Detecting Grand Tours of Europe with Geo-Tags

Conrad Lee, Derek Greene, Pádraig Cunningham
Clique Research Cluster, University College Dublin, Ireland

1 Introduction/Abstract

In his recent, light-hearted article "The Grand Tour", *New Yorker* journalist Evan Osnos writes of his time with a Chinese tour group in Europe [12]. Osnos describes a group of thirty eight closely-chaperoned tourists undertaking a breakneck bus tour, covering five countries in ten days. He argues that Chinese tourists have formed a distinctively Chinese "grand tour" of Europe. This tour includes places like Trier, Germany, the birthplace of Karl Marx, described in a Chinese guide book as "the Mecca of the Chinese people"; it includes some willows on Cambridge University's campus that were described in a famous Chinese poem as "young brides in the setting sun." Furthermore, Osnos notes that for nearly all of their meals, the tourists went to Chinese restaurants; the tour guide advised that "in general, one should steer clear of the local food."

Osnos' portrait of the Chinese grand tour raises the question of whether most people belong to distinct groups that display idiosyncratic preferences for various *points of interest* (POIs). Tourist boards are also interested in this question; in fact, gathering information on the interests and behavior of tourists is an extensive, expensive activity [1] and tourism behavior is itself a research topic [15]. In recent years, attention has focused on using internet and mobile phone data as a source of data for tourism research [14, 3]. However, using this data raises significant privacy concerns.

Here, we set out to detect the distinct "grand tours" of Europe that are undertaken by non-Europeans. We avoid the overhead and small scale of manual surveys and instead collect a massive dataset of travel itineraries on a global scale by collecting the metadata of 95 million Flickr photos for which precise geographic coordinates (*geo-tags*) are known. There is a growing body of work which makes use of Flickr geo-tagged photos [6, 5, 4]. In the next section, we describe how we collected this data and turned it into a list of POIs visited by each user. We then demonstrate how co-clustering tourists and POIs using Non-negative Matrix Factorization (NMF) [9] allows us to detect groups of individuals with distinct tourism preferences.

2 Methods

Collecting photo and home location data: The data for the present work consists of metadata associated with photographs that have been geo-tagged and posted on the popular photo-sharing website Flickr. These photographs are geo-tagged, either automatically by cameras (such as GPS-enabled smartphones), or manually using the Flickr interface. We collected metadata on 95 million geo-tagged images using the Flickr API. After discarding images with low geographic accuracy (indicated in the metadata) we were left with 83 million photographs, 59 million of which had an accuracy of 14 or greater (roughly street level). These photographs belong to 935,046 distinct users.

The Flickr API provided the home location for 186,827 of these users. As these locations were in the form of free text, we used the Yahoo! Placemaker API, to convert 153,069 of the location strings into geo-located metropolitan areas.

From lists of photographs to the tourist-POI matrix: To cluster tourists and the places they visit, we create a tourist-POI matrix. We now cover the two steps involved in creating this matrix: (a) mapping photographs to POIs, and (b) normalizing the tourist-POI matrix.

Due to its success and scalability, we choose to follow closely Crandall *et al.* 's lead and use mean-shift clustering with their parameter settings, setting the bandwidth to 0.001 decimal degrees (111 meters) and using their seeding method. We required a POI to have been visited by at least five distinct users. For details as well as a complete explanation of mean-shift, we refer the reader to their paper [5].

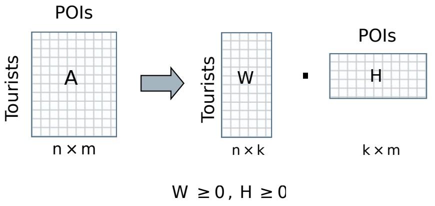


Figure 1: To cluster tourists and locations, we first create a normalized tourist-POI matrix $\mathbf{A}$. NMF on $\mathbf{A}$ produces two factors, $\mathbf{W}$ and $\mathbf{H}$, whose product approximates $\mathbf{A}$. The rows in $\mathbf{W}$ indicate how much each tourist belongs to each cluster, while the columns in $\mathbf{H}$ indicate the degree to which each POI belongs to each of the k clusters.

Each photo is assigned to the nearest POI within 111 meters; if it is is not within 111 meters of any POI it is filtered out. We create a binary matrix with users as rows and POIs as columns, assigning a value of 1 to (i, j) if the user associated with row i has visited the POI associated with column j. We include only those tourists who have visited at least five POIs. As in document clustering, we apply *tf-idf* normalization to the rows of this matrix and also take the euclidean norm. Whereas in document clustering this normalization assigns lower weights to words that are very common and therefore poor at distinguishing documents from each other (such as "the" or "of"), in this case the idea is to assign lower weights to POIs that all groups visit (e.g., the Eiffel Tower among tourists in Paris), and higher weights to POIs that distinguish one cluster from another.

Dimension reduction by co-clustering users and POIs with NMF: The resulting normalized matrix is extremely sparse. There are several reasons for this, including the unfortunate lack of time and resources that most people have for vacations. To aid in clustering this data, we reduce the dimensionality of the matrix. We choose NMF for this task because its output is readily interpretable as an "additive parts based" representation of the data [13, 9]. Fig. 1 displays the basic idea behind NMF, which is to approximate the matrix $\mathbf{A}$ (in this case the tourist-POI matrix) with the product of two low-dimensional factors, $\mathbf{W}$ and $\mathbf{H}$. We applied the projected gradient method for NMF proposed in [10] and ran it until convergence as recommended by the authors.

One important parameter in this analysis is k, which is the number of dimensions we use for the NMF (which also sets the width of the matrix in fig. 2). This parameter is difficult to set, but we noticed that when it was higher (with values of 40 or 50), then each dimension closely corresponded with a single city. Because we are interested in whether users belong to tours (collections of cities) rather than individual cities, we set $k = 10$ but any similarly value would be just as valid. We obtained qualitatively similar results with $k = 8$ and $k = 12$.

Clustering users by their preferences for NMF dimensions: After applying NMF, we have a matrix similar to the one the pictured on the next page, but whose rows have not yet been ordered. If we found that every user had all of his weight in one of the NMF dimensions, then no further clustering would be necessary because the NMF dimensions would be a good description of the "grand tours" we set out to find. However, while a subset of users are well-described by a single NMF dimension, many are better expressed as a mixture of these dimensions. For this reason, we define a "user group" as the people who are represented by a similar mixture of the NMF dimensions; we will use this terminology throughout the rest of the paper to avoid confusion between the terms NMF dimensions and clusters discovered through the subsequnt clustering.

There are many clustering methods we could use to find user groups (this task is equivalent to ordering the rows of the matrix in fig. 2 such that blocks emerge). We choose agglomerative hierarchical clustering using euclidean distance and the "average" linkage method [11] because we are interested in the possibility of a hierarchical taxonomy of tourist itineraries.

Before performing hierarchical clustering, we perform a filtering and normalization step so that the euclidean norm between two rows will function as desired. Upon inspection. we found many users who are very close to each other in terms of euclidean distance because they have near-zero values in each of the NMF dimensions. Despite this appearance of similarity, these users may in fact have visited diverse locations that simply did not align with any of the NMF dimensions; this is one consequence of using a low value for k. Since these users have very little signal in any of the NMF dimensions, we filter them out, leaving only those users whose total cluster membership weights in the NMF dimensions sum to ≥ 0.03. For each user vector not filtered out, we apply L1-normalization (so that each row in fig. 2 sums to one—i.e., all users have an equal amount of total preference).

Figure 2 displays the dendrogram produced by the hierarchical clustering alongside the matrix $\mathbf{W}$. The colors of branches in the dendrogram are created by simply taking a flat cut at a level using the default strategy of the SciPy software library [8] (Matlab uses the same default strategy), which produces a flat clustering of users that will serve as our user groups. The reader can judge the quality of these clusters by inspecting how well the partition corresponds to segments in the heatmap. These are the sets of users encompassed by red bands in the heatmap; each user group is identified by a number on the y-axis just to the right of the heatmap.

3 Characterizing grand tours of international tourists in Europe

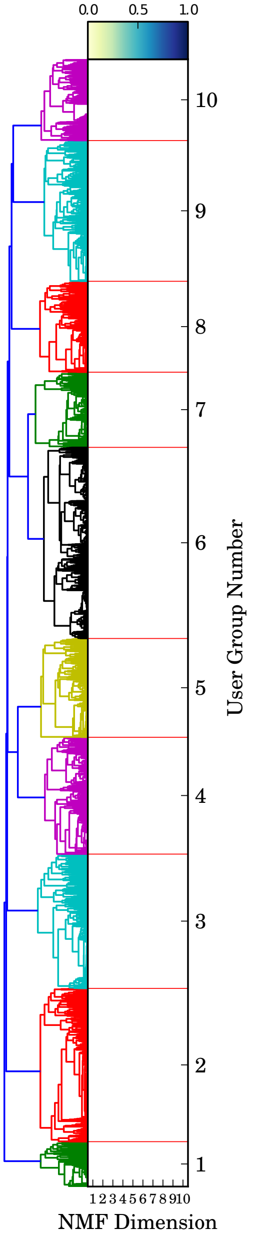


Figure 2: User groups detected by hierarchically clustering the matrix **W** from fig. 1.

In the following analysis, we focus on a subset of the data: only those photos that non-Europeans have taken when visiting Europe. Because we know the home locations of only a minority of the users, this leaves us with a relatively small sample of the data: there were 5,853 known non-Europeans who took photos of at least European five of the 31,538 POIs detected in Europe. After filtering out users with low weights in **W** as described above, 3,869 users remained.

We now characterize user groups in terms of POIs. First, we take the centroid of each user group, which can be thought of as representing the "average member" of the user group. We then multiply this centroid (which is a row vector) by the POI membership matrix (matrix **H** from fig. 1), an operation which gives us a single-row matrix that contains one weight for every POI. This product can be thought of as indicating how much a user group prefers each POI, so we can describe it as a user group's POI preference vector.

Given a user group's POI preference vector, we can calculate the total percentage of weight assigned to each POI. We can also collapse this vector by aggregating all of the POIs contained in each city. For five of the user groups, we have constructed a table listing the top cities (as well as the top POIs within those cities) that characterize that group. All of the numbers indicate the percentage of weight from that user group's POI preference vector for either a city or a POI in that city.

It is also possible to characterize user groups in terms of the attributes of their members. In this case, we know only one attribute: nationality. We check whether any groups are enriched in particular nationalities using Fisher's Exact Test [7]. For each user group, we test whether any nationality is over-represented. We do not correct for the fact that we are testing hundreds of hypohtheses, so the p-values are not meant to be taken at face value, but rather as an indicator of enrichment.

4 Results and discussion

We begin by discussing nationalities that were enriched in the user groups. The group with the largest number of enriched nationalities was user group 3, in which the following countries had a p-value of less than 0.05 (using the one-sided test Fisher's Exact Test): Mexico, Brazil, Peru, Puerto Rico, Argentina, Venezuela, Panama, and Chile. We note that user group 3 is oriented towards POIs in Spain (cf. section 4)—this preference could be explained by many factors, such as a common language (aside from Brazil) or the fact that many of the Latin American airlines have their European hub in Spain. The USA (the nationality with by far the most users) was enriched in only user group 6 (which is London oriented, cf. section 4). A few more nationalities were enriched in other user groups—see the supplemental download for a complete list [2].

Section 4 shows the top cities for each user group. (Again, see the supplemental download for the complete list). The weights are in terms of total percentage of weights, as described above. For each city, we can also see a user group's top POIs. We note first that three of the user groups are very focused on one particular city: group 7 has 49% of its weight in Rome, 18 has 58% of its weight in Paris, and 20 has 57% of its weight in London. The other groups are more diverse in their preferences for cities: no other group had more than 35% of its weight on the top city, and eight clusters have less than 20% of their weight in the top city.

City	Weight	Tag of top POIs
	User group 3	
Barcelona, ES	25.72	sagrada, güell, milà
Paris, FR	9.82	eiffel, notredamedeparis, triomphe
Madrid, ES	8.85	plazamayor, madroño, almudena
London, GB	5.31	greatcourt, bigben, londoneye
Berlin, DE	1.95	brandenburggate, holocaustmahnmal, re...
	User group 5	
Roma, IT	36.81	coliseo, pantheon, piazzasanpietro
Paris, FR	9.83	eiffel, triomphe, notredamedeparis
Florence, IT	4.12	santamariadelfiore, pontevecchio, log...
Venice, IT	3.44	stmarkssquare, rialtobridge, puntadel...
London, GB	3.32	towerbridge, bigben, britishairways
	User group 6	
London, GB	48.43	nelsonscolumn, turbinehall, greatcour...
Paris, FR	6.92	eiffel, notredamedeparis, triomphe
Edinburgh, GB	2.36	edinburghcastle, stgiles, victoriastr...
Bristol, GB	1.81	bathabbey, pulteney, royalcrescent
Dublin, IE	1.80	generalpostoffice, christchurchcathed...
	User group 7	
London, GB	42.19	bigben, londoneye, whitetower
Paris, FR	6.78	eiffel, notredamedeparis, triomphe
Roma, IT	5.10	coliseo, stpetersbasilica, pantheon
Venice, IT	1.65	stmarkssquare, rialtobridge, canalgra...
Amsterdam, NL	1.57	centraalstation, amsterdam, leliegrac...

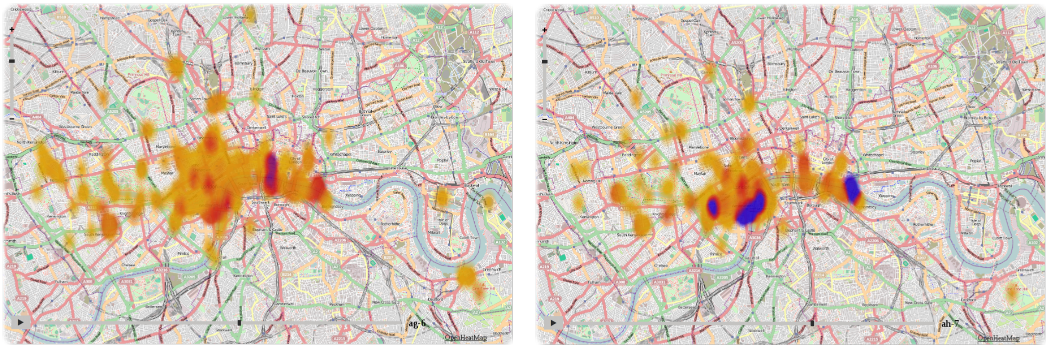
In section 4 we note that although both user group 6 (which is enriched with Canadians, Australians, and Malaysians) and user group 7 (enriched with South Africans and Australians) have London as their most prefered city, there are considerable difference in the other cities they visit and in the places they visit within the cities they share in common. This observation begs the question: when members of each of these group visit London, do they look at largely different POIs?

To answer this question, we plot the distribution of their weights over POIs in central London in fig. 3. We observe that user group 6 prefers a more difuse set of POIs, including peripheral attractions such as the Prime Meridian in Greenwich and the neighborhood Kensington. User group 7, on the other hand, prefers a few central POIs along the River Thames such as Westminster, Big Ben, and Whitehall Court; perhaps the members this group are inclined to take boat tours. An interactive version of these and other maps is available in the supplement.

Limitations & Future Work: Traditional research carried out by tourist boards involves random surveys of travellers. While such surveys are labor intensive, they can claim to be representative of all tourists. Perhaps the most major limitation of the Flickr is that it is not likely to be representative of the general tourist population. In future work, we could measure the representativeness of our data by comparing it with data gathered by a tourist agency.

In addition to not providing a representative sample of any tourist market, some tourist markets may be left out altogether. For example, few Asians use Flickr, leaving us with little to say about this quickly expanding demographic. Furthermore, we found that the Chinese users were very similar to the US, Canadian, and Australian users. It could be the case Flickr users in China are of a Western background, such as expats, while native Chinese use other photo sharing websites.

We know little about the tourists whom we have clustered because most Flickr users have incomplete profile pages. In future work we could learn more about users by analyzing the tags that they use. With an appropriate classifier and training data (available from the users who do fill in their profiles) a user's choice of tags may be able accurately reveal her primary language or home location.



(a) User Group 6

(b) User Group 7

Figure 3: User groups 6 and 7 are similar in that they both prefer London, but different in that they prefer a different set of POIs.

References

[1] For example, see the reports at http://www.statistics.gov.uk/hub/people-places/people/tourism and http://www.failteireland.ie/Research-Statistics/Surveys-and-Reports.

[2] See http://mlg.ucd.ie/datasets for supplemental information.

[3] R. Ahas, A. Aasa, A. Roose, Ü. Mark, and S. Silm. Evaluating passive mobile positioning data for tourism surveys: An estonian case study. *Tourism Management*, 29(3):469–486, 2008.

[4] M. Clements, P. Serdyukov, A.P. de Vries, and M.J.T. Reinders. Using flickr geotags to predict user travel behaviour. In *Proc. 33rd International SIGIR conference on Research and development in information retrieval*, pages 851–852. ACM, 2010.

[5] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proc. 18th international conference on World Wide Web (WWW '09)*, page 761. ACM, 2009.

[6] M. De Choudhury, M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel, and C. Yu. Automatic Construction of Travel Itineraries using Social Breadcrumbs. In *Proc. 21st ACM conference on Hypertext and hypermedia*. ACM, 2010.

[7] R.A. Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[8] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–.

[9] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–91, 1999.

[10] C.J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–79, October 2007.

[11] D. Müllner. Modern hierarchical, agglomerative clustering algorithms. *Arxiv preprint arXiv:1109.2378*, 2011.

[12] E. Osnos. The grand tour. *The New Yorker*, pages 50–60, April 2011.

[13] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[14] A.M. Schonland and P.W. Williams. Using the internet for travel and tourism survey research: Experiences from the net traveler survey. *Journal of Travel Research*, 35(2):81, 1996.

[15] A.V. Seaton and C. Palmer. Understanding VFR tourism behaviour: the first five years of the United Kingdom tourism survey. *Tourism Management*, 18(6):345–355, 1997.

This work is supported by Science Foundation Ireland under grant 08/SRC/I140: Clique: Graph and Network Analysis Cluster.