Shared Information for a Markov Chain on a Tree

Sagnik Bhattacharya and Prakash Narayan

***Abstract*—Shared information is a measure of mutual dependence among multiple jointly distributed random variables with finite alphabets. For a Markov chain on a tree with a given joint distribution, we give a new proof of an explicit characterization of shared information. The Markov chain on a tree is shown to possess a global Markov property based on graph separation; this property plays a key role in our proofs. When the underlying joint distribution is not known, we exploit the special form of this characterization to provide a multiarmed bandit algorithm for estimating shared information, and analyze its error performance.**

***Index Terms*—Global Markov property, Markov chain on a tree, multiarmed bandits, mutual information, mutual information estimation, shared information.**

I. Introduction

LET $X_1, \ldots, X_m$, $m \geq 2$ be random variables (rvs) with finite alphabets $\mathcal{X}_1, \ldots, \mathcal{X}_m$, respectively, and joint probability mass function (pmf) $P_{X_1 \cdots X_m}$. The *shared information* $\mathrm{SI}(X_1, \ldots, X_m)$ of the rvs $X_1, \ldots, X_m$ is a measure of mutual dependence among them; and for $m = 2$, $\mathrm{SI}(X_1, X_2)$ particularizes to mutual information $\mathrm{I}(X_1 \wedge X_2)$. Consider m terminals, with terminal i having privileged access to independent and identically distributed (i.i.d.) repetitions of X_i, $i = 1, \ldots, m$. Shared information $\mathrm{SI}(X_1, \ldots, X_m)$ has the operational meaning of being the largest rate of *shared common randomness* that the m terminals can generate in a distributed manner upon cooperating among themselves by means of interactive, publicly broadcast and noise-free communication[1]. Shared information measures the maximum rate of common randomness that is (nearly) independent of the open communication used to generate it.

The (Kullback-Leibler) divergence-based expression for $\mathrm{SI}(X_1, \ldots, X_m)$ was discovered in [21, Example 4], where it was derived as an upper bound for a single-letter formula for the "secret key capacity of a source model" with m terminals, a concept defined by the operational meaning above. The upper bound was shown to be tight for $m = 2$ and 3. Subsequently, in a significant advance [8], [9], [15], tightness of the upper bound was established for arbitrary m, thereby imbuing $\mathrm{SI}(X_1, \ldots, X_m)$ with the operational significance of being the mentioned maximum rate of shared secret common randomness. The potential for shared information to serve as a natural measure of mutual dependence of $m \geq 2$ rvs, in the manner of mutual information for $m = 2$ rvs, was suggested in [34]; see also [35].

A comprehensive and consequential study of shared information, where it is termed "multivariate mutual information" [11], examines the role of secret key capacity as a measure of mutual dependence among multiple rvs and derives important properties including structural features of an underlying optimization along with connections to the theory of submodular functions.

In addition to constituting secret key capacity for a multiterminal source model ([8], [15], [21]), shared information also affords operational meaning for: maximal packing of edge-disjoint spanning trees in a multigraph ([36], [37]; see also [10], [19], [11] for variant models); optimum querying exponent for resolving common randomness [42]; strong converse for multiterminal secret key capacity [42], [43]; and also undirected network coding [9], data clustering [14], among others.

As argued in [11], shared information also possesses several attributes of measures of dependence among $m \geq 2$ rvs proposed earlier, including Watanabe's total correlation [45] and Han's dual total correlation [26] (both mentioned in Section II). For $m = 2$ rvs, measures of common information due to Gács-Körner [23], Wyner [46] and Tyagi [41] have operational meanings; extensions to $m > 2$ rvs merit further study (an exception [32] treats Wyner's common information).

For a given joint pmf $P_{X_1 \cdots X_m}$ of the rvs $X_1, \ldots, X_m$, an explicit characterization of $\mathrm{SI}(X_1, \ldots, X_m)$ can be challenging (see Definition 1 below); exact formulas are available for special cases (cf. e.g., [21], [37], [11]). An efficient algorithm for calculating $\mathrm{SI}(X_1, \ldots, X_m)$ is given in [11].

Our focus in this paper is on a Markov chain on a tree (MCT) [24]. Tree-structured probabilistic graphical models are appealing owing to desirable statistical properties that enable, for instance, efficient algorithms for exact inference [28], [39]; decoding [33], [28]; sampling [22]; and structure learning [16]. An MCT can serve as a tractable tree-structured approximation to a given joint distribution arising in applications such as omniscience and secrecy generation [13], [21], and signal clustering [12]. The mentioned tractability facilitates exact calculation of associated rate quantities. We take the tree structure of our model to be known; algorithms exist already for learning tree structure from data samples [16], [17]. We exploit the special form of $P_{X_1 \cdots X_m}$ in the setting of an MCT to obtain a simple characterization of shared information. When the joint pmf $P_{X_1 \cdots X_m}$ is not known but the tree structure is, the said characterization facilitates an estimation of shared information.

In the setting of an MCT [24], our contributions are threefold. First, we derive an explicit characterization of shared information for an MCT with a given joint pmf $P_{X_1 \cdots X_m}$

S. Bhattacharya and P. Narayan are with the Department of Electrical and Computer Engineering and the Institute for Systems Research, University of Maryland, College Park, MD 20742, USA. E-mail: {sagnikb, prakash}@umd.edu. This work was supported by the U.S. National Science Foundation under Grants CCF 1910497 and CCF 2310203. A version of this paper was presented in part at the 2022 IEEE International Symposium on Information Theory [3].

[1]Our preferred nomenclature of shared information is justified by its operational meaning.

by means of a direct approach that exploits tree structure and Markovity of the pmf. A characterization of shared information had been sketched already in [21]; our new proof does not seek recourse to a secret key interpretation of shared information, unlike in [21]. Also, our proof differs in a material way from that in prior work [14] with a similar objective.

Second, we show an equivalence between the (weaker) original definition of an MCT [24] and a (stronger) global one based on separation in a graph [31, Section 3.2.1]. When $P_{X_1,\cdots,X_m}$ is assumed to be strictly positive, the two definitions are equivalent by the Hammersley-Clifford Theorem [31, Theorem 3.9]. We prove this equivalence even without said assumption, taking advantage of the underlying tree structure of the MCT; our proof method potentially is of independent interest.

Third, when $P_{X_1\cdots X_m}$ is not known, with the mentioned characterization serving as a linchpin, we provide an approach for estimating shared information for an MCT. Formulated as a correlated bandits problem [6], this approach seeks to identify the best arm-pair across which mutual information is minimal. Using a uniform sampling of arms, redolent of sampling mechanisms in [5], we provide an upper bound for the probability of estimation error and associated sample complexity. Our uniform sampling algorithm is similar to that in [47], [6]; however, our modified analysis takes into account estimator bias, a feature that is not common in known bandit algorithms. Also, this approach can accommodate more refined bandit algorithms as also alternatives to the probability of error criterion such as regret [7].

Section II contains the preliminaries. Useful properties of an MCT are elucidated in Section III. An explicit characterization of shared information for an MCT with a given $P_{X_1\cdots X_m}$ is provided in Section IV. Section V describes our approach for estimating shared information when $P_{X_1\cdots X_m}$ is not known. Section VI contains closing remarks.

II. Preliminaries

Let $X_1,\ldots,X_m$, $m \geq 2$, be rvs with finite alphabets $\mathcal{X}_1,\ldots,\mathcal{X}_m$, respectively, and joint pmf $P_{X_1\cdots X_m}$. For $A \subseteq \mathcal{M} = \{1,\ldots,m\}$, we write $X_A = (X_i, i \in A)$ with alphabet $\mathcal{X}_A = \prod_{i\in A}\mathcal{X}_i$. Let $\pi = (\pi_1,\ldots,\pi_k)$ denote a k-partition of $\mathcal{M}$, $2 \leq k \leq m$. All logarithms and exponentiations are with respect to the base 2, except $\ln$ and $\exp$ that are with respect to the base e.

Definition 1 (Shared information)**.** *The* shared information *of* $X_1,\ldots,X_m$ *is defined as*

$$\begin{aligned}&\mathrm{SI}(X_{\mathcal{M}})\\ &= \min_{2\leq k\leq m}\ \min_{\pi=(\pi_u,u=1,\cdots,k)}\ \frac{1}{k-1}D(P_{X_{\mathcal{M}}} \parallel \prod_{u=1}^{k} P_{X_{\pi_u}})\\ &= \min_{2\leq k\leq m}\ \min_{\pi=(\pi_u,u=1,\cdots,k)}\ \frac{1}{k-1}\left[\sum_{u=1}^{k}\mathrm{H}(X_{\pi_u}) - \mathrm{H}(X_{\mathcal{M}})\right].\end{aligned} \tag{1}$$

For a partition π of $\mathcal{M}$ with $2 \leq |\pi| \leq m$ atoms, it will be convenient to denote

$$\mathcal{I}(\pi) = \frac{1}{|\pi|-1}D(P_{X_{\mathcal{M}}} \parallel \prod_{u=1}^{|\pi|} P_{X_{\pi_u}}) \tag{2}$$

so that $\mathrm{SI}(X_{\mathcal{M}}) = \min_{2\leq|\pi|\leq m}\mathcal{I}(\pi)$.

Example 1. For $\mathcal{M} = \{1,2\}$, we have

$$\mathrm{SI}(X_1,X_2) = \text{mutual information } \mathrm{I}(X_1 \wedge X_2)$$

and for $\mathcal{M} = \{1,2,3\}$, it is checked readily that $\mathrm{SI}(X_1,X_2,X_3)$ is the minimum of $\mathrm{I}(X_1 \wedge X_2, X_3)$, $\mathrm{I}(X_2 \wedge X_1, X_3)$, $\mathrm{I}(X_3 \wedge X_1, X_2)$ and

$$\frac{1}{2}\left[\mathrm{H}(X_1)+\mathrm{H}(X_2)+\mathrm{H}(X_3)-\mathrm{H}(X_1,X_2,X_3)\right],$$

and can be inferred from [21, Examples 3,4].

When $X_1,\ldots,X_m$ form a Markov chain $X_1 \multimap \ldots \multimap X_m$, it is seen that $\mathrm{SI}(X_{\mathcal{M}}) = \min_{1\leq i\leq m-1}\mathrm{I}(X_i \wedge X_{i+1})$, the minimum mutual information between a pair of adjacent rvs in the chain [21].

Shared information possesses several properties befitting a measure of mutual dependence among multiple rvs. Clearly $\mathrm{SI}(X_{\mathcal{M}}) \geq 0$ and equality holds iff $P_{X_{\mathcal{M}}} = P_{X_A}P_{X_{A^c}}$ for some $A \subsetneq \mathcal{M}$; the latter follows from [21, Theorem 5] and [8], [15], [9]. When $X_1,\ldots,X_m$ are bijections of each other, i.e., $\mathrm{H}(X_i \,|\, X_j) = 0$, $1 \leq i \neq j \leq m$, then $\mathrm{SI}(X_{\mathcal{M}}) = \mathrm{H}(X_1)$, as expected [11].

Next, the secret key capacity interpretation of $\mathrm{SI}(X_{\mathcal{M}})$ [8], [9], [15], [21], [35] implies that upon grouping the rvs $X_1,\ldots,X_m$ into disjoint teams represented by the atoms of any k-partition $\pi = (\pi_1,\ldots,\pi_k)$ of $\mathcal{M}$, $2 \leq k \leq m$, the resulting shared information of the teamed rvs $X_{\pi_1},\ldots,X_{\pi_k}$ can be only larger, i.e.,

$$\mathrm{SI}(X_{\pi_1},\ldots,X_{\pi_k}) \geq \mathrm{SI}(X_1,\ldots,X_m). \tag{3}$$

Suppose that $\pi^* = (\pi_1^*,\ldots,\pi_l^*)$, $l \geq 2$, attains $\mathrm{SI}(X_{\mathcal{M}}) > 0$ (not necessarily uniquely) in Definition 1, i.e,

$$\mathrm{SI}(X_{\mathcal{M}}) = \frac{1}{l-1}D(P_{X_{\mathcal{M}}} \parallel \prod_{u=1}^{l} P_{X_{\pi_u^*}}). \tag{4}$$

A simple but useful observation based on Definition 1, (3) and (4) is that upon agglomerating the rvs in each atom of an optimum partition $\pi^* = (\pi_1^*,\ldots,\pi_l^*)$, the resulting shared information $\mathrm{SI}(X_{\pi_1^*},\ldots,X_{\pi_l^*})$ of the teams $X_{\pi_1^*},\ldots,X_{\pi_k^*}$ equals the shared information $\mathrm{SI}(X_{\mathcal{M}})$ of the (unteamed) rvs $X_1,\ldots,X_m$, i.e, for these special teams (3) holds with equality. This property has benefited information-clustering applications (cf. e.g., [11], [14]).

Shared information satisfies the data processing inequality [11]. For $X_{\mathcal{M}} = (X_1,\ldots,X_m)$, consider $X'_{\mathcal{M}} = (X'_1,\ldots,X'_m)$ where for a fixed $1 \leq j \leq m$, $X'_i = X_i$ for $i \in \mathcal{M}\setminus\{j\}$ and X'_j is obtained as the output of a stochastic matrix $W : \mathcal{X}_j \to \mathcal{X}_j$ with input X_j. Then, $\mathrm{SI}(X'_{\mathcal{M}}) \leq \mathrm{SI}(X_{\mathcal{M}})$.

It is worth comparing $\mathrm{SI}(X_{\mathcal{M}})$ with two well-known measures of correlation among $X_1, \ldots, X_m$, $m \geq 2$, of a similar vein. Watanabe's *total correlation* [45] is defined by

$$\mathcal{C}(X_{\mathcal{M}}) = D(P_{X_{\mathcal{M}}} \parallel \prod_{i=1}^{m} P_{X_i}) = \sum_{i=1}^{m-1} \mathrm{I}(X_{i+1} \wedge X_1, \ldots, X_i) \tag{5}$$

and equals $(m-1)\,\mathcal{I}(\pi)$ for the partition $\pi = (\{1\}, \ldots, \{m\})$ of $\mathcal{M}$ consisting of singleton atoms. By (1) and (5), clearly

$$\mathrm{SI}(X_{\mathcal{M}}) \leq \frac{1}{m-1}\, \mathcal{C}(X_{\mathcal{M}}). \tag{6}$$

Han's *dual total correlation* [26] is defined (equivalently) by

$$\begin{aligned} \mathcal{D}(X_{\mathcal{M}}) &= \sum_{i=1}^{m-1} \mathcal{D}iv P_{X_i} P_{X_{i+1}\cdots X_m} P_{X_1 \cdots X_{i-1}} \\ &= \sum_{i=1}^{m} \mathrm{H}(X_{\mathcal{M}\setminus\{i\}}) - (m-1)\,\mathrm{H}(X_{\mathcal{M}}) \\ &= \mathrm{H}(X_{\mathcal{M}}) - \sum_{i=1}^{m} \mathrm{H}(X_i \,|\, X_{\mathcal{M}\setminus\{i\}}) \\ &= \sum_{i=1}^{m-1} \mathrm{I}(X_i \wedge X_{i+1}, \ldots, X_m \,|\, X_1, \ldots, X_{i-1}), \end{aligned} \tag{7}$$

(with conditioning vacuous for $i = 1$) where the expression in (7) is from [1]. By a straightforward calculation, these measures are seen to enjoy the sandwich

$$\frac{\mathcal{C}(X_{\mathcal{M}})}{m-1} \leq \mathcal{D}(X_{\mathcal{M}}) \leq (m-1)\, \mathcal{C}(X_{\mathcal{M}}) \tag{8}$$

whereby we get from (6) and the first inequality in (8) that

$$\mathrm{SI}(X_{\mathcal{M}}) \leq \frac{\mathcal{C}(X_{\mathcal{M}})}{m-1} \quad \text{and} \quad \mathrm{SI}(X_{\mathcal{M}}) \leq \mathcal{D}(X_{\mathcal{M}}). \tag{9}$$

This makes $\mathrm{SI}(X_{\mathcal{M}})$ a leaner measure of correlation than $\mathcal{C}(X_{\mathcal{M}})$ (upon setting aside the fixed constant $1/(m-1)$) or $\mathcal{D}(X_{\mathcal{M}})$. Significantly, the notion of an optimal partition in $\mathrm{SI}(X_{\mathcal{M}})$ in (1) makes shared information an appealing measure for "local" as well as "global" dependencies among the rvs $X_1, \ldots, X_m$.

Remark 1. When $\mathcal{M} = \{1, 2\}$,

$$\mathrm{SI}(X_1, X_2) = \mathcal{C}(X_1, X_2) = \mathcal{D}(X_1, X_2) = \mathrm{I}(X_1 \wedge X_2).$$

Our focus is on shared information for a Markov chain on a tree.

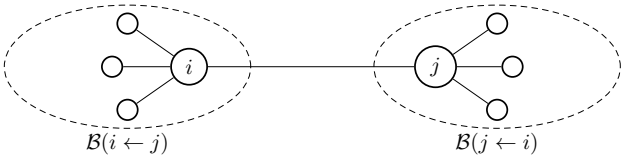

Fig. 1. Notation for a Markov chain on a tree.

Definition 2 (Markov Chain on a Tree)**.** *Let $\mathcal{G} = (\mathcal{M}, \mathcal{E})$ be a tree with vertex set $\mathcal{M} = \{1, \ldots, m\}$, $m \geq 2$, i.e., a connected graph containing no circuits. For (i, j) in the edge set $\mathcal{E}$, let $\mathcal{B}(i \leftarrow j)$ denote the set of all vertices connected with j by a path containing the edge (i, j). Note that $i \in \mathcal{B}(i \leftarrow j)$ but $j \notin \mathcal{B}(i \leftarrow j)$. See Figure 1. The rvs $X_1, \ldots, X_m$ form a* Markov Chain on a Tree *(MCT) $\mathcal{G}$ if for every $(i, j) \in \mathcal{E}$, the conditional pmf of X_j given $X_{\mathcal{B}(i\leftarrow j)} = \{X_l : l \in \mathcal{B}(i \leftarrow j)\}$ depends only on X_i. Specifically, X_j is conditionally independent of $X_{\mathcal{B}(i\leftarrow j)\setminus\{i\}}$ when conditioned on X_i. Thus, $P_{X_{\mathcal{M}}}$ is such that for each $(i, j) \in \mathcal{E}$,*

$$P_{X_j \,|\, X_{\mathcal{B}(i\leftarrow j)}} = P_{X_j \,|\, X_i}, \tag{10}$$

or, equivalently,

$$X_j \multimap X_i \multimap X_{\mathcal{B}(i\leftarrow j)\setminus\{i\}}. \tag{11}$$

When $\mathcal{G}$ is a chain, an MCT reduces to a standard Markov chain.

Remark 2. With an abuse of terminology, we shall use $\mathcal{G} = (\mathcal{M}, \mathcal{E})$ to refer to a tree and also to the associated MCT.

Example 2. Let $m = 2^l - 1$ for some positive integer l. Consider a balanced binary tree with l levels. Label the nodes progressively at each level and downwards, with the root node (at level 1) being 1 and the 2^{l-1} leaves (at level l) being $2^{l-1}, \ldots, 2^{l-1} + 2^{l-1} - 1 = 2^l - 1 = m$. Let $X_1, Z_1, \ldots, Z_{m-1}$ be mutually independent rvs where $X_1 = \mathrm{Ber}(0.5)$ and $Z_i = \mathrm{Ber}(p_i)$ with $0 < p_i < 0.5$, $i = 1, \ldots, m-1$. For $i = 2, \ldots, m$, set $X_i = X_{\lfloor i/2 \rfloor} + Z_{i-1}$ where "+" denotes addition modulo 2; note that X_i is determined by $(X_1, Z_1, \ldots, Z_{i-1})$, and $P(X_i = 0) = P(X_i = 1) = 0.5$.

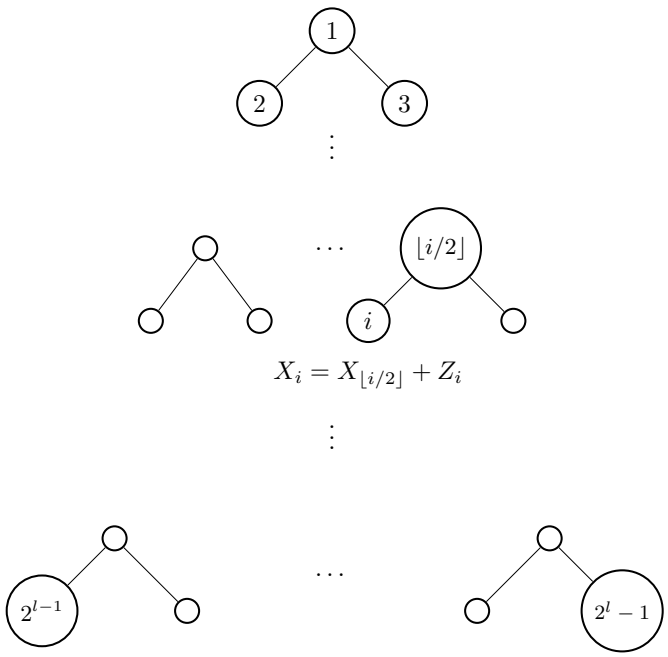

Fig. 2. Example 2 of an MCT.

Assign rv X_i to vertex i, $i = 1, \ldots, m$. See Figure 2. Then $X_1, \ldots, X_m$ form an MCT. Specifically, for any edge (i, j)

where i is the parent of $j \geq 2$, we have

$$\begin{aligned}
&P(X_j = x_j \,|\, X_{\mathcal{B}(i\leftarrow j)} = x_{\mathcal{B}(i\leftarrow j)}) \\
&= P(X_j = x_j \,|\, X_i = x_i, X_{\mathcal{B}(i\leftarrow j)\setminus\{i\}} = x_{\mathcal{B}(i\leftarrow j)\setminus\{i\}}) \\
&= P(x_i + Z_{j-1} = x_j \,|\, X_i = x_i, X_{\mathcal{B}(i\leftarrow j)\setminus\{i\}} = x_{\mathcal{B}(i\leftarrow j)\setminus\{i\}}) \\
&= P(Z_{j-1} = x_j + x_i) \\
&= P(X_j = x_j \,|\, X_i = x_i)
\end{aligned}$$

where the last two inequalities are by the independence of Z_{j-1} and $X_{\mathcal{B}(i\leftarrow j)}$, the latter rv being a function of $X_1, Z_{\{1,\ldots,m-1\}\setminus\{j-1\}}$.

III. Properties of an MCT

We develop properties of an MCT that will play a role in characterizing shared information, and also are of independent interest. These include the concept of an agglomerated MCT, and notions of local and global Markov properties.

The main conclusion of this section is that the MCT as defined in (11) has the global Markov property, which, in turn, implies (11); the proofs in this section, however, are based on (11).

We begin with agglomeration.

Lemma 1. *For the MCT $\mathcal{G} = (\mathcal{M}, \mathcal{E})$, for every $(i, j) \in \mathcal{E}$,*

$$\mathrm{I}(X_{\mathcal{B}(i\leftarrow j)} \wedge X_{\mathcal{B}(j\leftarrow i)}) = \mathrm{I}(X_i \wedge X_j), \tag{12}$$

i.e.,

$$X_{\mathcal{B}(i\leftarrow j)\setminus\{i\}} \multimap\!\circ\!\!-\!\!\circ\!\multimap X_i \multimap\!\circ\!\!-\!\!\circ\!\multimap X_j \multimap\!\circ\!\!-\!\!\circ\!\multimap X_{\mathcal{B}(j\leftarrow i)\setminus\{j\}}. \tag{13}$$

Proof. The assertion (12) is proved in Section A. Turning to (13), by the chain rule (12) is equivalent to

$$\begin{aligned}
&X_{\mathcal{B}(i\leftarrow j)\setminus\{i\}} \multimap\!\circ\!\!-\!\!\circ\!\multimap X_i \multimap\!\circ\!\!-\!\!\circ\!\multimap X_j \\
&\text{and } X_{\mathcal{B}(i\leftarrow j)} \multimap\!\circ\!\!-\!\!\circ\!\multimap X_j \multimap\!\circ\!\!-\!\!\circ\!\multimap X_{\mathcal{B}(j\leftarrow i)\setminus\{j\}}
\end{aligned}$$

which, in turn, are together tantamount to (13). □

Definition 3 (Agglomerated Tree)**.** *Consider a k-partition $\pi = (\pi_1, \ldots, \pi_k)$ of $\mathcal{M}$, $2 \leq k \leq m-1$, where each π_i, $1 \leq i \leq k$, is connected. The tree $\mathcal{G}' = (\mathcal{M}', \mathcal{E}')$ with vertex set $\mathcal{M}' = \{\pi_1, \ldots, \pi_k\}$ and edge set*

$$\mathcal{E}' = \{(\pi_{i'}, \pi_{j'}) : \exists i \in \pi_{i'}, j \in \pi_{j'} \text{ s.t. } (i, j) \in \mathcal{E}\}$$

is termed an agglomerated tree. Note that $\mathcal{G}' = (\mathcal{M}', \mathcal{E}')$ is a tree since $\mathcal{G} = (\mathcal{M}, \mathcal{E})$ is a tree and each π_i, $1 \leq i \leq k$, is connected.

Lemma 2. *Consider the agglomerated tree $\mathcal{G}' = (\mathcal{M}', \mathcal{E}')$ in Definition 3. If $X_1, \ldots, X_m$ form an MCT $\mathcal{G} = (\mathcal{M}, \mathcal{E})$, then $X_{\pi_1}, \ldots, X_{\pi_k}$ form an MCT $\mathcal{G}' = (\mathcal{M}', \mathcal{E}')$.*

Proof. By an obvious extension of Definition 2 to the agglomerated tree $\mathcal{G}' = (\mathcal{M}', \mathcal{E}')$, the lemma would follow upon showing that for every $(\pi_{i'}, \pi_{j'}) \in \mathcal{E}'$, it holds that

$$X_{\pi_{j'}} \multimap\!\circ\!\!-\!\!\circ\!\multimap X_{\pi_{i'}} \multimap\!\circ\!\!-\!\!\circ\!\multimap X_{\mathcal{B}(\pi_{i'}\leftarrow\pi_{j'})\setminus\pi_{i'}}. \tag{14}$$

For any $(\pi_{i'}, \pi_{j'}) \in \mathcal{E}'$, there exist by Definition 3 $i \in \pi_{i'}$, $j \in \pi_{j'}$ with $(i, j) \in \mathcal{E}$. By Lemma 1,

$$X_{\pi_{j'}} \multimap\!\circ\!\!-\!\!\circ\!\multimap X_i \multimap\!\circ\!\!-\!\!\circ\!\multimap X_{\mathcal{B}(i\leftarrow j)\setminus\{i\}} \tag{15}$$

which, upon noting that $X_{\mathcal{B}(\pi_{i'}\leftarrow\pi_{j'})\setminus\pi_{i'}} \subseteq X_{\mathcal{B}(i\leftarrow j)\setminus\{i\}}$ and applying the chain rule of mutual information to (15), gives (14). □

Next, we turn to local and global Markovity of an MCT.

Definition 4 (Separation)**.** *For a tree $\mathcal{G} = (\mathcal{M}, \mathcal{E})$, let A, B and S be disjoint, nonempty subsets of $\mathcal{M}$. Then S* separates *A and B if for every $a \in A$ and $b \in B$, the path connecting a and b has at least one vertex $s = s(a, b)$ in S.*

The notion of maximally connected subset will be used below: $A' \subseteq A$ is a maximally connected subset of A if A' is connected and the addition to A' of any vertex $u \in A \setminus A'$ renders $A' \cup \{u\}$ to be disconnected. By convention, we shall take singleton elements to be connected. In general, any connected component of A that is not maximally connected can be enlarged to absorb vertices outside it in A that do not render the union to be disconnected.

Theorem 3 (Global Markov property)**.** *For an MCT $\mathcal{G} = (\mathcal{M}, \mathcal{E})$, let A, B and S be disjoint, nonempty subsets of $\mathcal{M}$ such that S separates A and B. Then*

$$X_A \multimap\!\circ\!\!-\!\!\circ\!\multimap X_S \multimap\!\circ\!\!-\!\!\circ\!\multimap X_B. \tag{16}$$

Conversely, if $X_1, \ldots, X_m$ are assigned to the vertices of a tree $\mathcal{G} = (\mathcal{M}, \mathcal{E})$ and satisfy (16) *for every such A, B, S, they form an MCT.*

Remark 3. The Hammersley-Clifford Theorem [31, Theorem 3.9] implies the equivalence above for strictly positive joint pmfs, i.e., with $P(x_{\mathcal{M}}) > 0$ for every $x_{\mathcal{M}} \in \mathcal{X}_{\mathcal{M}}$. Theorem 3 shows that for an MCT, this equivalence holds also for a joint pmf $P_{X_{\mathcal{M}}}$ that is *not* strictly positive.

Remark 4. The "global" Markov property (16) [31, Section 3.2.1] clearly implies (10), (11) in Definition 2. Theorem 3 asserts that for an MCT, (16) is, in fact, equivalent to (10), (11).

Our proof of Theorem 3 in Appendix B relies on showing that an MCT has the "local Markov property" of Lemma 4 below. The notion of *neighborhood* is pertinent. Lemma 4, too, is proved in Appendix B.

Definition 5 (Neighborhood)**.** *For each $i \in \mathcal{M}$, its neighborhood is $\mathcal{N}(i) = \{j \in \mathcal{M} : (i, j) \in \mathcal{E}\}$; note that $i \notin \mathcal{N}(i)$. Similarly, the neighborhood of $A \subsetneq \mathcal{M}$ is $\mathcal{N}(A) = \cup_{i\in A}\mathcal{N}(i) \setminus A$.*

Lemma 4 (Local Markov property)**.** *Consider an MCT $\mathcal{G} = (\mathcal{M}, \mathcal{E})$. For each $i \in \mathcal{M}$,*

$$X_i \multimap\!\circ\!\!-\!\!\circ\!\multimap X_{\mathcal{N}(i)} \multimap\!\circ\!\!-\!\!\circ\!\multimap X_{\mathcal{M}\setminus(\{i\}\cup\mathcal{N}(i))}. \tag{17}$$

Furthermore, for every $A \subsetneq \mathcal{M}$ with no edge connecting any two vertices in it,

$$X_A \multimap\!\circ\!\!-\!\!\circ\!\multimap X_{\mathcal{N}(A)} \multimap\!\circ\!\!-\!\!\circ\!\multimap X_{\mathcal{M}\setminus(A\cup\mathcal{N}(A))}. \tag{18}$$

Remark 5. For $A \subsetneq \mathcal{M}$ as in Lemma 4, $\mathcal{N}(A) = \cup_{i\in A}\mathcal{N}(i)$.

The relationships among the MCT definition (11), and local and global Markov properties are summarized by the following lemma.

Lemma 5. *It holds that*
(i) MCT and the global Markov property are equivalent;
(ii) MCT implies the local Markov property;
(iii) the local Markov property implies neither the global Markov property nor the MCT.

Proof. (i), (ii) The proofs are contained in Theorem 3 and Lemma 4, respectively.
(iii) The following example is used in [31, Example 3.5] to show that local Markovity does not imply global Markovity. Let U and Z be independent $\mathrm{Ber}(0.5)$ rvs, and let $W = U$, $Y = Z$ and $X = WY$. Consider the graph

$$U - W - X - Y - Z$$

which has the local Markov property. Since

$$\begin{aligned}\mathrm{I}(Z \wedge U, W \,|\, X) &= \mathrm{I}(Z \wedge U \,|\, UZ) \\ &= \mathrm{H}(Z \,|\, UZ) \\ &= 0.75\ \mathrm{H}(Z \,|\, UZ = 0) \\ &= 0.75\, h(1/3),\end{aligned}$$

the claim in (iii) is true. □

IV. Shared Information for a Markov Chain on a Tree

We present a new proof of an explicit characterization of $\mathrm{SI}(X_{\mathcal{M}})$ for an MCT. The expression in Theorem 6 below was obtained first in [21] relying on its secrecy capacity interpretation. Specifically, it was computed using a linear program for said capacity and seen to equal an upper bound corresponding to shared information ([21, Examples 4, 7]). The new approach below works directly with the definition of shared information in Definition 1. Also, it differs materially from the treatment in [14, Section 4] for a model that appears to differ from ours.

While the upper bound for $\mathrm{SI}(X_{\mathcal{M}})$ below is akin to that involving secret key capacity in [21], the proof of the lower bound uses an altogether new method based on the structure of a "good" partition π in Definition 1.

Theorem 6. *Let $\mathcal{G} = (\mathcal{M}, \mathcal{E})$ be an MCT with pmf $P_{X_{\mathcal{M}}}$ in* (10)*. Then*

$$\mathrm{SI}(X_{\mathcal{M}}) = \min_{(i,j)\in\mathcal{E}} \mathrm{I}(X_i \wedge X_j). \tag{19}$$

Remark 6. For later use, let $(\bar{i}, \bar{j})$ be the (not necessarily unique) minimizer in the right-side of (19)

Example 3. For the MCT in Example 2, $\mathrm{SI}(X_{\mathcal{M}}) = 1 - h(p^*)$, where $p^* = \max_{1\le i\le m-1} p_i < 0.5$. Thus, the 2-partition obtained by cutting the (not necessarily unique) weakest correlating edge attains the minimum in Definition 1.

Proof. As shown in [21],

$$\mathrm{SI}(X_{\mathcal{M}}) \le \min_{(i,j)\in\mathcal{E}} \mathrm{I}(X_i \wedge X_j) \tag{20}$$

and is seen as follows. For each $(i,j) \in \mathcal{E}$, consider a 2-partition of $\mathcal{M}$, viz. $\pi = \pi((i,j)) = (\pi_1, \pi_2)$ where $\pi_1 = \mathcal{B}(i \leftarrow j)$, $\pi_2 = \mathcal{B}(j \leftarrow i)$. Then,

$$\begin{aligned}\mathrm{I}(X_{\pi_1} \wedge X_{\pi_2}) &= \mathrm{I}(X_{\mathcal{B}(i\leftarrow j)} \wedge X_{\mathcal{B}(j\leftarrow i)}) \\ &= \mathrm{I}(X_i \wedge X_j), \text{ by Lemma 1.}\end{aligned} \tag{21}$$

Hence,

$$\mathrm{SI}(X_{\mathcal{M}}) \le \mathrm{I}(X_{\pi_1} \wedge X_{\pi_2}) = \mathrm{I}(X_i \wedge X_j), \qquad (i,j) \in \mathcal{E}$$

leading to (20).

Next, we show that

$$\mathrm{SI}(X_{\mathcal{M}}) \ge \min_{(i,j)\in\mathcal{E}} \mathrm{I}(X_i \wedge X_j). \tag{22}$$

This is done in two steps. First, we show that for any k-partition π of $\mathcal{M}$, $2 \le k \le m$, *with (individually) connected atoms*,

$$\mathcal{I}(\pi) \ge \min_{(i,j)\in\mathcal{E}} \mathrm{I}(X_i \wedge X_j).$$

Second, we argue that for any k-partition $\pi = (\pi_1, \ldots, \pi_k)$ containing disconnected atoms, there exists a k'-partition $\pi' = (\pi', \ldots, \pi'_{k'})$, possibly with $k' \ne k$, and with fewer disconnected atoms such that $\mathcal{I}(\pi') \le \mathcal{I}(\pi)$.

Step 1: Let $\pi = (\pi_1, \ldots, \pi_k)$, $k \ge 2$, be a k-partition with each atom being a connected set. By Lemma 2, $X_{\pi_1}, \ldots, X_{\pi_k}$ form an agglomerated MCT $\mathcal{G}' = (\mathcal{M}', \mathcal{E}')$ as in Definition 3. Furthermore, by Lemma 1, if π_u and π_v in $\mathcal{M}'$ are connected by an edge $(\pi_u, \pi_v) \in \mathcal{E}'$, then there exist $i \in \pi_u$, $j \in \pi_v$, say, such that $(i,j) \in \mathcal{E}$, whence

$$\mathrm{I}(X_{\pi_u} \wedge X_{\pi_v}) = \mathrm{I}(X_i \wedge X_j). \tag{23}$$

Now, let $\pi_1, \ldots, \pi_k$ be an enumeration of the atoms, obtained from a breadth-first search [18, Ch. 22] run on the agglomerated tree with π_1 as the root vertex. Then,

$$\begin{aligned}\mathcal{I}(\pi) &= \frac{1}{k-1} D(P_{X_{\mathcal{M}}} \,\|\, \prod_{u=1}^{k} P_{X_{\pi_u}}) \\ &= \frac{1}{k-1}\left[\sum_{u=1}^{k} \big(\mathrm{H}(X_{\pi_u}) - \mathrm{H}(X_{\pi_u} \,|\, X_{\pi_1}, \ldots, X_{\pi_{u-1}})\big)\right] \\ &= \frac{1}{k-1} \sum_{u=2}^{k} \mathrm{I}(X_{\pi_u} \wedge X_{\pi_1}, \ldots, X_{\pi_{u-1}}) \\ &= \frac{1}{k-1} \sum_{u=2}^{k} \mathrm{I}(X_{\pi_u} \wedge X_{\mathrm{parent}(\pi_u)}) \qquad (24) \\ &\ge \min_{(\pi_u,\pi_v)\in\mathcal{E}'} \mathrm{I}(X_{\pi_u} \wedge X_{\pi_v}) \\ &\ge \min_{(i,j)\in\mathcal{E}} \mathrm{I}(X_i \wedge X_j).\end{aligned}$$

By the breadth-first search algorithm [18, Ch. 22], $\pi_1, \ldots, \pi_{u-1}$ are either at the same depth as π_u or are above it (and include $\mathrm{parent}(\pi_u)$). This, combined with Theorem 3, gives (24). The last inequality is by (23).

Step 2: Consider first the case $k = 2$. Take any 2-partition $\pi = (\pi_1, \pi_2)$ with possibly disconnected atoms, where $\pi_1 = \cup_{\rho=1}^{r} C_\rho$ and $\pi_2 = \cup_{\sigma=1}^{s} D_\sigma$ are unions of disjoint

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2024.3353769

components. Since π_1 is connected to π_2, some C_ρ and D_σ must be connected by some edge (i,j) in $\mathcal{E}$, so that

$$\begin{aligned}\mathcal{I}(\pi) = \mathrm{I}(X_{\pi_1} \wedge X_{\pi_2}) &\geq \mathrm{I}(X_{C_\rho} \wedge X_{D_\sigma}) \\ &\geq \mathrm{I}(X_i \wedge X_j) \geq \mathrm{I}(X_{\bar{i}} \wedge X_{\bar{j}})\end{aligned}$$

where the final lower bound, with $\bar{i}, \bar{j}$ as in Remark 6, is achieved by the 2-partition with connected atoms $(\mathcal{B}(\bar{i} \leftarrow \bar{j}), \mathcal{B}(\bar{j} \leftarrow \bar{i}))$ as in (21).

Next, consider a k-partition $\pi = (\pi_1, \ldots, \pi_k)$, $k \geq 3$, and suppose that the atom π_1 is not connected. Without loss of generality, assume π_1 to be the (disjoint) union of maximally connected subsets $A_1, \ldots, A_t$, $t \geq 2$, of π_1 (which, at an extreme, can be the individual vertices constituting π_1).

Take any A_l, say $A_l = A_{\bar{l}}$, and consider all its boundary edges, namely those edges for which one vertex is in $A_{\bar{l}}$ and the other outside it. As $A_{\bar{l}}$ is maximally connected in π_1, for each boundary edge the outside vertex cannot belong to π_1 and so must lie in $\mathcal{M} \setminus \pi_1$. Also, every such outside vertex associated with $A_{\bar{l}}$ must be the root of a subtree and, like $A_{\bar{l}}$, every A_l, $l \neq \bar{l}$, too, must be a subset of one such subtree linked to $A_{\bar{l}}$ – owing to connectedness within $A_{\bar{l}}$. Furthermore, since $A_1, \ldots, A_t$ are connected, and only through the subtrees rooted in $\mathcal{M} \setminus \pi_1$, there must exist at least one A_l such that all $A_{l'}$s, $l' \neq l$, are subsets of one subtree linked to A_l. In other words, denoting this A_l as A, we note that A has the property that

$$\pi_1 \setminus A = \bigcup_{\substack{l \in \{1, \cdots, t\}: \\ A_l \neq A}} A_l$$

is contained entirely in a subtree rooted at an outside vertex associated with A and lying in $\mathcal{M} \setminus \pi_1$. Let this vertex be $j \in \mathcal{M} \setminus \pi_1$, and let $\pi_u \in \pi$ be the atom that contains j. Since vertex j separates A from $\pi_1 \setminus A$, so does π_u. By Theorem 3, it follows that

$$A \multimap\!\circ\!\!-\, \pi_u \,-\!\circ\!\multimap\, \pi_1 \setminus A$$

whereby, using the data processing inequality,

$$\mathrm{I}(X_A \wedge X_{\pi_1 \setminus A}) \leq \mathrm{I}(X_{\pi_u} \wedge X_{\pi_1 \setminus A}) \leq \mathrm{I}(X_{\pi_u} \wedge X_{\pi_1}). \quad (25)$$

Next, consider the $(k-1)$-partition π' and the $(k+1)$-partition π'' of $\mathcal{M}$, defined by

$$\pi' = \left(\pi_1 \cup \pi_u, \{\pi_v\}_{v \neq 1, v \neq u}\right), \quad (26)$$

$$\pi'' = \left(\pi_1 \setminus A, A, \pi_u, \{\pi_v\}_{v \neq 1, v \neq u}\right). \quad (27)$$

Then,

$$\mathcal{I}(\pi) = \frac{1}{k-1}\Bigg[\mathrm{H}(X_{\pi_1}) + \mathrm{H}(X_{\pi_u}) + \sum_{v \neq 1, v \neq u} \mathrm{H}(X_{\pi_v}) - \mathrm{H}(X_{\mathcal{M}})\Bigg],$$

$$\mathcal{I}(\pi') = \frac{1}{k-2}\Bigg[\mathrm{H}(X_{\pi_1 \cup \pi_u}) + \sum_{v \neq 1, v \neq u} \mathrm{H}(X_{\pi_v}) - \mathrm{H}(X_{\mathcal{M}})\Bigg],$$

$$\mathcal{I}(\pi'') = \frac{1}{k}\Bigg[\mathrm{H}(X_{\pi_1 \setminus A}) + \mathrm{H}(X_A) + \mathrm{H}(X_{\pi_u}) + \sum_{v \neq 1, v \neq u} \mathrm{H}(X_{\pi_v}) - \mathrm{H}(X_{\mathcal{M}})\Bigg].$$

We claim that

$$\mathcal{I}(\pi) \geq \min\{\mathcal{I}(\pi'), \mathcal{I}(\pi'')\}. \quad (28)$$

Referring to (26) and (27), we can infer from the claim (28) that for a given k-partition π with a disconnected atom π_1 as above, merging a disconnected atom with another atom (as in (26)) or breaking it to create a connected atom (as in (27)), lead to partitions π' or π'', of which at least one has a lower $\mathcal{I}$-value than π. This argument is repeated until a final partition with connected atoms is reached that has the following form: considering the set of all maximally connected components of the atoms of $\pi = (\pi_1, \ldots, \pi_k)$, the final partition will consist of *connected* unions of such components. (A connected π_i already constitutes such a component.)

It remains to show (28). Suppose (28) were not true, i.e.,

$$\mathcal{I}(\pi) < \min\{\mathcal{I}(\pi'), \mathcal{I}(\pi'')\}.$$

Then,

$$\begin{aligned}\mathcal{I}(\pi) < \mathcal{I}(\pi') &\Leftrightarrow (k-2)\,\mathcal{I}(\pi) < (k-2)\,\mathcal{I}(\pi') \\ &\Leftrightarrow \mathrm{I}(X_{\pi_u} \wedge X_{\pi_1}) < \mathcal{I}(\pi), \quad (29)\end{aligned}$$

and similarly,

$$\begin{aligned}\mathcal{I}(\pi) < \mathcal{I}(\pi'') &\Leftrightarrow k\,\mathcal{I}(\pi) < k\,\mathcal{I}(\pi'') \\ &\Leftrightarrow \mathcal{I}(\pi) < \mathrm{I}(X_{\pi_1 \setminus A} \wedge X_A) \quad (30)\end{aligned}$$

where the second equivalences in (29) and (30) are obtained by straightforward manipulation. By (29) and (30),

$$\mathrm{I}(X_{\pi_u} \wedge X_{\pi_1}) < \mathrm{I}(X_{\pi_1 \setminus A} \wedge X_A)$$

which contradicts (25). Hence, (28) is true. □

V. Estimating $\mathrm{SI}(X_{\mathcal{M}})$ for an MCT

We consider the estimation of $\mathrm{SI}(X_{\mathcal{M}})$ when the pmf $P_{X_{\mathcal{M}}}$ of $X_{\mathcal{M}} = (X_1, \ldots, X_m)$ is unknown to an "agent" who, however, is assumed to know the tree $\mathcal{G} = (\mathcal{M}, \mathcal{E})$. We assume further in this section that $\mathcal{X}_1 = \cdots = \mathcal{X}_m = \mathcal{X}$, say, and also that the minimizing edge $(\bar{i}, \bar{j})$ on the right side of (19) is unique. By Theorem 6, $\mathrm{SI}(X_{\mathcal{M}})$ equals the minimum mutual information across an edge in the tree $\mathcal{G}$. Treating the determination of this edge as a correlated bandits problem of best arm pair identification, we provide an algorithm to home in on it, and analyze its error performance and associated sample complexity. *The estimate of shared information is taken to be the mutual information across the best arm-pair thus identified.* Our estimation procedure is motivated by the form of $\mathrm{SI}(X_{\mathcal{M}})$ in Theorem 6.

A. Preliminaries

As stated, estimation of $\mathrm{SI}(X_{\mathcal{M}})$ for an MCT will entail estimating $\mathrm{I}(X_i \wedge X_j)$, $(i,j) \in \mathcal{E}$. We first present pertinent tools that will be used to this end.

Let $(X_t, Y_t)_{t=1}^n$ be $n \geq 1$ independent and identically distributed (i.i.d.) repetitions of rv (X,Y) with (unknown) pmf P_{XY} of *assumed full support* on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are finite sets. For $(\mathbf{x}, \mathbf{y})$ in $\mathcal{X}^n \times \mathcal{Y}^n$, let $Q_{\mathbf{xy}}^{(n)}$ represent its joint type on $\mathcal{X} \times \mathcal{Y}$ (cf. [20, Ch. 2, 3]). Also, let $Q_{\mathbf{x}}^{(n)}$ (resp. $Q_{\mathbf{y}}^{(n)}$) represent the (marginal) type of $\mathbf{x}$ (resp. $\mathbf{y}$).

A well-known estimator for $\mathrm{I}(X \wedge Y) = \mathrm{I}_{P_{XY}}(X \wedge Y)$ on the basis of $(\mathbf{x}, \mathbf{y})$ in $\mathcal{X}^n \times \mathcal{Y}^n$ is the *empirical mutual information* (EMI) estimator $\mathrm{I}_{\mathsf{EMI}}^{(n)}$, based on EMI [25], [20, Ch. 3], defined by

$$\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{x} \wedge \mathbf{y}) = \mathrm{H}(Q_{\mathbf{x}}^{(n)}) + \mathrm{H}(Q_{\mathbf{y}}^{(n)}) - \mathrm{H}(Q_{\mathbf{xy}}^{(n)}). \tag{31}$$

Throughout this section, $(\mathbf{X}, \mathbf{Y})$ will represent n i.i.d. repetitions of the rv (X,Y).

Lemma 7 (Bias of EMI estimator)**.** *The bias*

$$\mathrm{Bias}(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X} \wedge \mathbf{Y})) \triangleq \mathbb{E}_{P_{XY}}\left[\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X} \wedge \mathbf{Y})\right] - \mathrm{I}(X \wedge Y)$$

satisfies

$$-\log\left(1 + \frac{|\mathcal{X}|-1}{n}\right)\left(1 + \frac{|\mathcal{Y}|-1}{n}\right) \leq \mathrm{Bias}(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X} \wedge \mathbf{Y})) \leq \log\left(1 + \frac{|\mathcal{X}|\,|\mathcal{Y}|-1}{n}\right).$$

Proof. The proof follows immediately from [38, Proposition 1]. □

A concentration bound for the estimator $\mathrm{I}_{\mathsf{EMI}}^{(n)}$ in (31) using techniques from [2], is given by

Lemma 8. *Given $\epsilon > 0$ and for every $n \geq 1$,*

$$P_{XY}\left(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X} \wedge \mathbf{Y}) - \mathbb{E}_{P_{XY}}\left[\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X} \wedge \mathbf{Y})\right] \geq \epsilon\right) \leq \exp\left(-\frac{2n\epsilon^2}{36\log^2 n}\right).$$

The same bound applies upon replacing $\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X} \wedge \mathbf{Y})$ by $-\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X} \wedge \mathbf{Y})$ above.

Proof. The empirical mutual information $\mathrm{I}_{\mathsf{EMI}}^{(n)} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}^+ \cup \{0\}$ satisfies the bounded differences property, namely

$$\max_{\substack{(\mathbf{x},\mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n \\ (x_i', y_i') \in \mathcal{X} \times \mathcal{Y}}} \left| \mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{x} \wedge \mathbf{y}) - \mathrm{I}_{\mathsf{EMI}}^{(n)}((x_1^{i-1}, x_i', x_{i+1}^n) \wedge (y_1^{i-1}, y_i', y_{i+1}^n)) \right| \leq \frac{6\log n}{n} \tag{32}$$

for $1 \leq i \leq n$, where for $l < k$, $x_l^k = (x_l, x_{l+1}, \ldots, x_k)$.

To see this, we note that changing

$$(\mathbf{x}, \mathbf{y}) = ((x_1, \ldots, x_n), (y_1, \ldots, y_n)) \to ((x_1^{i-1}, x_i', x_{i+1}^n), (y_1^{i-1}, y_i', y_{i+1}^n))$$

amounts to changing at most two components in the joint type $Q_{\mathbf{xy}}^{(n)}$ and marginal types $Q_{\mathbf{x}}^{(n)}$ and $Q_{\mathbf{y}}^{(n)}$; in each of these three cases, the probability of one symbol or one pair of symbols decreases by $1/n$ and that of another increases by $1/n$. The difference between the corresponding empirical entropies is given in each case by the sum of two terms. For instance, one such term for the joint empirical entropy is given by

$$\left| Q_{\mathbf{xy}}^{(n)}(x_i, y_i) \log Q_{\mathbf{xy}}^{(n)}(x_i, y_i) - \left(Q_{\mathbf{xy}}^{(n)}(x_i, y_i) - \frac{1}{n}\right) \log \left(Q_{\mathbf{xy}}^{(n)}(x_i, y_i) - \frac{1}{n}\right) \right|.$$

Each of these terms is $\leq \log n/n$, using the inequality [2]

$$\left| \frac{j+1}{n} \log \frac{j+1}{n} - \frac{j}{n} \log \frac{j}{n} \right| \leq \frac{\log n}{n}, \qquad 0 \leq j < n.$$

The bound in (32) is obtained upon applying the triangle inequality twice in each of the three mentioned cases. The claim of the lemma then follows by a standard application of McDiarmid's Bounded Differences Inequality [44, Theorem 2.9.1]. □

Since we seek to identify the edge with the smallest mutual information across it, we next present a technical lemma that bounds above the probability that the estimates of the mutual information between two pairs of rvs are in the wrong order. Our proof uses Lemma 8. Let (X,Y) and (X',Y') be two pairs of rvs with pmfs P_{XY} and $P_{X'Y'}$, respectively, on the (common) alphabet $\mathcal{X} \times \mathcal{Y}$, such that $\mathrm{I}(X \wedge Y) < \mathrm{I}(X' \wedge Y')$. Let

$$\Delta = \mathrm{I}(X' \wedge Y') - \mathrm{I}(X \wedge Y) > 0. \tag{33}$$

By Lemma 7, $\mathrm{I}_{\mathsf{EMI}}^{(n)}$ is asymptotically unbiased and, in particular, we can make $\mathrm{Bias}(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X} \wedge \mathbf{Y})), \mathrm{Bias}(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X}' \wedge \mathbf{Y}')) < \Delta/2$ by choosing n large enough, for instance,

$$n > \max\left\{\frac{|\mathcal{X}|^2 - 1}{2^{\Delta/2} - 1}, \frac{|\mathcal{X}| - 1}{2^{\Delta/4} - 1}\right\}. \tag{34}$$

The upper bound on the probability of ordering error depends on the bias of $\mathrm{I}_{\mathsf{EMI}}^{(n)}$ and decreases with decreasing bias.

Lemma 9. *With (X,Y), (X',Y') and n as in* (34)*,*

$$P\left(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X} \wedge \mathbf{Y}) \geq \mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X}' \wedge \mathbf{Y}')\right) \leq 2\max\left\{\exp\left(-\frac{2n\left(\Delta/2 - \mathrm{Bias}\left(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X} \wedge \mathbf{Y})\right)\right)^2}{36\log^2 n}\right), \exp\left(-\frac{2n\left(\Delta/2 - \mathrm{Bias}\left(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X}' \wedge \mathbf{Y}')\right)\right)^2}{36\log^2 n}\right)\right\}.$$

Proof. Recalling (33), we have

$$\begin{aligned}
&P\left(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X}\wedge\mathbf{Y}) \geq \mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X}'\wedge\mathbf{Y}')\right)\\
&= P\Big(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X}\wedge\mathbf{Y})\\
&\qquad - \mathrm{I}(X\wedge Y) - \mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X}'\wedge\mathbf{Y}') + \mathrm{I}(X'\wedge Y') \geq \Delta\Big)\\
&\leq P\left(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X}\wedge\mathbf{Y}) - \mathrm{I}(X\wedge Y) \geq \Delta/2\right)\\
&\qquad + P\left(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X}'\wedge\mathbf{Y}') - \mathrm{I}(X'\wedge Y') \leq -\Delta/2\right) \quad (35)
\end{aligned}$$

Using Lemma 8, and in view of (33), (34),

$$\begin{aligned}
&P\left(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X}\wedge\mathbf{Y}) - \mathrm{I}(X\wedge Y) \geq \Delta/2\right)\\
&= P\Big(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X}\wedge\mathbf{Y}) - \mathbb{E}\left[\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X}\wedge\mathbf{Y})\right]\\
&\qquad\qquad \geq \Delta/2 - \mathrm{Bias}\left(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X}\wedge\mathbf{Y})\right)\Big)\\
&\leq \exp\left(-\frac{2n\left(\Delta/2 - \mathrm{Bias}\left(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X}\wedge\mathbf{Y})\right)\right)^2}{36\log^2 n}\right), \quad (36)
\end{aligned}$$

and similarly,

$$\begin{aligned}
&P\left(\mathrm{I}(X'\wedge Y') - \mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X}'\wedge\mathbf{Y}') \geq \Delta/2\right)\\
&\leq \exp\left(-\frac{2n\left(\Delta/2 - \mathrm{Bias}\left(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X}'\wedge\mathbf{Y}')\right)\right)^2}{36\log^2 n}\right). \quad (37)
\end{aligned}$$

The claimed bound follows by using (36) and (37) in (35). □

B. Bandit algorithm for estimating $\mathrm{SI}(X_{\mathcal{M}})$

The following bandit-based method identifies the best arm pair corresponding to the edge of the MCT across which mutual information is minimal. For an introduction to the fundamentals of bandit algorithms, see [30].

In the parlance of banditry, the environment has m arms, one arm corresponding to each vertex in $\mathcal{G} = (\mathcal{M}, \mathcal{E})$. The agent can pull, in any step, two arms that are connected by an edge in $\mathcal{E}$. Each action of the agent is specified by the pair (i, j), $1 \leq i < j \leq m$, $(i, j) \in \mathcal{E}$, with associated reward being the realizations $(X_i = x_i, X_j = x_j)$. The agent is allowed to pull a total of N pairs of arms, say, using *uniform sampling*, where N will be specified below. A pulling of a pair of arms can be viewed also as pulling the corresponding connecting edge, thereby rendering it a traditional stochastic bandit problem. We resort to a uniform sampling strategy for the sake of simplicity.

Definition 6 (Uniform sampling)**.** *In uniform sampling, pairs of rvs corresponding to edges of the tree are sampled equally often. Specifically, each pair of rvs (X_i, X_j), $(i, j) \in \mathcal{E}$, is sampled n times over nonoverlapping time instants. Hence, an agent pulls a total of N pairs of arms, where $N = |\mathcal{E}|\, n$.*

By means of these actions, the agent seeks to form estimates of all two-dimensional marginal pmfs $P_{X_iX_j}$ and of the corresponding $\mathrm{I}(X_i \wedge X_j)$ for (i, j) as above, and subsequently identify $(\bar{i}, \bar{j}) \in \mathcal{E}$ (see Remark 6). Let $X_{\mathcal{M}}^N$ denote N i.i.d. repetitions of $X_{\mathcal{M}} = (X_1, \ldots, X_m)$. Specifically, the agent must produce an estimate $\hat{e}_N = \hat{e}_N(X_{\mathcal{M}}^N) \in \mathcal{E}$ of $(\bar{i}, \bar{j}) \in \mathcal{E}$ at the conclusion of N steps so as to minimize the error probability $P(\hat{e}_N \neq (\bar{i}, \bar{j}))$. The following notation is used. Write

$$\mathrm{I}(i \wedge j) = \mathrm{I}(X_i \wedge X_j), \quad (i, j) \in \mathcal{E}$$

for simplicity, and let

$$\mathrm{I}_{\mathsf{EMI}}^{(n)}(i \wedge j) \triangleq \mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{X}_i \wedge \mathbf{X}_j)$$

be the estimate of $\mathrm{I}(i \wedge j)$. At the end of $N = |\mathcal{E}|\, n$ steps, set

$$\hat{e}(X_{\mathcal{M}}^N) = \arg\min_{(i,j)\in\mathcal{E}} \mathrm{I}_{\mathsf{EMI}}^{(n)}(i, j) = (i^*, j^*), \text{ say} \quad (38)$$

with ties being resolved arbitrarily. Correspondingly, the estimate of shared information is

$$\mathrm{SI}_{\mathsf{EMI}}^{(N)}(X_{\mathcal{M}}^N) \triangleq \mathrm{I}_{\mathsf{EMI}}^{(n)}(i^* \wedge j^*). \quad (39)$$

Denote

$$\Delta_{ij} = \mathrm{I}(X_i \wedge X_j) - \mathrm{I}(X_{\bar{i}}, X_{\bar{j}}), \qquad (i, j) \in \mathcal{E}$$

and

$$\Delta_1 = \min_{\substack{(i,j)\in\mathcal{E}\\(i,j)\neq(\bar{i},\bar{j})}} \mathrm{I}(X_i \wedge X_j) - \mathrm{I}(X_{\bar{i}} \wedge X_{\bar{j}}),$$

where the latter is the difference between the second-lowest and lowest mutual information across edges in $\mathcal{E}$. Note that $\Delta_1 > 0$ by the assumed uniqueness of the minimizing edge $(\bar{i}, \bar{j})$.

The shared information estimate $\mathrm{SI}_{\mathsf{EMI}}^{(N)}(X_{\mathcal{M}}^N)$ converges almost surely and in the mean. This is shown in Theorem 11 below. To that end, we first provide an upper bound for the probability of arm misidentification with uniform sampling.

Proposition 10. *For uniform sampling, the probability of error in identifying the optimal pair of arms is*

$$P\left(\hat{e}_N(X_{\mathcal{M}}^N) \neq (\bar{i}, \bar{j})\right) \leq 2\,|\mathcal{E}| \exp\left(\frac{-(N/\,|\mathcal{E}|)\Delta_1^2}{648\log^2(N/\,|\mathcal{E}|)}\right)$$

for all

$$N > |\mathcal{E}| \max\left\{\frac{|\mathcal{X}|^2 - 1}{2^{\Delta_1/3} - 1}, \frac{|\mathcal{X}| - 1}{2^{\Delta_1/6} - 1}\right\}. \quad (40)$$

Proof. With $N = |\mathcal{E}|\, n$, let $x_{\mathcal{M}}^N$ represent a realization of $X_{\mathcal{M}}^N$. For each $(i, j) \in \mathcal{E}$, the agent computes the empirical mutual information estimate $\mathrm{I}_{\mathsf{EMI}}^{(n)}(\mathbf{x}_i \wedge \mathbf{x}_j)$ of $\mathrm{I}(X_i \wedge X_j)$. Note that the sampling of arm pairs occurs over nonoverlapping time instants. By Lemma 7 and (34)

$$\left|\mathrm{Bias}(\mathrm{I}_{\mathsf{EMI}}^{(n)}(i \wedge j))\right| \leq \frac{\Delta_1}{3} \leq \frac{\Delta_{ij}}{3} < \frac{\Delta_{ij}}{2} \quad \text{for } (i, j) \neq (\bar{i}, \bar{j}),$$

for all N as in (40). Then, we have

$$\begin{aligned}
&P\left(\hat{e}_N(X_{\mathcal{M}}^N) \neq (\bar{i},\bar{j})\right)\\
&= P\left(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\bar{i}\wedge\bar{j}) \geq \mathrm{I}_{\mathsf{EMI}}^{(n)}(i\wedge j) \text{ for some } (i,j)\neq(\bar{i},\bar{j})\right)\\
&\leq \sum_{(i,j)\neq(\bar{i},\bar{j})} P\left(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\bar{i}\wedge\bar{j}) \geq \mathrm{I}_{\mathsf{EMI}}^{(n)}(i\wedge j)\right)\\
&\leq \sum_{(i,j)\neq(\bar{i},\bar{j})} 2\exp\left(\frac{-n\Delta_{ij}^2}{648\log^2 n}\right), \qquad \text{by Lemma 9}\\
&\leq 2\left|\mathcal{E}\right|\exp\left(\frac{-(N/\left|\mathcal{E}\right|)\Delta_1^2}{648\log^2(N/\left|\mathcal{E}\right|)}\right). \qquad \square
\end{aligned}$$

Theorem 11. *For uniform sampling (as in Definition 6), the estimate* $\mathrm{SI}_{\mathsf{EMI}}^{(N)}(X_{\mathcal{M}}^N)$ *converges as* $N\to\infty$ *to* $\mathrm{SI}(X_{\mathcal{M}})$ *almost surely and in the mean.*

Proof. Let $0<\epsilon<\Delta_1$. By (34), we can choose n large enough such that $\mathrm{Bias}(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\bar{i}\wedge\bar{j}))<\epsilon/2$ (see Remark 6). For all such n,

$$\begin{aligned}
&P\left(\left|\mathrm{SI}_{\mathsf{EMI}}^{(N)}(X_{\mathcal{M}}) - \mathrm{SI}(X_{\mathcal{M}})\right|>\epsilon\right)\\
&\leq P\left(\left|\mathrm{SI}_{\mathsf{EMI}}^{(N)}(X_{\mathcal{M}}^N) - \mathrm{I}(\bar{i}\wedge\bar{j})\right|>\epsilon, \hat{e}_N(X_{\mathcal{M}}^N)=(\bar{i},\bar{j})\right)\\
&\qquad + P\left(\hat{e}_N(X_{\mathcal{M}}^N)\neq(\bar{i}\wedge\bar{j})\right)\\
&\leq P\left(\left|\mathrm{I}_{\mathsf{EMI}}^{(n)}(\bar{i}\wedge\bar{j}) - \mathrm{I}(\bar{i}\wedge\bar{j})\right|>\epsilon\right) + P\left(\hat{e}_N(X_{\mathcal{M}}^N)\neq(\bar{i},\bar{j})\right)\\
&= P\Bigg(\Bigg|\mathrm{I}_{\mathsf{EMI}}^{(n)}(\bar{i}\wedge\bar{j}) - \mathbb{E}\left[\mathrm{I}_{\mathsf{EMI}}^{(n)}(\bar{i}\wedge\bar{j})\right]\\
&\qquad + \mathbb{E}\left[\mathrm{I}_{\mathsf{EMI}}^{(n)}(\bar{i}\wedge\bar{j})\right] - \mathrm{I}(\bar{i}\wedge\bar{j})\Bigg|>\epsilon\Bigg)\\
&\qquad + P\left(\hat{e}_N(X_{\mathcal{M}}^N)\neq(\bar{i},\bar{j})\right)\\
&\leq P\Bigg(\left|\mathrm{I}_{\mathsf{EMI}}^{(n)}(\bar{i}\wedge\bar{j}) - \mathbb{E}\left[\mathrm{I}_{\mathsf{EMI}}^{(n)}(\bar{i}\wedge\bar{j})\right]\right|\\
&\qquad + \left|\mathrm{Bias}\left(\mathrm{I}_{\mathsf{EMI}}^{(n)}(\bar{i}\wedge\bar{j})\right)\right|>\epsilon\Bigg)\\
&\qquad + P\left(\hat{e}_N(X_{\mathcal{M}}^N)\neq(\bar{i},\bar{j})\right)\\
&\leq \exp\left(\frac{-2(N/\left|\mathcal{E}\right|)(\epsilon/2)^2}{36\log^2(N/\left|\mathcal{E}\right|)}\right)\\
&\qquad + 2\left|\mathcal{E}\right|\exp\left(\frac{-(N/\left|\mathcal{E}\right|)\Delta_1^2}{648\log^2(N/\left|\mathcal{E}\right|)}\right) \qquad (41)
\end{aligned}$$

by Lemma 8 and Proposition 10. Almost sure convergence follows from (41) and the Borel-Cantelli Lemma (cf. e.g., [29, Lemma 7.3]); and furthermore, since $\mathrm{SI}_{\mathsf{EMI}}^{(N)}(X_{\mathcal{M}}^N)\leq\log|\mathcal{X}|$ for all N, almost sure convergence implies convergence in the mean by the Dominated Convergence Theorem (cf. e.g., [29, Theorem 3.27]). $\square$

The following corollary specifies the sample complexity of $\mathrm{SI}_{\mathsf{EMI}}^{(N)}$ in terms of a minimum requirement on N for which the estimation error is small with high probability.

Corollary 12. *For* $0<\epsilon<1/2$ *and* $\delta<1/e$*, we have*

$$P\left(\left|\mathrm{SI}_{\mathsf{EMI}}^{(N)}(X_{\mathcal{M}}^N) - \mathrm{SI}(X_{\mathcal{M}})\right|>\epsilon\right)\leq\delta$$

for sample complexity $N=N(\epsilon,\delta)$ *that obeys*[2]

$$\begin{aligned}
N \gtrsim \left|\mathcal{E}\right|\Bigg[&\frac{\left|\mathcal{X}\right|}{\epsilon} + \frac{1}{\epsilon^2}\ln\left(\frac{1}{\delta}\right)\log^2\left(\frac{1}{\epsilon^2}\ln\left(\frac{1}{\delta}\right)\right)\\
&+\frac{1}{\Delta_1^2}\ln\left(\frac{\left|\mathcal{E}\right|}{\delta}\right)\log\left(\frac{1}{\Delta_1^2}\ln\left(\frac{\left|\mathcal{E}\right|}{\delta}\right)\right)\\
&+\frac{1}{\Delta_1^2}\ln\left(\frac{\left|\mathcal{E}\right|}{\delta}\right)\log^2\left(\frac{1}{\Delta_1^2}\ln\left(\frac{\left|\mathcal{E}\right|}{\delta}\right)\right)\Bigg]. \qquad (42)
\end{aligned}$$

The proof of the corollary relies on the following technical lemma, which is similar in spirit to [40, Lemma A.1].

Lemma 13. *It holds that*

$$x\geq c\ln^2 x, \quad c\geq 1,\ x\geq\max\left\{1, 4c\ln 2c + 16c\ln^2 c\right\}.$$

Proof. See Appendix C. $\square$

Proof of Corollary 12. From (41),

$$P\left(\left|\mathrm{SI}_{\mathsf{EMI}}^{(N)}(X_{\mathcal{M}}^N) - \mathrm{SI}(X_{\mathcal{M}})\right|>\epsilon\right)\leq\delta,$$

for $n=N/\left|\mathcal{E}\right|$ satisfying

$$n\geq\frac{\left|\mathcal{X}\right|}{\epsilon},\ \frac{n}{\log^2 n}\geq\frac{1}{\epsilon^2}\ln\left(\frac{1}{\delta}\right),\ \frac{n}{\log^2 n}\geq\frac{1}{\Delta_1^2}\ln\left(\frac{\left|\mathcal{E}\right|}{\delta}\right) \qquad (43)$$

up to numerical constant factors. Each of the inequalities in (43) yields one or more lower bounds for $n=N/\left|\mathcal{E}\right|$; the first does so directly, and the latter two upon writing them as $n\geq c\log^2 n$ (where c does not depend on n) and using Lemma 13. The conditions on ϵ and δ in Corollary 12, allow us to drop one of the bounds since it is always weaker than another. Combining all the lower bounds obtained from (43) and using $N=\left|\mathcal{E}\right|n$ finally results in (42). $\square$

VI. Closing Remarks

While the Hammersley-Clifford Theorem [31, Theorem 3.9] can be used to show the equivalence in Theorem 3 between the MCT definition and the global Markov property, when the joint pmf of $X_{\mathcal{M}}$ is strictly positive, we show for an MCT that it holds even for pmfs that are *not* strictly positive. The tree structure plays a material role in our proof. In particular, agglomeration of connected subsets of an MCT form an MCT as in Definition 3 and Lemma 2. The MCT property only involves verifying the Markov condition (10) or (11) for each edge in the tree, and therefore is easier to check than the global Markov property (16). Joint pmfs that are not strictly positive arise in applications such as function computation when a subset of rvs are determined by another subset.

Theorem 6 shows that for an MCT, a simple 2-partition achieves the minimum in Definition 1. While the result in Theorem 6 was known [21], our proof uses new techniques and further implies that for *any* partition π with disconnected atoms, there is a partition with connected atoms that has $\mathcal{I}$-value (see (2)) less than or equal to that of π. This structural property is stronger than that needed for proving Theorem 6.

[2]The approximate form of (42) considers only the significant terms depending on $|\mathcal{X}|$, $|\mathcal{E}|$, Δ_1, ϵ and δ.

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2024.3353769

Our proof technique for Theorem 6 can serve as a stepping stone for analyzing SI for more complicated graphical models in which the underlying graph is not a tree; see [4]. In particular, the tree structure was used in the proof of Theorem 6 only in Step 1 and (25) in Step 2.

In Section V, we have presented an algorithm for best-arm identification with biased (and asymptotically unbiased) estimates. A uniform sampling strategy and the empirical mutual information estimator were chosen for simplicity. Using bias-corrected estimators like the Miller-Madow or jack-knifed estimator for mutual information [38] would improve the bias performance of the algorithm. However, it hurts the constant in the bounded differences inequality that appears in Lemma 8. Polynomial approximation-based estimators [27] could also improve sample complexity. Moreover, a successive rejects algorithm [47], [6] could yield a better sample complexity than uniform sampling for a fixed estimator, as hinted by [47], [6] in different settings. The precise tradeoff afforded by the choice of a better estimator remains to be understood, as does the sample complexity of more refined algorithms for best arm identification in our setting. Both demand a converse result that needs to take into account estimator bias; this remains under study in our current work. A converse would also settle the question of optimality of the $\exp(-O(N/\log^2 N))$ decay in the probability of error in Proposition 10.

Appendix A
Proof of Lemma 1

We have

$$\begin{aligned}
&\mathrm{I}(X_{\mathcal{B}(i\leftarrow j)} \wedge X_{\mathcal{B}(j\leftarrow i)}) \\
&= \mathrm{I}(X_i \wedge X_j) + \mathrm{I}(X_i \wedge X_{\mathcal{B}(j\leftarrow i)\setminus\{j\}} \,|\, X_j) \\
&\qquad + \mathrm{I}(X_{\mathcal{B}(i\leftarrow j)\setminus\{i\}} \wedge X_j \,|\, X_i) \\
&\qquad\quad + \mathrm{I}(X_{\mathcal{B}(i\leftarrow j)\setminus\{i\}} \wedge X_{\mathcal{B}(j\leftarrow i)\setminus\{j\}} \,|\, X_i, X_j) \\
&= \mathrm{I}(X_i \wedge X_j) + \mathrm{I}(X_{\mathcal{B}(i\leftarrow j)\setminus\{i\}} \wedge X_{\mathcal{B}(j\leftarrow i)\setminus\{j\}} \,|\, X_i, X_j) \\
&= \mathrm{I}(X_i \wedge X_j) + \mathrm{H}(X_{\mathcal{B}(i\leftarrow j)\setminus\{i\}} \,|\, X_i) \\
&\qquad - \mathrm{H}(X_{\mathcal{B}(i\leftarrow j)\setminus\{i\}} \,|\, X_i, X_{\mathcal{B}(j\leftarrow i)}) \quad (44)
\end{aligned}$$

where the previous two inequalities are by (11).

The claim of Lemma 1 would follow from (44) upon showing that

$$\mathrm{H}(X_{\mathcal{B}(i\leftarrow j)\setminus\{i\}} \,|\, X_i, X_{\mathcal{B}(j\leftarrow i)}) = \mathrm{H}(X_{\mathcal{B}(i\leftarrow j)\setminus\{i\}} \,|\, X_i). \quad (45)$$

Without loss of generality, set j to be the root of the tree; this defines a *directed* tree whose leaves are from among the vertices (in $\mathcal{M}$) with no descendants. Denote the parent of i' in the (directed) tree by $p(i')$. Note that $p(i) = j$ in (45). We shall use induction on the *height* of i', i.e., the maximum distance of i' from a leaf of the directed tree, to show that

$$\begin{aligned}
&\mathrm{H}(X_{\mathcal{B}(i'\leftarrow p(i'))\setminus\{i'\}} \,|\, X_{i'}, X_{\mathcal{B}(p(i')\leftarrow i')}) \\
&\qquad = \mathrm{H}(X_{\mathcal{B}(i'\leftarrow p(i'))\setminus\{i'\}} \,|\, X_{i'}), \quad (46)
\end{aligned}$$

which proves (45) upon setting $i' = i$ and $p(i') = p(i) = j$.

First, assume that i' is a leaf. Then $\mathcal{B}(i' \leftarrow p(i'))\setminus\{i'\} = \varnothing$ and (46) holds trivially.

Next, assume the induction hypothesis that (46) is true for all vertices at height $< h$, and consider a vertex i' at height h.

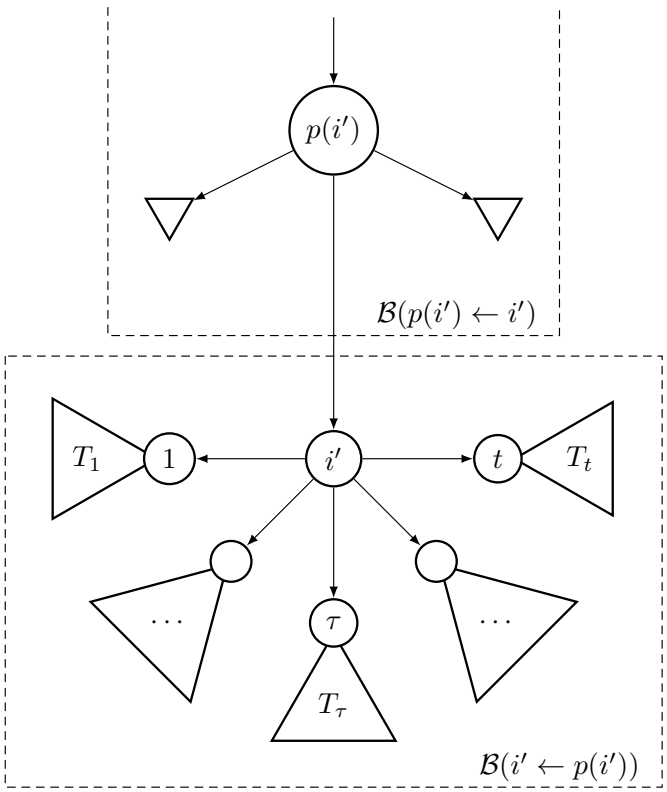

Fig. 3. Schematic for proof of (46).

Let i' have children $1, \ldots, t$; each of these vertices is the root of subtree $T_\tau = \mathcal{B}(\tau \leftarrow i')$, $1 \le \tau \le t$. See Figure 3. Further, each vertex τ, $1 \le \tau \le t$, has height $< h$. Then in (46),

$$\begin{aligned}
&\mathrm{H}(X_{\mathcal{B}(i'\leftarrow p(i'))\setminus\{i'\}} \,|\, X_{i'}, X_{\mathcal{B}(p(i')\leftarrow i')}) \\
&= \mathrm{H}\left((X_{T_\tau\setminus\{\tau\}}, X_\tau)_{1\le\tau\le t} \,|\, X_{i'}, X_{\mathcal{B}(p(i')\leftarrow i')}\right) \\
&= \sum_{\tau=1}^{t} \Big[\mathrm{H}\left(X_\tau \,|\, (X_{T_\sigma})_{1\le\sigma\le\tau-1}, X_{i'}, X_{\mathcal{B}(p(i')\leftarrow i')}\right) \\
&\quad + \mathrm{H}\left(X_{T_\tau\setminus\{\tau\}} \,|\, X_\tau, (X_{T_\sigma})_{1\le\sigma\le\tau-1}, X_{i'}, X_{\mathcal{B}(p(i')\leftarrow i')}\right)\Big]. \quad (47)
\end{aligned}$$

In (47), for each τ, $1 \le \tau \le t$, the first term within $[\cdot]$ is

$$\begin{aligned}
&\mathrm{H}\left(X_\tau \,|\, (X_{T_\sigma})_{1\le\sigma\le\tau-1}, X_{i'}, X_{\mathcal{B}(p(i')\leftarrow i')}\right) \\
&\qquad = \mathrm{H}(X_\tau \,|\, X_{i'}) = \mathrm{H}\left(X_\tau \,|\, (X_{T_\sigma})_{1\le\sigma\le\tau-1}, X_{i'}\right) \quad (48)
\end{aligned}$$

by (11) since

$$\left(\bigcup_{\sigma=1}^{\tau-1} T_\sigma, \mathcal{B}(p(i') \leftarrow i')\right) \subseteq \mathcal{B}(i' \leftarrow \tau)$$

(see Figure 3). In the second term in $[\cdot]$, we apply the induction hypothesis to vertex τ which is at height $h-1$. Note that $p(\tau) = i'$. Since

$$\begin{aligned}
&X_{T_\tau\setminus\{\tau\}} = X_{\mathcal{B}(\tau\leftarrow p(\tau))\setminus\{\tau\}} \\
&\text{and } \left((X_{T_\sigma})_{1\le\sigma\le\tau-1}, X_{i'}, X_{\mathcal{B}(p(i')\leftarrow i')}\right) \subseteq \mathcal{B}(p(\tau) \leftarrow \tau),
\end{aligned}$$

by the induction hypothesis at vertex τ, we get

$$\begin{aligned}&\mathrm{H}\left(X_{T_\tau\setminus\{\tau\}}\,|\,X_\tau,(X_{T_\sigma})_{1\le\sigma\le\tau-1},X_{i'},X_{\mathcal{B}(p(i')\leftarrow i')}\right)\\&=\mathrm{H}\left(X_{T_\tau\setminus\{\tau\}}\,|\,X_\tau\right)\\&=\mathrm{H}\left(X_{T_\tau\setminus\{\tau\}}\,|\,X_\tau,(X_{T_\sigma})_{1\le\sigma\le\tau-1},X_{i'}\right)\end{aligned}\tag{49}$$

with the last equality being due to (11). Substituting (48), (49) in (47), we obtain

$$\begin{aligned}&\mathrm{H}(X_{\mathcal{B}(i'\leftarrow p(i'))\setminus\{i'\}}\,|\,X_{i'},X_{\mathcal{B}(p(i')\leftarrow i')})\\&=\sum_{\tau=1}^{t}\Big[\mathrm{H}\left(X_\tau\,|\,(X_{T_\sigma})_{1\le\sigma\le\tau-1},X_{i'}\right)\\&\qquad+\mathrm{H}\left(X_{T_\tau\setminus\{\tau\}}\,|\,X_\tau,(X_{T_\sigma})_{1\le\sigma\le\tau-1},X_{i'}\right)\Big]\\&=\sum_{\tau=1}^{t}\mathrm{H}\left(X_{T_\tau}\,|\,(X_{T_\sigma})_{1\le\sigma\le\tau-1},X_{i'}\right)\\&=\mathrm{H}(X_{\mathcal{B}(i'\leftarrow p(i'))\setminus\{i'\}}\,|\,X_{i'})\end{aligned}$$

(see Figure 3) which is (46). □

Appendix B
Proof of Lemma 4 and Theorem 3

Proof of Lemma 4. Considering first (17), suppose that vertex $i\in\mathcal{M}$ has k neighbor, with $\mathcal{N}(i)=\{i_1,\ldots,i_k\}$, $1\le k\le m-1$. Then

$$\mathcal{M}\setminus(\{i\}\cup\mathcal{N}(i))=\bigcup_{l=1}^{k}\mathcal{B}(i_l\leftarrow i)\setminus\{i_l\}\,.$$

The claim of the lemma is

$$X_i\multimap(X_{i_u})_{1\le u\le k}\multimap\left(X_{\mathcal{B}(i_l\leftarrow i)\setminus\{i_l\}}\right)_{1\le l\le k}.\tag{50}$$

We have

$$\begin{aligned}&\mathrm{I}\left(X_i\wedge\left(X_{\mathcal{B}(i_l\leftarrow i)\setminus\{i_l\}}\right)_{1\le l\le k}\,|\,(X_{i_u})_{1\le u\le k}\right)\\&=\sum_{l=1}^{k}\mathrm{I}\left(X_i\wedge X_{\mathcal{B}(i_l\leftarrow i)\setminus\{i_l\}}\,|\,\left(X_{\mathcal{B}(i_j\leftarrow i)\setminus\{i_j\}}\right)_{1\le j\le l-1},(X_{i_u})_{1\le u\le k}\right)\\&\le\sum_{l=1}^{k}\mathrm{I}\Big(\left[X_i,\left(X_{\mathcal{B}(i_j\leftarrow i)\setminus\{i_j\}}\right)_{1\le j\le l-1},(X_{i_u})_{1\le u\ne l\le k}\right]\\&\qquad\wedge X_{\mathcal{B}(i_l\leftarrow i)\setminus\{i_l\}}\,|\,X_{i_l}\Big).\end{aligned}\tag{51}$$

For each l, $1\le l\le k$, the rvs within $[\cdot]$ above have indices that lie in $\mathcal{B}(i\leftarrow i_l)\setminus\{i_l\}$. Hence, by Lemma 1 (specifically (13)), each term in the sum in (51) equals zero. This proves (50). See Figure 4.

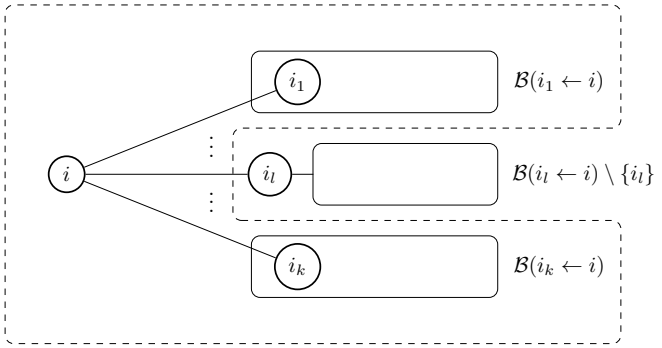

Fig. 4. Schematic for the proof of Lemma 4.

Turning to (18), we have

$$\begin{aligned}&\mathrm{I}\left(X_A\wedge X_{\mathcal{M}\setminus(A\cup\mathcal{N}(A))}\,|\,\mathcal{N}(A)\right)\\&=\mathrm{I}\left((X_i,i\in A)\wedge X_{\mathcal{M}\setminus\bigcup_{u\in A}(\{u\}\cup\mathcal{N}(u))}\,\big|\,X_{\bigcup_{v\in A}\mathcal{N}(v)}\right)\\&\le\sum_{i\in A}\mathrm{I}\Big(X_i\wedge X_{\bigcup_{j\in A\setminus\{i\}}(\{j\}\cup\mathcal{N}(j))},\\&\qquad\qquad X_{\mathcal{M}\setminus\bigcup_{u\in A}(\{u\}\cup\mathcal{N}(u))}\,\big|\,X_{\mathcal{N}(i)}\Big)\\&=\sum_{i\in A}\mathrm{I}\Big(X_i\wedge X_{\left(\bigcup_{j\in A\setminus\{i\}}(\{j\}\cup\mathcal{N}(j))\right)\setminus\mathcal{N}(i)},\\&\qquad\qquad X_{\mathcal{M}\setminus\bigcup_{u\in A}(\{u\}\cup\mathcal{N}(u))}\,\big|\,X_{\mathcal{N}(i)}\Big)\\&=0\end{aligned}$$

by (17) since for each $i\in A$,

$$\begin{aligned}&\left(\left(\bigcup_{j\in A\setminus\{i\}}(\{j\}\cup\mathcal{N}(j))\right)\setminus\mathcal{N}(i)\right)\\&\qquad\cup\left(\mathcal{M}\setminus\bigcup_{u\in A}(\{u\}\cup\mathcal{N}(u))\right)\subseteq\mathcal{M}\setminus(\{i\}\cup\mathcal{N}(i)).\ \square\end{aligned}$$

Proof of Theorem 3. The converse claim is immediately true upon choosing: for every $(i,j)\in\mathcal{E}$, $A=\mathcal{B}(i\leftarrow j)\setminus\{i\}$, $S=i$, $B=\{j\}$.

Turning to the first claim, let

$$A=\bigsqcup_{\alpha=1}^{a}A_\alpha,\quad B=\bigsqcup_{\beta=1}^{b}B_\beta,\quad S=\bigsqcup_{\sigma=1}^{s}S_\sigma$$

be representations in terms of maximally connected subsets of A, B and S, respectively. With $N=\mathcal{M}\setminus(A\cup B\cup S)$, let $N=\sqcup_{\nu=1}^{n}N_\nu$ be a decomposition into maximally connected subsets of N. Denote

$$\begin{aligned}&\mathcal{A}=\{A_\alpha,1\le\alpha\le a\}\,,\quad\mathcal{B}=\{B_\beta,1\le\beta\le b\}\,,\\&\mathcal{S}=\{S_\sigma,1\le\sigma\le s\}\,,\quad\mathcal{N}=\{N_\nu,1\le\nu\le n\}\,.\end{aligned}$$

Referring to Definition 3 and recalling Lemma 2, the tree $\mathcal{G}'=(\mathcal{M}',\mathcal{E}')$ with vertex set $\mathcal{M}'=\mathcal{A}\cup\mathcal{B}\cup\mathcal{S}\cup\mathcal{N}$ and edge set in the manner of Definition 3 constitutes an agglomerated MCT.

Next, we observe that since each $N_\nu\in\mathcal{N}$, $1\le\nu\le n$, is maximally connected in N, the neighbors of N_ν in $\mathcal{G}'$ cannot

be in $\mathcal{N}$. Therefore, neighbors of a given N_ν in $\mathcal{G}'$ that are not in $\mathcal{S}$ must be in $\mathcal{A}$ or $\mathcal{B}$. However, N_ν cannot have a non$\mathcal{S}$ neighbor in $\mathcal{A}$ and also one in $\mathcal{B}$, for then A and B would not be separated by S in $\mathcal{G}$. Accordingly, for *each* N_ν in $\mathcal{N}$, if its non$\mathcal{S}$ neighbors in $\mathcal{G}'$ are only in $\mathcal{A}$, add N_ν to $\mathcal{A}$; let N' be the union of all such N_νs. Consider $A' = A \cup N'$ and write $A' = \sqcup_{\alpha=1}^{a'} A'_\alpha$ where the A'_αs are maximally connected subsets of A'. Let $\mathcal{A}' = \{A'_\alpha, 1 \le \alpha \le a'\}$.

Now note that $\mathcal{A}'$ and $\mathcal{B}$ are separated in $\mathcal{G}'$ by $\mathcal{S}$. Thus, to establish (16), it suffices to show the (stronger) assertion

$$X_{\mathcal{A}'} -\!\circ\!- X_{\mathcal{S}} -\!\circ\!- X_{\mathcal{B}}. \tag{52}$$

By the description of $\mathcal{A}'$, each of its components (maximal subsets of A') has its neighborhood in $\mathcal{G}'$ that is contained *fully* in $\mathcal{S}$. Let $\tilde{\mathcal{S}} \subseteq \mathcal{S}$ denote the union of all such neighborhoods. Then, by Lemma 4 ((18)) applied to the agglomerated tree $\mathcal{G}'$, since there is no edge in $\mathcal{G}'$ that connects any two elements of $\mathcal{A}'$,

$$X_{\mathcal{A}'} -\!\circ\!- X_{\tilde{\mathcal{S}}} -\!\circ\!- X_{\mathcal{M}'\setminus(\mathcal{A}'\cup\tilde{\mathcal{S}})}$$

so that

$$\begin{aligned} 0 &= \mathrm{I}(X_{\mathcal{A}'} \wedge X_{\mathcal{M}'\setminus(\mathcal{A}'\cup\tilde{\mathcal{S}})} \,|\, X_{\tilde{\mathcal{S}}}) \\ &= \mathrm{I}(X_{\mathcal{A}'} \wedge X_{\mathcal{M}'\setminus(\mathcal{A}'\cup\tilde{\mathcal{S}})}, X_{\mathcal{S}\setminus\tilde{\mathcal{S}}} \,|\, X_{\tilde{\mathcal{S}}}) \\ &\ge \mathrm{I}(X_{\mathcal{A}'} \wedge X_{\mathcal{M}'\setminus(\mathcal{A}'\cup\mathcal{S})} \,|\, X_{\mathcal{S}}) \end{aligned} \tag{53}$$

since $\tilde{\mathcal{S}} \subseteq \mathcal{S}$. Finally, (53) implies (52) as $\mathcal{B} \subseteq \mathcal{M}' \setminus (\mathcal{A}' \cup \mathcal{S})$. □

Appendix C
Proof of Lemma 13

Proof. For $1 \le c \le 1.2$, $x \ge c\ln^2 x$ holds unconditionally; so assume that $c \ge 1.2$. Consider the function $f(x) = x - c\ln^2 x$. Then, using [40, Lemma A.1], $x \ge 4c\ln 2c$ implies $x \ge 2c\ln x$ which, in turn, implies $f'(x) \ge 0$. Therefore, for $x \ge 4c\ln c$, $f(x)$ is increasing in x. It is easy to check numerically that $f(16c\ln^2 c)$ is positive for $c \ge 1.2$. Thus, for all $x \ge \max\{4c\ln 2c, 16c\ln^2 c\}$, $f(x) \ge 0$ and so $x \ge c\ln^2 x$. □

References

[1] S. A. Abdallah and M. D. Plumbley, "A measure of statistical complexity based on predictive information," *ArXiv*, vol. abs/1012.1890, 2010.

[2] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Structures & Algorithms*, vol. 19, no. 3-4, pp. 163–193, 2001.

[3] S. Bhattacharya and P. Narayan, "Shared information for a Markov chain on a tree," in *2022 IEEE International Symposium on Information Theory*, 2022, pp. 3049–3054.

[4] ——, "Shared information for the cliqueylon graph," in *2023 IEEE International Symposium on Information Theory (ISIT)*. IEEE, Jun. 2023.

[5] V. P. Boda and P. Narayan, "Universal sampling rate distortion," *IEEE Transactions on Information Theory*, vol. 64, no. 12, pp. 7742–7758, Dec. 2018.

[6] V. P. Boda and L. A. Prashanth, "Correlated bandits or: How to minimize mean-squared error online," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97. PMLR, 6 2019, pp. 686–694.

[7] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[8] C. Chan, "On tightness of mutual dependence upperbound for secret-key capacity of multiple terminals," *ArXiv*, vol. abs/0805.3200, 2008.

[9] ——, "The hidden flow of information," *2011 IEEE International Symposium on Information Theory Proceedings*, pp. 978–982, 2011.

[10] ——, "Linear perfect secret key agreement," in *2011 IEEE Information Theory Workshop*, 2011, pp. 723–726.

[11] C. Chan, A. Al-Bashabsheh, J. B. 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, 2015.

[12] C. Chan, A. Al-Bashabsheh, and Q. Zhou, "Agglomerative info-clustering: Maximizing normalized total correlation," *IEEE Transactions on Information Theory*, vol. 67, no. 3, pp. 2001–2011, 2021.

[13] C. Chan, A. Al-Bashabsheh, Q. Zhou, N. Ding, T. Liu, and A. Sprintson, "Successive omniscience," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3270–3289, 2016.

[14] C. Chan, A. Al-Bashabsheh, Q. Zhou, T. Kaced, and T. Liu, "Info-clustering: A mathematical theory for data clustering," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, pp. 64–91, 2016.

[15] C. Chan and L. Zheng, "Mutual dependence for secret key agreement," in *2010 44th Annual Conference on Information Sciences and Systems (CISS)*, 2010, pp. 1–6.

[16] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory*, pp. 462–467, 1968.

[17] C. Chow and T. Wagner, "Consistency of an estimate of tree-dependent probability distributions (corresp.)," *IEEE Transactions on Information Theory*, vol. 19, pp. 369–371, 1973.

[18] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms, Third Edition*, 3rd ed. The MIT Press, 2009.

[19] T. A. Courtade and T. R. Halford, "Coded cooperative data exchange for a secret key," *IEEE Transactions on Information Theory*, vol. 62, pp. 3785–3795, 2016.

[20] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.

[21] I. Csiszár and P. Narayan, "Secrecy capacities for multiple terminals," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3047–3061, Dec. 2004.

[22] W. Feng, N. K. Vishnoi, and Y. Yin, "Dynamic sampling from graphical models," *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, 2019.

[23] P. Gács and J. Körner, "Common information is far less than mutual information," *Problems of Control and Information Theory*, vol. 2, 01 1973.

[24] H.-O. Georgii, *Gibbs Measures and Phase Transitions*. De Gruyter, 2011.

[25] V. D. Goppa, "Nonprobabilistic mutual information without memory," *Problems of Control and Information Theory*, vol. 4, pp. 97–102, 1975.

[26] T. S. Han, "Nonnegative entropy measures of multivariate symmetric correlations," *Information and Control*, vol. 36, no. 2, pp. 133–156, 1978.

[27] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2835–2885, 2015.

[28] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[29] L. Koralov, Y. Sinai, and Â. Sinaj, *Theory of Probability and Random Processes*, ser. Universitext (Berlin. Print). Springer, 2007.

[30] T. Lattimore and C. Szepesvari, "Bandit algorithms," 2017.

[31] S. L. Lauritzen, *Graphical Models*. Oxford University Press, 1996.

[32] W. Liu, G. Xu, and B. Chen, "The common information of n dependent random variables," *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 836–843, 2010.

[33] D. J. C. MacKay, *Information Theory, Inference & Learning Algorithms*. USA: Cambridge University Press, 2002.

[34] P. Narayan, "Omniscience and secrecy," 2012, plenary Talk, *IEEE International Symposium on Information Theory*, Cambridge, MA.

[35] P. Narayan and H. Tyagi, "Multiterminal secrecy by public discussion," *Foundations and Trends in Communications and Information Theory*, vol. 13, no. 2-3, pp. 129–275, 2016.

[36] S. Nitinawarat and P. Narayan, "Perfect omniscience, perfect secrecy, and Steiner tree packing," *IEEE Transactions on Information Theory*, vol. 56, pp. 6490–6500, 2010.

[37] S. Nitinawarat, C. Ye, A. Barg, P. Narayan, and A. Reznik, "Secret key generation for a pairwise independent network model," *IEEE*

Transactions on Information Theory, vol. 56, no. 12, p. 6482–6489, 12 2010.

[38] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, p. 1191–1253, 6 2003.

[39] J. Pearl, "Reverend Bayes on inference engines: A distributed hierarchical approach," in *Proceedings of the Second AAAI Conference on Artificial Intelligence*, ser. AAAI'82. AAAI Press, 1982, p. 133–136.

[40] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014.

[41] H. Tyagi, "Common information and secret key capacity," *IEEE Transactions on Information Theory*, vol. 59, pp. 5627–5640, 2013.

[42] H. Tyagi and P. Narayan, "How many queries will resolve common randomness?" *IEEE Transactions on Information Theory*, vol. 59, no. 9, pp. 5363–5378, 2013.

[43] H. Tyagi and S. Watanabe, "Converses for secret key agreement and secure computing," *IEEE Transactions on Information Theory*, vol. 61, pp. 4809–4827, 2015.

[44] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[45] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of Research and Development*, vol. 4, no. 1, pp. 66–82, 1960.

[46] A. Wyner, "The common information of two dependent random variables," *IEEE Transactions on Information Theory*, vol. 21, no. 2, pp. 163–179, 1975.

[47] J. yves Audibert, S. Bubeck, and R. Munos, "Best arm identification in multi-armed bandits," in *Proceedings of the Twenty-Third Annual Conference on Learning Theory*, 2010, pp. 41–53.

Sagnik Bhattacharya received the Bachelor of Technology in Electrical Engineering from the Indian Institute of Technology Kanpur, India, in 2019. He is currently a PhD candidate in the Department of Electrical and Computer Engineering at the University of Maryland, College Park. His research interests are in information theory, statistical learning, and their practical applications.

Prakash Narayan received the Bachelor of Technology degree in Electrical Engineering from the Indian Institute of Technology, Madras in 1976. He received the M.S. degree in Systems Science and Mathematics in 1978 and the D.Sc. degree in Electrical Engineering in 1981, both from Washington University, St. Louis, MO.

He is Professor of Electrical and Computer Engineering at the University of Maryland, College Park, with a joint appointment at the Institute for Systems Research. His research interests are in network information theory, coding theory, communication theory, communication networks, statistical learning, and cryptography.