ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 3, March 2020

The Efficient Mechanism of Cloaking Used by Hackers to Misguide Search Engine Crawlers and Spiders

Dipanshu Parashar[1]

Founder at The Cyber Agents, Mayur Public School, New Delhi, India[1]

ABSTRACT: Search engines have been making their security better and improvising their mechanism of Search Engine Bots to detect unforbidden activities that are violating the Internet Laws. Cloaking is one such activity which as per the definition says, a technique of delivering a certain set of pages to search engine crawlers, while at the same time delivering a completely different set of pages to human visitors. It has widely impacted the most famous Search Engine Industries and social media companies.This is one of the major backbones for hackers who run the malicious webpages unethically over the Internet and violates the terms and conditions of any Internet authority. This paper attempts to provide detailed research on cloaking and its functioning. The proposed algorithm can also be used to reduce such activities and highlights the Black hat SEO techniques.

KEYWORDS: Search Engine Optimization [SEO], Internet Protocols, Cloaking, Phishing, Spamming, Webmaster Console, Cybersecurity.

I. INTRODUCTION

Cybercrimes have been tremendously increased in the last decade and most of the spammers and hackers have arrived through the means of black hat techniques and violated the legalities along with harminginnocent internet users. One of the most common ways these attackers are targeting the Netizens [Net + Citizens] is by giving them fake offers which work as honeypots for these victims. There is a need to combat these situations and a lot of case studies have proved that these cyberattacks are usually becoming the source of mental issues too.

We are here to throw light upon one such mechanism which has paved the way to hackers for intruding in your systems and leaking your personal data or information. Cloaking is an art where the website content is different for search engine crawlers and human visitors. This technique works like a fake mirror for the bots. These bots can belong to any leading internet authority. Cloaking is based on an algorithm which when applied,can work on any social media platform and on any search engine. It is better explained in Figure 1 with the pictorial representation.

Spiders or Crawlers have their well-defined IP. To implement cloaking, what attackers do is use PHP script in the index page of a genuine website. To better index their websites through the webmaster console, the script is then responsible to show different content to these defined IPs of crawlers. This technique can be extended to specific countries, specific time zones and are different for various social media sites. Google, Facebook, and many such sites have their bots to crawl on the link sent to anyone if it finds suspicious, sharing the link is then not allowed. But suppose an attacker has applied to cloak then the bots will see a genuine site but the victim or human visitor to the link will be prone to a malicious website. It is completely violating the conditions of any Search Console.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 3, March 2020

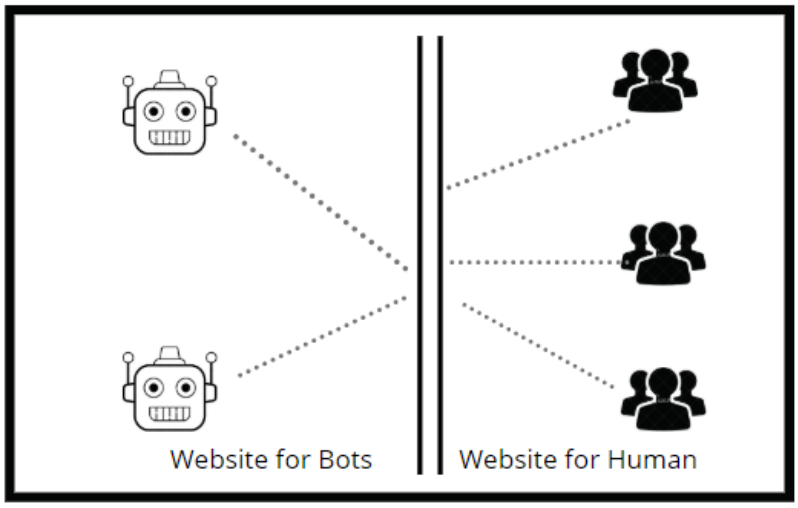

Figure 1

These are considered in black hat Search Engine Optimization [SEO] Techniques which can be identified by a search engine but most of them are unable to do so. SEO is of two types and here below, few of the major tricks are given. SEO technique that is not approved of by the search engines is termed as Black Hat SEO or spamdexing. Search engines can easily identify Black Hat SEO practices which will impede you from getting any benefits. Black Hat SEO practices include:

(i) SEO practices that involve deception and are disapproved by the search engines.

(ii) Redirecting users to human-friendly pages from search engine friendly pages or redirecting users to any page which is different from the page which was earlier ranked by the search engine.

(iii) Using cloaking SEO in which one version of the page is served to search engine spiders/bots and another version is served to human visitors.

(iv) Employing meta tag stuffing in which keywords are repeated in meta tags but the content is not related to those keywords.

(v) Employing keyword stuffing in which keywords are placed in a calculated manner in a content.

(vi) Using Doorway or Gateway pages in which low-quality web pages contain very little content which is stuffed with the same kinds of keywords and phrases.

(vii) Using mirror websites in which multiple websites or different URLs use conceptually similar content.

Cloaking doesn't require any kind of special requirement, but the knowledge of PHP along with Digital marketing is sufficient to cloak any webpage. Cloaking has not only been utilized by the hacker individuals but since using cloaking can help the person to rank the website at top in any search engine also many SEO practicing people use it for industrial use.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 3, March 2020

II. REQUIREMENTS AND FUNCTIONING

The black hat SEO only works on PHP scripts and majorly doesn't require any sperate hosting provider. Now let's get clear by understanding a scenario. A website is running on the internet because a minimum amount of hosting space has been allocated to the domain. All the webpages made in HTML, CSS, JS are uploaded on the webserver mostly in the Public_Html directory of FTP. But in the case of cloaking, an attacker needs to have a virtual private server for hosting and cannot successfully run on a Windows-based Hosting provider. Cloaking requires the Ubuntu-based platform to be applied correctly and without any error. So, an attacker usually gets a Virtual Private Server [VPS] from various sources such as digital ocean and then to create a cloaking application, the server is connected to the webpage using a mediator. In this case, Server Pilot works as the mediator.

The script used for applying cloaking or to fool the crawlers, spiders and bots, is based on PHP. This script includes the detailed information of IP, countries, time zone, CMS, technology and can be even customized to be as competitive comparatively to industries. Figure 2 represents a basic code of execution to fool google bots or in brief, this is a demo google cloaking script. Including this in the index page will cloak the website.

```
<?php

$redirect="http://www.example.com/";

include "robotIPs.php";
include "robots.php";

$robots=explode("\n",trim($robotList));
$seips=explode("\n",trim($seips));

if(!in_array($_SERVER['HTTP_USER_AGENT'],$robots)&&!in_array($_SERVER['REMOTE_ADDR
'],$seips)){
 header("Location: ".$redirect);
 exit;
}

?>
<html>
 <head>
  <title>Search engine content</title>
 </head>
 <body>
 <p style="text-align:center; font-size:18px; line-height:24px; font-weight:bold;">
 This is the content search engines<br />
 will see when they visit the page</p>
 </body>
</html>
```

Figure 2

This PHP code includes the redirect URL where the attacker will use the malicious website hosted on a VPS. Now, robots are here referring the bots of google which ping the website to check the originality and harm status. Before moving on the proposed algorithms and their design consideration, there is a need to understand the Industrial Impact of Cloaking and how it is affecting the human visitors not only by ranking at top in Search Index but by providing fake and malicious content to the victims.

III. INDUSTRIAL IMPACT CASE STUDIES

A lot of search engines and almost every social media site is running not only on the database of Users but also by the advertisement campaigns they have allowed to business minds and social media experimentalists. Their main source of income is these advertisements. But as per the terms and conditions, these websites do not allow to run malicious advertisements and campaigns.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 3, March 2020

Before approving any kind of campaign proposed by the user, the sites crawl the link but the cloaking script redirects the bots specifically made for that particular advertisement to a legitimate website and thus these bots find the campaign genuine and accept it. But when the human visitor enters the campaign shown on different sites such as Facebook, Instagram, they will be redirected to a spam site. Figure 3 depicts where users are investing and attracting viewers through the means of Social Media Ad Campaigns.

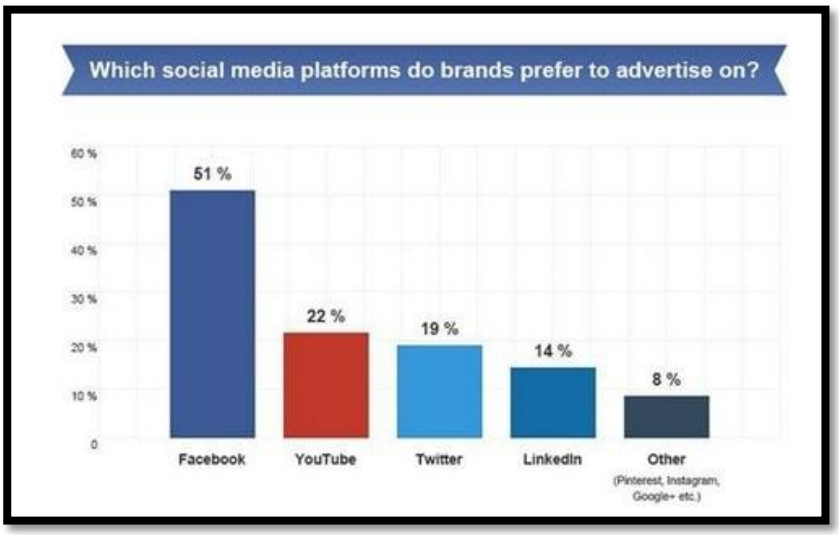

Figure 3

Cloaking is not limited only to desktop or laptop version sites. It can be customized for various devices where the user agent can either be android, MAC or windows. Taking a scenario, advertisers most likely to invest in mobile ads since, social media is mostly working on mobile devices.

Mobile users contribute to a majority of social advertising revenue. In fact, 94% of the Facebook advertising revenue for Q3 of 2019 came from mobile.The average ad spend per internet user also sees a gradual increase on mobile, jumping from $13.49 in 2018 to $15.40 in 2019. In 2020, advertisers will likely spend $16.85 per mobile internet user. Figure 4 is representing the investments made on mobile and desktop ad campaigns.

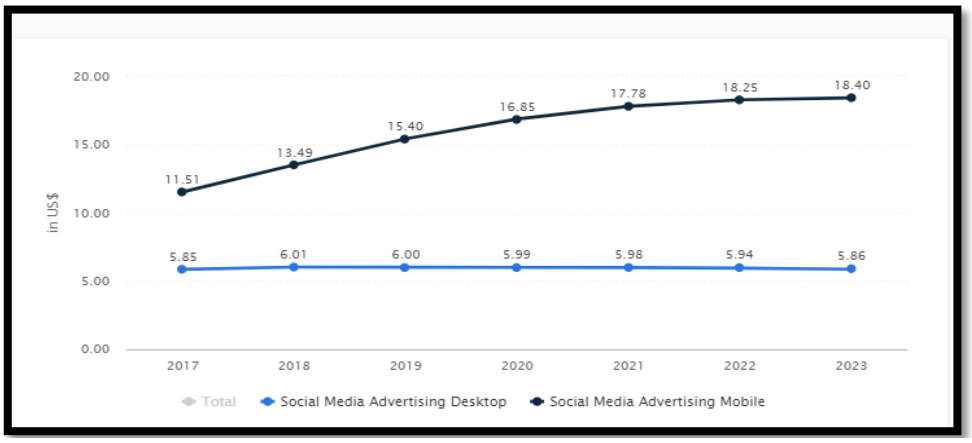

Figure 4

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 3, March 2020

These statistics clearly state that almost 50% of social media are depending on promoting products. This means almost every day at least 25%-30% of social media users who are clicking on the cloaked ads will be prone to malicious websites and the scams. Hence, here the proposed algorithm based on Python will help Search Engines to detect cloaking with few basic modifications.

IV. PROPOSED ALGORITHM

A. *Design Considerations:*

- The suggested design is based on Python and User-Agent, Location technology. The algorithm will ping the web application in various ways to check the difference between the contents of the application.
- The idea is utilizing external sources for now to change location but with few basic modifications can be integrated with Social Media Platforms.

B. *Description of the Proposed Algorithm:*

Aim of the proposed algorithm is to minimize the spams and hacks going around the world through the online marketing means and to minimize the effects of black hat SEO, Cloaking. This tool actually pinging the website or any application from different locations using proxies. The cloaking works on location bases, like what to show in which country. Proxies establish a fake connection between the web server and your local machine [Computer], this will access the website from different location on specific time delays.

If the content of website remains same then the website is perfectly fine for the use, otherwise, it is a cloaked website. In this case, the script is using external modules for efficient use. A] Request module is to request the website access or to allow the code to ping website and B] Beautiful soup module is used to work with proxies and manage the data effectively. Figure 5 shows the code to ping website on particular delays and from different locations using the proxies available on internet for temporary use.

```
import requests
from bs4 import BeautifulSoup
import random
def getproxy():
 with requests.get("https://www.sslproxies.org") as r:
  soup=BeautifulSoup(r.content,'html5lib')
  return {'https' : random.choice(list(map(lambda x:x[0]+':'+x[1], list(zip(map(lambda
x:x.text, soup.findAll('td')[::8]),map(lambda x:x.text, soup.findAll('td')[1::8])))))),'http' :
random.choice(list(map(lambda x:x[0]+':'+x[1], list(zip(map(lambda x:x.text,
soup.findAll('td')[::8]),map(lambda x:x.text, soup.findAll('td')[1::8])))))}
def proxyrequest(url):
 while 1:
  try:
   proxy=getproxy()
   p=requests.get(url,proxies=proxy,timeout=5)
   break
  except:
   pass
 return p.text
def normalrequest(url):
 q=requests.get(url)
 return q.text
url1=str(raw_input("Enter the url(Ex: http://www.google.com) :"))
if(proxyrequest(url1)==normalrequest(url1)):
 print("The Website isn't cloaked")
else:
 print("The website is cloaked.")
```

Figure 5

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 3, March 2020

Once the request is made, the content of the website will be matched to the content gained by another request to ping website after a specific delay and from different locations. If the content of website is similar in both the cases, then the website is considered as Cloaking free. This can be even modified to cross-check the data or content of website captured by any search engine bots. Figure 6 is a representation of how cloaking can be detected on a minor basis.

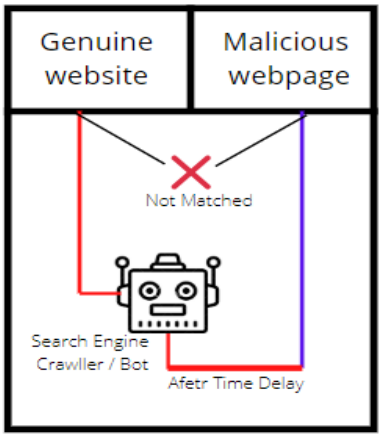

Figure 6

Violating the Internet Protocols by presenting a different sets of pages to search engine crawlers and spiders, and showing different set of ages to human visitors can be disastrous for Netizens. Using the proposed idea, usual Internet visitors can get an idea of evil lurking going behind their computer screens. With basic modification in database connection with the code and as per particular requirements, just by using the concept, search engines can provide efficient security and can be reliable to the visitors.

Cloaking is not about any script or location; it is a wide area full of opportunities for hackers to target their victims. The basic concept behind this is to ping a website from different User-Agent (the request maker), different locations, using proxies and Virtual Private Servers to identify whether you are provided with authentic work or not.

V. CONCLUSION AND FUTURE WORK

The Black Hat Search Engine Optimization techniques have been widely used by spammers to target specific sets of victims. Cybersecurity in industrial control systems (ICS) has received much attention lately, not least in the wake of the Stuxnet malware. ICSs are used to control facilities such as water plants, industrial production, electrical power distribution, and oil refineries. The complexity of these systems makes testbeds and simulations attractive for cyber situational awareness. Cloaking can be categorized broadly as advanced level phishing where the victim is served with fake website. It violates the Search Engine terms and conditions. Many Social media agencies and Search engines are unintentionally approving non-necessary advertisements which can be the source of payload intrusion, tech-support scams, Online frauds and much more. Cybersecurity has much more stuff to explore. To provide a white hat ethical security to the systems, passing through the black hat phase has become a necessary requirement.

REFERENCES

1. Search Engine Journal, Loren Baker, why to avoid Search Engine Cloaking
2. Catchupdates.com Black hat SEO
3. Social Media Statistics, Sprout Social, sproutsocial.com
4. Thecyberagents.com
5. Cyber situational awareness e A systematic review of the literature, Ulrik Franke
6. Marketing91, Social Media Advertisement Statistics
7. Hoot Suit, Social Media Statistics and advertisements