Info-Clustering: A Mathematical Theory for Data Clustering

Chung Chan, Ali Al-Bashabsheh, Qiaoqiao Zhou, Tarik Kaced and Tie Liu

Abstract—We formulate an info-clustering paradigm based on a multivariate information measure, called multivariate mutual information, that naturally extends Shannon's mutual information between two random variables to the multivariate case involving more than two random variables. With proper model reductions, we show that the paradigm can be applied to study the human genome and connectome in a more meaningful way than the conventional algorithmic approach. Not only can infoclustering provide justifications and refinements to some existing techniques, but it also inspires new computationally feasible solutions.

Index Terms—Genome, connectome, data clustering, multivariate mutual information, principal sequence of partitions

I. Introduction

Clustering is the process of grouping similar objects together while separating dissimilar ones apart. This simple idea has a wide range of applications in different areas of scientific research. In *bioinformatics*, clustering can identify co-expressed genes that work together for the same metabolic pathway [3, 4]. In *neuroscience*, clustering can also identify regions of neurons in the brain that are physically or functionally connected [5–7]. Both the *human genome* and the *human connectome* are highly complex systems, with about 23,000 protein-coding genes in the human genome [8] and 16×10^9 neurons composing the cerebral cortex of the human

Preliminary results published in [2] and presented at the Claude Shannon Centenary Workshop in Hong Kong and the Claude Shannon's Centennial Day in Shanghai in 2016. To appear in the special issue of the IEEE Transactions on Molecular, Biological, and Multi-Scale Communications on Biological Applications of Information Theory.

C. Chan (email: cchan@inc.cuhk.edu.hk, chungc@alum.mit.edu), A. Al-Bashabsheh, Q. Zhou are with the Institute of Network Coding at the Chinese University of Hong Kong, the Shenzhen Key Laboratory of Network Coding Key Technology and Application, China, and the Shenzhen Research Institute of the Chinese University of Hong Kong.

T. Kaced is with with Université Paris-Est Créteil at the Algorithmic, Complexity and Logic Laboratory.

T. Liu is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (email: tieliu@tamu.edu).

The work described in this paper was supported by a grant from University Grants Committee of the Hong Kong Special Administrative Region, China (Project No. AoE/E-02/08), and supported partially by a grant from Shenzhen Science and Technology Innovation Committee (JSGG20160301170514984), the Chinese University of Hong Kong (Shenzhen), China.

The work of C. Chan was supported in part by The Vice-Chancellor's One-off Discretionary Fund of The Chinese University of Hong Kong (Project Nos. VCF2014030 and VCF2015007), and a grant from the University Grants Committee of the Hong Kong Special Administrative Region, China (Project No. 14200714).

The work of T. Liu was supported in part by the National Science Foundation under Grant CCF-13-20237. Part of the work was done while T. Liu was visiting the Institute of Network Coding at the Chinese University of Hong Kong.

brain [9]. Therefore, the ability to group similar genes or neurons together based on their interactions is very helpful, as it reduces the complex systems into smaller, and so, more manageable subsystems for further studies.

There are different techniques for capturing the detailed physical structure and functional interaction in a biological system. For the human genome, the expression levels of different genes in different individuals (or tissues) can be measured by the microarray analysis [10] or RNA-sequencing [11, 12]. For the physical connections of the neurons, called the physical connectome, electron microscopy (EM) has been used to map out the entire structural interconnections of the neurons in a small living creature called the nematode Caenorhabditis elegans (C. elegans) [13–16]. For the human brain of living subjects, EM does not apply, but a magnetic resonance imaging (MRI) technique called diffusion spectral imaging (DSI) can be used instead [17-19]. The functional connectome of the neurons can also be studied by capturing the stimulation patterns of the neurons directly using electroencephalography (EEG) or indirectly using another MRI technique called functional MRI (fMRI). However, with the huge volume and variety of data available [20–22], the main challenge is to automate the clustering process using a mathematical criterion that leads to meaningful, yet arithmetically simple to compute, clusters.

We believe that the key to this challenge lies in a better understanding of what information is, and how we can measure mutual information quantitatively. In this work, we propose a novel information-theoretic approach to clustering, called infoclustering, and show that it applies to the study of the complex biological systems of the genes and neurons. The idea is to regard each object as a piece of information, and then group subsets of the objects together if their mutual information exceeds a certain threshold. By varying the threshold value, a hierarchy of clusters can be obtained.

A. Motivation: Human genome and connectome

An application of info-clustering is in the study of the human genome. We know that the biological information of a human being is encoded entirely in its DNA sequences. A DNA sequence is further divided into segments called the genes. Some genes are protein-coding in the sense that they express themselves in the form of gene products such as enzymes, hormones and receptors. These proteins carry out important functions that sustain different metabolic pathways. However, it is not entirely clear

- 1) how the genes work together to sustain the metabolic pathways, and
- how do mutations of the genes cause a certain disease such as cancer.

Clustering is a helpful first step in studying the metabolic pathways and disease pathology. This is because it helps identify smaller subsets of related genes that work closely together. More precisely, although different genes express differently in different people, or even in different tissues of the same person, genes that are co-regulated tend to have similar expression patterns [10, 23]. Such co-expression of the genes means that there is mutual information among the genes. If we have a way to measure such information, then we can cluster the genes according to their mutual information.

Another application of info-clustering is in the study of the human brain. We know that the brain carries out important tasks such as perception, emotion, thought and memory. The way it works is that, the human brain consists of many cells called neurons. These neurons are physically wired together by fiber-like projections called axons. The neurons stimulate each other in some pathway circuitries to carry out the important brain functions. More precisely, when a neuron is excited by an external stimulation, it sends an electrical signal down the axon, which stimulates one or more target neurons through the synapses.

The stimulation mechanism of the neurons has inspired a family of learning models in artificial intelligence called artificial neural networks. In machine learning, such artificial neural networks can be used by deep learning methods to perform complicated tasks such as image recognition [24]. The performance of such methods often superior to alternative approaches. However, exactly why the technique works so well is not entirely clear. It remains a mystery as to

- how the stimulations of neurons lead to the complicated brain functions, and
- how the damages or anomalies in the brain lead to mental disorders such as schizophrenia, bipolar disorder, autism and attention deficit hyperactivity disorder.

We believe information theory [25-29] lies close to the heart of these problems because the stimulation mechanism by electrical and chemical signals are simply transmission and processing of information. It was recently discovered [30-32] that the brain segregates into tightly connected regions, and there are important network hubs, called the rich-clubs, that connect between the different regions. Most neural signals pass through those network hubs, and therefore, damages to such network hubs can be detrimental. On the other hand, such information super highways were found to improve the performance of artificial neural networks [33], because they allow many layers of neurons to communicate effectively with each other. Indeed, the formation of communities and the small-world topology [34] are observed in social networks where people interact by communicating information. Since neurons also interact by transmitting information, we believe info-clustering can be applied to these information systems to discover or explain the communities with a large amount of intra-cluster communications as well as network hubs that support important inter-cluster communications.

B. Contributions

In this section, we give a summary of the contributions of this work and a brief survey of previous works pointing out, whenever possible, similarities and differences between info-clustering and existing approaches. This brief survey is neither complete nor intended to present info-clustering as a replacement of existing approaches, but rather to motivate info-clustering and help properly position it relative to existing works.

Many clustering algorithms have been proposed, even for gene clustering [3, 4, 35]. However, the conventional approach has been typically of a heuristic nature with a primary focus on algorithmic simplicity [36]. Such an algorithmic approach suffers several shortcomings, as it was already indicated by some researchers [37, 38]. For example, the well-known kmeans clustering algorithm and self-organizing map require prior knowledge of the number of clusters, which is a wellknown difficult task. For the k-means clustering algorithm, the similarity between objects is measured by the distance between the data points associated with the objects. This raises the concern that there are several different choices for defining the distance between two points or two clusters. Various mathematical criteria have been proposed. However, such criteria appear to be "easy to fool" in the sense that there are examples for which the resulting clustering solution is obviously not the desired one [39]. The problem is that distance is fundamentally a pairwise measure, and there is no clear unique extension to the case involving more than two data points.

There are clustering techniques that do not require any prior knowledge of the clusters, but their objective functions are often too difficult to compute. As a concrete example, correlation clustering [40] specifies the similarity structure by a simple graph, with positive edges between similar nodes and negative edges between dissimilar nodes. The objective is to cluster the nodes in a way that minimizes the total number of pairs of similar nodes in different clusters and dissimilar nodes in the same cluster. Despite the conceptual simplicity in its formulation, the problem was shown to be NP-hard [40]. This motivated the search for an approximation solution, such as the randomized 3-approximation algorithm in [41], which was also recently extended to a parallel version for clustering big data [42]. However, the obtained clusters are not reproducible since the randomization can result in very different-looking clusters. While there are indexes that evaluate the quality of the clusters, and algorithms that combine different clustering solutions together, a coherent theoretical ground is desired.

The problem of clustering is quite unique in the sense that it attempts to discover unknown patterns in the data. Indeed, [37] raised the question of whether clustering is more of an art than a science, because the existing methods of evaluating a clustering solution are not entirely justified. However, rather than declaring no satisfying solution to the problem, or jumping too quickly to a specific algorithm or

dataset, we believe it is more important to *lay a rigorous* theoretical ground, upon which many meaningful and practical implementations can be developed. Such a paradigm should be general enough to capture complicated similarity structures, and be able to reduce to computationally feasible algorithms under verifiable simplifying model assumptions.

Indeed, information theory has already been considered in some previous works on data clustering [3, 43–47]. In particular, for gene clustering, the well-known Shannon's mutual information [100] was used as a measure of similarity between two genes in the clustering algorithm by Mutual Information Relevance Networks (MIRN) [3]. The measure was reported to be less sensitive to outliers, among other benefits.

Unfortunately, Shannon's mutual information only measures the amount of information mutual to two random variables and so, its use for the multivariate case involving multiple random variables in [3] was not properly justified. As an illustration of this, we give a concrete example where the clustering by MIRN fails to return the desired cluster.

Many other information-theoretic frameworks make use of a proposed multivariate extension of Shannon's mutual information, called the total correlation [48]. Even with this choice of similarity measure, there have been very different approaches. For example, the hierarchical clustering by mutual information in [43] made use of the grouping property of the total correlation for three random variables. The correlation explanation algorithm in [45] used the conditional total correlation directly in the objective function to partition the random variables according to a latent tree model. In [49], the total correlation was further broken down into a sum of the so-called interaction multi-information. While these works consider information theory to be a promising framework for machine learning, a rigorous common theoretical ground is still missing. For instance, the clustering solutions in [43] and [49] have algorithmic characterizations which do not lead to a unique clustering solution. The approaches are mainly supported by experimental rather than theoretical results.

Instead of Shannon's mutual information or the total correlation, info-clustering makes use of a multivariate information measure called the multivariate mutual information (MMI) that can capture the higher-order correlation among multiple random variables. The MMI originates from the divergence upper bound in [50] for the capacity of the secret key agreement problem. Although the bound was shown by [51] to be slack (in the case with helpers), [51] also identified the rather general (no-helper) case when the bound is tight, and interpreted the corresponding expression as a measure of mutual dependency among multiple random variables. This established an alternative characterization of the secret key capacity that was formally studied as a measure of mutual information in [1], where many interpretations and properties of the measure were discovered to naturally extend those of Shannon's mutual information. The expression was therefore named and regarded as the same notion of mutual information as Shannon has defined in his seminal work [100], but extended to the multivariate case. We pause here to make some important remarks on the MMI:

- 1) The MMI has various concrete operational meanings. Indeed, it was shown in [52] to be precisely the capacity of the secret key agreement problem in [50] and the max-flow min-cut characterization of network coding throughput [53]. It is also related to the source coding problem of communication for omniscience [50] and the problem's extension to successive omniscience [54].
- 2) Among other information-theoretic properties, the MMI satisfies the well-known data processing inequality, which has been used in [1, 54–56] to derive new results or resolve some conjectures in other multiterminal information theory problems [57, 58].
- 3) The term MMI has also been used (though not very widely) to refer to McGill's multiple information. As we will explain in §III-A, there is an issue with such an extension of Shannon's mutual information, causing it to be negative for the example shown in Fig. 2b. A correction of this extension will lead to the non-negative MMI we consider. The MMI was also called the minimum partition information in [46], but the name was based on the characterization of the MMI by partitions [52], which is only one of the many possible characterizations. e.g., an axiomatic formulation of the MMI is given in [1] using the so-called mutually correlative property.
- 4) A more abstract mathematical form of the MMI for a sub-modular set function instead of multiple random variables appeared in the work [59] of Fujishige and the work [60] of Narayanan on the principal lattice of partitions of a submodular function. The MMI enriches the abstract mathematical structure with precise information-theoretic meaning by specializing the submodular function to the entropy function.

The MMI has also been applied to clustering by [46] and was shown to be superior to Shannon's mutual information under the proposed framework in [46]. However, unlike infoclustering, the work did not go deep into the information-theoretic interpretations of the MMI, and therefore, did not identify the clustering solution we found. Instead, it considered clustering as a universal communication problem, with a decoder that recovers patterns of the transmitted message as clusters. This idea is interesting although it is unclear whether this model assumption is fruitful or limiting, and whether the universal communication problem can lead to an efficient clustering solution.

The theoretical underpinning of the MMI is a mathematical structure called the principal sequence of partitions (PSP) [60]. On the one hand, this structure enables the MMI and the clusters to be computed in strongly polynomial time [59, 60] (see also [61]), and adds a new dimension to multi-terminal information theory [54–58, 62–66]. On the other hand, the MMI enriches the abstract mathematical structure with information-theoretic meanings.

There is also an existing clustering algorithm, called the minimum average cost (MAC) clustering [61], which builds implicitly upon the principal sequence of partitions to con-

struct the clusters. However, the exact formulation is based on an abstract mathematical criterion that minimizes certain average of a submodular cost function, which differs from that of the principal sequence of partitions. We show by concrete examples that the MAC clustering is different from info-clustering in general. Instead of building our clustering solution on the abstract mathematical structure of the PSP, we start with a seemingly different but more meaningful formulation and eventually connect it to the PSP using the properties of the MMI. We also prove the hierarchical structure of info-clustering separately based on a general property of the MMI, so that potentially other information measures satisfying such property can be applied. Building upon this abstract mathematical framework, a duality result was recently proved in [67] relating the info-clustering problem with the feature selection problem. The info-clustering formulation was also extended slightly there to map to the more elaborate structure of the principal lattice of partitions (PLP) instead of just the PSP.

The info-clustering paradigm is general. Under some simplifying assumptions on the correlation structure, we show that the solution reduces to the clustering solution by MIRN [2] for gene clustering. Another common model reduction is by assuming a jointly Gaussian distribution, as in the gene clustering method called the clustering identification by connectivity kernel (CLICK) [4]. We show that, under the jointly gaussian assumption, info-clustering reduces to a clustering solution that depends only on the covariance matrix through the spectra of its submatrices. This appears to be a new spectral clustering technique different from the spectral clustering algorithm in [68], which was only used as an approximate solution to the NP-hard problem of finding the minimum normalized cut for image segmentation [68].

Under the pairwise independent network (PIN) model [69], where the random variables have a graphical correlation structure, the MMI reduces to the partition connectivity for tree packing [70], a well-known notion in combinatorial optimizations [71]. The MMI was also shown to be equal to the maximum multicast throughput of an undirected network, giving it the usual connectivity notion of max-flow min-cut for graphs, which can be further extended to information flows over hypergraphs and, more generally, matroids [53, 72–74]. We show that under the PIN model, the clustering solution corresponds to the PSP of graphs, and the idea is extended further to hypergraphs and more general channel models, following the usual extension of commodity flow to information flows in network coding. Because the physical interconnections among the neurons can be specified by a graph or a hypergraph, with edges being channels that transmit information, the graphical reduction of the info-clustering algorithm can potentially be applied to identify regions of tightly connected neurons in the brain with high intra-cluster communication rates.

The infomap [47] is another clustering algorithm applied to cluster the human connectome [6]. The idea, like one of the interpretations of info-clustering, is to decompose the network by information flows. However, different from info-clustering, it uses a random walk over a graph as an analogy to information flows over a network. The clusters are obtained

by optimizing a special two-stage source coding of the random walk. Unfortunately, the optimization is difficult, and can only be solved approximately. The two-stage source code is also far from the optimal source coding scheme that achieves the entropy rate [75].

Another information-theoretic approach, called the integrated information theory (IIT), has also been proposed in neuroscience [76] to study consciousness based on the structure and dynamics of the brain. A measure called the integrated information was defined to measure how integrated the subsystems are within a large system. Another informationtheoretic measure is defined in [5] to measure the segregation of a large system into separate subsystems. The motivation of such measures is the construction of a whole-brain computational model that can help explain some important features of the brain. However, despite similarity to the info-clustering paradigm, the proposed measures do not have clear operational meanings because some distribution, normalization factors and parameters are chosen in a rather ad-hoc manner. We will show that the info-clustering paradigm leads to a more meaningful measure of segregation and integration. Indeed, the info-clustering paradigm is not limited to biological systems. It can also apply to other information systems or social networks, like the measure of segregation proposed based on social interactions in [77].

In summary, info-clustering has the following advantages:

- The clustering procedure is driven by a new multivariate information measure called the MMI, which extends Shannon's mutual information between two random variables to the mutual information among multiple random variables. Like Shannon's mutual information [78, 100], the MMI has concrete operational meanings in various information-theoretic problems, including source coding, network communication and security.
- 2) The clusters can be computed in strongly polynomial time due to the underlying mathematical structure called the PSP. However, unlike the related MAC clustering algorithm, info-clustering has a meaningful formulation not based directly upon the abstract mathematical structure.
- The clustering solution is unique and well-defined, unlike many other algorithmic formulations that may require an initial solution or an assumption on the number of clusters.
- 4) Under the Markov tree model, info-clustering reduces to an existing gene clustering algorithm called the clustering by MIRN. This shows that info-clustering can apply to gene clustering and help justify existing clustering algorithms with the concrete operational meanings of info-clustering.
- 5) For some non-Markov tree models, the clustering by MIRN fails to capture the higher-order statistical dependency among multiple random variables, while infoclustering succeeds to identify the correct clusters.
- 6) Under a hypergraphical source model, info-clustering reduces to the PSP of hypergraphs. It gives the PSP a concrete operational meaning as clustering by network

- information flow. Such a model can be applied to cluster the neurons by their physical connections.
- 7) Under the jointly Gaussian assumption, info-clustering reduces to a method of clustering by the covariance matrix. Compared to the existing spectral clustering method, it is a different algorithm that has concrete informationtheoretic meaning.
- 8) A meaningful measure of integration and segregation can be derived in a more rigorous way than the integrated information theory (IIT), with applications beyond biological systems such as social networks.

While there are many practical approximations and implementations possible for info-clustering, the focus of this paper is on the theoretical development and its potential biological applications in the study of human genome and connectome.

Organization: The paper will be organized as follows. The info-clustering paradigm will be formulated in \S II and characterized in \S III, with the detailed clustering procedures implemented in Algorithm 1, 2 and 3. Its biological applications are through the model reductions in \S IV.

Notations: Throughout this paper, unless otherwise specified, we use sans-serif upper-case letters (e.g., Z, X, etc.) to denote random variables and calligraphic font upper-case letters (e.g., C, F, etc.) to denote collections of sets. For any collection of sets F whose elements are subsets of some finite set, we use maximal F to denote the inclusion-wise maximal elements of F, i.e.,

$$\operatorname{maximal} \mathcal{F} := \{ B \in \mathcal{F} \mid \not \exists B' \supsetneq B, B' \in \mathcal{F} \}. \tag{1.1}$$

II. HIERARCHICAL CLUSTERING FORMULATION

The info-clustering formulation in this paper separates into two main components. The first component is a formulation using a threshold constraint on the MMI. We show that the solution is hierarchical, and so, an iterative algorithm can be used to compute the clusters. We will make the proof general using only a simple property of the MMI rather than its detailed definition, i.e., the results herein hold for any multivariate information measure that satisfies such a property. We use the two terms "multivariate information" and "multivariate mutual information (MMI)" for two distinct meanings, where the former refers to a general information measure for multiple random variables as detailed in this section while the latter refers to a specific information measure defined in the next section as (3.3). The second component of the formulation is a refinement of the clustering solution based on further properties of the MMI as detailed in the next section. The reason for the two-step characterization is not only for theoretical elegance but also for practical implementations of info-clustering in subsequent work. The more general hierarchical solution developed in this section may allow the MMI to be approximated and estimated from data more efficiently with a tunable level of computational and sample complexity.

A. Threshold constraint

To cluster objects using information theory, we first associate each object we want to cluster, say i, with all the information that describes it. The information is represented by a random variable, say Z_i , which can be viewed as a file containing some measurements of the object i. Then, we cluster all the objects based on the mutual information among the random variables Z_i 's.

As a motivating example, consider clustering the random variables Z_i 's defined in Fig. 1a. The random variables are correlated in the sense that they share some uniformly random independent bits X_a, X_b, X_c and X_d .¹ It is desirable to group Z_1, Z_2 and Z_3 in a cluster because they share a common bit X_a , but it is desirable to group Z_4 and Z_5 in a different cluster because they share a different independent bit X_b . It is also desirable to group Z_1 and Z_2 as a smaller cluster (compared to $\{1,2,3\}$) because they share the additional bit X_d (in addition to the bit X_a).

More generally, let V be a finite set of objects we want to cluster and

$$\mathsf{Z}_V := (\mathsf{Z}_i \mid i \in V)$$

be the vector of random variables associated with the objects. For every subset $B \subseteq V$ of at least two random variables, we measure their shared information by some finite real number $I(Z_B)$, i.e.,

$$I(\mathsf{Z}_B) \in \mathbb{R}$$
 for $B \subseteq V : |B| \ge 2$.

This multivariate information quantity should depend on the joint distribution P_{Z_B} of the random vector Z_B , but its precise definition will be postponed to \S III-A because the combinatorial structure of the following hierarchical clustering formulation of info-clustering does not depend on the particular choice of this measure, allowing potentially other measures to be used:

Definition 2.1 (Clusters) For a *threshold* $\gamma \in \mathbb{R}$, the set of *clusters* is defined as

$$C_{\gamma}(\mathsf{Z}_V) := \max\{B \subseteq V \mid |B| > 1, I(\mathsf{Z}_B) > \gamma\}.$$
 (2.1a)

The maximality (1.1) requirement ensures consistency among the clusters, i.e., a cluster B with $I(\mathsf{Z}_B) > \gamma$ does not separate apart any other highly correlated random variables in a larger set $B' \supsetneq B$ that also satisfies $I(\mathsf{Z}_{B'}) > \gamma$. For notational convenience,

$$C(\mathsf{Z}_V) := \bigcup_{\gamma \in \mathbb{R}} C_{\gamma}(\mathsf{Z}_V) \tag{2.1b}$$

denotes the collection of all clusters at different thresholds. \Box

In words, given a threshold γ , we consider subsets B of two or more elements from V, such that the random variables Z_B indexed by the elements of B have the multivariate information quantity $I(\mathsf{Z}_B)$ strictly larger than the threshold γ . Out of all such non-singleton subsets satisfying the threshold constraint, we pick the inclusion-wise maximal subsets to be

¹Each random bit is uniformly random over $\{0,1\}$, and the random bits are mutually independent.

(a) Source model.

(b) Clusters for different thresholds.

Fig. 1: An example of clustering a set of random variables based on their shared information. The statistical dependency of the random variables Z_i 's to be clustered are defined in terms of a set of independent uniformly random bits X_i 's.

the clusters. To put it simply, the idea of clustering is to group together random variables, as many as possible, such that the group has increasingly larger amount of multivariate information. The desired level of multivariate information is specified by the threshold γ .

The following is an illustration of the threshold-constraint formulation when applied to the motivating example above.

Example 2.1 Let $V := \{1, ..., 6\}$ and define Z_V in terms of the independent uniformly random bits X_a, X_b, X_c and X_d as shown in Fig. 1a. For this example, the dependency structure is simple, and so, let us define the information measure as the number of shared bits:

$$I(\mathsf{Z}_B) = \begin{cases} 2 & B = \{1, 2\} \\ 1 & B \in \{\{1, 2, 3\}, \{4, 5\}, \{2, 3\}, \{1, 3\}\} \\ 0 & \text{otherwise.} \end{cases}$$
 (2.2)

For instance, $I(Z_{\{1,2,3\}})=1=I(Z_{\{4,5\}})$ because Z_1 , Z_2 , and Z_3 share the common bit X_a while Z_4 and Z_5 share the common bit X_b . Similarly, $I(Z_{\{1,2\}})=2$ because Z_1 and Z_2 share the common bit X_d in addition to the bit X_a . Finally, $I(Z_V)=0$ because Z_6 is independent of all other random variables Z_i for $i\neq 6$.

If $\gamma=0$, then (2.1a) asks for the maximal subsets with shared information strictly larger than zero. By (2.2), the sets whose shared information is larger than zero are $\{4,5\}$, and any subset of two or more elements from $\{1,2,3\}$. The maximal among such subsets are the clusters at threshold zero, i.e., $\mathcal{C}_0(\mathsf{Z}_V)=\{\{1,2,3\},\{4,5\}\}$. If $\gamma=1$ instead, then (2.1a) asks for the maximal subsets of random variables with more than 1 bit of shared information. The only choice is the subset $\{1,2\}$, and so it is the only cluster at threshold 1, i.e., $\mathcal{C}_1(\mathsf{Z}_V)=\{\{1,2\}\}$. For other values of γ :

- For $\gamma < 0$, the entire set V is the only cluster because $I(\mathsf{Z}_V) = 0 > \gamma$ and V is the maximal set trivially.
- For $\gamma \in [0,1)$, there are two clusters $\{1,2,3\}$ and $\{4,5\}$ because $I(\mathsf{Z}_{\{1,2,3\}}) = I(\mathsf{Z}_{\{4,5\}}) = 1 > \gamma$, and each of the sets is maximal.
- For $\gamma \in [1,2)$, the set $\{1,2\}$ is the only cluster because it is the only set with the shared information $I(\mathsf{Z}_{\{1,2\}}) = 2 > \gamma$.

²The maximality constraint can be relaxed slightly in [67] to prove a stronger result on the duality between data clustering and feature selection. For simplicity, we limit the scope of this paper to the maximality constraint.

• There are no clusters for $\gamma \geq 2$ because no set of random variables has more than 2 bits of shared information.

The complete clustering solution is illustrated in Fig. 1b. \Box

There are various important properties we can observe from the clustering solution in Fig. 1b:

- The set of clusters changes at a finite set of threshold values, namely the set of thresholds $\{0,1,2\}$. The threshold values are the shared information of the clusters and each cluster occupies a contiguous interval between two of the threshold values. (E.g., the threshold value 2 is the shared information of the cluster $\{1,2\}$ that appears over the interval $\gamma \in [1,2)$.) In particular, the smallest threshold value 0 is the shared information of the entire set of random variables, and the largest threshold value 2 is the maximum shared information over all subsets that contain at least two random variables.
- For each threshold γ , the clusters are disjoint. For two different thresholds, two clusters are either disjoint or the larger-threshold cluster is a proper subset of the smaller-threshold cluster. (E.g., the cluster $\{1,2,3\}$ at $\gamma=0.5$ does not intersect with the other cluster $\{4,5\}$ at the same threshold, but it contains the cluster $\{1,2\}$ that arises at the larger threshold $\gamma=1.5$.)
- There is an iterative relationship among the clusters: the cluster of a cluster of Z_V is also a cluster of Z_V . For example, if we consider clustering the random variables from the cluster $\{1,2,3\}$, then $\{1,2\}$ is a cluster of $\{1,2,3\}$. Note that $\{1,2\}$ is also a cluster of Z_V .

In the next subsection, we will show that the above hierarchical structure holds more generally.

B. Hierarchical structure

For convenience, we will use γ^- (γ^+) to represent the value that is arbitrarily close to but strictly smaller (larger) than γ . More precisely, we write $\mathcal{C}_{\gamma^-}(\mathsf{Z}_V)$ ($\mathcal{C}_{\gamma^+}(\mathsf{Z}_V)$) for the limit of $\mathcal{C}_t(\mathsf{Z}_V)$ as t increases (decreases) to γ from below (above). The limit exists because there is only a finite number of clusters, i.e.,

$$|\mathcal{C}(\mathsf{Z}_V)| < |\{B \subset V\}| = 2^{|V|} < \infty,$$
 (2.3)

even though the set $C(Z_V)$ of all clusters (2.1b) is a union over all real threshold values γ . Define the set of *critical values* for

clustering Z_V as

$$\Gamma(\mathsf{Z}_V) := \{ \gamma \in \mathbb{R} \mid \mathcal{C}_{\gamma^-}(\mathsf{Z}_V) \neq \mathcal{C}_{\gamma^+}(\mathsf{Z}_V) \}. \tag{2.4}$$

This is the set of threshold values of interest, because the set of clusters changes at those values. The following theorem asserts that the set of critical values is also finite.

Theorem 2.1 (Discreteness) *The set of critical values can be written as*

$$\Gamma(\mathsf{Z}_V) = \{\gamma_i : 1 \le i \le N\}$$

for some positive integer N with $\gamma_i < \gamma_{i+1}$ for $1 \le i < N$. Furthermore, assuming $\{C_{\gamma_i}(\mathsf{Z}_V) : 1 \le i \le N\}$ is the collection of the sets of clusters for the critical values, then the complete clustering (2.1a) of Z_V is given as

$$C_{\gamma}(\mathsf{Z}_{V}) = \begin{cases} \{V\} & \gamma < \gamma_{1} \\ C_{\gamma_{i}}(\mathsf{Z}_{V}) & \gamma \in [\gamma_{i}, \gamma_{i+1}), 1 \leq i < N \\ \emptyset & \gamma \geq \gamma_{N}. \end{cases}$$
 (2.5a)

Finally, the first and last critical values are

$$\gamma_1 = I(\mathsf{Z}_V) \tag{2.5b}$$

$$\gamma_N = \max_{B \subseteq V: |B| > 1} I(\mathsf{Z}_B) \tag{2.5c}$$

and $C_{\gamma_{N-1}}(Z_V)$ is the set of all maximal B achieving the maximum amount of multivariate information (2.5c).

Definition 2.2 (Critical values) We will use $\gamma_i(\mathsf{Z}_V)$ and $\mathcal{N}(\mathsf{Z}_V)$ to denote the *i*-th critical value and the number of critical values for Z_V respectively. For simplicity and when there is no ambiguity, we may drop the dependency on Z_V and write, e.g., $\mathcal{C}_{\gamma_1}(\mathsf{Z}_V)(\mathsf{Z}_V)$ as $\mathcal{C}_{\gamma_1}(\mathsf{Z}_V)$.

Applying the definition of clusters (2.1a) to an arbitrary subset $B \subseteq V$ with size at least two, the definition above extends to any such subset, where in this case the simplified notation $\mathcal{C}_{\gamma_1}(\mathsf{Z}_B)$ will mean $\mathcal{C}_{\gamma_1}(\mathsf{Z}_B)$.

Next, we show that every cluster B' of Z_V can be obtained by computing the set of clusters of $Z_{B''}$ for some larger (previous) cluster B'' of Z_V :

Theorem 2.2 (Iterative relation) For each $B' \in \mathcal{C}_{\gamma'}(\mathsf{Z}_V)$ and $\gamma' \geq \gamma_1(\mathsf{Z}_V)$,

$$\exists B'' \in \mathcal{C}_{\gamma''}(\mathsf{Z}_V) : \gamma'' < \gamma', B' \in \mathcal{C}_{\gamma_1}(\mathsf{Z}_{B''}). \tag{2.6}$$

That is, we can obtain B' by computing clusters that correspond to the first critical value of an earlier cluster B''. \Box

According to Theorem 2.2, we can compute the complete solution to the clustering problem if we can compute the first set $\mathcal{C}_{Y_1}(\mathsf{Z}_B)$ of clusters for all subsets $B\subseteq V:|B|>1$. However, without any additional properties of the multivariate information measure, it is unclear whether the iterative algorithm can be computed efficiently due to the following issues:

- 1) While Theorem 2.2 states that every cluster B' is in the first set $\mathcal{C}_{\gamma_1}(\mathsf{Z}_{B''})$ of clusters of a larger cluster B'', the converse may not be true. That is, a cluster of $\mathsf{Z}_{B''}$ in $\mathcal{C}_{\gamma_1}(\mathsf{Z}_{B''})$ may not be a cluster of Z_V , because the maximality in (2.1a) needs to be verified in addition.
- 2) The total number of clusters under a general multivariate information measure can be large. By (2.1a), $C_{\gamma_1}(Z_V)$ is an *antichain* in the sense that a cluster in $C_{\gamma_1}(Z_V)$ cannot be a subset of another cluster in $C_{\gamma_1}(Z_V)$. Without any other restriction, the size of an antichain can be exponential in |V| by *Sperner's theorem* [79].

To illustrate the issues above more clearly:

Example 2.2 Consider $V := \{1, 2, 3, 4\}$ and let the multivariate information quantity be defined as follows: For $B \subseteq V$: |B| > 1,

$$I(\mathsf{Z}_B) = \begin{cases} 0 & B = V \\ 2 & B = \{2, 3\} \\ 3 & B = \{2, 3, 4\} \\ 1 & \text{otherwise.} \end{cases}$$

From this, it follows that the set of clusters is given by

$$C_{\gamma}(\mathsf{Z}_{V}) = \begin{cases} \{\{1, 2, 3, 4\}\} & \gamma < 0 \\ \{B \subseteq V : |B| = 3\} & \gamma \in [0, 1) \\ \{\{2, 3, 4\}\} & \gamma \in [1, 3) \\ \emptyset & \gamma \ge 3. \end{cases}$$

Consider now the cluster $\{1,2,3\}$ of Z_V . It is not hard to verify that $\{2,3\}$ is a cluster (in the first set of clusters) of $Z_{\{1,2,3\}}$. However, the set $\{2,3\}$ is not a cluster of Z_V because the proper superset $\{2,3,4\}$ has a larger multivariate information of 3.

Next, we show that both issues can be resolved if the following simple, but in our opinion fundamental, property is satisfied by the multivariate information measure:

$$I(\mathsf{Z}_{B_1 \cup B_2}) \ge \min\{I(\mathsf{Z}_{B_1}), I(\mathsf{Z}_{B_2})\}$$
 (2.7)

for all $B_i \subseteq V : |B_i| > 1$, $i \in \{1,2\}$ and $B_1 \cap B_2 \neq \emptyset$. This property holds for the MMI we will consider in (3.3) ([1, Corollary 5.1]), but it also holds for some other multivariate information quantities.³

Theorem 2.3 (Laminarity under (2.7)) For any multivariate information measure that satisfies the property (2.7), the collection of all clusters $C(Z_V)$ forms a laminar family [71], i.e.,

$$B_1 \cap B_2 \in \{\emptyset, B_1, B_2\}$$
 (2.8)

for all clusters $B_1, B_2 \in \mathcal{C}(\mathsf{Z}_V)$. In particular, for every $\gamma \in \mathbb{R}$, the set $\mathcal{C}_{\gamma}(\mathsf{Z}_V)$ consists of disjoint clusters.

PROOF Consider two clusters $B_1, B_2 \in \mathcal{C}(\mathsf{Z}_V)$ with $B_1 \cap B_2 \neq \emptyset$. Without loss of generality, assume $I(\mathsf{Z}_{B_1}) \leq I(\mathsf{Z}_{B_2})$. By (2.7), $I(\mathsf{Z}_{B_1 \cup B_2}) \geq I(\mathsf{Z}_{B_1})$ and so $B_1 \cup B_2 = B_1$ or it would contradict the maximality of B_1 . That is, $B_2 \subseteq B_1$ and so we have (2.8).

Under the context of data clustering, the aforementioned laminarity is usually known as hierarchical clustering or dendrogram. By Theorem 2.3, any multivariate information measure that satisfies the property (2.7) will necessarily lead to a clustering solution that is guaranteed to be hierarchical. If we define a similarity relation $i \sim_{\gamma} j$ to mean that there exists $C \subseteq V$ containing both i and j such that $I(\mathbb{Z}_C) > \gamma$, then it can be shown that \sim_{γ} is an equivalence relation for any threshold γ . In particular, (2.7) implies that the relation is transitive. The set $\mathcal{C}_{\gamma}(\mathbb{Z}_V)$ of clusters can be shown to be precisely the set of non-singleton equivalence classes, and so the clusters are disjoint.

Our next result shows that under the condition (2.7), the complete solution to the clustering problem can indeed be computed iteratively from the first set of clusters of a previous cluster.

Theorem 2.4 (Iterative relation under (2.7)) For any multivariate information measure that satisfies the property (2.7),

$$C(\mathsf{Z}_V) = \{V\} \cup \bigcup_{B' \in \mathcal{C}_{\mathsf{Y}_1}(\mathsf{Z}_V)} C(\mathsf{Z}_{B'}), \tag{2.9}$$

where for any $B \subseteq V$, the set $C(Z_B)$ is the collection of all clusters of Z_B (similar to (2.1b)).

PROOF See Appendix A-3.

Corollary 2.1 For any multivariate information measure that satisfies the property (2.7), the total number of clusters

$$|\mathcal{C}(\mathsf{Z}_V)| < |V| - 1, \tag{2.10}$$

which is linear in the number of random variables to be clustered.

PROOF Consider proving (2.10) by an induction on the size of V. In particular, consider the non-trivial case when $\mathcal{C}_{\gamma_1}(\mathsf{Z}_V)$ is non-empty. The base case |V|=2 holds trivially. By (2.9),

$$\begin{split} |\mathcal{C}(\mathsf{Z}_V)| &= 1 + \sum_{B' \in \mathcal{C}_{Y_1}(\mathsf{Z}_V)} |\mathcal{C}(\mathsf{Z}_{B'})| \\ &\leq 1 + \sum_{B' \in \mathcal{C}_{Y_1}(\mathsf{Z}_V)} (|B'| - 1) \leq |V| - 1, \end{split}$$

where the first inequality is by the inductive hypothesis and the last is because $C_{\gamma_1}(Z_V)$ consists of disjoint proper subsets of V by Theorem 2.3.

Hence, given an algorithm that can compute the first critical value γ_1 and the first set \mathcal{C}_{γ_1} of clusters of any given set of random variables, we can compute the entire clustering solution by applying the algorithm at most |V| times. The pseudocode is given in Algorithm 1.

³Watanabe's total correlation (3.2), Han's dual total correlation and the multivariate Wyner's common information described in [1] satisfy (2.7) because they are non-decreasing. However, their corresponding clustering solutions are also trivial due to the monotonicity. The normalized version of the total correlation considered in [1, (6.2)] satisfies (2.7), as can be shown by [1, Corollary 5.4]). It is not monotonic and therefore gives non-trivial clustering solutions with the laminar structure by Theorem 2.3. McGill's multiple information and the multivariate Gács–Körner common information both fail to satisfy (2.7).

Algorithm 1: Hierarchical clustering by iteration.

```
Data: Statistics of Z_V sufficient for calculating
              FirstClusters(B) for all B \subseteq V : |B| > 1.
    Result: S is a list of (I(Z_C), C) for C \in \mathcal{C}(Z_V), which
                gives \Gamma(\mathsf{Z}_V) = \{ \gamma' \mid (\gamma', B) \in \mathcal{S} \} and
                C_{\gamma}(\mathsf{Z}_V) = \text{maximal}\{B \mid (\gamma', B) \in \mathcal{S}, \gamma' > \gamma\}.
 1 \mathcal{S}, \mathcal{T} \leftarrow empty queues;
 2 enqueue V to \mathcal{T};
3 while T is non-empty do
         B \leftarrow \text{dequeue } \mathcal{T};
         (\gamma, \mathcal{C}) \leftarrow \text{FirstClusters}(B);
5
         enqueue (\gamma, B) to \mathcal{S};
         enqueue all elements of C to T;
8 end
9 function FirstClusters (B):
         return (\gamma_1(\mathsf{Z}_B), \mathcal{C}_{\gamma_1}(\mathsf{Z}_B));
11 end
```

The algorithm computes the list S of all clusters $B \in \mathcal{C}(\mathsf{Z}_V)$ and their associated values $I(\mathsf{Z}_B)$. It calls the function FirstClusters (B) iteratively to obtain the first critical value γ and the first set \mathcal{C} of clusters of every previously discovered cluster stored temporarily in \mathcal{T} . The newly discovered clusters in \mathcal{C} are further added to \mathcal{T} .

III. CLUSTERING BY THE MMI

In this section, we focus on the clustering solution under the MMI measure in [1]. Although a general property, namely, property (2.7), suffices for a laminar hierarchical clustering solution in \S II, the resulting clusters may be trivial or meaningless if the multivariate information measure is not chosen properly. (For instance, any choice of $I(Z_B)$ that is non-decreasing in B will satisfy (2.7) but will only produce the trivial cluster $\{V\}$.) We will explain the meaning of the MMI measure precisely and show that the corresponding hierarchical clustering formulation is related to a non-trivial mathematical structure called the principal sequence of partitions (PSP) [60] of the entropy function. Consequently, the solution can be computed by some well-studied submodular function optimization techniques [60] that can run in strongly polynomial time.

A. Multivariate mutual information

Recall that for the simple example considered earlier in Fig. 1a, the mutual information of a given set was measured by the number of bits shared by the random variables in the set. For a general source model, *Shannon's mutual information* [100] provides a well-accepted measure in the bivariate

case involving only two random variables:⁴

$$I(\mathsf{Z}_1 \wedge \mathsf{Z}_2) := D\left(P_{\mathsf{Z}_1,\mathsf{Z}_2} \| P_{\mathsf{Z}_1} P_{\mathsf{Z}_2}\right),$$
 (3.1)

where $D(P_1||P_2)$ denotes the Kullback-Liebler divergence [83] between the distributions P_1 and P_2 . The divergence on the R.H.S. of (3.1) can be interpreted as a statistical distance to independence, because it equals zero if and only if the two random variables are independent.

A straightforward extension of Shannon's mutual information (3.1) to the multivariate case is Watanabe's total correlation [48]:

$$J_{\mathrm{T}}(\mathsf{Z}_{V}) := D\left(P_{\mathsf{Z}_{V}} \mid\mid \prod_{i \in V} P_{\mathsf{Z}_{i}}\right),\tag{3.2}$$

which is equal to zero if and only if the random variables Z_i for $i \in V$ are mutually independent. While the total correlation captures the mutual independence among the random variables, it fails to capture many other forms of independence relation. The work in [1] aims at a more precise understanding of this and the formulation of the MMI, as follows, capable of capturing any form of independence that might exist among three or more random variables.

Let $\Pi'(V)$ be the collection of all possible partitions of V that splits V into at least two nonempty disjoint subsets. (In other words, $\Pi'(V)$ is the collection of all set partitions of V except the trivial partition $\{V\}$.) For any partition $\mathcal{P} \in \Pi'(V)$ of V, the product distribution $\prod_{C \in \mathcal{P}} P_{Z_C}$ specifies an independence relation, namely, that the agglomerated random variables Z_C 's are mutually independent. A well-constructed measure needs to ensure that the mutual information is measured at zero as long as an independence relation exists among the random variables from V, not just when all the random variables from V are mutually independent.

We now introduce the MMI measure from [1]:

$$I(\mathsf{Z}_V) := \min_{\mathcal{P} \in \Pi'(V)} I_{\mathcal{P}}(\mathsf{Z}_V) \quad \text{where}$$
 (3.3a)

$$I_{\mathcal{P}}(\mathsf{Z}_V) := \frac{1}{|\mathcal{P}| - 1} D\left(P_{\mathsf{Z}_V} \middle\| \prod_{C \in \mathcal{P}} P_{\mathsf{Z}_C}\right).$$
 (3.3b)

Clearly, by the above definition, we have $I(Z_V) = 0$ if and only if there exists an independence relation among the random variables from V.

Example 3.1 For the example considered earlier in Fig. 1a, we have $I(\mathbf{Z}_V) = 0$ because of the independence relation

$$P_{\mathsf{Z}_{\{1,2,3,4,5,6\}}} = P_{\mathsf{Z}_{\{1,2,3\}}} P_{\mathsf{Z}_{\{4,5\}}} P_{\mathsf{Z}_6}$$

and so
$$I_{\mathcal{P}}(\mathsf{Z}_V) = 0$$
 with $\mathcal{P} = \{\{1, 2, 3\}, \{4, 5\}, \{6\}\}.$

The divergence expression (3.3) of the MMI derives from a divergence upper bound [50, (26) in Example 4] on the

⁴For the bivariate case, there are also other measures of shared information such as the Wyner's common information [80] and Gács-Körner common information [81]. Shannon's mutual information was also described by Shannon [100] as the amount of information "common" to two random variables. However, Wyner's common information and Gács-Körner common information are not as widely used as Shannon's mutual information measures. They measure more specific kind of shared information and also have their own multivariate extension in [50, 56, 82].

secrecy capacity for the multiterminal secret key agreement problem. The bound was derived in the general case with helpers, and was considered as a heuristically meaning upper bound to the LP characterization of the capacity in [50]. The bound was shown to be tight for the case involving 2 or 3 users even involving helpers, but it was left open in [50] whether the bound is tight beyond 3 users. [51, 52] extended the brute-force search of [50] and showed with the help of a computer program that the bound is tight for 4 or 5 users. However, a counter-example involving 6 users with 3 helpers was also discovered, showing that the bound is loose with the presence of helpers and therefore does not have the same meaning as the secrecy capacity. Nevertheless, it was identified and proved in [51, Theorem 1][52, Theorem 1.1][1, Theorem 2.1] that the bound is tight in the no-helper case even under a general private source distribution, using only the wellknown submodularity of entropy. This establishes the concrete operational meaning of the MMI as the secrecy capacity in the general no-helper case, much like the way Shannon's mutual information was shown to characterize the channel capacity in the seminal work [100] of Shannon. A first attempt to interpret the MMI as a measure of mutual information among multiple random variables (and to explain the normalization factor of $|\mathcal{P}| - 1$ in (3.3)) appeared in [51, Section IV].

It is useful to compute I_P by rewriting the divergence in terms of Shannon's entropy or mutual information as follows:

$$D\left(P_{\mathsf{Z}_{V}} \middle\| \prod_{C \in \mathcal{P}} P_{\mathsf{Z}_{C}}\right) = \sum_{C \in \mathcal{P}} H(\mathsf{Z}_{C}) - H(\mathsf{Z}_{V}) \qquad \textbf{(3.4a)}$$
$$= \sum_{i=1}^{k-1} I(\mathsf{Z}_{C_{1} \cup \cdots \cup C_{i}} \wedge \mathsf{Z}_{C_{i+1}}) \quad \textbf{(3.4b)}$$

where we used C_1, \ldots, C_k to denote the blocks of the partition \mathcal{P} . The measure $I_{\mathcal{P}}$ is also written more explicitly in [1] as

$$I_{\mathcal{P}}(\mathsf{Z}_V) = I(\mathsf{Z}_{C_1} \wedge \mathsf{Z}_{C_2} \wedge \cdots \wedge \mathsf{Z}_{C_k}).$$

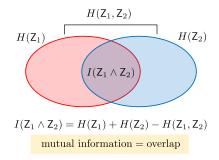
Through (3.4b), we can verify that the MMI measure defined as above is consistent with the measure of shared bits used in Example 2.2 for the special source model in Fig. 1a.

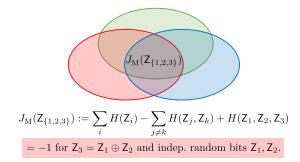
Example 3.2 Consider the example in Fig. 1a. The values of $I_{\mathcal{P}}(\mathsf{Z}_{\{1,2,3\}})$ for different partitions \mathcal{P} are

$$\begin{split} &I(\mathsf{Z}_1 \wedge \mathsf{Z}_{\{2,3\}}) = 2 \\ &I(\mathsf{Z}_2 \wedge \mathsf{Z}_{\{1,3\}}) = 2 \\ &I(\mathsf{Z}_3 \wedge \mathsf{Z}_{\{1,2\}}) = 1 \\ &I(\mathsf{Z}_1 \wedge \mathsf{Z}_2 \wedge \mathsf{Z}_3) = \frac{I(\mathsf{Z}_1 \wedge \mathsf{Z}_2) + I(\mathsf{Z}_{\{1,2\}} \wedge \mathsf{Z}_3)}{2} = \frac{3}{2}, \end{split}$$

where the last term is obtained by applying (3.4b) with $C_i = \{i\}$. Hence, $I(\mathsf{Z}_{\{1,2,3\}}) = 1$ with $\{\{1,2\},\{3\}\}$ being the unique optimal partition.

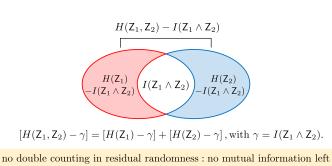
At the first sight, the normalization factor $|\mathcal{P}|-1$ on the R.H.S. of (3.3b) may appear arbitrary. This factor is not included in other proposed information measures involving the divergence, such as the total correlation (3.2) and the

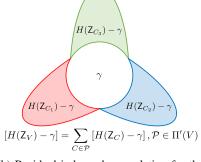




(a) Venn diagram for Shannon's mutual information. (See (3.5).) (b) Information diagram for McGill's multiple information.

Fig. 2: Interpretation of mutual information as the amount of overlap in randomness.





(b) Residual independence relation for the multi-(a) Residual independence relation for Shannon's mutual information. (See (3.6).)

Fig. 3: MMI as the smallest value satisfying the residual independence relation.

integrated information [76]. However, it turns out that such a factor has an important information-theoretic meaning, which relates it to the non-trivial, but polynomial-time solvable, mathematical structure of the PSP. Indeed, the normalization factor is often overlooked in other proposed multivariate information measures based on the independence relations, such as the total correlation in (3.2). This is because the factor only affects the measure when the independence relations do not hold, i.e., when the measure is non-zero.

To help understand the reasoning behind such a normalization, we will introduce the residual independence relation [1] by extending the well-known Venn diagram interpretation of Shannon's mutual information shown in Fig. 2a according to the identity

$$I(Z_1 \wedge Z_2) = H(Z_1) + H(Z_2) - H(Z_1, Z_2).$$
 (3.5)

From the Venn diagram, the mutual information has the meaningful interpretation as the amount of overlap in the randomness of the individual random variables. This interpretation has been extended by [84] to the Information diagram, and the amount of overlap can be measured by the McGill's multiple information [85] using the inclusion-exclusion principle. (See Fig. 2b.) Unfortunately, the McGill's multiple information can be negative even for a very simple example involving three random variables [84], contradicting the basic intuition that mutual information should be non-negative.

To "fix" this problem, one may rewrite (3.5) equivalently as

$$H(\mathsf{Z}_1, \mathsf{Z}_2) - I(\mathsf{Z}_1 \wedge \mathsf{Z}_2) = [H(\mathsf{Z}_1) - I(\mathsf{Z}_1 \wedge \mathsf{Z}_2)] + [H(\mathsf{Z}_2) - I(\mathsf{Z}_1 \wedge \mathsf{Z}_2)]. \quad (3.6)$$

Note that the L.H.S. is the total *residual* randomness after removing the mutual information, and the equality states that the total residual randomness is equal to the sum of the individual residual randomness in each random variable, as illustrated in Fig. 3a. The important interpretation of the equality is that:

No double counting in the sum means precisely that there is no mutual information left in the residual randomness.

The above idea can be extended to the multivariate case as follows. Consider a partition $\mathcal{P} \in \Pi'(V)$ and define the residual independence relation (RIR) as

$$H(\mathsf{Z}_V) - \gamma = \sum_{C \in \mathcal{P}} \left[H(\mathsf{Z}_C) - \gamma \right], \tag{3.7}$$

i.e., the total residual randomness after removing some real value $\gamma \in \mathbb{R}$ is equal to the sum of the individual residual randomness of the agglomerated random variables Z_C 's. This is illustrated in Fig. 3b.

We can now interpret the MMI (3.3) as the *smallest* $\gamma \in \mathbb{R}$ such that the RIR (3.7) holds for some partition $\mathcal{P} \in \Pi'(V)$.

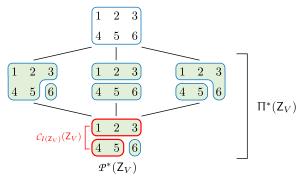


Fig. 4: Each partition in green is an optimal partition of the MMI (3.8) (corresponding to an independence relation) for the source in Fig. 1a. The non-singleton elements of the fundamental partition give some of the clusters shown in Fig. 1b, namely, the first set of clusters at the threshold γ_1 . (See Theorem 3.1.)

To see this, a simple re-arrangement of the terms in (3.7) gives

$$\gamma = \frac{1}{|\mathcal{P}| - 1} \left[\sum_{C \in \mathcal{P}} H(\mathsf{Z}_C) - H(\mathsf{Z}_V) \right] = I_{\mathcal{P}}(\mathsf{Z}_V)$$

where the last equality follows from (3.4a). Minimizing over all possible $\mathcal{P} \in \Pi'(V)$ gives rise to the MMI, and the normalization factor $|\mathcal{P}|-1$ appears naturally under the context of RIR due to the fact that any information mutual to the set of random variables should appear in every agglomerated random variable, and is therefore over-counted in the divergence. Note also that the residual randomness can be shown to be always non-negative as expected when the partition is the minimizing partition, and when the random variables are discrete.⁵

B. Fundamental partition

Back to the clustering problem, how do we compute the clusters efficiently using the MMI? Surprisingly, the optimal partition achieving the MMI (3.3a) gives us the desired clustering solution. The optimal partition achieving the MMI may not be unique, but it can be shown that the set of optimal partitions form a semi-lattice, which together with the partition $\{V\}$ is referred to as the *Dilworth truncation lattice* [60]. The lattice structure means that there exists a unique finest partition, which is called the *fundamental partition* [1]. It turns out that the non-singleton elements of the fundamental partitions give us the desired clusters.

The proof is rather involved, but the idea can be illustrated using the previous example in Fig. 1a. As explained in Example 3.1, $I(Z_V) = 0$ because of the independence relation $P_{Z_V} = P_{Z_{\{1,2,3\}}} P_{Z_{\{4,5\}}} P_{Z_6}$. There are also other independence relations, which correspond to further merging of the agglomerated random variables as shown in Fig. 4. Hence, the fundamental (finest optimal) partition is

$$\mathcal{P}^*(\mathsf{Z}_V) = \{\underbrace{\{1,2,3\},\{4,5\}}_{\mathit{C}_I(\mathsf{Z}_V)},\{6\}\}.$$

Note that, the non-singleton subsets in the fundamental partition, namely $\{1,2,3\}$ and $\{4,5\}$, are the clusters in $C_{I(Z_V)}(Z_V) = C_0(Z_V)$ in Fig. 1b. If we take the cluster $\{1,2,3\}$ and further compute its MMI, we have (from Example 3.2) $I(Z_{\{1,2,3\}}) = 1$ with the unique optimal partition

$$\mathcal{P}^*(\mathsf{Z}_{\{1,2,3\}}) = \{\underbrace{\{1,2\}}_{C_{I(\mathsf{Z}_{\{1,2,3\}})}(\mathsf{Z}_{\{1,2,3\}})}, \{3\}\}.$$

Note that the non-singleton subset $\{1,2\}$ in the fundamental partition of the cluster $\{1,2,3\}$ is a cluster of the entire set of random variables at the larger threshold of $I(Z_{\{1,2,3\}})$. In other words, clusters at larger thresholds can be obtained from the fundamental partition of some clusters at smaller thresholds.

Below, we will show more generally that computing the fundamental partitions iteratively gives the complete infoclustering solution. Recall from Theorem 2.1 that the first critical value $\gamma_1(\mathsf{Z}_V)$ is always given by the MMI $I(\mathsf{Z}_V)$ of the entire set. To determine the corresponding first set $\mathcal{C}_{\gamma_1}(\mathsf{Z}_V)$ of clusters, denote the set of optimal partitions to (3.3a) by

$$\Pi^*(\mathsf{Z}_V) := \{ \mathcal{P} \in \Pi'(\mathsf{Z}_V) : I_{\mathcal{P}}(\mathsf{Z}_V) = I(\mathsf{Z}_V) \}. \tag{3.8}$$

In general, $\Pi^*(Z_V)$ may contain more than one optimal partition, but the optimal partitions are related by the following partial order defined for any partitions \mathcal{P} and \mathcal{P}' as

$$\mathcal{P} \prec \mathcal{P}' \quad \text{iff} \quad \forall C \in \mathcal{P}, \ \exists C' \in \mathcal{P}' : C \subseteq C'.$$
 (3.9)

We say that \mathcal{P} is finer/smaller than \mathcal{P}' whenever $\mathcal{P} \leq \mathcal{P}'$, and we use \prec to denote the strict inequality. The following result is known from [1].

Proposition 3.1 ([1, Theorems 5.2 and 5.3]) There is a unique finest optimal partition, which we denote by $\mathcal{P}^*(\mathsf{Z}_V)$ and refer to as the fundamental partition for Z_V . (Namely, $\mathcal{P}^*(\mathsf{Z}_V) = \min \Pi^*(\mathsf{Z}_V)$, where the minimum is with respect to \preceq in (3.9).) Furthermore, the fundamental partition $\mathcal{P}^*(\mathsf{Z}_V)$ with the singletons removed, i.e., $\mathcal{P}^*(\mathsf{Z}_V) \setminus \{\{i\} \mid i \in V\}$, is the set of maximal subsets $B \subseteq V$ with $I(\mathsf{Z}_B) > I(\mathsf{Z}_V)$.

Since we have $\gamma_1(\mathsf{Z}_V) = I(\mathsf{Z}_V)$ by (2.5b), and the non-singleton elements from $\mathcal{P}^*(\mathsf{Z}_V)$ are, by Proposition 3.1, the maximal subsets of random variables with mutual information strictly larger than $\gamma_1(\mathsf{Z}_V)$, we have by (2.1a) that:

Theorem 3.1
$$C_{\gamma_1}(\mathsf{Z}_V) = \mathcal{P}^*(\mathsf{Z}_V) \setminus \{\{i\} \mid i \in V\}$$
 with the first critical value $\gamma_1(\mathsf{Z}_V) = I(\mathsf{Z}_V)$.

⁵ [1, Theorem 6.3] also provides an axiomatic characterization of the MMI measure, but the RIR interpretation appears more intuitive; it seems to be more information-theoretically appealing than the mathematical form (3.3) as it explains the normalization factor not considered by previous multivariate information measures; and it serves as a fix to the Venn diagram extension of Shannon's mutual information. Furthermore, similar to the way Shannon's mutual information provides a theoretical limit to the communication rate [100], the MMI has various operational meanings [1]. For instance, it specifies the limit of the amount of common secret key (mutual information) extractable individually from different but correlated random sources.

By Proposition 3.1, the MMI measure defined in (3.3a) is guaranteed to satisfy the key property (2.7) by [1, Corollary 5.1]. This can be argued as follows. Suppose to the contrary of (2.7) that both $I(\mathsf{Z}_{B_1})$ and $I(\mathsf{Z}_{B_2})$ are strictly larger than $I(\mathsf{Z}_{B_1\cup B_2})$. Then, the non-singleton elements in the fundamental partition $\mathcal{P}^*(\mathsf{Z}_{B_1\cup B_2})$ must consist of a superset of B_1 as well as a superset of B_2 . However, the two supersets cannot be disjoint as $B_1\cap B_2\neq\emptyset$, and they cannot be $B_1\cup B_2$ either, which is a contradiction.

The following result thus follows immediately from Theorems 2.3 and 2.4.

Theorem 3.2 *Under the MMI measure* (3.3a), *the clustering solution* (2.1a) *is guaranteed to be hierarchical. Furthermore, all clusters can be obtained by computing the fundamental partition iteratively for every previously obtained cluster.*

Given an algorithm that computes the fundamental partition exactly or approximately, we can compute the entire infoclustering solution following the iterative procedure in Algorithm 1. This is stated more precisely in Algorithm 2. The complexity is again |V| times the complexity in calculating the fundamental partition by (2.10).

Algorithm 2: Clustering by fundamental partition.

```
Data: Statistics of Z_V sufficient for calculating
            FundamentalPartition (B) for all
            B \subseteq V : |B| > 1.
   Result: S is a list of (I(Z_C), C) for C \in C(Z_V), which
              gives \Gamma(\mathsf{Z}_V) = \{\gamma' \mid (\gamma', B) \in \mathcal{S}\} and
              C_{\gamma}(\mathsf{Z}_V) = \text{maximal}\{B \mid (\gamma', B) \in \mathcal{S}, \gamma' > \gamma\}.
1 S, T \leftarrow empty queues;
2 enqueue V to \mathcal{T};
3 while \mathcal{T} is non-empty do
        B \leftarrow \text{dequeue } \mathcal{T}:
        (\gamma, \mathcal{P}) \leftarrow \text{FundamentalPartition}(B);
5
        enqueue (\gamma, B) to \mathcal{S};
6
        enqueue all non-singleton elements of \mathcal{P} to \mathcal{T};
8 end
9 function FundamentalPartition (B):
        return (I(Z_B), \mathcal{P}^*(Z_B));
11 end
```

Similar to Algorithm 1, the algorithm computes the list \mathcal{S} of all clusters $B \in \mathcal{C}(\mathsf{Z}_V)$ and their associated values $I(\mathsf{Z}_B)$. It calls the function FundamentalPartition (B) iteratively to obtain the first critical value γ and the fundamental partition \mathcal{P} of every previously discovered cluster stored temporarily in \mathcal{T} . The non-singleton elements of \mathcal{P} are the desired clusters further added to \mathcal{T} .

C. Principal sequence of partitions of entropy function

As pointed out in [1, 53, 86], the MMI and the fundamental partition can both be computed in polynomial time assuming the entropies of arbitrary subsets of the random variables in hand are also computable in polynomial time. This result is based on the property that the (conditional) mutual information is non-negative, or equivalently, the entropy is submodular [87]. Hence, the iterative algorithm in the previous section can discover the info-clustering solution in polynomial time.

Quite surprisingly, based on the RIR (3.7) interpretation of the MMI in §III-A, we find that the info-clustering solution can be mapped to the polynomial-time solvable mathematical structure of the principal sequence of partitions (PSP). The implication is that one can compute the general info-clustering solution more efficiently than the iterative algorithm, using techniques such as [60, 61]. This understanding will also allow us to compare the info-clustering solution to the closely related approach of MAC clustering [61]. The study of PSP from an information-theoretic perspective appears to be new, and we are beginning to discover more information-theoretic interpretations in other problems [54–56].

Define for $\gamma \in \mathbb{R}$ the residual entropy function [1]:

$$h_{\gamma}(B) := h(B) - \gamma$$
 for $B \subseteq V$, (3.10)

where $h(B) := H(\mathsf{Z}_B)$ is the usual entropy function [25]. $h_\gamma(B)$ measures the residual randomness of Z_B introduced in §III-A. For notational simplicity, the dependency on Z_B is implicit here.

The entropy function is well-known⁷ to be submodular [87], i.e., for all $B_1, B_2 \subseteq V$,

$$h(B_1) + h(B_2) \ge h(B_1 \cap B_2) + h(B_1 \cup B_2),$$
 (3.11)

and so, it is clear that the residual entropy function h_{γ} (3.10) is also submodular. The Dilworth truncation of the submodular residual entropy function is defined as

$$\hat{h}_{\gamma}(B) := \min_{\mathcal{P} \in \Pi(B)} h_{\gamma}[\mathcal{P}] \quad \text{for } B \subseteq V, \text{ where}$$
 (3.12a)

$$h_{\gamma}[\mathcal{P}] := \sum_{C \in \mathcal{P}} h_{\gamma}(C) \tag{3.12b}$$

and $\Pi(B)$ is the collection of all partitions of B into non-empty subsets. Note that the difference between $\Pi(B)$ and $\Pi'(B)$ is that $\Pi(B)$ includes the trivial partition $\{B\}$ as well, i.e., $\Pi'(B) = \Pi(B) \setminus \{\{B\}\}$. The Dilworth truncation is itself a submodular set function, and can be calculated efficiently in strongly polynomial time [88] for any given set using Edmonds' greedy algorithm and the *submodular function minimization (SFM)*. The running time is $O(|V|\operatorname{SFM}(|V|))$, where $\operatorname{SFM}(|V|)$ is the running time of the submodular function minimization over the ground set V. (See [71] and [1].) $\operatorname{SFM}(|V|)$ can be strongly polynomial assuming that the

 $^{^6\}mathrm{Indeed},$ we will see in the next section how the fundamental partition can be computed from the PSP of the entropy function. In fact, the additional factor of |V| can be saved by using the PSP rather than the iterative algorithm to compute the clusters. Nevertheless, the iterative algorithm is useful as it potentially allows us to compute the entire clustering solution approximately based on an approximate algorithm of computing the fundamental partition.

⁷The submodularity follows directly from the non-negativity of the conditional mutual information $I(Z_{B_1} \wedge Z_{B_2} | Z_{B_1 \cap B_2}) \ge 0$ [25].

entropy function can be evaluated efficiently for every given subset of random variables.

To characterize the info-clustering solution, we will focus on the Dilworth truncation evaluated at V:

$$\hat{h}_{\gamma}(V) = \min_{\mathcal{P} \in \Pi(V)} h_{\gamma}[\mathcal{P}]$$
 (3.13)

and think of it as a function of γ . More precisely, it is a minimization of the function

$$h_{\gamma}[\mathcal{P}] = \sum_{C \in \mathcal{P}} h_{\gamma}(C) = \sum_{C \in \mathcal{P}} H(\mathsf{Z}_C) - \gamma |\mathcal{P}|,$$

which is linear in γ with

$$\text{slope} = -|\mathcal{P}|, \qquad \text{y-intercept} = \sum\nolimits_{C \in \mathcal{P}} H(\mathsf{Z}_C).$$

Since $\hat{h}_{\gamma}(V)$ is a minimization over a finite collection of linear curves, it must be piecewise linear. More explicitly, for a given γ , let Π^* be the set of partitions attaining the minimization in (3.13), then at γ^+ the Dilworth truncation is given as the curve with the minimum slope among Π^* . Thus, $\hat{h}_{\gamma}(V)$ is piecewise linear in γ with slopes decreasing from -1 to -|V| and taking only integer values, as shown in Fig. 5. Since there is a finite number of partitions of V, the curve can be characterized by the set of turning points p_i 's where the slope changes. Denote the turning points as

$$p_1 = (\gamma_1, y_1), \dots, p_N = (\gamma_N, y_N)$$
 (3.14a)

for some positive integer N, and call

$$\gamma_1 < \gamma_2 < \dots < \gamma_N \tag{3.14b}$$

the critical values for the Dilworth truncation $\hat{h}_{\gamma}(V)$.

Example 3.3 Consider the example in Fig. 1a. To compute $\hat{h}_{\gamma}(C_1^*)$ for $C_1^* := \{1, 2, 3\}$, note that the values of $h_{\gamma}[\mathcal{P}]$ for different partitions \mathcal{P} are

$$\underbrace{ \underbrace{ b(\{1,2,3\})}_{2} - \gamma }_{b(\{1\}) + h(\{2,3\})} - 2\gamma \qquad \mathcal{P} = \{\{1,2,3\}\}$$

$$\underbrace{ b(\{1\}) + h(\{2,3\})}_{4} - 2\gamma \qquad \mathcal{P} = \{\{1\},\{2,3\}\}$$

$$\underbrace{ b(\{2\}) + h(\{1,3\})}_{4} - 2\gamma \qquad \mathcal{P} = \{\{2\},\{1,3\}\}$$

$$\underbrace{ b(\{3\}) + h(\{1,2\})}_{3} - 2\gamma \qquad \mathcal{P} = \{\{3\},\{1,2\}\}$$

$$\underbrace{ b(\{1\}) + h(\{2\}) + h(\{3\})}_{5} - 3\gamma \qquad \mathcal{P} = \{\{1\},\{2\},\{3\}\}.$$

The minimum, $\hat{h}_{\gamma}(C_1^*)$, of the above lines is plotted in Fig. 5a. For $\gamma < 1 = I(\mathsf{Z}_{C_1^*})$, the minimum is achieved uniquely by $\mathcal{P} = \{C_1^*\} = \{\{1,2,3\}\}$. For $\gamma \in (1,2)$, the minimum is achieved uniquely by $\mathcal{P} = \mathcal{P}^*(\mathsf{Z}_{C_1^*}) = \{\{3\},\{1,2\}\}$. For $\gamma > 2$, the minimum is achieved uniquely by the partition into singletons.

Similarly, $\hat{h}_{\gamma}(V)$ can be plotted as the minimum of a set of lines in Fig. 5b. For $\gamma < 0 = I(\mathsf{Z}_V)$, the minimum is achieved uniquely by $\mathcal{P} = \{V\} = \{1, \dots, 6\}$. For $\gamma \in (0, 1)$,

the minimum is achieved uniquely by $\mathcal{P}=\mathcal{P}^*(\mathsf{Z}_V)=\{\{1,2,3\},\{4,5\},\{6\}\}$. For $\gamma\in(1,2)$, the minimum is achieved uniquely by $\mathcal{P}=\{\{1,2\},\{3\},\{4\},\{5\},\{6\}\}\}$. For $\gamma>2$, the minimum is achieved uniquely by the partition into singletons.

The connection of the Dilworth truncation to the MMI is through the RIR, as shown in [1, Theorem 5.1]. When γ is sufficiently small, $\hat{h}_{\gamma}(V) = h_{\gamma}(V)$ because $h_{\gamma}[\mathcal{P}]$ has the largest slope of -1 when $\mathcal{P} = \{V\}$. More precisely,

$$h_{\gamma}[\mathcal{P}] - h_{\gamma}[\{V\}] = D\left(P_{\mathsf{Z}_{V}} \mid\mid \prod_{C \in \mathcal{P}} P_{\mathsf{Z}_{C}}\right) + (1 - |\mathcal{P}|)\gamma,$$

which will be positive, i.e., $h_{\gamma}[\mathcal{P}] > h_{\gamma}[\{V\}] = h_{\gamma}(V)$, for $|\mathcal{P}| > 1$ (or $\mathcal{P} \neq \{V\}$) and γ is sufficiently small. Therefore, it follows from (3.13) that p_1 is the intersection between $h_{\gamma}(V)$ and $\min_{\mathcal{P} \in \Pi'(V)} h_{\gamma}[\mathcal{P}]$, and so $\gamma = \gamma_1$ satisfies the equation

$$h_{\gamma}(V) = \min_{\mathcal{P} \in \Pi'(V)} h_{\gamma}[\mathcal{P}], \tag{3.15}$$

which translates directly to the RIR in (3.7). Hence, we have $\gamma_1 = I(\mathsf{Z}_V)$. Furthermore, p_1 lies on $h_\gamma[\mathcal{P}]$ if and only if $\mathcal{P} \in \Pi^*(\mathsf{Z}_V) \cup \{\{V\}\}$, since $\Pi^*(\mathsf{Z}_V)$ defined in (3.8) is the set of solutions to the minimization in (3.15), as can be seen from the RIR interpretation of the MMI. Since the fundamental partition $\mathcal{P}^*(\mathsf{Z}_V)$ is the unique finest partition in $\Pi^*(\mathsf{Z}_V)$, $h_\gamma[\mathcal{P}^*(\mathsf{Z}_V)]$ has the smallest slope and therefore uniquely defines the line segment following p_1 .

Note that, it is not clear a priori that the critical values (3.14b) defined for the Dilworth truncation $\hat{h}_{\gamma}(V)$ are precisely the critical values in $\Gamma(Z_V)$ (2.4) defined for infoclustering (2.1a), even though the above result from [1] shows that it is the case for the first critical value. We will show the stronger result that not only the two sets of critical values match, but that the line segments of the Dilworth truncation give the desired info-clustering solution. E.g., from Fig. 5b, the critical values of the Dilworth truncation can be verified to be precisely the critical values for the info-clustering solution in Fig. 1b. Furthermore, the sequence of partitions defining the line segments in Fig. 5b contains all the clusters in Fig. 1b as its non-singleton elements. This sequence of partitions is the PSP, which will be defined more precisely below.

Let $\Pi_i \subseteq \Pi(V)$ be the set of solutions to the minimization in $\hat{h}_{\gamma_i}(V)$. The elements of Π_i form a lattice:

Proposition 3.2 ([60, Theorem 3.5]) The set of optimal solutions to the Dilworth truncation $\hat{f}(V)$ of a submodular function $f: 2^V \mapsto \mathbb{R}$ forms a lattice (with respect to the partial order in (3.9)) called the Dilworth truncation lattice.

For instance, when we specialize the submodular function to the entropy function h for Z_V , the first critical value is $\gamma_1 = I(Z_V)$ and the associated lattice of partitions is $\Pi_1 = \Pi^*(Z_V) \cup \{\{V\}\}.$

Let $\min \Pi_i$ and $\max \Pi_i$ be, respectively, the (unique) minimum and maximum partitions in the lattice Π_i . The following proposition asserts that, for all i, the extreme partitions $\min \Pi_i$ and $\max \Pi_{i+1}$ are equal. (In particular, the fundamental partition is $\mathcal{P}^*(\mathsf{Z}_V) = \min \Pi_1 = \max \Pi_2$.) Furthermore, the

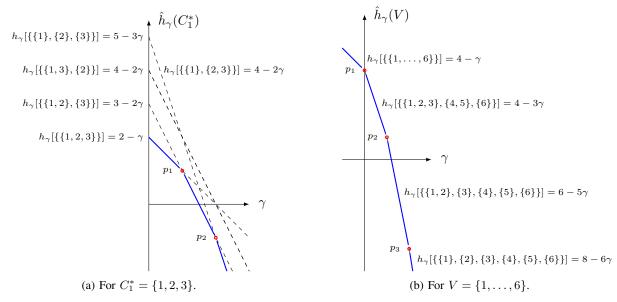


Fig. 5: Plots of Dilworth truncation for Example 3.3 as the minimum of the lines $h_{\gamma}[\mathcal{P}]$ over different partitions.

extreme partitions for different values of i form a sequence of successively finer partitions, referred to as the PSP.

Proposition 3.3 ([60, Theorem 3.7]) There is a unique sequence of partitions with respect to the partial order (3.9)

$$\mathcal{P}_0 \succ \mathcal{P}_1 \cdots \succ \mathcal{P}_N \in \Pi(V),$$
 (3.16a)

called the principal sequence of partitions (PSP), which satisfies

$$\mathcal{P}_{i-1} = \max \Pi_i \quad and \quad \mathcal{P}_i = \min \Pi_i$$
 (3.16b)

for $i \in \{1, ..., N\}$. More explicitly, $\mathcal{P}_0 = \max \Pi_1 = \{V\}$,

$$\min \Pi_i = \mathcal{P}_i = \max \Pi_{i+1} \quad \text{for } i \in \{1, \dots, N-1\},$$

and
$$\mathcal{P}_N = \min \Pi_N = \{\{i\} \mid i \in V\}.$$

Same the argument in [61], the PSP is computable in strongly polynomial time in $O(|V|^2 \operatorname{SFM}(|V|))$. For completeness, we include a simple proof below.

PROOF For $i \in \{1, \dots, N-1\}$, the line segment of $\hat{h}_{\gamma}(V)$ for $\gamma \in (\gamma_i, \gamma_{i+1})$ is defined by $h_{\gamma}[\mathcal{P}_i]$ for some partition $\mathcal{P}_i \in \Pi_i \cap \Pi_{i+1}$ because it passes through both turning points p_i and p_{i+1} . Since it has the smallest and largest slopes among all other lines through p_i and p_{i+1} respectively, \mathcal{P}_i is the minimum in Π_i and maximum in Π_{i+1} as desired. $\mathcal{P}_0 = \{V\}$ and $\mathcal{P}_N = \{\{i\} \mid i \in V\}$ because $h_{\gamma}[\mathcal{P}]$ has the largest and smallest slopes when \mathcal{P} is \mathcal{P}_0 and \mathcal{P}_N respectively.

The desired connection between info-clustering and the PSP of the entropy function follows from the main result below, which gives an interpretation to every critical value of the Dilworth truncation using the PSP.

Theorem 3.3 The *i*-th critical value of $\hat{h}_{\gamma}(V)$ (3.13) is

$$\gamma_{i} = \min_{\substack{\mathcal{P} \in \Pi(V): |\mathcal{P}| > |\mathcal{P}_{i-1}| \\ C \in \mathcal{P}_{i-1}: |C| > 1}} \frac{h[\mathcal{P}] - h[\mathcal{P}_{i-1}]}{|\mathcal{P}| - |\mathcal{P}_{i-1}|}$$

$$= \min_{\substack{C \in \mathcal{P}_{i-1}: |C| > 1}} I(\mathsf{Z}_{C}).$$
(3.17a)
$$(3.17b)$$

The set of optimal solutions to (3.17a) is $\Pi_i \setminus \{\mathcal{P}_{i-1}\}$. The set of optimal solutions to (3.17b), denoted as \mathcal{C}_{i-1}^* , is equal to $\mathcal{P}_{i-1} \setminus \mathcal{P}_i$, or equivalently,

$$\mathcal{P}_i = \left(\mathcal{P}_{i-1} \setminus \mathcal{C}_{i-1}^*\right) \cup \bigcup_{C \in \mathcal{C}_{i-1}^*} \mathcal{P}^*(\mathsf{Z}_C). \tag{3.18}$$

Furthermore, with the product of set families \mathcal{F} and \mathcal{G} defined as $\mathcal{F} \times \mathcal{G} := \{ \{F, G\} \mid F \in \mathcal{F}, G \in \mathcal{G} \}$, we have

$$\Pi_i = \prod_{C \in \mathcal{L}_{i-1} \setminus \mathcal{L}_i} [\Pi^*(\mathsf{Z}_C) \cup \{\{C\}\}] \times \prod_{C \in \mathcal{L}_{i-1} \cap \mathcal{L}_i} \{\{C\}\} \quad \textbf{(3.19)}$$

which consists of refinements of \mathcal{P}_{i-1} by successively partitioning one or more blocks $C \in \mathcal{P}_{i-1} \setminus \mathcal{P}_i$ according to $\Pi^*(\mathsf{Z}_C)$.

When i=1, (3.17a) reduces to $\gamma_1=I(\mathsf{Z}_V)$ because $\mathcal{Z}_{i-1}=\{V\}$ and $h[\mathcal{P}]-h[\mathcal{Z}_{i-1}]=D(P_{\mathsf{Z}_V}\|\prod_{C\in\mathcal{P}}P_{\mathsf{Z}_C})$. (3.19) reduces to $\Pi_1=\Pi^*(\mathsf{Z}_V)\cup\{\{V\}\}$ with $\mathcal{Z}_{i-1}\setminus\mathcal{Z}_i=\{V\}$ and $\mathcal{Z}_{i-1}\cap\mathcal{Z}_i=\emptyset$.

For $i \geq 1$, (3.17b) means that the other critical values can be obtained simply by iteratively computing the MMI for the non-singleton blocks of the fundamental partitions. (3.18) is essentially the iteration in (2.9) to obtain the clusters iteratively. Therefore, the critical values for the Dilworth truncation coincide with the critical values for the set of clusters, and the clusters are the non-singleton elements of the partitions in the PSP. This is summarized in the following corollary. (See Definition 2.2) for some of the notations.)

Corollary 3.1 For $1 \le i \le \mathcal{N}(\mathsf{Z}_V)$, we have $\mathcal{C}_{\gamma_i}(\mathsf{Z}_V) = \mathcal{P}_i(\mathsf{Z}_V) \setminus \{\{j\} : j \in V\}$ with the critical value $\gamma_i(\mathsf{Z}_V)$ being the *i*-th critical value for $\hat{h}_{\gamma}(V)$.

Since the info-clustering solution maps to the entire PSP of the entropy function, we can compute the clustering solution in strongly polynomial time as well. The algorithm is given in Algorithm 3, which is based on the algorithm of [61].

Algorithm 3: Clustering by PSP for entropy function. **Data**: Statistics of Z_V sufficient for calculating the

```
entropy function h(B) for B \subseteq V := \{1, \dots, m\}.
     Result: The array L contains the values in \Gamma(Z_V). The
                     array PSP contains the PSP \mathcal{P}_i's. More precisely,
                     \mathcal{P}_i is stored in PSP[|\mathcal{P}_i|], and \gamma_i is stored in
                     L[|\mathcal{P}_{i-1}|]. Hence, \mathcal{C}_{\gamma}(\mathsf{Z}_V) is the set the
                     non-singleton values of PSP[s] where s is the
                     smallest index with L[s] > \gamma. C_{\gamma}(Z_V) = \emptyset if no
                     such s exists.
 1 L, PSP\leftarrow empty arrays of size m;
 2 Q \leftarrow \{V\}, \mathcal{P} \leftarrow \{\{i\} \mid i \in V\};
 3 \text{ PSP}[|\mathcal{Q}|] \leftarrow \mathcal{Q};
 4 Split (Q, P);
 \begin{array}{ll} \textbf{5 procedure} \ \texttt{Split} \ (\mathcal{Q}, \mathcal{P}) \textbf{:} \\ \textbf{6} & \quad \gamma' \leftarrow \frac{1}{|\mathcal{P}| - |\mathcal{Q}|} (h[\mathcal{P}] - h[\mathcal{Q}]); \\ \textbf{7} & \quad h' \leftarrow \frac{1}{|\mathcal{P}| - |\mathcal{Q}|} (|\mathcal{P}| h[\mathcal{Q}] - |\mathcal{Q}| h[\mathcal{P}]); \end{array} 
            \mathcal{P}' \leftarrow \emptyset, \mathbf{x} \leftarrow all-zero array of size m;
 8
            for l=1 to m do
                   (\alpha,T) \leftarrow \text{SubmodularFnMin}(B \mapsto
10
                   h_{\gamma'}(\mathsf{Z}_B) - \sum_{i \in B} \mathsf{x}[i], l;
                   add \alpha to x[\overline{l}];
11
                   foreach C in \mathcal{P}' do
12
                          if C \cap T \neq \emptyset then
13
                                  T \leftarrow T \cup C;
14
                                  remove C from \mathcal{P}';
15
                          end
16
                   end
17
                   add T to \mathcal{P}';
18
19
            \begin{array}{ll} \text{if} & h' = \sum_{i=1}^m \mathbf{x}[i] \text{ then} \\ \big| & \mathbf{L}[|\mathcal{Q}|] \leftarrow \gamma'; \end{array}
20
21
            else
22
                   \mathsf{PSP}[|\mathcal{P}'|] \leftarrow \mathcal{P}';
23
                   Split(Q, P');
24
                   Split (\mathcal{P}', \mathcal{P});
25
            end
26
    end
27
28 function SubmodularFnMin (f,l):
            U \leftarrow \{1, \ldots, l\};
29
            return (\min_{B \subset U: l \in B} f(B), \arg\min_{B \subset U: l \in B} f(B))
30
31 end
```

Algorithm 3 computes the sequence of critical values and the PSP, and stores them in the arrays L and PSP respectively. The desired clusters can then be obtained from the nonsingleton subsets in the PSP. As in [61], the procedure Split starts with two partitions $\mathcal{Q} \succ \mathcal{P}$ in the PSP, and then check if there is any other partition \mathcal{P}' in the PSP with $\mathcal{Q} \succ \mathcal{P}' \succ \mathcal{P}$. To do so, it first computes the intersection point (γ',h') of the two lines $h_{\gamma}[\mathcal{Q}]$ and $h_{\gamma}[\mathcal{P}]$, and then check whether h' is equal to the Dilworth truncation $\hat{h}_{\gamma'}(V)$ (which is computed by lines 8–19 and stored in $\sum_{i=1}^{m} \mathbf{x}[i]$). If they are equal

(line 20), then \mathcal{Q} and \mathcal{P} are two consecutive partitions with no other partition between them in the PSP, and so γ' is a critical value. Otherwise, the optimal partition \mathcal{P}' achieving the Dilworth truncation must be a partition in the PSP satisfying $\mathcal{Q} \succ \mathcal{P}' \succ \mathcal{P}$. In this case, the procedure Split can be invoked in a recursive manner to further identify other partitions in the PSP that may lie between \mathcal{Q} and \mathcal{P}' , and between \mathcal{P}' and \mathcal{P} .

The complexity of the algorithm is mainly due to the computations of the Dilworth truncation (lines 8–19) by the submodular function minimization SubmodularFnMin [71]. The number of such computations is at most |V|-1, and each has a complexity of $O(|V|\operatorname{SFM}(|V|))$. Therefore, the overall complexity is $O(|V|^2\operatorname{SFM}(|V|))$.

Indeed, [61] also proposed the MAC clustering algorithm that builds upon the algorithm for finding the PSP for a submodular cost function. Although we have shown that infoclustering is also intimately connected to the PSP of the entropy function, the two clustering approaches are different in two ways:

- 1) Unlike info-clustering where the MMI is specified as a measure of mutual information under a meaningful hierarchical clustering formulation, the formulation of the MAC clustering does not specify how one should choose the submodular cost function for clustering. Hence, the mathematical criterion of MAC does not have a concrete operational meaning, that is, it is unclear in what sense are the elements in the same cluster are similar.
- 2) Unlike info-clustering where the solution maps precisely to the *entire* PSP of the entropy function, the solution of the MAC clustering is sensitive to shifts of the cost function by a constant, and is therefore not identical to the PSP of the submodular cost function.

In Appendix D, we give detailed explanations with concrete examples differentiating the two algorithms.

IV. MODEL REDUCTIONS

In this section, we show that info-clustering reduces to simpler clustering solutions under some special models. Model reduction is important for practical implementations because learning the entropy function from data, and even evaluating the entropy of an arbitrary distribution can take exponential time with respect to the number of random variables.

In the following, we show that the clustering algorithm by mutual information relevance networks (MIRN) [3] is a special case when Z_V forms a Markov tree. We also show that, if Z_V is jointly Gaussian, the clustering solution will depend only on the covariance matrix, which may be estimated more easily from data. Finally, if Z_V has a hypergraphical correlation, then info-clustering reduces to the procedure of computing the PSP for hypergraphs, which is useful in clustering the human connectome.

⁸The fundamental partition is obtained as a special case since it is a partition in the PSP. Although we do not know of a faster exact algorithm to compute the fundamental partition for the general source model, faster approximation algorithms may be possible, in which case the iterative procedure in Algorithm 2 can be used to approximate the entire clustering solution.

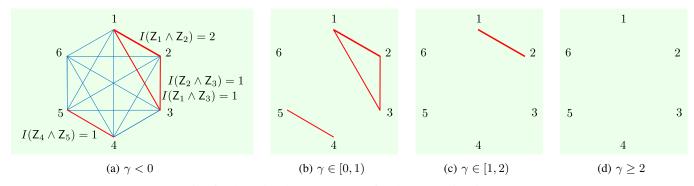


Fig. 6: Clustering by MIRN (4.3) for the source in Fig. 1a.

A. Clustering by Chow-Liu tree approximation

We first introduce the clustering by MIRN in [3] for gene clustering. This clustering algorithm first constructs a weighted complete graph, where the nodes represent the genes to be clustered. The weight of the edge between the nodes i and j is equal to the Shannon's mutual information $I(Z_i \wedge Z_j)$, which may be estimated from measurements of the expression level Z_i and Z_j of the corresponding genes i and j respectively. An example of such a graph is shown in Fig. 6a for the simple source model in Fig. 1a. In Fig. 6a, each blue edge has weight zero. Each red edge has weight one, except for the red edge between node 1 and 2, which has weight two.

Given a threshold γ , the algorithm filters the edges by removing all edges with weights no larger than γ . The clusters at threshold γ are then defined as the non-singleton components of the resulting graphs. Such non-singleton components are called the MIRN. For instance, in the case of Fig. 6, the edge removal (or clustering) for different γ 's is as follows:

- When $\gamma < 0$, we have the complete graph since the mutual information is non-negative. Consequently, we have the trivial cluster V.
- For $\gamma \in [0,1)$, all the blue edges are removed since they have weight equal to 0. Hence, we have the two clusters $\{1,2,3\}$ and $\{4,5\}$.
- For $\gamma \in [1,2)$, only the edge between nodes 1 and 2 remains, and so, we have the cluster $\{1,2\}$.
- For $\gamma \geq 2$, all edges are removed and so we have no clusters.

Note that the clusters we obtained from the above edgefiltering procedure are precisely the clusters we obtained by info-clustering in Fig. 1b. We can show more generally that, info-clustering reduces to the clustering by MIRN when the random variables Z_i 's form a Markov tree. In this example, we indeed have a Markov chain structure, namely,

$$Z_1 - Z_2 - Z_3 - Z_4 - Z_5 - Z_6$$

which is a special case of the Markov tree. If the random variables do not form a Markov tree, then the MIRN solution turns out to correspond to applying info-clustering after approximating the correlation structure by a Markov tree. More precisely, for a set of random variables whose distribution does not necessarily factor according to a Markov tree, the clustering solution by MIRN corresponds to the solution resulting from applying info-clustering to any Markov tree

obtained via the Chow-Liu tree approximation [89] of the distribution.⁹

To explain the reduction above between info-clustering and clustering by MIRN, we first define the clustering by MIRN more formally using some graph-theoretic notations. For a simple graph G with the vertex set V, we denote its edge set by $\mathcal{E}(G)\subseteq\{B\subseteq V:|B|=2\}$ (with the calligraphic font used for set families). For $i,j\in V$, we write $i\sim_G j$ to indicate that j is reachable from i via a path in G. Note that \sim_G is an equivalence relation, and we denote the set of equivalence classes as:

$$\mathcal{P}(G) := \max \{ B \subseteq V \mid i \sim_G j, \forall i, j \in B \} \in \Pi(V),$$
(4.1)

where $\Pi(V)$ denotes the collection of all partitions of V into non-empty disjoint sets. Each element in $\mathcal{P}(G)$ is the vertex set of a connected component of G, which will be considered as a cluster by MIRN as we describe below.

For any threshold $\gamma \in \mathbb{R}$, let $K_{\gamma}(\mathsf{Z}_V)$, or simply K_{γ} , be a graph with vertex set V and edge set

$$\mathcal{E}(K_{\gamma}) := \{\{i, j\} \mid i, j \in V, i \neq j, I(\mathsf{Z}_i \wedge \mathsf{Z}_j) > \gamma\}. \quad (4.2)$$

In words, we think of K as a complete graph and associate each edge $\{i,j\}$ with weight $I(\mathsf{Z}_i \wedge \mathsf{Z}_j)$. Then, the graph K_γ can be obtained from K by removing the "light" edges, i.e., edges with weight no larger than γ .

Definition 4.1 (MIRN [3]) The non-singleton connected components of $K_{\gamma}(\mathsf{Z}_V)$ are called the mutual information relevance networks (MIRN). The corresponding clusters are given by:

$$\mathcal{P}(K_{\gamma}) \setminus \{\{i\} \mid i \in V\},\tag{4.3}$$

where $\mathcal{P}(K_{\gamma})$ is the partition of the vertices of K_{γ} according to the connected components of K_{γ} .

Next, we introduce the Chow–Liu tree approximation under which (4.3) can be obtained from info-clustering. Consider a tree T with vertex set V. A dependency-tree approximation [89] to Z_V , denoted as Z_V^T , can be written in terms of the marginal distributions Z_B for $|B| \leq 2$ as:

$$P_{\mathsf{Z}_{V}^{T}}(z_{V}) := \left(\prod_{i \in V} P_{\mathsf{Z}_{i}}(z_{i})\right) \prod_{\{i,j\} \in \mathcal{E}(T)} \frac{P_{\mathsf{Z}_{i},\mathsf{Z}_{j}}(z_{i},z_{j})}{P_{\mathsf{Z}_{i}}(z_{i})P_{\mathsf{Z}_{j}}(z_{j})},$$
(4.4)

⁹The preliminary result has been published in [2].

for $z_V \in Z_V$. Such a distribution forms a Markov tree or a Bayesian network (in which the in-degree of every vertex is at most one) with respect to T, i.e., we can relabel the indices in V to $\{1, \ldots, |V|\}$ such that

$$\begin{split} P_{\mathsf{Z}_V^T}(z_V) := \prod_{i \in V} P_{\mathsf{Z}_i | \mathsf{Z}_{\mathsf{p}_i}}(z_i | z_{\mathsf{p}_i}) \\ \text{where } \mathsf{p}_1 = \emptyset, \, \mathsf{p}_i < i, \, \text{and} \, \left\{i, \mathsf{p}_i\right\} \in \mathcal{E}(T) \, \, \text{for} \, \, i > 1. \quad \textbf{(4.5)} \end{split}$$

Definition 4.2 (Chow-Liu trees [89]) The set of Chow-Liu trees is defined as

$$\mathcal{T}^*(\mathsf{Z}_V) := \arg\min_{T} D(P_{\mathsf{Z}_V} \| P_{\mathsf{Z}_V^T}) \qquad \text{where} \qquad \textbf{(4.6)}$$

$$D(P_{\mathsf{Z}_{V}} \| P_{\mathsf{Z}_{V}^{T}}) \stackrel{\text{(a)}}{=} D\left(P_{\mathsf{Z}_{V}} \| \prod_{i \in V} P_{\mathsf{Z}_{i}}\right) - \sum_{e \in \mathcal{E}(T)} I(\mathsf{Z}_{e}).$$
 (4.7)

Here, (a) follows from (4.4). For any $T \in \mathcal{T}^*(\mathsf{Z}_V)$, Z_V^T is called a Chow-Liu tree approximation to Z_V .

The celebrated Chow-Liu algorithm [89] computes a Chow-Liu tree as a maximum weight spanning tree since the minimization in (4.6) corresponds to maximizing the second term on the right-hand side of (4.7), which is the total weight of the tree.

The main result of this subsection is the following theorem on the equivalence between the clustering by MIRN and the clustering by MMI under the Chow-Liu tree approximation.

Theorem 4.1 The clustering of Z_V by MMI (2.1a) under the Chow-Liu tree approximation (4.6) is

$$C_{\gamma}(\mathsf{Z}_{V}^{T}) = \mathcal{P}(K_{\gamma}) \setminus \{\{i\} \mid i \in V\}$$
 (4.8)

for any $\gamma \in \mathbb{R}$ and any $T \in \mathcal{T}^*(\mathsf{Z}_V)$. Such a solution is identical to the clustering by MIRN (4.3) and independent of the choice of $T \in \mathcal{T}^*(\mathsf{Z}_V)$.

PROOF See Appendix C-1.

The proof of the equivalence makes use of the following result which evaluates the MMI for any dependency-tree distribution.

Theorem 4.2 (MMI of dependency-tree distributions) For any tree T on the vertex set V,

$$I(\mathsf{Z}_V^T) = \min_{e \in \mathcal{E}(T)} I(\mathsf{Z}_e) \quad and \tag{4.9}$$

$$\begin{split} I(\mathsf{Z}_V^T) &= \min_{e \in \mathcal{E}(T)} I(\mathsf{Z}_e) \quad \textit{and} \\ \mathcal{P}^*(\mathsf{Z}_V^T) &= \mathcal{P}(T_{I(\mathsf{Z}_V^T)}), \end{split} \tag{4.9}$$

where, as in (4.2), T_{γ} denotes the tree T with edges e of weight $I(Z_e) \leq \gamma$ removed.

(4.9) was discovered in [50] to be the secrecy capacity for Markov trees but (4.10) is new.

PROOF See Appendix C-2.

In other words, the connected components of T_{γ} for $\gamma = I(\mathsf{Z}_V^T)$ characterize the fundamental partition for any dependency-tree distribution Z_V^T .

The following theorem shows that T_{γ} in fact characterizes the entire hierarchical clustering (2.1a) of Z_V^T for different values of γ :

Theorem 4.3 (Clustering of dependency-tree distributions) For any tree T on V and any $\gamma \in \mathbb{R}$,

$$C_{\gamma}(\mathsf{Z}_{V}^{T}) = \mathcal{P}(T_{\gamma}) \setminus \{\{i\} \mid i \in V\}. \tag{4.11}$$

Furthermore, the critical value γ_i in (3.17b) is the i-th smallest value in $\{I(Z_e) \mid e \in \mathcal{E}(T)\}$ and the partition \mathcal{P}_i in (3.18) is $\mathcal{P}(T_{\gamma_i})$, for $1 \leq i \leq N$ and Z_C^T in place of Z_C .

PROOF See Appendix C-3.

However, the Chow-Liu tree approximation incurs a loss. Indeed, the clustering by MIRN fails to capture higher-order statistics beyond pairwise mutual information, because the algorithm only requires the knowledge of the pairwise mutual information. The following is a concrete example where the clustering by MIRN fails, while the general info-clustering without the Chow-Liu tree approximation succeeds.

Example 4.1 Let $V = \{1, 2, 3, 4\}$,

$$Z_1 := X_a, Z_2 := X_b, Z_3 := X_a \oplus X_b$$
 and $Z_4 := X_c$

where X_i 's are independent uniformly random bits. It can be shown that Z_i 's are pairwise independent, and so their pairwise mutual information are all zero. The clustering by MIRN will construct a complete graph with zero weight on the edges. Hence, it will not return any cluster for threshold $\gamma > 0$, because all the edges get removed in K_0 (4.2).

However, we know that Z_1 , Z_2 and Z_3 share some mutual information, because Z_3 can be completely determined by Z_1 and Z_2 . Indeed, it can be shown that

$$I(\mathsf{Z}_{\{1,2,3\}}) = \overbrace{\frac{H(\mathsf{Z}_1)}{H(\mathsf{Z}_1)} + \overbrace{H(\mathsf{Z}_2)}^{=1} + \overbrace{H(\mathsf{Z}_3)}^{=1} - \overbrace{H(\mathsf{Z}_1,\mathsf{Z}_2,\mathsf{Z}_3)}^{=2}}_{3-1} = \frac{1}{2},$$

and so the random variables share non-negative mutual information. The general info-clustering algorithm will correctly find the cluster $\{1,2,3\}$ at threshold $\gamma \in [0,\frac{1}{2})$.

B. Clustering by covariance matrix

The Gaussian distribution is often used as a simplifying assumption because the distribution is completely characterized by its mean and covariance, both of which can be estimated quite efficiently from data. The measure of segregation in [5], for instance, is simply the differential entropy of a set of random variables assuming a jointly Gaussian distribution. The CLICK algorithm [4] for gene clustering also makes certain assumption about the distribution being Gaussian, e.g., in the computation of the parameters and the threshold test. However, these assumptions are often mixed with other simplifications that make it rather difficult to tract the validity or the impact of the Gaussian assumption. For instance, the measure of segregation has a noise variance that is chosen in an adhoc manner to make the differential entropy in the desired range. For the CLICK algorithm, the clustering solution is defined as the end result of an algorithmic procedure, but only some of the steps are justified by the Gaussian assumption. In other words, the clustering solution does not appear to be uniquely defined from the mixture of algorithmic procedure and Gaussian assumption.

In contrast, we will derive a unique info-clustering solution assuming the random vector Z_V is jointly Gaussian with zero mean and covariance matrix Σ_V . It follows that any random subvector Z_B for $B \subseteq V$ is also jointly Gaussian with zero mean and covariance matrix Σ_B where Σ_B is the submatrix of Σ_V with the rows and columns indexed by elements outside B removed. In the following, we use |A| to denote the determinant of any square matrix A.

Proposition 4.1 For the jointly Gaussian source Z_V defined above,

$$I(\mathsf{Z}_{V}) = \min_{\mathcal{P} \in \Pi'(V)} \frac{\sum_{C \in \mathcal{P}} \log |\mathbf{\Sigma}_{C}| - \log |\mathbf{\Sigma}_{V}|}{|\mathcal{P}| - 1}$$
(4.12a)
$$= \min_{\mathcal{P} \in \Pi'(V)} \frac{\sum_{C \in \mathcal{P}} \sum_{i \in C} \log \lambda_{i}^{C} - \sum_{i \in V} \log \lambda_{i}^{V}}{|\mathcal{P}| - 1}$$
(4.12b)
$$= \min_{\mathcal{P} \in \Pi'(V)} \frac{\log |\mathbf{\Sigma}_{\mathcal{P}}| - \log |\mathbf{\Sigma}_{V}|}{|\mathcal{P}| - 1}$$
(4.12c)

where $(\lambda_i^C \mid i \in C)$ is a vector of the eigenvalues of Σ_C and $\Sigma_{\mathcal{P}}$ is the matrix Σ_V but with the entry at row i and column j forced to 0 if i and j belong to different blocks in P.

PROOF (4.12a) is obtained by substituting the following differential entropy into (3.3a),

$$H(\mathsf{Z}_C) = \frac{1}{2} \left(|C| \log 2\pi e + \log |\mathbf{\Sigma}_C| \right). \tag{4.13}$$

(4.12b) follows from the fact that the determinant $|\Sigma_C|$ is the product $\prod_{i \in C} \lambda_i^C$ of the eigenvalues. (4.12c) is because, by possibly reordering the indices in $V, \Sigma_{\mathcal{P}}$ can be written as a block diagonal matrix with Σ_C for $C \in \mathcal{P}$ being the blocks in the main diagonal. Hence, $|\Sigma_{\mathcal{P}}| = \prod_{C \in \mathcal{P}} |\Sigma_C|$.

Proposition 4.2 For the jointly Gaussian source Z_V defined above, the clusters are the non-singleton subsets from the PSP of the submodular function $C \mapsto \log |\Sigma_C|$.

From (4.12b), the clustering solution can be regarded as spectral clustering in the sense that it depends on the spectrum of the submatrices of the covariance matrix. However, it is a new clustering method different from the usual spectral clustering solution such as the one for approximately minimizing the normalized cuts [68].

C. Clustering by Network information flow

In order to make info-clustering applicable to the clustering of neurons based on their physical connectome, we need to convert the deterministic physical connections of neurons to a random source Z_V . We will show that this conversion is possible by reducing the info-clustering solution under the hypergraphical source model [52, 90]. To explain the idea, we start with the emulated source model.

Definition 4.3 (Emulated source [52, Definition 2.1]) For $i \in V$, let $Z_i = (X_i, Y_i)$ such that,

$$P_{\mathsf{X}_{V}\mathsf{Y}_{V}} = \prod_{i \in V} P_{\mathsf{X}_{i}} P_{\mathsf{Y}_{i}|\mathsf{X}_{V}}.$$
 (4.14)

The vector Z_V is called an emulated source network.

We can think of every $i \in V$ as a terminal that can send an input signal X_i independently over a channel that returns the output signal $Y_i = f_i(X_V, N_i)$ to terminal i, where f_i is deterministic and N_i 's are independent channel noises that satisfy $P_{N_V|X_V} = \prod_{i \in V} P_{N_i}$. (Note that the observation Y_i of terminal i may depend on the input specified by other terminals.) Since Z_i captures all the information in the input and output signals associated with terminal i, the MMI among Z_i 's reflects the mutual information among the terminals, and so we can cluster the terminals accordingly. The MMI has the following special form:

Proposition 4.3 ([52, Proposition 2.1]) For the emulated source network in Definition 4.3,

$$I(\mathsf{Z}_V) = \min_{\mathcal{P} \in \Pi'(V)} \frac{1}{|\mathcal{P}| - 1} \sum_{C \in \mathcal{P}} I(\mathsf{X}_{V \setminus C} \wedge \mathsf{Y}_C | \mathsf{X}_C), \quad \textbf{(4.15)}$$

which is an achievable secret key rate under a multiterminal channel model [91, §II-B].

In network information theory, the mapping to the conditional mutual information, $C \mapsto I(X_{V \setminus C} \land Y_C | X_C)$ in (4.15), is a cut function (evaluated at the cut set C) that measures the total amount of information flow from the terminals in $V \setminus C$ to the terminals in C. Similar to the usual graphical cut function, this cut function is also submodular with respect to C. Therefore, the info-clustering algorithm will return the non-singleton subsets in the PSP of the cut function as the clusters.

A special case of interest is when the channel $P_{Y_V|X_V}$ consists of a set of broadcast links among the subsets of the terminals. More precisely, consider a hypergraph with vertex set V, edge set E, and edge function $\phi: E \to 2^V \setminus \{\emptyset\}$. Each hyperedge $e \in E$ is regarded as a broadcast link with sender specified by $\rho(e) \in \phi(e)$ and receivers being the terminals in $\phi(e) \setminus \{\rho(e)\}\$. ρ is called the orientation of the edge e. The capacity of the broadcast link is specified by the non-negative weight c(e). More precisely, the emulated source Z_V is defined using

$$\begin{split} \mathsf{X}_i := \left(\mathsf{X}_i^e \mid e \in E, i = \rho(e)\right) & \text{for } i \in V \quad \textbf{(4.16a)} \\ \mathsf{Y}_j := \left(\mathsf{Y}_j^e \mid e \in E, j \neq \rho(e) \in (e)\right) & \text{for } j \in V, \quad \textbf{(4.16b)} \end{split}$$

$$Y_{i} := (Y_{i}^{e} \mid e \in E, j \neq \rho(e) \in (e))$$
 for $j \in V$, (4.16b)

and the input-output relationship of each broadcast link e:

$$\mathsf{Y}_{j}^{e} = \mathsf{X}_{\rho(e)}^{e}$$
 $\forall j \in \phi(e), \text{ and}$ (4.16c)

$$\log |X_{\rho(e)}^e| = c(e) \tag{4.16d}$$

where the first equation says that the outputs of the broadcast link are equal to its input, and the second equation means that the capacity c(e) is the log cardinality of the input alphabet set, which is the maximum amount of information that can be sent across the broadcast link. For instance, such a broadcast link can be used as a simple model for the physical connection between neurons because a neuron broadcasts signals to one or more neurons through the gap junctions and chemical synapses. The weight c(e) can be obtained from the number of synapses. More elaborate models, such as the interference link in [52], the ADT network in [92], and the matroidal network link model [72–74], can also be considered.

It is easy to argue that the MMI is maximized by the uniform input distribution, and the emulated source can be equivalently defined as follows without depending on the orientation ρ :¹⁰

Definition 4.4 (Broadcast Network [52, Definition 2.4])

A broadcast network with respect to the hypergraph $H:=(V,E,\phi)$ is defined as

$$Z_i := \{ Z^e : e \in E, i \in \phi(e) \}$$
 for $i \in V$, (4.17)

with
$$(Z^e: e \in E)$$
 uniformly distributed and $H(Z^e) = c(e)$. \Box

The fact that the source model does not depend on the orientation ρ means that one needs not distinguish between directed and undirected links for info-clustering. For instance, even though the gap junction in neurons is undirected and the chemical synapses are directed, the direction does not affect the clustering. This is because each link, directed or not, leads to a piece of information shared *symmetrically* among both the sender and the receivers.

The choice of the uniform input distribution can also be justified more rigorously. In the secret key agreement problem under the channel model [91, \S VI-B], the uniform distribution on the input was shown to achieve the secrecy capacity, which is precisely the MMI $I(Z_V)$. Furthermore, the MMI can also be written in the form of a max-flow min-cut expression that characterizes the maximum multicast rate of network coding [53]. The MMI can be written in terms of the directed cut function for the hypergraph:

Proposition 4.4 ([52, Proposition 2.4]) The MMI of the broadcast network (4.17) is

$$I(\mathsf{Z}_{V}) = \min_{\mathcal{P} \in \Pi'(V)} \frac{\sum_{C \in \mathcal{P}} \sum_{e \in \delta_{H^*}^{-}(C)} c(e)}{|\mathcal{P}| - 1}$$

$$= \min_{\mathcal{P} \in \Pi'(V)} \frac{\sum_{e \in E} c(e) (|\pi_{\mathcal{P}}(\phi(e))| - 1)}{|\mathcal{P}| - 1}, \quad \textbf{(4.18b)}$$

where $H^* := (V, E, \phi, \rho)$ is a hypergraph of H with an arbitrary choice of the orientation ρ for each edge,

$$\delta_{H^*}^-(C) := \{e \in E \mid \rho(e) \in C^c \not\supseteq \phi(e)\} \tag{4.19}$$

$$\pi_{\mathcal{P}}(\phi(e)) := \{ C \cap \phi(e) \mid C \in \mathcal{P} \} \setminus \{\emptyset\}$$
 (4.20)

are the set of in-coming edges into C and the partition of e respectively. $\ \ \Box$

Even though the MMI does not depend on the orientation ρ , as shown in (4.18b), it is informative to consider the alternative form in (4.18a) that is stated with an arbitrary choice of the

orientation ρ . In particular, from (4.18a), we can deduce that:

Proposition 4.5 For the hypergraphical source Z_V defined above, the clusters are the non-singleton subsets from the PSP of the submodular in-cut function $C \mapsto c(\delta_{H^*}^-(C))$.

Indeed, the physical connectome may be simplified as a graph instead of a hypergraph because the polyadic synapses that connect one neuron to multiple neurons are rare [13, 14]. In the special case when the hypergraph is a graph $G=(V,E,\theta)$ with $|\theta(e)|=2$, the broadcast network in (4.17) reduces to the graphical network called the pairwise independent network (PIN) [69]. it is straightforward to show that the MMI in (4.18a) can be further written as the strength of the graph:

$$I(\mathsf{Z}_V) = \min_{\mathcal{P} \in \Pi'(V)} \frac{\sum_{C \in \mathcal{P}} c(\delta_G(C))}{2(|\mathcal{P}| - 1)}$$
(4.21a)

where $C \mapsto c(\delta_G(C))$ is the submodular undirected cut function with the edge cut

$$\delta_G(C) := \{ e \in E \mid \emptyset \neq C \cap \theta(e) \subseteq C \}. \tag{4.21b}$$

The factor of 2 in the denominator of (4.21a) comes from the fact that an edge that crosses \mathcal{P} overlap with two disjoint subsets in \mathcal{P} , so it is doubly counted in the numerator. Since the factor does not affect the PSP, we have the following result:

Proposition 4.6 For the graphical source Z_V defined above, the clusters are the non-singleton subsets from the PSP of the undirected cut function $C \mapsto c(\delta_G(C))$.

By the Tutte–Nash-Williams tree packing theorem, the strength of a graph has the meaningful interpretation as the maximum amount of fractional tree packings of the graph [71, 93], which can also be extended to more general notion of partition connectivity for hypergraphs [52, 94, 95]. It can be shown that the principal sequence for graphs correspond to successive packing of forests, with the first critical value being the strength of the graph and the last critical value being the fractional arboricity, defined as the maximum amount of forests one can fractionally pack in the graph [96].

V. APPLICATIONS TO BIOLOGICAL DATASETS

In this section, we provide some discussions on how infoclustering can be used for the clustering of genes and neurons. For concreteness, we will describe some available datasets, and explain what one may potentially learn from them.

A. Gene clusering

As described in Section IV-A, the clustering by MIRN [3] is a special case of info-clustering under the Chow–Liu tree approximation. Therefore, the experimental results in [3] can be regarded as preliminary results of info-clustering, which may potentially be improved by considering higher-order correlation beyond pairwise mutual information as shown in Example 4.1.

 $^{^{10}}$ The result of [52] is modified slightly to include edge weight c.

The work in [3] considered the dataset from [97], which involves 2467 genes of a species of yeast called saccharomyces cerevisiae. The expression level of each gene was measured under 79 different conditions, including different stages of the cell cycle, temperatures, and time points. With V denoting the set of all genes, the different expression levels of gene $i \in V$ were regarded as i.i.d. realizations of a random variable Z_i that can be used for info-clustering. Towards this end, the pairwise mutual information $I(Z_i \wedge Z_j)$ between genes i and j was estimated using the empirical joint distribution of Z_i and Z_j after uniform quantization (since the expression levels are real-valued). We note that the empirical entropy after quantization can also be approximated without computing the empirical distribution [98].

Similarly, the MMI beyond the pairwise mutual information can be estimated from the empirical distribution of the quantized expression levels. The idea is to compute the empirical entropies of subsets of random variables Z_B after quantization, and use them in (3.3a) to estimate the MMI. The MMI of the quantized random variables is shown to approach the MMI of the continuous random variables in [91, Appendix B], and the details of the quantization can be found therein. However, computing the empirical joint distribution of a subset of random variables or estimating the joint entropy from the data samples takes exponential time with respect to the size of the subset [98]. This seems to suggest that some heuristics might be needed to tackle the problem of estimating the MMI. For example, an approach considered in [43] was to use a file compression algorithm to return the file size after compressing the data associated with the subset of random variables. Alternatively, one may consider other model reduction techniques so that the simplifying assumption made is clear.

B. Physical connectome

As described in Section IV-C, info-clustering can be specialized to cluster graphical networks. When applied to the physical connectome of neurons, it can identify clusters of tightly connected neurons, as well as the important intercluster connections, the damage of which may cause detrimental effects. While the physical connectome may not represent the functional connectome, i.e., the task-specific stimulation patterns of the neurons, the resulting clusters may be studied in conjunction with the functional connectome to understand how neurons work together to manifest consciousness and to carry out brain functions [76].

Instead of looking at the human connectome data, as an illustration, we will consider a small and nearly complete physical connectome dataset in [14] for a small creature called nematode *C. elegans*. A set of 279 neurons in the somatic nervous system was considered, and the dataset is in the form of an adjacency matrix, recording the total number of synaptic contacts (gap junction and chemical synapses) between every pairs of neurons.

The adjacency matrix defines the weighted graph for infoclustering. We can compute the cut function of the graph from the adjacency matrix, and then obtain the desired clusters from the PSP of the cut function as described in Section IV-C. It is worth pointing out that, owing to the similarity between MAC clustering algorithm of [61] and info-clustering as pointed out in Appendix D, we expect the performance of info-clustering to be close to that of the MAC clustering for cut functions, which was shown in [61] to be competitive with the existing leading algorithms for clustering graphical networks.

VI. MEASURES OF INTEGRATION AND SEGREGATION

Based on the info-clustering paradigm, we can derive some meaningful measures to describe the clustering solutions. For example, the MMI $I(\mathsf{Z}_V)$ naturally measures how integrated the objects in V are. This is because the more interaction among the objects in a system, the larger the mutual information they share. Such an argument is supported by the concrete operational meanings of $I(\mathsf{Z}_V)$ as the secrecy capacity for the multiterminal secret key agreement problem, the multicast throughput for the network coding problem, and the partition connectivity for hypergraphical or graphical models.

A measure called the integrated information was proposed in [76, (2B)] to measure how integrated a system is. This measure may appear similar to the MMI in the sense that it is defined as the divergence from the joint distribution of the overall system to the product of the marginal distributions of some subsystems. (The subsystems are obtained by partitioning the system according to what is called the minimum information partition.) However, there are two fundamental differences between the two definitions:

- 1) In contrast to the fundamental partition $\mathcal{P}^*(\mathsf{Z}_V)$, the minimum information partition is obtained by an additional normalization factor that forces the partition to be more balanced. However, this additional factor makes the problem intractable. This is similar to the normalized-cut minimization problem, which is NP-hard to solve. In general, the cluster size has nothing to do with the amount of information mutual to the elements in the cluster. Thus, this additional factor can steer the clustering procedure away from finding a small cluster that has high mutual information.
- 2) Unlike the MMI, the divergence expression in the integrated information with respect to the minimum information partition, say P, is not normalized by the factor |P| 1. As we have described using the concept of residual independence relation, the factor is needed to account for the double counting in the mutual information in each subsystem. Therefore, the integrated information does not have the desired information-theoretic meaning.

In addition, the integrated information is computed from the a posteriori probability that is marginalized using a uniform input distribution. However, the a posteriori distribution can be viewed as a channel, which can be handled by info-clustering as in the emulated source model in §IV-C. Moreover, instead choosing a uniform distribution by assumption, we can justify such a choice as one that maximizes the MMI in the case of the hypergraphical model.

A measure of segregation was also proposed in [5]. However, there are two issues of the formulation:

- The measure assumes a jointly Gaussian distribution rather than a general source distribution, so it is unclear how the measure can capture a more general correlation structure.
- 2) The measure is normalized using a noise variance, which is chosen in an ad-hoc manner without a concrete interpretation. The normalization is also done in a way different from the usual signal-to-noise ratio for the MIMO Gaussian channel [99].

We believe that the measure of segregation is simply a dual to the measure of integration, i.e., the MMI can be used to measure segregation and there is no need to define another fundamental quantity. More precisely, we can measure the segregation of a cluster C of V as

$$1 - \frac{I(\mathsf{Z}_V)}{I(\mathsf{Z}_C)} \in (0, 1]. \tag{6.1}$$

The index is non-negative because $I(\mathsf{Z}_C) > I(\mathsf{Z}_V)$ by the formulation (2.1a) of clusters, and it is upper bounded by 1 because of the non-negativity of the MMI. The index is large (the ratio $\frac{I(\mathsf{Z}_V)}{I(\mathsf{Z}_C)}$ is small) if the cluster C is more integrated than the entire set V, that is to say, C is more segregated from the rest of the nodes in $V \setminus C$. Depending on the application, one may further compute the average, minimum, or maximum segregation among a set of clusters to show how segregated the clusters are from each other.

VII. CONCLUSION

In this work, we proposed a new information-theoretic approach to clustering biological systems. In particular, we formulated the info-clustering paradigm and showed how it can be applied to study the human genome and connectome. Compared to the conventional algorithmic approaches, infoclustering follows a bottom-up theoretical approach for clustering. Rather than justifying the algorithm purely by data, which was shown to have many issues, we believe that it is more important to lay a rigorous mathematical theory before algorithmic simplifications. In particular, the info-clustering is formulated in a meaningful way without requiring any prior knowledge of the number of clusters nor an initial solution to start the clustering algorithm. The solution is shown to be unique, with meaningful information-theoretic interpretations as well as an elegant mathematical structure for efficient computation.

More precisely, we formulated the clustering problem (2.1a) using a threshold test on the MMI, and showed that the solution is hierarchical under a simple, but general, property (2.7) of the MMI, which also holds for some other choices of multivariate mutual information measures. The clustering solution is characterized by a finite set of critical values and their corresponding finite sets of clusters (2.5a). The formulation is different from the classical one in the sense that the set of clusters is not required a priori to form a partition. Instead, the set of all clusters is shown to be laminar (2.8) using the general property (2.7) of the MMI. Consequently, the complete clustering solution can be computed iteratively in Algorithm 1.

Using the precise definition (3.3a) of the MMI, we further showed that the clustering solution maps to the PSP of the entropy function. More precisely, the set (2.4) of critical values for info-clustering is precisely the set of critical values (3.14b) for the Dilworth truncation (3.12) of the residual entropy function (3.10). The corresponding set of clusters are the non-singleton subsets from the PSP (3.16) of the entropy function. This connection is non-trivial. It is based on the iterative relation (2.9) among the clusters and the iterative relation (3.17b) among the PSP. This connection not only enriches the abstract mathematical structure of the PSP with the concrete operational meanings from information theory, but also provides a concrete clustering solution that can be computed from the PSP in strongly polynomial time.

Indeed, we showed that info-clustering reduces to simpler and more practical algorithms under some special source models. Unlike the approximation algorithms of many clustering formulations, which focuses mainly on algorithmic simplicity, the model reduction for info-clustering specifies precisely what kind of correlation structure is assumed in return for the algorithmic simplicity. Consequently, we can verify whether the simplifying model applies to the case of interest, and identify the weaknesses of the simplified algorithm.

In particular, we showed that under the Markov tree model info-clustering reduces to the gene clustering algorithm by MIRN (4.8). If the correlation structure is not a Markov tree, the clustering by MIRN corresponds to the info-clustering algorithm under the Chow–Liu tree approximation. This shows that not only can info-clustering apply in practice to gene clustering, but it can also be used to justify existing techniques such as clustering by MIRN properly, with a concrete example showing how the Chow–Liu tree approximation may fail to capture the more complex multivariate correlation beyond the pairwise mutual information.

We also considered the usual Gaussian assumption, which simplifies the info-clustering solution to a clustering algorithm by the covariance matrix, or more specifically, the eigenvalues of the submatrices of the covariance matrix (Proposition 4.2.). This is a new spectral clustering technique that follows precisely from the info-clustering paradigm without any approximation.

For the study of the human connectome, we also examined the specification of info-clustering to the hypergraphical model, which can capture the possibility of polyadic physical connections among neurons. In this case, the solution reduces to the PSP of hypergraphs and graphs (Proposition 4.5 and 4.6), which can be computed more efficiently than the PSP of the entropy function of a general source model. In addition to the algorithmic simplicity, the solution also has a meaningful interpretation as the network information flow: Clusters are simply subnetworks that support large information flows.

Finally, using the info-clustering paradigm, we also demonstrated how the MMI can be used as a measure of the integration of a cluster, which can further be used to measure how segregated a cluster is from the other objects or clusters (6.1). The measures do not assume any particular source model or choice of parameters. Their values can be computed and justified from the info-clustering solution.

$\begin{array}{c} A \text{PPENDIX A} \\ \text{PROOF OF THEOREMS IN } \S II \end{array}$

1. Proof of Theorem 2.1

The following is a necessary and sufficient condition for a set to be a cluster:

Proposition A.1 A non-empty non-singleton subset of V is a cluster of Z_V if and only if it cannot be enlarged without reducing multivariate information quantity, i.e.,

$$B \in \mathcal{C}(\mathsf{Z}_V) \iff I(\mathsf{Z}_{B'}) < I(\mathsf{Z}_B) \quad \forall B' \subseteq V : B' \supseteq B \quad \textbf{(A.1)}$$

for
$$B \subseteq V : |B| > 1$$
.

PROOF Suppose the R.H.S. of (A.1) holds. Then, we have $B \in \mathcal{C}_{I(\mathsf{Z}_B)^-}(\mathsf{Z}_V)$ (and therefore the L.H.S. of (A.1)) because B is a maximal subset with multivariate information at least the threshold $I(\mathsf{Z}_B)^-$.

Suppose the R.H.S. of (A.1) does not hold, i.e., there exists a proper superset $B' \supseteq B$ with $I(\mathsf{Z}_{B'}) \ge I(\mathsf{Z}_B)$. It follows that $I(\mathsf{Z}_{B'}) > \gamma$ whenever $I(\mathsf{Z}_B) > \gamma$ and so B cannot be maximal in (2.1a) for any threshold $\gamma \in \mathbb{R}$. Therefore, the L.H.S. of (A.1) does not hold either.

Proposition A.2 $\Gamma(\mathsf{Z}_V) = \{I(\mathsf{Z}_B) : B \in \mathcal{C}(\mathsf{Z}_V)\}$, consisting of the multivariate information quantities of the clusters. \square

PROOF If $B \in \mathcal{C}_{\gamma}(\mathsf{Z}_V)$ for some $\gamma \in \mathbb{R}$, then $B \in \mathcal{C}_{I(\mathsf{Z}_B)^-}(\mathsf{Z}_V) \setminus \mathcal{C}_{I(\mathsf{Z}_B)^+}(\mathsf{Z}_V)$ and so $I(\mathsf{Z}_B) \in \Gamma(\mathsf{Z}_V)$ by definition (2.4).

Consider any $\gamma \in \Gamma(\mathsf{Z}_V)$. Then, by (2.4), we have one of the following two cases:

- 1) There exists $B \in \mathcal{C}_{\gamma^{-}}(\mathsf{Z}_{V}) \setminus \mathcal{C}_{\gamma^{+}}(\mathsf{Z}_{V})$, i.e., a cluster that disappears at γ . We must have $\gamma^{-} < I(\mathsf{Z}_{B}) \le \gamma^{+}$ by (2.1a), and so $\gamma = I(\mathsf{Z}_{B})$ as desired.
- 2) There exists $B \in \mathcal{C}_{\gamma^+}(\mathsf{Z}_V) \setminus \mathcal{C}_{\gamma^-}(\mathsf{Z}_V)$, i.e., a cluster that appears at γ . By Proposition A.1, this happens only if there is a larger cluster $B' \supsetneq B$ that disappears at γ , which reduces to the previous case.

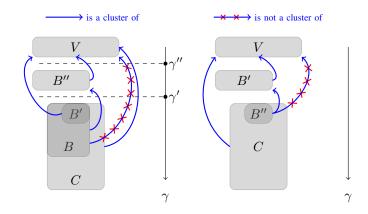
We are now ready to prove Theorem 2.1. For $\gamma < I(\mathsf{Z}_V)$, the set $\mathcal{C}_{\gamma}(\mathsf{Z}_V)$ contains V by definition (2.1a). Indeed, V is the unique cluster because it is the largest subset of V. It follows that $\gamma_1 = I(\mathsf{Z}_V)$ and $\mathcal{C}_{\gamma}(\mathsf{Z}_V) = \{V\}$ for $\gamma < \gamma_1$.

By the definition of critical values (2.4), the cluster $\mathcal{C}_{\gamma}(\mathsf{Z}_V)$ must remain unchanged for γ between consecutive critical values. Therefore, $\mathcal{C}_{\gamma}(\mathsf{Z}_V) = \mathcal{C}_{\gamma_i}(\mathsf{Z}_V)$ for $\gamma \in [\gamma_i, \gamma_{i+1})$ and $1 \leq i < N$.

When $\gamma \geq \max_{B \subseteq V: |B| > 1} I(\mathsf{Z}_B)$, we have $\mathcal{C}_{\gamma}(\mathsf{Z}_V) = \emptyset$ because no solution B to (2.1a) can have $I(\mathsf{Z}_B)$ larger than the maximum (2.5c), which must therefore be the last critical value. Since the clusters in $\mathcal{C}_{\gamma_{N-1}}(\mathsf{Z}_V)$ remain to be clusters for $\gamma \in [\gamma_{N-1}, \gamma_N)$, they must achieve the maximum value of the multivariate information quantity.

2. Proof of Theorem 2.2

To help understand the proof, the readers may refer to Fig. 7a for a summary of the relationship among the clusters in



(a) Theorem 2.2. Existence of C (b) Theorem 2.4. Existence of contradicts $I(Z_{B''})$ is maximized. C contradicts laminarity (2.8).

Fig. 7: Illustration of clusters in proofs by contradiction.

the proof. Let B'' be a cluster of Z_V that is a proper superset of B', i.e., we have

$$\gamma'' \in \mathbb{R}, B'' \supseteq B', B'' \in \mathcal{C}_{\gamma''}(\mathsf{Z}_V). \tag{A.2}$$

Such a choice of B'' exists because V is a feasible choice, but there can be multiple feasible choices. We choose any one that maximizes $I(Z_{B''})$. We will show that B'' and γ'' satisfy (2.6).

Note that $I(\mathsf{Z}_{B''}) > \gamma''$ and $I(\mathsf{Z}_{B'}) > \gamma'$ because $B'' \in \mathcal{C}_{\gamma''}(\mathsf{Z}_V)$ and $B' \in \mathcal{C}_{\gamma'}(\mathsf{Z}_V)$ respectively. We also have $I(\mathsf{Z}_{B''}) \leq \gamma'$ because, otherwise, $B'' \supsetneq B'$ contradicts the maximality of $B' \in \mathcal{C}_{\gamma'}(\mathsf{Z}_V)$. Altogether, we have

$$\gamma'' \stackrel{\text{(a)}}{<} \gamma_1(\mathsf{Z}_{B''}) \stackrel{\text{(b)}}{=} I(\mathsf{Z}_{B''}) \stackrel{\text{(c)}}{\leq} \gamma' \stackrel{\text{(d)}}{<} I(\mathsf{Z}_{B'})$$

where the equality (b) is by (2.5b).

(a), (b) and (c) implies $\gamma'' < \gamma'$ as desired by (2.6). Furthermore, $I(\mathsf{Z}_{B''}) < I(\mathsf{Z}_{B'})$ (from (c) and (d)) and the fact that B' is a cluster of Z_V implies that B' is also a cluster of $\mathsf{Z}_{B''}$. However, to establish (2.6), we need to show the stronger statement that $B' \in \mathcal{C}_{\gamma_1}(\mathsf{Z}_{B''})$.

Now, $C_{\gamma_1}(\mathsf{Z}_{B''}) \neq \emptyset$ because we at least have $B' \subseteq B''$ with $I(\mathsf{Z}_{B'}) > \gamma_1(\mathsf{Z}_{B''})$ (from (b), (c) and (d)). Therefore, we have $B' \in C_{\gamma_1}(\mathsf{Z}_{B''})$ as desired by (2.6) unless there exists

$$B \in \mathcal{C}_{\gamma_1}(\mathsf{Z}_{B''}) : B' \subsetneq^{(e)} B.$$

Suppose to the contrary that such a subset B exists. Then,

$$\gamma_1(\mathsf{Z}_{B''}) \overset{(\mathsf{f})}{<} I(\mathsf{Z}_B)$$

because $B \in \mathcal{C}_{\gamma_1}(\mathsf{Z}_{B''})$. We will show that, regardless of whether B is a cluster of Z_V or not, there is a contradiction to the maximality of $I(\mathsf{Z}_{B''})$ among all feasible subsets B'' satisfying (A.2).

- 1) Suppose $B \in \mathcal{C}(\mathsf{Z}_V)$. Then, (A.2) holds with B'' replaced by B and so $I(\mathsf{Z}_B) > I(\mathsf{Z}_{B''})$ (from (b) and (f)) contradicts the choice of B''.
- 2) Suppose $B \notin \mathcal{C}(\mathsf{Z}_V)$. Then, there exists

$$C \in \mathcal{C}(\mathsf{Z}_V) : B \subsetneq^{(\mathsf{g})} C, I(\mathsf{Z}_C) \stackrel{(\mathsf{h})}{\geq} I(\mathsf{Z}_B).$$

This is the complete scenario shown in Fig. 7a. It follows that $C \supseteq B'$ (from (e) and (g)) and $I(\mathsf{Z}_C) > \gamma''$ (from (a), (b), (f) and (h)). Therefore, (A.2) holds with B'' replaced by $C,^{11}$ but $I(\mathsf{Z}_C) > I(\mathsf{Z}_{B''})$ (from (b), (f) and (h)) contradicts the choice of B''.

3. Proof of Theorem 2.4

It suffices to show that for any cluster B' of Z_V , we have

$$C(\mathsf{Z}_{B'}) = \{ B \in C(\mathsf{Z}_V) : B \subseteq B' \}, \tag{A.3}$$

i.e., a cluster of $\mathsf{Z}_{B'}$ must be a cluster of Z_V . Then, from (A.3), a simple induction on γ over the finite set $\Gamma(\mathsf{Z}_V)$ will immediately lead to (2.9).

To prove (A.3), consider any B that is in the R.H.S. of (A.3). Then, by definition (2.1a), B is a maximal subset of V with $I(\mathsf{Z}_B) > \gamma$ for some $\gamma \in \mathbb{R}$. Since $B \subseteq B' \subseteq V$, we also have that B is a maximal subset of B' with $I(\mathsf{Z}_B) > \gamma$, i.e., B is a cluster of $\mathsf{Z}_{B'}$, belonging to the set in the L.H.S. of (A.3). Therefore, \supseteq holds for (A.3).

It remains to show the reverse inclusion \subseteq for (A.3). To help understand the proof, the readers may refer to Fig. 7b for a summary of the relationship among the clusters. Suppose to the contrary that a cluster B'' of $\mathsf{Z}_{B'}$ is not a cluster of Z_V . Note that $B'' \neq B'$ because B' is a cluster of Z_V but B'' is not. Therefore, we have the strict inequality $I(\mathsf{Z}_{B''}) > I(\mathsf{Z}_{B'})$. By (A.1), there exists

$$C \in \mathcal{C}(\mathsf{Z}_V) : B'' \subsetneq C, I(\mathsf{Z}_C) \geq I(\mathsf{Z}_{B''}).$$

which implies that

$$C \neq B'$$
 and $I(\mathsf{Z}_C) \geq I(\mathsf{Z}_{B''}) > I(\mathsf{Z}_{B'}).$

We will show that $C \cap B' \notin \{\emptyset, C, B'\}$, contradicting laminarity (2.8).

- 1) $C \cap B' \neq \emptyset$ because both C and B' contains the non-empty set B''.
- 2) $C \cap B' \neq C$ or simply $C \not\subseteq B'$ because, if to the contrary that $C \subseteq B'$, then $B'' \subsetneq C$ and $I(\mathsf{Z}_C) \geq I(\mathsf{Z}_{B''})$ assumed above contradicts the fact that B'' is a cluster of $\mathsf{Z}_{B'}$.
- 3) $C \cap B' \neq B'$ or simply $B' \not\subseteq C$ because, if to the contrary that $B' \subseteq C$, then $C \neq B'$ and $I(\mathsf{Z}_C) > I(\mathsf{Z}_{B'})$ derived above contradict the fact that B' is a cluster of Z_V .

APPENDIX B PROOF OF THEOREM 3.3

The line segment preceding the $p_i = (\gamma_i, y_i)$ is $h_{\gamma}[\mathcal{P}_{i-1}]$ since $\mathcal{P}_{i-1} = \max \Pi_i$ by (3.16b). The curve after p_i has a strictly smaller slope than $-|\mathcal{P}_{i-1}|$ by the definition of a turning point, and so $\gamma = \gamma_i$ is a solution to

$$h_{\gamma}[\mathcal{P}_{i-1}] = \min_{\mathcal{P} \in \Pi(V): |\mathcal{P}| > |\mathcal{P}_{i-1}|} h_{\gamma}[\mathcal{P}]$$

¹¹We can also redefine γ'' to be the previous value of $I(Z_{B''})$, in which case (a) need not be used to argue the contradiction.

where the R.H.S. corresponds to $\hat{h}_{\gamma}(V)$ for $\gamma \geq \gamma_i$, with the set of optimal partitions at $\gamma = \gamma_i$ being $\Pi_i \setminus \{\mathcal{P}_{i-1}\}$. Rearranging the terms, it follows that

$$\begin{split} 0 &= \min_{\mathcal{P} \in \Pi(V): |\mathcal{P}| > |\mathcal{P}_{i-1}|} h_{\gamma_i}[\mathcal{P}] - h_{\gamma_i}[\mathcal{P}_{i-1}] \\ &= \min_{\mathcal{P} \in \Pi(V): |\mathcal{P}| > |\mathcal{P}_{i-1}|} h[\mathcal{P}] - h[\mathcal{P}_{i-1}] - \gamma_i (|\mathcal{P}| - |\mathcal{P}_{i-1}|) \\ &= \min_{\mathcal{P} \in \Pi(V): |\mathcal{P}| > |\mathcal{P}_{i-1}|} \frac{h[\mathcal{P}] - h[\mathcal{P}_{i-1}]}{|\mathcal{P}| - |\mathcal{P}_{i-1}|} - \gamma_i, \end{split}$$

which implies (3.17a). The last expression is obtained from the previous by multiplying $|\mathcal{P}| - |\mathcal{P}_{i-1}| \geq 1$, which preserves both the minimum value of 0 and the set of minimum solutions, namely $\Pi_i \setminus \{\mathcal{P}_{i-1}\}$. Since $\mathcal{P}_{i-1} = \max \Pi_i$, every optimal solution is finer than \mathcal{P}_{i-1} , and so it does not lose optimality to impose $\mathcal{P} \prec \mathcal{P}_{i-1}$ in (3.17a) or equivalently,

$$\mathcal{P} = \bigcup\nolimits_{C \in \mathcal{P}_{i-1}} \mathcal{P}^C \quad \text{ for some } \mathcal{P}^C \in \Pi(C).$$

Using the above, (3.17a) can be rewritten as

$$\gamma_{i} = \min_{\mathcal{P} \prec \mathcal{P}_{i-1}} \frac{\sum_{C \in \mathcal{P}_{i-1}} \left[h[\mathcal{P}^{C}] - h(C) \right]}{\sum_{C \in \mathcal{P}_{i-1}} (|\mathcal{P}^{C}| - 1)}$$

$$\leq \min_{C \in \mathcal{P}_{i-1}: |C| > 1} \min_{\mathcal{P}^{C} \in \Pi'(C)} \frac{D(P_{\mathsf{Z}_{C}} \| \prod_{C' \in \mathcal{P}^{C}} P_{\mathsf{Z}_{C'}})}{|\mathcal{P}^{C}| - 1}$$

The last expression is obtained by imposing $\mathcal{P}^C = \{C\}$ for all but one $C \in \mathcal{P}_{i-1}$ with |C| > 1, and substituting $h[\mathcal{P}^C] - h(C) = D(P_{Z_C} \| \prod_{C' \in \mathcal{P}^C} P_{Z_{C'}})$. It is equal to the R.H.S. of (3.17b) by the definition of I (3.3a). (The existence of a block $C \in \mathcal{P}_{i-1} : |C| > 1$ is guaranteed by Proposition 3.3 since \mathcal{P}_{i-1} is coarser than the partition into singletons, i.e., $\mathcal{P}_{i-1} \succ \mathcal{P}_N$.) To show the reverse inequality, let γ be the R.H.S. of (3.17b) and Q be the set of optimal solutions. Then, for all $C \in \mathcal{P}_{i-1}$ and $\mathcal{P}^C \in \Pi(C)$, we have $h[\mathcal{P}^C] - h(C) \geq \gamma(|\mathcal{P}^C| - 1)$ and so

$$\gamma_{i} = \min_{\mathcal{P} \prec \mathcal{P}_{i-1}} \frac{\sum_{C \in \mathcal{P}_{i-1}} \left[h[\mathcal{P}^{C}] - h(C) \right]}{\sum_{C \in \mathcal{P}_{i-1}} (|\mathcal{P}^{C}| - 1)}$$
$$\geq \frac{\sum_{C \in \mathcal{P}_{i-1}} \gamma(|\mathcal{P}^{C}| - 1)}{\sum_{C \in \mathcal{P}_{i-1}} (|\mathcal{P}^{C}| - 1)} = \gamma$$

Equality happens if and only if, for all $C \in \mathcal{P}_{i-1}$, either we have $\mathcal{P}^C = \{C\}$ or we have $C \in \mathcal{Q}$ and $\mathcal{P}^C \in \Pi^*(\mathsf{Z}_C)$. This implies $\mathcal{Q} = \mathcal{P}_{i-1} \setminus \mathcal{P}_i$ and therefore (3.19).

$\label{eq:appendix} \mbox{Appendix C} \\ \mbox{Proof of Theorems in } \S IV\text{-}A \\ \mbox{}$

1. Proof of Theorem 4.1

The proof relies on Theorem 4.3 proved in Appendix C-2, which relies on Theorem 4.2 proved in Appendix C-3. First note that to prove Theorem 4.1, it suffices to prove $\mathcal{P}(T_\gamma) = \mathcal{P}(K_\gamma)$, because this implies (4.8) by (4.11). In other words, we want to show that the vertex sets of the connected components of T_γ are the same as those of K_γ , i.e., for any $i,j \in V$, we have $i \sim_{T_\gamma} j$ if and only if $i \sim_{K_\gamma} j$. The direct part (only if) is obvious, because T_γ is a subgraph of K_γ . To prove the converse (if) part, we will use the following exchange property for spanning trees.

Lemma C.1 ([71, Theorem 39.12]) Consider two spanning trees T and T' on the vertex set V. For any $e \in \mathcal{E}(T) \setminus \mathcal{E}(T')$, there exists $e' \in \mathcal{E}(T') \setminus \mathcal{E}(T)$ such that the graph T with e replaced by e', denoted as T - e + e', is a spanning tree. \Box

Now, suppose to the contrary that $i \not\sim_{T_{\gamma}} j$ but $i \sim_{K_{\gamma}} j$. Let \mathcal{E}' be the set of edges in a path from i to j in K_{γ} . Let e be an edge in the path from i to j in T but with $I(\mathsf{Z}_e) \leq \gamma$, and therefore not in $\mathcal{E}(K_{\gamma})$ nor \mathcal{E}' . Such an edge exists by the assumption $i \not\sim_{T_{\gamma}} j$.

Let G be the graph with edge set $\mathcal{E}(T)\cup\mathcal{E}'\setminus\{e\}$. There exists a spanning tree T' of G since G is connected, which follows from the facts that T is spanning and \mathcal{E}' connects the incident vertices of the removed edge e. Since $e\in\mathcal{E}(T)\setminus\mathcal{E}(T')$, we have by Lemma C.1 that there exists $e'\in\mathcal{E}(T')\setminus\mathcal{E}(T)$ such that T-e+e' is a spanning tree. The tree T-e+e' has a larger weight than T because $I(\mathsf{Z}_{e'})>\gamma\geq I(\mathsf{Z}_e)$, as $e'\in\mathcal{E}(T')\setminus\mathcal{E}(T)\subseteq\mathcal{E}'\subseteq\mathcal{E}(K_\gamma)$. This contradicts the maximality of $T\in T^*(\mathsf{Z}_V)$. We have thus completed the proof of Theorem 4.1.

2. Proof of Theorem 4.2

We remark that while the above arguments are purely of a graph-theoretical nature, in proving Theorems 4.2, and subsequently Theorem 4.3, we rely on some information-theoretic properties of the MMI (3.3a). The following is a lower bound on $I_{\mathcal{P}}$ specific to the dependency-tree distributions:

Lemma C.2 Consider the notation in (4.5). For $\mathcal{P} \in \Pi'(V)$,

$$I_{\mathcal{P}}(\mathsf{Z}_V^T) \ge \frac{1}{|\mathcal{E}_{\mathcal{P}}|} \sum_{e \in \mathcal{E}_{\mathcal{P}}} I(\mathsf{Z}_e)$$
 (C.1)

where $\mathcal{E}_{\mathcal{P}} := \{ \{ \min C, \mathsf{p}_{\min C} \} \colon 1 \not\in C \in \mathcal{P} \}$. Moreover, equality holds if we have

$$p_i \in C$$
 for all $C \in \mathcal{P}$ and $i \in C$ such that $i \neq \min C$.

(C.2)

Note that (C.2) simply means that the subgraph of T induced on each $C \in \mathcal{P}$ is a subtree.

PROOF By (3.4a), we can express $I_{\mathcal{P}}(\mathsf{Z}_V^T)$ (3.3b) in terms of the entropies as $\frac{\sum_{C\in\mathcal{P}}H(\mathsf{Z}_V^T)-H(\mathsf{Z}_V^T)}{|\mathcal{P}|-1}$. It follows from the definition of $\mathcal{E}_{\mathcal{P}}$ that $|\mathcal{E}_{\mathcal{P}}|=|\mathcal{P}|-1$, and so

$$\begin{split} &|\mathcal{E}_{\mathcal{P}}|I_{\mathcal{P}}(\mathsf{Z}_{V}^{T}) \\ &= \sum_{C \in \mathcal{P}} \underbrace{H(\mathsf{Z}_{C}^{T})}_{(\mathbf{i})} - \underbrace{H(\mathsf{Z}_{V}^{T})}_{(\mathbf{i}\mathbf{i})} - \underbrace{\sum_{i \in V} \underbrace{H(\mathsf{Z}_{i}^{T}|\mathsf{Z}_{\{j \in C|j < i\}}^{T})}_{(\mathbf{i}\mathbf{i}\mathbf{i})} - \underbrace{\sum_{i \in V} \underbrace{H(\mathsf{Z}_{i}^{T}|\mathsf{Z}_{\{j \in V|j < i\}}^{T})}_{(\mathbf{i}\mathbf{i}\mathbf{i}\mathbf{i})} \\ &\stackrel{\text{(b)}}{\geq} \sum_{C \in \mathcal{P}} \left[H(\mathsf{Z}_{\min C}) + \sum_{i \in C \setminus \{\min C\}} H(\mathsf{Z}_{i}^{T}|\mathsf{Z}_{\mathsf{p}_{i}}^{T}) \right] \\ &- \sum_{i \in V} H(\mathsf{Z}_{i}^{T}|\mathsf{Z}_{\mathsf{p}_{i}}^{T}) \\ &= \sum_{C \in \mathcal{P}} \left[H(\mathsf{Z}_{\min C}) - H(\mathsf{Z}_{\min C}|\mathsf{Z}_{\mathsf{p}_{\min C}}) \right] \end{split}$$

which is equal to $\sum_{e \in \mathcal{E}_{\mathcal{P}}} I(\mathsf{Z}_e)$, completing the proof of (C.1). To obtain (a), we applied the chain rule (i) = (i)' and (ii) =

(ii)'. To obtain (b), we used (iv) $= H(\mathsf{Z}_{\mathsf{p}_i}^T|\mathsf{Z}_{\mathsf{p}_i}^T)$ by the Markov relation (4.5). (iii) $= H(\mathsf{Z}_{\min C})$ when $i = \min C$, and (iii) $\geq H(\mathsf{Z}_i^T|\mathsf{Z}_{\{j \in C \mid j < i\} \cup \{\mathsf{p}_i\}}^T) = H(\mathsf{Z}_i^T|\mathsf{Z}_{\mathsf{p}_i}^T)$ for $i > \min C$ by the Markov relation and the fact that conditioning reduces entropy. Equality holds if and only if (C.2) holds, again due to the Markov relation (4.5).

We are now ready to prove Theorem 4.2. By (C.1) in Lemma C.2,

$$I_{\mathcal{P}}(\mathsf{Z}_{V}^{T}) \geq \frac{1}{|\mathcal{E}_{\mathcal{P}}|} \sum_{e \in \mathcal{E}_{\mathcal{P}}} I(\mathsf{Z}_{e}) \overset{\text{(a)}}{\geq} \min_{e \in \mathcal{E}_{\mathcal{P}}} I(\mathsf{Z}_{e}) \overset{\text{(b)}}{\geq} \min_{e \in \mathcal{E}(T)} I(\mathsf{Z}_{e})$$

where (a) is because the minimum edge weight on the right is no larger than the average on the left; (b) is because $\mathcal{E}_{\mathcal{P}} \subseteq \mathcal{E}(T)$. The above implies \geq for (4.9) by the definition (3.3a) of MMI.

To prove the reverse inequality, let $\mathcal{P} := \mathcal{P}(T_{\gamma})$ and $\gamma := \min_{e \in \mathcal{E}(T)} I(\mathsf{Z}_e)$. We shall argue that:

$$I_{\mathcal{P}}(\mathsf{Z}_{V}^{T}) \stackrel{\text{(c)}}{=} \frac{1}{|\mathcal{E}_{\mathcal{P}}|} \sum_{e \in \mathcal{E}_{\mathcal{P}}} I(\mathsf{Z}_{e}) \stackrel{\text{(d)}}{=} \min_{e \in \mathcal{E}(T)} I(\mathsf{Z}_{e}).$$

- (c) is because the equality condition (C.2) holds. More
 precisely, every C∈ P is the vertex set of a connected
 component, and so, for all i∈ C, we have p_i∈ C unless
 i = min C.
- To argue (d), it suffices to show that $\mathcal{E}_{\mathcal{P}} = \mathcal{E}(T) \setminus \mathcal{E}(T_{\gamma})$, because $I(\mathsf{Z}_e) = \gamma$ for all $e \in \mathcal{E}(T) \setminus \mathcal{E}(T_{\gamma})$. For any $C \in \mathcal{P}$, we have $\mathsf{p}_{\min C} \not\in C$ because $\mathsf{p}_{\min C} < \min C$ by (4.5). Therefore, any edge $\{\min C, \mathsf{p}_{\min C}\} \in \mathcal{E}_{\mathcal{P}}$ is also in $\mathcal{E}(T) \setminus \mathcal{E}(T_{\gamma})$. Conversely, consider any edge $\{i, \mathsf{p}_i\} \in \mathcal{E}(T) \setminus \mathcal{E}(T_{\gamma})$ and $C \in \mathcal{P} : i \in C$ for some i > 1 in V. Suppose to the contrary that $\{i, \mathsf{p}_i\} \not\in \mathcal{E}_{\mathcal{P}}$. By the equality condition (C.2) proved earlier, we have $i = \min C$, contradicting $\{i, \mathsf{p}_i\} \not\in \mathcal{E}_{\mathcal{P}}$.

Now that (4.9) is proved, we have $\gamma = I(\mathsf{Z}_V^T)$ and $\mathcal{P} = \mathcal{P}(T_{I(\mathsf{Z}_V^T)})$. To prove (4.10), suppose to the contrary that $\mathcal{P}^*(\mathsf{Z}_V^T) \neq \mathcal{P}$, i.e., there exists $\mathcal{P}' \in \Pi^*(\mathsf{Z}_V)$ with $|\mathcal{P}'| > |\mathcal{P}|$. We have

$$\min_{e \in \mathcal{E}(T)} I(\mathsf{Z}_e) \stackrel{\text{(e)}}{=} \gamma \stackrel{\text{(f)}}{=} I(\mathsf{Z}_V^T) \stackrel{\text{(g)}}{=} I_{\mathcal{P}'}(\mathsf{Z}_V^T),$$

where (e) is by the definition of γ ; (f) is by (4.9); and (g) is because $\mathcal{P}' \in \Pi^*(\mathsf{Z}_V)$. By (C.1), every edge in $\mathcal{E}_{\mathcal{P}'}$ has weight γ , and so $\mathcal{E}_{\mathcal{P}'} \subseteq \mathcal{E}(T) \setminus \mathcal{E}(T_\gamma)$, implying $|\mathcal{E}_{\mathcal{P}'}| \leq |\mathcal{E}(T) \setminus \mathcal{E}(T_\gamma)|$. However,

$$|\mathcal{E}_{\mathcal{P}'}| = |\mathcal{P}'| - 1 > |\mathcal{P}| - 1 = |\mathcal{E}_{\mathcal{P}}| = |\mathcal{E}(T) \setminus \mathcal{E}(T_{\gamma})|,$$

which contradicts $\mathcal{E}_{\mathcal{P}'} \subseteq \mathcal{E}(T) \setminus \mathcal{E}(T_{\gamma})$. Note here that the last equality follows from the proof of (d) above. This completes the proof of Theorem 4.2

3. Proof of Theorem 4.3

We shall prove by induction that, for $1 \le i \le N$, γ_i is the i-th smallest value of $I(\mathsf{Z}_e)$ for $e \in \mathcal{E}(T)$, and $\mathcal{P}_i = \mathcal{P}(T_{\gamma_i})$ with \mathcal{P}_i defined in (3.18) for Z_C^T in place of Z_C . This will imply (4.11) by Corollary 3.1.

By (3.17b) and (3.18) with i = 1, we have $\gamma_1 = I(Z_V^T)$ and $\mathcal{P}_1 = \mathcal{P}^*(\mathsf{Z}_V^T)$. This implies the base case under (4.9) and (4.10), namely that, γ_1 is the smallest $I(\mathsf{Z}_e)$ and $\mathcal{P}_1 = \mathcal{P}(T_{\gamma_1})$.

Let T(C) be the subgraph of T induced on the subset $C \subseteq$ V of vertices. By (3.17b), for $1 < i \le N$,

$$\begin{split} \gamma_i &= \min_{C \in \mathcal{P}_{i-1} \colon |C| > 1} I(\mathsf{Z}_C^T) \\ &\stackrel{\text{(a)}}{=} \min_{C \in \mathcal{P}(T_{\gamma_{i-1}})} \min_{e \in \mathcal{E}(T(C))} I(\mathsf{Z}_e) \\ &\stackrel{\text{(b)}}{=} \min_{e \in \mathcal{E}(T_{\gamma_{i-1}})} I(\mathsf{Z}_e). \end{split}$$

Here, (a) is by the inductive hypothesis $\mathcal{P}_{i-1} = \mathcal{P}(T_{\gamma_{i-1}})$ as well as (4.9) that $I(\mathsf{Z}_C^T) = \min_{e \in \mathcal{E}(T(C))} I(\mathsf{Z}_e)$. (b) is because $\mathcal{E}(T_{\gamma})$ is the union of $\mathcal{E}(T(C))$ over $C \in \mathcal{P}(T_{\gamma})$ (for $\gamma =$ γ_{i-1}).

The above equalities implies that γ_i is the *i*-th smallest value of $I(Z_e)$ for $e \in \mathcal{E}(T)$ because the R.H.S. of (b) is, by the inductive hypothesis that γ_{i-1} is the (i-1)-st smallest value, and the fact that $T_{\gamma_{i-1}}$ contains all edges in $\mathcal{E}(T)$ with weights strictly larger than γ_{i-1} .

It remains to show $\mathcal{P}_i = \mathcal{P}(T_{\gamma_i})$. From (3.18), we have

$$\begin{split} & \mathcal{P}_i \overset{\text{(d)}}{=} \left(\mathcal{P}(T_{\gamma_{i-1}}) \setminus \mathcal{C}_{i-1}^* \right) \cup \bigcup_{C \in \mathcal{C}_{i-1}^*} \mathcal{P}(T_{\gamma_i}(C)) \\ & = \bigcup_{C \in \mathcal{P}_{i-1} \setminus \mathcal{C}_{i-1}^*} \{C\} \cup \bigcup_{C \in \mathcal{C}_{i-1}^*} \mathcal{P}(T_{\gamma_i}(C)) \\ & \overset{\text{(e)}}{=} \bigcup_{C \in \mathcal{P}(T_{\gamma_{i-1}})} \mathcal{P}(T_{\gamma_i}(C)) \overset{\text{(f)}}{=} \mathcal{P}(T_{\gamma_i}). \end{split}$$

Here, (d) is by applying (4.10) to (3.18). (e) is by rewriting $\mathcal{P}(T_{\gamma_{i-1}})\setminus\mathcal{C}_{i-1}^*$ as $\bigcup_{C\in\mathcal{P}_{i-1}\setminus\mathcal{C}_{i-1}^*}\{C\}$ and then applying $\{C\} = \mathcal{P}(T_{\gamma_i}(C))$ because $C \notin \mathcal{C}_{i-1}^*$ means that every edge of $T_{\gamma_i}(C)$ has weight strictly larger than γ_i by (3.17b). Finally, (f) follows from $\mathcal{E}(T_{\gamma_i})=\bigcup_{C\in \mathscr{Q}(T_{\gamma_{i-1}})}\mathcal{E}(T_{\gamma_i}(C)),$ which can be argued as follows. \supseteq is obvious because $\mathcal{E}(T_{\gamma_i}) \supseteq \mathcal{E}(T_{\gamma_i}(C))$. To prove the reverse inclusion, note that $\gamma_{i-1} < \gamma_i$ and so the edge in T_{γ_i} must be in a connected component of $T_{\gamma_{i-1}}$, namely, a subtree $T_{\gamma_i}(C)$ induced on some $C \in \mathcal{P}(T_{\gamma_{i-1}})$. This completes the proof of Theorem 4.3.

APPENDIX D CLUSTERING BY MINIMUM AVERAGE COST

The objective of minimum average cost (MAC) clustering is to obtain a partition \mathcal{P} of size $|\mathcal{P}| > k$ for some threshold k as the set of clusters, and the singleton elements in the partition are also regarded as clusters in satisfying the constraint $|\mathcal{P}| > k$. To solve this problem using our clustering solution by multivariate mutual information, it is natural to use

$$\mathcal{P} = \mathcal{P}_i$$
 such that $|\mathcal{P}_{i-1}| < k < |\mathcal{P}_i|$. (D.1)

where \mathcal{P}_i 's form the PSP of the entropy function in (3.18). \mathcal{P}_i is the coarsest partition from the PSP with more than k parts. The clustering solution proposed by [61] obtains the partition by solving the following minimum average cost constraint

$$\gamma = \min_{\mathcal{P} \in \Pi(V): |\mathcal{P}| > k} \frac{\sum_{C \in \mathcal{P}} f(C)}{|\mathcal{P}| - k}, \text{ or } (\textbf{D.2a})$$
$$-k\gamma = \min_{\mathcal{P} \in \Pi(V): |\mathcal{P}| > k} \sum_{C \in \mathcal{P}} [f(C) - \gamma] (\textbf{D.2b})$$

$$-k\gamma = \min_{\mathcal{P} \in \Pi(V): |\mathcal{P}| > k} \sum_{C \in \mathcal{P}} [f(C) - \gamma]$$
 (D.2b)

where f is a submodular function that needs to be chosen appropriately. The question of interest is, whether there is an obvious choice of f in terms of the entropy function h for which the two clustering solutions in (D.1) and (D.2) are the same. The similarity is more apparent by thinking of \mathcal{P} in (D.1) as the solution to (3.17a), namely,

$$\gamma_i = \min_{\mathcal{P} \in \Pi(V): |\mathcal{P}| > |\mathcal{P}_{i-1}|} \frac{h[\mathcal{P}] - h[\mathcal{P}_{i-1}]}{|\mathcal{P}| - |\mathcal{P}_{i-1}|},$$

which is similar to (D.2a) except for the numerator and kin place of $|\mathcal{P}_{i-1}|$. Note that the choice of \mathcal{P}_{i-1} depends on k according to (D.1). In particular, $|\mathcal{P}_{i-1}| \leq k$ but equality is not *needed* so long as the solution \mathcal{P}_i to the above minimization satisfies $|\mathcal{P}_i| > k$ as required by (D.1).

For graphical networks, [61] chooses f to be the cut function of the graph, which is also the case for info-clustering by information flow in Section IV-C. The following is a concrete example that distinguishes info-clustering from MAC clustering.

Example D.1 Consider a weighted graph G with vertex set $V = \{1, 2, 3, 4\}, \text{ edge set } E = \{e_{12}, e_{23}, e_{34}\}, \text{ and } E = \{e_{12}, e_{23}, e_$

$$\begin{split} \theta(e_{12}) &= \{1,2\}, &\quad \theta(e_{23}) &= \{2,3\}, &\quad \theta(e_{34}) &= \{3,4\} \text{ and } \\ c(e_{12}) &= 2, &\quad c(e_{23}) &= 3, &\quad c(e_{34}) &= 4, \end{split}$$

where θ and c are the edge and weight functions as in Section IV-C. The PSP of the cut function $C \mapsto c(\delta_G(C))$ (see (4.21b)) can be shown to be

$$\mathcal{P}_0 = \{\{1, 2, 3, 4\}\}
\mathcal{P}_1 = \{\{1\}, \{2, 3, 4\}\}
\mathcal{P}_2 = \{\{1\}, \{2\}, \{3, 4\}\}
\mathcal{P}_3 = \{\{1\}, \{2\}, \{3\}, \{4\}\}.$$

For k = 2, info-clustering will return \mathcal{P}_2 according (D.1), since \mathcal{P}_2 is the coarsest partition with more than k parts. However, choosing f to be the cut function, MAC clustering does not return the same solution because the average cost (D.2a) of \mathcal{P}_3 is strictly smaller than that of \mathcal{P}_2 :

$$\frac{\sum_{C \in \mathcal{P}_2} f(C)}{|\mathcal{P}_2| - k} = \frac{c(e_{12}) + c(e_{23})}{3 - 2} = 5$$

$$\frac{\sum_{C \in \mathcal{P}_3} f(C)}{|\mathcal{P}_3| - k} = \frac{c(e_{12}) + c(e_{23}) + c(e_{34})}{4 - 2} = 4.5 < 5.$$

Indeed, it can be shown that \mathcal{P}_3 achieves the minimum average cost among all other partitions of V, and so MAC clustering will return the less intuitive clustering by \mathcal{P}_3 instead of \mathcal{P}_2 .

Actually, f was assumed to be non-negative in [61], because then, the constraint $|\mathcal{P}| > k$ can be dropped from (D.2b) without changing the solution. Doing so reduces (D.2b) to computing the Dilworth truncation, which can be done efficiently and guaranteed to return a partition in the PSP, despite the possibility of returning one that is finer than required, as shown in the previous example (since cut function is non-negative). In the general case when f can be negative, removing the constraint $|\mathcal{P}| > k$ from (D.2b) can potentially change the solution to something outside the PSP, and so it is unclear whether the clusters can be computed efficiently. In the following, we will further compare MAC clustering to info-clustering without assuming f to non-negative.

It can be shown that constant scaling of f does not change the solution to (D.2), but constant shift does. To ensure submodularity, a reasonable choice of f is $f = h_s$ for some appropriate constant shift s. We will show that there is a choice of s such that the clustering solutions for k = 1 are the same for (D.1) and (D.2). However, there is no choice of s for which the complete clustering solutions for different k are the same.

More precisely, the fundamental partition $\mathcal{P}_1 = \mathcal{P}^*(\mathsf{Z}_V)$ in our clustering solution (D.1) can be obtained from (D.2) with

$$f(B) := h_{h(V)}(B) = -H(\mathsf{Z}_{V \setminus B}|\mathsf{Z}_B)$$
 for $B \subseteq V$.

To see this, rewrite (D.2a) with the above choice and k = 1:

$$\begin{split} \gamma &= \min_{\mathcal{P} \in \Pi(V): |\mathcal{P}| > 1} \frac{\sum_{C \in \mathcal{P}} - H(\mathsf{Z}_{V \setminus C} | \mathsf{Z}_C)}{|\mathcal{P}| - 1} \\ &= \min_{\mathcal{P} \in \Pi'(V)} \frac{\sum_{C \in \mathcal{P}} H(\mathsf{Z}_C) - H(\mathsf{Z}_V)}{|\mathcal{P}| - 1} - H(\mathsf{Z}_V). \end{split}$$

The first term on the R.H.S. is $I(Z_V)$ by (3.4a) and (3.3), and so the finest optimal partition is $\mathcal{P}^*(Z_V)$ as desired.

Note that we have allowed f to be negative above $H(\mathsf{Z}_B) < H(\mathsf{Z}_V)$. However, it turns out that, even if we allow $f = h_s$ to be negative, there is no choice of s for which the complete clustering solutions in (D.1) and (D.2) are the same:

Example D.2 Let $V = \{1, 2, 3, 4\}$ and

$$Z_1 := (X_a, X_c)$$
 $Z_4 := (X_d, X_e),$ $Z_2 := (X_a, X_b)$ $Z_3 := (X_b, X_c)$

where X_i are independent uniformly random bits. The Dilworth truncation $\hat{h}_{\gamma}(V)$ is plotted in Fig. 8a. Our clustering solution (D.1) consists of the PSP:

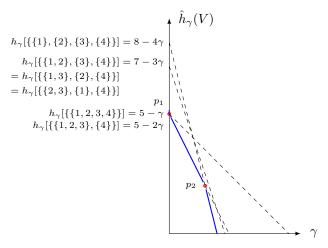
$$\begin{split} \mathcal{P}_1 &= \{\{1,2,3\},\{4\}\} & \text{for } k = 1 \\ \mathcal{P}_2 &= \{\{i\} \mid i \in V\} & \text{for } k > 1 \end{split}$$

The solution respects the symmetry of the correlation in $Z_{\{1,2,3\}}$ and the independence between Z_4 and $Z_{\{1,2,3\}}$.

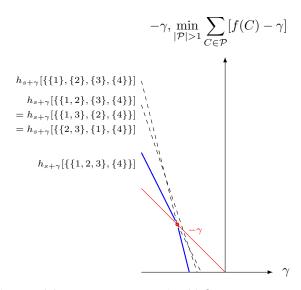
To compute the MAC clustering in (D.2), rewrite (D.2b) with $f=h_{\rm s}$

$$-k\gamma = \min_{\mathcal{P} \in \Pi(V): |\mathcal{P}| > k} h_{s+\gamma}[\mathcal{P}].$$

For k=1 and s=3.5, the L.H.S. and R.H.S. are plotted in Fig. 8b. Since the curve for the L.H.S. intersect the curve on the R.H.S. along the line segment $h_{s+\gamma}[\mathcal{P}_1]$, the partition \mathcal{P}_1 is an optimal solution to the R.H.S.. If s<3.5, then $-\gamma$ will intersect the line segment corresponding to $h_{s+\gamma}[\mathcal{P}_2]$ instead of $h_{s+\gamma}[\mathcal{P}_1]$. Therefore, in order to have \mathcal{P}_1 to be the solution, we must have $s\geq3.5$. For k=2 and s=3.5, the plot in Fig. 8c shows that the optimal partition to the R.H.S. is not a partition in the PSP. This is the case even for $s\geq3.5$ because



(a) For PSP (D.1).



(b) For minimum average cost (D.2) with k = 1, s = 3.5.

$$-2\gamma, \min_{|\mathcal{P}|>2} \sum_{C \in \mathcal{P}} [f(C) - \gamma]$$

$$h_{s+\gamma}[\{\{1\}, \{2\}, \{3\}, \{4\}\}\}]$$

$$= h_{s+\gamma}[\{\{1, 3\}, \{2\}, \{4\}\}]$$

$$= h_{s+\gamma}[\{\{2, 3\}, \{1\}, \{4\}\}]$$

(c) For minimum average cost (D.2) with k = 2, s = 3.5.

Fig. 8: Computing the clustering solutions to Example D.2.

increasing γ will only move the intersection point to the left further away from $h_{s+\gamma}[\mathcal{P}_2]$. The optimal partition, such as $\{\{1,2\},\{3\},\{4\}\}$, does not appear to respect the symmetry in the correlation among $Z_{\{1,2,3\}}$.

ACKNOWLEDGMENT

The authors would like to thank Prof. Raymond W. Yeung, the Co-Director of the Institute of Network Coding (INC) at the Chinese University of Hong Kong, for his generous support of our research on information theory; Prof. Lav R. Varshney, Prof. Rosanna Y-Y. Chan, and Prof. Chen Change Loy for their suggestions of relevant works in machine learning and neuroscience; Dr. Javad B. Ebrahimi, Dr. Ravi K. Raman, and Dr. Ni Ding for their helpful discussions; Prof. Frank Kschischang, Prof. Devavrat Shah, and the colleagues at INC, whose comments have helped significantly improve the presentation of the paper. We would also like to thank the Associate Editor Prof. Peter Thomas and the reviewers for their detailed reading and insightful comments.

The first author would like to thank Prof. Imre Csiszár for the discussion on the divergence upper bound for secret key agreement and the issue of tightness, and Prof. Prakash Narayan for his recognitions of the contribution of this work. He would also like to thank his Ph.D. advisor, Prof. Lizhong Zheng, for leading him to the field of information theory.

REFERENCES

- C. Chan, A. Al-Bashabsheh, J. Ebrahimi, T. Kaced, and T. Liu, "Multivariate mutual information inspired by secret-key agreement," Proceedings of the IEEE, vol. 103, no. 10, pp. 1883–1913, Oct 2015.
- [2] C. Chan and T. Liu, "Clustering of random variables by multivariate mutual information under Chow-Liu tree approximations," in *Fifty-Third Annual Allerton Conference on Communication, Control, and Computing*, Allerton Retreat Center, Monticello, Illinois, pp. 993–999, Sep. 2015.
- [3] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Pac Symp Biocomput*, vol. 5, 2000, pp. 418–429.
- [4] R. Sharan and R. Shamir, "CLICK: a clustering algorithm with applications to gene expression analysis," in *Proc Int Conf Intell Syst Mol Biol*, vol. 8, 2000, pp. 307–316.
- [5] G. Deco, G. Tononi, M. Boly, and M. L. Kringelbach, "Rethinking segregation and integration: contributions of whole-brain modelling," *Nature Reviews Neuroscience*, vol. 16, no. 7, pp. 430–439, 2015.
- [6] M. Hinne, M. Ekman, R. J. Janssen, T. Heskes, and M. van Gerven, "Probabilistic clustering of the human connectome identifies communities and hubs," *PloS one*, vol. 10, no. 1, 2015, e0117179.
- [7] G. Tononi, A. R. McIntosh, D. P. Russell, and G. M. Edelman, "Functional clustering: identifying strongly interactive brain regions in neuroimaging data," *Neuroimage*, vol. 7, no. 2, pp. 133–149, 1998.
- [8] I. H. G. S. Consortium et al., "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, pp. 931–945, 2004.
- [9] F. A. Azevedo, L. R. Carvalho, L. T. Grinberg, J. M. Farfel, R. E. Ferretti, R. E. Leite, R. Lent, S. Herculano-Houzel *et al.*, "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain," *Journal of Comparative Neurology*, vol. 513, no. 5, pp. 532–541, 2009.
- [10] P. O. Brown and D. Botstein, "Exploring the new world of the genome with DNA microarrays," *Nature genetics*, vol. 21, pp. 33–37, 1999.
- [11] R. D. Morin, M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. J. Pugh, H. McDonald, R. Varhol, S. J. Jones, and M. A. Marra, "Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing," *Biotechniques*, vol. 45, no. 1, p. 81, 2008.
- [12] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature reviews genetics*, vol. 10, no. 1, pp. 57–63, 2009

- [13] J. White, E. Southgate, J. Thomson, and S. Brenner, "The structure of the nervous system of the nematode Caenorhabditis elegans: the mind of a worm," *Philos Trans R Soc Lond B Biol Sci*, vol. 314, pp. 1–340, 1986
- [14] L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, and D. B. Chklovskii, "Structural properties of the Caenorhabditis elegans neuronal network," *PLoS Comput Biol*, vol. 7, no. 2, 2011, e1001066.
- [15] S. W. Emmons, "The beginning of connectomics: a commentary on white et al. (1986) 'The structure of the nervous system of the nematode Caenorhabditis elegans'," *Phil. Trans. R. Soc. B.*, vol. 370, no. 1666, p. 20140309, 2015.
- [16] T. A. Jarrell, Y. Wang, A. E. Bloniarz, C. A. Brittin, M. Xu, J. N. Thomson, D. G. Albertson, D. H. Hall, and S. W. Emmons, "The connectome of a decision-making neural network," *Science*, vol. 337, no. 6093, pp. 437–444, 2012.
- [17] V. J. Wedeen, P. Hagmann, W.-Y. I. Tseng, T. G. Reese, and R. M. Weisskoff, "Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging," *Magnetic resonance in medicine*, vol. 54, no. 6, pp. 1377–1386, 2005.
- [18] V. J. Wedeen, R. Wang, J. D. Schmahmann, T. Benner, W. Tseng, G. Dai, D. Pandya, P. Hagmann, H. D'Arceuil, and A. J. de Crespigny, "Diffusion spectrum magnetic resonance imaging (DSI) tractography of crossing fibers," *Neuroimage*, vol. 41, no. 4, pp. 1267–1277, 2008.
- [19] P. Hagmann, M. Kurant, X. Gigandet, P. Thiran, V. J. Wedeen, R. Meuli, and J.-P. Thiran, "Mapping human whole-brain structural networks with diffusion MRI," *PloS one*, vol. 2, no. 7, 2007, e597.
- [20] L. of Neuro Imaging and M. C. for Biomedical Imaging at Massachusetts General Hospital, "Human Connectome Project," 2009, [Online; accessed Apr-2016]. [Online]. Available: http://www.humanconnectomeproject.org
- [21] J. T. Vogelstein, "Open Connectome Project," 2011, [Online; accessed Apr-2016]. [Online]. Available: http://www.openconnectomeproject.org
- [22] R. A. Poldrack, "MyConnectome Project," 2011, [Online; accessed Apr-2016]. [Online]. Available: http://myconnectome.org
- [23] T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler et al., "Repeated observation of breast tumor subtypes in independent gene expression data sets," Proceedings of the National Academy of Sciences, vol. 100, no. 14, pp. 8418–8423, 2003.
- [24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural* information processing systems, 2012, pp. 1097–1105.
- [25] R. W. Yeung, Information Theory and Network Coding. Springer 2008
- [26] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed., ser. Wiley Series in Telecommunications and Signal Processing. John Wiley & Sons, 2012.
- [27] I. Csiszar and J. Körner, Information theory: coding theorems for discrete memoryless systems, 2nd ed. Cambridge University Press, 2011.
- [28] R. G. Gallager, Information theory and reliable communication. Springer, 1968, vol. 2.
- [29] D. J. MacKay, Information theory, inference and learning algorithms. Cambridge university press, 2003.
- [30] O. Sporns, "Network attributes for segregation and integration in the human brain," *Current opinion in neurobiology*, vol. 23, no. 2, pp. 162–171, 2013.
- [31] M. P. van den Heuvel and O. Sporns, "Network hubs in the human brain," *Trends in cognitive sciences*, vol. 17, no. 12, pp. 683–696, 2013.
- [32] S. Nigam, M. Shimono, S. Ito, F.-C. Yeh, N. Timme, M. Myroshny-chenko, C. C. Lapish, Z. Tosi, P. Hottowy, W. C. Smith *et al.*, "Richclub organization in effective connectivity among cortical neurons," *The Journal of Neuroscience*, vol. 36, no. 3, pp. 670–684, 2016.
- [33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv preprint arXiv:1512.03385, 2015, presented at 2016 IEEE Conference on Computer Vision and Pattern Recognition.
- [34] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.
- [35] A. Thalamuthu, I. Mukhopadhyay, X. Zheng, and G. C. Tseng, "Evaluation and comparison of gene clustering methods in microarray analysis," *Bioinformatics*, vol. 22, no. 19, pp. 2405–2412, 2006.
- [36] M. Pelillo, "What is a cluster? perspectives from game theory," in Proc. of the NIPS Workshop on Clustering Theory, 2009.
- [37] U. Von Luxburg, R. C. Williamson, and I. Guyon, "Clustering: Science or art?" in *ICML Unsupervised and Transfer Learning*, 2012, pp. 65– 80.
- [38] P. Agarwal, M. A. Alam, and R. Biswas, "Issues, challenges and tools

- of clustering algorithms," arXiv preprint arXiv:1110.2610, 2011.
- [39] R. Kannan, S. Vempala, and A. Vetta, "On clusterings: Good, bad and spectral," *Journal of the ACM (JACM)*, vol. 51, no. 3, pp. 497–515, 2004.
- [40] N. Bansal, A. Blum, and S. Chawla, "Correlation clustering," *Machine Learning*, vol. 56, no. 1-3, pp. 89–113, 2004.
- [41] N. Ailon, M. Charikar, and A. Newman, "Aggregating inconsistent information: ranking and clustering," *Journal of the ACM (JACM)*, vol. 55, no. 5, p. 23, 2008.
- [42] X. Pan, D. Papailiopoulos, S. Oymak, B. Recht, K. Ramchandran, and M. I. Jordan, "Parallel correlation clustering on big graphs," in *Advances in Neural Information Processing Systems*, 2015, pp. 82–90.
- [43] A. Kraskov and P. Grassberger, "MIC: Mutual information based hierarchical clustering," in *Information Theory and Statistical Learning*, F. Emmert-Streib and M. Dehmer, Eds. Springer US, 2009, pp. 101– 123.
- [44] M. Aghagolzadeh, H. Soltanian-Zadeh, B. Araabi, and A. Aghagolzadeh, "A hierarchical clustering based on mutual information maximization," in 2007 IEEE International Conference on Image Processing, vol. 1, Sept 2007, pp. I – 277–I – 280.
- [45] G. Ver Steeg and A. Galstyan, "Discovering structure in highdimensional data through correlation explanation," in Advances in Neural Information Processing Systems, 2014, pp. 577–585.
- [46] V. Misra, "Universal communication and clustering," Ph.D. dissertation, Stanford University, Dept. of Electrical Engineering, 2014.
- [47] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.
- [48] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of Research and Development*, vol. 4, no. 1, pp. 66–82, 1960.
- [49] I. Nemenman, "Information theory, multivariate dependence, and genetic network inference," arXiv preprint q-bio/0406015, 2004.
- [50] I. Csiszár and P. Narayan, "Secrecy capacities for multiple terminals," IEEE Trans. Inf. Theory, vol. 50, no. 12, pp. 3047–3061, Dec. 2004.
- [51] C. Chan, "On tightness of mutual dependence upperbound for secretkey capacity of multiple terminals," arXiv preprint arXiv:0805.3200, 2008
- [52] C. Chan and L. Zheng, "Mutual dependence for secret key agreement," in *Information Sciences and Systems (CISS)*, 2010 44th Annual Conference on, March 2010, pp. 1–6.
- [53] C. Chan, "The hidden flow of information," in Proc. IEEE Int. Symp. on Inf. Theory, St. Petersburg, Russia, Jul. 2011.
- [54] C. Chan, A. Al-Bashabsheh, Q. Zhou, N. Ding, T. Liu, and A. Sprintson, "Successive omniscience," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3270–3289, June 2016.
- [55] C. Chan, A. Al-Bashabsheh, and Q. Zhou, "Incremental and decremental secret key agreement," in *Proc. of IEEE Int. Symp. on Inf. Theory*, July 2016, pp. 2514–2518.
- [56] C. Chan, M. Mukherjee, N. Kashyap, and Q. Zhou, "When is omniscience a rate-optimal strategy for achieving secret key capacity?" Sep. 2016, to appear in IEEE Information Theory Workshop (ITW).
- [57] M. Mukherjee, N. Kashyap, and Y. Sankarasubramaniam, "Achieving SK capacity in the source model: When must all terminals talk?" in Proc. of IEEE Int. Symp. on Inf. Theory, June 2014.
- [58] H. Zhang, Y. Liang, and L. Lai, "Secret key capacity: Talk or keep silent?" in Proc. of IEEE Int. Symp. on Inf. Theory, June 2015.
- [59] S. Fujishige, "Optimization over the polyhedron determined by a submodular function on a co-intersecting family," *Mathematical Pro*gramming, vol. 42, no. 1-3, pp. 565–577, 1988.
- [60] H. Narayanan, "The principal lattice of partitions of a submodular function," *Linear Algebra and its Applications*, vol. 144, no. 0, pp. 179 – 216, 1990.
- [61] K. Nagano, Y. Kawahara, and S. Iwata, "Minimum average cost clustering." in NIPS, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 1759–1767.
- [62] M. Mukherjee, C. Chan, N. Kashyap, and Q. Zhou, "Bounds on the communication rate needed to achieve SK capacity in the hypergraphical source model," in *Proc. of IEEE Int. Symp. on Inf. Theory*, July 2016, pp. 2504–2508.
- [63] M. Mukherjee and N. Kashyap, "The communication complexity of achieving SK capacity in a class of PIN models," in *Proc. of IEEE Int.* Symp. on Inf. Theory, June 2015, pp. 296–300.
- [64] M. Mukherjee, N. Kashyap, and Y. Sankarasubramaniam, "On the public communication needed to achieve sk capacity in the multiterminal source model," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 3811–3830,

- July 2016.
- [65] N. Ding, C. Chan, Q. Zhou, R. A. Kennedy, and P. Sadeghi, "Fairness in communication for omniscience," in *Proc. of IEEE Int. Symp. on Inf. Theory*, July 2016, pp. 2314–2318.
- [66] N. Ding, R. Kennedy, and P. Sadeghi, "Iterative merging algorithm for cooperative data exchange," in *Proc. of IEEE Int. Symp. on Net. Cod.*, June 2015, pp. 41–45.
- [67] C. Chan, A. Al-Bashabsheh, Q. Zhou, and T. Liu, "Duality between feature selection and data clustering," arXiv preprint arXiv:1609.08312, 2016, preliminary results published at the Fifty-Fourth Annual Allerton Conference on Communication, Control, and Computing, Allerton Retreat Center, Monticello, Illinois.
- [68] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.
- [69] S. Nitinawarat, C. Ye, A. Barg, P. Narayan, and A. Reznik, "Secret key generation for a pairwise independent network model," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 6482–6489, Dec 2010.
- [70] S. Nitinawarat and P. Narayan, "Perfect omniscience, perfect secrecy, and steiner tree packing," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 6490–6500, Dec. 2010.
- [71] A. Schrijver, Combinatorial Optimization: Polyhedra and Efficiency. Springer, 2002.
- [72] C. Chan, "Matroidal undirected network," in *Proc. of IEEE Int. Symp. on Inf. Theory*, July 2012, pp. 1498–1502.
- [73] ——, "Cyclic linking network," in Proc. of IEEE Int. Symp. on Inf. Theory, July 2013, pp. 789–793.
- [74] C. Chan, K. W. Shum, and Q. T. Sun, "Combinatorial flow over cyclic linear networks," in *IEEE Information Theory Workshop (ITW)*, Sep. 2013, pp. 1–5.
- [75] T. M. Cover and J. A. Thomas, Elements of Information Theory. Wiley-Interscience Publication, 1991.
- [76] D. Balduzzi and G. Tononi, "Integrated information in discrete dynamical systems: motivation and theoretical framework," *PLoS Comput Biol*, vol. 4, no. 6, 2008, e1000091.
- [77] F. Echenique and R. G. Fryer Jr, "A measure of segregation based on social interactions," *The Quarterly Journal of Economics*, pp. 441–485, 2007
- [78] C. E. Shannon, "Communication theory of secrecy systems," The Bell System Technical Journal, vol. 28, no. 4, pp. 656–715, 1949.
- [79] R. P. Stanley, Enumerative Combinatorics, 2nd ed. Cambridge University Press, 2012, vol. I.
- [80] A. Wyner, "The common information of two dependent random variables," *IEEE Trans. Inf. Theory*, vol. 21, no. 2, pp. 163–179, Mar 1975.
- [81] P. Gács and J. Körner, "Common information is far less than mutual information," *Problems of Control and Information Theory*, vol. 2, no. 2, pp. 149–162, Feb. 1972.
- [82] W. Liu, G. Xu, and B. Chen, "The common information of N dependent random variables," in Forty-Eighth Annual Allerton Conference on Communication, Control, and Computing, Sept 2010, pp. 836–843.
- [83] S. Kullback and R. A. Leibler, "On information and sufficiency," The annals of mathematical statistics, vol. 22, no. 1, pp. 79–86, 1951.
- [84] R. W. Yeung, "A new outlook on Shannon's information measures," IEEE Trans. Inf. Theory, vol. 37, no. 3, pp. 466–474, May 1991.
- [85] W. J. McGill, "Multivariate information transmission," *Psychometrika*, vol. 19, no. 2, pp. 97–116, Jun. 1954.
- [86] N. Milosavljevic, S. Pawar, S. El Rouayheb, M. Gastpar, and K. Ramchandran, "Deterministic algorithm for the cooperative data exchange problem," in *Proc. of IEEE Int. Symp. on Inf. Theory*, Jul. 2011.
- [87] S. Fujishige, "Polymatroidal dependence structure of a set of random variables," *Information and Control*, vol. 39, no. 1, pp. 55 – 72, 1978.
- [88] F. D. J. Dunstan, "Matroids and submodular functions," *Quarterly Journal of Mathematics*, vol. 27, no. 3, pp. 339–348, 1976.
- [89] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 462–467, May 1968.
- [90] C. Chan, "Generating secret in a network," Ph.D. dissertation, Massachusetts Institute of Technology, 2010.
- [91] C. Chan and L. Zheng, "Multiterminal secret key agreement," *IEEE Trans. Inf. Theory*, vol. 60, no. 6, pp. 3379–3412, June 2014.
 [92] A. Avestimehr, S. Diggavi, and D. Tse, "Wireless network information
- [92] A. Avestimehr, S. Diggavi, and D. Tse, "Wireless network information flow: A deterministic approach," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 1872–1905, April 2011.
- [93] C. S. J. Nash-Williams, "Edge-disjoint spanning trees of finite graphs," J. London Math. Soc., vol. 1, no. 1, pp. 445–450, 1961.
- [94] A. Frank, T. Király, and M. Kriesell, "On decomposing a hypergraph into k-connected sub-hypergraphs," Discrete Applied Mathematics, vol.

- 131, no. 2, pp. 373-383, Sep. 2003.
- [95] J. Bang-Jensen and S. Thomassé, "Decompositions and orientations of hypergraphs," Preprint no. 10, Department of Mathematics and Computer Science, University of Southern Denmark, May 2001.
- [96] C. S. J. Nash-Williams, "Decomposition of finite graphs into forests," J. London Math. Soc., vol. 1, no. 1, pp. 12–12, 1964.
- [97] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings* of the National Academy of Sciences, vol. 95, no. 25, pp. 14863– 14868, 1998.
- [98] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3702–3720, 2016.
- [99] E. Telatar, "Capacity of multi-antenna gaussian channels," European transactions on telecommunications, vol. 10, no. 6, pp. 585–595, 1999.
- [100] C. E. Shannon, "A mathematical theory of communication," The Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, July 1948.



Chung Chan received the B.Sc., M.Eng. and Ph.D. from the EECS Department at MIT in 2004, 2005 and 2010 respectively. He is currently a Research Assistant Professor at the Institute of Network Coding, the Chinese University of Hong Kong. His research is in the area of information theory, with applications to network coding, multiple-terminal source coding and security problems that involve high-dimensional statistics. He is currently working on machine learning applications such as data clustering and feature selection.



Ali Al-Bashabsheh received a B.Sc. (2001) and an M.Sc (2005) in electrical engineering from Jordan University of Science and Technology, an M.Sc. (2012) in mathematics from Carleton University, and a Ph.D. (2014) in electrical engineering from the University of Ottawa. Since April 2014, he has been a postdoctoral fellow at the Institute of Network Coding at the Chinese University of Hong Kong. His research interests include graphical models, coding theory, and information theory.



Qiaoqiao Zhou received his B.B.A. in business administration and M.S. in electrical engineering from Beijing University of Post and Telecommunication, China, in 2011 and 2014, respectively. From 2014 to 2015, he was a research assistant at the Institute of Network Coding, the Chinese University of Hong Kong. He is currently a Ph.D. candidate at the Department of Information Engineering, the Chinese University of Hong Kong. His research interests include information-theoretic security and machine learning.



Tarik Kaced was born in France, he received his B.Sc. in Fundamental Computer Science from École Normale Supérieure de Lyon in 2007, and his M.Sc. from Université de Nice Sophia-Antipolis in 2009. He completed his Ph.D. degree in Computer Science in 2012 at Université de Montpellier 2 in the ESCAPE team from LIRMM. He has been a post-doctoral fellow at the Institute of Network Coding at The Chinese University of Hong Kong for two years. He was a post-doctoral fellow at in Université Paris-Est Créteil at the Algorithmic, Complexity and

Logic Laboratory.

His research interests include information theory, Kolmogorov complexity, matroid theory, computability, information inequalities, combinatorics, error correcting codes and secret sharing.



Tie Liu was born in Jilin, China in 1976. He received his B.S. (1998) and M.S. (2000) degrees, both in Electrical Engineering, from Tsinghua University, Beijing, China and a second M.S. degree in Mathematics (2004) and Ph.D. degree in Electrical and Computer Engineering (2006) from the University of Illinois at Urbana-Champaign. Since August 2006 he has been with Texas A&M University, where he is currently an Associate Professor with the Department of Electrical and Computer Engineering. His primary research interest is in the area of information

theory and statistical information processing.

Dr. Liu received an M. E. Van Valkenburg Graduate Research Award (2006) from the University of Illinois at Urbana-Champaign and a Faculty Early Career Development (CAREER) Award (2009) from the National Science Foundation. He was a Technical Program Committee Co-Chair for the 2008 IEEE Global Communications Conference (GLOBECOM) and a General Co-Chair for the 2011 IEEE North American School of Information Theory. He currently serves as an Associate Editor for Shannon Theory for the IEEE Transactions on Information Theory.