Improving Individual Predictions using Social Networks Assortativity

Dounia Mulders, Cyril de Bodt, Michel Verleysen {NAME.SURNAME}@UCLOUVAIN.BE
ICTEAM institute, Université catholique de Louvain, Place du Levant 3, 1348 Louvain-la-Neuve, Belgium

Johannes Bjelland JOHANNES.BJELLAND@TELENOR.COM
Telenor Research, Snarøyveien 30, N-1360 Fornebu, Norway

Alex (Sandy) Pentland PENTLAND@MIT.EDU
MIT Media Lab, Massachusetts Institute of Technology, 77 Mass. Ave., E14/E15, Cambridge, MA 02139, USA

Yves-Alexandre de Montjoye DEMONTJOYE@IMPERIAL.AC.UK
Data Science Institute, Imperial College London, 180 Queen's Gate, London SW7 2AZ, U.K.

Keywords: Belief propagation, assortativity, homophily, social networks, mobile phone metadata.

Abstract

Social networks are known to be assortative with respect to many attributes such as age, weight, wealth, ethnicity and gender. Independently of its origin, this assortativity gives us information about each node given its neighbors. It can thus be used to improve individual predictions in many situations, when data are missing or inaccurate. This work presents a general framework based on probabilistic graphical models to exploit social network structures for improving individual predictions of node attributes. We quantify the assortativity range leading to an accuracy gain. We also show how specific characteristics of the network can improve performances further. For instance, the gender assortativity in mobile phone data changes significantly according to some communication attributes.

1. Introduction

Social networks such as Facebook, Twitter, Google+ and mobile phone networks are nowadays largely studied for predicting and analyzing individual demographics (Traud et al., 2012; Al Zamal et al., 2012; Magno & Weber, 2014). Demographics are indeed a key input for the establishment of economic and social policies, health campaigns, market segmentation, etc. (Magno & Weber, 2014; Frias-Martinez et al., 2010; Sarraute et al., 2014). Especially in developing countries, such statistics are often scarce, as local censuses are costly, rough, time-consuming and hence rarely up-to-date (de Montjoye et al., 2014).

Social networks contain individual information about their users (e.g. generated tweets for Twitter), in addition to a graph topology information. The assortativity of social networks is defined as the nodes tendency to be linked to others which are similar in some sense (Aral et al., 2009). This assortativity with respect to various demographics of their individuals such as gender, age, weight, income level, etc. is well documented in the literature (McPherson et al., 2001; Madan et al., 2010; Wang et al., 2013; Smith et al., 2014; Newman, 2003). This property has been theorized to come either from influences or homophilies or a combination of both. Independently of its cause, this assortativity can be used for individual prediction purposes when some labels are missing or uncertain, e.g. for demographics prediction in large networks. Some methods are currently developed to exploit that assortativity (Al Zamal et al., 2012; Herrera-Yagüe & Zufiria, 2012). However, few studies take the global network structure into account (Sarraute et al., 2014; Dong et al., 2014). Also, to our best knowledge, no research quantifies how the performances are related to the assortativity strength. The goal of this work, already published, is to overcome these shortcomings (Mulders et al., 2017).

2. Method

We propose a framework based on probabilistic graphical models to exploit a social network structure, and

Appearing in *Proceedings of Benelearn 2017*.

especially its underlying assortativity, for individual prediction improvement in a general context. The network assortativity is quantified by the assortativity coefficient of the attribute to predict, denoted by r (Newman, 2003). The method can be applied with the only knowledge of the labels of a limited number of pairs of connected users in order to evaluate the assortativity, as well as class probability estimates for each user. These probabilities may for example be obtained by applying a machine learning algorithm exploiting the node-level information, after it has been trained on the individual data of the users with known labels. As described in (Mulders et al., 2017), a loopy belief propagation algorithm is applied on a Markov random field modeling the network to improve the accuracy of these prior class probability estimates. The model is able to benefit from the strength of the links, quantified for example by the number of contacts. The estimation of r allows to optimally tune the model parameters, by defining synthetic graphs. These simulations permit (1) to prevent overfitting a given network structure, (2) to perform the parameter tuning off-line and (3) to avoid requiring the labeled users to form a connected graph. These simulations also allow to quantify the assortativity range leading to an accuracy gain over an approach ignoring the network topology.

3. Mobile phone network

The methodology is validated on mobile phone data to predict gender (M and F resp. for male and female). Since our model exploits the gender homophilies, its performances depend on r. In the worst case of a randomly mixed network, $r = 0$. Perfect (dis-)assortativity leads to $r = (-)1$. In our network, $r \approx 0.3$, but Fig. 1 shows that it can change according to some communication attributes. The strongest edges (with many texts and/or calls) are more anti-homophilic, allowing to partition the edges into strong and weak parts, respectively disassortative and assortative ($r \approx 0.3$ in the weak part, whereas r can reach -0.3 in the strong one while retaining $\approx 1\%$ of the edges). This partition is exploited to improve the predictions by adapting the model parameters in the different parts of the network. Fig. 2 shows the accuracy and recall gains of our method, over simulated initial predictions with varying initial accuracies resulting from sampled class probability estimates. The highest accuracy gains are obtained in the range $[70, 85]\%$ of initial accuracy, covering the accuracies reached by state-of-the-art techniques aiming to predict gender using individual-level features (Felbo et al., 2015; Sarraute et al., 2014; Frias-Martinez et al., 2010). These gains overcome the results obtained with Sarraute et al's *reaction-diffusion* algorithm (2014).

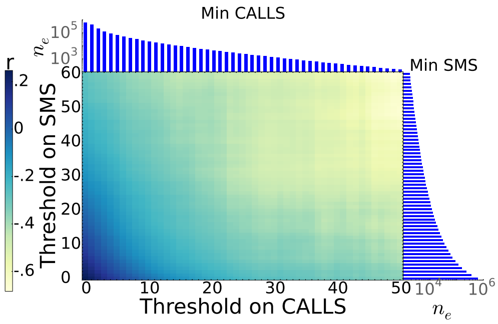

Figure 1. Gender assortativity coefficient in a mobile phone network when only the edges with a number of texts (SMS) and a number of calls (CALLS) larger than some increasing thresholds are preserved. The top (resp. right) histogram gives the number of edges (n_e) with CALLS (resp. SMS) larger than the corresponding value on the x-axis (resp. y-axis), on a log. scale.

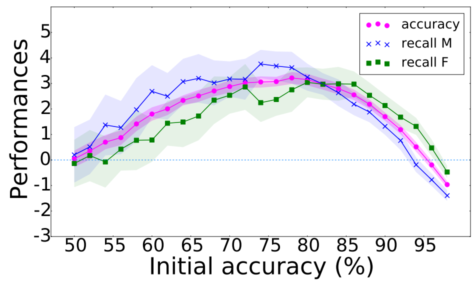

Figure 2. Accuracy and recall gains when varying the initial accuracy in a mobile phone network, averaged over 50 random simulations of the first predictions. The filled areas delimit intervals of one standard deviation around the mean gains.

4. Conclusion

This work shows how assortativity can be exploited to improve individual demographics prediction in social networks, using a probabilistic graphical model. The achieved performances are studied on simulated networks as a function of the assortativity and the quality of the initial predictions, both in terms of accuracy and distribution. Indeed, the relevance of the network information compared to individual features depends on (1) the assortativity amplitude and (2) the quality of the prior individual predictions. The graph simulations allow to tune the model parameters. Our method is validated on a mobile phone network and the model is refined to predict gender, exploiting both weak, homophilic and strong, anti-homophilic links.

Acknowledgments

DM and CdB are Research Fellows of the Fonds de la Recherche Scientifique - FNRS.

References

Al Zamal, F., Liu, W., & Ruths, D. (2012). Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. *ICWSM, 270.*

Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, *106*, 21544–21549.

de Montjoye, Y.-A., Kendall, J., & Kerry, C. F. (2014). Enabling humanitarian use of mobile phone data. *Brookings Center for Tech. Innovation.*

Dong, Y., Yang, Y., Tang, J., Yang, Y., & Chawla, N. V. (2014). Inferring user demographics and social strategies in mobile social networks. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 15–24).

Felbo, B., Sundsøy, P., Pentland, A., Lehmann, S., & de Montjoye, Y.-A. (2015). Using deep learning to predict demographics from mobile phone metadata. *arXiv preprint arXiv:1511.06660.*

Frias-Martinez, V., Frias-Martinez, E., & Oliver, N. (2010). A gender-centric analysis of calling behavior in a developing economy using call detail records. *AAAI spring symposium: artificial intelligence for development.*

Herrera-Yagüe, C., & Zufiria, P. J. (2012). Prediction of telephone user attributes based on network neighborhood information. *International Workshop on Machine Learning and Data Mining in Pattern Recognition* (pp. 645–659).

Madan, A., Moturu, S. T., Lazer, D., & Pentland, A. S. (2010). Social sensing: obesity, unhealthy eating and exercise in face-to-face networks. *Wireless Health 2010* (pp. 104–110).

Magno, G., & Weber, I. (2014). International gender differences and gaps in online social networks. *International Conference on Social Informatics* (pp. 121–138).

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 415–444.

Mulders, D., de Bodt, C., Bjelland, J., Pentland, A., Verleysen, M., & de Montjoye, Y.-A. (2017). Improving individual predictions using social networks assortativity. *Proceedings of the 12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM+).*

Newman, M. E. (2003). Mixing patterns in networks. *Physical Review E*, *67*, 026126.

Sarraute, C., Blanc, P., & Burroni, J. (2014). A study of age and gender seen through mobile phone usage patterns in mexico. *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on* (pp. 836–843).

Smith, J. A., McPherson, M., & Smith-Lovin, L. (2014). Social distance in the united states: Sex, race, religion, age, and education homophily among confidants, 1985 to 2004. *American Sociological Review*, *79*, 432–456.

Traud, A. L., Mucha, P. J., & Porter, M. A. (2012). Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, *391*, 4165–4180.

Wang, Y., Zang, H., & Faloutsos, M. (2013). Inferring cellular user demographic information using homophily on call graphs. *INFOCOM, 2013 Proceedings IEEE* (pp. 3363–3368).