An Economic Model of Consensus on Distributed Ledgers*

Hanna Halaburda† Zhiguo He‡ Jiasun Li§

May 16, 2023

Abstract

The designs of many new blockchains are inspired by the Byzantine fault tolerance (BFT) problem. While traditional BFT protocols assume most system nodes behave honestly, we recognize that blockchains are deployed in environments where nodes are subject to strategic incentives. This paper thus develops an economic framework for analyzing distributed consensus formation with explicit incentive considerations. We formalize the consensus formation process in a dynamic game with imperfect information and preplay communication where non-Byzantine nodes are *Knightian uncertain* about Byzantine actions, and characterize all of its symmetric equilibria. Our findings enrich those from traditional BFT algorithms, offer guidance for designing blockchains in trustless environments, and also provide a theoretical framework bridging distributed consensus and game theoretical modeling.

Keywords: Ambiguity aversion, Blockchain, Byzantine fault tolerance, Distributed consensus, Game Theory, Higher-Order Beliefs

*We thank Yackolley Amoussou-Guenou, Alessandro Bonatti, Matthieu Bouvard, Andrea Canidio, Engin Iyidogan, Peter Klibanoff, Jacob Leshno, Ye Li, Simon Mayer, David Parkes, Doron Ravid, Diandian Ren, Thomas Rivera, Marciano Siniscalchi, Gerry Tsoukalas, Rakesh Vohra and Leifu Zhang, as well as seminar participants at AEA, ABFER, CBER, Federal Reserve Board Money Week, Hong Kong Conference for Fintech, AI, and Big Data in Business, Bank of Canada Payment Conference XI, IMF/WB Blockchain Summit, UNC-Chapel Hill DeFi Conference, Stony Brook Game Theory Conference, Fields-CFI Workshop on Mathematical Finance and Cryptocurrencies, Tel Aviv University Blockchain conference, Warwick Gillmore Fintech Conference, UWA Blockchain conference, CICF, INFORMS Annual Meeting, CESC at UC Berkeley, Massachusetts Fintech Hub, Global AI Finance Research Conference, Tokenonomics, Columbia CryptoEconomics Workshop, SKEMA-ESSEC Finance Workshop, 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.

1 Introduction

Blockchains feature computer nodes that rely on peer-to-peer communication to maintain their respective ledgers and achieve consensus, that is, to ensure that their respective ledgers keep the same record, even though some nodes may be faulty or hijacked by hackers (such nodes are called *Byzantine faulty*, often abbreviated to *Byzantine*). 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 many new blockchains (e.g., Ethereum's recent upgrade from a proof-of-work system to a proof-of-stake one).

However, a significant differentiation exists between conventional BFT protocols and blockchains. Traditional BFT protocols do not integrate incentives into their design, operating under the assumption that nodes are inherently "honest"—that is to say, non-strategic. This assumption holds true in most distributed systems implemented within a single organization. In contrast, blockchain nodes are often unaffiliated entities that strive to maximize their individual payoffs. The presence of such strategic behaviors presents a new challenge when designing blockchain systems based on BFT protocols. Consequently, incentives must be an integral part of the protocol design, and we need economic analysis to shed light on the relevant forces at play, offer guidance on designing BFT protocols with explicit incentives, and identify trade-offs that arise in such environments. These considerations motivate our paper.

From an economics perspective, classic BFT protocols in the computer science literature have three key features: First, Byzantine nodes may behave arbitrarily, and both the system and non-Byzantine nodes concern the "worst case" scenario regarding Byzantine nodes' arbitrary actions (consensus edge cases). Second, due to their distributed nature, each node only has and thus acts upon "local" information rather than "global" knowledge.[1] Finally, nodes are treated like machines rather than "rational" participants who operate with incentive considerations, as non-Byzantine

[1]Here, we follow the network literature (e.g., Galeotti, Goyal, Jackson, Vega-Redondo and Yariv (2010)) and use "local" information to indicate information that a node holds exclusively; 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).

nodes are all assumed to willingly follow "honest" strategies the protocol stipulates for them. Our focus is on the third key feature.

Specifically, 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 explicitly 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. The framework results in a multiple-stage game that features preplay communications: 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 messages 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. Consistent with typical practices of BFT protocols, every message contains its sender's signature so nodes cannot impersonate others. Consensus is then defined as an outcome in which all rational nodes commit 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. Considering the reward and penalty as an explicit payoff associated with certain outcomes is another point of departure from traditional BFT models. With "honest" nodes following protocol-recommended strategies, traditional BFT protocols have no need for setting such rewards and penalties. These are, however, essential in commit decisions of rational nodes maximizing their expected payoff.

We fully characterize all symmetric equilibria within the game: First, there always exists a set of "gridlock equilibria" in which nodes discard preplay communications and never commit to new messages. Second, under some conditions there also exist "consensus equilibria" in which consensus on the message is reached. In these consensus equilibria, each rational node uses information learned from communication to Bayesian-update the posterior probability of the leader being rational or Byzantine. We show that a Byzantine leader, coordinating with other Byzantine backups, may lead a rational node into a "wrong" commit decision (that is, committing to a message that does not obtain consensus). Seeing this possibility, rational nodes who are ambiguity averse to

Byzantine nodes' strategies prefer not committing when they know the leader is Byzantine. As a result, a rational node commits only if her communication outcome is consistent with the leader being rational. These consensus equilibria exists only when the reward from successfully achieving consensus is sufficiently high compared to the penalty for a wrong commit decision. We characterize conditions on reward and penalty for such equilibria to exist. Finally, we categorize all consensus equilibria into two classes: one in which rational leaders always send messages to all nodes, and the other in which rational leaders withhold messages from some nodes. While traditional BFT protocols resemble the former, we point out that this class of equilibria is not robust to potential message losses once we account for incentives. The two classes of equilibria have distinct conditions on the reward and penalty. Depending on the level of the reward set in the system, different number of equilibria may exist. This dependence reveals new tradeoffs that arise when explicitly accounting for payoffs and incentives in BFT framework.

Our results demonstrate how outcome-dependent payoffs, like the reward and penalty scheme, determine consensus success in a distributed system with rational nodes. In our model, we take the reward and penalty as given, which captures environments in which they are exogenous. Yet, to some extend these values may also be chosen by protocols designers. For example, blockchains typically offer "reward" in the form of block reward, which can only be realized when consensus succeeds.[2] At the same time, they impose "penalty", for example the opportunity cost of staking and severity of slashing policies. In these cases, our results also provide guidance on how to set the reward level as an integral part of the blockchain design. For example, a protocol designer should choose reward to be high enough to ensure existence of a commitment equilibrium, but even higher reward creates additional equilibria, which may not be desirable.

In sum, inspired by widely used BFT consensus protocols in the computer science literature and yet explicitly addressing incentive considerations, this paper develops an economic framework for analyzing BFT consensus protocols in strategic settings as seen in many new blockchain applications. A key departure of our analysis from the mainstream computer science literature is introduction of rational non-Byzantine nodes and incorporation of payoffs, as we show that the

[2] A node can only "cash out" her reward if other nodes agree that she controls these coins.

existence and structure of (multiple) equilibria depend on the payoffs the nodes receive when the consensus is reached or not. Besides offering guidance to blockchain protocol designers to set appropriate incentives for participants in consensus processes, we hope our framework also lays the foundation for more 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 to develop a practical Byzantine fault tolerant (PBFT) protocol. More recent developments in BFT protocols include Buterin and Griffith (2017), Buchman (2016), Pass and Shi (2018), Yin, Malkhi, Reiter, Gueta and Abraham (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.[3]

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, Koutsoupias, Kyropoulou and Tselekounis (2016), Budish (2018), Biais, Bisière, Bouvard and Casamatta (2019*b*), Leshno and Strack (2020), Hinzen, John and Saleh (2022), Cong, He and Li (2021)), and similarly in other permissionless consensus protocols such as proof of stake (e.g, Gans and Gandal (2019), John, Rivera and Saleh (2020, 2021), Saleh (2021), Roşu and Saleh (2021), and He, Li and Wu (2023)), such studies in BFT protocols are more scarce.[4]

A prominent example of incentive analysis in BFT protocols is Amoussou-Guenou, Biais, Potop-

[3]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.

[4]For other works in economics that study the broader implications of blockchain technology, see Abadi and Brunnermeier (2018), Cong and He (2019) and Halaburda, Sarvary and Haeringer (2022), among others. See Halaburda, Haeringer, Gans and Gandal (forthcoming) for an overview of this literature.

Butucaru and Tucci-Piergiovanni (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.[5] 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-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.

Auer, Monnet and Shin (2021) consider a voting-based consensus system, similar to Amoussou-Guenou, Biais, Potop-Butucaru and Tucci-Piergiovanni (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, Biais, Potop-Butucaru and Tucci-Piergiovanni (2020) and Auer, Monnet and Shin (2021), analyzing the incentives to follow the protocol, we look beyond the cost to validate and send messages, and we focus on the role of communication and "local" information, which is essential to any distributed system. 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 mech-

[5] Motivating deviations from protocol prescriptions by operational costs has also been used in the computer science literature, see e.g. the BAR model (Aiyer, Alvisi, Clement, Dahlin, Martin and Porth (2005) and Clement, Li, Napper, Martin, Alvisi and Dahlin (2008)).

anism. 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 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.

Finally, our model also relates to a large literature in economics that study games with preplay communication. Examples include Forges (1990), Bárány (1992), Ben-Porath (2003), Gerardi (2004), Renault, Renou and Tomala (2014), Renou and Tomala (2012), and Rivera (2018), etc.

2 The Model

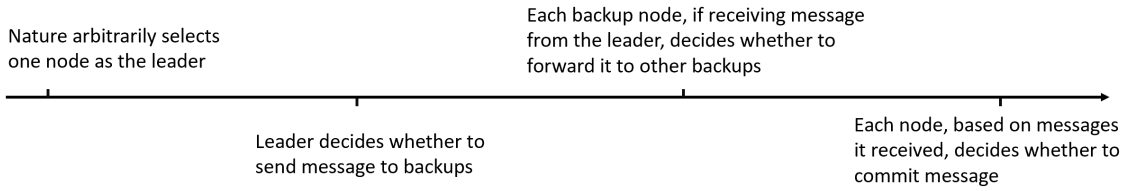
This section lays out the model ingredients and formalizes our equilibrium concept.

2.1 Sequence of Moves

We study a *consensus game* among a measure of n computer nodes with the following sequence of moves:[6] First, nature arbitrarily 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 (Lamport, Shostak and Pease (1982)), `message` can be interpreted as "leader orders to attack," while in the context of a transaction ledger, `message` can be interpreted as a set of new transactions to be added to the ledger. Following the tradition of BFT protocols, every `message` from the leader contains a

[6] We adopt a continuum setup solely for simplicity, although some large-scale consensus protocols like Ethereum (with > 0.5M validators at the time of writing) do provide some justification. Also 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) 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.

Figure 1: Sequence of moves in the consenus game



digital signature that others cannot forge. 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 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 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 draws the sequence of moves in the consensus game.

2.2 Agents

There are a measure of n nodes in the system; we explain the role of the "continuum" toward the end of Section 2.4. Following the literature on BFT protocols, we differentiate between two types of nodes. First, there exists a measure f of *Byzantine* (faulty) nodes, who may together have an

"arbitrary" strategy profile denoted by B that describes all Byzantine nodes' 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. Section 2.3 first gives a formal definition of *consensus*, based on which Section 2.4 provides more details about these rational 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.*

Definition 1 can also be viewed as a characterization of how consensus is typically defined in traditional BFT literature. For example, in the original Byzantine generals problem, consensus on `message` implies that all rational players "attack". In the context of transaction ledgers, consensus on `message` implies (almost) all rational nodes update their local ledgers to include `message`.[7] 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 do. To capture such behaviors, we construct rational nodes' preferences so that they prefer committing to `message` if and only if they believe it would

[7] As a result, (almost) all rational nodes agree on the same state, corresponding to the "safety" requirement in traditional BFT protocols. Furthermore, (almost) all rational nodes make progresses on their local ledgers by updating the status quo, corresponding to the "liveness" requirement in traditional BFT protocols.

reach consensus. We thus assign the following utilities: When a rational node commits to `message`, she receives a positive reward $R > 0$ if consensus succeeds and a penalty $c > 0$ if consensus fails.[8] A rational node who does not commit always gets 0. This utility specification is illustrated as follows:

	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 who commits is rewarded if all her rational peers commit 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 directly enter utilities, the game also has a "cheap talk" flavor à la Crawford and Sobel (1982). Unlike cheap-talk games, however, available committing actions depend on the communication stage, as a node cannot commit if she never receives any `message`.

Our game is dynamic as laid out in Figure 1, in the sense that the to-be-imposed sequential

[8] We will see later that only the rewards to backups enter into subsequent analysis, so we can interpret R as rewards to backups only while accommodating a different reward R_L to the leader, as is commonly observed in practice.

rationality requires a rational node i's sending, forwarding, and committing decisions to be all optimal. We show that the core of the analysis is i's commitment decision; and in fact, all other decisions before commitment stage will satisfy the sequential rationality requirement as an outcome of the equilibrium analysis. This is because first, 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 (1). Second, with a 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, that preplay communication does not enter utility directly, which is intrinsic to consensus games in general, together with the continuum assumption, which we specifically impose for our model, significantly simplifies 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. 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 also relates to the literature on robust mechanism design (e.g., Bergemann and Morris (2005)).

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

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

where $\mathbb{E}_i[\cdot]$ indicates the expectation conditional on node-i's local information. 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, contingent on whether the leader is Byzantine or not. In the computer science tradition, Byzantine nodes are assumed to be able to perfectly coordinate.[9] 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 PBE in our setup is defined over every rational node i's strategy $\tilde{a}_i \equiv \{p_i, q_i, \tilde{C}_i\}$, where

- $p_i \in [0, 1]$ denotes node i's probability of sending `message` to all backups (in an i.i.d. fashion) when being a leader.
- $q_i \in [0, 1]$ denotes node i's probability of forwarding `message` to all other peer nodes (in an i.i.d. fashion) when being a backup who has received `message` from the leader. Formally, if $z \in \{0, 1\}$ denotes receiving `message` from the leader ($z=1$) or not ($z=0$), then $\tilde{q}_i : \{0, 1\} \to [0, 1]$, with $\tilde{q}_i(z = 1) = q_i$ while $\tilde{q}_i(z = 0) = 0$, i.e., the backup cannot forward `message` without receiving one. For the ease of exposition we denote this part of forwarding strategy by q_i.
- $\tilde{C}_i : \{0, 1\} \times [0, 1] \to \{\text{commit, not commit}\}$ 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 $k \in [0, n]$ denotes the measure of `messages` collected from communications. Note that a backup would only be able to commit `message` if $k > 0$.

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

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*
   (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,*

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

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

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

where the expected utility is based on i*'s belief over whether the leader is Byzantine as well as the realizations of* A_{-i}^* *consistent with Bayesian updating given any Byzantine strategy profile* B*.*[10]

Condition (3) implies that node i chooses optimal sending/forwarding decisions, and more importantly, optimal commit decision $\tilde{C}(I_i)$ when facing information set I_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

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

whether she has received the message from the leader, as this fact carries information about whether the leader is Byzantine or not.

Throughout the paper we keep most of the proofs in the main text; these proofs are intuitive logical extensions of a sequence of key lemmas, whose full technical proofs are available in the Appendix.

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, it is easily verified that there always exist gridlock equilibria where $\mathcal{E} = \emptyset$, i.e., rational nodes choose to not commit to `message`, regardless of what happens during the communication stage. However, we are more interested in the existence of consensus equilibria. Hence, this section characterizes the set $\mathcal{E}$ for any symmetric consensus equilibrium with a given pair of (p, q). For clarity of exposition, our analysis focuses on $p > 0$ and $q > 0$.[11]

Since in any equilibrium with given (p, q), a backup can receive at most $(n-f)q+f$ `messages`,[12] without loss of generality, we assume that a rational node with an off-equilibrium $k > (n-f)q+f$ believes that no other nodes commit and thus does not commit either.[13]

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 analyze backup's Bayesian updating in the multiprior framework.

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

Given other rational nodes' equilibrium strategy profile A^*_{-i} (characterized by p, q and $\mathcal{E}$) and

[11] Section 4.4 below gives a brief comment on $q=0$ when discussing the role of peer communication, and Appendix B shows that for $p=0$ no consensus equilibrium exists.

[12] This case occurs when the leader (a Byzantine one when $p < 1$) sends `message` to everyone and all Byzantine backups forward `message` to everyone. Byzantine nodes however cannot make rational backups forward `message` more often than q.

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

i's information set I_i, by (2) the rational backup i's utility from committing to `message` is:

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

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.

Rational Nodes' Information Sets In this section, we introduce the notation for rational nodes' information sets. 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. Define $\mathcal{I}^R$ as the collection of commit-stage information sets that are consistent with a rational node being chosen as the leader, we then have

Lemma 1. *In an equilibrium with $p \in (0, 1]$ and $q \in (0, 1]$, if the leader is rational, then*

$$\mathcal{I}^R \equiv \begin{cases} \{z, k\} : z \in \{0, 1\} \text{ and } k \in S(p, q), & \text{if } p \in (0, 1); \\ \{z, k\} : z = 1 \text{ and } k \in S(1, q), & \text{if } p = 1, \end{cases} \tag{6}$$

where $\mathcal{S}(p, q)$ is a set indexed by p and q defined as

$$\mathcal{S}(p, q) \equiv [(n - f)pq, (n - f)pq + fp]. \tag{7}$$

To see Lemma 1, note that by Definition 2, in an equilibrium with p and q, when the leader is rational, $(n - f)p$ rational backups receive `message`, who each forwards it with probability q; fp Byzantine backups receive `message`, who arbitrary choose whether to forward or not. Therefore,

a rational node receives $(n-f)pq$ `messages` from rational peers and any number within 0 to fp from Byzantine peers, and thus all rational nodes will receive $k \in \mathcal{S}(p,q)$ `messages`.

Expression (6) then 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. Thus, for them z can 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

$$\mathcal{I}^0 \equiv \big\{\{z,k\} : \{z,k\} \in \mathcal{I}^R \text{ and } z = 0\big\} \quad \text{and} \quad \mathcal{I}^1 \equiv \big\{\{z,k\} : \{z,k\} \in \mathcal{I}^R \text{ and } z = 1\big\}. \tag{8}$$

So, $\mathcal{I}^0 \cup \mathcal{I}^1 = \mathcal{I}^R$ 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 Rational Nodes' Inference about Other Rational Nodes' Information Sets

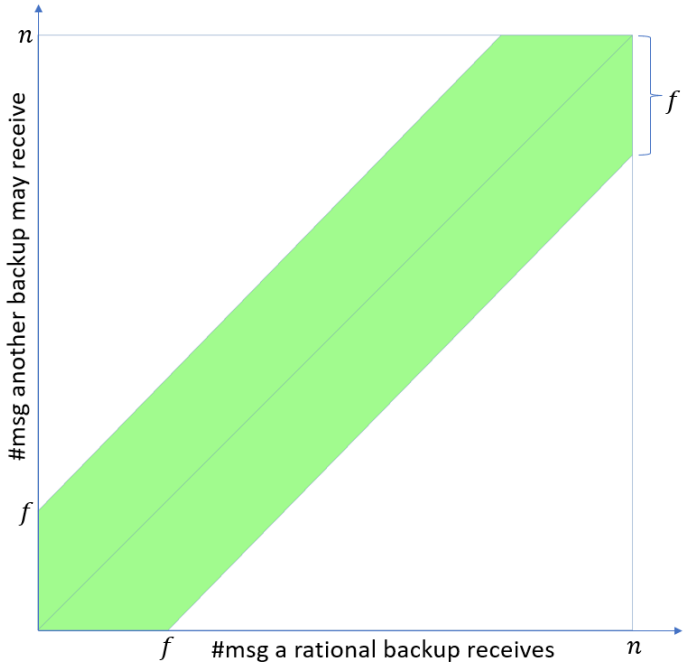
We now take note of another important observation about how a rational node would infer about other rational nodes' information sets.

Lemma 2. *There exists a particular set of Byzantine strategy profiles $\mathcal{B}^z(k)$ for any $k \in [0, (n-f)q + f]$, so that when the leader is Byzantine, any Byzantine strategy profile $B \in \mathcal{B}^z(k)$ leads to a rational node i receiving k `messages` (with or without the leader's, indicated by z) while other rational nodes receive an arbitrary measure of `messages` within $[\max\{0, k-f\}, k)$.*

Figure 2 illustrates a key takeaway of Lemma 2, that is, when the leader is Byzantine, a backup who receives k `messages` infers that the number of `messages` any other node receives must be within $[k-f, k+f]$, subject to a lower bound of 0 and upper bound of n. In other words, with a Byzantine leader, the pair of the number of `messages` a node receives and her inference about the number of `messages` any other node receives must be within the green region in Figure 2.

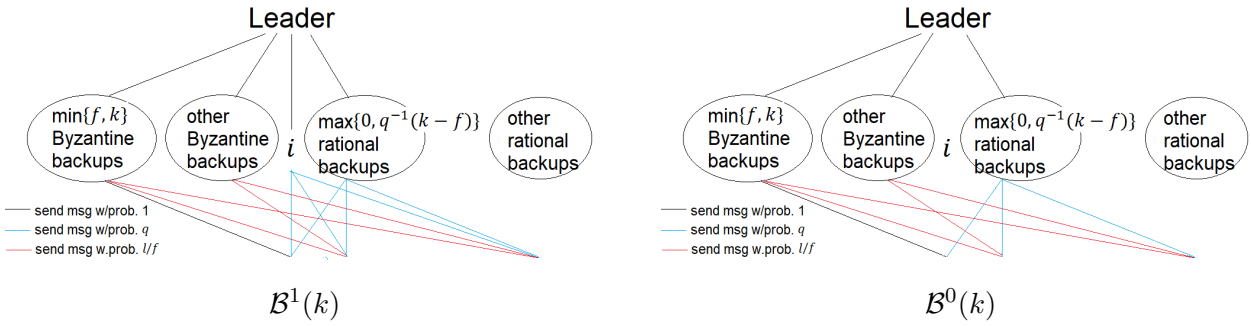
To prove Lemma 2, we can construct $\mathcal{B}^z(k)$ as follows: A strategy profile $B \in \mathcal{B}^z(k)$ specifies

Figure 2: Illustration of Lemma 2



This figure illustrates the key takeaway of Lemma 2: When the leader is Byzantine, the pair of the actual number of `messages` a node receives and her inference about the number of `messages` any other node receives must be within the green region in the figure.

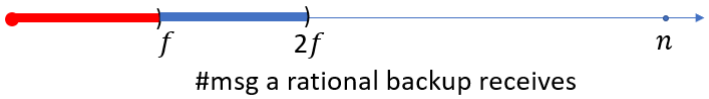
Figure 3: Illustration of $\mathcal{B}^1(k)$ and $\mathcal{B}^0(k)$



$\mathcal{B}^1(k)$ $\mathcal{B}^0(k)$

The left (right) figure illustrates $\mathcal{B}^1(k)$ ($\mathcal{B}^0(k)$): The leader is Byzantine, and he sends `message` to $\max\left\{0, \frac{k-f}{q}\right\}$ rational backups *including* (*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 $\mathcal{B}^1(k)$ ($\mathcal{B}^0(k)$) have the following outcome: Node i receives k `messages`, with (without) one from the leader, while other rational nodes receive an arbitrary measure of $l + \max\{0, k - f\} \in [\max\{0, k - f\}, k)$ `messages`.

Figure 4: Illustration of Lemma 3



This figure illustrates the intuition behind the first two induction steps in the proof of Lemma 3.

that when the leader is Byzantine, he 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 $l \in [0, \min\{f, k\})$. Such a strategy B leads to i receiving k `messages` while other rational nodes receiving $l + \max\{0, k - f\} \in [\max\{0, k - f\}, k)$ `messages` under a Byzantine leader. The set $\mathcal{B}^z(k)$ will play a special role in later proofs. Figure 3 illustrates the strategy profiles.

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 3 starts with an iterated elimination of strictly dominated strategies 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 3. *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 an edge case in which* $p = 1$ *and she receives exactly* $k = (n - f)q + f$ `messages`*.*

A formal proof of Lemma 3 with all details covered is given in Appendix C. The proof is based on an induction argument, for which Figure 4 illustrates its first two steps. A rough intuition goes as follows: First, when a backup receives any number within $[0, f)$ of `messages` (the red region), if she believes that the leader is Byzantine, then by Lemma 2 she will find it possible that other rational backups receive no `messages` and thus do not commit; Therefore, in this case, any rational

and ambiguity averse backup receiving $[0, f)$ `messages` will find it optimal to not commit. Second, when a backup receives any number within $[f, 2f)$ of `messages` (the blue region), if she believes that the leader is Byzantine, then by Lemma 2 she will find it possible that other rational backups receive $[0, f)$ `messages` and thus do not commit (per the previous step). Therefore, in this case, any rational and ambiguity-averse backup receiving $[f, 2f)$ `messages` will also find it optimal to not commit.

The above logic applies to the entire interval $[0, n]$ when the leader is known to be Byzantine, except for one edge case: when $p = 1$ and a rational backup i receives $k = (n - f)q + f$ `messages`, she knows that the leader—even if he is Byzantine—must have sent the `message` to everyone (this is the only possibility for her to receive $k = (n - f)q + f$ `messages`). In other words, in this situation even if the leader is known to be Byzantine, the rational backup with $k = (n - f)q + f$ `messages` still infers from her own knowledge that the leader has been behaving like a rational leader (and further analysis is needed). While our subsequent analyses still discuss this edge case 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 3, the node would not commits in this case. Specifically, it implies that if the number of `messages` a rational backup receives lies in $[0, n]\backslash\mathcal{S}(p, q)$ (recall the Definition in (7)), she can immediately infer that the leader is Byzantine (by Lemma 1) and therefore decide not to commit.

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 consensus equilibrium exists, and further shows that the reverse is also true. Formal proof of the proposition is in Appendix D.

Proposition 1. *In a symmetric consensus 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 $\mathcal{E}^0 = \{(n - f)q + f\}$.*

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 leads to our Proposition 1 only relies on the "worst-case scenario," rather than the expectation over potentially possible priors (which nests the consideration of the worst-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 implicit in the computer science literature on BFT protocols.

4 Equilibrium Characterization

Section 3 has laid out the necessary structures of a symmetric consensus equilibrium. We now further characterize conditions under which symmetric consensus equilibria indeed exist.

4.1 Bayesian Updating and Multiprior Expected Utilities

In Section 3.3, we have pointed out that in a symmetric consensus equilibrium with $p \in (0,1]$ and $q \in (0,1]$, 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, 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.

The above logic implies that given the conjectured symmetric consensus equilibrium with $p \in (0,1]$ and $q \in (0,1]$, a rational node i has only two payoff-relevant potential information partition for her commitment decision: either $I_i \in \mathcal{I}^R$, or $I_i \notin \mathcal{I}^R$; and, conditional on $\mathcal{I}^R$, she needs to calculate the probability of the leader being rational within a multiprior framework in Eq. (4) laid out in Section 3.1 as a part of her expected utility from committing.[14]

Recall in expression (8) we define $\mathcal{I}^z$ as the collection of commit-stage information sets that are consistent with the leader being rational and $z \in \{0,1\}$ (i.e., receiving `message` from the leader).

[14] Belief-updating in a multi-prior framework has a long-standing literature, because of some of its undesired features like "dilation," causing the unappealing feature that "all news are bad news;" see, e.g., Seidenfeld and Wasserman, 1993. By focusing on the information sets with coarsest partitions, our analysis circumvents this issue in the canonical multi-prior setting (Gilboa and Schmeidler, 1993) so that we can invoke the standard Bayesian updating. For some related studies in this topic, see Gul and Pesendorfer (2021).

We have

Lemma 4. *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*

$$\min_{B\in\mathcal{B}} \mathbb{P}(\mathcal{R}|B,\mathcal{I}^z) = \mathbb{P}(\mathcal{R}|B\in\mathcal{B}^z(k),\mathcal{I}^z) = \begin{cases} \frac{p(n-f)}{p(n-f)+f}, & \text{if } z=1; \\ \frac{(1-p)(n-f)}{(1-p)(n-f)+f}, & \text{if } z=0. \end{cases} \tag{9}$$

Proof. Consider a symmetric perfect Bayesian equilibrium with $p \in (0,1]$ and $q \in (0,1]$. When $z=1$, notice that for any $B \in \mathcal{B}$

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

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

$$\begin{aligned} \mathbb{P}(\mathcal{R}|B,\mathcal{I}^0) &= \frac{\mathbb{P}(\mathcal{I}^0|B,\mathcal{R})\mathbb{P}(\mathcal{R})}{\mathbb{P}(\mathcal{I}^0|B,\mathcal{R})\mathbb{P}(\mathcal{R}) + \mathbb{P}(\mathcal{I}^0|B,\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\ &= \frac{(1-p)\mathbb{P}(\mathcal{R})}{(1-p)\mathbb{P}(\mathcal{R}) + \mathbb{P}(\mathcal{I}^0|B,\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\ &= \frac{(1-p)(n-f)}{(1-p)(n-f)+\mathbb{P}(\mathcal{I}^0|B,\overline{\mathcal{R}})f} \geq \frac{(1-p)(n-f)}{(1-p)(n-f)+f}, \end{aligned} \tag{11}$$

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

Lemma 5. *In a symmetric consensus 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 consensus 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,

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

By Lemma 4, both inequalities obtain equality in the above equation when $B \in \mathcal{B}^z(k)$. Furthermore, by Lemma 3, the term $(*)$ equals $-c$—except for when $p = 1$ and $I_i = \{z, (n-f)q + f\}$ for either $z \in \{0,1\}$; there, $(*)$ equals R. The last point holds true, by Proposition 1 because the only way for i to receive $k = (n-f)q + f$ given q is for the measure $(n-f)$ of non-Byzantine nodes to have received `message` from the leader (recall the discussion after Lemma 3). If so, almost all non-Byzantine nodes get $z = 1$ and $k \in S(1,q)$, and therefore commit. Then, committing yields R for i in such a case. □

With the probabilities characterized in Lemma 4 and utilities from committing `message` stated in Lemma 5, we can pin down conditions under which a consensus equilibrium exists.

4.2 Existence of Equilibria with Successful Consensus on `message`

A consensus equilibrium exists if and only if the utility from committing is larger than utility from not committing when other nodes are playing committing 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 consensus equilibrium with* $p = 1$

if and only if

$$\frac{f}{n}(-c) + \frac{n-f}{n}R \geq 0. \tag{12}$$

Proof. To show the existence of an equilibrium, we will 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 $\mathcal{I}^1$ 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 4 and Lemma 5, her utility from committing `message` if all other nodes commit to `message` is

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

And i's best response is to commit to `message` if and only if condition (12) 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 (12) holds. □

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

$$\begin{cases} \frac{f}{p(n-f)+f}(-c) + \frac{p(n-f)}{p(n-f)+f}R \geq 0, \\ \frac{f}{(1-p)(n-f)+f}(-c) + \frac{(1-p)(n-f)}{(1-p)(n-f)+f}R \geq 0. \end{cases} \tag{13}$$

Proof. To prove equilibrium existence, we show that under condition (13), for a 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 4 and 5, the utility from committing to `message` is

$$\min_{B \in \mathcal{B}} \mathbb{E}[u_i(commit, A^*_{-i}, B)|\mathcal{I}^0] = \frac{(1-p)(n-f)}{(1-p)(n-f)+f}\, R - \frac{f}{(1-p)(n-f)+f}\, c\,. \tag{14}$$

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

$$\min_{B\in\mathcal{B}} \mathbb{E}[u_i(commit, A^*_{-i}, B)|\mathcal{I}^1] = \frac{p(n-f)}{p(n-f)+f} R - \frac{f}{p(n-f)+f} c\,. \tag{15}$$

Both (14) and (15) are positive if and only if condition (13) 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.[15] To complete the equilibrium characterization, we need to also identify which p and q can constitute an equilibrium. We are especially interested in consensus 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 consensus equilibrium when the existence condition in Proposition 2 is satisfied. Neither the leader nor the backups have incentives to deviate. To see this, a rational node chosen as the leader knows that if he 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. For backups, since each backup is of measure 0, the deviation from q would not impact anyone's utilities, including her own. Thus, a profitable deviation is not possible. The logic for $p \in (0, 1)$ follows similarly.

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

Theorem 1. *All symmetric equilibria are completely characterized as follows:*

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

2. ***Interval-$\mathcal{E}^0$-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 $q \in$*

[15] While in our analysis in the previous sections we have focused on backups' commit decisions, the rational leader also commits when $k \in S(p, q)$. Note that being a rational leader is covered by the case of $z = 1$, as we can view the leader gets his own message automatically.

$(0, 1]$, *and a backup commits if and only if receiving* $k \in \mathcal{S}(p, q) \equiv [(n-f)pq, (n-f)pq + fp]$ *messages, regardless of whether receiving from the leader. That is,* $\mathcal{E}^0 = \mathcal{E}^1 = \mathcal{S}(p, q)$.

3. ***Singleton-$\mathcal{E}^0$-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 received) with probability* $q \in (0, 1]$, *and a backup commits if and only if receiving* $k \in \mathcal{S}(1, q) = [(n-f)q, (n-f)q + f]$ `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 gridlock equilibrium survives. In addition, singleton-$\mathcal{E}^0$-equilibria are "knife-edge" ones in that if rational nodes' `messages` are only delivered with some probability $\alpha < 1$, then this equilibrium is eliminated, as analyzed in Appendix A.

The condition $\frac{1}{2}(n-f)R \geq fc$ for the existence of interval-$\mathcal{E}^0$-equilibria is more strict than the condition for the existence of singleton-$\mathcal{E}^0$-equilibria. This is because in the former, there is a need to incentivize rational backups who receive `messages` within $\mathcal{S}(p, q)$ to commit both when they received `message` directly from the leader and when they did not.

4.4 Implications for Designing Blockchain Systems

In order to see how explicit incentive considerations affect the system design, notice that in a version of our game with honest nodes, the payoffs would be irrelevant. Moreover, with honest nodes setting any $p \in (0, 1), q > 0$ and $\mathcal{E}^1 = \mathcal{E}^0 = S(p, q)$ or $p = 1, q \geq 0$, $\mathcal{E}^1 = S(1, q)$ and $\mathcal{E}^0 = \{(n-f)q + f\}$ results in the same outcome – reaching consensus on `message` anytime the leader is rational, and failing consensus anytime a Byzantine leader deviates from prescribed strategy.

Any consensus equilibrium in a blockchain system similarly reaches consensus on `message` when the leader is rational and fails when a Byzantine leader deviates from a rational leader's strategy. However, in a blockchain system, where the nodes are rational and individually maximizing their expected payoff, for any recommended parameters, the level of reward needs to be high enough to ensure existence of a consensus equilibrium. At the same time, no level of reward can assure

a consensus equilibrium. This is because with rational nodes, there always exists the gridlock equilibrium in which non-Byzantine nodes discard all preplay communications and do not commit to any messages, while such a situation does not occur with honest nodes.

Moreover, consensus equilibria with different p's require different level of reward. In some cases, the protocol designer's objective may be to achieve consensus at a lowest cost, i.e., with a lower R. For a given environment, characterized by n, f and c,[16] singleton-$\mathcal{E}^0$-equilibria with $p = 1$ require the lowest R. But when an equilibrium with $p = 1$ is not available (as in the case of, say, message loss analyzed in Appendix A), a consensus may be reached with at a lower cost, i.e., with a lower R requirement, when the protocol prescribes p closer to $1/2$, i.e., lower than technically possible.

This relation between the reward requirement and information withholding occurs because when $p < 1$ and a node receives a number of peer messages consistent with a rational leader (i.e., in $S(p, q)$), the node has to have the incentive to commit if she got `message` from the leader or not. Therefore, the more informative z is, i.e., the further p is from $1/2$, the higher reward is needed to incentivize commit decision. If p is very close to 1, a node with $z = 0$ puts very high probability on the leader being Byzantine, and thus needs very high R to compensate for the risk of commit decision. Similarly, when p is close to 0 and the node received `message` from the leader, i.e., $z = 1$.

In contrast, when $p = 1$, reward needs to only satisfy one condition – to incentivize the node to commit when $z = 1$. Note that while a consensus equilibrium with $p = 1$ exists even with $q = 0$, i.e. no forwarded messages,[17] such peer communication is necessary when $p < 1$. This is because with no peer communication a positive mass of rational nodes would not receive any `message` and thus could never commit.

5 Further Discussions

Our model has made many abstractions to highlight its key insight. This section further expands on various important conceptual issues including equivocation, forks, and multiple views.

[16] Depending on the context, c could be exogenous (e.g., cost of running the system and sending messages) or chosen by the protocol designer (e.g., level of staking and slashing policies). R is typically chosen by the designer (e.g. block reward).

[17] In this equilibrium, 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.

5.1 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 `message` and some different $\texttt{message}'$ 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 Byzantine, 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.

5.2 Robustness to the Presence of Honest Nodes

Our model assumes that all non-Byzantine nodes are rational, that is, they all behave to maximize their payoffs. In practice, it is also possible that not all non-Byzantine nodes are rational, and some of them may indeed behave like "honest" nodes, in that they loyally follow prescribed strategies and do not deviate. The presence of "honest" behavior can be rationalized when the protocol-stipulated strategies are written in some default software, so that deviations may require additional modifications to the software, for which nodes may either have little expertise or limited attention.

It is easy to see that all equilibria characterized in the paper are robust to the presence of honest nodes. Intuitively, to verify whether a candidate strategy profile constitutes an equilibrium, one checks that every rational node has no incentives to deviate, holding others' strategies unchanged.

Since honest nodes by assumption stick to their prescribed strategy (in the candidate strategy profile), their presence does not change rational nodes' strategic considerations. Therefore, while we present our model among rational and Byzantine nodes, our findings extend to applications with rational, Byzantine, and honest nodes.

5.3 Forks

Given the well-known double-spending problem, preventing forks is at the core of any blockchain system. This section explains how forks manifest within our setup.

First, we note that forks may have different meanings in BFT consensus-based and Nakamoto consensus-based systems. A widely held view is that in 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 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.

As a result, the literature also interprets forks differently: It may describe a situation where some but not all rational nodes commit to a certain `message` while the remaining rational nodes do not, and such an interpretation of forks is captured in our framework by the probabilistic "bad" commit decision (when the leader is Byzantine) and penalty $-c$; It may 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), and our model does not consider this possibility because allowing a Byzantine leader to send different `messages` does not change our results (as shown in the previous section); Forks may also refer to all rational nodes agreeing to revise certain history, and this possibility can be accommodated within our framework by expanding the `message` space to include "removing certain history" as a specific message.[18] Finally, the DAO-type of forks that result in two coexisting chains is ruled out by assumption, as we assume that

[18] Biais, Bisiere, Bouvard and Casamatta (2019*a*) 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."

nodes get positive payoffs if and only if the consensus is unanimous.[19]

5.4 Uncertainty, Risk, and Ambiguity Aversion

Our framework combines ambiguity aversion and expected utility: Rational nodes are ambiguity averse over Byzantine actions, but form expectations over whether the leader is rational or Byzantine. This assumption is crucial for obtaining a successful consensus on `message`. 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 gridlock equilibrium exists). This is because every rational node who receives k `messages` always deems 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. One reason why the consensus on `message` always fails in our model under full ambiguity aversion is that we do not allow for the possibility of replacing potentially Byzantine leaders. Such leader replacement processes are called "view changes" in traditional BFT protocols, and the next section discusses this possibility.

5.5 Future Directions

We close the section discussing a few future research directions for our framework.

Multiple Views Consensus formation in general features a safety-liveness trade-off: If nodes are too aggressive with their commit decisions, they tend to commit prematurely, creating inconsistent commit decisions across nodes and leading to a safety failure. On the other hand, if they are too cautious, they tend to be indecisive, causing the protocol to get stuck and leading to a liveness failure. BFT protocols in the computer science literature are thus designed to strike the right balance between being neither too aggressive nor too cautious, and achieve 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

[19] This assumption is related to some established results in the literature. For example Saleh (2021) shows that the notorious "nothing-at-stake" problem of proof-of-stake (PoS) blockchains is resolved if nodes get higher payoffs when under a unanimous consensus than when multiple branches coexist. The payoff in Saleh (2021) comes from the price of coins – it is assumed that the coin price drops when multiple branches are perpetuated.

do not simply deem the consensus on `message` as have failed, but rather replace the leader and play the consensus game again. As the consensus game is repeatedly played, under a "partial synchrony" assumption, consensus on `message` will be reached within an adequate time after GST.[20]

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 (i.e. non-gridlock equilibrium) outcomes even with full ambiguity aversion. Explicit modeling view-changes may also accommodate additional directions for future research, as we further explain below.

Multiple rounds of communication Our current model restricts attention to one round of peer communication, although many BFT protocols do feature several rounds. Intuitively, a key friction in consensus protocols is that each node has to make decisions based on only local information learned from communication. While more rounds of communication give nodes more local information and tend to help with consensus in traditional BFT protocols without strategic incentives, it is less obvious with rational nodes as more rounds also increase the space of strategic actions and thus the complexity of "global" information. Therefore, we envision the insight from our current model to be still relevant when we accommodate more rounds, and future work can evaluate this intuition.

Analogy with Email Game An interesting direction is to probe potential analogies between our setup with that in 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

[20] Although GST's arrival may not be common knowledge so the consensus game may have to be played forever — A partial synchrony network assumes that GST will arrive at an unknown future time, after which $\alpha = 1$. This fact together with *view-changes* ensures all honest nodes know that some future leader (potentially after many view-changes) is non-Byzantine.

feature an equilibrium outcome that "coordination always fails," once the nodes behave as rational economic agents do. That said, the leader replacement feature of "view changes" in standard computer science settings but not in the email game may help coordination in this dynamic system.[21]

Equivocation Finally, one may explicitly consider equivocation in an expanded framework with view changes. Although we have explained in Section 5.1 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.

6 Conclusion

While BFT protocols have been proposed for applications in blockchains powered by multiple self-interested parties, challenge arises as traditional BFT protocols stipulate "honest" behaviors, leaving no room for incentives analysis. In this paper, we provide a framework to analyze the incentives of the nodes in maintaining a reliable distributed ledger: We model rational nodes as being ambiguity averse to Byzantine strategies, and focus on frictions such as peer-to-peer information transmission and local information-based commit decisions.

We show that accounting for non-Byzantine nodes' rational incentives gives rise to multiple equilibria in the BFT consensus game. There always exist gridlock equilibria, in which no new information is added to the blockchain. When individual payoffs from achieving consensus are large enough, there may also exist a variety of equilibria in which consensus on new information is achieved. These equilibria differ in nodes' messaging and committing strategies. Furthermore, in some cases consensus may be achieved at a lower cost if the leader decreases the probability of sending messages.

While BFT protocols in the traditional computer science literature does not need to consider

[21] 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.

equilibrium multiplicity thanks to the "honest" node assumption, the design of blockchain applications that rely on independent parties to maintain shared ledgers have to take these concerns into account. As our model incorporates rational incentives yet stays close to existing assumptions in traditional BFT protocols, we provide a framework for future work on the strategic analysis of distributed consensus protocols.

References

Abadi, Joseph, and Markus Brunnermeier. 2018. "Blockchain economics." National Bureau of Economic Research. 4

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. 4

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." 4, 5

Angeletos, George-Marios, and Iván Werning. 2006. "Crises and prices: Information aggregation, multiplicity, and volatility." *american economic review*, 96(5): 1720–1736. 1

Auer, Raphael, Cyril Monnet, and Hyun Song Shin. 2021. "Permissioned distributed ledgers and the governance of money." 5

Bárány, Imre. 1992. "Fair distribution protocols or how the players replace fortune." *Mathematics of Operations Research*, 17(2): 327–340. 6

Benhaim, Alon, Brett Hemenway Falk, and Gerry Tsoukalas. 2021. "Scaling Blockchains: Can Elected Committees Help?" *Available at SSRN 3914471*. 5, 6

Ben-Porath, Elchanan. 2003. "Cheap talk in games with incomplete information." *Journal of Economic Theory*, 108(1): 45–71. 6

Bergemann, Dirk, and Stephen Morris. 2005. "Robust mechanism design." *Econometrica*, 1771–1813. 10

Biais, Bruno, Christophe Bisiere, Matthieu Bouvard, and Catherine Casamatta. 2019*a*. "The blockchain folk theorem." *Review of Financial Studies*, 32(5): 1662–1715. 27

Biais, Bruno, Christophe Bisière, Matthieu Bouvard, and Catherine Casamatta. 2019*b*. "Blockchains, coordination, and forks." Vol. 109, 88–92. 4

Buchman, Ethan. 2016. "Tendermint: Byzantine fault tolerance in the age of blockchains." PhD diss. 4

Budish, Eric. 2018. "The Economic Limits of Bitcoin and the Blockchain." National Bureau of Economic Research. 4

Buterin, Vitalik, and Virgil Griffith. 2017. "Casper the friendly finality gadget." *arXiv preprint arXiv:1710.09437*. 4

Castro, Miguel, and Barbara Liskov. 1999. "Practical Byzantine fault tolerance." *Proceedings of the third symposium on Operating systems design and implementation*, 173–186. 4, 6, 7

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. 4

Cong, Lin William, Zhiguo He, and Jiasun Li. 2021. "Decentralized mining in centralized pools." *The Review of Financial Studies*, 34(3): 1191–1235. 4

Crawford, Vincent P, and Joel Sobel. 1982. "Strategic information transmission." *Econometrica: Journal of the Econometric Society*, 1431–1451. 9

Eliaz, Kfir. 2002. "Fault tolerant implementation." *The Review of Economic Studies*, 69(3): 589–610. 10

Epstein, Larry G, and Martin Schneider. 2003. "Recursive multiple-priors." *Journal of Economic Theory*, 113(1): 1–31. 10

Forges, Françoise. 1990. "Equilibria with communication in a job market example." *The Quarterly Journal of Economics*, 105(2): 375–398. 6

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. 1

Gans, Joshua S, and Neil Gandal. 2019. "More (or less) economic limits of the blockchain." National Bureau of Economic Research. 4

Gerardi, Dino. 2004. "Unmediated communication in games with complete and incomplete information." *Journal of Economic Theory*, 114(1): 104–131. 6

Gilboa, Itzhak, and David Schmeidler. 1993. "Updating ambiguous beliefs." *Journal of economic theory*, 59(1): 33–49. 10, 19

Groce, Adam, Jonathan Katz, Aishwarya Thiruvengadam, and Vassilis Zikas. 2012. "Byzantine agreement with a rational adversary." 561–572, Springer. 11

Gul, Faruk, and Wolfgang Pesendorfer. 2021. "Evaluating ambiguous random variables from Choquet to maxmin expected utility." *Journal of Economic Theory*, 192: 105129. 19

Halaburda, Hanna, Guillaume Haeringer, Joshua S Gans, and Neil Gandal. forthcoming. "The microeconomics of cryptocurrencies." *Journal of Economic Literature*. 4

Halaburda, Hanna, Miklos Sarvary, and Guillaume Haeringer. 2022. *Beyond bitcoin: Economics of digital currencies and blockchain technologies.* Palgrave Macmillan. 4

Hanany, Eran, Peter Klibanoff, and Sujoy Mukerji. 2020. "Incomplete information games with ambiguity averse players." *American Economic Journal: Microeconomics*, 12(2): 135–87. 10

He, Zhiguo, Jiasun Li, and Zhengxun Wu. 2023. "Don't Trust, Verify: The Case of Slashing from a Popular Ethereum Explorer." *WWW '23 Companion*, 1078–1084. New York, NY, USA:Association for Computing Machinery. 4

Hinzen, Franz J, Kose John, and Fahad Saleh. 2022. "Bitcoin's limited adoption problem." *Journal of Financial Economics*, 144(2): 347–369. 4

John, Kose, Thomas J Rivera, and Fahad Saleh. 2020. "Economic implications of scaling blockchains: Why the consensus protocol matters." *Available at SSRN 3750467*. 4

John, Kose, Thomas J Rivera, and Fahad Saleh. 2021. "Equilibrium staking levels in a proof-of-stake blockchain." *Available at SSRN 3965599*. 4

Kiayias, Aggelos, Elias Koutsoupias, Maria Kyropoulou, and Yiannis Tselekounis. 2016. "Blockchain mining games." 365–382. 4

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. 4

Lamport, Leslie, Robert Shostak, and Marshall Pease. 1982. "The Byzantine Generals Problem." *ACM Transactions on Programming Languages and Systems*, 4(3): 382–401. 4, 6

Leshno, Jacob, and Philipp Strack. 2020. "Bitcoin: An impossibility theorem for proof-of-work based protocols." *American Economic Review: Insights*. 4

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. 10

Morris, Stephen, and Hyun Song Shin. 2002. "Social value of public information." *american economic review*, 92(5): 1521–1534. 1

Pass, Rafael, and Elaine Shi. 2018. "Thunderella: Blockchains with optimistic instant confirmation." 3–33, Springer. 4

Renault, Jérôme, Ludovic Renou, and Tristan Tomala. 2014. "Secure message transmission on directed networks." *Games and Economic Behavior*, 85: 1–18. 6

Renou, Ludovic, and Tristan Tomala. 2012. "Mechanism design and communication networks." *Theoretical Economics*, 7(3): 489–533. 6

Rivera, Thomas J. 2018. "Incentives and the Structure of Communication." *Journal of Economic Theory*, 175: 201–247. 6

Roşu, Ioanid, and Fahad Saleh. 2021. "Evolution of shares in a proof-of-stake cryptocurrency." *Management Science*, 67(2): 661–672. 4

Rubinstein, Ariel. 1989. "The Electronic Mail Game: Strategic Behavior Under" Almost Common Knowledge"." *American Economic Review*, 385–391. 7, 29, 30

Saleh, Fahad. 2021. "Blockchain without waste: Proof-of-stake." *The Review of financial studies*, 34(3): 1156–1190. 4, 28

Seidenfeld, Teddy, and Larry Wasserman. 1993. "Dilation for sets of probabilities." *The Annals of Statistics*, 21(3): 1139–1154. 19

Shi, Elaine. 2020. *Foundations of Distributed Consensus and Blockchains.* Book manuscript, Available at https://www.distributedconsensus.net. 4

Siniscalchi, Marciano. 2011. "Dynamic choice under ambiguity." *Theoretical Economics*, 6(3): 379–421. 10

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*. 4

Appendix

A Extension: 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.[22] In this section, we accommodate such possibilities by allowing `messages` to be lost.

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

[22]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.

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

$$\mathcal{S}(p\alpha^2, q) = [(n-f)qp\alpha^2, (n-f)qp\alpha^2 + fp\alpha^2]. \tag{16}$$

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, α^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 use the law of large numbers given idiosyncratic `message` losses.

Inferences and Bayesian Updating Potential `message` losses affect rational backups' inferences. As Eq. (10) and (11) in the proof of Lemma 4 suggest, 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

$$\begin{aligned}
\mathbb{P}(\mathcal{R}|\mathcal{I}^0) &= \frac{\mathbb{P}(\mathcal{I}^0|\mathcal{R})\mathbb{P}(\mathcal{R})}{\mathbb{P}(\mathcal{I}^0|\mathcal{R})\mathbb{P}(\mathcal{R}) + \mathbb{P}(\mathcal{I}^0|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&= \frac{(1-p\alpha)\mathbb{P}(\mathcal{R})}{(1-p\alpha)\mathbb{P}(\mathcal{R}) + \mathbb{P}(\mathcal{I}^0|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&\geq \frac{(1-p\alpha)\mathbb{P}(\mathcal{R})}{(1-p\alpha)\mathbb{P}(\mathcal{R}) + \mathbb{P}(\overline{\mathcal{R}})} = \frac{(1-p\alpha)(n-f)}{(1-p\alpha)(n-f)+f},
\end{aligned} \tag{17}$$

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

$$\begin{aligned}
\mathbb{P}(\mathcal{R}|\mathcal{I}^1) &= \frac{\mathbb{P}(\mathcal{I}^1|\mathcal{R})\mathbb{P}(\mathcal{R})}{\mathbb{P}(\mathcal{I}^1|\mathcal{R})\mathbb{P}(\mathcal{R}) + \mathbb{P}(\mathcal{I}^1|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&= \frac{p\alpha\mathbb{P}(\mathcal{R})}{p\alpha\mathbb{P}(\mathcal{R}) + \mathbb{P}(\mathcal{I}^1|\overline{\mathcal{R}})\mathbb{P}(\overline{\mathcal{R}})} \\
&\geq \frac{p\alpha\mathbb{P}(\mathcal{R})}{p\alpha\mathbb{P}(\mathcal{R}) + \mathbb{P}(\overline{\mathcal{R}})} = \frac{p\alpha(n-f)}{p\alpha(n-f)+f}.
\end{aligned} \tag{18}$$

Committing Decisions and Equilibra Characterization Consensus on `message` requires unanimous commit from all rational nodes.[23] 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, rational backups will commit only when both conditions (19) and (20) are satisfied:

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

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

or equivalently,

$$\frac{R}{c} \geq \frac{f}{n-f} \cdot \max\left\{\frac{1}{p\alpha}, \frac{1}{1-p\alpha}\right\}. \tag{21}$$

The next theorem, which parallels Theorem 1, summarizes all symmetric equilibria when facing idiosyncratic risks of `messages` not being delivered.

Theorem 2. *If all `messages` sent are delivered with probability $\alpha < 1$, we have the following characterization of all symmetric equilibria.*

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

2. *Interval-$\mathcal{E}^0$-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*

$$p \in \left[\frac{1}{\alpha}\frac{fc}{(n-f)R}, \frac{1}{\alpha}\left(1 - \frac{fc}{(n-f)R}\right)\right] \cap [0,1], \tag{22}$$

a rational backup forwards `message` (if received) with probability q, and a rational backup commits if and only if she receives $k \in \mathcal{S}(p\alpha^2, q)$ `messages`, regardless of whether she 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 (The-

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

orem 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 (α, which takes a value of 1 in our baseline model of Theorem 1).

Second, and perhaps with greater economic content, Theorem 2 reveals that Case 3 (singleton-$\mathcal{E}^0$-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.

Because singleton-$\mathcal{E}^0$-equilibria are nongeneric, from now on our analysis focuses on interval-$\mathcal{E}^0$-equilibria, which correspond to Case 2 in both Theorem 1 and 2.

Welfare Analysis Given equilibria multiplicity, a planner (e.g. one designing the protocol) may select endogenous `message` sending/forwarding strategies (p and q) to maximize welfare.

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:

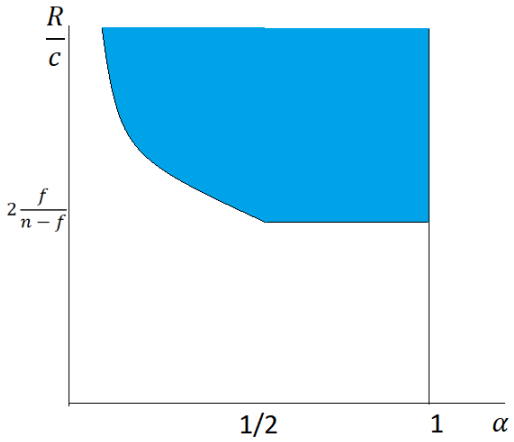
$$W \equiv \max_{p\in\left[\frac{1}{\alpha}\frac{fc}{(n-f)R},\frac{1}{\alpha}\left(1-\frac{fc}{(n-f)R}\right)\right]\cap[0,1]} \underbrace{(n-f)}_{\#\text{rational nodes}} \underbrace{\left(\frac{n-f}{n}R+\frac{f}{n}(-c)\right)}_{\text{expected payoff from committing}} \underbrace{\mathbb{1}_{\frac{R}{c}\geq\max\{\frac{f}{p\alpha(n-f)},\frac{f}{(1-p\alpha)(n-f)}\}}}_{\text{if commits}}. \tag{23}$$

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

$$V \equiv \max_{p\in\left[\frac{1}{\alpha}\frac{fc}{(n-f)R},\frac{1}{\alpha}\left(1-\frac{fc}{(n-f)R}\right)\right]\cap[0,1]} \mathbb{1}_{\frac{R}{c}\geq\max\{\frac{f}{p\alpha(n-f)},\frac{f}{(1-p\alpha)(n-f)}\}}. \tag{24}$$

Problem (24) and problem (23) share the same solution when we view welfare as a function of α. However, as the planner may attach an arbitrary surplus to the consensus, the objective in (24) potentially permits broader interpretations: for instance, the system's safety may serve other purposes with significant social value (say payment); and some key parameters R or c might be

Figure 5: Illustration of V



This figure illustrates V in the parameter space of R/c, with solid area taking a value of 1.

viewed as transfers, and hence part of them should not be counted in welfare.

The solution to problem (24) 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 α.
- If $\alpha < \frac{1}{2}$, the welfare-maximizing equilibrium has $p = 1$. In this case, welfare increases in α.

Figure 5 illustrates the objective V in (23), with the solid area taking a value of 1, in the parameter space of R/c and α. We observe that better communication technology (a higher α) improves the chance of reaching consensus in the system.

Further Comment on the Role of Peer Communication The welfare analysis also further demonstrates 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 $\alpha < 1$. Without peer communication, 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`.

B Proof that consensus 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 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. □

C Proof of Lemma 3

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 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 $l = 0$ (as illustrated in Figure 3) would result in all other rational backups nodes receiving no `messages`, which makes 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 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`

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 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 $k' < (n-f)pq + f$.[24] Thus, node i's utility from committing is $-c$, and she does not commit.

In Step 2 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 be well defined. However, since receiving $k > (n-f)q + f$ is off equilibrium path, by our earlier

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

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. □

D Proof of Proposition 1

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 then an immediate outcome of Lemma 3: If a rational node i's commit-stage information set is not consistent with a rational leader, i.e. $I_i \notin \mathcal{I}^R$, 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 3, 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 3 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

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

$$\leq \mathbb{P}(\mathcal{R}|\hat{B}, I_i) \underbrace{u_i(commit, A^*_{-i}; \hat{B}; \mathcal{R})}_{=-c} + \mathbb{P}(\overline{\mathcal{R}}|\hat{B}, I_i) \underbrace{u_i(commit, A^*_{-i}; \hat{B}; \overline{\mathcal{R}})}_{-c} = -c < 0.$$

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 3), 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. □