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

Grant McKenzie*, Benjamin Adams**, and Krzysztof Janowicz*

Department of Geography*
National Center for Ecological Analysis and Synthesis (NCEAS)**
University of California, Santa Barbara
grant.mckenzie@geog.ucsb.edu, adams@nceas.ucsb.edu, jano@geog.ucsb.edu

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

**Keywords:** Location-based social networking, user similarity, topic modeling, diagnosticity

## 1 Introduction

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

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

The abundance of data published through online sources provides an exceptional foundation from which to investigate user similarity. To many users of these OSNs, the benefits of allowing access to this personal information is worth the cost of privacy. From a research perspective, these data offer an unprecedented opportunity to observe human behavior and design new methods for exploring the similarity between individuals. Studying similarity is important for several reasons. First, it can be used to suggest new contacts and thus, enrich the social network of a user. Second, as similar users are more likely to share similar interests, user similarities play a key role in recommender systems [12] and geographic information retrieval [6]. For instance, the *Last.fm* music platform offers social networking functions by which users can explore their *musical compatibility* with others and listen to their personalized radio stations. Third, and of most importance for our work, the information available about users, their locations, and activities is still sparse. User similarities can be exploited to predict *types* of activities and places preferred by a user based on those of users with similar preferences.

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

The remainder of the paper is organized as follows. In Section 2, we discuss related work on user similarity and location-based social networks. Section 3 focuses on data mining and the methods used for defining user similarity. In Section 4, we present results based on actual user data. Section 5 discusses a few of the limitations we faced in conducting this research and Section 6 presents our conclusions and points out directions for future work.

## 2 Related Work

Assessing user similarity has become an important topic in information retrieval and recommender systems over the past few years. The motivations for developing user similarity measures range considerably, from recommendation systems [2, 5] and dating sites [4] to location and activity prediction [10, 14].

A number of recent studies have focused on measuring user similarity through trajectory comparison [7, 9, 19]. In [7], Lee et al. explore a geometric approach to trajectory similarity by exploiting three types of distance measures in order to group trajectories. While their *Partition-and-Group* framework is unique, it is limited to the geospatial realm, overlooking the types of activities and social information related to the activity locations. Similarly, Li et al. [9] focused on the spatial components of user trajectories. Their method employs hierarchical trajectory sequence matching to determine similar users. Making use of GPS tracks, Li et al. extract *stay points* at which a user's activity is determined based on the affordances of a specific location.

While the above methods measure user similarity based on geospatial aspects of user trajectories, we argue that an understanding of the semantics of an activity space are essential. Ye et al. [18] investigate the concept of semantic annotations for venue categorization. In developing a semantic signature for a categorized place based on *check-in* behavior, similar, uncategorized places could be discovered. This concept of semantic signatures may also be applied to assessing user similarity through semantic trajectories. In this vein, [19], Ying et al. measured semantic similarity between user trajectories in order to developed a *friend recommendation system.* This work focuses on the type of activities completed by each user and the sequence in which these activities take place. Akin to the *stay point* work presented by Li et al. [9], the authors focus on *stay cells* and obtaining a semantic understanding of the types of activities conducted within the cells. From there, a semantic trajectory is formed and patterns are assessed and compared between users.

Activity prediction research can also benefit from exploring user similarity. Based on check-in data gathered through *Foursquare*, Noulas et al. [14] exploit factors such as transition between types of places, mobility flows between venues and spatial-temporal characteristics of user check-in patterns to build a supervised model for predicting a user's next check-in. This method, while exploring previous check-ins across users, does not assess similarity between users in predicting future locations, an aspect that our research suggests is beneficial. Traditional work in collaborative filtering (e.g., Amazon recommendations) has also focused on measuring user similarity, but typically concentrates on "structured" data such as numerical (star) ratings [11, 3].

Recently, Lee and Chung [8] presented a method for determining user similarity based on LBSN data. While the authors also made use of check-in information, they concentrated on the hierarchy location categories supplied by *Foursquare* in conjunction with the frequency of check-ins to determine a measure of similarity. By comparison, our approach is novel in that it makes use of an abundance of unstructured descriptive text (tips) provided by visitors of specific venues rather than a single categorical value.

## 3 Methodology

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

### 3.1 Data Source

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

**Venue Tips** An additional feature of *Foursquare*, is the ability for a user to contribute text-based *tips* to a venue. *Tips* consist of user input on a specific venue and can range from a restaurant review to a hiking recommendation. Lacking any official descriptive text for venues on *Foursquare*, these unstructured tips describe and define the venue and location. As with most crowdsourced data, the length, content, and number of tips vary significantly throughout the *Foursquare* venue data set. Of the 9 million *Foursquare* venues available in the continental United States, approximately 22.8% included at least one tip. Taking only venues that have had more than ten unique user check-ins, this value jumps to 54.0%. Of the venues to which our sample population checked in, 77.0% include at least one tip with the mean length of a single tip being 74 characters (stdev = 49.3). Table 1 shows a few examples of tips left at different venues.

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

**Table 1.** Example tips

### 3.2 Data Collection

Publicly geotagged *Foursquare* check-ins were accessed via the *Twitter API* for 6000 users over a period of 128 days. Check-ins to venues with less than ten tips were removed as well as users with an overall check-in count less than 16.

[1] https://foursquare.com/

This resulted in a dataset totaling 24,788 check-ins over 11,915 venues for 797 users (mean of 31.1 check-ins per user). From a geosocial perspective, we define an individual's activity identity as an amalgamation of the venues to which she checks in.

### 3.3 Themes

In this work, we use a Latent Dirichlet Allocation (LDA) topic model to extract a finite number of descriptive themes (topics) from the user-generated tips assigned to venues in our Foursquare dataset. While numerous topic models are discussed in the literature, LDA is a state-of-the art generative probabilistic topic model that can be used to infer the latent topics in a large textual corpus in an unsupervised manner [1]. A topic is a multinomial distribution over terms, where the distribution describes the probabilities that a topic will generate a specific word. LDA models each document as a mixture of these topics based on a Dirichlet distribution. Several mature implementations of LDA with improvements exist; for this work we employ the implementation in the MALLET toolkit [13].

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

**Venue Themes** Using this model we are able to express each venue as a mixture of a given number of topics. The model was tested with 40 topics at 2000 iterations. Future work could involve running similarity models with a varied number of topics. A few of the topics are concerned with a specific type of food, while others are focused on tourism and even baseball. Table 2 shows four examples of topics, based on top terms, extracted using LDA.

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

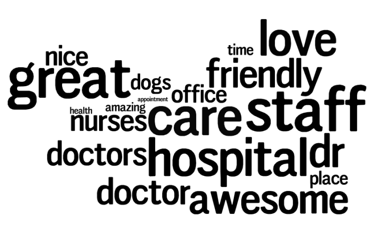
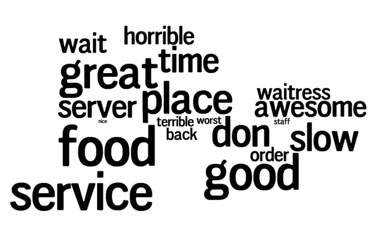
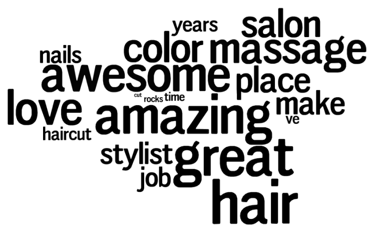
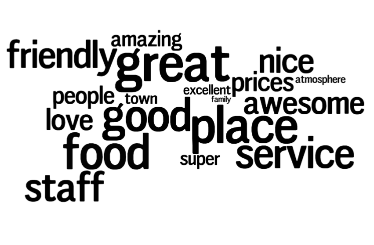
| High Entropy | Low Entropy |
|---|---|
|  nice<br>time<br>love<br>great<br>dogs<br>office<br>friendly<br>health<br>appointment<br>amazing<br>nurses<br>care<br>staff<br>doctors<br>hospital<br>dr<br>place<br>doctor<br>awesome |  wait<br>horrible<br>time<br>great<br>server<br>place<br>waitress<br>awesome<br>nice<br>terrible<br>worst<br>staff<br>back<br>food<br>don<br>slow<br>order<br>good<br>service |
|  years<br>salon<br>nails<br>color<br>massage<br>awesome<br>place<br>cut<br>rocks<br>time<br>love<br>amazing<br>make<br>ve<br>haircut<br>stylist<br>great<br>job<br>hair |  amazing<br>friendly<br>great<br>nice<br>people<br>town<br>excellent<br>prices<br>atmosphere<br>family<br>love<br>good<br>place<br>awesome<br>food<br>super<br>service<br>staff |

**Table 2.** Sample topics derived from Tip text represented as word clouds, where larger words are higher probability words for the topic.

A topic *signature* is computed for each of the collections via equation 1 where $T_i$ is one topic in the collective topic distribution, $n$ is the number of venues in the collection, $\#V_j$ is the number of times the same venue appears in the collection and $t_i^{Vj}$ is a single topic probability of Venue $j$. It is important to note that this method takes the frequency of check-ins to a unique venue into consideration. This ensures that multiple check-ins to a single location do not over-influence the topic distribution.

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

### 3.4 Variability vs. Commonality Weighting

This approach to calculating the topic signature for a collection of venues puts an equal amount of emphasis on all topics. This is not ideal when measuring the similarity between signatures as some topics are more prevalent across all venues than others. In order to augment the similarity model, we compute the entropy for each topic across all venues. In Table 2, two of the word clouds are examples of topics showing high entropy while the other two represent topics with low entropy.

Let $t_i$ be the weight of topic $t$ for venue $i$. A new discrete variable is defined for topics over venues by normalizing each $t_i$ to $t'_i$ by setting $t'_i = \frac{t_i}{\sum_{j=1}^{N} t_j}$, where $n$ is the number of venues. The topic's entropy over all venues, $E_T$, is defined in equation 2.

$$E_T = -\sum_{j=1}^{n} t'_j \log_2 t'_j. \tag{2}$$

Given this set of entropy values, a method for incorporating them as weights in a user similarity model must be assessed. This leads to questioning the role of topic prevalence in constructing a model for assessing user similarity. The approaches we present in the following subsections are influenced by literature in the cognitive sciences that examined the role of context (or framing) in human similarity assessments. Tversky [17] found that when two objects are compared for similarity, the set of objects from which the two objects are selected has the effect of making some properties more or less salient in the similarity judgment. The properties that are more salient are termed to be more 'diagnostic'. Tversky argued that two factors contribute to the *diagnosticity* of a property. The first is *variability*, which finds that the properties that vary across the elements of the context set are used more to determine the similarity (or dissimilarity) of two objects. The second factor *commonality*, is the opposite, that properties that are shared by most elements of the context set are the important properties, because they help explain what is important in the domain of discourse.

Although this context effect is well-studied in the cognitive sciences most computer science similarity measurements are without context in this sense. A notable exception is the *Matching-Distance Similarity Measure* (MDSM), created to compare similarity of spatial entity classes [16]. MDSM defines commonality and variability metrics for feature-based classes. In the following sections we adopt these notions to the venue topic signatures.

**Variability** One approach postulates that though the commonality topics remain critical in defining the venue (or user), they are less valuable in determining the similarities between two users. For example, if all venues in a dataset are high in a topic related to coffee, this topic does little in determining which two users are most similar. It is the less ubiquitous topics which are more *diagnostic* in the similarity model. Based on the literature on similarity [17], we call this type of diagnosticity, the *variability* weight.

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

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

The influence of entropy on topics using this commonality method involves taking the top ten topics with the lowest entropy and building our similarity model based purely on those topics. Again, the collective topic distribution is normalized in order to sum to 1.

### 3.5 Comparing Users

Since each aggregate venue signatures consist of a distribution over an equal number of topics, a divergence metric may be used to measure the similarity between our *focal user* and all other users at at any given activity. Using the *Jensen-Shannon divergence* ($JSD$) (Equation 3), we compute a dissimilarity metric between each user's topic distribution and the *focal user's* respective topic signature. $U1$ and $U2$ represent the topic signatures for User 1 and User 2 respectively, $M = \frac{1}{2}(U1 + U2)$ and $KLD(U1 \parallel M)$ and $KLD(U2 \parallel M)$ are *Kullback-Leibler divergences* as shown in Equation 4.

$$JSD(U1 \parallel U2) = \frac{1}{2}KLD(U1 \parallel M) + \frac{1}{2}KLD(U2 \parallel M) \tag{3}$$

$$KLD(P \parallel Q) = \sum_{i} P(i) \log_2 \frac{P(i)}{Q(i)} \tag{4}$$

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

## 4 Results & Discussion

Selecting a *focal user* at random from the 797 users, we first run the basic $JSD$ dissimilarity model without including an entropy weight. In order to keep the number of topics uniform across all models, a set of ten topics are randomly selected for comparison. Figure 1 shows the dissimilarity metrics at activity level resolution for 3 individuals compared to the *focal user.* As one can see, *User A's* similarity to the *focal user* generally decreases as the day progresses, with late evening proving to be the most similar time of day, *User B* is similar around lunchtime and quite dissimilar in the morning. Lastly, *User C* mirrors the average for most of the day with a small bump in the morning and a sharp peak of similarity at around 16:30.

In comparison, Figure 2 shows the effect of including the entropy measure with the purpose of emphasizing more diagnostic topics within the venue distributions. The same three users are compared to our *focal user*, but this time the venue distribution is composed of topics high in variability. The most visible outcome of the variability weight inclusion is an increase in range of similarity measures across users. Each of the three users is completely dissimilar to our *focal user* at some point during the day and the average dissimilarity across all users has increased.

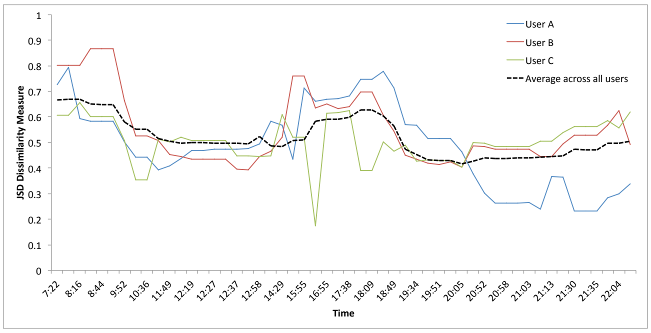


**Fig. 1.** Similarity of User A, B & C to Focal User (randomly selected topics)

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

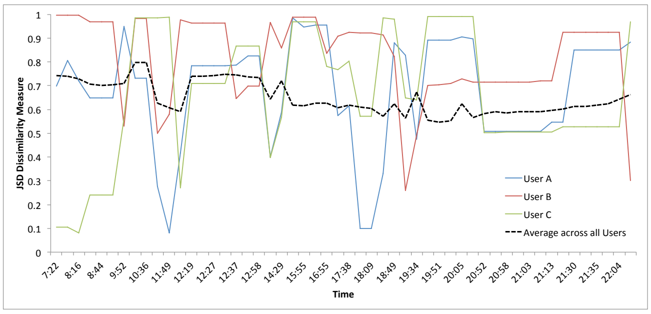


**Fig. 2.** Similarity of User A, B & C to Focal User (topics with highest entropy)

The *commonality* model offers a very different perspective. Figure 3 shows that on average, the similarity between all users and the *focal user* increased. While some semblance of the random-topics figure still exists, the users appear more uniform in their similarity to our *focal user*.

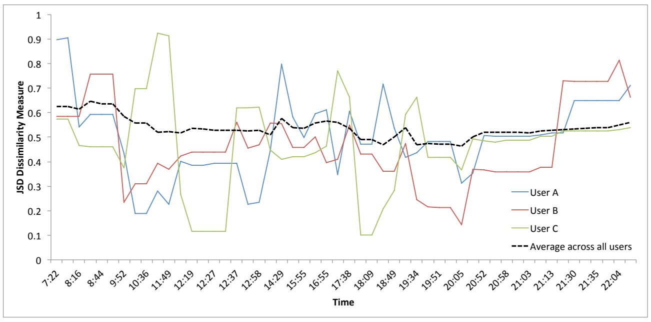


**Fig. 3.** Similarity of User A, B & C to Focal User (topics with lowest entropy)

### 4.1 Validating the model

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

To start, the topic distributions for the top-k most similar users for each check-in are combined using equation 5. The influence of each user on the combined topic distribution ($HV$) is calculated by multiplying the topic by the similarity value $sim$ where $sim = 1 - dissimilarity$. This ensures that more similar users have a larger impact on the overall topic signature.

$$HV_{T_i} = \sum_{j=1}^{n}((sim_j) * T_i^j) / \sum_{i=1}^{m} T_i \qquad (5)$$

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

The 30 sample venues are ranked in order of similarity to the *hypothetical venue* and the position of the real check-in venue within this ranked set is recorded. Figure 4 shows an example with graduated symbol markers representing the dissimilarity of each venue (large dark color = low dissimilarity). In this example, the top 5 most similar venues are labeled with the actual check-in venue resulting in 1 (the most similar venue to the *hypothetical venue*). This process is run across all check-ins for all users with the three levels of weighting. Table 3 shows an ordered-position table based on 3188 sampled check-ins over the 797 users in our dataset (4 randomly sampled check-ins per user). Both the 40 Topic model and the 30 Topic model are present in this table, showing the results for the *Variability*, *Commonality* and *No weight* models.

The *Commonality* weighted model produced the best results with over 77% of the *hypothetical venues* contributing to a correct estimation of the actual venue.

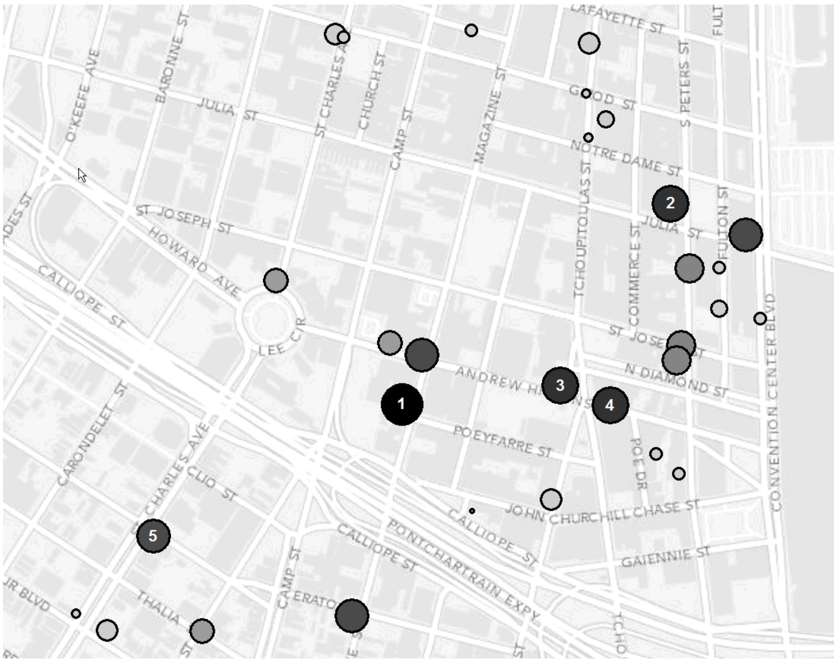


**Fig. 4.** Map fragment showing graduated symbols for the 30 nearest venues

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

## 5 Limitations

While the methods presented in this paper offer a promising approach to assessing user similarity through unstructured data, there are a number of limitations. Since the topic models are built on crowdsourced data (*tips*) from users of the application, the standard bias and errors of crowdsourcing are present. There is no way to ensure that a user submitting a tip has ever been to the venue or

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

**Table 3.** Placement of actual venue based on similarity to *Hypothetical Venue*

is offering a truthful tip. Additionally, since all tips for a single venue are combined in order to run the LDA model, those tips with more content have a large impact on the overall generation of topics. While there has been an increase in the number of people using LBSN applications, it should be noted that one's *Foursquare* check-in history does not account for every single activity that the user conducts throughout her day; the average user does not *check in* to every venue that she visits. It is more likely that a user checks in to locations that are unique or different from those to which she normally checks in. To some users, one venue might offer more social capital [15] than another (e.g., nightclub vs. hospital) and user's opinions range on what is *unique*. However, the limitations discussed here also hold for most other methods designed based on volunteered geographic information and are a research challenge.

## 6 Conclusion and Future Work

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

Future work in this area will flow in a number of directions. With an increase in the amount of user check-ins, the data will allow for further temporal factoring to reflect day of the week and month. It is expected that a user's activity patterns are not limited to hours within a day, but also reflects days of the week. The addition of temporal components will further enhance the ability of the model to discover similar users. Exploring the factors that contribute to this measure of user similarity will be a next step in this area of research as well. Analysis involving the correlation between location types and similarity measurements should be examined as well as outside factors that may contribute to similarity between users (e.g., demographic data, climate, etc).

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

## References

1. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
2. Ido Guy, Inbal Ronen, and Eric Wilcox. Do you know? recommending people to invite into your social network. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 77–86, 2009.
3. J.L. Herlocker, J.A. Konstan, L.G. Terveen, and J.T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

4. Günter J. Hitsch, Ali Hortaçsu, and Dan Ariely. Matching and sorting in online dating. *The American Economic Review*, 100(1):130–163, 2010.
5. T. Horozov, N. Narasimhan, and V Vasudevan. Using location for personalized poi recommendations in mobile environments. *SAINT*, page 124129, 2006.
6. C.B. Jones and R.S. Purves. Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3):219–228, 2008.
7. Jae-gil Lee, Jiawei Han, and K.-Y. Whang. Trajectory Clustering : A Partition-and-Group Framework . In *International Conference on Management of Data*, pages 593–604, 2007.
8. M Lee and C Chung. A user similarity calculation based on the location for social network services. *DASFAA*, pages 38–52, 2011.
9. Quannan Li, Yu Zheng, Xing Xie, Yukun Chen, Wenyu Liu, and Wei-Ying Ma. Mining user similarity based on location history. *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems - GIS '08*, page 34, 2008.
10. A. Lima and M. Musolesi. Spatial dissemination metrics for location-based social networks. In *UbiComp 2012*, 2012.
11. G. Linden, B. Smith, and J. York. Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80, 2003.
12. C. Matyas and C. Schlieder. A spatial user similarity measure for geographic recommender systems. *GeoSpatial Semantics*, pages 122–139, 2009.
13. Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.
14. Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. International conference on data mining. In *Mining User Mobility Features for Next Place Prediction in Location-based Services*, 2012.
15. Edward Pultar, Stephan Winter, and Martin Raubal. Location-based social network capital. In *GIScience, Extended Abstracts*, 2010.
16. M. Andrea Rodriguez and Max J. Egenhofer. Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. *International Journal of Geographical Information Science*, 18(3):229–256, 2004.
17. A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.
18. Mau Ye, Dong Shou, Wang chien Lee, Peifeng Yin, and Krzysztof Janowicz. On the semantic annotation of places in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge discovery and data mining*, pages 520–528, 2011.
19. Josh Jia-Ching Ying, Eric Hsueh-Chan Lu, Wang-Chien Lee, Tz-Chiao Weng, and Vincent S. Tseng. Mining user similarity from semantic trajectories. *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks - LBSN '10*, pages 19–26, 2010.