Estimating Determinants of Attrition in Eating Disorder Communities on Twitter: An Instrumental Variables Approach

Tao Wang¹,², MSc*; Emmanouil Mentzakis¹, PhD; Markus Brede³, PhD; Antonella Ianni¹, PhD;

¹Department of Economics, University of Southampton, UK
²The Alan Turing Institute, UK
³Department of Electronics and Computer Science, University of Southampton, UK

Corresponding Author:
Tao Wang, MSc
Department of Economics
University of Southampton
Southampton, SO17 1BJ
United Kingdom
Phone: +44 7864373080
Email: t.wang@soton.ac.uk

Abstract

Background: The use of social media as key health-information source has increased steadily among people affected by eating disorders. Research has examined characteristics of individuals engaging in online communities, while little is known about discontinuation of engagement and the phenomenon of participants dropping out of these communities.

Objective: This study aims to investigate characteristics of dropout behaviors among eating disordered individuals on Twitter and to estimate the causal effects of personal emotions and social networks on dropout behaviors.

Methods: Using a snowball sampling method, we collected a set of individuals who self-identified with eating disorders in their Twitter profile descriptions, as well as their tweets and social networks, leading to 241,243,043 tweets from 208,063 users. Individuals’ emotions are measured from their language use in tweets using an automatic sentiment analysis tool, and network centralities are measured from users’ following networks. Dropout statuses of users are observed in a follow-up period 1.5 years later (from Feb. 11, 2016 to Aug. 17, 2017). Linear and survival regression instrumental variables models are used to estimate the effects of emotions and network centrality on dropout behaviors. The average levels of attributes among an individual’s followees (i.e., people who are followed by the individual) are used as instruments for the individual’s attributes.

Results: Eating disordered users have relatively short periods of activity on Twitter, with one half of our sample dropping out at 6 months after account creation. Active users show more negative emotions and higher network centralities than dropped-out users. Active users tend to connect to other active users, while dropped-out users tend to cluster together. Estimation results suggest that users’ emotions and network centralities have causal effects on their dropout behaviors on Twitter. More specifically, users with positive emotions are more likely to drop out and have shorter-lasting periods of activity online than users with negative emotions, while central users in a social network have longer-lasting participation than peripheral users. Findings on users’ tweeting interests further show that users who attempt to recover from eating disorders are more likely to drop out than those who promote eating disorders as a lifestyle choice.

Conclusions: Presence in online communities is strongly determined by individual’s emotions and social networks, suggesting that studies analyzing and trying to draw condition and population characteristics through online health communities are likely to be biased. Future research needs to examine in more detail the links between individual characteristics and participation patterns if better understanding of the entire population is to be achieved. At the same time, such attrition dynamics need to be acknowledged and controlled for when designing online interventions so as to accurately capture their intended populations.
KEYWORDS
Medical Informatics; Eating Disorders; Social Media; Attrition; Emotions; Social Network; Causality; Instrumental Variables

Introduction

Eating disorders (ED), such as anorexia and bulimia, are complex mental disorders defined by extreme obsessions with body weight or shape, and unusual eating behaviors [1]. These diseases have the highest mortality rate of any mental illness [2], intractable co-morbidities [3] and worldwide prevalence [4], having become a major public health concern. Although a variety of treatment options have emerged over recent years [5], populations affected by ED are often hard to reach through traditional healthcare services. This is mainly due to fear of stigma or a feeling of shame; many sufferers conceal their ED symptoms and never seek professional treatment or support [6, 7]. To keep struggles with illnesses private, people often seek health-related information and support through online peer-to-peer communities, particularly via social media sites like Twitter and Facebook. Participation in online communities is common in ED populations [8] and has been suggested as a screening factor for ED [3]. This provides an opportunity for healthcare professionals to deliver health support to these hard-to-reach populations through online communities [9–13]. Moreover, as online communities present a relatively anonymous environment for individuals to naturally self-disclose and socialize [14], user-generated data online provides a large amount of records about individuals’ concerns, thoughts, emotions and social interactions [15–17], which may complement traditional data sources (e.g., surveys and interviews) in understanding risk factors of ED. Hence, growing research has focused on characterizing individuals’ behavioral patterns in online communities [15–19], so as to better understand ED and promote population-level well-being.

One notable characteristic of online ED communities is their participants having widely different stances on ED [8, 20, 21]. Some communities encourage members to discuss their struggles with ED, share treatment options and offer support towards recovery from ED, so called pro-recovery communities [20–22]. There are also many anti-recovery or pro-ED communities in which members often deny ED being a disorder and instead promote ED as a healthy lifestyle choice [8, 23]. These pro-ED communities can negatively affect health and quality of life among people with and without ED, through reinforcing an individual’s identity around ED [24], promoting thin ideals [25], and disseminating harmful practices for weight loss [8]. Recent studies have shown that individuals’ language use online strongly indicate their pro-ED or pro-recovery stances [15, 17, 20], as well as emotions of depression, helplessness and anxiety that reflect their mental disorders [16]. Other studies have also examined interactions between pro-ED and pro-recovery communities on Flickr [21], anorexia-related misinformation [18], sentiments of comments on ED-related videos on YouTube [26], characteristics of removed pro-ED content [27] and lexical variation of pro-ED tags on Instagram [19, 28]. Yet, prior studies have largely focused on examining how people engage in and maintain an online ED community, while little is known about how people drop out of such a community. As a dynamic process, people who join and actively engage in a community at earlier stages can have less participation and leave the community at later stages. Understanding the attrition processes of online communities can enhance our knowledge of the dynamics in these communities.

Studying the attrition process of an online community can also have practical implications for disease prevention and health interventions. Given the ease of accessibility of social media for many individuals (e.g., via mobile devices), increasing attention has focused on using online communities to deliver health interventions [9–13, 29, 30]. One of the most popular approaches is to deliver health lessons and behavior-changing instructions via online communities [9–13, 29]. Although pilot studies based on small samples have demonstrated the effectiveness of these approaches in reducing body dissatisfaction and disordered eating [12, 13], evidence from interventions for a variety of health behaviors (e.g., smoking, diet, exercise and sexual health) suggests that attrition (i.e., participant loss) is one of the most common challenges in online interventions [10, 29]. This is known as the “law of attrition” of online interventions [31]. A recent study has shown a high attrition rate in an online intervention for ED [32], though this intervention is delivered via a purposely designed website rather than a general social media site. Thus, an important goal in conducting successful interventions via online communities is to improve members’ retention, as members who remain longer are more likely to receive these interventions and have more opportunities to promote a target behavior change. To achieve this goal, a critical first step is to understand what factors influence members’ retention in an online community.

Prior studies have shown that people’s decisions of retention or dropout in online communities are associated with a variety of factors [33, 34], including personality traits (e.g., shyness and the Big-Five traits) [35, 36], interests [37], recognition in a community [38–41] and support from others [42, 43]. However, such an association is not adequate to conclude the presence of a causal relationship [44, 45] between an individual’s attributes and her/his online participation. This is because an association...
can arise from non-causal relationships. For example, most prior studies focus on the use of self-reported surveys and rely on participants’ reports of their own personality, concerns and behaviors [35, 39, 46]. This can introduce considerable retrospective bias and measurement errors, leading to a coincidental association between two unrelated variables, particularly in small samples. Even if variables are measured rather than self-reported [43, 47], participation in an online community is inherently self-selected (e.g., sharing common interests) and members can drop out for many different reasons (e.g., effect of an online or offline event). Thus, unobservable factors (i.e., confounding variables) may affect both a main predictor and participation outcomes, causing a spurious association. Moreover, in some cases reverse causality can lead to an association, e.g., prior studies suggest that feelings of social isolation are linked to frequent social media use [35, 48] whereas recent studies indicate that social media use is linked to increased feelings of social isolation [49]. Technically speaking, the issues of measurement errors, confounding variables and reverse causality can cause endogeneity which refers to an explanatory variable of interest being correlated with the error term in a regression model [44]. In these cases, traditional methods such as ordinary least squares (OLS) give biased and inconsistent estimates of the effect of interest. It is therefore not surprising that mixed results exist in prior studies, e.g., a positive association between individuals’ expertise and online participation was found in [46] while a negative association was found in [38].

This study aims to estimate determinants of dropout in an online ED community, while addressing the endogeneity issues by using an instrumental variables (IV) approach [44]. Specifically, we analyze tweeting activities for a large set of individuals who self-identified with ED on Twitter over 1.5 years and identify the presence of dropout if a user ceased to post tweets in the observation period. We explore determinants of a user’s dropout based on incentive theory [34, 50] which argues that people’s engagement in an activity can be driven by (i) intrinsic motivation which refers to doing something because it is interesting or enjoyable, and (ii) extrinsic motivation which refers to doing something because it earns an external reward. We focus on intrinsic motivation captured by personal emotions and extrinsic motivation captured by sociometric status in an online peer-to-peer community. Rather than using self-reports [35, 39, 46], we measure users’ emotions based on their emotional expressions in tweets using sentiment analysis techniques [51] and quantify users’ sociometric statuses by network centrality [52] in the social network of an ED community on Twitter. Based on these measured variables, IV estimators both for the decision to drop out and for the time to dropout are implemented to achieve consistent estimates of the effects of personal emotions and network centrality on dropout in an online ED community. To better understand the estimation results, we further examine heterogeneity in tweeting interests (i.e., topics discussed in tweets) of users with differing levels of characteristics (e.g., emotions) and dropout outcomes, respectively. To our knowledge, this study is the first to systematically characterize the determinants of dropout behaviors in online ED communities. Three research questions are examined: (i) what are the general characteristics of the attrition process in an online ED community, (ii) how do intrinsic and extrinsic factors affect the decision of an individual to drop out of the community, (iii) how do these factors affect the duration of time until the occurrence of dropout?

**Methods**

**Data Collection**

Our data is collected from Twitter, a microblogging platform that allows millions of users to self-disclose and socialize. As many social media platforms like Facebook and Instagram have taken moderation actions to counteract pro-ED content and user accounts [28], Twitter has not yet enforced actions to limit such content [53]. This makes Twitter a unique platform to study the attrition process naturally happening in an online ED community and allows us to examine individuals’ behaviors in a non-reactive way. Our study protocol was approved by the Ethics Committee at the University of Southampton. All data used in our study is public information on Twitter and available through the Twitter APIs (application programming interfaces). No personally identifiable information is used in this study. Our data collection process includes three phases.

- First, we collect a set of individuals who self-identified with ED on Twitter using a snowball sampling approach. Specifically, we track the public tweet stream using “eating disorder”, “anorexia”, “bulimia” and “EDNOS” from Jan. 8 to 15, 2016. This results in 1,169 tweets that mention ED. From the authors of these tweets, we identify 33 users who self-reported both ED-related keywords (e.g., “eating disorder”, “anorexia” and “bulimia”) and personal bio-information (e.g., body weight and height) in her profile descriptions (i.e., a sequence of user-generated text describing their accounts below profile images). Starting from these seed users, we expand the user set using snowball sampling through their social networks of followees/followers. At each sampling stage, we filter out non-English speaking accounts and finally obtain.
3,380 unique ED users who self-report ED-related keywords and bio-information in their profile descriptions. Note that our focus in this work is studying individuals who are affected by ED rather than those who are related to ED. The inclusion of bio-information in user sampling allows us to filter out ED-related therapists, institutes or organizations, as these users often display ED-related keywords but do not show bio-information in their Twitter profile descriptions. Details about the data collection of ED users can be found in our prior work [54].

- Then, we collect all friends (including followees and followers) of each ED user, leading to a large social network consisting of 208,063 users. For each user, we retrieve up to 3,200 (the limit returned from Twitter APIs) of their most recent tweets and obtain 241,243,043 tweets in total. The data collection process finished on Feb. 11, 2016.

- Finally, we open a follow-up observation period for all users on Aug. 17, 2017 to obtain measurements on users’ activities online. In the second observation, we only collect users’ profile information which includes users’ last posted statuses.

To verify the quality of our collected sample, two members of the research team classified a random sample of 1,000 users on whether they were likely to be a true ED user based on their posted tweets, images and friends’ profiles. Users are classified as “disordered” if they frequently and intensively post their body weights, details of their dietary regimen (e.g., calories), struggles with eating (e.g., “I want to eat but cannot”), pictures of themselves, self-reports of being disordered or in recovery in tweets and follow ED-related friends (e.g., user profiles with ED-related keywords). The process revealed a 95.2\% match between the identified ED individuals in the data collection stage and those classified as ED during inspection. Although it is impossible to diagnose individuals’ disorders based on their online behaviors, this inspection provides a strong indication that the collected users are likely to be affected by ED rather than those who merely talk about ED online. See [54] for details of data validation.

**Estimation Framework**

Two different models are specified to estimate the effects of emotions and network centrality on dropout. First, we specify a linear probability model on the whole sample to estimate the effects of individuals’ characteristics observed in the first-observation period on the probability of dropping-out in the second-observation period. Second, we estimate survival models to explore the effects of individuals’ characteristics observed in the first observation on the time to dropout in the second observation (i.e., the duration from our first observation to the dropout in our second observation). However, like all social media studies, only a limited number of individuals’ characteristics are available for our estimations and these are mostly observed through user-generated data online. This leads to confounding variable bias, since unobservable factors can be correlated with both the main explanatory variables (i.e., emotions and network centrality) and dropout outcomes. For example, undergoing hospital treatment can simultaneously affect a person’s emotional state and the use of social media. Further, prior studies have shown that social media use is associated with increased depression [55], social anxiety [49] and body dissatisfaction [56, 57], implying an effect of online participation on individuals’ emotions (i.e., reverse causality). Both confounding variables and reverse causality result in biased estimates of the effects of emotions and network centrality on dropout. This problem can be addressed by using a randomized controlled trial in which emotions or network centralities are randomly assigned to users by researchers [58]. Such a trial, however, is not feasible, due to ethical and practical limitations [59].

Here, we utilize an alternative approach for estimating the effects of interest that is based on instrumental variables (IV) regression, an econometric technique to infer causal relations from observational data [44]. This technique has been applied to a variety of contexts, from identifying the causal effect of education on earning [60], the effect of a health treatment [61], to estimating social contagion effects on both online [59] and offline behaviors [62]. Formally, consider a model \( Y = \beta_1 X_1 + \beta_2 X_2 + u \), where \( X_1 \) is endogenous, \( X_2 \) is exogenous, \( u \) is a random error term and \( \beta \)'s are effects to be estimated. IV methodology uses an instrument \( Z \) (which is (i) not contained in the explanatory equation, (ii) correlated with \( X_1 \), i.e., \( \text{cov}(Z, X_1) \neq 0 \), and (iii) uncorrelated with \( u \), i.e., \( \text{cov}(Z, u) = 0 \), conditional on the other covariates such as \( X_2 \) ) and runs a first stage reduced-form regression \( X_1 = \gamma_1 Z + \gamma_2 X_2 + v \), where \( v \) is a random error. The causal effect of \( X_1 \) on \( Y \) is then given in a second stage regression \( Y = \beta_1 \hat{X}_1 + \beta_2 X_2 + u \), where \( \hat{X}_1 \) is the predicted values of \( X_1 \) from the first stage. See [44] for more details.

**Measures**

A number of variables are needed for estimations. All independent variables and IV are measured in the first-observation period (unless otherwise stated), while dependent variables are measured in the second-observation period.
Dropout Outcomes as Dependent Variables

Following previous studies [42, 43], we identify the presence of dropout if a user ceases to post tweets. Specifically, in the linear probability models, we encode the dropout status of a user as 0 (denoting non-dropout) if the user has updated posts in our second observation, and 1 (denoting dropout) otherwise.

In the survival models, each user has a two-variable outcome: (i) a censoring variable denoting whether the event of dropout occurs, and (ii) a variable of survival duration denoting the duration of time until the occurrence of dropout. We censor the occurrence of a “dropout event” in two ways. First, users are said to drop out if they have not posted tweets for more than a fixed threshold interval \( \pi \) before our second observation (so called identical-interval censoring). As people use social media platforms with different activity levels, e.g., some users post every several hours while other users only post once every couple of days, our second censoring method further accounts for personalized posting activities of individuals (called personalized-interval censoring). In this method, users are said to drop out if they have not posted tweets for more than a variate threshold interval \( \lambda \pi + (1 - \lambda) \bar{I}_i \) before our second observation, where \( \pi \) is a fixed threshold, \( \bar{I}_i \) is the average posting interval of individual \( i \) in our first observation period, and \( \lambda \) is a tunable parameter to control the effects of individual activities. We tune the parameters by maximizing the agreement between the estimated dropout states based on users’ activities in our first observation and the observed states in our second observation. See Supplementary Information (SI) for details. For users who are censored as dropped-out, we set their survival times as the durations from our first observation to their last postings in our second observation. For those who are censored as non-dropped-out, we set their survival times as the whole time period between our two observations.

Emotions and Network Centrality as Main Explanatory Variables

Individuals’ emotions are measured through their language used in tweets. There is a variety of sentiment analysis algorithms to measure emotional expressions in texts [51, 63]. In this study, we use SentiStrength [51] as (i) it has been used to measure the emotional content in online ED communities and shown good inter-rater reliability [26]; (ii) it is designed for short informal texts with abbreviations and slang, and thus suitable to process tweets [51]. After removing mention marks, hashtags and URLs, each tweet is assigned a scaled value in \([-4, 4]\) by SentiStrength, where negative/positive scores indicate the strength of negative/positive emotions respectively, and 0 denotes neutral emotions. We quantify a user’s emotional state by the average score of all tweets posted by the user. All re-tweets are excluded, as re-tweets reflect more the emotions of their original authors than those of their re-tweeters. For robust results from the language processing algorithms, we only consider users who have more than 10 tweets and post more than 50 words.

Network centrality measures the importance of a person in a social network; people well-recognized by their peers often have high centralities in a group [52]. To measure a user’s centrality in the ED community, we build a who-follows-whom network among ED users and their friends, where a directed edge runs from node A representing user A to node B representing user B if A follows B on Twitter. While there are various measures of network centrality, we focus on coreness centrality [64] as it has been shown to outperform other measures such as degree and betweenness centrality [52] in detecting influential nodes in complex networks [65] and cascades of users leaving an online community [66, 67]. We measure the sociometric status of a user in the ED community by the in-coreness centrality [68] of a node in the generated network using the package igraph 0.7.0 [69].

Aggregated Emotions and Network Centrality of Friends as Instrumental Variables

As IV for a user’s attributes, we use average emotions and network centrality over all followees of the user, i.e., people who are followed by the user. The choice of these IV is based on the following considerations. First, we consider the relevance assumption of our instruments requiring that the characteristics of followees are correlated to the user’s characteristics, i.e., \( \text{cov}(Z, X_1) \neq 0 \). We expect that followees’ updates act as information sources for a user, and followees’ behaviors as well as emotions manifested in their tweets can influence the user. Prior work [54] has shown the presence of homophily among ED users on Twitter suggesting that users who share similar emotional and network attributes tend to follow one another. Further, the empirical existence and strength of the relevance property are tested in a first stage regression and presented along with the structural estimates of the models.

Second, we examine the exogeneity requirement (i.e., \( \text{cov}(Z, u) = 0 \)), where followees’ emotions and centrality must not be have a direct effect on the drop-out decision of the user other than through their effect on the user’s emotions. While we take such assumption to be reasonable, we identify a pathway through which direct links could arise. Followees’ attributes (e.g.,
emotions) could affect a user’s dropout through their effects on followees’ own dropouts, e.g., followees’ emotional states may affect their own dropouts, and a feeling of loneliness due to friends’ leaving may then drive the target user to drop out. To control for this channel, we measure the proportion and durations of followees that remain active in our second observation (regardless of whether the target user drops out or not). Further, we change the definition of followees (that are used to create the instruments) to those who are followed by a user but do not follow the user back (called single-way followees). In this setting, the reverse causality of a user’s dropout on followees’ attributes is nullified, which strengthens the exogeneity assumption on IV and controls.

Table 1. Covariates used in estimations.

<table>
<thead>
<tr>
<th>Control effect</th>
<th>Covariate</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social capital</td>
<td>#Followees</td>
<td>Number of total followees</td>
</tr>
<tr>
<td></td>
<td>#Posts</td>
<td>Number of total posts, including tweets and re-tweets</td>
</tr>
<tr>
<td></td>
<td>#Followers</td>
<td>Number of total followers</td>
</tr>
<tr>
<td>Activity level</td>
<td>Active days</td>
<td>Number of days from account creation to last posting</td>
</tr>
<tr>
<td></td>
<td>#Followee/day</td>
<td>Average number of followees per day</td>
</tr>
<tr>
<td></td>
<td>#Posts/day</td>
<td>Average number of posts per day</td>
</tr>
<tr>
<td></td>
<td>#Followers/day</td>
<td>Average number of followers per day</td>
</tr>
<tr>
<td>Observational bias</td>
<td>#Tweets in use</td>
<td>Number of tweets in use to measure emotions</td>
</tr>
<tr>
<td></td>
<td>#Followees in use</td>
<td>Number of followees whose attributes are used as instruments</td>
</tr>
<tr>
<td>Alternative causal channel</td>
<td>%Active followees</td>
<td>Proportion of followees being active between two observations</td>
</tr>
<tr>
<td></td>
<td>⟨Followee durations⟩</td>
<td>Average days of followees being active between two observations</td>
</tr>
</tbody>
</table>

Estimation Covariates

Our estimates control for several covariates that may affect users’ tweeting activities, as listed in Table 1. We first measure users’ social capital on Twitter (e.g., the numbers of social connections and the levels of engagement in sharing content) to capture the fact that people with different levels of popularity may have different tendencies to share content online [70]. Note that, although the numbers of followees and followers can be regarded as the in- and out-degree centralities of a user in the whole social network on Twitter (i.e., the “global” social capital), we are interesting in the “local” network centrality in the ED-specific communities. Second, as prior studies show an association between social media use and depression [55], we measure historical activity levels of users (i.e., active days) to capture effects that previous engagement may relate to both users’ emotions and their future engagement. We also measure users’ activity frequencies (e.g., posting frequency) to capture their patterns of Twitter usage. Third, the covariates on observational bias are used to control for effects caused by incomplete observations, e.g., a limited number of tweets are retrieved and used to measure emotions for a user. All variables on social capital, activity level and observational bias are measured from users’ profile information and tweets collected in our first observation. Moreover, as discussed above, we include the proportion and average durations of followees that are active in our second observation to capture the channel that followees’ emotions affect a user’s dropout through their effects on followees’ own dropouts.

Model Estimations

IV Estimation in Linear Regression Model

We use standard two-stage least squares (2SLS) estimators for linear probability models. In the first stage, we run an auxiliary regression and predict the endogenous variables (i.e., an individual’s emotional state and network centrality) based on IV and exogenous covariates. In the second stage regression, we substitute the endogenous variables of interest with their predicted values from the first stage. Estimation is conducted through the AER package [71] and robust standard errors are computed.
IV Estimation in Survival Model

We use a Kaplan-Meier estimator [72] to estimate the survival function from data. Aalen’s additive hazards model [73] is used to estimate the effects of users’ attributes on the time to dropout. Compared to the proportional hazards models in which the ratios of hazard functions (i.e., hazard ratios) for different strata are assumed to be constant over time [74], the additive model is more flexible and applies under less restrictive assumptions. To compute an IV estimator in an additive hazards model, we use a control-function based approach which is proposed by Tchetgen et al. [61]. The TIMEREG package [75] is used for the implementation of the estimation algorithm. Standard errors are obtained through non-parametric bootstrap.

Results

Descriptive Statistics

We obtain 2,906 users who posted more than 10 tweets (excluding re-tweets) and 50 words in our data, where 2,459 (85%) users had no posting activities during our two observation periods. Among 357 users who self-reported gender information in their Twitter profile descriptions, 84% of them \( (n = 300) \) are female. The mean age is 17.3 among ED users who self-reported age \( (n = 1,015) \). Based on the timestamps of account creation and last posting, we use the Kaplan-Meier estimator to estimate the “lifetime” of a user on Twitter, i.e., the duration from account creation to the last posting. The estimated median lifetime of these users on Twitter is 6 months, i.e., one half of the entire cohort drops out at 6 months after creating an account. Figure 1 visualizes the social network between dropouts and non-dropouts among ED users. We note that users with the same dropout states tend to cluster together. Computing Newman’s homophily coefficient \( r [76] \) of this network by users’ dropout states, we find \( r = 0.09 \) \( (z = 16.84 \text{ and } P < .001 \text{ compared to a null model, see } SI) \), suggesting that users with the same dropout states tend to befriend one another. See SI for details of data statistics.

Estimation Results of Linear Probability Models

Table 2 shows estimated results in the linear models with two different IV specifications. In the first specification, we use all followees of a user to create IV for the user’s attributes. The results are given in columns 2-3, in which both OLS and IV
estimators show that positive emotions are associated with a higher probability of dropout \((\beta = 0.044, P = .007\) and \(\beta = 0.29, P < .001\), respectively), with largely comparable coefficients for covariates. Compared to the OLS estimator, the IV estimator of the effect of emotions on dropout is remarkably stronger. The Wu-Hausman test further shows a significant difference between the OLS and IV estimators \((P < .01)\), suggesting the presence of endogeneity. These results indicate that ignoring endogeneity in the OLS estimation leads to an underestimation of the effect of interest. Moreover, the \(F\)-statistics in the first stage regressions show that the relevance of IV exceeds the conventional standard of \(F = 10\) [78], indicating the validity of our IV.

Table 2. Estimated effects of emotions on dropout using OLS and IV models.

<table>
<thead>
<tr>
<th></th>
<th>All followees</th>
<th></th>
<th>Single-way followees</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>IV</td>
<td>OLS</td>
<td>IV</td>
</tr>
<tr>
<td>(\beta)</td>
<td>.044</td>
<td>.005</td>
<td>.290</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>#Followees</td>
<td>-.0004</td>
<td>.01</td>
<td>-.0002</td>
<td>.18</td>
</tr>
<tr>
<td>#Posts</td>
<td>-.000000</td>
<td>.18</td>
<td>-.000000</td>
<td>.35</td>
</tr>
<tr>
<td>#Followers</td>
<td>0.00001</td>
<td>.53</td>
<td>0.00001</td>
<td>.58</td>
</tr>
<tr>
<td>Active days</td>
<td>-.0003</td>
<td>&lt; .001</td>
<td>-.0003</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>#Followee/day</td>
<td>0.001</td>
<td>.01</td>
<td>0.002</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>#Posts/day</td>
<td>0.0002</td>
<td>.72</td>
<td>-.0001</td>
<td>.38</td>
</tr>
<tr>
<td>#Followers/day</td>
<td>-.003</td>
<td>.01</td>
<td>-.005</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>#Tweets in use</td>
<td>-.000004</td>
<td>.002</td>
<td>-.000004</td>
<td>.003</td>
</tr>
<tr>
<td>#Followees in use</td>
<td>0.0004</td>
<td>.03</td>
<td>0.0002</td>
<td>.38</td>
</tr>
<tr>
<td>%Active followees</td>
<td>-1.159</td>
<td>&lt; .001</td>
<td>-.812</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>⟨Follower durations⟩</td>
<td>.001</td>
<td>.004</td>
<td>.001</td>
<td>.16</td>
</tr>
<tr>
<td>Constant</td>
<td>1.270</td>
<td>&lt; .001</td>
<td>1.273</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>

Observations 2,906 2,906 2,898 2,898
First-stage \(F\)-statistic\(^a\) 440.26 \((P < .001)\) 158.21 \((P < .001)\)
Wu-Hausman test\(^b\) 42.24 \((P < .001)\) 14.54 \((P < .001)\)

\(^a\)\(F\)-statistic tests the significance of the instrument from a first-stage regression of a user’s emotions on followees’ emotions (i.e. the instrument) and the rest of the covariates.

\(^b\)Test the difference in estimates between OLS and IV; rejecting the null hypothesis suggests the presence of endogeneity.

\(^c\)P-values are computed based on heteroscedasticity-consistent standard errors.

Columns 4-5 show results of the second IV specification in which only single-way followees are used to create IV. Users who have no any single-way followees are excluded as instruments for these users’ attributes are not available. Thus, the number of observations decreases as compared to that in the first IV specification. Moreover, as data on a smaller number of friends is used in the second IV specification, the relevance of IV becomes weaker but still passes the conventional test in the first stage regression. Despite such changes, the two specifications produce largely similar results. Computing Wald tests of equality of coefficients between the two IV models, we find that the estimated effects of emotions on dropout are statistically the same across different IV specifications \((P = .8)\), potentially suggesting robustness of the results.

Note that network centrality is excluded from the linear models. This is because many users had dropped out long before our first observation (see Figure S2, SI), and the social networks of such users might largely change from the dates of their dropouts to our first observation, e.g., a user might be followed by new followers when these followers were unaware of the dropout of this user. In these cases, network centralities in the future are used to explain dropouts in the past, which can produce misleading results in the linear models. Nevertheless, including network centrality and instrumenting for it return statistically insignificant effect of centrality on the dropout decision, confirming our argument above on the irrelevance of centrality on this binary decision to drop out or not.
Table 3. Estimated effects of emotions and centrality on survival time using Aalen’s additive hazards models\textsuperscript{a}.

<table>
<thead>
<tr>
<th></th>
<th>All followees (95% CI)</th>
<th>Single-way followees (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Identical-interval censoring</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emotions</td>
<td>-0.018 (-0.037, 0.0002)</td>
<td>-0.043 (-0.083, -0.004)</td>
</tr>
<tr>
<td>Centrality</td>
<td>0.001 (0.0008, 0.0011)</td>
<td>0.001 (0.0007, 0.0011)</td>
</tr>
<tr>
<td><strong>Personalized-interval censoring</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emotions</td>
<td>-0.016 (-0.034, 0.0031)</td>
<td>-0.038 (-0.08, 0.0022)</td>
</tr>
<tr>
<td>Centrality</td>
<td>0.001 (0.0008, 0.0011)</td>
<td>0.001 (0.0008, 0.0012)</td>
</tr>
<tr>
<td><strong>Observations</strong></td>
<td>447</td>
<td>447</td>
</tr>
<tr>
<td><strong>First-stage F-statistic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emotions</td>
<td>66.11 (P &lt; .001)</td>
<td></td>
</tr>
<tr>
<td>Centrality</td>
<td>27.85 (P &lt; .001)</td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{a}All models are estimated controlling for the full list of covariates but are omitted from the tables due to space concerns. Results are available from the authors.

\textsuperscript{b}Confidence intervals (CI) for coefficients are obtained from 1,000 bootstrap replicates. A coefficient is significant at \( P < .05 \) if 0 is not in 95% CI.

\textsuperscript{c}F-statistic tests the joint significance of the two instruments from a first-stage regression of a user’s emotions on followees’ emotions and followees’ centralities (i.e., the instruments) plus the rest of the covariates.

\textsuperscript{d}F-statistic tests the joint significance of the two excluded instruments from a first-stage regression of a user’s centrality on followees’ emotions and followees’ centralities (i.e., the instruments) plus the rest of the covariates.

Estimation Results of Survival Models

In the survival models, we only consider users who were active past our first observation period, so as to examine the effect of network centralities in our first-observation on users’ activities in the second-observation period. Table 3 shows mean coefficients of emotions and network centrality in the survival models. Following [61], the effects of all covariates are assumed to be time dependent in estimations. Both the standard and IV models on the identical-interval censored data show that (i) positive emotions lead to a shorter survival time (\( P < .05 \) in the IV model), and (ii) a core position in social networks is associated with a longer survival time (\( P < .05 \) in both models). Estimations on the personalized-interval censored data and using different IV specifications give similar results. The strong relevance of IV in the first stage regressions confirms the validity of IV across different models. A comparison of results between the linear and survival models further shows that these models have consistent estimators for the effect of emotions on dropout, i.e., positive emotions increase the likelihood to drop out.

Underlying Connection between Emotions and Dropout

To better understand the relationships between emotions and dropout, we examine posting interests among users with different dropout statuses and emotional states based on hashtags used in users’ tweets (see SI for details). We find non-dropouts are interested in advocating a thin ideal (e.g., using hashtags “mythinspo” and “skinny4xmas”) and promoting a pro-ED identity (e.g., “edlogic” and “beautiful”). In contrast, dropouts engage in discussing their health problems (e.g., “selfharmprobz”, “bulimicprobz” and “anorexicprobz”) and offering emotional support for others (e.g., “anasisters” and “stayingstrong”), which implies a tendency of these users to recover from disorders [20–22]. Similarly, we split all ED users into three equal-size sets based on their emotional scores and examine hashtags used by each set of users. We find that users with negative emotions often engage in promoting thin ideals (e.g., “bonespo” and “mythinspo”), showing largely overlapping interests with the non-dropouts. In contrast, users with neutral and positive emotions are more interested in discussing their health problems (e.g., “anorexicprobz” and “bulimicprobz”), opposing pro-ED promotions (e.g., “reversethinspo”) and encouraging healthier body image and behaviors (e.g., “fitfam” and “fitness”), showing similar interests with the dropouts. See SI for more detailed lists of hashtags.

Measuring the Spearman rank correlation \( \rho \) between pairwise lists of hashtags posted by users with a given state (e.g., dropped-out or not, and positive or negative), we find a positive correlation between negative users and non-dropouts in hashtag usage (\( \rho = 0.36, P = .003 \) in Table 4), indicating similar posting interests among these users. A similar pattern occurs between positive users and dropouts. In contrast, users with other pairs of states show a negative correlation or non-correlation in hashtag usage, indicating their discrepancies in posting interests. These results reveal a possible underlying connection between positive emotions and dropout. Compared to users with positive emotions, those with negative emotions have more similar interests to active members (i.e., non-dropouts) in the ED community. Finding similarities with other members in a community can enhance
Discussion

It is not surprising that negative users are less likely to drop out than positive users in our estimations. A sense of belonging to the community and positively increase intention to engage in community activities [33, 37]. Therefore, it is not surprising that negative users are less likely to drop out than positive users in our estimations.

Principal Findings

This study provides the first estimates of the effects of personal emotions and interpersonal social networks on dropout in online ED communities. The present work has several strengths. First, we base our analysis on incentive theory to explore determinants of users’ online behaviors (i.e., dropout), allowing us to study users’ behaviors in a more systematic way than most prior studies that focus on a single type of determinant (e.g., individual attributes [35, 36] or social attributes [42, 43, 67]). Second, we use automated sentiment analysis techniques to measure users’ emotions and network analysis methods to quantify users’ sociometric statuses in an online community, leading to higher efficiency than traditional research methods such as surveys [35, 37, 39, 41, 48]. Third, we apply an IV approach to both linear probability and survival models, which enables us to achieve a more consistent estimate of human behavior in online settings than traditional methods (e.g., OLS) used in prior studies [39, 41, 47]. Overall, we find that positive emotions increase the likelihood of dropout in ED individuals and accelerate the dropout process on Twitter. In contrast, a central position in the social network of ED individuals at an earlier stage is associated with prolonged participation of an individual at a later stage. These findings are verified across a variety of robustness checks.

Despite differences in methodology, our findings align with prior studies in psychological and social media research [5, 33, 35]. Our results suggest that ED users with negative emotions have high levels of participation on Twitter. This aligns with prior survey studies on social media use (e.g., Facebook use), where people with social anxiety and shyness (i.e., personality traits that are often correlated with multiple negative emotions such as feeling lonely, isolated and unhappy [80]) are found to spend more time online [35, 48, 81]. An explanation for this is the online disinhibition effect [82], i.e., because of anonymity in online interactions, people with social inhibitions (e.g., those who are socially anxious or shy and those with a stigmatized health problem [83]) might be more willing share personal feelings and reveal themselves in online interactions than offline interactions, in order to meet their social and intimacy needs [48]. Additional analyses on users’ posting interests reveal that users with negative emotions share similar interests with active users. This allows us to confirm the validity of our results via the social capital theory [39, 40], i.e., sharing common attributes (e.g., interests and visions) with other members can enhance a sense of belonging and positive feeling toward a community, which drives people to actively engage in the community.

Consistent with positive associations between network centrality and active participation in other online communities [67, 70], we find that central users in the social network of an ED community tend to have a longer-lasting participation in the community. This result is to be expected for several reasons. First, users who are centrally embedded in a group have a relatively high number of social ties to other members, which can lead these users to feel being socially accepted and approved, as well as a strong sense of belonging to the group. Prior studies have consistently shown that recognition from other members and identification within an online community increase an individual’s commitment to the community [34, 39–41]. Second, information shared by central users is likely to spread to the majority of a community through social ties, and their central positions in the community may promote other members to trust such information [70]. This implies that central users have a greater potential than peripheral

\[ \rho \]

\[ \rho \] values testing for non-correlation are reported in parentheses.

Table 4. Spearman rank correlations between pairwise lists of hashtags posted by users with a given dropout state and by users with a given emotional state respectively.

<table>
<thead>
<tr>
<th></th>
<th>Negative (n = 61)</th>
<th>Neutral (n = 108)</th>
<th>Positive (n = 110)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-dropout (n = 54)</td>
<td>0.36 (P = .003)</td>
<td>-0.21 (P = .03)</td>
<td>-0.66 (P &lt; .001)</td>
</tr>
<tr>
<td>Dropout (n = 227)</td>
<td>-0.33 (P &lt; .001)</td>
<td>-0.04 (P = .57)</td>
<td>0.12 (P = .07)</td>
</tr>
</tbody>
</table>

\( \rho \) values testing for non-correlation are reported in parentheses.

\[ \rho \] values testing for non-correlation are reported in parentheses.

\[ \rho \] values testing for non-correlation are reported in parentheses.
users in influencing members’ opinions, emotions and behaviors in online communities [84]. Thus, compared to peripheral users, feeling influential may provide an additional incentive for central users to continue participating.

In line with prior studies on online ED communities [15, 17, 21], we find that ED users on Twitter have different stances on ED, where users with negative emotions often share pro-ED content and those with positive emotions often share pro-recovery content. As pro-ED content often contains thin-ideal images and harmful tips for weight loss/control [8, 24, 25], this result aligns with clinical evidence on ED treatment showing that more emotional distress is associated with a higher risk to learn and develop dysfunctional coping behaviors among ED sufferers [5]. Thus, as suggested by prior studies [85], engaging in pro-ED content may serve as a coping mechanism to deal with emotional pressures and stress of ED. A possible explanation for the association between engaging in harmful online content and coping with stress is sensation seeking [86], a basic personality trait defined as the seeking of varied, novel, complex and intense sensations and experiences, and the willingness to take risks. Several studies have shown that sensation seeking is prominent in adolescence (i.e., the age that disordered eating often develops [1]) and closely related to pathological Internet use, such as use of violent sites [87] and Internet dependence [88].

Our study also offers new insights into online ED communities. First, ED users have a high dropout rate (85% in our sample) and a short lifespan between an account creation to lost posting on Twitter (with 6 months of median time to drop out). This aligns with views of online ED communities as hidden, secretive groups [30], but also indicates the dynamic characteristics of these communities. Second, users who discuss their health problems and share pro-recovery content (i.e., pro-recovery users) have lower levels of posting activities (i.e., a higher dropout rate) than those who share pro-ED content (i.e., pro-ED users) on Twitter. This can be explained as follows. Due to common interests in ED, pro-recovery and pro-ED groups are likely to be connected in the same social networks, and content shared within a group is hence likely to be visible to the other group. However, exposure to content from the antagonist group can have distinct effects in pro-ED and pro-recovery groups. Exposure to pro-ED content is harmful for pro-recovery users and can impede their recovery process [3, 24], while exposure to pro-recovery content can instead stimulate harmful behaviors in pro-ED users (e.g., actively sharing pro-ED content) [21]. Thus, pro-recovery users might tend to leave such an online community to avoid a risk of further deterioration or relapse. Our finding may also explain why pro-ED content is found being more pervasive than pro-recovery content across social media sites [15, 17, 21], e.g., almost five times in terms of unique publishers on Tumblr [15]. Third, ED users tend to connect with others with the same dropout states on Twitter. This implies that whether an individual drops out from online communities depends on whether others in the individual’s social networks drop out. In other words, dropout in online ED communities is not only a function of individual experience or individual choice but also a property of group interactions, e.g., homophily [89] and social contagion effects [59].

Implications

Our findings are of practical relevance to the promotion of public health over social media. First, the decision to maintain active participation in an online community can be caused by intrinsic and extrinsic characteristics/traits of the participants, e.g., personal emotions, interests and social networks. Such self-selection bias can lead to the sample not being representative of the whole population, and hence researchers need to consider both active and dropped-out users for a well-rounded picture of online health communities. This is particularly important for public health officials to make special efforts to reach these dropouts and offer more intensive support when they are trying to recover. Second, high attrition rates are often regarded as negative outcomes in online interventions, particularly in those delivered over a purposely designed website [11, 31, 32]. However, this may or may not be the case in interventions over general social media sites (e.g., Twitter) depending on how targeted populations use these sites. For example, when an intervention is delivered in an online community in which members often shared harmful content, a high attrition rate (i.e., members dropping out of the harmful community) may not be a negative outcome. Using automated data-mining techniques to track users’ behaviors (e.g., emotions and posting interests), as used in this work, can provide more detailed information about people’s use of online health communities and improve our understanding of attrition in online interventions. Third, interventions that recommend content containing positive emotions to ED users (not limited to ED-related content but more general content containing happiness and inspiration) may reduce their engagement in a harmful online community. This aligns with Fredrickson’s broaden–and–build model which argues that cultivating positive emotions is useful to prevent and treat mental health problems [90]. Finally, intervention strategies could be tailored for different individuals depending on their positions in the social network of an online community. For example, identifying central individuals as change agent might enhance the efficacy and cost effectiveness of an intervention, due to their greater influence potential through larger numbers of social ties [91], but also their longer-lasting effects through longer-term participation in the community.
Limitations
First, we recognize that self-diagnosis information on Twitter may be itself self-censored by users to align with their personality traits and perceptions of their audience on the platform. People may not use tags like “eatingdisorder” to self-report their experience of illness and would be excluded by our collection methods. Also, although over 208K users and over 241M tweets are studied in this work, a small sample of rich social media data is used to explore the attrition of ED communities on Twitter. Thus, generalization of our results to all ED-related online communities should be cautious. Second, our measures of dropout are based on posting activity, while some people primarily use Twitter to receive outside information but rarely post their own information. We have little activity data on these users and hence less understanding of the characteristics of their dropout. This thus raises important issues that need further research to enhance our understanding of attrition in online health communities, such as consensus and clarity about the definition of dropout. Third, our study focuses on the Twitter platform, without validation on other platforms. However, stopping using a platform can be related to the attractiveness of the platform. Hence, future research is also in need to examine many other factors that we did not explore but can affect dropout on social media, such as individual personality, physical health states, perceptions and purposes of using a particular social media platform. Fourthly, user accounts on a social media site are often not unique—an individual may have multiple accounts on the same site. Thus, we cannot be certain whether individuals who stopped using an account will engage in the same or similar online communities through other accounts. That is, the dropout of a user account may not necessarily imply that an individual abandons a specific identity (e.g., pro-ED) shared within a community. Finally, although our methodology allows us to establish a causal link from emotions to dropout behavior, it offers limited insights into the pathways through which this link exactly operates and future work is needed to explore such issues in detail.

Conclusions
This study presents a systematic characterization of attrition in an ED community on Twitter. Our analysis offers the first attempt towards the estimation of the effects of personal emotions and network centrality on dropout behaviors in individuals affected by ED on Twitter. Our results provide new insights into the trajectories that ED communities develop online, which can help public health officials to better understand individual needs in using online ED communities and provide tailored support for individuals with different needs.

Acknowledgement
This work is supported by ESRC Doctoral Training Centre (NO. ES/J500161/1), Institute for Life Sciences, WSI-RCSF and SocSCI-SIRF, University of Southampton, and The Alan Turing Institute, UK.

Conflicts of Interest
None declared.

References


