# Covertly Probing Underground Economy Marketplaces

Hanno Fallmann, Gilbert Wondracek and Christian Platzer
{fallmann|gilbert|cplatzer}@iseclab.org

Vienna University of Technology
Secure Systems Lab

**Abstract.** Cyber-criminals around the world are using Internet-based communication channels to establish trade relationships and complete fraudulent transactions. Furthermore, they control and operate publicly accessible information channels that serve as marketplaces for the underground economy. In this work, we present a novel system for automatically monitoring these channels and their participants. Our approach is focused on creating a stealthy system, which allows it to stay largely undetected by both marketplace operators and participants. We implemented a prototype that is capable of monitoring IRC (Internet Relay Chat) and web forum marketplaces, and successfully performed an experimental evaluation over a period of 11 months. In our experimental evaluation we present the findings about the captured underground information channels and their characteristics.

## 1   Introduction

In recent years, there has been a significant rise in dubious or even outright criminal activity performed via the Internet [1,2]. For example, cyber-criminals conduct credit card fraud, trade compromised user accounts, or openly sell stolen banking credentials online. To communicate with each other and to coordinate themselves, cyber-criminals make use of online communication channels, such as chat rooms, instant messaging, e-mail or web forums. In particular, media like IRC (Internet Relay Chat) chatrooms or Internet forums are frequently used as *underground marketplaces*, virtual places where goods and services that are related to cyber-crime are being offered and traded. These marketplaces appear to be popular among criminals, as they are easily accessible, highly frequented and typically offer a high degree of anonymity to their participants. Clearly, the ability to monitor such underground economy marketplaces would allow researchers and law enforcement to gain new insights into the internals of the existing underground economy and to more efficiently predict or counter cyber-crime. The main contributions presented in this work are the following:

1. We present a novel system for covertly and automatically identifying and monitoring a large number of underground marketplaces simultaneously.

2. We performed an experimental evaluation of our implementation to prove the usability of the developed system.
3. Based on a dataset which spans a period of approximately one year, we present a comprehensive overview on currently established communication methods and channels within the underground community.

In the following section, we give an overview of existing work related to automatically monitoring the underground economy.

## 2   Related Work

Observing the underground economy is not a new topic, and several related studies have been previously published.

For example, in their study on the underground economy, Thomas and Martin [3] examine IRC based marketplaces. However, as the authors focus on a high-level analysis of the underground economy's structure and actors, they collected relatively little data from these marketplaces.

A more extensive study was conducted by Franklin et al. [1]. In their work, the authors present findings on the underground economy that they derived from the message data of an IRC channel. While a significant amount of data was collected, the scope of the marketplace observation is limited to a single data source.

Interestingly, Herley and Florencio recently published work [4] that claims that the underground economy trading places are classic examples of lemon markets [5], i.e. prices in the underground economy do not necessarily reflect the quality of the offered goods. Furthermore, the authors claim that IRC is mainly used by lesser-skilled cyber-criminals.

In a study [6] by the security company Symantec, both IRC and web forums were covered. The authors claim to have collected 44 million messages over one year in IRC and web forums. Unfortunately, no details on the methods and techniques used for collecting the data are given.

A different approach to underground marketplace monitoring is presented by Holz et al. [2]. Instead of observing the marketplaces directly, the authors analyze data that they have extracted from "dropzones", i.e. places where criminals collect stolen user data.

In the work of Zhuge et al. [7], aspects related to the Chinese underground market are described. The authors' focus deals primarily with the impact of malicious websites. To estimate the volume of criminal activity, they crawl a single black market forum and only one business platform.

## 3   Underground Marketplaces

To be able to determine the design criteria for an efficient system for monitoring the underground economy, it is necessary to first understand the characteristics of underground marketplaces. While, in theory, it is possible to use arbitrary

communication channels (such as e-mail lists) as underground marketplaces, our real-world observations indicate that only two specific types are widely used by cyber-criminals, IRC rooms and web forums.

### 3.1 IRC Rooms

IRC (Internet Relay Chat) is a well-known, popular, text-based chat protocol that is specified in an RFC document [8]. However, in order that IRC networks can develop new features as well as protect their users from spammers and automated malicious users, they extend the original RFC specification and create their own protocol rules. Unfortunately, these additions complicate an automated aggregation of messages for our monitoring system since our probes are automated as well. Furthermore, underground related IRC channels (i.e. chat rooms) are actively policed by the channel operator to get rid of unwanted participants, e.g. *rippers* who are fraudulent traders and who scam vendors and buyers alike. In order to solve these challenges, we mimic human behavior and aim at creating as little annoyance as possible to chat participants, while preventing our system from causing excessive resource usage for server operators.

### 3.2 Web Forums

Web-based forums are the second dominant medium used for underground marketplaces. They are often based on popular off-the-shelf software (e.g., phpBB [9], vBulletin [10]) that is commonly used for benign forums and message boards. In contrast to IRC rooms, forums exhibit a more restricted access policy. Typically, users who wish to participate have to first create user accounts and authenticate themselves via credentials (nickname and password), before they can write or sometimes even read messages. Communication is structured in forum *threads*, which represent a communication topic and consist of a list of messages posted by users, i.e. each new message in a thread is attached to the end of the list.

## 4 System Design

The overall aim of our system is to observe a large number of underground economy related communication channels. To this end, we deploy a number of sensors for distinct types of communication media. Furthermore, we aim at a system that can be easily extended (i.e., adding sensors should require little effort). In the scope of this document, a *probe* is a software agent within our system that executes surveillance tasks on a specific type of communication medium and that is managed by a *probe pool*. Moreover, the probes are able to collaborate on a given task in a coordinated manner. To this end, probes have the ability to communicate with the main system. For example, a probe can notify the main system if it is unable to continue its task, thus activating a replacement probe. Additionally, our system can incorporate many different network interfaces, making it more stealthy and flexible.

### 4.1 IRC Sensor

Our general aim in observing IRC networks is to covertly detect underground trading channels and to retrieve a maximum amount of information from them. To this end, it is necessary to observe these channels for longer periods of time (i.e. at least several days), while capturing all public messages during this time frame. In practice, this is a non-trivial task, as it requires our system to be resilient against being intentionally blocked or banned from communication channels by administrators. At the same time, our system should be able to collect user data from individual participants within a channel, while appearing as a genuine user itself.

**Information Gathering Methods.** Besides recording messages from IRC channels, the IRC sensor probes have the ability to retrieve information about users directly. To this end we developed several methods that differ in the returned information and their "visibility", i.e. some can be regarded as common IRC operations while others might appear more suspicious to a channel operator.

For example, by sending an IRC `whois` request we retrieve, amongst other values, information about the "real" name (designated by the user), the IP address, the channels the user has currently joined, as well as the information if the user has IRC operator privileges. Additionally, we can make use of protocols built on top of the IRC infrastructure (e.g. CTCP (Client To Client Protocol) [11] and DCC (Direct Client to Client) [12] protocol) to learn more about the adversaries and the IRC clients they are using. Moreover, if we collect the IP address of a user, we can apply tools like the geolocation database GeoIP [13] (to pinpoint the IP address to a geographical location), or Nmap [14] (to acquire specific information about the computer of the adversary).

**Observation Strategies.** To control the behavior of individual probes, the system assigns an *observation strategy* to each probe instance. Each strategy determines which information gathering techniques are used by the probes, and how "aggressive" they are in pursuing their goals.

*Chain Strategy.* The chain strategy aims at automatically extending the original observation scope (defined by the probe's initial list of channels) by finding additional, interesting channels. To this end, probes following the chain strategy periodically request the list of joined channels from each user in the currently observed channels by sending IRC `whois` requests. For each newly found channel, the size of the intersection set with users in the already observed channels is computed. The channel with the largest intersection set is regarded as the most "popular" channel, and will be added to the probe's target list.

*Swap Strategy.* We found that IRC users who are too passive, i.e. who do not participate in conversation at all, are frequently removed from IRC channels (e.g. to prevent resource waste or to get rid of "zombie" peers who did not log out

properly). To prevent this from happening to our observation probes, the *swap strategy* adds additional probes to observed channels after a certain amount of time. Before the original probe leaves the channel, it waits for a random, intentional overlap time to obscure the "swap".

*Chat Strategy.* A considerably large proportion of underground economy channel messages are from announcement-bots that advertisers use to draw attention to their offers and requests. Typically, users are asked to send a private message to learn about details of these business offers. As soon as the strategy encounters this type of message, the promoter will be directly engaged using the A.I.M.L. [15] chat system. This chat system locates proper responses to incoming messages using a library consisting of entries written in a XML dialect called Artificial Intelligence Markup Language.

*Sensor Strategy.* The purpose of this strategy is to cover channels with names and topics that match denoted patterns. To this end, the strategy dispatches an IRC `list` command to retrieve the channels of the network and assigns one probe for each matched channel.

*Combinations.* Various compositions of strategies can be constructed with different grades of observation behavior, ranging from *passive* to *aggressive*. For example, the combination of a chain strategy with a swap strategy with an aggressive observation attitude using `whois` and DCC requests to concentrate on users, leads to a more adaptive strategy that rotates the probes between channels and expands the observation coverage.

**Supervising Information Accumulation - The Right Strategy for the Right Job.** To probe underground communication channels in an IRC network they must be discovered first. For each network a network supervisor is initiated that starts the sensor strategy with include and exclude patterns specifically designed for the purpose of recognizing fraudulent trade channels. Our aim is to quickly find these channels in the beginning and then expand the observation with additional methods. If, for example, a credit card trading channel is masked as a sports channel, it will be initially ignored with this method, but will be discovered by another technique.

As soon as a new channel has entered our observation scope a channel supervisor is loaded. Since the intention is not to annoy innocent users and cause needless traffic, the channel supervisor starts the surveillance in a passive manner. After a designated time, all the messages belonging to the channel are automatically assessed on the relation to underground economy context. An SVM (Support Vector Machine) [16], with a training set tailored to recognize fraudulent content, makes the classification possible. Based on the affinity of a channel to underground economy, the channel supervisor adapts its observation strategies and mechanisms.

Besides using the *pattern matching* approach to broaden the observation scope, the channel supervisor applies the chain strategy on fraudulent channels to

observe other *popular channels* of the current users. Additionally, if a designated quota of maximum channels is not reached, *random channels* are being joined and dismissed as soon as the SVM reasoner classified them as being benign to make room for new random channels.

## 4.2 Web Forum Sensor

According to Guo et al. [17], web forums exhibit a number of characteristics that make it non-trivial to extract structured data with standard web crawlers. The major problem is that the same content of a forum can have a multitude of URIs addressing it. One reason, is the dynamic nature of a web forum. For example, two requests, that differ in the URI, can lead the forum engine into generating the same content. Another reason is the existence of "noisy links", i.e. URIs that contain functions like ordering posts or searching content. This problem can lead a standard web crawler to be redirected in a circular way (so-called *spider-traps*). A further intricacy we faced is the diversity of different forum engines and versions that require a general solution. Taking these challenges into consideration, we decided to use the approach described by Yang et al. [18].

**Crawling Underground Economy Forums.** Unlike benign forums (i.e. non-underground), most underground economy related forums employ some sort of additional authentication measures or counter-measures that prevent automatic crawling. The following list highlights the most frequently employed mechanisms, and how we address them in our implementation.

1. The content is only viewable for registered users. Since we enhanced the crawler with login functionality, we only have to manually register users to the forum one time.
2. Reputation-based trust systems that allow only users who gained a high enough status to view the content. An example are *escrow services*, i.e. forum administrators charge a fee to verify the integrity and quality of trade offers and monetary transactions before any goods are exchanged. We are not able to automatically gain these privileges for a user, but if the content seems to be valuable for analysis, this can be done manually.
3. Individual users can only view a certain amount of pages per time unit, sometimes coupled with a limitation of recovered pages per network address. The solution to evade these restriction of viewable pages is similar to the IRC swap strategy: Each registered probe has a dedicated IP address assigned. During the crawl of the forum the probes are being changed (logoff old probe, login new probe) after a random amount of acquired pages.

## 5 Experimental Evaluation

We started an observation of underground economy marketplaces in March 2009 and collected data for the following 11 months.

### 5.1 Coverage and Proportion of IRC Networks

To find interesting IRC networks for our experiment, we initially extracted known server addresses from the server list of mIRC [19], a popular IRC client, and from a website [20] that is dedicated to finding IRC networks. Additionally, we complemented this list with servers that we manually extracted from announcements in underground economy IRC channels.

In our crawling experiments, we examined a total of 26,207 distinct IRC channels from 291 networks. Of these, 2,677 channels contained chat messages and 4.7% of them have been recognized to be related to underground economy content. A chat message is a public channel message that excludes server notifications such as join (i.e. entering a channel), part (i.e. leaving a channel) or kick (i.e. expel a user from a channel). The content identification has been done with an SVM (Support Vector Machine) [16]. The categorization results were manually checked and we found zero false positive recognized channels and 67 false negative recognized channels. However, the SVM module is exchangeable and an improved version can be effortless integrated. We found 23,530 channels to contain no chat messages.

In Sect. 4.1 we have described three different basic methods to provide a reasonable coverage of fraudulent channels in the search space. The *pattern matching* approach retrieves the obvious channels and has the biggest hit rate with nearly 58% of all covered fraudulent channels. The chain strategy provides all the *popular channels* we missed and constitutes with a scope of over 40% together with the pattern matching approach, 98% of all exposed channels. Only two channels have been detected with the joining of *random channels*. However, this method allowed us to exclude over 22,000 channels not being used for criminal activities.

### 5.2 IRC Observation Results

In 291 IRC Networks 495,939 distinct user names have been accumulated. Using 14,526 probes we gathered 15 GB of data for which the statistic is listed in Table 1. *Kicks* denotes the number of received expulsion over all users where as *Kicks of Probes* only takes our observing users into account and *Distinct Kicks of Probes* counts manifold expulsions of a probe in a channel as one. *Channel Bans* consists of the number of distinct channels we were banned from. Comparing the total value of the kick rate with the rate that affected only our probes, it is clearly visible that our strategies significantly reduced the potential of being expelled from a channel. Because traders advertise their goods with a high message rate, they are responsible for a big fraction of all the messages we collected. We can clearly confirm the presence of these traders and the extension of their actions in IRC.

### 5.3 Web Forum Observation Results

First of all, before the observation can start, we have to locate addresses of web forums in which illicit trade is taking place. To this end, we initially gather URIs

**Table 1.** Statistics of the IRC observation regarding the user behavior.

|  | Malignant | Per Channel | Benign | Per Channel |
|---|---|---|---|---|
| Channels | 126 | - | 2,551 | - |
| Chat Messages | 43,148,421 | 342,447.79 | 2,950,208 | 1,156.49 |
| Joins | 550,685 | 4,370.52 | 562,002 | 220.31 |
| Parts | 100,354 | 796.46 | 169,670 | 66.51 |
| Kicks | 25,298 | 200.78 | 2,792 | 1.09 |
| Kicks of probes | 79 | 0.63 | 1,996 | 0.78 |
| Distinct Kicks of probes | 42 | 0.33 | 1,105 | 0.43 |
| Channel Bans | 26 | 0.21 | 681 | 0.27 |

via keywords entered in web search engines. After underground economy related forums have been crawled, it is possible to derive forum addresses from them. Additionally the sensor system provides a global search on all communication channels: if a forum address is posted in an IRC channel, an observation can be started immediately on it and vice versa. To test the web forum sensor, we crawled eleven different forums multiple times. All in all we gathered from 11 web forums over 127 GB of pages that contain over one million forum posts.

### 5.4 Classification and Analysis of Web Forums related to Underground Economy

Web forums are being used differently by miscreants: Advertisers use spamming tools on mostly innocent forums to promote messages similar to those on IRC channels. Some forums are only used to exchange knowledge, to provide tutorials for beginners or to find new contacts. Other forums include trading sections as well, including auctions of stolen goods.

Table 2 shows examples for different usages of such forums are listed. A low value of *Posts per User* is an indication for a high proportion of different users with a low post count. This can be, either due to the fact that a forum is fairly new, or if it is being abused by spammers. Depending on the vigilance of the forum admins, these newly created users and posts have a short life-time. Since the majority of the spamming tools create a new thread and post one spam message into it, the percentage of threads with only one post can additionally be used to figure out the current state of the forum regarding spam. The forum `www.talk-hyip.com` with an average rate of 2.75 posts per user and a proportion of 94% of threads containing only one post, is an example of a lost battle against spammers.

The time span of the forums, from the date of the first post till the date of the last post, reflects on the forum type as well. In our data, the lifetime of fraudulent forums with trading and discussion sections is shorter when compared to spammed forums. To get the data of underground economy forums with an active community we visited some of the sites more than once, to find out how

**Table 2.** Examples of different forums and how they are being used in the underground economy. (The letter $d$ stand for discussion of underground economy, $t$ for the trading of goods, and $s$ for a spammed forum).

| Forum | Boards | Threads | Posts | Users | Posts/User | 1 Post/Thread | d | t | s |
|---|---|---|---|---|---|---|---|---|---|
| blackhatpalace.com | 22 | 424 | 1,323 | 160 | 8.27 | 49.92% | ✓ | | |
| forum.rorta.net | 21 | 2,643 | 53,731 | 843 | 63.74 | 5.60% | ✓ | | |
| www.carders.cc | 67 | 6,290 | 65,312 | 2,411 | 27.09 | 16.96% | ✓ | ✓ | |
| www.hack-info.ru | 47 | 27,221 | 207,020 | 9,891 | 20.93 | 34.80% | ✓ | ✓ | |
| www.clicks.ws | 26 | 9,463 | 19,975 | 2,681 | 7.45 | 76.71% | | | ✓ |
| www.hotsurfs.com | 62 | 39,297 | 61,287 | 2,161 | 28.36 | 88.78% | | | ✓ |
| www.talk-hyip.com | 9 | 3,884 | 5,966 | 2,166 | 2.75 | 94.00% | | | ✓ |
| www.wifi-forum.com | 22 | 109,221 | 610,658 | 32,084 | 19.03 | 59.00% | | | ✓ |

many new posts have been committed. Occasionally it happened that a site was offline, either because they were completely shut down, or because they were just not reachable for a few months.

## 6 Conclusion

In this work, we presented a novel system for automatically monitoring adversary information channels. For example, in the domain of the online underground economy, researchers who study online crime or law enforcement agencies have a vital interest in acquiring data from related sources, such as underground marketplaces or chatrooms used by criminals. To the best of our knowledge, our system is the first to include specific features to monitor information channels related to the underground economy. Furthermore, our system can mimic (human) user behavior to remain stealthy, i.e. avoid being detected by administrators.

For an experimental evaluation we have implemented a prototype that can observe IRC channels and web forums, the most widely spread information media used by cyber criminals. During a period of 11 months, our system has managed to collect a dataset of more than 43 million chat messages and approximately one million forum entries from underground sources without experiencing any problems. This demonstrates that our system can be effectively used in a real-world setting to acquire vital information on cybercrime, which can be used for investigations or research in the area.

## 7 Acknowledgements

# References

1. Franklin, J., Paxson, V., Savage, S., Perrig, A.: An Inquiry into the Nature and Causes of the Wealth of Internet Miscreants. In: ACM Conference on Computer and Communications Security (CCS), ACM (November 2007)
2. Holz, T., Engelberth, M., Freiling, F.C.: Learning More about the Underground Economy: A Case-Study of Keyloggers and Dropzones. In: ESORICS. (2009) 1–18
3. Thomas, R., Martin, J.: The Underground Economy: Priceless. In: USENIX ; LOGIN:. (2006)
4. Herley, C., Florencio, D.: Nobody Sells Gold for the Price of Silver: Dishonesty, Uncertainty and the Underground Economy. Technical report, Microsoft Research (2009)
5. Akerlof, G.A.: The Market for "Lemons": Quality Uncertainty and the Market Mechanism. The Quarterly Journal of Economics (3) (1970)
6. Symantec: Symantec Report on the Underground Economy. `http://eval.symantec.com/mktginfo/enterprise/white_papers/b-whitepaper_underground_economy_report_11-2008-14525717.en-us.pdf` (2008)
7. Zhuge, J., Holz, T., Song, C., Guo, J., Han, X., Zou, W.: Studying Malicious Websites and the Underground Economy on the Chinese Web. Technical report (2008)
8. Oikarinen, J., Reed, D.: RFC 1459: Internet Relay Chat Protocol. Technical report (May 1993)
9. Online: phpBB. `http://www.phpbb.com/` (Accessed: April 2010)
10. Online: vBulletin. `http://www.vbulletin.com/` (Accessed: April 2010)
11. Zeuge, K., Rollo, T., Mesander, B.: Client To Client Protocol (CTCP). `http://www.irchelp.org/irchelp/rfc/ctcpspec.html`
12. Zeuge, K., Rollo, T., Mesander, B.: Direct Client Connection (DCC). `http://www.irchelp.org/irchelp/rfc/dccspec.html`
13. Online: GeoIP. `http://www.maxmind.com/` (Accessed: April 2010)
14. Online: Network Tool Nmap. `http://nmap.org/` (Accessed: April 2010)
15. Wallace, R.: The Elements of AIML Style. Technical report, ALICE A.I. Foundation (2003)
16. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: European Conference on Machine Learning (ECML), Berlin, Springer (1998) 137–142
17. Guo, Y., Li, K., Zhang, K., Zhang, G.: Board Forum Crawling: A Web Crawling Method for Web Forum. In: WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Washington, DC, USA, IEEE Computer Society (2006) 745–748
18. Yang, J.M., Cai, R., Wang, Y., Zhu, J., Zhang, L., Ma, W.Y.: Incorporating site-level knowledge to extract structured data from web forums. In: WWW '09: Proceedings of the 18th international conference on World wide web, New York, NY, USA, ACM (2009) 181–190
19. Online: mIRC server list. `http://www.mirc.com/servers.ini` (Accessed: April 2010)
20. Online: IRC netsplit. `http://irc.netsplit.de/` (Accessed: April 2010)