Decoupled Smoothing on Graphs

Alex Chin
Stanford University
ajchin@stanford.edu

Yatong Chen
Stanford University
yatong@stanford.edu

Kristen M. Altenburger
Stanford University
kaltenb@stanford.edu

Johan Ugander
Stanford University
jugander@stanford.edu

ABSTRACT

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 is less well-studied in the literature. In this work, we demonstrate that for social networks, the basic friendship graph itself may often not be the appropriate graph for predicting node attributes using graph smoothing. 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," as observed in recent empirical work. Our model results in a rigorous extension of the Gaussian Markov Random Field (GMRF) models that underlie graph smoothing, interpretable as smoothing on an appropriate auxiliary graph of weighted or unweighted two-hop relationships.

KEYWORDS

Semi-supervised learning; graph smoothing; attribute prediction

ACM Reference Format:
Alex Chin, Yatong Chen, Kristen M. Altenburger, and Johan Ugander. 2019. Decoupled Smoothing on Graphs. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3308558.3313748

1 INTRODUCTION

Graph-based learning describes a broad class of problems in which 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 "birds of a feather flock together" [22]. Many varieties of models derived from this assumption have been studied and successfully applied. The most popular methods include the work of Zhu, Ghahramani, and Lafferty (ZGL) on semi-supervised learning using Gaussian Markov Random Fields [33], Zhou et al.'s related method for random walk smoothing [31], and Xu, Dyer, and Owen's work connecting the former two methods to methods for kriging from spatial statistics [30]. See [32] for a survey of semi-supervised learning methods on graphs.

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 $x_i \in \mathbb{R}^D$ associated with response values $y_i \in \mathbb{R}$ (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 $\mathbb{R}^D$ 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.

In this work, we focus on graph-based learning problems on social networks, where the relationship between the graph and the response variables can be quite different than the basic similarity relationship between responses on synthetic graphs produced by standard induction methods such as k-nearest neighbors. We rely on the recent observation that there are contexts where the fundamental assumption of correlation that drives graph-based learning—that connected nodes are correlated in their responses—may not be true or necessary for inference to succeed [1, 25]. The work in [1], which focused on the social structure of gender, observed a fundamental difference between similarities with "the company you keep" and "the company you're 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. In the present work, we develop these differing social assumptions into an alternative semi-supervised learning problem, where correlations on the social networks are based on a decoupling of separate "real" and "target" responses of individuals corresponding to separated notions of social identity and social preference. The key principle is to allow the outcomes of an individual to be possibly different from the outcomes with which the individual prefers to associate.

The consequences of this decoupling between real and target responses is that individuals connected in the graph are only correlated in their responses if the real and target responses of the individuals are correlated with each other. But even if they are entirely uncorrelated, a subtler correlation is nonetheless induced, as pointed out in [1], which can be particularly useful for inference: even if an individual is not correlated with their graph neighbors, the fact that they have preferences for a target response will still imply that their neighbors are correlated with each other. As a

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313748

result, individuals can be similar to their friends-of-friends without necessarily being similar to their friends.

In this work we show that the above decoupling can be modeled in a very natural semi-supervised learning problem for an appropriately constructed auxiliary similarity matrix encoding two-hop similarities. The two-hop similarity matrix that we consider is derived from previously unrelated literature on combining estimators [8, 11, 12, 19, 23, 26], since the "real" and "target" state of each node must be estimated from the dispersed evidence surrounding them in the graph. By making this connection, our work also introduces an iterative procedure for the combining estimators problem when data is being collected on a graph.

As a byproduct of our investigation, we also strengthen the known connections between random walk-based methods (such as ZGL) and Kriging-based methods, introduce a procedure that maps any Kriging problem to a ZGL problem. This connection between the methods enables us to take a "decoupled" view of either problem class, and also allows us to use computationally expedient iterative solvers for standard ZGL-type problems to solve Kriging problems, coupled or decoupled.

2 GRAPH SMOOTHING PRELIMINARIES

In this section we review two standard formulations of graph smoothing, the semi-supervised learning problem of [33], which we refer to here simply as smoothing, and the graph regularization approach involving noisy observations that we refer to as soft smoothing. We review the closed form solutions, and later give recurrences that converge on the closed form solutions for both problems.

We assume that we observe a social network of n people with connections represented by an undirected, unweighted graph $G = (V, E)$. We let A and D denote the adjacency matrix and degree matrix of the graph, respectively, where D has diagonal entries $d_i = \sum_j A_{ij}$. We denote the neighborhood of node i by $\mathcal{N}_i = \{j : A_{ij} > 0\}$. Each person i has a outcome value of interest $\theta_i \in \mathbb{R}$; we write $\theta \in \mathbb{R}^n$ for the vector of response values. We only observe the labels θ_i for a subset of the nodes; we denote the set of labeled nodes V_0 and the set of unlabeled nodes $V_1 = V \setminus V_0$. Let $\mathcal{N}_i^0 = \{j : A_{ij} > 0, j \in V_0\}$ be the set of neighbors of i for which the labels are observed, and $\mathcal{N}_i^1 = \{j : A_{ij} > 0, j \in V_1\}$ be the set of neighbors of i for which the labels are unobserved. Correspondingly define $d_i^0 = |\mathcal{N}_i^0|$ and $d_i^1 = |\mathcal{N}_i^1|$, the "labeled degree" and "unlabeled degree," respectively. The vectors $\theta_0 \in \mathbb{R}^{|V_0|}$ and $\theta_1 \in \mathbb{R}^{|V_1|}$ represent the restriction of the response vector θ to the observed and unobserved node sets. In general, we reserve the subscripts 0 and 1 for subsetting on labeled and unlabeled nodes, respectively.

2.1 Smoothing

The standard formulation of graph smoothing, proposed in [33], 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 [20, 30]. Suppose we place a Gaussian Markov Random Field (GMRF) on the node set by placing a prior $\theta \sim N\left(0, \tau^2(D - \gamma A)^{-1}\right)$ on θ. This prior is the *conditional autoregressive* (CAR) model popular in the spatial statistics literature [5, 6], and has the property that θ_i conditional on the other values of θ follows the distribution

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

Under this GMRF prior, the Bayes estimator of θ, 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.

2.2 Soft smoothing

A second approach to graph smoothing has its origins in graph-based regularization, as studied in [4]. In this problem, we no longer observe θ_0 directly but rather observe $y_0 = \theta_0 + \varepsilon$ where $\varepsilon \sim N(0, \zeta^2 I)$. We will refer to this problem as soft smoothing, to emphasize the close relationship to ordinary smoothing. Consider the estimator

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \|y_0 - \theta_0\|_2^2 + \lambda\theta^\top(D - A)\theta. \tag{4}$$

Under the prior assumption $\theta \sim N(0, \tau^2(D - \gamma A)^{-1})$, which is equivalent to the conditional specification of the GMRF on the node set in Equation (3), the estimator $\hat{\theta}$ in Equation (4) is the Bayes estimator for this model with $\lambda = \zeta^2/\tau^2$ and $\gamma \to 1$. This estimator $\hat{\theta}$ has the following closed form solution when $\lambda > 0$:

$$\hat{\theta} = (J + \lambda L)^{-1}Jy^*, \text{ where } J = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \tag{5}$$

and y^* is a vector that agrees with the labeled points and takes on any value on the unlabeled points. If $\zeta^2 = 0$ (whereby $\lambda = 0$) then the optimization problem in Equation (4) reduces to the noiseless graph smoothing problem in Equation (1) [30]. Later, we will show that even when $\zeta^2 > 0$ there is a direct mapping between soft and "hard" smoothing problem instances.

3 DECOUPLED GRAPH SMOOTHING

Crucial to the performance of graph smoothing estimators like Equation (2) is a smoothness assumption positing that the response values vary smoothly across the topology of the graph. This assumption in often appropriate in synthetic graph applications like those constructed for image segmentation and changepoint detection problems, and smooth kernel edge weights are often assumed in theoretical studies of semi-supervised learning performance [13, 24].

However, this assumption can fail to hold in social networks. Indeed, [1] shows that even when homophily is minimal or nonexistent, additional variation known as *monophily* among friend-of-friend (two-hop) relations can be exploited to predict outcomes. Intuitively, if an unlabeled individual i is friends with person j who tends to befriend a high proportion of female individuals, then we might be more certain that the unlabeled individual i is female as well. In this case, what matters for classifying individual i is not the label of individual j, but rather individual j's preferences in her friends. In this work we propose *decoupling* the true parameter of interest θ_j from a *target parameter* ϕ_j that captures the true parameters of the neighbors of j. We now study a model that gives rise to such a decoupling.

Suppose we have a weight matrix W, and denote the row sums by $z_i = \sum_j W_{ij}$ and the column sums by $z'_j = \sum_i W_{ij}$. Our approach is to relate the real θ_i and target ϕ_i via W as follows:

$$\theta_i \approx \frac{1}{z_i} \sum_{j=1}^{n} W_{ij}\phi_j, \tag{6}$$

$$\phi_j \approx \frac{1}{z'_j} \sum_{i=1}^{n} W_{ij}\phi_i. \tag{7}$$

Note that W need not be symmetric. The choice of W is to be set by the researcher and may depend on the particular problem being solved. For example, if W is set to be the adjacency matrix A, then ϕ_i and θ_i become unweighted averages of peer values. Section 5 contains a discussion of alternative options for setting W, where we argue that more complex choices of W may indeed be desirable.

Expressions (6) and (7) are to be understood in an "average" and informal sense. Formally, we consider the Gaussian Markov random field model

$$\theta_i|\phi \sim N\left(\frac{\gamma}{z_i}\sum_{j=1}^{n} W_{ij}\phi_j, \frac{\tau^2}{z_i}\right), \quad \phi_j|\theta \sim N\left(\frac{\gamma}{z'_j}\sum_{i=1}^{n} W_{ij}\theta_i, \frac{\tau^2}{z'_j}\right), \tag{8}$$

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.

Theorem 3.1. *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

$$\Sigma = \tau^2 \begin{pmatrix} Z & -\gamma W \\ -\gamma W^\top & Z' \end{pmatrix}^{-1}, \tag{9}$$

where $Z = \mathrm{diag}(z_1, \ldots, z_n)$ and $Z' = \mathrm{diag}(z'_1, \ldots, z'_n)$.

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

$$\min_\theta \theta^\top L' \theta, \qquad \text{subject to } \theta|_{V_0} = \theta_0 \tag{10}$$

for the modified Laplacian

$$L' = (Z - \gamma^2 W Z'^{-1} W^\top). \tag{11}$$

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 $\tilde{A} = WZ'^{-1}W^\top$ 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

$$\hat{\theta}_1 = (I - P_{11})^{-1} P_{10}\theta_0, \tag{12}$$

as before in Equation (2), but now with $P = Z^{-1}(Z - \gamma^2 W Z'^{-1} W^\top)$. The closed form solution to the soft decoupled smoothing objective, as in Equation (5), becomes

$$\hat{\theta} = (J + \lambda(Z - \gamma^2 W Z'^{-1} W^\top))^{-1} J y^*. \tag{13}$$

We have intentionally kept the weight matrix W generic in these derivations. In Section 5, we use ideas from the literature on expert aggregation to motivate our choice of W.

3.1 Coupling $\theta = \phi$ reduces to the standard smoother

Here we further motivate the sense in which the above decoupled smoothing problem is strongly connected to the basic smoothing problem: by conditioning the distribution of the GMRF upon $\theta = \phi$, we recover the basic smoothing objective.

Under the transformation $\nu = \theta - \phi$, the log-likelihood of the vector (θ, ν) is

$$p(\theta, \nu) \propto \exp\Big\{ -\frac{1}{2\tau^2}\Big(\theta^\top(Z + Z' - 2\gamma W)\theta - 2\theta^\top(Z' - \gamma W)\nu + \nu^\top Z'\nu\Big)\Big\}.$$

If we let the weight matrix W be equal to the adjacency matrix A, so that the row and column sums reduce to the degree matrix $Z = Z' = D$, then the joint parameters (θ, ν) have the covariance matrix

$$\tau^2 \begin{pmatrix} 2(D - \gamma A) & -(D - \gamma A) \\ -(D - \gamma A) & D \end{pmatrix}^{-1}.$$

The upper left block, $2\tau^2(D - A)$, is the precision matrix of θ conditional on ν, which shows that θ follows the ordinary smoother distribution in Section 2.1 with τ^2 replaced by $2\tau^2$.

4 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 then contribute a reduction which shows that any *soft* smoothing problem can be written as a *hard* smoothing problem, and can thus be written iteratively as well. We formulate the recurrence relation that underlies the decoupled smoothing problem studied in this work. In the next section, Section 5, 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.

4.1 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 [7, 21] 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

$$\hat{\theta}_1^t = P_{10}\theta_0 + P_{11}\hat{\theta}_1^{t-1}, \qquad \hat{\theta}_0^t = \theta_0, \tag{14}$$

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).

4.2 Iterative formulation of soft smoothing

We now contribute an iterative algorithm for soft smoothing by showing that a soft smoothing solution is equivalent to a hard smoothing solution on an appropriate auxiliary graph. Consider an augmented graph $\tilde{G}$ formed by attaching a degree-one node to each labeled node on the original graph, with edge weight $\alpha > 0$. The vertex set of the resulting graph contains a copy of the labeled nodes, denoted $\tilde{V}_0$, in addition to the original labeled node set V_0 and unlabeled node set V_1.

This new graph has adjacency matrix

$$\tilde{A} = \begin{pmatrix} 0 & \alpha I & 0 \\ \alpha I & A_{00} & A_{01} \\ 0 & A_{10} & A_{11} \end{pmatrix},$$

where the first block corresponds to $\tilde{V}_0$, the second block corresponds to V_0, and the third block corresponds to V_1. If we compute the hard smoothing solution on this augmented graph treating $\tilde{V}_0$ as the labeled set and the entire original vertex set V as the unlabeled set, the corresponding random walk transition matrix is

$$\tilde{P} = \tilde{D}^{-1}\tilde{A} = \begin{pmatrix} 0 & I & 0 \\ \tilde{D}_0^{-1}A_{10} & \tilde{D}_0^{-1}A_{00} & \alpha\tilde{D}_0^{-1} \\ D_1^{-1}A_{10} & D_1^{-1}A_{11} & 0 \end{pmatrix},$$

where D_1 is the degree matrix of the original graph restricted to the unlabeled node set V_1, and $\tilde{D}_0 = D_0 + \alpha I$ is the degree matrix restricted to the labeled node set V_0, adjusted to account for the new auxiliary nodes.

The hard smoothing solution in Equation (15) is $\hat{\theta} = (I - P_{11})^{-1}P_{10}\theta_0$. Applying this solution to the above problem, we have

$$\begin{aligned} P_{11} &= (D + \alpha J)^{-1}A, \\ P_{10}\theta_0 &= \alpha(D_0 + \alpha I)^{-1}\theta_0 = \alpha(D + \alpha J)^{-1}Jy^*, \end{aligned}$$

which results in the estimator

$$\begin{aligned} \hat{\theta} &= (I - (D + \alpha J)^{-1}A)^{-1}\alpha(D + \alpha J)^{-1}Jy^* \\ &= (\alpha J + D - A)^{-1}\alpha Jy^* \\ &= (J + \alpha^{-1}L)^{-1}Jy^*. \end{aligned}$$

This estimator is precisely the soft smoothing closed-form solution in Equation (5) when $\alpha = \lambda^{-1}$. Notice the non-equivalence between this solution and the solution in the case of $\zeta^2 = 0$ (for which $\lambda = 0$).

For soft smoothing, the iterative update, Equation (14), then amounts to

$$\begin{aligned} \hat{\theta}^t &= \alpha(D + \alpha J)^{-1}Jy^* + (D + \alpha J)^{-1}A\hat{\theta}^{t-1} \\ &= (D + \alpha J)^{-1}(\alpha Jy^* + A\hat{\theta}^{t-1}), \end{aligned} \tag{16}$$

where $D + \alpha J$ is diagonal and therefore easily invertible, making for an expedient computational procedure for solving soft smoothing problems. This interpretation of soft smoothing connects to a related used of "shadow nodes" in the *adsorption* method of graph smoothing [3].

4.3 Iterative formulation of decoupled smoothing

Examining the decoupled Laplacian in Equation (11) 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 3. Hence the degree matrix of $\tilde{A}$ is Z, and the solution to the decoupled smoothing problem in Equation (10) 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. The reduction from soft smoothing to ordinary smoothing in Section 4.2 also gives an iterative formulation of soft decoupled smoothing.

4.4 Improving majority vote with local 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.

Limiting noise in early steps. 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 element-wise as

$$\hat{\theta}_i^t = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \hat{\theta}_j^{t-1}. \tag{17}$$

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 friends labelled at the previous time-step only. For $t = 1$ this technique gives:

$$\hat{\theta}_i^1 = \frac{1}{|\mathcal{N}_i^0|} \sum_{j \in \mathcal{N}_i^0} \theta_j, \tag{18}$$

for $i \in V_1$. Recall that $\mathcal{N}_i^0$ is the set of neighbors of i for which a label was observed. For nodes $i \in V_1$ with no labelled neighbors, we do not update $\hat{\theta}_i^1$. More generally, we can postpone making assignments $\hat{\theta}_i^t$ until a node has some neighbor with an estimate at time $t-1$. If using $\hat{\theta}^t$ to make predictions, for missing values we then guess randomly with the class proportions of the training sample.

This strategy, which sidesteps dependence on random initial $\hat{\theta}_1^0$, tends to lead to a slight bump in performance in early iterations; see Section 6 for example illustrations.

Local regularization towards true labels. 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

$$\hat{\theta}_i^t = \lambda_i^t \frac{1}{d_i^0} \sum_{j \in N_i^0} \theta_j + (1 - \lambda_i^t) \frac{1}{d_i^1} \sum_{j \in N_i^1} \hat{\theta}_j^{t-1}, \tag{19}$$

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 (17) as

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

we see that we recover equation (19) from equation (17) when $\lambda_i^t = d_i^0 / d_i$, 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. Different 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.

Consider parameterizing $\lambda_i^t = f_i(t)$ for a function $f_i(\cdot)$ that reduces the number of parameters. For example one may consider the choice $\lambda_i^t = (d_i^0 / d_i)^t$, which represents exponential decay in t. This extension can be written in matrix-vector form as follows. Let $D_1 = \mathrm{diag}(d_i^1)$ and $D_0 = \mathrm{diag}(d_i^0)$ be diagonal degree matrices restricted to the unlabeled and labeled node sets, respectively. Consider the transition matrices $T_{10} = (D_1^{-1} A)_{10}$ and $T_{11} = (D_0^{-1} A)_{11}$. Then we may write the update rule as

$$\hat{\theta}_1^t = \lambda_t T_{10} \theta_0 + (1 - \lambda_t) T_{11} \hat{\theta}_1^{t-1},$$

where $\lambda_t \in [0, 1]^U$ is a vector of penalty parameters for time t. Depending on the form of λ_t, this recursion may not reduce cleanly to a power series that can be written in closed form.

This approach is a form of regularization in the sense that it constrains the complexity of the fitting function θ and restricts how far it can deviate from the observed values θ_0.

The choice of weight $\lambda_i^t = (d_i^0 / d_i)^t$ is related to the sequence of weights commonly used on the popular personalized PageRank diffusion kernel for community detection [2], although it is used in a different way. While personalized PageRank uses the kernel to model decaying influence as a function of graph distance, here we use this kernel to define how much to regularize the solution towards labeled values as a function of time, thus representing the amount of trust placed in labeled nodes. The limiting solution, therefore, will not be the same as the personalized PageRank diffusion, but this conceptual link can be instructive for suggesting analogs of other diffusion kernels that may be used as well [17, 18]. While the discussion of this regularization has centered on standard (one-hop) graph smoothing, the same ideas can be applied to decoupled smoothing as well.

5 AGGREGATING EXPERT INFORMATION

In this section we suggest possible ways of selecting the weights W for the method described in Section 3. Recall that in the decoupled smoothing model, equation (8), we view the target attributes ϕ_j as capturing information from the true responses in the neighborhood $\mathcal{N}_j$, which are useful for estimating θ_i. As a thought experiment, supposing the target parameters ϕ are known. Then the conditional mean (when $\gamma = 1$) is

$$\mathbf{E}[\theta_i | \phi] = \frac{1}{z_i} \sum_{j=1}^{n} W_{ij} \phi_j. \tag{20}$$

Most simply, it may be reasonable in many cases to take $W = A$ to agree with the network structure so that $\hat{\theta}_i$ is just an unweighted average of the estimated target parameters of the neighbors unit i. More generally, however, the row $\{W_{ij} : j = 1, \ldots, n\}$ of weights should be chosen in a way that highlights the nodes j where ϕ_j is most useful for prediction θ_i. In many graph contexts there are highly-skewed degree distributions, and as a result we expect some friends of i to provide more accurate estimates than others of i's true state. Notice that W_{ij} may be asymmetric in general because

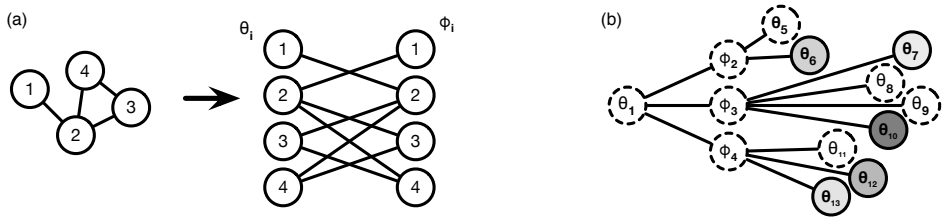


Figure 1: (a) An illustration of how an undirected graph can be decoupled into a bipartite graph with real responses θ_i and target responses ϕ_i. (b) An illustration of how two hop information can be used to estimate the expert opinions ϕ_j, which in turn inform the estimate of the unknown value θ_1.

the information provided by ϕ_i for estimating θ_j may not be the same as the information provided by ϕ_j for estimating θ_i, which is one of the key insights of the decoupled approach.

In this section we discuss how one may use an otherwise unrelated literature on expert opinion aggregation in the graph smoothing context. In this discussion we assume W has the same sparsity pattern as the adjacency matrix A, so that $W_{ij} \neq 0$ whenever $A_{ij} \neq 0$, although this need not generally be true; sometimes *not* being friends with someone can be highly informative as well.

The overall procedure of how a network is converted into a decoupled graph of real and target responses is illustrated schematically in Figure 1(a), and a schematic illustration of how the expert opinion aggregation pulls in information to a single node is illustrated in Figure 1(b).

5.1 Combining independent estimators

Consider that the information contributed by each expert (friend) j for estimating θ_i is in the form of the "observations" $\{\theta_k : k \in \mathcal{N}_j\}$, 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 [8, 11, 12, 14, 19, 23, 26].

Explicitly, suppose the variables in the set $\{\theta_k : k \in \mathcal{N}_j\}$ 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 (the "expert"). Then the weight matrix entry reduces to $W_{ij} = A_{ij}/\sigma_j^2$, with row sum $z_i = \sum_{\ell \in \mathcal{N}_i} \sigma_\ell^{-2}$ 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 3, the auxiliary graph with this diagonal covariance specification has an adjacency matrix with entries

$$\tilde{A}_{ij} = \sum_k A_{ik}A_{jk}/(d_k\sigma_k^2), \tag{21}$$

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:

$$\begin{aligned}\hat{\theta}_i^t &= (Z^{-1}\tilde{A}\hat{\theta}^{t-1})_i = \frac{1}{z_i}\sum_{j=1}^n \tilde{A}_{ij}\hat{\theta}_j^{t-1} \\ &= \frac{1}{\sum_{\ell\in\mathcal{N}_i}\sigma_\ell^{-2}} \sum_{k\in\mathcal{N}_i}\frac{1}{d_k\sigma_k^2}\sum_{j\in\mathcal{N}_k}\hat{\theta}_j^{t-1}. \end{aligned}\tag{22}$$

By viewing the aggregation as performed on a graph, we can in fact turn this standard estimation procedure into an iterative procedure, linking to smoothing as shown in Section 4.

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

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

is the sample variance of the labeled nodes in the neighborhood of j (recall that $\mathcal{N}_j^0$ is the labeled neighborhood and $d_j^0 = |\mathcal{N}_j^0|$ is the labeled degree). We then use $\hat{\sigma}_j^2$ as a plugin estimate in the update rule in Equation (22), giving

$$\hat{\theta}_i^t = \frac{1}{\sum_{\ell\in\mathcal{N}_i}(S_\ell^2/d_\ell)^{-1}}\sum_{k\in\mathcal{N}_i}\frac{1}{S_k^2}\sum_{j\in\mathcal{N}_k}\hat{\theta}_j^{t-1}. \tag{23}$$

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

$$\hat{\theta}_i^t = \frac{1}{\sum_{\ell\in\mathcal{N}_i} d_\ell}\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{N}_k}\hat{\theta}_j^{t-1}. \tag{24}$$

For exposition here we have let d_ℓ represent the total graph degree of unit ℓ, which disregards the number of labeled nodes. The shortcomings of this choice discussed in Section 4.4 still hold, and the improvements discussed in that section can be applied here.

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. The simple recurrence in equation (24) 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.

5.2 Risk analysis and Bayesian interpretation

This idea of combining information from different expert sources is also related to methods developed in the risk analysis literature [10]. Here we discuss graph analogs of the model and analysis of Halperin [15], later derived independently by Winkler [29]. In this model, expert opinions ϕ_j are viewed as unbiased samples from the true unknown response θ_i. That is, the expert predictions have the form

$$\phi_{\mathcal{N}_i} = \theta_i \mathbf{1} + \varepsilon_{\mathcal{N}_i},$$

where $\phi_{\mathcal{N}_i} \in \mathbb{R}^{d_i}$ is the restriction of the target vector ϕ to the neighborhood $\mathcal{N}_i$, $\mathbf{1}$ is the ones vector of length d_i, and $\varepsilon_{\mathcal{N}_i} \sim N(0, \Sigma_{\mathcal{N}_i})$ is Gaussian noise. The covariance matrix $\Sigma_{\mathcal{N}_i} \in \mathbb{R}^{d_i \times d_i}$ captures the variance and dependencies among the expert opinions, and is a generalization of the discussion in Section 5.1 in that expert opinions may be correlated with each other. Under this model and a flat prior on θ_i, the posterior distribution of θ_i given $\phi_{\mathcal{N}_i}$ with known covariance $\Sigma_{\mathcal{N}_i}$ is given by

$$\theta_i | \phi_{\mathcal{N}_i} \sim N\left(\frac{\mathbf{1}^\top \Sigma_{\mathcal{N}_i}^{-1} \phi_{\mathcal{N}_i}}{\mathbf{1}^\top \Sigma_{\mathcal{N}_i}^{-1} \mathbf{1}}, \frac{1}{\mathbf{1}^\top \Sigma_{\mathcal{N}_i}^{-1} \mathbf{1}} \right).$$

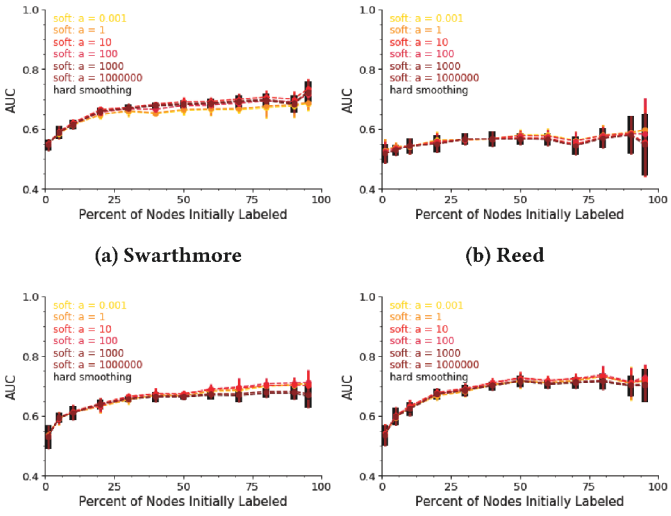
The Bayes estimator is then the posterior mean,

$$\hat{\theta} = (\mathbf{1}^\top \Sigma_{\mathcal{N}_i}^{-1} \phi_{\mathcal{N}_i}) / (\mathbf{1}^\top \Sigma_{\mathcal{N}_i}^{-1} \mathbf{1}).$$

From comparing to the conditional mean in Equation (20), the Bayes estimator suggests taking the weight W_{ij} to be element of the vector $(\mathbf{1}^\top \Sigma_{\mathcal{N}_i}^{-1})$ that corresponds to node j if i is connected to j, and $W_{ij} = 0$ otherwise. The normalization $z_i = \sum_{j \in \mathcal{N}_i} W_{ij} = \mathbf{1}^\top \Sigma_{\mathcal{N}_i}^{-1} \mathbf{1}$ then falls out naturally.

More general covariance structures. A principal outstanding question is determining which assumptions to place on each $\Sigma_{\mathcal{N}_i}$, which are unknown in practice. We notice that determining the $\Sigma_{\mathcal{N}_i}$ is equivalent to defining a single covariance matrix for the expert opinions $\Sigma \in \mathbb{R}^{n \times n}$ for the entire graph, where Σ_{jk} represents the covariance between the expert opinions provided by individuals j and k. The weight matrix results by taking $W = A\Sigma^{-1}$. The specification of the covariance matrix Σ is crucial for the form of the resulting decoupled smoothing estimator.

In practice on social networks, dependencies between expert opinions are bound to occur due to intersecting friendship neighborhoods. if the estimation target is i, and two neighbors j and k of i share many mutual friends other than i, then their estimates of θ_i will be dependent. The presence of such correlations indicates that aggregation and smoothing methods that are able to harness these arbitrary covariance matrices may yield a powerful tool for node classification in some application areas, and we leave a detailed study of such extensions to future work.



(a) Swarthmore (b) Reed

(c) Amherst (d) Haverford

Figure 2: Predicting gender based on soft smoothing. Horizontal axis displays initial proportion of unlabeled nodes and vertical axis displays AUC scores. Soft smoothing is essentially identical in performance to hard smoothing for the gender classification task.

6 EMPIRICAL ILLUSTRATIONS

We perform experiments on a sample of undergraduate college networks collected from a single-day snapshot of Facebook in September 2005 [27, 28]. 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.

We varied the percentage of initially labelled nodes by selecting a labelled sample uniformly at random. We trained our models varying the percentage of initially labelled nodes in the network. For a given fixed percent of labelled individuals (training dataset), we measure classification performance (by AUC) on the remaining unlabelled nodes (testing dataset), using the same train/test splits across the different inference methods. For all plots in this section we attempt classification 10 times based on different independent labelled subsets of nodes using stratified sampling (stratified based on gender). We show the average AUC with error bars denoting the standard deviation across the 10 runs. We also provide results for synthetic networks generated from the stochastic block model [16] and the overdispersed stochastic block model [1].

6.1 Soft smoothing

We first run our soft smoothing algorithm on the four Facebook networks. As shown in Section 4, soft smoothing is equivalent to hard smoothing on a modified graph in which each originally labeled node has its label removed and is instead connected to a new degree-one node via an edge with weight $\alpha > 0$. Here we

choose a range of different α values to study its influence on the performance of soft smoothing method. The baseline algorithm is the hard smoothing approach of [33], which is the solution to Equation (15). Observe that α quantifies the impact of the new attached nodes on the original labeled nodes; therefore, the larger the value of α, the closer the soft smoothing result resembles the hard smoothing result.

In our experiments we observe that the outputs from soft smoothing are essentially identical to those from hard smoothing. As Figure 2 shows, since our goal is to classify gender, which is self-reported as a binary variable, adding a "noise" term to the labeled nodes has relatively little effect. Our hypothesis is that the soft smoothing method may achieve better performance when applied to real-valued outcomes.

6.2 Decoupled smoothing

In Figure 3 we see our experiments with decoupled smoothing, which indicate that the two-hop majority vote update given by Equation (24) outperforms both the standard 1-hop majority vote estimator and the corresponding (ZGL) smoothing estimator in terms of AUC, regardless of the percentage of initially labeled nodes. Meanwhile we also observe that decoupled smoothing performs slightly worse than the much simpler 2-hop majority vote estimator in some situations (namely `Amherst` and `Haverford`). Recall from Section 4.4 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 (gender prediction) than global information, and the smoothing algorithms are inappropriately synthesizing information from local and global sources. This hypothesis motivates the improvements discussed in Section 4.4, the results for which are discussed in the following section.

6.3 Regularized iterations

In Section 4.4 we considered a modified iterated majority vote algorithm that includes a local regularization $\lambda_i^t = (d_i^L/d_i)^t$ 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 equation (18) as the first iteration's update rule instead of (17) 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 block model graph as well as the on the Facebook networks.

6.3.1 Improved iterative hard smoothing. We first display results on a graph sampled from a stochastic block model (SBM). The particular SBM we use has two blocks with 500 nodes in each block, representing 500 male and 500 female individuals. The expected average degree is 42, so as to achieve the same edge density as used in [1]. In Figure 4 we plot the AUC corresponding to the predictions given by equation (14) for each iteration, $t = 1, 2, 3, \ldots$. Interestingly, in Figure 4(a), we see that some early iterations (in this case, iteration 2) actually outperform the limiting ZGL/hard smoothing solution. This suggests that improvements can be made

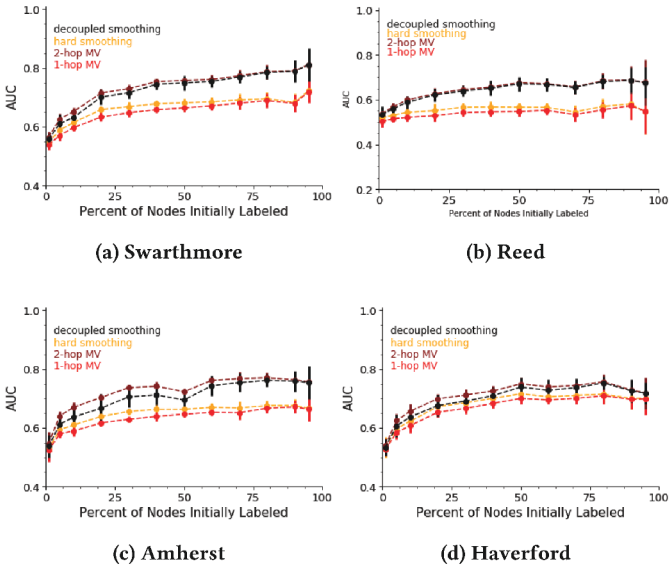


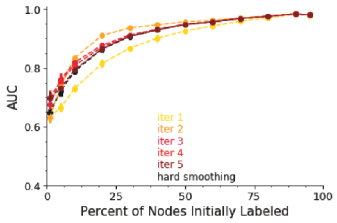
Figure 3: Predicting gender based on decoupled smoothing, compared with hard smoothing (ZGL) and 1-hop and 2-hop majority vote. The estimators based on two-step neighborhood information clearly outperform those based on one-step information, but 2-hop majority vote sometimes outperforms decoupled smoothing.

by teaching the classifier to hone in on these successful iterations and give them more weight, which is what the regularizer seeks to do. We next compare the original and regularized versions of the hard smoothing iterations. As shown in Figure 4(b), the regularization proposed in Section 4.4 improves the overall prediction accuracy for hard smoothing under the stochastic block model. The method is essentially converged after five iterations.

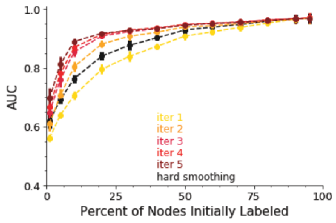
Results are more mixed on empirical networks; in Figure 4(c-d) we display performance for the `Amherst` Facebook network. We see that regularization does not significantly improve the performance over hard smoothing for gender prediction on this network.

6.3.2 Improved iterative decoupled smoothing. Our regularization ideas can also be applied to the decoupled smoother. 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 [1]. Here we compare the iterative method results for the original decoupled smoothing method against the regularized iterative decoupled smoothing method. As shown in Figure 5b, the regularization improves the overall prediction accuracy for decoupled smoothing under the overdispersed stochastic block model.

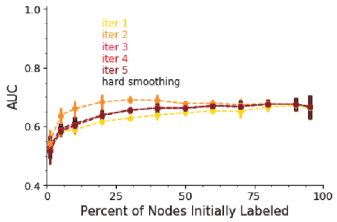
On the Facebook `Amherst` network we use the regularization $\lambda_i^t = (d_i^0/(rd_i))^{t-1}$, where r is the initial percent of labeled nodes.



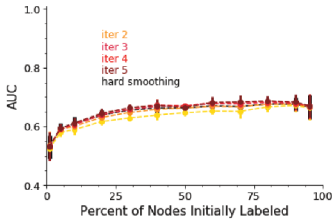
(a) SBM: hard smoothing



(b) SBM: regularized hard smoothing

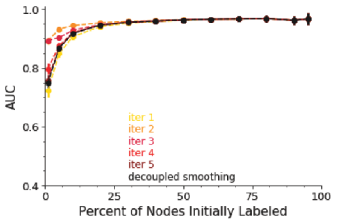


(c) Amherst: hard smoothing

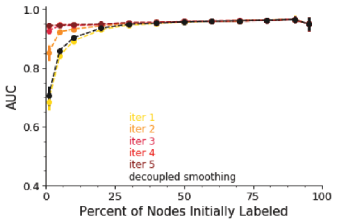


(d) Amherst: regularized hard smoothing

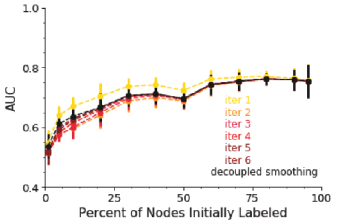
Figure 4: Hard smoothing iterations, with and without regularization, for classifying gender on an SBM graph and the `Amherst` dataset. Regularization mostly improves the predictive performance.



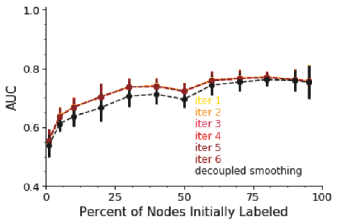
(a) oSBM: decoupled smoothing



(b) oSBM: regularized decoupled smoothing



(c) Amherst: decoupled smoothing



(d) Amherst: regularized decoupled smoothing

Figure 5: Decoupled smoothing iterations, with and without regularization, for classifying gender on an oSBM and the `Amherst` network. Regularization mostly improves the predictive performance.

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 way λ_i^t is parameterized, it is possible that performance can be further improved.

7 DISCUSSION

In this work we investigate the use of graph-based smoothing for node attribute prediction on 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 observations in [1], which highlights the distinction between "the company you keep" and "the company you're kept in." We develop a model for what we call *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 iterated 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 best choices for regularization has yet to be fully explored and may well vary depending on the particular domain of application.

Graph-based semi-supervised learning is broadly applied to classification and prediction problems on derived graphs, typically from the nearest-neighbor graphs of point clouds in $\mathbb{R}^n$. In this work, we observe that classification and prediction problems involving explicit graph structure, such as predicting node attributes in social networks, can benefit from a careful consideration of how the prediction target quantity and the graph structure may be related. While our evaluation only touches gender, other attribute prediction tasks in social networks such as predicting abusive or spam accounts could also benefit from similar careful consideration. While our empirical improvements for gender are modest, our results on synthetic graphs suggest that decoupled graph smoothing has the potential to offer considerable improvements in other diverse network prediction tasks.

Code availability. Notebooks and code available online: https://github.com/YatongChen/decoupled_smoothing_on_graphs. All code was run using Python version 3.6.2.

Acknowledgements. This work was supported in part by NSF grant IIS-1657104 and an ARO Young Investigator Award. KMA was supported in part by a National Defense Science and Engineering Graduate Fellowship.

REFERENCES

[1] Kristen M Altenburger and Johan Ugander. 2018. Monophily in social networks introduces similarity among friends-of-friends. *Nature Human Behaviour* 2, 4 (2018), 284.

[2] Reid Andersen, Fan Chung, and Kevin Lang. 2006. Local graph partitioning using PageRank vectors. In *null*. IEEE, 475–486.

[3] Shumeet Baluja, Rohan Seth, D Sivakumar, Yushi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravichandran, and Mohamed Aly. 2008. Video suggestion and discovery for YouTube: taking random walks through the view graph. In *Proceedings of the 17th International Conference on World Wide Web*. ACM, 895–904.

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

[5] Julian Besag. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)* (1974), 192–236.

[6] Julian Besag, Jeremy York, Annie Mollié, et al. 1991. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 43, 1 (1991), 1–20.

[7] Smriti Bhagat, Graham Cormode, and S Muthukrishnan. 2011. Node classification in social networks. In *Social Network Data Analytics*. Springer, 115–148.

[8] Leo Breiman. 1996. Stacked regressions. *Machine Learning* 24, 1 (1996), 49–64.

[9] D Brook. 1964. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika* 51, 3/4 (1964), 481–483.

[10] Robert T Clemen and Robert L Winkler. 1999. Combining probability distributions from experts in risk analysis. *Risk Analysis* 19, 2 (1999), 187–203.

[11] William G Cochran. 1937. Problems arising in the analysis of a series of similar experiments. *Supplement to the Journal of the Royal Statistical Society* 4, 1 (1937), 102–118.

[12] William G Cochran. 1954. The combination of estimates from different experiments. *Biometrics* 10, 1 (1954), 101–129.

[13] Ahmed El Alaoui, Xiang Cheng, Aaditya Ramdas, Martin J Wainwright, and Michael I Jordan. 2016. Asymptotic behavior of\ell_p-based laplacian regularization in semi-supervised learning. In *Conference on Learning Theory*. 879–906.

[14] William R Fairweather. 1972. A method of obtaining an exact confidence interval for the common mean of several normal populations. *Applied Statistics* (1972), 229–233.

[15] Max Halperin. 1961. Almost linearly-optimum combination of unbiased estimates. *J. Amer. Statist. Assoc.* 56, 293 (1961), 36–43.

[16] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. 1983. Stochastic blockmodels: First steps. *Social networks* 5, 2 (1983), 109–137.

[17] Kyle Kloster and David F Gleich. 2014. Heat kernel based community detection. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1386–1395.

[18] Isabel M Kloumann, Johan Ugander, and Jon Kleinberg. 2017. Block models and personalized PageRank. *Proceedings of the National Academy of Sciences* 114, 1 (2017), 33–38.

[19] Frédéric Lavancier and Paul Rochet. 2016. A general procedure to combine estimators. *Computational Statistics & Data Analysis* 94 (2016), 175–192.

[20] Tianxi Li, Elizaveta Levina, and Ji Zhu. 2016. Prediction models for network-linked data. *arXiv preprint arXiv:1602.01192* (2016).

[21] Sofus A Macskassy and Foster Provost. 2007. Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research* 8, May (2007), 935–983.

[22] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 1 (2001), 415–444.

[23] Ronny Meir. 1995. Bias, variance and the combination of least squares estimators. In *Advances in Neural Information Processing Systems*. 295–302.

[24] Boaz Nadler, Nathan Srebro, and Xueyuan Zhou. 2009. Semi-supervised learning with the graph Laplacian: The limit of infinite unlabelled data. (2009).

[25] Leto Peel. 2016. Graph-based semi-supervised learning for relational networks. *arXiv preprint arXiv:1612.05001* (2016).

[26] JNK Rao and Kathleen Subrahmaniam. 1971. Combining independent estimators and estimation in linear regression with unequal variances. *Biometrics* (1971), 971–990.

[27] Amanda L Traud, Eric D Kelsic, Peter J Mucha, and Mason A Porter. 2011. Comparing community structure to characteristics in online collegiate social networks. *SIAM review* 53, 3 (2011), 526–543.

[28] Amanda L Traud, Peter J Mucha, and Mason A Porter. 2012. Social structure of Facebook networks. *Physica A: Statistical Mechanics and its Applications* 391, 16 (2012), 4165–4180.

[29] Robert L Winkler. 1981. Combining probability distributions from dependent information sources. *Management Science* 27, 4 (1981), 479–488.

[30] Ya Xu, Justin S Dyer, and Art B Owen. 2010. Empirical stationary correlations for semi-supervised learning on graphs. *The Annals of Applied Statistics* (2010), 589–614.

[31] Denny Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. 2004. Learning with local and global consistency. In *Advances in neural information processing systems*. 321–328.

[32] Xiaojin Zhu. 2006. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison* 2, 3 (2006), 4.

[33] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*. 912–919.

APPENDIX

1 PROOF OF GMRF EQUIVALENCE

Theorem 3.1. *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

$$\Sigma = \tau^2 \begin{pmatrix} Z & -\gamma W \\ -\gamma W^\top & Z' \end{pmatrix}^{-1}, \tag{9}$$

where $Z = \mathrm{diag}(z_1, \ldots, z_n)$ and $Z' = \mathrm{diag}(z'_1, \ldots, z'_n)$.

Proof. We use Brook's lemma [9], which states that for any distribution such that $p(x) > 0$ for all x, then for any x and x',

$$\frac{p(x)}{p(x')} = \prod_{i=1}^{n} \frac{p(x_i|x_1, \ldots, x_{i-1}, x'_{i+1}, \ldots, x'_n)}{p(x'_i|x_1, \ldots, x_{i-1}, x'_{i+1}, \ldots, x'_n)}.$$

Applying Brook's lemma with $x = (\theta, \phi)$ and $x' = 0$, we see that

$$\begin{aligned} p(\theta, \phi) &\propto \prod_{i=1}^{n} \Bigg(\frac{p(\theta_i|\theta_1, \ldots, \theta_{i-1}, 0_{i+1}, \ldots, 0_{2n})}{p(0_i|\theta_1, \ldots, \theta_{i-1}, 0_{i+1}, \ldots, 0_{2n})} \times \\ &\qquad \frac{p(\phi_i|\theta_1, \ldots, \theta_n, \phi_1, \ldots, \phi_{i-1}, 0_{i+1}, \ldots, 0_{2n})}{p(0_i|\theta_1, \ldots, \theta_n, \phi_1, \ldots, \phi_{i-1}, 0_{i+1}, \ldots, 0_{2n})} \Bigg) \\ &= \prod_{i=1}^{n} \frac{p(\theta_i|\phi = 0)}{p(\theta_i = 0|\phi = 0)} \frac{p(\phi_i|\theta)}{p(\phi_i = 0|\theta)} \end{aligned}$$

where we have indexed the 0s for clarity. Substituting in the likelihood from the conditional models in Theorem 3.1,

$$\begin{aligned} p(\theta, \phi) &\propto \frac{\exp\left\{-\frac{z_i}{2\tau^2}\theta_i^2\right\}}{1} \frac{\exp\left\{-\frac{z'_i}{2\tau^2}\left(\phi_i - \frac{\gamma}{z'_i}\sum_j W_{ji}\theta_j\right)^2\right\}}{\exp\left\{-\frac{z'_i}{2\tau^2}\left(\frac{\gamma}{z'_i}\sum_j W_{ji}\theta_j\right)^2\right\}} \\ &= \exp\left\{-\frac{1}{2\tau^2}\sum_{i=1}^{n}\left(z_i\theta_i^2 + z'_i\phi_i^2 - 2\gamma\phi_i\sum_j W_{ji}\theta_j\right)\right\} \\ &= \exp\left\{-\frac{1}{2\tau^2}\left(\theta^\top Z\theta + \phi^\top Z'\phi - 2\gamma\theta^\top W\phi\right)\right\}, \end{aligned}$$

which is the likelihood corresponding to a mean zero Gaussian random vector with covariance given by Equation (9). □