Learning Darknet Markets

Luis Armona Daniel Stackman

New York Federal Reserve Bank

December 15th, 2014

Abstract

We construct an original dataset by regularly scraping narcotics listings from a Darknet search engine. Due to their illicit nature, there is a paucity of data on illegal drug markets; consequently, little is known about the distributions of price, volume, and other important characteristics of these marketplaces. After hand-labeling a training set of 8,902 listings, we train several classes of SVMs on a variety of listing characteristics, to distinguish genuine drug listings from the rest of the results returned by our keyword searches. In our sample, a spectrum kernel defined over the listing names, combined with a linear kernel over the non-text features, yields optimal performance, classifying 96.3% of the data accurately. Applying predictions generated by this SVM to our entire dataset, we extract and analyze over 87,000 marijuana listings from the raw data.

1 Introduction

1.1 A Brief History of Darknet Markets

Silk Road was an online market place for drugs, weapons, counterfeit, and other illicit goods, established in 2011. To our knowledge, it was the first of its kind: accessible solely through the anonymous Tor web browser, with all transactions carried out in Bitcoin. A June 2011 profile of the website by Gawker [2] first brought the website into the public eye. From late 2011 until mid 2012, an estimated $1.2 million of business was conducted over Silk Road each month [3]. On October 2, 2013, the website was seized and shut down by the FBI, Europol, and Interpol [1], and its founder indicted. Since then, dozens of similar websites have sprung up, including Silk Road 2.0, which was seized and shut down on November 6, 2014 [5, 14].

1.2 Grams: Google for the Darknet

Grams is a search engine for the Darknet, self-consciously styled after the regular net's most popular search engine, Google [15]. According to its anonymous creator, Grams uses a ranking algorithm similar to Google's to allow users to search the Darknet, returning results from multiple Darknet sites [17]. Over our sample, Grams returned results from nine different Darknet markets: Silk Road 2.0, Agora, and Evolution were the three largest, with 12,524, 31,007, and 55,097 (marijauana) listings respectively.

1.3 Data description

We built a web crawler that feeds Grams a list of search terms, and scrapes the results it returns. Over a period of several months, we scraped Grams once every two or three days, storing hundreds of thousands of listings. In this project, we analyze listings returned by the search terms "cannabis", "marijuana", and "weed". Some fraction of the 118,513 listings these search terms collectively returned are not actually listings for marijuana: our raw data contains listings for heroin, methamphetamine, and cocaine, as well as various drug paraphernalia such as grinders and ziplock bags. Additionally, we chose to exclude listings for hashish, THC concentrate liquids, and marijauna-infused edibles from our definition of "marijuana". After collecting our data, and deciding on our exclusion criterion, we labeled a randomly chosen subset of 8,902 examples (about a 7.5% sample) by hand. The data scraped for each listing includes name, description, price, vendor, market, and location. We extracted weight from the text listing by searching for number-unit pairs within the text. Weights were successfully extracted from over 80% of the training data, and over 90% of the unlabeled data. The price and weight variables, were normalized to have mean 0 and standard deviation 1. Vendor, location, and market data, were each mapped to 0-1 indicator variables for each unique vendor/location/marketplace. For example, the location variable for "United States" will be 0 for listings that are not from the United States, and 1 for listings that are. Table 1 below provides summary statistics for the main variables used in our analysis. We trained a variety of SVMs on our labeled data, using LIBSVM [12] and the Python computing language (in particular the PyML library, built on LIBSVM). All code used in this project, can be accessed at `https://www.dropbox.com/sh/ps9sgdzl32rsix8/AABiuaKATXAS9jtIAVabvUH6a?dl=0`.

Table 1: Summary Statistics of Data

Statistics	Training Sample	Unlabeled Sample
price-mean (BTC)	1.60	2.16
price-median (BTC)	0.23	0.23
price-s.d. (BTC)	28.98	47.12
weight (g) -mean	227.65	178
weight (g) -median	7.00	10
weight (g) -s.d.	14456.52	10000.3
character length of listings-mean	34.60	37.75
character length of listings-median	36.00	37.00
character length of listings-s.d.	16.66	16.59
number of observations	8902	118513

1.4 Previous Literature

To our surprise, a small literature on darknet marketplaces already exits. Christin [3] collected data from the original Silk Road from late 2011 until mid 2012. He tracks the growth of the site over his sample, noting that it was almost exclusively a market for narcotics. Lau [11] performs a similar analysis of Silk Road 2.0. We extend the literature by scraping Grams, which searches the whole Darknet for drug listings. In contrast to Christin and Lau, who examine 24,000 and 3,585 listings, respectively, we have at our disposal over 118,000 search results for marijuana alone. The wealth of data at our disposal allows us to better characterize the entire Darknet drug economy, and necessitates the use of best-in-class machine learning algorithms in classifying our listings.

2 String kernels

Ever since the seminal paper of Cortes and Vapnik [4], support vector machines have occupied a central role in the classification literature. As a reminder, the SVM optimization problem can be stated mathematically as

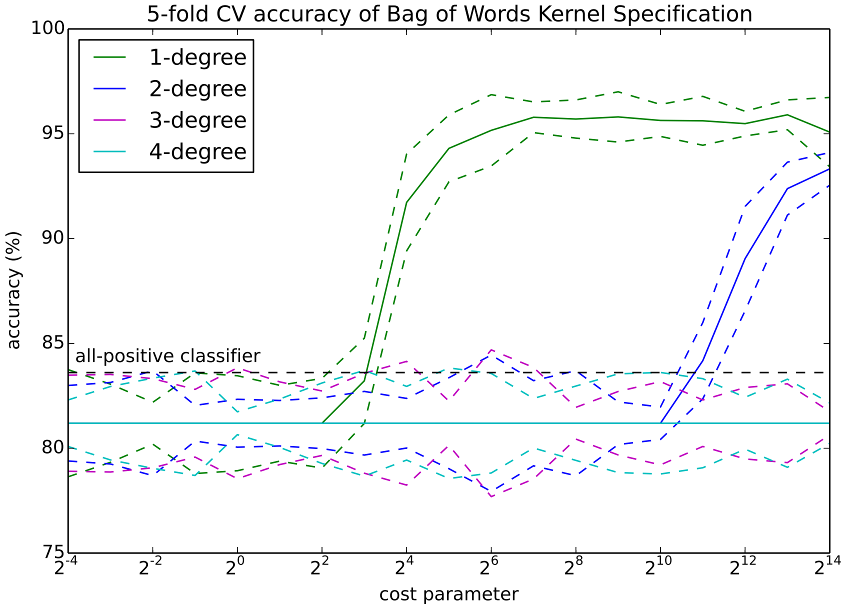
$$\max_{\beta} \sum_{i=1}^{m} \beta_i - \sum_{i=1}^{m} \sum_{i=1}^{m} \beta_i \beta_j y_i y_j K(x_i, x_j),$$

where x_i are the data, y_i are the labels, β_i are the slack variables, and K is a positive definite symmetric kernel.

SVMs were first extended to the problem of text classification by Joachims [8]. In Joachims, the text is transformed into a list of individual word-stems (e.g., “runner” and “running” would both map to “run”), commonly used “stop-words” (e.g., “the”) are removed, and then the counts of each word-stem in the document are represented as vectors. At that point, it is straightforward to train any number of polynomial or gaussian radial basis function kernel SVMs, just as one would with numeric data. This type of model is called “Bag of Words”, since any information contained in the ordering of the words, or their proximity to one another, is lost in the mapping from the text to the count vectors: it is the first model we apply to our data.

2.1 Bag of Words

“Bag of Words” refers to a general class of models, including k-means, nearest neighbor, and naive Bayes, as well as SVMs, that represent the frequencies of individual words (or word-stems) in a text as count vectors. Our version of Bag of Words is not a string kernel, per say; rather, it is a mapping from text to a vector space, followed by the application of a more standard kernel (e.g. polynomial, gaussian radial, etc.). In our preferred specification, all non-alphabetical characters are removed from the text, along with any extraneous white-space, and then the listing names and descriptions are tokenized into individual words. These words are collected into a dictionary, and the frequencies of each word in a particular listing are stored in a vector. These word-count vectors are combined with weight, price, and indicator vectors for vendor, darknet market, and seller location, to form a feature matrix. Feature matrix in hand, we try our a variety of polynomial kernels, varying cost by powers of two, and polynomial degree.



As is clear from the figure above, the linear kernel dramatically outperforms its competitors, achieving a maximum accuracy of just under 96%. The quadratic kernel also performs respectably, but at much higher cost. Interestingly, the cubic and quartic polynomials both perform abysmally, failing even to outperform a classifier that labels every data point positively. Gaussian kernels (results not shown) also underperform, with much lower accuracies and larger standard deviation bands than either the linear or quadratic specifications. While Bag of Words performed quite well on our data, we also wanted to explore approaches that took advantage of more of the information contained in our listings. To do that, we needed to expand our horizons to include more exotic kernels.

2.2 Spectrum Kernel

The spectrum kernel, proposed originally by Leslie et al. in 2002 [9] for classifying sequences of proteins, is a sequence kernel that bases its measure of similarity between two sequences on the number of k-mers that both sequences share (k-mers are all possible subsequences of length k within a sequence). Precisely, we consider a string of length l to be a sequence $\mathbf{x} = x_1, x_2, ..., x_l$ where each x_i belongs to the set Σ, known in the literature as the alphabet of the sequence. In our dataset, the alphabet for string sequences is simply the 256-character set of ASCII characters. For a subsequence/substring to be counted as contained within a string, the spectrum kernel requires it appears in the string contiguously, i.e. that no elements of the sequence are skipped in order to locate subsequence. For example, the substring CAT would be counted as appearing in the string $CATEGORY$ but not in $CART$. The kernel can be represented as:

$$K_k(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{u} \in \Sigma^k} \#_{\mathbf{u} \in \mathbf{x}} \#_{\mathbf{u} \in \mathbf{y}}$$

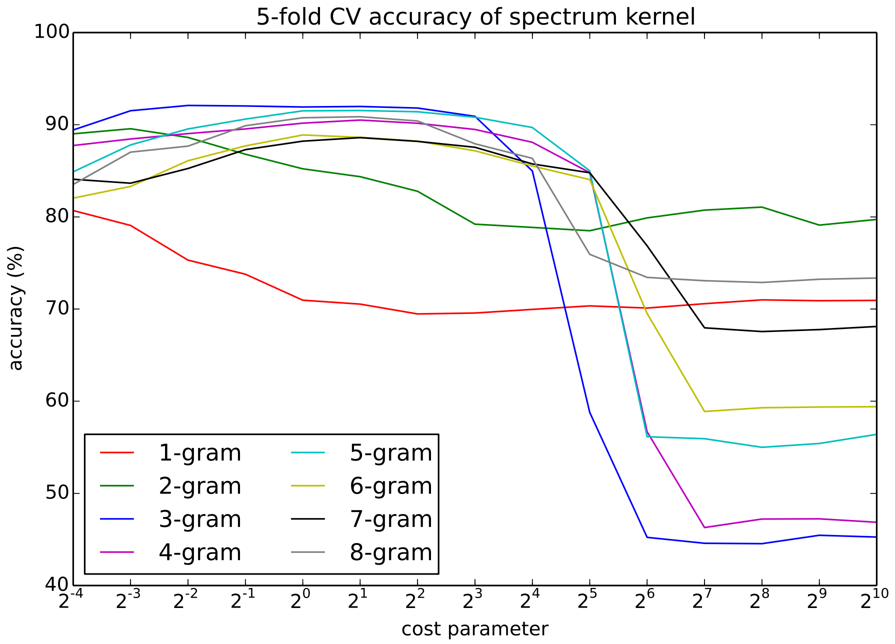
where $\#_{\mathbf{u} \in \mathbf{s}}$ is the number of times the subsequence $\mathbf{u}$ appears in the sequence $\mathbf{s}$, and k is the length of subsequences considered. This can be equivalently expressed as a dot product $K_k(\mathbf{x}, \mathbf{y}) = \psi_k(\mathbf{x}) \cdot \psi_k(\mathbf{y})$, where $\psi_k(.) : \Sigma^k \rightarrow \mathbb{N}^{|\Sigma|^k}$ is the feature mapping of a sequence to the frequencies of all possible k-mers within it. Because this kernel admits a representation from the dot product of a feature mapping, it is subsequently Positive Definite Symmetric (PDS) and therefore can be used in conjunction with SVMs in place of a traditional dot product.

2.3 Implementation of Spectrum Kernel

In order to apply the spectrum kernel to our classification problem, we first preprocess the strings to limit confounding of the spectrum kernel's similarity measure. In particular, we remove 174 stop-words (provided by R's `tm` package [6]; stop-words, e.g., "the", contain little information about the similarity of two texts), convert all alphabet letters to lowercase (so that substrings differing only by uppercase and lowercase instances of the same letter are treated equivalently), and replace all consecutive occurrences of whitespace with a single space ' ' (excess whitespace is not useful information for our classification problem).

Having prepared our text listings, we use SVM classification with a spectrum kernel to classify the data. To pick the optimal hyperparameters k (the length of the substrings searched for) and C (the cost parameter in SVM classification), we perform 5-fold cross validation on the hand-labeled dataset, ranging the cost parameter from 2^{-4} to 2^{10} in powers of 2, and we consider n-grams from $n = 1$ to $n = 8$. Mean 5-fold cross validation accuracies for the different specifications are reported in the figure below.

In general, we find the spectrum kernel performs better at low cost, and in all cases performs poorly for cost greater than 32. The maximal CV accuracy achieved is 92.1%, at cost C=0.5 and substring length $k = 3$. Overall, the 3-gram specification performs the best, achieving the highest accuracy at every cost from 2^{-4} to 2^3 . The 5-gram specification performs comparably to the 3-gram, and even exceeds it for intermediate cost choices.



2.4 Combining Kernels

While the spectrum kernel performs well, it only takes advantage of information from the text description of the listing, neglecting our information on the vendor, the darknet market where it was listed, the vendor's geographic location, and the listing price. We would like to make use of all the information at our disposal, and for this purpose also use a linear combination of multiple kernels to classify our data. In particular, we explore linear combinations of a sequence kernel, used to calculate similarities between our text data, and a second kernel $K_p(\cdot)$, that classifies on the basis of our real-valued features. The combined kernel we consider has the following form:

$$\begin{aligned} K_c(\mathbf{x}, \mathbf{y}) = & \quad \alpha K'_{k^*}(\mathbf{x}, \mathbf{y}) + (1-\alpha) K'_{p^*}(\mathbf{x}, \mathbf{y}) \\ = & \quad \alpha \frac{K_{k^*}(\mathbf{x},\mathbf{y})}{K_{k^*}(\mathbf{x},\mathbf{x}) K_{k^*}(\mathbf{y},\mathbf{y})} + (1-\alpha) \frac{K_{p^*}(\mathbf{x},\mathbf{y})}{K_{p^*}(\mathbf{x},\mathbf{x}) K_{p^*}(\mathbf{y},\mathbf{y})}, \end{aligned}$$

where $K'(\cdot)$ is the normalized kernel, (i.e. $K'(\mathbf{x}, \mathbf{y}) = \frac{K(\mathbf{x},\mathbf{y})}{K(\mathbf{x},\mathbf{x}) K(\mathbf{y},\mathbf{y})}$), and $\alpha \in [0, 1]$ is the weight on the first kernel. Normalization of the individual kernels before combining is necessary because otherwise one may be in a completely different scale of the other, rendering the two incomparable. So long as both input kernels are themselves PDS, we know that our combination kernel will be PDS because the PDS property is closed under addition, and therefore it may be used with SVMs to classify our data (see Theorem 5.3 in Mohri et al [13]) To find the optimal combined kernel, we first separately find the optimal kernels K_{n^*} and K_{p^*} that perform best when considered individually, via cross validation on the labeled data. Next, fixing the optimal hyperparameters for the input kernels, we perform grid search on the choice of α and C in order to solve the optimization problem:

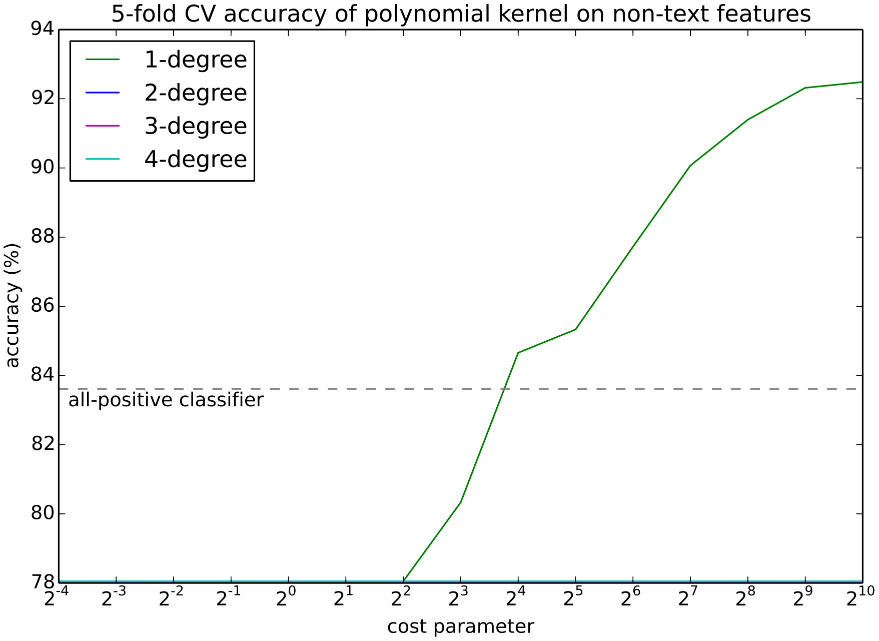
$$\min_{\alpha \in [0,1], C \geq 0} \widehat{R}_C(K_c(\mathbf{x}, \mathbf{y}))$$

where $\widehat{R}_C$ is the empirical error of a two-class SVM with cost parameter C using a kernel $K_c(.)$. This roughly follows the two-step training method proposed in Gönen et al. (2011) [7] for combining kernels.

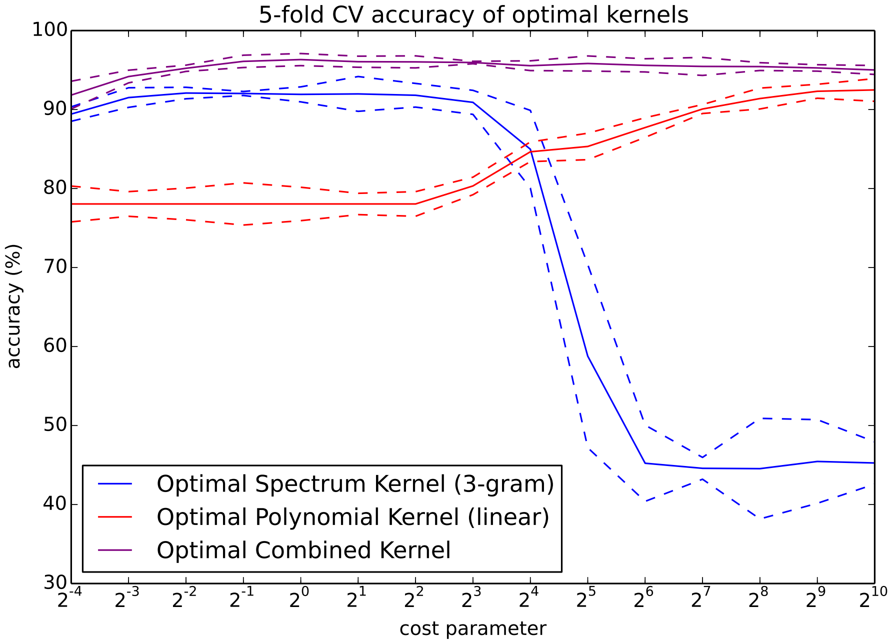
2.5 Implementation of Combined Kernels

For our string data, we use the spectrum kernel of 3-grams, which we found to be optimal among all spectrum kernel specifications considered, as explained in Section 2.3. For our real-valued data, we consider the families of both Gaussian and polynomial kernels. We found polynomials significantly outperform Gaussian kernels, as was the case with Bag of Words results discussed in Section 2.1. Below, we plot the 5-fold CV accuracies of polynomial kernels of degree 1 to degree 4, defined over the real-valued data, as a function of cost.

Considered in isolation, the polynomial kernels are all inferior to the optimal spectrum kernel, particularly at low cost, where, with accuracies of around 78%, they perform even worse than a simple rule of classifying all the results as positive (about 84% of our hand-labeled listings are actually listings for marijuana). As we increase the cost parameter, the linear kernel (simple dot product) begins to improve its performance, eventually yielding accuracy up to 92%. It is not obvious to us why higher-order polynomials perform so poorly on our data. One possible explanation is that the majority of our real-valued features are 0-1 indicators (for vendor, market, and location), and are therefore unchanged when raised to higher powers. Furthermore, we speculate the the information contained in price, vendor, market, and location are relatively orthogonal, leaving little room for the importance of interactions between different feature vectors. Whatever the cause, the linear kernel's superior performance dictates that we choose it to represent the non-text features, in combination with the spectrum kernel.



Because both the spectrum kernel and the linear kernel are PDS, the resulting combined kernel is PDS, and may be used in conjunction with SVMs, substituting for the dot product, as mentioned previously. After performing cross-validation on a variety of weights, we find that the optimal weight choice is $\alpha^* = 0.8$, which places a much heavier weight on the spectrum kernel than the linear kernel. The best accuracy achieved is 96.3%, at a cost of $C^* = 1$, which substantially improves on the performance of both constituent kernels, considered individually. A comparison of the optimal polynomial kernel, the optimal spectrum kernel, and the optimal combined kernel, is presented below.



Clearly, using of all the information at our disposal substantially improves the performance of the SVM.

3 Results

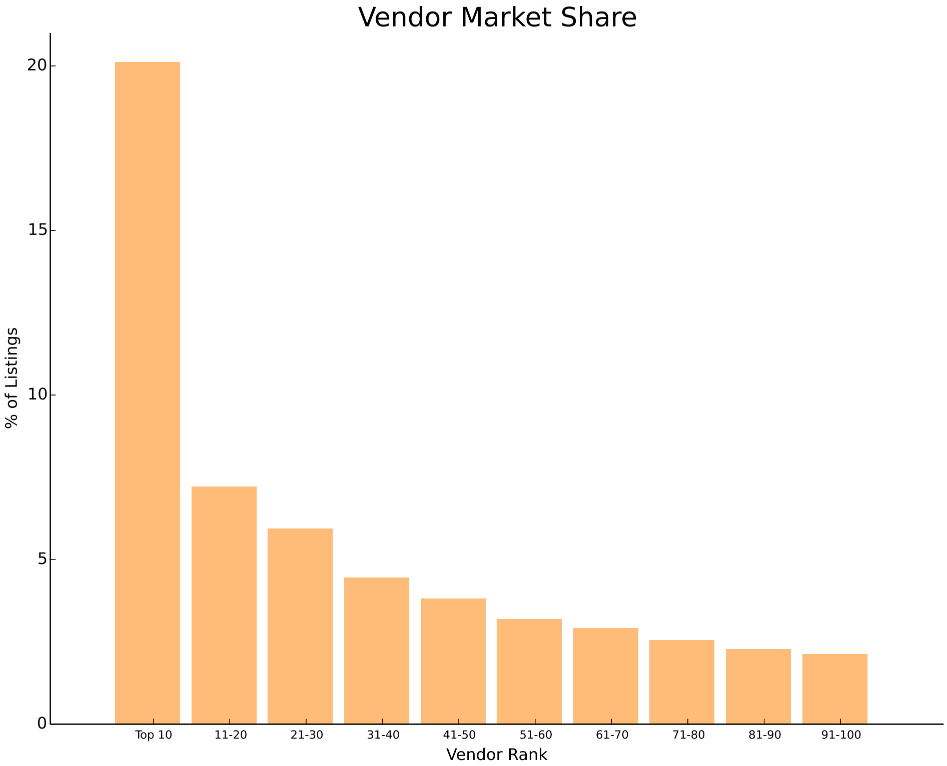
3.1 Discussion of Results

Both the bag of words specification, and the spectrum kernel, when combined with the other features of the listings, achieved very respectable 5-fold CV accuracies. While the bag of words representation was adept at inferring classification status from "buzzwords" contained in listings, the spectrum kernel, which could identify stems *within* words, and could consider sequences that followed one another, interrupted by whitespace, had a much greater scope to identify similarity, which possibly accounts for its improved accuracy rate.

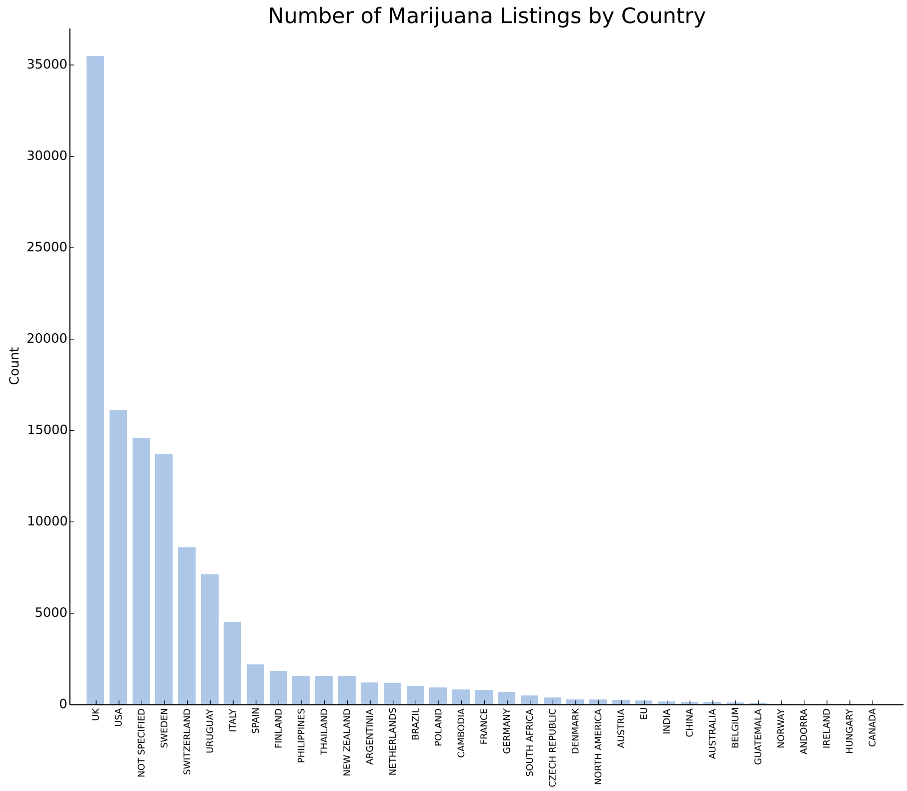
When choosing between specifications, in addition to cross-validation error, we also considered computational costs and false positive rates. Using bag of words requires populating an $L \times W$ matrix, where L is the number of listings in our sample, and W is the number of distinct words in our training data. In this case, with $L = 118513$ in our full sample, and $W > 4000$, this computation, while only a few lines of code, was very time-intensive. In contrast, once the spectrum kernel has been optimized on the training data, it can be applied to the full, unlabeled sample without any additional computations. Our concern with false positives arises because false positives would be much more likely to skew our analysis of the market than false negatives. Misclassifying a few non-marijuana listings as marijuana could dramatically affect quantities of interest, such as the mean and standard deviation of prices. In contrast, a few false negatives merely decreases our sample size, which is not a major concern given the number of raw listings we collect. The optimal combined kernel performed well with respect to this metric, with a positive predicted value (PPV) of 98%.

3.2 The online market for marijuana

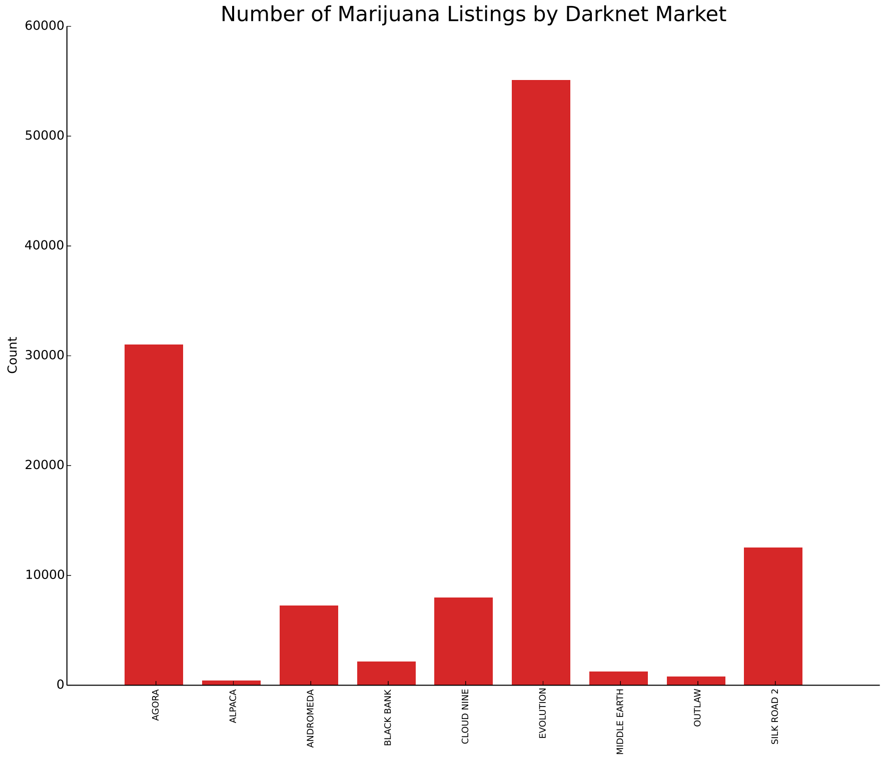
Having successfully predicted the labels of over 96% of our unlabeled sample, we can now give a detailed characterization of the Darknet market for marijuana. The most striking feature of the market is its concentration, which manifests itself in several different ways. The most obvious measure of concentration is the market share (defined as number of listings by one vendor over the total number of listings) of individual vendors, displayed below.



The top 10 vendors (by number of listings) alone account for 20% of the total listing volume. Cumulatively, the top 100, who in number represent only 13% of the 775 vendors, post more than half of the listings, while the bottom 500 post less than 20%. This pattern also shows up if we look at the geographic location of the vendors:



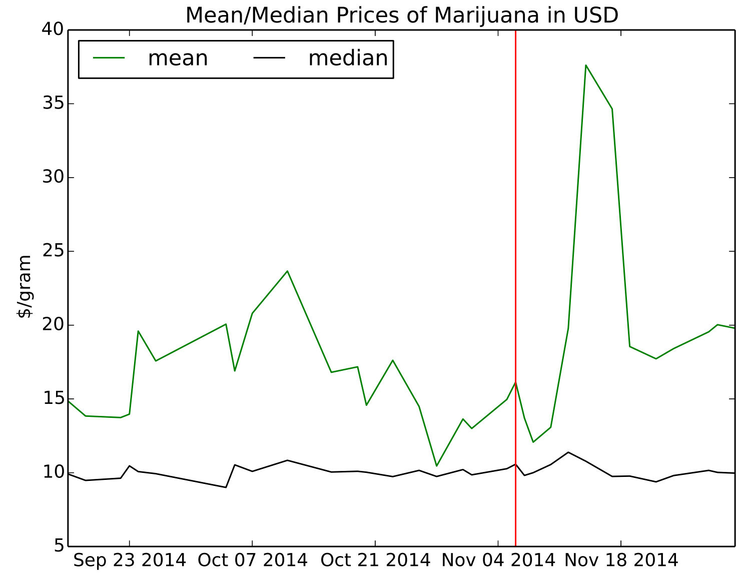
The UK dominates our sample, with more than twice as many listings as the second place country, the US. The distribution of listings between markets, picture below, is also extremely unequal. Despite its high profile and first-mover status, Silk Road 2 is only the third largest darknet market for marijuana, lagging far behind Agora and Evolution.

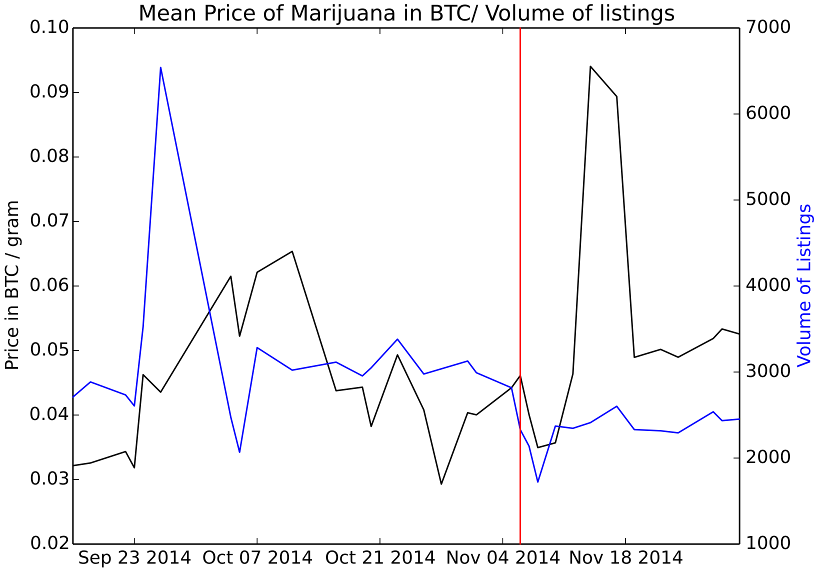


3.3 Prices and Volume

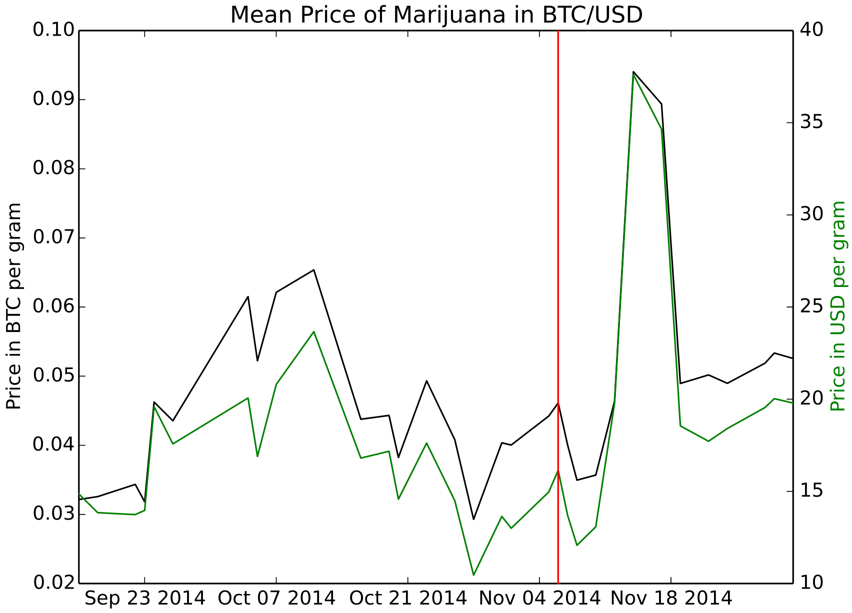
Marijuana prices are fairly stable in our sample, with mean price varying between \$10 and \$25 per gram, and median price nearly constant at \$10 a gram. We suspect that the higher level and volatility of the mean versus the median, which also shows up in our summary table above, is a result of a few very large listings. The notable exception to this pattern of relative stability occurs in the days after the shutdown of Silk Road 2 (red line), when prices skyrocket.

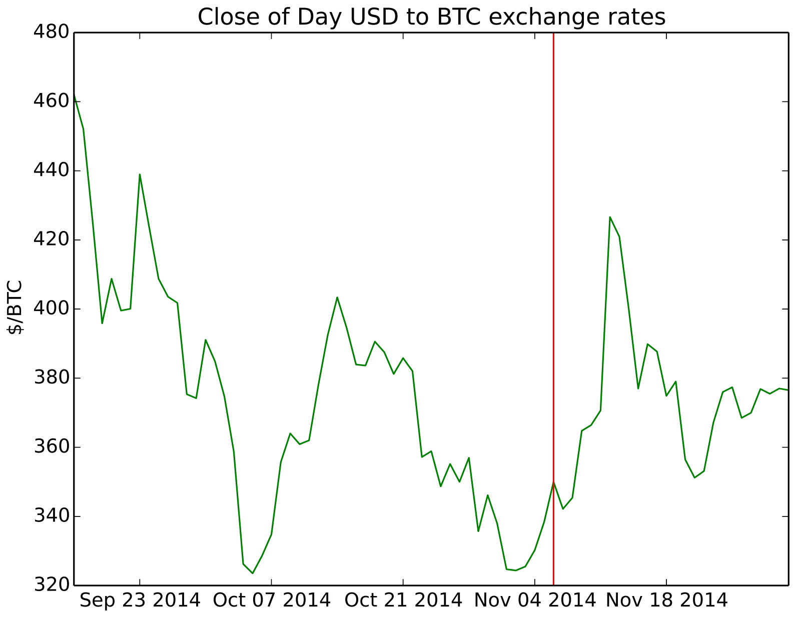
A plot of volume (measured as number of listings per day) and prices shows that volume dips noticeably after the shutdown of Silk Road 2, but quickly recovers. This makes sense given the size of Silk Road relative to other Darknet markets (see the bar chart above). However, it is a bit surprising that there is no obvious relationship in the data between volume and price. In particular, the decline in volume following the shutdown cannot explain the subsequent price jump, which appears to have occurred several days later.


Mean/Median Prices of Marijuana in USD
mean
median
$/gram
40
35
30
25
20
15
10
5
Sep 23 2014
Oct 07 2014
Oct 21 2014
Nov 04 2014
Nov 18 2014


Mean Price of Marijuana in BTC/ Volume of listings
Price in BTC / gram
Volume of Listings
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
7000
6000
5000
4000
3000
2000
1000
Sep 23 2014
Oct 07 2014
Oct 21 2014
Nov 04 2014
Nov 18 2014

There are several possible explanations for this result. The simplest is that the Grams listings were several days old, had stale prices, and only reflected the shock to the market a few days after the shutdown occurred. But even if this were true, the relative stability of volume that coincides with the sharp increase in price makes this explanation incomplete at best. Another possibility is that there was a significant reduction in the supply of Bitcoin when Silk Road 2 was seized. Given that the arrest of the original Silk Road's founder led to the largest Bitcoin seizure in history [1], it is not too much to suppose that a large number of Bitcoins were seized in this episode as well. The dramatic increase in marijuana prices may been driven by spillovers from the Bitcoin market, where a suddenly decreased supply led to an increase in price, rather than by a decrease in the supply of marijuana.





The plots above show that the sharp increase in prices coincided with a similarly sharp appreciation in Bitcoin's price. However, Bitcoin has experienced many episodes of volatility over its short history, so we should be cautious in our interpretation of this particular spike.

4 Conclusion

In conclusion, we apply SVMs to the problem of binary classification on an original dataset of darknet drug listings. After experimenting with a variety of kernels, we settle on a combined linear and spectrum kernel, which performs optimally on our training data in terms of 5-fold cross valuation accuracy, PPV rate, and computational feasibility. After settling on our classifier, we predict labels for our entire dataset of 118,513 marijuana listings, and analyze 87,000 positively-labeled listings to characterize the nature and behavior of the online market for marijuana. We look forward to conducting further research, delving deeper in the the nature of this, and other, illicit online economies.

References

[1] Ball, James, Charles Arthur, and Adam Gabbatt. “FBI Claims Largest Bitcoin Seizure after Arrest of Alleged Silk Road Founder.” The Guardian. N.p., 2 Oct. 2013.

[2] Chen, Adrian. “The Underground Website Where You Can Buy Any Drug Imaginable.” Gawker. Gawker Media, 1 June 2011. Web. 12 Dec. 2014. http://gawker.com/the-underground-website-where-you-can-buy-any-drug-imag.

[3] Christin, Nicolas. *Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace*. Retrieved from: http://arxiv.org/abs/1207.7139v2.

[4] Cortes, Corinna, and Vladimir Vapnik. “Support-vector Networks.” *Machine Learning* 20.3 (1995): 273-97.

[5] Cox, Joseph. “Silk Road 2.0 Was Just Shut Down by the FBI.” Motherboard. VICE, 6 Nov. 2014. Web. 13 Dec. 2014. http://motherboard.vice.com/read/silk-road-2-has-been-seized-by-the-fbi.

[6] Feinerer, Ingo and Kurt Hornik (2013). tm: Text Mining Package. R package version 0.5-8.3. http://CRAN.R-project.org/package=tm

[7] Gönen, Mehmet, and Ethem Alpaydn.“Multiple Kernel Learning Algorithms.” *Journal of Machine Learning Research* 12 (2011): 2211-268.

[8] Joachims, Thorstein. “Text Categorization with Support Vector Machines: Learning with Many Relevant Features”. *LS8-Report* 23, Universitt Dortmund, LS VIII-Report, 1997.

[9] Leslie, Christina S., Eleazar Eskin, and William Stafford Noble. ”The spectrum kernel: A string kernel for SVM protein classification.” Pacific symposium on biocomputing. Vol. 7. 2002.

[10] Lodhi, Huma, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. “Text Classification Using String Kernels.” *Journal of Machine Learning Research* 2 (2002): 419-44.

[11] Lau, Daryl. *Analyzing Trends in Silk Road 2.0*. Retrieved from: http://lau.im/articles/analyzing-silk-road-2-0-part-1/.

[12] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

[13] Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of machine learning. MIT press, 2012.

[14] Mullin, Joe. "Silk Road 2.0, Infiltrated from the Start, Sold $8M per Month in Drugs." Arstechnica. N.p., 7 Nov. 2014. Web. 13 Dec. 2014. http://arstechnica.com/tech-policy/2014/11/silk-road-2-0-infiltrated-from-the-start-sold-8m-per-month/-in-drugs.

[15] Neal, Meghan. "I Used the Dark Net's First Search Engine to Look for Drugs." Motherboard. VICE, 17 Apr. 2014. Web. 13 Dec. 2014. http://motherboard.vice.com/en_uk/read/i-let-grams-guide-me-through-the-dark-nets-illegal-bazaars.

[16] Sonnenburg, Soren, Gunnar Ratsch, and Konrad Rieck. "Large Scale Learning with String Kernels." N.p., 2007. Web. 13 Dec. 2014. Retrieved from: http://sonnenburgs.de/soeren/publications/SonRaeRie07.

[17] Zetter, Kim. "New 'Google' for the Dark Web Makes Buying Dope and Guns Easy." Wired.com. Conde Nast Digital, 17 Apr. 2014. Web. 13 Dec. 2014. http://www.wired.com/2014/04/grams-search-engine-dark-web/.