Noname manuscript No.
(will be inserted by the editor)

Targeted Interest-Driven Advertising in Cities Using Twitter

Aris Anagnostopoulos · Fabio Petroni · Mara Sorella

Received: date / Accepted: date

Abstract Targeted advertising is a key characteristic of online as well as traditional-media marketing. However it is very limited in outdoor advertising, that is, performing campaigns by means of billboards in public places. The reason is the lack of information about the interests of the particular passersby, except at very imprecise and aggregate demographic or traffic estimates. In this work we propose a methodology for performing targeted outdoor advertising by leveraging the use of social media. In particular, we use the Twitter social network to gather information about users' degree of interest in given advertising categories and about the common routes that they follow, characterizing in this way each zone in a given city. Then we use our characterization for recommending physical locations for advertising. Given an advertisement category, we estimate the most promising areas to be selected for the placement of an ad that can maximize its targeted effectiveness. We show that our approach is able to select advertising locations better with respect to a baseline reflecting a current ad-placement policy. To the best of our knowledge this is the first work on offline advertising in urban areas making use of (publicly available) data from social networks.

Keywords Targeted outdoor advertising · Twitter · Geotag

A preliminary version of this work appeared as a poster paper in the Proceedings of the 10th International AAAI Conference on Web and Social Media, (ICWSM 2016)

This research was partially supported by the Google Focused Research Award "Algorithms for Large-Scale Data Analysis" and by the EU FET project MULTIPLEX 317532.

Aris Anagnostopoulos, Fabio Petroni, Mara Sorella
Sapienza University of Rome
E-mail: {aris, petroni, sorella}@dis.uniroma1.it

1 Introduction

The computer-science research community has been involved significantly in the study of online advertising, with several workshops and conference tracks being dedicated to it. This is quite natural, as online advertising forms the main revenue source for many large or small internet companies. Yet, *outdoor advertising* (e.g., billboards or distribution of leaflets), remains the main advertising medium for several offline companies.

Numerous studies in marketing and advertising have demonstrated the effectiveness of outdoor advertising and have studied the effect of different characteristics of the form, location, and so on [7, 10, 16, 20, 34]. Indeed, especially for small companies, outdoor advertising is a very effective advertising medium [2]; however, one of its main drawbacks has been the limited opportunity to target a particular audience [20]. Indeed, it is easy for one to make a case for the effects of *targeted advertising* [8, 14]. A famous quote by John Wanamaker, a pioneer in advertising, states: “Half the money I spend on advertising is wasted. The trouble is, I don’t know which half.” Although social networks and media have been of tremendous help for online advertising, leading to a high degree of targeting and tailoring, its outdoor counterpart has only relied on traffic data and rough demographic estimates: other than allowing for limited targeting power, these strategies select highly crowded areas for billboards, leading to an overcluttering effect where the attention of customers, exposed to a high number of co-occurring ads, is lost. This lack of verified data on audience characteristics, has reportedly [27] limited the growth of the outdoor-advertising industry, preventing many advertisers from investing heavily in it.

The availability of geolocated social data provides the potential to change this situation: the proliferation of location and movement-tracing devices, such as accelerometers and GPS devices, the development of location based social networking services such as Foursquare, and the wide use of micro-blogging services such as Twitter, are able to provide a detailed characterization of user interests. Furthermore, information such as that collected by telephone companies on their volume load on cell phone antennas, as well as traffic and other types of sensors, can give a reasonably accurate estimate on the presence of citizens at various city streets and other locations. Such information, if processed and filtered appropriately, can be invaluable to advertisers for targeting potential clients, a process that is currently performed typically manually by outdoor advertising companies.

As an example, consider the city of New York. To be effective, the advertisements that one should place in SOHO are different from the ones that she should place in Park Avenue, and the ads to place in Astoria are different from Harlem. These examples pinpoint the opportunity for targeting, assuming we can have a way to characterize city locations based on the number of people passing every day, their interests, and so on.

We tackle the problem by proposing a new technique, which leverages public information from Twitter: we collect tweets’ geotags to obtain information about user trajectories and then we perform user profiling to identify the degree of interest of each user towards different topics corresponding to a predefined set of advertising categories. Intuitively, interests drive the way in which people are influenced by an ad. We combine this information with the collective mobility patterns of users sharing the same interests, to estimate, for each category, the most promising areas to place a relevant ad.

To assess the quality of the solution we perform validation on a test portion of the users to verify if those users, interested in some topics, will or not pass by the corresponding identified zones (thus having a chance to see the targeted ad). Furthermore, we use mobile communication usage data to measure how crowded is each zone, using it both as baseline, and to understand the difference between the zones found by our algorithms and the simply crowded areas. Our results show that even with a low budget in terms of the number of zones in which we can place an ad, for all the categories, we are able to cover a consistently higher portion of the interested users compared to simply placing ads at top crowded areas. Furthermore, we present some anecdotal evidence of the targeted interestingness of the discovered zones, both suggesting a possibly higher influencing effect, and giving insights on the applicability of this approach to achieve a characterization of urban dynamics of city areas shaped by people with common interests.

Contributions To summarize, we present the first work (to the best of our knowledge) that leverages the use of publicly available social-media data for the purpose of targeted outdoor advertising. As part of our approach, we present a method for characterizing city zones based on social-media data. We also present a set of algorithms for targeted outdoor advertising placement and we show that they outperform the most natural and mostly used baseline applying our evaluation measures. We believe that our approach can have a significant effect, especially for small- or medium-size companies or nonprofit organizations, where usually the available budget for advertising is limited.

In the next section we present some related work. In Section 3 we expose our approach and in Section 4 we evaluate it. We conclude with some discussion and some ideas for future work in Section 5.

2 Related Work

Despite being a natural application, there exists surprisingly little work by the computer-science research community on the use of social-sensing data for the use in outdoor advertising. Instead, there has been significant work on the use of social data for user profiling, as well as for the characterization of city neighborhoods. Naturally, there is also work by the marketing community in outdoor advertising. Here we summarize briefly the areas that are more related to our work.

User profiling in Twitter. A key element of our approach is understanding the interests of Twitter users. Many past works rely either on the the text of the tweets issued by an user himself, or the users he follows. Early works of this kind are based on bag-of-words and statistical approaches [4], whereas more recent works use topic-modeling techniques such as latent Dirichlet allocation and its derivatives [33]. The main problem behind such frequency-based text-mining approaches is that tweets are short, they are enforcing grammar rules loosely, and they often contain conversations about daily activities of users [31], making difficult the identification of meaningful topics. Wagner et al. [31] test different types of user-related information (tweets, retweets, bio, and lists) to understand if they convey interest-specific information and found that bio and list membership are the most discriminative to identify topical interests and expertise.

Lists are an organizational feature of Twitter, which allows users to create and manage curated lists of other users. A few works exploited Twitter lists to find latent attributes of Twitter users: lists are treated as topical containers to infer the interest of the contained users [9]. In practice, list names and descriptions can be translated into crowdsourced "tags" [35]: users are put in relevant lists because they are judged as topical authorities or "experts" by other users. Along this line, some works use lists meta-data to find experts [9, 31] and interests [3]. In Section 3 we use a similar approach to infer the interests of a user.

Urban computing using geotagged data. Urban computing [36] is an emerging field promoting acquisition and analysis of big and heterogeneous data generated by a diversity of sources in urban spaces (e.g., traffic flow, human mobility) to tackle the major issues that cities face, such as air pollution, increased energy consumption, and traffic congestion, for the purpose of understanding and improving the urban environment. Among the works that leverage geotagged data for this purpose, of particular interest are those that involve (1) finding events in a city [18,24,30,32]; (2) assisting in safety-critical situations (fires, floods, etc.) [29]; (3) finding local experts on Twitter [5], and (4) characterizing city areas such as neighborhoods in terms of the local activities [6,17]. These last works are the most relevant to ours; they use data from location-based services such as Foursquare to characterize city areas based on the type of venues present. Nevertheless, it may be the case that some people can frequent an area for latent reasons that cannot simply be captured by the venues or point of interests contained therein. Furthermore, data from location-based services are generally not public. For these reasons we characterize city zones by looking at the interests of the users who transit in them, as expressed by their Twitter activity.

Outdoor advertising. Outdoor advertising forms a crucial part of marketing science and it has attracted a very large attention by researchers in the area (e.g., [7,10,16,20,34]). Shannon et al. [26] describe a system that uses information obtained from a person's Facebook profile to deliver targeted ads to single users moving in pervasive environments, identified by their bluetooth devices. The authors are upfront with several privacy issues that arise in this context: advertisements that are highly targeted to one single user will undoubtedly be observed by other customers, which raises the likelihood of an ad being shown to a wider audience than desired, compromising some user's identity. Indeed, advertisements based on the aggregate interests and characteristics of the people in its environment would be safer to display on public screens. Furthermore, the applicability of the approach can be challenged: one could argue that not all devices have bluetooth turned on, and that users are generally not willing to share access to their Facebook profile. Quercia et al. [23] propose a system that infers people's preferences by combining location estimates from their mobile phones with listings of public events (like football games or music festivals), then it builds clusters of neighborhoods with similar composition of preferences. The main differences with respect to our work are that they employ a less granular scale of both datasets (mobile presence estimates versus geotags) and solutions (neighborhood clusters versus city grid cells) and the fact that the interests of the users are estimated using their mobility patterns themselves, whereas in our case we use an independent source of information, that is, topical information consumption habits on Twitter, which allows us to cover more advertising categories. Also, because of the lack of ground truth, the evaluation of the approach is mainly qualitative.

Finally, Liebig et al. [19] study the problem of untargeted indoor advertising using manually collected traces of users in public stations, which they use to train a pedestrian flow model (which can optionally integrate also GPS signals for mixed in/out traces). Indeed our setting is different as we consider targeted (interest-based) advertising.

We proposed a preliminary version of this work as a poster paper [1]. This new version contributes a revised problem definition, a set of preliminary analyses of our datasets, new evaluation metrics, and a more in-depth analysis considering also a new baseline. To the best of our knowledge our work is the first to make use of only public social-media data to perform targeted outdoor advertising.

3 Interest-Driven Urban Zone Ranking

In this section we formally define the problem and we describe a novel methodology to solve it.

3.1 Problem Definition

We have as input a set of geotagged tweets T made by a set of Twitter users U for whom we have access to profile information (Twitter username) during a given time period, and a fixed set of categories I to which both user interests and ads conform. For instance, we may have $I = \{\textit{Food}, \textit{Cinema}, \textit{Sports}, \dots\}$.

We partition the area spanned by the tweets into a set of n non overlapping city zones $Z = \{z_1, \dots, z_n\}$, such that each tweet's coordinates included in the geotag belong to a single zone. Previous works have considered various zone shapes: squared cells [22], convex hulls [15], or Voronoi polygons [25] resulting from spatial clustering. Our approach is oblivious to the particular choice, and in our experimental evaluation in Section 4 we adopt the squared, equally sized cells to conform to the granularity of a dataset used for the evaluation. Denote by $Z_u \subseteq Z$ the set containing all the zones where user $u \in U$ has issued at least one tweet, which we refer to as the *trace* of the user.

For a given interest i and zone z we define the value $\mathcal{E}(z,i)$ as the number of impressions to people interested in i that pass by zone z. We will discuss about how we estimate this value after we define the problem—for now we assume that it is known.

For each zone $z \in Z$ we assume that it costs $c(z)$ to perform advertising in it, by placing billboard ads or by distributing leaflets during a fixed time period. Finally, we have available some budget B, which we can use for our advertising campaign.

Our goal is, given a category $i \in I$, to compute a set of zones, and provide the top-k zones Z_i^* that will be the candidates for targeted advertising, while respecting the budget constraint. We, therefore, define:

Problem 1 (OfflineAdvertising) Given an interest category $i \in I$ and a budget B, select a set of k zones $Z^* = \{z_1^*, \dots, z_k^*\}$, with $z_j^* \in Z$, such that

$$\sum_{z \in Z^*} c(z) \leq B$$

that maximizes

$$\sum_{z \in Z^*} \mathcal{E}(z,i).$$

As we see later, to solve the problem, given a category, we rank the zones according to our estimate of the expected effectiveness (in terms of interested users reached) of an ad placed in the zone.

3.2 Discussion

Before we proceed into the details of how we tackle the OfflineAdvertising problem, we make some comments.

Combinatorially, the problem, in its general form, is NP-hard: it can model the knapsack problem. We assume to have integer or discretizable costs in the following, an assumption that can naturally be applied to currency-related costs. Likewise, we assume that B is integral. With these assumptions, the knapsack problem can be solved in pseudo-polynomial time and we have implemented such an algorithm for the study case in Section 4.3, which turns out to be sufficiently fast for the dataset sizes that we possess.

A much harder problem is the fact that the values $\mathcal{E}(z,i)$ are unknown and practically impossible to estimate: they would require knowing the interests of passengers and whether they have noticed an advertising sign. Therefore, in Section 3.3.2 we present different proxies for it, which are based on social-media data—this is actually one of the main contributions of this work. As a result, our approaches will not optimize the objective function but the modified objective functions in which the proxies of Section 3.3.2 are being used. In the experimental section we evaluate the effectiveness of the various proxies.

In this paper we mostly study the *unweighted* version of the problem. This is an important special case, in which we have $c(z) = 1$ for each $z \in Z$. Then we have that the number k of zones selected is fixed and equal to B: for instance, if $B = 10$ this means that we can target 10 city zones. In the unweighted case, a greedy algorithm that simply ranks the zones by decreasing effectiveness and picks the top-k zones is optimal. This choice allows us to study many aspects of the problem, while ignoring the specifics on the pricing in a particular city. For the unweighted settin, we will often refer to k as the budget, given that $B = k$. Subsequently, in Section 4.3, we study in detail the weighted case by considering the real costs for placing ads. The findings about our algorithms are similar to the unweighted case.

3.3 Methodology

Our approach consists of two components: (1) a method to identify the interests of the users towards the identified category set I and (2) a procedure to find, for each category, a ranking over the city zones. We describe them next.

3.3.1 Inferring User Interests

Our first goal is to infer users' interests. For this we started by profiling users based on their tweets' content. However, our preliminary findings indicated that these approaches were less effective for our objectives (they allowed us to profile only a small amount of users), so we decided to use the approach that we describe next.

To get information about the interests of Twitter users, we use the *follower–followee* relationship of Twitter. The main motivation behind this choice is that following a user is a form of subscribing to the information produced by the followee and, thus, an indication of interest to topics that interest the latter as well. Based on this idea, we use a technique similar to the one of Bhattacharya et al. [3]. We exploit Twitter *lists*, an organizational feature of Twitter, which allow users to create and manage curated lists of other users. Each list is characterized by a name and an optional description. Lists are mainly used to group followed or simply popular accounts under topical themes. For instance, a user can create a list called "*Music and Bands*," and add accounts such as @YahooMusic, @radiohead, or @katyperry. Given a target user $u \in U$, we obtain the set F_u of all the users he follows. The objective is to categorize each followed user $f \in F_u$ into some *topics*, using the lists in which f was (possibly) added by some other Twitter user. To this end, for each user $f \in F_u$ we gathered all lists containing f: we refer to this set as L_f. Indeed, users'

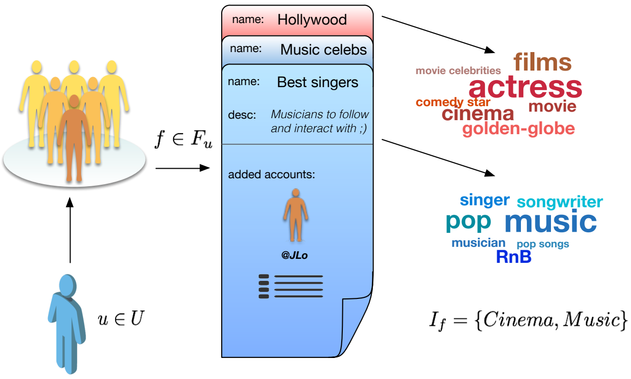


Fig. 1: First step of the process of interest inference for a user u. For each followed user f we gather all the lists containing f and look for top occurring topics (right). In this case the followed user *Jennifer Lopez* (@JLo) will contribute for a positive interest in *Music* and *Cinema*.

list memberships do not necessarily reflect topical interests [31]—some lists relate to personal feelings or beliefs towards the followed users (e.g., "great people," "interesting twitter users") or how they relate with them (e.g., "my family," "colleagues"); for this reason we consider users f who belong to at least 10 lists. We consider as topic all unigrams and bigrams, composed by only nouns and adjectives (as recognized by a standard part-of-speech tagger), found in all the descriptions and names of each list $l \in L_f$, rejecting all topics that do not appear in at least 10 lists. Furthermore, we keep only the top 100 most frequent topics for user f, and we manually classify them in the categories set I. We then associate to user f a set of categories $I_f \subseteq I$, such that for each $i \in I_f$ there is at least one topic classified in the corresponding category i. We, therefore, informally consider user f as *expert* (or *authority*) in each category $i \in I_f$. Figure 1 provides an example of this entire process: a user u has in his list of followers F_u (depicted in yellow/orange) the user *Jennifer Lopez* (@JLo); @JLo appears in a set of lists (three of them reported in the figure: "Hollywood," "Music celebs," and "Best singers"); from the names and descriptions of these lists we extract a set of unigrams and bigrams (topics); these topics are classified in two categories, that are Music (red font) and Cinema (blue font); @JLo is then considered an authority in the Cinema and Music categories.

We now need a way to derive from this information the actual interest degree of user u in the various categories. We make the assumption that the more experts that user u follows on a certain category the more he is likely to be interested in that category. We denote as $E_u^i = \{f \in F_u : i \in I_f\}$ the set of users followed by u who are expert in category i. Finally, we associate with the original user u an $|I|$-dimensional vector $interest_u$ of interest scores, one for each considered interest category. The score of each user relative to a specific category j will be the fraction of experts on category j he follows, normalized over all followed experts:

$$\mathit{interest}_u[i] = \frac{|E_u^i|}{\sum_{j \in I} |E_u^j|}.$$

We have also considered unnormalized interest score vectors, but we eventually used normalized score vectors for the reasons explained in Section 4.1.

The manual tagging methodology can be a bottleneck in the interest inference process, for cities having a Twitter user base greater than the one in our dataset. However, because the list-based interest extraction is based on classifying unigrams and bigrams, and the list of interest category is fixed, to overcome this limitation in case of very large scale datasets, we believe that this step could be implemented using a classification engine based on ontologies (like Freebase or DBPedia) or lexicon databases like Wordnet [28].

3.3.2 Top-k Zone Ranking

In this phase we compute a ranking over the considered zones for each category so as to select the most promising locations for advertising. Intuitively, our approach is to use the user traces and project the amount of users' interests towards the different topics on the various city zones, thus exploiting the power of this collective signal to drive our ranking. Let $U_z = \{u : z \in Z_u\}$ identify the set of users u who have passed through zone z and *Freq*(u,z) be the number of geotagged tweets issued by user u in zone z.

Given an interest category i, a zone z, and the set U_z of the users who have the zone in their traces, to evaluate the *targeted effectiveness* of the city zone z for category i we consider four different scoring functions:
All: sum of $\mathit{interest}_u[i]$ scores, of *all* users $u \in U_z$

$$\mathcal{E}_{\mathrm{A}}(z,i) = \sum_{u \in U_z} \mathit{interest}_u[i].$$

Primary: sum of $\mathit{interest}_u[i]$ scores, considering only users u for whom i is the category of *primary* interest, that is, the set of users $\overline{U}_z \subseteq U_z$, $\overline{U}_z = \{u \in U_z : \mathit{interest}_u[i] > \mathit{interest}_u[j], \forall j \neq i\}$

$$\mathcal{E}_{\mathrm{P}}(z,i) = \sum_{u \in \overline{U}_z} \mathit{interest}_u[i].$$

AllFreq: sum of the product of $\mathit{interest}_u[i]$ scores and the number of geotagged tweets by each user u in zone z

$$\mathcal{E}_{\mathrm{AF}}(z,i) = \sum_{u \in U_z} \mathit{Freq}(u,z) \cdot \mathit{interest}_u[i].$$

PrimaryFreq: like *AllFreq*, but considering only the users for which i is the category of *primary* interest

$$\mathcal{E}_{\mathrm{PF}}(z,i) = \sum_{u \in \overline{U}_z} \mathit{Freq}(u,z) \cdot \mathit{interest}_u[i].$$

The output of the algorithms is a ranked list of k zones Z_i^*, for each category $i \in I$, containing the top-k zones according to the ranking provided by the function $\mathcal{E}(z,i)$. For each algorithm and for each category $i \in I$, we then output a ranked list of the top-k zones Z_i^*.

Polarization. In the task of evaluating the score of a zone for a category, all the previous measures work explicitly over only the set of users interested in that category: the optimization problem for a category is treated as independent of the other categories. Nevertheless, our goal of finding a ranking of the best zones could potentially benefit from a criterion that also considers how the interests *overlap*, decreasing the score of zones frequented by users with many different interests. In other words, it can be interesting to consider the *polarization* of zone z towards category i, defined as the relative frequency of users interested in category i that the zone z could attract in the future, given the observed data. We estimate it with a Beta distribution, a continuous distribution that is widely used for modeling uncertainty on processes with binary outcomes [11]. It is very flexible for modeling proportions as its density can have quite different shapes depending on the values of the two parameters α and β, and it is easy to update when new information is provided.

The process with binary outcomes in this setting is given by the fact that a user u (who has zone z in his trace) can be either interested (i.e., true) or not (i.e., false) in category i.

Specifically, for each zone z and for each category i, given the set of users U_z who have zone z in their traces, the unknown relative frequency of users interested in category i that the zone z could attract in the future has a probability distribution expressed by a Beta function with parameters α_z^i and β_z^i, where α_z^i is the number of users $u \in U_z$ interested in category i (plus one, to be in accordance with the typical use of the beta distribution) and β_z^i is the number of users $u \in U_z$ not interested in category i (plus one); that is:

$$\begin{aligned}
\alpha_z^i &= |\{u \in U_z : \mathit{interest}_u[i] > 0\}| + 1, \\
\beta_z^i &= |\{u \in U_z : \mathit{interest}_u[j] > 0 \text{ for } j \neq i\}| + 1.
\end{aligned}$$

The expected value of the Beta distribution $\mathbb{E}\big(\mathrm{Beta}(\alpha_z^i,\beta_z^i)\big)$ can be interpreted as the expected value of the relative frequency of users interested in category i that the zone z could attract in the future. The standard error (SE) of the Beta distribution, $\mathrm{SE}\big(\mathrm{Beta}(\alpha_z^i,\beta_z^i)\big)$, is an estimate of the standard deviation of the expected value. This value is important to indicate the reliability of an estimation. Intuitively, the more representative is the subset of users, the lower the SE and the more accurate the estimation of the expectation.

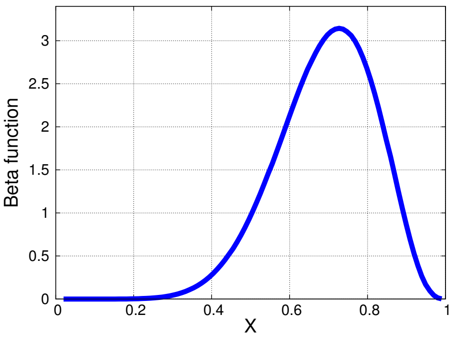


Fig. 2: Beta function after 8 users interested in category i and 3 users not.

Figure 2 shows the Beta function for a scenario where 11 users have zone z in their trace. Among them 8 are interested in category i and 3 are not interested. This curve expresses the probability that zone z could attract users interested in category i in the future. The expectation of the distribution is 0.7. For instance, the system expects that 7 of the next 10 users who transit in z will be interested in category i. The SE of the distribution is roughly 0.04. Using Chebyshev's inequality [13], the SE can be interpreted as saying that the system expects that in the next 100 users who transit in z, between 62 and 78 will be interested in category i, with probability bigger than 0.75.

To derive a polarization score for each zone $z \in Z$ and for each category $i \in I$ we adopt a worst case estimation approach by subtracting a quantity $\mu \cdot \mathrm{SE}$ from the expectation, as follows:

$$\mathcal{P}(z,i) = \mathbb{E}\big(\mathrm{Beta}(\alpha_z^i,\beta_z^i)\big) - \mu \cdot \mathrm{SE}\big(\mathrm{Beta}(\alpha_z^i,\beta_z^i)\big) .$$

The aim of the worst-case estimation approach is to prevent that an inaccurate estimation will corrupt the model. In fact, the expectation of zones where few users transit can be really high just by chance. However, such zones also have a relatively high SE. Therefore we subtract $\mu \cdot \mathrm{SE}$ from the expectation to penalize the polarization score of zones for which we do not have enough information. A similar approach has been proposed in [21]. In our experiments we set the value of μ equal to 4.

Having defined the polarization score $\mathcal{P}(z,i)$ we can consider variants of all the algorithms described so far, in a way that explicitly considers the polarization effect, as follows:

$$\hat{\mathcal{E}}_{alg}(z,i) = \mathcal{E}_{alg}(z,i) \cdot \mathcal{P}(z,i),$$

and will add the suffix *Pol* to their names (i.e., *PrimaryPol*, *AllPol*). An important aspect behind enforcing polarization is that it may allow for a lower overlap of the zone rankings for different interest categories: this promotes zone specificity and alleviates the phenomenon of *overcluttering*, the presence of many co-occurring ads. Such overload may lead to a loss of attention and therefore of effectiveness of the ad [12, 20].

4 Experimental Evaluation

In this section we describe the datasets used for the evaluation of the proposed solution as well as other collateral analyses we performed on them. Then we discuss our results and some interesting properties of the zones found.

4.1 Experimental Setup

Datasets. Our main dataset is a collection of geotagged tweets gathered from the Twitter Firehose for the two-month period of November and December 2013, obtained specifying as boundary region (or *bounding-box*) the city of Milan, Italy and its suburbs. It consists of a total of 477,913 tweets by 31,356 users. By restricting the set of users to those having at least 10 tweets, we end up with 404,077 tweets and 5,086 total users. The city zones $z \in Z$ have the shape of square cells, each of $235m^2$, for a total of 10,000 cells. For each zone we also possess mobile-telecommunication usage data spanning the same two-month observation period for each cell. The telecommunication data consist of the normalized level of interaction of the users with the mobile phone network, considering inbound and outbound calls, SMS messages, and Internet usage.[1]
Evaluation Methodology. The ideal evaluation of our methodology would require the experimentation with a real advertising campaign, which would promote some product (preferably more than one) for each category and place ads in cities based on different strategies. Although such an approach falls within the scope of our work, it is beyond our capabilities. Therefore, we now describe an alternative way to measure how we can target zones suitable for ads of a given category.

We consider a set of nine different advertising categories, namely, $I =$ {*Food, Art–Photography, Shopping–Fashion, Music, Cinema–TV, Technology, Home-Design, Sport, Motors*}. To evaluate the rankings computed by our techniques we identify, for each category $i \in I$, the set $U_i \subset U$ of representative users as the users whose corresponding interest score for interest category i is greater than 0 (that is, users interested in that topic).

To recall, the interest score of a user towards a category is determined by the normalized fraction of the number of users that he follows who are found to be representative for that category, following the procedure detailed in Section 3.3.1. The choice of normalization is not obvious, and was made for the following reason: the distribution of the number of users followed by people in our dataset is uneven: there is a majority of users who follow just a handful of other people, a small group of users (around 5%) who follow a very large number of users, and an almost continuous variation in between. We therefore opted for normalization to keep the interest scores comparable. Furthermore, on a tentative use of an unnormalized interest scores, we noticed

[1] All the datasets used for this work are available at `https://dandelion.eu/datamine/open-big-data/`, released by Telecom Italia, the main Italian telecommunication provider, for the international competition *Telecom Big Data Challenge 2014*.

that some zones scores were boosted by just one of such "power" Twitter users, overall leading to low performance. As a measure of sample representativeness we introduce a cutoff parameter ζ: we filter out of the study all users whose trace contains fewer than ζ zones, and we filter as well all zones that appear in fewer than ζ user traces (the impact of this parameter on the sample size of the various categories can be observed in Figure 4).

We evaluate our approach performing a 5-*fold cross validation*. To this end, for each category $i \in I$, we divide U_i in 5 folds of equal size and we use in turn one of these folds as test set and the remaining as training set. The reported results are the average of 5 independent runs, one for each possible fold chosen as test set.

The information about what zones are selected by advertisers in practice is not publicly available. Therefore, we consider two baselines with which we compare our approaches.

The first baseline is a strategy that selects the most crowded zones (e.g., train stations, busy streets, main squares), as one of the more natural approaches, which is known to be used in practice [34]. To estimate the zones' crowdedness we leverage the data in our possession about mobile telecommunications activity in the zones. In particular, we compute the average daily activity per zone. Note that this baseline is aligned with a standard measure for ad effectiveness [23], called *daily effective circulation*, developed by the Traffic Audit Bureau for Media Measurement and reported as the estimated number of people who have the opportunity to see a billboard in one day. Not surprisingly, we observed that the zones with the highest telecommunication activity are indeed the most crowded ones: the central station and the main square of Milan. We refer to this baseline approach as *Telco*.

The second baseline is also natural: it uses the Twitter user data, where the score associated to each zone is the total number of tweets in that zone. Intuitively this interest-agnostic baseline shows what part of the proposed algorithms' performance is due alone to the crowdedness of zones and not to the user interests. We refer to this baseline approach as *AllTweets*. We will therefore test the performance of all our algorithms described in section 3.3.2 (*Primary*, *All*, *PrimaryFreq*, *AllFreq*, *PrimaryPol*, *AllPol*) and the baselines *Telco* and *AllTweets* on the 5-fold cross validated test sets, using the metrics described in the following section.

Metrics. As we mentioned previously, we consider the trace of a user as a proxy for his movements. It is the set $Z_u \subseteq Z$ containing all the zones where the user has issued at least one tweet.

To evaluate and compare the performance of all algorithms and the baselines we evaluate the corresponding ranked list Z_i^* of top-k zones (candidates for targeted advertising) computed by each algorithm for each advertising category i, using the following metrics, defined over the set of users U_i in the test portions:

Coverage: the fraction of users in the test set who passed in at least one of the *top-k* zones in the solution:

$$\mathrm{Cov} = \frac{1}{|U_i|} \sum_{u \in U_i} x_{u,i},$$

where $x_{u,i}$ is 1 if $Z_u \cap Z_i^* \neq \emptyset$, and 0 otherwise.

Precision: the fraction of *top-k* zones in the solution where a user passed, averaged over all the users in the test set:

$$P@k = \frac{1}{|U_i|} \sum_{u \in U_i} \frac{|Z_u \cap Z_i^*|}{|Z_i^*|}.$$

Recall: the fraction of zones in the user trace that are also in the solution, averaged over all the users in the test set:

$$Rec = \frac{1}{|U_i|} \sum_{u \in U_i} \frac{|Z_u \cap Z_i^*|}{|Z_u|}.$$

Mean Average Precision (MAP): the average precision of the ranking for a given ranking size:

$$\mathrm{MAP} = \frac{\sum_{i=1}^{k} P@i}{k}.$$

A high coverage indicates that the selected zones are effective spots to place an advertisement because a high number of interested users can be potentially reached. The precision and recall measures refine this estimation by considering the actual number of top-k zones where the representative users passed. The MAP metric additionally reflects the quality of the ranking in the top-k zones: the passage along a highly ranked zone is accounted with a higher weight with respect to a passage in a lowly ranked zone.

Note that these measures take into consideration only the number of distinct users and not the frequency of their visits. We decided to not use frequency-based measures because, despite the effort of removing the home location for a user (see next section), the presence of work and close-to-home locations would induce a high bias.

Moreover, we also make use of a metric to compute the similarity (i.e., the overlap) between pairs of solutions Z_i^* and Z_j^* where i and j are different ad categories:

Jaccard similarity index the ratio of size of the intersection to the size of the union of two solutions:

$$Jac(i,j) = \frac{\left|Z_i^* \cap Z_j^*\right|}{\left|Z_i^* \cup Z_j^*\right|}.$$

4.2 Evaluation Results

We start by describing some analyses that we performed on our dataset, and we continue describing the results of our evaluation.

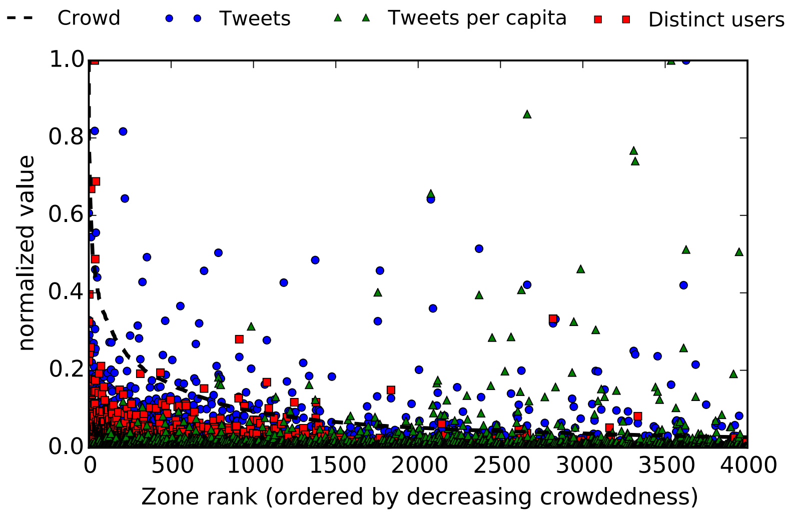


Fig. 3: Analysis of *tweetiness* versus *crowdedness* of the different zones. The x-axis represents the rank of the zones given by the *Telco* algorithm. On the y-axis, are instead reported the normalized values of some tweet-related measures (i.e., the value of the activity divided by its maximum value, as found in our dataset): number of tweets, tweets per capita, and number of distinct users present in our dataset for each zone. Notice, that whereas the zones are ordered into decreasing order of crowdedness, the other three measures are not decreasing.

4.2.1 Dataset Analysis

In this section we describe some analyses of the datasets and on the outcome of the profiling process (Section 3.3.1). More specifically, Figure 3 correlates the ranking of the first 4,000 zones obtained using the telecommunication activity (as a proxy for determining the crowdedness of the zones), corresponding to the outcome of the *Telco* baseline, to some measures related to the geotagged tweets issued in the corresponding areas (i.e., the normalized number of tweets, tweets per capita, and number of distinct users). We notice that, despite the expected general tendency of the Twitter-related measures to conform to the direction of decreasing crowdedness, there are a number of zones that are ranked low by crowdedness but indeed give rise to a very high number of tweets and distinct users: this motivates the idea that seeking for solutions that do not rely only on crowd estimates might be successful for the advertising application (for which some less crowded zones may be more economical).

We deepen our investigation into the outcome of the profiling process by measuring the distribution of user interests.

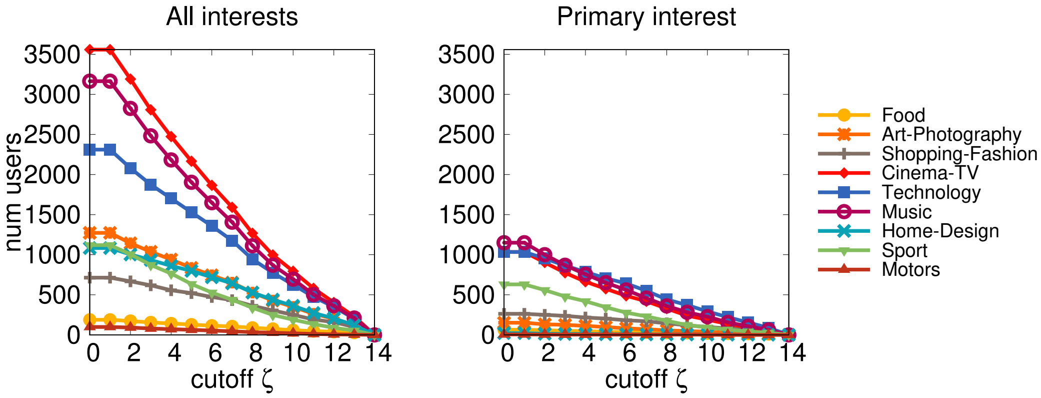


Fig. 4: Number of users in the set U_i for each category $i \in I$ versus the cutoff parameter ζ.

Figure 4 shows the number of users per category with respect to the cutoff parameter ζ, when considering *all interests* (left) or just the *primary interest* (right). The latter shows how many users exhibit their main interest in a given category $i \in I$ (i.e., the maximum score in the user interest vector is associated to category i).

The left plot shows that the most popular categories for the considered dataset are *Cinema–TV*, *Music*, and *Technology*, followed by *Art–Photography*, *Home-Design*, and *Sports*. Instead, if we consider only the primary interest of the users (right), the categories *Cinema–TV*, *Music*, *Technology* and *Sports* have still a relatively high number of users, whereas the number for *Art–Photography* and *Home-Design* drops. A possible explanation for this is that these latter categories are less likely to be the primary interest for the users since they compete with more common categories (such as *Cinema–TV*, *Music*, *Technology*, and *Sports*), in general and specifically on Twitter (where the most popular accounts belong to musicians, players or actors[2]). For the rest of our experiments we use $\zeta = 4$.

We now inspect the user-interest vectors. In particular, we study how interests in categories correlate with each other. Figure 5 shows the correlation matrix (computed using Pearson's r correlation) of the users interest vectors, for all categories. For instance, we observe that the *Cinema–TV* and *Music* categories are correlated, as well as the *Home* and *Art–Photography*, or the *Motors* and *Sport*. However, overall, the interests appear quite independent, that is, most of the correlations are close to zero (lightest color in the matrix).

[2] Source: Twittercounter `http://twittercounter.com/pages/100`

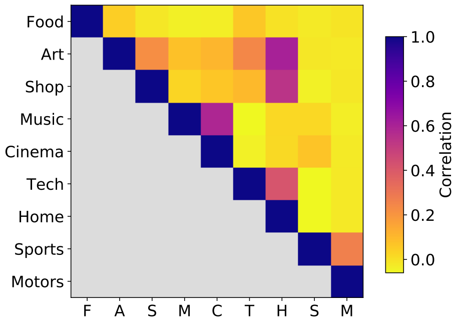
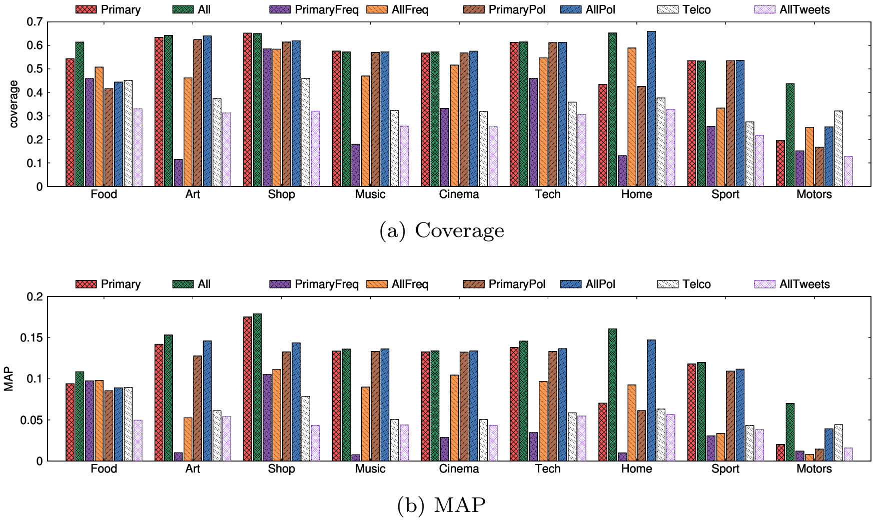


Fig. 5: Interest vectors correlation matrix. Since the matrix is symmetric, only the upper triangle is shown.



(a) Coverage

(b) MAP

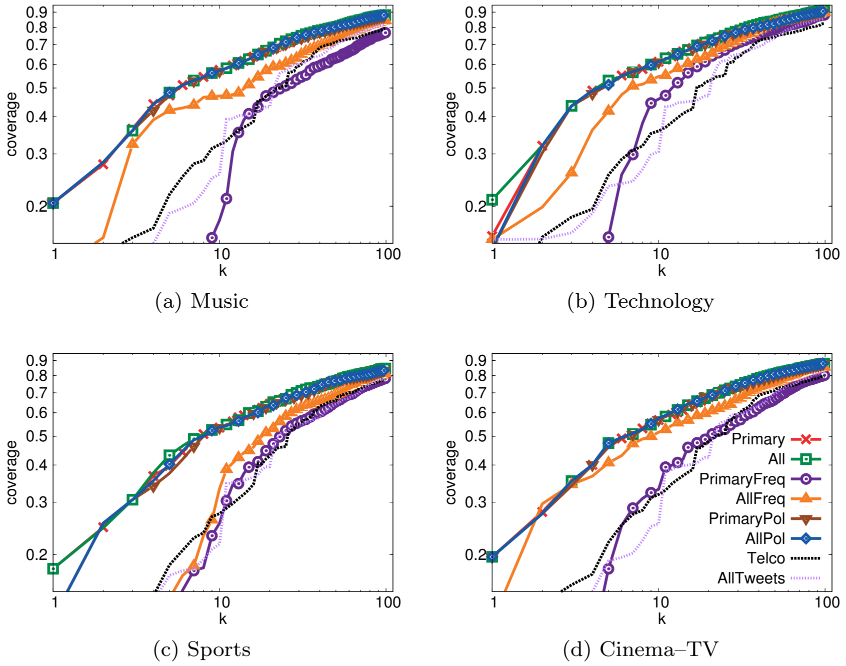
Fig. 6: Coverage and MAP with $k = 10$.

4.2.2 Performance

Now we present the results of our experimental study for the unweighted problem setting. Figure 6a depicts the coverage values of the solutions for all the categories, for a fixed budget $k = 10$.

Figure 7 compares coverage with varying k (in log–log scale), achieved by the various algorithms for the categories with more users (i.e., *Cinema–TV*, *Music*, *Technology*, and *Sports*).

All algorithms that ignore the *frequency* of the tweeting activity (i.e., *Primary*, *All*, *PrimaryPol* and *AllPol*) achieve a high coverage, outperforming the baseline solutions *Telco* and *AllTweets* by a consistent margin in all the considered categories. We omit the polarized versions of the frequency-based algo-



(a) Music

(b) Technology

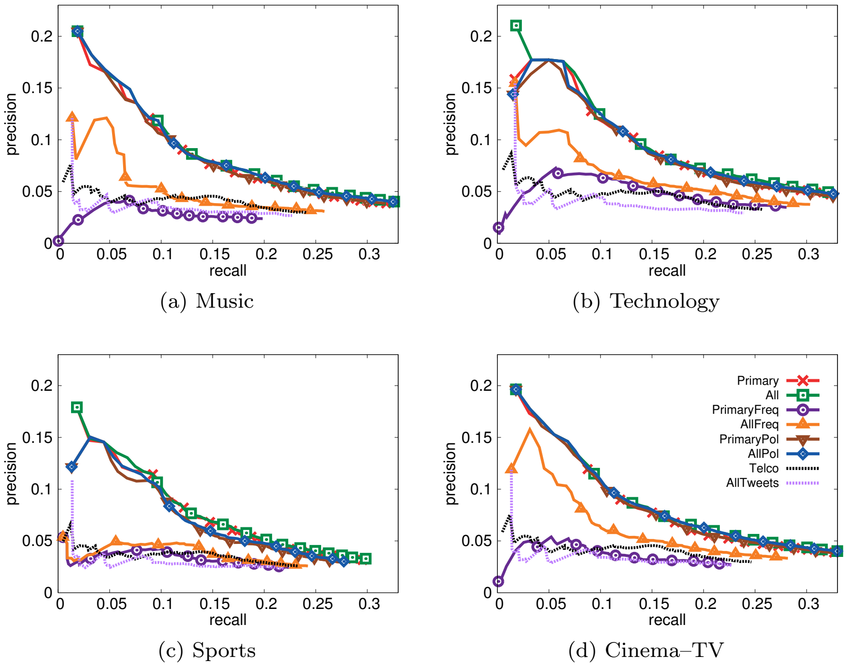
(c) Sports

(d) Cinema–TV

Fig. 7: Coverage varying k, log–log scale.

rithms because their performance is essentially identical to their non-polarized versions. Algorithms *PrimaryFreq* and *AllFreq*, instead, perform poorly. A possible explanation for this behavior is that these algorithms boost the ranking of users' *everyday zones* (e.g., home, work, gym, favorite bars), which are often unique to the specific user and may not be a good indicator for a global perspective. To mitigate this issue, we filtered the *home* location for each user (which we consider to be the zone where he tweets the most). This does not completely solve the problem. Moreover, algorithms based on tweet frequency are affected by a common bias underlying the use of geotagged tweets as a proxy for the movements: some spots are more suitable than others for tweeting activity (e.g., bars, parks, rest places). By taking into account frequency, these zones are even more privileged in the ranking. We notice that the second baseline *AllTweets*, behaves in a way that is in the middle between the other baseline *Telco* and the frequency-based algorithms. This is expected, as it targets areas where highly active Twitter users are present.

The coverage of *Telco* reaches its peak for the Shopping–Fashion category. We believe that this result is related to the fact that the main shopping zones often correspond to the main areas of a city, usually the most crowded. In our case study, the city of Milan, this is indeed true.



(a) Music

(b) Technology

(c) Sports

(d) Cinema–TV

Fig. 8: Recall vs Precision.

Next we study the performance of the algorithms in terms of average precision and recall. In particular, we run them with different values of k and we collect the precision–recall value pair achieved in each run. In Figure 8 we report the curves obtained by plotting all the precision–recall value pairs collected for $k = 1, \ldots, 100$. Both average precision and average recall values are generally low (yet up to 40 times better than the baseline). This is expected: although with our approach we can hope to obtain interest-related zones that work at a user aggregated level, when we sum the contributions of many users, we cannot assume that a high portion of zones in the single user traces reflect personal interests—simply, users that may be interested in cinema may have not visited the top *Cinema–TV* areas of our dataset, in the two-month period of our observations. Put more simply, our setting differs from the classical information-retrieval scenario: even a perfect algorithm cannot achieve a precision or a recall score equal to 1; it will be much lower. Overall, the relative performance of the algorithms agrees with the one on the coverage metric. However, analyzing the quality of the ranking, we observe a difference in the performance of the frequency-agnostic algorithms. Figure 6b depicts the MAP score achieved by the various algorithms on all the categories. We can observe that the algorithms that use all the user interests (i.e., *All* and *AllPol*) perform slightly (but consistently) better than the those considering only primary interests (i.e., *Primary* and *PrimaryPol*). Intuitively, on the one hand, when a user is represented only by his primary interest, the algorithms may lose some important information, thus lowering the overall performance. On the other hand, taking into account all the user interests may increase the overlap between the rankings pertaining to different categories, because, for a given user, the same user trace is used as representative for all the categories associated with his interests. For this reason the *Primary* algorithm may be considered a valid alternative when the specificity of the zones is of concern.

We investigate more this aspect in our next analysis, studying to what extent the zones in the top-k solutions (Z_i^*) overlap between different categories. We compute the Jaccard similarity index between pairs of top-k solutions (i.e. Z_i^* and Z_j^*, for $i, j \in I$). Figure 9 shows the Jaccard similarity index of frequency-agnostic algorithms with a heatmap, for $k = 10$. The results confirm our intuition: *Primary* is able to differentiate better the solutions among different categories (i.e., low similarity among them) compared to *All*. However, the overlap among the rankings can be further reduced by considering the polarization of the zones (i.e., *AllPol* and *PrimaryPol* algorithms), by boosting the rank of those zones where the interest in that specific category is more significant with respect to the others. In particular, *PrimaryPol* provides highly differentiated solutions per category, where a slight overlap is present only for those categories that are a priori correlated (see Figure 5), such as *Cinema–TV* and *Music*. We also include the *Telco* solution, to assess how much the rankings found by our algorithms differ with respect to the baseline (untargeted) solution that ranks the zones by crowdedness.

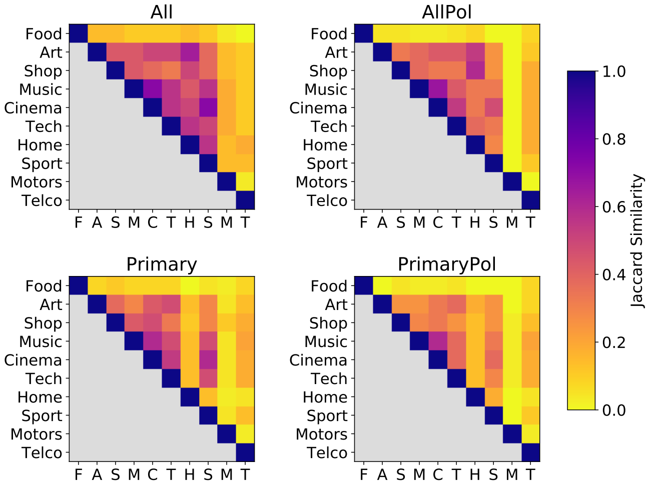
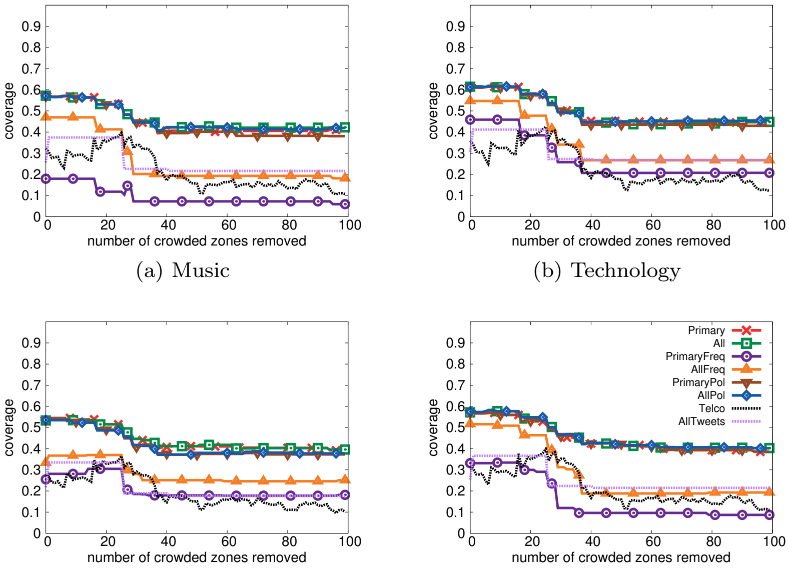


Fig. 9: Jaccard similarity matrices for the rankings of the different algorithms with $k = 10$.



(a) Music (b) Technology

(c) Sports (d) Cinema–TV

Fig. 10: Resilience in coverage of the top-10 solutions to the removal of top crowded zones.

In our next experiment we deepen in the comparison with the *Telco* baseline. In particular, we are interested in assessing whether our algorithms are able to identify unexpected zones that are not among the most crowded cells, but nonetheless are frequented by people with a specific interest. Positioning an ad in such zones may represent an advantage in capturing user attention, as crowded areas are typically overloaded with ads. In addition, they are in general more economical. To investigate the resilience of our rankings to the removal of crowded areas from the potential candidates, we remove for all categories, an increasing number x of generally crowded cells, the top-x zones provided by algorithm *Telco*. The objective is to try to understand what is the loss in coverage that we incur if we cannot consider as candidates the most crowded cells (because of a supposedly high cost, because of their ads overload, etc.). Although it might seem counterintuitive, we also perform this filtering for the two baselines; for *Telco*, this corresponds to shifting the selection of top-k solutions by x positions (i.e., ignoring the top-x ranked zones). We present the results in Figure 10 for $k = 10$, by plotting the coverage against the number x of solutions removed.

The figure shows that the coverage for all algorithms is initially stable, it slightly decreases when x is in the interval [10,40], and then it stabilizes again. The coverage for *Telco* shows a surprising initial growth, as we remove the first cells of the ranking (which on the contrary, are expected to be the more effective), reaching a global peak around the value $x = 25$, and dropping thereafter. This peak corresponds to the inclusion of a few cells where the number of distinct Twitter users in our dataset is higher. When these cells are discarded, and therefore can't be part of the solution, the coverage of *Telco* decreases abruptly: in fact, as shown and discussed in Figure 3, the number of tweets and distinct users is very peaky and not monotonically decreasing in the direction of decreasing crowdedness. This is a sign that the top crowded zones may be less effective to cover users in the test set. We can notice a similar effect for the *AllTweets* baseline, although it applies only to the first two cells.

To conclude, our algorithms appear to be overall resilient to the removal of crowded cells, experiencing an average loss in coverage of just about 25% when the top-100 crowded zones are ignored (as a reference, the area of the whole city center corresponds to about 200 cells). This result is an indicator that they are indeed able to identify nontrivial zones.

4.2.3 Anecdotal Results

To have more insights on the proposed approach we conclude the unweighted setting by performing some qualitative analysis. Figure 11 (center) shows the actual top-10 zones identified by the Algorithm *All* for each considered category in the city of Milan. We can see that such zones do not fall exclusively in the city center, but they span the entire considered area. To help the qualitative assessment of the relevance of the solutions for the specific category, we report some points of interest (POIs) found in some zones. The identified POIs are actually highly related to the corresponding category, as, for instance, the *Triennale* exhibition[3], for the *Art–Photography* category, or the Computer Engineering building of the Politecnico di Milano for the Technology category. By manual inspection, we also found other less obvious, yet relevant places, highlighted with the crosshatch, with corresponding pictures on the left and right sides of Figure 11.

4.3 Cost of Advertising

As we discuss in Section 3.2, our main emphasis is on the simpler, unweighted case. In this section, we delve into the specifics of a particular medium of advertising: placing banners in public spaces of the city of Milan. Using public information from the open data website of Comune di Milano[4] we have obtained the prices for placing banners for the two-month period of our study. For highly populated municipalities like Milan, the costs of advertising varies on the basis of the specific street, along two categories: *regular* for most streets, and *special* for some highly trafficked streets. Indicatively, the cost for placing

[3] http://en.wikipedia.org/wiki/Triennale

[4] `http://www.comune.milano.it/wps/portal/ist/it/servizi/tributi/pubblicita/pubblicita_impostapubblicita`

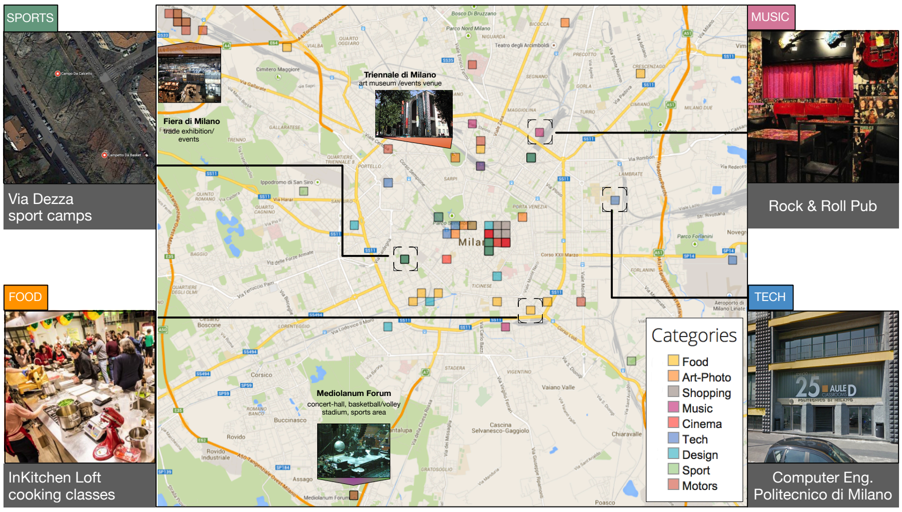


Fig. 11: Center: Actual top-10 zones identified by the *All* algorithm for each considered category in the city of Milan. Some POIs are also displayed. Sides: pictures of some less central venues found in the zones (crosshatch). Left: Sports: *Via Dezza* (public sport camps) and Food: *InKitchen Loft* (cooking master classes school). Right: Music: *Rock&Roll* (live music pub); Technology: Computer Engineering building (Politecnico di Milano)

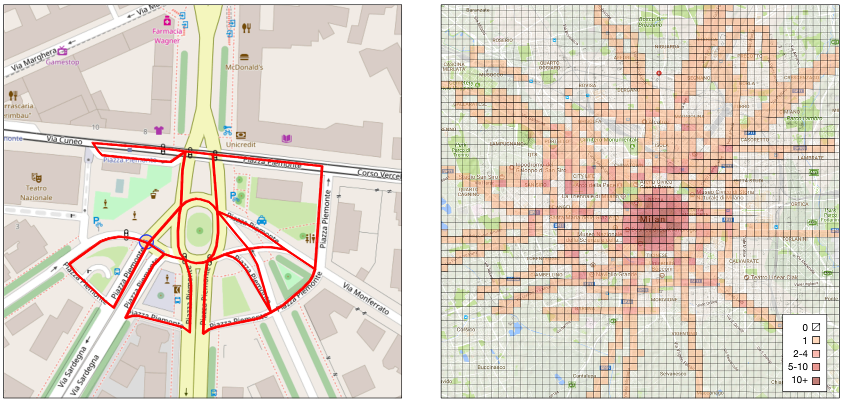
a banner of one square meter placed for the two-month period of our study in a regular zone was 93 EUR, and in a special was 235 EUR. We obtained the special-street coordinates polylines using Overpass Turbo[5] a mining tool that exposes an API (and a query language) to obtain geographic data from OpenStreetMap[6]. Figure 12a shows an example of the polylines extracted for *Piazza Piemonte*, a square that is part of the special streets list.

Because we work with the granularity of city zones, we then identified the intersections of special streets with our city zones (cells): Figure 12b shows a portion of the Milan map where the color of each cell is proportional to the number of special streets it intersects with (the transparent cells have 0 special streets, whereas the darker cells have more than 10 special streets). After having identified the streets that intersect with each zone z we declare a zone z as special (and we set $c(z) = 235$) if it intersects with at least one special street, otherwise we declare it as regular (and we set $c(z) = 93$). (We have tried also other approaches for pricing city zones, but the findings are qualitatively similar.)

Next, we consider the corresponding knapsack instance where the items are the city zones $z \in Z$, their value $\mathcal{E}(z,i)$ for a given interest category i is

[5] `http://overpass-turbo.eu`

[6] `https://www.openstreetmap.org`



(a) Extracted polyline for *Piazza Piemonte*.

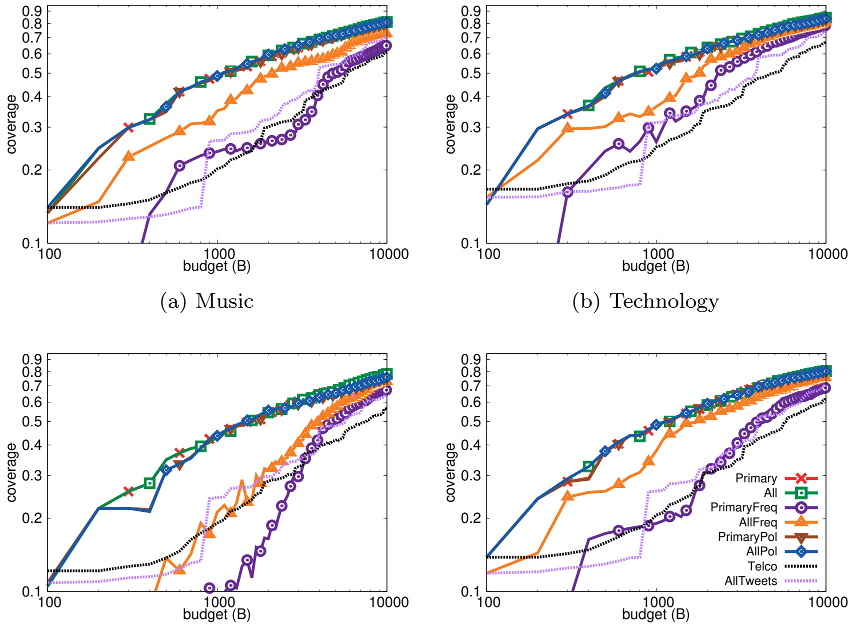
(b) Heatmap showing the number of special streets contained in each zone, for a portion of the map of Milan.

Fig. 12: Steps of the cost extraction process: polyline extraction and zone-polyline intersection.

estimated by our algorithms (and the baselines) and their cost $c(z)$ is the one described in the previous paragraph.

To evaluate our algorithms we solve multiple instances of the problem considering budgets ranging from 100 (allowing to take 1 normal cell in the solution) to 10,000 (allowing for about 100 normal cells or about 40 special cells), considering the two extremes for the sake of comparison. Because the size of our problem is small (max n=10,000 cells, max B=10,000) we implemented the solver using the well-known dynamic programming approach for the knapsack problem, ensuring an $O(n \cdot W)$ time and space costs.

Figure 13 depicts the coverage with varying B (in a log–log scale), achieved by the various algorithms for the categories *Cinema*, *Music*, *Technology*, and *Sports*. The coverage results are quite similar (although overall expectedly slightly lower because of the budget scale) compared to the unweighted case. All our algorithms (except for the ones based on frequency) consistently outperform the baselines (especially Telco, as expected) with a higher margin with respect to the unweighted case, also for increasing values of B. The considerations made in the unweighted case for all the algorithms stay the same. Note also that, differently from the unweighted setting, here the cardinality of the solution is not fixed. This is the reason behind the non-monotone behavior of the curves in Figure 13: for example, a run of the algorithm with a higher budget despite having a higher estimated value of the solution, can have a lower cardinality (i.e., can include a smaller number of cells) compared to an execution with a lower budget, thus potentially leading to a lower coverage of the test set. The results for the other metrics are very similar, and, therefore, we omit them here.



(a) Music

(b) Technology

(c) Sports

(d) Cinema

Fig. 13: Coverage varying the budget B.

5 Discussion and Conclusion

To our knowledge this is the first work leveraging the use of publicly available social-media data for the purpose of targeted outdoor advertising. We believe, and hope, that work towards this direction can have a significant effect, especially for small- or medium-size companies or nonprofit organizations, where usually the available budget for advertising is limited. Our technique shows that a targeted selection of ad placement can result in a better use of resources by reaching users interested (as measured by Twitter activity) into topics related to the advertising.

Of course, our work is only a first step in this direction and there are several limitations. First, note that we use social media as a proxy of what areas users with a particular interest frequent. Even though such assumptions have been done in almost all the previous works on event detection or city characterization that are based on social media, nevertheless they introduce biases because of the specific social medium used: in our case, considering only signals relative to Twitter-affine users. For instance there may be bias towards bars, or restaurants where users sit and tweet. Even though we provided substantial evidence for the effectiveness of our approach, ultimately one needs to perform real in-field experiments. Such experiments are definitely nontrivial and require the collaboration of real vendors. Yet, a simple experiment that we are considering is putting hashtags on the advertising signs and measure the possible information spreads. This could reveal insights on how the gap between offline and online media is crossed. Other extensions we are thinking take into account the viral-marketing effects; for example, taking into consideration the centrality of the users that frequent a particular location. Yet, another important extension would be to take into account the well known fact in marketing [20] that repeated exposures increase advertising recall. This means that we could study trajectories of users and select spots for ad placement that are common across various trajectories (a trivial example would be along Broadway street in NY), leading to some hard and interesting combinatorial problems. Finally, our approach can have many other applications, such as tourism [17].

Acknowledgements We would like to thank the anonymous reviewers for their valuable comments.

References

1. Anagnostopoulos, A., Petroni, F., Sorella, M.: Targeted interest-driven advertising in cities using Twitter. In: Proceedings of the 10th International AAAI Conference on Web and Social Media (ICWSM), pp. 527–530 (2016)
2. Belch, G.E., Belch, M.A.: Advertising and Promotion: An Integrated Marketing Communications Perspective. McGraw-Hill (2011)
3. Bhattacharya, P., Zafar, M.B., Ganguly, N., Ghosh, S., Gummadi, K.P.: Inferring user interests in the Twitter social network. In: Proceedings of the 8th ACM Conference on Recommender systems (RecSys), pp. 357–360. ACM (2014)
4. Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.: Short and tweet: experiments on recommending content from information streams. In: Proceedings of the 28th ACM Conference on Human Factors in Computing Systems (SIGCHI), pp. 1185–1194. ACM (2010)
5. Cheng, Z., Caverlee, J., Barthwal, H., Bachani, V.: Who is the barbecue king of texas?: a geo-spatial approach to finding local experts on Twitter. In: Proceedings of the 37th international ACM Conference on Research & Development in Information Retrieval (SIGIR), pp. 335–344. ACM (2014)
6. Cranshaw, J., Schwartz, R., Hong, J.I., Sadeh, N.M.: The livehoods project: Utilizing social media to understand the dynamics of a city. In: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM), pp. 58–65 (2012)
7. Donthu, N., Cherian, J., Bhargava, M.: Factors influencing recall of outdoor advertising. Journal of Advertising Research **33**(3), 64–73 (1993)
8. Farahat, A., Bailey, M.C.: How effective is targeted advertising? In: Proceedings of the 21st International Conference on World Wide Web (WWW), pp. 111–120. ACM (2012)
9. Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N., Gummadi, K.: Cognos: crowdsourcing search for topic experts in microblogs. In: Proceedings of the 35th International ACM Conference on Research & development in Information Retrieval (SIGIR), pp. 575–590. ACM (2012)
10. Gulmez, M., Karaca, S., Kitapci, O.: The Effects Of Outdoor Advertisements On Consumers: A Case Study. Studies in Business and Economics **5**(2), 70–88 (2010)
11. Gupta, A.K., Nadarajah, S.: Handbook of beta distribution and its applications. CRC press (2004)
12. Ha, L., McCann, K.: An integrated model of advertising clutter in offline and online media. International Journal of Advertising **27**(4), 569–592 (2008)

13. Huber, P.J.: The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1, pp. 221–233 (1967)
14. Iyer, G., Soberman, D., Villas-Boas, J.M.: The targeting of advertising. Marketing Science **24**(3), 461–476 (2005)
15. Kisilevich, S., Krstajic, M., Keim, D., Andrienko, N., Andrienko, G.: Event-based analysis of people's activities and behavior using flickr and panoramio geotagged photo collections. In: Proceedings of the 14th IEEE International Conference on Information Visualization (IV), pp. 289–296. IEEE (2010)
16. Kumar, A.: Dimensionality of consumer beliefs toward billboard advertising. Journal of Marketing & Communication **8**(1), 22–26 (2012)
17. Le Falher, G., Gionis, A., Mathioudakis, M.: Where is the soho of rome? measures and algorithms for finding similar neighborhoods in cities. In: Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM), pp. 228–237 (2015)
18. Lee, R., Sumiya, K.: Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In: Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, pp. 1–10. ACM (2010)
19. Liebig, T., Stange, H., Hecker, D., May, M., Kórner, C., Hofmann, U.: A general pedestrian movement model for the evaluation of mixed indoor–outdoor poster campaigns. In: Proceedings of the 3rd International Conference on Applied Operation Research (ICAOR), pp. 289–300 (2011)
20. Osborne, A.C., Coleman, R.: Outdoor advertising recall: A comparison of newer technology and traditional billboards. Journal of Current Issues & Research in Advertising **30**(1), 13–30 (2008)
21. Petroni, F., Querzoni, L., Beraldi, R., Paolucci, M.: Lcbm: a fast and lightweight collaborative filtering algorithm for binary ratings **117**, 583–594 (2016)
22. Piórkowski, M.: Sampling urban mobility through on-line repositories of gps tracks. In: Proceedings of the 1st ACM International Workshop on Hot Topics of planet-scale Mobility Measurements, pp. 1–6. ACM (2009)
23. Quercia, D., Di Lorenzo, G., Calabrese, F., Ratti, C.: Mobile phones and outdoor advertising: Measurable advertising. IEEE Pervasive Computing **2**(10), 28–36 (2011)
24. Rozenshtein, P., Anagnostopoulos, A., Gionis, A., Tatti, N.: Event detection in activity networks. In: Proceedings of the 20th International ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1176–1185. ACM (2014)
25. Saravanou, A., Valkanas, G., Gunopulos, D., Andrienko, G.: Twitter floods when it rains: A case study of the uk floods in early 2014. In: Proceedings of the 24th ACM International Conference on World Wide Web (WWW), pp. 1233–1238. ACM (2015)
26. Shannon, R., Stabeler, M., Quigley, A., Nixon, P.: Profiling and targeting opportunities in pervasive advertising. In: Proceedings of the 1st Workshop on Pervasive Advertising (PerAd) (2009)
27. Shimp, T., Andrews, J.C.: Advertising promotion and other aspects of integrated marketing communications. Cengage Learning (2013)
28. Varga, A.: Exploiting domain knowledge for cross-domain text classification in heterogeneous data sources. Ph.D. thesis, University of Sheffield (2014)
29. Vieweg, S., Hughes, A.L., Starbird, K., Palen, L.: Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In: PProceedings of the 28th ACM Conference on Human Factors in Computing Systems (SIGCHI), pp. 1079–1088. ACM (2010)
30. Villatoro, D., Serna, J., Rodríguez, V., Torrent-Moreno, M.: The tweetbeat of the city: microblogging used for discovering behavioural patterns during the mwc2012. In: Citizen in Sensor Networks, pp. 43–56. Springer (2013)
31. Wagner, C., Liao, V., Pirolli, P., Nelson, L., Strohmaier, M.: It's not in their tweets: Modeling topical expertise of twitter users. In: Proceedings of the 4th IEEE International Conference on Privacy, Security, Risk and Trust (PASSAT), pp. 91–100. IEEE (2012)
32. Wang, X., Zhang, Y., Zhang, W., Lin, X.: Efficiently identify local frequent keyword co-occurrence patterns in geo-tagged Twitter stream. In: Proceedings of the 37th international ACM Conference on Research & Development in Information Retrieval (SIGIR), pp. 1215–1218. ACM (2014)

33. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM), pp. 261–270. ACM (2010)
34. Woodside, A.G.: Outdoor advertising as experiments. Journal of the Academy of Marketing Science **18**(3), 229–237 (1990)
35. Yamaguchi, Y., Amagasa, T., Kitagawa, H.: Tag-based user topic discovery using Twitter lists. In: Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 13–20. IEEE (2011)
36. Zheng, Y., Capra, L., Wolfson, O., Yang, H.: Urban computing: concepts, methodologies, and applications. ACM Transactions on Intelligent Systems and Technology (TIST) **5**(3), 38 (2014)