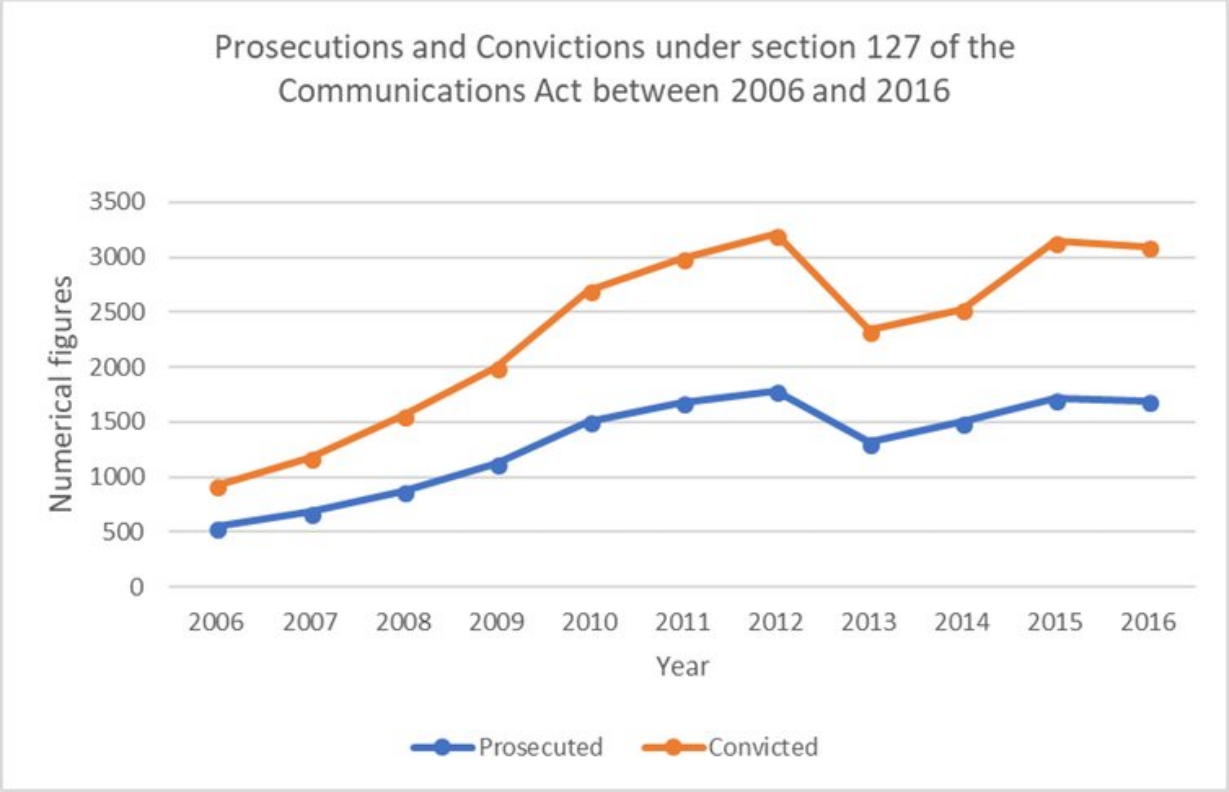
What Facebook isn't telling us about its fight against online abuse

May 22 2018, by Laura Bliss



Prosecutions for offensive messages. Credit: Based on figures obtained from the Ministry of Justice, Author provided

Facebook has for the first time made available data on the scale of abusive comments posted to its site. This may have been done under the growing pressure by organisations for social media companies to be

more transparent about online abuse, or to gain credibility after the Cambridge Analytica data scandal. Either way, the figures do not make for pleasurable reading.

In a six-month period from October 2017 to March 20178, 21m sexually explicit pictures, 3.5m graphically violent posts and 2.5m forms of hate speech were removed from its site. These figures help reveal some striking points.

As expected, the data indicates that the problem is getting worse. For instance, between January and March it was estimated that for every 10,000 messages online, between 22 and 27 contained graphic violence, up from 16 to 19 in the previous three months. This puts into sharp relief the fact that in the UK, prosecutions for online abuse have been decreasing, as demonstrated in the graph below.

Yet what Facebook hasn't told us is just as significant.

The social network has been under growing pressure to combat abuse on its site, in particular, the removal of terrorist propaganda after events such as the 2017 Westminster attack and Manchester Arena bombing. Here, the company has been proactive. Between January and March 2018, Facebook removed 1.9m messages encouraging terrorist propaganda, an increase of 800,000 comments compared to the previous three months. A total of 99.5% of these messages were located with the aid of advancing technology.

At first glance, it looks like Facebook has successfully developed software that can remove this content from its server. But Facebook hasn't released figures showing how prevalent terrorist propaganda is on its site. So we really don't know how successful the software is in this respect.

Removing violent posts

Facebook has also used technology to aid the removal of graphic violence from its site. Between the two three-month periods there was a 183% increase in the amount of posts removed that were labelled graphically violent. A total of 86% of these comments were flagged by a computer system.

But we also know that Facebook's figures also show that up to 27 out of every 10,000 comments that made it past the detection technology contained graphic violence. That doesn't sound like many but it's worth considering the sheer number of total comments posted to the site by its more than 2 billion active users. One estimate suggests that 510,000 comments are posted every minute. If accurate, that would mean 1,982,880 violent comments are posted every 24 hours.

To make up for the failures in its detection software, Facebook, like other social networks, has for years relied on self-regulation, with users encouraged to report comments they believe should not be on the site. For example, between January and March 2018, Facebook removed 2.5m comments that were considered hate speech, yet only 950,000 (38%) of these messages had been flagged by its system. The other 62% were reported by users. This shows that Facebook's technology is failing to adequately combat hate speech on its network, despite the growing concern that social networking sites are fuelling hate crime in the real world.

How many comments are reported?

This brings us to the other significant figure not included in the data released by Facebook: the total number of comments reported by users. As this is a fundamental mechanism in tackling online abuse, the amount

of reports made to the company should be made publicly available. This will allow us to understand the full extent of abusive commentary made online, while making clear the total number of messages Facebook doesn't remove from the site.

Facebook's decision to release data exposing the scale of abuse on its site is a significant step forward. Twitter, by contrast, was asked for similar information but refused to release it, claiming it would be misleading. Clearly, not all comments flagged by users of social networking sites will breach its terms and conditions. But Twitter's failure to release this information suggests the company is not willing to reveal the scale of abuse on its own site.

However, even Facebook still has a long way to go to get to total transparency. Ideally, all social networking sites would release annual reports on how they are tackling abuse online. This would enable regulators and the public to hold the firms more directly to account for failures to remove online abuse from their servers.

This article was originally published on The Conversation. Read the original article.

Provided by The Conversation

Citation: What Facebook isn't telling us about its fight against online abuse (2018, May 22) retrieved 10 April 2024 from https://phys.org/news/2018-05-facebook-isnt-online-abuse.html