Before the Ban

A decentralized system for online identity management and deplatforming insurance

Matt Asher[1]

(Version 1.0. October 1, 2018)

Social media users often assume they will never lose access to their usernames, connections, or content, but they have no contractual guarantee of this. People can, and do, lose access to their accounts for reasons related to terms of use violations or accidental deletion. More recently, social media platforms have been terminating the accounts of high profile users who post controversial content.

In this paper we present Before the Ban, a system designed to safeguard the identity of social media users, helping their connections and content persist even after a ban.[2]

DISCLAIMER: Nothing in this paper should be interpreted as a defense or endorsement of individual persons or their content.

NOTE: We begin this paper with a detailed look at the trends and forces that make a system like Before the Ban indispensable and inevitable. If you are willing to accept as a premise the need for such a service, and are uninterested in an analysis of how such a need arose, you can skip to the section titled "Introducing Before the Ban".

PUBLISHING GETS DECENTRALIZED

The long arc of publishing history shows a strong tendency towards decentralization. In the earliest days of recorded history, publishing was the exclusive right of rulers and gods (or their presumptive agents on earth). The flow of information went in one direction only. In the story of Moses, stone tablets were literally brought down from on high to the people. And, as a New Yorker cartoon observed, the tablets had no comments section.[3]

With the rise of the internet, the road to accessible, decentralized, and free publishing seemed to have reached its final destination. Anyone with a laptop, tablet or smartphone

---

[1] me@mattasher.com

[2] Several paragraphs in this paper were adapted from the EveryBit.js white paper, authored by Matt Asher and Dann Toliver. See https://github.com/EveryBit-com/everybit.js/tree/develop/whitepaper.

[3] Tom, T. (2013, July 29). Moses with ten commandments on two tablets [Cartoon]. New Yorker.

is now able to set up a blog or social media account for free. Within minutes users can begin posting content, with a potential audience of billions.

The internet used to be a highly decentralized publishing space. Bloggers published to their own domain names and readers aggregated the content with RSS, creating their own ad-hoc lists of people to follow. Email newsletters and mailing lists sent content to users who were largely free to opt-in or out as they pleased. In those days, online publishing was challenging for non-technical users, but the spaces they published to, and users' digital identities, were largely under their control.[4]

PUBLISHING GETS CENTRALIZED

Over time, the internet saw the rise of services that brought down the barriers to publishing. Users flocked en mass to platforms offering user friendly tools for publishing content, building an audience, and aggregating interesting news. As part of the bargain, however, users have given up control over their digital identities. These platforms have become gatekeepers and gateways, setting rules for what billions of people can publish, and filtering news feeds according to their hidden, and censorious, algorithms.

During the ascendancy of the now dominant social media platforms, a light touch was taken in asserting control over users and their content.[5] As recently as 2015, CEO Jack Dorsey declared that, "Twitter stands for freedom of expression".[6] Usernames were implied to be free and guaranteed for life, and revocations or involuntary transfers were rare. Even the loss of access to inactive accounts was rare – and controversial – enough to make news, and might lead to the return of the username.[7]

Users' presumption that they had a lifetime guarantee to usernames and unrestricted communication was key to the early adoption of the now-dominant platforms. It's well known that farmers who own their land work harder on it than those who are mere tenants or sharecroppers[8]; similarly, networks like Facebook and Twitter have a vested interest in getting users to believe that their space within those networks permanently belongs to them. The network's *de facto* protection of a user's right to continued, free use of their identity has encouraged users to spend hundreds or thousands of hours

---

[4] A notable exception here is email, which historically was tied to a workplace or a university.

[5] In 2012, the general manager of Twitter in the UK declared that the platform saw itself as "the free speech wing of the free speech party". (Halliday, Josh. 'Twitter's Tony Wang: "We are the free speech wing of the free speech party"'. *The Guardian* www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech (accessed September 26, 2018)).

[6] Dorsey, Jack. Twitter Post. October 5, 2015, 4:59AM. https://twitter.com/jack/status/651003891153108997.

[7] In 2014 Kathleen Hoff's inactive username was taken, then returned to her, by Instagram. (D'Onfro, Jillian. "Instagram Might Give Away Your Account Handle Without Telling You First If You're Inactive For Too Long" *Business Insider* https://www.businessinsider.com/instagram-give-away-inactive-accounts-2014-4 (accessed September 26, 2018).

[8] Burns, Christopher et al. "Farmland Values, Land Ownership, and Returns to Farmland, 2000-2016" *United States Department of Agriculture* 245 (February 2018).

developing content and building up followers, with little worry that they have no legal title to their identities within these platforms.

Today, the fiction that a user's social media identities belong to them as a matter of right has become increasingly difficult to maintain, and is now rarely asserted by the dominant platforms.[9] In the next section, we explore the rise of censorship and banning on social media platforms, provide some explanations for this rise, and discuss the history and implications of "reputational integration".[10]

LEGAL PROTECTIONS AGAINST DEPLATFORMING

In the United States, consumers take for granted that no matter where they go, and no matter who they are, they will be attended to at a store or restaurant. There are broad legal and social protections which enforce this presumption.[11]

While objectionable from the perspective of civil liberties,[12] one benefit of such anti-discrimination laws is that they provide social and economic cover for revenue-maximizing establishments to admit all customers, without having to defend their decision to let people into their store who belong to groups that others despise.[13]

These legal decisions were supported by broad cultural support for "dignity culture", at least in the North. In contrast to "honor culture", where stigma acts like a kind of communicable disease, in a dignity culture, the owner of a business suffers no loss of reputation by serving even socially marginalized customers. Likewise, a reputational barrier tends to protect consumers from the stigma of buying products from a vendor who is considered to be a bad person. A person's identity is considered an inviolable

---

[9] In 2017, an employee declared that, "Making Twitter a safer place is our primary focus". Clearly this assertion implies that users and content will be removed, as needed, to make others feel safe. (Ho, Ed. Twitter Post. January 30, 2017. 4:00 PM. https://twitter.com/mrdonut/status/826218619147165696).

[10] We are not overly concerned with the distinction between government and private censorship because we are not arguing that first amendment protections have been broken. Private censorship can be legal yet still problematic, and can certainly feel unethical to service users. Consumers have the right to complain about companies which stand against their interests and beliefs, and the right to vote with their feet by using alternatives, a topic we discuss in the section titled "Project Diaspora Kicks Off the Diaspora".

[11] The recent ruling in Masterpiece Cakeshop v. Colorado Civil Rights Commission , 584 U.S. (2018) might seem to go against this principle, but this case could be interpreted as a request for a service **not** offered, specifically for a kind of wedding cake not on the baker's menu.

[12] Anti-discrimination statutes limit the freedom of individuals to freely choose who they do business with, an apparent violation of Bill of Rights protections, but considered to be superseded by the Commerce Clause. See Heart of Atlanta Motel, Inc. v. United States, 379 U.S. 241 (1964).

[13] Imagine a fine dining restaurant in the Jim Crow era South that wishes to serve black customers. A law prohibiting racial discrimination reduces backlash from bigoted customers, who can no longer demand a whites-only dining experience, there or anywhere. On the flip side, laws restricting specific kinds of discrimination can lead to the presumption that anyone who falls outside of specific legal protections is fair game for exclusion. This parallels the founders' concern that by enumerating specific protected rights, they would create the perception that *only* those rights were worthy of protection. (See Best, James. *Principled Action: Lessons from the Origins of the American Republic*, pp 88-89. Tucson, Wheatsheaf, Inc, 2012).

(sovereign) part of their human dignity; shame is attached to the person's actions, not to their essential being, and certainly not to the people who treat them as worthy of shopping at their stores.

In honor cultures, reputational barriers bleed over from customer to vendor, and vice versa. Diminishment and outright revocation of identity are used as tools of social control.[14] Recently, the term "deplatforming" has been coined to mean the suspension or banning of a user from a social media service, but this concept is hardly new. Over the years, deplatforming has taken a number of forms, from excommunication and exile to sitting shiva for a person who isn't literally dead.

The results of deplatforming can be so negative that we have prohibited it among some identity providers. For example, monopoly communications providers are considered "common carriers". These companies have "a special duty to offer their services to anyone on just and reasonable terms without discrimination".[15] In exchange for giving up the right to discriminate, common carriers were given protection from being held responsible for the speech of their customers (in effect, the reputational barrier was made bilateral). Even more so than private companies, government identity providers are largely required to serve all clients: even felons are entitled to mail delivery, state IDs, and social security numbers.

BLURRING OF THE REPUTATIONAL BOUNDARIES

While there are still strong legal protections against identity discrimination within certain contexts, the social commitment to hold the reputations of consumers and vendors as independent has eroded significantly. A combination of cultural forces, recent events, and the particularities of internet publishing has contributed to this erosion, which has had a significant impact on the security of online identities, and the ability for users to publish freely on social media platforms.

This reputational blurring has been bilateral. For consumers, the willingness to buy from a company is now increasingly dependent on the company's reputation, not just the quality or cost of the product. The rise of "Green" and "Eco" businesses is a reaction to demands that companies present themselves as environmentally responsible. Increased partisan intolerance has induced a reputational bifurcation of many businesses into being perceived as "on the side of", or "against", one of the two main political "tribes". This bifurcation is sometimes encouraged by the companies themselves, by embracing their

---

[14] This revocation of identity is often used in the current judicial system, for example when prisoners are stripped of their names and referred to by number.

[15] "Common carriers, such as telephone and telegraph companies, differ from the printing press and broadcasters. Common law imposed on them a special duty to offer their services to anyone on just and reasonable terms without discrimination. The status of 'common carriers' was later recognized in telecommunication regulations. The law imposes liability on common carriers that discriminate content without cause. It requires that all customers be served. Common carriers traditionally enjoyed immunity for defamatory statements over their networks since they have no editorial control." (Lavi, Michal. "Content Providers' Secondary Liability." *Fordham Intellectual Property, Media and Law Journal* 26 (2016): 865).

support for certain causes,[16] or, even more dramatically, refusing to serve customers based on their political affiliations.[17]

Social support for these acts of discrimination is often found in the ascendant "social justice" (SJ) movement and culture. This movement strongly embraces aspects of the honor culture approach to reputation. The world is divided into classes of disadvantaged victims (who are to be treated as honorable, and in fact morally shielded by their status as victims), and privileged aggressors (who are stigmatized, along with anyone who sides with them). Perceived slights, even ones that might seem small, are derided as "not OK". Any statement considered to be racist or hateful towards disadvantaged groups immediately brands the speaker as morally tainted. Even the slightest hint of bigotry can be enough to ruin a person's reputation.

In this context, one of the things viewed as absolutely, positively, "not OK" is to provide a platform for these "deplorables" and "dregs of society".[18] Allowing these people to speak, or even worse, inviting them to speak (in effect affirming their identities as legitimate participants in ideological debates), is seen as so harmful to disadvantaged groups that censoring them is more important than protecting principles like free speech and open dialogue.

The cultural push of the current SJ ethos is distinguished from traditional honor culture in that the threshold for victimization is much lower, and does not require that the transgressions be intentional.[19]

---

[16] For example, fried-chicken chain Chick-Fil-A has publicly endorsed groups fighting same-sex marriage. (Mimms, Sarah. "Why Republicans Can't Stop Eating Chick-Fil-A." *The Atlantic*, March 13, 2015. https://www.theatlantic.com/politics/archive/2015/03/why-republicans-cant-stop-eating-chick-fil-a/449493/ (accessed September 26, 2018)).

[17] On June 23, 2018, White House Press Secretary Sarah Huckabee was asked to leave The Red Hen restaurant in Lexington, Virginia by the restaurant co-owner. The reason given for Huckabee's ejection was her connection to President Trump. (Vanden Brook, Tom. "Trump spokeswoman Sarah Huckabee Sanders kicked out of Virginia restaurant by owner." *USA Today*, June 23, 2018. https://www.usatoday.com/story/news/politics/2018/06/23/trump-spokeswoman-sarah-sanders-virginia-restaurant/727972002/ (accessed September 26 2018).

[18] In a speech on September 9, 2016, Democratic presidential candidate Hillary Clinton referred to many of the supporters of her opponent, Donald Trump, as the "basket of deplorables". Two years later, former Vice President Joe Biden described President Trump's allies as "virulent people"', going on to elaborate that "Some of them [are] the dregs of society"'. (Barry, Richard. "From Deplorables to Dregs of Society." *1420 WBSM*, September 18, 2018. http://wbsm.com/from-deplorables-to-dregs-of-society-opinion/ (accessed September 26, 2018)).

[19] Mary Spellman, dean of students at Claremont McKenna College in California, was forced to step down following an email in which she referred to non-white students who "don't fit our CMC mold". The wider context of the email was that it was written to reach out to a student who felt slighted for race-related reasons. Spellman was agreeing to a meeting to discuss the importance of these issues. (Mathis-Lilley, Ben. "Claremont McKenna Dean Resigns After Implying Nonwhite Students 'Don't Fit Our Mold'". *Slate*, November 13, 2015. https://slate.com/news-and-politics/2015/11/claremont-mckenna-dean-resigns-for-don-t-fit-our-mold-email.html (accessed September 26, 2018).

This new (or hybrid) form of honor culture[20] represents the final collapse of reputational barriers between partners in commercial, and sometimes human, transactions. Even in honor cultures, the damned can be given charity or allowed do some business (under diminished circumstances) without harming the reputation of the other person. For example, slaves were allowed to shop, but they had to wait at the back of the line, and those of even lower status had to enter through the back door, hidden from view of the other customers.[21]

In honor cultures, victimization leads to a diminishment of reputation. Restoring that reputation requires an evening of the score, in the form of violent revenge or an invitation to duel. But not every slight ends with bloodshed, in part because the possibility of escalation to this stage serves as a powerful incentive to find a less drastic resolution, or to brush off the slight as unintentional or unimportant. Note that duels expose both the perceived transgressor and the victim to risk. No such bilateral jeopardy exists (yet) in the new SJ culture, which in part explains how quick the aggrieved are to call out perceived aggressors. All reputational risk falls on the person (or company) being called out. For the accuser, the worst consequence of making unsubstantiated or weak accusations is to have them ignored. In this context, accusations of bigotry against fellow users of a communications platform proliferate, and the pressure is on companies to respond (or else see their own reputation suffer).

The nature of the dominant social media platforms

To some extent, social media platforms are uniquely vulnerable to reputational collapse between themselves and their customers/consumers. Platforms, especially the very few dominant ones like Facebook, exist somewhere in the gray zone between common carriers (neutral service providers) and publishers (who exercise editorial judgment over content). In congressional testimony, CEO Mark Zuckerberg called Facebook a "technology company" that builds products and services for other people.[22]

---

[20] Writing in The Atlantic, Conor Friedersdorf refers to the concept of "victimhood culture". Quoting the sociologists Bradley Campbell and Jason Manning, authors of a study into the concept of victimhood culture, Friedersdorf writes: "People increasingly demand help from others, and advertise their oppression as evidence that they deserve respect and assistance." (Friedersdorf, Conor. "The Rise of Victimhood Culture". *The Atlantic*, September 11, 2015. https://www.theatlantic.com/politics/archive/2015/09/the-rise-of-victimhood-culture/404794/ (accessed September 26, 2018)).

[21] This is, arguably, the compromise staked out at Reddit for the large, yet broadly despised, community known as `the_donald`. While generally allowed to express their views without restriction within the confines of this "sub-Reddit", posts in the `the_donald` are excluded from appearing on the home page of Reddit, regardless of how many upvotes they get (see https://mashable.com/2017/02/16/reddit-new-front-page-popular/#zRoTUg1mAEqc). The desire to keep `the_donald` users from leaving their ghetto and "infecting" the broader community is so strong that a popular browser plugin exists just to identify `the_donald` users who dare to participate in other sub-Reddits (see https://chrome.google.com/webstore/detail/reddit-masstagger/ebjdimopaogdkhiagbgmkjjhehmooheo).

[22] Castillo, Michelle. "Zuckerberg tells Congress Facebook is not a media company: 'I consider us to be a technology company'". *CNBC*, April 11, 2018.

This isn't untrue, but it's also the case that those services come with – and, by their nature, *must* come with – rules about who can post what and under which conditions, who is allowed to keep their identity within the system, and who will be expelled. They also come with algorithms to decide how newsfeeds should be ordered, which ads are allowed to be shown, and how those ads can be targeted. In other words, these technological products and services embed editorial decisions, which makes Facebook a publisher. And publishers shouldn't be giving a platform to repugnant or devious users, should they?

If Zuckerberg would prefer Facebook to be viewed as a (neutral) service provider, the company hasn't done itself any favors in recent years. To handle content moderation, it developed an extensive rule book of allowed and forbidden content. The rules may be arbitrary (e.g. breastfeeding pics of humans feeding humans can stay up, but not humans breastfeeding feeding animals, no matter what the context), but they were designed to be enforceable in a nearly automated way. More recently though, when the rules would have removed specific content that was deemed to be important or newsworthy, exceptions were made.[23]

Once Facebook began caving to pressure over politically sensitive content, it opened the doors to increased lobbying over content policy, with the assumption that any specific content allowed (or disallowed) must be the result of alignment (or nonalignment) with the views of those in power within the company. Facebook had torn down the reputational wall between themselves and the users they allowed on the platform.

This presumption of editorial influence has grown even stronger as platforms like Facebook appear to be banning people not for posting content that violates the rules (or, at least, not any more than what others post), but for holding views that fall outside the window of allowable opinion.[24] At this point, anyone not banned appears to have the tacit approval of the platform itself. In essence, now that "deplorables" are being deplatformed, anyone not deplatformed must be considered OK by the service provider. From an SJ perspective, until *all* the awful people have had their identities revoked, the platform itself is tainted. The purge must continue until purity is achieved.

Putting platforms into play

Imagine that Facebook and other social media channels are like board games, no different from Monopoly or The Game of Life. Every player has an agenda: to accumulate followers, to push a political viewpoint, to get likes, to pick on the people

---

https://www.cnbc.com/2018/04/11/mark-zuckerberg-facebook-is-a-technology-company-not-media-company.html (accessed September 26 2018).

[23] For a look at the evolution of Facebook's content rulebook, see the Radiolab podcast episode "Post No Evil". (Adler, Simon et al. "Post No Evil". *Radiolab*. Podcast audio, August 17, 2018. Retrieved from https://www.wnycstudios.org/story/post-no-evil).

[24] All societies have a limit to the beliefs that can be expressed without strong pushback. This limited range of acceptable discourse is sometimes called the "Overton Window". (See Robertson, Derek. "How an Obscure Conservative Theory Became the Trump Era's Go-to Nerd Phrase". *Politico Magazine*, February 25, 2018. https://www.politico.com/magazine/story/2018/02/25/overton-window-explained-definition-meaning-217010 (accessed September 28, 2018)).

they hate. Each version of this game has rule differences and its own culture. Facebook encourages the most active players of their game by giving them more exposure for their pages and posts. Twitter encourages combative dialogue that makes users feel compelled to monitor their feed and respond to attacks.

As with all games, the direct approach to winning is to out-compete the opposition. The indirect way is to bribe the refs or have the rules changed to favor you or your team. If a platform signals that it isn't just a neutral arbiter of rules written in stone, the platform itself is put in play. This shifts activity from on the field, to off the field. The game is now political, and participants begin to focus on lobbying the refs to have their opponents disqualified or in some way diminished. In the case of social media, this diminishment can take the form of shadow bans,[25] revocation of verified user checkmarks, or removal from lists of suggested people to follow.

In the case of Twitter, once it put itself into play by caving to political forces to ban certain high profile users, CEO Jack Dorsey went from being a ref who refuses to remove a despised player like Alex Jones (no matter how much the crowd boos him), to a ref who invites only those players he personally likes to join the game. Once this shift in perception happened, the reputational damage to Dorsey became too strong to bear, and a reason was found to give Jones a red card and ban him from the field[26].

In this paper we take no position on whether the transition from neutral arbiters to malleable refs is good for Twitter or Facebook as *companies*. We simply note that the process is well underway, that it will tend to feed on itself, and will be extremely hard to reverse. We also note that this trend has strong consequences for the users of the system. If users could previously assume (however wrongly) that usernames were as strong as land titles and platforms were neutral forums for communication, those assumptions now need to be examined closely.

PROJECT DIASPORA KICKS OFF THE DIASPORA

For reasons related to banning, censorship, political alignment, or concerns about control over their newsfeeds, more and more users are looking for alternatives to the dominant platforms. A number of alternatives have gained a significant level of traction, starting with Project Diaspora in 2010. Social network Ello was launched in 2014 with the motto "The anti-Facebook", and more recently a number of federated and decentralized services have been founded, most notably Mastodon (a Twitter competitor that allows anyone to create an "instance" and join the system), Steemit (a blockchain-based online forum that rewards contributors with digital currency), and Minds.com (a decentralized "crypto social network").

As more and more users join alternative social media platforms, publishing once again moves in the direction of decentralization. It also becomes more fragmented and

---

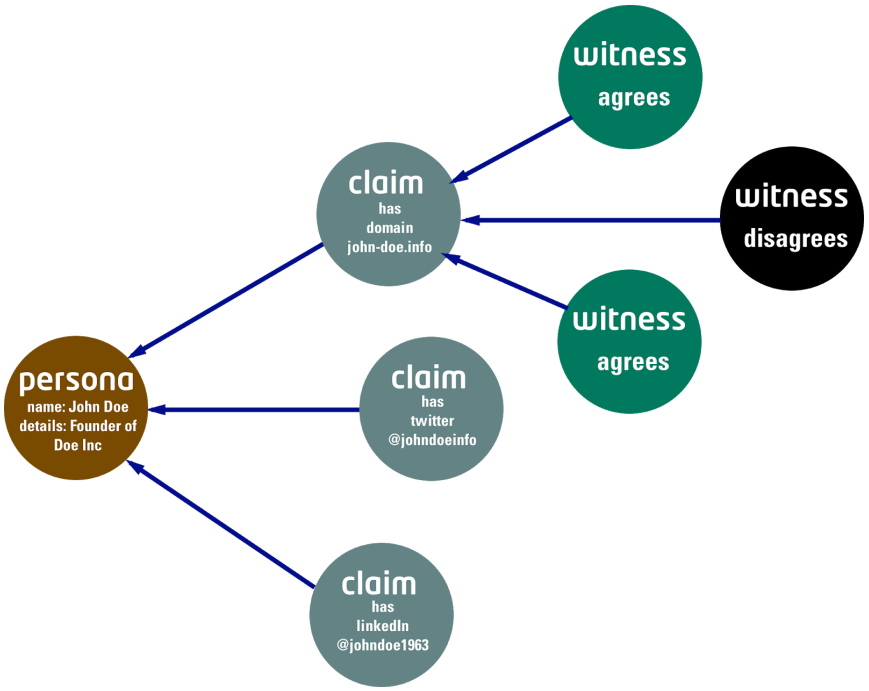
[25] In a shadow ban, access to the platform isn't restricted, but the user's content is hidden from appearing in search results and newsfeeds.

[26] "Twitter bans Alex Jones and Infowars for abusive behaviour". *BBC*, September 6, 2018. https://www.bbc.com/news/world-us-canada-45442417 (accessed October 1, 2018)

chaotic. Keeping track of the hundreds or thousands of people you wish to follow and maintaining access to their current and past content, using dozens of social media platforms, will become ever more difficult.

In this context, we believe the time is ripe for a system that protects users' identities and allows them to preserve their content and followers as much as possible, no matter which platforms they join or are excluded from.

INTRODUCING BEFORE THE BAN



Before the Ban (BtB) is a decentralized, cross-platform index of users and where to find their content. The BtB system consists of three main parts: Personas, Claims, and Witnesses.

Personas are named identities. A persona could be related to a real, named human (e.g. James Roberts), a pseudonym (e.g. Satoshi Nakamoto), or an organization (e.g. WikiLeaks). All BtB actions are tied back to individual personas. Anyone can create a persona, either for themselves or for someone else.

Claims are made by personas, on behalf of the same persona or another one. Claims assert ownership (or control) over some part of the persona's online identity. A typical claim might be that the persona with the name James Roberts has the Twitter handle `@jamesroberts78`, or that this persona runs the blog `JamesRobertsThoughts.com`.

Claims can be backed up with evidence, which takes the form of a URL or the hash of a document. For example, to show that the persona with the name James Roberts

controls `@jamesroberts78`, a post could be made by that Twitter account referencing the James Roberts persona on BtB.

Witnesses are personas acting in judgment of a claim. Witnesses help validate (or invalidate) claims by moderating them; in effect, they vote up or down a claim's legitimacy.

Each of these three parts has a corresponding document (or **record**) that takes the form of a cryptographically signed transaction. These transactions reference each other and are linked (specifically, they form a DAG[27]). The interpretation of this DAG in terms of claimed information is straightforward: anyone looking at the records can see which social media accounts are claimed to belong to which personas, and which personas agree or disagree with those claims. For technical details about these three types of records, and how they are linked together, see Appendix B: The Structure of Reports.

All the persona, claim, and witness records are open and available for anyone to inspect or analyze. See Privacy Concerns below for a discussion of the privacy implication of BtB.

Witnesses and the web of trust

BtB is an open, decentralized system. Anyone can use it, and no central authority validates information submitted to the system. Instead, by witnessing claims, a web of trust is formed with each user (persona) at its center.

As a BtB user, the trust I put in a witness is related to how often I agree with the validity of their claims. In other words, if witnessing claims is like moderation, the more my votes agree with another moderator, the more I trust their votes on claims that I haven't witnessed myself.

This system encourages users to witness other claims, and to do so accurately, as the quality of a user's web of trust rankings grows along with the number and quality of the other witnesses they agree with. If a user's activity as a witness is aligned with witnesses who tend to be accurate, then that user's web of trust will weight accurate witnesses more heavily.

As a reference implementation for web of trust rankings, BtB builds on the EiganTrust algorithm.[28] In this model, trust is extended out transitively from the users you trust to the users trusted by those users. While each participant (in our case, persona) has their own view of the trustworthiness of the other participants, those views can be shown to converge under certain conditions, resulting in an absolute score for each participant.

Before the Ban records are meant to replicate our real world experience with identity and trustworthniess. We judge and evaluate others directly, and we also rely on a variety of services and personal networks to help us judge and assign authority (or trustworthiness) to people and institutions.

---

[27] DAG stands for Directed Acyclic Graph. In this mathematical structure, nodes point to each other, and the flow of pointers never completes a cycle. Each node can have multiple parents and multiple children.

[28] Kamvar, Sepandar D., et al. "The EigenTrust Algorithm for Reputation Management in P2P Networks". *Stanford*, May 2003.

One algorithm is not enough

While the developers of BtB strive to make this trust algorithm as strong as possible, it should be stressed that this is only a *reference implementation*. The trust algorithm is designed to be a **modular** part of BtB. It can be tweaked or swapped out for an entirely different approach to calculating trust.

All systems can be thought of like a game. As such, BtB is by nature collaborative and at times competitive. The main "goal" of BtB at large is to create a directory of all public personas and the outlets they claim as their own. The formal rules of this game are highly limited: they specify the required format and cryptographic signing for several types of records (persona, claims, witnesses), and rules for valid chains of records. Beyond that, the interpretation of the records, and the implied trust of each claim, is left open.

We view this modularity of the trust algorithm as a core strength of BtB. Multiple algorithms will make it harder to game the system, and the open source nature of the trust ranking module makes it easy to update or fork in response to a coordinated attack on the system. Given a diverse ecosystem of trust algorithms, it would be impossible to game all of them at once, as any strategy targeted at certain algorithms would necessarily be poorly optimized for others. In effect, gaming the system becomes a highly complex effort to maximize some function in a multidimensional field of gradient vectors. And this is in the ideal case, where the gamer has complete control over all claims and witnesses.

Modularity should also make it clear that BtB itself is not passing judgment on which claims are the most credible. There is no single most (or least) verified identity, where an identity is comprised of a combination of persona and claims. Each algorithm creates its own partial (or total) order on the reliability of claims.

We believe this shifts the focus from winning by maximizing a universal metric, to winning by making sure your persona record, and the records of the people you follow, all have claims on them you know to be valid, and each of those claims is properly witnessed.

A solution for content

By itself, BtB doesn't solve the problem of persisting user content after a user has lost access to their archive of posts. However, by allowing for decentralized (and self-service) control over the pointers to content, BtB makes it possible to decouple content and platform. Items of content can be posted anywhere, and a claim can be made that links this content repository with that persona. For example, a copy of a user's posts could be published to IPFS[29]. Each post could be signed by a cryptographic key associated with the related persona record, and these posts could be linked together with IPNS[30] as the decentralized "domain name" where the posts occur. Alternatively, people could use a

[29] InterPlanetary File System, see https://ipfs.io.

[30] InterPlanetary Name System. (See https://medium.com/coinmonks/how-to-add-site-to-ipfs-and-ipns-f121b4cfc8ee).

centralized archiving service to host copies of the posts. Either way, as long as a BtB claim links the persona and the location of their current content archive, access to the content is preserved.

RSS all the things

Given that each social media platform has a different format for their posts (and allows different media types), we recommend that all items be indexed with a common format. Most blogging software already does this, using RSS. Ideally, every user of BtB would have an RSS feed for each of the platforms they publish to, either because that platform supports RSS directly, or because a tool exists to convert their feed into RSS format. Some tools already exist to do this.[31] We are working to extend these tools and develop new ones so that all of the major (centralized and decentralized) platforms are covered.

Your feed, the way you want it

In a world where all public social media posts are accessible through RSS, regardless of which platform they were published to originally, many new possibilities open up for aggregated newsfeeds. By decoupling content from platform, users can create personalized feeds that cross platforms, and sort them in any order they wish, including by most recent, without taking into account popularity.[32] Before the Ban opens up the possibility of completely portable, cross-platform, ad-hoc social networks defined by a single configuration file of options (e.g., follow James Roberts on Twitter, Jane Reynolds on Instagram, use the Probabilistic Trust Inference v1.3 to determine the top rated claim about where these feeds are located, and sort the resulting items by oldest first).

WEAKNESSES AND LIMITATIONS TO BTB

Every system has limitations and weaknesses. In this section we consider limitations related to incentivizing storage and the lack of a username system. We discuss key loss, privacy issues, and potential attacks on the system and ways to mitigate them.

Incentivizing storage

BtB has baked-in incentives for users to store and make available the core record files (personas, claims, and witnesses). Finding and sharing these files is key to locating user content and calculating web of trust rankings for witnesses, which will keep these records in circulation. The number of (valid) records should grow roughly linearly with the number of personas in the system, as each user of the system will be associated with a

---

[31] See https://rsshub.app/

[32] After Instagram changed their feed to no longer respect strict chronological order, many users were upset. (See Kraft, Amy. “Backlash continues over Instagram’s new algorithm”. *CBS News*, March 28, 2016.
https://www.cbsnews.com/news/backlash-continues-over-instagrams-new-algorithm/ (accessed October 1, 2016).

limited number of personas, claims, and witness records. With the proper spam protection, pruning, and in-browser sharing using WebRTC, all vital core records should be preservable (see "Scammers and gamers" below).

Individual content items are a different story. The core goal of BtB is not to provide a content repository, but to help users find the current location for a user's content. BtB will provide users with tools to convert individual posts into (immutable, cryptographically signed) flat files which can be stored in decentralized networks such as IPFS, or centralized hosting solutions like Dropbox or a CDN. However, persistence of these items at these services is not guaranteed.

As banning events, content removal, and mobility between social networks increases, we expect that more and more content mirrors like archive.is and tweetsave.com, and personal backup tools like digi.me will be leaned on to prevent individual pieces of content from disappearing into the memory hole. The challenge will be content discovery, and maintaining an up-to-date aggregated feed of all of the people you want to follow, across whatever platforms they are currently using.

No username system

In terms of scope, BtB does not attempt to solve the problem of creating a universal username system. It may seem like a small leap from providing everyone with a decentralized identity to letting people choose usernames for those identities, but in fact this additional requirement would have imposed significant additional burdens.

Some of the trickiest problems with username systems stem from the need to prioritize ownership claims. At a minimum, BtB would have required a consensus mechanism to guarantee a (total) order for registration claims. In Bitcoin, arriving at consensus over the timing of transactions represents almost all of the very large cost of maintaining the network.[33] As a (partially ordered) DAG of transactions, BtB can only make guarantees about the relative timing of linked records. However, as the trust algorithm is not concerned with finding a single "true" James Roberts persona, the inability to determine the first record with his name doesn't harm the system.

In addition to timing issues, username systems have issues related to theft, monopolistic pricing, adjudication of competing claims, and resolving trademark issues.

As a result of these challenges, every attempt so far to create a universal system has created yet another username silo, each with its own significant drawbacks. There is long history of such attempts, every one ending in failure[34] or at best short term windfalls for their creators.[35]

---

[33] Hardware and electricity costs for "mining" Bitcoin are in the billions of dollars per year. See https://www.marketwatch.com/story/in-one-chart-heres-how-much-it-costs-to-mine-bitcoin-in-your-state-2017-12-15.

[34] This was learned the hard way by one of the authors of this paper, who championed a universal username system as part of the EveryBit.js project.

[35] AOL's keywords was an early attempt at a universal system of owned tags (effectively, usernames). Others of note have included Namecoin (complicated to use and hard to integrate with other system) and Ethereum's ENS (centralized, and vulnerable to censorship, blackmail, and hostile takeover forking).

The closest thing we have to universal usernames are domain names. In many ways the domain name system works quite well, but it's hardly cheap, and as a federated system with centralized dispute resolution, it's vulnerable to censorship,[36] reverse domain hijacking,[37] and outright theft.[38]

With BtB, implementing a username system would have also distracted from the true object of value: digital identity. We define digital identity as the link between someone (or some organization) and their content. Generally, usernames at individual platforms are the gateway for finding someone's content, and they may even be brands unto themselves. Before the Ban can't protect someone's specific Twitter handle, but it can ensure that the people who followed `@jamesroberts78` can continue to follow James Roberts if he leaves or is forced off the platform. James Roberts' (digital) identity, as defined by his content and his right to communicate with followers, is preserved.

Key loss and recovery

Instead of usernames and logins, BtB uses private and public keys (addresses). All records (personas, claims, and witnesses) are digitally signed, and those signatures are validated to make sure they correspond to the indicated persona. The benefit of this system is that no central authority has access to a user's passwords. All signing happens client-side (e.g. in the user's web browser), so no passwords are sent over the network. No network administrator has the power to impersonate their clients.

One downside of this key-based system is that there is no built-in way to recover a lost key. This is a known problem in the space; to solve it, BtB will be leaning on the many solutions being developed for other projects, including backup systems, hardware wallets, and using several trusted parties to store parts of the private key.[39]

We emphasize that unlike with cryptocurrency and other key-based systems, loss of access to a private key does not mean a loss of access to BtB, nor does it mean a person's established decentralized identity will go away. All of the claims made about an identity (as well as all the witnessing of those claims) remains. What is lost is the ability for that persona to make new claims (or provide witnessing) using the reputation built up for that address. This reputation will have to be created anew, and each of the trust algorithms will view this persona as a brand new user until they have established a history at this new address.

---

[36] The domain name goatse.cx, which once hosted a notoriously explicit image, was revoked by its registrar in 2014 (https://en.wikipedia.org/wiki/Goatse.cx#Domain_suspension_and_sale_of_domain_name). The US government has seized domain names on allegations of hosting illegal content (see https://thehill.com/policy/technology/130763-homeland-security-dept-seizes-domain-names-).

[37] In reverse domain hijacking, a company tries to abuse the domain dispute resolution process (UDRP) to take away a domain from its current, legitimate, owner.

[38] Most famously, in the case of the domain name sex.com.

[39] Before the Ban uses a cryptographic scheme similar to the one used by Bitcoin, so any key management solution that works for Bitcoin should work for Before the Ban.

Privacy concerns

As a public system, BtB does not directly address the issue of privacy. All of the records related to personas, claims, and witnesses are public. The presumption is that if someone (or some organization) publishes content to a social network or URL, and they make that content public, then helping users find the current location where that content is being published is a positive service, as much so as any search engine.

If a user controls a social media account and they wish to prevent it from being associated with them as a person, BtB by itself will neither protect nor expose their true identity. Right now, anyone who has the information to "dox" (out) a pseudonymous user can do so by spreading the information online. BtB is just one more place a doxer could publish their claims.

There may be some privacy benefits that arise from BtB in the long term. For example, the system could be used as a resource for public key lookups, or finding out how to contact someone using a secure messaging service. This could facilitate the exchange of more messages in an encrypted way.[40] However, it should be noted that before sending sensitive data based on BtB claims related to public keys, it would be wise for users to independently verify the claims.

Spammers and gamers

As an open system with no restrictions on who can use it or how many records can be created, BtB is vulnerable to various forms of abuse, including spamming, gaming, and malicious attacks.

There are approximately 5 billion internet users. If each one generates 50 BtB records (assuming an average of two personas, 10 claims, and several dozen witness records), and if each record occupies 1KB of space, the total storage for all records would cap out at about 250 terabytes of uncompressed data. If a single entity wanted to cache all of the data at once, the cost would be about USD $5,000,[41] a relatively small amount to pay for a directory of everyone's public social media usernames. If the information were split up among users of the system and shared using WebRTC, each browser would only need to store about a megabyte to achieve a high level of redundancy.

Clearly, the challenge isn't storing legitimate records, it's recognizing the bad ones and letting them fall away.

As a first pass, we can remove any records that are poorly signed. If the signature doesn't match the claimed `createdBy` address, we throw it out immediately. We also require that all claims and witnesses be made by someone who has created a persona. To do otherwise would be the equivalent of allowing anonymous users to moderate content on a web forum. Even though there may be no restriction on signing up for an account, the additional friction provides some barrier, and the ability to track back all actions to a persona lets the web of trust algorithm detect attempts to game the system (for example, a

---

[40] At present, the website keybase.io provides this service in a centralized way, with its own username system.

[41] See https://www.backblaze.com/blog/hard-drive-cost-per-gigabyte/

ring of personas who all validate each other's claims, but no one outside the ring is agreeing with their claims).

We also remove redundant information. Duplicate claims are discarded, keeping the earlier one and discarding the new one. This prevents users from gaming the system by introducing a new claim to wipe away old witnesses. Likewise, there's no reason to keep more than one witness record by the same persona of the same claim. In this case, though, we keep the most recent record, as a user might wish to update their witness record if they made a mistake, or a claim that was once valid is no longer true (e.g. a persona's domain name expires or is sold).

The timestamp field `createdAt` provides another opportunity to prune bad records. Before the Ban will refuse to store any record that comes in with a timestamp significantly later than the present moment (allowing for slight clock inaccuracies). Except for the persona records, all the other records reference another, existing, record. We make sure that the referencing timestamp is later than the record it references, and throw it out if not. We also discard any records that refer to non-existent records (effectively, bad links).

After applying all the formal methods of detecting bad data, we can set a threshold based on our web of trust algorithm and its rules. For example, we might prune any persona with no related claims after a certain period of time, say one year. Pruning of unwitnessed claims could also be done.

Note that because the system explicitly invites multiple web of trust algorithms, different users will have different sets of records that meet the threshold for saving. The presumption is that with a diverse, high quality pool of trust algorithms, it will be unlikely that quality records disappear, while at the same time difficult for bad actors to flood the system with invalid ones.

Should all these measures fail to limit abuse of the system to a sustainable level, a "proof of burn" requirement could be instituted. Under such a scheme, each transaction to add a persona or claim would have to be submitted with proof that the user who created it destroyed some amount of cryptocurrency on an existing blockchain. To destroy (or "burn") the currency, users would send it to a non-existent address, along with a memo linking the transaction to the BtB record they wish to add.

WHY EVERYONE SHOULD CARE

It might not be apparent why social media users who see their content as uncontroversial would wish to use BtB. Beyond using it to find content from other users who *are* controversial, why should they care?

We present here an argument that even users who see their content as completely inoffensive might wish to insure their identity with BtB.

For starters, the content users see as ordinary might still get flagged by a filtering algorithm. For example, in 2016, author and artist Dennis Cooper found that his blog had

been deactivated by Google for hosting explicit artwork (nude paintings, among other content). The ban resulted in the loss of 14 years of work, including an unfinished book.[42]

Even if the things you post right now are uncontroversial enough to avoid upsetting humans or machines, they may someday be taken as a sign that you are an undesirable person to have on a social media platform (see Appendix A: The Myth of Inoffensiveness). Should your content remain uncontroversial to most future users, it may still become a problem for the platform you use if the people who disagree with you are particularly intolerant.[43]

No matter how good the algorithm being used to find rogue users, there is no guarantee that the algorithm itself won't go rogue, and start banning people for unclear or mistaken reasons. For example, algorithms have a hard time distinguishing jokes, sarcasm and meta-commentary from directly offensive content.[44]

Finally, a user might already be banned and not know it. Under a "shadow ban," everything about a service would appear to work fine for someone, but their content is hidden from others in search results or other locations.

The importance of identity

In the digital age, identity is the linchpin of freedom.

Our digital identities are the artifacts that allow us to be found online, and to communicate with others. They are the usernames with which we post, the domain names people use to find our content, the email addresses with which we send messages.

Identity is what allows us to build up a following and stake out a claim to digital content. It is our title to the virtual plot of land we occupy, and the directory that lets others find the structures we build on it. To the extent that our digital identities can be hidden, revoked, or censored, our freedom to express ourselves online is at risk, as is the freedom of others to access or content.

ACKNOWLEDGEMENTS

Both David Harrison and Anne Gruel contributed valuable edits to this paper. I'd also like to thank the team at Crypto Media Hub for their feedback.

---

[42] Sidahmed, Mazin. "Dennis Cooper fears censorship as Google erases blog without warning". *The Guardian*, July 14, 2016. https://www.theguardian.com/books/2016/jul/14/dennis-cooper-google-censorship-dc-blog (accessed September 26, 2018)

[43] In his book Skin in the Game, Nassim Nicholas Taleb explains how the most intolerant group, even if very small, can often get the majority to accommodate their beliefs, to the detriment of the more tolerant majority. (Taleb, Nassim Nicholas. *Skin in the Game: Hidden Asymmetries in Daily Life*. New York, Random House, 2018).

[44] In order to highlight media bias, conservative commentator Candice Owens replaced the word "white" with "black" (or "Jewish") in some tweets from recently named New York Times editorial board member Sarah Jeong. Owens was promptly banned from Twitter, even though she was trying to highlight the hatefulness of the original posts, not actually endorsing the message her posts contained. Owens had her account quickly restored, but only because Twitter was flooded with complaints from her large and vocal fan base.

APPENDIX A: THE MYTH OF INOFFENSIVENESS

Given a diverse enough population, your beliefs are almost certain to already be offensive to some sub group. As HackerNews user nostrademons put it, “One interesting consequence of the Internet is that we’re becoming very aware that for every label you could ascribe to yourself, there is some group out there who holds a deep, visceral hate for that label, so deep that they wish you would just cease to exist.”[45]

If you've posted, and wish to continue to post, about any about the following topics:

- Meat eating
- Abortion
- Capital punishment
- Circumcision
- Zoos
- Any religion
- Politics

you may find yourself subject to a ban, especially if the people who disagree with you have political power within the network or are particularly loud and intolerant.

APPENDIX B: STRUCTURE OF THE REPORTS

Before the Ban records, or “transactions”, are JSON files. They have the following structures:

```
// Personas
persona: {
    ...
    address: "", // Address of record owner
    name: "John Smith",
    description: "Founder of the Arizona Widget Company",
}

// Claims
claim: {
    ...
    accountType: "twitter",
    accountURI: "@JohnSmithWidgetGuy",
    evidence: "https://twitter.com/statuses/salfd", // URL or hash or
document that demonstrates control over this account
    personaSig: "", // Signature of the related persona record
}
```

[45] See https://news.ycombinator.com/item?id=16609847

```
// Witnesses
witness: {
    ...
    witness: true, // Does this witness agree with the claim?
    claimSig: "", // Signature of the related claim
}
```

All records contain the following fields:

`version`: The version of the spec being used

`createdBy`: The public address (public key) of record creator

`createdAt`: The time of record creation

`sig`: A cryptographic signature of the record, validated against the `createdBy` address