# Exact Age Prediction in Social Networks

Bryan Perozzi
Stony Brook University
Department of Computer Science
bperozzi@cs.stonybrook.edu

Steven Skiena
Stony Brook University
Department of Computer Science
skiena@cs.stonybrook.edu

## ABSTRACT

Predicting accurate demographic information about the users of information systems is a problem of interest in personalized search, ad targeting, and other related fields. Despite such broad applications, most existing work only considers age prediction as one of classification, typically into only a few broad categories.

Here, we consider the problem of exact age prediction in social networks as one of regression. Our proposed method learns social representations which capture community information for use as covariates. In our preliminary experiments on a large real-world social network, it can predict age within 4.15 years on average, strongly outperforming standard network regression techniques when labeled data is sparse.

## Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: Information Networks

## Keywords

Social Networks, Network Regression, Latent Representations, User Profiling

## 1. INTRODUCTION

In recent years, there has been an abundance of work devoted to predicting hidden attributes of users from their interactions with information systems (e.g. from their search queries[5], blog posts[6], and social interactions[1, 3, 4, 8]). Surprisingly much of this existing work eschews exact prediction of user's ages, instead framing the problem as one of classification - which can mask poor model performance in the presence of outliers and noisy data.

In this paper, we consider several methods which use social interactions to estimate hidden quantities in social networks. Our preliminary experiments on a large social network shows that accurate age prediction is possible, even when as little as 5% of users have shared their age.


*WWW 2015 Companion,* May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742765.

## 2. NETWORK REGRESSION

The generalized regression problem frames the output variable $y$ as a linear combination of the inputs variables $\mathbf{x}$, plus a noise term $\epsilon$.

$$y = \mathbf{w}^T \mathbf{x} + \epsilon. \tag{1}$$

When an output variable has dependences with other 'nearby' response variables (i.e., $y$ is *auto-correlated*), the general regression model can be extended with additional covariates to model these local effects. In the social network setting, the dependencies between variables are given by the graph $G = (V, E)$ of social interactions between users. These interdependencies between output variables are typically modeled as a Markov Random Field (MRF). In a MRF, it is assumed that particular response variable $y_i$ depends on its neighborhood $\mathcal{N}_i$ (the Markov assumption). In network regression, the output variables are usually assumed to be Gaussian, which leads to the following regression problem:

$$y_i = w_{\mathcal{N}} \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} y_j + \epsilon. \tag{2}$$

We propose modeling these social dependencies instead with low dimensional social representations. Using the approach detailed in [7], we learn a mapping function $\Phi\colon i \in V \mapsto \mathbb{R}^{|V| \times d}$ which encodes each user $i$ in a $d$-dimensional representation that captures similarities between vertex co-occurrences in short random walks over the graph $G$. Our corresponding regression problem is then

$$y_i = \mathbf{w}^T \mathbf{\Phi}(i) + \epsilon, \tag{3}$$

which unlike Eq. 2 can be solved through standard methods, such as Ordinary Least Squares (OLS).

## 3. EXPERIMENTS AND RESULTS

In order to test the effectiveness of our proposed method, we conducted experiments on POKEC[9], the most popular social network in Slovakia. POKEC is interesting to study for network analysis, as it is large ($|V| = 1,632,803$, $|E| = 30,622,564$), contains user profile data, and is complete (all users are present). The age distribution of POKEC is shown in Figure 1. We consider the following methods for network regression[1]:

- **Linear Regression + DeepWalk**: Our proposed method learns social representations [7], and performs

[1]Code available at `www.perozzi.net`

| | 5% train | | 20% train | | 50% train | | 80% train | | 95% train | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *MAE* | $R^2$ | *MAE* | $R^2$ | *MAE* | $R^2$ | *MAE* | $R^2$ | *MAE* | $R^2$ |
| Linear Regression + DeepWalk | **4.15** | **0.485** | **4.15** | **0.486** | **4.15** | **0.485** | **4.15** | **0.485** | 4.13 | **0.488** |
| Iterative Algorithm | 5.97 | 0.131 | 5.17 | 0.284 | 4.43 | 0.401 | 4.16 | 0.435 | **4.09** | 0.444 |
| Neighborhood Average | 6.17 | 0.080 | 5.22 | 0.218 | 4.58 | 0.333 | 4.29 | 0.388 | 4.18 | 0.410 |
| Predict Mean | 7.015 | 0.000 | 7.015 | 0.000 | 7.016 | 0.000 | 7.015 | 0.000 | 7.015 | 0.000 |

Table 1: Age prediction performance as evaluated through both mean absolute error (*MAE*) and $R^2$ on POKEC.

an OLS regression. We use the hyperparameters (dimensionality $d$=128, window size $w$=10, walks per node $\gamma$=10, and walk length $t$=40).

- **Iterative Algorithm**: This method uses an iterative solver (in the spirit of Besag's iterative conditional modes [2]) to find a solution to Eq. 2. Such iterative approaches deal with label sparsity by propagating information within their neighborhoods, but can suffer from the presence of noisy labels.
- **Neighborhood Average**: Also known as the weighted-vote Relational Neighbor (wvRN), this algorithm assumes that each node is the average of its neighbors. In the absence of labeled neighbors, it predicts the mean.
- **Predict Mean** - This naive algorithm simply predicts the mean value of the training data.

We test these methods by taking the 1,138,314 nodes from POKEC which have ages entered in their profile, and splitting them into training and testing sets. We vary the amount of data available for training, from 5% to 95%, and test on the remainder of the data. We repeat this process 5 times and present averages. To evaluate the performance on the testing set we use the mean absolute error (*MAE*) and the coefficient of determination ($R^2$). The *MAE* corresponds to the average age in years that our predictions are off by, while $R^2$ gives an indication of how much of the total variance is captured by the model.

The results of this experiment are shown in Table 1. We see that using linear regression on DeepWalk features provides the best *MAE* until 95% of training data is used, and that it can predict a user's age within 4.15 years on average. We note that OLS+DeepWalk's $R^2$ consistently outperforms all methods as the training data is increased from 5% to 95%. This strong performance means that a more complicated regression technique may be able to extract additional gains from the representations.

For the other methods, we see that the Iterative Algorithm is better able to deal with label sparsity, and consistently has a lower *MAE* and higher $R^2$ than the Neighborhood Average. At 95% training data, the Iterative Algorithm manages to achieve a lower *MAE* than using DeepWalk features, but has a worse $R^2$ score. This indicates that the Iterative Algorithm is making more extreme errors than OLS, perhaps the result of propagating incorrect information.

# 4. CONCLUSIONS AND FUTURE WORK

These results show that accurate age detection is possible in social networks without the use of any extra profile data, blog posts, or web history. The strong performance of our method when only a very small number (5%) of nodes' ages are provided is both interesting and also worrying, raising privacy concerns.

Our continued work in the area has two focuses. First, we would like to improve prediction accuracy further - for example, could we predict the exact month and year a user was born? Second, we would like to expand upon the types of user attributes that we can infer.

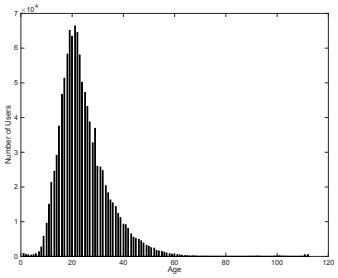


Figure 1: Age Distribution in POKEC

**Acknowledgments**: This research was partially supported by NSF Grants DBI-1355990 and IIS-1017181, a Google Faculty Research Award, and the Institute for Computational Science at Stony Brook University.

# 5. REFERENCES

[1] F. Al Zamal, W. Liu, and D. Ruths. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *ICWSM*, 2012.

[2] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, 48(3):259–302, 1986.

[3] B. Bi, M. Shokouhi, M. Kosinski, and T. Graepel. Inferring the demographics of search users: Social data meets search queries. In *WWW*, pages 131–140, 2013.

[4] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla. Inferring user demographics and social strategies in mobile social networks. In *KDD*, pages 15–24, 2014.

[5] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. In *WWW*, pages 151–160, 2007.

[6] C. Peersman, W. Daelemans, and L. Van Vaerenbergh. Predicting age and gender in online social networks. In *3rd International Workshop on Search and Mining User-generated Contents*, SMUC '11, pages 37–44, 2011.

[7] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 701–710, New York, NY, USA, 2014. ACM.

[8] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 2013.

[9] L. Takac and M. Zabovsky. Data analysis in public social networks. In *International Scientific Conference AND International Workshop Present Day Trends of Innovations*, 2012.