Speech and Safety Laboratories

DRAFT // do not distribute or cite without permission

Nicholas Bramble (nbramble@google.com)

For discussion at Yale Free Expression Scholars Conference (May 4, 2014)

Abstract: Along with spawning numerous diverse laboratories for free expression, the Internet has also generated a number of user-driven strategies for fighting abuse and promoting user safety. Many networks rely heavily on users to alert them about abusive content via flags or other reporting mechanisms. However, an increasing number of networks are also developing digital toolkits for users to take action on their own to minimize their exposure to content that shocks or offends them. In addition, users themselves are increasingly taking the opportunity to confront, engage, refute, and satirize speech that they deem to be negative or hateful.

This paper surveys the evolving landscape of abuse-fighting strategies across several online services and forums. It finds that rather than adopting a one-size-fits-all enforcement regime, different networks have evolved to handle abuse in different, frequently speech-protective ways. As a consequence of these diverse network-driven and user-driven strategies for minimizing abuse, users now have a wide range of choices not just for where and how to speak, but also for how to protect themselves and their speech online. Because networks and users alike have generally been free to experiment with new strategies, users now have more options to manage and structure their online experiences, along with more ways to speak their mind, respond to problematic content, and help one another out.

I. Introduction — Speech Laboratories

On the Internet, people have the ability to speak and be heard in more ways than ever before. The statistics speak for themselves, but it's worth reviewing these to get a sense for the immense scale of users, viewers, creators, and developers who are out there creating content on different networks.

Google Search has now crawled over 60 trillion URLs, including pages from over 230 million different domains. Over 100 billion searches take place each month, and every day 15% of web searches are brand new. Knowledge Graph, the part of Search that now provides quick answers and connections, contains over 50 billion facts about how 570 million things in the real world are connected.[1]

Beyond finding information, users can share their thoughts and experiences instantly with a wide audience, and are doing so in ever increasing numbers. There are over one billion unique monthly users on YouTube, and these users are uploading more than 100 hours of video to YouTube every minute.[2] 1.5 billion photos are uploaded every week to Google+.[3]

According to public statistics, five hundred million tweets are sent per day.[4] There are 757 million daily active users on Facebook, 81% of whom are outside the United States and Canada.[5] Approximately 10 million edits are made on Wikipedia each

[1] http://insidesearch.blogspot.com/2013/09/fifteen-years-onand-were-just-getting.html
[2] http://www.youtube.com/yt/press/statistics.html
[3] http://googleblog.blogspot.com/2013/10/google-hangouts-and-photos-save-some.html
[4] http://www.ted.com/talks/del_harvey_the_strangeness_of_scale_at_twitter
[5] http://newsroom.fb.com/company-info/

month,[6] and there are an average of 86 edits on each article.[7] 53 million reviews have been written on Yelp.[8] Reddit receives about 115 million unique visitors per month.[9]

These and other networks and forums have cultivated different userbases and promoted different kinds of user interests and methods for communicating. People have the ability to choose forums for expression based on numerous factors, including:

- vastness of audience;
- diversity of their fellow speakers;
- ability to interact directly with readers and commenters (and degree to which those interactions are antagonistic or peaceful);
- ability to provide context and link up individual images/stories to a broader discussion, community, or narrative;[10]
- whether one's speech is linked to an anonymous, pseudonymous, or real-world identity;
- degree to which an operator/moderator of a forum or network is able to promote or inhibit different kinds of speech.

Given the staggering amount of speech that is out there on this wide range of forums, this naturally raises the question: when speech turns nasty, or hateful, or dangerous, or pornographic, or spammy, what actions can users take to limit their exposure to abuse, to flag or report abuse, or to counter or challenge speech that concerns them? As the following section suggests, the answers to these questions depend on the nature of the

[6] http://stats.wikimedia.org/EN/TablesDatabaseEdits.htm
[7] http://stats.wikimedia.org/EN/TablesArticlesEditsPerArticle.htm
[8] http://www.yelp.com/about
[9] http://www.reddit.com/about/
[10] *See, e.g.*, http://youtube-global.blogspot.com/2013/02/context-is-king-share-your-story.html

network, the nature of the affordances given to users, and the character of the user communities that develop on different networks.

II. Safety Laboratories

As forums for speech have proliferated, so have strategies for addressing abusive or offensive speech. Information providers and intermediaries have implemented a wide range of systems for handling and addressing the presence of abusive content on their networks. At the same time, users themselves have taken many actions and initiated a variety of strategies to respond to harmful content online.

A. User Flags and Report Abuse Flows

Many networks rely heavily on users to alert them about abusive or problematic content via flags or other reporting mechanisms. Here's an example of how this works. You, the user, see a doctored photograph of a friend on the beach with an inappropriate word written across the photo. You believe that this image violates a service's or platform's content policies or community guidelines. You click the flag or drop-down arrow next to the photo, and you report the content to the network as "harassment or bullying."

When we receive one of these reports of potential abuse on a Google product like Blogger, Google+, or YouTube, we have teams around the world working twenty-four hours per day to review flagged content. Note that this is a very large-scale operation, particularly when you consider that tens of thousands of videos on YouTube (and similarly large amounts of content on other services) are flagged each day.

When we find that content violates our policies, we remove it promptly, but we also recognize strong exceptions in our policies for educational content, documentary/newsworthy content, scientific content, and artistic content.

Automated rules exist as well for some forms of content, such as massively distributed spam or malware, depending on the service.

However, for the really hard cases around hate speech, harassment, or violence in the context of political speech, and other low-volume but high-risk and often context-sensitive abuse types, we rely on users to flag this content and identify the reason why it might violate our policies. Then we assess these complaints individually and take appropriate action under our policies.

There are three primary challenges here: building scalable systems, understanding context, and recognizing that good policies shouldn't be overbroad or over-enforced.

The first challenge involves building scalable solutions across many hundreds of millions of users (and hundreds of languages), locating the really problematic needles within the haystack of otherwise awesome content that is out there, and trying to avoid removing valid content (false positives) while at the same time making sure all serious policy violations get quick and correct verdicts after we're notified about that content. Thousands of people at Google are working on solving these hard challenges around abuse and quality on different products, including building effective appeals systems for users to challenge our actions or remedy their policy violations.[11]

[11] *See, e.g.*, https://support.google.com/plus/answer/3100745, https://support.google.com/youtube/answer/185111, https://support.google.com/googleplay/android-developer/contact/appappeals

But it's not always easy to scale these solutions, given the second challenge around understanding context. Outside of the more obvious policy violations described above (pornography, spam, graphic violence, etc), it is sometimes difficult to understand why a given piece of content is objectionable or abusive for a given user. People are astoundingly different from one another, and even the most thoroughly trained and culturally aware reviewer might not be able (at first glance) to grasp the abusive or objectionable intent behind a given post, photo, or video. As a result, abuse reporting flows are frequently built to elicit additional information from users as to why a given piece of content is objectionable. This additional information (e.g., a flag or report that is specifically about harassment, rather than a spam flag) improves reviewers' ability to assess context.

Finally, there's a basic challenge in ensuring that policies strike a good balance between promoting free speech and protecting users from abusive speech. Different networks approach this question in different ways. For example, Blogger's policies include the following language: "We respect our users' ownership of and responsibility for the content they choose to share. It is our belief that censoring this content is contrary to a service that bases itself on freedom of expression. In order to uphold these values, we need to curb abuses that threaten our ability to provide this service and the freedom of expression it encourages."[12]

Different services identify different values when striking the balance between free expression and prohibited abuse. But one common thread within most policies is the recognition of a serious difference between (a) speech that attacks or threatens

[12] https://www.blogger.com/content.g?hl=en

someone and (b) speech that is merely negative or mean-spirited. Preserving that balance is one way that these and many other services and platforms avoid overbroad enforcement of their policies and protect user speech.

B. Empowering Users to Stop Abuse

Beyond abuse reporting, users also have a number of tools and options at their disposal on different networks to stop, avoid, or respond to abuse on their own. Empowering users to work out solutions to problems on their own can be a win for speech, and a win for increasing respect and decreasing abuse. Some of the tools that are available to users include:

- blocking tools (e.g., the "block" feature on Google+, which prevents someone who has been blocked from being able to comment on content or +mention the user who has initiated the block);[13]

- moderation tools (e.g., options that allow a YouTube video owner to remove comments from a given video or hide comments until they are approved for public view);[14]

- user-to-user communication, which can take place directly between users, or can be facilitated by an intermediary or moderator (e.g., via tools for users to reach out to trusted third parties about upsetting content);

- user involvement in counterspeech and counter-messaging (discussed in more detail below);

- community dispute resolution and moderation;[15]

- solicitation of user preferences regarding racy or sensitive content. For example, Safety Mode helps filter out potentially objectionable content on

[13] https://support.google.com/plus/answer/1047934
[14] https://support.google.com/youtube/answer/111870
[15] http://en.m.wikipedia.org/wiki/Wikipedia:DR

YouTube,[16] while SafeSearch does the same for sexually explicit content that you might encounter through Google search results.[17]

- tools and settings that developers and uploaders can use to rate or assess their content along different metrics for violence, adult themes, or sexual content.[18]

The number of tools and strategies identified above indicates that there is no one-size-fits-all solution when it comes to user options and affordances for stopping abuse. Instead, we have seen the evolution of a complex and diverse landscape of abuse-fighting strategies, including the more experimental and user-generated strategies described below.

C. User-to-User Resolution

For certain forms of low-level abuse, annoying posts, or unflattering photos, an effective response sometimes involves simply reaching out to the person who posted that content and asking them to take the content down. Early evidence from this form of intervention has revealed some promising increases in action rates.

Different networks have enabled user-to-user communication in different ways. YouTube offers a process that encourages users to express their concerns about a given video directly to the video's uploader.[19] Facebook's "social reporting" flow relies heavily on providing templates for users to contact one another and request removal of content. Google+ offers users who are reporting abuse on a photo the option to send a

[16] https://support.google.com/youtube/answer/174084
[17] https://support.google.com/websearch/answer/510
[18] *See, e.g.*, https://support.google.com/googleplay/answer/1075738
[19] https://support.google.com/youtube/answer/142768

private post to the photo uploader, and prepopulates the message box with the following language: “Hi [name], I don't like this photo. Would you mind removing it?”[20]

These low-impact communications can help address the range of potentially problematic content out there that doesn’t fit neatly into any particular policy. The impact of a concerned statement or outreach information can be much more immediate when it comes from a friend, rather than a distant moderator or network administrator. Other positive impacts include the increased sense of community ownership and mutual support that can come when users communicate directly with one another and work out solutions on their own.

Of course, direct communication between users isn’t the right solution in all cases, particularly if there’s already serious harassment taking place between these users or other content present that would be taken down for policy violations. But in situations involving low-level forms of abuse or annoyance, user-to-user communication can be a good “ground-up” way to resolve problems and disputes.

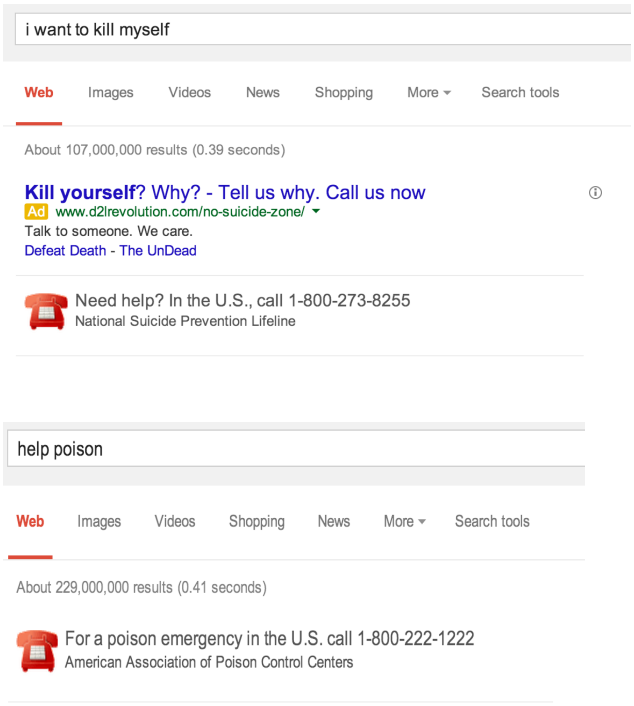
D. More experimental examples

It’s outside the scope of this paper to cover every new anti-abuse strategy evolving on different networks, or to predict what will emerge in the future. But it’s worth pointing quickly to a few final examples of interesting models for making networks safer and more engaging to users.

[20] https://support.google.com/plus/answer/1047388

For users who are concerned about harassment, Wikipedia offers a wide range of informal dispute resolution strategies,[21] many of which require users to identify the specific diffs that are the subject of their dispute.[22] When dispute resolution fails, Wikipedia also offers more formal mediation processes.[23]

Anti-abuse strategies aren't just about enabling users to report or respond to abusive content. Sometimes, the purpose of these strategies is to help users obtain useful information and resources during a difficult time. For example, Google has integrated resources for suicide, self-harm, and poison control hotlines into search results for certain terms:



[21] http://en.m.wikipedia.org/wiki/Wikipedia:DR
[22] http://en.m.wikipedia.org/wiki/Help:Diff
[23] http://en.m.wikipedia.org/wiki/Wikipedia:Requests_for_mediation

One could imagine similar strategies that might be deployed to enable connectivity between users who are in a position to help one another out, or that might allow one concerned user to provide useful feedback about a potentially problematic post from another user.

Finally, beyond all the tools that different services and platforms have provided to help their communities of users moderate their experience, it's important to recognize that these communities also frequently evolve and create their own democratic principles. For example, there have been many powerful examples of people simply choosing to stand up and refute—through facts, satire, or counter-argument—content that offends them. Attempts to expose and challenge controversial or hateful sentiments range from large-scale initiatives and responses such as the "It Gets Better" campaign (which inspired over 50,000 video uploads)[24] and the Honey Maid "Love" video,[25] to smaller-scale response videos and campaigns.

These examples help demonstrate that sweeping negative messages under the rug isn't always the right solution. Furthermore, the process of engaging with people who hold divergent views (and/or attempting to refute those views) can also help communicate something even more essential about democracy: that we live in a society where people can openly debate ideas, even when offensive, even if no consensus is reached. Counterspeech demonstrates that we can withstand hateful views in order to debunk them, rather than be compelled to censor them in order to avoid offense.

[24] http://www.itgetsbetter.org/pages/about-it-gets-better-project/
[25] https://www.youtube.com/watch?v=cBC-pRFt9OM

III. Conclusion

The strategies identified above represent a subset of the anti-abuse strategies that have emerged on different networks. Rather than being subject to a universal or one-size-fits-all enforcement framework, these networks have been free to experiment with flexible, community-focused, and often speech-protective solutions for minimizing abuse. As a result of this laboratory-based approach, we have seen the development of a wide range of tools and strategies that make it simpler, safer, and more effective for a given user to speak their mind and respond to problematic content.