Protecting Yourself from Revenge Pornography – Share Your Images Before Your Abuser Does!

Andy Phippen and Maggie Brennan
Plymouth University, UK. *Andy.phippen@plymouth.ac.uk*
University College Cork, Ireland. *Margaret.brennan@ucc.ie*

There has been considerable legislative development in the UK and other jurisdictions around the protection of the victims of "Revenge Pornography". While both England and Wales (section 33 of the Serious Crime and Courts Bill 2015) and Scotland (section 2(1) of the Abusive Behaviour and Sexual Harm (Scotland) Act 2016) legislation have already shown a significant response in the courts[1], this remain a last resort and potentially costly way of resolution for victims. The Crown Prosecution Service guidance[2] suggests that court proceedings will only be pursued by the state if there is clear evidence of intent *to cause distress* and the person in the image has not consented to the image being shared. While the current issues around sentencing and issues with the legislation lie outside of the scope of this article, the fact that intent to cause distress is so clearly defined by the CPS, we can see there will may discrepancies between what we might argue to be revenge pornography and that which can be tried in a courtroom. As is typical with any social problem facilitated by technology, legislation, with its precise definitions and case based interpretations, can only go so far to protect victims of harm.

Service Providers Need to "Do More"

As such, there remains pressure on social media companies to "do more" to protect victims of harm that arises from the non-consensual sharing of indecent images in a more pro-active manner. There have been calls[3] to ensure that this media is not posted in the first place or to automate the take down process so that victims do not have to experience repeated discoveries of the shared images and make numerous take down reports. There is certainly evidence, from organisations such the Revenge Pornography Helpline[4], that social media companies are far more responsive to take down requests than they have been historically. However, there is still a requirement for a report to take down the image before an investigation is mounted and only then, in the event that "community standards" have been breached, the media will be removed.

There have been calls for technology that might algorithmically determine whether an image is indecent, or a post contains jihadist materials. The algorithm could then prevent its posting, often through the application of fabled artificial intelligence solutions called for by UK politicians such as Amber Rudd[5]. However, some providers have declared such automated solutions are still a way off[6]. There seems little surprise that this is the case –

---

[1] http://www.bbc.co.uk/news/uk-37278264
[2] http://www.cps.gov.uk/legal/a_to_c/communications_sent_via_social_media/
[3] http://www.dailymail.co.uk/news/article-4410462/Facebook-not-haven-paedophiles.html
[4] https://revengepornhelpline.org.uk
[5] http://www.telegraph.co.uk/news/2017/11/10/amber-rudd-urges-social-media-firms-use-ai-block-extremist-content/
[6] https://newsroom.fb.com/news/2017/06/how-we-counter-terrorism/

while artificial intelligence is currently being extolled for solutions for everything from self-driving cars to determining whether a post on social media reflects the mental health of the poster, these techniques generally rely on pattern matching and classification within large scale data processing. Algorithms are "trained" by controlled data so it can "learn" similarity and hopefully lead to matching when shown a different but similar piece of data. Therefore, theoretically, an algorithm fed with a training set of indecent images would identify others with similarities in their data composition. However, given the complexities in law of determining whether an image or video is indecent it seems a high and unrealistic expectation to do this algorithmically – how could a training set ever be sufficiently large but still consistent in data composition to encompass everything that might be considered indecent in law? How large would such a set need to be? Who decided that the data set represents indecency in the legal sense of the term?

More Realistic Expectations on Service Provider Responsibly

Therefore, we might have some sympathy with platform providers in the never-ending policy makers quest to make them "do more", particularly if they are being asked to do things that are technically extremely difficult in a live, large-scale operation platform. However, there are things that platform providers can do, in the event an image has already been reported and taken down, given it would have therefore been interpreted by the provider as not having met community standards. In an ongoing case in Belfast a girl who, at the age of 14, had indecent self-generated images shared on Facebook and taken down following her reporting them, discovered that they had been shared again[7]. The girl's legal team argued that Facebook had responsibility to prevent the reposting of images already taken down and that the technology to do this is far less complex than mythical "indecency detection" algorithms – a data processing technique called file hashing[8].

Hashing is a method used in many computer based applications to verify the validity of data. It is a mathematical technique that takes data in its requisite form (for example a file, a communication, elements within a structured database) and through the application of an algorithm, processes this data into a unique alphanumeric string. The algorithms are developed to produce a unique value for any given data artefact, and are very reliable ways to compare the validity of data, or whether two data artefacts are identical in composition. For example, in a scenario where a file might be communicated between two devices, a "hash" of the file prior to transmission, and the subsequent hash of the file at the receiving end, allows the data communications protocol to check whether it has remained intact during transmission. In the case of image checking the hash is generated and stored for a given image and, if that image is posted again, the hashing algorithm will generate the same unique alphanumeric string, and the platform will know the image has been shared before.

[7] http://www.belfastlive.co.uk/news/belfast-news/picture-naked-14-year-old-11861938

[8] Konheim, Alan (2010). "7. HASHING FOR STORAGE: DATA MANAGEMENT". *Hashing in Computer Science: Fifty Years of Slicing and Dicing*. Wiley-Interscience. ISBN 9780470344736.

This is the central argument in the Belfast case – the indecent images of the girl had already been posted on Facebook and then taken down. Why, argue the legal team, could the service provider not hash any image that has been taken down so if someone tries to repost, the system will detect this and prevent posting? Since this case has moved to a full hearing, Facebook have announced a project in the US that will do what has been called for in the case for extremist content, but not indecent, non-consensually shared indecent photographs[9].

At the start of November 2017, Facebook, along with the Australian eSafety Commission, announce a pilot project that would add another level of proactivity to addressing revenge pornography concerns[10]. The proposed pilot system, initially solely in Australia, provides the facilities for users to share their self-generated indecent images with Facebook and use hashing techniques to ensure that further reposting is not possible. The basic steps of the proposed system are:

- *Australians can complete an online form on the eSafety Commissioner's official website.*
- *To establish which image is of concern, people will be asked to send the image to themselves on Messenger.*
- *The eSafety Commissioner's office notifies us of the submission (via their form). However, they do not have access to the actual image.*
- *Once we receive this notification, a specially trained representative from our Community Operations team reviews and hashes the image, which creates a human-unreadable, numerical fingerprint of it.*

Facebook stress that the image is not retained for any longer that it takes to evaluate and hash the image, the only thing that is stored long term is the hash itelf. As a final step in the process once the hash is complete and stored, the sender of the image is informed so they can remove the image from Messenger, and at that point Facebook will remove the image from their database.

Therefore, if there is a case of someone else trying to share that image on Facebook, when that image is hashed and compared against the user uploaded indecent image hashes, there will be a match, the image will not be posted and the intended poster will be informed their upload infringes community standards. In that way, the approach "protects" the person in the image from exposure by others.

Throwing Tech at Social Problems

While this is an interesting concept, and does show how existing proven technology can be used to provide an element of protection for those wishing to protect themselves from risk of exposure, it does also demonstrate how, if left to service providers, their only route to protection is technological (unless such a platform employs human intervention prior to *anything* being uploaded, which is possibly exorbitantly resource intensive). This

[9] https://newsroom.fb.com/news/2016/12/partnering-to-help-curb-spread-of-online-terrorist-content/

[10] https://newsroom.fb.com/news/h/non-consensual-intimate-image-pilot-the-facts/

subsequently does not provide a complete solution while once again demonstrating Ranum's Law[11] – *"You can't solve social problems with software."*

We can reflect upon a few potential issues with this approach, by way of illustration of this point. Firstly, a fundamental issue lies in the age of the poster of images. We are fully aware that teen "sexting" is a complex and challenging issue for a lot of young people[12] and there is a high likelihood that images shared by young people among their peers may end up being posted on social media. Therefore, one can realistically assume that young people might view the proposed approach by Facebook as one that could be used to ensure they are protected from the risk of spreading images of themselves. Of course, this poses a serious problem to the "solution" provider – if young people are sharing self-generated images with Facebook, the provider is, in essence, facilitating the distribution of indecent images of minors. The subsequent storing of these images further breaks child protection laws – the duration of storage is not something defined in the covering statutes, it is any storage of indecent images of minors that defines the criminality.

While the provider might argue that, given there is a vetting process which will check the profile of the user who is uploading the images, they verify age via the profile, much has been written about the struggle all social media providers face with age verification on their systems. There are many "under aged" users on these social media platforms, and the means by which most social media providers verify age (asking at sign up for a date of birth) is hardly rigorous. And while a provider might argue that any user lying about their age on sign up is invalidating the terms and conditions of use and therefore their right to protection, it would be a very brave defence in the event of providing a "solution" to revenge pornography that potentially encourages the distribution of indecent images of minors.

We then move into the perennial legal and moral debate around who determines the indecency of the image and makes the judgement on whether it should be hashed and stored. While Facebook has stated that "*a specially trained representative from our Community Operations team reviews and hashes the image",* the review process itself is not defined, nor are the thresholds for indecency. What would be the situation where a user flagged an image for hashing they believed was placing them at risk, but the reviewer decided that the image was not indecent "enough" to be hashed. Who makes the decision on decency? The image owner, or the platform provider? We know from our own work that there are cultural differences around indecency – for example, a western perspective on indecency might be different from one from an Asian one. While nudity might be a threshold for some, cleavage might be sufficient in another. And while legal indecency may not be met, the impact on the victim in the event of the sharing of such images may still equally severe.

Following on from this, the owner of the image also raises some concerns – for example, if we are to state that the image has to be self-generated in order for Facebook to be able to hash the image and store it, can we assume that indecent images of individuals taken,

[11] https://utcc.utoronto.ca/~cks/space/blog/tech/SocialProblemsAndTechnicalDecisions
[12] Phippen, A. (2017). "Children's Online Behaviour and Safety: Policy and Rights Challenges". Palgrave, UK.

for example, by a partner are not covered? What would happen then if a vexatious ex-partner, with a collection of images they had taken of the victim, decided to share those images on the platform? We must assume that this protection mechanism would not be open to the victim in this scenario, because copyright for the image lies with the taker of the image, not the subject in the image (as defined in the Copyright, Designs and Patents Act 1988) as long as the subject consented, at the time, to the image being taken. In essence, this protection will only work for "selfies". Once potential abusers are aware of this, their approach to abuse will simply switch to sharing images for which they retain copyright.

There are also issues with hashing itself – given the aim of a hashing algorithm is to produce a unique hash based upon the whole of the data within an image, any modification to the source image will produce a different hash, even if this alteration is a single pixel. While Microsoft's PhotoDNA (the proposed hashing technology in the Facebook trial) has some built in resilience to this, such as resizing or recolouring[13], it will still produce a different hash if the image is cropped or edited in a more complex way. Therefore, if a potential abuser is sufficiently aware of this "protection mechanism", it is very easy to bypass by creating a new image that is perhaps a crop of the original image. While hashing is a very good way of comparing an exact copy of an image, it will be far less successful if such images are modified in some way.

Responsibility for Protection vs Going After the Abusers

Even in this brief review of the issues, from a legislative and technical perspective we can see that the "solution" is not without significant problems. And while we should applaud efforts by service providers to think about how technology can be used to help victims, we can also see that technology is always going to struggle to find solutions within complex social systems. In essence, the pressure on service providers to "do more" is resulting in surrogates implemented in code rather than addressing the issues with a multi-stakeholder approach drawn from effective social policy. However, the more fundamental issue around this "solution" that lies at the heart of victim blaming within sexting and revenge pornography scenarios – responsibility.

In our own work with RP victims, both adults and minors, what is clear is that they will often blame themselves for the redistribution of the image. Many victims will state that by taking the image they placed themselves at risk, and it was therefore their fault their abuser had the image to share or use for further harassment. With our work in sexting, recent research presented at the Australian eSafety Commissioner's Conference in November 2017[14] over 70% of young people surveyed (n=700) say the responsibility of the image lies with the maker (and usually subject) of the image.

By communicating to the population that the way to protect yourself from the dangers of abuse from revenge pornography is to share the image with a third party who will put in measures to protect you, surely one is saying the responsibility for the control and management of the image lies with victim? There are two messages this given out by

[13] https://news.microsoft.com/features/microsofts-photodna-protecting-children-and-businesses-in-the-cloud/

[14] ttps://www.esafety.gov.au/on-the-edge-17

this. Firstly, the provider is saying *“we have provided you with a means to protect yourself from someone else sharing images you might have taken of yourself on our platform. if you don’t engage with this solution, it will be your fault if someone posts it on our platform”.* However, perhaps the more confusing message is that *“if you want to protect yourself from someone sharing images of you, you need to firstly share the image with someone else. But don’t worry, we can be trusted. But don’t share with anyone else, because they can’t be, unless you share with us first!”*. These messages are surely just further ways of showing the victim that responsibility for the non-consensual sharing of an image lies with them, unless they do something further to protect themselves. This harks back to the sort of messages young people talk about when they raise the non-consensual sharing of images with adults supposedly responsible for their safeguarding – “well, you shouldn’t have done that should you!?”.

Surely the focus of control, and technical intervention, should lie with ensuring that others cannot non-consensually share indecent images of others and, if they do there will be consequences for doing so. This should be implemented in line with more effective public communication, legislation, and stronger sentencing of those who chose to share images of other without consent. We fear that this “solution” just places more pressure on the victim to ensure they protect themselves and the responsibility for protection lies with them.