

## MIT Open Access Articles

*An Efficient Message-Passing Algorithm for  
Optimizing Decentralized Detection Networks*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

**Citation:** Kreidl, O.P., and A.S. Willsky. "An Efficient Message-Passing Algorithm for Optimizing Decentralized Detection Networks." IEEE Transactions on Automatic Control 55.3 (2010): 563–578. © Copyright 2010 IEEE

**As Published:** <http://dx.doi.org/10.1109/tac.2009.2039547>

**Publisher:** Institute of Electrical and Electronics Engineers (IEEE)

**Persistent URL:** <http://hdl.handle.net/1721.1/73604>

**Version:** Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

**Terms of Use:** Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.



# An Efficient Message-Passing Algorithm for Optimizing Decentralized Detection Networks

O. Patrick Kreidl and Alan S. Willsky, *Fellow, IEEE*

**Abstract**—A promising feature of emerging wireless sensor networks is the opportunity for each spatially-distributed node to measure its local state and transmit only information relevant to effective global decision-making. An equally important design objective, as a result of each node's finite power, is for measurement processing to satisfy explicit constraints on, or perhaps make selective use of, the distributed algorithmic resources. We formulate this multi-objective design problem within the Bayesian decentralized detection paradigm, modeling resource constraints by a directed acyclic network with low-rate, unreliable communication links. Existing team theory establishes when necessary optimality conditions reduce to a convergent iterative algorithm to be executed *offline* (i.e., before measurements are processed). Even so, this offline algorithm has exponential complexity in the number of nodes, and its distributed implementation assumes a fully-connected communication network. We state conditions under which the offline algorithm admits an efficient *message-passing* interpretation, featuring linear complexity and a natural distributed implementation. We experiment with a simulated network of binary detectors, applying the message-passing algorithm to optimize the achievable tradeoff between global detection performance and network-wide online communication. The empirical analysis also exposes a design tradeoff between constraining in-network processing to preserve resources (per online measurement) and then having to consume resources (per offline reorganization) to maintain detection performance.

**Index Terms**—Bayes procedures, cooperative systems, directed graphs, distributed detection, iterative methods, message passing, multi-sensor systems, networks, signal processing, trees (graphs).

## I. INTRODUCTION

THE vision of collaborative self-organizing wireless sensor networks, a confluence of emerging technology in both miniaturized devices and wireless communications, is of growing interest in a variety of scientific fields and engineering applications e.g., geology, biology, surveillance, fault-monitoring [1], [2]. Their promising feature is the opportunity for each spatially-distributed node to receive measurements from its local environment and transmit information that is relevant for effective global decision-making. The finite power available to each node creates incentives for prolonging operational lifetime, motivating measurement processing strategies that satisfy

Manuscript received March 28, 2007; revised September 30, 2008. First published February 02, 2010; current version published March 10, 2010. This work was supported by the U.S. Army Research Office under MURI Grants W911NF-05-1-0207 and W911NF-06-1-0076. Recommended by Associate Editor S. Dey.

The authors are with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: opk@mit.edu; willsky@mit.edu).

Digital Object Identifier 10.1109/TAC.2009.2039547

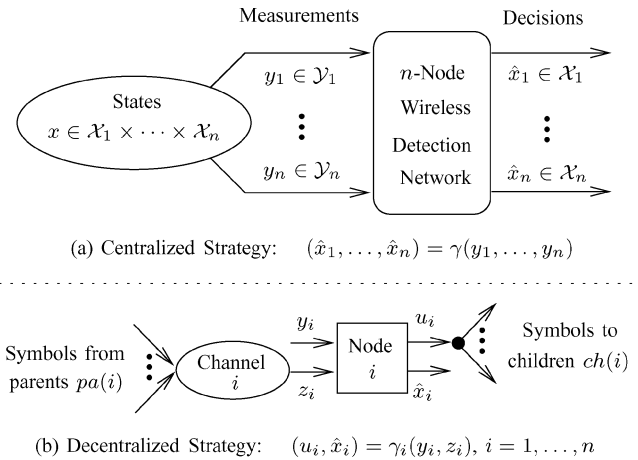


Fig. 1. The  $n$ -sensor detection model (described in Section II) assuming (a) a centralized strategy for processing a measurement vector  $y$  to generate a decision vector  $\hat{x}$  about the (hidden, discrete-valued) state vector  $x$  and (b) a decentralized strategy also subject to explicit network constraints defined on an  $n$ -node directed acyclic graph, each edge representing a unidirectional finite-rate (and perhaps unreliable) communication link between two sensors.

explicit resource constraints (e.g., on communication, computation, memory) in the network layer. One also anticipates intermittent reorganization by the network to stay connected (due, for example, to node dropouts or link failures), implying that resource constraints can change accordingly and creating an incentive for intermittent re-optimization in the application layer. So, unless the *offline* optimization algorithm is itself amenable to efficient distributed implementation, there is little hope for maintaining application-layer decision objectives without also rapidly diminishing the network-layer resources that remain for actual *online* measurement processing.

We explore these design challenges assuming that the decision-making objective is optimal Bayesian detection and the dominant measurement processing constraints arise from the underlying communication medium. Our main model is illustrated in Fig. 1, extending the canonical team-theoretic decentralized detection problem [3]–[5] in ways motivated by the vision of wireless sensor networks. Firstly, the network topology can be defined on any  $n$ -node directed acyclic graph  $\mathcal{G}$ , each edge  $(i, j)$  representing a feasible point-to-point, low-rate communication link from node  $i$  to node  $j$  [6], [7]. Secondly, each node  $i$  can selectively transmit a (perhaps different) finite-alphabet symbol to each of its downstream neighbors, or *children*  $ch(i)$ , in  $\mathcal{G}$  [8], [9] and the multipoint-to-point channel into each node  $i$  from its upstream neighbors, or *parents*  $pa(i)$ , in  $\mathcal{G}$  can be unreliable (e.g., due to uncoded interference, packet loss) [10]. Thirdly, each node  $i$  receives a noisy measurement  $y_i$  related only to its local (discrete-valued) hidden state process  $X_i$ , the

latter correlated with the hidden states local to all other nodes,  $X_{-i} = \{X_j; j \neq i\}$  [7]. Finally, the global decision objective can itself be spatially-distributed e.g., when each node's state-related decision  $\hat{x}_i$  should indicate whether an object is located in the subregion local to node  $i$ .

Team decision problems are known to be NP-hard, in general [5]. Also known is a relaxation of the problem that, under certain model assumptions, analytically reduces to a convergent iterative algorithm to be executed offline [5], [11]. The (generally multiple and non-unique) fixed-points of this iterative algorithm correspond to different so-called *person-by-person optimal* processing strategies, each known to satisfy necessary (but not always sufficient) optimality conditions for the original problem. This offline algorithm strives to couple all nodes' local rules such that there is minimal total performance loss from the online processing constraints. However, the algorithm generally requires all nodes to be initialized with common knowledge of global statistics and iterative per-node computation scales exponentially with the number of nodes.

We identify a class of models for which the convergent offline algorithm admits an efficient *message-passing* interpretation, itself equivalent to a sequence of purely-local computations interleaved with only nearest-neighbor communications [12], [13]. In each offline iteration, every node adjusts its local rule (for subsequent online processing) based on incoming messages from its neighbors and, in turn, sends adjusted outgoing messages to its neighbors. The message schedule consists of repeated forward-backward sweeps through the network: the forward messages received by each node from its parents define, in the context of its local objectives, a "likelihood function" for the symbols it may receive online (e.g., "what does the information from my neighbors mean to me") while the backward messages from its children define, in the context of all other nodes' objectives, a "cost-to-go function" for the symbols it may transmit online (e.g., "what does the information from me mean to my neighbors"). Each node need only be initialized with local statistics and iterative per-node computation is invariant to the number of nodes (but still scales exponentially with the number of neighbors, so the algorithm is best-suited for sparsely-connected networks).

The end result of this (offline) message-passing algorithm can be thought of as a distributed *fusion protocol*, in which the nodes have collectively determined their individual online rules for interpreting information transmitted by their parents and then transmitting information to their children. This protocol takes into account explicitly the limits on available online communication resources and, in turn, produces highly-resourceful measurement processing strategies. For instance, as we will show, the absence of transmission is interpreted as an extra symbol on each link, which remains of value even when active transmissions are of negligible cost or unreliable.

The prospect of a computationally-efficient algorithm to optimize large-scale decentralized detection networks is complementary to other recent work, which focuses on asymptotic analyses [14]–[19] typically under assumptions regarding network regularity or sensor homogeneity. The message-passing algorithm we propose here offers a tractable design alternative for applications in which such assumptions cannot be made, especially if network connectivity is also sparse

and the detection objective is itself spatially-distributed. A tractable linear programming approximation to the offline design problem is developed in [20], but it lacks the natural distributed implementation featured by our message-passing solution. Finally, numerous other authors have advocated iterative message-passing algorithms to solve decision problems in wireless sensor networks (e.g., [21], [22]), but with the key difference that they execute on a per-measurement basis during *online* processing. These solutions perform poorly (or become infeasible) given the severe online resource constraints considered here. It is worth noting, however, that the resources required by our offline message-passing solution (and much of the problem structure under which it is derived) are comparable to that required by these online message-passing solutions.

The next section reviews the theory of decentralized Bayesian detection in the generality implied by Fig. 1, providing examples and notation in preparation for subsequent developments. Section III defines the class of models for which the offline iterative algorithm described in Section II admits its efficient message-passing interpretation. In Section IV, we consider a simulated network of binary detectors and apply the offline message-passing algorithm to quantify the tradeoff between online communication and detection performance. The empirical analysis also exposes a design tradeoff between constraining in-network processing to preserve resources (per online measurement) and then having to consume resources (per offline reorganization) to maintain effective detection performance. Conclusions and future research are discussed in Section V.

## II. DECENTRALIZED DETECTION NETWORKS

This section summarizes the theory of decentralized Bayesian detection [4], [5] in the generality treated by subsequent sections of the paper. It starts with a brief review of well-known concepts in classical detection theory, using notation that may at first seem unnecessarily cumbersome: we do this in preparation for the developments in subsequent sections of the paper. Throughout, for a random variable  $A$  that takes its values in a discrete (or Euclidean) set  $\mathcal{A}$ , we let  $p_A : \mathcal{A} \rightarrow [0, \infty)$  denote its probability mass (or density) function; similarly, let  $c_A : \mathcal{A} \rightarrow [0, \infty)$  denote a cost function associated to the random variable  $A$ . The subscript notation is suppressed when the random variable involved is implied by the functional argument; that is, we let  $p(a) \equiv p_A(a)$  and  $c(a) \equiv c_A(a)$  for every  $a$  in  $\mathcal{A}$ . Also note that  $p(A)$  and  $c(A)$  are themselves well-defined random variables, each taking values in  $[0, \infty)$  according to a distribution derived from  $A$  and the functions  $p_A$  and  $c_A$ , respectively. The expectation of  $A$  is denoted by  $\mathbf{E}[A]$ .

### A. Classical Bayesian Formulation

Let us first focus on Fig. 1(a), supposing the hidden state  $x$  and observable measurement  $y$  take their values in, respectively, a discrete product space  $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$  and Euclidean product space  $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n$ . Subject to design is the function  $\gamma : \mathcal{Y} \rightarrow \mathcal{X}$  by which the network generates its state-related decision  $\hat{x} \in \mathcal{X}$  based on any particular measurement vector  $y \in \mathcal{Y}$ . The classical Bayesian problem formulation [23] (i) describes the (hidden) state process  $X$  and (observed) measurement process  $Y$  by a given joint distribution  $p(x, y)$  and

(ii) assigns a numerical cost  $c(\hat{x}, x)$  to every possible state-decision outcome. The performance of any induced decision process  $\hat{X} = \gamma(Y)$  is then measured by

$$J_d(\gamma) = \mathbf{E} [c(\hat{X}, X)] = \mathbf{E} [\mathbf{E} [c(\gamma(Y), X) | Y]]. \quad (1)$$

Note that the penalty  $J_d(\gamma)$  specializes to (i) the error probability by choosing  $c(\hat{x}, x)$  to be the indicator function on  $\{(\hat{x}, x) \in \mathcal{X} \times \mathcal{X} | \hat{x} \neq x\}$  and (ii) the sum-error probability (i.e., the expected number of nodes in error) by choosing  $c(\hat{x}, x) = \sum_i c(\hat{x}_i, x_i)$  where, for every  $i$ , the local costs  $c(\hat{x}_i, x_i)$  correspond to the indicator function on  $\{(\hat{x}_i, x_i) \in \mathcal{X}_i \times \mathcal{X}_i | \hat{x}_i \neq x_i\}$ . The special case of an  $m$ -ary hypothesis test corresponds to  $|\mathcal{X}_i| = m$  for every  $i$  and prior probabilities  $p(x)$  such that  $\Pr[X_1 = X_2 = \dots = X_n] = 1$ .

Before formalizing the network-constrained processing model depicted in Fig. 1(b), we highlight two special types of processing strategies that rely on only the classical theory. The first we call the optimal centralized strategy  $\bar{\gamma}$ , which minimizes the detection penalty in (1) but pays no regard to the possibility of online communication constraints; the second we call the myopic decentralized strategy  $\underline{\gamma}$ , which strictly enforces zero online communication overhead (and, of course, generally yields a greater penalty). The set of feasible decentralized strategies we define thereafter will explicitly exclude the former yet always include the latter, so their respective penalties can be viewed as lower and upper baselines to that of any sensible decentralized solution. Moreover, the manner with which these baseline strategies are presented (in (2) and (5), respectively) puts our problem into a form suitable and natural for the message-passing algorithm we develop in Section III.

1) *Optimal Centralized Strategy*: Because  $p(x|y)$  is proportional to  $p(x, y) = p(x)p(y|x)$  for every  $y \in \mathcal{Y}$  such that  $p(y)$  is nonzero, it follows that  $\bar{\gamma}$  minimizes (1) if and only if

$$\bar{\gamma}(Y) = \arg \min_{\hat{x} \in \mathcal{X}} \sum_{x \in \mathcal{X}} p(x)c(\hat{x}, x)p(Y|x) \quad (2)$$

with probability one. Note that (i) the likelihood function  $p(Y|x)$ , taking its values  $L(y) = p(y|x)$  in the product set  $\mathcal{L} = [0, \infty)^{|\mathcal{X}|}$ , provides a sufficient statistic of the *online* measurement process  $Y$  and (ii) the parameter matrix  $\theta \in [0, \infty)^{|\mathcal{X}| \times |\mathcal{X}|}$ , where the optimal values are given by  $\bar{\theta}(\hat{x}, x) = p(x)c(\hat{x}, x)$ , can be determined *offline* (i.e., before receiving actual measurements). One views the optimal detector in (2) as a particular partition of the set  $\mathcal{L}$  into the regions  $\mathcal{L}^1, \dots, \mathcal{L}^{|\mathcal{X}|}$ , always choosing  $\hat{x}$  such that  $L(y) \in \mathcal{L}^{\hat{x}}$ . This equivalence stems from the fact that (2) implies

$$p(\hat{x}|y; \bar{\gamma}) = \begin{cases} 1, & \text{if } \hat{x} = \bar{\gamma}(y) \\ 0, & \text{otherwise} \end{cases}$$

and, because  $p(y, \hat{x}|x; \bar{\gamma}) = p(y|x)p(\hat{x}|y; \bar{\gamma})$ , the identity

$$p(\hat{x}|x; \bar{\gamma}) = \int_{y \in \{y' \in \mathcal{Y} | L(y') \in \mathcal{L}^{\hat{x}}\}} p(y|x) dy \quad (3)$$

from which the achieved penalty is determined according to

$$J_d(\bar{\gamma}) = \sum_{x \in \mathcal{X}} p(x) \sum_{\hat{x} \in \mathcal{X}} c(\hat{x}, x)p(\hat{x}|x; \bar{\gamma}). \quad (4)$$

Of course, the sums and integrals over product spaces  $\mathcal{X}$  and  $\mathcal{Y}$  in (2)–(4) become difficult to compute as  $n$  grows large, especially in the absence of additional problem structure.

The following elementary examples will help the reader relate the general notation employed above to the more familiar presentation of classical detection theory. They are also relevant to our experimental setup in Section IV.

*Example 1 (Binary Detection)*: Suppose the hidden state process  $X$  takes just two values, which we will label as  $-1$  and  $+1$ . Making the natural assumption that an error event  $\hat{X} \neq x$  is more costly than an error-free event  $\hat{X} = x$  for either possible value of  $x$ , the function in (2) is equivalent to the so-called *likelihood-ratio threshold rule*

$$\frac{p_{Y|X}(y|+1)}{p_{Y|X}(y|-1)} \equiv \Lambda(y) \begin{matrix} \hat{x} = +1 \\ > \\ < \\ \hat{x} = -1 \end{matrix} \bar{\eta} \equiv \frac{\bar{\theta}(+1, -1) - \bar{\theta}(-1, -1)}{\bar{\theta}(-1, +1) - \bar{\theta}(+1, +1)}.$$

The distribution  $p(\hat{x}|x; \gamma)$  in (3) is then defined by the *false-alarm* and *detection* probabilities, or  $\Pr[\Lambda(Y) > \eta | X = x]$  with  $x = -1$  and  $x = +1$ , respectively.

*Example 2 (Linear Binary Detectors)*: The special case of a linear-Gaussian measurement model allows the decision regions in measurement space  $\mathcal{Y}$  to retain the polyhedral form of their counterparts in likelihood space  $\mathcal{L}$ , simplifying the multi-dimensional integrals that must be solved in (3). Starting with the binary problem in Example 1, denote by  $\mu^-$  and  $\mu^+$  the real-valued signal vectors associated to the two possible states and assume the measurement process is  $Y = \mu^X + W$ , where the additive noise process  $W$  is a zero-mean Gaussian random vector with (known) covariance matrix  $\Sigma$ . The likelihood-ratio threshold rule then specializes to

$$\begin{matrix} \hat{x} = +1 \\ (\mu^+ - \mu^-)' \Sigma^{-1} y > \\ < \\ \hat{x} = -1 \end{matrix} \log(\eta) + \frac{\mu^{+'} \Sigma^{-1} \mu^+ - \mu^{-'} \Sigma^{-1} \mu^-}{2}$$

which is linear in the measurement vector  $y$ . If  $Y$  is also scalar (in which case so is the signal  $\mu^X$  and noise variance  $\sigma^2$ ), then the threshold rule with parameter  $\eta$  (in likelihood space) simplifies to a threshold rule in measurement space, comparing  $y$  to parameter  $\tau = (\sigma^2 / (\mu^+ - \mu^-)) \log(\eta) + ((\mu^- + \mu^+) / 2)$ . Accordingly, the false-alarm and detection probabilities simplify to  $\Pr[Y > \tau | X = x] = 1 - \Phi((\tau - \mu^x) / \sigma)$  with  $\Phi$  denoting the cumulative distribution function of a zero-mean, unit-variance Gaussian random variable.

*Example 3 (Binary Detection With Non-Binary Decisions)*: Consider the binary problem in Example 1 but where the decision space, call it  $\mathcal{U}$ , can have cardinality  $d \geq 2$ . Any given rule parameters  $\theta \in [0, \infty)^{d \times 2}$  define a particular partition of the likelihood-ratio space  $[0, \infty)$  into (at most)  $d$  subintervals, characterized by  $d - 1$  threshold values satisfying

$$0 \leq \eta^1 \leq \eta^2 \leq \dots \leq \eta^{d-1} \leq \infty.$$

This *monotone threshold rule* alongside the natural assumption that the  $d$  elements of  $\mathcal{U}$  are labeled such that

$$\frac{\Pr[U = u | X = +1]}{\Pr[U = u | X = -1]} \leq \frac{\Pr[U = u + 1 | X = +1]}{\Pr[U = u + 1 | X = -1]}$$

for every  $u = 1, \dots, d-1$  simplifies to making the decision  $u \in \mathcal{U}$  such that  $\Lambda(y) \in [\eta^{u-1}, \eta^u)$ , taking  $\eta^0 = 0$  and  $\eta^d = \infty$ . In the special case of a scalar linear binary detector (see Example 2), we retain the analogous partition in measurement space  $(-\infty, \infty)$  with respective thresholds

$$-\infty \leq \tau^1 \leq \tau^2 \leq \dots \leq \tau^{d-1} \leq \infty$$

determined by  $\tau^u = (\sigma^2/(\mu^+ - \mu^-)) \log(\eta^u) + ((\mu^- + \mu^+)/2)$ . The rule simplifies to deciding  $u$  such that  $y \in [\tau^{u-1}, \tau^u)$ , so  $\Pr[U = u | X = x] = \Phi((\tau^u - \mu^x)/\sigma) - \Phi((\tau^{u-1} - \mu^x)/\sigma)$ .

2) *Myopic Decentralized Strategy*: Assume each sensor  $i$  is initialized knowing only its *local* model for Bayesian detection i.e., the distribution  $p(x_i, y_i)$  and a cost function  $c(\hat{x}_i, x_i)$ , and, using only this local model, determines its component estimate  $\hat{x}_i$  as if in isolation i.e., the rule at each node  $i$  is

$$\underline{\gamma}_i(Y_i) = \arg \min_{\hat{x}_i \in \mathcal{X}_i} \sum_{x_i \in \mathcal{X}_i} \underbrace{p(x_i) c(\hat{x}_i, x_i)}_{\phi_i(\hat{x}_i, x_i) \in \mathbb{R}} p(Y_i | x_i). \quad (5)$$

That is, the myopic strategy is a particular collection of single-sensor rules  $\underline{\gamma} = (\underline{\gamma}_1, \dots, \underline{\gamma}_n)$  specified offline by parameters  $\underline{\phi} = (\phi_1, \dots, \phi_n)$ , where no one node transmits nor receives information and total online computation scales linearly in  $n$ . It is easy to see that the myopic strategy is sub-optimal, meaning  $J_d(\underline{\gamma}) \geq J_d(\bar{\gamma})$ . Equality is achieved only in certain degenerate (and arguably uninteresting) cases, including the zero cost function i.e.,  $c(\hat{x}, x) = 0$  for all  $(\hat{x}, x) \in \mathcal{X} \times \mathcal{X}$ , or the case of  $n$  unrelated single-sensor problems i.e.,  $p(x, y) = \prod_{i=1}^n p(x_i, y_i)$  and  $c(\hat{x}, x) = \sum_{i=1}^n c(\hat{x}_i, x_i)$ . More generally, the extent to which the myopic strategy  $\underline{\gamma}$  falls short from optimal performance, or the *loss*  $J_d(\underline{\gamma}) - J_d(\bar{\gamma})$ , remains a complicated function of the global detection model i.e., the distribution  $p(x, y)$  and cost function  $c(\hat{x}, x)$ .

While the optimal centralized strategy  $\bar{\gamma}$  and the myopic decentralized strategy  $\underline{\gamma}$  are both functions that map  $\mathcal{Y}$  to  $\mathcal{X}$ , the different processing assumptions amount to different size- $|\mathcal{X}|$  partitions of the likelihood space  $\mathcal{L}$ . In particular, assuming myopic processing, the strategy-dependent conditional distribution in the integrand of (3) inherits the factored structure  $p(\hat{x}|y; \underline{\gamma}) = \prod_{i=1}^n p(\hat{x}_i | y_i; \underline{\gamma}_i)$ , where each  $i$ th term involves variables only at the individual node  $i$ . This structure has desirable computational ramifications: suppose only the decision at node  $i$  is costly, captured by choosing  $c(\hat{x}, x) = c(\hat{x}_i, x_i)$  for all  $(\hat{x}, x) \in \mathcal{X} \times \mathcal{X}$ . Then, the strategy  $\bar{\gamma}_i : \mathcal{Y} \rightarrow \mathcal{X}_i$  defined by selecting the  $i$ th component of  $\hat{x} = \bar{\gamma}(y)$  is the global minimizer of (1), achieving penalty

$$J_d(\bar{\gamma}_i) = \sum_{x_i} p(x_i) \sum_{\hat{x}_i} c(\hat{x}_i, x_i) \overbrace{\sum_{x_{-i}} \sum_{\hat{x}_{-i}} p(x_{-i} | x_i) p(\hat{x}_{-i} | x_i; \bar{\gamma})}^{p(\hat{x}_{-i} | x_i; \bar{\gamma}_i)}.$$

In contrast, the myopic rule  $\underline{\gamma}_i$  minimizes (1) over only the subset of rules of the form  $\gamma_i : \mathcal{Y}_i \rightarrow \mathcal{X}_i$ , achieving penalty

$$J_d(\underline{\gamma}_i) = \sum_{x_i} p(x_i) \sum_{\hat{x}_i} c(\hat{x}_i, x_i) \overbrace{\int_{y_i} p(y_i | x_i) p(\hat{x}_i | y_i; \underline{\gamma}_i) dy_i}^{p(\hat{x}_i | x_i; \underline{\gamma}_i)} \quad (6)$$

regardless of the non-local conditional distribution  $p(x_{-i}, y_{-i} | x_i, y_i)$  and the collective strategy  $\gamma_{-i} : \mathcal{Y}_{-i} \rightarrow \mathcal{X}_{-i}$  of all other nodes. Thus, assuming myopic processing constraints and focusing on a cost function local to node  $i$ , the global penalty  $J_d$  involves sums and integrals over only local random variables  $(X_i, Y_i, \hat{X}_i)$ . This simplification foreshadows the key problem structure to be exploited in our later developments, seeking to retain a similarly tractable decomposition of the general  $n$ -sensor sums and integrals, yet also relaxing the constraint of zero online communication and considering costs that can depend on all sensors' decisions.

## B. Network-Constrained Online Processing Model

We now turn to the formal description of Fig. 1(b), which in contrast to the preceding subsection assumes that the decision vector  $\hat{x} \in \mathcal{X}$  is generated component-wise in the forward partial order of a given  $n$ -node directed acyclic graph  $\mathcal{G}$ . Each edge  $(i, j)$  in  $\mathcal{G}$  indicates a (perhaps unreliable) low-rate communication link from node  $i$  to node  $j$ . In particular, each node  $i$ , observing only the component measurement  $y_i$  and the symbol(s)  $z_i$  received on incoming links with all parents  $pa(i) = \{j | \text{edge } (j, i) \text{ in } \mathcal{G}\}$  (if any), is to decide upon both its component estimate  $\hat{x}_i$  and the symbol(s)  $u_i$  transmitted on outgoing links with all children  $ch(i) = \{j | \text{edge } (i, j) \text{ in } \mathcal{G}\}$  (if any). The collections of received symbols  $z$  and transmitted symbols  $u$  take their values in discrete product spaces  $\mathcal{Z} = \mathcal{Z}_1 \times \dots \times \mathcal{Z}_n$  and  $\mathcal{U} = \mathcal{U}_1 \times \dots \times \mathcal{U}_n$ , respectively. However, the exact cardinality of the symbol spaces  $\mathcal{Z}_i$  and  $\mathcal{U}_i$  local to each node  $i$  will depend upon its in-degree  $|pa(i)|$  and out-degree  $|ch(i)|$  in  $\mathcal{G}$ , respectively, as well as the specifics of the channel models and transmission schemes.

Suppose each edge  $(i, j)$  in  $\mathcal{G}$  is assigned an integer  $d_{i \rightarrow j} \geq 2$  denoting the size of the symbol set supported by this link (i.e., the link rate is  $\log_2 d_{i \rightarrow j}$  bits per measurement). The symbol(s)  $u_i$  transmitted by node  $i$  can thus take *at most*  $\prod_{j \in ch(i)} d_{i \rightarrow j}$  distinct values. For example, a scheme in which node  $i$  may transmit a different symbol to each child is modeled by a finite set  $\mathcal{U}_i$  with cardinality equal to  $\prod_{j \in ch(i)} d_{i \rightarrow j}$ , while a scheme in which node  $i$  transmits the same symbol to every child corresponds to  $|\mathcal{U}_i| = \min_{j \in ch(i)} d_{i \rightarrow j}$ . In any case, the convention here is that each node must somehow compress its local data into a relatively small number of logical outgoing symbols (e.g., one symbol per outgoing link). In turn, the cardinality of each node's set of incoming symbols  $\mathcal{Z}_i$  will reflect the joint cardinality  $|\mathcal{U}_{pa(i)}| = \prod_{j \in pa(i)} |\mathcal{U}_j|$  of its parents' transmissions, but the exact relation is determined by the given multipoint-to-point channel into each node  $i$ . In any case, each such channel is modeled by a conditional distribution  $p(z_i | x, y, u_{pa(i)})$ , describing the information  $Z_i$  received by node  $i$  based on its parents' transmitted symbols  $u_{pa(i)} = \{u_j \in \mathcal{U}_j | j \in pa(i)\}$ .<sup>1</sup>

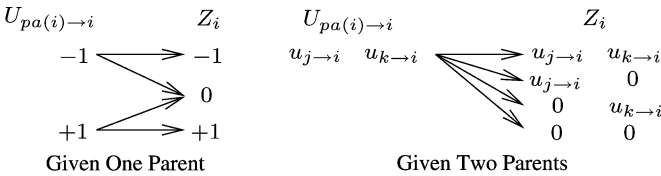
The following examples demonstrate how different classes of transmission schemes and channel models manifest themselves in different discrete symbol sets  $\mathcal{Z}_i$  and  $\mathcal{U}_i$ . They are described

<sup>1</sup>Here, we have also allowed the channel model to depend on the processes  $(X, Y)$  of the environment external to the network. Whether such generality is warranted will, of course, depend on the application (e.g., the sensor seeks to detect the presence of a malicious jammer), and later sections will indeed sacrifice some of this generality in the interest of scalability.

assuming independence of the processes  $(X, Y)$  external to the network i.e.,  $p(z_i|x, y, u_{pa(i)}) = p(z_i|u_{pa(i)})$ .

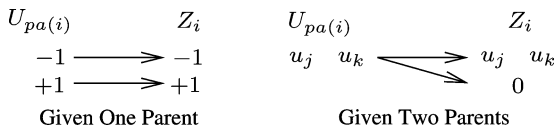
*Example 4 (Peer-to-Peer Binary Comms With Erasures):* Associate each edge  $(i, j)$  in directed graph  $\mathcal{G}$  with a unit-rate communication link, meaning  $d_{i \rightarrow j} = 2$ . If  $u_{i \rightarrow j} \in \{-1, +1\}$  denotes the actual symbol transmitted by node  $i$  to its child  $j \in ch(i)$ , then the collective communication decision  $u_i$  takes its values in  $\mathcal{U}_i = \{-1, +1\}^{|ch(i)|}$ . On the receiving end, let  $z_{j \rightarrow i} \in \{-1, 0, +1\}$  denote the actual symbol received by node  $i$  from its parent  $j \in pa(i)$ , where the value “0” indicates an erasure and otherwise  $z_{j \rightarrow i} = u_{j \rightarrow i}$ . It follows that the aggregate symbol  $z_i$  received by node  $i$  takes values in  $\mathcal{Z}_i = \{-1, 0, +1\}^{|pa(i)|}$ . Let us denote the collection of all symbols transmitted to a particular node  $j$  by  $u_{pa(j) \rightarrow j} = \{u_{i \rightarrow j}; i \in pa(j)\}$ . If we assume all incoming links from parents  $pa(i)$  experience independent and identically-distributed erasures, each such link erased with probability  $q_i \in [0, 1]$ , then the local channel model is  $p(z_i|u_{pa(i)}) = p(z_i|u_{pa(i) \rightarrow i}) = \prod_{j \in pa(i)} p(z_{j \rightarrow i}|u_{j \rightarrow i})$  with

$$p(z_{j \rightarrow i}|u_{j \rightarrow i}) = \begin{cases} 1 - q_i, & z_{j \rightarrow i} = u_{j \rightarrow i} \\ q_i, & z_{j \rightarrow i} = 0. \end{cases}$$



*Example 5 (Broadcast Binary Comms With Interference):* As in Example 4, let  $d_{i \rightarrow j} = 2$  for each edge  $(i, j)$  in  $\mathcal{G}$ . However, now assume each node  $i$  always transmits the same binary-valued symbol to all of its children, meaning  $\mathcal{U}_i = \{-1, +1\}$ . On the receiving end, suppose there are two possibilities: either  $z_i = u_{pa(i)}$  or, when there are two or more parents, none of the incoming symbols are received due to inter-symbol interference. Denoting the latter event by  $z_i = 0$ , it follows that  $\mathcal{Z}_i = \{-1, +1\}^{|pa(i)|} \times \{0\}$ . If we assume all incoming links from parents  $pa(i)$  collectively interfere with one another, each such link blocking the others with probability  $q_i \in [0, 1]$ , then the local channel model is

$$p(z_i|u_{pa(i)}) = \begin{cases} (1 - q_i)^{|pa(i)|-1}, & z_i = u_{pa(i)} \\ 1 - (1 - q_i)^{|pa(i)|-1}, & z_i = 0. \end{cases}$$



### C. Decentralized Formulation With Costly Communication

As in the classical Bayesian formulation, let distribution  $p(x, y)$  jointly describe the hidden state process  $X$  and the noisy measurement process  $Y$ . Recall that in Fig. 1(a), the only other random vector is the decision process  $\hat{X} = \gamma(Y)$ , derived from  $Y$  by a function of the form  $\gamma : \mathcal{Y} \rightarrow \mathcal{X}$ . In Fig. 1(b), however, there also exist the communication-related random vectors  $Z$  and  $U$ , collectively derived from  $Y$  by the successive application of local channels  $p(z_i|x, y, u_{pa(i)})$  and local rules  $(U_i, \hat{X}_i) = \gamma_i(Y_i, Z_i)$  in the forward partial order

of network topology  $\mathcal{G}$ . Denote by  $\bar{\Gamma}$  the set of all functions of the form  $\gamma : \mathcal{Y} \times \mathcal{Z} \rightarrow \mathcal{U} \times \mathcal{X}$ , which includes all centralized strategies that remain agnostic about  $Z$  and  $U$ , and denote by  $\Gamma \subset \bar{\Gamma}$  only the *admissible* subset in  $\bar{\Gamma}$ , which is equivalent to  $\Gamma = \Gamma_1 \times \dots \times \Gamma_n$  with each  $\Gamma_i$  denoting the set of all feasible rules  $\gamma_i : \mathcal{Y}_i \times \mathcal{Z}_i \rightarrow \mathcal{U}_i \times \mathcal{X}_i$  local to node  $i$ .

The Bayesian performance criterion is essentially the same as in the centralized detection problem, accounting also for the communication-related decision process  $U$ . Specifically, let every possible realization of the joint process  $(U, \hat{X}, X)$  be assigned a cost of the form  $c(u, \hat{x}, x) = c(\hat{x}, x) + \lambda c(u, x)$ , where non-negative constant  $\lambda$  specifies the unit conversion between detection costs  $c(\hat{x}, x)$  and communication costs  $c(u, x)$ . In turn, the global penalty function is given by

$$J(\gamma) = \mathbf{E} [c(U, \hat{X}, X)] = \mathbf{E} [\mathbf{E} [c(\gamma(Y, Z), X) | Y, Z]] \quad (7)$$

and the decentralized design problem is to find the strategy  $\gamma^* \in \Gamma \subset \bar{\Gamma}$  such that

$$J(\gamma^*) = \min_{\gamma \in \bar{\Gamma}} J_d(\gamma) + \lambda J_c(\gamma) \text{ subject to } \gamma \in \Gamma \quad (8)$$

where functions  $J_d : \bar{\Gamma} \rightarrow \mathbb{R}$  and  $J_c : \bar{\Gamma} \rightarrow \mathbb{R}$  quantify the *detection penalty* and *communication penalty*, respectively. Viewing (8) as a multi-objective criterion parameterized by  $\lambda$ , the achievable design tradeoff is then captured by the *Pareto-optimal* planar curve  $\{(J_c(\gamma^*), J_d(\gamma^*)); \lambda \geq 0\}$ .

The formulation in (8) specializes to the centralized design problem when online communication is both unconstrained and unpenalized i.e.,  $\Gamma$  is the set of all functions  $\gamma : \mathcal{Y} \rightarrow \mathcal{X}$  and  $\lambda = 0$ . Given any (finite-rate) directed network, however, the function space  $\Gamma$  excludes the optimal centralized strategy  $\bar{\gamma}$  in (2), but still includes the myopic decentralized strategy  $\underline{\gamma}$  in (5). The non-ideal communication model also manifests itself as a factored representation within the distribution underlying (7). By construction, fixing a rule  $\gamma_i \in \Gamma_i$  is equivalent to specifying

$$p(u_i, \hat{x}_i | y_i, z_i; \gamma_i) = \begin{cases} 1, & \text{if } (u_i, \hat{x}_i) = \gamma_i(y_i, z_i) \\ 0, & \text{otherwise.} \end{cases}$$

It follows that fixing a strategy  $\gamma \in \Gamma \subset \bar{\Gamma}$  specifies

$$p(u, z, \hat{x} | x, y; \gamma) = \prod_{i=1}^n p(z_i | x, y, u_{pa(i)}) p(u_i, \hat{x}_i | y_i, z_i; \gamma_i) \quad (9)$$

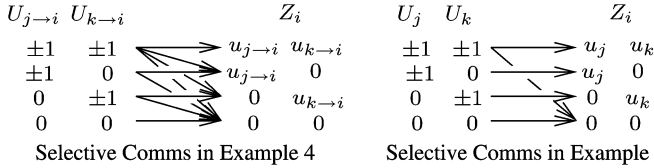
reflecting the sequential processing implied by the directed network topology  $\mathcal{G}$ . In turn, the distribution that determines the global penalty function  $J(\gamma)$  in (7) becomes

$$p(u, \hat{x}, x; \gamma) = \int_{\mathcal{Y}} p(x, y) \prod_{i=1}^n p(u_i, \hat{x}_i | x, y, u_{pa(i)}; \gamma_i) dy \quad (10)$$

where the summation over  $\mathcal{Z}$  is taken inside the product i.e., for each node  $i$ , we have  $p(u_i, \hat{x}_i | x, y, u_{pa(i)}; \gamma_i) = \sum_{z_i} p(z_i | x, y, u_{pa(i)}) p(u_i, \hat{x}_i | y_i, z_i; \gamma_i)$ . Note that the integration over  $\mathcal{Y}$  cannot be decomposed in the absence of additional model assumptions, a possibility we explore subsequently.

*Example 6 (Selective Binary Comms):* As in Examples 4 & 5, let  $d_{i \rightarrow j} = 2$  for each edge  $(i, j)$  in  $\mathcal{G}$ . A selective communication scheme refers to each node having the option to suppress transmission, or remain silent, on one or more of its outgoing

links. We denote this option by the symbol “0,” and we assume it is always both cost-free and reliably received. In Example 4, this implies any communicating node  $i$  selects from an augmented decision space of  $\mathcal{U}_i = \{-1, 0, +1\}^{|ch(i)|}$ . Meanwhile, upon receiving  $z_{i \rightarrow j} = 0$ , any child  $j \in ch(i)$  is then uncertain as to whether node  $i$  selected silence or link  $(i, j)$  experienced an erasure; on the other hand, if  $z_{i \rightarrow j} \neq 0$ , then child  $j$  knows neither selective silence nor an erasure has occurred. In Example 5, we let  $\mathcal{U}_i = \{-1, 0, +1\}$  for node  $i$  and capture interference effects only among the subset of actively transmitting parents.



#### D. Team-Theoretic Solution

In general, it is not known whether the strategy  $\gamma^*$  in (8) lies in a finitely-parameterized subspace of  $\Gamma$ . The team-theoretic approximation used here is to satisfy a set of *person-by-person* optimality conditions, each based on a simple observation: if a decentralized strategy  $\gamma^* = (\gamma_1^*, \dots, \gamma_n^*)$  is optimal over  $\Gamma$ , then for each  $i$  and assuming rules  $\gamma_{-i}^* = \{\gamma_j^* \in \Gamma_j | j \neq i\}$  are fixed, the rule  $\gamma_i^*$  is optimal over  $\Gamma_i$  i.e., for each  $i$

$$\gamma_i^* = \arg \min_{\gamma_i \in \Gamma_i} J_d(\gamma_{-i}^*, \gamma_i) + \lambda J_c(\gamma_{-i}^*, \gamma_i). \quad (11)$$

Simultaneously satisfying (11) for all  $i$  is (by definition) a necessary optimality condition, but it is not sufficient because, in general, it does not preclude a decrease in  $J$  via joint minimization over multiple nodes. Under certain model assumptions, however, finding a solution to the  $n$  coupled optimization problems in (11) reduces analytically to finding a fixed-point of a particular system of nonlinear equations.

In this and subsequent sections we introduce a sequence of model assumptions, each of which introduces additional problem structure that we exploit to construct our efficient offline algorithm. We do this in steps to help elucidate the value and impact of each successive assumption.

*Assumption 1 (Conditional Independence):* Conditioned on the state process  $X$ , the measurement  $Y_i$  and received symbol  $Z_i$  local to node  $i$  are mutually independent as well as independent of all other information observed in the network, namely the measurements  $Y_{-i}$  and symbols  $Z_{-i}$  received by all other nodes i.e., for every  $i$

$$p(y_i, z_i | x, y_{-i}, z_{-i}, u_{-i}) = p(y_i | x) p(z_i | x, u_{pa(i)}). \quad (12)$$

Assumption 1 is satisfied if e.g., each measurement  $Y_i$  is a function of  $X$  corrupted by noise, each received symbol  $Z_i$  is a function of  $X$  (and transmitted symbols  $U_{pa(i)}$ ) corrupted by noise, and all such noise processes are mutually independent.

*Lemma 1 (Factored Representation):* Let Assumption 1 hold. For every strategy  $\gamma \in \Gamma$ , (10) specializes to

$$p(u, \hat{x}, x; \gamma) = p(x) \prod_{i=1}^n p(u_i, \hat{x}_i | x, u_{pa(i)}; \gamma_i)$$

where for every  $i$

$$p(u_i, \hat{x}_i | x, u_{pa(i)}; \gamma_i) = \sum_{z_i} p(z_i | x, u_{pa(i)}) \times \int_{y_i} p(y_i | x) p(u_i, \hat{x}_i | y_i, z_i; \gamma_i) dy_i. \quad (13)$$

*Proof:* Substituting (12) into (9) and (10) results in

$$p(u, \hat{x} | x; \gamma) = \sum_z \int_y \prod_{i=1}^n p(y_i | x) \times p(z_i | x, u_{pa(i)}) p(u_i, \hat{x}_i | y_i, z_i; \gamma_i) dy_i.$$

Because only the  $i$ th factor in the integrand involves variables  $(y_i, z_i)$ , global marginalization over  $(Y, Z)$  simplifies to  $n$  local marginalizations, each over  $(Y_i, Z_i)$ . ■

*Proposition 1 (Person-by-Person Optimality):* Let Assumption 1 hold. The  $i$ th component optimization in (11) reduces to

$$\gamma_i^*(Y_i, Z_i) = \arg \min_{(u_i, \hat{x}_i) \in \mathcal{U}_i \times \mathcal{X}_i} \sum_{x \in \mathcal{X}} \theta_i^*(u_i, \hat{x}_i, x; Z_i) p(Y_i | x) \quad (14)$$

where, for each  $z_i \in \mathcal{Z}_i$  such that  $p(Y_i, z_i; \gamma_{-i}^*) > 0$ , the parameter values  $\theta_i^*(z_i) \in \mathbb{R}^{|\mathcal{U}_i| \times |\mathcal{X}_i| \times |\mathcal{X}|}$  are given by

$$\theta_i^*(u_i, \hat{x}_i, x; z_i) = p(x) \sum_{u_{-i} \in \mathcal{U}_{-i}} p(z_i | x, u_{pa(i)}) \times \sum_{\hat{x}_{-i} \in \mathcal{X}_{-i}} c(u, \hat{x}, x) \prod_{j \neq i} p(u_j, \hat{x}_j | x, u_{pa(j)}; \gamma_j^*). \quad (15)$$

*Proof:* The proof follows the same key steps by which (2) is derived in the centralized case, but accounting for a composite measurement  $(Y_i, Z_i)$  and a cost function that also depends on non-local decision variables  $(U_{-i}, \hat{X}_{-i})$ . Assumption 1 is essential for the parameter values  $\theta_i^*$  to be independent of the local measurement  $Y_i$ . See [24]. ■

It is instructive to note the similarity between a local rule  $\gamma_i^*$  in Proposition 1 and the centralized strategy in (2). Both process an  $|\mathcal{X}|$ -dimensional sufficient statistic of the available measurement with optimal parameter values to be computed offline. In rule  $\gamma_i^*$ , however, the computation is more than simple multiplication of probabilities  $p(x)$  and costs  $c(u, \hat{x}, x)$ : parameter values  $\theta_i^* \in \mathbb{R}^{|\mathcal{U}_i| \times |\mathcal{X}_i| \times |\mathcal{X}| \times |\mathcal{Z}_i|}$  in (15) now involve conditional expectations, taken over distributions that depend on the fixed rules  $\gamma_{-i}^*$  of all other nodes  $j \neq i$ . Each such fixed rule  $\gamma_j^*$  is similarly of the form in Proposition 1, where fixing parameter values  $\theta_j^*$  specifies  $p(u_j, \hat{x}_j | x, u_{pa(j)}; \theta_j^*)$  local to node  $j$  through (13) and (14). Each  $i$ th minimization in (11) is thereby equivalent to minimizing

$$J(\gamma_{-i}^*, \gamma_i) = \sum_{x \in \mathcal{X}} p(x) \sum_{u \in \mathcal{U}} \sum_{\hat{x} \in \mathcal{X}} c(u, \hat{x}, x) p(u, \hat{x} | x; \theta_{-i}^*, \theta_i)$$

over the parameterized space of distributions defined by

$$p(u, \hat{x} | x; \theta_{-i}^*, \theta_i) = p(u_i, \hat{x}_i | x, u_{pa(i)}; \theta_i) \times \prod_{j \neq i} p(u_j, \hat{x}_j | x, u_{pa(j)}; \theta_j^*).$$

It follows that the simultaneous satisfaction of (11) at all nodes corresponds to solving for  $\theta^* = (\theta_1^*, \dots, \theta_n^*)$  in a system of nonlinear equations expressed by (13)–(15). Specifically, if we let  $f_i(\theta_{-i}^*)$  denote the right-hand-side of (15), then offline computation of a person-by-person optimal strategy reduces to solving the fixed-point equations

$$\theta_i = f_i(\theta_{-i}), \quad i = 1, \dots, n. \quad (16)$$

*Corollary 1 (Offline Iterative Algorithm):* Initialize parameters  $\theta^0 = (\theta_1^0, \dots, \theta_n^0)$  and generate the sequence  $\{\theta^k\}$  by iterating (16) in any component-by-component order e.g.,

$$\theta_i^k := f_i(\theta_1^k, \dots, \theta_{i-1}^k, \theta_{i+1}^{k-1}, \dots, \theta_n^{k-1}), \quad i = 1, \dots, n$$

in iteration  $k = 1, 2, \dots$ . If Assumption 1 holds, then the associated sequence  $\{J(\gamma^k)\}$  is non-increasing and converges.

*Proof:* By Proposition 1, each operator  $f_i$  is the solution to the minimization of  $J$  over the  $i$ th coordinate function space  $\Gamma_i$ . Thus, a component-wise iteration of  $f$  is equivalent to a coordinate-descent iteration of  $J$ , implying  $J(\gamma^k) \leq J(\gamma^{k-1})$  for every  $k$  [25]. Because the real-valued, non-increasing sequence  $\{J(\gamma^k)\}$  is bounded below, it has a limit point. ■

In the absence of additional technical conditions (e.g.,  $J$  is convex,  $f$  is contracting [25]), it is *not* known whether the sequence  $\{J(\gamma^k)\}$  converges to the optimal performance  $J(\gamma^*)$ , whether the achieved performance is invariant to the choice of initial parameters  $\theta^0$ , nor whether the associated sequence  $\{\theta^k\}$  converges. Indeed, the possibility of a poorly performing person-by-person-optimal strategy is known to exist (see [26] for such crafted examples). These theoretical limitations are inherent to nonlinear minimization problems, in general, where second-order optimality conditions can be “locally” satisfied at many points, but only one of them may achieve the “global” minimum [25]. Nonetheless, the iterative algorithm is often reported to yield reasonable decision strategies, which has also been our experience (in experiments to be described in Section IV) providing the initialization is done with some care.

Corollary 1 also assumes that every node  $i$  can *exactly* compute the local marginalization of (13). Some measurement models of practical interest lead to numerical or Monte-Carlo approximation of these marginalizations at each iteration  $k$ , and the extent to which the resulting errors may affect convergence is also not known. This issue is beyond the scope of this paper and, as such, all of our experiments will involve sensor models in which such complications do not arise (e.g., local instantiations of the models in Examples 1 to 3).

### III. MESSAGE-PASSING ALGORITHM

Online measurement processing implied by Proposition 1 is, by design, well-suited for distributed implementation in network topology  $\mathcal{G}$ . However, two practical difficulties remain: firstly, convergent offline optimization requires global knowledge of probabilities  $p(x)$ , costs  $c(u, \hat{x}, x)$  and statistics  $\{p(u_i, \hat{x}_i | x, u_{pa(i)}; \theta_i^k)\}$  in every iteration  $k$ ; and secondly, total (offline and online) memory/computation requirements scale exponentially with the number of nodes  $n$ . In this section, we establish conditions so that convergent offline optimization

can be executed in a recursive fashion: each node  $i$  starts with local probabilities  $p(x_{pa(i)}, x_i)$  and local costs  $c(u_i, \hat{x}_i, x_i)$ , then in each iteration computes and exchanges rule-dependent statistics, or *messages*, with only its neighbors  $pa(i) \cup ch(i)$  in  $\mathcal{G}$ . We will interpret this message-passing algorithm as an instance of Corollary 1 under additional model assumptions. Thus, when these additional assumptions hold, it inherits the same convergence properties. Moreover, total memory/computation requirements scale only linearly with  $n$ .

Our message-passing algorithm is primarily built upon the computational theory discussed in [5]–[7], [9]–[11], albeit each of these considers only certain special cases of Fig. 1 and so one contribution in this paper is the generality with which the results apply. For example, in contrast to [6], our derivation need not assume *from the start* that all nodes must employ local likelihood-ratio tests, nor that the penalty function  $J$  is differentiable with respect to the threshold parameters. (Local likelihood-ratio tests in the general form discussed in Section II are appropriate, however, so we have not negated the algorithm in [6] but rather broadened its applicability.) Our main contribution, however, stems from our emphasis not just on preserving algorithm correctness as we make these generalizations, but also on preserving algorithm efficiency. As will be discussed, a new insight from our analysis is the extent to which the graphical structure underlying the hidden state process may deviate from the network topology without sacrificing either algorithm correctness or efficiency. Moreover, the local recursive structure of the message-passing equations can be applied to topologies beyond those for which it is originally derived, providing a new approximation paradigm for large irregular networks of heterogeneous sensors in which the general algorithm of Corollary 1 is intractable and conclusions based on asymptotic analyses [14]–[19] are not necessarily valid.

#### A. Efficient Online Processing

We first introduce an assumption that removes the exponential dependence on the number of nodes  $n$  of the online computation i.e., the actual operation of the optimized strategy as measurements are received and communication and decision-making takes place. This exponential dependence is due to the appearance of the global state vector  $X$  in (14), specifically the summation over the full product space  $\mathcal{X}$ . The following assumption reduces this online processing to a dependence only on the local state component of each node. The supporting offline computation, however, continues to scale exponentially (in  $n$ ): additional assumptions to reduce this offline complexity are provided in the subsequent subsection.

*Assumption 2 (Measurement/Channel Locality):* In addition to the conditions of Assumption 1, the measurement and channel models local to node  $i$  do not directly depend on any of the non-local state processes  $X_{-i}$  i.e., for every  $i$

$$p(y_i, z_i | x, y_{-i}, z_{-i}, u_{-i}) = p(y_i | x_i) p(z_i | x_i, u_{pa(i)}). \quad (17)$$

*Corollary 2 (Online Efficiency):* If Assumption 2 holds, then (14) and (15) in Proposition 1 specialize to

$$\gamma_i^*(Y_i, Z_i) = \arg \min_{(u_i, \hat{x}_i) \in \mathcal{U}_i \times \mathcal{X}_i} \sum_{x_i \in \mathcal{X}_i} \phi_i^*(u_i, \hat{x}_i, x_i; Z_i) p(Y_i | x_i) \quad (18)$$



and

$$\begin{aligned} \phi_i^*(u_i, \hat{x}_i, x_i; z_i) &= \sum_{x_{-i} \in \mathcal{X}_{-i}} p(x) \sum_{u_{-i} \in \mathcal{U}_{-i}} p(z_i | x_i, u_{pa(i)}) \\ &\times \sum_{\hat{x}_{-i} \in \mathcal{X}_{-i}} c(u, \hat{x}, x) \prod_{j \neq i} p(u_j, \hat{x}_j | x_j, u_{pa(j)}; \gamma_j^*) \end{aligned} \quad (19)$$

respectively.

*Proof:* Recognizing (17) to be the special case of (12) with  $p(y_i | x) = p(y_i | x_i)$  and  $p(z_i | x, u_{pa(i)}) = p(z_i | x_i, u_{pa(i)})$  for every  $i$ , (13) in Lemma 1 similarly specializes to

$$\begin{aligned} p(u_i, \hat{x}_i | x_i, u_{pa(i)}; \gamma_i) &= \sum_{z_i} p(z_i | x_i, u_{pa(i)}) \\ &\times \int_{y_i} p(y_i | x_i) p(u_i, \hat{x}_i | y_i, z_i; \gamma_i) dy_i \end{aligned} \quad (20)$$

for every  $i$ . We then apply Proposition 1 with  $\phi_i^*(u_i, \hat{x}_i, x_i; z_i) = \sum_{x_{-i}} \theta_i^*(u_i, \hat{x}_i, x; z_i)$ . ■

It is instructive to note the similarity between  $\gamma_i^*$  in Corollary 2 and the myopic rule in (5). Online computation is nearly identical, but with  $\gamma_i^*$  using parameters that reflect the composite decision space  $\mathcal{U}_i \times \mathcal{X}_i$  and depend explicitly on the received information  $Z_i = z_i$ . This similarity is also apparent in the offline computation implied by (20) for fixed parameters  $\phi_i^*$  in (18), which per value  $z_i \in \mathcal{Z}_i$  involves the same local marginalization over  $Y_i$  highlighted in (6) for the myopic rule.

Assumption 2 is a special case of Assumption 1, which was discussed to be applicable when the measurement and channel noise processes local to all nodes are mutually independent. Assumption 2 applies when also, for example, the sensing and communication ranges of each node are on the order of their spatial distances to their neighbors. Then, for every node  $i$ , the peripheral state processes  $X_{-i}$  will, in comparison to the local state process  $X_i$ , have negligible bearing on the local measurement and channel models. Note that the channel models defined in Examples 4 & 5 satisfy Assumption 2, as do the measurement models defined in the following example.

*Example 7 (Linear Binary Detectors in Noise):* Let each  $i$ th such sensor be a scalar linear binary detector (see Example 2) with *local* likelihood function given by

$$p(y_i | x_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(y_i - \frac{x_i r_i}{2}\right)^2\right).$$

Here, we have chosen a state-related signal  $\mu^{X_i} = X_i(r_i/2)$  and a unit-variance noise process  $W_i$  so that the single parameter  $r_i \in [0, \infty)$  captures the effective signal strength e.g., measurements by sensor  $i$  are less noisy, or equivalently more informative on the average, than measurements by sensor  $j$  if  $r_i > r_j$ . If the Gaussian noise processes  $W_1, \dots, W_n$  are mutually uncorrelated (i.e., the case of a diagonal covariance matrix  $\Sigma$  in Example 2), then the observable processes  $Y_1, \dots, Y_n$  are also mutually independent conditioned on the global hidden process  $X = x$  i.e., we have  $p(y | x) = \prod_{i=1}^n p(y_i | x_i)$ .

### B. Efficient Offline Optimization

Efficiency in the offline iterative algorithm—i.e., in the algorithm for optimizing at all nodes the decision rules for subsequent online processing—requires not only the statistical locality of Assumption 2 but a bit more, namely that the overall

cost function decomposes into a sum of per-node local costs and that the network topology is a polytree.

*Assumption 3 (Cost Locality):* The Bayesian cost function is additive across the nodes of the network i.e.,

$$c(u, \hat{x}, x) = \sum_{i=1}^n c(u_i, \hat{x}_i, x_i). \quad (21)$$

*Assumption 4 (Polytree Topology):* Graph  $\mathcal{G}$  is a polytree i.e., has at most one (directed) path between any pair of nodes.

*Proposition 2 (Offline Efficiency):* If Assumptions 2 to 4 hold, then (18) applies with (19) specialized to

$$\phi_i^*(u_i, \hat{x}_i, x_i; z_i) \propto p(x_i) P_i^*(z_i | x_i) [c(u_i, \hat{x}_i, x_i) + C_i^*(u_i, x_i)] \quad (22)$$

where (i) the *likelihood function*  $P_i(z_i | x_i)$  for received information  $Z_i$  is determined by the forward recursion

$$\begin{aligned} P_i^*(z_i | x_i) &= \begin{cases} 1, & pa(i) \text{ empty} \\ \sum_{x_{pa(i)}} p(x_{pa(i)} | x_i) \sum_{u_{pa(i)}} p(z_i | x_i, u_{pa(i)}) \\ \quad \times \prod_{j \in pa(i)} P_{j \rightarrow i}^*(u_j | x_j), & \text{otherwise} \end{cases} \end{aligned} \quad (23)$$

with *forward message* from each parent  $j \in pa(i)$  given by

$$P_{j \rightarrow i}^*(u_j | x_j) = \sum_{z_j} P_j^*(z_j | x_j) \sum_{\hat{x}_j} p(u_j, \hat{x}_j | x_j, z_j; \gamma_j^*) \quad (24)$$

and (ii) the *cost-to-go function*  $C_i(u_i, x_i)$  for transmitted information  $U_i$  is determined by the backward recursion

$$C_i^*(u_i, x_i) = \begin{cases} 0, & ch(i) \text{ empty} \\ \sum_{j \in ch(i)} C_{j \rightarrow i}^*(u_i, x_i), & \text{otherwise} \end{cases} \quad (25)$$

with *backward message* from each child  $j \in ch(i)$  given by

$$\begin{aligned} C_{j \rightarrow i}^*(u_i, x_i) &= \sum_{x_j} \sum_{u_j} \sum_{\hat{x}_j} [c(u_j, \hat{x}_j, x_j) + C_j^*(u_j, x_j)] \\ &\quad \times Q_{j \rightarrow i}^*(u_j, \hat{x}_j, x_j | u_i, x_i) \end{aligned} \quad (26)$$

and statistic  $Q_{j \rightarrow i}^*(u_j, \hat{x}_j, x_j | u_i, x_i)$  given by

$$\begin{aligned} Q_{j \rightarrow i}^*(u_j, \hat{x}_j, x_j | u_i, x_i) &= \sum_{x_{pa(j)-i}} p(x_{pa(j)}, x_j | x_i) \\ &\quad \times \sum_{u_{pa(j)-i}} p(u_j, \hat{x}_j | x_j, u_{pa(j)}; \gamma_j^*) \\ &\quad \times \prod_{m \in pa(j)-i} P_{m \rightarrow j}^*(u_m | x_m). \end{aligned} \quad (27)$$

*Proof:* We provide only a sketch here; see [24] for details. Assumption 2 implies the global likelihood function for received information  $Z_i$  is independent of the rules and states local to nodes other than  $i$  and its *ancestors* (i.e., the parents  $pa(i)$ , each such parent's parents, and so on). Assumption 3 implies the global penalty function takes an additive form over all nodes, where terms local to nodes other than  $i$  and its *descendants* (i.e., the children  $ch(i)$ , each such child's children, and so on) cannot be influenced by local decision

$(u_i, \hat{x}_i)$  and, hence, have no bearing on the optimization of rule  $\gamma_i$ . Assumption 4 implies the information observed and generated by all ancestors is independent (conditioned on  $X$ ) of the information *to be* observed and generated by all descendants. This conditional independence between the “upstream” likelihood statistics and the “downstream” penalty terms specializes the parameter values  $\phi_i^*$  of Corollary 2 to the proportionality of (22). Assumption 4 also guarantees no two parents have a common ancestor, implying that upstream likelihoods decompose multiplicatively across parent nodes, and no two children have a common descendant, implying that downstream penalties decompose additively across child nodes. Altogether, it is the structural implications of Assumptions 2 to 4 that yield the recursive formulas (23)–(27).

Proposition 2 has a number of important implications. The first is that parameters  $\phi_i^*$  at node  $i$  are now completely determined by the incoming messages from its neighbors  $pa(i) \cup ch(i)$ . Specifically, we see in (22) that the global meaning of received information  $Z_i$  manifests itself as a Bayesian correction to the myopic prior  $p(x_i)$ , while the global meaning of transmitted information  $U_i$  manifests itself as an additive correction to the myopic cost  $c(u_i, \hat{x}_i, x_i)$ . The former correction requires the likelihood function  $P_i^*$  expressed by (23), uniquely determined from the incoming forward messages  $P_{pa(i) \rightarrow i}^* = \{P_{j \rightarrow i}^*; j \in pa(i)\}$  from all parents, while the latter involves the cost-to-go function  $C_i^*$  expressed by (25), uniquely determined from the incoming backward messages  $C_{ch(i) \rightarrow i}^* = \{C_{j \rightarrow i}^*; j \in ch(i)\}$  from all children. Thus, after substitution of (23) and (25), we see that the right-hand-side of (22) can be viewed as an operator  $f_i(P_{pa(i) \rightarrow i}^*, C_{ch(i) \rightarrow i}^*)$ . Similarly, person-by-person optimality at every node other than  $i$  requires the outgoing messages from node  $i$  to its neighbors  $pa(i) \cup ch(i)$ . The outgoing forward messages  $P_{i \rightarrow ch(i)}^* = \{P_{i \rightarrow j}^*; j \in ch(i)\}$  are collectively determined by the right-hand-side of (24), which after substitution of (23) and (20) we denote by the operator  $g_i(\phi_i^*, P_{pa(i) \rightarrow i}^*)$ . The outgoing backward messages  $C_{i \rightarrow pa(i)}^* = \{C_{i \rightarrow j}^*; j \in pa(i)\}$  are collectively determined by the right-hand-side of (26), which after substitution of (25) and (20) we denote by the operator  $h_i(\phi_i^*, P_{pa(i) \rightarrow i}^*, C_{ch(i) \rightarrow i}^*)$ . Altogether, we see that Proposition 2 specializes the nonlinear fixed-point equations in (16) to the block-structured form

$$\begin{aligned}
 \phi_i &= f_i(P_{pa(i) \rightarrow i}, C_{ch(i) \rightarrow i}) \\
 P_{i \rightarrow ch(i)} &= g_i(\phi_i, P_{pa(i) \rightarrow i}) \quad i = 1, \dots, n. \\
 C_{i \rightarrow pa(i)} &= h_i(\phi_i, P_{pa(i) \rightarrow i}, C_{ch(i) \rightarrow i}) \quad (28)
 \end{aligned}$$

**Corollary 3 (Offline Message-Passing Algorithm):** Initialize all rule parameters  $\phi^0 = (\phi_1^0, \dots, \phi_n^0)$  and generate the sequence  $\{\phi^k\}$  by iterating (28) in a repeated forward-backward pass over  $\mathcal{G}$  e.g., iteration  $k = 1, 2, \dots$  is

$$P_{i \rightarrow ch(i)}^k := g_i(\phi_i^{k-1}, P_{pa(i) \rightarrow i}^k)$$

from  $i = 1, 2, \dots, n$  and

$$\begin{aligned}
 \phi_i^k &:= f_i(P_{pa(i) \rightarrow i}^k, C_{ch(i) \rightarrow i}^k) \\
 C_{i \rightarrow pa(i)}^k &:= h_i(\phi_i^k, P_{pa(i) \rightarrow i}^k, C_{ch(i) \rightarrow i}^k)
 \end{aligned}$$

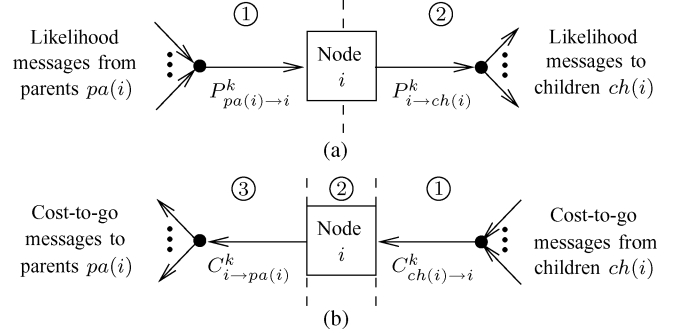


Fig. 2. Distributed message-passing interpretation of the  $k$ th iteration in the offline algorithm of Corollary 3, each node  $i$  interleaving its purely-local computations with only nearest-neighbor communications. (a) Forward pass at node  $i$ : “receive & transmit”; (b) backward pass at node  $i$ : “receive, update & transmit.”

from  $i = n, n-1, \dots, 1$  (see Fig. 2). If Assumptions 2 to 4 hold, then the associated sequence  $\{J(\gamma^k)\}$  converges.

**Proof:** By virtue of Proposition 2, a sequence  $\{\phi^k\}$  is the special case of a sequence  $\{\theta^k\}$  considered in Corollary 1. Each forward-backward pass in the partial-order implied by  $\mathcal{G}$  ensures each iterate  $\phi^k$  is generated in the node-by-node coordinate descent fashion required for convergence. ■

Proposition 2 also implies that, to carry out the iterations defined in Corollary 3, each node no longer needs a complete description of the global state distribution  $p(x)$ . This is arguably surprising, since we have not yet made a restrictive assumption about the state process  $X$ . As seen from (22)–(27), it is sufficient for each node  $i$  to know the joint distribution  $p(x_{pa(i)}, x_i)$  of only the states local to itself and its parents. In our work here, we assume that these local probabilities are available at initialization. However, computing such local probabilities for a general random vector  $X$  has exponential complexity and must often be approximated. Of course, if process  $X$  is itself defined on a graphical model with structure commensurate with the polytree network topology  $\mathcal{G}$ , then the distributed computation to first obtain the local priors  $p(x_{pa(i)}, x_i)$  at each node  $i$  is straightforward and tractable e.g., via belief propagation [12].

A final implication of Proposition 2 is the simplicity with which the sequence  $\{J(\gamma^k)\}$  can be computed. Specifically, the global penalty associated to iterate  $\phi^k$  is given by

$$J(\gamma^k) := \sum_i \sum_{x_i} p(x_i) \sum_{u_i} \sum_{\hat{x}_i} c(u_i, \hat{x}_i, x_i) p(u_i, \hat{x}_i | x_i; \phi^k) \quad (29)$$

with  $p(u_i, \hat{x}_i | x_i; \phi^k) = \sum_{z_i} P_i^{k+1}(z_i | x_i) p(u_i, \hat{x}_i | x_i, z_i; \phi_i^k)$  for every  $i$ . That is, given that the likelihood function  $P_i^{k+1}$  is known local to each node  $i$  (which occurs upon completion of the forward pass in iteration  $k+1$ ), each  $i$ th term in (29) can be locally computed by node  $i$  and, in turn, computation of the total penalty  $J(\gamma^k)$  scales linearly in  $n$ .

As was the case for Corollary 1, the choice of initial parameter vector  $\phi^0$  in Corollary 3 can be important. Consider, for example, initializing to the myopic strategy  $\underline{\gamma} = (\underline{\gamma}_1, \dots, \underline{\gamma}_n)$ , where every node employs the rule in (5) that both ignores its received information and transmits no information (or, equivalently, always transmits the same zero-cost symbol so that  $J_c(\underline{\gamma})$  is zero): given Assumption 2 and Assumption 3 both hold and also assuming  $c(u_i, \hat{x}_i, x_i) = c(\hat{x}_i, x_i) + \lambda c(u_i, x_i)$  for every  $i$ ,

it turns out that this myopic strategy is person-by-person optimal! That is, the parameter vector  $\underline{\phi} = (\phi_1, \dots, \phi_n)$  is itself a fixed-point of (28), and as such the algorithm will make no progress from the associated myopic (and typically sub-optimal) performance  $J(\underline{\gamma}) = J_d(\underline{\gamma})$ . While most details will vary for different classes of models, one general guideline is to initialize with a strategy such that every possible transmission/state pair  $(u_i, x_i)$  at every node  $i$  has a nonzero probability of occurrence. This will ensure that the algorithm explores, at least to some degree, the cost/benefit tradeoff of the online communication, making convergence to the myopic fixed-point likely only when  $\lambda$  is so large in (8) that communication penalty  $J_c(\underline{\gamma}^*)$  should be zero, as will be demonstrated in Section IV.

Assumption 4 is arguably the most restrictive in Proposition 2, in the sense that satisfying it in practice must contend with non-local network connectivity constraints. For example, while any node may have more than one parent node, none of those parents may have a common ancestor. In principle, as illustrated in Fig. 3, this restriction can be removed by merging such parent nodes together into single “super-nodes,” but doing this recognizes the associated need for direct “offline” communication among these merged parent nodes while designing the decision rules (even though these decision rules continue to respect the online network topology  $\mathcal{G}$ ). Combining such parent nodes also leads to increasing complexity in the offline computation local to that super-node (as we must consider the joint states/decisions at the nodes being merged); however, for sparse network structures, such merged state/decision spaces will still be of relatively small cardinality. Alternatively, there is nothing that prevents one from applying the message-passing algorithm as an approximation within a general directed acyclic network, an idea we illustrate for a simple model in Section IV.

#### IV. EXAMPLES AND EXPERIMENTS

This section summarizes experiments with the offline message-passing algorithm presented in Section III. Throughout, we take the global measurement model to be the spatially-distributed linear binary detectors defined in Example 7 and the network communication model to be the unit-rate selective peer-to-peer transmission scheme with erasure channels defined in Example 6. The overarching decision objective is simultaneously to avoid state-related decision errors at a specified subset of all nodes, which we call the “gateway” nodes (and nodes not in the gateway we call “communication-only” nodes), and to avoid active symbol transmissions on all links of the network. Specifically, we choose the global costs  $c(\hat{x}, x)$  and  $c(u, x)$  so that detection penalty  $J_d$  and communication penalty  $J_c$  in (8) measure precisely the gateway “node-error-rate” and network-wide “link-use-rate,” respectively. Our purpose is to characterize the value and overhead of our message-passing solutions to the decentralized problem formulation of Section II-C, quantifying the tradeoffs relative to the benchmark centralized and myopic solutions discussed in Section II-A. Our procedure is, for a given network, to sample the range of parameter  $\lambda$  in (8), or the weight on communication penalty  $J_c(\underline{\gamma})$  relative to detection penalty  $J_d(\underline{\gamma})$ , in each case applying the message-passing algorithm and recording (i) the achieved online performance  $(J_d^\lambda, J_c^\lambda)$  and (ii) the number of offline iterations  $k^\lambda$  to convergence.

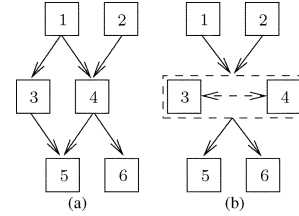


Fig. 3. Example of (a) a non-polytree topology  $\mathcal{G}$  and (b) its equivalent polytree topology for which Proposition 2 is applicable. Specifically, the parents of node 5, namely nodes 3 and 4, have node 1 as a common ancestor so we “merge” nodes 3 and 4. This is done at the (strictly offline) expense of requiring both direct communication between nodes 3 and 4 and increased local computation by nodes 3 and 4, so the message-passing algorithm in Corollary 3 can jointly consider the random variables  $X_3, X_4, U_3$  and  $U_4$ .

Sections IV-B to IV-D present our results across different network topologies, different levels of measurement/channel noise and different prior probability models. These results demonstrate how the decentralized strategy produced by the message-passing algorithm consistently balances between the two different decision penalties. Firstly, in all examples, the optimized strategy degenerates gracefully to the myopic strategy as  $\lambda$  gets large i.e.,  $\lim_{\lambda \rightarrow \infty} (J_d^\lambda, J_c^\lambda) \rightarrow (J_d(\underline{\gamma}), 0)$ . Secondly, in every example that satisfies all conditions of Proposition 2, the optimized strategy achieves a monotonic tradeoff between the two penalties i.e., if  $\lambda_1 < \lambda_2$ , then  $J_d^{\lambda_1} \leq J_d^{\lambda_2}$  and  $J_c^{\lambda_1} \geq J_c^{\lambda_2}$ . The experimental results also illustrate more subtle benefits of the optimized strategy: for example, even when actual symbols can be transmitted without penalty (i.e., when  $\lambda = 0$ ), each node will continue to employ selective silence to resourcefully convey additional information to its children, which remains of value even if erasure probabilities are nonzero.

While these desirable performance tradeoffs are not surprising given the developments in Section III, the experiments in Section IV-C also demonstrate that they are difficult to ensure when the decentralized strategy is selected heuristically (i.e., based on reasonable intuition instead of the proscribed offline iterative procedure). On the other hand, recognizing that per offline iteration each link  $(i, j)$  must reliably compute and communicate messages  $P_{i \rightarrow j}$  and  $C_{j \rightarrow i}$ , each a collection of up to  $|\mathcal{X}_i \times \mathcal{U}_i|$  real numbers, these performance benefits must also be weighed against the number  $k^\lambda$  of offline iterations to convergence. Our consideration of this offline overhead makes, we believe, an important point in understanding the value and feasibility of self-organizing sensor networks, as it allows us to assess the price of adaptive organization, or re-organization. In particular, our analysis emphasizes that for such offline organization to be warranted, it must be that the price of performing it can be amortized over a substantial number of online usages, or equivalently that the network resources consumed for organization represent only a modest fraction of the resources available over the total operational lifetime.

##### A. Basic Experimental Setup

To apply the offline message-passing algorithm developed in Section III, each node  $i$  in the given network  $\mathcal{G}$  requires the following local models: likelihoods  $p(y_i|x_i)$ , channels  $p(z_i|x_i, u_{pa(i)})$ , costs  $c(u_i, \hat{x}_i, x_i)$ , priors  $p(x_{pa(i)}, x_i)$  and

an initial rule  $\gamma_i^0 \in \Gamma_i$ . This subsection describes the parametric forms of the local models that are in common with all experiments to be described in the following subsections. In particular, only the local priors will be different across these experiments, so we now describe all other such local models and leave the description of priors for later subsections.

As mentioned above, the likelihoods  $p(y_i|x_i)$  are those of the linear Gaussian binary detectors defined in Example 7 and the channels  $p(z_i|x_i, u_{pa(i)}) \equiv p(z_i|u_{pa(i)})$  are those of the unit-rate selective peer-to-peer transmission scheme with erasures defined in Example 6. In all experiments to follow, however, we will assume homogeneity, meaning that the signal strength  $r_i \equiv r \in [0, \infty)$  and erasure probability  $q_i \equiv q \in [0, 1]$  are the same for every node  $i$  in  $\mathcal{G}$ .

As also mentioned above, the costs  $c(u_i, \hat{x}_i, u_i)$  local to node  $i$  are defined such that the detection penalty  $J_d$  and communication penalty  $J_c$  in (8) equal the gateway “node-error-rate” and network-wide “link-use-rate,” respectively. Specifically, letting  $\kappa_i = 1$  denote that node  $i$  is in the gateway and  $\kappa_i = 0$  denote otherwise, we choose  $c(u_i, \hat{x}_i, x_i) = \kappa_i c(\hat{x}_i, x_i) + \lambda \sum_{j \in ch(i)} c(u_{i \rightarrow j})$ . Here, the detection-related costs indicate node errors and the communication-related costs indicate link uses i.e.,

$$c(\hat{x}_i, x_i) = \begin{cases} 0, & \hat{x}_i = x_i \\ 1, & \hat{x}_i \neq x_i \end{cases} \text{ and } c(u_{i \rightarrow j}) = \begin{cases} 0, & u_{i \rightarrow j} = 0 \\ 1, & u_{i \rightarrow j} \neq 0. \end{cases}$$

As discussed in Example 6, the event  $U_{i \rightarrow j} = 0$  indicates that node  $i$  suppresses the transmission on the outgoing link to child  $j$ , so it is associated to zero communication cost. Also note that the myopic threshold for each gateway node reduces to  $\eta_i = p_{X_i}(-1)/p_{X_i}(+1)$  in likelihood-ratio space, and to  $\tau_i = \log(\eta_i)/r$  in measurement space (see Example 2).

A final consideration is the initial rule  $\gamma_i^0$  local to node  $i$ . As remarked after Corollary 3 in Section III, initializing to the myopic rule in (5) would prohibit the offline algorithm from making progress. Fig. 4 illustrates our typical choice of initial rule  $\gamma_i^0$ . We partition the real  $y_i$ -axis by a central threshold of  $\tau_i^0 = \log(\eta_i)/r$  to decide  $\hat{x}_i \in \{-1, +1\}$ , followed by  $2(2^{|ch(i)} - 1)$  thresholds evenly spaced within an interval  $r$  about  $\tau_i^0$  to decide  $u_i \in \{-1, 0, +1\}^{|ch(i)|}$ . In essence, each node  $i$  is initialized to (i) ignore all information received on the incoming links, (ii) myopically make a maximum-*a-posteriori* estimate of its local state and (iii) make a binary-valued decision per outgoing link  $(i, j)$ , remaining silent (with  $u_{i \rightarrow j} = 0$ ) when the measurement is near its least-informative values or transmitting its local state estimate (with  $u_{i \rightarrow j} = \hat{x}_i$ ) otherwise. This initial strategy is clearly sub-optimal, achieving the myopic detection penalty but with nonzero communication penalty, yet the message-passing algorithm is observed to make reliable progress from this initialization as long as the induced statistics at every node  $i$  satisfy  $p(u_i|x_i; \gamma_i^0) > 0$  for all  $(x_i, u_i) \in \{-1, +1\} \times \{-1, 0, +1\}^{|ch(i)|}$ .

The initial rule of Fig. 4 belongs to the class of monotone threshold rules for linear-Gaussian binary detectors described in Example 3. This class applies directly to a node without parents and readily extends to a node with parents, in which there can be  $|Z_i|$  such partitions of the likelihood-ratio space  $[0, \infty)$ , or in our experiments one set of such thresholds per symbol value  $z_i \in \{-1, 0, +1\}^{|pa(i)|}$ . The initialization in Fig. 4 makes no use of this received information, but the optimized strategy certainly

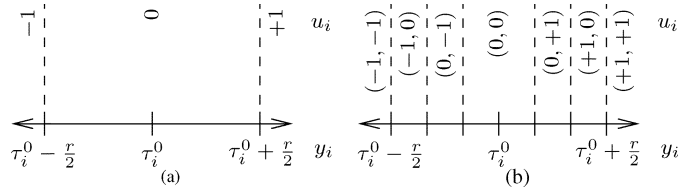


Fig. 4. Initial rule  $\gamma_i^0$  typically used in our experiments with scalar linear binary detectors of signal strength  $r$  and the central threshold (in measurement space) of  $\tau_i^0 = \log(\eta_i)/r$  with  $\eta_i = p_{X_i}(-1)/p_{X_i}(+1)$ . (a) Given node  $i$  has one child; (b) given node  $i$  has two children.

should. This motivates consideration of a more elaborate initial strategy, taking information  $Z_i$  into account. One such initialization, which we experiment with in Section IV-C, is to utilize the neighborhood prior  $p(x_{pa(i)}, x_i)$  and interpret each received symbol  $z_{j \rightarrow i} \in \{-1, +1\}$  as the correct value of state process  $X_j$  i.e., node  $i$  assumes that  $X_{pa(i)} = z_i$ , marginalizing over each state  $X_j$  corresponding to a parent for which  $z_{j \rightarrow i} = 0$ , and accordingly adjusts the central threshold in Fig. 4 to

$$\tau_i^0(z_i) = \frac{1}{r} \log \left( \frac{p_{X_{pa(i)}, X_i}(z_i, -1)}{p_{X_{pa(i)}, X_i}(z_i, +1)} \right).$$

## B. A Small Illustrative Network

This subsection assumes the local models discussed in the preceding subsection and considers the prior probability model  $p(x)$  and the network topology  $\mathcal{G}$  depicted in Fig. 5. Specifically, let state process  $X$  be defined by a probability mass function

$$p(x) \propto \prod_{(i,j) \in \mathcal{E}} \psi(x_i, x_j) \quad (30)$$

where  $\mathcal{E}$  denotes the edge set of the *undirected* graph illustrated in Fig. 5(a) and the non-negative function

$$\psi(x_i, x_j) = \begin{cases} w, & x_i = x_j \\ 1 - w, & x_i \neq x_j \end{cases}$$

captures the correlation (i.e., negative, zero, or positive when  $w \in (0, 1)$  is less than, equal to, or greater than 0.5, respectively) between neighboring binary-valued states  $X_i$  and  $X_j$ . Note that, with just  $n = 12$  nodes, both the normalization implied by (30) and the marginalizations of  $p(x)$  to obtain  $p(x_{pa(i)}, x_i)$  for each node  $i$  can be performed directly. Also observe that, in this example, the links in the network topology are a proper subset of the edges in the (loopy) undirected graph upon which random vector  $X$  is defined.

Fig. 6 displays the tradeoff curves between node-error-rate  $J_d$ , where every node is in the gateway (so that the maximal node error rate is twelve), and link-use-rate  $J_c$  achieved by the message-passing algorithm across different model parameters. We take the nominal case to be  $(w, r) = (0.9, 1.0)$  and the low-correlation case to be  $(w, r) = (0.8, 1.0)$ , considering for each such case three different erasure probabilities, namely  $q = 0.0$ ,  $q = 0.3$  and  $q = 0.6$ . In each of these six parameter settings, we obtain the associated tradeoff curve by sampling  $\lambda$  in increments of  $10^{-4}$ , starting with  $\lambda = 0$ , and declaring convergence in iteration  $k$  when  $J(\gamma^{k-1}) - J(\gamma^k) < 10^{-3}$ . Per instance of parameters  $(w, r)$ , we see that the three curves always start from a common point, corresponding to  $\lambda$  being

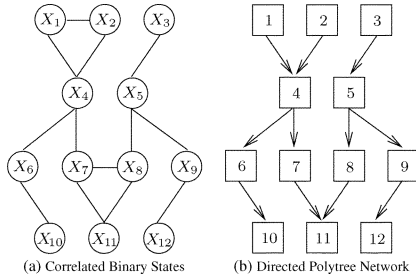


Fig. 5. Small ( $n = 12$ ) decentralized detection network used in the experiments of Section IV-B: (a) the (undirected) graph upon which the spatially-distributed state process  $X$  is defined via (30) and (b) a tree-structured (directed) network topology that spans the vertices in (a).

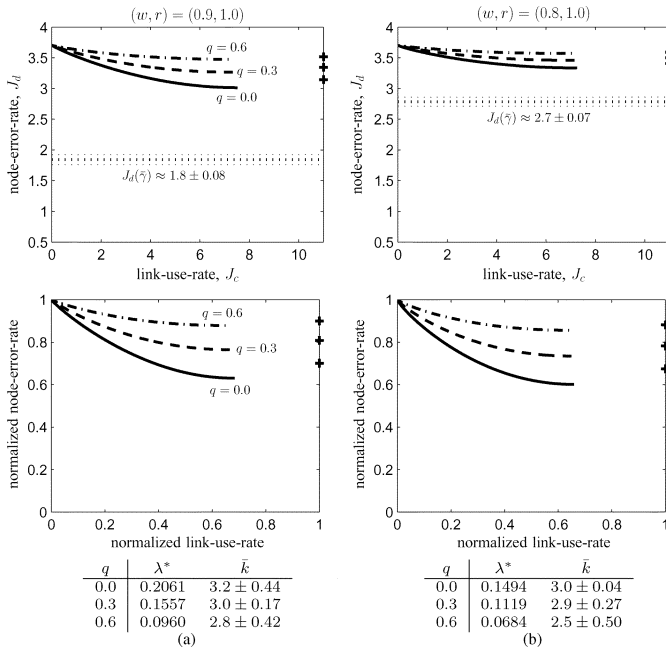


Fig. 6. Optimized tradeoff curves for the network in Fig. 5 given (a) a nominal environment and (b) low state correlation, each such environment with three different link erasure probabilities  $q = 0$  (solid line),  $q = 0.3$  (dashed line) and  $q = 0.6$  (dash-dotted line). The second row of figures uses the same data as the first, normalizing the penalties to better compare across the different model instances. The tables contain the quantities  $\lambda^*$  and  $\bar{k}$  we record while computing each curve. See Section IV-B for more discussion.

large enough so that zero link-use-rate (and thus myopic node-error-rate  $J_d(\bar{\gamma})$ ) is optimal. The smallest value of  $\lambda$  achieving this myopic point, call it  $\lambda^*$ , can be interpreted (for that model instance) as the maximal price (in units of detection penalty) that the optimized network is willing to pay per unit of communication penalty. For  $\lambda$  less than  $\lambda^*$ , we see that the message-passing algorithm smoothly trades off increasing link-use-rate with decreasing node-error-rate. Not surprisingly, this tradeoff is most pronounced when the erasure probability  $q$  is zero, and approaches the myopic detection penalty as  $q$  approaches unity. Also shown per instance of parameters  $(w, r)$  is a Monte-Carlo estimate of the optimal centralized performance  $J_d(\bar{\gamma})$ , computed using 1000 samples from  $p(x, y)$  and simulating the strategy in (2).

The second row of curves displays the same data as in the first row, but after (i) normalizing the achieved link-use-rate by

its capacity (namely eleven unit-rate links) and (ii) expressing the achieved node-error-rate on a unit-scale relative to the benchmark centralized detection penalty and the myopic detection loss  $J_d(\bar{\gamma}) - J_d(\bar{\gamma})$  recovered via offline coordination). The curves show that, subject to less than eleven bits (per global estimate) of online communication, up to 40% of the centralized performance lost by the purely myopic strategy can be recovered. These rescalings also emphasize that the maximum link-use-rates on each optimized curve are well below network capacity and that the message-passing algorithm consistently converges to a strategy that exploits the selective silence: intuitively, each node in the cooperative strategy is able to interpret “no news as providing news.” For further comparison, consider the model with selective communication disabled, meaning each node must always transmit either a  $+1$  or  $-1$  to each of its children and, in turn, link-use-rate is at 100% capacity. Applying the message-passing algorithm to these models yields the points indicated by “+” marks: indeed, we see that selective communication affords up to an additional 10% recovery of detection performance while using only 70% of the online communication capacity.

The tables in Fig. 6 list two key quantities recorded during the generation of each of the six tradeoff curves, namely  $\lambda^*$  and  $\bar{k}$  denoting the lowest value of  $\lambda$  for which the myopic point is the solution and the average number of offline iterations to convergence, respectively. As discussed above, the former can be interpreted as the “fair” per-unit price of online communication: indeed, from the tables, we see that  $\lambda^*$  is inversely related to erasure probability  $q$ , quantifying the diminishing value of active transmission as link reliability degrades. Moreover, comparing  $\lambda^*$  in (a) and (b), we see that lower state correlation similarly diminishes the value of active transmission. The empirical value of  $\bar{k}$  is related to the price of offline self-organization: we see that over all uses of the algorithm in the network of Fig. 5(b), it measures between three and four iterations, implying that maintaining the optimized online tradeoff depends (per offline reorganization) upon the exact computation and reliable communication of 684 to 912 real numbers in total, or roughly 57 to 76 numbers per node.

### C. Large Randomly-Generated Networks

This subsection performs a similar analysis as in Section IV-B, but for a collection of randomly-generated model instances of more realistic size and character. Fig. 7 illustrates a typical output of our model generation procedure: it starts with  $n = 100$  nodes, each randomly positioned within a unit-area square and connected to a randomly selected subset of its spatial neighbors. The vector state process  $X$  is described by a *directed* graphical model, constructed such that the correlation between neighboring states reflects the spatial proximity of the neighbors. Specifically, let  $\bar{p}a(i)$  denote the parents of node  $i$  (in the directed acyclic graph underlying process  $X$ , which we call  $\bar{\mathcal{G}}$ ) and let  $D(i, j)$  denote the spatial distance between node  $i$  and node  $j$ . The global prior probabilities are equal to

$$p(x) = \prod_{i=1}^n p(x_i | x_{\bar{p}a(i)}) \quad (31)$$

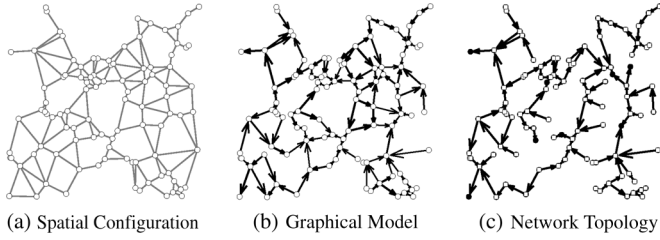


Fig. 7. Typical 100-node detection network generated randomly for the experiments in Section IV-C: (a) the spatial configuration of all nodes in the unit-area square, (b) an arbitrary directed acyclic graph upon which the spatially-distributed state process  $X$  is defined via (31) and (c) the polytree network topology with ten randomly-selected gateway nodes (filled markers).

where, for each parentless node  $i$ , we choose  $p(x_i)$  to be uniform and, otherwise, choose

$$p(x_i|x_{\bar{p}a(i)}) = \begin{cases} 1 - \rho(x_{\bar{p}a(i)}), & x_i = -1 \\ \rho(x_{\bar{p}a(i)}), & x_i = +1, \end{cases}$$

$$\rho(x_{\bar{p}a(i)}) = \frac{\sum_{j \in \bar{p}a(i)} 1_j(x_{\bar{p}a(i)}) D(i, j)^{-1}}{\sum_{j \in \bar{p}a(i)} D(i, j)^{-1}},$$

$$1_j(x_{\bar{p}a(i)}) = \begin{cases} 1, & x_j = +1 \\ 0, & x_j = -1. \end{cases}$$

In words, each  $i$ th factor  $p(x_i|x_{\bar{p}a(i)})$  biases the (conditional) distribution of local state process  $X_i$  to take the same value as the (given) states of its spatially-nearest parent nodes (in  $\bar{\mathcal{G}}$ ).

The next step of the model generation procedure is to construct a polytree network topology  $\mathcal{G}$  and derive the associated priors  $p(x_{pa(i)}, x_i)$  for every node  $i$  required by the offline message-passing algorithm. First, we select ten gateway nodes at random and construct  $\mathcal{G}$  such that (i) these gateway nodes are childless and (ii) its undirected counterpart is an embedded spanning tree of the undirected counterpart to the probability graph  $\bar{\mathcal{G}}$ . This is accomplished using Kruskal's max-weight spanning tree algorithm, where we choose edge weights proportional to the pairwise correlation between the states sharing each edge in  $\bar{\mathcal{G}}$  (and any node pair not sharing an edge in  $\bar{\mathcal{G}}$  is assigned the weight  $-\infty$ ). These pairwise correlations can be computed from the so-called clique marginals  $p(x_{\bar{p}a(i)}, x_i)$  associated with the factorization in (31), which we obtain via Murphy's *Bayesian Network Toolbox in MATLAB* [27]. We further exploit the Markov properties implied by (31) to find the neighborhood marginals (still in  $\bar{\mathcal{G}}$ ) via

$$p(x_{\bar{n}e(i)}, x_i) = p(x_{\bar{p}a(i)}, x_i) \prod_{j \in \bar{c}h(i)} p(x_j|x_i). \quad (32)$$

Finally, because network topology  $\mathcal{G}$  is embedded in the undirected counterpart of  $\bar{\mathcal{G}}$ , we have that  $pa(i) \subseteq \bar{n}e(i)$  for every  $i$  and, in turn, each distribution  $p(x_{pa(i)}, x_i)$  can be found by appropriate marginalization of (32).

Fig. 8 depicts the average-case performance achieved by the message-passing algorithm over 50 randomly-generated model instances. Each plot consists of four clusters of points, three corresponding to the optimized point assuming three different values of  $\lambda$  and one corresponding to the point achieved by the heuristic strategy, essentially interpreting each incoming

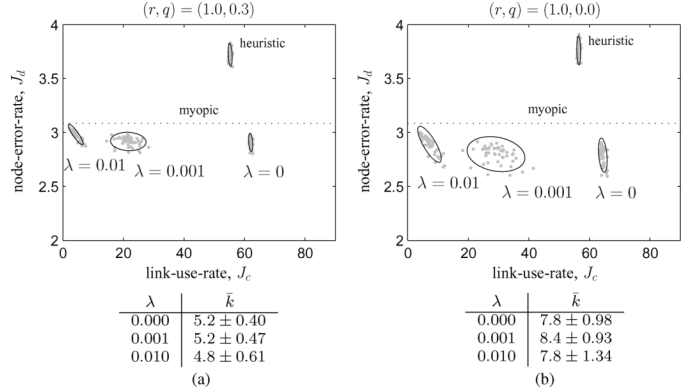


Fig. 8. Performance of different strategies for 50 randomly generated networks of the type shown in Fig. 7 given (a) a nominal environment and (b) zero channel noise. See Section IV-C for more discussion.

symbol as indicating the true value of the parents' local states as described in Section IV-A. In fact, in these experiments, we initialized our algorithm with this heuristic strategy, though it's seen to fail catastrophically in the sense that communication penalty is nonzero and yet the detection penalty is larger than even that of the myopic strategy! This unsatisfactory heuristic performance underscores the value of our offline message-passing algorithm, which via parameter  $\lambda$  consistently decreases global detection penalty (from that of the myopic strategy) as global communication penalty increases despite the poorly performing initializations.

Each table in Fig. 8 lists the average number  $\bar{k}$  of message-passing iterations to convergence, which underscores the price of our offline coordination in the same sense discussed in Section IV-B. We see that roughly eight iterations can be required in the 100-node models, in comparison to roughly three iterations in the twelve-node models of the previous subsection, suggesting the price of offline coordination scales sublinearly with the number of nodes. It is worth noting that the communication overhead associated with each offline iteration also depends on the connectivity of the network topology, each node exchanging a number of messages that scales linearly with its degree (whereas the local computation of these messages scales exponentially with its degree).

#### D. A Small Non-Tree-Structured Network

The preceding experiments focused on models that satisfy all assumptions under which the offline message-passing algorithm is derived. We now discuss experiments for a model in which the network topology is *not* a polytree. In such cases the local fixed-point equations in Corollary 3 are no longer guaranteed to be equivalent to the general fixed-point equations in Corollary 1. In turn, the message-passing algorithm no longer necessarily inherits the general convergence and correctness guarantees discussed for Corollary 1. As remarked in Section III, the team-optimal solution can be computed by aggregating nodes in the original topology so as to form a polytree to which our message-passing algorithm can be applied. Of course, this approach implicitly requires communication among nodes that have been aggregated but are not neighbors in the original topology; moreover, it is computationally tractable only if a small number of nodes need to be aggregated.

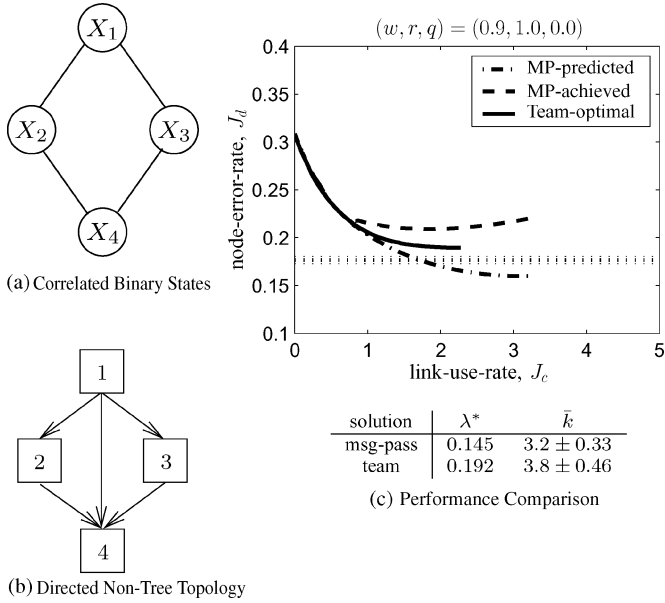


Fig. 9. Small ( $n = 4$ ) detection network discussed in Section IV-D: (a) the (undirected) graph upon which the spatially-distributed state process  $X$  is defined via (30); (b) the (directed) network topology that spans the vertices in (a); and (c) the performance comparison between the team-optimal solution and the message-passing approximation. In (c), three tradeoff curves are shown, dashed being that achieved by the message-passing approximation, solid being that achieved by the team-optimal solution, and dash-dotted being that predicted by the message-passing approximation (but based on incorrect assumptions).

For the above reasons, in non-polytree networks it is useful to understand what the truly team-optimal solution can achieve in comparison to what is achieved if one rather applies the local message-passing algorithm. In this section, we present and discuss experiments on a small example in order to explore these questions. Even in such small models, the team-optimal solution produces rather sophisticated signaling strategies, exploiting the non-tree network structure in ways that are not available in the message-passing approximation. Nonetheless, at least in this small example, the local message-passing algorithm still achieves competitive decision performance.

Let us consider a model of the same type as in Section IV-B, except involving only four nodes in the non-tree configuration depicted in Fig. 9. For illustration, we fix  $r = 1$ ,  $q = 0$ , and  $w = 0.9$ , so that all measurements have the same noise, all channels have zero erasure probability and the states are positively-correlated. Assume node 4 is the lone gateway node, while nodes 1, 2 and 3 are communication-only nodes. The team objective boils down to having the communication-only nodes collectively generate the “most-informative-yet-resourceful” signal to support the gateway node’s final decision. Indeed, we anticipate node 1 to play the dominant role in any such signaling strategy, given its direct link to every other node in the communication network topology of Fig. 9(b). Note, in particular, that this communication topology includes a direct path from node 1 to node 4 which is not present in the probability graph of Fig. 9(a). Thus, this example also allows us to illustrate the value of longer-distance messaging than would be found, for example, if loopy belief propagation [12], [13] were applied to this problem.

Fig. 9 also displays the tradeoff between node-error-rate  $J_d$  and link-use-rate  $J_c$  achieved by both the team-optimal solu-

tion and the message-passing approximation. There are in fact two performance curves associated with the message-passing approximation, namely that which it actually achieves and that which it *predicts* to achieve. Specifically, this prediction corresponds to the performance  $J(\gamma^k)$  computed via (29) as a function of the forward messages  $P_i^{k+1}(z_i|x_i)$ , which can be incorrect because each such message is no longer necessarily equivalent to the distribution  $p(z_i|x_i; \gamma^k)$ . To compute the actual performance here, we rather determine the full distribution  $p(u, \hat{x}, x; \gamma^k)$  via (10) and evaluate the expected cost directly. As in earlier experiments, each curve is obtained by sampling  $\lambda$  in increments of  $10^{-3}$ , starting with  $\lambda = 0$ , and declaring convergence in iteration  $k$  when  $J(\gamma^{k-1}) - J(\gamma^k) < 10^{-3}$  (but using the sequence of predicted performances in the message-passing approximation). Also shown is the empirical estimate (plus or minus one standard deviation based on 10000 samples) of the centralized performance.

Notice first in Fig. 9(c) that the message-passing and team-optimal curves coincide at very low link-use-rates, a regime in which enough links remain unused so that the network topology is effectively tree-structured. For higher link-use-rates, we see that the message-passing prediction is consistently over-optimistic, eventually even suggesting that the achieved node-error-rate surpasses the optimal centralized performance in the actual network; meanwhile, the actual performance achieved by the message-passing approximation is consistently inferior to that of the team-optimal solution, yet for this simple model still a reliable improvement relative to myopic detection performance. Also notice how the message-passing approximation does not produce a monotonic tradeoff curve, in the sense that it permits link-use-rates to increase beyond the range over which the node-error-rate remains non-increasing. The team-optimal solution is, of course, monotonic in this sense, with peak link-use-rate well below that determined by the message-passing approximation. Finally, the table in Fig. 9(c) shows that the team-optimal solution is (i) more resourceful with its link usage, as quantified by  $\lambda^*$ , and (ii) takes on-average more iterations to converge, as quantified by  $\bar{k}$ . The latter is arguably surprising, considering it is the message-passing approximation that comes without any theoretical guarantee of convergence. Indeed, these particular experiments did not encounter a problem instance in which the message-passing algorithm failed to converge.

We conjecture that algorithm convergence failures will be experienced when the message-passing approximation is applied to more elaborate non-tree-structured models. To help justify this point, Fig. 10 depicts the key discrepancy between the team-optimal solution and the message-passing approximation. As each node performs each of its local message-passing iterations, it neglects the possibility that any two parents could have a common ancestor (or, equivalently, that any two children could have a common descendant), implicitly introducing fictitious replications of any such neighbors and essentially “double-counting” their influence. This replication is reminiscent of the replications seen in the so-called *computation tree* interpretation of loopy belief propagation [28]. However, there are important differences in our case, as this replication is both in upstream nodes that provide information to a specific node *and* in downstream nodes whose decision costs must be propagated back to the node in question. Moreover, the nature of these replications

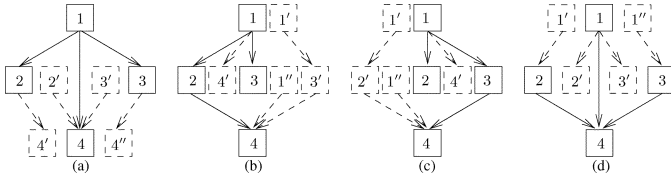


Fig. 10. Tree-based message-passing approximation from the perspective of each node in the non-tree network of Fig. 9. Nodes and links drawn with dashed lines represent the fictitious nodes introduced by the approximation, which neglects the possibility that any two parents could have a common ancestor (or, equivalently, any two children could have a common descendant). (a) At node 1; (b) at node 2; (c) at node 3; (d) at node 4.

is itself node-dependent, meaning each offline iteration may be cycling over  $n$  different assumptions about the global network structure. The potential for these different perspectives to give rise to erroneous message iterates lies at the heart of the possibility for convergence difficulties in more elaborate models.

Though we observed no convergence difficulty, the nature of the approximation illustrated in Fig. 10 still manifests itself in a performance difference between the solutions compared in Fig. 9(c), most apparent for small values of  $\lambda$ . We have more carefully inspected the different strategies and, while both yield signaling in which node 1 takes a leadership role, the team-optimal strategy consistently uses nodes 2 and 3 in a more resourceful way, ultimately allowing gateway node 4 to receive better side information for its final decision. For instance, in the team-optimal solution, node 1 typically signals exclusively to node 4 or exclusively to node 3, and only for the most discriminative local measurement will it signal to both nodes 2 and node 4; that is, node 1 never signals all three other nodes and, in turn, the signaling rules used by nodes 2 and 3 are asymmetric. In the message-passing approximation, however, node 1 typically uses either none or all of its links, in the latter case transmitting the same symbol to all other nodes; in turn, nodes 2 and 3 employ identical signaling rules to node 4 in which, given node 1 has communicated, the presence or absence of signal indicates agreement or disagreement, respectively, with the symbol broadcasted by node 1. The message-passing approximation cannot recognize the value of introducing asymmetry and, consequently, determines that a larger networkwide link-use-rate is necessary to achieve a comparable gateway node-error-rate.

## V. CONCLUSION

A key challenge in modern sensor networks concerns the inherent design tradeoffs between application-layer decision performance and network-layer resource efficiency. In this paper we explored such tradeoffs for the decision-making objective of optimal Bayesian detection, assuming in-network processing constraints are dominated by a low-rate unreliable communication medium. Mitigating performance loss in the presence of such constraints demands an offline “self-organization” algorithm by which the processing rules local to all nodes are iteratively coupled in a manner driven by global problem statistics. We showed that, assuming (i) online measurement processing is constrained to a single forward sweep in a finite-rate, sparsely-connected polytree network, (ii) the measurement/channel noise processes are spatially-independent and (iii) the global decision

criterion decomposes additively across the nodes, this offline computation admits interpretation as an efficient forward-backward message-passing algorithm. Each forward sweep propagates likelihood messages, encoding what online communication along each link means from the transmitter’s perspective, while each backward sweep propagates cost-to-go messages, encoding what online communication along each link means from the receiver’s perspective. In each iteration, both types of incoming messages influence how each node updates its local rule parameters before it engages in the next iteration.

The key steps by which we obtain these results can be traced to a collection of earlier works in the abundant decentralized (team) detection literature. As was discussed, however, each of these earlier works considered only a special case of the model considered here, typically employing a proof technique not immediately applicable to our more general case. For example, our results hold for noisy channel models that include a dependence on the local hidden state (e.g., for detecting the presence or absence of a jamming signal) or the composite transmissions of all parent nodes (e.g., for modeling the effects of multipoint-to-point interference). Our results also shed new light on the extent to which the graphical structure underlying the spatially-distributed hidden state process may deviate from the communication network topology without sacrificing either algorithm correctness or efficiency.

Experiments with the efficient message-passing algorithm underscored how: (i) the algorithm can produce very resourceful cooperative processing strategies in which each node becomes capable of using the absence of communication as an additional informative signal; (ii) design decisions to reduce online resource overhead by imposing explicit in-network processing constraints must be balanced with the offline resource expenditure to optimize performance subject to such constraints; and (iii) the message-passing algorithm can be applied to models that do not necessarily satisfy all of the assumptions under which it is originally derived. Our empirical success with the latter idea motivates a number of new research questions, including (i) whether the additional model assumptions introduced in going from the general algorithm in Corollary 1 to the message-passing algorithm in Corollary 3 allow for stronger convergence guarantees than those presented here; and (ii) whether there exist bounds on the performance loss resulting from applying the message-passing algorithm to more general topologies.

There are other open questions related to application of the message-passing algorithm beyond the scope of its derivation. One such question is the degree to which quantization errors in the offline messages can be tolerated. Another line of questioning concerns the adverse effects of mismatches between the local models assumed at any particular node from the true ones. Related questions are how much is lost when not all noise processes are spatially independent or when the cost function does not decompose additively across the nodes. Better understanding of such robustness properties is the first step towards addressing the difficult problem of when a detection network should reorganize i.e., when the network topology or the local models have changed enough to merit re-optimization. Other avenues for research are whether offline message-passing solutions exist for decentralized decision problems involving



- (i) more elaborate online communication architectures (e.g., [29], [30]) than just the single forward sweep analyzed here or  
 (ii) continuous-valued state processes.

#### ACKNOWLEDGMENT

The authors thank J. N. Tsitsiklis for helpful discussions.

#### REFERENCES

- [1] C.-Y. Chong and S. P. Kumar, "Sensor networks: Evolution, opportunities, and challenges," *Proc. IEEE*, vol. 91, no. 8, pp. 1247–1256, Aug. 2003.
- [2] A. M. Sayeed *et al.*, "Self-organizing distributed collaborative sensor networks," *IEEE J. SAC*, vol. 23, no. 4, pp. 689–692, Apr. 2005.
- [3] R. R. Tenney and N. R. Sandell, Jr., "Detection with distributed sensors," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-17, no. 4, pp. 501–510, Sep. 1981.
- [4] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors: Part I," *Proc. IEEE*, vol. 85, no. 1, pp. 54–63, Jan. 1997.
- [5] J. N. Tsitsiklis, "Decentralized detection," in *Advances in Statistical Signal Processing*. Stamford, CT: JAI Press, 1993, vol. 2, pp. 297–344.
- [6] Z. Tang, K. Pattipati, and D. Kleinman, "Optimization of detection networks: Part II," *IEEE Trans. SMC*, vol. 23, no. 1, pp. 211–221, Jan./Feb. 1993.
- [7] A. Pete, K. R. Pattipati, and D. L. Kleinman, "Optimization of decision networks in structured task environments," *IEEE Trans. SMC*, vol. 26, no. 6, pp. 739–748, Nov. 1996.
- [8] J. D. Papastavrou and M. Athans, "A distributed hypothesis-testing team decision problem with communications cost," in *Proc. IEEE CDC*, 1986, pp. 219–225.
- [9] C. Rago, P. Willett, and Y. Bar-Shalom, "Censoring sensors: A low-communication-rate scheme for distributed detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 32, no. 2, pp. 554–568, Apr. 1996.
- [10] B. Chen *et al.*, "Channel aware decision fusion in wireless sensor networks," *IEEE Trans. Signal Processing*, vol. 52, no. 12, pp. 3454–3458, Dec. 2004.
- [11] Z.-B. Tang, K. R. Pattipati, and D. L. Kleinman, "An algorithm for determining the decision thresholds in a distributed detection problem," *IEEE Trans. SMC*, vol. 21, no. 1, pp. 231–237, Jan./Feb. 1991.
- [12] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. San Francisco, CA: Morgan Kaufmann, 1988.
- [13] A. S. Willsky, "Multiresolution markov models for signal and image processing," *Proc. IEEE*, vol. 90, no. 8, pp. 1396–1458, Aug. 2002.
- [14] J. N. Tsitsiklis, "Decentralized detection by a large number of sensors," *Math. Control, Signals Syst.*, vol. 1, pp. 167–182, 1988.
- [15] J. D. Papastavrou and M. Athans, "Distributed detection by a large team of sensors in tandem," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 28, no. 3, pp. 639–653, Jul. 1992.
- [16] J.-F. Chamberland and V. V. Veeravalli, "Decentralized detection in sensor networks," *IEEE Trans. Signal Processing*, vol. 51, no. 2, pp. 407–416, Feb. 2003.
- [17] J.-J. Xiao and Z.-Q. Luo, "Universal decentralized detection in a bandwidth constrained sensor network," *IEEE Trans. Signal Processing*, vol. 53, no. 8, pp. 2617–2624, Aug. 2005.
- [18] S. A. Aldosari and J. M. F. Moura, "Detection in decentralized sensor networks," in *Proc. IEEE ICASSP*, 2004, vol. 2, pp. 277–280.
- [19] W.-P. Tay, J. N. Tsitsiklis, and M. Z. Win, "Data fusion trees for detection: Does architecture matter?," *IEEE Trans. Inform. Theory*, vol. 54, no. 9, pp. 4155–4168, Sep. 2008.
- [20] R. Cogill and S. Lall, "Decentralized stochastic decision problems and polynomial approximation," in *Proc. Amer. Control Conf. (ACC)*, 2005, pp. 2709–2714.
- [21] V. Saligrama, M. Alanyali, and O. Savas, "Distributed detection with packet losses and finite capacity links," *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4118–4132, Nov. 2006.
- [22] A. Dogandžić and B. Zhang, "Distributed estimation and detection for sensor networks using hidden markov random fields," *IEEE Trans. Signal Processing*, vol. 54, no. 8, pp. 3200–3215, Aug. 2006.
- [23] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*. New York: Wiley, 1968, vol. 1.
- [24] O. P. Kreidl, "Graphical Models and Message-Passing Algorithms for Network-Constrained Decision Problems," Ph.D. dissertation, EECS Dept., MIT, Cambridge, MA, 2008.
- [25] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA: Athena Scientific, 1995.
- [26] R. Cogill and S. Lall, "An approximation algorithm for the discrete team decision problem," *SIAM J. Control Optim.*, vol. 45, no. 4, pp. 1359–1368, 2005.
- [27] K. P. Murphy, "The bayes net toolbox for Matlab," *Comput. Sci. Stat.*, vol. 33, 2001 [Online]. Available: <http://people.cs.ubc.ca/~murphyk/Papers/bnt.pdf>
- [28] S. Tatikonda and M. I. Jordan, "Loopy belief propagation and gibbs measures," in *Proc. UAI*, 2002, pp. 493–500.
- [29] S. Alhakeem and P. K. Varshney, "A unified approach to the design of decentralized detection systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 31, no. 1, pp. 9–20, Jan. 1995.
- [30] P. F. Swaszek and P. Willett, "Parley as an approach to distributed detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 31, no. 1, pp. 447–457, Jan. 1995.



**O. Patrick Kreidl** received the B.S. degree from George Mason University (GMU), Fairfax, VA, in 1994 and the S.M. and Ph.D. degrees from the Massachusetts Institute of Technology (MIT), Cambridge, in 1996 and 2008, respectively, all in electrical engineering.

He has been a Research Affiliate at MIT's Laboratory for Information and Decision Systems since 2008, and a Research Engineer at BAE Systems Advanced Information Technologies (formerly Alphatech, Inc.) since 1996. He also held an adjunct faculty appointment within GMU's Electrical and Computer Engineering Department from 1998 to 2001. His research interests are in statistical signal processing, stochastic control, and numerical optimization.



**Alan S. Willsky** (S'70–M'73–SM'82–F'86) joined the Massachusetts Institute of Technology (MIT), Cambridge, in 1973 and is the Edwin Sibley Webster Professor of Electrical Engineering and Co-Director of the Laboratory for Information and Decision Systems. He was a Founder of Alphatech, Inc. and Chief Scientific Consultant (a role in which he continues at BAE Systems Advanced Information Technologies). From 1998 to 2002, he served on the U.S. Air Force Scientific Advisory Board. He has delivered numerous keynote addresses and is co-author of the text *Signals and Systems*. His research interests are in the development and application of advanced methods of estimation, machine learning, and statistical signal & image processing.

Dr. Willsky received the 1975 American Automatic Control Council Donald P. Eckman Award, the 1979 ASCE Alfred Noble Prize, the 1980 IEEE Browder J. Thompson Memorial Award, the IEEE Control Systems Society Distinguished Member Award in 1988, the 2004 IEEE Donald G. Fink Prize Paper Award, and the Doctorat Honoris Causa from Université de Rennes in 2005.