Once upon a time, we were disconnected online.
So, we formed social computing systems and online communities to connect us to each other.
We met new friends, created net culture, and shared our ideas with the world. Life was happy.
But then, anti-social behavior arose. It grew and grew until it threatened to destroy the social computing systems.
The people were trolled and flamed. Their communities fractured. What could help?
Today: anti-social computing

How can a community manage anti-social behavior?

Premeditated anti-social behavior: trolling
Non-premeditated anti-social behavior: flaming
The darkest of the dark: beyond trolling
Last time: growing pains

Communities can't maintain the same design as they grow. Newcomers change the dynamics, even if they absorb the norms—and oftentimes they don’t absorb the norms.

Growth begets contention and rulemaking, which can push off newcomers

Ranking: a common approach to help manage attention as the content in the system grows
Premeditated anti-social behavior

a.k.a., trolling 😈
How Trolls Are Ruining the Internet

When Will the Internet Be Safe for Women?

FAKE NEWS IS ABOUT TO GET EVEN SCARIER THAN YOU EVER DREAMED

Examples via Justin Cheng [Time 2016; The Atlantic 2016; Vanity Fair 2017]
47% of online users have been harassed

Examples via Justin Cheng [Data & Society 2017]
WHY WE'RE SHUTTING OFF OUR COMMENTS

We're turning comments off for a bit

Sick of Internet comments? Us, too - here's what we're doing about it

Examples via Justin Cheng [Popular Science 2013; The Verge 2015; Chicago Sun-Times 2014]
‘Zoombombing’ Becomes a Dangerous Organized Effort

Zoom, the videoconferencing app, has become a target for harassment and abuse coordinated in private off-platform chats.
What is trolling?

Intentional disruption of an online community [Schwarz, NYT 2008]

Behavior that falls outside the acceptable bounds of the community [Binns 2012; Hardaker 2010]

People who habitually engage in trolling are known as trolls, as in the grumpy monsters who hide under bridges.
Is trolling worse online?

It’s certainly more well-publicized. There are two reasons we might run into it more online than offline:

(1) Scale: a single troll can impact many communities, or a single highly visible community, in ways that their reach would otherwise be limited.

(2) People troll more online than they do offline.

Are these true? [2min]
Online disinhibition effect

[Suler 2004]

When we interact online, we say and do things that we would not do offline and in-person. We self-disclose more, and we act out more.

This is known as the online disinhibition effect: we have less inhibition when online.

Online disinhibition would imply that we do troll more online than offline.
Dealing with disinhibition

Reasons include:

Anonymity: dissociation from my real identity, so fewer consequences

Few social cues: no facial expressions, reactions, etc.; a socially opaque system

Asynchronicity: conversations never cool off

So, design interventions might be:

Re-individuate by associating actions with an identity that I care about

Re-introduce social cues: e.g., when I reply to a mean comment, they often soften up

Take it offline: don’t try to manage fights online if possible
Blocking

If a few people are responsible for the majority of the deleterious content, then blocking them should silence most of the negative behavior.

However, blocking just becomes whack-a-mole if it’s easy for participants to create another account. So, this only works if identities are expensive to create.
Why do trolls troll?

One argument: trolls are born that way.

Inveterate trolls do, on average, register strong personality dispositions such as high self-report scores in three of the four Dark Tetrad of personality traits: sadism, psychopathy, narcissism, and Machiavellianism [Buckels, Trapnell, and Paulhus 2014]

Reasons given range from boredom [Varjas et al. 2010], to doing it for fun [Shachaf and Hara 2010], to venting [Lee and Kim 2015].
How much do trolls troll?

Proportion of Banned Users vs. Proportion of Deleted Posts

CNN.com data: some trolls will troll almost exclusively.
How much do trolls troll?

CNN.com data: some trolls will troll almost exclusively… and some only rarely.

Why? [2min]
How much do trolls troll?

CNN.com data: some trolls will troll almost exclusively… and some only rarely.

Why? [2min]
Unpremeditated anti-social behavior

a.k.a., flaming 🔥
What is flaming?

Flaming: uninhibited hostile behavior directed at another person or group [Kayany 1998, Kiesler 1986]

Common examples: swearing, calling names, ridiculing, insulting

While trolling usually refers to someone intentionally riling people up, flaming usually refers to someone who lost self-control.
Flaming and disinhibition

Online disinhibition would also predict that we flame more online than in person. We are more likely to act out, others are more likely to act out in response, and so on: a flame war.

But does the environment suppress or amplify the amount of trolling that people engage in?
“Well, that escalated quickly.”

Human observers can guess with 72% accuracy from just the first post on a Wikipedia talk page whether the conversation is going to turn toxic. [Zhang et al. 2018]

We are not good at predicting how others will read our comments. [Chang et al. 2020]

People are more likely to perceive opinions than to intend them when they write the comment.

If I intended to share a fact but you perceived me as sharing an opinion, the conversation is more likely to derail into a flamefest.
Recall: environment matters

I’m Voting for Hillary Because of My Daughter
Back in the 2008 primary season, I supported Hillary Clinton. That choice...

Top Comments Sorted by Best
User1337 · 2 hours ago
I’m a woman, and I don’t think you should vote for a woman just because she is a woman. Vote for her because you believe she deserves it.
6 ▲ | ▼ · Reply

User9054 · 3 hours ago
Personally, I’d vote for whoever I think is the best and...

Top Comments Sorted by Best
User1337 · 2 hours ago
Oh yes. By all means, vote for a Wall Street sellout - - a lying, abuse-enabling, soon-to-be felon as our next President. And do it for your daughter. You’re quite the role model.
1 ▲ | ▼ · Reply

User9054 · 3 hours ago
Hillary is a cunt. I am voting with my dick for Putin. (s

Positive comments
Result: 35% troll comments

[Cheng et al. 2017]

Negative comments
Result: 47% troll comments
(Relative increase of one third compared to the 35% baseline)
The environment is something the designer has control of. But people also bring their own affective state to a social computing system.

Being in a bad mood reduces self-regulation [Leith and Baumeister 1996] and results in less favorable impressions of others [Forgas and Bower 1987].
Antisocial behavior tracks human diurnal mood patterns

Proportion of flagged posts on CNN.com

Daily negative affect

[Gold & Macy 2011]
Mood influences behavior [Cheng et al. 2017]

Placed in a good mood by doing well on an easy test
Result: 35% troll comments

Placed in a negative mood by doing poorly on a difficult test
Result: 49% troll comments
(Same effect as seeing troll comments!)
The effects compound.
Can antisocial behavior spiral?

Do people improve their behavior after community feedback?

[Cheng, Danescu-Niculescu-Mizil, Leskovec 2014]

Five upvotes, four downvotes

Positively evaluated >=90% upvotes

Negatively evaluated <=20% upvotes
Can antisocial behavior

[Cheng, Danescu-Niculescu-Mizil, Leskovec 2014]

Posts from before and upvoted/downvoted posts are matched via a trained comment quality classifier (via propensity score matching). So the users and critical comments are similar, only the votes are different.
Can antisocial behavior

[Cheng, Danescu-Niculescu-Mizil, Leskovec 2014]

After a negative evaluation, community members…

Post worse content going forward

Post more frequently

Evaluate others more negatively ↓

A self-fulfilling prophecy!
Internet mobs

Accusations can cascade, and without due process, the wrong people can be harmed.

Ex: Reddit misidentified Sunil Tripathi as the Boston Marathon Bomber

Misinformation spreads: the Boston Bomber rumors were corrected, but the corrections spread too slowly. [Starbird et al. 2014]
Bill Gates Is Targeted by Social-Media Mobs

Coronavirus crisis intensifies conspiracy theories and online harassment of Microsoft co-founder
Does trolling worsen over time?

![Graph showing the mean proportion of flagged posts and users over time from December to August. The graph indicates a trend of increasing flagged posts and users over the months.](CNN.com)

[Cheng et al. 2017]
Design opportunities
What does the design encourage?

Systems that reward short-term engagement are likely to produce snark and flame, since these activities are the most likely ones to raise an affective response.

[Image from Niloufar Salehi]
Many independent signals can combine to create a hostile or negative environment

[https://stackoverflow.blog/2019/07/18/building-community-inclusivity-stack-overflow/]
Early detection of off-the-rails conversations

[Chang and Danescu-Niculescu-Mizil 2019]

Two threads on the Wikipedia discussion for the Dyadlov Pass incident:

**A1:** Why there’s no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources some require it wouldn’t even come close, not to mention having some really weak points, but it doesn’t mean that it doesn’t exist.

**B1:** Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I’m going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don’t think this article should rely on one so-so source.

**A2:** So what you’re saying is we should put a bad source in the article because it exists?

**B2:** I would assume that it’s as reliable as any other mainstream news source.
Reducing flaming

Assuming the environment and norms aren’t changeable, then one possible mechanism is to manage mood. Moods pass, so consider a cool-off period before posting a flame post.

Example from Instagram

Or: “we will ask you again in 20 minutes if you really want to post this”
Blockbots

[Geiger 2016]

Community-maintained block lists: harassers can get added to the blocklist, then are automatically blocked from any user's account that subscribes to the blocklist.

Details

Block Together is designed to reduce the burden of blocking when many accounts are attacking you, or when a few accounts are attacking many people in your community. It uses the Twitter API. If you choose to share your list of blocks, your friends can subscribe to your list so that when you block an account, they block that account automatically. You can also use Block Together without sharing your blocks with anyone.

Protections from incorrect blocks

Block Together will never block an account you follow, even if it meets one of the above criteria.

Propagating forgiveness

If you subscribe to a block list, you will block every account currently on that list. Later, if the author of that block list blocks more accounts, you will automatically block those as well. If the author of the list unblocks an account, you will automatically unblock it too, but only if you originally blocked it through the shared list. If you blocked the account independently, it will not be unblocked.
Rather than defaults that focus on repair after the fact, what if designs aimed for affirmative consent?
Dealing with norm breakers

Imagine you were a Stanford admin/RF/RA/etc. and it was brought to your attention that someone was engaging in substantial amounts of antisocial behavior in your community’s online spaces.

What would you do? [1 min]
Face-saving

[Kiesler et al. 2012]

People self-regulate if they can do so without having to admit that they deliberately violated norms. MIT’s warning email to students:

Someone using your account did [whatever the offense is]. Account holders are responsible for the use of their accounts. If you were unaware that your account was being used in this way, it may have been compromised. User Accounts can help you change your password and re-secure your account.

Many would change their password and the practice would stop, even if MIT knew from eyewitnesses that they had done it.

Calling them out instead prompted people to assert the behavior as within their rights and continue doing it to challenge authority.
The darkest of the dark

Content warning: rape, bullying, doxing, revenge porn, partner abuse
Anti-social behavior gets (even more) personal

We typically think of anti-social behavior as hurling insults at each other, but this behavior escalates.

In this final section, we’ll survey some of the most troubling behaviors and what we know about them. You may already have heard of some of these behaviors.
In LambdaMOO, a text-based online spaces (a la a textual MMO), a character named Mr. Bungle developed a piece of software that allowed him to command other characters to perform actions.

He then forced two other avatars to perform sexual acts on him and on each other, and to violate their own bodies.

Mr. Bungle was eventually banned from the server after a community meeting, but the damage had been done.
Cyberbullying

Adolescent bully on social media — for example in public or private messages — are particularly problematic because they follow you home, not just at school. [Li 2006]

Women and LGBTQ are more likely to be victims, and perpetrators are more likely to be male. [Aboujaoude et al. 2015]

Being cyberbullied increases suicidal ideation and the probability of attempting suicide. [Hinduja and Patchin 2010]
Cyberbullying

Designs typically focus on blocking, but this doesn’t erase the issues

Example from Instagram

Are you having a problem with divdivk?

- Protect yourself from unwanted interactions without them knowing.
- Only you and divdivk will be able to see their new comments.
- They won’t see when you’re online or when you’ve read their messages.

Restrict Account
Revenge porn

Revenge porn is the nonconsensual distribution of sexual photos of a person.

Typically, the “revenge” in revenge porn means that the pair used to be partners and the photos were initially shared consensually, but a breakup or other event prompted one party to release the other person’s sexual photos to hurt the other party.
Doxing

Doxing refers to releasing the personal information (e.g., address, name, photo) of an individual online against their will. The term originated from the idea of sharing “docs”, or documents, of someone.

One shared, the information cannot easily be taken back.
Intimate partner violence

[Freed et al. 2018]

Abusers in intimate partner violence utilize technology to intimidate, monitor, impersonate, and harass. Often, the victims are married to their abusers and share social networks and physical space.

- Owning the device or paying for the family plan, threatening to remove it
- Installing or authorizing software to track the victim (e.g., Find My Friends)
- Forcing victims to disclose social media passwords, monitoring messages

Generally, security approaches are not designed to combat attackers who know the victim intimately.
What do we do?

There is no permanent solution here. New behaviors arise over time. People manipulate the system to achieve their goals.

Step one: ensure that there are serious consequences for this kind of behavior — possibly legal ones.

Step two: find confidential and trusted means for users to report abuse. Develop a trusted adjudication system.
Summary

Anti-social behavior is a fact of life in social computing systems. Trolling is purposeful; flaming may be due to a momentary lack of self-control.

The environment and mood can influence a user’s propensity to engage in anti-social behavior: but (nearly) anybody, given the wrong circumstances, can transform into a troll.

Changing the environment, allowing mood to pass, and allowing face-saving can help reduce anti-social behavior.

Dark behavior exists: be prepared to respond.
Social Computing
CS 278 | Stanford University | Michael Bernstein

Creative Commons images thanks to Kamau Akabueze, Eric Parker, Chris Goldberg, Dick Vos, Wikimedia, MaxPixel.net, Mescon, and Andrew Taylor.

Slide content shareable under a Creative Commons Attribution-NonCommercial 4.0 International License.