Improving Individual Predictions using Social Networks Assortativity

Dounia Mulders*, Cyril de Bodt*, Johannes Bjelland†, Alex (Sandy) Pentland‡,
Michel Verleysen* and Yves-Alexandre de Montjoye§,‡

* ICTEAM institute, Université catholique de Louvain, {dounia.mulders, cyril.debodt, michel.verleysen}@uclouvain.be

† Telenor Research, johannes.bjelland@telenor.com

‡ MIT Media Lab, Massachusetts Institute of Technology, pentland@mit.edu

§ Data Science Institute and Department of Computing, Imperial College London, deMontjoye@imperial.ac.uk

***Abstract*—Social networks are known to be assortative with respect to many attributes, such as age, weight, wealth, level of education, ethnicity and gender. This can be explained by influences and homophilies. Independently of its origin, this assortativity gives us information about each node given its neighbors. Assortativity can thus be used to improve individual predictions in a broad range of situations, when data are missing or inaccurate. This paper presents a general framework based on probabilistic graphical models to exploit social network structures for improving individual predictions of node attributes. Using this framework, we quantify the assortativity range leading to an accuracy gain in several situations. We finally show how specific characteristics of the network can improve performances further. For instance, the gender assortativity in real-world mobile phone data changes significantly according to some communication attributes. In this case, individual predictions with 75% accuracy are improved by up to 3%.**

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

I. Introduction

Social networks currently drive an increasing attention in the research community, as they are found in diverse situations and are described by huge amounts of data notably collected through the web and mobile devices. Facebook, Twitter, Google+, mobile phone networks and other large-scale social graphs are nowadays largely studied for predicting and analyzing individual demographics [1]–[3]. This type of information is indeed a key input for the establishment of economic and social policies, health campaigns, market segmentation, etc. [3]–[5]. Nevertheless, especially (but not exclusively) in developing countries, such statistics are often scarce or even lacking, as local censuses are costly, rough, time-consuming and hence rarely up-to-date [6]. This is the reason why recent researches address this problem by inferring demographics from large social networks [4], [7], in order to ease the access of policy makers and NGO's toward more reliable information.

Social networks contain individual information about their users (e.g. generated tweets for Twitter), in addition to a graph topology information. These graphs present specific structures carrying many different characteristics, such as small-worldness or heterogeneous degree distribution [8]. The assortativity of social networks, defined as the nodes tendency to be linked to others which are similar in some sense [9], with respect to various demographics of their individuals such as gender, age, weight, income level, education, race, religion, etc. is well documented in the literature [10]–[14]. This property has been theorized to come either from influences or homophilies or a combination of both. For instance, Rosenquiest et al. show that social influence can enhance the spreading of alcohol consumption [15] and Madan et al. find that weight changes in an individual can be influenced by exposure to overweight peers with unhealthy habits or inactive lifestyles [11]. On the other hand, the concept of homophily is easily understood as the saying goes: "birds of a feather flock together", which means that people sharing some characteristics tend to more communicate. For instance, we observe more connections between people of the same age and gender [10].

978-1-5090-6638-4/17/$31.00 © 2017 IEEE

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 [2], [16]. However, few studies take the global network structure into account [5], [17]. Also, to the best of our knowledge, no research quantifies how the performances are related to the assortativity strength.

We here propose a framework based on probabilistic graphical models (PGMs) to exploit the social network structure for individual prediction improvement in a general context. The method can be applied while only knowing the labels of a limited number of pairs of connected users in order to evaluate the assortativity, and 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. 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 the network assortativity allows to optimally tune the model parameters, by defining synthetic graphs. The latter simulations permit (1) to prevent overfitting a given (real) 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 structure. The methodology is validated on mobile phone data to predict gender. As the assortativity required to significantly improve the quality of the prior class

Please cite as: 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+)*, pp. 127-134.

probabilities might not always be reached in practice, we show that the assortativity significantly changes according to some communication attributes, which can in turn be exploited to improve the predictions by appropriately adapting the model parameters in different parts of the network.

The paper is organized as follows. The general methodology to improve attribute predictions in a network is detailed in Section II. Its key parameters are highlighted and their tuning based on simulated assortative networks is detailed. In Section III, we introduce the real-world data sets which are studied, analyze their underlying gender homophilies and assess the performances of our method compared to a baseline algorithm. Section IV then discusses the results, while summarizing the related work. Conclusions are drawn in Section V.

In the following, uppercase and lowercase letters denote respectively random variables and observed values.

II. Method

Given an arbitrary social network $\mathcal{G}$, the goal is to exploit its assortativity to infer, for each user i, an individual scalar attribute (or class) Y_i taking values in a finite alphabet $\mathcal{Y}$. This class can be, for instance, the age or gender of each individual. $\mathcal{G}$ is defined as a pair $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ are respectively the sets of nodes (one for each user) and edges (connecting each pair of individuals who are in contact), with $|\mathcal{V}| = N$. The available individual information about user i is denoted by the vector X_i. In the case of Twitter, x_i consists in the tweets generated by user i, and possibly in public profile details (e.g. the user's name). It is assumed that estimates $\widehat{p}_{Y_i|X_i}(y_i|x_i)$ of the class membership probabilities $p_{Y_i|X_i}(y_i|x_i)$ are provided. These can be seen as "initial predictions" for each user $i \in \mathcal{V}$, which can encode deterministic information (known labels) or which can be outputted by a machine learning algorithm applied on the individual features x_i to predict the class y_i. If such information is missing for some users, uniform class probabilities are used.

Our inference model is built in Section II-A based on the social network. Next, in Section II-B, by simulating individual predictions and synthetic networks, we assess how the performance enhancement is related to the network assortativity and to the quality of the initial set of predictions, both in terms of accuracy and distribution. The latter procedure permits to determine the best model parameters.

A. Probabilistic graphical model

In order to improve the predictions $\widehat{p}_{Y_i|X_i}(y_i|x_i)$, we use an undirected PGM (also called Markov random field, MRF) with one node (resp. one edge) for each user (resp. link) in the social network. The random variables Y_i that we want to infer are assigned to the nodes of the network; each link represents a conditional dependency between two of them. As indicated in Fig. 1, the graphical model contains N additional nodes associated to the X_i's, each one being linked to its corresponding Y_i (as in [18] for instance). The relationships between the individual data X_i and the label Y_i of each user i are hence captured, as well as the direct mutual influence of adjacent users. We choose an undirected graphical model to reflect the statistical dependencies between the considered random variables, since there is no causal link between the labels in the social network which could be represented with a directed PGM (also, their joint distribution does not admit a natural factorization through conditional probabilities) [19]. Instead, our MRF represents conditional independencies. As indicated by the graph separation property [20], the joint probability distribution $p(Y, X)$ modeled by the PGM, where Y (resp. X) concatenates all the Y_i's (resp. X_i's), admits the factorization $p_{Y,X}(y, x) = p_Y(y) \cdot \prod_i p_{X_i|Y_i}(x_i|y_i)$. The underlying assumption is that X_i given Y_i is independent from Y_j and X_j, for all $j \neq i$.

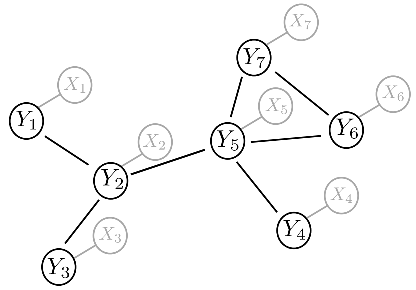


Fig. 1. Toy example of the MRF. There are two nodes per user i in the graph, Y_i being her class and X_i her individual data.

We choose a nonmaximal cliques representation through pairwise interactions for the distribution $p_Y(y)$, allowing to considerably reduce the inference cost at the expense of taking into account higher order relationships, leading to

$$p_{Y,X}(y, x) = \frac{1}{Z} \prod_i \psi(y_i, x_i) \prod_{(j,k)\in\mathcal{E}} \Psi(y_j, y_k), \tag{1}$$

where Z is a normalization constant and ψ and Ψ are called the node and edge potentials respectively. By identification with the previous factorization, the i^{th} node potential $\psi(y_i, x_i) = p_{X_i|Y_i}(x_i|y_i) \propto \frac{p_{Y_i|X_i}(y_i|x_i)}{p_{Y_i}(y_i)}$ corresponds to the likelihood of the i^{th} user's individual data knowing her class. It can be estimated using the first predicted class probability $\widehat{p}_{Y_i|X_i}(y_i|x_i)$ and the estimated class prior $\widehat{p}_{Y_i}(y_i)$ defined as the proportion of users initially predicted as y_i. In order to reflect the assortativity of the links, a simple way of defining the edge potential $\Psi(y_j, y_k)$ for each pair of adjacent users j and k is given by

$$\Psi(y_j, y_k) = \begin{cases} s_{jk}, & \text{if } y_j = y_k \\ 1 - s_{jk}, & \text{if } y_j \neq y_k \end{cases} \tag{2}$$

with $s_{jk} \in [0, 1]$ and y_j, $y_k \in \mathcal{Y}$. This way, if s_{jk} is greater than 0.5, it will encourage users j and k to share the same class. At the opposite, an s_{jk} value smaller than 0.5 will favor neighboring users j and k to have different labels (anti-homophilic contacts). This parameter can hence be interpreted as the probability for edge (j, k) to be homophilic. Depending on the application, one may have access to some edge weights, which can be used to model these s_{jk}. Section III-C provides an example of such a modeling in the context of a real-world application. Another option is to employ the same s_{jk} value for all the edges.

Along with the factorization of the joint probability distribution, the defined PGM structure allows to infer the users' class by estimating the posterior probabilities $p_{Y_i|X}$ at low cost. Exact inference on the loopy MRF is intractable, as it would require to use the junction tree algorithm [20] which, even if all the maximal cliques in $\mathcal{G}$ were identified, has an exponential complexity in the size of the largest one. This motivates the use of factorization (1), with pairwise potentials only and leads us to use the loopy belief propagation (LBP) algorithm [20]. The latter provides estimates of the posterior probabilities $\widehat{p}_{Y_i|X}(y_i|x)$ for each node i in the graph and for all $y_i \in \mathcal{Y}$. These estimates approximate the true posterior probabilities in the Bethe-Kikuchi sense [19]. The predicted class for user i is then given by $\arg\max_{y_i \in \mathcal{Y}} \widehat{p}_{Y_i|X}(y_i|x)$.

B. Parameter tuning

In our model, the s_{jk} values of the edge potential (2) have to be assigned. As these parameters reflect the confidence in the (dis-)assortative character of the edges, their tuning should be related to the network assortativity. The latter quantity has hence to be quantified. For this purpose, Newman introduced the assortativity coefficient of a network, denoted by r [14]. It allows to assess the correlations between the attributes of adjacent nodes such as the node degree, the gender or the user age[1]. It can be derived thanks to the mixing matrix $M = [m_{ij}]_{i,j=1}^{L}$, where m_{ij} is the fraction of edges connecting a vertex of class i to a vertex of class j, and L is the total number of classes. For an undirected graph, M is symmetric. Each of its row sums, denoted by m_i, gives the proportions of ends of edges from class i. In the case of a binary attribute, the mixing matrix becomes

$$M = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}, \tag{3}$$

and the assortativity coefficient is defined as

$$r = \frac{m_{11} + m_{22} - m_1^2 - m_2^2}{1 - m_1^2 - m_2^2}. \tag{4}$$

If all the edges lie between pairs of people of the same class, the network is perfectly assortative and it is straightforward to see that $r = 1$. At the opposite, in a perfectly disassortative network, there can only be as many people from each of the two classes at the ends of the edges, since each edge is between two users from distinct classes. Hence, $m_1 = m_2 = 0.5$ and r is equal to -1. In the intermediate case, a random mixing occurs when the classes of two connected users are independent. Hence $m_{11} = m_1^2$ and $m_{22} = m_2^2$, which implies that $r = 0$. Many studies show that social networks tend to be more assortative than other ones (e.g. technological or biological) [21], with positive assortativity coefficients ranging up to 0.6 [8] for attributes like race of partners in a bipartite graph of sexual partnerships[2].

As it is most of the time unknown, r should be reliably estimated in a real setting. One possibility consists in edge sampling, as shown in Section III-C in the case of gender prediction in a mobile phone network. We thus assume in the following that an accurate estimate of r is provided.

For a given network, the model parameters s_{jk} of the edge potential Ψ can be optimized according to our confidence in the (dis-)assortativity of each link (j, k). If our sole knowledge about assortativity is r, the same s_{jk} value (denoted by s) can be used for all the edges. This s expresses our confidence in the network information, which is proportional to $|r|$: as indicated by (1), large $|0.5 - s|$ values dilute the initial predictions (implied in the node potential ψ) and give a heavy weight to the network, while at the opposite an s value close to 0.5 will not change the initial predictions very much, since Ψ will take similar values for homo- and heterogeneous edges. The definition of synthetic networks with assortativity coefficients close to a given r allows us to find optimal s. To this aim, a grid search is performed: LBP is applied on the MRF using each value from the grid and the one achieving the highest average performances on different synthetic networks is kept as optimal.

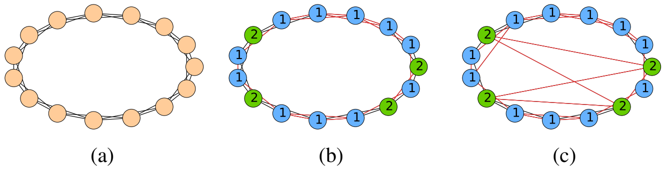


Fig. 2. (a) Regular lattice, (b) binary label assignation and (c) final graph obtained after some edges rewiring, with 15 nodes, a mean degree $k = 4$ and $r \approx 0.3$. The homogeneous edges are depicted in red.

The construction of the synthetic networks relies on the same principle as the Watts-Strogatz small-world graphs [22]. The idea is to begin with a regular circular lattice $\mathcal{G}_R = (\mathcal{V}_R, \mathcal{E}_R)$, each of the n nodes being linked to its k closest neighbors in a ring topology, where k is even. The attribute values y_i's that we want to infer are randomly assigned to each node i to follow a given distribution. Some edges are then rewired in the graph until the obtained assortativity coefficient is close enough to the targeted one, denoted by r. This last step is detailed by the following procedure, illustrated in Fig. 2.

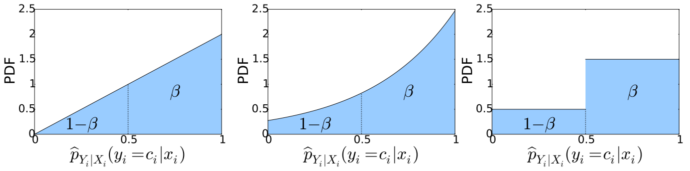
```
$r_R \leftarrow$ assortativity of $\mathcal{G}_R$ ;
while $|r_R - r| >$ tolerance do
    if $r_R < r$ then
        Randomly select an edge $(i, j) \in \mathcal{E}_R$ which is not a bridge and such that $y_i \neq y_j$
        $\mathcal{E}_R \leftarrow \mathcal{E}_R \setminus (i, j)$
        Add a random edge $(i, l)$ in $\mathcal{G}_R$ such that $y_i = y_l$
    else
        Randomly select an edge $(i, j) \in \mathcal{E}_R$ which is not a bridge and such that $y_i = y_j$
        $\mathcal{E}_R \leftarrow \mathcal{E}_R \setminus (i, j)$
        Add a random edge $(i, l)$ in $\mathcal{G}_R$ such that $y_i \neq y_l$
    end if
    $r_R \leftarrow$ assortativity of $\mathcal{G}_R$;
end while
```

It remains to endow the synthetic networks nodes with prior class probability estimates $\widehat{p}_{Y_i|X_i}(y_i|x_i)$. In a given application, a machine learning algorithm predicting the classes y_i from the individual features x_i gives us access to such prior information for all the users of the real network. We can then sample values from the distribution of the provided $\widehat{p}_{Y_i|X_i}(y_i|x_i)$ and assign them to the nodes of the synthetic graphs. Nevertheless, in order to analyze the behavior of our method in a broad range of situations, we here generate these prior probabilities according to three synthetic distributions (linear, exponential and bi-uniform, which are depicted in Fig. 3). Inverse transform sampling then allows to study how the performances are related to the quality of the initial set of predictions both in terms of accuracy and distribution.

As an example, the results of the parameter tuning procedure are depicted in Fig. 4 for an arbitrary binary attribute.

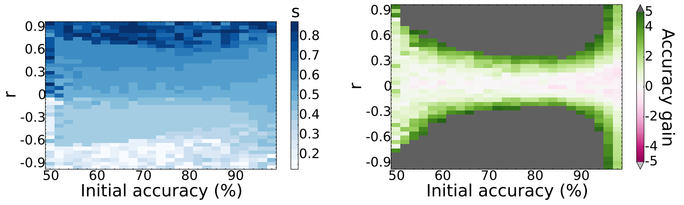
[1] For such ordered multi-valued attributes, a numeric coefficient is defined.

[2] According to McPherson et al., this attribute is among the most homophilic ones [10].



(a) (b) (c)

Fig. 3. (a) Linear, (b) exponential and (c) bi-uniform distributions of true class prior probabilities, leading to an initial accuracy of β. c_i denotes the true class of user i.



(a) (b)

Fig. 4. (a) Optimal s parameter of the edge potential (2) (chosen in a grid with a step of 0.05) and (b) mean accuracy gain (in %) over 50 random networks with 200 nodes and a mean degree $k = 8$ each, as a function of the initial accuracy (in %) of the predictions and the assortativity coefficient for a binary attribute. The initial predictions are simulated with a linear distribution.

The best s value and the corresponding mean accuracy gain are provided as a function of the assortativity and the initial accuracy of the predictions (β). For each pair of β and r, the optimal s value is selected as the one maximizing the average accuracy over 50 random networks with 200 vertices containing as many nodes from each one of the two classes (the randomness covers the edge rewiring in the networks, the attribute assignations and the sampling of the prior probabilities). The prior probabilities are in this case simulated using the linear distribution.

It can be observed that the optimal s values are almost independent of β and hence the parametrization mainly depends on the assortativity coefficient. Also, the chosen s evolves in a consistent way as a function of r, increasing from smaller values for disassortative networks to higher values for assortative ones. Using these s values, Fig. 4b shows that the accuracy gain is almost always positive, except for some particular pairs of r and β, especially when the assortativity is within the range $[-0.1, 0.1]$. This observation is consistent as our PGM is designed to exploit the assortativity, which is absent if $r = 0$ (random mixing). We finally note that the variability of the accuracy gains on the random synthetic networks do not affect the choice of the optimal s (corresponding results not shown).

The latter analysis has also been conducted with the other prior distributions. Similar conclusions are drawn for the exponential one but much lower accuracy gains are obtained in the case of the bi-uniform one, for reasons detailed in section III-C3. The extension to the case of non-binary attributes is straightforward, possibly using the numeric assortativity coefficient (e.g. in the case of the age attribute).

III. Mobile phone networks

The validation task considered in this section consists in gender prediction using two undirected and weighted mobile phone networks from a developed European country, denoted by $\mathcal{G}_S$ and $\mathcal{G}_L$. The data analysis of this work is only conducted on $\mathcal{G}_L$, while the performances assessment is performed on $\mathcal{G}_S$. This allows to avoid overfitting the particular network $\mathcal{G}_L$.

Predicting gender is of great interest to assess a demographic structure. For instance, this information is required to study gender disparities in diverse countries, allowing to refine or even undermine the available reports using social networks such as Google+ [3], Twitter or a mobile phone network. Among social networks, mobile phone data currently raise the interest of the research community and practitioners, as they become more and more ubiquitous, while being freely accessible at massive scale, automatically collected in real-time and powerful indicators of people behaviors [23]. They also often consist in the most accessible type of population information in developing countries. A shortcoming to their use however is that they often lack even the most basic information about their carrier, such as the gender, age or socioeconomic status. Indeed, most of the mobile phone connections worldwide are prepaid, as well in developing as in developed countries. Although these connections provide fine-grained information about the mobile phone usage, they do not give access to basic demographics.

In this section, the data sets are introduced and some of their features highlighted, showing significant gender homophilies which will be exploited through the inference process. In this case, X_i is the individual metadata of user i and Y_i is the random variable for her gender, defined on the alphabet $\mathcal{Y}= \{F, M\}$ with F and M resp. for a female and male.

A. Data description

In both networks $\mathcal{G}_L$ and $\mathcal{G}_S$, each node refers to one individual and an undirected edge binds any pair of users who exchanged at least one phone call or text during a fixed time period. The gender is known for the majority of the users. The communication attributes (extracted from the CDRs) of any edge e are the number of texts (SMS), the number of calls (CALLS) and the total duration of the calls (CALL_DUR). Different functions of these edge attributes can be defined. For example, the sum of SMS and CALLS is denoted by S_AND_C and counts the number of contacts between two given persons, which is well-suited to define the strength of a contact [24].

Table I provides general features of both networks, as well as the values of the three edge attributes distinguished by gender. As indicated, the communication patterns differ between hetero- and homogeneous (M-M and F-F) contacts. This reflects the stronger relationships occurring within the couples. Indeed, for instance in $\mathcal{G}_L$, there are on average 6.4 and 9.7 contacts (calls and texts) respectively between any homo- and heterogeneous pairs during the observation period. The same behavior is observed for the number of texts or calls distinctly. The mobile phone use of each individual according to her gender is not analyzed in this study, since this kind of information is exploited to provide the individual predictions. As the gender is binary, its assortativity is defined by (4). In $\mathcal{G}_S$ and $\mathcal{G}_L$, a moderate gender assortative mixing is observed.

B. Observational analysis

Since the strength of the heterogeneous communications (in terms of number of texts and calls exchanged) tend to overcome the one of the homogeneous contacts, the weights of an edge might give clues on its likelihood to be rather hetero- or homogenous. The subset of the strongest edges might have a completely different assortativity than the whole

TABLE I. SOME FEATURES OF THE NETWORKS. IF "EDGE TYPE" IS OMITTED, THE CHARACTERISTIC CONCERNS THE WHOLE NETWORK. "HOMOGENEOUS" (HOMO.) AND "HETEROGENEOUS" (HETERO.) REFER TO THE GENDER OF THE PERSONS LINKED BY THE EDGES.

	Edge type	Network $\mathcal{G}_L$	Network $\mathcal{G}_S$
Covered time period		15 days	3 months
Number of nodes		160818	19779
Number of edges		390778	78441
r (for gender)		0.3	0.26
Prop. of homo. edges (%)		66.47	63.5
Prop. of male nodes (%)		56.38	53.44
Mean SMS	homo.	3.58	15
	hetero.	5.74	25.8
Mean CALLS	homo.	2.84	5.3
	hetero.	3.96	7.9
Mean CALL_DUR	homo.	13min 40s	16min
	hetero.	15min 20s	19min 20s

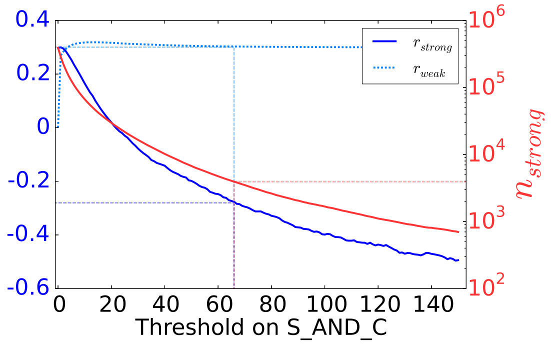


Fig. 5. Gender assortativity coefficient in $\mathcal{G}_L$ when the edges with S_AND_C values larger than some increasing thresholds are kept (strong part) or discarded (weak part). The red curve (right y-axis in log. scale) indicates the number of edges in the strong part, denoted by n_{strong}.

network. As the performance gains increase with the assortativity amplitude, identifying stronger (anti-)homophilic subgroups is of great interest. This section shows that r can indeed significantly change when considering subsets of the edges with specific attribute values.

We analyze the evolution of the assortativity coefficient when sub-graphs are constructed by only considering the edges with a scalar combination of their attributes above a threshold, the latter being progressively increased. For a given threshold and attribute combination, the strongest edges (according to this combination) constitute the strong part of the graph, while the weaker part refers to the rest. Several attribute combinations have been considered, including the attributes themselves. The most significant evolution of r is obtained using S_AND_C as a measure of link strength and is depicted in Fig. 5. The assortativity coefficient in the strong part (i.e. with edges such that S_AND_C is higher than the threshold on the x-axis) is denoted by r_{strong}, while r_{weak} is the one of the weak part. n_{strong} refers to the number of edges in the strong part. The dotted lines indicate the threshold and the corresponding r_{weak}, r_{strong} and n_{strong} values such that there are 1% of the edges in the strong part of $\mathcal{G}_L$. Using this partition, r_{weak} is still equal to about 0.3, but r_{strong} reaches -0.3 meaning that the strong part is rather anti-homophilic, as suggested by Table I. From a more general point of view, as the threshold on S_AND_C increases, r_{strong} decreases toward disassortative values. r_{weak} remains quite stable since the major part of the edges has small S_AND_C, as indicated by the evolution of n_{strong} in logarithmic scale.

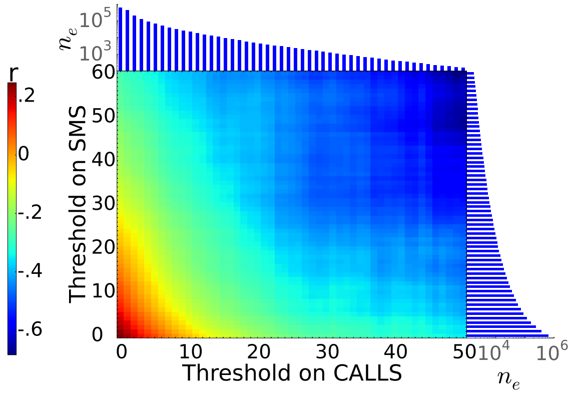


Fig. 6. Gender assortativity coefficient in $\mathcal{G}_L$ when only the edges with SMS and CALLS values larger than some increasing thresholds are preserved. The top (resp. right) histogram gives the number of edges (n_e) with CALLS (resp. SMS) larger than the corresponding value on the x-axis (resp. y-axis), on a log. scale.

A refinement of the previous analysis consists in combining two thresholds on two different edge attributes in order to observe how r_{strong} behaves. Fig. 6 depicts such an evolution using the SMS and CALLS attributes. The evolution of r_{weak} as a function of the two thresholds is negligible: it stays around 0.3, as in Fig. 5. Again, this figure highlights that the strongest edges are more disassortative. However, the strong part cannot be very large and have a significantly negative r in the mean time, as most of the edges have low SMS and CALLS values.

In section II-B, we show how to select a common s_{jk} parameter for all the edges of a network with a given r. On the other hand, the above analysis tells us that the assortativity significantly changes in distinct parts of a mobile phone network, decreasing as the strength of the links increase. We can exploit this information by using different s values in the strong and weak parts of the network, respectively denoted by s_{strong} and s_{weak}. However modeling s_{jk} as a step function is questionable. Indeed s_{jk} is the posterior probability for the edge (j, k) to be homophilic given its weights. Since this posterior probability is unlikely to abruptly change for some weight value, a smooth function should model it, with upper and lower plateau values corresponding to s_{weak} and s_{strong} respectively. Determining whether the edge (j, k) is hetero- or homophilic can be seen as a binary classification problem, with the edge weights as features. Hence, inspired by logistic regression, we model s_{jk} with a sigmoid function parametrized by a fixed linear combination S_AND_C of the edge weights,

$$s_{jk}(\text{S\_AND\_C}) = \frac{s_{weak} - s_{strong}}{1 + e^{G\cdot(\text{S\_AND\_C} - x_0)}} + s_{strong}, \quad (5)$$

where G and x_0 are two parameters to assign. Following the previous observations, the strong part of the network is defined as the set of the 1% strongest edges in terms of number of contacts. The plateaus s_{weak} and s_{strong} are tuned using the synthetic networks (with constant s_{jk} values) according to r_{weak} and r_{strong}. Let us further denote by x_U and x_L the x-values at which the sigmoid reaches $s_{strong} + 0.99\,(s_{weak} - s_{strong})$ and $s_{strong} + 0.01\,(s_{weak} - s_{strong})$. G and x_0 are fixed such that there are approximately 1% of the edges with a number of contacts lower (resp. higher) than x_U (resp. x_L). Fig. 7 presents the resulting smooth model of the s_{jk}'s for $\mathcal{G}_S$, which will be used later on to assess our methodology. The estimated

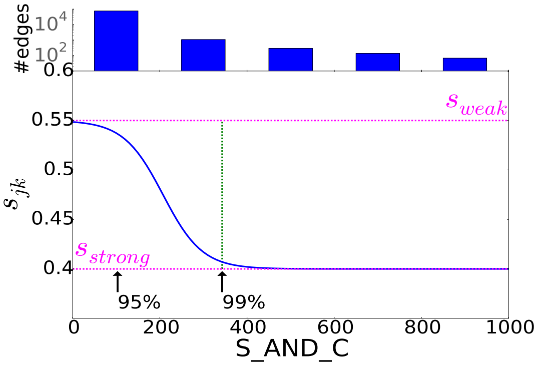


Fig. 7. Sigmoid function defining the s_{jk} values of the edge potential used for $\mathcal{G}_S$. The threshold on the number of contacts defining the edges as strong, indicated by the vertical green line, is determined to have 1% of strong edges. The top histogram gives the distribution of S_AND_C in $\mathcal{G}_S$ (in log. scale).

r_{weak} and r_{strong} in $\mathcal{G}_S$ lead us to choose $s_{weak} = 0.55$ and $s_{strong} = 0.4$. The percentages below the curve indicate quantiles of the S_AND_C distribution.

C. Results

The overall assortative character of $\mathcal{G}_L$, along with the observed differences between its strong and weak parts, indicates that the genders of the neighbors of an individual might be useful to predict its own gender. Our methodology is now tested on $\mathcal{G}_S$, while simulating individual prior predictions. The obtained performances are compared with the results of a baseline method, termed the *reaction-diffusion* algorithm [5].

1) Reaction-Diffusion algorithm: The reaction-diffusion (R-D) algorithm iteratively updates the predicted gender probabilities of each user by computing a weighted sum of its current gender probabilities and the ones of her neighbors, starting from a prior information as in our setting. R-D is equivalent to the consistency method [25], with a regularization parameter fixed to 0.5. The notation $p_i^t := \widehat{p}_{X_i}(M)$ denotes the estimated probability for user i to be a male at iteration t. These probability estimates are updated at each iteration for each user $i \in \mathcal{V}$ in the following way:

$$p_i^{t+1} = \frac{1}{2} \cdot \left(p_i^0 + \frac{1}{|\mathcal{N}(i)|} \cdot \left(\sum_{j \in \mathcal{N}(i)} p_j^t \right) \right) \qquad \forall i \in \mathcal{V} \quad (6)$$

until convergence, with $\mathcal{N}(i)$ the set of neighbors of user i.

2) Estimating the assortativity: The best edge potential for a given r can be estimated using the synthetic networks. However the assortativity of a given real network still needs to be estimated. To this end, we propose to collect the gender of an a priori fixed number of pairs of adjacent users in the considered graph $\mathcal{G}$, for example by carrying out a mobile phone survey, and then to use these edges to compute an estimate of r in $\mathcal{G}$. This procedure has been tested on $\mathcal{G}_L$, since it is larger than $\mathcal{G}_S$, which allows to consider more independent edge samplings. Fig. 8 presents the results. The assortativity estimates are roughly unbiased, while the variance of the estimator decreases toward 0.029, 0.022 and 0.014 when the gender of respectively $1,000$, $2,000$ and $5,000$ pairs of adjacent users are known. Hence knowing the gender of about 2k pairs of neighbors is sufficient to reliably estimate r. Indeed, an error of 0.05 on its estimation induces at worst an error of 0.05 on the s_{jk} value, as indicated by Fig. 4a. A sensitivity analysis (not presented) shows that such an error on the s_{jk} leads to negligible performance losses.

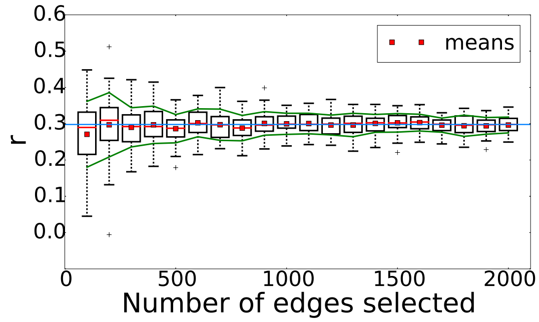


Fig. 8. Estimated r as a function of the number of randomly selected pairs of adjacent users with known gender in $\mathcal{G}_L$. For each number in abscissa, the edge selection is performed 50 times. The vertical distance between each mean estimated r (red squares) and the green lines gives the standard deviation of the estimation. The horizontal blue line indicates the true r in $\mathcal{G}_L$ (= 0.3).

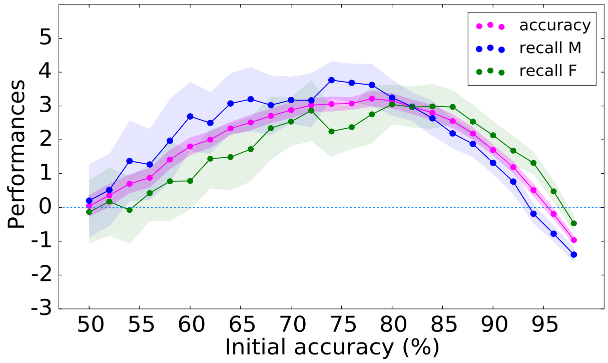


Fig. 9. Accuracy and recall gains when varying the initial accuracy β in $\mathcal{G}_S$, averaged over 50 random simulations of the first predictions using a linear distribution. The filled areas delimit intervals of one standard deviation around the mean gains.

It is noteworthy that using distinct edge potential parameters s_{strong} and s_{weak} in the strong and weak parts of the network requires to estimate r within these two parts. As the strong part tends to be significantly smaller, the estimation of r_{strong} in a real setting should be carefully performed. Meanwhile, the edges used to estimate r may be, for instance, used as a training set to provide individual gender predictions.

3) Performances: Fig. 9 shows the accuracy and recall gains of LBP using our model on $\mathcal{G}_S$, over simulated initial predictions using the linear distribution, for varying β. The well balanced recalls indicate that the weighting by the class prior in the node potential ψ is effective, avoiding to favor the dominant class (M) to the expense of the other one. Although optimal s_{jk} values are quite independent of β, the performances are not, with highest accuracy gains in the range $[70, 85]$%. This range covers the accuracies reached by state-of-the-art techniques aiming to predict gender using individual-level features [4], [5], [26]. Likewise, for an assortativity coefficient similar to the one of $\mathcal{G}_S$ (≈ 0.25), the accuracy gains on synthetic networks are significant when $\beta \in [0.62, 0.92]$. This result is natural, as near-perfect initial accuracies do not let many opportunities to improve the predictions, while almost random ones induce too rough node potentials.

TABLE II. MEAN PERFORMANCES ON $\mathcal{G}_S$ OF THE BASELINE UPDATE (R-D) SCHEME (6) AND OF LBP, FOR 50 DIFFERENT ASSIGNATIONS OF THE FIRST PREDICTIONS. THE THREE DEFINED DISTRIBUTIONS OF THE FIRST PREDICTIONS ARE CONSIDERED WITH AN INITIAL ACCURACY $\beta = 75\%$. Δ REFERS TO THE GAINS OVER THE PRIOR PREDICTIONS.

			LBP	R-D	LBP − R-D
Initial distribution	Linear	ΔAccuracy	3.17	2.01	**1.16**
		ΔRecall$_M$	3.78	2.73	**1.05**
		ΔRecall$_F$	2.47	1.18	**1.29**
	Exponential	ΔAccuracy	2.54	1.57	**0.97**
		ΔRecall$_M$	2.75	2.14	**0.61**
		ΔRecall$_F$	2.3	0.91	**1.39**
	Bi-uniform	ΔAccuracy	-0.52	-1.4	**0.88**
		ΔRecall$_M$	-1.12	-0.95	-0.17
		ΔRecall$_F$	0.16	-1.92	**2.08**

Table II gives the average accuracy and recall gains of both the baseline (R-D) and LBP in $\mathcal{G}_S$ over initial predictions with $\beta = 0.75$. Neither the baseline nor LBP allows to improve the predictions when the bi-uniform distribution is used. On the other hand, LBP increases the accuracy by more than 3 and 2.5% when the linear and exponential distributions are respectively chosen, overcoming the R-D algorithm. The results for the bi-uniform distribution can be explained by the fact that, in this case, only the sign of $\widehat{p}_{Y_i|X_i}(y_i = M|x_i) - 0.5$ brings information. On the other hand, its amplitude also matters when using the two other distributions, which would also be probably the case in a real setting. From this respect, the bi-uniform distribution might not be very realistic.

IV. RELATED WORK AND DISCUSSION

Different researches exploit mobile phone data for healthcare, epidemics containment or marketing study purposes [27]–[29]. In each context, the knowledge of the users' gender is of great importance. Most of the recent works on demographics prediction use classical machine learning algorithms on mobile phone data for predicting gender, age, income level or even personality [4], [16], [26], [30]. These algorithms rely on features defined for each user and reflecting their mobile phone usage at an individual scale, such as the recharge rate of their prepaid cards, spending speed, total call duration, etc. Some studies further refine such standard metadata by deriving diverse behavioral indicators [23], allowing to enhance the prediction capabilities of the models. All these studies are thus based on an "individual" part of the mobile phone data. For example, Felbo et al. and Sarraute et al. predict the gender with respectively 79.7% and 77.1%[3] accuracy, either by harnessing their temporal information using deep learning or by using linear SVM and logistic regression [5], [26]. Other works tackle the gender prediction problem in a similar way using different kinds of data sets, such as Twitter or LinkedIn data, the first name of a person or even chat texts [31]–[33]. Closer to our work, Al Zamal et al. exploit the homophily in a Twitter network to predict the users' gender, age, and political affiliation [2], by analyzing how the knowledge of the data from some immediate friends of a given user can improve the prediction quality. This question is studied in a usual machine learning framework: feature vectors are defined for each user, either augmented with data from her neighbors or not. Considering the neighbors' information in the feature vectors allows to improve the accuracy from 3 to 5% for the age and political affiliation prediction, whereas including the immediate neighbors' features does not improve the gender predictions.

Since the former approaches do not take the global network structure into account, they could further benefit from its assortativity. Fewer studies consequently consider the network directly to predict demographics, either by assuming that the labels of some individuals in the graph are known and the remaining ones are missing [17] or by adopting a two-step approach, first computing uncertain predictions using the individual part of the data, in the same spirit as the former researches, and then improving them using the network [5]. Dong et al. introduce a double dependent-variable factor graph model in order to jointly predict the users' age and gender by taking profit of the links between these two demographic attributes in a network [17]. Knowing 50% of the labels, the remaining unknown genders are predicted with up to 80% accuracy. However, as they do not quantify the assortativity of their network, these performances are not easy to compare to our study. Our results may nevertheless qualitatively explain the success of their approach, at least partly. Combining age and gender implicitly delineates in an automated manner some rather (anti-)homophilic sub-graphs, as illustrated by their data analysis. As highlighted by our work, this definition of strong and weak network parts with accentuated (anti-)homophilies improves the inference performances. The latter observation is essential, as several studies mention that gender assortativity is generally rather weak [2], [10] and thus not sufficient to infer the gender. For instance, the reaction-diffusion algorithm introduced by Sarraute et al. is used to infer the age group of some users, but not their gender [5]. Their network indeed bears a strong age homophily. When 70% of the known age labels (a fraction of all the users) are propagated through the network to infer the 30% remaining nodes, the age group among four categories is predicted with 43.4% accuracy.

Compared to the two aforementioned studies, in addition to exploiting the global network topology, our study makes use of an objective measure of the assortativity to provide guarantees about the performances generalization. This quantitative measure of the network homophilies is typically not provided by graphical representations. It allows us to describe to which extent the sole network information improves individual demographics prediction, as a function of the assortativity. The proposed methodology also easily permits to take profit of some known labels, as well as first individual predictions performed using individual data. Finally, the model can benefit from assortativity variations in different sub-graphs, highlighted by the edge weights. By modeling the statistical dependencies between adjacent labels, it can favor heterogeneous as well as homogeneous contacts (by opposition to so-called consistency methods only favoring homogeneity [25], [34]).

V. CONCLUSION

This work shows how assortativity can be exploited to improve individual demographics prediction in social networks. To this aim, a general approach is introduced, 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

[3] But with only 25% coverage.

assortativity amplitude and (2) the quality of the prior individual predictions (poor prior information is misleading, while excellent one does not leave much room for improvement). The graph simulations allow to tune the model parameters. Our method is further validated on a real-world mobile phone network and the model is refined to predict gender, exploiting both weak, homophilic and strong, anti-homophilic links. In this particular case, the approach allows to improve individual-based gender predictions by up to 3%.

The analysis performed on synthetic networks illustrates that a strong assortativity can be easily exploited through our methodology. Also, an almost randomly mixed network might still be composed of several parts which are, if considered in isolation, assortative and disassortative. Thus even in the latter configuration, the network topology might still be useful. As a further work, the generalization of the proposed methodology to multivariate predictions would be of great interest. The model could then benefit from the relationships between the target variables, and automatically make use of sub-networks presenting more pronounced homophilies.

Acknowledgment

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

References

[1] A. L. Traud, P. J. Mucha, and M. A. Porter, "Social structure of facebook networks," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 16, pp. 4165–4180, 2012.

[2] F. Al Zamal, W. Liu, and D. Ruths, "Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors.," *ICWSM*, vol. 270, 2012.

[3] G. Magno and I. Weber, "International gender differences and gaps in online social networks," in *International Conference on Social Informatics*, pp. 121–138, Springer, 2014.

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

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

[6] Y.-A. de Montjoye, J. Kendall, and C. F. Kerry, "Enabling humanitarian use of mobile phone data," *Brookings Center for Tech. Innovation*, 2014.

[7] J. Blumenstock, G. Cadamuro, and R. On, "Predicting poverty and wealth from mobile phone metadata," *Science*, vol. 350, no. 6264, pp. 1073–1076, 2015.

[8] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[9] S. Aral, L. Muchnik, and A. Sundararajan, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," *Proceedings of the National Academy of Sciences*, vol. 106, no. 51, pp. 21544–21549, 2009.

[10] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, pp. 415–444, 2001.

[11] A. Madan, S. T. Moturu, D. Lazer, and A. S. Pentland, "Social sensing: obesity, unhealthy eating and exercise in face-to-face networks," in *Wireless Health 2010*, pp. 104–110, ACM, 2010.

[12] Y. Wang, H. Zang, and M. Faloutsos, "Inferring cellular user demographic information using homophily on call graphs," in *INFOCOM, 2013 Proceedings IEEE*, pp. 3363–3368, IEEE, 2013.

[13] J. A. Smith, M. McPherson, and L. Smith-Lovin, "Social distance in the united states: Sex, race, religion, age, and education homophily among confidants, 1985 to 2004," *American Sociological Review*, vol. 79, no. 3, pp. 432–456, 2014.

[14] M. E. Newman, "Mixing patterns in networks," *Physical Review E*, vol. 67, no. 2, p. 026126, 2003.

[15] J. N. Rosenquist, J. Murabito, J. H. Fowler, and N. A. Christakis, "The spread of alcohol consumption behavior in a large social network," *Annals of internal medicine*, vol. 152, no. 7, pp. 426–433, 2010.

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

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

[18] Y. Weiss and W. T. Freeman, "On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs," *Information Theory, IEEE Transactions on*, vol. 47, no. 2, pp. 736–744, 2001.

[19] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008. DOI: 10.1561/2200000001.

[20] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[21] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics," *Reviews of modern physics*, vol. 81, no. 2, p. 591, 2009.

[22] M. E. Newman, "Models of the small world," *Journal of Statistical Physics*, vol. 101, no. 3-4, pp. 819–841, 2000.

[23] Y.-A. de Montjoye, L. Rocher, and A. S. Pentland, "bandicoot: a python toolbox for mobile phone metadata," *Journal of Machine Learning Research*, vol. 17, no. 175, pp. 1–5, 2016.

[24] V. Palchykov, K. Kaski, J. Kertész, A.-L. Barabási, and R. I. Dunbar, "Sex differences in intimate relationships," *Scientific rep.*, vol. 2, 2012.

[25] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *NIPS*, vol. 16, pp. 321–328, 2003.

[26] B. Felbo, P. Sundsøy, A. Pentland, S. Lehmann, and Y.-A. de Montjoye, "Using deep learning to predict demographics from mobile phone metadata," *arXiv preprint arXiv:1511.06660*, 2015.

[27] L. Bengtsson, X. Lu, A. Thorson, R. Garfield, and J. Von Schreeb, "Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in haiti," *PLoS Med*, vol. 8, no. 8, p. e1001083, 2011.

[28] A. J. Tatem, Y. Qiu, D. L. Smith, O. Sabot, A. S. Ali, B. Moonen, *et al.*, "The use of mobile phone data for the estimation of the travel patterns and imported plasmodium falciparum rates among zanzibar residents," *Malar J*, vol. 8, no. 1, pp. 10–1186, 2009.

[29] R. Montoliu and D. Gatica-Perez, "Discovering human places of interest from multimodal mobile phone data," in *Proceedings of the 9th Internat. Conf. on Mobile and Ubiquitous Multimedia*, p. 12, ACM, 2010.

[30] Y.-A. de Montjoye, J. Quoidbach, F. Robic, and A. S. Pentland, "Predicting personality using novel mobile phone-based metrics," in *Social computing, behavioral-cultural modeling and prediction*, pp. 48–55, Springer, 2013.

[31] W. Liu and D. Ruths, "What's in a name? using first names as features for gender inference in twitter.," in *AAAI Spring Symposium: Analyzing Microtext*, vol. 13, p. 01, 2013.

[32] C. Peersman, W. Daelemans, and L. Van Vaerenbergh, "Predicting age and gender in online social networks," in *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pp. 37–44, ACM, 2011.

[33] A. Kokkos and T. Tzouramanis, "A robust gender inference model for online social networks and its application to linkedin and twitter," *First Monday*, vol. 19, no. 9, 2014.

[34] D. Zhou, J. Huang, and B. Schölkopf, "Learning from labeled and unlabeled data on a directed graph," in *Proceedings of the 22nd international conference on Machine learning*, pp. 1036–1043, ACM, 2005.