ICM 2018
RIO DE JANEIRO

Proc. Int. Cong. of Math. – 2018
Rio de Janeiro, Vol. 4 (3883–3910)

GRAPHICAL MODELS IN MACHINE LEARNING, NETWORKS AND UNCERTAINTY QUANTIFICATION

Andrea L. Bertozzi

Abstract

This paper is a review article on semi-supervised and unsupervised graph models for classification using similarity graphs and for community detection in networks. The paper reviews graph-based variational models built on graph cut metrics. The equivalence between the graph mincut problem and total variation minimization on the graph for an assignment function allows one to cast graph-cut variational problems in the language of total variation minimization, thus creating a parallel between low dimensional data science problems in Euclidean space (e.g. image segmentation) and high dimensional clustering. The connection paves the way for new algorithms for data science that have a similar structure to well-known computational methods for nonlinear partial differential equations. This paper focuses on a class of methods build around diffuse interface models (e.g. the Ginzburg–Landau functional and the Allen–Cahn equation) and threshold dynamics, developed by the Author and collaborators. Semi-supervised learning with a small amount of training data can be carried out in this framework with diverse applications ranging from hyperspectral pixel classification to identifying activity in police body worn video. It can also be extended to the context of uncertainty quantification with Gaussian noise models. The problem of community detection in networks also has a graph-cut structure and algorithms are presented for the use of threshold dynamics for modularity optimization. With efficient methods, this allows for the use of network modularity for unsupervised machine learning problems with unknown number of classes.

1 Similarity Graphs and Spectral Clustering

Graphical models provide a mathematical structure for high dimensional data problems that yield important latent information in the data. They are also the basic building block

This work was supported by NSF grants DMS-1737770, DMS-1417674, NIJ grant 2014-R2-CX-0101, and ONR grant N00014-16-1-2119.

MSC2010: primary 65K10; secondary 35Q56, 49M20, 6209, 91C20, 91D30.

Keywords: diffuse interfaces, graphical models, graph Laplacian, machine learning, uncertainty quantification, social networks, community detection, data clustering, modularity, MBO scheme.

for network analysis. Graphical models yield useful information about connections between pieces of data from pairwise comparisons of that data, most notably via a *similarity graph* in which nodes represent pieces of data and edge weights are related to pairwise comparisons of the data. For machine learning methods, a major challenge is the inherent $O(N^2)$ computational complexity of the weights (for N nodes) unless the graph is sparse. Another source of complexity is the number of classes. Furthermore, most machine learning methods, including those developed for complex graphical models, are based on linear algebra and linear models. Graph-based structures have the potential to provide a useful framework that is inherently nonlinear providing a broader framework for the data structures. For classification in machine learning there are basic methods like Support Vector Machine, which identifies a hyperplane separating different classes of data. This is a **supervised** algorithm involving a lot of training data with small amounts of unknown data. Kernel methods allow for unknown nonlinear mappings to be computed as part of the methodology. Still this restricts that data to have a certain form and for the mapping to be learned or computed.

In contrast, a similarity graph allows analysis of the data by performing operations on the graph itself, thus removing the original high- dimensionality of the problem. Linear structures have been studied, most notably the graph Laplacian matrix of the form $L = D - W$ where W is the weight matrix of off-diagonal elements w_{ij} and the diagonal matrix D has each entry d_i equal to the sum of the weights connected to node i. Spectral clustering is an **unsupervised** method in which clusters are determined by a k-means method applied to a small set of eigenfunctions of the graph Laplacian matrix von Luxburg [2007]. Spectral clustering can be paired with a random sampling method using the Nyström extension, that allows for an approximately $O(N)$ low-rank approximation of the graph Laplacian matrix. Spectral clustering in machine learning requires the graph to be constructed from data. Similarity graphs are well-known in machine learning and have each node corresponding to a feature vector V_i comprised of high-dimensional data to be classified, and the weights w_{ij} between nodes are computed as a pairwise comparison between the feature vectors. Some examples include:

1. The Gaussian function

$$w_{i,j} = \exp(-||V_i - V_j||^2/\tau) \tag{1}$$

Depending on the choice of metric, this similarity function includes the Yaroslavsky filter Yaroslavsky [1985] and the nonlocal means filter Buades, Coll, and Morel [2005].

2. Gaussian with cosine angle

$$w_{i,j} = \exp - \frac{(1 - \frac{<V_i, V_j>}{|V_i||V_j|})^2}{2\sigma^2} \tag{2}$$

is a common similarity function used in hyperspectral imaging. In this case one is interested in alignment of feature vectors rather than their Euclidean distance.

3. Zelnik-Manor and Perona introduced local scaling weights for sparse matrix computations Zelnik-Manor and Perona [2004]. Given a metric $d(V_i, V_j)$ between each feature vector, they define a local parameter $\sqrt{\tau(V_i)}$ for each V_i. The choice in Zelnik-Manor and Perona [ibid.] is $\sqrt{\tau(V_i)} = d(V_i, V_M)$, where V_M is the Mth closest vector to V_i. The similarity matrix is then defined as

$$w_{i,j} = \exp\left(-\frac{d(V_i, V_j)^2}{\sqrt{\tau(V_i)\tau(V_j)}}\right). \tag{3}$$

This similarity matrix is better at segmentation when there are multiple scales that need to be segmented simultaneously.

There are two popular normalization procedures for the graph Laplacian, and the normalization has segmentation consequences F. R. K. Chung [1996] and von Luxburg [2007]. The normalization that is often used for the nonlocal means graph for images is the symmetric Laplacian L_s defined as

$$L_s = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}. \tag{4}$$

The symmetric Laplacian is named as such since it is a symmetric matrix. The random walk Laplacian is another important normalization given by

$$L_w = D^{-1} L = I - D^{-1} W. \tag{5}$$

The random walk Laplacian is closely related to discrete Markov processes.

A novel example of spectral clustering applied to social science data is presented in van Gennip, Hunter, et al. [2013]. LAPD Field Interview (FI) cards provide a unique opportunity to investigate the relationships between individual use of space, social networks and group identities, specifically criminal street gang affiliation. FI cards are completed when a patrol officer comes into contact with a member of the public. They record spatio-temporal data about where and when the stop occurred, individual characteristics (e.g., name and home address) and demographic characteristics (e.g., age, sex, ethnic group). FI cards also record information about criminal activity and gang affiliation, if applicable. Critical here is information on gang membership. Known or suspected members of gangs have their gang affiliation recorded, gang moniker if known, and information on the duration of gang membership (e.g., member since 2004). FI cards also record instances where two or more gang members were stopped and interviewed together. Thus, each FI with two or more gang members represents a spatial sample of occasions when nodes in a social

network interacted. We developed a graphical model using both social network information from raw observations and spatial coordinates of these observations. Figure 1 shows results of spectral clustering using the composite graph with both information - the result finds latent groups of individuals that differ from the known gang affiliations as illustrated in the Pie chart. The work in Figure 1 used standard spectral clustering methods to identify

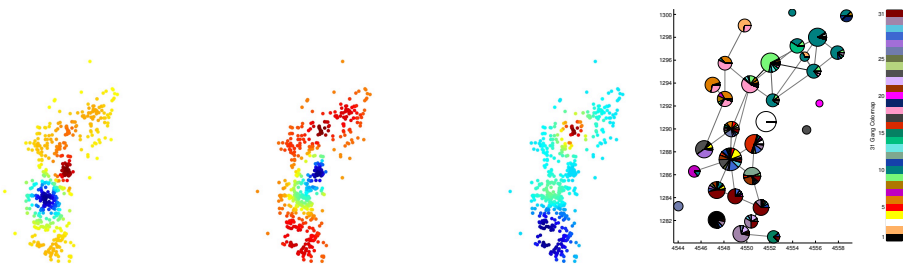


Figure 1: Spectral clustering applied to LAPD Hollenbeck Division Field Interview card data for 2009 van Gennip, Hunter, et al. [2013]. Left eigenvalues associated with event data shown with geographic placement. (right) Pie charts showing clusters identified by spectral clustering compared with known ground truth of gang affiliation (shown in colors) for the 31 gangs in Hollenbeck.

latent groups in the geo-social set space. The results show that natural groupings Hollenbeck sometimes are comprised of mostly one gang but othertimes, especially in areas with different gangs in high spatial proximity, can have members of multiple gangs affiliated with the same observed group.

2 Data classification and the Ginzburg-Landau functional on graphs

The Author and Arjuna Flenner developed the first suite of binary classifiers for semi-supervised machine learning (minimal training data) using a fully nonlinear model for similarity graphs Bertozzi and Flenner [2012]. We proposed the Ginzburg-Landau (GL) functional as a smooth relaxation of graph total variation (equivalent to graph cuts), as a regularizer for semi- supervised learning. For large datasets, we incorporate efficient linear algorithms into a nonlinear PDE-based method for non-convex optimization. Our work has been republished as a SIGEST paper Bertozzi and Flenner [2016]. Over 50 new

papers and methods have arises from this work including fast methods for nonlocal means image processing using the MBO scheme Merkurjev, Kostic, and Bertozzi [2013], multiclass learning methods Garcia-Cardona, Merkurjev, Bertozzi, Flenner, and Percus [2014] and Iyer, Chanussot, and Bertozzi [2017], parallel methods for exascale-ready platforms Meng, Koniges, He, S. Williams, Kurth, Cook, Deslippe, and Bertozzi [2016], hyperspectral video analysis Hu, Sunu, and Bertozzi [2015], Merkurjev, Sunu, and Bertozzi [2014], Meng, Merkurjev, Koniges, and Bertozzi [2017], and W. Zhu, Chayes, Tiard, S. Sanchez, Dahlberg, Bertozzi, Osher, Zosso, and Kuang [2017], modularity optimization for network analysis Hu, Laurent, Porter, and Bertozzi [2013] and Boyd, Bai, X. C. Tai, and Bertozzi [2017], measurement techniques in Zoology Calatroni, van Gennip, Schönlieb, Rowland, and Flenner [2017], generalizations to hypergraphs Bosch, Klamt, and Stoll [2016], Pagerank Merkurjev, Bertozzi, and F. Chung [2016] and Cheeger cut based methods Merkurjev, Bertozzi, Yan, and Lerman [2017]. This paper reviews some of this literature and discusses future problem areas including crossover work between network modularity and machine learning and efforts in uncertainty quantification.

Given a phase field variable u, the Ginzburg-Landau energy, introduced for Euclidean space in the last century, involves a competition between the convex functional $\int (\nabla u)^2 dx$ that induces smoothing, with a double well function $\int W(u)dx$, that separates its argument into phases. The Bertozzi-Flenner graph model replaces the first term with the graph Dirichlet energy, $\sum_{ij} w_{ij}(u_i - u_j)^2$, equivalent to the inner product of Lu with u where L is the graph Laplacian:

$$E_{GL}(f) = \frac{1}{\epsilon}\langle Lf, f\rangle + \epsilon \sum_i (W(f_i)). \tag{6}$$

For a variant of the GL functional in Equation (6) one can prove Gamma convergence of the vanishing ϵ limit to the graph TV functional van Gennip and Bertozzi [2012], equivalent to the graph cut energy:

$$E_{TV}(u) = \sum_{ij} w_{ij}|u_i - u_j|, \tag{7}$$

for u defining a graph partition. More recent work extending these results is Thorpe and Theil [2017]. An equivalent result has been known for Euclidean space for several decades Kohn and Sternberg [1989]. Another variant involves a wavelet GL functional Dobrosotskaya and Bertozzi [2008] which has a Wulff shape energy as its sharp interface Gamma-limit Dobrosotskaya and Bertozzi [2010]. The GL functional is useful, in lieu of L1 compressed sensing methods for minimizing total variation, because the computationally expensive graph information only arises in the Dirichlet energy, leveraging optimization algorithms that can exploit efficient approximations of the graph Laplacian.

The nonlinear structure can be reduced to simple calculations such as local thresholding, as shown in the MBO scheme below. The graph cut functional or equivalent TV functional can be incorporated into a semi-supervised or unsupervised learning problem. Without additional terms in the energy, the minimizer of the energy is trivial - simply pick u to be a constant, one of the minimizers of the well W. However nontrivial solutions can be found by modifying the energy to include a semi-supervised penalty term or additional balance terms in the case of unsupervised learning problems. For semi-supervised learning we consider an L^2 penalty for known training data (defined to be set S and with values u_0 along with a graph cut term to minimize the sum of the weights between unlike classes:

$$E_1(u) = |u|_{TV} + \sum_{i \in S} \frac{\lambda}{2}(u_0(i) - u(i))^2 \approx E_{GL}(u) + \sum_{i \in S} \frac{\lambda}{2}(u_0(i) - u(i))^2.$$

The second term is for semi-supervision and the first is for the graph cut. The parameter λ provides a soft constraint for semi-supervision. In many applications discussed below the supervision involves a small amount of training data, e.g. 10% or less, compared to the majority of the data for supervised learning such as SVM.

The semi-supervised learning problem described above can be minimized quickly on very large datasets using a pseudo-spectral method involving the eigenfunctions and eigenvalues of the graph Laplacian and convex splitting methods Schönlieb and Bertozzi [2011] from nonlinear PDE. The important eigenfunctions can be computed very quickly for large datasets using sub-sampling methods, e.g. the Nyström extension Belongie, Fowlkes, F. Chung, and Malik [2002], Fowlkes, Belongie, F. Chung, and Malik [2004], and Fowlkes, Belongie, and Malik [2001]. What is remarkable is that the entire TV minimization problem can be solved without computing all the weights of the graph (which can be prohibitive in the case of e.g. nonlocal means used in image processing with textures) Buades, Coll, and Morel [2005], Gilboa and Osher [2007, 2008], and Merkurjev, Sunu, and Bertozzi [2014]. While there are other fast algorithms out there for TV minimization (e.g. the split Bregman method Goldstein and Osher [2009]) none of them can easily be adapted to use the fast algorithms for eigenfunctions that rely on having a symmetric matrix. Indeed the algorithms presented in this paper only require the knowledge of the important eigenfunctions of the graph Laplacian and do not require the computation of a "right hand side" that arises in more general TV minimization algorithms, such as split Bregman.

An unsupervised learning model can be constructed as a generalization of the piecewise constant Mumford-Shah model from image segmentation, applied to a graphical data model. We recall the the piecewise constant Mumford-Shah model T. Chan and Vese [2001], Vese and T. F. Chan [2002], and Esedoğlu and Tsai [2006] involves the identification of a contour Φ that divides the image up into $\hat{n}$ regions Ω_r. The energy to minimize

is

$$E(\Phi, \{c_r\}_{r=1}^{\hat{n}}) = |\Phi| + \lambda \sum_{r=1}^{\hat{n}} \int_{\Omega_r} (u_0 - c_r)^2$$

where u_0 is the observed image data, c_r denotes a constant approximation of the image in the set Ω_r and $|\Phi|$ denotes the length of the contour Φ. The works Hu, Sunu, and Bertozzi [2015] and Meng, Merkurjev, Koniges, and Bertozzi [2017] present a generalization of this to graphical models. The examples studied are largely hyperspectral imagery however the idea could be applied to other high dimensional vectors. The data is used both to create the similarity graph and to solve the clustering problem because the constants will be chosen in the high dimensional space to approximate the high dimensional vectors within each class. This is different from the previous example in which the clustering can be computed outside of the high dimensional data space once the graph is known (or approximated) and the training data is known. More specifically, we consider the energy

$$E_2 = \frac{1}{2}|f|_{TV} + \lambda \sum_{r=1}^{\hat{n}} \sum_{i} f_r(n_i)\|u_0(n_i) - c_r\|^2,$$

where f is a simplex constrained vector value that indicates class assignment:

$$f : G \to \{0,1\}^{\hat{n}}, \sum_{r=1}^{\hat{n}} f_r(n_i) = 1\}.$$

Specifically if $f_r(n_i) = 1$ for some r then the data at node n_i belongs to the $r-th$ class. For each f we have a partition of the graph into at most $\hat{n}$ classes. The connection to the original piecewise constant Mumford-Shah model is that f_r is the characteristic function of the rth class and thus $\sum_i f_r(n_i)\|u_0(n_i) - c_r\|^2$ is analogous to the term $\int_{\Omega_r}(u_0 - c_r)^2$ while the TV norm on graphs is the analogue of the length of the boundary in the Euclidean space problem.

2.1 The MBO scheme on Graphs. Rather than minimizing the GL functional, using an efficient convex splitting method such as in Bertozzi and Flenner [2016], we can use an even more efficient MBO method. Using the original Euclidean GL functional and classical PDE methods, Esedoğlu and Tsai [2006] developed a simple algorithm for piecewise-constant image segmentation that alternated between evolution of the heat equation and thresholding. That paper built on even earlier work by Merriman, Bence, and Osher [1992] (MBO) for motion by mean curvature. Motivated by this work, the MBO computational scheme was extended to the graphical setting by Merkurjev, Kostic, and Bertozzi [2013] for binary classification and methods that build on binary classification such as bit-wise

greyscale classification for inpainting of greyscale images. The Graph MBO scheme for semi-supervised learning consists of the following two steps:

1. Heat equation with forcing term. Propagate using

$$\frac{u^{(n+1/2)} - u^{(n)}}{dt} = -L_u^{(n)} - \lambda(i)(u^{(n)} - u_0)$$

2. Threshold.

$$u^{(n+1)} = \begin{cases} 1 & \text{if } u^{(n+1/2)} \geq 0 \\ 0 & \text{if } u^{(n+1/2)} < 0. \end{cases}$$

The results in Merkurjev, Kostic, and Bertozzi [2013] showed significant speed-up in run time compared to the Ginzburg-Landau method developed here and also faster run times than the split-Bregman method applied to the Osher-Gilboa nonlocal means graph for the same datasets. Both the GL and MBO methods for binary learning were extended to the multiclass case in Garcia-Cardona, Merkurjev, Bertozzi, Flenner, and Percus [2014]. The MBO scheme in particular is trivial to extend - the algorithm is the same except that the classes are defined taking the range of u in $\hat{n}$ dimensions where $\hat{n}$ is the number of classes and thresholding to the corners of the simplex. The MBO scheme is quite fast and in most cases finds the global minimum. For unusual problems that require a provably optimal solution, we have considered methods built around max flow and ADMM methods that are less efficient than MBO, but they can guarantee a global optimal solution for the binary, semi-supervised segmentation problem Merkurjev, Bae, Bertozzi, and X.-C. Tai [2015].

As an example of a high dimensional problem with multiple classes, consider the classification of hyperspectral pixels in a video sequence. Figure 2 shows data from standoff detection of a glass plume using 128 spectra in the Long Wave Infrared (LWIR) from the Dugway Proving Ground. The graph weights are computed with spectral angle. The Nyström extension provides eigenfunctions of the graph Laplacian, which can run in Matlab in 2 minutes on a modest laptop. The actual classification runs in seconds. The Nyström method and the MBO scheme have recently been optimized on an exascale-ready platform at the National Energy Research Supercomputing Center (NERSC) Meng, Koniges, He, S. Williams, Kurth, Cook, Deslippe, and Bertozzi [2016].

Inspired by the work in Esedoğlu and Otto [2015], one can translate the MBO scheme into a discrete time approximate graph cut minimization method. In Hu, Sunu, and Bertozzi [2015] and van Gennip, Guillen, Osting, and Bertozzi [2014] it is shown that the diffusion operator $\Gamma_\tau = e^{-\tau L}$ where L is the graph Laplacian defined above and τ is the timestep of the MBO scheme, then the discrete energy

$$E_{MBO}(u) = \frac{1}{\tau} < 1 - u, \Gamma_\tau u >$$

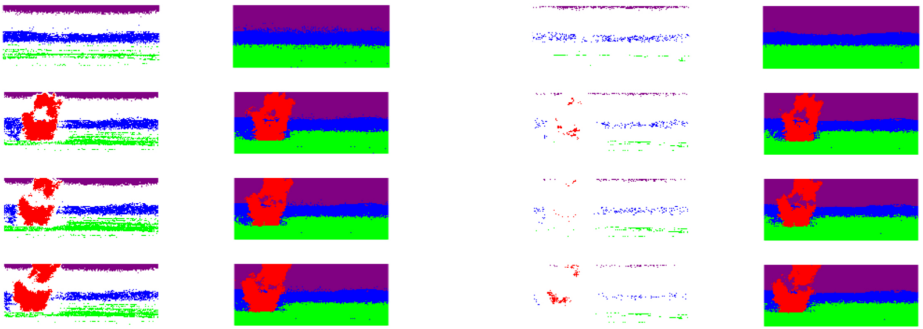


Figure 2: Feature vectors V_i are 128 dimensional hyperspectral pixels taken from a video sequence from standoff detection of gas plume in the Dugway Proving Ground. Shown are 4 of 7 video frames with $N = 280,000$ graph nodes. The colors correspond to four classes: plume (red), sky (purple), foreground (green), mounain (blue). (left) 36% training data and the resulting classification - as in Merkurjev, Sunu, and Bertozzi [2014]. (right) The same calculation with only 4% training data. In each case the training data is shown to the left of the fully segmented video. Code for this calculation is published online in Meng, Merkurjev, Koniges, and Bertozzi [2017]

decreases on each timestep and also approximates the graph TV energy.

2.2 Volume Penalties. In the case of unsupervised classification it is often desireable to have volume constraints. So rather than just minimizing the size of the graph cut, i.e. $|u|_{TV}$, one can put in a penalty that forces the size of the classes to be reasonably distributed. Two such normalizations are the ratio cut and the normalized cut, the problem is to find a subset S of the graph to minimize $cut(S,\bar{S})R(S)$ where R is $(1/|S|+1/|\bar{S}|)$ for the ratio cut and $(1/vol(S)+1/vol(\bar{S}))$ for the normalized cut. Here the volume of the graph is the sum of the degrees of the all the vertices and the degree of the node is the sum of the weights connected to that node. Another normalization is the Cheeger cut in which $R = (min(|S|,|\bar{S}|)^{-1}$. All three of these functionals are linear in the graph cut term and nonlinear in the volumetric constraints. The energy blows up as the size of S or $\bar{S}$ goes to zero, thus ensuring a balance cut. There are several important papers related to clustering with volume penalties by Bresson, Szlam, Laurent, von Brecht. These works use other methods than the ones described above. In Szlam and Bresson [2010], study a relaxation of the Cheeger cut problem with connections between the energy of the relaxed problem and well studied energies in image processing. Authors of Bresson, Laurent, Uminsky, and von Brecht [2012] detail two procedures for the relaxed Cheeger cut problem. The

first algorithm is a steepest descent approach, and the second one is a modified inverse power method. In Bresson, Laurent, Uminsky, and von Brecht [2013], develop another version of the method shown in Bresson, Laurent, Uminsky, and von Brecht [2012] using a new adaptive stopping condition. The result is an algorithm that is monotonic and more efficient. The GL functional on graphs has been extended to these product-form volume penalized problems. The paper Merkurjev, Bertozzi, Yan, and Lerman [2017] uses a diffuse interface approach along with the graph Laplacian to solve the fully nonlinear ratio cut problem and the Cheeger cut problem. The results are shown to be very efficient with the efficiency partly achieved through the use of the Nyström extension method. The main idea is to approximate the cut term using the GL functional and then use PDE-based methods for gradient descent in a spectral approach. Jacobs, Merkurjev, and Esedoğlu [2018] have a very efficient MBO-based method for solving the volume-constrained classification problem with different phases and with prescribed volume constraints and volume inequalities for the different phases. This work combines some of the best features of the MBO scheme in both Euclidean space and on graphs with a highly efficient algorithm of Bertsekas [1979] for the auction problem with volume constraints.

One of the challenges in machine learning is the case where the sizes of the classes are unknown. Volume constraints could perchance become incorporated as building blocks for solutions to complex data sorting problems, where the amount of data is so large that it becomes physically impossible for a human to verify all the results by inspection. An example of such large data currently under collection by law enforcement agencies around the world are video feeds from body worn cameras. The author and collaborators have been working with such a dataset provided by the Los Angeles Police Department and have developed classification methods based on the MBO scheme. The BW camera poses unusual challenges - typically the goal is to identify what is going on in the scene, both in terms of the wearer of the camera and his or her interaction with the scene. Thus the task requires understanding both the scene and the ego-motion, i.e. the motion of the individual to whom the camera is mounted. In Meng, J. Sanchez, Morel, Bertozzi, and Brantingham [2017], the authors develop an algorithm for the ego-motion classification, combining the MBO scheme for multi-class semi-supervised learning with an inverse compositional algorithm Sánchez [2016] to estimate transformations between successive frames. Thus the video is preprocessed to obtain an eight dimensional feature vector for each frame corresponding to the Left-Right; Up-Down; Rotational; and Forward-Backward motions of the camera wearer along with the frequencies of each of these motions. This is a gross reduction of the action of the video to a very low dimensional vector. These ideas have been extended by students at UCLA to higher dimensional feature vectors encoding both the egomotion and information from the scene Akar, Chen, Dhillon, Song, and T. Zhou [2017]. Studies of the effect of class size are made possible by an extensive effort during a summer REU to hand classify sufficient video footage to provide ground truth for a

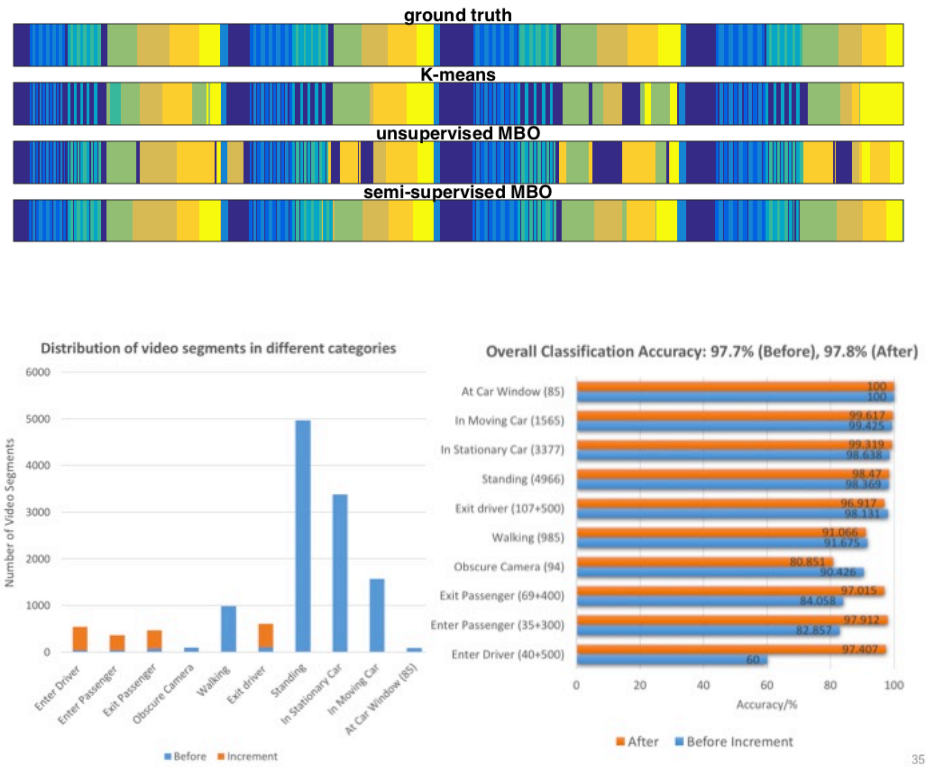


Figure 3: Top: Ego-motion classification results of the QUAD video Meng, J. Sanchez, Morel, Bertozzi, and Brantingham [2017]. The 9 colors represent 9 different ego-motion classes: standing still (dark blue), turning left (moderate blue), turning right (light blue), looking up (dark green) and looking down (light green), jumping (bud green), stepping (aztec gold), walking (orange), runing (yellow). Copyright © 2018 Springer Nature. Published with permission of Springer Nature. Bottom: Semi-supervised MBO on LAPD Body Worn Cameras, using a complex motion-feature vector for each frame. Both examples use 10% of the data as training Akar, Chen, Dhillon, Song, and T. Zhou [2017].

larger study. Figure 3 shows results from Meng, J. Sanchez, Morel, Bertozzi, and Brantingham [2017] for the QUAD video in Baker and Matthews [2004] for various clustering algorithms described here and results on LAPD body worn camera data from Akar, Chen, Dhillon, Song, and T. Zhou [2017]. In the second example there are catagories of activity with relatively small sample sizes and those are predominantly misclassified by this method. As a test, the smaller samples were aumented to an incremental level (shown in orange) by simply duplicating the video frame feature vectors, thereby increasing both the size of the classes and the size of the training data. The results show a marked increase in accuracy. Most research papers make use of scripted datasets with real-world data being relatively scarce for use in basic research (compared to its rate of capture in real-world applications). This case study shows the need for more algorithm development and theoretical results centered around the realities of real datasets. Privacy and proprietary reasons often hinder the use of such datasets in basic research and reproduceability can be hindered by the lack of public access to such data, nevertheless there is a strong societal need for more work to be done on real-world datasets and published in the scientific literature. Another important point related to this study is the fact that the data compression of entire video footage into low dimensional feature vectors (less than 100 dimensions per frame or group of frames) can serve as a tool for anonymizing sensitive data in order to develop computational algorithms on online computational platforms in shared workspaces, which can be forbidden to directly handle sensitive data. Such steps are imperative if one is to work with real-world data in an academic environment.

3 Uncertainty quantification (UQ) for graphical metrics

Semi-supervised learning combined both unlabeled data with labeled data; however, as the BWV example elucidates, in many applications there are so many unlabelled data points that one can not hand label everything. Moreover, there is a myriad of work in computer science and applied mathematics addressing the development of algorithms and a modest amount of work addressing performance of these methods in terms of convergence for problems such as clustering. For real-world applications, existing work does not address many obvious concerns - it is common to use methodologies ‘out of the box’ with measurements of performance of the methods based on ground truth information for toy/test problems but little information available regarding the likelihood of the results in general for real-world applications when ground truth is not available or when the existing ground truth is limited to a small percentage of training data. The graphical models and methods are particularly appropriate for the development of new mathematical methodology for uncertainty quantification, because of their organization around graph Laplacian matrices and nonlinear functionals that directly use these operators. In Blum and Chawla

[2001], using a graph min-cut problem for binary semi-supervised learning. This is equivalent to a maximum a posteriori (MAP) estimation on Bayesian posterior distribution for a Markov random field (MRF) over the discrete state space of binary labels X. Zhu [2005]. Inference for multi-label discrete MRFs is typically intractable Dahlhaus, Johnson, Papadimitriou, Seymour, and Yannakakis [1992]. Some approximate algorithms have been developed for the multi-label case Y. Boykov, Veksler, and Zabih [2001, 1998] and Madry [2010], with application to imaging tasks Y. Y. Boykov and Jolly [2001], Berthod, Kato, Yu, and Zerubia [1996], and Li [2012]. In X. Zhu, Ghahramani, Lafferty, et al. [2003], relaxed the discrete state space to a continuous real-variable setting, and modeled the semi-supervised learning problem as a Gaussian random field. D. Zhou, Bousquet, Lal, Weston, and Schölkopf [2004] generalized the model to handle label noise, and also generalized it to the case of directed graphs D. Zhou, Hofmann, and Schölkopf [2004]. We note that this earlier work of Zhou was a precursor to the nonloncal means graph developed by Buades Coll and Morel Buades, Coll, and Morel [2005] and further developed by Gilboa and Osher Gilboa and Osher [2007, 2008] that inspired some of the methods in the work of the Author and collaborators for the MBO scheme on graphs Merkurjev, Kostic, and Bertozzi [2013].

The probit classification method in C. K. I. Williams and Rasmussen [1996] uses the same prior as in X. Zhu, Ghahramani, Lafferty, et al. [2003] but the data takes on binary values, found from thresholding the underlying continuous variable, and thereby provides a link between the combinatorial and continuous state space approaches. The probit methodology is often implemented via MAP optimization – that is the posterior probability is maximized rather than sampled – or an approximation to the posterior is computed, in the neighborhood of the MAP estimator. In the context of MAP estimation, the graph-based terms act as a regularizer, in the form of the graph Dirichelet energy $\frac{1}{2}\langle u, Lu\rangle$, with L the symmetrized graph Laplacian. A formal framework for graph-based regularization can be found in Belkin, Matveeva, and Niyogi [2004] and Belkin, Niyogi, and Sindhwani [2006]. More recently, other forms of regularization have been considered such as the graph wavelet regularization Shuman, Faraji, and Vandergheynst [2011] and Hammond, Vandergheynst, and Gribonval [2011].

The author and collaborators Bertozzi, Luo, Stuart, and Zygalakis [2017] have developed UQ methodologies for graph classification based on optimization over real-valued variables developed in the works discussed in Bertozzi and Flenner [2016]. The UQ approach builds on the following ideas: (a) that a Bayesian formulation of the classification problem gives UQ automatically, (b) that fully Bayesian sampling is possible if one develops a methodology that scales well with respect to large graph size. The existing work on scalable algorithms for minimizing the graph GL functional is critical for (b). We have results for Gaussian noise models for binary classifiers that leverage several Bayesian models extended to classification on graphs; via the posterior distribution on the labels,

these methods automatically equip the classifications with measures of uncertainty. These models build on well-know Bayesian models in Euclidean space with fewer dimensions that arise in machine learning.

The probit classification C. K. I. Williams and Rasmussen [1996] nicely extends to graphical models. This involves defining a Gaussian measure through the graph Dirichlet energy and choosing a likelihood function involving thresholding a continuum latent variable plus noise. Bayes theorem provides a direct calculation of the posterior and its PDF. One can then compute the maximum a posteriori estimation (MAP) which is the minimizer of the negative of the log posterior, a convex function in the case of probit. The probit model may not be the most appropriate when the data is naturally segmented into groups, in which the variation is often understood to occur within the group. The next step is to build on several additional methods that have this structure; they are the level set method for Bayesian inverse problems Iglesias, Lu, and Stuart [2015], atomic noise models, and the Ginzburg-Landau optimization-based classifier Bertozzi and Flenner [2012] and van Gennip and Bertozzi [2012], which by virtue of its direct use of the Dirichlet energy, is tractable to generalize to a Bayesian setting. In all cases the posterior $P(u|y)$ has the form

$$P(u|y) \propto \exp(-J(u)), \quad J(u) = \frac{1}{2c}\langle u, Lu\rangle + \Phi(u)$$

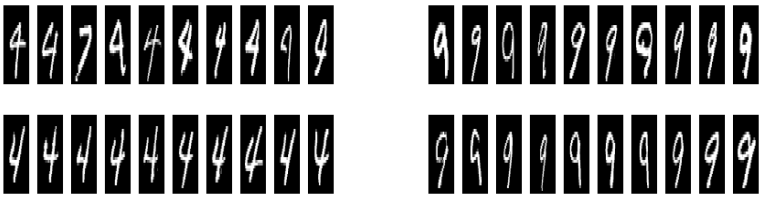
for some function Φ, different for each of the four models - and for which the Ginzburg-Landau case, the independent variable is a real-valued relaxation of label space, rather than an underlying latent variable which may be thresholded by $S(\cdot)$ into label space.) Here L is the graph Laplacian and c is a known scaling constant. The choice of scaling of L should be consistent with the scaling used for one of the learning methods (without UQ) discussed in the previous sections. Furthermore, the MAP estimator is the minimizer of J. Φ is differentiable for the Ginzburg-Landau and probit models, but not for the level set and atomic noise models. We are interested in algorithms for both sampling and MAP estimation.

In Bertozzi, Luo, Stuart, and Zygalakis [2017] the authors develop efficient numerical methods, suited to large data-sets, for both MCMC-based sampling as well as gradient-based MAP estimation. In order to induce scalability with respect to size of the graph, we consider the pCN method described in Cotter, G. O. Roberts, Stuart, and White [2013] and introduced in the context of diffusions by Beskos in Beskos, G. Roberts, Stuart, and Voss [2008] and by Neal in the context of machine learning Neal [1998]. The standard random walk Metropolis (RWM) algorithm suffers from the fact that the optimal proposal variance or stepsize scales inverse proportionally to the dimension of the state space G. O. Roberts, Gelman, Gilks, et al. [1997], which is the graph size N in this case. The pCN method was designed so that the proposal variance required to obtain a given acceptance

probability scales independently of the dimension of the state space, hence in practice giving faster convergence of the MCMC when compared with RWM. For graphs with a large number of nodes N, it is prohibitively costly to directly sample from the distribution μ_0, since doing so involves knowledge of a complete eigen-decomposition of L. In machine learning classification tasks it is common to restrict the support of u to the eigenspace spanned by the first ℓ eigenvectors with the smallest non-zero eigenvalues of L (hence largest precision) and this idea may be used to approximate the pCN method. The Author and collaborators have made use of both low rank Fowlkes, Belongie, F. Chung, and Malik [2004] approximations of nonsparse matrices and fast algorithms for computing the smallest non-zero eigenvalues of sparse matrices Anderson [2010]. The upshot is a confidence score for the class assignment for binary classifiers, based on the node-wise posterior mean of the thresholded variable.

An example is shown in Bertozzi, Luo, Stuart, and Zygalakis [2017] with the MNIST database consists of $70,000$ images of size 28×28 pixels containing the handwritten digits 0 through 9; see LeCun, Cortes, and Burges [1998] for details. The nodes of the graph are the images and as feature vectors one uses the leading 50 principal components given by PCA; thus the feature vectors at each node have length $d = 50$. We construct a K-nearest neighbor graph with $K = 20$ for each pair of digits considered. Namely, the weights a_{ij} are non-zero if and only if one of i or j is in the K nearest neighbors of the other. The non-zero weights are set using a local rescaling as in Equation (3). For more details see Bertozzi, Luo, Stuart, and Zygalakis [2017]. The noise variance γ is set to 0.1, and 4% of fidelity points are chosen randomly from each class. The probit posterior is used to compute a node-wise posterior mean. Figure 4 shows that nodes with scores posterior mean closer to the binary ground truth labels ± 1 look visually more uniform than nodes with score far from those labels. This illustrates that the posterior mean can differentiate between outliers and inliers that align with human perception.

There are a number of natural avenues to explore building on the work in Bertozzi, Luo, Stuart, and Zygalakis [ibid.]; (a) there is a natural question of whether one works in label space, or a relaxation of it, as in GL, or with a latent variable as in probit - more investigation of the models on toy problems should elucidate this; (b) the models proposed above are rather simplistic and may not be best tuned to real datasets - it would be interesting to develop a preprocessing method to probe the data and to learn something about the data - preliminary results using probit as a preprocessing step for GL show some benefit to such hybrid methods; (c) these binary classification models will be extended to multiclass - the GL methodology is nicely extended in Garcia-Cardona, Merkurjev, Bertozzi, Flenner, and Percus [2014] but not in a Bayesian setting, whereas the other methods do not directly extend as easily although recursive methods can be helpful; (d) all of the methods described here involve a combination of graph Laplacian diffusion and thresholding analogous to the MBO scheme on graphs developed by the PI and collaborators Merkurjev, Kostic, and



(a) Fours in MNIST (b) Nines in MNIST

Figure 4: "Hard to classify" vs "easy to classify" nodes in the MNIST $(4, 9)$ dataset under the probit model. Here the digit "4" is labeled +1 and "9" is labeled -1. The top (bottom) row of the left column corresponds to images that have the lowest (highest) values of the node-wise posterior mean out of all the "4" digist. The right column is organized in the same way for images with ground truth labels 9 except the top row now corresponds to the highest values of posterior mean. Higher posterior mean indicates higher confidence that the image is a 4 and not a "9", hence the top row could be interpreted as images that are "hard to classify" by the current model, and vice versa for the bottom row. See Bertozzi, Luo, Stuart, and Zygalakis [2017] for more details.

Bertozzi [2013]. Those algorithms also involve graph diffusion plus thresholding in a different way from the Bayesian statistical methods - and some measurement of similarity or difference should be made; (e) furthermore, one can consider unsupervised problems - for example the hybrid method in the paper Hu, Sunu, and Bertozzi [2015] that considers k-means plus the MBO scheme for clustering; (f) finally there are natural UQ questions that will arise from the other thrusts of the project. For example, for data fusion methods, the development of multimodal graphical models provides a natural context in which to extend the UQ methodology to these more complex data problems, providing not only insight into the results but also insight into the best choice of models for the fusion.

4 Network Analysis

The above discussion of uncertainty quantification was mainly directed at graphs that arise from machine learning problems involving similarity matrices that result from pairwise comparisons of high dimensional data. Another natural class of graphs are networks - for example social network graphs such as those that arise from social media, transportation networks, and other examples M. E. J. Newman [2010]. Mathematical models and algorithms for structure in networks have led to a large body of work, for example in the

physics literature, that has largely happened independent of the work carried out in machine learning. There is a need to develop novel ideas in both areas and in some cases, especially with security applications, there is a need to have models that fit both machine learning (big data) problems and network problems.

There are several papers in the literature that connect network clustering to machine learning, and for brevity we mention a few methods here, including issues that arise when viewing network analysis methods in the context of machine learning: (a) in Peel, Larremore, and Clauset [2016] the authors consider metadata as 'ground truth' and prove a general "No Free Lunch" theorem for community detection, implying that no algorithm can perform better than others across all inputs; (b) Newman M. E. J. Newman [2013] considers spectral methods for three different problems - network modularity (discussed below), statistical inference, and normalized graph partitioning, concluding that algorithmically the spectral methods are the same for each class of problems; (c) Devooght et. al. consider random-walk based modularity applied to semi-supervised learning Devooght, Mantrach, Kivimäki, Bersini, Jaimes, and Saerens [2014] focusing on paths on the graph rather than edges. A review of clustering of graphs, including attributes (semi-supervision) from a network perspective is Bothorel, Cruz, Magnani, and Micenkova [2015]. A recent review of community detection methods on networks can be found in Fortunato and Hric [2016].

A few years ago the Author and collaborators developed the first paper to directly connect network modularity optimization and total variation minimization on graphs, using the null model introduced by Newman and Girvan in Girvan and M. E. J. Newman [2004]. To explain in more detail, the modularity of a network partition measures the fraction of total edge weight within communities versus what one might expect if edges were placed randomly according to some null model. More specifically, the objective is to **maximize** the modularity

$$Q = \frac{1}{2m}\sum_{ij}(w_{ij} - \gamma P_{ij})\delta(g_i, g_j)$$

over all possible partitions where g_i is the group assignment for node i. Here P_{ij} is a probability null model (e.g. $P_{ij} = k_i k_j/2m$) where $k_j = \sum_i w_{ij}$ and $2m$ is the total volume of the graph ($\sum_i k_i$) and γ is a resolution parameter. Our work Hu, Laurent, Porter, and Bertozzi [2013] shows that maximizing Q is equivalent to a graph cut problem that can be rewritten using the TV functional:

$$Min_{u:G\to V^{\hat{n}}} E(u) = |u|_{TV} - \gamma|u - m_2(u)|^2_{L_2}$$

for the case of $\hat{n}$ classes where $V^{\hat{n}}$ are the end nodes of the $\hat{n}$-dimensional simplex and m_2 denotes a simple moment whose constraint can be introduced in a computationally tractable forcing term. Her u denotes the class assignment and takes vales on the corners

of the simplex. One can then use the above ideas to minimize this functional over all possible numbers of clusters $\hat{n}$. We note that our work also sheds new light on some of the other papers mentioned above. For example the TV-modularity connection is a direct relationship between the graph cuts and modularity, beyond the connection between spectral algorithms. Furthermore, the method used in Hu, Laurent, Porter, and Bertozzi [2013] builds on the graph heat equation, which is, roughly speaking, a mean field limit of a random walk dynamics. It uses directly the MBO scheme on graphs Merkurjev, Kostic, and Bertozzi [2013] from Section 2.1 and multiclass methods for TV-minimization on graphs Garcia-Cardona, Merkurjev, Bertozzi, Flenner, and Percus [2014]. The idea in these recent papers is to develop algorithms for graph clustering, in particular TV minimization, which is equivalent to graph cut minimization on weighted graphs when applied to partition functions. In the case of modularity optimization, the main idea is that maximizing the modularity functional, when applied to a fixed number of classes, is equivalent to minimization an energy for the assignment function, comprised of the graph total variation minus a second moment term. This opens the door to apply compressed sensing ideas to modularity optimization, a superior but computationally more complex method than spectral clustering. More can be done in this area and we propose to work on problems of direct relevance to multimodal graphs such as those that arise from composite information such as spatial nonlocal means, as in the example above, social networks, and latent information such as text-content topics from twitter.

The method is very scalable, which allows the algorithmic approach to go far beyond sparse network analysis, providing a new tool for analyzing large similarity graphs in machine learning. For example, the MNIST dataset LeCun, Cortes, and Burges [1998] of 70,000 handwritten digits, with tens of thousands of nodes, this approach is 10-100 times faster computationally than the GenLouvain algorithm Jutla, Jeub, and Mucha [n.d.] and produces comparable quality results, exceeding that of basic fast greedy algorithms such as M. E. Newman [2006] and Blondel, Guillaume, Lambiotte, and Lefebvre [2008b] and outperforming all other unsupervised clustering methods that we are aware of. What is most striking is the ability to correctly classify and identify the number of classes, in a fairly short amount of computational time. In general unsupervised clustering without prior knowledge of the number of classes is a very difficult problem for large datasets. So methodologies that are efficient enough to be useful for large data (including scalability) are needed. For example, in Hu, van Gennip, Hunter, Bertozzi, and Porter [2012] the GenLouvain code Jutla, Jeub, and Mucha [n.d.] was tested on the nonlocal means graph for a basic color image with excellent segmentation results for unsupservised clustering but with a run time that was neither practical nor scalable. Although for completeness one should compare with other methods such as the C++ implementations of Blondel of the Louvain method Blondel, Guillaume, Lambiotte, and Lefebvre [2008a]. We note that even the soft clustering methods like Nonnegative Matrix Factorization and Latent

Dirichlet Allocation require the user to specify the number of classes and still have the restriction that they are built around a linear mixture model.

4.1 Network analysis and machine learning. An interesting line of inquiry is to unify the graphical models developed independently by the machine learning community and by the network science community for unsupervised learning. We believe that there is an opportunity to improve unsupervised learning algorithms (built on similarity graphs) for data science as well as to further understand the link between network structure and algorithm type. Starting from our earlier work on network modularity as a constrained multiclass graph cut problem, we address modularity as a constrained balanced cut problem in which convex methods can be used apart from the constraint. In a new work Boyd, Bai, X. C. Tai, and Bertozzi [2017] we have identified four different equivalent formulations of the modularity problem which we term soft balanced cut, penalized balanced cut, balanced total variation (TV) minimization, and penalized TV minimization.

Theorem 1 (Equivalent forms of modularity Boyd, Bai, X. C. Tai, and Bertozzi [ibid.])**.** *For any subset S of the nodes of G, define $\mathrm{vol}\, S = \sum_{i \in S} k_i$. Then the following optimization problems are all equivalent:*

(8)

$$\textit{Std. form:} \quad \underset{\hat{n} \in \mathbb{N}, \{A_\ell\}_{\ell=1}^{\hat{n}} \in \Pi(G)}{\operatorname{argmax}} \quad \sum_{\ell=1}^{\hat{n}} \sum_{ij \in A_\ell} w_{ij} - \gamma \frac{k_i k_j}{2m},$$

(9)

$$\textit{Bal. cut (I):} \quad \underset{\hat{n} \in \mathbb{N}, \{A_\ell\}_{\ell=1}^{\hat{n}} \in \Pi(G)}{\operatorname{argmin}} \quad \sum_{\ell=1}^{\hat{n}} \left(\mathrm{Cut}\left(A_\ell, A_\ell^c\right) + \frac{\gamma}{2m} \left(\mathrm{vol}\, A_\ell\right)^2 \right),$$

(10)

$$\textit{Bal. cut (II):} \quad \underset{\hat{n} \in \mathbb{N}, \{A_\ell\}_{\ell=1}^{\hat{n}} \in \Pi(G)}{\operatorname{argmin}} \quad \sum_{\ell=1}^{\hat{n}} \left(\mathrm{Cut}\left(A_\ell, A_\ell^c\right) + \frac{\gamma}{2m} \left(\mathrm{vol}\, A_\ell - \frac{2m}{\hat{n}} \right)^2 \right) + \gamma \frac{2m}{\hat{n}}$$

(11)

$$\textit{Bal. TV (I):} \quad \underset{\hat{n} \in \mathbb{N}, u \in \Pi(G)}{\operatorname{argmin}} \quad |u|_{TV} + \frac{\gamma}{2m} \left\| k^T u \right\|_2^2$$

(12)

$$\textit{Bal. TV (II):} \quad \underset{\hat{n} \in \mathbb{N}, u \in \Pi(G)}{\operatorname{argmin}} \quad |u|_{TV} + \frac{\gamma}{2m} \left\| k^T u - \frac{2m}{\hat{n}} \right\|_2^2 + \gamma \frac{2m}{\hat{n}}.$$

Each of the preceding forms has a different interpretation. The original formulation of modularity was based on comparison with a statistical model and views communities as regions that are more connected than they would be if edges were totally random. The cut formulations represent modularity as favoring sparsely interconnected regions with balanced volumes, and the TV formulation seeks a piecewise-constant partition function u whose discontinuities have small perimeter, together with a balance-inducing quadratic penalty. The cut and TV forms come in pairs. The first form (labelled "I") is simpler to write but harder to interpret, while the second (labelled "II") has more terms, but the nature of the balance term is easy to understand, as it is minimized (for fixed $\hat{n}$) when each community has volume $\frac{2m}{\hat{n}}$.

In addition to providing a new perspective on the modularity problem in general, this equivalence shows that modularity optimization can be viewed as minimizing a convex functional but subject to a binary constraint. These methodologies provide a direct connection between modularity and other balance cut problems such as the Cheeger or Ratio cut and a connection to convex optimization methods already developed for semi-supervised learning on graphs Merkurjev, Bae, Bertozzi, and X.-C. Tai [2015] and Bae and Merkurjev [2016]. A significant emphasis on spectral algorithms exists in the literature on graph cut methods for networks, see e.g. M. E. Newman [2006] for spectral methods for modularity vs. other spectral methods applied to networks, and a large literature on accuracy of spectral approximations for the Cheeger cut (e.g. Ghosh, Teng, Lerman, and Yan [2014]). What distinguishes our approach from other efforts is the focus on non-network data using a network approach. There are many reasons to do this. At the forefront is the ability to do unsupervised clustering well, without knowing the number of clusters.

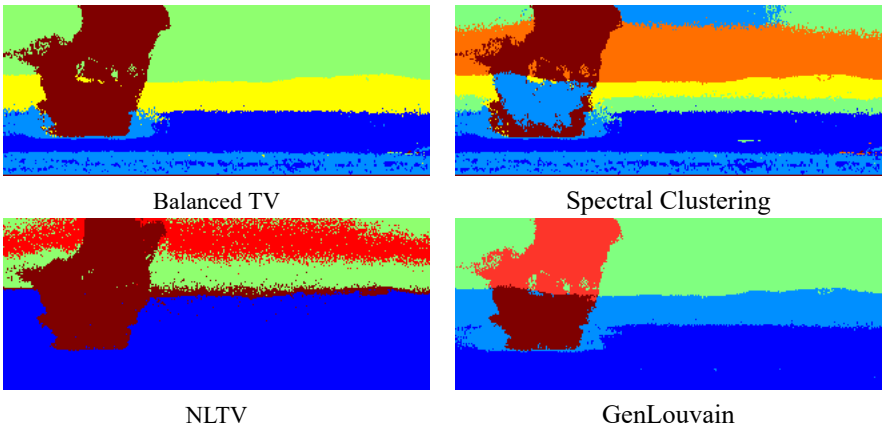


Figure 5: Segmentations Boyd, Bai, X. C. Tai, and Bertozzi [2017] of the plume hyperspectral video using different methods. The Balanced TV is the only method that has the whole plume into a single class without any extraneous pixels (NLTV method from W. Zhu, Chayes, Tiard, S. Sanchez, Dahlberg, Bertozzi, Osher, Zosso, and Kuang [2017]).

The ideas developed in Boyd, Bai, X. C. Tai, and Bertozzi [2017] show that, while modularity optimization is inherently nonconvex, that working with it as a constrained convex optimization problem produces results that are noticeably improved compared to prior methods that do not use such a formulation, including the formulation in Hu, Laurent, Porter, and Bertozzi [2013]. Another relevant recent work is Merkurjev, Bae, Bertozzi, and X.-C. Tai [2015] that develops convex optimization methods to find global minimizers of graph cut problems for semi-supervised learning. This work is loosely related to Boyd, Bai, X. C. Tai, and Bertozzi [2017] and serves as a resource for the use of L1 compressed sensing methods and max flow methods for constrained cut problems. Regarding benchmark testing, we note that Bazzi, Jeub, Arenas, Howison, and Porter [2016] has developed a new class of benchmark networks that can be tested with algorithms in addition to the LFR benchmarks.

4.2 Data fusion, multilayer graphs and networks. There are many works in the literature for data fusion that do not use graphs - they require a specific connection between the information and are typically not flexible to extend to unrelated data fusion problems. One such example pan sharpening of remote sensing images in which a panchromatic sensor has higher spatial resolution than a multispectral sensor Möller, Wittman, Bertozzi, and Burger [2012]. Graphical methods for data fusion are still in their infancy with a few ideas

in the literature but very little theoretical understanding of these approaches. Some examples from the works of the Author include (a) a homotopy parameter for social network vs. spatial embedding used to study Field Interview cards from Los Angeles van Gennip, Hunter, et al. [2013] (Figure 1 in this paper), (b) a variational model for MPLE for statistical density estimation of crime data Woodworth, Mohler, Bertozzi, and Brantingham [2014] that uses a nonlocal means-based graphical model for high spatial resolution housing density information as a regularization parameter, and (c) a threshold-based similarity graph for combined LIDAR and multispectral sensors Iyer, Chanussot, and Bertozzi [2017]. The network science community has different methods for fusing network data compared to traditional methods used in sensor fusion. One might explore the similarities and differences between the network science models and the sensor fusion models and to examine and identify opportunities to bring ideas from one community into the other through the use of graphical models, along with related rigorous analytical results of relevance.

Another problem is to develop algorithms based on models for more complex networks - for example multi-layer modularity optimization as proposed by Porter and colleagues Mucha, Richardson, Macon, Porter, and Onnela [2010] (Science 2010) and more recent papers that build on that work (e.g. Bazzi, Porter, S. Williams, McDonald, Fenn, and Howison [2016] and M. E. J. Newman and Peixoto [2015]). The multilayer approach can give much better granularity of clustering in social networks however it is even more computationally prohibitive than regular modularity in the case of larger datasets (e.g. tens of thousands of nodes). Multilayer models are able to work with more complex similarity graphs, such as those that might arise from multimodal data, although little work has been done unifying these ideas. As an example, for the LAPD field interview cards studied in van Gennip, Hunter, et al. [2013], one might analyze what additional information might be encoded in a multilayer network structure compared to a parametric homotopy model on a single layer graph. For multilayer graphs, we expect TV minimization methods to handle structures within a layer, however different methods may be required when strong connections arise across layers. We expect that different issues may arise when considering such graphs for data fusion rather than complex network applications.

For multilayer graph models models one could explore hybrid schemes that leverage the ultrafast segmentation that can be done for large clusters using something like MBO while using the combinatorial methods (e.g. Gen Louvain Jutla, Jeub, and Mucha [n.d.]) for the network structure that has some granularity. This is a main challenge when working with complex data such as the artificial LFR benchmarks Lancichinetti, Fortunato, and Radicchi [2008] that have a power law community distribution. One can also compare against the new benchmark graphs in Bazzi, Jeub, Arenas, Howison, and Porter [2016] using their code. Future work might involve a hybrid method that will have components of TV minimization methods such as the MBO scheme Merkurjev, Kostic, and Bertozzi [2013],

components of GenLouvain and possible post-processing steps such as Kernighan-Lin node-swapping M. E. Newman [2006], Porter, Onnela, and Mucha [2009], and Richardson, Mucha, and Porter [2009].

5 Final Comments

The Author and collaborators have developed rigorous analysis for the dynamics of the graphical MBO iteration scheme for semi-supervised learning van Gennip, Guillen, Osting, and Bertozzi [2014] and and this work could be extended to unsupervised, multiclass, classification methods such as those that arise in network modularity. The Author and Luo have developed theoretical convergence estimates Luo and Bertozzi [2017] for the Ginzburg-Landau convex-splitting method for semi-supervised learning for various versions of the graph Laplacian discussed above. For example, for the standard graph Laplacian we have maximum norm convergence results for minimizers of the Ginzburg-Landau energy using a combination of L^2-energy estimates and maximum principle results for the Laplacian operator Luo and Bertozzi [ibid.]. The GL energy is a non-convex functional, so those results prove convergence to a local minimizer rather than a global one and can require modest *a posteriori* estimates to guarantee convergence; these are ones that can be built directly into the code. One of the rigorous results proved in Luo and Bertozzi [ibid.] is that the *convergence and stability of the scheme are independent of the size of the graph, and of its sparseness*, an important feature for scalability of methods.

Another issue that is rarely discussed for either the semi-supervised or unsupervised cases, regarding similarity graphs, is whether to thin the graph before performing classification or to use the fully connected graph in connection with a low rank approximation of the matrix such as the Nyström extension, discussed above. Research is needed to develop rigorous estimates related to the thinning of the graph in conjunction with models for clustering data - for example we can take examples models built on the Gaussian priors in the previous section on UQ and develop estimates for what is lost from the matrix when removing edges with smaller weights, a common process using e.g. a k-nearest neighbor graph. This problem involves the role of the graph structure on optimization problems and can also benefit from existing results from the network literature.

Acknowledgments. The author thanks Z. Meng for help with Figure 2 and Mason Porter for useful comments. This work discusses a body of research that would not have been possible without the perserverence and insight of many students and postdocs including Egil Bae, Zach Boyd, Julia Dobrosotskaya, Cristina Garcia-Cardona, Nestor Guillen, Huiyi Hu, Blake Hunter, Geoffrey Iyer, Tijana Kostic, Hao Li, Xiyang Luo, Zhaoyi Meng, Ekaterina

Merkurjev, Braxton Osting, Carola Schönlieb, Justin Sunu, and Yves van Gennip. Numerous ideas came from interactions with many colleagues including Chris Anderson, P. Jeffrey Brantingham, Tony Chan, Jocelyn Chanussot, Fan Chung, Arjuna Flenner, Thomas Laurent, Kristina Lerman, Stanley Osher, J. M. Morel, Mason Porter, George Tita, Andrew Stuart and Xue-Cheng Tai, and Luminita Vese.

References

O. Akar, H. Chen, A. Dhillon, A. Song, and T. Zhou (2017). "Body Worn Video". Technical report from 2017 summer REU on classification of BWV from LAPD; Andrea L. Bertozzi, M. Haberland, H. Li, P. J. Brantingham, and M. Roper faculty mentors (cit. on pp. 3892–3894).

Christopher R. Anderson (Sept. 2010). "A Rayleigh-Chebyshev Procedure for Finding the Smallest Eigenvalues and Associated Eigenvectors of Large Sparse Hermitian Matrices". *J. Comput. Phys.* 229.19, pp. 7477–7487 (cit. on p. 3897).

E. Bae and E. Merkurjev (2016). "Convex Variational Methods for Multiclass Data Segmentation on Graphs" (cit. on p. 3902).

Simon Baker and Iain Matthews (Feb. 2004). "Lucas-Kanade 20 Years On: A Unifying Framework". *International Journal of Computer Vision* 56.3, pp. 221–255 (cit. on p. 3894).

M. Bazzi, L. G. S. Jeub, A. Arenas, S. D. Howison, and M. A. Porter (2016). "Generative benchmark models for mesoscale structure in multilayer networks" (cit. on pp. 3903, 3904).

M. Bazzi, M. A. Porter, S. Williams, M. McDonald, D. J. Fenn, and S. D. Howison (2016). "Community Detection in Temporal Multilayer Networks, with an Application to Correlation Networks". *Multiscale Model. Simul.* 14.1, pp. 1–41 (cit. on p. 3904).

Mikhail Belkin, Irina Matveeva, and Partha Niyogi (2004). "Regularization and semi-supervised learning on large graphs". In: *International Conference on Computational Learning Theory*. Springer, pp. 624–638 (cit. on p. 3895).

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani (2006). "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples". *Journal of machine learning research* 7.Nov, pp. 2399–2434 (cit. on p. 3895).

Serge Belongie, Charless Fowlkes, Fan Chung, and Jitendra Malik (2002). "Spectral partitioning with indefinite kernels using the Nyström extension". In: *European Conference on Computer Vision*, pp. 531–542 (cit. on p. 3888).

Marc Berthod, Zoltan Kato, Shan Yu, and Josiane Zerubia (1996). "Bayesian image classification using Markov random fields". *Image and vision computing* 14.4, pp. 285–295 (cit. on p. 3895).

Andrea L. Bertozzi and Arjuna Flenner (2012). "Diffuse interface models on graphs for classification of high dimensional data". *Multiscale Modeling & Simulation* 10.3, pp. 1090–1118 (cit. on pp. 3886, 3896).

– (2016). "Diffuse Interface Models on Graphs for Classification of High Dimensional Data". *SIAM Review* 58.2, pp. 293–328 (cit. on pp. 3886, 3889, 3895).

Andrea L. Bertozzi, X. Luo, A. M. Stuart, and K. C. Zygalakis (2017). "Uncertainty Quantification in the Classification of High Dimensional Data" (cit. on pp. 3895–3898).

Dimitri Bertsekas (1979). "A Distributed Algorithm for the Assignment Problem". Technical report, MIT (cit. on p. 3892).

A. Beskos, G. Roberts, A. M. Stuart, and J. Voss (2008). "MCMC methods for diffusion bridges". *Stochastics and Dynamics* 8.03, pp. 319–350 (cit. on p. 3896).

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre (2008a). *Louvain Method: Finding Communities in Large Networks* (cit. on p. 3900).

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre (2008b). "Fast unfolding of communities in large networks". *J. Stat. Mech: Theory Exp.* 2008.10, P10008 (cit. on p. 3900).

Avrim Blum and Shuchi Chawla (2001). "Learning from labeled and unlabeled data using graph mincuts". *Proc. 18th Int. Conf. Mach. Learning (ICML)* (cit. on p. 3894).

J. Bosch, S. Klamt, and M. Stoll (2016). "Generalizing diffuse interface methods on graphs: non-smooth potentials and hypergraphs" (cit. on p. 3887).

Cécile Bothorel, Juan David Cruz, Matteo Magnani, and Barbora Micenkova (2015). "Clustering attributed graphs: models, measures and methods". *Network Science* 3.3, pp. 408–444 (cit. on p. 3899).

Z. Boyd, E. Bai, X. C. Tai, and Andrea L. Bertozzi (2017). "Simplified energy landscape for modularity using total variation" (cit. on pp. 3887, 3901, 3903).

Yuri Y Boykov and M-P Jolly (2001). "Interactive graph cuts for optimal boundary & region segmentation of objects in ND images". In: *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*. Vol. 1. IEEE, pp. 105–112 (cit. on p. 3895).

Yuri Boykov, Olga Veksler, and Ramin Zabih (1998). "Markov random fields with efficient approximations". In: *Computer vision and pattern recognition, 1998. Proceedings. 1998 IEEE computer society conference on*. IEEE, pp. 648–655 (cit. on p. 3895).

– (2001). "Fast approximate energy minimization via graph cuts". *IEEE Transactions on pattern analysis and machine intelligence* 23.11, pp. 1222–1239 (cit. on p. 3895).

Xavier Bresson, Thomas Laurent, David Uminsky, and James H. von Brecht (2012). "Convergence and Energy Landscape for Cheeger Cut Clustering". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., pp. 1385–1393 (cit. on pp. 3891, 3892).

Xavier Bresson, Thomas Laurent, David Uminsky, and James H. von Brecht (2013). "An Adaptive Total Variation Algorithm for Computing the Balanced Cut of a Graph" (cit. on p. 3892).

Antoni Buades, Bartomeu Coll, and J-M Morel (2005). "A non-local algorithm for image denoising". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 2. IEEE, pp. 60–65 (cit. on pp. 3884, 3888, 3895).

Luca Calatroni, Yves van Gennip, Carola-Bibiane Schönlieb, Hannah M. Rowland, and Arjuna Flenner (Feb. 2017). "Graph Clustering, Variational Image Segmentation Methods and Hough Transform Scale Detection for Object Measurement in Images". *Journal of Mathematical Imaging and Vision* 57.2, pp. 269–291 (cit. on p. 3887).

T. Chan and L. A. Vese (2001). "Active Contours without Edges". *IEEE Trans. Image Process.* 10, pp. 266–277 (cit. on p. 3888).

Fan R. K. Chung (1996). *Spectral Graph Theory*. American Mathematical Society (cit. on p. 3885).

S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White (2013). "MCMC methods for functions: modifying old algorithms to make them faster." *Statistical Science* 28.3, pp. 424–446 (cit. on p. 3896).

Elias Dahlhaus, David S Johnson, Christos H Papadimitriou, Paul D Seymour, and Mihalis Yannakakis (1992). "The complexity of multiway cuts". In: *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*. ACM, pp. 241–251 (cit. on p. 3895).

Robin Devooght, Amin Mantrach, Ilkka Kivimäki, Hugues Bersini, Alejandro Jaimes, and Marco Saerens (2014). "Random Walks Based Modularity: Application to Semi-supervised Learning". In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14. New York, NY, USA: ACM, pp. 213–224 (cit. on p. 3899).

Julia A. Dobrosotskaya and Andrea L. Bertozzi (2008). "A Wavelet-Laplace Variational Technique for Image Deconvolution and Inpainting". *IEEE Trans. Imag. Proc.* 17.5, pp. 657–663 (cit. on p. 3887).

– (2010). "Wavelet analogue of the Ginzburg-Landau energy and its Γ−convergence". *Int. Free Boundaries* 12.4, pp. 497–525 (cit. on p. 3887).

Selim Esedoğlu and Felix Otto (2015). "Threshold Dynamics for Networks with Arbitrary Surface Tensions". *Communications on Pure and Applied Mathematics* 68.5, pp. 808–864 (cit. on p. 3890).

Selim Esedoğlu and Yen-Hsi Richard Tsai (2006). "Threshold dynamics for the piecewise constant Mumford-Shah functional". *Journal of Computational Physics* 211.1, pp. 367–384 (cit. on pp. 3888, 3889).

Santo Fortunato and Darko Hric (2016). "Community detection in networks: A user guide". *Physics Reports* 659. Community detection in networks: A user guide, pp. 1–44 (cit. on p. 3899).

Charless Fowlkes, Serge Belongie, Fan Chung, and Jitendra Malik (2004). "Spectral grouping using the Nyström method". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.2, pp. 214–225 (cit. on pp. 3888, 3897).

Charless Fowlkes, Serge Belongie, and Jitendra Malik (2001). "Efficient spatiotemporal grouping using the Nyström method". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* Vol. 1, pp. I–231 (cit. on p. 3888).

Cristina Garcia-Cardona, Ekaterina Merkurjev, Andrea L. Bertozzi, Arjuna Flenner, and Allon G Percus (2014). "Multiclass data segmentation using diffuse interface methods on graphs". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.8, pp. 1600–1613 (cit. on pp. 3887, 3890, 3897, 3900).

Yves van Gennip and Andrea L. Bertozzi (2012). "Γ-convergence of graph Ginzburg-Landau functionals". *Advances in Differential Equations* 17.11–12, pp. 1115–1180 (cit. on pp. 3887, 3896).

Yves van Gennip, Nestor Guillen, Braxton Osting, and Andrea L. Bertozzi (2014). "Mean curvature, threshold dynamics, and phase field theory on finite graphs". *Milan J. of Math* 82.1, pp. 3–65 (cit. on pp. 3890, 3905).

Yves van Gennip, Blake Hunter, et al. (2013). "Community detection using spectral clustering on sparse geosocial data". *SIAM J. Appl. Math.* 73.1, pp. 67–83 (cit. on pp. 3885, 3886, 3904).

R. Ghosh, S.-H. Teng, K. Lerman, and X. Yan (2014). "The interplay between dynamics and networks: centrality, communities, and Cheeger inequality". *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (cit. on p. 3902).

Guy Gilboa and Stanley Osher (2007). "Nonlocal linear image regularization and supervised segmentation". *Multiscale Modeling & Simulation* 6.2, pp. 595–630 (cit. on pp. 3888, 3895).

– (2008). "Nonlocal operators with applications to image processing". *Multiscale Modeling & Simulation* 7.3, pp. 1005–1028 (cit. on pp. 3888, 3895).

M. Girvan and M. E. J. Newman (2004). "Finding and evaluating community structure in networks". *Phys. Rev. E.* 69 (cit. on p. 3899).

Tom Goldstein and Stanley Osher (2009). "The split Bregman method for L1-regularized problems". *SIAM Journal on Imaging Sciences* 2.2, pp. 323–343 (cit. on p. 3888).

David K Hammond, Pierre Vandergheynst, and Rémi Gribonval (2011). "Wavelets on graphs via spectral graph theory". *Applied and Computational Harmonic Analysis* 30.2, pp. 129–150 (cit. on p. 3895).

H. Hu, Y. van Gennip, B. Hunter, Andrea L. Bertozzi, and M. A. Porter (2012). "Multislice Modularity Optimization in Community Detection and Image Segmentation". *Proc. IEEE International Conference on Data Mining (Brussels), ICDM'12*, pp. 934–936 (cit. on p. 3900).

Huiyi Hu, Thomas Laurent, Mason A. Porter, and Andrea L. Bertozzi (2013). "A Method Based on Total Variation for Network Modularity Optimization using the MBO Scheme". *SIAM J. Appl. Math.* 73.6, pp. 2224–2246 (cit. on pp. 3887, 3899, 3900, 3903).

Huiyi Hu, Justin Sunu, and Andrea L. Bertozzi (2015). "Multi-class Graph Mumford-Shah Model for Plume Detection Using the MBO scheme". In: *Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, pp. 209–222 (cit. on pp. 3887, 3889, 3890, 3898).

Marco A Iglesias, Yulong Lu, and Andrew M Stuart (2015). "A Bayesian Level Set Method for Geometric Inverse Problems". arXiv: 1504.00313 (cit. on p. 3896).

G. Iyer, J. Chanussot, and Andrea L. Bertozzi (2017). "A graph-based approach for feature extraction and segmentation in multimodal images". *Proc. Int. Conf. Image Proc., Beijing*, pp. 3320–3324 (cit. on pp. 3887, 3904).

Matt Jacobs, Ekaterina Merkurjev, and Selim Esedoğlu (2018). "Auction dynamics: A volume constrained MBO scheme". *Journal of Computational Physics* 354.Supplement C, pp. 288–310 (cit. on p. 3892).

Inderjit S. Jutla, Lucas G. S. Jeub, and Peter J. Mucha (n.d.). *A generalized Louvain method for community detectio implemented in MATLAB* (cit. on pp. 3900, 3904).

R. V. Kohn and P. Sternberg (1989). "Local minimisers and singular perturbations". *Proc. Roy. Soc. Edinburgh Sect. A* 111, pp. 69–84 (cit. on p. 3887).

A. Lancichinetti, S. Fortunato, and F. Radicchi (2008). "Benchmark graphs for testing community detection algorithms". *Phys. Rev. E* 78.04, p. 046110 (cit. on p. 3904).

Yann LeCun, Corinna Cortes, and Christopher JC Burges (1998). *The MNIST database of handwritten digits* (cit. on pp. 3897, 3900).

Stan Z Li (2012). *Markov random field modeling in computer vision*. Springer Science & Business Media (cit. on p. 3895).

X. Luo and Andrea L. Bertozzi (2017). "Convergence Analysis of the Graph Allen-Cahn Scheme". *J. Stat. Phys.* 167.3, pp. 934–958 (cit. on p. 3905).

Ulrike von Luxburg (2007). "A tutorial on spectral clustering". *Statistics and computing* 17.4, pp. 395–416 (cit. on pp. 3884, 3885).

Aleksander Madry (2010). "Fast approximation algorithms for cut-based problems in undirected graphs". In: *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*. IEEE, pp. 245–254 (cit. on p. 3895).

Z. Meng, E. Merkurjev, A. Koniges, and Andrea L. Bertozzi (2017). "Hyperspectral Image Classification Using Graph Clustering Methods". *Image Processing Online (IPOL)* 7. published with code, pp. 218–245 (cit. on pp. 3887, 3889, 3891).

Zhaoyi Meng, Alice Koniges, Yun (Helen) He, Samuel Williams, Thorsten Kurth, Brandon Cook, Jack Deslippe, and Andrea L. Bertozzi (2016). "OpenMP Parallelization and Optimization of Graph-based Machine Learning Algorithm". *Proc. 12th International Workshop on OpenMP (IWOMP)* (cit. on pp. 3887, 3890).

Zhaoyi Meng, Javier Sanchez, Jean-Michel Morel, Andrea L. Bertozzi, and P. Jeffrey Brantingham (2017). "Ego-motion Classification for Body-worn Videos". accepted in the Proceedings of the 2016 Conference on Imaging, vision and learning based on optimization and PDEs, Bergen, Norway (cit. on pp. 3892–3894).

E. Merkurjev, E. Bae, Andrea L. Bertozzi, and X.-C. Tai (2015). "Global binary optimization on graphs for classification of high dimensional data". *J. Math. Imag. Vis.* 52.3, pp. 414–435 (cit. on pp. 3890, 3902, 3903).

E. Merkurjev, Andrea L. Bertozzi, and F. Chung (2016). "A semi-supervised heat kernel pagerank MBO algorithm for data classification" (cit. on p. 3887).

E. Merkurjev, J. Sunu, and Andrea L. Bertozzi (2014). "Graph MBO method for multiclass segmentation of hyperspectral stand-off detection video". In: *Proc. Int. Conf. Image Proc. (ICIP) Paris*. IEEE, pp. 689–693 (cit. on pp. 3887, 3888, 3891).

Ekaterina Merkurjev, Andrea L. Bertozzi, Xiaoran Yan, and Kristina Lerman (2017). "Modified Cheeger and ratio cut methods using the Ginzburg-Landau functional for classification of high-dimensional data". *Inverse Problems* 33.7, p. 074003 (cit. on pp. 3887, 3892).

Ekaterina Merkurjev, Tijana Kostic, and Andrea L. Bertozzi (2013). "An MBO scheme on graphs for classification and image processing". *SIAM Journal on Imaging Sciences* 6.4, pp. 1903–1930 (cit. on pp. 3887, 3889, 3890, 3895, 3897, 3900, 3904).

B. Merriman, J. Bence, and S. Osher (1992). "Diffusion generated motion by mean curvature". *Proc. Comput. Crystal Growers Workshop*, pp. 73–83 (cit. on p. 3889).

Michael Möller, Todd Wittman, Andrea L. Bertozzi, and Martin Burger (2012). "A Variational Approach for Sharpening High Dimensional Images". *SIAM Journal on Imaging Sciences* 5.1, pp. 150–178 (cit. on p. 3903).

Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela (2010). "Community structure in time-dependent, multiscale, and multiplex networks". *Science* 328.5980, pp. 876–878 (cit. on p. 3904).

R. Neal (1998). "Regression and classification using Gaussian process priors". *Bayesian Statistics* 6, p. 475 (cit. on p. 3896).

M. E. J. Newman (Oct. 2013). "Spectral methods for community detection and graph partitioning". *Phys. Rev. E* 88, p. 042822 (cit. on p. 3899).

Mark E J Newman (2010). *Networks: An Introduction*. Oxford, UK: Oxford University Press (cit. on p. 3898).

Mark E J Newman and Tiago P Peixoto (Aug. 2015). "Generalized Communities in Networks". English. *Phys. Rev. Lett.* 115.8, p. 088701 (cit. on p. 3904).

Mark EJ Newman (2006). "Modularity and community structure in networks". *Proc. Nat. Acad. Sci.* 103.23, pp. 8577–8582 (cit. on pp. 3900, 3902, 3905).

L. Peel, D. B. Larremore, and A. Clauset (2016). "The ground truth about metadata and community detection in networks" (cit. on p. 3899).

M. A. Porter, J.-P. Onnela, and P. J. Mucha (2009). "Communities in networks". *Notices Amer. Math. Soc.* 56.9, pp. 1082–1097, 1164–1166 (cit. on p. 3905).

Thomas Richardson, Peter J. Mucha, and Mason A. Porter (Sept. 2009). "Spectral tripartitioning of networks". *Phys. Rev. E* 80, p. 036111 (cit. on p. 3905).

Gareth O Roberts, Andrew Gelman, Walter R Gilks, et al. (1997). "Weak convergence and optimal scaling of random walk Metropolis algorithms". *The annals of applied probability* 7.1, pp. 110–120 (cit. on p. 3896).

J. Sánchez (2016). "The Inverse Compositional Algorithm for Parametric Registration". *Image Processing On Line*, pp. 212–232 (cit. on p. 3892).

Carola-Bibiane Schönlieb and Andrea L. Bertozzi (2011). "Unconditionally stable schemes for higher order inpainting". *Comm. Math. Sci.* 9.2, pp. 413–457 (cit. on p. 3888).

David I Shuman, Mohammadjavad Faraji, and Pierre Vandergheynst (2011). "Semi-supervised learning with spectral graph wavelets". In: *Proceedings of the International Conference on Sampling Theory and Applications (SampTA)*. EPFL-CONF-164765 (cit. on p. 3895).

Arthur Szlam and Xavier Bresson (2010). "Total Variation and Cheeger Cuts". In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. USA: Omnipress, pp. 1039–1046 (cit. on p. 3891).

M. Thorpe and F. Theil (2017). "Asymptotic Analysis of the Ginzburg-Landau Functional on Point Clouds". to appear in the Proceedings of the Royal Society of Edinburgh Section A: Mathematics, 2017 (cit. on p. 3887).

L. A. Vese and T. F. Chan (2002). "A multiphase level set framework for image segmentation using the Mumford-Shah model". *Int. J. Comput. Vis.* 50, pp. 271–293 (cit. on p. 3888).

Christopher K. I. Williams and Carl Edward Rasmussen (1996). *Gaussian Processes for Regression*. MIT (cit. on pp. 3895, 3896).

J. T. Woodworth, G. O. Mohler, Andrea L. Bertozzi, and P. J. Brantingham (2014). "Non-local Crime Density Estimation Incorporating Housing Information". *Phil. Trans. Roy. Soc. A* 372.2028 (cit. on p. 3904).

L. P. Yaroslavsky (1985). *Digital Picture Processing. An Introduction*. Springer-Verlag (cit. on p. 3884).

Lihi Zelnik-Manor and Pietro Perona (2004). "Self-tuning spectral clustering". In: *Advances in neural information processing systems*, pp. 1601–1608 (cit. on p. 3885).

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf (2004). "Learning with local and global consistency". *Advances in neural information processing systems* 16.16, pp. 321–328 (cit. on p. 3895).

Dengyong Zhou, Thomas Hofmann, and Bernhard Schölkopf (2004). "Semi-supervised learning on directed graphs". In: *Advances in neural information processing systems*, pp. 1633–1640 (cit. on p. 3895).

W. Zhu, V. Chayes, A. Tiard, S. Sanchez, D. Dahlberg, Andrea L. Bertozzi, S. Osher, D. Zosso, and D. Kuang (2017). "Unsupervised Classification in Hyperspectral Imagery With Nonlocal Total Variation and Primal-Dual Hybrid Gradient Algorithm". *IEEE Transactions on Geoscience and Remote Sensing* 55.5, pp. 2786–2798 (cit. on pp. 3887, 3903).

Xiaojin Zhu (2005). "Semi-supervised learning literature survey". Technical Report 1530, Computer Sciences, Univ. of Wisconsin-Madison (cit. on p. 3895).

Xiaojin Zhu, Zoubin Ghahramani, John Lafferty, et al. (2003). "Semi-supervised learning using Gaussian fields and harmonic functions". In: *ICML*. Vol. 3, pp. 912–919 (cit. on p. 3895).

Received 2017-12-02.

Andrea L. Bertozzi
Department of Mathematics
UCLA
bertozzi@math.ucla.edu