THE WEAKEST FAILURE DETECTORS TO SOLVE QUITTABLE CONSENSUS AND NONBLOCKING ATOMIC COMMIT

RACHID GUERRAOUI†, VASSOS HADZILACOS‡, PETR KUZNETSOV§, AND SAM TOUEG‡

Abstract. We define quittable consensus, a natural variation of the consensus problem, where processes have the option to agree on “quit” if failures occur, and we relate this problem to the well-known problem of nonblocking atomic commit. We then determine the weakest failure detectors for these two problems in all environments, regardless of the number of faulty processes.

Key words. asynchronous distributed computing, atomic commit problem, consensus, fault-tolerant algorithms, failure detection

AMS subject classifications. 68Q17, 68Q25, 68Q85

DOI. 10.1137/070698877

1. Introduction. Nonblocking atomic commit (NBAC) is a well-known problem that arises in distributed transaction processing [24]. Informally, the set of processes that participate in a transaction must agree on whether to commit or abort that transaction. Initially each process votes Yes (“I am willing to commit”) or No (“we must abort”), and eventually processes must reach a common decision, Commit or Abort. The decision to Commit can be reached only if all processes voted Yes. Furthermore, if all processes voted Yes and no failure occurs, then the decision must be Commit. NBAC is similar to the classical problem of consensus, where each process initially proposes a value, and eventually processes must reach a common decision on one of the proposed values.

It is well known that NBAC and consensus are unsolvable in asynchronous systems with process crashes (even if communication is reliable) [20]. One way to circumvent such impossibility results is through the use of unreliable failure detectors [10]. Intuitively, a failure detector provides each process with some (possibly incomplete and inaccurate) information about failures, e.g., a list of processes currently suspected to have crashed.

Failure detectors can be compared by “reduction”: Intuitively, failure detector $D$ is weaker than failure detector $D'$ if there is an algorithm that transforms $D'$ into $D$. Note that if $D$ is weaker than $D'$, any problem that can be solved with $D$ can also be solved with $D'$. For any problem $P$, a natural question is to determine the weakest failure detector to solve $P$, i.e., to determine the failure detector $D^*$ such that (a) there is an algorithm that uses $D^*$ to solve $P$, and (b) $D^*$ is weaker than any failure detector $D$ that can be used to solve $P$. Finding the weakest failure detector to solve a problem $P$ amounts to determining the minimum amount of information about failures that is necessary to solve $P$. It also provides important intuition about systems in which $P$ is solvable: $P$ is solvable in any system where the weakest failure detector
for \( P \) can be implemented. Such a system may be defined in terms of partial synchrony assumptions, or in terms of other assumptions, e.g., the number and timing of failures.

Chandra, Hadzilacos, and Toueg [9] determined the weakest failure detector to solve consensus in systems with a majority of correct processes, while Delporte-Gallet, Fauconnier, and Guerraoui [15] generalized this result to all systems, regardless of the number of correct processes.

As with consensus, failure detectors can be used to solve NBAC [25, 22]. It was an open problem, however, whether there is a weakest failure detector to solve NBAC and, if so, what that failure detector is. In this paper we resolve this problem. To do so,

(a) we define a natural variation of consensus, called quittable consensus (QC);
(b) we establish a close relationship between QC and NBAC;
(c) we determine the weakest failure detector to solve QC; and
(d) we use (b) and (c) to derive the weakest failure detector to solve NBAC.

Informally, QC is like consensus except that, if a failure occurs, processes have the option (but not the obligation) to agree on a special value \( Q \) (for “quit”). This weakening of consensus is appropriate for applications where, if a failure occurs, processes are allowed to agree on that fact (rather than on an input value) and resort to a default action.

Despite their apparent similarity, QC and NBAC are different in important ways. In NBAC the two possible input values, Yes and No, are not symmetric: A single vote of No is enough to force the decision to abort. In contrast, in QC (as in consensus) no input value has a privileged role. Another way in which the two problems differ is that the semantics of the decision to abort (in NBAC) and the decision to quit (in QC) are different. In NBAC the decision to abort is sometimes inevitable (e.g., if a process crashes before voting); in contrast, in QC the decision to quit is never inevitable, it is only an option. Moreover, in NBAC the decision to abort signifies that either a failure occurred or someone voted No; in contrast, in QC the decision to quit is allowed only if a failure occurred.

We now describe in more detail our results, which involve the following three failure detectors:

- The leader failure detector \( \Omega \) outputs the identity (id) of a process at each process. If there is a correct process, then there is a time after which \( \Omega \) outputs the id of the same correct process at all correct processes [9].
- The quorum failure detector \( \Sigma \) outputs a set of processes at each process. Any two sets (output at any times and by any processes) intersect, and eventually every set output at any correct process consists of only correct processes [15].
- The failure signal failure detector \( FS \) outputs green or red at each process. As long as there are no failures, \( FS \) must output green at every process; once a failure occurs, and only if it does, \( FS \) must eventually output red permanently at every correct process [12, 25].

We show that there is a weakest failure detector to solve QC. This failure detector, which we denote \( \Psi \), is closely related to the weakest failure detector to solve consensus, namely \((\Omega, \Sigma)\) [15], and to \( FS \). Intuitively, \( \Psi \) behaves as follows: For an initial period of time the output of \( \Psi \) at each process is \( \bot \). Eventually, however, \( \Psi \) behaves either like the failure detector \((\Omega, \Sigma)\) at all correct processes or like the failure detector \( FS \) at all
correct processes. The switch from $\perp$ to $(\Omega, \Sigma)$ or $\mathcal{FS}$ need not occur simultaneously at all processes, but the same choice is made by all processes. Furthermore, $\Psi$ can switch from $\perp$ to $\mathcal{FS}$ only if a failure occurred. This result has an intuitively appealing interpretation: To solve QC, a failure detector must eventually either truthfully inform all the correct processes that a failure occurred, in which case they can decide $Q$, or be powerful enough to allow processes to solve consensus on their proposed values. This matches the behavior of $\Psi$.

We also prove that NBAC is in some sense equivalent to QC modulo the failure detector $\mathcal{FS}$. Intuitively, (a) given $\mathcal{FS}$, any QC algorithm can be converted to an algorithm for NBAC, and (b) any algorithm for NBAC can be converted to an algorithm for QC, and can also be used to implement $\mathcal{FS}$.

Using this equivalence we prove that $(\Psi, \mathcal{FS})$ is the weakest failure detector to solve NBAC. This result applies to any system, regardless of the number of faulty processes.

Related work. The model of asynchronous systems augmented with failure detectors was introduced in [10] as one way to circumvent the impossibility result of [20]. Chandra, Hadzilacos, and Toueg [9] proved that $\Omega$ is the weakest failure detector to solve consensus in systems with a majority of correct processes. Delporte-Gallet, Fauconnier, and Guerraoui [15] generalized this result to prove that $(\Omega, \Sigma)$ is the weakest failure detector to solve consensus in any system, regardless of the number of correct processes. Failure detectors have been used to capture the minimum information about failures that is necessary to solve other basic problems in distributed computing, such as set agreement [18, 38, 23], mutual exclusion [17], boosting obstruction-freedom to wait-freedom [26], implementing an atomic register in message-passing systems [15], and implementing uniform reliable broadcast [5, 30] in systems with lossy communication links. It is worth noting that the result that $\Omega$ is the weakest failure detector for solving consensus led to the discovery of several consensus algorithms for other important models, in particular, for several weak models of partial synchrony (e.g., see [2, 1, 34, 31, 6]). This was done by implementing $\Omega$ in such systems, and then combining this implementation with any algorithm that solves consensus using $\Omega$, thus exploiting the modularity of the failure detector approach.

The NBAC problem has been studied extensively in the context of transaction processing [24, 37]. Its relation to consensus was first explored in [28]. Charron-Bost and Toueg [12] and Guerraoui [25] showed that despite some apparent similarities, in asynchronous systems NBAC and consensus are in general incomparable—i.e., a solution for one problem cannot be used to solve the other.\footnote{An exception is the case where at most one process may fail. In this case, any algorithm that solves NBAC can be converted into one that solves consensus, but the reverse does not hold.} The problem of determining the weakest failure detector to solve NBAC was explored and settled in special settings. Fromentin, Raynal, and Tronel [22] determine that to solve NBAC between every pair of processes in the system, one needs a perfect failure detector [10]. Guerraoui and Kouznetsov [27] determine the weakest failure detector for NBAC for a restricted class of failure detectors. From results of [12] and [25] it follows that in the special case where at most one process may crash, $\mathcal{FS}$ is the weakest failure detector to solve NBAC. The general question, however, remained open until our results appeared, in preliminary form, in [16].\footnote{That paper contained additional results by Delporte, Fauconnier, and Guerraoui, which have since appeared in full form in [15].}

Quittable consensus is closely related to the detectable broadcast problem introduced and studied by Fitzi et al. [21] in a different setting, namely, synchronous
systems with arbitrary process failures. Roughly speaking, in the detectable broadcast problem, correct processes either agree on the broadcast value or, if failures occur, may agree to “reject” the broadcast; furthermore, if any correct process rejects the broadcast, then the “adversary gets no information about the sender’s input”—a privacy requirement that is relevant in the case of arbitrary failures.

Quittable consensus is also related to the abortable consensus problem that Chen [13] defined in the context of message-passing systems with probabilistic message delays and losses. Roughly speaking, in abortable consensus some processes are allowed to abort when the behavior of the system degenerates (e.g., there are many process failures or message delays or losses). In contrast to quittable consensus, however, abortable consensus does not require agreement: some processes may decide the same value while others abort.

Other weakenings of the consensus problem were studied in the context of obstruction-free object implementations in shared-memory systems. For example, Attiya et al. [7] defined objects that may reply with a special value “pause” or “fail” to some processes if there is step contention. Similarly, Aguilera et al. [3] defined abortable objects that may return “abort” in the event of interval contention. In both works, when the object is consensus, agreement in not required: Some processes may “pause” or “abort” while others agree on the same value. Furthermore, pausing, failing, or aborting is allowed when there is contention, not failures. In contrast, in quittable consensus the decision to quit must be agreed on by all processes and is allowed only in the case of failures.

Roadmap. The rest of the paper is organized as follows: In section 2 we review the model of computation. Sections 3 and 4 contain the precise specifications of the failure detectors used in this paper, and of QC and NBAC, the two problems we consider. In section 5 we show that QC and NBAC are closely related. In that section we also identify the weakest failure detector Ψ to solve QC and prove that (Ψ,F) is the weakest failure detector to solve NBAC. In section 6 we show that Ψ is sufficient to solve QC. Sections 7 and 8 contain the proof that Ψ is necessary to solve QC. We conclude with some final remarks in section 9.

2. The model. Our model of asynchronous computation is the one described in [9], which augments the model of Fischer, Lynch, and Paterson [20] with failure detectors. Henceforth, we assume a discrete global clock to which the processes do not have access. The range of this clock’s ticks is \( \mathbb{N} \).

2.1. Systems. We consider distributed message-passing systems with a set of \( n \geq 2 \) processes \( \Pi = \{1, 2, \ldots, n\} \). Processes execute steps of computation asynchronously; i.e., there is no bound on the delay between steps. (Section 2.4 describes what a process does in each step.) Each pair of processes is connected by a reliable link. The links transmit messages with finite but unbounded delay. They are modeled as a set \( M \), called the message buffer, that contains triples of the form \((p, \text{data}, q)\) indicating that \( p \) has sent the message \text{data} to \( q \), and \( q \) has not yet received it. We assume that each message sent by a process \( p \) to a process \( q \) is unique; this can be guaranteed by having the sender include a counter with each message.

2.2. Failures, failure patterns, and environments. We consider crash failures only: Processes fail only by halting prematurely. A failure pattern is a function \( F : \mathbb{N} \rightarrow 2^{\Pi} \), where \( F(t) \) is the set of processes that have crashed through time \( t \). Since processes never recover from crashes, \( F(t) \subseteq F(t + 1) \). Let \( \text{faulty}(F) = \bigcup_{t \in \mathbb{N}} F(t) \) be the set of faulty processes in a failure pattern \( F \), and let \( \text{correct}(F) = \Pi - \text{faulty}(F) \) be
the set of correct processes in $F$. When the failure pattern $F$ is clear from the context, we say that process $p$ is correct if $p \in \text{correct}(F)$, and $p$ is faulty if $p \in \text{faulty}(F)$.

An environment, denoted $E$, is a set of failure patterns. Intuitively, an environment $E$ describes the number and timing of failures that can occur in the system. Thus, a result that applies to all environments is one that holds regardless of the number and timing of failures. We denote by $E^*$ the set of all failure patterns. Intuitively, in a system with environment $E^*$ each process may crash, and it may do so at any time.

### 2.3. Failure detectors

A failure detector history $H$ with range $\mathcal{R}$ describes the behavior of a failure detector during an execution. Formally, it is a function $H : \Pi \times \mathbb{N} \rightarrow \mathcal{R}$, where $H(p,t)$ is the value output by the failure detector module of process $p$ at time $t$.

A failure detector $D$ with range $\mathcal{R}$ is a function that maps every failure pattern $F$ to a nonempty set of failure detector histories with range $\mathcal{R}$. $D(F)$ is the set of all possible failure detector histories that may be output by $D$ in a failure pattern $F$. Typically we specify a failure detector by stating the properties that its histories satisfy.

Given two failure detectors $D$ and $D'$, we denote by $(D,D')$ the failure detector whose output is an ordered pair in which the first element corresponds to an output of $D$, and the second element corresponds to an output of $D'$. More precisely, if $\mathcal{R}$ and $\mathcal{R}'$ are the ranges of $D$ and $D'$, respectively, then the range of $(D,D')$ is $\mathcal{R} \times \mathcal{R}'$. For all failure patterns $F$,

$$(D,D')(F) = \{ H'' | \exists H \in D(F), \exists H' \in D'(F), \forall p \in \Pi, \forall t \in \mathbb{N} : H''(p,t) = (H(p,t),H'(p,t)) \}.$$ 

### 2.4. Algorithms

An algorithm $\mathcal{A}$ is modeled as a collection of $n$ deterministic automata. There is an automaton $\mathcal{A}(p)$ for each process $p$. Computation proceeds in steps of these automata. In each step, a process $p$ atomically
- receives a single message $m$ from the message buffer $M$, or the empty message $\lambda$;
- queries its local failure detector module and receives a value $d$;
- changes its state; and
- sends a message to every process.

The state transition and the messages that $p$ sends are all uniquely determined by the automaton $\mathcal{A}(p)$, the state of $p$ at the beginning of the step, the received message $m$, and the failure detector value $d$. Formally, a step is a tuple $e = (p,m,d,\mathcal{A})$, where $p$ is the process taking step $e$, $m$ is the message received by $p$ during $e$, $d$ is the failure detector value seen by $p$ in $e$, and $\mathcal{A}$ is the algorithm being executed.

The message received in a step is nondeterministically selected from $M \cup \{\lambda\}$. This reflects the asynchrony of the communication channels: a process $p$ may receive the empty message despite the existence of unreceived messages addressed to $p$.

We assume that each process $p$ has a read-only input variable, denoted $IN_p$, and a write-once output variable, denoted $OUT_p$. Technically, these variables are components of the states of the automaton $\mathcal{A}(p)$. In each initial state of $\mathcal{A}(p)$, the input variable $IN_p$ has some value in $\{0,1\}^*$, and the output variable $OUT_p$ is initialized to the special value $\bot \notin \{0,1\}^*$ (to denote that it was not yet written by $p$).

### 2.5. Configurations

A configuration of an algorithm $\mathcal{A}$ is a pair $(s,M)$, where $s$ is a function that maps each process $p$ to a state of $\mathcal{A}(p)$, and $M$ is the message
buffer. Recall that $M$ is a set of triples $(p, data, q)$, where $p$ sent message $data$ to $q$, which has not yet received it. An initial configuration of algorithm $A$ is a pair $(s, M)$, where $M = \emptyset$ and $s(p)$ is an initial state of the automaton $A(p)$.

A step $(p, m, d, A)$ is applicable to a configuration $C = (s, M)$ if and only if $m \in M \cup \{\lambda\}$. If $e$ is a step applicable to configuration $C$, $e(C)$ denotes the configuration that results when we apply $e$ to $C$. This is uniquely determined by the automaton $A(p)$ of the process $p$ that takes step $e$.

2.6. Schedules. A schedule $S$ of an algorithm $A$ is a finite or infinite sequence of steps of $A$. We denote by $\text{participants}(S)$ the set of processes that take at least one step in schedule $S$. The $i$th step in schedule $S$ is denoted by $S[i]$. A schedule $S$ is applicable to a configuration $C$ if $S$ is the empty schedule, or $S[1]$ is applicable to $C$, $S[2]$ is applicable to $S[1](C)$, etc. If $S$ is finite and is applicable to $C$, $S(C)$ denotes the configuration that results when we apply schedule $S$ to configuration $C$.

Let $S$ be a schedule applicable to an initial configuration $I$ of an algorithm $A$, and let $i, j$ be positive integers such that $i, j \leq |S|$. We say that step $i$ causally precedes step $j$ in $S$ with respect to $I$ if and only if one of the following holds [33]:

- $S[i]$ and $S[j]$ are steps of the same process and $i < j$;
- $S[i]$ is a step in which a message $m$ is sent and $S[j]$ is a step in which $m$ is received, i.e., step $S[i]$ applied to configuration $S[1]\cdots S[i-1](I)$ results in the sending of $m$ and $S[j] = (-, m, -, A)$; or
- there is a positive integer $k \leq |S|$ such that step $i$ causally precedes step $k$, and step $k$ causally precedes step $j$ in $S$ with respect to $I$.

Note that if $S[i]$ and $S[j]$ are steps involving the sending and receipt of the same message $m$, then $i < j$ (because if $j < i$, then $S[j]$ would be receiving $m$ before $m$ is sent in $S[i]$, contradicting the fact that $S$ is applicable to $I$). This implies the following observation.

**Observation 1.** If step $i$ causally precedes step $j$ in $S$ with respect to $I$, then $i < j$.

2.7. Runs. A run of algorithm $A$ using failure detector $D$ in environment $E$ is a tuple $R = (F, H, I, S, T)$ where $F$ is a failure pattern in $E$, $H$ is a failure detector history in $D(F)$, $I$ is an initial configuration of $A$, $S$ is a schedule of $A$, and $T$ is a list of times in $\mathbb{N}$ (informally, $T[i]$ is the time when step $S[i]$ is taken) such that the following hold:

1. $S$ is applicable to $I$.
2. $S$ and $T$ are both finite sequences of the same length, or are both infinite sequences.
3. For all positive integers $i \leq |S|$, if $S[i] = (p, -, d, A)$, then $p \notin F(T[i])$ and $d = H(p, T[i])$.
4. For all positive integers $i < j \leq |S|$, $T[i] \leq T[j]$.
5. For all positive integers $i, j \leq |S|$, if step $i$ causally precedes step $j$ in $S$ with respect to $I$, then $T[i] < T[j]$.

Property (3) states that a process does not take steps after crashing, and that the failure detector value seen in a step is the one dictated by the failure detector history $H$. Property (4) states that the sequence of times when processes take steps in a schedule is nondecreasing, and property (5) states that these times respect causal precedence.

---

*The symbol "-" in a field of a tuple indicates an arbitrary permissible value for that field of the tuple. We use this convention throughout the paper.*
A run whose schedule is finite (respectively, infinite) is called a finite (respectively, infinite) run. An admissible run of algorithm $A$ using failure detector $D$ in environment $E$ is an infinite run $R = \langle F, H, I, S, T \rangle$ of $A$ using $D$ in $E$ with two additional properties:

1. Every correct process takes an infinite number of steps in $S$.
2. Each message sent to a correct process is eventually received. More precisely, for every finite prefix $S' \subseteq S$, and every $q \in \text{correct}(F)$, if the message buffer in configuration $S'(I)$ contains a message $m = (q, m, -, q)$, then for some $i \in \mathbb{N}$, $S[i] = (q, m, -, A)$.

The input and output of a run $R = \langle F, H, I, S, T \rangle$ of an algorithm $A$ are defined as follows. The input of $R$, denoted $I(R)$, is the vector $(I_1, \ldots, I_n)$ where $I_p$ is the value of the input variable $IN_p$ in the initial configuration $I$ of $R$. The output of $R$, denoted $O(R)$, is the vector $(O_1, \ldots, O_n)$ where $O_p$ is the pair $(v, t)$ such that $p$ writes $v$ in its output variable $OUT_p$ at time $t$ in run $R$ ($O_p = \bot$ if $p$ never writes $OUT_p$ in run $R$).

### 2.8. Problems

We consider input/output problems, i.e., problems where each process has an input value and produces an output value. We can specify such a problem $P$ as a set of triples of the form $\langle F, I, O \rangle$: Intuitively, $\langle F, I, O \rangle \in P$ if and only if, when the failure pattern is $F$ and the processes’ input is $I$, the processes’ output $O$ is acceptable, i.e., it “satisfies” problem $P$. More precisely, a problem $P$ is a set of triples $\langle F, I, O \rangle$ where $F$ is a failure pattern, $I$ is a vector $(I_1, \ldots, I_n)$ of input values (each one in $\{0,1\}^*$), and $O$ is a vector $(O_1, \ldots, O_n)$ where each $O_p$ is either $\bot$ or a pair $(v, t)$ such that $v$ is an output value in $\{0,1\}^*$ and $t$ is a time in $\mathbb{N}$. We say that $I$ is an input vector of $P$ if $\langle F, I, O \rangle \in P$ for some $F$ and $O$.

### 2.9. Solving a problem

Let $P$ be a problem, $A$ an algorithm, $D$ a failure detector, and $E$ an environment. We say the following:

1. A run $R = \langle F, H, I, S, T \rangle$ of $A$ using $D$ in $E$ satisfies $P$ if and only if $\langle F, I(R), O(R) \rangle \in P$, or there is no $O$ such that $\langle F, I(R), O \rangle \in P$.
2. $A$ solves $P$ using $D$ in $E$ if and only if
   a. every admissible run $R$ of $A$ using $D$ in $E$ satisfies $P$, and
   b. for every input vector $I = (I_1, \ldots, I_n)$ of $P$, there is an initial configuration $I$ of $A$ with this input (i.e., in configuration $I$ we have $IN_p = I_p$ for every process $p$).
3. $D$ can be used to solve $P$ in $E$ (or simply $P$ can be solved with $D$ in $E$) if and only if there is an algorithm that solves $P$ using $D$ in $E$.

### 2.10. Comparing failure detectors

Intuitively, a failure detector $D'$ is weaker than a failure detector $D$ if processes can use $D$ to emulate $D'$; so if they can solve a problem with $D'$, they can also solve it with $D$. We say that processes can use $D$ to emulate $D'$ in an environment $E$ if there is an algorithm that transforms $D$ to $D'$ in $E$ as follows. The transformation algorithm, denoted $T_{D \rightarrow D'}$, uses $D$ to maintain a variable $D'_{-output}_p$ at every process $p$; $D'_{-output}_p$ functions as the output of the emulated failure detector $D'$ at $p$. For each admissible run $R$ of $T_{D \rightarrow D'}$, let $O_R$ be the history of all the $D'_{-output}$ variables in $R$; i.e., $O_R(p, t)$ is the value of $D'_{-output}_p$ at time $t$ in $R$. Algorithm $T_{D \rightarrow D'}$ transforms $D$ to $D'$ in environment $E$ if and only if for every admissible run $R = \langle F, H, I, S, T \rangle$ of $T_{D \rightarrow D'}$ using $D$ in $E$, $O_R \in D'(F)$.

---

5Intuitively, this means that when the failure pattern is $F$ and the input is $I(R)$ the problem $P$ does not care what the output is.
We say that $\mathcal{D}'$ is weaker than $\mathcal{D}$ in $\mathcal{E}$ if there is an algorithm $\mathcal{T}_{\mathcal{D} \rightarrow \mathcal{D}'}$ that transforms $\mathcal{D}$ to $\mathcal{D}'$ in $\mathcal{E}$. It is easy to see that if $\mathcal{D}'$ is weaker than $\mathcal{D}$ in $\mathcal{E}$, then every problem that can be solved with $\mathcal{D}'$ in $\mathcal{E}$ can also be solved with $\mathcal{D}$ in $\mathcal{E}$. We say that two failure detectors are equivalent in $\mathcal{E}$ if each is weaker than the other in $\mathcal{E}$.

2.11. Weakest failure detector. A failure detector $\mathcal{D}^*$ is the weakest failure detector to solve problem $P$ in environment $\mathcal{E}$ if and only if the following hold:

**Sufficiency.** $\mathcal{D}^*$ can be used to solve $P$ in $\mathcal{E}$.

**Necessity.** For any failure detector $\mathcal{D}$, if $\mathcal{D}$ can be used to solve $P$ in $\mathcal{E}$, then $\mathcal{D}^*$ is weaker than $\mathcal{D}$ in $\mathcal{E}$.

Note that there may be several distinct failure detectors that are the weakest to solve a problem $P$. It is easy to see, however, that they are all equivalent: If $\mathcal{D}$ and $\mathcal{D}'$ are two failures detectors that are weakest to solve the same problem $P$, $\mathcal{D}'$ can be used to solve $P$ (by sufficiency of $\mathcal{D}'$), and so $\mathcal{D}$ is weaker than $\mathcal{D}'$ (by necessity of $\mathcal{D}$). Symmetrically, $\mathcal{D}'$ is weaker than $\mathcal{D}$, and so $\mathcal{D}$ and $\mathcal{D}'$ are equivalent. For this reason, we speak of the weakest, rather than a weakest, failure detector to solve $P$.

3. The failure detectors used in this paper. We now define the failure detectors $\Omega$, $\Sigma$, $\mathcal{FS}$, and $\Psi$ that we informally described in section 1:

- At each process, the leader failure detector $\Omega$ outputs the id of a process; furthermore, if a correct process exists, then there is a time after which $\Omega$ outputs the id of the same correct process at every correct process. Formally, the range of $\Omega$ is $\Pi$, and for every failure pattern $F$,

$$\Omega(F) = \{ H \mid \text{correct}(F) \neq \emptyset \Rightarrow (\exists q \in \text{correct}(F) \forall p \in \text{correct}(F), \exists t \in \mathbb{N}, \forall t' \geq t : H(p,t') = q) \}.$$

- The quorum failure detector $\Sigma$ outputs a set of processes at each process. Any two sets output at any times and by any processes intersect, and eventually every set output at any correct process consists of only correct processes. Formally, the range of $\Sigma$ is $2^\Pi$, and for every failure pattern $F$,

$$\Sigma(F) = \{ H \mid (\forall p,p' \in \Pi, \forall t,t' \in \mathbb{N} : H(p,t) \cap H(p',t') \neq \emptyset) \land (\forall p \in \text{correct}(F), \exists t \in \mathbb{N}, \forall t' \geq t : H(p,t') \subseteq \text{correct}(F)) \}.$$

- The failure signal failure detector $\mathcal{FS}$ outputs green or red at each process. As long as there are no failures, $\mathcal{FS}$ outputs green at every process; once a failure occurs, and only if it does, $\mathcal{FS}$ eventually outputs red permanently at every correct process. Formally, the range of $\mathcal{FS}$ is $\{\text{green, red}\}$, and for every failure pattern $F$,

$$\mathcal{FS}(F) = \{ H \mid \forall p \in \Pi, \forall t \in \mathbb{N} : (H(p,t) = \text{red} \Rightarrow F(t) \neq \emptyset) \land (\text{faulty}(F) \neq \emptyset \Rightarrow \forall p \in \text{correct}(F), \exists t \in \mathbb{N}, \forall t' \geq t : H(p,t') = \text{red}) \}.$$

- The failure detector $\Psi$ initially outputs $\perp$ and may eventually switch to behaving permanently like $(\Omega,\Sigma)$ or like $\mathcal{FS}$. This switch has the following properties: (a) It must occur at all correct processes; (b) it must be consistent (it is not possible for $\Psi$ to behave like $(\Omega,\Sigma)$ at a process $p$ at time $t$ and like $\mathcal{FS}$ at process $p'$ at time $t'$); and (c) $\Psi$ may start behaving like $\mathcal{FS}$ at a process only if a failure occurred. Formally, the range of $\Psi$ is...
{⊥} ∪ {green, red} ∪ {(p, P) | p ∈ Π ∧ P ∈ 2^Π}, and for every failure pattern F,

\[ \Psi(F) = \{ H \mid \exists H' ∈ (Ω, Σ)(F) ∪ FS(F) \forall p ∈ Π : \]
\[ (p ∈ faulty(F) ∧ ∀t ∈ N : H(p, t) = ⊥) ∨ \]
\[ ∃t ∈ N : (∀t' < t : H(p, t') = ⊥ ∧ ∀t' ≥ t : H(p, t') = H'(p, t') ∧ H' ∈ FS(F) ⇒ F(t) ≠ ∅) \} \]

4. Specification of consensus, QC, and NBAC. In this section we define the three problems considered in this paper, namely, consensus, QC, and NBAC. Each of these problems is an input/output problem that can be formally specified as explained in section 2.8; our definitions are more informal here.

4.1. Consensus and quittable consensus. In the consensus problem, each process p has some input value \( v ∈ V = \{0, 1\}^* \) (we say that \( p \) proposes \( v \)) and must write some output value \( v ∈ V \) (we say that \( p \) decides \( v \)) such that the following properties hold.

Termination: Every correct process eventually decides some value.

Uniform agreement: No two processes (whether correct or faulty) decide different values.

Validity: If a process decides \( v \), then some process proposes \( v \).

Quittable consensus is similar to consensus, except that processes are allowed to decide a special value \( Q \notin V \) (which means “quit”) if a failure occurred. More precisely, QC has the same requirements as consensus, except that the above validity property is replaced by the following one.

Validity: Each process may only decide some value in \( V ∪ \{Q\} \), where \( Q \notin V \).

Moreover:

(i) If a process decides \( v \neq Q \), then some process proposes \( v \).
(ii) If a process decides \( Q \), then a failure occurred.\(^6\)

A straightforward proof by indistinguishable scenarios leads to the following.

Observation 2. Let \( A \) be an algorithm that solves consensus, or QC, using a failure detector \( D \) in an environment \( E \). In every run \( R \) of \( A \) using \( D \) in \( E \), if a process decides some value \( v ∈ V \) at some time \( t \), then there is a process that proposes \( v \) and takes at least one step in \( R \) by time \( t \).

4.2. Nonblocking atomic commit. In the NBAC problem, each process \( p \) has some input value \( v ∈ \{\text{Yes, No}\} \) (we say that \( p \) votes \( v \)) and must write some output value \( v ∈ \{\text{Commit, Abort}\} \) (we say that \( p \) decides \( v \)) such that the following properties hold.

Termination: Every correct process eventually decides some value.

Uniform agreement: No two processes (whether correct or faulty) decide different values.

Validity: Each process may only decide Commit or Abort. Moreover:

(i) If a process decides Commit, then all processes vote Yes.
(ii) If a process decides Abort, then either some process votes No or a failure occurred.

\(^6\)Throughout this paper, when we say “if event \( x \) occurs, then event \( y \) occurred” we mean, more precisely, “if event \( x \) occurs at time \( t \), then event \( y \) occurred by time \( t' ≤ t \)."
As with Observation 2, an obvious proof by indistinguishable scenarios leads to the following observation.

**Observation 3.** Let \( A \) be an algorithm that solves NBAC using a failure detector \( D \) in an environment \( E \). In every run \( R \) of \( A \) using \( D \) in \( E \), if a process decides Commit at some time \( t \), then all processes vote Yes and take at least one step in \( R \) by time \( t \).

### 4.3. Using a consensus, QC, or NBAC algorithm inside another algorithm

An algorithm \( A \) can use an algorithm \( A_c \) that solves consensus by emulating \( A_c \) as follows. In the pseudocode of \( A \), a process \( p \) can execute a statement of the form "$d := \text{propose}(v)$" to start an emulated execution of \( A_c \) with input (i.e., proposal) value \( v \). This statement first sets up the initial state of \( A_c(p) \) to correspond to the input value \( v \), and then starts to execute the steps of \( A_c(p) \) with this initial state. If and when \( p \) decides in this emulated execution of \( A_c(p) \), the decision value is assigned to the variable \( d \), and \( p \) resumes executing the steps of \( A \). Concurrently, \( p \) continues to execute the steps of \( A_c(p) \) until \( A_c(p) \) halts. Similar comments apply for an algorithm \( A \) that uses a QC or NBAC algorithm.

### 5. Relating NBAC and QC and their weakest failure detectors

NBAC is in some sense equivalent to the combination of QC and failure detector \( FS \). More precisely, we have the following.

**Theorem 4.** In every environment \( E \):

1. If a failure detector \( D \) can be used to solve QC in \( E \), then \((D,FS)\) can be used to solve NBAC in \( E \).

2. If a failure detector \( D' \) can be used to solve NBAC in \( E \), then
   - \( D' \) can be used to solve QC in \( E \), and
   - \( D' \) can be transformed to \( FS \) in \( E \).

**Proof.** Let \( E \) be an arbitrary environment.

1. Suppose that failure detector \( D \) can be used to solve QC in \( E \), i.e., there is an algorithm \( A_{qc} \) that uses \( D \) to solve QC in \( E \). Figure 1 shows an algorithm \( A_{nbac} \) that uses \((D,FS)\) to solve NBAC in \( E \). \( A_{nbac} \) works as follows. Each process \( p \) sends its vote to every process, and waits until it receives a vote from every process or the \( FS \) component of \((D,FS)\) indicates that a failure occurred. If \( p \) receives a vote from every process and all the votes are Yes, it sets \( myproposal \) to 1; otherwise some vote was No or a failure occurred, and \( p \) sets \( myproposal \) to 0. Then, in line 7, \( p \) participates in an execution of the
QC algorithm $A_{qc}$ (which uses the $D$ component of $(D,FS)$) where $p$'s initial value is set to $\text{myproposal}$ (as explained in section 4.3). If $p$ decides 1 in this execution of $A_{qc}$, then $p$ decides Commit for NBAC; if $p$ decides 0 or Q in this execution of $A_{qc}$, then $p$ decides Abort for NBAC.

We now prove that, in every admissible run, algorithm $A_{nbac}$ satisfies all the properties of NBAC.

**Termination.** This property holds trivially if all processes are faulty, so assume that some process is correct. Let $p$ be any correct process. Since every correct process executes the algorithm in Figure 1, if $p$ never receives the vote of some process $q$, then $q$ must have crashed. In that case, by the specification of the failure detector $FS$, eventually $FS_p = \text{red}$ forever. Thus, $p$ eventually completes the wait statement on line 2. Therefore, eventually every correct process starts the execution of $A_{qc}$ in line 7 (as explained in section 4.3). By the termination property of QC, every correct process completes its execution of line 7, and eventually decides.

**Uniform agreement.** This follows from the uniform agreement property of QC.

**Validity.** Let $p$ be any process.

(a) Suppose $p$ decides Commit (line 9). Then $p$ decided 1 in its execution of $A_{qc}$ on line 7. By part (i) of validity of QC, some process $q$ proposes 1 on line 7 (i.e., $q$ starts its emulation of $A_{qc}$ with initial value 1 on that line). Before doing so, $q$ must have received Yes votes from all processes (see lines 3–4). So, if a process decides Commit, all processes vote Yes.

(b) Suppose $p$ decides Abort (line 11). Then $p$ decided 0 or $Q$ in its execution of $A_{qc}$ on line 7. If $p$ decided $Q$, then by part (ii) of validity of QC, a failure occurred. If $p$ decided 0, then, by Observation 2, there is a process $q$ that proposed 0 and took a step in the emulation of $A_{qc}$ on line 7. Before doing so, $q$ must have received a vote No from some process or found that $FS_q = \text{red}$ (see lines 3, 5, 6). The latter can happen only if a failure occurred. We conclude that if a process decides Abort, some process votes No or a failure occurred.

(2) Suppose that failure detector $D'$ can be used to solve NBAC in $\mathcal{E}$; i.e., there is an algorithm $B_{nbac}$ that uses $D'$ to solve NBAC in $\mathcal{E}$.

(a) $D'$ can be used to solve QC in $\mathcal{E}$. An algorithm $B_{qc}$ that uses $D'$ to solve QC in $\mathcal{E}$ is shown in Figure 2. Informally, it works as follows. Each process $p$ first sends its QC proposal, some value $v \in V$, to all processes. Then, in line 2, $p$ participates in an execution of the NBAC algorithm $B_{nbac}$ (which uses $D'$) with initial value Yes, i.e., an execution of NBAC where $p$ votes Yes. If this execution returns Abort, $p$ decides $Q$; if it returns Commit, $p$ waits to receive a proposal from every process and decides the smallest of these proposals.

We now prove that, in every admissible run, algorithm $B_{qc}$ satisfies all the properties of QC.

**Termination.** This property holds trivially if all processes are faulty, so assume that some process is correct. Every correct process executes the statement $d := \text{vote(Yes)}$ on line 2 to participate in an execution of NBAC (as explained in section 4.3). By the termination property of NBAC, all correct processes eventually decide, i.e., they complete the execution of this statement. If line 2 sets $d$ to Abort, then $p$ decides $Q$ on line 4. Otherwise, it must set $d$ to Commit. By Observation 3, every
process q voted Yes and took a step in the emulation of $B_{nbac}$ on line 2. Before doing so, q sent its QC proposal to all processes (on line 1). So, $p$ eventually receives a proposal from every process, completes the wait statement on line 6, and decides some value for QC on line 7. 

Uniform agreement. By the uniform agreement property of NBAC, all the processes that set their variable $d$ in line 2 set it to the same value. Thus, all the processes that decide some value (for QC) do so on line 4 or they all decide on line 7. In the first case they all decide Q, and in the second case they all decide the smallest proposal of all processes in $\Pi$. So no two processes decide differently.

Validity. Let $p$ be any process. If $p$ decides $v \neq Q$ (on line 7), then $v$ is the smallest proposal that $p$ received, and thus some process proposes $v$. Now suppose $p$ decides Q (on line 4). Thus, $p$'s execution of the statement $d := \text{vote}(\text{Yes})$ on line 2 sets $d$ to Abort. By part (ii) of validity of NBAC, either some process votes No or a failure occurred. But no process votes No. Thus, a failure occurred.

(b) $D'$ can be transformed to $FS$ in $E$ (this result can be found in [12, 25]).

The transformation algorithm is shown in Figure 3. At each process $p$, the variable $FS$-output$_p$ (which emulates the output of $FS$ at $p$) is initially green. Processes emulate consecutive and independent executions of $B_{nbac}$ using $D'$ to solve consecutive instances of NBAC while voting Yes in every instance. If and when a process $p$ decides Abort in an instance of NBAC, then $p$ sets $FS$-output$_p$ to red, and never changes $FS$-output$_p$ thereafter.

From the agreement and termination properties of NBAC, it is easy to show by induction that the following holds (the proof is omitted here). 

CLAIM 4.1. Either all correct processes execute the repeat-until loop of lines 2–4 infinitely many times, or they all exit this loop and execute line 5.
Suppose no failures occur. Since (a) all the processes are correct, and (b) they all vote Yes in every instance of NBAC executed in line 2, then by part (ii) of the validity property of NBAC no process ever decides Abort on line 2, and so $FS$-output remains green at all processes forever.

Suppose a failure occurs. Then there is a process $p$ that crashes and a $k$ such that $p$ does not participate (i.e., does not take any step) in the $k$th instance of NBAC. We claim that every correct process eventually sets $FS$-output to red on line 5. Suppose, for contradiction, that some correct process never sets $FS$-output to red on line 5. By Claim 4.1, it must be that all correct processes execute the repeat-until loop of lines 2–4 infinitely many times, and so they participate in the $k$th instance of NBAC. Since $p$ takes no steps in this instance, then, by Observation 3, correct processes cannot decide Commit in that instance. So, by Claim 4.1, they all decide Abort in the $k$th instance of NBAC, and then they exit the repeat-until loop—a contradiction. Thus, every correct process eventually sets $FS$-output to red on line 5.

Finally, suppose some process $p$ sets $FS$-output to red on line 5 at some time $t$. Then $p$ must have decided Abort in an instance of NBAC on line 2 by time $t$. By part (ii) of validity of NBAC, some process votes No in that instance of NBAC, or a failure occurred by time $t$. Since no process ever votes No, a failure occurred by time $t$.

The close relationship between NBAC and QC established in Theorem 4 allows us to relate the weakest failure detectors to solve these problems.

**Theorem 5.** For every environment $E$, if $D$ is the weakest failure detector to solve QC in $E$, then $(D,FS)$ is the weakest failure detector to solve NBAC in $E$.

**Proof.** Let $E$ be an arbitrary environment, and let $D$ be the weakest failure detector to solve QC in $E$. This means that (i) $D$ can be used to solve QC in $E$, and (ii) if a failure detector $D'$ can be used to solve QC in $E$, then $D'$ can be transformed to $D$ in $E$.

To prove that $(D,FS)$ is the weakest failure detector to solve NBAC in $E$, we now show two facts:

1. $(D,FS)$ can be used to solve NBAC in $E$. This follows directly from (i) and Theorem 4(1).
2. If a failure detector $D'$ can be used to solve NBAC in $E$, then $D'$ can be transformed to $(D,FS)$.

To see this, let $D'$ be a failure detector that can be used to solve NBAC in $E$. By Theorem 4(2),

- $D'$ can be used to solve QC in $E$. So by (ii) above, $D'$ can be transformed to $D$ in $E$.
- $D'$ can be transformed to $FS$ in $E$.

Thus, $D'$ can be transformed to $(D,FS)$ in $E$. □

**The weakest failure detectors to solve QC and NBAC.** In section 6 we show that $Ψ$ can be used to solve QC in every environment (Theorem 8). In section 8 we show that, in every environment $E$, any failure detector that can be used to solve QC in $E$ can be transformed to $Ψ$ in $E$ (Theorem 30). From these two facts, we have the following corollary.

**Corollary 6.** For every environment $E$, $Ψ$ is the weakest failure detector to solve QC in $E$.

Theorem 5 relates the weakest failure detector to solve QC to the weakest failure detector to solve NBAC. So by Corollary 6 and Theorem 5, we have the next corollary.
Code that process \( p \) executes to propose \( v \) for QC:

1. wait until \( [\Psi \neq \bot] \)
2. if \( \Psi \in \{\text{green, red}\} \) then /* \( \Psi \) behaves like \( \mathcal{FS} \) and, thus, a failure occurred */
3. \( \text{decide} \ Q \)
4. else /* henceforth \( \Psi \) behaves like \( (\Omega, \Sigma) \) */
5. \( d := \text{propose}(v) \) /* execute \( \mathcal{A}_c \) using \( \Psi \) to solve an instance of consensus */
6. \( \text{decide} \ d \)

FIG. 4. Using \( \Psi \) to solve QC.

**Corollary 7.** For every environment \( \mathcal{E} \), \( (\Psi, \mathcal{FS}) \) is the weakest failure detector to solve NBAC in \( \mathcal{E} \).

**6. \( \Psi \) is sufficient to solve quittable consensus.** Recall that, intuitively, \( \Psi \) behaves as follows (see section 3 for a precise definition). For an initial period of time the output of \( \Psi \) at each process is \( \bot \). Eventually, however, \( \Psi \) behaves either like the failure detector \( (\Omega, \Sigma) \) at all correct processes or like the failure detector \( \mathcal{FS} \) at all correct processes. The switch from \( \bot \) to \( (\Omega, \Sigma) \) or \( \mathcal{FS} \) is consistent at all processes, and a switch from \( \bot \) to \( \mathcal{FS} \) can happen only if a failure occurred.

In Figure 4 we show an algorithm that uses \( \Psi \) to solve QC in any environment. This algorithm uses an algorithm \( \mathcal{A}_c \) that solves consensus using \( (\Omega, \Sigma) \) in any environment. Delporte-Gallet, Fauconnier, and Guerraoui [15] have shown that such an algorithm exists.

Informally, the algorithm in Figure 4 works as follows. To propose some value \( v \in V \) for QC, a process \( p \) waits until \( \Psi_p \) ( \( p \)'s module of failure detector \( \Psi \)) outputs a value different from \( \bot \). At that time, either \( \Psi_p \) starts behaving like \( \mathcal{FS} \) or it starts behaving like \( (\Omega, \Sigma) \). If \( \Psi_p \) behaves like \( \mathcal{FS} \) (which happens only if a failure occurred), then \( p \) decides Q. If, on the other hand, \( \Psi_p \) behaves like \( (\Omega, \Sigma) \), then \( p \) participates in an execution of the consensus algorithm \( \mathcal{A}_c \), where it proposes \( v \) (it does so by executing the \( d := \text{propose}(v) \) statement on line 5, as explained in section 4.3). Process \( p \) adopts the decision value of this execution of \( \mathcal{A}_c \), as its decision for QC.

**Theorem 8.** For every environment \( \mathcal{E} \), the algorithm in Figure 4 uses \( \Psi \) to solve QC in \( \mathcal{E} \).

*Proof.* Consider any admissible run of the algorithm in Figure 4. We will prove that this run satisfies the properties of QC.

**Termination.** This property holds trivially if all processes are faulty, so assume that some process is correct. Let \( p \) be any correct process. By the specification of \( \Psi \), there is a time after which \( \Psi_p \) has values in the range of either \( \mathcal{FS} \) or \( (\Omega, \Sigma) \); thus, \( p \) completes the wait statement on line 1. If eventually \( \Psi_p \) has values in the range of \( \mathcal{FS} \), then \( p \) decides Q (see lines 2–3). Otherwise, \( \Psi \) never outputs values in the range of \( \mathcal{FS} \) at any process, and there is a time after which \( \Psi \) outputs only values in the range of \( (\Omega, \Sigma) \) at all correct processes. Thus, eventually every correct process executes the statement \( d := \text{propose}(v) \) for some \( v \) on line 5; i.e., every correct process participates in an execution of \( \mathcal{A}_c \). By the termination property of consensus, this execution terminates, and so \( p \) decides \( d \) on line 6.

**Uniform agreement.** By the specification of \( \Psi \), it is not possible that \( \Psi \) outputs a value in the range of \( \mathcal{FS} \) at one process and a value in the range of \( (\Omega, \Sigma) \) at another. From this observation and the fact that \( \mathcal{A}_c \) satisfies uniform agreement (for consensus), it follows that the algorithm in Figure 4 satisfies uniform agreement (for QC).
Validity. Let $p$ be any process.

(i) Suppose $p$ decides some value $v \neq Q$ for QC (on line 6). Thus, $p$ also decides $v$ in its execution of the consensus algorithm $A_c$ on line 5. From Observation 2, at least one process $q$ starts to execute the statement $d := \text{PROPOSE}(v)$ on line 5. Therefore, process $q$ executes the algorithm in Figure 4 with QC proposal $v$. So, if a process decides $v \neq Q$ (for QC), some process proposes $v$ (for QC).

(ii) Suppose that $p$ decides Q for QC at some time $t$ (on line 3). Thus, $\Psi_p \in \{\text{green, red}\}$ by time $t$. By the specification of $\Psi$, a failure occurred by time $t$. □

7. Some auxiliary results. In this section we present some technical lemmas used in our proof that in every environment $\Psi$ is necessary to solve QC, presented in section 8. The lemmas in section 7.2 appeared in [9], sometimes in different form.

7.1. Mergeable runs. Several proofs in distributed computing employ a technique known as the “partition argument.” At the heart of this technique is the ability to combine two different runs $R_0$ and $R_1$ of an algorithm $A$ that involve disjoint sets of processes $P_0$ and $P_1$, respectively, into a single run of $A$ in which the processes in $P_0$ behave as in $R_0$ and the processes in $P_1$ behave as in $R_1$. We now formalize this and prove that in our model it is possible to combine such “mergeable” runs in this manner.

Let $R_0 = (F, H, I, S, S_0, T_0 \cdot T_0)$ and $R_1 = (F, H, I, S, S_1, T_1 \cdot T_1)$ be two finite runs of an algorithm $A$ using failure detector $D$ in some environment $E$, such that $|T_0| = |T_1| = |S|$ and $\text{participants}(S_0) \cap \text{participants}(S_1) = \emptyset$. Note that the schedules of these two runs start with the same prefix $S$, while their continuations $S_0$ and $S_1$ involve disjoint sets of processes. A merging of two such runs is a tuple $R = (F, H, I, S, S, T \cdot T)$, where (a) $T$ is whichever of $T_0$ or $T_1$ has the smaller last element (either one, if both have the same last element); (b) $T$ is the sequence consisting of the times in $T_0$ and $T_1$ in nondecreasing order; and (c) $S$ is the sequence consisting of the steps in $S_0$ and $S_1$ merged in the same order as the elements of $T_0$ and $T_1$ were merged into $T$. For example, suppose that $S_0 = a_1, a_2, a_3$, $T_0 = 3, 5, 7$; and $S_1 = b_1, b_2, b_3, b_4$, $T_1 = 2, 4, 5, 6$. Then $T = 2, 3, 4, 5, 6, 7$, and the two possibilities for $S$ are $b_1, a_1, b_2, b_3, a_2, b_4, a_3$ or $b_1, a_1, b_2, a_2, b_3, b_4, a_3$. More formally, the requirements for $R = (F, H, I, S, S, T \cdot T)$ to be a merging of $R_0$ and $R_1$ are as follows:

- $|S| = |S_0| + |S_1|$ and $|T| = |T_0| + |T_1|$;
- $\hat{T} = T_b$ for some $b \in \{0, 1\}$ such that the last element of $T_b$ is less than or equal to the last element of $T_b$;
- $T$ is nondecreasing:
  - for each $b \in \{0, 1\}$ and each $i \in \{1, 2, \ldots, |S_b|\}$ there is a $j \in \{1, 2, \ldots, |S|\}$ such that $S_b[i] = S[j]$ and $T_b[i] = T[j]$; and
  - for each $j \in \{1, 2, \ldots, |S|\}$ there is a $b \in \{0, 1\}$ and an $i \in \{1, 2, \ldots, |S_b|\}$ such that $S[j] = S_b[i]$ and $T[j] = T_b[j]$.

Lemma 9. Let $R_0 = (F, H, I, S, S_0, T_0 \cdot T_0)$ and $R_1 = (F, H, I, S, S_1, T_1 \cdot T_1)$ be two finite runs of an algorithm $A$ using failure detector $D$ in some environment $E$ such that $|T_0| = |T_1| = |S|$ and $\text{participants}(S_0) \cap \text{participants}(S_1) = \emptyset$. Let $R = (F, H, I, S, S, T \cdot T)$ be a merging of $R_0$ and $R_1$. Then:

- (a) $R$ is also a run of $A$ using $D$ in $E$.
- (b) For each $b \in \{0, 1\}$ and each process $p \in \text{participants}(S_b)$, the state of $p$ is the same in $S \cdot S(I)$ as in $S \cdot S_b(I)$.

The proof of Lemma 9 is straightforward though somewhat tedious; it is given in the appendix.
7.2. Directed acyclic graphs and simulations. To complete the proof that, for any environment $E$, $\Psi$ is the weakest failure detector to solve QC in $E$, it remains to show that any failure detector that can be used to solve QC in $E$ can be transformed to $\Psi$ in $E$. In this section we review a technique for proving statements of this type. The technique was introduced by Chandra, Hadzilacos, and Toueg [9], who used it to prove that any failure detector that can be used to solve consensus can be transformed to $\Omega$. We will use it in this paper to prove that any failure detector that can be used to solve QC can be transformed to $\Psi$ (see section 8).

Suppose we want to prove that $D^*$ is the weakest failure detector to solve some problem $P$ in some environment $E$. Let $D$ be any failure detector that can be used to solve $P$ in $E$, i.e., there is an algorithm $A$ that uses $D$ to solve $P$ in $E$. We need to show that $D$ can be transformed to $D^*$. The proof technique of [9] shows how to use $D$ and $A$ to emulate $D^*$ in $E$. This emulation consists of two interacting components: the communication component and the computation component. In the communication component, each process continuously “samples” its local module of $D$ and exchanges messages with other processes to construct an ever-growing directed acyclic graph (DAG) of failure detector samples of $D$. In the computation component, $p$ uses this DAG to simulate schedules of the algorithm $A$ (which uses $D$ to solve $P$). Based on these simulated schedules, $p$ simulates the output of the failure detector $D^*$ that we want to emulate.

We now explain in more detail how each process builds its DAG of failure detector samples and how it uses this DAG to simulate schedules of $A$.

7.2.1. Building directed acyclic graphs of failure detector samples. The DAG-building algorithm, denoted $A_{DAG}$, is shown in Figure 5. In our algorithm descriptions, which we give in pseudocode, we use the following conventions. Variables of process $p$ are subscripted with $p$. If $D$ is a failure detector, then $D_p$ denotes the function call by which $p$ can access its local module of $D$: this call returns the current value of $p$'s local module of $D$. The pseudocode of each process begins with an initialize clause, which defines the process's state in the initial configuration. (Variables whose values are not explicitly set in this clause can be assigned arbitrary values in the initial configuration.)

In $A_{DAG}$, each process $p$ maintains a DAG of failure detector samples of $D$ in the variable $G_p$. Each node of this DAG is of the form $(q, d, k)$; such a triple indicates that process $q$ obtained value $d$ when it queried its failure detector module $D_q$ for the $k$th time. (The third component is included to ensure that distinct samplings of the failure detector result in distinct nodes.) We call such triples samples; a sample $(q, −, −)$ is said to be of or taken by process $q$. We use the terms “node (of the DAG)” and “sample” interchangeably.

Process $p$ periodically performs the following actions:
(a) it receives a message, which is either a DAG previously sent to $p$ by another process, or the empty message (line 5);
(b) it queries its local failure detector module $D_p$, receiving a value that it stores in variable $d_p$ (line 6);
(c) it updates its DAG $G_p$ by first adding to it the DAG that it received in (a), and then adding to it a new node with the failure detector value it got in (b), as well as edges from all other nodes to the new node (lines 7–10); and
(d) it sends the updated $G_p$ to all processes (line 11).

Note that this sequence of actions (receiving a message, querying the local failure detector module, changing local state, and sending messages to other processes)
corresponds exactly to the sequence of actions taken in a single step in our model. Thus, each iteration of the loop in Figure 5 is executed as a single step.

We now present some properties of the DAGs of samples computed by algorithm \( \mathcal{A}_{DAG} \). In the following, we consider an arbitrary admissible run \( R = (F,H,I,S,T) \) of \( \mathcal{A}_{DAG} \) using failure detector \( \mathcal{D} \) in some arbitrary environment \( \mathcal{E} \). We use the following notation throughout this section: In the context of a given run of an algorithm, the value of variable \( x_p \) at time \( t \) is denoted \( x_p^t \); if \( p \) takes a step at time \( t \), then \( x_p^t \) is the value of \( x_p \) after that step.

We start with some simple observations, in each of which \( p \) is an arbitrary process. Since \( p \) never removes any nodes or edges from \( G_p \), the DAG contained in this variable is monotonically nondecreasing. That is, we have the following observation.

**Observation 10.** For all \( t,t' \in \mathbb{N} \), if \( t \leq t' \) then \( G_p^t \) is a subgraph of \( G_p^{t'} \).

We define the limit DAG of a process \( p \) to be \( G_p^\infty = \bigcup_{t \in \mathbb{N}} G_p^t \).

In the same step that a process updates its DAG (line 10), it also sends the new DAG to all processes (line 11); thus each correct process will eventually receive that DAG and will incorporate it into its own. Thus, we have the next observation.

**Observation 11.** For every correct process \( p \), every process \( q \), and every time \( t \in \mathbb{N} \), \( G_p^t \) is a subgraph of \( G_p^\infty \).

From this, the next observation follows immediately.

**Observation 12.** If \( p \) and \( q \) are correct processes, then \( G_p^\infty = G_q^\infty \).

Since \( k_p \) is incremented in each iteration of \( p \)'s loop, when \( p \) takes sample \( (p,-,k) \), it has already taken samples \( (p,-,k') \) for all \( k' < k \); and, at that time, it adds edges from all such nodes to \( (p,-,k) \). Thus, we have the following observation.

**Observation 13.** If \( v = (p,-,k) \) and \( v' = (p,-,k') \) are nodes of \( G_p^\infty \) and \( k \geq k' \), then \( v \) is a descendant of \( v' \) in \( G_p^\infty \).

Let \( v = (q,d,k) \) be any node of \( G_p^\infty \). It is obvious from the code of \( \mathcal{A}_{DAG} \) that process \( q \) received \( d \) from its failure detector module in its \( k \)th step. Let \( \tau(v) \) be the time when \( q \) takes this step. More precisely, if \( S[i] \) is the \( k \)th step of \( q \) in \( S \), then \( \tau(v) = T[i] \). (Recall that \( S \) is the schedule and \( T \) is the sequence of times of the run \( R \) of \( \mathcal{A}_{DAG} \) that we are considering.) From property (3) of runs, we have the following observation.

**Observation 14.** If \( v = (q,d,k) \) is a node of \( G_p^\infty \), then \( q \notin F(\tau(v)) \) and \( d = H(q,\tau(v)) \).
From the algorithm \( A_{	ext{DAG}} \), it is clear that if \((u, v)\) is an edge of the limit DAG \( G_p^\infty \), then the step in which sample \( u \) was taken causally precedes the step in which sample \( v \) was taken in schedule \( S \) with respect to \( I \) (the initial configuration of run \( R \)).

From property (5) of the runs of \( A_{	ext{DAG}} \) (see section 2.7), it follows that \( \tau(u) < \tau(v) \).

By induction we can generalize this observation from single edges to finite or infinite paths of \( G_p^\infty \).

**Observation 15.** If \( g = v_0, v_1, \ldots \) is a finite or infinite path in \( G_p^\infty \), then the sequence of times \( \tau(v_0), \tau(v_1), \ldots \) is strictly increasing.

Let \( G \) be any DAG; if \( v \) is a node of \( G \), then \( G|v \) is the subgraph of \( G \) induced by the descendants of \( v \) in \( G \); otherwise, \( G|v \) is the empty graph. Informally, the next lemma states that any finite path in process \( p \)'s limit DAG eventually appears permanently in \( p \)'s DAG.

**Lemma 16.** Let \( p \) be a process and \( v \) be a node of \( G_p^\infty \). For each finite path \( g \) in \( G_p^\infty |v \), there is a time \( t \) such that, for all \( t' \geq t \), \( g \in G_p^t|v \).

**Proof.** In \( G_p^\infty |v \), let \( g \) be any finite path, \( g' \) be a finite path from \( v \) to the first node of \( g \), and \( h \) be the path consisting of \( g' \) followed by \( g \). Since \( G_p^\infty = \cup_{t \in \mathbb{N}} G_p^t \), it is clear that for each edge \( e \) of \( h \) there is a time \( t(e) \) such that \( e \in G_p^{t(e)} \). Let \( t = \max \{ t(e) : e \text{ is an edge of } h \} \). By Observation 10, every edge \( e \) of \( h \) (and hence \( v \) and the entire path \( g \)) is in \( G_p^t \) for all \( t' \geq t \). Since \( g \) is in \( G_p^\infty |v \), every node in \( g \) is a descendant of \( v \). Thus, \( g \) is in \( G_p^t|v \), for all \( t' \geq t \). \( \square \)

Since faulty processes eventually crash and cease to take steps, from a certain point on, only correct processes take samples. This is the basic intuition underlying the next lemma.

**Lemma 17.** For every correct process \( p \), there is a sample \( v^* \) of \( p \) in \( G_p^\infty \) such that \( G_p^\infty |v^* \) contains only samples of correct processes. Furthermore:

(a) There is a time after which any node \( v \) in variable \( v_p \) (line 9) is a descendant of \( v^* \) in \( G_p^\infty \).

(b) For any descendant \( v \) of \( v^* \) in \( G_p^\infty \) and any \( t \in \mathbb{N} \), \( G_p^t|v \) contains only samples of correct processes.

**Proof.** Since \( p \) is correct, it takes infinitely many steps. Let \( t^* \) be the first time that \( p \) takes a step after all faulty processes have crashed, and let \( v^* \) be the sample that \( p \) takes in that step. Consider any node \( v \) of \( G_p^\infty |v^* \). Since \( v \) is a descendant of \( v^* \) in \( G_p^\infty \), by Observation 15, \( \tau(v) \geq \tau(v^*) = t^* \). Since all faulty processes have crashed by time \( t^* \), the process that takes sample \( v \) (at time \( \tau(v) \geq t^* \)) must be correct. So, \( G_p^\infty |v^* \) contains only samples of correct processes.

(a) Let \( v^* = (p, -, k^*) \). Since \( k_p \) increases in each iteration of \( p \)'s loop, eventually \( k_p \) has values that are more than \( k^* \). Therefore, eventually only nodes whose third entry is more than \( k^* \) are assigned to \( v_p \). By Observation 13 all these nodes are descendants of \( v^* \) in \( G_p^\infty \).

(b) Consider any descendant \( v \) of \( v^* \) in \( G_p^\infty \) and any \( t \in \mathbb{N} \). Clearly, \( G_p^t|v \) is a subgraph of \( G_p^\infty |v \), and \( G_p^\infty |v \) is a subgraph of \( G_p^\infty |v^* \). Since \( G_p^\infty |v^* \) contains only samples of correct processes, so does its subgraph \( G_p^t|v \). \( \square \)

Since correct processes keep taking samples and exchanging their DAGs forever, every correct process’s limit DAG has an infinite path with infinitely many samples of each correct process. This observation is formalized by Lemma 19. To prove it, it is convenient to prove the following lemma first.

**Lemma 18.** Suppose \( p \) is a correct process, and let \( G \) be a subgraph of \( G_p^t \) for some time \( t \). For every correct process \( q \), there is a time \( t' \) such that \( G_p^t \) contains a sample \( w \) of \( q \) and an edge from every node of \( G \) to \( w \).
Proof. Let $s$ be the first step that $p$ takes after time $t$. By Observation 10, $G$ is still in $p$'s DAG just before this step. There are two cases:

$p = q$. In step $s$, $p$ adds to its DAG a new sample $w = (p, -, -)$, and adds edges from every other node in its DAG (in particular, from every node in $G$) to $w$. Thus, when this step is completed, say at time $t'$, $G_p'$ has the desired properties.

$p \neq q$. In step $s$, $p$ sends to all processes a DAG that contains $G$. Now consider the steps in which $q$ receives that DAG. In that step, $q$ first incorporates the DAG it receives, which contains $G$, into its own DAG. Then $q$ adds to its DAG a new sample $w = (q, -, -)$, and adds edges from every other node in its DAG (in particular, from every node in $G$) to $w$. Finally, $q$ sends the resulting DAG to all processes. Consider the step in which $p$ receives that DAG. When it does so, $p$ incorporates the DAG it receives into its own DAG. Thus, when this step is completed, say at time $t'$, $G_p'$ has the desired properties. 

Lemma 19. If $p$ is a correct process and $v$ is a node of $G_p^\infty$, then $G_p^\infty$ has an infinite path that starts with $v$ and contains infinitely many samples of each correct process.

Proof. Since $v$ is a node of $G_p^\infty$, there is a time $t_0$ such that $v$ is in $G_p^{t_0}$. By repeated application of Lemma 18, there is an infinite sequence of times $t_0, t_1, \ldots$ and an infinite sequence of paths $g^0, g^1, \ldots$ such that for all $i \in \mathbb{N}$, (a) $g^i$ is in $G_p^{t_i}$ and starts with $v$, (b) $g^i$ is a prefix of $g^{i+1}$, and (c) each correct process has at least $i$ samples in $g^i$.

Let $g^\infty$ be the “limit” of sequence $g^0, g^1, \ldots$; that is, $g^\infty$ is the infinite path which, up to length $|g^i|$, is identical to $g^i$. (This is well defined because of (b).) It is now easy to see that $g^\infty$ is a path in $G_p^\infty$ that starts with $v$ and contains infinitely many samples of each correct process.
Let $G$ be any DAG of samples and $I$ be any initial configuration of $A$. \textbf{Sch}$(G,I)$
denotes the set of schedules of $A$ that are compatible with some path in $G$ and are
applicable to $I$. Note that if $G$ is finite, then \textbf{Sch}$(G,I)$ contains a finite number of
finite schedules.

We now present some properties of simulated schedules. In the following, we
consider an arbitrary admissible run $R$ of $A_{\text{DAG}}$ using failure detector $D$ in some
arbitrary environment $E$. Let $F \in E$ be the failure pattern of this run, and let
$H \in D(F)$ be its failure detector history.

The first lemma justifies the name “simulated schedules”; it states that these
schedules really are schedules of runs of algorithm $A$ using $D$, with failure pattern $F$
and failure detector history $H$.

\textbf{Lemma 20}. Let $p$ be a process, $t \in \mathbb{N} \cup \{\infty\}$, $G$ be a subgraph of $G_p^t$,
and $I$ be an initial configuration of algorithm $A$. For each schedule $S \in \textbf{Sch}(G,I)$, there is a list
of times $T$, all at most $t$, such that $R_A = (F,H,I,S,T)$ is a run of $A$ using $D$ in $E$.

Proof. Let $S$ be any schedule in $\textbf{Sch}(G,I)$. Thus, $S$ is a schedule of $A$ that
is applicable to $I$ and compatible with some path $g = v_1,v_2,\ldots$ in $G$. Let $T =
\tau(v_1),\tau(v_2),\ldots$. Recall that for each positive integer $i \leq |S|$, $\tau(v_i)$ is the time when
sample $v_i$ was taken. A sample cannot appear in any DAG until the time it is taken.
Since $v_i$ is a node in a subgraph of $G_p^t$, $\tau(v_i) \leq t$.

We claim that $R_A = (F,H,I,S,T)$ is a run of $A$ using $D$ in $E$. Since $F \in E$,
$H \in D(F)$, and $I$ is an initial configuration of $A$, it suffices to verify that $R_A$
satisfies properties (1)–(5) of runs. $S$ is applicable to $I$ (property (1)) by definition of
\textbf{Sch}$(G,I)$. $S$ and $T$ have the same length (property (2)) because each of them has
the same length as $g$. The fact that in $R$ no process takes a step after it has crashed,
and that the failure detector value in each step is consistent with the history $H$ (property (3)) follows from Observation 14, since $S$ is compatible with path $g = v_1,v_2,\ldots$
and $T = \tau(v_1),\tau(v_2),\ldots$. Observation 15 implies that $T$ is strictly increasing, and
so property (4) is also satisfied. To show property (5), we must prove that if step $i$
causally precedes step $j$ in $S$ with respect to $I$, then $T[i] < T[j]$. This follows from
Observation 1 and the fact that $T$ is strictly increasing. \qed

By Lemma 20, every infinite schedule $S^\infty \in \textbf{Sch}$(G_p^\infty,I)$ is a schedule of an infinite
run of $A$ using $D$ in $E$. However, $S^\infty$ is not necessarily a schedule of an
admissible run, i.e., a run where each correct process takes an infinite number of steps (property (6))
and eventually receives every message sent to it (property (7)). The next lemma,
however, states that every finite schedule $S \in \textbf{Sch}(G_p^\infty,I)$ can be extended to some
infinite schedule $S^\infty \in \textbf{Sch}$(G_p^\infty,I)$ of an admissible run of $A$.

\textbf{Lemma 21}. Suppose $p$ is a correct process, and let $I$ be an initial configuration of
$A$. For any finite schedule $S \in \textbf{Sch}(G_p^\infty,I)$ there is a schedule $S^\infty \in \textbf{Sch}$(G_p^\infty,I)$
that extends $S$ and a list of times $T^\infty$ such that $R_A = (F,H,I,S^\infty,T^\infty)$ is an admissible
run of $A$ using $D$ in $E$. Furthermore, for any node $u$ in $G_p^\infty$, $S^\infty$ can be chosen so
that $S^\infty = S \cdot \sigma^\infty$ for a schedule $\sigma^\infty$ that is compatible with a path in $G_p^\infty|u$.

Proof. Let $S$ be any finite schedule in $\textbf{Sch}$(G_p^\infty,I)$, and let $u$ be any node in $G_p^\infty$. Thus $S$
is applicable to $I$ and compatible with a finite path $g$ of $G_p^\infty$. By Lemma 18 (applied with $g = p$
and $G$ consisting of the path $g$ and the node $u$) and the monotonicity of the DAGs (Observation 10),
$G_p^\infty$ contains a sample $v$ of $p$ and an edge from every node of $g$ and from $u$ to $v$. By Lemma 19, $G_p^\infty$
has an infinite path $\gamma$ that starts with $v$ and contains infinitely many samples of each correct process.
Note that $g\cdot \gamma$ is a path in $G_p^\infty$ (because there are edges from every node in $g$ to the first
node of $\gamma$), and $\gamma$ is a path in $G_p^\infty|u$ (because there is an edge from $u$ to the first node
of $\gamma$).
We define an infinite sequence of schedules $\sigma^0, \sigma^1, \ldots$ such that for each $i \in \mathbb{N}$, (a) $\sigma^i$ has length $i$, (b) $\sigma^i$ is compatible with the path consisting of the first $i$ nodes of $\gamma$, (c) $\sigma^i$ is applicable to $S(I)$, and (d) if $i > 0$, $\sigma^{i-1}$ is a prefix of $\sigma^i$. The definition is by induction.

**Basis.** $\sigma^0$ is the empty schedule. It is obvious that this has the required properties.

**Induction step.** Let $i$ be an arbitrary positive integer, and assume that $\sigma^{i-1}$ with the required properties has been defined. Let the $i$th node of $\gamma$ be $(p,d,-)$. Then $\sigma^i = \sigma^{i-1} \cdot (p,m,d,A)$, where $m$ is the message defined as follows: If the message buffer of configuration $S \cdot \sigma^{i-1}(I)$ has no message for $p$ (i.e., no message of the form $(-, -, p)$), then $m = \lambda$; otherwise, $m$ is the oldest message sent to $p$ in the message buffer of $S \cdot \sigma^{i-1}(I)$ (i.e., there is no message $m'$ for $p$ in the message buffer of $S \cdot \sigma^{i-1}(I)$ and prefix $S'$ of $S \cdot \sigma^{i-1}$ such that the message buffer of $S'(I)$ contains $m'$ but not $m$). It is straightforward to verify that $\sigma^i$ has the required properties: It has length $i$, it is compatible with the first $i$ nodes of $\gamma$, it is applicable to $S(I)$, and it is an extension of $\sigma^{i-1}$.

Now define $\sigma^\infty$ to be the “limit” of the sequence $\sigma^0, \sigma^1, \ldots$, i.e., the infinite schedule whose prefix of length $i$ is $\sigma^i$. (This is well defined because, for all $i \in \mathbb{N}$, $\sigma^i$ has length $i$ and is a prefix of $\sigma^{i+1}$.) Clearly $\sigma^\infty$ is compatible with $\gamma$ and applicable to $S(I)$. Let $S^\infty = S \cdot \sigma^\infty$. We have:

- $S^\infty \in \text{Sch}(G^\infty_p, I)$. This follows from the fact that $S^\infty$ is compatible with path $g \cdot \gamma$ in $G^\infty_p$, and $S^\infty$ is applicable to $I$ (because $S$ is applicable to $I$, and $\sigma^\infty$ is applicable to $S(I)$).
- $\sigma^\infty$ is compatible with path $\gamma$ in $G^\infty_p | u$.

By Lemma 20, there is a time list $T^\infty$ such that $R_A = (F, H, I, S^\infty, T^\infty)$ is a run of $A$ using $D$ in $E$. It remains to prove that $R_A$ is admissible. We first note that each correct process takes infinitely many steps in $R_A$: this is because $S^\infty$ is compatible with $g \cdot \gamma$ and $\gamma$ contains infinitely many samples of each correct process. Furthermore, from the way we choose the message received in each step of $\sigma^\infty$, every message sent to a correct process is eventually received in $R_A$. So, $R_A$ has the required properties (6) and (7) of admissible runs.

The following lemma is an immediate consequence of Observation 11 and the definition of $\text{Sch}(-,-)$.

**Lemma 22.** For every correct process $p$, every process $q$, every time $t \in \mathbb{N}$, and every initial configuration $I$ of $A$, $\text{Sch}(G^\infty_p, I) \subseteq \text{Sch}(G^\infty_p, I)$.

The following lemma is an immediate consequence of Lemma 16 and the definition of $\text{Sch}(-,-)$.

**Lemma 23.** Let $p$ be a process and $I$ be an initial configuration of $A$. For each finite schedule $S \in \text{Sch}(G^\infty_p, I)$, there is a time $t$ such that, for all $t' \geq t$, $S \in \text{Sch}(G^\infty_p, I)$.

8. **$\Psi$ is necessary to solve quittable consensus.** In this section, we show that $\Psi$ is necessary to solve QC. Let $D$ be any failure detector that can be used to solve QC in some environment $E$; i.e., there is an algorithm $A$ that uses $D$ to solve QC in $E$. We must show that there is an algorithm that transforms $D$ into $\Psi$ in $E$. We do so by giving a transformation algorithm that uses $A$ and $D$ to emulate the output of $\Psi$—a failure detector that initially outputs $\bot$ and later behaves either like $(\Omega, \Sigma)$ or like $FS$. This transformation algorithm, denoted $T_{D \rightarrow \Psi}$, is shown in Figures 6–7, and is explained below.
Code for each process $p$:

initialize
1. $\Omega$-output$_p$ ← $p$
2. $\Sigma$-output$_p$ ← $\Pi$
3. $\Psi$-output$_p$ ← $\bot$
4. $k_p$ ← 0
5. $G_p$ ← empty graph
6. decision$_p$ ← $\bot$

 begun
10. /* Thread 1 — Build DAG of $D$-samples */
11. loop
12. receive a message $m$
13. $d_p$ ← $D_p$
14. if $m \neq \lambda$ then
15. $G_p$ ← $G_p \cup m$
16. $k_p$ ← $k_p + 1$
17. $v_p$ ← $(p, d_p, k_p)$
18. add node $v_p$ to $G_p$ and an edge from every other node in $G_p$ to $v_p$
19. send $G_p$ to every process

20. /* Thread 2 — Choose behavior of $\Psi$ */
21. $\forall j \in [0..n]$,
22. $I^j$ ← initial configuration of $A$ where the initial state of each $q \in [1..j]$ corresponds to proposal 1 and the initial state of each $q \in [j+1..n]$ corresponds to proposal 0
23. repeat
24. $\forall j \in [0..n]$,
25. $\text{Sch}(G_p, I^j)$ ← set of schedules of $A$ compatible with $G_p$ and applicable to $I^j$
26. $\forall j \in [0..n-1]$, let $S^j_p \in \text{Sch}(G_p, I^j)$ and $x^j_p$ be such that $p$ decides $x^j_p$ in $S^j_p(I^j)$
27. /* execute $A$ using $D$ to solve an instance of QC */
28. decision$_p$ := $\text{Propose}(0)$ /* propose 0 in this instance of QC */
29. else
30. let $i \in [0..n-1]$ be such that $x^i_p = 0$ and $x^{i+1}_p = 1$
31. decision$_p$ := $\text{Propose}(I^i, I^{i+1}, S^i_p, S^{i+1}_p)$ /* propose $(I^i, I^{i+1}, S^i_p, S^{i+1}_p)$ in this instance of QC */

32. /* Thread 3 — Output value of $\Psi$ */
33. wait until decision$_p$ $\neq \bot$
34. if decision$_p$ = 0 or decision$_p$ = $Q$ then / * a failure occurred */
35. $\Psi$-output$_p$ ← $\bot$
36. else /* $p$’s decision is a tuple with two initial configurations and two schedules */
37. $(I_0, I_1, S_0, S_1) :=$ decision$_p$
38. $u_p$ ← $v_p$
39. loop
40. $\Omega$-output$_p$ ← $\text{extract-leader}()$ /* see Figure 7 */
41. $\Sigma$-output$_p$ ← $\text{extract-quorum}()$ /* see Figure 7 */
42. $\Psi$-output$_p$ ← $(\Omega$-output$_p, \Sigma$-output$_p)$
43. coend

Fig. 6. Algorithm $T_{D \rightarrow \Psi}$. 

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.
8.1. Overview of the transformation. To make the presentation clearer, the code of each process $p$ in algorithm $T_{D\rightarrow \Psi}$ is given by three concurrent threads.\footnote{It is straightforward to write the code of $p$ as the code of a sequential process that can be directly expressed in our model (e.g., $p$ can execute the three concurrent threads in round-robin fashion).}

In Thread 1, $p$ builds a DAG of samples of failure detector $D$ using the algorithm discussed in section 7.2.1. In Thread 2, $p$ uses its current DAG to determine whether (a) it is legitimate for $\Psi$ to behave like $FS$ and output red permanently (because a failure occurred in the current run), or (b) it is possible to “extract” $(\Omega, \Sigma)$ in the current run. Then $p$ participates in an instance of QC to reach agreement with the other processes on (a) or (b). In Thread 3, $p$ produces the output of failure detector $\Psi$ according to the agreement reached in Thread 2. We now explain $T_{D\rightarrow \Psi}$ in more detail.

First recall that in the algorithm $A$ that solves QC, the value that a process $p$ proposes (i.e., its input value) is encoded in the initial state of $A(p)$. For each $j \in [0..n]$, let $I^j$ be the initial configuration of $A$ in which every process $q \in [1..j]$ proposes 1 and every process $q \in [j+1..n]$ proposes 0. Thus in any run starting from $I^0$, all processes propose 0, and starting from $I^n$ they all propose 1.

In $T_{D\rightarrow \Psi}$, each process $p$ starts by outputting $\bot$ (line 4), and then it executes three concurrent threads:

- In Thread 1, $p$ builds $G_p$, a DAG of failure detector values seen by processes in the current run.
- In Thread 2, $p$ repeatedly examines $n + 1$ sets of simulated schedules of algorithm $A$, namely $\text{Sch}(G_p, I^0), \ldots, \text{Sch}(G_p, I^n)$ (line 21), until it finds that for every $j \in [0..n]$, $\text{Sch}(G_p, I^j)$ contains a schedule $S_p^j$ such that $p$ decides some value $x_p^j$ in $S_p^j(I^j)$ (line 22). If for some $j \in [0..n]$, $x_p^j = Q$, then a failure occurred (in the failure pattern of the current run), and so $p$ knows that it is legitimate to extract $FS$ and output red in this run. Otherwise for every $j \in [0..n]$, $x_p^j$ is either 0 or 1, and in this case $p$ determines that it is.
possible to extract \((\Omega, \Sigma)\) in the current run.

At this point \(p\) participates in an instance of QC (by using the algorithm \(A\) and failure detector \(D\)) to agree with the other processes on whether to extract \(\mathcal{F}\) or to extract \((\Omega, \Sigma)\). Specifically, if \(p\) has determined that it is legitimate to output \(\text{red}\), then it proposes 0 in this execution of QC (lines 25–26). Otherwise, \(p\) proposes a tuple \((I^i, I^{i+1}, S^i_p, S^{i+1}_p)\), where \(i \in [0..n - 1]\) is such that \(x_p^i = 0\) and \(x_p^{i+1} = 1\) (lines 27–29). Note that such an index \(i\) must exist: Since \(p\) reaches line 27, by the condition on line 25 every \(x_p^j\) is either 0 or 1; by the validity property of QC, \(x_p^0 = 0\) and \(x_p^n = 1\); thus, for some \(i \in [0..n - 1]\), \(x_p^i = 0\) and \(x_p^{i+1} = 1\).

- In Thread 3, \(p\) computes the output values of \(\Psi\) according to the decision of the QC executed in Thread 2. If this decision is 0 or Q, then \(p\) stops outputting \(\perp\) and outputs \(\text{red}\) from that time on (lines 32–33). If the decision is some tuple \((I_0, I_1, S_0, S_1)\) (line 35), then \(p\) stops outputting \(\perp\) and starts extracting \(\Omega\) (line 38) and \(\Sigma\) (line 39), combining these two outputs into the output of \(\Psi\) (line 40). \(\Omega\) is extracted as in [9] (see section 8.2) and \(\Sigma\) is extracted using novel techniques (see section 8.3).

Note that in \(\mathcal{T}_{D \rightarrow \Psi}\), processes use the algorithm \(A\) for QC in two different ways and for different purposes. First, each process uses its DAG of failure detector samples to simulate many schedules of \(A\) to determine whether it is legitimate to output \(\text{red}\) or it is possible to extract \((\Omega, \Sigma)\) in the current run. Then processes actually participate in a real execution of \(A\) to reach a common decision on whether to output \(\text{red}\) or to extract \((\Omega, \Sigma)\). Finally, if processes decide to extract \((\Omega, \Sigma)\), they resume simulating schedules of \(A\) to effect this extraction.

For the remainder of section 8 we consider an arbitrary admissible run of algorithm \(\mathcal{T}_{D \rightarrow \Psi}\) using \(D\) in some arbitrary environment \(E\). Let \(F \in E\) be the failure pattern of this run and \(H \in \mathcal{D}(F)\) be its failure detector history.

### 8.2. Extracting \(\Omega\)

The specification of \(\Omega\) requires the following: At each process, \(\Omega\) outputs the id of a process; furthermore, if a correct process exists, then there is a time after which \(\Omega\) outputs the id of the same correct process \(p^*\) at every correct process. Note that this specification is trivially satisfied in runs where all processes are faulty. So in the rest of section 8.2, we assume that there is at least one correct process in the run under consideration, i.e, \(\text{correct}(F) \neq \emptyset\).

If in \(\mathcal{T}_{D \rightarrow \Psi}\) processes decide to extract \(\Omega\), they do so by executing the algorithm that extracts \(\Omega\) described in [9], as we now explain. As in [9], processes build a DAG of samples of failure detector \(D\) (Thread 1 in Figure 6). More precisely, processes build a DAG of the failure detector values that they see in \(H \in \mathcal{D}(F)\). By Observation 12, the limit DAG of all correct processes is the same. Let \(G^\infty\) denote that DAG. \((G^\infty)\) is well defined since a correct process exists.

Recall that \(I^i\) is the initial configuration of \(A\) in which the initial state of each process \(q \in [1..j]\) corresponds to proposal 1 and the initial state of each process \(q \in [j + 1..n]\) corresponds to proposal 0. Furthermore \(\text{Sch}(G^\infty, I^i)\) is the set of schedules of \(A\) that is compatible with \(G^\infty\) and applicable to \(I^i\). For each \(i \in [0..n]\), we organize the set of schedules in \(\text{Sch}(G^\infty, I^i)\) into a tree \(T^i\), called the limit tree (for initial configuration \(I^i\)). The nodes of this tree are the schedules in \(\text{Sch}(G^\infty, I^i)\), and there is an edge from node \(S\) to node \(S'\) if and only if there is a step \(e\) such that \(S' = S \cdot e\). We also define the limit forest \(\Upsilon\) to be the set of limit trees \(\{T^0, T^1, \ldots, T^n\}\).

In [9], algorithm \(A\) solves the binary version of consensus where processes propose only 0 or 1. So, by the validity property of consensus, the only possible decisions
are 0 or 1. In [9], it is shown that the root of each tree \( \Upsilon^i \) of the limit forest \( \Upsilon \) has a descendant \( S \) such that some correct process decides in \( S(I^i) \). The root of \( \Upsilon^i \) is \( v \)-valent for \( v \in \{0, 1\} \) if it has no descendant \( S \) such that some correct process decides \( u \neq v \) in \( S(I^i) \); the root of \( \Upsilon^i \) is multivalent if it is not \( v \)-valent for any \( v \in \{0, 1\} \). It is clear that the root of \( \Upsilon^i \) either is \( v \)-valent for exactly one value \( v \) or is multivalent. The limit forest \( \Upsilon \) has a critical index \( i \in [1..n] \) if and only if the root of \( \Upsilon^{i-1} \) is \( v \)-valent and the root of \( \Upsilon^i \) is \( u \)-valent for \( u \neq v \) (in which case index \( i \) is univalent critical) or the root of \( \Upsilon^i \) is multivalent (in which case index \( i \) is multivalent critical).

At a high level, the extraction of \( \Omega \) in [9] works as follows:

(a) First, it is shown that the limit forest \( \Upsilon \) has a critical index \( i \). This part of the proof uses the validity property of consensus.

(b) Then it is shown that for each critical index \( i \) of \( \Upsilon \), one can identify a corresponding process \( j \) that is necessarily correct in the failure pattern \( F \) of the current run.\(^8\) This part of the proof uses only the termination and uniform agreement properties of consensus; in particular, it does not rely on the validity property.

(c) Finally, it is shown how all correct processes can eventually converge on the smallest critical index \( i \) of \( \Upsilon \), and on the correct process \( j \) that corresponds to \( i \). This part of the proof also does not use the validity property of consensus.

The above three steps outline the extraction of \( \Omega \) when the given algorithm \( A \) solves binary consensus. Here we want to extract \( \Omega \) when \( A \) solves QC, and therefore \( A \) also solves binary QC where proposals are only 0 or 1. Note that binary consensus and binary QC share the same uniform agreement and termination properties, and they differ only in their validity property.

By the validity property of binary QC, there are now three possible decisions, 0, 1, or Q (instead of only 0 or 1). The definitions of \( v \)-valent or multivalent nodes remain the same, except that now \( v \in \{0, 1, Q\} \). The definitions of univalent and multivalent critical index \( i \) also remain the same.

To extract \( \Omega \) here, one may try to apply steps (a), (b), and (c) exactly as in [9]. Unfortunately, this does not quite work: with binary QC it is not always the case that the limit forest \( \Upsilon \) has a critical index. This is because, in contrast to consensus, the validity property of QC allows processes to decide Q if failures occur. To see why \( \Upsilon \) may not have a critical index, suppose some process crashes (in the failure pattern \( F \) of the current run). With QC, all the processes that decide “in the limit forest \( \Upsilon \)” may decide Q. In this case, the roots of all the trees in \( \Upsilon \) are Q-valent, \( \Upsilon \) has no critical index, and we cannot apply steps (b) and (c) above to extract the id of a correct process.

This is why, in our transformation algorithm of Figure 6, processes do not always attempt to extract \( \Omega \). As Lemma 24 below shows, however, if processes actually attempt to extract \( \Omega \) (on line 38), then a critical index does exist in the limit forest \( \Upsilon \). It is important to note that if \( \Upsilon \) has a critical index, then processes can converge on the identity of a correct process by applying steps (b) and (c) above, exactly as in [9]: This is because the correctness of steps (b) and (c) does not rely on the validity property of consensus (which is the only difference between consensus and QC).

**Lemma 24.** If any process reaches line 34 of algorithm \( T_{D \rightarrow \Psi} \), then the limit

---

\(^8\)If \( i \) is univalent critical, process \( i \) is necessarily correct; if \( i \) is bivalent critical, the limit tree \( \Upsilon^i \) contains a subgraph that reveals the identity of a process \( j \) that is necessarily correct.
forest Υ has a critical index.

Proof. If a process reaches line 34, then it decided some value \( v \neq \{0, \alpha, \beta, \delta, \gamma, \delta \} \) in the instance of QC that it executed in Thread 2 (line 29). In this instance of QC, each process \( p \) can only propose \( 0 \) (line 26) or a tuple of the form \((I^i, I^{i+1}, S_p, S_p\) (line 29). Thus, by part (i) of the validity property of QC, it must be that some process \( q \) proposed \((I^i, I^{i+1}, S_q, S_q\) for some index \( i \in [0..n] \) (line 29).

**Claim 24.1.** There are finite schedules \( S_0 \in \textbf{Sch}(G^\infty, I^i) \) and \( S_1 \in \textbf{Sch}(G^\infty, I^{i+1}) \) such that some correct process decides \( 0 \) in \( S_0(I^i) \) and some correct process decides \( 1 \) in \( S_1(I^{i+1}) \).

**Proof of Claim 24.1.** We prove the existence of \( S_0 \); the proof for \( S_1 \) is symmetric.

Since \( q \) proposed \((I^i, I^{i+1}, S_q, S_q\) on line 29, it must be that when \( q \) executed line 23, say at time \( t \), there is a schedule \( S_q^0 \in \textbf{Sch}(G_q^i, I^i) \) such that \( q \) decides \( 0 \) in \( S_q^0(I^i) \). Let \( p \) be any correct process. By Lemma 22, \( S_q^0 \in \textbf{Sch}(G_p^\infty, I^i) \). By Lemma 21, there is a schedule \( S^\infty \in \textbf{Sch}(G_p^\infty, I^i) \) that extends \( S_q^0 \), such that \( R_A = (F, H, I^i, S^\infty) \) is an admissible run of \( A \) (which solves QC) using \( D \) in \( E \). By the termination property of QC, there is a finite prefix \( S_0 \) of \( S^\infty \) such that \( p \) decides in \( S_0(I^i) \). Since \( S^\infty \in \textbf{Sch}(G_p^\infty, I^i) \), it follows that \( S_0 \in \textbf{Sch}(G_p^\infty, I^i) \) and, since \( G_p^i = G^\infty \), \( S_0 \in \textbf{Sch}(G^\infty, I^{i+1}) \). Since both \( S_0 \) and \( S_q^0 \) are prefixes of \( S^\infty \), one of them is a prefix of the other. Since \( q \) decides \( 0 \) in \( S_q^0(I^i) \), by the uniform agreement property of QC, \( p \) also decides \( 0 \) in \( S_0(I^i) \).}

By Claim 24.1, the root of \( \Upsilon^i \) is either 0-valent or multivalent, and the root of \( \Upsilon^{i+1} \) is either 1-valent or multivalent, or the root of \( \Upsilon^i \) is 0-valent and the root of \( \Upsilon^{i+1} \) is 1-valent. So, in all cases, there is a critical index in the limit forest \( \Upsilon \).

**8.3. Extracting \( \Sigma \).** To extract \( \Sigma \), \( p \) must continuously output a set of processes (quorum) such that the quorums of all processes always intersect, and eventually the quorums of correct processes contain only correct processes. This is done by the function extract-quorum() (lines 45–53) as follows.

Function extract-quorum() is called only on line 39, at which point \( p \) has agreed with other processes on a tuple \((I_0, I_1, S_0, S_1)\) (line 35). Process \( p \) maintains in variable \( u_p \) a "recent" failure detector sample of its own. This is initialized to \( u_p \)'s most recent sample when \( p \) executes line 36, and is updated to \( u_p \)'s most recent sample each time \( p \) outputs a new quorum (lines 49–50).

To determine the quorum to output, \( p \) examines every prefix \( S \) of \( S_0 \) and \( S_1 \), looking for a schedule \( \sigma_S \) that (a) uses only failure detector samples that are "fresher" than \( u_p \), (b) can be appended to \( S \) so that \( S \cdot \sigma_S \) is a simulated schedule of \( A \), and (c) \( p \) decides at the end of that schedule. More precisely, if \( S \) is a prefix of \( S_0b \) where \( b \in \{0, 1\} \), \( \sigma_S \) is required to be a schedule in \( \textbf{Sch}(G_p[u_p, S(I_b)]) \) (i.e., compatible with a path of samples at least as recent as \( u_p \) and applicable to \( S(I_b) \)), and \( p \) must decide in \( S \cdot \sigma_S(S(I_b)) \) (see the condition on line 47). If such a schedule \( \sigma_S \) can be found for every prefix \( S \) of \( S_0 \) and \( S_1 \), \( p \) computes a new quorum consisting of all processes that take steps in these \( \sigma_S \)'s (line 49). Otherwise, \( p \)'s quorum remains unchanged (lines 52–53).

Note how the sample in \( u_p \) acts as a "freshness barrier": \( p \)'s new quorum contains only processes that have taken samples at least as recent as \( u_p \). As we will see in the proof of Lemma 26 below, this (together with the fact that \( u_p \) contains ever more recent samples) ensures the completeness property of \( \Sigma \): The quorum output by a correct process \( p \) eventually contains only correct processes.

We will also see in the proof of Lemma 28 that this way of choosing quorums
ensures the intersection property of $\Sigma$: Every two quorums output by any two processes at any times intersect. Intuitively, this follows from the uniform agreement property of QC: If two quorums do not intersect, we would be able to construct an admissible run of the algorithm $A$ in which two different values in $\{0, 1, Q\}$ are decided, establishing a contradiction.

To prove that the completeness property holds, we first prove that the “freshness barrier” $u_p$ is updated infinitely often (line 50), and consequently a new quorum is also computed infinitely often (line 49).

**Lemma 25.** Every correct process $p$ that reaches line 34 of algorithm $T_{D\rightarrow\Psi}$ assigns a quorum to new-quorum$_p$ (line 49) and a node to $u_p$ (line 50) infinitely often.

**Proof.** Suppose some correct process $p$ reaches line 34. It is clear from the algorithm that $p$ either assigns both new-quorum$_p$ and $u_p$ infinitely often, or assigns both of them only a finite number of times (see lines 50 and 51). Suppose, for contradiction, that $p$ assigns $u_p$ only a finite number of times. Since $p$ reaches line 34, it also reaches line 36, and so it assigns $u_p$ at least once. Let $u$ be the last node of $G_p$ that $p$ assigns to $u_p$.

In the next two paragraphs we show that there is a time after which the condition of the if statement on line 47 is true forever. That is, for each prefix $S$ of $S_b$, where $b \in \{0, 1\}$, there is a finite schedule $\sigma_S$, compatible with a path in $G_p|u$, such that $S \cdot \sigma_S \in \text{Sch}(G_p, I_b)$ and $p$ decides in $S \cdot \sigma_S(I_b)$.

Since $p$ reaches line 34, it decided a value different from 0 or Q in the instance of QC that it executed in Thread 2 (line 26 or line 29). By part (i) of the validity property of QC, this decision value must be some tuple $(I_0, I_1, S_0, S_1)$ that some process $q$ proposed in Thread 2 (line 29). Thus, at some time $t$, for each $b \in \{0, 1\}$, $S_b \in \text{Sch}(G_p, I_b)$ (see line 23). By Lemma 22, for each $b \in \{0, 1\}$, $S_b \in \text{Sch}(G_p, I_b)$.

Consider any prefix $S$ of $S_b$, where $b \in \{0, 1\}$. Since $S_b \in \text{Sch}(G_p, I_b)$, it follows that $S \in \text{Sch}(G_p, I_b)$. By Lemma 21 there is a schedule $S^\infty \in \text{Sch}(G_p, I_b)$ that extends $S$ such that $R_A = (F, H, I_b, S^\infty, -)$ is an admissible run of $A$ (which solves QC) using $D$ in $E$. Furthermore, $S^\infty$ can be chosen so that $S^\infty = S \cdot \sigma^\infty_S$, for a schedule $\sigma^\infty_S$ that is compatible with a path in $G_p^\infty|u$. By the termination property of QC, there is a finite prefix $\sigma_S$ of $\sigma^\infty_S$ such that $p$ decides in $S \cdot \sigma_S(I_b)$. Since $S \cdot \sigma_S$ is a finite prefix of $S^\infty = S \cdot \sigma^\infty_S$, and $S^\infty \in \text{Sch}(G_p, I_b)$, by Lemma 23 it follows that there is some time $t_s$ such that, for all $t \geq t_s$, $S \cdot \sigma_S \in \text{Sch}(G_p, I_b)$. Also, since $\sigma^\infty_S$ is compatible with a path in $G_p^\infty|u$, and $\sigma_S$ is a finite prefix of $\sigma^\infty_S$, by Lemma 16, there is a time $t_S$ such that, for all $t \geq t_S$, $\sigma_S$ is compatible with a path in $G_p^t|u$. Let $t_1 = \max\{t_s, t_S \cdot S \text{ is a prefix of } S_0 \text{ or } S_1\}$. Let $t_2$ be the time of the last assignment to $u_p$, and let $t^* = \max(t_1, t_2)$. Thus, for all $t \geq t^*$, it is true that $u^*_p = u$ and, for every prefix $S$ of $S_b$, where $b \in \{0, 1\}$, there is a finite schedule $\sigma_S$, compatible with a path in $G_p^t|u^*_p$, such that $S \cdot \sigma_S \in \text{Sch}(G_p, I_b)$ and $p$ decides in $S \cdot \sigma_S(I_b)$. In other words, after $t^*$, the condition of the if statement on line 47 is always satisfied.

Since $p$ is correct and reaches line 37, it executes line 47 infinitely often. The first time after $t^*$ that $p$ executes that line, it finds that the condition of the if statement is satisfied, and assigns a node to $u_p$ on line 50. This occurs after time $t_2$, contradicting the definition of $t_2$.

**Lemma 26.** For every correct process $p$ that reaches line 34 of algorithm $T_{D\rightarrow\Psi}$, there is a time after which $\Sigma$-output$_p$ contains only correct processes.

**Proof.** Suppose some correct process $p$ reaches line 34. By Lemma 17, there is a sample $v^*$ of $p$ in $G_p^\infty$ such that $G_p^\infty|v^*$ contains only samples of correct processes.
By Lemma 17(a), there is a time after which any node \( v \) contained in variable \( v_p \) is a descendant of \( v^* \) in \( G_p^\infty \). By Lemma 25, there are infinitely many assignments to \( u_p \); in all of these, \( u_p \) is assigned the node in \( v_p \) (see lines 36 and 50). Thus, there is a time \( t^* \) such that for all \( t \geq t^* \), \( u_p^t \) is a descendant of \( v^* \) in \( G_p^\infty \). By Lemma 17(b), for all \( t \geq t^* \), \( G_p^t|u_p^t \) contains only samples of correct processes.

By Lemma 25, \( p \) computes a new quorum on line 49 infinitely often after time \( t^* \). Every quorum assigned to \( \Sigma -output_p \) (other than the initialization) is computed on line 49. Thus, it suffices to prove that any quorum assigned to \( new-quorum_p \) on line 49 after time \( t^* \) contains only correct processes.

Consider any such assignment, say at time \( t \geq t^* \) (see lines 47–49). The quorum assigned to \( new-quorum_p \) at time \( t \) is the union of certain sets of the form \( participants(\sigma) \), where \( \sigma \) is a schedule compatible with \( G_p^t|u_p^t \). Since \( t \geq t^* \), \( G_p^t|u_p^t \) contains only samples of correct processes. This implies that all processes in each such set \( participants(\sigma) \) are correct. Therefore, the quorum assigned to \( new-quorum_p \) at time \( t \) contains only correct processes.

We now prove that the quorums output by \( T_{D \rightarrow q} \) satisfy the intersection property of \( \Sigma \). Intuitively, we do so by showing that if two quorums do not intersect, then there are two runs of algorithm \( A \) such that (a) processes decide differently in these two runs, and (b) these two runs can be merged into a single run of \( A \)—a contradiction to the uniform agreement property of QC. To carry out this proof, we need the lemma that allows us to merge certain runs (Lemma 9 of section 7.1). More precisely, we use the following corollary of this lemma.

**Corollary 27.** Let \( R_0 = (F, H, I, S, I_0, -) \) and \( R_1 = (F, H, I, S, I_1, -) \) be two finite runs of \( A \) using \( D \) in \( E \) such that \( participants(S_0) \cap participants(S_1) = \emptyset \). If some process decides \( v_0 \) in \( R_0 \) and some process decides \( v_1 \) in \( R_1 \), then \( v_0 = v_1 \).

**Proof.** The proof is immediate from Lemma 9, the fact that \( A \) uses \( D \) to solve QC in \( E \), and the uniform agreement property of QC.

**Lemma 28.** For all processes \( p \) and \( q \), any two quorums assigned to \( \Sigma -output_p \) and \( \Sigma -output_q \) in algorithm \( T_{D \rightarrow q} \) intersect.

**Proof.** Suppose, for contradiction, that for some processes \( p \) and \( q \), there is a time when \( \Sigma -output_p = P \) and a time when \( \Sigma -output_q = Q \), but \( P \cap Q = \emptyset \).

First, observe that any set that \( p \) assigns to \( new-quorum_p \) on line 49 cannot be empty. This is because this set must include the participants of a schedule \( \sigma \) that is applicable to some initial configuration \( I_0 \) such that \( p \) decides in \( \sigma(I_0) \). It is easy to see that no process decides in any initial configuration, and so \( participants(\sigma) \neq \emptyset \). Thus, any set that \( p \) assigns to \( new-quorum_p \) on line 49 is nonempty. Similarly, any set that \( q \) assigns to \( new-quorum_q \) on line 49 is also nonempty.

Note that at any time, \( \Sigma -output_p = \Pi \) (at initialization) or \( \Sigma -output_p = new-quorum_p \). Similarly, \( \Sigma -output_q = \Pi \) or \( \Sigma -output_q = new-quorum_q \). Since \( new-quorum_p \) and \( new-quorum_q \) are not empty, it must be that the nonintersecting quorums \( P \) and \( Q \) are assigned to \( new-quorum_p \) and \( new-quorum_q \), respectively, on line 49.

Since \( p \) and \( q \) reach line 49, they also reach line 34, and so they decide a value different from 0 or \( Q \) in the instance of QC they execute in Thread 2 (line 26 or line 29). By the validity and uniform agreement properties of QC, it must be that \( decision_p = decision_q = (I_0, I_1, S_0, S_1) \) such that some process proposed \( (I_0, I_1, S_0, S_1) \) in Thread 2 (line 29). Note that \( I_0 \) and \( I_1 \) are initial configurations of algorithm \( A \) that differ only in the initial state of a single process, and \( S_0 \) and \( S_1 \) are schedules of \( A \) such that some process decides 0 in \( S_0(I_0) \) and 1 in \( S_1(I_1) \) (see lines 28–29).

For the notation defined in this and the next paragraph, see Figure 8. Let \( S_0 = e_1 \ldots e_l \) and \( S_1 = f_1 \ldots f_m \), where the \( e_i \)'s and \( f_j \)'s are steps. Let \( C_0 = I_0 \) and \( C_i = e_i(C_{i-1}) \) for \( i \in [1..l] \); similarly, \( D_0 = I_1 \) and \( D_j = f_j(D_{j-1}) \) for \( j \in [1..m] \).
Let $t$ be the time when $p$ first assigns $P$ to *new-quorum* on line 49. By the condition on line 47, for each $i \in [0..\ell]$, there is a schedule $\sigma_i^p$ such that $e_1 \ldots e_i \cdot \sigma_i^p \in \text{Sch}(G_{p}^{t}, I_0)$ and $p$ decides some value $x_i^p$ in $e_1 \ldots e_i \cdot \sigma_i^p(I_0)$.

Similarly, for each $j \in [0..m]$, there is a schedule $\tau_j^p$ such that $f_1 \ldots f_j \cdot \tau_j^p \in \text{Sch}(G_{p}^{t}, I_1)$ and $p$ decides some value $y_j^p$ in $f_1 \ldots f_j \cdot \tau_j^p(I_1)$. The quorum $P$ is the union of the participants in the $\sigma_i^p$'s and $\tau_j^p$'s. Similarly, let $t'$ be the time when $q$ first assigns $Q$ to *new-quorum*.

We define $\sigma_i^q$, $x_i^q$, $\tau_j^q$, and $y_j^q$ in an analogous manner. The quorum $Q$ is the union of the participants of the $\sigma_i^q$'s and $\tau_j^q$'s.

**Claim 28.1.** For all $i \in [0..\ell]$, $x_i^p = x_i^q$, and for all $j \in [0..m]$, $y_j^p = y_j^q$.

**Proof of Claim 28.1.** Since $e_1 \ldots e_i \cdot \sigma_i^p \in \text{Sch}(G_{p}^{t}, I_0)$, by Lemma 20, there is a run $R_0 = (F, H, I_0, e_1 \ldots e_i \cdot \sigma_i^p, -)$ of $A$ using $D$ in $\mathcal{E}$. Process $p$ decides $x_i^p$ in $R_0$. Similarly, there is a run $R_1 = (F, H, I_0, e_1 \ldots e_i \cdot \sigma_i^q, -)$ of $A$ using $D$ in $\mathcal{E}$, in which $q$ decides $x_i^q$. Since $P$ and $Q$ are disjoint, so are their subsets $\text{participants}(\sigma_i^p)$ and $\text{participants}(\sigma_i^q)$. Thus, by Corollary 27 (applied with $I = I_0$, $\hat{S} = e_1 \ldots e_i$, $S_0 = \sigma_i^p$, $S_1 = \sigma_i^q$, $v_0 = x_i^p$, and $v_1 = x_i^q$), we have that $x_i^p = x_i^q$. The proof that $y_j^p = y_j^q$ is analogous.

By Claim 28.1, we can now define $x_i = x_i^p = x_i^q$ and $y_j = y_j^p = y_j^q$.

**Claim 28.2.** For all $i \in [0..\ell-1]$, $x_{i+1} = x_i$, and for all $j \in [0..m-1]$, $y_{j+1} = y_j$.

**Proof of Claim 28.2.** Consider any $i \in [0..\ell-1]$, and let $r$ be the process that takes step $e_{i+1}$. Since $P$ and $Q$ are disjoint, $r \notin P$ or $r \notin Q$. Without loss of generality, suppose that $r \notin P$. In particular, $r \notin \text{participants}(\sigma_i^p)$. Also, again because $P$ and $Q$ are disjoint, so are their subsets $\text{participants}(\sigma_i^p)$ and $\text{participants}(\sigma_i^q)$. Therefore, $\text{participants}(\sigma_i^p)$ and $\text{participants}(e_{i+1} \cdot \sigma_{i+1}^q)$ are disjoint.

Since $e_1 \ldots e_i \cdot \sigma_i^p \in \text{Sch}(G_{p}^{t}, I_0)$, by Lemma 20, there is a run $R_0 = (F, H, I_0, e_1 \ldots e_i \cdot \sigma_i^p, -)$ of $A$ using $D$ in $\mathcal{E}$. Process $p$ decides $x_i$ in $R_0$. Also, since $e_1 \ldots e_{i+1} \cdot \sigma_{i+1}^q \in \text{Sch}(G_{q}^{t'}, I_0)$ there is a run $R_1 = (F, H, I_0, e_1 \ldots e_{i+1} \cdot \sigma_{i+1}^q, -)$ of $A$ using $D$ in $\mathcal{E}$, in which $q$ decides $x_{i+1}$. Thus, by Corollary 27 (applied with $I = I_0$, $\hat{S} = e_1 \ldots e_i$, $S_0 = \sigma_i^p$, $S_1 = e_1 \ldots e_{i+1} \cdot \sigma_{i+1}^q$, $v_0 = x_i$, and $v_1 = x_{i+1}$), we have that $x_i = x_{i+1}$. The proof that $y_j = y_{j+1}$ is analogous.

**Claim 28.3.** $x_0 = 0$ and $y_0 = 1$.

**Proof of Claim 28.3.** Recall that some process decides 0 in $S_0(I_0)$, and $p$ decides $x_i$ in $S_0 \cdot \sigma_i^p(I_0)$. Since $S_0 \cdot \sigma_i^p \in \text{Sch}(G_{p}^{t'}, I_0)$, by Lemma 20, there is a run $(F, H, I_0, S_0 \cdot \sigma_i^p, -)$ of $A$ using $D$ in $\mathcal{E}$. In this run, some process decides 0 and $p$ decides $x_i$. Since $A$ solves QC using $D$ in $\mathcal{E}$, by the uniform agreement property of QC, we have $x_i = 0$. By Claim 28.2 and a trivial induction, $x_i = 0$ for all $i \in [0..\ell]$. In particular, $x_0 = 0$. The proof that $y_0 = 1$ is analogous.

Since $P$ and $Q$ are disjoint, so are their subsets $\text{participants}(\sigma_0^p)$ and $\text{participants}(\tau_0^q)$. Let $r$ be the process such that $I_0$ and $I_1$ differ only in the initial state of $r$. Process $r$ does not take a step in at least one of $\sigma_0^p$ and $\tau_0^q$. Without loss of generality, assume that $r$ does not take a step in $\sigma_0^p$. Thus, $\sigma_0^p$ is also applicable to $I_1$, and $p$ decides the same value, $x_0$, in $\sigma_0^p(I_1)$ as in $\sigma_0^p(I_0)$. Since $\sigma_0^p \in \text{Sch}(G_{p}^{t'}, I_0)$, we also have that $\sigma_0^p \in \text{Sch}(G_{p}^{t'}, I_1)$. By Lemma 20, there is a run $R_0 = (F, H, I_1, \sigma_0^p, -)$ of $A$ using $D$ in $\mathcal{E}$. Process $p$ decides $x_0$ in $R_0$. Since $\tau_0^q \in \text{Sch}(G_{q}^{t'}, I_1)$, again by Lemma 20, there is a run $R_1 = (F, H, I_1, \tau_0^q, -)$ of $A$ using $D$ in $\mathcal{E}$. Process $q$ decides $y_0$ in $R_1$.

By Corollary 27 (applied with $I = I_1$, $\hat{S}$ being the empty schedule, $S_0 = \sigma_0^p$, $S_1 = \tau_0^q$, $v_0 = x_0$, and $v_1 = y_0$), we have that $x_0 = y_0$. This contradicts Claim 28.3 and completes the proof of Claim 28.

---

9We adopt the convention that, for $i = 0$, $e_1 \ldots e_i$ is the empty schedule.
8.4. Correctness of the transformation. We are now ready to show the following theorem.

**Theorem 29.** Algorithm $T_{D \rightarrow \Psi}$ transforms $D$ to $\Psi$.

**Proof.** Recall that algorithm $A$ uses $D$ to solve QC in $E$. As before, we consider an arbitrary admissible run of $T_{D \rightarrow \Psi}$ in $E$, where $F \in E$ is the failure pattern and $H \in D(F)$ is the failure history of this run.

To show that $T_{D \rightarrow \Psi}$ transforms $D$ to $\Psi$, we must prove that the values of the variables $\Psi$-output$_p$ conform to the specification of $\Psi$. By inspection of $T_{D \rightarrow \Psi}$, it is clear that $\Psi$-output$_p$ is either $\bot$, or red (in which case we say that it is of type $FS$), or a pair $(q, Q)$ where $q \in \Pi$ and $Q \subseteq \Pi$ (in which case we say that it is of type $(\Omega, \Sigma)$).

1. For each process $p$, $\Psi$-output$_p$ is initially $\bot$ (line 4). If $\Psi$-output$_p$ ever changes value, it becomes of type $FS$ forever (line 33) or of type $(\Omega, \Sigma)$ forever (line 40).

This follows by inspection of $T_{D \rightarrow \Psi}$.

2. For all distinct processes $p$ and $q$, it is impossible for $\Psi$-output$_p$ to be of type $FS$ and $\Psi$-output$_q$ to be of type $(\Omega, \Sigma)$.

This is because, by the uniform agreement property of QC, $p$ and $q$ cannot decide different values on lines 26 and 29; thus, they cannot execute in different branches of the if-then-else statement of lines 32–34.

3. For each correct process $p$, eventually $\Psi$-output$_p \neq \bot$.

To prove this, we first show that every correct process $p$ eventually completes the loop on lines 20–22. By Lemma 21 (taking $S$ to be the empty schedule), for each $j \in [0..n]$ there is a schedule $S^\infty \in \text{Sch}(G^\infty_p, P)$ such that $(F, H, P', S^\infty, -)$ is an admissible run of algorithm $A$ (which solves QC) using $D$ in $E$. By the termination property of QC, there is a finite prefix $S^j$ of $S^\infty$ such that $p$ decides in $S^j(P')$. By Lemma 23, there is a time $t^j$ such that for all $t \geq t^j$, $S^j \in \text{Sch}(G^j_p, P')$. Thus, after time $\max\{t : j \in [0..n]\}$, the exit condition on line 22 is true forever, and so eventually $p$ completes the loop.

We claim that, after completing the loop on lines 20–22, every correct process $p$ executes line 26 or line 29.

To show this claim, first note that since $p$ completes this loop, for every $j \in [0..n]$, there is a time $t_j$ and a schedule $S^j_p \in \text{Sch}(G^j_p, P)$ such that $p$ decides some value $x^j_p$ in $S^j_p(P)$. By Lemma 20, for all $j \in [0..n]$, there is a run $R^j_p = (F, H, P', S^j_p, -)$ of $A$ using $D$ in $E$. Since (a) $A$ solves QC using $D$ in $E$, (b) processes can propose only 0 or 1 in $R^j_p$, and (c) $p$ decides $x^j_p$ in $R^j_p$, then by the validity property of QC, $x^j_p \in \{0, 1, Q\}$. Furthermore, since no process proposes 1 in the run $R^j_p$, whose initial configuration is $I^0$, we have $x^0_p \in \{0, Q\}$. Similarly, $x^p_n \in \{Q\}$.

There are two possible cases:

- There is a $j \in [0..n]$ such that $x^j_p = Q$. In this case, $p$ executes line 26.
- For all $j \in [0..n]$, $x^j_p \neq Q$. In this case, for all $j \in [0..n]$, $x^0_p \in \{0, 1\}$; moreover, $x^0_p = 0$ and $x^0_p = 1$. So there must be some $i \in [0..n-1]$ such that $x^i_p = 0$ and $x^{i+1}_p = 1$. Thus, $p$ executes line 29.

Thus, $p$ executes line 26 or line 29, which shows the claim.

From this claim, all correct processes propose some value (on line 26 or 29) in an instance of QC executed in Thread 2. By the termination property of QC, all correct processes eventually decide in that instance, and so they all complete line 31. Thus, eventually every correct process $p$ sets $\Psi$-output$_p$ to a non-$\bot$ value on line 33 or 40.
For each process $p$ and time $t$, if $\Psi$-output$_p^t = \text{red}$, then a failure occurred by time $t$.

To see this, let $p$ be a process and $t$ be a time such that $\Psi$-output$_p^t = \text{red}$ (line 33). By lines 31–32, $p$ decided 0 or Q on line 26 or 29 at some time $t^* \leq t$. There are two possible cases:

(a) $p$ decided Q at time $t^* \leq t$. Then, by part (ii) of the validity property of QC, a failure occurred by time $t^* \leq t$.

(b) $p$ decided 0 at time $t^* \leq t$. Then, by Observation 2, there must be at least one process $q$ that proposes 0 and executes a step of the QC algorithm on line 26 by time $t^* \leq t$. This implies that there is a time $t'' \leq t'$, an index $j \in [0..n]$, and a schedule $S_q^j \in \text{Sch}(G_q^{t''}, P)$, such that $q$ decides Q in $S_q^j(I)$ (see lines 23–25). By Lemma 20, there is a list of times $T$, all at most $t''$, such that $(F, H, P, S_q^j, T)$ is a run of $A$ using $D$ in $E$. By part (ii) of the validity property of QC, this implies that a failure occurred by time $t'' \leq t' \leq t$.

If the $\Psi$-output variable of some process $p$ becomes of type $FS$ at time $t$, then we have the following:

- A failure occurred by time $t$.

  If $\Psi$-output$_p^t$ is of type $FS$, then, by inspection of $T_{D\rightarrow \Psi}$, $\Psi$-output$_p^t = \text{red}$. By (4), a failure occurred by time $t$.

- For every correct process $q$, there is a time after which $\Psi$-output$_q = \text{red}$.

  By (2) and (3), for every correct process $q$, there is a time after which $\Psi$-output$_q$ is of type $FS$. In $T_{D\rightarrow \Psi}$, the variable $\Psi$-output$_q$ can become of type $FS$ only by being set to $\text{red}$.

If the $\Psi$-output variable of some process becomes of type $(\Omega, \Sigma)$, then we have the following:

- (i) For every process $p$ and every time $t \in \mathbb{N}$, $\Omega$-output$_p^t \in \Pi$; furthermore, (ii) if a correct process exists, then there is a correct process $p^*$ and a time $t^*$ such that, for every correct process $p$ and every time $t \geq t^*$, $\Omega$-output$_p^t = p^*$.

  Part (i) is immediate by inspection of $T_{D\rightarrow \Psi}$. Part (ii) is trivial if all processes are faulty, so suppose that some correct process exists. By assumption, the $\Psi$-output variable of some process becomes of type $(\Omega, \Sigma)$. Then, by (2) and (3) above, eventually the $\Psi$-output variable of every correct process also becomes of type $(\Omega, \Sigma)$. So every correct process sets its $\Omega$-output variable repeatedly on line 38 using the extraction procedure described in [9]. Since some process reaches line 34, by Lemma 24, the limit forest $Y$ has a critical index. Thus, as we explained in section 8.2, we can now apply steps (b) and (c) of the proof of [9] to show that there is some correct process $p^*$ and a time $t^*$ such that, for every correct process $p$ and time $t \geq t^*$, $\Omega$-output$_p^t = p^*$. The only difference is that whenever the proof in [9] refers to a bivalent node, we now refer to a multivalent one, and whenever [9] refers to $u$-valent versus $v$-valent nodes for some distinct $u$ and $v$ in $\{0, 1\}$, here $u$ and $v$ are in $\{0, 1, Q\}$.

- (i) For every correct process $p$, there is a time after which $\Sigma$-output$_p$ contains only correct processes, and (ii) for all processes $p$ and $q$, any two quorums assigned to $\Sigma$-output$_p$ and $\Sigma$-output$_q$ intersect.

  Part (i) was shown in Lemma 26 and part (ii) in Lemma 28.

From (1)–(6) above, it follows that the values of the variables $\Psi$-output conform to the specification of $\Psi$, as defined in section 3: Initially, $\Psi$-output = ⊥
at each process; eventually, however, \( \Psi \text{-output} \) behaves either like the failure detector \((\Omega, \Sigma)\) at all correct processes or like the failure detector \(FS\) at all correct processes. The switch from \( \perp \) to \((\Omega, \Sigma)\) or \(FS\) is consistent at all processes, and a switch from \( \perp \) to \(FS\) can happen only if a failure occurred. \( \square \)

Since Theorem 29 holds for any environment \( E \) and any failure detector \( D \) that can be used to solve QC in \( E \), we conclude the following theorem.

**Theorem 30.** For every environment \( E \), if failure detector \( D \) can be used to solve QC in \( E \), then \( D \) can be transformed to \( \Psi \) in \( E \).

### 8.5. Binary versus multivalued quittable consensus

Our proof that \( \Psi \) is the weakest failure detector to solve QC uses the fact that, in QC, each process can propose any value in the infinite set \( \{0,1\}^* \); i.e., the proof used the fact that QC is *multivalued*.\(^\text{10}\) So one may ask whether \( \Psi \) is also the weakest failure detector to solve the binary version of QC where processes can propose only 0 or 1. The answer is affirmative.

To prove this, we use an algorithm by Mostéfaoui, Raynal, and Tronel [35] that converts any algorithm that solves binary consensus into an algorithm that solves multivalued consensus. With a straightforward modification, this conversion algorithm also works with quittable consensus: It converts any algorithm that solves binary QC (using some failure detector \( D \)) into one that solves multivalued QC (using the same failure detector \( D \)). This gives us the following theorem.

**Theorem 31.** For every environment \( E \), if failure detector \( D \) can be used to solve binary QC in \( E \), then \( D \) can be used to solve QC in \( E \).

Therefore, we have the following corollary.

**Corollary 32.** For every environment \( E \), \( \Psi \) is the weakest failure detector to solve binary QC in \( E \).

**Proof.** Let \( E \) be any environment.

(a) \( \Psi \) can be used to solve binary QC in \( E \). This is obvious since, by Theorem 8, \( \Psi \) can be used to solve QC in \( E \).

(b) Suppose \( D \) can be used to solve binary QC in \( E \). By Theorem 31, \( D \) can be used to solve QC in \( E \). So, by Theorem 30, \( D \) can be transformed to \( \Psi \) in \( E \). \( \square \)

### 9. Final remarks. Failure detector emulations

Intuitively, a failure detector \( D \) is weaker than a failure detector \( D' \) if processes can use \( D' \) to emulate \( D \). Two technical definitions of failure detector emulation have been proposed in the literature [9, 32]. In this paper we adopted the original definition of emulation given in [9] since we used parts of the proof given in that paper. As we explain below, however, our results also hold with the definition of emulation given in [32].

With the original definition of emulation [9], an implementation of \( D \) must maintain local variables that mirror the output of \( D \) at all times. The definition of emulation given in [32] is weaker: With this definition, an implementation of \( D \) is required to behave like \( D \) only when it is actually queried.\(^\text{11}\) The failure detectors \( \Psi \) and

\(^{10}\)Specifically, in Thread 2 of \( T_{D \rightarrow \Psi} \), processes may propose tuples of the form \((I, I', S, S')\) for some initial configurations and finite schedules of algorithm \( A \).

\(^{11}\)More precisely, if the implementation of \( D \) is queried at time \( t_1 \) and it replies with a value \( d \) at time \( t_2 \), then \( d \) must be a valid value of \( D \) at some time \( t \in [t_1, t_2] \); so, it is as if the query/reply occurred *atomically* at some time \( t \) within the interval of time during which the query/reply actually occurred. In other words, the behavior of the implementation of \( D \) is *linearizable* with respect to the specification of \( D \).
(Ψ, FS), which we proved here to be the weakest for QC and NBAC under the original definition of failure detector emulation, are also weakest for these problems under the definition of emulation given in [32]. In a nutshell, this is because (a) all the algorithms that we give here also work under the model of [32], and (b) if processes can emulate a failure detector \( D \) according to the strong definition of emulation of [9] (i.e., \( p \) is able to maintain a variable \( D\text{-output}_p \) that always mirrors the output of \( D \)), then processes can also emulate \( D \) according to the weaker definition of emulation of [32]: Whenever it is queried, \( p \) can just return the value of \( D\text{-output}_p \). For the same reasons, all the failure detectors that we are aware of to be weakest for a problem under the definition of emulation of [9] are also weakest for these problems under the definition of emulation of [32]; this includes the weakest failure detectors for consensus [9] and nonuniform consensus [19], set agreement [18, 38, 23], implementing an atomic register [15], and boosting obstruction-freedom to wait-freedom [26].

The newer definition of emulation given in [32] has two advantages over the original one of [9]. First, the original definition of emulation is more stringent than necessary: When using an emulated failure detector \( D \), it is sufficient that the emulated \( D \) behaves correctly only when it is queried—which is exactly what the newer definition stipulates. Second, the definition of emulation given in [32] is reflexive; i.e., for every failure detector \( D \), processes can use \( D \) to emulate \( D \). In contrast, as remarked by [32] and later in [11], the original definition of emulation is not reflexive: If the output of a failure detector \( D \) is sensitive to time, the processes, because they are asynchronous, may not be able to maintain variables that mirror the output of \( D \) at all times as the original definition of emulation requires. The nonreflexivity of the failure detector emulation under the original definition of emulation of [9] has no bearing on the results of this paper or on the other weakest failure detector results cited above: As we explained above, the same results also hold with the newer definition of emulation given in [32] which does satisfy reflexivity.

**Granularity of steps.** As in the models of Fischer, Lynch, and Paterson [20] and Chandra, Hadzilacos, and Toueg [9], in our model a process can send a message \( m \) to every process in an atomic step. Since a sender cannot fail in “the middle” of a step that sends \( m \) to all, our model has the following property: If any process receives \( m \), then every correct process eventually receives \( m \) (*). One may ask whether our results also hold in another model, let’s call it model \( B \), where a process can send a message \( m \) to only one process in an atomic step. To answer this question, note that in model \( B \) (where processes may crash but links are reliable) one can implement Uniform Reliable Broadcast (URB) [29], a communication primitive that provides the property (*) of our model. Since model \( B \) can emulate the atomic “send \( m \) to all” of our model, it is easy to see that our results also hold in model \( B \).

**Systems that are not asynchronous.** It is worth noting that since our model is that of an asynchronous system augmented with failure detectors, the algorithms that emulate failure detectors are also asynchronous [11]. So the weakest failure detectors that result from such emulations are also asynchronous in the sense that their output values could be delayed for any finite time. Some previous works explored failure detectors in systems that are not purely asynchronous. For example Aguilera et al. investigated the use of “fast” failure detectors to speed up agreement algorithms in some synchronous systems [4]. In another body of work, researchers considered the definition and implementation of failure detectors for systems where message delays and losses follow some probability distribution [14, 8, 36]. It may be interesting to investigate QC and NBAC in systems that are not asynchronous, and to determine
whether these problems have weakest failure detectors in these systems. This, however, is beyond the scope of this paper.

**Systems with a majority of correct processes.** In environments where a majority of processes are correct it is easy to implement the quorum failure detector Σ: Each process periodically sends “join-quorum” messages and takes as its present quorum any majority of processes that respond to that message. Therefore, in such environments Ψ is equivalent to a simpler failure detector, one which outputs just Ω instead of (Ω, Σ).

**Future failures.** Our definitions of QC and NBAC do not allow a process to quit or abort because of a future failure. We could have defined these problems in a way that allows such behavior, as in fact is the case in some specifications of NBAC in the literature. Our results also hold with these definitions, provided we make a corresponding change to the definitions of the failure detectors FS and Ψ: They are now allowed to output red in executions with failures even before a failure occurs.

**Appendix A. Proof of Lemma 9.** We prove Lemma 9 from section 7.1.

**Lemma 9.** Let R₀ = (F, H, I, S · S₀, T₀ · T₀) and R₁ = (F, H, I, S · S₁, T₁ · T₁) be two finite runs of an algorithm A using failure detector D in some environment E such that |T₀| = |T₁| = |S| and participants(S₀) ∩ participants(S₁) = Ø. Let R = (F, H, I, S · S, T · T) be a merging of R₀ and R₁. Then:

(a) R is also a run of A using D in E.

(b) For each b ∈ {0, 1} and each process p ∈ participants(Š · S₀), the state of p is the same in Š · S(I) as in Š · S₀(I).

**Proof.** To show that R is a run of A using D in E, we first note that F ∈ E, H ∈ D(F), and I is indeed an initial configuration of A. It now suffices to show that R satisfies properties (1)–(5) of runs (see section 2.7). The fact that Š · S and Š · T have the same length (property (2)) is obvious from the definition of R. The fact that in R no process takes a step after it has crashed, and that the failure detector value in each step is consistent with the history H (property (3)), follows from the way R is constructed from R₀ and R₁, and the fact that R₀ and R₁ have this property. Š · T is nondecreasing (property (4)) because each of T₀ · T₀ and T₁ · T₁ is nondecreasing, Š is chosen to be whichever of T₀ and T₁ has the smallest maximum element, and T is obtained by merging T₀ and T₁ in nondecreasing order. The times of the steps in R respect the causal precedence relation (property (5)) because R₀ and R₁ have this property, and no process takes a step in both S₀ and S₁. It remains to prove that Š · S is applicable to I (property (1)).

For the purposes of this proof, if σ is a schedule and i ∈ {0, 1, . . . , |σ|}, we denote by σ𝑖 the prefix of σ that has length i (σ₀ is the empty schedule). Also, for the suffix S of the schedule of the merged run R (i.e., the portion of the schedule of R produced by merging S₀ and S₁), and b ∈ {0, 1}, let f₀(b(i)) be the number of steps of S(b) that come from S₀. Using a straightforward induction, we can show the following for all i ∈ {0, 1, . . . , |S|}:

(i) For all b ∈ {0, 1}, the set of messages between processes in participants(Š · S₀) (i.e., messages of the form (p, −, q) where p, q ∈ participants(Š · S₀)) in the message buffer of configuration Š · S′(I) is equal to the set of messages between processes in participants(Š · S₀) in the message buffer of configuration Š · S′₀(b(i))(I).

(ii) For all b ∈ {0, 1}, the state of any process p ∈ participants(Š · S₀) is the same in Š · S′(I) as in Š · S′₀(b(i))(I).
Below we use (i) to show that, for each \( i \in \{1, 2, \ldots, |S|\} \), \( S[i] \) is applicable to \( \hat{S} \cdot S^{i-1}(I) \). This proves that \( \hat{S} \cdot S \) is applicable to \( I \).

Let \( S[i] = (p, m, d, A) \). Let \( b \in \{0, 1\} \) be such that \( p \in \text{participants}(S_b) \) (such a \( b \) exists because every step of \( S \) is in either \( S_0 \) or \( S_1 \)). Thus, \( (p, m, d, A) \) is step \( f_b(i) \) of \( S_b \). Since \( R_b \) is a run, \( \hat{S} \cdot S_b \) is applicable to \( I \). In particular, step \( (p, m, d, A) \) of \( S_b \) is applicable to \( \hat{S} \cdot S_b^{f_b(i-1)}(I) \). Note that \( f_b(i-1) = f_b(i)-1 \). So, \( (p, m, d, A) \) is applicable to \( \hat{S} \cdot S_b^{f_b(i-1)}(I) \). Thus, \( m \) is in the message buffer of \( \hat{S} \cdot S_b^{f_b(i-1)}(I) \). Furthermore, it is a message between processes in \( \text{participants}(\hat{S} \cdot S_b) \). This is because (1) being in the message buffer of \( \hat{S} \cdot S_b^{f_b(i-1)}(I) \), \( m \) was sent by a process in \( \text{participants}(\hat{S} \cdot S_b^{f_b(i-1)}) \), and (2) \( p \), the recipient of \( m \), is the process that takes the \( f_b(i) \)th step of \( S_b \). By (i), \( m \) is in the message buffer of \( \hat{S} \cdot S^{i-1}(I) \). So, \( (p, m, d, A) \) is applicable to \( \hat{S} \cdot S^{i-1}(I) \), as wanted.

Part (b) of the lemma follows directly from (ii), taking \( i = |S| \). □

REFERENCES


[38] P. Zielinski, Anti-\( \Omega \): The weakest failure detector for set agreement, Distributed Computing, 22 (2010), pp. 335–348.