Can Anyone Become an Internet Troll?

Jure Leskovec (@jure)

Including joint work with Justin Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil
Two Metaphors for the Web

Web as a library

Web as a social place
Web as a Social Place

Crucial for the success of the Web is the ability of users to contribute content and provide feedback.
We worried that Byrd was gonna get lonely if we didn't set up a cat cam, this is what we saw.

Our new kitten has officially stolen my heart.

Posted about Manfred earlier this week. He's super cute.

Reddit, meet Hercules. This is what he does in the bathroom.

My Best Little Buddy. :)

Original Content
Stack Overflow is a question and answer site for professional and enthusiast programmers. Registration required.

Can C++ code be valid in both C++03 and C++11 but

Is it possible for C++ code to conform to both the C++03 standard and the C++11 standard? Different things depending on under which standard it is being compiled?

13 I'm pretty sure the `auto` keyword could result in a situation like this – OMGtechy Apr 13 at 19:15

6 Yes. One example is `>>` when used in a template. You can come up with a situation where it is valid for both standards. Another one that I'm sure would be easy to find changes for is `free` Apr 13 at 19:15

4 Here's a nice article on the `>>` situation: gustedt.wordpress.com/2013/12/15/… –

5 @OMGtechy: I don't think `auto` can cause this. With the old meaning, an `auto` type name; with the new meaning, a type name is not permitted. – Keith Thompson

2 Closing this as too broad would be a shame, this is question likely to be asked many times. It would be good to have a solid canonical question to point to that is not closed. – Shafik Yassine

7 Answers
Facebook privacy now defaults to friends only

By Doug Gross, CNN
updated 3:39 PM EDT, Thu May 22, 2014 | Filed under: Social Media

Hi Charlie! Before you post, we wanted to make sure you’re aware of how your posts impact your audience.

On Thursday Facebook bowed to pressure and announced changes to its privacy settings. The changes aren’t revolutionary, but they will make it harder for people to post information without their friends knowing about it.

Facebook users will still be able to post updates to their timelines, but those updates will default to “friends” instead of “friends and friends of friends.”

Facebook users will be able to toggle between the two settings, but the new default will be “friends.”

This change is part of a larger effort to make Facebook more transparent and user-friendly. Facebook has also announced changes to its news feed, which will make it easier for users to see posts from friends and family.

It’s unclear how these changes will affect Facebook’s bottom line, but they are likely to be well-received by users who have been concerned about their privacy.

In other news, Facebook has also announced a new feature that allows users to see what content they have shared with their friends in the past. This feature could be useful for users who want to keep track of what they have posted online.

Overall, these changes are likely to make Facebook a more user-friendly platform. Facebook users will have more control over what they share, and the changes will make it easier for them to see what their friends are posting.

Facebook is a social networking platform that allows users to connect with friends and family online. It was founded in 2004 and has become one of the most popular social media platforms in the world.

Facebook currently has over 2.3 billion monthly active users, and it is used by people of all ages and backgrounds. It is a popular platform for sharing photos, videos, and other content, and it has become an important tool for businesses and brands.

In recent years, Facebook has faced criticism for its handling of user data. In 2018, the Cambridge Analytica scandal revealed that Facebook had shared user data with third-party apps without the users’ knowledge.

Facebook has since made changes to its privacy settings, and it has announced new features to improve user privacy. These changes are likely to help Facebook maintain its position as one of the most popular social media platforms in the world.
Štirje predsedniki menijo, da je Slovenija na pravi poti

Slovenija mora biti vključujoča družba

19. decembra 2016 ob 13:45
Ljubljana - MMC RTV SLO/STA

Predsedniki republike, vlade, državnega zbora in državnega sveta, Borut Pahor, Miro Cerar, Milan Brglez in Mitja Bervar, se strinjajo, da je Slovenija na pravi poti in da razlogi za utemeljen optimizem so.

Stabilnost je treba izkoristiti za sprejetje potrebnih reform, krepiti pa je treba tudi zaupanje ljudi v delovanje političnih institucij, so menili na srečanju štirih predsednikov.

Tako so po Pahorjevih besedah danes ugotovili, da Slovenija po letih surove gospodarske in finančne krize okreira, na kar kažejo tudi glavni gospodarski in socialni kazalniki. Politično stabilnost je po njegovih besedah treba izkoristiti za sprejetje reform, ki so potrebne, da bi Slovenija dodatno okrepla svojo gospodarsko in socialno moč ter da bi država postala odpomejša ob morebitni novi zaostriti razmer v mednarodnem okolju.
Two Metaphors for the Web

Web as a library
- One static version
- Everyone’s experience is fixed
- No user interactions

Web as a social place
- Complex dynamic feedback effects
- One user’s experience is a function of other users’ experiences
- User interactions
Challenges

Want users to engage and contribute (good) content (and make others’ experiences better)

- How do we keep users engaged?
- How do we promote good behavior?
- How do we ensure health of the community and the content?
Challenges

- We know a lot about overall average user behavior in online communities.
- But many times only a few users can have a disproportionate (negative) effect on the entire community.
- How do we study and model behavior of such users?
- Have do we use the insights to improve online communities?
Antisocial Behavior

Internet Troll is a person who sows discord on the Internet

- by starting arguments and upsetting people
- by posting inflammatory, extraneous, or off-topic messages in an online community

[wikipedia]
Facebook: The encyclopedia of beauty?

By Amanda Enayati, Special to CNN

Updated 9:13 PM ET, Fri March 16, 2012

Demi Lovato and Miley Cyrus have used social media to send positive messages about body image.

Story highlights

Social networks have young women competing against each other for the best body

Networks influence self-esteem, although experts disagree on whether its for better or worse

Heaviest users of media are the ones who have the most distorted beliefs about society

One day last year, Amanda Coleman decided to quit Facebook.

It wasn’t one precipitating event that led her to her decision, but a slow build -- a series of disturbing, sometimes anguished conversations.

Coleman, a college student and the president of her sorority, found herself spending a lot of time counseling young girls, many of them freshmen at the university.

Perceptions of Beauty

Dieting companies target men

Men face the issue at the same rate as women, but until recently, ads for weight loss products mostly featured female faces and voices.

Facebook: Encyclopedia of beauty?

Because of social networks, the field of competition has expanded dramatically. Now you’re competing with the best pictures and the ebullient status updates of every girl you know.

Fat: New ugly on the playground

Children as young as 3 worry about being fat. An obsession with skinny ideals is trickle down to our youngest generation.

Sex, lies and media

There’s a new wave of activists challenging our society’s notions of beauty.
Girls are focused on their image way too much. This isn't new. And when they talk about other girls 'fat', 'ugly', and 'b*tch' are recurring words.
It also shows that Islam and Christianity teaching women to dress modest could be right after all.
It also shows that Islam and Christianity teaching women to dress modest could be right after all.

Religious nut alert

Clearly that is the only logical conclusion to this article. Now if you'll excuse me, I need to iron my tarp. I have work on Monday, and I want to appear 'modest'.

fail at life. go bomb yourself.
Both of these little skanks are ugly to the bone.
Both of these little skanks are ugly to the bone.

Troll
**PROBLEM:** Trolls disrupt online discussions

Antisocial Behavior in Online Discussion Communities
AAAI ICWSM, 2015.
We examine posts of users who will be blocked from the community in the future.

What do we learn?
Facebook: The encyclopedia of beauty?

Facebook is for losers with no friends in real life.

Less positive
Soccer star Muamba still 'critical' after on-pitch collapse

Suck a f****ng c@ck u d@uchebag

More profanity
Ex-NBA star returns to inner city, brings hoop dreams

*sniff* This had me tearing up in the office. Great job Penny! You were always a class act on and off the court. Your Grandmother is applauding you again in heaven!

Fantastic story. Kudos to Penny Hardaway. This is what you call giving back.

Penny, once again you show why you are one of the best of the league. Always a class-act. Always a good example.
Penny, once again you show why you are one of the best of the league. Always a class-act. Always a good example.

I have a question for CNN.......why as a reader do I not see any articles similar to this about white NBA basketball players?......every single touchy feely story is about a black ball player and then all you do is blast Tim Tebow.........YOU GUYS MAKE ME SICK AS A READER !

Off-topic: Less similar to previous posts

How many white NBA players grew up
many stories of charitable efforts by white NBA players, but this isn't a story about a Black NBA player...it's a story about someone who came back to their roots to contribute.

b\c young black men need to see examples from their own race. They need to see role models who look like them. It allows for mentorship and guidance.

A story about a millionaire helping kids in his poor neighborhood personally, and being a positive role model makes you sick as a reader. Gotta love conservatives.
If you claim you want less government but want to control the bedroom, you're a Republican; If you want to cut Education, you're a Republican; If you want to cut Medicare and Medicaid, you're a Republican; If you want to cut Social Security, you're a Republican.

If you want to cut Medicare and Medicaid, you're a Republican; If you want to cut Social Security, you're a Republican; If you want to cut the bedroom, you're a Republican; If you want to cut Social Security, you're a Republican.
Discussion Communities

Complete data over 18 months,
~1.7 million users generated
~40 million posts and
~100 million votes
### How Common is Bad Behavior?

<table>
<thead>
<tr>
<th></th>
<th>CNN</th>
<th>IGN</th>
<th>BREITBART</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Post Deletion Rate</strong></td>
<td>21.4%</td>
<td>2.3%</td>
<td>2.7%</td>
</tr>
<tr>
<td><strong>User Ban Rate</strong></td>
<td>3.3%</td>
<td>1.7%</td>
<td>2.2%</td>
</tr>
</tbody>
</table>
This Talk: Trolling

- Large-scale investigation of trolling in online discussion communities

- Common theme: Complementing observational data analysis with human experiments
This Talk: Questions

1) Who are trolls?
   - Mechanisms behind trolling behaviors: Mood and discussion context

2) Does feedback mitigate trolling?
   - Votes lead to cascading of bad behaviors

3) Can we detect trolling early?
   - Predicting whether a given user will be blocked from the community
Who are these people?
What makes them troll?
Why is trolling so common?

Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions
Confessions of a former internet troll

by Emmett Rensin on December 16, 2014

Internet Trolls Really Are Horrible People

REDDIT WANTS TO EXILE TROLLS. BUT GROWING UP IS HARD

Behind the online comments: the psychology of internet trolls
Studies on Trolling

Prior research suggests:

- Trolls are born and not made
  [Baker ‘01; Herring et al. ‘11; Shachaf-Hara ‘10]
- Trolls have particular personality
  [Buckels et al. ’14; Raine ‘02]
- Trolls are sociopathic individuals
  [Rensin ’14; Schwartz ‘08]
- Q: Are trolls really special people or can anyone become a troll?
Our Work:
Can Anyone Become a Troll?
Hypotheses: Mood & Context

H1: Negative mood increases a user’s likelihood of trolling

H2: The discussion context affects a user’s susceptibility to troll.
Designing an Experiment

MTukers only care about being paid.

How do we design an experiment?
Designing an Experiment

Two-by-two experimental design:

- **Step 1: Mood – Take a test (5min)**
  - Easy test & tell people they did well
  - Hard test & tell people they did poorly

- **Step 2: Context – Be in a discussion**
  - Neutral context of neutral comments
  - Negative context of negative comments

People do not know the real purpose. They are debriefed at the end.
Qualification Test

Instructions

- Below is a series of simple questions, many of which you should be able to answer correctly.
- You will have **five minutes** to complete all the questions.
- Currently, average performance is **8 or more correct answers**.
- You are allowed to use pen and paper, but not any electronic aids (including the Internet).
- Your performance on this task will not affect your payment on the task.

Unscramble the following letters to form an English word:
"P A P H Y"
Type in your answer.

Subtract three thousand from five thousand. Write your answer in words.
Type in your answer.

How many 'l's are there in these sentences?

266.5 seconds left
Controlling Mood: NEGMOOD

Qualification Test

Instructions

- Below is a series of simple questions, many of which you should be able to answer correctly.
- You will have **five minutes** to complete all the questions.
- Currently, average performance is **8 or more correct answers**.
- You are allowed to use pen and paper, but not any electronic aids (including the Internet).
- Your performance on this task will not affect your payment on the task.

Unscramble the following letters to form an English word:
"D E A N Y O N"
Type in your answer.

Subtract six thousand, seven hundred eighty-three from eight thousand, eleven. Write your answer in words.
Type in your answer.

How many 'l's are there in these sentences?

266.5 seconds left
News of the Day

I’m Voting for Hillary Because of My Daughter

Back in the 2008 primary season, I supported Hillary Clinton. That choice wasn’t easy for me, especially as Barack Obama advanced and became ever more compelling as a candidate. I felt spoiled as the year progressed: two great,

Top Comments Sorted by Best

User1337 · 14 weeks 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 · 14 weeks ago
Personally, I’d vote for whoever I think is the best and most qualified candidate, but that’s just me.
1 ▲ ▼
News of the Day

I’m Voting for Hillary Because of My Daughter

Back in the 2008 primary season, I supported Hillary Clinton. That choice wasn’t easy for me, especially as Barack Obama advanced and became ever more compelling as a candidate. I felt spoiled as the year progressed: two great,

Top Comments Sorted by Best

User1337 · 9 weeks 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 · 9 weeks ago
Hillary is a [redacted]. I am voting with my [redacted] for Putin. /s

0 ▲ | ▼
Mechanical Turk Experiment

667 participants, 40% female, age 34.2,
54% Democrat, 25% Moderate, 21% Republican
Contributed 791 posts (37.8 words/post), 1392 votes
Results: Test Performance

- **PosMOOD (easy test):**
  11 out of 15 questions correct
  - Outperforming stated avg. score of 8

- **NEGMOOD (hard test):**
  2 out of 15 questions correct
  - After below stated avg. score of 8

Does the test score affect user’s mood?
Results: Does it affect mood?

- After the test: 65 questions on how people were feeling
  - Based on the Profile of Mood States

- People in NEGMOOD felt worse:
  - Higher anger, confusion, depression, fatigue, and tension scores

- How good were the comments?
### Comments: MOOD

<table>
<thead>
<tr>
<th></th>
<th>Prop. Troll Posts</th>
<th>POSMood</th>
<th>NEGMood</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NeutralContext</strong></td>
<td>0.35</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td><strong>NegContext</strong></td>
<td>0.47</td>
<td>0.68</td>
<td></td>
</tr>
</tbody>
</table>

- Users in the negative mood condition are more likely to troll (~80% higher odds, p < 0.05)
Users in the negative context condition are more likely to troll (~60% higher odds, p < 0.05)
Comments: Negativity

Negative Affect (LIWC)

<table>
<thead>
<tr>
<th>NeutralContext</th>
<th>PosMood</th>
<th>NegMood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.011</td>
<td>0.014</td>
</tr>
<tr>
<td>NegContext</td>
<td>0.023</td>
<td>0.029</td>
</tr>
</tbody>
</table>

- Negative mood and context lead to more negative posts
Bad mood and context make people write bad posts.

How do we combat that?
Sites allow users to vote on the content and hopefully bad posts get voted down
We worried that Byrd was gonna get lonely, so we set up a cat cam, this is what we saw

Our new kitten has officially stolen my heart.

Posted about Manfred earlier this week. He's really friendly.

Reddit, meet Hercules. This is what he does in the bathroom.

My Best Little Buddy. :)
Can C++ code be valid in both C++03 and C++11 but

Is it possible for C++ code to conform to both the C++03 standard and the C++11 standard, and behave differently depending on under which standard it is being compiled?

I'm pretty sure `auto` could result in a situation like this – OMGtechy Apr 13 at 19:15

Yes. One example is `>>` when used in a template. You can come up with a situation where it works for both standards. Another example that I'm sure would be easy to find changes for is `while` Apr 13 at 19:15

Here's a nice article on the `>` situation: gustedt.wordpress.com/2013/12/15/… –

@OMGtechy: I don't think `auto` can cause this. With the old meaning, an `auto` type name; with the new meaning, a type name is not permitted. – Keith Thompson

Closing this as too broad would be a shame, this is question likely to be asked many times and good to have a solid canonical question to point to that is not closed. – Shafik Y
Tom · 7 hours ago
If you're posting something to Facebook, it shouldn't be anything you wouldn't print and tape on the front door of a local grocery store.

ccw101 · Tom · 7 hours ago
I hate Facebook for the fact the only person you have control over is yourself. I have seen full grown adults get angry at their own children and rip them a new one on their Facebook home page!
If adults can be so ST**id then what do kids do?
Facebook is scary. And has given people the opportunity to use it to cause home break in's, ruined reputations, fights, suicides etc.

IAMNOTATROLL · ccw101 · 7 hours ago
Come now, I thoroughly enjoy watching my in-laws publicly argue and shred each other to pieces over Facebook.

Furby · IAMNOTATROLL · 6 hours ago
I had a distant cousin try to blackmail my mom on FB publicly. Me and her became real close after that - and not in the way you want to get close to someone. Some people are just plain dumb
When a user evaluates a comment, she is indirectly evaluating the author.

Do such evaluations help authors contribute better content?

How Community Feedback Shapes User Behavior
J. Cheng, C. Danescu-Niculescu-Mizil, J. Leskovec
AAAI ICWSM, 2014
Positive Vote

Negative Vote
Evaluation of a Post

- A post is **positively/negatively** evaluated when the proportion of up-votes is higher/lower than a given threshold.

\[ \frac{P}{P+N} = \frac{9}{9+1} = 0.9 \geq \text{High Threshold} \]
defined by the proportion of up-votes

How does positive/negative feedback influence subsequent behavior?
Effects of Evaluations

- Post quality
  - How well you write
- Community bias
  - How people evaluate you
- Posting frequency
  - How regularly you post
- Voting behavior
  - How you vote on others
Do Users Improve?

Operant conditioning predicts that feedback would guide authors towards better behavior:

- up-votes are “reward” stimuli
- down-votes are “punishment” stimuli

Or Do They Get Worse?

Feedback can have negative effects:

- People given only positive feedback tend to become complacent
- Also, bad impressions are quicker to form and more resistant to disconfirmation

Matching Similar Users

Compare pairs of users who were evaluated differently on similar content.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects.
Details of Matching

Match pairs of users where one got positive and one got negative evaluation

- Match based on
  - history text quality
  - number of posts
  - post length
  - overall proportion of up-votes

Text quality determined by training a machine learning model using text features, validated using crowd workers.
Results: Neg-Votes Increase

Before

After

Positive evaluation

Negative evaluation

Proportion of up-votes

Results:

Neg - Votes

Increase

63
Content vs. Community

Can valuations be explained by textual or community effects?

Down-voting because of post content

I hate …

Down-voting because the community dislikes the author

We don’t like you.
Results: Negativity Bias

- Text quality drops significantly after a negative evaluation but does not change after a positive evaluation.

$p < 0.05$ in all communities.

To learn more about these types of effects, see Kanouse, D. E., & Hanson Jr, L. R. (1987). Negativity in evaluations.
Evaluations can Affect…

…Community Bias

(How people perceive you)

- How does community perception of a user change after an evaluation?
<table>
<thead>
<tr>
<th></th>
<th>Actual Evaluation P/(P+N)</th>
<th>Text Quality</th>
<th>Community Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up-votes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Down-votes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text Quality</td>
<td>0.9</td>
<td>0.8</td>
<td>0.9-0.8</td>
</tr>
<tr>
<td>Community Bias</td>
<td></td>
<td></td>
<td>+0.1</td>
</tr>
</tbody>
</table>
Compare community bias before and after the evaluation
Communities Discriminate

Posts made after a negative evaluation were perceived worse than those made after a positive evaluation.

- On CNN, on average, your posts are being evaluated 48% worse than would be predicted by text quality, p < 0.05 in all communities.
Summary of Effects

Before

More Positive

Similar History

Positive Eval.

Similar Text Quality

Worse Text

Worse Perception

Negative Eval.

After
Evaluations can Affect…

…Posting Frequency

(How often you post)

- Does feedback regulate posting quantity?
Feedback and Posting Activity

- Do users post more after a positive/negative evaluation?
Negativity Increases Activity

Users who receive negative feedback post more frequently
Evaluations can Affect...

...Voting Behavior

(How you vote on others)

- Does feedback result in subsequent backlash of negative behavior?
More Down-Votes Given

Users who receive negative feedback are more likely to down-vote others.
Summary so far…

- Negatively-evaluated users
  - Write worse (and more!),
  - Are evaluated worse by the community
  - And evaluate others worse

- Positively-evaluated users don’t do any better
Is there a downward spiral in communities?
Downward Spiral in CNN

Proportion of down-votes is increasing over time
Can we identify antisocial behavior quickly and prevent downward spirals?

Antisocial Behavior in Online Discussion Communities
AAAI ICWSM, 2015.
Prediction Task

- **Task:** Detect which users will be banned in the future

- Post # of words ...
- Activity posts/day ...
- Community % upvoted ...
- Moderator deletions ...

- ~19k users CNN, balanced dataset, first 10 posts
Prediction Results

<table>
<thead>
<tr>
<th>Method</th>
<th>ROC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag of Words</td>
<td>0.70</td>
</tr>
<tr>
<td>Deletion Rate</td>
<td>0.74</td>
</tr>
</tbody>
</table>

0.5  0.6  0.7  0.8  0.9
Prediction Results

- Bag of Words: 0.70
- Deletion Rate: 0.74
- Post: 0.62

ROC AUC

Values: 0.5, 0.6, 0.7, 0.8, 0.9
Prediction Results

- Bag of Words: 0.70
- Deletion Rate: 0.74
- Post: 0.62
- + Activity: 0.73
- + Community: 0.83
- + Moderator: 0.84

ROC AUC
Time to Detection

User lifetime:  
- Red: $\geq N$  
- Blue: 10–20  
- Purple: 40–80  
- Orange: 5–10  
- Green: 20–40  
- Yellow: 80+

AUC vs. Using First N Posts
## Results: Cross-Domain

<table>
<thead>
<tr>
<th>Tested on</th>
<th>CNN</th>
<th>IGN</th>
<th>BREITBART</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>0.84</td>
<td>0.74</td>
<td>0.76</td>
</tr>
<tr>
<td>IGN</td>
<td>0.69</td>
<td>0.83</td>
<td>0.74</td>
</tr>
<tr>
<td>BREITBART</td>
<td>0.74</td>
<td>0.75</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Cross-Domain ROC AUC
Conclusion

- Web is a social place!
  But people are less inhibited online
- Important to understand interaction mechanisms and behavioral patterns
- Towards healthier and more welcoming online communities
Summary

Large-scale longitudinal analysis of antisocial behavior in discussion forums

Combination of data analysis and human experiments

Trolls are not “special” people

- Mood and discussion context increase the probability of trolling
Summary

Bad behavior can spread

- The effects of negative feedback are more pronounced
  - Write worse and more
  - Are evaluated worse by others
  - Evaluate others worse

- Positively-evaluated users, on the other hand, don’t do any better
Summary

- **Prediction task:** Predicting whether a user will be banned in the future.
- After observing just 10 posts, we can accurately predict whether the user will be banned in the future.
- Generalization performance across communities.
Further Questions

- **Social feedback loops**: Are trolls socially created by harsh feedback?
- **Reputation**: Effects of user reputation
- **Role of moderators**: Are moderators effectively helping and moderating
- **Linguistic norms**: Linguistic analysis
- **Feedback mechanism**: Communities with only positive feedback (likes)
KEEP CALM AND DON'T FEED THE TROLL
References

