Article:
Gerrard, Y. (2018) Beyond the hashtag : circumventing content moderation on social media. New Media and Society, 20 (12). pp. 4492-4511. ISSN: 1461-4448

https://doi.org/10.1177/1461444818776611

Beyond the hashtag: Circumventing content moderation on social media.

Abstract

Social media companies make important decisions about what counts as 'problematic' content and how they will remove it. Some choose to moderate *hashtags*, blocking the results for certain tag searches and issuing public service announcements (PSAs) when users search for troubling terms. The hashtag has thus become an indicator of where problematic content can be found, but this has produced limited understandings of how such content actually circulates. Using pro-eating disorder (pro-ED) communities as a case study, this paper **explores the practices of circumventing hashtag moderation in online pro-ED communities.** It shows how: (1) untagged pro-ED content can be found without using the hashtag as a search mechanism, (2) users are evading hashtag and other forms of platform policing, devising signals to identify themselves as 'pro-ED', and (3) platforms' recommendation systems recirculate pro-ED content, revealing the limitations of hashtag logics in social media content moderation.

Keywords

Algorithms, content moderation, eating disorders, hashtags, Instagram, Pinterest, pro-ana, social media, Tumblr.

Introduction

Social media companies encourage their users to share content about themselves but downplay decisions about how they moderate problematic posts and why they choose to do so. Platforms often make decisions about moderation when they face public pressures, like accusations that they host pro-eating disorder (pro-ED) content[1]. This is what happened in February 2012, when a *Huffington Post* writer published a widely read exposé on the 'secret world' of Tumblr's thinspiration blogs (Gregoire, 2012). Other publications like *The Atlantic* (Greenfield, 2012) and *Jezebel* (Ryan, 2012) joined the debate, criticizing platforms like Instagram, Pinterest and Tumblr for failing to intervene in online eating disorder communities. By May 2012, all three platforms had publicly announced their plans to minimize the spread of pro-ED content. Pro-ED has long held the attention of the popular press, academic researchers and Web companies, and these three stakeholders have often claimed to share concerns that people use the Web - be it through homepages, forums, social media or otherwise - to promote and glorify eating disorders.

Pro-ED communities are a longstanding societal concern. To be pro-ED is to promote an eating disorder 'as a "lifestyle choice" rather than as a "disease" (Paquette 2002), thus challenging medical and psychiatric conceptualizations which position the "sufferer" as passive and helpless' (Day and Keys, 2008, p.5). The relationship between social media and eating disorders has become more important in recent years because of the rise in reported cases of eating disorders amongst young women in the United Kingdom, for whom social media

platforms play a meaningful role, and because of sensationalist press discourses telling readers that 'social media is to blame' (Dugan, 2014). While some in medical circles and elsewhere condemn pro-ED spaces, others - particularly feminist scholars, myself included - also recognize their value as (cyber)spaces of support, mostly free from the social stigmatization that accompanies forms of disordered eating (see for example Dias, 2003; Bell, 2009; Ging and Garvey, 2017). It can also be difficult to interpret these spaces as '*pro*-ED' at all, as such content often sits alongside pro-recovery, not-pro-anything, and other complex positionalities, and it is the **shared codes and circumvention techniques that perhaps define this community of users**. It is thus too simplistic to read these spaces as either 'good' or 'bad' (Bell, 2009). **But by announcing new rules for content related to eating disorders, platforms have decided that these communities can only occupy a certain kind of space on social media: one that is located *at the margins.***

Instagram, Pinterest and Tumblr were uncharacteristically vocal about their decisions to police pro-ED content (Gillespie, 2015), announcing their interventions through a series of blog posts and press releases. The platforms enforce their rules in fairly similar ways – by May of 2012, all three began to issue public service announcements (PSAs) when users search for troubling hashtags, like #proana (pro-anorexia) and #thinspiration, and Instagram began to block the results of certain hashtag searches. These rules were still in place at the time of writing and work alongside user-driven forms of moderation such as 'flagging'. **When a user flags a post on social media, it passes through an automated system which matches it against 'known databases of unwanted content, facial recognition, and "skin filters,"**

which screen photos or videos for flesh tones and then flag them as pornography' (Roberts, 2017b, n.p.). If a flagged post is not already known to a platform, it will be sent to a human commercial content moderator (CCM) who will use a set of guidelines provided by the company (and not publicly available) to decide whether it should stay or go (Gillespie, 2017; Roberts, 2017a,b). In their Community Guidelines and similar public-facing policies, the platforms explain that they prohibit user accounts and individual posts that 'glorify', 'promote' or 'encourage' eating disorders (Instagram, 2012; Pinterest, 2017a; Tumblr, 2012a,b), but they do not provide clear definitions of these terms, making it difficult to know how users might break the rules.

Instagram, Pinterest and Tumblr have joined a much longer debate about mediated depictions of disordered and typically young, female and white bodies (Bordo, 2003; Bell, 2009), following the standards set by older social networking sites (SNSs) like MySpace and Xanga, and Web hosts like Yahoo!. But what made the 2012 iteration so unique was *how* the platforms chose to intervene: through hashtag moderation. The hashtag - the hash or pound symbol (#) followed by a string of alphanumeric characters (for example, #promia or #size00) - is used widely across social media platforms. Hashtags are an appealing point of intervention for various reasons. They are convenient tools for aggregating relevant content between users outside of each other's follower/followee networks (Schmidt, 2014), they circumvent the difficulties of algorithmically tagging visual imagery to categorize it as 'pro-ED', and they help CCMs to interpret posts within the seconds they have to decide whether they should stay or go (Roberts, 2017a). Users do this work for the platform by tagging their posts with ED-related

terms, but a hashtag cannot tell the whole story about a given phenomenon. For instance, some users have developed a set of signals to indicate their content as pro-ED to other interested users, in careful and mostly untagged ways that allow them to evade moderation. Platforms also present users with pro-ED content through automated recommendation systems, pointing to the very limited solution that hashtag monitoring represents.

In this article, **I explore the circumvention of hashtag moderation in online pro-ED communities.** While discussions of content moderation tend to focus on the human labor (Roberts, 2017a,b) and broader politics of platforms' interventions (Gillespie, 2015, 2018), I ask how an already-marginalized community of users works around these techniques and why they might be doing so. This article also responds to recent calls for more methodological approaches to obtaining untagged content, however difficult this work may be (for similar arguments, see Mitchell et al, 2015; Bruns et al, 2016; D'heer et al, 2017). I use an innovative methodological approach to reveal the importance of untagged and evasive communication between pro-ED users. Indeed, only 779 of the 2612 posts I analyzed included hashtags, suggesting that hashtags are not an especially powerful or trustworthy communicative tool for users within the pro-ED community. I discuss the methods I used after explaining why platforms privilege the hashtag as a form of moderation. I then examine how users learn to recognize and signal each other as pro-ED in the absence of hashtags, before turning to the role of recommendation systems in suggesting pro-ED content to users *despite* hashtag bans. This makes hashtag moderation a rather ineffective intervention.

Privileging the hashtag in a pro-eating disorder 'problem'

There is a reason why platforms use hashtags as a mechanism through which to police problematic posts: because non-tagged content is more difficult to find. Hashtags are perhaps the most *visible* form of social media communication, connecting content between users 'who have no preexisting follower/followee relationship' (Schmidt, 2014, p.6). Hashtags' visibility makes them distinct from other forms of social media engagement such as liking and commenting, but identifying untagged pro-ED content from Instagram, Pinterest and Tumblr's userbase is a difficult, perhaps impossible task for both human content moderators and platforms' automated moderation mechanisms. There are of course tools to algorithmically tag visual imagery on social media, but these methods are notoriously unreliable. For example, image-hosting platform Flickr rolled out an auto-tagging system in May 2015, but quickly faced backlash from users after it incorrectly tagged images of Dachau concentration camp with the 'jungle gym' and 'sport' tags (Hern, 2015).

Human content moderators face the same problem. For example, on its Community Guidelines Pinterest states that it will 'remove anything that promotes self-harm, such as self-mutilation, eating disorders or drug abuse' (Pinterest, 2017a). It gives an example of an image (see below) that would be acceptable, claiming 'It's okay because the focus is on nutrition and fitness':

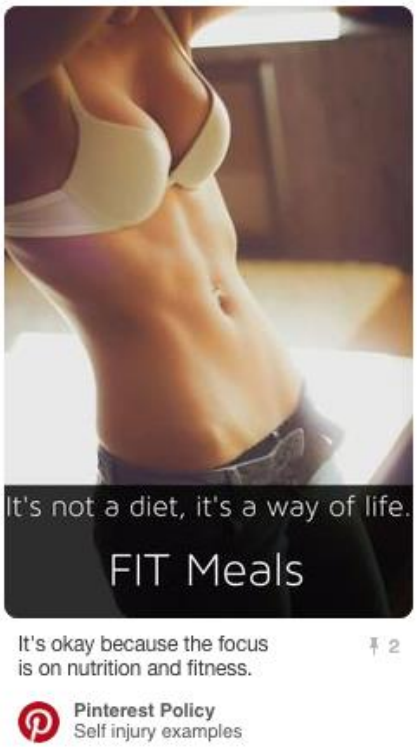

Figure 1: An example of an 'acceptable' image of a female body, provided by Pinterest in its Community Guidelines (Pinterest, 2017b).

The image's overlaid text 'It's not a diet, it's a way of life. FIT Meals' de-couples it from pro-eating disorder discourses, but plenty of images like the above could be shared as both 'thinspiration' or 'fitspiration' posts (Lewallen and Behm-Morawitz, 2016). The absence of hashtags and text overlays would make it difficult for both automated image tagging systems and human content moderators to make the distinction between 'thinspo' and 'fitspo' imagery, if indeed the latter should be understood as less socially problematic. By including hashtags in a

post, users are telling platforms - intentionally or otherwise - what the post is about. Hashtags are perhaps the most visible form of social media communication, making them vulnerable to platforms' interventions, especially if they are controversial. But they are also versatile, ready to be re-shaped or even abandoned by users in response to platforms' rules, **which is precisely what some users in the pro-ED community are doing.**

The over-emphasis on hashtags extends to research in the social sciences. Scholars pay attention to hashtags for many of the same reasons that platforms do: the convenience of accessing data from what appears to be a diverse range of users, and their prominence in current debates about eating disorders and social media. Hashtags are frequently used as entry points for data collection, but as Bruns et al explain:

> Hashtag datasets [...] constitute the low-hanging fruit in social media data, which has led to an abundance of research building on such datasets, compared to a relative dearth of studies drawing on less instantly accessible sources (Burgess and Bruns, 2015). [...] There is a strong need to put hashtag use into better perspective also by comparing the patterns of engagement around topical hashtags with the broader patterns of activity relating to these topics outside of the hashtags themselves - however methodologically difficult such work may turn out to be. (2016, p.21)

The difficulty of finding untagged social media data is reflected in an over-reliance on research about tagged pro-ED social media content. In the social sciences, hashtags have been

used to locate pro-ED content for analyses of tagged images (Seko and Lewis, 2016; Ging and Garvey, 2017). In the computer sciences, Chancellor et al (2016) have identified a range of hashtags that Instagram users coined to work around the platform's hashtag ban (for example, #thighgap became #thyghgapp), and Moreno et al (2016) have found a number of deliberately ambiguous non-suicidal self-injury (NSSI) tags on Instagram, like #secretsociety123. **But missing from the growing body of research on social media and pro-ED is an examination of the circumvention of hashtag moderation, including non-hashtag use and signalling techniques, combined with the re-circulation of pro-ED content through platforms' recommendation systems**. Hashtags need to be put *in perspective* as they represent only a narrow subset of social media's many communicative layers (Bruns and Moe, 2014). I now discuss my approach to locating pro-ED content without relying on hashtag searches.

Finding untagged pro-eating disorder content on Instagram and Tumblr

To find new content on social media, users might search for hashtags and keywords through in-platform search engines. This makes the hashtag an important wayfinding mechanism. But Instagram and Tumblr moderate searches for problematic hashtags, making it hard for users (and researchers) to find untagged pro-ED and other moderated posts. This has led to an understandable but nonetheless problematic over-reliance on tagged data and methods in social media research. There are also ethical considerations for researching untagged data, as such posts might signal a user's desire to be excluded from a broader conversation about a given

topic and minimize their visibility to particular actors (Larsson, 2015). **This is especially true for pro-ED users who often work hard to avoid detection. While research with these users can be done ethically – for example, I do not substantively quote from social media posts or include usernames – it is important to reflect on the tensions of investigating and exposing the secrets of a community operating at the margins of social media.**

It is ironic that I am emphasizing the need for these communities to have secrecy while simultaneously revealing what they are doing. This tension is difficult to reconcile and there should be sound justifications for exposing a marginalized communities' secrets. People with eating disorders, for example, often turn to social media to discuss their conditions and escape surveillant press, medical and other discourses. But given the difficulties of accessing eating disorder patients to talk to them about their relationship to social media (Lavis, 2015), platforms like Instagram, Pinterest and Tumblr offer rare spaces for researchers to gather knowledge about this phenomenon: one that I intend to address the complexities of pro-ED identities and move away from a 'good' and 'bad' dialectic (Bell, 2009). I discuss my findings about users' evasive techniques later in this paper, but first describe how I located untagged posts on Instagram and Tumblr: an innovative methodological approach that could be applied to other moderated phenomena.

I began my research by creating new accounts on Instagram and Tumblr, though I already had personal accounts on both platforms. I also created a separate account on Pinterest but save my discussion of this platform for the final section on recommendation systems. As

people typically access social media through apps rather than Web browsers (Light et al, 2016), I walked through the apps on an iPhone. I took a platform-specific approach, navigating my way through Instagram and Tumblr according to their unique architectures and rules around moderation. For example, users of the Instagram app can use its in-platform search engine to search for top posts, people, tags, or places. Instagram currently moderates the results for pro-ED and other in-platform searches in four main ways: (1) (semi)-permanent blocks, where a search returns no tagged content, (2) new posts moderated, where the platform shows you thirty-six 'top posts', (3) a 'no posts yet' error message, telling users how many posts have been tagged with a term but not returning any content, and (4) a public service announcement (PSA) shown before search results are returned (Suzor, 2016). Instagram, Pinterest and Tumblr have not produced exhaustive lists of all banned pro-ED terms and they also vary on a daily basis (Suzor, 2016). This echoes Gillespie's (2018) argument that decisions about content moderation are built into platforms' closed codes.

Although Instagram's hashtag moderation is mostly successful in restricting access to pro-ED content, there are other routes into these communities. I found that Instagram does not block or even restrict access to searches for top posts or for people. At the time of my analysis, Instagram returned search results for users whose account names or biographies feature the following pro-ED terms: proana, proanorexia, thinspiration, thinspo, and thighgap[2]. I identified ninety-six Instagram users through this method and whose accounts were related to eating disorders, a message they communicated through their typically-pseudonymous usernames, profile biographies, and/or captioned content. Seventy-four of these accounts were public and

twenty-two were private. Over a two-week period, I manually coded 1612 posts from the public accounts. I counted the numbers of tagged and untagged posts within the dataset and found that only 561 of the 1612 posts included hashtags in their captions, suggesting that hashtags are not an especially important communicative tool for these users. I made field notes about the content and captions of the untagged posts, and of the keywords in public and private profile biographies, which I developed into a number of themes and discuss in the forthcoming sections.

Tumblr's approach to content moderation is different. Tumblr does not ban any search terms and instead issues a PSA when users search for certain tags, which currently reads 'Everything okay? It sounds like you could use some kind words right about now. We suggest Koko, an anonymous support community made up of nice, caring people like you'. Unlike Instagram, Tumblr returns content for all pro-ED searches and users can simply scroll past the PSA to view the tagged content. But Tumblr does not issue a PSA if you 'follow' certain keywords in the same way you would a blog. The platform lets users follow certain topics to ensure they have content on their dashboard even if they are not connecting with specific users. This is similar to Instagram's 'explore' function, which shows users content from accounts they are not yet following but might be interested in[3]. I began to follow ED-related terms like 'bulimia' on Tumblr and received the following automated message when I clicked 'follow': 'Lovely. All the best things about bulimia will automatically show up on your dashboard'. This message was positioned beneath a PSA, as shown in the image below:

Figure 2: A screenshot from the Tumblr app.

Once I had followed ED-related terms - anorexia, anorexic, bulimia, bulimic, thinspiration, thinspo, proana, purge, purging - the platform delivered this content to me through my dashboard and also via email. Tumblr showed me relevant posts and suggested a list of users whose accounts I should follow. As some of these terms are not straightforwardly *pro*-ED (unlike, for example, proana), I was presented with blogs identifying as 'pro-recovery' in their biographies. But I excluded these blogs from the dataset as they were not the focus of my analysis. Tumblr recommended blogs that were, for example, 'big in proana' or 'like' other popular blogs. I identified fifty pro-ED users through this method. I analyzed twenty posts per

user, giving me a total of 1000 Tumblr posts. I manually coded these posts over a two-week period, making field notes on keywords in their profile biographies and the numbers of tagged and untagged posts within the dataset. Similar to Instagram, only 218 of the 1000 Tumblr posts I analyzed included hashtags, again telling us that tags are not an especially important form of communication for pro-ED users.

The kinds of moderation offered by Instagram and Tumblr suggest that the platforms are trying to make such content unsearchable and less visible under the watchful eye of press and other commentators. Visibility leads to accountability, and hashtag moderation tells concerned parties that social media companies take this issue seriously. Even if the platforms were to correct the moderation gaps I have noted in this section, already-networked pro-ED users are unlikely to rely on in-platform search engines to find new posts and users. They learn to navigate platforms in all sorts of ways and they know how to break the rules. I will now present my analysis of untagged posts to show how pro-ED users **recognize and circumvent** content moderation.

Hiding in plain sight: Signalling the pro-eating disorder userbase

Users who are conscious about content moderation - of which there are many - must go beyond the hashtag to find new ways of being visible to those who they wish to be seen by. Given how loudly Instagram, Pinterest and Tumblr voiced their interventions through various

blog posts and press statements, it is no surprise that the pro-ED community is aware of and thus tries to circumvent moderation. It also comes as no surprise because eating disorders have long been socially stigmatized, censured and erased from view (Ferreday, 2003), meaning pro-ED users are 'forced to deploy mechanisms of denial and disguise' (Cobb, 2017, p.192). But how do researchers learn how to see communities and users who do not want to be seen, or who want to be seen only by the 'right' people? **For what reasons should researchers be doing this kind of work, ethically speaking?** And how do social media users learn how to see each other under these conditions?

Donath's (2007) work on signalling theory is useful for exploring how users identify content as pro-ED in the absence of hashtags and other obvious markers. She argues that people often rely on signals rather than directly observable traits to learn about each other and to 'indicate the presence of those hidden qualities' (Donath, 2007, p.233). Cobb makes a similar point about online pro-anorexia communities:

> In recent years pro-ana online spaces have dispersed and become increasingly more difficult to find (Ferreday, 2009); because of censorship, these spaces are renowned for engaging in 'an elaborate game of cat and mouse to remain one step ahead of the "authorities"' (Crowe and Watts, 2016, p.381). (2017, p.190)

But it takes time to learn how to read these subtle signals, a job that might be difficult for a human content moderator who is given 'only a few seconds' (Roberts, 2017a, p.2) to

decide whether a post should stay or go, albeit informed by a social media company's closely guarded guidelines (Gillespie, 2017). It is precisely because hashtags are valuable methods of contextualization for platforms and other concerned parties that users have developed a set of signals to subtly indicate their content as pro-ED.

Many users are aware that pro-ED content is a target for moderation, and one of the most obvious ways to deflect attention is to simply not use hashtags. Indeed, only 779 of the 2612 posts I collected from Instagram and Tumblr included one or more hashtags. But users must still indicate their content as pro-ED if they want to signal those who are in the know, and so they frequently use in-group signals to identify their profiles and content as pro-ED. They also talk back to moderation through their profile biographies by posting what I call 'disclaimers of denial': **phrases used in Instagram and Tumblr profile biographies which disavow pro-eating disorder identities (for example, 'I'm not pro-anything') to** reassure moderators and non-in-group users that their accounts are unproblematic. In what follows, I describe the richness that scholars will not see if we continue an over-reliance on tagged datasets and ignore the reasons why people do not use hashtags, which includes an awareness of social media content moderation.

'One like = one hour of fasting': Pro-ED in-group signals

Because pro-ED hashtags are scrutinized, members of this community have developed a set of non-tagged signals to indicate their identities to likeminded users. Members of the pro-ED community want to be partially visible: visible enough to find other users and content and to also *be* found online, but sufficiently hidden to avoid moderation. **The use of coded language is, of course, not always related to content moderation.** Lingel (2017) **for example explores how** members of New Brunswick's punk rock community use a shared set of signals across social media to organize an underground network of music shows. **There is great value in being visible to a particular community but not to a broader set of users, however for the pro-ED community – which has long been policed by various parties – signalling techniques provide an extra layer of protection.** Within the online pro-ED community, much of this signalling work is done through users' profile biographies. Instagram, for example, allows users to choose a profile name, username, provide a link to their website or other online presence and write a 'bio'. Instagram users can see these details even if an account is set to private. Tumblr's architecture is very similar, allowing users to choose a title for their blog and write a description of it. Both of these biographical spaces are important yet volatile communicative tools for pro-ED users as they are used to 'effectively exclude outsiders, parents, or those with censorial privilege, while simultaneously signalling to fellow pro-anas that such content can be found therein' (Cobb, 2017, p.195). These signals, along with the other techniques I discuss below, become important precisely because pro-ED users no longer trust the hashtag to keep them safe.

A popular biographical signal in the pro-ED community is a list of users' target weights: their starting weight (SW), current weight (CW), goal weight(s) (GW, GW1, GW2) and sometimes an ultimate goal weight (UGW). These acronyms - which are intended to document users' weight loss journeys - are not exclusive to the pro-ED community. In my analysis, I found that the same language is often used on fitness, health, and nutrition accounts. What can differentiate this pro-ED language from other contexts are the weights listed. For example, a user I found on Instagram listed their UGW as forty-three kilograms. According to the UK's National Health Service (NHS) body mass indicator (BMI) calculator, achieving their ultimate weight would make this particular user extremely underweight and put them at risk of death. **But would a commercial content moderator – who has only a few seconds to make a decision about a user's flagged profile – decide that their account should be removed from the platform given the presence of goal weights in their biography? The recent leak of Facebook's guidelines for CCMs tells us that decisions about what should be removed and what should stay are done in secret, and that much of this work is interpretive (Gillespie, 2017).** Weight loss should also be understood as a symptom of only *some* eating disorders, thus not all pro-ED users will list dangerously low weights in their bios, making it even more difficult to identify users with eating disorders.

A slightly more reliable indicator that a user identifies as pro-ED is their participation in 'fasting games'. The image below shows a game played on Instagram, where for every 'like' the image receives the poster will fast for one hour. When I took this screenshot, the user was planning to fast for eighteen hours:

Figure 3: An example of a fasting game, taken from Instagram.

Other users post images relating to diet plans - like the Ana Boot Camp Diet (ABC Diet) (Fleming-May and Miller, 2010) - which severely reduce a person's daily calorie intake. Ging and Garvey categorize these practices as 'gamified and interactive' (2017, pp. 6-7). Again, a human content moderator is unlikely to have the time to do the calculations behind these diet plans and decide they 'promote', to borrow the language often used in platforms' Terms of Use, eating disorders. With innocuous names like 'the ABC Diet', a content moderator who is not in the know might decide that this content should stay, if it were to be reported by another user. **It is also difficult to determine whether this kind of content encourages *other* users to embrace an eating disorder, a behavior the three platforms categorize as a form of 'self-harm' through their Terms of Use and similar public-facing policies. The same can be said about users' goal weights – these metrics are expressions of self-identity and do not straightforwardly 'encourage' harmful behaviors. Not only are posts like these difficult to**

interpret as pro-ED in the absence of hashtags, but it is also not clear whether they would break platforms' ever-changing and opaque rules around content moderation.

To support each other through these games and diet plans, users within the pro-ED community will also request an 'ana buddy': a person who supports someone through their anorexia, not to seek treatment or recovery but to worsen the condition by losing more weight (Ging and Garvey, 2017). Some users also offer and request 'meanspo', short for 'mean inspiration', where users post negative comments about other users to discourage them from eating. Some users ask for 'tips' to lose weight quickly, hide their condition from parents, teachers and friends, and to purge effectively (Yom-Tov and boyd, 2014). Another way to identify an Instagram or Tumblr post as pro-ED is that they often have a certain visual aesthetic. Ging and Garvey have recently developed a visual typology of pro-ED content on Instagram, identifying categories like 'black and white and bleached out colours', which 'not only accentuates bone protrusion but also references art photography and the kind of aesthetic frequently associated with high-end or designer fashion' (2017, p.12). The Instagram and Tumblr images I analyzed were often aestheticized in this way, helping me to recognize posts as pro-ED in the absence of hashtags and other clear markers.

It is arguably difficult to uncover the hidden meanings behind these signals in the absence of hashtags. As Donath notes, 'it is important to keep in mind that the interpretation of any signal is subtle and subjective' (2007, p.238) and it also takes time to learn how to read them. I analyzed 2612 Instagram and Tumblr posts, looking at the images themselves and also

the poster's caption. It took me a long time to recognize that these signals - metrics about weight, the ana buddy system, requests for tips, and decisions about image aesthetics - were unique to certain communities of users. There was also the chance that I would get this interpretative work wrong, just as platforms sometimes do. But users understand that this interpretive work takes time for those who are not in-the-know, which **might explain** why they do not use hashtags.

'Not-pro-anything': Talking back to moderators through disclaimers of denial

Many users within Instagram and Tumblr's pro-ED communities are aware their practices are unwelcome, not just by society at large but also by platforms. It is for this reason that users often explicitly disavow a 'pro-ED' identity in their profile biographies, coining terms like 'not-pro-anything' to reassure an imagined third party – a content moderator, platform policy-maker, concerned user, or even a troll – that their account is unproblematic. Disclaimers of denial offer a new way for researchers to identify pro-ED content on social media but might confuse moderators who have not spent as much time familiarizing themselves with these subtle and continuously evolving discourses. The phrase most commonly used by pro-ED users in my dataset was 'not-pro-anything' or 'not promoting anything', which means they do not affiliate with a *pro*-eating disorder identity. Some users claim they made their accounts 'for myself', 'for motivation' or 'for personal purposes', explicitly distancing themselves from the wider pro-ED community. This makes it difficult to neatly recognize and define 'pro-eating disorder'

communities, as content that might be understood to encourage and promote disordered eating often sits alongside pro-recovery, not-pro-anything, and other complex positionalities.

One reading of this practice is that users are disavowing socially stigmatized identities like 'anorexic'. Cobb makes this point about pro-ana bloggers' disclaimers that they are 'not pro-ana', arguing that:

> users create a distinction between pro-ana and thinspo, suggesting that pro-ana is pathological but thinspo is acceptable. […] For instance, one blogger describes herself as 'Ana [anorexic] Mia [bulimic] and addicted to thinspiration,' yet adds immediately after, 'This is *not* pro-ana' (Thinspo3), presumably in an attempt to distance herself from what has been decreed a contentious phenomenon. (2017, p.195, emphasis in original)

Although social stigmatization might play a role in this kind of disavowal, these practices could also be read as users' savvy attempts to talk back to content moderation and retain a façade that their behaviours are acceptable to those who 'watch' social media activity. This is a form of obfuscation and an 'in the know' comment – a performance, a disclaimer.

Users also acknowledge platforms' potential interventions by telling people in their network that they have created a back-up account in case their main account is shut down, or that a past account has been removed. Some users include this text in their profile biographies,

while others include it in their posts. Signals like these offer a way for users to moderate other users' behavior, thus content moderation might also be understood as something that users do - as *user-led* moderation. Although these users might not share the same initial motivation as platforms - economic, institutional, legal, and so on - they are trying to rid their accounts of, to borrow from Tumblr's policy against self-harm blogs, specific kinds of content that 'aren't welcome' (Tumblr, 2012a). I now turn to a discussion of another way pro-ED content circulates beyond the hashtag: through social media's algorithmic recommendation systems.

Trending in anorexia this week: Platforms as recommendation systems

I argued in my introduction that hashtag moderation is an ineffective intervention into pro-ED communities. This is partly because social media users **circumvent it**, but it is also because part of the work of platforms is to *recommend* content to their users. While users are often aware of and are not wholly conditioned by algorithms (Bucher, 2017), much of what they see on social media is chosen for them. Once someone is embedded in a pro-ED or other network - through their followers/followees, the content they share, like, save, comment on, their clickstreams, and other forms of mined social media data - they do not need to rely on hashtags to find new content. Instead, platforms begin to recommend it to them. In my investigation - in which I did not post any content, or follow or engage with any users such as by liking their posts - platforms presented me with pro-ED content through my algorithmically-organized Instagram, Pinterest and Tumblr feeds, and also via email. **Users tdo not always**

have to strategically circumvent content moderation and can instead simply enact a pro-ED identity on social media to see this kind of content.

Although all three platforms have recommendation algorithms, they encourage different forms of communication from their users, affecting how content is seen and experienced. For example, on Pinterest:

> The most common interaction two users will have with one another is not commenting or following but simply repinning, or adding another user's pin to one's own collection (Hall and Zarro, 2012). […] Although it is true that Pinterest permits users to follow one another and comment upon pins and boards, one's followers seem not to be the primary audience. Rather, it is a curation of the self (e.g. Donald and Zheng, 2009, p. 507). (Friz and Gehl, 2016, p.691; p.695)

Instructions about followers/following are also not included in Pinterest's sign-up interface, meaning the focus is placed squarely on a user's interests (Friz and Gehl, 2016). Pinterest prioritizes content curation over creation and communication thus its recommendation algorithms play a central role in how users experience the platform and learn about new content.

When a user finds an image on the Pinterest app, they can scroll down the page to view other recommended content. The platform also suggests alternative yet related phrases that users might want to search for, or 'ideas you might love', as seen in the image below:

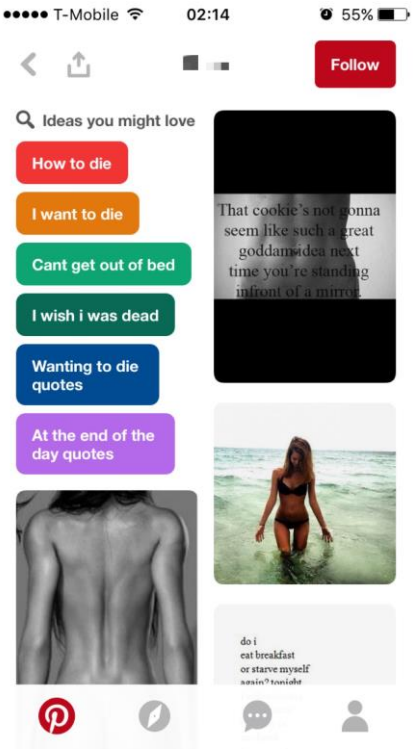

Figure 4: A screenshot of a Pinterest recommendation.

To reach this page, I browsed various ED-related pin boards. Two of the images depict slender but muscular female bodies and may be coded as 'fitspo' (Lewallen and Behm-Morawitz, 2016). But one of the images has the black-and-white pro-ED aesthetic identified by Ging and Garvey (2017) and reads: 'That cookie's not gonna seem like such a great goddamn idea next time you're standing infront of a mirror'. Pinterest also suggested I might 'love' to view other 'ideas', all of which are connected to death and suicide, such as 'how to die' and 'wanting to die quotes'. Pinterest is thus algorithmically aligning pro-ED imagery with discourses like death, suicide and self-harm, which resonates with the framing of eating

disorders in the three platforms' policy wordings. Instagram categorizes the 'embrace of anorexia, bulimia, or other eating disorders' as a form of self-harm (Instagram, 2012), Tumblr aligns 'blogs that actively promote self-harm' with 'blogs that glorify or promote anorexia, bulimia, and other eating disorders' (Tumblr, 2012a), and Pinterest claims to 'remove anything that promotes self-harm, such as self-mutilation, eating disorders or drug abuse' (Pinterest, 2017a). This reveals an intimate connection between platforms' public-facing policies and closed codes.

Pinterest also recommended content to me through email updates. It sent me 'popular Pins for you', 'Hip bones, Workout music and other topics you might love' and also matched me with a user:

Figure 5: A screenshot of an automated email from Pinterest.

I received similar and almost daily emails from Pinterest during my investigation, and although I had to carefully develop a methodology to find some pro-eating disorder content, it became almost inescapable once I was embedded in such spaces. These automated messages reveal important contradictions between platforms' policies against pro-eating disorder and self-harm content and their technologies, which are developed to create a personalized experience for each user. Platforms have not yet algorithmically reconciled their moral stances on eating disorders and self-harm, meaning they simultaneously push and deny problematic content to their users. Tufekci (2018) makes a similar point about YouTube's recommendation algorithm, an under-discussed aspect of the platform she argues contributes to the radicalization of its users.

Once I started behaving like a pro-ED user on Instagram and Tumblr, these platforms also started to recommend such content to me. Instagram, for example, allows users to save posts and create their own content collections. As users are currently not able to tell when you have saved their post (Instagram, 2017), I could save posts and behave like a pro-ED user without causing reactivity. Instagram then began recommending similar content to me through its Explore function. These findings reveal how recommendation systems work in direct opposition to platforms' other mechanisms of control, like PSAs. Recommendation algorithms like those discussed above raise important questions about the reasons why platforms problematize pro-ED. This form of hashtag moderation appears to be designed to protect new

users who are at risk of joining pro-ED and other such networks, rather than those who are already embedded within them. Moderating search results suggests that platforms are protecting users who are curious but as-yet unaffected by eating disorders - those who are at risk of *contagion* (Burke, 2006). Bell makes a similar argument:

> It is not simply that these texts are of concern to those against pro-anorexia, but also that seemingly unsuspecting or uncritical computer users/readers/viewers – who are assumed to be female – will be exposed to their infectious content. Moreover, women are seen as distinctly susceptible to this kind of media transmission. (2009, p.155)

If women can indeed '"catch" anorexic behaviours from looking at each other and images of other women' (Burke, 2006, p.316), these discourses – of contagion, infection, virality – help platforms to justify content moderation.

For users who are already situated within pro-ED networks, PSAs and hashtag bans are ineffective because they will not be seen. The hashtag is insufficient as a mechanism for moderation, just as it is insufficient to focus only on tagged pro-ED cultures, as researchers (for example Chancellor et al, 2016; Moreno et al, 2016; Seko and Lewis, 2016; Ging and Garvey, 2017) and press commentators have done so far. But it would be a mistake to assume that platforms want to remove all of this content. The success of platforms' algorithmic recommendations for delivering pro-ED content reveals the politics behind their interventions. Hashtag moderation is not a method for platforms to remove all pro-ED and other kinds of

content. Instead, it tells concerned parties that platforms are willing to intervene and attempt to solve a problem of which they are a part. It is also almost impossible to map the range of ways people immerse themselves in networked pro-ED cultures, such as the dark web, private accounts, private messaging, and ephemeral forms of communication like Instagram Stories and Snapchat. Despite this, the hashtag gets privileged as a *way of seeing* the relationship between eating disorders and social media. Algorithmic recommendations give users precisely the kinds of content that content moderators and their critics have so far missed.

Concluding remarks

This article **has explored the circumvention of hashtag moderation in online pro-ED communities. It has** put forward a case for paying closer attention to non-hashtag use on social media, and for recognizing the limitations of *only* talking about hashtags when we talk about pro-eating disorder content moderation. Of the 2612 Instagram and Tumblr posts I analyzed, only 779 included hashtags. Hashtags, it would seem, are not reliable indicators of where pro-ED and perhaps other kinds of content can be found. Non-hashtag use is an important communicative tool for the pro-ED userbase, likely because users recognize hashtags' vulnerability to interventions by platforms and concerned third-parties. Platforms' efforts to moderate hashtags, combined with subsequent press commentaries about their interventions and an over-reliance on tagged pro-ED posts in social and computer science research, have produced limited understandings of how such content actually circulates on platforms. This

paper has opened up discussions about untagged pro-ED posts and paints a fuller picture of users' **evasive** practices that go beyond the hashtag. The findings tell us that members of the pro-ED community are savvy: they often do not use hashtags and have devised a set of signals to indicate their identities and content as pro-ED in the absence of clear, tagged markers. These signals are deliberately obfuscating, meaning commercial content moderators (CCMs), who have only a few seconds to decide whether a flagged post should stay or go (Roberts, 2017a), might struggle to identify it as pro-ED. Some users also disavow a pro-ED identity in their profile biographies, using disclaimers of denial to reveal both the importance of the biographical space as a communicative tool and their efforts to reassure third parties that their account is unproblematic (even if, by the platform's standards, it is).

The findings also tell us that blanket bans on hashtags, a logic put forth by Instagram, do not work. But perhaps they are not intended to. This form of moderation appears to be designed to protect new users who are at risk of joining pro-ED and other such networks rather than those who are already embedded within them. Hashtag logics protect social media users who are curious but as-yet unaffected by eating disorders and who still rely on in-platform search engines to find new content. Once a user is embedded within such a network – albeit at the margins of social media – Instagram, Pinterest, and Tumblr's recommendation systems will continue to suggest pro-ED content to them. **This marginalization might be read as problematic, especially because eating disorders are socially stigmatized, but it offers somewhat of a compromise for those who want to turn to online communities. My analysis also reveals the complexities of the 'pro-ED' identity – it is not singular, and users' feeds**

comprise a range of discourses which include pro-eating disorder, pro-recovery, and not-pro-anything, from people helping users with their recovery to those requesting ana buddies, posting tips and playing fasting games. It is the shared codes and circumvention techniques that define this community of users, thus perhaps it is not possible nor wise to police ED-related content in a systematic way.

The hashtag has become a way of seeing a complex socio-technical phenomenon, but this logic misses the other important ways people engage with ED-related content on social media. Given the ethical difficulties of accessing patients with eating disorders to ask them about their relationship to social media (Lavis, 2015), future directions for research should aim to provide a deeper analysis of the pro-ED content hidden within social media's many communicative layers (Bruns and Moe, 2014). As reported cases of eating disorders are on the rise in the United Kingdom (Dugan, 2014), social media can provide us with a rich source of knowledge. As-yet unaddressed issues include: (1) analyses of users' comments on pro-ED posts, (2) a cross-platform analysis to understand any socio-technical variation between different pro-ED cultures, and (3) analyses of pro-ED users' self-representations, which are often enacted pseudonymously. Future research on pro-ED should move away from a reliance on tagged datasets (see Mitchell et al, 2015; Bruns et al, 2016; D'heer et al, 2017 for similar arguments) and aim to produce knowledge about social media users that might *only* be understood by analyzing untagged posts.

[1] I use the term 'pro-ED' throughout this article to capture the range of known eating disorders – anorexia, avoidant restrictive food intake disorder (ARFID), binge eating disorder, bulimia, eating disorder not otherwise specified (EDNOS), orthorexia, and others (NEDA, 2017). The online spaces I discuss are sometimes called 'pro-ana' (pro-anorexia), particularly in press discussions, but the users in these spaces do not only discuss anorexia.

[2] To run my experiment, I used the tags listed by the three platforms when they announced their ban on pro-ED content: thinspiration, probulimia, proanorexia (Instagram, 2012) anorexia, anorexic, bulimia, bulimic, thinspiration, thinspo, proana, purge, purging, promia (Tumblr, 2012a).

[3] After I conducted the research, Instagram started letting users 'follow' hashtags and show them posts from the hashtag collection in their main feed (Popper, 2017).

References

Bell M (2009) '@ the Doctor's Office': pro-anorexia and the medical gaze. *Surveillance and Society* 6(2): 151-162.

Bordo S (2003) *Unbearable weight: feminism, Western culture, and the body.* 10th anniversary edition. Berkeley: University of California Press.

Bruns A and Moe H (2014) Structural layers of communication on Twitter. In: Weller K, Bruns, A, Burgess, J, Mahrt, M and Puschmann, C (eds.) *Twitter and Society*. New York: Peter Lang, 15-28.

Bruns A, Moon B, Paul A and Münch F (2016) Towards a typology of hashtag publics: a large-scale comparative study of user engagement across trending topics. *Communication Research and Practice* 2(1): 20-46.

Bucher T (2017) The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication and Society* 20(1): 30-44.

Burke E (2006) Feminine visions: anorexia and contagion in pop discourse. *Feminist Media Studies* 6(3): 315-330.

Chancellor S, Pater JA, Clear T, Gilbert E, and De Choudhury M (2016) #thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. *CSCW '16 Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing*. Available at: http://www.munmund.net/pubs/cscw16_thyghgapp.pdf (Accessed 6 December 2016).

Cobb G (2017) 'This is not pro-ana': denial and disguise in pro-anorexia online spaces. *Fat Studies* 6(2): 189-205.

D'heer E, Vandersmissen B, De Neve W, Verdegem P and Van de Walle R (2017) What are we missing? An empirical exploration in the structural biases of hashtag-based sampling on Twitter. *First Monday* 22(2): DOI http://dx.doi.org/10.5210/fm.v22i2.6353.

Day K and Keys T (2008) Starving in cyberspace: a discourse analysis of pro-eating disorder websites. *Journal of Gender Studies* 17(1): 1-15.

Dias K (2003) The ana sanctuary: women's pro-anorexia narratives in cyberspace 4(2): 31-45.

Donath J (2007) Signals in social supernets. *Journal of Computer-Mediated Communication* 13(1): 231-251.

Dugan E (2014, 26 January) Exclusive: eating disorders soar among teens - and social media is to blame. *The Independent*. Available at: http://www.independent.co.uk/life-style/health-and-families/health-news/exclusive-eating-disorders-soar-among-teens-and-social-media-is-to-blame-9085500.html (accessed 22 January 2018).

Ferreday D (2003) Unspeakable bodies: erasure, embodiment and the pro-ana community. *International Journal of Cultural Studies* 6(3): 227-295.

Fleming-May RA and Miller LE (2010) 'I'm scared to look. But I'm dying to know': information seeking and sharing on pro-ana weblogs. *Proceedings of the Association for Information Science and Technology* 47(1): 1-9.

Friz A and Gehl RW (2016) Pinning the feminine user: gender scripts in Pinterest's sign-up interface. *Media, Culture and Society* 38(5): 686-703.

Gillespie T (2015) Platforms intervene. *Social Media and Society*, 1(1): 1-2.

Gillespie T (2017, 24 May) Facebook can't moderate in secret anymore. *Medium*. Available at: https://points.datasociety.net/facebook-cant-moderate-in-secret-any-more-ca2dbcd9d2 (accessed 5 April 2018).

Gillespie T (2018, forthcoming) *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*. London; New Haven: Yale University Press.

Ging D and Garvey S (2017) 'Written in these scars are the stories I can't explain': a content analysis of pro-ana and thinspiration image sharing on Instagram. *New Media and Society*: 1-20. DOI: 10.1177/1461444816687288.

Greenfield R (2012, 19 March) Pinterest has an anorexia problem now. *The Atlantic*. Available at: https://www.theatlantic.com/technology/archive/2012/03/pinterest-has-anorexia-problem-now-too/330315/ (accessed 14 June 2017).

Gregoire C (2012, 9 February) THE HUNGER BLOGS: a secret world of teenage "thinspiration". *Huffington Post*. Available at: http://www.huffingtonpost.co.uk/entry/thinspiration-blogs_n_1264459 (accessed 14 June 2017).

Hern A (2015, 20 May) Flickr faces complaints over 'offensive' auto-tagging for photos. *The Guardian*. Available at: https://www.theguardian.com/technology/2015/may/20/flickr-complaints-offensive-auto-tagging-photos (accessed 21 July 2017).

Instagram (2012, 20 April) Instagram's new guidelines against self-harm images and accounts. Available at: http://blog.instagram.com/post/21454597658/instagrams-new-guidelines-against-self-harm (accessed 24 May 2017).

Instagram (2017) How can I save posts I see on Instagram? Available at: https://help.instagram.com/1744643532522513 (accessed 13 December 2017).

Larsson AO (2015) Studying big data – ethical and methodological considerations. In: Fossheim H and Ingierd H (eds.) *Internet Research* Ethics. Oslo: Cappelen Damm Akademisk, 141–157.

Lavis A (2015, June 4) Jumping to blame social media for eating disorders is dangerous. *The Conversation*. Available at: https://theconversation.com/jumping-to-blame-social-media-for-eating-disorders-is-dangerous-42777 (accessed 13 December 2017).

Lewallen J and Behm-Morawitz E (2016) Pinterest or Thinterest?: Social comparison and body image on social media. *Social Media and Society* 2(1).

Light B, Burgess J and Duguay S (2016) The walkthrough method: an approach to the study of apps. *New Media and Society*: 1-20. DOI: 10.1177/1461444816675438.

Lingel J (2017) *Digital countercultures and the struggle for community*. Cambridge, MA: MIT Press.

Mitchell P, Bruns A and Münch F (2016) The (net)work of mourning: emotional contagion, viral performativity, and the death of David Bowie. Paper presented at *AoIR 2016: The 17th Annual Conference of the Association of Internet Researchers*. Berlin, Germany. Available at: http://spir.aoir.org (accessed 17 August 2017).

Moreno MA, Ton A, Selkie E and Evans Y (2016) Secret Society 123: understanding the language of self-harm on Instagram. *Journal of Adolescent Health* 58(1): 78-84.

National Eating Disorders Association (NEDA) (2017) By eating disorder. Available at: https://www.nationaleatingdisorders.org/learn/by-eating-disorder (accessed 12 December 2017).

Pinterest (2017a) Community guidelines. Available at: https://policy.pinterest.com/en-gb/community-guidelines (accessed 20 August 2017).

Pinterest (2017b) Self injury examples. Available at: https://www.pinterest.co.uk/pinterestpolicy/self-injury-examples/ (accessed 20 August 2017).

Popper B (2017, 12 December) Instagram gets more #interesting: the social network now lets you follow hashtags. *The Verge.* Available at: https://www.theverge.com/2017/12/12/16763502/instagram-hashtag-follow-new-feature-announced (accessed 13 December 2017).

Reynolds E (2016, 14 March) Instagram's pro-anorexia ban made the problem worse. *Wired*. Available at: http://www.wired.co.uk/article/instagram-pro-anorexia-search-terms (accessed 31 January 2018).

Roberts ST (2017a) Aggregating the unseen. In: Byström A and Soda M (eds.) *Pics or it Didn't Happen*. Munich: Prestel, 17-21.

Roberts ST (2017b, 8 March) Social media's silent filter. *The Atlantic*. Available at: https://www.theatlantic.com/technology/archive/2017/03/commercial-content-moderation/518796/ (accessed 28 September 2017).

Ryan EG (2012, 19 March) The scary, weird world of Pinterest thinspo boards. *Jezebel*. Available at: https://jezebel.com/5893382/the-scary-weird-world-of-pinterest-thinspo-boards (accessed 14 June 2017).

Schmidt J (2014) Twitter and the rise of personal publics. In: Weller K, Bruns A, Burgess J, Mahrt M and Puschmann C (eds.) *Twitter and Society*. New York: Peter Lang, 3-14.

Seko Y and Lewis SP (2016) The self - harmed, visualized, and reblogged: remaking of self-injury narratives on Tumblr. *New Media and Society*, 1-16. DOI: 10.1177/1461444816660783.

Suzor N (2016, 17 September) How does Instagram censor hashtags? *Medium*. Available at: https://digitalsocialcontract.net/how-does-instagram-censor-hashtags-c7f38872d1fd (accessed 29 July 2017).

Tufekci Z (2018, 10 March) YouTube, the great radicalizer. *New York Times*. Available at: https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html (accessed 5 April 2018).

Tumblr (2012a, 23 February) A new policy against self-harm blogs. Available at: https://staff.tumblr.com/post/18132624829/self-harm-blogs (accessed 2 June 2017).

Tumblr (2012b, 1 March) Follow-up: Tumblr's new policy against pro-self-harm blogs. Available at: https://staff.tumblr.com/post/18563255291/follow-up-tumblrs-new-policy-against (accessed 2 June 2017).

Yom-Tov E and boyd d (2014) On the link between media coverage of anorexia and pro-anorexic practices on the Web. *International Journal of Eating Disorders* 47(2): 196-202.