Supplementary material for "Factor Analysis on a Graph"

A Proof for Theorem 1

We discuss the relation between the graph connectivity and our kernel $\widehat{\boldsymbol{\Sigma}}$, by using covariance decomposition of Jones and West (2005), which was originally proposed for analyzing paths on a graphical model. The (i, j) element of the covariance matrix can be decomposed as a weighted sum of products of conditional correlations of consecutive node pairs on all possible paths between i and j.

Theorem 4 (Jones and West (2005))**.** *Let $\mathcal{P}_{ij}$ be a set of paths between nodes i and j on the graph. A path $\mathcal{P} \in \mathcal{P}_{ij}$ is defined by a set of nodes ordered from i to j, i.e., $\mathcal{P} := \{(p_1, \ldots, p_m) | p_1 = i, p_m = j, m \le d\}$. We then have*

$$\boldsymbol{\Sigma}_{ij} = (-1)^{m+1} \boldsymbol{\Theta}_{p_1,p_2} \boldsymbol{\Theta}_{p_2,p_3} \ldots \boldsymbol{\Theta}_{p_{m-1},p_m} \frac{\det\left(\boldsymbol{\Theta}_{\setminus \mathcal{P}}\right)}{\det\left(\boldsymbol{\Theta}\right)} \tag{10}$$

According to the decomposition (10) and $\det(\boldsymbol{\Sigma}_{\mathcal{P}}) = \det\left(\boldsymbol{\Theta}_{\setminus \mathcal{P}}\right) / \det(\boldsymbol{\Theta})$, we obtain Theorem 1 in the main text.

B Optimality Condition of Factor Loading Matrix

The optimality condition of factor loading matrix $\boldsymbol{A}$ is as follows:

Lemma 1 (e.g., Jöreskog 1967)**.** *Assuming that we already have $\widehat{\boldsymbol{\Psi}}$, defined as the maximum likelihood estimate for $\boldsymbol{\Psi}$, then the maximum likelihood solution for $\boldsymbol{A}$ satisfies the following equation:*

$$(\widehat{\boldsymbol{\Psi}}^{-1/2} \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\Psi}}^{-1/2})(\widehat{\boldsymbol{\Psi}}^{-1/2} \boldsymbol{A}) = (\widehat{\boldsymbol{\Psi}}^{-1/2} \boldsymbol{A})(\boldsymbol{I} + \boldsymbol{A}^\top \widehat{\boldsymbol{\Psi}}^{-1} \boldsymbol{A}).$$

Suppose that $\boldsymbol{A}^\top \widehat{\boldsymbol{\Psi}}^{-1} \boldsymbol{A}$ is a diagonal matrix (This can be achieved by post-multiplying $\boldsymbol{A}$ by an orthogonal matrix, which does not affect the value of the likelihood), the equation can be regarded as an eigenvalue decomposition by which we obtain the estimator $\widehat{\boldsymbol{A}}$ for $\boldsymbol{A}$ as follows:

$$\widehat{\boldsymbol{A}} := \widehat{\boldsymbol{\Psi}}^{1/2} \boldsymbol{U}_k (\boldsymbol{\Lambda}_k - \boldsymbol{I})^{1/2}, \tag{11}$$

where $\boldsymbol{\Lambda}_k := \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ for the first k largest eigenvalues $\lambda_1, \ldots, \lambda_k$ of $\widehat{\boldsymbol{\Psi}}^{-1/2} \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\Psi}}^{-1/2}$, and $\boldsymbol{U}_k \in \mathbb{R}^{d \times k}$ is a set of the corresponding eigenvectors.

We now derive the first order condition of $\boldsymbol{A}$ in the above lemma. The derivative of the objective function of factor analysis in the main text in terms of $\boldsymbol{A}$ is

$$\begin{aligned} &2(\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Psi})^{-1} \boldsymbol{A} - (\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Psi})^{-1} \widehat{\boldsymbol{\Sigma}} (\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Psi})^{-1} \boldsymbol{A} \\ &= 2(\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Psi})^{-1} (\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Psi} - \widehat{\boldsymbol{\Sigma}}) (\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Psi})^{-1} \boldsymbol{A}. \end{aligned}$$

The first order condition is written as

$$(\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Psi})^{-1} (\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Psi} - \widehat{\boldsymbol{\Sigma}}) (\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Psi})^{-1} \boldsymbol{A} = \boldsymbol{0}.$$

Multiplying this equation by $(\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Psi})$ from the left, we obtain

$$(\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Psi} - \widehat{\boldsymbol{\Sigma}}) (\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Psi})^{-1} \boldsymbol{A} = \boldsymbol{0}.$$

Using Woodbury formula, we see

$$\begin{aligned} &(\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Psi})^{-1} \boldsymbol{A} \\ &= (\boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1} \boldsymbol{A} (\boldsymbol{I} + \boldsymbol{A}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{\Psi}^{-1}) \boldsymbol{A} \\ &= \boldsymbol{\Psi}^{-1} \boldsymbol{A} (\boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{A}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{A}) \\ &= \boldsymbol{\Psi}^{-1} \boldsymbol{A} (\boldsymbol{I} + \boldsymbol{A}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{A})^{-1} (\boldsymbol{I} + \boldsymbol{A}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{A} - \boldsymbol{A}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{A}) \\ &= \boldsymbol{\Psi}^{-1} \boldsymbol{A} (\boldsymbol{I} + \boldsymbol{A}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{A})^{-1}. \end{aligned}$$

By this transformation, the first order condition can be written as

$$\begin{aligned}(\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Psi} - \widehat{\boldsymbol{\Sigma}})\boldsymbol{\Psi}^{-1}\boldsymbol{A}(\boldsymbol{I} + \boldsymbol{A}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{A})^{-1} &= \boldsymbol{0}\\ (\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{\Psi} - \widehat{\boldsymbol{\Sigma}})\boldsymbol{\Psi}^{-1}\boldsymbol{A} &= \boldsymbol{0}\\ \boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{A} + \boldsymbol{A} - \widehat{\boldsymbol{\Sigma}}\boldsymbol{\Psi}^{-1}\boldsymbol{A} &= \boldsymbol{0}.\end{aligned}$$

To derive the second equation in the above, we multiplied through by $(\boldsymbol{I} + \boldsymbol{A}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{A})$ from the right. Arranging this equation, we obtain

$$\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Psi}^{-1}\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{I} + \boldsymbol{A}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{A}^\top).$$

Multiplying this equation by $\boldsymbol{\Psi}^{-1/2}$ from the left, we finally see

$$(\boldsymbol{\Psi}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Psi}^{-1/2})(\boldsymbol{\Psi}^{-1/2}\boldsymbol{A}) = (\boldsymbol{\Psi}^{-1/2}\boldsymbol{A})(\boldsymbol{I} + \boldsymbol{A}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{A}^\top).$$

C Spectral Relaxation of Weighted Kernel k-means

Then the objective function of weighted kernel k-means is defined by

$$\sum_{i=1}^{k}\sum_{j\in\mathcal{C}_i}\widehat{\psi}_j^{-1}\left\|\boldsymbol{\phi}_j - \boldsymbol{\mu}_i\right\|_2^2, \tag{12}$$

where $\mathcal{C}_i$ for $i = 1, \ldots, k$ is an index set of the i-th cluster, and $\boldsymbol{\mu}_i$ is a centroid of the i-th cluster. In this objective function (12), the squared error between each $\boldsymbol{\phi}_i$ and its centroid is weighted by ψ_i^{-1}, which means that if the i-th dimension of the factor analysis error term $\boldsymbol{\epsilon}$ has a smaller variance, a corresponding $\boldsymbol{\phi}_i$ is penalized more strongly. Using an indicator matrix $\boldsymbol{Z}$, in which the (i, j) element takes 1 if the i-th instance belongs to the j-th cluster or takes 0 otherwise, this function can be re-written as:

$$\operatorname{trace}\left\{(\boldsymbol{\Phi} - \boldsymbol{Z}\boldsymbol{M}^\top)^\top\widehat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\Phi} - \boldsymbol{Z}\boldsymbol{M}^\top)\right\}, \tag{13}$$

where $\boldsymbol{M} := [\ \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\ ]$.

We consider a spectral relaxation of this weighted kernel k-means. Given a cluster assignment $\mathcal{C}_i$, the centroid which minimizes the squared error is the weighted average of the instances: $\sum_{j\in\mathcal{C}_i}\widehat{\psi}_j^{-1}\boldsymbol{\phi}_j / \sum_{j\in\mathcal{C}_i}\widehat{\psi}_j^{-1}$. Then, the set of centroids can be written as

$$\boldsymbol{M} = \boldsymbol{\Phi}^\top\widehat{\boldsymbol{\Psi}}^{-1}\boldsymbol{Z}\boldsymbol{C},$$

where $\boldsymbol{C} = \operatorname{diag}(1/\sum_{j\in\mathcal{C}_1}\widehat{\psi}_j^{-1}, \ldots, 1/\sum_{j\in\mathcal{C}_k}\widehat{\psi}_j^{-1})$. Substituting this into (13), the objective function can be transformed into

$$\begin{aligned}&\operatorname{trace}\left\{(\boldsymbol{\Phi} - \boldsymbol{Z}\boldsymbol{M}^\top)^\top\widehat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\Phi} - \boldsymbol{Z}\boldsymbol{M}^\top)\right\}\\ &= \operatorname{trace}\left(\boldsymbol{\Phi}^\top\widehat{\boldsymbol{\Psi}}^{-1}\boldsymbol{\Phi} - \boldsymbol{\Phi}^\top\widehat{\boldsymbol{\Psi}}^{-1}\boldsymbol{Z}\boldsymbol{C}\boldsymbol{Z}^\top\widehat{\boldsymbol{\Psi}}^{-1}\boldsymbol{\Phi}\right).\\ &= \operatorname{trace}\left(\boldsymbol{\Phi}^\top\widehat{\boldsymbol{\Psi}}^{-1}\boldsymbol{\Phi} - \boldsymbol{C}^{1/2}\boldsymbol{Z}^\top\widehat{\boldsymbol{\Psi}}^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\widehat{\boldsymbol{\Psi}}^{-1}\boldsymbol{Z}\boldsymbol{C}^{1/2}\right).\end{aligned}$$

Here, we used $\boldsymbol{Z}^\top\widehat{\boldsymbol{\Psi}}^{-1}\boldsymbol{Z} = \boldsymbol{C}^{-1}$, and the first term is now constant. Defining $\boldsymbol{V}_k := \widehat{\boldsymbol{\Psi}}^{-1/2}\boldsymbol{Z}\boldsymbol{C}^{1/2}$, which leads $\boldsymbol{V}_k^\top\boldsymbol{V}_k = \boldsymbol{C}^{1/2}\boldsymbol{Z}\widehat{\boldsymbol{\Psi}}^{-1}\boldsymbol{Z}\boldsymbol{C}^{1/2} = \boldsymbol{C}^{1/2}\boldsymbol{C}^{-1}\boldsymbol{C}^{1/2} = \boldsymbol{I}$, the following spectral relaxation of the weighted kernel k-means can be derived:

$$\begin{aligned}\max_{\boldsymbol{V}_k\in\mathbb{R}^{d\times k}} \quad & \operatorname{trace}\left(\boldsymbol{V}_k^\top\widehat{\boldsymbol{\Psi}}^{-1/2}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Psi}}^{-1/2}\boldsymbol{V}_k\right)\\ \text{s.t.} \quad & \boldsymbol{V}_k^\top\boldsymbol{V}_k = \boldsymbol{I}.\end{aligned} \tag{14}$$

D Proof for Theorem 2

Theorem 2 can be derived from the optimality condition for the factor loading matrix $\boldsymbol{A}$ written in supplementary appendix B.

Since the set of eigenvectors corresponding to the k largest eigenvalues of $\widehat{\boldsymbol{\Psi}}^{-1/2}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Psi}}^{-1/2}$ is an optimal solution to (14), we see $\boldsymbol{V}_k = \boldsymbol{U}_k$. We therefore obtain the relation $\widehat{\boldsymbol{Z}} = \widehat{\boldsymbol{A}}\boldsymbol{D}$, where $\boldsymbol{D} := \boldsymbol{C}^{-1/2}(\boldsymbol{\Lambda}_k - \boldsymbol{I})^{-1/2}$, which is a diagonal matrix.

E Proof for Theorem 3

Replacing $\boldsymbol{V}_k$ in $\widehat{\boldsymbol{Z}}$ (5) (written the main text) by $\boldsymbol{V}_k\boldsymbol{Q}$ keeps the objective of kernel k-means (14) optimal, and we obtain $\widehat{\boldsymbol{Z}} = \widehat{\boldsymbol{\Psi}}^{1/2}\boldsymbol{V}_k\boldsymbol{Q}\boldsymbol{C}^{-1/2}$. Then, we see $\widehat{\boldsymbol{Z}}\boldsymbol{C}^{1/2} = \widehat{\boldsymbol{A}}(\boldsymbol{\Lambda}_k - \boldsymbol{I})^{-1/2}\boldsymbol{Q}$ (Note that $\boldsymbol{C}$ is diagonal). The invariance of the likelihood can be easily seen by $\widehat{\boldsymbol{A}}_{\mathrm{rot}}(\boldsymbol{Q}^\top(\boldsymbol{\Lambda}_k - \boldsymbol{I})\boldsymbol{Q})\widehat{\boldsymbol{A}}_{\mathrm{rot}}^\top = \widehat{\boldsymbol{A}}\widehat{\boldsymbol{A}}^\top$.

F Formulation of Lap-PCA

Let $\boldsymbol{W}$ be an adjacency matrix of the graph $\mathcal{G}$ in which the (i,j) element is $W_{ij} = 1$ if $(i,j) \in \mathcal{E}$, and $W_{ij} = 0$ otherwise.

For factor analysis and PCA, In our case, the graph structure can be incorporated into the matrix $\boldsymbol{A}$ by the following formulation: in factor analysis or PCA by

$$\min_{\boldsymbol{A}\in\mathbb{R}^{d\times k}, \boldsymbol{\Psi}\in\mathcal{D}_+^d} \quad (1-\alpha)\ \ell(\boldsymbol{A}, \boldsymbol{\Psi}) + \alpha \sum_{k'=1}^{k} \sum_{(i,j)\in\mathcal{E}} W_{ij}(A_{ik'} - A_{jk'})^2, \tag{15}$$

where ℓ is a loss function (negative log-likelihood), $\boldsymbol{L} \in \mathbb{R}^{d\times d}$ is the graph Laplacian matrix (see, e.g., Chung, 1997, for detail), and $\alpha \in [0,1]$ is a regularization parameter. For PCA, $\boldsymbol{\Psi}$ has an additional constraint $\boldsymbol{\Psi} = \sigma^2\boldsymbol{I}$.

In experiments, we chose the best regularization parameter α in (15) out of $\{0.25, 0.5, 0.75\}$ in terms of each result.

References

M. Ashburner. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

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.

D. M. Boyd and N. B. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007.

E. G. Cerami and et al. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39:685–690, 2011.

F. R. K. Chung. *Spectral Graph Theory.* American Mathematical Society, 1997.

R. Cohen and S. Havlin. *Complex networks: Structure, Robustness and Function.* Cambridge University Press, 2010.

I. S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means: Spectral clustering and normalized cuts. In *Proc. of the 10th ACM SIGKDD*, pages 551–556. ACM, 2004.

A. B. Goldberg, X. Zhu, and S. J. Wright. Dissimilarity in graph-based semi-supervised classification. In M. Meila and X. Shen, editors, *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, volume 2, pages 155–162. JMLR.org, 2007.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in NIPS 19*, pages 513–520. MIT Press, 2007.

H. H. Harman. *Modern Factor Analysis*. The university of chicago press, 1960.

T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(suppl 1):S233–S240, 2002.

B. Jiang, C. Ding, B. Luo, and J. Tang. Graph-Laplacian PCA: Closed-form solution and robustness. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3492–3498, 2013.

I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 2002.

B. Jones and M. West. Covariance decomposition in undirected gaussian graphical models. *Biometrika*, 92(4): 779–786, 2005.

K. G. Jöreskog. Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32(4):443–482, 1967.

H. F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1): 27–30, 2000.

B. Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *PHYSICAL REVIEW E*, 83:016107, Jan 2011.

R. Kosala and H. Blockeel. Web mining research: A survey. *SIGKDD Explorations Newsletter*, 2(1):1–15, June 2000.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1): 79–86, 1951.

C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

M. Meila and J. Shi. A random walks view of spectral segmentation. In *Proceedings of the 8th International Workshop on Artifical Intelligence and Statistics*. Morgan Kaufmann, 2001.

M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in NIPS 14*, pages 849–856. MIT Press, 2001.

J. S.-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge university press, 2004.

M. Saerens, F. Fouss, L. Yen, and P. Dupont. The principal components analysis of a graph, and its relationships to spectral clustering. In *Proc. of the 15th ECML*, volume 3201 of *LNCS*, pages 371–383. Springer Berlin Heidelberg, 2004.

G. Sales, E. Calura, D. Cavalieri, and C. Romualdi. graphite - a bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*, 13(1):20, 2012.

T. Sandler, J. Blitzer, P. P. Talukdar, and L. H. Ungar. Regularized learning with networks of features. In *Advances in NIPS 21*, pages 1401–1408, 2008.

S. E. Schaeffer. Survey: Graph clustering. *Computer Science Review*, 1(1):27–64, Aug. 2007.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

A. J. Smola and I. R. Kondor. Kernels and regularization on graphs. In *Proceedings of the Annual Conference on Computational Learning Theory*, pages 144–158, 2003.

R. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011.

M. J. van de Vijver and et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley Publishing, 1990.

L. Wu, X. Ying, X. Wu, A. Lu, and Z.-H. Zhou. Spectral analysis of k-balanced signed graphs. In *Advances in Knowledge Discovery and Data Mining - 15th Pacific-Asia Conference (PAKDD)*, volume 6635 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2011.

S. Yang, L. Yuan, Y.-C. Lai, X. Shen, P. Wonka, and J. Ye. Feature grouping and selection over an undirected graph. In *Proc. of the 18th ACM SIGKDD*, pages 922–930. ACM, 2012.

H. Zha, X. He, C. Ding, H. Simon, and M. Gu. Spectral relaxation for k-means clustering. In *Advances in NIPS 14*, pages 1057–1064. MIT Press, 2001.

M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, and D. Cai. Graph regularized sparse coding for image representation. *Image Processing, IEEE Transactions on*, 20(5):1327–1336, 2011.

D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems (NIPS) 16*. MIT Press, 2004.

X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In T. Fawcett and N. Mishra, editors, *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 912–919. AAAI Press, 2003.