



**HAL**  
open science

## A Connectivity Model for Agreement in Dynamic Systems

Carlos Gómez-Calzado, Arnaud Casteigts, Alberto Lafuente, Mikel Larrea

► **To cite this version:**

Carlos Gómez-Calzado, Arnaud Casteigts, Alberto Lafuente, Mikel Larrea. A Connectivity Model for Agreement in Dynamic Systems. 21st International Conference on Parallel Processing (EURO-PAR 2015), Aug 2015, Vienne, Austria. pp.333 - 345, 10.1007/978-3-662-48096-0\_26 . hal-01472943

**HAL Id: hal-01472943**

**<https://hal.science/hal-01472943>**

Submitted on 21 Feb 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# A Connectivity Model for Agreement in Dynamic Systems <sup>★</sup>

Carlos Gómez-Calzado<sup>1</sup>, Arnaud Casteigts<sup>2</sup>, Alberto Lafuente<sup>1</sup>, and Mikel Larrea<sup>1</sup>

<sup>1</sup> University of the Basque Country UPV/EHU, Spain  
{carlos.gomez, alberto.lafuente, mikel.larrea}@ehu.es

<sup>2</sup> LaBRI, University of Bordeaux, France  
arnaud.casteigts@labri.fr

**Abstract.** The consensus problem is a fundamental paradigm in distributed systems, because it captures the difficulty to solve other agreement problems. Many current systems evolve with time, e.g., due to node mobility, and consensus has been little studied in these systems so far. Specifically, it is not well established how to define an appropriate set of assumptions for consensus in dynamic distributed systems. This paper studies a hierarchy of three classes of time-varying graphs, and provides a solution for each class to the problem of Terminating Reliable Broadcast (TRB). The classes introduce increasingly stronger assumptions on timeliness, so that the trade-off between weakness versus implementability and efficiency can be analysed. Being TRB equivalent to consensus in synchronous systems, the paper extends this equivalence to dynamic systems.

## 1 Introduction

The consensus problem is a central paradigm in distributed systems, as it represents many agreement problems, e.g., leader election, atomic commitment and total-order broadcast. Solving consensus has attracted a lot of attention in dependable computing and has generated fundamental results. In this regard, it is known that in crash-prone asynchronous distributed systems it is impossible to solve consensus deterministically due to the impossibility of distinguishing between “slow” processes and crashed ones, a result known as FLP impossibility [8]. Alternatively, consensus can be easily solved in synchronous systems, where perfect failure detection can be implemented [5].

Most of the research on consensus has considered a static distributed system with permanent connectivity among nodes. In many current distributed systems, however, these assumptions are not valid any more. Instead, these new systems exhibit a dynamic behavior, with nodes joining the system, leaving it or just moving, which implies uncertain connectivity conditions. Indeed, and unlike in classical static systems, these events are no longer considered incorrect or sporadic behaviors, but rather the natural dynamics of the system.

---

<sup>★</sup> Research supported by the Spanish Research Council, grant TIN2013-41123-P, the Basque Government, grant IT395-10, and the University of the Basque Country UPV/EHU, grant UFI11/45. Carlos Gómez-Calzado is recipient of a doctoral fellowship from the Basque Government. This work has been carried out during a research stay of Carlos Gómez-Calzado at the University of Bordeaux. The authors want to thank Antonio Fernández Anta for his valuable comments.

Clearly, even the synchrony assumptions of classical (static) models of distributed systems are not enough to solve agreement problems in dynamic systems. For example, having an upper bound on link latencies is pointless if the link is not available at the time of transmission of the message. Note however that the nodes could still communicate using an alternative path in the network. Thus, assumptions should consider the overall system connectivity, which encourages for a holistic approach to model dynamic distributed systems.

In recent years there was a rising interest in modeling dynamic distributed systems from the perspective of graph theory. In this regard, there exist several works that study the solvability of deterministic problems, including consensus, in highly-dynamic systems [2,3,7,11,12]. However, regarding consensus, none of them lowers the assumptions to the realm of *temporal* connectivity, i.e., not requiring that the graph be connected at every instant, but only that paths exist over time and space (temporal path, aka *journeys*). The time-varying graph formalism [4] (TVG, for short) provides a useful qualitative framework to model dynamic distributed systems. In this formalism, the dynamic network is represented as a graph, together with a presence function that tells whether a given edge is present at a given time and a latency function that tells how long it takes to cross a given edge at a given time. In [4], Casteigts et al. define a hierarchy of classes of dynamic networks, most of which are based on temporal connectivity concepts. Among them, the *recurrent connectivity* class requires that a journey exists between any two nodes infinitely often (that is, recurrently). Nevertheless, this class lacks the necessary timeliness (i.e. time bounds in communication) to describe the specific assumptions that are required by synchronous agreement algorithms, such as TRB, to terminate. One of the goals of our paper is to extend some of the existing TVG classes by introducing timeliness constraints, together with practical considerations, and analyze the impact of these new constraints on solving consensus.

## Our contribution

In this paper, we address timeliness in evolving systems (i.e., time-varying graphs, TVG) from a synchronous point of view, i.e., systems where the transmission delay of messages is bounded and the bound is known a priori by the processes. The resulting set of concepts and mechanisms makes it possible to describe system dynamics at different levels of abstraction and with a gradual set of assumptions.

We first formulate a very abstract property on the temporal connectivity of the TVG, namely, that the temporal diameter (i.e. maximum duration of a foremost journey) of a component in the TVG is always bounded by  $\Delta$ . We refer to such a component as a  $\Delta$ -component, and define the concept of correct process in terms of this component. We then specify a version of the Terminating Reliable Broadcast problem (TRB) for  $\Delta$ -components, which we relate to the ability of solving agreement at component level.

Although  $\Delta$ -components are proven to be a sufficient concept at the most abstract level, they rely on non implementable communication patterns in message-passing systems. Indeed, the solution to TRB proposed in this abstract model relies on an oracle that provides the algorithm with instantaneous knowledge of the appearance of an edge. Unfortunately, this oracle does not have a straightforward implementation in terms of real processes and communication links. Therefore, we introduce a first constraint to force the existence of journeys whose edges presence duration is lower-bounded by some duration  $\beta$  (which holds a relation

to the maximal latency of a link), thereby enabling repetitive communication attempts to succeed eventually. These journeys are called  $\beta$ -journeys and their existence makes it possible to implement the TRB algorithm without oracle. We then look at a further constrained class of TVG, inspired by the work of Fernández-Anta et al [7], whereby the local appearance of the edge used by every next hop of (at least one of the possibly many)  $\beta$ -journeys also must be bounded by some duration  $\alpha$ , yielding to the concept of  $(\alpha, \beta)$ -journeys. The existence of recurrent  $(\alpha, \beta)$ -journeys allows the nodes to stop sending a message  $\alpha$  time after they receive it, which is much more efficient.

The rest of the paper is organized as follows. Section 2 introduces basic time-varying graph notations, used in Section 3 to define the abstract timely connectivity model based on  $\Delta$ -components. In the same section we redefine the TRB problem with respect to  $\Delta$ -components and give a solution to it. Then, in Section 4, we introduce  $\beta$ -journeys (and the corresponding  $\beta$ -components), which we show to be sufficient to implement an effective (i.e., oracle-free) version of the algorithm. We then define  $(\alpha, \beta)$ -journeys and components, and discuss their advantages (and disadvantages) over  $\beta$ -journeys. In Section 5, we describe how consensus can be solved by using the TRB implementations introduced in Sections 3 and 4. We finally conclude in Section 6 with open questions and future work.

## 2 Time-Varying Graphs

A recent framework called *time-varying graphs*, proposed by Casteigts et al. [4], aims to provide a precise formalism for describing dynamic networks. As usual, the entities of the system and the communication links between them are represented as a graph. More specifically, a time-varying graph (TVG, for short) is defined as a tuple  $\mathcal{G} = (V, E, \mathcal{T}, \rho, \zeta)$ , where:

- $V$  is the set of communicating entities (or nodes, or processes, interchangeably).
- $E$  is the set of edges (or links, interchangeably) that interconnect the nodes in  $V$ . In this work, all edges are undirected.
- $\mathcal{T}$  is the *lifetime* of  $\mathcal{G}$ , i.e. the interval of time over which the graph is defined. It is a subset of the temporal domain  $\mathbb{T}$ , itself being  $\mathbb{N}$  or  $\mathbb{R}^+$  depending on whether time is discrete or continuous (in this work, it is continuous). For convenience, both endpoints of  $\mathcal{T}$  are referred to as  $\mathcal{T}^-$  and  $\mathcal{T}^+$ , the latter being possibly  $+\infty$ .
- $\rho : E \times \mathcal{T} \rightarrow \{true, false\}$ , i.e., the *presence* function, indicates whether a given edge is present at a given time (i.e.,  $\rho(e, t) = true$  iff edge  $e$  is present at time  $t$ )
- $\zeta : E \times \mathcal{T} \rightarrow \mathbb{T}$ , i.e., the *latency* function, indicates how long it takes to send a message across a given edge for a given emission time (assuming the edge is present at that time)

The kind of network we are addressing is possibly disconnected at every instant. Still, a form of communication can be achieved over time by means of *journeys* (a.k.a. *temporal path*). Formally, a journey  $\mathcal{J} = \{(e_1, t_1), (e_2, t_2), \dots, (e_k, t_k)\}$  is a sequence such that  $(e_1, e_2, \dots, e_k)$  is a valid path in the underlying graph  $(V, E)$ , and (1) for every  $i \in [1, k]$  edge  $e_i$  is present at time  $t_i$  long enough to send a message across (formally,  $\rho(e_i, t_i + \delta) = true$  for all  $\delta \in [0, \zeta(e_i, t_i)]$ ), and (2) the times when edges are crossed (we also say *activated*) and the corresponding latencies allow a sequential traversal (formally,  $t_{i+1} \geq t_i + \zeta(e_i, t_i)$  for all  $i \in [1, k)$ ). What makes this form of connectivity *temporal* is the fact that a journey can pause in between hops, e.g. if the next link is not yet available.

Given a journey  $\mathcal{J}$ ,  $\text{departure}(\mathcal{J})$  and  $\text{arrival}(\mathcal{J})$  denote respectively its starting time  $t_1$  and its ending time  $t_k + \zeta(e_k, t_k)$ . Journeys can be thought of as paths over time, having both a *topological length*  $k$  (i.e., the number of *hops*) and a *temporal length* (i.e., a duration)  $\text{arrival}(\mathcal{J}) - \text{departure}(\mathcal{J}) = t_k + \zeta(e_k, t_k) - t_1$ . Note that journeys describe *opportunities* of communication between an emitter and a receiver.  $\mathcal{J}_{\mathcal{G}}^*$  is the set of all such opportunities over  $\mathcal{G}$ 's lifetime, while  $\mathcal{J}_{(p,q)}^* \subseteq \mathcal{J}_{\mathcal{G}}^*$  are those journeys from  $p$  to  $q$ . A simplified way of denoting the existence of a journey between a process  $p$  and a process  $q$ , when the context of  $\mathcal{G}$  is clear, is  $p \rightsquigarrow q$ . Finally, the graph is *temporally connected* if for every  $p, q \in V, p \rightsquigarrow q$ .

An induced sub-TVG  $\mathcal{G}' \subseteq \mathcal{G}$  is obtained by restricting either the set of vertices  $V' \subseteq V$  or the lifetime  $\mathcal{T}' \subseteq \mathcal{T}$ , resulting in the tuple  $(V', E', \mathcal{T}', \rho', \zeta')$  such that:

- $(V', E')$  is the subgraph of  $(V, E)$  induced (in the usual sense) by  $V'$
- $\rho' : E' \times \mathcal{T}' \rightarrow \{\text{true}, \text{false}\}$  where  $\rho'(e, t) = \rho(e, t)$
- $\zeta' : E' \times \mathcal{T}' \rightarrow \mathbb{T}$  where  $\zeta'(e, t) = \zeta(e, t)$

If only the lifetime is restricted, say to some interval  $[t_a, t_b)$ , then the resulting graph  $\mathcal{G}'$  is called a *temporal* subgraph of  $\mathcal{G}$  and denoted  $\mathcal{G}_{[t_a, t_b)}$ . The *temporal diameter* of a graph  $\mathcal{G}$  at time  $t$  is the smallest duration  $d$  such that  $\mathcal{G}_{[t, t+d)}$  is temporally connected.

Finally, following Bhadra and Ferreira in [1], we consider a temporal variant of connected components (hereafter, simply called *components*), which are maximal sets of nodes  $V' \subseteq V$  such that  $\forall p, q \in V', p \rightsquigarrow q$ . Two variants are actually considered, whether the corresponding journeys can also use nodes that are in  $V \setminus V'$  (*open* components) or not (*closed* components). Observe that a close component is equivalent to an induced sub-TVG being temporally connected.

### 3 A Timely Model for Dynamic Systems

This section focuses on the analysis of timeliness in dynamic systems at the most abstract level, i.e. considering only a general communication bound  $\Delta$  for end-to-end communication. We first provide a set of definitions related to this bound, which leads to the formulation of a new class of TVGs that is a strict subset of Class 5 (*recurrent connectivity*) in [4]. We then specify a solution to the problem of Terminating Reliable Broadcast (TRB) in the corresponding context.

#### 3.1 Definitions

We define the concept of bounded-time journey as follows:

**Definition 1.** A journey  $\mathcal{J}$  is a  $\Delta$ -journey if and only if  $\text{arrival}(\mathcal{J}) - \text{departure}(\mathcal{J}) \leq \Delta$ .

Based on  $\Delta$ -journeys we define the concept of bounded-time component. Unlike components, we require here that connectivity also be recurrent by definition.

**Definition 2.** A  $\Delta$ -component in  $\mathcal{G} = (V, E, \mathcal{T}, \rho, \zeta)$  is a set  $V' \subseteq V$  such that for every  $t$  in  $[\mathcal{T}^-, \mathcal{T}^+ - \Delta]$ , for every  $p, q$  in  $V'$ , there exists a  $\Delta$ -journey from  $p$  to  $q$  in  $\mathcal{G}_{[t, t+\Delta)}$ .

Similarly to components,  $\Delta$ -components can be open or closed, depending on whether the  $\Delta$ -journeys use nodes in  $V \setminus V'$ . Observe that, a graph behaving in an open way provides flexibility in mobility, and therefore, a model allowing open  $\Delta$ -components is weaker (in the sense that it requires less assumption) than a model strictly based on closed  $\Delta$ -components. Henceforth we assume that in our system model  $\Delta$ -components are by default open.

Informally,  $\Delta$ -components allow us to think about subsets of nodes behaving timely with each other. Hence, nodes in a  $\Delta$ -component are also *timely connected*. We define the (parametrized) class of timely (and recurrently) connected TVGs  $\mathcal{TC}(\Delta)$  as follows:

**Definition 3.**  $\mathcal{G} \in \mathcal{TC}(\Delta) \iff V$  is a  $\Delta$ -component.

### 3.2 Terminating Reliable Broadcast in $\mathcal{TC}(\Delta)$

According to [6], consensus is equivalent to Terminating Reliable Broadcast in static synchronous systems. We take this as a starting point and describe here a solution for TRB in the scope of a  $\Delta$ -component.

We assume that processes know a global time. Processing times are negligible with respect to communication time. The system is composed by processes that can crash/recover, and leave/join the system. Processes that crash or leave the system, even if they recover or join again later, are by definition excluded from any  $\Delta$ -component, however since we assume the existence of open  $\Delta$ -components, they can punctually take part on various journeys.

Recall that a distributed system  $\mathcal{G}$  may have several  $\Delta$ -components. There may exist values of  $\Delta$  for which a process belongs to different components, which are thus overlapping. However, since every component is *recurrently* connected, then overlapping components become naturally merged as the value for  $\Delta$  increases, and transitively, there must exist a sufficiently large value of  $\Delta$  such that all remaining components are disjoint. Henceforth, we consider  $\Delta$  to be (an upper bound on) such a value.

We define now which processes are *correct* in terms of the classical terminology. In a classical partitioned system it can be considered that a process  $p$  behaves correctly in its partition, and incorrectly with respect to the other partitions in the system. Similarly, in our  $\Delta$ -component based system a process  $p$  behaves correctly with respect to the  $\Delta$ -component  $p$  belongs to, e.g.  $C$ . Hence, a member of a  $\Delta$ -component  $C$  is by definition correct with respect to  $C$ . However  $p$  could still sporadically communicate timely with *some* processes in another  $\Delta$ -component,  $C'$ . Obviously, we consider  $p$  incorrect with respect to  $C'$ , but a message  $m$  from  $p$  received by some process in  $C'$  should either be delivered by all processes in  $C'$ , or by none of them in order to hold the agreement property of reliable broadcast.

Thus, in  $\mathcal{TC}(\Delta)$  a set of properties should be hold by a process *with respect to* a  $\Delta$ -component in order to provide  $\Delta$ -TRB:

- $\Delta$ -Termination: Every process in the same  $\Delta$ -component eventually delivers some message.
- $\Delta$ -Validity: If a process in a  $\Delta$ -component broadcasts a message  $m$ , then all processes in the same  $\Delta$ -component eventually deliver  $m$ .
- $\Delta$ -Agreement: If a process in a  $\Delta$ -component delivers a message  $m$ , then all processes in the same  $\Delta$ -component eventually deliver  $m$ .

- $\Delta$ -Integrity: For any message  $m$ , every process in the same  $\Delta$ -component delivers  $m$  at most once, and if it delivers  $m \neq SF$  (sender faulty) then the sender( $m$ ) must have broadcast  $m$ .

As usual, the broadcast at time  $t_{init}$  of a message  $m$  is considered in the scope of  $m$ .

To guarantee the  $\Delta$ -Agreement property we should correctly understand when a message  $m$  broadcast by  $p \notin C$  should be delivered by all processes in  $C$ . If  $p$  has been able to propagate  $m$  to some process  $q \in C$ , then we assume that there exists a  $\Delta$ -journey from  $p$  to  $q$ . Observe that this assumption is consistent with the fact that our model allows the existence of open  $\Delta$ -components.

```

To  $\Delta$ -TRBroadcast a message  $m$  at time  $t_{init}$ :
if  $p = p_B$  then
  for all edge  $e = (p_B, -)$  s.t.  $\rho(e, t_{init}) = true$  do
     $send(m)$  on  $e$  at  $t_{init}$ 

On appearance of  $e = (p_B, -)$  at time  $t \in [t_{init}, t_{init} + \Delta)$ :
 $send(m)$  on  $e$  at  $t$ 

On reception of a message  $m$  for the first time at time  $t_{rec} \in [t_{init}, t_{init} + 2\Delta)$ :
if  $p \neq p_B$  then
  for all edge  $e = (p, -)$  s.t.  $\rho(e, t_{rec}) = true$  do
     $send(m)$  on  $e$  at  $t_{rec}$ 

On appearance of  $e = (p, -)$  at time  $t \in [t_{init}, t_{init} + 2\Delta)$ :
if a message  $m$  has been previously received then
   $send(m)$  on  $e$  at  $t$ 

At time  $t_{init} + 2\Delta$ :
if a message  $m$  has been previously received then
   $\Delta$ -TRDeliver( $m$ )
else
   $\Delta$ -TRDeliver( $SF$ )

```

**Fig. 1.** Terminating Reliable Broadcast for  $\mathcal{TC}(\Delta)$ .

A solution to the TRB problem is described in Figure 1 (see [9] for details). Informally, the distinguished process  $p_B$   $\Delta$ -TRBroadcasts a message  $m$  by sending  $m$  on all its active edges at time  $t_{init}$ . Whenever an edge in  $p_B$ 's neighborhood appears<sup>3</sup>,  $p_B$  also sends  $m$  on that edge. Every other process  $p$ , upon reception of  $m$  for the first time, forwards  $m$  on all its active edges, as well as upon the appearance of a new edge. Finally, at time  $t_{init} + 2\Delta$  every process  $p$   $\Delta$ -TRBdelivers either  $m$  (if  $m$  has been received) or  $SF$ .

We explain next why a time of  $2\Delta$  is necessary and sufficient to deliver  $m$ .

<sup>3</sup> We assume here the existence of an abstract *oracle* to capture events of edge appearance. In the next section we will board the implementation of such an oracle.

Observe that, since we are assuming that  $p_B$  could be not in  $C$ ,  $p_B$  could not be able to communicate to all nodes in  $C$  in  $\Delta$  time, (otherwise  $p_B \in C$ ), thus, after  $m$  is resent by  $q$ , every process in  $C$  will receive  $m$  into a second  $\Delta$  time interval. Henceforth the bound for a process in  $C$  to  $TRDeliver$  a message is  $2\Delta$ .

**Theorem 1.** *The specification in Figure 1 satisfies the properties of  $\Delta$ -TRB in  $\mathcal{TC}(\Delta)$ .*

## 4 Implementability of TRB

The specification of TRB provided in Figure 1 relies on an “oracle” available at every process  $p$ , which informs  $p$  instantaneously upon appearance of a new edge in its neighbourhood. Such an abstraction has been recently used by Raynal et al. [13] to implement a broadcast algorithm for recurrent dynamic systems. However, a strict implementation of this oracle in a real system is far from being trivial, as we discuss now.

Observe that the only temporal assumption on  $\Delta$ -journeys is that they satisfy a given upper-bound  $\Delta$  in its temporal length, thus the duration of an edge may be as short as the latency of the message. In consequence, an implementation of this oracle should be able to allow the sending of a message at the very same time that the edge get activated, which is unrealistic since the oracle should be able to predict the behaviour of the links in a real network. Alternatively, an algorithm could continuously send message  $m$  along the whole time interval in the hope that one of the sending attempts will success in the appearance of an edge. Observe, however, that this iteration would require a period of time zero between two consecutive sends. In other words, the algorithm should be able to send an infinite number of messages per unit of time, which is impossible.

Therefore, additional assumptions should be introduced in order to provide an implementation for the above specification of TRB. Specifically we first propose an extra assumption that allow to maintain active the edge not only for communicating the message but also to detect its appearance.

### 4.1 (Lower)-bounding the edge stability

We assume that the edge latency is bounded, i.e, there exist a bound on  $\max\{\zeta(e, t) : t \in \mathcal{T}, e \in E\}$ , that we call  $\zeta_{MAX}$ . Additionally, we assume that edges are active at least  $\beta$  time. Let us call  $\beta$ -edge an edge that fulfils this bounded disposability. For this new model we define  $\beta$ -journeys as follows:

**Definition 4.** *A  $\beta$ -journey  $\mathcal{J} = \{(e_1, t_1), \dots, (e_k, t_k)\}$  is a  $\Delta$ -journey such that:*

1.  $\zeta_{MAX} < \beta \leq \Delta$ .
2.  $\forall i \in [1, k], e_i$  is a  $\beta$ -edge.
3. *the times when edges are activated and their corresponding latencies allow a bounded sequential traversal (formally,  $\forall i \in [1, k], t_{i+1} \geq t_i + \beta$ ).*

We now define  $\beta$ -components as a subset of  $\Delta$ -components that uses  $\beta$ -journeys. Formally:

**Definition 5.** A  $\beta$ -component is a  $\Delta$ -component where a set  $V' \subseteq V$  satisfies that  $\forall t \in [\mathcal{T}^-, \mathcal{T}^+ - \Delta]$ ,  $V'$  is a  $\beta$ -journey based temporal component in  $\mathcal{G}_{[t, t+\Delta]}$ .

We define the parametrized timely connectivity class  $\mathcal{TC}'(\beta)$  as follows:

**Definition 6.**  $\mathcal{G} \in \mathcal{TC}'(\beta) \iff V$  is a  $\beta$ -component.

### TRB in $\mathcal{TC}'(\beta)$

We give now a TRB algorithm for the  $\mathcal{TC}'(\beta)$  model, which is shown in Figure 2.

```

1  $W \leftarrow \text{value} \in (0, \beta - \zeta_{MAX}]$ 
2 To  $\Delta$ -TRBroadcast a message  $m$  at time  $t_{init}$ :
3   if  $p = p_B$  then
4     while  $\text{now}() < t_{init} + \Delta$  do
5       send( $m$ ) to all
6       wait( $W$ )
7 On reception of a message  $m$  for the first time at time  $t_{rec} \in [t_{init}, t_{init} + 2\Delta]$ :
8    $\Delta$ -TRDeliver( $m$ )
9   if  $p \neq p_B$  then
10    while  $\text{now}() < t_{rec} + \Delta$  do
11      send( $m$ ) to all
12      wait( $W$ )
13 At time  $t_{init} + 2\Delta$ :
14   if  $p$  has not  $\Delta$ -TRDelivered any message then
15      $\Delta$ -TRDeliver( $SF$ )

```

**Fig. 2.** Terminating Reliable Broadcast for  $\mathcal{TC}'(\beta)$ .

In the algorithm proposed in Figure 2 a process  $p_B$  sends at time  $t_{init}$  a message  $m$  by  $\Delta$ -TRBroadcasting it, and  $p_B$  keeps sending  $m$  each  $W$  time in order to assure the correct send of  $m$  by every  $\beta$ -journey. Observe that, according to the definition of  $\beta$ -edge, for a  $\beta$ -edge  $e = (p, q)$  in a  $\beta$ -journey, if process  $p$  sends a message  $m$  on  $e$  each  $W \leq \beta - \zeta_{MAX}$  time during  $\Delta$ ,  $q$  will receive  $m$  at least once (see [9] for the complete correctness proof). When a process  $p$  receives the message  $m$  automatically  $\Delta$ -TRDelivers  $m$ , and additionally, if  $p \neq p_B$ ,  $p$  sends  $m$  each  $W$  time during  $\Delta$ . Finally, if a process does not receive the message  $m$ , at time  $t_{init} + 2\Delta$ , it  $\Delta$ -TRDelivers the special message  $SF$ .

**Theorem 2.** The specification in Figure 2 satisfies the properties of  $\Delta$ -TRB in  $\mathcal{TC}'(\beta)$ .

## 4.2 (Upper)-bounding the edge appearance

Observe that, in the algorithm in Figure 2 messages are forwarded during the whole  $\Delta$  interval. This is necessary because the ending edge of a  $\beta$ -journey could be activated at a time as

late as  $t_{init} + \Delta - \beta$ . It is apparent that more efficient implementations of a TRB algorithm in terms of number of messages could be envisaged if stronger connectivity assumptions are introduced in the model. Specifically, in this section we introduce an additional timely assumption on the appearance of edges.

We adopt the assumption of [7], where, besides  $\beta$ , a bound  $\alpha$  on the appearance of links is defined. We define a new type of journey, that we call  $(\alpha, \beta)$ -journey. Formally:

**Definition 7.** A  $(\alpha, \beta)$ -journey  $\mathcal{J} = \{(e_1, t_1), \dots, (e_k, t_k)\}$  is a  $\beta$ -journey such that:

1. The appearance of  $e_1$  is bounded by  $\alpha$ .
2. The appearance of the subsequent edges are also bounded by  $\alpha$ . Formally,  $t_{i+1} \leq t_i + \zeta(e_i, t_i) + \alpha$  for all  $i \in [1, k]$ .

We define a  $(\alpha, \beta)$ -component as follows:

**Definition 8.** A  $(\alpha, \beta)$ -component is a  $\beta$ -component where a set  $V' \subseteq V$  satisfies that  $\forall t \in [\mathcal{T}^-, \mathcal{T}^+ - \Delta]$ ,  $V'$  is a  $(\alpha, \beta)$ -journey based temporal component in  $\mathcal{G}_{[t, t+\Delta]}$ .

We define the parametrized timely connectivity class  $\mathcal{TC}''(\alpha, \beta)$  as follows:

**Definition 9.**  $\mathcal{G} \in \mathcal{TC}''(\alpha, \beta) \iff V$  is a  $(\alpha, \beta)$ -component.

### TRB in $\mathcal{TC}''(\alpha, \beta)$

The algorithm in Figure 3, describes a TRB algorithm executable in a  $\mathcal{TC}''(\alpha, \beta)$  dynamic system.

The new bound  $\alpha$ , altogether with the latency bound  $\beta$  and  $\zeta_{MAX}$ , allows to calculate global system bounds, namely the period  $W$  and a time to deliver  $\Gamma$ , strictly in terms of specific network parameters. In the algorithm proposed in Figure 3 a process  $p_B$   $\Delta$ -TRBroadcast a message  $m$  at time  $t_{init}$  by sending each  $W$  time  $m$  until the time is strictly higher than  $t_{init} + \alpha$ , in order to assure the correct sending of  $m$  by every  $(\alpha, \beta)$ -journey. When a process  $p$  receives the message  $m$  at time  $t_{rec}$  for the first time, automatically  $\Delta$ -TRDelivers  $m$  and, additionally, if  $p \neq p_B$ ,  $p$  sends  $m$  each  $W$  until the time is strictly higher than  $t_{rec} + \alpha$ . Finally, if any of the process in  $p$  does not receive the message  $m$  at time  $t_{init} + \Gamma$ ,  $\Delta$ -TRDelivers the special message  $SF$  denoting the sender failure.

A detailed explanation of how  $\Gamma$  is obtained can be found in the extended version of this paper [9]. It is important to note that in the TRB algorithm for  $\mathcal{TC}''(\alpha, \beta)$ , differently to the previous classes, processes need to know the network diameter, which is bounded by  $|V| - 1$ . This is a consequence of the fact of considering strictly local bounds in  $\mathcal{TC}''(\alpha, \beta)$ . Instead, both  $\mathcal{TC}(\Delta)$  and  $\mathcal{TC}'(\beta)$  rely on a system-wide bound,  $\Delta$ .

**Theorem 3.** The specification in Figure 3 satisfies the properties of  $\Delta$ -TRB in  $\mathcal{TC}''(\alpha, \beta)$ .

### 4.3 Relating timely classes

We have defined a hierarchy of classes with increasingly stronger timely assumptions. Being  $\mathcal{TC}(\Delta)$ ,  $\mathcal{TC}'(\beta)$  and  $\mathcal{TC}''(\alpha, \beta)$  the parametrized classes, we define now for each one the union of all its possible instances:

```

1  $W \leftarrow \text{value} \in (0, \beta - \zeta_{MAX}]$ 
2  $\Gamma \leftarrow (\lceil \frac{\alpha}{W} \rceil + (|V| - 2)\lceil \frac{\zeta_{MAX} + \alpha}{W} \rceil)W + \zeta_{MAX}$ 
3 To  $\Delta$ -TRBroadcast a message  $m$  at time  $t_{init}$ :
4   if  $p = p_B$  then
5     send( $m$ ) to all
6   repeat
7     wait( $W$ )
8     send( $m$ ) to all
9   until  $now() > t_{init} + \alpha$ 

10 On reception of a message  $m$  for the first time at time  $t_{rec}$ :
11    $\Delta$ -TRDeliver( $m$ )
12   if  $p \neq p_B$  then
13     send( $m$ ) to all
14   repeat
15     wait( $W$ )
16     send( $m$ ) to all
17   until  $now() > t_{rec} + \alpha$ 

18 At time  $t_{init} + \Gamma$ :
19   if  $p$  has not  $\Delta$ -TRDelivered any message then
20      $\Delta$ -TRDeliver( $SF$ )

```

**Fig. 3.** Terminating Reliable Broadcast for  $\mathcal{TC}''(\alpha, \beta)$ .

$$\begin{aligned} \mathcal{G} \in \mathcal{TC}^* &\iff \exists \Delta \neq \infty : \mathcal{G} \in \mathcal{TC}(\Delta) \\ \mathcal{G} \in \mathcal{TC}'^* &\iff \exists \beta \neq \infty : \mathcal{G} \in \mathcal{TC}'(\beta) \\ \mathcal{G} \in \mathcal{TC}''^* &\iff \exists \alpha, \beta \neq \infty : \mathcal{G} \in \mathcal{TC}''(\alpha, \beta) \end{aligned}$$

In spite of the different strength of the parametrized classes, we show in [9] that  $\mathcal{TC}''^* \equiv \mathcal{TC}'^*$ . Besides,  $\mathcal{TC}'^* \subset \mathcal{TC}^*$  and  $\mathcal{TC}''^* \subset \mathcal{TC}^*$ .

## 5 From $\Delta$ -TRB to $\Delta$ -Consensus in Dynamic Systems

In this section we analyse the equivalence between TRB and consensus, originally stated for synchronous static systems [6], in terms of a dynamic system as the one we have modelled.

In the previous sections we have presented three  $\Delta$ -TRB algorithms in the scope of respectively  $\Delta$ -,  $\beta$ - and  $(\alpha, \beta)$ -components. We show now how the consensus problem can be reduced<sup>4</sup> to a  $\Delta$ -TRB problem. We will refer as  $\Delta$ -Consensus to this kind of consensus in the scope of  $\Delta$ -components.

By the properties of  $\Delta$ -TRB, it is straightforward to define the  $\Delta$ -Consensus properties as follows:

<sup>4</sup> We say that a problem A can be reduced to a problem B if A can be solved using B.

- $\Delta$ -Termination: Every process in the  $\Delta$ -component eventually decides.
- $\Delta$ -Agreement: Every process in the  $\Delta$ -component decides the same value.
- $\Delta$ -Validity: The decided value is a proposed one.

Without loosing generality we focus here on  $\Delta$ -Consensus using the  $\Delta$ -TRB specification of Figure 1 for the  $\mathcal{TC}(\Delta)$  Class.

```

1 Vector  $V_p(i) \leftarrow \perp : i \in [0, |V|)$ 
2 To  $\Delta$ -Propose  $v$  at time  $t_{init}$ :
3    $\Delta$ -TRBroadcast( $v$ )
4 On  $\Delta$ -TRDeliver( $m$ ) by  $q$ :
5    $V_p(q) \leftarrow m$ 
6 At time  $t_{init} + 2\Delta$ :
7    $\Delta$ -decide( $V_p(\min(i : V_p(i) \neq SF))$ )

```

**Fig. 4.**  $\Delta$ -TRB based  $\Delta$ -Consensus algorithm for  $\mathcal{TC}(\Delta)$ .

The resulting  $\Delta$ -Consensus algorithm is shown in Figure 4. Every process  $p$  holds a vector  $V_p$  initialized to  $\perp$ . At time  $t_{init}$ ,  $|V|$  instances of  $\Delta$ -TRB are started, one per process, being each process the sender in one instance. Every process  $p$  records in vector  $V_p(q)$  the message  $m_q$  delivered from process  $q$  (or  $SF$  in case  $m_q$  has not been received on time). At time  $t_{init} + 2\Delta$ ,  $p$  decides on the first non- $SF$  value of  $V_p$ .

Note that solving consensus at system level would require a second  $\Delta$ -TRB round to agree on the decision of the majority, provided that the temporal interval  $[\mathcal{T}^-, \mathcal{T}^+]$  in which the  $\Delta$ -component is defined covers both rounds. In other words, the stability of  $\Delta$ -components must be temporally extended to solve consensus at system level.

## 6 Conclusions

In this paper we studied how to introduce timeliness in evolving systems so that the resolution of agreement problems (specifically consensus) is possible. On the basis of previous works, we have adopted the concept of journey or temporal path and have introduced the necessary timeliness (i.e., time bounds) to describe the specific assumptions that are required by an agreement algorithm to terminate and satisfy the consensus properties.

We have first proposed a general class,  $\mathcal{TC}(\Delta)$ , with a very abstract property on the temporal connectivity of the TVG to provide the necessary stability conditions, namely, that the temporal diameter of a recurrent component in the TVG is bounded. We refer to such a component as a  $\Delta$ -component. To approach the consensus problem we have defined a TRB specification in terms of  $\Delta$ -components,  $\Delta$ -TRB. However,  $\Delta$ -TRB is not implementable in  $\mathcal{TC}(\Delta)$  by message-passing without zero processing time assumptions. Henceforth, by introducing increasingly stronger connectivity assumptions, we have provided two *implementable*

connectivity classes, namely  $\mathcal{TC}'(\beta)$  and  $\mathcal{TC}''(\alpha, \beta)$ , as well as two respective implementations of  $\Delta$ -TRB in these classes. Finally, we have shown that consensus at  $\Delta$ -component level is easily reduced to  $\Delta$ -TRB.

An open issue is the search of the weakest connectivity class that allows to implement  $\Delta$ -TRB (and henceforth consensus) in message-passing systems. Of additional interest is to extend the proposed classes to partially synchronous models. In this regard, in [10], leader election is implemented in a partially synchronous system with dynamic partitions that could be modelled as  $\Delta$ -components.

## References

1. Bhadra, S., Ferreira, A.: Complexity of connected components in evolving graphs and the computation of multicast trees in dynamic networks. *Ad-Hoc, Mobile, and Wireless Networks*, pp. 259–270. Springer (2003)
2. Biely, M., Robinson, P., Schmid, U.: Agreement in directed dynamic networks. *Proceedings of the 19th International Colloquium on Structural Information and Communication Complexity, SIROCCO 2012*, Reykjavik, Iceland. *Lecture Notes in Computer Science*, vol. 7355, pp. 73–84. Springer (2012)
3. Casteigts, A., Flocchini, P., Mans, B., Santoro, N.: Shortest, Fastest, and Foremost Broadcast in Dynamic Networks. *International Journal of Foundations of Computer Science*, to appear (accepted January 2015), World Scientific.
4. Casteigts, A., Flocchini, P., Quattrociocchi, W., Santoro, N.: Time-varying graphs and dynamic networks. *International Journal of Parallel, Emergent and Distributed Systems* 27(5), 387–408 (2012)
5. Chandra, T.D., Toueg, S.: Unreliable failure detectors for reliable distributed systems. *Journal of the ACM* 43(2), 225–267 (1996)
6. Feldman, P., Micali, S.: Optimal algorithms for byzantine agreement. *Proceedings of the 20th Annual ACM Symposium on Theory of Computing, STOC 1998*, pp. 148–161. ACM (1988)
7. Fernández-Anta, A., Milani, A., Mosteiro, M.A., Zaks, S.: Opportunistic information dissemination in mobile ad-hoc networks: The profit of global synchrony. *Distributed Computing* 25(4), 279–296 (2012)
8. Fischer, M.J., Lynch, N.A., Paterson, M.: Impossibility of distributed consensus with one faulty process. *Journal of the ACM* 32(2), 374–382 (1985)
9. Gómez-Calzado, C., Casteigts, A., Lafuente, A., Larrea, M.: A Connectivity Model for Agreement in Dynamic Systems. *Technical Report EHU-KAT-IK-01-15*, University of the Basque Country UPV/EHU (February 2015), available at <http://www.sc.ehu.es/acwlaalm/research/EHU-KAT-IK-01-15.pdf>
10. Gomez-Calzado, C., Lafuente, A., Larrea, M., Raynal, M.: Fault-tolerant leader election in mobile dynamic distributed systems. *Proceedings of the 19th IEEE Pacific Rim International Symposium on Dependable Computing, PRDC 2013*, pp. 78–87 (2013)
11. Kuhn, F., Lynch, N.A., Oshman, R.: Distributed computation in dynamic networks. *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010*, Cambridge, Massachusetts, USA, 5–8 June 2010, pp. 513–522. ACM (2010)
12. Kuhn, F., Oshman, R.: Dynamic networks: models and algorithms. *ACM SIGACT News* 42(1), 82–96 (2011)
13. Raynal, M., Stainer, J., Cao, J., Wu, W.: A simple broadcast algorithm for recurrent dynamic systems. *Proceedings of the IEEE 28th International Conference on Advanced Information Networking and Applications, AINA 2014*, pp. 933–939 (2014)