Reaching diagnostic agreement in Multi-Agent Diagnosis

Nico Roos
IKAT
Universiteit Maastricht
P.O.Box 616
6200 MD Maastricht
roos@cs.unimaas.nl

Annette ten Teije
Faculty of Sciences
Free University
De Boelelaan 1081A
1081 HV Amsterdam
annette@cs.vu.nl

Cees Witteveen
EWI
Delft University of Technology
PO Box 5031
2600 GA Delft
c.witteveen@its.tudelft.nl

Abstract

We consider the problem of finding a commonly agreed upon diagnosis for errors observed in a system monitored by a number of different expert agents. Each agent is assumed to have its own specialized (expert) view on the system and collectively, the agents have to agree on one or more diagnoses based on their views. Reaching an agreement is complicated by the two factors: (i) different specialisms need not distinguish the same fault modes of a component and (ii) knowledge of different specialisms need not be correct in some cases. This paper analyzes these problems and presents protocols that enable the agents to deal with these issues.

1. Introduction

A traditional diagnostic tool can be viewed as a single *diagnostic agent* having a model of the whole system to be diagnosed. In some applications, however, such a single agent approach is infeasible or at least undesirable. For example, the integration of knowledge into one model of the system is infeasible if the system is too large, too dynamic or distributed over different legal entities. Integration is undesirable if it concerns the combination of knowledge from different fields of expertise. In this latter case, where knowledge is called to be *semantically distributed*[1] [6], it would be better to introduce specialized agents communicating about anomalies detected.

The introduction of specialized (expert) agents immediately raises the problem how to reach an agreement on the cause of observed anomalies. As was pointed out in [12], assuming a fixed maximum number of broken components, there exists a polynomial time protocol for reaching an agreement between the agents in case of a semantic knowledge distribution. This protocol is rather straight forward. A more difficult situation arises if the knowledge of some agents is *incomplete* in the sense that the agents have no behavioral knowledge about some fault modes, or if the knowledge of some agents is *incorrect* in the sense that the agents have incompatible knowledge about the behaviors of components. In this paper, we will address both issues.

Remark We do not consider formulations based on Discrete Event Systems [3, 10]. These formulations emphasize more the the dynamical aspects of failure events on an abstract level. We neither consider diagnosing disagreements, plan-failure, organizational problems and misbehavior in a group of collaborating agents [4, 7, 8].

This paper is organized as follows. Section 2 specifies the diagnostic setting, which is extended to multi-agent diagnosis in section 3. Section 4 introduces a protocol for determining the global diagnoses on which all agent can agree. Section 5 addresses the probability of the derived diagnoses and section 6 evaluates in a series of experiments whether the most probable diagnoses will contain the correct diagnosis. Section 7 concludes the paper.

2. The diagnostic setting

A system to be diagnosed is a tuple $S = (C, M, Id, Sd, Ctx, Obs)$ where C is a set of components, $M = \{M_c \mid c \in C\}$ is a specification of possible behavior modes per component, Id is a set of identifiers p of connection points between components, Sd is the system description, Ctx is a specification of input values of the system that are determined outside the system by the environment and Obs is a set of observed values of the system. A component in C has a normal mode $nor \in M_c$, one general fault mode $ab \in M_c$ and possibly several specific fault

1 Besides a semantic knowledge distributed, we also distinguish a *spatial knowledge distribution*: knowledge of system behavior is distributed over the agents according to the spatial distribution of the system's components. The latter has been discussed in [13].

modes. We assume that all components have *in-* and *outputs*.[2]

The system description $Sd = Str \cup Beh$ consists of a structural description Str and a behavioral description Beh of the components. The structural description Str consists of instances of the form $p = in(x, c)$ or $p = out(x, c)$ where x is an in- or an output identification of a component c and $p \in Id$ is a connection point identifier[3]. Of course, a connection point $p \in Id$ is connected to at most one output of some component; i.e. if $p = out(x, c)$ and $p = out(y, c')$, then $x = y$ and $c = c'$. A connection point has a value, which is determined by the output of a component or a system input. The function $value(p)$ denotes the value of the connection point. The set of input connection points $Id^{\text{in}} \subset Id$ is defined as $Id^{\text{in}} = \{p \in Id \mid \forall x, c : (p = out(x, c)) \notin Str\}$.

The set $Beh = \bigcup_{c \in C} Beh_c$ specifies a behavior for each component $c \in C$. The behavior description Beh_c of a component specifies the component's behavior for each mode m in M_c as an implication of the form $mode(c, m) \rightarrow \Phi$[4] where the predicate $mode(c, m)$ is used to denote the mode $m \in M_c$ of a component c. The formula Φ describes the component's behaviour given its mode $m \in M_c$. In the special case that $m = ab$, the behavioral description does not specify a specific behavior, i.e., $mode(c, ab) \rightarrow \top$ is the behavioral description for the general fault mode.

The set Ctx describes the values of system inputs Id^{in} that are determined by the environment. Hence Ctx consists of instances of the form $value(p) = v$ where $p \in Id^{\text{in}}$ is a connection point and v is a value.

Finally, let $Id^{\text{obs}} \subseteq Id$ be the set of connection points that are observed by the diagnostic agent. Like Ctx, the set Obs describes the *values* of those connection points and consists of instances of the form $value(p) = v$ where $p \in Id^{\text{obs}}$ and v is a value.

A *candidate diagnosis* is a set D of instances of the predicate $mode(,)$ such that for every component $c \in C$ there is exactly one mode in $m \in M_c$ such that $mode(c, m) \in D$. A *diagnosis* is defined as follows:

Definition 1 *Let $S = (C, M, Id, Sd, Ctx, Obs)$ be the system to be diagnosed and let $\mid\sim$ to denote the possibly limited reasoning capabilities of a diagnostic system.*[5] *Moreover, let Obs^{con}, $Obs^{\text{abd}} \subseteq Obs$ be subsets of observations and let D be a candidate diagnosis. Then D is a diagnosis for S iff*

$$D \cup Sd \cup Ctx \mid\sim \bigwedge_{\varphi \in Obs^{\text{abd}}} \varphi \text{ and}$$
$$D \cup Sd \cup Ctx \cup Obs^{\text{con}} \not\mid\sim \bot.$$

Remark In the literature two types of diagnoses are distinguished: *consistency based* [9, 11] and *abductive* [1] diagnosis. Both can be combined into one more general diagnostic definition [2]. This latter definition is used here.

3. Multi-agent diagnosis

A (knowledge) distribution of a system S over a set A of k agents $\{A_i\}_{i=1}^k$ induces a division of S into k subsystems S_i. In the case of a *semantical* knowledge distribution, each agent A_i diagnoses S from a different *perspective* S_i on S. Here, we define such a perspective S_i on S as a subsystem $S_i = (C, M, Id, Sd_i, Ctx, Obs_i, In_i)$ of S that is related to S as follows:

The components C are known to all agents but the system description $Sd_i = Str_i \cup Beh_i$ may differ from one agent to the other.

The connections between components may only be relevant from specific perspectives; e.g. connection for electrical signals and connection for conducting heat. We therefore define Str_i as the subset of instances $p = in(x, c)$ and $p = out(x, c)$ that occur in Str where the value of p falls in the perspective i. Of course, $Str = \bigcup_{i=1}^k Str_i$.

Distributing the structural description Str of S over the agents implies that also the observations Obs must be distributed over the agents. Even if agents consider the same connections, it may be necessary to distribute the observations since agents may look from different perspectives to the value of a connection. For instance, an electrical signal can be divided in an DC-component and an AC-component thereby creating different perspectives. Hence, with each perspective i there corresponds a set of observations Obs_i where $Obs \equiv \bigwedge_{i=1}^k Obs_i$ and for each perspective i and each $(value(p) = v) \in Obs_i$, there is a $p = out(x, c) \in Str_i$.[6]

The set Beh_i specifies the behaviour from the perspective of agent i and each component $c \in C$ has a specific behavior $mode(c, m) \rightarrow \Phi_i \in Beh_{c,i}$ for each behavior mode $m \in M_c$. These behaviours are related to the components total behaviour Beh_c as follows: If $mode(c, m) \rightarrow \Phi_i \in Beh_{c,i}$ for $i \in \{1, ..., k\}$ and if $Beh_c = mode(c, m) \rightarrow \Phi$, then $\Phi \equiv (\bigwedge_{i=1}^k \Phi_i)$. That is, for the normal mode and for each fault mode, the complete behavioral description Φ is equivalent to the conjunction of the behavioral descriptions given in each perspective.

2 This assumption is not valid in every system. We can, however, transform most systems to a system consisting components with only inputs and outputs (see for instance [5]).

3 A connection between components is modeled by *connection point* that is shared by one or more inputs and an output. Note that a physical connection should be modeled by component.

4 Note that we may use a single description for a class of components. Instances of this description must imply the form of description give here.

5 I.e $\{\varphi \mid \Sigma \mid\sim \varphi\} \subseteq \{\varphi \mid \Sigma \vdash \varphi\}$.

6 Note that the introduction of different perspectives implies that connection point can have different values but no more than one for each perspective.

Though agents need not know the connections that fall outside there perspective, the values of a component's inputs that are determined by such a connection may be relevant to predict the behavior given a mode m from a perspective i. E.g. the temperature of the environment of a component may be relevant for its electrical behavior. Other agents must provide the agent A_i with the values of these inputs. Therefore, we add the set In_i that denote the connection points the (input) values of which are provided by other agents.

The diagnosis of one agent Each agent A_i in the multi-agent system must be able to make a diagnosis of the subsystem $S_i = (C, M, Id, Sd_i, Ctx, Obs_i, In_i)$. This can be viewed a single agent diagnosis if values of the inputs and outputs of the subsystem are known. We use the set V_i to denote value assignments $value(p) = v$, with $p \in In_i$, to the inputs. Hence, V_i is the local context of the subsystem S_i that is determined by the outputs of other subsystems. The following definition is a simple extension of Definition 1 to handle a diagnosis of a subsystem S_i.

Definition 2 *Let $S_i = (C, M, Id, Sd_i, Ctx, Obs_i, In_i)$ be a subsystem from the perspective of agent A_i. Let Obs_i^{con}, $Obs_i^{\mathrm{abd}} \subseteq Obs_i$ be subsets of the observations, and let V_i be a description of the values of the (input) connection points In_i. Finally, let D_i be a candidate diagnosis of S_i. Then D_i is a diagnosis for S_i iff*

$$D_i \cup Sd_i \cup Ctx \cup V_i \mid\sim \textstyle\bigwedge_{\varphi \in Obs_i^{\mathrm{abd}}} \varphi \text{ and}$$
$$D_i \cup Sd_i \cup Ctx \cup V_i \cup Obs_i^{\mathrm{con}} \not\mid\sim \bot.$$

The diagnosis of multiple agents If there is a need for using multiple diagnostic agents, an important question is whether we loose information using a multi agent approach with respect to a single agent approach. To answer this questions we assume *there are no principal conflicts between the knowledge of the different agents*; i.e. there always exists a diagnosis D such that: $D \cup Sd \cup Cxt \cup Obs$ is consistent. We need this assumption because single agent diagnosis requires consistent knowledge.

Proposition 1 [7] *Let $S_1, ..., S_k$ be the subsystems generated by the perspectives on S, let D be a single agent diagnosis of S, and let $V_i = \{(value(p) = v) \mid p \in In_i,\ D \cup Sd \cup Ctx \mid\sim (value(p) = v)\}$ be the local context in S_i.*

Then D is a diagnosis of S_i.

Proposition 2 *Let $S_1, ..., S_k$ be the subsystems generated by the perspectives on S.*

Then, D is a single-agent diagnosis if D diagnosis of every subsystem S_i given the local contex $V_i = \{(value(p) = v) \mid p \in In_i, D_j \cup Sd_j \cup Ctx \cup V_j \mid\sim (value(p) = v)\}$.

7 The proofs are omitted because of lack of space.

The above propositions show that multi-agent diagnosis is possible. Note, however, that given a global candidate diagnosis D, predicting the values of all connection points is an NP-Hard problem [12]. When knowledge of the system is semantically distributed over the agents, often there are only a few connection points between the subsystems managed by different agents. Moreover, if the connections between subsystems do not form cycles, the distribution of knowledge over the agents does not contribute significantly to the time complexity of predicting the system's behavior given a diagnosis. Since usually, there are not many connections between different behavioral aspects of the system, in the remainder of this paper, we will assume that the prediction of the system's behavior is not an issue.

A single agent approach is based on the implicit assumption that an agent has complete and consistent knowledge of a component's behavior given its known behavioral modes. Without this assumption, a single agent cannot make a diagnosis using Definition 1. However, when knowledge is semantically distributed, this assumption need not be valid. Therefore, we must study the consequences of incomplete and incorrect knowledge on establishing a global diagnosis.

3.1. Agents with incomplete knowledge

When agents look at different aspects of a component, they may not have the same detailed knowledge for every aspect. Concerning the electrical aspects of an integrated circuit for instance, an agent may distinguish many specialized fault modes for which knowledge concerning the thermodynamic aspects of the circuit is lacking. Hence, for a component c an agent A_i may only have behavioral knowledge for *some* of the component's fault modes $M_{c,i} \subseteq M_c$.

The lack of knowledge about a component's behavior for some fault modes raises a problem: the agents may not be able to reach an agreement. To overcome this problem an agent A_i may just *assume* a behavior for each behavior mode $m \in (M_c - M_{c,i})$. But then the problem is, which behaviors can validly be assumed? If the behavior of a less specific fault mode would be known, this behavior may be used. Since a set of behavior modes $M_{c,i}$ always contains the least specific fault mode ab for which no behavior is known, we may assume the existence of a hierarchy of modes ordered with respect to specificity. We call such a hierarchy an *abstraction hierarchy*. For an example, see figure 1

Definition 3 *Let $c \in C$ be a component and M_c its set of behavior modes. An abstraction hierarchy on M_c is a strict partial order $(M_c, \succ)$, where the intuitive meaning of $m \succ m'$ $m, m' \in M_c$ is that m is more specific than m' and where ab is the unique least specific element in the hierarchy, i.e. for all $m \in M_c - \{ab\}$: $m \succ ab$.*

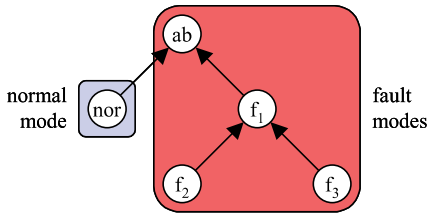

Figure 1. An abstraction hierarchy.

Intuitively, a more specific mode implies a more specific description of the behavior of the component. Moreover, more specific modes should be mutually exclusive. Therefore, the following requirements must hold.
For every $m, m' \in M_{c,i}$*: if* $m \succ m'$, $mode(c, m) \rightarrow \Phi \in Beh_{c,i}$ *and* $mode(c, m') \rightarrow \Phi' \in Beh_{c,i}$, *then* $\Phi \models \Phi'$.
For every $m, m', m'' \in M_c$ *if* $m' \succ m$, $m'' \succ m$, $mode(c, m') \rightarrow \Phi' \in Beh_{c,i}$, $mode(c, m'') \rightarrow \Phi'' \in Beh_{c,i}$, *then* $\Phi', \Phi'' \models \bot$.

Definition 4 *Let* $\Phi_{i,\mathrm{nor}}$ *be the normal behavior from the perspective* i *of a component* c

An abstraction hierarchy is complete *iff for each a fault mode* m_0 *that is not a most specific fault mode, there is a set of fault modes* $m_1, ..., m_\ell$ *such that* $m_j \succ m_0$ *for* $j \geq 1$, $mode(c, m_j) \rightarrow \Phi_{i,j} \in Beh_{c,i}$ *and* $\models \Phi_{i,0} \leftrightarrow (\Phi_{i,1} \vee ... \vee \Phi_{i,\ell})$.

The abstraction hierarchy on the fault modes defines a similar abstraction hierarchy on the diagnoses.

Definition 5 *Let* D, D' *be two candidate diagnoses.* D *is at least as specific as* D', $D \succeq D'$, *iff for every* $mode(c, m) \in D$ *there is a* $mode(c, m') \in D'$ *such that* $m \succeq m'$.

Note that agents that wish to give a best possible explanation for the observed anomalies, should focus on the most specific diagnoses. Whether the agents only determine the most specific diagnoses depends on the type of diagnosis they use; i.e. consistency based or abductive based:

Proposition 3 *Pure abductive diagnosis may not produces less specific diagnoses.*

Pure consistency based diagnosis returns every less specific diagnosis.

Proposition 4 *Let* $S_1, ..., S_k$ *be the subsystems that make up the system* S *and let the abstraction hierarchy of fault modes be complete. Moreover, let* D *be a most specific diagnosis of* S.

Then there exists a set of most specific diagnoses $D_1, ..., D_k$ *for, respectively* $S_1, \ldots S_k$, *such that* $D = \bigcup_{i=1}^{k} D_i$.

The behavioral description $Beh_{c,i}$ of a component c with respect to a perspective of agent A_i might be incomplete w.r.t. some fault modes, i.e. there might not be a behavior specification for each fault mode in M_c. In order for the agent to establish a global diagnosis, therefore these missing behaviors have to be added. The following assumption serves this purpose.

Assumption *A fault mode* m *of a component* c *for which an agent* A_i *has no behavioral specification, is assumed to have the same behavior as the most specific mode* $m' \in M_{c,i}$ *with* $m \succ m'$ *for which a behavior is known.*

The purpose of this assumption is to extend the behavioral description, making the behavioral knowledge of every fault mode of every component complete for all aspects. Hence, the results of propositions 1 and 2 apply.

3.2. Agents with incorrect knowledge

Agents lacking knowledge about behavior modes is not the only problem that may arise in a multi-agent system. Knowledge of agents about the components' behaviors may in some situation be incorrect. As a result, agents need not agree on the components that can be broken and if they do agree on the components that are broken, they need not agree on the fault modes of the broken components.

A robust multi agent system should be able to handle situations in which agents do not agree on global diagnoses. One possibility to overcome these problems proposed by Schroeder and Wagner [14] is the use of voting. However, if agents look from different perspectives at the system, voting offers no solution. Moreover, voting requires the communication of all local diagnoses of all agents. The number of these diagnoses may be exponential in the number of components.

The abstraction hierarchy on the fault modes also makes it possible to handle problems that arise because agent have incorrect knowledge about a components behavior. Given a *complete* abstraction hierarchy of fault modes agents must be able to agree on one or more most specific diagnoses When we are unable to reach an agreement on a most specific diagnosis, they should be able to agree on a less specific diagnosis. Such an agreement is always possible since $m \succ ab$ for every behavior mode $m \in M_c$. If m is the most specific fault mode of a component c on which the agents agree, there must be a diagnosis D with $mode(c, m) \in D$ and D is a consistency-based diagnosis of the subsystem S_i managed by agent A_i.

We can identify the presents of incorrect knowledge by the absent of a most specific diagnosis if the abstraction hierarchy is complete. This does not imply that the knowledge is correct if agents agree on a most specific diagnoses in all circumstances.

Proposition 5 *Let $S_1, ..., S_k$ be the subsystems that make up the system S and let the abstraction hierarchy of behavior modes be complete.*

The knowledge of an agent about the behavior of a component is incorrect if for some context Cxt and some set of observations Obs no most specific diagnosis exits on which all agents agree.

4. A protocol for diagnostic agreement

The agents may determine a global diagnosis by first determining all fault modes $M_c = \bigcup_{i=1}^{m} M_{c,i}$ as well as the abstraction hierarchy $\succ$ on M_c for each component c, and second exchanging their local diagnoses. The first step is straight forward and will not be discussed here because of space limitations. The second step is more problematic. The number of diagnoses to be exchanged between the agents can be quite high and can be exponential in the number of component is the worst case. In order to control the complexity, agents should focus on numerical minimum, inclusion minimal or probable diagnoses.

Since a locally diagnosis need not be a global diagnosis, the agent proposing the diagnosis needs to receive feedback when a proposed diagnosis is rejected by other agents. Subsequently, the agent can generate a new diagnosis taking into account the diagnoses that have been rejected.

The generation of new diagnoses can be improved if agents supply the reasons for rejecting a proposed diagnosis. When agent A_1 proposes a partial diagnosis D_1, agents $A_2, ..., A_k$ might reject the diagnosis because some (combination of) modes is inconsistent with its observations. For $i = 2, \ldots k$, let $R_i \subseteq D_1$ be such (a combination of) modes. Then for R_i it should hold that R_i is a smallest subset of D_1 such that: $R_i \cup Sd_i \cup Ctx \cup V_i \cup Obs_i \not\vdash \bot$ for $2 \leq i \leq k$.

Note that an agent A_i might determine more than one smallest subset R_i. If SR_i is the set of all smallest subsets R_i, agent A_1 can use this information $TR = \bigcup_{2 \leq i \leq k} SR_i$ as a set of constraints in its search for a next diagnosis. It may not select a new diagnosis D_1' containing any $R_i \in TR$ as a subset.

The protocol in figure 2 shows how the agents may proceed. To gain robustness, eventually, always one of the agents takes the initiative to establishes the global diagnoses. In the protocol, the agent that takes the initiative is agent A_1.

Agent	**Action**
A_1	$TR := \varnothing$;
A_1	finished := false;
A_1	while not finished do
A_1	generate the next most specific local diagnosis D_1 of S_1 such that for no $R \in TR$: $R \subseteq D_1$;
A_1	finished := not diagnosis_found;
A_1	while diagnosis_found, for $i := 2$ to k do
A_1	send 'propose D_1' to A_i;
A_i	receive 'propose D_1' from A_1;
A_i	determine a most specific local diagnosis D_i of S_i such that $D_1 \succeq D_i$;
A_i	if a diagnosis D_i exists then;
A_i	send 'accept D_i' to A_1;
A_i	else
A_i	send 'reject SR_i' to A_1;
A_i	end;
A_1	if received 'accept D_i' from A_i then
A_1	$D_1 := D_i$;
A_1	else if received 'reject SR_i' from A_i then
A_1	$TR := TR \cup SR_i$;
A_1	diagnosis_found := false;
A_1	end;
A_1	end;
A_1	if diagnosis_found then
A_1	store D_1;
A_1	end;
A_1	end;

Figure 2. Establishing global diagnoses

5. Probable diagnoses

The protocol proposed in the previous section determines a more abstract diagnosis in case the agents cannot agree on a most specific diagnosis. In this way agents can reach an agreement even if the knowledge of some agent predicts the wrong behavior given the current context and current observations. What we would like to know is how this affects the probability of a diagnoses, especially if the knowledge of some agents is incorrect given the current context and current observations.

What to measure Before applying probabilities measures, it is important to first determine what exactly we wish to measure. First of all we wish to know which components are broken. Therefore, we should determine the probability that some components $B \subseteq C$ are broken while others $C - B$ are not. Hence, if $\Delta(Obs)$ is the set of diagnoses given the agents' observations Obs, we should determine the probability of the diagnoses $\Delta(B, Obs) = \{D \in \Delta(Obs) \mid B = \{c \in C \mid mode(c, nor) \notin D\}\}$. Let $L_B = \{mode(c, nor) \mid c \in C - B\} \cup \{mode(c, ab) \mid c \in B\}$ the least specific diagnosis in which the components B are broken. Then $\Delta(B, Obs) \neq \varnothing$ iff $L_B \in \Delta(B, Obs)$ iff: $L_B \in \Delta(Obs)$. Therefore it suffices to determine the probability $P(L_B \mid Obs)$ for every $L_B \in \Delta(Obs)$.

The number of diagnoses $L_B \in \Delta(Obs)$ can be exponential in the number of components C. Fortunately, we

can reduce this number. Since the a priori probability that a component is not broken is in general much greater than the probability that it is broken, it suffices to determine the probabilities of the least specific *subset-minimal* diagnoses. If the fault probabilities are very small, we may restrict ourselves to the diagnoses of minimum cardinality.

A priori probabilities To determine the probability of diagnoses, we assume that fault probabilities are known for every fault mode of every component. If $spec(m) = \{n \mid n \succ m, \forall \ell \succ m : n \not\succ \ell\}$ denotes the set of fault mode that directly refines a fault mode m, then the following properties hold.

- For every behavior mode $m \in M_c$ of a component c, $spec(m)$ is a set of mutual exclusive fault modes.
- If the abstraction hierarchy of behavior modes for a component c is complete, then for every $m \in M_c$:

  $$P(c, m) = \sum_{n \in spec(m)} P(c, n),$$

  and if it is incomplete , then for every $m \in M_c$:

  $$P(c, m) \geq \sum_{n \in spec(m)} P(c, n).$$

From the probabilities of fault modes, the a priori probabilities of diagnoses can be derived. Given a diagnoses D,

$$P(D) = \prod_{mode(c,m) \in D} P(c, m).$$

For the probabilities of diagnoses similar properties as above hold. Let $spec(D) = \{D' \mid D' \succ D, \forall D'' \succ D : D' \not\succ D''\}$ be the set of diagnoses that directly specialize D. Then we have:

- For every candidate diagnosis D, $spec(D)$ is a set of mutual exclusive diagnoses.
- If the abstraction hierarchy of behavior modes is complete for every component in C, then for every candidate diagnosis D and for every set of observations Obs:

  $$P(D \mid Obs) = \sum_{D' \in spec(D)} P(D' \mid Obs),$$

  and if it is incomplete, then for every candidate diagnosis D and for every set of observations Obs:

  $$P(D \mid Obs) \geq \sum_{D' \in spec(D)} P(D' \mid Obs).$$

A posterior probabilities Deriving the a posterior probability of a diagnosis is more complicated, since the set of diagnoses given the agents' observations need not be mutually exclusive; if D is a diagnosis, then so is D' with $D \succ D'$. In deriving the probability measure, we will first assume that the knowledge of the agents is correct. Hence, we must first determine the a priori probability of a diagnosis $D \in \Delta(Obs)$ knowing that every more specific candidate diagnosis $D' \succ D$ with $D' \notin \Delta(Obs)$ is *no* diagnosis. We will use $\Gamma(D, \Delta(Obs))$ to denote this event. Then we have:

$$P(\Gamma(D, \Delta(Obs))) = P(D) - \sum_{D' \in spec(D)} P(D') + \sum_{D' \in spec(D) - \Delta(Obs)} P(\Gamma(D', \Delta(Obs)))$$

Note that in a complete hierarchy $P(D) - \sum_{D' \in spec(D)} P(D') = 0$ if D is not a most specific diagnosis.

The a posterior probability of a diagnosis D given the agents' observations Obs can now be determined by normalizing the a priori probabilities $P(\Gamma(D, \Delta(Obs)))$ of $D \in \Delta(Obs)$.

$$P(D \mid Obs) = \frac{P(\Gamma(D, \Delta(Obs)))}{\sum_{L_B \in \Delta(Obs), B \subseteq C} P(\Gamma(L_B, \Delta(Obs)))}$$

Note that for every $D \succ L_B$, $D \in \Delta(Obs)$ implies $L_B \in \Delta(Obs)$. Also note that the denominator summates over an exponential number of subsets $B \subseteq C$. If the fault probabilities are sufficiently small, we can ignore all non numerical minimum subsets.

Till now the agents $A = \{A_i\}_{i=1}^{k}$ were assumed to have perfect knowledge of the system in their area of expertise. Without this assumption, we must take into consideration that local diagnoses may be incorrect because of incorrect knowledge about the behaviour of subsystems. Hence, if D is the correct diagnosis, D need not be among the local diagnoses of some agents. The agents will, however, agree on the least specific diagnosis $L_B \prec D$. Without taking into consideration the probability that behavioral descriptions are incorrect, however, the a posterior probability of L_B is 0 if the abstraction hierarchy is complete.

If D is a most specific diagnosis supported by a subset $A(D) \subset A$ of the agents, there must be errors in the behavioral descriptions used by the agents $A - A(D)$. Therefore, given the local diagnoses $\Delta_1(Obs_1), ..., \Delta_k(Obs_k)$ of the agents, we must determine the probability of $P(\Gamma(D, \Delta_1(Obs_1), ..., \Delta_k(Obs_k)))$ by also taking into account the probability $P(sup(D) \mid Obs)$ that the agents $A(D)$ are correct in supporting the diagnosis D while the agents in $A - A(D)$ are wrong in not supporting D.

$$P(\Gamma(D, \Delta_1(Obs_1), ..., \Delta_k(Obs_k))) = (P(D) - \sum_{D' \in spec(D)} P(D')) \cdot P(sup(D) \mid Obs) + \sum_{D' \in spec(D)} P(\Gamma(D', \Delta_1(Obs_1), ..., \Delta_k(Obs_k)))$$

Since an agent's knowledge describes for each mode of a component its behavior, the probability that an agent is correct in supporting the diagnosis D or wrong in not supporting D, depends on the number of components on which the agent (dis)agrees with D. To determine this probability, we need to know the minimal number of components on which an agent disagrees with D. This number follows from the most specific diagnoses D' that resolves the disagreement with D. The set $agree(D, \Delta_i(Obs_i)) = \{D' \in \Delta_i(Obs_i) \mid D \succeq D',$ for no $D'' \in \Delta_i(Obs_i)$:$D \succeq D'' \succ$

$D'\}$ specifics these diagnoses. For each diagnosis $D' \in agree(D, \Delta_i(Obs_i))$, agent A_i agrees on $mode(c,m) \in D$ if $mode(c,m) \in D \cap D'$ and disagrees if $mode(c,m) \in D - D'$. If agent A_i disagrees on $mode(c,m) \in D$, then $P(fault(c,m))$ denotes the probability that a fault in description of the behavior of a component c with behavior mode m leads to an incorrect prediction of the component's behavior. Hence,

$$P(sup(D) \mid Obs) = \prod_{i=1}^{k} P(sup_i(D) \mid Obs_i) \text{ and}$$

$$P(sup_i(D) \mid Obs_i) = \sum_{D' \in agree(D, \Delta_i(Obs_i))} \Big(\prod_{mode(c,m) \in D - D'} P(fault(c,m)) \cdot \prod_{mode(c,m) \in D \cap D'} (1 - P(fault(c,m))) \Big)$$

Selecting leaf diagnoses After determining the a posterior probabilities of the diagnoses $L_B \in \Delta(Obs)$ the agents know which components are likely to be broken. Next, it is important to determine the most probable leaf diagnoses $D \succ L_B$. These diagnoses are important if components are not to be replaced but will be repaired. Here two issues play a role. First, agents can agree on several most specific diagnoses. In that case there is uncertainty about the correct one and agent can simply choose the most probable one. Second, agents may not agree on a most specific diagnosis. Hence the knowledge about the components' behaviors of some agent must be incorrect if L_B is a correct diagnosis. Since the above described a posterior probabilities of diagnoses take into account the agents supporting a diagnosis, again the agents can simply choose the most probable most specific diagnosis $D \succ L_B$.

If there are more than two most specific diagnoses, the most probable one can be less probable the sum of the probabilities of the other most specific diagnoses. Hence choosing the most probable most specific diagnosis need not the best choice. Instead, the agents should choose the most specific diagnosis $D \succ L_B$ for which there holds:

$$\frac{P(D \mid Obs)}{P(L_B \mid Obs)} > \theta > 0.5$$

where θ is some threshold value.

6. Experiments

In a series of experiments, we have evaluated how well the proposed approach is in determining the cause of a problem. Here, knowing the components that are broken, i.e. determine the least-specific diagnoses, is the first and most important step. Since no behavior is specified for the mode ab, the correct least-specific diagnosis will always be among the agent's diagnoses. An important question is, however, whether the correct[8] least-specific diagnosis is among the most probable least-specific diagnoses, especially if for some agents the description of a component's behavior is not free of errors.

To answer the question, 80000 systems were randomly generated and were diagnosed by three agents. We used three agents since this is the smallest number to make one leaf diagnosis significantly more probable if one of the agents disagrees with the others, while using more agents would have simplified the diagnostic problem.[9] Each generated system consisted of 40 components. Each component had one output and two inputs. An input was either corrected to one of the four system inputs or to an output of a randomly chosen component. The system was generated in such a way that it contained no cycles.

The normal behavior of a component was a modulo n adder for each to the three perspectives. Besides, a component had faulty behaviors, namely ab and two specific faulty behaviors f_1 and f_2. The corresponding abstraction hierarchy is shown in figure 3. In both fault modes f_1 and f_2 a fault value was added modulo n the output of the component. These faults values were randomly chosen for each combination of a component, a fault mode and an agent. Finally, for every component c, the same value was used for the probabilities of the fault modes $P(c, f_1)$ and $P(c, f_2)$ as well as the probabilities that the specified behavior of a behavior mode is incorrect $P(fault(c, f_1))$ and $P(fault(c, f_2))$.[10]

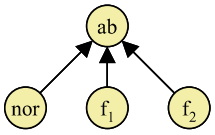

Figure 3. The hierarchy in the experiments.

To create a diagnostic problem, in each generated system one component was chosen to be the broken component and one of the fault mode f_1 or f_2 was selected for the component. In one of the three perspectives, however, the component behaved according to the other fault mode, i.e. the knowledge of the agent using this perspective was incorrect in the current situation.

Whether the least specific *correct* diagnosis is among the most probable diagnoses, mainly depends on the number of observation points and the number of values in- or output can have. We therefore varied these numbers in the experi-

8 The *correct* diagnosis is the one that we used to create the faulty behavior of the system.

9 Different perspectives provide more information.

10 If the probability value is small enough, the actual value becomes irrelevant.

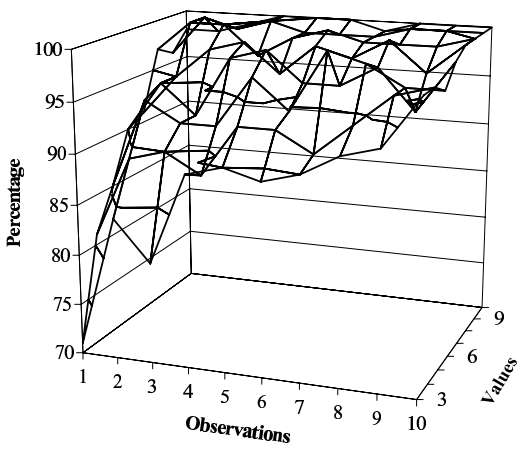

Figure 4. Distinguishability.

ments. To create the most difficult problems, the agents all used the same randomly chosen observation points. Figure 4 shows the percentage of problems in which the correct diagnosis is among the most probable diagnoses and figure 5 shows the average number of most probable diagnoses.

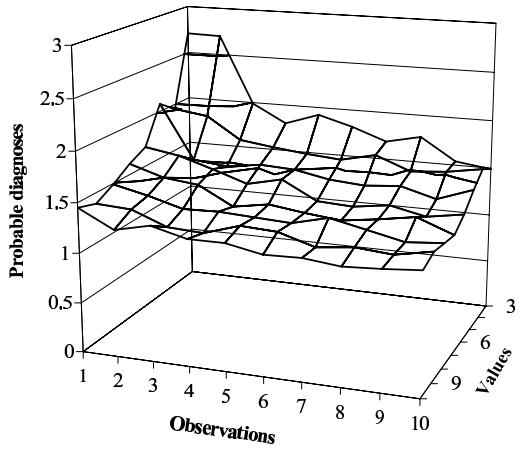

Figure 5. Probable diagnoses.

7. Conclusion

In this paper, we analyzed the problem of multi-agent diagnosis when knowledge is semantically distributed over the agents. Especially the case that the agents' knowledge concerning the faulty behavior of some components, is incorrect has been considered. A solution based on an abstraction hierarchy on the behavior modes has been proposed and a protocol for determining the global diagnoses with a minimal number of broken components has been given. Moreover, probabilistic correctness measures for diagnoses have been derived for the case that the agents' knowledge is correct and the case that the agents' knowledge is not always correct. These measure enable the agents to identify the most probable diagnoses. Finally, we investigated whether the most probable diagnoses may contain the correct diagnoses if one of the agent's knowledge contains errors. The results show that if the components have enough (> 6) output values this will be the case.

References

[1] L. Console and P. Torasso. Hypothetical reasoning in causal models. *International Journal of Intelligence Systems*, 5:83–124, 1990.

[2] L. Console and P. Torasso. A spectrum of logical definitions of model-based diagnosis. *Computational Intelligence*, 7:133–141, 1991.

[3] R. Debouk, S. Lafortune, and D. Teneketzis. Coordinated decentralized protocols for failure diagnosis of discrete-event systems. *Journal of Discrete Event Dynamical Systems: Theory and Application*, 10:33–86, 2000.

[4] C. Dellarrocas and M. Klein. An experimental evaluation of domain-independent fault-handling services in open multi-agent systems. In *ICMAS-2000*, pages 95–102, 2000.

[5] J. J. v. Dixhoorn. Bond graphs and the challenge of a unified modelling theory of physical systems. In F. E. Cellier, editor, *Progress in Modelling & Simulation*, pages 207–245. Academic Press, 1982.

[6] P. Frohlich, I. de Almeida Mora, W. Nejdl, and M. Schroeder. Diagnostic agents for distributed systems. In J.-J. C. Meyer and P.-Y. Schobbens, editors, *Formal Models of Agents. ESPRIT Project ModelAge Final Report Selected Papers. LNAI 1760*, pages 173–186. Springer-Verlag, 2000.

[7] B. Horling, B. Benyo, and V. Lesser. Using Self-Diagnosis to Adapt Organizational Structures. In *Proceedings of the 5th International Conference on Autonomous Agents*, pages 529–536. ACM Press, 2001.

[8] M. Kalech and G. A. Kaminka. On the design ov social diagnosis algorithms for multi-agent teams. In *IJCAI-03*, pages 370–375, 2003.

[9] J. d. Kleer and B. C. Williams. Diagnosing multiple faults. *Artificial Intelligence*, 32:97–130, 1987.

[10] Y. Pencolé, M. Cordier, and L. Rozé. Incremental decentralized diagnosis approach for the supervision of a telecommunication network. In *DX01*, 2001.

[11] R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–95, 1987.

[12] N. Roos, A. ten Teije, A. Bos, and C. Witteveen. An analysis of multi-agent diagnosis. In *AAMAS 2002*, pages 986–987, 2002.

[13] N. Roos, A. ten Teije, and C. Witteveen. A protocol for multi-agent diagnosis with spatially distributed knowledge. In *AAMAS 2003*, pages 655–661, 2003.

[14] M. Schroeder and G. Wagner. Distributed diagnosis by vivid agents. In *Proceedings of the first conference on Autonomous Agents*, pages 268–275, 1997.