Does Online Anonymous Market Vendor Reputation Matter?

Alejandro Cuevas
Carnegie Mellon University

Nicolas Christin
Carnegie Mellon University

Abstract

Reputation is crucial for trust in underground markets such as online anonymous marketplaces (OAMs), where there is little recourse against unscrupulous vendors. These markets rely on eBay-like feedback scores and forum reviews as reputation signals to ensure market safety, driving away dishonest vendors and flagging low-quality or dangerous products. Despite their importance, there has been scant work exploring the correlation (or lack thereof) between reputation signals and vendor success. To fill this gap, we study vendor success from two angles: (i) longevity and (ii) future financial success, by studying eight OAMs from 2011 to 2023. We complement market data with social network features extracted from a OAM forum, and by qualitatively coding reputation signals from over 15,000 posts and comments across two subreddits. Using survival analysis techniques and simple Random Forest models, we show that feedback scores (including those imported from other markets) can explain vendors' longevity, but fail to predict vendor disappearance in the short term. Further, feedback scores are not the main predictors of future financial success. Rather, vendors who quickly generate revenue when they start on a market typically end up acquiring the most wealth overall. We show that our models generalize across different markets and time periods spanning over a decade. Our findings provide empirical insights into early identification of potential high-scale vendors, effectiveness of "reputation poisoning" strategies, and how reputation systems could contribute to harm reduction in OAMs. We find in particular that, despite their coarseness, existing reputation signals are useful to identify potentially dishonest sellers, and highlight some possible improvements.

1 Introduction

Reputation and feedback systems are ubiquitous to facilitate trust and trade in online marketplaces such as eBay, Amazon, AirBnB, and Uber. A functioning reputation system benefits honest vendors and pushes dishonest vendors out of the platform. High-quality vendors are rewarded with signals of their trustworthiness, which ought to make them more appealing to future buyers. On the other hand, low-quality vendors ought to see their sales dry up after a series of disgruntled buyers report their negative experiences.

On markets where "seller anonymity is guaranteed, and no legal recourse exists against scammers, one would expect a certain amount of deception." [8] Yet, the market capitalization of online anonymous marketplaces (OAMs) has massively grown since their inception in 2011, with individual vendors in these platforms that operate multi-million dollar operations [35, 46]. This alone seems to indicate that the reputation and feedback systems in place in these marketplaces are overall working as expected.

However, scam stories abound in underground forums. Goods that do not match their description, dangerously adulterated drugs, and unfulfilled orders are among the most common complaints. So, which is it? Do these marketplaces provide enough signals for buyers to distinguish between high and low quality vendors? Or do buyers have to resort to other signals to make this determination?

Answering these questions is especially important in the context of underground markets, where hazardous substances (e.g., narcotics) are often being sold. A key argument in defense of these markets is that, by enabling buyers to avoid dangerous vendors and/or products, reputation systems help with harm reduction compared to alternatives (e.g., street sales). However, this claim assumes that these reputation systems provide a useful signal.

Surprisingly, despite substantial research demonstrating the importance of reputation in driving sales in traditional online marketplaces [2, 38], there has been significantly less exploration of what drives success in OAMs. While prior work has found some correlations between market or forum-derived features and performance [10, 18, 45], they have only studied narrower contexts: carding forums [10, 18], or B2B cybercrime vendors in a single market [45]. Furthermore, despite ample evidence that buyers use Reddit-like forums to provide additional vendor reviews, no prior work studies the link between forum-derived features and success in OAMs. Last,

despite prior work examining listing and vendor longevity in OAMs [8, 35], no prior work tests *which* factors impact survivability of vendors in these markets.

We fill these gaps by exploring the predictive power of various signals on the financial success or longevity of a vendor from a OAM. We 1) use multivariate survivability models to test the role of various covariates on the disappearance of a vendor, and 2) use explainable machine learning models to predict the disappearance and wealth tier that a vendor will belong to in a future state of the market. We conduct our experiments on eight OAMs and two types of forums, with activity spanning from 2011 to 2023.

Ultimately, long-term vendor success, as determined by accrued wealth and permanence in the market, is a good proxy for the vendor selling acceptable products. As such, the ability to predict this success likely helps predicting the risk associated with a specific vendor.

We offer the following contributions:

- We quantify the impact of various market and forum-derived features on vendor longevity and find that feedback scores (including imported product reviews from other markets) have a significant impact on increasing longevity across most markets we study;
- We find that (both positive and negative) reputation signals from forums explain vendor survivability, but overall have little predictive power for vendor success;
- We demonstrate we can build a *generalizable* model to predict, more accurately than raw feedback, which vendors may leave the market in the short-term (1–3 months);
- We find that future financial success is predictable, particularly for the top/bottom 25% of vendors, and even *on previously unseen markets*.
- We find that features external to the market, and time-series representations of features not only fail to increase the predictive power, but instead often *decrease* it.

Our results have several implications. First, our models can be used by law enforcement agencies for early identification of important vendors on emerging markets. In particular, by achieving high predictive accuracy even when applied to a new, previously unseen market, our models can make monitoring and intervention efforts targeting online criminal ecosystems more efficient. Furthermore, our results shed light on the viability of strategies that involve "poisoning" the reputation of vendors inside OAMs and across forums.

Second, our results empirically validate the role of reputation systems in OAMs. On the one hand, we find evidence that a functioning feedback system may help online marketplaces reduce harm for drug consumers—and that it can be improved by looking at other signals. On the other hand, we find that discourse and reputation signals from external forums may not be as useful to identify bad actors.

2 Background and Related Work

We next provide an overview of reputation systems in the broad context of general online commerce, before focusing on idiosyncracies of anonymous markets; and discuss measurement and inferencing work on OAMs and forums.

2.1 Reputation & Feedback Systems

Online marketplaces initially faced significant skepticism, particularly from economic theorists. The asymmetric information between buyers and sellers, as well as the lack of incentive from one party to guard against risk, can indeed drive markets to failure [38]. Traditional online marketplaces (e.g., eBay) overcame these challenges by employing reputation systems. Reputation systems are essential in creating trust, particularly in two-sided marketplaces (i.e., markets that serve as platforms to connect independent buyers with independent vendors, such as eBay). The promise that good (resp. bad) behavior in the present may be rewarded (resp. penalized) in the future by increased (resp. decreased) sales is how reputation systems incentivize buyers to act in good faith. There is substantial economic literature in conventional markets that empirically demonstrate how vendors with better reputation attract more buyers and higher prices, while the converse holds true for disreputable vendors [2].

OAMs face similar challenges as conventional online marketplaces, but with some particularities. First, dispute resolution is less robust. None of the parties (particularly buyers) have any legal recourse when facing a scam. Second, vendors often have access to buyers' private information (e.g., shipping address) and can leak this information in retaliation [27]. Third, illicit goods—particularly, narcotics—typically have high price and quality dispersion [34]. This quality uncertainty is exacerbated by the lack of incentive for buyers to guard buyers against risk (moral hazard).

Nonetheless, OAMs have persisted and thrived, which indicates that they have managed to create systems of trust between buyers and sellers. Platforms offer a variety of features to create trust, including escrow, discussion forums, feedback scores, automated reviews, and various signaling mechanisms such as badges [25, 41]. In two surveys, OAM buyers reported that the existence of reputation systems fostered their engagement [3, 4]. Yet, it is unclear which specific signals are most important in creating trust and drive vendor success.

2.2 Performance in Criminal Markets

Prior work [5, 10, 18, 45] has attempted to measure and explain the factors that drive success in criminal markets, particularly in OAMs and sales-driven criminal forums.[1] While OAMs

[1] We distinguish between sales-driven criminal forums whose primary intent is to connect buyers and vendors in private transactions, and forums that serve a complementary role to OAMs, e.g., to discuss vendor experiences.

and criminal forums offer slightly different transaction experiences for buyers, they share many similarities and have been broadly studied using similar theories. For instance, researchers have analyzed vendor signals through Gambetta's signaling theory [14] to identify and explain buyer preference in carding forums [10, 18], while van Wegberg et al. applied it to explain B2B vendor performance in OAMs [44]. Several papers have attempted to characterize vendor trajectories in OAMs [5, 44], or have studied conversations and actors in forums to identify "key players" [7, 22, 33, 49]. Similarly, others have found links between observable features (e.g., vendor position in their social network) and private features (e.g., amount of private messages received) [28, 30, 37].

Ultimately, this body of research attempts to identify which vendors will become successful directly (e.g., sales volume when feedback can be used as a proxy) or indirectly (e.g., number of private messages when sales proxies are elusive). Unfortunately, the results have not yielded a clear picture of what drives financial success. Van Wegberg et al. posited that who the vendor is matters more than product differentiators [45]. Holt et al. found that signals like badges in forums seemed to drive more feedback [18]. Décary-Hétu et al. found correlations between vendors' sales and their network features but not with their forum features [10]. Furthermore, even though buyers have long used forums to review OAM vendors [19, 20, 24, 47], the literature shows a gap on how reputation and/or influence signals from forums affect OAM vendor success. Despite prior work modeling the survivability of vendors and listings [8, 35], factors that accelerate the disappearance of vendors and listings remain unknown. Last, Bradley explored the resiliency of the OAM ecosystem, as well as that of vendors within it. Closest to our work, they observed how reputational damage may reduce vendor capacity to trade [6]. They also employed a qualitative approach on forum data to assess the impact of law enforcement operations [6]. We use similar techniques but apply them toward vendor financial success.

3 Methodology

We next describe how we obtained and processed data from the markets and forums we analyze, how we extract the features our analysis uses, and discuss data validation.

3.1 Data Sources

Marketplaces: High-confidence inference from web scrapes requires robust processing and validation strategies [9]. Hence, we use peer-reviewed and validated datasets when possible. For the Silk Road, Pandora, Silk Road 2.0, Agora, and Evolution markets, we use the Soska and Christin [35] dataset; for Hansa Market, the Cuevas et al. [9] dataset; for Alphabay, the van Wegberg et al. [46] dataset. In addition, we collected and processed a market active at the time of writing, Nemesis, along with its internal forum. Figure 1 shows the revenue of all markets (scaled to be on the same time axis).

Subreddits: Many OAMs used Reddit as a discussion platform until they got banned in 2018 [12].[2] We collected and processed data from the subreddit `/r/HansaDarknet-Market` which contains 264 posts, 3,613 comments, from September 2015 to September 2017. This subreddit was used to discuss matters related to the Hansa marketplace (e.g., news, policy updates), by vendors to advertise products, and by buyers to describe their experiences with vendors. We also collected and processed `/r/DarkNetMarkets`, with 125,300 posts, and 1,850,533 comments, ranging from October 2013 to September 2017. Similar to `/r/HansaDarknet`, this subreddit discussed vendor quality across a variety of markets, among other topics.

Nemesis Forum: The Nemesis forum similarly employs a Reddit-style interface, with various sub-forums such as `/n/AskNemesis` for platform questions and `/n/Cocaine`, for discussions related to cocaine vendors. Creating a Nemesis marketplace account also creates a Nemesis forum account, so that marketplace and forum handles are identical (for both buyers and sellers). We collected 4,018 posts and 12,710 comments from March 2022 to February 2023.

3.2 Data Processing and Validation

We scraped Nemesis from November 18th, 2022 to February 1st, 2023 at a rate of about 32 pages/min (or roughly 46,000 pages per day). We employed a Scrapy-based breadth-first scraper.[3] Similar to previous work, we attempted to proxy sales by matching feedback to item listings. Given that we began scraping the market relatively early in its development, we were able to match over 99% of collected feedback to listings. This is facilitated by Nemesis' design: feedback left on vendor pages links to the item page featuring the review. However, Nemesis presents a unique challenge: some individual item listings feature various quantity options (e.g., a listing "High-quality Cocaine" may offer "1g at $10," "15g at $125," and "1kg at $5,000"). The most conservative approach would be to assume that each sale is for the lowest priced option, giving us a lower bound on sales, but potentially vastly underestimating the sales. Instead, we experiment with taking the mean and the median price when a list of options is provided. While Nemesis vendors do not seem to use "holding prices," i.e., abnormally high prices signifying a lack of stock, we applied the same heuristic as Soska and Christin to filter these out, such that our data is consistent with theirs [35].

Furthermore, to validate our processing and inference, both authors independently parsed the raw HTML scrapes and esti-

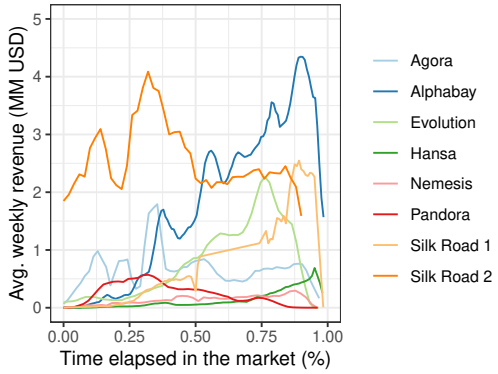
[2] After the 2018 ban, a Reddit alternative, Dread, emerged, but it does not feature data relevant to the markets we study—in particular, Nemesis discourse is all but banned on Dread due to a feud between administrators.

[3] Due to parallelization across multiple scraping agents, the breadth-first order is not always respected in practice.

Table 1: Overview of collected and processed marketplace data.

Marketplace	#Vendors	#Feedbacks	Est. Revenue	First Seen	Last Seen	Activity Length	#Snapshots
Silk Road	2,336	605,744	$62,334,431	2011-11-27	2013-08-19	631 days	133
Pandora	459	89,065	$12,239,165	2013-11-02	2014-10-13	345 days	140
Silk Road 2.0	1,202	687,375	$121,529,265	2013-11-27	2014-10-29	336 days	195
Agora	1,961	234,272	$40,857,567	2013-12-24	2015-02-11	414 days	161
Evolution	2,352	464,146	$43,993,997	2014-01-13	2015-02-18	401 days	43
Alphabay	6,101	1,736,127	$218,971,605	2014-12-31	2017-05-26	877 days	33
Hansa	1,309	153,400	$13,149,373	2015-08-21	2017-07-15	694 days	14
Nemesis	372	18,794	$6,388,411	2022-03-09	2023-01-31	328 days	10

Figure 1: Revenue over time for all markets scaled to the same time axis. Each point is a four-week rolling window average.



mated the revenue for each vendor. Our estimates of revenue using the minimum listing price are within 6% of each other. We did not find public analyses of Nemesis to which we could compare our revenue against as Cuevas et al. did for the Hansa market [9].

3.3 Extracting Features

Our goal is to understand the impact of reputation systems on the financial success or disappearance of vendors across markets. Over time, markets have displayed a variety of attributes and badges for vendors, such as measures of "level," "score," "experience," and/or whether a vendor has undergone some measure of verification. Markets have also employed a variety of feedback scales (e.g., 1–5 stars, positive/neutral/negative, etc.). Furthermore, vendors have also utilized various differentiators, such as alternate media of communication (e.g., Telegram, ICQ, etc.), describing terms of service, refund policies, and avenues for customer support. Prior work has found that some platform-specific attributes may be used to explain sales performance (e.g., customer support), for a type of goods (i.e., cybercrime-related), within a specific market (Alphabay) [46].

We hypothesize that we can use generalizable features or attributes to explain and predict the performance of vendors across OAMs. We focus on features common across markets, using the basic objects that support these markets: feedback/reviews, listings/items, and vendors [9]. We also explore the impact of capturing time variations across these features, as the market evolves. Lastly, we investigate the impact of forum-based features.

3.4 Ethics of Data Collection and Release

Our Institutional Review Board (IRB) deemed our study not to be human-subject research. Nonetheless, our work still has important ethical implications. The collection and release of our data follows the principles outlined by Martin and Christin [26], and the same approach as previous OAM research, especially Soska and Christin [35] and Cuevas et al. [9]. Whenever possible, we prioritized the use of existing peer-reviewed datasets for replicability and to abide by the same ethical considerations as prior work. However, we also collected data from a new marketplace, Nemesis market. For this, we balanced data accuracy with stealth (to avoid impacting the studied ecosystem) and low impact on the Tor network (using a light-weight crawler). We also contribute fast Tor relays with long uptime to compensate for our use.

Marketplace and forum data contain discussion of potentially illicit activities; and forum data may inadvertently leak information about buyers and/or sellers. After consulting our IRB and general counsel, an unlimited public release is undesirable. However, we can follow the lead of other researchers. Indeed, data for seven of the eight marketplaces we study are already publicly available [9, 35] *upon request*, for non-commercial use, using the IMPACT portal.[4] This allows researchers to vet possible uses of the data before releasing it. We will adopt the same strategy for our own (Nemesis) data.

Last, we also relied on Reddit data, which was publicly available through Pushshift [1] at the time of writing. However, since then, Reddit updated its API policies which has affected the availability of these data through Pushshift [43].

[4] https://www.impactcybertrust.org.

Instead, these data may be accessed through independently hosted torrents [36].

3.4.1 Base Features

As a starting point, we define and extract features common across our markets. Our initial set of feature categories are:

- **Revenue features**: mean/median order value, cumulative revenue, time of first sale.
- **Feedback features**: count, average count per week, and feedback score.
- **Listing features**: item diversity, within-category price z-score, time of first listing, main category of goods sold.

3.4.2 Temporal Features

We can make some of the above features more expressive by adding a time dimension. Using just the base features defined above, yields a matrix of shape $(nr\_vendors \times nr\_features)$ up to a time T in the market. However, we could also build time series for several of the above features by tracking their evolution over time. That is, we break T into a series of time steps t_i, resulting in a matrix of shape $(nr\_vendors \times nr\_timesteps \times nr\_features)$. For example, for period of length T, rather than just having the total revenue up to time T, we instead consider the revenue per time step (e.g., week) t_i up to T.

3.4.3 Forum Features

Forums provide a platform for customers to discuss experiences with vendors, or suspicions that a vendor has been compromised by law enforcement [19, 24]. Forums also provide signals on vendor notoriety (e.g., if a vendor's posts garner a lot of attention) or influence (e.g., by looking at their interaction network size). As such, forum signals may help predict vendor success and longevity.

For Hansa, we consider `/r/HansaDarknet-Market` and `/r/DarkNetMarkets`, similar to prior work [6]. Posts we consider in `/r/DarkNetMarkets` refer to vendors who existed in Hansa, but may not always directly relate to a sale taking place on Hansa. Further, there is no definitive way of mapping users from Reddit to the Hansa marketplace. For this reason, we do not attempt to build interaction networks between users. Instead, we only extract comment "sentiment" (good, neutral, and bad) about vendors. We first try automated methods: named-entity recognition for vendor discovery, and sentiment analysis. However, pilot testing showed that these methods perform poorly in these forums, as described below.[5] Instead, we opt for a manual analysis process.

[5] Whether typical sentiment analysis packages could be made to work, using specialized training sets, is an open question, that is likely to be answered in the affirmative. However, performing such retraining would have required a labeled OAM forum dataset in the first place, which was not available.

We first use a fuzzy matching search for vendor names across posts to find a set of candidate posts and comments that may refer to a given vendor. We find 2,294 posts and a total of 13,002 comments under these posts. One coder independently goes through the posts and comments and 1) confirms that the match was appropriate, and 2) determine the sentiment of the comment or post given to a vendor. For validation, we randomly select 10% of the posts and comments and have a second coder qualify these posts. That way, we also ensured that the first coder was not missing entries. The coders then compared their results and derived a Cohen's Kappa of $\kappa = 0.58$, moderate agreement. The disagreements mainly stemmed from three types of issues:

- **Unclear interpretation (e.g., conflicting sentiment)**. "[REDACTED]'s bud would actually look reallly nice if his buds werent so compressed."
- **Unclear attribution (e.g., acronym mapping)**. "I remember KK having issues with oily batches." posts anymore?"
- **Lack of understanding in lingo**. "50% FE [from this vendor]seems tarded to me. Anyone else?"

While a moderate agreement is not ideal, the examples above illustrate the difficulty of both attributing and interpreting signals in fora, even when done manually. Unsurprisingly, our automated efforts to extract entities (through named-entity recognition models) and to extract sentiment (by leveraging sentiment analysis models) failed to provide useful results. To mitigate the sources of disagreement, we introduced a "neutral" code for comments, rather than just "positive" and "negative." However, we chose to exclude "neutral" mentions as we found them to be of little use (e.g., "Please give me one example of shilling for [REDACTED]" conveys little signal). Furthermore, when the attribution was not clear, we decided to omit the comment. Using these guidelines, the first coder coded the rest of the dataset, and we focused on comments that had clearer signals, such as "I love that Yoda out of all those strains!! That Skywalker from [REDACTED] is fire as well!!" In total we found 843 positive (677 unique vendors) and 263 (210 unique vendors) negative comments.

For Nemesis we can derive social networks of vendors and buyers given that the forum and marketplace aliases are the same. We can see a vendor listings, as well as their posts and comments. Thus, we create a directed interaction network for comments, whereby an edge is formed when a comment is left as a reply to a comment/post. We also quantify the number of posts and comments made by each vendor, as well as the up-votes they receive. Due to the large number of interactions between buyers and vendors on the forum, we did not attempt to manually code the sentiment of these interactions.

3.4.4 Listing Categorization

Each market provides a different categorization of goods. For cross-market comparability, we use the listing category clas-

sifier from Soska and Christin [35]. This classifier predicts whether a listing pertain to one of the following categories: Opioids, Ecstasy, Psychedelics, Cannabis, Digital Goods (e.g., malware, cybercrime, carding), Prescription-based Drugs, stimulants, Benzodiazepines, Dissociatives, Other (which combines drug paraphernalia, weapons, electronics, tobacco, sildenafil, and steroids [35]), and Miscellaneous (everything that does not fit in any of the above categories).

4 Survivability Drivers

We first explore the impact of reputation scores on vendor survivability using a Cox proportional-hazards regression. We include in our model covariates that capture the main type of goods that the vendor offers, as well as whether they operate in a different market. Last, we control for the effects of wealth by stratifying our experiments.

Past work has measured the survivability of vendors by employing Kaplan-Meier models and using the observability (i.e., reachability of the page or last observed activity[6]) of vendors and listings to define the "death" event [5, 8, 35]. However, Kaplan-Meier models are univariate and do not allow us to observe the effect of various covariates, nor can they be used with continuous variables.

4.1 Experimental Setup

We define our death event to be the last week that a vendor has an observable sale (as observed by the feedback timestamp) in the market. If the vendor had a sale in the last two weeks of the market, we consider the vendor to have remained alive until the market end. We do this to account for collection errors during the days preceding a market takedown operation. Using statistical terminology, all vendors who did not die prior to the end of the market are "right-censored."

We are interested in the effects of reputation scores on the survivability of a vendor. To explore this effect, we also include covariates that may impact the survivability, namely, the main type of goods sold by the vendor, as well as their presence in other markets. To determine presence in other markets, we matched case-insensitive handles. Tai et al. show this approximation is acceptable, given the absence of ground truth and infrequent occurrences of impersonation [39]. Last, we account for the wealth tier the vendor belonged to during our last observations. More specifically, we encode our variables as follows:

- **Average feedback value (FB)**: the mean value of all the feedback the vendor received. If the market does not use a 5-point scale, we transform the scores using a min-max scaler.
- **Presence in other markets (POM)**: encoded as an indicator variable, 1 indicates the vendor's name exists in a different (contemporary or earlier) market, 0 and if not.
- **Main category**: Soska and Christin [35]'s classifier distinguishes between 10 categories of goods. To reduce the number of covariates in our model, we re-label the categories into a smaller set considering the potential harm to users [29]. We distinguish between category A drugs (potentially more harmful): opioids, ecstasy, prescription, stimulants, benzodiazepines, and dissociatives; category B drugs: psychedelics and cannabis (potentially less harmful); and digital goods (D). We exclude miscellaneous goods such as counterfeit goods and weapons, as their sales volumes are very small. We then create three indicator variables, where a 1 indicates the main category of goods sold by the vendor, between category A drugs (MA), category B drugs (MB), and digital goods (MD).
- **Wealth tier**: vendors are divided into quartiles based on the revenue they accumulated at the time of our last observation. We encode this as 1 to 4, where 4 corresponds to the highest 25% earners. Because this variable is correlated with survivability, we do not include it as a covariate. Instead, we stratify our model based on the four tiers.

The hazards regression formula for all markets is then:

$$h(t) = exp(\alpha + \text{FB} + I(POM) + I(MA) + I(MB) + I(MD)) \ .$$

Last, we run two additional experiments with the features derived from forum data, namely the `/r/HansaDarknet-Market` and `/r/DarkNetMarkets` subreddits and the internal Nemesis forum. For Hansa, we encode the variables as two indicator variables which capture negative and positive mentions. We choose indicator variables as the encoding for two reasons. First, plenty of users refer to vendors by aliases or abbreviations (e.g., "YD" for YOURDEALER). Our fuzzy matcher is not able to catch these instances so such users are underrepresented. Further, we noticed that in some threads, users almost exclusively mention a vendor by name, whereas in other threads vendors are introduced by name once and subsequent comments only refer to them using pronouns. Thus, we smoothen the effect with indicator variables. For Nemesis, we add as covariates the vendors' degree and various centralities, as well as the number of posts made and the number of posts deleted. However, experiments involving betweenness, eigenvector, and closeness failed to converge, so we omit them.

The hazards regression formula for the extended variables in Hansa and Nemesis are as follows:

$$h(t) = exp(\alpha + I(\text{Pos.Mention}) + I(\text{Neg.Mention})) \ ,$$

and

$$h(t) = exp(\alpha + \text{Deg.Cent.} + \text{Bet.Cent.Nr.Posts} + \text{Nr.Del.Posts}) \ .$$

[6] Some marketplaces present a "last seen" field in vendor profiles that seems to track login activity.

Table 2: Cox Proportional Hazards Regression across all 8 markets, where exp(c) indicates the hazard rate increase per unit increment. The regression was stratified based on the wealth quartile the vendors belonged to at the end of the market.

	Silk Road 1				**Pandora**				**Silk Road 2**				**Agora**			
Covariates	exp(c)	SE	z	p	exp(c)	SE	z	p	exp(c)	SE	z	p	exp(c)	SE	z	p
Avg. Feedback Value	0.50	0.16	-4.25	**<.005**	0.69	0.24	-1.53	0.13	0.51	0.10	-6.34	**<.005**	0.61	0.07	-7.48	**<.005**
Presence in Other Mkt.	–	–	–	–	0.85	0.14	-1.16	0.24	0.59	0.09	-5.95	**<.005**	0.67	0.07	-5.73	**<.005**
Mainly Digital	0.89	0.19	-0.62	0.53	0.78	0.37	-0.69	0.49	0.56	0.26	-2.24	**0.02**	0.80	0.21	0.29	0.29
Mainly Category A Drugs	1.12	0.18	0.61	0.54	1.10	0.34	0.29	0.77	0.87	0.23	-0.59	0.55	1.31	0.19	1.42	0.16
Mainly Category B Drugs	1.30	0.17	1.50	0.13	0.91	0.34	-0.27	0.78	0.78	0.24	-1.06	0.29	1.10	0.20	0.50	0.62
	Evolution				**Alphabay**				**Hansa**				**Nemesis**			
Covariates	exp(c)	SE	z	p	exp(c)	SE	z	p	exp(c)	SE	z	p	exp(c)	SE	z	p
Avg. Feedback Value	0.58	0.12	-4.44	**<.005**	0.67	0.05	-7.91	**<.005**	0.50	0.19	-3.59	**<.005**	0.36	0.18	-5.48	**<.005**
Presence in Other Mkt.	0.59	0.07	-7.26	**<.005**	0.68	0.05	-7.29	**<.005**	0.60	0.14	-3.77	**<.005**	1.32	0.31	0.89	0.37
Mainly Digital	0.64	0.21	-2.14	**0.03**	0.47	0.11	-7.09	**<.005**	0.78	0.41	-0.61	0.54	0.33	0.63	-1.75	0.08
Mainly Category A Drugs	0.69	0.21	-1.74	0.08	0.75	0.11	-2.69	**0.01**	1.90	0.40	1.62	0.10	0.61	0.65	-0.76	0.45
Mainly Category B Drugs	0.74	0.21	-1.43	0.15	0.75	0.11	-2.70	**0.01**	1.58	0.40	1.14	0.26	0.45	0.65	-1.23	0.22

Table 3: Cox Proportional Hazards Regression on forum features extracted for Hansa and Nemesis.

	Hansa-Extended			
Covariates	exp(c)	SE	z	p
Positive Mention	0.70	0.15	-2.43	**0.01**
Negative Mention	0.80	0.24	-0.93	0.35
	Nemesis-Extended			
Covariates	exp(c)	SE	z	p
In Degree	0.99	0.01	-1.18	0.24
Out Degree	0.02	1.02	1.71	0.09
Nr. of Posts	0.99	0.01	-0.47	0.64
Nr. of Upvotes	1.00	0.00	-0.26	0.80
Nr. of Del. Posts	1.04	0.06	0.65	0.52

4.2 Results

We find that the average reputation score of each vendor is significantly ($p < .005$) associated with a decrease in the hazard rate across all markets except Pandora, as seen in Table 2. The interpretation for the exponential of the coefficient (exp(c)) is that, for example, a one-unit increase in the average feedback value on Silk Road 1, corresponds to a 50% decrease in the hazard rate. We also observe the same significant reduction in the hazard rate on vendors who had a presence in other markets. We find more mixed effects on the category of drugs being sold. That is, whether the seller mainly class A "harder" or class B "softer" drugs has mixed impact on the hazard rate across markets. Vendors who focused on digital goods, however, were more consistently correlated with lower hazards with some significant effects ($p < .05$) observed in Silk Road 2, Evolution, and Alphabay.

In our extended experiments for Hansa, we found that positive mentions of vendors across subreddits decreased the hazard rate by 30% significantly ($p = .01$), as observed in Table 3. In the case of Nemesis, we did not observe significant effects across the measures of centrality that we tested, nor across the number of posts or deleted posts that vendors had.

4.3 Reputation Slander Attack

By leveraging the results from our survivability analysis we can conceptualize the cost and potential impact of a reputation attack. Past work suggested interventions that exacerbate information asymmetries in these markets to push them to failure [21, 32]; and showed that reduction in reputation may affect vendors' trade capabilities [6]. An example proposed by Franklin et al. in IRC-based markets was to use Sybils to slander the reputation of vendors [13].

Our results indicate that a slander campaign may only work if done through product reviews within the market and not in forums. In our model, we did not observe that negative mentions had an effect on survivability. Forum signals may in fact be too noisy to a prospective buyer. For instance, vendor visibility across posts could also help advertising. Likewise, negative comments are not always unilaterally accepted, instead they often draw debate and alternative experience reports from other buyers. This phenomenon was also noted by Morselli et al., when exploring conflict resolution techniques in criminal forums [27]. On the other hand, product review scores have a marked impact on survivability. Based on these results, we can infer the theoretical cost and impact of the attack as follows: we calculate the cost of decreasing a unit of average review score based on the lowest item cost. Let CA be a vendor's current average score, TF the total number of reviews they have received, L the lowest review that can be given, and F the number of feedback needed for the attack. Then,

$$\frac{CA \times TF + L \times F}{TF + F} = CA - 1 \ ,$$

and

$$Cost = F \times \text{Item Cost}\ .$$

Solving for F gives us the cost of a reputational attack on a given OAM vendor by increasing their hazard. As an example, the vendor "YOURDEALER" (one of the largest vendors in Nemesis, at the time of writing) has an average feedback score of 4.99 from a total of 742 reviews, and their lowest priced item is $9. It would take 254 1-star reviews for a total cost of $\sim \$2,286$ to reduce their average rating by 1 unit and thus increase their (predicted) hazard by 64%. In practice, less 1-star reviews might be sufficient to cause fear in future vendors. Furthermore, the cost could be further reduced by conducting these attacks early in a vendor's career.

5 Predicting Success and Longevity

We now explore whether we are able to predict the financial success of a vendor, and the variables that drive their success. For interpretability, we use standard decision tree-based models. We train and test a standard prediction model which does not capture time variation across variables, and a model which does. We then repeat our experiments with the additional variables from Hansa and Nemesis. Last, we explore the generalizability of our models by training and testing with different market combinations.

5.1 Predicting Future Financial Success

Given a set of observable features from a vendor at a given state of the market, our first goal is to predict the wealth tier (i.e., revenue quartile) to which the vendor will belong at some point in the future. We then repeat this process by incorporating temporal features and forum-derived features for Hansa and Nemesis.

We do not attempt to predict revenue directly because revenue estimates are noisy and can often be heavily biased by collection and inference factors [9]. Consider the case of Nemesis, where vendors can choose to create a listing with various price options, or create one listing per offering. Using feedback as proxy for sales, we have no way of inferring which option the buyer used. Thus, the range of potential revenue that we could estimate for the vendor is wide, depending on what price we choose to use for our proxy. Furthermore, using quartiles allows for evenly balanced prediction targets.

5.1.1 Experiments Setup

For each market, we split the market into weekly intervals. We label each vendor with the quartile they belong at the end of the market (i.e., the last week the market was active prior to a takedown, or in the case of Nemesis, the last week for which we have data collection). Then, we iteratively split our dataset into observation intervals up to a given week. At each time step, we train a model based on the state of the market at that time. As we include observations of the market, new vendors appear and the features evolve.

We first train a Random Forest Classifier (RF) on the observable vendors' base features (described in Section 3.4.1). A Random Forest model is an ensemble estimator that fits decision trees to various sub-samples of the data [31]. We also train two additional classifiers on Hansa and Nemesis with the additional features extracted from their corresponding forums.

We hypothesize that time and time variation of features carry signals which will improve our estimation task. For instance, we may want to capture vendors with first-mover advantage, or the momentum of sales that a vendor has from one time step to the next. Our base features can be made more expressive by adding a time dimension. Using only the base features defined in Section 3.4.1, we have a matrix of shape $(nr\_vendors \times nr\_features)$ up to time T in the market. However, we could also build a time series for some of the features by tracking the evolution of features over time, as described in Section 3.4.2. That is, we break T into a series of time steps t_i, and end up with a matrix of shape $(nr\_vendors \times nr\_timesteps \times nr\_features)$.

To conduct a classification task on our time series data, we train a Time Series Forest classifier (TSF). A TSF model extends a RF classifier by sub-sampling the input time series into slices of random lengths (denoted as "windows" in the model) and extracting the mean, the standard deviation, and the slope. Each of these windows can provide insights into the temporal characteristics of the input time series, allowing us to explore what windows and features were the most relevant in the prediction [11]. Similar to the RF classifier, the sub-trees in TSF choose a label using hard voting.

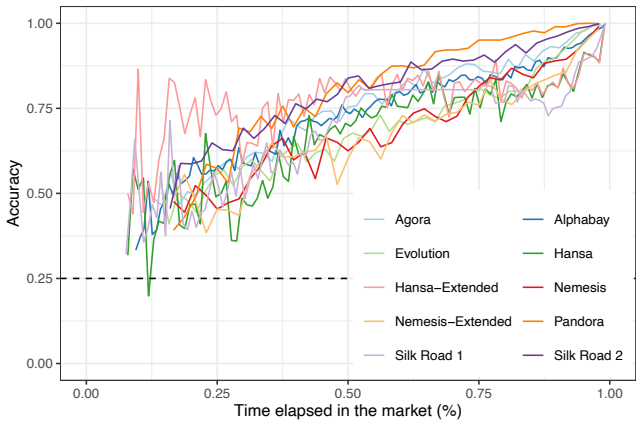
We repeat the same process we defined with our RF model for all markets. We use out-of-the-box parameters for our models: 100 estimators and maximum depth of 4 for both the RF and the TSF. TSF has an additional parameter: the number of windows. For this, we choose the number of timestamps as the number of windows. We use a 75/25% train test split.

5.1.2 Results

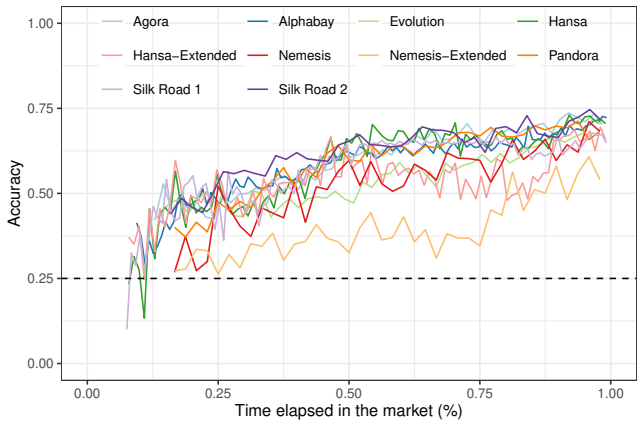
Our models perform better *without* the temporal features, as observed in Figure 2. Even when the model has access to almost a complete view of the market, the average accuracy plateaus at 70% for our TSF model. This indicates that having temporal features is detrimental to the model's performance, which is possibly caused by the hard voting mechanism the TSF model uses. The model may be learning features from earlier portions of a vendor's performance that may seem to indicate future success. Since it weighs these features equally to more recent data, the newer observations are unable to affect the final prediction.

On the other hand, we see that the RF model converges to

Figure 2: Accuracy in predicting the end of market wealth quantile a vendor belongs to across markets. Labels are balanced. Experiments were conducted in absolute time. However, for visual representation, markets' timesteps were scaled to the percent elapsed; 0.5 represents halfway through a market's lifetime. "Extended" indicates that we used additional subreddit/forum features. Decreases in accuracy are due to new vendor entrancy.



(a) RF model using base features.



(b) TSF model using temporal features.

a perfect accuracy as the market evolves. At 20% of the market's lifetime, we achieve over 40% of accuracy in predicting the vendors who would accrue the most wealth by the end of the market. At 40% of the market's life time our accuracy is mostly over 60%. And by 80% of the market, we have over 75% accuracy, and over 90% accuracy for two markets.

Last, we find that the additional forum features seem to have little effect on the prediction accuracy. Hansa's forum features decrease the accuracy of the model. Nemesis shows the opposite. In either case, the effect is small in the RF model. On the other hand, we see a significantly higher negative effect in the TSF model. In the case of "Nemesis-Extended," we see an accuracy of 5–10% decrease at each time step, as well as less convergence towards the end of the market. In this case as well, the forum activity habits of vendors of different sizes may not be sufficiently distinct for these features to carry a meaningful signal, which ultimately confounds the model.

5.2 Predicting Vendor Disappearance

We now attempt to build a model to predict whether a vendor will leave the market or is at risk of doing so. Similar to before, we consider a vendor to have disappeared from the market at the time they stop receiving feedback. To do this, we employ the base features we described in Section 3.4.1. Furthermore, we design our experiments to combine observations across markets for generalizability.

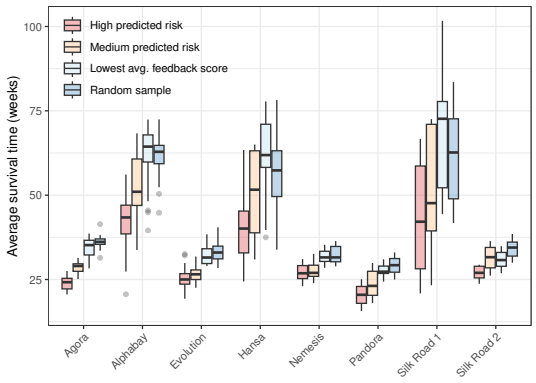
5.2.1 Experiment Setup

Our goal is to identify the vendors who are on the brink of leaving the market. To do this we design a classifier that attempts to predict one of the following: 1) whether a given vendor will leave the market in the next month (high-risk), 2) whether the vendor will leave the market after the first month but before the third month, or 3) whether the vendor will still be active after the third month.

For each market, we split the market lifetime into weekly intervals. At each stage of the market, we label the vendors according to the labels above. We then combine data from different markets. However, given that our prediction goal is not end-of-market revenue, we do not combine them based on the percent of revenue accrued by the market. Instead, we naïvely combine vendors from different markets based on the amount of time elapsed in the market. That is, we combine observations from a vendor from market M at week W with a vendor from market M' at week W. Further, as vendors disappear from the market, we remove them from our sample (so as to not overfit on already disappeared vendors).

We then train and evaluate an RF model and a TSF model with default settings, similar to our experiment setup in Section 5.1.1, at each timestep. We note, however, that both models perform poorly ($F_1 < 0.25$ for high/mid-risk vendors) due to class imbalance (i.e., not a lot of vendors disappear within 1–3 months). Therefore, instead of evaluating our model based on label prediction, we collect the label probabilities for each class. To do this, we pick the classifier with the best F_1 score for high/mid-risk labels, trained on a subset of data from all markets. We then use this classifier to assemble, for each timestep, the 20% of vendors with: 1) the highest probabilities in the high-risk class, 2) the highest probabilities in the mid-risk class, 3) the lowest average feedback score, 4) and a random sample. We take the average survival time for each of these groups, at each timestep of each market. We chose

Figure 3: Average survival time in weeks for each group, across all market timesteps. Vendors who stop having sales after a given week are removed from the sample. The high/mid-risk groups across markets were assembled with the same classifier, which was trained with samples from all markets.



20% as it provided a big enough sample size for each market, while reducing the overlap between vendors across groups.

5.2.2 Results

Across all markets, the group of vendors assigned the highest probabilities of being at "high-risk" indeed had shorter lifespans as compared to the other groups, as seen in Figure 3. Vendors in the "mid-risk" category also had shorter lifespans than the other groups, except for Silk Road 2. We observe that a low average feedback score seems to carry some signal of quality, given that vendors with low average feedback score have, for the most part, slightly shorter lifespans than a random sample. However, we observe that a low average feedback score, alone, may not be clearly indicative of a near-term disappearance from the market. For instance, established vendors may have a dip in their average feedback score in a given week, but that may not necessarily shorten their lifespan significantly. Our main finding is that making a prediction on the lifespan of a vendor, may depend on more variables beyond just average feedback scores.

6 Generalizability and Feature Importance

For our vendor disappearance model, we trained our model by mixing vendor observations from different markets because the event of interest is whether a vendor stopped receiving feedback. That is, the labels are not significantly different across markets. We claim generalizability for this model, given that we used a single classifier, trained on traces from all markets to do predictions for each of these markets across each of the markets' lifetimes.

For our financial success model, however, our labels are the end-of-market revenue quantiles for a given market. Thus, we cannot directly combine vendor observations from different markets For instance, vendor V from market M may belong to Q1 with \$1M revenue, whereas vendor V' from market M' may belong to Q3 with \$1M of revenue. Furthermore, in our vendor disappearance model we were able to directly combine our traces based on the time elapsed in the market. However, revenue is trickier. Consider the case of Silk Road and Hansa. Silk Road was the first successful market, facing little competition in its early stages. Hansa on the other hand, was a market that had little traction for over a year, and gained most of its revenue following the Alphabay takedown. If we combine vendors' data from the first month of Silk Road with the data from the first month of Hansa, we are combining two disparate market environments. Instead, we combine market's data when their environments were most similar.

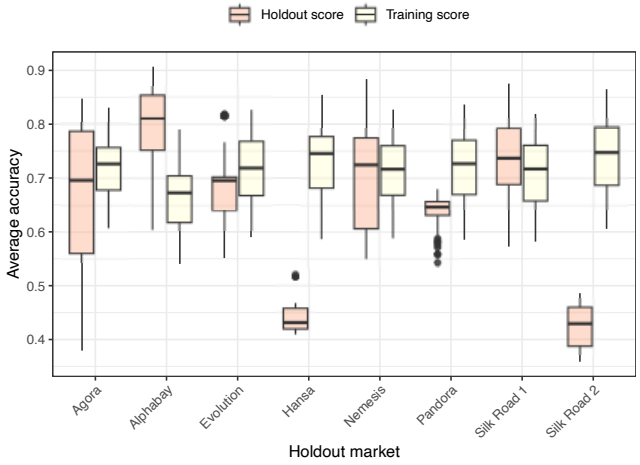
Thus, to explore the generalizability of our financial prediction model. We design an experiment where we train a model on $n-1$ markets and predict on an unseen market. Further, we combine cross-market observations by combining traces at stages where the markets had accrued a similar revenue percentage. We perform this experiment with all vendors, and also by segmenting vendors by the main category they sold. Last, we discuss feature importances across each model.

6.1 Experiment Setup

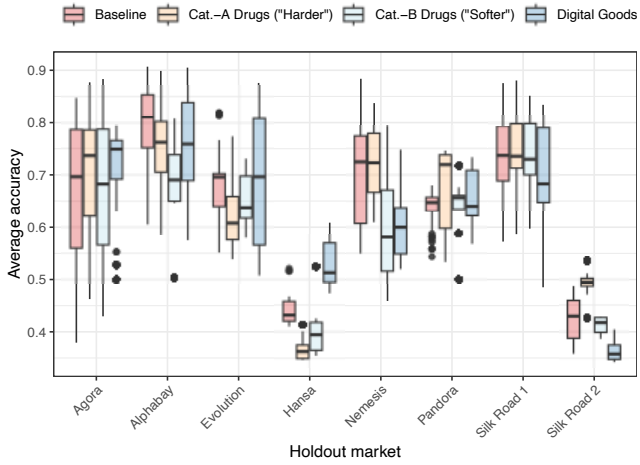
We want to repeat the experiments defined in Section 5.1.1 by training a model on a set of markets and testing our prediction on an unseen market. To combine observations across different markets, we explore a simple heuristic: splitting the data by the percent of revenue accrued by the market. That is, we iteratively split each markets' data at the time they accrued 10%, 20%, ..., 90% of the revenue at the time of their last observation. We then iteratively combine the data of $n-1$ markets to train our model, leaving one market out completely (which we call our *holdout market*). We train and evaluate each model on these $n-1$ markets using a 75/25% train/test split. We then test the performance of our model on our holdout market. Finally, for each model, we conduct an ablation study by iteratively removing each of the feature categories described in Section 3.4.1: listing features, revenue features, and feedback/reputation features.

We hypothesized that our classifier accuracy could be improved by segmenting vendors by category. We used the same labeling of Section 4. We combined vendors who sell mainly category A ("harder") drugs, B ("softer") drugs, and digital goods. For each market, we followed the same procedure as mentioned above, except that we only trained and tested on one category at a time.

Figure 4: Average prediction accuracy for the revenue quantile that a vendor belongs to by the end of the market. The prediction accuracy is averaged across various revenue stages of the market (average accuracy). We train and evaluate our model on $n-1$ markets and test on the holdout market.



(a) Comparison of train/test scores. The training score is the performance on the split of $n-1$ markets, whereas the test score is the score on the holdout.



(b) Comparison of test scores across categories. Category segmentation uses only vendors whose primary good sold is in the given category. Baseline is without segmentation.

6.2 Results

Our financial success model generalizes well to 6/8 other markets even when using a naïve heuristic to combine markets' data, as seen in Figure 4a. Across all markets, the average accuracy during evaluation stayed consistent. This means that even when we shuffled vendors from different markets during our train/test split, we were able to maintain a consistent accuracy of over 70% for all markets except Alphabay, which was 67%. Furthermore, in 5/8 markets we observed similar performance between the accuracy during training/evaluation and the accuracy during testing. This means that our model was able to perform well when doing prediction on vendors from a completely unseen market. The results from Nemesis indicate generalizability across time, given that Nemesis is significantly more recent than some markets (e.g., it appeared 10 years later than Silk Road).

With regards to our category segmentation approach, we observe mixed results across markets, as seen in Figure 4b. In general, we do not see significant improvements/deterioration in performance over our baseline across markets. This could be due the categories being too broad, due to a reduction in the size of the training set, due to our current approach at combining market segments, and/or due to different category performance dynamics across markets (e.g., market X is more popular for drug A, whereas market Y is more popular for drug B). Nonetheless, we believe some form of segmentation is useful, but will likely require market-specific optimizations.

In Table 4, we show the precision, recall, and F1 scores for our experiments with each holdout market, across our 4 revenue quantiles. The model performs best when doing predictions on the lowest/highest earners (Q1 and Q4). Because the overall market revenue follows a power law distribution, the middle portion (Q2 and Q3) are harder to distinguish. Last, in Table 5, we show that the absence of revenue-related features decreases accuracy the most. When only reputation features are excluded, accuracy is barely affected. When revenue and reputation features are both excluded, the model suffers the biggest loss. Listing-related features have little impact on the model.

6.3 Explaining and Improving Performance

We hypothesize that the poor performance on Silk Road 2 and Hansa is due to the unique environments that these markets faced, as seen in Figure 1. Essentially, vendors in these markets may be considered to be out of distribution. Hansa had unremarkable economic activity until two months before its takedown. Following the Alphabay takedown, about 5,000 users a day flocked to Hansa [15]. Intuitively, models that were trained in economic activity from markets that did not experience the same trajectory are bound to have poor performance, as observed in Figure 2. In the case of SR2, this market had strong performance from the beginning likely due to its brand recognition after the original Silk Road's takedown. However, its performance gradually degraded due to a series of issues (arrest of moderators and a hack) [23], as opposed to gradually ramping up. Because of this, a naïve model that trains on markets dissimilar to Silk Road 2 yields lower quality results.

Table 4: Average classification metrics across our 4 labels (wealth tiers). The holdout is the market on which we predict while training on the others. Labels are balanced across classes. Each metric is the average score obtained across our 10 experiments.

Wealth Tier	**Q1 ($x \leq 25\%$)**			**Q2 ($25\% < x \leq 50\%$)**			**Q3($50\% < x \leq 75\%$)**			**Q4($75\% < x$)**		
Holdout Market	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Silk Road	0.77	0.82	0.79	0.63	0.72	0.67	0.67	0.62	0.64	0.88	0.80	0.83
Pandora	0.83	0.55	0.65	0.47	0.50	0.48	0.54	0.64	0.58	0.81	0.84	0.82
Silk Road 2.0	0.76	0.31	0.43	0.22	0.16	0.18	0.23	0.19	0.20	0.54	0.93	0.68
Agora	0.64	0.76	0.67	0.53	0.66	0.58	0.63	0.62	0.62	0.92	0.69	0.77
Evolution	0.74	0.99	0.84	0.57	0.78	0.65	0.56	0.52	0.54	1.00	0.54	0.69
Alphabay	0.81	0.96	0.88	0.74	0.78	0.76	0.77	0.64	0.69	0.87	0.83	0.85
Hansa	0.48	1.00	0.65	0.26	0.38	0.31	0.28	0.20	0.23	1.00	0.40	0.57
Nemesis	0.68	0.96	0.79	0.65	0.72	0.68	0.66	0.65	0.64	0.99	0.62	0.76
Average:	0.71	0.79	0.71	0.51	0.59	0.54	0.54	0.51	0.52	0.88	0.71	0.75

Table 5: Ablation study of our revenue prediction model on holdout markets. We exclude combinations of features and quantify the accuracy decrease on the model.

Excluded Feature(s)		**Avg. Accuracy Decrease**			
Feature Set 1	**Feature Set 2**	**Min.**	**Max.**	**Mean**	**Std.**
Revenue	–	0.01	0.31	0.16	0.08
Reputation	–	<0.01	<0.01	<0.01	<0.01
Listing	–	<0.01	0.02	<0.01	<0.01
Revenue	Reputation	0.05	0.44	0.27	0.12
Revenue	Listing	0.04	0.31	0.16	0.09
Reputation	Listing	0.02	0.02	<0.01	<0.01

While segmentation may not offer substantial improvements, a set of approaches can be adopted at other stages of the pipeline. For consistency, we conducted our holdout experiments by training on $n-1$ markets. However, some markets have uncommon trajectories (e.g., SR2, Hansa). In a practical setting, we may need to curate our training set based on the target market. For example, if the target market was born in response to a takedown, or faces more/less competitors, we ought train our models on markets with similar characteristics. With regards to our prediction goal, we naïvely consider everybody in a quartile to have the same label (a classification task). Instead, we could design our model to be a regression tree over vendor revenue percentiles to preserve relative ordering within vendors; we could also define arbitrary cutoffs (e.g., top 5% of vendors). Last, our models can be improved with traditional machine learning optimizations: testing other models (potentially trading explainability for accuracy) and finetuning parameters.

7 Discussion

Our results from Section 4 indicate that reputation, derived from feedback scores, plays a role both in the financial success as well as in the longevity of vendors, although in different forms. Our proportional-hazards regression shows that average feedback scores in the market have a significant impact on the survivability of a vendor. Across the board, we see that a 1-unit increase in average reputation reduces the hazard rate of vendors across the markets. However, this regression leverages a full view of the market. That is, as a whole, feedback scores can explain the disappearance of vendors.

On the other hand, our results from Section 5.2 show that the average feedback score is not the best predictor of a vendor leaving the market in the short term. That is, as the market progresses, vendors with lowest average score may not necessarily leave the market. This effect surfaces on the markets that have a longer lifetime (i.e., SR1, Hansa, Alphabay). Instead, our model, by leveraging more vendor features, better identifies vendors at a higher risk of disappearing. Furthermore, our model generalizes across markets and time, given that it was trained on vendor observations from 8 different markets spanning 12 years.

With regards to the financial success of vendors, we demonstrated that our predictions generalize across most markets. The average feedback score seems to play a role in predicting their future wealth. However, it is not the main predictor. Rather, past financial performance is a better predictor of future financial performance. In part, we hypothesize that this is the case because scaling criminal operations is hard, particularly for drug-related items [16]. Thus, vendors who demonstrate capacity to scale their business early (as demonstrated by large sales) often become dominant vendors. Another reason why sales volume and history are likely drivers of success is because these signals are hard to fake. Décary-Hétu et al. noted that signals which could be cheaply purchased had little impact in predicting sales [10]. Frequent sales, over time, ultimately create an attractive signal for buyers who want to reduce risk. When segmenting vendors by category, however, we do not observe a significant difference across markets. We believe that segmentation may help, but may need to rely on other approaches to combining market data and accounting for vendor offering diversity. Similarly, our time-series

model performed significantly worse and may also benefit from ddiferent feature engineering approaches. Remarkably, we did not perform any parameter tuning; instead, we employed out-of-the-box defaults. We did not employ sophisticated feature transformations nor models. Rather we focused on explainability and establishing a performance lower bound.

Across our experiments we did not observe a significant effect from signals derived from forums, neither from the co-located forum (for Nemesis), nor from the external forums (subreddits). During our manual analysis, we observed that forum signals are predominantly noisy. We observed small vendors that frequently used forums for advertising. A vendor who posts a lot can easily build an impressive social network through their interactions, despite not driving sales. We also observed large vendors who were not mentioned even once, and who also did not engage in any of the forums. Furthermore, negative reviews in forums were often not unilaterally accepted but often raised discussions from other users in the community, a similar finding to Morselli et al. [27].

Interventions and Policy Takeaways Our results can help improve interventions in two ways. First, our prediction model can readily be used in new markets to identify vendors who will become big earners. Early identification allows for monitoring efforts to be more efficient, particularly as OAMs and criminal forums increasingly adopt adversarial anti-scraping mechanisms [40]. Cuevas et al. demonstrated that focusing scraping efforts on more popular vendors using a naïve algorithm improved coverage and inferences substantially [9]; our results build on that approach. Our prediction model ought to be taken probabilistically: not as a definitive answer, but as a tool that can help navigate uncertainty. Second, we show that slander attacks may be viable and cost-effective, particularly when done early in a vendor's career. Our findings suggest, however, that slander attacks ought to be done through low score feedback orders and not through slander in forums.

With regards to market design and policy, our results demonstrate that existing reputation systems within these markets carry a signal that can help reduce harm in the long run. However, this signal is imperfect and may not have a strong enough effect in the short term. On one hand, the continued success of these markets are testament to the fact that existing reputation systems are, however crudely, culling out low-quality vendors. On the other hand, our simple classifier demonstrates that there are other signals which seem to more readily identify vendors who might disappear from the market. While there are a variety of benign reasons why a vendor may leave a market, there are some quite harmful ones, such as vendors who sell dangerously adulterated drugs. A model or signals which can more quickly alert buyers of these situations can substantially reduce harm in the long run. Policies which consider the regulation of two-sided marketplaces (particularly for drugs), ought to consider the reputation system design as well.

Limitations and Future Work First, we do not test a large number of covariates through our proportional hazards model, because a "one-in-ten/twenty" rule (1 covariate for every 10/20 deaths) is advised for proportional hazards model [17]. Thus, while we identified a set of meaningful covariates contributing to vendor survivability, there may be other latent factors which our model does not capture. Second, our financial success prediction model only predicts the wealth quantile that a vendor will belong to. Within the top 25% of vendors there may be significant variance in revenue. Third, we only tested the impact of external reputation signals for one market (Hansa) and social network features for one market (Nemesis). Our manual review of these signals indicates high noise, particularly as it relates to the success of vendors. However, these features may correlate with other vendor attributes which future work may explore. In our study, we saw less accuracy from the TSF model which sought to capture time-based feature changes. However, it may be useful to explore other feature engineering approaches that incorporate temporal features. Furthermore, we leveraged qualitative analysis to extract signals from forums in an effort to collect high-fidelity signals. However, scaling this work manually is inefficient. Current off-the-shelf named-entity recognition and sentiment analysis techniques did not perform well on our dataset. However, advances in large language models, particularly for coding textual data [42, 48], may allow our forum analyses to scale.

8 Conclusion

We conducted a set of experiments to understand the role of reputation on the financial success and longevity of vendors across OAMs. By leveraging manually-coded reputation signals for one market, and interaction signals from a forum in another market, we found little evidence that reputation signals outside of the market have an impact on a vendor's financial success and/or longevity. Instead, feedback scores (from product reviews within the market) seem to, in the long run, push out low-quality vendors.

However, in the short-term, feedback scores may not be good enough to predict whether a vendor will leave the market. With a simple model, we could better identify which vendors were more likely to leave a market in the short term. Furthermore, we found that we can predict which vendors will become the biggest earners with decent accuracy early in the market. While reputation helps, it is not the main signal of success. Rather, vendors who demonstrate that they can bring in significant revenue early on, end up becoming the largest vendors. Both of our models generalize across markets and time periods, with out-of-the-box parameters and without extensive feature engineering.

In summary, current reputation signals play a role in vendor success and longevity, but are coarse and can be improved, which would help both policing and harm reduction.

Acknowledgments

This research was partially supported by the Defence Science and Technology Agency (DSTA) and by a Carnegie Mellon CyLab Presidential Fellowship. We thank our anonymous reviewers and shepherd for suggested changes that greatly improved this paper. This work originally stemmed from discussions with Kyle Soska, Xiao Hui Tai, and Behtash Banihashemi. It subsequently benefited tremendously from many suggestions (and careful reads) by Rolf van Wegberg, Fieke Miedema, and their group at TU Delft. Finally, Rohan Sonecha helped with the manual coding of forum post sentiment.

References

[1] Pushshift. https://pushshift.io/. Accessed: September 19, 2023.

[2] H. Bar-Isaac and S. Tadelis. Seller reputation. *Foundations and Trends in Microeconomics*, 4(4):273–351, 2008.

[3] M. J. Barratt, J. A. Ferris, and A. R. Winstock. Use of silk road, the online drug marketplace, in the united kingdom, australia and the united states. *Addiction*, 109(5):774–783, 2014.

[4] M. J. Barratt, J. A. Ferris, and A. R. Winstock. Safer scoring? cryptomarkets, social supply and drug market violence. *International Journal of Drug Policy*, 35:24–31, 2016.

[5] T. M. Booij, T. Verburgh, F. Falconieri, and R. S. van Wegberg. Get rich or keep tryin' trajectories in dark net market vendor careers. In *IEEE European Symposium on Security and Privacy Workshops (EuroS&PW'21)*, pages 202–212, Virtual Conference, 2021.

[6] C. Bradley. *On the Resilience of the Dark Net Market Ecosystem to Law Enforcement Intervention*. PhD thesis, UCL (University College London), 2019.

[7] A. Caines, S. Pastrana, A. Hutchings, and P. J. Buttery. Automatically identifying the function and intent of posts in underground forums. *Crime Science*, 7(1):19, 2018.

[8] N. Christin. Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace. In *The ACM Web Conference (WWW'13)*, pages 213–224, Rio de Janeiro, Brazil, May 2013.

[9] Alejandro Cuevas, Fieke Miedema, Kyle Soska, Nicolas Christin, and Rolf van Wegberg. Measurement by proxy: On the accuracy of online marketplace measurements. In *31st USENIX Security Symposium (USENIX Security'22)*, pages 2153–2170, Boston, USA, 2022.

[10] David Décary-Hétu and Anna Leppänen. Criminals and signals: An assessment of criminal performance in the carding underworld. *Security Journal*, 29:442–460, 2016.

[11] H. Deng, G. Runger, E. Tuv, and M. Vladimir. A time series forest for classification and feature extraction. *Information Sciences*, 239:142–153, 2013.

[12] L. Franceschi-Bicchierai. Reddit bans subreddits dedicated to dark web drug markets and selling guns. VICE, 2018.

[13] J. Franklin, A. Perrig, V. Paxson, and S. Savage. An inquiry into the nature and causes of the wealth of internet miscreants. In *14th ACM Conference on Computer and Communications Security (CCS'07)*, volume 7, pages 375–388, Virginia, USA, October 2007.

[14] Diego Gambetta. *Codes of the Underworld: How Criminals Communicate*. Princeton University Press, 2009.

[15] A. Greenberg. How dutch police took over hansa, a top dark web market. Accessed: September 19, 2023.

[16] E. Hammersvik, S. Sandberg, and W. Pedersen. Why small-scale cannabis growers stay small: Five mechanisms that prevent small-scale growers from going large scale. *International Journal of Drug Policy*, 23(6):458–464, 2012.

[17] F. E. Harrell, K. L. Lee, and D. B. Mark. Multivariate prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387, 1996.

[18] T. J. Holt. Examining the forces shaping cybercrime markets online. *Social Science Computer Review*, 31(2):165–177, 2013.

[19] T. J. Holt. Exploring the social organisation and structure of stolen data markets. *Global Crime*, 14(2-3):155–174, 2013.

[20] T. J. Holt, O. Smirnova, Y. T. Chua, and H. Copes. Examining the risk reduction strategies of actors in online criminal markets. *Global Crime*, 16(2):81–103, 2015.

[21] T. J. Holt, O. Smirnova, and A. Hutchings. Examining signals of trust in criminal markets online. *Journal of Cybersecurity*, 2(2):137–145, 2016.

[22] J. Hughes, B. Collier, and A. Hutchings. From playing games to committing crimes: A multi-technique approach to predicting key actors on an online gaming forum. In *APWG Symposium on Electronic Crime Research (eCrime'19)*, pages 1–12, Pennsylvania, USA, November 2019.

[23] R. Mac. Silk road 2.0's blake benthall arrested, charged with running the massive dark web drug site. https://www.forbes.com/sites/ryanmac/2014/11/06/silk-road-2-blake-benthall-fbi-shutdown/?sh=7bdde138170f, 2014.

[24] A. Maddox, M. J. Barratt, M. Allen, and S. Lenton. Constructive activism in the dark web: Cryptomarkets and illicit drugs in the digital 'demimonde'. *Information, Communication & Society*, 19(1):111–126, 2016.

[25] J. Martin. *Drugs on the Dark Net: How Cryptomarkets are Transforming the Global Trade in Illicit Drugs*. Springer, 2014.

[26] J. Martin and N. Christin. Ethics in cryptomarket research. *International Journal of Drug Policy*, 25:84–91, 2016.

[27] Carlo Morselli, David Décary-Hétu, Masarah Paquet-Clouston, and Judith Aldridge. Conflict management in illicit drug cryptomarkets. *International Criminal Justice Review*, 27(4):237–254, 2017.

[28] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker. An analysis of underground forums. In *ACM Internet Measurement Conference (IMC'11)*, pages 71–80, Berlin, Germany, November 2011.

[29] D. J. Nutt, L. A. King, L. D. Phillips, et al. Drug harms in the uk: A multicriteria decision analysis. *The Lancet*, 376(9752):1558–1565, 2010.

[30] R. Overdorf, C. Troncoso, R. Greenstadt, and D. McCoy. Under the underground: Predicting private interactions in underground forums. arXiv (1805.04494), 2018.

[31] M. Pal. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222, 2005.

[32] Masarah Paquet-Clouston, David Décary-Hétu, and Carlo Morselli. Assessing market competition and vendors' size and scope on alphabay. *International Journal of Drug Policy*, 54:87–98, 2018.

[33] S. Pastrana, A. Hutchings, A. Caines, and P. Buttery. Characterizing eve: Analysing cybercrime actors in a large underground forum. In *International Symposium on Research in Attacks, Intrusions, and Defenses (RAID'18)*, pages 207–227, Heraklion, Greece, September 2018.

[34] P. Reuter and J. P. Caulkins. Illegal 'lemons': Price dispersion in cocaine and heroin markets. *Bulletin on Narcotics*, 56(1-2):141–165, 2004.

[35] K. Soska and N. Christin. Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. In *24th USENIX Security Symposium (USENIX Security'15)*, pages 33–48, Washington (DC), USA, August 2015.

[36] stuck-in-the matrix and Watchful1. 'reddit comments/submissions 2005-06 to 2022-12'.

[37] Z. Sun, C. E. Rubio-Medrano, Z. Zhao, T. Bao, A. Doupé, and G. Ahn. Understanding and predicting private interactions in underground forums. In *9th ACM Conference on Data and Application Security and Privacy (CODASPY'19)*, pages 303–314, Texas, USA, March 2019.

[38] S. Tadelis. Reputation and feedback systems in online platform markets. *Annual Review of Economics*, 8(1):321–340, 2016.

[39] X. H. Tai, K. Soska, and N. Christin. Adversarial matching of dark net market vendor accounts. In *25th ACM SIGKDD Conference on Knowledge, Discovery, and Data Mining (KDD'19)*, pages 1871–1880, Alaska, USA, August 2019.

[40] Kieron Turk, Sergio Pastrana, and Ben Collier. A tight scrape: methodological approaches to cybercrime research data collection in adversarial environments. In *5th IEEE European Symposium on Security and Privacy Workshops (EuroS&PW'20)*, pages 428–437, Virtual Conference, 2020.

[41] M. Tzanetakis, G. Kamphausen, B. Werse, and R. von Laufenberg. The transparency paradox. building trust, resolving disputes and optimising logistics on conventional and online drugs markets. *International Journal of Drug Policy*, 35:58–68, 2016.

[42] P. Törnberg. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. arXiv (2304.06588), 2023.

[43] 'u/lift_ticket83'. Reddit data api update: Changes to pushshift access. Website. Accessed: September 19, 2023.

[44] J. van de Laarschot and R. van Wegberg. Risky business? investigating the security practices of vendors on an online anonymous market using ground-truth data. In *30th USENIX Security Symposium (USENIX Security'21)*, pages 4079–4095, Virtual Conference, 2021.

[45] R. van Wegberg, F. Miedema, U. Akyazi, A. Noroozian, B. Klievink, and M. van Eeten. Go see a specialist? predicting cybercrime sales on online anonymous markets from vendor and product characteristics. In *The ACM Web Conference (WWW'20)*, page 816–826, Taipei, Taiwan, April 2020.

[46] R. Van Wegberg, S. Tajalizadehkhoob, K. Soska, U. Akyazi, C. H. Ganan, B. Klievink, and N. Christin. Plug and prey? measuring the commoditization of cybercrime via online anonymous markets. In *27th USENIX Security Symposium (USENIX Security'18)*, pages 1125–1141, Maryland, USA, August 2018.

[47] F. Wehinger. The dark net: Self-regulation dynamics of illegal online markets for identities and related services. In *European Intelligence and Security Informatics Conference (EISIC'11)*, pages 209–213, Athens, Greece, September 2011.

[48] Z. Xiao, X. Yuan, Q. V. Liao, R. Abdelghani, and P. Oudeyer. Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. In *28th ACM International Conference on Intelligent User Interfaces (IUI'23)*, pages 75–78, Sydney, Australia, 2023.

[49] Y. Zhang, Y. Fan, Y. Ye, L. Zhao, and C. Shi. Key player identification in underground forums over attributed heterogeneous information network embedding framework. In *28th ACM International Conference on Information and Knowledge Management (CIKM'19)*, pages 549–558, Beijing, China, November 2019.