# Byzantine Computing on Unknown Randomized External Data

Kelsey Knowlson*, Victoria Lien†, Bryce Palmer‡, Matthew Rackley§
Department of Computer Science, University of New Mexico, Albuquerque, NM

***Abstract*—This paper investigates a robust method for achieving consensus using randomized external data in the presence of Byzantine faults. The concepts extend Augustine et al.'s Data Retrieval (DR) model by the introduction of an error-bounded blacklisting protocol and multiplicative weights. Both mechanisms incorporate robust statistical techniques such as the Median of Means (MoM) to estimate global means under sub-Gaussian or Gaussian assumptions, even when up to a third of peers are Byzantine. The error-bounded blacklisting protocol statistically validates peer estimates before blacklisting nodes. The Multiplicative Weights algorithm is used to resist the influence of Byzantine sources by dynamically adjusting weights of sources based on their deviation from a consensus estimate. The framework provides theoretical guarantees via Chernoff bounds and demonstrates practical potential for secure aggregation in distributed systems. The framework was empirically evaluated through simulations of the Multiplicative Weights method, which demonstrated strong performance and closely aligned with theoretical predictions.**

## I. Introduction

In adversarial environments with the presence of Byzantine nodes, robust methods of reaching consensus are still a relevant issue. Consensus challenges are particularly difficult when data is external, unknown, and randomized, as can be the case in many real-world systems. Without prior knowledge of the distribution of data or the trustworthiness of the sources, typical aggregation methods can be exploited by adversaries that aim to prohibit consensus.

This paper analyzes two complementary approaches to achieve resilient consensus under the aforementioned conditions, built upon the Data Retrieval (DR) model proposed by Augustine et al.[2], which allows nodes to query external data sources in a round-based, synchronous, and symmetric setting. We extend this model with an Error-Bounded Blacklisting Protocol that strengthens resilience by using repeated sampling and statistical guarantees to estimate source means. It improves reliability by performing multiple queries per source and using tight statistical error bounds to differentiate between natural sampling noise and malicious deviations. Additionally, a Multiplicative Weights (MW) approach is used on data sources to identify and reduce the influence of Byzantine nodes. The method aims to manipulate the weights for each data source adaptively, reducing the weight of unreliable sources.

Theoretical analysis provides convergence guarantees and probabilistic bounds on the estimation error, while simulations of the MW method confirm the practicality and robustness of the proposed framework under varying levels of adversarial influence. Together, these techniques offer statistically grounded methodologies for Byzantine-resilient consensus on external, randomized data.

## II. Related Works

This project builds upon several key contributions in distributed systems and Byzantine-resilient computation. The following works have informed both the theoretical foundation and practical motivations for our research, with the most significant influence being the paper by Augustine et al., *Byzantine Resilient Distributed Computing on External Data*.[2] Their work introduces the Data Retrieval (DR) model, a framework designed to address consensus and data access in distributed systems that rely on an external, trusted data source. The DR model assumes a round-based synchronous setting where a fixed number of peers, up to a $\beta$ fraction of which may be Byzantine, can query a shared external array at a cost, while communicating freely with one another. The authors focus on problems as the Download problem, which retrieves the full array, and they propose randomized and deterministic protocols that achieve strong theoretical guarantees even under adversarial conditions.

Our work extends this model in two important aspects. First, we relax the assumption of a known or fixed data distribution. In real-world settings, external data is often unstructured, non-Gaussian, or adversarially generated. Second, while Augustine et al. focus on symmetric trust, we aim to explore settings with asymmetric trust and data-dependent behavior, where nodes may differ in reliability and sample differently over time. Building on their idea of blacklisting and statistical verification, we introduce an Error-Bounded Blacklisting Protocol that uses repeated queries and concentration inequalities to make more principled decisions about trust.

We also incorporate a Multiplicative Weights approach that adaptively reduces the influence of inconsistent sources over time.

Beyond the DR model, other works provide supporting perspectives. Cachin and Zanolini's *From Symmetric to Asymmetric Asynchronous Byzantine Consensus*[3] is relevant for its treatment of asymmetry in trust and asynchronous communication, both of which are closer to practical distributed environments. Although it does not involve external data sources, the consensus model it presents helps contextualize our shift from fully symmetric settings.

Another paper referenced is by Ao et al., *On Precision Bound of Distributed Fault-Tolerant Sensor Fusion Algorithms*.[1] This work studies how accurately one can infer truth in the presence of faulty or adversarial sensors, using methods such as Marzullo's interval approach and the Brooks-Iyengar algorithm. Their results on precision limits and fault tolerance are highly relevant to our environment, where querying an external source acts like sensor sampling, and Byzantine peers resemble misbehaving sensors. The parallels offer insight into the trade-offs between redundancy, accuracy, and robustness in adversarial settings.

Hou et al.'s work, *Randomized View Reconciliation in Permissionless Distributed Systems*[5], contributes an analysis of Sybil attacks and how adversaries can create large numbers of fake identities to mislead the system. Although our model assumes authenticated peers and thus avoids traditional Sybil vulnerabilities, the risk of view divergence due to coordinate Byzantine behavior remains relevant. Their work emphasizes the importance of reconciling inconsistent system views and suppressing outliers, which are objectives we attempt to meet with both our blacklisting and MW-based filtering techniques.

The paper *Byzantine Agreement with Optimal Resilience via Statistical Fraud Detection* by Huang et al.[6] proposes protocols that achieve optimal Byzantine fault tolerance using statistical anomaly detection. Their strategy of using statistical deviations to detect misbehavior aligns closely with our blacklisting mechanism, where source estimates are compared against statistically derived thresholds to prevent false positive while isolating malicious peers. Although their focus is on achieving consensus in an asynchronous setting, the underlying principle of resilience via statistical filtering is shared across both approaches.

In addition, this project incorporates techniques from probabilistic analysis and robust statistics to support its theoretical guarantees. *A Short Note on the Median-of-Means Estimator* by Chen[4] offers a method for mean estimation that remains reliable even when a portion of the data is corrupted. Its use is particularly effective in adversarial settings where traditional averaging fails. To establish error bounds for sampling and loss accumulation, the analysis relies on Chernoff bounds, a tool for quantifying tail probabilities in randomized processes. These bounds are detailed in the *Probability and Computing: Randomized Algorithms and Probabilistic Analysis* textbook by Mitzenmacher and Upfal[7] and further elaborated in Saia's *Bitcoin Consensus* lecture notes.[8] Both provide important context for concentration results used throughout this project.

Together, these works form the theoretical and practical foundation for our project. Building on the framework introduced by Augustine et al.'s DR model, our work extends it to handle more realistic and dynamic environments, where data is uncertain, adversaries are adaptive, and trust must continually adapt over time.

## III. Model

This project took two approaches that work together and build from each other: an Error-Bounded Blacklisting Protocol and Multiplicative Weights (MW).

### A. Estimating Guassian Means with Chernoff Bounds

To begin, a baseline for estimating means from external data sources under simple conditions without Byzantine behavior is established. The goal is to determine the minimum number of queries necessary to reliably estimate the mean of each external source's data distribution.

We consider $n$ external data sources, each corresponding to an independent real-valued random variable $x_i$ drawn from an unknown Gaussian distribution $\mathcal{N}(\mu_i, 1)$, where $0 \leq \mu_i \leq 1$. Each source has its own unique mean $\mu_i$, but all share a variance of 1. A single peer (without interaction with others) is tasked with estimating each $\mu_i$ up to a designated additive error $\varepsilon$, with high probability guarantees.

Specifically, for each source $i$, the peer queries the source multiple times and computes a sample mean $\hat{\mu}_i$. The estimation is considered successful if:

$$|\hat{\mu}_i - \mu_i| \leq \varepsilon$$

with probability at least $1-1/n^c$ for a chosen constant $c \geq 1$.

*1) Baseline Query Complexity:* Using Chernoff bounds for Gaussian random variables, it is shown that $m = O\left(\frac{c \log n}{\varepsilon^2}\right)$ queries per source are sufficient to achieve the desired accuracy guarantee for each individual mean. A union bound across all $n$ sources then implies that with probability at least $1 - 1/n$, all $n$ sample means simultaneously satisfy the target $\varepsilon$-deviation bound.

Thus, the baseline query complexity per source is $O(\log n)$, and the total query complexity over all sources is $O(n \log n)$.

### B. Error-Bounded Blacklisting Protocol

We now extend the baseline model to a distributed setting involving $k$ peers, of which at most a $\beta$-fraction may behave arbitrarily due to Byzantine faults. In this section, we assume $\beta < 1/3$, which aligns with the standard threshold that enables distributed consensus guarantees in the presence of adversarial agents. We describe a modification of the Download protocol introduced by Augustine et al.[2], where each peer previously queried each external source exactly once and blacklisted peers based on inconsistent reported values. In our variant, each peer instead performs $O(\log n)$ queries per source to obtain a more robust estimate of the mean, and blacklisting is based on deviations exceeding a carefully reduced threshold.

Each peer queries each external source $O(\log n)$ times, obtaining a personal estimate $\hat{\mu}_i^{(p)}$ for each source $i$. Using Chernoff bounds, each peer can guarantee that their estimate satisfies

$$|\hat{\mu}_i^{(p)} - \mu_i| \leq \varepsilon$$

with probability at least $1 - 1/(kn^c)$, for a chosen constant $c \geq 1$. This ensures that a union bound over all $n$ sources and $k$ peers yields an overall failure probability of at most $1/n$, even in the presence of up to $\beta k$ Byzantine peers.

Peers validate each other by exchanging their estimated sample means. A peer $p$ will blacklist another peer $q$ if $|\hat{\mu}_i^{(p)} - \hat{\mu}_i^{(q)}| > \varepsilon$ for any source $i$. However, because even honest peers' estimates can differ due to independent sampling error, it is necessary to modify the threshold to prevent false blacklisting.

*1) Shrinking the Error Thresholds:* To ensure that honest peers do not mistakenly blacklist each other, each peer's individual estimation guarantee is increased to an error tolerance of $\varepsilon/3$. That is, each peer ensures:

$$|\hat{\mu}_i^{(p)} - \mu_i| \leq \varepsilon/3$$

for all $i$ with high probability.

By the triangle inequality, this modification ensures:

$$|\hat{\mu}_i^{(p)} - \hat{\mu}_i^{(q)}| \leq |\hat{\mu}_i^{(p)} - \mu_i| + |\hat{\mu}_i^{(q)} - \mu_i| \leq 2\varepsilon/3$$

for any two honest peers $p$ and $q$. Thus, two honest peers will not falsely blacklist each other when blacklisting is triggered only if the reported means differ by more than $2\varepsilon/3$.

To ensure that no honest peer is falsely blacklisted while maintaining maximal sensitivity to Byzantine deviations, we set the estimation error bound for each peer to $\varepsilon/3$ and define the blacklisting threshold as $2\varepsilon/3$. These values are chosen to satisfy two key constraints: first, that the combined deviation between any two honest peers remains below the blacklisting threshold; and second, that the sum of the estimation error and the tolerated deviation does not exceed the error tolerance $\varepsilon$. This choice guarantees correctness while maximizing the gap between honest variation and malicious behavior.

This reduced error tolerance requires at most a constant factor increase in the number of queries per source, resulting in a per-peer query complexity of $O(n \log^2 n/k)$. Specifically, the number of queries per peer per source becomes

$$m = O\left(\frac{c \log n + \log k}{\varepsilon^2}\right),$$

where $c$ is a constant depending on the desired probability of success. The small additional $\log k$ term that would arise from union bounding over $k$ peers is negligible under the assumption that $k \ll n$, and is absorbed into the overall $O(\log n)$ term.

*2) Final Aggregation of Surviving Estimates:* After the blacklisting phase, each source $i$ has a surviving committee of peers whose sample means $\hat{\mu}_i^{(p)}$ were not flagged by honest peers. We denote the set of surviving sample means as:

$$S_i = \left\{ \hat{\mu}_i^{(q)} \;\middle|\; q \notin \texttt{blacklist}_i \right\}.$$

Due to the $\varepsilon/3$ estimation guarantee and blacklisting threshold of $2\varepsilon/3$, this set contains only estimates from honest peers with high probability. From the committee guarantees of Augustine et al.[2], the number of such surviving peers per source is at least:

$$s = \Theta(\log n)$$

with probability at least $1 - O(1/n)$.

*3) Simple Averaging Aggregator:* Given that all remaining estimates in $S_i$ are independent, sub-Gaussian, and bounded within $\varepsilon/3$ of the true mean $\mu_i$, we compute the final estimate via a simple average:

$$\hat{\mu}_i^{\text{final}} = \frac{1}{|S_i|} \sum_{q \notin \texttt{blacklist}_i} \hat{\mu}_i^{(q)}.$$

*4) Error Bound via Chernoff Concentration:* Because all values in $S_i$ are bounded and independent with mean $\mu_i$ and sub-Gaussian tails, we can apply standard concentration bounds to the average:

$$\mathbb{P}\left(|\hat{\mu}_i^{\text{final}} - \mu_i| > \delta_{\text{final}}\right) \leq \delta,$$

where

$$\delta_{\text{final}} = O\left(\sqrt{\frac{\log(1/\delta)}{s}}\right).$$

Setting $s = \Theta(\log n)$ and $\delta = 1/n$, we obtain:

$$\delta_{\text{final}} = O\left(\sqrt{\frac{\log n}{\log n}}\right) = O(1).$$

*5) Summary of Final Results:* Under the revised protocol, each peer performs $O(\log n)$ queries per source and estimates each mean $\mu_i$ to within an error tolerance of $\varepsilon/3$ using Chernoff bounds. Peers then exchange their local estimates and apply a blacklisting rule: a peer $q$ is blacklisted by peer $p$ if their reported estimates differ by more than $2\varepsilon/3$ for any source. This ensures that:

- No two honest peers blacklist each other (by triangle inequality).
- All or most Byzantine peers are excluded from the final aggregation.

Each peer participates in $O(n/k)$ committees and performs $O(\log n)$ queries per committee, resulting in a total per-peer query complexity of $O(n \log^2 n/k)$, or $\tilde{O}(n/k)$ ignoring logarithmic factors.

After blacklisting, aggregation is performed over the surviving set of sample means $S_i$, which contains $s = \Theta(\log n)$ honest values with high probability. Because these surviving means are individually bounded by $\varepsilon/3$ and drawn from sub-Gaussian sources, a simple average over $S_i$ suffices to achieve the final estimate:

$$\hat{\mu}_i^{\text{final}} = \frac{1}{|S_i|} \sum_{\hat{\mu}_i^{(q)} \in S_i} \hat{\mu}_i^{(q)}.$$

Applying concentration bounds, the average of $s = \Theta(\log n)$ sub-Gaussian estimates yields a deviation bound of:

$$\mathbb{P}\left(|\hat{\mu}_i^{\text{final}} - \mu_i| > \delta_{\text{final}}\right) \leq \frac{1}{n}, \quad \text{where } \delta_{\text{final}} = O(1).$$

Each final estimate remains within a constant additive error of the true mean with high probability. This matches the accuracy of centralized estimation, despite the presence of Byzantine agents and only partial information at each peer.

Thus, the protocol guarantees both robustness and near-optimal query complexity while achieving a constant bounded estimation error using only local samples and simple peer-to-peer verification.

### C. Multiplicative Weights

This project has taken a Multiplicative Weights (MW) approach to estimate an unknown sub-Gaussian distribution $P$ in the presence of Byzantine data points and data sources. The method aims to dynamically assign weights to data sources based on majority behavior, penalizing and suppressing the influence of consistently abnormal sources. The problem is composed as follows:

There are n external data sources providing the points $\{x_1, x_2, ...x_n\}$ from which the fraction (1-$\beta$) are honest sources that draw values from an unknown sub-Gaussian distribution $P$. The value $\beta$ represents the fraction of the sources which are Byzantine and provide arbitrary or possibly adversarial values.

*1) Variance:* We assume a variance ($\sigma^2$) of honest sources is either known or bounded by a universal constant, which enables normalization of deviations in the pseudo-loss function. When the variance is unknown in simulation, a robust estimate was obtained by using the Median Absolute Deviation (MAD) due to its resilience to Byzantine corruption. MAD is defined as:

$$MAD = \text{median}(|x_i - \text{median}(x)|)$$

*2) Aggregates:* The Multiplicative Weights framework functions by maintaining for each data source $i$ a weight $w_i(t)$ is initialized to 1, where $t$ is the time. At the beginning of each round the robust aggregate $\hat{\mu}(t)$ is calculated using the collection $\{x_1(t), x_2(t), ..., x_n(t)\}$.

The robust aggregate was calculated using the Median of Means (MoM) method as it is more robust to Byzantine than other methods. It functions by partitioning data into $k = O(\log(1/\delta))$ random buckets, and the aggregate $\hat{\mu}(t)$ is

obtained by taking the coordinate-wise median of the bucket means. MoM requires $\beta < 1/4$ corruption to guarantee concentration. Under these conditions, the aggregation method satisfies the high-probability concentration bound:

$$\mathbb{P}\left(|\hat{\mu}(t) - \mu| > O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)\right) \leq \delta.$$

Thus, $\hat{\mu}(t)$ can be treated as a reliable estimate of $\mu$ across rounds of the Multiplicative Weights update, ensuring effective separation between honest and adversarial sources.

In addition to the robust aggregation method, a standard Multiplicative Weights (MW) approach is considered where the aggregate $\hat{\mu}(t)$ is simply the weighted mean of all data points:

$$\hat{\mu}_{\text{MW}}(t) = \frac{\sum_{i=1}^{n} w_i(t)x_i(t)}{\sum_{i=1}^{n} w_i(t)}.$$

This method is highly responsive; deviations in the data immediately influence the center of mass, leading to potentially faster convergence when adversarial contamination is low or absent. However, standard MW aggregation is *not robust*, as adversarial sources can shift $\hat{\mu}_{\text{MW}}(t)$ significantly by concentrating their weight mass strategically. Provable concentration guarantees do not exist if the fraction of adversaries, $\beta$, is nontrivial. In the worst case, the center of mass itself becomes adversarially biased, misguiding the weight updates and allowing malicious points to survive. Thus, while standard MW may converge faster under benign conditions, it lacks any worst-case resilience guarantees and is vulnerable to collapse under moderate or sophisticated Byzantine attacks. However, both methods were examined in the simulation experimentation.

*3) Loss Function:* The pseudo-loss function, $\ell_i(t)$ is defined for theoretical purposes as the Softmin:

$$\ell_i(t) = \frac{e^{-(x_i(t)-\hat{\mu}(t))^2/c}}{\sum_j e^{-(x_i(t)-\hat{\mu}(t))^2/c}}$$

where $c$ is a scale parameter set to the estimated sub-Gaussian variance in simulation. However, in practice, the Softmin was less effective, leading to the use of:

$$\ell_i(t) = \min\left(1, \frac{|x_i(t) - \hat{\mu}(t)|}{c}\right)$$

*4) Weight Updates:* The weights are updated using the formula:

$$w_i(t+1) = w_i(t)(1 - \eta\ell_i(t))$$

where $\eta \in (0, 1)$ is a learning rate. The variable $\tau$ was set to be a blacklisting threshold should any data source's weight fall below it. However, experimentation was conducted with and without blacklisting data sources, as the weights of the data sources impact their overall influence over the robust aggregate.

*5) Analysis of Weight Decay:* It should be noted that the weight decay analysis is based on the conjecture that by using a correct loss function over time, honest nodes will trend towards $\ell_i(t) = 0$ and byzantine $\ell_i(t) = 1$ since byzantine nodes can lie arbitrarily far from $\hat{\mu}(t)$.

In honest sources, if $x_i(t)$ is drawn from $P$, then with high probability $|x_i(t) - \hat{\mu}(t)| = O(\sqrt{\log(1/\delta)/n})$. Choosing $c$ appropriately ensures that $\ell_i(t) \approx 0$ for honest sources. Thus, their weights satisfy

$$w_i(t+1) \approx w_i(t)(1 - \eta \cdot 0) = w_i(t).$$

That is, honest weights remain approximately stable over time.

Conversely, adversarial sources produce $x_i(t)$ that can lie arbitrarily far from $\hat{\mu}(t)$. Hence, for many rounds,

$$\ell_i(t) \approx 1,$$

leading to updates

$$w_i(t+1) = w_i(t)(1 - \eta).$$

Thus, after $T$ rounds,

$$w_i(T) \approx (1 - \eta)^T.$$

To ensure $w_i(T) \leq \tau$ (blacklisting threshold), we require

$$T \geq \frac{\log(\tau)}{\log(1-\eta)} \approx \frac{1}{\eta}\log\left(\frac{1}{\tau}\right).$$

*6) Chernoff Bound for Honest Weight Preservation:* To ensure that honest sources are not mistakenly blacklisted, we apply a Chernoff bound to control their cumulative pseudo-loss over time.

Assume that for an honest source $i$, the pseudo-loss $\ell_i(t)$ is a bounded random variable satisfying $\ell_i(t) \in [0, 1]$ and $\mathbb{E}[\ell_i(t)] \leq \epsilon$, where $\epsilon$ is small.

Define the cumulative pseudo-loss up to round $T$:

$$S_i(T) = \sum_{t=1}^{T} \ell_i(t).$$

Then, using LOE:

$$\mathbb{E}[S_i(T)] = \mathbb{E}[\sum_{t=1}^{T} \ell_i(t)] = \sum_{t=1}^{T} \mathbb{E}[\ell_i(t)]$$

Since $\mathbb{E}[\ell_i(t)] \leq \epsilon$, we can apply:

$$\sum_{t=1}^{T} \mathbb{E}[\ell_i(t)] \leq \sum_{t=1}^{T} \epsilon$$

$$\mathbb{E}[S_i(T)] \leq \epsilon T.$$

Applying the multiplicative Chernoff bound for sums of bounded random variables:

$$\mathbb{P}\left(S_i(T) > (1+\delta)\epsilon T\right) \leq \exp\left(-\frac{\delta^2 \epsilon T}{3}\right).$$

*7) Theoretical Guarantees:* To establish theoretical guarantees, we can utilize the results of the Chernoff bound for honest weights. Let $\epsilon, \delta > 0$ and assume that for each honest source $i$, the expected pseudo-loss satisfies $\mathbb{E}[\ell_i(t)] \leq \epsilon$ uniformly over time. Then, for

$$T = O\left(\frac{1}{\epsilon} \log\left(\frac{n}{\delta}\right)\right)$$

rounds, with probability at least $1 - \delta$, all honest sources retain weights above the blacklisting threshold $\tau$.

We can prove that the above is true by defining $S_i(T) = \sum_{t=1}^{T} \ell_i(t)$ as the cumulative pseudo-loss of source $i$ up to round $T$. Since $\ell_i(t) \in [0, 1]$ and $\mathbb{E}[\ell_i(t)] \leq \epsilon$, application of the multiplicative Chernoff bound yields:

$$\mathbb{P}\left(S_i(T) > (1+\delta')\epsilon T\right) \leq \exp\left(-\frac{\delta'^2 \epsilon T}{3}\right).$$

Choosing $\delta' = 1$ gives:

$$\mathbb{P}\left(S_i(T) > 2\epsilon T\right) \leq \exp\left(-\frac{\epsilon T}{3}\right).$$

To ensure this failure probability is at most $\delta/n$, we require

$$T \geq \frac{3}{\epsilon} \log\left(\frac{n}{\delta}\right).$$

Applying a union bound across all $n$ sources guarantees that with probability at least $1 - \delta$, all honest sources satisfy $S_i(T) \leq 2\epsilon T$.

For the weight updates, observe that

$$w_i(T) = \prod_{t=1}^{T} (1 - \eta \ell_i(t)) \geq (1 - \eta)^{S_i(T)}.$$

Using the inequality $1 - \eta \geq e^{-2\eta}$ valid for $\eta$ sufficiently small, we deduce

$$w_i(T) \geq e^{-2\eta S_i(T)} \geq e^{-4\eta \epsilon T}.$$

Thus, selecting $\eta$ and $T$ to satisfy

$$e^{-4\eta \epsilon T} > \tau$$

ensures that the weights of honest sources remain above the blacklisting threshold throughout.

*8) Blacklisting Byzantine Sources Convergence Time:* To create a theoretical convergence time for blacklisting, assume that $\tau$ is the blacklisting threshold and $\eta$ is the MW learning rate. Then any Byzantine source with persistent deviation will be blacklisted after at most

$$T = O\left(\frac{1}{\eta} \log\left(\frac{1}{\tau}\right)\right)$$

rounds.

To prove this claim, assume that for a Byzantine source, the pseudo-loss $\ell_i(t)$ is approximately 1 at each round, due to the adversarial deviations typically being large compared to the honest consensus.

Using this premise, the weight evolution satisfies:

$$w_i(t+1) = w_i(t)(1 - \eta \ell_i(t)) \approx w_i(t)(1 - \eta).$$

After $T$ rounds:

$$w_i(T) \approx (1 - \eta)^T.$$

Taking logarithms:

$$\log w_i(T) = T \log(1 - \eta).$$

Using the inequality $\log(1 - \eta) \leq -\eta$ for $\eta \in (0, 1)$:

$$\log w_i(T) \leq -\eta T,$$

$$w_i(T) \leq \exp(-\eta T).$$

The goal is $w_i(T) \leq \tau$, where $\tau$ is the blacklisting threshold. Setting $\exp(-\eta T) \leq \tau$ and taking the natural logarithm:

$$-\eta T \leq \log(\tau),$$

Thus:

$$T \geq \frac{1}{\eta} \log\left(\frac{1}{\tau}\right).$$

Hence, the convergence time to blacklist a Byzantine source is $T = O\left(\frac{1}{\eta} \log\left(\frac{1}{\tau}\right)\right)$.

*9) Pseudocode:* In our simulation, the algorithm was implemented as follows:

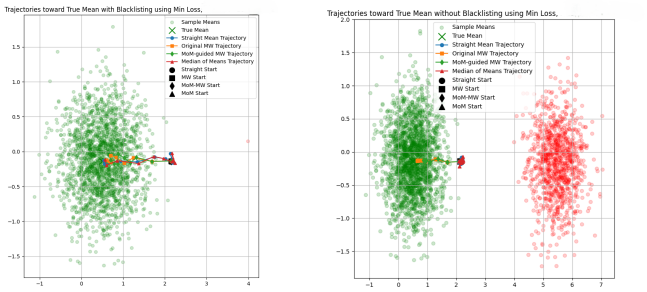
**Algorithm 1** Multiplicative Weights Robust Estimation

Initialize $w_i(0) \leftarrow 1$ for all sources $i$.
**for** each time step $t = 1$ to $T$ **do**
  Collect incoming data points $x_1(t), \ldots, x_n(t)$.
  Compute robust estimate $\hat{\mu}(t)$ (median, trimmed mean, or median of means).
  **for** each source $i$ **do**
    Compute pseudo-loss $\ell_i(t) \leftarrow \min\left(1, \frac{|x_i(t) - \hat{\mu}(t)|}{c}\right)$.
    Update weight $w_i(t+1) \leftarrow w_i(t)(1 - \eta \ell_i(t))$.
    **if** Blacklisting **then**
      **if** $w_i(t+1) < \tau$ **then**
        Blacklist source $i$.
      **end if**
    **end if**
  **end for**
**end for**

## IV. Results

Using a simulation to test based on the pseudocode previously shown, the theoretical results were compared with empirical results. The testing was carried out with and without explicit blacklisting. That is, in some test runs, the sources deemed Byzantine were removed entirely from the equation. In other cases, the weights were allowed to significantly reduce the influence of the Byzantine sources, rather than to remove them entirely.

Figure 1 illustrates a blacklisting test shown in addition to a non-blacklisting test in which the green and red clusters represent the population of honest and Byzantine sources, respectively. As can be seen, the MW algorithm effectively adjusts the mean towards the true mean, despite Byzantine data sources that skew the aggregate mean. However, in the case of blacklisting the less robust methods of statistical analysis, such as a basic mean, also move towards the true mean as expected. The lack of a red cluster in the blacklisting example is evidence of the algorithm to identify and eliminate Byzantine sources.




(a) Trajectory of aggregate methods towards the true mean while using blacklisting of Byzantine data sources.

(b) Trajectory of aggregate methods towards the true mean without using blacklisting of Byzantine data sources.

Fig. 1: Trajectory towards true mean.

Figures 2 and 3 show the loss of the aggregate methods over ten rounds while using blacklisting methods and non-blacklisting methods. These results are another way to visualize the effects of blacklisting on the aggregate methods used. It can be seen that the blacklisting method works more effectively to reduce the error in non-robust methods but does not improve performance in MW assisted methods, which indicates that the weights of the Byzantine agents are being reduced sufficiently to remove influence on the system with or without blacklisting.

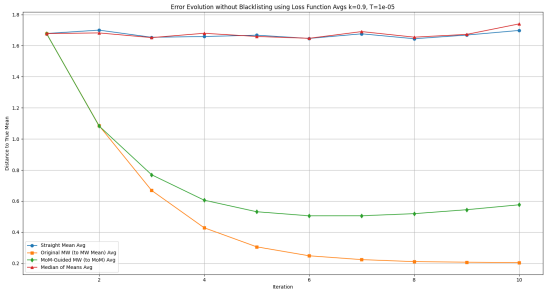



Fig. 2: Evaluation of aggregate methods in reaching the true mean without using blacklisting of Byzantine data sources.

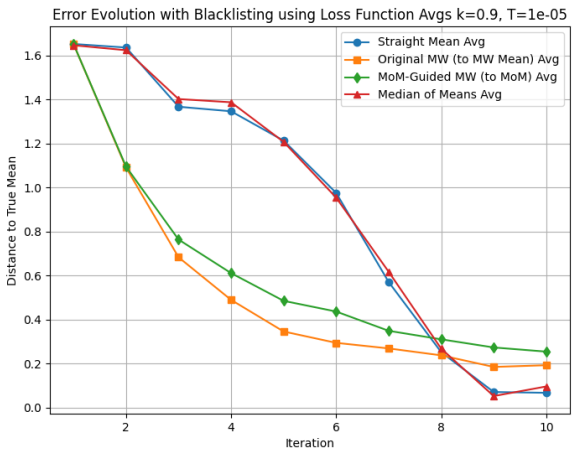



Fig. 3: Evaluation of aggregate methods in reaching the true mean using blacklisting of Byzantine data sources.

Figure 4 illustrates the likelihood of failure, particularly that Byzantine sources are significantly more likely to fail, since with the correct eta and tau, the probability of being discovered is 1. In contrast, honest sources are much less likely to be identified as malicious, reaching a worst-case probability of 0.6.

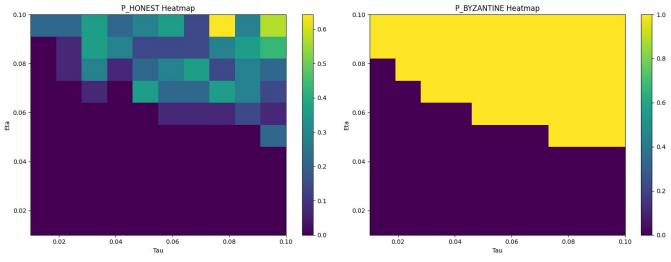



Fig. 4: Comparison of the blacklisting threshold values ($\tau$) and learning rate ($\eta$) controlling the probability of failure between Byzantine and honest data sources.

Figure 5 depicts the number of rounds needed to reach the best-case curves compared to theoretical expectations and empirical results. It illustrates that the theoretical and empirical round that must occur in order to reach the best weight curves differ by only 1.

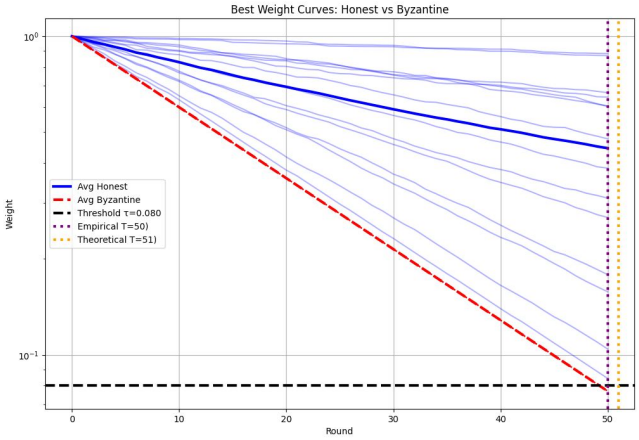



Fig. 5: Graph showing the best weight curves in honest and Byzantine processes to analyze empirical vs theoretical round in which they are reached.

Figure 6 illustrates the comparison between theoretical and empirical rounds to achieve the best case convergence time, supporting the theoretical hypothesis. This figure underscores the similarity of the theoretical expectations and empirical results.

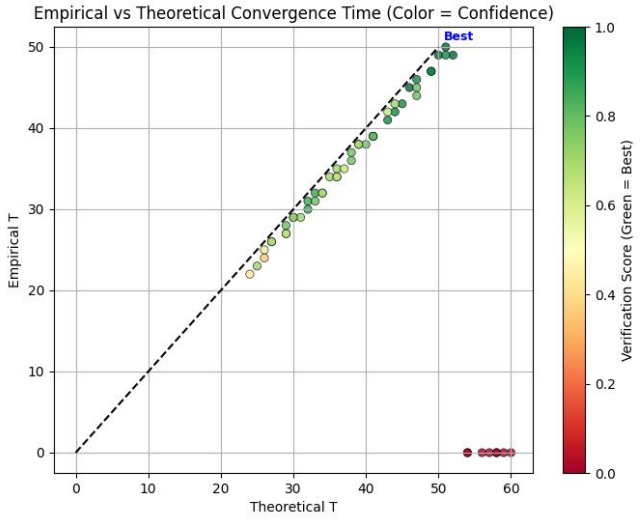



Fig. 6: Comparison of theoretical and empirical rounds to reach convergence.

Figure 7 depicts the ideal line that empirical values should follow to reach convergence based on theoretical calculations. As can be seen, the results of the simulation follow theoretical expectations very closely. The generated fitted line has a $R^2$ value of 0.9981, which is extremely well fitted and visually clear that the results were on par with expectations. Table I shows the slope, intercept, and $R^2$ of the fitted line.

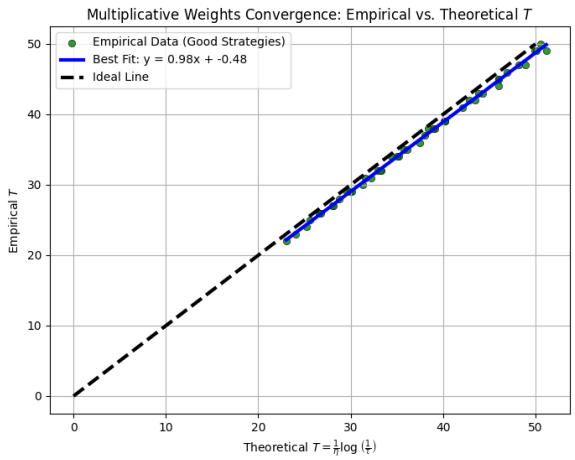



Fig. 7: Comparison of ideal line and the fitted line on the empirical data.

| Metric | Value |
|---|---|
| Slope | 0.9844 |
| Intercept | -0.4758 |
| $R^2$ | 0.9981 |

TABLE I: Data analysis metrics from best fit line on empirical data.

## V. Limitations

While the proposed methods demonstrate promising theoretical and empirical results, several challenges were encountered throughout the project, showing opportunities for further refinement and improvement.

First, although the Multiplicative Weights (MW) algorithm is appealing in theory due to its adaptability and convergence guarantees, its practical implementation proved to be more complex. In real-world settings, the performance of MW is sensitive to parameter choices, particularly the learning rate and loss function formulation. Small miscalibrations can lead to unstable weight updates or insufficient separation between honest and adversarial sources, which limits the robustness of the consensus process.

Second, the effectiveness of the Error-Bounded Blacklisting Protocol depends on the selection of appropriate thresholds for both estimation error and peer divergence. In environments where data distributions are unknown or highly variable, determining a universally optimal blacklisting threshold is challenging. A threshold that is too strict can result in honest peers being incorrectly blacklisted due to natural statistical variance, while a threshold that is too lenient allows adversarial peers to continue influencing consensus outcomes. This trade-off makes it challenging to balance responsiveness to malicious behavior with the risk of penalizing honest nodes, especially in unpredictable or unstructured data settings.

Together, these challenges reveal the difficulty in constructing Byzantine-resilient consensus mechanisms that operate without manual parameter tuning in adversarial settings with unpredictable external data.

## VI. Future Work

There are several directions in which this work can be extended. First, future efforts could focus on fully integrating the Multiplicative Weights framework into Augustine et al.'s Data Retrieval model. This would require adapting MW to function in a synchronous, round-based environment, where peers could iteratively adjust source weights in response to observed deviations. Such integration could enable a more unified protocol that combines the adaptability of MW with the DR model's formal guarantees, supporting dynamic trust evolution and consensus over time.

Another extension involves enhancing the Error-Bounded Blacklisting Protocol through the use of more robust statistical aggregators, such as the Median of Means (MoM).[4] This could improve the protocol's tolerance to noise or adversarial manipulation by ensuring that outlier estimates have limited influence on blacklist decisions. Formal analysis of the aggregate's deviation from the true source mean would be essential to establish tighter error bounds and inform the design of blacklisting thresholds that maintain both precision and resilience.

Additionally, a more thorough empirical analysis of the blacklisting mechanism would be beneficial. Simulations could explore the trade-offs between query complexity, false positive rates, and resilience as parameters such as $\epsilon$, the number of peers, and the fraction of Byzantine nodes vary. These experiments would enable the calibration of threshold parameters more accurately and validate theoretical assumptions in more realistic environments.

It would also be valuable to evaluate the system's behavior in scenarios where the fraction of Byzantine peers is below the classical $1/3$ threshold. Understanding the performance and failure modes in these more favorable conditions could provide insight into the boundaries of the protocol's resilience and help identify opportunities for practical optimization. This analysis is relevant to both the blacklisting framework and to the MW-based filtering method, as each

reacts differently under varying levels of adversarial influence.

Beyond the synchronous setting of the DR model, future work could explore how these techniques generalize to asynchronous or partially synchronous environments, where peers do not share a global clock and message delays are unpredictable. Adapting blacklisting and weighting strategies to asynchronous settings introduces additional challenges, particularly in maintaining consistent trust evaluations across peers. However, doing so would significantly broaden the applicability of the protocol to real-world systems, such as permissionless blockchains and decentralized oracles, where timing cannot be globally coordinated.

Finally, a potential extension is the development of a fully online consensus mechanism that combines MW updates with adaptive blacklisting in a decentralized manner. In such a system, trust scores and estimates would evolve continuously as new data arrives, and peers would dynamically reconfigure their behavior to reflect the evolving reliability of information sources. This would enable Byzantine resilient consensus over external, randomized data sources in settings where both the data and the adversary are adaptive and unpredictable.

## VII. Conclusion

This research presented a framework for achieving Byzantine resilient consensus with external, randomized data sources. Building on Augustine et al.'s Data Retrieval model, we introduced two complementary techniques: an Error-Bounded Blacklisting Protocol and a Multiplicative Weights (MW) approach for robust mean estimation. Both methods are designed to tolerate adversarial behavior by dynamically adjusting peer influence or rejecting inconsistent reports based on statistical guarantees.

The blacklisting protocol enhances resilience by allowing peers to estimate source means through repeated sampling and identify inconsistencies usually carefully chosen thresholds grounded in concentration bounds. In parallel, the MW framework reduces the impact of Byzantine sources over time by adaptively decreasing the weight of sources that consistently deviate from robust aggregates. Statistical tools, Median of Means (MoM) and Median Absolute Deviation (MAD), were used to mitigate the influence of outliers, offering added robustness in unpredictable or non-Gaussian settings.

Theoretical analysis provided probabilistic guarantees on convergence, honest weight preservation, and Byzantine source suppression. Empirical simulations focused on the MW method, confirming its effectiveness and close alignment with theoretical expectations across a variety of adversarial conditions. While the blacklisting protocol was not directly evaluated in simulation, its development is grounded in rigorous statistical analysis and could be explored in future empirical testing.

Although there is room for further refinement, particularly in parameter tuning, asynchronous settings, and adaptive trust integration, this work provides a groundwork for achieving robust distributed consensus over noisy and untrusted external data. The combined use of adaptive weighting and statistical validation demonstrates a practical approach to building fault-tolerant protocols for decentralized systems operating under adversarial conditions.

## REFERENCES


[1] Buke Ao et al. "On Precision Bound of Distributed Fault-Tolerant Sensor Fusion Algorithms". In: *ACM Computing Surveys* 49.1 (May 2016). Accessed: 2025-04-01, Article 5, 23 pages. DOI: 10.1145/2898984. URL: http://dx.doi.org/10.1145/2898984.

[2] John Augustine et al. "Byzantine Resilient Distributed Computing on External Data". In: *38th International Symposium on Distributed Computing (DISC 2024)*. Ed. by Dan Alistarh. Vol. 319. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024, 3:1–3:23. ISBN: 978-3-95977-352-2. DOI: 10.4230/LIPIcs.DISC.2024.3. URL: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.DISC.2024.3.

[3] Christian Cachin and Jovana Zanolini. "From Symmetric to Asymmetric Asynchronous Byzantine Consensus". In: *arXiv preprint arXiv:2005.08795v2* (2020). URL: https://arxiv.org/abs/2005.08795v2.

[4] Yen-Chi Chen. *A Short Note on the Median-of-Means Estimator*. https://faculty.washington.edu/yenchic/short_note/note_MoM.pdf. Accessed: 2025-04-01. 2020.

[5] Ruomu Hou et al. *Randomized View Reconciliation in Permissionless Distributed Systems*. Tech. rep. TR12/17. Accessed: 2025-04-01. Computing 1, 13 Computing Drive, Singapore 117417: The National University of Singapore, School of Computing, 2017. URL: https://dl.comp.nus.edu.sg/server/api/core/bitstreams/d4468729-7b25-4059-bf41-6244d59fc4ca/content.

[6] Shang-En Huang, Seth Pettie, and Leqi Zhu. "Byzantine Agreement with Optimal Resilience via Statistical Fraud Detection". In: *Proceedings of the ACM Symposium on Principles of Distributed Computing (PODC)* (2024). Accessed: 2025-04-01. DOI: 10.1145/3639454. URL: https://dl.acm.org/doi/abs/10.1145/3639454.

[7] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Accessed: 2025-04-01. Cambridge University Press, 2005. ISBN: 9780521835404. URL: https://doi.org/10.1017/CBO9780511813603.

[8] Jared Saia. *Bitcoin Consensus*. https://www.cs.unm.edu/~saia/classes/591-EconDistrib-s25/lec/lec-BitcoinConsensus.pdf. Lecture notes for CS 591 - Distributed Econ, University of New Mexico. Accessed: 2025-04-01. 2025.