Improving attribute prediction through Network-Augmented Attribute Prediction

Aaron Zimmerman* Tyler McCormick*‡ Ali Shojaie†* Hedwig Lee‡
*Dept. of Statistics - †Dept. of Biostatistics - ‡Dept. of Sociology
University of Washington
Seattle, WA 98195
azimmer@uw.edu

***Abstract*—We propose a method for predicting individuals' attributes based on partially observed social network data. The Network-Augmented Attribute Prediction (NAAP) procedure uses observed individuals' nodal and network attributes to infer unobserved network connections and then uses these predicted network connections to predict unobserved characteristics of individuals. We demonstrate that inclusion of such inferred network attributes can increase the accuracy of predictive modeling using data from a household survey of villages in Karnataka, a state in southwestern India.**

***Index Terms*—Attribute prediction, Link prediction, Network-augmented Attribute Prediction (NAAP)**

I. Introduction

A large body of research has begun to establish the importance of social network information in predicting attributes and behaviors. However, in many cases, it is not possible to obtain complete network information for the sampling population of interest, and this is especially likely when the data have been collected on hard to reach populations (e.g. homeless, sex-workers, people below the poverty line) [1]. Furthermore, missing links can have an effect on predicted network level statistics [2] and in the case of both missing links and missing attributes, careful attention needs to be paid to all steps of imputation and error propagation.

In response to the missing network data, much research has been developed to deal with the classical link predication problem in which we observe one instance of a social network at time t and aim to predict links in the network at some future time t' [3]. Our setting is slightly different in that we assume a *partially observed static network*. Given some subset of true links, we need to both impute the missing links and the missing attributes of interest, assuming only missingness and no temporal network dynamics. In many settings, a static network is capable of capturing most of the relevant network features. For example, epidemiological studies often assume that a static network will capture the relevant information for rapidly spreading diseases [4]. While there has been some other work using both node attributes and known links to infer missing links in a static network ([5], [6], [7]), less attention has been paid to joint link and attribute prediction in the static network setting.

We propose a network augmentation procedure to infer unknown social network links between individuals in the fixed network while also predicting missing node attributes of interest. The proposed method uses partially known network information, as well as individual characteristics of actors, with the premise that the inclusion of the predicted links improves the accuracy for predicting node attributes.

We illustrate this procedure by predicting financial savings behavior in a developing country using incomplete network data. In poorer countries, a bad financial year can easily lead to housing instability, malnutrition, or disease. To avoid the risks of having bad years, an individual needs to find a way to smooth her/his consumption between income fluctuations. It is known that precautionary savings can help smooth consumption [8], and in order to assess certain types of risk (e.g., an individual's uninsured risk) it is important to know whether or not an individual has a savings account to aid in the process of risk reducing savings. Unfortunately, such information is often unavailable, and prediction methods are needed to determine whether an individual is likely to have a savings account.

II. Data

The Social Networks and Microfinance dataset was collected in rural southern Karnataka, India. The dataset contains network information at both the individual and household level on 75 villages in the region, as well as questionnaire data at the individual and household levels. Prior investigation by Bharatha Swamukti Samsthe (BSS, a microfinance institution) identified these 75 villages as locations to set up microfinance infrastructure. Shortly after the survey, BSS began operating in some of the villages and in the next two years it spread to over half of them. Data pertaining to involvement in the BSS microfinance programs were then linked to collected household level data.

The networks are constructed based on specific (often finance oriented) activities between households or individuals within the community. For example, individuals were asked questions like from whom they borrow kerosene or rice, or who they go to for medical advice. Each of these questions was used to create a different network within the villages. There were 12 such questions, and each of the 75 villages has

24 different networks, 12 at the individual level and 12 at the household level. The household level demographic questions include questions about the home (e. g. latrine type, roof type, number of rooms), while the individual level demographic questions pertain to individuals within the home (such as age, gender, or caste). The household questions were administered to all households within a village while the network questions and individual level questions were given to just under half of the households. More information and the data may be found here: <http://economics.mit.edu/faculty/eduflo/social>.

For the remainder of the paper we deal with individual level link and attribute prediction and leave inclusion of household level data for future work. Based on initial data exploration, only a subset of individual-level node attributes were used and they were chosen to be the respondents age (integer), gender (binary), religion (categorical: Hinduism, Islam, Christianity), caste (categorical: Scheduled Caste, Scheduled Tribe, Other Backward Caste, General), education level (categorical with 16 categories), and rationcard (categorical: have a ration card, don't have one, it's missing). As mentioned in the introduction, we predict a categorical node attribute indicating whether or not an individual has a savings account ('savings') and we predict links for the social network indicating if individuals go to temple together ('templecompany').

III. Methods

We present a Network-augmented Attribute Prediction (NAAP) procedure, which *infers* individuals' network attributes based on existing training data and their personal attributes. These inferred network features are then used to predict unobserved attributes of individuals. An approximate error propagation scheme, based on measurement error models from the statistics literature [9], can be utilized to ensure that uncertainty in the attribute prediction stage accounts for confidence in the network inference stage.

Our proposed NAAP procedure has two stages: (i) network inference and (ii) attribute prediction. First, consider a subset $\mathcal{G}_0$ of graph $\mathcal{G}$ consisting of vertices V_0. Let $Y_{ij} = Y_{ji}$ indicate presence/absence of association among nodes i and j. We define $\mathcal{N}_{V_0}$ as neighbors of V_0, among unknown connections. Our goal in network inference is to estimate Y_{jk} for $j, k \in \mathcal{N}_{V_0}$ using structure of $\mathcal{G}_0$, as well as attributes X_j, X_k for $j, k \in \mathcal{N}_{V_0}$. The new graph, $\tilde{\mathcal{G}}$ includes available and predicted links among $V_0 \cup \mathcal{N}_{V_0}$. We exploit the penalized multi-attribute exponential random graph model (ERGM) [10], and let $Z_{l,i} = Z_{l,i}(Y), l = 1, \ldots, q$ be graph-derived attributes for node i. Attributes could be node specific (e.g. degree), or global (e.g. number of triangles in the graph). We can also incorporate non-graph based attributes, $W_{k,i} = W_{k,i}(X), k = 1, \ldots, p$ which may include characteristics such as age or gender. For this report, we focus on prediction within one network, but the dataset has 12 different networks based on 12 different social interactions. Missingness in multiple network structures can also be dealt with by using the framework in multi-attribute ERGM.

Consider the ERGM model [11] where $Y_{i,j}$ and $Y_{i',j'}$ are independent, conditioned on the values of Z_l, W_k. We use penalized pseudo likelihood estimation, which can be written as a logistic regression:

$$\log\left[\frac{\mathbb{P}_{\beta,\theta}(Y_{ij} = m \,|\, Z_{ij} = z_{ij}, W_i = w_i, W_j = w_j)}{\mathbb{P}_{\beta,\theta}(Y_{ij} = 0 \,|\, Z_{ij} = z_{ij}, W_i = w_i, W_j = w_j)}\right] = \theta^T z + \beta^T w + J(\theta, \beta) \tag{1}$$

where $J(\theta, \beta)$ is a regularization penalty, $m = 1, \ldots, M$ denote different edge types (in the simple case, $M = 1$). Although we may use fully observed node attributes, we cannot use any partially observed attributes which we plan to model and predict in the second stage.

We now move to the second stage of our method, attribute prediction. Begin by defining a binary attribute r_i where we define $P(r_i = 0 | X_i, \mathcal{G})$ as our measure of risk for individual i (e.g. to not have a savings account which would help smooth consumption risk). We note also that our model and computations are easily generalized to any exponential family likelihood and thus easily accommodates multi-class attributes.

Denote $f(\tilde{\mathcal{G}}_0)$ as the set of features computed based on the inferred graph. In our example, these features consist of, for example, the number of connections an individual has with other individuals in one of the 12 different known or predicted networks. Given $f(\tilde{\mathcal{G}}_0)$ we construct a logistic regression likelihood for outcome $r_i \sim$ Bernoulli(p_i):

$$\log\left(\frac{p_i}{1 - p_i}\right) = \mu + \beta f(\tilde{\mathcal{G}}_0) + \gamma g(\mathbf{X}) \tag{2}$$

where $\mathbf{X}$ is a set of non-network features that may overlap with the features used in the previous stage. We fit the model using a Bayesian approach where $\mu, \beta,$ and$, \gamma$ have independent Gaussian priors with mean zero and variances σ^2_μ, σ^2_β, and σ^2_γ, respectively.

Since our outcome of interest in the dataset is binary (as are the links we need to predict) we use the training data in both stages to cross-validate a cut-off probability for predicting the outcomes. While a standard cutoff of 0.5 maybe used as a starting value, in the face of an unbalanced design we choose the cutoff probability to minimize an overall measure of the misclassification error.

IV. Model Fitting

The ultimate goal is to predict node attributes at the individual level. In the first stage of model fitting, we perform link prediction using both known node and network attributes. In the second stage of model fitting we use known node attributes as well as a set of combined known and predicted network attributes to predict unknown node attributes.

In the first stage, we build up a set of dyad level attributes from the known node attributes. An example of a dyad level attribute might be an indicator variable representing whether or not nodes N_i and N_j share the same categorical realization for a specific nodal categorical covariate, or in the case of 'age,' it could be the continuous measure of difference in age between the two nodes. Using the dyad-level node attributes and known network information we model each link in the villages as our outcomes. Using a 50% testing-training split by village, the model was fit using binomial regression with a logit link and a standard normal prior on the randomly selected training villages. The model was then applied to the villages in the testing data, and the accuracy of the model was measured using the F1 score: $F1 = 2 \times \frac{precision \cdot recall}{precision + recall}$. The F1 score is an overall measure of classification accuracy and is the harmonic mean between precision and recall [12]. We predict the links for the "templecompany" network (who you go to temple with), a significant variable with the largest impact among network attributes from the model using both fully known demographic and known network attributes (in particular the saving status was not included in the link prediction model). At this point, the training half of the data were removed, and the the second stage of modeling was performed on the testing data using the predicted "templecompany" links.

In the second stage, we model the node attribute of interest, "savings", as a function of nodal covariates. To convert the known and predicted networks to nodal values we use the degree of node N_i in a specific network (e.g. the predicted 'templecompany' network) as the network covariates. In this way we capture information about how the number of connections in a particular activity affect an individual's prevalence for a savings account. The attribute prediction model was fit three times: (1) using only the network attributes as predictors, (2) using only demographic attributes, (3) using both network and demographic attributes.

To perform model fitting for attribute prediction, the villages were again split into testing and training groups (half of the data was removed by this point as training data for the link prediction). As a result, the link prediction model had only half the amount of data to train and perform accuracy checks, compared to the model with all known network attributes. Penalized logistic regression with an appropriately chosen tuning parameter was used to fit the model. By repeating this process using different randomly selected training and testing data subsets, we were able to obtain an initial representation of the uncertainty in the prediction accuracy which is shown by the range of accuracy values and confidence intervals in the boxplots in Fig. 1.

V. Results

In this section we describe the performance of our model for inferring attributes in the Karnataka data. To examine the performance of our model, we consider both fully known network information as well as augmented network information using predicted links. The goal of our experiments is to see if network attributes, even when inferred through the modeling process, are useful in predicting "savings."

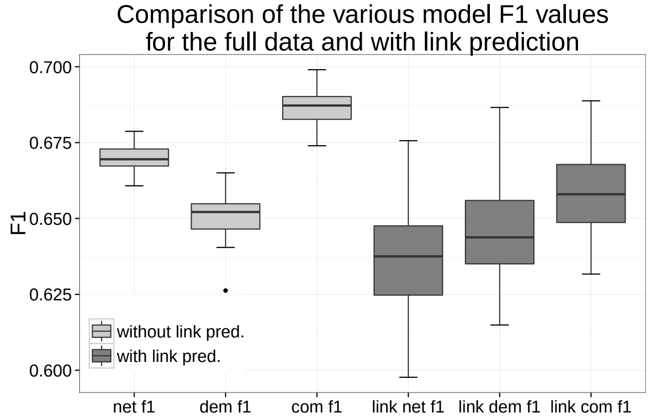

Fig. 1. Comparison of model performance using an F1 measure for models fit using known networks and using predicted networks. Boxplots for known network attributes are on the left side of the plot while models using link prediction are on the right and denoted with 'link.' 'Net' stands for models using only network features for attribute prediction, 'dem' stands for models using only other known node attributes for attribute prediction, and 'com' stands for the combined model using both network and known node attributes for missing attribute prediction.

We compare three different models:

1) A model using only a chosen subset of the demographic attributes: age, race, gender, caste, ration card (binary, possession of ration card)
2) A model using only network attribute covariates
3) A model using both network and demographic attributes.

The left-hand side of Fig. 1 demonstrates that with fully known network information, the inclusion of network information can significantly improve the accuracy of predicting individual attributes. More so, in this case, the network information by itself does a better job of attribute prediction when compared to using only other known node attributes to predict the missing 'savings' node attribute.

The right-hand side of Fig. 1 demonstrates that in the presence of missing links, attribute prediction accuracy can still be increased by using both predicted and known network information. In this example, the most informative network, "templecompany," was predicted (after using half the network data for training) among the testing data subset in stage I and used in conjunction with less informative networks to add to the predictive accuracy of the node attribute of interest in stage II.

The range of F1 values and the 95% confidence intervals in the right-hand side of Fig. 1 are much larger than the results from the models which used the fully observed network information. This is in part due to increased uncertainty caused by the predicted links but is also due to the fact that the we modeled the savings outcome in the predicted link setting using only half of the full dataset (we performed link prediction using 50% training data which was then removed). Note that the difference between methods with demographic information only can be attributed to the smaller sample size of the training data in the results of the right-hand side. Nonetheless, comparing the F1 values among each set of results clearly indicates the advantage of incorporating network attributes, even when these attributes are derived from partially observed networks. The improved performance of the proposed augmented procedure highlights the potential benefits of estimating the incomplete network information based on available information.

VI. Related Work

As mentioned earlier, there has been significant interest in developing methods for link prediction in the case of temporally dynamic networks. Even in this setting, there has been relatively little work done on predicting both links and nodal attributes simultaneously. Recent work by Gong et al. [13] extends the Social-Attribute Network (SAN) developed by Yin et al. [14] which performs inference on an attribute-augmented network. The proposed method adds extra nodes to the social network to represent attributes and reduces the problem of predicting node attributes to the problem of predicting attribute links. The method is clever and exciting, especially in terms of accuracy and scalability, but it leaves questions of uncertainty assessment unanswered. While prediction is important, modeling the underlying mechanism and determining the uncertainty of the predictions and of the covariate effects are of the utmost interest to modelers.

Other work on link prediction ([5], [6], [7]) in the static network setting has been developed but has not generally been extended to settings where both links and nodal attributes need to be predicted.

VII. Discussion

In this paper, we develop a method for predicting individual attributes using estimated network relations. This method uses a two-stage model-fitting procedure where missing network links are first inferred based on actor covariates and observed links. We then predict attributes based on these inferred network relations. While predicting links naturally results in some loss of predictive accuracy for individuals' attributes, using NAAP-predicted links in conjunction with an individual's attributes generally improves performance over a model with only demographics and in some cases matches performance using only known attributes. In our application to data from villages in rural southern India, we inferred all of the edges for the individuals with unknown attributes (corresponding to inferring an entire village). Our method would also work in situations where a subset of an individual's network is observed. In such cases, we would first predict each actor's unobserved links and then use this information to predict attributes. This situation arises in many survey-based network data collection work where respondents are asked to list a subset of their network ties (their "five best friends" for example).

Our results indicate that the primary cost associated with missing links is increased variability in out-of-sample attribute predictions, as demonstrated in Fig. 1. This situation arises since, when inferring links, we have uncertainty about both the network statistics used for attribute prediction and about the unobserved attributes. The Bayesian framework presented here for attribute prediction is a first attempt to quantify this uncertainty, though additional work in this area could be a fruitful area for future work. In particular, additional experiments are needed to quantify the increased level of uncertainty in predictive performance, as well as the improvement through the augmented link prediction procedure as the amount of available network information varies.

VIII. Acknowledgements

This material is based upon work supported by, or in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under contract/grant number W911NF1210379.

References

[1] H. R. Bernard, T. Hallett, A. Iovita, E. C. Johnsen, R. Lyerla, C. McCarty, M. Mahy, M. Salganik, T. Saliuk, O. Scutelniciuc, G. Shelley, P. Sirinirund, S. Weir and D. Stroup, "Counting hard-to-count populations: the network scale-up method for public health", *Sexually Transmitted Infections*, vol. 86, pp. ii11-ii15, 2010.

[2] G. Kossinets, "Effects of missing data in social networks", *Social Network*s, vol. 28, pp. 247-268, 2006.

[3] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks", In CIKM, pp. 556-559, 2003.

[4] Lloyd, A.L. and Valeika, S. "Network models in epidemiology: an overview", In: Complex Population Dynamics: Nonlinear Modeling in Ecology, Epidemiology and Genetics, B. Blasius, J. Kurths and L. Stone (eds.), World Scientific, 2007.

[5] A. Popescul and L. Ungar, "Statistical relational learning for link prediction", In Workshop on Learning Statistical Models from Relational Data at the International Joint Conference on Artificial Intelligence, 2003.

[6] B. Taskar, M-F Wong, P. Abbeel, and D. Koller. "Link prediction in relational data", In *N*eural Information Processing Systems, vol. 15. 2003.

[7] D. S. Goldberg and F. P. Roth, "Assessing experimentally derived interactions in a small world", In Proceedings of the National Academy of Sciences USA, vol. 100, pp.4372-4376, April 2003.

[8] A. Deaton, "Saving in developing countries: theory and review", In Proceedings of the World Bank Annual Conference on Development Economics, World Bank, Washington DC, 1989.

[9] P. Gustafson, "Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments," vol. 13. CRC Press, 2004.

[10] A. Shojaie, "A penalized multi-attribute exponential random graph model for link prediction in biological networks", unpublished.

[11] O. Frank and D. Strauss, "Markov Graphs", *JASA*, vol 395, pp. 832-842, 1986.

[12] C. J. van Rijsbergen, "Information Retrieval", Butterworth-Heinemann, London, 2nd edition, 1979.

[13] N. Z. Gong, A. Talwalkar, L. Mackey, L. Huang, E. Chul R. Shin, E. Stefanov and D. Song,“Jointly predicting links and inferring attributes using a social-attribute network (SAN)”, arXiv preprint arXiv:1112.3265v9, 2012.

[14] Z. Yin, M. Gupta, T. Weninger and J. Han, “A unifed framework for link recommendation using random walks”, In *A*SONAM, 2010.