

Who's Afraid of the Dark? Hype Versus Reality on the Dark Web

By Juan Sanchez and Garth Griffin




Key Findings

- The collection of onion sites that is sometimes called the dark web is often portrayed as a vast and mysterious part of the internet. In reality, the number of onion sites is tiny compared to the size of the surface web. Our count of live reachable onion site domains comes to less than 0.005% of the number of surface-web site domains. Out of about 55,000 onion domains that we found, only around 8,400 onion domains had a live site (15%). The popular iceberg metaphor that describes the relationship of the surface web and dark web is upside down.
- These onion sites are disorganized and unreliable. Scams are prevalent, such as a typosquatting scam that claims to have successfully defrauded users of over 400 popular onion sites, netting thousands of dollars in Bitcoin from victims. Uptime even on popular dark web sites is well below the 99.999% “five nines” availability that is expected for reputable companies on the surface web, and onion sites regularly disappear permanently with or without explanation.
- From a language standpoint, onion sites are more homogeneous than the surface web. We observed that 86% of onion sites have English as their primary language, with the next two most common being Russian with 2.8% and German with 1.6%. On the surface web, researchers report English is at the top with only 54%.
- The idea of a dark web that is hidden and mysterious is more likely an extrapolation of a tiny portion of these onion sites — a set of invitation-only and unpublicized communities buried in the most shadowy corners of this part of the internet. On the surface web, popular websites will attract inbound link counts in the millions or more. In our onion site crawl, the site with the highest inbound link count was a popular market with 3,585 inbound links. An onion site offering help setting up onion servers had 279 inbound links. In contrast, we looked at what we view as the top eight onion sites most respected in the criminal community and found that the most visible had a maximum of 15 inbound links with an average of only 8.7 inbound links per site. It is this tiny slice of the dark web that is truly dark.

The dark web has long been shrouded in mystique and is a frequent topic of interest to our users. But what's the real story?

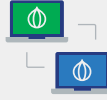
THE DARK WEB

- Any World Wide Web content that requires specific software, configurations, or authorization to access
- Includes the Tor network



TOR

- Free, open-source software initially developed by the U.S. military and designed for anonymous communication
- Network consists of onion domains and connections between them in the form of direct links
- Websites that are able to be reached without these kinds of specific software or network configurations are known as the clear web

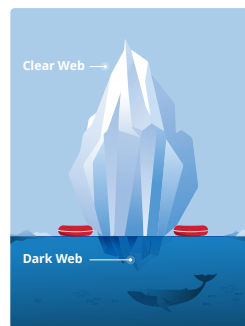


Recorded Future spidered about 260,000 onion pages to approximate the full reachable Tor network from a starting set of onion sites that we pulled from public lists and our own content.

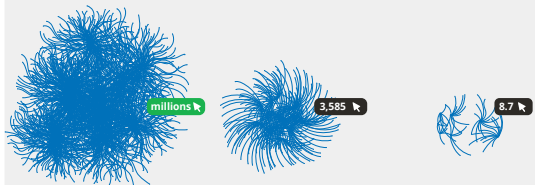
Though the dark web is often portrayed as a vast and mysterious part of the internet, it's actually tiny compared to the "clear web."

The popular iceberg metaphor that describes the relationship of the clear and dark web is upside down.

- Our count of live reachable onion domains comes to less than 0.005% of the number of clear web site domains.
- Only around 8,400 onion domains had a live site out of about 55,000 onion domains that we found (15%).
- Uptime even on popular dark web sites is well below the 99.999%+ (or "five nines") availability that is expected for reputable companies on the clear web, and sites regularly disappear permanently with or without explanation.



The idea of a dark web that is hidden and mysterious is more likely an extrapolation of a tiny portion of the dark web — a set of invitation-only and unpublicized communities buried in the most shadowy corners of the internet.



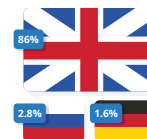
CLEAR WEB	DARK WEB	DARK WEB (TOP TIER)
On the clear web, popular websites will attract inbound link counts in the millions or more.	In our dark web crawl, the site with the highest inbound link count was a popular market with 3,585 inbound links.	In contrast, we looked at what we view as the top eight onion sites most respected in the criminal community.
		The most visible sites had a maximum of 15 inbound links with an average of only 8.7 inbound links per site.

OUR DEEP DIVE OF THE DARK WEB FOUND ...



Scams are prevalent.

A dark web typosquatting scam claims to have successfully defrauded users of over 400 popular onion sites, netting thousands of dollars in Bitcoin from victims.



From a language standpoint, the dark web is more homogeneous than the clear web.

- 86% of sites in the dark web have English as their primary language, followed by Russian (2.8%) and German (1.6%).
- On the clear web, researchers report English is at the top with only 54%.

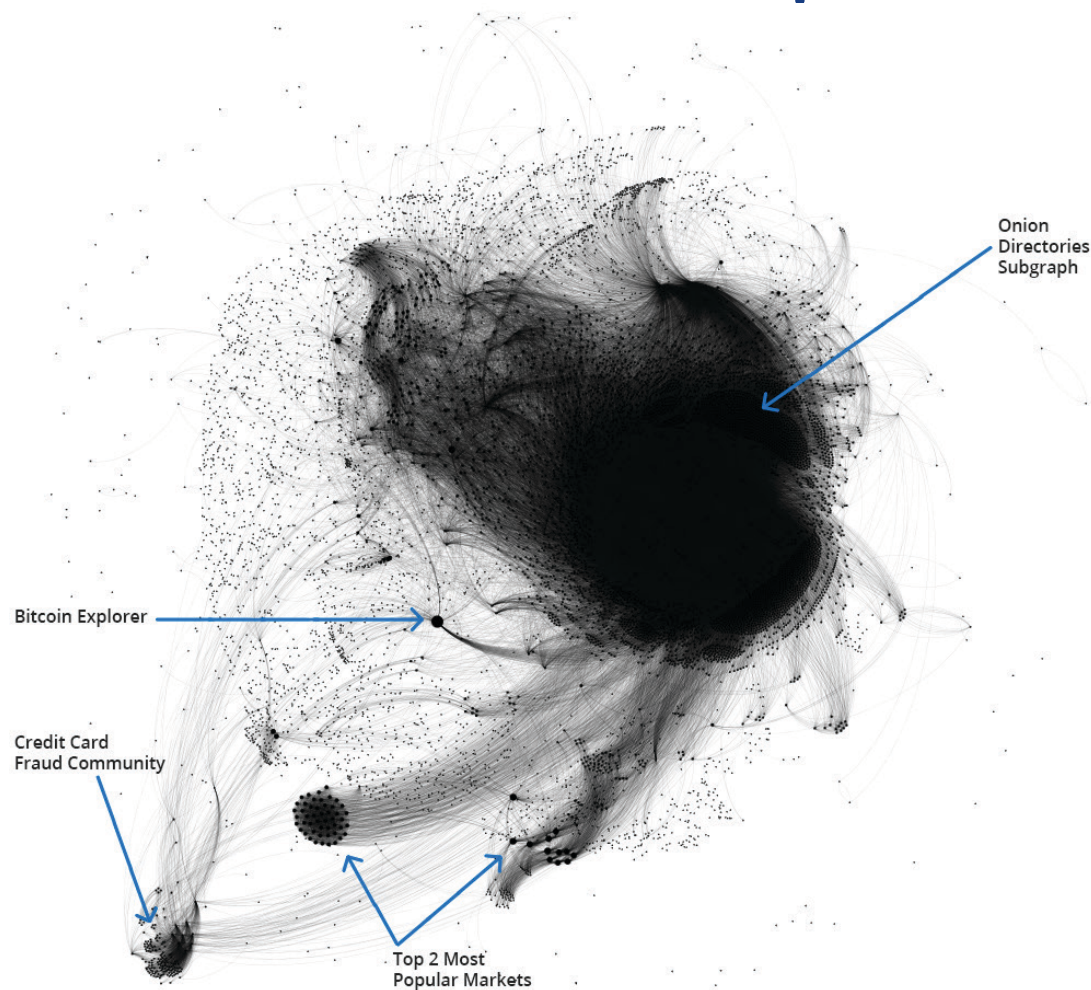
What Is the Dark Web?

The dark web is a frequent topic of interest for anyone who cares about cybersecurity, but its mystique has given rise to a number of popular misconceptions, and “[dark web](#)” can be a muddled term. To make a more concrete assessment of one precise definition of the term “dark web,” this blog presents our findings of a spider specifically for those sites that are accessible within the Tor network of onion domains. There are plenty of varied definitions for the dark web, the deep web, the criminal underground, and other related concepts, but for this investigation, our exclusive focus is on onion sites.

According to [Wikipedia](#), the dark web can be described as any web content that requires specific software, configurations, or authorization to access. This definition overlaps with another common term, the “deep web,” which is commonly used to refer to all the parts of the internet not indexed by search engines.

The dark web is also often conflated with the cybercriminal underground, implying that it is solely a place where people traffic illicit and sordid goods and services. While that kind of activity makes up a significant proportion of content on the dark web, the fact that the Tor browser can circumvent surveillance measures also makes it useful for legitimate activities in certain circumstances, like free expression from political dissidents in authoritarian countries. Some prominent surface websites host mirrors of their content on Tor sites for exactly this reason, including The New York Times and Facebook. On the other side of the coin, [Insikt Group's research](#) has shown that much criminal activity happens on sites not requiring any special protocols to access, such as public social media sites like Twitter or messaging services like WhatsApp and Telegram.

In this research, we investigated a few things about this network of onion sites: how big it really is, the languages in which it's written, and how reliable it is to use in terms of uptime and trustworthiness. We spidered about 260,000 onion pages to approximate the full reachable Tor network from a starting set of onion sites that we pulled from public lists and our own content.

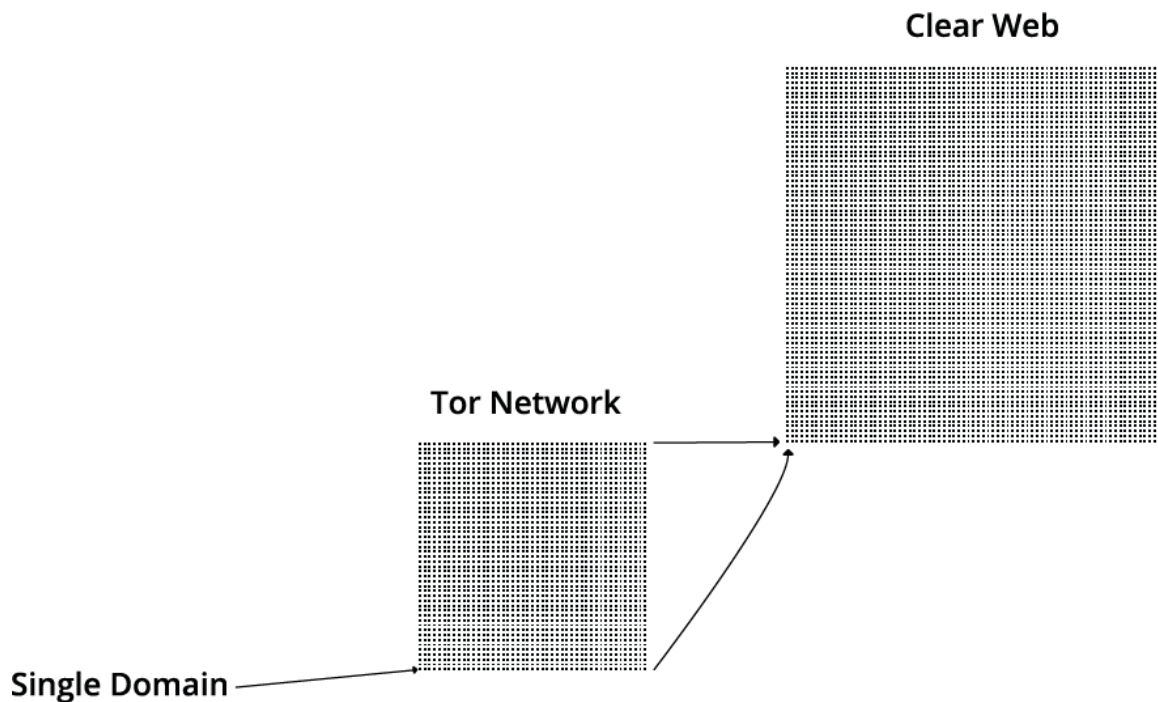


Graphical representation of 8,416 onion domains.

The Dark Web Is Tiny

The dark web is often portrayed as vast and mysterious, implying that there is a large number of onion sites on Tor, but this is not what we find. This misperception may be in part due to the fact that there are many tragic and horrible things that take place under the anonymity Tor provides. While we cannot contradict the sad reality that those things do happen, we find that in terms of size, the network of onion sites is tiny compared to the surface web, and the part with real threat intelligence value is smaller still. Our crawling found 55,828 different onion domains, but only 8,416 were observed to be live on the Tor network during our crawl.

Our [findings disprove](#) the misconception that the relationship between the surface web and dark web has an iceberg shape, with the surface web being a small portion of the World Wide Web above the water and the dark web below the visible surface accounting for the majority. The truth is that this iceberg shape is upside down.

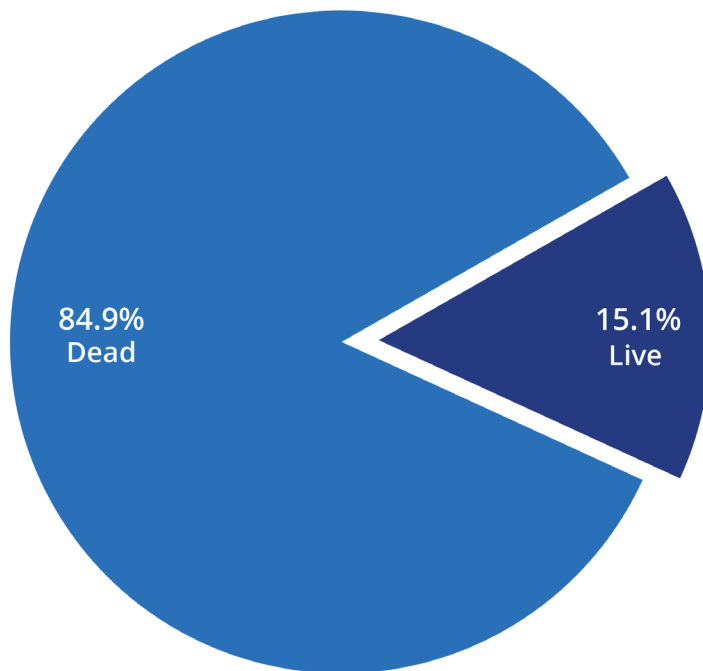


Pixel representation of the dark web versus the surface web.

There are an estimated 200 million unique [surface web domains](#) that are active, which positions the current live onion site network at less than 0.005% of the size of the World Wide Web.

Onion sites are prone to disappearing from the network, which will cause any attempt to reach the page to fail. The ratio of live to total onion domains was about 15% in our results. A similar ratio (about 15%) holds for the [surface web](#). This number, which provides an estimate for the size of the Tor network, complements the findings of an Onionscan report from 2017, which reported a live rate of [4,400 live sites out of 30,000](#). Others also claim that [the network is shrinking](#). While we cannot directly compare against their numbers because their approach was not as broad as our spider, we do find that the ratio of live to dead continues to be similar to these previous findings, with about 15% of the sites being live.

State of Onion Domains Found With Spider



Percentage of onion domains that will succeed in loading a page.

We also found that this tiny network of onion sites is tightly connected. For 82% of the live domains in the network that we've crawled, the average degrees of separation from a popular link hub like the Hidden Wiki is 2.47. The data suggests that if you visit the Hidden Wiki onion page, you'd be about three clicks away from 82% of live onion sites. This measure is tighter than might be expected in the surface web. For example, the Facebook social graph has been reported to have an [average degree of separation](#) of 3.57 between pairs of users.

It's also notable that the other 18% of crawled domains were completely disconnected from the Hidden Wiki, which might indicate the presence of isolated communities separate from the rest of the network. While this opens the possibility of there being swaths of sites that our approach could not discover, we believe this is unlikely due to our broad starting base, which included all onion domains seen anywhere in our vast open source data as well as our extensive collection focused on the criminal underground.

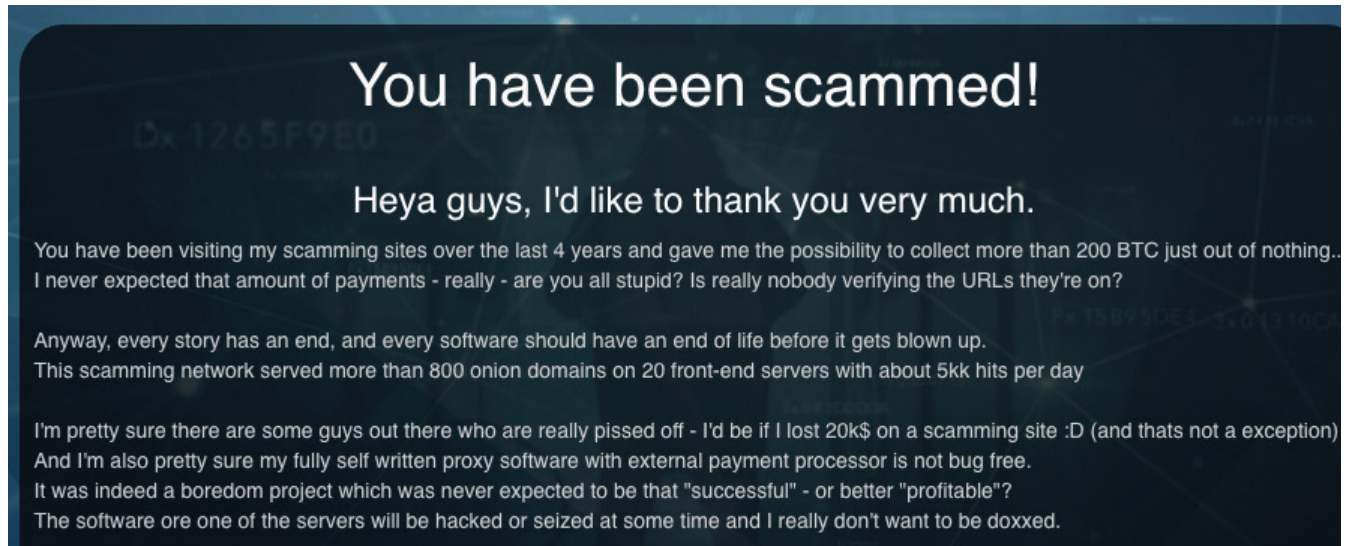
The Dark Web Is Disorganized and Unreliable

The dark web is plagued by flakiness. As criminal activity has proliferated across onion sites, so have scams and attacks. The servers of onion websites are taken down when they fall victim to attacks. A prominent example is a site called Daniel's Hosting, which used to provide Tor hosting services to about 6,500 onion websites. This site was hacked in 2018, causing a [massive outage](#) of onion sites. The infrastructure was compromised using a PHP zero-day vulnerability that allowed the hacker to gain access to the full database of sites and delete all the accounts inside.

While it was eventually recovered, the victimization and prolonged downtime is a typical example of the level of service found on onion sites. Even popular dark web markets can have uptime well below 90%, with one well-known market having about [65% uptime](#) as of this article. Sites can be down for weeks at a time, which would be unthinkable for reputable service providers on the surface web. For comparison, [Facebook's uptime](#) is measured at 99.95%, and the gold standard is 99.999% availability, known as "five nines." Onion sites are typically far below that level, and some simply disappear for days, for weeks, or for good.

[Typosquatting](#) is a tactic used by malicious actors on the surface web, and this has been taken to onion sites as well. Typosquatting is a technique where a malicious actor registers a domain that users of a legitimate website might easily mistake for the website of the service they're trying to use, which is then exploited by the actor hosting malicious content on the typosquatted domain (for example, a fake login page at "aple[.]com" or "apple[.]co").

We found a blatant example of onion site typosquatting that we're calling the "Thank You" scam. Our spider found numerous copies of onion sites hosting only a simple banner from someone that claims to have earned more than 200 BTC by hosting slightly modified domain names for over 800 popular onion sites. We speculate that the perpetrator might have asked for user credentials and [profited from stealing](#) them, but this is unclear, as the scam landing pages are no longer visible and all the sites instead show the gloating message. Well-known Bitcoin mixers and markets were included in the list of typosquat victims.



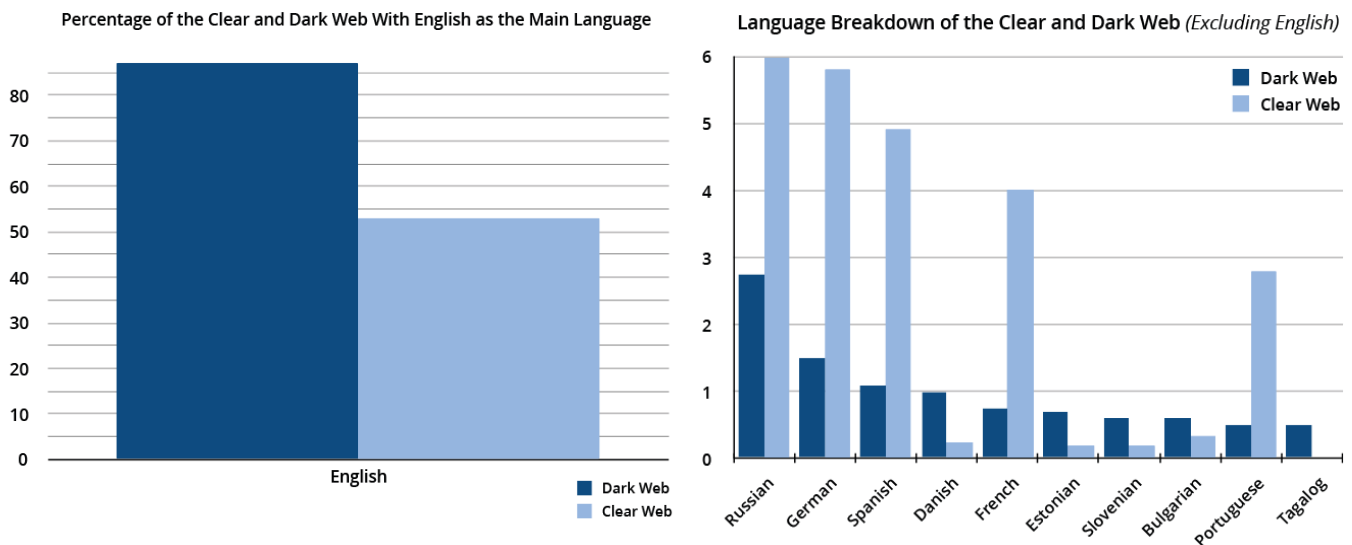
Recent screenshot for a fake site which was part of the "Thank You" scam.

From the "more than 800" fake domains referenced by the scammer, our spider found 430 live sites, all with a landing page where the perpetrator communicates his retirement and thanks the viewer for their money. If indeed there are as many as the banner claims, we believe that the remaining 370 are no longer live.

Typosquatting is even easier on onion sites than the surface web due to the way that onion domains work. Onion domains are hashes, so they typically contain many characters that appear entirely random to a human user. For example, the onion domain 7rmath4ro2of2a42.onion does not correspond in any visual sense to the site that it loads, a news site called SoylentNews. This makes it hard for a Tor user to distinguish between a real onion domain and a typosquat. Sharing written onion typosquats would be an effective way to spread them, as many Tor users will not be familiar enough with the real domain to tell the difference. In addition, many fake domains were added to [Daniel's Onion Link List](#), a popular site for hosting and listing onion domains. Finding phishing links is common enough for [Deep Dot Web](#) to make a post warning about it. Even without considering the content of the sites, these factors give the entire network of onion sites a sense of untrustworthiness.

Language Usage on the Dark Web is More Homogeneous Than on the Surface Web

Various studies have estimated the language breakdown of the surface web, but such measurements of the dark web are rare. Spidering the Tor network provides a way to measure the breakdown of written languages on onion sites. We estimate that English is the main language for 86% of onion sites, a higher proportion than the [surface web](#), in which English accounts for only 54%. Following English comes Russian at 2.8%, German at 1.6%, and Spanish at 1%. The languages below those in frequency account for less than 1% each and 8.6% as a whole. While the percentages differ, the order of the top four languages by popularity is the same as the order for the surface web. After that, the order diverges as the percentages get smaller.



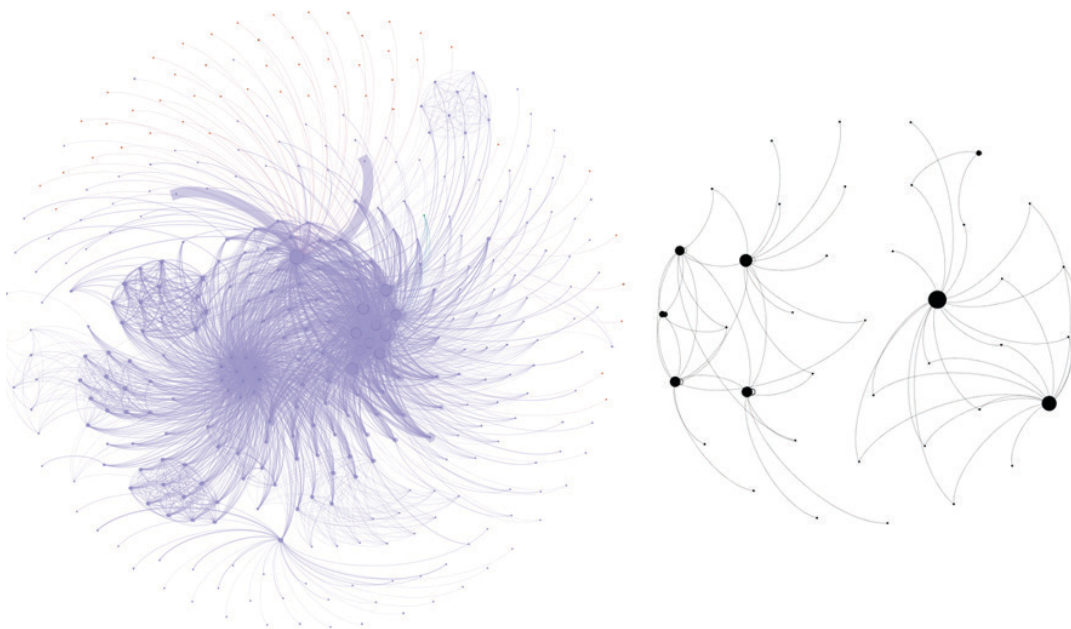
Language usage on the surface web versus onion sites.

We formed these estimates using stratified sampling for our spidered data, selecting random pages from each crawled domain and assigning a main language for the whole domain based on a majority vote of the languages detected across the pages.

The Hidden Dark Web

The idea of a dark web that is hidden and mysterious is better exemplified by a tiny portion of onion sites, a set of invitation-only and generally unpublicized communities buried in the most shadowy corners of the internet. To understand just how hidden these sites are, we measured how many unique onion domains had a link pointing to a given site. This measurement can then be compared to popular sites to evaluate their relative visibility. Popular surface-web sites have inbound link counts in the millions or more.

The site with the highest inbound link count across all our crawled onion domains was a popular market with 3,585 inbound links. An onion site providing help with hosting onion servers had 279 inbound links. We chose eight sites that in our qualitative expertise we view as top-tier criminal sites with significant barriers to entry and a high level of obscurity. For these eight sites, we measured an average of 8.7 domains with links to them, and the highest inbound link count for one of these sites was 15 — a stark contrast with the link counts for well-known sites. It is sites like these that are truly dark, and sites like these that have the most value for threat research on the dark web.



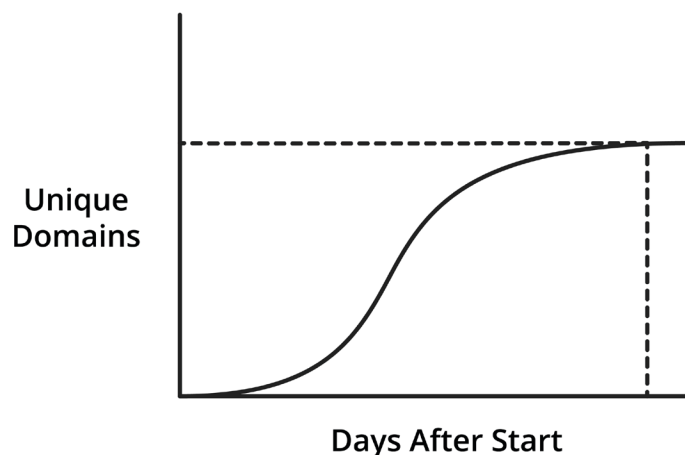
Inbound links into a popular dark web market (left) versus the entire network of inbound links for eight example sites we view as top-tier criminal sites (right).

Methodology

Tor (“The Onion Router”) is free, open-source software [initially developed by the U.S. military](#) and designed for [anonymous communication](#). The network consists of onion domains and connections between them in the form of direct links. For the purposes of our research, we use the term “dark web” exclusively to refer to websites on onion domains. Websites that are able to be reached without these kinds of specific software or network configurations are known as the surface web.

For years, Recorded Future has collected targeted dark web content that is relevant to our clients. For this project, we aimed to collect data from the whole Tor network without regard to whether a site likely contains useful information for threat intelligence data or is just junk. The approach was a web crawler (“spider”) that uses a Tor browser simulator. Our spider has been crawling new onion pages since December 2018. The spider was started on lists of known onions like the Hidden Wiki and as onion pages seen in Recorded Future’s existing data holdings.

Live Domains Found With Spider



Rate of encountering new onion domains in spider results.

Estimating the size of the Tor network required two procedures. First, we had to run the spider for enough time to crawl the majority of the onion sites. Second, we had to remove any duplicates from the count. For the former, we measured the rate of new, live domains found per day. This started out at around 2,000 new domains per day and leveled off about two months later. While we do still find some new domains, the overall rate is small enough that there is a high probability that we have found the vast majority of sites that are reachable from our current set of onion pages. It is possible that there are sites that are not reachable from our starting lists of onion sites, which the crawler will never find. While we cannot rule that out, the breadth of our starting lists gives us confidence that we have found the vast majority of onion sites that exist.

To count domains after data was collected, we removed any duplicates. One of the largest sources of duplication was 5,941 duplicates of the Deep Dot Web onion site. For an unknown reason, there are thousands of variations for the onion domain for this site using different placements of a non-printing character in the URL. The domains vary only by this inclusion of a unicode character that is not printable. This character, the “soft hyphen,” or “SHY” in unicode, is not visible in the URL bar when copying and pasting the domain. It also appears to have no effect on the returned site, with the same webpage returned regardless of the SHY character. From a human’s point of view, the modified Tor site URL will be an exact copy and will load the same site, but the non-printing characters are visible when the URL is rendered as raw characters, such as when viewing the raw HTML for the site containing the link.



```
href="http://deepdot&shy;&shy;&shy;3&shy;5w&shy;&shy;vmeyd5.onion/feed/"  
href="http://dee&shy;&shy;p&shy;&shy;dot&shy;&shy;&shy;35wvmeyd5.onion/comments/feed/"
```

It is unclear why Deep Dot Web has decided to use such a great number of different spellings of their domain that are all indistinguishable visually. One possible explanation could be that they are trying to prevent others from indexing their site. In 2015, Deep Dot Web reported having to aggressively shut down fake copies of their onion site that had the onion urls of popular markets [replaced by phishing links](#). We did not attempt to evaluate this strange behavior further, and just removed the duplicate domains from our counts.

We did not attempt to determine how many unique servers were underlying the domains we observed. Given that some hosting services may host thousands of sites, like in the case of Daniel's hosting service, we estimate that the number of different servers is in the hundreds or low thousands. Additional work would be required to obtain greater certainty.

To load onion urls, we only used browser-default ports 443 and 80. It is possible that a portion of failed urls will load correctly if requested via different ports. This is another potential future expansion on this work as the spider continues its ongoing scraping.

The contents of all live onion pages scraped with the spider are added to the Recorded Future® Platform.

Conclusion

The dark web is many things, but it is not the vast sprawling network of steely-eyed, hardened criminals that some might imagine it to be. Its 8,400 live onion domains are a tiny fraction of the surface web, with only 15% being live out of a mere 55,000 onion sites total. Onion sites are easy prey for attacks and scams like the “Thank You” typosquatting scam. It is more homogeneous, with 86% of onion sites primarily in English. The part of the dark web that does live up to its reputation is the set of top-tier criminal forums. In-link analysis of a select set of sites that we view as top-tier confirmed that they do indeed have less visibility, measured by a reduced number of links pointing to them.

If you’re curious for more, with the Recorded Future platform, you can see all of our spidered content yourself and get a deeper sense of what the dark web really is.

About Recorded Future

Recorded Future arms security teams with the only complete threat intelligence solution powered by patented machine learning to lower risk. Our technology automatically collects and analyzes information from an unrivaled breadth of sources and provides invaluable context in real time and packaged for human analysis or integration with security technologies.