Epistemology of Decoy Systems; Probing the Attacks on the Privacy of the Monero Blockchain

Nathan Borggren

CompDec

(Dated: May 10, 2023)

In an effort to address adversarial attacks on the privacy of the Monero blockchain, in which an adversary utilizes leaked or privileged information to gain information about the global network, a mathematical and accompanying computational environment will be developed that is amenable to analytical, empirical, and simulated analysis. The proposed work will quantify, measure, and analyze the structure and sensitivity Monero has with respect to the fungibility of the Monero coins, the connectivity of merchants and their customers, and how this structure can change as the microeconomics of repeated transactions is correlated and associated. Results, code, and recommendations will be shared with the community with the goal of quantifying, qualifying and improving its privacy capabilities.

I. INTRODUCTION

This proposal seeks financial and technical support to provide analysis in abstracto, in silico, and ultimately in situ of Monero's privacy features. We seek to address the concerns raised about decoy systems in [1]. The particular attacks raised in that work are the

- **Overseer Attack** - In which a merchant (receiver) uses their privileged transaction information to deduce repeat customers and build up a consumer model of their customers. This may already be a severe problem given the fraction of transactions that involve an exchange as one of the counter-parties.

- **Flashlight Attack** - A sender repeats transactions to an address or collection of addresses, providing them the means to link stealth addresses in the blockchain, and follow the taint tree forward to identify funds being aggregated (perhaps at a subpoenaable exchange).

- **Tainted Dust Attack** - Similar to a Flashlight Attack, but may involve a large number of candidate addresses, with an effort of identifying which address leads to a particular user-pattern.

though other related information leaks can also occur:

- **Anomalous or unique wallet configurations.** Users may inadvertently identify themselves by making unique choices for RingCT or number of outputs, use of flags, additional information etc. In previous machine learning blockchain analysis, features involving these choices were a primary source of leaks, as evidenced by their preference by random forests, neural networks, and other classification schemes.

- **Anomalous or unique decoy selection.** If a wallet or user uses non-standard decoy selection, the distributions used can often be teased apart from standard decoy selections by common statistical means.

- **Broadcasting of private information**. ShapeShift's API leaked a mass of information for a number of years. Mining pools tend to broadcast information about coinbase transactions.

- **Mordinals and anomalous** *tx_extra***.** Some users are indifferent or otherwise misaligned from the privacy use-case of Monero and seek to immortalize themselves on the blockchain. This behavior has a network effect of reducing the effective ring size of subsequent transactions.

All of these memory leaks share the common properties of leaking information about the ring inputs, the tx outputs, or both. If the decoy system is working ideally and as intended then these effects should be localized and protect both senders and receivers from each other as well as third parties. However, much work needs to be done to quantify, qualify, and improve the systems. This proposal seeks to do so with a Three-Pronged Approach: Theoretical, Computational, and Experimental.

The theoretical framework will build upon previous graph theoretical work which is largely concerned with asymptotic and/or stationary behavior. The computational work proposes to use statistical machinery like pymc, to efficiently explore a vast number of scenarios in a controlled manner, allowing information to be incorporated into the models in a Bayesian sense and simulate the effect of introducing dramatic changes. The empirical framework will take advantage of the test network and if deemed prudent, the real network to analyze the transaction graphs as they appear on actual blockchains.

An advantage of the multi-pronged approach is to incorporate other interested parties into the analysis as will be discussed in the final *Proposal Request* section.

II. THEORETICAL FRAMEWORK

Previously there has been graph theoretical work into graph-based attacks that have provided some comfort into the viability of decoy systems [2, 3], but more effort needs to be made to see this established in the wild. In particular the privacy loss from repeated transactions with another party needs to be quantified and measured. Additional privacy features beyond decoys might be necessary.

Topological Data Analysis (TDA) is an approach to algebraic topology that prioritizes the analysis of real world data. [4, 5] As graphs are the '1-skeleton' of the simplicial structures of algebraic topology a connection can be made with the graph theory of previous work. In addition to the vertices and edges present in graph theory, TDA allows for higher order simplices to be constructed. For example faces consisting of three vertices (V_1, V_2, V_3) can

aid in incorporating the additional information gained in situations like the Overseer Attack and the Eve-Alice-Bob-Eve (EABE) Attack and related situations.

To be more clear consider a merchant who has received payment from two different stealth addresses. Only the merchant can know that the public blockchain info (t_1, v_1) and (t_2, v_2) can be extended to include (t_1, v_2), (t_2, v_1), (t_1, t_2, v_2), (t_1, t_2, v_1), (t_1, v_1, v_2), (t_2, v_1, v_2), and (t_1, t_2, v_1, v_2). This additional structure can shorten path lengths, change connectivity and alter other topological variants non-trivially. All of this additional information advantages the merchant seeking to draw inferences about their clients.

The advantage is a technical one, opening up connections to mathematical formalisms, as well as computational tools. These tools include further graph analysis [6, 7] and multiple python packages [8, 9]. It is our belief that Graph Theory, (including bipartite graphs and hyper graphs), Topological Data Analysis, Optimal Transport, and probabilistic methods are the most capable of revealing any cracks that may be present in the Monero facade.

Additionally these structures will be able to serve as *domains* for functional calculations. Probability and statistics can be introduced in various ways. One such approach is that of Optimal Transport. In this context the transport is of the coins themselves from their genesis in a coinbase transaction to the space of users. The coin values play the role of mass in the theory. We are imagining the equiprobable choice of priors, where the decoy systems are behaving optimally, to be an important reference frame to evaluate the enhancement to the probabilities that the information gains bring. We also expect that although there are a vast number of paths, the paths still cluster around this reference frame. This situation occurs in statistical mechanics out of equilibrium and is the analog of Maximum Entropy in the non-equilibrium context, known as Maximum Caliber.

An additional statistical choice is that of Bayesian Networks, where distributions are defined over the various structures: like the edges, vertices, and faces. For example, although a transaction is of unknown value on the Monero blockchain, this very transaction value could be the subject of a leaked transaction. Even without the leak the transaction value is constrained by the number of coinbase transactions that eventually lead to its inputs, as decoys or otherwise. The Bayesian framework is a natural choice to incorporate the information gain from this conditioned information. We will elaborate in the computational framework section where we introduce the computational tools to actually perform these tasks [10].

III. COMPUTATIONAL FRAMEWORK

The power of TDA comes through its use of Persistent Homology. A filtration is introduced to enable the calculation of a persistence diagram which quantifies noise and structures at all hierarchical levels. In the setting of Monero the scalar parameter behind this filtration can be the 'probability of being the real transaction.' As this probability parameter is sweeped from zero to one various topological features come and go. The interval of the parameter where these features 'persist' is the subject of persistent homology. These persistence diagrams reside in a metric space where the distance is given by the Wasserstein distance. It will be the object of the proposed research to formalize these ideas but we would like to sketch our initial approach.

Although the graph of the entire Monero blockchain is large, it may be within the realm of computational feasibility that a graph analysis is performed on the entire blockchain. Furthermore the subgraphs that arise out of the aided information in the case of the overseer attacks would be tractable computationally and we can presume readily available to an adversary, thus it is necessary to understand what information could have been leaked.

To move some of the theoretical work into the realm of a messy reality various packages are available, some of which will also be useful and appropriate in the Empirical Framework section. In my previous explorations of the subject I felt as though having rings composed of entirely decoys (removing the ability to know for certain that some input was being spent) or having decoy outputs (red-herring outputs where the entire subsequent taint tree is obfuscating the truth) would be very problematic for blockchain analysts. With two outputs, both parties are of microeconomic interest; both were involved in the transaction. Adding additional decoy outputs not only provides more choices in subsequent transactions, but insures that these choices are working towards the obfuscation of the transaction graph. These and other hypothesis can be explored.

IV. EMPIRICAL FRAMEWORK

In a previous study [11], [12] we staged a Monero blockchain with 10-50 agents interacting with each other to probe connectivity of the resulting transactions. The machine learning therein was followed up on in another Magic Grant funded project [13]. I think the similar

research going forward should utilize the test network rather than building a new blockchain from scratch. I would also like to draw attention to the correlation results therein, it signifies a new attack not addressed by the decoy system in use.

The Monero test network and stage net will be used to perform an empirical analysis of example Overseer attacks to avoid transaction costs and other considerations. Cat-and-Mouse games will be setup to mimic various microeconomic considerations in analogy to the attacks:

- Sender repeats transactions to a Receiver on a schedule.
- Receiver aggregates a collection of previous transactions on a timed schedule
- Receiver aggregates a collection of previous transactions when a threshold amount has been reached.
- Sender sends transactions to a collection of addresses in hopes to identify which addresses belong to a particular user.

the resulting transaction graphs can then be analyzed and algorithm performance evaluated. It is possible that the decoy selection process may require such graph analysis, drawing from specific clusters in order to increase the fungibility between the clusters. We aim to connect and contribute to other ongoing Monero Research Lab research as fitting and able.

If the effort can be done safely, so as not to compromise existing consumers, the real Monero blockchain can also be analyzed. Motifs can be identified and histogrammed and previous leaks and residue from previous forks can be analyzed to see to what extent the blockchain has already been compromised. Churning strategies can be recommended or decoy selection can be motivated to diffuse this information loss over the continuing blockchain.

V. INTEGRATING THE FRAMEWORKS

Although it is the subject of the research to fully formalize these ideas, we can begin to see how the research may unfold and how the insights from the different approaches integrated. The information garnished from these analysis are of different varietals, deterministic, Bayesian, frequentist. Some aspects of which are more important depending on the questions being asked and the motives of those asking them. For example some level of

de-anonymization might be sufficient to get a picture of what the microeconomic activity occurring on the blockchain is like but insufficient to identify or link the specific entities involved.

The graph theoretical and TDA approaches, in their basic form, are deterministic theories, they can potentially reveal certainties with respect to connectivity. The Monero blockchain itself, after many statistical choices like decoy selection, stealth address creation are made become part of Natural History, a fossil record of the specific things that occurred. It is a basic premise of Monero that 'the one true subgraph' consisting of all real transactions between real users can be sufficiently obfuscated with decoys, hidden in plain sight, so as to be rendered indiscoverable. The performance of this hypothesis in real world conditions versus powerful mathematical and computational machinery is what we are seeking to address. These conditions involve zooming into a potentially very small subset of transactions and diligently using all available resources to prune this tree into a clear picture of economic activity.

Indeed the fear behind the Overseer Attack is that these 'red-handed' connections can be made given enough leaked information. In the same way a user has certainty over the transactions involved in their own activity, can an exchange achieve that same level of certainty on the origins of some funds through the systematic pruning of the edges their privileged information enables? This question can be asked deterministically and addressed through theoretical and empirical means. In which case the graph/simplex decompositions results in either a unique explanation of the scenario or a subgraph which still has a host of decoys present. This latter case is the more general case and the one anticipated by myself and others.

As mentioned in [1] this level of certainty isn't obligatory for the purveyors of these attacks and they may proceed with their queries at the level of probability of their choosing. At this stage we introduce probabilities and ask how the weights have changed on the decoy sets with the conditioned information. In the setting of Overseer Attack this investigation may take on similarities of the motif-hunting in [6] with the added calculation of likelihoods that these motifs and their subgraphs originated from the decoy selection process. A narrative of accuracy also begins when these probabilities get introduced. A bias towards recall over precision can be present in the case where False positives are of little importance; if the unscrupulous-attack-purveyor is not as much concerned with busting down the wrong door

so much as the time wasted in the process. It is thus important to evaluate algorithms over the spectrum of relative priority between false positives and false negatives; algorithms with a high rate of false positive can still be quite dangerous for victims of these attacks.

It is most natural to think of how that evaluation process would proceed in the empirical framework section. We will actually try to execute a statistically significant number of these attacks on our own test network. Having the blockchain data as well as the private wallet data will allow us to evaluate the performance of yet-to-be-determined algorithms and mathematics and considering the false positive and false negative rates. Simulating the scenarios as well will allow us to establish a ground truth off of which our recommendations can be tested and evaluated.

VI. PROPOSAL REQUEST

The author is seeking a rate of 2/3 of 1 XMR/hour or approximately \$100/*hour* and has the means to put up to a full time effort (30-40 hours/week) in June-July 2023 and a part time effort thereafter of approximately 15-20 hours/week. This rate is slightly less than the author's usual rate for data science work at the PhD level, but commensurate with previous funding from the Magic Fund. The rate is also less than comparable work at Chainalysis, Elementus, TRM Labs, and US Government whose ongoing work on Monero analysis does not necessarily feedback into the improvement of the protocol and indeed can be antithetical to the currency's objective. The total request is **220 XMR** over three months. Although this total is larger than a typical reward perhaps it can be combined with CCS or other mechanisms and is a competitive bid for the requested work.

This timeline is short compared to usual peer-reviewed research timescales but I think is sufficient amount of time to understand the breadth and depth of the issues and inform the Monero community of a path forward. It should also be clear at the end of this time if the Monero Foundation is interested into continuing the research, at which time new proposals can be submitted and further funding arrangements made as it sees fit.

I'd like to acknowledge the anonymous Monero community members that reached out to others and myself in order to create a small community of Monero/Graph Theory/Statisticians who have expressed interest and capabilities towards the problems these adversarial attacks present. Although this proposal does not seek particulars of the funding

arrangements for those researchers, I will be grateful and open to their continuous feedback throughout the research process and contributing feedback to their efforts. My research efforts for the entire duration of the proposed work will be transparent to this working group to facilitate this feedback loop.

Part of the partitioning into *Theoretical, Statistical, and Empirical* components is to be able to scale the effort in a multitude of directions based on the needs of the community as well as the interests and capabilities of the other interested researchers. Having lead teams of data scientists from prestigious universities with skill levels from undergraduate, graduate, post-doc and even professorial levels, I feel equipped and available to lead the effort of scaling these analysis to whatever level the community deems appropriate.

VII. UPDATE

Versions of this proposal were released to the working group mentioned as well as the Monero Research Lab matrix channel. Isthmus, Xmrack and Rucknium responded with feedback. There seems to be a preference for using the transaction graph of the real network as opposed to the test network. My worry of compromising other transactions through the analysis of our own seemed not to be shared. Doing the tests on the real network would be fine but would take cooperation and some Monero to provide those transactions. I don't personally want to steward Monero beyond dust transactions for this purpose, but perhaps the foundation or users who are willing to share transaction history from their wallets can help play the role of E, A, as well as B in the EABE attacks. It is my belief at this stage that the patterns created by these attacks will be consistent on either the test network or the real network, but that the patterns may be more 'hidden' with the added transactions on the real network.

My assertion about the timeline being short was also seconded, and third-ed. Analogous government contracts would likely be for a year, and indeed 6 months to 1 year might be more reasonable for true resolution of these issues. I am just trying to be consistent with the previous award amounts, but would be delighted to continue the work for longer at the same rate. I would certainly agree to **440 XMR** over 6 months as well, but wanted to give the foundation the option to make an assessment early in the research, and also an opportunity for the researchers that are enthusiastic about helping me with this work to apply and be

compensated for their efforts as well.

---

[1] Blockchain Privacy; Equal Parts Theory and Practice, 2023. URL `https://zfnd.org/blockchain-privacy-equal-parts-theory-and-practice/`.

[2] Christoph Egger, Russell WF Lai, Viktoria Ronge, Ivy KY Woo, and Hoover HF Yin. On defeating graph analysis of anonymous transactions. *Cryptology ePrint Archive*, 2022.

[3] Saravanan Vijayakumaran. Analysis of cryptonote transaction graphs using the dulmage-mendelsohn decomposition. *Cryptology ePrint Archive*, 2021.

[4] R. Rabadan and A.J. Blumberg. *Topological Data Analysis for Genomics and Evolution: Topology in Biology*. Cambridge University Press, 2019. ISBN 9781107159549.

[5] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2): 255–308, 2009.

[6] Stephen Ranshous, Cliff A Joslyn, Sean Kreyling, Kathleen Nowak, Nagiza F Samatova, Curtis L West, and Samuel Winters. Exchange pattern mining in the bitcoin transaction directed hypergraph. In *Financial Cryptography and Data Security: FC 2017 International Workshops, WAHC, BITCOIN, VOTING, WTSC, and TA, Sliema, Malta, April 7, 2017, Revised Selected Papers 21*, pages 248–263. Springer, 2017.

[7] Aric Hagberg and Drew Conway. Networkx: Network analysis with python. *URL: https://networkx. github. io*, 2020.

[8] Guillaume Tauzin, Umberto Lupo, Lewis Tunstall, Julian Burella Pérez, Matteo Caorsi, Anibal M Medina-Mardones, Alberto Dassatti, and Kathryn Hess. giotto-tda: A topological data analysis toolkit for machine learning and data exploration. *The Journal of Machine Learning Research*, 22(1):1834–1839, 2021.

[9] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. Pot: Python optimal transport. *The Journal of Machine Learning Research*, 22(1):3571–3578, 2021.

[10] Anand Patil, David Huard, and Christopher J Fonnesbeck. Pymc: Bayesian stochastic modelling in python. *Journal of statistical software*, 35(4):1, 2010.

[11] N. Borggren, H.-Y. Kim, L. Yao, and G. Koplik. Simulated blockchains for machine learning

traceability and transaction values in the Monero network, 2020.

[12] N. Borggren and L. Yao. Correlations of multi-input Monero transactions, 2020.

[13] An empirical analysis of monero's ring signature resilience to artificially intelligent attacks, 2022. URL https://github.com/MAGICGrants/Monero-Fund/issues/15#issuecomment-1086122008.