Data Fusion and Mis-information Removal In Social Networks

Vikram Krishnamurthy and Maziyar Hamdi
Department of Electrical and Computer Engineering
University of British Columbia
Vancouver, Canada
Email: {vikramk, maziyarh}@ece.ubc.ca

Abstract— A key issue in state estimation presented in a multi-agent social network scenario is the inadvertent multiple re-use of data also known as mis-information propagation. We formulate this mis-information propagation in a graph theoretic setting and give a necessary and sufficient conditions on the adjacency matrix so that the underlying state can be estimated. A sub-optimal algorithm is also presented when the information flow graph is not known.

Index Terms—Social Network, mis-information propagation, Bayesian estimation, data incest.

I. INTRODUCTION

In social networks, each group of individuals use received information from the other groups to evaluate a belief about an economic or a social parameter such as quality of a restaurant or a political party. Suppose that there are \( S \) groups in a social network. Let \( x \in \mathbb{R}^d \) represent a state of nature (economic or social parameter) that individuals in the social network aim to estimate. Assume \( x \) has prior distribution \( \pi_0 \). To estimate \( x \), each group \( s \in S \) in the social network obtains a \( M \)-dimensional measurement vector at each time \( k = 1, 2, \ldots \)

\[
z_{[s,k]} = H_{[s,k]} x + v_{[s,k]} \quad s \in \{1, \ldots, S\}, \quad k = 1, 2, \ldots
\]  

\( 1 \)

Here \( H_{[s,k]} \) is observation matrix and \( v_{[s,k]} \) denotes the observation noise. Assume that the \( z_{[s,k]} \) given \( x \) are independent random variables with respect to type \( s \) and time \( k \). Based on their local observation \( z_{[s,k]} \), each group \( s \) combines this measurement with information received from other groups in social network to compute an estimate of \( x \). They then communicate this estimate to other groups in the social network. An important parameter that characterizes the information exchange is the delay. Individuals take different amounts of time to form beliefs and communicate them. The most important consequence of this delay is mis-information (or rumor) propagation as we will explain shortly. Let us first formulate the information exchange and associated delays using graph-theoretic notation.

Information Exchange Protocol: Let \( G_k = (V_k, E_k) \), \( k = 1, 2, \ldots \) denote a sequence of time-dependent directed graph of information flow in the social network until and including time \( k \). Here \( V_k \), the set of vertices,

\[
V_k = \{(s,t) | i \leq k, s \in \{1, 2, \ldots, S\}\}
\]  

\( 2 \)

and \( E_k \subseteq V_k \times V_k \) is the set of edges which depicts the connections between vertices in \( G_k \). For example if \( (s,t), (s',t') \) \( \in E_k \), it means that the information from group (node) \( s \) at time \( i \) is available at group \( s' \) at time \( i' \) \( (i \leq i' \leq k) \). Typically \( E_k \) depicts the random delays.

The aim is to estimate the underlying state of nature \( x \). This estimation problem can be expressed in the following abstract form:

\[
\begin{align*}
G_k &= (V_k, E_k) \quad k = 1, 2, \ldots \\
z_{[s,k]} &= H_{[s,k]} x + v_{[s,k]}, \quad \text{(observation process)} \quad (3) \\
\theta_{[s,k]} &= \mathcal{A}(\Theta_{\mathcal{A}}(G_k), z_{[s,k]}), \quad \text{(information exchange protocol)}
\end{align*}
\]

\( \mathcal{A} \) denotes the column corresponding to vertex \( (s,k) \) of the adjacency matrix of the information flow graph \( G_k \).

Here \( \mathcal{A}(G_k) \) denotes the column corresponding to vertex \( (s,k) \) of the adjacency matrix of the information flow graph \( G_k \).

Therefore, \( \Theta_{\mathcal{A}}(G_k) \) in (3) denotes the set of beliefs from other nodes at previous times available at node \( s \) at time \( k \) which depends on the information flow network \( G_k \). \( \theta_{[s,k]} \) denotes the sufficient statistic of the estimate for \( x \) given \( \Theta_{\mathcal{A}}(G_k) \) and \( z_{[s,k]} \). \( \mathcal{A} \) denotes the algorithm used by each node to compute the estimate \( \theta_{[s,k]} \).

Mis-information Propagation: The aim of this paper is to construct the information exchange algorithm \( \mathcal{A} \) in (3) so that it yields the filtered estimate of \( x \) given the most up-to-date information. If \( \mathcal{A} \) is not constructed properly, then rumor or mis-information can propagate in the social network. For example, assume that the estimates of group 1 at time 1 reach group 2 at time 2. Also suppose the estimates from social group reach social group 1 at time 3. Since group 2 used the estimate from group 1 at time 1, the estimate generated by group 2 is a function of the estimate of group 1 at time 1. Therefore, if group 1 naively combines the estimate of group 2 received at time 3, with its own local estimates, it would have double counted its estimate at time 1. In the above graph theoretic notation, we can depict graph \( G_3 \) as

\[
\begin{align*}
(1,1) &\rightarrow (1,2) \rightarrow (1,3) \\
(2,1) &\rightarrow (2,2) \rightarrow (2,3)
\end{align*}
\]  

\( 4 \)

where the two-tuples denote vertices defined in (2) and the arrows denote edges. The fact that are two distinct paths between (1,1) and (1,3) in the graph of (4) shows that information in (1,1) is double counted leading to mis-information propagation can be viewed as the destructive re-use of measurement information. It leads to an overconfidence phenomena i.e. the variance is under-estimated. So we need to design the algorithm \( \mathcal{A} \) to cope with this mis-information.
**Benchmark Protocol:** To obtain a benchmark for how to manage mis-information, consider the following idealized scenario. Instead of transmitting sufficient statistic \( \theta_{s,k} \), each group transmits its own local measurements \( z_{s,k} \) and all raw measurement received over the network. Thus each node has the entire available observation history up to time \( k \) based on all \( S \) nodes, which we will denote as \( Z^{r}(G_{k}) \). Therefore, the estimation of each node is free of mis-information. \( r^{f}(G_{k}) \) is the column of transitive closure matrix \( 1^{\circ} \) corresponding to vertex \((s,k)\) which shows all bundles \((i,t)\) in \( V_{k} \) that there is a path between \((i,t)\) and \((s,k)\) in \( G_{k} \). The aim is to estimate the underlying state if the nature \( x \) with prior \( \pi_{0} \). The problem is a simple filtering problem that can be summarized as follows:

\[
\begin{align*}
G_{k} = (V_{k}, E_{k}) \quad & \text{is given.} \\
\hat{z}_{s,k} &= H_{s,k} \times y_{s,k} + v_{s,k}, \quad \text{(observation process)} \\
\hat{y}_{s,k} &= \mathcal{F}(Z^{r}(G_{k}), z_{s,k}), \quad \text{(ordinary filtering problem)}
\end{align*}
\]

The estimation \( \hat{y}_{s,k} \) is free of mis-information because at node \( s \) uses the all available raw measurement (and not estimations) form other nodes at previous times. The recursive nature of Bayesian estimation in decentralized fusion requires careful design to cope with the possible re-use of information such that the estimate \( \hat{\theta}_{s,k} \) is equal to mis-information free estimate \( \hat{y}_{s,k} \) in the optimal scenario. In more realistic problems considered in this paper, there are multiple nodes together with random delays in the network. For such cases, mis-information management is a non-trivial problem.

**Main Results:** With the above models, we are now ready to outline the main results of this paper: In this paper, we address the following questions:

1) **Existence Problem:** Under what constraints, is the algorithm \( \mathcal{A} \) able to completely remove the mis-information?

2) **Design Problem:** How can we design algorithm \( \mathcal{A} \) such that \( \hat{\theta}_{s,k} = \hat{y}_{s,k} \) i.e., the mis-information propagation is prevented.

3) **Reconstruction Problem:** If the information flow graph, \( G_{k} \), is not completely known at each time, how can we design an algorithm to mitigate the mis-information propagation?

**Related Work:** For the motivation of the mis-information problem, we refer to [1], [2], [3], [4], [5] in sensor networks – where the term ‘data incest’ is used. The key requirement is to fuse estimates sharing a common information set. A first approach consists of formulating this problem as fusion with unknown correlation. A consistent estimate can be derived using the Covariance Intersection developed by Julier and Uhlmann [1]. However, this algorithm based on convex combination of information is by nature sub-optimal. Also the optimal estimates free of mis-information can be derived for some particular network topologies. An optimal solution can be computed for the case of connected tree networks by combining a Decentralized Information Filter and a Channel Filter [6]. In this paper, we consider mis-information propagation through a social network with arbitrary network topologies. Each node in this social network records their observation of a parameter with any arbitrary noise. We used a combination of graph theory and Bayesian estimation to remove the mis-information removal generated by different delays in links. Moreover, the paper derived a sufficient condition on the topology of the network which guarantees that mis-information problem can be solved. Chong et al introduced graph information in [7] which is a directed graph to identify common information.

The rest of the paper is organized as follows. We model a social network with a type-based graph and show in Section II that Directed Acyclic graphs are the key tools to understand the topology of the network and consequently to derive optimal estimates free of mis-information propagation.

In Sec.III, we present a necessary and sufficient condition (called Property 1) on the topology of the network that guarantees mis-information free optimal estimates. We showed that with full knowledge of information flow graph, Property 1 leads to an algorithm for exact mis-information removal using optimal Bayesian estimation defined in Sec. III. An algorithm for mis-information removal when the information flow graph is not completely known at each time is proposed in Sec. IV. Numerical results that show the effect of mis-information propagation and also the excellent performance of the proposed mis-information removal algorithm is presented in Sec. V. Finally Sec. VI concludes the paper and presents directions for future work.

**II. INFORMATION FLOW IN SOCIAL NETWORKS**

In this section we discuss the spread of the information through a social network. In this paper we used the Directed Acyclic graphs to model the information flow in social network. To introduce the model, we first provide some graph theoretical definitions and then later in this section we discuss the information flow network in more details.

**Definition 1:** A graph \( G_{N} \) comprising of \( N \) nodes is a pair \((V,E)\), where \( V = \{v_{1},\ldots,v_{N}\} \) is a set of nodes (also called vertices), and \( E \subset V \times V \) is a set of edges between the nodes. \( G_{N} \) is a directed graph if for any \((v_{i},v_{j}) \in E \) then \((v_{j},v_{i}) \notin E \). A **path** is an alternating sequence of nodes and edges, beginning and ending with an edge, in which each node is incident to the two edges that precede and follow it in sequence. A **Directed Acyclic Graph (DAG)** is a directed graph with no path that starts and ends at the same node.

**Definition 2:** Let \( G_{N} = (V,E) \) denote a graph with \( N \) nodes \( V = \{v_{1},\ldots,v_{N}\} \).

The Transitive Closure Matrix \( T \) of \( G_{N} \) is an \( N \times N \) matrix whose elements \( T(i,j) \) are given by \( T(i,i) = 1 \), and

\[
T(i,j) = \begin{cases} 
1 & \text{if there is a path between } v_{i} \text{ and } v_{j}, \\
0 & \text{otherwise}
\end{cases}
\]
**Definition 3 (Family of DAGs):** A family of DAGs $\mathcal{G}_N$ is defined as a set of DAGs $\{G_1, \ldots, G_N\}$ where $G_n$ is the sub-graph of $G_{n+1}$ such that
\[
\begin{align*}
V_n &= V_{n+1}/V_{n+1}^0, \\
E_n &= E_{n+1}/\{(v_i, v_{n+1}) \in E_{n+1}|v_i \in V_{n+1}\}.
\end{align*}
\]

**Notation:** For each graph $G_n \in \mathcal{G}_N$, let the $n \times n$ matrices $A_n$ and $T_n$, respectively, denote the Adjacency Matrix and Transitive Closure Matrix. For future reference, define the following:
- $t_n \in \{0, 1\}^{1 \times (n-1)}$: transpose of first $n-1$ elements of $n$th column of $T_n$.
- $a_n \in \{0, 1\}^{1 \times (n-1)}$: transpose of first $n-1$ elements of $n$th column of $A_n$.

### A. Modeling a social network with a type-based graph

In this section, we map a typical social network to a graph we named type-based representative of a social network. In the type-based graph, as described in Sec. I, each node represents a component of individuals in the social network with same characteristics. Let $H$ denote the graph corresponding to the underlying social network. Recall that in this graph each vertex represents an individual in the social network and each edge depicts a relation (such as friendship) between two vertices. The graph $H$ can be divided into $S$ cluster of nodes of the same type that share the same characteristics about the social or economic parameter under investigation. Each cluster, $H_i$, is a sub-graph of social network which includes all nodes with type $i$. These components should satisfy the following:

- $H = \bigcup_{i,j=1}^S H_i$
- $H_i \cap H_j = \emptyset$ if $i \neq j$

We represent each component in the graph of social network with a node in the type-based graph. In this paper we assume that if two individuals are connected to each other, then the beliefs (opinions about the investigated social parameter) of one individual becomes available at the other node after a time delay. This models a social networks where people are interacting with others and sharing their opinions. It is obvious that if the total number of connections between two clusters is higher, the delay of the link connecting those two clusters is low and if two clusters are isolated (or slightly connected) the delay time is much higher. Because of the statistical characteristics of the evolution of social networks, the total number of edges between two clusters and as a result the delay of the corresponding link in the type-based graph is random.

### B. Decentralized Data Fusion Model and Communication Protocols

Up to now, we described how we can model a social network with a type based graph. Here, we describe the information flow among different types in the social network. To formulate the mis-information problem informally described in Sec. I, consider two types of information flow:

1. **Full information flow network.** This is an idealization that assumes that nodes broadcast their local raw measurements over the network. In this protocol each node records its observation and then combine it with the received measurements of other nodes to update its estimation and then transfer its own raw measurement over the network. The full information flow network serves as a theoretical benchmark against which we can compare a second type of information flow that characterizes the real social networks. In this type of information flow, as a result, a node cannot double count measurements from other nodes and therefore, mis-information cannot propagate.

2. **Constrained information flow network.** This refers to the actual social network where as described in Sec. I, only sufficient statistics of aposteriori distribution are broadcasted. It is in such a constrained information flow network that mis-information propagation arises.

Since full information protocol serves as an idealized benchmark, its efficiency is irrelevant. However, it can be made more efficient by requiring each node $s$ to broadcast only observations which have not been already integrated in the global estimates computed by the other nodes in the network. However, node $k$ will have at least to transmit is current observation $z_{[s,k]}$. As described earlier this is not suitable for human-related social networks.

### III. Optimal Mis-information Propagation Removal Algorithm

Recall from Sec. I, we assumed that each node measures a noisy version of $x \in \mathbb{R}^d$ with the apriori distribution is $x \sim \pi_0(\cdot)$ and $z_{[s,k]}$ denote the $M$-dimensional measurement vector observed by node of type $s$, at time instant $k$ as defined in (1). For notational simplicity, instead of $[s,k]$, the following scalar index $n$ is used:

\[
n \triangleq [s,k] \triangleq s+S(k-1), \quad s \in \{1, \ldots, S\}, \quad k \in \{1, \ldots, K\}.
\]

Notice $n$ is a composite of time $k$ and node with type $s$. Subsequently, we will refer to $n$ as a “node” of a time dependent graph namely information flow graph. Consequently, following (8), the observation equation (1) at node $n$ can be rewritten as

\[
z_n = H_n x + v_n, \quad \text{for } n = [s,k]
\]
measurements is ignored easily and thus all the estimates are clear, by transmitting the raw measurements, any repetition of investigation this protocol is that there is no room for mis-information. However, the reason that we nodes. This information flow protocol is not possible in the human-related social networks. Recall the optimal full information flow scenario defined in Sec. I, there might be some abusive structure removal of the mis-information. Lets get back to the full information network can be presented as finding an algorithm for $\Theta_{n+1}$ as a wise combination of $\Theta_{n}(G_n)$ and $z_n$ such that it assures

$$p(x|\Theta_{n}(G_n), z_n) = p(x|Z_{n+1}(G_n), z_n).$$

(12)

The following lemma gives a mathematical proof for existence of such optimal algorithm.

**Lemma 1**: Consider a probability space $\Omega$, probability measure $\mu$, and $\sigma$-algebras $\mathcal{F}$, $\mathcal{H} \subseteq 2^{\Omega}$ then

$$\forall x \in \Omega, \quad p(x|\mathcal{F}) = p(x|\mathcal{H}) \iff \mathcal{F} = \mathcal{H}$$

(13)

**Proof**: Proof can be found in any theory of probability book.

The Lemma 1 assures that if the sigma algebra generated by $\Theta_{n}(G_n)$ is equal to the sigma algebra generated by $Z_{n+1}(G_n)$ then the mis-information removal problem has solution. It also iterates that in any arbitrary constrained information flow network even with full knowledge of the information flow graph we may not be able to mitigate the mis-information propagation completely. Later on this section we present the condition as a topological constraint on the information flow network.

Now we are trying to find the algorithm $\mathcal{A}$ that assure the removal of the mis-information. Now get back to the full information flow network. In this protocol each node uses the Bayesian estimation to update the probability distribution of $x$ given the set of observations. The following recursive equation is used at each node to update the estimation of the probability distribution at node $n+1$:

$$p(x|Z_{n+1}(G_n), z_{n+1}) = \frac{p(x|Z_{n+1}(G_n))p(z_{n+1}|x)}{p(z_{n+1}|Z_{n+1}(G_n))}. \tag{14}$$

The denominator is a constant relative to $x$ therefore, Eq. (14) can be written as following:

$$p(x|Z_{n+1}(G_n), z_{n+1}) = \chi p(x|Z_{n+1}(G_n))p(z_{n+1}|x). \tag{15}$$

where $\chi = \frac{1}{p(z_{n+1}|Z_{n+1}(G_n))}$ is usually ignored in practice. Similarly, we can write the normalized Bayesian updates for $p(x|Z_{n+1}(G_n))$ as follows

$$p(x|Z_{n+1}(G_n)) = p(x|Z_{n+1}(G_n)/i)p(z_{i}|x) \quad i \in v(G_n). \tag{16}$$

Re-iterating Eq. (16) yields:

$$p(x|Z_{n+1}(G_n), z_{n+1}) = \pi_0 p(z_{n+1}|x) \prod_{i \in Z_{n+1}(G_n)} p(z_{i}|x). \tag{17}$$

To find the algorithm for mis-information removal problem, we assume that each node is estimating the logarithm of the probability distribution of $x$ given the set of available observations. Let $y_n^{full} = \log \left( p(x|Z_{n+1}(G_n), z_{n+1}) \right)$ be the mis-information free sufficient statistics of the parameter $x$ in social network.
Proposition 1: Let \( \mathcal{G}_N \) be a family of DAGs representing a full information flow network represented by full information protocol. With \( t_n \) defined in Sec. II, the optimal estimate free of mis-information at node \( n \) is:

\[
\hat{y}^{\text{full}}_n = (t_n \otimes I_d) i_{1:n-1} + i_n,
\]

where \( i_n \) denotes \( \log(p(z_n|x)) \) and \( i_{1:n-1} \triangleq [i_1', \ldots, i_{n-1}'] \in \mathbb{R}^{(n-1)d \times 1} \). Here \( \otimes \) denotes Kronecker (tensor) product and \( I_d \) denotes the \( d \times d \) identity matrix.

Proof: Proof can be found in Appendix A.

According to Proposition 1, the optimal estimate free of mis-information can be expressed as a linear function of the parameters via the Transitive Closure Matrix \( T_n \) of graph \( G_n \). This formula is quite intuitive. Recall that a node represents a node of type-based graph at a specific instant time. In the protocol for full information flow network described in Sec. II-B, a node broadcasts its own raw measurements and also pass the raw measurements received from others so that each node has the most accurate estimate. Consequently, the optimal estimate free of mis-information at node \( n \) is the sum of the information collected by the nodes such that there are paths between all these nodes and \( n \) (recall that \( G_n \) represents the information flow between the \( n \) first nodes).

We now present our main result. We return to the realistic model of a network with constrained information flow modeled by constrained information flow defined in Sec. II-B. In Proposition 2, we prove that the estimates in constrained information flow network and the full information flow network are equal, if and only if the topology of the network graph has a specific structure satisfying (21) defined below.

The Optimal Combination Scheme: Define new variable \( \hat{y}_n = \log \theta_n \) as the mis-information free sufficient statistics of the parameter \( x \) in social network. Our proposed combination algorithm states that the estimate should be constructed as a weighted sum of estimates from previous nodes \( \hat{y}_{1:n-1} \) and the current information \( i_n \). Therefore, with the \( n-1 \) dimensional vector \( w_n \) below denoting a weight vector (a more precise construction is given below), and \( i_n \) defined in (18), the probability distribution estimates at node \( n \) has the following form:

\[
\hat{y}_n = (w_n \otimes I_d) \hat{y}_{1:n-1} + i_n,
\]

where \( \hat{y}_{1:n-1} \triangleq [\hat{y}_1', \ldots, \hat{y}_{n-1}'] \). Therefore, \( \theta_n = \mathcal{A}(\Theta_n(G_n), z_n) \) is defined in such way that the logarithm of \( \theta_n \)'s satisfy (19). In few paragraphs we compute the optimal value for the weighting factor \( w_n \).

Next we introduce the following constraints on the \( n-1 \) dimensional weight vector \( w_n \) which imposes the condition that due to constrained information flow, some estimates are not available to node \( n \) due to the graph topology.

Constraint 1: Let \( \mathcal{G}_N = \{G_1, \ldots, G_N \} \) denote a family of DAGs representing the information flow in a node network. Consider \( a_n \) defined in Sec. II. Then the set of weights \( \{w_n\}_{n \in \{1, \ldots, N\}} \) in (19) satisfies the topological constraint for constrained information flow network if \( \forall j \in \{1, \ldots, n-1 \} \) and \( \forall n \in \{1, \ldots, N\} \)

\[
a_n(j) = 0 \implies w_n(j) = 0.
\]

According to (19), the estimate at node \( n \) is a weighted sum of the estimates computed by the previous nodes i.e. \( \hat{y}_{1:n-1} \) and the information collected \( i_n \) at node \( n \). However, according to the communication topology described by Adjacency Matrix \( A_n \), some nodes do not transmit their estimates to node \( n \). Consequently, these estimates are not available at node \( n \) and must not be used to compute estimate \( \hat{y}_n \) at node \( n \). An obvious way to introduce this constraint in the estimation of \( \hat{y}_n \) is to set the weight related to an unavailable estimate to zero. This is constraint (20). Therefore, it is also clear that (20) is a necessary condition for exact mis-information removal. Now, assuming that the estimates of the \( n-1 \) latest nodes are optimal estimates free of mis-information i.e. \( \hat{y}_{1:n-1} = \hat{y}^{\text{full}}_{1:n-1} \), is it possible to find a vector \( w_n \) under constraint 1 so that \( \hat{y}_n \) is equal to the optimal estimates free of mis-information i.e. \( \hat{y}^{\text{full}}_n \)? Proposition 2 below gives a necessary and sufficient condition on \( w_n \) for \( \hat{y}_n = \hat{y}^{\text{full}}_n \).

Proposition 2: Let \( \mathcal{G}_N = \{G_1, \ldots, G_N \} \) be a family of DAGs representing constrained information flow in the network represented by Protocol 1. Consider the class of estimators (19). Then the following optimality property holds for the estimates \( \hat{y}_n \) in the constrained information flow network:

\[
\hat{y}_n = \hat{y}^{\text{full}}_n \iff \left\{ \begin{array}{l}
w_n = t_n ((T_{n-1})')^{-1} \\
\text{and } w_n \text{ satisfies constraint (20)}
\end{array} \right.
\]

In words: A necessary and sufficient condition for the estimates \( \hat{y}_n \) of the constrained information flow network to be identical to the estimates \( \hat{y}^{\text{full}}_n \) of the full information flow network, is that the weight vector \( w_n \) satisfies \( w_n = t_n ((T_{n-1})')^{-1} \) and simultaneously satisfies topological constraint (20). Alternatively, a necessary and sufficient condition for exact mis-information removal is that the right hand side of (21) holds. Here \( t_n \) is defined in Sec. II and \( T_{n-1} \) is the Transitive Closure Matrix of \( G_{n-1} \).

Proof: We show first that the left hand side of (21) (i.e., \( \hat{y}_n = \hat{y}^{\text{full}}_n \)) implies the right hand side of (21). Start with (19) for \( \hat{y}_n \) and replacing \( \hat{y}_n \) with \( \hat{y}^{\text{full}}_n \) yields

\[
\hat{y}^{\text{full}}_n = (w_n \otimes I_d) \hat{y}^{\text{full}}_{1:n-1} + i_n.
\]

Using Proposition 1 it follows that \( \hat{y}^{\text{full}}_{1:n-1} = (T_{n-1}' \otimes I_d) i_{1:n-1} \). Then, incorporating this into (22) yields

\[
\hat{y}^{\text{full}}_n = (w_n \otimes I_d) (T_{n-1}' \otimes I_d) i_{1:n-1} + i_n.
\]

From Proposition 1, we have \( \hat{y}^{\text{full}}_n = (t_n \otimes I_d) i_{1:n-1} + i_n \). Equating the right hand sides of this equation and (23) yields

\[
(t_n \otimes I_d) i_{1:n-1} = (w_n \otimes I_d) (T_{n-1}' \otimes I_d) i_{1:n-1} = ((w_n T_{n-1}') \otimes I_d) i_{1:n-1}.
\]

The last equality above follows from the distributive property of tensor products. (24) is true for any information vector.
\(i_{1:n-1}\). This implies \(t_n = w_n T_{n-1}' \implies w_n = t_n (T_{n-1}')^{-1}\), since \(T_n\) is a upper triangular matrix with ones on the diagonal and so invertible. To complete the proof that the left hand side of (21) implies the right hand side, recall that the information structure for constrained flow is such that if \(\epsilon_n(j) = 0\) then certain components of the vector \(\hat{y}_{n|1:n-1}\) are not available to node \(n\). If the corresponding weight \(w_n(j)\) is non-zero it is impossible to reconstruct \(\hat{y}_n\) according to (19) to be equal to \(\hat{y}_{n|1:n}^{\text{all}}\). This is simply the topological constraint(20).

Showing that the right hand side of (21) implies the left hand side is very similar to the above proof and is omitted. \(\blacksquare\)

IV. SUB-OPTIMAL MIS-INFORMATION REMOVAL
ALGORITHM IN AN ARBITRARY NETWORK

In Sec. III, we proposed an optimal algorithm to exactly remove the mis-information in constrained information flow network at each time. However, this algorithm requires full knowledge of the information flow graph which can be achieved by transmitting additional information about the routing path of each message. In this section, we relax that assumption and consider the case that each node does not know the path of each message it received. Therefore, in addition to the social parameter \(x\), each node should estimate the information flow graph at each time. Therefore the mis-information removal problem in this case can be presented as follows:

\[
\begin{align*}
G_n &= (V_n, E_n), \\
E_n' &= g(E_{n-1}, \epsilon_{n-1}), \\
\hat{z}_n &= H_n x + v_n, \\
(\hat{\theta}_n, G_n) &= \mathcal{B}(\Theta_{u(G_n)}, \hat{z}_n, g(\cdot)),
\end{align*}
\]  

(25)

where \(\epsilon_{n-1}\) is random variable which models the statistical characteristics of \(E_n\) caused by the random delay. In Algorithm \(\mathcal{B}(\cdot)\), using the dynamics of the graph \(g(\cdot)\), the recorded observation \(\hat{z}_n\), and the information received from other nodes \(\Theta_{u(G_n)}\), each node should run a filter to estimate the information flow network, \(G_n\), and the sufficient statistics of parameter \(x\). Note that even in the case without full knowledge of information flow network, \(u(G_n)\) is known (because \(u(G_n)\) denote the nodes who are directly connected to node \(n\) in the information flow graph, i.e., each individual knows the type of who she received information from in social network).

To relieve the mis-information, the algorithm \(\mathcal{B}\) should be designed in a way that the distance between \(p(x|\Theta_{u(G_n)}, \hat{z}_n)\) and the optimal mis-information free estimates of the full information flow network (defined in (11)) is minimized. Therefore the mis-information problem without complete knowledge of information flow network can be defined formally as follows: How can we define the optimal algorithm \(\mathcal{B}\) to jointly combine the received beliefs of other nodes and the local recorded observation to estimate the information flow network \(G_n\) and minimize \(||p(x|\Theta_{u(G_n)}, \hat{z}_n) - p(x|\Theta_{u(G_n)}, \hat{z}_n)||\) . Finding a closed form optimal algorithm to solve seems impossible therefore, we propose a sub-optimal algorithm to minimize the expected mis-information at each time.

In the optimal scenario defined in Sec. III, \(v(G_n)\), all nodes who have a path to node \(n\), is know at each node \(n\). Therefore, using \(v(G_n)\) and the Transitive Closure matrix of the information flow graph, an algorithm is proposed that assures \(p(x|\Theta_{u(G_n)}, \hat{z}_n) = p(x|\hat{v}_n(G_n), \hat{z}_n)\) . In the general scenario, \(u(G_n)\) is given at each node \(n\) but \(v(G_n)\) is not completely known. So if we compute the probability of having a path to node \(n\), we can estimate the most probable nodes connected to node \(n\) (indirectly) namely \(\hat{v}(G_n)\) . Knowing \(\hat{v}(G_n)\), we can use the algorithm defined in Proposition 2 to remove the expected mis-information. We can summarize the sub-optimal mis-information removal problem without knowledge of information flow network as follows:

\[
\begin{align*}
\hat{G}_n &= (V_n, E_n), \\
\hat{E}_n &= g(E_{n-1}, \epsilon_{n-1}), \\
\hat{z}_n &= H_n x + v_n, \\
(\hat{\theta}_n, \hat{G}_n) &= \mathcal{B}(\Theta_{u(G_n)}, \hat{z}_n, \hat{v}(G_n)),
\end{align*}
\]

\(\hat{G}_n\) is the first \(n \rightarrow 1, \ldots, N\) elements of \(T_n\) defined in Appendix B. The following proposition presents the necessary and sufficient condition to remove the mis-information and the sub-optimal weighting factor.

**Proposition 3:** Let \(\mathcal{D}_n = \{G_1, \ldots, G_N\}\) be a family of DAGs representing constrained information flow in the network represented by Protocol 1 and \(\hat{v}(G_n)\) denote the hard estimate if the all nodes connected to node \(n\) in \(G_n\). Consider the sub-optimal estimates presented in (27) and the following class of estimator:

\[
\hat{y}_n = \hat{w}_n \otimes \mathbf{I}_d \hat{y}_{1:n-1} + \hat{r}_n.
\]  

(28)

Then the following optimality property holds for the estimates \(\hat{y}_n\), \(\hat{y}_{sub}\) in the constrained information flow network (for \(n = 1, 2, \ldots, N\)):

\[
\hat{y}_n = y_{sub} \iff \hat{w}_n = \hat{r}_n \left(\hat{T}_{n-1}\right)^{-1}
\]  

(29)

In the case without knowledge of information flow graph using the dynamic of the information flow graph and the statistical characteristics of \(\epsilon_{n-1}\), a set of nodes is estimated who has the relatively high probability of contribution in the estimation of nodes connected to node \(n\). Then, an algorithm is proposed to remove the mis-information expected to generate by these nodes. A necessary and sufficient condition for the estimates \(\hat{y}_n\) of the constrained information flow network to be identical
to the estimates $y_{nub}$ of the full information flow network, is that the weight vector $\bar{w}_n$ satisfies $\bar{w}_n = \bar{t}_n \left( \bar{T}_{n-1} \right)^{-1}$ and simultaneously satisfies topological constraint (20). Here $\bar{t}_n$ is defined in (27) and $\bar{T}_{n-1}$ is defined in Appendix B.

**Algorithm 1** Algorithm for mis-information removal

For $n = 1, 2, \ldots$

1) Reconstruct the expected graph of information flow, $\bar{G}_n$ as described in Appendix B.

2) Using the expected adjacency matrix and the expected transitive closure matrix of information graph, find all nodes who have a path to node $n$ with probability more than a threshold, $\lambda_{th}$, namely $\bar{v}(\bar{G}_n)$.

3) Find the sub-optimal weighting vector $\bar{w}_n$ as described in Proposition 3.

**V. NUMERICAL EXAMPLES**

In this section, numerical results are given to illustrate the effect of the mis-information and verify the results from Sec. III. We start with showing the effect of mis-information in the performance of the multi-agent Bayesian estimation. Consider a social network consisting of $S = 5$ different types. These nodes are estimating a scalar parameter $x$ in the network with a given apriori distribution. We simulate a social network that the delay of each link connecting different clusters is random quantized integer in $\{1, 2, \ldots, 5\}$. We assume that all these delays are equiprobable. The prior distribution of $x$ is uniform distribution $U[0, 4]$ and the observation noise is zero-mean normal distribution $N(0, 1)$.

Fig. 1 illustrates the effect of mis-information propagation in the Bayesian estimators. As can be seen in this figure, the performance of the Bayesian estimator is ruined in the existence of the mis-information. The blue line shows the expected value of $x$ using given the set of available (due to the topology of the information flow graph) raw measurements (of all previous nodes) in the full information protocol defined in II. The red line that converges to a slightly different value (than $x$ showed by dashed line) is obtained by naive recursive Bayesian estimation at each node. Fig. 2 shows the excellent performance of the mis-information removal algorithm presented in Proposition 2. As expected from Sec. III, the estimates of the algorithm defined in Proposition 2 is very close to those of the optimal mis-information free estimate of the full information flow protocol depicted by the blue line.

**VI. CONCLUSIONS AND FUTURE WORKS**

In this paper we addressed the problem of mis-information propagation among different types of the social networks. We considered the most general case with any type of observation noise and any apriori distribution of the parameter under investigation. A sufficient and necessary condition for mis-information removal problem is derived based on the topology of the information flow network. Also the performance of the proposed mis-information removal algorithm is illustrated in numerical examples.

An extension to this work can be considering the topologies that do not satisfy Constraint 19 and finding a probably sub-optimal approach for mis-information removal in this scenario.

**APPENDIX**

**A. Proof of Proposition 1**

Proof: Each $j \in v(G_n)$ is the index of the $j$th node connected to node $n$ (elements of $v(G_n)$), it means that there is a path between node $j$ and node $n$ and from the definition
of $t_n$, $t_n(j) = 1$ Therefore, $(t_n \otimes \mathbf{I}_d) t_{1:n-1}$ can be written as
\[
(t_n \otimes \mathbf{I}_d) t_{1:n-1} = \sum_{j \in \mathcal{V}(G_n)} i_j \\
= \sum_{j \in \mathcal{V}(G_n)} \log(p(z_j|x)) \\
= \sum_{j \in \mathcal{V}(G_n)} \log(p(z_j|x)).
\] (30)

Taking logarithm form (17), yields:
\[
\log(p(x|z_{\mathcal{V}(G_n)},z_k)) = \sum_{j \in \mathcal{V}(G_n)} \log(p(z_j|x)).
\] (31)

Using (31) and (30), we can rewrite $\gamma_{n}^{\text{full}}$ and complete the proof,
\[
\gamma_{n}^{\text{full}} = \sum_{j \in \mathcal{V}(G_n)} \log(p(z_j|x)) \\
= \sum_{j \in \mathcal{V}(G_n)} \log(p(z_j|x)) + \log(p(z_n|x)) \\
= (t_n \otimes \mathbf{I}_d) t_{1:n-1} + t_n.
\] (32)

B. Constructing the Expected Adjacency Matrix of Information Flow Graph

As described in Sec. II-A, individuals in a social network are divided into different types based on their characteristics. Nodes with the same type share the same economic, social, or political characteristic therefore we can represent a cluster (a component of graph $H$ which represents the social network) of nodes with type $s$ at time $k$ with a single node $n$ in the “type-based” graph where $n = s + S(k-1)$. In this graph each link has a propagation delay which shows the time it takes that a message propagates form one node to another. This delay depends on the total numbers of connections between two clusters in the graph that represents the social network. In addition to the delay each message face to transmit from one cluster to another, it takes time for each belief to diffuse through all individuals within that specific cluster. It is assumed that this delay corresponding to inter-connection within each cluster depends on the average degree of nodes in that cluster. We assume that the delay is quantized and can have values in $\{1, 2, \ldots, \Delta\}$. We assumed that the propagation delay cannot exceed $\Delta$. The statistical distribution of the propagation delay between cluster $i$ and cluster $j$, $\tau_{ij}$, is function of the total number of connections between cluster $i$ and cluster $j$. It also depends the average degree of nodes in both clusters. Therefore we can say that the probability distribution of $\tau_{ij}$ can be written as following:
\[
p_{ij}^{\tau} = f(d_{i}^{\text{avg}}, d_{j}^{\text{avg}}, d_{ij}),
\] (33)

where $d_{i}^{\text{avg}}$, the average degree of nodes in cluster $i$, and $d_{ij}$, the total number of connections between cluster $i$ and cluster $j$, can be easily computed using the adjacency matrices of graph $H$ and subgraphs $H_{i}$ and $H_{j}$. Using the probability distribution defined in (33), we can say that for example the message of node with type $i$ at time $k$ may be available at node with type $j$ at time $k_1$ with probability $p_1$ and at time $k_2$ with probability $p_2$ and at other times with zero probability. Therefore, the expected adjacency matrix of information flow can be reconstructed using the probability of having a link between all nodes. In the above example, we know the element on the row $i+S(k-1)$ and column $j+S(k_1-1)$ is $p_1$ and the element on the same row and columns $j+S(k_2-1)$ is $p_2$. All the steps to construct the expected adjacency matrix of the information flow graph are summarized all in Algorithm 2.

Algorithm 2 Algorithm for constructing the expected adjacency matrix of the information flow graph

For $n = 1, 2, \ldots$

1) find the corresponding time and type to each value of $n$ (21) (recall $n \equiv [s,k]$).

2) Using the adjacency matrix of the type based graph, $L$, find all neighbors of node with type $s$; namely $N_{\text{Neighbor}}(s)$

3) For each type $s' \in N_{\text{Neighbor}}(s)$ and each $j$, $1 \leq j \leq \Delta$

- find $p_{n}^{s'}$, the probability of receiving the message of node $n$ at node of type $s'$ at time $k+j$ using (33)
- $\tilde{A}_{n,s'+S(k+j-1)} = p_{n}^{s'}$

Each elements of $\tilde{A}$ shows the probability of having an edge in the information flow network. The probability of having a path between two nodes can be found by multiplying the probabilities of adjacent edges between two nodes and the inclusion-exclusion principle. Let $\tilde{T}_n$ denote the expected transitive closure matrix with elements $t_{ij}$ which shows the probability of having a path between node $i$ and node $j$. $\tilde{T}_n$ and $T_n$ are used to find the sub-optimal weighting vector for mis-information removal problem.

REFERENCES