Black-market archives - Gwern.net

Mirrors of Tor-Bitcoin black-markets & forums 2011-2015, and related material (*Bitcoin*, *Silk Road*)
created: *1 Dec 2013*; modified: *01 Oct 2016*; status: *finished*; belief: *highly likely*

> Dark Net Markets (DNM) are online markets typically hosted as Tor hidden services providing escrow services between buyers & sellers transacting in Bitcoin or other cryptocoins, usually for drugs or other illegal/regulated goods; the most famous DNM was Silk Road 1, which pioneered the business model in 2011. From 2013-2015, I scraped/mirrored on a weekly or daily basis all existing English-language DNMs as part of my research into their usage, lifetimes/characteristics, & legal riskiness; these scrapes covered vendor pages, feedback, images, etc. In addition, I made or obtained copies of as many other datasets & documents related to the DNMs as I could. This uniquely comprehensive collection is now publicly released as a 50GB (~1.6TB uncompressed) collection covering 89 DNMs & 37+ related forums, representing <4,438 mirrors, and is available for any research. This page documents the download, contents, interpretation, and technical methods behind the scrapes.

I have been involved in DNMs since June 2011 when Adrian Chen published his famous *Gawker* article proving that Silk Road 1 was, contrary to my assumption when it was announced in January/February 2011, not a scam and had become a functional drug black-market; fascinated, I signed up, made my first order, and began documenting how to use SR1 and then a few months later, began documenting the first known SR1-linked arrests. Monitoring DNMs was easy because SR1 was overwhelmingly dominant and BlackMarket Reloaded was a distant second-place market, with a few irrelevancies like Deepbay or Sheep and then the flashy Atlantis.

This idyllic period ended with the raid on SR1 in October 2013, which ushered in a new age of chaos in which centralized markets battled for dominance, the would-be successor Silk Road 2 was crippled by arrests and turned into a ghost-ship carrying scammers, and the multisig breakthrough went begging. The tumult made it clear to me that no market or forum could be counted on to last as long as SR1, and research into the DNM communities and markets, or even simply the memory of their history, was threatened by bitrot: already in November 2013 I was seeing pervasive myths spread throughout the media - that SR1 had $1 billion in sales, that you could buy child pornography or hitmen services on it, that there were multiple Dread Pirate Roberts - and other dangerous beliefs in the community (that PGP was paranoia, markets could be trusted not to exit-scam, that FE was not a recipe for disaster, that SR2 was not infiltrated despite the staff arrests & even media coverage of a SR1 mole, that guns & poison sellers were not extraordinarily risky to purchase from, that buyers were never arrested).

And so, starting with the SR1 forums, which had not been taken down by the raid (to help the mole? I wondered at the time), I began scraping all the new markets, doing so weekly and sometimes daily starting in December 2013. These are the results.

Download

The full archive is available for download from the Internet Archive as a torrent (magnet; item page)[1]. (If the

download does not start, it may be a Torrent client problem related to Getright-webseeding-support; if the torrent does not work, all files can be downloaded normally over HTTP from the IA item page, but if possible, torrents are recommended for reducing the bandwidth burden & error-checking.)

- providing information on vendors across markets like their PGP key and feedback ratings
- identifying arrested and flipped sellers (eg the Weaponsguy sting on Agora)
- individual drug and category popularity
- total sales per day, with consequent turnover and commission estimates; correlates with Bitcoin or black-market-related search traffic, subreddit traffic, Bitcoin price or volume, etc
- seller lifetimes, ratings, over time and by product sold
- losses to black-market exit scams, or seller exit scams
- reactions to exogenous shocks like Operation Onymous
- survival analysis, and predictors of exit-scams (early finalization volume; site downtime; new vendors; etc)
- topic modeling of forums
- compilations of forum posts on lab tests estimating purity and safety
- compilations of forum-posted Bitcoin addresses to examine the effectiveness of market tumblers
- stylometric analysis of posters, particular site staff (what is staff turnover like? do any markets ever change hands?)
- deanonymization and information leaks (eg GPS coordinates in metadata, usernames reused on the clearnet, valid emails in PGP public keys)
- security practices: use of PGP, lifetime of individual keys, accidental posts of private rather than public keys, malformed or unusable public keys, etc
- anthologies of real-world photos of particular drugs compiled from all sellers of them
- simply browsing old listings, remembering the good times and bad times, the fallen and the free

Works using this dataset

Papers:

Posts or articles:

Citing

Please cite this resource as:

- Gwern Branwen, Nicolas Christin, David Décary-Hétu, Rasmus Munksgaard Andersen, StExo, El Presidente, Anonymous, Daryl Lau, Sohhlz, Delyan Kratunov, Vince Cakic, Van Buskirk, & Whom. “Dark Net Market archives, 2011-2015”, 12 July 2015. Web. [access date] `www.gwern.net/Black-market%20archives`

Donations

A dataset like this owes its existence to many parties:

- the black-markets could not exist without volunteers and nonprofits spending the money to pay for the bandwidth used by the Tor network; these scrapes collectively represent terabytes of consumed bandwidth. If you would like to donate towards keeping Tor servers running, you can donate to Torservers.net or the Tor Project itself
- the Internet Archive hosts countless amazing resources, of which this is only one, and is a unique Internet resource; they accept Bitcoin
- collating and creating these scrapes has absorbed an enormous amount of my time & energy due to the need to solve CAPTCHAs, launch crawls on a daily or weekly basis, debug subtle glitches, work around site defenses, periodically archive scrapes to make disk space available, provide hosting for some scrapes released publicly etc (my arbtt time-logs suggest >200 hours since 2013); I subsist primarily on donations, and I too accept Bitcoins: `1GWERNi49LgEb5LpvxxGFSuVYo2K3BDRdo`

Contents

There are ~89 markets, >37 forums and ~5 other sites, representing <4,438 mirrors of >43,596,420 files in ~49.4GB of 163 compressed files, unpacking to >1548GB; the largest single archive decompresses to <250GB. (It can be burned to 3 25GB BDs or 2 50GB BDs; if the former, it may be worth generating additional FEC.)

These archives are xz-compressed tarballs (optimized with the sort-key trick); typically each subfolder is a single date-stamped (`YYYY-MM-DD`) crawl using wget, with the default directory/file layout. The majority of the content is HTML, CSS, and images (typically photos of item listings); images are space-intensive but I feel that images are useful to allow browsing the markets as they were and may be highly valuable in their own right as research material, so I tried to collects images where applicable. (Child porn is not a concern as all DNMs & DNM forums ban that content.) Archives sourced from other people follow their own particular conventions. Mac & Windows users may be able to uncompress using their built-in OS archiver, 7zip, Stuffit, or WinRAR; the PAR2 error-checking can be done using `par2`, QuickPar, Par Buddy, MultiPar or others depending on one's OS.

If you don't want to uncompress all of a particular archive, as they can be very large, you can try extracting specific files using archiver-specific options; for example, a SR2F command targeting a particular old forum thread:

```
tar --verbose --extract --xz --file='silkroad2-forums.tar.xz' --no-anchored --
wildcards '*topic=49187*'
```

Overall Coverage

Most of the material dates from 2013 to 2015; some archives sourced from other people (before I began crawling) may date 2011-2012.

Specifically:

- - 1776
  - Abraxas

- Agape
- Agora
- Alpaca
- AlphaBay
- Amazon Dark
- Anarchia
- Andromeda
- Area51
- Armory (not to be confused with the original Silk Road 1 weapons site which closed for lack of sales; this is a much later, independent site which was probably a scam)
- Atlantis
- BlackBank Market
- Black Goblin
- BlackMarket Reloaded
- Black Services Market
- Bloomsfield
- Blue Sky Market
- Breaking Bad
- bungee54
- BuyItNow
- Cannabis Road 1
- Cannabis Road 2
- Cannabis Road 3
- Cantina
- Cloud9
- Crypto Market / Diabolus
- DarkBay
- Darklist
- Darknet Heroes
- DBay
- Deepzon
- Doge Road
- Dream Market
- Drugslist
- East India Company
- Evolution
- FreeBay
- Freedom Marketplace
- Free Market
- GreyRoad
- Havana/Absolem
- Haven

- Horizon
- Hydra
- Ironclad
- Kiss
- Middle Earth
- Mr Nice guy 2
- Nucleus
- Onionshop
- Outlaw Market
- Oxygen
- Panacea
- Pandora
- Pigeon
- Pirate Market
- Poseidon
- Project Black Flag
- Sheep
- Silk Road 1
- Silk Road 2
- Silk Road Reloaded (I2P)
- Silkstreet
- Simply Bear
- The BlackBox Market
- The Majestic Garden
- The Marketplace
- The RealDeal
- Tochka
- TOM
- Topix 2
- TorBay
- TorBazaar
- TorEscrow
- TorMarket
- Tortuga 2
- Underground Market
- Utopia
- Vault43
- White Rabbit
- Zanzibar Spice

- Abraxas forums
- Agora forums

- Andromeda forums
- Black Market Reloaded forums
- BlackBank Market forums
- bungee54 forums
- Cannabis Road 2 forums
- Cannabis Road 3 forums
- DarkBay forums
- Darknet heroes forums
- Diabolus forums
- Doge Road forums
- Evolution forums
- Gobotal
- GreyRoad forums
- Havana/Absolem forums
- Hydra forums
- Kingdom forums
- Kiss forums
- Mr Nice Guy 1 forums
- Nucleus forums
- Outlaw Market forums
- Panacea forums
- Pandora forums
- Pigeon forums
- Project Black Flag forums
- Revolver forums
- Silk Road 1 forums
- Silk Road 2 forums
- TOM forums
- The Cave
- The Hub forums
- The Majestic Garden forums
- The RealDeal forums
- TorEscrow forums
- TorBazaar forums
- Tortuga 1 forums
- Underground Market forums
- Unitech
- Utopia forums

- Assassination Market
- Cryuserv
- DNM-related documents (primarily the Ross Ulbricht trial evidence exhibits)

- DNStats
- Grams
- Pedofunding
- SR2doug's leaks

Interpreting & analyzing

Scrapes can be difficult to analyze. They are large, complicated, redundant, and highly error-prone. They cannot be taken at face-value.

No matter how much work one puts into it, one will never get an exact snapshot of a market at a particular instant: listings will go up or down as one crawls, vendors will be banned and their entire profile & listings & all feedback vanish instantly, Tor connection errors will cause a nontrivial % of page requests to fail, the site itself will go down (Agora especially), and Internet connections are imperfect. Scrapes can get bogged down in a backwater of irrelevant pages, spend all their time downloading a morass of on-demand generated pages, the user login expire or be banned by site administrators, etc. If a page is present in a scrape, then it probably existed at some point; but if a page is not present, then it may not have existed or existed but did not get downloaded for any of a myriad of reasons. At best, a scrape is a lower bound on how much was there.

So any analysis *must* take seriously the incompleteness of each crawl and the fact that there is a lot and always will be a lot of missing data, and do things like focus on what can be inferred from 'random' sampling or explicitly model incompleteness by using markets' category-count-listings. (For example, if your download of a market claims to have 1.3k items but the categories' claimed listings sum to 13k items, your download is probably highly incomplete & biased towards certain categories as well.) There are many subtle biases: for example, there will be upward biases in markets' average review ratings because sellers who turn out to be scammers will disappear from the market scrapes when they are banned, and few of their customers will go back and revise their ratings; similarly if scammers are concentrated in particular categories, then using a single snapshot will lead to biased results as the scammers have already been removed, while uncontroversial sellers last a lot longer (which might lead to, say, e-book sellers seeming to have many more sales than expected).

The contents cannot be taken at face-value either. Some vendors engage in review-stuffing using shills. Metadata like categories can be manipulated (a category labeled "Musical instruments" may contain listings for prescription drugs - beta blockers - or modafinil or Adderall may be listed in both a "Prescription drugs" and "Stimulants" category). Many things said on forums are lies or bluffing or scams. Market operators may deliberately deceive users (Ross Ulbricht claiming to have sold SR1, the SR2 team engaging in "psyops") or conceal information (the hacks of SR1; the second SR2 hack) or attack their users (Sheep Marketplace and Pandora). Different markets have different characteristics: the commission rate on Pandora was unilaterally raised after it was hacked (causing sales volume to fall); SR2 was a notorious scammer haven due to inactive or overwhelmed staff and lacking a working escrow mechanism; etc. There is no substitute here for domain knowledge.

Knowing this, analyses should have some strategy to deal with missingness. There are a couple tacks:

- attempt to exploit "ground truths" to explicitly model and cope with varying degrees of missingness; there are a

number of ground-truths available in the form of leaked seller data (screenshots & data), databases (leaked, hacked), official statements (eg the FBI's quoted numbers about Silk Road 1's total sales, number of accounts, number of transactions, etc)
- assume missing-at-random and use analyses insensitive to that, focusing on things like ratios
- work with the data as is, writing results such that the biases and lower-bounds are explicit & emphasized

Individual archives

Some of the archives are unusual and need to be described in more detail.

Aldridge & Décary-Hétu SR1

The September SR1 crawl is processed data stored in SPSS `.sav` Data Files. There are various libraries available for reading this format (in R, using the `foreign` library like `library(foreign); sellers <- read.spss("Sellers -- 2013-09-15.sav", to.data.frame=TRUE)`.)

DNStats

DNStats is a service which periodically pings hidden services and records the response & latency, generating graphs of uptime and allowing users to see how long a market has been down and if an error is likely to be transient. The owner has provided me with two SQL exports of the ping database; this database could be useful for comparing downtime across markets, examining the effect of DoS attacks, or regressing downtime against things like the Bitcoin exchange rate (presumably if the markets still drive more than a trivial amount of the Bitcoin economy, downtime of the largest markets or market deaths should predict falls in the exchange rate).

For example, to graph an average of site uptime per day and highlight as an exogenous event Operation Onymous, the R code would go like this:

```
dnmUptime <- read.delim("dnstats-20150712.sql", na.strings="NULL",
                        nrows=6000000, colClasses=c("factor", "factor",
"factor", "integer",
                                                    "factor", "numeric",
"numeric", "POSIXct"))
markets <- dnmUptime[dnmUptime$type==1,] # type 1 = markets
dnmUptime <- NULL # save RAM due to dataset size
markets$Date <- as.Date(markets$timestamp)
markets$Up <- markets$httpcode == 200
daily <- aggregate(Up ~ Date + sitename, markets, mean)
library(ggplot2)
qplot(Date, sitename, color=Up, data=daily) +
geom_vline(xintercept=as.Date("2014-11-05"), color="red")
```

The service is a useful one and accepts donations: `1DNstATs59JANuXjbpS5ngWHqvApAhYHBS`.

Grams

Grams (subreddit) is a service primarily specializing in searching market listings; they can pull listings from API exports provided by markets (Evolution, Cloud9, Middle Earth, Bungee54, Outlaw), or they may use their own custom crawls (the rest). They have generously given me near-daily CSV exports of the current state of listings in their search engine, ranging from 2014-06-09 to 2015-07-12 for the first archive and 2015-07-14 to 2016-04-17 for the second. Grams coverage:

1. first:

The Grams archive has three virtues:

1. while it doesn't have any raw data, the CSVs are easy to work with. For example, to read in all the Grams SR2 crawls, then count & graph total listings by day in R:

```
DIR <- "blackmarket-mirrors/archive/grams"
# Grams's SR2 crawls are named like "grams/2014-06-13/SilkRoad.csv"
gramsFiles <- list.files(path=DIR, pattern="SilkRoad.csv", all.files=TRUE,
full.names=TRUE, recursive=TRUE)
# schema of SR2 crawls eg:
##
"hash","market_name","item_link","vendor_name","price","name","description","im
age_link","add_time", \
## "ship_from",
## "2-11922","Silk Road 2","http://silkroad6ownowfk.onion/items/220-fe-only-tw-
x-mb","$220for28grams", \
## "0.34349900", "220 FE Only TW X MB","1oz of the same tw x mb as my other
listing FE only. Not shipped \
##  until finalized. Price is higher for non FE
listing.","","1404258628","United States",...
# read in each CSV, note what day it is from, and combine into a single data-
frame:
grams <- data.frame()
for (i in 1:length(gramsFiles)) {
    log <- read.csv(gramsFiles[i], header=TRUE)
    log$Date <- as.Date(gsub("/SilkRoad.csv", "", gsub(paste0(DIR,"/"), "",
gramsFiles[i])))
    grams <- rbind(grams,log)
}
totalCounts <- aggregate(hash ~ Date, length, data=grams)
summary(totalCounts)
#       Date                 hash
#  Min.   :2014-06-09   Min.   : 2846.00
#  1st Qu.:2014-07-05   1st Qu.: 9584.25
```

```
#  Median :2014-08-26   Median :10527.50
#  Mean   :2014-08-21   Mean   : 9651.44
#  3rd Qu.:2014-09-29   3rd Qu.:11165.00
#  Max.   :2014-11-07   Max.   :19686.00
library(ggplot2)
qplot(Date, hash, data=totalCounts)
# https://i.imgur.com/ucPMvJQ.png
```

Other included datasets which are in structured formats that may be easier to deal with for prototyping: the Aldridge & Décary-Hétu 2013 SR1 crawl; the SR1 sales spreadsheet (original is a PDF but I've created a usable CSV of it); the BMR feedback dumps are in SQL, as is DNStats and Christin et al 2013's public data (but note the last is so heavily redacted & anonymized as to support few analyses); and Daryl Lau's SR2 work may be in a structured format.

2. the crawls were conducted independent of other crawls and they can be used to check each other
3. the market data sourced from the APIs can be considered close to 100% complete & accurate, which is rare

The main drawbacks are:

- the largest markets can be split across multiple CSVs (eg `EV0.csv` & `EV02.csv`), complicating reading the data in somewhat
- the export each time is of the current listings, which means that different days can repeat the same identical crawl data if there was not a successful crawl by Grams in between
- exports are not available for every day, and some gaps are large. The 2015-01-09 to 2015-02-21 gap is due to a broken Grams export during this period before I noticed the problem and requested it be fixed; other gaps may be due to transient errors with the cron job:

  ```
  @daily ping -q -c 5 google.com && torify wget --quiet --continue
  http://grams7enufi7jmdl.onion/gwernapi/$SECRETKEY
              -O ~/blackmarket-mirrors/grams/`date '+\%Y-\%m-\%d'`.zip
  ```

  so if my Internet was down, or Grams was down, or the download was corrupted halfway through, then there would be nothing that day.

Information leaks

Diabolus/Crypto Market

Diabolus/Crypto Market are two markets run by the same team off, apparently, the same server. Crypto Market had an information leak where any attempt to log in as an existing user revealed the status bar of that Diabolus account, listing their current number of orders, number of PMs, and Bitcoin balance, and hence giving access to ground-truth estimates of market turnover and revenue. Using my Diabolus crawls to source a list of vendors, I set up a script to automatically download the leaks daily until the hole was finally closed.

Simply Bear

Upon launch, the market Simply Bear made the amateur mistake of failing to disable the default Apache `/server-status` page, which shows information about the server such as what HTML pages are being browsed and the connecting IPs. Being a Tor hidden service, most IPs were localhost connections from the daemon, but I noticed the administrator was logging in from a local IP (the `192.168.1.x` range) and curious whether I could de-anonymize him, I set up a script to poll `/server-status` every minute or so, increasing the interval as time passed. After two or three days, no naked IPs had appeared yet and I killed the script.

TheRealDeal

TheRealDeal was reported on Reddit in late June 2015 to have a info leak where any logged-in user could browse around a sixth of the order-details pages (which were in a predictable incrementing whole-number format) of all users without any additional authentication, yielding the Bitcoin amount, listing, and all Bitcoin multisig addresses for that order. TRD denied that this was any kind of problem, so I collected order information for about a week.

Modafinil

As part of my interest in the stimulant modafinil, I have been monthly collecting by hand scrapes of all modafinil/armodafinil/adrafinil listings across the black-markets; the modafinil archive contains the saved files in MHT or MAFF format from 2013-05-28 to 2015-07-03. Sampled markets include:

- Abraxas
- Agora
- Alpaca
- AlphaBay
- Andromeda
- Black Bank
- Blue Sky
- Cloud-Nine
- Crypto/Diabolus
- Diabolus
- Dream
- East India Company
- Evolution
- Haven
- Hydra
- Middle Earth
- Nucleus
- Outlaw
- Oxygen
- Pandora
- Sheep
- SR2
- TOM

Pedofunding

A crowdfunding site for child pornography, "Pedofunding", was launched in November 2014. It seemed like possibly the birth of a new black-market business model so I set up a logged-out scrape to archive its beginnings (sans any images), collecting 20 scrapes from 2014-11-13 to 2014-12-02, after which it shut down, apparently having found no traction. (A followup in 2015 tried to use some sort of Dash/Darkcoin mining model; it's unclear why they don't simply use Darkleaks.)

Silk Road 1 (SR41)

Sources:

- appendix to Van Buskirk et al
- 2013 scrape provided me by anonymous

SR1F

This archive of the Silk Road 1 forums is composed of 3 parts, all created during October 2013 after Silk Road 1 was shut down but before the Silk Road 1 forums went offline some months later:

1. StExo's archive, released anonymously

   This excludes the Vendor Roundtable (VRT) subforum, and is believed to have been censored in various respects such as removing many of StExo's own posts.
2. Moustache's archived pages

   Unknown source, may be based on StExo archives
3. consolidated `wget` spider

   After the SR1 bust and StExo's archiving, I began mirroring the SR1F with `wget`, logged in as a vendor with access to the Vendor Roundtable; unfortunately due to my inexperience with the forum software Simple Machines, I did not know it was possible to revoke your own access to subforums with `wget` and failed to blacklist the revocation URL. Hence the VRT was incompletely archived. I combined my various archives into a single version.

   Simultaneously, qwertyoruiop was archiving the SR1F with a regular user account and a custom Node.js script. I combined his spider with my version to produce a final version with reasonable coverage of the forums (perhaps 3/4s of what was left after everyone began deleting & censoring their past posts).

SR2

Sources:

- in January 2014, Sohhlz made & distributed a scrape of SR2 vendor pages akin to StExo's SR1 vendor dump

SR2Doug

In 2015, a pseudonym claiming to be a SR2 programmer offered for sale, using the Darkleaks protocol, what he claimed was the username/password dump and SR2 source code. The Darkleaks protocol requires providing encrypted data and then the revelation of a random fraction of it. This archive is all the encrypted data, decryption keys, and revealed usernames I was able to collate. (The auction did not seem to go well as the revealed data was not a compelling proof, and it's unclear whether he was the genuine article.)

Copyright

The copyright status of crawls of websites, particularly ones engaged in illegal activities, is unclear.

- to the extent I hold any copyright in the contents, I release my work under the Creative Commons CC0 "No Rights Reserved" license
- the SR1 Christin et al 2013 dataset is licensed under the CC BY-NC
- other authors may reserve other rights

(I request that users respect the spirit of this archive and release their own source code & derived datasets to public, but I will not legally demand it.)

Previous releases

Some of these archives have been released publicly before and are now obsoleted by this torrent:

Verification

Integrity of the archive can be verified using PAR2: `par2verify ecc.par2` Up to 10% of file damage/loss can be repaired using the supplied PAR2 files for FEC and `par2repair`; see the man page for details.

Signed SHA-256 hashes of the archives:

```
-----BEGIN PGP SIGNED MESSAGE-----
Hash: SHA512

8b05d5fcba36db6889af4fe23d1117a48c39b0808332d32919f9d7c835380721  1776.tar.xz
cc6f54d5818e13fb585b14d6c414fcdbf4d20a4e1ab3aa398f5ce05287a1d1b0  2015-sr2doug-
claimedsr2leaks.tar.xz
6e082846f83dc9e06950fc29095491d303f5b336d65bbe6760db2c03d969cf02  abraxas-
forums.tar.xz
3dcb6ba24bc3e4f75e13827bb1e2f0632ed269b10e6158bdb554cc50983f1204  abraxas.tar.xz
4231b81aa12d529f4502129683f8d5f1e0ef1f813d252d6edcce9d3b75eecdd2  agape.tar.xz
4838969a87610fe80678ae72a3d631ab2aaa5a6b219cd67226f528d96c4fc958  agora-forums-
20140421-whom-astorposts.tar.xz
f6afe2df9238ce5cecea6dac70fd7c4b67a444824eccf07667ca46b15a167734  agora-forums-
2014093020141016-rasmusandersen.tar.xz
```

```
5730cc4e7e34138aeee934985b937ba8a2ae78f23580ba9a666348fb04fb3583  agora-
forums.tar.xz
4e7d5d4f63be66956037d4c27f3b97c0b980addd3ed5029b24904ab69f705c9d  agora.tar.xz
ab9fc0d2324ddbd03fcf5a9e8b9213fc6c650fcb1f7e99f9d3b7a63cd67923af  alpaca.tar.xz
1bbb33eda2094f662d982cad045033541a5fb22e850359883fa3decb5a0d81d2  alphabay.tar.xz
7a61ae8945322455f9b6d0afdad2751847f9a294b951920ea6cccaa8f3b06d86
amazondark.tar.xz
19e634813d8038474460d72e0c5311a7d97a9a2e9e9089eab32a719cf4a0c377  anarchia.tar.xz
8da899bae2e51384afa8d4f839a45371a1b1c5b22a52685f698aced1dba5adbd  andromeda-
forums.tar.xz
0c95881e291bde995dc33ae8ee516ca7c8b200cb8dd3b967f8dc62ec5a36b6b2
andromeda.tar.xz
3466f8f9637aab4f2d74ef9c242be7aeff08d5adfadcffe7ca69ce58392a62a9  area51.tar.xz
d9f4f00dba4a44cc7bb45b19d9967046be56b83328c0149697cdf44862438ef2  armory.tar.xz
d9e887e1370f690724e9a178287baf5c85e5e8a900e9e9dae019b795e2afdb6e
assassinationmarket.tar.xz
c959b430f7aef932d26fe389498c6f4d3d7d02421e9d05c204803b009317869b  atlantis-
20130921-christin.tar.xz
e1539816b1318badf183152960783697f234ce6c972e90ed2830b119d620313a
blackbankmarket-forums.tar.xz
c9e4940b16078ad2982a55c4c1221054ad3b6a2cac99517d55fc24063a71efdd
blackbankmarket.tar.xz
cbb17ccd867d242ce571ea692a4672474c8330679d3a41e2fff7ebaa511ffd58
blackgoblin.tar.xz
f68f7bb73b47161d8d0499eb062ddd8b4f7b267cad9b2c9179b3a6d309ac9d2b
blackmarketreloaded-20131017-userlist.sql.xz
eed272069f2f057dc6894bbb078041c4bf64db3936a1218cf9f9db9c42518839
blackmarketreloaded-20131225-feedback-wousd.sql.xz
6b0a07ea3cbf67cd60c743a52cdf0427a3e4e587655e3950a75c48fad2f57085
blackmarketreloaded-elpresidente.tar.xz
63d95bc6baa947842247084f0332e8e5ccc465ad112df2fe4d88e1a024aeb5fc
blackmarketreloaded-forums.tar.xz
84598eccbc428ce0325327618f2d7566e55ab799f46e030a1c5b8295e0397fd0
blackservicesmarket.tar.xz
9d0f068823a37eb405b2bf6014ba3051a6cddfb78997111ae1a0c7507c60dd3e
bloomsfield.tar.xz
a4477cf586ff6b18df649e5bfb47d825f2c604c3913b934c235eafa514d0025b  bluesky.tar.xz
8e9b225be42d4f3cff9f835e7f24ba414a6d72e3131d77655f3fc7d05c3b6208
breakingbad.tar.xz
9bb37c2f8b68730b02d38ddf3be04154384f2c79a70505a3324fb8b973e4553c  bungee54-
forums.tar.xz
78f5599807f5adc1a068cb86f8a8c7ad194d67d28ef5f451076a40a8587f1776  bungee54.tar.xz
9afedc1135e8a96a61974fb663eaaabef2476bafbc4193dc9f6744402573c98c  buyitnow.tar.xz
f9559a82359cc33f9e9b093d5aa7a6d8b4deebb39aa13841c2fb91ea6f6fdac5  cannabisroad2-
```

```
forums.tar.xz
db133bef60e5c338757af23809175a8f64a9b4ca1dbebcbf3d8930af590a924a
cannabisroad2.tar.xz
9fca953f118c80f6e61264b513872404ab67b51e06e544bba35284b1fcf8defd  cannabisroad3-
forums.tar.xz
173d4f60232941b18a5cdef0c04d45a678fd1f9c4ff0a4a1158266cd1f15c4fa
cannabisroad3.tar.xz
5feeb4f56b4b2c0ab058e45d82543588ec09386f50a3663af53109abb72d66c6
cannabisroad.tar.xz
e0b5355ac6fc07b53dd6ae6767783462173d0e5a62f77b3ca23b699d5f59ce25  cantina.tar.xz
a2db7e54af153958d9d0bac0bf4088ff371e28c7e5510e5fae6b850af88dda8f
cloudnine.tar.xz
9010bcfd779f01508075d341e278dbd412c2350d9fba41bb96a1345494956b40
cryptomarket.tar.xz
66d0236a256059df1ae4f0c6da5e7ded59f83f4534e2293c576575ad0191262e  cryuserv.tar.xz
1dd482381d3a4ff8b30c4750696f1de1fbceb19ce29061ad39f5ce33092239f3  darkbay-
forums.tar.xz
366e30bdb6d84e6cbe5d54909d2f49a7f95e0f232ecd886ea53e729f479104e0  darkbay.tar.xz
d7f666e3fd244c299621c6fb7beb20111690e4e7c8786161f1534c23c7836d51  darklist.tar.xz
c6d2478c2a0f860c4b1e8507a5925f699ee39edf8dead1df2cec5d0d94b51af2  darknetheroes-
forums.tar.xz
1197eae4c7cb83ed97aa5374365a26b67beea75bf053a9927b2e8948393fe58d
darknetheroes.tar.xz
623ff7d3509727be5936f27ab95cd2b40432f25b0f07e20df7062e5e2cd55217
darknetnation.tar.xz
23e4932551b2a56c12d151d2f14140d5c9a7c25407b766b34d48456c5dbab589  dbay.tar.xz
f8b3cd5c861e7c32147ad720538728f113bcda0f41760ef7475ffbaf26037490  deepzon.tar.xz
2199f5062ad587d355ed683b894ada4dd1529ec50c5f5761b523cdaff9c20b5c  diabolus-
cminfoleak-20150220-20150311.tar.xz
f1f6df5855287def19443db64082aa1c7df507991a6968dca6f5f097b024e253  diabolus-
forums.tar.xz
42d1a476d9eb6b9b4807789ba08c5791054d41f3d6b9e7506a78a309603bad78  diabolus.tar.xz
ddeed8ce25ef813814522bffe2224f390c84dcdca4dcd0c3023b49d0a63a8b5a  dnstats-
20150712.sql.xz
649e311c427398006bf390f7827fe3534026c730a905766cb9f3e78bad82b520
documents.tar.xz
2f2523f4125e64acaa86ebacb8fe2f08fc640608aabc95d747e9319bf9446e12  dogeroad-
forums.tar.xz
78079f03495ba405a04860fb546421780f9bc1cdcf06025e7abd29033f77c450  dogeroad.tar.xz
768482dd0aae12fab023497cda437fd290657ac1e9df29a6b65f1b142d1ce8af
dreammarket.tar.xz
229373106b35aa6d72a71f7dc48e90d1da47647cc58348ee0cb768a3926294c4
drugslist.tar.xz
f8a324d215858918d781436a09d51bfaa88c2b9bd59ef6af4a75f52c81891a6c
```

```
eastindiacompany.tar.xz
23449de611a42899bcb27db8186d194f7b805ee7e55034ec5ab17adee226aecd  evolution-
forums-2014093020141016-rasmusandersen.tar.xz
109eb980c11ed37b29321f6403cb5e95614f3c44525a549164d95d0a52eb94cf  evolution-
forums.tar.xz
a6a0ccd588635903f1e914390f36bb9a56f562d37b9e92d6e58dac6364b35b8a
evolution.tar.xz
0b2e5eac28bad63ca832aeeebb8a759dec21bbf2b52eb5f816dc010ab5a825f3  freebay.tar.xz
336c43eb0794174bb8c58cb8b018a8e019a4dd1719a298051b0c0e4ba04a7109
freedommarketplace.tar.xz
61f2037e6245d2e0a23f87df142ff53c0736da26844a3a3f7d869fdd1b835202
freemarket.tar.xz
af4dd8003b015519677c802cc3c19f0910cb79541876be0be719e0c176fe7f5e  galaxy.tar.xz
0d963a63009ef5b581ce705555a608997cfc7220971a26236d8f12b6268c224c  gobotal-
20140818-20141102.tar.xz
0cecd5e78416328caf06614ee6a8fabee0d91b8aecddd9ca2d67f059ff7497d6  grams.tar.xz
2dccb3df553b89dfceb5ba4930269ffff4fcd39dc6c876ca6cfc9e85c98bda9a
grandtrunk.tar.xz
2fe55a93c6c7b69b40a5bfe1c1dcd7c0cc4601045696870f1b4dad460c93ea70  greyroad-
forums.tar.xz
419e97c0c28784e6077f296746bf2ae5b4899cc0fef2756108c3b5c3d5ed9b13  greyroad.tar.xz
d7624f290f63642d3d875d0b94baf84af89cd63e2abab57c1889bf8d18883596  havanaabsolem-
forums.tar.xz
94bafe76779807cdf7cc86d0534da64155b22e40db79f1bb801e865becd44fc6
havanaabsolem.tar.xz
32475d62c6ff9cce00063b6473576782a2941bf1dc2e05a0f9a6bc9880ed91c3  haven.tar.xz
b69715d148fa02e87af8143d36152f4deda57b39f85fe4da47e8090e5e93c348  horizon.tar.xz
b06b7f272934b661920eae5ba9cc3ac8480c8e94ca86d7ab039988cdbf348f2a  hydra-
forums.tar.xz
0cf4eda89b71d17a9a539599053e06f4fed4322c0ea306edb6e30c950ab0d16b  hydra.tar.xz
cebec4d92f705475a61ab0fe66c905d509c737139276e96c4c8826539bdd2e07  ironclad.tar.xz
deb71f9e282bbc477c16c922ea8731ecc8817244808619fe881c22467df1d213  kingdom-
forums.tar.xz
466772600b49a37d6f5078c1534d889f0b3d3d7ccb165228292e1121217395fd  kiss-
forums.tar.xz
74436c0b38dab5007ad212e5c8bb7f1d67708fbdfbbaf6488a80ea637cdcd912  kiss.tar.xz
73ed19cbc40d0d313cf91ed68c7c8f931438238605076bea95c6db7e41a382bd
middleearth.tar.xz
69e783616806f90715b3a63b8f8623ca7ea83f81a48b71e0fadbfa85dfca214f
modafinil.tar.xz
fc29a84ba388a0bf7aa7c27437ea2e53462bfdb527f00c45958b2d15a43237ef
mrniceguy2.tar.xz
796fa38de4eae84797ce07c30a158123b61224dffdb6e94dfd5be39f8a96a187  mrniceguy-
forums.tar.xz
```

```
146f2ae90fd4fa25932f43596e621065204a07ca5b8149d4e6af142abea32597  mtgox-2011-
usernamepasswordleak.csv.xz
0d4136f8e59a4cedfbfac30da33a846d42ed1c9e6e1af8ed030be8ac42e42522  mtgox-20140309-
leak.tar.xz
e22b5c83f04ac244e4e77bad4e91588642373a371b3b5606c311a5021bd2eba2  nucleus-
forums.tar.xz
87fb7a67bfd55f25f882fbf10e10c82bf2872721109f47728192b5be0e830252  nucleus.tar.xz
ff975d6dc3c91c5b2fd42a86c54acecfed17616dcd80ba5a320ff4b4df2e89fd
onionshop.tar.xz
1b95c06289b081c1dc674dc5d4e055f61fd1609b8a75d5a65a51134407639c11  outlawmarket-
forums.tar.xz
4d7d1c24197c89252d515e35ef1bc3c80543180e952ed3e6aae821eb48d17d4c
outlawmarket.tar.xz
11327c8c1915e802cd6083e590217e8e93b19767c9453fc62291e24b96a0a420  oxygen.tar.xz
5355211f6e1b8a338115ef10b2c8498af3b4ee494405b51147f1ffe27645d7b5  panacea-
forums.tar.xz
58a76cba9c7ca06c4d92ce03bb39bddf24f15dabeee508f2004f0158bf1aca70  panacea.tar.xz
ed17677aa7269d725cdd81fc1832655a76b3ab701a0ca356b1182443622bedd7  pandora-
elpresidente.tar.xz
9f9de82834b46973a5712a6b1dcabe3cb2af1b3c42348d3f2ab4534b59f64dc6  pandora-forums-
20140421-whom-astorposts.tar.xz
29bb6c5add500b077b3545559871eda0515887f8847380f1024072ce6cc785aa  pandora-
forums.tar.xz
d6e00fb115cecb5739e72c994243edf3199a7b2c9524ebe1e55983bcd2dbc894  pandora.tar.xz
0dfcfdac5d359b508efae9c50cb861f5403924e047de00831db758841a469bfa
pedofunding.tar.xz
427bc78c1e466a7bdc7f0b667d125aced3de76da7bfd8fed5fce564f44421372  pigeon-
forums.tar.xz
6fe6fd24b0b604ec70b9e56610743f3bdf91683d24e6ade3a149ecd61b7b787f  pigeon.tar.xz
bd634bf2b2943fb1d01c548f1d731d86c8344d319b799a03a9197874e8e01772
piratemarket.tar.xz
f8dbee89392ebced3a529a972e19c5146aaa3cfe8ce9d25005f538d41b47c2ed  poseidon.tar.xz
71b44fc678bebb8122ddfdba02e2ef80335f72eaf49b4f11ef3204ee7f29ec35
projectblackflag-20131103-anonymous-logsdump.tar.xz
0000462319ea6467b0a25f070f659124966518da3adce1a0fa92d81a84a24e59
projectblackflag-forums.tar.xz
b2ec62fbe54b8148f7e6e7738b84d0d7d45c6b7a91b951494a9a8ab20769e24b  revolver-
forums.tar.xz
4f8573bded758c065f86c1eae189d69c1ad622fb6558d10d4aef780e699e09c2  sheep-
elpresidente.tar.xz
073829fc8ae4fe9e6920b2c3232bc253ebe6c877b29264a569651e5d76c3b191  sheep.tar.xz
4099f3d49d74d8828b12d8ff532979531c5ca31092985457e93f5f5e9fafbdc1  silkroad1-
20111103-delyankratunov.tar.xz
57b641200c30bf6a801fe2faf462d507fcc99c678567943f25af9d0c51970879  silkroad1-
```

```
20120722-vanbuskirk.docx
59e72f95201726cc46d9680f97a53f44c45f242b57a96567916c4cb76a863d5e  silkroad1-
20120723-christin-censored.tar.xz
da8726427d1b13f850a9647a34757ee95be000c036a5ec370e8f43b01fde6609  silkroad1-
20130703-anonymous.tar.xz
a3fe8ec72186e7ec02fe206f92616688fae07b756f06a555bd8f306a92b0451b  silkroad1-
20130915-aldridgehetu.tar.xz
12876b0783fb928a9c982dff048155fae331b174e08847e66a3100a9f74c9369  silkroad1-
forums-20130703-anonymous.tar.xz
5533a90285c0d072d62ebf681cfe717987dfe595f13b96e1e8dc9ae1ed7274ab  silkroad1-
forums-20131103-gwernrasmusandersen.tar.xz
3a28097c243843cc69d365b1c6456075679bfa09cd3a50daa6105a0c7f4df837  silkroad1-
forums-anonymous.tar.xz
37db1b2eab69923e22cb0d2ee65426152cb11ab09d92d1d6013a2fe7f20aa7d0  silkroad1-
forums-stexo.tar.xz
eac0013182b996b4a77f446a28ffabd74f23ea0fa32eeaa6f3bc499081c372c8  silkroad1-
forums.tar.xz
ab1ffac3b85b9cbb2d7ff80ed28a1899561f945758196ba3976dbb2e5b8b4c21  silkroad1-
vendorprofiles-stexo.tar.xz
2df744013fedfdacfd349472e05981316dbf392ccb56e627ff6d6f09b4ad7a8a  silkroad1-
wiki.tar.xz
1c8e643eade9750b39485c5e101f65d2c12ec977cb7b681cd8df064eccf4c0e7  silkroad2-
20140129-sohhlz-vendors.tar.xz
3381cd4305c4cd909aa86cf218a1022e6be5ed227d6eb728603c41b9956c7a28  silkroad2-
20140927-daryllau.tar.xz
7367dc56f15f61212d8567033a4d3a9468622e05f86d38607a70d5686164648a  silkroad2-
forums-20140419-whom-astorposts.tar.xz
0900093d7100b4faf983707b4b1e0ec1fae3c4b18270eaa8eedfe4f8b69a6e23  silkroad2-
forums-2014093020141016-rasmusandersen.tar.xz
a473132cb8eec64aea2066628a24628a0c1eb38c195c9945c700dd19f1f972f2  silkroad2-
forums.tar.xz
2abc793c7fdfce31d375db11307b66aa69cb91f4c684408840d546bf4e61e41b
silkroad2.tar.xz
3384789112185d81544dcad5bc69967cd44b097b7a772da48f5a1226b43155de
silkroadreloaded.tar.xz
ed9d47ecc9afce0f541386471da9894c436833b89da06663ffbc5ab6de2beacf
silkstreet.tar.xz
7e254452405543c27ee47c0bf6a455fe34443a6fa335a904e086fef61cf6f330
simplybear.tar.xz
80c759f67a5eac57b6345417dff1181690a80ecb965a14ce812ab79d315f2f2d  tcf.tar.xz
6f0775201cb379bb0845c60fde22e66b8aa7d5319d6046987202cdc9065b0591
theblackboxmarket.tar.xz
c25c1f2b35d1cf1f38f1f009b40d559f5a0aaf484248d98aed7b9942fade20a8  thecave.tar.xz
078cc6e61cb37c56f671b6d87ca243e885c2a37a17645d73d26c01e56b28afe4  thehub-forums-
```

```
20140420-whom-astorposts.tar.xz
5620dae0fac58b30bff4efbf116ce9674d071c3d43fe7cef2f5f84c2950b4182  thehub-
forums.tar.xz
c542fed2541d059c466d0b9dc402465952a778b1ef584a3af73e7ad34d953f7e
themajesticgarden-forums.tar.xz
a8a57924768c5f7ad4062fe0b6931722a078caab91b65a515b554817b2e4c1dc
themajesticgarden.tar.xz
8deee8650c55fbd4cfb8366a4f8b5e8a5370b525f676769de34f81a8864e92d2
themarketplace.tar.xz
420889ca017ac87c92a0ff774d21dc79c3abc1958c8dee0dcc11e1af59fd680d  therealdeal-
forums.tar.xz
b1ee23d727b30c486c3d197212ac91ac16f18b78b30ba5346854bedf81e6b821
therealdeal.tar.xz
70cf9c9a75815e9a514d4a5eb69aef77df862f3c8e36aff19feed8dae7c1e1cc  tochka.tar.xz
32acbc1289525785c12f179a7da9ce76a838e5a13a4dbaa6fb16c3f1870f9d98  tom-
forums.tar.xz
3f62941a988c166ebcec9c788069de1d30a3c365f0b1da1921d342c8a4df3a35  tom.tar.xz
6c50bd480914e0c257b6e85a3e22a087e0e058614d465f7269e2ebd1f867a35a  topix2.tar.xz
fee6a7cd032648bebaae7752045bcd64c0a069c0abd311c53686323103fe7ede  torbay.tar.xz
76fdc6da85a4d697e2e5ed5b9c3d608c5d1ac33a0831fd0701cfd0c6c922e9db  torbazaar-
forums.tar.xz
5b9b457c2e541fc618461b69c14511b03fff886daed25ba1e0cb49a89c5b749c
torbazaar.tar.xz
0f3c3a34496feeb44f258e07ee46704a38f856e975e394bcf689e03a18d263ca  torescrow-
forums.tar.xz
7e4bf1ef60826367375ab419b068ce1b61daf231cda407594f595ec3bffc6d50
torescrow.tar.xz
1b911a07423900ee4ef9ff71e9d1f4752bfa89ad9c473b760263314f56c7a021  tormarket-
20131213-dpr2-dbdump.mht
e229859ffa92bb7c142d2d54317d4b571e48dcc030d412fc93489a3f5aaa9faa  tormarket-
elpresidente.tar.xz
55b50e6e9283df50e68d1843db0d07360cc0e6c7d2d032dc00de2c04a00cd489
tormarket.tar.xz
f81a11e6dd8779a4bf077f9bc833740536ed202d2dca106ab5122d758784bf74  tortuga1-
forums.tar.xz
15c7d2ad0b525a9f3ae417dc63a670698204ac755a28bd98f104b0b240f3a4fd  tortuga2.tar.xz
0bb2324c424faa0481a3ca5b4004e57493eacfb7a521a7018edb40c3b467037b
undergroundmarket-forums.tar.xz
2153d48e75b60942cb7287a06b93c43b2968fb175af7b4f82fff59577674e9f6
undergroundmarket.tar.xz
13bb5eda0762a41aecc74caf3f3a527035b0015ea71019ba4d2d2363aeaf86d3  unitech.tar.xz
2811a120a4db56907498b2758b0b5d8b2d43c2167a40b2bf0c6e432ba383ff55  utopia-
forums.tar.xz
c64666bf5ea4218f7b69d366243ce13a1c8fc21a68d4e24a6ac8c7c3d8bf6908  utopia.tar.xz
```

```
9278f2ed7191642cf736bc4dc88c2ccbe7c0b1af6cc6e6ffcb283263a4aef729  vault43.tar.xz
8087f7b4a7781ffc634d0baa2ac4a7cec7b7b1bd5a619f89cb43d49faae002b7
whiterabbit.tar.xz
dc64656700ad46505bd02412d7af5a04d60aba138c713720a00d80cc4bd20000
zanzibarspice.tar.xz
-----BEGIN PGP SIGNATURE-----

iQIcBAEBCgAGBQJVoq+QAAoJEH3Oo4eJxYjM52IP/3ZMzulM6TuwKfkcsGDrFe4Q
X3gQL4Ru2N80jWWcUj3hA/SxEyhs5gWA/xnLZr1HFPPEOXZQRMZb5G3tVQ7clhxL
dH2q7YPl+1L151iqtZHATYMcK8kSB7gbs8S33JU5SkS+y7R0tOXI9fpVuhnaD6HN
q3nGEKrSXI0CaC2o4bBxmUh/1WsimTySiNbcErdj0jMns10MKeYwTq98E+6yc+XQ
ItsMqS9gfSVlGN0yLRedc+kI+Y3M4ujLzY5aHC7PDv2RnpZhRMV68cSbsTc4FD7m
A7AOFKHukUhDPBqp1d3BEU/IiNqY4YhfIkmDMIQ8y2ioYG+rkk0SMojb3OYXgv0p
ioO0QuHNsJSomXYe9OkNoF9y2Tb99nJr7Wr6TFyJ4Geeow9B9p0j2LWFwfrpD3oq
eevXcIQruyi1AG4sK3/F6UG+GAZ3ZgsvcECoRc0+zytXNF0sn14WNcnyqGmtyfo1
/Y0KcDA0RCiWyvUTyAHWjjv0xOxVGDij8r9aqDM+8UgTsECIL6tlTo/Ifhm/k4a6
qF0adhyCpeFPAhmW2kz7BYsmtM0TzWDV/eD3h3mrpo8bn0ILgZr4MpEpLn3WPjY/
D+ZepCz12epZSURHV+6SWFteO6PM44fU895ezBq/iU5ZIRK8uvTShR6KEtPivJFp
fYrFFbOhBc6KRQbNJ8o2
=U0bP
-----END PGP SIGNATURE-----
```

How to crawl markets

The bulk of the crawls are my own work, and were generally all created in a similar way.

My setup is a Debian testing Linux system with Tor, Privoxy, and Polipo installed. For browsing, I used Iceweasel; useful FF extensions included LastPass, Flashblock & NoScript, Live HTTP Headers, Mozilla Archive Format, User Agent Switcher & switchproxytype, and RECAP. See the Tor guides.

1. when a new market opens, I learn of it typically from Reddit or The Hub, and browse to it in Firefox configured to proxy through `127.0.0.1:8123` (Polipo)
2. create a new account

   The username/password are not particularly important but using a password manager to create & store strong passwords for throwaway accounts has the advantage of making it easier to authenticate any hacks or database dumps later. (Given the poor security record of many markets, it should go without saying that you should not use your own username or any password which is used anywhere else.)
3. I locate various 'action' URLs: login, logout, 'report vendor', 'settings', 'place order', 'send message', and add the URL prefixes (sometimes they need to be regexps) into `/etc/privoxy/user.action`; Privoxy, a filtering proxy running on `127.0.0.1:8118`, will then block any attempt to download URLs which match those prefixes/regexps

   A good blacklist is critical to avoid logging oneself out and immediately ending the crawl, but it's also important

to avoid triggering any on-site actions which might cause your account to be banned or prompt the operators to put in anti-crawl measures you may have a hard time working around. A blacklist is also invaluable for avoiding downloading superfluous pages like the same category page sorted 15 different ways; Tor is high latency and you cannot afford to waste a request on redundant or meaningless pages, which there can be many of. Simple Machine Forums are particularly dangerous in this regard, requiring at least 39 URLs blacklisted to get an efficient crawl, and implementing many actions as simply HTTP links that a crawler will browse (for example, if you have managed to get access to a private subforum on a SMF, you will *delete your access to it* if you simply turn a crawler like wget or HTTrack loose, which I learned the hard way).

4. where possible, configure the site to simplify crawling: request as many listings as possible on each page, hide clutter, disable any options which might get in the way, etc.

   Forums often default to showing 20 posts on a page, but options might let you show 100; if you set it to display as much as possible, the crawls will be faster, save disk space, *and* be more reliable because the crawl is less likely to suffer from downtime.
5. in Firefox, I export a `cookies.txt` using the FF extension Export Cookies. (I also recommend NoScript to avoid JavaScript shenanigans, Live HTTP Headers to assist in debugging by showing the HTTP headers and requests FF is actually sending to the market, and User Agent Switcher to lock your FF into showing a consistent TorBrowser user-agent)
6. with a valid cookie in the `cookies.txt` and a proper blacklist set up, mirrors can now be made with wget, using commands like thus:

```
alias today="date '+%Y-%m-%d'" # prints out current date like "2015-07-05"
cat ~/blackmarket-mirrors/user-agent.txt
## Mozilla/5.0 (Windows NT 6.1; rv:31.0) Gecko/20100101 Firefox/30.0

cd ~/blackmarket-mirrors/cryptomarket/
fgrep --no-filename '.onion' ~/cookies.txt ~/`today`/cookies.txt >
./cookies.txt
http_proxy="localhost:8118" wget --mirror
    --tries=5 --retry-connrefused --waitretry=1 --read-timeout=20 --timeout=15
--tries=10
    --load-cookies=cookies.txt --keep-session-cookies
    --max-redirect=1
    --referer="http://cryptomktgxdn2zd.onion"
    --user-agent="$(cat ~/blackmarket-mirrors/user-agent.txt)"
    --append-output=log.txt --server-response
    'http://cryptomktgxdn2zd.onion/category.php?id=Weed'
mv ./cryptomktgxdn2zd.onion/ `today`
mv log.txt ./`today`/
rm cookies.txt
```

To unpack the commands:

- the `fgrep` invocation minimizes the size of the local cookies.txt and helps prevent accidental release of a full cookies.txt while packing up archives and sharing them with other people
- wget:

  - we direct it to download only through Privoxy in order to benefit from the blacklist. Warning: wget has a blacklist option but it does *not* work, because it is implemented in a bizarre fashion where it downloads the blacklisted URL (!) and then deletes it; this is a known >12-year-old bug in wget. For other crawlers, this behavior should be double-checked so you don't wind up inadvertently logging yourself out of a market and downloading gigabytes of worthless front pages.
  - we throw in a number of options to encourage wget to ignore connection failures and retry; hidden servers are slow and unreliable
  - we load the cookies file with the authentication for the market, and in particular, we need `--keep-session-cookies` to keep around all cookies a market might give us, particularly the ones which change on each page load.
  - `--max-redirect=1` helps deal with a nasty market behavior where when one's cookie has expired, they then quietly redirect, without errors or warnings, all subsequent page requests to a login page. Of course, the login page should also be in the blacklist as well, but this is extra insurance and can save one round-trip's worth of time, which will add up. (This isn't always a cure, since a market may serve a requested page without any redirects or error codes but the content will be a transcluded login page; this apparently happened with some of my crawls such as Black Bank Market. There's not much that can be done about this except some sort of post-download regexp check or a similar post-processing step.)
  - some markets seem to snoop on the "referer" part of a HTTP request specifying where you come from; putting in the market page seems to help
  - the user-agent, as mentioned, should exactly match however one logged in, as some markets record that and block accesses if the user-agent does not match exactly. Putting the current user-agent into a centralized text file helps avoid scripts getting out of date and specifying an old user-agent

- logging of requests and particularly errors is important; `--server-response` prints out headers, and `--append-output` stores them to a log file. Most crawlers do not keep an error log around, but this is necessary to allow investigation of incompleteness and observe where errors in a crawl started (perhaps you missed blacklisting a page); for example, "Evaluating drug trafficking on the Tor Network: Silk Road 2, the sequel", Dolliver 2015, failed to log errors in their few HTTrack crawls of SR2, and so wound up with a grossly incomplete crawl which led to nonsense conclusions like 1-2% of SR2's sales were drugs. (I speculate the HTTrack crawl was stuck in the ebooks section, which was always clogged with spam, and then SR2 went down for an hour or two, leading to HTTrack's default behavior of quickly erroring out and finishing the crawl; but the lack of logging means we may never know what went wrong.)

7. once the wget crawl is done, then we name it whatever day it terminated on, we store the log inside the mirror, and clean up the probably-now-expired cookies, and perhaps check for any unusual problems.

This method will permit somewhere around 18 simultaneous crawls of different black-market sites or forums before you begin to risk Privoxy throwing errors about "too many connections". A Privoxy bug may also lead to huge logs being stored on each request. Between these two issues, I've found it helpful to have a daily cron job

reading `rm -rf /var/log/privoxy/*; /etc/init.d/privoxy restart` so as to keep the logfile mess under control and occasionally start a fresh Privoxy.

Crawls can be quickly checked by comparing the downloaded sizes to past downloads; markets typically do not grow or shrink more than 10% in a week, and forums' downloaded size should monotonically increase. (Incidentally, that implies that it's more important to archive markets than forums.) If the crawls are no longer working, one can check for problems:

- is your user-agent no longer in sync?
- does the crawl error out at a specific page?
- do the headers shown by wget match the headers you see in a regular browser using Live HTTP Headers?
- has the target URL been renamed?
- do the URLs in the blacklist match the URLs of the site, or did you log in at the right URL? (for example, a blacklist of "www.abraxas…onion" is different from "abraxas…onion"; and if you logged in at a onion with `www.` prefix, the cookie may be invalid on the prefix-free onion)
- did the server simply go down for a few hours while crawling? Then you can simply restart and merge the crawls.
- has your account been banned? If the signup process is particularly easy, it may be simplest to just register a fresh account each time.

Despite all this, not all markets can be crawled or present other difficulties:

- Blue Sky Market did something with HTTP headers which defeated all my attempts to crawl it; it rejected all my wget attempts at the first request, before anything even downloaded, but I was never able to figure out exactly how the wget HTTP headers differed in any respect from the (working) Firefox requests
- Mr Nice Guy 2 breaks the HTTP standard by returning all pages gzip-encoded, whether or not the client says it can accept gzip-encoded HTML; as it happens, wget cannot read gzip-encoded HTML and parse the page for additional URLs to download, and so mirroring breaks
- AlphaBay, during the DoS attacks of mid-2015, began doing something odd with its HTTP responses, which makes Polipo error out; one must browse AlphaBay after switching to Privoxy; Poseidon also did something similar for a time
- Middle Earth rate-limits crawls per session, limiting how much can be downloaded without investing a lot of time or in a CAPTCHA-breaking service
- Abraxas leads to peculiarly high RAM usage by wget, which can lead to the OOM killer ending the crawl prematurely

See also the comments on crawling in "Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem", Soska & Christin 2015.

Crawler wishlist

In retrospect, had I known I was going to be scraping so many sites for 3 years, I probably would have worked on writing a custom crawler. A custom crawler could have simplified the blacklist part and allowed some other

desirable features (in descending order of importance):

- CAPTCHA library: if CAPTCHAs could be solved automatically, then each crawl could be scheduled and run on its own.

  The downside is that one would need to occasionally manually check in to make sure that none of the possible problems mentioned previously have happened, since one wouldn't be getting the immediate of noticing a manual crawl finishing suspiciously quickly (eg a big site like SR2 or Evolution or Agora should take a single-threaded normal crawl at least a day and easily several days if images are downloaded as well; if a crawl finishes in a few hours, something went wrong).
- supporting parallel crawls using multiple accounts on a site
- optimized tree traversal: ideally one would download all category pages on a market first, to maximize information gain from initial crawls & allow estimates of completeness, and then either randomly sample items or prioritize items which are new/changed compared to previous crawls; this would be better than generic crawlers' defaults of depth or breadth-first
- removing initial hops in connecting to the hidden service, speeding it up and reducing latency (does not seem to be a config option in Tor daemon but I'm told something like this is done in Tor2web)
- post-download checks: a market may not visibly error out but start returning login pages or warnings. If these could be detected, the custom crawler could log back in (particularly with CAPTCHA-solving) or at least alert the user to the problem so they can decide whether to log back in, create a new account, slow down crawling, split over multiple accounts, etc

Other datasets

A number of other datasets are known to exist but are unavailable, including:

- law enforcement scrapes (see the Force briefing), seized server images