National Cybersecurity and Communications Integration Center

05 December 2013

Google "Dorking"

Executive Summary

Search engines are powerful tools that allow individuals to locate content across the web by searching for specific key words (queries). These tools use large indexes and algorithms to determine which websites contain content that most closely matches the keywords for which an end user has searched. Though search engines provide an efficient and effective way of locating intentionally available information on the web, they can also be valuable assets to malicious actors attempting to locate data that organizations may have never intended to be openly available to the public. Therefore, understanding the more advanced functions of search engines can make locating desirable information much easier for common end users, vulnerability testers and those with malicious intent. Using the advanced search functions is not necessarily difficult and with a plethora of information and resources that explain and/or automate the process, those with little experience can use these functions to quickly locate information they're seeking. It is important then to understand how advanced searches work so users may apply this knowledge to their overall approach to security.

Google

Google, ranked by Alexa as the world's most commonly visited website, is a search engine that allows its users to locate information and resources including webpages, images, and videos.[1] In 2012 the average number of Google searches per day exceeded 5 billion.[2] Google, as well as other prominent search engines like Bing and Yahoo, are able to quickly provide users with the information they are seeking by utilizing a process called web indexing. Web indexing is accomplished using software commonly referred to as a spider (aka crawler, bot), to filter through web pages, locating and tagging website keywords and adding that information to the index.[3] When a user enters a keyword into a search engine, the tool will access its index and identify matching keywords. Search engines then use special algorithms to rank results associated with these keywords in an attempt to present the most relevant results at the top of the search engine results page (SERP).[4] Google's confidential algorithm uses a PageRank system that assigns a score to all the matches it finds in its index and presents the highest ranked websites first.[5] The more websites that can be crawled and indexed, the better the chance a search engine will locate the information an individual is looking for. As of 2012, Google's index was estimated to contain over 50 billion websites and has continued to increase annually.[6] In addition to standard key word queries, search engines also allow individuals to use specialized search parameters referred to as advanced operators to locate very specific information.[7,8]

Google Dorking

Google dorking (aka Google hacking) is a term used to define the use of Google's advanced search operators to aid in locating vulnerable data or misconfigured websites.[9,10] The availability of this data is usually the result of weak security and can expose information including:

- Admin login pages
- Username and passwords
- Vulnerable entities
- Bank account details
- Sensitive documents
- Govt/military data
- Email lists

While many of the advanced operators used in Google dorking will perform the same function in other search engines, not all are directly interchangeable.[11] Therefore, this product will focus on Google as it is the most commonly used search engine and is most often associated with this process.[12] Google dorks follow the syntax, *"Operator:keyword" (where the keyword relates to the information a user wishes to locate)* and can contain one or many operators. Advanced operators and the information they are intended to locate include the following:[13,14]

Example	Operator	Specific term/string	Searches for
allintext:Google Dorking	allintext	Google Dorking	Only websites containing both the words Google and Dorking within the page text
link:www.googledorking.com	link	www.googledorking.com	Webpages that point to www.googledorking.com
allinurl:Google Dorking	allinurl	Google Dorking	Only URLs containing Google and Dorking within the URL
filetype:xls	filetype	xls	Pages whose name end in .xls indicating the page is an excel spreadsheet

Individuals can use a single operator value in a search query such as *"inurl:Dork"* or combine multiple operators in the same string to return more narrowed and targeted results. For example: *"filetype:xls inurl:google"* will locate webpages with the file extension .xls that contains the word Google in their uniform resource locator (URL). Adding new operators can be useful when an initial search returns results that an end user may wish to further filter.[15]

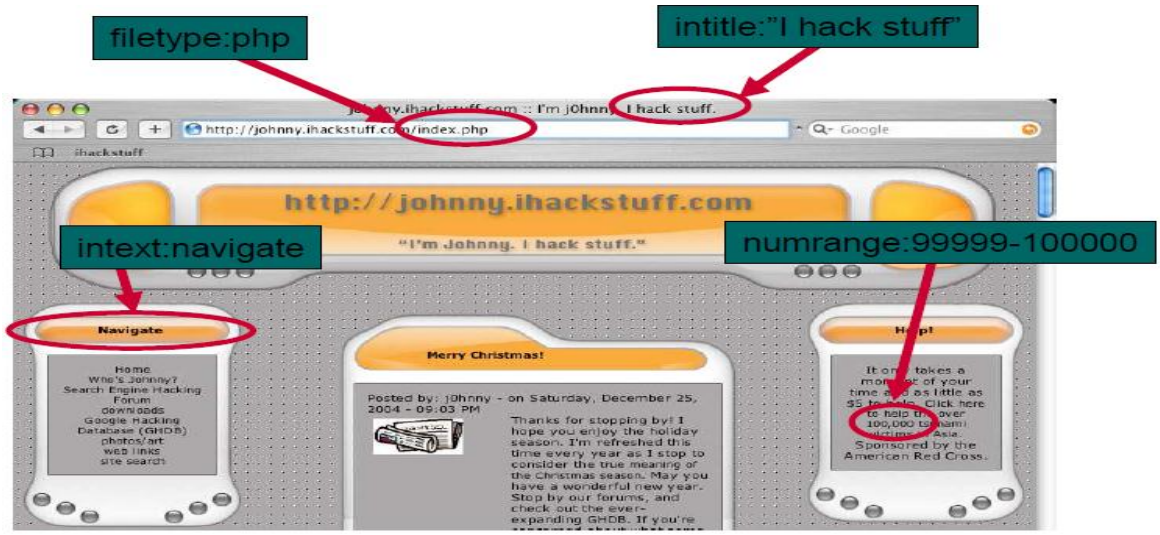

Figure 1: Illustration of Advanced Operators

While the examples provided are very basic, the Google Hacking Database (GHDB), found at http://www.exploit-db.com/google-dorks contains thousands of ready-made Google dork queries that can be copied and pasted by any user to locate a wide array of sensitive information. This database was created as a resource for penetration testers and vulnerability researchers to serve as a comprehensive collection of exploits gathered through direct submissions, mailing lists, and other public sources. This database is broken down into categories including: files containing usernames, vulnerable servers, files containing passwords, and advisories and vulnerabilities. The last of these examples provides search

queries that allow individuals to locate potentially vulnerable servers based on information collected from published security advisories. Information obtained via dorking can be valuable for those conducting penetration testing as well as malicious actors looking for vulnerabilities or exposed information.

Potential Google Dorking Objectives

If Google dorking is used as part of a security teams approach to network defense it can help uncover security weaknesses or vulnerable data before those weaknesses are discovered by malicious actors. For malicious actors, Google dorking is sometimes the first step in a larger campaign as it can be used to scan the web for targets of opportunity that have left information openly available somewhere within their site. Dorking can also be used to conduct reconnaissance against targeted sites prior to an attack. While most Google dorks search the web as a whole, entering the operator *"site:examplesite.com"* at the end of the query string will limit its scans to a single domain name. One of the key features Google dorking provides malicious actors over other forms of web scanning is that it allows a user to scan websites without ever sending packets directly to the target. This allows those with malicious intent to use Google's servers to obfuscate their activity. Some further detailed examples of the objectives which may prompt dorking searches include:

Dorking to Gather Sensitive Data for Subsequent Attacks

Information including passwords, email addresses or credit card data can be a valuable asset to future campaigns. For example, if an actor locates an excel spreadsheet containing usernames and passwords by entering a query such *as "filetype:xls intext:username"*, they can attempt to access various websites with those user accounts. Passwords may also be added to an attacker's password cracking list for dictionary style attacks.

> *A dictionary attack is a method in which actors attempt to crack passwords by using words or phrases with a higher possibility of success[16]. These words can come from many sources including the words in a dictionary or passwords that have been obtained through previous data leaks. In addition, these attacks are generally performed through an automated process rather than manually attempting each password.*

Actors who locate email addresses associated with a targeted organization or agency using a query like *"filetype:xls intext:@email.com"*, can use those emails in a future spearphishing campaign.

> *Spearphishing is a form of email fraud in which actors send specially crafted malicious emails to targeted individuals or organizations and attempt to portray themselves as a trusted source that victims normally receive emails from.[17]*

Dorking to Find Sites Vulnerable to SQLi

Google can also be used to help an individual locate websites potentially vulnerable to structured query language injection (SQLi) attacks.[18] To begin the process, a user could enter dork queries that search for websites using the PHP scripting language. Some of these include:

inurl:index.php?id= inurl:trainers.php?id= inurl:buy.php?category= inurl:article.php?ID= inurl:play_old.php?id= inurl:declaration_more.php?decl_id= inurl:pageid=	inurl:games.php?id= inurl:page.php?file= inurl:newsDetail.php?id= inurl:gallery.php?id= inurl:article.php?id= inurl:show.php?id= inurl:staff_id=	inurl:newsitem.php?num= andinurl:index.php?id= inurl:trainers.php?id= inurl:buy.php?category= inurl:article.php?ID= inurl:play_old.php?id= inurl:declaration_more.php?decl_id=

According to one security researcher, PHP sites are a valuable as they can be established by everyday users using content management systems (CMS) like WordPress, and often contain valuable data. After an individual enters these queries, they would then need to confirm if the sites returned do in fact contain a SQLi flaw. To do this a user could enter an apostrophe (‘) at the end of the URL. If the response from the server indicates a SQL error, the actor knows the site is potentially vulnerable to SQLi.[19]

Dorking to Find Sites Containing a Known Vulnerability

In some cases, an actor may already know how to exploit a particular vulnerability or misconfiguration. For instance, if an actor knows an exploit for vulnerability in WordPress, they could search *“inurl:wp-content/”* to locate sites that run that CMS.[20,21] From here, actors can compile a list of the sites returned and attempt to exploit each of them.

Dorking for Network Mapping and Port Scanning

Dorking can also be used to map a targeted network. In this case, an actor can make use of the *“site:example.com”* operator to locate all indexed domains and subdomains associated with a targeted entity to provide a clearer picture of a targets network. Adding additional site operator function can further refine results and may lead actors to vulnerable pages including user login pages. Dorks can also be used for port scanning in which actors search for sites that can be accessed through a specific port.

Dorking to Feed other Tools

Google dorking can be used in conjunction with other automated tools to allow actors with low skill levels to conduct successful campaigns. Actors that have compiled lists of potentially vulnerable websites through Google dorking can run those sites through other automated scanning programs like Havij or SQLmap which will then locate and exploit any identified vulnerabilities.

> *Havij is a database scanning tool which runs on the Microsoft Windows OS. Actors using the tool simply enter a targeted site’s URL into Havij’s user friendly graphical user interface (GUI) and click analyze. Havij can perform back-end database fingerprinting, retrieve database management system (DBMS) user names and password hashes, dump tables and columns, fetch data from the database, run SQL statements, and even access the underlying file system and executing commands on the OS.*
>
> *SQLmap is a freely available open source penetration testing tool written in python which automates the detection and exploitation of flaws vulnerable to SQL injection. Rather than a GUI, SQLmap uses the command line as its user interface. SQLmap is equipped with a powerful detection engine and broad range of features from database fingerprinting, data fetching from databases, accessing underlying file systems and executing commands on the OS via out-of-band connections.*

While these examples highlight some ways in which the information obtained through dorking can be used, it only grazes the surface of this methods potential.

Automated Dorking Tools

While databases like the GHDB provide users with various dorks they can use to locate desired information, there are also freely available web applications that grant users the ability to run automated scans using multiple dork queries. These tools can be useful to both security personnel and malicious actors as they provide a simple way to perform thousands of scans that could uncover potentially dangerous or damaging vulnerabilities. While some of these automated programs may exist for malicious purposes, there are legitimate automated scanners available for penetration testing and network defense.

The major difference among these programs is their use of Google's application programming interface (API).

The use of automated search tool is considered a breach in Google's user policy and can result in IP blacklisting. Google's policy states, "You may not send automated queries of any sort to Google's system without express permission in advance from Google. These automated queries include:[22]

- using any software which sends queries to Google to determine how a website or webpage "ranks" on Google for various queries;
- "meta-searching" Google; and
- Performing "offline" searches on Google."

Only tools developed with Google's application programing interface (API) are permitted to run automated scans. This API allows developers to use Google web search results on their websites or applications. However, programs that use Google's API are limited to 1,000 queries a day. With thousands of available dorks, this is a significant limitation.[23]

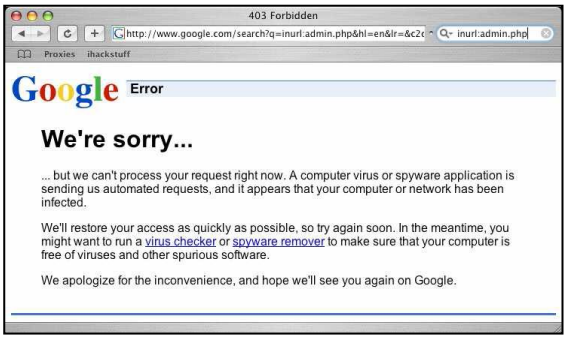

Figure 2: Dorking Detection

However, if individuals choose to perform scans with a tool that does not use Google's API, they run the risk that Google will block their IP address from accessing the search engine. Reports indicate that when Google's search engine detects what appears to be automated scanning, it presents users with a capcha screen to validate the user is in fact a human. When users hit this roadblock, further scanning becomes more difficult and may result in IP blacklisting/blocking. Therefore, these tools are not recommended for use by legitimate organizations unless users receive express consent from Google prior to scanning activity. These tools include Site Digger, which uses the API, and Goolag which does not use the API.

Site Digger

Site Digger is a scanning tool offered as a free download by a division of McAffee known as Foundstone. This tool, created with the Google API, provides its users a simple graphical user interface (GUI) which can be used to run various Google dorks against the sites of their choice. This tool is equipped with a list of predefined dorks but additional queries can be added from sites like the GHDB mentioned earlier. In addition, users can add their own queries directly to the program. Individuals can tailor their scans to run all signatures, groups of signatures targeting specific vulnerabilities, or individual signatures. After scanning is complete, the tool provides an HTML report listing its findings to include the problem URL, summary of issues, and detailed descriptions.[24] While it's not marketed as an all-encompassing solution to web security needs, it can be a

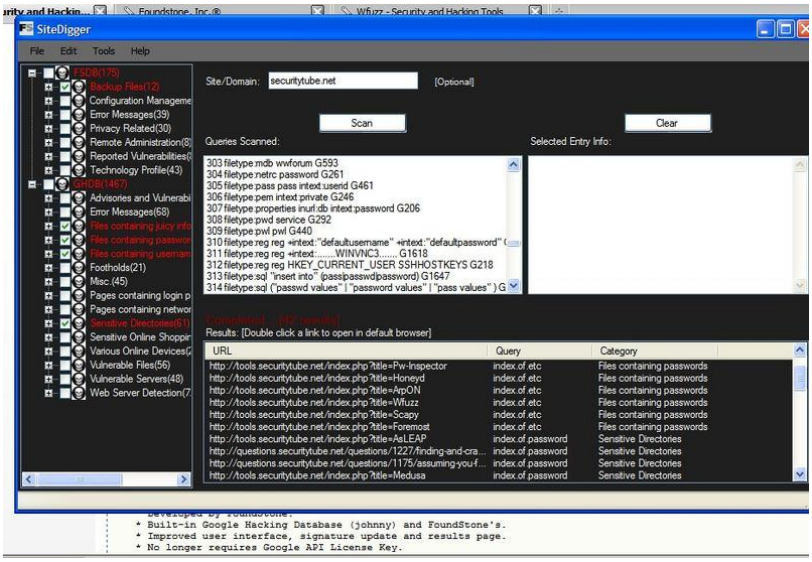

Figure 3: SiteDigger GUI

useful tool to help security experts quickly identify vulnerabilities that can be exposed through search engines like Google.[25]

Goolag

Goolag, which was released in 2008, allows individuals to run automated Google dork scans across the web or directed at an individual site.[26] Goolag applies the dorks maintained within the GHDB. When the user chooses their target, they select the dorks they wish to use and begin the scan. However, Goolag does not use Google's API.[27] Therefore, as mentioned above, those who use it run the risk of having their IPs blocked or blacklisted. Other Google scanning tools that exist which do not use the Google API include Gooscan, Athena and MaxiSploit Scanner. While these tools are often used for malicious purposes, they can be used by security professionals so long as they have coordinated appropriately and received permission to do so from Google.

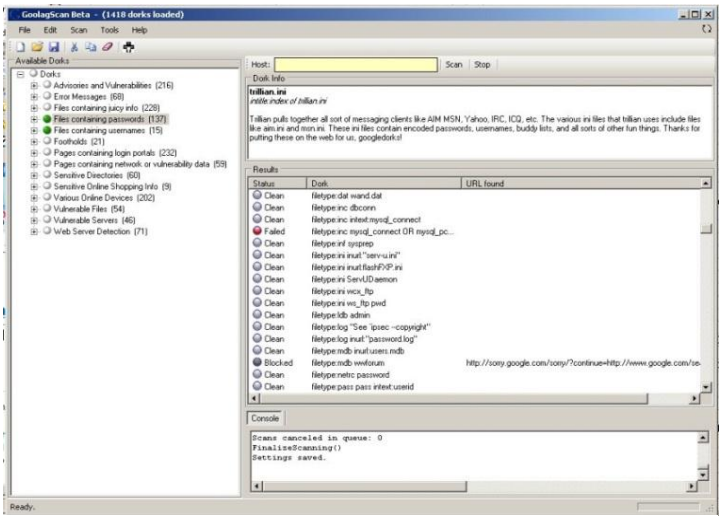

Figure 4: Goolag GUI

Google Dorking Incidents

As Google dorking may only be used as the initial step in an attack campaign and does not send packets to its intended victims, it is not always easy to determine whether it was employed in any early stage of an attack. However, some incidents that were believed to have involved Google dorking include the following:

- In an article published 16 October 2013, one security researcher explained how google dorking methods, were likely used to locate various websites running vulnerable versions of vBulletin. These vulnerabilities allowed attackers to add new administrator accounts by sending a specially crafted message to a site. The researcher was able to locate the sites affected by this campaign by using dorking techniques they believed to have been similar to those used by the attacker.[28]

- In August 2011, Google dorking was reported to have been used to uncover personally identifiable information (PII) for approximately 43,000 Yale University faculty, staff, students and alumni. This was accomplished when actors located an unprotected File Transfer Protocal (FTP) server.[29]

Defending Against Dorking

In response to the amount of malicious activity that makes use of Google dorking techniques, researchers have created defensive tools including the Google Hack Honeypot (GHH).[30]

> *A honeypot is a system made to look vulnerable so as to attract malicious activity. This activity is then monitored and recorded to provide researchers with pertinent data that can be studied and analyzed to help aid in understanding the tactics techniques and procedures used by various malicious actors.*[31]

Much like a regular honeypot, the GHH can be implemented on an individual's website and is designed to provide security researchers or site administrators a method of security reconnaissance. The tool is

powered by the Google search engine index and the GHDB and gives insight into how malicious actors are using Google dorking techniques against sites in an attempt to locate vulnerabilities or unprotected files.[32] This information can be used to notify service providers of the malicious activity observed originating from their networks, to denying future access, or used for statistical analysis.

In addition, security experts and Google dorking pioneer Johnny Long, provide the following steps web administrators can take to protect potentially sensitive information accessible via one of the thousands of Google queries existing in the GHDB.

- The first and most obvious recommendation is to never put sensitive data on the web. Even if data is placed on a website temporarily, it's possible that it may be forgotten or that it could remain on the site long enough for Google's crawler to locate and index it. Instead, share sensitive information using secure communications like secure shell (SSH) or encrypted email.

- Security personnel should use Google dorking against their own site to identify vulnerabilities that may be discovered or exploited by malicious actors by making use of the multitude of existing automated Google dorking scanning programs. However, as noted above, when using a program that utilizes the Google API, scans will be limited to 1,000 a day. If security personnel use a scanning service that does not utilize the Google API it is strongly recommended to get advanced permission from Google. Also, as new dork queries are continually added to repositories such as the GDHB, individuals should continually monitor these resources and update their own lists to remain current. In addition, some commercially available web scanning tools have Google dorking capabilities built in. These scanning tools can run Google dork queries obtained from the GHDB against the crawled pages of the website being scanned.

- As dorking is made possible through search engines use of site indexing, some defenders may wish to take steps to remove their sites from this index or ensure certain sites never become indexed. There are a few ways this can be accomplished. Methods include deleting pages from Google's index and using robots.txt to stop Google from crawling pages on a website.

    - For sites which an individual has verified ownership, Google provides webmaster tools which can be used to perform various functions including removing an entire site, individual URLs, directories, and cached copies of a website from Google's index.[33] These options are available in the diagnostics tab within the Webmaster tool available at https://www.google.com/webmasters/tools/home?hl=en and clicking the URL removals option.

    - Additional options include placing a NOINDEX meta tag in the header of the page an individual wishes to remove from Google's index. When Google crawls the site, this tag should be picked up.

    - However, as this may take weeks to occur, users can also access the "Crawl" menu in their Webmaster Tool and select "remove page from index". From here, users simply add the page they want to delete. This will only work if the user has already placed the NOINDEX tag in the header of the same site or page. This second step will ensure the site is removed more quickly.[34]

- Lastly, Web administrators can use the robots.txt file to inform Google what sites or pages it does not want crawled and indexed. This file can be created by anything that produces a text file

including notepad, wordpad, and textedit and should be placed in the top-level directory of the web server.[35] Below are some examples of robots.txt file entries and what they do.[36]

User-agent: *
Disallow:
Result: As nothing is represented after (:) robots or "crawlers" can view all files

User-agent: *
Disallow: /
Result: This entry will keep all robots out of all directories as directory paths use (/)

User-agent: *
Disallow: /example/
Result: stops robots from crawling specific directories. In addition, not including the trailing / will stop robots from crawling files as well.

Points of Contact

For all inquiries pertaining to this product, please contact the NCCIC Duty Officer or NCCIC O&I Analysis at NCCIC@hq.dhs.gov or 1(800) 282-0870.

Can I share this product?

- Recipients may share TLP: GREEN information with peers and partner organizations within their sector or community, but not via publicly accessible channels.

References

[1] http://www.alexa.com/topsites
[2] http://www.statisticbrain.com/google-searches/
[3] http://kinetickknowledge.com/google-crawl-and-business-blog/
[4] http://www.howstuffworks.com/google-algorithm.htm
[5] http://computer.howstuffworks.com/google-algorithm1.htm
[6] http://www.statisticbrain.com/google-searches/
[7] http://techcrunch.com/2008/07/25/googles-misleading-blog-post-on-the-size-of-the-web/
[8] http://www.statisticbrain.com/total-number-of-pages-indexed-by-google/
[9] http://www.acunetix.com/websitesecurity/google-hacking/
[10] http://www.techrepublic.com/blog/it-security/google-hacking-its-all-about-the-dorks/
[11] http://www.bruceclay.com/advancedsearches.htm
[12] http://www.alexa.com/topsites
[13] http://resources.infosecinstitute.com/google-hacking-for-fun-and-profit-i/
[14] http://searchsecurity.techtarget.com/tip/Protect-your-business-from-a-Google-hack
[15] https://www.blackhat.com/presentations/bh-europe-05/BH_EU_05-Long.pdf
[16] http://www.webopedia.com/TERM/D/dictionary_attack.html
[17] http://www.fbi.gov/news/stories/2009/april/spearphishing_040109
[18] https://www.udemy.com/blog/sql-injection-tutorial/
[19] https://www.udemy.com/blog/sql-injection-tutorial/
[20] http://searchsecurity.techtarget.com/definition/Google-hacking
[21] http://hackertarget.com/google-dorking-wordpress/
[22] http://www.google.com/terms_of_service.html
[23] http://pdf.textfiles.com/security/googlehackers.pdf
[24] http://www.infoworld.com/d/security-central/sitedigger-unearths-web-information-leaks-721
[25] http://www.infoworld.com/d/security-central/sitedigger-unearths-web-information-leaks-721
[26] http://www.enterprisenetworkingplanet.com/netsecur/article.php/3771686/Use-Goolag-to-Find-Your-Inner-Dork.htm

[27] http://www.watchguard.com/RSS/showarticle.aspx?pack=RSS.Goolag.Scanner
[28] http://www.scmagazine.com/vulnerability-in-vbulletin-grants-website-admin-privileges/article/316622/
[29] http://content.usatoday.com/communities/technologylive/post/2011/08/google-hacking-exposes-large-caches-of-personal-data/1
[30] http://ghh.sourceforge.net/
[31] http://ghh.sourceforge.net/
[32] http://ghh.sourceforge.net/userfaq.php
[33] http://googlewebmastercentral.blogspot.com/2007/04/requesting-removal-of-content-from-our.html
[34] http://www.pearanalytics.com/blog/2010/how-to-remove-pages-from-googles-index/
[35] http://www.robotstxt.org/faq.html
[36] http://resources.infosecinstitute.com/defending-from-google-hackers/