BOYD, C. J. 2010. The relationship between discrimination and substance use disorders among lesbian, gay, and bisexual adults in the united states. *Journal Information 100,* 10.

MCPHERSON, J. M. AND RANGER-MOORE, J. R. 1991. Evolution on a dancing landscape: Organizations and networks in dynamic blau space. *Social Forces 70,* 1, 19–42.

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*.

MOON, I.-C. AND CARLEY, K. M. 2007. Modeling and simulating terrorist networks in social and geospatial dimensions. *Intelligent Systems, IEEE 22,* 5, 4049.

NOULAS, A., SCELLATO, S., LAMBIOTTE, R., PONTIL, M., AND MASCOLO, C. 2012a. A tale of many cities: universal patterns in human urban mobility. *PloS one 7,* 5, e37027.

NOULAS, A., SCELLATO, S., LATHIA, N., AND MASCOLO, C. 2012b. Mining user mobility features for next place prediction in location-based services. In *2012 IEEE 12th International Conference on Data Mining (ICDM)*. 1038–1043.

NOULAS, A., SCELLATO, S., MASCOLO, C., AND PONTIL, M. 2011. 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, 570–573.

OPSAHL, T. AND PANZARASA, P. 2009. Clustering in weighted networks. *Social networks 31,* 2, 155–163.

ROBINS, G., PATTISON, P., KALISH, Y., AND LUSHER, D. 2007. An introduction to exponential random graph (p*) models for social networks. *Social Networks 29,* 2, 173–191.

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, Seattle, Washington, USA, 723–732.

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, San Diego, California, USA, 1046–1054.

SCHELLING, T. C. 1971. Dynamic models of segregation. *The Journal of Mathematical Sociology 1,* 2, 143–186.

SCOTT, D. W. 2009. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons.

SELTMAN, H. J. 2012. *Experimental design and analysis*. Carnegie Mellon University, Pittsburgh, PA.

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

SNIJDERS, T. A. 2011. Statistical models for social networks. *Annual Review of Sociology 37,* 1, 131–153.

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, 85–94.

VASCONCELOS, M. A., RICCI, S., ALMEIDA, J., BENEVENUTO, F., AND ALMEIDA, V. 2012. Tips, dones and todos: uncovering user profiles in foursquare. In *Proceedings of the fifth ACM international conference on Web search and data mining*. 653–662.

VEELEN, R. V., OTTEN, S., AND HANSEN, N. 2013. A personal touch to diversity: Self-anchoring increases minority members identification in a diverse group. *Group Processes & Intergroup Relations*.

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, 1105–1112.

WASSERMAN, L. 2003. *All of Statistics: A Concise Course in Statistical Inference*. Springer.

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, 186–194.

ZHENG, Y. 2011. Location-based social networks: Users. In *Computing with Spatial Trajectories*, Y. Zheng and X. Zhou, Eds. Springer New York, 243–276.

Received October 2012; revised ; accepted