Leveraging Homophily to Infer Demographic Attributes: Inferring the Age of Twitter Users Using Label Propagation

Peter Li PETER.LI@NYU.EDU
New York University, Center for Data Science
HRL Laboratories

Jiejun Xu and Tsai-Ching Lu {JXU,TLU}@HRL.COM
HRL Laboratories

1. Introduction

Homophily is a basic organizing principle in social networks that states connections between similar people occur at a more frequent rate than between dissimilar people. In other words, people tend to associate with others who are similar to them.

Since the 1920's, social scientists have repeatedly observed this behaviour across different social networks and along multiple socio-demographic attributes. Given such a strong organizing principle in social networks, we wonder if homophily can be used in the task of latent attribute inference on social media accounts. In this task, we "use available, unstructured online data generated by individuals to infer demographic attributes" (Al Zamal et al., 2012). This is a well studied problem where researchers have worked to infer many demographic attributes e.g., age (Al Zamal et al., 2012) (Chen et al., 2015) (Nguyen et al., 2013) (Culotta et al., 2015), ethnicity (Mohammady & Culotta, 2014)(Culotta et al., 2015), gender (Al Zamal et al., 2012)(Culotta et al., 2015), politicial orientation (Al Zamal et al., 2012), etc. In this paper, we present a method to infer demographic attributes using label propagation. As a case study, we will focus on the problem of inferring the age of Twitter users.

Previous approaches to this problem typically focus on the *content* and *behaviour* of a user. These models typically follow a two-stage supervised machine learning framework. First, users are represented as a set of features engineered from user content and/or behaviour e.g., k-top n-grams, k-top hashtags, tweet frequency, retweet frequency, etc. These features are then used to train a learning algorithm such as support vector machines or random forests. Broadly speaking, these models can be seen as trying to answer the question: Can we infer your age based on *what* you Tweet?

In this paper, we approach the problem from a different angle. If social media users are truly homophilous, then we would expect users with strong social ties to also share similar demographic attributes. We propose a model that uses these social ties, directly, to infer demographic attributes. In the following sections, we present the following:

- A method to construct a graph that measures the strength of social ties between users - the *@mention network*
- Evidence that age homophily exists in the @mention network
- A graph-based algorithm that leverages homophily by spreading age labels on the the @mention network

Previous papers have used label propagation techniques to spread labels on a graph but they differ in the network type and label inference methods (Brea et al., 2014) (Speriosu et al., 2011) . To the best of our knowledge, this is the first label propagation based approach to infer the age of Twitter users.

2. Model

Traditionally, machine learning tasks can be divided into two categories - supervised learning and unsupervised learning. In supervised learning, the goal is to learn a mapping from x to y. To accomplish this, the algorithm is given a set of *examples* $X = (x_1, ..., x_n)$ and the corresponding *labels* $Y = (y_1, ..., y_n)$. In unsupervised learning, the labels are taken away and the task is to find some interesting structure in X.

Semi-supervised learning lies in between supervised and unsupervised learning. In a standard semi-supervised learning task, we are given a set of examples, $X_L = (x_1, ..., x_\ell)$, their corresponding labels $Y_L = (y_1, ..., y_\ell)$, *and* a set of unlabelled examples $X_U = (x_\ell, ..., x_{\ell+u})$.

The idea is that, even though examples in X_U do not have labels, they nonetheless contain useful information. Algorithms differ in how they use X_U but they typically make the following assumptions (Bengio et al., 2006):

- smoothness assumption - if two points x_i and x_j in a high density region are close, then so should their labels y_i and y_j
- cluster assumption - if points are in the same cluster, they are likely to be of the same class
- manifold assumption - The data lie (roughly) on a low-dimensional manifold

Label Propagation [1]

Label Propagation is one of the first instances of a graph-based, semi-supervised algorithm (Zhu & Ghahramani, 2002). The idea behind these models is to build a graph $G = (V, E)$ where the vertices, $V = X_L \cup X_U$, represent examples and the edges, E, encode some measure of similarity. Using the induced geometry and a set of know labels Y_L, graph-based algorithms attempt to assign labels, $\hat{Y}$, to all vertices.

In Label Propagation, the edge weights are computed using the RBF kernel. For two vertices, x_i and x_j, the edge between them is $e_{ij} = exp(\frac{-\|x_i - x_j\|^2}{\sigma^2})$. For each unlabelled example, the label is computed by taking a weighted (by normalized edge weight) average of its neighbours' labels. This process is repeated until convergence.

Algorithm 1 Label Propagation

Input: Graph, $G = (V, E)$
Define weight matrix, $\mathbf{W}_{ij} = e_{ij}$
Compute degree matrix, $\mathbf{D}$, where $\mathbf{D}_{ii} \leftarrow \Sigma_j \mathbf{W}_{ij}$
Initialize $\hat{Y} \leftarrow (y_1, ..., y_\ell, 0, ..., 0)$
repeat
  $\hat{Y} \leftarrow \mathbf{D^{-1}W}\hat{Y}$
  $\hat{Y}_L \leftarrow Y_L$
until convergence
$\hat{Y} \leftarrow sign(\hat{Y})$

Label Propagation On Social Networks

In its original form, Label Propagation computes edge weights as a function of the distance between examples. Since examples in a social network are users, there is no well defined metric. Instead we will set our edge weights using the strength of social ties between users. If the principle of homophily holds true, then examples with high edge weights should also have similar labels. Using this graph, we will use Label Propagation to spread age labels to unlabelled examples.

Graph Construction

To construct our graph $G = (V, E)$ - the @mention network - we first associate each user with a vertex, $v_i \leftarrow user_i$. To compute the edge weights, we follow the procedure proposed in (Compton et al., 2014) and use the number of reciprocated @mentions between users. @mentions occur when a user appends the at sign (@) to the mentioned user's name in the body of a tweet. Since this action requires an active choice and cuts into the Tweet character limit, we take this as a good measure of the strength of a social tie.

First, we count the number of mentions between all pairs of users. Let m_{ij} be the number of times $user_i$ mentions $user_j$. Then, we calculate the edge weight between v_i and v_j as a function of the number of mentions between users, $e_{ij} = f(m_{ij}, m_{ji})$. If mentions are not reciprocated, i.e., at least one of the counts is zero, we do not count that edge.

3. Experiments

In this section, we present our experiment set-up and results. For our analysis, we use a 10% sample of publicly available tweets collected between 2012 - 2014.

Graph

Following the procedure outlined above, we constructed two graphs G_{min} and G_{max} by setting $f = min()$ and $f = max()$ respectively. Both graphs have 110 million vertices and 1 billion edges. As a control, we also generated a random graph, G_{rand}, by shuffling the vertices on G_{min}. This way, we removed any homophily structure, but preserved the distribution of edge weights.

User Age Labels

The second necessary component of our model is user age. We needed age labels for an initial set of users to seed the learning algorithm. We also needed to know age to do any type of model evaluation.

Unfortunately, age data is typically not available for Twitter users[2]. To collect a large number of age labels we followed the methodology used in (Al Zamal et al., 2012). We analysed the Tweet text for users wishing themselves a happy

[1] Here, we present the algorithm for a binary classification problem. Each labelled example has a scalar label $y \in Y_L$ of either -1 or 1. Unknown labels are initialized to 0. For multi-class classification, each label will be a vector representing label distribution. In this case, each label vector will be normalized during each iteration of the algorithm.

[2] Twitter added a birthday field in July 2015. However, this is an optional feature and users can choose not to share birth year.

Id:--975	happy 20th birthday to me!! http://t.co/***
Id:--2336	happy 17th birthday to meeeeee!!!!!!!
Id:--934	happy 20th birthday to me.. #rock on! http://t.co/***
Id:--11	happy 21st birthday to meee!!!! http://t.co/***
Id:--477	happy 15th birthday to me ♡ #15 http://t.co/***
Id:--14	just got id'd for my drink at this bar in canada. happy 19th birthday to me
Id:--520	well, after many fights and tears, i'm going to bed. happy 23rd birthday to me...

Figure 1. Sample Tweets. Due to non-standard spellings, matching "Happy Birthday to Me" exactly filtered out many usable labels. Using regular expressions, we also allowed for repeated letters. This allowed us to capture Tweets like the one in row 2.

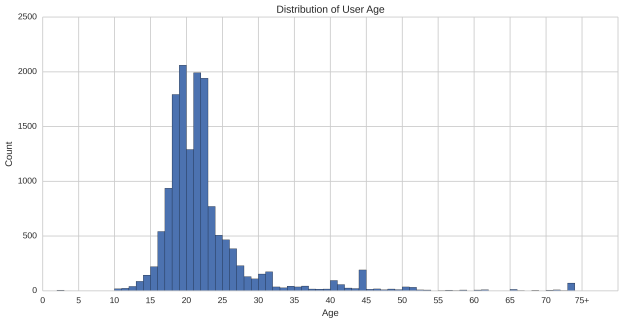

Figure 2. Distribution of User Ages (as of Dec. 2014)

birthday. Specifically, we tried to find the phrase " Happy # Birthday to Me" in the Tweet body. Since Twitter users often use non-standard spellings, we modified the approach slightly to try to match the pattern while allowing for typos and creativity. As sample of Tweets we collected are shown in Figure 1.

Using this method, we collected the ages for 14,903 users. For our experiments, we separated users into two classes - older or younger than the median age of 21. The distribution of collected ages is show in Figure 2.

Evidence of Homophily

Before running our model, we needed to verify that Twitter users are actually homophilous. To do this, we counted the number of connections between users of different ages. Theses counts were stored in a @mention matrix M, where M_{ij} was the number of times a user of age_i mentioned a user of age_j. If age homophily existed, we would expect M to have higher values along the diagonal. We plot a heatmap of M in Figure 3 and see that this is, in fact, the case.

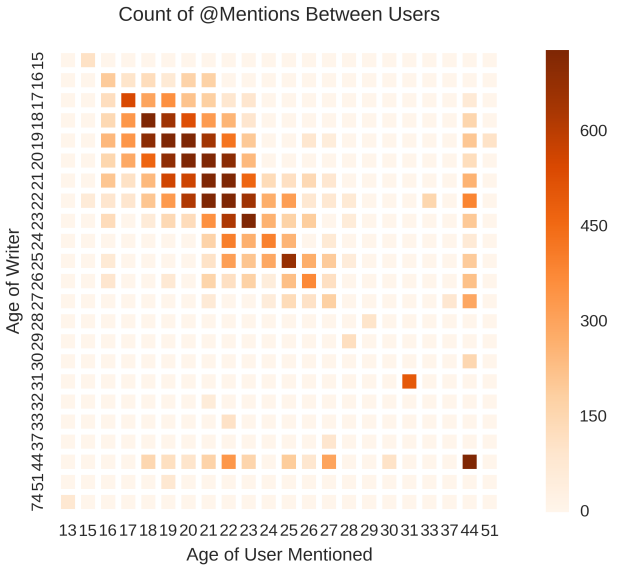

Figure 3. @mention matrix. We count the number of times a user of age_i mentions a user of age_j. We expect higher values to lie on the diagonal. For plotting, we filter out pairs with less than 50 mentions.

Model Performance

We ran our models on the @mention networks using approximately 80% of the labels to seed the model. The remaining 20% of the data was used for evaluation. We ran 5 trials and report the average performance metrics as well as standard errors. The results are show in Table 1.

We achieved 77% test accuracy using 80% of available age labels to seed the @mention network. Since there is no standard data set for this task, it is impossible to make a direct comparison to previous works. However, we can say that our performance is in line with many of the previous results for this task [3].

Table 1. Performance Metrics.

	G_{min}	G_{max}	G_{rand}
ACCURACY	0.773 (0.017)	0.773 (0.020)	0.504 (0.011)
PRECISION	0.790 (0.031)	0.796 (0.026)	0.524 (0.030)
RECALL	0.766 (0.029)	0.770 (0.025)	0.450 (0.095)
F_1 SCORE	0.777 (0.029)	0.778 (0.021)	0.483 (0.041)

[3] (Al Zamal et al., 2012) studied 386 Twitter users in a binary age classification task and reported an accuracy of 0.805 on a model using neighbourhood data (homophily). Their model without homophily achieve an accuracy of 0.751. (Chen et al., 2015) used LDA on user profiles to achieve an accuracy of 0.601 on three age categories.

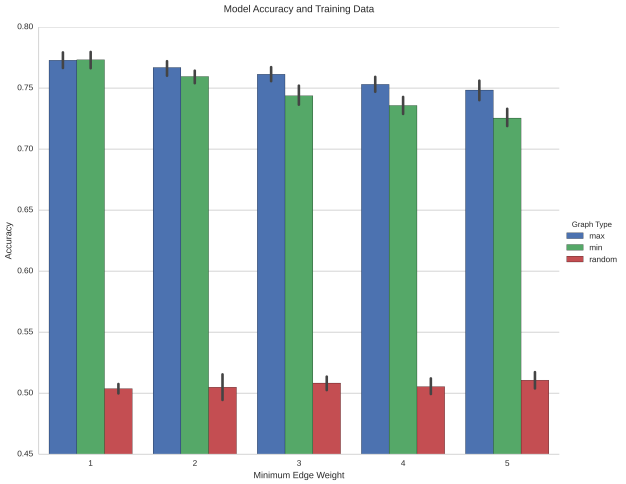

Figure 4. Model Performance on Different Networks.

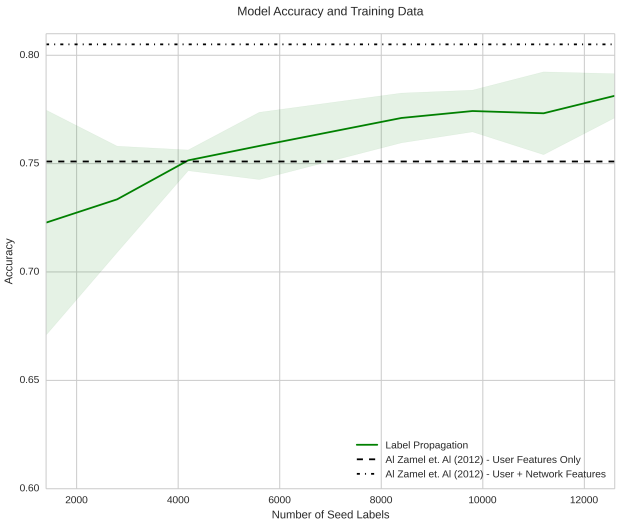

Figure 5. Model Performance Using Different Amounts of Seed Labels. Average test accuracy and standard errors using G_{min}.

Experiment I: Label Propagation on Different Networks

In the previous results, we ran our models on the three @mentions graphs. Given the importance of graph structure, we experimented with Label Propagation on different graphs.

Starting with G_{min}, G_{max}, and G_{rand} we constructed new graphs by discarding any edges less than a specified threshold. In our experiments, the threshold was set as high as 5. This pruning left us with approximately 17% of the original edges for G_{min} and G_{rand}. For G_{max} we had 33% of the original edges. The results are show in Figure 4.

Reducing the size of the graph reduced accuracy, but not by too much. When we pruned away edge less than 5, we saw a 5% drop in accuracy. However, this was achieved on a graph that was less than one-fifth the original size.

Experiment II: Varying the Size of the Seed Set

We also examined model performance using different amounts of data to seed the model. Using between 10% - 90% of the data, we ran Label Propagation using G_{min}. The results are show in Figure 5.

We saw an approximately 10% drop in accuracy between using 90% and 10% of the data. So, Label Propagation can achieve reasonable performance even when labelled examples are scarce. This is especially important, since labelled examples are often difficult to acquire.

4. Conclusion

We presented a method to infer the demographic attributes of social media users using Label Propagation on social networks. Our method achieves performance in line with previous, content-based models, but uses social tie information instead. We show that this method is relatively robust to network size and the amount of training data.

In future research, it would be interesting to further explore different network types (e.g. followers network or Retweet network) and the propagation of multiple labels.

References

Al Zamal, Faiyaz, Liu, Wendy, and Ruths, Derek. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *ICWSM*, 2012.

Bengio, Yoshua, Delalleau, Olivier, and Le Roux, Nicolas. Label propagation and quadratic criterion. *Semi-supervised learning*, 10, 2006.

Brea, Jorge, Burroni, Javier, Minnoni, Martin, and Sarraute, Carlos. Harnessing mobile phone social network topology to infer users demographic attributes. In *Proceedings of the 8th Workshop on Social Network Mining and Analysis*, pp. 1. ACM, 2014.

Chen, Xin, Wang, Yu, Agichtein, Eugene, and Wang, Fusheng. A comparative study of demographic attribute inference in twitter. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

Compton, Ryan, Jurgens, David, and Allen, David. Geotagging one hundred million twitter accounts with total variation minimization. In *Big Data (Big Data), 2014 IEEE International Conference on*, pp. 393–401. IEEE, 2014.

Culotta, Aron, Ravi, Nirmal Kumar, and Cutler, Jennifer. Predicting the demographics of twitter users from website traffic data. In *Proceedings of the International Conference on Web and Social Media (ICWSM), in press. Menlo Park, California: AAAI Press*, 2015.

Marsden, Peter V. Homogeneity in confiding relations. *Social networks*, 10(1):57–76, 1988.

McPherson, Miller, Smith-Lovin, Lynn, and Cook, James M. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pp. 415–444, 2001.

Mohammady, Ehsan and Culotta, Aron. Using county demographics to infer attributes of twitter users. *ACL 2014*, pp. 7, 2014.

Nguyen, Dong, Gravel, Rilana, Trieschnigg, Dolf, and Meder, Theo. " how old do you think i am?"; a study of language and age in twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. AAAI Press, 2013.

Speriosu, Michael, Sudan, Nikita, Upadhyay, Sid, and Baldridge, Jason. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pp. 53–63. Association for Computational Linguistics, 2011.

Ugander, Johan, Karrer, Brian, Backstrom, Lars, and Marlow, Cameron. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.

Wellman, Beth. The school childs choice of companions. *The Journal of Educational Research*, 14(2):126–132, 1926.

Zhu, Xiaojin and Ghahramani, Zoubin. Learning from labeled and unlabeled data with label propagation. Technical report, Citeseer, 2002.