DEEPLIGHT

SHINING A LIGHT ON THE DARK WEB

AN INTELLIAGG
REPORT: 2016

Intelliagg

CONTENTS

GLOSSARY OF TERMS

ALGORITHM – a step-by-step set of mathematical operations that perform functions such as calculation, data processing or even automated reasoning.

WEB BROWSER – a software application such as Firefox, Internet Explorer, Safari and Google Chrome that retrieves and presents information found on the web.

COLLECTION SOFTWARE – a 'spider' or software application that crawls through the web following links in order to compile an index of its pages.

DARK WEB – web content that cannot be reached without the use of specialized encryption software, which means it is not indexed by search engines and those who use it can do so with complete anonymity.

DEEP WEB – everything on the web that exists behind firewalls and paywalls (including the 'Internet of things') that cannot be 'indexed', that is, found by search engines using keywords and metadata.

ENCRYPTION – the process of encoding messages or information in such a way that only authorized parties can read it, thereby preventing anyone who intercepts an intended communication from understanding its content.

INDEXING – any method of compiling an index of the contents of a website or of the Internet as a whole, which is crucial to the process of searching it.

MACHINE-LEARNING – the use of algorithms that can 'learn' about data from a model they are given as an example, enabling them to then make decisions rather than just follow simple programmatic instructions.

MACHINE-LEARNING INTELLIGENCE CLASSIFICATION SYSTEM – an array of complex algorithms that are 'trained' by humans then sent off to classify large amounts of data automatically.

ROUTERS – networking devices or software that steer data between end users.

TOR NETWORK – a network of routers that adds encryption to conceal a web user's location and usage so that these are resistant to surveillance and hence are truly anonymous. The domain names of these hidden sites all end in '.onion' and they are only accessible by using a Tor browser. TOR stands for 'The Onion Router'.

INTRODUCTION

THIS REPORT HAS BEEN COMPILED BY EXPERTS AT **INTELLIAGG**, A LEADING EUROPEAN PROVIDER OF INTELLIGENCE ON CYBER THREATS, AND ITS U.S. COUNTERPART, **DARKSUM**.

OUR RESEARCH:

- Maps what is on the dark web for the first time;
- Reveals that the dark web is much smaller than commonly thought;
- Reveals that much of the dark web's content is entirely legitimate, with only half of it containing content likely to be illegal under U.K. and U.S. law.

Using complex algorithms that 'learn' as they go, we have mapped the 'dark web' – a poorly understood realm of cyberspace that has come under growing scrutiny for posing unprecedented security threats to organizations. Sites on the dark web cannot be reached without the use of specialized encryption software, enabling those who use it to operate with complete anonymity.

Although partial efforts to map the dark web have been attempted, Intelliagg and Darksum are the first to employ hard data to paint a comprehensive picture of its contents. We are experts on the dark web: we monitor it on behalf of our clients to provide them with actionable intelligence about emerging threats to their business.

We believe it is important for the public to gain a better understanding of the contents of the dark web in order for there to be a proper debate about its nature, dangers – and potential benefits. Misunderstanding about the dark net is rife, and has been fuelled by often misleading media coverage. This, in turn, has influenced policy debates based on incorrect assumptions and hyperbole.

Intelliagg and Darksum believe there is a need for more transparency about the contents of the dark web if it is to make a positive contribution to the broader web ecosystem. To that end, and by issuing regular mapping reports of this kind, we will provide thought leadership informing the wider debate about the dark web's future.

WHAT IS THE DARK WEB?

The growth in use of the dark web is analogous to that of the early Internet in the 1990s. Adopters of the Internet at that time included curious students, marginalized groups and criminals, who quickly grasped its potential for communicating and doing business.

While the Internet has since become a vast public realm, sites on the dark web and the individuals and groups that use it remain fully anonymous. This can, of course, lend itself to illegal activity – from the sale of stolen credit card information, for example, to hate crime and the exchange of extreme pornography. But it can also be of importance to freedom of speech as a refuge from unwelcome surveillance. Some journalists, for example, use the dark web to investigate sensitive matters in the public interest.

However, a lack of understanding about the dark web helps to explain why there has been a common misconception in press coverage that it is vastly bigger than the Internet. This is often because the dark web is confused with what is known as the 'deep web'. The deep web includes everything behind firewalls and paywalls as well as the IOT (Internet of things) that cannot be 'indexed', i.e. found by search engines using keywords and metadata.

Whereas the Internet comprises about 1.3 billion indexed web sites, the deep web is thought to be up to 500 times larger than that.

All internet traffic is directed between users by 'routers', networking devices or software that steer data. The dark web exists on an extra layer of routers that together comprise the dark network. The largest dark network is what is called the Tor network. Tor stands for 'The Onion Router' and originated in the U.S. security establishment in the 1990s. It adds encryption to conceal a web user's location and usage so that these are resistant to surveillance and hence are truly anonymous. The domain names of these hidden sites all end in '.onion' and they are only accessible by using a Tor browser.

Our research has mapped the Tor network's services, which until now have not been comprehensively catalogued. Our experts have discovered that the Tor-based dark web is very limited in size, with only approximately 30,000 '.onion' addresses active at any one time.

RESEARCH METHODOLOGY

We compiled our census of the dark web using the Darksum 'collection software', a 'spider' or software application that crawls through the web following links in order to compile an index of its pages, and Intelliagg's 'machine-learning intelligence classification system' – complex algorithms that are 'trained' by humans then sent off to classify data automatically.

DISCOVERY

First, the team compiled a list of hidden dark web addresses from several different sources: by using spiders; by consulting existing lists of Tor links compiled over time by individuals; and by monitoring the Tor network itself (further information available on request). On discovery of a previously unseen '.onion' address, it is a simple matter to see if this newly discovered hidden service responds on a given 'port', the dedicated access point into a computer from a network. For this study, only the web ports 80 (http) and 443 (https) were investigated.

INDEXING AND DATA RETRIEVAL

Once a hidden service was discovered, our automated web browser indexed it based on its contents. Many sites on the dark web have an obvious interest in not being indexed, so the browser emulates a human employing a Tor browser. The data obtained in this way was fed into our automated classification system.

SITE CLASSIFICATION AND AUTOMATION

Our classification system was 'trained' using data that had been classified manually from 1,000 sites on the dark web. It proceeded to classify the remaining data automatically without human supervision. This automated method proved to be 94% as accurate as it would have been had this process been entirely done by hand, meaning that nine times out of 10 our algorithms came to the same conclusion as an experienced analyst.

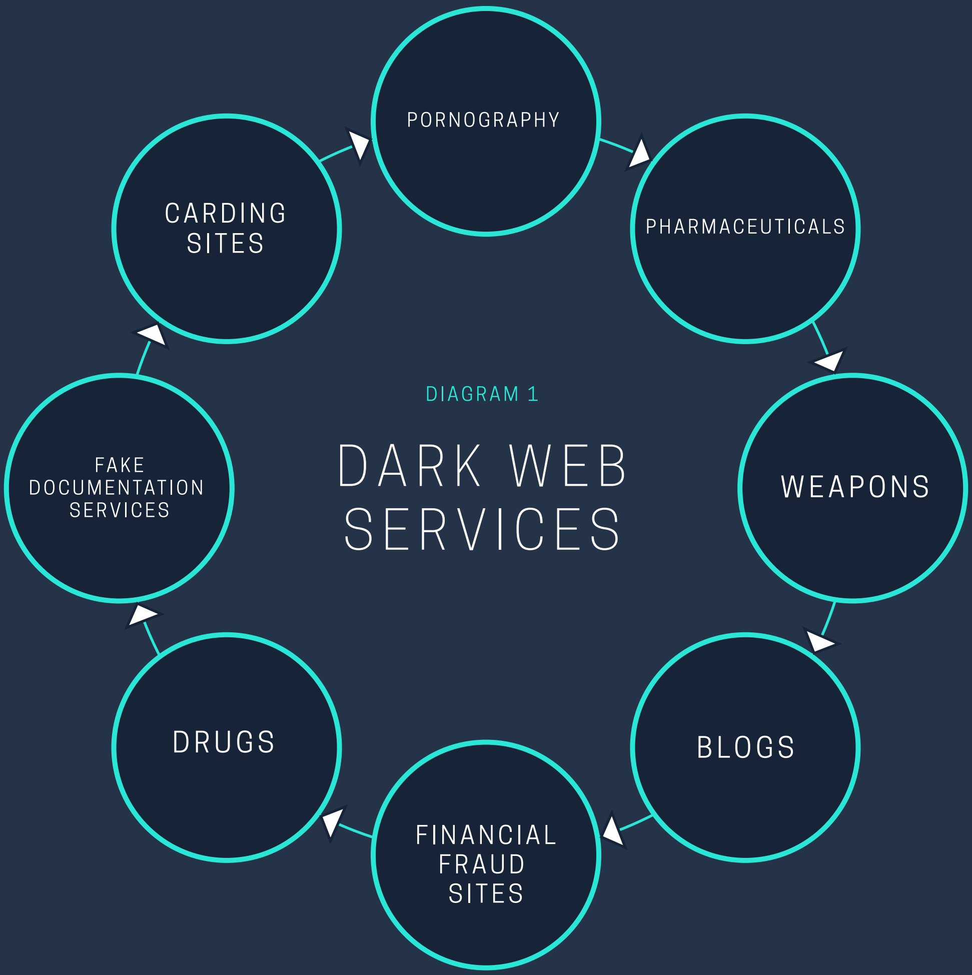
DATA OUTPUT

Data from this 'machine-learning' exercise was collected over two weeks in February 2016 (raw data is available on request to qualified parties interested in further research).

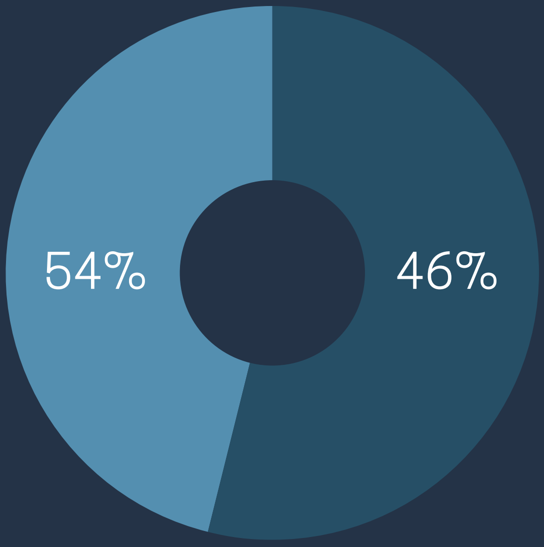
DARK WEB CATEGORIES

The content of the dark web observed in the course of our research was noticeably different from the content of the Internet. A high proportion of content that would not be illegal under the laws of the U.K. or U.S.A. if it were to appear on the Internet itself originated among social groups in the developed world with interests that are best described as 'niche'. This is especially pronounced when it came to pornographic and political content, which often consists of material from groups that are or feel persecuted or socially ostracized, such as 'fetishists' and conspiracy theorists.

Some dark web content is clearly illegal in that it would break U.K. or U.S. law if it was accessible on the Internet. The illegal contents of Tor hidden services spans the full range of criminal activity, from certain forms of pornography and the retailing of drugs and weapons to inciting violence and hate speech.



MAPPING OVERVIEW



● INACCESSIBLE ● ACCESSIBLE

A total of 29,532 '.onion' addresses were identified during the sampling period. Of these, fewer than half were accessible at some point during this period. The remaining 54% (which were not analysed further) were probably only up on the dark web for a very short period of time. This could be for many reasons: commonly that they were addresses relating to 'command and control' servers used to manage malicious software, chat clients, or file-sharing applications.

LANGUAGES DETECTED

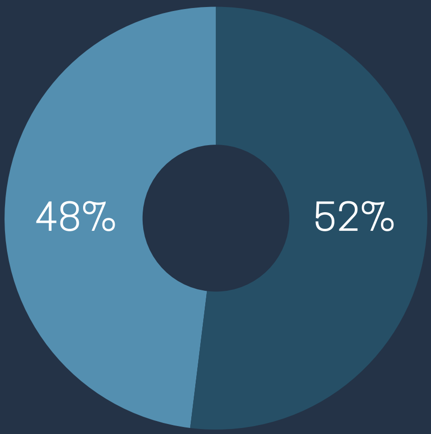
Material in a total of 32 different languages was detected and classified by our method. Not surprisingly, the vast majority of information on the hidden services network is in English, followed by German and Chinese.

ENGLISH:	GERMAN:	CHINESE:
76%	4%	3.7%

Other detected languages representing 3% or less, presented in descending order:

French
Russian
Spanish
Dutch
Quechuan
Portugese
Italian
Tai-kadai
Polish
Norwegian
Danish
Swedish
Finnish
Armenian
Kinyarwanda
Maltese
Irish
Greek
Bulgarian
Luxembourgish
Kurdish
Korean
Amharic
Georgian
Malay
Kannada
Romanian
Hebrew

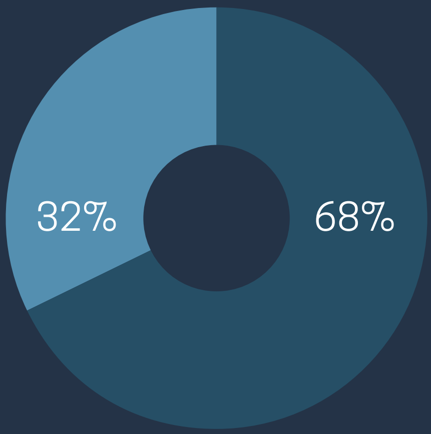
DEMOGRAPHICS & ANALYSIS



LEGAL/ILLEGAL

Of the total number of sites analysed with our 'machine-learning' classification method, about half of the contents were classified as legal under U.K. and U.S. law.

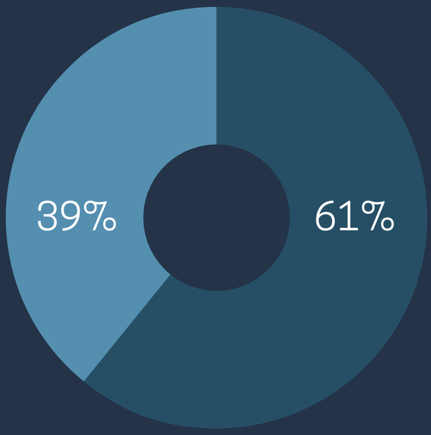
● ILLEGAL ● LEGAL



MANUAL ONLY

After classifying the content of 1,000 sites manually, our analysts found that about 68% of this would be illegal under U.K. and U.S. law.

● LEGAL ● ILLEGAL



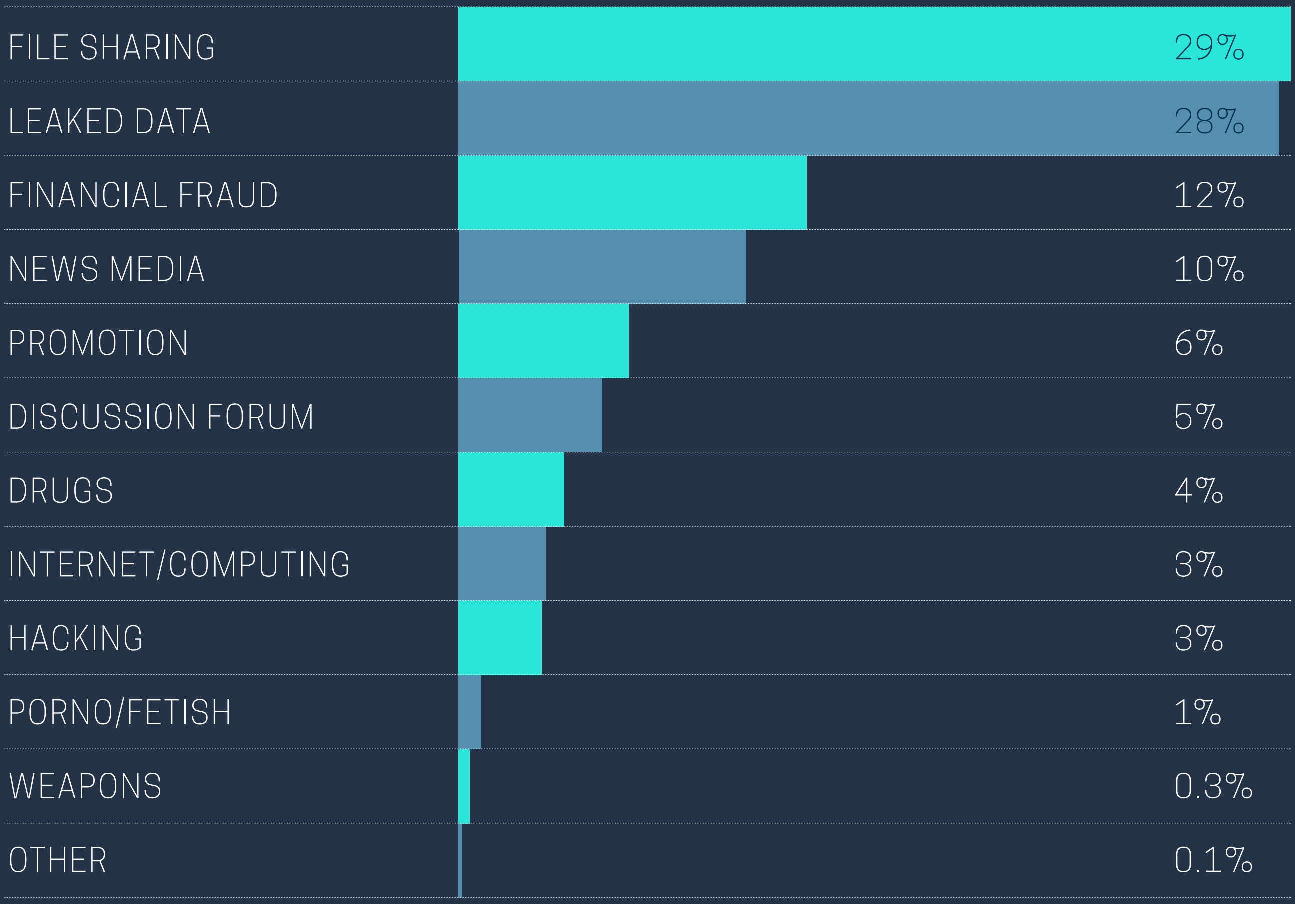
LINKED/UNLINKED

Of the total sites analysed, 61% were located through links in material analysed in the dataset, whereas the remaining 39% were discovered through other means. This indicates that a significant portion of the sites cannot be found easily. However, this analysis does not take into account whether a site could be linked to from sources on the indexed Internet.

● UNLINKED ● LINKED

CATEGORIZATION OF RESPONSIVE SITES

Filesharing, leaked data, and financial fraud made up large categories of content in the sample. A total of 29,532 ‘.onion’ addresses were identified during the sampling period. Of these, fewer than half were reachable at some point during this period, with the remaining 54% probably only up on the dark web for a very short period. This underlines an important characteristic of the dark web, which is transitory and changeable.



CONCLUSION

THIS GROUNDBREAKING RESEARCH BY **INTELLIAGG** AND **DARKSUM** CONSTITUTES THE FIRST COMPREHENSIVE EFFORT TO MAP THE DARK WEB AND OFFERS VALUABLE INSIGHTS INTO THIS POORLY UNDERSTOOD REALM OF CYBERSPACE.

Our research shows that:

- With just 30,000 sites, the dark web is much smaller than commonly thought – and only a small fraction of the size of the Internet;
- half of the dark web's content is entirely legitimate, with only about 15,000 sites that contain material that would be illegal under U.K. or U.S. law.

Tor hidden services have a number of properties that make them useful to those worried about 'man in the middle attacks' on the Internet – where a communication is changed or intercepted, something sometimes done by government agencies. However, this legitimate use of Tor has yet to be employed widely. Nonetheless, we expect it to increase over time and, eventually, the technologies developed by the Tor project to become mainstream, with positive benefits for security and privacy.

Intelliagg and Darksum recommend that efforts are made to further increase transparency about the contents of the dark web in order for it to become the positive force for security and privacy that we believe it can. Ironically, for the dark net to prosper it needs to become lighter.

We will continue to monitor the evolution of the dark net regularly, and this report represents the first of a series to identify trends and developments.

ABOUT US

FOUNDED IN 2011, INTELLIAGG PROVIDES INTELLIGENCE AND WORKS WITH ORGANIZATIONS IN ORDER TO CONTROL OR EVADE DATA LOSS, REPUTATIONAL DAMAGE AND TARGETED CYBERCRIME.

We collect and aggregate information that will help our clients deal with cyber threats throughout the entire online world, regardless of language, source or classification, and we provide a suite of professional services to manage their responses to incidents.

Intelliagg

Produced in association with Onyx.

For further information and interviews please contact:

Anne Cantelo
onyx-pr.com
anne@onyx-pr.com
+44 (0)20 7048 2700

ONYX