Discrimination of Toxic Flow in Uniswap V3: Part 2

CrocSwap

This post is a new installment in an ongoing series by @0xfbifemboy on Uniswap liquidity pools, concentrated liquidity, and fee dynamics. It is the second of multiple posts in a subsequence which aims to focus on the characterization of toxic flow in ETH/USDC swap data and potential implementations of price discrimination or flow segmentation mechanisms.

Introduction

In the previous post, we began to characterize toxic and non-toxic swap flow on Uniswap V3's ETH/USDC pools in greater detail. We identified several patterns of interest; for example, we observed that 'fresh' wallets with limited trading history were likely to make uninformed trades (profitable for the liquidity pool), whereas a very small subset of wallets (fewer than 1%), trading frequently and at size, originated the majority of toxic flow and the lion's share of the pool's losses.

Although we were able to produce some interesting observations about the nature of toxic and nontoxic flow, our post left a number of critical questions unanswered. How robust is our method for identifying sources of toxic flow? Can we say anything deeper about the behavior of wallets that originate toxic versus nontoxic flow? Finally, and perhaps most critically, how can we bring our insights together into a practical methodology for price discrimination based on the source of a swap?

In this post, we attempt to bring some of these loose strands together. We explore alternate definitions of toxicity, find them to be largely consistent with our prior work, and create a consolidated classification of wallets into three toxicity levels. We analyze the trading of these wallets in greater depth. Closing out, we sketch out a tentative proposal for the design of a system that could be used to upcharge toxic flow or, equivalently, give discounts to retail traders.

An alternate definition of toxicity

Recall that initially we attempted to segment swap flow on a per-wallet basis by looking at aggregated wallet statistics. In particular, we looked at each wallet's average notional swap size versus its average PnL, for wallets with at least 250 swaps:

We also calculated the autocorrelation of swap PnL across buckets of 50 swaps each. Noticing the presence distinct clusters in the above plot, we may make the following tentative definitions:

- High toxicity: Wallets with at least 250 swaps, mean PnL < 0 basis points, mean notional > 10k USD, swap PnL autocorrelation > 0.9
- Medium toxicity: Wallets with at least 250 swaps, mean PnL < 0 basis points, mean notional > 10k USD, swap PnL autocorrelation ≤ 0.9
- Low toxicity: All other wallets

It appeared, based on this categorization, that the bulk of the liquidity pool's losses originated from the high and medium toxicity groups, suggesting that our grouping of wallets is valid.

However, when trying to analyze situations with relatively unknown "ground truth," it is always useful to ask the following: How robust are our results to alternate metrics or definitions? We will attempt to identify toxic wallets in a different way and see how the results compare.

Instead of starting with aggregated statistics, can we instead look at the wallets from which toxic swaps originate? Recall our original observation, where we found that the vast majority of the liquidity pool's losses came from large swaps with negative PnL:

We restrict specifically to the set of wallets with at least five ETH/USDC swaps, where all of those swaps were between 80th and 95th percentile in notional size and all of which had a markout PnL of -5 basis points or less. This yields a list of merely 1,776 wallets, a very small fraction of the over 450k distinct wallets recorded in the ETH/USDC swap data.

Tentatively classifying these wallets as "toxic wallets" and plotting the distribution of each wallet's average notional swap size, we do see that these wallets have much higher notional swap sizes than usual:

This is an unsurprising result, as having multiple large swaps was a precondition to being included in this group of wallets to begin with!

One might then ask if the same trend holds for swap PnL. Below, we plot the distribution of average swap PnL per wallet, restricting to wallets with at least 20 swaps in the ETH/USDC pool for simplicity (this eliminates a great deal of noise from the results):

Interestingly, it appears that our group of toxic wallets exhibits a bimodal distribution in swap PnL. One of the peaks has, on average, positive PnL for the liquidity pool, much like the larger distribution of non-toxic wallets; however, the other peak is deeply into negative territory, with a mean of -5 basis points or so. The second (leftmost) peak in the distribution probably reflects wallets that consistently generate swap toxicity for ETH/USDC liquidity providers!

Let us take this leftmost peak of toxic swappers with negative average PnL. To refresh, this subset consists of 817 wallets with the following characteristics:

- Average negative PnL across all swaps
- At least five swaps with PnL below -5 basis points and notional size between the 80th and 95th percentiles

If we look at the swaps originating specifically from these wallets, do we find anything interesting? We can once again segment by notional swap size:

Swap PnL indeed declines as notional swap size increases, and here the rate of decline is quite regular and begins at a lower percentile of notional swap size than when looking at the totality of all swaps. Additionally, the PnL seems quite noisy for low notional sizes, leading us to suspect that most of the swaps from these wallets are concentrated at higher notional swap sizes:

As expected, the vast majority of the swaps made by these wallets are at very high notional trade sizes!

In aggregate, the notional PnL realized by the liquidity pool as a result of all the swaps from this small group of 817 swappers is a whopping -165 million USD. Recall that the overall profitability of the liquidity pool using short-term Binance markouts was merely -43 million USD, meaning that the aggregate PnL from all other swappers is positive 122 million USD! We see once again the compelling value of effective discrimination between toxic and non-toxic flow: if all the swaps originating from these 817 swappers had been charged five additional basis points, the aggregate liquidity pool PnL would almost certainly be deep “into the black!”

Consolidating our definitions

Now, on top of our first classification of wallets as possessing high, medium, or low toxicity, we now have an equally plausible division of wallets based on the following definitions:

- Toxic wallets: Wallets with negative average PnL which also possess at least 5 swaps, all of which are between the 80th and 95th percentile in notional swap size and all of which have PnL of -5 basis points or less
- Nontoxic wallets: All other wallets

Both sets of definitions are relatively similar, but not identical. How do they compare? It turns out that:

- The high toxicity wallets in the first definition are all classified as toxic wallets in the second definition
- The medium toxicity wallets in the first definition are almost all (>97%) classified as toxic wallets in the second definition
- The low toxicity wallets in the first definition are almost all (>99%) classified as nontoxic wallets in the second definition

We have almost perfect concordance between the two classifications, with two caveats:

- The high toxicity group in the first definition, which looks at autocorrelation of bucketed swap profitability, does seem to be picking out an important and special subset of toxic wallets overall
- There are 1,408 wallets classified as toxic in the second definition but as low toxicity in the first definition; while a small fraction of the whole, these wallets do seem to genuinely originate toxic flow, and they are missed by the first categorization mainly because they have fewer than 250 swaps overall

We therefore generate a consolidated ranking of wallets into three toxicity groups:

- High toxicity: Classified as high toxicity in the first definition
- Medium toxicity: Classified as medium toxicity in the first definition or classified as toxic wallets in the second definition
- Low toxicity: All others

While this may seem a little pedantic, it is nevertheless valuable to show that, approaching an unclear problem from two distinct perspectives, we still end up at roughly the same conclusion. It gives us confidence, furthermore, that our results will be robust across different choices of explicit parameters or thresholds. We recognize, naturally, that this is fundamentally an imperfect categorization. Certainly there will be real arbitrageur wallets we have missed and, similarly, real instances of retail flow that we may have miscategorized as being high or medium toxicity! Nevertheless, we believe that reasonable analyses

of this data will largely concur with our results here, even if not with perfect precision.

Characteristics of toxic flow

Now that we have a consolidated classification of wallets as possessing high, medium, or low toxicity, what else can we say about the swap behavior of these wallets?

One natural question to ask is how frequently swaps come in. Intuitively, we expect automated trading strategies with systematic alphas to fire off swaps at regular intervals (whenever trading opportunities or price dislocations are detected). On the other end of the spectrum, a retail trader might fire off swaps at relatively more random, infrequent times: perhaps a trade here when they feel like it, another trade over the weekend… and so on.

Accordingly, we examine the distribution of each wallet's median time between consecutive swaps (we use the median here to avoid biasing estimates from temporary pauses in trading and so on, and restrict to wallets with more than 20 swaps recorded):

It appears, consistent with our expectations, that high toxicity wallets trade extremely frequently (typically an average of 2 minutes or so between swaps). Interesting, medium toxicity wallets have a more "spread out" distribution of between-swap intervals, taking slightly longer on average but still usually placing orders at a fairly rapid clip. This might suggest that the high-toxicity wallets participate in arbitrage trades of a more 'consistent' nature, perhaps atomic, on-chain arbitrage for example, whereas perhaps medium-toxicity wallets tend to focus on statistical arbitrage in comparison, with alphas at a larger variety of timescales. Such a model would also be consistent with our prior observation that the high-toxicity wallets tend to make swaps with notional sizes 1–2 orders of magnitude than wallets in the medium-toxicity group.

Beyond looking at swap frequency alone, we can also look at how frequently each wallet switches the direction of its swaps. One might naively expect that high-frequency trading strategies would switch swap direction very often; for example, they might buy ETH at a low price then sell at a higher price the next block; alternatively, even if only "one side" of each trade happens on Uniswap, we would expect many types of statistical alphas to give us essentially a random distribution of ETH buys vs. sells. In contrast, one might expect that a retail trader would exhibit a great deal of momentum (from FOMO, etc.) in their trades, leading to longer durations of time between switching swap direction.

We directly examined the distribution of each wallet's average "run length," where we look at the length (in terms of numbers of swaps) of "runs" of consecutive, same-direction swaps in each wallet's swap history:

(Note that to control the presence of extreme outliers, we are again restricting to wallets with more than 20 swaps.)

Clearly, high-toxicity wallets switch swap directionality far more frequently than either the medium- or low-toxicity wallets! It is actually quite interesting that the medium-toxicity wallets switch swap direction so much less frequently than the high-toxicity wallets; this would be consistent with the medium-toxicity wallets focusing on statistical arbitrage, which resolves on longer time horizons than the high-toxicity wallets' strategies and, perhaps, even involve multiple swaps in the same direction as part of a trading response to the same signal.

This naturally brings to mind a subsequent question: on what timescale exactly, do the alphas of high or medium toxicity wallets persist? For a retail trader who simply decides to buy or sell some quantity of ETH on a whim, it does not really matter (in a sense) if they execute their trade now, or in ten seconds, or in ten minutes, or perhaps even in ten hours or ten days; some undesirable variance is introduced via random price movements, of course, but on the whole, they do not expect to have any particular ability to 'time' ETH/USDC prices especially well, and should be mostly indifferent to swap execution now versus in the next hour.

On the other hand, high-frequency traders seeking to exploit statistical correlations, predicted movements, and dislocations between venues should exhibit much higher sensitivity to the choice of timescale. If Binance price leads ETH price discovery, for example, then Uniswap pools will only remain 'mispriced' after a large movement on Binance for a single-digit number of minutes (if even that long). They should certainly very much not be indifferent to placing a swap now versus placing a swap in the next hour.

We can try to analyze this question by looking at how the average PnL per wallet varies across different markout horizons:

(In the above graph, we are taking the notional-weighted average of the PnL of each wallet's swaps. The motivation here is that the calculation of each wallet's PnL should take into account the relative "bet sizing" of each trade, rather than assuming that notional sizes of different trades are completely independent of each other.)

As expected, low-toxicity wallets have a high average PnL (profitable for the liquidity pool), and the PnL of these wallets is largely independent of markout horizon. Intuitively, this makes sense: if you are a trader with zero edge, then your expected PnL in the future, whether we check 1 minute or 10 minutes after your trade, should still be exactly zero (albeit with higher variance for longer markouts). Interestingly, though, we do see that high and medium toxicity wallets — especially high toxicity wallets — seem to suffer severe alpha decay on the sub-1-minute horizon, with PnL largely plateauing afterwards. This does suggest that the trading signals that profitable systematic traders use in the ETH/USDC pool decay to zero alpha throughout the first minute or so (4–5 blocks on Ethereum mainnet).

We can refine our understanding of these wallets' trading behaviors even further by calculating wallet PnL using markouts based on the marginal Uniswap ETH/USDC pool price rather than Binance data:

Notice here that the high-toxicity group now appears to exhibit alpha decay on the order of >5 basis points after 10–15 minutes, rather than mostly decaying in the first minute of data. To see the significance of this observation, imagine that you are an arbitrageur and you notice that ETH is trading 20 basis points higher on Binance, which leads in ETH price discovery, than on Uniswap. You want to arbitrage the difference, so you swap through the 0.05% fee pool. However, you only buy up enough ETH such that the price of ETH goes up to 15 basis points higher at the margin and no more; any closer to the Binance price, and you are now overpaying for ETH after taking into account the 5 basis point trading fee.

Eventually, because the price feeds of ETH on Binance and Uniswap are tightly cointegrated processes, you do expect the Uniswap price to converge to the Binance price; however, this might take place over the course of, say, 10 minutes or so, rather than a sub-1-minute timescale. Because the gap between the post-arbitrage Uniswap price and the Binance price was 5 basis points, you would expect to see the Uniswap pool price move more 5 basis points in the same direction as the arbitrage swap over the next 10+ minutes, making the arbitrage swap's PnL seemingly increase in the arbitrageur's favor by 5 extra basis points if we calculate markout PnLs based on markout durations in that time interval. This is exactly the empirical pattern we see in the plot above!

Similarly, if a zero-edge retail trader's swap creates a backrun opportunity, the markout PnL of that swap will decline over the next minute or so as the backrun opportunity is inevitably filled by an arbitrageur. This, too, is exactly the pattern we see in the plot above — the markout PnL of low-toxicity wallets' swaps increases in the pool's favor over the first couple blocks' worth of time after the swap before completely plateauing.

All in all, we have managed to make some fascinating observations about the behavior of wallets at different toxicity levels. However, can these findings be applied practically, or are they largely of academic interest?

Implementation of a discriminator mechanism

Let us take a step back and appreciate the problem before us. How do we actually implement an effective program of price discrimination? Although we have discussed quite a few characteristics of toxic and non-toxic flow in this post so far, it would be challenging and gas-intensive to implement these in an actual smart contract. Address whitelisting or blacklisting is trivially checked and evaded; explicit conditions can be reverse-engineered and gamed; finally, performing computation on the EVM is simply a very expensive endeavor, and so overly complex fee discrimination schemes may increase gas costs for swappers well past the point of tolerability.

We are in active exploration of various implementation methods for fee discrimination schemes. One compelling way to approach the problem is shifting the framing of the question: instead of asking, How can we implement this in a way compatible with the EVM's computational requirements?, one might instead ask, How do we shift the computational burden elsewhere so that we can take full advantage of our characterization of toxic flow? It is not strictly necessary for the protocol itself to perform price discrimination at the smart contract level; instead, one could take advantage of the rich network structure inherent in modern blockchains and set up systems of incentives that allow price discrimination to occur in a more decentralized fashion.

To make this concrete: suppose that the protocol has a whitelist of certain relayers which receive a privileged (discounted) fee rate. Because relayers profit from being allowed to relay transactions, they will want to stay on the protocol whitelist, so that they are chosen by more swappers. If protocol governance periodically monitors the PnL of whitelisted relayers to ensure that whitelisted relayers are consistently sending in nontoxic swap flow, this creates a system of natural incentives for relayers to implement arbitrary complex wallet profiling schemes. (Hopefully, the research in these posts will serve useful for that purpose!)

Conversely, one might one ask: How could wallets signal signs of nontoxicity, and how can we detect and privilege wallets which signal sufficient nontoxicity? One natural example comes directly to us from the analysis of alpha decay in the prior section! Recall that the alphas of high and medium toxicity wallets appear to decay (relative to Binance markouts) on the sub-1-minute timescale. Therefore, if a wallet is willing to execute its swaps on a delayed timescale, for example delaying 5 or 10 minutes after the time of swap submission, that is actually a very strong signal of swap nontoxicity, and the wallet's swap should likely be incentivized with a fee discount! One may imagine the incentivized-relayer setup previously described easily implementing such a method of flow segmentation.

Further methods of utilizing a blockchain's network structure can be devised; for example, one could imagine block builders playing some role in such a system. Regardless of the exact implementation, however, it seems that partial decentralization of the discriminatory element into a multi-actor system is a potentially very powerful method of characterizing and segmenting incoming ETH/USDC swap flow.

Conclusion

In this post, we have made considerable progress in our study of toxic and nontoxic flow on Uniswap ETH/USDC pools. In particular, we have shown that our prior identification method for toxic wallets was fairly robust with respect to alternate

definitions of wallet toxicity. After doing so, we generated a ‘consolidated’ categorization of wallets as having high, medium, or low toxicity. We were able to look more deeply at the swapping behavior of these wallets, finding that:

- High-toxicity wallets swap more frequently, and change swap directionality more frequently, than either medium or low toxicity wallets
- High-toxicity and medium-toxicity wallets have trading alphas that decay on the timescale of ~1 minute, although it takes on the order of ~10 minutes for Uniswap prices to fully converge to the prevailing Binance price

We also proposed a relayer-based scheme for swap flow segmentation and gave a simple example of how wallets might be able to signal nontoxicity to a relayer or to the underlying protocol. In sum, we have worked out an “end-to-end” example of how a protocol might in principle implement a working price discrimination mechanism to give nontoxic flow a substantial fee discount.

However, we have still left many stones unturned. For example, now that we have a fairly good classification of wallets in terms of their swap toxicity, we might begin to ask: how predictable is the PnL of each group of wallets if we look at retrospective data, such as the volatility of ETH prices in the last 5 minutes? We hope to explore such questions, and more, in future installments.

-0xfbifemboy