Ads and the City: Considering Geographic Distance Goes a Long Way

Diego Saez-Trumper*
Web Research Group
Universitat Pompeu Fabra
Spain
diego.saez@upf.edu

Daniele Quercia
Computer Laboratory
University of Cambridge
United Kingdom
daniele.quercia@cl.cam.ac.uk

Jon Crowcroft
Computer Laboratory
University of Cambridge
United Kingdom
jon.crowcroft@cl.cam.ac.uk

ABSTRACT

Social-networking sites have started to offer tools that suggest "guests" who should be invited to user-defined social events (e.g., birthday parties, networking events). The problem of how to recommend people to events is similar to the more traditional (recommender system) problem of how to recommend events (items) to people (users). Yet, upon Foursquare data of "who visits what" in the city of London, we show that a state-of-the-art recommender system does not perform well - mainly because of data sparsity. To fix this problem, we add domain knowledge to the recommendation process. From the complex system literature in human mobility, we learn two insights: 1) there are special individuals (often called power users) who visit many places; and 2) individuals go to a venue not only because they like it but also because they are close-by. We model these insights into two simple models and learn that: 1) simply recommending power users works better than random but is far from producing the best recommendations; 2) an item-based recommender system produces accurate recommendations; and 3) recommending places that are closest to a user's geographic center of interest produces recommendations that are as accurate as, if not more accurate than, item-based recommender's. This last result has practical implications as it offers guidelines for designing location-based recommender systems and for partly addressing cold-start situations.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms

Keywords

Advertisements, Mobile, Social Marketing

*This work was done while visiting the Computer Laboratory of the University of Cambridge.

RecSys'12, September 9–13, 2012, Dublin, Ireland.

1. INTRODUCTION

Social media sites have been recently testing features that return lists of people ("guests") users might want to consider inviting to their events (e.g., law firm parties, birthday parties, *PR*'s club invitations) [5]. Guests are selected based on relevance to the event and to the other fellow guests.

The problem of predicting relevant "guests" for venues or events has thus started to receive attention on the Web but has not been fully explored on mobile-social media platforms such as Foursquare, as discussed in Section 6 on "Related Work". One way of recommending venues to people is to use existing Web-based collaborative filtering algorithms. In Section 2, we show that such algorithms are not effective, mainly because of data sparsity: a venue is visited, on average, by very few users. Therefore, we propose two simple techniques for "recommending guests" that are reasonably accurate and scalable, and whose recommendations are easy to explain. In so doing, we make two main contributions:

- We put forward two proposals - a Bayesian model and a linear regression - that incorporate domain knowledge from the literature of human mobility and that cope with data sparsity (Section 3).
- We evaluate how the models perform against Foursquare data for the whole city of London (Section 4). We find that the simplest model - linear regression - returns the most accurate recommendations for all types of venues.

Before placing this work in the context of relevant literature (Section 6) and concluding (Section 7), we discuss some open questions (Section 5), including that of when our models do *not* work (and, consequently, where future work should go).

2. COLLABORATIVE TARGETING: UNFIT

To begin with, we state our research problem.

> **Problem Statement:** Given a venue (e.g., Italian restaurant), select individuals who are likely to visit it.

This simple problem, if solved, might enable a variety of applications, which include target advertising, commercial property evaluation, and social marketing (as we shall discuss in Section 5).

The problem might be formulated in simple "recommender system" terms - that is, it is the problem of how to recommend venues (items) to people (users). One way of solving it is to run a state-of-the-art matrix factorization algorithm on the inverted *venue-by-people* matrix (whose value m_{ij} is 1, if user j checked-in in venue i; 0 otherwise) and obtain, for each venue, a list of people who

Category	#Venues	#Users
food	1,293	1,566
nightlife	1,075	1,207
travel	850	1,744
home/work/etc.	411	1,037
shops	362	878
arts&entertainment	348	841
parks&outdoors	184	363
education	49	117
Total	4,572	3,110

Table 1: London Foursquare Data. Number of users and venues across venue categories.

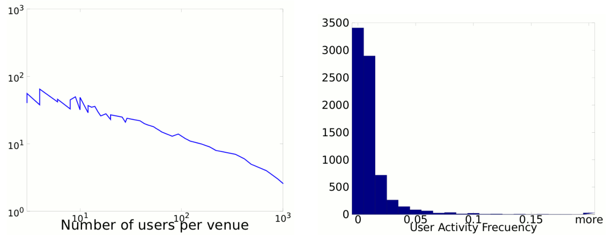
Category	Precision@10	Recall@10
food	0.013	0.012
nightlife	0.019	0.018
travel	0.004	0.005
shops	0.003	0.003
home/work/etc.	0.001	0.001
arts&entertainment	0.000	0.000
parks&outdoors	0.000	0.000
education	0.000	0.000

Table 2: Implicit SVD's Precision and Recall across Categories.

might like to visit it. We do just that: we use the state-of-the art *Implicit SVD* method introduced by Hu, Koren and Volinsky [10] and implemented within the Mahout framework. To evaluate its effectiveness, we measure its precision and recall on the following dataset.

Dataset. Foursquare is a mobile social-networking application that allows registered users to share their presence in a venue (e.g., share their "check-in" in a restaurant) with their social contacts. Users can share their check-ins not only on Foursquare but also on Twitter and Facebook. Each venue is associated with a category (e.g., "nightlife", "food") and a sub-category (e.g., "bar", "club", "Italian restaurant"). In 2011, Cheng *et al.* collected 22 million check-ins of 225,098 users [3]. We take the 228,625 check-ins in Greater London, which are generated by 29,044 users across 7,205 venues. To this data in the form *(user,venue)* pairs, with further crawling, we add each venue's category and subcategory. After considering venues and users that disjointly appear at least twice in our *(user,venue)* pairs, we end up with 3,110 users and 4,572 venues in the city of London. Table 1 breaks statistics about users and venues down into the different categories. One can, for example, see that food venues are numerous and attract many users, while educational venues are rare but proportionally attract more users.

Implicit SVD Performance. We arrange this data in a *venue-by-user* matrix and measure the *Implicit SVD*'s precision and recall. For each venue, precision is the probability that a recommended user is relevant ($\frac{relevant \cap recommended}{recommended}$), while recall is the probability that a relevant user will be recommended ($\frac{relevant \cap recommended}{relevant}$). By relevant, we mean users who visited the venue. Also, we consider that the recommendation list for each venue contains the *top-10* recommended users. The results reported in Table 2 shows that precision and recall are extremely low - for some categories, they are even zero. These appalling results have a clear explanation - the data is sparse. There are too few people



(a) Visitors per Venue (b) User Activity

Figure 1: *(a)* Number of Visitors per Venue; *(b)* Frequency Distribution of User Activity: this is a user's fraction of visited locations over the total ones.

going to the same venue; indeed, the number visitors per venue is power law (Figure 1(a)).

It thus seems that an alternative mechanism for recommending people is needed. But what sort of mechanism should we use? The widely-used classification algorithm of *SVM* does not work in the presence of data sparsity [18]. Therefore, we need a solution that: 1) is robust to sparsity; and 2) integrates domain knowledge (after all, our goal is to model how people "move" as much as is to model their preferences).

3. DOMAIN-AWARE RECOMMENDATION

We take these two requirements and translate them into a solution that unfolds in three steps:

1. Incorporate domain knowledge from the complex system literature in human mobility (Section 3.1);
2. Deal with data sparsity by using item-based collaborative filtering to model user preferences (Section 3.2);
3. Integrate the previous two steps into a Bayesian model and a linear regression (Section 3.3).

3.1 (Individual) Closeness

For starters, one might go to a venue not only because one likes it but also because one is nearby. Thus, leaving out the users' taste from a moment, one can model the probability of an individual visiting a venue as $p(go|close)$ - i.e., the probability of going to a venue given that it is close - and can do so using Bayes' Law:

$$p(go|close) \propto p_{close} \cdot p_{go} \quad (1)$$

where p_{close} is the probability of the user being close (being at a certain (log) distance), and p_{go} is the probability of a user going to any venue:

$$p_{go} = \frac{\text{\#venues visited by user } u}{\text{total \#venues}} \quad (2)$$

This latter probability reflects the general activity of a given user, which is a skewed distribution (Figure 1(b)), as one would expect: the vast majority of users visit few places, while a tiny fraction of (power) users (0.3%) visited roughly 20% of the London venues (within a category).

Literature: How people move. Scientists have long wondered how to measure something as ephemeral as movement. Early studies suggested that humans wander in a random fashion, similar to a so-called "Levy flight" pattern displayed by foraging animals. In 2006, to track human movements, researchers used more than half a million US one-dollar bills as a proxy measure and analyzed their movements as they were passed around over five years [2]. They found many short movements and occasional longer ones. Similar patterns were found by Gonzalez *et al.* who studied the trajectories of 100,000 mobile phone users tracked for six months [7]. These researchers found that people are regular, in that, the vast majority of them move around over a very short distance (from 5 to 10km) and make regular trips to the same few destinations such as work and home on a daily basis (70 percent of the time they were found in their two most frequently visited locations); people occasionally make longer trips when they, for example, go on vacation. More recently, Cheng *et al.* analyzed the movement of Foursquare users across venues and found similar patterns: a mixture of short, random movements with occasional long jumps. As such, the vast majority of users had a small radius of gyration - typically less than 10 miles [3].

Considering Geographic Closeness. To sum up, upon different types of movement (derived from dollar bills, mobile phones, and mobile social-networking applications), researchers in different disciplines have independently concluded that people rarely stray from familiar areas - they travel to a limited number of nearby locations and, consequently, short-range movements are more frequent than long-range ones (i.e., the frequency distribution of distance is exponentially distributed). This is also the case in our London data: Figure 2 plots the probability of one's traveling a certain distance for different venue categories. The distributions (for different categories) are very skewed and all fit the same distribution:

$$p_{close} = k_1 \frac{1}{d_{ui}^{\alpha}} \quad (3)$$

where d_{ui} is the distance between the user's (u's) center of geographic interest - which is center of mass or barycenter computed considering the locations where the user has previously checked-in - and the venue i. Interestingly, different venue categories are associated with different α, and the higher α, the less distance matters in one's choice when visiting a venue. Table 3 reports the α's for the different categories. The highest α (2.22) is associated with venues in the category "travel": those include train stations and bus stations, and it makes sense that people travel farther when going to places of limited supply (e.g., not all neighborhoods have a train station). The lowest α's are registered for venues in the categories "nightlife" and "home/work/etc.". That is, one's center of geographic interest revolves around home and work locations, and when going to bars, one goes to nearby ones.

Considering Power Users. Another conclusion from the literature is that not all mobile users are equally mobile. Individuals display significant regularity, yet, when compared to each other, there are few users who travel a lot, while the vast majority have limited travel activity. By framing the problem probabilistically, expression (1) is able to account for those special (power) users. It does so with p_{go} in expression (2), which reflects the extent to which one is a power user or not.

3.2 Likes

The model in expression (1) has only considered whether one user is close or not and whether is a power user or not; but the model has not taken into account personal preferences. To fix that, we need to compute $p(like|go)$ - we need to compute the extent to which a user visits venues that are predictable from his/her past visits/likes. However, to do so, we need a way to measures a user's $like$s. Since our data is sparse (Section 2), we measure likes not

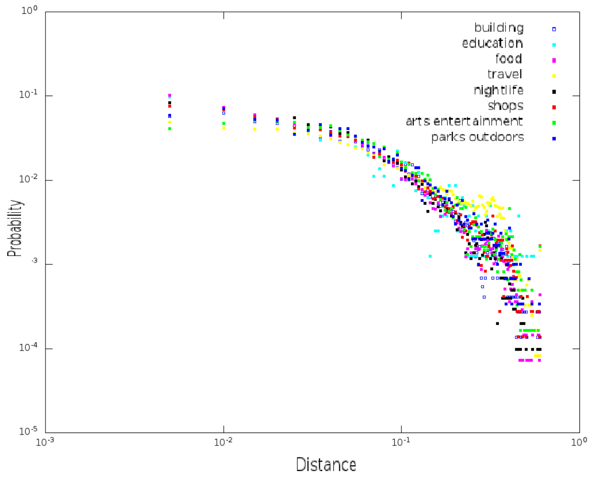


Figure 2: Probability of one's traveling a certain distance across different types of venues (best seen in color).

Category	α
food	1.64
nightlife	1.61
travel	2.22
home/work/etc.	1.62
shops	1.64
arts&entertainment	1.64
parks&outdoors	1.68
education	1.93

Table 3: Why People Visit Different Types of Venues. The higher α, the more one travels farther than usual to reach the venue in that category.

based on similarity among users but among venues. That is, we use an *item*-based collaborative filtering [21], which has been found to work well in such situations: "Unlike traditional collaborative filtering, the algorithm also performs well with limited user data, producing high-quality recommendations based on as few as two or three items." [15]. Rather than matching the user to other similar users, item-to-item collaborative filtering matches each of the user's venues with similar venues. A common way of computing the similarity between two venues is to compute the cosine similarity between two binary vectors: each vector reflects a venue, and a vector's i^{th} position reflects whether the i^{th} user visited the venue or not. Upon a so-constructed venue similarity table, the algorithm finds, for each user, the venues similar to the ones previously visited by the user.

We apply the item-based collaborative filtering algorithm on the $user$-by-$venue$ matrix and obtain a rating l_{ui} for each user u and venue i. Figure 3 shows the distribution of the predicted ratings. Upon these ratings, we compute $p(like = l_{ui}|go)$, which is the fraction of venues i visited by u that have predicted ratings l_{ui}:

$$p(like = l_{ui}|go) = \frac{\text{\#venues visited by user } u \text{ with rating } l_{ui}}{\text{total \#venues visited by user } u} \quad (4)$$

3.3 Putting All Together

Having users' whereabouts and preferences at hand, we now

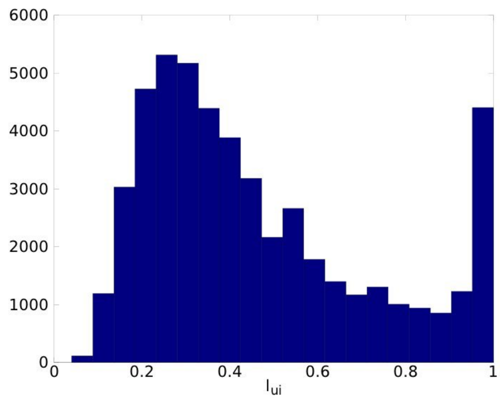


Figure 3: Distribution of Predicted Ratings.

need to predict which users are likely to be at a certain venue. We do so using a Naive Bayesian model, a Bayesian model, and a linear regression.

Naive Bayesian modeling. One simple way of modeling all the three factors together is to compute $p(go|like, close)$ using Bayes' Law:

$$\begin{aligned} p \quad & (go|like, close) \propto \\ \propto \quad & p_{like} \cdot p_{close} \cdot p_{go} \\ \propto \quad & p_{like} \cdot p_{go} \cdot k_1 \frac{1}{d_{ui}^{\alpha}} \end{aligned}$$

For each pair *(user, venue)*, we compute p_{close} with expression (3) and p_{like} with (4); and for each user, we compute p_{go} with (2). The importance of venue i for user u is then proportional to the above $p(go|like, close)$, and we call it $rank_{u,i}$.

Bayesian modeling. The previous model assumes that whereabouts and preferences are independent. This might well be not the case: those addicted to luxury goods will often be found near Bond Street (a major shopping street in the West End of London with many high price fashion shops). Here preference and whereabout go hand in hand. To go beyond independence, we could model jointly the two attributes:

$$p(go|like, close) = \frac{p_{like|go,close} \cdot p_{go|close}}{p_{like|close}}$$

where:

$$p_{like|go,close} = \frac{\text{\#venues visited by user } u \text{ at distance } d_{ui} \text{with rating } l_{ui}}{\text{\#venues visited by user } u \text{ at distance } d_{ui}}$$

$$p_{go|close} = \frac{\text{\#venues at distance } d_{ui} \text{ visited by user } u}{\text{\#venues at distance } d_{ui}}$$

$$p_{like|close} = \frac{\text{\#venues at distance } d_{ui} \text{with rating } l_{ui}}{\text{\#venues at distance } d_{ui}}$$

Linear Regression. Another approach for combining preferences and whereabouts is to run a linear regression:

$$rank_{u,i} = \alpha + \beta_1 I_{like} + \beta_2 I_{close} + \beta_3 I_{close} \cdot I_{like}$$

where I's are normalized values of whereabouts and preferences: I_{close} is $\frac{1}{log(d_{ui})}$ (the logarithm because the frequency distribution of distance is very skewed), and I_{like} is l_{ui}. The product $I_{close} \cdot I_{like}$ controls for interaction effects between whereabouts and preferences.

4. EVALUATION

The goal of this work is to predict which users are more likely to visit a given venue. To ascertain the effectiveness of our proposed techniques at meeting this goal, we need to select a desirable metric, measure it, and interpret those measurements. We execute these three steps next.

Metric. We need to find a measure that reflects the extent to which the predicted users for a venue are those who actually visited the venue. One such measure is called percentile-ranking [10]. The percentile-ranking $rank_{u,i}$ of user u for venue i ranges from 0% to 100%: it is 0%, if user u is first in venue i's recommendation list; it is 100%, if the user is last. Percentile-ranks have the advantage over absolute ranks of being independent of the number of users. Our quality measure is then the total average percentile-ranking:

$$\overline{rank} = \frac{\sum_{u,i} gone_{u,i} \cdot rank_{u,i}}{\sum_{u,i} gone_{u,i}} \tag{5}$$

where $gone_{u,i}$ is a flag that reflects whether user u was in venue i: it is 0, if u was not there; otherwise, it is 1. The lower $\overline{rank}$ for a list, the better the list's quality. For random predictions, the expected value for $rank_{u,j}$ is 50% (averaging infinite placements of users for a venue returns the middle position of the list). Therefore, $\overline{rank} < 50\%$ indicates an algorithm better than random. To ease illustration, we covert percentile ranking into ranking accuracy, which is 1, if the percentile ranking is 0% (best); and it is 0, if the percentile ranking is 50% (random):

$$accuracy = \frac{50\% - \overline{rank}}{50\%} \tag{6}$$

Accuracy would be 0 for a random predictor (baseline), and would be 1 for an ideal (oracle) predictor.

Execution. To measure the ranking accuracy, we run a *10-fold* cross validation. That is, we divide the dataset into 10 segments, we take one segment s at a time, consider it to be the testing set, and go through the following steps:

1. For each venue in the *training* set (the venues in all segments other than s), associate it with the users who visited that venue.

2. Train the model using the venues (and corresponding visitors) in the training set.

3. Use the trained model to then infer a rank list of users who are likely to go to each venue in the *testing* set (the venues in s).

We finally compare the users predicted for each venue to those who actually visited it (those who are in the ground truth).

Results. Figure 4 reports the ranking precisions for the individual components of the Bayesian models (first three bars in each venue category) and for the overall models (Naive in the fourth bar, Bayesian in the fifth, and Linear Regression in the sixth). Starting from the first bar in each category (p_{go}), one sees that recommending power users works better than random (accuracy is always well above zero): the more so for shops (.38) than for arts&entertainment venues (.24). Considering only nearby places (second bar in each set) returns more accurate rankings - again, more for shops (.60) than for arts&entertainment venues (.38). However, if one consider only past user preferences (third bar p_{like}), then accuracy is

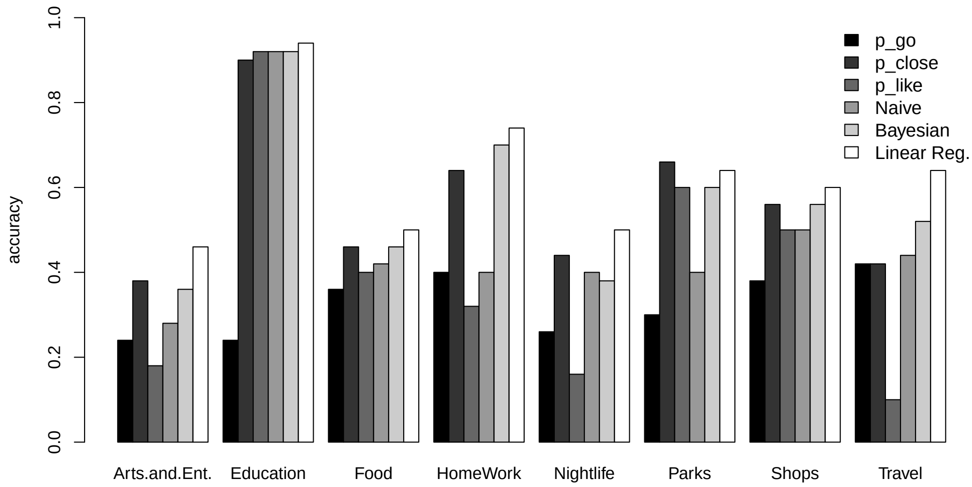


Figure 4: Rank Precision. The rank goes from 0 (random baseline predictions) to 1 (relevant user always ranked first in the recommendation list).

comparable to that of recommendations based on proximity (second and third bars do not differ much). This suggests that the simple concept of geographic distance is as important as that of the user's taste in all venue categories. It also suggests that, by only knowing where a user usually hangs out (without any information on the user's taste), one can produce reasonable recommendations (ideal for cold start situations). If we then combine these previous elements in a Naive Bayesian model, results do not improve; on the contrary, they are worse than those offered by simple geographic proximity for venues in the categories "food" and "arts&entertainment". That might be because the model treats its components as though they were completely independent. However, on average, the Pearson correlation coefficients ρ between each pairs of components are small: $\rho(p_{go}, p_{like}) = .13$, $\rho(p_{go}, p_{close}) = .05$, and $\rho(p_{like}, p_{close}) = .21$. Yet, looking at the fifth bar in each set, one registers improvements with the traditional Bayesian model (in which dependencies are model). Another common reason for which Naive Bayesian does not work well in certain situations is that the addition of redundant components and arbitrary discretization of the random variables skew the learning process, and that seems to be the case here. Indeed, the linear regression (last bar) - which just models taste, whereabouts, and interactions between the two - works best in all categories. As one would expect, for categories characterized by less data sparsity and periodic patterns (e.g., education buildings), the models perform extremely well (accuracy above .90): the performance tend to be comparable to, if not better than, those registered in Web applications.

5. DISCUSSION

Putting Results into Context. For the case of recommending shows on set top boxes, Hu *et al.* had 17K of unique programs (roughly 2× our number of venues) and 32M non-zero ratings (140× ours). In that context of less sparsity, they managed to achieve a ranking accuracy as good as .80 (upon learning from 200 distinct factors). Thus our results with the linear regression (always above .50 and above .60 for categories such as 'shops' and 'parks' and 'travel') are comparable to those reported in the literature in far more favorable contexts (140× less sparsity). Also, the percentile rankings are expected to slightly improve in more 'realistic' situations. To see why, consider that our data has been collected within a limited time window; by contrast, if one were to crawl the entire Foursquare history, then the resulting data would be still sparse but less so, and, as such, the prediction results would improve, as we have already registered with the category "educational" venues for which the accuracy was above .90.

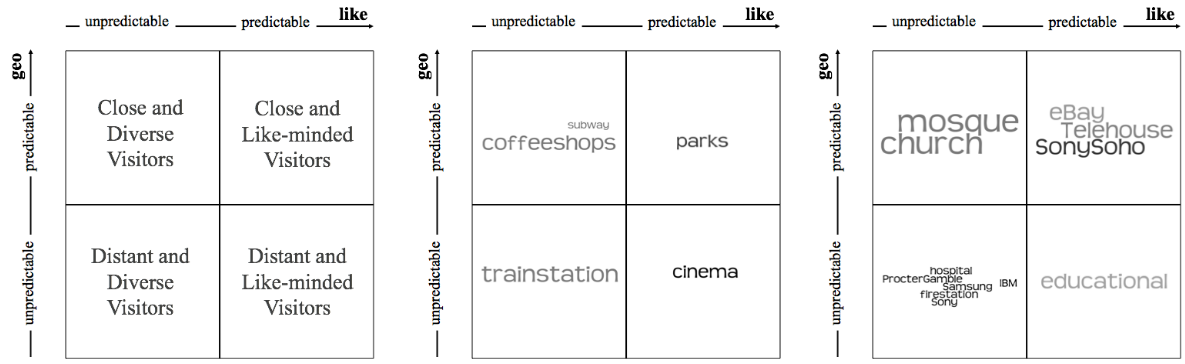
When It Does not Work. When putting forward new predictive models, one often tends to focus on favorable situations in which predictions are best. Next, we briefly focus on the opposite case - we focus on situations in which prediction are worst. The idea behind this exercise is to find out which aspects future models should consider to increase accuracy. To this end, we run a qualitative study. For each venue i, we compute four predictability and unpredictability measures upon the following quantities: $gone_{ui}$, which reflects whether user u visited venue i; the geographic decay constant α taken from Table 3; the predicted rating l_{ui} for user u and venue i; and the distance d_{ui} between u's geographic center of interest and venue i. More specifically, upon these quantities, for each venue i, we compute:

Geo Predictability. The higher it is, the more the venue's visitors are predictable based on distance. It is higher for venues (e.g., bakery shops) whose visitors travel nearby:

$$P^i_{geo} = \frac{\sum_u \frac{1}{log(d^{\alpha}_{ui})} \cdot gone_{ui}}{\sum_u gone_{ui}}$$

It is the average inverse (log) distance for the venue's visitors.

Geo Unpredictability. The higher it is, the less its visitors are predictable based on distance. It is higher for venues (e.g.,



(a) Illustration (b) All Categories (c) Category 'Building'

Figure 5: Four-quadrant Predictability Box. Quadrants are defined by the venue's unpredictability and predictability measures, which are based on visitors' geographic closeness (rows) and likes (columns).

airports, high-end restaurants) whose visitors travel farther:

$$U^i_{geo} = \frac{\sum_u log(d_{ui}) \cdot gone_{ui}}{\sum_u gone_{ui}}$$

It is the average (log) distance for the venue's visitors.

Like Predictability. The higher it is, the more its visitors are predictable based on past preferences (past likes). It is higher for venues whose visitors have common preferences:

$$P^i_{like} = \frac{\sum_u l_{ui} \cdot gone_{ui}}{\sum_u gone_{ui}}$$

It is the average predicted ratings for the venue's visitors.

Like Unpredictability. The higher it is, the less its visitors are predictable based on past preferences. It is higher for venues whose visitors have diverse preferences:

$$U^i_{like} = \frac{\sum_u \frac{1}{l_{ui}} \cdot gone_{ui}}{\sum_u gone_{ui}}$$

It is the average inverse predicted ratings for the venue's visitors.

We create four tables that contain the *top*-10 venues ranked by each of those four measures and ask three coders (three Londoners with diverse background - architect, barrister, and medical doctor) to build predictability boxes of the kind in Figure 5(a). For them, that translated into ordering venue categories that are predicted (hard to predict) by geographic distance based on the table ranked by P^i_{geo} (by U^i_{geo}), and categories that are predicted (hard to predict) by user preferences based on the table ranked by P^i_{like} (by U^i_{like}). We consider only the answers for which two out of three coders or all three have independently agreed. In Figures 5(b) and 5(c), word size is proportional to the coders' agreement. For all venue categories (Figure 5(b)), the unpredictable venues (predicted neither by closeness nor by taste) are train stations. That is because train stations are often far from where one hangs out and do not reflect a specific taste in, say, music, bars, clubs, or food. By contrast, local parks and outdoor activities are predictable either by closeness or by taste, suggesting that people prefer their local parks over bigger parks (they stay close), and that residents of the same area tend to be like-minded (a tendency often called "geographic sorting" [1]). Closeness is more informative for predicting visits to coffee shops (one tends to go to local coffee shops); while user taste is more informative for cinemas in central London areas, where diversified choice of movies motivates visitors to travel farther than usual. For the specific category "buildings" (Figure 5(c)), the unpredictable venues (predicted neither by closeness nor by taste) are companies such as IBM, Procter&Gamble, Samsung whose headquarters are in suburban areas where people with diverse background work but do not hangout, not least because of limited supply of amenities. By contrast, the behavior of employees (mostly interaction designers) of Sony, eBay, Telehouse working in central areas like Soho is predictable either by closeness or by taste. Finally, closeness is more informative for predicting visits to mosques and churches (one tends to go to local religious venues); while user taste is more informative for visitors of university (e.g., UCL's, Birkbeck's) facilities in central areas. From these qualitative results, one can extrapolate two key insights:

1. Predictable situations are those in which people: a) stay close because they have what they need at hand; or b) congregate in places where other like-minded people tend to be (e.g., local parks and cinemas).

2. By contrast, unpredictable situations are those in which people: a) travel because they do not have what they want at hand; or b) go to places that attract individuals of very diverse backgrounds (e.g., coffee shops, train stations).

Future work should go into models that are able to simultaneously account for these (at times) conflicting situations.

Applications. The practical implications of this work go beyond traditional applications of recommender systems:

Target Advertising. The first step when promoting new nightclubs, bars and restaurants is often to identify the target market. Thus, knowing the kind of people who are willing to go to, say, certain restaurants or bars (which is what this work is about) translates into low-cost marketing strategies for bars and restaurants that are willing to attract new crowds.

Commercial Property Evaluation. This is the process of identifying and quantifying the value of commercial properties and is generally carried out by experts who analyze properties similar to the one being valued. A primary factor that affects this assessment is location, yet this factor is generally quantified based on the valuer's expert knowledge of a locality. More recently, well-informed ways of valuing properties have been proposed, and they rely on the *creation and maintenance* of GIS-based property valuation databases. These databases (especially those for commercial properties) might well be enriched by this work - in particular, by knowing how close a venue is to its target audience (the higher the number of potential users who like a type of venue in a neighborhood, the higher the venue's value).

Social Marketing. Social marketing can be defined as a research-driven approach to promote voluntary behavior change in a priority population. A case in point is "Stop the Sores", a social marketing campaign designed to increase syphilis testing in Los Angeles County [17]. Social marketing has its foundation in consumer marketing and consists of three key elements: market research [24], audience segmentation [8], and branding [12]. The second element of segmentation is related to this work and is essential for developing campaign messages that resonate with the target population and helps in identifying the largest or highest-risk subgroup (e.g., swingers, men having sex with men) at minimal cost.

Scalability. The two main parts of this work - which model whereabouts and preferences - are highly scalable:

Whereabout Part. This requires to know a geographic point for each user (where an individual usually hangs out) and one single decay constant α (which is universal in that it equally applies to all users). Learning a point per user and a constant for all is extremely scalable. In addition to being scalable, the models are likely to be generalizable, not least because they have been built upon previous general rules of people's wanderings [2, 3, 7], and, being general, they are also likely to work for any instance of mobility (not only for Foursquare users).

Preference Part. This translates into item-based collaborative filtering. The (computationally) expensive part of this algorithm (venue similarity table) can be computed offline, while what needs to be computed online - matching the user's venues with similar ones - scales independently of the total number of venues and total number of users, in that, it only depends on the number of venues each single user has visited (which is generally extremely low).

6. RELATED WORK: WEB DOCET

The problem of recommending events has been initially tackled on the Web. In this context, researchers have mainly worked on detecting and tracking events [11, 13]. They initially considered how textual content evolves over time and left out network effects. Zhu and Sasha then started to model social interactions and topic evolutions by treating these two elements separately [26]. More recently, Lin *et al.* built a model that considers these two elements simultaneously and showed that it worked upon two very different types of data - Twitter and DBLP [14]. After detecting events, one can then recommend them. That is what Daly and Geyer *et al.* did: they built a system that recommends events in an internal event management service and proposed a new way of recommending events to new users [4]. Before that, Minkov *et al.* had run large user studies in which they evaluated the effectiveness of different strategies for recommending academic talks [16]. They found that, in a situation of limited data sparsity, collaborative filtering approaches work better than content-based ones. The recommendation process generally relies on user ratings but has also been enriched by social networks at times. A case in point is Golbeck *et al.* who built a recommender system that integrates social networks to offer well-informed movie recommendations [6].

Hence past work on recommending events has mostly gone into Web platforms, while mobile ones have been investigated only recently. Takeuchi and Sugimoto proposed a system that recommends shops based on past visited locations, and found item-based collaborative filtering to work reasonably well [23]. Ricci and Nguyen proposed a system that recommend nearby restaurants using a critique-based model [20]. More recently, for major mobile social-networking services, Scellato *et al.* studied their geographic properties at scale and suggested that these properties could well inform venue recommendation in large cities [22]. Upon mobile phone data in the metropolitan area of Boston, Quercia *et al.* studied strategies for recommending large-scale events (e.g., concerts, baseball matches) and showed how different types of events require different recommendation strategies [19].

Shifting attention from recommending events to recommending people, one sees that most of the work has again gone into Web platforms. Within an enterprise social network, Guy *et al.* proposed ways to recommend people a user is not likely to know but might be interested in [9]. Few months ago, Facebook launched a new feature called "suggested guests" [5]: this returns a list of people (three at the time) a user might want to consider inviting to their event, and the list is compiled based on relevance to the event and to the people who are attending. Since work on recommending people for events has just started on the Web, it comes as no surprise that little work about it has gone into mobile social-networking platforms.

7. CONCLUSION

We have studied different strategies for recommending "guests" for real-world venues and, not surprisingly, found that results are best not only for venues with considerable historical data (e.g., educational institutions) but also for venues that are visited regularly (e.g., work locations). For other types of venues such as restaurants and bars, geographic closeness plays a very important role. Combining user preferences and geographic closeness has the expected result of offering more accurate recommendations, and that result can be achieved by using very simple models - Bayesian or linear regression. Being simple, these models not only are scalable and cost efficient but also produce recommendations that are easy to explain. The main criticism for the new Facebook "suggested guests" feature has been that it "does not offer... any sort of context" [5]. Our recommendations - which depend on whether one has visited similar locations or whether one often hangs out in certain neighborhoods - are likely to be easier to explain than those produced by black-box approaches. In the future, we will work in this direction: on how to recommend "guests" in ways that are easy to explain and that increase serendipitous encounters [25].

Acknowledgments. We thank Tamas Jambor for his help on the *Implicit SDV* algorithm and his valuable insights, and the reviewers for their constructive comments. This work was funded by RCUK through the Horizon Digital Economy Research grant (EP/G065802/1) and by the Spanish Economy and Competitiveness Ministry through the HIPERGRAPH project (TIN2009-14560-C03-01).

8. REFERENCES

[1] B. Bishop. *The Big Sort: Why the Clustering of Like-Minded America Is Tearing Us Apart*. Houghton Mifflin, May 2008.

[2] Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439, January 2006.

[3] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring Millions of Footprints in Location Sharing Services. In *Proceedings of the* 5^{th} *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[4] E. M. Daly and W. Geyer. Effective event discovery: using location and social information for scoping event recommendations. In *Proceedings of the* 5^{th} *ACM Conference on Recommender Systems (RecSys)*, 2011.

[5] B. Darwell. Facebook tests 'suggested guests' for events. In *Inside Facebook*, February 2012.

[6] J. Golbeck. Trust and nuanced profile similarity in online social networks. *ACM Transactions on the Web*, 3(4):12:1–12:33, September 2009.

[7] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196), June 2008.

[8] S. Grier and C. A. Bryant. Social marketing in public health. *Annual Review of Public Health*, 26(1), 2005.

[9] I. Guy, S. Ur, I. Ronen, A. Perer, and M. Jacovi. Do you want to know?: recommending strangers in the enterprise. In *Proceedings of the ACM Conference on Computer supported Cooperative Work (CSCW)*, 2011.

[10] Y. Hu, Y. Koren, and C. Volinsky. Collaborative Filtering for Implicit Feedback Datasets. In *Proceedings of the* 8^{th} *IEEE International Conference on Data Mining (ICDM)*, 2008.

[11] A. Ihler, J. Hutchins, and P. Smyth. Adaptive event detection with time-varying poisson processes. In *Proceedings of the* 12^{th} *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.

[12] K. L. Keller. Branding perspectives on social marketing. *Advances in Consumer Research*, 25(1), 1998.

[13] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the* 8^{th} *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2002.

[14] C. X. Lin, B. Zhao, Q. Mei, and J. Han. PET: a statistical model for popular events tracking in social communities. In *Proceedings of the* 16^{th} *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2010.

[15] G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[16] E. Minkov, B. Charrow, J. Ledlie, S. Teller, and T. Jaakkola. Collaborative future event recommendation. In *Proceedings of the* 19^{th} *ACM International Conference on Information and Knowledge Management (CIKM)*, 2010.

[17] A. Plant, J. A. Montoya, H. Rotblatt, P. R. Kerndt, K. L. Mall, L. G. Pappas, C. K. Kent, and D. Klausner. Stop the Sores: The Making and Evaluation of a Successful Social Marketing Campaign. *Health Promotion Practice*, 11(1), 2010.

[18] D. Quercia, H. Askham, and J. Crowcroft. TweetLDA: Supervised Topic Classification and Link Prediction in Twitter. In *Proceedings of the* 4^{th} *ACM International Conference on Web Science (WebSci)*, 2012.

[19] D. Quercia, N. Lathia, F. Calabrese, G. D. Lorenzo, and J. Crowcroft. Recommending Social Events from Mobile Phone Location Data. In *Proceedings of the* 10^{th} *IEEE International Conference on Data Mining (ICDM)*, 2010.

[20] F. Ricci and Q. N. Nguyen. Acquiring and Revising Preferences in a Critique-Based Mobile Recommender System. *IEEE Intelligent Systems*, 22(3):22–29, May 2007.

[21] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the* 10^{th} *ACM Conference on World Wide Web (WWW)*, 2001.

[22] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. In *Proceedings of the* 5^{th} *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[23] Y. Takeuchi and M. Sugimoto. CityVoyager: An Outdoor Recommendation System Based on User Location History. In *Ubiquitous Intelligence and Computing*, Lecture Notes in Computer Science, 2006.

[24] D. C. Walsh, R. E. Rudd, B. A. Moeykens, and T. W. Moloney. Social marketing for public health. *Health Affairs*, 12(2), 1993.

[25] Y. C. Zhang, D. O. Séaghdha, D. Quercia, and T. Jambor. Auralist: introducing serendipity into music recommendation. In *Proceedings of the* 5^{th} *ACM International Conference on Web Search and Data Mining (WSDM)*, 2012.

[26] D. Zhou, X. Ji, H. Zha, and C. L. Giles. Topic evolution and social interactions: how authors effect research. In *Proceedings of the* 15^{th} *ACM International Conference on Information and Knowledge Management (CIKM)*, 2006.