CS224W final report: Node Classification in Social Networks Using Semi-supervised Learning

Yatong Chen [SuID:yatong] *

December 9, 2018

Code can be found here: *https://github.com/YatongChen/decoupled-smoothing-on-graphs-code-.git*

1 Introduction

Graph-based learning describes a broad class of problems where response values are observed on a subset of the nodes of a graph, and the learning objective is to infer responses for the unlabeled nodes. Inference methods for graph-based learning nearly unanimously derive their success from an assumption that connected nodes are correlated in their responses, akin to the social phenomenon of homophily whereby bird of a feather flock together. Many variations on models derived from this assumption have been studied and applied with great success. While presented as graph-based methods, the graphs that underlie the typical applications of these methods are often synthetic in nature. For example, they may be derived from high-dimensional text or image data. These typical applications begin with a semi-supervised learning problem studying high-dimensional data points $x_i \in R^D$ associated with response values $y_i \in R$ (such as images x_i associated with quality scores y_i) and then induce a graph between the data points by taking a k-nearest neighbor graph in the space to obtain a sparse similarity graph. Despite the synthetic nature of these graphs, graph-based learning methods have been highly effective for solving machine learning problems. Graph smoothing methods are an extremely popular family of approaches for semi-supervised learning. The choice of graph used to represent relationships in these learning problems is often a more important decision than the particular algorithm or loss function used, yet this choice has not been well-studied in the literature. In this work we demonstrate that for social networks, the basic friendship graph may often not be the appropriate graph for the problem of predicting node attributes. More specifically, standard graph smoothing is designed to harness the social phenomenon of homophily whereby individuals are similar to "the company they keep." We present a *decoupled* approach to graph smoothing that decouples notions of "identity" and "preference," resulting in an alternative social phenomenon of monophily whereby individuals are similar to "the company they're kept in." Our model results in a rigorous extension of the GMRF models that underlie graph smoothing, interpretable as smoothing on an appropriate auxiliary graph of weighted or unweighted two-hop relationships.

*This is a joint work with Alex Chin, Kristen M Altenburger and Johan Ugander.

2 Problem Statement

We consider the general problem of learning from labeled and unlabeled data. Given a point set $X = x_1 \cdots x_l; x_{l+1} \cdots x_n$ and a label set $L = \{1, 2 \cdots c\}$, the first l points have labels $\{y_1, \cdots y_l\} \in L$ and the remaining points are unlabeled. The goal is to predict the labels of the unlabeled points. The performance of an algorithm is measured by the error rate on these unlabeled points only. Here in our work, we will focus on predicting the gender for individuals in the network, which means that we will have two different components: male and female in the label set.

3 Related Work

In the project proposal, we discussed four papers in the field of semi-supervised learning and node or link labeling problem in this report, spanning a time frame of one decades. The first paper we considered was authored by Zhu, Ghahramani and Lafferty(ZGL) in 2003, entitled Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions [1]. It is one of the first few works to use a random walk based method for node labeling problem. Before their pioneering work, most of the node labeling methods are in the framework of iterative method [6]. We then consider the paper by Zhou, Bousquet, Lal, Weston and Scholkopf, entitled Learning with Local and Global Consistency in 2004 [2]. Different from ZGL, the keynote of their methods is to let every point iteratively spread its label information to its neighbors until a global stable state is achieved, which can help achieve a better overall prediction result. The last paper we discussed was about Monophily phenomenon is social network by Altenburger and Ugander in 2017, which introduced the concept of Monophily. The author of the paper observed a fundamental difference between similarities with the company you keep and the company youre kept in in social networks. That work found that the two-hop similarities implied by the latter can exist in the complete absence of any one-hop similarities, which served as the fundamental inspiration of the concept of decouple which is mentioned below.

4 Description of Dataset

We analyzed populations of networks from the FB100 network dataset. FB100 consists of online friendship networks from Facebook collected in September 2005 from 100 US colleges primarily consisting of college-aged individuals. Traud et al. provide extensive documentation of the descriptive statistics of these networks. We will exclude Wellesley College, Smith College and Simmons College from our analysis, all of which are single-sex institutions with $\geq 98\%$ female nodes in the original network datasets. For all networks, we restricted the analysis to only nodes that disclose their attributes, completely removing those with missing labels. We also restricted the analyses to nodes in the largest (weakly) connected component to benchmark against classification methods that assume a connected graph.

5 Graph smoothing preliminaries

In this section we review the standard formulations of graph smoothing, the semi-supervised learning problem of [1], which we refer to here simply as smoothing. We review the closed

form solutions. Later we will talk about the new concept of decoupled smoothing graph and will provide its closed form solution.

5.1 Smoothing

The standard formulation of graph smoothing, proposed in [1], is to solve the optimization problem

$$\min_{\theta} \sum_{(i,j)\in E} A_{ij}(\theta_i - \theta_j)^2, \qquad \text{subject to } \theta|_{V_0} = \theta_0. \tag{1}$$

The loss function in Equation (1) is $\theta^\top L\theta$, where $L = D - A$ is the graph Laplacian.

If we define the transition matrix $P = D^{-1}A$ and identify blocks of P according to the labeled nodes V_0 and unlabeled nodes V_1, the closed-form solution to Equation (1) for the unlabeled nodes is then:

$$\hat{\theta}_1 = (I - P_{11})^{-1}P_{10}\theta_0, \text{ where } P = \begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix}. \tag{2}$$

This solution has a Bayesian interpretation [3]. Suppose we place a Gaussian Markov Random Field (GMRF) on the node set by placing a prior $\theta \sim N(0, \tau^2(D - \gamma A)^{-1})$ on θ. This prior is the *conditional autoregressive* (CAR) model popular in the spatial statistics literature, and has the property that θ_i conditional on the other values of θ follows the distribution

$$\theta_i \mid (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n) \sim N\left(\frac{\gamma}{d_i}\sum_{j\in_i} \theta_j, \frac{\tau^2}{d_i}\right). \tag{3}$$

Under this GMRF prior, the Bayes estimator conditional on having observed the labels θ_i, $i \in V_0$, is the solution to Equation (1), when $\gamma \to 1$. The parameter $\gamma < 1$ is a correlation parameter that is necessary for the distribution to be non-degenerate. In practice it is common to add a small ridge to the diagonal of the Laplacian when solving Equation (1) for numerical stability, which achieves a similar purpose.

5.2 Decoupled graph smoothing

In this work we propose *decoupling* the true parameter of interest θ_i from a *target parameter* that is close to the true parameters of the neighbors of i. We now study a model that gives rise to such a decoupling.

Suppose we have an asymmetric weight matrix W, and denote the row sums by $z_i = \sum_j W_{ij}$ and the column sums by $z'_j = \sum_i W_{ij}$. Consider the Gaussian Markov random field model

$$\theta_i|\phi \sim N\left(\frac{\gamma}{z_i}\sum_{j=1}^{n} W_{ij}\phi_j, \frac{\tau^2}{z_i}\right), \tag{4}$$

$$\phi_j|\theta \sim N\left(\frac{\gamma}{z'_j}\sum_{i=1}^{n} W_{ij}\theta_i, \frac{\tau^2}{z'_j}\right), \tag{5}$$

where γ and τ^2 are constants. We now establish that this model is equivalent to marginally specifying the joint Gaussian distribution for θ and ϕ as follows. A proof of this equivalence is found in the appendix.

theoremthmgmrf Let W be a weight matrix with row sums $z_i = \sum_j W_{ij}$ and column sums $z'_j = \sum_i W_{ij}$. Let $\tau^2 > 0$ and $\gamma \in (0, 1)$. Then the conditional specifications

$$\theta_i|\phi \sim N\left(\frac{\gamma}{z_i}\sum_{j=1}^{n} W_{ij}\phi_j, \frac{\tau^2}{z_i}\right) \qquad \phi_j|\theta \sim N\left(\frac{\gamma}{z'_j}\sum_{i=1}^{n} W_{ij}\theta_i, \frac{\tau^2}{z'_j}\right)$$

define a valid, non-degenerate probability distribution over θ and ϕ with marginal distribution $\begin{pmatrix}\theta \\ \phi\end{pmatrix} \sim N(\mu, \Sigma)$, where $\mu = 0$ and

$$\Sigma = \tau^2 \begin{pmatrix} Z & -\gamma W \\ -\gamma W^\top & Z' \end{pmatrix}^{-1}. \tag{6}$$

Because our goal is to obtain predictions for the real attributes θ, we view the target attributes ϕ as nuisance parameters and marginalize them out. By studying the precision matrix $M = \Sigma^{-1}$ and applying the standard 2×2 block matrix inversion Schur complement

$$(M^{-1})_{11} = (M_{11} - M_{12}M_{22}^{-1}M_{21})^{-1},$$

we find the marginal prior for θ is then Gaussian with mean 0 and covariance matrix $\tau^2\left(Z - \gamma^2 W Z'^{-1} W^\top\right)^{-1}$. Therefore, minimizing the posterior log-likelihood conditional on observing values θ_i for $i \in V_0$ reduces to the optimization problem

$$\min_\theta \theta^\top L'\theta, \qquad \text{subject to } \theta|_{V_0} = \theta_0 \tag{7}$$

for the modified Laplacian

$$L' = (Z - \gamma^2 W Z'^{-1} W^\top). \tag{8}$$

We call this modified Laplacian the *decoupled Laplacian*, to emphasize the decoupling between the real responses θ and the target responses ϕ in the underlying model.

From this expression for the decoupled Laplacian we can view $\tilde{A} = WZ'^{-1}W$ as a weighted adjacency matrix for an auxiliary graph that is essentially connecting nodes to their two-hop neighbors with appropriately weighted edges. With this modified auxiliary matrix, the solution to the decoupled smoothing objective is then

$$\hat{\theta}_1 = (I - P_{11})^{-1} P_{10}\theta_0, \tag{9}$$

as before in Equation (2), but now with $P = Z^{-1}(Z - \gamma^2 W Z'^{-1} W^\top)$.

5.3 Combining independent estimators

Consider that the information contributed by each friend j for estimating θ_i is in the form of the "observations" $\{\theta_k : k \in_j\}$, which are values located two steps away from unit i. One way to think about combining this information has been studied extensively in the statistics literature in the context of estimating a common location parameter from samples of varying precision. Explicitly, suppose the variables in the set $\{\theta_k : k \in_j\}$ follow a distribution with mean θ_i and variance σ_j. That is, all observations contribute unbiased information for estimating θ_i, but they have varying precisions which are modulated by unit j. Then the weight matrix entry reduces to $W_{ij} = A_{ij}/\sigma_j^2$, with row sum $z_i = \sum_{\ell \in_i} \sigma_\ell^{-2}$ and column sum

$z'_j = d_j/\sigma_j^2$. In this case, we now show that we obtain a concise recurrence recognizable as a particular weighting of 2-hop majority vote.

From Section 5.2, the auxiliary graph with this diagonal covariance specification has an adjacency matrix with entries

$$\tilde{A}_{ij} = \sum_k A_{ik}A_{jk}/(d_k\sigma_k^2), \tag{10}$$

with the smoothing update rule being $\hat{\theta}^t = Z^{-1}\tilde{A}\hat{\theta}^{t-1}$. For an unlabeled node i, if we employ the weights derived here we obtain the recurrence:

$$\begin{aligned}\hat{\theta}_i^t &= (Z^{-1}\tilde{A}\hat{\theta}^{t-1})_i = \frac{1}{z_i}\sum_j \tilde{A}_{ij}\hat{\theta}_j^{t-1} \\ &= \frac{1}{\sum_{\ell\in_i}\sigma_\ell^{-2}}\sum_{k\in_i}\frac{1}{d_k\sigma_k^2}\sum_{j\in_k}\hat{\theta}_j^{t-1}.\end{aligned} \tag{11}$$

By viewing the aggregation as performed on a graph, we can in fact turn this standard estimation procedure into an iterative procedure. As a generic problem of aggregating estimators, if we observe $X_{jk} \sim N(\theta, \zeta_j^2)$, $k = 1, \ldots, d_j$, then the minimum variance, linear unbiased estimator (MVLUE) of θ when the ζ_j^2 are known, is $\hat{\theta} = \sum_j w_j\bar{X}_j$ with weights given by $w_j = (d_j/\zeta_j^2)/\sum_k(d_k/\zeta_k^2)$. Our formulation of expert aggregation aligns with this view, where the expert variances are $\sigma^2 = \zeta_i^2/d_i$ and higher degree nodes therefore having appropriately more precise information.

In order to estimate σ_j^2 we can notice that it essentially represents the standard error for the expert estimate. Hence we can use the regular standard error estimate for the Gaussian sample mean, $\hat{\sigma}_j^2 = S_j^2/d_j$, where (recall that ${}^0_j$ is the labeled neighborhood and $d_j^0 = |{}^0_j|$ is the labeled degree)

$$S_j^2 = \frac{1}{d_j^0}\sum_{k\in_j^0}\left(\theta_k - \frac{1}{d_j^0}\sum_{\ell\in_j^0}\theta_\ell\right)^2$$

is the sample variance of the labeled nodes in the neighborhood of j. We then use $\hat{\sigma}_j^2$ as a plugin estimate in the update rule in Equation (11), giving

$$\hat{\theta}_i^t = \frac{1}{\sum_{\ell\in_i}(S_\ell^2/d_\ell)^{-1}}\sum_{k\in_i}\frac{1}{S_k^2}\sum_{j\in_k}\hat{\theta}_j^{t-1}. \tag{12}$$

Alternatively, we can directly impose homogeneous standard errors, $\sigma_i^2 = \sigma^2/d_i$, in which case the normalization term reduces to $1/\sum_{\ell\in_i}\sigma_\ell^{-2} = 1/\sum_{\ell\in_i}d_\ell$, the number of nodes in the two-step neighborhood of i, and we obtain the update rule

$$\hat{\theta}_i^t = \frac{1}{\sum_{\ell\in_i}d_\ell}\sum_{k\in_i}\sum_{j\in_k}\hat{\theta}_j^{t-1}. \tag{13}$$

For exposition here we have let d_ℓ represent the total graph degree of unit ℓ, which disregards the number of labeled nodes. We thus see how iterating a simple two-hop majority vote update can be motivated for graph smoothing, despite initial appearances as defining a "non-physical" process whereby information bypasses individuals. This simple recurrence emerges as the MVLUE under the assumption that expert friends contribute independent opinions, an assumption which appears to be reasonable for the graph-based learning problems we study.

6 Iterative perspective on smoothing

In this section we outline how the closed form solutions to the smoothing problems discussed in this work can be formulated as the solutions to the iterative application of recurrence relations. We first review the known iterative formulation of smoothing. We formulate the recurrence relation that underlies the decoupled smoothing problem studied in this work. In the next section, we will show how this recurrence can be interpreted in the language of expert opinion aggregation, giving us an intuition for how to choose the previously unspecified weight matrix W in the recurrence we derive here.

6.1 Iterative formulation of smoothing

The closed form solution to the smoothing objective in Equation (1) is known to arise from a repeated application of majority vote in the following sense: define the time 0 estimate $\hat{\theta}^0$ to agree with the true labels on V_0. Take the transition matrix $P = D^{-1}A$ and perform the updates

$$\hat{\theta}_1^t = P_{01}\theta_0 + P_{11}\hat{\theta}_1^{t-1}, \qquad \hat{\theta}_0^t = \theta_0, \tag{14}$$

where $P = \begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix}$ has been partitioned into labeled and unlabeled blocks, as before. In other words, the time t estimate is the majority vote estimate using the time $t-1$ predictions, where after each step we replace the labeled predictions by their original, true labels. In the limit,

$$\hat{\theta}_1 = \lim_{t\to\infty} \hat{\theta}_1^t = (I - P_{11})^{-1}P_{10}\theta_0, \tag{15}$$

which is the solution to Equation (1) given in Equation (2).

6.2 Iterative formulation of decoupled smoothing

Examining the decoupled Laplacian in Equation (8) alongside the iterative smoothing formulation provides an iterative algorithm for the decoupled smoother. We define an auxiliary weighted, directed graph with weighted adjacency matrix $\tilde{A} = WZ'^{-1}W^\top$, which has edge weight $\tilde{A}_{ij} = \sum_k \frac{W_{ik}W_{jk}}{z'_k}$. The out-degree of node i reduces to $\sum_j \tilde{A}_{ij} = z_i$, where z_i is the same row sum defined in Section 5.2. Hence the degree matrix of $\tilde{A}$ is Z, and the solution to the decoupled smoothing problem in Equation (7) results from performing the iterative one-hop majority vote updates, Equation (14), on the auxiliary, directed graph.

By employing the update equations in Equation (14) with the transition matrix $P = Z^{-1}WZ'^{-1}W^\top$, we can see that decoupled smoothing amounts to an iterative update of a weighted *two-hop* majority vote.

6.3 Improving majority vote with regularization

The iterative perspective is not only useful for computational purposes but also gives insights into how to improve the basic iterated majority vote. Here we describe an improvement to the basic smoothing algorithm, inspired by the details of implementing the iterative algorithm, which can be applied in either the standard, soft, or decoupled setting.

Since iterative majority vote is recursively defined, it relies on defining an initial set of guesses for the unlabeled nodes; when $t = 1$, equation (14) requires a value for $\hat{\theta}_1^0$ which

can be safely set to random initial labels without compromising the limiting result. Then, equation (14) can also be written elementwise as

$$\hat{\theta}_i^t = \frac{1}{d_i} \sum_{j \in_i} \hat{\theta}_j^{t-1}. \tag{16}$$

for every unlabeled node $i \in V_1$. From here, one sees that the performance of the first few iterations can be quite unsatisfactory, because it depends strongly on the initial noise input $\hat{\theta}_1^0$. An alternative strategy is to set the first iteration of the unlabeled nodes to be the average value of labeled friends only:

$$\hat{\theta}_i^1 = \frac{1}{d_i^0} \sum_{j \in_i^0} \theta_j \tag{17}$$

for $i \in V_1$.

This is a reasonable choice because it avoids corrupting the early estimates with noise, and indeed this modification tends to lead to a slight bump in performance in early iterations; see Section 7 for example illustrations.

However, we can further generalize this idea of upweighting the true labels when they should be trusted more than haphazard (random) guesses. Consider the convex combination update

$$\hat{\theta}_i^t = \lambda_i^t \frac{1}{d_i^0} \sum_{j \in N_i^0} \theta_j + (1 - \lambda_i^t) \frac{1}{d_i^1} \sum_{j \in N_i^1} \hat{\theta}_j^{t-1}, \tag{18}$$

where $\lambda_i^t \in [0, 1]$ are weight parameters that control the amount of trust to place in the guesses of previous iterations. This places weight λ_i^t on the true labels and weight $1 - \lambda_i^t$ on the predicted values for iteration $t - 1$. Most generally λ_i^t may be indexed by both the unit i and the time step t, as it is reasonable to expect that this weight should be personalized to individuals (e.g., vary based on degree) and that estimates of later iterations should be trusted more (which would have λ_i^t decreasing in time t).

Decomposing the sum in equation (16) as

$$\hat{\theta}_i^t = \frac{1}{d_i} \left[\sum_{j \in_i^0} \theta_j + \sum_{j \in_i^1} \hat{\theta}_j^{t-1} \right],$$

we see that equation (18) reduces to the one-hop majority vote iteration for the choice of weights $\lambda_i^t = d_i^0/d_i$, which is constant in t.

The search space of weights λ_i^t is quite large and we leave a formal analysis of this space to future work, restricting ourselves here to providing intuition for choices of λ_i^t that appear to work well in our empirical experiments. The goal is to place more weight on labeled nodes in the early stages and less weight on labeled nodes at later iterations, which suggests λ_i^t decaying in t. Consider parametrizing $\lambda_i^t = f_i(t)$ for a function $f_i(\cdot)$ that reduces the number of parameters. For example one may consider the choice $f_i(t) = (d_i^0/d_i)^t$, which represents exponential decay in t. The choices of λ_i^t will lead to different limiting values $\lim_{t \to \infty} \hat{\theta}^t$, some of which appear to outperform the basic version of majority vote.

7 Empirical Illustrations

7.1 decoupled smoothing

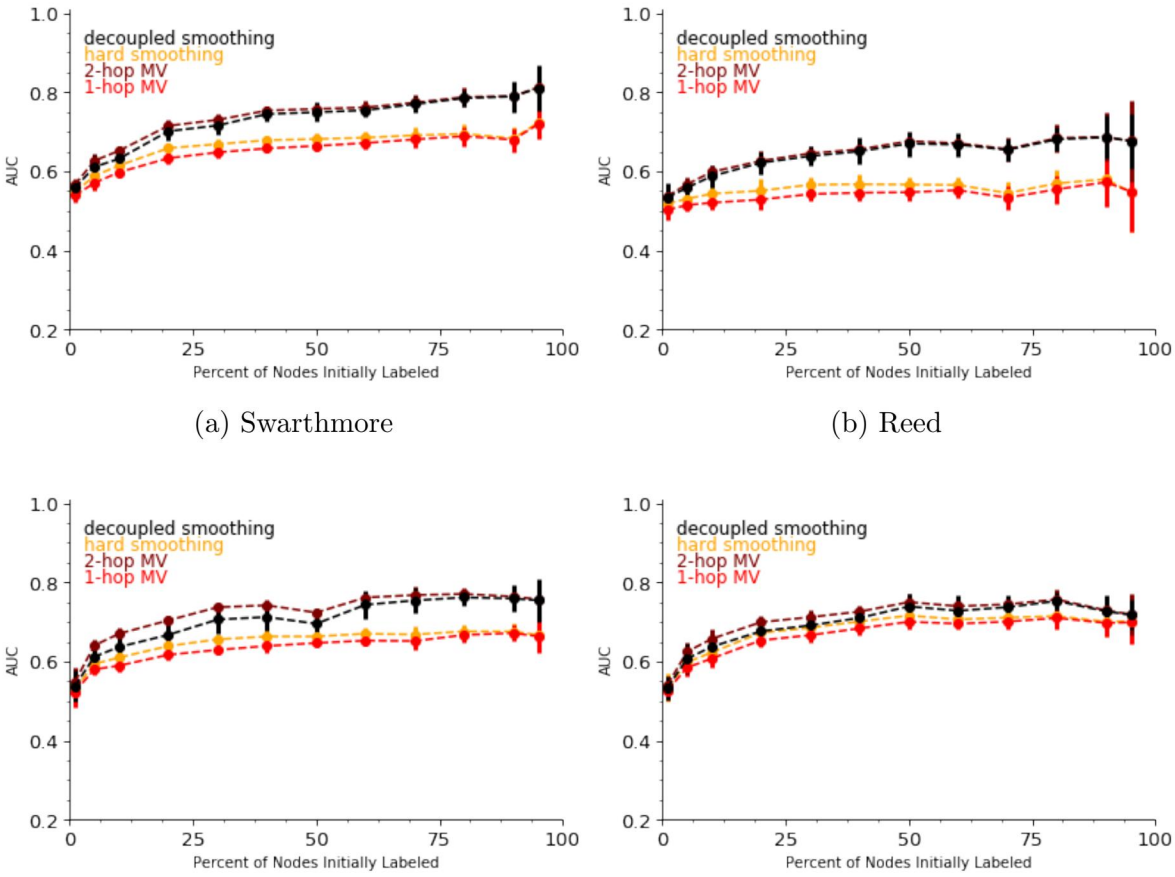
We perform experiments on a sample of undergraduate college networks collected from a single-day snapshot of Facebook in September 2005. We focus on the task of gender classification in these networks, restricting our analyses to the subset of nodes that self-reported their gender to the platform. We use the largest connected components from four medium-sized colleges, `amherst`, `reed`, `haverford`, and `swarthmore`. Amherst has 2032 nodes and 78733 edges, Reed has 962 nodes and 18812 edges, Haverford has 1350 nodes and 53904 edges, and Swarthmore has 1517 nodes and 53725 edges. For all plots in this section we attempt classification 10 times based on different independent labelled subsets of nodes. The plots show the average AUC with error bars denoting the standard deviation across the 10 runs. In Figure 1 we see our experiments with decoupled smoothing, which indicate that the two-hop majority vote update given by Equation (13) outperforms both the standard 1-hop majority vote estimator and the corresponding (ZGL) smoothing estimator in terms of classification accuracy, regardless of the percentage of initially labeled nodes. Meanwhile we also observe that decoupled smoothing performs worse than the much simpler 2-hop majority vote estimator in some situations (namely `amherst` and `haverford`). Recall from Section 4.3 that decoupled smoothing can be interpreted as iterated 2-hop majority vote, but with randomly initialized guesses. We suspect that the better performance of the plain 2-hop majority vote is due to the fact that local information is more pertinent for this particular task than global information, and the smoothing algorithms are inappropriately synthesizing information from local and global sources.

7.2 Regularized iterations

In Section 6.3 we considered a modified iterated majority vote algorithm that includes a regularization term $\lambda_i^t = (d_i^L/d_i)^t$ for each unlabeled node i. This modification was inspired by the empirical observation that 2-hop majority vote outperforms the limiting iterated smoother. As a secondary inspiration, using (17) as the first iteration's update rule instead of (16) greatly reduces the number of iterations needed for convergence. In this section, we present experimental results from applying these modifications for both hard smoothing and decoupled smoothing on a synthetic stochastic blockmodel graph as well as the on the Facebook networks.

7.2.1 Improved iterative decoupled smoothing

We first test our modification in an overdispersed stochastic block model (oSBM), an extension of the stochastic blockmodel that contains an additional parameter to model *monophily*. It is thus designed to capture aspects of the network that are particularly well suited for 2-hop estimators. Again, we use two blocks with 500 nodes in each block representing 500 males and 500 females. The expected average degree is 42 and dispersion rate is 0.004, giving the same edge density and dispersion rate as in [4]. Here we compare the iterative method results for the original decoupled smoothing method against the regularized iterative decoupled smoothing method. As shown in Figure 2b, the regularization improves the overall prediction accuracy for decoupled smoothing under the overdispersed stochastic block model.



(a) Swarthmore (b) Reed

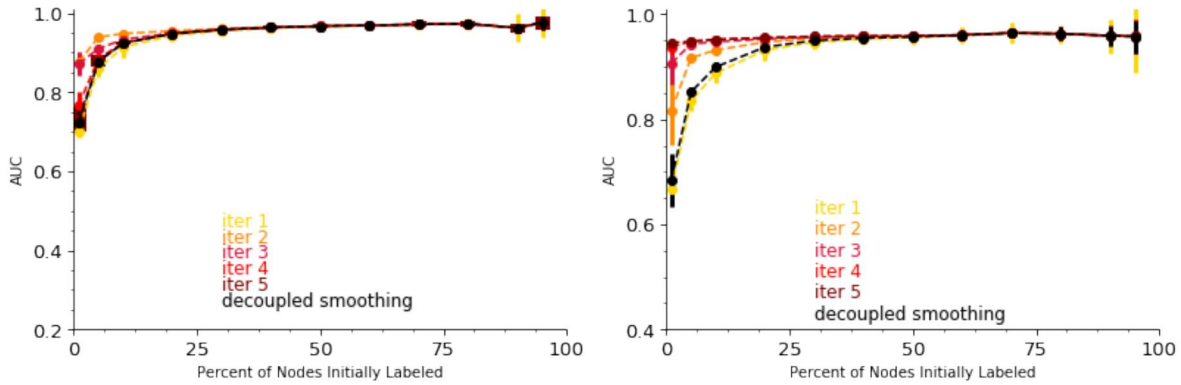
(c) Amherst (d) Haverford

Figure 1: Decoupled smoothing performance for classification accuracy of different iterative estimators of gender, compared with hard smoothing (ZGL) and 1-hop and 2-hop majority vote. The estimators based on two-step neighborhood information clearly outperform those based on one-step information, but 2-hop majority vote sometimes outperforms decoupled smoothing.

On the Facebook `Amherst` network we use the regularization $\lambda_i^t = (d_i^0/(rd_i))^{t-1}$, where r is the initial percent of labeled nodes. This choice is motivated by the fact that the relative importance of local to global information should depend on the proportion of labeled nodes; if there is little local information available, then it makes sense to pull in information from farther away. We can see that with this particular regularization term, the smoother modestly improves the overall prediction accuracy for decoupled smoothing. It is encouraging that from pure intuition we can see better results; with a more careful optimization over the λ_i^t space it is possible that performance can be further improved.

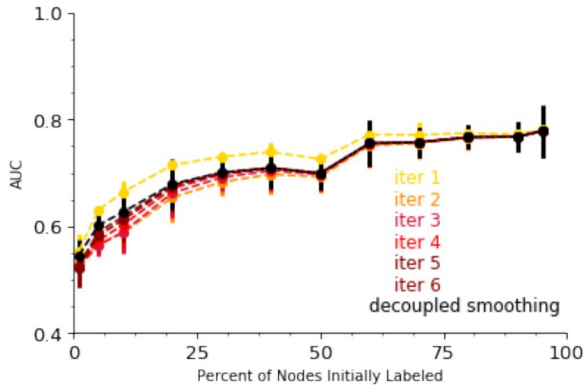
8 Discussion

In this work we investigate the use of graph-based learning problems for social networks, where a thoughtful understanding of social forces that underlie network formation can help inform the choice of the smoothing model. Our work is motivated by the investigation into empirical social phenomena in [4], which highlights the distinction between "the company you

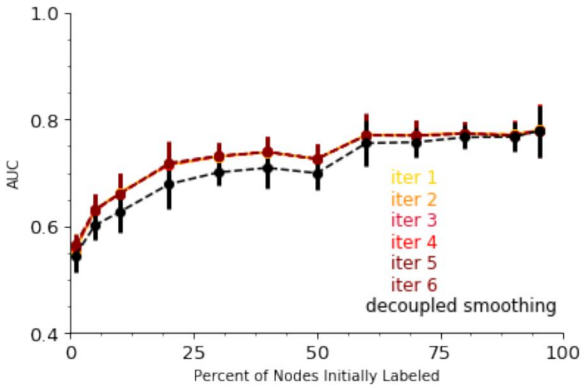


(a) oSBM: decoupled smoothing

(b) oSBM: regularized decoupled smoothing



(c) Amherst: decoupled smoothing



(d) Amherst: regularized decoupled smoothing

Figure 2: Decoupled smoothers, with and without regularization, for classifying gender on an oSBM and the `Amherst` dataset.

keep" and "the company you're kept in." We develop a model for decoupled graph smoothing that links this empirical observation to graph smoothing, semi-supervised learning, and diffusion algorithms popularly used for node classification tasks. We provide a Bayesian viewpoint of this model which is related to the literature on expert opinion aggregation.

As a part of our analysis, we contribute an iterative algorithm for soft smoothing, which allows us to solve soft smoothing problems efficiently on large datasets. We find that a close examination into the form of iteration is not only crucial for computational efficiency but also for efficacy of predictive performance, as the basic majority vote algorithms make suboptimal choices in the initial iterations. We contribute a generalization that allows one to place greater weight on and regularize toward labeled values. This method displays improved performance on some simulated and real datasets. This generalization is flexible enough that the practitioner has a lot of control over the resulting algorithm. The optimality of such choices has yet to be fully explored and may well vary depending on the particular domain of application.

References

[1] X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. *ICML*, volume 20, pages 912–919, 2003.

[2] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schlkopf. Learning with local and global consistency. *NIPS*, 2004.

[3] Ya Xu, Justin Dyer, and Art B. Owen. Empirical stationary correlations for semi-supervised learning on graphs. 2010.

[4] K. M. Altenburger and J. Ugander. Bias and variance in the social structure of gender. ArXiv e-prints, May 2017. arXiv:1705.04774

[5] L. Backstrom and J. Leskovec. Supervised random walks: Predicting and recommending links in social networks. In Proc. of ACM WSDM11, pages 635644, Hong Kong, China, 2011.

[6] J. Neville and D. Jensen. Iterative classication in relational data. In Workshop on Learning Statistical Models from Relational Data, AAAI, 2000.

[7] Belkin, Mikhail and Matveeva, Irina and Niyogi, Partha. Regularization and semi-supervised learning on large graphs. International Conference on Computational Learning Theory, 624–638, 2004,Springer.