Solvability-Based Comparison of Failure Detectors

Srikanth Sastry Josef Widder
Google, Inc. TU Wien

May 14, 2014

Abstract

Failure detectors are oracles that have been introduced to provide processes in asynchronous systems with information about faults. This information can then be used to solve problems otherwise unsolvable in asynchronous systems. A natural question is on the "minimum amount of information" a failure detector has to provide for a given problem. This question is classically addressed using a relation that states that a failure detector $\mathcal{D}$ is stronger (that is, provides "more, or better, information") than a failure detector $\mathcal{D}'$ if $\mathcal{D}$ can be used to implement $\mathcal{D}'$. It has recently been shown that this classic implementability relation has some drawbacks. To overcome this, different relations have been defined, one of which states that a failure detector $\mathcal{D}$ is stronger than $\mathcal{D}'$ if $\mathcal{D}$ can solve all the time-free problems solvable by $\mathcal{D}'$. In this paper we compare the implementability-based hierarchy of failure detectors to the hierarchy based on solvability. This is done by introducing a new proof technique for establishing the solvability relation. We apply this technique to known failure detectors from the literature and demonstrate significant differences between the hierarchies.

1 Introduction

Failure detectors [CT96] provide an oracular mechanism to circumvent the impossibility of several problems in fault-prone asynchronous systems [FLP85, FR03]. Intuitively, the idea is to enrich asynchronous systems with information about failures that may be useful to overcome the difficulties posed by process crashes. Chandra and Toueg [CT96] and Chandra, Hadzilacos, and Toueg [CHT96] demonstrated landmark results relating to failure detectors: the results in [CT96] demonstrated the use of failure detectors to solve consensus and other related problems, while the results in [CHT96] showed that any failure detector that can be used to solve consensus can also be used to implement a failure detector called Ω. Since Ω is also sufficient to solve consensus, it is the weakest failure detector to solve consensus. To arrive at these important results, [CT96] and [CHT96] introduced a relation to compare the "power" of failure detectors: denoted by $\mathcal{D} \succeq^{CT} \mathcal{D}'$, a failure detector $\mathcal{D}$ to said to be stronger than $\mathcal{D}'$, if $\mathcal{D}$ can be used to *implement* $\mathcal{D}'$.

Since [CHT96], the relation $\succeq^{CT}$ has been used to prove similar results for several other problems and has motivated the view that failure detectors could be used as "computability benchmark" [FGK11]; that is, an answer to the question on the weakest failure detector to solve a problem P is said to provide the minimal synchrony assumptions necessary to solve P in fault-prone systems [CHT96, FGK11]. This viewpoint is based on several implicit assumptions, one of which is that the hierarchy of failure detectors induced by the relation $\succeq^{CT}$ is similar to hierarchies induced by other natural relations, in other words, that it is robust.

In the work presented here, we focus on this assumption and explore the nature of relations that compare failure detectors. Incidentally, the robustness of the $\succeq^{CT}$ relation has been challenged in recent work [JT08, CBHW10], where it was observed that the relation has several drawbacks; for instance, $\succeq^{CT}$ is not reflexive. To overcome the drawbacks of the $\succeq^{CT}$ relation, new relations have been proposed in [JT08] and [CBHW10].

Jayanti and Toueg introduced a new relation in [JT08], which we denote $\succeq^{JT}$, with a different notion of what it means to *implement* a failure detector. The new relation $\succeq^{JT}$ extends $\succeq^{CT}$ and avoids several drawbacks of the $\succeq^{CT}$ relation[1]. Based on the $\succeq^{JT}$ relation, Jayanti and Toueg then demonstrate that every problem has a weakest failure detector. The results in [JT08] actually holds true for a specific class of problems, and in fact, later work by Bhatt and Jayanti [BJ09] shows that there exist a different class of problems that do not have a weakest failure detector. The apparent contradiction[2] between [JT08] and [BJ09] regarding the existence of weakest failure detectors demonstrates the significant dependence of weakest failure detector results on the definition of a "problem" and choice of the failure detector comparison relation.

In [CBHW10], Charron-Bost et al. advocate a new comparison relation denoted by $\succeq^{s}$. By definition, $\mathcal{D} \succeq^{s} \mathcal{D}'$ if every (time-free) problem solvable by $\mathcal{D}'$ is also solvable by $\mathcal{D}$. In contrast to the $\succeq^{CT}$ and $\succeq^{JT}$ relations, which are based on implementing one failure detector using another, the $\succeq^{s}$ relation depends on the set of problems solvable by each failure detector. If $\mathcal{D} \succeq^{CT} \mathcal{D}'$, or $\mathcal{D} \succeq^{JT} \mathcal{D}'$, then any problem solvable with $\mathcal{D}'$ can be solved using $\mathcal{D}$. Consequently, it is straightforward that $\succeq^{s}$ extends $\succeq^{CT}$ and $\succeq^{JT}$. However, given two failure detectors $\mathcal{D}$ and $\mathcal{D}'$, [CBHW10] provides no mechanism for demonstrating $\mathcal{D} \succeq^{s} \mathcal{D}'$ without having to establish $\mathcal{D} \succeq^{CT} \mathcal{D}'$ or $\mathcal{D} \succeq^{JT} \mathcal{D}'$. In effect, it is not clear how the $\succeq^{s}$ relation differs from the $\succeq^{CT}$ and $\succeq^{JT}$ relations.

Summary of results. In this paper, we address the aforementioned issues by providing a new proof technique to establish the $\succeq^{s}$ relation. Our approach is based on algorithm transformations, and to our knowledge, we are first to do so in the context of failure detector comparison. Although, from a technical viewpoint, the proofs are similar to existing proofs that establish $\succeq^{CT}$ and $\succeq^{JT}$ relations, the relationships resulting from our proof technique differ significantly from existing relationships among some failure detectors.

In order to illustrate the difference between $\succeq^{CT}$ and $\succeq^{s}$, we consider three families of failure detectors: the perfect failure detector $\mathcal{P}$ [CT96], the Marabout failure detector $\mathcal{M}$ [Gue01], and the $\mathcal{P}_k$ sequence [BJ09].[3]

The results in [Gue01] established that $\mathcal{M} \not\succeq^{CT} \mathcal{P}$ and $\mathcal{P} \not\succeq^{CT} \mathcal{M}$. In contrast, we show that $\mathcal{M}$ may be used to solve all the problems solvable using $\mathcal{P}$, and furthermore, there are problems that are solvable using $\mathcal{M}$ but not solvable using $\mathcal{P}$. In other words, we show that $\mathcal{M} \succeq^{s} \mathcal{P}$ and $\mathcal{P} \not\succeq^{s} \mathcal{M}$.

The results in [BJ09] show that $\mathcal{P}^k \succeq^{CT} \mathcal{P}^{k+1}$ and $\mathcal{P}^{k+1} \not\succeq^{CT} \mathcal{P}^k$. In contrast, we show that $\mathcal{P}^k$ and $\mathcal{P}^{k+1}$ can be used to solve the same set of problems, that is, for any k, the failure detectors $\mathcal{P}^k$ and $\mathcal{P}^{k+1}$ are equivalent with respect to the $\succeq^{s}$ relation.

The results in [Gue01] and [BJ09] employ the $\succeq^{CT}$ relation[4] to prove that certain failure detectors cannot be the weakest ones to solve the given problem. In contrast, our results show that these conclusions drawn in [Gue01] and [BJ09] do not hold if the failure detectors

[1] We provide detailed descriptions of the $\succeq^{CT}$ and $\succeq^{JT}$ relations in Section 4.

[2] There is no real contradiction here. The reconciliation between [JT08] and [BJ09] is explained in [BJ09].

[3] We describe these failure detectors in detail and give their definitions in Section 6.

[4] Using arguments similar to the ones presented in [Gue01] and [BJ09], one can easily show that $\mathcal{M}$ and $\mathcal{P}$ are incomparable and $\mathcal{P}^k$ is strictly stronger than $\mathcal{P}^{k+1}$ with respect to the $\succeq^{JT}$ relation as well.

are compared using the $\succeq^s$ relation. Thus, different natural relations to compare failure detectors lead to significantly different results.

2 The failure detector model

We recall the basic definitions of the failure detector model [CT96]. Informally, it consists of a set of crash-prone processes that are connected via reliable asynchronous links and have access to a failure-detector oracle that provides information. In this paper, we only consider failure detectors where this information has the form of a subset of the processes in the system.

More formally, the system consists of a finite set of *processes* Π. We assume that each process p_i in Π has a link $l_{(i,j)}$ to every process p_j in Π over which *messages* can be sent. There is a discrete global time base $\mathcal{T}$, and for simplicity we assume its range of values is the natural numbers $\mathbb{N}$.

Failures and failure patterns. A *failure pattern* is a function $F\colon \mathcal{T} \to 2^{\Pi}$. This means that if $p_i \in F(t)$ then p_i has failed by time t. We consider crash faults only, and so $F(t) \subseteq F(t+1)$, for all times t. We say that p_i is live at time t if $p_i \notin F(t)$, and define the set of live processes at time t as $\mathit{live}(F,t) = \Pi \setminus F(t)$. A process p_i is correct in F if p_i is always live, that is, $p_i \in \mathit{correct}(F) = \bigcap_{t\in\mathcal{T}} \mathit{live}(F,t)$. We say processes that are not correct are faulty — or crashed — and we abbreviate $\mathit{faulty}(F) = \Pi \setminus \mathit{correct}(F)$. An *environment* $\mathcal{E}$ is defined as a non-empty set of failure patterns. In this paper, we consider the environment that consists of all failure patterns for Π.

Failure detectors. A failure detector history H is a function $H\colon \Pi \times \mathcal{T} \to 2^{\Pi}$.[5] If $\mathcal{H}$ denotes the set of all possible histories, then a *failure detector* is a function $\mathcal{D}\colon \mathcal{E} \to 2^{\mathcal{H}} \setminus \emptyset$.

States and configurations. Each process is modeled as a (possibly infinite) state machine A_i over the set of states Q_i for each process $p_i \in \Pi$. An algorithm A is a collection of all such state machines $(A_i)_{p_i\in\Pi}$. There exists a non-empty set of states $\hat{Q}_i \subseteq Q_i$ that are the *initial states* of p_i.

Each communication link $l_{(i,j)}$ is also represented by a set of states, and the state of each link $l_{(i,j)}$, denoted $s_{(i,j)}$ is the set of messages in transit from p_i to p_j. The state of a link with no messages in transit is said to be the *initial state* of the link.

The *configuration* of a system is a vector $C = (s_0, \ldots, s_{n-1}, s_{(0,0)}, s_{(0,1)}, \ldots, s_{(n-1,n-1)})$ where s_i is the state of p_i and $s_{(i,j)}$ is the state of the link $l_{(i,j)}$. Then a configuration in which all the processes and links are in initial states is called an *initial configuration*. The set of all configurations of a system is denoted $\mathcal{C}$ and the set of all initial configurations is denoted $\mathcal{I}$. The notation $C|_i$ denotes the state of p_i in configuration C, and $C|_\Pi$ denotes the vector of states of the processes in Π in configuration C. Similarly, the notation $C|_{(i,j)}$ denotes the state of the link $l_{(i,j)}$ in configuration C.

Steps. Each transition of the state machine A_i — or *step* of the process p_i — takes as input the current state s of the process, zero or one message m_r (the "received" message), and an output d from the failure detector; it produces as output a new state s' for the process and may send a message m_s to another process p_k via the corresponding communication link

[5] The failure detectors considered in this paper always output a set of processes. So we do not need the more general original definition [CT96] here.

(the "sent" message). Incidentally, the receipt of a message by a process p_i from p_j removes the message from the link $l_{(j,i)}$ and the sending of a message by p_i to p_k adds the message to the link $l_{(i,k)}$; this step can then be identified by the tuple (p_i, s, m, d, s', m'), where m is $\perp$ if no message is received and (p_j, m_r) otherwise, and similarly, m' is $\perp$ if no message is sent and (p_k, m_s) otherwise.

Schedules. A schedule Φ of an algorithm A is a sequence of steps taken by processes executing A; the ℓth step of Φ is denoted $\Phi[\ell]$. A *projection* of a schedule Φ over a process p_i is the subsequence of Φ consisting of only the steps executed by p_i and is denoted $\Phi|_i$.

Time-Sequences. A time-sequence T is a sequence of increasing values in $\mathcal{T}$; the ℓth element in T is denoted $T[\ell]$ (which represents the time at which the step $\Phi[\ell]$ occurs). Again, we define a *projection* of a time-sequence T over a process p_i as a subsequence of T consisting of only the times at which p_i executes steps and is denoted $T|_i$.

Runs. A run R of an algorithm A using a failure detector $\mathcal{D}$ is a tuple $\langle F, H, I, \Phi, T\rangle$, where F is a failure pattern, $H \in \mathcal{D}(F)$ is a failure detector history, $I \in \mathcal{I}$ is an initial configuration of A, Φ is a schedule of A, and T is a time-sequence. Run R is *valid for* A — or just *valid* for short — if correct processes take an infinite number of steps and if for each $\ell \geq 1$, the step $\Phi[\ell] \equiv (p_i, s, m, d, s', m')$ satisfies the following properties.

- The process p_i is live at time $T[\ell]$; that is, $p_i \notin F(T[\ell])$.
- d is an output of the failure detector $\mathcal{D}$ at time $T[\ell]$; formally, $d = H(p_i, T[\ell])$.
- There are no spurious messages, that is, if m is of the form (p_j, m_r), then there exists some $k < \ell$ such that m_r is a message that was sent by p_j to p_i in step $\Phi[k]$ identified by $(p_j, *, *, *, *, (p_i, m_r))$.
- Message transmission is reliable, that is, if m' is of the form (p_j, m_s), then there is at most one $k > \ell$ such that step $\Phi[k]$ is of the form $(p_j, *, (p_i, m_s), *, *, *)$. Furthermore, if p_j is correct, then there is exactly one such step.
- If $\Phi[\ell]$ is the first step of process in p_i in run R, then $s = I|_i$.
- The state of a process does not change between consecutive steps by that process; that is, if p_i takes another step, then the first step of p_i after $\Phi[\ell]$ is of the form $(p_i, s', *, *, *, *)$.

Configuration sequences induced by runs. Given a run $R = \langle F, H, I, \Phi, T\rangle$, the configuration of the system after k steps are taken is given by $\gamma(I, \Phi, k)$. The sequence $\gamma(I, \Phi, 0), \gamma(I, \Phi, 1), \ldots$ is the *configuration sequence of run* R. The state of process p_i after p_i takes k steps in the run is given by $\gamma_i(I, \Phi, k)$; if process p_i crashes and takes only k steps, then we use the convention that $\gamma_i(I, \Phi, \ell) = \gamma_i(I, \Phi, k)$ for $\ell \geq k$.

Note that if two runs share the same I and Φ (but differ, for instance, at the times steps are taken), then they induce the same configuration sequence.

3 Solving problems

We now define the notion of a *problem* and what it means to *solve* a problem. Problems traditionally depend on initial values (as in consensus [FLP85]) and transitions to certain

states depending on the initial values. So we have to define a problem by referring to problem states. Problems also depend on the correctness of processes. For instance, faulty processes are not required to make progress. In the failure-detector model, faults are modeled by failure patterns, which define after what time faulty processes must not take steps. However, before that, processes need not take steps. As we want to get rid of all time dependencies in the problem definition, it is hence natural to restrict problems by the set of processes that appear in the failure pattern rather than restricting the problems by the times at which processes appear in the failure pattern. This is done in the *crash time independence* property described later.

Moreover, as we define problems to be solvable in asynchronous systems, we have to consider the nature of runs in such systems. Since message delays and process speeds are unconstrained in asynchronous systems, processes may take finitely many idempotent or no-op steps while waiting for a message, or while waiting on some local predicate to become true. To reflect this, we require that problems are tolerant to *finite stuttering* which is described after the following preliminary definitions.

We start by defining σ as a set of *problem states*. By $\hat{\sigma}$ we denote the set of *initial problem states*, with $\hat{\sigma} \subseteq \sigma$. A *problem configuration* Σ for a system of size n is an n-dimensional vector of problem states. We denote by $\Sigma|_i$, the problem state associated with process p_i in the problem configuration Σ. A problem configuration consisting only of initial problem states is called an *initial problem configuration* $\hat{\Sigma}$. We denote $\Sigma^\star$ to be the set of all possible problem configurations, and we denote $\hat{\Sigma}^\star$ to be the set of all possible initial problem configurations; note that $\hat{\Sigma}^\star \subseteq \Sigma^\star$. We denote $W(\hat{\Sigma}^\star, \Sigma^\star)$ to be the set of all sequences of problem configurations that start with an initial problem configuration.

Further, let w_{pre} be a finite problem configuration sequence starting with an initial problem configuration, let w_{suff} be a problem configuration sequence, and let Σ and Σ' be two problem configurations. Let Σ_{mid} be any problem configuration such that for each process p_i, either $\Sigma_{mid}|_i = \Sigma|_i$ or $\Sigma_{mid}|_i = \Sigma'|_i$. Then, for any problem configuration sequence $w = w_{pre} \cdot \Sigma \cdot \Sigma' \cdot w_{suff}$, the sequence $w' = w_{pre} \cdot \Sigma \cdot \Sigma_{mid} \cdot \Sigma' \cdot w_{suff}$ is a 1-*stutter* of w denoted by $w \sqsubset_1 w'$. Inductively for each $n > 1$, we define w' to be an n-*stutter* of w, denoted by $w \sqsubset_n w'$, if there is a sequence v such that $w \sqsubset_{n-1} v \ \wedge \ v \sqsubset_1 w'$. Further, we define w' to be a *stutter* of w, denoted by $w \sqsubseteq w'$, if either $w = w'$ or there is an n, $0 < n < \infty$, such that $w \sqsubset_n w'$.

Problems. Briefly, a problem is a predicate over a problem configuration sequence that starts with an initial problem configuration, and a fault pattern. More precisely, a *time-free problem* P over $W(\hat{\Sigma}^\star, \Sigma^\star)$ in fault environment $\mathcal{E}$ — or just *problem* for short — is a predicate P on $W(\hat{\Sigma}^\star, \Sigma^\star) \times \mathcal{E}$ with the following properties:

- *Crash time independence.* For all failure patterns F and F' in $\mathcal{E}$ and for all w in $W(\hat{\Sigma}^\star, \Sigma^\star)$, $correct(F) = correct(F')$ implies $P(w, F) = P(w, F')$.
- *Finite stuttering.* For any failure pattern F, and any two problem configuration sequences w and w' in $W(\hat{\Sigma}^\star, \Sigma^\star)$, $w \sqsubseteq w'$ implies $P(w, F) = P(w', F)$.

Solving a problem Let A be an algorithm, and let a problem P be defined for $W(\hat{\Sigma}^\star, \Sigma^\star)$ and $\mathcal{E}$. Let an *interpretation* V_i be a function that maps the states Q_i of A to σ (the problem states that constitute $W(\hat{\Sigma}^\star, \Sigma^\star)$), such that the initial states of the algorithm $\hat{Q}_i$ are mapped onto $\hat{\sigma}$ (surjective). This naturally extends to a function V_Π that maps configurations $C|_\Pi$ to problem configurations. An *interpreted run* is a sequence of problem configurations obtained by applying V_Π to the configuration sequence of a valid run

$R = \langle F, H, I, \Phi, T\rangle$ of A; it is denoted by $ir(R, V_\Pi)$. Further, the set of all interpreted runs of algorithm A using $\mathcal{D}$ with failure pattern F interpreted by V_Π is denoted by $IR(A, F, \mathcal{D}, V_\Pi)$.

Algorithm A solves a problem P using failure detector $\mathcal{D}$ in environment $\mathcal{E}$, if there is a function V_Π such that for all F in $\mathcal{E}$ and any $w \in IR(A, F, \mathcal{D}, V_\Pi)$, the predicate $P(w, F)$ holds. If there is an algorithm that solves problem P using failure detector $\mathcal{D}$ we say that failure detector $\mathcal{D}$ *can be used to solve* P, or in other words P *is solvable using* $\mathcal{D}$.

The definition of a problem encompasses many common problems in distributed computing, including classic agreement problems. The set of problem states of consensus, for instance, can be defined as $\sigma = \{(p, d)\colon p \in \{0, 1\} \wedge d \in \{\perp, 0, 1\}\}$. A problem state (p, d) at process p_i signifies a state where a process p_i has p as its proposed initial value, and d is its decision; if p_i has not yet decided, then $d = \perp$, and otherwise d is p_i's final decision. The set of initial problem configurations $\hat{\Sigma}^\star$ is the set of all n-element vectors where each i-th element is a problem state of p_i and is of the form $(p, \perp) \in \sigma$. One can then naturally define the consensus properties agreement, termination, and validity as predicates on problem configuration sequences, and consensus as the conjunction of these predicates.

4 Comparison relations

Chandra-Toueg relation. We recall from [CT96, CHT96] that $\mathcal{D} \succeq^{CT} \mathcal{D}'$ is defined via failure detector transformation as follows. An algorithm $T_{\mathcal{D}\to\mathcal{D}'}$ uses $\mathcal{D}$ to maintain a variable out_i at every process p_i. This variable emulates the output of $\mathcal{D}'$ at p_i. Let O_R be the history of all the out_i variables in run R, that is, $O_R(p_i, t)$ is the value of out_i at time t in run R. Algorithm $T_{\mathcal{D}\to\mathcal{D}'}$ *transforms* $\mathcal{D}$ *into* $\mathcal{D}'$ if for every valid run $R = \langle F, H, I, \Phi, T\rangle$ of $T_{\mathcal{D}\to\mathcal{D}'}$ using $\mathcal{D}$, $O_R \in \mathcal{D}'(F)$. If such an algorithm A exists, then $\mathcal{D} \succeq^{CT} \mathcal{D}'$.

Jayanti-Toueg relation. The relation $\succeq^{JT}$, introduced in [JT08], differs from $\succeq^{CT}$ in that the notion of what it means to transform a failure detector is different from the one used in [CT96]; partly by changing the computational model. Instead of using the failure detector value at the time the step occurs, the "query mechanism" is modeled via a query to the failure detector at time t and a response from the failure detector at some time $t' > t$. Specifically, an algorithm $T_{\mathcal{D}\to\mathcal{D}'}$ uses $\mathcal{D}$ and transforms $\mathcal{D}$ to $\mathcal{D}'$ if and only if, for every valid run of $T_{\mathcal{D}\to\mathcal{D}'}$, there exists a history H of $\mathcal{D}'$ under the failure pattern of the run such that the following is true. For each process p_i, and for each query by p_i to $T_{\mathcal{D}\to\mathcal{D}'}$ which happens at some time t, $T_{\mathcal{D}\to\mathcal{D}'}$ responds with an output out at some time $t' \geq t$, and $out \in \{H(p_i, s)\colon s \in [t, t']\}$. Hence, the definition of transformation does not require maintaining a variable out_i but rather requires ensuring consistency of the query and response events.

Solvability relation. The relation $\succeq^s$, introduced in [CBHW10], states that a failure detector $\mathcal{D}$ is stronger than $\mathcal{D}'$ with respect to the solvability relation, denoted $\mathcal{D} \succeq^s \mathcal{D}'$, if $\mathcal{D}$ can be used to solve any problem solvable using $\mathcal{D}'$.

The definitions of $\succeq^{CT}$ and $\succeq^{JT}$ provide a straightforward proof technique to demonstrate the claims $\mathcal{D} \succeq^{CT} \mathcal{D}'$ and $\mathcal{D} \succeq^{JT} \mathcal{D}'$. In order to prove $\mathcal{D} \succeq^{CT} \mathcal{D}'$ or $\mathcal{D} \succeq^{JT} \mathcal{D}'$ one has to provide an algorithm $T_{\mathcal{D}\to\mathcal{D}'}$ that has the properties described above.

If $\mathcal{D} \succeq^{CT} \mathcal{D}'$ then every problem solvable with $\mathcal{D}'$ is solvable with $\mathcal{D}$ [CT96, CHT96] and thus $\succeq^s$ extends $\succeq^{CT}$. Similarly, one sees that $\succeq^s$ extends $\succeq^{JT}$ as well. However, if $\mathcal{D} \not\succeq^{CT} \mathcal{D}'$, no proof technique has been given so far to establish $\mathcal{D} \succeq^s \mathcal{D}'$.

5 New technique for proving the solvability relation

Our approach is based on the following idea. If a problem P is solvable using $\mathcal{D}'$, then there exists an algorithm A that uses $\mathcal{D}'$ and solves P. If we can transform A to another algorithm $\widetilde{A}$ such that $\widetilde{A}$ uses $\mathcal{D}$ and solves P, then we have shown that problem P is also solvable using $\mathcal{D}$. Furthermore, if we demonstrate the aforementioned result for every problem solvable using $\mathcal{D}'$, then we have shown that $\mathcal{D} \succeq^s \mathcal{D}'$.

More generally, the proof technique focuses on defining a transformation function $\mathfrak{F}$ whose domain is the set of all algorithms that use $\mathcal{D}'$ and whose range is the set of algorithms that use $\mathcal{D}$ such that if algorithm A uses $\mathcal{D}'$ to solve P, then $\mathfrak{F}(A)$ uses $\mathcal{D}$ and solves P.

In order to prove that the function $\mathfrak{F}$ actually has this desired property, we consider an arbitrary problem P solvable using $\mathcal{D}'$. We do so by considering an algorithm A that solves P using $\mathcal{D}'$. By definition, such an algorithm must exist. Moreover, there is a function V_Π which maps configurations of each valid run R of A using $\mathcal{D}'$ to a sequence of problem configurations that satisfy P. Using V_Π, we define a new function $\widetilde{V}_\Pi$ that maps the configurations of $\mathfrak{F}(A)$ to problem configurations. We then have to show that for any interpreted run $w \in IR(\mathfrak{F}(A), F, \mathcal{D}, \tilde{V}_\Pi)$, the predicate $P(w, F)$ holds.

6 Failure detectors under consideration

6.1 Definitions

In this section we define the three kinds of failure detectors that we are going to use in this paper. The *perfect failure detector* $\mathcal{P}$ was originally proposed in [CT96]. Informally, $\mathcal{P}$ eventually and permanently suspects crashed processes and never suspects live processes. More precisely, $\mathcal{P}$ is defined to ensure *strong completeness*:

$$\forall F \in \mathcal{E},\ \forall H \in \mathcal{P}(F),\ \forall p_j \in \mathit{faulty}(F),\ \forall p_i \in \mathit{correct}(F),\ \exists t' \in \mathcal{T},\ \forall t > t' \colon\ p_j \in H(p_i, t),$$

and *strong accuracy*:

$$\forall F \in \mathcal{E},\ \forall H \in \mathcal{P}(F),\ \forall t \in \mathcal{T},\ \forall p_i, p_j \in \mathit{live}(F, t) \colon\ p_j \notin H(p_i, t).$$

The *Marabout* failure detector $\mathcal{M}$ was introduced in [Gue01][6], and it always outputs the set of faulty processes. It is defined as:

$$\forall F \in \mathcal{E},\ \forall H \in \mathcal{M}(F),\ \forall t \in \mathcal{T},\ \forall p_i \in \mathit{live}(F, t) :\ H(p_i, t) = \mathit{faulty}(F).$$

The $\mathcal{P}_k$ failure detector was introduced in [BJ09] (using the notation "$\mathcal{D}_k$" which we find somewhat inconsistent with the rest of our notations). Informally, $\mathcal{P}_k$ can provide arbitrary information about processes that crash before or at time k. For correct processes and processes that crash after time k, $\mathcal{P}_k$ never suspects these processes before they crash, and $\mathcal{P}_k$ eventually and permanently suspects these processes after they crash. Formally, $\mathcal{P}_k$ satisfies the properties k-Completeness:

$$\forall F \in \mathcal{E},\ \forall H \in \mathcal{P}_k(F),\ \forall p_i, p_j \in \Pi,\ \exists t' \in \mathcal{T},\ \forall t > t' :\\ (p_j \in \mathit{live}(F, k) \wedge p_j \in \mathit{faulty}(F) \wedge p_i \in \mathit{correct}(F)) \Rightarrow p_j \in H(p_i, t),$$

[6] Although the definition printed in [Gue01] is slightly different (only failure detector outputs of correct processes instead of live processes are restricted), we claim that actually the definition given here is used in the proof sketches in [Gue01]. Otherwise, for instance, the proof sketch of [Gue01, Proposition 3.3] would fail; one could easily construct a case where a process that is going to crash in the future decides differently from a correct process.

and k-Accuracy:

$$\forall F \in \mathcal{E},\ \forall H \in \mathcal{P}_k(F),\ \forall p_i, p_j \in \Pi,\ \forall t \in \mathcal{T} : (p_j \in \mathit{live}(F,k) \wedge p_j \notin F(t)) \Rightarrow p_j \notin H(p_i,t).$$

6.2 Comparing $\mathcal{M}$ and $\mathcal{P}$.

In [Gue01] it was shown that $\mathcal{P}$ and $\mathcal{M}$ are not comparable with respect to $\succeq^{CT}$. Informally, the arguments for the result are as follows. No algorithm can tell by message exchange or from looking at the output of $\mathcal{P}$ at a certain time which processes will eventually crash (in the future), therefore $\mathcal{P} \not\succeq^{CT} \mathcal{M}$. For showing $\mathcal{M} \not\succeq^{CT} \mathcal{P}$, note that faulty processes should not be put into the set of suspected processes too early by $\mathcal{P}$, as this would violate strong accuracy. However, by strong completeness of $\mathcal{P}$, crashed processes have to be added to the set eventually. The outputs of $\mathcal{M}$ do not allow us to reconcile these two requirements. Hence, no algorithm that queries $\mathcal{M}$ can implement $\mathcal{P}$; in other words, $\mathcal{M} \not\succeq^{CT} \mathcal{P}$. Similar arguments also apply to the $\succeq^{JT}$ relation, and it can be shown that $\mathcal{M}$ and $\mathcal{P}$ are incomparable with respect to the $\succeq^{JT}$ relation as well.

In this paper, we show for the solvability relation, that $\mathcal{P} \not\succeq^{s} \mathcal{M}$ and $\mathcal{M} \succeq^{s} \mathcal{P}$. Demonstrating $\mathcal{P} \not\succeq^{s} \mathcal{M}$ is straightforward. It is sufficient to give a problem solvable using $\mathcal{M}$ and not solvable using $\mathcal{P}$. Consider the following variant of consensus, called *strong consensus*, which requires that all the correct processes have to output the input value of some unique *correct* process in the system, if there is a correct process, and otherwise output anything.

Solving this problem using $\mathcal{M}$ is straightforward. Each process sends its input to all the processes and waits for inputs from the set of processes not suspected by $\mathcal{M}$. Since the processes not suspected by $\mathcal{M}$ are the correct processes, if each process decides on the input of the correct process with the smallest ID, the problem is solved. However, as $\mathcal{P}$ does not provide information on process crashes in the future, we can show that there is no algorithm that solves strong consensus using $\mathcal{P}$. So we conclude that $\mathcal{P} \not\succeq^{s} \mathcal{M}$.

In order to establish that $\mathcal{M}$ is strictly stronger than $\mathcal{P}$, it remains to show that $\mathcal{M} \succeq^{s} \mathcal{P}$. We shall do so in Section 7 in which we introduce a general transformation *Stall-on-Suspect* that transforms any algorithm A using $\mathcal{P}$ into an algorithm $\tilde{A}$ using $\mathcal{M}$. Intuitively, Stall-on-Suspect ensures that faulty processes do not participate in the algorithm. Given an algorithm A, each process first queries $\mathcal{M}$ to determine whether it is correct or faulty. If a process p_i queries $\mathcal{M}$ and discovers that it is faulty, then p_i stops participating in the algorithm by performing only no-op steps and sends no messages until it crashes. Otherwise, process p_i follows the original algorithm A faithfully. We show in Section 7 that each valid run of the modified algorithm using $\mathcal{M}$ is indistinguishable from some valid run of the original algorithm using $\mathcal{P}$ where faulty processes crash initially, at time 0. Since, by assumption, the original algorithm solves the problem using $\mathcal{P}$, the same problem is solvable by $\mathcal{M}$ as well. Thus, we show that every problem solvable by $\mathcal{P}$ is also solvable by $\mathcal{M}$.

6.3 Comparing $\mathcal{P}_k$ failure detectors

In [BJ09], the series of $\mathcal{P}_k$ failure detectors were proposed to solve FCFS mutual exclusion. Note that various values of k instantiate different failure detectors, and it was shown in [BJ09] for all $k \geq 0$ that $\mathcal{P}_k \succeq^{CT} \mathcal{P}_{k+1}$ and $\mathcal{P}_{k+1} \not\succeq^{CT} \mathcal{P}_k$. The proof of the former is based on the observation that the trivial transformation (namely, at each step, write the current failure detector output into out_i) is sufficient to implement $\mathcal{P}_{k+1}$ using $\mathcal{P}_k$; intuitively, correctness follows because the histories of $\mathcal{P}_k$ are a strict subset of the histories of $\mathcal{P}_{k+1}$.[7]

[7] This argument is in general not sufficient to prove $\succeq^{CT}$ as shown in [CBHW10]. It works in this case, as $\mathcal{P}_k$ belongs to the class of failure detectors called "time-free" in [CBHW10]; they allow finite stuttering.

The latter ($\mathcal{P}_{k+1} \not\succeq^{CT} \mathcal{P}_k$) is established by showing that no algorithm that queries $\mathcal{P}_{k+1}$ can reliably detect if some process has crashed at time $k+1$, which is a necessary requirement to implement $\mathcal{P}_k$. Similar arguments show for all $k \geq 0$ that ($\mathcal{P}_k \succeq^{JT} \mathcal{P}_{k+1}$) and ($\mathcal{P}_{k+1} \not\succeq^{JT} \mathcal{P}_k$)

In this paper, we show for all $k \geq 0$ that ($\mathcal{P}_k \succeq^{s} \mathcal{P}_{k+1}$) $\wedge$ ($\mathcal{P}_{k+1} \succeq^{s} \mathcal{P}_k$). Demonstrating $\mathcal{P}_k \succeq^{s} \mathcal{P}_{k+1}$ is straightforward and it follows from the result $\mathcal{P}_k \succeq^{CT} \mathcal{P}_{k+1}$ from [BJ09] and the observation that $\succeq^{s}$ extends $\succeq^{CT}$ [CT96].

Therefore, it remains to be shown that $\mathcal{P}_{k+1} \succeq^{s} \mathcal{P}_k$. We do so in Section 8 using a general transformation *Delay-a-Step* which just adds a no-op step at the beginning of each execution for each algorithm. Given an algorithm A that solves some problem P using failure detector $\mathcal{P}_k$, in the delay-a-step transformation, each process p_i first executes a no-op step in which p_i neither receives nor sends any message; thereafter, p_i executes the algorithm A but queries $\mathcal{P}_{k+1}$ instead of $\mathcal{P}_k$. We show in Section 8 that each valid run of the modified algorithm using $\mathcal{P}_{k+1}$ induces an interpreted run that is also an interpreted run (with "shifted" failure pattern) of the original algorithm using $\mathcal{P}_k$. Since, by assumption, the original algorithm solves P using $\mathcal{P}_k$, problem P is solvable by $\mathcal{P}_{k+1}$ as well. Thus, we show that every problem solvable by $\mathcal{P}_k$ is also solvable by $\mathcal{P}_{k+1}$.

7 Every problem solvable using $\mathcal{P}$ is solvable using $\mathcal{M}$

7.1 Algorithmic transformation: Stall-on-Suspect

Informally, the *Stall-on-Suspect* transformation (SoS) converts an algorithm A to an algorithm $\widetilde{A}$ such that $\widetilde{A}$ at a process p_i behaves exactly like A if the failure detector at p_i does not suspect itself initially. Otherwise, $\widetilde{A}$ goes into a special stall state in which it remains for the remainder of the execution.

More precisely, the SoS transformation is defined by a function $\mathfrak{F}_{SoS}(A)$ that maps an algorithm $A = (A_i)_{\forall p_i \in \Pi}$ that uses a failure detector that outputs a list of suspected processes to a new algorithm $\widetilde{A} = (\widetilde{A}_i)_{\forall p_i \in \Pi}$. The new algorithm $\widetilde{A}$ is constructed as follows. First, for each process p_i, we add a new set of states $S_i^{\dagger}$ to the states of A_i, such that $|S_i^{\dagger}| = |\hat{Q}_i|$. The states in $S_i^{\dagger}$ are not initial states in $\widetilde{A}_i$. We define a bijective function $\mathit{stall}_i \colon \hat{Q}_i \to S_i^{\dagger}$ that maps the initial states of process p_i to states in $S_i^{\dagger}$.

The state transitions in $\widetilde{A}_i$ differ only in the transitions from initial states: If a process p_i of $\widetilde{A}_i$ is in state $q \in \hat{Q}_i$, and if the failure detector output of a step of p_i contains p_i, then p_i sends no message and goes into state $\mathit{stall}_i(q)$. Otherwise, p_i's step is the one specified by A_i. If a process p_i of $\widetilde{A}_i$ is in $s \in S_i^{\dagger}$, then p_i sends no message and remains in state s in each step.

7.2 Solving P using $\mathfrak{F}_{SoS}(A)$

Consider the algorithm $\widetilde{A} = \mathfrak{F}_{SoS}(A)$. Let $\widetilde{R} = \langle F, H, I, \widetilde{\Phi}, \widetilde{T} \rangle$ be an arbitrary valid run of $\widetilde{A}$ using $\mathcal{M}$. Let Φ and T be the schedule and time sequence obtained by removing the entries corresponding to steps of processes in $\mathit{faulty}(F)$ from $\widetilde{\Phi}$ and $\widetilde{T}$, respectively.

Proposition 7.1. *If $\widetilde{R} = \langle F, H, I, \widetilde{\Phi}, \widetilde{T} \rangle$ is a valid run of $\widetilde{A}$ using failure detector $\mathcal{M}$, then $R = \langle F, H, I, \Phi, T \rangle$ is a valid run of A using $\mathcal{M}$ where no faulty process takes a step.*

Proof. To show this proposition, one has to check that the consistency requirements of a valid run from Section 2 are met in R. Since the output of $\mathcal{M}$ at a faulty process always

suspects itself, in the first step of a faulty process in $\widetilde{A}$, the process transitions to a state in $S^\dagger$ and never sends a message. Therefore, faulty processes do not send messages in run $\widetilde{R}$ of $\widetilde{A}$. Since correct processes never suspect themselves, they take the same steps in R and $\widetilde{R}$ by construction. Consequently, R does not contain any steps in which a message from a faulty process is received. Apart from this, the consistency of R follows from the consistency of $\widetilde{R}$. □

Given a failure pattern F, let F^0 be the *initial crash scenario*, that is, the failure pattern where $F^0(0) = \mathit{faulty}(F)$ and for any $t > 0$, $F^0(t) = F^0(0)$.

Proposition 7.2. *If $R = \langle F, H, I, \Phi, T\rangle$ is a valid run of A using $\mathcal{M}$ where no faulty process takes a step, then $R^0 = \langle F^0, H, I, \Phi, T\rangle$ is a valid run of A using $\mathcal{M}$.*

Proof. We prove this proposition by showing that R^0 satisfies the consistency conditions of a valid run as specified in Section 2. Note that in R^0 all faulty processes crash at time 0; therefore, no faulty process takes a steps in R^0. Since $\mathit{correct}(F) = \mathit{correct}(F^0)$, the history H is a valid history of $\mathcal{M}$ for fault pattern F^0. Since R and R^0 share the same schedule Φ and R is a valid run of A using $\mathcal{M}$, remaining consistency conditions for R^0 follows from the consistency of R. □

From the definition of $\mathcal{M}$ and $\mathcal{P}$ one observes that in initial crash scenarios, the history of $\mathcal{M}$ is in the set of allowed histories of $\mathcal{P}$, and therefore we find:

Proposition 7.3. *If $R^0 = \langle F^0, H, I, \Phi, T\rangle$ is a valid run of A using $\mathcal{M}$, then R^0 is a valid run of A using $\mathcal{P}$.*

From the three propositions above we infer

Theorem 7.4. *For any valid run $\widetilde{R} = \langle F, H, I, \widetilde{\Phi}, \widetilde{T}\rangle$ of $\widetilde{A}$ using $\mathcal{M}$ there is a valid run $R^0 = \langle F^0, H, I, \Phi, T\rangle$ of A using $\mathcal{P}$.*

Next, we argue that if algorithm A solves problem P using $\mathcal{P}$, then $\widetilde{A}$ solves P using $\mathcal{M}$. Assuming that A solves P, there is an interpretation V_Π such that for all F in $\mathcal{E}$ and any $w \in \mathit{IR}(A, F, \mathcal{D}, V_\Pi)$, the predicate $P(w, F)$ holds. As any interpreted run of A using $\mathcal{P}$ satisfies the problem, and since by Theorem 7.4 every valid run of $\widetilde{A}$ using $\mathcal{M}$ can be mapped to a valid run of A using $\mathcal{P}$, we have to show that the mapping from $\widetilde{\Phi}$ to Φ ensures that $\widetilde{A}$ also solves the problem using $\mathcal{M}$.

To this end, we obtain $\widetilde{V}_\Pi$ by defining for each process p_i a new function $\widetilde{V}_i$ as a mapping of each state of p_i in $\widetilde{A}$ to a problem state: for states $s \in S_i^\dagger$ we define $\widetilde{V}_i(s) = V_i(\mathit{stall}_i^{-1}(s))$, and for all other states s of p_i we define $\widetilde{V}_i(s) = V_i(s)$.

As $\mathit{faulty}(F) = \mathit{faulty}(F^0)$, we just speak of faulty (or correct) processes in the following, as no confusion may occur.

Proposition 7.5. *If $\widetilde{R} = \langle F, H, I, \widetilde{\Phi}, \widetilde{T}\rangle$ is valid run of $\widetilde{A}$ using failure detector $\mathcal{M}$ and if $R^0 = \langle F^0, H, I, \Phi, T\rangle$ is a valid run of A using $\mathcal{P}$, then for any correct process p_i and for any index $\ell \geq 0$:*

$$\widetilde{V}_i(\gamma_i(I, \widetilde{\Phi}, \ell)) = V_i(\gamma_i(I, \Phi, \ell)).$$

Proof. Since, Φ is constructed from $\widetilde{\Phi}$ by deleting the no-op steps taken by faulty processes, we know that each correct process p_i follows the same sequence of states in $\widetilde{\Phi}$ and Φ. That is, $\gamma_i(I, \widetilde{\Phi}, \ell) = \gamma_i(I, \Phi, \ell)$. Since p_i is correct, p_i is never suspected by both $\mathcal{M}$ and $\mathcal{P}$. Therefore, in $\widetilde{R}$, p_i is never in any state in $S^\dagger$. Hence, for each state s that p_i is in $\widetilde{R}$, $\widetilde{V}_i(s) = V_i(s)$. In other words, $\widetilde{V}_i(\gamma_i(I, \widetilde{\Phi}, \ell)) = V_i(\gamma_i(I, \Phi, \ell))$. □

Proposition 7.6. *If $\widetilde{R} = \langle F, H, I, \widetilde{\Phi}, \widetilde{T}\rangle$ is valid run of $\widetilde{A}$ using failure detector $\mathcal{M}$ and if $R^0 = \langle F^0, H, I, \Phi, T\rangle$ is a valid run of A using $\mathcal{P}$, then for any faulty process p_i and for any index $\ell \geq 0$:*

$$\widetilde{V}_i(\gamma_i(I, \widetilde{\Phi}, \ell)) = V_i(\gamma_i(I, \Phi, \ell)).$$

Proof. Since faulty processes do not take any steps in R^0, we know that for each faulty process p_i, and each index $\ell \geq 0$ in run R^0, $\gamma_i(I, \Phi, \ell) = I|_i$.

In run $\widetilde{R}$, we know from the construction of algorithm $\widetilde{A}$ that each faulty process p_i, initially, in state $\hat{q}_i \in \hat{Q}_i$, enters a state $s_i^\dagger \in S_i^\dagger$ in its first step where $s_i^\dagger = \mathit{stall}(\hat{q}_i)$, and remains there until it crashes. Therefore, for each faulty process p_i, and each index $\ell \geq 0$ in run $\widetilde{R}$, $\gamma_i(I, \Phi, \ell) \in \{\hat{q}_i, s_i^\dagger\}$.

From the definition of $\widetilde{V}_i$, we know that $\widetilde{V}_i(\hat{q}_i) = V_i(\hat{q}_i)$, and $\widetilde{V}_i(s_i^\dagger) = V_i(\mathit{stall}_i^{-1}(s))$. As $s_i^\dagger = \mathit{stall}(\hat{q}_i)$, we obtain $\widetilde{V}_i(s_i^\dagger) = V_i(\hat{q}_i)$. Therefore, for each faulty process p_i, and each index $\ell \geq 0$ in run $\widetilde{R}$, $\widetilde{V}_i(\gamma_i(I, \Phi, \ell)) = V_i(\hat{q}_i)$.

Since each process p_i is in the same initial state in $\widetilde{R}$ and R^0, we have $\hat{q}_i = I|_i$. Therefore, $\widetilde{V}_i(\gamma_i(I, \Phi, \ell)) = V_i(\hat{q}_i) = V_i(\gamma_i(I, \Phi, \ell))$. □

Theorem 7.7. *If A solves P using $\mathcal{P}$ then $\widetilde{A} = \mathfrak{F}_{SoS}(A)$ solves P using $\mathcal{M}$.*

Proof. Since A solves P using $\mathcal{P}$, we know that there exists a function V_Π such that for any $w \in \mathit{IR}(A, F, \mathcal{P}, V_\Pi)$, the predicate $P(w, F)$ is true.

Let $\widetilde{A} = \mathfrak{F}_{SoS}(A)$, and let $\widetilde{R} = \langle F, H, I, \widetilde{\Phi}, \widetilde{T}\rangle$ be an arbitrary valid run of $\widetilde{A}$ using $\mathcal{M}$. Let $R^0 = \langle F^0, H, I, \Phi, T\rangle$ be a valid run of A using $\mathcal{P}$ where $\forall t \in \mathcal{T} : F^0(t) = \mathit{faulty}(F)$, Φ and T are obtained by deleting the entries associated with faulty processes in $\widetilde{\Phi}$ and $\widetilde{T}$, respectively. From Propositions 7.1, 7.2 and 7.3, we know that R^0 is a valid run of A using $\mathcal{P}$. Therefore, each $w \in \mathit{IR}(A, F^0, \mathcal{P}, V_\Pi)$ satisfies $P(w, F^0)$.

Let $\widetilde{V}_\Pi$ be a function derived from V_Π as described earlier in this section. From Propositions 7.5 and 7.6, we conclude that for all processes p_i and all indexes ℓ in runs $\widetilde{R}$ and R^0, $\widetilde{V}_i(\gamma_i(I, \Phi, \ell)) = V_i(\gamma_i(I, \Phi, \ell))$. Note that there is no re-ordering of steps of correct processes between Φ and $\widetilde{\Phi}$; however, steps of faulty processes may be missing in R^0. Thus, we infer $\mathit{ir}(R^0, V_\Pi) \sqsubseteq \mathit{ir}(\widetilde{R}, V_\Pi)$. From the finite stuttering property of problems and Theorem 7.4, we conclude that if A solves P using $\mathcal{P}$ then $\widetilde{A} = \mathfrak{F}_{SoS}(A)$ solves P using $\mathcal{M}$. □

Corollary 7.8. *$\mathcal{M} \succeq^s \mathcal{P}$ and $\mathcal{P} \not\succeq^s \mathcal{M}$.*

8 Equivalence among $\mathcal{P}_k$ failure detectors

8.1 Algorithmic transformation: Delay-a-Step

Informally, the *Delay-a-Step* transformation (DaS) converts an algorithm A to an algorithm $\widetilde{A}$ such that in $\widetilde{A}$ each process p_i first executes a single no-op step, and subsequently p_i behaves exactly like it does in A. We define a transformation function $\mathfrak{F}_{DaS}$ that maps an algorithm $A = (A_i)_{\forall p_i \in \Pi}$ to a new algorithm $\widetilde{A} = (\widetilde{A}_i)_{\forall p_i \in \Pi}$. The new state space of $\widetilde{A}$ is constructed as follows. For each process p_i, we add a new set of states $S_i^\star$, which are the initial states of $\widetilde{A}_i$, such that $|S_i^\star| = |\hat{Q}_i|$, to obtain the set of states for $\widetilde{A}_i$. This implies that the states in $\hat{Q}_i$ are *not* initial states of $\widetilde{A}_i$. We define a bijective function $\mathit{delay}_i \colon S_i^\star \to \hat{Q}_i$.

The state transitions of $\widetilde{A}$ are the state transition of A and the following rules for initial states $S_i^\star$: if a process p_i is in state $s \in S_i^\star$ when it takes a step, then p_i neither receives nor sends messages and goes into state $\mathit{delay}_i(s)$.

8.2 Showing $\mathcal{P}_{k+1}$ is at least as strong as $\mathcal{P}_k$

Let A be an algorithm that solves some problem P using a failure detector $\mathcal{P}_k$, and let $\widetilde{A} = \mathfrak{F}_{DaS}(A)$. The remainder of this section shows that $\widetilde{A}$ solves P using the failure detector $\mathcal{P}_{k+1}$.

Let $\widetilde{R} = \langle \widetilde{F}, \widetilde{H}, \widetilde{I}, \widetilde{\Phi}, \widetilde{T} \rangle$ be a valid run of $\widetilde{A}$ using $\mathcal{P}_{k+1}$. In the following, we construct (in several steps) a new initial configuration I, a new schedule Φ, a new time-sequence T, a new failure pattern F, and a new history H such that the run $R = \langle F, H, I, \Phi, T \rangle$ is a valid run of A using the failure detector $\mathcal{P}_k$. We then show that if R is a valid run of A using $\mathcal{P}_k$, then $\widetilde{A}$ solves problem P using $\mathcal{P}_{k+1}$.

First, we construct the initial configuration I as follows. For each process p_i, $I|_i = \mathit{delay}_i(\widetilde{I}|_i)$.

Next, we construct the new schedule Φ and a new time-sequence T' as follows. For each process $p_i \in \Pi$, let $\mathit{no\text{-}op}(i)$ denote the index of the first entry of the form $(p_i, *, *, *, *, *)$ in $\widetilde{\Phi}$. The schedule Φ is obtained by deleting for each process p_i the step $\widetilde{\Phi}[\mathit{no\text{-}op}(i)]$ from $\widetilde{\Phi}$. A time-sequence T' is obtained by deleting for each process p_i the entry $\widetilde{T}[\mathit{no\text{-}op}(i)]$ from $\widetilde{T}$.

Proposition 8.1. *If $\widetilde{R} = \langle \widetilde{F}, \widetilde{H}, \widetilde{I}, \widetilde{\Phi}, \widetilde{T} \rangle$ is a valid run of $\widetilde{A}$ using $\mathcal{P}_{k+1}$ then $R' = \langle \widetilde{F}, \widetilde{H}, I, \Phi, T' \rangle$ is a valid run of A using $\mathcal{P}_{k+1}$.*

Proof. By construction, the first step of each process p_i in $\widetilde{A}$ is of the form $(p_i, *, \perp, d, *, \perp)$, and all the subsequent steps of p_i are the same as in A. Since $\widetilde{\Phi}$ is a schedule of $\widetilde{A}$, we see that for each process p_i, $\widetilde{\Phi}[\mathit{no\text{-}op}(i)]$ is the first step of p_i executing $\widetilde{A}_i$, and is therefore a no-op step of the form $(p_i, *, \perp, d, *, \perp)$. Also, note that upon executing a no-op step from state $\widetilde{I}|_i$, process p_i transitions to state $\mathit{delay}_i(\widetilde{I}|_i)$ which, by construction, is equal to the state $I|_i$.

Hence, by deleting the $\widetilde{\Phi}[\mathit{no\text{-}op}(i)]$ step for each process p_i from $\widetilde{\Phi}$, we obtain a valid schedule for A; that is, Φ is a valid schedule for a run of A. Similarly, by deleting the times at which the $\widetilde{\Phi}[\mathit{no\text{-}op}(i)]$ step occurred for each process p_i from $\widetilde{T}$, we obtain a valid time-sequence for A; that is, T' is a valid time-sequence for the schedule Φ in a run of A. The proposition follows. □

Then we define the new failure pattern F by $F(t) = \widetilde{F}(t+1)$, for $t \in \mathcal{T}$. Intuitively, each faulty process crashes one time unit earlier in F than in $\widetilde{F}$. Similarly, the new history H is defined by $H(p_i, t) = \widetilde{H}(p_i, t+1)$, for all $p_i \in \Pi$ and $t \in \mathcal{T}$.

Proposition 8.2. *If $\widetilde{H} \in \mathcal{P}_{k+1}(\widetilde{F})$ then $H \in \mathcal{P}_k(F)$.*

Proof. Since $\widetilde{H} \in \mathcal{P}_{k+1}(\widetilde{F})$, it follows from k-Accuracy that

$$\forall p_i, p_j \in \Pi,\ \forall t \in \mathcal{T} : (p_j \notin \widetilde{F}(k+1) \wedge p_j \in \widetilde{H}(p_i, t+1)) \Rightarrow p_j \in \widetilde{F}(t+1), \tag{1}$$

and from k-Completeness

$$\forall p_i, p_j \in \Pi,\ \exists t' \in \mathcal{T},\ \forall t > t' : \\ (p_j \notin \widetilde{F}(k+1) \wedge p_j \notin \mathit{correct}(\widetilde{F}) \wedge p_i \in \mathit{correct}(\widetilde{F})) \Rightarrow p_j \in \widetilde{H}(p_i, t). \tag{2}$$

Since $\forall t \in \mathcal{T} : F(t) = \widetilde{F}(t+1)$, and $\forall p_i \in \Pi, \forall t \in \mathcal{T} : H(p_i, t) = \widetilde{H}(p_i, t+1)$, substituting these functions in Equations (1) and (2) we obtain

$$\forall p_i, p_j \in \Pi,\ \forall t \in \mathcal{T}\ : (p_j \notin F(k) \wedge p_j \in H(p_i, t)) \Rightarrow p_j \in F(t), \tag{3}$$

and since $correct(F) = correct(\widetilde{F})$,

$$\forall p_i, p_j \in \Pi,\ \exists t' \in \mathcal{T},\ \forall t > t' :$$
$$(p_j \notin F(k) \wedge p_j \notin correct(F) \wedge p_i \in correct(F)) \Rightarrow p_j \in H(p_i, t). \quad (4)$$

We observe that the failure detector whose histories are as described in Equations (3) and (4) satisfies k-Accuracy and k-Completeness. □

Because T' is obtained by removing the time of the first step of each process, it follows that for any ℓ, $T'[\ell] > 0$. We may thus define the new time-sequence T as $T[\ell] = T'[\ell] - 1$ with $\ell \in \mathbb{N}$.

Proposition 8.3. *If $R' = \langle \widetilde{F}, \widetilde{H}, I, \Phi, T' \rangle$ is a valid run of A using $\mathcal{P}_{k+1}$, then $R = \langle F, H, I, \Phi, T \rangle$ is a valid run of A using $\mathcal{P}_k$.*

Proof. From the construction of T, we know that in run R, each process p_i takes the same steps as in R', but each step taken at time t in R' is taken at time $t-1$ in R. From the construction of H, we see that the output of the failure detector queried in run R' at a time t is identical to the output of the failure detector queried in run R at time $t-1$. Similarly, in the failure pattern F, each process that crashes at time t in $\widetilde{F}$ crashes at time $t-1$ in F. Therefore, the run R is the run R' after every step and the associated failure detector output in R' is moved earlier in time by 1 unit.

Also, recall that $H \in \mathcal{P}_k(F)$, from Proposition 8.2. Therefore, if R' is a valid run of A using failure detector $\mathcal{P}_{k+1}$, then R is a valid run of A using $\mathcal{P}_k$. □

As A solves P using $\mathcal{P}_k$, for each process p_i there exists a function V_i that maps each state of p_i to a problem state. For each process p_i we define a new function $\widetilde{V}_i$ as follows. For each (initial) state $s \in S_i^\star$, $\widetilde{V}_i(s) = V_i(delay_i(s))$, and for each state $s \notin S_i^\star$, $\widetilde{V}_i(s) = V_i(s)$.

Theorem 8.4. *If A solves problem P using failure detector $\mathcal{P}_k$, then Algorithm $\widetilde{A}$ solves problem P using failure detector $\mathcal{P}_{k+1}$.*

Proof. Let $\widetilde{R} = \langle \widetilde{F}, \widetilde{H}, \widetilde{I}, \widetilde{\Phi}, \widetilde{T} \rangle$ be a valid, run of $\widetilde{A}$ using $\mathcal{P}_{k+1}$. Applying Propositions 8.1, 8.2, and 8.3, we see that from $\widetilde{R}$ we can construct a unique run $R = \langle F, H, I, \Phi, T \rangle$ that is a valid run of A using $\mathcal{P}_k$.

Note that by assumption A solves problem P using failure detector $\mathcal{P}_k$. Hence there is an interpretation V_Π which ensures that $P(ir(R, V_\Pi), F)$ holds. Since $correct(F) = correct(\widetilde{F})$, applying the crash time independence property from Section 3, we obtain that $P(ir(R, V_\Pi), \widetilde{F})$ is true.

Note that for each process p_i, p_i is never is a state $s_i \in S_i^\star$ in run R, and for each state $s \notin S_i^\star$, $\widetilde{V}_i(s) = V_i(s)$. Therefore, $ir(R, V_\Pi) = ir(R, \widetilde{V}_\Pi)$.

Also, note that for each process p_i, for each state $s \in S_i^\star$, $\widetilde{V}_i(s) = V_i(delay_i(s))$, and $delay_i(s) \in \hat{Q}_i$; therefore, $\widetilde{V}_i(delay_i(s)) = V_i(delay_i(s))$. In effect, $ir(R, \widetilde{V}_\Pi) \sqsubseteq ir(\widetilde{R}, \widetilde{V}_\Pi)$. So we apply the finite stutter property from Section 3 and see that since $P(ir(R, V_\Pi), F)$ is true, $P(ir(\widetilde{R}, \widetilde{V}_\Pi), \widetilde{F})$ is also true.

We thus have shown that for any interpreted run $w \in IR(\widetilde{A}, \widetilde{F}, \mathcal{P}_{k+1}, \widetilde{V}_\Pi)$, the predicate $P(w, \widetilde{F})$ holds. In other words, $\widetilde{A}$ solves P using failure detector $\mathcal{P}_{k+1}$. □

Corollary 8.5. *$\mathcal{P}_k \succeq^s \mathcal{P}_{k+1}$ and $\mathcal{P}_{k+1} \succeq^s \mathcal{P}_k$.*

9 Conclusion

In this paper, we introduced a new proof technique that compares failure detectors and does not depend on the ability of one failure detector to implement another. Instead, we propose a novel approach which is based on algorithm transformation so that for every algorithm A that solves some problem using failure detector $\mathcal{D}'$ we derive a new algorithm $\widetilde{A}$ which solves the same problem using $\mathcal{D}$ instead, and thus we show $\mathcal{D} \succeq^s \mathcal{D}'$, where $\succeq^s$ is the solvability relation introduced in [CBHW10].

We demonstrated the utility of the new proof technique by presenting two new results. First, we showed that the $\mathcal{P}$ and $\mathcal{M}$ failure detectors, which are incomparable with respect to the $\succeq^{CT}$ and $\succeq^{JT}$ relations, are strictly ordered with respect to the $\succeq^s$ relation; $\mathcal{M}$ is strictly stronger than $\mathcal{P}$. Second, we showed that the $\mathcal{P}_k$ series of failure detectors (denoted by $\mathcal{D}_k$ in [BJ09]), which were shown to be strictly ordered as $\mathcal{P}_k \succeq^{CT} \mathcal{P}_{k+1}$ for all k, are equivalent to each other with respect to the $\succeq^s$ relation.

Significance. The primary motivation for the introduction of the $\mathcal{M}$ failure detector in [Gue01] was to show that $\mathcal{P}$ is not the weakest failure detector for certain problems such as non-blocking atomic commitment or terminating reliable broadcast. This was done by showing that $\mathcal{M}$ and $\mathcal{P}$, despite being incomparable with respect to $\succeq^{CT}$, can be used to solve the aforementioned problems under consideration. However, we have shown that $\mathcal{M}$ and $\mathcal{P}$ can be strictly ordered with respect to $\succeq^s$. This shows that the reasoning used in [Gue01] is limited only to the $\succeq^{CT}$ relation.[8]

Similarly, the $\mathcal{P}_k$ sequence of failure detectors was introduced in [BJ09] in order to demonstrate that FCFS mutual exclusion does not have a weakest failure detector. The proof relies on the fact that for any k, $\mathcal{P}_k$ is strictly stronger than $\mathcal{P}_{k+1}$ with respect to $\succeq^{CT}$ while every such $\mathcal{P}_k$ is sufficient to solve FCFS mutual exclusion. However, we have shown that all the $\mathcal{P}_k$ failure detectors are equivalent with respect to $\succeq^s$ and, therefore, these failure detectors solve the same set of time-free problems.

The above two examples show that some results on weakest failure detectors based on the $\succeq^{CT}$ and $\succeq^{JT}$ relation do not carry over to the $\succeq^s$ relation. This, in conjunction with the seemingly contradictory results regarding the (non)existence of weakest failure detectors in [JT08] and [BJ09], leaves open the possibility that the use of failure detectors as "computability benchmark" [FGK11] may not be appropriate until we have resolved the question of the "right" comparison relation to order failure detectors.

Comparison to standard proofs. From a technical viewpoint, our new proof technique is quite similar to proofs that establish the $\succeq^{CT}$ relation. In both, one argues about an algorithm using some failure detector. In $\succeq^{CT}$ proofs, one usually gives an algorithm more or less explicitly, while we give an algorithm $\widetilde{A}$ as function of another algorithm A. In $\succeq^{CT}$ proofs, one shows that the states the algorithm goes through are related to histories of the implemented failure detector. In our $\succeq^s$ proofs, we show that the states the algorithm goes through are related to problem configuration sequences.

The differences in the comparison relations discussed above then come from the fact that we relate to a schedule of algorithm A which is within the world of asynchronous runs,

[8] It was later shown in [Lar03] that failure detectors that are weaker with respect to $\succeq^{CT}$ than both $\mathcal{M}$ and $\mathcal{P}$ are sufficient to solve non-blocking atomic commitment and terminating reliable broadcast. However, our motivation was not to find a weakest failure detector for a given problem, but rather to make explicit that certain proofs are limited to the $\succeq^{CT}$ relation.

while $\succeq^{CT}$ proofs relate to a failure detector history, which is defined with respect to time, and is hence outside the world of asynchronous runs.

Future Work. Our results are preliminary and provide multiple avenues for future work. We present two such open questions. First, note that the proof technique introduced here does not necessarily characterize the $\succeq^s$ relation completely. That is, there might be other proof techniques which establish the $\succeq^s$ relation between two failure detectors in the cases where our proposed technique does not lead to the required result. Thus, there is scope for complete characterization of the $\succeq^s$ relation. Second, since different comparison relations establish different relationships among various failure detectors, an obvious question presents itself: is there a "right" comparison relation for failure detectors? If yes, which one is it?

Acknowledgement. We would like to thank Jennifer Welch and Martin Hutle for their comments, suggestions, and criticisms that greatly helped improve this article.

References

[BJ09] Vibhor Bhatt and Prasad Jayanti. On the existence of weakest failure detectors for mutual exclusion and k-exclusion. In *Proceedings of the 23rd International Symposium on Distributed Computing*, pages 311–325, 2009.

[CBHW10] Bernadette Charron-Bost, Martin Hutle, and Josef Widder. In search of lost time. *Information Processing Letters*, 110(21), 2010.

[CHT96] Tushar Deepak Chandra, Vassos Hadzilacos, and Sam Toueg. The weakest failure detector for solving consensus. *Journal of the ACM*, 43(4):685–722, 1996.

[CT96] Tushar Deepak Chandra and Sam Toueg. Unreliable failure detectors for reliable distributed systems. *J. ACM*, 43(2):225–267, 1996.

[FGK11] Felix C. Freiling, Rachid Guerraoui, and Petr Kuznetsov. The failure detector abstraction. *ACM Comput. Surv.*, 43:9:1–9:40, February 2011.

[FLP85] Michael J. Fischer, Nancy A. Lynch, and Michael S. Paterson. Impossibility of distributed consensus with one faulty process. *Journal of the ACM*, 32(2):374–382, 1985.

[FR03] Faith Fich and Eric Ruppert. Hundreds of impossibility results for distributed computing. *Distributed Computing*, 16(2-3):121–163, 2003.

[Gue01] Rachid Guerraoui. On the hardness of failure-sensitive agreement problems. *Information Processing Letters*, 79(2):99–104, 2001.

[JT08] Prasad Jayanti and Sam Toueg. Every problem has a weakest failure detector. In *Proceedings of the 27^{th} ACM symposium on Principles of distributed computing (PODC)*, pages 75–84, 2008.

[Lar03] Mikel Larrea. On the weakest failure detector for hard agreement problems. *Journal of Systems Architecture*, 49(7-9):345 – 353, 2003.