March 2024 Suspected Black Marble Flooding Against Monero: Privacy, User Experience, and Countermeasures

Draft v0.1

Rucknium

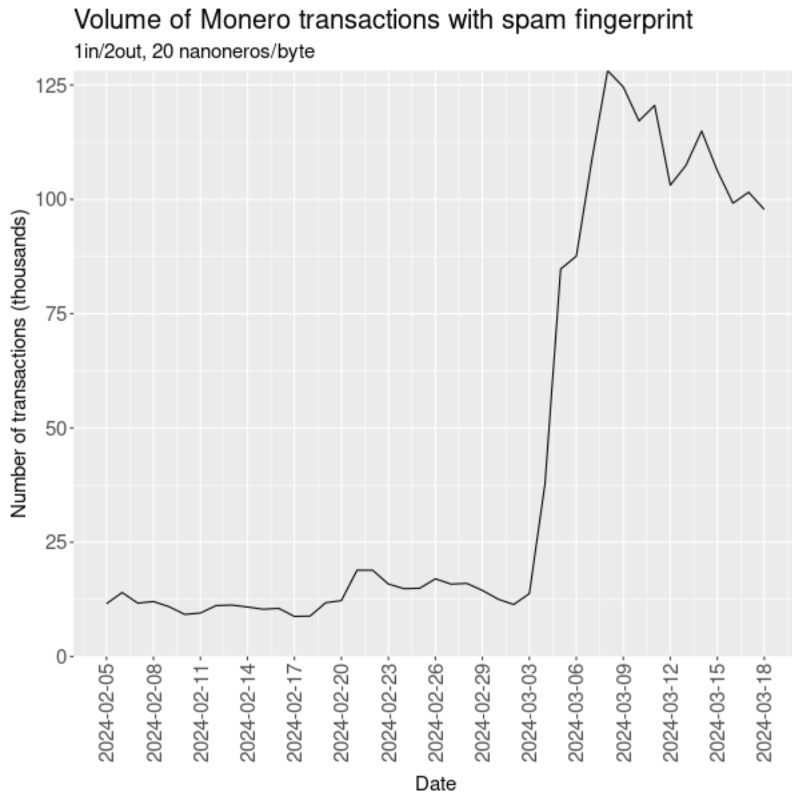
March 20, 2024

Abstract

On March 4, 2024, aggregate Monero transaction volume suddenly almost tripled. This note analyzes the effect of the large number of transactions, assuming that the transaction volume is an attempted black marble flooding attack by an adversary. According to my estimates, mean effective ring size has decreased from 16 to 5.5 if the black marble flooding hypothesis is correct. At current transaction volumes, the suspected spam transactions probably cannot be used for “chain reaction” analysis to eliminate all ring members except for the real spend for a large number of rings. Effects of increasing Monero’s ring size above 16 are analyzed.

1 March 4, 2024: Sudden transaction volume

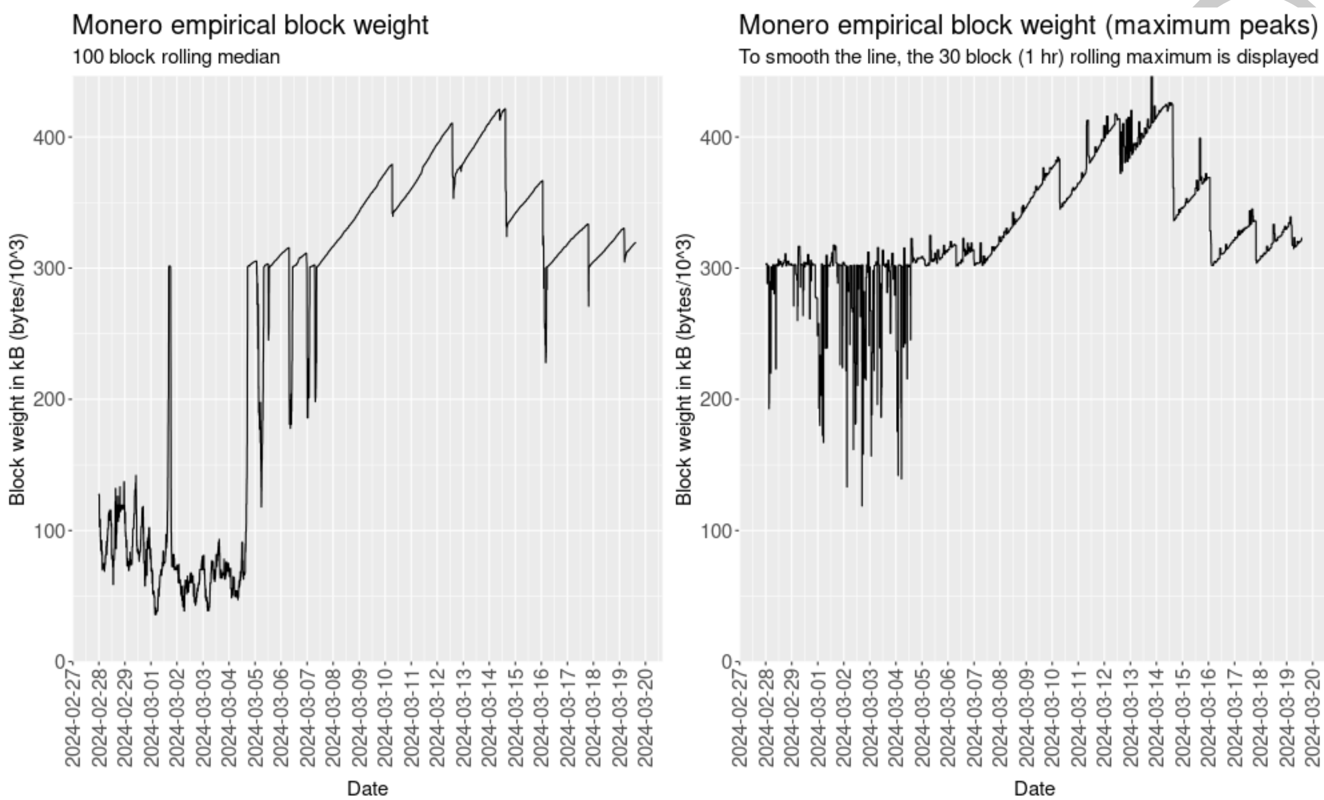
Figure 1: Volume of Monero transactions with spam fingerprint



On March 4, 2024 at approximately block height 3097764 (15:21:24 UTC), the number of 1input/2output minimum fee (20 nanoneros/byte) transactions sent to the Monero network rapidly increased. Figure 1 shows daily volume of this type of transaction increasing from about 15,000 to over 100,000.

The large volume of these transactions was enough to entirely fill the 300 kB Monero blocks mined about every two minutes. Monero's dynamic block size algorithm activated. The 100 block rolling median block size slowly increased to adjust for the larger number of transactions that miners could pack in blocks. Figure 2 shows the adjustment. The high transaction volume raised the 100 block median gradually for period of time. Then the transaction volume reduced just enough to allow the 100 block median to reset to a lower level. Then the process would restart. Block sizes have usually remained between 300 kB and 400 kB. Occasionally, high-fee transactions would allow miners to get more total revenue by giving up some of the 0.6 XMR/block tail emission and including more transactions in a block. The "maximum peaks" plot shows this phenomenon.

Figure 2: Monero empirical block weight



The sudden transaction volume rise may originate from a single entity. The motive may be spamming transactions to bloat the blockchain size, increase transaction confirmation times for real users, perform a network stress test, or execute a black marble flooding attack to reduce the privacy of Monero users. I will focus most of my analysis on the last possibility.

2 Literature review

The very first research bulletin released by the Monero Research Lab described black marble transaction flooding. [Noether et al., 2014] points out that the ring signature privacy model requires rings to contain transaction outputs that are could be plausible real spends. If a single entity owns a large share of outputs (spent or not), it can use its knowledge to rule out ring members in other users' transactions that cannot be the real spend. Since the entity knows that itself did not spend the output(s) in a particular ring, the effective ring size that protects other users' privacy can be reduced — even to an effective ring size of 1 when the entity knows the real spend with certainty. Rings with known real spends can be leveraged to determine the real spend in other rings in a "chain reaction" attack.

[Noether et al., 2014] gave the name "black marble" to the outputs owned by an anti-privacy adversary since they modeled the problem using a marble draw problem with a hypergeometric distribution. When a specific number of marbles are drawn *without* replacement from an urn containing a specific number of white and black marbles, the hypergeometric distribution describes the probability of drawing a specific number of black marbles. In my modeling I use the binomial distribution, which is the same as the hypergeometric except marbles are drawn *with* replacement. The binomial distribution makes more sense now ten years after [Noether et al., 2014] was written. The total number of RingCT outputs on the blockchain that can be included in a ring is over 90 million. The hypergeometric distribution converges to the binomial distribution as the total number of marbles increases to infinity. Moreover, Monero's current decoy selection algorithm does not select all outputs with equal probability. More recent outputs are selected with much higher probability. The hypergeometric distribution cannot be used when individual marbles have unequal probability of being selected.

[Chervinski et al., 2021] simulates a realistic black marble flood attack. They consider two scenarios. The adversary could create 2input/16output transactions to maximize the number of black marble outputs per block or the adversary could create 2input/2output transactions to make the attack less obvious. The paper uses Monero transaction data from 2020 to set the estimated number of real outputs and kB per block at 41 outputs and 51 kB respectively. The nominal ring size at this time was 11. The researchers simulated filling the remaining 249 kB of the 300 kB block with black marble transactions. A "chain reaction" algorithm was used to boost the effectiveness of the attack. In the 2in/2out scenario, the real spend could be deduced (effective ring size 1) in 11% of rings after one month of spamming black marbles. Later I will compare the results of this simulation with the current suspected spam incident.

[Krawiec-Thayer et al., 2021] analyze a suspected spam incident in July-August 2021. Transactions' inputs, outputs, fees, and ring member ages were plotted to evaluate evidence that a single entity created the spam. The analysis concluded, "All signs point towards a single entity. While transaction homogeneity is a strong clue, a the [sic] input consumption patterns are more conclusive. In the case of organic growth due to independent entities, we would expect the typically semi-correlated trends across different input counts, and no correlation between independent users' wallets. During the anomaly, we instead observed an extremely atypical spike in 1–2 input txns with no appreciable increase in 4+ input transactions."

TODO: A few papers like [Ronge et al., 2021, Egger et al., 2022] discuss black marble attacks tool

3 Black marble theory

The binomial distribution describes the probability of drawing x number of "successful" items when drawing a total of n items when the probability of a successful draw is p. It can be used to model the number of transaction outputs selected by the decoy selection algorithm that are not controlled by a suspected adversary.

The probability mass function of the binomial distribution with $n \in \{0, 1, 2, \ldots\}$ number of draws and $p \in [0, 1]$ probability of success is

$$f(x,n,p) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ where } \binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{1}$$

The expected value (the theoretical mean) of a random variable with a binomial distribution is np.

Monero's standard decoy selection algorithm programmed in `wallet2` does not select outputs with equal probability. The probability of selecting each output depends on the age of the output. Specifics are in [citation]. The probability of a single draw selecting an output that is not owned by the adversary, p_r, is equal to the share of the probability mass function occupied by those outputs: $p_r = \sum_{i \in R} g(i)$, where R is the set of outputs owned by real users and $g(x)$ is the probability mass function of the decoy selection algorithm.

3.1 Spam assumptions

There is some set of criteria that identifies suspected spam. The early March 2024 suspected spam transactions: 1) have one input; 2) have two outputs; 3) pay the minimum 20 nanoneros per byte transaction fee. The normal volume of these transactions produced by real users must be estimated. The volume in excess of the normal volume is assumed to be spam. I followed this procedure:

1. Compute the mean number of daily transactions that fit the suspected spam criteria for the four weeks that preceded the suspected spam incident. A separate mean was calculated for each day of the week (Monday, Tuesday,...) because Monero transaction volumes have weekly cycles. These volume means are denoted $v_{r,m}, v_{r,t}, v_{r,w}, \ldots$ for the days of the week.

2. For each day of the suspected spam interval, sum the number of transactions that fit the suspected spam criteria. Subtract the amounts found in step (1) from this sum, matching on the day of the week. This provides the estimated number of spam transactions for each day: $v_{s,1}, v_{s,2}, v_{s,3}, \ldots$

3. For each day of the suspected spam interval, randomly select $v_{s,t}$ transactions from the set of transactions that fit the suspected spam criteria, without replacement. This randomly selected set is assumed to be the true spam transactions.

4. During the period of time of the spam incident, compute the expected probability p_r that one output drawn from the `wallet2` decoy distribution will select an output owned by a real user (instead of the adversary) when the wallet constructs a ring at the point in time when the blockchain tip is at height h. [the closed form formula is in x]

5. The expected effective ring size of each ring constructed at block height h is $1 + 15 \cdot p_r$. The coefficient on p_r is the number of decoys.

Figure 3 shows the results of this methodology. The mean effective ring size settled at about 5.5 by the fifth day of the large transaction volume. On March 12 and 13 there was a large increase in the number of 1input/2output transactions that paid 320 nanoneros/byte (the third fee tier). This could have been the spammer switching fee level temporarily or a service that uses Monero increasing fees to avoid delays. I used the same method to estimate the spam volume of these 320 nanoneros/byte suspected spam. The 1in/2out 320 nanoneros/byte transactions displaced some of the 1in/2out 20 nanoneros/byte transactions because miners preferred to put transactions with higher fees into blocks. Other graphs and analysis will consider only the 1in/2out 20 nanoneros/byte transactions as spam unless indicated otherwise.

Figure 3: Estimated mean effective ring size

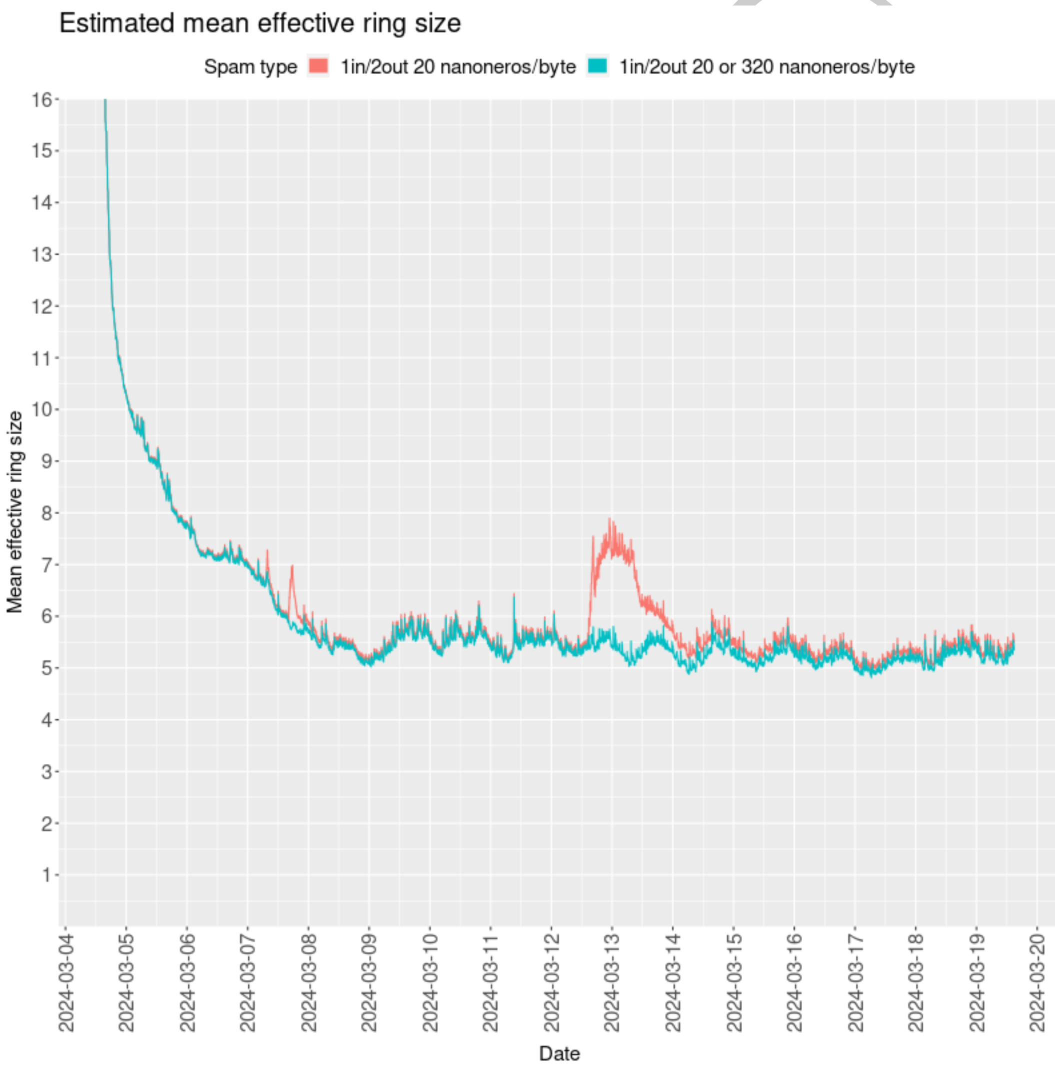
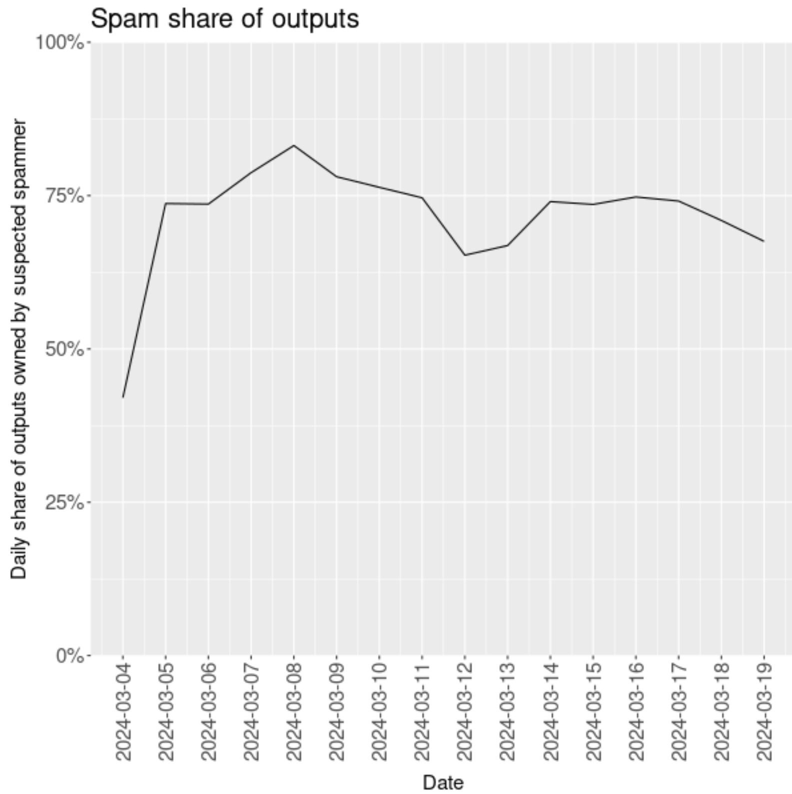


Figure 4 shows the daily share of outputs on the blockchain that are owned by the suspected spammer. The mean share of outputs since the suspected spam started is about 75 percent.

Figure 4: Spam share of outputs



3.2 Long term projection scenarios at different ring sizes

Fix the number of outputs owned by real users at r. The analysis will let the number s of outputs owned by the adversary vary. The share of outputs owned by real users is

$$p_r = \frac{r}{r+s} \tag{2}$$

The 2 expression can be written $p_r = \frac{1}{r} \cdot \dfrac{r}{1+\frac{1}{r}s}$, which is the formula for hyperbolic decay with the additional $\frac{1}{r}$ coefficient at the beginning of the expression [Aguado et al., 2010].

Let n be the nominal ring size (16 in Monero version 0.18). The number of decoys chosen by the decoy selection algorithm is $n-1$. The mean effective ring size for a real user's ring is one (the real spend) plus the ring's expected number of decoys owned by other real users.

$$\mathrm{E}\left[n_e\right] = 1 + (n-1) \cdot \frac{r}{r+s} \tag{3}$$

The empirical analysis of Section 3.1 considered the fact that the `wallet2` decoy selection algorithm draws a small number of decoys from the pre-spam era. Now we will assume that the spam incident has continued for a very long time and all but a negligible number of decoys are selected from the spam era. We will hold constant the non-spam transactions and vary the number of spam transactions and the ring size. Figures 5, 6, and 7 show the results of the simulations.

Figure 5: Long-term projected mean effective ring size

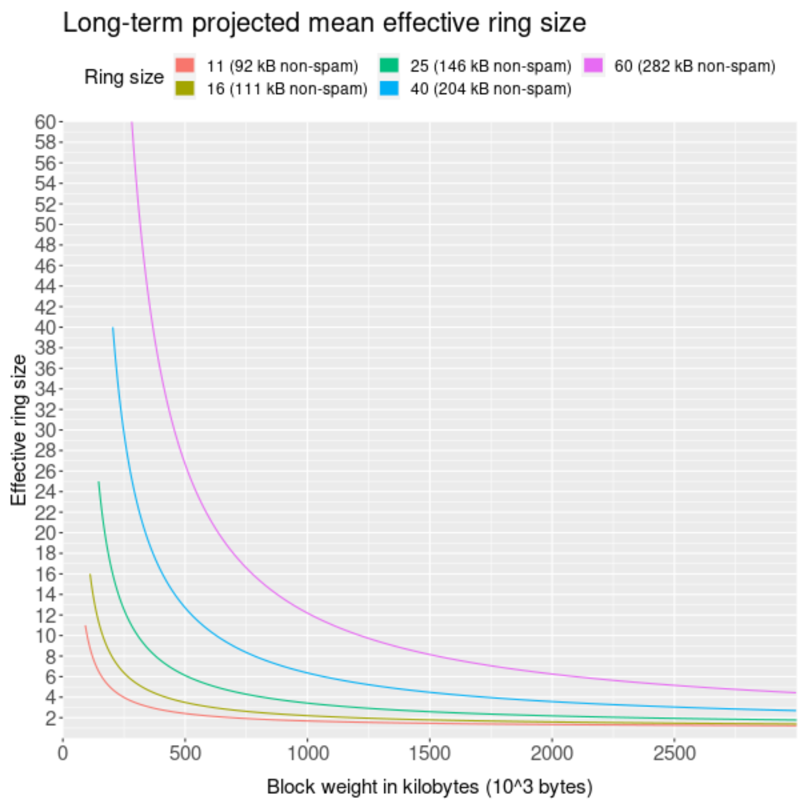


Figure 6: Long-term projected mean effective ring size (log-log scale)

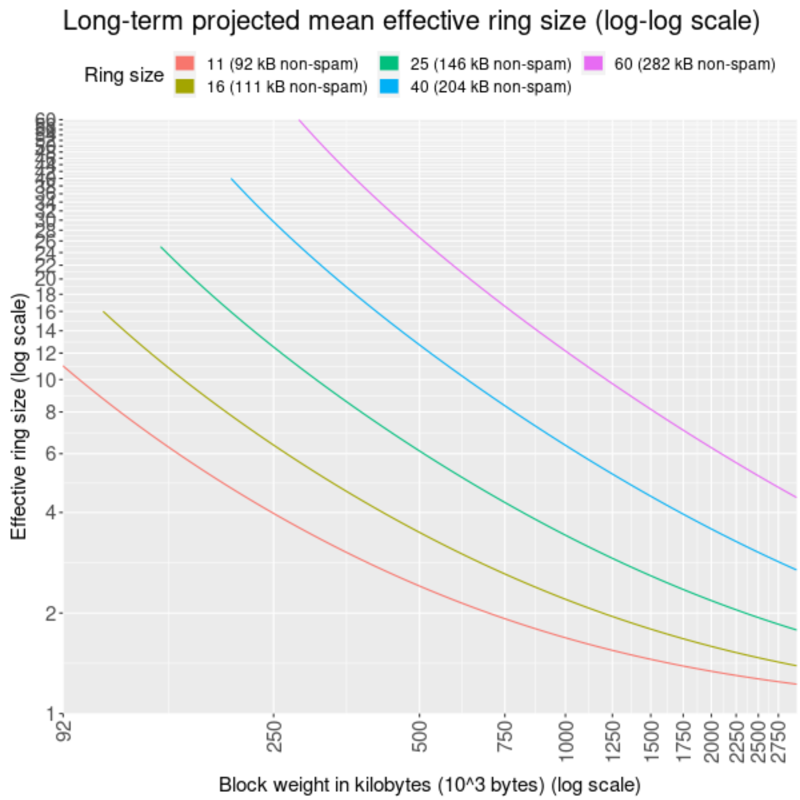
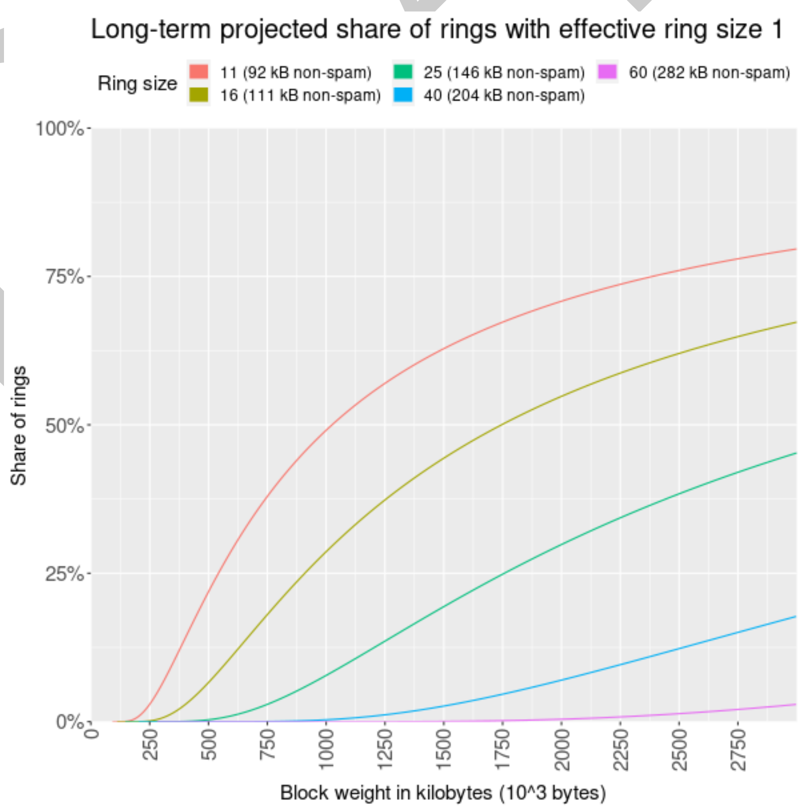


Figure 7: Long-term projected share of rings with effective ring size 1



3.3 Guessing the real spend using a black marble flooder's simple classifier

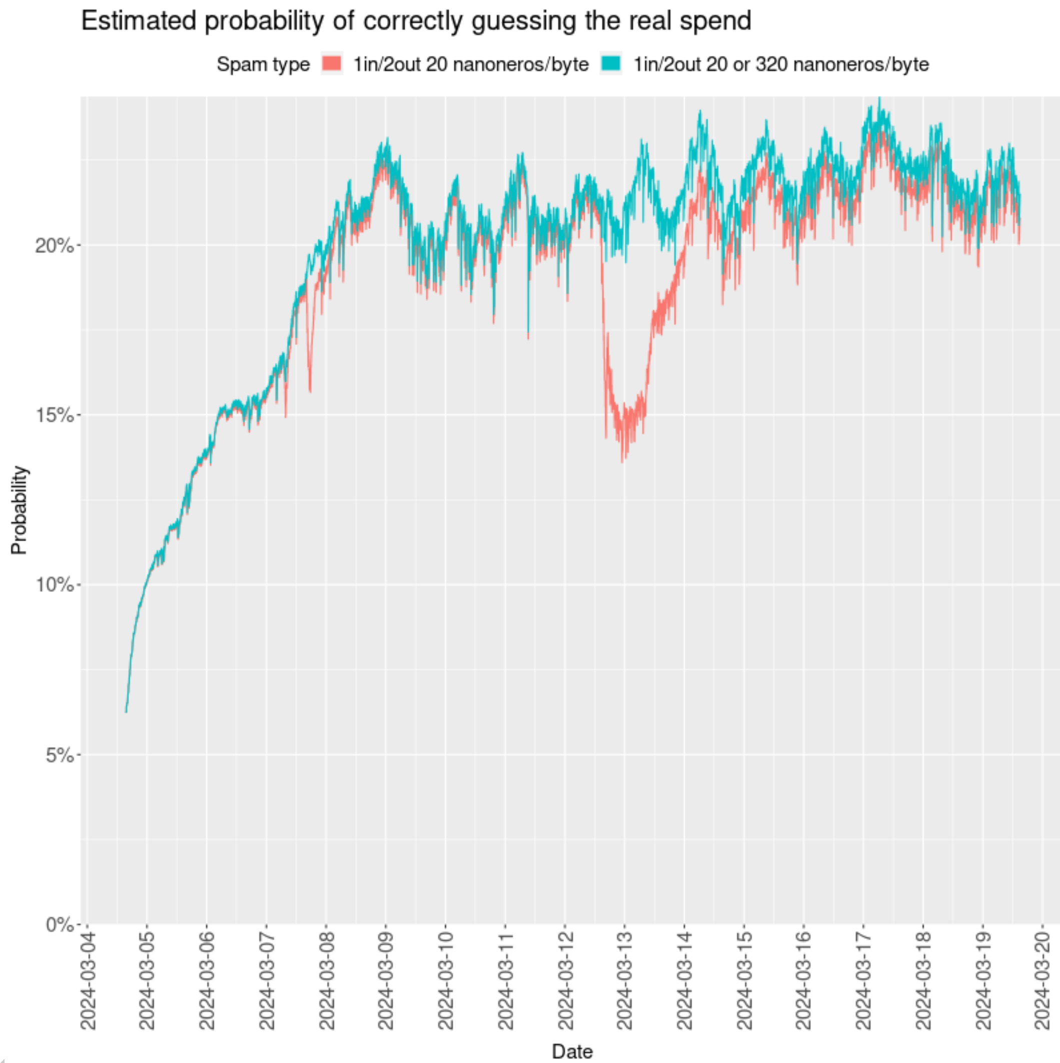
The adversary carrying out a black marble flooding attack could use a simple classifier to try to guess the real spend: Let n be nominal ring size and n_s be the number of outputs in a given ring that are owned by the attacker. n_s is a random variable because decoy selection is a random process. The adversary can eliminate n_s of the n ring members as possible real spends. The attacker guesses randomly with uniform probability that the ith ring member of the $n - n_s$ remaining ring members is the real spend. The probability of correctly guessing the real spend is $\frac{1}{n-n_s}$. If the adversary owns all ring members except for one ring member, which must be the real spend, the probability of correctly guessing the real spend is 100%. If the adversary owns all except two ring members, the probability of correctly guessing is 50%. And so forth.

The mean effective ring size is $\mathrm{E}\left[n_e\right]$ from 3. Does this mean that the mean probability of correctly guessing the real spend is $\frac{1}{\mathrm{E}[n_e]}$? No. The $h(x) = \frac{1}{x}$ function is strictly convex. By Jensen's inequality, $\mathrm{E}\left[\frac{1}{n_e}\right] > \frac{1}{\mathrm{E}[n_e]}$. The mean probability of correctly guessing the real spend is

$$\mathrm{E}\left[\frac{1}{n_e}\right] = \sum_{i=1}^{n} \frac{1}{i} \cdot f(i-1, n-1, \frac{\mathrm{E}\left[n_e\right]-1}{n-1}) \tag{4}$$

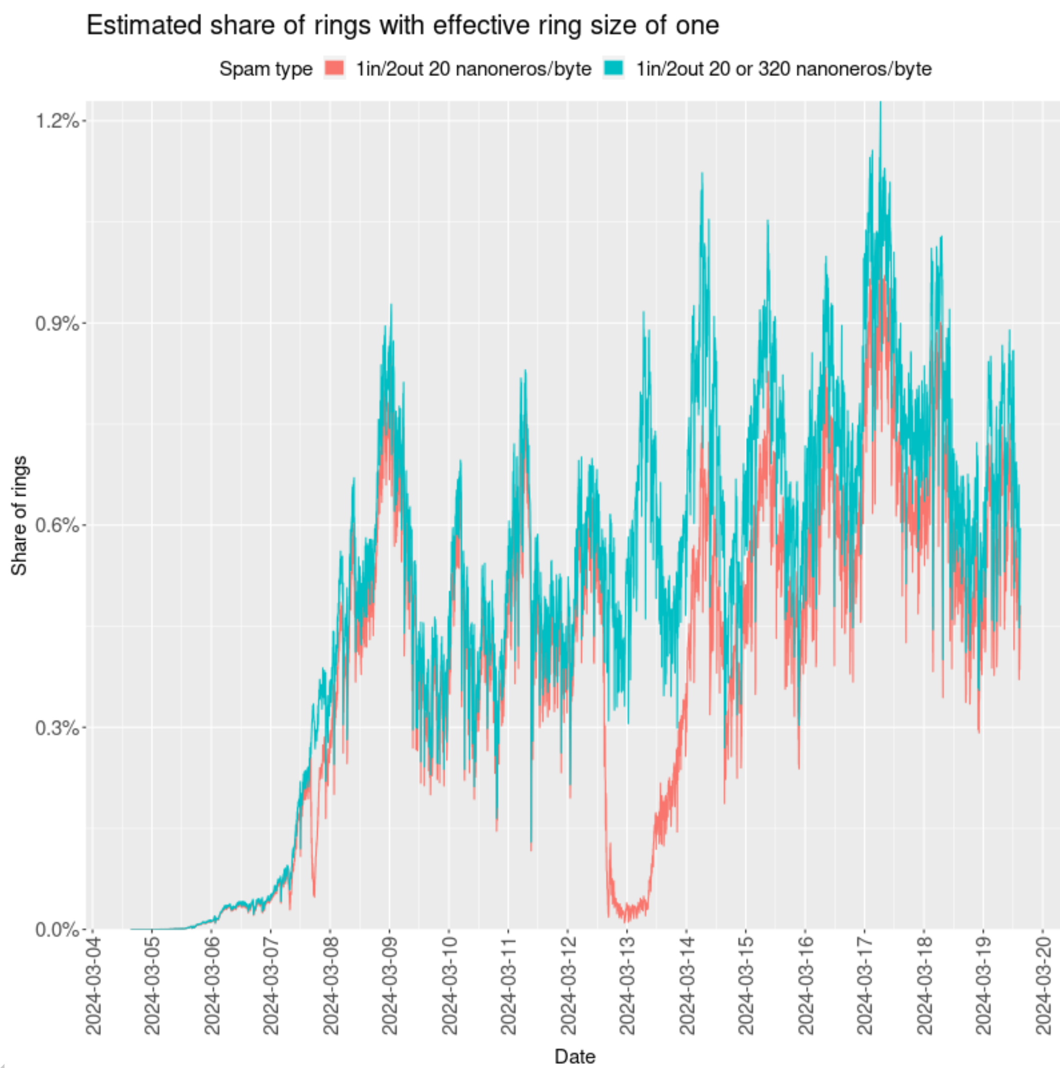
$\frac{1}{i}$ is the probability of correctly guessing the real spend when the effective ring size is i. f is the probability mass function of the binomial distribution. It calculates the probability of the decoy selection algorithm selecting $i-1$ decoys that are owned by real users. The total number of decoys to select is $n-1$ (that is the argument in the second position of f). The probability of selecting a decoy owned by a real user is $\frac{\mathrm{E}[n_e]-1}{n-1} = \frac{r}{r+s}$.

Figure 8: Estimated probability of correctly guessing the real spend



The probability of a given ring having all adversary-owned ring members except for the real spend is $f\left(0, n-1, \frac{\mathrm{E}[n_e]-1}{n-1}\right)$. Figure 9 plots the estimated share of rings with effective ring size one.

Figure 9: Estimated share of rings with effective ring size of one



4 Chain reaction graph attacks

TODO

5 Countermeasures

See https://github.com/monero-project/research-lab/issues/119

TODO

6 Estimated cost to suspected spammer

1in/2out 20 nanoneros/byte spam definition: 42.5 XMR in total fees. 2.1 GB total size of transactions.

1in/2out 20 and 320 nanoneros/byte spam definition: 47.6 XMR in total fees. 2.2 GB total size of transactions.

TODO

7 Transaction confirmation delay

TODO

8 Real user fee behavior

TODO

References

[Aguado et al., 2010] Aguado, J., Cid, C., Saiz, E., & Cerrato, Y. (2010). Hyperbolic decay of the dst index during the recovery phase of intense geomagnetic storms. *Journal of Geophysical Research: Space Physics*, 115(A7). `https://doi.org/https://doi.org/10.1029/2009JA014658`

[Chervinski et al., 2021] Chervinski, O. J., Kreutz, D., & Yu, J. (2021). Analysis of transaction flooding attacks against monero. *2021 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, 1–8. `https://doi.org/10.1109/ICBC51069.2021.9461084`

[Egger et al., 2022] Egger, C., Lai, R. W. F., Ronge, V., Woo, I. K. Y., & Yin, H. H. F. (2022). On defeating graph analysis of anonymous transactions. *Proceedings on Privacy Enhancing Technologies*, 2022(3). `https://petsymposium.org/2022/files/papers/issue3/popets-2022-0085.pdf`

[Krawiec-Thayer et al., 2021] Krawiec-Thayer, M. P., Neptune, Rucknium, Jberman, & Carrington (2021). *Fingerprinting a flood: forensic statistical analysis of the mid-2021 monero transaction volume anomaly.* `https://mitchellpkt.medium.com/fingerprinting-a-flood-forensic-statistical-analysis-of-the-mid-2021-monero-transaction-volume-a` Available at https://mitchellpkt.medium.com/fingerprinting-a-flood-forensic-statistical-analysis-of-the-mid-2021-monero-transaction-volume-a19cbf41ce60

[Noether et al., 2014] Noether, S., Noether, S., & Mackenzie, A. (2014). *A note on chain reactions in traceability in cryptonote 2.0.* Research Bulletin. `https://www.getmonero.org/resources/research-lab/pubs/MRL-0001.pdf`

[Ronge et al., 2021] Ronge, V., Egger, C., Lai, R. W. F., Schröder, D., & Yin, H. H. F. (2021). Foundations of ring sampling. *Proceedings on Privacy Enhancing Technologies*, 2021(3), 265–288. `https://doi.org/doi:10.2478/popets-2021-0047`