Security through Inefficiency

Leveraging Social Friction to Combat Online Manipulation

Joel Miller
University of Illinois at Chicago
jmill54@uic.edu

Chris Kanich
University of Illinois at Chicago
ckanich@uic.edu

ABSTRACT

The economies of scale of the modern Internet have enabled a dizzying array of delightful products delivered at infinitesimal marginal cost. At the same time, bots and disinformation plague social networks and have significantly altered political discourse in America. In this paper we propose the paradigm of *security through inefficiency* as an approach to understanding and combatting the latter while preserving the former. We compare online communities with real-world communities and argue that online communities are susceptible to manipulation partially due to their lack of *social friction*, the soft socio-technical boundaries that underpin communication within non-online communities. We examine how social friction can provide security for communities and argue that while efficient scaling is the dogma of modern software engineering, it is possible that the judicious application of social friction can increase the overall functioning of these systems, especially when it comes to resistance to online manipulation. We then analyze two social networks, Mastodon and Nextdoor, and theorize that some of their structural elements can elucidate the benefits of introducing social friction to online spaces. Fundamentally, we seek to question the longstanding dominance of technical efficiency in system design, and wish to prioritize socio-technical outcomes that are contrary to such scaling as a means to combat online disinformation and manipulation.

CCS CONCEPTS

• **Information systems** → *Social networks*; • **Security and privacy** → **Human and societal aspects of security and privacy**.

KEYWORDS

Privacy; Security; Social Networks; Sybil Attacks; Political Discourse; Propaganda

ACM Reference Format:
Joel Miller and Chris Kanich. 2020. Security through Inefficiency: Leveraging Social Friction to Combat Online Manipulation. In *Proceedings of New Security Paradigms Workshop 2020 (NSPW 2020).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

NSPW 2020, 26-29 October 2020, North Conway, NH

ACM ISBN 978-1-4503-XXXX-X/18/06…$0.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

From a societal perspective, the information & communication technologies (ICT) of the contemporary era have brought about incredibly substantial change in the latency, frequency, and audience sizes of interpersonal communication: with suitable connectivity and hardware, high-definition intercontinental video chat is possible with latency at a nontrivial fraction of the speed of light, and instantaneous broadcast is possible to millions or billions of potential audience members.

Perhaps even more of an upheaval however is the economic impact of ICT: once the physical infrastructure is deployed, the marginal cost of delivering a computation-based service is infinitesimal, and thus companies like Google and Facebook can operate at truly global economies of scale. The raw benefits of these technologies is clear: software products like Facebook or Google Search easily have development costs of several billion dollars, and are available without monetary cost to billions of users - in no small part due to how little additional cost is associated with providing the service to an additional user. Beyond this benefit, security also benefits from these economies of scale: these companies (and their products, and their users) not only benefit from having the organizational capacity to devote significant resources to security (exemplified by Google's Project Zero,[1] which devotes substantial resources to fixing security bugs in other companies' products for free), but they also benefit from the immense visibility afforded to companies that operate ICT-based services at global scale, being able to record every request made to every piece of software within their datacenters.

It would be outlandish to argue that these economies of scale are anything besides a net positive for the people who can take advantage of them. However, they are not without their downsides. Most obvious is the shift toward a more centralized mass media and communications infrastructure: in the modern advertising-driven information economy, control over things like local media have arguably become more centralized than they have been since the invention of the printing press. Additionally, while disinformation and information warfare have existed for quite some time, the manipulation happening in contemporary online discourse is wreaking significant havoc for society's ability to effectively coordinate and regulate some semblance of a shared truth.

1.1 Online disinformation

Academically grounded concerns about online media's use as a tool for manipulation date at least as early as 1988 [23], and concerns about *social* media's use for the same purposes date at least as early as 2006 [24]. Indeed, although social media has been used effectively

[1] https://googleprojectzero.blogspot.com/

to empower social movements and stimulate democratic conversation [14, 22, 59], much recent popular discourse has focused on the threats that bots, propaganda, and unauthorized data collection pose to healthy discussion on social media networks, especially discussions about politics [36, 37, 53, 54].

Throughout the 2010s, research was conducted on the dangers of manipulative social media practices. Facebook was found to be vulnerable to large-scale infiltration by botnets [8], Twitter influence metrics were shown to be far from robust to manipulation [33], and researchers started to study political polarization and disinformation campaigns on Twitter [15, 47]. Meanwhile, Kramer *et al* showed that emotional states spread within social networks [29], which in turn suggests that botnets which gain access to large swaths of a network can influence the emotions of real users through the spread of inflammatory content.

In American public discourse, much attention has been given to the role of bots, propaganda, foreign influence, and unauthorized data collection in the 2016 presidential election [44, 53, 54]. Between election day and January 2018, Twitter identified 50,258 bot accounts which both tweeted political content during election season and were linked to foreign actors [43]. In the same vein, an investigation by United States special counsel Robert Mueller found that foreign actors sought to influence the outcome of the election, partially through various social media campaigns [34], and Bessi *et al* used the bot-detection program *BotOrNot* [18] to estimate that bots accounted for around one fifth of generated content in the political discussion leading up to the 2016 election [7]. Several other studies have produced evidence that bot activity and propaganda significantly effected social media discussions in advance of the 2016 election [3, 25, 49, 61]. The months preceding the election also saw a surge in unauthorized data collection for the purpose of creating targeted political ads [54].

However, the notion of using human or automated accounts to spread propaganda via social media is not unique to the 2016 election in America: countries around the world deploy "cyber troops" to influence online discussions [11], and Ratkiewicz *et al* found that botnets were used to spread political content in advance of the 2010 US Midterm elections as well [46].

Misuse of social media tools in human hands can also lead to the unraveling of social movements. Writing of the 2011-2012 Indignados protests in Spain [35], Rone found that trolling behavior, the hijacking of social media accounts, the manipulation of voting systems, and the creation of fake accounts to infiltrate closed groups were all significant factors contributing to the dissolution of the movement [48]. Social networks have also been used to radicalize people to terrorist organizations [38, 52]. All in all, social media has become a hotbed for manipulative practices and the spread of political disinformation, via bots or otherwise[6].

1.2 Security through Inefficiency

We hypothesize that these online manipulation campaigns owe a large portion of their success to the mismatch between global scale ICT-supported human communication and the techniques humans have used to enforce group membership at the traditional community scale. In this paper we propose the paradigm of *security through inefficiency:* the reconsideration of efficiencies of scale for the sake of maintaining the security of an ICT based system. While it would be simple to take a Luddite–inspired path and reject these economies of scale wholesale, we claim that security can be improved through an intentional application of *social friction* that is aligned with an anthropological and sociological understanding of human coordination and communication.

The rest of the paper proceeds as follows:

- In Section 2, we define the notion of *social friction*, show its connections to anthropology and sociology, and discuss how it can act as a security mechanism that keeps unwanted outsiders from joining social groups.
- In Section 3, we examine the ways in which social friction manifests itself on the internet, critically analyze the effect that mainstream perceptions of the internet have had on the ability of bad-faith actors to spread propaganda and disinformation, and propose that building networks with more social friction could mollify some of these problems.
- In Section 4, we discuss two already-existing social networks, Mastodon and Nextdoor, which each satisfy some of the desiderata we see as integral to our model of social friction in online spaces.
- In Section 5, we propose a taxonomy of social network platforms based on their frictive properties, and propose research directions predicated on exploring the concept of social friction in social media.

2 SOCIAL FRICTION

We use the term *social friction* to describe the natural barriers to entry an individual may face when joining a community. If the community is a group of friends, a newcomer seeking to join the friend group will generally need to gain the trust of most group members before they are accepted. If the community is a town, a newcomer must buy or rent property in that town to be considered a resident. If the community is a family, then a newcomer must marry a family member (or be born of a family member) to be considered a member of that family. In each of these scenarios, a newcomer to the community needs to meet certain criteria (accruing trust, buying property, courtship and marriage) before they can attain community membership. We use the term *social friction* to describe the challenges encountered in meeting those criteria.

Social friction is related to the concepts of *social barriers* and *rites of passage* from Anthropology and Sociology. The concept of a *rite of passage* emerged from Anthropological research and has since seen use in other disciplines [4, 58]. Generally speaking, rites of passage are special events that "symbolize the transition of an individual or a group from one status to another" [58]. While the idea of a rite of passage does bear similarity to the idea of overcoming social friction, the former term is much more specific. A rite of passage denotes a well-defined ritual which occurs at a specific point in time, whereas the actions needed to overcome social friction (e.g. gaining trust) are not necessarily demarcated by any one event and can occur gradually over time. Nevertheless, some of the actions necessary to overcome social friction could be seen as rites of passage (e.g. a meeting to finalize the purchase of a house), so there is some overlap between the two concepts.

The term *social barrier*, on the other hand, is used in the literature to describe a non-technological hindrance that makes it more difficult for an individual or group to complete an action, whether that action be joining a social movement [27], visiting a city park [16], accepting a new energy source[41], or achieving economic empowerment[60]. The varied contexts in which the idea of a social barrier appears highlight a key difference between that term and our idea of social friction: scope. Where a "rite of passage" is too specific of a term to mesh well with our concept of social friction, a "social barrier" is instead too broad of an idea. As the above examples show, a social barrier can refer to a wide set of phenomena, whereas social friction is more specific. In fact, one could think of social friction as the specific social barriers that hinder an individual from joining a community (rather than, say, accepting a new energy source).

We also seek to distinguish the idea of social friction from the idea of social barriers in an perceptual sense. In the papers we encountered during our literature review, a social barrier was always framed by the authors as something negative that should be torn down if possible, and with good reason – many of these social barriers did indeed stand in the way of positive changes. But we do not see social friction as an inherently negative phenomenon. In fact, social friction is beneficial to communities in so far as it provides a natural vetting process against potentially malicious outsiders. Imagine, for example, if a stranger walked into your house and announced themselves as a new member of your family. You might have some misgivings about that stranger's intentions, and you might not want to give them all the privileges you give to other family members (a copy of the key to the living space, access to a shared bank account, etc.). Social friction, manifesting here as the requirement that this stranger gain the trust and love of an unmarried family member before joining the family, is what keeps such an intrusion from happening.

One can see social friction as a mechanism through which communities can provide security for themselves – a tool wielded by community members to make sure that only people they trust join the community. Specific communities can, either consciously or subconsciously, calibrate the amount of social friction they surround themselves with to their own specifications. In fact, one can notice that the amount of social friction surrounding a community is often correlated with the amount of harm that an unwanted community member could cause: attending an academic talk from a visiting lecturer is a process with relatively little friction, whereas joining a family is a high-friction process. If we consider social friction as a source of security for communities, this correlation makes perfect sense: it is natural to put more effort into protecting something that is more valuable.

3 FRICTION ON THE INTERNET?

In our view, the design of popular social networks encourages the formation of communities surrounded by very little friction. Any Twitter account can tweet at any other account, as long as both are public. Facebook friend requests are often accepted with relative ease in comparison to the effort involved in forming an offline friendship [45], and the same could be said of acceptance into Facebook groups [40].

The relative openness of these platforms should come as no surprise when one considers the context in which they were built: the internet has been (rightly) praised for, and achieved global relevance due to the way it allows ideas and digital artifacts to travel between people and communities with relatively little hindrance [9, 10]. In general, openness is seen as a defining characteristic of the internet [5, 17, 31]. That attitude has accordingly manifested itself in the domain of social networks: besides the evidence of social media openness mentioned above [40, 45], a recent survey of Sybil attack defenses in social networks cites openness as a fundamental property of online social networks [1]. At Facebook's 2016 F8 developer conference[2], the theme of CEO Mark Zuckerberg's keynote presentation was *"give everyone the power to share anything with anyone"* [55].

Most research on bot detection and Sybil attack defense in the context of social networks follows this same ideological trend. Many proposed solutions are *reactive* in nature, meaning that the authors take the openness of the system as a given, and build tools to figure out which actors within it are bots/Sybil identities. Of the 19 techniques surveyed in a 2017 article [1], all but three [56, 62, 63] were reactive. Research on bot detection (e.g. [18]) is also reactive, but such work is aimed at having a direct material impact on current social networks, so it is understandable that this vein of research would take the status quo as a given.

The nature of these approaches aside, the presence of bots is not the only factor contributing to the political disarray of most social networks anyways: real humans ("cyber troops") can also take up the task of spreading disinformation and propaganda, sometimes for pay [11]. Also problematic are the data collection practices of firms that use personal data to create targeted political messaging. The American public's negative response to such practices are perhaps best exemplified in the Facebook-Cambridge Analytica scandal [54]. Neither of these transgressions can be classified as Sybil attacks, but both of them involve bad-faith actors either giving (in the case of cyber troops) or taking (in the case of data collection) information to/from a user in a way that is not necessarily respectful of the user's views or in the spirit of healthy discourse. In both cases, the relative openness of online spaces is what allows these transgressions to occur.

At the end of the previous section, we noted an anecdotal correlation between the amount of social friction surrounding a real-world community and relative harm a malicious outsider could do to that community, were they able to gain membership. Given the turmoil caused by digital interference in social networks (see Section 1) and the way that social networks influence political decisions with material consequences, it would appear that many online communities do not adhere to this trend – that is to say, they have low social friction, but unwanted actors who enter can cause a lot of harm.

Overall, we worry that in the context of social media, and more specifically in the context of political discussions on social media, the prioritization of the openness of the internet has allowed for more harm than good. Specifically, the ability of any user to receive incoming information from any other source has been abused by

[2] https://www.f8.com/

malicious parties who inundate regular users with so much bad-faith information that distinguishing between the good from the bad is all but impossible. A recent Pew Research study supports this claim – the survey found that only 40 percent of respondents felt somewhat confident they could recognize a bot account on social media, and only 7 percent were very confident they could [12]. But again, bots are not even the whole picture – real users can also spread disinformation and agitation. Anthropological research suggests that humans cannot keeps track of more than (around) 150 relationships at a time [64], which further suggests that social media users simply do not have the mental capacity to maintain detailed trust information about all the accounts they know about online, be they bot or human. Moreover, foundational research on Sybil attacks shows that reactive approaches are far less effective than approaches that gate who can enter the community, except under exceedingly rare circumstances [19].

All of this evidence suggests that in certain situations it could be useful or even necessary to leverage friction as a security tool. Much current security research takes the openness of social networks as a given and tries to reduce abuse given that framework, but there are both technological [19] and psychological [12, 64] limits to that approach. On the other hand, an approach to security that incorporates friction as a foundational element has the potential to mollify the harms that arise out of bad-faith actors abusing the openness of social media.

4 NETWORK STRUCTURE AND SOCIAL FRICTION

In this section we analyze two existing social networks, Mastodon and Nextdoor, and discuss some of their structural elements which we feel would be conducive to incorporating social friction into online spaces for improved security.

4.1 Mastodon and distributed administration/ community Hosting

Mastodon[3] is microblogging service similar to Twitter. Users share short messages called "toots" (an analog to Tweets) which they can favorite, reply to, and "boost" (an analog to Twitter's retweet functionality). Mastodon does not have the recommendation features that Twitter has (i.e. "who to follow" suggestions), but recent work has shown that implementing recommender systems on top of Mastodon is possible [57]. Like Twitter, Mastodon is primarily developed by a centralized team. However, Mastodon's software is open source.

The major difference between Mastodon and Twitter (and most other social networks) is the nature in which its communities and administrative privileges are distributed. Mastodon is made up of many interconnected servers, each hosted by an individual or party who need not be connected to the developers of Mastodon, and when a user joins Mastodon they choose a specific server to join (although they can migrate their content to an account on another server later if they so choose), and in general, there is nothing stopping a user on one server from interacting with a user on another server *a priori*. Moreover, the developers of Mastodon hold none of the administrative privileges. Instead, the parties who run each individual mastodon server have administrative priveleges over what goes on on their server, including power over who is allowed to join and who is allowed to see posts from users on that server. In this way, Mastodon is decentralized with respect to its distribution of administrative privileges and its distribution of community-hosting responsibility.

This naturally allows for a framework by which communities can create social friction around themselves. Specifically, each server is free to set their own guidelines on who is allowed to enter. These guidelines can be social (e.g. community members discuss whether or not they want to let the outsider in), technological (e.g. a user is let into the server that residents of a town use to talk about local politics if they can provide a crpytographic proof of their residence in the town), or a mix of both.

Letting communities set their own security and membership guidelines has another advantage: it reduces the scalability of attacks. Any bot-detection algorithm implemented by an open social network like Twitter will have the drawback that an any successful evasion of the algorithm will scale quite well. That is to say, since all communities on Twitter are uniform with regards to their protection under a bot-detection algorithm, a strategy for evading the algorithm in one community will also work in any other community. In contrast, letting every decentralized server create their own membership requirements is a security strategy that will not lend itself well to wide-scale attacks. In a world where each server is encouraged to customize their membership requirements, an attack that works well on one server is by no means guaranteed to work on any other server. This is an important advantage because it greatly reduces the economic efficiency of any potential attack. Moreover, since under our model the development privileges are still centralized, the developers of a system like Mastodon could still create a bot-detection algorithm and let servers adopt it for an extra layer of security.

Lastly, we note that letting the developers of a social media network hold administrative privileges represents an potential conflict of interest with regards to the curtailing of bots and propaganda. Specifically, barring the negative effects of public outrage over heavily publicized scandals involving bots and propaganda, one might imagine that from the developers' perspective, bot activity and propaganda is good for business. Inflammatory content is shared with higher frequency on social media [51], and bots are more likely to share inflammatory content [50]. Therefore, if one measures the success of a social network by the amount of traffic its users generate, then it would appear that curtailing bots might not always be in the best economic interests of the managers of the network. Of course, it is also possible that a period of intense public backlash to bot activity constitutes a financial threat more severe than the financial gain of allowing bots, and in this case the developers of a social network would instead be incentivized to crack down on bot activity. But in either case, centralizing a network's administrative privileges to the same people who run the network leads to scenarios in which the administrators wield those privileges in the way that will provide them with the most financial benefit, rather than the way that will create the best discursive environment.

We do not mean to bring up Mastodon in order to suggest that it needs to become the center of research or development efforts

[3] https://joinmastodon.org/

centered around social friction in particular, but only to suggest that its operational model, and specifically its decentralization of administration and hosting responsibilities, has many potentially positive qualities with regard to enabling communities to provide themselves with increased security.

4.2 Nextdoor and location-specific networks

Given that we propose geographically-based online communities as a main use-case for the idea of social friction, we now discuss Nextdoor[4], a social networking platform tailored to individual neighborhoods. We examine the security Nextdoor provides to its users and review academic work on the quality of its communities.

Nextdoor is a social networking platform where users join communities specific to the neighborhood they live in. In order to join one such community, a user must either send Nextdoor a picture of their driver's license or enter a code on a postcard mailed to their address. In this way, Nextdoor centralizes an important administrative privilege: the power to decide who can join a neighborhood's network. However, power users can gain some other administrative privileges. Unlike Mastodon, Nextdoor also centralizes hosting responsibility.

Masden *et al* conducted interviews with 13 Nextdoor users across various neighborhoods in a metropolitan area and found that participants reported strong community engagement and "a lack of divisive or combative content" on the platform, however they also reported that privacy concerns and disagreements about the boundaries of specific neighborhoods were a cause for tension [32]. We are heartened by Masden's positive findings, and feel that participants' anxieties about privacy and neighborhood boundaries could largely be mollified in a decentralized system that offered increased privacy protections (partially possible due to a reduced need to please advertisers) and the ability for self-sovereign communities to change their boundaries over time, rather than having those boundaries controlled by a centralized source.

Payne also critiques the rigid geographic boundaries imposed by Nextdoor [42]. Again, under a decentralized scheme, online communities would not need to be so discretely divided, even if they were tied to specific locations.

On the other hand, Kurwa conducted an exploratory analysis in which he found that Nexdoor "has become an important platform for the surveillance and policing of race in residential space" [30]. In section 7, we outline future work to better understand this phenomenon and the extent to which it is endemic to neighborhood-specific online communities.

We also note that the entry requirements Nextdoor places on its users are perhaps at the upper bound of how strict a community could be about its entrance requirements – in the world of social friction, these requirements may be analogous to very coarse sandpaper. We imagine a network of communities in which each community is free to define its own membership requirements, and for a geographically-based community, those membership requirements could be a proof of residence, but they could also be something more lenient. For example, an online community centered around a town may allow people from that town as well as neighboring towns to join, or it might allow past residents to join, or residents from a neighboring town that at least n residents can vouch for, etc.

In section 7, we also outline future work to leverage cryptography (specifically zero-knowledge proofs) to allow users to make statements like "I spend at least 50 percent of my time in this town", which could be used as inputs into a community's scheme of entrance requirements. Zero-knowledge proofs of statements like this could be especially useful with respect to more complicated geographic situations where a user's place of residence does not necessarily reflect the entire scope of their political interests. For example, one could imagine a situation where many people live in area A but commute to area B for work – perhaps these people should have a say in discussions about the economic policies of area B. Nextdoor does not support this type of fine grained and community-specific specialization of entrance requirements.

Overall, although some research Nextdoor is hopeful with regards to the discursive environments of geographically-centered online networks, we ultimately seek a system with more flexibility than what Nextdoor provides. Furthermore, we note that Nextdoor's centralization of administrative capability is likely a factor contributing to its inflexibility of entrance requirements across communities – for a centralized team, it is much more economical to create one set of entrance requirements that can be applied in any context.

5 RESEARCH DIRECTIONS

In this paper, we have hypothesized that enabling communities to establish social friction around themselves will increase security and lead to richer and more productive deliberatory experiences inside said communities. Moreover, we have hypothesized that the centralization or decentralization of a network's administrative and hosting responsibilities has a profound impact on the way communities can create social friction around themselves.

To better understand the axes along which various social networks are centralized or decentralized, we have taxonomized several popular social networks (Table 1). We then leverage this taxonomy to outline plans for future work to empirically validate and further explore these ideas.

5.1 A taxonomy of networks

In Table 1, we taxonomize several popular online networks according to their centralization or decentralization of development responsibility, administrative responsibility, and community hosting responsibility.

Of these networks, three have decentralized administrative privileges (Reddit, Slack/Discord, and Mastodon), and of those three only Mastodon is decentralized with regards to community hosting. But without decentralized community hosting, administrative privileges are not *truly* decentralized, since the power given to administrators can still be altered by the central entity who hosts all the communities. For example, there is nothing stopping Reddit from changing its policy on subreddit administration and transferring some moderator power to a centralized in-house team.

We can also see the "report" functions on Twitter and Facebook (via which anyone can flag an inappropriate post for moderator review) as an attempt to distribute some administrative capability

[4] https://nextdoor.com/

Table 1: A taxonomy of online networks

Platform	Development responsibility	Administrative responsibility	Community hosting responsibility
Twitter	Centralized	Centralized	
Facebook	Centralized	Centralized (although, within Facebook *groups*, moderators have some administrative privileges)	Centralized
Reddit	Centralized	Decentralized	Centralized
Slack/ discord	Centralized	Decentralized	Centralized
Group Chats	Centralized	N/A – no one has administrative privileges	Centralized
Mastodon	Decentralized (open source)	Decentralized	Decentralized
Nextdoor	Centralized	Semi-centralized: power users can gain some administrative privileges	Centralized
Blockchain Apps	Decentralized (open source)	N/A – application-dependent	Decentralized

throughout the network. However, any reports made under such a scheme must still pass through a centralized bottleneck of in-house moderators.

This taxonomy is useful in guiding plans for future work, but we also note that expanding it to catalog networks in different ways might constitute a research direction in its own right.

5.2 Validating friction

To test our hypothesis on the effectiveness of social friction, we will conduct studies that measure the quality of conversations across online communities that exhibit various amounts of social friction. One promising set of communities to examine are Reddit's "subreddits", many online communities each with their own sets of rules (a property made possible by Reddit's distribution of administrative privileges). By using sentiment analysis to measure the nature of conversations, we can compare trends across different subreddits with different entrance requirements and codes of conduct, and explore the extent to which these factors correlate with better conversations.

5.3 Examining Nextdoor

Nextdoor is of particular interest to us. It would appear that very little academic literature has examined Nextdoor: in our literature review, we were only able to find three papers [30, 32, 42], all of which were discussed in section 4.2. Importantly, Masden *et al* 's interviews and Kurwa's analysis paint very different pictures of Nextdoor communities, with Masden *et al* suggesting that Nextdoor can be a positive force for communities [32] and Kurwa arguing that Nextdoor can be abused as a tool for discriminatory surveillance [30]. We hope to conduct more interviews to better understand the ways in which different communities use Nextdoor, the harms that can arise out of its misuse, and the extent to which those harms are endemic to location-specific online communities.

5.4 Overlaying friction on other networks

The network effects [26] exhibited by large social networks mean that wide-scale adoption of other platforms might come slowly, or not at all. But the idea of social friction can also be used to build overlays on existing social networks, which presents a potentially appealing compromise.

For example, one could imagine a tool that lets Twitter users define their own metric of trust relative to another user (e.g. number of followers in common, distance away in the follow graph, etc) and augments the user's Twitter feed by only showing the user tweets from other users who attain a threshold trust score, under whatever metric the user defines. Since each user is free to customize their own metric of trust, we gain an advantage from decentralization similar to the advantage gained by letting each individual Mastodon server set its own membership rules. That is to say, for users Alice and Bob with different trust metrics, a strategy that lets an attacker create many bots that can gain Alice's trust will not necessarily succeed in creating bots that can unfairly gain Bob's trust.

We are interested in building tools like the one mentioned above, partially for their immediate utility and partially because doing so will allow us to further explore the ways that friction can be applied as a security primitive in online spaces.

But we also note that it might not be appropriate to incorporate the idea of friction into every social network. In many scenarios, the ability to receive information from a previously unknown or untrusted source should be seen as a net positive.

5.5 Evaluating the formation of filter bubbles

Enabling communities to surround themselves with friction could lead to filter bubbles [39]. Also colloquially referred to as "ideological echo chambers", filter bubbles are online spaces where users only associate with other users whose politics closely align with theirs.

Filter bubbles are not necessarily endemic to networks with a lot of friction, as they can be found in current social networks [21]. Other research suggests that factors underlying the formation of filter bubbles are innate to human psychology [28]. This suggests that the relationship between the friction surrounding a community and its ideological uniformity is not as clear-cut as conventional thinking might suggest. We aim to critically examine the actual trade-off between the security of an online community (in terms of friction) and the extent to which that community exhibits filter bubble-like properties.

However, we also note that in the context of political discussion, our goal is not even to necessarily to enable the creation communities that are divided along political lines, but instead to enable the

creation of communities that are free of bad-faith outsiders who might wish to sway the conversation in one direction or another.

A fitting analog might be Fishkin's concept of Deliberative Democracy [20], which has been implemented most recently via the America in One Room project [13]. This event saw Americans from all walks of life invited to a single location in Texas where they debated topics germane to American politics with each other. The organizers of this event did not seek to exclusively invite participants with a specific political orientation, but they *did* seek to only invite participants who were from America, given that the event was centered around debating American politics. In the same way, our hope is that online communities can use social friction to keep out bad-faith outsiders while still allowing for healthy and ideologically diverse debate.

Furthermore, we note that if one was able to build such an online community, its ideological diversity (assuming diversity of the underlying population) could make it *less* of an echo chamber than current sub-communities of popular social networking cites – as stated above, these communities have already been found to exhibit filter bubble-like properties [21].

5.6 Cryptography and geographic authentication

Suitably private and secure location attestation is a problem that could be well served by cryptographic primitives. Before designing such a system, it is important to carefully consider which facts must be attested and how they can be used to create social friction without sacrificing the efficiency gains of information technology.

One possible approach draws inspiration from the privacy preserving contract tracing specification released by Apple and Google [2]. The specification was developed for the purpose of informing users if they spent time near someone diagnosed with COVID-19, without revealing the identity of the diagnosed individual. Essentially, the specification describes a scheme in which phones within a close proximity to each other exchange random-looking numbers over Bluetooth. Each phone records a list of numbers it has both sent and received. When an individual is diagnosed with COVID-19, they can choose to anonymously upload their list of *sent* random numbers to a central server. All phones query the server daily and check if any uploaded numbers appear in that phone's list of received numbers – if a match occurs, then that phone alerts its user that they came into contact with someone who had COVID-19.

A similar and simple scheme could be used to support location attestation as follows: businesses set up bluetooth-enabled devices to communicate with customer phones. When a customer visits a business, they send the store's bluetooth device a random number. The business appends some geographic data (like the name of the town the business is in) to the number, hashes it, and sends the result to a public server to which only businesses have write access. Then, in order to prove that they were in a specific location, a user can reveal their random number and the relevant geographic data, and others can reconstruct the hashed value and check that it appears on the server. To prove long-term residence, a user could make many such claims, and if businesses also appended a timestamp before hashing then users could attest to their location over time. This scheme is not robust to businesses and customers conspiring together, but provides a promising starting point for future work.

REFERENCES

[1] Muhammad Al-Qurishi, Mabrook Al-Rakhami, Atif Alamri, Majed Alrubaian, Sk Md Mizanur Rahman, and M Shamim Hossain. 2017. Sybil defense techniques in online social networks: a survey. *IEEE Access* 5 (2017), 1200–1219.

[2] Inc. Apple. 2020. *Privacy-Preserving Contract Tracing.* Retrieved May 22, 2020 from https://www.apple.com/covid19/contacttracing

[3] Adam Badawy, Emilio Ferrara, and Kristina Lerman. 2018. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).* IEEE, 258–265.

[4] Alan Barnard and Jonathan Spencer (Eds.). 2010. *The Routledge Encyclopedia of Social and Cultural Anthropology, Second Edition.* Emerald Group Publishing Limited. 616–617 pages.

[5] Anja Bechmann and Stine Lomborg. 2014. *The ubiquitous internet: user and industry perspectives.* Routledge. https://books.google.com/books/about/The_Ubiquitous_Internet.html?id=-VZWBQAAQBAJ

[6] Yochai Benkler, Robert Faris, and Hal Roberts. 2018. *Network propaganda: Manipulation, disinformation, and radicalization in American politics.* Oxford University Press. https://doi.org/10.1093/oso/9780190923624.001.0001

[7] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday* 21, 11 (Nov. 2016). https://doi.org/10.5210/fm.v21i11.7090

[8] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. 2013. Design and Analysis of a Social Botnet. *Comput. Netw.* 57, 2 (Feb. 2013), 556–578. https://doi.org/10.1016/j.comnet.2012.06.006

[9] Sarah Box. 2016. Internet Openness and Fragmentation: Toward Measuring the Economic Effects. (2016). https://www.cigionline.org/publications/internet-openness-and-fragmentation-toward-measuring-economic-effects

[10] Sarah Box and Jeremy K West. 2016. Economic and social benefits of internet openness. (2016). https://doi.org/10.1787/20716826

[11] Samantha Bradshaw and Philip Howard. 2017. Troops, trolls and troublemakers: A global inventory of organized social media manipulation. *CompProp, OII, Working Paper* (2017). https://comprop.oii.ox.ac.uk/research/troops-trolls-and-trouble-makers-a-global-inventory-of-organized-social-media-manipulation/

[12] Pew Research Center. 2018. *Social Media Bots Draw Public's Attention and Concern.* Retrieved May 20, 2020 from https://www.journalism.org/2018/10/15/social-media-bots-draw-publics-attention-and-concern/

[13] Stanford center for Deliberative Democracy. 2019. *America in One Room - CDD.* Retrieved May 17, 2020 from https://cdd.stanford.edu/2019/america-in-one-room/

[14] Michael D Conover, Emilio Ferrara, Filippo Menczer, and Alessandro Flammini. 2013. The digital evolution of occupy wall street. *PloS one* 8, 5 (05 2013), e64679. https://doi.org/10.1371/journal.pone.0064679

[15] Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media.* https://www.semanticscholar.org/paper/Detecting-and-Tracking-Political-Abuse-in-Social-Ratkiewicz-Conover/7e8a5e0a87fab337d71ce04ba02b7a5ded392421

[16] Bethany B Cutts, Kate J Darby, Christopher G Boone, and Alexandra Brewis. 2009. City structure, obesity, and environmental justice: an integrated analysis of physical and social barriers to walkable streets and park access. *Social science & medicine* 69, 9 (2009), 1314–1322.

[17] Leslie Daigle. 2015. On the Nature of the Internet. (2015). https://www.cigionline.org/publications/nature-internet

[18] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. BotOrNot: A System to Evaluate Social Bots. In *Proceedings of the 25th International Conference Companion on World Wide Web* (Montréal, Québec, Canada) *(WWW '16 Companion).* International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 273–274. https://doi.org/10.1145/2872518.2889302

[19] John R. Douceur. 2002. The Sybil Attack. In *Peer-to-Peer Systems*, Peter Druschel, Frans Kaashoek, and Antony Rowstron (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 251–260.

[20] James S. Fishkin. 1991. *Democracy and Deliberation: New Directions for Democratic Reform.* Yale University Press. http://www.jstor.org/stable/j.ctt1dt006v

[21] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18).* International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 913–922. https://doi.org/10.1145/3178876.3186139

[22] Sandra González-Bailón, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. 2011. The dynamics of protest recruitment through an online network. *Scientific reports* 1 (2011), 197. https://doi.org/10.1038/srep00197

[23] Edward S Herman and Noam Chomsky. 1988. *Manufacturing consent: The political economy of the mass media.* Random House.

[24] Philip N Howard et al. 2006. *New media campaigns and the managed citizen.* Cambridge University Press. https://doi.org/10.1017/CBO9780511615986

[25] Philip N Howard, Gillian Bolsover, Bence Kollanyi, Samantha Bradshaw, and Lisa-Maria Neudert. 2017. Junk news and bots during the US election: What were Michigan voters sharing over Twitter. *CompProp, OII, Data Memo* (2017). 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/

[26] Michael L. Katz and Carl Shapiro. 1994. Systems Competition and Network Effects. *Journal of Economic Perspectives* 8, 2 (June 1994), 93–115. https://doi.org/10.1257/jep.8.2.93

[27] Bert Klandermans and Dirk Oegema. 1987. Potentials, networks, motivations, and barriers: Steps towards participation in social movements. *American sociological review* (1987), 519–531.

[28] Silvia Knobloch-Westerwick and Jingbo Meng. 2011. Reinforcement of the Political Self Through Selective Exposure to Political Messages. *Journal of Communication* 61, 2 (04 2011), 349–368. https://doi.org/10.1111/j.1460-2466.2011.01543.x arXiv:https://academic.oup.com/joc/article-pdf/61/2/349/22324723/jjnlcom0349.pdf

[29] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America* 111, 24 (06 2014), 8788–8790. https://doi.org/10.1073/pnas.1320040111

[30] Rahim Kurwa. 2019. Building the Digitally Gated Community: The Case of Nextdoor. *Surveillance & Society* 17, 1/2 (2019), 111–117. https://doi.org/10.24908/ss.v17i1/2.12927

[31] L. Lessig. 2002. *The Future of Ideas: The Fate of the Commons in a Connected World.* Knopf Doubleday Publishing Group. https://books.google.com/books?id=dWf-p2SkGQ8C

[32] Christina A. Masden, Catherine Grevet, Rebecca E. Grinter, Eric Gilbert, and W. Keith Edwards. 2014. Tensions in Scaling-up Community Social Media: A Multi-Neighborhood Study of Nextdoor. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14).* Association for Computing Machinery, New York, NY, USA, 3239–3248. https://doi.org/10.1145/2556288.2557319

[33] Johnnatan Messias, Lucas Schmidt, Ricardo Oliveira, and Fabrício Benevenuto. 2013. You followed my bot! Transforming robots into influential users in Twitter. *First Monday* 18, 7 (Jun. 2013). https://doi.org/10.5210/fm.v18i7.4217

[34] Robert S Mueller and Man With A. Cat. 2019. *Report on the investigation into Russian interference in the 2016 presidential election.* Vol. 1. US Department of Justice Washington, DC. https://www.hsdl.org/?view&did=824221

[35] BBC News. 2012. *Spain's Indignados protest here to stay.* Retrieved May 18, 2020 from https://www.bbc.com/news/world-europe-18070246

[36] BBC News. 2017. *Bots used to bias online political chats.* Retrieved May 20, 2020 from https://www.bbc.com/news/technology-40344208

[37] BBC News. 2017. *Pro-Indian 'fake websites targeted decision makers in Europe.* Retrieved May 20, 2020 from https://www.bbc.com/news/world-asia-india-50749764

[38] Simon O'Rourke. 2007. Virtual radicalisation: Challenges for police. In *8th Australian Information Warfare and Security Conference.* School of Computer and Infomation Science, Edith Cowan University, Perth. https://doi.org/10.4225/75/57a83c57befab

[39] Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think.* Penguin.

[40] Namsu Park, Kerk F Kee, and Sebastián Valenzuela. 2009. Being immersed in social networking environment: Facebook groups, uses and gratifications, and social outcomes. *CyberPsychology & Behavior* 12, 6 (2009), 729–733. https://doi.org/10.1089/cpb.2009.0003

[41] Martin J Pasqualetti. 2011. Social barriers to renewable energy landscapes. *Geographical Review* 101, 2 (2011), 201–223.

[42] Will Payne. 2017. Welcome to the Polygon: Contested Digital Neighborhoods and Spatialized Segregation on Nextdoor. *Computational Culture* 6 (2017). http://computationalculture.net/welcome-to-the-polygon-contested-digital-neighborhoods-and-spatialized-segregation-on-nextdoor/

[43] Twitter Public Policy. 2018. *Update on Twitter's review of the 2016 US election.* Retrieved May 18, 2020 from https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html

[44] Quartz. 2017. *Twitter has a serious bot problem, and Wikipedia might have the solution.* Retrieved May 20, 2020 from https://qz.com/1108092/twitter-has-a-serious-bot-problem-and-wikipedia-might-have-the-solution/

[45] Hootan Rashtian, Yazan Boshmaf, Pooya Jaferian, and Konstantin Beznosov. 2014. To Befriend Or Not? A Model of Friend Request Acceptance on Facebook. In *10th Symposium On Usable Privacy and Security (SOUPS 2014).* USENIX Association, Menlo Park, CA, 285–300. https://www.usenix.org/conference/soups2014/proceedings/presentation/rashtian

[46] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. 2011. Truthy: Mapping the Spread of Astroturf in Microblog Streams. In *Proceedings of the 20th International Conference Companion on World Wide Web* (Hyderabad, India) *(WWW '11).* Association for Computing Machinery, New York, NY, USA, 249–252. https://doi.org/10.1145/1963192.1963301

[47] Jacob Ratkiewicz, Michael D Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer Menczer. 2011. Detecting and tracking political abuse in social media. In *In Proceedings of the 5th AAAI International Conference on Weblogs and Social Media (ICWSM'11).*

[48] Julia Rone. 2019. Fake profiles, trolls, and digital paranoia: digital media practices in breaking the Indignados movement. *Social Movement Studies* 0, 0 (2019), 1–17. https://doi.org/10.1080/14742837.2019.1679108 arXiv:https://doi.org/10.1080/14742837.2019.1679108

[49] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications* 9, 1 (2018), 1–9. https://doi.org/10.1038/s41467-018-06930-7

[50] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. 2018. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences* 115, 49 (2018), 12435–12440. https://doi.org/10.1073/pnas.1803470115

[51] Stefan Stieglitz and Linh Dang-Xuan. 2013. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems* 29, 4 (2013), 217–248. https://doi.org/10.2753/MIS0742-1222290408

[52] Robin Thompson. 2011. Radicalization and the Use of Social Media. *Journal of Strategic Security* 4, 4 (2011), 167–190. http://www.jstor.org/stable/26463917

[53] New York Times. 2018. *Chatbots Are a Danger to Democracy.* Retrieved May 20, 2020 from https://www.nytimes.com/2018/12/04/opinion/chatbots-ai-democracy-free-speech.html?auth=login-email&login=email

[54] The New York Times. 2018. *Cambridge Analytica and Facebook: The Scandal and the Fallout So Far.* Retrieved May 17, 2020 from https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

[55] USA Today. 2016. *Facebook's Mark Zuckerberg steps into political fray.* Retrieved May 21, 2020 from https://www.usatoday.com/story/tech/2016/04/12/zuckerbergs-10-year-plan-expand-facebook-empire/82936814/

[56] N. Tran, J. Li, L. Subramanian, and S. S. M. Chow. 2011. Optimal Sybil-resilient node admission control. In *2011 Proceedings IEEE INFOCOM.* 3218–3226.

[57] Jan Trienes, Andrés Torres Cano, and Djoerd Hiemstra. 2018. Recommending Users: Whom to Follow on Federated Social Networks. *CoRR* abs/1811.09292 (2018). arXiv:1811.09292 http://arxiv.org/abs/1811.09292

[58] Rodanthi Tzanelli. 2007. The Blackwell Encyclopedia of Sociology. , 3940—3941 pages.

[59] Onur Varol, Emilio Ferrara, Christine L. Ogan, Filippo Menczer, and Alessandro Flammini. 2014. Evolution of Online User Behavior during a Social Upheaval. In *Proceedings of the 2014 ACM Conference on Web Science* (Bloomington, Indiana, USA) *(WebSci '14).* Association for Computing Machinery, New York, NY, USA, 81–90. https://doi.org/10.1145/2615569.2615699

[60] Michael Woolcock. 2000. Removing Social Barriers and Building Social Institutions. *World Development Report* 2001 (2000).

[61] Samuel C Woolley and Douglas Guilbeault. 2017. Computational propaganda in the United States of America: Manufacturing consensus online. *CompProp, OII, Working Paper* (2017). https://comprop.oii.ox.ac.uk/research/working-papers/computational-propaganda-in-the-united-states-of-america-manufacturing-consensus-online/

[62] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao. 2010. SybilLimit: A Near-Optimal Social Network Defense Against Sybil Attacks. *IEEE/ACM Transactions on Networking* 18, 3 (2010), 885–898.

[63] Haifeng Yu, Michael Kaminsky, Phillip B. Gibbons, and Abraham D. Flaxman. 2008. SybilGuard: Defending against Sybil Attacks via Social Networks. *IEEE/ACM Trans. Netw.* 16, 3 (June 2008), 576–589. https://doi.org/10.1109/TNET.2008.923723

[64] W-X Zhou, Didier Sornette, Russell A Hill, and Robin IM Dunbar. 2005. Discrete hierarchical organization of social group sizes. *Proceedings of the Royal Society B: Biological Sciences* 272, 1561 (2005), 439–444. https://doi.org/10.1098/rspb.2004.2970