#thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities

Stevie Chancellor **Jessica Pater** **Trustin Clear** **Eric Gilbert** **Munmun De Choudhury**
School of Interactive Computing, Georgia Institute of Technology, Atlanta GA 30332
{schancellor3, pater, trustin}@gatech.edu, {gilbert, mchoudhu}@cc.gatech.edu

ABSTRACT
Pro-eating disorder (pro-ED) communities on social media encourage the adoption and maintenance of disordered eating habits as acceptable alternative lifestyles rather than threats to health. In particular, the social networking site Instagram has reacted by banning searches on several pro-ED tags and issuing content advisories on others. We present the first large-scale quantitative study investigating pro-ED communities on Instagram in the aftermath of moderation – our dataset contains 2.5M posts between 2011 and 2014. We find that the pro-ED community has adopted non-standard lexical variations of moderated tags to circumvent these restrictions. In fact, increasingly complex lexical variants have emerged over time. Communities that use lexical variants show increased participation and support of pro-ED (15-30%). Finally, the tags associated with content on these variants express more toxic, self-harm, and vulnerable content. Despite Instagram's moderation strategies, pro-ED communities are active and thriving. We discuss the effectiveness of content moderation as an intervention for communities of deviant behavior.

Author Keywords
Instagram; social media; lexical variation; eating disorder.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
Online connectivity has changed our experiences of health disorders, both for good and for bad. On one hand, the web provides a candid and emotionally supportive network for communities with socially stigmatized illnesses, e.g., depression [12,31]. On the other, online platforms have connected people in ways that can enable and amplify the destructive power of eating disorders [19]. Once socially or physically isolated, individuals with eating disorders can now connect with other sufferers online. Sometimes, these users connect in "pro-eating disorder" (pro-ED) communities that share content, advice, and provide social support for disordered or unusual eating choices as a reasonable lifestyle alternative [7]. Social sharing of such behaviors is dangerous not only for those with eating disorders but also represents contagion threats to those who do not currently have these conditions but may be vulnerable [7].

CSCW '16, February 27-March 02, 2016, San Francisco, CA, USA
 ISBN 978-1-4503-3592-8/16/02…$15.00
DOI: http://dx.doi.org/10.1145/2818048.2819963

Instagram is a photo-sharing social network founded in 2010. The platform is unique in that it does not have formalized community structures, like forums or private groups. Instead, communities form around more amorphous, public tags. In the case of the pro-ED community on Instagram, users cluster around tags relating to eating disorders (e.g., "anorexia", "proana").

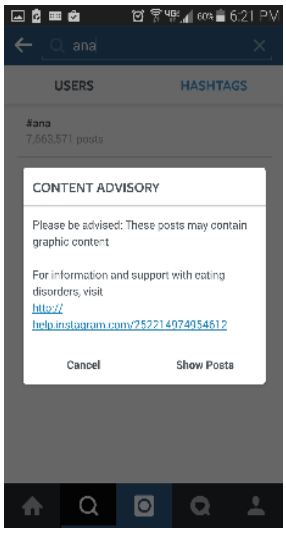

Figure 1. A content advisory is issued on searches for "ana".

Instagram, along with other social media platforms like Tumblr, has been challenged with the proliferation of such content. In response to media scrutiny in 2012 [32], Instagram began to publicly ban some of the most common tags associated with pro-ED [24] with the stated goal that such restrictions would discourage pro-ED content. Banned tags can still be used in posts, but posts will not be returned if a user searches for any of these tags. In addition, Instagram issues content advisories that serve as public service announcements on searches around eating disorder-related tags (Figure 1). We will refer to these practices by Instagram as "content moderation."

In response to such moderation, the pro-ED community has adopted tagging conventions to circumvent restrictions on accessing pro-ED content. One popular technique used by the community is adopting non-standard linguistic variants of moderated tags [10,13], what we call "lexical variants." These variants include adding or deleting characters in tags ("anorexiaa"), substituting letters ("thynsporation"), or deliberate misspellings ("anarexic") but keeping the semantics of the tag consistent.

In this paper, we investigate the adoption of lexical variation in tags used by the pro-ED community before and after Instagram began moderating pro-ED content. Our research is the first large-scale quantitative study that examines the effectiveness of such content moderation over time. This study has four aims – to:

- Study the emergence and evolution of lexical variations of moderated tags, focusing on the period following changes to Instagram's community policy in 2012.
- Explore how communities adopting lexically variant tags change over time.
- Quantify how the greater community engages with the content associated with lexical variants.
- Examine the topical context of lexical variants and contrast it with that of the moderated tags.

Our study uses 2.5 million pro-ED Instagram posts from half a million users, shared between 2011 and 2014. After content moderation, Lexical variants emerged for all 17 pro-ED tags that underwent initial moderation in 2012. Many lexical variants were adopted by the pro-ED community following the enforcement of content moderation – an average of almost 40 variants emerged corresponding to each moderated tag. Further, engagement on these variant tags through 'likes' and comments was 15-30% higher compared to the original moderated tags. While the size of communities adopting the variations was often smaller and largely non-overlapping with the moderated tags, certain lexical variations reached dramatic sizes (2 to 40 times larger) relative to the initial tag. In fact, lexical variants of tags with content advisories grew by 22% following Instagram's moderation of pro-ED content. We also find that the content associated with lexical variants reflected heightened vulnerability to self-harm and isolation from the greater community of sufferers of eating disorders on Instagram.

Our quantitative investigation suggests that Instagram's current moderation practices are not effective at dispersing the pro-ED community or in controlling the propagation of pro-ED behavior on the platform. Moderation might in fact be amplifying the destructive power of pro-ED posts. Our research offers insights into avoidance mechanisms of platform-imposed moderation for pro-ED communities. These insights can inform whether moderation is a viable intervention mechanism for pro-ED, and if not, how to craft more effective ways to help vulnerable communities. Beyond eating disorders, we hope our findings to encourage deeper discussions around the role of policing and moderating content to curb deviant behavior.

Privacy and Ethics. In this paper, given the sensitivities around the topic of investigation, we use only public data collected via Instagram's official API. We also do not report activities of specific users, their postings, or any information that could potentially be personally identifiable. Since our methods involved no interaction with the users and public data was used, our work did not qualify for institutional review board approval.

PRIOR WORK AND RESEARCH QUESTIONS

Eating disorders are a group of psychosocial disorders characterized by abnormal behaviors in eating and exercise. These disorders negatively affect both mental and physical health and include symptoms of binging, restricting, purging, obsessing, or other forms of extreme emotional responses to the procurement and ingestion of food, exercise, or body modification [38]. Anorexia nervosa and bulimia nervosa are the two well-known eating disorders specified in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). In the US, it is estimated that roughly 20 million women and 10 million men suffer from an eating disorder at some point in their life [41]. Eating disorders have the highest mortality rate of any mental disorder [2].

Online Communities and Eating Disorders

Previous research has examined the content of pro-ED communities on blogs and related social networks [7,19,27,34]. Most of these studies use qualitative coding schemes to analyze content. They categorize the various support structures these postings offer community members [7,19], analyze search patterns for pro-ED content [27], and look at the ethical situations surrounding pro-ED in online communities [34]. One preliminary study examined Tumblr as a content portal for pro-ED behavior [9], but very few have deeply delved into the membership and structure of these user-generated and amorphous networks. Quantitative examinations of pro-ED communities are limited aside from the work in [42], which examines images and community dynamics. With the exception of [11], few studies have explored pro-ED communities that have emerged on social media, such as Tumblr and Instagram.

Instagram has unique affordances that make it an appropriate platform to examine pro-ED behavior. The demographics of Instagram and the demographics of the common eating disorder patient are similar. Approximately 70% of Instagram users are female and roughly half of all Internet-using young adults (12-18 years) are using Instagram [14] compared to typical eating disorder patients who are women ages 15-24 [34]. In addition, the visual nature of Instagram itself may predispose pro-ED communities to stay. A 2010 study found that 69% of American girls five to 12-years old say pictures influence their concept of ideal body shape and 47% report that images make them want to lose weight [29]. Further, the use of tags on Instagram makes the social network a likely target for deviant behavior. Pro-ED communities are often hidden in plain sight; that is, their activities are generally cut off from the mainstream activity of users but are easily accessible by searching for related tags/keywords.

Root tag	#Var.	Lexical Variants
ana	9	anaa, anna, anaaa, anaaaa, annaa, annna, annaaa, anaaaaaa, anaaaaa
anorexia	99	anorexic, anorexie, anoressia, anorexi, anorexia, anorexique, anorexica, anorectic, anorexia, anoretic
anorexianervosa	62	anorexianervousa, anorexianerviosa, anoressianervosa, anorexianevosa, anorexicnervouse, anorexianevrosa, anorexicnervosa, anorexinervosa, anorexianervose, anorexianervosia
bonespo	6	bonespoo, bonespoooo, bonespooo, bonesspo, bonesporation, bonessspo
bulimia	49	bulimic, bulima, bulimie, bulimi, bulimia, bulimica, bulimc, bulimiaaa, bulimic, bulimist
eatingdisorder	97	eatingdisorders, eatingdissorder, eatingdisoder, eatingdis, eatingdisorter, eatingdisoreder, eatingdisorde, eatingdisorderrr, eatingdisordered, eating_disorder
mia	3	miaa, miaaa, miaaaa
proana	11	proanaa, proanna, proanaaa, proanaaaa, pro_ana , prooana, proaana, proannaa, proanaaaaaa
proanorexia	1	proanorexic
probulimia	1	probulimic
promia	4	promiaa, promiaaa, promiaaaa, proomia
secretsociety	55	secret_society123, secretsociety_123, secretsociety123, secret_society, secret_society_123, secretsociety1234, secret_society1234, secret_society124, thesecretsociety, secretsociety124
skinny	18	skinny, skiny, skinny, skinny, skinny, skinnyyy, skini, skynni, skinnyyyy, skinnnyyy
thighgap	107	thighgaps, thygap, thighgapp, thigh_gap, thightgap, thyghgap, thighgappp, thegap, thigap, thighgapss
thin	9	thyn, thinn, thynn, thinnn, thynnn, thiin, thiiin, thinnnn, thyyn
thinspiration	101	thynspiration, thinsperation, thinspire, thynsporation, thinsporation, thinspiring, thinspirationn, thinspirational, thinsparation, thynsperation
thinspo	40	thinspoooo, thynspo, thynspoo, thynspooo, thinspoo, thinspooo, thinspooooo, thynspoooo, thinnspo, thinspooooo
Total root tags		17
Total variant tags		672

Table 1. Root tags, total number of variants in each tag chain, and 10 most frequent lexical variants.

Social Media Content Moderation

Various social media platforms moderate and remove content for legal or political reasons [35]. Some decisions are driven by the legalities of the country where they operate. All US social media sites, for instance, ban child pornography as well as content that commits copyright infringement. Platforms may also abide by censorship standards imposed by governments. Several studies have examined attributes and impacts of Chinese censorship on social media [3,17,22,25,28] or social media censorship more broadly [38]. The impact of censorship on information sharing, propagation, accessibility, and journalistic practices was discussed in [41] in the context of socio-political protests in authoritarian regimes.

Beyond these, social media sites may also choose to remove content for social, moral, or community reasons. Facebook, Instagram, Tumblr, and YouTube moderate general pornographic content, and Facebook bans hateful and violent speech [18]. In the context of eating disorders, while there is no obvious moderation on eating disorder-related content on Twitter, YouTube, or Reddit, other platforms like Pinterest and Facebook more rigorously ban tags and terms around it [10]. Tumblr issues public service announcements on searches on pro-ED terms and Instagram has banned several pro-ED tags and provides content advisories on others [21]. Instagram's regulation of pro-ED content falls into this broad social space and our research presents one of the first quantitative insights into the effectiveness of platform-enforced moderation of pro-ED behavior.

Language Variation in Social Media

Language variation has been of great interest to researchers for many decades. Social media has become a popular medium to explore, model, and detect a variety of linguistic variations [15] and to understand the emergence of linguistic conventions [23,26] over contexts such as geography, demographics, and style.

Automated detection of language variation has been methodologically challenging. Most quantitative work in this area focuses on identifying a hand-curated small set of variable pairs (actual term and variant term) and measuring their frequencies, except [15] which uses a latent variable model for the purpose. *Lexical variation*, in particular, is challenging to measure because it is often difficult to assess what could be in the possible universe of all variants – social media is known for use of non-standard terms (*smh, jk, ima, wassup*). Lexical variants often do not follow any regular, expected patterns, conventions or rules as they deviate from their actual terms.

Note that the precise definition of lexical variation in the literature is varied and often depends on the specific research question under investigation. Eisenstein et al. [15] defined lexical variation to be the differences in the use of different linguistic constructs (e.g., words) and proposed

methods to detect how such constructs vary with geography. Bamman et al. [4] extended these investigations of lexical variations in Twitter to gender identity. Schwartz et al. [33] found differences in lexical constructs across populations on Twitter. The lengthening of sentiment words as a form of lexical variation was examined by Brody and Diakopoulos in [8]. In this paper, we address these issues by developing a lexical variation detection method that combines automated natural language processing techniques with human annotations. Further, prior literature did not focus on the unique circumstances of adoption of lexical variation to engage in deviant behavior – our contributions lie in examining the nature of changes in one particular deviant behavior community, pro-ED, following the adoption of lexical variation.

In our work, we define lexical variation in the light of tagging strategies adopted by the pro-ED community in the aftermath of content moderation enforcement by Instagram.

Research Questions

In light of the above prior work and our focus on the social media Instagram, we examined lexical, behavioral, and topical changes associated with the emergence of lexical variation in Instagram's pro-ED communities. We address the following research questions:

RQ 1. (*Lexical Changes*) How do lexical variations of moderated pro-ED tags evolve over time?

RQ 2. (*Behavioral Changes*) How does posting activity and support manifested in pro-ED posts evolve as lexical variations are adopted?

RQ 3. (*Topical Changes*) What topics characterize posts with lexically variant tags, and how do they contrast to the set of posts with the moderated tags?

DEFINITIONS, DATA, AND METHODS

Defining Lexical Variation

Because there is no standard definition or a set of "gold standard labels" on tag variations in analyses of pro-ED communities, we offer a definition for lexical variation for this paper. We began our investigation with anecdotal observations made in popular media on this topic, e.g., "thinspoo" was identified to be a variation that emerged following moderation of "thinspo" [10, 13]. Variations that emerged out of moderated tags included lexical additions, deletions, substitutions, or permutations of characters. However, we noticed that these variant tags kept similar semantic meaning and structure. For example, "anatips" and "anaaaaa" are both tags with Levenshtein edit distance of 4 [30] with respect to the moderated tag "ana," have additions and permutations, and could, with traditional metrics [15], be considered variants. However our qualitative observations indicated "anatips" and "anaaaaa" are used for different purposes – the former tag for gathering advice on the maintenance of anorexic lifestyle, while the latter as a description of anorexia. As also observed by [8], standard lemmatization methods or spell-correction techniques that are based on edit distance were therefore not appropriate for selecting our initial set of variants for the moderated tags.

Based on these observations, we offer a set of general rules to define lexical variants. We consider a tag (t_j) to be a "lexical variant" of another tag (t_i) if:

1) t_j is lengthened by repeating any of t_i's characters or other newly added characters.
2) Some of the characters in t_i are permuted to create t_j.
3) Some of the characters in t_i are eliminated to create t_j.
4) One or more characters not in t_i (including alphanumeric characters) are added to or substituted in t_j.
5) A combination of the above criteria is used to create t_j

These rules are relatively more restrictive compared to those used in existing literature on language variation [15]; however, they allow us to define a form of variation in which the *semantic structure is unchanged*, and the variation is limited to the lexical elements of a tag. These rules provide a much-needed scope to examine tag variants in pro-ED communities.

Based on these criteria, we formally define the following two terms that are used throughout the paper:

a. Root tag: A tag t_i which serves as a basis for us to discover and understand lexical variations of tag use, is referred to as a "root tag". We assume the root tag t_i to be the canonical form of lexical variants t_j. Root tags are the original version of a tag; in our case they are the tags which underwent moderation by Instagram in 2012.

b. Tag chain: The set of all the lexical variants t_j of each root tag t_i, as obtained through the rules above.

Data Collection

We used Instagram's official API[1] to collect over eight million public posts in the pro-ED space. However, Instagram's API does not return any posts when queried with banned tags. Our data gathering occurred in three steps to work around this limitation: sampling for pro-ED tags that co-occurred with banned tags in posts, a larger data collection, and creating a candidate pro-ED post set by removing noisy, ambiguous or irrelevant content.

First, we obtained post counts for nine "seed tags"[2] known to be related to eating disorders [11]. We collected all posts for each of these nine tags over 30 days. The resulting sample contained 434K posts with 234K unique tags. We used this to establish co-occurrence probabilities for all tag pairs. Sorting tags in order of decreasing probability of co-occurrence identified 222 tags with at least a 1% occurrence

[1] http://instagram.com/developer/

[2] Seed tags include: "ed", "eatingdisorder", "ednos", "ana", "anorexia", "anorexic", "mia", "bulimia", and "bulimic".

rate, collectively associated with tens of millions of posts dating back as far as January 2011.

With this co-occurrence tag list, we then excluded tags that were not related to eating disorders. This step needed to be done manually to find tags semantically related to eating disorders, not the closely related communities of mental health and eating disorder recovery. Our selection criteria excluded tags that were broad enough to be used by the general population or be applied to another mental disorder. Tags that were too broad include "fat", "beautiful", and "whale" as well as tags related to other mental disorders like "anxiety" and "depression." We also excluded any obvious recovery tags like "anarecovery" – this is because we wanted to specifically focus on the behavior of the pro-ED community that promoted/reinforced eating disorders. This reduced the dataset from 222 tags to 72 known eating disorder tags. Next, we collected our dataset, which contained all available posts tagged with any of these 72 tags from November 2014 as far back as January 2011. This dataset contained over 8 million posts.

Finally, we created a candidate set of posts from this raw set that we confirmed to be related to pro-ED behavior. We removed any posts with three tags ("mia", "ana", and "ed") that did not also contain another tag from our list of 72 tags.

Tag Chain	Status	Posts (All)	Posts (Root)	Posts (Variants)
ana	Advisory	1654530	1617455	37075 (↓)
anorexia	Advisory	2137204	1333694	803510 (↓)
anorexianervosa	Advisory	121037	116125	4912 (↓)
bonespo	Advisory	35371	34587	784 (↓)
bulimia	Advisory	1169581	773704	395877 (↓)
eatingdisorder	Advisory	748204	683115	65089 (↓)
mia	Advisory	964083	948164	15919 (↓)
proana	Banned	17593	13170	4423 (↓)
proanorexia	Banned	365	303	62 (↓)
probulimia	Banned	219	168	51 (↓)
promia	Banned	4470	4124	346 (↓)
secretsociety	Banned	332287	8166	324121 (↑)
skinny	Advisory	521933	519852	2081 (↓)
thighgap	Banned	88457	14572	73885 (↑)
thin	Advisory	304684	293318	11366 (↓)
thinspiration	Banned	68474	21254	47220 (↑)
thinspo	Banned	206473	62380	144093 (↑)
Total posts (roots + variants)				2416272
Mean change in #variant posts compared to #root posts				-70%

Table 2. Summary statistics of the tag chains as well as the moderation status of each root tag. Downward arrows indicate chains where moderation results in fewer posts with variants. Upward arrows indicate an increase.

Qualitative observation showed that these tags were strongly associated with the pro-ED community on Instagram but were also commonly used as first names or for referencing popular celebrities ("ed" for Ed Sheeran). This filtering created a dataset of 6.5 million posts.

Identifying Root Tags

Following our data collection, we devised an approach to identify a set of root tags relevant to the pro-ED community that underwent moderation. Instagram does not publish a centralized resource for all moderated tags, and third-party sources on the same are scarce and only include banned tags, not the ones with content advisories. To overcome these limitations, we first constructed a tag usage frequency distribution to identify frequent tags in all crawled posts. For the top 200 tags, two researchers who are Instagram users manually checked for bans or content advisories on these tags. This produced 17 tags that uniquely characterized pro-ED content and have either a ban or content advisory placed by Instagram. These 17 tags served as our set of moderated root tags on which we base our ensuing analyses of lexical variation.

Identifying Lexical Variants

Finally, we identified lexical variants of our 17 root tags in our dataset. For the purpose, we constructed a matching regular expression in line with the rules stated earlier in the section "Defining Lexical Variation". Our regular expressions were intentionally broad to capture any potential variants. This returned a rough list of potential variants for our root tags.

Two researchers familiar with Instagram and pro-ED content independently participated in a binary rating task to remove spurious and unrelated variants (recall the "anatips" and "anaaaaa" example from before). Each candidate variant was rated as "yes" or "no" – "yes" indicated a valid variant, whereas "no" did not. The researchers then pooled their responses, and Cohen's κ of interrater agreement was observed to be very high (.98). Our analysis uses variants where both raters agreed "yes."

Table 1 gives a list of the 17 root tags along with the number of lexically variant tags obtained through the method above (672 total). We also show the top 10 lexical variants found to be most frequent in our pro-ED post set. In Table 2, we further report the moderation status of these 17 tags and the total posts for the root and all variant tags. As Table 2 shows, different styles of tag variants, ranging from arbitrary word lengthening (e.g., "thinspoo") to permutations of letters in a word (e.g., "anoreixa"), to elimination and addition of arbitrary characters (e.g., "bulimkc") characterized the pro-ED communities following moderation.

Our final dataset contained all posts from our candidate set that were tagged with any moderated tags and any of their

lexical variants. It has more than 2.4 million posts and had over half a million users.

RESULTS

RQ1 (Lexical Changes): Evolution of Lexical Variations

To answer RQ1, we investigate the pattern and evolution of lexical variations associated with the root tags. Levenshtein edit distance between two words is the minimum number of single-character edits (i.e. insertions, deletions, or substitutions) required to change one word into the other [30]. In Figure 2 we show scatter plots of the Levenshtein edit distance for variants of "anorexia", "eatingdisorder" and "thighgap" over time.

Figure 2 shows that all chains, edit distance of a variant tag compared to the root *increases* over time – linear trend (least squares) fits to the edit distances of all variants for "anorexia", "eatingdisorder", and "thighgap" yield R^2=.2 (p=.002), R^2=.27 (p=.001) and R^2=.34 (p=.0005) respectively. As newer variants emerged over time after the root, they were increasingly more syntactically distinct ("thighgap" → "thyghgapss"). The mean and maximum edit distance over all variants per root tag are reported in Table 3 – we note that the mean edit distances are higher than one and the maximum at nine characters, indicating considerable lexical variation in the tag chains. However, it is important to note that there is no positive correlation between the mean edit distance of the variants and activity (i.e., volume of posts, ref. Table 2) on the corresponding moderated tag (Pearson correlation coefficient ρ=.045; p=.19). For instance, mean edit distance is highest for the "thinspo" and "ana" tag chains, however lower for "anorexia" and "bulimia"; however, the latter two have some of the largest proportion of posts in our data. This shows that the increased dispersion in lexical elements (indicated by high edit distance in the variants) is likely not an artifact of the moderated tag being a more popular tag in the pro-ED community.

Tag chain	Max. edit dist.	Mean edit dist.	Momentum
ana	5	2.556 ±1.257	1.281
anorexia	7	1.939 ±1.043	1.285
anorexianervosa	8	1.629 ±1.096	1.192
bonespo	6	2.500 ±1.708	1.367
bulimia	7	1.755 ±0.980	1.203
eatingdisorder	5	1.629 ±0.778	1.156
mia	3	2.000 ±0.816	1.750
proana	5	2.000 ±1.348	1.492
promia	3	1.750 ±0.829	1.278
secretsociety123	6	2.255 ±1.239	1.321
skinny	4	1.722 ±0.870	1.221
thighgap	5	2.084 ±1.006	1.218
thin	3	1.889 ±0.737	1.188
thinspiration	6	2.307 ±1.318	1.396
proanorexia	1	1.000 0.000	-
probulimia	1	1.000 0.000	-
thinspo	9	3.125 ±1.952	1.383
Mean momentum			1.3

Table 3. Variation patterns among tags in a chain, with respect to the root. Momentum indicates the rate of change of edit distance of variants over time of their emergence. "proanorexia" and "probulimia" each had one variant, so there was no momentum measured for these tag chains.

In each chain, we further define a rate of change metric momentum [23], given as: $(1/N)\Sigma_i(e(t_i) \,/\, e(t_{i-1}))$, i.e., the mean ratio between edit distance of the i^{th} tag t_i to the tag t_{i-1} appearing in the time slot before it, where N is the total number of variant tags corresponding to a root. All 17 tag chains show increased edit distance momentum of the variants with mean momentum of 1.3 across all chains (a value of 1 would indicate the rate of change is constant). Interestingly, based on a Mann-Whitney U-test, there is no statistically significant differences between the edit distance momentum of variants of banned tags and those of the advisory tags (p=.35). We conjecture the pro-ED community adopts increasing lexical variance in their tags to avoid In-

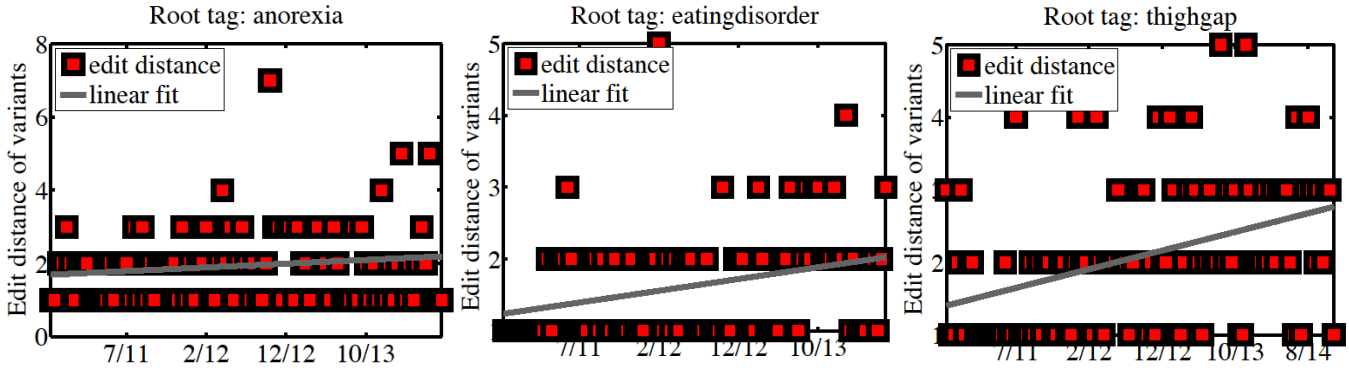

Figure 2. Changes in Levenshtein's edit distance with emergence of newer lexical variations over time – shown for "anorexia", "eatingdisorder" and "thighgap" tag chains. Each data point in the scatter plot corresponds to the edit distance of a particular variant at a certain point in time.

Tag chain	Root users	Variant users	Overlap (%)
ana	87575	2792 (↓)	2.12
anorexia	86631	57837 (↓)	39.06
anorexianervosa	5156	547 (↓)	4.81
bonespo	2107	115 (↓)	2.80
bulimia	49468	25758 (↓)	36.61
eatingdisorder	40605	9622 (↓)	9.11
mia	53880	684 (↓)	0.97
proana	2338	355 (↓)	3.59
proanorexia	24	9 (↓)	8.33
probulimia	10	1 (↓)	10.00
promia	672	51 (↓)	1.79
secretsociety	852	15215 (↑)	65.73
skinny	55639	564 (↓)	0.66
thighgap	973	5931 (↑)	5.86
thin	27386	865 (↓)	2.25
thinspiration	2919	3534 (↑)	17.71
thinspo	9304	9289 (↓)	17.79
Total unique users (roots + variants)			496498
Mean change in #variant users from #root users			-68%

Table 4. Number of users who used the root tag, all variants, and the percentage overlap, and their overlap. Down arrows are tag chains where the number of variant users decreased after moderation, whereas up arrows indicate an increase.

stagram's moderation of tags, beginning with closer variants to the root tag and becoming increasingly complex.

RQ2 (Behavioral Changes): Posting Activity & Support

In RQ2, we explore temporal changes in posting activity, users, and engagement/support around root and variant tags.

Comparing Activity

Figure 3 shows the changes in normalized proportions of posts that correspond to six moderated root tags and the same for three of their most common variants. To determine this normalized proportion of posts, we divided the total number of users who posted on a root tag or any of its lexical variants by the number of users that posted on any tag during the same time slot. Normalizing posts was necessary to prevent effects of overly active users as well as to circumvent disproportionate distribution of posts obtained from Instagram over the course of our three yearlong analysis. Our time slots were one week.

After changing community policies and introducing content moderation in April 2012, posting activity changed in ways we consider both varied and surprising. For the banned tags ("thighgap," "thinspo," and "thinspiration"), the proportion of posts sharply drops when Instagram reported changing its community policies. This is consistent across the other banned tags (not shown for brevity) – the use of banned tags decreased 13-78% after April 2012 (mean 52%).

However, for root tags with content advisories, we see a surprising *increase* in the proportion of posts after the policy change ("ana", "mia", "eatingdisorder"). This increase ranges between 9 and 37% (mean 22%). The emergence and substantial adoption of lexical variant tags only happens after April 2012. While a causal effect may not be directly derived, we believe that this shows a deliberate strategy by the pro-ED community to circumvent content moderation policies and to continue to organize and sustain themselves. Next, while lexical variants did emerge for the moderated tags, in some cases, the proportion of posts on variants is lower than the posts on the root tag (Table 2). In fact, on average there is a 70% decrease in proportion of variant posts compared to that of the root tag posts. This shows that Instagram's moderation policy did reduce activity on these tags. However, certain tag chains also increase in size (in terms of posts) through the adoption of lexical variants – e.g., "secretsociety" increases by more than 4000%, "thighgap" by 500%, and "thinspo" more than 200%. This increased activity shows that the pro-ED community continues to thrive even though overall participation dropped on some tags.

Likes						
Tag Chain	**Mean (Root)**		**Mean (Variants)**		***z***	
eatingdisorder	53	±55.28	44	±72.87	-36.21	***
mia	44	±46.37	56	±46.42	32.79	***
thighgap	36	±39.02	52	±49.00	38.55	***
thinspiration	31	±26.35	58	±57.86	64.12	***
thinspo	33	±34.47	53	±50.58	87.16	***
Change in #likes in variant posts vs. root posts					30.6%	
Comments						
Tag Chain	**Mean - Root**		**Mean – Variant**		***t* Stat.**	
eatingdisorder	2	±4.80	2	±4.01	-23.76	***
thighgap	1	±3.05	2	±3.97	27.85	***
thinspiration	1	±3.01	1	±3.62	24.50	***
thinspo	1	±3.22	2	±3.95	38.54	***
Change in #comments in variant posts vs. root posts					15.1%	

Table 5. Engagement (likes, comments) on the roots and their variants. Tag chains with most significant change in mean likes and comments are shown. Statistical significance is tested based on Mann Whitney U-tests. Bonferroni correction (α/17), where α=.05 (*), .01 (***), and .001 (***), is adopted to control for familywise error rate.

Comparing Users and Support

Next, we examine the volume of unique users associated with the root tags and their variants as well as the Jaccard similarity/overlap of users between the two (Table 4). In general, there are some tag chains where there is considerable overlap of users between the root tags and adopters of their variants (e.g, "bulimia," "secretsociety"). However, most tag chains have little overlap (e.g., "ana," "thighgap"). We believe this shows a shift in users who adopt these variations to overcome moderation restrictions enforced by Instagram. It also implies that adoption of lexical variation in tag usage might be an intrinsic individual characteristic;

More freq. w/ roots	LLR	More freq. w/ variants	LLR	Equally freq.	LLR
ednos	5.52	insecure	-6.62	depressed	.72
feelugly	5.19	cutting	-6.44	mentalhealth	.55
starve	4.87	loathemyself	-6.43	tired	.93
anamia	4.63	killingmeinside	-6.37	worthless	.49
anatips	4.15	lifeispointless	-5.91	life	.20
anaaccounts	4.09	bloodsecret123	-5.75	hate	.19
disappeared	4.01	nobodylikesme	-5.57	perfection	.04
darkangel	4.01	skinnyplease	-5.28	dead	.71
purge	3.89	trigger	-5.24	sad	.70
thinstagram	3.82	selfhate	-5.17	blithe	.79

Table 6. Top 10 tags co-occurring with roots and variants with the highest, lowest and near zero log likelihood (LLR).

that is, the users likely to embrace this strategy are perhaps a small fraction of those who use the root tags. Alternatively, it may also indicate the propensity of a certain segment of the pro-ED community to adopt the lexical variations in their content sharing, perhaps to avoid discoverability more broadly, build and maintain social cohesion, and to even "hide in plain sight" following the enforcement of the moderation policy.

To compliment this analysis, we examine how the pro-ED community engages and supports posts in the root and variant tags. To measure engagement and support, we use mean 'likes' and mean comments on root posts and variant posts (Table 5). There is a statistically significant increase in likes and comments for most variants when compared to the base tags. The mean number of likes in variant posts is higher by 30% compared to the root posts, while comments are 15% higher in variants (statistically significant through Mann Whitney U-tests). Despite a drop in the user base in some tags and the content moderation efforts of Instagram, there is continued support in the pro-ED community on lexical variants.

RQ3 (Topical Changes): Comparison of Topical Context

Finally, in RQ3, we investigate how the context of root tag use in posts differs from posts containing variant tags. We consider the context of use to be the tags that co-occur with roots or variants in posts. In our data, there are 194,421 tags that co-occur with roots three or more times, while 225,282 tags co-occur with variants three or more times. Before we could compare topical content, we determined that the two sets of tags are considerably different. Mean normalized mutual information (NMI) between the two co-occurrence tag distributions is .32 (high NMI implies high correlation), and a Mann-Whitney U-test notes this difference to be statistically significant (z=-2.93; p=.002). Further, the frequencies of co-occurrence of the tags with roots and variants are also different – Kendall's τ between the frequency distributions of the two sets is .28.

To explore these differences, we report a sample of 10 tags with lowest and highest log likelihood ratios between the two sources (excluding co-occurring tags that are roots or variants themselves (Table 6)). The Log likelihood ratio of a tag t_i is computed as: $\text{LLR}(t_i) = \log(P(t_i|\{\text{roots}\})/P(t_i|\{\text{variants}\})$, i.e., a measure proportional to the ratio between probability of co-occurrence of t_i with any of the

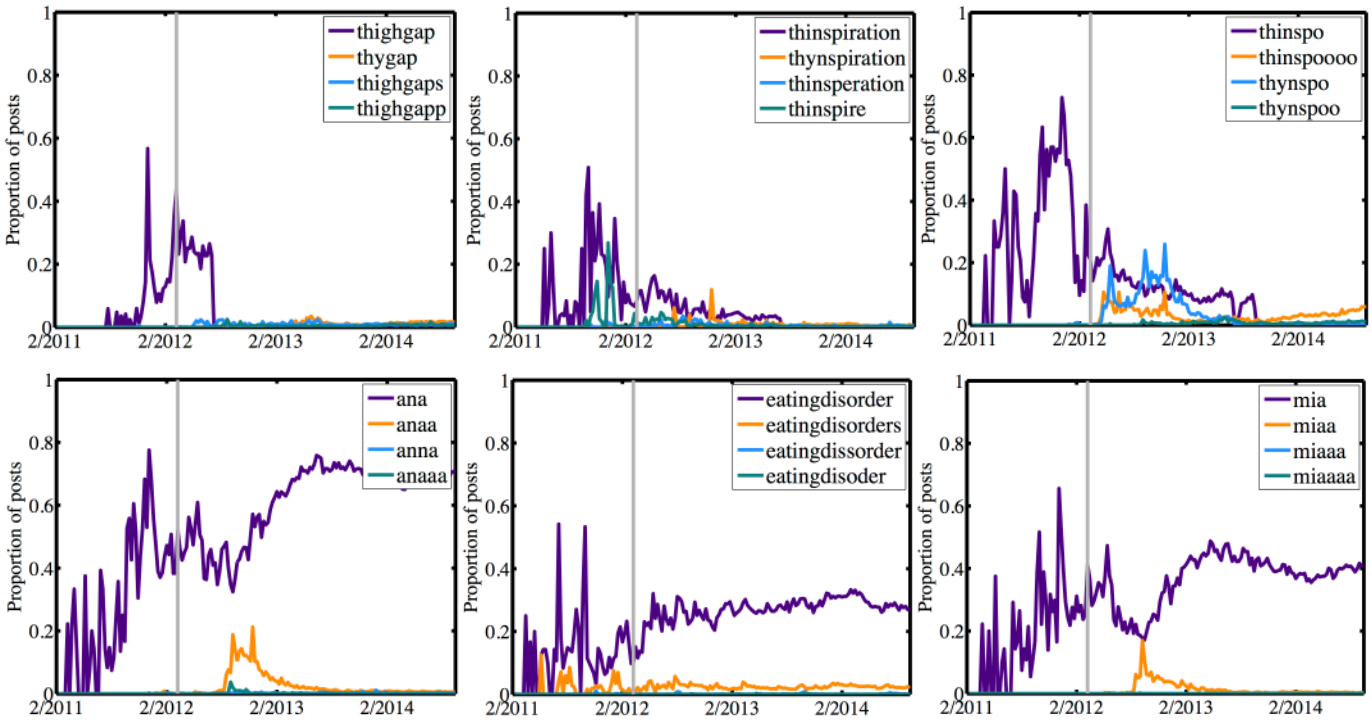

Figure 3. Normalized proportion of weekly posts for six root tags and their corresponding three most frequent variants over time. The vertical grey lines indicate time when Instagram publicly reported change in its community policies (Apr 2012).

roots to the probability of its co-occurrence with any of the variants. Large positive values of LLR imply that a tag more frequently co-occurs with root tags, while large negative values of LLR imply that it co-occurs more with the variants. A value of zero LLR implies it is equally frequent in both sources.

As shown in Table 6, there are considerable differences between the tags co-occurring with roots and those co-occurring with variants. Are these systematic themes that encompass these co-occurring tags of roots and variants and which are likely responsible for the observed differences?

Clustering Co-occurrence Tags. To answer this, we detect clusters of topics in an unsupervised manner in the set of tags co-occurring with the roots and those co-occurring with the variants. We use the normalized spectral clustering algorithm [35] on two graphs constructed out of the two sets. For instance, the root co-occurrence tag graph $G_r(V,E)$ comprises the tags t_i as nodes, such that t_i co-occurs with one of the root tags in a post and e_{ij} is in E if the tag t_i has co-occurred with the tag t_j at least five times in posts containing a root tag. The algorithm partitions the data by mapping the original space to an eigen space. Spectral clustering has been used in prior work to partition graphs [34] and is robust because it does not make any assumptions on the form of the clusters. This algorithm works well when the graph is sparse, which is the case with both of our tag co-occurrence graphs.

Extracting Themes in Co-Occurrence Tag Clusters. To examine the most dominant themes in the tag co-occurrence graphs, we analyze the two clusters corresponding to the first two eigenvalues of the Laplacian matrix given by spectral clustering (Table 7). Clustering algorithms, like spectral clustering, do not associate descriptive thematic elements with inferred clusters, so we incorporate human annotations into our analysis.

Two researchers familiar with pro-ED social media content and Instagram validated the set of tags in these clusters. They used an open coding approach to develop a codebook and extracted descriptive topical themes for the clusters (Cohen's κ was observed to be .84; we also tested for overall marginal homogeneity using the Stuart-Maxwell test [6], which is often useful in analyzing interrater agreement). During codebook development, the two annotators referred to prior literature on content characterization relating to eating disorders [7,42]. In Table 7, we report a sample of the 15 most frequent tags in each of the two clusters for the root and variant cases.

Themes in Root and Variant Tag Clusters. The clusters of tags that appear with root tags depict negative emotions and feelings known to be associated with pro-ED. The first cluster of tags co-occurring with root tags depicts expression of sadness and pain ("alone," "alwayssad," "broken,") and attributes of eating disorder and anorexia ("pain," "anaxiety," "sadstagram"). The pro-ED experience is associated with introversion, avoidance, and negative experience of social relationships – attributes supported by the use of the tags in this cluster [42]. The second cluster is associated with thinness and body image depiction where users describe physical attributes of their body ("collarbones," "hipbones"). Tags like "nofood," "mustbesmaller," and "skinnyangels" indicate the desire to practice the pro-ED lifestyle by suggesting unusual dieting strategies and emotionally justifying pro-ED as a legitimate choice.

Tags co-occurrent w/ roots		**Tags co-occurrent w/ variants**	
Topic I	**Topic II**	**Topic I**	**Topic II**
alone	bodycheck	suicide	smoke
alwayssad	nofood	selfharrrrm	failure
lifesucks	bones	selfmutilation	depression
pain	flatstomach	cutaddict	depressedquotes
unhappy	collarbones	cuts	deadinside
emptyfeeling	skinnyangels	harmingmyself	notgoodenough
anaxiety	thinstagram	scar	addiction
broken	mustbesmaller	razor	wishiweredead
emogirl	fat	bloodsecret123	abandon
sad	tiny	blades	paranoid
sadstagram	assbutt	cutting	callmemistaken
sadsmile	fatty	beautifulpain	useless
anxiety	hipbones	slicemywrists	letmeleave
sorry	beautiful	blood	lost
im_not_okay	pale	die	crying

Table 7. 15 most frequent tags in two dominant clusters extracted from the root and variant co-occurrence graphs.

The content of the variant tag clusters depict more vulnerable, toxic, and "triggering" content. The first cluster contains tags that bear a tone of self-loathing and self-harm ("suicide," "selfharmmm," "cuts"). We believe it comes from the community's constant dissatisfaction and discomfort with their objectified sense of physical appearance and attributes – self-harm takes such thoughts and emotions to an extreme. These tags also describe depression and reduced self-esteem more dramatically than the other cluster (e.g., "depression," "deadinside," "notgoodenough"). Literature indicates such pro-ED behavior to be consequences of disturbed interpersonal relations, difficulties with impulse control, and feelings of anxiety and failure [7].

These two distinctive clustering patterns show a tendency of the variant communities to adopt the lexical variations perhaps as a way to subvert Instagram attention on sharing of triggering, self-harm, and vulnerable content. Moreover, since we observed earlier that there is little overlap of the root tag and the variant tag communities, the user base who use these lexical variations may be the segment of the pro-ED community who intend to use the platform for sharing and promoting self-harm.

DISCUSSION

Summary of Findings. Our research has explored linguistic, behavioral, and topical changes in pro-ED communities in the aftermath of Instagram's moderation of pro-ED tags in 2012. Moderation in April 2012 led to the emergence of lexical variants of banned tags and tags with advisories. In fact, lexical variation showed a monotonic increase over time, indicating a desire on the part of the community to avoid outside attention and operate as an isolated, closed group (RQ 1). Next, while in general the sizes of these communities adopting lexical variant tags were smaller relative to the corresponding root tag, some lexical variation communities disproportionally increased in size (RQ 2). Communities adopting lexical variants were also found to show increased social participation and engagement compared to those around the moderated tags, revealing a tendency of the variant communities to continue to reinforce their pro-ED belief systems. Finally, these variants were extensively used to continue to share information encouraging adoption and maintenance of pro-ED lifestyles, often to also share more triggering, vulnerable, and self-harm content (RQ 3).

Is Moderation Effective? Overall, Instagram-enforced moderation was associated with negative consequences on Instagram's long-term strategy to remove pro-ED content. We observed increased use of lexical variation and expression of heightened toxic and vulnerable behavior over time. While social support and cohesion are generally linked to improved well-being, the pro-ED community situates such social cohesion to strengthen harmful attitudes towards body and health. Thus, content moderation has been mostly ineffective at decelerating the dissemination and proliferation of pro-ED behavior on the platform.

Moderation and Polarization. We note that banning a pro-ED tag does not remove or automatically delete posts that contain the tag; it only makes the tag unsearchable. Such moderation practices thus pose a genuine risk of these communities moving to the periphery of Instagram where any intervention techniques will be increasingly difficult to implement. In other words, if the community continues to move to increasing lexically variant tags and away from the more common pro-ED tags, as our results show, it would be difficult to discover them and thereby report, remove, or bring help to them. Additionally, these policies risk polarizing the pro-ED community and favoring pro-ED content. As our findings indicate, content on the variant tags is more triggering and vulnerable and relates to topics like self-harm. As users move away from the broader and more common pro-ED tags, they are less likely to be exposed to alternate views on eating disorders outside of their "echo chambers" or "filter bubbles" – especially views that can alter their perceptions of pro-ED as a lifestyle choice or raise awareness the dangerous effects of eating disorders on physical and emotional health.

Pro-ED as a Form of Deviance. This kind of adversarial adoption of non-conventional practices to subvert content moderation, as practiced by the pro-ED communities on Instagram, has been observed in other contexts as well. For example, citizens of authoritarian regimes avoid censorship by embracing different linguistic variation [3,17,25,28]. Further, several communities of deviant behavior have been known to avoid oversight of moderation by adopting a variety of agreed upon unorthodox norms, such as communities engaging in cyberbullying and online harassment [39], as well as those involved in socially unacceptable or damaging activities (human trafficking, drug abuse, violence, organized crime) [1].

Our research corroborates what has been observed scientifically and anecdotally in these other communities. We expand on this knowledge and show that platforms that use this as a strategy to disrupt dysfunctional communities may not be successful. Broadly, our work offers some of the first quantitative insights into the effectiveness of intervention strategies towards deviant behavior on social media, with pro-ED being a specific instance of deviance.

Design Implications

Previous research has shown the importance of sensitive communities as emotional "safety valves" of negative behavior, allowing disinhibiting discourse to avoid more drastic/dangerous actions [16]. However, use of safety valve communities can also have detrimental effects on the health of a community [12]. Our findings show that moderation of content may not be the most appropriate intervention. Communities simply adapt their social norms and conventions and share more vulnerable content. Rather than suppressing such content, social media platforms need to consider alternative intervention techniques that both provide this safety valve and promote *recovery* from pro-ED. We present some design considerations for alternative (both preventive and remedial) intervention techniques:

(1) Exposure to Recovery Content. Platforms could more critically examine the strategies that they use to moderate content. Prohibiting pro-ED content from being discoverable at all (banning precludes searches on a tag) was followed by increased activity and social participation. A more nuanced intervention strategy that does not ban content, for instance, could be issuing public service announcements, with pointers to eating disorder support communities or to an eating disorder hotline/resources (e.g., the National Eating Disorders Association NEDA website: *http://www.national eatingdisorders.org/*). This might create less incentive to migrate to different tags and more chances of continued use of the popular ones. This might also have beneficial effects on the community, since prior work indicates the recovery community to often attempt to "permeate" into the pro-ED community by using frequent and prominent tags in their content [11,42]. In essence, this

kind of intervention can promote pro-ED users' likelihood of being exposed to healthier behaviors than what is possible via banning.

(2) Recommending Healthy Behaviors. Many social networks like Instagram include recommendation systems that find similar content to a user's posting history or their social ties with other users. However, when these systems suggest content related to pro-ED or other damaging behaviors, pro-ED behavior gets reinforced to these vulnerable populations. We already observed that engagement and support on variant tags in the aftermath of enforcement of moderation increased over time. Therefore, recommending similar pro-ED content would only fortify such attitudes towards pro-ED. With appropriate modifications to the recommendation algorithms, Instagram and other platforms could limit the exposure of content associated with pro-ED tags. Instead, platforms could introduce recovery-related content in the suggested recommendations and to help disseminate information on the benefits and importance of ED recovery.

(3) Preemptively Detecting Emergent Pro-ED Tags. As we have noted, automatic discovery of the lexically variant tags can be challenging. As a way to tackle this, social networks could, for instance, detect emergent pro-ED tags, including the variant tags that are found to co-occur with known pro-ED tags. One method they could employ is identifying trending topics within the community over time. Such efforts would allow platforms to monitor the "health" of specific tags and communities. They could also monitor sentiment and attitudes across different communities and make appropriate adjustments to content advisories and notification strategies.

(4) Social and Clinical Help on Vulnerable Content. Our findings showed that variant tags were used by a segment of the pro-ED community with more vulnerable behavior. Social computing system designers could work with clinicians, therapists, and trusted/identified family members, and close friends to examine how to bring timely, appropriate, and privacy-preserving help to such groups alter their attitudes about the impact of pro-ED behaviors on health. In fact, recovery from pro-ED is a challenging experience and many individuals undergo conflicting perceptions of identity during recovery attempts, including revelation of vulnerability [11]. Intervention tools may specifically focus on the needs of such groups, for instance, providing psychosocial support in response to expression of vulnerable behavior in social media content.

Ethical Considerations

Social networks and platforms do not have any obligation to intervene in the case of the pro-ED community or other vulnerable populations. However, eating disorders are unique in that body perception and self-esteem are negatively impacted by social comparison enabled by social platforms as well as consumption of images of idealized physical appearance [5]. Unlike other health conditions, there is a collective opportunity for social media designers and researchers to rethink the affordances around discoverability and sharing of pro-ED content, not only for the dissipation of such behaviors but also to promote recovery and treatment of eating disorders.

We note that designing intervention strategies for users who participate in these communities is challenging on many practical and ethical fronts. Interventions must be delicately crafted. But at what point do interventions on social media become counterproductive or possibly manipulative? It is also important to balance these public health impacts alongside privacy concerns. To what extent can we notify trusted friends, family, and clinicians that someone may be suffering from an eating disorder? We would expect that our suggested strategies would be implemented with privacy-protecting standards in mind.

Detection and intervention will always be reactionary to new trends of deviant communities to avoid detection and hide in plain sight. Thus, any kind of intervention technique is a "game of cat and mouse" for many social networks. Pro-ED is only one example of a community strategically avoiding oversight; however, our research shows that, for this particular instance of deviant behavior, moderating content does not remove or reduce the proliferation of the community. Through our findings and this discussion, we hope to spur conversations in social media research and design communities towards crafting effective intervention systems for sensitive populations like pro-ED.

Limitations and Future Work

We acknowledge limitations of our research. This study used Instagram's official API for data collection, which is limited by Instagram's content moderation policies. The API does not return any data for banned tags. Our current dataset could only consider those posts where the banned tag also co-occurred with at least one other non-banned tag. While we are confident our findings hold given the overall size of the dataset, further investigations could incorporate alternative methods for broader data collection.

We also note that we investigated the adaptation of behavior in a specific community, pro-ED, following platform-enforced moderation of tags. However Instagram bans or provides advisory on tags spanning a variety of other topics too (e.g., pornography, suicide). Does the kind of adversarial attitude in our findings generalize to those settings as well? Future work will be able to explore to what extent our results were characteristic of the pro-ED community, or if it was a response by the broader Instagram user population towards content moderation

Our dataset is also limited by the public accessibility of Instagram content. We cannot access private posts, those

that have been formally removed by Instagram, or deleted or hidden by the users themselves. Importantly, from a statistical perspective, we suggest caution in deriving causality. While our findings do indicate increased lexical change in variant tags over time as well as heightened vulnerability manifested through their use in the aftermath of enforcement of moderation, there might be other latent factors relating to pro-ED behavior that might contribute to the observations in our data. Identifying such latent factors constitute a promising direction for future research. Finally, our research does not make any claim to attributing a diagnosis of an eating disorder to the posters of Instagram we study. It is not clear to what extent these posters actually met clinical criteria on eating disorders defined by DSM-5.

Future research could also incorporate a mixed methods approach toward developing deeper understanding of the intent and motives of Instagram and social media usage by the pro-ED communities. A complementary research question could examine how users came to know, understand, and agree on lexically variant tags. Detecting the lexical variant tags automatically through machine learning methods is another direction towards methodological innovation. Certainly, future work will require sensitivity and balancing privacy and other ethical concerns alongside goals of reducing the incidence of eating disorders. Collaborations between social media researchers and clinicians will be essential in developing future studies that examine the pro-ED space and other controversial communities.

CONCLUSION

In this paper, we offered the first quantitative analysis of pro-ED communities and their adoption of lexical variation on Instagram. Overall, we observed Instagram's content moderation policy to curb the sharing of pro-ED content to be ineffective. While some tags experience drop in usage after moderation, activity and engagement increased in others. More importantly, we showed that in lexical variants, topics of conversation move towards topics of self-harm, self-loathing, and other negative topics compared to root tags. Our findings thus raise interesting questions as to whether content moderation is the most effective means of intervention in the pro-ED community. Given the controversial nature of pro-ED content, social media design needs to consider broadly the impact of content moderation on deviant behavior and in social networks.

REFERENCES

1. Ronald L. Akers, Marvin D. Krohn, Lonn Lanza-Kaduce, and Marcia Radosevich. 1979. Social learning and deviant behavior: A specific test of a general theory. *American Sociological Review*, 636-655.
2. Jon Arcelus, Alex J Mitchell, Jackie Wales, and Sren Nielsen. 2011. Mortality rates in patients with anorexia nervosa and other eating disorders: a meta-analysis of 36 studies. *Archv. Gen. Psychiatry 68*, 7 (2011), 724–731.
3. David Bamman, Brendan O'Connor, and Noah Smith. 2012. Censorship and deletion practives in Chinese social media. *First Monday*, 17(3).
4. David Bamman, Jacob Eisenstein and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics,* 18(2), 135-160.
5. Anna M Bardone-Cone and Kamila M Cass. 2006. Investigating the impact of pro-anorexia websites: A pilot study. European Eating Disorders Review 14, 4 (2006), 256–262.
6. Yvonne M. Bishop, Stephen E. Feinberg, and Paul W. Holland. 2007. *Discrete multivariate analysis: theory and practice.* Springer Science & Business Media.
7. Dina LG Borzekowski, Summer Schenk, Jenny L Wilson, and Rebecka Peebles. 2010. e-Ana and e-Mia: A content analysis of pro–eating disorder web sites. *American journal of public health* 100, 8 (2010), 1526.
8. Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooo lllllllllllll!!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proc. EMNLP*. 562-570.
9. Jane Callaghan. 2013. Research in online spaces: 'Tumblr' and eating 'disorder'. Symposium presented to: Children and Young People's Mental Health. The University of Northampton.
10. Justine Costanza. Instagram 'Thinspo' Ban Won't Combat Pro-Eating Disorder Web Content. 2013. Retrieved November 15, 2014 from http://www.ibtimes.com/instagram-thinspo-ban-wont-combat-pro-eating-disorder-web-content-994366
11. Munmun De Choudhury. 2015. Anorexia on Tumblr: A Characterization Study. In *Proc. Digital Health*.
12. Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on Reddit: Self-Disclosure, Social Support, and Anonymity. In *Proc. ICWSM*.
13. Lauren Duca. Can Thinspiration Really Be #Banned From Instagram?. 2013. Retrieved November 16, 2014 from *http://www.huffingtonpost. com/lauren-duca/thinspiration-banned-from-instagram_b_3829155.html*
14. Maeve Duggan, Nicole B. Ellison, Cliff Lampe, Amanda Lenhart, and Mary Madden. 2015. Social Media Update 2014. *Pew Research Internet Proj.*
15. Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proc. EMNLP*.
16. Tobit Emmens and Andy Phippen. 2010. Evaluating Online Safety Programs. *Harvard Berkman Center for Internet & Society.*

17. Antonio M. Espinoza and Jedidiah R. Crandall. 2011. Work-in-progress: Automated named entity extraction for tracking censorship of current events. In *USENIX Workshop on Free & Open Comm. on the Internet*.
18. Facebook. Statement of Rights and Responsibilities. 2013. Retrieved January 10, 2015 from *https://www.facebook.com/legal/terms*
19. Rachel A Fleming-May and Laura E Miller. 2010. I'm scared to look. But I'm dying to know: Information seeking and sharing on Pro-Ana weblogs. In *Proc. ASIST* 47, 1 (2010), 1–9.
20. Stephan Gouws, Donald Metzler, Congxing Cai, and Eduard Hovy. 2011. Contextual bearing on linguistic variation in social media. In *Proc. Workshop on LSM*. 20-29.
21. Heba Hasan. Instagram bans Thinspo Content. 2012. Retrieved November 15, 2014 from *http://newsfeed.time.com/2012/04/26/instagram-bans-thinspo-content/*
22. Chaya Hiruncharoenvate, Zhiyuan Lin, and Eric Gilbert. "Algorithmically Bypassing Censorship on Sina Weibo with Nondeterministic Homophone Substitutions." 2015. In *Proc. ICWSM*.
23. Yuheng Hu, Kartik Talamadupula, & Subbarao Kambhampati. 2013. Dude, srsly?: The Surprisingly Formal Nature of Twitter's Language. In *Proc. ICWSM*.
24. Instagram. Instagram's New Guidelines Against Self-Harm Images & Accounts. 2012. Retrieved November 17, 2014 from *http://blog.instagram.com/post/21454597658/ instagrams-new-guidelines-against-self-harm*
25. Gary King, Jennifer Pan, & Margaret E. Roberts. 2013. How censorship in China allows government criticism but silences collective expression. *American Political Science Rev*, *107*(02), 326-343.
26. Farshad Kooti, Haeryun Yang, Meeyoung Cha, P. Krishna Gummadi, and Winter A. Mason. 2012. The Emergence of Conventions in Online Social Networks. In *Proc. ICWSM*.
27. Stephen P. Lewis and Alexis E. Arbuthnott. 2012. Searching for thinspiration: The nature of internet searches for pro- eating disorder websites. *Cyberpsychology, Behavior, and Social Networking*, 15(4), 200-204.
28. Rebecca MacKinnon. 2008. Flatter world and thicker walls? Blogs, censorship and civic discourse in China. *Public Choice* 134, 1-2 (2008), 31–46.
29. Jeanne B Martin. 2010. The development of ideal body image perceptions in the United States. *Nutrition Today* 45, 3 (2010), 98–110.
30. Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM comp. surveys*, *33*(1), 31-88.
31. Minsu Park, David W McDonald, and Meeyoung Cha. 2013. Perception Differences between the Depressed and Non-depressed Users in Twitter. *In Proc. ICWSM*.
32. Alice Philipson. Concerns raised over Instagram after app allows users to see photos promoting anorexia. 2013. Retrieved November 16, 2014 from *http://www.telegraph.co.uk/technology/social-media/9775559/Concerns-raised-over-Instagram-after-app-allows-users-to-see-photos-promoting-anorexia.html*
33. Hansen Andrew Schwartz, Johannes C. Eichstaedt., Margaret L. Kern, Lukasz Dziurzynski, Richard E. Lucas, Megha Agrawal, Gregory J. Park, et al. 2013. Characterizing Geographic Variation in Well-Being Using Tweets. In *Proc. ICWSM*.
34. Leslie Regan Shade. 2003. Weborexics: The Ethical Issues Surrounding Pro-Ana Websites. *SIGCAS Comput. Soc*. 33, 4 (Dec. 2003).
35. Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *22*(8), 888-905.
35. Clay Shirky. 2011. The political power of social media: Technology, the public sphere, and political change. *Foreign affairs*, 28-41.
36. Irina Shklovski and Nalini Kotamraju. 2011. Online contribution practices in countries that engage in Internet blocking and censorship. In *Proc. CHI*. 1109-1118.
37. Frédérique RE Smink, Daphne van Hoeken, and Hans W Hoek. 2012. Epidemiology of eating disorders: incidence, prevalence and mortality rates. *Current psychiatry reports* 14, 4 (2012), 406–414.
38. John R Suler and Wende L Phillips. 1998. The bad boys of cyberspace: Deviant behavior in a multimedia chat community. *CyberPsychology & Behavior* 1, 3 (1998), 275–294.
39. Zeynep Tufekci, & Christopher Wilson. 2012. Social media and the decision to participate in political protest: Observations from Tahrir Square. *Journal of Comm*, *62*(2), 363-379.
40. Tracey D. Wade, Anna Keski-Rahkonen., & James I. Hudson. 2011. Epidemiology of Eating Disorders. In *Textbook in Psychiatric Epidemiology*. Wiley, 343–360.
41. Elad Yom-Tov, Luis Fernandez-Luque, Ingmar Weber, and P Steven Crain. 2012. Pro-Anorexia and Pro-Recovery Photo Sharing: A Tale of Two Warring Tribes. J Med Internet Res (2012).
42. Yafu Zhao and William Encinosa. 2009. *Hospitalizations for Eating Disorders from 1999-2006.* Agency for Healthcare Policy and Research, Rockville, MD.