Chapter 2

AUTOMATED ANALYSIS OF UNDERGROUND MARKETPLACES

Aleksandar Hudic, Katharina Krombholz, Thomas Otterbein, Christian Platzer and Edgar Weippl

Abstract Cyber criminals congregate and operate in crowded online underground marketplaces. Because forensic investigators lack efficient and reliable tools, they are forced to analyze the marketplace channels manually to locate criminals – a complex, time-consuming and expensive task. This paper demonstrates how machine learning algorithms can be used to automatically determine if a communication channel is used as an underground marketplace. Experimental results demonstrate that the classification system, which uses features related to the cyber crime domain, correctly classifies 51.3 million messages. The automation can significantly reduce the manual effort and the costs involved in investigating online underground marketplaces.

Keywords: Underground marketplaces, automated analysis, machine learning

1. Introduction

Cyber criminals routinely use online underground marketplaces to communicate and trade stolen or illegal goods and services. Typically, publicly-accessible chatrooms and web forums are used as marketplaces by criminals who openly hawk their goods and initiate contractual agreements. Recent research has shown that underground marketplaces are a significant security risk because they provide venues for buying and selling stolen credentials, credit card numbers and other sensitive data [6]. Detecting these marketplaces and investigating the criminal activities being conducted are tedious and time-consuming tasks. Automating this process could significantly enhance the ability of forensic analysts to investigate criminal activities conducted in underground marketplaces. Unfortunately, the large number of online marketplaces and their ad hoc nature and volatility prevent naive detection approaches such as web crawling systems from being effective. Furthermore, criminals often "hijack" benign websites (e.g., websites containing classified ads and abandoned forums) instead of using dedicated underground websites.

This paper demonstrates how machine learning can be used to automatically detect underground marketplaces. An experimental evaluation is presented based on eleven months of real-world Internet Relay Chat (IRC) and web forum communications. The results show that the classification system can successfully locate and monitor communication channels used by cyber criminals, significantly reducing the manual effort and the costs involved in investigating online underground marketplaces.

2. Background

While any type of communication channel could be used as an underground marketplace, the two most common types of channels are IRC chatrooms and web forums. IRC chatrooms and web forums are very popular communication channels and have multitudes of legitimate users.

A large number of publicly-accessible IRC networks are accessible over the Internet (e.g., QuakeNet, IRCnet, Undernet, EFnet, Rizon, Ustream and IRC-Hispano). In most cases, they do not have user authentication mechanisms, which unfortunately means that there is no straightforward means of attribution. Cyber criminals exploit IRC networks to advertise their goods and services. While some IRC networks appear to be specifically designated for criminal activities, benign networks are abused as well. The organizers simply create channels with names that are known to insiders. For example, channel names with the prefix `#cc` (short for credit card) are often used by criminals involved in credit card fraud.

Cyber criminals also operate underground marketplaces on websites that contain forums and message boards. The forums organize individual messages (i.e., posts) in the form of "threads" (i.e., lists of messages belonging to the same topic). Unlike IRC networks, the contents of these forums are persistent and users can communicate in a more organized manner, e.g., by replying to specific posts or to groups of users. Forums generally have stricter admission procedures than IRC networks (e.g., users have to sign-up to receive login credentials). Also, they offer "convenience" services to their members such as escrow and private messaging.

3. Related Work

Research related to underground marketplaces has focused on the evaluation of message content [9] and the acquisition of reliable data from underground marketplaces [6, 16].

Franklin, *et al.* [3] conducted a systematic study of IRC channels exploited as underground marketplaces. They evaluated the content using machine learning techniques and demonstrated that underground marketplaces have considerable security implications; they also presented approaches for disrupting underground marketplaces. A Symantec report [13] on the underground economy provides useful analysis based on a significant amount of data collected from IRC networks and web forums over a period of one year; however, detailed information is not provided about the methodologies used to collect and analyze the data. Thomas and Martin [14] have studied the structure and players of the underground economy by examining IRC-based marketplaces. They describe the infrastructure established by criminals along with their activities, alliances and advertising methods.

Zhuge, *et al.* [16] have presented an overview of the underground market and malicious activities on Chinese websites based on a black market bulletin board and an online business platform. Holz, *et al.* [6] have studied "dropzones" that trade stolen digital credentials; they also evaluated a method that enables the automated analysis of impersonation attacks.

In contrast, Herley and Florencio [5] argue that marketplaces such as IRC channels and web forums do not have a significant impact. Instead, they describe them as standard markets for lemons where the goods are hard to monetize and the only people who derive benefits from the markets are the rippers.

Fallmann, *et al.* [2] have presented a novel system for automatically monitoring IRC channels and web forums. Furthermore, they extracted information and performed an experimental evaluation of the monitored environments.

4. Locating Underground Marketplaces

Finding underground marketplaces is a manual task that is complex and time-consuming. We present a novel classification system for automatically discovering and monitoring underground marketplaces, even when they are hidden in benign information channels.

Figure 1 presents an overview of the training process. The learning algorithm of the classifier approximates the optimal function $f : D \rightarrow C$ that maps document vectors $d \in D$ to a specific class $c \in C$ based on the training set, where class c is either benign or criminal.

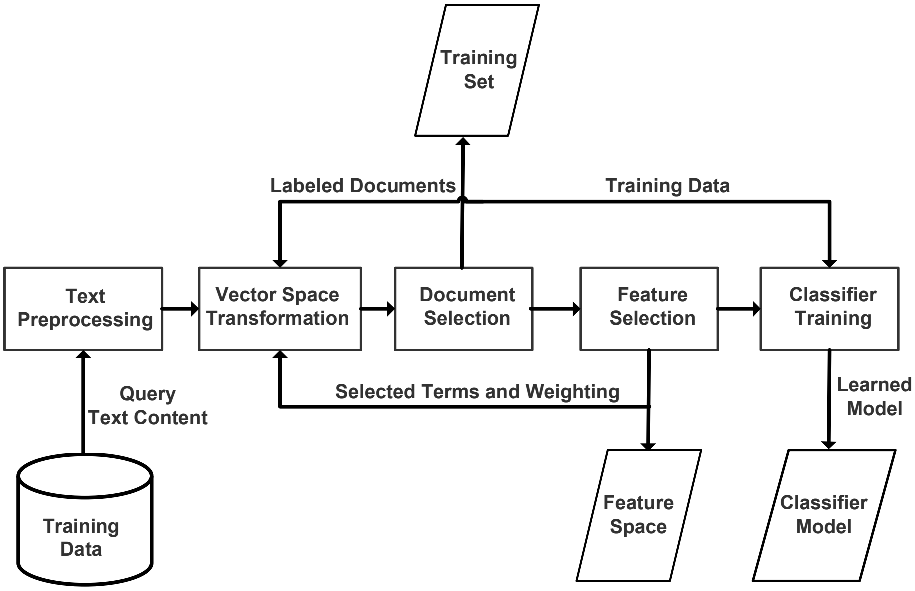


Figure 1. Training process.

In the classification process, a "document" is assumed to be an IRC chatroom or a web forum (thread), and "terms" are the words in IRC messages or web forum posts. The terms are mapped from each document to a numeric vector via the bag of words (BoW) model [4]. This model is agnostic to the exact ordering of terms within a document and interprets the terms as a set for each document. The resulting vector space model allows different weightings of the frequencies of individual terms.

4.1 Text Preprocessing

The raw training data contained noise and content that was not relevant to classification. The first step involved the extraction of the plain-text content. During this step, HTML elements and specific character encodings were eliminated.

4.2 Vector Space Transformation

The second step involved the generation of the vector space using tokenization [8]. Tokenization separates chunks of text with specific semantic values. A word-based model was employed because it has been shown to have the best performance [1, 11].

The next step in vector space transformation was to tag semantically-meaningful units that carry domain-relevant information. Various labels were attached to uniform resource identifiers (URIs), domain names, IP addresses, e-mail addresses, and numbers and dates to help identify the content. The tagging process helps reduce the feature space, e.g., by substituting frequently-changing values such as dates with a single date label.

The terms in the vector space model were weighted using tf-idf (term frequency – inverse document frequency) [10]. The term frequency $tf_{t,d}$ is the frequency of term t in document d and the inverse document frequency idf_t indicates the importance of term t to the document corpus. The tf-idf weighting scheme reduces the impact of common words (i.e., words that appear with high frequencies in the document). Since the documents had different lengths, the vectors were normalized using cosine normalization [12].

4.3 Document Selection

The third step involved the selection of appropriate documents according to their relevance. The document selection was based on hierarchical agglomerative clustering (HAC), a commonly-used deterministic bottom-up clustering algorithm that does not require a pre-specified number of clusters as input. HAC merges documents with the highest similarity into a cluster. The similarity of a merged cluster is called the combination similarity. Our HAC prototype implementation supports single-link and complete-link clustering. Single-link clustering defines the combination similarity in terms of the most similar members; the merge criterion is, therefore, local in nature. Complete-link clustering, on the other hand, defines the similarity of two merged clusters in terms of the similarity of the most dissimilar members and merges clusters using a non-local criterion. The algorithm merges documents into clusters until a predefined cutoff similarity value is reached.

The document selection process currently supports two methods for choosing a representative for each cluster. The first selects the document that represents the centroid based on the Euclidean distance, while the second is based on a definable score function.

Similarity. The cosine similarity measure is commonly used to compute the similarity between two documents in vector space:

$$sim(d_1, d_2) = \cos\theta = \frac{\vec{V}(d_1) * \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|}.$$

Cosine similarity measures the similarity of the relative distribution of the terms by finding the cosine of the angle between the two document vectors $\vec{V}(d_1)$ and $\vec{V}(d_2)$. The cosine of the angle θ between the two document vectors ranges from zero to one, where zero indicates that the documents are independent and one means that the two document vectors are identical.

Distance. The Euclidean distance is another measure for comparing two vectors. This distance is more appropriate when the lengths of the documents are considered. For example, the Euclidean distance measure can be used to compute the nearest neighbors or, in our case, to determine the centroid of the cluster during the document selection process.

4.4 Feature Selection

Feature selection involves the selection of a subset of terms from the training set that is used for the vector space model. This process decreases the cardinality of the vector space and reduces the computation time.

In our case, features that occurred less than three times in the training set of the document corpus were removed (as proposed by Joachims [7]). Each term t was also ranked according to its information gain (IG) with regard to class c using the equation:

$$IG(c,t) = H(c) - H(c|t)$$

where H is the entropy. Selecting terms based on their information gain produces more accurate results. In this case, IG-based feature selection retained the top 10,000 terms from the IRC data collection.

4.5 Classification

The SVM-Light classifier [7] from the Weka toolkit [15] was used as the classifier. SVM-Light with a linear kernel function was chosen as the classifier because it performed better than Naive Bayes (NB), IBk (a k-nearest neighbor classifier), SMO (which implements the sequential minimal optimization algorithm) and the J48 algorithm (which is based on a pruned C4.5 decision tree).

Figure 2 presents an overview of the classification process. The initial stage involved the preprocessing of text as in the training phase. Following this, the data was transformed to a vector space model. Finally, the corresponding features were weighted according to the feature space

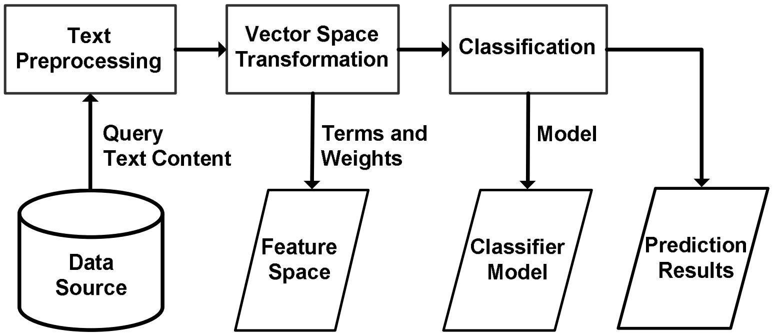


Figure 2. Classification process.

model and classified using the classifier constructed during the training phase.

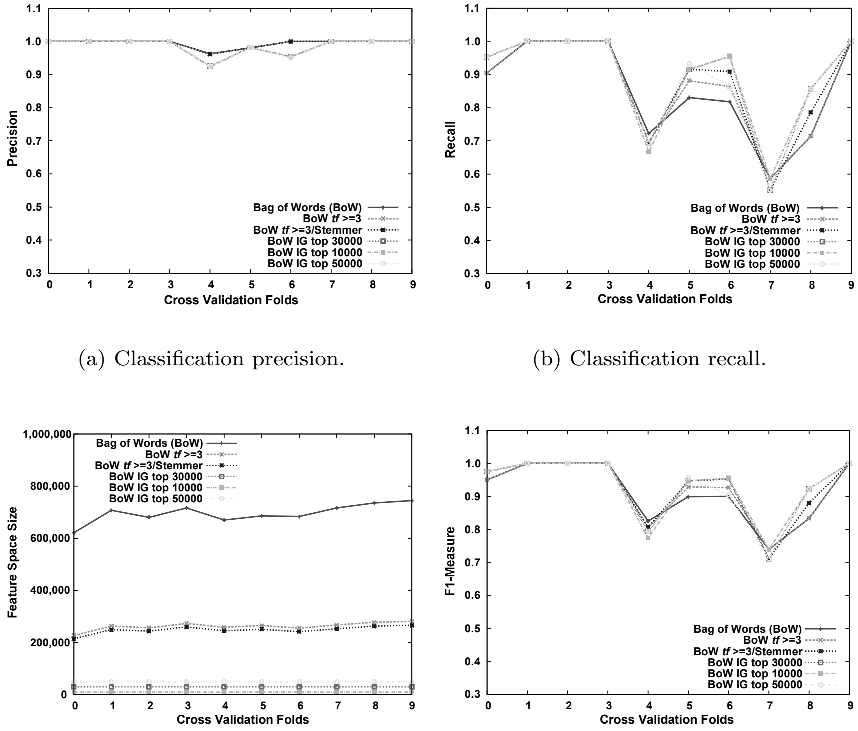
5. Performance Evaluation

The IRC data corpus was collected over a period of eleven months using an observation framework [2]; the corpus contains 51.3 million IRC messages transmitted over 2,693 channels on 246 networks. The web forum data corpus was collected by crawling through more than 203,000 threads in ten forums. This section evaluates the performance of the classification system and the effectiveness of various vector space models and document selection methods.

5.1 IRC Channels

To evaluate the performance of the classification system on IRC channels, we manually labeled all 2,693 IRC channels based on their relationship to the underground economy and performed k-fold cross-validation. Figure 3 shows the cross-validation results for the various vector space models.

Figure 3(a) shows that the SVM classifier maintains a consistently high precision, which means that the predicted results do not contain many false positives. The drop in the recall rate in Figure 3(b) is mostly due to channels in which the underground-economy-related content accounts for a fraction of the exchanged messages and are mistakenly classified as false negatives. In general, removing terms with $tf < 3$ combined with the English stop word list and the Porter stemmer produced an average precision of 99.43% and increased the recall from the initial 85.76% to an average of 88.09%. The feature selection based on the top 10,000 terms ranked by the IG reduced the vector space to 4% of the



(a) Classification precision.

(b) Classification recall.

(c) Feature space size.

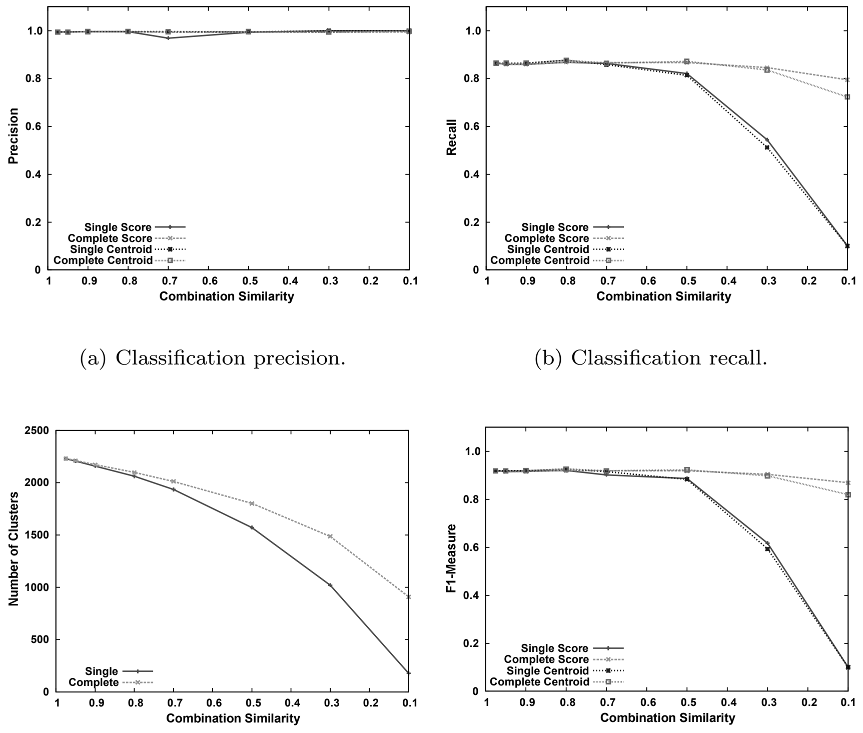
(d) Classification F_1-measure.

Figure 3. Underground marketplace detection in IRC channels.

noise-filtered space and yielded the best score with an average precision of 98.59% and recall of 89.32%. This results in an average F_1-measure of 93.14% and an average accuracy of 97.84%. Thus, the classification system performs very well despite the noisy content of IRC channels.

Additionally, we evaluated the performance of document selection for different similarity values. To this end, the IRC channels were merged into clusters determined by the combination similarity cutoff value. The document selection evaluation also analyzed the selection methods for the cluster representative and compared the centroid-based method against the score function approach, which is defined as the ratio of unique textual content to the number of messages in the channel.

A k-fold cross-validation was also performed based on the training sets generated by document selection. Figure 4 shows the average performance results.



(a) Classification precision.

(b) Classification recall.

(c) Number of clusters.

(d) Classification F_1-measure.

Figure 4. Classification performance of document selection IRC channels.

While single-link clustering reduces the number of clusters for the given similarity values more rapidly, it produces a significant accuracy loss. In contrast, complete-link clustering can reduce the number of needed training samples to less than 40% with minimal loss of recall. As shown in Figure 4(d), the selection methods for the cluster representative, which would be added to the training set, performed equally well for the upper interval of the combination similarity. Ultimately, the deviation between the two methods is only visible for very low combination similarity, where the score function based on the content information performed slightly better.

5.2 Web Forums

To evaluate the performance of the classification system on web forums, we manually labeled 300 randomly selected threads from the web forum `www.clicks.ws` depending on whether or not the posts were re-

Table 1. Average results of classification performance on web forums.

	Size	**Precision**	**Recall**	**Accuracy**	**F_1-Measure**
BoW	$\lfloor$34,890$\rfloor$	96.79%	83.55%	94.02%	89.40%
BoW, $tf < 3$	$\lfloor$14,391$\rfloor$	96.95%	83.75%	94.20%	89.72%
BoW, $tf < 3$, Stemmed	$\lfloor$11,720$\rfloor$	97.22%	84.58%	94.38%	90.32%
BoW, IG Top 5,000	5,000	95.87%	83.04%	94.01%	88.90%
BoW, IG Top 3,000	3,000	94.66%	81.60%	93.38%	87.37%
BoW, IG Top 1,000	1,000	94.81%	82.31%	93.47%	87.93%

lated to the underground economy. In addition, we extended the training set by another 100 randomly-selected threads from each of the other nine web forums. Table 1 shows the average performance of the classification system for the k-fold cross-validation of the web forum test set.

The classification system was very effective for the web forum threads, but unfortunately not quite as effective as for the IRC channels. The IRC channel contents involved more structured discussions and the threads were less noisy, which made it easier to extract information. The loss of accuracy is mostly caused by the dissimilarity of the selected samples, especially due to the German web forum `www.carders.cc`. As highlighted in Table 1, the approach with $tf < 3$ and English stop word filtering combined with the Porter stemmer performed best with an average F_1-measure of 90.32%. The IG-based feature selection did not show its advantages, but it is clearly not necessary in this case because of the dimensionality of the vector space. In conclusion, the vector space models show similar behavior as in the case of the IRC channel evaluation, demonstrating that the system is effective at detecting web forums used by cyber criminals.

6. Conclusions

A machine-learning-based classification system can be very effective at detecting underground marketplaces that use venues such as IRC chatrooms and web forums. Indeed, automatically identifying and monitoring these marketplaces can greatly enhance investigations of online criminal activities. The classification system described in this paper detected underground marketplaces in a collection of 51.3 million IRC messages with an average accuracy of 97%. Furthermore, the system classified a subset of threads from ten web forums, ranging from underground economy discussion forums to hijacked benign web forums, with an average accuracy of 94%.

Acknowledgement

This research was funded by the Austrian Research Promotion Agency (FFG) under COMET K1.

References

[1] L. Baker and A. McCallum, Distributional clustering of words for text classification, *Proceedings of the Twenty-First International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 96–103, 1998.

[2] H. Fallmann, G. Wondracek and C. Platzer, Covertly probing underground economy marketplaces, *Proceedings of the Seventh International Conference on the Detection of Intrusions and Malware and Vulnerability Assessment*, pp. 101–110, 2010.

[3] J. Franklin, A. Perrig, V. Paxson and S. Savage, An inquiry into the nature and causes of the wealth of Internet miscreants, *Proceedings of the Fourteenth ACM Conference on Computer and Communications Security*, pp. 375–388, 2007.

[4] Z. Harris, Distributional structure, *Word*, vol. 10(23), pp. 146–162, 1954.

[5] C. Herley and D. Florencio, Nobody sells gold for the price of silver: Dishonesty, uncertainty and the underground economy, *Proceedings of the Eighth Annual Workshop on the Economics of Information Security*, pp. 33–53, 2009.

[6] T. Holz, M. Engelberth and F. Freiling, Learning more about the underground economy: A case-study of keyloggers and dropzones, *Proceedings of the Fourteenth European Conference on Research in Computer Security*, pp. 1–18, 2009.

[7] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, *Proceedings of the Tenth European Conference on Machine Learning*, pp. 137–142, 1998.

[8] P. McNamee and J. Mayfield, Character n-gram tokenization for European language text retrieval, *Information Retrieval*, vol. 7(1-2), pp. 73–97, 2004.

[9] J. Radianti, E. Rich and J. Gonzalez, Using a mixed data collection strategy to uncover vulnerability black markets, presented at the *Second Pre-ICIS Workshop on Information Security and Privacy*, 2007.

[10] G. Salton and C. Buckley, Term-weighting approaches in automatic text retrieval, *Information Processing and Management*, vol. 24(5), pp. 513–523, 1988.

[11] F. Sebastiani, Machine learning in automated text categorization, *ACM Computing Surveys*, vol. 34(1), pp. 1–47, 2002.

[12] A. Singhal, C. Buckley and M. Mitra, Pivoted document length normalization, *Proceedings of the Nineteenth International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 21–29, 1996.

[13] Symantec, Symantec Report on the Underground Economy, July 07–June 08, Technical Report, Mountain View, California, 2008.

[14] R. Thomas and J. Martin, The underground economy: Priceless, *;login*, vol. 31(6), pp. 7–16, 2006.

[15] I. Witten, E. Frank and M. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Elsevier, Amsterdam, The Netherlands, 2011.

[16] J. Zhuge, T. Holz, C. Song, J. Guo, X. Han and W. Zou, Studying malicious websites and the underground economy on the Chinese web, *Proceedings of the Seventh Annual Workshop on the Economics of Information Security*, pp. 225–244, 2008.