Suppose that an attacker with >=50% of the validator set attempts to carry out a censorship attack where they block all prepares and commits coming from the other validators. From the point of view of inside the protocol, this is indistinguishable from a scenario where the complement of the attacker (ie. the victim minority) is offline due to their own fault.

There are two possible schools of thought for how to design incentives in such a scenario. The first is to heavily punish both the majority

(who is possibly a censoring attacker, and is possibly innocent) and the minority

(who is possibly innocent, and possibly nodes that are offline because they are lazy or malicious). If we treat the validators as a captive and unchanging set, this creates strong incentives for all sides to act correctly. However, in the real world, validation is voluntary, and validators are unwilling to join games where they are likely to lose, and will be readily willing to leave such games. Hence, this opens up vulnerability to discouragement attacks, where an attacker carries out a low-grade attack which costs both themselves and victims just enough to make validation unprofitable (and in fact lossy), waits for victims to leave, and then attacks against a much smaller validator set.

The alternative approach is to only penalize nodes that appear offline, and leave other nodes alone, even though it may be the case that it is the online nodes that are at fault because they are censoring

. In the case of 51% attacks, a “just” outcome can be reached through minority chain splits and market-based adjudication. We assume that honest validators simply ignore chains that appear to be censoring their own prepares and commits, and so we can expect these honest validators to form their own chain. This leads to two chains: chain A, run by the majority validators, where the majority keeps their deposits and the minority loses a large fraction of their deposits, and chain B, run by the minority validators, where the minority keeps their deposits and the majority loses a large fraction of their deposits. Note that the PREPARE_COMMIT_CONSISTENCY

slashing condition makes it difficult for a validator to repeatedly “switch sides”; once they commit on one side, they will be stuck on that side until the other side reaches the point where it has justified a checkpoint.

We then rely on the market to adjudicate between the two chains, relying on the assumption that the market prefers a chain where attackers have less sway, and so whichever of the two chains is honest is the one that becomes dominant.

The problem is, however, that this kind of market-based adjudication is expensive and risky

; if a community gets too used to it, then it may pose a centralization risk, and it is certainly a serious usability hurdle every time it happens. So we want it to be expensive to cause a fault that causes the chain to split and devolve to market-based adjudication

. One possible compromise is to rely on the “penalize both sides” approach for low-grade faults, ensuring that attacking small minorities is unprofitable, but then limit the amount of money that validators who appear online can lose, effectively switching to the market-based adjudication approach in the specific case of high-grade faults where large damage to protocol execution is caused.

What are people’s thoughts on this?