The Economics of Confrontational Conversation

Shane Greenstein, *Harvard Business School, Boston, MA, 02163, USA*

My favorite panel from Randall Munroe's one-panel comic, xkcd, is labeled "duty calls." It shows a lone stick figure at a desk, hunched over, intensely staring at his computer screen, while engaging in a staccato conversation with an offscreen companion. She says: "Are you coming to bed?" Him: "I can't. This is important." Her: "What?" Him: "Someone is *wrong* on the internet."

Like all good satire, this comic elicits both laughs and winces. Nobody would ever engage in this behavior in any physical place where a veneer of social politeness predominates, such as standing in line at a cashier or sitting in an airplane seat. On the internet surfers jettison much of their social restraint, confronting, and correcting perfect strangers. It leads to, for example, edit wars on Wikipedia, condescending insults on Reddit, and righteous putdowns on Twitter.

This behavior invites plenty of legal analysis, angry editorializing, and technological proposals, but rarely economic analysis. Let us address that gap. What economic factors make confrontational conversation more or less likely in our era?

TENSIONS WITH COMPROMISE

The first piece of relevant economics is the low cost of scale. It is inexpensive to host terabytes of data, and dirt-cheap to serve millions of users. Software can be replicated many times at almost no cost, making it possible for a platform to scale.

The second relevant economic factor complements scale, and goes by the label "network effects." These are self-reinforcing advantages affiliated with being a focal platform. Simplifying, a platform becomes focal in one of two ways. In one form a platform attracts more apps or content, and that attracts more users, which then attracts more apps or content, and so on. In another form a platform attracts more sellers, which attracts more buyers, which attracts more sellers and so on. In either case, those attractions become self-reinforcing.

A third crucial piece of economics shapes focal platforms with large scale. Some confrontation is inevitable, and undermines the functionality of most (not all!) platforms. It is possible to write volumes on how to address confrontation with human moderation or algorithmic processes, and whether specific practices are legal or effective. Take a step back from that discussion, and recognize the broad economic facts. No matter how it gets implemented, the processes are expensive and imperfect.

ON THE INTERNET SURFERS JETTISON MUCH OF THEIR SOCIAL RESTRAINT, CONFRONTING, AND CORRECTING PERFECT STRANGERS. IT LEADS TO, FOR EXAMPLE, EDIT WARS ON WIKIPEDIA, CONDESCENDING INSULTS ON REDDIT, AND RIGHTEOUS PUTDOWNS ON TWITTER. THIS BEHAVIOR INVITES PLENTY OF LEGAL ANALYSIS, ANGRY EDITORIALIZING, AND TECHNOLOGICAL PROPOSALS, BUT RARELY ECONOMIC ANALYSIS.

Two features of the present era drive up those costs and exacerbate the imperfections. Anything with long video is inherently expensive to moderate, as managers at YouTube and Facebook can attest. In addition, bots and misinformation farms have flooded the modern experience, especially at focal platforms—just ask Twitter and Facebook. Holding those in check has become an endless and expensive whack-a-mole for large platforms.

More to the point, addressing content moderation has become this era's key to achieving scale. Different platforms have taken different approaches to this challenge, and each approach comes with different upsides and drawbacks.

0272-1732 © 2021 IEEE
Digital Object Identifier 10.1109/MM.2021.3060295
Date of current version 26 March 2021.

DIFFERENT APPROACHES

Apple's approach to content moderation is the easiest to understand, namely, it "sanitizes" its content during a preapproval process for apps. Apple treats violent content and confrontational language just like it treats porn. It bans everything that violate some rules and bounds, with the (possible) exception of some video games.

Is that expensive? Yes, because humans must do it. Moreover, the approval process is arduous, and many developers dislike it. So why do it? Because it protects the brand. Would most parents buy an iPhone for a teenage child if it included an app for quasi-Nazi nonsense? Of course not. As in any other business, Apple targets a primary customer, and aligns their operations with serving these customers.

Parts of this approach has migrated to other platforms. Most platforms want nothing to do with Holocaust deniers and flat-earthers or violent terrorists of any sort, so they ban them, using their terms of service as a legal shield.

DURING THE ELECTION A NEW TYPE OF BREAKAWAY EMERGED WHEN POLITICAL OPINION SHAPERS COUCHED THEIR APPEAL IN RIGHTEOUS PATRIOTISM. LOVE THEM OR HATE THEM, THEIR EXPERIENCE ILLUSTRATES THE ECONOMICS.

Sanitizing alone does not work on all platforms in all situations. That mixed success motivates another approach, channeling the confrontations with clever design accompanied by a few sanitizing rules about indecent language. For example, a customer may hate a restaurant for a good or crazy reason, but Yelp found a way to guide the confrontations, and make them valuable to readers. Yelp displays the average as stars, and spotlights a colorful description, which users can read or ignore.

A whole range of ratings sites use different variations on this approach. However, fake reviews and offensive reviews (that just skirt the rules) have become the pest in today's version of whack-a-mole. Without naming names, it undermines the usefulness of some ratings sites.

Facebook illustrates a different approach, namely, using cleverly designed algorithms to manipulate who takes part in the confrontational conversation. One might call it "divide but engage" because this approach tends to divide participants into like-minded groups. Facebook came to this approach through testing and refining its algorithms for the news feed to maximize engagement, showing users content that kept them for longer and motivated them to come back.

As has been widely noted, this approach has one major drawback. It creates insulated conversation bubbles of outrage among like-minded friends. Moreover, a flood of misinformation and bots make this problem worse. Some time ago Facebook started to feel the downside in terms of user fatigue and disgust with the site overall, so it has been experimenting with policing the misinformation. It banned many fake sites. Like other platforms, Facebook also implemented labels for lies and for unverified rumors. Like every imperfect approach, this one was costly, and it frustrated plenty of users.

And, yet, they survive and still thrive. The economic lesson? The site held together because the network effects are strong enough.

A final approach melds clever design and sanitizing into something more muddled. While there are many examples, the most successful of these is Google's approach to its search engine. Google sanitizes its ads, banning alcohol, gambling, porn, and illegal activities, as well as indecent language. At the same time, Google has long taken a more permissive approach to its organic search results. If a user searched for something shady they got back something, well, shady. Users filter the offensive results if they want to. This was (and still is) a clever and effective way to find anything except the dark web.

Its downsides are also well known: It remains vulnerable to special pleading. Google fields endless complaints about its organic results. Many governments around the world wants to regulate it, while many politicians want to manipulate it.

PRESSURES AGAINST SCALE

We just described settings where network effects are strong. What happens when they are not? Simply stated, a breakaway community forms if users find an entrepreneur who creates an alternative platform that users prefer.

Breakaway communities with facets of indecency have been a part of the internet for a long time. For example, Reddit, 4chan, or many places in the dark web wear their confrontational language as a badge of honor. More to the point, the skills to manage one of these sites are common, and it is a lot cheaper to operate if the moderation is light. It is not my taste

and it may not be yours, but there is a large supply of these breakaways.

During the election a new type of breakaway emerged when political opinion shapers couched their appeal in righteous patriotism. Love them or hate them, their experience illustrates the economics.

It started with, say, Alex Jones and Infowars, who got banned because it went beyond the pale. Many mainstream sites thought it was the right thing to do, and it aligned with the economics because the losses of user were small compared to the risks to offending too many disgusted customers. Once banned, Alex Jones took his users with him and formed his own community elsewhere.

AS DEEPFAKES BECOME MORE COMMON, WHO WILL ADJUDICATE WHETHER A DEEPFAKE OF A POLITICIAN OR CELEBRITY IS REAL OR NOT? HOW CAN ANYBODY DO THAT IF ONLINE USERS HAVE SORTED THEMSELVES INTO VARIOUS GROUPS THAT DO NOT TRUST ONE ANOTHER?

Parler's recent experience exemplifies the state of play today. As a reminder, Parler came into existence because many white nationalist groups tired of Facebook's and Twitter's attempts to label their content as factually problematic and potentially violent. Parler opened on AWS, and made lack of moderation a feature instead of a bug, and it attracted a following quickly.

The other mainstream platforms eventually reacted in predictable ways. Apple was among the earliest to ban Parler's app, consistent with its sanitizing approach. AWS eventually they took a similar action after notifying Parler of their concerns about lack of moderation. Once Parler gained prominence as a coordinator of the capital riot, AWS saw no benefit to having its brand associated with this group, and that was that.

Parler's breakaway did not go well because management cut many corners and did not design a hack-proof site, but here is the rub. Parler is trying again. Even if they fail again, it is a good bet that somebody else will get a version of this up and running.

Summarizing, the increasing frequency of breakaways is a symptom that they are becoming cheaper to build. Ergo, we should expect mainstream sites to face increasing pressures towards fragmentation.

CONCLUSION

The trends toward fragmentation worries anyone who wants to maintain civil society. Who will encourage the confrontations that settle the public conversations?

Most worrisome, misinformation, and deep fakes are becoming more widespread in breakaway communities, and especially on the dark web. Right now, most users of deepfakes entertain themselves (You do not really want to know the details.), but, as with any frontier software, it will become mainstream soon enough.

As deepfakes become more common, who will adjudicate whether a deepfake of a politician or celebrity is real or not? How can anybody do that if online users have sorted themselves into various groups that do not trust one another?

SHANE GREENSTEIN is a Professor at the Harvard Business School. Contact him at sgreenstein@hbs.edu.