Mining Complex Hyperspectral ALMA Cubes for Structure with Neural Machine Learning

Erzsébet Merényi, senior member
Department of Statistics
Department of Electrical and Computer Engineering
Rice University
Houston, Texas 77005
Email: erzsebet@rice.edu

Joshua Taylor, student member
Department of Statistics
Rice University
Houston, Texas 77005
Email: jtay@rice.edu

Andrea Isella
Department of Physics and Astronomy
Rice University
Houston, Texas 77005
Email: isella@rice.edu

***Abstract*—Astronomy is producing the largest "Big Data" sets today and in the near future, with instruments such as the Atacama Large Millimeter and sub-millimeter Array (ALMA), the Large Synoptic Survey Telescope (LSST), and the Square Kilometer Array (SKA). These observations afford a deeper, wider, and more dynamic glimpse into the structure and composition of the universe than ever before. However, in addition to unprecedented volume, the data also exhibit unprecedented complexity, mandating new approaches for extracting and summarizing relevant information. ALMA data, in particular, challenges with very high dimensionality (measurements in a large number of spectral channels) where the dimensions represent both compositional information and velocities, and the high spectral resolution allows detailed interpretation of the kinematic structure of sources such as molecular clouds or protoplanetary disks. Traditional tools like moment maps can no longer fully exploit and visualize the rich information in these data. We present a neural map-based clustering approach that can utilize all spectral channels simultaneously and is capable of finding clusters of widely varying statistical properties, which are expected in these complex data sets. Many clustering methods, including modern graph segmentation algorithms, run into limitations when encountering such data. We demonstrate our tools, collectively named "NeuroScope", through structure mining from an ALMA image of the protoplanetary disk HD142527. We highlight the advantages for both the emerging details and visualization. In addition, we explore an augmentation of leading graph segmentation algorithms with NeuroScope products, which can lead to efficient full automation of our clustering process for fast distillation of large data sets on-board or in archives.**

I. Background on Data Challenge

Next-generation telescopes such as the Large Synoptic survey Telescope (LSST) and the Square Kilometer Array (SKA), along with the currently most advanced radio telescope, the Atacama Large Millimeter and sub-millimeter Array (ALMA), produce the largest "Big Data" in the early decades of the 21st century [1]. ALMA, in particular, opened an era of a new level of data complexity in radio and millimeter observations. Hyperspectral data cubes are becoming the norm, with simultaneously recorded, spatially resolved and co-registered images of a target in many different molecular lines. Each line is resolved by dozens to hundreds of spectral (velocity) channels providing detailed information about the kinematic behavior of multiple molecular species in objects such as protoplanetary disks, molecular clouds, interstellar medium, nearby galaxies, and more. Figure 1 shows spectral responses in two molecular lines, $C^{18}O$ J=3-2 and ^{13}CO J=3-2 each comprising 100 channels, which are concatenated (or stacked) along the spectral axis. This example is from a high-resolution ALMA data cube of the protoplanetary disk HD142527 observed by [2]. The lines have a width of about 6 MHz and are observed with a spectral resolution of about 65 KHz per channel (corresponding to 0.11 km/sec velocity resolution). Spectra from two different spatial locations are shown in Figure 1. The peaks within each molecular line appear shifted relative to the rest frequency (green vertical line), the frequency at which the observed source emits radiation at rest in the observer's reference frame. The rest frequencies of the $C^{18}O$ J=3-2 and ^{13}CO J=3-2 lines are 329.33055 GHz and 330.58797 GHz, respectively. The shift from the rest frequency indicates the relative velocity of the source, which can vary across spatial locations. The radiation intensity — the height of the peaks — is typically different for different molecules, as it depends on the temperature and density of the emitting gas. Furthermore, the relative intensity difference between species can also vary by spatial location as illustrated in Figure 1. The stacked spectral signatures capture the complex variations of the combined compositional and kinematic properties within a spatially resolved source. The challenge is to extract and map these variations, and to produce results efficiently enough for on-board analyses or for discovery in large archives.

A. Traditional Methods of Structure Discovery in Radio Astronomy

Before ALMA, the previous generation of radio telescopes had more limited spectral and spatial resolution due to smaller bandwidth and lower sensitivity. The resulting observations were relatively simple, containing a few spectral lines with a dozen or so channels each. In ALMA observations the number of spectral lines can be dozens to hundreds, at greatly increased velocity resolution, easily amounting to thousands of channels per image cube. This greatly increased spectral coverage and resolution offers a magnifying lens for our understanding of the kinematics of atomic and molecular gas, as well as of the distribution of solid particles. However, current techniques seem inadequate for visualization, analysis,

978-1-5090-4240-1/16/$31.00 ©2016 IEEE

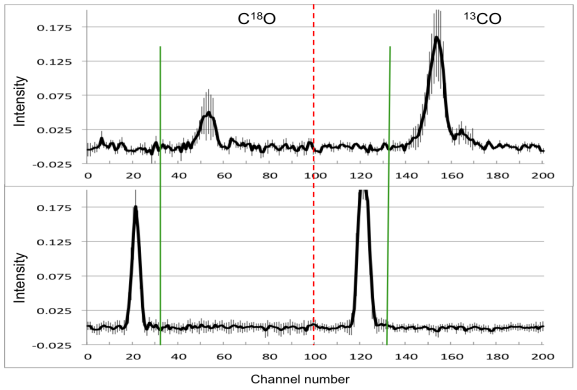


Fig. 1. Sample spectral signatures (means of spectra from small areas) from a data cube that comprises all channels of two molecular lines stacked in the order of increasing frequency. Each line is resolved in 100 channels (a total of 200 channels) in this example. The red dotted line shows where the two lines were concatenated. The green vertical lines indicate the channel of the rest frequency for each of the molecular species. The shift of the peaks relative to the rest frequency expresses a velocity difference — the Doppler shift — between observer and source. The Doppler shift is different in the two spectral samples, which are from two different spatial locaions of the protoplanetary disk HD142527. The radiation intensity (height of the peaks) as well as the relative difference in the peak intensities between the two species is also different at the two locations, adding to the complexity of the structure of gas across the disk. Data from [2]. The 200-element stacked spectra from individual pixels are the 200-D input vectors to our clusterings.

and interpretation of these enormous and complex data sets, which limits the full realization of the sensor potential.

Traditionally, two approaches have been used for the extraction of physical (velocity) structure from 3-D (3-dimensional) data cubes generated by radio and millimeter telescopes like ALMA. One is to visually inspect simultaneously displayed images of each spectral channel (a "channel map") within a spectral line. A second technique is to project, or integrate, the data along one dimension. A commonly adopted procedure is to integrate along the spectral dimension of a single line to calculate the so-called "intensity moments", at all pixel locations. The first three intensity moments correspond to the spectrally integrated intensity (moment 0), the velocity corresponding to the center of the line (moment 1), and the width of the line emission assumed to have Gaussian shape (moment 2). The moment 0 image maps the spatial distribution of the emitting gas, the moment 1 image informs about the motion of the gas on spatial scales larger than the spatial resolution of the observations, and the moment 2 informs about the motion of the gas on spatial scales smaller than the resolution of the observations. Alternatively, 3-D data cubes can be integrated along one spatial dimension. This leads to the so-called "position-velocity" diagrams, which are sensitive to kinematical properties of the gas such as inflows or outflows. Figure 2 shows the moment maps of two molecular lines of the protoplanetary disk HD142527, while Figure 3 shows the position-velocity diagram for the ^{13}CO line. This object is particularly interesting because the bright red and yellow ring in the moment 0 maps is thought to have been formed by a newborn planetary system inside the dark blue center of about 100 AU diameter, comparable to that of our Solar System [3].

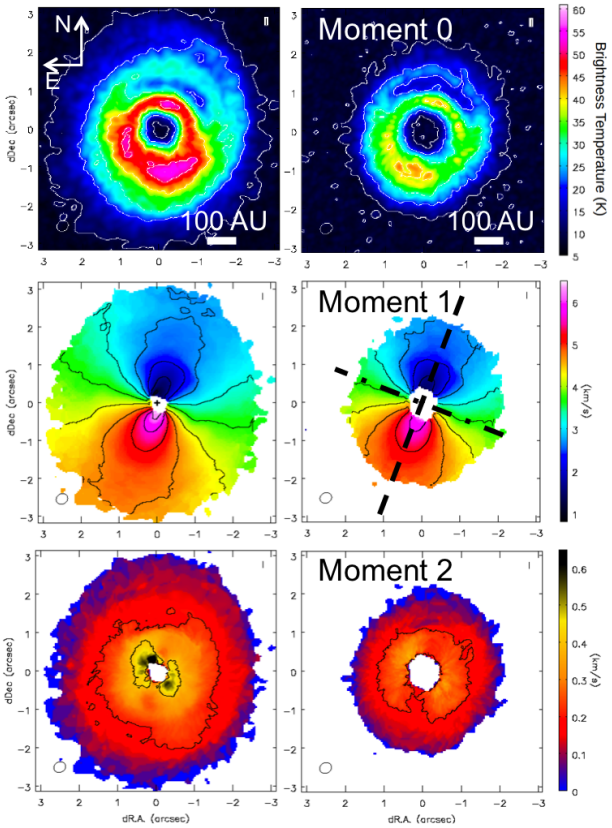


Fig. 2. From [2]. Traditional moment maps generated separately from two molecular lines, ^{13}CO J=3-2 at left, and C^{18}O J=3-2 at right, from high-resolution observations of the protoplanetary disk HD142527. The Moment 0 map shows the intensity of the line emission, expressed in units of brightness temperature defined as the temperature of a black body with an energy density equal to that observed in the emission line. The brightness temperature corresponds to the gas kinetic temperature if the line emission is optically thick, as in the case of HD142527. The moment 1 map shows the velocity of the emitting gas relative to the observer. The dot-dashed line indicates the rotation axis of the disk, and the dashed line shows the apparent major axis of the disk. The moment 2 map shows the width of the line emission.

For precursors of ALMA, moment maps and position-velocity diagrams were suitable tools to visualize in two dimensions the main characteristics of the molecular line emission, at least for simple objects such as protoplanetary disks. However, these traditional methods do not scale up well to the much more complex ALMA cubes and, in particular, they might lead to erroneous conclusions if the gas kinematics along a line of sight cannot be represented as a Gaussian line, as in the case of gas moving at different velocities. At the same time, with potentially dozens of spectral lines consisting of several thousand channels, visual analysis of channel maps becomes unfeasible, especially comparing kinematics across multiple spectral lines of many species.

B. Recent Structure Discovery Methods in Radio Astronomy

In recent years, a number of procedures were developed to overcome the limitations of the moment analysis. They include Clumpfind [4], Cloudprops [5], dendogram object identification [6], [7], and Discrete Persistent Structures Ex-

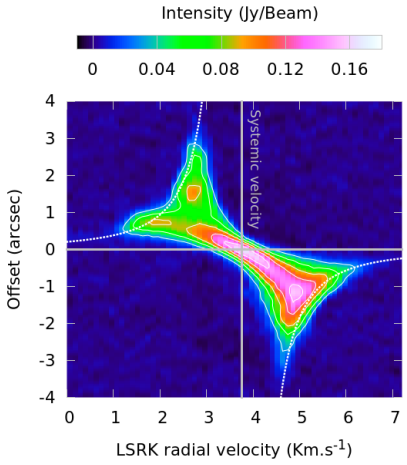


Fig. 3. From [2]. Position-Velocity diagram for the ^{13}CO J=3-2 line emission measured toward HD142527. The x-axis shows the offset with respect to the center of the disk measured along the major axis of the disk. The y-axis shows the velocity along the line of sight of the emitting gas relative to the observer. The white dotted curves indicate the expected velocity for gas rotating at Keplerian velocity around the central star. The vertical line shows the systemic velocity, i.e., the velocity along the line of sight of the star+disk system relative to the observer.

tractor (DisPerSE), [8]. Some of these methods rely on fitting 2-D or 3-D Gaussian distributions (e.g., Clumpfind and its kin) thus limiting the discovery to simple structures. Others are well suited for identification of particular structures (e.g., filamentary structures in the case of DisPerSE) but do not generalize to different ones. Or, some methods require the noise across the image to be uniform and well characterized (as for the dendogram analysis).

In this paper we aim to demonstrate a clustering approach with *NeuroScope*, a set of neural machine learning tools whetted in other domains, for the identification and visualization of spatial regions with different/distinct patterns of motion and display them in one integrated map. This approach also allows characterization of the complex kinematics of multiple species from their combined (stacked) spectral lines. The emerging details indicate fuller exploitation of the rich ALMA data than by the traditional methods. The NeuroScope approach is insensitive to most of the limitations discussed above and it is therefore a more robust and flexible tool of analysis of ALMA data. We discuss some specific complementarities between the above leading methods and NeuroScope after a brief technical background below.

II. Structure Discovery with NeuroScope Tools

Our approach is based on clustering and pattern recognition, therefore it returns regions of any shape that exhibit coherent behavior in terms of the combined spectral signatures.

A. Cluster Discovery

Clustering with NeuroScope tools involves manifold learning with particular variants of Self-Organizing Maps (SOMs) and a recent similarity measure that facilitates interpretation of the SOM's knowledge based on the *connectivity* properties of the data manifold. As we explain below, this combination allows deeper exploitation of relevant details than customary uses of SOMs and common similarity metrics (e.g., Euclidean distance or spectral angle), which in turn allows sensitive distinction of clusters with subtle but consistent spectral properties. We briefly review these tools below.

The Self-Organizing Map (SOM) [9] is a popular unsupervised neural network learning agorithm for discovery of clusters (groups of similar patterns, e.g., similar spectra). It mimics the information summarization and organization that takes place in various cortical areas of natural brains (e.g., visual, auditory, somatosensory cortex). A SOM consists of a rigid (usually 2-D) lattice of artificial neurons each of which is connected to an input layer by an n-D weight vector, also called a *prototype* vector. The prototypes are typically initialized to random values. The SOM learns the structure of the n-D input data space by cycling thorough the following steps many times: i) a randomly selected input vector is compared to the prototypes (weight vectors) of all SOM neurons, and the neuron with the most similar prototype "wins" the input. Then the winner neuron and — usually to a lesser extent — neurons in its lattice neighborhood adapt their prototypes to become more similar to the presented input vector. The neighborhood can be defined by various functions. Most frequently a 2-D Gaussian neighborhood is used, which means that the prototypes of all neurons are updated in every learning step but the extent of the update decreases with the lattice distance from the winner. This iterative learning process accomplishes two things. One is adaptive vector quantization of the input data: the weight vectors of SOM neurons become prototypes of similar input patterns. SOM learning moves the prototypes in data space such that they follow the data distribution (more prototypes are allocated to dense regions than to sparse regions), therefore the data summarization by the SOM prototypes captures the salient details of the data distribution. Simultaneously, a topologically ordered map of the prototypes is formed on the SOM lattice. As a result, neurons neighboring in the lattice collectively represent groups of similar data vectors after sufficient learning. The SOM expresses an intelligent summarization of both the statistics (the n-D density distribution) and the topology of the data manifold. The learning does not require a pre-specified number of clusters. Clusters can be extracted from a learned SOM by evaluating the similarity relationships of prototypes neighboring in the SOM grid, and segmenting the SOM into groups of similar prototypes. For capturing complex cluster structure we use a recently developed similarity measure, CONN (Figure 4), which is derived from the converged SOM and expresses manifold connectivity rather than data space distances [10], [11]. Using these tools, the main steps of finding clusters are

A) Learn the data manifold with a SOM. This is automatic, and needs little parameter tuning. We use specific advanced variants of SOMs for our goals as noted below.

B) Represent the knowledge of the SOM through its connectivity graph (CONN), which is also automatic. Default parameters for filtering unimportant connections (inconsequential

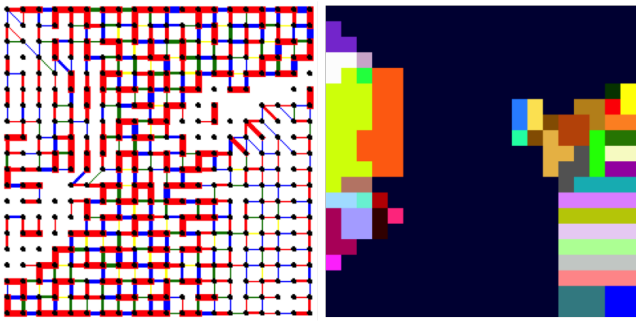


Fig. 4. Left: SOM lattice of 20 x 20 neurons (black dots), with CONN graph representation of the learned manifold structure of the protoplanetary disk HD142527 using 200-D input vectors stacked from two molecular lines as in Figure 1. A cell with no dot has no data vectors mapped to it (has an empty prototype). The thickness of the line segments between two prototypes signifies the absolute strength of their connection. Connection strength is measured, during a full recall on the data set after the SOM has converged, as the number of data vectors that choose one prototype as the SOM winner and the other as second winner. Colors indicate the relative importance of the connections to other prototypes. Red is most-connected, blue is second most-connected, followed by green, yellow, and grey shades (not present for this data). The combination of global connection strengths and their local ranking provides rich information about where the manifold is strongly woven and where it is disconnected or thin. Cluster boundaries are found between regions that are strongly connected inside and have thin or no connections to other regions. A threshold to automatically cut "unimportant connections is also computed from CONN statistics and applied. For visualization, a non-linear binning is applied to aid the human eye where the bin boundaries are derived from the connectivity statistics. Details on this and the cluster extraction procedure are given in [10]. Right: SOM lattice with clusters of similar prototypes extracted interactively from the CONN graph representation. Each color represents a different similarity group of prototypes. The largest and also strongest-connected cluster (dark blue) comprises the sky background. While visually overwhelming in the CONN graph at left, it does not affect the extraction of smaller clusters with more subtle differences, by analysis of the local relations.

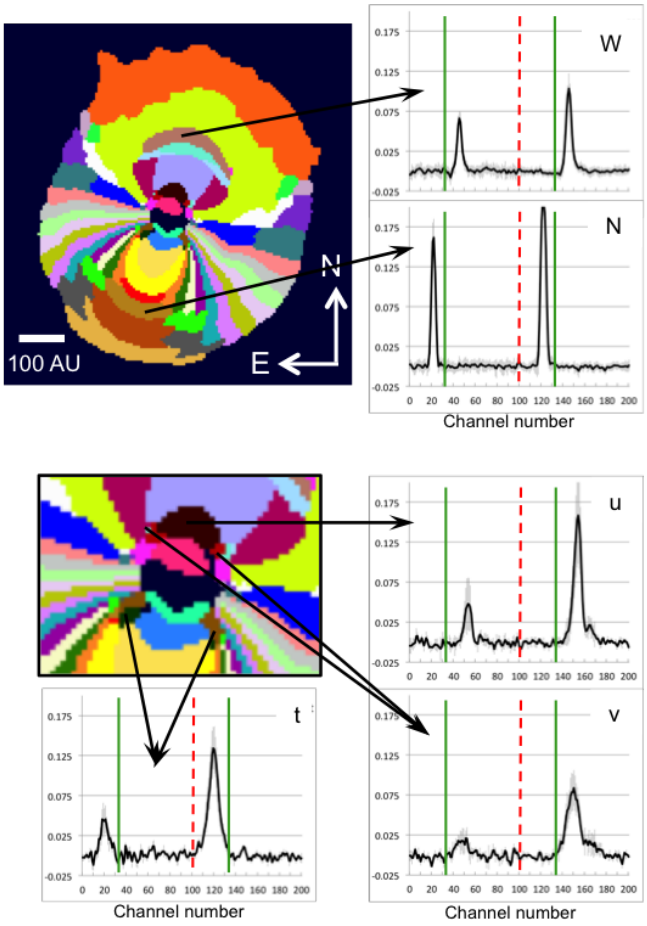


Fig. 5. Top panel: The clusters shown in the disk of HD142527, where each pixel is colored as its prototype in the SOM lattice in Figure 4. The colors are chosen for contrast and do not express relative grades of similarity. (This is not a heat map.) The mean spectra of two selected clusters (labeled W and N, two different brown shades in the map) at symmetric locations along the major (NW-SE) axis are also shown at right. These illustrate an example of the lack of expected symmetry captured by this clustering. Red and green lines as in Figure 1. Bottom panel: The magnified center region surrounded by the mean signatures of three interesting clusters. Cluster u, and the tiny whisker-like clusters v and t flanking the arch-shaped clusters g (hot pink, North of the center) and j (turquoise, South of the center), respectively, exhibit double or widened peaks in the ^{13}CO line (approx. between channels 140 - 175 for u and v; channels 101 - 150 for t). These can indicate deviation from the Keplerian motion.

for clustering) are automatically computed based on the data statistics, as described in [10], [12].

C) Segment the SOM, i.e., cluster the prototypes based on their similarities. (Data points mapped to a prototype cluster make up a data cluster.) This is the most challenging step for data with complex structure.

Figure 4 illustrates steps B) and C). Figure 5 shows the cluster structure found in the protoplanetary disk HD142527, using the combined (stacked) molecular lines C^{18}O and ^{13}CO (as in Figure 1) as input data vectors. The data have undergone the standard ALMA data reduction to correct for atmospheric and instrumental effects [2] but no other preprocessing. The relatively simple kinematical structure of this object enables us to validate our technique and appreciate its potential for radio astronomy data.

The first thing to notice is that the cluster map shows structures similar to that shown in the moment 1 map. In particular, the clusters reveal regions of the disk moving at similar velocities. Since the gas kinematics is dominated by rotation around the star, the cluster map is roughly symmetric with respect to the axis of rotation of the disk (the dot-dashed line on the moment 1 image). Looking more carefully, clusters at two symmetric positions with respect to the minor axis (see, e.g., those indicated by the arrows in the top panel of Figure 5) show an asymmetry in intensities, which agrees with the intensity differences in the moment 0 map. The cluster means additionally reveal that the relative intensity change for the two gas species is also different at the two symmetric locations.

The neural map-based clustering is superior to the moments visualization in identifying kinematical deviations from Keplerian rotation. For example, Figure 5 shows clusters (u, and t) of pixels characterized by a double peak profile, or a widened peak (cluster v) in the ^{13}CO line. The main peak in cluster u (at approximately channel 153) arises from gas moving at a velocity of 2.2 km/sec relative to the star in the direction of the observer, and traces gas orbiting the central star at Keplerian velocity. The second, minor peak (at channel 165) arises from gas moving at 3.5 km/sec relative to the star in

the direction of the observer, and might arise from gas blown away by a stellar wind. Although low-intensity, the second peak emerges cleanly in the average cluster signature with very small standard deviations. Similarly for cluster t (the pair of small downward pointing whisker-like dark green features marked by arrows at either sides of the arch-shaped turquoise cluster), where the main and minor peaks (at channels 120 and 148, respectively) are more distinct and have larger velocity difference (-1.43 km/sec vs 1.7 km/sec) than in cluster u. Here the two velocities have opposite signs, which may be caused by two gas components moving in opposite directions. While the moment 2 map indicates a widening of the line at the location of clusters u and v, it does not provide information of the shape of the line. Furthermore, the moment 2 fails to highlight the location of cluster t. This example shows how NeuroScope clustering can enlarge the discovery space by fully exploiting the transformational imaging capabilities provided by current telescopes.

There are important advantages of clustering analysis compared to the traditional moment visualization. First, the capability of combining multiple lines at once allows one to identify correlation and anticorrelation between different gas tracers. Second, clustering analysis naturally combines signals from similar regions augmenting the capabilities to see faint structures. Third, the proposed — data-driven — technique does not assume priors for the line emission and therefore deliver an unbiased interpretation of the observations.

For successful clustering and interpretation of the data, the correctness of the manifold learning (the placement of the SOM prototypes according to the data distribution, and the preservation of the data topology in the SOM lattice), and the quality of cluster extraction from a learned SOM are of crucial importance. This can be challenging to ascertain for complicated large data such as noisy spectral imagery with many inherent clusters. Many of these issues have been addressed by [12], [11], and references therein. This includes measures of topology preservation, to separate serious topology violations from those that are inconsequential for cluster detection and can be ignored.

In NeuroScope we use Conscience Self-Organizing Maps (CSOM; [13]), which achieve better data density matching (more faithful representation of the data distribution) than the original Kohonen SOM. For high-dimensional input data this was verified by [14]. The CSOM, at the same time, is computationally less expensive because it only has to update the immediate lattice neighbors in each learning step, in contrast to updating all SOM prototypes in the Kohonen SOM.

SOM learning is robust, partly because the summarization of data by the prototypes greatly reduces noise, and partly because the topological ordering of the prototypes facilitates the preservation of subtle differences despite possibly small vector distances between spectra. SOM learning is not directly limited by the dimensionality of the data. We can use the full spectral information (all channels) as input to clustering. There is no need for prior noise reduction, dimensionality reduction, or prior assessment of statistical properties of emission lines. We can use the spectral cubes coming straight out of the ALMA data reduction pipeline (like in this study) without needing additional preprocessing.

B. Relationship to the State-of-the-Art

The dendogram object classification has similarities with NeuroScope clustering in that both cluster the spectral signatures. The original application of the dendrogram method (to structure finding in molecular clouds) in 2-D provides the lineage of spectral relations (velocity) but loses the size and shape information of the component clouds [6]. The subsequent extension to a 3-D dendrogram method recovers the size and shape information by displaying the iso-surfaces, identified by the dendrogram, on the channel map that is closest to the velocity of the particular cloud component [7]. This, however, can be challenging to synthesize and interpret for a large number of channels. More importantly, since the dendogram analysis relies on identifying iso-surfaces, it is not well suited to low signal-to-noise data and it does not naturally extend to the simultaneous analysis of data cubes containing multiple lines.

NeuroScope clustering produces a map of distinct homogeneous source regions in terms of spectral properties. For example, three different cloud components traced by the same molecule and having different velocities would appear as three distinct clusters (differently colored regions) on a single 2-D view of the cloud system, outlining the shape and size of the sources while indicating that they differ in velocity. This result is similar to that delivered by dendogram analysis. However, NeuroScope can naturally use multiple molecular tracers by stacking the spectral signatures, in which case the cluster map would show distinct homogeneous regions based on both velocity and composition as in in Figure 5. Furthermore, differently from the dendrogram algorithm, NeuroScope clustering does not require any prior knowledge of the noise and it does not require specification of intensity thresholds for the identification of clusters. The integrated yet detailed view provided by the NeuroScope cluster map facilitates (at least alerts to) discovery at-a-glance.

Clustering with NeuroScope tools can make a significant difference for structure discovery in high-dimensional, complex data. While comparison with other clustering methods is outside the scope of this paper, we point to a comparison with K-means clustering for multi- and hyperspectral data containing many clusters in [15]. For the relatively low-dimensional and less complex multi-spectral data K-means produces reasonable clustering. However, for the hyperpsectral data, which also discriminates more clusters, the study shows confusion and missing of many relevant material clusters by K-means while the SOM-based approach delineates the known materials with an accuracy and sensitivity that allows direct match with field spectra for material identification.

C. Efficiency for Large Data

SOM learning is efficient for large data sets because it provides intelligent summarization, shrinking the data volume

by a large factor spatially while retaining the manifold characteristics relevant for cluster discovery and keeping the spectral resolution in the prototypes. As an example, the protoplanetary disk and immediate surroundings in Figure 5 comprise approx. 56,000 pixels. These are represented and characterized by 400 SOM prototypes (a factor of 140 down-sizing) without loss of relevant knowledge of the compositional and structural distinctions. We can further compress the data for certain uses by only storing the cluster means and a few additional vectors such as the minimum, maximum, and standard deviation across the pixels in a cluster. This is a significant advantage for large data sets. SOMs are inherently parallel algorithms. Parallel (FPGA, ASIC, GPU) implementation of the learning process can provide magnitudes faster speed than sequential implementations. A specialized large-scale SOM accelerator capable of handling hyperspectral data cubes has been tested and is being considered for use in NeuroScope [16].

III. Automation Approach

While SOM-based clustering using our tools can result in excellent structure discovery, full automation is needed for processing vast archives or for near-real time on-board analyses. However, SOM segmentation (step C in Section II.A) is currently generally more successful when done interactively, from expressive visualizations of the SOM's knowledge such as in Figure 4. Existing automated methods work well for some type of data and low cluster complexity while underperforming for other types of data. For example, our function SOMcluster (under development) produced clean cluster maps for Mars Exploration Rover imagery (similar to that shown in [17]) but poor segmentation of SOMs of functional Magnetic Resonance Images for brain mapping [18]. The reasons include the substantial difference in the general characteristics of these two types of data including the noise, the great difference in the number of inherent clusters (approx. 20 vs 70) and the level of the average similarity (variance) of the data vectors. Such differences are expected between different domains, especially for large and complex data. Characterization of the data and using that to inform tools can help achieve domain-specific acuity for the automation of SOM segmentation.

Here we explore an approach based on informing leading graph segmentation algorithms with the SOM's knowledge of the data. Graph segmentation algorithms have been proposed for automated clustering of "Big Data", but their use for large data sets requires huge computational resources since they start with a graph with N^2 edges for N data points (representing the pairwise distances of all data points). Instead, we give the SOM prototypes to the graph segmentation algorithms (approximately $\sqrt{N}$ data vectors, generating N graph edges), and provide the CONN similarity values as "distances" for the algorithms. We show that this combination significantly enhances the graph segmentation algorithms' ability to find relevant clusters, and the computing time is very small. We compare the resulting clusterings of the HD142527 data cube with the cluster map in Figure 5.

A. Graph Segmentation Overview

Graph segmentation (also called graph partitioning, community detection or graph clustering) aims to identify a subgraph structure such that each subgraph is densely connected within itself and sparsely connected to other subgraphs, where the concept of vertex *connectivity* (binary) is replaced by vertex *similarity* (graded) for weighted graphs. Decades of research in this area have resulted in many different classes of algorithms. The so-called *cut* methods (mininum cut, normalized cut, ratio cut) attempt a recursive bipartitioning (a cut) of the graph driven by minimizing an objective function which represents the similarity between the two resulting sub-clusters. *Spectral* methods aim to approximately minimize this same cost function via eigenspace decompositions of the graph Laplacian matrix. *Modularity* based methods maximize the "quality" of a partition as measured by the modularity function, which compares a given graph partition to a distribution of random graphs with the same structure (number of edges and vertex degrees), favoring cluster structure that occurs from more than "chance" alone. Many other partitioning schemes have been adapted from other domains such as agglomerative and partitive data clustering from statistics, Markov chain formulations from probability theory and equilibrium models from statistical physics. See [19] for a more complete overview.

We experimented with a number of leading algorithms with mixed results. The greedy agglomerative modularity algorithm of [20], the Multilevel algorithm of [21] and the eigenspace modularity approximation of [22] produced clusterings of generally poor quality while the `Walktrap` [23] and `Infomap` [24] algorithms identified more meaningful cluster structure of the ALMA data. We assess an algorithm's suitability for automated cluster detection by comparing its clustering with default parameters to the interactive clustering in Figure 5 based on a) visual inspection; b) the percentage of matching pixels; c) cluster size distribution; and d) the Jaccard similarity coefficient (JSC) and the Jensen-Shannon divergence. The JSC between two clusterings C_1 and C_2 is the proportion of pixel pairs which are assigned to the same cluster in both C_1 and C_2. Thus $JSC = 1$ indicates complete agreement between clusterings while $JSC = 0$ indicates complete discord. The Jensen-Shannon (J-S) divergence measures the similarity between two distributions relative to their mean (and consequently is a symmetrized version of the Kullback-Leibler divergence). All algorithms we experimented with require few (1 or 2) parameters, which is important for automation. Both highlighted methods are freely distributed in the `igraph` package [25] and require negligible runtimes ($\approx$ 1 second on an ordinary MacBook Pro for the SOM-based graphs described in section III-B) in our experiments.

The `Walktrap` algorithm utilizes the method of *random walks on graphs* to create a unique distance measure that is then used in conjunction with Ward's classical agglomerative clustering scheme. In this case, the random walk on the graph is represented by a Markov chain whose state space is initially equal to the set of all vertices and transition probability matrix

P is equal to the matrix of vertex similarities (edge weights) with row-sums normalized to 1. Thus the probability of transitioning from vertex i to vertex j in a walk of length t is P^t_{ij}. If vertex i and vertex j belong to the same cluster we expect P^t_{ij} to be relatively high, at least for short walks (with the converse also expected). Note, however, that this formulation alone does not produce a clustering of the vertices. To accomplish this, a probability based distance between vertices i and j is defined from P as $d_{ij} = \sqrt{\sum_{k=1}^{N} \frac{(P^t_{ik} - P^t_{jk})^2}{\deg(k)}}$ where N is the total number of vertices in the graph and $\deg(k)$ gives the degree (sum of edge weights with one endpoint in vertex k) of its argument. Since the rows of P^t specify probability distributions, this can be thought of as the "L^2" distance between the distributions describing the movement from vertex i and vertex j. Put simply, if a walk transiting from i has roughly the same distribution of destination states as a walk transiting from j, vertices i and j are considered similar in terms of this metric. Ward's algorithm [26] is then used to choose two vertices to merge into one cluster. P is recalculated to represent the Markov chain on this reduced state space, with transition probabilities between the merged community and other vertices inherited from its members. This process is repeated $N - 1$ times, with the terminal result the final partitioning of the graph. The default walk length parameter $t = 4$ was used to obtain our results.

Information theory guides the clustering of the `Infomap` algorithm, which is also based on random walks on graphs. Instead of directly analyzing transition probabilities, P (same as `Walktrap`) motivates an entropy-based cost function L which, given any partitioning C of the graph into c clusters, describes the total entropy of movement both between and within clusters: $L(C) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{r=1}^{c} p^r_{\circlearrowright} H(\mathcal{P}^r)$, where $q_{\curvearrowright}$ is the probability of moving between clusters and $p^r_{\circlearrowright}$ is the probability of movement within cluster r, $H(\mathcal{Q})$ is the entropy of between-cluster movement and $H(\mathcal{P}^r)$ is the entropy of movement within cluster r. These quantities can all be derived directly from the entries of P. Since minimum entropy corresponds to the most information about a stochastic system, the goal is thus to *minimize* L with respect to C. Direct minimization of L is in most cases computationally intractable so a greedy agglomerative partitioning scheme is devised. Initially, each vertex comprises its own cluster. A cluster is picked at random and merged with its neighbor (i.e., a cluster with which it shares an edge) to produce the largest overall decrease in L. After all clusters have been considered, a new graph is constructed comprising only one vertex per merged cluster (and, as in `Walktrap`, parent clusters inherit their transition probabilities from their children). The process begins anew and is repeated until no further decrease in L is possible. The final clustering, an approximate minimizer of L, is then the one describing the structure inherent to the graph with most information. Since neighbor evaluation is initiated randomly, the algorithm likely produces a local minimizer of L. To account for this, the entire process is repeated $num.trials$ times (which is the only parameter required) and an approximate global minimizer of L is selected from the $num.trials$ candidates. We used $num.trials$ = 10.

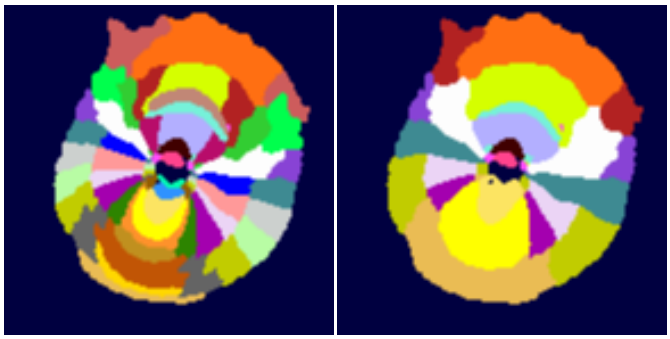
B. The SOM as a Graph

Positing the SOM as a graph is natural and straightforward but a distinction must be made between the SOM output space, which is traditionally visualized on a two-dimensional lattice, and the weighted graphs we derive as inputs to graph segmentation algorithms. In both, vertices represent data prototype vectors but the latter is not a lattice; its connectivity structure is dictated by the function we choose to represent similarities between prototypes. For comparison, we consider both the CONN values and the (inverse) Euclidean distance between prototype vectors (IEDP) as similarity measures. Since CONN is an asymmetric measure (unlike Euclidean distance), it additionally admits the formulation of directed graphs. (This is not revealed in Figure 4, see [10] for details.) Both directed and undirected variations of the CONN-based graph are considered where appropriate.

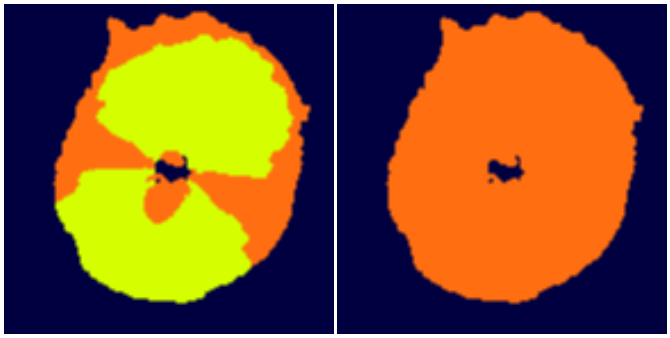
Graphs based on prototype representations of the data, as opposed to graphs based on individual data points (in our case, the spectral signatures of the pixels of the HD142527 data cube), offer three benefits for automated segmentation algorithms. From a computational standpoint, it is intractable to compute and store the similarity measure between all pixel pairs of a large cube emanating from real world spectral imaging. For example, our focus area of the HD142527 data cube is 236×236 pixels (requiring $\approx$ 56,000 vertices and on the order of 10^9 edge weights if using a symmetric similarity measure such as Euclidean distance). Storage for even this small region exceeded the capacity of a 3GHz dual core processor with 16GB of RAM. Edge sparsity, if available, can lower this demand somewhat but that typically requires some *a priori* knowledge or preprocessing of the data. In contrast, processing the graph of the 400 SOM prototypes in Figure 4 takes less than 1 sec. Practical constraints aside, prototype-based neural learning schemes have the added benefit of noise reduction, boosting the signal-to-noise ratio in their representation of the underlying data distribution. Most importantly, such schemes permit the introduction of new similarity measures (i.e., CONN) which are unavailable in the data domain itself. Of course, individual similarity measures tailored to a particular task may be available through consultation with domain experts, but then we are no longer in the realm of automated data analysis. We outline below the marked benefit of the CONN similarity measure over traditional choices.

C. Results

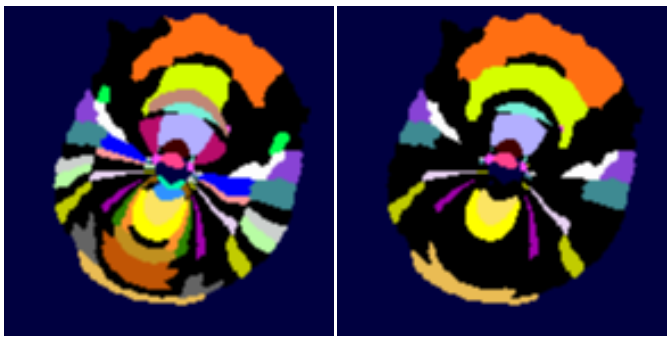
Figure 6 displays the clusterings by the `Walktrap` (6a,6c) and `Infomap` (6b,6d) algorithms applied to graph representations of the SOM prototypes. We note first that all algorithms we considered perform extremely poorly when inverse Euclidean distance (IEDP) is used as a similarity (i.e., edge weight) between graph vertices. This is evident for the two algorithms highlighted here when comparing 6c and



(a) Walktrap-CONN (b) Infomap-CONN



(c) Walktrap-IEDP (d) Infomap-IEDP



(e) Walktrap-CONN (Agreement) (f) Infomap-CONN (Agreement)

Fig. 6. Automated clusterings from the SOM prototypes of the HD142527 data cube by the (column 1) Walktrap and (column 2) Infomap algorithms using CONN values (row 1) and the inverse Euclidean distance (IEDP, row 2) between prototypes as similarity measure. The agreement images in row 3 are relative to Figure 5, with mismatching pixels masked in black.

6d to Figure 5, where little more than the general shape of the protoplanetary disk is identified. The CONN similarity measure, on the other hand, helps these algorithms discern clusterings (6a,6b) which are much more similar to Figure 5. One possible explanation as to why is the edge sparsity induced by CONN when compared to IEDP. To control for sparsity effects, we allowed each algorithm to attempt a clustering of a graph whose edge weights are specified by IEDP, but edge existence is specified by CONN. These results (not shown) generally produce only very modest improvement, indicating CONN's superiority in this case is due to more than sparsity. The rest of the discussion will focus on CONN-based clustering results.

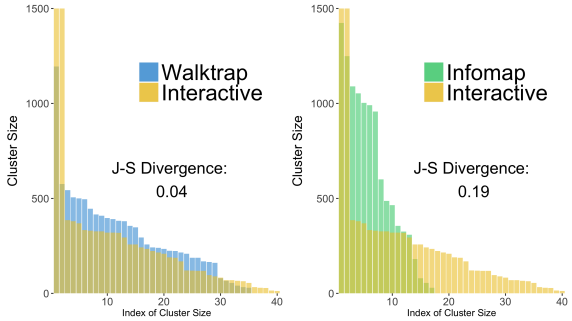


Fig. 7. Distribution of cluster sizes by segmentation method

Statistical assessments also support the visual comparisons. First, Walktrap-CONN clustering matches 60% and Infomap matches 44% of the pixels in Figure 5 (approximately). Second, Walktrap-CONN produces more clusters, and the clusters they both produce are generally tighter (have smaller envelope), with smaller standard deviation in the Walktrap-CONN clustering than in that from Infomap-CONN. The envelope and standard deviation of the Walktrap-CONN clusters are close to those of the reference clusters (not shown here). Measured by the Jaccard index, the clustering produced by Walktrap ($JSC = .29$) shows greater fidelity to Figure 5 than does the clustering produced by Infomap ($JSC = 0.25$). Indeed, the distribution of cluster sizes (i.e., number of pixels in each cluster) from the Walktrap clustering more closely matches that of Figure 5 as seen in the left panel of Figure 7. The Jensen-Shannon (J-S) divergence of the Infomap and Walktrap cluster size distributions are 0.19 and 0.04, respectively, confirming that, of the two, Walktrap produces a clustering with distributional characteristics more faithful to Figure 5.

The agreement images from Walktrap and Infomap (Figures 6e and 6f, respectively) facilitate a visual assessment of the cluster structures relative to Figure 5 by showing only the regions where each clustering algorithm agrees with Figure 5. Walktrap appears as the most similar of the two, especially in the center of the image. In fact, the small, and perhaps most interesting kinematic details we discussed in Figure 5 are matched remarkably by Walktrap-CONN, whereas Infomap-CONN misses most of that fine structure. Both algorithms are able to discern the radial structure emanating from the center, but Infomap does so with less granularity. Walktrap, on the other hand, appears to favor more sub-clustering of the radial arcs. This can be explained by the fact that (as also seen from the moment 0 maps) the gas intensity drops at the radius where the "breaks" are seen in the automated clustering results. These algorithms are apparently more influenced by the intensities than by the velocity differences.

Despite their differences, the Walktrap and Infomap segmentations generally capture at least the high level structure in Figure 5 fairly well, and we stress again that we used off-the-shelf default parameters with no further tuning. The quality

of clusters can be evaluated more rigorously — with and without a reference — and this is planned for a more involved study, but the simple comparison above provides promising initial results from these automated clustering procedures. In our experiments, however, this success is wholly dependent on the combination of a prototype-based representation of the data and the CONN similarity measure.

IV. Discussion, Conclusion and Outlook

We demonstrated on a relatively simple astronomical object the advantages of neural map-based clustering over traditional moment maps for finding structure. Our tools can highlight regions of distinct combinations of kinematics and gas densities for multiple molecular species in a single integrated map, alerting to potential discoveries. More exhaustive validation and interpretation of our clustering results in Figure 5 is an ongoing effort and will be presented in a follow-up paper. We showed that using the learned SOM prototypes and CONN similarity measure (manifold connectivity derived from the learned SOM) as input to leading graph segmentation algorithms `Walktrap` and `Infomap` we can dramatically enhance their ability to produce automated cluster extraction results comparable to our interactive clustering in negligible processing time, while relying on their default parameters with no further tuning. While these clusterings are not yet good enough to replace our interactive process, we expect to improve the results from `Walktrap` and `Infomap` in subsequent work by studying their parameters and exploiting the directional edge weighting afforded by CONN. The goal is to achieve full automation with the same or very close quality to that of the interactive processing. Follow-up work will also target more complex astronomical objects (such as molecular clouds) where the kinematics are not dominated by a regular motion, and signatures of multiple sources may be superimposed. This will be the next level of challenge for our tools. In conclusion and based on the study we presented, we anticipate that clustering methods that can deal with the complexity and richness of the data involved in next-generation radio astronomy observations, as well as keep up with the data volume, will play an essential role in uncovering the intricate processes of the universe.

Acknowledgment

A. I. acknowledges support from the NASA Origins of Solar Systems Program, through award NNX15AB06G. We thank two anonymous reviewers for insightful comments.

References

[1] P. Estevez, "Big data era challenges and opportunities in astronomy: How SOM/LVQ and related learning methods can contribute?" 2016, invited talk, posted at http://wsom2016.rice.edu/files/2016/01/WSOM-2016PE2-1cqipnv.pdf.

[2] Y. Boehler, A. Isella, E. Weaver, C. Grady, J. Carpenter, L. Perez, and L. Ricci, "A close-up view of the horseshoe disk HD 142527," 2016, submitted.

[3] A. Isella, L. M. Pérez, J. M. Carpenter, L. Ricci, S. Andrews, and K. Rosenfeld, "An Azimuthal Asymmetry in the LkHα 330 Disk," *Astrophys. J.*, vol. 775, p. 30, Sep. 2013.

[4] J. Williams, E. de Geus, and L. Blitz, "Determining structure in molecular cloud," *The Astrophysical Journal*, vol. 428, p. 693, 1994.

[5] E. Rosolowsky and A. Leroy, "Bias-free measurement of giant molecular cloud properties," *Publications of the Astronomical Society of the Pacific*, vol. 118, pp. 590–610, April 2006.

[6] P. Houlahan and J. Scalo, "Recognition and characterization of hierarchical interstellar structure. ii. structure three statistics," *The Astrophysical Journal*, vol. 393, pp. 171–187, July 1992.

[7] E. Rosolowsky, J. Pineda, J. Kaufmann, and A. Goodman, "Structural analysis of molecular clouds: Dendrograms," *The Astrophysical Journal*, vol. 678, pp. 1338–1351, June 1 2008.

[8] T. Sousbie, "Disperse: robust structure identification in 2d and 3d," *arXiv*, no. 1302.6221v1, February 2013.

[9] T. Kohonen, *Self-Organization and Associative Memory*. New York: Springer-Verlag, 1988.

[10] K. Tasdemir and E. Merényi, "Exploiting data topology in visualization and clustering of Self-Organizing Maps," *IEEE Trans. on Neural Networks*, vol. 20, no. 4, pp. 549–562, 2009.

[11] ——, "A validity index for prototype based clustering of data sets with complex structures," *IEEE Trans. Systems, Man and Cybernetics, Part B*, vol. 41, no. 4, pp. 1039–1053, August 2011, dOI: 10.1109/TSMCB.2010.2104319.

[12] E. Merényi, K. Tasdemir, and L. Zhang, "Learning highly structured manifolds: Harnessing the power of SOMs," in *Similarity-Based Clustering*, ser. Lecture Notes in Computer Science, M. Biehl, B. Hammer, M. Verleysen, and T. Villmann, Eds. Berlin Heidelberg: Springer Verlag, 2009, vol. 5400, pp. 138–168.

[13] D. DeSieno, "Adding a conscience to competitive learning," in *Proc. IEEE Int'l Conference on Neural Networks (ICNN), July 1988*, vol. I, New York, 1988, pp. I–117–124.

[14] E. Merényi, A. Jain, and T. Villmann, "Explicit magnification control of self-organizing maps for "forbidden" data," *IEEE Trans. on Neural Networks*, vol. 18, no. 3, pp. 786–797, May 2007.

[15] E. Merényi, B. Csató, and K. Taşdemir, "Knowledge discovery in urban environments from fused multi-dimensional imagery," in *Proc. IEEE GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas (URBAN 2007).*, P. Gamba and M. Crawford, Eds., Paris, France, 11–13 April 2007, pp. 1–13.

[16] J. Lachmair, E. Merényi, M. Porrmann, and U. Rückert, "A reconfigurable neuroprocessor for self-organizing feature maps," *Neurocomputing*, vol. 112, pp. 189–199, 2013.

[17] K. Tasdemir and E. Merényi, "Cluster analysis in remote sensing spectral imagery through graph representation and advanced SOM visualization," in *Proc. 11th Intl Conf. on Discovery Science (DS-2008)*, ser. Lecture Notes in Computer Science, vol. LNCS 5255/2008. Budapest, Hungary: Springer, October 13–16 2008, pp. 259–271.

[18] P. O'Driscoll, E. Merényi, C. Karmonik, and R. Grossman, "The effect of SOM size and similarity measure on identification of functional and anatomical regions in fMRI data," in *Advances in Self-Organizing Maps adn Learning Vector Quantization*, ser. Proc. 11th International Workshop WSOM 2016, E. Merényi, M. Mendenhall, and P. O'Driscoll, Eds. Springer, January 6–8, pp. 251–263.

[19] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75 – 174, 2010.

[20] A. Clauset, M. Newman, and C. Moore, "Finding community structure in very large networks," *pre-print*, vol. 70, no. 6, p. 066111, Dec. 2004.

[21] V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 10, p. 10008, Oct. 2008.

[22] M. Newman, "Finding community structure in networks using the eigenvectors of matrices," *pre*, vol. 74, no. 3, p. 036104, Sep. 2006.

[23] P. Pons and M. Latapy, "Computing communities in large networks using random walks (long version)," *ArXiv Physics e-prints*, Dec. 2005.

[24] M. Rosvall and C. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Science*, vol. 105, pp. 1118–1123, Jan. 2008.

[25] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal*, vol. Complex Systems, p. 1695, 2006. [Online]. Available: http://igraph.org

[26] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845