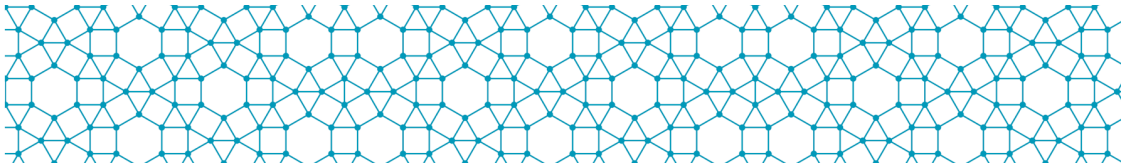
TERBIUM LABS

Separating Fact from Fiction:

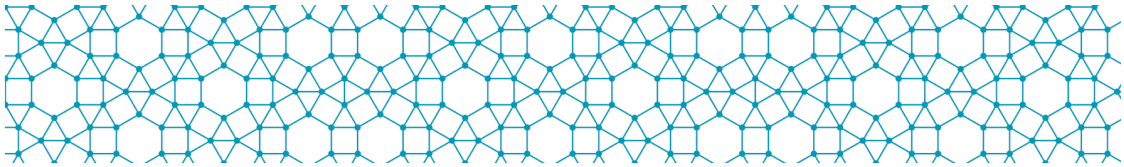
THE TRUTH ABOUT THE DARK WEB



About This Study

As a dark web data intelligence company, we often get the question "What's actually on the dark web?"

We're glad you asked.



About Terbium Labs

Terbium Labs is a dark web data intelligence company based in Baltimore, Maryland. Terbium runs **Matchlight**, the world's first fully private, fully automated dark web data intelligence system. Matchlight alerts customers the moment a trace of their information appears online where it shouldn't, without customers ever needing to reveal their information to anyone – not even Terbium.

Matchlight relies on two key enabling technologies. The first is **Data Fingerprinting**, a patented technology that uses a fuzzy hashing protocol to generate one-way cryptographic hashes. Those hashes, or Data Fingerprints, are generated on our clients' systems and are the only information ever sent to Terbium. The second is our **large scale crawler**, constantly indexing the dark web and generating billions of new fingerprints of dark web data every day.

With these technologies at our disposal we set out to answer the question we get asked most often: **What is actually on the dark web?**

This report contains the results of our research.

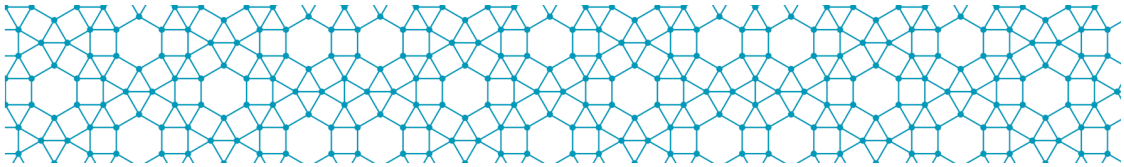
About The Authors

Dr. Clare Gollnick, Chief Data Scientist

Clare Gollnick is the Chief Data Scientist at Terbium Labs. As a statistician and engineer, she designs the algorithms that direct Terbium's automated crawl of the dark web and leads the crawler engineering team. Before turning her attention to the dark web, Clare was a neuroscientist. Her academic publications focus on information processing within neural networks and validation of statistical methods. Clare holds a PhD in biomedical engineering from Georgia Tech and a B.S. in bioengineering from UC Berkeley.

Emily Wilson, Director of Analysis

Emily Wilson is the Director of Analysis at Terbium Labs. Emily is responsible for managing a team of analysts, while tracking industry news and trends among actors on the dark web, including specific breach operations, popular targets, and the appearance of new sites for trading or discussing stolen data. Emily provides analysis on the appearance of fraud, drugs, weapons, extremism, and other information appearing on the dark web. Emily has a background in Eastern European politics, and holds a Bachelor of Arts in International Relations from The College of William & Mary.



Introduction

The dark web is a mysterious place.

Most people never access the dark web. For those of us that do, there are few resources that allow for exploration or meticulous study. Even seemingly straightforward questions like what kind and how much of different types of content exist on the dark web do not have reliable answers. Past industry reports have superficially addressed these questions, but often include numbers and conclusions without explanations of where the data came from or how those numbers were computed.

In this report, we aim to bring a more rigorous and scientific approach to analyzing content on the dark web. Just like a scientific paper, our report comes with a complete description of methods and an explicit discussion of the limitations of our approach. By being transparent, we provide an opportunity for others to review and criticize our data and to interpret our research with a level of skepticism that would be expected of a scientific study.

This is the Terbium Labs' ethos. Research is hard. Data is not necessarily objective or accurate. Methodology is everything. The details matter.

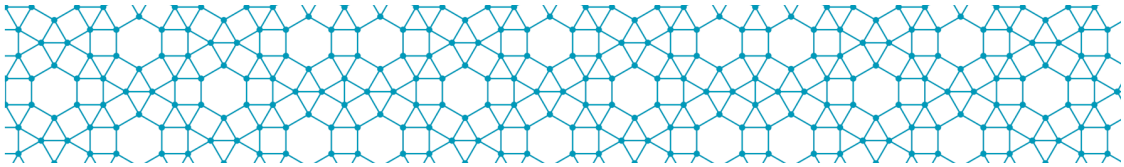
Why Is This Type of Research So Hard?

Selection Bias

There is no definitive list of all sites of the dark web. Work in this space often relies on human analysts to choose which sites to review, or depends on known lists of Tor Hidden Services curated by anonymous but interested parties. As we learn over and over again each election season, the quality of a study or a poll is defined by the quality of its random sample. Using a human-curated list instead of a random draw is equivalent to only polling viewers of a single TV channel on their preferred candidate. The results will be biased and inaccurate. By using a random draw from the population of Tor Hidden Services found by our automated big-data crawler, our approach minimizes the effect of selection bias. To learn more about how we optimized our sampling methodology to ensure our sample was representative, check out our Methodology section (page 28).

Definitions Matter, Even Humans Do Not Agree

Classification of content or topic is not as simple as it may seem. Humans disagree on the definitions of categories, and disagree on what content belongs in each category. Even the most rigorously designed classification guidebook cannot take into account every possible outlying content type. At what point do weapons become weapons of mass destruction? When does political or religious rhetoric cross into extremism? That's why simple statements like "the algorithm was 90% accurate" are misleading. Changing these category definitions will dramatically



alter the quantitative results. Most sites do not fit securely within one category or another.

There are multiple other definitions that matter in this study, including defining the dark web itself. We discuss this in more depth later in the report (skip to Definitions Matter on page 30)

A Quick Overview

Complete methodology is included at the end of the report. But here's what you need to know to get started.

We reviewed a sample of 400 URLs from a single day in our automated crawler's history. URLs (as opposed to domains) were used as the independent unit within the sample. The sample was selected at random from the population of URLs known to our unrivaled big-data infrastructure that crawls the dark web continuously. For each of these URLs, a team of analysts classified the page content into one of 15 predefined categories.

We classified sites into the following categories (skip to Appendix on page 35 for detailed definitions):

- Legal
- Explicit
- Drugs
- Pharmaceuticals
- Fraud
- Multiple Categories (Illicit)
- Falsified Documents & Counterfeits
- Exploitation
- Hacking & Exploits
- Weapons
- Extremism
- Weapons of Mass Destruction
- Other Illicit Activity
- Unknown/Site Down
- Downloadable File

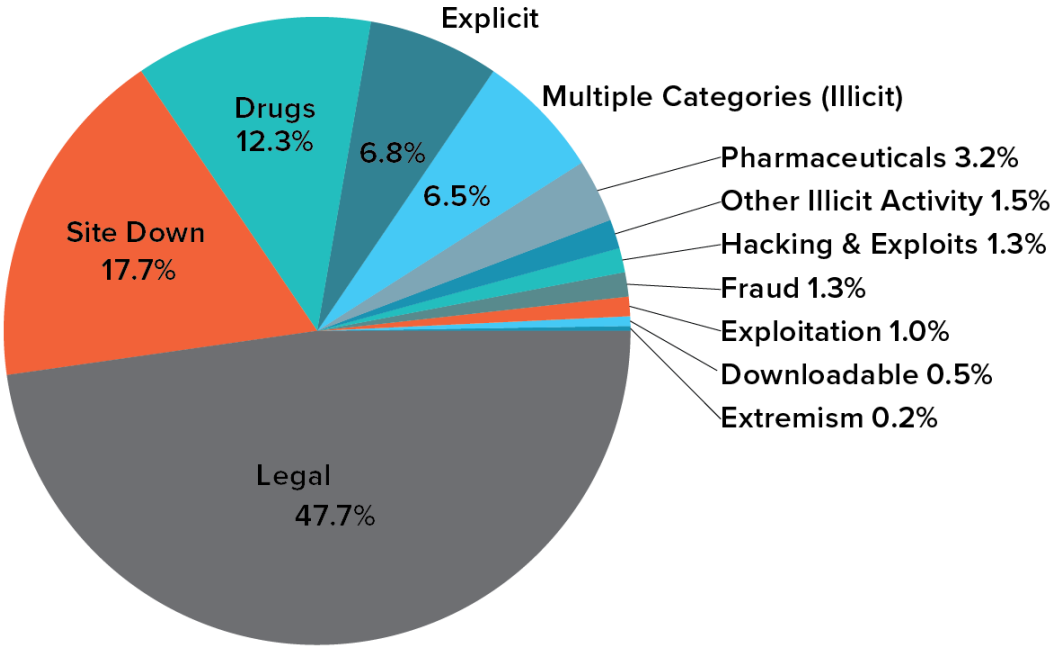
We provide summary statistics (percentages) for each of these categories. For those so inclined, estimates of error and confidence intervals for each measure are included at the end of the report (skip to Confidence Intervals on page 31).

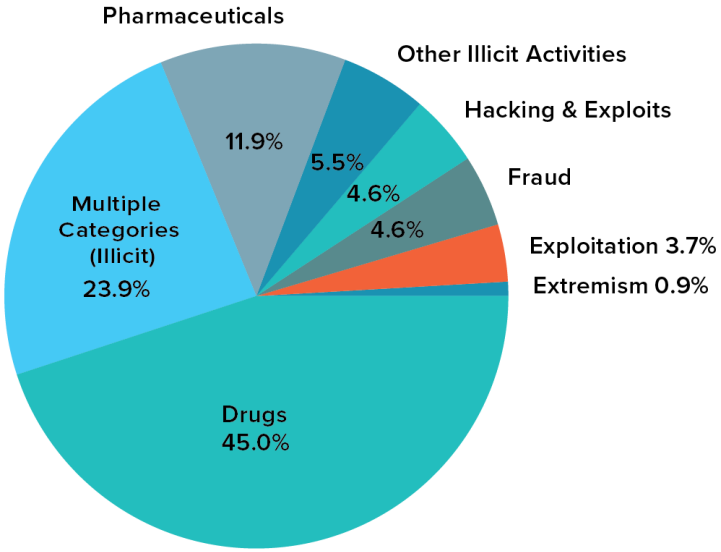
We consider each category in a separate section. Within each section, we sliced our data in a couple of different ways. First, we computed percentages relative to the full data set: numbers including both legal and illegal content. When it comes to the dark web, we know that most of the mystery and intrigue stems from the illicit material. We also compute the percentages with respect to only the total illicit content.

Second, we split our analysis based on URL counts and domain counts. These numbers are correlated, but allow for some nuance of interpretation. Should a big domain with many millions of URLs be more important that a single-page domain? It depends on who you ask. We sliced the data both ways.

TOTAL CONTENT (BY URL)

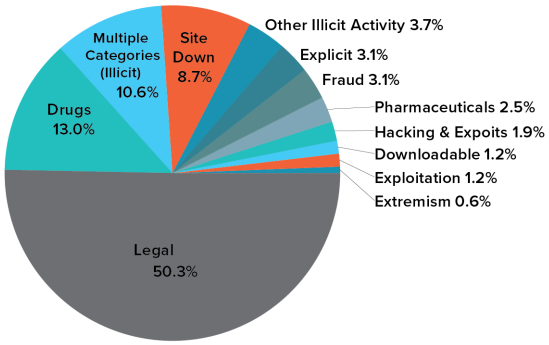
This graph contains total findings for all content, measured by URL.

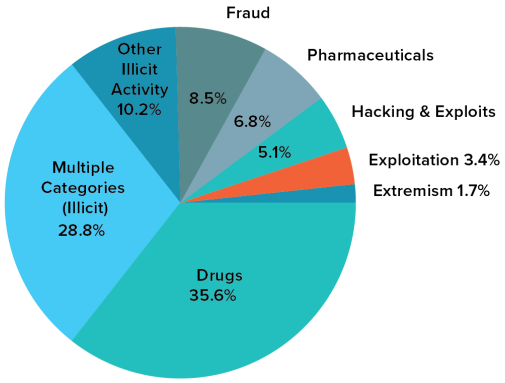




ILLICIT CONTENT (BY URL)

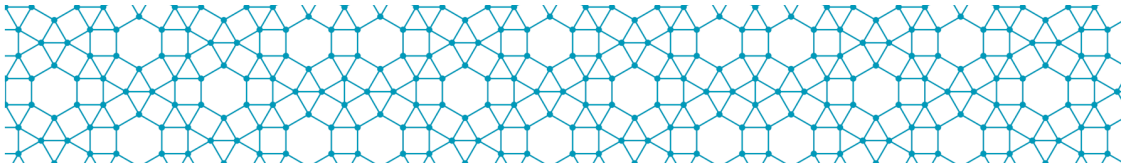
This graph is derived from the above graph but removes all reference to legal content.





CONTENT (BY DOMAIN)

These two graphs correspond to the graphs above, except instead of counting URLs we count domains. The graph on the left contains total results for all categories, including legal content. The graph on the right considers only illegal content.



Legal Content On The Dark Web

The dark web has a bad reputation because of its structural anonymity and emphasis on privacy.

We hate to break it to you: the dark web is mostly legal.

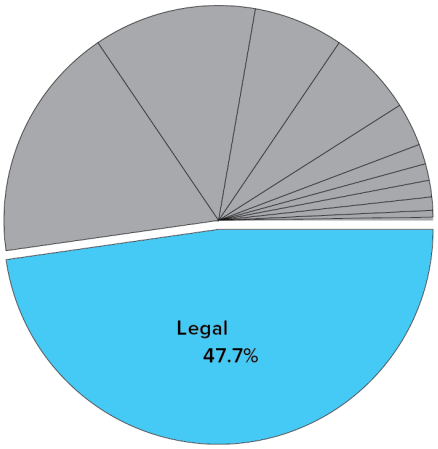
Legal content comprises 53.4% of all domains and 54.5% of all URLs in our sample.

Legal

We defined **Legal** as any activity or discussion that was not explicitly illegal. We also classified a page as **Legal** if the content on the page was legal, even if the broader site may have contained illegal information, goods, or services.

Findings:

Total by URL: 47.7%
Total Percent of Domains: 50.3%



What compromises that %?

The dark web receives a fair amount of negative attention because of the anonymity it provides. To outside observers, the desire for anonymity goes hand-in-hand with criminal activity, and many summaries of the dark web focus exclusively on this criminal activity. As our data shows, however, the majority of the dark web contains perfectly legal content.

What does legal content on the dark web look like? These Tor Hidden Services play host to Facebook, European graphic design firms, Scandinavian political parties, personal blogs about security, and forums to discuss privacy, technology, even erectile dysfunction. Anonymity does not equate criminality, merely a desire for privacy.

Most discussions of the dark web entirely gloss over the existence of legal content. As the numbers above indicate, legal content compromises over half of all of the domains sampled for our research, and nearly half of all pages we classified.

Scenes From The Dark Web

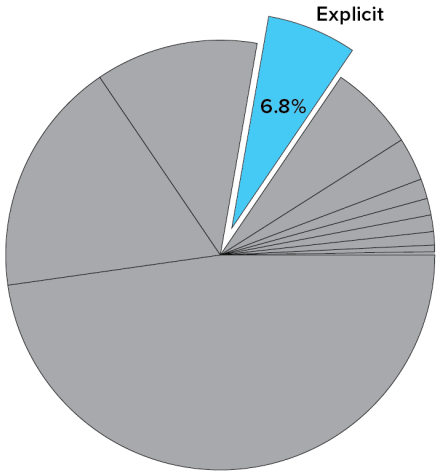
In case you didn't get enough of a fix from I Can Has Cheezburger, we submit "Tor Kittenz" for your consideration.

Explicit

Explicit includes any non-exploitation explicit content - effectively, pornography. We included **Explicit** as a category to account for the percentage of the dark web that contains legal pornography, available for free or for purchase on Tor Hidden Services.

Findings:

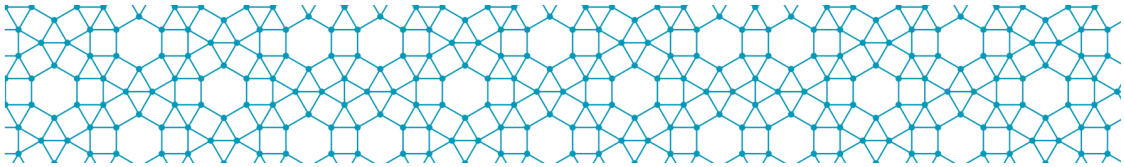
Total by URL: 6.75%
Total Percent of Domains: 3.1%



What compromises that %?

The dark web contains good old-fashioned porn. Not all explicit content on the dark web is exploitation. Perfectly legal photos, videos, and written material: one or more consenting adults, available for free or for purchase. The dark web is a natural extension of the internet and the desire for pornography does not end at the edge of the clear web.

The dark web also offers a certain illicit thrill to visitors who view Tor Hidden Services as a way to explore the "hidden" parts of the internet. These explicit sites will also receive a certain amount of traffic purely by having a functional, legal porn site available exclusively to people who are using the Tor browser.



Illegal Content On The Dark Web

Anonymity does facilitate crime.

Drugs dominate the illegal content on the dark web, followed by sites with a combination of multiple illicit materials.

Drugs constitute 44.5% of all illegal content on the dark web – 56% if you include pharmaceuticals.

Drugs

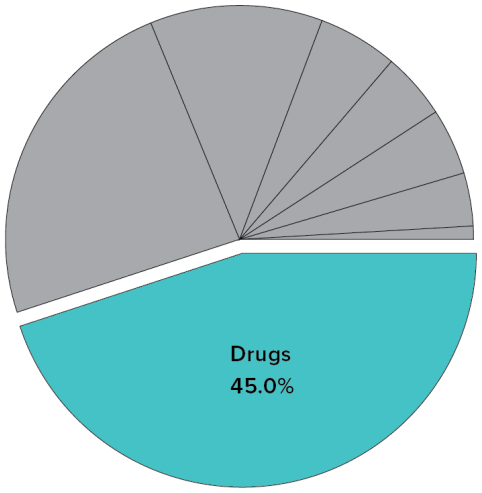
We defined **Drugs** as any non-pharmaceutical drug or substance bought or sold for recreational purposes. To provide a more detailed breakdown of the kinds of drugs available on the dark web, we separately classified any **Pharmaceuticals** available for sale as well. We include marijuana as a drug and not a pharmaceutical for the purposes
of this study.

Findings (total):

Total by URL: 12%
Total Percent of Domains: 13.04%

Findings (Legal excluded):

Total by URL: 44.95%
Total Percent of Domains: 35.59%

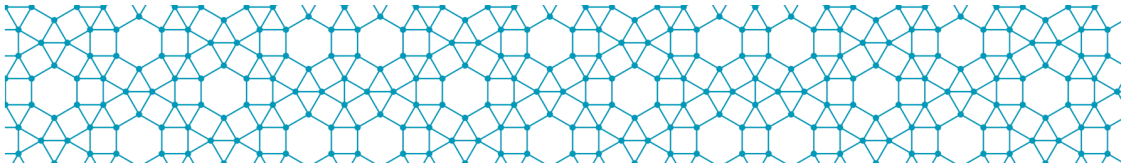


What compromises that %?

The dark web drug trade, if we can call it that, is far more organized and mundane than you might expect. Purchases take place on moderated and well-maintained markets where buyers can drill down into categories and subcategories in search of the right vendor or product.

The drug community is a self-regulating, even a self-policing group. Vendors rely on reviews and commentary to build their brand and drive orders. New vendors are quick to offer free samples, and the community is more than happy to test a new product. Reviews follow a standard template, where users rank the stealth, shipping time, purity, high, and overall experience with the vendor. The community quickly shuns know scammers as warnings pop up about disappearing packs or when buyers get less than half of what they paid for.

Beyond the markets, drugs are a frequent topic of conversation in forums and blogs, as we suspect might also be true on the clear web.



Scenes From The Dark Web

A vendor commenting on one buyer's meth review pointed out that you can - with a discerning sense of taste - tell what type of battery provided the acid used to cook a batch of meth. One popular battery brand apparently lends a taste of lime, coffee, and cigar.

Ask your sommelier about that one.

Pharmaceuticals

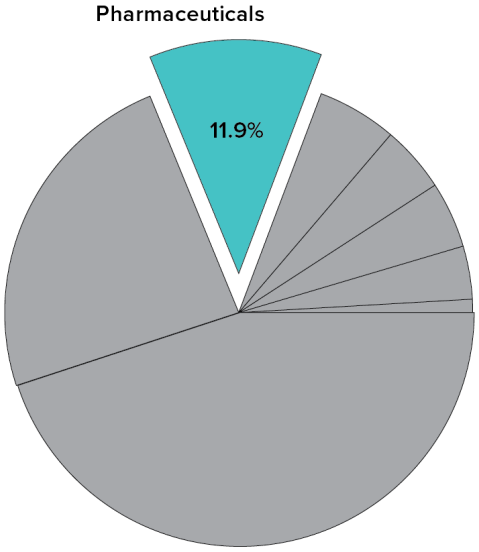
Pharmaceuticals include any kind of drug that a doctor might prescribe, excluding painkillers and their derivatives. For our classification, **Pharmaceuticals** include ADD/ADHD and anti-anxiety medications, even though these medications are often used recreationally.

Findings (total):

Total by URL: 3.25%
Total Percent of Domains: 2.48%

Findings (Legal excluded):

Total by URL: 11.92%
Total Percent of Domains: 6.77%



What compromises that %?

No prescriptions, unlimited refills, and no questions asked. Dark web pharmacies provide unfettered access to prescription medications, recalled over-the-counter drugs, and unregulated supplements.

Take, for example, human growth hormones. Through the dark web, buyers in United States have access to a host of steroids that have not met FDA approval, or are pending approval in the United States. Unlike recreational drugs, pharmaceutical vendors are more likely to invoke name brands or branded packaging in an effort to add legitimacy to their products, at least in part because pharmaceuticals are more difficult (if not impossible) for vendors to manufacture at home.

Fraud

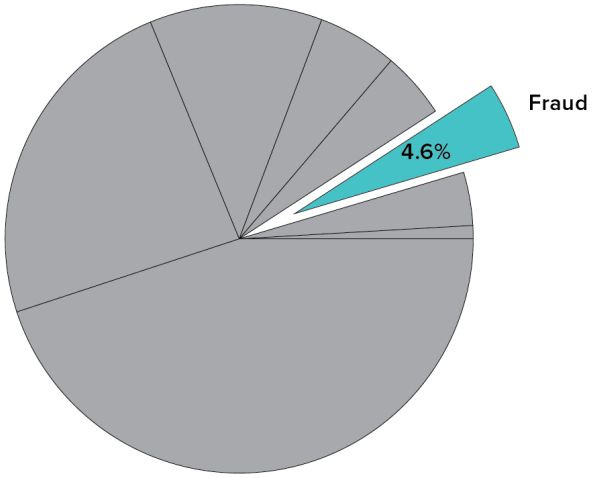
We defined **Fraud** as any kind of content based on personal information, credentials, or accounts. We distinguish **Fraud** from **Falsified Documents & Counterfeits** and **Other Illicit Activity** in order to capture instances where an **existing Falsified Documents & Counterfeits** was compromised, leaked, or sold for the purposes of committing fraudulent acts (this is separated from, for example, doxxing, which would be classified as **Other Illicit Activity**).

Findings (total):

Total by URL: 1.25%
Total Percent of Domains: 3.1%

Findings (Legal excluded):

Total by URL: 4.58%
Total Percent of Domains: 8.47%



What compromises that %?

Dark web vendors have
thoroughly systematized the fraud trade, offering everything from bank drops (bank accounts) to fullz (full identities) to credit cards or music-streaming credentials.

On certain dark web carding sites vendors sell batches of cards with few additional details; users simply purchase a certain number of cards from the vendor's stockpile and go on their way.

On other more sophisticated carding markets, users can sort and filter by bank, card type, city, state, issuer - a host of different options to narrow down a specific set of cards for purchase. Both vendors and buyers are facing falling validity rates on batches of cards as banks increasingly identify common points of purchase and shut down any active cards that may have been involved in a breach.

Scenes From The Dark Web

Many of the most prolific fraud sites technically exist on the clear web.

These clear web sites often operate on top level domains based in countries less likely to shut down sites hosting illegal activity: Western Samoa (.ws), Cameroon (.cm), Cocos Islands (.cc), and Oman (.om) are all favorite TLDs for fraud marketplaces.

Multiple Categories (Illicit)

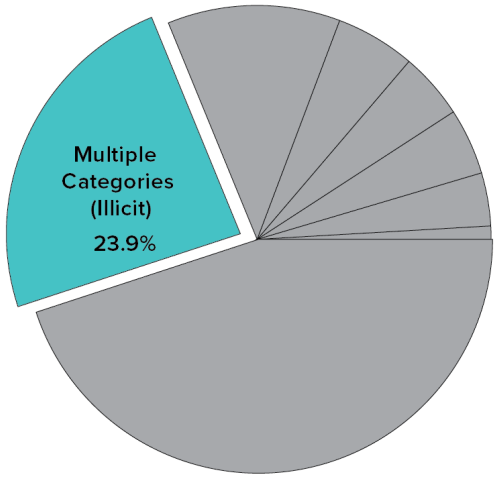
We included **Multiple Categories (Illicit)** as a way to measure instances where a series of illicit activities overlap rather than capturing the different categories discretely by allowing a single page to carry multiple classifications.

Findings (total):

Total by URL: 6.5%
Total Percent of Domains: 10.55%

Findings (Legal excluded):

Total by URL: 23.8%
Total Percent of Domains: 28.81%



What compromises that %?

A prime example of **Multiple Categories (Illicit)** is the home page of any major dark web market, where vendors are selling or advertising a wide variety of illicit products. Alphabay is perhaps the largest dark web market still running, thanks to a series of law enforcement shutdowns and exit scams on other major markets. This "Amazon of the dark web" contains multiple categories and subcategories detailing the goods and services available for sale: drugs, fraud, exploits, physical goods (watches, jewelry, electronics, even counterfeit currency), pornography, e-books. These listings all appear alongside each other, creating a single hub for all buyers.

Falsified Documents & Counterfeits

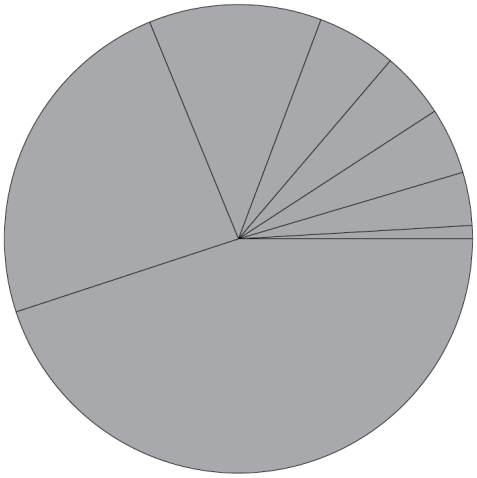
We defined **Falsified Documents & Counterfeits** as any materials related to creating or falsifying an identity. We distinguished **Falsified Documents & Counterfeits** from **Fraud** and **Other Illicit Activity** as a way to capture the proportion of the dark web specifically related to creating or stealing an identity for reasons other than marketed monetary gain. We also used **Falsified Documents & Counterfeits** as a means to account for resources that would be useful in human trafficking.

Findings (total):

Total by URL: 0%
Total Percent of Domains: 0%

Findings (Legal excluded):

Total by URL: 0%
Total Percent of Domains: 0%



What compromises that %?

For the purposes of this research, we originally broke Falsified Documents & Counterfeits out from Fraud as a means to differentiate between standard fraud and materials that could facilitate human trafficking. We wanted a means to capture any market listings, discussions, or information that would allow a buyer to obtain a new identity for themselves or someone else. Falsified Documents & Counterfeits on the dark web, where they exist, include fake passports or utility documents, social security numbers, birth certificates, drivers' licenses, or other national identification materials.

Falsified Documents & Counterfeits do exist on the dark web, but appear infrequently enough that we did not observe a single instance in our sample. For more about unobserved categories, see our Absence of Evidence section on page 31. For details about our consideration of error (skip to Confidence Intervals on page 31)

Scenes From The Dark Web

Looking for an under-21 fake ID? You can find it on the dark web.

A vendor on Alphabay started selling these under-21 IDs as a joke, and it turned into a lucrative side product. Never pay full price for a gym membership or hockey tickets again.

Exploitation

Terbium Labs reports all exploitation material to the proper authorities, including the National Center for Missing and Exploited Children (ncmec.org).

We defined **Exploitation** as any kind of content containing pornographic, violent, or otherwise abusive or illegal content involving children. **Exploitation** includes the discussion of or distribution of child pornography, along with links to files, chat rooms, or websites, promoting the rights of children and/or abusers for this purpose, hypothetical or fictional content/ fantasies, and drawings or depictions of any individual believed to be or suggested to be underage.

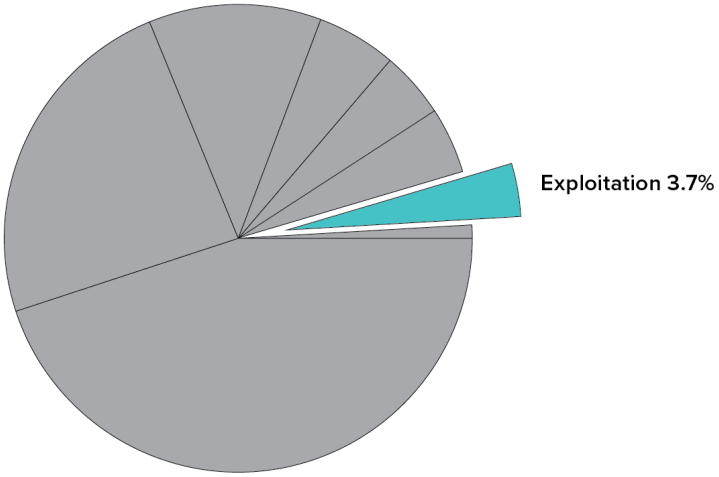
For the purposes of our research, all sites were classified with browser images turned off. Any URL or domain containing exploitation material was added to a blacklist. For full details on our blacklisting procedures, please see the Methodology section on page 28.

Findings (total):

Total by URL: 1%
Total Percent of Domains: 1.24%

Findings (Legal excluded):

Total by URL: 3.66%
Total Percent of Domains: 3.89%



What compromises that %?

Unfortunately, exploitation exists in measurable quantities; it is present more than extremism, more than weapons, even almost equal to fraud. This is a legitimate and real concern on the dark web, and is not as infrequent as you might hope it to be.

Hacking & Exploits

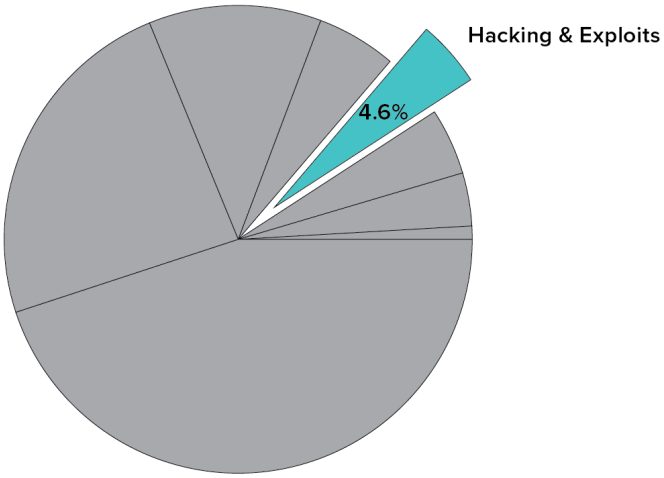
We designated **Hacking & Exploits** to capture any content that would typically be considered "cybercrime", including malicious software, exploit kits, and hacker-for-hire services.

Findings (total):

Total by URL: 1.25%
Total Percent of Domains: 1.86%

Findings (Legal excluded):

Total by URL: 4.58%
Total Percent of Domains: 5.08%



What compromises that %?

Hacking & Exploits on the dark web take on two primary forms: information and tools for sale and tradecraft building. Listings on major markets and dark web forums include a range of offers for technological crime, ranging from Distributed Denial of Service (DDoS) attacks for sale to malware or ransomware kits, even offers to expose or attack individuals. For a price, vendors on the dark web will arrange to automatically call any phone number every hour with pre-recorded threatening messages.

In addition to these materials for sale, the dark web is also home to a plethora of forums designed to build tradecraft and institutional knowledge regarding exploits and developing new and creative ways to infiltrate secure systems.

Weapons

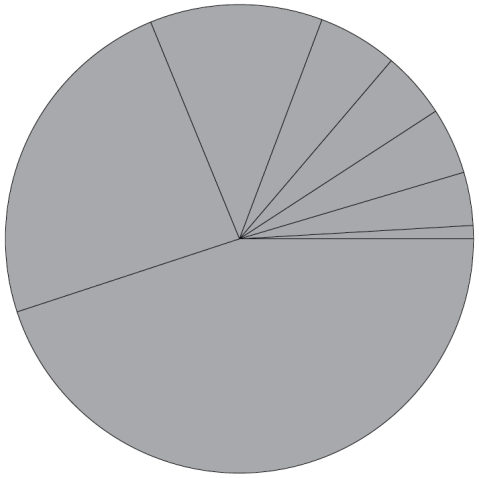
We defined **Weapons** as any kind of personal weapon that can be used to harm an individual or otherwise small number of people. **This category** includes items like guns, bullets, knives, and other individualized weapons. We separated **Weapons** from **Weapons of Mass Destruction** to capture separate concerns over handheld weapons (primarily firearms) from larger scale, traditional WMDs.

Findings (total):

Total by URL: 0%
Total Percent of Domains: 0%

Findings (Legal excluded):

Total by URL: 0%
Total Percent of Domains: 0%



What compromises that %?

Can I get a gun on here? The image of major arms dealers selling crates of Kalashnikovs is one of the most persistent dark web myths. Unlike ISIS planning forums, weapons aren't quite a myth: they do exist in isolated pockets. So isolated, in fact, that our sample of dark web sites did not include a single weapons site. For more information, see our discussion on absence of evidence and estimates of error.

At any given moment there are a handful of independent sellers, outside of the major markets, offering handguns, rifles, ammunition, even the occasional grenade launcher. While these items do appear for sale, the dark web community largely regards the gun sellers as scammers.

Scenes From The Dark Web

In the wake of the terrorist attacks in late 2015 and early 2016, many of the major markets removed firearm listings from their markets.

You can still buy a flashlight Taser, though.

Extremism:

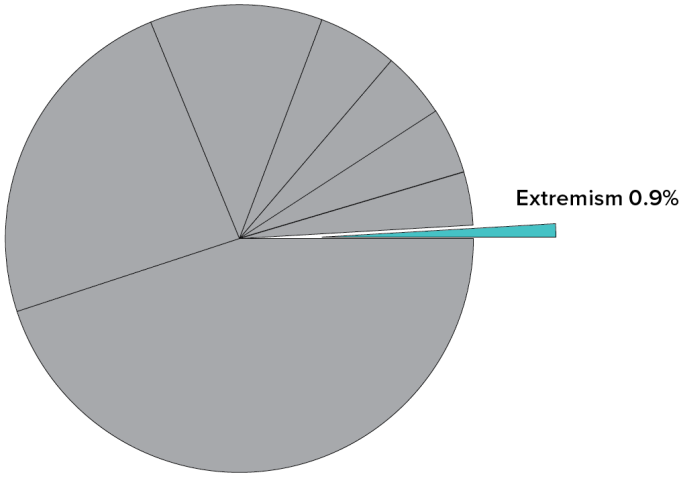
We defined **Extremism** as any kind of radical rhetoric, particularly as relates to political or religious calls to action. **Extremism** covers any generally accepted extreme group rhetoric, including Islamic extremism, neo-Nazism, white supremacy, militant groups, etc. **Extremism** differs from **Other Illicit Activity** in that extremism is driven by an ideology rather than a personal, specific desire. **Extremism** is used to classify extreme rhetoric or calls to action that originate from an adherence to a cause larger than self.

Findings (total):

Total by URL: 0.25%
Total Percent of Domains: 0.6%

Findings (Legal excluded):

Total by URL: 0.91%
Total Percent of Domains: 1.69%



What compromises that %?

We observed one instance of extremism in the 400 URLs we reviewed for this study. Extremism is a difficult category to capture, because (anecdotally) we don't observe a large amount of extremist activity on the dark web. Extremist content does appear from time to time, typically in the form of religious or political rhetoric (as seen in our sample). A few outliers of note: In 2015, an official ISIS .onion site appeared and was quickly taken down by Anonymous. In early 2016, a highly-stylized ISIS-sponsored guide to bomb-making also began to make the rounds.

These are notable because they are rare and because they are at odds with the kind of content typically found on the dark web.

Weapons of Mass Destruction

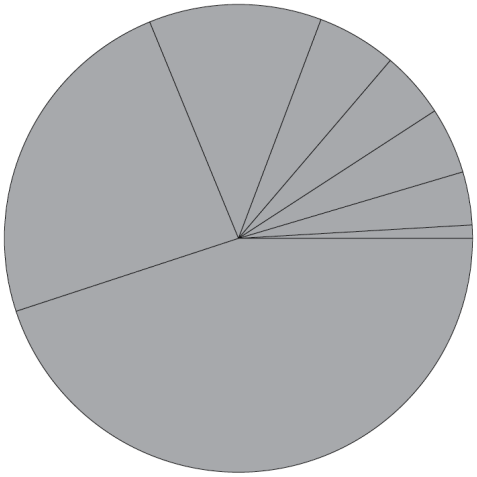
We defined **Weapons of Mass Destruction (WMD)** as any kind of biological agent, chemical, weapon, conventional weapon, or any other kind of technology designed to target, injure, or kill a large number of people. A vendor selling Anthrax, or a guide providing details on how to create Chlorine Gas would fall under **WMD**.

Findings (total):

Total by URL: 0%
Total Percent of Domains: 0%

Findings (Legal excluded):

Total by URL: 0%
Total Percent of Domains: 0%



What compromises that %?

Individuals outside of the dark web community often ask about the appearance of Weapons of Mass Destruction (along with Extremism and Human Trafficking) in relation to the dark web. The expectation is that the dark web, as a place of illegal commerce, would facilitate trade in dangerous materials.

While the dark web does facilitate trade in illicit materials, the anonymity of Tor Hidden Services does not make it easier to ship ostentatiously large items (e.g., nuclear warheads), and does not circumvent the issues in shipping dangerous chemicals or biological weapons (e.g., smallpox). These limitations might explain why the dark web is not a main thoroughfare for trade in Weapons of Mass Destruction.

Other Illicit Activity

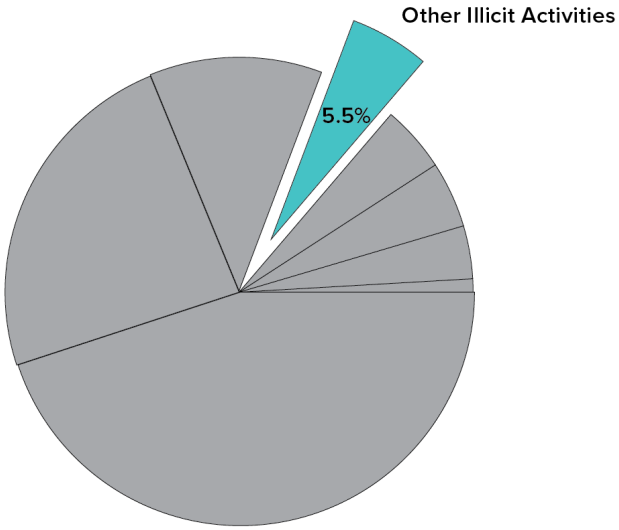
We defined **Other Illicit Activity** as any kind of *single* illicit activity that did not fit into one of our other categories.

Findings (total):

Total by URL: 1.5%
Total Percent of Domains: 3.7%

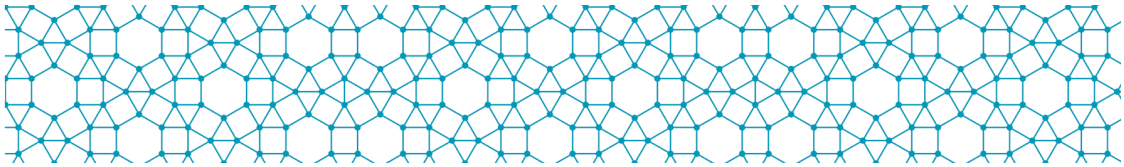
Findings (Legal excluded):

Total by URL: 5.5%
Total Percent of Domains: 10.16%



What compromises that %?

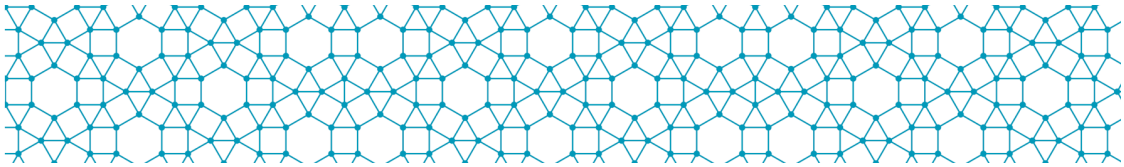
The content that our analysts ultimately classified as Other Illicit Activity included what some might think of as more traditional crime, typically in the form of anarchy guides or recommendations on how to otherwise wreak havoc. These guides are relatively rare, and include instructions on the best way to harass your neighbor (perhaps some small-scale arson).



Methodology & Limitations

Data is not necessarily objective or accurate.
Methodology is everything.

The details matter.



Methodology

Sampling Strategy and Classification

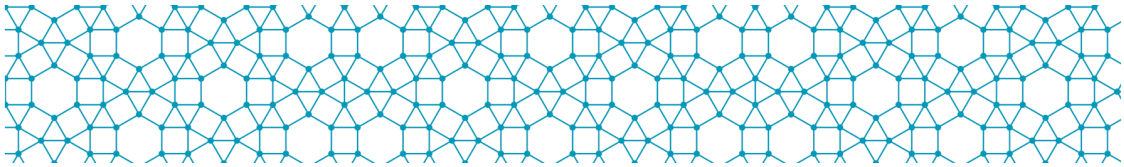
A sample of 400 unique URLs were selected from the population of Tor Hidden Service URLs that our crawler successfully accessed in one day (August 5, 2016). The population was restricted to a single day as the dark web is highly transient, with sites appearing/disappearing regularly. There is no *a priori* reason to think that all types of content are equally transient. By restricting our analysis to a single day, we can represent the relative proportions of different content at a representative moment in time. A URL is considered a dark web URL when its top level domain is '*.onion*'. URLs that were not successfully accessed on August 5, 2016 or did not have the ".onion" top level domain were excluded from the study population. The random sample was drawn from the set of unique URLs that met our exclusion criterion, allowing multiple URLs from the same domain to be independently selected. During the following week, a team of human analysts visited each of the 400 unique URLs and classified the contents of the site as belonging to one of 15 categories. The category definitions are included in the Appendix (page 35) . URLs are only classified into one category. Websites with content from multiple illicit categories are classified as "Multiple Categories (Illicit)".

Handling Illegal Content and Exploitation

In accordance with Terbium Labs' policy, plain-text content from the dark web was not stored in order to complete this study. Extreme care was taken in handling material that may contain evidence of exploitation content. During the study, we maintained a blacklist of URLs and domains that are known to contain Exploitation content. While these URLs are included as data points for statistical analysis, they are never shown to an analyst. To further protect our human analysts from unsavory material, images and movies are disabled in the Tor browser during the classification process. If a human analyst encounters a previously unknown site with evidence of exploitation, the analyst marks the URL or domain as "blacklisted". The URL and domain were immediately blocked. The appropriate authorities, including the National Center for Missing and Exploited Children (ncmec.org), were immediately notified.

Statistical Analysis

Category proportions were computed with respect to URL and domain. 95% confidence intervals were computed using the methods proposed by Sison and Glaz in their 1995 paper for simultaneous calculation of confidence intervals on multinomial distributions. Custom code was developed and analysis was done in python. To do this calculation, we also used 'MultinomialCI' package in R.

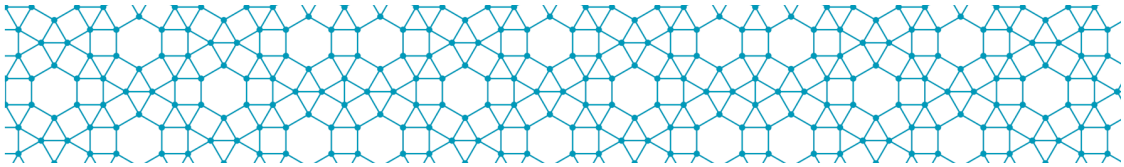


Sison C, Glaz J. Simultaneous confidence intervals and sample size determination for multinomial proportions. Journal of the American Statistical Association 1995 90:366-369.

Why We Used a Random Sample

In the age of big data, it is tempting to assume more data and more automation results in increased in confidence and accuracy. In reality, an abnormally large sample size or a poor machine learning model can amplify sources of selection bias. If a study considers samples or training data that is chosen by (or curated by) a human, the data will reflect implicit biases. A fraud analyst may include more fraud sites. Analysts searching for personal information are more likely to include sites that are associated with data dumps. This will result in an over-estimation of the relative amount of fraud and leaked data while underestimating generally uninteresting content such as legal porn, news sites, and other legal uses of Tor. Most machine-learning classification algorithms also consider the *prior*, the relative frequencies of different categories in the training data. That knowledge is used implicitly when automatically classifying new data. In this way, human biases are programmed directly into the automated algorithms.

Selection bias also results from technical limitations. Much of the dark web is hidden behind captchas and logins. These hurdles are designed to keep web crawlers and humans from accessing a site. Some types of content are more likely to be hidden or restricted by a login or captcha than others. Market content is meant to be publically advertised, while other content is meant to be kept secret. When defining a representative sample, it is important not to exclude sites simply because they are hard to find or access automatically. Most web crawlers do not interact with a website before extracting new URLs. Given the prevalence of these privacy features on the dark web, this type of availability bias is particularly important. Terbium's propriety crawler gets around captchas and behind some logins, ensuring a more representative sample population.



Quantifying the Dark Web is Hard and Researchers Can Disagree

We are always skeptical of data (including ours).

Terbium Labs has an academic mindset with strong roots in experimental science. We know that a single study is not definitive and that the scientific method is best applied with a healthy amount of skepticism. Data never speaks for itself; data is not objective. The following is a discussion of the limitations of our study, and others like it. We outline the assumptions we made in order to carry out this research. Taken together with the data itself, this allows for a more nuanced interpretation of our results.

Definitions Matter

The dark web is not well-defined. Human beings rarely create content that fits into simple categories. This makes the job of a quantitative researcher more difficult, creating a layer of subjectivity.

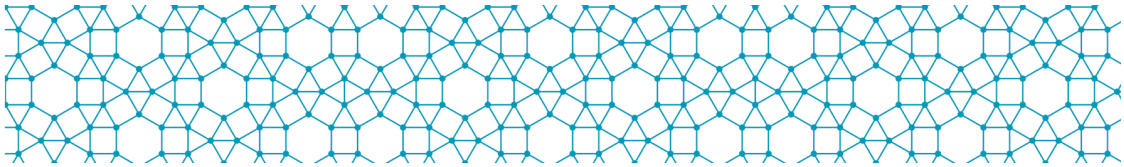
In the presence of gray areas, definitions matter. A metric cannot exist without a definition of what is being measured. Boundary conditions can arise from many sources: the requirements of a specific use-case, a shared definition among experts, or an arbitrary distinction of convenience. At multiple points in the design of this study, we made decisions about definitions and boundaries that influence the interpretation of our results.

Boundaries of the Dark Web

It is unclear where the dark web starts and ends. Here at Terbium, we often think of the dark web as "anywhere you would not want your sensitive data to appear on the internet." But that definition is not practical as it is impossible to measure. Another definition we considered was "the part of the internet that is not indexed by Google". That definition would also include many websites (such as those behind clear-web logins) that are definitively not part of the dark web.

For this study, we define the dark web specifically as sites hosted as a Tor Hidden Service, indicated by the '.onion' top-level domain in the URL. There is certainly room for criticism in this definition, but its practicality (easy-to-measure) made it preferable to many other possible boundary definitions.

If we used a more inclusive definition of the dark web, would our numbers be different? Absolutely. For instance, carding markets are often considered part of the dark web. Many major carding markets are not hosted on as a Tor Hidden Service and consequently were not represented in our population of study.



Transience and the Unit of the Dark Web

Quantitative metrics have units. At first glance, this sounds like a simple statement. In practice, it is not. Determining which unit is relevant for a given study or use case is sometimes trivial, but often extremely difficult. In this study, we fall squarely into the second category. When you measure distance, you can choose miles or meters without much thought because it is easy to convert between the two. What is the unit of the dark web? Do we care about individual URLs, or only the number of domains? Do we care only about sites that are visited, or is the fact that content exists sufficient? There is no deterministic way to convert between these units.

We measured the dark web by the unit of URLs. This means that a single domain that has many URLs will be given more weight in the final numbers than a single page domain.

This definition has real consequences. For example, when we attempt to quantify transience of the dark web. Consider the "Site Down" category. When computed with respect to URLs the relative number of transient sites is much higher than when considered by domain. This is evidence that there was at least one domain with multiple URLs in our sample population. When the domain went down, all the URLs went with it. Were each of these URLs important? Did they ever really exist or was it a crawler trap? We cannot know. The most we can do is reiterate, definitions matter. Data is not objective.

Absence of Evidence

A key conclusion of this study is that certain types of illicit activity, specifically weapons of mass destruction and extremism largely do not exist on the dark web. No amount of data can ever support a claim of "absence". This is not a limitation unique to our study, but a truth about the limits of inference. It is always possible that these sites are isolated from the rest of the dark web, or that our sample was not large enough to observe a specific example. Even if we increased our sample size, the same logic still applies. This a problem that, unfortunately, plagues all data-driven endeavors. Our confidence intervals reflect this reality, providing acceptable ranges that should still be considered consistent with our results.

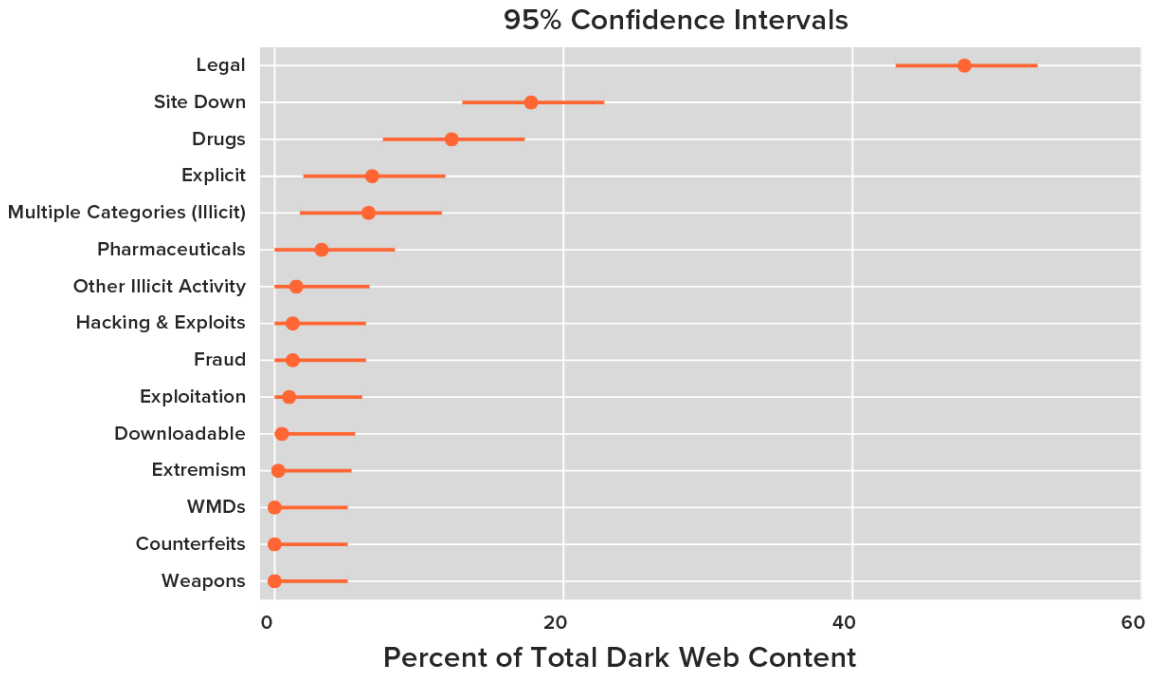
The absence of the evidence problem extends to other categories within this study. We also did not observe categories in our sample that our analysts know with certainty do exist on the dark web, such as counterfeits and weapons. When you take a random sample, there is error. We make an estimates of this error in the next section.

Error and Confidence Intervals

Measurements have error. Statisticians spend considerable effort debating how to accurately or efficiently estimate sampling error. This often results in a confidence

interval, or as it is more commonly known, a margin of error. If we were to repeatedly perform our study in exactly the same way (with the same selection criterion and population) we would expect our observed metrics to fall within the confidence interval 95% of the time.

Our study has multiple categories (a multinomial distribution) and a finite sample (N is not large). This requires simultaneous estimation of confidence intervals across all the categories. Multiple methods are available for this calculation; we chose to use the method outlined by Sison and Glaz in their 1995 paper. Below is a graphical representative our 95% confidence intervals.

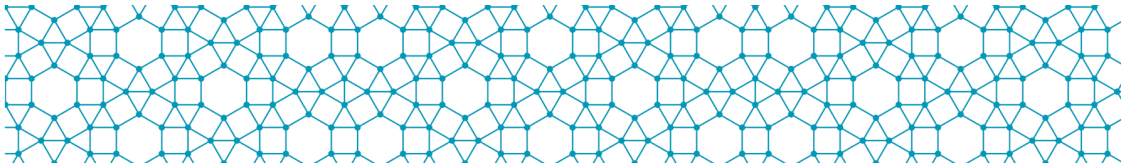


There are a couple notable inferences to draw from this data.

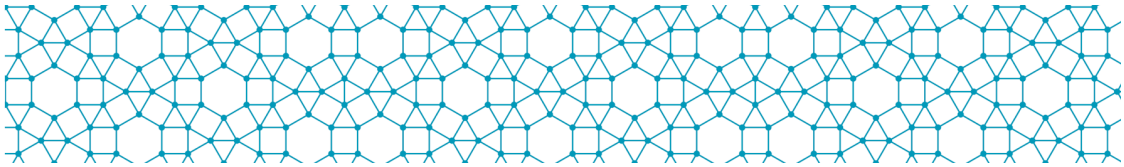
First, categories which we did not observe in our sample still have a confidence interval. This relates the absence of evidence issue but also relates to the fact that our sample size was finite. Even though we did not observe any weapons of mass destruction URLs, they might still exist. Based on this study, we estimate that the percentage of the dark web URLs that might be contain content related to weapons of mass destruction is up to 5%.

With the exception of legal, explicit, and drug categories, most categories' confidence intervals include zero. This means that if we were to redraw a new random sample, we could fail to observe any of these categories some of the time.

Despite the fact that weapons and weapons of mass destruction are both unobserved categories, they are not equally unlikely. Confidence intervals reflect sampling error. No one (including the statisticians!) know what "ground truth"



really is. Our team of analysts know for certain that weapons and counterfeits do exist on the dark web. Our team is not aware of clear examples of weapons of mass destruction or violent extremism. If we were to take the Bayesian approach to this inference, we could have used this existing knowledge to update and strengthen our inferences. Data, as we mentioned before, is not inherently objective. Research is hard.



Conclusion

The dark web is host to primarily legal, even mundane content. The sections that aren't legal are dominated by drugs, fraud, and combinations of illegal activity - primarily the major dark web markets.

Just because the majority of the content is legal, though, doesn't mean it's safe. Legal content has the potential to be damaging, even dangerous. At Terbium, we refuse to make judgement calls about the kind of information that would be interesting to our customers. That's why we built out a big data infrastructure, backed by an unrivaled crawler. We want our customers to see everything that's on the dark web, not just the sites we think may be of interest.

There's always more research to be done. Within each of the categories we cover here, what types of content actually exist? What percentage of the illegal content we captured is for sale, and what percentage is shared as free information? What would these numbers look like if we expanded our sample beyond Tor Hidden Services?

Research is hard, and data is not magic. At Terbium, we like hard questions.

Appendix

Legal: We defined Legal as any activity or discussion that was not explicitly illegal. We also classified a page as Legal if the content on the page was legal, even if the broader site may have contained illegal information, goods, or services.

Explicit: Explicit includes any non-exploitation explicit content - effectively, pornography. We included explicit as a category to account for the percentage of the dark web that contains perfectly legal pornography, available for free or for purchase on Tor Hidden Services.

Multiple Categories (Illicit): We included Multiple Categories (Illicit) as a way to measure instances where a series of illicit activities overlap rather than capturing the different categories discretely by allowing a single page to carry multiple classifications.

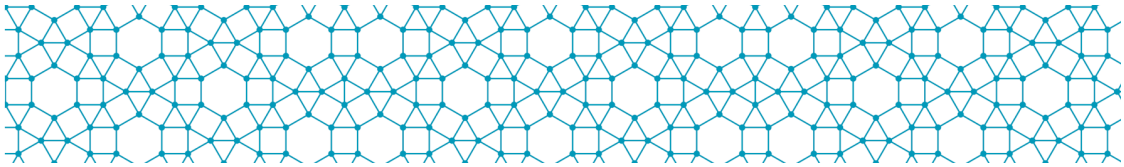
Drugs: We defined Drugs as any non-pharmaceutical drug or substance bought or sold for recreational purposes. To provide a more detailed breakdown of the kinds of drugs available on the dark web, we separately classified any Pharmaceuticals available for sale as well. We include marijuana as a drug and not a pharmaceutical for the purposes of this study.

Pharmaceuticals: Pharmaceuticals include any kind of drug that a doctor might prescribe, excluding painkillers and their derivatives. For our classification, Pharmaceuticals include ADD/ADHD and anti-anxiety medications, even though these medications are often used recreationally.

Fraud: We defined Fraud as any kind of content based on set credentials or accounts. We distinguish Fraud from Falsified Documents & Counterfeits and Other Illicit Activity in order to capture instances where an existing Falsified Documents & Counterfeits was compromised, leaked, or sold for the purposes of committing fraudulent acts. We defined Fraud this way to account for instances where identities are being sold or marketed for fraudulent purposes separate from either personally identifiable information made available for vandalism (e.g., doxxing, classified as Other Illicit Activity), or instances where the fraudulent act involved creating a new Falsified Documents & Counterfeits (e.g., a fake passport, classified as Falsified Documents & Counterfeits).

Falsified Documents & Counterfeits: We defined Fraud as any kind of content based on personal information, credentials, or accounts. We distinguish Fraud from Falsified Documents & Counterfeits and Other Illicit Activity in order to capture instances where an existing Falsified Documents & Counterfeits was compromised, leaked, or sold for the purposes of committing fraudulent acts (this is separated from, for example, doxxing, which would be classified as Other Illicit Activity).

Exploitation: We defined Exploitation as any kind of content containing pornographic, violent, or otherwise abusive or illegal content involving children.



Exploitation includes the discussion of or distribution of child pornography, along with links to files, chat rooms, or websites, promoting the rights of children and/or abusers for this purpose, hypothetical or fictional content/ fantasies, and drawings or depictions of any individual believed to be or suggested to be underage.

Hacking & Exploits: We designated Hacking & Exploits to capture any content that would typically be considered "cybercrime", including malicious software, exploit kits, and hacker-for-hire services.

Weapons: We defined Weapons as any kind of individual, personal weapon that can be used to harm an individual or otherwise small number of people. Weapons includes items like guns, bullets, knives, and other individualized weapons. We separated Weapons from Weapons of Mass Destruction to capture separate concerns over handheld weapons (primarily firearms) from larger scale, traditional WMDs.

Extremism: We defined Extremism as any kind of radical rhetoric, particularly as relates to political or religious calls to action. Extremism covers any generally accepted extreme group rhetoric, including Islamic extremism, neo-Nazism, white supremacy, militant groups, etc. Extremism differs from Other Illicit Activity in that extremism is driven by an ideology rather than a personal, specific desire. Extremism is used to classify extreme rhetoric or calls to action that originate from or adherence to a cause larger than self.

Weapons of Mass Destruction: We defined Weapons of Mass Destruction (WMD) as any kind of biological agent, chemical, weapon, conventional weapon, or any other kind of technology designed to target, injure, or kill a large number of people. A vendor selling Anthrax, or a guide providing details on how to create Chlorine Gas would fall under WMD.

Other Illicit Activity: We defined Other Illicit Activity as any kind of single illicit activity that did not fit into one of our other categories.