# College Towns, Vacation Spots, and Tech Hubs: Using Geo-Social Media to Model and Compare Locations

**Hancheng Ge, James Caverlee**
Department of Computer Science and Engineering
Texas A&M University, USA, 77840

## Abstract

In this paper, we explore the potential of geo-social media to construct location-based interest profiles to uncover the hidden relationships among disparate locations. Through an investigation of millions of geo-tagged Tweets, we construct a per-city interest model based on fourteen high-level categories (e.g., technology, art, sports). These interest models support the discovery of related locations that are connected based on these categorical perspectives (e.g., college towns or vacation spots) but perhaps not on the individual tweet level. We then connect these city-based interest models to underlying demographic data. By building multivariate multiple linear regression (MMLR) and neural network (NN) models we show how a location's interest profile may be estimated based purely on its demographics features.

## Introduction

Urbanization in the past century has transformed the world. Today, around 54% of the world's population lives in urban areas, and projections estimate that by 2050 around two-thirds of all people will live in cities (U.N. 2014). With this growth, a key task for urban planners, economists, social scientists, and political entities is to understand the people who live in these cities – for planning future development, building engaged communities, and so on.

Traditional approaches for collecting population data have often relied on labor-intensive, expensive, and slow collection methods. For example, the US census costs 13 billion dollars and provides in-depth data only every ten years (Economist 2011). More frequent surveys have been adopted in the past decade including the American Community Survey (in the US) and the National Household Survey (in Canada); these surveys aim to collect data every year, but only sample a few million households per year (Bureau 2014). In a separate direction, private corporations (and some research institutetions) have sought to collect and analyze proprietary datasets including cellphone call records (Greenwald, MacAskill, and Ackerman 2013), search engine query logs (Boytsov, Dean, and Sercinoglu 2012), in addition to marketing records that are created by companies like Wal-Mart, Amazon, and Target.

In contrast, geo-social media offers a potentially low-cost, scalable, and fine-grained window into the spatio-temporal activities of millions of people. Enabled by the widespread adoption of GPS-enabled tagging of social media content via smartphones and social media services (e.g., Facebook, Twitter, Foursquare), these "footprints" open new possibilities for understanding the dynamics of human behavior and the rhythm/pulsation of social life from local to global levels. However, gleaning knowledge about human behaviors with the geo-social media data is very challenging as the data is usually unstructured and thematically diverse (Croitoru et al. 2013). While we recently witnessed many compelling new studies to leverage the large geo-spatial footprint to explore dynamics of individuals or communities (Li et al. 2008; Scellato et al. 2011; Yin et al. 2011), few studies focus on exploring the dynamics of human behavior from the perspective of cities that is a typical scenario of urban computing (Zheng et al. 2014). For instance, we may ask what the distribution of interests of people living in Los Angles is; if it is similar with the distribution in Houston.

In this paper, our goal is to explore the viability of geo-social media to model, compare, and forecast the interest-based profiles of cities. Concretely, we explore the following questions: (1) how do we build the interest-based profiles of cities?; (2) is there the coherent characteristics between cities?; (3) Which factors could be related to these interest-based profiles?; and (4) how can we predict the interest-based profiles of cities? To answer these questions, we propose a geo-located tweet-driven framework to model cities by constructing *city-based interest models* over 14 high-level interest categories (e.g., Sports, Technology, Politics). Compared to prior work like ESRI's Tapestry Segmentation (ESRI 2014) (which breaks down US neighborhoods by socioeconomics and demographics) or efforts to segment regions of a city by human mobility (Yuan, Zheng, and Xie 2012), we focus on constructing *interest models* of cities based solely on widespread GPS-tagged social media. These interest models are important for multiple applications including: (i) profiling the revealed preferences of people in a particular location for targeted social analytics; (ii) early detection of shifting population interests, which can impact urban planning; and (iii) new data-driven resources for social scientists to study evolving social structure and social composition, among many others.

We evaluate the quality of these city-based interest models to identify coherent groupings of cities (e.g., college towns, tech hubs) via a clustering-based approach and then to estimate the type of city based purely on these interest models. We further link these city-based interest models to underlying demographic features of each city – to uncover the correlation between the composition of a city and the tweet-based interests of people in the city, which is important for location-based recommendation systems, targeted advertising, social analytics, and so forth. Our interest here is to investigate if there is a relationship between geo-social media based interest profiles and underlying factors (like demographics). In this study, we explore this question at the city level though we leave open to future investigations at more fine-grained levels.

As a step towards the vision of urban computing, this paper (1) Investigates the city-based interest models over a collection of 112 US cities, wherein we identify coherent clusters of cities based only on these interest models, validate the quality of these models against a ground truth generated from 100s of expert-curated lists, and test these models on an additional held-out 56 cities. (2) Links the city-based interest models to demographic features, toward identifying key characteristics of each city that provide explanatory power over its expressed interests. (3) Predicts a city's interest model based purely on key demographic features.

## Study Framework and Setup

Our study aims to build interest models for cities from geo-social media, so that we can understand what a population cares about, how cities are positioned relative to each other, and what the key factors are impacting these models. While geo-social media is an inherently biased source – due to factors like uneven adoption of mobile devices – the proposed framework is designed to demonstrate the potential of leveraging this rich source of population data. Beginning with a large set of geo-tagged social media posts during a period (in our case, around a year), we first create city-based interest models by applying a hybrid tweet classification approach. Based on these interest models, we group cities into coherent clusters (e.g., tech hubs, college towns), explore the salient demographic features that impact these interest models, and develop predictive models.

**Geo-Social Media Data.** Our work here focuses on an initial sample of over 1 billion geo-tagged tweets collected via the Twitter Streaming API from September 2012 to May 2013. We filter the tweets to keep only those originating from the continental US, in English, and containing at least ten words (for providing enough context for simplifying human judgments needed in later parts of the evaluation setting), resulting in a base collection of 60,531,279 tweets. We reverse geo-code each tweet, keeping only those that originate from one of the top-112 US cities by population, resulting in a final dataset of 4,845,316 tweets.

**Demographics Data.** We collected demographic data for these 112 cities from the 2010 Census Summary File and the American Community Survey (ACS) 5 Year Data. The census data contains population data including sex, age, household relationship, household type, family type, as well as housing items including occupancy status, vacancy status, and so on. The ACS contains a variety of social and economic characteristics. In total, we identify 827 demographic features (e.g., median family income, mean hours worked in past 12 months).

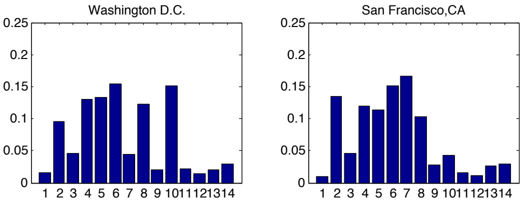

Figure 1: Example City-Based Interest Models

## Building City-Based Interest Models.

In this section, we introduce the city-based interest model for mapping from individual tweets to a high-level categorical representation. Given a set of tweets $T$ originating from a number of cities $S$, our goal is to construct an $n$-dimensional *city-based interest model* $\vec{s} = \{d_1, d_2, ..., d_n\}$ for a city $s \in S$, that represents the high-level interests of the individuals. For the purpose of this study, we adopt 14 high-level categories based on the categories in the Open Directory Project (DMOZ 2014) and further refine to reflect interests expressed on Twitter: Arts, Entertainment, Science and Education, Business, Food, Sports, Technology, Travel, News, Politics, Religion, Weather, Health, and Holidays.

### Multilevel Mixed Feature Classifier

To identify the city-based interest model based on these 14 categories, we propose and train a special multilevel mixed feature (MMF) classifier, as one type of cross-domain data fusion approach (Zheng 2015), for mapping from the content of each tweet to one of the 14 categories.[1] This hybrid classifier, combines: (i) a binary classifier for distinguishing between topical and non-topical tweets to filter out nonsensical and chat-oriented tweets; and (ii) a multinomial classifier for determining which of the 14 categories a topical tweet belongs to. The output of the hybrid classifier is a city-based interest model across these 14 categories, as illustrated in Figure 1, where the x-axis represents 14 interest categories and the y-axis is the percentage of tweets in each category. The proposed MMF is summarized in Algorithm 1.

**Topical vs. Non-topical Tweets** The first step is to isolate tweets that actually express a categorical interest for building the city-based interest model. We can view this step as a binary classification problem for distinguishing between topical and non-topical tweets. A key step for such a classification is feature selection, which is especially challenging

[1] Of course, the city-based interest model could be further refined to consider more top-level categories or nested categories.

```
Algorithm 1: Build city-based interest models by MMF
Input: A set of cities $S = \{s_1, s_2, ..., s_M\}$, a set of
       tweets $T = \{(t_i, s_i), i = 1, 2, ..., N, s_i \in S\}$,
       and topical categories $C = \{c_1, c_2, ..., c_L\}$.
Output: City-based interest models $\vec{s} = \{d_1, ..., d_L\}$
for $i \leftarrow 1$ to $N$ do
    $b_i \leftarrow GeneratingBinaryFeature\ (t_i)$
    $\gamma_i \leftarrow GeneratingTokenFeature\ (t_i)$
    $\bar{\xi_i} \leftarrow GeneratingNewFeatureVector\ (b_i, \gamma_i)$
Extracting a subset of tweets $T_\Psi \in T$ as the training
  dataset with manual labels indicating if the tweet $\bar{\xi_i}$ is
  topical (1) or not (0)
Training a binary classifier $RF$ using $T_\Psi$
for $i \leftarrow 1$ to $N$ do
    $f_i \leftarrow RF(\bar{t_i})$ distinguishing topical tweets
    $\xi'_i \leftarrow (f_i, b_i, \gamma_i)$ new feature vector for tweet $t_i$
Aggregating topical tweets $T_\Omega = \{t_i, f_i = 1, t_i \in T\}$
Extracting a subset $T_\Phi \in T_\Omega$ as the training dataset with
  manual labels indicating which topical categories
  $c_j \in C$ the tweet $t_i$ belongs to
Training two classifiers LG and SMO using $T_\Phi$
for $t_i$ in $T_\Omega$ do
    $l_i^1 \leftarrow LG(\xi'_i)$ get topic category from $LG$
    $l_i^2 \leftarrow SMO(\xi'_i)$ get topic category from $SMO$
    $l_i^3 \leftarrow MAX(\gamma_i)$ get topic category holding the
      maximum value in the token feature $\gamma_i$
    if $l_i^2 \neq l_i^2 \neq l_i^3$ then
        $l_i \leftarrow OneExtraDetermination(l_i^2, l_i^2, l_i^3)$
    else
        $l_i \leftarrow MajorityVote(l_i^2, l_i^2, l_i^3)$
    $j \leftarrow GetIndexOfTopicCategory(l_i, C)$
    $s_i \leftarrow Location(t_i)$
    $d_j \leftarrow d_j + 1, j \in \{1, 2, \ldots, L\}, d_j \in \vec{s_i}$
return $\vec{s_i}, i = 1, 2, ..., M$
```

for tweets that are 140 characters in length and often written informally. Here, we adopt two types of features: *binary features* representing various characteristics of the tweets and *token features* for capturing category-specific words.

*Binary Features.* Inspired by the work (Sriram et al. 2010), we define 12 binary features as follows: (·) Reference to another user at the beginning of a tweet; (·) Reference to another user within a tweet; (·) Reference to hashtag at beginning of a tweet; (·) Reference to hashtag within a tweet; (·) Personal pronoun at beginning of a tweet; (·) Opinion words; (·) Time and date information; (·) URLs; (·) Currency information; (·) Emphasis on words; (·) Emoticon; (·) Scores in sports games.

*Token Features.* By token features, a tweet is represented as a vector with 14 elements, each of which is the cumulative score of keywords and phrases on the text of this tweet in a corresponding category. For each of the 14 categories, three annotators identify a selection of keywords and phrases based on (EnchantedLearning 2013). For instance, for Politics we may select words like "Obama" and "war". Each keyword is weighted based on the number of categories it appears in. Hence, a keyword appearing in a single category is weighted 4, a keyword in up to four categories is weighted 2, and so on. Experimentally, we find that a weighting scheme of (4, 2, 1) results in the best performance after testing schemes like (3,2,1) and (5,3,1).

To test the topical vs. non-topical tweet classifier, we manually label a random sample of 8,200 tweets (3,200 topical and 5,000 non-topical tweets), where the ground truth categories are assigned by the majority vote of three labelers. Meanwhile, these 3,200 topical tweets are further labeled to 14 categories. If three labelers have different opinions on a tweet, one extra annotator is invited to make the final decision. Since there is an imbalance in the dataset, we additionally apply the resampling method of Synthetic Minority Oversampling TEchnique (SMOTE) (Chawla et al. 2002) to balance the number of samples in topical and non-topical tweets. We report the results of several classifiers using 10-fold cross validation in Table 1. We can see that Random Forest with both binary and token features provides a better performance with a precision of 89.1%. Moreover, SMOTE significantly improves performance of classifiers with a maximum improvement of 4.3%.

**Assigning Topical Tweets to Categories** Once we have separated out the non-topical tweets, our second step is to assign a category to each of the remaining topical tweets for building the city-based interest model. We propose a multinomial classifier to further classify them into 14 pre-defined categories. This multinomial classifier is a mixed one with three components by utilizing the rule of majority vote. First of all, we tested 6 classifiers (using the same setup as for the binary classifier) to identify the best two of them. Based on 3,200 manually labeled tweets as well as 10-fold cross validation, we find that Logistic and SMO perform better. Additionally, we find that directly considering the category with the highest score based purely on the token feature as the assigned category of the tweet, the precision is 93.12%, the recall is 92.54% and the F1-score is 92.83%. This method is named "MAX" and is our third component; note that "MAX" alone may lead to over-fitting. Hence, we combine the three components to find the final category as determined by the a majority vote of the three components. In cases of ties, we default to the choice of the "MAX" approach.

Finally, we apply the MMF classifier to the entire dataset of 4.8 million tweets. For evaluation, we randomly select a total of 1,000 of these tweets from each of 14 categories and manually label them. We find that our proposed approach performs good with an average precision of 85.24%, an average recall of 86.19%, and an average F1-score of 85.65%, and provides quality on par with prior studies that have aimed to label tweets (Lee et al. 2011; Huang and Mu 2014). Additionally, since our goal is to identify a macro-perspective city-based interest model, individual labeling errors can be tolerated with some extent. Ultimately, we can apply this MMF approach to create city-based interest models, as shown in Figure 1.

Table 1: Precision For Classifying Topical and Non-Topical Tweets

| **Classifier** | **No SMOTE** | | | **With SMOTE** | | | **BOW** |
|---|---|---|---|---|---|---|---|
| | *Binary* | *Token* | *Binary+Token* | *Binary* | *Token* | *Binary+Token* | *BOW* |
| *Naive Bayes* | 0.770 | 0.819 | 0.811 | 0.709 | 0.734 | 0.749 | 0.808 |
| *Logistic* | 0.797 | 0.833 | 0.832 | 0.736 | 0.810 | 0.861 | 0.778 |
| *RBF Network* | 0.776 | 0.817 | 0.785 | 0.703 | 0.746 | 0.775 | 0.782 |
| *C4.5* | 0.722 | 0.822 | 0.826 | 0.816 | 0.822 | 0.869 | 0.793 |
| *Random Forest* | 0.802 | 0.880 | 0.872 | 0.828 | 0.831 | ***0.891*** | 0.804 |
| *SMO* | 0.784 | 0.802 | 0.828 | 0.760 | 0.828 | 0.861 | 0.802 |

## Uncovering Intrinsic Interests among Cities

Given the city-based interest models derived from geo-social media, we explore in this section the quality of these models. Do they reveal meaningful relationships among cities? Can we identify coherent clusters of cities based on these interest models? Do the models generalize?

### Identifying Coherent Groups of Cities

To investigate these questions, we focus our initial efforts on the original 112 cities and employ the well-known spectral clustering algorithm over the city-based interest models. Since the spectral clustering takes as input a pre-specified number of clusters, we adopt the Cubic Clustering Criterion (CCC) which estimates the number of clusters based on the assumption that a uniform distribution on a hyperrectangle will be divided into clusters shaped roughly like hypercubes (Sarle 1983). The optimal number of clusters is commonly taken as the one with the largest CCC. We tested the number of clusters from 3 to 100, and as can be seen in the bottom-right subfigure in Figure 3, the best choice is $k = 10$ (the red point). Since the individual runs of the spectral clustering conduct slightly different clusters possible due to the random initialization, we run the algorithm multiple times (100 times in this study) and group together cities that are clustered more than 50% of the time. After that, we further employ Ward's hierarchical minimum variance method to refine these clusters.

With $k = 10$, we show the output clusters in Figure 3, where the x-axis in each figure is the category (i.e. 1 = Arts, 2 = Entertainment, ..., 14 = Holidays) and the y-axis is the relative difference of this category versus the mean percentage across all clusters. That is, cities in Cluster 1 score high in Politics (Category 10) relative to all other clusters where the mean is set to 0; similarly, cities in Cluster 2 score high in Technology (Category 7) relative to all other clusters. Based on manual inspection, we can observe that the clusters make intuitive sense. For example:

- *Political Cities:* The first figure corresponding to Cluster 1 is over-represented by Politics (Category 10) and Travel (Category 8) relative all other cities. Fitting our intuition, this cluster contains Washington DC and Arlington, VA.
- *College Towns:* The fifth cluster is college towns that emphasize on Science and Education (Category 3) and Food (Category 5) but low in Travel (Category 8) and Business (Category 4). Member cities include Athens, GA, Blacksburg, VA, College Station, TX, and Cambridge, MA.

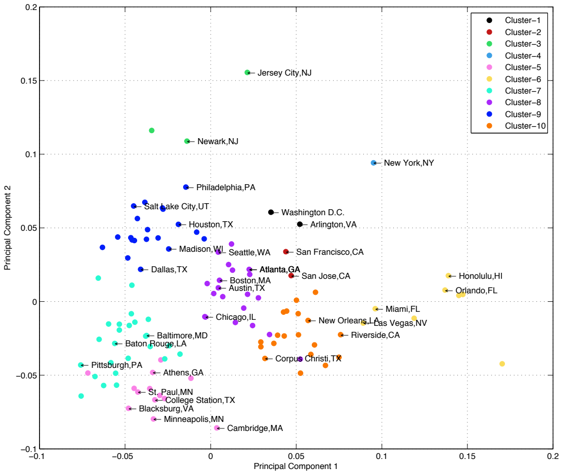

Figure 2: Scatter Plot for All 112 Cities

- *Tech Hubs:* The second cluster contains technology-related hubs like San Francisco, CA and San Jose, CA.
- *Tourist Towns:* The sixth cluster contains Orlando, FL, Miami, FL, Honolulu, HI, and Las Vegas, NV, scoring highest in Travel (Category 8) and Food (Category 5) and relatively low in Sports (Category 6).

So based purely on high-level interests expressed via social media, the city-based interest models can uncover clear coherence in how cities are related to one another. Focusing on specific cities, we find similarly encouraging results. For instance, Cambridge, MA has the largest degree of interest in Travel relative to other college towns in its cluster; with two world-class universities and closeness to a major city, this interest is unsurprising. In Cluster 8, Birmingham, AL has a high interest in religion relative to other cities in the same cluster. College Park, MD displays the highest interest in Politics relative to other college towns, most likely due to its proximity to Washington DC.

For a better visual inspection, we apply principal component analysis (PCA) to the city-based interest models of all 112 cities and plot the two cities according to their first two principal components in Figure 2. Here we can see that the 10 clusters identify contiguous regions of related cities.

### Validating Against Ground Truth

While these results are intuitively sensible, we now turn to an evaluation of the quality of the discovered city-based in-

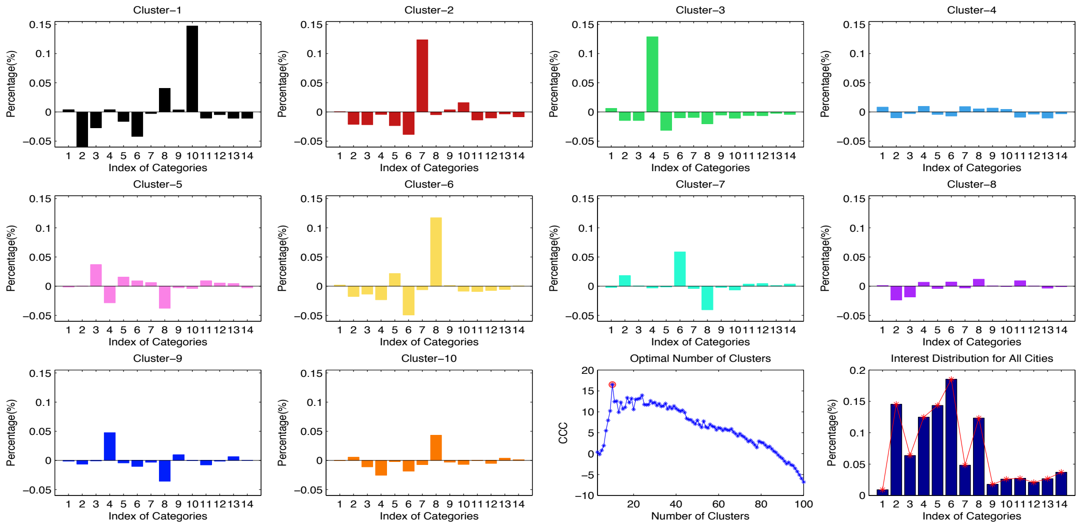

Figure 3: 10 Clusters Derived from City-Based Interest Models

Table 2: Average Similarity Scores for 10 Clusters

| Cluster | Avg. with-in Cluster | Avg. with All Others |
|---|---|---|
| *1* | 1.00 | 0.31 |
| *2* | 0.79 | 0.23 |
| *3* | 0.86 | 0.38 |
| *4* | None | 0.26 |
| *5* | 0.78 | 0.29 |
| *6* | 0.80 | 0.34 |
| *7* | 0.66 | 0.32 |
| *8* | 0.83 | 0.28 |
| *9* | 0.77 | 0.36 |
| *10* | 0.80 | 0.28 |

terest models and clusters via three experiments: (i) the first compares the discovered clusters versus a ground truth derived from expert-curated lists of related cities; (ii) the second assigns a group of 56 new cities to previously clustered groups of 112 cities and evaluates the quality of these assignment via three metrics (precision, recall, and F-measure); and (iii) the third compares the discovered clusters conducted through the proposed city-based interest models versus ones made by two alternative methods.

**Versus Ground Truth.** In the first experiment, we mine 800 expert-curated lists of related cities from the City-Data web resource (City-Data 2008). These lists span many areas, including top college cities, top sport cities, top religious cities, top technology cities, and so forth. Treating membership on a list as a signal of relatedness, we apply the Jaccard coefficient to numerically measure the similarity between two cities as a function of the number of lists the two co-occur on. The similarity score (WS) between two cities can be defined as:

$$WS(l_i, l_j) = \frac{|O_{l_i} \cap O_{l_j}|}{|O_{l_i} \cup O_{l_j}|} \qquad (1)$$

where $O_l$ denotes the set of lists the city $l$ occur on.

We compute the average similarity scores in each of the 10 clusters as shown in Table 2. Overall, the average similarity for cities within clusters is 0.81, as compared to 0.31 for cities in different clusters. In other words, the city-based interest models lead to groups of cities that conform to expert-curated lists of how cities are related. Across each discovered cluster, the results hold with stronger list-based similarity within clusters than across clusters. Note that Cluster 4 is an outlier, containing just one city (New York City) and so there is no within cluster assessment. These results support the assertion that city-based interest models built on geo-social media do meaningfully model cities.

**Testing New Cities.** Coupled with this investigation of the 112 cities in our initial dataset, we additionally consider a collection of 56 new cities that appear on 800 top lists, but not in our collected 112 cities. Following our approach, we construct for each city an interest profile by collecting tweets, classifying them into 14 pre-defined categories using the hybrid classifier, and aggregating them with their corresponding tagged geo-location information. We then assign each city to the cluster with the smallest Euclidean distance from its interest model. As in the case of the original 112 cities, we find intuitive and sensible results: college towns like Chapel Hill, NC, Auburn, AL, State College, PA, and Urbana-Champaign, IL are assigned to our original College Towns cluster (Cluster 5). Palo Alto, CA is assigned to the Tech Hubs (Cluster 2). Destinations like Napa, CA, Lahaina, HI, Palm Springs, CA, and Savannah, GA are assigned to Tourist Towns (Cluster 6).

We also quantitatively evaluate the quality of these assignments by considering two metrics – Average Precision and Average Recall – over the space of expert-curated lists:

$$AP = \sum_{l_i \in S_c} \frac{|S_c \bigcap S_{list}^{l_i}|}{|S_c|}, AR = \sum_{l_i \in S_c} \frac{|S_c \bigcap S_{list}^{l_i}|}{|S_{list}^{l_i}|},$$

where $S_{list}^{l_i}$ denotes the set of cities co-occurring with city

$l_i$ in the set of lists, and $S_c$ is a set of cities in the cluster $c$. Higher values indicate that the city-based interest models do reflect the underlying relationships among these cities.

In practice, we employ a distance metric, *Euclidean Distance*, to assign each of the additional 56 cities to one of the 10 clusters after testing different metrics. The results are reported in Table 3. Note that since there is no city assigned to clusters 1 and 4, and only one city to be assigned to cluster 2, 3 and 10, respectively, we are not able to calculate metrics for these five clusters. For the rest of the clusters, the Average F1 reaches 70.4% on average with the best of 100% in the cluster 6 and the worst of 34% in the cluster 9 to which only four cities are assigned. This result is a good indicator implying that most 56 cities are correctly assigned to 10 clusters based on the proposed city-based interest models.

Table 3: Performance for 56 Cities

| Cluster | Avg. Precision | Avg. Recall | Avg. F1 |
|---|---|---|---|
| *5* | 85% | 77% | 81% |
| *6* | 100% | 100% | 100% |
| *7* | 79% | 73% | 76% |
| *8* | 71% | 52% | 61% |
| *9* | 44% | 29% | 34% |

**Comparison with Alternative Methods.** While encouraging, are the proposed city-based models better at identifying coherent groups of cities than other methods? To address this questions, we consider two alternatives for grouping cities – one based on demographics data using the spectral clustering approach, similar to (Yin et al. 2011); and one based on topic vectors derived from Latent Dirichlet Allocation (LDA). For the demographics-based approach, we represent each of the 112 cities by 827 unique demographic features from the ACS 5 Year Data. For the LDA-based approach, the number of topics is determined by the method of Arun et al. (Arun et al. 2010) where LDA is treated as a matrix factorization mechanism. In both cases, the number of clusters is set as 10 as we did before. We average the average similarity score on all clusters as reported in Table 4. It is apparent that the coherent groups of cities derived from the proposed city-based interest model results in a higher average similarity scores in both with-in and with-out clusters are more reasonable than ones produced from the demographic-based and LDA-based approaches.

Table 4: Comparing Performance of Identifying Coherent Groups of Cities with Two Other Approaches

| Method | Avg. with-in Cluster⋆ | Avg. with All Others⋆ |
|---|---|---|
| *City-based* | 0.81 | 0.31 |
| *Demogr.-based* | 0.72 | 0.32 |
| *LDA-based* | 0.56 | 0.35 |

⋆Avg. means averaging the average similarity score on all clusters.

## Linking Demographics to Interest Models

So far we have seen that city-based interest models built over public geo-social media can provide a new window into modeling and comparing cities, and that latent groups of cities can naturally fall out of these models. We now focus our attention on linking the city-based interest models to demographic features of these cities, toward identifying key characteristics of each city that provide explanatory power over its expressed interests. By identifying the relationship between the underlying characteristics of cities and the interests of its population via social media, we can begin to utilize them in further analysis and applications such as building interest prediction models, location-based recommendation systems, targeted advertising, location selection and urban social analytics.

Consider a set of city demographic features $\Theta$ which might play an important role in shaping the interests of people. For city $s$, we denote its $k$ demographic features as $DG_s = \{dg_s^1, dg_s^2, ..., dg_s^k\}$, $DG_s \in \Theta$. Given $M$ clusters of cities $\{(s_i, sc_i), i \in \{1, 2, ..., N\}, sc_i \in \{1, 2, ..., M\}\}$ derived from geo-social media city-based interest models where $sc_i$ is the label of cluster for city $s_i$, we explore the salient demographic factors impacting these interest models. Recall that these interest models are based on tweets collected over the course of one year, so we anticipate that the interests embedded in these models reflect long-term population interests.

Concretely, we consider the original 112 cities and the 827 demographic features extracted from the 2010 Census and ACS 5 Year Data. But which of these demographic features actually contribute to the city-based interest models? Finding these salient demographic factors can be viewed as a feature selection problem. Since the number of features is significantly larger than the number of instances (cities), we adopt the Elastic Net (Zou and Hastie 2005), which has been shown to be a good tool for feature selection in the high-dimensional data. Similar to the work (Zhu and Hastie 2004), we employ a multinomial logistic regression model (known as the logit model) to this study. By considering the elastic net regularization, the negative log-likelihood can be defined as:

$$\begin{aligned} arg\min - \sum_{i=1}^{N}\Big(\sum_{k=sc_i} DG_i^T \boldsymbol{\beta}_k + log\sum_{j=1}^{M} e^{DG_i^T \boldsymbol{\beta}_j}\Big) \\ + \lambda_1 \sum_{j=1}^{M} |\boldsymbol{\beta}_j| + \lambda_2 \sum_{j=1}^{M} \|\boldsymbol{\beta}_j\|^2, \end{aligned} \qquad (2)$$

where $\boldsymbol{\beta}_i$ is a $k$-vector $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, ..., \beta_{ik})^T$, $\lambda_1$ and $\lambda_2$ are the shrinkage parameters. We can solve this model using coordinate descent (Friedman, Hastie, and Tibshirani 2009). It should be noted that each cluster is associated with their own $\boldsymbol{\beta}$, meaning that there are $M$ parameter vectors $\{\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \dots, \boldsymbol{\beta_M}\}$ if we have $M$ clusters in cities.

Taking into account the frequencies of selected features occurring in 10 clusters and their sizes of estimated coefficients, we identify the top-two salient features for each city-based interest model in Table 5. These are the most explanatory demographic features for the interest expressed by each city. For example, the interests expressed in tweets by members of College Towns (Cluster 5) can be best modeled by features (negatively) related to hours worked and age. Polit-

Table 5: Selected Demographic Features

| No. of Cluster | Selected Features | Coefficient |
|---|---|---|
| 1 | Median Family Income | 46.444 |
| | Mean Hours Worked in Past 12 Months for Workers 16 to 64 Years | 10.095 |
| 2 | Median Household Income by Age of Householder (Householder 25 to 44 Years) | 35.183 |
| | Race Population: Asian alone | 8.667 |
| 3 | Median Age by Means of Transportation to Work (Carpooled) | 129.643 |
| | Median Household Income by Age of Householder 65 Years and Over | -28.171 |
| 4 | Population of Workers by Means of Transportation to Work (Subway) | 10.657 |
| | Population of Workers by Means of Transportation to Work (Public Transportation) | 1.897 |
| 5 | Mean Hours Worked in Past 12 Months for Workers 16 to 64 Years | -164.662 |
| | Median Age by Means of Transportation to Work (Car, Truck or Van) | -65.153 |
| 6 | Median Age by Sex (Total) | 87.125 |
| | Median Number of Rooms | -40.936 |
| 7 | Median Number of Rooms | 61.657 |
| | Median Age by Means of Transportation to Work (Total) | 26.829 |
| 8 | Median Age by Sex (Total) | 63.572 |
| | Household Size by Vehicle Available (1-Person:2-Vehicles) | 20.954 |
| 9 | Mean Hours Worked in Past 12 Months for Workers 16 to 64 Years | 145.883 |
| | Median Monthly Housing Costs(Dollars) | -65.417 |
| 10 | Median Family Income | 18.651 |
| | Median Age by Means of Transportation to Work (Total) | 179.753 |

ical Cities (Cluster 1) are most impacted by median family income and mean working time. Tech Hubs (Cluster 2) are related to median income and population of Asians.

## Forecasting City Profiles

Finally, we investigate in this section the connection between these salient demographic features and a city's demographic information. The benefit of such an investigation is that accessing and collecting geo-social media data are much easier than collecting new demographic data of cities (which can take 1 to 10 years). So, can the city-based interest model be estimated based only on underlying demographics? Rather than claiming a causal relationship, our goal is to uncover the potential connection between underlying demographic factors and the positioning of a city as revealed through our city-based interest model clusters. With these models, we can isolate how demographic changes can alter the trajectory of a city's future interests.

Based upon a total of 64 selected salient features for all cities, we propose two models: a multivariate multiple linear regression (MMLR) model used to estimate the linear association between predictors and responses, and a neural network (NN) model, which is one of the most widely applied machine learning methods. Distributions of people's interests in all 168 cities (112 cities + 56 cities) are employed to develop prediction models in this section. Given a city's salient demographic data only, these two models can predict the city-based interest model. Specifically, the multivariate multiple linear regression (MMLR) model can be expressed as the following equation:

$$\vec{s_i} = DG_i * \beta_i + e_i, i = 1, ..., N \qquad (3)$$

where $\vec{s_i}$ is the city-based interest model for city $s_i$ as we introduced in this study; $DG_i$ is the demographic feature vector corresponding to city $s_i$, $e_i$ is the error term with multivariate normal distribution, and $N$ is the number of samples which is equal to 168 cities in our study.

The neural network (NN) model consists of two hidden layers in each of which there are 10 nodes with the standard back propagation (BP) algorithm. The sigmoid function is applied as the activation function in the neural network. In order to evaluate the performance of these two models, we adopt the Root Mean Square Error (RMSE). The 10-fold cross validation is applied in estimating parameters, which divides our data into 10 equally-sized sub-data sets, and performs 10 training and validation steps. In each step, 9 sub-data sets are utilized as training data and the remaining one is for validating. Each sub-data set can be only used as the validation once. Overall, the prediction results are summarized in Table 6.

Table 6: Performance of Models by RMSE

| Model | RMSE | Max Error | Min Error |
|---|---|---|---|
| *MMRL* | 0.0122 | 0.0591 | 0.0000232 |
| *NN* | 0.0186 | 0.0750 | 0.00006065 |

As we can see, both models can estimate the city-based interest model with very low error, with the MMRL model achieving better RMSE than the neural network model. This positive result demonstrates that these two prediction models work well and that it is possible to automatically predict the city-based interest profiles.

## Conclusion

Geo-social media uncovers a new window into not only individuals, but also to the cities in which they reside. Exploring the relationship between cities and social media is an emerging and important research area. In this paper, we have explored the potential of geo-social media to construct location-based interest profiles to uncover the hidden relationships among disparate locations. We have seen that these models can identify coherent clusters of cities that conform with expert-curated lists of related cities. We have also shown how these city-based interest models can be connected to the underlying demographics of these cities, opening new opportunities for forecasting the future of cities.

## References

Arun, R.; Suresh, V.; Madhavan, C. V.; and Murthy, M. N. 2010. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Advances in Knowledge Discovery and Data Mining*.

Boytsov, A.; Dean, J. A.; and Sercinoglu, O. 2012. Systems and methods for generating statistics from search engine query logs. US Patent 8,126,874.

Bureau, U. C. 2014. American community survey. http://www.census.gov/acs/www/.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*.

City-Data. 2008. Profiles of all u.s. cities. http://www.city-data.com.

Croitoru, A.; Crooks, A.; Radzikowski, J.; and Stefanidis, A. 2013. Geosocial gauge: a system prototype for knowledge discovery from social media. *International Journal of Geographical Information Science*.

DMOZ. 2014. The open directory project. http://www.dmoz.org/.

Economist, T. 2011. Censuses: Costing the count. http://www.economist.com/node/18772674.

EnchantedLearning. 2013. Vocabulary word lists (word banks) by theme. http://www.enchantedlearning.com/wordlist.

ESRI. 2014. Tapestry segmentation. http://doc.arcgis.com/en/esri-demographics/data/tapestry-segmentation.htm.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2009. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*.

Greenwald, G.; MacAskill, E.; and Ackerman, S. 2013. Nsa collecting phone records of millions of verizon customers daily. *The Guardian*.

Huang, D., and Mu, D. 2014. Topic detection in twitter based on label propagation model. In *DCABES*.

Lee, K.; Palsetia, D.; Narayanan, R.; Patwary, M. M. A.; Agrawal, A.; and Choudhary, A. 2011. Twitter trending topic classification. In *ICDMW*.

Li, Q.; Zheng, Y.; Xie, X.; Chen, Y.; Liu, W.; and Ma, W.-Y. 2008. Mining user similarity based on location history. In *SIGSPATIAL*.

Sarle, W. 1983. The cubic clustering criterion. Technical report, SAS Technical Report A-108.

Scellato, S.; Noulas, A.; Lambiotte, R.; and Mascolo, C. 2011. Socio-spatial properties of online location-based social networks.

Sriram, B.; Fuhry, D.; Demir, E.; Ferhatosmanoglu, H.; and Demirbas, M. 2010. Short text classification in twitter to improve information filtering. In *SIGIR*.

U.N. 2014. World urbanization prospects: The 2014 revision. *United Nations, Department of Economic and Social Affairs, Population Division*.

Yin, Z.; Cao, L.; Han, J.; Zhai, C.; and Huang, T. 2011. Geographical topic discovery and comparison. In *WWW*.

Yuan, J.; Zheng, Y.; and Xie, X. 2012. Discovering regions of different functions in a city using human mobility and pois. In *SIGKDD*.

Zheng, Y.; Capra, L.; Wolfson, O.; and Yang, H. 2014. Urban computing: concepts, methodologies, and applications. *ACM TIST*.

Zheng, Y. 2015. Methodologies for cross-domain data fusion: An overview. *IEEE Transactions on Big Data*.

Zhu, J., and Hastie, T. 2004. Classification of gene microarrays by penalized logistic regression. *Biostatistics*.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.