Inferring Interests from Mobility and Social Interactions

Anastasios Noulas
Computer Laboratory
University of Cambridge
an346@cam.ac.uk

Mirco Musolesi
Computer Laboratory
University of Cambridge
mm753@cam.ac.uk

Massimiliano Pontil
Dept. of Computer Science
University College London
m.pontil@cs.ucl.ac.uk

Cecilia Mascolo
Computer Laboratory
University of Cambridge
cm542@cam.ac.uk

Abstract

In recent years there has been an explosion in the availability of data sets about colocation between individuals and connectivity with specific network infrastructure access points, from which user location can be inferred. These traces are usually collected through mobile devices equipped with short-range radio interfaces, such as Bluetooth. Their potential is enormous as user movement data can be mapped onto the geographical space and the social interactions of individuals can be extrapolated from the colocation data. Quite interestingly, some of these data sets also contain a description of user profiles, such as the interests of the person, his/her age and gender and so on.

In this paper we show that mobility and colocation information (i.e., social interactions) can be used to infer user interests by applying standard machine learning techniques. We evaluate a supervised and a semi-supervised technique using two different data sets containing information of interactions amongst people at conferences. We assume different degrees of available information for the inference problem and show that we are able to predict people's interests with good accuracy also when only a small amount of information about user interests is available. While correlation of user interests with movement and proximity has already been investigated in social network research, this is the first work that uses machine learning to show this quantitatively.

1 Introduction

While social network studies have established the correlation between human movement and proximity with human interests [17], no quantitative large studies have ever been attempted to consolidate the credibility of this theory. With the soaring availability of data sets containing information about people's contacts and movements, empirical studies of human behaviour, social interaction and mobility have become possible [5, 7]. The available data sets [13] contain traces of contacts between mobile devices carried by humans, identifying the interactions between devices and the fixed infrastructure, thus giving the location of the individuals. Sometimes, these data sets also contain information about the user interests, gender, membership, institutions and so on, allowing for an analysis of the relationships between people social interactions and their (common) interests.

In this paper we investigate the problem of inferring *user interests* from information about their geographical location and social interaction over time, by means of supervised and semi-supervised machine learning techniques. The goal of this work is not to propose new machine learning techniques but to adapt existing ones to this novel application domain. First, we assume that full knowledge about user interests is available and we apply the *k nearest neighbours* [8] classification algorithm to infer them, proving that a priori knowledge about the movements and the interests of a population can be used to infer the interests of a generic user of that population given his/her movement patterns. More specifically, our approach is based on the exploitation of a similarity graph among the movement patterns of the individuals considering different representations based on frequency and residence duration in a certain location (measured by means of proximity with base stations) and close proximity (measured by means of short-radio technologies). Then, we assume that partial knowledge on people's interests is available and we apply a semi-supervised learning solution based on Gaussian Fields and Harmonic Functions [19] for label propagation in graphs. Both approaches are evaluated on two large data sets collected in two different conference environments (Infocom [18] and HOPE [2]) with the use of Bluetooth-equipped mobile devices and RFID transponders respectively. The data sets also contain information about the users, which was collected through questionnaires. The evaluation shows that we are able to predict people's interests with good accuracy also when only a small amount of information about user interests is available. We show that we are able to achieve an average accuracy around 74% and 80% respectively for the two data sets using the supervised learning tecniques. With respect to semi-supervised case, we demonstrate that 5 to 20% of information is sufficient to achieve results comparable to the supervised learning case for these two data sets.

2 Description of the Inference Algorithms

In this section we describe the algorithms that we use to infer user interests from colocation traces. We choose to use well established techniques and demonstrate how these can be employed in this problem domain of inferring people interests given knowledge of their interactions and geographical positions over time.

2.1 Similarity Graphs and Mobility Representations

As we mentioned, starting from data about the user location (e.g., its proximity to an access point) and users interaction (e.g., through the detection of each other's Bluetooth signals), our approach aims at identifying similarity of users and predicting their interests.

We define a data set $\mathcal{D} = \{(x_i, y_i), i = 1, ..., m\}$ with *m* equal to the number of users. We use a multi-dimensional vector, $x \in \Re^d$ to describe a person's mobility and interaction behaviour as explained below. $y_i \in \{+1, -1\}$ indicates an interest label. The value of y_i is equal to $+1$ if a user has expressed that interest, -1 otherwise. Using this model, we can construct an *m*-node similarity graph $G = (V, E)$. The weight associated to a generic edge between node i and node j is equal to the similarity values between two data points x_i, x_j. The similarity is defined as the inverse of the Euclidean distance between these two data points. More formally, considering two users *i* and *j* we define these quantities as $distance_{i,j} = \|x_i - x_j\|$ and $similarity_{i,j} = 1/\left(distance_{i,j} + 1\right)$.

Hence, edges between points which are close to each other in the Euclidean space will be assigned large weights in the graph. For our purposes, we define and experiment with three different similarity graphs which correspond to three mobility representations extracted from different movement measurements and depending on different constructions of the vector x:

Frequency: The definition of the *frequency* representation involves adding to each feature of x a value equal to the times a person has visited a specific location. The latter implies that the dimensionality of a vector is equal to the number of existing locations. This information can be extracted by means of fixed base stations.

Duration: Next, we build the *duration* representation where the frequency metric is replaced by the total residence time an individual has spent in a location. Also this information can be extracted by means of fixed base stations.

Colocation: Finally, we derive a third similarity metric, where each edge in the mobility graph *G* is being weighted according to the total time two users were collocated during the period of the experiment (i.e., *colocation*). The latter captures a person's close social interactions. This information can be extracted by means of short-range radio technologies such as Bluetooth (the transmission range is under 10 m).

2.2 Supervised Learning

Overview Supervised learning is a branch of Machine Learning that deals with the problem of defining a predictor function f that relates two different data spaces: $\mathcal{X}$, which is defined as the input space and $\mathcal{Y}$ which is the output space. In general, $\mathcal{X}$ and $\mathcal{Y}$ are presented in the form of the training set of input and output pairs $\mathcal{D} = \{(x_i, y_i), i = 1, ..., m\}$ with x_i and y_i being members of $\mathcal{X}$ and $\mathcal{Y}$ respectively.

The assessment of the solutions f in supervised learning requires two additional definitions. For any point x_i the solution of f predicts an answer $f(x_i)$. We are interested in how far this prediction is from the actual output label y_i. This distance is expressed through the loss function $\ell(f(x_i), y_i)$. In the context of binary classification the loss function can take two values; 0 for correct prediction or 1 otherwise. The empirical error, which is the application of the loss function over the whole training set, is then defined as follows:

$$R_{emp} = \frac{1}{m}\sum_{i=1}^{m} \ell(f(x_i), y_i)$$

In addition, we define the generalization error for points x which do not belong to the training set $\mathcal{D}$ as a mean for assessing the quality of the solution f for new inputs: $R_{gen}(f(x)) = E_{x,y}[\ell(f(x), y)]$, where $E_{x,y}[.]$ is the expectation with respect to pair (x, y). The generalization error is the principal metric used to assess the performance of a supervised learning solution. In the next paragraph we describe the predictor function f and in Section 3 a method to estimate its generalization error.

k Nearest Neighbors Based Interest Inference The *k nearest neighbors* algorithm (k-NN) [8] is one of the simplest Machine Learning algorithms. The prediction function $f(x_i)$ for a given data point x_i is equal to the average of the labels of the *k* closest points to x_i. We have implemented and run the algorithm for the different values of the parameter *k*.

We conjecture that a good strategy for inferring people's interests is to consider those that have similar mobility behaviour (i.e., the nearest to them in terms of mobility patterns). In this work we consider the prediction of each interest independent from each other, but the model can be extended to include the other interests in the prediction model. More formally, we predict a generic interest of user *i* as follows:

$$f(x_i) = \begin{cases} +1, & \text{if } \frac{1}{|N(i)|}\sum_{j \in N(i)} y_j \geq 0 \\ -1, & \text{otherwise.} \end{cases}$$

where y_j is equal to $+1$, if user j is interested and -1 otherwise. $N(i)$ is the set of the k closest neighbours of *i*.

2.3 Semi-supervised Learning

Overview In the semi-supervised learning scenario we investigate the case of predicting interest preferences of individuals when only a subset of labels is available with respect to the population of existing data points. This corresponds to the case where interest information is available only about a (potentially small) portion of users.

Let us consider the data set $\mathcal{D}$ and suppose that we have a set L with l label points $(x_1, y_1), ..., (x_l, y_l)$ and a set *U* with *u* unlabeled points x_{l+1}, ..., x_{l+u}. Moreover, let us suppose that we have a graph $G = (V, E)$ with nodes *V* including the set of all *m* data points ($V = L + U$). The edges E are weighted according to the similarity metrics defined in 2.1. The goal of a semi-supervised learning algorithm is to exploit the connectivity of all nodes in *G* in order to predict the labels of the unlabelled points in *U*.

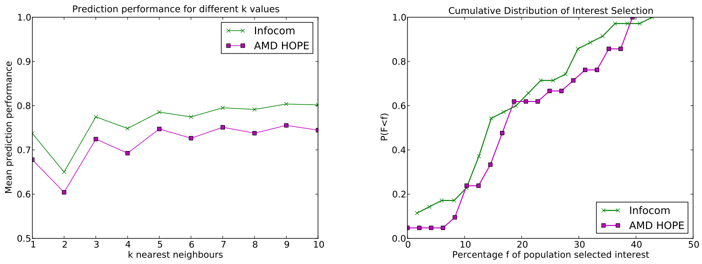

Figure 1: Performance of frequency representation over k neighbours and distribution of interests in population.

Label Propagation for Interest Prediction with Partial Availability of Information To solve the problem of predicting the labels of points in *U*, we consider a solution proposed in [19] which defines a Gaussian Random field model on the graph *G*, where the mean of the field is characterised using a harmonic function $f : V \longrightarrow \Re$. The intuition behind this approach is that the value of neighbouring points in the graph have similar values. The harmonic property of the function *f* suggests that the predicted value of an unlabelled data point can be calculated considering the average of the values of its neighbours:

$$f\left(x_i\right) = \frac{1}{d_i} \sum_{i \neq j} w_{ij} f\left(x_j\right), \; for \; i \; = l+1, ..., l+u$$

where d_i is the total sum of the weighted edges of node *i* in the graph, while w_{ij} is equal to the similarity value between nodes *i* and *j*. Since the definition of our problem falls into the binary classification paradigm and f returns real values, we simply predict a label $+1$ (i.e., the person is interested) if $f\left(x_i\right) \geq 0$ or -1 otherwise.

3 Evaluation

3.1 Data Sets

The two data sets (AMD HOPE and Infocom 2006) were collected independently and from different organizations. In both cases, experimenters used mobility tracking technologies to monitor the presence of people in a conference environment. Hence, these traces contain time-stamped information about the location of each user throughout the period of the conference. In addition to the mobility information, participants at the conferences were asked to respond on questionnaires relevant to the topic of the conference. In other words, these traces represent unique multi-dimensional data sets for the evaluation of the proposed inference algorithms.

Infocom 2006 This data set was collected at the IEEE Infocom 2006 conference. The event lasted 4 days. Scott et al. [18] distributed a set of imote devices to 70 students and researchers. Imotes are Bluetooth capable devices, which are able to record the presence of other devices close to them. Another set of long range static imotes were deployed at 17 key locations in the conference area. As in the AMD HOPE data set, using data collected from the static imotes, user location is known throughout the study. Additionally, a questionnaire (35 questions) was provided to the participants, calling them to express interests with respect to their area of expertise. As in the AMD HOPE data set, only a subset of users filled the questionnaire. For this data set, however, we have a bigger sample when compared to the total population (61 out of 70).

AMD HOPE AMD, or ”Attendee Meta-Data”, is a project that aims to explore potential uses of RFID technology. The AMD Last HOPE (Hackers On Planet Earth) conference [2] was an attempt to show how RFID tags could be used within a conference environment to enhance the experience of

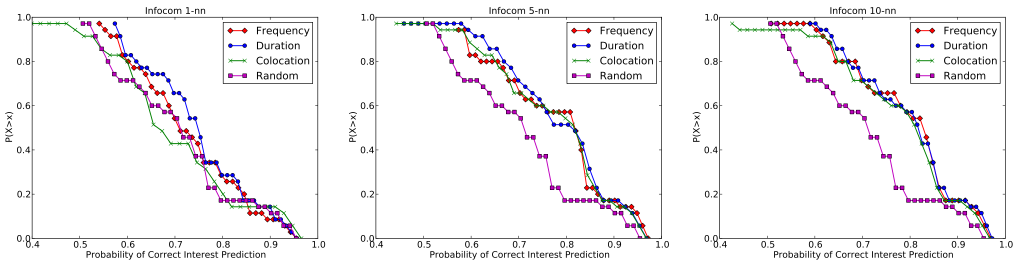

Figure 2: Infocom data set interest prediction results for 1, 5 and 10 nearest neighbor cases.

Interest:	"Ad Hoc Nets"	"Multimedia"	"Sensor Nets"	"Security"	"Traffic Analysis"	(mean)
Colocation	0.54	0.85	0.83	0.62	0.67	0.78
Frequency	0.62	0.85	0.83	0.70	0.60	0.78
Duration	0.72	0.86	0.85	0.70	0.63	0.79
Random	0.50	0.77	0.72	0.58	0.55	0.71

Table 1: Infocom 5-NN prediction performance for 5 interest samples & overall mean.

the attendees. People wearing tags were tracked for the course of three days. Moreover, participants were asked to express their interests on an online form. RFID and expressed interests were used by the organizers of the conference to help people in networking. RFID readers were deployed at 21 locations throughout the conference area for tracking the participants. The questionnaire form contained a list of 21 interests and users were called to choose at most 5 among those. Despite the fact that the overall number of users who have used RFID tags were 1281, only 410 of them decided to fill the questionnaire. Hence, for the purposes of the present work, we have used only this subset of users.

3.2 Experimental Results

We present our experimental results over the two data sets for the supervised and semi-supervised learning techniques. We solve a multi-label classification problem: for both cases the number of binary classification tasks is equal to the number of interests for each data set. That is 21 for AMD HOPE and 35 for Infocom 2006.

Supervised Learning We estimate the generalization error of the *k nearest neighbours* algorithm with the leave-one-out error technique comparing it to a random predictor (see also in [8]). The technique is an unbiased estimator of the generalization error of the learning algorithm. As the name implies, we calculate the error of the loss function l corresponding to point x_i by training the function f for all input-output pairs but (x_i, y_i). More formally, the leave-one-out error is defined as:

$$R_{loo}(f) = \frac{1}{m} \sum_{i=1}^{m} \ell(f^i(x_i), y_i)$$

where m is the number of data points and f^i suggests that f was derived by excluding pair (x_i, y_i) from the training set. We use the notion of *probability of correct interest prediction* which is complementary to the leave-one-out error estimator.

We plot curves with results for the three representations that are used for the construction of the mobility graphs: *frequency, duration* and *collocation*. We compare the performance of those with a *random* prediction case, in order to demonstrate how mobility information can be used to improve the accuracy of the inference task. The *random* prediction is calculated using the following equation:

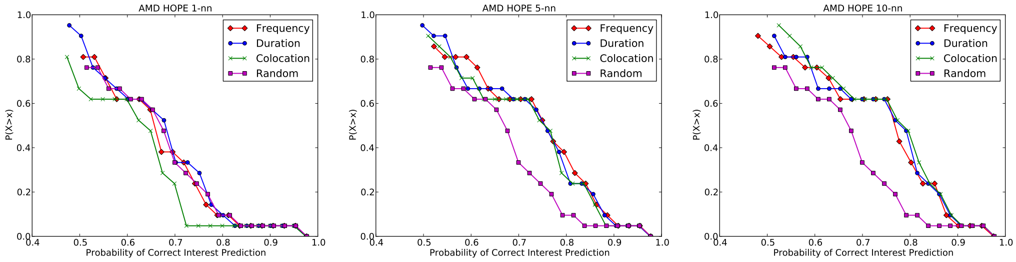

Figure 3: AMD HOPE data set interest prediction results for 1, 5 and 10 nearest neighbor cases.

Interest:	”New Technology”	”Ethics”	”Privacy”	”Cryptography”	”Network Security”	(mean)
Colocation	0.59	0.79	0.63	0.57	0.54	0.73
Frequency	0.49	0.83	0.61	0.61	0.58	0.74
Duration	0.53	0.84	0.65	0.65	0.52	0.74
Random	0.51	0.74	0.56	0.57	0.51	0.68

Table 2: AMD HOPE 5-NN prediction performance for 5 interest samples & overall mean.

$$P_{random}(i) = \left(\frac{t}{n}\right)^2 + \left(1 - \frac{t}{n}\right)^2$$

with n number of users and t the number of times an interest i was selected. The random predictor is built according to a probability distribution based on the prior knowledge of selection frequency of each interest; interests with unbalanced distribution in the population can be predicted more easily. In Figure 1 $(right)$ we plot the cumulative distribution of interest selection frequency in each data set (i.e., how many times an interest was selected). We can observe that some very popular interests are selected by almost half of the two populations and that a small set of interests concern only a few participants.

The parameter *k*, defining the number of neighbours, plays a key role in the accuracy of the algorithm. We experiment with $k = 1, ..., 10$. Figure 1 $(left)$ shows that the biggest improvement in terms of prediction performance correspond to the values k equal to 3 and 5. There is no significant increase in performance beyond 5-NN for both data sets. An explanation of this behaviour is that as we go beyond 5-NN, although the information over the distribution of interests in a population increases, we move away from a person's social ties which are expected to share similar interest preferences.

By comparing results presented in Figures 2 and 3 for the cases of 1, 5 and 10 nearest neighbours, we can observe that the accuracy increases as the number of neighbours increases. The predictor based on k-NN outperforms the random one for values of k bigger than 2 (with k equal to 1 and 2 we have an underfit). Another common observation for all graphs is that the three different mobility representations offer very similar performance. This is probably due to the fact that nodes that these three representations are highly correlated. The attendees mostly met inside the conference venue. Therefore, there is a high proportion of the contacts detected by means of the short-range technologies (i.e, colocation information) that is also recorded by the access points (i.e., residence interval duration information). Moreover, it is possible to observe a proportionality between the duration of the contacts and their frequency. However, we observe small variations on a per interest basis: in Tables 1 and 2 we provide a comparison of the three representations and the random predictor for a subset of interests. An additional observed characteristic in Figures 2 and 3 is that the curves related to the three mobility representations and to the random predictor overlap in two cases. First, as mentioned above there are a few interests with unbalanced distributions whose prediction is easy even for the random predictor. For example in the Infocom dataset we have prediction accuracy higher than 90% for 20% of the interests. The second scenario where random and mobility curves overlap is related to a subset of interests that are very hard to predict as their choice of preference is evenly distributed in the population and therefore the uncertainty associated with them is high.

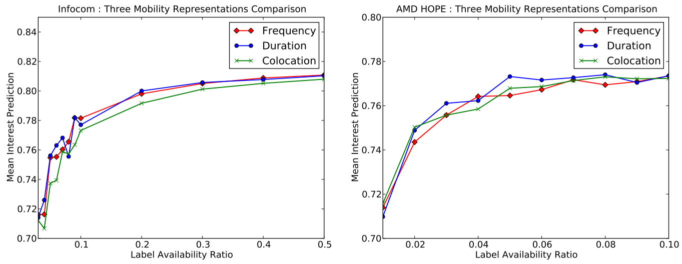

Figure 4: Semi-supervised Learning performance comparison for Infocom and AMD HOPE data sets.

We also observe that we achieve better results for the Infocom 2006 traces. This is due to the fact that the AMD HOPE data set is relatively sparse, since many users did not fill the questionnarie and were removed from the data set.

Semi-supervised Learning As far as the semi-supervised learning algorithm is concerned, for each run we calculate an evaluation score, which is equal to the number of unlabelled nodes that were predicted correctly over the total number of unlabeled nodes. With respect to the semi-supervised learning methodology we assume that only a portion of the interest labels are available in the data set. The availability percentages we have experimented with range from 1% for AMD HOPE and 3% for Infocom to 90% of the two populations. For each run of the experiment, we randomly select a subset of the nodes in our mobility graph and assign labels to them. We run the classification task 50 times for each interest for the different levels of partial knowledge available. There are two notable points with respect to the experimental results presented in Figure 4. First, even with small number of labels available high performance can be achieved, which remain steady as we increase the label availability ratio. That is 20% for the Infocom (12 labels out of 61 users) and 5% of labels for the AMD HOPE (20 labels out of 410 users) dataset. Second, the three mobility representations offer almost identical performance, as observed for the supervised learning case.

4 Related Work

We view our work as part of the broader research areas of the analysis interconnections between social and technological networks [12] and learning over graphs [4, 9, 19]. Data from technological networks have been exploited to investigate small world phenomena in social networks [16], influence of friends in purchase decisions [10, 14] and dynamics of spread of information [1]. More specifically, machine learning techniques have been successfully applied in a number of cases of social network analysis solving problems on various aspects such as link prediction [15], group problem solving [11] and evolution of communities [3]. Most of these works focus on online social networks. Recently, there has also been an increasing interest in the analysis of mobile networks. The Reality Mining project [6] is a representative example: smart phones were given to staff and students of the MIT Media Lab and Sloan Business School to log location and proximity data over the course of 9 months. The authors use the information about user movements to infer social relationships and routines [5].

Our work investigates for the first time quantitatively the correlation between mobility and user profiles: we have shown how machine learning techniques can be applied to extract not only information about relationships among users but also about their profiles from social interactions detected by means of short-range radio technologies.

5 Conclusions

We have presented an investigation about how machine learning techniques can be used to infer user interests from mobility patterns in conference environments. Our approach is based on the construction of similarity graphs of mobility patterns among the users. We have considered two different cases, characterised by different percentage of available information as a priori knowledge for the inference algorithms and we have applied a k nearest neighbours algorithm for the case of full knowledge and a label propagation technique for the case of partial knowledge. We have evaluated the techniques using two real-world data sets from two different conference environments. We have shown that we can achieve an average accuracy around 74% and 80% respectively for the two data sets using the supervised learning tecniques. With respect to semi-supervised case, we have demonstrated that 5 to 20% of information is sufficient to achieve results close to the supervised learning case for these two data sets.

References

[1] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem at WWW'04*, New York, NY, 2004.

[2] aestetix and C. Petro. CRAWDAD data set hope/amd (v. 2008-08-07). Downloaded from http://crawdad.cs.dartmouth.edu/hope/amd, Aug. 2008.

[3] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. *Proceedings of the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (KDD '06)*, pages 44–54, 2006.

[4] M. Belkin, I. Matveeva, and P. Niyogi. Regularization and semi-supervised learning on large graphs. In *Proceedings of COLT '04*, pages 624–638. Springer, 2004.

[5] N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, August 2009.

[6] N. Eagle and A. S. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, May 2006.

[7] M. C. Gonzalez, C. A. Hidalgo, and A. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[8] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, second edition, 2005.

[9] M. Herbster, M. Pontil, and L. Wainer. Online learning over graphs. In *Proceedings of the 22nd International Conference on Machine Learning (ICML'05)*, 2005.

[10] S. Hill, F. Provost, and C. Volinsky. Network-based marketing: identifying likely adopters via consumer networks. *Statistical Science*, 22(2):256–275, 2006.

[11] M. Kearns, S. Suri, and N. Montfort. An experimental study of the coloring problem on human subject networks. *Science*, 313(5788):824–827, August 2006.

[12] J. Kleinberg. The convergence of social and technological networks. *Commun. ACM*, 51(11):66–72, 2008.

[13] D. Kotz and T. Henderson. CRAWDAD: A Community Resource for Archiving Wireless Data at Dartmouth. *IEEE Pervasive Computing*, 4(4):12–14, 2005.

[14] J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 1(1), May 2007.

[15] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM '03)*, pages 556–559, 2003.

[16] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, August 2005.

[17] M. Mcpherson, L. S. Lovin, and J. M. Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[18] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau. CRAWDAD trace cambridge/haggle/imote/infocom2006 (v. 2009-05-29). Downloaded from http://crawdad.cs.dartmouth.edu/cambridge/haggle/imote/infocom2006, May 2009.

[19] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 12th International Conference on Machine Learning (ICML'06)*, pages 912–919, 2003.