# Community-Based Geospatial Tag Estimation

Wei Niu, James Caverlee, Haokai Lu, Krishna Kamath
Department of Computer Science and Engineering
Texas A&M University
College Station, Texas, USA
{wei,caverlee,hlu,kykamath}@cse.tamu.edu

***Abstract*—This paper tackles the geospatial tag estimation problem, which is of critical importance for location-based search, retrieval, and mining applications. However, tag estimation is challenging due to massive sparsity, uncertainty in the tags actually used, as well as diversity across locations and times. Toward overcoming these challenges, we propose a community-based smoothing approach that seeks to uncover hidden conceptual communities which link multiple related locations by their common interests in addition to their proximity. Through extensive experiments over a sample of millions of geo-tagged Twitter posts, we demonstrate the effectiveness of the smoothing approach and validate the intuition that geo-locations have the tendency to share similar "ideas" in the formation of conceptual communities.**

## I. Introduction

The sharing of fine-grained geospatial footprints through smartphones and social media services (e.g., Facebook, Twitter) promises new insights into the dynamics of human behavior and new intelligent location-aware applications. Already, we have seen many research efforts aimed at enhancing web and social media search by integrating location signals [1], [2] and identifying "living neighborhoods" [3], [4]. Many of these scenarios are driven by social media posts that include latitude-longitude coordinates and a timestamp, often including a user-generated tag that provides additional context (e.g,. #sunset, #beach, #happy).

A critical challenge for search, retrieval, and mining applications built on these geo-tagged posts is accurate identification of popular tags and estimation of the tag distribution for a particular place of interest. For example, a real-time local search system may want to reflect what is currently "hot" in the area of interest (e.g., reaction to a political debate). Similarly, a geo-enabled advertising system may place targeted social media ads based on local interests (e.g., advertising jerseys during a football match). Unfortunately many locations may have no (or little) evidence of tags due to the sparsity of geo-enabled social media. Indeed, recent estimates suggest that only 1.5% of tweets have geo-coordinates [5]. And even for locations that are well-represented, there is inherent uncertainty around using social media to capture the overall distribution of interests in a particular location. Moreover, tags are diverse from location to location and can temporally fluctuate in popularity.

Toward overcoming these challenges, this paper tackles the *geospatial tag distribution estimation problem* for accurately estimating the tags in a particular location. The main intuition of the proposed approach is to "smooth" a location's tag distribution from a larger "conceptual" community in which a particular location belongs to, as well as geographically contiguous neighbor locations. Evidence of *homophily* in social networks [6] – wherein individuals tend to associate and bond with similar others – motivates our hypothesis that geo-locations (as comprised by individuals) may also share a similar tendency in the "ideas" (or hashtags) that are shared. Thus, it may be possible to rely on the denser (and richer) information from a larger community to shed light on individual locations, potentially alleviating the challenges of tag estimation. Concretely, we propose a community discovery framework and an approach for estimating the tags in a particular location. Through this framework, we investigate:

- The representation of locations for community discovery – both through the distribution of hashtags adopted and by how rapidly they are adopted;
- Methods for measuring the conceptual distance between locations – by two content-based methods and an adoption time metric, plus variations integrating physical distance;
- Methods for estimating the unknown hashtag distribution – through a novel community and neighbor-based smoothing method.

Through extensive experiments based on 100s of millions of geo-tagged Twitter posts, we demonstrate the effectiveness of conceptual community-based smoothing. We investigate the impact on precision and recall of multiple estimators, examine the impact of the number of communities on tag distribution estimation, and find an improvement of the proposed approach versus a neighbor-based baseline. These findings are encouraging for improving the quality and coverage of geospatial tag estimation, which is an important step toward providing intelligent location-aware search and recommendation systems.

## II. Related Work

Recently, there has been a rise of research aimed at location-revealing social networks like Facebook and Twitter [7]. Some works have represented locations as a bag-of-tags of geo-tagged photos [8] or checkins and venue features [4]. Cranshaw et al. present several methods for comparing cities as vectors of venue categories and then using hierarchical clustering to find similar cities [9]. Spectral clustering is used to detect geospatial neighborhood-like communities or similar locations based on user activity patterns or venue

IEEE/ACM ASONAM 2016, August 18-21, 2016, San Francisco, USA
978-1-5090-2846-7/16/$31.00 

features [3], [10]. In contrast, this paper uses tag distribution information to represent locations for uncovering hidden conceptual communities [11]. That is, we do not require users to necessarily check-in in one place and then another; instead, we aim to uncover locations that are tied by common topic interests at multiple granularities and potentially across physical proximity boundaries.

Another research thread is aimed at studying tags in social media as a form of user-generated metadata that can provide rich context for the tagged object. Example tags include Flickr-like tags on images and Twitter-style tags on text posts. Much effort has been devoted to tag recommendation [12], tag ranking [13], and tag enrichment [14]. In one direction, hashtags can be recommended to users to encourage additional appropriate tagging; much of this work has focused on content similarity between user's tweets and an existing hashtag topic [15], [16]. In another direction, researchers have studied the overall temporal and geospatial distribution of hashtags as they diffuse [17]. For example, there has been research aimed at predicting hashtag popularity from temporal and geospatial perspectives [18]. The work presented here – where geographic patterns of hashtag use uncover conceptual communities – could inform efforts on tag recommendation and geographic tag diffusion.

There is also recent work on estimation to overcome sparse data in social networks, such as social network user home location prediction [19], [20], [21]. To alleviate data sparsity [22], for example, researchers have proposed to use transfer learning in collaborative filtering [23]. Another widely used approach in language modeling for adjusting the maximum likelihood estimator to compensate for data sparseness is smoothing [24], [25]. While in our study, we use the community distribution to smooth over the target location, where the hashtag information is unknown, to deal with data sparsity and to focus on the overall correctness of estimating the hashtag distribution.

## III. Geospatial Tag Distribution Estimation

We begin by assuming that we have a collection of social media posts that include latitude-longitude coordinates, a timestamp and one or more *tags* (or *hashtags*; in the following we shall use hashtags). We view a tag on a post as a tuple $(h_i, t_i, g_i)$, where $h_i \in \mathcal{H}$ is the hashtag, $t_i$ is the timestamp, and $g_i$ is a latitude-longitude coordinate. We assume there is a mapping function that converts latitude-longitude coordinates into *locations*. A location here could correspond to a particular venue (e.g., a restaurant), a city block, equal or variable-sized grid cells, or some other domain-specific collection of latitude-longitude coordinates. We denote $\mathcal{L}$ as the set of all possible locations and $\mathcal{H}$ as the set of all distinct hashtags. Hence, we finally view a tagging action as the hashtag text, the timestamp, and its location: $(h_i, t_i, l_i)$. In total, we call this collection of tuples $\mathcal{P}$ (for posts).

The *hashtag distribution at a location $l$* is the probability distribution of all the hashtag occurrences at that location, which we denote as $\theta_l$:

$$\theta_l = \{[h_1, p(l, h_1)], [h_2, p(l, h_2)], \cdots, [h_m, p(l, h_m)]\}$$

where $p(l, h_i)$ is the probability of a hashtag $h_i$ in location $l$. In many scenarios, this distribution is unknown or incomplete due to data sparsity, uncertainty, and other factors. Hence, the *hashtag distribution estimation problem* is to find an estimated hashtag distribution $\widetilde{\theta}_l$ that matches the actual (but, unknown) distribution $\theta_l$ as closely as possible (where measures of closeness are defined in the experiments).

We propose to tackle the hashtag distribution estimation problem by considering multiple sources of evidence – the conceptual community which the location is associated with and the contiguous neighboring areas around a location – and by smoothing a location's tag distribution by incorporating evidence of both the conceptual community in which a particular location belongs to as well as neighboring locations. The key intuition of community-based smoothing estimation is that if a hashtag appears in a community $c$, which is a group of locations drawn from $\mathcal{L}$, then it is likely to appear in each individual location $l_i$, where $l_i \in c$.

In the following we address two key challenges to robust hashtag distribution estimation: (1) How do we identify these conceptual communities in the first place? and (2) Given a set of communities, how do we estimate the hashtag distribution for a particular location?

### A. Challenge 1: Finding Conceptual Communities

A *conceptual community* links multiple related locations by their common interests rather than by their geographic proximity (that is, without the constraint of being geographically contiguous). We define the *hidden conceptual community discovery problem* as: given a collection $\mathcal{P}$ of social media posts, identify the set of conceptual communities $\mathcal{C} = \{c_1, c_2, \cdots, c_k\}$, where each community $c_i \in \mathcal{C}$ is a group of locations drawn from $\mathcal{L}$. Toward discovering these conceptual communities, we investigate in this paper a clustering-based approach. In our preliminary research, we have tested several clustering algorithms – including spectral clustering [26], hierarchical clustering [27], Affinity Propagation [28], etc. We find that while there are qualitative differences in the clusters generated, that the k-means++ algorithm [29] provides a nice balance of efficiency and cluster quality (as measured in the experiments section). Hence, we focus our discussion here on k-means++, which operates by finding partitions of $n$ locations into $k$ clusters (corresponding to our hidden conceptual communities). It provides a way for choosing the initial centers for k-means instead of random assignment which often lead to poor performance. Some clustering approaches have an embedded method for determining k, while in our case, we prefer a flexible k value since the number of communities is impacted by the performance of tag distribution estimation.

In the following, we tackle two important questions – (i) what is the appropriate representation for a location as a basic unit of community discovery? and (ii) how can we compare two locations, to determine a meaningful notion of conceptual distance for use in identifying related locations?

**Representing a Location.** We consider two methods for representing a location: by hashtag frequency and by hashtag adoption time.

*By Hashtag Frequency.* This approach captures the overall distribution of "ideas" associated with each location by considering the relative frequency distribution of hashtags observed at that location. Formally, we can represent each location *l* by its hashtag frequency :

$$l = \{[h_1, \mathcal{F}(l, h_1)], [h_2, \mathcal{F}(l, h_2)], \cdots, [h_n, \mathcal{F}(l, h_n)]\}$$

where $\mathcal{F}(l, h_n)$ represent the number of occurence, or we say frequency of hashtag $h_n$ at location *l*.

*By Hashtag Adoption Time.* While the previous approach may meaningfully capture the overall interests of the location, it ignores the timing information embedded in the timestamps of each hashtag. Hence, this approach intended to capture these differences by considering the first occurrence time of each hashtag *h* at location *l*, $\tau(l, h)$ and a location can then be represented as:

$$l = \{[h_1, \tau(l, h_1)], [h_2, \tau(l, h_2)], \cdots, [h_n, \tau(l, h_n)]\}$$

Unlike hashtag frequency, the adoption time implicitly captures the *influence* of a location, which is helpful for distinguishing between locations with similar topics but different relative importance (as measured by adoption time of ideas).

**Measuring Distance Between Locations.** Most existing neighborhood finding approaches require nearby (geographic-distance constrained) locations [4], so that the resulting regions are spatially bounded in a single contiguous region. In many applications this is a reasonable assumption, since contiguous regions provide the basis of many real-world ways in which we organize space (e.g., state boundaries, congressional districts). Alternatively, we explore in this section methods for linking locations based on the conceptual distance between them, so that related locations may be connected regardless of geographic distance.

First we consider *content similarity* to characterize the location similarity. We apply well-known Jaccard similarity $\mathcal{S}im_{Jaccard}$ and cosine similarity $\mathcal{S}im_{cos}$ over each pair of locations vectors represented with hashtag frequency. The corresponding distance $\mathcal{D}_{Jaccard}$ and $\mathcal{D}_{cos}$ is simply defined as the inverse of the similarity. Additionally, we propose *temporal distance*, which based on the adoption time of hashtags at different locations. We can measure the adoption time difference between two locations $l_1$ and $l_2$ as:

$$T_a = \frac{1}{\|H(l_1) \cap H(l_2)\|} \sum_{h_i \in H(l_1) \cap H(l_2)} \tau_i^{l_1} - \tau_i^{l_2}$$

which can be considered a measure of relative influence between a pair of locations. $\tau_i^l$represent the first occurrence time of hashtag $h_i$ at location $l$. If we take absolute value over time difference $\tau_i^{l_1} - \tau_i^{l_2}$ in the above formula, then it measures the average hashtag adoption time between a pair of locations. We denote it as $\mathcal{D}_t$. When $\mathcal{D}_t$ is small for two locations, they are more closely related to each other, otherwise, we consider they are far apart. Note however, the adoption time approach does have the drawback of only considering pairwise influence, and so it may miss cases in which a third location is influencing the two locations of interest.

*Integrating Geographical Distance.* Finally, we augment each of the baseline distance metrics with a geographic distance-based damping factor. The idea is that we can provide a tunable parameter for biasing the conceptual distance between two locations by additionally considering the geographic distance between them. There are mainly two reasons. First, our intuition is drawn from the first law of geography, which states that "Everything is related to everything else, but near things are more related than distant things" [30]. Second, we aim to mitigate the impact of data sparsity on clustering. For example, if we only collected one hashtag "sports" at a particular location, then this location by content similarity, may be clustered with locations that have high frequency of "sports". To avoid degenerate clusters like that, we incorporate a factor damping with distance. We begin by considering the Haversine distance between two locations – where the Haversine distance $D_H$ [31] measures the shortest distance over the earth's surface between two points. We update each of the distance metrics by incorporating this Haversine distance, as shown in Table I. In each case, $\alpha$ is a user-defined distance decay coefficient. Larger values of $\alpha$ will increase the distance between geographically far apart locations, making them less likely to be grouped together.

### *B. Challenge 2: Hashtag Distribution Estimation*

Given a community *c* and a target location $l \in c$, we further investigate how to estimate the hashtag distribution for location *l*. In this section, we consider three approaches: (i) the first considers the conceptual community (the idea-based neighborhood) around a location; (ii) the second considers immediate geographic neighbors (ignoring the conceptual distance of these neighbors, as well as the potential conceptual closeness of more distant neighbors); and (iii) a hybrid approach which seeks to balance both conceptual and geographic distance.

**Community-Based Estimation.** In this first approach, we estimate the hashtag probability at location $l$ according to the hashtag probability at the conceptual community that $l$ belongs to:

$$\widetilde{p}(l, h) = \mathcal{F}(\mathcal{C}, h) / \sum\nolimits_i \mathcal{F}(\mathcal{C}, h_i)$$

where $\mathcal{F}(\mathcal{C}, h)$ is the frequency of hashtag $h$ in the community $\mathcal{C}$ that $l$ belongs to. The intuition is the hashtag distribution over the community tends to be coherent from location to location. Thus it emphasizes the overall hashtag popularity in the community, resulting in hashtags with high frequency being good candidates for the target location.

**Neighbor-Based Estimation.** In the second approach, we ignore the conceptual closeness of distant communities in favor of considering neighboring locations as sources of similar

| Distance | $\mathcal{D}_{\mathbf{Jaccard}}{}^*(l_1, l_2)$ | $\mathcal{D}_{\mathbf{cos}}{}^*(l_1, l_2)$ | $\mathcal{D}_{\mathbf{t}}{}^*(l_1, l_2)$ |
|---|---|---|---|
| | $\frac{1}{Sim_{Jaccard}}\alpha^{D_H(l_1,l_2)}$ | $\frac{1}{Sim_{cos}}\alpha^{D_H(l_1,l_2)}$ | $D_t\alpha^{D_H(l_1,l_2)}$ |

TABLE I: Distance metrics that integrate geographical distance

hashtag evidence. This method estimates the probability of a hashtag at location $l$ according to the aggregated hashtag distribution of neighboring locations that border with the target location $l$. The expectation is that locations contiguous with target location share a large portion of the hashtag as what people see and experience tend to be similar. That is, we have:

$$\widetilde{p}(l,h) = \mathcal{F}(\mathcal{N},h)/\sum\nolimits_i \mathcal{F}(\mathcal{N},h_i)$$

Where $\mathcal{N}$ represents the locations that border with $l$.

**Hybrid Approach.** Finally, we propose a smoothing approach that seeks to balance both community-based and neighbor-based estimation. The intuition is that neither approach in isolation is best for estimating the hashtags of a location; rather, we should adopt a more flexible model to integrate both sources of evidence that can vary from location to location.

Building on the previous two estimations approaches, the hybrid approach is:

$$\widetilde{p}(l,h) = \beta \cdot p(l_{community},h) + (1-\beta)\cdot p(l_{neighbors},h)$$

where $\beta$ is a weight equals to $N_d(h)/N_{max}$. $N_d(h)$ is the total number of distinct hashtags of target location in the dataset used for community discovery – and $N_{max}$ – the maximum number of distinct hashtags one location has in the same dataset. When $N_d(h)$ is large, neighboring locations may only share part of the hashtags with the target location due to the difference in hashtag density. Meanwhile, the hashtag in the target location is more likely to be prevalent in the community, since we expect the community contains some similar locations of the same scale. We can rely more on community distribution as a source of supplementary and richer information for better estimation. For example, for a location like New York City, the aggregated hashtag distribution of nearby locations is not sufficient to represent the hashtag distribution of the urban area. Alternatively, the community which also contains Boston and D.C. may provide more accurate hashtag information. When $N_d(h)$ is small, the target location is less influential in the community and meanwhile, it tends to be well represented by the neighboring locations. Thus more weight should be placed on neighboring distributions. To conclude, we increase $\beta$ when $N_d(h)$ is large and decrease $\beta$ as $N_d(h)$ is small.

## IV. Experimental Evaluation

In this section, we present a set of experiments to show the discovered conceptual communities over different granularities and investigate the strengths and weaknesses of hashtag distribution estimation.

### A. Dataset

All of our experiments are over a collection of geo-tagged social media posts sampled from Twitter. We collected 324 million hashtags from tweets that were annotated with a latitude-longitude coordinate over the course of two years (from February 2011 to March 2013). We crawled the dataset using the Twitter streaming API using a bounding box to only gather tweets from particular parts of the world. We use the Universal Transverse Mercator (UTM) [32] geographic coordinate system to grid the area of interest into homolographic subareas, each of which corresponds to a location.

### B. Evaluating Hashtag Distribution Estimation

To evaluate the quality of an estimated hashtag distribution, we consider weighted versions of precision, recall, and Jensen-Shannon Divergence [33].

*Weighted Precision.* We consider a weighted version of precision which considers the probability of a hashtag:

$$\mathcal{P}@n = \frac{\sum_n r(h_n)\widetilde{p}(l,h_n)}{\sum_n \widetilde{p}(l,h_n)}$$

where $\widetilde{p}(l,h)$ is the probability of occurrence of hashtag $h$ in the estimated distribution, and $r(h)$ is an indicator variable that is 1 when the estimated distribution contains $h$, and 0 otherwise. $\mathcal{P}@n$ is the percentage of first $n$ estimated hashtags that actually exist at target location. These hashtags are weighted by their probability in the estimated distribution. It is a measurement of relatedness of estimated hashtags. However, it only considers the estimated distribution and ignores the actual distribution.

*Weighted Recall.* The weighted recall is defined as:

$$\mathcal{R}@n = \frac{\sum_n r(h_n)\widetilde{p}(l,h_n)}{\sum_m \widetilde{p}(l,h^*_m)}$$

where $h^*$ represents all the hashtags in the actual distribution. $\mathcal{R}@n$ is the percentage of the actual hashtags that is covered by the estimated distribution. Hashtags are weighted by the probability in the estimated distribution. It is a measurement of how completely the estimated hashtags match the actual hashtags.

We use $P@n$ and $R@n$ as preliminary performance metrics and for the real distribution matching, we rely on Jensen-Shannon Divergence.

*Jensen-Shannon Divergence (JSD)*: JSD measures the similarity between two distributions. The JSD between two hashtag distribution is defined as:

$$JSD(\theta_m,\theta_n) = \frac{1}{2}\sum_i \ln(\frac{p(m,h_i)}{\bar{p}(h_i)})\cdot p(m,h_i) + \frac{1}{2}\sum_j \ln(\frac{p(n,h_j)}{\bar{p}(h_j)})\cdot p(n,h_j)$$

where $p(m,h)$ represents the probability of hashtag $h$ in distribution $m$ and $\bar{p}(h) = \frac{1}{2}(p(m,h) + p(n,h))$. Smaller JSD values indicate the two distributions are more similar.

### C. Conceptual Community Discovery

Here, we show some example patterns of the discovered communities. Since the community discovery is randomized, we present a typical output for each target area. We postpone the discussion of picking the number of communities.

We considered four different geo-spatial granularity with respect to the different areas. The detailed statistics are listed in Table II, for example, at a global-scale, we have tested a location with an accuracy of 100 km which correspond to an area of approximately $10^4 km^2$. The total number of hashtags being considered in this case is 324 million, and the number of distinct hashtags is 298,000.

| Target area | World | US | NYC | MHTN |
|---|---|---|---|---|
| Area ($km^2$) | $10^4$ | 2500 | 1 | $10^{-2}$ |
| No. of locations | 2,565 | 2,482 | 1,445 | 5,529 |
| Hashtag frequency | $324m$ | $29m$ | $3m$ | $0.8m$ |
| Distinct hashtags | $298k$ | $98k$ | $45k$ | $7k$ |

TABLE II: Four social media collections

Figure 1 shows us the 13 communities in the world using $\mathcal{D}_{cos^*}$. We observe a language-based pattern: for example, Brazil and Portugal are in the same community due to the common language. Overall, we observe that language (and culture) is the dominating factor for identifying communities based on content similarity.

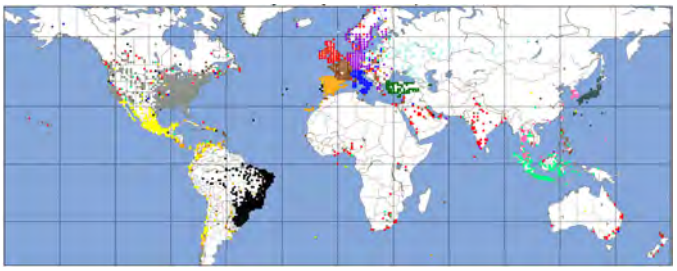


Fig. 1: (Color) World by Content Similarity (Cosine)

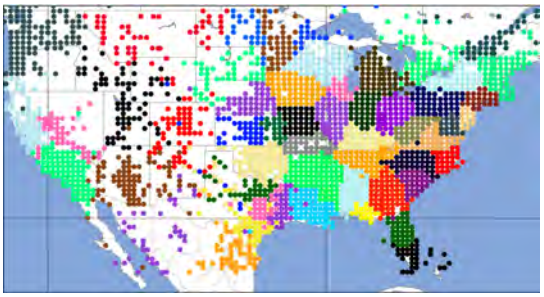
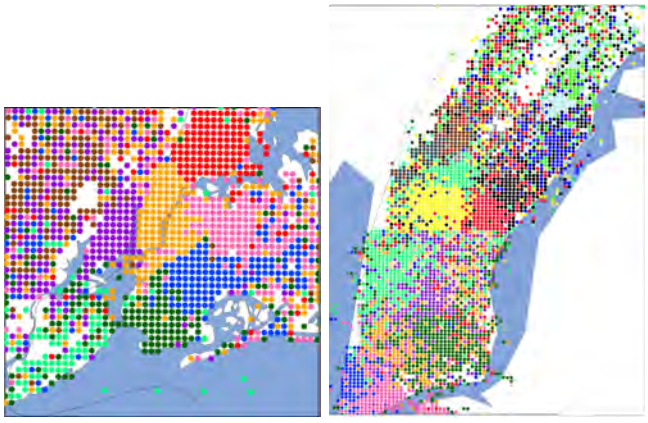


Fig. 2: (Color) US by Content Similarity (Cosine)

Moving to a more focused context, we next consider the communities discovered in the continental US with $\mathcal{D}_{cos^*}$ in Figure 2. The 60 communities closely match state contours and metropolitan areas: we can see Chicago area, New York and New Jersey area, Rocky Mountains and many more. In some cases, communities span borders: for example, Arizona and part of New Mexico are in the same community. These results indicate the strength of political boundaries on the "ideas" shared via social media, supporting the argument that culture is strongly impacted by political organization.

We next examine New York City. Here, Figure 3a show the discovered 8 communities in New York City by $\mathcal{D}_{cos^*}$. We see a clear delineation of boroughs and neighborhoods, where the contour is consistent with the race and ethnicity map of NYC according to 2010 Census data[1].



(a) (Color) NYC (b) (Color) Manhattan

Fig. 3: NYC and Manhattan by Content Similarity(Cosine)

Finally, we turn to the most narrowly scoped target area: Manhattan. We again look at cosine-based approaches. Figures 3b show the 22 communities discovered in Manhattan. In this figure, we can notice the neighborhood-level communities. As our data at this scale is very sparse, we notice a more scattered community pattern. These community covers blocks where people may hear the same news, see similar events, and share a similar lifestyle.

In summary, we find that locations at different granularities can be grouped into coherent communities based on the approaches defined above. At the world level, language is the dominating factor that influences community discovery. At country level, we see that the communities we find are influenced mostly by the cultural identity, as well as geographic reasons. At more fine-grained levels, we attribute the different communities to demographics and everyday activities. We see that Jaccard and cosine tend to discover similar communities, whereas adoption time provides a different flavor of finding communities. It is challenging to make a direct and fair comparison with existing neighborhood detection work mentioned in Section II due to the feature and goal differences. Thus we will further demonstrate the quality of these communities via the geo-spatial tag distribution estimation task in the following.

[1]https://www.flickr.com/photos/walkingsf/5559914315/

### D. Evaluating the Community-based Approach

We now turn to investigating the quality of conceptual community-based tag estimation, before turning in the following section to comparing the community-based estimator versus a neighbor-based one and a hybrid method. We evaluate the quality of hashtag distribution estimation over the 2,482 US locations based on three proposed community discovery approaches – by Jaccard, by cosine, by temporal distance. As a baseline, we consider a physical community that less than 200 miles in Haversine distance (HD) from target location; that is, the communities are necessarily linked only by distance and not by any tag-related information. For the conceptual community-based estimators, we find initial communities based on 50% of the dataset, then use the estimation methods to infer the estimated distribution for each location with the rest of the data, which is then compared with the actual distribution of the corresponding location for evaluation. This two-fold cross validation for distribution estimation is different from classification, in that we try to keep a balanced partition and avoid estimation data being too sparse, which won't reflect the true estimation performance. We consider all locations and report averages. In our initial experiments we consider $k = 90$ communities and then test the impact of varying $k$.

Interestingly, we see the precision and recall for the four approaches in Figure 4. The x-axis in all cases corresponds to the number of hashtags being considered, ranked by decreasing occurrence probability. In all cases, we see that the two content-based community discovery methods – Jaccard and cosine – outperform the temporal distance. We see that HD results in the best precision and recall, followed by cosine, then Jaccard, and finally temporal distance. These results indicate the strong locality effects of hashtag adoption as postulated by Tobler's first law of geography – in that communities composed of nearby locations may share more common "ideas" than those composed of distant locations. But can the conceptual communities complement this strong locality impact?

**Integrating Haversine Distance.** Next, we consider the impact on hashtag distribution estimation when we integrate Haversine distance into the conceptual distance; does forcing communities to be more compact improve the quality of hashtag estimation? Following experimental results presented in [34], we adopt a distance decaying coefficient $\alpha = 1.01$. In Figure 5, we report the precision and recall when the communities are discovered using the Haversine-integrated approach. We do see an improvement on the absolute performance: for example, the increase in precision and recall are especially apparent for Jaccard ($\mathcal{D}_j$) and the temporal distance ($\mathcal{D}_t$). Overall, we find using the Cosine+HD approach ($\mathcal{D}^*_{cos}$) yields the best precision and recall. So yes, integrating the strong locality of nearby locations with more distant conceptual communities can positively impact hashtag distribution estimation. These results confirm the importance of carefully identifying these distant conceptual communities, and of integrating them into more naive distance-based approaches.

**Varying the Number of Communities.** So far, we have considered a fixed number of communities. But what impact does varying the number of communities have on the hashtag estimation problem? Indeed, there are existing methods that aim to find a proper number of clusters $k$ for k-means [35]. But here, we would like to consider an application driven strategy to decide the number of communities, i.e. by the performance of estimation. Many small communities may favor precision in the hashtags that locations share, but without the overall perspective (and the additional hashtags) that may be present in a few larger communities that are composed of many locations. Conversely, when the number of communities is small, the hashtags shared by its constituent locations may be overly broad, resulting in poor estimation for a specific target location. Hence, in this experiment we vary the number of communities from 30 up to 240 using the Cosine+HD metric ($\mathcal{D}^*_{cos}$). Figure 6 presents the F1 score to characterize the performance: for example, in the 120 communities case, which has relatively smaller community size, the precision and recall are better than the cases of 30 or 60 communities. So, more smaller communities tend to result in better estimation. However, as the number of communities increases (and the size of each community decreases), we see that the quality of estimation actually degrades. For example, at 240 communities there is worse precision and recall than in the 120 and 180 community case. As the community size shrinks, the hashtags in each community becomes more fragmentary toward representing the target location, thus dragging the precision and overall performance down. These results demonstrate the importance of incorporating location (and domain) specific models of what constitutes a community, so as to balance the finer granularity of small communities with the richness inherent in larger ones.

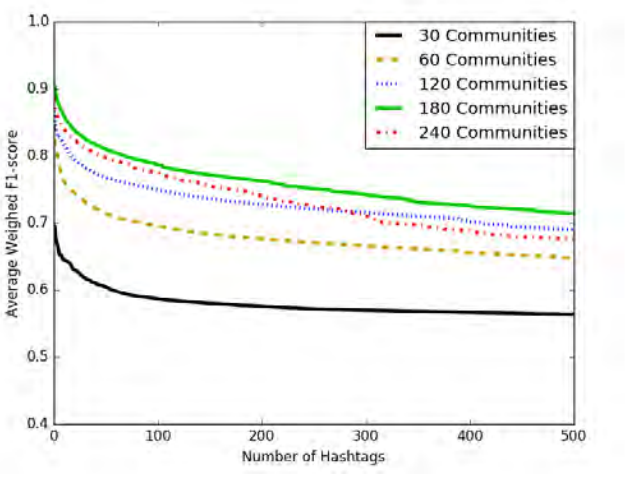



Fig. 6: Varying the Number of Communities

We further measure the distributional similarity between the estimated hashtag distribution and the actual distribution using Jensen-Shannon divergence in order to compare various distance metrics used for community discovery. For each

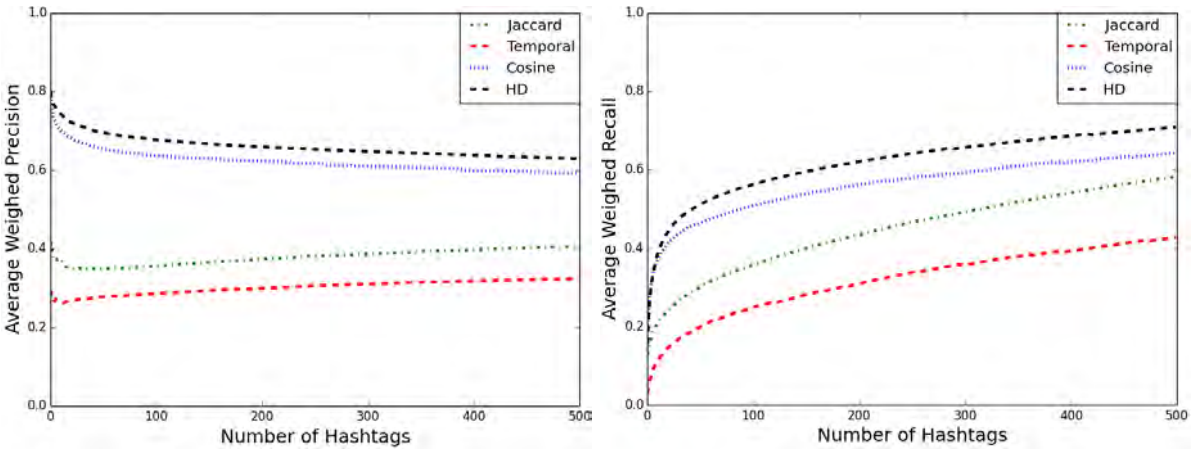



Fig. 4: Average Weighted Precision and Recall

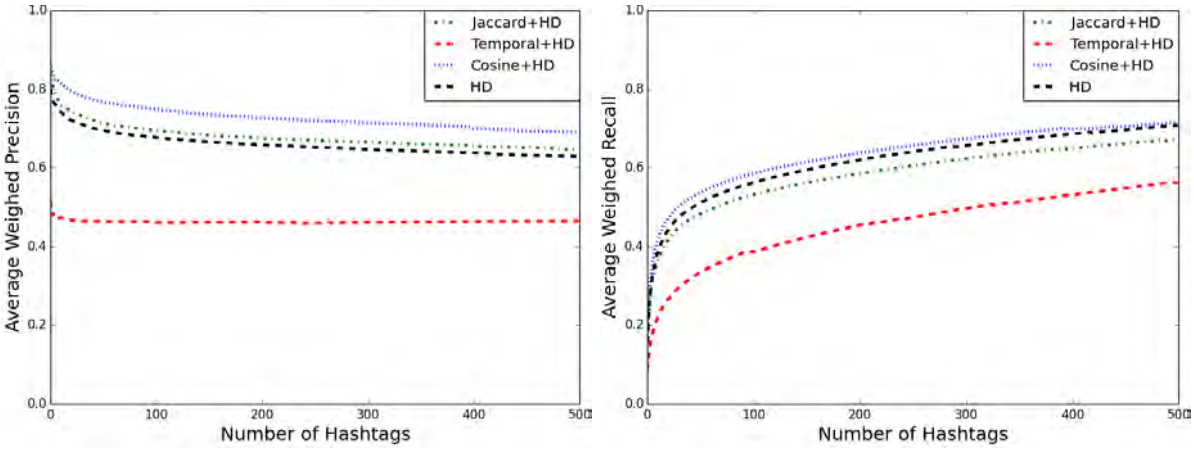



Fig. 5: Average Weighted Precision and Recall (Integrate Haversine Distance)

| Distance metrics | 10% | 20% | 50% | 100% |
|---|---|---|---|---|
| Cosine + HD | 0.632 | 0.581 | 0.494 | 0.353 |
| Jaccard + HD | 0.670 | 0.632 | 0.551 | 0.380 |
| Temporal + HD | 0.676 | 0.648 | 0.589 | 0.453 |

TABLE III: Average Jensen-Shannon Divergence @P% for Community-based Estimation

| Distance metrics | 10% | 20% | 50% | 100% |
|---|---|---|---|---|
| CB | 0.632 | 0.581 | 0.494 | 0.353 |
| NB | 0.628 | 0.583 | 0.489 | 0.386 |
| NB (within 200 miles) | 0.645 | 0.603 | 0.511 | 0.400 |
| Hybrid ($\beta = 0.5$) | 0.626 | 0.575 | 0.466 | 0.340 |
| Hybrid | 0.628 | 0.576 | 0.463 | 0.330 |

TABLE IV: Average Jensen-Shannon Divergence @P%

location $l$ in the community $c$, we compare the distribution of hashtag at $l$ with the distribution of equal number of hashtag suggested by the community distribution, and then we calculate the corresponding JSD. Again in Table III, we see that the Temporal+HD approach results in the highest divergence (and so, worst performance in terms of estimating the actual distribution). In all cases, we see that the Cosine+HD approach ($\mathcal{D}^*_{cos}$) results in the smallest divergence. This result demonstrates that Cosine+HD is the most effective distance metric, again indicating the strength of content-based similarity versus temporal-based similarity of locations.

### *E. Comparing Three Estimation Approaches*

Next we compare the performance of the three proposed approaches for estimating the hashtag distribution of the target location. For community-based (CB) estimation, we adopt Cosine+HD combination as it was shown to give the best estimation in all community-based variations. For the Neighbor-based (NB) approach, we consider two alternatives: only contiguous neighbors of the target location and all locations that are within 200 miles from the target location. For the hybrid approach, we also compare two alternatives: a version which has constant $\beta$ value and a version with varying $\beta$ as we defined previously. We report the average for locations that have more than 500 distinct hashtags.

The JSDs of the estimation methods are shown in Table IV. We observe two hybrid approaches, with JSD values 0.33 and 0.34, generally perform better than the strictly community-based approach and the neighbor-based approach. On the one

hand, neighbor-based estimation approach tends to identify hashtags that actually exist in a target location, however the popularity of these hashtags is unclear: they might either be popular, or be very local that only appear in a few locations. On the other hand, the community-based approach tends to identify overall popular hashtags that are possibly popular in target location as well. A combination of these two distributions effectively balances the hashtags discovered – leading to an increase in the accuracy of the estimated distribution.

We also observe that varying $\beta$ is better than a constant $\beta$, as we adjust the weight according to how influential the target location is. For a location with a large number of distinct hashtags, hashtags of neighbor locations tend to be sparse compared to the target and thus is prone to be missing and incomplete in representing the target location. By increasing the weight of the community distribution, we increase the probability of seeing the popular hashtags in the community, which are also likely to be popular in the target location. For a location with a small number of distinct hashtags, neighbor hashtags tend to be more precise than community hashtags, as the target location is not influential in the community and may share limited number of hashtags with the community. Increasing the weight to the neighbor distribution works better.

## V. Conclusion

We have proposed and evaluated a community-based framework for tackling the problem of geospatial tag distribution estimation, which is a key component of many new location-augmented search, retrieval, and mining applications. We have investigated two ways to represent locations. Additionally, we have compared three approaches for capturing the hidden conceptual distance between locations, and evaluated three smoothing strategies. Through experimental investigation, we found that geo-locations have a tendency of sharing similar "ideas" and forming geo-spatial communities. Meanwhile, we demonstrated how our community discovering approach and smoothing strategy leads to high-quality hashtag distribution estimation. In our future work, we are interested to study geospatial community formation in alternative social media platforms (e.g., Pinterest) and to incorporate alternative signals of community formation, including activity patterns, temporal changes of idea flow, and topic-sensitive signals (e.g., considering only political hashtags).

**Acknowledgments** This work was supported in part by NSF grant IIS-1149383 and a Google Research Award.

## References

[1] B. Shaw, J. Shea *et al.*, "Learning to rank for spatiotemporal search," in *Proceedings of WSDM*, 2013.

[2] J. Teevan, A. Karlson *et al.*, "Understanding the importance of location, time, and people in mobile local search behavior," in *MobileHCI*, 2011.

[3] J. Cranshaw, R. Schwartz *et al.*, "The livehoods project: Utilizing social media to understand the dynamics of a city," in *Proceedings of the ICWSM*, 2012.

[4] A. X. Zhang, A. Noulas *et al.*, "Hoodsquare: Modeling and recommending neighborhoods in location-based social networks," in *IEEE Conference on SocialCom*, 2013.

[5] L. B. Baltussen, M. Büchi *et al.*, "One Percent of Twitter, Part II: Geotags, Text Analysis, and Event Profiling," 2014.

[6] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, 2001.

[7] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, "Socio-spatial properties of online location-based social networks," in *Proceedings of the ICWSM*, 2011.

[8] P. Serdyukov, V. Murdock, and R. Van Zwol, "Placing flickr photos on a map," in *Proceedings of the SIGIR*, 2009.

[9] D. Preoţiuc-Pietro, J. Cranshaw, and T. Yano, "Exploring venue-based city-to-city similarity measures," in *Proceedings of the SIGKDD Workshop on Urban Computing*, 2013.

[10] A. Noulas, S. Scellato *et al.*, "Exploiting semantic annotations for clustering geographic areas and users in location-based social networks," in *The Social Mobile Web*, 2011.

[11] M. Kafsi, H. Cramer *et al.*, "Describing and understanding neighborhood characteristics through online social media," in *Proceedings of the WWW*, 2015.

[12] H. Wang, B. Chen, and W.-J. Li, "Collaborative topic regression with social regularization for tag recommendation," in *IJCAI*, 2013.

[13] D. Liu, X.-S. Hua *et al.*, "Tag ranking," in *Proceedings of the WWW*, 2009.

[14] P. Heymann, D. Ramage, and H. Garcia-Molina, "Social tag prediction," in *Proceedings of the SIGIR*, 2008.

[15] E. Zangerle, W. Gassler, and G. Specht, "Recommending#-tags in twitter," in *Proceedings of the Workshop on Semantic Adaptive Social Web CEUR Workshop Proceedings*, 2011.

[16] F. Godin, V. Slavkovikj *et al.*, "Using topic models for twitter hashtag recommendation," in *Proceedings of the 22nd international conference on World Wide Web companion*, 2013.

[17] K. Y. Kamath, J. Caverlee *et al.*, "Spatio-temporal dynamics of online memes: a study of geo-tagged tweets," in *Proceedings of the WWW*, 2013.

[18] Z. Ma, A. Sun, and G. Cong, "Will this# hashtag be popular tomorrow?" in *Proceedings of the SIGIR*, 2012.

[19] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proceedings of the CIKM*, 2010.

[20] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: improving geographical prediction with social and spatial proximity," in *Proceedings of the WWW*, 2010.

[21] D. Jurgens, T. Finethy *et al.*, "Geolocation prediction in twitter using social networks: a critical analysis and review of current practice," in *Proceedings of the ICWSM*, 2015.

[22] H. Saif, Y. He, and H. Alani, "Alleviating data sparsity for twitter sentiment analysis," in *CEUR Workshop Proceedings*, 2012.

[23] W. Pan *et al.*, "Transfer learning in collaborative filtering for sparsity reduction," in *AAAI*, 2010.

[24] J. Lin, R. Snow, and W. Morgan, "Smoothing techniques for adaptive online language models: Topic tracking in tweet streams," in *Proceedings of the SIGKDD*, 2011.

[25] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *TOIS*, 2004.

[26] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, 2007.

[27] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, 1967.

[28] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, 2007.

[29] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007.

[30] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Economic geography*, 1970.

[31] R. Sinnott, "Virtues of the haversine." 1984.

[32] J. W. Hager, J. F. Behensky, and B. W. Drew, "The universal grids: Universal transverse mercator (utm) and universal polar stereographic (ups)," Tech. Rep., 1989.

[33] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on 37.1 (1991)*, 1991.

[34] K. Y. Kamath and J. Caverlee, "Spatio-temporal meme prediction: learning what hashtags will be popular where," in *Proceedings of the CIKM*, 2013.

[35] D. Pelleg and A. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters." in *ICML*, 2000.