First Monday, Volume 24, Number 12 - 2 December 2019

first monday
PEER-REVIEWED JOURNAL ON THE INTERNET

Online content moderation and the Dark Web: Policy responses to radicalizing hate speech and malicious content on the Darknet

by Eric Jardine

Abstract

De-listing, de-platforming, and account bans are just some of the increasingly common steps taken by major Internet companies to moderate their online content environments. Yet these steps are not without their unintended effects. This paper proposes a surface-to-Dark Web content cycle. In this process, malicious content is initially posted on the surface Web. It is then moderated by platforms. Moderated content does not necessarily disappear when major Internet platforms crackdown, but simply shifts to the Dark Web. From the Dark Web, malicious informational content can then percolate back to the surface Web through a series of three pathways. The implication of this cycle is that managing the online information environment requires careful attention to the whole system, not just content hosted on surface Web platforms *per se*. Both government and private sector actors can more effectively manage the surface-to-Dark Web content cycle through a series of discrete practices and policies implemented at each stage of the wider process.

Contents

Introduction

On 3 August 2019, a lone shooter walked into a Walmart in El Paso, Texas. The gunman, Patrick Crusius, then proceeded to murder 22 people, injuring 24 others. Twenty-seven minutes before the attack, Crusius reportedly posted his white nationalist manifesto onto an Internet forum called 8Chan (Blankstein and Burke, 2019). The short 2,300-word screed drew heavily upon a number of other radicalizing concepts freely floating around online (Arango, *et al.*, 2019; Heft, *et al.*, 2019).

Events like those in El Paso illustrate one facet of society’s growing information problem. ‘Fake news’, disinformation, misinformation, hateful content, and outright lies are all too common and spread all too quickly on the proverbial information superhighway (Grinberg, *et al.*, 2019; Howard, *et al.*, 2017). The Internet has radically shifted the media landscape, displacing many traditional informational gatekeepers, allowing for increased peer-production of content (Gehl, 2011). The commercial World Wide Web (WWW) now leverages data analytics to quantify clicks and create bespoke profiles that feed tailored content directly to users (Pariser, 2011). In practice, the resulting filter bubbles often create discordant views of social, economic, and political issues that can lead to both polarization and radicalization (Barberá, *et al.*, 2015; Colleoni, *et al.*, 2014; Nguyen, *et al.*, 2014; Resnick, *et al.*, 2013).

Good information is one of the core foundations of democratic governance and a free society (Brown and Duguid, 2017; Dewey, 1923; Jardine, 2019a; Lazer, *et al.*, 2018). The Internet presents myriad challenges to the quality of society’s information environment (Cohen-Almagor, 2011; Oboler, *et al.*, 2012). Since 2016 and the election of U.S. President Donald Trump, concerns over the informational role of social media platforms, content aggregation sites such as Reddit, or intermediaries such as Google have grown particularly acute. With additional allegations of persistent Russian interference in domestic U.S. electoral processes, pressure on content platforms to more carefully control their information environments has grown further still

(Gillespie, 2018). Despite existing protections against intermediary liability and rhetorical commitments to freedom of expression, the platforms have responded with more extensive content moderation (Gillespie, 2018). While these reactions are imperfect in their effectiveness, a long litany of Internet companies now routinely filter and remove content in ways that are largely patterned upon earlier copyright violation takedowns — eliminating terrorist propaganda, banning accounts of radical users, removing potentially offensive content, and otherwise attempting to control (for the better, in theory) society’s informational ecosystem.

This paper illustrates how each restrictive action by social media sites, content delivery networks (CDNs), domain registrars, and content aggregation services, among other actors, can result in a displacement. Restricting content on social media platforms, for example, does not necessarily eliminate the content, it simply moves the message elsewhere (N.F. Johnson, *et al.*, 2019). These shifts are not random. Increasingly, the actors behind malicious content respond to bans, delistings, and other heavy-handed content restrictions with movement to alternative platforms where centralized points of infrastructural control are often lacking (DeNardis, 2012). These more distributed points of the Internet have a reduced reach, but are far more resilient and harder to moderate, allowing otherwise radicalizing, false, or hate-filled content to persist. The most favored destination of banned content is the Dark Web — an anonymized and unindexed portion of the global Internet.

Moderating surface Web content can improve the information environment at one level, but the move is incomplete and partial unless content hosted on the distributed Dark Web is similarly managed. The trouble here is that by its very design, the anonymized Dark Web has fewer centralized points of control that can work as content moderators (Dingledine, *et al.*, 2004). While individual Dark Web sites can actively administer the sort of content that is hosted on their platforms (Gehl, 2018, 2016), major Internet hubs like Google (YouTube), Facebook, and Twitter do not exist in the same way on the networks of the Dark Web (Facebook does have a .onion address of its popular social media application). Often, harmful and discordant memes, tropes, and ideas originate in the underbelly of the Internet and then migrate into the mainstream via traditional media, illustrating society’s permeable information ecosystem (Brown and Duguid, 2017; Phillips, 2016). How then can society best manage harmful informational content that is displaced to the Dark Web?

This paper responds to this question by first explaining the Dark Web from both a technological and informational perspective. The next section points to a dynamic cycle of content generation-moderation-displacement-and percolation. The third section discusses, through the course of four subsections, how various actors can intervene in the surface-to-Dark Web content cycle to improve society’s information environment. The last section provides a conclusion to observations noted in this paper.

The Dark Web from an informational perspective

The Dark Web is a suite of technologies that render users anonymous on the Internet (Dingledine, *et al.*, 2004). The Onion Router (Tor) is the most commonly used portal to the Dark Web, with roughly 400,000 direct connections per day in the United States and almost 2.5 million daily clients globally (Tor Project, 2018). Other systems such as I2P or Freenet exist, but are far less commonly used (Graham and Pitman, 2018).

The technology of Tor works as an overlay network on the Internet. As a result, Tor leverages the basic infrastructure and protocols of the Internet (such as a home ISP-provided Internet connection and TCP/IP) to relay a person’s query through a minimum of three randomly selected nodes in the globally distributed and volunteer-based Tor network. The process of hops creates the anonymity of the system. Alice might want to access content hosted by Bob, but rather than directly querying this information, Alice’s request is sent to a guard node (entry node) in the Tor network. This node then relays the request to the next node in the system, which passes it on to a third exit node. The exit node then unwraps the final layer of encryption and queries the content hosted by Bob. The information is retrieved and then relayed back to Alice through the Tor network by way of another series of randomized hops (Tor Project, 2019b).

Anonymity is produced by disassociating a user from the content they are trying to view. The entry node might record Alice’s IP address, but it does not know the final destination of the query. The exit node might know the content that is being requested (or, minimally, the address of the site hosting the content), but not who initiated the query. While attacks on the Tor network can sometimes de-anonymize blocks of traffic (A. Johnson, *et al.*, 2013), the system is generally a robust anonymity and censorship circumvention tool with wide appeal, especially for those engaging in illegal conduct, the privacy conscious and individuals in repressive political regimes (Jardine, 2018a, 2018b, 2015; Lindner and Xiao, 2018).

The anonymity of Tor allows for three interrelated informational functions. First, users of Tor can host content on .onion domains (the proverbial Darknet). These hidden service sites are hosted at random rendezvous points in the Tor network and housed on unique alphanumeric .onion domains. While estimates of the total number of available hidden services fluctuates due to the high rate of churn in available content (Owenson, *et al.*, 2018), the Tor Project records around 75,000 active .onion address during 2019 and systematic empirical estimations place the number of hidden services in the 35,000–65,000 range (Owen and Savage, 2015; Tor Project, 2019a).

Unlike the surface Web, this comparatively modest pool of .onion addresses are not indexed by search engines, meaning key term searches and similar functions from the Web are not an effective strategy for navigating available content. Instead, users need to either know the address that they want to visit or frequent a directory, wiki, or link aggregation site such as the now shutdown DeepDotWeb.

Several studies have indexed available hidden service content hosted on the Tor network (Faizan and Khan, 2019; Intelliagg, 2016; Moore and Rid, 2016; Owen and Savage, 2015). These studies provide supply-side indicators of available Dark Web content, recording effectively what is on offer upon the Darknet (Jardine, 2019b). Sites dedicated to drugs routinely represent the largest plurality of hidden service content (Moore and Rid, 2016; Owen and Savage, 2015). This result mirrors the prevalence of drugs upon individual Dark Web cryptomarket sites such as Silk Road, where as much as 90 percent of the ten most popular available products were related to drugs (Christin, 2013; Martin, 2014; Soska and Christin, 2015). Hosted alongside this drug-related content are sites dedicated to extremism, child abuse content, cryptocurrencies, fraud, information sharing, wikis, whistleblowing, and a plethora of other topics.

The second informational function of the Dark Web is the Tor browser, which can act as a pathway to hidden services content. Normal Web browsers cannot access Darknet content. Instead, to access the information hosted on hidden services, users have to use the Tor browser. The Tor browser can be easily downloaded from the Tor Project Web site (https://www.torproject.org) and is available for free. The Tor browser works as a powerful anonymity-granting informational tool by allowing users to access hosted hidden services content.

Aggregate site visits to .onion domains have been tracked by academics, providing a demand-side indicator of how available Darknet content is being consumed (Jardine, 2019b; Owen and Savage, 2015). In 2015, Owen and Savage volunteered nodes into the Tor network to categorize content and then track the pattern of site visits. Site visits tended to cluster on certain types of available hidden services — a result similar to users clustering on specific Dark Web cryptomarkets and the Internet as a whole (Barabási, 2014; Barabási and Albert, 1999; Christin, 2013; Jardine, 2019b, 2017; Soska and Christin, 2015). In this case, the two percent of available .onion sites dedicated to child abuse content received over 80 percent of recorded site visits. Traffic to other categories of content, such as drugs or wiki sites, made up comparatively miniscule proportions (Owen and Savage, 2015).

This pattern of site visits on Tor hidden services suggests a clear prevalence of malicious activity on Darknet sites. But the flow of traffic to hidden services also reveals the third informational function of the Dark Web: the Tor browser can be used to anonymously engage with surface Web content. Indeed, most Tor traffic flows to the surface Web, not Tor-hosted hidden services. The Tor Project, for example, published a blog in 2015 clarifying just how much overall network traffic actually goes to hidden services. According to their preliminary results, only around 3.4 to 6.1 percent of Tor traffic went to hidden services (Tor Project, 2015). The rest (upwards of 96.6 percent) flowed through the Tor network to surface Web sites. Essentially, many Tor users treat the Tor browser as a hyper-private, anonymity-granting Web browser, a step far more advanced than Chrome Incognito, using DuckDuckGo, or even employing a commercial VPN.

In combination, the suite of technologies underpinning the Dark Web allow users to engage anonymously with information as a host (hidden services), a consumer or producer of anonymously hosted content (hidden services visitor), or as an anonymous consumer and distributed producer of surface Web content (Tor browser to surface Web sites). These functions connect the Dark Web with the surface Web in an informational chain, with clear implications for attempts to manage the information environment online.

An emerging content cycle: Production, restriction, displacement, and percolation

Many users do not realize that their online experience is heavily curated (Powers, 2017; Rader and Gray, 2015). Yet moderation of content in any online space is inevitable. Historically, this moderation was limited to simple procedures, such as the sequence of content presentation (*e.g.*, oldest to newest). Increasingly, content curation has assumed a wider commercial aim as platforms attempt to drive retention of users and ad-based clicks (Gillespie, 2018; Jardine, 2017; Zuboff, 2019). With the rise of so-called ‘fake news’ and increasing public attention on the integrity of information online, the major Internet platforms of the surface Web have become even more active in curating the content available through their systems (Gillespie, 2018).

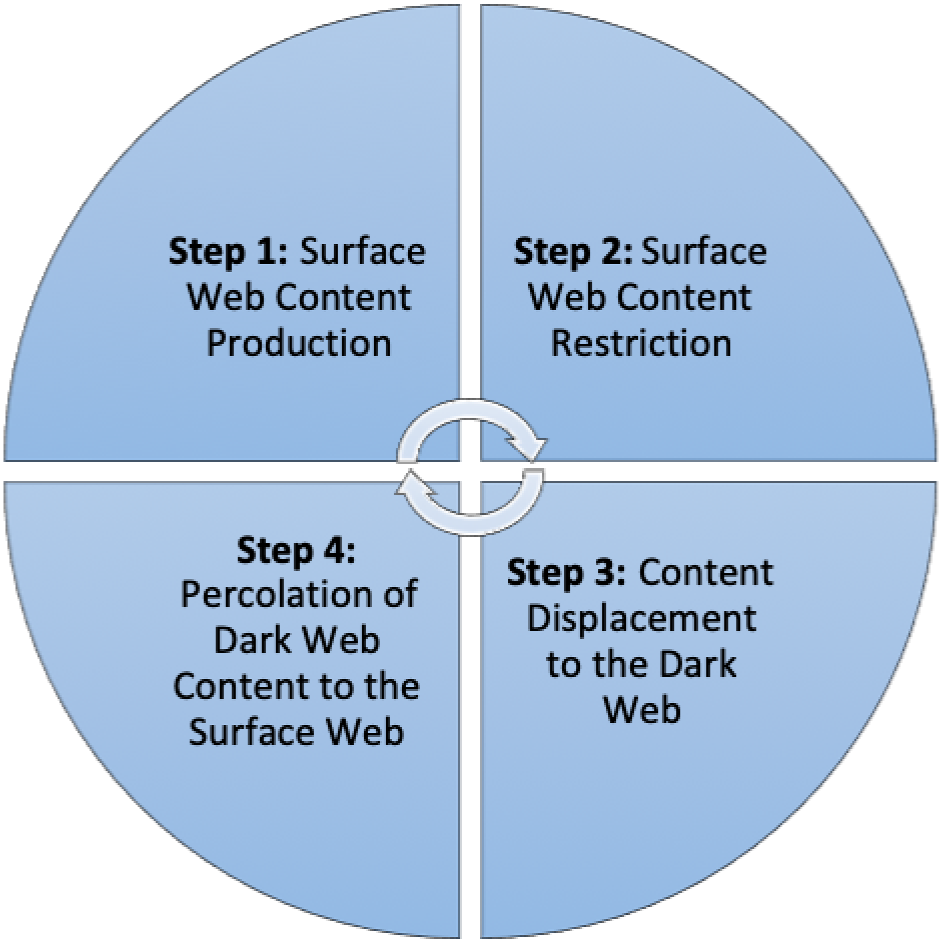

Step 1: Surface Web Content Production
Step 2: Surface Web Content Restriction
Step 3: Content Displacement to the Dark Web
Step 4: Percolation of Dark Web Content to the Surface Web

Figure 1: The surface-to-Dark Web content cycle.

This active curation/moderation process is cuing off an increasingly routine (non-linear or deterministic) cycle of content productionmoderationdisplacementand percolation (see Figure 1). In this cycle, surface Web platforms often act as an initial host for potentially false, hateful, or otherwise malicious content. The platform hosting the content can then move to limit the information through a variety of strategies such as delisting, demonetization, account bans, and de-hosting, among others. Often, while this content might migrate via so called "hate highways" to other surface Web platforms (N.F. Johnson, *et al.*, 2019) or in the short run be simply reposted on the same site, much of the most heavily moderated content eventually moves to the Dark Web in a process of displacement. Lastly, many of the ideas preserved and fostered in this distributed and anonymous environment then percolate back up to the surface Web, potentially starting the cycle anew. Small vignettes can illustrate the stages of this process.

In 2017, for example, the Neo-Nazi site, *The Daily Stormer*, helped to organize the "Unite the Right" protests in Charlottesville, Virginia. After the surrounding protests grew fatal, major Internet platforms responded by delisting the site. In August of 2017, GoDaddy delisted *The Daily Stormer* domain (CBS News, 2017). *Daily Stormer's* administrators first attempted to migrate their site to a new domain provided by Google, but the new location was also rapidly delisted and their content was banned from YouTube (Reuters, 2017). Following these cascading moderations, *Daily Stormer* administrators eventually moved the site and its hate-filled content to a .onion domain located on the Tor-hosted Darknet (Ling, 2017).

8Chan provides another example of the first three stages of the surface-to-dark Web content cycle. Shortly after it became clear that the mass shooter in El Paso, Texas, for example, had posted his white nationalist manifesto to the image board, many major Internet companies again moved to de-list the site — steps that had originally been proposed after similarly tragic shootings in Pittsburgh, Pennsylvania, Christchurch, New Zealand, and Poway, California (Online Hate Prevention Institute, 2019). CloudFlare, which provided DDoS protection for 8Chan through its CDN, removed its services (Timmer, 2019). Several other key Internet service providers followed suit, with 8chan also losing its domain name registrar (Cox, 2019). While shut down for a time, mirrored versions of the site then migrated to distributed Dark Web infrastructure and continue to act as a host for informational content, somewhat limiting ease of access but nevertheless preserving the hosted material online (Cox, 2019).

The first three steps of the surface-to-dark Web content cycle are easiest to directly observe with the recent rash of surface Web moderation efforts and the resulting displacement of content to the Darknet. The last step of the cycle — informational percolation from the Darknet to the surface Web — is less easy to directly observe without a longer duration between delistings, restrictions, and bans and extensive data collection efforts. Yet, three pathways exist by which malicious content that has been displaced to the Darknet can percolate back to the surface Web: 1) direct resurgence on the surface Web; 2) migration through multi-media content pieces (pictured memes; infographics; slogans); and, 3) percolation through idea networks.

In the first case, the operators of delisted sites or banned accounts might work opportunistically to again find purchase on the surface Web (N.F. Johnson, *et al.*, 2019). For example, *The Daily Stormer* continued after its delisting to try to find a favorable top-level domain to host its content. During this process, the site was briefly available on .ru (Russia), .al (Albania), .at (Austria) and .ws (Western Samoa) country-level TLDs (Lavin, 2018). Likewise, at a more micro scale, social media users tied to banned accounts can, with varying degrees of ease depending on the platform in question, simply return to the site under a new handle or username.

In the second pathway, the persistent core of content hosted on the Dark Web can sustain a community of ideas that produces novel (often collaboratively derived) multi-media memes, tropes, and slogans that can then migrate back into the surface Web. A user could, for example, frequent a displaced site on the Darknet, see a new multimedia meme and then use the Tor browser to anonymously post the image on a surface Web discussion forum, messaging board, or social media site (Oboler, 2012). Similar migration of memes from the recesses of the Internet to the mainstream has happened before, from trolls getting Opera Winfrey to say “9,000 penises” on air in 2008 to tropes about the birth place of U.S. President Barrack Obama (Phillips, 2016). This pathway of percolation is often exacerbated by mainstream media coverage (Phillips, 2016), as was the case in 2018 when the “NPC meme” depicting personality devoid automatons as a description of political liberals began being heavily shared on Twitter following coverage in the *New York Times* (Alexander, 2018).

The last pathway of percolation is similar to the second pathway, but more ephemeral. It captures, effectively, the movement of memes in its classic sense as defined by Richard Dawkins as a “unit of cultural transmission” (Dawkins, 2016) — that is, it is the more amorphous spread of ideas, as opposed to the sharing of reified cultural artifacts such as a hate-filled multimedia picture or video (Oboler, 2012). Informational content is two sided: it can always be both produced and consumed. When consumed it becomes ideas in people’s minds, which can reform or reinforce opinions and shape behavior. Here, users of the Tor browser might view displaced content on the Darknet and then have their worldview gradually shifted. Indeed, leaked style guides for the *Daily Stormer* indicate that this percolation of ideas is key:

> The goal is to continually repeat the same points, over and over and over again. The reader is at first drawn in by curiosity or the naughty humor, and is slowly awakened to the reality by repeatedly reading the same points. We are able to keep these points fresh by applying them to current events (cited in Feinberg, 2017).

In short, from an informational perspective, the surface Web and the Dark Web exist as an interconnected whole. Content moderation efforts on the surface Web are at best partial, with the Dark Web providing an alternative core from which malicious content producers and consumers can continue their routine. Working to correct for society’s growing information problem requires attention to the whole system, not just the happenings on the major surface Web hubs.

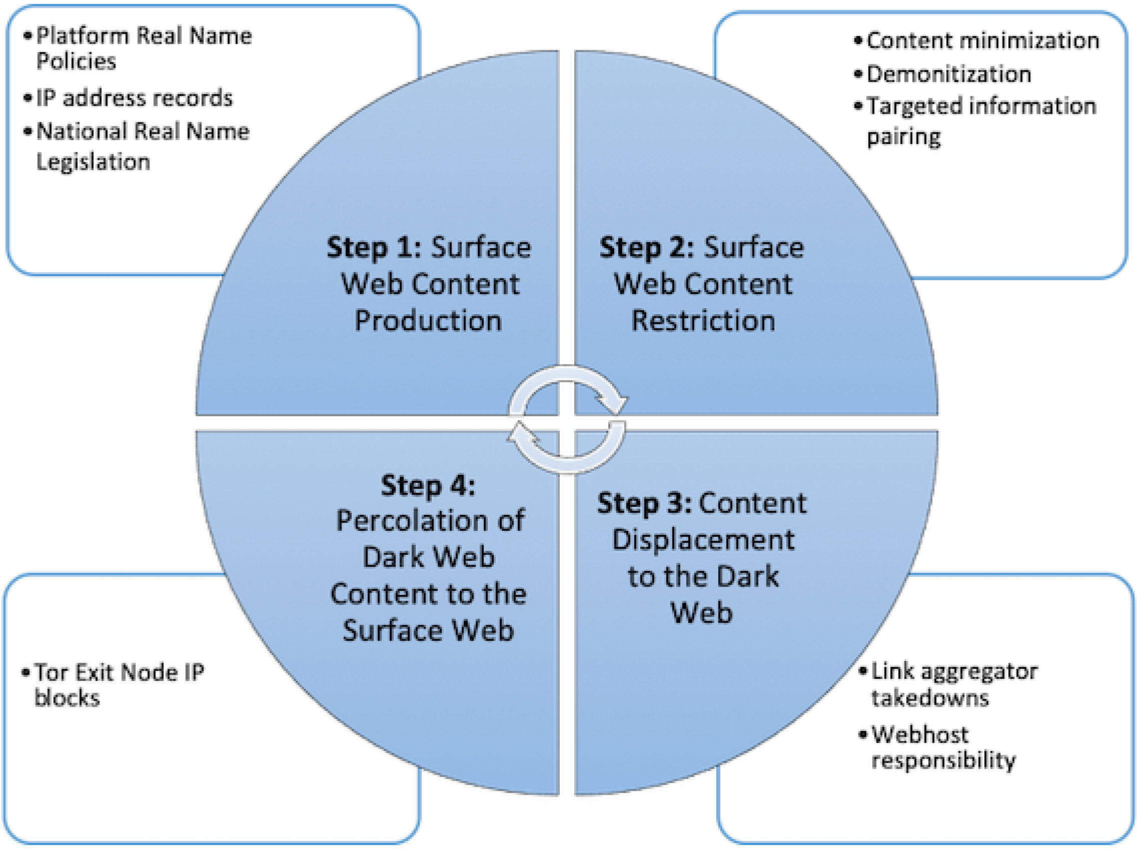
Tackling the surface-to-Dark Web content cycle

Viewing the online information environment holistically suggests that there is an essential tradeoff between the reach and resiliency of information online. Content hosted on commercial surface Web platforms can have a tremendous reach, but is invariably beholden to the whims of the platforms themselves. Content on Facebook, for example, can reach an extraordinarily large audience of up to 2.7 billion people. But, especially over more extended time intervals, Facebook can effectively exclude content from the newsfeed of users, making hosted content hard to find and significantly reducing the odds of spontaneous exposure.

In contrast, content hosted on the Darknet is often difficult to find and taps into a smaller (if potentially more motivated) overall user base, but is also hosted in an anonymized and distributed way that maximizes its resiliency. As a simple descriptive point, content on the Darknet is not indexed and keyword searchable. Limited search functions, when compared with the surface Web, make finding Darknet-hosted content comparatively challenging. Additionally, simply by dint of numbers, content on the Darknet is likely to be seen by fewer eyes. Compared to Facebook’s 2.7 billion users, there are only around 2.5 million Tor clients per day globally (which is not a measure of unique users, but merely client sessions). Additionally, up to 96.6 percent of that traffic goes to the surface Web instead of Darknet .onion addresses, implying that only around 85,000 Tor clients per day venture toward content hosted on the Darknet proper (Tor Project, 2018, 2015).

Yet, what is lost in reach is returned in resiliency. Content hosted on the Darknet is anonymized and distributed, making it less subject to points of centralized control, be they the major content platforms, infrastructure providers, or governments exercising sovereign authority over their territory. In the end, content hosted on Darknet .onion hidden services is harder to find and will appear before fewer eyes, but is potentially more persistent overall than content on the surface Web.

With such a reach/resiliency tradeoff in mind, attempting to address society's information problem can leverage discrete practices and policies at various points in the overall surface-to-Dark Web content cycle. A multitude of actors, as befits a multistakeholder system (Bradshaw, *et al.*, 2015; Raymond and DeNardis, 2015), can intervene in different ways throughout the information cycle to minimize fake news, and hateful or radicalizing content. The next four sections break down the menu of policy options according to each discrete stage of the cycle. Figure 2 presents a summary of the points of intervention at each discrete stage.



•Platform Real Name Policies
•IP address records
•National Real Name Legislation
•Content minimization
•Demonitization
•Targeted information pairing
Step 1: Surface Web Content Production
Step 2: Surface Web Content Restriction
Step 4: Percolation of Dark Web Content to the Surface Web
Step 3: Content Displacement to the Dark Web
•Tor Exit Node IP blocks
•Link aggregator takedowns
•Webhost responsibility

Figure 2: Policies/practice to manager the surface-to-Dark Web content cycle.

Step one: Malicious content emergence on the surface Web

The first step in the surface-to-Dark Web cycle is the initial emergence of malicious content on surface Web platforms — platforms such as domain registrars, Facebook or Twitter's newsfeed, Reddit threads, Google search results, or YouTube recommended videos, to name a few. Countering the emergence of radicalized or hate-filled content at this stage of the cycle could be done through several steps that are unified around a single principle: limiting (to varying degrees) anonymous participation/production of content.

Anonymous participation on the Web certainly does not necessarily equal negative participation. People can be anonymous yet still behave well. However, from early parables of the Ring of Gyges in Plato's *Republic* (Plato, 1974) and H.G. Wells' tale of *The Invisible Man* (Wells, 1927) to modern psychological research into online deindividuation effects (Hinduja, 2008; N.A. Johnson, *et al.*, 2009; Lowry, *et al.*, 2016), disassociating a person's actions from their identity often undermines moral behavior. The early Web, before its commercialization and the expansion of a complex surveillance infrastructure (Zuboff, 2019), presented an initial testbed for whether anonymity leads to negativity, at least some of the time.

Often, but again not always, anonymity does indeed foster a context for bad behaviors. As Lawrence Lessig put it when recounting how an anonymous poster, IBEX, in an early run of his "The Law of Cyberspace" course, destroyed the emergent community of an online discussion group, "Just as anonymity might give you the strength to state an unpopular view, it can also shield you if you post an irresponsible, or slanderous, or hurtful view. ... [Anonymity] had made the community what it was. But the same anonymity that created the community gave birth to IBEX as well, and thus took the community away" [1].

Controlling anonymity on surface Web platforms can take a variety of forms and include actors ranging from the platform operators to government, depending on the context. Facebook's real name policy is one example of an anonymity-tempering step. Ensuring that users of a site create accounts that are linked to a physical identity can limit bad behavior stemming from the protective shield of anonymity. Such a policy may not be legally or politically practicable in all settings, since national rules about privacy and free expression vary widely by country (Kastrenakes, 2018).

Yet platform real name policies and account verification procedures do tend to improve the quality of online discussion. For instance, the online technology news Web site *TechCrunch* switched from an anonymous comment function to an identity-verified comment function on 1 January 2010. Analysis of data from both before and after the switch shows that the more a poster's identity is revealed, either voluntarily in the pre-verification dataset or obligatorily in the post-change dataset, the better the comments. Stated simply, the more a person's identity is tied to their online activities, the more relevant their posts become and the "less swearing, less anger, more affect words, more positive emotion words and less negative emotion words" are used in discussion comments [2].

A subtler step than requiring the use of a real name to engage with surface Web content is the collection of Internet Protocol (IP) address information from users, even if those individuals post under archetypically pseudonymous handles. Given the underlying IP infrastructure of the Internet, anonymous speech beyond the screen can only be approached if users can employ privacy-enhancing technologies, such as virtual private networks (VNPs) or Tor. These technologies allow users to disassociate their online behaviors from their IP addresses. When using a normal Internet connection, technical anonymity — as oppose to a sense of screen anonymity to be found behind a pseudonymous handle — is far from guaranteed. Typically, a user's ISP dynamically assigns an IP address to a user for a session. That address is recorded and stored in a dataset somewhere, alongside a physical world address and an account name. Likewise, Web site operators often track the IP addresses of devices that visit their servers, helping to create a digital profile of users. These steps hamper the exercise of truly anonymous speech, which is far more difficult to obtain online than in the physical world. Given the nature of this infrastructure and assuming reasonable probable cause, governments can subpoena Web site operators and ISPs to provide a record of who visited what site and posted what content, illustrating the typical lack of real online anonymity.

A Baltimore Department of Homeland Security investigation into user-initiated chatter on the content aggregator Reddit showcases the point. In 2015, federal agents issued an administrative warrant to Reddit asking the platform to turn over records for five users of the popular r/darknetmarkets subreddit. These users posted on the subreddit under pseudonyms to avoid easy detection, while discussing where to find Dark Web drug cryptomarkets and what to purchase on these markets. Reddit, which stores IP address information of users for up to 90 days, eventually complied with the warrant and turned this identifying information over to authorities. Despite not signing their names, the surface Web posts by these five individuals were not anonymous (Knibbs, 2015).

Beyond the policies and practices of platforms, governments could in some cases get more directly involved with the aim of legislating rules that minimize anonymous online participation. Such rules might be legally fraught or politically complicated in many jurisdictions. In the United States, for example, the First Amendment guarantees a weak right to anonymous speech online, as initially outlined in a number of off-line cases involving anonymous political leaflets (Diaz, 2016) and which are now deemed to generally apply online (*Reno v. American Civil Liberties Union*, 1997). Of course, qualifications to the general rule often apply, such as in the case of speech deemed to be "integral to criminal conduct," "fighting words," child abuse content, "obscenity," incitements to "crime" or "violence," and "true threats," all of which can obviate potential legal protections for anonymous online participation in the United States (Diaz, 2016; Volokh, 2015). Likewise, within the European Union, the *Europe Declaration on Freedom of Communication on the Internet* includes a specific provision (Principle 7: Anonymity), which stipulates that: "member states should respect the will of users of the Internet not to disclose their identity." Such a principle might inhibit legislation dictating real name policies or mandatory IP address collection and retention by platforms, but is also tempered by a second Principle 7 stipulation wherein respect for the online anonymity of users "does not prevent member states from taking measures and co-operating in order to trace those responsible for criminal acts" (Council of Europe, 2003).

While the legal and political hurtles surrounding a restriction of anonymous online activities are not inconsiderable, regulations reducing anonymous online participation have tended to improve the civility of online discussion. One empirical example of the effect of such a policy comes from South Korea's 2007 passage of a Real Name Verification Law. The law was aimed at reducing bullying behavior in online posting. In practice, the law reduced aggregate "uninhibited" behaviors, such as posting with swear words, and did not, over the long term, have an effect on the volume of posting or online activity (Cho, 2013; Cho, *et al.*, 2012). Put otherwise, the law positively changed, in aggregate, how people posted, not how much they posted. Of course, the effects of such a policy depends on the wider political context within which the rules are implemented. A real name policy in a liberal democracy might reduce abusive posts, but equivalent rules in less politically free regimes can stifle political discussion, as was the case with a Real Name Registration Policy on Weibo, which led some micro bloggers to discuss politics less frequently than before (Fu, *et al.*, 2013).

Step two: Moderating content on surface Web platforms

Even as platform design choices and national regulatory regimes can manage anonymity and improve the ways in which people behave online, Internet platform operators will invariably, at some point, have to make choices about how they are going to curate or moderate the content that does get displayed on Web sites, users feeds, search results, or recommendation lists (Gillespie, 2018). While content platforms often prefer to present themselves as neutral in the content moderation process, each must make a variety of choices that affect how users engage with their systems. Social networking sites, search engines, and content aggregation sites, for example, need to decide how their algorithms rank, weight, and ultimately display content to users. These choices often happen in the background and many users remain largely unaware that their online experience is heavily moderated, despite the ubiquity and necessity of such steps (Eslami, *et al.*, 2015; Powers, 2017; Rader and Gray, 2015).

The precise nature of these steps affects both the reach of potentially malicious content and the probability that such informational content will be displaced to the Darknet. Steps that outright remove content, ban users, or de-platform services, as happened with 8Chan and *Daily Stormer*, limit the reach of information, but are very likely to lead to displacement. A number of alternative moderation steps, however, can either limit the reach of malicious content or blunt a user's receptivity to the ideas being promulgated, without necessarily causing a movement of content from the surface Web to the Darknet. Content minimization, demonetization, and informational inoculation each work to reduce the reach (and uptake) of potentially malicious content.

Content minimization approaches, for example, can reduce the reach of information, but stop short of outright content removal. Social media sites are expert at such steps. The deep fake video of Speaker Nancy Pelosi, for example, was not removed from Facebook. However, it was eventually minimized to such a degree that it no longer appears on user newsfeeds, making direct navigation to the page of someone who had already shared the video the only way to view the content (Wagner, 2019). Such a policy uses algorithmic sorting to minimize the reach of content, but gives rise to fewer risks of displacement to alternative sites or the Dark Web than outright removal.

Demonetization, used extensively by YouTube, does not limit the spread of content, but does condition the incentives that might surround the production of content in the first place. Content demonetization removes a potential financial reward associated with the production of malicious content. Of course, several limitations to the effectiveness of demonetization exist. First, the boundaries of demonetization can catch content that is perhaps not truly fake news, malicious hate speech, or otherwise objectionable content. Second, platform demonetization can only remove ad-based financial rewards for content producers. It does nothing about potential alternative channels of financial support such as direct donations to content producers via Patreon or similar peer-to-peer financial services. Third, demonetization targets the financial incentives surrounding particular videos (on say YouTube), but content producers could balance the loss of revenue from one video against an increase in subscribers which will produce more economic returns over the longer term. Lastly, the producers of genuinely malicious, fake, or otherwise hate-filled content might not have financial reward as their primary motivation and will happily produce demonetized content for surface Web sites in order to spread information as widely as possible, even if they need to do so at a strict financial loss.

The last moderation step that can minimize the reach of malicious information targets neither exposure to nor production of content — as with minimization or demonetization — but instead aims at blunting user receptivity to the message. Such processes work similar to viral immunizations (Cook, *et al.*, 2017; Roozenbeek and van der Linden, 2019; van der Linden, *et al.*, 2017). Users might see malicious or misleading content on their newsfeed, say, but that content is then algorithmically paired with additional information that works to blunt its effects. Such additional information could be as simple as a flag attesting to the potential lack of integrity found in a particular post or story. The additional information could also involving the pairing of potentially malicious informational content with hyperlinks or stories presenting alternative narratives. For example, posts denying climate change could be paired with information pointing to the scientific consensus that surrounds climate science. This sort of informational inoculation tends to blunt user belief that climate change is not real or lacking in scientific validity (van der Linden, *et al.*, 2017), although such actions can also potentially increase user perceptions of bias toward social networking service newsfeed algorithms and lead to greater calls for content moderation.

Step three: Tackling displacement

While certain moderation policies can be used by various actors at points of concentration and control on the surface Web to minimize the reach of malicious information without causing a displacement to the Dark Web, some content will invariably be de-platformed, some user accounts will be banned or IP blocked, and truly hateful content will eventually migrate, when these steps are undertaken, to the anonymized and distributed architecture of the Darknet. At this stage of the surface-to-Dark Web content cycle, centralized points of infrastructural control are weaker (DeNardis, 2012). Nevertheless, a couple of discrete steps could be implemented that leverage the distributed and unindexed nature of the Tor-hosted Darknet to both counter the resiliency of content on the Dark Web and further blunt its already diminished reach.

The first step that can be taken to manage content displaced to the Darknet is to leverage informational artifacts that are produced as a result of the unindexed nature of Tor .onion addresses. The Dark Web is not keyword searchable in the same way as the Web. In practice, the limited discoverability features of the Dark Web restrict what content users can access overall, but also how users find what they are looking for. Surface Web link aggregation or wiki sites are a common feature of the Darknet content discovery process. For example, if a user wants to visit a cryptomarket to purchase drugs, they would need to find the unique .onion address that is currently hosting the site. Link aggregation sites and wikis provide a primary means of doing so.

While these sites are useful for law enforcement in identifying locations of criminal misdeeds, they also potentially illustrate key pressure points that can disrupt the Dark Web ecosystem and its informational connections to the surface Web (Zabihimayvan, *et al.*, 2019). One such example is the 2019 takedown of the Dark Web news and link aggregator site DeepDotWeb by the FBI and Europol. The operators of the site had been reportedly profiting from paid .onion url placement on their site. The shuttering of the site and the arrest of the site operators was justified by the charge that the administrators were knowingly profiting from the criminal enterprises linked to the activity at the other end of the addresses that they were curating (Greenberg, 2019). Similar takedown initiatives can be undertaken to target the operation of directory or link aggregation sites based upon legal principles such “contributory infringement,” if the operator knows they are linking to stolen information, or if it can be shown that the operators are in other ways “aiding and abetting” criminal enterprises (Dalal, 2010). Of course, variation in national laws and the location of the operators of such sites will determine the practicality of such a policy as a way to limit user access to Darknet content.

The second practice that could be used to disrupt the displacement of content to the Darknet targets the hosting of hidden services. Individuals wishing to establish a .onion address can take two non-exclusive paths to hosting the content. First, they can design and host the hidden services site locally upon their own machines. Second, and far more commonly, they could also host their site on a third-party service. The first route is both technically and logistically complicated for individuals, but minimizes the amount of control others can exert over the resultant hidden service. The second route is technically easier in many ways, but makes the site operator subject to the whim (and fate) of the hosting service provider.

Daniel’s Hosting, a Web hosting company, shows the overarching popularity of the second route to hosting hidden services content. In November 2018, malicious actors compromised the servers behind Daniel’s Hosting, potentially through the exploit of a php vulnerability. As a result, some 6500 hidden services sites hosted on Daniel’s Hosting were destroyed (Cimpanu, 2018). While the Tor Project own metrics point to there being around 90,000–100,000 hidden services active at this time (Tor Project, 2019a), the extraordinarily high rate of intraday churn in hidden services suggests that the real total volume was likely only around 40 percent of this official estimate (Owenson, *et al.*, 2018). At the high end, these discounted numbers imply that there might have been around a total of 40,000 .onion hidden services in operation in November of 2018 when Daniel’s Hosting fell. Potentially, then, the disruption of this single hosting provider resulted in a loss of around 16.25 percent of the entire Tor-hosted Darknet.

The concentration of .onion addresses on centralized hosting services also highlights another potential point of content moderation and control within the Darknet ecosystem. The operators of the various hosting services could identify the .onion sites hosted on their servers and do a periodic scan to determine the nature of the hosted content. From there, content that violated local laws or that promulgated hate speech and other radicalizing content could be taken down. Daniel’s Hosting again provides a good example. The hosting information page indicates the rules associated with using the service:

> “Rules
>
> - No child pornography!
> - No terroristic propaganda!
> - No illegal content according to German law!
> - No malware! (*e.g.* botnets)
> - No phishing, scams or spam!
> - No mining without explicit user permission! (*e.g.* using coinhive)
> - No shops, markets or any other sites dedicated to making money! (This is a FREE hosting!)
> - No proxy scripts! (You are already using TOR and this will just burden the network)
> - No IP logger or similar de-anonymizer sites!
> - I preserve the right to delete any site for violating these rules and adding new rules at any time.
> - Should you not honor these rules, I will (have to) work together with Law Enforcement!
>
> (Daniel’s Hosting, 2019).”

Of course, not all Web host operators will be willing to work to moderate hosted Darknet content, as was the case with Freedom Hosting, which ran a huge proportion of the child abuse sites hosted on the Tor network back in early 2010s (Hampson and Jardine, 2016). Many individuals can also, in a pinch, opt to host locally, if professional hosting services start to clamp down on the content housed on .onion addresses. Yet, individual web hosting providers such as ‘Daniel’ can still work to reduce the resiliency of Darknet content by monitoring and potentially moderating what is being housed on their systems. Additionally, within some jurisdictions, governments could even regulate local Web hosts to compel them to do random inspections hosted .onion content, deleting any that are in violation of national laws.

Step four: Controlling percolation

While content on the Darknet is significantly more resilient than surface Web content, the limited reach of anonymous content on the Dark Web makes it an undesirable platform for terrorist/extremist propaganda and the dissemination of hate speech. However, through the pathways pointed to before, information on the Dark Web can leech back onto surface Web platforms, where the information can again spread to the multitudes.

Mitigating content percolation from the Darknet is complicated by the highly diffuse nature of the spread of ideas. Users of the Tor browser who view Darknet content on say *The Daily Stormer* may then bring these ideas, tropes, and memes with them back to Reddit, Twitter, or any number of other sites. Steps by surface Web platforms that restrict anonymous online participation could again help to manage the behavior of users and the percolation of ideas. But one additional step could be done that would work beyond real name policies or IP address collection to actually blunt one of the outstanding informational uses of the Tor Browser: namely, the ability to use Tor to engage and create surface Web content.

An effective moderation step in this case is to block Tor exit node traffic to surface Web sites, thereby limiting the extent to which users can engage and produce content behind the wider protections of the Dark Web. Such a practice would largely fall to either content delivery networks (CDNs) or the major platforms operating their own Web servers. CloudFlare made just such a move in 2016, observing that, "Based on data across the CloudFlare network, 94% of requests that we see across the Tor network are *per se* malicious (Price, 2016)." In response to this observation (which was contested by the Tor Project), CloudFlare imposed time-consuming CAPTCHAs on all traffic coming from known Tor exit relays and attempting to access Web sites protected by the company. Similar steps voluntarily taken by other dominant Internet platforms could further restrict the usefulness of Tor as a tool of anonymous speech and would blunt, but in no way eliminate, the percolation of malicious content from the Darknet to the surface Web.

Conclusion

The Dark Web is certainly not all malicious content (Gehl, 2018, 2016; Hampson and Jardine, 2016; Jardine, 2018b, 2015; Lindner and Xiao, 2018). Indexing efforts routinely find that only around a half of available content is illicit (Faizan and Khan, 2019; Intelliagg, 2016; Lindner and Xiao, 2018; Moore and Rid, 2016; Owen and Savage, 2015). It is especially useful in highly repressive regimes as a tool of privacy and censorship circumvention (Jardine, 2018b).

Yet the proliferation of malicious content on the surface Web is leading to increased moderation by major Internet platforms and providers. This moderation results in a displacement of content, with shifts from the comparatively controllable surface Web to the distributed, anonymous, and harder to control Dark Web. From this more protected position, malicious informational content can then percolate back up to the surface Web. What remains are piecemeal moderation steps at every stage of the cycle that can manage and correct for the excess of the Dark Web, without providing any permanent solutions. Society's information ecosystem is one big system and needs to be treated as such. A number of policy steps by government and surface web actors can disrupt the surface web-to-Dark Web content cycle and help to improve society's information environment.

About the author

Eric Jardine is Assistant Professor of Political Science at Virginia Tech, and Senior Fellow, Centre for International Governance Innovation (CIGI) in Waterloo, Ontario, Canada.
E-mail: ejardine [at] vt [dot] edu

Acknowledgements

I would like to thank *First Monday's* anonymous reviewers for their insightful comments and Leanna Ireland, Priya Dixit, and Rachel Hulvey for reading earlier drafts of the paper. Lingering mistakes are my own.

Notes

[1] Lessig, 2006, pp. 104, 106.

[2] Omernick and Sood, 2013, p. 533.

References

J. Alexander, 2018. "The NPC meme went viral when the media gave it oxygen," *Verge* (23 October), at https://www.theverge.com/2018/10/23/17991274/npc-meme-4chan-press-coverage-viral, accessed 13 November 2019.

T. Arango, N. Bogel-Burroughs, and K. Benner, 2019. "Minutes before El Paso killing, hate-filled manifesto appears online," *New York Times* (3 August), at https://www.nytimes.com/2019/08/03/us/patrick-crusius-el-paso-shooter-manifesto.html, accessed 13 November 2019.

A.-L. Barabási, 2014. *Linked: How everything is connected to everything else and what it means for business, science, and everyday life*. New York: Basic Books.

A.-L. Barabási and R. Albert, 1999. "Emergence of scaling in random networks," *Science*, volume 286, number 5439 (15 October), pp. 509–512.
doi: https://doi.org/10.1126/science.286.5439.509, accessed 13 November 2019.

P. Barberá, J.T. Jost, J. Nagler, J.A. Tucker, and R. Bonneau, 2015. "Tweeting from left to right: Is online political communication more than an echo chamber?" *Psychological Science*, volume 26, number 10, pp.

1,531–1,542.
doi: https://doi.org/10.1177/0956797615594620, accessed 13 November 2019.

A. Blankstein and M. Burke, 2019. "El Paso shooting: 20 people dead, 26 injured, suspect in custody, police say," *NBC News* (3 August), at https://www.nbcnews.com/news/us-news/active-shooter-near-el-paso-mall-police-responding-n1039001, accessed 13 November 2019.

S. Bradshaw, L. DeNardis, F.O. Hampson, E. Jardine, and M. Raymond, 2015. "The emergence of contention in global Internet governance," *Global Commission on Internet Governance, Paper Series*, number 17, at https://www.cigionline.org/publications/emergence-contention-global-internet-governance, accessed 13 November 2019.

J.S. Brown and P. Duguid, 2017. *The social life of information: Updated, with a new preface*. Boston, Mass.: Harvard Business Review Press.

CBS News, 2017. "Daily Stormer being dumped by GoDaddy" (14 August), at https://www.cbsnews.com/news/daily-stormer-being-dumped-by-godaddy-apparently-seized-by-anonymous/, accessed 13 November 2019.

D. Cho, 2013. "Real name verification law on the Internet: A poison or cure for privacy?" In: B. Schneier (editor). *Economics of Information Security and Privacy III*. New York: Springer, pp. 239–261.
doi: https://doi.org/10.1007/978-1-4614-1981-5_11, accessed 13 November 2019.

D. Cho, S. Kim, and A. Acquisti, 2012. "Empirical analysis of online anonymity and user behaviors: The impact of real name policy," *HICSS '12: Proceedings of the 2012 45th Hawaii International Conference on System Sciences*, pp. 3,041–3,050.
doi: https://doi.org/10.1109/HICSS.2012.241, accessed 13 November 2019.

N. Christin, 2013. "Traveling the silk road: A measurement analysis of a large anonymous online marketplace," *WWW '13: Proceedings of the 22nd International Conference on World Wide Web*, pp. 213–224.
doi: https://doi.org/10.1145/2488388.2488408, accessed 13 November 2019.

C. Cimpanu, 2018. "Popular Dark Web hosting provider got hacked, 6,500 sites down," *ZDNet* (17 November), at https://www.zdnet.com/article/popular-dark-web-hosting-provider-got-hacked-6500-sites-down/, accessed 13 November 2019.

R. Cohen-Almagor, 2011. "Fighting hate and bigotry on the Internet," *Policy & Internet*, volume 3, number 3, pp. 1–26.
doi: https://doi.org/10.2202/1944-2866.1059, accessed 13 November 2019.

E. Colleoni, A. Rozza, and A. Arvidsson, 2014. "Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data," *Journal of Communication*, volume 64, number 2, pp. 317–332.
doi: https://doi.org/10.1111/jcom.12084, accessed 13 November 2019.

J. Cook, S. Lewandowsky, and U.K.H. Ecker, 2017. "Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence," *PLoS ONE*, volume 12, number 5, e0175799.
doi: https://doi.org/10.1371/journal.pone.0175799, accessed 13 November 2019.

Council of Europe, 2003. "Declaration on freedom of communication on the Internet" (28 May), at https://www.osce.org/fom/31507, accessed 13 November 2019.

J. Cox, 2019. "8chan forced to move to obscure Dark Web service," *Vice* (6 August), at https://www.vice.com/en_us/article/wjwe34/8chan-forced-to-move-to-obscure-dark-web-service, accessed 13 November 2019.

A. Dalal, 2010. Protecting hyperlinks and preserving First Amendment values on the Internet, *University of Pennsylvania Journal of Constitutional Law*, volume 13, number 4, pp. 1,017–1,078, at https://scholarship.law.upenn.edu/jcl/vol13/iss4/4/, accessed 13 November 2019.

Daniel's Hosting, 2019. "Hosting — Info," at https://hosting.danwin1210.me/, accessed 13 November 2019.

R. Dawkins, 2016. *The selfish gene*. Fortieth anniversary edition. New York: Oxford University Press.

L. DeNardis, 2012. "Hidden levers of Internet control: An infrastructure-based theory of Internet governance," *Information, Communication & Society*, volume 15, number 5, pp. 720–738.
doi: https://doi.org/10.1080/1369118X.2012.659199, accessed 13 November 2019.

J. Dewey, 1923. *Democracy and education: An introduction to the philosophy of education*. New York: Macmillan.

F.L. Diaz, 2016. "Trolling & the First Amendment: Protecting Internet speech in the era of cyberbullies & Internet defamation," *Journal of Law, Technology & Policy*, volume 2016, pp. 135–159, and at http://illinoisjltp.com/journal/wp-content/uploads/2016/06/Diaz.pdf, accessed 13 November 2019.

R. Dingledine, N. Mathewson, and P. Syverson, 2004. "Tor: The second-generation onion router," *SSYM'04: Proceedings of the 13th Conference on USENIX Security Symposium*, volume 13, p. 21–21; version at https://svn-archive.torproject.org/svn/projects/design-paper/tor-design.pdf, accessed 13 November 2019.

M. Eslami, A. Rickman, K. Vaccaro, A. Aleyasen, A. Vuong, K. Karahalios, K. Hamilton, and C. Sandvig, 2015. "'I always assumed that I wasn't really that close to [her]': Reasoning about invisible algorithms in news feeds," *CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 153–162.
doi: https://doi.org/10.1145/2702123.2702556, accessed 13 November 2019.

M. Faizan and R.A. Khan, 2019. "Exploring and analyzing the dark Web: A new alchemy," *First Monday*, volume 24, number 5, at https://firstmonday.org/article/view/9473/7794, accessed 13 November 2019.
doi: https://doi.org/10.5210/fm.v24i5.9473, accessed 13 November 2019.

A. Feinberg, 2017. "This is the Daily Stormer's playbook," *Huffpost* (13 December), at https://www.huffpost.com/entry/daily-stormer-nazi-style-guide_n_5a2ece19e4b0ce3b344492f2, accessed 13 November 2019.

K.-w. Fu, C. Chan, and M. Chau, 2013. "Assessing censorship on microblogs in China: Discriminatory keyword analysis and the real-name registration policy," *IEEE Internet Computing*, volume 17, number 3, pp. 42–50.
doi: https://doi.org/10.1109/MIC.2013.28, accessed 13 November 2019.

R.W. Gehl, 2018. *Weaving the Dark Web: Legitimacy on Freenet, Tor, and I2P*. Cambridge, Mass.: MIT Press.

R.W. Gehl, 2016. "Power/freedom on the Dark Web: A digital ethnography of the Dark Web social network," *New Media & Society*, volume 18, number 7, pp. 1,219–1,235.
doi: https://doi.org/10.1177/1461444814554900, accessed 13 November 2019.

R.W. Gehl, 2011. "The archive and the processor: The internal logic of Web 2.0," *New Media & Society*, volume 13, number 8, pp. 1,228–1,244.
doi: https://doi.org/10.1177/1461444811401735, accessed 13 November 2019.

T. Gillespie, 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven, Conn.: Yale University Press.

R. Graham and B. Pitman, 2018. "Freedom in the wilderness: A study of a Darknet space," *Convergence* (18 October).
doi: https://doi.org/10.1177/1354856518806636, accessed 13 November 2019.

A. Greenberg, 2019. "Feds dismantled the Dark Web drug trade — But it's already rebuilding," *Wired* (9 May), at https://www.wired.com/story/dark-web-drug-takedowns-deepdotweb-rebound/, accessed 13 November 2019.

N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, 2019. "Fake news on Twitter during the 2016 U.S. presidential election," *Science*, volume 363, number 6425 (25 January), pp. 374–378.
doi: https://doi.org/10.1126/science.aau2706, accessed 13 November 2019.

F.O. Hampson and E. Jardine, 2016. *Look who's watching: Surveillance, treachery and trust online*. Waterloo, Ontario: CIGI/Centre for International Governance Innovation, and at https://www.cigionline.org/publications/look-whos-watching-surveillance-treachery-and-trust-online, accessed 13 November 2019.

A. Heft, E. Mayerhöffer, S. Reinhardt, and C. Knüpfer, 2019. "Beyond Breitbart: Comparing rightwing digital news infrastructures in six Western democracies," *Policy & Internet*.
doi: https://doi.org/10.1002/poi3.219, accessed 13 November 2019.

S. Hinduja, 2008. "Deindividuation and Internet software piracy," *CyberPsychology & Behavior*, volume 11, number 4, pp. 391–398.
doi: https://doi.org/10.1089/cpb.2007.0048, accessed 13 November 2019.

P.N. Howard, G. Bolsover, B. Kollanyi, S. Bradshaw, and L.-M. Neudert, 2017. "Junk news and bots during the U.S. election: What were Michigan voters sharing over Twitter?" *Oxford Project on Computational Propaganda, Data Memo*, 2017.1, at https://comprop.oii.ox.ac.uk/research/working-papers/junk-news-and-bots-during-the-u-s-election-what-were-michigan-voters-sharing-over-twitter/, accessed 13 November 2019.

Intelliagg, 2016. "Deeplight: Shining a light on the Dark Web," at https://media.scmagazine.com/documents/224/deeplight_(1)_55856.pdf, accessed 13 November 2019.

E. Jardine, 2019a. "Beware fake news: How influence operations challenge liberal democratic governments," *Centre for Internet Governance Innovation*, at https://www.cigionline.org/articles/beware-fake-news, accessed 13 November 2019.

E. Jardine, 2019b. "The trouble with (supply-side) counts: The potential and limitations of counting sites, vendors or products as a metric for threat trends on the Dark Web," *Intelligence and National Security*, volume 34, number 1, pp. 95–111.
doi: https://doi.org/10.1080/02684527.2018.1528752, accessed 13 November 2019.

E. Jardine, 2018a. "Privacy, censorship, data breaches and Internet freedom: The drivers of support and opposition to Dark Web technologies," *New Media & Society*, volume 20, number 8, pp. 2,824–2,843.
doi: https://doi.org/10.1177/1461444817733134, accessed 13 November 2019.

E. Jardine, 2018b. “Tor, what is it good for? Political repression and the use of online anonymity-granting technologies,” *New Media & Society*, volume 20, number 2, pp. 435–452.
doi: https://doi.org/10.1177/1461444816639976, accessed 13 November 2019.

E. Jardine, 2017. “‘Something is rotten in the state of Denmark:’ Why the Internets advertising business model is broken,” *First Monday*, volume 22, number 7, at https://firstmonday.org/article/view/7087/6328, accessed 13 November 2019.
doi: https://doi.org/10.5210/fm.v22i7.7087, accessed 13 November 2019.

E. Jardine, 2015. “The Dark Web dilemma: Tor, anonymity and online policing,” *Global Commission on Internet Governance, Paper Series*, number 21, at https://www.cigionline.org/publications/dark-web-dilemma-tor-anonymity-and-online-policing, accessed 13 November 2019.

A. Johnson, C. Wacek, R. Jansen, M. Sherr, and P. Syverson, 2013. “Users get routed: Traffic correlation on Tor by realistic adversaries,” *CCS ’13: Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, pp. 337–348.
doi: https://doi.org/10.1145/2508859.2516651, accessed 13 November 2019.

N.A. Johnson, R.B. Cooper, and W.W. Chin, 2009. “Anger and flaming in computer-mediated negotiation among strangers,” *Decision Support Systems*, volume 46, number 3, pp. 660–672.
doi: https://doi.org/10.1016/j.dss.2008.10.008, accessed 13 November 2019.

N.F. Johnson, R. Leahy, N. Johnson Restrepo, N. Velasquez, M. Zheng, P. Manrique, P. Devkota, and S. Wuchty, 2019. “Hidden resilience and adaptive dynamics of the global online hate ecology,” *Nature*, volume 573, number 7773 (12 September), pp. 261–265.
doi: https://doi.org/10.1038/s41586-019-1494-7, accessed 13 November 2019.

J. Kastrenakes, 2018. “German court says Facebook’s real name policy is illegal,” *Verge* (12 February), at https://www.theverge.com/2018/2/12/17005746/facebook-real-name-policy-illegal-german-court-rules, accessed 13 November 2019.

K. Knibbs, 2015. “Feds want Reddit to give up personal info of Darknet market Redditors,” *Gizmodo* (30 March), at https://gizmodo.com/feds-want-reddit-to-give-up-personal-info-of-darknet-ma-1694608548, accessed 13 November 2019.

T. Lavin, 2018. “The neo-Nazis of the Daily Stormer wander the digital wilderness,” *New Yorker* (7 January), at https://www.newyorker.com/tech/annals-of-technology/the-neo-nazis-of-the-daily-stormer-wander-the-digital-wilderness, accessed 13 November 2019.

D.M.J. Lazer, M.A. Baum, Y. Benkler, A.J. Berinsky, K.M. Greenhill, F. Menczer, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S.A. Sloman, C.R. Sunstein, E.A. Thorson, D.J. Watts, J.L. Zittrain, 2018. “The science of fake news,” *Science*, volume 359, number 6380 (9 March), pp. 1,094–1,096.
doi: https://doi.org/10.1126/science.aao2998, accessed 13 November 2019.

L. Lessig, 2006. *Code*. Version 2.0. New York: Basic Books.

A.M. Lindner and T. Xiao, 2018. “When the public seeks anonymity online: How news, ondustry, and socio-political conditions shape interest in the Tor anonymity network, 2006–2015,” *SocArXiv* (2 July), at https://osf.io/preprints/socarxiv/79mek, accessed 13 November 2019.

J. Ling, 2017. “Neo-nazi site The Daily Stormer moves to the Dark Web, but promises a comeback,” *Vice News* (15 August), at https://news.vice.com/en_us/article/gydmdj/neo-nazi-site-the-daily-stormer-moves-to-the-darkweb-but-promises-a-comeback, accessed 13 November 2019.

P.B. Lowry, J. Zhang, C. Wang, and M. Siponen, 2016. “Why do adults engage in cyberbullying on social media? An integration of online disinhibition and deindividuation effects with the social structure and social learning model,” *Information Systems Research*, volume 27, number 4, pp. 962–986.
doi: https://doi.org/10.1287/isre.2016.0671, accessed 13 November 2019.

J. Martin, 2014. “Lost on the *Silk Road*: Online drug distribution and the ‘cryptomarket’,” *Criminology & Criminal Justice*, volume 14, number 3, pp. 351–367.
doi: https://doi.org/10.1177/1748895813505234, accessed 13 November 2019.

D. Moore and T. Rid, 2016. “Cryptopolitik and the Darknet,” *Survival*, volume 58, number 1, pp. 7–38.
doi: https://doi.org/10.1080/00396338.2016.1142085, accessed 13 November 2019.

T.T. Nguyen, P.-M. Hui, F. Maxwell Harper, L. Terveen, and J.A. Konstan, 2014. “Exploring the filter bubble: The effect of using recommender systems on content diversity,” *WWW ’14: Proceedings of the 23rd International Conference on World Wide Web*, pp. 677–686.
doi: https://doi.org/10.1145/2566486.2568012, accessed 13 November 2019.

A. Oboler, 2012. “Aboriginal memes & online hate,” *Online Hate Prevention Institute, Report*, IR12–2, at http://ohpi.org.au/reports/IR12-2-Aboriginal-Memes.pdf, accessed 13 November 2019.

A. Oboler, K. Welsh, and L. Cruz, 2012. “The danger of big data: Social media as computational social science,” *First Monday*, volume 17, number 7, at https://firstmonday.org/article/view/3993/3269, accessed

13 November 2019.
doi: https://doi.org/10.5210/fm.v17i7.3993, accessed 13 November 2019.

E. Omernick and S.O. Sood, 2013. “The impact of anonymity in online communities,” *SOCIALCOM ’13: Proceedings of the 2013 International Conference on Social Computing*, pp. 526–535.
doi: https://doi.org/10.1109/SocialCom.2013.80, accessed 13 November 2019.

Online Hate Prevention Institute, 2019. “San Diego synagogue attack” (28 April), at https://ohpi.org.au/san-diego-synagogue-attack/, accessed 13 November 2019.

G. Owen and N. Savage, 2015. “The Tor Dark Net,” *Global Commission on Internet Governance, Paper Series*, number 20, at https://www.cigionline.org/publications/tor-dark-net, accessed 13 November 2019.

G. Owenson, S. Cortes, and A. Lewman, 2018. “The Darknet’s smaller than we thought: The life cycle of Tor hidden services,” *Digital Investigation*, volume 27, pp. 17–22.
doi: https://doi.org/10.1016/j.diin.2018.09.005, accessed 13 November 2019.

E. Pariser, 2011. *The filter bubble: What the Internet is hiding from you*. New York: Penguin Press.

W. Phillips, 2016. *This is why we can’t have nice things: Mapping the relationship between online trolling and mainstream culture*. Cambridge, Mass.: MIT Press.

Plato, 1974. *Plato’s Republic*. Translated by G.M.A. Grube. Indianapolis: Hackett Publishing.

E. Powers, 2017. “My news feed is filtered? Awareness of news personalization among college students,” *Digital Journalism*, volume 5, number 10, pp. 1,315–1,335.
doi: https://doi.org/10.1080/21670811.2017.1286943, accessed 13 November 2019.

M. Price, 2016. “The trouble with Tor” (30 March), at https://new.blog.cloudflare.com/the-trouble-with-tor/, accessed 13 November 2019.

E. Rader and R. Gray, 2015. “Understanding user beliefs about algorithmic curation in the Facebook news feed,” *CHI ’15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 173–182.
doi: https://doi.org/10.1145/2702123.2702174, accessed 13 November 2019.

M. Raymond and L. DeNardis, 2015. “Multistakeholderism: Anatomy of an inchoate global institution,” *International Theory*, volume 7, supplement 3, pp. 572–616.
doi: https://doi.org/10.1017/S1752971915000081, accessed 13 November 2019.

Reno v. American Civil Liberties Union, 1997. U.S. Supreme Court, docket number 96-511, at https://www.law.cornell.edu/supremecourt/text/521/844, accessed 13 November 2019.

P. Resnick, R Kelly Garrett, T. Kriplean, S.A. Munson, and N.J. Stroud, 2013. “Bursting your (filter) bubble: Strategies for promoting diverse exposure,” *CSCW ’13: Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion*, pp. 95–100.
doi: https://doi.org/10.1145/2441955.2441981, accessed 13 November 2019.

Reuters, 2017. “Google cancels neo-Nazi site registration soon after it was dumped by GoDaddy,” *CNBC* (14 August), at https://www.cnbc.com/2017/08/14/godaddy-boots-the-daily-stormer-because-of-what-it-wrote-about-charlottesville-victim.html, accessed 13 November 2019.

J. Roozenbeek and S. van der Linden, 2019. “The fake news game: Actively inoculating against the risk of misinformation,” *Journal of Risk Research*, volume 22, number 5, pp. 570–580.
doi: https://doi.org/10.1080/13669877.2018.1443491, accessed 13 November 2019.

K. Soska and N. Christin, 2015. “Measuring the longitudinal evolution of the online anonymous marketplace ecosystem,” *SEC’15: Proceedings of the 24th USENIX Conference on Security Symposium*, pp. 33–48, and at https://www.usenix.org/node/190887, accessed 13 November 2019.

J. Timmer, 2019. “Cloudflare has had enough, cutting off 8chan,” *Ars Technica* (4 August), at https://arstechnica.com/tech-policy/2019/08/cloudflare-has-had-enough-cutting-off-8chan/, accessed 13 November 2019.

Tor Project, 2019a. “Onion services,” at https://metrics.torproject.org/hidserv-dir-onions-seen.html, accessed 13 November 2019.

Tor Project, 2019b. “Tor: Overview,” at https://2019.www.torproject.org/about/overview.html.en, accessed 13 November 2019.

Tor Project, 2018. “Users,” at https://metrics.torproject.org/userstats-relay-country.html?start=2018-03-20&end=2018-03-22&country=all&events=points, accessed 13 November 2019.

Tor Project, 2015. “Some statistics about onions” (26 February), at https://blog.torproject.org/some-statistics-about-onions, accessed 13 November 2019.

S. van der Linden, A. Leiserowitz, S. Rosenthal, and E. Maibach, 2017. “Inoculating the public against misinformation about climate change,” *Global Challenges*, volume 1, number 2.

doi: https://doi.org/10.1002/gch2.201600008, accessed 13 November 2019.

E. Volokh, 2015. “The speech integral to criminal conduct exception,” *Cornell Law Review*, volume 101, number 4, pp. 981–1,052, and at http://scholarship.law.cornell.edu/clr/vol101/iss4/3, accessed 13 November 2019.

K. Wagner, 2019. “Facebook CEO: Company was too slow to respond to Pelosi deepfake,” *Bloomberg* (26 June), at https://www.bloomberg.com/news/articles/2019-06-26/facebook-ceo-company-was-too-slow-to-respond-to-pelosi-deepfake, accessed 13 November 2019.

H.G. Wells, 1927. *Complete short stories*. London: Benn.

M. Zabihimayvan, R. Sadeghi, D. Doran, and M. Allahyari, 2019. “A broad evaluation of the Tor English content ecosystem,” *arXiv* (18 February), at https://arxiv.org/abs/1902.06680, accessed 13 November 2019.

S. Zuboff, 2019. *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. New York: PublicAffairs.

Editorial history

Received 4 September 2019; revised 31 October 2019; revised 4 November 2019; accepted 4 November 2019.

Online content moderation and the Dark Web: Policy responses to radicalizing hate speech and malicious content on the Darknet
by Eric Jardine.
First Monday, Volume 24, Number 12 - 2 December 2019
https://firstmonday.org/ojs/index.php/fm/rt/printerFriendly/10266/8287
doi: http://dx.doi.org/10.5210/fm.v24i12.10266