Beyond “Local”, “Categories” and “Friends”: Clustering foursquare Users with Latent “Topics”

Kenneth Joseph
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15212
kjoseph@cs.cmu.edu

Chun How Tan
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15212
chunhowt@cmu.edu

Kathleen M. Carley
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15212
kathleen.carley@cs.cmu.edu

ABSTRACT

In this work, we use foursquare check-ins to cluster users via topic modeling, a technique commonly used to classify text documents according to latent “themes”. Here, however, the latent variables which group users can be thought of not as themes but rather as factors which drive check in behaviors, allowing for a qualitative understanding of influences on user check ins. Our model is agnostic of geo-spatial location, time, users’ friends on social networking sites and the venue categories- we treat the existence of and intricate interactions between these factors as being latent, allowing them to emerge entirely from the data. We instantiate our model on data from New York and the San Francisco Bay Area and find evidence that the model is able to identify groups of people which are of different types (e.g. tourists), communities (e.g. users tightly clustered in space) and interests (e.g. people who enjoy athletics).

Author Keywords

location-based service, foursquare, topic modeling

ACM Classification Keywords

H.5.m Information Interfaces and Presentation: (e.g. HCI); J.4 Social and Behavioral Sciences: Sociology

INTRODUCTION

There has long been an interest in understanding how, when, where and why people move from place to place [14]. In recent years, such studies have begun to focus more on the large amount of geo-spatially tagged data being produced from mobile devices, as such data allows one to approach questions of societal-level interest in an entirely data-driven manner [26, 7, 27]. Data drawn from mobile devices on the whereabouts of their users has led to an influx of interesting findings in explaining patterns of human mobility [6, 17, 5], predicting friendship on social networking sites based on location data [23, 20], and better understanding different aspects of cities, both at the city level as a whole [24] and at the neighborhood level [7, 8].

A significant amount of previous work on human mobility has come to the conclusion that people tend to stay within relatively small geographic areas for the majority of their time [18, 17, 16, 6]. These areas, particularly in cities, are often representative of different neighborhoods - areas in cities with dynamic, fuzzy boundaries [7, 8] whose residents exhibit homophilic tendencies, both in their demographics and social interactions [9]. Because of the importance of neighborhoods in a diverse set of social processes (e.g. [21]), a natural way of conceptualizing groups of people in cities is to cluster them based on the neighborhoods in which they reside or are most active.

Yet while heterogeneity in a population can to a large extent be explained by closeness of social and geodesic distances [4] (a closeness inherent in neighborhoods), there are other ways of defining “groups” of people in cities. For example, one could consider tourists, who almost by definition are not bound to specific parts of a city, as being a group of interest. Similarly, sports enthusiasts, bound not to neighborhoods but more to the specific places (e.g. stadiums) they frequent, may be of interest as well. These two examples represent interesting and useful groupings of people moving within cities which the concept of a neighborhood cannot fully capture.

In this work, we consider an alternate approach to defining groups of people in a city by characterizing people simply by the places they go. Such an approach has previously been found effective for uncovering social and interest relationships between users as well as for location and friend recommendations, though the approach taken also considered temporal information and utilized a different type of location data [10, 27]. We work with a large dataset of foursquare check ins, obtained from the authors of [7], which details where and when people were at specific locations around different U.S. cities. While the dataset gives a diverse set of information, we describe each user simply by the places they go and how often they go there, thus choosing to ignore geospatial and social information which exists in the data. In addition, we ignore information on the category of different places, as explained in later sections.

Using such a simple feature set for users is somewhat counterintuitive, however, we do so with a specific purpose. First, the relationship between social ties, geodesic distances between people and their temporal coevolution are intertwined in an intricate manner which has only recently begun to be understood in a strong quantitative manner [4, 9, 10, 19]. By only implicitly considering these variables in our data rep-

UbiComp ’12, Sep 5-Sep 8, 2012, Pittsburgh, USA.
978-1-4503-1224-0/12/09

resentation, we allow for them to emerge as latent factors whose association we need not explicitly predefine. This allows a further understanding of how these variables may affect the check in behaviors of users. In addition, by not specifying any presumed factors to be responsible for similar check in locations between users, we avoid restrictions of the types of groups our model might find. For example, explicitly using geo-spatial features may restrict our ability to understand groups of users with similar interests which are spread throughout a city, such as the tourists described above.

Given that our feature set is simple and we desire an understanding of what causes users to check in at various locations, the model we use must allow us to posit that latent factors affect where users go within cities, quantify how each user is affected by them, and give some intuition as to what these latent factors might be. We use a clustering model based on the idea of topic modeling, a method of clustering which captures these very concepts. Specifically, our model assumes that every user can be represented by multiple hidden factors, and that each check in by that user is motivated by one or more of these hidden variables. These hidden factors may represent, for example, interests or needs of a user, but the range of their distinction is broad, as topic models force the researcher to determine the qualitative meaning behind the hidden factors it discovers. Users can then be grouped by how strongly they are affected by these hidden factors, and the hidden factors themselves can be defined by a certain set of locations (or venues) which are frequented by the users clustered together by it.

In order to understand what might be gained from a topic model approach to location-based social network data, we instantiate Latent Dirichlet Analysis (LDA) on foursquare data from check ins in New York City and the Bay Area. We find that our model produces latent collections of people which represent both geo-spatially close groups and people who appear to have similar interests, thus suggesting social factors are at play. Our model therefore captures drivers of user check in behavior found to be so important in grouping people in cities into neighborhoods, lending another piece of support to previous work in this area [7, 9, 4]. However, in addition to latent factors indicating groupings due to social and geo-spatial closeness, we also find clusters of different "types" of people, such as tourists, who tend to visit specific venues which do not seem, qualitatively, to have any clear geo-spatial or social relevance. Thus, by using a model agnostic of place category, geo-spatial information and friendship information, we create a model which is rich enough to incorporate all of them and extend beyond to include these user "types". This allows for a deeper understanding of user check in behavior on location based social networking sites.

After describing our findings, we discuss some potential applications of a topic-model approach, such as venue recommendation. We also consider ways in which our model is different from preexisting methods of recommendation using similar data [27, 10], with particular concern to note the limitations of the current approach.

RELATED WORKS

The Data

The data that we use, obtained from the authors of [7], is a set of approximately 18 million data points from across the United States of users of foursquare. Foursquare is a socially-driven location sharing application [11], where users can "check in" to different locations and have these check ins be shared with friends both on foursquare and on various other social networking sites. In addition to allowing users to share check ins with others, foursquare uses various gamification techniques to encourage contributions, including rewarding users with badges and points for various actions.

Indeed, these gamification techniques have been found to be a strong determinant in use of foursquare. Lindqvist et al. [11] found the most likely reason for a check in was for the gamification aspects of the site, followed by social aspects (e.g. to interact with friends), to visit and discover new places, and to keep track of personal history and accomplishments. Users were also asked for reasons why they would not check in at a location- these centered on privacy concerns and issues of self-representation. Self-representation concerns the fact that users have a desire to be represented as being of a certain type, leading to the possibility that a user's check ins could mis-represent his or her interests. For example, users surveyed tended to not want to check in to places which they perceived to be uninteresting (e.g. work) or embarrassing (e.g. fast food). Such findings are pertinent to our understanding of the clusters resulting from our topic model in that they must be analyzed from the perspective of the typical user- one that is at least partially interested in gamification and self-representation.

Each of our 18 million data points represents one check in which was published to Twitter by a foursquare user. In the data set used, a check in provides a unique user Twitter id, the timestamp of the user check in, an optional user description (e.g. "the coolest place ever!"), and also the venue id of the check in location. Using this venue's id, the original data collectors also obtain the venue's name, geo-location, and "category" information. These categories are drawn from a set of hierarchical categories given by foursquare itself - there are over three hundred categories, the full list of which can be found by querying the foursquare API [1]. We utilize these categories extensively in defining the hidden factors which our topic model generates. From the data set collected by the authors of [7], we consider check ins located in the metropolitan areas of New York City and the San Francisco Bay Area.

Human Mobility

Noulas et al, in [17], study distances users travel between successive check ins, noting that nearly 80% of the total check ins for a user occur within 10 kilometers of the previous check in. Though larger than the typical neighborhood, this lends support to the idea that people tend to stay within small sections of cities and hence can be grouped in this manner. Similarly, user displacement, or distance between two successive check ins, follows a power law which

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

can be modeled by a Lévy Flight [5]. Work in [16] shows that across various American cities, the density of the city is negatively correlated with user displacement.

While such work suggests that users tend to stay within reasonably small areas, particularly in dense areas, they provide little evidence of the specific places people are traveling to at various distances. Cho et al. [6] explain mobility using the concept of two places, home and work, but do not go in depth into travel to places which might represent user interests. In contrast, Cranshaw et al., in [7], present an approach which utilizes geographic proximity in combination with user check in history, thus incorporating a better understanding of the specific places people travel to. This information is utilized to understand how the existence of "neighborhoods" can be approximated by foursquare check ins. Our work is different from the work of Cranshaw et al. in that we focus on clustering users, as opposed to venues, into groups. It is important to notice, however, that the method in [7] can be applied without modification to show a variety of neighborhoods around a city which similar users might frequent. Such a model suggests that incorporating venue categories as a feature describing user movement restricts the formation of neighborhood clusters- each neighborhood has its own set of venues within a variety of categories.

The question of user clustering, as opposed to venue clustering, has been previously approached [12, 22, 10], most noteably in [10], where a hierarchical, temporally aware user clustering mechanism (HGSM) is proposed. This method is extended to show its abilities as a recommender system in [27], where it is shown to perform well on tasks involving recommending both places and social connections for users. We discuss how this model compares to the one presented here in Section .

One might be tempted to assume that human movement can be much better understood if it is conditioned on the movement of friends. Indeed, much work has been done to show that one can reliably predict the location of a user based on the location of their friends (see [20, 23] for recent examples). Furthermore, recent work has shown that neighborhoods implied by census boundaries can be inferred from social graphs [9]. However, evidence from [6] suggests that while location prediction is possible, predicting where a user will go based on where their friends on location-based services go is not as straightforward. Cho et al. [6] state that people who are friends on Brightkite and Gowalla have a check in in common less than ten percent of the time. Furthermore, the authors find that travel over short distances is not heavily impacted by the social network structure - friendship links on location-sharing social networks only can explain about 30% of all check ins. These findings suggest to us that explicitly adding in features of friendship on social networking sites may restrict clusters to community-based groups, perhaps overpowering other latent variables such as user interests which exist in the data.

MODEL DESCRIPTION

To cluster foursquare users into meaningful groups which are representative of different factors driving check in behavior, we apply the idea of topic modeling. Specifically, we apply a topic model known as Latent Dirichlet Allocation (LDA), first introduced in [3]. LDA is a latent space model commonly used to better understand text corpora by representing a large collection of documents in a much more compact set of hidden topics. In a typical LDA model (as discussed in [3]) 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 hidden topics, and each hidden topic can be described by the words which are most heavily associated with it. For example, a document about the opening of a new Italian restaurant might contain the words "restaurant" and "dinner", associated with Topic 1, and the words "pizza" and "spaghetti", associated with Topic 2. LDA would give us information on how heavily the document was related to the two topics, and we could understand what these two topics were about by considering the words which are associated with them (e.g. Topic 2 is likely about "food" or "Italian Food"). By considering documents which are highly associated with the same topic, we can begin to understand "clusters" (or groups) of documents in the corpora, where each cluster represents a set of documents which are related to a given topic.

Ref. [8] has successfully applied LDA to location check in data. Specifically, [8] applied a topic-model approach on socially-tagged data from a location-sharing social network in order to understand boundaries that might exist on neighborhoods and characteristics of these neighborhoods. In their topic model, the "documents" were regions generated by splitting geo-spatial coordinates into grids, their "words" were venue category tags, and their topics were hypothesized to be archetypal neighborhoods. Note, however, that LDA was still performed on text.

In contrast, we use an instantiation of the data which does not revolve around the concept of themes in text. Rather, in order to model user check in behaviors using LDA, we use the analogy of a document to represent a user, and thus each check in for a user can be thought of as a word in a document. As each venue has a unique identifier, we can model each as a unique word. This means, for instance, that the Starbucks on 5th Street will be different than the Starbucks on 10th Street. Similar to text documents, where documents can have the same word multiple times, we define a multinomial distribution for the check ins for each user by using the counts of check ins for each venue as features.

Using our representation of user check in behavior, we obtain a set of hidden topics which can each be described by a set of venues (words), and which can be used to categories users (documents) according to these topics. Because these topics have to do with check ins at different venues, we can associate them not with textual themes but rather with factors which drive users to check in at various locations (e.g. interest). More specifically, for each user, we obtain a set of weights corresponding to each hidden topic, allowing us to understand multiple facets of the behavior of each user. Thus, a benefit of using LDA is that each user can be represented as a distribution of a variety of drives in behavior. This coincides well with our intuition that check in behavior is driven by multiple underlying factors, each of which may be used to correlate the behavior of a user with others and thus may help to better understand user check in behavior at a general level.

In the case studies below, we set the number of hidden topics to be twenty. A shortcoming of LDA, addressed in later topic models (as shown in [3]) is that an arbitrary number of hidden topics need be chosen. We complete sensitivity tests, as suggested in [3], and find that our model is most effective and most interpretable when we use twenty clusters. In addition, we remove those users with less than 5 unique venue check ins and those venues with less than 10 check ins, repeating the pruning iteratively until all such venues and users are removed. This approach of pruning data points is common in document modeling, as in imperfect documents there tend to be spelling mistakes which occur rarely and are obviously not of interest. Similarly, in our case, those users who have few check ins might be newcomers to foursquare who quickly stop using the application and thus might not be well-represented by their check ins. However, this pruning criteria is selected arbitrarily and future work in the direction of data selection is important.

RESULTS

In this section, we present two exploratory case studies of the results of running LDA on check in data from New York City and the San Francisco Bay Area. The New York data set initially had a total of 448,156 check ins, 36,388 users and 44,312 venues. After pruning, we were left with 288,029 check ins, 10,459 users and 7,432 venues. Note that we still keep more than half of the check ins, although the number of venues and users decreases significantly. While future work may make use of these data points, we find that including them in the model makes clusters more difficult to interpret, as is often the case when incorporating analogous words and documents when running LDA on text corpora. Similarly, for San Francisco Bay Area data, we initially have a total of 181,572 check ins, 18,650 users and 20,844 venues, and were left with 102,851 check ins, 4,269 users and 3,439 venues after applying the same pruning criteria.

In our analysis, we examine the "top" venues in each cluster, as given by weights from the LDA. By observing these venues, we are able to better understand the latent factor which is representative of the users in each cluster. In the following sections, we provide a qualitative analysis of three different "kinds" of latent factors that our model uncovers as being hidden drivers in where users check in. We develop this intuition by considering the geo-spatial distribution and categorical information (garnered from foursquare category information) of the venues which represent each cluster. Thus, the cluster types that we generate are in some cases quite similar to the typical category of the venues within that cluster, though we will show that this is not always the case.

Category	**Venue Name**
"Sport Enthusiast" Cluster	
Baseball Stadium	Yankee Stadium
Football Stadium	MetLife Stadium
Baseball Stadium	Stadium Citi Field
Hockey,Basketball Arena	Madison Square Garden
"Art Enthusiast" Cluster	
Performing Art Venue	NBC Studio 1A
Art Museum	Brooklyn Museum
Art Museum	Metropolitan Museum of Art
Art Museum	Museum of Modern Art (MOMA)

Table 1. Top venues of two clusters found in the New York data. These two clusters can be thought of as clustering users based on similar interests

Interest Factors

We define interest factors as those where the top venues within that cluster are all associated with a specific action which can be performed, such as eating ice cream or watching a sporting event. Two examples of such clusters in the New York data are shown in Table 1 - similar clusters, not shown, are observed in the San Francisco data. In many respects, one would question any model of user behavior which does not in some way account for the interests of the user, and as such the fact that our model discovers interests as a hidden driver of user action is not of particular surprise. However, such clusters are of interest in suggesting that certain arguments recently put forth for describing human mobility are too simplistic. In particular, claims that simple concepts of geo-spatial phenomenon, as suggested in [16], are sufficient to explain human mobility cities should be treated with some skepticism - our model suggests contextual factors, including points of interest at different locations in cities are clearly of high importance [27]. This can be observed by the relatively wide geo-spatial spread of the two interest clusters from the New York data, as can be seen in the clusters with these names in Figure 1.

In addition to casting doubt on the simplicity of human mobility modeling, the existence of clusters of users driven by what we presume to be similar interests suggests that social factors have a strong underlying effect on locations that users check in to. This claim is based on the well-supported notion that social acquaintances tend to have similar interests [13]. Such a finding indicates that a topic model approach to clustering users may be an effective route to generating friendship recommendations [27] with the added benefit of being able to give reasons for the recommendation which are less intrusive. For example, instead of suggesting to two users in New York that they may want to become friends because they have both visited Yankee Stadium and MetLife stadium within the past few months, a topic model approach (with the assistance of a practitioner) could instead suggest that these two users become friends because they are both "Sport Enthusiasts". Though other approaches might provide similar functionality, it would likely be on a more case-by-case basis.

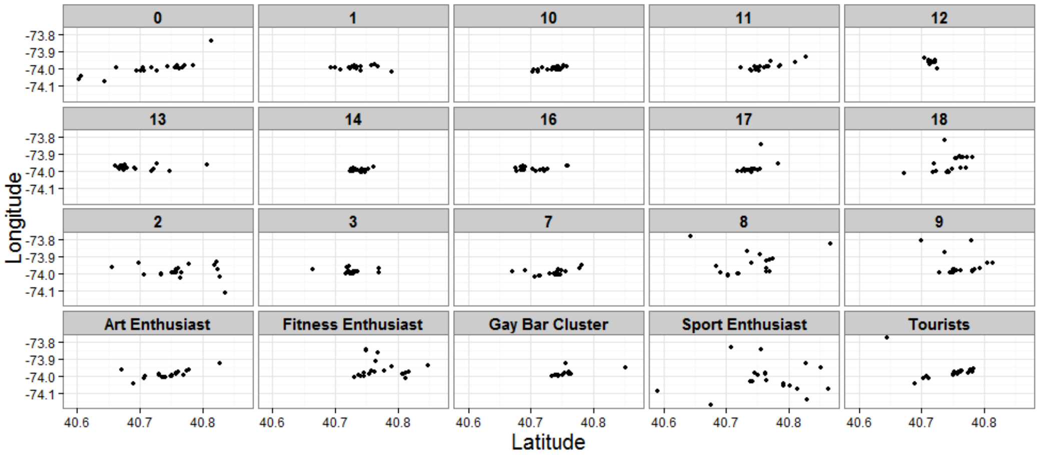


Figure 1. The geo-spatial distribution of the twenty clusters our model discovers in New York - each point represents one of the top twenty venues in each cluster. Those clusters assessed which could be qualitatively associated with a name are labeled with these names, as discussed in the paper. Those with numbers are not qualitatively assessed in this work-they are shown to give a better understanding of spatial distribution of clusters

Category	Venue Name
Bridge	George Washington Bridge
Gay Bar	Therapy NYC
Gay Bar	Boxers NYC Sportsbar
Gay Bar	Ritz Bar and Lounge
Gay Bar	Splash Bar
American Restaurant	Elmo Restaurant and Lounge
Gym	Equinox
Gay Bar	Posh
Train Station	New York Penn Station
Gay Bar	Pieces Bar
Coffee Shop	Starbucks
Gay Bar	XES Lounge
Gay Bar	Barrage

Table 2. A cluster consisting mainly of gay bars, found in the New York data

Category	Venue Name
Gay Bar	Toad Hall
Park	Mission Dolores Park
Gay Bar	Badlands
Gay Bar	QBar
Gay Bar	Club Trigger
Gay Bar	The Lookout
Burger Joint	Harvey's
Gay Bar	440 Castro
Gay Bar	Blackbird Bar
Gay Bar	The Mix
Supermarket	Safeway
Movie Theater	AMC Loews Metreon 16
Gay Bar	Moby Dick
Train Station	Castro MUNI Metro Station

Table 3. The San Francisco "gay bar" cluster. These venues are all found in The Castro, an area with a large gay population. Notice that venues of other types are included in the cluster

Community Factors

Given the previous work which suggests that geo-spatial (and thus social [4]) factors influence user mobility, it is also not surprising to see several clusters which are tightly clustered in space. However, as our model ignores the geo-spatial coordinates of venues, it is interesting to note that such clusters are purely the result of a group of users which are all driven by some hidden factor driving them to check in within a small geographical area. It is important to note, however, that this factor could be representative of issues of self-representation, or to some genuine factor influencing users to stay within that area. We thus define these clusters as "communities", in that users either feel strongly that they are associated with this specific area, or are indeed frequenters of the area.

In the results from both the New York and San Francisco Bay Area data, a "community" cluster exists where the representative venues are nearly all of the category "Gay Bar". A list of places which describe these clusters in the two data sets are shown in Table 2 and Table 3, and the closeness of these places in space can be seen for the New York data in Figure 1. Thee San Francisco gay bar cluster is similarly close in space- in fact nearly all venues in the cluster are located in "The Castro", a neighborhood well-known for its gay population. This can be seen in Figure 2, where each marker represents one of the top twenty venues associated with the given cluster. What is particularly interesting is that the observed hidden factor associated with these clusters correlates well with a segment of the population which is heavily discriminated against, fitting traditional notions which suggest that people who are discriminated against tend to coalesce into tight communities [2]. Indeed, while many other types of venues carrying explicit demographic information about their users, such as churches, exist in foursquare's categories, this was the only one to repeatedly appear as a topic across both cities and various model configurations. The ability of foursquare data to reveal such segregations even when geo-spatial properties of venues are ignored is a rather interesting finding which we hope to explore in later work.

User Type Factors

The final kind of cluster we uncover in our results groups users by hidden factors we refer to as a "type". We define type clusters as those which group users into a recognizable form which is clearly distinguishable quantitatively but rep-

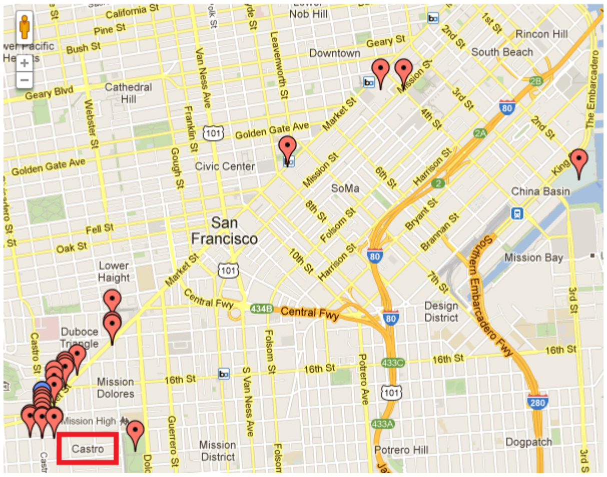


Figure 2. The geo-spatial distribution of the community cluster found in the San Francisco data. Each marker denotes a location, and the name of the traditionally gay area in San Francisco, the Castro, is boxed in red for the reader.

Category	Venue Name
Electronics Store	Apple Store
Train Station	New York Penn Station
Train Station	Grand Central Terminal
Park	Central Park
Airport Terminal	Terminal 5
Art Museum	Museum of Modern Art (MOMA)
Park	Bryant Park
Art Museum	Metropolitan Museum of Art
Department Store	Macy
Bridge	Brooklyn Bridge
Plaza	Rockefeller Center
Science Museum	American Museum of Nat. History
Historic Site	National September 11 Memorial
Toy or Game Store	FAO Schwarz
Monument	Statue of Liberty
Hotel	Hilton New York
Art Museum	Guggenheim Museum

Table 4. A cluster consisting mainly of tourist attractions in New York

resents users across a varied geo-spatial setting and across a variety of possible interests. As such, these type clusters could be considered a kind of "catch-all", however in noting that we only suggest that a few clusters qualify, we do not consider it as such. A type cluster for New York can be seen in Table 4, where the users grouped into this cluster appear to be of the type "tourist". We make this claim based on the fact that most of the places representing this cluster are sites which tourists would visit, and includes several travel venues, in particular the airport. Similarly in the San Francisco Bay Area data, we identify a type cluster corresponding to Stanford students, as shown in Table 5. Here, the top venues are either establishments within Stanford University or common places that college students might visit in San Francisco (e.g. movie theaters and bars), along with a few public transport stations.

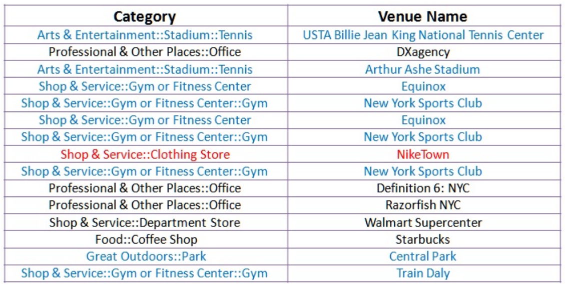
Type clusters are interesting in that they show the ability of a latent model to capture relationships between users which cannot be easily expressed in a parameterized model. As such, our model can be seen here to transcend simple categorical, geo-spatial and social factors which influence users to check in a different locations, and thus gives evidence of topic models as being a useful approach for location based data. In particular, as many users utilize foursquare to present themselves as being a specific type of person [11], topic models which expose different characteristic types of people, in addition to user interests, may be much more apt to make recommendations based not on superficial factors but on more internalized ones such as geo-spatial homophily. While such clusters are interesting in defining non-obvious, latent factors affecting where users check in, a drawback to type clusters is that they do not have as distinctive features, such as common venue categories or tight geo-spatial locations, as the other kinds of clusters we observe. Thus, these clusters clearly require increased knowledge of the city at hand to justify their existence, and as such should be approached with caution until more quantified means of analyzing their existence are examined.

Category	Venue Name
Subway Station	Civic Center BART Station
Subway Station	Balboa Park BART Station
University	The Quad
University	Jordan Hall
University	Gates CS Building
University	Hoover Tower
Nightclub	The Ambassador
Movie Theater	AMC Bay Street 16 and IMAX
Bridge	San Francisco-Oakland Bay Bridge
Light Rail	BART - Transbay Tube
Stanford	Stanford Golf Course
Stanford	Stanford University
Plaza	The Claw
Sculpture Garden	Rodin Sculpture Garden
Movie Theater	Regal Emery Bay 10
Movie Theater	AMC Loews Metreon 16

Table 5. A cluster from the San Francisco Bay Area which seems to consist of Stanford students

FUTURE WORK AND LIMITATIONS

There are several application areas which could be pursued with the output from our model. One obvious use of the clusters discovered would be to design a recommendation system, as is done in [27] with the user similarity metric developed first in [10]. As noted, our model can recommend related places which are not necessarily of same category, leading to robust and interesting recommendations based on representative latent factors which have driven previous check ins of users. For example, Figure 3 shows a cluster of users that seem to frequent fitness centers. We observe that these "fitness enthusiasts" also frequent Nike-Town, an athletic clothing store which would make for a very reasonable recommendation. One advantages to our approach in location recommendation over previous user similarity approaches in [27, 15] is that we cluster specifically on Points of Interest (POI), as opposed to geographic coordinates, thus allowing us to recommend areas without the need to have an additional filtering step. In addition, users are already grouped according to factors which drive their interests or needs, thus allowing us to avoid costly user similarity calculations in situations which are time critical, such as on-the-fly recommendations.



Category	Venue Name
Arts & Entertainment::Stadium::Tennis	USTA Billie Jean King National Tennis Center
Professional & Other Places::Office	DXagency
Arts & Entertainment::Stadium::Tennis	Arthur Ashe Stadium
Shop & Service::Gym or Fitness Center	Equinox
Shop & Service::Gym or Fitness Center::Gym	New York Sports Club
Shop & Service::Gym or Fitness Center	Equinox
Shop & Service::Gym or Fitness Center::Gym	New York Sports Club
Shop & Service::Clothing Store	NikeTown
Shop & Service::Gym or Fitness Center::Gym	New York Sports Club
Professional & Other Places::Office	Definition 6: NYC
Professional & Other Places::Office	Razorfish NYC
Shop & Service::Department Store	Walmart Supercenter
Food::Coffee Shop	Starbucks
Great Outdoors::Park	Central Park
Shop & Service::Gym or Fitness Center::Gym	Train Daly

Figure 3. A cluster found by LDA using New York data that consists mainly of gyms (in blue), yet also including a sporting apparel store (red).

Venue Names
Century Cinema 16
AT&T Park
Stanford University
Cafe Borrone
Hacker Dojo
Googleplex
LinkedIn
Google San Francisco
Facebook
The Company Store
San Francisco Caltrain
Computer History Museum
Stanford Stadium
In-N-Out Burger
Mission Bay Conference Center
Apple Inc.
Microsoft SVC
Googleplex - 43
Googleplex - Charlie's Cafe
Plug And Play Tech Center
Stanford Shopping Center
Mountain View Caltrain

Table 6. A cluster in the San Francisco Bay Area data made up of check ins from what one would expect to be a tech event(s).

Another possible use of our model, if applied in an online form [1], would be to better understand the dynamics of check ins due to groups of users being in cities for various events. An example of a group formed by a possible event is observed in the data from San Francisco in Table 6. The cluster discovered is represented by a collection of places in the San Francisco Bay Area which relate strongly to famous technology sites across the city, transportation and conference centers. While we could not confirm it, we believe that this cluster is representative of a technology event which brought technology fans into the city, who in turn toured important sites in the field around the Bay Area.

The existence of a possibly fleeting group of users points to one clear limitation of our model - by ignoring temporal information in the data, we assume that groupings of users (and thus the factors affecting their check in behaviors) are heavily static, which is likely not the case. Topic models which consider temporal information, such as periodicity [25], may be able to garner interesting clusters over time. We also ignore temporal information with respect to sequences of user actions, a significant shortcoming of our model as compared to the user similarity model proposed in [10, 27]. At least five other limitations exist. First, the results of the model are not always interpretable - it is difficult, particularly if one is not familiar with the city in which the check ins occurred, to understand certain clusters. Foursquare categories help, but can only explain so much about the intricacies of user behavior.

A second problem of our model is that, in addition to being difficult to interpret, the resulting clusters are also not predictable - because the model is probabilistic, results can change slightly with each run. Furthermore, results are not predictable across cities, for example, we did not find a tourist cluster in San Francisco Bay Area. Third, the model is sensitive to the amount of data it is given. In noting that San Francisco Bay Area had much less data compared to New York, we also find that the clusters are not as well defined. This was also true for cities for which we had even less data, such as Pittsburgh and Chicago. One possible solution would be to incorporate other, similar data types, such as Yelp data. Fourth, in the text modeling domain, "stop words" are often removed, words such as "a" and "the" which are highly frequent. It might behoove a model of places to do the same. However, while it might make sense to remove uninteresting places such as airports and bus stations, it is unclear if popular places representative of interests, like stadiums, should really be removed. While we considered this avenue, we did not obtain rigorous findings in this direction. Finally, a hierarchical approach, as implemented in [10], would allow us to extend beyond the current categorizations.

CONCLUSIONS

The model we present is simplistic in the features of the data it incorporates. We group users into clusters based only on the places they go and thus do not incorporate explicit representations of geo-spatial, categorical or social aspects of our data. Because of this simplistic methodology, significant limitations, suggested above, exist in the practical usage of the model we present. Indeed, we do not discuss how such a model would compare to those which incorporate more rich features, and discuss how in some ways, previous user similarity metrics are more desirable than the one presented here. Future work, particularly those keen on understanding the applicability of our approach to recommendation technologies, should indeed incorporate a user study which allows for the comparison of a topic model approach, using various feature sets, to different mechanisms for recommendation.

However, the simplicity of our model allows us to generate a more data-driven understanding than has been previously explored of the latent factors which may be driving user check in behavior in data from location based social networks. Our findings confirm that geo-spatial and social homophily are powerful factors in grouping user into different types, interests and communities, thus supporting a large amount of work which suggests the same (e.g. [7, 4, 9, 13]). However, in addition to supporting previous work, we extend their efforts in two ways. First, we find that by typifying different users with a categorical, qualitative type such as "tourist", one can understand check in behavior beyond patterns in social, geo-spatial and venue categories. Second, for those groups which are in fact bonded by social and geo-spatial factors, our model allows for interpretation of the groupings beyond these variables to specific traits, such as homosexuality, which define a part of the community itself.

ACKNOWLEDGEMENTS

We would like to thank Justin Cranshaw for the data and for his assistance in developing the concepts presented. We would also like to thank Robert Kraut, Jason Hong and our

anonymous reviewers for their invaluable suggestions.

REFERENCES

1. L. Al Sumait, D. Barbara, and C. Domeniconi. On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, Proc. ICDM'08, pages 3–12. IEEE Computer Society.

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

3. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

4. C. Butts. *Space and Structure: Methods and Models for Large-Scale Interpersonal Networks*. Springer, 2012, expected.

5. Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring millions of footprints in location sharing services. In *Proc. ICWSM '11*, pages 81–88. AAAI, 2011.

6. E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proc. SIGKDD '11*, pages 1082–1090. ACM, 2011.

7. J. Cranshaw, R. Schwartz, J. I. Hong, and N. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proc. ICWSM '12*. AAAI, 2012.

8. J. Cranshaw and T. Yano. 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, 2010.

9. J. R. Hipp, R. W. Faris, and A. Boessen. Measuring neighborhood: Constructing network neighborhoods. *Social Networks*, 34(1):128 – 140, 2012. Capturing Context: Integrating Spatial and Social Network Analyses.

10. Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W. Ma. Mining user similarity based on location history. In *Proc. SIGSPATIAL '08*, GIS '08, pages 34:1–34:10. ACM, 2008.

11. J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman. I'm the mayor of my house: examining why people use foursquare - a social-driven location sharing application. In *Proc. CHI '11*, pages 2409–2418. ACM, 2011.

12. M. Loecher, D. Rosenberg, and T. Jebara. Citysensetm: Multiscale space time clustering of gps points and trajectories. In *Joint Statistical Modeling*, JSM '09, 2009.

13. M. McPherson, L. Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, (1):415–444, 2001.

14. S. Milgram. A Psychological Map of New York City. *American Scientist*, 60:194–200, Mar. 1972.

15. J. Moore. Building a recommendation engine, foursquare style, Mar. 2011.

16. A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: universal patterns in human urban mobility. *ArXiv e-prints*, Aug. 2011.

17. A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. In *Proc. ICWSM '11*, pages 570–573. AAAI, 2011.

18. R. Park and E. Burgess. *Introduction to the Science of Sociology*. University of Chicago, 1921.

19. M. T. Rivera, S. B. Soderstrom, and B. Uzzi. Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms. *Annual Review of Sociology*, 36(1):91–115, 2010.

20. A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *Proc. WSDM '12*, pages 723–732. ACM, 2012.

21. R. J. Sampson, S. W. Raudenbush, and F. Earls. Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277(5328):918–924, 1997.

22. C. Sato, S. Takeuchi, and N. Okude. Experience-based curiosity model: Curiosity extracting model regarding individual experiences of urban spaces. In A. Marcus, editor, *Design, User Experience, and Usability. Theory, Methods, Tools and Practice*, volume 6770 of *Lecture Notes in Computer Science*, pages 635–644. Springer Berlin / Heidelberg, 2011.

23. S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proc. SIGKDD '11*, pages 1046–1054. ACM, 2011.

24. S. Wakamiya, R. Lee, and K. Sumiya. Crowd-based urban characterization: extracting crowd behavioral patterns in urban areas from twitter. In *Proc. SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 77–84. ACM, 2011.

25. Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Lpta: A probabilistic model for latent periodic topic analysis. In *Proc. ICDM '11*, pages 904 –913, dec. 2011.

26. Y. Zheng. Location-Based social networks: Users. In Y. Zheng and X. Zhou, editors, *Computing with Spatial Trajectories*, pages 243–276. Springer New York, 2011.

27. Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W. Ma. Recommending friends and locations based on individual location history. *ACM Trans. Web*, 5(1):5:1–5:44, Feb. 2011.