Uncovering the Location of Twitter Users

Conference Paper · October 2013

DOI: 10.1109/BRACIS.2013.47

CITATIONS
11

READS
302

5 authors, including:

Érica Alvarenga Crespo Rodrigues
Universidade Federal de Ouro Preto
533 PUBLICATIONS **22,638** CITATIONS
SEE PROFILE

Assuncao Renato Martins
Federal University of Minas Gerais
146 PUBLICATIONS **4,333** CITATIONS
SEE PROFILE

Gisele L. Pappa
Federal University of Minas Gerais
154 PUBLICATIONS **1,942** CITATIONS
SEE PROFILE

Renato Miranda
Federal University of Juiz de Fora
47 PUBLICATIONS **440** CITATIONS
SEE PROFILE

Uncovering the location of *Twitter* users

Erica Rodrigues
Departamento de Estatística
Universidade Federal de Ouro Preto
Ouro Preto, Brazil
ericarodrigues@iceb.ufop.br

Renato Assunção, Gisele L. Pappa, Renato Miranda, Wagner Meira Jr.
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
{assuncao,glpappa,renato.miranda,meira}@dcc.ufmg.br

***Abstract*—Social networks, like *Twitter* and *Facebook*, are valuable sources to monitor real-time events, such as earthquakes and epidemics. For this type of surveillance the user's location is an essential piece of information, but a substantial number of users choose not to disclose their geographical information. However, characteristics of the users' behavior, such as the friends they associate with and the types of messages published can hint on their spatial location. In this paper, we present a method to infer the spatial location of *Twitter* users. Unlike the approaches presented so far, we incorporate two sources of information to learn the geographical position: the text posted by users and their friendship network. We propose a probabilistic approach that jointly models the geographical labels and the *Twitter* texts of the users organized in the form of a graph representing the friendship network. We use Markov random field probability model to represent the network and learning is carried out through a Markov chain Monte Carlo simulation technique to approximate the posterior probability distribution of the missing geographical labels. We demonstrate the utility of this model in a large dataset of *Twitter* users, where the ground truth is the location given by GPS dispositives. The method is evaluated and compared to two basline algorithms that user these two types of information separatdly, The accuracy rates obtained are significantly better than those of the baseline methods.**

***Index Terms*—Network Learning; Geographic Targeting; Geolocation Estimation; Spatial Data Mining.**

I. Introduction

Created in 2006, *Twitter* is a microblog through which users can post messages up to 140 characters, called *tweets*. Users of this service can act as a kind of radar that provides real-time information about events such as earthquakes [1], epidemics [2], etc. To carry out this type of surveillance, an essential piece of information is the user's location. In Twitter, a user location can be provided in three different ways, each one with different degrees of precision and accuracy. In the first case, the user can manually fill in in his profile where he lives. Because this field is freely filled, a large volume of invalid (Mars) or low precision (Brazil) locations are reported by users. The second way is to obtain the geographic location from the computer IP address. This type of georeferencing is not very reliable and needs to be continually updated. In Brazil, for example, this service correctly locates 72% of IP's within a radius of 40 kilometers [3]. The third source of information is obtained from GPS coordinates of mobile phones. This third kind is the one with best accuracy and reliability, since it is restricted to those cases in which the user posts a message from a mobile device with GPS and allows such information to be disclosed. However, such geographical information is present in only a small fraction of *tweets*. In some countries like Brazil, this proportion is below 5% [].

Despite the geographic information not being explicit in most cases, some aspects of user behavior can give us hints about his location. For example, the set of tweets he publishes may provide us with information about where he lives. Some works have been developed in this direction [4], [5]. [5], for instance, estimate the user's location by identifying words that characterize certain locations. For example, the term "rocket" is typically from the city of Houston. The authors define these local words as those with high frequency at a given point in space but falling quickly as we move away.

In addition to the text, relationships of follower/followed between users can also bring geographical information. Especially in countries where the spoken language is not English, friendship relationships in *Twitter* tend to reflect the geographical proximity between users [6]. Taking this into consideration, the friendship network can be used as a source of information for the inference process. [7] proposed an estimation method whereby a user's location is set as the one which is the most frequent among his friends. In determining the friendship relationships, they consider that two users are friends only if they follow each other reciprocally. This prevents institutional pages or celebrity profiles to disturb the inference process. Some of the problems encountered by [7] refer to the reduced number of friends that some users have, making the inference process very difficult. Moreover, users with many friends are also a source of errors, because such friendship relationships most likely do not reflect geographical proximity. The method proposed by [7] is quite simple, but requires calibration of parameters. Other methods of classification for data organized in graphs have been presented in the literature and a comprehensive review on the subject can be found in [8].

Although tweets content and the friendship network have been previously used isolated to infer the location of a Twitter user, to the best of our knowledge, so far no published work has presented an approach that integrates these two types of data. In this direction, this paper presents a method that explores the content of the messages, available for all users, with the friendship relations and location information from a subset of the users. The users with known locations include only GPS acquired information. The method is based on a Gibbs Sampler, which is a simple Markov chain Monte Carlo

algorithm.

The proposed method was tested in a set of 8,477 users from three different cities and more than one million connections. The results were compared with methods using content and the friendship graphs separately, namely Naive Bayes and MRW (MultiRankWalk). The results show that the aggregation of both types of information can improve the overall accuracy results from 60.46 to 73.98%.

The results obtained are better than those attained by the previous algorithms proposed in the literature. This demonstrates the value brought by the combination of the two evidence sources in learning the missing locations. The rest of the paper is organized as follows. Section III describes our probabilistic model and Section IV presents the experimental results obtained so far. In Section V, we present the main conclusions and future work.

II. Related Work

This section reviews works in the literature that consider the text, information of user profile or the relationship graph when predicting the greographical location of a user.

Among the works that consider text is [5], where a classifier is used to automatically identify words within the tweets that are strongly related to a local geographic scope, and then user locations can be estimated using a smoothing model that searches for the identified words in the messages.

[4], in contrast, infers the user's location based on text together with tweeting behavior (volume of tweets per time unit), and external location knowledge (e.g., dictionary containing names of cities and states). They use an ensemble of classifiers to explore the features aforementioned. The use of dictionaries is a very popular approach when looking for locations in Web text in general, as showed by [9].

With respect to works using the Twitter friendship network, a comprehensive review about learning from graph relationships can be found in [8]. Regarding works focused on the Twitter graph, [7] proposed an estimation method whereby a user's location is set as the one which is the most frequent among his friends.

[10] proposed WRW (MultiRankWalk), a method able to classify sites in a semi-supervised manner, i.e. where only few vertices of the graph need to be labeled. Their idea is to use an algorithms similar to PageRank [11] to infer the labels of the vertices, creating multiple rankings using random walks from seed (labeled) instances. This approach is the one we compare with the method proposed here.

Crandall et al. [12] proposed a very different approach where, knowing that two users have been in approximately the same geographic location at approximately the same time, on multiple occasions, it estimates the probability of them knowing each other.

Finally, Gonzales et al. [6] investigated the effects of locality in Twitter, focusing specially in user/followers location relations. One of their results shows that, in countries where English is not the first language, there is a high intra-country locality among users and their followers, while English-speaking countries suffer from what they call external locality effect, having many of their followers in the U.S..

III. A probabilistic model for inferring missing location

This section describes the proposed method to infer users location based on both the content of his tweets and his friendship graph. Let N be the total number os users, θ_i the i-th user location and $\boldsymbol{\theta} = (\theta_L \cup \theta_U)$, where $\theta_L = (\theta_1, \theta_2, \ldots, \theta_k)$ is the set of k users with known location (labeled nodes) and $\theta_U = (\theta_{k+1}, \theta_{k+2}, \ldots, \theta_N)$ the set of users with unknown location (unlabeled nodes). Denote by $\boldsymbol{\theta}_{-i}$ the $N-1$-dimensional vector with all the location labels except that of the i-th user. The friendship graph, our first data source, is defined by $G =< V, E >$, where V represents the set of N users (vertices) and E the mutual follower relationships between pairs of users.

Now let $\mathbf{w}_i$ be the vector of the most common words utilized by the user in its past m tweets, where m is defined according to Twitter data collection restrictions. $\mathbf{w}_i$ represents our second data source. For a user i whose location is unknown, our aim is to find the most likely value of θ_i. In order to find this value, we need the posterior probability distribution of $\boldsymbol{\theta}$, which we approximate using the Gibbs sampler algorithm [13].

The Gibbs sampler algorithm is one of the simplest Markov chain Monte Carlo algorithms, and simulates successively from the complete conditional distribution

$$P(\theta_i|\boldsymbol{\theta}_{-i}, \mathbf{w}_i)\ .$$

to generate a sample that is approximately selected from the joint $P(\boldsymbol{\theta}|\mathbf{w}_1, \ldots, \mathbf{w}_N)$.

Initially, we factorize the complete conditional distribution

$$P(\theta_i|\boldsymbol{\theta}_{-i}, \mathbf{w}_i) \propto P(\theta_i|\boldsymbol{\theta}_{-i})P(\mathbf{w}_i|\theta_i, \boldsymbol{\theta}_{-i})\ .$$

We reduce the probability distribution of the geographical locations assuming a sparse representation for $\boldsymbol{\theta}$ based on a Markov random field model. The sparsity is induced by the assumption that, providing the information about location of a user's friends, the rest of information contained the network can be ignored.

Thus $P(\theta_i|\boldsymbol{\theta}_{-i})$ is simplified to

$$P(\theta_i|\boldsymbol{\theta}_{-i}) = P(\theta_i|\boldsymbol{\theta}_{\partial i})$$

where the vector $\boldsymbol{\theta}_{\partial i}$ contains the location of all the neighbors of i.

As the possible locations form a finite set of labels, we modeled the Markov field using Potts Model [14], a generalization of the celebrated Ising model used in image restoration. According to Potts' model, the joint probability of a given configuration depends on its energy measured by the degree of similarity between neighboring sites. The induced conditional probability that a site belongs to a particular class is an

increasing function of the number of his neighbors which pertain to the same class, ie

$$P(\theta_i|\boldsymbol{\theta}_{\partial i}) \propto \exp\left(\beta \sum_{j:j\in\partial_i} \sigma_{ij}\right)$$

where σ_{ij} is an indicator function that takes value 1 if i and j belong to the same class, and zero otherwise. The parameter β is known as the temperature of the model and measures the degree of interaction between the sites. For $\beta > 0$ we have an attractive model, ie, neighboring sites tend to belong to the same class.

The Markov random field allows for the correlation among labels exploring the graph connectivity between users. To appreciate the usefulness of this probability model, consider a situation where an user has no geographical label and is surrounded in the connection graph by other users with no geolocation information. The Markov model can still infer the geolocation of the center user by automatically looking at farther apart neighbors. In this process it also automatically takes into account the entire graph topology, such that not every user enters equally likely in predicting a given label. This is reached by means of an approximate inference algorithm, based on Markov Chain Monte Carlo methods, the Gibbs sampler. In this algorithm, a probability distribution for the entire set of labels is obtained, conditioned on those labels that are known and the text produced by all users. Therefore, for a given user with missing location, its marginal conditional probability will assign probabilities for all possible labels and the inference can be based on the maximum a posteriori or, more demanding, on the posterior mean. We adopted the second estimator for our label prediction.

Focusing on the term $P(\mathbf{w}_i|\theta_i, \boldsymbol{\theta}_{-i})$, we note that to predict a user's text, geographic information about his friends is conditionally independent given that we know his location. Thus, this probability is simplified to

$$P(\mathbf{w}_i|\theta_i, \boldsymbol{\theta}_{-i}) = P(\mathbf{w}_i|\theta_i)\ .$$

To find the value of $P(\mathbf{w}_i|\theta_i)$ we use the *Naive Bayes* classifier by considering the words posted by the user as independent of each other. That is,

$$P(\mathbf{w}_i|\theta_i) = \prod_j P(\mathbf{w}_{ij}|\theta_i)\ .$$

where w_{ij} denotes the j-th word published by the i-th user. This assumption is obviously only an approximation. Despite its simplicity, the method obtained very good results when this information was coupled with the graph structure summarized in the Markov random field model, as we show in Section IV. Each probability $P(w_{ij}|\theta_i)$ is estimated by the empirical frequency that the word w_{ij} appears among all the words published by users residing in location θ_i.

With these assumptions, the conditional probability $P(\theta_i|\boldsymbol{\theta}_{-i}, \mathbf{w}_i)$ turns out to be

$$P(\theta_i|\boldsymbol{\theta}_{-i}, \mathbf{w}_i) \propto P(\theta_i|\boldsymbol{\theta}_{\partial i}) \prod_j P(w_{ij}|\theta_i)\ .$$

TABLE I
RESULTS OF ACCURACY (%) FOR THE PROPOSED METHOD (INTEGRATED-DATA APPROACH - IDA) AND TWO BASELINES THAT CONSIDER THE INFORMATION ABOUT THE TWO DATA SOURCES INDEPENDENTLY

Method	Belo Hor.	São Paulo	Rio	Total
Class Distr.	16.54	47.91	35.55	100
MRW	23.14	55	39.78	50.27
Naive Bayes	42.15	63.7	55.99	60.46
IDA (tf)	38.02	82.95	59.26	73.98
IDA (tf-idf)	33.88	84.01	59.26	74.79

The values of θ_i for those users whose location is unknown are updated by the Gibbs Sampler algorithm.

IV. EXPERIMENTAL RESULTS

In order to verify the performance of our probabilistic model and algorithm we consider a set of 8,477 *Twitter* users residents in the Brazilian cities of Belo Horizonte, Rio de Janeiro and São Paulo. The total number of users collected in these three cities were, respectively, 1,402, 4,061 and 3,014. We consider first only these three locations because there is not enough information in small towns to do the inference. These 8,477 users form a graph composed by 1,401,715 edges, shown in Figure 1. Note that there is a strongly connected component in this graph and many isolated users, with less than two friends. Information about content was extracted from the 200 most recent tweets posted by the user. In order to validate the method, 70% of the users were randomly selected to compose the training set (5,933 users), and the remaining 30% the validation set (2,544 users).

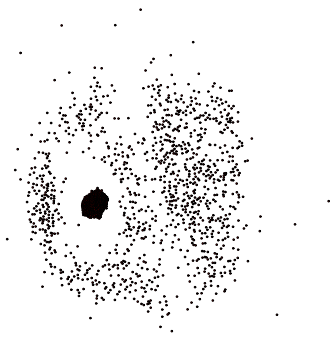

Fig. 1. Friendship graph based on 8,477 Twitter users.

The results were compared with two other methods, the first based on the friendship graph and the and the second on tweets content. The first baseline was MRW, the method proposed by [10]. We implemented their algorithm using the *Igraph* package of R [15], and set the probability of teleportation to 0.5. The second baseline was the Naive Bayes Classifier, using only the content of tweets to infer the user location. The vector of terms given to the algorithms was obtained by excluding the *stopping words* together with those words with frequency

less than three. The latter were excluded because they are most likely typos or nonexistent words. After this process, we ended up with 5,557,173 words.

Table I shows the results obtained by the three methods considering the 2,544 users in the validation set. The second line (Class Distr.) represents the data distribution along the three cities considered. Note that MRW correctly predicted the location of 50.27% of users, while Naive Bayes correctly predicted 60.46%. Hence, the information from the messages content is more relevant to predict the location tham the friendship graph by itself. Here we need to take into account that is easier for MRW to predict the locations of places with a greater number of users.

The results obtained by the proposed method are reported in the last two lines of Table I (IDA-Integrated-data approach), and the location the i-th user is estimated by that with the highest probability among all possible locations. For now let us focus on the results of IDA (tf), which considers the frequency of the words in the tweets. Overall, the method correctly infers the location of 73.98% of users. For users living in Belo Horizonte, São Paulo and Rio de Janeiro the proportion of corrected inference were, respectively, 38.02%, 82.95% and 59.26%. Combining both types of data improved the overall accuracy but, for Belo Horizonte, the precision decreased from 42.15 to 38.02. This may be due to sampling variance only, not reflecting any intrinsic aspect of the problem. Indeed, the difference between the success rates of the two methods is not statistically significant. This can be seen by building the 95% confidence interval for the rate obtained by Naive Bayes method, which is given by [33.21,50.79], covering the new success rate 38.02%

As the classes are unbalanced, Figure 2 presents two confusion matrices for the recall (top) and precision (botton) measures. The recall matrix show the probability of the user be classified as being from city i, given that he is in fact from city i. For example, a user from Rio de Janeiro has a probability of 82.95% to be classified as being from Rio de Janeiro; 4.32% of chance of being classified as being from Belo Horizonte and 12.73% of chance of being classified from São Paulo. The second matrix shows the values of precision, ie, the probability of a user being from city i, given that it was classified as being from that city. For example, if a user is classified as being of Rio de Janeiro, he has 82.46% of chance to be actually from Rio de Janeiro, 2.47% to be from Belo Horizonte and 15.07% of chance to be from São Paulo. Note, that for São Paulo, most of the error consist in classifying the user as being from Rio de Janeiro. This can occur due to greater interaction between users of these two cities.

A disadvantge of the preprocessing performed in the tweets is that it uses a large set of terms, most of which probably do not help to differentiate one city from another. Thus, we changed the way we assign weights to terms replacing the term frequence by the $tf - idf$ (term frequency-inverse document frequency) [16]. The value of $tf-idf$ is high when the term is rare across the base but very common in the document under analysis. Let us denote by D the complete set of documents

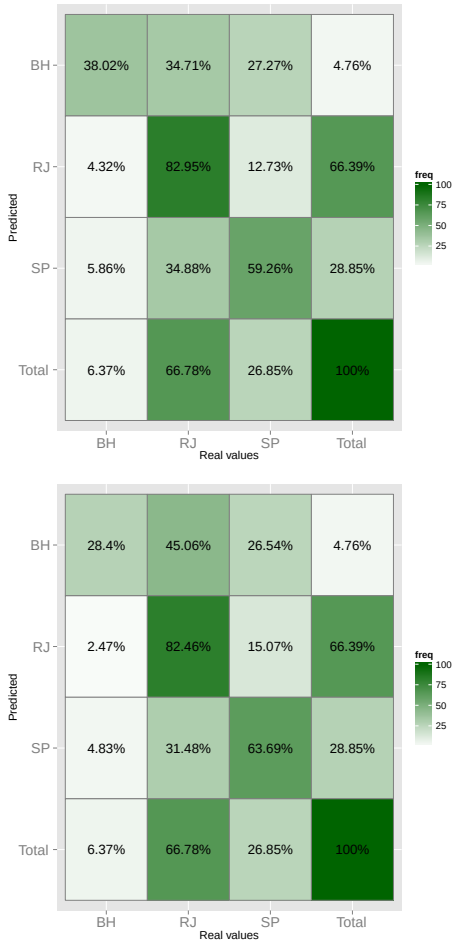

Fig. 2. Confusion matrices of the results obtained by applying the proposed methodology. The first matrix shows the recall measures and the second one, the precision measure of the method.

(in our case, set of tweets), d a particular document (tweet) and t, a specific term. The $tf-idf$ is composed of two parts. The first is the term frequency and is given by

$$tf(t,d) = \frac{f(t,d)}{\max\{f(w,d) : w \in d\}}$$

where $f(t,d)$ is the frequency of the term t in a document d. The second part is the inverse document frequency and is defined as

$$idf(t,D) = \log \frac{|D|}{\{d \in D : t \in d\}}$$

where $|D|$ is the total number of documents and $\{d \in D : t \in d\}$ is the number of documents where the term t occurs. The $tf-idf$ is given by a combination of these two terms, defined as

$$tfidf(t,d) = tf(t,d) \times idf(t,D)\ .$$

Figure 3 presents the values of recall and precision obtained using $tf-idf$. Note that the results obtained are very similar to those we got without any weighting factor. Hence, perhaps

a smart selection of the most relevant attributes should be performed before estimating the probabilities. This is left for future work.

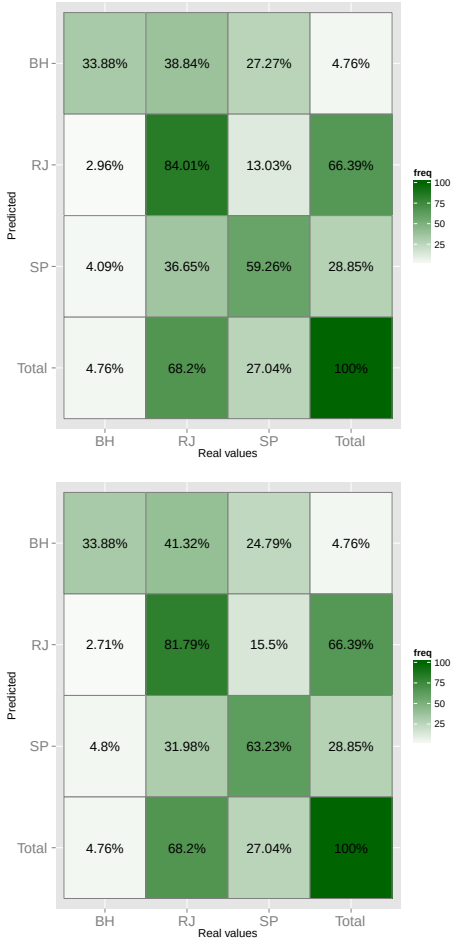

Fig. 3. Confusion matrices of the results obtained by applying the proposed methodology and correcting by Tf-Ifd factor. The first matrix shows the recall measures and the second one, the precision measure of the method.

V. Conclusions and Future Work

This work presented a probabilistic model and an algorithm to infer the location of *Twitter* users. By integrating information from the user texts and their friendship network, we were able to infer their location with substantially higher accuracy rates than baseline competitor methods proposed in the literature. Our model is based on a Markov random field distribution for the geographical labels, which is partially observed. For the text, we assume a bag-of-words model. One important novelty in this work is the use of the Markov random field model, a much more sophisticated probability description than those proposed so far by other researchers. In this model, even farther apart users in the connection graph can impact one's label probability distribution. The model automatically takes into account the distance in the neighborhood graph as well as the presence of other neighboring and redundant users. A second innovation is the merge of the two sources of information one could use to infer location: the text each user sends by *Twitter* and the friendship network represented by the mutual follower relationship. Our experimental results show that using the two sources in a combined way is worthwhile as compared to using each one of them separately.

As future work, we intend to apply our method to a larger database, with more locations and users. If the dataset is large enough to contain data with small geographical scale and, at the same time, large geographical extent, we envision that a low signal in a small town could be amplified by low signals in nearby towns, where most of the friends of a typical user could reside. Alternatively, for towns with a small number of users and, as a consequence, where spatial information is very sparse, we will seek to integrate data from other social networks like *Facebook* and *FourSquare*. Borrowing information from these other sources could have the same leverage effect as we found in using the text posted by users. Furthermore, the methodology used in this work can also be extended to analyze data from other social networks such as *Facebook*, *Flickr*, and *Instagram*. Therefore, our proposed probabilistic model have other applications beyond the specific *Twitter* case study we presented here. In all these possible applications, the main driving idea is to borrow information from one information source to help infer other hidden variables. The degree of success will depend on the correlation degree of the two information sources.

Acknowledgment

The authors would like to thank CNPq and FAPEMIG for financial support.

References

[1] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*, ser. WWW '10. ACM, 2010, pp. 851–860.

[2] J. Gomide, A. Veloso, W. Meira, V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira, "Dengue surveillance based on a computational model of spatio-temporal locality of twitter," in *Proceedings of the ACM*, ser. WebSci'11, 2011, pp. 1–8.

[3] "Geoip city accuracy for selected countries," http://www.maxmind.com/en/city_accuracy, accessed: 18/03/2013.

[4] J. Mahmud, J. Nichols, and C. Drews, "Where is this tweet from? inferring home locations of twitter users." in *ICWSM*, J. G. Breslin, N. B. Ellison, J. G. Shanahan, and Z. Tufekci, Eds. The AAAI Press, 2012, pp. 511–514.

[5] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, ser. CIKM '10. ACM, 2010, pp. 759–768.

[6] R. Gonzalez, R. C. Rumín, Á. Cuevas, and C. Guerrero, "Where are my followers? understanding the locality effect in twitter," *CoRR*, vol. abs/1105.3682, 2011.

[7] C. A. Davis Jr., G. L. Pappa, D. R. R. de Oliveira, and F. de L. Arcanjo, "Inferring the location of twitter messages based on user relationships," *Transactions in GIS*, vol. 15, no. 6, pp. 735–751, 2011. [Online]. Available: http://dx.doi.org/10.1111/j.1467-9671.2011.01297.x

[8] S. A. Macskassy and F. Provost, "Classification in networked data: A toolkit and a univariate case study," *Journal of Machine Learning Research*, vol. 8, pp. 935–983, 2007.

[9] E. Amitay, N. Har'El, R. Sivan, and A. Soffer, "Web-a-where: geotagging web content," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '04. New York, NY, USA: ACM, 2004, pp. 273–280. [Online]. Available: http://doi.acm.org/10.1145/1008992.1009040

[10] F. Lin and W. Cohen, "Semi-supervised classification of network data using very few labels," in *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 192–199.

[11] A. L. Langville and C. D. Meyer, *Google's PageRank and beyond - the science of search engine rankings*. Princeton University Press, 2006.

[12] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg, "Inferring social ties from geographic coincidences," *Proceedings of the National Academy of Sciences*, vol. 107, no. 52, pp. 22 436–22 441, 2010.

[13] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan, "An Introduction to MCMC for Machine Learning," *Machine Learning*, vol. 50, 2003.

[14] S. Z. Li, *Markov random field modeling in image analysis*, 3rd ed. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2009.

[15] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal*, vol. Complex Systems, p. 1695, 2006.

[16] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.