Singularly Optimal Randomized Leader Election

Shay Kutten
Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, Haifa, Israel
kutten@technion.ac.il

William K. Moses Jr.
Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, Haifa, Israel
wkmjr3@gmail.com

Gopal Pandurangan
Department of Computer Science, University of Houston, Houston, TX, USA
gopal@cs.uh.edu

David Peleg
Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel
david.peleg@weizmann.ac.il

Abstract
This paper concerns designing distributed algorithms that are *singularly optimal*, i.e., algorithms that are simultaneously time and message optimal, for the fundamental leader election problem in networks. Our main result is a randomized distributed leader election algorithm for asynchronous complete networks that is essentially (up to a polylogarithmic factor) singularly optimal. Our algorithm uses $O(n)$ messages with high probability and runs in $O(\log^2 n)$ time (with high probability) to elect a unique leader. The $O(n)$ message complexity should be contrasted with the $\Omega(n \log n)$ lower bounds for the deterministic message complexity of leader election algorithms (regardless of time), proven by Korach, Moran, and Zaks (TCS, 1989) for asynchronous algorithms and by Afek and Gafni (SIAM J. Comput., 1991) for synchronous networks. Hence, our result also separates the message complexities of randomized and deterministic leader election. More importantly, our (randomized) time complexity of $O(\log n)$ for obtaining the optimal $O(n)$ message complexity is significantly smaller than the long-standing $\tilde{\Theta}(n)$ time complexity obtained by Afek and Gafni and by Singh (SIAM J. Comput., 1997) for message optimal (deterministic) election in asynchronous networks. Afek and Gafni also conjectured that $\Theta(n)$ time would be optimal for message-optimal asynchronous algorithms. Our result shows that randomized algorithms are significantly faster.

Turning to synchronous complete networks, Afek and Gafni showed an essentially singularly optimal deterministic algorithm with $O(\log n)$ time and $O(n \log n)$ messages. Ramanathan et al. (Distrib. Comput. 2007) used randomization to improve the message complexity, and showed a randomized algorithm with $O(n)$ messages but still with $O(\log n)$ time (with failure probability $O(1/\log^{O(1)} n)$). Our second result shows that synchronous complete networks admit a *tightly* singularly optimal randomized algorithm, with $O(1)$ time and $O(n)$ messages (both bounds are optimal). Moreover, our algorithm's time bound holds with certainty, and its message bound holds with high probability, i.e., $1 - 1/n^c$ for constant $c$.

Our results demonstrate that leader election can be solved in a simultaneously message and time-efficient manner in asynchronous complete networks using randomization. It is open whether this is possible in asynchronous general networks.

2012 ACM Subject Classification Theory of computation → Distributed algorithms; Mathematics of computing → Probabilistic algorithms; Mathematics of computing → Discrete mathematics

Keywords and phrases Leader election, Asynchronous systems, Randomized algorithms, Singularly optimal, Complete networks

---

1 Throughout, “with high probability” means probability at least $1 - 1/n^c$, for a constant $c > 0$. 
1 Introduction

Leader election is a classical and fundamental problem in distributed computing with numerous applications; see e.g., [30, 12, 11, 15, 23, 25, 26, 28, 47, 27, 31, 32, 46, 20, 33, 38, 13, 24, 41, 3, 42, 16, 29, 2]. The goal is to select a unique node, called the leader, from a set of nodes. An arbitrary subset of nodes can wake up spontaneously at arbitrary times and start the election algorithm by sending messages over the network. When the algorithm terminates, a unique node \( v \) must be elected as leader and be known to all nodes.

Election is important both theoretically and because of the multiple applications such as implementing databases and data centers [19, 9, 48], locks [7], file servers [14, 10], broadcast and multicast [40, 8], and virtually every global task [5]. In many of these applications, the network is treated as being virtually complete. With the advent of large-scale and resource-constrained networks such as peer-to-peer systems [22, 21, 5, 13, 50] and ad hoc and sensor networks (e.g., [13, 50]), it is often desirable to achieve low cost and scalable leader election. Leader election has been studied extensively over the years in various distributed computing models starting with the standard CONGEST model of networks (in particular, ring, complete networks, and arbitrary networks), radio networks, peer-to-peer networks, population protocols, and programmable matter to name a few.

Our goal in this paper is to design leader election algorithms in distributed networks such that each is efficient with respect to both of the two fundamental complexity measures, namely messages and time. Unfortunately, designing distributed network algorithms that are simultaneously time- and message-efficient has proved to be a challenging task. Consequently, research in the last three decades has focused mainly on optimizing either one of the two measures separately, typically at the cost of neglecting the other. There has been significant recent progress in obtaining algorithms that are essentially optimal in both measures (or at least work well under both measures to the extent possible) for various problems such as leader election, minimum spanning tree and shortest paths [27, 37, 11, 17]. In particular, as defined in [37] (see also [15]), the following two notions are of interest for any given problem:

— Singular optimality: A problem enjoys singular optimality if it has a distributed algorithm that is optimal (or at least optimal up to a polylogarithmic factor) with respect to both measures simultaneously.

— Time-message trade-off: A problem exhibits a time-message trade-off if it fails to admit a singularly optimal solution, namely, algorithms of better time complexity for it necessarily incur higher message complexity and vice versa.

The singularly optimal results mentioned earlier for leader election, minimum spanning tree, and shortest paths [27, 37, 11, 17], crucially apply only to synchronous networks. A main motivation for this work is to study whether leader election admits singularly optimal algorithms or exhibits a time-message tradeoff in asynchronous networks. Unlike synchronous networks which admit a leader election algorithm which is singularly optimal (up to a logarithmic factor) [27], it is not known whether asynchronous networks admit such an
algorithm (see “Background and Prior Work”).

In this paper, we focus on the classical problem of leader election in complete networks, which itself has been studied extensively for nearly four decades. In such a network, every pair of nodes is connected by a bidirectional communication link and in the beginning, no node has any knowledge of any other nodes, including their identities (if any). Like many prior works (e.g., [1, 47, 18, 12, 2], and others), this paper focuses on the more challenging asynchronous model and provides the first-known singularly optimal algorithm.

It is clear that the best possible time bound for leader election in complete networks is constant if there is no restriction on the message complexity. The situation with respect to the message complexity is more nuanced. (The message lower bound is trivially $\Omega(n)$ since the leader has to be known to all nodes.) For deterministic algorithms, there is a well-known lower bound of $\Omega(n \log n)$ messages (even in the synchronous setting when an adversary decides which nodes to wake up and when) [11, 26, 25] where $n$ is the number of network nodes. Moreover, Afek and Gafni [11] show that any deterministic algorithm that is message optimal (i.e., takes $\Theta(n \log n)$ messages) needs $\Omega(\log n)$ time; they present such a deterministic algorithm that takes $O(n \log n)$ messages and $O(\log n)$ time in synchronous networks. Hence for synchronous networks, there exists a deterministic algorithm that is essentially singularly optimal (up to a logarithmic factor).

The situation is less clear in the asynchronous setting and for randomized solutions. First, the $\Omega(n \log n)$ message lower bound was shown for deterministic algorithms, and it was not clear whether by using randomization one can breach this bound. Second, Afek and Gafni [11] present an asynchronous leader election algorithm that takes $O(n \log n)$ messages (which is message optimal), but takes $O(n)$ time. They conjecture that this is the best possible time bound for message-optimal asynchronous algorithms. Singh [47] (mildly) disproves this conjecture by presenting an algorithm that runs in $O(n/\log n)$ time and is also message optimal ($O(n \log n)$ messages). This is still a far cry from being singularly optimal, as the time bound is essentially linear. When it comes to the use of randomness, Afek and Matias [2] develop a randomized algorithm that succeeds with probability $1 - \varepsilon$ and runs in $O(\log n)$ time using $O((n/\varepsilon) \log^2(1/\varepsilon))$ messages. If we allow for a constant probability of success, i.e., if we set $\varepsilon$ to a fixed constant, then the algorithm runs in $O(\log n)$ rounds and uses $O(n)$ messages. However, for randomized algorithms, it is typically desired that algorithms succeed with high probability. In such a scenario if we set $\varepsilon$ to $1/n^c$ for some constant $c > 0$, the message complexity significantly increases to $O(n^{1+c} \log^2 n)$.

The results of the current paper (detailed later on) break the long-standing barrier of $\Omega(n \log n)$ messages for deterministic algorithms and show that there exist singularly optimal randomized asynchronous algorithms that succeed with high probability.

**Distributed Computing Model.** We consider a system represented as an undirected complete graph $G = (V, E)$, $|V| = n$, similar to the models of [11, 18, 23, 25, 26], except that processors can access private unbiased coins. Our upper bounds do not require unique identities, and in particular, nodes can be anonymous. If nodes do have unique identifiers, then we assume, as in prior works on complete networks (see [11]), that nodes initially do not know the unique identifiers of other nodes.²

---

² In the paper, it is claimed that they develop an algorithm with termination detection that runs in $O(\log n)$ time for complete graphs. However, it is unclear if indeed the running time is $O(\log n)$ for an algorithm with termination detection or rather $O(n)$ time when termination detection is required.

³ Otherwise (in the KT model), where each node knows the unique identifiers of other nodes), leader election is trivial in complete networks.
In the synchronous communication setting, the computation is divided into discrete lockstep time units called *rounds*, and every message sent over an edge arrives at the receiver after a fixed delay of one time unit, namely, at the end of the current round. In contrast, in the asynchronous communication setting, messages sent on edges incur unpredictable but finite delay, in an error-free and FIFO manner (i.e., messages will arrive in sequence). Nevertheless, for the sake of time analysis it is assumed that a message takes *at most one time unit* to be delivered. For both modes, we make the usual (and here, very reasonable) assumption that local computation within a node is instantaneous and free. We assume the standard **CONGEST** model [39], where a node can send at most one $O(\log n)$ bit message on each edge in each round.

Following the standard assumption in asynchronous protocols (see [1, 12, 47]), nodes are initially asleep. A node enters the execution when it is woken up by the environment (at most once), or upon receiving messages from awakened neighbors. In the asynchronous setting, once a node enters execution, it performs all the computations required of it by the algorithm, sends out messages to neighbors as specified by the algorithm. In the synchronous setting, we assume the above adversarial wake up assumption as well.[4]

The message complexity of an algorithm is the worst-case total number of $O(\log n)$ bit messages sent during its execution. The time complexity is the worst-case total number of time units since the first node is woken up to the last message transmission due to the algorithm. Note that a time unit in the synchronous model corresponds to one round, whereas in the asynchronous model it is an upper bound on the transmission time of a message over an edge. Hence, it is generally more difficult to design efficient algorithms in asynchronous systems than in synchronous systems.

Following the standard approach to modeling distributed networks (see [1]), we represent the environmental uncertainties by means of an adversary controlling some of the execution parameters. Specifically, we assume an adversarial wake up mode, where node wake-up is scheduled by an adversary (who may decide to keep some nodes dormant). A node can also be woken up by receiving messages from other nodes. In addition to the wake-up schedule, the adversary also decides for how long to delay each message. These decisions are done *adaptively*, i.e., when the adversary makes a decision to wake up a node or delay a message, it has access to the results of all previous coin flips. Finally, recalling that initially the nodes are unaware of which neighbor is connected to each of their outgoing edges, the adversary also controls the graph structure (i.e., the mapping of outgoing edges to neighbors). Here, we consider the adversary to be oblivious, i.e., if nodes use randomness to choose an outgoing edge, the adversary must choose endpoints for all such outgoing edges prior to the first use of randomness in this way.[5]

**Background and additional Prior Work.** The complexity of the leader election problem and (especially deterministic) algorithms for it have been very well-studied in distributed networks. Various algorithms and lower bounds are known in different models with synchronous/asynchronous communication and in networks of varying network topologies such as a cycle, a complete graph, or some arbitrary topology (e.g., see [20, 23, 27, 28, 33, 38, 45, 49] and the references therein). The problem was first studied in context of a ring network by

---

4 This should be contrasted with the *simultaneous wake up* model, where all nodes are assumed to be awake at the beginning of computation; this is typically assumed in design of synchronous protocols (see e.g., [1, 25, 27]).

5 To appreciate the implications of allowing the adversary to construct the graph *after* seeing the random choices, see the lower bound proof techniques for message complexity in [26, 1]. These techniques are for deterministic algorithms, but will work when the adversary has adaptive edge mapping abilities.
and discussed for general graphs in the influential paper of Gallager, Humblet, and Spira [30]. Kutten et al. [27] presented a singularly optimal (up to a logarithmic factor) randomized leader election algorithm for general synchronous networks that ran in $O(D)$ time and used $O(m \log n)$ messages (where $D$, $m$, and $n$ are the network diameter, number of edges, and the number of nodes respectively). We note that $\Omega(D)$ and $\Omega(m)$ are lower bounds for time and messages for leader election even for randomized algorithms [27]. It is not known whether similar bounds can be achieved for general asynchronous networks, although one can obtain algorithms that are separately time optimal [38] and message optimal [12].

Leader election in the class of complete networks — which is the focus of this paper — has come to occupy a special position of its own and has been extensively studied [1, 2, 18, 23, 25, 26, 28, 47]; see also [31, 32, 46] for leader election in complete networks where nodes have a sense of direction. While $\Omega(n)$ is an obvious lower bound on the message complexity of leader election when the leader’s identity should be known for all nodes, an $\tilde{O}(\sqrt{n})$ (i.e., sublinear) message complexity can be obtained for a related but different problem in synchronous complete networks with simultaneous wake up [28] where we do not require that the nodes not elected know who is the leader (nor which of their ports lead to it). The above result crucially uses randomization to break the linear message complexity bound that applies for deterministic algorithms.

Afek and Gafni (as a part of presenting a tradeoff between time and message complexity) showed that the time complexity of a message optimal synchronous algorithm was $\Theta(\log n)$, while for a message optimal asynchronous algorithm, they only demonstrated an $O(n)$ time upper bound (improving previous time bounds for message optimal algorithms). They conjectured that in the asynchronous case, time complexity of any message optimal algorithm is $\Omega(n)$. Singh [47] (as a part of presenting a different tradeoff), presented a somewhat better ($O(n/\log n)$) time for message optimal asynchronous algorithms, but still posed as an important open problem the question whether $\Omega(n/\log n)$ time was optimal for such algorithms. The current paper demonstrates that this is not the case, at least for randomized algorithms that succeed w.h.p., by giving a $O(\log^2 n)$ time bound for message optimal $O(n)$ algorithm.

The earlier mentioned papers [31, 32, 46] on “sense of direction” demonstrated that when the nodes possessed additional knowledge on the topology, it was possible to reduce the number of messages to $O(n)$. In the same vein, note that for deterministic algorithms in synchronous networks and under the strong assumption of simultaneous wake up, there exists an $O(n)$ messages algorithm [27]. In contrast, an $\Omega(n \log n)$ message lower bound in synchronous networks was shown by Afek and Gafni [11] under the more common assumption of adversarial wake up. The message complexity in the current paper is $O(n)$ without using sense of direction or simultaneous wakeup assumptions (but using randomized algorithms).

For anonymous networks under some reasonable assumptions, deterministic leader election was shown to be impossible, using symmetry arguments [3]. Randomization comes to the

---

6 This variant of leader election is sometimes called “implicit”, as opposed to the version studied here, where all nodes need to know the identity of the leader.
Our Main Results. The main focus of this paper is on studying how randomization can help in designing singularly optimal algorithms for leader election in asynchronous as well as synchronous networks. Our results are summarized in Table 1. Our main result is a randomized asynchronous leader election algorithm for complete networks that runs in $O(\log^2 n)$ time and uses only $O(n)$ messages to elect a unique leader with high probability (Section 2). This is a significant improvement over the $\Omega(n \log n)$ messages needed for any deterministic algorithm, and an even larger improvement over the time complexity of previous message optimal ($O(n \log n)$) deterministic algorithms for asynchronous networks, which required $O(n)$ or $O(n/\log n)$ time [11, 47]. In addition, we show that for the synchronous setting too, we can obtain a singularly optimal algorithm that tightly matches the best possible asymptotic time and message lower bounds. We present a randomized algorithm in

\footnote{In contrast, the synchronous algorithm presented in the current paper succeeds with high probability, its time complexity is constant, and its complexity bounds hold with probability $1 - O(1/n^{(1)})$.}

Table 1 Comparison of previous upper bound results for leader election in complete networks of $n$ nodes along with our contributions. The variables $2 \leq c \leq n$, $\log n \leq k \leq n$, and $0 < \varepsilon < 1$ are parameters provided to the respective algorithms.

| Paper | Message Complexity | Time Complexity | Communication Mode | Type of Solution |
|-------|---------------------|-----------------|-------------------|-----------------|
| [17]  | $O(n \log n)$       | $O(n \log n)$  | Asynchronous      | Deterministic   |
| [1]   | $O(n \log n)$       | $O(n)$          | Asynchronous      | Deterministic   |
| [2]   | $O(nk)$             | $O(n/k)$        | Asynchronous      | Deterministic   |
| This paper | $O(n \log^{(1/\varepsilon)})$ | $O(\log n)$ | Asynchronous | Randomized, success prob. $1 - \varepsilon$ |
| [42]  | $O(n \log n)$       | $O(\log n)$    | Synchronous       | Deterministic   |
| This paper | $O(n)$               | $O(\log n)$   | Synchronous       | Randomized, success prob. $1 - O(1/\log^{(1)} n)$ |

*The algorithm always succeeds when nodes have unique identifiers (as opposed to anonymous networks).
synchronous networks that takes \(O(1)\) time and \(O(n)\) messages with high probability (see Section 3). Note that our algorithms succeed w.h.p. in anonymous networks and always succeed when each node has a unique identifier.

Our results, while providing near-optimal message and time bounds for complete asynchronous networks, are a step towards designing singularly optimal algorithms for general asynchronous networks.

2 A Randomized Algorithm for Asynchronous Networks & Analysis

In this section, we first give some high level intuition of how we use randomization to overcome the barriers of time and message complexity posed to deterministic asynchronous algorithms. Subsequently, we present a randomized algorithm that solves leader election in \(O(\log^2 n)\) time with high probability using \(O(n)\) messages with high probability.

**High level intuition behind the use of randomness.** In [1], an \(\Omega(n \log n)\) message lower bound is presented. A key idea in the proof is that the adversary is able to control the destination of messages sent by a node. It can do this for a deterministic algorithm because the adversary may predict which edges will be used in the course of the algorithm and construct the initial graph accordingly. Intuitively, if multiple clusters of nodes are active, a leader cannot be determined until they interact. If the adversary can consistently control the destination of messages over unknown links, the adversary may blow up the number of messages until the clusters interact. We bypass this issue by having nodes choose the edges they use to communicate at random, similar to the idea used in [28]. Since the adversary cannot predict which edge will be used, it cannot initially construct an undesirable graph. Our second use of randomness is to have each node, once awake, choose its \(RANK\) uniformly at random from \([1, n^4]\). This helps us achieve a significantly smaller running time compared to the deterministic case. If \(RANK\)s are not randomly chosen but deterministically assigned, then the adversary may wake up nodes such that it takes \(O(n)\) time from the time the first node is woken up until the algorithm terminates. This is because, in our algorithm, a node’s \(RANK\) plays an important role in deciding if it will go on to become the leader when interacting with another node or if it will stop trying to be the leader.

**The Randomized algorithm.** We now describe the algorithm. Each awake node may play up to two roles in the algorithm, namely (i) a candidate and (ii) a referee. During the process, candidates attempt to progress towards becoming the leader. The position of a candidate \(u\) in this process is represented by the pair \((RANK_u, PH_u)\), where \(RANK_u\) is \(u\)’s rank and \(PH_u\) is its current phase. Intuitively, we say that candidate \(v\) is ahead of candidate \(u\), and \(u\) is behind \(v\), if \(v\) has a higher phase number, or, in the case of a tie, \(v\) has a higher \(RANK\). Formally, we say that \(v \gg u\) if either (i) \(PH_v > PH_u\), or (ii) \(PH_v = PH_u\) and \(RANK_v > RANK_u\).

During the process, candidates gradually either progress or retire, eventually resulting in a single un-retired candidate who then becomes the leader. Essentially, a candidate \(u\) retires upon encountering another candidate \(v\) that is ahead of it. Such an encounter occurs when a referee learns of both candidates, and it is the referee’s task to make one of the candidates retire. (The referee is typically a third party, but may also be \(u\) or \(v\).) Hence, candidates fundamentally drive the algorithm forward, while referees serve a complementary role.

---

8 If nodes have unique identifiers, then in the case of two nodes in the same phase with the same rank, their unique identifiers can be used to break ties. Furthermore, a node will append its unique identifier to its \(RANK\) to avoid ambiguity.
role of guardians, assisting candidates in deciding whether to retire or to proceed. Each node is in one of three candidate states \textit{Candidate}, \textit{Non-elected}, \textit{Elected}, stored in the variable \textit{CAND-STATE}, corresponding to whether the node is still a candidate in the running to become a leader, is retired, or is elected, respectively.

If a node is woken up by the adversary, then it plays the role of a candidate and may eventually become the leader. If a node is woken up by a message from a neighbor, then it is immediately considered to be a retired candidate. When a node receives a message, depending on the type of message received, the node may either act in the role of a candidate or in the role of a referee, with an appropriate procedure called to handle the message. Thus, during the execution, a node either acts as just a referee or as both a candidate and a referee.

Each node \( u \), once woken up, checks if it was woken by the adversary (i.e., spontaneously wakes up itself) or by a message from a neighbor. If \( u \) was woken by a neighbor, \( u \) sets its candidate state to \textit{Non-elected}. If \( u \) was woken by the adversary, \( u \) sets its candidate state to \textit{Candidate}, chooses an integer in \([1, n^3]\) uniformly at random as its \( ID_u \) and attempts to progress. The initialization is described in Algorithm 1.

\begin{algorithm}
\begin{algorithmic}
\State If at any time \( u \) receives a message \((\text{RANK}_u, 0, \text{LEADER})\) from another node \( v \), then \( u \) sets the leader as \( v \), sets \textit{CAND-STATE} \leftarrow \textit{Non-elected}, and terminates
\State \text{REF-STATE} \leftarrow C0
\State Run Procedures \textit{Candidate}, \textit{Referee_Request_Response}, \textit{Referee_Dispute_Response}, and \textit{Decide} in parallel until termination, according to the following rules:
\State \textbf{if} \( u \) spontaneously woke up \textbf{then}
\State \hspace{1em} \textit{CAND-STATE} \leftarrow \textit{Candidate}
\State \hspace{1em} Choose integer in \([1, n^3]\) uniformly at random to be \( u \)'s rank \( \text{RANK}_u \)
\State \hspace{1em} Run Procedure \textit{Candidate}
\State \textbf{else}
\State \hspace{1em} \textit{CAND-STATE} \leftarrow \textit{Non-elected}
\State \hspace{1em} \textbf{if} \( u \) receives a message of the form \((\cdot, \cdot, \text{REQUEST})\), run Procedure \textit{Referee_Request_Response}
\State \hspace{1em} \hspace{1em} \textbf{if} \( u \) receives a message of the form \((\cdot, \cdot, \text{WINS})\circ(\cdot, \cdot, \text{LOSES})\) or \((\cdot, \cdot, \text{LOSES})\circ(\cdot, \cdot, \text{WINS})\), run Procedure \textit{Referee_Dispute_Response}
\State \hspace{1em} \hspace{1em} \textbf{if} \( u \) receives a message of the form \((\cdot, \cdot, \text{DECIDE})\), run Procedure \textit{Decide}
\end{algorithmic}
\end{algorithm}

A candidate \( u \) progresses through at most \( K = \lceil \log \sqrt{4n \log n} \rceil + 1 \) phases until it either changes its candidate state to \textit{Non-elected} or completes phase \( K \) and stays in candidate state \textit{Candidate}, in which case it declares itself leader. At the beginning of each phase \( i \) from 1 to \( K - 1 \), a candidate node \( u \) chooses a set \( S \) of \( \min\{10 \cdot 2^i, \lceil \sqrt{4n \log n} \rceil\} \) nodes uniformly at random designated as its \textit{referees} and seeks their approval to progress\footnote{We assume that node \( u \) treats itself as a referee as well in addition to the nodes of \( S \).}. In phase \( i = K \), \( u \) sets \( S \leftarrow V \), i.e. all nodes of the network form set \( S \). Node \( u \) sends a \textit{request} message \( \langle \text{RANK}_u, PH_u, \text{REQUEST} \rangle \) to each referee in \( S \).

When node \( u \) receives replies from all nodes in \( S \), it checks if any of those replies is a decline of the form \( \langle \text{RANK}_u, PH_u, \text{DECLINED} \rangle \). If so, \( u \) changes its candidate state to \textit{Non-elected}\footnote{The candidate state of \( u \) can also be changed to \textit{Non-elected} as a result of processing a \textit{DECIDE} message, described later.} in case \( u \) received no declines, it may proceed. In particular, if \( i < K \), then \( u \)
increases $PH_u$ to $i+1$ and starts the next phase, and if $i = K$ and $u$ still retains candidate state Candidate, then $u$ declares itself as leader, namely, it changes its candidate state to Elected, broadcasts the final announcement $⟨\text{RANK}_u, 0, \text{LEADER}⟩$ and terminates. The above process is described in Procedure 2.

**Algorithm 2** Procedure Candidate, run by each candidate $u$.

1: for $i ← 1$ to $K$
2: if $i < K$ then
3: Choose a set $S$ of $\min\{10 \cdot 2^i, \lceil \sqrt{\ln n} \rceil \}$ neighbors uniformly at random
4: else
5: Set $S ← V$
6: Send message $⟨\text{RANK}_u, i, \text{REQUEST}⟩$ to each node in $S$
7: Wait for replies from all referees in $S$.
8: (* Other procedures may run in parallel and do other things in the meantime.*)
9: if any reply is of the form $⟨\text{RANK}_u, PH_u, \text{DECLINED}⟩$ then
10: $\text{CAND-STATE} ← \text{Non-elected}$
11: if $\text{CAND-STATE} = \text{Non-elected}$ then
12: Exit procedure
13: if $\text{CAND-STATE} = \text{Candidate}$ then
14: $\text{CAND-STATE} ← \text{Elected}$
15: Broadcast $⟨\text{RANK}_u, 0, \text{LEADER}⟩$ and terminate

Each referee $r$ helps candidates retire (until only one is left) by comparing candidate pairs, keeping the more advanced one, and instructing the other to retire. Referee $r$ can be in one of four referee states stored in the variable $\text{REF-STATE}$.

- $\text{REF-STATE} = C0$ holds when $r$ has not been approached by any candidate yet.
- $\text{REF-STATE} = C1$ holds when $r$ has one approved candidate, referred to as its chosen candidate, and has declined every other candidate that approached it so far. A record containing the position of the chosen candidate $v$, namely, $P(v) = ⟨\text{RANK}_v, PH_v⟩$, is kept in the variable Chosen.
- $\text{REF-STATE} = C2$ holds when $r$ currently keeps track of two candidates: the chosen $v$ (in the variable Chosen) and a contender $w$ (in the variable Contender), such that $w \gg v$, and all other candidates that approached $r$ so far were declined. Moreover, a dispute is currently in progress between $v$ and $w$. This state is typically reached when $r$ has a chosen candidate $v$ that was approved by it, and later it gets a request from another candidate $w$ such that $w \gg v$. In this situation, $r$ cannot decline $w$ (since it is ahead of its current chosen), but at the same time it cannot approve $w$, since it may be that $v$ has progressed in the meantime, and it is now ahead of $w$. To resolve this uncertainty, $r$ declares a dispute, and sends a DECIDE message containing $w$’s position $⟨\text{RANK}_w, PH_w⟩$ to $v$, asking it to make a comparison between $w$ and itself (based on its current phase $PH_v$). While waiting for $v$’s response, $r$ keeps a record containing the details of $w$, namely, $P(w) = ⟨\text{RANK}_w, PH_w⟩$, in the variable Contender.
- $\text{REF-STATE} = C3$ is similar to $C2$, i.e., it holds when $r$ currently keeps track of a chosen $v$ and a contender $w$, all other candidates that approached $r$ so far were declined, and a dispute is currently in progress. The difference, however, is that the on-going dispute does not involve $w$. Rather, it is between $v$ and some previous contender $z$ such that $w \gg z \gg v$. This state is typically reached when a new candidate $w$ approaches $r$ while
If \( r \) is in referee state \( C2 \) with a dispute in progress between a chosen candidate \( v \) and a contender \( z \gg v \), and \( r \) discovers that \( w \gg z \). This allows \( r \) to decline the current contender \( z \) immediately (since even if the outcome of the dispute favors \( z \) over \( r \), the new candidate \( w \) is ahead of \( z \), so \( z \) must retire). Now \( w \) takes \( z \)'s place as the contender.

If \( r \) receives a message \( \langle RANK_u, PH_u, REQUEST \rangle \) from node \( u \) over some edge \( e \), it responds as follows.

\( r \) is in referee state \( C0 \) if it replies as follows.

\( r \) is in referee state \( C1 \) if it registers \( u \) as its chosen candidate, sends back an approval message, and switches its referee state to \( C1 \).

\( r \) is in referee state \( C2 \) if it updates the record stored in \( \text{Chosen} \). If it is another node \( v \) such that \( v \gg u \), then \( r \) sends \( u \) a decline message. Otherwise (i.e., if \( u \gg v \)), \( r \) registers \( u \) as the contender and initiates a dispute by sending a \text{DECIDE} message to \( v \), requesting it to compare its current position with that of \( u \). It also switches its referee state to \( C2 \).

\( r \) is in referee state \( C3 \) if it does the same as in referee state \( C2 \).

\( r \) is in referee state \( C3 \) if \( r \) never replies as follows.

\( r \) is in referee state \( C3 \) if \( r \) never replies as follows.

The pseudocode for the referee’s actions on receiving a request is given in Procedure 3.

When a node \( v \) which is currently the chosen candidate of the referee \( r \) (but may have possibly retired since the time it was approved by \( r \)) receives from \( r \) an \( \langle RANK_u, PH_u, \text{DECIDE} \rangle \) message, it must decide whether to end its candidacy (if it is still a candidate) or that of the other candidate. To do so, it compares its own current position with that of the contender \( u \).

If \( v \) is in candidate state \text{Non-elected} or \( v \ll u \), then \( v \) returns the message \( \langle RANK_u, PH_u, \text{LOSSES} \rangle \).

If \( v \) is in candidate state \text{Candidate} and is ahead of \( u \), it returns the message \( \langle RANK_u, PH_u, \text{WINS} \rangle \). Pseudocode for the chosen’s actions on receiving a \text{DECIDE} message from a referee is given in Procedure 4.

Once the chosen’s reply message is received by \( r \), it is processed as follows.

If \( v \) replied that the contender \( u \) has won the dispute, \( r \) sends an approval message to the current contender (which is \( u \) if the referee state is \( C2 \), and another candidate if the referee state is \( C3 \)), makes it the chosen, and switches to referee state \( C1 \).

If \( v \) replied that it has progressed beyond the contender \( u \), so \( u \) has lost the dispute and must retire, then there are two sub-cases to consider. If the referee state is \( C2 \), \( r \) sends a decline message to the contender \( u \), and switches to state \( C1 \). Now suppose the referee state is \( C3 \) and the current contender is some candidate \( w \). If \( v \)'s current position is such that \( v \gg w \), then \( r \) sends a decline message to \( w \) and switches to referee state \( C1 \).

Otherwise (\( w \gg v \)), a new dispute is required, this time between \( v \) and \( w \).

The referee’s actions on receiving a reply from the chosen about an ongoing dispute are given in Procedure 5.

**Analysis of the algorithm.** We establish the following theorem.

**Theorem 1.** Consider a complete anonymous network \( G \) of \( n \) nodes in the \text{CONGEST} model. Assume communication is asynchronous with adversarial wakeup. Then there is a
Algorithm 3 Procedure Referee\_Request\_Response, run by each referee $r$ on receiving a message $\langle \text{RANK}_u, \text{PH}_i, \text{REQUEST} \rangle$ from candidate $u$.

1: if $\text{REF-STATE} = C0$ (* $u$ is the first candidate to approach $r$ *) then
2:   Set $\text{Chosen} \leftarrow \mathcal{P}(v) = \{\text{RANK}_u, \text{PH}_u\}$
3:   Send message $\langle \text{RANK}_u, \text{PH}_i, \text{APPROVED} \rangle$ to $u$
4:   Set $\text{REF-STATE} \leftarrow C1$
5: else if $\text{Chosen} = u$ (* $u$ is the current chosen candidate, from an earlier phase *) then
6:   Set $\text{Chosen} \leftarrow \{\text{RANK}_u, \text{PH}_u\}$
7: else if $\text{REF-STATE} = C1$ (* There is a chosen candidate $v = \text{Chosen}$; all other candidates that approached $r$ so far were rejected *) then
8:   if $u \prec v$ then
9:     Send message $\langle \text{RANK}_u, \text{PH}_u, \text{declined} \rangle$ to $u$
10:   else
11:     Set $\text{Contender} \leftarrow \mathcal{P}(u) = \{\text{RANK}_u, \text{PH}_u\}$
12:     Send message $\langle \text{RANK}_u, \text{PH}_u, \text{decide} \rangle$ to the chosen $v$
13:     Set $\text{REF-STATE} \leftarrow C2$
14: else if $\text{REF-STATE} = C2$ (* A dispute is in progress between the current chosen $v = \text{Chosen}$ and the current contender $w = \text{Contender}$, $w \gg v$ *) then
15:     if $u \prec w$ then
16:         Send message $\langle \text{RANK}_u, \text{PH}_u, \text{declined} \rangle$ to $u$
17:     else
18:         Send message $\langle \text{RANK}_u, \text{PH}_w, \text{declined} \rangle$ to $w$
19:         Set $\text{Contender} \leftarrow \mathcal{P}(u) = \{\text{RANK}_u, \text{PH}_u\}$
20:         Set $\text{REF-STATE} \leftarrow C3$
21: else if $\text{REF-STATE} = C3$ (* A dispute is in progress between the current chosen $v = \text{Chosen}$ and some predecessor $z$ of the current contender $w = \text{Contender}$, $w \gg z \gg v$ *) then
22:     if $u \prec w$ then
23:         Send message $\langle \text{RANK}_u, \text{PH}_u, \text{declined} \rangle$ to $u$
24:     else
25:         Send message $\langle \text{RANK}_w, \text{PH}_w, \text{declined} \rangle$ to $w$
26:         Set $\text{Contender} \leftarrow \mathcal{P}(u) = \{\text{RANK}_u, \text{PH}_u\}$

Algorithm 4 Procedure Decide, run by a node $v$ upon receiving the message $\langle \text{RANK}_u, \text{PH}_v, \text{decide} \rangle$ from node $r$.

1: if $\text{CAND-STATE} = \text{Non-elected}$ then
2:   Send the message $\langle \text{RANK}_u, \text{PH}_v, \text{wins} \rangle \circ \langle \text{RANK}_v, \text{PH}_v, \text{loses} \rangle$ to $r$
3: else if $\text{CAND-STATE} = \text{Candidate}$ then
4:   if $u \gg v$ then
5:     $\text{CAND-STATE} \leftarrow \text{Non-elected}$
6:     Send the message $\langle \text{RANK}_u, \text{PH}_v, \text{wins} \rangle \circ \langle \text{RANK}_v, \text{PH}_v, \text{loses} \rangle$ to $r$
7: else
8:     Send the message $\langle \text{RANK}_u, \text{PH}_v, \text{loses} \rangle \circ \langle \text{RANK}_v, \text{PH}_v, \text{wins} \rangle$ to $r$
Algorithm 5 Procedure Referee_Dispute_Response, run by a referee \( r \) on receiving a reply to a message \( \langle RANK_u, PH_u, DECIDE \rangle \) from the chosen \( v \). This message only arrives while the referee is in state \( C_2 \) or \( C_3 \).

1: Let \( w \) be the Contender \( \text{(* which may be different from } u \text{ *)} \)
2: if reply is \( \langle RANK_u, PH_u, WINS \rangle \circ \langle RANK_v, PH_v, LOSES \rangle \) \( \text{(* } u \text{ won the dispute *)} \) then
3: Send message \( \langle RANK_w, PH_w, APPROVED \rangle \) to node \( w \)
4: Set Chosen \( \leftarrow \) Contender; Contender \( \leftarrow \perp \)
5: Set REF-STATE \( \leftarrow C_1 \)
6: else \( \text{(* The chosen } v \text{ won the dispute over } u \text{ *)} \)
7: if REF-STATE \( = C_2 \) OR (REF-STATE \( = C_3 \) AND \( v \gg w \)) then \( \text{(* } v \text{ remains the chosen *)} \)
8: Send message \( \langle RANK_w, PH_w, DECLINED \rangle \) to node \( w \)
9: Contender \( \leftarrow \perp \); Set REF-STATE \( \leftarrow C_1 \)
10: else \( \text{(* REF-STATE } = C_3 \text{ but } w \gg v \text{, so a new dispute is required *)} \)
11: Send message \( \langle RANK_w, PH_w, DECIDE \rangle \) to the chosen \( v \)
12: Set REF-STATE \( \leftarrow C_2 \)

randomized algorithm to solve leader election with high probability in \( O(\log^2 n) \) time with high probability using \( O(n) \) messages with high probability. If each node has a unique identifier, then the algorithm always succeeds. All nodes terminate at the end of the algorithm.

We show three properties: exactly one leader is chosen w.h.p. for anonymous networks and always when nodes have unique identifiers and all nodes subsequently terminate, the time complexity is \( O(\log^2 n) \) w.h.p., and the message complexity is \( O(n) \) w.h.p.

Before we continue with the proof, we make an important observation.

Observation 1. For each node \( u \) that participates in the algorithm, if \( u \) begins phase \( i \), then \( u \) will finish phase \( i \).

Observation 1 is used implicitly in the remaining proof whenever a node is mentioned to finish some phase and possibly perform some calculation as a result of completing that phase.

Lemma 2. By the end of the algorithm, exactly one node is in candidate state Elected w.h.p. for anonymous networks and always when nodes have unique identifiers and the remaining nodes are in candidate state Non-elected. Furthermore, all nodes eventually terminate.

Proof. Let us consider a network where each node has a unique identifier. We show that exactly one leader is always chosen by arguing that exactly one \( RANK \) is chosen as the leader. It is clear to see that if we then consider an anonymous network where each node chooses its \( RANK \) from \( [1,n^4] \) and nodes do not have access to unique identifiers to break ties, the \( RANK \) chosen as the leader will belong to not more than one node w.h.p.

We prove by induction that at least one node is a candidate at the end of phase \( i \) for \( 1 \leq i \leq K \). For the induction basis, consider, for the sake of the proof, that when a node wakes up, it is in phase 0, but it immediately moves to phase 1. Clearly, this does not change the outcome of the algorithm. Now the lemma holds for phase zero since at least one node is woken up by the adversary initially, and the induction is proven with phase zero as base case.

Assume that the claim holds true up to the end of some phase \( k < K \). Let \( u \) be the candidate with largest \( RANK \) in phase \( k + 1 \). If \( u \) is not retired by any other node in phase \( k + 1 \), then it will be in candidate state Candidate at the end of phase \( k + 1 \). Hence the claim holds for \( k + 1 \). Otherwise, if \( u \) is retired in phase \( k + 1 \), then necessarily either \( u \) came
into contact with (i) some candidate \( u' \) in a higher phase \( PH_{u'} > k + 1 \) or a referee to such a candidate, (ii) a candidate \( u'' \) with higher rank \( RANK_{u''} > RANK_u \) in the same phase \( k + 1 \) or a referee to such a candidate, or (iii) a candidate \( v \) that completed all phases and is still in candidate state Candidate or a message from such a candidate \( v \). In all cases, the induction claim still holds true for \( k + 1 \).

We subsequently show that exactly one node will change its candidate state to Elected. By the previous claim, at least one candidate completes phase \( K - 1 \). Let \( u_{\text{max}} \) be the node with the largest RANK that is still a candidate at the beginning of phase \( K \). In phase \( K \), \( u_{\text{max}} \) will approach all other nodes, receive an approval from each of them, and subsequently change its candidate state to Elected. Any other candidate \( v \) will approach \( u_{\text{max}} \), receive a declined message from it, and thus change its candidate state to Non-elected. Thus exactly one node will change its candidate state to Elected and broadcast a \( \langle \cdot, 0, \text{LEADER} \rangle \) message.

Finally, we argue that all nodes terminate. It is clear that if exactly one node changes its candidate state to Elected and terminates the algorithm. All other nodes change their candidate state to Non-elected and terminate the algorithm upon receiving this message. Thus, the proof is complete.

We now prove the time complexity bound. First, we prove the following useful lemma about the amount of time one phase for a candidate takes.

\textbf{Lemma 3.} If a node \( u \) starts some phase \( i \) as a candidate, then it exits the phase (either by increasing its phase to \( i + 1 \) or by retiring) within \( O(1) \) time.

\textbf{Proof.} For any candidate \( u \), for each phase \( i \), the largest delay until that phase ends occurs when candidate \( u \) sends a message to a referee \( r \), which in turn sends a message to its current chosen \( v \). Subsequently, \( v \) sends a reply to \( r \), which in turn sends a reply to \( u \). If the congestion on both these edges is \( O(1) \) messages in the time interval from start to end of phase \( i \), then the total time duration of the phase is \( O(1) \). Let \( e_{uv} \) and \( e_{rv} \) denote edges between \( u \) and \( r \) and between \( r \) and \( v \) resp. We show at most \( O(1) \) messages need to be transmitted on both \( e_{uv} \) and \( e_{rv} \) during phase \( i \), thus allowing each phase to take \( O(1) \) time.

Consider edge \( e_{uv} \). In phase \( i \), \( u \) can send an invite to \( r \) and receive a reply to the invite from \( r \). Furthermore, \( u \) can also be the chosen for \( r \) and receive a DECEIVE message from \( r \) and subsequently send back a reply.\(^{11}\) Additionally, \( r \) may also be a candidate with \( u \) as its referee or \( r \) may be the chosen for \( u \), resulting in at most a doubling of messages over the edge. Thus at most \( O(1) \) messages are sent across that edge for phase \( i \) of \( u \). Now consider edge \( e_{rv} \). A similar analysis as above renders \( O(1) \) an upper bound on the messages on the edge. Thus, we see that each phase takes \( O(1) \) time units.

In order to bound the running time, we also make use of the following useful property from [36], denoted in the reference as Problem C.2 in Appendix C.6. We slightly modify the statement to suit our needs.

\textbf{Lemma 4.} (Problem C.2 in [36]) Let \( a_1, a_2, \ldots, a_n \) be a sequence of \( n \) values. Each value \( a_i \) is independently and randomly chosen from a fixed distribution \( D \). Let \( m_i = \max \{ a_1, a_2, \ldots, a_i \} \), i.e., the maximum of the first \( i \) values. Let random variable \( Y \) denote

\(^{11}\) Recall that once some node \( u \), in phase \( i \), is the chosen of some node \( r \) and receives a DECEIVE message, either \( u \) is retired or \( r \) learns that \( u \) is in phase \( i \) and will not send any further messages from other candidates in phase \( < i \). If another DECEIVE message is subsequently sent to \( u \) from \( r \), then either \( u \) wins (implying it is in a higher phase), or else \( u \) is retired and no further messages are sent along the edge.
the number of times the maximum value is updated, i.e., the number of times \( m_i \neq m_{i+1} \). Then \( E[Y] = O(\log n) \).

**Lemma 5.** The running time of the algorithm is \( O(\log^2 n) \) with high probability.

**Proof.** For any given execution of the algorithm, eventually one candidate, \( C_t \), woken up by the adversary, becomes the leader. We show two properties for any such \( C_t \). First, once \( C_t \) is woken up by the adversary, the algorithm takes an additional at most \( O(K) \) time until \( C_t \) broadcasts that it is the leader and the algorithm terminates for every node. Second, we bound the time from when the first node is woken up by the adversary to when candidate \( C_t \) is woken up as \( O(K \cdot \log n) \) with high probability.

Once \( C_t \) is woken up, it must complete \( K \) phases before it broadcasts that it is the leader. By Lemma 3, each of its phases takes \( O(1) \) time to complete. Thus, once \( C_t \) is woken up by the adversary, it takes at most \( O(K) \) time until all nodes terminate.

Now, we bound the time until \( C_t \) is woken up. For this proof, we say that two candidates \( u \) and \( v \) *encounter* each other at time \( t \) when a referee (which could be either \( u \) or \( v \) itself) is made aware of both \( u \) and \( v \) for the first time. If, as a result of the encounter, \( v \) is retired, then we call \( u \) the *winner* of the encounter and \( v \) the *loser*. Note the following observation.

**Observation 2.** Consider some execution of the algorithm where a candidate \( u \) while in phase \( k \) retires candidate \( w \). Subsequently, \( u \) reaches phase \( i \) and is retired by another candidate \( v \) in phase \( j \). When \( u \) and \( v \) encounter each other, then either \( j > i \) or \( v \) has a larger \( RANK \) than \( u \) and is also in phase \( i \). Furthermore, from the time \( u \) retired \( w \), at most \( O(i - k) \) time units have passed until the encounter between \( u \) and \( v \) (by Lemma 3) and an additional \( O(1) \) time units pass until \( u \) retires.

Observation 2 has the following implications. Either the winner of an encounter has a larger \( RANK \) than the loser or else the phase number of the winner is larger than that of the loser when they encounter one another. Furthermore, the amount of time that passes between two encounters involving a common candidate can be bounded by the difference in phases of that candidate plus \( O(1) \) time units.

We now construct a new graph where the nodes are the subset of the nodes that are woken up by the adversary (and which become candidates). There is an edge from candidate \( C_i \) to \( C_j \) if candidate \( C_i \) encounters and subsequently is retired by \( C_j \). Notice that this graph is a directed acyclic graph since once a candidate \( C_i \) is retired by some other node \( C_j \), \( C_i \) cannot go on to retire \( C_j \) or any candidate that retires \( C_j \). Notice also that a candidate may encounter multiple other candidates, resulting in nodes with an in degree or out degree (or both) that is \( > 1 \). Finally, note that candidates that do not retire any other candidates are sources and \( C_t \) is a sink. Consider a path in the graph from a source to \( C_t \) such that the node first woken up by the adversary is one of the nodes in this path. If there are multiple such paths, choose one arbitrarily. Label the nodes from the source to the node just before \( C_t \) as \( C_1, C_2, \ldots, C_w \). Let \( C_1 \) be in some phase \( p \) at the time of its encounter with \( C_2 \). Let \( R \) be the time that pass from the first encounter until the final encounter in the chain. Then, the total time from when the first node was woken up until \( C_t \) was woken up is upper bounded by \( O(p + R) = O(K + R) \). We now bound the value of \( R \).

With each of the candidates \( C_i \), \( 1 \leq i \leq w \), associate a bit \( b_i \) which indicates how \( C_i \) was retired. Specifically, \( b_i \) is set to 0 if \( C_{i+1} \) was in a higher phase than \( C_i \) at the time of the encounter. \(^{12} \)

Bit \( b_i \) is set to 1 if \( C_{i+1} \) has a higher \( RANK \) than \( C_i \) and is in the same phase.

---

\(^{12} \) Note that we are referring to the actual phases \( C_i \) and \( C_{i+1} \) are in at this time, not necessarily the
as it at the time of the encounter. By Observation 2 and its implications, we see that every bit \( b_i, 1 \leq i \leq w \) is set to 0 or 1.

Consider the bit string \( B = b_1 b_2 \ldots b_n \). We will show that \( R \) is upper bounded \( O(K + |B|) \). In order to bound the size of \( B \), we first bound the number of 0’s that can be present in \( B \), and then we bound the number of 1’s that can be present between any two 0’s in \( B \).

Denote by \( P_i \) the phase of \( C_i \) when it was retired. Observe that, regardless of whether \( b_i = 0 \) or 1, the phase \( P_{i+1} \) of the subsequent candidate \( C_{i+1} \) can never be less than that of the candidate \( C_i \) it just retired. Furthermore, when \( b_i = 0 \), \( P_{i+1} > P_i \). Thus, there can be at most \( K \) bits set to 0 in \( B \). We now bound the number of bits set to 1 between any two bits that are set to 0.

Between any two bits set to 0, at most \( n - 1 \) nodes can be woken up by the adversary to trigger an encounter leading to a bit being set to 1. An encounter where a bit is set to 1 involves the \( RANK \) of the awakened node being higher than that of the currently considered candidate, resulting in the higher \( RANK \) node becoming the currently considered candidate. By Lemma 4, we see that on expectation, \( O(\log n) \) such encounters can thus be triggered. Applying a simple Chernoff bound, we see that \( O(\log n) \) is in fact a high probability upper bound on the number of such encounters, and by extension the number of bits set to 1 between any two bits set to 0. Since there are at most \( K \) bits set to 0, \( |B| = O(K \cdot \log n) \).

Each encounter of two candidates contributes \( O(1) \) time towards the running time. Furthermore, we must account for the time between encounters as well. Between any two encounters, the phase of a candidate may increase. We have already seen that if candidate \( C_i \) is in phase \( P_i \), then for subsequent candidates \( C_j \) where \( j > i \), \( P_j \geq P_i \). From Lemma 3, we see that for any candidate a phase takes \( O(1) \) time. Therefore, the total number of time units between all encounters is at most \( O(K) \) time units. Thus, \( R \) can be upper bounded by \( O(K + |B|) = O(K \cdot \log n) \) and the total running time is \( O(K + R) = O(\log^2 n) \) time with high probability.

We now prove that the message complexity of the randomized algorithm is \( O(n) \) with high probability. We first show that the number of candidates that can participate in every phase is reduced by a factor of four from one phase to the next up to phase \( \rho = K - [\log \log n] - 5 \) with high probability.

\[ \textbf{Lemma 6.} \text{ The total number of candidates that participate in phase } i \text{ of the algorithm is at most } \left\lceil \frac{n}{4^{i-1}} \right\rceil \text{ with high probability for } 1 \leq i \leq \rho. \]

\textbf{Proof.} We prove the claim by induction. Initially, even if the adversary wakes up all nodes, at most \( n \) candidates participate in phase 1. Assume that the claim holds true until some phase \( k < \rho \). We prove that the claim holds in phase \( k + 1 \) when the number of candidates participating in phase \( k \) is upper bounded by \( \left\lceil \frac{n}{4^k} \right\rceil \).

Notice that if \( \leq \left\lfloor \frac{n}{4^k} \right\rfloor \) candidates participate in phase \( k \), then the claim holds immediately for phase \( k + 1 \). Now, let us assume that the number of candidates participating in phase \( k \) lies in \( (\left\lfloor \frac{n}{4^k} \right\rfloor, \left\lceil \frac{n}{4^k} \right\rceil] \).

Organize the candidates in increasing order of their \( RANKs \). Group the top one-sixteenth of these candidates into the set \( top \) and group the remaining candidates into the set \( bottom \). We ignore the retirement of candidates from \( top \) in phase \( k \) by assuming that none of them are retired in this phase and show that a sufficient number of candidates from \( bottom \) are
retired for the claim to hold in phase $k+1$.

In this phase, there are at least $1/16 \cdot \lceil n/4^k \rceil$ candidates in top, each of which makes $10 \cdot 2^k$ requests. Thus, altogether, candidates from top make at least $m_0 = (\lceil n/4^k \rceil/16) \cdot 10 \cdot 2^k \geq 5n/2^{k+3}$ requests uniformly at random.

Consider a candidate $u$ from bottom. Let $r$ be some specific referee approached by $u$. Call $r$ top-free if none of the candidates from top has approached $r$ and $r$ itself is not in top. Then the probability that $r$ is top-free is at most $p_0 = (1 - 1/n)^{m_0} \cdot 15/16 \leq (1 - 1/n)^{m_0}$. Note that $u$ will be retired if even one of its requests is to a non-top-free referee. Therefore, the probability that $u$ is not retired in this phase is at most $p_0^{10 \cdot 2^k} \leq (1 - 1/n)^{(5n/2^{k+3}) \cdot 10 \cdot 2^k} \leq e^{-25/4}$. Thus, denoting the number of candidates from bottom that are not retired by $NR_B$, we have

$$
\mathbb{E}[NR_B] \leq \frac{(15/16) \cdot \lceil n/4^{k-1} \rceil \cdot e^{-25/4}}{(15/2)e^{-25/4} \cdot n/4^k} \leq \frac{(1/16) \cdot (n/4^k)}.
$$

As $NR_B$ is the sum of independent Bernoulli trials, we can apply a Chernoff bound (second bound of Theorem 4.4 in [35]). When $k < \rho$, the probability that $NR_B$ exceeds twice the expected value is upper bounded by $1/n^3$. Thus the total number of candidates that can participate in phase $k+1$ with high probability is at most $(1/16) \cdot \lceil n/4^{k-1} \rceil + (2/16) \cdot (n/4^k) \leq \lceil n/4^k \rceil$. Note that for each phase, the upper bound on the number of candidates who are not retired holds with probability $1 - 1/n^3$, assuming that the bound held in the previous phase. Applying a union bound over all phases of the induction, we see that this upper bound holds for phase $\rho$ with probability $1 - O(1/n^2)$, i.e. w.h.p., and for each previous phase with a larger probability. Thus the claim is true for all $1 \leq i \leq \rho$ with high probability.

We next show that in the final phase of the algorithm, only a single candidate will not be retired with high probability.

**Lemma 7.** At the end of phase $K - 1$, exactly one node will be in candidate state Candidate with high probability.

**Proof.** Consider the candidate with the largest RANK, $u_{\text{max}}$, in phase $K - 1$. We show that for each candidate $v \neq u_{\text{max}}$ in this phase, the intersection of referees approached by $v$ and $u_{\text{max}}$ is non-zero with high probability. Thus $u_{\text{max}}$ will retire each such candidate $v$ with high probability. Notice that $10 \cdot 2^k \geq \lceil \sqrt{4n \log n} \rceil$ and so each candidate in this phase makes exactly $\lceil \sqrt{4n \log n} \rceil$ requests. Then

$$
\mathbb{P}[v \text{ is not retired by } u_{\text{max}}] = \frac{(1 - 1/n)\lceil \sqrt{4n \log n} \rceil \cdot \lceil \sqrt{4n \log n} \rceil}{1/n^3} \leq 1/n^4.
$$

---

13 Notice that the claim is an upper bound on the number of candidates in each phase. By showing that the claim holds when none of the candidates from top is retired, it is easy to see that the claim also holds when at least one candidate from top is retired.

14 Notice that for all phases $i$ that the claim applies to, $10 \cdot 2^i \leq \lceil \sqrt{4n \log n} \rceil$. Thus, any candidate in one of these phases will make $10 \cdot 2^i$ requests.

15 Note that the requests by a candidate in a single phase are made without repetition, i.e. a candidate does not send more than one request along the same edge in a single phase. During the subsequent analysis, we consider them to be made with repetition so as to simplify the analysis. This assumption only decreases the probability of a candidate from bottom inviting a referee that was approached by one of the top candidates and thus this simplifying assumption is acceptable.
By Lemma 6, there are at most \(2^{13}\log n\) candidates participating in phase \(\rho\). As the number of candidates participating in subsequent phases cannot increase beyond this value, it acts as an upper bound for the number of candidates participating in phase \(K - 1\).

We see that the probability that any of these candidates is not retired is at most \(1/n^3\) through the use of a union bound. Thus, w.h.p. \(u_{\max}\) retires every other candidate in this phase. Thus only \(u_{\max}\) ends the phase in candidate state Candidate w.h.p.

**Lemma 8.** The message complexity of the algorithm is \(O(n)\) messages with high probability.

**Proof.** Let \(\psi = [\log(1/10 \cdot \sqrt{4n\log n})]\). We start by analysing the message complexity that can be “charged” to candidates in phases 1 to \(\psi\). In each phase \(i \leq \psi\), each candidate \(u\) generates and receives up to \(2 \cdot 10 \cdot 2^i\) messages from its referees. Furthermore, for each of its referees in a given phase \(i\) and from all previous phases, a candidate may receive a message to decide candidacy, resulting in at most \(\sum_{j=1}^{i-1} 2 \cdot 10 \cdot 2^k \leq \sum_{j=1}^{i-1} 2 \cdot 10 \cdot 2^{j+1} = 5 \cdot 2^{i+4}\) additional messages received and generated by the candidate in this phase. (For a candidate \(u\) and a referee \(r\) chosen by \(u\) in phase \(i\) or a previous one, only one decide message will be sent from \(r\) to \(u\).)

In each phase \(1 \leq i \leq \rho\), there are at most \([n/4^{i-1}]\) candidates w.h.p. by Lemma 6. As calculated earlier, we have \(5 \cdot 2^{i+2} + 5 \cdot 2^{i+4} = 25 \cdot 2^{i+2}\) messages per candidate in each phase \(i\), resulting in a total of \(O(n)\) messages w.h.p. generated across all candidates in all phases.

In phases \(\rho + 1 \leq i \leq \psi\), there are at most \(2^{13}\log n\) candidates in each phase with high probability by Lemma 6 each making \(10 \cdot 2^i\) requests in that phase. Thus, the total number of messages over these phases is \(\sum_{i=\rho+1}^{\psi} [2^{13}\log n]25 \cdot 2^{i+2} = O(\sqrt{n}\log^{3/2} n)\) w.h.p.

Now we analyze the message complexity attributed to candidates in phases \(\psi\) and above. In each phase \(i > \psi\), each candidate \(u\) generates and receives \(2 \cdot [\sqrt{4n\log n}]\) messages from its referees. Furthermore, for each of its referees in a given phase \(i\) and from all previous phases, \(u\) may receive a message to decide candidacy, resulting in at most \(2[\sqrt{4n\log n}] (i - \psi)\) additional messages due to decide messages from referees of phases \(j > \psi\), plus \(5 \cdot 2^{\psi+4}\) messages due to decide messages from referees of phases \(j \leq \psi\). Thus, at most \(2 \cdot (i + 5 - \psi) \cdot [\sqrt{4n\log n}]\) messages are received and generated by \(u\) in this phase.

In phases \(i > \psi\), there are at most \(2^{13}\log n\) candidates in each phase w.h.p. by Lemma 6. Thus, the total number of messages across all these phases is at most \(\sum_{i=\psi+1}^{K} [2^{13}\log n] \cdot 2 \cdot (i + 5 - \psi) \cdot [\sqrt{4n\log n}] = O(\sqrt{n}\log^{5/2} n)\) w.h.p.

By Lemma 7 exactly one candidate will complete all the first \(K - 1\) phases and remain in candidate state Candidate w.h.p. This candidate will generate \(O(n)\) messages in phase \(K\).

Thus, over all phases of the algorithm, there are totally \(O(n)\) messages w.h.p.

**3 A (Tightly) Singly Optimal Synchronous Algorithm**

In this section, we present a message and time optimal algorithm for the synchronous setting, i.e., an algorithm that takes \(O(1)\) time and \(O(n)\) messages with high probability. Note that these upper bounds are tight (hence the term “tightly” in the section title). Recall that we are still dealing with adversarial wake up, so a node may still not be woken up by the adversary, or may wake up late.

Like in the asynchronous algorithm of Section 2 when a node wakes up, it chooses a RANK from \([1, n^4]\) uniformly at random. Also, nodes participate in the algorithm as candidates or referees. However, here there are two types of candidates, silent and active. Specifically, a node \(v\) that is woken up spontaneously (by the adversary) becomes a silent candidate. A silent candidate turns into an active candidate following a successful coin
toss, as described later. A node $v$ can also be woken up by another node $w$ who is an active candidate and who approaches $v$ with an approval request. In that case, $v$ becomes a referee and will not propose its candidacy. Finally, a node $v$ can also be woken up by a \textsc{Winner}(\textsc{Rank}) message announcing the rank of a winner, in which case it does nothing.

If a silent candidate $v$ is approached at any time by an active candidate $w$ appointing it as its referee and requesting its approval (or by a \textsc{Winner}(\textsc{Rank}) message), then $v$ immediately retires (and acts as a referee if requested).

A node $v$ that is woken up at time $t$ and becomes a silent candidate, makes three attempts to transform into an active candidates. Each attempt corresponds to a biased coin toss. If the coin comes up heads, the candidate transforms from a silent candidate into an active one. At time $t$, $v$ selects itself randomly as an active candidate with probability $n^{-2/3}$. If this attempt fails, then $v$ remains silent for $3$ time units and (assuming it was not retired in the meantime by some active candidate) tries again at time $t + 3$, this time with probability $n^{-1/3}$. If this attempt fails as well, then $v$ remains silent for $3$ additional time units and (assuming it was not retired) tries again at time $t + 6$, this time with probability 1.

Once a silent candidate $v$ becomes an active candidate, it actively attempts to become the leader. To do so, $v$ chooses u.a.r. (uniformly at random) $[2\sqrt{n}\log n]$ neighbors as referees, and sends them an approval request containing its rank, \textsc{Rank}_v. For simplicity, $v$ always appoints itself as its own referee in addition to the previously chosen ones.

Note that there may be other active candidates trying to get elected at the same time. Among them, the one with the largest \textsc{Rank} will win the election. Each referee $w$, approached by a group $C_w$ of active candidates ($|C_w| \geq 1$) at some time $t$, selects the largest \textsc{Rank}$_v$ among all $v \in C_w$, and sends a message back to each $v \in C_w$ with this \textsc{Rank}. (Note that $w$ itself may be one of the competing active candidates.)

If an active candidate receives its own \textsc{Rank} back from all its referees, then $v$ declares itself a \textsc{winner} by broadcasting \textsc{Winner}(\textsc{Rank}_v). (Note that $v$ is not guaranteed to be the elected leader, since there may be more than one winner, but w.h.p. there will be a single winner.) Otherwise (i.e., if $v$ hears $\textsc{Rank}_w \neq \textsc{Rank}_v$ from some referee $w$, where $\textsc{Rank}_w \neq \textsc{Rank}_v$), $v$ retires.

At any time, a node $v'$ receiving one or more \textsc{Winner}(\textsc{Rank}_v) messages finds the largest \textsc{Rank} among these winners and sets \textit{leader} $\leftarrow$ \textsc{Rank}. This holds even if $v'$ itself is currently an active candidate and $\textsc{Rank}_w > \textsc{Rank}$. (This can happen if $v'$ became active at a later time than the earliest active candidates, and still did not send its own \textsc{Winner} message.) We establish the following theorem.

\textbf{Theorem 9.} Consider a complete anonymous network $G$ of $n$ nodes in the \textsc{CONGEST} model, with synchronous communication and adversarial wakeup. There is a randomized algorithm to solve leader election w.h.p. using $O(n)$ messages w.h.p. in 9 rounds (deterministically). (If nodes have unique identifiers, then the algorithm always elects a leader.)

\textbf{Proof.} We first argue about the time and correctness. We then prove that the bound on message complexity holds w.h.p.

\textbf{Lemma 10.} The algorithm elects exactly one leader w.h.p. for an anonymous network and always elects exactly one leader when nodes have unique identifiers. Furthermore, the algorithm terminates within at most 9 time units from the time the first node is woken up.

\footnote{If nodes have unique identifiers, then in case two active candidates have the same \textsc{Rank}, their unique identifiers will be used to break ties. Furthermore, their unique identifiers will be appended to any messages involving their \textsc{Rank} to avoid ambiguity.}
Proof. Suppose the adversary wakes up the first node at time $t_0$. Note that the adversary may wake up more than one node at time $t_0$, as well as in any later time $t > t_0$. However, no messages are sent until the first time $\hat{t}$ in which some silent candidate succeeds in becoming active.

Observe that $\hat{t} \leq t_0 + 6$, since (i) from time $t_0$ and after there is at least one silent candidate, (ii) a silent candidate that is not retired will necessarily become active at most 6 time units after waking up, by which time its success probability is increased to 1, and (iii) a silent candidate can only be made to retire (directly or indirectly) on some time $t$ by an active candidate who necessarily became active earlier than $t$.

Observe that once at least one active candidate exists, a leader will be elected for sure within at most 3 additional time units. In particular, denoting by $Z$ the set of silent candidates that became active at time $\hat{t}$, and letting $z$ be the highest $\text{RANK}$ node among them, $z$ will surely announce itself a leader.

Note that messages will be sent out only at times $\hat{t}, \hat{t} + 1$ and $\hat{t} + 2$. Specifically, at time $\hat{t}$, all nodes in $Z$ will send their requests to their referees, the referee replies will be sent in the following time, and at time $\hat{t} + 2$ each elected leader will send out its announcement.

Note that with high probability, $z$ will be the only winner. To see this, observe that each active candidate chooses some $\left[2\sqrt{n}\log n\right]$ referees u.a.r. For each other active candidate $v_i \in Z$, define a bad event $E_i$ occurring when there is no overlap between $v_i$’s and $z$’s referees. If $E_i$ occurs, $v_i$ could possibly become a winner. However,

$$\Pr[E_i] = (1 - 2\sqrt{n}\log n/n)^{2\sqrt{n}\log n} = (1 - 2\log n/\sqrt{n})^{2\sqrt{n}\log n} \leq e^{-2\log n} = 1/n^2.$$ 

As there are at most $n - 1$ active candidates in $Z$, by the union bound, the probability that at least one bad event occurs is at most $(n - 1)/n^2 \leq 1/n$. Thus, w.h.p. no bad event occurs and only $z$ becomes a winner.

In the unlikely case when multiple candidates become winners and broadcast their rank at time $\hat{t} + 2$, if nodes have unique identifiers, then exactly one of the winners will become the elected leader of everyone in the network, since node $\text{RANKs}$ would then be distinct.

Note that some silent candidate $q$ may become active at time $\hat{t} + 1$ or $\hat{t} + 2$ and approach its referees. Such a candidate will not win over $z$ even if $\text{RANK}_q > \text{RANK}_z$, since it will get the message $\text{Winner}(\text{RANK}_z)$ one or two time units before sending out its own announcement.

Lemma 11. The message complexity of the algorithm is $O(n)$ with high probability.

Proof. As mentioned earlier, throughout the execution, no messages are sent except at times $\hat{t}, \hat{t} + 1$ and $\hat{t} + 2$.

Each winner generates $O(n)$ $\text{Winner}$ messages in its final broadcast. As argued earlier, with high probability, at most one node will send $\text{Winner}$ messages, so the contribution of such messages to the total message complexity is $O(n)$.

Each of the candidates in $Z$ is responsible for $O(\sqrt{n}\log n)$ “election messages” between itself and its referees. In addition, such messages may be sent by candidates who become active at time $\hat{t} + 1$ or $\hat{t} + 2$ (although these candidates will not become winners). Denote the sets of these late-coming active candidates by $Z'$ and $Z''$ respectively. Hence to bound the number of messages, it suffices to bound $|Z| + |Z'| + |Z''|$. Let us first bound $|Z|$. Partition the set $Z$ into three subsets, $Z = \bigcup_{\ell=1}^{3} Z_{\ell}$, where $Z_1$ consists of nodes that woke up at time $\hat{t}$ and succeeded in becoming active candidates on their first attempt, $Z_2$ consists of nodes that woke up at time $\hat{t} - 3$ and succeeded in becoming active candidates at time $\hat{t}$ on their second attempt, and $Z_3$ consists of nodes that woke up
at time \( t - 6 \) and succeeded on their third attempt. For \( \ell = 1, 2, 3 \), denote by \( W_\ell \) the set of silent candidates awoken by the adversary at time \( t - 3(\ell - 1) \), and let \( n_\ell = |W_\ell| \). Note that 
\[
\mathbb{E} [|Z_\ell|] = n_\ell \cdot n^{\ell/3-1}.
\]

Consider first the active candidates of \( Z_1 \) (i.e., that woke up at time \( t \)). Since \( n_1 \leq n \), their expected number is 
\[
\mathbb{E} [|Z_1|] = n_1 \cdot n^{-2/3} \leq n^{1/3}.
\]

Therefore, with high probability, the actual number of active candidates in \( Z_1 \) satisfies 
\[
|Z_1| \leq 6n^{1/3}.
\]

(To prove this, consider the third inequality from Theorem 4.4 in [35], i.e., 
\[
\text{Prob}(X \geq R) < 2^{-R} \quad \text{for } R \geq 6\mathbb{E}[X],
\]

where \( X \) is the sum of independent Poisson trials. Setting \( R = 6n^{1/3} \), we get the desired high probability bound.)

Next, consider candidates from \( Z_2 \). Note that with high probability, 
\[
n_2 \leq n^{7/9}.
\]

To see why this holds, suppose towards contradiction that \( n_2 > n^{7/9} \). At time \( t - 3 \), each of the \( n_2 \) silent candidates in \( W_2 \) attempted to become active with probability \( n^{-2/3} \). Therefore, the expected number of silent candidates from \( W_2 \) that should have become active at time \( t - 3 \) was 
\[
n_2 \cdot n^{-2/3} > n^{7/9} \cdot n^{-2/3} = n^{1/9}.
\]

This implies that with high probability, there should have been at least one actual active candidate from \( W_2 \) at time \( t - 3 \), contradicting the definition of \( t \).

By Eq. (1), the expected number of active candidates in \( Z_2 \) is thus 
\[
\mathbb{E} [|Z_2|] = n_2 \cdot n^{-1/3} \leq n^{7/9} \cdot n^{-1/3} \leq n^{4/9}.
\]

Therefore, with high probability, \( |Z_2| \leq 6n^{4/9} \).

Finally, consider the active candidates in \( Z_3 \). By arguments similar to those used to show Eq. (1), we have that with high probability, \( n_3 \leq n^{4/9} \). It follows that the expected number of active candidates in \( Z_3 \) is 
\[
\mathbb{E} [|Z_3|] = n_3 \cdot 1 \leq n^{4/9}.
\]

Therefore, with high probability, \( |Z_3| \leq 6n^{4/9} \).

In summary, 
\[
|Z| = |Z_1| + |Z_2| + |Z_3| \leq 6n^{1/3} + 6n^{4/9} + 6n^{4/9} \leq 18n^{4/9}
\]

with high probability.

We now turn to bounding \( |Z'| \). This is done in a very similar manner, by partitioning the set \( Z' \) into three subsets, \( Z' = \bigcup_{\ell=1}^3 Z'_\ell \), where \( Z'_1 \) consists of nodes that woke up at time \( t + 1 \) and succeeded in becoming active candidates on their first attempt, \( Z'_2 \) consists of nodes that woke up at time \( t - 2 \) and succeeded in becoming active candidates at time \( t + 1 \) on their second attempt, and \( Z'_3 \) consists of nodes that woke up at time \( t - 5 \) and succeeded on their third attempt. A similar analysis reveals that \( Z'_1 \) is no larger than \( 6n^{1/3} \) with high probability, and each of the other two sets, \( Z'_2 \) and \( Z'_3 \), must be no larger than \( 6n^{4/9} \) with high probability, since otherwise an active candidate would have emerged w.h.p at a time earlier than \( t + 1 \). It follows that 
\[
|Z'| \leq 18n^{4/9}
\]

with high probability. The same analysis yields that also \( |Z''| \leq 18n^{4/9} \) with high probability.

It follows that the total number of election messages due to the active candidates in these sets is at most 
\[
(|Z| + |Z'| + |Z''|) \cdot O(\sqrt{n} \log n) \lesssim O(n^{4/9}) \cdot O(\sqrt{n} \log n) \lesssim O(n),
\]

for sufficiently large \( n \).

This completes the proof of Theorem 9.\( \blacksquare \)
4 Conclusion

In the asynchronous setting, no singularly optimal deterministic leader election algorithm is known to exist. In contrast, we have shown that using randomization, a singularly optimal algorithm (whose message complexity is asymptotically optimal and whose time complexity is optimal up to a polylogarithmic factor) can be obtained.

One open question is whether we can improve the time complexity of our message-optimal ($O(n)$) randomized asynchronous algorithm from the current $O(\log^2 n)$ time, to say, $O(\log n)$ time. Also, can we obtain such a (almost) singularly optimal deterministic algorithm?

Another important question is to find out whether (and when) one can construct time and message-efficient asynchronous algorithms also for general graphs, in particular algorithms that are (essentially) singularly optimal. In general graphs, that would mean algorithms with $O(m)$ (or, at least, $\tilde{O}(m)$) messages and $\tilde{O}(D)$ (or, at least, $\tilde{O}(D)$) time. This was shown for the case of synchronous networks in [27].

For synchronous complete networks, an essentially (up to a logarithmic factor) singularly optimal deterministic algorithm was known. We obtained a tightly singularly optimal randomized one (that is, with no extra logarithmic factors).

References

1. Yehuda Afek and Eli Gafni. Time and message bounds for election in synchronous and asynchronous complete networks. *SICOMP*, 20(2):376–394, 1991.
2. Yehuda Afek and Yossi Matias. Elections in anonymous networks. *Information and Computation*, 113(2):312–330, 1994.
3. Dana Angluin. Local and global properties in networks of processors (extended abstract). In *STOC*, pages 82–93, 1980.
4. John Augustine, Gopal Pandurangan, Peter Robinson, and Eli Upfal. Towards robust and efficient distributed computation in dynamic peer-to-peer networks. In *SODA*, pages 551–569, 2012.
5. Baruch Awerbuch, Israel Cidon, and Shay Kutten. Communication-optimal maintenance of replicated information. In *Proceedings [1990] 31st Annual Symposium on Foundations of Computer Science*, pages 492–502. IEEE, 1990.
6. Baruch Awerbuch, Oded Goldreich, Ronen Vainish, and David Peleg. A trade-off between information and communication in broadcast protocols. *J. ACM*, 37:238–256, 1990.
7. Mike Burrows. The chubby lock service for loosely-coupled distributed systems. In *7th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2006.
8. Miguel Castro, Peter Druschel, A-M Kermarrec, and Antony IT Rowstron. Scribe: A large-scale and decentralized application-level multicast infrastructure. *IEEE Journal on Selected Areas in communications*, 20(8):1489–1499, 2002.
9. Tushar D Chandra, Robert Griesemer, and Joshua Redstone. Paxos made live: an engineering perspective. In *Proceedings of the twenty-sixth annual ACM symposium on Principles of distributed computing*, pages 398–407, 2007.
10. Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C Hsieh, Deborah A Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E Gruber. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, 26(2):1–26, 2008.
11. Michael Elkin. A simple deterministic distributed mst algorithm, with near-optimal time and message complexities. In *PODC*, pages 157–163, 2017.
12. Robert G. Gallager, Pierre A. Humblet, and Philip M. Spira. A distributed algorithm for minimum-weight spanning trees. *ACM Trans. Programming Languages & systems*
22 Singularly Optimal Randomized Leader Election

(TOPLAS), 5(1):66–77, January 1983. URL: http://doi.acm.org/10.1145/357195.357200
doi:10.1145/357195.357200

13 Saurabh Ganeriwal, Ram Kumar, and Mani B. Srivastava. Timing-sync protocol for sensor networks. In Proceedings of the 1st International Conference on Embedded Networked Sensor Systems, SenSys 2003, Los Angeles, California, USA, November 5-7, 2003, pages 138–149, 2003.

14 Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In Proceedings of the nineteenth ACM symposium on Operating systems principles, pages 29–43, 2003.

15 Robert Gmyr and Gopal Pandurangan. Time-message trade-offs in distributed algorithms. In 32nd International Symposium on Distributed Computing, DISC 2018, New Orleans, LA, USA, October 15-19, 2018, pages 32:1–32:18, 2018.

16 Indranil Gupta, Robbert van Renesse, and Kenneth P. Birman. A probabilistically correct leader election protocol for large groups. In Proceedings of the 14th International Conference on Distributed Computing, DISC ’00, pages 89–103, 2000. URL: http://dl.acm.org/citation.cf?id=645957.675964.

17 Bernhard Haeupler, D Ellis Hershkowitz, and David Wajc. Round-and message-optimal distributed graph algorithms. In PODC, pages 119–128, 2018.

18 Pierre A. Humblet. Selecting a leader in a clique in $O(n \log n)$ messages. Memo, Lab. for Information & Decision Systems, MIT, 1984.

19 Michael Isard. Autopilot: automatic data center management. ACM SIGOPS Operating Systems Review, 41(2):60–67, 2007.

20 Maleq Khan, Fabian Kuhn, Dahlia Malkhi, Gopal Pandurangan, and Kunal Talwar. Efficient distributed approximation algorithms via probabilistic tree embeddings. In Proceedings of the twenty-seventh ACM symposium on Principles of distributed computing, PODC ’08, pages 263–272, New York, NY, USA, 2008. ACM. URL: http://doi.acm.org/10.1145/1400751.1400787
doi:http://doi.acm.org/10.1145/1400751.1400787

21 Valerie King, Jared Saia, Vishal Sanwalani, and Erik Vee. Scalable leader election. In SODA, pages 990–999, 2006.

22 Valerie King, Jared Saia, Vishal Sanwalani, and Erik Vee. Towards secure and scalable computation in peer-to-peer networks. In FOCS, pages 87–98, 2006.

23 Ephraim Korach, Shay Kutten, and Shlomo Moran. A modular technique for the design of efficient distributed leader finding algorithms. ACM Trans. Programming Languages & Systems (TOPLAS), 12(1):84–101, January 1990. URL: http://doi.acm.org/10.1145/77606.77610
doi:10.1145/77606.77610

24 Ephraim Korach, Shlomo Moran, and Shmuel Zaks. Tight lower and upper bounds for some distributed algorithms for a complete network of processors. In Proceedings of the third annual ACM symposium on Principles of distributed computing, pages 199–207, New York, NY, USA, 1984. ACM. URL: http://doi.acm.org/10.1145/800222.806747
doi:10.1145/800222.806747

25 Ephraim Korach, Shlomo Moran, and Shmuel Zaks. The optimality of distributive constructions of minimum weight and degree restricted spanning trees in a complete network of processors. SIAM J. Computing, 16(2):231–236, 1987. URL: http://epubs.siam.org/doi/abs/10.1137/0216019 arXiv:http://epubs.siam.org/doi/pdf/10.1137/0216019
doi:10.1137/0216019

26 Ephraim Korach, Shlomo Moran, and Shmuel Zaks. Optimal lower bounds for some distributed algorithms for a complete network of processors. Theoretical Computer Science, 64(1):125 – 132, 1989. URL: http://www.sciencedirect.com/science/article/pii/0304397589901035
doi:10.1016/0304-3975(89)90103-5

27 Shay Kutten, Gopal Pandurangan, David Peleg, Peter Robinson, and Amitabh Trehan. On the complexity of universal leader election. J. ACM, 62:7, 2015.
28 Shay Kutten, Gopal Pandurangan, David Peleg, Peter Robinson, and Amitabh Trehan. Sublinear bounds for randomized leader election. *Theoretical Computer Science*, 561:134–143, 2015.
29 Shay Kutten and Dmitry Zinenko. Low communication self-stabilization through randomization. In *DISC*, pages 465–479. Springer, 2010.
30 Gérard Le Lann. Distributed systems - towards a formal approach. In *IFIP Congress*, pages 155–160, 1977.
31 Michael C. Loui, Teresa A. Matsushita, and Douglas B. West. Election in a complete network with a sense of direction. *IPL*, 22(4):185–187, 1986.
32 Michael C. Loui, Teresa A. Matsushita, and Douglas B. West. Election in a complete network with a sense of direction. *IPL*, 28(6):327, 1988.
33 Nancy Lynch. *Distributed Algorithms*. Morgan Kaufman Publishers, Inc., San Francisco, USA, 1996.
34 Dahlia Malkhi, Michael Reiter, and Rebecca Wright. Probabilistic quorum systems. In *PODC 1997*, pages 267–273, New York, NY, USA, 1997. ACM. URL: [http://doi.acm.org/10.1145/259380.259458](http://doi.acm.org/10.1145/259380.259458).
35 Michael Mitzenmacher and Eli Upfal. Probability and computing: randomization and probabilistic techniques in algorithms and data analysis. Cambridge university press, 2017.
36 Gopal Pandurangan. Distributed network algorithms. [https://sites.google.com/site/gopalpandurangan/dna](https://sites.google.com/site/gopalpandurangan/dna). Accessed: 2020-02-14.
37 Gopal Pandurangan, Peter Robinson, and Michele Scquizzato. A time-and message-optimal distributed algorithm for minimum spanning trees. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 743–756. ACM, 2017.
38 David Peleg. Time-optimal leader election in general networks. *J. Parallel & Distributed Computing*, 8(1):96 – 99, 1990. URL: [http://www.sciencedirect.com/science/article/pii/074373159090074Y](http://www.sciencedirect.com/science/article/pii/074373159090074Y) doi:10.1016/0743-7315(90)90074-Y
39 David Peleg. *Distributed Computing: A Locality-Sensitive Approach*. SIAM, 2000.
40 Radia Perlman. *Interconnections: bridges, routers, switches, and internetworking protocols*. Addison-Wesley Professional, 2000.
41 Gary Peterson. Efficient algorithms for elections in meshes and complete graphs. Technical report, TR 140, Dept. of CS, Univ. Rochester, 1985.
42 Murali Krishna Ramanathan, Ronaldo A. Ferreira, Suresh Jagannathan, Ananth Grama, and Wojciech Szwankowski. Randomized leader election. *Distributed Computing*, pages 403–418, 1997.
43 Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Shenker. A scalable content-addressable network. In *SIGCOMM 2001*, pages 161–172, New York, NY, USA, 2001. ACM. URL: [http://doi.acm.org/10.1145/383059.383072](http://doi.acm.org/10.1145/383059.383072) doi:http://doi.acm.org/10.1145/383059.383072
44 Antony I. T. Rowstron and Peter Druschel. Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. In *Proceedings of the IFIP/ACM International Conference on Distributed Systems Platforms Heidelberg, Middleware '01*, pages 329–350. Springer-Verlag, 2001. URL: [http://dl.acm.org/citation.cfm?id=646591.697650](http://dl.acm.org/citation.cfm?id=646591.697650).
45 Nicola Santoro. *Design and Analysis of Distributed Algorithms (Wiley Series on Parallel and Distributed Computing)*. Wiley-Interscience, 2006.
46 Gurdip Singh. Efficient leader election using sense of direction. *Distributed Computing*, 10(3):159–165, 1997.
47 Gurdip Singh. Leader election in complete networks. *SIAM J. Comput.*, 26(3):772–785, 1997.
48 Andrew S. Tanenbaum and Maarten Van Steen. *Distributed systems: principles and paradigms*. Prentice-Hall, 2007.
49 Gerard Tel. *Introduction to distributed algorithms*. Cambridge University Press, New York, NY, USA, 1994.
Yong Yao and Johannes Gehrke. The cougar approach to in-network query processing in sensor networks. *SIGMOD Record*, 31(3):9–18, 2002.

Ben Y. Zhao, Ling Huang, Jeremy Stribling, Sean C. Rhea, Anthony D. Joseph, and John D. Kubiatowicz. Tapestry: a resilient global-scale overlay for service deployment. *IEEE J. Selected Areas in Communications*, 22(1):41 – 53, January 2004. doi:10.1109/JSAC.2003.818784.