# An Economic Model of Consensus on Distributed Ledgers*

Hanna Halaburda† Zhiguo He‡ Jiasun Li§

November 19, 2021

**Abstract**

In recent years, the designs of many new blockchain applications have been inspired by the Byzantine fault tolerance (BFT) problem. While traditional BFT protocols assume that most system nodes are honest (in that they follow the protocol), we recognize that blockchains are deployed in environments where nodes are subject to strategic incentives. This paper develops an economic framework for analyzing such cases. Specifically, we assume that 1) non-Byzantine nodes are *rational*, so we explicitly study their incentives when participating in a BFT consensus process; 2) non-Byzantine nodes are ambiguity averse, and specifically, *Knightian uncertain* about Byzantine actions; and 3) decisions/inferences are all based on *local* information. The consensus game then resembles one with preplay "cheap talk" communications. We characterize all equilibria, some of which feature rational leaders withholding messages from some nodes in order to achieve consensus. These findings enrich those from traditional BFT algorithms, where an honest leader always sends messages to all nodes. We also study how the progress of communication technology (i.e., potential message losses) affects the equilibrium consensus outcome.

**Keywords**: Ambiguity aversion, Blockchain, Byzantine fault tolerance, Distributed consensus

*We thank Peter Klibanoff and Marciano Siniscalchi, as well as seminar participants at ABFER and IC3 for helpful comments. Zhiguo He is grateful for support from the John E. Jeuck Endowment at the University of Chicago Booth School of Business. Zhiguo He and Jiasun Li are grateful for research grants from the Paris-Dauphine Partnership Foundation.

†hhalaburda@gmail.com. Stern School of Business, New York University, 44 W 4th St, New York, NY 10012.

‡zhiguo.he@chicagobooth.edu. Booth School of Business, University of Chicago and NBER, 5807 South Woodlawn Ave, Chicago, IL 60637.

§jli29@gmu.edu. George Mason University, 4400 University Drive, MSN 1B8, Fairfax, VA 22030, USA.



Electronic copy available at: https://ssrn.com/abstract=3965955

# 1 Introduction

Bitcoin's rise in popularity has in recent years inspired many other general blockchain applications, which aim to provide better resilience to centralized systems by removing single points of failure. Examples include Ethereum 2.0, Diem led by Facebook (formerly known as Libra) and Cosmos (based on the Tendermint protocol). These applications feature distributed ledgers in which computer nodes rely on peer-to-peer communication to maintain their respective ledgers: Nodes may send different messages to peers, and messages may get lost; some nodes may also be faulty or hijacked by hackers (such nodes are called *Byzantine faulty*, often abbreviated to *Byzantine*). The challenge is to achieve consensus in such an environment, that is, to ensure that all nodes keep the same record in their respective ledgers.

For decades, extensive research in the computer science literature has developed numerous results on how to tackle this challenge of reaching consensus even in the presence of Byzantine faulty nodes. These results are commonly known as Byzantine fault tolerant (BFT) protocols and have been major inspirations for designing the many new blockchains mentioned above.

From an economist's perspective, classic BFT protocols have the following three key features. First, there are some Byzantine nodes who behave arbitrarily; the system and non-Byzantine nodes concern the "worst case" scenario regarding the Byzantine nodes' actions. Second, by the nature of a distributed system, each node only has and thus acts upon "local" information rather than "global" knowledge.[1] Finally, various BFT protocols in the computer science literature all stipulate "honest" strategies for non-Byzantine nodes and assume that they all willingly follow prescribed strategies. In other words, nodes are treated like machines rather than "rational" participants who operate with incentive considerations.

Presumably, it is easier to enforce "honest" behaviors assumed by traditional BFT protocols in a trusted environment, which matches well with the distributed systems that are typically implemented within the same company. However, nodes in the many new blockchain applications are

[1]Here, we follow the network literature (e.g. Galeotti et al. (2010)) and use "local" information to indicate information that only the node knows; our paper is about network communication among a set of computer nodes. "Local" versus "global" information is similar to private versus public information (à la Morris and Shin (2002); in Angeletos and Werning (2006), public information is provided via a centralized financial market rather than peer communication as in our model).



Electronic copy available at: https://ssrn.com/abstract=3965955

independent entities with individual (potentially conflicting) interests, and these new applications alter the trusted environments of traditional BFT problems to adversarial ones. This shift, therefore, calls for a better understanding of incentives in BFT protocols, so they can be successfully applied to blockchain systems with large stakes involved.

In this paper, we develop an economic framework incorporating the key elements of traditional BFT protocols, while explicitly modeling nodes' incentives. Specifically, we assume that (i) non-Byzantine nodes are *rational*, so we explicit study their incentives when participating in a BFT consensus process; (ii) non-Byzantine nodes are ambiguity averse, and specifically, *Knightian uncertain* about non-Byzantine actions; and (iii) inferences and, thus, decisions are all based on *local* information. In this game, consensus is defined as all rational nodes *committing* to the same value; when a node commits, she receives a reward only when consensus is reached — she incurs a penalty instead when consensus fails.

Our consensus game then resembles a game with preplay "cheap talk" communications. More specifically, our game contains multiple stages. In the first stage, one of the nodes is selected as a "leader" and sends a message to other "backup" nodes. In the second stage, these backup nodes confirm each other's message received via peer communication. In the final stage, based on her local knowledge after such communications, each node decides whether to commit to her received message, that is, to regard her received message as a consensus value.

We fully characterize all symmetric equilibria within our model. First, in line with well-established results from "cheap talk" games (see e.g. Crawford and Sobel (1982)), there always exists a set of babbling equilibria in which nodes discard preplay communications and never commit to new messages. Second, when the reward from successfully achieving consensus is sufficiently high compared to the penalty for a "wrong" commit decision (that is, committing to a message that does not obtain consensus), there also exist nonbabbling equilibria in which consensus could be reached. We characterize sharp conditions on reward and penalty for such equilibria to exist.

In these nonbabbling equilibria, each rational node uses information from communication to Bayesian update the posterior probability of the leader being rational or Byzantine. We show that a Byzantine leader, together with other Byzantine backups, may happen to coordinate and



Electronic copy available at: https://ssrn.com/abstract=3965955

lead a rational node into a wrong commit decision. Seeing this possibility, rational nodes who are ambiguity averse to Byzantine nodes' strategies will prefer not committing when they know the leader is Byzantine. As a result, a rational node commits if and only if the total measure of messages she receives lies in a certain interval, as only such communication outcomes are consistent with the leader being rational.

Finally, we show that all nonbabbling symmetric equilibria in our model can be categorized into two classes: in one class of equilibria rational leaders always send messages to all nodes, and in the other class rational leaders withhold messages from some nodes. While traditional BFT protocols resemble the former, we show that this class of equilibria is nongeneric and not robust to potential message losses.

We also analyze how technological parameters, e.g., the probability that a message sent may be lost affects our results. We show that idiosyncratic message losses tend to decrease the probability of reaching consensus, yet systematic message losses have an inverted U-shaped effect on the probability of reaching consensus. More specifically, when there exists a "bad" aggregate state in which messages will be lost randomly, any rational node—say Alice—without seeing the realized aggregate state needs to not only worry about other rational nodes'—say Bob's—local knowledge, but also their higher-order beliefs, i.e., how Bob thinks about Alice's local knowledge. We show that the unanimity requirement of consensus has such a strong bite in this inference problem that any small prior probability of this bad aggregate state can prevent the system from reaching a consensus. This negative outcome for any small probability of possible message losses is related to the celebrated result in Rubinstein (1989).

In sum, inspired by widely used BFT consensus protocols in the computer science literature and yet explicitly tackling incentive considerations, this paper develops an economic framework for analyzing BFT consensus protocols in strategic settings as seen from many new blockchain applications. A key departure of our analysis from the mainstream computer science literature is the incorporation of payoffs, as we find that the existence and structure of (multiple) equilibria depend on the payoffs the nodes receive when the consensus is reached or not. This result could provide guidance for how blockchain protocol designers can set incentives—including both reward



Electronic copy available at: https://ssrn.com/abstract=3965955

and penalties—for participants in the consensus process. We hope that our framework lays the foundation for further research on connecting game theoretical modeling and distributed consensus.

**Related Literature** Studies of Byzantine fault tolerant consensus mechanisms start with Lamport, Shostak and Pease (1982), who formulated the Byzantine generals problem and showed that consensus is possible. Castro and Liskov (1999) further streamline the consensus algorithm as a practical Byzantine fault tolerant (PBFT) mechanism. More recent developments in BFT protocols include Buterin and Griffith (2017), Buchman (2016), Pass and Shi (2018), Yin et al. (2018), etc. See Shi (2020) for a summary. While this literature develops algorithms for achieving consensus in the presence of Byzantine faulty nodes, it does so by assuming that the nonfaulty nodes are "honest," i.e., follow the prescribed protocol without incentive considerations.[2]

In contrast, an emerging literature in economics concerns whether the nonfaulty nodes would find it optimal to follow prescribed protocols, and recognizes that they can deviate from prescribed protocols if they find it beneficial. That is, the nonfaulty nodes are "rational" rather than "honest." While incentives in consensus formation have been studied quite extensively in the context of permissionless proof-of-work (PoW) protocols including Bitcoin (e.g., Kroll, Davey and Felten (2013), Kiayias et al. (2016), Budish (2018), Leshno and Strack (2020), Cong, He and Li (2021)), and similarly in other permissionless consensus protocols such as proof of stake (e.g, Saleh (2021)) or proof of presence (Branderburger and Steverson (2020), such studies in BFT protocols are more scarce.[3]

A prominent example of incentive analysis in BFT protocols is Amoussou-Guenou et al. (2020). The authors recognize that non-Byzantine nodes do not need to follow the protocol if they do not find it beneficial. Specifically, the nodes find it costly to check the validity of the proposed message and send the confirmation to other nodes.[4] They benefit when the consensus is reached, i.e., when a sufficiently large fraction of nodes vote in favor of the message. This combination creates free-

[2]There are attempts in the computer science literature to bring rationality into BFT analysis, see Abraham, Alvisi and Halpern (2011) for a review. These papers take a mechanism design perspective and check whether certain centralized systems can be decentralized. However, they do not characterize all possible equilibria as we do here.

[3]For other papers that study the broader implications of blockchain technology, see Cong and He (2019) and Abadi and Brunnermeier (2018), among others. See Halaburda et al. (forthcoming) for an overview of this literature.

[4]Motivating deviations from protocol prescriptions by operational costs has also been used in the computer science literature, see e.g. the BAR model (Aiyer et al. (2005) and Clement et al. (2008)).



Electronic copy available at: https://ssrn.com/abstract=3965955

riding incentives and a coordination problem, which results in a possible equilibrium where no node takes action, and thus the messages are not added to the ledger. Thus, like us, Amoussou-Guenou et al. (2020) show that rational non-Byzantine nodes in BFT protocol may lead to the babbling equilibrium, though driven by different forces. We also identify a variety of other equilibria.

Auer, Monnet and Shin (2021) consider a voting-based consensus system, similar to Amoussou-Guenou et al. (2020), in the context of permissioned distributed ledgers. Costly message verification and sending also leads to coordination and free-riding problems. These problems are solved if the nodes are sufficiently compensated for participation. While in the classical BFT formulation some nodes are Byzantine, in Auer, Monnet and Shin (2021) all nodes are rational, but they can be bribed to introduce false messages. Auer, Monnet and Shin (2021) derive conditions when the nodes would find it more beneficial to follow the protocol than to take the bribe.

In contrast to Amoussou-Guenou et al. (2020) and Auer, Monnet and Shin (2021), we look at incentives to follow the protocol even when there is no cost to validate and send messages. The BFT protocol prescribes that nodes send the same messages to all the other nodes, but it recognizes that Byzantine nodes can send different messages to different recipients, including sending no message to some. We analyze possible equilibria recognizing that rational nodes also decide whether to send messages to everyone or only to selected recipients.

Outside of the consensus game within a committee once it has been formed, Benhaim, Hemenway Falk and Tsoukalas (2021) look at the committee formation process and provide an interesting connection between voting and BFT mechanisms in the context of delegated proof-of-stake mechanism. The participants who own the stake in the blockchain do not directly participate in the validation of the blocks. Instead, the blocks are validated by a committee of *block producers* via BFT mechanisms, and the stakeholders vote on which of the block producers will be on the committee, utilizing their private information about each block producer's type. The block producers can be either honest or malicious, but the stakeholders are rational and strategic in their voting. Benhaim, Hemenway Falk and Tsoukalas (2021) study optimal voting strategies where the stakeholder's objective is to select a committee that is composed of at least two-thirds honest block producers. They show that even with little private information, stakeholders can still elect robust



Electronic copy available at: https://ssrn.com/abstract=3965955

committees. Our analysis, however, is rather concerned with what happens after the committee is set, if we relax the assumption that some block producers always follow the protocol.

The rest of the paper proceeds as follows. Section 2 describes our baseline consensus game, assuming that all messages sent will be delivered for sure. Section 3 and 4 characterize all symmetric equilibria of this baseline model. An extension with potential message losses is considered in Section 5. Section 6 connects our model to practical BFT protocol and discusses directions for future research. Section 7 concludes.

# 2 The Model

This section lays out the model ingredients and formalizes our equilibrium concept.

## 2.1 Sequence of Moves

We study a *consensus game* among a measure of $n$ computer nodes with the following sequence of moves:[5] First, nature randomly selects one node as the *leader*, and designates all other nodes as *backups*. The leader then decides whether to *send* a `message` to each backup. The content of `message` is application specific. For example, in the original Byzantine generals problem, `message` can be interpreted as "leader orders to attack," while in the context of a transaction ledger, `message` can be interpreted as a new transaction (or set of transactions) to be added to the ledger. Following the tradition of BFT protocols, every `message` from the leader contains her digital signature that others cannot forge. Note that the leader may send `message` to some backups but not others.

Each backup who receives `message` then decides, for each other node, whether to *forward* `message`, while a backup not receiving `message` does nothing. Because of the leader's digital signature, in the forwarding stage, a backup cannot fabricate a message that is different from what she has received from the leader, or make up one if she did not receive any in the first place. Each forwarded `message` also contains the forwarding backup's digital signature, so for any given backup

[5]For simplicity, we study one round of synchronous peer communication in a single view. Lamport, Shostak and Pease (1982) study $f$ rounds of peer communication. Castro and Liskov (1999) (PBFT) study two rounds of potentially asynchronous communication with view changes. We also assume adequately close message delivery speeds to justify simultaneous moves in each step.



Electronic copy available at: https://ssrn.com/abstract=3965955

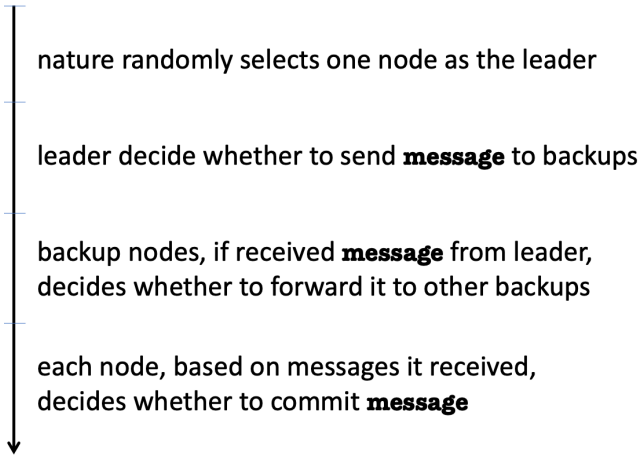



Figure 1: Sequence of moves

$i$, no other nodes can impersonate $i$ and forward `messages` on $i$'s behalf.

After the previous steps, each node decides whether to *commit* to `message` based on her local information. A commit decision can be interpreted as taking a certain application-specific action. For example, in the original Byzantine generals problem, committing to `message` can be interpreted as "attacking," while in the context of a transaction ledger (or more generally, any state machine replication problem, e.g., Castro and Liskov (1999)), a node's commit decision can be interpreted as the node adding the transactions in `message` to her own local ledger (or updating her local database). We will be studying the second context, so that a node that has received no `messages` cannot commit. Note that this is different from a traditional coordination game (e.g., the traditional Byzantine generals problem and the email game in Rubinstein (1989)), in which agents' action spaces are not affected by their information.

Figure 1 provides a detailed timeline of the sequence of moves in the consensus game.

## 2.2 Agents

There are a measure of $n$ nodes in the system; we will explain the role of the "continuum" toward the end of Section 2.4. Following the literature on Byzantine fault tolerance protocols, we differentiate between two types of nodes. First, there exists a measure of $f$ *Byzantine* nodes, who may together have an "arbitrary" strategy profile denoted by $B$, describing all Byzantine nodes'



Electronic copy available at: https://ssrn.com/abstract=3965955

sending, forwarding, and committing decisions. The set of all feasible Byzantine strategy profiles is denoted $\mathcal{B}$.

Second, the remaining measure $n-f$ of nodes are non-Byzantine. In traditional Byzantine fault tolerance protocols, these nodes are often called *honest* as they are assumed to loyally follow the strategies prescribed by the protocol. A key contribution of our study is to relax this "honesty" assumption so that non-Byzantine nodes will behave according to certain well-defined preferences rather than blindly follow protocol prescriptions. Hence, in the rest of the paper, we refer to these non-Byzantine nodes as *rational* nodes. We will first give a formal definition of *consensus* in Section 2.3 and then build on this in Section 2.4 to provide more details about these nodes' preferences.

### 2.3 Consensus

*Consensus* is a central concept in the proper functioning of distributed systems and will also be a desirable outcome of our game. Throughout the paper we define consensus as follows.

**Definition 1** (Consensus)**.** *Consensus on* `message` *succeeds, or is reached, if and only if "almost all" (measure $n-f$) rational nodes commit. Otherwise, consensus fails.*

In the original Byzantine generals problem, consensus on `message` implies that all rational players "attack". In the context of transactions ledgers, consensus on `message` implies (almost) all rational nodes agree on the same state across their local ledgers. Consensus has to be be reached via peer communications described in the previous section, since there is no centralized "reference point" coordinating it.

### 2.4 Payoffs

Traditional BFT protocols prescribe strategies so that an "honest" node only commits to `message` when she knows that other honest nodes also commit to `message`. To capture such behaviors, we assign utilities to rational nodes so that they prefer committing to `message` if and only if they believe it would reach consensus. We thus construct the following utilities: When a rational node commits to `message`, she receives a positive reward $R > 0$ if consensus succeeds and a penalty



Electronic copy available at: https://ssrn.com/abstract=3965955

$c > 0$ if consensus fails. A rational node who does not commit always gets 0. The following table illustrates this utility specification.

| | If consensus on `message` succeeds | If consensus on `message` fails |
|---|---|---|
| Commit to`message` | $R > 0$ | $-c < 0$ |
| Not commit to`message` | 0 | 0 |

Formally, denote a rational node $i$'s action by $a_i$, which consists of a tuple of $(p_i, q_i, C_i)$ within the action space $\mathcal{A} \equiv [0,1]^2 \times \{\text{commit}, \text{not commit}\}$. Here, $p_i \in [0,1]$ indicates that $i$ sends `message` to all backups with i.i.d. probability $p_i$ when she is selected as a leader, $q_i \in [0,1]$ indicates that $i$ forwards the leader's `message` (if received) to all other peer nodes with i.i.d. probability $q_i$ when she is selected as a backup, and $C_i \in \{\text{commit}, \text{not commit}\}$ denotes $i$'s eventual commit decision. Then, for a given action profile $A_{-i} \equiv \{a_j\}_{j \neq i}$ of other rational nodes and Byzantine nodes' strategy profile $B$, a rational node $i$'s utility in the consensus game is given by:

$$u_i(a_i, A_{-i}; B) = \mathbb{1}_{\text{commit} \in a_i} \cdot \left( \mathbb{1}_{|j:\text{commit} \notin a_j| = 0} \cdot R + \mathbb{1}_{|j:\text{commit} \notin a_j| > 0} \cdot (-c) \right), \tag{1}$$

where the term "commit $\in a_i$" denotes that node $i$ commits to `message`, and $|j : \text{commit} \notin a_j|$ denotes the measure of rational nodes who do not commit.

According to the utility specification in (1), a rational node is rewarded if she commits together with all her rational peers and is penalized otherwise. Thus, our game resembles a standard coordination game. On the other hand, since only committing actions but no sending/forwarding actions enter utilities, the game also has a "cheap talk" flavor à la Crawford and Sobel (1982).

Even though the dynamic nature of our game and the to-be-imposed sequential rationality requires a rational node $i$'s sending, forwarding, and committing decisions to be all optimal, eventually we will only need to be concerned about $i$'s commitment decision. First, the "cheap talk" nature of our game ensures that $i$'s sending strategy as a leader and forwarding strategy as a backup receiving `message` do not directly affect $i$'s utility as specified in Eq. (1). Second, with a



Electronic copy available at: https://ssrn.com/abstract=3965955

continuum of nodes, each single (zero-measure) backup's forwarding strategy does not affect other rational nodes' information sets, and thus their equilibrium actions. Therefore, the "cheap-talk" nature, which is intrinsic to consensus games in general, and the continuum assumption, which we specifically impose for our model, will significantly simplify our equilibrium characterization later.

### 2.5 Ambiguity Aversion toward Byzantine Strategies

Our game is one with imperfect information as each node acts upon her local information set after communications. We thus incorporate Byzantine behaviors into the well-established solution concept of perfect Bayesian equilibrium (PBE). Recall that a PBE specifies a set of strategies and beliefs that satisfy (i) sequential rationality, i.e., a rational node's strategy maximizes her expected utility given her belief at every information set, and (ii) belief consistency, i.e., a node's belief follows Bayesian updating at every information set. The presence of Byzantine nodes who may take arbitrary actions, however, complicates both requirements above. Regarding sequential rationality, the issue is how to set expectation for Byzantine node's uncertain actions. Regarding belief consistency, the issue is how to Bayesian update from a Byzantine node's uncertain actions. To address both challenges, we follow the ambiguity-aversion literature (Gilboa and Schmeidler (1993), Epstein and Schneider (2003), Siniscalchi (2011), Hanany, Klibanoff and Mukerji (2020), etc. See Machina and Siniscalchi (2014) for a review.) and adopt a multiprior framework in which rational nodes are Knightian uncertain about all Byzantine nodes' strategy profile and have max-min utilities over them, while having expected utilities over the state of nature. Our modelling approach is similar to Eliaz (2002) and is also related to the literature on robust mechanism design (e.g., Bergemann and Morris (2005)).

Formally, a rational node $i$ at any information set $I_i$ who is ambiguity averse towards Byzantine strategies in $\mathcal{B}$ chooses action $a_i \in \mathcal{A}$ to maximize

$$\min_{B \in \mathcal{B}} \mathbb{E}[u_i(a_i, A_{-i}; B)|I_i]. \tag{2}$$

The Byzantine nodes' strategy profile $B$ specifies the actions of a Byzantine leader (if the leader happens to be Byzantine) as well as how Byzantine backups forward the leader's messages, con-



Electronic copy available at: https://ssrn.com/abstract=3965955

tingent on whether the leader is Byzantine or not. In the computer science tradition, Byzantine nodes are assumed to be able to perfectly coordinate.[6] With rational nodes being ambiguity-averse toward Byzantine nodes' strategies, our setting accommodates the possibility of coordinated Byzantine nodes, but does not necessarily assume so. This is because rational nodes max-min over all possible $B$'s in $\mathcal{B}$, which includes the strategies where the Byzantine nodes coordinate.

### 2.6 Equilibrium Definition

A perfect Bayesian equilibrium in our environment is defined over every rational node $i$'s strategy $\tilde{a}_i \equiv \{p_i, q_i, \tilde{C}_i\}$. The first two elements in the tuple directly follow the action $a_i$: $p_i \in [0,1]$ denotes node $i$'s i.i.d. probability of sending `message` to all backups when being a leader, and $q_i \in [0,1]$ denotes node $i$'s i.i.d. probability of forwarding `message` to all other peer nodes when being a backup who has received `message` from the leader. The third element $\tilde{C}_i : \{0,1\} \times [0,1] \to \{\text{commit, not commit}\}$ denotes node $i$'s commit strategy when being a backup: It maps from a specific information set $I_i \equiv \{z, k\}$ to a decision of whether to commit or not, where $z \in \{0,1\}$ represents whether `message` is received from the leader ($z=1$) or not ($z=0$) and $k \in [0,n]$ denotes the measure of `messages` collected from communications. Note that a backup would only be able to commit `message` if she receives at least one `message`.

We focus on symmetric perfect Bayesian equilibria, where "symmetry" requires every *rational* node to follow the same strategy (while Byzantine nodes may have arbitrary strategy profiles). Hence, we can define a symmetric perfect Bayesian equilibria in our setup as follows:

**Definition 2** (Symmetric perfect Bayesian equilibrium)**.** *A symmetric equilibrium consists of a profile of rational nodes' strategies* $\{\tilde{a}_i^*\}_{i=1}^n$ *and beliefs over whether the leader is Byzantine or not, so that* $\forall i$, $\tilde{a}_i^* = \{p, q, \tilde{C}\}$ *where*

1. *a rational leader sends* `message` *to each backup with probability* $p \in [0,1]$*;*
2. *a rational backup who receives* `message` *from the leader forwards it with probability* $q \in [0,1]$*;*
3. *a rational node commits to* `message` *if and only if it receives*

[6] One variant in the computer science literature is Groce et al. (2012), who studies consensus among honest nodes and rational adversaries, and thus assumes away Byzantine behaviors.



Electronic copy available at: https://ssrn.com/abstract=3965955

*(a)* $k \in \mathcal{E}^1 \subseteq [0, n]$ `messages`*, with one from the leader, or*

*(b)* $k \in \mathcal{E}^0 \subseteq [0, n]$ `messages`*, and none of which is from the leader,*

$$\textit{that is, } \tilde{C}(z,k) = \begin{cases} \textit{commit}, & \textit{if } k \in \mathcal{E}^z \\ \textit{not commit}, & k \notin \mathcal{E}^z \end{cases} \qquad \textit{for } z = \{0,1\}.$$

*Given other rational nodes' equilibrium strategies* $\tilde{A}^*_{-i} \equiv \{\tilde{a}^*_j\}_{j \neq i}$*,* $\tilde{a}^*_i$ *maximizes* $i$*'s multiprior expected utility*

$$\tilde{C}(I_i) \in \arg\max_{a_i \in \mathcal{A}} \min_{B \in \mathcal{B}} \mathbb{E}[u_i(a_i, \tilde{A}^*_{-i}; B) | I_i], \tag{3}$$

*where the expected utility is based on* $i$*'s belief over whether leader is Byzantine or not as well as the realizations of* $A^*_{-i}$ *consistent with Bayesian updating given any Byzantine strategy* $B$*.*[7]

Condition (3) seemingly only verifies $i$'s committing decision, even though sequential rationality requires one's $p$, $q$, and $\tilde{C}$ choices to all be optimal. Note, however, that $p$ and $q$ enter $i$'s committing decision in (3) through $\tilde{A}^*_{-i}$.

The key to solving the equilibria is to characterize two sets $\mathcal{E}^1$ and $\mathcal{E}^0$, i.e., the measures of `messages` that convince the rational node to commit. Thus characterizing $\mathcal{E}^1$ and $\mathcal{E}^0$ fully defines the commit strategy $\tilde{C}$ given the commit-stage information set. Here we have used "symmetry" so the identities of forwarders do not matter. Naturally, the node's commit decision depends on whether she has received the message from the leader, as this fact carries information about whether the leader is Byzantine or not.

## 3 Characterizing Sets $\mathcal{E}^0$ and $\mathcal{E}^1$ in Equilibria

Denote $\mathcal{E} \equiv \mathcal{E}^0 \cup \mathcal{E}^1$. For any $p$ and $q$, there always exists a babbling equilibrium where $\mathcal{E} = \emptyset$, i.e., rational nodes choose to not commit to `message`, regardless of what happens during the communication stage.[8] However, we are more interested in the existence of nonbabbling equilibria. Hence, this section characterizes the set $\mathcal{E}$ for any symmetric nonbabbling equilibrium with a given

[7] For strategies $B$ that are "inconsistent" with $I_i$, i.e., $\mathbb{P}(B|I_i) = 0$, we follow the convention of $u_i(a_i, \tilde{A}^*_{-i}; B) = \infty$.

[8] The existence of a babbling equilibrium is a standard result in cheap talk games (e.g., Crawford and Sobel, 1982).



Electronic copy available at: https://ssrn.com/abstract=3965955

pair of $(p, q)$. For clarity of exposition, our analysis focuses on $p > 0$ and $q > 0$.[9]

Since in any equilibrium for a given $(p, q)$, a backup can receive at most $(n-f)q+f$ `messages`,[10] Without loss of generality, we assume that a rational node with an off-equilibrium path $k > (n-f)q+f$ believes that no other nodes commit and thus does not commit either.[11]

### 3.1 Utility and Information Sets of Rational Nodes

Based on the formulation in (2), we study a rational backup $i$'s optimal decision by analyzing her payoff from either committing to `message` or not, in which a key step in our derivation is to conduct Bayesian updating in a multiprior framework.

**Utility under Ambiguity Aversion** We separate the event in which the leader is rational, which we denote as $\mathcal{R}$, and the event in which the leader is Byzantine, which we denote as $\overline{\mathcal{R}}$.

Given other rational nodes' equilibrium strategy profile $A^*_{-i}$ (i.e., $p$, $q$ and $\mathcal{E}$) and information $I_i$ from (2), we have a rational backup $i$'s utility from committing to `message` as:

$$\min_{B\in\mathcal{B}} \mathbb{E}[u_i(commit, A^*_{-i}, B)|I_i] = \min_{B\in\mathcal{B}} \left\{ \underbrace{\mathbb{P}(\mathcal{R}|B, I_i)u_i(commit, A^*_{-i}; B; \mathcal{R})}_{\text{When the leader is rational}} + \underbrace{\mathbb{P}(\overline{\mathcal{R}}|B, I_i)u_i(commit, A^*_{-i}; B; \overline{\mathcal{R}})}_{\text{When the leader is Byzantine}} \right\}. \tag{4}$$

Here, $\mathbb{P}(\mathcal{R}|B, I_i)$ (or $\mathbb{P}(\overline{\mathcal{R}}|B, I_i)$) denotes $i$'s inferred posterior probability of the leader being rational (or Byzantine) conditional on information $I_i$ and a given Byzantine strategy profile $B$, with

$$\mathbb{P}(\overline{\mathcal{R}}|B, I_i) = 1 - \mathbb{P}(\mathcal{R}|B, I_i), \tag{5}$$

and $u_i(commit, A^*_{-i}; B; \mathcal{R})$ denotes (with a slight abuse of notation) $i$'s payoff when she commits, other rational nodes follow $A^*_{-i}$, Byzantine nodes follow $B$, and the leader is rational; $u_i(commit, A^*_{-i}; B; \overline{\mathcal{R}})$ is defined analogously.

[9] Section 4.4 below will give a brief comment on $q=0$ to discuss the role of peer communication, and Appendix C will show that for $p=0$ no nonbabbling equilibrium exists.

[10] This case happens when the leader is Byzantine and sends `message` to everyone, and all Byzantine backup nodes forward `message` to everyone. However, Byzantine nodes cannot make rational backups forward `message` more often than $q$.

[11] Recall that a PBE does not restrict beliefs on off-equilibrium paths.



Electronic copy available at: https://ssrn.com/abstract=3965955

**Information Sets of Rational Nodes** In this section, we introduce the notation for rational nodes' information sets. Define a class of sets indexed by $p$ and $q$:

$$\mathcal{S}(p,q) \equiv [(n-f)pq, (n-f)pq + fp]\,. \tag{6}$$

By Definition 2, in an equilibrium with $p$ and $q$, if the leader is rational, all rational nodes will receive $k \in \mathcal{S}(p,q)$ `messages`. Our main analysis focuses on $p \in (0,1]$ and $q \in (0,1]$; we consider the special cases of $p = 0$ or $q = 0$ later.

In an equilibrium with $p \in (0,1]$ and $q \in (0,1]$, define $\mathcal{I}^R$ as the collection of commit-stage information sets that are consistent with a rational node being chosen as the leader. Then,

$$\mathcal{I}^R \equiv \begin{cases} \{z,k\} :\ z \in \{0,1\} \text{ and } k \in S(p,q), & \text{if } p \in (0,1); \\ \{z,k\} :\ z = 1 \text{ and } k \in S(1,q), & \text{if } p = 1. \end{cases} \tag{7}$$

Expression (7) distinguishes the two cases of $p \in (0,1)$ and $p = 1$ because when $p \in (0,1)$, even under a rational leader only a fraction $p \in (0,1)$ of rational backups directly receive `message` from the leader. They thus consider it to be possible for $z$ to be either 0 or 1. When $p = 1$, however, (almost) all rational backups receive `message` from the leader, that is $z = 1$.

For ease of exposition, we also partition $\mathcal{I}^R$ by whether $z = 0$ or $z = 1$, so that

$$\mathcal{I}^0 \equiv \big\{\{z,k\} :\ \{z,k\} \in \mathcal{I}^R \text{ and } z = 0\big\} \quad \text{and} \quad \mathcal{I}^1 \equiv \big\{\{z,k\} :\ \{z,k\} \in \mathcal{I}^R \text{ and } z = 1\big\}.$$

Notice that $\mathcal{I}^0 \cup \mathcal{I}^1 = \mathcal{I}^R$ and $\mathcal{I}^0 \cap \mathcal{I}^1 = \emptyset$.

A rational backup node $i$ with information $I_i \notin \mathcal{I}^R$ at the commit-stage can infer that the leader is definitely Byzantine, i.e., $\mathbb{P}(\overline{\mathcal{R}}|B, I_i \notin \mathcal{I}^R) = 1$. Commit-stage information $I_i \in \mathcal{I}^R$, however, does not guarantee a rational leader, as a Byzantine leader may also give $I_i \in \mathcal{I}^R$ to node $i$.

### 3.2 Key Byzantine Strategy Profiles

Among many possible Byzantine strategies, we consider a particular set of Byzantine strategy profiles $\mathcal{B}^z(k)$ for any $k \in [0, (n-f)q + f]$. Any strategy profile $B \in \mathcal{B}^z(k)$ specifies that when the



Electronic copy available at: https://ssrn.com/abstract=3965955

leader is Byzantine, she sends `message` to max $\left\{0, \frac{k-f}{q}\right\}$ rational backups (excluding $i$ if $z = 0$ or including $i$ if $z = 1$, and all Byzantine backups); $\min\{f, k\}$ Byzantine backups forward `message` to $i$; and all Byzantine backups forward `message` to all other rational backups with probability $l/f$, where $l \in [0, \min\{f, k\})$. Figure 2 illustrates the strategy profiles.

The set $\mathcal{B}^z(k)$ plays a special role in later proofs as any $B \in \mathcal{B}^z(k)$ leads to node $i$ receiving $k$ `messages` (with or without the leader's, indicated by $z$) while other rational nodes receive an arbitrary measure of $l + \max\{0, k - f\} < k$ `messages`.

### 3.3 Relation between $\mathcal{E}$, $\mathcal{I}^R$, and $\mathcal{S}(p, q)$

In this section, we characterize the relation between the commit sets $\mathcal{E}$, set $\mathcal{I}^R$, and $S(p, q)$. Lemma 1 starts with an iterated elimination of strictly dominated strategies (IESDS) argument and shows that all rational nodes who know the leader is Byzantine (except for a zero measure of them) have a payoff of $-c$ from committing to `message` and thus do not commit.

**Lemma 1.** *A rational backup who knows the leader is Byzantine has a multiprior expected utility from committing to* `message` *as* $\min_{B \in \mathcal{B}} u_i(commit, A^*_{-i}; B; \overline{\mathcal{R}}) = -c$ *and thus does not commit* `message`*, except for when* $p = 1$ *and she receives exactly* $k = (n - f)q + f$ `messages`*.*

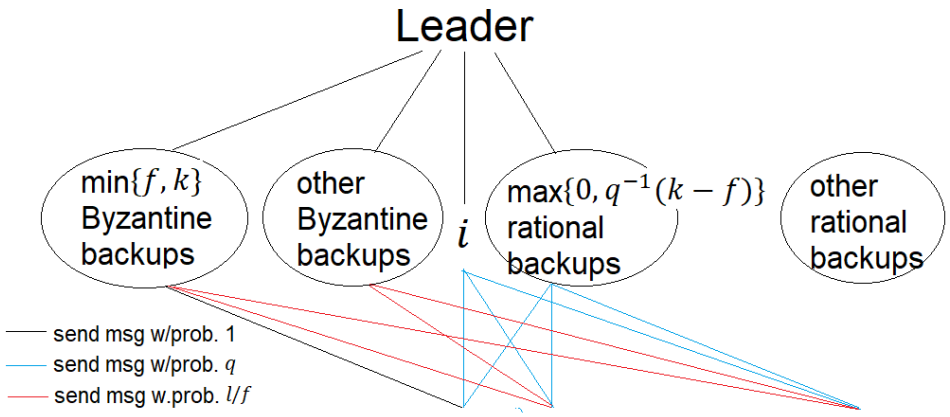
*Proof.* We first prove by induction that if a rational node $i$ knows the leader is Byzantine and has information $I_i = \{z, k\}$ where $k < (n - f)pq + f$, then there exists a Byzantine strategy in $\mathcal{B}^z(k)$ such that a positive measure of rational nodes do not commit.

In Step 1 of of the induction argument, consider a rational node $i$ who knows the leader is Byzantine and receives some $k^0 < f$ `messages`. Byzantine strategy profile $B \in \mathcal{B}^z(k^0)$ with $l = 0$ would result in all other rational backups nodes receiving no `messages`, and hence make it impossible for them to commit. If this is the case, there is no consensus on `message`, and thus, $\min_{B \in \mathcal{B}} \mathbb{E}\left[u_i(commit, A^*_{-i}, B) | \{z, k^0\}\right] \leq u_i(commit, A^*_{-i}; B \in \mathcal{B}^z(k^0); \overline{\mathcal{R}}) = -c$. Compared to the utility 0 from not committing, rational node $i$ would strictly prefer not committing to `messsage`.

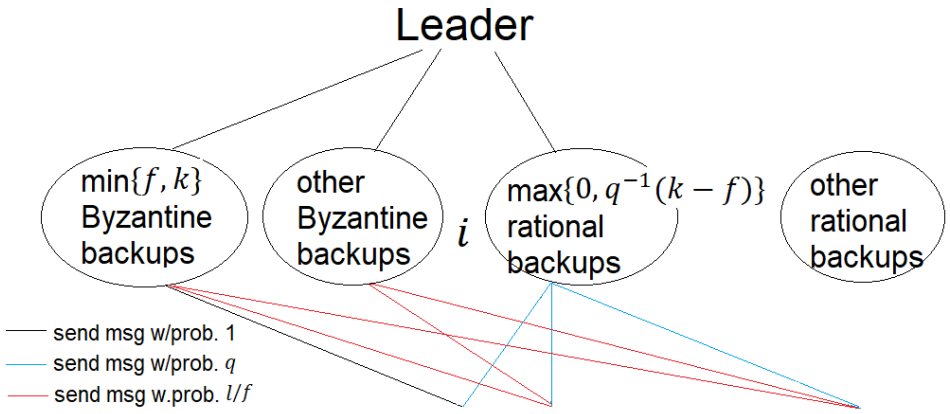
In Step 2 of of the induction argument, assuming that any rational node who receives $k^{m-1} \in [(m-1)f, mf) \cap [0, (n-f)pq+f)$ `messages` and knows that the leader is Byzantine does not commit, we prove that a rational node $i$ receiving $k^m \in [mf, (m + 1)f) \cap [0, (n - f)pq + f)$ `messages`



Electronic copy available at: https://ssrn.com/abstract=3965955




$\mathcal{B}^1(k)$




$\mathcal{B}^0(k)$

Figure 2: Illustration of $\mathcal{B}^1(k)$ and $\mathcal{B}^0(k)$

The upper figure illustrates $\mathcal{B}^1(k)$: The leader is Byzantine, and she sends `message` to max $\left\{0, \frac{k-f}{q}\right\}$ rational backups *including* $i$ and all Byzantine backups; $\min\{f, k\}$ Byzantine backups forward `message` to $i$; all Byzantine backups forward `message` to all other rational backups with probability $l/f$ where $l \in [0, \min\{f, k\})$. The set of strategies $\mathcal{B}^1(k)$ have the following outcome: Node $i$ receives $k$ `messages`, with one from the leader, while other rational nodes receive an arbitrary measure of $l + \max\{0, k - f\} < k$ `messages`.

The lower figure illustrates $\mathcal{B}^0(k)$: The leader is Byzantine, and she sends `message` to max $\left\{0, \frac{k-f}{q}\right\}$ rational backups *excluding* $i$ and all Byzantine backups; $\min\{f, k\}$ Byzantine backups forward `message` to $i$; all Byzantine backups forward `message` to all other rational backups with probability $l/f$ where $l \in [0, \min\{f, k\})$. The set of strategies $\mathcal{B}^0(k)$ have the following outcome: Node $i$ receives $k$ `messages`, without one from the leader, while other rational nodes receive an arbitrary measure of $l + \max\{0, k - f\} < k$ `messages`.



Electronic copy available at: https://ssrn.com/abstract=3965955

and who knows the leader is Byzantine also strictly prefers not committing. This is because a Byzantine strategy profile $B \in \mathcal{B}^z(k^m)$ with $l = 0$ would result in all other rational nodes receiving $k^m - f \in \big[(m-1)f, mf\big) \cap [0, (n-f)pq)$ `messages`. Since $k^m - f < (n-f)pq$, these nodes definitely know that the leader is Byzantine as neither $\{0, k^m - f\}$, nor $\{1, k^m - f\}$ are within $\mathcal{I}^R$, and thus they do not commit by the induction assumption. Then, $\min_{B \in \mathcal{B}} \mathbb{E}[u_i(commit, A^*_{-i}, B) | \{z, k^m\}] \leq u_i(commit, A^*_{-i}; B \in \mathcal{B}^z(k^m); \overline{\mathcal{R}}) = -c$ and backup $i$ does not commit to `message`.

We next prove by induction that when $p < 1$, if a rational node $i$ knows the leader is Byzantine and has information $I_i = \{z, k\}$ where $k \geq (n-f)pq + f$, then there also exists a Byzantine strategy in $\mathcal{B}^z(k)$ such that a positive measure of rational nodes do not commit.

In Step 1 of of the induction argument, consider a rational node $i$ who knows the leader is Byzantine and receives some $k^0 \in [(n-f)pq + f, (n-f)pq + pf + f)$ `messages`. There exists $B \in \mathcal{B}^z(k^0)$ within which all other rational backups nodes receive $k' = (n-f)pq + pf + \epsilon \in ((n-f)pq + pf, (n-f)pq + f)$ `messages`, so they infer that the leader is Byzantine and do not commit by the first part on $k' < (n-f)pq + f$.[12] Thus, node $i$'s utility from committing is $-c$, and she does not commit.

In Step 2 of of the induction argument, assuming that any rational node who receives $k^{m-1} \in [(n-f)pq + pf + (m-1)f, (n-f)pq + pf + mf) \cap [(n-f)pq + f, (n-f)q + f]$ `messages` and knows that the leader is Byzantine does not commit, we prove that a rational node $i$ receiving $k^m \in [(n-f)pq + pf + mf, (n-f)pq + pf + (m+1)f) \cap [(n-f)pq + f, (n-f)q + f]$ `messages` and who knows the leader is Byzantine also strictly prefers not committing. This is because a Byzantine strategy profile $B \in \mathcal{B}^z(k^m)$ with $l = 0$ would result in all other rational nodes receiving $k^m - f \in \big[(n-f)pq + pf + (m-1)f, (n-f)pq + pf + mf\big) \cap [(n-f)pq + f, (n-f)q + f]$ `messages`. Since for $p < 1$, $k^m - f \geq (n-f)pq + f > (n-f)pq + pf$, these nodes definitely know that the leader is Byzantine, and thus do not commit by the induction assumption. Then, $\min_{B \in \mathcal{B}} \mathbb{E}[u_i(commit, A^*_{-i}, B) | \{z, k^m\}] \leq u_i(commit, A^*_{-i}; B \in \mathcal{B}^z(k^m); \overline{\mathcal{R}}) = -c$ and backup $i$ does not commit to `message`.

Note that we cannot use the induction argument for $k > (n-f)q + f$, as then $\mathcal{B}^z(k)$ would not

[12] This happens when $l = f - (k^0 - k')$.



Electronic copy available at: https://ssrn.com/abstract=3965955

be well defined. However, since receiving $k > (n-f)q+f$ is off equilibrium path, by our earlier specification, a rational node expects that a positive measure of rational nodes do not commit. Therefore, for any $z$ and any $k$, we obtain that rational node $i$'s expected utility of committing `message` is $-c$, if she knows that the leader is Byzantine and the node does not commit. □

When $p = 1$, from the first part of the proof we have that a rational node $i$ receiving $k < (n-f)q+f$ `messages` and who knows the leader is Byzantine gets $-c$ from committing and thus does not commit. On the other hand, an exception arises when a rational node $i$ receives $k = (n-f)q+f$ `messages` and knows that the leader is Byzantine. This is because such case can occur only if the leader has sent `message` to everyone, so all nodes got information sets within $\mathcal{I}^R$ and cannot tell that the leader is Byzantine. We will later see that in a symmetric nonbabbling equilibrium with $p = 1$ a rational node who gets $k = (n-f)q+f$ `messages` may prefer to commit. That said, these nodes are always of measure zero and thus their strategies would not affect equilibrium outcomes. While our subsequent analyses may still discuss such cases for completeness, one may simply ignore this exception.

At the commit stage, a node's information includes how many `messages` she has received and whether she receives `message` from the leader. If this information is inconsistent with a rational leader's strategy given $p$ and $q$, the node infers that the leader is Byzantine. By Lemma 1, the node never commits in such a case. Therefore, a rational node commits to `message` only if her information set is consistent with the leader being rational. Proposition 1 characterizes commit decisions if a nonbabbling equilibrium exists.

**Proposition 1.** *In a symmetric nonbabbling equilibrium with $p \in (0,1]$ and $q \in (0,1]$, we have*

- *for $p \in (0,1)$, $\mathcal{E}^0 = \mathcal{E}^1 = S(p,q)$;*
- *for $p = 1$, $\mathcal{E}^1 = S(1,q)$ and $\mathcal{E}^0 = \{(n-f)q+f\}$.*

*Proof.* We start by showing that for $p < 1$ a rational backup commits if and only if her local information is consistent with the leader being rational, i.e., $k \in \mathcal{E}^z \iff \{z,k\} \in \mathcal{I}^R$.

The "only if" part, i.e., $k \in \mathcal{E}^z \implies \{z,k\} \in \mathcal{I}^R$, is an immediate outcome of Lemma 1: If a rational node $i$'s commit-stage information set is not consistent with a rational leader, i.e. $I_i \notin \mathcal{I}^R$,



Electronic copy available at: https://ssrn.com/abstract=3965955

then $i$ infers that the leader is definitely Byzantine, i.e., $\mathbb{P}(\overline{\mathcal{R}}|B, I_i \notin \mathcal{I}^R) = 1$. By Lemma 1, node $i$ does not commit, thus $\{z,k\} \notin \mathcal{I}^R \implies k \notin \mathcal{E}^z$ or equivalently, $k \in \mathcal{E}^z \implies \{z,k\} \in \mathcal{I}^R$.

We prove the "if" part by contradiction: for any $z = \{0,1\}$, we show that if there exists $g$ such that $\{z,g\} \in \mathcal{I}^R$ and $g \neq \mathcal{E}^z$, then $\mathcal{E}^z = \emptyset$.

Fix $z$. Suppose that there exists $g$ such $\{z,g\} \in \mathcal{I}^R$ and $g \neq \mathcal{E}^z$. Any rational node with a commit-stage information set $\{z,k\} \in \mathcal{I}^R$ knows that the leader can be either Byzantine or rational. If the leader is Byzantine, then by Lemma 1 committing to `message` yields utility $-c$. If the leader is rational, there exists a strategy for the Byzantine backup nodes such that a positive measure of rational nodes $j \neq i$ end up with $I_j = \{z,g\}$. For example, when all Byzantine nodes forward `messages` to $i$ with probability $b(k)$ and all other rational nodes with probability $b(g)$, where $(n-f)pq + b(k)\,pf = k$ and $(n-f)pq + b(g)\,pf = g$, then almost all rational nodes receive $g$ `messages`, and a positive measure of them will get $\{z,g\}$ and thus do not commit by assumption.

Denote $\hat{B}$ as a Byzantine strategy profile so that if the leader is Byzantine, $\hat{B} \in \mathcal{B}^z(k)$ and a positive measure of rational nodes receive $k < (n-f)pq$, and if the leader is rational, then a positive measure of rational nodes receive $g$ `messages`. In such a case, for any $I_i = \{z,k\} \in \mathcal{I}^R$ we have

$$\begin{aligned}
\min_{B\in\mathcal{B}} \mathbb{E}[u_i(commit, A^*_{-i}, B)|I_i] =& \min_{B\in\mathcal{B}} \left\{\mathbb{P}(\mathcal{R}|B, I_i)u_i(commit, A^*_{-i}; B; \mathcal{R}) + \mathbb{P}(\overline{\mathcal{R}}|B, I_i)u_i(commit, A^*_{-i}; B; \overline{\mathcal{R}})\right\} \\
\leq& \mathbb{P}(\mathcal{R}|\hat{B}, I_i) \underbrace{u_i(commit, A^*_{-i}; \hat{B}; \mathcal{R})}_{=-c} + \mathbb{P}(\overline{\mathcal{R}}|\hat{B}, I_i) \underbrace{u_i(commit, A^*_{-i}; \hat{B}; \overline{\mathcal{R}})}_{-c} \\
=& -c < 0.
\end{aligned}$$

When $p = 1$, the above proof logic directly applies for a node with $z = 1$. Those with $z = 0$ would infer the leader is Byzantine, and thus (i) does not commit if $k < (n-f)q + f$ (Lemma 1), or (ii) commit if $k = (n-f)q + f$, because she infers that all other rational nodes (other than a zero measure) have $\{z,k\} \in \mathcal{I}^1$ and thus commit. □

It is worth noting that although we have formulated the rational nodes' utility under ambiguity aversion based on the multiprior approach (Gilboa and Schmeidler (1993)), the key argument that



Electronic copy available at: https://ssrn.com/abstract=3965955

leads to our Proposition 1 only relies on the "worse-case scenario," rather than the expectation over potentially possible priors (which nests the consideration of the worse-case scenario only). In other words, we have shown that a rational node whose information set is *inconsistent* with a rational leader's strategy will see the possibility of "the leader being Byzantine" in Eq. (4), and hence does not commit to avoid the penalty of $-c$. This worst-case scenario argument is widely used in the computer science literature on BFT protocols.

# 4 Equilibrium Characterization

Section 3 has laid out the necessary structures of a symmetric nonbabbling equilibrium. This section further characterizes conditions under which symmetric nonbabbling equilibria indeed exist.

## 4.1 Bayesian Updating and Multiprior Expected Utilities

In Section 3, we have shown that a rational node $i$ with information set $I_i \notin \mathcal{I}^R$ infers that the leader is definitely Byzantine and thus always envisions a worst-case payoff $-c$ from committing to `message`. However, in a symmetric nonbabbling equilibrium with $p \in (0,1]$ and $q \in (0,1]$, a rational node $i$ with an information set $I_i \in \mathcal{I}^R$ may still see the leader as rational or Byzantine with positive probabilities. This section calculates such probabilities within a multiprior framework (Lemma 2), and characterizes a rational node's multiprior expected utility from committing `message` when she has information in $\mathcal{I}^R$ (Lemma 3). These results are the building blocks toward deriving conditions for a nonbabbling equilibrium to exist.

**Lemma 2.** *In a symmetric equilibrium with $p \in (0,1)$ and $q \in (0,1]$, a rational node $i$ with an information set in $\mathcal{I}^z$ has a posterior probability of the leader being rational given by*

$$\min_{B \in \mathcal{B}} \mathbb{P}(\mathcal{R}|B, \mathcal{I}^z) = \mathbb{P}(\mathcal{R}|B \in \mathcal{B}^z(k), \mathcal{I}^z) = \begin{cases} \frac{p(n-f)}{p(n-f)+f}, & \text{if } z = 1; \\ \frac{(1-p)(n-f)}{(1-p)(n-f)+f}, & \text{if } z = 0. \end{cases} \tag{8}$$

*Proof.* Suppose that in a symmetric perfect Bayesian equilibrium with $p \in (0,1]$ and $q \in (0,1]$, a rational node $i$'s commit-stage information set is in $\mathcal{I}^z$ for $z = \{0,1\}$.



Electronic copy available at: https://ssrn.com/abstract=3965955

Suppose $z = 1$. Notice that for any $B \in \mathcal{B}$

$$\begin{aligned}
\mathbb{P}(\mathcal{R}|B, \mathcal{I}^1) &= \frac{\mathbb{P}(\mathcal{I}^1|B, \mathcal{R})\mathbb{P}(\mathcal{R})}{\mathbb{P}(\mathcal{I}^1|B, \mathcal{R})\mathbb{P}(\mathcal{R}) + \mathbb{P}(\mathcal{I}^1|B, \overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&= \frac{p\mathbb{P}(\mathcal{R})}{p\mathbb{P}(\mathcal{R}) + \mathbb{P}(\mathcal{I}^1|B, \overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&= \frac{p(n-f)}{p(n-f) + \mathbb{P}(\mathcal{I}^1|B, \overline{\mathcal{R}})f} \geq \frac{p(n-f)}{p(n-f)+f}, \qquad (9)
\end{aligned}$$

where the last equality holds when $B \in \mathcal{B}^1(k)$. In contrast, when $z = 0$, we have for any $B \in \mathcal{B}$

$$\begin{aligned}
\mathbb{P}(\mathcal{R}|B, \mathcal{I}^0) &= \frac{\mathbb{P}(\mathcal{I}^0|B, \mathcal{R})\mathbb{P}(\mathcal{R})}{\mathbb{P}(\mathcal{I}^0|B, \mathcal{R})\mathbb{P}(\mathcal{R}) + \mathbb{P}(\mathcal{I}^0|B, \overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&= \frac{(1-p)\mathbb{P}(\mathcal{R})}{(1-p)\mathbb{P}(\mathcal{R}) + \mathbb{P}(\mathcal{I}^0|B, \overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&= \frac{(1-p)(n-f)}{(1-p)(n-f) + \mathbb{P}(\mathcal{I}^0|B, \overline{\mathcal{R}})f} \geq \frac{(1-p)(n-f)}{(1-p)(n-f)+f}, \qquad (10)
\end{aligned}$$

where the last equality holds when $B \in \mathcal{B}^0(k)$. □

**Lemma 3.** *In a symmetric nonbabbling equilibrium with $p \in (0, 1]$ and $q \in (0, 1]$, a rational node $i$ with an information set in $\mathcal{I}^R$ gets the following utility from committing to* `message`*:*

$$\min_{B\in\mathcal{B}} \mathbb{E}[u_i(commit, A^*_{-i}, B)|\mathcal{I}^z] = \min_{B\in\mathcal{B}}\{\mathbb{P}(\mathcal{R}|B, \mathcal{I}^z)\}R + \left(1 - \min_{B\in\mathcal{B}}\{\mathbb{P}(\mathcal{R}|B, \mathcal{I}^z)\}\right)(-c),$$

*except for when $p = 1$ and $k = (n-f)q + f$, in which case $\min_{B\in\mathcal{B}} \mathbb{E}[u_i(commit, A^*_{-i}, B)|\mathcal{I}^z] = R$.*

*Proof.* First, notice that if $\mathcal{E} \neq \emptyset$, then $\forall B \in \mathcal{B}, u_i(commit, A^*_{-i}; B; \mathcal{R}) = R$. This is because when the leader is rational, a rational node $i$ knows that in an equilibrium with $p \in (0, 1]$ and $q \in (0, 1]$, all rational nodes receive $\{z, k\} \in \mathcal{I}^R$ `messages` regardless of Byzantine backups' strategies. By Proposition 1 if a nonbabbling equilibrium exists, all rational nodes who receive $\{z, k\} \in \mathcal{I}^R$ commit to `message`. Thus, for $i$, committing to `message` yields $R$. Then,

$$\begin{aligned}
\mathbb{E}[u_i(commit, A^*_{-i}, B)|\mathcal{I}^z] &= \mathbb{P}(\mathcal{R}|B, \mathcal{I}^z)R + (1 - \mathbb{P}(\mathcal{R}|B, \mathcal{I}^z))u_i(commit, A^*_{-i}; B; \overline{\mathcal{R}}) \\
&= \mathbb{P}(\mathcal{R}|B, \mathcal{I}^z)\left(R - u_i(commit, A^*_{-i}; B; \overline{\mathcal{R}})\right) + u_i(commit, A^*_{-i}; B; \overline{\mathcal{R}}, )
\end{aligned}$$



Electronic copy available at: https://ssrn.com/abstract=3965955

$$
\begin{aligned}
&\geq \min_{B\in\mathcal{B}}\{\mathbb{P}(\mathcal{R}|B,\mathcal{I}^z)\}\left(R - u_i(commit, A^*_{-i}; B; \overline{\mathcal{R}})\right) + u_i(commit, A^*_{-i}; B; \overline{\mathcal{R}}) \\
&= \min_{B\in\mathcal{B}}\{\mathbb{P}(\mathcal{R}|B,\mathcal{I}^z)\}R + \left(1 - \min_{B\in\mathcal{B}}\{\mathbb{P}(\mathcal{R}|B,\mathcal{I}^z)\}\right) u_i(commit, A^*_{-i}; B; \overline{\mathcal{R}}) \\
&\geq \min_{B\in\mathcal{B}}\{\mathbb{P}(\mathcal{R}|B,\mathcal{I}^z)\}R + \left(1 - \min_{B\in\mathcal{B}}\{\mathbb{P}(\mathcal{R}|B,\mathcal{I}^z)\}\right) \underbrace{\min_{B\in\mathcal{B}}\left\{u_i(commit, A^*_{-i}; B; \overline{\mathcal{R}})\right\}}_{(*)}
\end{aligned}
$$

By Lemma 2, both inequalities obtain equality when $B \in \mathcal{B}^z(k)$. Furthermore, by Lemma 1, the term $(*)$ equals $-c$, except for when $p = 1$ and $I_i = \{0, (n-f)q + f\}$ or $\{1, (n-f)q + f\}$. In these exceptions, $(*)$ equals $R$. □

With the probabilities characterized in Lemma 2 and utilities from committing `message` stated in Lemma 3, we can pin down conditions under which a nonbabbling equilibrium exists.

### 4.2 Existence of Equilibria with Successful Consensus on `message`

A nonbabbling equilibrium exists if and only if the utility from committing is larger than utility from not committing when other nodes are playing nonbabbling strategies. In light of Proposition 1, we distinguish $p = 1$ and $p \in (0, 1)$.

**Proposition 2** (Existence when $p = 1$)**.** *There exists a symmetric nonbabbling equilibrium with $p = 1$ if and only if*

$$
\frac{f}{n}(-c) + \frac{n-f}{n}R \geq 0. \tag{11}
$$

*Proof.* To show the existence of an equilibrium, we will show that under condition (11), for any rational node $i$ if all other nodes $j \neq i$ commit to `message` if and only if they have information set in $\mathcal{I}^1$ or $\{z, k\} = \{0, (n-f)q + f\}$, then $i$ also finds it optimal to commit to `message` if and only if she has information set in $\mathcal{I}^1$ or $\{z, k\} = \{0, (n-f)q + f\}$.

Consider a rational node $i$ with commit-stage information set in $\mathcal{I}^1$. By Lemma 2 and Lemma 3, her utility from committing `message` if all other nodes commit to `message` is

$$
\min_{B\in\mathcal{B}} \mathbb{E}[u_i(commit, A^*_{-i}, B)|\mathcal{I}^1] = \frac{n-f}{n}\,R - \frac{f}{n}\,c.
$$



Electronic copy available at: https://ssrn.com/abstract=3965955

And $i$'s best response is to commit to `message` if and only if condition (11) holds. For $\{z,k\} = \{0,(n-f)q+f\}$, the expected utility from committing is $R$. But since a positive measure of rational nodes have information set in $\mathcal{I}^1$, node $i$ does not commit unless condition (11) holds. □

**Proposition 3** (Existence when $p \in (0,1)$)**.** *There exists a symmetric nonbabbling equilibrium with $p \in (0,1)$ if and only if*

$$\begin{cases} \frac{f}{p(n-f)+f}(-c) + \frac{p(n-f)}{p(n-f)+f}R \geq 0, \\ \frac{f}{(1-p)(n-f)+f}(-c) + \frac{(1-p)(n-f)}{(1-p)(n-f)+f}R \geq 0. \end{cases} \tag{12}$$

*Proof.* To show the existence of an equilibrium, we show that under condition (12), for any rational node $i$ if all other nodes $j \neq i$ commit to `message` if and only if they have information set in $\mathcal{I}^0$ or $\mathcal{I}^1$, then $i$ also finds it optimal to commit to `message` if and only if she has information set in $\mathcal{I}^0$ or $\mathcal{I}^1$.

Suppose that all other rational nodes $j \neq i$ commit to `message` when they have an information set in $\mathcal{I}^0$ or $\mathcal{I}^1$. Then for a rational node $i$ with commit-stage information set $\mathcal{I}^0$, by Lemma 2 and 3, the utility from committing to `message` is

$$\min_{B\in\mathcal{B}} \mathbb{E}[u_i(commit, A^*_{-i}, B)|\mathcal{I}^0] = \frac{(1-p)(n-f)}{(1-p)(n-f)+f}\,R - \frac{f}{(1-p)(n-f)+f}\,c\,. \tag{13}$$

Similarly, for a rational node $i$ with commit-stage information set $\mathcal{I}^1$,

$$\min_{B\in\mathcal{B}} \mathbb{E}[u_i(commit, A^*_{-i}, B)|\mathcal{I}^1] = \frac{p(n-f)}{p(n-f)+f}\,R - \frac{f}{p(n-f)+f}\,c\,. \tag{14}$$

Both (13) and (14) are positive if and only if condition (12) holds. □

### 4.3 A Complete Equilibria Characterization

Looking back at Definition 2, so far we have focused on characterizing the commit strategies $\tilde{C}$ by characterizing $\mathcal{E}^0$ and $\mathcal{E}^1$ for given $p$ and $q$. To complete the equilibrium characterization, we need to also identify which $p$ and $q$ can constitute an equilibrium. We are especially interested in



Electronic copy available at: https://ssrn.com/abstract=3965955

nonbabbling equilibria, where $\mathcal{E} \neq \emptyset$.

The strategies $p$ and $q$ are decided by the nodes knowing how they could impact the number of `messages` sent and the commit strategies afterwards. For $p = 1$, any $q \in (0, 1]$ constitutes a nonbabbling equilibrium when the existence condition in Proposition 2 is satisfied. Neither the leader, nor the backups have incentive to deviate. If backups expect $p = 1$, and a rational node chosen as the leader deviates to lower $p_i < 1$, a positive measure of backups would end up with $z = 0$, and not commit. Thus the leader's payoff would be strictly lower than $R$. Since each backup is of measure 0, the deviation from $q$ would not impact anyone's utilities, including his own. Thus, a profitable deviation is not possible.

By similar reasoning, for any $p \in (0, 1)$ satisfying the existence conditions in Proposition 3, there is a nonbabbling equilibrium for any $q \in (0, 1]$. Neither the leader, nor the backups have incentive to deviate from $p$ and $q$. The leader could end up spoiling consensus by deviating from $p$, and a backup's deviation would not have an impact on anyone's utilities.

While in our analysis in the previous sections we focused on backups' committing decisions, the leader also commits when $k \in S(p, q)$. Being a leader means that $z = 1$.

Based on the above analysis, we obtain the following result.

**Theorem 1.** *With set $\mathcal{S}(p, q)$ defined as $[(n-f)pq, (n-f)pq+fp]$, we have the following complete characterization of all symmetric equilibria.*

1. *A "babbling" equilibrium always exists, in which nodes never commit regardless of the communication. That is, $p \in (0, 1]$, $q \in (0, 1]$ and $\mathcal{E} = \emptyset$.*

2. *Fractional-p-equilibria exist when $\frac{1}{2}(n-f)R \geq fc$. In this continuum of equilibria, a rational leader sends* `message` *to each backup with probability $p \in \left[\frac{fc}{(n-f)R}, 1 - \frac{fc}{(n-f)R}\right]$, a rational backup forwards* `message` *(if received) with probability $q \in (0, 1]$, and a backup commits if and only if receiving $k \in \mathcal{S}(p, q)$ messages, regardless of whether receiving from the leader. That is, $\mathcal{E}^0 = \mathcal{E}^1 = \mathcal{S}(p, q)$.*

3. *Unitary-p-equilibria exist when $(n - f)R \geq fc$. In this continuum of equilibria, a rational leader sends* `message` *to each backup with $p = 1$, a rational backup forwards* `message` *(if*



Electronic copy available at: https://ssrn.com/abstract=3965955

*received) with probability* $q \in (0,1]$*, and a backup commits if and only if receiving* $k \in \mathcal{S}(p=1,q)$ `messages`*, with one from the leader or* $(n-f)q+f$ `messages` *without any from the leader. That is,* $\mathcal{E}^0 = \{(n-f)q+f\}$ *and* $\mathcal{E}^1 = \mathcal{S}(1,q)$*.*

Note that as $c \to +\infty$, only the babbling equilibrium survives, consistent with the computer science literature (which restores nonbabbling equilibria by allowing additional views). In addition, unitary-$p$-equilibria are "knife-edge" ones in that if rational nodes' `messages` are only delivered with some probability $\alpha < 1$, then this equilibrium will be eliminated—this case will be discussed in the extensions below.

The condition $\frac{1}{2}(n-f)R \geq fc$ in case 2 is a result of requiring rational backups who receive `messages` within $\mathcal{S}(p,q)$, including both those who receive `message` directly from the leader and those who do not, to commit. We will see this condition again in later sections.

### 4.4 Discussion of Economic Implications

We provide two discussions on the economic implications of Theorem 1.

**Economic Incentives and Connection to the Computer Science Literature** Our analysis shows that BFT problems with explicit incentive considerations tend to accommodate multiple equilibria. In particular, there always exists a "babbling" equilibrium in which non-Byzantine nodes always discard all preplay communications and do not commit to any messages. On the other hand, the traditional computer science literature on BFT protocols typically requires a *safety* condition that effectively stipulates consensus success on `message` as a unique equilibrium outcome. This is because these papers do not explicitly consider nodes' incentives: By forcing nodes to behave honestly, the babbling equilibrium is artificially ruled out by a "no-triviality" assumption. Relatedly, the level of $R$ compared to other variables affects the number of possible equilibria. All else equal, the higher $R$ is, the more equilibria there are. For protocol designers, $R$ should be chosen to be not too low to ensure (nonbabbling) equilibrium existence. Finally, in our framework, a successful consensus on `message` crucially relies on whether the leader is Byzantine or not, since a Byzantine leader can always disrupt consensus on `message` (that is, consensus on `message` definitely



Electronic copy available at: https://ssrn.com/abstract=3965955

fails if the leader is known to be Byzantine), while on the other hand, a non-Byzantine leader can always ensure a successful consensus on `message`.

**The Role of Peer Communication** Peer communications among rational backups after the leader's messaging stage help rational backups make more informed commit decisions. To see this, consider a simpler game in which backups have to immediately make a commit decision after the leader's messaging stage, rather than waiting until after another round of peer communications. This is as if we assume that $q = 0$.

It is easy to verify that when $(n - f)R \geq fc$, this simpler game has the following unique nonbabbling symmetric equilibrium: A rational leader chooses $p = 1$, and all rational backups who receive `message` from the leader immediately commit, while those who do not receive `message` from the leader immediately choose not to commit.[13]

In this simpler game, a Byzantine leader who sends `message` to a rational backup (but not all other backups) always "tricks" this backup into a bad commit decision, while in our full model with peer communications, this "wrong" decision may be avoided if the rational backup does not receive the appropriate number of forwarded `messages` from her peers. Therefore, the peer-communication stage increases the ex post payoff to rational nodes.[14]

That said, since in our model, a rational leader can ensure a successful consensus on `message` and a Byzantine one (together with Byzantine backups) can cause a failed consensus on `message` as a worst-case outcome, both games would lead to the same ex ante total surplus to rational nodes.

The identical ex ante and higher ex post total payoff from peer communication do not conflict with each other, as rational nodes are ambiguity averse so that their ex ante surpluses always focus on the worst-case outcome while ex post payoffs concern all cases (not necessarily the worst one). This ex ante welfare equivalence, however, relies on the rational leader being able to ensure consensus by choosing $p = 1$. As a result, it is not robust to the possibility of message losses; in

[13] If $p < 1$, then there exists no nonbabbling equilibria, because a positive measure of rational nodes get no `message` and cannot commit.

[14] To see this, a rational node who does not commit always gets 0, regardless of whether peer communication is allowed. If a rational node gets $R$ from committing when peer communication is disallowed, then the leader (either rational of Byzantine) must have sent `message` to all rational nodes, then given the same Byzantine strategy profile, when peer communication is allowed, all rational nodes would receive $k$ `messages` with $k \in \mathcal{S}(p, q)$, and they will also get $R$ from committing.



Electronic copy available at: https://ssrn.com/abstract=3965955

such a scenario, peer-communication always helps, as shown in Section 5.1.

# 5 Extensions: Introducing Message Losses

Our discussions so far have assumed that all `messages` sent will be delivered with certainty. However, in practice, a central issue in the design of distributed consensus systems is the possibility of `messages` lost in the delivery process, reflecting certain technological constraints.[15] In this section, we first study the case in which each `message` sent may only be delivered with some probability but in an idiosyncratic way; we then further allow systematic risk over the random `message` deliveries.

## 5.1 Idiosyncratic `message` Losses

Suppose that all `messages` sent are delivered probabilistically, following an identical and independent (binary) distribution with a fixed probability $\alpha \in (0, 1)$. As before we consider a candidate symmetric equilibrium in which a rational leader sends `message` to each backup with probability $p$ and each rational backup forwards `message` (if received) with probability $q$.

Based on the earlier definition of $S(p, q)$, we have

$$\mathcal{S}(p\alpha^2, q) = [(n-f)qp\alpha^2, (n-f)qp\alpha^2 + fp\alpha^2]. \tag{15}$$

Conditional on the leader being rational, a rational backup receives the leader's `message` with probability $p\alpha$. Regardless of whether the leader's `message` was received, any rational backup expects to receive $k \in \mathcal{S}(p\alpha^2, q)$ `messages` from other backups. Here, $\alpha^2$ captures the fact that `message` loss could occur when the leader sends the `message` as well as when backups forward the `message` (see Figure 1); and we have used the law of large numbers given idiosyncratic `message` losses.

[15] The assumption of all `messages` sent being delivered within a fixed time is what typically known in the computer science literature as the *synchronous* network assumption. Many BFT protocols used in practice often assume a weaker assumption of *partial synchrony*, which does not explicitly allow `messages` to be lost, but only arbitrarily delayed. That said, in practical implementations such protocols are designed to proceed differently depending on whether `messages` are delivered or not within some preset time limits.



Electronic copy available at: https://ssrn.com/abstract=3965955

**Inferences and Bayesian Updating** The potential `message` loss caused by technological conditions affects rational backups inferences. As in Eq. (10) and (9) in the proof of Lemma 2, any rational backup who receives $k \in \mathcal{S}(p\alpha^2, q)$ `messages` but misses the leader's ($z = 0$) infers that the leader is rational with a posterior probability of

$$\begin{aligned}
\mathbb{P}(\mathcal{R}|\mathcal{I}^0) &= \frac{\mathbb{P}(\mathcal{I}^0|\mathcal{R})\mathbb{P}(\mathcal{R})}{\mathbb{P}(\mathcal{I}^0|\mathcal{R})\mathbb{P}(\mathcal{R}) + \mathbb{P}(\mathcal{I}^0|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&= \frac{(1-p\alpha)\mathbb{P}(\mathcal{R})}{(1-p\alpha)\mathbb{P}(\mathcal{R}) + \mathbb{P}(\mathcal{I}^0|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&\geq \frac{(1-p\alpha)\mathbb{P}(\mathcal{R})}{(1-p\alpha)\mathbb{P}(\mathcal{R}) + \mathbb{P}(\overline{\mathcal{R}})} = \frac{(1-p\alpha)(n-f)}{(1-p\alpha)(n-f)+f},
\end{aligned} \tag{16}$$

while a rational backup who receives $k \in \mathcal{S}(p\alpha^2, q)$ `messages` with $z = 1$ infers that the leader is rational with a posterior probability of

$$\begin{aligned}
\mathbb{P}(\mathcal{R}|\mathcal{I}^1) &= \frac{\mathbb{P}(\mathcal{I}^1|\mathcal{R})\mathbb{P}(\mathcal{R})}{\mathbb{P}(\mathcal{I}^1|\mathcal{R})\mathbb{P}(\mathcal{R}) + \mathbb{P}(\mathcal{I}^1|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&= \frac{p\alpha\mathbb{P}(\mathcal{R})}{p\alpha\mathbb{P}(\mathcal{R}) + \mathbb{P}(\mathcal{I}^1|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&\geq \frac{p\alpha\mathbb{P}(\mathcal{R})}{p\alpha\mathbb{P}(\mathcal{R}) + \mathbb{P}(\overline{\mathcal{R}})} = \frac{p\alpha(n-f)}{p\alpha(n-f)+f}.
\end{aligned} \tag{17}$$

**Committing Decisions and Equilibra Characterization** Consensus on `message` requires unanimous commit from all rational nodes.[16] When the leader is rational, although all rational backups receive a number of `messages` within the interval $\mathcal{S}(p\alpha^2, q)$, potential `message` losses imply that only a fraction of them receive `message` from the leader ($\mathcal{I}^1$) while the others do not ($\mathcal{I}^0$). Hence, all rational backups will commit only when both conditions (18) and (19) are satisfied:

$$\left(\frac{(1-p\alpha)(n-f)}{(1-p\alpha)(n-f)+f}\right)\cdot R \geq \left(1-\left(\frac{(1-p\alpha)(n-f)}{(1-p\alpha)(n-f)+f}\right)\right)\cdot c \tag{18}$$

$$\left(\frac{p\alpha(n-f)}{p\alpha(n-f)+f}\right)\cdot R \geq \left(1-\frac{p\alpha(n-f)}{p\alpha(n-f)+f}\right)\cdot c. \tag{19}$$

[16] More precisely, rational nodes who do not commit are of measure zero.



Electronic copy available at: https://ssrn.com/abstract=3965955

That is, we require

$$\frac{R}{c} \geq \frac{f}{n-f} \cdot \max\left\{\frac{1}{p\alpha}, \frac{1}{1-p\alpha}\right\}. \tag{20}$$

The next theorem, which parallels with Theorem 1, summarizes our result in this section.

**Theorem 2.** *Facing idiosyncratic risks of* `messages` *not being delivered—that is, all* `messages` *sent are delivered with probability $\alpha < 1$, we have the following characterization of all symmetric equilibria.*

1. *A "babbling" equilibrium always exists, in which nodes never commit regardless of the communication. That is, $\mathcal{E} = \emptyset$.*

2. *Fractional-p-equilibria exist when $(n-f)R \geq \max\left\{2, \frac{1}{\alpha}\right\} \cdot fc$. In this continuum of equilibria, a rational leader sends* `message` *to each backup with probability*

$$p \in \left[\frac{1}{\alpha}\frac{fc}{(n-f)R}, \frac{1}{\alpha}\left(1 - \frac{fc}{(n-f)R}\right)\right] \cap [0,1], \tag{21}$$

*a rational backup forwards* `message` *(if received) with probability $q$, and a rational backup commits if and only if it receives $k \in \mathcal{S}(p\alpha^2, q)$* `messages`*, regardless of whether it receives anything from the leader. That is, $\mathcal{E}^0 = \mathcal{E}^1 = \mathcal{S}(p\alpha^2, q)$.*

There are two key differences between the equilibria with idiosyncratic `message` losses (Theorem 2) and the equilibria without (Theorem 1). First, as expected, the interval-$\mathcal{E}^0$ equilibria in both theorems are the same except with $p$ replaced by $p\alpha$. Intuitively, the effective `message` delivery probability is the product of the strategic `message` delivery probability ($p$) and technological `message` delivery probability ($\alpha$, which takes a value of 1 in our baseline).

Second, which perhaps has greater economic content, Theorem 2 reveals that Case 3 (unitary-$p$-equilibria) in Theorem 1 is a nongeneric "knife-edge" case. For every rational node to commit, this class of equilibria requires them to not only send/forward but also always receive these `messages`. Theorem 2 establishes that these equilibria do not survive when we perturb the system to have $(1-\alpha)$-chance of `message` delivery failure.



Electronic copy available at: https://ssrn.com/abstract=3965955

Because the unitary-$p$-equilibria are nongeneric, from now on our analysis focuses on fractional-$p$-equilibria, which are Case 2 in both Theorem 1 and 2.

**Welfare Analysis** We now move on to study welfare in this system. Given the nature of equilibria multiplicity, we assume that the endogenous strategies on `message` sending and forwarding ($p$ and $q$) can be selected by the planner who aims to maximize the expected welfare. This can be implemented with a pregame stage in coordinating all rational nodes to play the best equilibrium.

We measure welfare by (expected) successful consensus on `message` from the perspective of a planner with similar preferences as rational nodes (i.e., ambiguity-averse to Byzantine behaviors). More specifically, the planner solves the following problem:

$$W \equiv \max_{p\in\left[\frac{1}{\alpha}\frac{fc}{(n-f)R},\frac{1}{\alpha}\left(1-\frac{fc}{(n-f)R}\right)\right]\cap[0,1]} \underbrace{(n-f)}_{\#\text{rational nodes}} \underbrace{\left(\frac{n-f}{n}R+\frac{f}{n}(-c)\right)}_{\text{expected payoff from committing}} \underbrace{\mathbb{1}_{\frac{R}{c}\geq\max\{\frac{f}{p\alpha(n-f)},\frac{f}{(1-p\alpha)(n-f)}\}}}_{\text{if commits}}. \tag{22}$$

An alternative welfare $V$ captures whether the system could reach consensus or not:

$$V \equiv \max_{p\in\left[\frac{1}{\alpha}\frac{fc}{(n-f)R},\frac{1}{\alpha}\left(1-\frac{fc}{(n-f)R}\right)\right]\cap[0,1]} \mathbb{1}_{\frac{R}{c}\geq\max\{\frac{f}{p\alpha(n-f)},\frac{f}{(1-p\alpha)(n-f)}\}}. \tag{23}$$

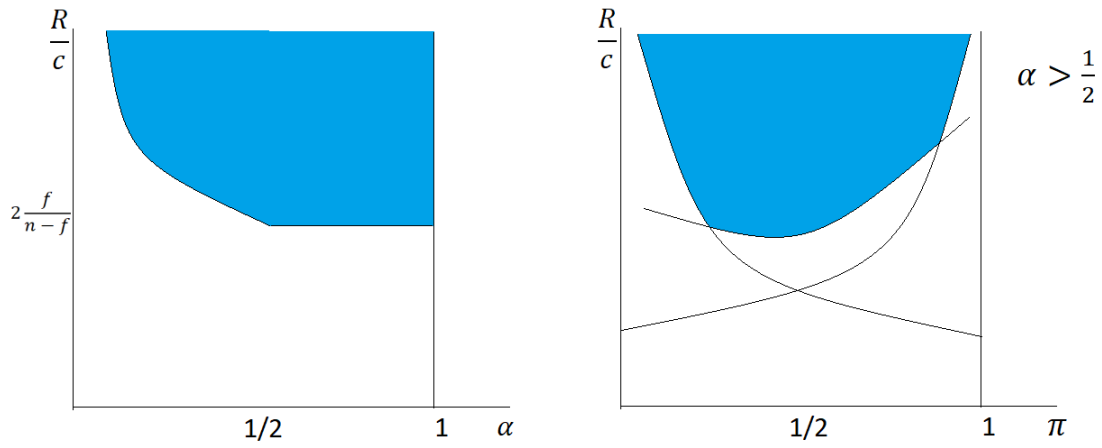
Problem (23) and problem (22) share the same solution when we view welfare as a function of $\alpha$. However, as the planner may attach an arbitrary surplus to the consensus, the objective in (23) potentially permits broader interpretations: for instance, the system's safety may serve other purposes with other significant social value (say payment); and some key parameters $R$ or $c$ might be viewed as transfers, and hence part of them should not be counted in welfare.

The solution to problem (23) is given as follows:

- If $\alpha \geq \frac{1}{2}$, the welfare-maximizing equilibrium has $p$ such that $p\alpha = \frac{1}{2}$. In this case, welfare is invariant with $\alpha$.
- If $\alpha < \frac{1}{2}$, the welfare-maximizing equilibrium has $p$ such that $p = 1$. In this case, welfare is increasing in $\alpha$.



Electronic copy available at: https://ssrn.com/abstract=3965955




Panel A Panel B

Figure 3: This figure illustrates $V$ in the parameter space of $R/c$, with solid area taking a value of 1. Panel A is with respect to $\alpha$ with idiosyncratic `message` losses, while Panel B is with respect to $\pi$ with systematic `message` losses (for a given level of $\alpha$). The latter will be discussed later in Section 5.2.

Panel A of Figure 3 illustrates the objective $V$ in (22), with the solid area taking a value of 1, in the parameter space of $R/c$ and $\alpha$. We observe that better communication technology (a higher $\alpha$) improves the chance of reaching consensus in the system. As we will show shortly, this is in contrast to the case of systematic risk of `message` losses.

**Further Comment on the Role of Peer Communication** The welfare analysis also demonstrates further why potential `message` losses necessitate peer communications. As we have pointed out toward the end of Section 4.4, the ex ante total surplus to all rational backups in our baseline model ($\alpha = 1$) is identical to that in a simpler game without peer communications. This result, however, is not robust when $\alpha < 1$. In this case, backups have to make a commit decision immediately upon receiving (or not) `message` from the leader, so consensus on `message` will always fail: Those who do not receive `message` from the leader will not commit, while those who do receive `message` from the leader, recognizing a positive measure of rational backups not committing, will also choose to not commit. By allowing one additional round of communication among rational backups, they are given the ability to make more informed commit decisions, and as result are more likely to reach a successful consensus on `message`.



Electronic copy available at: https://ssrn.com/abstract=3965955

### 5.2 Systematic Risk of `message` Losses

Whether the potential `message` losses are idiosyncratic or systematic plays a significant role in determining the consensus game's equilibria.

Consider the following extension that features two aggregate states of the world. With probability $\pi \in (0,1)$ the state is "good," so that all `messages` sent will be delivered; this state corresponds to the world after *global stabilization time (GST)* in the terminology of the computer science literature. Otherwise, with probability $1-\pi$ a "bad" state with *network congestion* occurs, in which all `messages` sent will be delivered with probability $\alpha \in \left[\sqrt{\frac{n-f}{n}}, 1\right)$.[17] We are interested in how the probability $\pi$—the good GST state—affects the ex ante welfare.

**Inferences and Bayesian Updating** We use $\mathcal{G}$ to denote the event of the good state occurring. Naturally, $\mathcal{G}\&\mathcal{R}$ denotes the event of the state being GST and the leader being rational; in this event, any rational backup could receive a `message` from the leader (with probability $p$) or not (with probability $1-p$), but she should receive $k \in \mathcal{S}(p,q)$ `messages`. Similarly, denote by $\overline{\mathcal{G}}\&\mathcal{R}$ the event of the state being non-GST and the leader being rational. In this case, any rational backup could receive a `message` from the leader (with probability $p\alpha$) or not (with probability $1-p\alpha$), but she should receive $k \in \mathcal{S}(p\alpha^2, q)$ `messages`.

There are other two possible underlying events with a Byzantine leader: $\mathcal{G}\&\overline{\mathcal{R}}$ denotes the state being GST and the leader being Byzantine, in which any rational backup could receive $k \in [0, (n-f)q+f]$ `messages` together with a leader's `message` with any probability, while $\overline{\mathcal{G}}\&\overline{\mathcal{R}}$ denotes the non-GST state and the leader being Byzantine, in which any rational backup could receive $k \in \left[0, (n-f)q\alpha^2 + f\alpha^2\right]$ `messages` and the leader's `message` with probability in $[0,\alpha]$. In the following analysis we combine these two payoff-equivalent events as event $\overline{\mathcal{R}}$ (leader being Byzantine), as in both events the rational backup's payoff is always $-c$.

Table 1 applies Bayes's rule and calculates a rational backup's posterior probability over the above three states ($\mathcal{G}\&\mathcal{R}$, $\overline{\mathcal{G}}\&\mathcal{R}$, and $\overline{\mathcal{R}}$), conditional on $k$ and $z$. Two points are worth noting. First, since the three events $\mathcal{G}\&\mathcal{R}$, $\overline{\mathcal{G}}\&\mathcal{R}$, and $\overline{\mathcal{R}}$ are mutually exclusive, and at least one of them happens,

[17] As we will see later, this condition ensures that the two states are not easily distinguishable.



Electronic copy available at: https://ssrn.com/abstract=3965955

the posterior probability of $\overline{\mathcal{R}}$ is immediately obtained as one minus the posterior probabilities of $\mathcal{G}\&\mathcal{R}$ and $\overline{\mathcal{G}}\&\mathcal{R}$. Second, the equilibrium sets $\mathcal{E}$ under two different underlying states (i.e., $\mathcal{S}(p,q)$ under GST and $\mathcal{S}(p\alpha^2,q)$ under non-GST) differ but overlap; this yields three partitions $\mathcal{S}(p\alpha^2,q)\backslash\mathcal{S}(p,q)$, $\mathcal{S}(p\alpha^2,q)\cap\mathcal{S}(p,q)$, and $\mathcal{S}(p,q)\backslash\mathcal{S}(p\alpha^2,q)$ in Table 1.

Consider Case 1 first. When $k\in\mathcal{S}(p\alpha^2,q)\backslash\mathcal{S}(p,q)=[(n-f)pq\alpha^2,(n-f)pq)$, any rational backup rules out the event $\mathcal{G}\&\mathcal{R}$: if it were $\mathcal{G}\&\mathcal{R}$, she should receive $k\in\mathcal{S}(p,q)$, but Case 1 falls strictly outside of the interval $\mathcal{S}(p,q)$. For the event of $\overline{\mathcal{G}}\&\mathcal{R}$, i.e., non-GST and rational leader, the rational backup with $z=0$ forms a posterior probability of

$$\begin{aligned}\mathbb{P}(\overline{\mathcal{G}}\&\mathcal{R}|\mathcal{I}^0) &= \frac{\mathbb{P}(\mathcal{I}^0|\overline{\mathcal{G}}\&\mathcal{R})\mathbb{P}(\overline{\mathcal{G}})\mathbb{P}(\mathcal{R})}{\mathbb{P}(\mathcal{I}^0|\overline{\mathcal{G}}\&\mathcal{R})\mathbb{P}(\overline{\mathcal{G}})\mathbb{P}(\mathcal{R})+\mathbb{P}(\mathcal{I}^0|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})}=\frac{(1-p\alpha)(1-\pi)\mathbb{P}(\mathcal{R})}{(1-p\alpha)(1-\pi)\mathbb{P}(\mathcal{R})+\mathbb{P}(\mathcal{I}^0|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})}\\ &\geq \frac{(1-p\alpha)(1-\pi)\mathbb{P}(\mathcal{R})}{(1-p\alpha)(1-\pi)\mathbb{P}(\mathcal{R})+\mathbb{P}(\overline{\mathcal{R}})}=\frac{(1-p\alpha)(1-\pi)(n-f)}{(1-p\alpha)(1-\pi)(n-f)+f},\end{aligned}\tag{24}$$

where we used $\mathbb{P}(\mathcal{I}^0|\overline{\mathcal{G}}\&\mathcal{R})=1-p\alpha$ and $\mathbb{P}(\mathcal{I}^0|\overline{\mathcal{R}})\leq 1$.[18] A similar calculation applies to Case 1′:

$$\mathbb{P}(\overline{\mathcal{G}}\&\mathcal{R}|\mathcal{I}^1) = \frac{\mathbb{P}(\mathcal{I}^1|\overline{\mathcal{G}}\&\mathcal{R})\mathbb{P}(\overline{\mathcal{G}}\&\mathcal{R})}{\mathbb{P}(\mathcal{I}^1|\overline{\mathcal{G}}\&\mathcal{R})\mathbb{P}(\overline{\mathcal{G}}\&\mathcal{R})+\mathbb{P}(\mathcal{I}^1|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})}\geq\frac{p\alpha(1-\pi)(n-f)}{p\alpha(1-\pi)(n-f)+f}.\tag{25}$$

Case 3 and 3′ with $k\in\mathcal{S}(p,q)\backslash\mathcal{S}(p\alpha^2,q)$ follow similarly. (See derivations in Appendix.)

In Case 2, $k\in\mathcal{S}(p\alpha^2,q)\cap\mathcal{S}(p,q)=[(n-f)pq,(n-f)pq\alpha^2+fp\alpha^2]$. A rational backup with $z=0$ infers that both $\mathcal{G}\&\mathcal{R}$ and $\overline{\mathcal{G}}\&\mathcal{R}$ are possible. For the event $\overline{\mathcal{G}}\&\mathcal{R}$, the calculation is identical to (24), while the posterior for $\mathcal{G}\&\mathcal{R}$ is:

$$\begin{aligned}\mathbb{P}(\mathcal{G}\&\mathcal{R}|\mathcal{I}^0) &= \frac{\mathbb{P}(\mathcal{I}^0|\mathcal{G}\&\mathcal{R})\mathbb{P}(\mathcal{G}\&\mathcal{R})}{\mathbb{P}(\mathcal{I}^0|\mathcal{G}\&\mathcal{R})\mathbb{P}(\mathcal{G}\&\mathcal{R})+\mathbb{P}(\mathcal{I}^0|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})}=\frac{(1-p)\pi\mathbb{P}(\mathcal{R})}{(1-p)\pi\mathbb{P}(\mathcal{R})+\mathbb{P}(\mathcal{I}^0|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})}\\ &\geq \frac{(1-p)\pi\mathbb{P}(\mathcal{R})}{(1-p)\pi\mathbb{P}(\mathcal{R})+\mathbb{P}(\overline{\mathcal{R}})}=\frac{(1-p)\pi(n-f)}{(1-p)\pi(n-f)+f}.\end{aligned}\tag{26}$$

Similarly, we have for Case 2′ that

$$\mathbb{P}(\mathcal{G}\&\mathcal{R}|\mathcal{I}^1) = \frac{\mathbb{P}(\mathcal{I}^1|\mathcal{G}\&\mathcal{R})\mathbb{P}(\mathcal{G}\&\mathcal{R})}{\mathbb{P}(\mathcal{I}^1|\mathcal{G}\&\mathcal{R})\mathbb{P}(\mathcal{G}\&\mathcal{R})+\mathbb{P}(\mathcal{I}^1|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})}\geq\frac{p\pi(n-f)}{p\pi(n-f)+f}.\tag{27}$$

[18] The equality holds when the leader is Byzantine and Byzantine nodes mimic the situation under a rational leader.



Electronic copy available at: https://ssrn.com/abstract=3965955

| | #`messages` received $k \in$ | leader's `message`? $z =$ | $\mathcal{G}\&\mathcal{R}$ | $\overline{\mathcal{G}}\&\mathcal{R}$ |
|---|---|---|---|---|
| 0 | $[0, (n-f)pq\alpha^2)$ | 0 | 0 | 0 |
| $0'$ | $[0, (n-f)pq\alpha^2)$ | 1 | 0 | 0 |
| 1 | $[(n-f)pq\alpha^2, (n-f)pq)$ | 0 | 0 | $\geq \frac{(1-p\alpha)(1-\pi)(n-f)}{(1-p\alpha)(1-\pi)(n-f)+f}$ |
| $1'$ | $[(n-f)pq\alpha^2, (n-f)pq)$ | 1 | 0 | $\geq \frac{p\alpha(1-\pi)(n-f)}{p\alpha(1-\pi)(n-f)+f}$ |
| 2 | $[(n-f)pq, (n-f)pq\alpha^2 + fp\alpha^2]$ | 0 | $\geq \frac{(1-p)\pi(n-f)}{(1-p)\pi(n-f)+f}$ | $\geq \frac{(1-p\alpha)(1-\pi)(n-f)}{(1-p\alpha)(1-\pi)(n-f)+f}$ |
| $2'$ | $[(n-f)pq, (n-f)pq\alpha^2 + fp\alpha^2]$ | 1 | $\geq \frac{p\pi(n-f)}{p\pi(n-f)+f}$ | $\geq \frac{p\alpha(1-\pi)(n-f)}{p\alpha(1-\pi)(n-f)+f}$ |
| 3 | $((n-f)pq\alpha^2 + fp\alpha^2, (n-f)pq + fp]$ | 0 | $\geq \frac{(1-p)\pi(n-f)}{(1-p)\pi(n-f)+f}$ | 0 |
| $3'$ | $((n-f)pq\alpha^2 + fp\alpha^2, (n-f)pq + fp]$ | 1 | $\geq \frac{p\pi(n-f)}{p\pi(n-f)+f}$ | 0 |
| $0''$ | $((n-f)pq + fp, (n-f)q + f]$ | 0 | 0 | 0 |
| $0'''$ | $((n-f)pq + fp, (n-f)q + f]$ | 1 | 0 | 0 |

Table 1: Posterior probabilities with systematic risk

This table summarizes a rational backup's posterior probability of $\mathcal{G}\&\mathcal{R}$ and $\overline{\mathcal{G}}\&\mathcal{R}$ conditional on how many `messages` she receives and if she receives `message` from the leader. Intervals increase from upper rows to lower ones.

**Commit Decisions and Equilibra Characterization** With posterior probabilities in Table 1 we study rational nodes' commit decisions. We have the following four cases:

- When receiving $k \in [0, (n-f)pq\alpha^2)$ or $k \in ((n-f)pq + fp, (n-f)q + f]$ `messages` (case 0, $0'$, $0''$ and $0'''$ or $k \in [0, n] \backslash (\mathcal{S}(p\alpha^2, q) \cup \mathcal{S}(p, q))$), do not commit;

- When receiving $k \in \mathcal{S}(p\alpha^2, q) \backslash \mathcal{S}(p, q) = [(n-f)pq\alpha^2, (n-f)pq]$ `messages` (case 1 and $1'$), commit if those in case 1, $1'$, 2, and $2'$ also commit;

- When receiving $k \in \mathcal{S}(p, q) \backslash \mathcal{S}(p\alpha^2, q) = [(n-f)pq\alpha^2 + fp\alpha^2, (n-f)pq + fp]$ `messages` (case 3 and $3'$), commit if those in case 2, $2'$, 3, and $3'$ also commit;

- When receiving $k \in \mathcal{S}(p\alpha^2, q) \cap \mathcal{S}(p, q) = [(n-f)pq, (n-f)pq\alpha^2 + fp\alpha^2]$ `messages` (case 2 and $2'$), commit if those in all cases from 1 to $3'$ also commit.

To understand these results, consider a rational node in cases $\{1, 1'\} = \mathcal{S}(p\alpha^2, q) \backslash \mathcal{S}(p, q)$. She infers that the system is in the non-GST state and hence other rational backups could be facing $\{1, 1', 2, 2'\}$. Similarly, a rational node who find herself in cases $\{3, 3'\} = \mathcal{S}(p, q) \backslash \mathcal{S}(p\alpha^2, q)$ knows



Electronic copy available at: https://ssrn.com/abstract=3965955

that GST definitely occurs and other rational backups could be facing $\{2, 2', 3, 3'\}$. Finally, a rational node in cases $\{2, 2'\} = \mathcal{S}(p, q) \cap \mathcal{S}(p\alpha^2, q)$ observes that both GST and non-GST are possible and therefore her fellow rational nodes could face all possibilities $\{1, 1', 2, 2', 3, 3'\}$.

Recall that consensus requires unanimous commit to `message` from all rational nodes; this implies that nonbabbling equilibrium requires rational nodes to commit in *all* cases. To see this, consider a rational node in Case 1; she plays commit only if she knows that rational nodes in Case 2 commit; but this in turn requires that rational nodes in all cases to commit, as explained above. We highlight that this observation captures the idea of local knowledge and higher order beliefs, a reasoning that is reminiscent of the key logic in global games (Morris and Shin (2003)).

As a result, any nonbabbling symmetric equilibrium requires that

$$\mathbb{P}_{\text{posterior}} \cdot R \geq (1 - \mathbb{P}_{\text{posterior}}) \cdot c \tag{28}$$

always holds, where $\mathbb{P}_{\text{posterior}}$ takes any value in (24)–(27) when $p < 1$. This implies that

$$\frac{R}{c} \geq \max\left\{ \frac{f}{p\pi(n-f)}, \frac{f}{p\alpha(1-\pi)(n-f)}, \frac{f}{(1-p)\pi(n-f)}, \frac{f}{(1-p\alpha)(1-\pi)(n-f)} \right\}. \tag{29}$$

When $p = 1$, we do not require (28) to hold for $\mathbb{P}_{\text{posterior}} = (26)$. This is because a rational node in Case 3 knows for sure that the system is in GST, and on equilibrium paths all rational nodes understand that a rational leader can ensure everyone receives `message` from the leader.

We now present the main result of this section.

**Theorem 3.** *When there exist a systematic risk of* `message` *losses, all symmetric equilibria are charaterized as follows.*

1. *A "babbling" equilibrium always exists, in which nodes never commit regardless of the communication. That is, $\mathcal{E} = \emptyset$.*

2. *Fractional-p-equilibria exist if and only if*

$$\frac{R}{c} \geq \max\left\{ \frac{2f}{\pi(n-f)}, \frac{2f}{(1-\pi)(n-f)}, \left(\frac{1}{\pi} + \frac{\alpha^{-1}}{1-\pi}\right)\frac{f}{n-f} \right\}. \tag{30}$$



Electronic copy available at: https://ssrn.com/abstract=3965955

*In this class of equilibria, a rational leader sends* `message` *with probability*

$$p \in \left[\frac{c}{R}\frac{f}{\pi(n-f)}, 1-\frac{c}{R}\frac{f}{\pi(n-f)}\right] \cap \left[\frac{c}{\alpha R}\frac{f}{(1-\pi)(n-f)}, \frac{1}{\alpha}\left(1-\frac{c}{R}\frac{f}{(1-\pi)(n-f)}\right)\right], \quad (31)$$

*a rational backup forwards* `message` *(if received) with probability* $q$*, and a backup commits if and only if she receives* $k \in S(p\alpha^2, q) \cup S(p, q)$ `messages`*, regardless of whether anything is received from the leader. That is,* $\mathcal{E}^0 = \mathcal{E}^1 = S(p\alpha^2, q) \cup S(p, q)$*.*

3. *Unitary-*$p$*-equilibria, which exist when*

$$\frac{R}{c} \geq \max\left\{\frac{f}{\pi(n-f)}, \frac{f}{(1-\pi)\alpha(n-f)}, \frac{f}{(1-\pi)(1-\alpha)(n-f)}\right\}.$$

*In this class of equilibria, a rational leader sends* `message` *to each backup with* $p = 1$*, a rational backup forwards* `message` *(if received) with probability* $q$*, and a backup commits if and only if either (i) she receives* $k \in \mathcal{S}(\alpha^2, q)$ `messages`*, regardless of whether anything is received from the leader, or (ii)* $k \in \mathcal{S}(\alpha^2, q) \cup \mathcal{S}(1, q)$ `messages`*, with one from the leader. That is,* $\mathcal{E}^0 = \mathcal{S}(\alpha^2, q)$ *and* $\mathcal{E}^1 = \mathcal{S}(\alpha^2, q) \cup \mathcal{S}(1, q)$*.*

Two points are noteworthy. First, in fractional-$p$-equilibria, to ensure (31) is nonempty we impose the necessary and sufficient condition (30) on model primitives. Second, unlike Theorem 2 the unitary-$p$ equilibrium exists here given systematic `message` losses, because nodes who receive more than $(n-f)q\alpha^2 + f\alpha^2$ `messages` can infer GST for sure, under which an equilibrium with $p = 1$ exists as shown in Theorem 1.

**Welfare Analysis: The Role of Systematic `message` Losses.** For better comparison with Theorem 2 in Section 5.1, we focus on fractional-$p$-equilibria. A salient feature emerges from Theorem 3: Consensus on `message` becomes much harder to achieve when the system faces a systematic risk of `message` losses. To see this, note that the equilibrium $p$ in (31) must lie in the intersection of two intervals; in fact, each of them corresponds to the relevant condition in one of the aggregate states (GST or non-GST), just as given by (21) in Theorem 2 without systematic risk. Consistent with this intuition, the commit set $\mathcal{E}$, which is independent of $\pi$, is simply the



Electronic copy available at: https://ssrn.com/abstract=3965955

union of two sets $\mathcal{S}(p\alpha^2, q)$ and $\mathcal{S}(p, q)$. In other words, with systematic risk of `message` losses, the need to satisfy equilibria conditions in both aggregate states shrinks the set of equilibria.

In fact, condition (30) implies that fractional-$p$-equilibria exist only when the probability $\pi \in [0, 1]$ of the GST state takes some intermediate value.[19] In the extreme, this class of equilibria fails to exist when $\pi \to 1$, implying that our equilibrium profile is not left-continuous as $\pi \to 1$: The limiting case of $\pi = 1$ corresponds to the baseline case with a "certain" GST state, with well-behaved fractional-$p$-equilibria given in Theorem 1.

The economic intuition is rooted in the unanimous consensus requirement, and the interaction between local knowledge and high-order beliefs (like in global games à la Morris and Shin (2003); and the celebrated result in Rubinstein (1989) that we discuss in Section 6.4). With systematic risk of `message` losses, any rational node without seeing the realized aggregate state—call her Alice—need to not only worry about other rational nodes' (call one of them Bob) local knowledge—i.e., the measure of `messages` that Bob receives), but also how Bob thinks about the aggregate state and Alice's local knowledge. Take the example of $\pi = 1-\epsilon$; in this case, with a probability close to unity all rational nodes receive $k \in \mathcal{S}(p, q)$, however they remain uncertain whether the state is actually non-GST—in which some rational nodes will receive $k \in \mathcal{S}(p\alpha^2, q)\backslash\mathcal{S}(p, q)$ `messages`. However, rational nodes with $k \in \mathcal{S}(p\alpha^2, q)\backslash\mathcal{S}(p, q)$ perceive overwhelming probabilities of the leader being Byzantine, which dissuades them from committing and in turn dissuades those rational nodes with $k \in \mathcal{S}(p, q)$ from committing.

To visualize this effect, take as an example the indicator function of "consensus" $V$:

$$V = \max_{p \text{ s.t. } (31)} \mathbb{1}_{\frac{R}{c} \geq \max\left\{\frac{f}{p\pi(n-f)}, \frac{f}{p\alpha(1-\pi)(n-f)}, \frac{f}{\pi(1-p)(n-f)}, \frac{f}{(1-p\alpha)(1-\pi)(n-f)}\right\}}, \tag{32}$$

where the planner chooses $p$ that indexes the equilibrium given in Theorem 3. As explained, the ex ante welfare is nonmonotone in the `message` delivery probability $\pi$. This is shown in Panel B of Figure 3, in contrast to the `message` delivery probability $\alpha$ when the `message` losses are idiosyncratic (Panel A of Figure 3). The nonmonotonicity is driven by two competing forces: On

[19] If the distribution of $\alpha$'s has a wider support that covers the edge 0 or 1, then consensus on `message` definitely fails. As explained below, this is because the stringent requirement that commit decisions be unanimous across all rational nodes.



Electronic copy available at: https://ssrn.com/abstract=3965955

the one hand, a higher $\pi$ makes it more possible for nodes to receive `messages`, so those who receive the leader's `message` are more likely to commit; however, it also leads those who do not receive the leader's `message` not to commit, making a successful consensus on `message` harder.

To illustrate this point further, we consider a "smooth" version of (32) by calculating the expected welfare, when we face "random" $R/c$ that follows some distribution. Denote the resulting ex ante welfare function by $\hat{V}(\pi)$.[20] We have the following proposition.

**Proposition 4.** *For any fully supported distribution of $\frac{R}{c}$, the ex ante welfare is $\hat{V}(\pi)$ single-peaked in $\pi$ with the single peak given by*

$$\arg\max_{\pi} \hat{V}(\pi) = \frac{\sqrt{\alpha}}{1+\sqrt{\alpha}}.$$

Seeing the single-peakedness is straightforward given what we have learned from Panel B in Figure 3. To see the intuition of $\arg\max_{\pi} \hat{V}(\pi) = \frac{\sqrt{\alpha}}{1+\sqrt{\alpha}}$, we show in the proof that when non-monotonicity occurs in Panel B Figure 3, the welfare-maximizing equilibrium has $p = \frac{\pi}{(1-\pi)\alpha+\pi}$. In this case, $V = \mathbb{1}_{\frac{R}{c} \geq \frac{((1-\pi)\alpha+\pi)f}{(1-\pi)(n-f)\pi\alpha}}$, and the threshold $\frac{((1-\pi)\alpha+\pi)f}{(1-\pi)(n-f)\pi\alpha}$ for $R/c$ takes a minimum value at $\pi = \frac{\sqrt{\alpha}}{1+\sqrt{\alpha}}$.

# 6 Further Discussions

There are many abstractions we make to highlight the key insight of our model. In the next section, we provide more expansive discussions about various important conceptual issues including equivocation, forks, and multiple views.

## 6.1 Robustness to Equivocation

In the literature, Byzantine behaviors typically also include equivocation, that is, sending different `messages` to peer nodes even when the protocol stipulates sending a unique one. Specifically, equivocation in our setup would take the form of a Byzantine leader simultaneously sending a

[20] The same result holds for ex ante expected surplus $\hat{W}(\pi)$, which corresponds to $W$ in Eq. (22).



Electronic copy available at: https://ssrn.com/abstract=3965955

`message` and some different `message`$'$ to backup nodes. The ability to equivocate typically gives Byzantine nodes more power to disrupt distributed consensus formation.

Although we do not explicitly model the possibility of equivocation, as the leader is only allowed to either send `message` or not, we can reason that introducing the possibility of equivocation would not change the consensus outcome in our baseline model. This is because when a rational backup's information set is compatible with the leader being rational (that is, when she only receives a unique value from $k \in \mathcal{E}$ `messages`), she expects the leader to be rational, i.e., event $\mathcal{R}$ (or irrational, i.e., event $\overline{\mathcal{R}}$) with probability $\frac{n-f}{n}$ (or $\frac{f}{n}$). Committing to the value she has received thus gives $R$ (or $-c$) in the former case (or in the worst-case scenario of the latter case), which is the same as when Byzantine nodes cannot equivocate.

Intuitively, in our setup, a rational leader can always ensure consensus success, while a Byzantine leader can always disrupt consensus, even without the possibility of equivocation. Therefore, enhancing Byzantine nodes with the ability to equivocate would not improve or harm the outcome.

### 6.2 Forks

Motivated by the well-known double-spending problem, the prevention of forks (that is, different rational nodes committing to different records) is at the core of the development of a blockchain system. It is, however, worth noting that forks have different meanings in permissioned BFT consensus–based and permissionless Nakamoto consensus–based systems. The widely held view is that in mainstream (permissioned) BFT protocols, forks never happen because nodes will never change a committed decision, and they only commit when they are sure that other nodes either have committed or will commit to the same value. The BFT literature refers to this property as safety. On the other hand, forks can always happen in permissionless Nakamoto consensus–based systems like Bitcoin because nodes in Nakamoto consensus never reach the type of strong consensus required by BFT protocols; rather nodes only reach "asymptotic" consensus, in that the probability of any blocks being overturned is never zero, but only decreases exponentially over time.

Therefore, forking may have different interpretations in this literature. It could describe a situation where some but not all rational nodes commit to a certain `message`, while the remaining



Electronic copy available at: https://ssrn.com/abstract=3965955

rational nodes do not; and this case is captured by the probabilistic "bad" commit decision (when the leader is Byzantine) and penalty $-c$ in our framework. It could also describe a situation where a Byzantine leader sends different `messages` to rational nodes who therefore commit to different `messages` (by following their equilibrium strategies). In our model, we do not consider the possibility that a Byzantine leader sends different `messages`. However, as shown in the previous section, adding this possibility would not change our results, implying that the rational nodes will never commit to different `messages` even in this extension. Finally, there is also a possibility that rational nodes all agree to revise a certain history. This case can be accommodated within our model framework by expanding the `message` space by adding "remove certain history."[21]

It is important to stress that the DAO-type of forking where two systems coexist is ruled out by our assumption, as we assume that nodes get positive payoff if and only if the consensus is unanimous. This feature is related to some established results on how to avoid forks in permissionless systems. For example Saleh (2021) shows that the notorious "nothing-at-stake" problem of proof-of-stake (PoS) blockchains is resolved if the nodes' payoffs are higher when the consensus is unanimous rather than when multiple branches coexist.[22]

### 6.3 Uncertainty, Risk, and Ambiguity Aversion

Our framework combines ambiguity aversion and expected utility. The rational nodes are ambiguity averse over Byzantine actions, but they form expectations over whether the leader is rational or Byzantine. This assumption is crucial for obtaining a successful consensus on `message` in the model analyzed here. If we instead assume that rational nodes are also ambiguity averse about whether the leader is rational, then the consensus on `message` will always fail (that is, only the babbling equilibrium exists). This is because every rational node who receives $k$ `messages` always considers the following worst case scenario to be possible: 1) The leader is Byzantine and 2) Byzantine nodes' strategy profile falls within $\mathcal{B}^1(k)$ or $\mathcal{B}^0(k)$. Thus, a rational node would always choose not to commit.

[21] Biais et al. (2019) study this type of fork by multiplicity of equilibrium outcomes in settings of Bitcoin-like proof-of-work blockchains, which they call "annihilation of certain history."

[22] The payoff vehicle in Saleh (2021) is the price of coins. It is assumed that the price of coins significantly drops when multiple branches are perpetuated.



Electronic copy available at: https://ssrn.com/abstract=3965955

One of the reasons why the consensus on `message` always fails in our model under full ambiguity aversion is because we do not allow for the possibility of replacing potentially Byzantine leaders. Such leader replacement processes are called "view changes" in BFT protocols, and the next section discusses this possibility.

### 6.4 Multiple Views

In the standard computer science setting there are multiple rounds of communication. BFT protocols in the computer science literature are characterized by a safety-liveness trade-off: If nodes are too aggressive in commit decisions, they tend to commit prematurely, creating inconsistent commit decisions across nodes leading to a safety failure. On the other hand, if nodes are too cautious in commit decisions, they tend to be indecisive, leading the protocol to get stuck—in other words, a liveness failure. BFT protocols thus are designed in such a way to strike the right balance between neither being too aggressive, nor too cautious in commit decisions, achieving safety and liveness simultaneously. As a part of not being too aggressive, BFT protocols typically feature a view-change process, so that when local information is not adequate to justify a commit decision, nodes do not simply deem the consensus on `message` to fail, but rather they replace the leader and play the consensus game again. Under a "partial synchrony" assumption, as the consensus game is repeatedly played, consensus on `message` will be reached within an adequate time after GST, even though this fact may not be common knowledge so the consensus game may have to be played forever.[23]

The model we have analyzed is effectively a consensus game with one view. A fruitful future research direction is to investigate whether a repeated game (without a deterministic end) that explicitly models view changes may obtain nontrivial consensus outcomes (i.e. not babbling) even with full ambiguity aversion. Explicit modeling view-changes may also accommodate additional directions for future research, as we explain now.

[23] A partial synchrony network assumes that GST will arrive at an unknown time in the future, after which $\alpha = 1$. This fact together with *view-changes* ensures that all honest nodes know that some future leader (potentially after many view-changes) is non-Byzantine.



Electronic copy available at: https://ssrn.com/abstract=3965955

**Analogy with Email Game** It would be particularly interesting to probe the potential analogy between our result with that of an "email" game (Rubinstein (1989)), which is an interesting application of "almost common knowledge" and closely connects to the global games literature. More specifically, in the email game with expected utility, if the game has to stop after a (commonly known) finite number of rounds, coordination fails probabilistically; while if the game repeats indefinitely, then coordination definitely fails. Our current setup of one view corresponds to a finite period game, while allowing view-changes as in the computer science literature corresponds to an infinitely repeated game. This seems to suggest that the commonly used setting in computer science may feature an equilibrium outcome that "coordination always fails," once the nodes behave as rational economic agents do. That said, "view changes" that exist in the standard computer science setting but not in the email game may help coordination in this dynamic system.[24]

**Equivocation** Finally, one may explicitly consider equivocation in an expanded framework with view changes. Although we have explained in Section 6.1 that in the baseline model of our current setup, introducing equivocation (i.e. `message` and `message`$'$) does not change the consensus outcome, this conclusion may be revised when multiple views are introduced. This is because with view changes, a previous leader who equivocates may have nodes inherit different values in a new view, complicating the consensus process.

# 7 Conclusion

BFT protocols have been proposed for permissioned blockchains powered by multiple self-interested parties. However, a challenge arises as traditional BFT protocols impose "honest" behaviors, leaving no room for incentives analysis. In this paper, we provide a framework to analyze the incentives of the nodes in maintaining reliable distributed ledger. We model rational nodes as ambiguity averse to Byzantine strategies, and focus on frictions such as peer-to-peer information transition as well as commit decisions based on local information, and thus stay close to traditional assumptions of BFT protocols. Our model thus provides a framework for future work in the

[24] Besides, our paper adopt the ambiguity averse preference, which features expected utility that max-minimize over multiple priors, as opposed to standard expected utility in Rubinstein (1989). It is unclear about the role of ambiguity aversion in a dynamic setting with multiple views.



Electronic copy available at: https://ssrn.com/abstract=3965955

strategic analysis of BFT protocols in specific and distributed consensus in general.

We show that accounting for rational non-Byzantine nodes gives rise to multiple equilibria in the BFT consensus game. In babbling equilibria, which always exist, no information is added to the blockchain. There are a variety of equilibria where consensus on new information is achieved, which differ in the nodes' messaging strategies. These equilibria exist only if the individual payoffs for achieving consensus are large enough.

The traditional treatment of BFT consensus in the computer science literature does not need to concern itself with multiplicity of equilibria and payoffs because of the algorithmically prescribed behavior of the honest nodes. However, as blockchain applications often rely on independent parties to maintain a shared ledger, the design of blockchains should take these concerns into account.

## References


**Abadi, Joseph, and Markus Brunnermeier.** 2018. "Blockchain economics." National Bureau of Economic Research. 5

**Abraham, Ittai, Lorenzo Alvisi, and Joseph Y Halpern.** 2011. "Distributed computing meets game theory: combining insights from two fields." *Acm Sigact News*, 42(2): 69–76. 5

**Aiyer, Amitanand S, Lorenzo Alvisi, Allen Clement, Mike Dahlin, Jean-Philippe Martin, and Carl Porth.** 2005. "BAR fault tolerance for cooperative services." 45–58. 5

**Amoussou-Guenou, Yackolley, Bruno Biais, Maria Potop-Butucaru, and Sara Tucci-Piergiovanni.** 2020. "Committee-based Blockchains as Games Between Opportunistic players and Adversaries." 5, 6

**Angeletos, George-Marios, and Iván Werning.** 2006. "Crises and prices: Information aggregation, multiplicity, and volatility." *american economic review*, 96(5): 1720–1736. 2

**Auer, Raphael, Cyril Monnet, and Hyun Song Shin.** 2021. "Permissioned distributed ledgers and the governance of money." 6




Electronic copy available at: https://ssrn.com/abstract=3965955

**Benhaim, Alon, Brett Hemenway Falk, and Gerry Tsoukalas.** 2021. "Scaling Blockchains: Can Elected Committees Help?" *Available at SSRN 3914471*. 6

**Bergemann, Dirk, and Stephen Morris.** 2005. "Robust mechanism design." *Econometrica*, 1771–1813. 11

**Biais, Bruno, Christophe Bisiere, Matthieu Bouvard, and Catherine Casamatta.** 2019. "The blockchain folk theorem." *Review of Financial Studies*, 32(5): 1662–1715. 41

**Branderburger, Adam, and Kai Steverson.** 2020. "Using "Proof-of-Presence" to Coordinate." 5

**Buchman, Ethan.** 2016. "Tendermint: Byzantine fault tolerance in the age of blockchains." PhD diss. 5

**Budish, Eric.** 2018. "The Economic Limits of Bitcoin and the Blockchain." National Bureau of Economic Research. 5

**Buterin, Vitalik, and Virgil Griffith.** 2017. "Casper the friendly finality gadget." *arXiv preprint arXiv:1710.09437*. 5

**Castro, Miguel, and Barbara Liskov.** 1999. "Practical Byzantine fault tolerance." *Proceedings of the third symposium on Operating systems design and implementation*, 173–186. 5, 7, 8

**Clement, Allen, Harry Li, Jeff Napper, Jean-Philippe Martin, Lorenzo Alvisi, and Mike Dahlin.** 2008. "BAR primer." 287–296, IEEE. 5

**Cong, Lin William, and Zhiguo He.** 2019. "Blockchain disruption and smart contracts." *The Review of Financial Studies*, 32(5): 1754–1797. 5

**Cong, Lin William, Zhiguo He, and Jiasun Li.** 2021. "Decentralized mining in centralized pools." *The Review of Financial Studies*, 34(3): 1191–1235. 5

**Crawford, Vincent P, and Joel Sobel.** 1982. "Strategic information transmission." *Econometrica: Journal of the Econometric Society*, 1431–1451. 3, 10, 13



Electronic copy available at: https://ssrn.com/abstract=3965955

**Eliaz, Kfir.** 2002. "Fault tolerant implementation." *The Review of Economic Studies*, 69(3): 589–610. 11

**Epstein, Larry G, and Martin Schneider.** 2003. "Recursive multiple-priors." *Journal of Economic Theory*, 113(1): 1–31. 11

**Galeotti, Andrea, Sanjeev Goyal, Matthew O Jackson, Fernando Vega-Redondo, and Leeat Yariv.** 2010. "Network games." *The review of economic studies*, 77(1): 218–244. 2

**Gilboa, Itzhak, and David Schmeidler.** 1993. "Updating ambiguous beliefs." *Journal of economic theory*, 59(1): 33–49. 11, 20

**Groce, Adam, Jonathan Katz, Aishwarya Thiruvengadam, and Vassilis Zikas.** 2012. "Byzantine agreement with a rational adversary." 561–572, Springer. 12

**Halaburda, Hanna, Guillaume Haeringer, Joshua S Gans, and Neil Gandal.** forthcoming. "The microeconomics of cryptocurrencies." *Journal of Economic Literature*. 5

**Hanany, Eran, Peter Klibanoff, and Sujoy Mukerji.** 2020. "Incomplete information games with ambiguity averse players." *American Economic Journal: Microeconomics*, 12(2): 135–87. 11

**Kiayias, Aggelos, Elias Koutsoupias, Maria Kyropoulou, and Yiannis Tselekounis.** 2016. "Blockchain mining games." 365–382. 5

**Kroll, Joshua A, Ian C Davey, and Edward W Felten.** 2013. "The economics of Bitcoin mining, or Bitcoin in the presence of adversaries." Vol. 2013, 11. 5

**Lamport, Leslie, Robert Shostak, and Marshall Pease.** 1982. "The Byzantine Generals Problem." *ACM Transactions on Programming Languages and Systems*, 4(3): 382–401. 5, 7

**Leshno, Jacob, and Philipp Strack.** 2020. "Bitcoin: An impossibility theorem for proof-of-work based protocols." *American Economic Review: Insights*. 5

**Machina, Mark J, and Marciano Siniscalchi.** 2014. "Ambiguity and ambiguity aversion." In *Handbook of the Economics of Risk and Uncertainty*. Vol. 1, 729–807. Elsevier. 11



Electronic copy available at: https://ssrn.com/abstract=3965955

**Morris, Stephen, and Hyun Song Shin.** 2002. "Social value of public information." *american economic review*, 92(5): 1521–1534. 2

**Morris, Stephen, and Hyun Song Shin.** 2003. "Global games: theory and applications." 36, 38

**Pass, Rafael, and Elaine Shi.** 2018. "Thunderella: Blockchains with optimistic instant confirmation." 3–33, Springer. 5

**Rubinstein, Ariel.** 1989. "The Electronic Mail Game: Strategic Behavior Under" Almost Common Knowledge"." *American Economic Review*, 385–391. 4, 8, 38, 43

**Saleh, Fahad.** 2021. "Blockchain without waste: Proof-of-stake." *The Review of financial studies*, 34(3): 1156–1190. 5, 41

**Shi, Elaine.** 2020. *Foundations of Distributed Consensus and Blockchains.* Book manuscript, Available at https://www.distributedconsensus.net. 5

**Siniscalchi, Marciano.** 2011. "Dynamic choice under ambiguity." *Theoretical Economics*, 6(3): 379–421. 11

**Yin, Maofan, Dahlia Malkhi, Michael K Reiter, Guy Golan Gueta, and Ittai Abraham.** 2018. "HotStuff: BFT consensus in the lens of blockchain." *arXiv preprint arXiv:1803.05069.* 5

# A Posterior probabilities with systematic risk

We discuss all the following cases sequentially:

- If $k \in [(n-f)pq\alpha^2, (n-f)pq)$ and not receiving `message` from the leader, the rational backup sees GST and the leader being rational (event $\mathcal{G}\&\mathcal{R}$) with probability 0, as well as non-GST and the leader being rational (event $\overline{\mathcal{G}}\&\mathcal{R}$) with a posterior probability of

$$\mathbb{P}(\overline{\mathcal{G}}\&\mathcal{R}|\mathcal{I}^0) = \frac{\mathbb{P}(\mathcal{I}^0|\overline{\mathcal{G}}\&\mathcal{R})\mathbb{P}(\overline{\mathcal{G}}\&\mathcal{R})}{\mathbb{P}(\mathcal{I}^0|\overline{\mathcal{G}}\&\mathcal{R})\mathbb{P}(\overline{\mathcal{G}}\&\mathcal{R}) + \mathbb{P}(\mathcal{I}^0|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})}$$



Electronic copy available at: https://ssrn.com/abstract=3965955

$$
\begin{aligned}
&= \frac{(1-p\alpha)(1-\pi)\mathbb{P}(\mathcal{R})}{(1-p\alpha)(1-\pi)\mathbb{P}(\mathcal{R})+\mathbb{P}(\mathcal{I}^0|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&\geq \frac{(1-p\alpha)(1-\pi)\mathbb{P}(\mathcal{R})}{(1-p\alpha)(1-\pi)\mathbb{P}(\mathcal{R})+\mathbb{P}(\overline{\mathcal{R}})} = \frac{(1-p\alpha)(1-\pi)(n-f)}{(1-p\alpha)(1-\pi)(n-f)+f}
\end{aligned} \tag{33}
$$

- If $k \in [(n-f)pq\alpha^2, (n-f)pq)$ and receiving `message` from the leader, the rational backup sees GST and the leader being rational (event $\mathcal{G}\&\mathcal{R}$) with probability 0, as well as non-GST and the leader being rational (event $\overline{\mathcal{G}}\&\mathcal{R}$) with a posterior probability of

$$
\begin{aligned}
\mathbb{P}(\overline{\mathcal{G}}\&\mathcal{R}|\mathcal{I}^1) &= \frac{\mathbb{P}(\mathcal{I}^1|\overline{\mathcal{G}}\&\mathcal{R})\mathbb{P}(\overline{\mathcal{G}}\&\mathcal{R})}{\mathbb{P}(\mathcal{I}^1|\overline{\mathcal{G}}\&\mathcal{R})\mathbb{P}(\overline{\mathcal{G}}\&\mathcal{R})+\mathbb{P}(\mathcal{I}^1|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&= \frac{p\alpha(1-\pi)\mathbb{P}(\mathcal{R})}{p\alpha(1-\pi)\mathbb{P}(\mathcal{R})+\mathbb{P}(\mathcal{I}^1|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&\geq \frac{p\alpha(1-\pi)\mathbb{P}(\mathcal{R})}{p\alpha(1-\pi)\mathbb{P}(\mathcal{R})+\mathbb{P}(\overline{\mathcal{R}})} = \frac{p\alpha(1-\pi)(n-f)}{p\alpha(1-\pi)(n-f)+f}
\end{aligned} \tag{34}
$$

- If $k \in ((n-f)pq\alpha^2 + fp\alpha^2, (n-f)pq + fp]$ and not receiving `message` from the leader, the rational backup sees non-GST and the leader being rational (event $\overline{\mathcal{G}}\&\mathcal{R}$) with probability 0, as well as GST and the leader being rational (event $\mathcal{G}\&\mathcal{R}$) with a posterior probability of

$$
\begin{aligned}
\mathbb{P}(\mathcal{G}\&\mathcal{R}|\mathcal{I}^0) &= \frac{\mathbb{P}(\mathcal{I}^0|\mathcal{G}\&\mathcal{R})\mathbb{P}(\mathcal{G}\&\mathcal{R})}{\mathbb{P}(\mathcal{I}^0|\mathcal{G}\&\mathcal{R})\mathbb{P}(\mathcal{G}\&\mathcal{R})+\mathbb{P}(\mathcal{I}^0|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&= \frac{(1-p)\pi\mathbb{P}(\mathcal{R})}{(1-p)\pi\mathbb{P}(\mathcal{R})+\mathbb{P}(\mathcal{I}^0|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&\geq \frac{(1-p)\pi\mathbb{P}(\mathcal{R})}{(1-p)\pi\mathbb{P}(\mathcal{R})+\mathbb{P}(\overline{\mathcal{R}})} = \frac{(1-p)\pi(n-f)}{(1-p)\pi(n-f)+f}
\end{aligned} \tag{35}
$$

- If $k \in ((n-f)pq\alpha^2 + fp\alpha^2, (n-f)pq + fp]$ and receiving `message` from the leader, the rational backup sees non-GST and the leader being rational (event $\overline{\mathcal{G}}\&\mathcal{R}$) with probability 0, as well as GST and the leader being rational (event $\mathcal{G}\&\mathcal{R}$) with a posterior probability of

$$
\begin{aligned}
\mathbb{P}(\mathcal{G}\&\mathcal{R}|\mathcal{I}^1) &= \frac{\mathbb{P}(\mathcal{I}^1|\mathcal{G}\&\mathcal{R})\mathbb{P}(\mathcal{G}\&\mathcal{R})}{\mathbb{P}(\mathcal{I}^1|\mathcal{G}\&\mathcal{R})\mathbb{P}(\mathcal{G}\&\mathcal{R})+\mathbb{P}(\mathcal{I}^1|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&= \frac{p\pi\mathbb{P}(\mathcal{R})}{p\pi\mathbb{P}(\mathcal{R})+\mathbb{P}(\mathcal{I}^1|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&\geq \frac{p\pi\mathbb{P}(\mathcal{R})}{p\pi\mathbb{P}(\mathcal{R})+\mathbb{P}(\overline{\mathcal{R}})} = \frac{p\pi(n-f)}{p\pi(n-f)+f}
\end{aligned} \tag{36}
$$



Electronic copy available at: https://ssrn.com/abstract=3965955

Based on the above derivations, we can further get that:

- If $k \in [(n-f)pq, (n-f)pq\alpha^2 + fp\alpha^2]$ and not receiving `message` from the leader, by the same logic behind (26) and (24) we get that the rational backup sees GST and the leader being rational (event $\mathcal{G}\&\mathcal{R}$) with a (worst case) posterior probability of $\frac{(1-p)\pi(n-f)}{(1-p)\pi(n-f)+f}$, as well as non-GST and the leader being rational (event $\overline{\mathcal{G}}\&\mathcal{R}$) with a (worst case) posterior probability of $\frac{(1-p\alpha)(1-\pi)(n-f)}{(1-\pi)(1-p\alpha)(n-f)+f}$
- If $k \in [(n-f)pq, (n-f)pq\alpha^2 + fp\alpha^2]$ and receiving `message` from the leader, by the same logic behind (27) and (25) we get that the rational backup sees GST and the leader being rational (event $\mathcal{G}\&\mathcal{R}$) with a (worst case) posterior probability of $\frac{p\pi(n-f)}{p(1-\pi)(n-f)+f}$, as well as non-GST and the leader being rational (event $\overline{\mathcal{G}}\&\mathcal{R}$) with a (worst case) posterior probability of $\frac{p\alpha(1-\pi)(n-f)}{p\alpha(1-\pi)(n-f)+f}$
- Finally, if $k$ is outside of $\mathcal{S}(p\alpha^2, q) \cup \mathcal{S}(p, q)$, then a rational node can immediately infer that the leader is Byzantine.

# B Omitted proofs from the main text

*Proof of Proposition 4.* As before we can solve the problem in (32) as follows:

Case 1: If $\frac{1}{\pi} > \frac{1}{(1-\pi)\alpha}$, the welfare-maximizing equilibrium has $p = \frac{1}{2}$, and $V = \mathbb{1}_{\frac{R}{c} \geq \frac{2f}{\pi(n-f)}}$;

Case 2: If $\alpha \leq \frac{1}{2}$, $\frac{1}{\pi} \leq \frac{1}{(1-\pi)\alpha}$; or $\alpha > \frac{1}{2}$, $\frac{1}{\pi} \leq \frac{1}{(1-\pi)\alpha}$, $\frac{\pi}{(1-\pi)\alpha+\pi} < \frac{1}{2\alpha}$, the welfare-maximizing equilibrium has $p = \frac{\pi}{(1-\pi)\alpha+\pi}$, and $V = \mathbb{1}_{\frac{R}{c} \geq \frac{((1-\pi)\alpha+\pi)f}{(1-\pi)\pi\alpha(n-f)}}$;

Case 3: If $\alpha > \frac{1}{2}$, $\frac{\pi}{(1-\pi)\alpha+\pi} \geq \frac{1}{2\alpha}$, the welfare-maximizing equilibrium has $p = \frac{1}{2\alpha}$, and $V = \mathbb{1}_{\frac{R}{c} \geq \frac{2f}{(1-\pi)(n-f)}}$.

Figure 4 illustrates the $(\pi, \alpha)$ regions that correspond to the three cases above.

In case 1, the expected ex ante welfare $\int_0^{+\infty}(n-f)\left(\frac{n-f}{n}R + \frac{f}{n}(-c)\right) \cdot \mathbb{1}_{\frac{R}{c} \geq \frac{2f}{\pi(n-f)}} d\frac{R}{c}$ equals $\int_{\frac{2f}{\pi(n-f)}}^{+\infty}(n-f)\left(\frac{n-f}{n}R + \frac{f}{n}(-c)\right) d\frac{R}{c}$, which strictly increases in $\pi$.



Electronic copy available at: https://ssrn.com/abstract=3965955

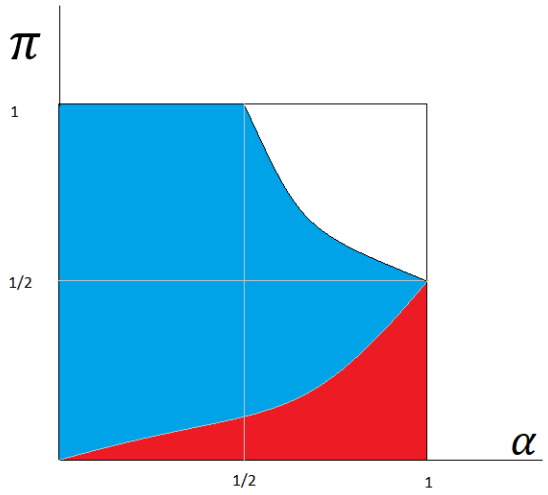



Figure 4: The red region corresponds to case 1, blue case 2, and white case 3. Note that by assumption, we will only focus on areas where $\alpha \geq \sqrt{\frac{n-f}{n}}$. This restriction, however, does not change our conclusion.

In case 2, the expected ex ante welfare $\int_0^{+\infty}(n-f)\left(\frac{n-f}{n}R+\frac{f}{n}(-c)\right)\cdot \mathbb{1}_{\frac{R}{c}\geq\frac{((1-\pi)\alpha+\pi)f}{(1-\pi)\pi\alpha(n-f)}}d\frac{R}{c}$ equals $\int_{\frac{((1-\pi)\alpha+\pi)f}{(1-\pi)\pi\alpha(n-f)}}^{+\infty}(n-f)\left(\frac{n-f}{n}R+\frac{f}{n}(-c)\right)d\frac{R}{c}$, which strictly increases in $\pi$ if $\pi<\frac{\sqrt{\alpha}}{1+\sqrt{\alpha}}$ and strictly decreases in $\pi$ if $\pi>\frac{\sqrt{\alpha}}{1+\sqrt{\alpha}}$. The welfare thus obtains maximum at $\pi=\frac{\sqrt{\alpha}}{1+\sqrt{\alpha}}$.

In case 3, the expected ex ante welfare $\int_0^{+\infty}(n-f)\left(\frac{n-f}{n}R+\frac{f}{n}(-c)\right)\cdot \mathbb{1}_{\frac{R}{c}\geq\frac{2f}{(1-\pi)(n-f)}}d\frac{R}{c}$ equals $\int_{\frac{R}{c}\geq\frac{2f}{(1-\pi)(n-f)}}^{+\infty}(n-f)\left(\frac{n-f}{n}R+\frac{f}{n}(-c)\right)d\frac{R}{c}$, which strictly decreases in $\pi$.

Finally, we show that the ex ante welfare is continuous at the two boundaries between case 1 and 2 as well as between case 2 and 3: (i) At the boundary between case 1 and 2, $\pi=\frac{\alpha}{1+\alpha}$, and the expected welfare equals $\int_{\frac{2(1+\alpha)f}{\alpha(n-f)}}^{+\infty}(n-f)\left(\frac{n-f}{n}R+\frac{f}{n}(-c)\right)d\frac{R}{c}$ in both case 1 and 2; (ii) At the boundary between case 2 and 3, $\pi=\frac{\alpha}{3\alpha-1}$, and the expected welfare equals $\int_{\frac{R}{c}\geq\frac{2(3\alpha-1)f}{(2\alpha-1)(n-f)}}^{+\infty}(n-f)\left(\frac{n-f}{n}R+\frac{f}{n}(-c)\right)d\frac{R}{c}$ in both case 2 and 3. Therefore, the expected welfare is single-peaked with respect to $\pi$ over $[0,1]$. □

## C Proof that nonbabbling equilibrium does not exist when $p=0$

*Proof.* We prove by induction. First, any rational node $i$ who receives some $k^0<f$ invokes $B\in\mathcal{B}^z(k^0)$ and sees it possible that other rational nodes do not receive any `messages` and thus do



Electronic copy available at: https://ssrn.com/abstract=3965955

not commit. Therefore $i$ does not commit either.

Now suppose any rational node $i$ who receives some $k^{m-1} < mf$ `messages` does not commit. Then for any rational node $i$ who receives some $k^m < (m+1)f$ `messages`, she can invoke $B \in \mathcal{B}^z(k^m)$ and sees it possible that other rational nodes receive fewer than $k^{m-1}$ `messages` and thus do not commit. Therefore $i$ does not commit, either. □



Electronic copy available at: https://ssrn.com/abstract=3965955