Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem

Kyle Soska and Nicolas Christin, ***Carnegie Mellon University***

https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/soska

This paper is included in the Proceedings of the 24th USENIX Security Symposium

August 12–14, 2015 • Washington, D.C.

ISBN 978-1-939133-11-3

Open access to the Proceedings of the 24th USENIX Security Symposium is sponsored by USENIX

Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem

Kyle Soska and Nicolas Christin
Carnegie Mellon University
{ksoska, nicolasc}@cmu.edu

Abstract

February 2011 saw the emergence of Silk Road, the first successful online anonymous marketplace, in which buyers and sellers could transact with anonymity properties far superior to those available in alternative online or offline means of commerce. Business on Silk Road, primarily involving narcotics trafficking, rapidly boomed, and competitors emerged. At the same time, law enforcement did not sit idle, and eventually managed to shut down Silk Road in October 2013 and arrest its operator. Far from causing the demise of this novel form of commerce, the Silk Road take-down spawned an entire, dynamic, online anonymous marketplace ecosystem, which has continued to evolve to this day. This paper presents a long-term measurement analysis of a large portion of this online anonymous marketplace ecosystem, including 16 different marketplaces, over more than two years (2013–2015). By using long-term measurements, and combining our own data collection with publicly available previous efforts, we offer a detailed understanding of the growth of the online anonymous marketplace ecosystem. We are able to document the evolution of the types of goods being sold, and assess the effect (or lack thereof) of adversarial events, such as law enforcement operations or large-scale frauds, on the overall size of the economy. We also provide insights into how vendors are diversifying and replicating across marketplaces, and how vendor security practices (e.g., PGP adoption) are evolving. These different aspects help us understand how traditional, physical-world criminal activities are developing an online presence, in the same manner traditional commerce diversified online in the 1990s.

1 Introduction

In February 2011, a new Tor hidden service [16], called "Silk Road," opened its doors. Silk Road portrayed itself as an online anonymous marketplace, where buyers and sellers could meet and conduct electronic commerce transactions in a manner similar to the Amazon Marketplace, or the fixed price listings of eBay. The key innovation in Silk Road was to guarantee stronger anonymity properties to its participants than any other online marketplace. The anonymity properties were achieved by combining the network anonymity properties of Tor hidden services—which make the IP addresses of both the client and the server unknown to each other and to outside observers—with the use of the pseudonymous, decentralized Bitcoin electronic payment system [33]. Silk Road itself did not sell any product, but provided a feedback system to rate vendors and buyers, as well as escrow services (to ensure that transactions were completed to everybody's satisfaction) and optional hedging services (to buffer fluctuations in the value of the bitcoin).

Embolden by the anonymity properties Silk Road provided, sellers and buyers on Silk Road mostly traded in contraband and narcotics. While Silk Road was not the first venue to allow people to purchase such goods online—older forums such at the Open Vendor Database, or smaller web stores such as the Farmer's Market predated it—it was by far the most successful one to date at the time due to its (perceived) superior anonymity guarantees [13]. The Silk Road operator famously declared in August 2013 in an interview with Forbes, that the "War on Drugs" had been won by Silk Road and its patrons [18]. While this was an overstatement, the business model of Silk Road had proven viable enough that competitors, such as Black Market Reloaded, Atlantis, or the Sheep Marketplace had emerged.

Then, in early October 2013, Silk Road was shut down, its operator arrested, and all the money held in escrow on the site confiscated by law enforcement. Within the next couple of weeks, reports of Silk Road sellers and buyers moving to Silk Road's ex-competitors (chiefly, Sheep Marketplace and Black Market Reloaded) or starting their own anonymous marketplaces started to surface. By early November 2013, a novel incarnation

of Silk Road, dubbed "Silk Road 2.0" was online—set up by former administrators and vendors of the original Silk Road.[1] Within a few months, numerous marketplaces following the same model of offering an online anonymous rendez-vous point for sellers and buyers appeared. These different marketplaces offered various levels of sophistication, durability and specialization (drugs, weapons, counterfeits, financial accounts, ...). At the same time, marketplaces would often disappear, sometimes due to arrests (e.g., as was the case with Utopia [19]), sometimes voluntarily (e.g., Sheep Marketplace [34]). In other words, the anonymous online marketplace ecosystem had evolved significantly compared to the early days when Silk Road was nearly a monopoly.

In this paper, we present our measurements and analysis of the anonymous marketplace ecosystem over a period of two and a half years between 2013 and 2015. Previous studies either focused on a specific marketplace (e.g., Silk Road [13]), or on simply describing high-level characteristics of certain marketplaces, such as the number of posted listings at a given point in time [15].

By using long-term measurements, combining our own data collection with publicly available previous efforts, and validating the completeness of our dataset using capture and recapture estimation, we offer a much more detailed understanding of the evolution of the online anonymous marketplace ecosystem. In particular, we are able to measure the effect of the Silk Road takedown on the overall sales volume; how reported "scams" in some marketplaces dented consumer confidence; how vendors are diversifying and replicating across marketplaces; and how security practices (e.g., PGP adoption) are evolving. These different aspects paint what we believe is an accurate picture of how traditional, physical-world criminal activities are developing an online presence, in the same manner traditional commerce diversified online in the 1990s.

We discover several interesting properties. Our analysis of the sales volumes demonstrates that as a whole the online anonymous marketplace ecosystem appears to be resilient, on the long term, to adverse events such as law enforcement take-downs or "exit scams" in which the operators abscond with the money. We also evidence stability over time in the types of products being sold and purchased: cannabis-, ecstasy- and cocaine-related products consistently account for about 70% of all sales. Analyzing vendor characteristics shows a mix of highly specialized vendors, who focus on a single product, and sellers who sell a large number of different products. We also discover that vendor population has long-tail characteristics: while a few vendors are (or were) highly successful, the vast majority of vendors grossed less than $10,000 over our entire study interval. This further substantiates the notion that online anonymous marketplaces are primarily competing with street dealers, in the retail space, rather than with established criminal organizations which focus on bulk sales.

The rest of this paper is structured as follows. Section 2 provides a brief overview of how the various online marketplaces we study operate. Section 3 describes our measurement methodology and infrastructure. Section 4 presents our measurement analysis. We discuss limitations of our approach and resulting open questions in Section 5, before introducing the related work in Section 6 and finally concluding in Section 7.

2 Online Anonymous Marketplaces

The sale of contraband and illicit products on the Internet can probably be traced back to the origins of the Internet itself, with a number of forums and bulletin board systems where buyers and sellers could interact.

However, online markets have met with considerable developments in sophistication and scale, over the past six years or so, going from relatively confidential "classifieds"-type of listings such as on the Open Vendor Database, to large online anonymous marketplaces. Following the Silk Road blueprint, modern online anonymous markets run as Tor hidden services, which gives participants (marketplace operators and participants such as buyers and sellers) communication anonymity properties far superior to those available from alternative solutions (e.g., anonymous hosting); and use pseudonymous online currencies as payment systems (e.g., Bitcoin [33]) to make it possible to exchange money electronically without the immediate traceability that conventional payment systems (wire transfers, or credit card payments) provide.

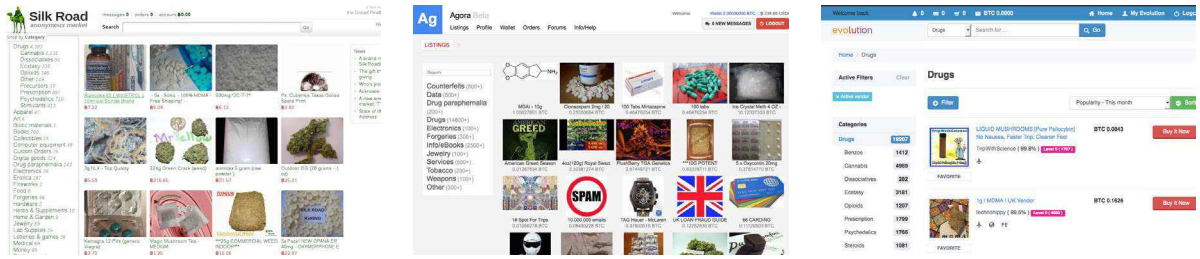
The common point between all these marketplaces is that they actually are not themselves selling contraband. Instead, they are risk management platforms for participants in (mostly illegal) transactions. Risk is mitigated on several levels. First, by abolishing physical interactions between transacting parties, these marketplaces claim to reduce (or indeed, eliminate) the potential for physical violence during the transaction.

Second, by providing superior anonymity guarantees compared to the alternatives, online anonymous marketplaces shield – to some degree[2] – transaction participants from law enforcement intervention.

Third, online anonymous marketplaces provide an escrow system to prevent financial risk. These systems are very similar in spirit to those developed by electronic

[1] Including, ironically, undercover law enforcement agents [7].

[2] Physical items still need to be delivered, which is a potential intervention point for law enforcement as shown in documented arrests [4].



(a) Silk Road (b) Agora (c) Evolution

Figure 1: **Example of marketplaces.** Most marketplaces use very similar interfaces, following the original Silk Road design.

commerce platforms such as eBay or the Amazon Marketplace. Suppose Alice wants to purchase an item from Bob. Instead of directly paying Bob, she pays the marketplace operator, Oscar. Oscar then instructs Bob that he has received the payment, and that the item should be shipped. After Alice confirms receipt of the item, Oscar releases the money held in escrow to Bob. This allows the marketplace to adjudicate any dispute that could arise if Bob claims the item has been shipped, but Alice claims not to have received it. Some marketplaces claim to support Bitcoin's recently standardized "multisig" feature which allows a transaction to be redeemed if, e.g., two out of three parties agree on its validity. For instance, Alice and Bob could agree the funds be transferred without Oscar's explicit blessing, which prevents the escrow funds from being lost if the marketplace is seized or Oscar is incapacitated.[3]

Fourth, and most importantly for our measurements, online anonymous marketplaces provide a feedback system to enforce quality control of the goods being sold. In marketplaces where feedback is mandatory, feedback is a good proxy to derive sales volumes [13]. We will adopt a similar technique to estimate sales volumes.

At the time of this writing the Darknet Stats service [1] lists 28 active marketplaces. As illustrated in Fig. 1 for the Evolution and Agora marketplaces, marketplaces tend to have very similar interfaces, often loosely based on the original Silk Road user interface. Product categories (on the right in each screen capture) are typically self-selected by vendors. We discovered that categories are sometimes incorrectly chosen, which led us to build our own tools to properly categorize items. Feedback data (not shown in the figure) comes in various flavors. Some marketplaces provide individual feedback per product and per transaction. This makes computation of sales volumes relatively easy as long as one can determine with good precision the time at which each piece of feedback was issued. Others provide feedback per vendor; if we can then link vendor feedback to specific items, we can again obtain a good estimate for sales volumes, but if not, we may not be able to derive any meaningful numbers. Last, in some marketplaces, feedback is either not mandatory, or only given as aggregates (e.g., "top 5% vendor"), which does not allow for detailed volume analysis.

[3]The Evolution marketplace claimed to support multisig. However, Evolution's operators absconded with escrow money on March 17th, 2015 [9]; it turns out that their multisig implementation did not function as intended, and was rarely used. Almost none of the stolen funds have been recovered so far.

3 Measurement methodology

Our measurement methodology consists of 1) crawling online anonymous marketplaces, and 2) parsing them. Table 1 lists all the anonymous marketplaces for which we have data. We scraped 35 different marketplaces a total of 1,908 times yielding a dataset of 3.2 TB in size. The total number of pages obtained from each scrape ranged from 27 to 331,691 pages and performing each scrape took anywhere from minutes up to five days.

The sheer size of the data corpus we are considering, as well as other challenging factors (e.g., hidden service latency and poor marketplace availability) led us to devise a custom web scraping framework built on top of Scrapy [3] and Tor [16], which we discuss first. We then highlight how we decide to parse (or ignore) marketplaces, before touching on validation techniques we use to ensure soundness of our analysis.

3.1 Scraping marketplaces

We designed and implemented the scraping framework with a few simple goals in mind. First, we want our scraping to be carried out in a *stealthy* manner. We do not want to alert a potential marketplace administrator to our presence lest our page requests be censored, by either modifying the content in an attempt to deceive us or denying the request altogether.

[4] The November 2011–July 2012 Silk Road data comes from a previously reported collection effort, with publicly available data [13].

Marketplace	Parsed?	Measurement dates	# snap.
Agora	Y	12/28/13–06/12/15	161
Atlantis‡	Y	02/07/13–09/21/13	52
Black Flag‡	Y	10/19/13–10/28/13	9
Black Market Reloaded†	Y	10/11/13–11/29/13	25
Tor Bazaar*	Y	07/02/14–10/15/14	27
Cloud 9*	Y	07/02/14–10/28/14	27
Deep Bay‡	Y	10/19/13–11/29/13	24
Evolution‡	Y	07/02/14–02/16/15	43
Flo Market‡	Y	12/02/13–01/05/14	23
Hydra*	Y	07/01/14–10/28/14	29
The Marketplace†	Y	07/08/14–11/08/14	90
Pandora‡	Y	12/01/13–10/28/14	140
Sheep Marketplace‡	Y	10/19/13–11/29/13	25
Silk Road*[4]	Y	11/22/11–07/24/12	133
	Y	06/18/13–08/18/13	31
Silk Road 2.0*	Y	11/24/13–10/26/14	195
Utopia*	Y	02/06/14–02/10/14	10
AlphaBay	N	03/18/15–06/02/15	17
Andromeda‡	N	07/01/14–11/10/14	30
Behind Blood Shot Eyes‡	N	01/31/14–08/27/14	56
BlackBank	N	07/02/14–05/16/15	56
Blue Sky*	N	12/25/13–06/10/14	126
Budster‡	N	12/01/13–03/11/14	56
Deep Shop‡	N	01/31/14–03/09/14	20
Deep Zone†	N	07/01/14–07/08/14	10
Dutchy‡	N	01/31/14–08/07/14	86
Area 51‡	N	11/20/14–01/20/15	14
Freebay†	N	12/31/13–03/11/14	36
Middle Earth	N	11/21/14–06/02/15	15
Nucleus	N	11/21/14–05/26/15	22
Outlaw	N	01/31/14–04/20/15	99
White Rabbit†	N	01/14/14–05/26/14	61
The Pirate Shop‡	N	01/14/14–09/17/14	102
The Majestic Garden	N	11/21/14–06/02/15	23
Tom Cat†	N	11/18/14–12/08/14	11
Tor Market	N	12/01/13–12/23/13	24

Table 1: **Markets crawled.** The table describes which markets were crawled, the time the measurements spanned, and the number of snapshots that were taken. * denote market sites seized by the police, † voluntary shutdowns, and ‡ (suspected) fraudulent closures (owners absconding with escrow money).

Second, we want the scrapes to be *complete*, *instantaneous*, and *frequent*. Scrapes that are instantaneous and complete convey a coherent picture about what is taking place on the marketplace without doubts about possible unobserved actions or the inconsistency that may be introduced by time delay. Scraping very often ensures that we have high precision in dating when actions occurred, and reduces the chances of missing vendor actions, such as listing and rapidly de-listing a given item.

Third we want our scraper to be *reliable* even when the marketplace that we are measuring is not. Even when a marketplace is unavailable for hours, the scraper should hold state and retry to avoid an incomplete capture.

Fourth, the scraper should be capable of handling *client-side state* normally kept by the users browser such as cookies, and be *robust* enough to avoid any detection schemes that might be devised to thwart the scraper. We attempt to address these design objectives as follows.

Avoiding censorship Before we add a site to the scraping regimen, we first manually inspect it and identify its layout. We build and use as input to the scraper a configuration including regular expressions on the URLs for that particular marketplace. This allows us to avoid following links that may cause undesirable actions to be performed such as adding items to a cart, sending messages or logging out. We also provide as input to the scraper a session cookie that we obtain by manually logging into the marketplace and solving a CAPTCHA; and parameters such as the maximum desired scraping rate.

In addition to being careful about what to request from a marketplace, we obfuscate how we request content. For each page request, the scraper randomly selects a Tor circuit out of 20 pre-built circuits. This strategy ensures that the requests are being distributed over several rendezvous points in the Tor network. This helps prevent triggering anti-DDoS heuristics certain marketplaces use.[5] This strategy also provides redundancy in the event that one of the circuits being used becomes unreliable and speeds up the time it takes to observe the entire site.

Completeness, soundness, and instantaneousness The goal of the data collection is to make an observation of the entire marketplace at an instantaneous point in time, which yields information such as item listings, pricing information, feedback, and user pages. Instantaneous observations are of course impossible, and can only be approximated by scraping the marketplace as quickly as possible. Scraping a site aggressively however limits the stealth of the scraper; We manually identified sites that prohibit aggressive scraping (e.g., Agora) and imposed appropriate rate limits.

Scrape completeness is also crucial. A partial scrape of a site may lead to underestimating the activities taking place. Fortunately, since marketplaces leverage feedback to build vendor reputation, old feedback is rarely deleted. This means that it is sufficient for an item listing and its feedback to be eventually observed in order to know that the transaction took place. Over time, the price of an item may fluctuate however, and information about when the transaction occurred often becomes less precise, so it is much more desirable to observe feedback as soon as possible after it is left. We generally attempted a scrape for each marketplace once every two to three days unless the marketplace was either unavailable or the previous scrape had not yet completed; having collected most of the data we were interested in by that time, we scraped considerably less often toward the end of our data collection interval (February through May 2015).

Many marketplaces that we observed have quite poor reliability, with 70% uptime or lower. It is very difficult

[5] However some marketplaces, e.g., Agora, use session cookies to bind requests coming from different circuits, and require additional attention.

to extract entire scrapes from marketplaces suffering frequent outages. This is particularly true for large sites, where a complete scrape can take several days. As a workaround, we designed the scraping infrastructure to keep state and retry pages using an increasing back-off interval for up to 24 hours. Using such a system allowed the scraper to function despite brief outages in marketplace availability. Retrying the site after 24 hours would be futile as in most cases, the session cookie would have expired and the scrape would require a manual login, and thus a manual restart.

Most marketplaces require the user to log in before they are able to view item listings and other sensitive information. Fortunately, creating an account on these marketplaces is free. However, one typically needs to solve a CAPTCHA when logging in; this was done manually. The process of performing a scrape begins with manually logging into the marketplace, extracting the session cookie, and using it as input to the scrape to continue scraping under that session. In many cases the site will fail to respond to requests properly unless multiple cookies are managed or unless the user agent of the scraper matches the user agent of the browser that generated the cookie. We managed to emulate typical browser behavior in all but one case (BlueSky). We were unable to collect meaningful data on BlueSky, as an anti-scraping measure on the server side was to annihilate any session after approximately 100 page requests, and get the user to log in again.

3.2 Parsing marketplaces

The raw page data collected by the scraper needs to be parsed to extract information useful for analysis. The parser first identifies which marketplace a particular page was scraped from; it then determines which type of page is being analyzed (item listing, user page, feedback page, or any combination of those).

Each page is then parsed using a set of heuristics we manually devised for each marketplace. We treat the information extracted as a single *observation* and record it into a database. Information that does not exist or cannot be parsed is assigned default values.

The heuristics for parsing can often become quite complicated as many marketplaces observed over long periods of time went through several iterations of page formats. This justified our conscious decision to decouple scraping from parsing so that we could minimize data loss. Because of the high manual effort associated with creating and debugging new parsers for marketplaces, we only generated parsers for marketplaces that we perceived to be of significance. While observing the scrapes of several marketplaces, it became apparent that their volume was either extremely small (<$1,000) or was not measurable by observing the website (e.g., because feedback is not mandatory). These marketplaces were omitted without greatly affecting the overall picture; their analysis is left for future work.

3.3 Internally validating data analysis

To ensure that the analysis we performed was not biased, and as a safety against egregious errors, both authors of this paper concurrently and independently developed multiple implementations of the analysis we present in the next section. During that stage of the work, the two authors relied on the same data sources, but used *different* analysis code and tools and did not communicate with each other until all results were produced.

We then internally confirmed that the independent estimations of total market volumes varied by less than 10% at any single point in time, and less than 5% on average, well within expected margin of errors for data indirectly estimated from potentially noisy sources (user feedback).[6] The independent reproducibility of the analysis is important since, as we will show, estimating market volumes presents many pitfalls, such as the risk of double-counting observations or using a holding price as the true value of an item.

3.4 Validating data completeness

The poor availability of certain marketplaces (e.g., Agora), combined with the large amount of time needed to fully scrape very large marketplaces raises concerns about data completeness. We attempt to estimate the amount of data that might be missing through a process known as marking and recapturing.

The basic idea is as follows. Consider that a given site scrape at time t contains a number M of feedback. Since we do not know whether the scrape is complete, we can only assert that M is a lower bound on the total number of feedback F actually present on the site at time t. Now, consider a second scrape (presumably taken after time t), which contains n pieces of feedback left at or before time t. The number n is another lower bound of F. We then estimate F as $\hat{F} = nM/m$, where m is the number of feedback captured in the first scrape that we *also* observe in the second scrape ($m \leq M$).

The Schnabel estimator [36] extends the above technique to estimate the size of a population to multiple samples, and is thus well-suited to our measurements. For n samples, if we denote by C_t the number of feedback in sample t, by M_t the total number of unique previously observed feedback in sample $(t-1)$, and by R_t the

[6] These minor discrepancies can be attributed to slightly different filtering heuristics, which we discuss later.

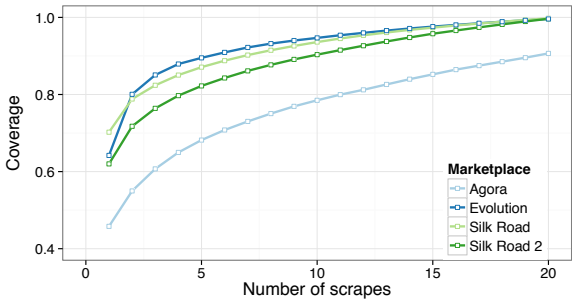


Figure 2: **Coverage of Agora, Silk Road 1, Silk Road 2, and Evolution.** This plot estimates the fraction of all feedback we obtain for a given time, as a function of the number of scrapes we collect.

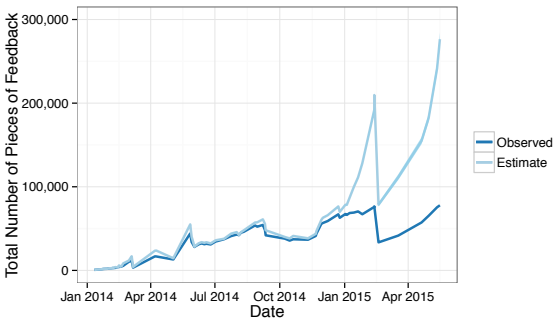


Figure 3: **Observed and estimated number of feedback present on Agora over time.** The lower and upper bounds for the estimate are nearly indistinguishable from the estimate itself.

number of previously observed feedback during sample t, we estimate the total number of feedback at time t as:

$$\hat{F} = \frac{\sum_{t=1}^{n} C_t M_t}{\sum_{t=1}^{n} R_t} .$$

The Schnabel estimator implicitly assumes that the distribution is time-invariant and that samples are drawn uniformly. To help ensure time invariance, the estimator begins with a sample at time t. Pieces of feedback with timestamps greater than t are omitted from all samples taken in the future ($t + \tau$). It is also important not to consider samples from too far into the future since items are occasionally de-listed and the corresponding feedback destroyed. To help minimize the impact of feedback deleted in the future, we only use samples within 60 days of t in our estimate.

We illustrate this estimate in Figure 2 for Agora, Silk Road 1, Silk Road 2, and Evolution after multiple observations have been made. Agora has relatively poor reliability and, on average, a single scrape will not manage to capture even half of the feedback present at that time on the site. On other marketplaces it is typical on the first visit to see as much as 60% of the entire population, or higher. After ten or more independent scrapes, we can expect to obtain a dataset that approaches 90% coverage or higher.

Figure 3 further illustrates our point, by comparing the number of pieces of feedback observed on Agora to its estimate. For most of the observed lifetime of Agora, the data that we have is very close to what we estimate the total to be. This is because information about a marketplace at a particular (past) point in time benefits from subsequent observations. Most recent observations do not have this benefit and therefore suffer from poor coverage, leading to significant divergence from their estimate. This results in potentially large underestimations towards the very end of our dataset, which will require us to censor some of this data when estimating volumes.

4 Analysis

We next turn to data analysis. We first estimate the overall evolution of the sales volumes in the entire ecosystem over the past couple of years. We then move to an assessment of the types of products being sold over time. Last, we discuss findings about vendor activity and techniques.

4.1 Sales volumes

The first important question that our analysis answers is how much product in terms of money is being bought and sold on online anonymous marketplaces. While we cannot directly measure the money being transacted from buyers to sellers, or packages being shipped from vendors to customers, we do make frequent observations of product feedback left for particular item listings on the marketplaces. Similar to prior work [13], we use these observations of feedback as a proxy to estimate a lower bound for sales.

Caveats In many marketplaces (e.g., Silk Road, Silk Road 2.0, Agora, Evolution among others) customers are required to leave feedback for a vendor whenever they receive their order of one of the vendor's items. An order for an item may be of varying quantity, so a customer that purchases a single quantity of a product, and a customer that purchases multiple quantities of a product will both leave a single feedback. In an effort to be conservative, we make the assumption that for every feedback observed, only a single quantity was purchased.

Our prudent strategy of estimating sales volume from confirmed observations of feedback diverges from other,

simpler approaches, such as counting the number of item listings offered (see, e.g., [15]). For instance, over the observed lifetime of Evolution, a few of the most successful item listings had feedback entries that indicated over 1 million dollars had been spent on each of them. The presence of these highly influential item listings suggests that simply counting the total number of listings on a site is a very poor indicator of sales volume. This claim is compounded by the observation that the average sales per item listing per day on Evolution in early July of 2014 was $85.14; but by September 2014, after new vendors and item listings had entered, the sales per item listing had declined to $19.42. Such volatile behavior is particularly common in marketplaces that are small or are going through periods of rapid growth.

Estimation We derived the estimates for the total amount of money transacted in three steps. We first took the set of all feedback observations that had been collected and removed any duplicates. For example, on two consecutive scrapes of a particular marketplace, the same item listing and its entire feedback history were observed and recorded twice. It would be incorrect to count two different observations of the same feedback twice. We thus developed a criterion for uniqueness for each marketplace—typically enforcing uniqueness of fields such as feedback message body, the vendor for which the feedback was left, the title of the item listing and the approximate date the feedback was left. Two pieces of feedback are considered different if and only if they differ in at least one of these categories.

The second step was to identify the the point in time at which the feedback was left. This time is an upper bound on when the transaction occurred. We obtained this estimate by noting the time of the observation and utilizing any information available about the age of the feedback. Different marketplaces have varying precision information about feedback timestamps. In the most precise instances, the time that the feedback was left is specified within the hour; in the most ambiguous cases, we can only infer the month in which feedback was deposited. Fortunately, due to our rather high sampling rate of the marketplaces, in most instances we have roughly a 24-hour accuracy on feedback time.

The third and final step is to identify the value of the transaction that each feedback represents. This involves pairing each feedback observation with a single observation of an item listing and its advertised price. Careful attention must be paid here as a few caveats exist, namely that the advertised price of an item listing varies with time, and that, in some rare cases, the corresponding item is never observed, leaving us unable to identify the value of the transaction.

Item prices change for two different reasons. The first and most common reason is that the vendors responsible for selling items are subject to standard free market pressures and may raise or lower their prices in response to competition, supply, demand, or other factors. The second reason is that when a vendor temporarily wishes to halt sales of an item with the expectation of selling it again in the future, instead of de-listing the item and losing all of the reviews and ratings that have accumulated over time, the vendor instead raises the price to something prohibitively high in order to discourage any sales. This is what we call a *holding price*. Holding prices are particularly dangerous for our analysis, because they can be in excess of millions of dollars. So, mistaking a holding price for an actual price just once could have dramatic consequences on the overall analysis.

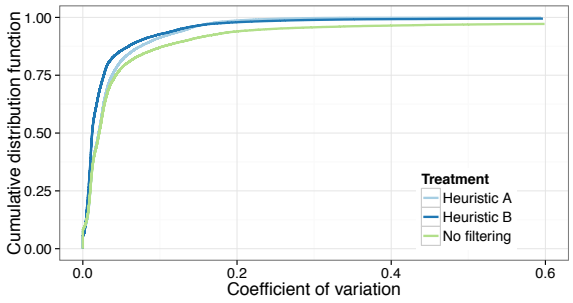


Figure 4: **C.d.f. of Coefficient of Variation for sets of observations of item listings** Both heuristics perform very similarly.

Dealing with holding prices Given a particular feedback and a set of observations of the corresponding product listing, the objective becomes to determine which observation yields the most accurate price for that feedback. Independent analysis (see Section 3.3) yielded two different heuristics for solving this problem. In the first heuristic (Heuristic A), we dismissed observations of the listing where the price was greater than $10,000 USD as well as observations that showed prices of zero (free). We then dismissed observations that were greater than 5 times the median of the remaining samples as well as observations that were less than 25% the value of the median. We manually observed thousands of product listings and identified that only in some very rare cases were the assumptions violated.

The second heuristic (Heuristic B) proceeded by removing observations with a price >$10,000 USD, as well as the upper quartile and any observations that were more than 100 times greater than the observation corresponding to the cheapest, non-zero price. To understand the effect that these heuristics had on observations, we calculated the coefficient of variation defined as $c_v = \sigma/\mu$ (standard deviation over mean) for the set

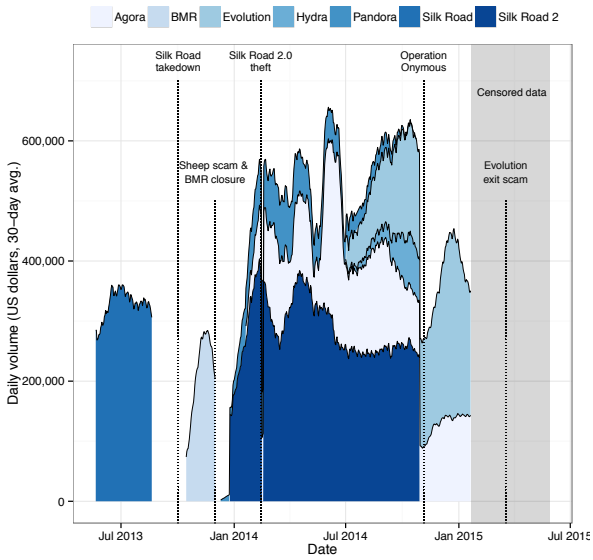


Figure 5: **Sales volumes in the entire ecosystem.** This stacked plot shows how sales volume vary over time for the marketplaces we study.

of observations for each item listing and plotted its cumulative distribution function.

Figure 4 shows that without any filtering, about 5% of all item listings were at some point sampled with highly variable prices, which suggests that a holding price was observed for this listing. Both heuristics produce relatively similar filtering; we ended up using Heuristic A in the rest of the analysis.

After applying the filter, there is still some smaller variation in the pricing of many listings which is consistent with the fluctuation in prices due to typical market pressures but it is clear that no listings with extremely high variations remain. 79,512 total unique item listings were identified, 1,003 (1.26%) of which had no valid observations remaining after filtering, meaning that the output of the heuristic was the empty set, the remaining 78,509 item listings returned at least one acceptable observation.

After filtering the listing observations, we pair each feedback with one of the remaining listing samples. To minimize the difference in estimated price of the feedback from the true price, we select the listing observation that is closest to the feedback in time. At this point we have a set of unique pieces of feedback, each mapped to a price at some point in time; from there, we can construct an estimate for the sales volumes.

Results We present our results in Figure 5 where we show the total volume, per marketplace we study, over time. The plot is stacked, which means that the top line corresponds to the total volume cleared by all marketplaces under study. In early 2013, we only have results for Silk Road, which at that point grossed around $300,000/day, far more than previously estimated for 2012 [13]. This number would project to over $100M in a year; combined by the previous $15M estimate [13] for early 2012, and "filling in" gaps for data we do not have in late 2012, appears consistent with the (revised) US Government calculations of $214M of total grossed income by Silk Road over its lifetime, based on Bitcoin transaction logs. These calculations were presented during the trial of the Silk Road founder (evidence GX940).

We then have a data collection gap, roughly corresponding to the time Silk Road was taken down. (We do not show volumes for Atlantis, which are negligible, in the order of $2,000–3,000/day.) Shortly after the Silk Road take-down we started measuring Black Market Reloaded, and realized that it has already made up for a vast portion of the volumes previously seen on Silk Road. We do not have sales data for Sheep Marketplace due to incomplete parses, but we do believe that the combination of both markets made up for the loss of Silk Road. Then, both Sheep and Black Market Reloaded closed – in the case of Sheep, apparently fraudulently. There was then quite a bit of turmoil with various markets starting and failing quickly. Only around late November 2013 did the ecosystem find a bit more stability, as Silk Road 2.0 had been launched and was rapidly growing. In parallel Pandora, Agora, and Evolution were also launched. By late January 2014, volumes far exceeded what was seen prior to the Silk Road take-down. At that point, though, a massive scam on Silk Road 2.0 caused dramatic loss of user confidence, which is evidenced by the rapid decrease until April 2014, before it starts recovering. Competitors however were not affected. (Agora does show spikes due to very imprecise feedback timing at a couple of points.) Eventually, in the Fall of 2014, the anonymous online marketplace ecosystem reached unprecedented highs. We started collecting data from Evolution in July, so it is possible that we miss quite a bit in the early part of 2014, but the overall take-away is unchanged. Finally, in November 2014, Operation Onymous [38] resulted in the take-down of Silk Road 2 and a number of less marketplaces. This did significantly affect total sales, but we immediately see a rebound by people going to Evolution and Agora. We censor the data we obtained from February 2015: at that point we only have results for Agora and Evolution, but coverage is poor, and as explained in Section 3, is likely to underestimate volumes significantly. We did note a short volume decrease prior to the Evolution "exit scam" of March 2015. We have not analyzed data for other smaller marketplaces (e.g., Black Bank, Middle Earth, or Nucleus) but suspect the volumes are much smaller. Fi-

nally, more recent marketplaces such as AlphaBay seem to have grown rapidly after the Evolution exit scam, but feedback on AlphaBay is not mandatory, and thus cannot be used to reliably estimate sales volumes.

In short, the entire ecosystem shows resilience to scams – Sheep, but also Pandora, which, as we can see started off very well before losing ground due to a loss in customer confidence, before shutting down. The effect of law enforcement take-downs (Silk Road 1&2, Operation Onymous) is mixed at best: the ecosystem relatively quickly recovered from the Silk Road shutdown, and appears to have withstood Operation Onymous quite well, since aggregate volumes were back within weeks to more than half what they were prior to Operation Onymous. We however caution that one would need longer term data to fully assess the impact of Operation Onymous.

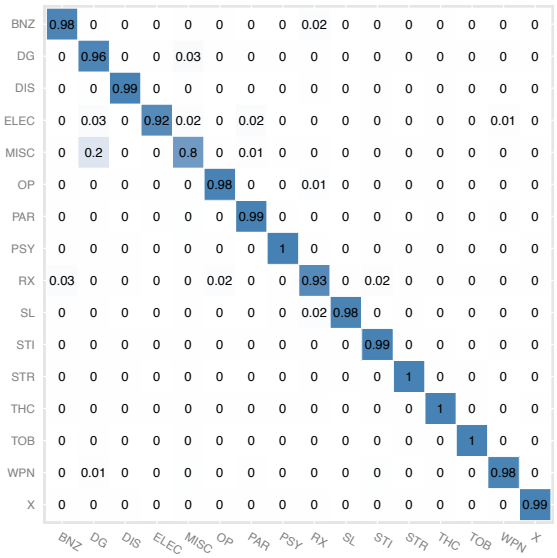


Figure 6: **Classifier confusion matrix**. BNZ: Benzos, DG: Digital Goods, DIS: Dissociatives, ELEC: Electronics, MISC: Miscellaneous, OP: Opioids, PAR: Drug Paraphernalia, PSY: Psychedelics, RX: Prescription drugs, SL: Sildenafil, STI: Stimulants, STR: Steroids, THC: Cannabis, TOB: Tobacco, WPN: Weapons, X: Ecstasy.

4.2 Product categories

In addition to estimating the value of the products that are being sold, we strived to develop an understanding of what is being sold. Several marketplaces such as Agora and Evolution include information on item listing pages that describe the nature of the listing as provided by the vendor that posted it. Unfortunately these descriptions are often too specific, conflict across marketplaces, and in the case of some sites, are not even available at all.

For our analysis, we need a consistent and coherent labeling for all items, so that we could categorize them into broad mutually exclusive categories. We thus implemented a machine learning classifier that was trained and tested on samples from Agora and Evolution, where ground truth was available via labeling. We then took this classifier and applied it to item listings on all marketplaces to answer the question of what is being sold.

We took 1,941,538 unique samples from Evolution and Agora, where a sample is the concatenation of an item listing's title and all descriptive information about it that was parsed from the page. We tokenized each sample under the assumption that the sample is written in English, resulting in a total of 162,198 unique words observed. We then computed a *tf-idf* value for each of the 162,198 words in the support for each sample, and used these values as inputs to an L2-Penalized SVM under L2-Loss implemented using Python and *scikit-learn*.

We evaluated our classifier using 10-fold cross validation. The overall precision and recall were both (roughly) 0.98. We also evaluated the classifier on Agora data when trained with samples from Evolution and vice-versa to ensure that the classifier was not biased to only perform well on the distributions it was trained on. The confusion matrix in Figure 6 shows that classification performance is very strong for all categories. Only "Misc" is occasionally confused with Digital Goods and Prescriptions are occasionally confused with Benzos (which in fact is not necessarily surprising). We believe that these errors are most likely caused by mislabeled test samples. Although we drew our samples from Evolution and Agora which provide a specific label for each listing, the label is selected by the vendor and may be erroneous, particularly for listings that are hard to place. Manual inspection revealed that several of the errors came from item listings that offered US $100 Bills in exchange for Bitcoin.

We then applied the classifier to the aggregate analysis performed earlier. In addition to placing a particular feedback in time, and pairing it with an item listing observation to derive the price, we predicted the class label of that listing and aggregated the price by class label. Figure 7 shows the normalized market aggregate by category. Drug paraphernalia, weapons, electronics, tobacco, sildenafil, and steroids were collapsed into a category called 'Other' for clarity.

Over time the fraction of market share that belongs to each category is relatively stable. However, around October of 2013, December 2013, March 2014, and January 2015, cannabis spikes up to as much as half of the market share. These spikes correspond to the earlier mentioned 1) take-down of Silk Road, 2) closure of Black Market

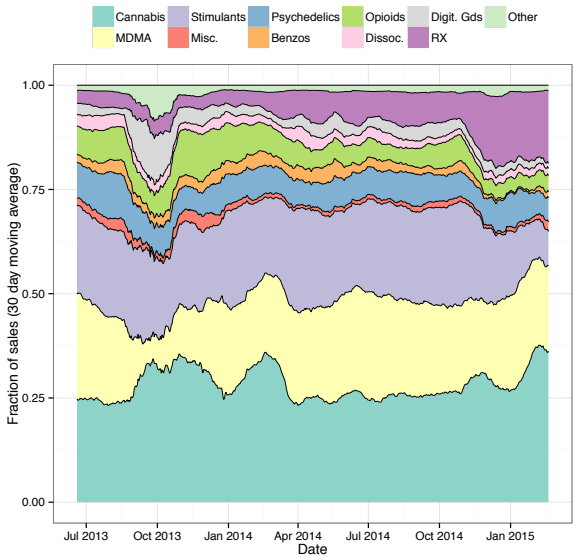


Figure 7: **Fractions of sales per item category.**

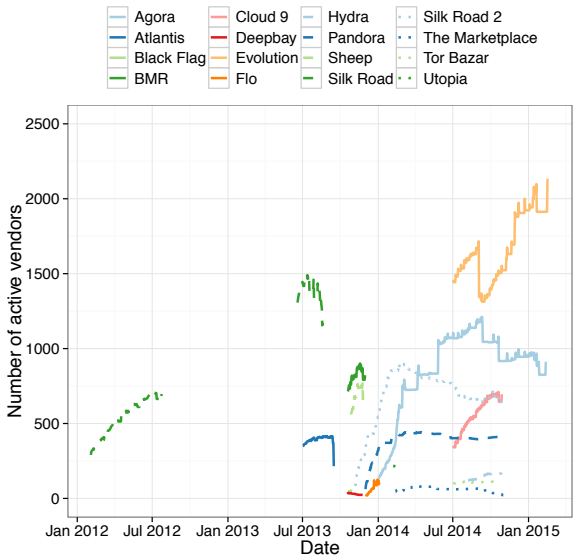


Figure 8: **Evolution of the number of active sellers over time.** Each "seller" here corresponds to a unique marketplace-vendor name pair. Certain sellers participate in several marketplaces and are thus counted multiple times here.

Reloaded and Sheep scam, 3) Silk Road 2.0 theft [5], and 4) Operation Onymous respectively. These are all events that generated substantial doubts in both vendors and consumers regarding the safety and security of operating on these marketplaces. At these times the perceived risk of operation was higher, which may have exerted pressure towards buying and selling cannabis as opposed to other products for which the punishment if caught is much more severe. We can also see that digital goods take an unusually high market share in times of uncertainty, which is most obvious around October 2013: this is not surprising as digital goods are often a good way to quickly accumulate large numbers of listings on a new marketplace.

Figure 7 shows that after an event such as a take-down or large scale scam occurs, it takes about 2–3 months before consumer and vendor confidence is restored and the markets converge back to equilibrium. At equilibrium, cannabis and MDMA (ecstasy) are about 25% of market demand each with stimulants closely behind at about 20%. Psychedelics, opioids, and prescription drugs are a little less than 10% of market demand each, although starting in November 2014, prescription drugs have gained significant traction—perhaps making anonymous marketplaces a viable alternative to unlicensed online pharmacies.

4.3 Vendors

Online anonymous marketplaces are only successful when they manage to attract a large enough vendor population to provide a critical mass of offerings. At the same time, vendors are not bound to a specific marketplace. Anecdotal evidence shows that certain sellers list products on several marketplaces at once; likewise, certain sellers "move" from marketplace to marketplace in response to law enforcement take-down or other marketplace failures. Here, we try to provide a good picture of the vendor dynamics across the entire ecosystem.

Number of sellers Figure 8 shows, over time, the evolution of the number of active sellers on all the marketplaces we considered. For each marketplace, a seller is defined as active at time T is we observed her having at least one active listing at time $t \leq T$, and at least one active listing (potentially the same) at a time $t \geq T$. This is a slightly different definition from that used in Christin [13] which required an active listing at time t to count a seller as active. For us, active sellers include sellers that may be on vacation but will come back, whereas Christin did not include such sellers. As a result, our results for Silk Road are very slightly higher than his.

The main takeaway from Figure 8 is that the number of sellers overall has considerably increased since the days of Silk Road. By the time Silk Road stopped activities in 2013, it featured around 1,400 sellers; its leading competitors, Atlantis and Black Market Reloaded (BMR) were much smaller. After the Silk Road take-down (October 2013) and Atlantis closure, we observe that both BMR and the Sheep marketplace rapidly pick up a large

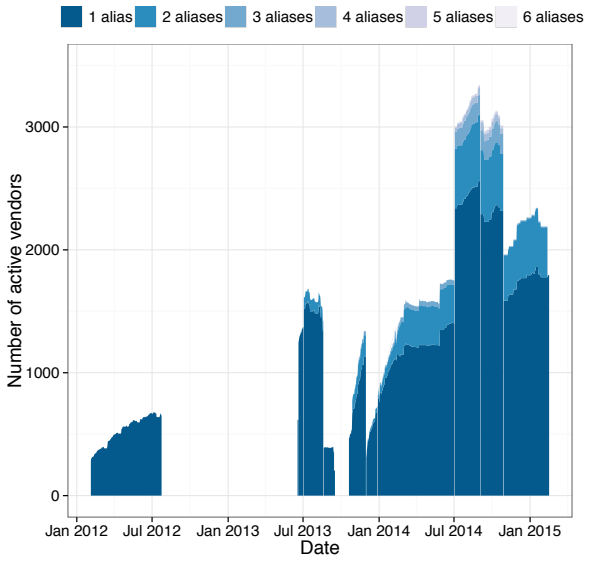


Figure 9: **Number of aliases per seller.** This plot shows the evolution of the number of aliases per seller across all marketplaces, over time. The contour of the curve denotes the total number of sellers overall.

influx of sellers. In parallel, Silk Road 2.0 also grows at a very rapid pace. Successful newcomers like Pandora, Agora, and Evolution also see quick rises in the number of sellers. After a certain amount of time, however, per-marketplace population tends to stabilize, even in the most popular marketplaces. On the other hand, we also observe that some marketplaces never took off: The Marketplace, Hydra, Deepbay, and Tor Bazaar, for instance, consistently have a small number of vendors. In other words, we see very strong network effects: Either marketplaces manage to get initial traction and then rapidly flourish, or they never manage to take off.

Sellers and aliases After Silk Road was taken down, a number of sellers reportedly moved to Black Market Reloaded or the Sheep Marketplace. More generally, nothing prevents a vendor from opening shop on multiple marketplaces; in fact, it is probably a desirable strategy to hedge against marketplace take-downs or failures. As a result, a given seller, Sally, may have multiple vendor accounts on several marketplaces: Sally may sell on Silk Road 2 as "Sally," on Agora as "sally" and on Evolution as "Easy Sally;" she may even have a second Evolution account ("The Real Easy Sally").

We formally define an *alias* as a unique (*vendor nickname*, *marketplace*) pair, and link different aliases to the same vendor using the combination of the following three heuristics. We first consider vendor nicknames on different marketplaces with only case differences as belonging to the same person (e.g., "Sally" and "sally"). We then use the InfoDesk feature of the Grams "DarkNet Markets" search engine [2] to further link various vendor nicknames.[7] We filter out vendor nicknames consisting only of a common substring (e.g., "weed," "dealer," "Amsterdam," ...) used by many vendors prior to conducting the search. Finally, we link all vendor accounts that claim to be using the same PGP key. Clearly, our linking strategy is very conservative – in the sense that minor variations like "Sally" and "Sally!" will not be linked absent a common PGP key.

Using this set of heuristics, from a total of 29,258 unique aliases observed across our entire measurement interval, we obtain a list of 9,386 sellers. In Figure 9, we show, over time, the number of vendors that have one, two or up to six aliases active at any given time T (where we use the same definition of "active" as earlier, i.e., the alias has at least one listing available before and after T). The plot is by definition incomplete since we can only take into account, for each time t, the marketplaces that we have crawled (and parsed) at time t.

For instance, the earlier part of the data show a complete monopoly: this is not surprising since we only have data for Silk Road at that time, even though Black Market Reloaded was also active at the same time. We observe in the summer of 2013 that a few vendors sell simultaneously on Silk Road and Atlantis, but the practice of having multiple vendor accounts on several sites seems to only really take hold in 2014, after many marketplaces failed in the Fall of 2013 (including Silk Road, and many of its short-lived successors). The second jump in July 2014 corresponds to our starting to collect data for the very large Evolution marketplace. Finally, the decrease observed in late 2014 is due to Operation Onymous [38], which – besides Silk Road 2.0 – took down a relatively large number of secondary marketplaces, such as Cloud 9.

Besides the relatively robust rise is the number of sellers to take-downs and scams, the main takeaway from this plot is that the majority of sellers appear to only use one alias – but this may be a bit misleading, as (as we will see later) a large number of vendors sell extremely limited quantities of products. An interesting extension would be to check whether "top" vendors diversify across marketplaces or not.

We complement this analysis by looking into the "survivability" functions of aliases and sellers, which we report in Figure 10. Here the survival function is defined as the probability $p(\tau)$ that a given seller (resp. alias) observed at time t be still active at time $t+\tau$. The figure shows the survival function, derived from a Kaplan-Meier estimator [24] to account for the fact that we have

[7] It is not clear how the Grams search engine is implemented; we suspect the vendor directory is primarily based on manual curation.

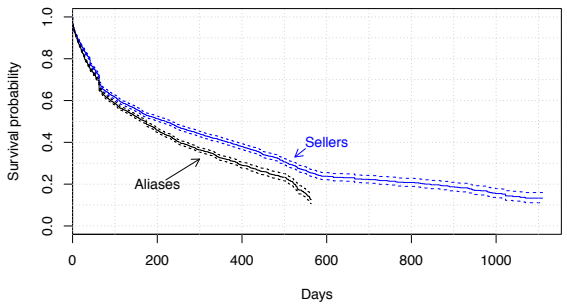


Figure 10: **Seller survivability analysis.** The plot describes the probability a given alias is still active after a certain number of days; and the probability a given seller (regardless of which alias it is using) is still active after a certain number of days. On average, sellers are active for 220 days, while aliases remain active for 172 days.

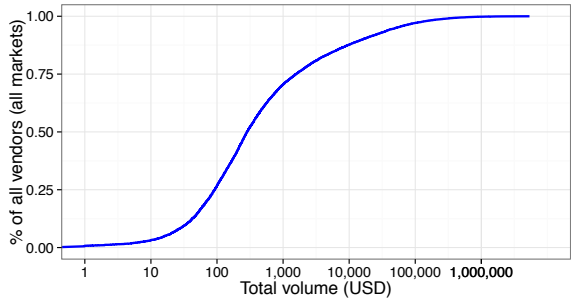


Figure 11: **Seller volumes.** A very small fraction of sellers generate significant profit. On average, a typical seller only makes a couple of hundreds dollars.

finite measurement intervals, along with 95% confidence intervals. The key findings here are that half of the sellers are only present for 220 days or less; half of the aliases only exist for 172 days or less. More interesting is the "long-tail" phenomenon we observe: a number (more than 10%) of sellers have been active *throughout the entire measurement interval*. More generally approximately 25% of all sellers are "in it for the long run," and remain active (with various aliases on various marketplaces) for years.

Volumes per vendor In an effort to obtain a more clear understanding of how vendors operate, we aggregated unique feedback left for products by vendor. We used this to calculate the total value of the transactions for items sold by each vendor and then grouped these vendor aliases to yield the total value of transactions for each seller. Figure 11 plots the CDF of sellers by the total value of their transactions. About 70% of all sellers never managed to sell more than $1,000 worth of products. Another 18% of sellers were observed to sell between $1,000 and $10,000 but only about 2% of vendors managed to sell more than $100,000. In fact, 35 sellers were observed selling over $1,000,000 worth of product and the top 1% most successful vendors were responsible for 51.5% of all the volume transacted. Some of these sellers, like "SuperTrips" (or to a lesser extent, "Nod") from Silk Road, have been arrested, and numbers released in connection with these arrests are consistent with our findings [4, 6].

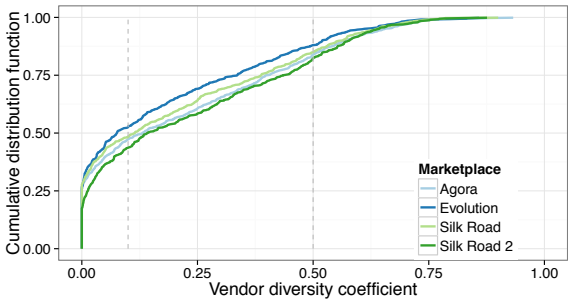


Figure 12: **Vendor diversity**

There is a clear discrepancy between sellers that experiment in the marketplaces and those who manage to leverage it to operate a successful business. Going forward, we define any seller that we have observed selling in excess of $10,000 to be successful. This allows us to draw conclusions only about vendors that have had a meaningful impact on the marketplace ecosystem. Now that we know how much sellers are selling, we wish to understand *what* they are selling. Once again we group feedback by vendor but this time we also use the classifier to categorize the items that were being sold and aggregate by category. Let $\mathscr{C}$ be the set of normalized item categories for each seller and $\mathscr{S}$ be the set of all sellers across all marketplaces. So, $|\mathscr{C}| = 16$, and $|\mathscr{S}| = 9,386$. Define $\mathscr{C}_i(s_j)$ as the normalized value of the i-th category for seller j such that $\forall s_j \in S$, $\sum_{i=1}^{|\mathscr{C}|} \mathscr{C}_i(s_j) = 1$. Then, we define the coefficient of diversity for a seller s_j as:

$$c_d = \left(1 - \max_i \left(\mathscr{C}_i(s_j)\right)\right) \frac{|\mathscr{C}|}{|\mathscr{C}| - 1} .$$

Intuitively, the coefficient of diversity is measuring how invested a seller is into their most popular category, normalized so that $c_d \in [0,1]$. When evaluating the categories that different sellers are invested in, it only makes sense to consider successful sellers as less significant sellers are volatile and greatly influenced by an individual sale in some category.

Figure 12 plots the CDF of the coefficient of diversity for sellers from Evolution, Silk Road, Silk Road 2 and

Agora that sold more than $10,000 total. From Figure 12 we argue that there are roughly three types of sellers. The first type of seller with a coefficient of diversity between 0 and 0.1 is highly specialized, and sells exactly one type of product. About half of all sellers are highly specialized and indicates that the seller has access to a steady long-term supply of some type of product. About one third of all vendors who specialize sell cannabis, another third sell digital goods, and the last third sell in the various other categories. While digital goods is a relatively small share of the total marketplace ecosystem, it tends to attract vendors that specialize. This is likely due to the domain expertise required for actions such as manufacturing fake IDs or stealing credit cards. The second type of seller has a diversity coefficient of between 0.1 and 0.5 and generally specializes in two or three types of products. The most common two categories to simultaneously specialize in are ecstasy and psychedelics – i.e., primarily recreational and club drugs. The third type of vendor has a diversity coefficient greater than 0.5 and has no specialty but rather sells a variety of items. These types of sellers may be networks of users with access to many different sources, or may be involved in arbitrage between markets.

PGP deployment We conclude our discussion of vendor behavior by looking in more detail at their security practices. While we cannot easily assess their overall operational security, we consider a very simple proxy for security behavior: the availability of a valid PGP key. From our data set, we extracted 7,717 PGP keys. Most vendors use keys of appropriate length, even though we did observe a couple of oddities (e.g., a 2,047-bit key!) that might indicate an incorrect use of the software. Inspired by Heninger et al. [20] and Lenstra et al. [25] we checked all pairs of keys to determine whether or not they had common primes. We did not find any, which either suggests that GPG software was always properly used and with a good random number generator, or, more likely, that our dataset is too small to contain evidence of weak keys.

We then plot in Figure 13 the fraction of vendors, over time, that have (at least) one usable PGP key. We take an extremely inclusive view of PGP deployment here: as long as a vendor has advertised a valid PGP key for one or her active aliases, we consider they are using PGP. As vendors deal with highly sensitive information such as postal delivery addresses of their customers, we would expect close to 100% deployment. We see that, despite improvements, this is not the case. In the original Silk Road, only approximately 2/3 to 3/4 of vendors had a valid PGP key listed. During the upheaval of the 2013 Fall, with many marketplaces opening and shutting down quickly, we see that PGP deployment is very low. When the situation stabilizes in January 2014, we observe an

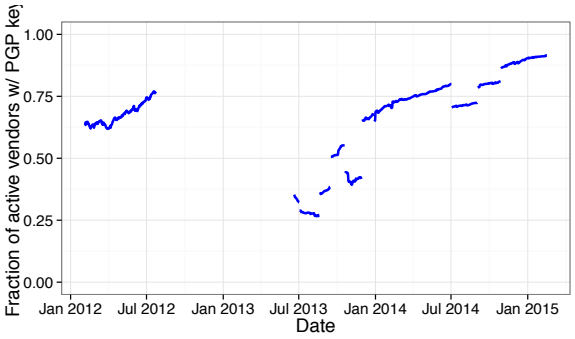


Figure 13: **PGP deployment over time.**

increase in PGP adoption; interestingly, *after* Operation Onymous, adoption seems even higher, which can be construed as an evolutionary argument: marketplaces that support and encourage PGP use by their sellers (such as Evolution and Agora) might have been also more secure in other respects, and more resilient against takedowns. Shortly before the Evolution shutdown, PGP deployment on Agora and Evolution was close to 90%.

5 Discussion

A study of this kind brings up a number of important discussion points. We focus here on what we consider are the most salient ones: validation, ethics, and potential public policy take-aways.

5.1 Validation

Scientific measurements should be amenable to validation. Unfortunately, here, ground truth is rarely available, which in turn makes validation extremely difficult. Marketplace operators indeed generally do not publish metrics such as seller numbers or traffic volumes. However, in certain cases, we have limited information that we can use for spot-checking estimates.

Ross Ulbricht trial evidence (Silk Road) In October 2013, a San Francisco man by the name of Ross Ulbricht was arrested and charged as being the operator of Silk Road [8]. A large amount of data was subsequently entered into evidence used during his trial, which took place in January 2015. In particular, evidence contained relatively detailed accounting entries found on Mr. Ulbricht's laptop, and claimed to pertain to Silk Road. Chat transcripts (evidence GX226A, GX227C) place weekly volumes at $475,000/week in late March 2012 for instance: this is consistent with the data previously reported [13] and which we use for documenting Silk Road. Evidence GX250 contains a personal ledger

which apparently faithfully documents Silk Road sales commissions. Projecting the data listed during the time of the previous study [13] ($680,279) over a year yields a yearly projection of about $1.2M; Christin's estimates were of $1.1M [13]. This hints that the technique of using feedback as a sales proxy, which we reuse here, produces reliable estimates.

Blake Benthall criminal complaint (Silk Road 2) In November 2014, another San Francisco man by the name of Blake Benthall was arrested and charged with being "Defcon," the Silk Road 2.0 administrator. The criminal complaint against Mr. Benthall [7] reports that in September 2014, the administrator, talking to an undercover agent actually working on Silk Road 2's staff, reports around $6M of monthly sales; and later amends this number to $8M. This corresponds to a daily sales volume of $200,000–$250,000 which is very close to what we report in Figure 5 for Silk Road 2 at that given time.

Leaked Agora seller page In December 2014, it was revealed that an Agora vendor page had been scraped and leaked on Pastebin [21]. This vendor page in particular contains a subset of all the vendor's transactions; one can estimate precisely the amount for that specific vendor on June 5, 2014 to $3,460. Checking in our database, our instantaneous estimate credits that seller with $3,408 on the day – which, considering Bitcoin exchange fluctuations is pretty much identical to the ground truth.

5.2 Ethics of data collection

We share much of the ethical concerns and views documented in previous work [13]. Our data collection, in particular, is massive, and could potentially put some strain on the Tor network, not to mention marketplace servers themselves. However, even though it is hard to assess we believe that our measurements represent a small fraction of all traffic that is going to online anonymous marketplaces. As discussed in Section 3 we are attempting to balance accuracy of the data collection with a light-weight enough crawling strategy to avoid detection – or worse, impacting the very operations we are trying to measure. In addition, we are contributing Tor relays with long uptimes on very fast networks to "compensate" for our own massive use of the network. Our work takes a number of steps to remain neutral. We certainly do not want to facilitate vendor or marketplace operator arrests. This is not just an ethical question, but is also a scientific one: our measurements, to be sound, should not impact the subject(s) being measured [23].

5.3 Public-policy take-aways

The main outcome of this work, we hope, is a critical evaluation of meaningful public policy toward online anonymous marketplaces. While members of Congress have routinely called for the take down of "brazen" online marketplaces, it is unclear that this is the most pragmatic use of taxpayer money.

In fact, our measurements suggest that the ecosystem appears quite resilient to law enforcement take-downs. We see this without ambiguity in response to the (original) Silk Road take-down; and while it is too early to tell the long-lasting impacts of Operation Onymous, its main effect so far seems to have been to consolidate transactions in the two dominant marketplaces at the time of the take-down. More generally, economics tell us that because user demand for drugs online is present (and quite massive), enterprising individuals will seemingly always be interested in accommodating this demand.

A natural question is whether the cat-and-mouse game between law enforcement and marketplace operators could end with the complete demise of online anonymous marketplaces. Our results suggest it is unlikely. Thus, considering the expenses incurred in very lengthy investigations and the level of international coordination needed in operations like Operation Onymous, the time may be ripe to investigate alternative solutions.

Reducing demand through prevention is certainly an alternative worth exploring on a global public policy level, but, from a law enforcement perspective, even active intervention could be much more targeted, e.g., toward seizing highly dangerous products while in transit. A number of documented successes in using traditional police work against sellers of hazardous substances (e.g., [35]) or large-scale dealers (e.g., [4, 6] among many others) show that law enforcement is not powerless to address the issue in the physical world.

6 Related work

The past decade has seen a large number of detailed research efforts aiming at gathering actual measurements from various online criminal ecosystems in order to devise meaningful defenses; see, e.g., [13, 14, 22, 26, 27, 28, 29, 32, 40, 41]. Anderson et al. [11] and Thomas et al. [37] provide a very good overview of the field. Closest among these papers to our work, McCoy et al. obtained detailed measurements of online pharmaceutical affiliates, showing that individual networks grossed between USD 12.8 million/year to USD 67.7 million/year. In comparison, the long-term rough average we see here is in the order of $150–180M/year for the entire online anonymous marketplace ecosystem. In other words, online marketplaces have seemingly surpassed more "traditional" ways of de-

livering illicit narcotics.

With respect to specific measurements of online anonymous marketplaces, the present paper builds up on our previous work [13]. Surprisingly few other efforts exist attempting to quantitatively characterize the economics of online anonymous marketplaces. Of note, Aldridge and Décary-Hétu [10] complement our original volume estimates by showing revised numbers of around $90M/year for Silk Road in 2013 right before its take-down. This is roughly in line with our own measurements, albeit slightly more conservative (Figure 5 shows about $300K/day for Silk Road in summer 2013.) More recent work by Dolliver [17] tries to assess the volumes on Silk Road 2.0. While she does not report volumes, her seller numbers are far smaller than ours, and we suspect her scrapes might have been incomplete. Looking at the problem from a different angle, Meiklejohn et al. [31] provide a detailed analysis of transaction traceability in the Bitcoin network, and show which addresses are related to Silk Road, which in turn could be a useful way of assessing the total volumes of that marketplace. A follow up paper [30] shows that purported Bitcoin "anonymity" (i.e., unlinkability) is greatly overstated, even when using newer mixing primitives.

On the customer side, Barratt et al. [12] provide an insightful survey of Silk Road patrons, showing that a lot of them associate with the "party culture," which is corroborated by our results showing that cannabis and ecstasy correspond to roughly half of the sales; likewise Van Hout and Bingham provide valuable insights into individual participants [39]. Our research complements these efforts by providing a macro-level view of the ecosystem.

7 Conclusions

Even though anonymous online marketplaces are a relatively recent development in the overall online crime ecosystem, our longitudinal measurements show that in the short four years since the development of the original Silk Road, total volumes have reached up to $650,000 daily (averaged over 30-day windows) and are generally stable around $300,000-$500,000 a day, far exceeding what had been previously reported. More remarkably, anonymous marketplaces are extremely resilient to take-downs and scams – highlighting the simple fact that economics (demand) plays a dominant role. In light of our findings, we suggest a re-evaluation of intervention policies against anonymous marketplaces. Given the high demand for the products being sold, it is not clear that take-downs will be effective; at least we have found no evidence they were. Even if one went to the impractical extreme of banning anonymous networks, demand would probably simply move to other channels, while some of the benefits associated with these markets (e.g., reduction in risks of violence at the retail level) would be lost. Instead, a focus on reducing consumer demand, e.g., through prevention, might be worth considering; likewise, it would be well-worth investigating whether more targeted interventions (e.g., at the seller level) have had measurable effects on the overall ecosystem. While our paper does not answer these questions, we believe that the data collection methodology we described, as well as some of the data we have collected, may enable further research in the field.

Acknowledgments

This research was partially supported by the National Science Foundation under ITR award CCF-0424422 (TRUST) and SaTC award CNS-1223762; and by the Department of Homeland Security Science and Technology Directorate, Cyber Security Division (DHS S&T/CSD), the Government of Australia and SPAWAR Systems Center Pacific via contract number N66001-13-C-0131. This paper represents the position of the authors and not that of the aforementioned agencies. We thank our anonymous reviewers and our shepherd, Damon McCoy, for feedback that greatly improved the manuscript.

References

[1] Darknet stats. https://dnstats.net/.

[2] Grams: Search the darknet. http://grams7enufi7jmdl.onion.

[3] Scrapy: An open source web scraping framework for Python. http://scrapy.org.

[4] United States of America vs. Steven Lloyd Sadler and Jenna M. White, Nov. 2013. United States District Court, Western District of Washington at Seattle. Criminal Complaint MJ13-487.

[5] Silk Road 2.0 'hack' blamed on Bitcoin bug, all funds stolen, Feb. 2014. http://www.forbes.com/sites/andygreenberg/2014/02/13/silk-road-2-0-hacked-using-bitcoin-bug-all-its-funds-stolen/.

[6] Silk Road online drug dealer pleads guilty to trafficking, May 2014. http://www.cbsnews.com/news/silk-road-online-drug-dealer-pleads- guilty-to-trafficking/.

[7] United States of America vs. Blake Benthall, Oct. 2014. United States District Court, Southern District of New York. Sealed Complaint 14MAG2427.

[8] United States of America vs. Ross William Ulbricht, Feb. 2014. United States District Court, Southern District of New York. Indictment 14CRIM068.

[9] Bitcoin "exit scam": deep-web market operators disappear with $12m, Mar. 2015. http://www.theguardian.com/technology/2015/mar/18/bitcoin-deep-web-evolution-exit-scam-12-million-dollars/.

[10] Aldridge, J., and Décary-Hétu, D. Not an "Ebay for drugs": The cryptomarket "Silk Road" as a paradigm shifting criminal innovation. *Available at SSRN 2436643* (2014).

[11] ANDERSON, R., BARTON, C., BÖHME, R., CLAYTON, R., VAN EETEN, M. J., LEVI, M., MOORE, T., AND SAVAGE, S. Measuring the cost of cybercrime. In *The economics of information security and privacy*. Springer, 2013, pp. 265–300.

[12] BARRATT, M. J., FERRIS, J. A., AND WINSTOCK, A. R. Use of silk road, the online drug marketplace, in the united kingdom, australia and the united states. *Addiction 109*, 5 (2014), 774–783.

[13] CHRISTIN, N. Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace. In *Proceedings of the 22nd World Wide Web Conference (WWW'13)* (Rio de Janeiro, Brazil, May 2013), pp. 213–224.

[14] CHRISTIN, N., YANAGIHARA, S., AND KAMATAKI, K. Dissecting one click frauds. In *Proc. ACM CCS'10* (Chicago, IL, Oct. 2010).

[15] DIGITAL CITIZENS ALLIANCE. Busted, but not broken: The state of Silk Road and the darknet marketplaces, Apr. 2014.

[16] DINGLEDINE, R., MATHEWSON, N., AND SYVERSON, P. Tor: The second-generation onion router. In *Proceedings of the 13th USENIX Security Symposium* (San Diego, CA, Aug. 2004).

[17] DOLLIVER, D. Evaluating drug trafficking on the Tor network: Silk Road 2, the sequel. *International Journal of Drug Policy* (2015).

[18] GREENBERG, A. An interview with a digital drug lord: The Silk Road's Dread Pirate Roberts (Q&A), Aug. 2013. `http://www.forbes.com/sites/andygreenberg/2013/08/14/an-interview-with-a-digital-drug-lord-the-silk-roads-dread-pirate-roberts-qa/`.

[19] GREENBERG, A. Five men arrested in dutch crackdown on Silk Road copycat, Feb. 2014. `http://www.forbes.com/sites/andygreenberg/2014/02/12/five-men-arrested-in-dutch-crackdown-on-silk-road-copycat/`.

[20] HENINGER, N., DURUMERIC, Z., WUSTROW, E., AND HALDERMAN, J. A. Mining your Ps and Qs: Detection of widespread weak keys in network devices. In *Proceedings of the 21st USENIX Security Symposium* (Bellevue, WA, Aug. 2012).

[21] IMPOST_R. Boosie5150 questionable security practices - Agora account compromised in june. `https://www.reddit.com/r/DarkNetMarkets/comments/2oisq0/boosie5150_questionable_security_practices_agora/`.

[22] JOHN, J., YU, F., XIE, Y., ABADI, M., AND KRISHNAMURTHY, A. deSEO: Combating search-result poisoning. In *Proceedings of USENIX Security 2011* (San Francisco, CA, Aug. 2011).

[23] KANICH, C., LEVCHENKO, K., ENRIGHT, B., VOELKER, G., AND SAVAGE, S. The Heisenbot uncertainty problem: challenges in separating bots from chaff. In *Proceedings of USENIX LEET'08* (San Francisco, CA, Apr. 2008).

[24] KAPLAN, E., AND MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53* (1958), 457–481.

[25] LENSTRA, A., HUGHES, J. P., AUGIER, M., BOS, J. W., KLEINJUNG, T., AND WACHTER, C. Ron was wrong, Whit is right. Tech. rep., IACR, 2012.

[26] LEVCHENKO, K., CHACHRA, N., ENRIGHT, B., FELEGYHAZI, M., GRIER, C., HALVORSON, T., KANICH, C., KREIBICH, C., LIU, H., MCCOY, D., PITSILLIDIS, A., WEAVER, N., PAXSON, V., VOELKER, G., AND SAVAGE, S. Click trajectories: End-to-end analysis of the spam value chain. In *Proceedings of IEEE Security and Privacy* (Oakland, CA, May 2011).

[27] LI, Z., ALRWAIS, S., WANG, X., AND ALOWAISHEQ, E. Hunting the red fox online: Understanding and detection of mass redirect-script injections. In *Proceedings of the 2014 IEEE Symposium on Security and Privacy (Oakland'14)* (San Jose, CA, May 2014).

[28] LU, L., PERDISCI, R., AND LEE, W. SURF: Detecting and measuring search poisoning. In *Proceedings of ACM CCS 2011* (Chicago, IL, Oct. 2011).

[29] MCCOY, D., PITSILLIDIS, A., JORDAN, G., WEAVER, N., KREIBICH, C., KREBS, B., VOELKER, G., SAVAGE, S., AND LEVCHENKO, K. Pharmaleaks: Understanding the business of online pharmaceutical affiliate programs. In *Proceedings of USENIX Security 2012* (Bellevue, WA, Aug. 2012).

[30] MEIKLEJOHN, S., AND ORLANDI, C. Privacy-enhancing overlays in bitcoin. In *Proceedings of the 2015 BITCOIN research workshop* (Puerto Rico, Jan. 2015).

[31] MEIKLEJOHN, S., POMAROLE, M., JORDAN, G., LEVCHENKO, K., MCCOY, D., VOELKER, G. M., AND SAVAGE, S. A fistful of bitcoins: characterizing payments among men with no names. In *Proceedings of the ACM/USENIX Internet measurement conference* (Barcelona, Spain, Oct. 2013), pp. 127–140.

[32] MOORE, T., LEONTIADIS, N., AND CHRISTIN, N. Fashion crimes: Trending-term exploitation on the web. In *Proceedings of ACM CCS 2011* (Chicago, IL, Oct. 2011).

[33] NAKAMOTO, S. Bitcoin: a peer-to-peer electronic cash system, Oct. 2008. Available from `http://bitcoin.org/bitcoin.pdf`.

[34] SANKIN, A. Sheep marketplace scam reveals everything that's wrong with the deep web, Dec. 2013. `http://www.dailydot.com/crime/sheep-marketplace-scam-shut-down/`.

[35] STERBENZ, C. 20-year-old gets 9 years in prison for trying to poison people all over the world, Feb. 2014. `http://www.businessinsider.com/r-florida-man-gets-nine-years-prison -in-new-jersey-over-global-poison-plot-2015-2`.

[36] SUTHERLAND, W. J. *Ecological Census Techniques: A Handbook*. Cambridge University Press, 1996.

[37] THOMAS, K., HUANG, D., WANG, D., BURSZTEIN, E., GRIER, C., HOLT, T., KRUEGEL, C., MCCOY, D., SAVAGE, S., AND VIGNA, G. Framing dependencies introduced by underground commoditization. In *Proceedings (online) of the Workshop on Economics of Information Security (WEIS)* (June 2015).

[38] U.S. ATTORNEY'S OFFICE, SOUTHERN DISTRICT OF NEW YORK. Dozens of online "dark markets" seized pursuant to forfeiture complaint filed in Manhattan federal court in conjunction with the arrest of the operator of Silk Road 2.0, Nov. 2014. `http://www.justice.gov/usao/nys/pressreleases/November14/DarkMarketTakedown.php`.

[39] VAN HOUT, M. C., AND BINGHAM, T. silk road, the virtual drug marketplace: A single case study of user experiences. *International Journal of Drug Policy 24*, 5 (2013), 385–391.

[40] WANG, D., DER, M., KARAMI, M., SAUL, L., MCCOY, D., SAVAGE, S., AND VOELKER, G. Search + seizure: The effectiveness of interventions on seo campaigns. In *Proceedings of ACM IMC'14* (Vancouver, BC, Canada, Nov. 2014).

[41] WANG, D., VOELKER, G., AND SAVAGE, S. Juice: A longitudinal study of an SEO botnet. In *Proceedings of NDSS'13* (San Diego, CA, Feb. 2013).