Journal of Machine Learning Research 14 (2013) 3721-3752 Submitted 4/11; Revised 5/13; Published 12/13

Joint Harmonic Functions and Their Supervised Connections

Mark Vere Culp MVCULP@MAIL.WVU.EDU
Kenneth Joseph Ryan KJRYAN@MAIL.WVU.EDU
Department of Statistics
West Virginia University
Morgantown, WV 26506, USA

Editor: Sridhar Mahadevan

Abstract

The cluster assumption had a significant impact on the reasoning behind semi-supervised classification methods in graph-based learning. The literature includes numerous applications where harmonic functions provided estimates that conformed to data satisfying this well-known assumption, but the relationship between this assumption and harmonic functions is not as well-understood theoretically. We investigate these matters from the perspective of supervised kernel classification and provide concrete answers to two fundamental questions. (i) Under what conditions do semi-supervised harmonic approaches satisfy this assumption? (ii) If such an assumption is satisfied then why precisely would an observation sacrifice its own supervised estimate in favor of the cluster? First, a harmonic function is guaranteed to assign labels to data in harmony with the cluster assumption if a specific condition on the boundary of the harmonic function is satisfied. Second, it is shown that any harmonic function estimate within the interior is a probability weighted average of supervised estimates, where the weight is focused on supervised kernel estimates near labeled cases. We demonstrate that the uniqueness criterion for harmonic estimators is sensitive when the graph is sparse or the size of the boundary is relatively small. This sets the stage for a third contribution, a new regularized joint harmonic function for semi-supervised learning based on a joint optimization criterion. Mathematical properties of this estimator, such as its uniqueness even when the graph is sparse or the size of the boundary is relatively small, are proven. A main selling point is its ability to operate in circumstances where the cluster assumption may not be fully satisfied on real data by compromising between the purely harmonic and purely supervised estimators. The competitive stature of the new regularized joint harmonic approach is established.

Keywords: harmonic function, joint training, cluster assumption, semi-supervised learning

1. Introduction

The problem under consideration is semi-supervised learning in the graph-based setting. Observations are vertices on a graph, and edges provide similarity associations between vertices. Classification is required if vertices are labeled and the goal is to design a function to predict the labels. Local classifiers like k-NN or more generally kernel regression are ideal in the graph-based setting since they can operate directly on the similarity matrix and do not require $\boldsymbol{X}$-data support (Chapelle et al., 2006b; Lafferty and Wasserman, 2007). On the other hand, methods of prediction for observations without labels are arguably more complicated and less understood than those from classical supervised settings. A vertex corresponding to an observation without a label provides connections through it which are meaningful to the data structure, and unlabeled data increase

performance if used during training (Culp et al., 2009). The need to extend locally smooth functions into this graph-based setting is an important problem (Chapelle et al., 2006b; Abney, 2008). Applications of graph-based learning include text classification (McCallum et al., 2000), protein interaction (Yamanishi et al., 2004; Kui et al., 2002), chemogenomics in pharmaceuticals (Bredel and Jacoby, 2004), biology and chemistry networks (Lundblad, 2004; Culp et al., 2009), and web data/email (Koprinska et al., 2007). There are also applications where the edges of the graph were constructed using a similarity function generated from feature data (Carreira-Perpiñán and Zemel, 2005; Chapelle et al., 2006b; Jebara et al., 2009).

Harmonic functions provide a natural solution to the problem of extending local classifiers into semi-supervised learning. The definition of a harmonic function depends on two key terms, that is, the boundary (observed labels) and the interior (unlabeled). The boundary choice defines the harmonic function. With a given function estimate on the boundary, the harmonic solution achieves an equilibrium on the interior. Each interior case is an average of its and its neighbors' estimates, so an estimate for an interior observation does not change if averaged a second time. Currently, the authors are aware of only one harmonic approach in the semi-supervised literature. This estimator, referred to as the *clamped harmonic estimator*, sets the boundary equal to its observed labeling. The clamped harmonic estimator in semi-supervised learning was studied and applied to energy optimization (Chapelle et al., 2006b; Abney, 2008), graph-based smoothing (Culp et al., 2009), Gaussian processes (Zhu, 2008), iterative algorithms with large data (Subramanya and Bilmes, 2011), stability methods for transductive learning (Cortes et al., 2008), and other areas (Zhu and Goldberg, 2009).

The clamped harmonic estimator has known shortcomings. First, its performance degradation due to sensitivity to noise in either the support or labeling is well-known. Also, there is no way to estimate a residual, which renders the smoothing technique impossible to use for any inferential analysis, outlier detection, or descriptive analysis. Recent work suggests that the clamped harmonic solution also suffers in circumstances where the size of the boundary is much smaller than that of the interior. The main argument is that the harmonic solution converges to the zero function with spikes within the boundary as the size of the interior grows (Nadler et al., 2009; von Luxburg et al., 2010).

Applications where semi-supervised learning has solid performance as well as an abstraction of such applications into a set of mathematical assumptions is of recent interest (Lafferty and Wasserman, 2007; Azizyan et al., 2013). It is fairly well understood in semi-supervised learning that if two points x_1, x_2 are close in the intrinsic geometry of the probability distribution of X then learning can occur if the conditional probability distributions of $y \mid x_1$ and $y \mid x_2$ are similar. Such a characterization is commonly assumed in semi-supervised learning and often referred to as the *cluster assumption* (Chapelle et al., 2006b). Optimization problems involving minimax error bounds under the cluster and other similar smoothness assumptions is of recent interest (Rigollet, 2007; Lafferty and Wasserman, 2007; Singh et al., 2008). Lafferty and Wasserman (2007) further note the importance of separating semi-supervised smoothness assumptions from other seemingly similar assumptions in manifold learning (Hein et al., 2005; Aswani et al., 2010). The clamped harmonic estimator has been empirically validated to satisfy the cluster assumption, but this, to our knowledge, has not been established rigorously. A key contribution of this work is a condition on the boundary for when any harmonic function is guaranteed to satisfy the cluster assumption.

How semi-supervised approaches compare to supervised alternatives is a looming and important question. In the case of harmonic functions, we are primarily interested in articulating how these

approaches compare to supervised local smoothing classifiers. A significant contribution of this work is extensive analysis and development of harmonic functions in this capacity. In this regard, we show that any harmonic function, no matter how the boundary estimator is generated, can be decomposed as the reweighted average of soft local supervised estimates consisting only of unlabeled predictions. Specifically, the estimate for an interior observation is a weighted average of all the interior local supervised estimates. This work further establishes that interior observations nearest to the boundary carry the weight in the prediction of interior cases.

Harmonic functions and supervised local estimators each use two types of information that describe relationships between the boundary states (labeled) and the interior states (unlabeled). The first type, which we term *labeled adjacent*, involves direct kernel weighted distances from an unlabeled observation to each labeled observation/case. Local supervised approaches essentially form a weighted average of this labeled adjacent information even when an unlabeled case has small adjacency to each labeled case. The second type of information, which we term *labeled connective*, exploits interconnectivity within unlabeled cases to find other unlabeled cases that have stronger adjacency to labeled cases. Harmonic functions propagate the local supervised estimates from unlabeled cases with strong adjacency to some labeled cases to the other unlabeled cases. In short, harmonic functions in semi-supervised learning are purely labeled connective, while local supervised approaches are purely labeled adjacent.

Another key contribution of this work is a new harmonic function approach based off of a joint optimization criterion. The novel use of the joint optimization criterion allows for regularization within semi-supervised learning. Settings of a single regularization parameter can reproduce the extremes, that is, a labeled connective harmonic function estimator or the labeled adjacent soft local supervised estimator, but can also be tuned to any one of a continuum of semi-supervised estimators to compromise between the extremes. It is the only estimator to our knowledge that has been shown to balance between supervised learning and semi-supervised learning in this manner. The benefits of regularization in joint harmonic estimation are empirically assessed with strong results.

The paper is organized as follows. After a brief description of notational conventions in Section 2, the problem is formulated in Section 3. Care is taken to succinctly describe semi-supervised block matrix results in terms of their supervised counterparts, so the stage is set for our main contributions. General results on harmonic functions with regard to the cluster assumption and supervised learning are in Section 4. Section 5 includes the definition of the new regularized joint harmonic function approach and characterization of its mathematical properties. Sections 6 and 7 include empirical tests of the new approach. Section 8 has concluding remarks, and a proof of each Lemma, Proposition, and Theorem is in Appendix A.

2. Notational Conventions

It is common to let $\boldsymbol{A}_{ij}$ represent the entry of a matrix $\boldsymbol{A}$ in row i and column j. A generalization of this $\boldsymbol{A}_{ij}$ notation that is particularly useful in semi-supervised learning is to replace i and j with a list of rows and columns to represent the corresponding sub matrix, so if matrix $\boldsymbol{A}$ is $n \times n$ and sets $L = \{1,2,\ldots,l\}$ and $U = \{l+1,l+2,\ldots,n\}$, then

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{LL} & \boldsymbol{A}_{LU} \\ \boldsymbol{A}_{UL} & \boldsymbol{A}_{UU} \end{pmatrix}. \tag{1}$$

The usefulness of Partitioning (1) will become clear when attention turns back to discussion of the sets of labeled L and unlabeled U cases in the semi-supervised learning context of Section 3. Denote

$$\boldsymbol{A}^*_{LL} = \boldsymbol{A}_{LL} - \boldsymbol{A}_{LU}\boldsymbol{A}^{-1}_{UU}\boldsymbol{A}_{UL} \quad (\boldsymbol{A}_{UU} \text{ Block Schur Complement of } \boldsymbol{A}). \tag{2}$$

Note the important distinction between $\boldsymbol{A}_{LL}$ in Display (1) and $\boldsymbol{A}^*_{LL}$ in Display (2). Schur complements and some of their most basic properties given in Remark 1 play a key role in the methods to come as well as in the Appendix A proofs. Table 1 summarizes all of our matrix algebra conventions for future reference.

Notation	Definition
$\mathcal{N}(\boldsymbol{A})$	Null space of matrix $\boldsymbol{A}$.
$\boldsymbol{A} \geq 0$	Matrix $\boldsymbol{A}$ with all nonnegative entries ($>$ for positive).
$\boldsymbol{A} \succeq 0$	Positive semi-definite symmetric matrix $\boldsymbol{A}$ ($\succ$ for positive definite).
$\rho^{(i)}(\boldsymbol{A})$	ith largest modulus of the eigenvalues of a square matrix $\boldsymbol{A}$.
$\rho(\boldsymbol{A})$	Spectral radius of a square matrix $\boldsymbol{A}$, that is, $\rho(\boldsymbol{A}) = \rho^{(1)}(\boldsymbol{A})$.
$\boldsymbol{A}_{LL}$	Upper-left sub matrix in Partitioning (1) of a square matrix $\boldsymbol{A}$.
$\boldsymbol{A}^*_{LL}$	$\boldsymbol{A}_{UU}$ Block Schur Complement (2) of matrix $\boldsymbol{A}$ with Partitioning (1).

Table 1: List of notational conventions.

Remark 1 *Based on the Partitioning (1), it is well known that if $\boldsymbol{A}_{UU}$ is invertible then $\boldsymbol{A}$ is invertible if and only if $\boldsymbol{A}^*_{LL}$ is invertible. In the case that $\boldsymbol{A} \succeq 0$ (i.e., $\boldsymbol{A}$ is symmetric and positive semi-definite), this result becomes if $\boldsymbol{A}_{UU} \succ 0$ then $\boldsymbol{A} \succ 0$ if and only if $\boldsymbol{A}^*_{LL} \succ 0$.*

3. Problem Set-Up

In graph-based semi-supervised learning, partially labeled data are in the form of a weighted graph. Vertices $\{1,\cdots,n\}$ represent the n observations, and edges the values of a correspondence between each pair of observations. The $n \times n$ symmetric matrix $\boldsymbol{W}$ with $\boldsymbol{W}_{ij} \geq 0$ is the adjacency matrix of the *weighted graph* $(\{1,\cdots,n\}, \boldsymbol{W})$ or graph $\boldsymbol{W}$ for brevity. For this particular weighted graph, additionally assume $\boldsymbol{W}_{ij} \leq 1$ and $\boldsymbol{W}_{ii} = 1$. In some applications, $\boldsymbol{W}$ must be constructed from an $n \times p$ data matrix $\boldsymbol{X}$, for example,

$$\boldsymbol{W}_{ij} = K_\lambda(x_i, x_j),$$

where kernel function $K_\lambda(x_i, x_j)$ is applied to each pair of rows of $\boldsymbol{X}$ to form $\boldsymbol{W}$. Experimental Sections 6 and 7 include examples of each type, that is, $\boldsymbol{W}$ observed directly and $\boldsymbol{W}$ generated from $\boldsymbol{X}$. For now, simply assume that the symmetric matrix $\boldsymbol{W}$ is in hand.

The *training response* is

$$Y(Y_U) = \begin{pmatrix} Y_L \\ Y_U \end{pmatrix} \in \mathbb{R}^n, \text{ where } Y_U \in \mathbb{R}^{|U|}, \tag{3}$$

and the data partition into two observed subsets $\{1,\cdots,n\} = L \cup U$. Subset L is the set of all *boundary states*, whereas U is that for *interior states*. The subsets are distinguished by the labeling

function. The boundary states have an observed labeling vector Y_L, while the labelings for the interior states go unobserved. We assert the missing at random assumption and assume that L was initially a random subset of $\{1,\cdots,n\}$, but for ease of notation, the data were subsequently sorted so that boundary observations are first in the indexing. The vector of latent variables Y_U is comprised of the unknown labelings for the interior. Our joint optimization based method defined later in Section 5 involves the training response. The solution to this joint optimization problem provides the capacity for transductive or semi-supervised learning as will be illustrated later in Section 7.

Next, general graph theory results are discussed and applied to graph $\boldsymbol{W}$. In particular, Laplacian and stochastic smoother matrices corresponding to graph $\boldsymbol{W}$ are defined, and the relationships between these three matrices are discussed briefly. It is fundamental to think about the general idea being applied to graph $\boldsymbol{W}$ because later they will be applied to a particular graph with vertex set L in each of the Sections 3.1-3.3. These three graphs on L to be introduced in Sections 3.1-3.3 help one understand a semi-supervised technique through a decomposition of L to L connectivities in the larger graph $\boldsymbol{W}$ on $L \cup U$.

The *Laplacian* of $\boldsymbol{W}$ is $\boldsymbol{\Delta} = \boldsymbol{D} - \boldsymbol{W}$, where $\boldsymbol{D} = \text{diag}(\boldsymbol{W}\vec{1})$ is the *degree matrix* of $\boldsymbol{W}$. Proposition 7 is a well-known result on $\boldsymbol{\Delta}$ (Belkin et al., 2006).

Proposition 7 *Laplacian* $\boldsymbol{\Delta} \succeq 0$.

The square matrix $\boldsymbol{S} = \boldsymbol{D}^{-1}\boldsymbol{W}$ is a *stochastic smoother*, that is, $\boldsymbol{S} \geq 0$ and $\boldsymbol{S}\vec{1} = \vec{1}$, so 1 is an eigenvalue of $\boldsymbol{S}$. Proposition 8 further establishes that $\rho(\boldsymbol{S}) = 1$.

Proposition 8 *If* $\boldsymbol{W} \succeq 0$ *then each eigenvalue of* $\boldsymbol{S} = \boldsymbol{D}^{-1}\boldsymbol{W}$ *is an element of* $[0,1]$.

The identity $\boldsymbol{\Delta} = \boldsymbol{D}(\boldsymbol{I} - \boldsymbol{S})$ helps demonstrate that

$$\boldsymbol{\Delta}\nu = \vec{0} \iff \boldsymbol{S}\nu = \nu, \tag{4}$$

that is, $\mathcal{N}(\boldsymbol{\Delta})$ equals the eigenspace of $\boldsymbol{S}$ corresponding to eigenvalue 1. An eigenvalue decomposition of $\boldsymbol{\Delta}$ or $\boldsymbol{S}$ provides a way to compute the number of connected components in graph $\boldsymbol{W}$. One simply counts the multiplicity of eigenvalue 0 for $\boldsymbol{\Delta}$ by Remark 2 or equivalently eigenvalue 1 for $\boldsymbol{S}$ by Display (4).

The graphs in Sections 3.1-3.3 are based on partitioning the adjacency, stochastic smoother, and Laplacian matrices of graph $\boldsymbol{W}$ by L and U. Using Section 2 notation and Display (1) in particular, this is

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_{LL} & \boldsymbol{W}_{LU} \\ \boldsymbol{W}_{UL} & \boldsymbol{W}_{UU} \end{pmatrix}, \quad \boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{LL} & \boldsymbol{S}_{LU} \\ \boldsymbol{S}_{UL} & \boldsymbol{S}_{UU} \end{pmatrix}, \quad \boldsymbol{\Delta} = \begin{pmatrix} \boldsymbol{\Delta}_{LL} & \boldsymbol{\Delta}_{LU} \\ \boldsymbol{\Delta}_{UL} & \boldsymbol{\Delta}_{UU} \end{pmatrix}. \tag{5}$$

The entries of $\boldsymbol{W}_{LL}$ and $\boldsymbol{W}_{UU}$ are similarities within the boundary and interior, respectively, while $\boldsymbol{W}_{LU} = \boldsymbol{W}_{UL}^T$ contain the similarities between boundary and interior observations. Analogous interpretations extend to the other matrices partitioned in Display (5). For the diagonal degree matrix $\boldsymbol{D} \geq 0$, define the $|L| \times |L|$ diagonal matrices $\widetilde{\boldsymbol{D}}_{LL} = \text{diag}(\boldsymbol{W}_{LL}\vec{1}) \geq 0$ and $\widetilde{\boldsymbol{D}}_{LU} = \text{diag}(\boldsymbol{W}_{LU}\vec{1}) \geq 0$ and the $|U| \times |U|$ diagonal matrices $\widetilde{\boldsymbol{D}}_{UU} = \text{diag}(\boldsymbol{W}_{UU}\vec{1}) \geq 0$ and $\widetilde{\boldsymbol{D}}_{UL} = \text{diag}(\boldsymbol{W}_{UL}\vec{1}) \geq 0$, so that

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_{LL} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_{UU} \end{pmatrix} = \begin{pmatrix} \widetilde{\boldsymbol{D}}_{LL} + \widetilde{\boldsymbol{D}}_{LU} & \boldsymbol{0} \\ \boldsymbol{0} & \widetilde{\boldsymbol{D}}_{UL} + \widetilde{\boldsymbol{D}}_{UU} \end{pmatrix}.$$

Next, supervised, offset, and semi-supervised weighted graphs are studied in Sections 3.1-3.3 to assist in a deep understanding of a semi-supervised boundary estimation method.

Remark 2 *Vertices i and j are adjacent in graph $\boldsymbol{W}$ if $\boldsymbol{W}_{ij} > 0$, and are connected if there exists a sequence of vertices starting with i and ending with j such that consecutive vertices throughout the sequence are adjacent. The concept of connectedness partitions the vertices into some number of connected components, and each vertex in a connected component is connected to any other vertex in that component. Basic structure of a weighted graph includes the number of connected components and whether or not any given pair of vertices is in the same connected component. Both of these properties are encoded in particular eigenvectors of the graph's Laplacian matrix and stochastic smoother. Just take the binary vector in $\mathbb{R}^n$ that indicates observations in a connected component of $\boldsymbol{W}$. The set of all such binary vectors over all connected components is an orthogonal basis for $\mathcal{N}(\boldsymbol{\Delta})$, so the dimension of $\mathcal{N}(\boldsymbol{\Delta})$ equals the number of connected components. Furthermore, it is obvious that the vectors in this basis sum to $\vec{1} \in \mathcal{N}(\boldsymbol{\Delta})$.*

3.1 The Supervised Case

The supervised local kernel smoother at any point x_i is

$$\widetilde{f}(i) = \frac{\sum_{j\in L} K_\lambda(x_i,x_j)y_j}{\sum_{j\in L} K_\lambda(x_i,x_j)} \approx E[Y_i \mid X_i = x_i]$$

and is often called a Nadaraya-Watson kernel regression estimator (Hastie et al., 2001, Chapter 6). When applied to $L \cup U$, this estimator is

$$\widetilde{f} = \begin{pmatrix} \widetilde{f}_L \\ \widetilde{f}_U \end{pmatrix} = \begin{pmatrix} \widetilde{\boldsymbol{S}}_{LL} \\ \widetilde{\boldsymbol{S}}_{UL} \end{pmatrix} Y_L = \begin{pmatrix} \widetilde{\boldsymbol{D}}_{LL}^{-1}\boldsymbol{D}_{LL}\boldsymbol{S}_{LL} \\ \widetilde{\boldsymbol{D}}_{UL}^{-1}\boldsymbol{D}_{UU}\boldsymbol{S}_{UL} \end{pmatrix} Y_L, \tag{6}$$

where $\widetilde{\boldsymbol{S}}_{LL} = \widetilde{\boldsymbol{D}}_{LL}^{-1}\boldsymbol{W}_{LL}$ and $\widetilde{\boldsymbol{S}}_{UL} = \widetilde{\boldsymbol{D}}_{UL}^{-1}\boldsymbol{W}_{UL}$.

The supervised boundary estimator $\widetilde{f}_L = \widetilde{\boldsymbol{S}}_{LL}Y_L$ in Display (6) is based on the *supervised graph* $(L, \boldsymbol{W}_{LL})$. The supervised graph is the subgraph of $\boldsymbol{W}$ on L and has

$$\begin{array}{rcll} \widetilde{\boldsymbol{\Delta}}_{LL} & = & \widetilde{\boldsymbol{D}}_{LL} - \boldsymbol{W}_{LL} = \boldsymbol{\Delta}_{LL} - \boldsymbol{D}_{LL} & \text{(Supervised Laplacian)}, \\ \widetilde{\boldsymbol{S}}_{LL} & = & \widetilde{\boldsymbol{D}}_{LL}^{-1}\boldsymbol{W}_{LL} & \text{(Supervised Stochastic Smoother)}. \end{array}$$

The supervised smoothed value $\widetilde{f}_i$ for $i \in L$ is the probability weighted average of Y_L with weights from the ith row of $\widetilde{\boldsymbol{S}}_{LL}$, so $\widetilde{f}_i$ is based on relative strength of adjacencies within L, which might be depicted by $L \to L$. The supervised graph incorporates neither non-adjacent vertices nor U. Estimator $\widetilde{f}_L$ is also the solution to

$$\min_{f_L}(Y_L - f_L)^T\boldsymbol{W}_{LL}(Y_L - f_L) + f_L^T\widetilde{\boldsymbol{\Delta}}_{LL}f_L.$$

Supervised predictions of the interior from Display (6) are $\widetilde{f}_U = \widetilde{\boldsymbol{S}}_{UL}Y_L$. If $\widetilde{\boldsymbol{D}}_{UL_{ii}} = 0$ for some $i \in U$ then this supervised estimator is not defined for interior observation i, so this estimator exists for all $i \in U$ if and only if $\widetilde{\boldsymbol{D}}_{UL} \succ 0$, that is,

$$\nu^T\widetilde{\boldsymbol{D}}_{UL}\nu > 0 \text{ for any non-zero } \nu \in \mathbb{R}^{|U|}. \tag{7}$$

Condition (7) holds if and only if each unlabeled observation is adjacent to a labeled observation. This adjacency condition is a stringent requirement, especially when the proportion of labeled observations $|L|/n$ is small, and one might correctly guess that such a rigid requirement is not necessary if a semi-supervised harmonic function approach from Section 4 is taken.

3.2 The Offset Case

In this section, three $|L| \times |L|$ matrices $\boldsymbol{W}_{LUL}$, $\boldsymbol{\Delta}_{LUL}$, and $\boldsymbol{S}_{LUL}$ are defined, and it is shown that they correspond to the adjacency, Laplacian, and stochastic smoother matrices of a weighted graph on vertex set L, which we call the *offset graph*. These matrices are

$$\begin{array}{rcll} \boldsymbol{W}_{LUL} & = & \boldsymbol{\Delta}_{LU}\boldsymbol{\Delta}_{UU}^{-1}\boldsymbol{\Delta}_{UL} & \text{(Offset Graph with Vertex Set } L\text{)}, \\ \boldsymbol{\Delta}_{LUL} & = & \widetilde{\boldsymbol{D}}_{LU} - \boldsymbol{\Delta}_{LU}\boldsymbol{\Delta}_{UU}^{-1}\boldsymbol{\Delta}_{UL} & \text{(Offset Laplacian)}, \\ \boldsymbol{S}_{LUL} & = & \widetilde{\boldsymbol{D}}_{LU}^{-1}\boldsymbol{\Delta}_{LU}\boldsymbol{\Delta}_{UU}^{-1}\boldsymbol{\Delta}_{UL} & \text{(Offset Stochastic Smoother)}. \end{array}$$

Recall the necessary and sufficient adjacency condition in Display (7) for the uniqueness of the supervised estimator for all n observations. An intuitive condition for the uniqueness of a semi-supervised estimator for all n observations is that each connected component of $\boldsymbol{W}$ includes an observation from L, that is,

$$\nu^T \widetilde{\boldsymbol{D}}_{UL}\nu > 0 \text{ for any non-zero } \nu \in \mathcal{N}\left(\widetilde{\boldsymbol{D}}_{UU} - \boldsymbol{W}_{UU}\right). \tag{8}$$

Apply Remark 2 to subgraph $(U, \boldsymbol{W}_{UU})$ to justify this practical interpretation of Condition (8). The connectedness to L condition in Display (8) is less restrictive than the adjacency to L condition in Display (7), and Condition (8) implies that $\boldsymbol{W}$ has at most $|L|$ connected components. Proposition 10 establishes that Condition (8) is equivalent to the existence of $\boldsymbol{\Delta}_{UU}^{-1}$, a matrix involved in the definition of the offset graph.

Proposition 10 *If $\boldsymbol{W} \succeq 0$ then the following conditions are equivalent.*

(a) $\boldsymbol{\Delta}_{UU} \succ 0$.

(b) $\rho(\boldsymbol{S}_{UU}) < 1$.

(c) $\nu^T \widetilde{\boldsymbol{D}}_{UL}\nu > 0$ *for any non-zero* $\nu \in \mathcal{N}(\widetilde{\boldsymbol{D}}_{UU} - \boldsymbol{W}_{UU})$.

Condition (b) from Proposition 10 guarantees the convergence of the geometric matrix series with terms $\boldsymbol{S}_{UU}^{\ell} = \boldsymbol{\mathcal{O}}\boldsymbol{\mathcal{D}}^{\ell}\boldsymbol{\mathcal{O}}^{-1}$, where $\boldsymbol{\mathcal{O}}\boldsymbol{\mathcal{D}}\boldsymbol{\mathcal{O}}^{-1}$ is the eigendecomposition of $\boldsymbol{S}_{UU}$, so

$$\boldsymbol{D}_{LL}^{-1}\boldsymbol{W}_{LUL} = \boldsymbol{D}_{LL}^{-1}\boldsymbol{\Delta}_{LU}\boldsymbol{\Delta}_{UU}^{-1}\boldsymbol{\Delta}_{UL} = \boldsymbol{S}_{LU}\left(\boldsymbol{I} - \boldsymbol{S}_{UU}\right)^{-1}\boldsymbol{S}_{UL} = \sum_{\ell=0}^{\infty} \boldsymbol{S}_{LU}\boldsymbol{S}_{UU}^{\ell}\boldsymbol{S}_{UL} \geq 0, \tag{9}$$

where the inequality holds because $\boldsymbol{S}_{LU}\boldsymbol{S}_{UU}^{\ell}\boldsymbol{S}_{UL} \geq 0$ for each $\ell = 0, 1, \ldots$. Thus, $\boldsymbol{W}_{LUL} \geq 0$ is a valid weighted graph on L, since it's symmetric by definition. By the Laplacian property $\boldsymbol{\Delta}\vec{1} = \vec{0}$ and Partitioning (5), $\boldsymbol{\Delta}_{UL}\vec{1} = -\boldsymbol{\Delta}_{UU}\vec{1}$ and $\boldsymbol{\Delta}_{LU}\vec{1} = -\widetilde{\boldsymbol{D}}_{LU}\vec{1}$, so the degree matrix of $\boldsymbol{W}_{LUL}$ is $\text{diag}(\boldsymbol{W}_{LUL}\vec{1}) = \widetilde{\boldsymbol{D}}_{LU}$. Thus, the Laplacian and stochastic smoother of offset graph $\boldsymbol{W}_{LUL}$ are also established as matrices $\boldsymbol{\Delta}_{LUL}$ and $\boldsymbol{S}_{LUL}$ defined earlier.

The geometric matrix series in Display (9) provides a clear interpretation of each adjacency in offset graph $\boldsymbol{W}_{LUL}$. A pair of labeled observations is adjacent in $\boldsymbol{W}_{LUL}$ if and only if they are connected in $\boldsymbol{W}$ through a sequence of unlabeled observations; this type of connectedness might be depicted by $L \to U \leftrightarrow U \to L$. The offset boundary estimator is $(\boldsymbol{S}_{LUL}Y_L)_i$ for $i \in L$, that is, the probability weighted average of Y_L with weights from the ith row of $\boldsymbol{S}_{LUL}$. The probability weight on Y_{L_j} for $j \in L$ is $\boldsymbol{S}_{LUL_{ij}}$, and this weight will be relatively large if i has "strong" adjacencies to vertices in a "strongly adjacent" U network that is "strongly adjacent" to j. These are the only types of connectivity that matter in the offset case. For example, the adjacency between i and j simply does not factor into the offset based estimator.

3.3 The Semi-Supervised Case

The semi-supervised adjacency matrix is simply the sum of those from the supervised and offset cases, that is,

$$\boldsymbol{W}_{LL}+\boldsymbol{W}_{LUL} \quad \text{(Semi-Supervised Graph with Vertex Set } L\text{)}.$$

The semi-supervised Laplacian is thus the sum of positive semi-definite Laplacians

$$\begin{aligned}\boldsymbol{\Delta}^{\star}_{LL} &= \overbrace{\widetilde{\boldsymbol{D}}_{LL}-\boldsymbol{W}_{LL}}^{\substack{\text{Supervised}\\ \text{Laplacian}}}+\overbrace{\widetilde{\boldsymbol{D}}_{LU}-\boldsymbol{\Delta}_{LU}\boldsymbol{\Delta}_{UU}^{-1}\boldsymbol{\Delta}_{UL}}^{\text{Offset Laplacian}} && \text{(Semi-Supervised Laplacian)} \\ &= \boldsymbol{\Delta}_{LL}-\boldsymbol{\Delta}_{LU}\boldsymbol{\Delta}_{UU}^{-1}\boldsymbol{\Delta}_{UL} && (\boldsymbol{\Delta}_{UU} \text{ Block Schur Complement of } \boldsymbol{\Delta}).\end{aligned} \tag{10}$$

Refer to Section 2 and Display (2) for Schur complements.

The semi-supervised stochastic smoother is $\boldsymbol{M}_{LL}=\boldsymbol{D}_{LL}^{-1}(\boldsymbol{W}_{LL}+\boldsymbol{W}_{LUL})$. For more insight, first define the diagonal matrix $\boldsymbol{Q}_L=\boldsymbol{D}_{LL}^{-1}\widetilde{\boldsymbol{D}}_{LL}\succ 0$, which stores the proportion of each case's total similarities over all cases $L\cup U$ that is within L, that is,

$$\boldsymbol{Q}_{L_{ii}}=\frac{\sum_{j\in L}\boldsymbol{W}_{ij}}{\sum_{j\in L\cup U}\boldsymbol{W}_{ij}}.$$

Matrix $\boldsymbol{Q}_L$ provides the case-by-case probability weighted average compromise between the supervised and offset stochastic smoothers that is the semi-supervised stochastic smoother

$$\boldsymbol{M}_{LL}=\boldsymbol{Q}_L\widetilde{\boldsymbol{S}}_{LL}+(\boldsymbol{I}-\boldsymbol{Q}_L)\boldsymbol{S}_{LUL} \quad \text{(Semi-Supervised Stochastic Smoother)}.$$

More factorization produces yet another equivalent form

$$\boldsymbol{M}_{LL}=\boldsymbol{S}_{LL}+\boldsymbol{S}_{LU}(\boldsymbol{I}-\boldsymbol{S}_{UU})^{-1}\boldsymbol{S}_{UL} \quad (\boldsymbol{S}_{UU} \text{ Stochastic Complement of } \boldsymbol{S}). \tag{11}$$

Adjacencies accumulate in semi-supervised graph $\boldsymbol{W}_{LL}+\boldsymbol{\Delta}_{LU}\boldsymbol{\Delta}_{UU}^{-1}\boldsymbol{\Delta}_{UL}$ due to exactly two types of connectedness among the labeled observations in graph $\boldsymbol{W}$: (i) supervised $L\rightarrow L$ and (ii) offset $L\rightarrow U\leftrightarrow U\rightarrow L$. The prediction for a case $i\in L$ puts more weight on the supervised prediction for large $\boldsymbol{Q}_{L_{ii}}$ and on the offset prediction for large $1-\boldsymbol{Q}_{L_{ii}}$, so $\boldsymbol{M}_{LL}$ is always a practical probability weighted average of the estimators based on graphs $\boldsymbol{W}_{LL}$ and $\boldsymbol{W}_{LUL}$. The connectedness of labeled vertices in the semi-supervised graph is the same as that in the full graph $\boldsymbol{W}$, but types of connectedness outside (i) and (ii) don't get incorporated into semi-supervised predictions (see Remark 3).

The decomposition of the semi-supervised graph into supervised and offset graphs is displayed concisely in Figure 1. While it is not too hard to compute the Laplacian or stochastic smoother from the weighted graph, no other offset or semi-supervised representation can be fully recovered from just the Laplacian or just the smoother. However, it is possible to recover $\boldsymbol{W}$ from $\boldsymbol{S}$ because $\boldsymbol{W}_{ii}=1$ is known.

Additional insight into the inter-workings of the semi-supervised smoother is gleaned through analytical eigenvalue results. First, $\boldsymbol{\Delta}^{\star}_{LL}=\boldsymbol{D}_{LL}(\boldsymbol{I}-\boldsymbol{M}_{LL})$, so

$$\boldsymbol{\Delta}^{\star}_{LL}\nu=\vec{0} \iff \boldsymbol{M}_{LL}\nu=\nu$$

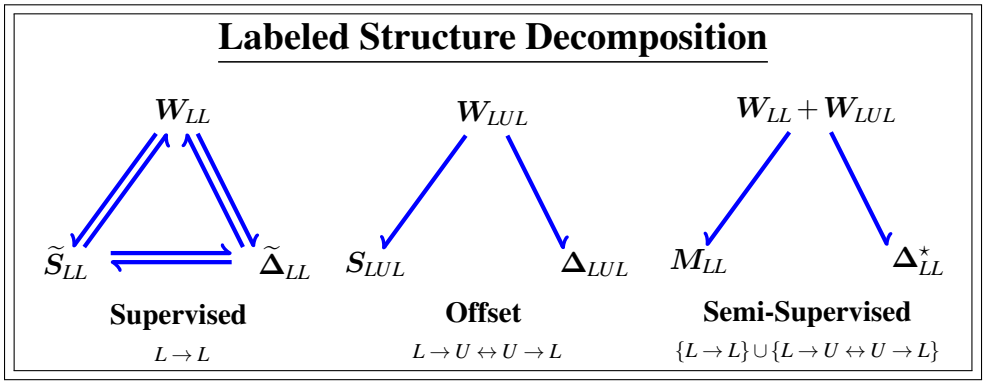

Figure 1: Matrix representations of weighted graphs each with vertex set L: adjacency (top), stochastic smoother (bottom left), and Laplacian (bottom right). Each semi-supervised labeled representation is a linear combination of the corresponding supervised and offset representations. Harpoons indicate that the representation after the barb can be computed from that on the other end.

provides a second example of the general relationship between a smoother and its Laplacian (reference Display (4) for that between $\boldsymbol{\Delta}$ and $\boldsymbol{S}$). To analytically break down $\mathcal{N}(\boldsymbol{\Delta}^{\star}_{LL})$ (and hence the eigenspace of $\boldsymbol{M}_{LL}$ corresponding to eigenvalue 1), first recall the decomposition of Laplacian $\boldsymbol{\Delta}^{\star}_{LL}$ in Equation (10) as the sum of positive semi-definite Laplacians. Thus,

$$\mathcal{N}(\boldsymbol{\Delta}^{\star}_{LL}) = \mathcal{N}\left(\widetilde{\boldsymbol{\Delta}}_{LL}\right) \cap \mathcal{N}(\boldsymbol{\Delta}_{LUL}) \subseteq I\!R^{|L|}.$$

Certainly $\vec{1} \in \mathcal{N}(\boldsymbol{\Delta}^{\star}_{LL})$, and a particular orthogonal basis of binary vectors for $\mathcal{N}(\boldsymbol{\Delta}^{\star}_{LL})$ is given by Remark 2. Each basis vector indicates vertices in a connected component of the semi-supervised graph, and so they partition L and sum to $\vec{1}$. Similarly, partitions of L corresponding to the connected components of the supervised and offset graphs correspond to orthogonal bases of binary vectors for $\mathcal{N}(\widetilde{\boldsymbol{\Delta}}_{LL})$ and $\mathcal{N}(\boldsymbol{\Delta}_{LUL})$. The operation of intersecting $\mathcal{N}(\widetilde{\boldsymbol{\Delta}}_{LL})$ and $\mathcal{N}(\boldsymbol{\Delta}_{LUL})$ can never increase the dimension of the resulting $\mathcal{N}(\boldsymbol{\Delta}^{\star}_{LL})$ and is equivalent to increasing connectivity by producing the coarsest possible partition of L that can be made by both partitions (of L corresponding to $\mathcal{N}(\widetilde{\boldsymbol{\Delta}}_{LL})$ and $\mathcal{N}(\boldsymbol{\Delta}_{LUL})$) via unions of their respective subsets.

Supervised graph $\boldsymbol{W}_{LL}$ is a subgraph of $\boldsymbol{W}$. They have the same adjacencies in L, but $\boldsymbol{W}_{LL}$ can only reduce connectivity in L relative to that in $\boldsymbol{W}$. The addition of the offset $\boldsymbol{W}_{LUL}$ to $\boldsymbol{W}_{LL}$ achieves the same level of connectedness in L as $\boldsymbol{W}$, but more importantly introduces offset adjacencies in the semi-supervised graph not found in the supervised graph. It is the adjacencies in the semi-supervised graph that determine non-zero smoother weights (see Remark 3). In spite of this, the connectedness structure of the semi-supervised graph is still important so that one understands the smoother properties via its eigenvalue decomposition. If a condition from Proposition 10 holds, then each connected component of $\boldsymbol{W}$ includes a vertex from L. In this case, the dimension of $\mathcal{N}(\boldsymbol{\Delta}^{\star}_{LL}) \subseteq I\!R^{|L|}$ equals the dimension of $\mathcal{N}(\boldsymbol{\Delta}) \subseteq I\!R^{|L \cup U|}$. Intuitively, we view $\boldsymbol{M}_{LL}$ as a labeled stochastic smoother with respect to the observed response Y_L, while $\boldsymbol{S}$ is a stochastic smoother with respect to the training response $Y(Y_U)$.

Remark 3 *Semi-supervised graph $\boldsymbol{W}_{LL}+\boldsymbol{W}_{LUL}$ on L keeps the meaningful connectedness structure of the full graph $\boldsymbol{W}$ on $L\cup U$. A pair of labeled observations are in the same connected component of one of these graphs if and only if the same is true in the other graph. This follows because adjacent boundary vertices in $\boldsymbol{W}_{LL}+\boldsymbol{W}_{LUL}$ are connected in $\boldsymbol{W}$ via either a sequence of labeled vertices (supervised) or a sequence of unlabeled vertices (offset), and sequences of these two types of connectivities in $\boldsymbol{W}$ can build any type connectivity that exists in $\boldsymbol{W}$ from an $i\in L$ to $j\in L$. It follows that*

$$\nu=\begin{pmatrix}\nu_L\\ \nu_U\end{pmatrix}\in\mathcal{N}(\boldsymbol{\Delta})\subseteq\mathbb{R}^{|L\cup U|}\implies\nu_L\in\mathcal{N}(\boldsymbol{\Delta}^{\star}_{LL})\subseteq\mathbb{R}^{|L|} \tag{12}$$

(refer to Remark 2).

Let $i\in L$ and $j\in L$. Probability weight $\boldsymbol{M}_{LL_{ij}}$ is that for Y_{L_j} in the semi-supervised smoothed value for Y_{L_i}. It should come as no surprise that a sufficient condition for $\boldsymbol{M}_{LL_{ij}}=0$ is that boundary vertices i and j are not in the same connect component of $\boldsymbol{W}$, but this condition is not necessary. The necessary and sufficient condition for $\boldsymbol{M}_{LL_{ij}}>0$ is that i and j are adjacent in at least one graph $\boldsymbol{W}_{LL}$ or $\boldsymbol{W}_{LUL}$. The hypothetical situation where i and j are in the same connect component of $\boldsymbol{W}$ and $\boldsymbol{M}_{LL_{ij}}=0$ is possible if boundary vertices i and j are connected in the full graph $\boldsymbol{W}$ but not through a pure sequence of all boundary (or of all interior) vertices.

4. Harmonic Functions in Semi-Supervised Learning

Harmonic functions form the basis for the connection between electrical networks and random walks (Doyle and Snell, 1984). The use of harmonic estimation in semi-supervised learning is discussed extensively in its relation to random walks, electrical networks, and energy optimization (Zhu et al., 2003).

A function $h:\mathcal{V}\to\mathbb{R}$ is *harmonic* with respect to a stochastic matrix $\boldsymbol{S}$ if

$$f_i=\sum_{\ell\in L\cup U}\boldsymbol{S}_{i\ell}f_\ell\quad\text{for each } i\in U, \tag{13}$$

where $f_i=h(i)$ (Zhu et al., 2003; Abney, 2008). In matrix form, the implication of Equation (13) on a resulting *harmonic estimator* $f\in\mathbb{R}^n$ is

$$\boldsymbol{S}f=\begin{pmatrix}\boldsymbol{S}_{LL}f_L+\boldsymbol{S}_{LU}f_U\\ \boldsymbol{S}_{UL}f_L+\boldsymbol{S}_{UU}f_U\end{pmatrix}=\begin{pmatrix}(\boldsymbol{S}f)_L\\ f_U\end{pmatrix}. \tag{14}$$

In the case of a harmonic estimator in Display (14), it follows by Display (12) that $(\boldsymbol{S}f)_L=f_L$ if and only if $f_L\in\mathcal{N}(\boldsymbol{\Delta}^{\star}_{LL})$. In other words, $\boldsymbol{S}f=f$ holds for a harmonic estimator f if and only if f_L is constant within the connected components of $\boldsymbol{W}$. This precise concept of when $\boldsymbol{S}f=f$ is in tandem with the practical application of a judiciously chosen harmonic estimator under the cluster assumption studied further in Section 4.1.

A question not addressed in the above discussion is the existence and uniqueness of a harmonic estimator f. This mathematical matter is solved in two cases $\rho(\boldsymbol{S}_{UU})<1$ and $\rho(\boldsymbol{S}_{UU})=1$, which are collectively exhaustive by Lemma 9 in Appendix A. First, consider the case of $\rho(\boldsymbol{S}_{UU})<1$ (or any other equivalent condition from Proposition 10), so that $(\boldsymbol{I}-\boldsymbol{S}_{UU})^{-1}$ exists. In this case, the unique estimator for the interior $f_U=(\boldsymbol{I}-\boldsymbol{S}_{UU})^{-1}\boldsymbol{S}_{UL}f_L$ is a linear transformation of the boundary

estimate. If one uses this unique solution for the interior as well as the stochastic complement representation of $\boldsymbol{M}_{LL}$ from Equation (11), then Equation (14) simplifies to

$$\boldsymbol{S}f = \begin{pmatrix} (\boldsymbol{S}f)_L \\ f_U \end{pmatrix} = \begin{pmatrix} \boldsymbol{M}_{LL} \\ (\boldsymbol{I} - \boldsymbol{S}_{UU})^{-1} \boldsymbol{S}_{UL} \end{pmatrix} f_L. \tag{15}$$

The left of Equation (15) is an $n \times n$ times $n \times 1$ matrix multiplication, whereas the right is an $n \times |L|$ times an $|L| \times 1$. Next, the case of $\rho(\boldsymbol{S}_{UU}) = 1$ implies that at least one connected component in $\boldsymbol{W}$ contains all interior observations, that is, Condition (8) does not hold. So with given estimate f_L, a harmonic estimate f_U exists, but is not unique because there is an arbitrary choice for a constant labeling within each pure interior connected component. The assumption $\rho(\boldsymbol{S}_{UU}) < 1$ used throughout most of Sections 3 and 4 avoids this arbitrary nature of harmonic estimators when $\rho(\boldsymbol{S}_{UU}) = 1$. The subtlety in the case of $\rho(\boldsymbol{S}_{UU}) = 1$ is directly overcome by methods of regularization presented later in Section 5.

The maximum principle states that a harmonic solution is bounded above and below by the boundary estimate (Doyle and Snell, 1984). The uniqueness principle, which applies in the case of $\rho(\boldsymbol{S}_{UU}) < 1$, states that if two harmonic functions are applied with the same boundary estimate f_L then they must produce the same interior estimate f_U. One thing that is clear from each of these principles is that a harmonic estimate f_U of the interior is a function of the boundary estimate f_L. While the semi-supervised boundary estimator $f_L = \boldsymbol{M}_{LL} Y_L$ was thoroughly developed in Section 3, the plethora of competing boundary estimators is a focus of Section 4.1.

4.1 The Cluster Assumption and Boundary Estimation

The *cluster assumption* states that observations close in proximity should have similar labels. Our main objective is to understand how this concept relates to classifiers. Let ψ be an arbitrary classifier trained with weighted graph $\boldsymbol{W}$ and arbitrary response Y_L. We say that ψ is a *cluster assumption classifier* if ψ is guaranteed to satisfy

$$\psi \in \mathcal{N}(\boldsymbol{\Delta}) \text{ and } \psi_L = Y_L \iff Y_L \in \mathcal{N}(\boldsymbol{\Delta}^\star_{LL}). \tag{16}$$

Suppose the response is constant within the connected components of $\boldsymbol{W}$. Condition (16) guarantees that a cluster assumption classifier classifies each interior observation with the unique label observed within its connected component (refer to Remarks 2 and 3).

Let f be a harmonic function trained from the weighted graph $\boldsymbol{W}$ and response Y_L. In order for f to also be a cluster assumption classifier, the boundary must be estimated with $f_L = Y_L$ for any $Y_L \in \mathcal{N}(\boldsymbol{\Delta}^\star_{LL})$, that is, $Y_L \in \mathcal{N}(\boldsymbol{\Delta}^\star_{LL}) \Longrightarrow \boldsymbol{S}f = f$ and $f \in \mathcal{N}(\boldsymbol{\Delta})$. Harmonic functions that are cluster assumption classifiers are also useful in circumstances when $\boldsymbol{W}$ has only one connected component. Suppose there are weak adjacencies less than some small $\varepsilon/n > 0$ between clusters, and pairs within clusters are connected by an edge path with adjacencies exceeding ε. Then decomposition $\boldsymbol{W} = \boldsymbol{W}_{\text{weak}} + \boldsymbol{W}_{\text{strong}}$, where $\boldsymbol{W}_{\text{weak}_{ij}} = \min\{\varepsilon/n, \boldsymbol{W}_{ij}\}$, produces connected components in the strong graph that correspond to clusters. The cluster assumption holds on the strong graph. Now, for any $f \in \mathcal{N}(\boldsymbol{\Delta}_{\text{strong}})$, $\boldsymbol{S}f \approx \boldsymbol{S}_{\text{strong}} f \in \mathcal{N}(\boldsymbol{\Delta}_{\text{strong}})$ because the smoother $\boldsymbol{S}$ is a row wise probability weighted average of the strong and weak smoothers that puts a low weight on the weak smoother. If $f_L = Y_L \in \mathcal{N}\left(\boldsymbol{\Delta}^\star_{LL_{\text{strong}}}\right)$ such that $Y_{L_i} = 1$ on a connected component of $\boldsymbol{W}_{\text{strong}}$ and $Y_{L_i} = -1$ elsewhere, then $\text{sign}(\boldsymbol{S}f) \in \mathcal{N}(\boldsymbol{\Delta}_{\text{strong}})$, so the hard labels classify in accordance with the cluster

assumption, which is consistent with the empirical evidence in the literature (Chapelle et al., 2006b; Abney, 2008).

The simplest boundary estimate for a harmonic estimator is the *clamped harmonic estimator* $f_L = Y_L$ (Zhu et al., 2003; Abney, 2008). The clamped harmonic estimator can be motivated as solving

$$\min_{f_L} (Y_L - f_L)^T (Y_L - f_L)$$

to obtain the boundary estimator $f_L = Y_L$ and then enforcing Equation (15) to define a harmonic estimator by setting $f_U = (\boldsymbol{I} - \boldsymbol{S}_{UU})^{-1} \boldsymbol{S}_{UL} f_L$.

This is not the only possible harmonic estimator because one can use any boundary estimator to develop a harmonic estimator. For example, consider

$$\min_{f} (Y_L - f_L)^T (\boldsymbol{W}_{LL} + \boldsymbol{W}_{LUL})(Y_L - f_L) + f^T \boldsymbol{\Delta} f, \tag{17}$$

where the loss function is based off of the semi-supervised graph developed in Section 3. The solution to Optimization (17) is a harmonic function with the boundary estimate $f_L = \boldsymbol{M}_{LL} Y_L$ from Section 3.3. The reason why Optimization (17) produces a harmonic function can be seen by studying the optimization of a generalized labeled loss function with penalty

$$\min_{f} L(Y_L, f_L) + \eta f^T \boldsymbol{\Delta} f, \tag{18}$$

where $L(Y_L, Y_L) \leq L(Y_L, f_L)$ for any f_L. Since this loss function is independent of f_U, the optimal estimate for the interior for any $\eta > 0$ is

$$\arg\min_{f_U} f^T \boldsymbol{\Delta} f = (\boldsymbol{I} - \boldsymbol{S}_{UU})^{-1} \boldsymbol{S}_{UL} f_L,$$

which is harmonic. For any harmonic function f,

$$\begin{aligned} f^T \boldsymbol{\Delta} f &= f_L^T \boldsymbol{\Delta}_{LL} f_L + 2 f_L^T \boldsymbol{\Delta}_{LU} f_U + f_U^T \boldsymbol{\Delta}_{UU} f_U \\ &= f_L^T \boldsymbol{\Delta}_{LL} f_L - 2 f_L^T \boldsymbol{\Delta}_{LU} \boldsymbol{\Delta}_{UU}^{-1} \boldsymbol{\Delta}_{UL} f_L + f_L^T \boldsymbol{\Delta}_{LU} \boldsymbol{\Delta}_{UU}^{-1} \boldsymbol{\Delta}_{UL} f_L \\ &= f_L^T \boldsymbol{\Delta}_{LL}^{\star} f_L, \end{aligned}$$

so Optimization (18) produces a harmonic function with boundary solving

$$\min_{f_L} L(Y_L, f_L) + \eta f_L^T \boldsymbol{\Delta}_{LL}^{\star} f_L, \tag{19}$$

or equivalently

$$\min_{f_L} \underbrace{L(Y_L, f_L) + \eta f_L^T \widetilde{\boldsymbol{\Delta}}_{LL} f_L}_{\text{Supervised Objective}} + \underbrace{\eta f_L^T \boldsymbol{\Delta}_{LUL} f_L}_{\text{Offset}}.$$

Furthermore, under Optimization (19) with a finite loss $L(\cdot,\cdot)$, the clamped estimate of $f_L = Y_L$ is optimal for all $\eta > 0$ if and only if $Y_L \in \mathcal{N}(\boldsymbol{\Delta}_{LL}^{\star})$. In general, the clamped harmonic estimator is not necessarily optimal among harmonic estimators.

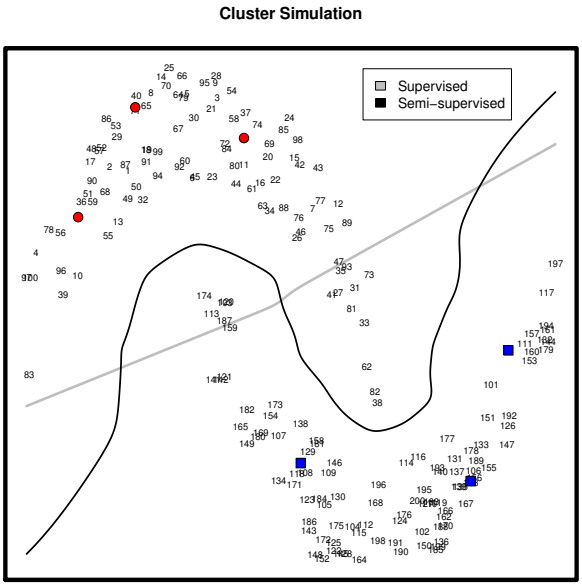

Figure 2: A "two moons" data set with $|L| = 6$ and $|U| = 200$. Label ● $= -1$ and ■ $= 1$.

4.2 Impact of Supervised Kernel Smoothing on Harmonic Estimators

Further examination of the cluster assumption is had by comparing the supervised kernel smoother (Section 3.1) to the semi-supervised harmonic estimator (Section 3.3). A goal is to understand why an observation $i \in U$ would sacrifice its own supervised estimate in favor of the cluster. Take the "two moons" example in Figure 2 that includes supervised and semi-supervised boundaries (see Remark 5). Focus on observation 38 in the downward pointing horn on right. According to the supervised rule this observation is ■ with probability 1. The semi-supervised prediction $f_{U_{38}} = -0.42$ is ● with probability 0.7, so the supervised estimate is overturned in favor of the cluster.

Any harmonic estimator with boundary f_L has the form $f_U = (\boldsymbol{I} - \boldsymbol{S}_{UU})^{-1} \boldsymbol{S}_{UL} f_L$. Assume $\widetilde{\boldsymbol{D}}_{UL} \succ 0$, so the supervised estimator exists (see Remark 4). Also, generalize the supervised predictions to $\widetilde{f}_U = \widetilde{\boldsymbol{S}}_{UL} f_L$, which we refer to as *soft supervised estimates*. Matrix $(\boldsymbol{I} - \boldsymbol{S}_{UU})^{-1} \boldsymbol{S}_{UL}$ is the product of the $|U| \times |U|$ stochastic matrix $(\boldsymbol{I} - \boldsymbol{S}_{UU})^{-1} \boldsymbol{D}_{UU}^{-1} \widetilde{\boldsymbol{D}}_{UL}$ and the $|U| \times |L|$ supervised prediction matrix $\widetilde{\boldsymbol{S}}_{UL} = \widetilde{\boldsymbol{D}}_{UL}^{-1} \boldsymbol{W}_{UL}$, that is,

$$\begin{aligned} f_U &= (\boldsymbol{I} - \boldsymbol{S}_{UU})^{-1} \boldsymbol{S}_{UL} f_L && (20) \\ &= (\boldsymbol{I} - \boldsymbol{S}_{UU})^{-1} \boldsymbol{D}_{UU}^{-1} \widetilde{\boldsymbol{D}}_{UL} \widetilde{\boldsymbol{S}}_{UL} f_L \\ &= (\boldsymbol{I} - \boldsymbol{S}_{UU})^{-1} \boldsymbol{D}_{UU}^{-1} \widetilde{\boldsymbol{D}}_{UL} \widetilde{f}_U. && (21) \end{aligned}$$

Equation (21) shows that any semi-supervised harmonic function is a probability weighted average of the soft supervised estimators of U, that is,

$$f_{U_i} = \sum_{j \in U} \boldsymbol{P}_{ij} \widetilde{f}_j,$$

where the weights come from the stochastic matrix

$$\boldsymbol{P} = (\boldsymbol{I} - \boldsymbol{S}_{UU})^{-1} \boldsymbol{D}_{UU}^{-1} \widetilde{\boldsymbol{D}}_{UL}. \qquad (22)$$

Determining which soft supervised predictions $\widetilde{f}_U$ get the larger probability weights in the semi-supervised predictions f_U makes practical sense. Such a determination is possible if one relates the stochastic matrix $\boldsymbol{P}$ in Equation (22) to an absorbing Markov chain probability model (Doyle and Snell, 1984).

Consider the $|U|+1$ state Markov chain with transition matrix

$$\begin{pmatrix} \boldsymbol{S}_{UU} & (\boldsymbol{I}-\boldsymbol{S}_{UU})\,\vec{1} \\ \vec{0}^T & \vec{1} \end{pmatrix}.$$

Boundary L is treated as an *absorbing state*, and the harmonic estimator of the interior is

$$f_{U_i} = e_i^T f_U = \left(\sum_{k=0}^{\infty} e_i^T \boldsymbol{S}_{UU}^k \boldsymbol{D}_{UU}^{-1} \widetilde{\boldsymbol{D}}_{UL} \right) \widetilde{f}_U \text{ with elementary vector } e_i. \tag{23}$$

Each term in geometric series from Display (23) is the probability of a particular sequence of transitions with a given starting point in the absorbing Markov Chain probability model.

1: The first transition absorption to L starting from $i \in U$ is the $1 \times |U|$ row vector $e_i^T \boldsymbol{D}_{UU}^{-1} \widetilde{\boldsymbol{D}}_{UL}$, which has one non-zero entry. This non-zero, column i entry is the probability that a chain starting at unlabeled state i is absorbed into L at the first transition. This probability is large if unlabeled case $i \in U$ has more total similarity with cases in L than that with cases in U.

2: The second transition absorption to L from $j \in U$ starting from $i \in U$ is the row vector $e_i^T \boldsymbol{S}_{UU} \boldsymbol{D}_{UU}^{-1} \widetilde{\boldsymbol{D}}_{UL}$. Its jth column entry is the probability that a chain starting at unlabeled state i goes to unlabeled state j at first transition and is absorbed into L at the second transition.

...

k: The k'th transition absorption to L from $j \in U$ starting from $i \in U$ is the jth column entry of row vector $e_i^T \boldsymbol{S}_{UU}^{k-1} \boldsymbol{D}_{UU}^{-1} \widetilde{\boldsymbol{D}}_{UL}$. It is the probability that a chain starting at $i \in U$ goes $k-1$ transitions in U ending at some state $j \in U$ before being absorbed into L at the kth transition.

By Equation (23), the probability weight on soft supervised prediction $j \in U$ in semi-supervised prediction $i \in U$ is just the probability that a chain starting at $i \in U$ is absorbed from $j \in U$. Therefore, the soft supervised predictions for $j \in U$ that are "strongly adjacent" to observations in L carry the majority of the weight.

Back to Figure 2 for case 38. The top ten cases, that is, 72, 129, 84, 69, 74, 71, 36, 108, 20, and 59, carry 68% of the weight in the semi-supervised prediction of case 38, and each is close to a labeled observation. This "top ten" provides the approximation $\sum_{i=1}^{200} \boldsymbol{P}_{38,j} \widetilde{f}_{U_j} \approx \sum_{i=1}^{10} \boldsymbol{P}_{38(j)} \widetilde{f}_{U_{(j)}} = -0.38$ of the semi-supervised estimate, where (j) is the column of $\boldsymbol{P}$ containing its jth largest value in the 38th row. Hence the label prediction for observation 38 is already determined as -1 from this "top 10" because the combined weight of the other 190 cases at 32% is not enough to reverse the sign -0.38 given a ± 1 labeling. Furthermore, the supervised estimate for observation 38 is 68'th in the order with weight of only 0.002 or 0.2% in its very own semi-supervised prediction.

Remark 4 *"Assumption"* $\widetilde{\boldsymbol{D}}_{UL} \succ 0$ *is not necessary. If* $\widetilde{\boldsymbol{D}}_{UL} \not\succ 0$*, there exists* $i \in U$ *such that* $\widetilde{\boldsymbol{D}}_{UL_{ii}} = 0$*, and the supervised estimate does not exist for such i. This does not affect Equation*

(20), but is required in Factorization (21). Let $\widetilde{\boldsymbol{D}}_{UL}^{+}$ *be the diagonal generalized inverse of* $\widetilde{\boldsymbol{D}}_{UL}$ *with the same number of zero entries. If* $\widetilde{\boldsymbol{D}}_{UL}^{+}$ *is substituted in place of the nonexistent* $\widetilde{\boldsymbol{D}}_{UL}^{-1}$ *so that nonexistent soft supervised estimators are set to* $\widetilde{f}_{U_i} = \left(\widetilde{\boldsymbol{D}}_{UL}^{+}\boldsymbol{W}_{UL}f_L\right)_i = 0$*, Factorization (21) and its ensuing interpretation hold.*

5. Regularized Joint Harmonic Functions

Briefly consider the case when the response y is observed for all n observations. The Nadaraya-Watson kernel estimator $f = \boldsymbol{S}y$ results if functional $(y-f)^T\boldsymbol{W}(y-f)+f^T\boldsymbol{\Delta}f$ is minimized. In the semi-supervised setting when Y_U is missing, we replace y with the training response $Y(Y_U)$ from Display (3) and jointly optimize for both f and Y_U. In particular, the *regularized joint harmonic estimator* is the solution to

$$\min_{Y_U,f}\ (Y(Y_U)-f)^T\boldsymbol{W}(Y(Y_U)-f)+f^T\boldsymbol{\Delta}f+\gamma Y_U^TY_U \quad \text{(Joint Optimization Problem).} \tag{24}$$

The regularized joint harmonic estimator, given in Proposition 12, includes an estimator for both Y_U and f. The form of the f portion of this estimator is established as harmonic when $\gamma=0$ in Section 5.1. Discussion of the stabilizing effect due to the additional term $\gamma Y_U^TY_U$ in the context of the Joint Optimization Problem (24) when $\gamma>0$ is deferred until Section 5.2.

Proposition 12 *Let* $\boldsymbol{W}\succeq 0$*. Assume* $(\boldsymbol{\Delta S})_{UU}\succ 0$ *when one selects* $\gamma=0$*; this additional assumption is not required when one selects some* $\gamma>0$*. The unique solution to the Joint Harmonic Optimization Problem (24) is* $(Y_U,f)=\left(\hat{Y}_{U_\gamma},\boldsymbol{S}Y\left(\hat{Y}_{U_\gamma}\right)\right)$*, where*

$$\hat{Y}_{U_\gamma}=-\left((\boldsymbol{\Delta S})_{UU}+\gamma\boldsymbol{I}\right)^{-1}(\boldsymbol{\Delta S})_{UL}Y_L.$$

Matrix $\boldsymbol{\Delta S}$ has many of the properties of $\boldsymbol{\Delta}$ from Section 3, for example, $\boldsymbol{\Delta}\vec{1}=\vec{0}$ and $\boldsymbol{\Delta S}\vec{1}=\vec{0}$. Moreover, it is easy to verify that $\mathcal{N}(\boldsymbol{\Delta S})=\mathcal{N}(\boldsymbol{\Delta})$. Proposition 11 establishes a result for the positive semi-definiteness of $\boldsymbol{\Delta S}$, which is analogous to $\boldsymbol{\Delta}$ and Proposition 7.[1]

Proposition 11 *If* $\boldsymbol{W}\succeq 0$ *then* $\boldsymbol{\Delta S}\succeq 0$*.*

By Proposition 11, $\boldsymbol{W}\succeq 0$ is a sufficient condition for the uniqueness of the joint harmonic estimator when $\gamma>0$, but the added condition $(\boldsymbol{\Delta S})_{UU}\succ 0$ from Proposition 12 is needed if $\gamma=0$. Case $\gamma=0$ is discussed further in Section 5.1, and case $\gamma>0$ in Section 5.2.

Remark 5 *The prediction of a novel case given its nonnegative similarities* $(w_1,\cdots,w_n)$ *and response estimate* $\hat{Y}_{U_\gamma}$ *is computed from the Nadaraya-Watson kernel based function*

$$\breve{h}(w_1,\cdots,w_n)=\frac{\sum_{i\in L\cup U}w_iY_i\left(\hat{Y}_{U_\gamma}\right)}{\sum_{i\in L\cup U}w_i},$$

where $\breve{h}:\mathbb{R}^n\to\mathbb{R}$*. Finding the points in* $\mathbb{R}^n$ *that satisfy* $\breve{h}(w_1,\cdots,w_n)=0$ *is how one finds boundaries like those superimposed on Figure 2.*

1. Proposition 11 is used to prove Proposition 12 in Appendix A, but order was reversed here for presentation.

5.1 Joint Harmonic Estimator $\gamma = 0$

Here the joint harmonic function requires $(\boldsymbol{\Delta S})_{UU} \succ 0$ for its uniqueness (see Proposition 12). Results to come later in this section show that its boundary estimator is built on the unlabeled-unlabeled Schur complements of $\boldsymbol{W}$ and $\boldsymbol{\Delta}$ (refer to Section 2). First, Proposition 16 establishes an equivalence between these Schur complements and $(\boldsymbol{\Delta S})_{UU} \succ 0$.

Proposition 16 *If $\boldsymbol{W} \succeq 0$ then*

$$(\boldsymbol{\Delta S})_{UU} \succ 0 \iff \boldsymbol{W}_{UU} \succ 0,\ \boldsymbol{\Delta}_{UU} \succ 0, \text{ and } (\boldsymbol{W}^{\star}_{LL} + \boldsymbol{\Delta}^{\star}_{LL}) \succ 0.$$

Conditions from Proposition 16 are necessary and sufficient for the existence of the smoother

$$\boldsymbol{\Gamma}_{LL} = (\boldsymbol{W}^{\star}_{LL} + \boldsymbol{\Delta}^{\star}_{LL})^{-1} \boldsymbol{W}^{\star}_{LL} \quad \text{(Joint Harmonic Smoother)}, \tag{25}$$

and Theorem 18 states that smoother $\boldsymbol{\Gamma}_{LL}$ is that for the joint harmonic estimator.

Theorem 18 *Let $\boldsymbol{W} \succeq 0$, and assume that $\boldsymbol{\Gamma}_{LL}$ exists. The solution to the Joint Harmonic Optimization Problem (24) with $\gamma = 0$ has*

$$f = \begin{pmatrix} f_L \\ (\boldsymbol{I} - \boldsymbol{S}_{UU})^{-1} \boldsymbol{S}_{UL} f_L \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Gamma}_{LL} \\ (\boldsymbol{I} - \boldsymbol{S}_{UU})^{-1} \boldsymbol{S}_{UL} \boldsymbol{\Gamma}_{LL} \end{pmatrix} Y_L,$$

so f is in-fact harmonic.

The work connecting the interior of a harmonic estimator to supervised estimators in Section 4.2 now applies to the joint harmonic estimator, that is, in particular recall $f_U = \boldsymbol{P}\widetilde{\boldsymbol{S}}_{UL}\boldsymbol{\Gamma}_{LL}Y_L$ with $\boldsymbol{P}$ from Display (22). One can view $\boldsymbol{\Gamma}_{LL}$ as a filter between the response Y_L and the supervised prediction smoother $\widetilde{\boldsymbol{S}}_{UL}$, which provides additional robustness for misspecified responses over that of using Y_L directly to form supervised predictions.

The boundary estimator is equivalently expressed as the solution to

$$\min_{f_L} (Y_L - f_L)^T \boldsymbol{W}^{\star}_{LL} (Y_L - f_L) + f_L^T \boldsymbol{\Delta}^{\star}_{LL} f_L. \tag{26}$$

Optimization (26) provides an interesting example of the labeled loss optimization problem from Display (18), where $\boldsymbol{W}^{\star}_{LL}$ allows unlabeled data to influence the weighted squared error loss functional independent of f_U. Hence, the harmonic result for labeled loss is still preserved, but the loss function is not independent of the unlabeled data. This also shows how this estimator generalizes the supervised case by replacing $\boldsymbol{W}_{LL}$ with $\boldsymbol{W}^{\star}_{LL}$ and $\widetilde{\boldsymbol{\Delta}}_{LL}$ with $\boldsymbol{\Delta}^{\star}_{LL}$. Furthermore, since $\boldsymbol{\Gamma}_{LL} = \boldsymbol{I} - (\boldsymbol{W}^{\star}_{LL} + \boldsymbol{\Delta}^{\star}_{LL})^{-1} \boldsymbol{\Delta}^{\star}_{LL}$,

$$\boldsymbol{\Delta}^{\star}_{LL} \nu = \vec{0} \iff \boldsymbol{\Gamma}_{LL} \nu = \nu,$$

so the joint harmonic estimator is a *cluster assumption classifier* (refer to Section 4.1). Proposition 19 provides further insight on the smoothing properties of $\boldsymbol{\Gamma}_{LL}$.

Proposition 19 *If $\boldsymbol{W} \succeq 0$ and $\boldsymbol{\Gamma}_{LL}$ exists then each eigenvalue of $\boldsymbol{\Gamma}_{LL}$ is an element of $[0,1]$.*

The above results for smoother $\boldsymbol{\Gamma}_{LL}$ are weaker than those for the stochastic semi-supervised smoother $\boldsymbol{M}_{LL}$ from Figure 1. In general, $\boldsymbol{\Gamma}_{LL}$ is not stochastic, although it was stochastic in nearly every numerical example we considered. In cases when $\boldsymbol{\Gamma}_{LL}$ is stochastic, the stronger condition that $|e_i^T f_U| \leq |e_i^T Y_L|$ holds, by the maximum principle of harmonic functions (Doyle and Snell, 1984).

In applications such as those in Sections 6 and 7, assumptions for the uniqueness of the $\gamma = 0$ joint harmonic estimator are not likely to be satisfied. These assumptions are especially sensitive to circumstances where $\boldsymbol{W}$ is generated from $\boldsymbol{X}$ with a kernel function set to small λ. The breakdown tends to worsen when $|L|/n$ is small. On the other hand, the $\gamma > 0$ regularized joint harmonic estimators in Section 5.2 elegantly relax these assumptions by modifying the Schur complements on the right of Display (25).

5.2 Regularized Joint Harmonic Estimators $\gamma > 0$

If the Joint Optimization Problem (24) is regularized with some $\gamma > 0$, the resulting joint estimator is unique. This estimator is built off of "regularized Schur complements"

$$\boldsymbol{W}^{\star}_{LL_\gamma} = \boldsymbol{W}_{LL} - \boldsymbol{W}_{LU}\boldsymbol{W}^{-}_{UU_\gamma}\boldsymbol{W}_{UL}, \tag{27}$$

$$\boldsymbol{\Delta}^{\star}_{LL_\gamma} = \boldsymbol{\Delta}_{LL} - \boldsymbol{\Delta}_{LU}\boldsymbol{\Delta}^{-}_{UU_\gamma}\boldsymbol{\Delta}_{UL}, \tag{28}$$

where the "regularized inverses"

$$\boldsymbol{W}^{-}_{UU_\gamma} = \left(\boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UU} + \gamma\boldsymbol{I}\right)^{-1}\left(\boldsymbol{I} - \boldsymbol{S}_{UU}\right)^{T}, \tag{29}$$

$$\boldsymbol{\Delta}^{-}_{UU_\gamma} = \left(\boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UU} + \gamma\boldsymbol{I}\right)^{-1}\boldsymbol{S}^{T}_{UU}. \tag{30}$$

If $\gamma = 0$, $\rho(\boldsymbol{S}_{UU}) < 1$, and $\rho\left(\boldsymbol{I} - \boldsymbol{S}_{UU}\right) < 1$, then $\boldsymbol{W}^{-}_{UU_0} = \boldsymbol{W}^{-1}_{UU}$ and $\boldsymbol{\Delta}^{-}_{UU_0} = \boldsymbol{\Delta}^{-1}_{UU}$, so regularized Schur complements in Displays (27) and (28) simplify to the Schur complements on the right of Display (25). It is also easily verified that

$$\boldsymbol{\Gamma}_{LL_\gamma} = \left(\boldsymbol{W}^{\star}_{LL_\gamma} + \boldsymbol{\Delta}^{\star}_{LL_\gamma}\right)^{-1}\boldsymbol{W}^{\star}_{LL_\gamma} \quad \text{(Regularized Joint Smoother)}$$

exists for any $\gamma > 0$. Theorem 21 extends Theorem 18 from $\gamma = 0$ to $\gamma > 0$.

Theorem 21 *Let $\boldsymbol{W} \succeq 0$. Let f_γ denote the solution to the Joint Harmonic Optimization Problem (24) with $\gamma > 0$. Then*

$$f_\gamma = \begin{pmatrix} \boldsymbol{\Gamma}_{LL_\gamma} \\ -\left(\boldsymbol{\Delta}^{-}_{UU_\gamma}\right)^{T}\boldsymbol{\Delta}_{UL}\boldsymbol{\Gamma}_{LL_\gamma} + \left(\boldsymbol{I} - \left(\boldsymbol{\Delta}^{-}_{UU_\gamma}\right)^{T}\boldsymbol{\Delta}_{UU}\right)\boldsymbol{S}_{UL} \end{pmatrix} Y_L.$$

The Theorem 21 decomposition is a compromise between the semi-supervised harmonic estimator (labeled connective) and supervised kernel estimator (labeled adjacent)

$$f_{U_\gamma} = \underbrace{-\left(\boldsymbol{\Delta}^{-}_{UU_\gamma}\right)^{T}\boldsymbol{\Delta}_{UL}f_{L_\gamma}}_{\text{Harmonic Part}} + \underbrace{\left(\boldsymbol{I} - \left(\boldsymbol{\Delta}^{-}_{UU_\gamma}\right)^{T}\boldsymbol{\Delta}_{UU}\right)\boldsymbol{S}_{UL}Y_L}_{\text{Supervised Part}}.$$

In the case of $\gamma = 0$, the harmonic part reduces to the harmonic estimator, and the supervised part equals zero. On the other extreme, as $\gamma \to \infty$, the harmonic part converges to zero, while the supervised part has limit

$$f_\gamma \to f_\infty = \boldsymbol{S}Y\left(\vec{0}\right) = \begin{pmatrix} \boldsymbol{S}_{LL} \\ \boldsymbol{S}_{UL} \end{pmatrix} Y_L = \begin{pmatrix} \boldsymbol{Q}_L\widetilde{\boldsymbol{S}}_{LL} \\ \left(\boldsymbol{I} - \boldsymbol{Q}_U\right)\widetilde{\boldsymbol{S}}_{UL} \end{pmatrix} Y_L = \begin{pmatrix} \boldsymbol{Q}_L\widetilde{f}_L \\ \left(\boldsymbol{I} - \boldsymbol{Q}_U\right)\widetilde{f}_U \end{pmatrix}, \tag{31}$$

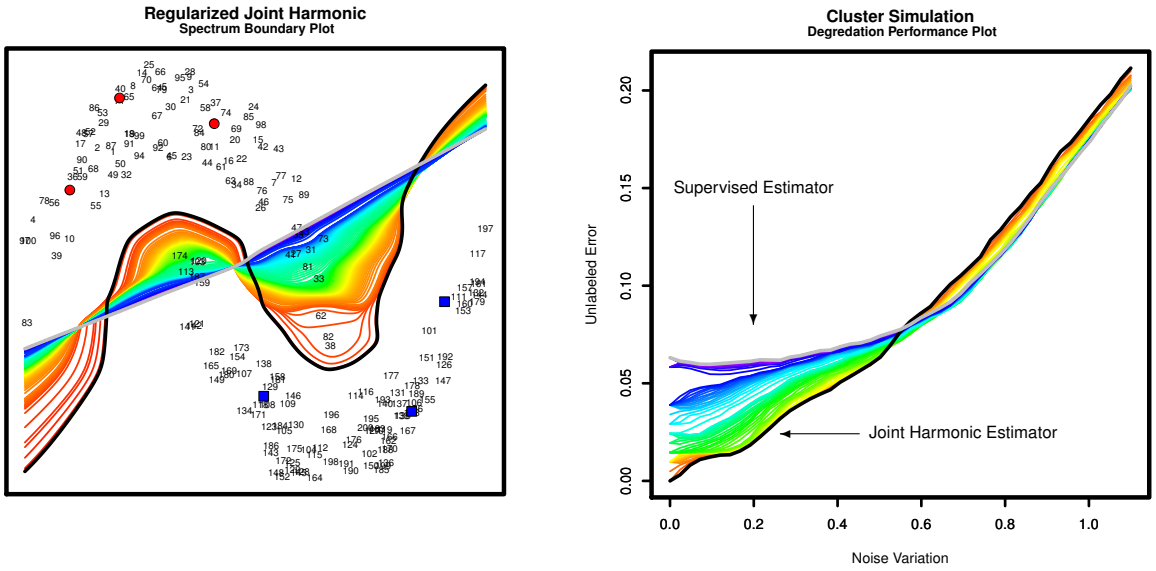

Figure 3: The "two moons" data from Figure 2 with regularized joint harmonic classification boundary curves on left. Noise degradation study on right. Black: *gamma* $= 0$ (harmonic extreme). Gray: $\gamma = \infty$ (supervised extreme). Rainbow spectrum: ordered by $\gamma \in (0, \infty)$.

where diagonal matrix $\boldsymbol{Q}_U$ has $\boldsymbol{Q}_{U_{ii}} = \sum_{j \in U} \boldsymbol{W}_{ij} / \sum_{j \in L \cup U} \boldsymbol{W}_{ij}$ for $i \in U$ and $\boldsymbol{Q}_L$ is defined analogously on L (apply Remark 4 when entries in $\widetilde{f}_U$ do not exist). Each estimator is a multiple of the supervised case by the right of Equation (31), so $\lim_{\gamma \to \infty} \text{sign}(f_{\gamma_i}) = \text{sign}(\widetilde{f}_i)$ for every i in the context of a classification problem with $Y_L \in \{-1, 1\}^{|L|}$.

The "two moons" data from Figure 2 are now revisited in Figure 3. The black joint harmonic function ($\gamma = 0$) and the gray supervised extreme ($\gamma = \infty$) borders in the left panel of Figure 3 correspond to the harmonic and supervised borders in Figure 2 as expected. The rainbow spectrum of borders rely less on the interior network and more on local supervised estimates as γ increases. Now, suppose the "two moons" data were instead observed with noise around each observation. Independent random samples from $\text{N}(0, \sigma^2)$ were added to each coordinate after scaling each axis in the left panel to sample standard deviation one. The regularized joint harmonic estimate was computed for each γ and σ over a grid, and unlabeled errors were recorded over this grid assuming the "truth" of a constant labeling by moon in the $\sigma = 0$ noiseless data on left. This was repeated 50 times, and average unlabeled error rates versus noise variation σ are plotted by γ in the right panel of Figure 3. While the joint harmonic function and the supervised solution are optimal for small and large σ, compromise solutions are best for data with an intermediate level of noise. Overall, the regularized joint harmonic estimator is a compromise between the harmonic estimator (which emphasizes unlabeled connectivity to labeled cases) and the supervised estimator (which requires unlabeled adjacency to labeled cases).

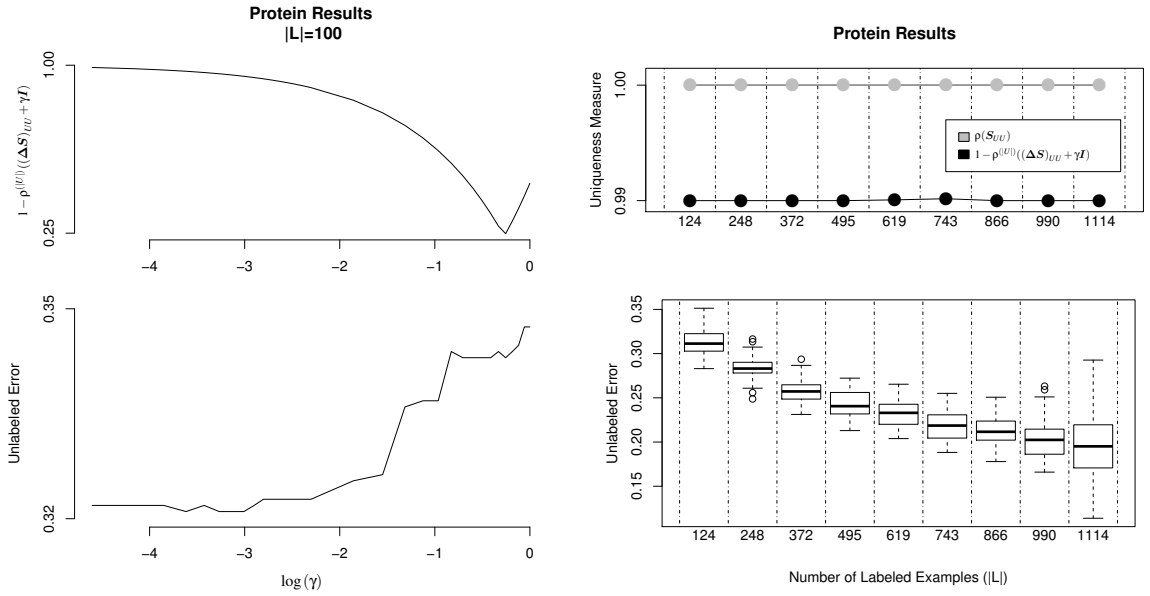

Figure 4: Regularized joint harmonic analyses of the protein data. Left: Uniqueness condition (top) and performance measure (bottom) versus regularization parameter $\log(\gamma)$ with a labeled set of size $|L| = 100$. Right: Uniqueness measure (top) and performance (bottom) for each of 50 replicates at each $|L|$ tested.

5.3 Joint Training Connections

The regularized joint harmonic estimator is the solution to a particular version of a *generalized joint training* optimization problem

$$\min_{Y_U, f} L(Y(Y_U), f) + \eta J_1(f) + \gamma J_2(Y_U) \tag{32}$$

with $L(y, f)$ a loss function, $J_1(f) \geq 0$ a penalty term independent of Y_U with $\eta \geq 0$, and $J_2(Y_U) \geq 0$ a penalty term independent of f with $\gamma \geq 0$. It is clear how to choose $L(\cdot,\cdot)$, $J_1(\cdot)$, and $J_2(\cdot)$ so that the generalized problem from Display (32) simplifies to the problem in Display (24). The S^3VM (Chapelle et al., 2006a) is approximated by setting $L(\cdot,\cdot)$ as a diagonally weighted hinge loss function with $L(Y(Y_U), f) = c_1 \sum_{i \in L}(1 + Y_i f_i)_+ + c_2 \sum_{i \in U}(1 + Y_i f_i)_+$ for $c_1, c_2 \in \mathbb{R}^+$, optimizing Y_U in a binary space, setting $J_1(f)$ as a quadratic ambient penalty, and forcing $\gamma = 0$. In this case, $\sum_{i \in U}(1 + Y_i f_i)_+$ is referred to as an interplay penalty between Y_U and f_U. The SSVM and SPSI algorithms are also construed as approximations of Optimization (32) (Wang and Shen, 2007). Lastly, linear joint training was proposed in Culp (2013) to extend the elastic net and other linear approaches into the semi-supervised setting.

6. Protein Interaction Data

Data on $n = 1237$ proteins from yeast organisms were collected. Each of 13 systems was used to detect the presence of protein-to-protein interactions (Kui et al., 2002). Adjacencies in $\boldsymbol{W}$ are taken

to be the proportion of systems detecting an interaction, so

$$\boldsymbol{W}_{ij} = \begin{cases} \frac{1}{13}\sum_s I_{\{\text{system } s \text{ detected an interaction between proteins } i,j\}} & i \neq j \\ 1 & i = j. \end{cases}$$

An important yet difficult problem is to classify whether or not a protein is located on the nucleus of a cell (Yamanishi et al., 2004). A number of analyses of $\boldsymbol{W}$ using the regularized joint harmonic estimator are presented in Figure 4. All 1237 proteins were included in each analysis, but the definition of the boundary was altered. The clamped harmonic and joint harmonic approaches are singular in each of these analyses, whereas the regularization strategy posed for the joint harmonic estimator provides the practical benefit of a well-defined classifier with a unique solution. Furthermore, the protein interaction graph $\boldsymbol{W}$ was observed directly, so there is no tuning parameter for either harmonic estimator.

Boundary L is 100 randomly selected proteins in the left panels of Figure 4. Since $\rho(\boldsymbol{S}_{UU}) = 1$, any harmonic estimator is singular. On the other hand, the regularized joint harmonic estimator is applicable with large enough γ so that $(\boldsymbol{\Delta S})_{UU} + \gamma \boldsymbol{I}$ is invertible, that is, when $\rho^{(|U|)}((\boldsymbol{\Delta S})_{UU} + \gamma \boldsymbol{I}) > 0$ in the top panel. The corresponding unlabeled error performance as a function of $\log(\gamma)$ is plotted in the bottom panel.

Consider now the analyses in the right panels of Figure 4. Proportion $|L|/n$ was varied from 0.1 to 0.9 by 0.1, and an analysis like that on the left was run for each of 50 randomly selected boundary sets at each $|L|$. The top right panel shows that the spectral radius uniqueness assumption was violated for any harmonic estimator, for example, the clamped or $\gamma = 0$, whereas regularization of the joint harmonic approach identified a well-defined classifier. The corresponding testing errors indicate a trend toward improved performance as the size of the labeled set increases in the bottom panel.

7. Machine Learning Data Sets

A comparison of procedures was based on three data sets from the UCI repository (Frank and Asuncion, 2010), that is, the ionosphere data set with $n = 351$ observations, thyroid data $n = 215$, and breast cancer data $n = 699$, and a publicly available pharmaceutical solubility data set with $n = 5631$ (Izenman, 2008). Missing values within the solubility data were handled by mean imputation. The $|L \cup U| \times |L \cup U|$ matrix $\boldsymbol{W}$ was computed from $\boldsymbol{X}$ feature data using the Gaussian kernel function, that is, $\boldsymbol{W}_{ij} = K_\lambda(x_i, x_j)$. Five-fold cross-validation was used to estimate $(\hat{\lambda}, \hat{\gamma})$ for the regularized joint harmonic function and $\hat{\lambda}$ for the clamped harmonic estimator. A semi-supervised SVM (S^3VM) with a linear kernel was also fit; its cost and gamma parameters were estimated using cross-validation with the *svm.tune* function from **R** library *e1071* (R Core Team, 2012; Meyer et al., 2012).

A transductive comparison is provided by Figure 5. The ionosphere and thyroid data were each randomly partitioned into L and U sets 50 times for each $|L| = 10, 20, 30, 40, 50$, and the techniques were all run on the same L and U partitions. The top and middle panels of Figure 5 summarize a particular example with $|L| = 20$ from the corresponding bottom panel. The clamped harmonic estimator is computationally singular and cannot be computed when $\rho(\boldsymbol{S}_{UU}) \approx 1$ (see Remark 6). This occurs for any $\lambda < 0.3$, that is, $\log(\lambda) < -1.2$, in the ionosphere application and for any $\lambda < 0.2$, that is, $\log(\lambda) < -1.6$, in the thyroid application. The joint harmonic estimator ($\gamma = 0$) requires the more stringent assumption $(\boldsymbol{\Delta S})_{UU} \succ 0$, and it was singular for all λ in the ionosphere

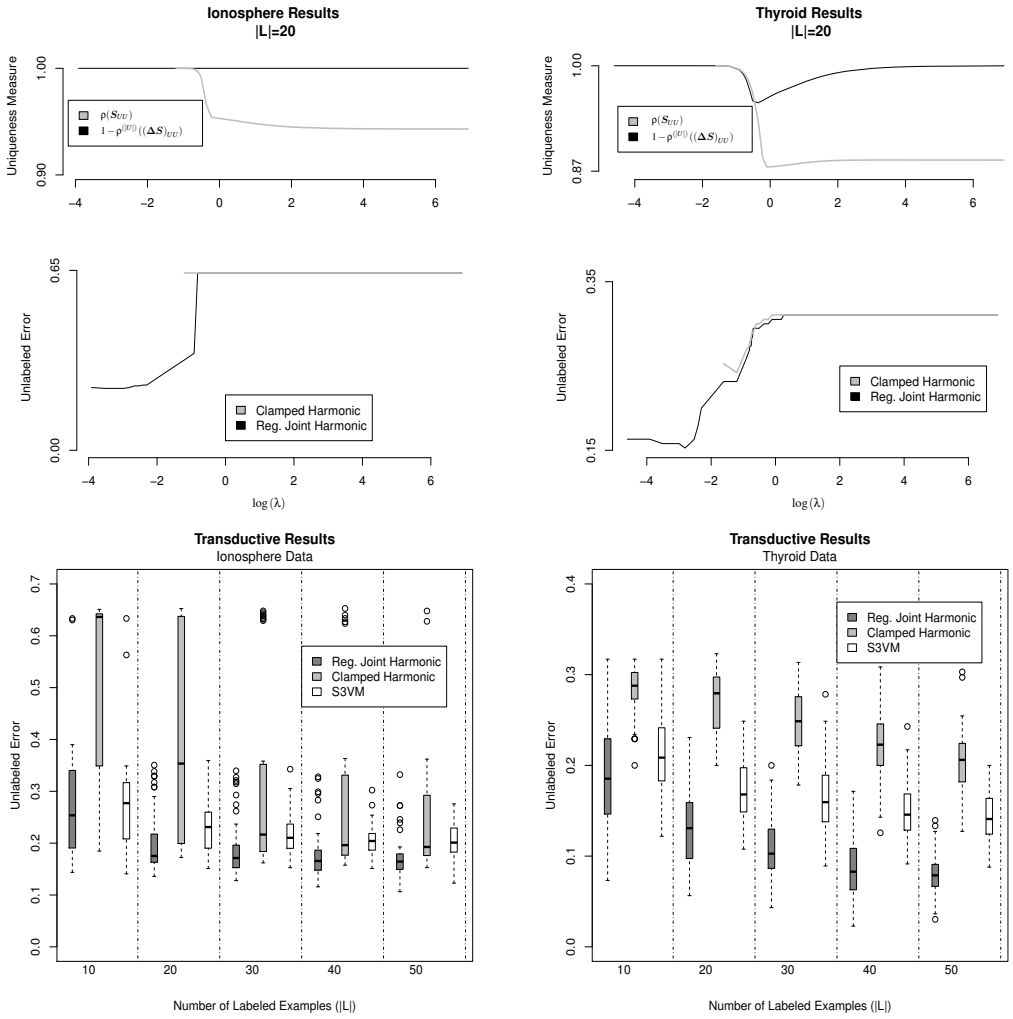

Figure 5: Transductive results for the ionosphere (left) and thyroid (right) data sets. Uniqueness measure (top) and unlabeled error performance (middle) each versus kernel parameter $\log(\lambda)$ for a particular analysis with $|L| = 20$ from the bottom panels. Unlabeled error rate performance (bottom) of the regularized joint harmonic, clamped harmonic, and S^3VM estimators for 50 randomly selected labeled sets L of each size $|L| = 10, 20, 30, 40, 50$.

application. However, estimates $\hat{\gamma} = 0.5$ and $\hat{\gamma} = 0.04$ in the ionosphere and thyroid applications were obtainable with the regularized joint harmonic estimator. Its access to a wider range of values λ, especially small λ, may yield substantial improvement in performance in other applications, like that seen in the bottom panels of Figure 5. As expected, a substantial performance gap exists between the regularized joint harmonic estimator and the clamped harmonic estimator. The S^3VM also outperformed the clamped harmonic estimator.

A semi-supervised comparison is provided by Figure 6. The data were first randomly partitioned into "seen" (25%) and "unseen" (75%) cases. The seen cases $L \cup U$ were then randomly partitioned into sets L and U of each size $|L| = 10, 20, 30, 40, 50$. The techniques were all run on the same

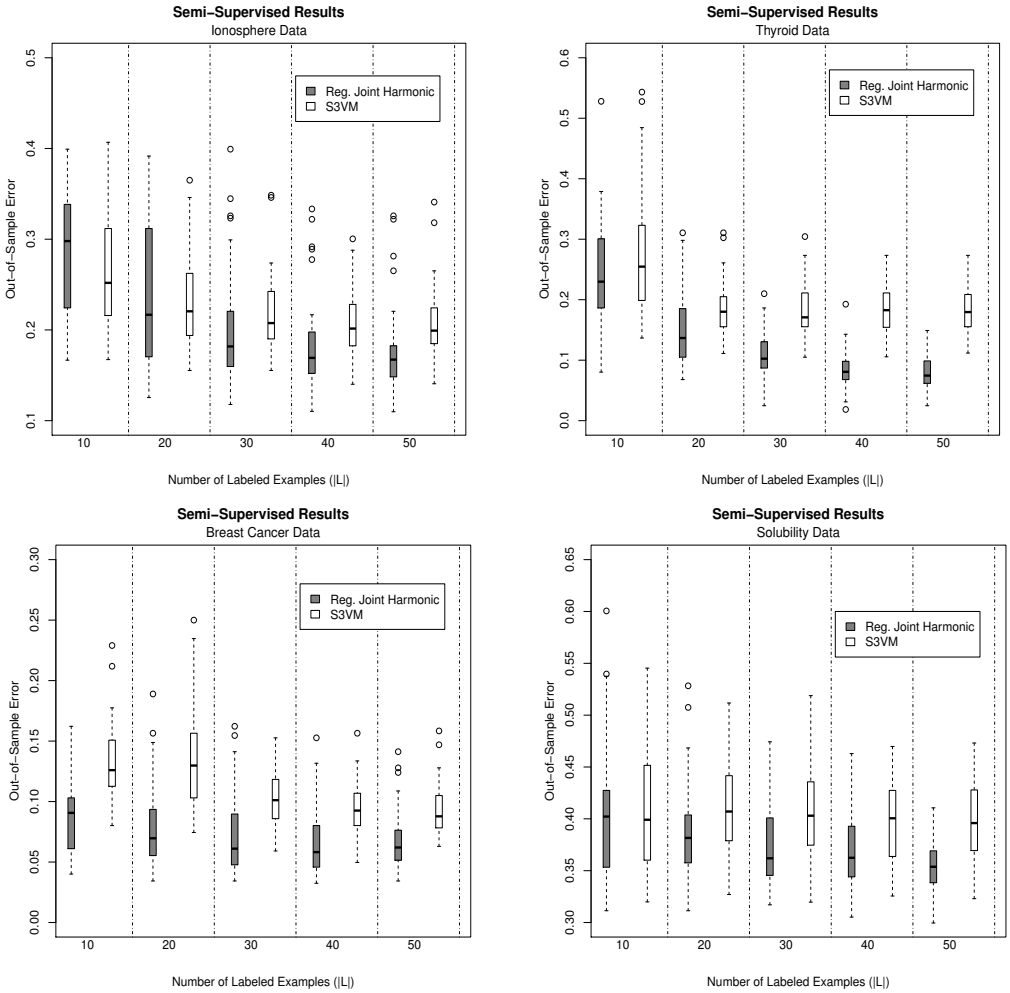

Figure 6: Semi-supervised out-of-sample error rate performance of the regularized joint harmonic and S^3VM estimators on four publicly available data sets. Each randomly obtained out-of-sample extension was 75% of cases. The other 25% were treated as $L \cup U$. Labeled sets L of each size $|L| = 10, 20, 30, 40, 50$ were also obtained randomly prior to cross-validation. This entire process was repeated 50 times.

"unseen," L, and U partitions, and the entire process was repeated 50 times. The clamped harmonic estimator is no longer applicable. Semi-supervised performance comparisons (Figure 6) of the regularized joint harmonic approach to the S^3VM are consistent with the transductive case (Figure 5), and the variability of the error measure increased in the out-of-sample extension as expected. In short, Figures 5 and 6 include real, low labeled sample size, transductive and semi-supervised applications, and the competitive stature of our proposed regularized joint harmonic estimator holds.

Remark 6 *Assumptions for the uniqueness of the clamped harmonic and the regularized joint harmonic approaches depend on the denseness or sparseness of the $\boldsymbol{W}_{UL}$ component of similarity graph $\boldsymbol{W}$. Sparseness makes the needed eigenvalue conditions more difficult to satisfy. One might expect a*

more sparse $\boldsymbol{W}_{UL}$ component when the labeled set size $|L|$ is small relative to the unlabeled set size $|U|$. As kernel parameter λ decreases, the off-diagonal elements of $\boldsymbol{W}$ approach 0, and this forces computational zeros in matrix $\boldsymbol{W}_{UL}$ leading to less stable estimators for any harmonic estimator. This follows since $\boldsymbol{S}\vec{1} = \vec{1}$, and so if $\boldsymbol{S}_{UL}\vec{1} \approx \vec{0}$, then $\boldsymbol{S}_{UU}\vec{1} \approx \vec{1}$. Hence, $\rho(\boldsymbol{S}_{UU}) \approx 1$ in the sparse case. On the other hand, larger values of λ allow the potential for a denser $\boldsymbol{W}_{UL}$ component which potentially makes the eigenvalue assumptions less stringent. The parameter λ is estimated using five-fold cross-validation, which does not account for assumptions on $\boldsymbol{W}_{UL}$. Regularization within the joint harmonic approach has the key advantage of a unique estimator for any $\lambda > 0$.

8. Conclusion

Semi-supervised harmonic estimation for graph-based semi-supervised learning was examined theoretically and empirically. A cluster assumption classifier was also defined, and it was shown that such classifiers assign labels to data that conform to the cluster assumption in the logical manner. Harmonic functions with a well-chosen boundary are examples of cluster assumption classifiers. In addition, harmonic functions were shown to be weighted averages of local supervised estimators applied to the interior. This work further established that harmonic estimators rely primarily on connectivity within the unlabeled network to form predictions using local supervised estimators; supervised estimates near labeled cases are up-weighted while supervised estimates deep within the network are down-weighted. Another key contribution, the development of the regularized joint harmonic function approach, used a joint optimization criterion with regularization to automate the trade-off between labeled connectivity versus labeled adjacency. Empirical results demonstrated the practical benefit gained by regularization of joint harmonic estimation.

Acknowledgments

The authors thank the AE and anonymous referees for their useful comments and suggestions. The work of Mark Vere Culp was supported in part by the NSF CAREER/DMS-1255045 grant. The work of Kenneth Joseph Ryan was supported in part by the U.S. Department of Justice 2010-DD-BX-0161 grant. The opinions and views expressed in this paper are those of the authors and do not reflect the opinions or views at either the NSF or the U.S. Department of Justice.

Appendix A. Proofs

Proofs of Lemmas, Propositions, and Theorems follow.

A.1 Problem Set-Up

Proposition 7 *Laplacian $\boldsymbol{\Delta} \succeq 0$.*

Proof Matrix $\boldsymbol{\Delta}$ satisfies $\boldsymbol{\Delta}_{ii} = \sum_{k=1}^{n} \boldsymbol{W}_{ik} I_{\{i \neq k\}} \geq \boldsymbol{W}_{ij} = -\boldsymbol{\Delta}_{ij} \geq 0$ for each $i \neq j$, and such symmetric, diagonally dominant Z-matrices are positive semi-definite. ■

Proposition 8 *If $\boldsymbol{W} \succeq 0$ then each eigenvalue of $\boldsymbol{S} = \boldsymbol{D}^{-1}\boldsymbol{W}$ is an element of $[0,1]$.*

Proof Matrices $\boldsymbol{S}$ and $\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2} \succeq 0$ have the same eigenvalues, so the eigenvalues of $\boldsymbol{S}$ are bounded below by 0. Proposition 7 implies $\boldsymbol{D}^{-1/2}\boldsymbol{\Delta}\boldsymbol{D}^{-1/2} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2} \succeq 0$, so the eigenvalues of $\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ and hence $\boldsymbol{S}$ are also bounded above by 1. ■

Lemma 9 *If $\boldsymbol{W} \succeq 0$ then each eigenvalue of $\boldsymbol{S}_{UU}$ is an element of $[0,1]$.*

Proof Define $\boldsymbol{I}_U = \operatorname{diag}\left(\mathbf{1}_{\{i \in U\}}\right)$ based on the binary vector $\mathbf{1}_{\{i \in U\}} \in \mathbb{R}^{|L \cup U|}$. Matrices $\boldsymbol{S}_{UU}$ and $\left(\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}\right)_{UU} = \boldsymbol{I}_U\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}\boldsymbol{I}_U \succeq 0$ have the same eigenvalues, so

$$\rho(\boldsymbol{S}_{UU}) = \rho\left(\boldsymbol{I}_U\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}\boldsymbol{I}_U\right) \leq \rho\left(\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}\right) \leq 1,$$

where the second inequality was justified during the proof of Proposition 8. ■

Proposition 10 *If $\boldsymbol{W} \succeq 0$ then the following conditions are equivalent.*

(a) $\boldsymbol{\Delta}_{UU} \succ 0$.

(b) $\rho(\boldsymbol{S}_{UU}) < 1$.

(c) $\nu^T \widetilde{\boldsymbol{D}}_{UL}\nu > 0$ *for any non-zero* $\nu \in \mathcal{N}(\widetilde{\boldsymbol{D}}_{UU} - \boldsymbol{W}_{UU})$.

Proof [(a) $\Longleftrightarrow$ (b)]: This equivalence follows by taking inverses of $\boldsymbol{\Delta}_{UU} = \boldsymbol{D}_{UU}\left(\boldsymbol{I} - \boldsymbol{S}_{UU}\right)$. Condition (a) implies $\mathcal{N}(\boldsymbol{\Delta}_{UU}) = \left\{\vec{1}\right\}$, so condition (b) follows because the Lemma 9 upper bound of 1 for the largest eigenvalue of $\boldsymbol{S}_{UU}$ cannot be achieved. Condition (b) implies the existence of $(\boldsymbol{I} - \boldsymbol{S}_{UU})^{-1}$ by a geometric matrix series, and so condition (a) follows.

[(a) $\Longleftrightarrow$ (c)]: Proposition 7 implies $\boldsymbol{\Delta}_{UU} \succeq 0$, so if $\nu \in \mathcal{N}\left(\widetilde{\boldsymbol{D}}_{UU} - \boldsymbol{W}_{UU}\right)$,

$$\nu^T\boldsymbol{\Delta}_{UU}\nu = \nu^T\widetilde{\boldsymbol{D}}_{UL}\nu + \nu^T\left(\widetilde{\boldsymbol{D}}_{UU} - \boldsymbol{W}_{UU}\right)\nu > 0 \iff \nu^T\widetilde{\boldsymbol{D}}_{UL}\nu > 0.$$

■

A.2 Regularized Joint Harmonic Functions

Proposition 11 *If $\boldsymbol{W} \succeq 0$ then $\boldsymbol{\Delta}\boldsymbol{S} \succeq 0$.*

Proof Define the matrix

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{W} & \boldsymbol{W} \\ \boldsymbol{W} & \boldsymbol{D} \end{pmatrix}, \text{ and let } \nu = \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} \in \mathbb{R}^{2|L \cup U|} \text{ with } \nu \neq \vec{0}. \tag{33}$$

Since $\nu^T\boldsymbol{V}\nu = \nu_1^T\boldsymbol{W}\nu_1 + \nu_1^T\boldsymbol{W}\nu_2 + \nu_2^T\boldsymbol{W}\nu_1 + \nu_2^T\boldsymbol{D}\nu_2 = (\nu_1+\nu_2)^T\boldsymbol{W}(\nu_1+\nu_2) + \nu_2^T\boldsymbol{\Delta}\nu_2 \geq 0$, the $\boldsymbol{D}$ block Schur complement of $\boldsymbol{V}$ is positive semi-definite, that is, $\boldsymbol{W} - \boldsymbol{W}\boldsymbol{D}^{-1}\boldsymbol{W} = \boldsymbol{\Delta}\boldsymbol{S} \succeq 0$. ■

Proposition 12 *Let $\boldsymbol{W} \succeq 0$. Assume $(\boldsymbol{\Delta S})_{UU} \succ 0$ when one selects $\gamma = 0$; this additional assumption is not required when one selects some $\gamma > 0$. The unique solution to the Joint Harmonic Optimization Problem (24) is $(Y_U, f) = \left(\hat{Y}_{U_\gamma}, \boldsymbol{SY}\left(\hat{Y}_{U_\gamma}\right)\right)$, where*

$$\hat{Y}_{U_\gamma} = -\left((\boldsymbol{\Delta S})_{UU} + \gamma \boldsymbol{I}\right)^{-1} (\boldsymbol{\Delta S})_{UL} Y_L.$$

Proof The solution is unique if the scores of the quadratic in (Y_U, f) objective function are non-degenerate. After some rearrangement, the scores with respect to Y_U and f are

$$\begin{aligned} \boldsymbol{S}_{UU}(\hat{Y}_{U_\gamma} - f_U) + \boldsymbol{S}_{UL}(Y_L - f_L) + \gamma \boldsymbol{D}_{UU}^{-1} \hat{Y}_{U_\gamma} &= \vec{0} \quad (34) \\ f(Y_U) &= \boldsymbol{SY}(Y_U), \quad (35) \end{aligned}$$

and plugging the f_U portion of Vector (35) into Unlabeled Score (34) produces

$$\begin{aligned} \boldsymbol{D}_{UU}^{-1}\left(\gamma \boldsymbol{I} + \boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UU} + \boldsymbol{\Delta}_{UL}\boldsymbol{S}_{LU}\right)\hat{Y}_{U_\gamma} &= -\boldsymbol{D}_{UU}^{-1}\left(\boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UL} + \boldsymbol{\Delta}_{UL}\boldsymbol{S}_{LL}\right) Y_L \\ \hat{Y}_{U_\gamma} &= -\left((\boldsymbol{\Delta S})_{UU} + \gamma \boldsymbol{I}\right)^{-1} (\boldsymbol{\Delta S})_{UL} Y_L. \end{aligned}$$

Matrix $(\boldsymbol{\Delta S})_{UU} + \gamma \boldsymbol{I} \succ 0$ by Proposition 11 when $\gamma > 0$ and by assumption when $\gamma = 0$, so its inverse exists. Substitution of $Y_U = \hat{Y}_{U_\gamma}$ into Equation (35) results in $f = \boldsymbol{SY}\left(\hat{Y}_{U_\gamma}\right)$. ■

A.3 Joint Harmonic Estimator $\gamma = 0$

Lemma 13 *If $\boldsymbol{W} \succeq 0$ then $\boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UU} = \boldsymbol{D}_{UU}\left(\boldsymbol{I} - \boldsymbol{S}_{UU}\right)\boldsymbol{S}_{UU} \succeq 0$. In addition,*

$$\boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UU} \succ 0 \iff \rho(\boldsymbol{S}_{UU}) < 1 \text{ and } \rho\left(\boldsymbol{I} - \boldsymbol{S}_{UU}\right) < 1.$$

Proof In Display (33), substitute $\boldsymbol{W}_{UU}$ for $\boldsymbol{W}$ and $\boldsymbol{D}_{UU}$ for $\boldsymbol{D}$ and take $\nu \in \mathbb{R}^{2|U|}$. Then

$$\boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UU} \succeq 0 \iff (\nu_1 + \nu_2)^T \boldsymbol{W}_{UU}(\nu_1 + \nu_2) + \nu_2^T \boldsymbol{\Delta}_{UU}\nu_2 \geq 0. \quad (36)$$

One can set $\nu_2 = \vec{0}$ or $\nu_1 + \nu_2 = \vec{0}$ such that $\nu \neq \vec{0}$, so both inequalities in Display (36) are strict if and only if $\boldsymbol{\Delta}_{UU} \succ 0$ and $\boldsymbol{W}_{UU} \succ 0$. Furthermore, $\boldsymbol{\Delta}_{UU} \succ 0 \iff \rho(\boldsymbol{S}_{UU}) < 1$ by Proposition 10, and $\boldsymbol{W}_{UU} \succ 0 \iff \rho\left(\boldsymbol{I} - \boldsymbol{S}_{UU}\right) < 1$ by Lemma 9. ■

Lemma 14 *Let $\boldsymbol{W} \succeq 0$. Also, assume $\boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UU} \succ 0$, so $\boldsymbol{A} = \boldsymbol{S}_{LU}\boldsymbol{S}_{UU}^{-1}\left(\boldsymbol{I} - \boldsymbol{S}_{UU}\right)^{-1}\boldsymbol{S}_{UL}$ exists by Lemma 13. Then each eigenvalue of $\boldsymbol{A}$ is an element of $[0,1]$.*

Proof Each eigenvalue of $\boldsymbol{S}_{UU}$ is an element of $(0,1)$ by Lemma 9, since $\boldsymbol{W}_{UU} \succ 0$ rules out eigenvalues of 0 and $\boldsymbol{\Delta}_{UU} \succ 0$ eigenvalues of 1 by Proposition 10. Furthermore, the UU block Schur complements $\boldsymbol{\Delta}_{LL}^\star$ and $\boldsymbol{W}_{LL}^\star$ are each positive semi-definite, so

$$\boldsymbol{B}_1 = \boldsymbol{D}_{LL}^{-1/2}\left(\boldsymbol{W}_{LL}^\star + \boldsymbol{\Delta}_{LL}^\star\right)\boldsymbol{D}_{LL}^{-1/2} \succeq 0. \quad (37)$$

By assumption (and application of Lemma 13), $\boldsymbol{D}_{UU}\left(\boldsymbol{I} - \boldsymbol{S}_{UU}\right)^{-1}\boldsymbol{S}_{UU}^{-1} \succ 0$, so since a row of $\boldsymbol{W}_{UL}$ could be all zeros,

$$\boldsymbol{B}_2 = \boldsymbol{D}_{LL}^{-1/2}\boldsymbol{W}_{LU}\left(\boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UU}\right)^{-1}\boldsymbol{W}_{UL}\boldsymbol{D}_{LL}^{-1/2} \succeq 0.$$

Although tedious to establish, there is a simple relationship between $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$; that is,

$$\begin{aligned}
\boldsymbol{B}_2 &= \boldsymbol{D}_{LL}^{-1/2}\boldsymbol{W}_{LU}\boldsymbol{S}_{UU}^{-1}\left(\boldsymbol{I}-\boldsymbol{S}_{UU}\right)^{-1}\boldsymbol{S}_{UL}\boldsymbol{D}_{LL}^{-1/2} \\
&= \boldsymbol{D}_{LL}^{-1/2}\boldsymbol{W}_{LU}\boldsymbol{S}_{UU}^{-1}\boldsymbol{S}_{UL}\boldsymbol{D}_{LL}^{-1/2}+\boldsymbol{D}_{LL}^{-1/2}\boldsymbol{W}_{LU}\left(\boldsymbol{I}-\boldsymbol{S}_{UU}\right)^{-1}\boldsymbol{S}_{UL}\boldsymbol{D}_{LL}^{-1/2} \qquad (38) \\
&= \boldsymbol{D}_{LL}^{-1/2}\boldsymbol{W}_{LU}\boldsymbol{W}_{UU}^{-1}\boldsymbol{W}_{UL}\boldsymbol{D}_{LL}^{-1/2}+\boldsymbol{D}_{LL}^{-1/2}\boldsymbol{\Delta}_{LU}\boldsymbol{\Delta}_{UU}^{-1}\boldsymbol{\Delta}_{UL}\boldsymbol{D}_{LL}^{-1/2} \\
&= \boldsymbol{I}-\boldsymbol{D}_{LL}^{-1/2}\left(\left(\boldsymbol{D}_{LL}-\boldsymbol{W}_{LL}-\boldsymbol{\Delta}_{LU}\boldsymbol{\Delta}_{UU}^{-1}\boldsymbol{\Delta}_{UL}\right)+\left(\boldsymbol{W}_{LL}-\boldsymbol{W}_{LU}\boldsymbol{W}_{UU}^{-1}\boldsymbol{W}_{UL}\right)\right)\boldsymbol{D}_{LL}^{-1/2} \\
&= \boldsymbol{I}-\boldsymbol{B}_1,
\end{aligned}$$

where equality holds in Display (38) because $\boldsymbol{S}_{UU}^{-1}\left(\boldsymbol{I}-\boldsymbol{S}_{UU}\right)^{-1}=\boldsymbol{S}_{UU}^{-1}+\left(\boldsymbol{I}-\boldsymbol{S}_{UU}\right)^{-1}$.

The eigenvalues of $\boldsymbol{B}_2$ are bounded below by 0 because $\boldsymbol{B}_2 \succeq 0$ and bounded above by 1 because $\boldsymbol{B}_1 \succeq 0$ and $\boldsymbol{B}_2=\boldsymbol{I}-\boldsymbol{B}_1 \succeq 0$. This proof concludes by noting that $\boldsymbol{B}_2$ and $\boldsymbol{A}$ have the same eigenvalues since $\boldsymbol{B}_2\phi=\lambda\phi \iff \boldsymbol{A}\breve{\phi}=\lambda\breve{\phi}$, where $\breve{\phi}=\boldsymbol{D}_{LL}^{-1/2}\phi$. ■

Lemma 15 *If $\boldsymbol{W} \succeq 0$ then the following conditions are equivalent.*

(a) $(\boldsymbol{\Delta S})_{UU} \succ 0$.

(b) $\rho(\boldsymbol{S}_{UU})<1$, $\rho\left(\boldsymbol{I}-\boldsymbol{S}_{UU}\right)<1$, *and* $\rho(\boldsymbol{A})<1$, *where* $\boldsymbol{A}=\boldsymbol{S}_{LU}\boldsymbol{S}_{UU}^{-1}\left(\boldsymbol{I}-\boldsymbol{S}_{UU}\right)^{-1}\boldsymbol{S}_{UL}$.

(c) $\boldsymbol{W}_{UU} \succ 0$, $\boldsymbol{\Delta}_{UU} \succ 0$, *and* $(\boldsymbol{W}_{LL}^{\star}+\boldsymbol{\Delta}_{LL}^{\star}) \succ 0$.

(d) $\boldsymbol{\Gamma}_{LL}=(\boldsymbol{W}_{LL}^{\star}+\boldsymbol{\Delta}_{LL}^{\star})^{-1}\boldsymbol{W}_{LL}^{\star}$ *exists.*

Proof [(a) $\iff$ (b)]: Matrix $(\boldsymbol{\Delta S})_{UU} \succeq 0$ by Proposition 11. Also,

$$(\boldsymbol{\Delta S})_{UU}=\boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UU}-\boldsymbol{W}_{UL}\boldsymbol{D}_{LL}^{-1}\boldsymbol{W}_{LU}$$

is the $\boldsymbol{D}_{LL}$ block Schur complement of

$$\boldsymbol{V}_2=\begin{pmatrix}\boldsymbol{D}_{LL} & \boldsymbol{W}_{LU} \\ \boldsymbol{W}_{UL} & \boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UU}\end{pmatrix},$$

so condition (a) $\iff \boldsymbol{V}_2 \succ 0$. Hence, it suffices to show $\boldsymbol{V}_2 \succ 0 \iff$ condition (b). This follows because $\boldsymbol{V}_2 \succ 0 \iff$ the $\boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UU}$ block Schur complement of $\boldsymbol{V}_2$ is positive definite, that is, $\left(\boldsymbol{D}_{LL}-\boldsymbol{W}_{LU}\left(\boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UU}\right)^{-1}\boldsymbol{W}_{UL}\right)=\boldsymbol{D}_{LL}(\boldsymbol{I}-\boldsymbol{A}) \succ 0$. Recall $(\boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UU})^{-1} \iff \rho(\boldsymbol{S}_{UU})<1$ and $\rho\left(\boldsymbol{I}-\boldsymbol{S}_{UU}\right)<1$ by Lemma 13. Furthermore, the existence of $(\boldsymbol{I}-\boldsymbol{A})^{-1} \iff \rho(\boldsymbol{A})<1$ by Lemma 14 because $\boldsymbol{A}\nu=\lambda\nu \iff (\boldsymbol{I}-\boldsymbol{A})\nu=(1-\lambda)\nu$.

[(b) $\iff$ (c)]: By Lemma 13, $\rho(\boldsymbol{S}_{UU})<1$ and $\rho\left(\boldsymbol{I}-\boldsymbol{S}_{UU}\right)<1 \iff \boldsymbol{W}_{UU} \succ 0$ and $\boldsymbol{\Delta}_{UU} \succ 0$. Either set of these equivalent conditions implies

$$\begin{aligned}
\boldsymbol{D}_{LL}(\boldsymbol{I}-\boldsymbol{A}) &= \boldsymbol{D}_{LL}\left(\boldsymbol{I}-\boldsymbol{S}_{LU}\boldsymbol{S}_{UU}^{-1}\left(\boldsymbol{I}-\boldsymbol{S}_{UU}\right)^{-1}\boldsymbol{S}_{UL}\right) \\
&= \boldsymbol{D}_{LL}\left(\boldsymbol{I}-\boldsymbol{S}_{LU}\boldsymbol{S}_{UU}^{-1}\boldsymbol{S}_{UL}-\boldsymbol{S}_{LU}\left(\boldsymbol{I}-\boldsymbol{S}_{UU}\right)^{-1}\boldsymbol{S}_{UL}\right) \\
&= \boldsymbol{D}_{LL}\left(\boldsymbol{I}-\boldsymbol{S}_{LU}\left(\boldsymbol{I}-\boldsymbol{S}_{UU}\right)^{-1}\boldsymbol{S}_{UL}-\boldsymbol{S}_{LL}\right)+\boldsymbol{D}_{LL}\left(\boldsymbol{S}_{LL}-\boldsymbol{S}_{LU}\boldsymbol{S}_{UU}^{-1}\boldsymbol{S}_{UL}\right) \qquad (39) \\
&= \left(\boldsymbol{\Delta}_{LL}-\boldsymbol{\Delta}_{LU}\boldsymbol{\Delta}_{UU}^{-1}\boldsymbol{\Delta}_{UL}\right)+\left(\boldsymbol{W}_{LL}-\boldsymbol{W}_{LU}\boldsymbol{W}_{UU}^{-1}\boldsymbol{W}_{UL}\right) \\
&= \boldsymbol{W}_{LL}^{\star}+\boldsymbol{\Delta}_{LL}^{\star},
\end{aligned}$$

so $(\boldsymbol{W}^{\star}_{LL}+\boldsymbol{\Delta}^{\star}_{LL})^{-1}$ exists $\iff \rho(\boldsymbol{A})<1$.

[(c) $\iff$ (d)]: This follows automatically. ■

Proposition 16 *If* $\boldsymbol{W}\succeq 0$ *then*

$$(\boldsymbol{\Delta S})_{UU}\succ 0 \iff \boldsymbol{W}_{UU}\succ 0,\ \boldsymbol{\Delta}_{UU}\succ 0, \text{ and } (\boldsymbol{W}^{\star}_{LL}+\boldsymbol{\Delta}^{\star}_{LL})\succ 0.$$

Proof This is a special case of Lemma 15. ■

Lemma 17 *Let* $\boldsymbol{W}\succeq 0$*, and assume that* $\boldsymbol{\Gamma}_{LL}$ *exists. An equivalent form to that in Proposition 12 for the labeled solution to the joint training problem in Display (24) with* $\gamma=0$ *is* $f_L=\boldsymbol{\Gamma}_{LL}Y_L$.

Proof By Proposition 12 with $\gamma=0$, the joint training labeled estimator is

$$f_L=\left(\boldsymbol{S}_{LL}-\boldsymbol{S}_{LU}(\boldsymbol{\Delta S})^{-1}_{UU}(\boldsymbol{\Delta S})_{UL}\right)Y_L. \tag{40}$$

Now, it follows from some matrix algebra that

$$\begin{aligned}-(\boldsymbol{\Delta S})^{-1}_{UU}(\boldsymbol{\Delta S})_{UL} &= ((\boldsymbol{I}-\boldsymbol{S}_{UU})\boldsymbol{S}_{UU}-\boldsymbol{S}_{UL}\boldsymbol{S}_{LU})^{-1}(\boldsymbol{S}_{UL}\boldsymbol{S}_{LL}-(\boldsymbol{I}-\boldsymbol{S}_{UU})\boldsymbol{S}_{UL})\\ &= (\boldsymbol{I}-\boldsymbol{F})^{-1}(\boldsymbol{E}-\boldsymbol{S}^{-1}_{UU}\boldsymbol{S}_{UL}),\end{aligned} \tag{41}$$

where

$$\begin{aligned}\boldsymbol{E} &= \boldsymbol{S}^{-1}_{UU}(\boldsymbol{I}-\boldsymbol{S}_{UU})^{-1}\boldsymbol{S}_{UL}\boldsymbol{S}_{LL},\\ \boldsymbol{F} &= \boldsymbol{S}^{-1}_{UU}(\boldsymbol{I}-\boldsymbol{S}_{UU})^{-1}\boldsymbol{S}_{UL}\boldsymbol{S}_{LU}.\end{aligned}$$

Further simplification is based on an identity involving $\boldsymbol{A}$ from Lemma 14 and $\boldsymbol{F}$, that is,

$$\begin{aligned}\boldsymbol{S}_{LU}(\boldsymbol{I}-\boldsymbol{F})^{-1} &= \boldsymbol{S}_{LU}\left(\sum_{\ell=0}^{\infty}\left(\boldsymbol{S}^{-1}_{UU}(\boldsymbol{I}-\boldsymbol{S}_{UU})^{-1}\boldsymbol{S}_{UL}\boldsymbol{S}_{LU}\right)^{\ell}\right) &(42)\\ &= \left(\sum_{\ell=0}^{\infty}\left(\boldsymbol{S}_{LU}\boldsymbol{S}^{-1}_{UU}(\boldsymbol{I}-\boldsymbol{S}_{UU})^{-1}\boldsymbol{S}_{UL}\right)^{\ell}\right)\boldsymbol{S}_{LU} &(43)\\ &= (\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{S}_{LU}.\end{aligned}$$

The geometric matrix series in Display (43) converges because $\rho(\boldsymbol{A})<1$ by Lemma 15. Since $\boldsymbol{F}\nu=\lambda\nu \Longrightarrow \boldsymbol{A}\boldsymbol{S}_{LU}\nu=\lambda\boldsymbol{S}_{LU}\nu$ and $\nu^T\boldsymbol{A}=\lambda\nu^T \Longrightarrow \nu^T\boldsymbol{S}_{LU}\boldsymbol{F}=\lambda\nu^T\boldsymbol{S}_{LU}$, $\boldsymbol{F}$ and $\boldsymbol{A}$ have the same non-zero eigenvalues, so the infinite series in Display (42) is also well-defined.

Substitutions of Display (41) and $\boldsymbol{S}_{LU}(\boldsymbol{I}-\boldsymbol{F})^{-1}=(\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{S}_{LU}$ produce

$$\begin{aligned}\boldsymbol{S}_{LL}-\boldsymbol{S}_{LU}(\boldsymbol{\Delta S})^{-1}_{UU}(\boldsymbol{\Delta S})_{UL} &= \boldsymbol{S}_{LL}+\boldsymbol{S}_{LU}(\boldsymbol{I}-\boldsymbol{F})^{-1}(\boldsymbol{E}-\boldsymbol{S}^{-1}_{UU}\boldsymbol{S}_{UL})\\ &= \boldsymbol{S}_{LL}+(\boldsymbol{I}-\boldsymbol{A})^{-1}(\boldsymbol{S}_{LU}\boldsymbol{E}-\boldsymbol{S}_{LU}\boldsymbol{S}^{-1}_{UU}\boldsymbol{S}_{UL})\\ &= \boldsymbol{S}_{LL}+(\boldsymbol{I}-\boldsymbol{A})^{-1}(\boldsymbol{A}\boldsymbol{S}_{LL}-\boldsymbol{S}_{LU}\boldsymbol{S}^{-1}_{UU}\boldsymbol{S}_{UL})\\ &= \left(\boldsymbol{I}+(\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{A}\right)\boldsymbol{S}_{LL}-(\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{S}_{LU}\boldsymbol{S}^{-1}_{UU}\boldsymbol{S}_{UL}\\ &= (\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{S}^{\star}_{LL}.\end{aligned}$$

Therefore, the equivalent form $f_L = \mathbf{\Gamma}_{LL} Y_L = (\boldsymbol{W}^\star_{LL} + \boldsymbol{\Delta}^\star_{LL})^{-1} \boldsymbol{W}^\star_{LL} Y_L$ for Equation (40) is established using $\boldsymbol{D}_{LL}(\boldsymbol{I} - \boldsymbol{A}) = \boldsymbol{W}^\star_{LL} + \boldsymbol{\Delta}^\star_{LL}$ from Display (39) and $\boldsymbol{S}^\star_{LL} = \boldsymbol{D}^{-1}_{LL} \boldsymbol{W}^\star_{LL}$. ■

Theorem 18 *Let $\boldsymbol{W} \succeq 0$, and assume that $\mathbf{\Gamma}_{LL}$ exists. The solution to the Joint Harmonic Optimization Problem (24) with $\gamma = 0$ has*

$$f = \begin{pmatrix} f_L \\ (\boldsymbol{I} - \boldsymbol{S}_{UU})^{-1} \boldsymbol{S}_{UL} f_L \end{pmatrix} = \begin{pmatrix} \mathbf{\Gamma}_{LL} \\ (\boldsymbol{I} - \boldsymbol{S}_{UU})^{-1} \boldsymbol{S}_{UL} \mathbf{\Gamma}_{LL} \end{pmatrix} Y_L,$$

so f is in-fact harmonic.

Proof The optimal $\hat{Y}_U$ satisfies the derivative score in Display (34) with $\gamma = 0$, so

$$\hat{Y}_U = f_U - \boldsymbol{S}^{-1}_{UU} \boldsymbol{S}_{UL} (Y_L - f_L)$$

after rearrangement. Finally, since the optimal f satisfies $f = \boldsymbol{S} Y(\hat{Y}_U)$, f_U satisfies

$$\begin{aligned} f_U &= \boldsymbol{S}_{UL} Y_L + \boldsymbol{S}_{UU} \hat{Y}_U \\ &= \boldsymbol{S}_{UL} Y_L + \boldsymbol{S}_{UU} f_U - \boldsymbol{S}_{UL} (Y_L - f_L) \\ &= (\boldsymbol{I} - \boldsymbol{S}_{UU})^{-1} \boldsymbol{S}_{UL} f_L, \end{aligned}$$

and the optimal f_L satisfies $f_L = \mathbf{\Gamma}_{LL} Y_L$ by Lemma 17. ■

Proposition 19 *If $\boldsymbol{W} \succeq 0$ and $\mathbf{\Gamma}_{LL}$ exists then each eigenvalue of $\mathbf{\Gamma}_{LL}$ is an element of $[0, 1]$.*

Proof Since $\boldsymbol{W}_{UU} \succ 0$ by Lemma 15, $\boldsymbol{W} \succeq 0 \iff \boldsymbol{W}^\star_{LL} \succeq 0$, so it is well-defined to set

$$\boldsymbol{V}_3 = \begin{pmatrix} \boldsymbol{I} & {\boldsymbol{W}^\star_{LL}}^{1/2} \\ {\boldsymbol{W}^\star_{LL}}^{1/2} & \boldsymbol{W}^\star_{LL} + \boldsymbol{\Delta}^\star_{LL} \end{pmatrix}.$$

The $\boldsymbol{I}$ block Schur complement of $\boldsymbol{V}_3$ is $\boldsymbol{\Delta}^\star_{LL} \succeq 0$, so the other block is positive semi-definite, that is,

$$\boldsymbol{I} - {\boldsymbol{W}^\star_{LL}}^{1/2} (\boldsymbol{W}^\star_{LL} + \boldsymbol{\Delta}^\star_{LL})^{-1} {\boldsymbol{W}^\star_{LL}}^{1/2} \succeq 0,$$

and $\mathbf{\Gamma}_{LL}$ and ${\boldsymbol{W}^\star_{LL}}^{1/2} (\boldsymbol{W}^\star_{LL} + \boldsymbol{\Delta}^\star_{LL})^{-1} {\boldsymbol{W}^\star_{LL}}^{1/2} \succeq 0$ have the same eigenvalues. ■

A.4 Regularized Joint Harmonic Estimators $\gamma > 0$

Lemma 20 *Let $\boldsymbol{W} \succeq 0$ and $\gamma > 0$ and define*

$$\mathbf{\Gamma}_{LL_\gamma} = \left(\boldsymbol{W}^\star_{LL_\gamma} + \boldsymbol{\Delta}^\star_{LL_\gamma} \right)^{-1} \boldsymbol{W}^\star_{LL_\gamma}.$$

The labeled solution to the Joint Optimization Problem (24) is equivalently given by $f_{L_\gamma} = \mathbf{\Gamma}_{LL_\gamma} Y_L$.

Proof The sum of "regularized inverses" in Displays (29) and (30)

$$\boldsymbol{C}_\gamma = \boldsymbol{W}_{\overline{UU}_\gamma} + \boldsymbol{\Delta}^-_{\overline{UU}_\gamma} = (\boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UU} + \gamma\boldsymbol{I})^{-1}$$

is positive definite by Proposition 11, and

$$((\boldsymbol{\Delta S})_{UU} + \gamma\boldsymbol{I})^{-1}(\boldsymbol{\Delta S})_{UL} \;=\; \boldsymbol{G}_\gamma + \boldsymbol{H}_\gamma, \tag{44}$$

where

$$\begin{aligned}\boldsymbol{G}_\gamma &= \left(\boldsymbol{I} - \boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\boldsymbol{S}_{LU}\right)^{-1}\boldsymbol{C}_\gamma\boldsymbol{\Delta}_{UL}\boldsymbol{S}_{LL},\\ \boldsymbol{H}_\gamma &= \left(\boldsymbol{I} - \boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\boldsymbol{S}_{LU}\right)^{-1}\boldsymbol{C}_\gamma\boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UL}.\end{aligned}$$

Thus, by Proposition 12, labeled estimator f_L depends on

$$\boldsymbol{S}_{LL} - \boldsymbol{S}_{LU}((\boldsymbol{\Delta S})_{UU} + \gamma\boldsymbol{I})^{-1}(\boldsymbol{\Delta S})_{UL} = \boldsymbol{S}_{LL} - \boldsymbol{S}_{LU}\boldsymbol{G}_\gamma - \boldsymbol{S}_{LU}\boldsymbol{H}_\gamma. \tag{45}$$

Simplification of terms on the right of Equation (45) is based on

$$\begin{aligned}\left(\boldsymbol{I} - \boldsymbol{S}_{LU}\boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\right)^{-1}\boldsymbol{D}_{LL}^{-1} &= \left(\boldsymbol{D}_{LL} - \boldsymbol{W}_{LU}\boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\right)^{-1}\\ &= \left(\boldsymbol{D}_{LL} - \boldsymbol{\Delta}_{LU}\boldsymbol{\Delta}^-_{\overline{UU}_\gamma}\boldsymbol{\Delta}_{UL} - \boldsymbol{W}_{LU}\boldsymbol{W}^-_{\overline{UU}_\gamma}\boldsymbol{W}_{UL}\right)^{-1}\\ &= \left(\boldsymbol{W}^\star_{\overline{LL}_\gamma} + \boldsymbol{\Delta}^\star_{\overline{LL}_\gamma}\right)^{-1}\end{aligned}$$

and on

$$\boldsymbol{S}_{LU}\left(\boldsymbol{I} - \boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\boldsymbol{S}_{LU}\right)^{-1} = \left(\boldsymbol{I} - \boldsymbol{S}_{LU}\boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\right)^{-1}\boldsymbol{S}_{LU}$$

if $\rho\left(\boldsymbol{S}_{LU}\boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\right) < 1$ by a geometric matrix series argument similar to that used to establish Displays (42) and (43). Because $\gamma > 0$ is shrinking the eigenvalues of $\boldsymbol{C}_\gamma$, $\rho\left(\boldsymbol{S}_{LU}\boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\right) < 1$ as a consequence of a generalization of Lemma 14 since $\boldsymbol{B}_1$ is unique even if arbitrary generalized inverses are used to compute the Schur complements in Display (37). Now, terms on the right of Equation (45) reduce to

$$\begin{aligned}\boldsymbol{S}_{LL} - \boldsymbol{S}_{LU}\boldsymbol{G}_\gamma &= \left(\boldsymbol{I} + \boldsymbol{S}_{LU}\left(\boldsymbol{I} - \boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\boldsymbol{S}_{LU}\right)^{-1}\boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\right)\boldsymbol{S}_{LL}\\ &= \left(\boldsymbol{I} + \left(\boldsymbol{I} - \boldsymbol{S}_{LU}\boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\right)^{-1}\boldsymbol{S}_{LU}\boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\right)\boldsymbol{S}_{LL}\\ &= \left(\boldsymbol{I} - \boldsymbol{S}_{LU}\boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\right)^{-1}\boldsymbol{D}_{LL}^{-1}\boldsymbol{W}_{LL}\\ &= \left(\boldsymbol{W}^\star_{\overline{LL}_\gamma} + \boldsymbol{\Delta}^\star_{\overline{LL}_\gamma}\right)^{-1}\boldsymbol{W}_{LL}\end{aligned} \tag{46}$$

and

$$\begin{aligned}\boldsymbol{S}_{LU}\boldsymbol{H}_\gamma &= \boldsymbol{S}_{LU}\left(\boldsymbol{I} - \boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\boldsymbol{S}_{LU}\right)^{-1}\boldsymbol{C}_\gamma\boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UL}\\ &= \left(\boldsymbol{I} - \boldsymbol{S}_{LU}\boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\right)^{-1}\boldsymbol{S}_{LU}\boldsymbol{C}_\gamma(\boldsymbol{I} - \boldsymbol{S}_{UU})^T\boldsymbol{W}_{UL}\\ &= \left(\boldsymbol{I} - \boldsymbol{S}_{LU}\boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\right)^{-1}\boldsymbol{D}_{LL}^{-1}\left(\boldsymbol{W}_{LU}\boldsymbol{W}^-_{\overline{UU}_\gamma}\boldsymbol{W}_{UL}\right)\\ &= \left(\boldsymbol{W}^\star_{\overline{LL}_\gamma} + \boldsymbol{\Delta}^\star_{\overline{LL}_\gamma}\right)^{-1}\boldsymbol{W}_{LU}\boldsymbol{W}^-_{\overline{UU}_\gamma}\boldsymbol{W}_{UL}.\end{aligned} \tag{47}$$

The right of Equation (45) simplifies to $\boldsymbol{\Gamma}_{\overline{LL}_\gamma}$ based on Equations (46) and (47). ■

Theorem 21 *Let $\boldsymbol{W} \succeq 0$. Let f_γ denote the solution to the Joint Harmonic Optimization Problem (24) with $\gamma > 0$. Then*

$$f_\gamma = \begin{pmatrix} \boldsymbol{\Gamma}_{LL_\gamma} \\ -\left(\boldsymbol{\Delta}_{UU_\gamma}^-\right)^T \boldsymbol{\Delta}_{UL}\boldsymbol{\Gamma}_{LL_\gamma} + \left(\boldsymbol{I} - \left(\boldsymbol{\Delta}_{UU_\gamma}^-\right)^T \boldsymbol{\Delta}_{UU}\right)\boldsymbol{S}_{UL} \end{pmatrix} Y_L.$$

Proof Matrix definitions and techniques from the proof of Lemma 20 are used here. Let

$$\begin{aligned} \boldsymbol{R}_\gamma &= \left(\boldsymbol{\Delta}_{UU_\gamma}^-\right)^T \boldsymbol{W}_{UL}\left(\boldsymbol{W}_{LL_\gamma}^\star + \boldsymbol{\Delta}_{LL_\gamma}^\star\right)^{-1} \boldsymbol{W}_{LU}\boldsymbol{W}_{UU_\gamma}^- \boldsymbol{W}_{UL} \\ &= \left(\boldsymbol{\Delta}_{UU_\gamma}^-\right)^T \left\{\boldsymbol{W}_{UL}\left(\boldsymbol{I} - \boldsymbol{S}_{LU}\boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\right)^{-1}\boldsymbol{S}_{LU}\boldsymbol{C}_\gamma\right\}\left(\boldsymbol{I} - \boldsymbol{S}_{UU}\right)^T \boldsymbol{W}_{UL} \\ &= \left(\boldsymbol{\Delta}_{UU_\gamma}^-\right)^T \left\{\left(\boldsymbol{I} - \boldsymbol{W}_{UL}\boldsymbol{S}_{LU}\boldsymbol{C}_\gamma\right)^{-1}\boldsymbol{W}_{UL}\boldsymbol{S}_{LU}\boldsymbol{C}_\gamma\right\}\boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UL}. \end{aligned} \tag{48}$$

Then

$$\begin{aligned} \boldsymbol{S}_{UU}\boldsymbol{G}_\gamma &= \boldsymbol{S}_{UU}\left(\boldsymbol{I} - \boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\boldsymbol{S}_{LU}\right)^{-1}\boldsymbol{C}_\gamma\boldsymbol{\Delta}_{UL}\boldsymbol{S}_{LL} \\ &= \boldsymbol{S}_{UU}\boldsymbol{C}_\gamma\boldsymbol{\Delta}_{UL}\left(\boldsymbol{I} - \boldsymbol{S}_{LU}\boldsymbol{C}_\gamma\boldsymbol{W}_{UL}\right)^{-1}\boldsymbol{S}_{LL} \\ &= \left(\boldsymbol{\Delta}_{UU_\gamma}^-\right)^T \boldsymbol{\Delta}_{UL}\left(\boldsymbol{W}_{LL_\gamma}^\star + \boldsymbol{\Delta}_{LL_\gamma}^\star\right)^{-1}\boldsymbol{W}_{LL} \\ &= \left(\boldsymbol{\Delta}_{UU_\gamma}^-\right)^T \boldsymbol{\Delta}_{UL}\left(\boldsymbol{W}_{LL_\gamma}^\star + \boldsymbol{\Delta}_{LL_\gamma}^\star\right)^{-1}\boldsymbol{W}_{LL}^\star + \boldsymbol{R}_\gamma \\ &= \left(\boldsymbol{\Delta}_{UU_\gamma}^-\right)^T \boldsymbol{\Delta}_{UL}\boldsymbol{\Gamma}_{LL_\gamma} + \boldsymbol{R}_\gamma. \end{aligned} \tag{49}$$

Equation (48) and $\boldsymbol{S}_{UU}\boldsymbol{H}_\gamma = \left(\boldsymbol{\Delta}_{UU_\gamma}^-\right)^T \left\{\left(\boldsymbol{I} - \boldsymbol{W}_{UL}\boldsymbol{S}_{LU}\boldsymbol{C}_\gamma\right)^{-1}\right\}\boldsymbol{\Delta}_{UU}\boldsymbol{S}_{UL}$ imply

$$\boldsymbol{S}_{UL} - \left(\boldsymbol{S}_{UU}\boldsymbol{H}_\gamma + \boldsymbol{R}_\gamma\right) = \left(\boldsymbol{I} - \left(\boldsymbol{\Delta}_{UU_\gamma}^-\right)^T \{\boldsymbol{I}\}\,\boldsymbol{\Delta}_{UU}\right)\boldsymbol{S}_{UL}. \tag{50}$$

Proposition 12 and Equation (44) result in the unlabeled estimator smoother

$$\boldsymbol{S}_{UL} - \boldsymbol{S}_{UU}(\boldsymbol{G}_\gamma + \boldsymbol{H}_\gamma) = \boldsymbol{S}_{UL} - \left(\boldsymbol{S}_{UU}\boldsymbol{H}_\gamma + \boldsymbol{R}_\gamma\right) - (\boldsymbol{S}_{UU}\boldsymbol{G}_\gamma - \boldsymbol{R}_\gamma), \tag{51}$$

and substitutions based on Equations (49) and (50) into the right of Equation (51) produce its desired form. The labeled estimator smoother $\boldsymbol{\Gamma}_{LL_\gamma}$ is given by Lemma 20. ■

References

S Abney. *Semisupervised Learning for Computational Linguistics*. Chapman and Hall, CRC, 2008.

A Aswani, P Bickel, and C Tomlin. Regression on manifolds: estimation of the exterior derivative. *Annals of Statistics*, 39(1):48–81, 2010.

M Azizyan, A Singh, and L Wasserman. Density-sensitive semisupervised inference. *Annals of Statistics*, 41(2):751–771, 2013.

M Belkin, P Niyogi, and V Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

M Bredel and E Jacoby. Chemogenomics: An emerging strategy for rapid target and drug discovery. *Nature Reviews Genetics*, 5(4):262–275, April 2004.

M Carreira-Perpiñán and R Zemel. Proximity graphs for clustering and manifold learning. In *Advances in NIPs 18*, pages 225–232, 2005.

O Chapelle, M Chi, and A Zien. A continuation method for semi-supervised SVMs. In *International Conference on Machine Learning*, 2006a.

O Chapelle, B Schölkopf, and A Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006b. URL `http://www.kyb.tuebingen.mpg.de/ssl-book`.

C Cortes, M Mohri, D Pechyony, and A Rastogi. Stability of transductive regression algorithms. In *International Conference of Machine Learning*, 2008.

M Culp. On the semi-supervised joint trained elastic net. *Journal of Computational Graphics and Statistics*, 22(2):300–318, 2013.

M Culp, G Michailidis, and K Johnson. On multi-view learning with additive models. *Annals of Applied Statistics*, 3(1):545–571, 2009.

P Doyle and J Snell. Random walks and electrical networks. *Mathematical Association of America*, 1984.

A Frank and A Asuncion. UCI machine learning repository, 2010. URL `http://archive.ics.uci.edu/ml`.

T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning (Data Mining, Inference and Prediction)*. Springer Verlag, 2001.

M Hein, J Audibert, and U von Luxburg. From graphs to manifolds–weak and strong pointwise consistency of graph Laplacians. In *Conference on Learning Theory*, pages 470–485, 2005.

A Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Verlag, 2008.

T Jebara, J Wang, and S Chang. Graph construction and b-matching for semi-supervised learning. In *International Conference of Machine Learning*, 2009.

I Koprinska, J Poon, J Clark, and J Chan. Learning to classify e-mail. *Information Science*, 177 (10):2167–2187, 2007. ISSN 0020-0255.

M Kui, K Zhang, S Mehta, T Chen, and F Sun. Prediction of protein function using protein-protein interaction data. *Journal of Computational Biology*, 10:947–960, 2002.

J Lafferty and L Wasserman. Statistical analysis of semi-supervised regression. In *Advances in NIPS*, pages 801–808. MIT Press, 2007.

R Lundblad. *Chemical Reagents for Protein Modification*. CRC Press Inc., 2004. ISBN 08493-1983-8.

A McCallum, K Nigam, J Rennie, and K Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval Journal*, 3:127–163, 2000.

D Meyer, E Dimitriadou, K Hornik, A Weingessel, and F Leisch. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2012. URL `http://CRAN.R-project.org/package=e1071`. R package version 1.6-1.

B Nadler, N Srebro, and X Zhou. Statistical analysis of semi-supervised learning: The limit of infinite unlabelled data. In *Advances in NIPs 22*, pages 1330–1338. MIT Press, 2009.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2012.

P Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8:1369–1392, 2007.

A Singh, R Nowak, and X Zhu. Unlabeled data: Now it helps, now it doesn't. In *Advanced in NIPS*, pages 1513–1520, 2008.

A Subramanya and J Bilmes. Semi-supervised learning with measure propagation. *Journal of Machine Learning Research*, 12:3311–3370, 2011.

U von Luxburg, A Radl, and M Hein. Hitting times, commute distances and the spectral gap for large random geometric graphs. *Computing Research Repository*, abs/1003.1266, 2010.

J Wang and X Shen. Large margin semi-supervised learning. *Journal of Machine Learning Research*, 8:1867–1897, 2007.

Y Yamanishi, J Vert, and M Kanehisa. Protein network inference from multiple genomic data: A supervised approach. *Bioinformatics*, 20:363–370, 2004.

X Zhu. Semi-supervised learning literature survey. Technical report, Computer Sciences, University of Wisconsin-Madison, 2008.

X Zhu and A Goldberg. *Introduction to Semi-Supervised Learning*. Morgan and Claypool Publishers, 2009.

X Zhu, Z Ghahramani, and J Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *International Conference on Machine Learning*, pages 912–919, 2003.