Platonic.Systems

for Project Ardana

Audit

Authors
Quinn Dougherty quinn.dougherty@platonic.systems
Foo Bar foo.bar@platonic.systems

Supervisor
Isaac Shapira isaac.shapira@platonic.systems

September 28, 2021

Table of Contents

0.1 Preamble to the audit

0.2 On datastream integrity

There are a couple places in the Ardana ecosystem that suggest **interaction with live datastreams**. A variety of software concerns such as API integration practices or keeping a stream open are ignored in the current document.

One of these such places is the oracle/bot from the governance layer. The oracle is to consume *third party* price data from something like coinbase, binance, etc. and produce transaction signatures.

The second such place is a proposal Morgan and I discussed on discord[1] for DEX subpool size limits. If we choose to do something functional/dynamic, rather than constant/static, for the subpool size limits the most principled choice is to *infer* them from data. > We can determine it empirically using data such as the data cited here[2] and analytical formulas for determining the pool sizes required to reach slippage targets for different swap sizes (Morgan)

While it's absolutely an option to download a static dataset once (and refresh it every few months or so), the natural question is *do we better serve the project by implementing online learning?* Here I am assuming that somewhere in the literature a formula is written down, making the actual model dead simple. But since the possibility of live datastream has been floated, I want to enumerate some of the pitfalls.

I will provide a few threatmodels that arise when a datastream is interacted with "online".

[1] https://discord.com/channels/844383474676662292/844387861251751978/885557683745349632

[2] https://www.mechanism.capital/liquidity-targeting/

Threatmodel 1: third parties compromised

The idea is *we do not trust the third party data sources*. Attackers here fall into two camps: those who are trying to corrupt the beliefs of the whole market and those who are trying to corrupt our beliefs in particular.

I'm envisioning something like the creation of artificial (even fraudulent) arbitrage opportunities by directly perturbing coinbase/binance's beliefs through hacking, and Ardana's behavior is *downstream* of coinbase/binance's beliefs because of where in our system we interact with their APIs.

Suppose we implement the online learning version of subpool size limit selection. An attacker may be able to arbitrage on pool sizes somehow iff they can force an irrational choice of pool sizes (and if our pool sizes are downstream of coinbase/binance data, all they'd need to do is hack into coinbase/binance).

Mitigation

Does coinbase/binance have a notification system that goes out to API users when they detect a breach? If so, we should handle it almost as an error and fallback to the last uncorrupted snapshot when such a notification is retrieved.

Test for agreement between multiple sources. The probability that an attacker compromises multiple data sources in exactly the same way is much lower than the probability that an attacker compromises one of them.

Threatmodel 2: edge case behavior of model

Even if model is dead simple, it could still go off the rails if it got bizarre, unforeseen inputs.

Analogy

In a more intricate model, **out-of-distribution robustness**[3] describes the resilience of that model to *shifts in input distribution* or, in the extreme case, an attacker eliciting behavior that the model creator does not want by finding inputs on which the model behaves pathologically. As you can imagine, it is easier for attackers to simply make the model fail (make wrong predictions, for instance) than it is for attackers to target behaviors that they desire (make the model benefit them in some way, for instance).

[3] https://en.wikipedia.org/wiki/Adversarial_machine_learning

Mitigation

Be extremely liberal in property tests, this is not a time to save testing resources from implausible cases, because implausible cases could be what hurts us.

Be extremely conservative in input validation/constraints, though beware a lot of inference/data decisions are being made when such constraints are imposed.

Threatmodel 3: positive feedback loop

Optimistically trades and prices made on our DEX system will be a part of the data from coinbase/binance. What does it mean when our behavior shows up in the data from which we derive our behavior?

A potential pitfall here isn't entirely unlike threatmodel 2. Under positive feedback, data can simply be sent off the rails into "edge case behavior" that we didn't think would show up in the operating of our model.

Analogy

Fraud detection at Stripe[4] is trained on the prior year's fraud data. The problem is that data is labeled by the earlier iteration of the model, so a perturbation in the model's behavior might lead to a (nonlinearly drastic) perturbation in the new data labels.

Mitigation

Via the Stripe example, we can do something called *counterfactual evaluation* which is an algorithm for generating data "as if" our behavior wasn't influencing the data. If this seems important/promising I can workshop with Bassam what this would look like for us.

0.3 Vampire attack

Wouldn't it be great if there were unique affordances by Cardano itself that prevented vampire attack?

[4] https://youtu.be/rHSpab1Wi9k

What is it?

According to my research, this attack happened once. Finematics' coverage is excellent, so I'll defer explanation of the base level occurrence to them[5]. But here I will attempt to provide a generalized notion.

[5] https://finematics.com/vampire-attack-sushiswap-explained/