REGULAR ARTICLE

Characterizing dynamic communication in online eating disorder communities: A multiplex network approach

Tao Wang[1,2*], Markus Brede[3], Antonella Ianni[4,5] and Emmanouil Mentzakis[4]

*Correspondence:
t.wang@soton.ac.uk
[1]ESRC Doctoral Training Centre, University of Southampton, Southampton, UK
[2]The Alan Turing Institute, London, UK
Full list of author information is available at the end of the article

Abstract

Growing evidence shows that social media facilitate diffusion of both pro-recovery and anti-recovery information among people affected by mental health problems, while little is known about the associations of people's activities in sharing different types of information. Our work explores this question by analyzing a large set of Twitter conversations among users who self-identified as eating disordered. We use clustering algorithms to identify topics shared in online conversations and represent interpersonal interactions by a multilayer network in which each layer represents user-to-user communication on a different topic. By measuring structural properties of the multilayer network, we find that (i) the same set of users form social networks with different structures in communicating different types of information and (ii) exposure to content on body image can reinforce individual engagement in anti-recovery communication and weaken engagement in pro-recovery communication. By measuring structural changes in a sequence of temporal, multilayer networks built based on users' conversations over time, we further find that (i) actors previously engaged in pro-recovery communication are likely to engage in anti-recovery communication in the future and (ii) actors in anti-recovery communication have frequent entries into and exits from such communication system. Our results shed light on the organization and evolution of communication in online eating disorder communities.

Keywords: eating disorders; social media; multilayer network; network dynamics

1 Introduction

Eating disorders (ED) such as anorexia and bulimia are complex mental illnesses that can cause serious health consequences and have the highest mortality rate of any mental illness [1, 2]. Despite such negative effects, there are many online pro-ED (often known as pro-anorexia and pro-ana) communities that actively promote ED as a legitimate lifestyle choice rather than a dangerous disease [3, 4, 5]. These pro-ED communities can negatively affect people with and without ED, through promoting unrealistic ideals of thinness, encouraging disordered eating behaviors, and sharing harmful tips on how to develop and maintain ED (known as "thinspiration" or "thinspo") [6, 7, 4, 8, 9]. As a public concern, pro-ED communities draw widespread criticism, particularly by so-called pro-recovery communities that aim to raise awareness of ED and offer support for people to recover [10, 11, 12, 13, 14]. Under pressures from these pro-recovery communities and the general public, several social media platforms have adopted censorship-based interventions for pro-ED

communities, e.g., banning pro-ED content and user accounts on Tumblr[1] and Instagram[2] [15, 16, 17]. However, the efficacy of these interventions is still uncertain. Concerns about this are heightened by recent findings that censoring pro-ED content leads to a wide spread of more harmful alternatives to such content (e.g., content sharing self-harm) [15, 16], and banning pro-ED users makes these individuals more "invisible", less reachable by health care providers and recovery-oriented information [17]. These findings highlight the importance of understanding how different types of information (not only pro-ED content but also ED-related content more generally) flow through an online ED community and how these information flows correlate with one another, before introducing interventions.

However, our understanding of information flows in online ED communities is limited, as prior studies in this field have often focused on content analysis and largely ignored interaction patterns. Examples of these analyses are examining the types of content shared in online ED communities [18, 19, 20, 21, 22, 23], characterizing linguistic styles of individuals in online self-presentation [10, 12], identifying diagnostic information from language use in pro-ED content [24, 13], detecting lexical variation of pro-ED content [15, 16], and measuring people's attitudes on pro-ED and pro-recovery information based users' emotional expressions in online comments [9, 11]. Yet, social interactions in information exchange and the resulting communication networks have been largely under-explored. As a result, little is known about the organizational structure of communication in online ED communities and the functional roles of individuals in communicating different types of information.

Although recent studies have turned attention from content analysis to network analysis [25, 26, 27, 28], they either focus solely on a single type of communication (e.g., sharing pro-ED content [28]) or do not distinguish different types of information shared in online ED communities [25, 26, 27]. It remains unclear how different types of communication correlate with one another in these communities. Insights into the correlations among different types of communication can facilitate predictions of an community's responses to interventions. For example, if users' activities in two types of communication have a highly positive correlation, blocking one type of communication is likely to promote the other type of communication.

In this work, we address these research gaps by using a multilayer network approach to systematically characterizing communication networks for a broad range of types of content in an online ED community. We analyze a large set of Twitter conversations (i.e., tweets with a "mention" or "reply") among individuals who self-identified by having ED in their Twitter profile descriptions and their online friends, involving 2,206,919 tweets posted by 55,164 users over 7 years (from March 2009 and March 2016). Three major research questions guide our analysis: (i) what types of content are often discussed in an online ED community? (ii) how do different types of content flow through interpersonal communication networks? and (iii) whether and how do different types of communication correlate with one another? The main contributions of this work are as follows.

First, we demonstrate the use of unsupervised clustering methods to identify the types of content discussed in online ED communities. Unlike previous studies that

[1] https://staff.tumblr.com/post/18563255291/follow-up-tumblrs-new-policy-against

[2] http://instagram.tumblr.com/post/21454597658/instagrams-new-guidelines-against-self-harm

assume a type of content with predetermined features (e.g., a set of keywords) [9, 11, 12, 13, 25, 26, 28], our approach allows themes of content to emerge from the data, which can reduce bias due to predetermined assumptions and provide an overall view of the full range of topics discussed in an online ED community.

Second, we propose to represent types of communication in online ED communities by a multilayer network in which each layer is a network representing interactions among the same set of users in discussing a specific topic. Compared to traditional monolayer networks [29], multilayer networks provide (i) a more natural representation of a communication system by capturing the multiplex nature of human interactions [30, 31, 32], and (ii) a more elegant and flexible way for incorporating multidimensional information. Based on this multilayer representation, we (i) characterize different types of communication by measuring structures of single-layer networks in the multilayer communication network, and (ii) examine interdependencies of different communication by measuring structural correlations of inter-layer networks.

Finally, we study dynamics of user communication and reveal underlying processes that lead to correlations of communication on different topics. By measuring structural changes and stability in a sequence of temporal, multilayer networks that are built based on users' conversations over time, we find that (i) actors previously engaged in pro-recovery communication are likely to engage in pro-ED communication in the future and (ii) actors engaged in sharing pro-ED content have frequent entries into and exits from the corresponding communication network.

2 Data

Our dataset is collected from Twitter, a microblogging platform that allows millions of people to interact by exchanging short tweet messages. Whereas many social media sites restrict pro-ED content [15], Twitter has not yet enforced any restrictions [26]. This makes Twitter a unique platform to examine communication naturally happening within online ED communities in a non-reactive way. All data used in this study is publicly accessible information on Twitter; no personally identifiable information is used. Next, we provide details about the dataset we used.

2.1 Collecting user sample

We use a snowball sampling method [33] to gather data about individuals affected by ED on Twitter. We first search for ED-related tweets by a set of keywords (e.g., "eating disorder", "anorexia" and "bulimia") via the Twitter application programming interfaces (APIs). From authors of 1,169 ED-related tweets, we identified 33 users who self-reported both ED-diagnosis information (e.g., "eating disorder", "edprob" and "proana") and personal bio-information (e.g., height and weight) in their Twitter profile descriptions. Starting with these seed users, we use a snowball sampling procedure through users' who-follows-whom networks to expand the user set. This results in 3,380 *ED users* who self-identified as disordered in their profile descriptions. Our data validations show that 95.2% of the ED users are likely to be affected by ED (i.e., a high precision, see [33] for more details). However, the above process does not ensure a high recall, as we miss users who did not disclose their disorders in Twitter profile descriptions.

To obtain a more representative sample of online ED communities, we further collect ED users' Twitter friends (including followees and followers) who posted ED-related content in tweets. To this end, we first crawl all friends of each ED user on Twitter, yielding 208,065 users (including the 3,380 ED users). For each user, we retrieve up to 3,200 (the limit returned from the Twitter APIs) of their most recent tweets, resulting in 241,243,043 tweets. This collection process finished on March 2, 2016. Then, we search for users who posted an ED-related hashtag in their historical tweets (see Appendix, Section 1 for details), resulting in 41,456 ED-related users.

2.2 Tracking interpersonal conversations

The other task of our data collection is to track interpersonal communication of ED-related users. We focus on users' communication via the *"mention"* and *"reply"* interactions as these interactions are the two main ways to conduct direct communication on Twitter[3]. Also, as users can discuss a topic by sending and replying to tweets over several rounds, a single tweet message often cannot provide complete context to understand human communication. For example, it may be hard to recognize that user A might dissuade user B from committing suicide based on a single tweet *"@XXX please don't do it, I love you so much!"*, without considering that this tweet is user A's reply to user B's tweet *"9 30pm on the 8th of July 2012 I will hopefully die, so n/r going to write my suicide note"*. Thus, to obtain a relatively complete context in a discussion, we shift attention from single tweets to conversations, i.e., aggregations of successive tweets in a discussion [35].

Specifically, for each user, we search for their tweets that contain a mention or reply. Then, we aggregate tweets into conversations based on the *"in_reply_to_status_id"* field returned by the Twitter APIs. Each conversation consists of a seed tweet, all tweets in reply to it, and replies to the replies, which can involve several tweets and users. Mentions that do not receive any replies are considered as individual conversations. We obtain 1,044,573 conversations consisting of 2,206,919 tweets. All re-tweets are excluded, since the mentions or replies in a re-tweet are conducted by the original author of the re-tweet, not by users who re-tweet it. Detailed statistics of these conversations are presented in Appendix, Section 2.

3 Results

In this section, we present our analyses of Twitter conversations in online ED communities. These analyses involve three steps: (i) characterizing the types of content in users' conversations; (ii) examining how different types of content flow through interpersonal interactions and measuring structural correlations among different types of communication; and (iii) exploring how different types of communication correlate by analyzing temporal information.

[3] Although "re-tweet" interactions are also widely used to spread information on Twitter, it is hard to trace communication pathways in a re-tweeting network based on Twitter APIs. This is because, in the settings of Twitter APIs, all re-tweets of a tweet in each cascade are directly linked to the original tweet [34]. That is, if Bob re-tweets Andy and then Cole re-tweets Bob, both Bob and Cole are linked to Andy in a re-tweeting network, even though Cole did not re-tweet Andy directly.

3.1 Content analysis

Common methods for characterizing the types of textual content are topic modeling (e.g., latent Dirichlet allocation models [36, 35]) and content-based clustering methods (e.g., bag-of-words and word/document embeddings [37, 38]). However, these methods generated topics that were hard to interpret in our preliminary experiments. Previous studies have shown that these methods perform poorly when applied to short and noisy tweets [35]. Although we aggregate short tweets into a conversation, most conversations are still short (on average 21.9 words in each conversation) and they are often dominated by general chats (e.g., "why do you follow me?"). Inspired by prior work [34], we here characterize the types of users' conversations by identifying topics of hashtags used in these conversations. As hashtags are often used to annotate the theme of a tweet, these clusters of hashtags have been shown to effectively indicate the underlying topics in tweets [39, 40, 34].

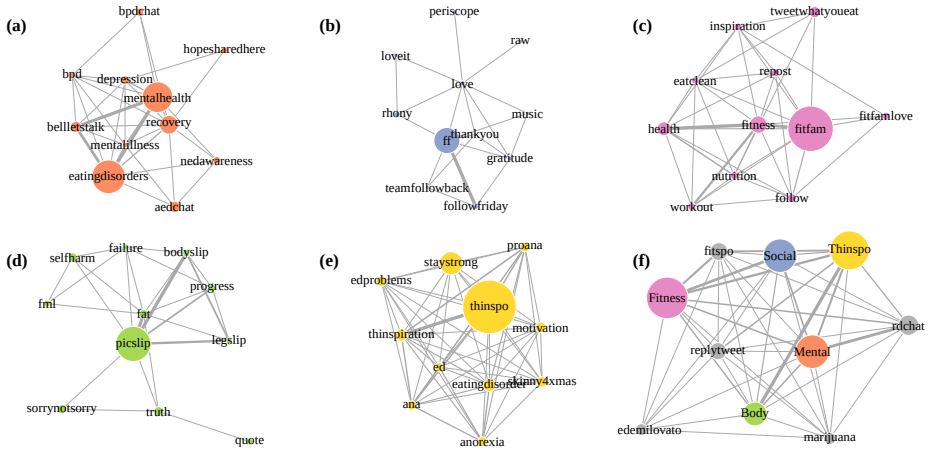

Figure 1 Topics in hashtag co-occurrence networks. (a-e) Co-occurrence networks of the most frequent hashtags in each of the five popular topics, where each node denotes a hashtag. The size of a node is proportional to the number of tweets with a hashtag and edge width is proportional to the number of co-occurrences between tags. (f) Co-occurrences of popular topics involving more than 1,000 users, where each node denotes a topic as labeled. The node size is proportional to the number of tweets mentioning a topic and edge width is proportional to the ratio of the number of tweets mentioning both topics over the number of tweets mentioning at least one topic.

We detect topics of hashtags by performing community detection in co-occurrence networks of hashtags. We build an undirected, weighted hashtag network based on the co-occurrences of hashtags in the tweets of users' conversations, where an edge is weighted by the co-occurrence count of hashtags. To filter out noise, only tags used by more than three distinct users and used in more than three tweets are considered. The resulting network contains 65,756 nodes and 109,663 edges, partitioned in 672 connected components, where 5,791 nodes are in the giant component and 4 in the second largest component. Due to the dominance, we focus on analyzing the giant component and obtain 26 topic clusters of hashtags by applying the Louvain method [41] to this network[4]. The resulting modularity is $Q = 0.51$ ($z = 6.63$ compared

[4] We also tried other well-established methods for community detection in networks, e.g., the Infomap algorithm [42]. These methods produced comparable results in our

to a random configuration model [43], $p < 0.001$ in a two-tailed test), indicating a clustered topic structure in the hashtag co-occurrence network. By examining the numbers of tweets and users related to each of the 26 topics identified above, we find that users have consistently high levels of engagement in five topics with IDs 2, 4, 8, 16 and 22 respectively, whereas other topics are much less popular (see Appendix, Section 3). To avoid analyzing topics of interest to a specific subgroup of online ED communities, we focus on the five popular topics in this work.

Figures 1(a-e) show the most frequent hashtags and their co-occurrence networks for each of the five popular topics. As shown in Figure 1(a), topic 2 is dominated by "#eatingdisorders", "#mentalhealth", "#recovery" and "#bellletstalk"[5], showing a clear tendency to support recovery from ED and promote mental health [25, 13]. We label this topic *mental*. In contrast, topic 4 (Figure 1(b)) is dominated by a single tag "#ff" which is likely to be an abbreviation of "#followfriday", given frequent co-occurrences between the two tags. These tags are often used in a weekly social events where people recommend their followers to follow more people on Twitter[6]. We thus label this topic *social*. Figure 1(c) shows that topic 8 is mainly concerned with fitness activities and diet (thus labeled *fitness*). Topic 16 (Figure 1(d)) is about "#picslip" which is often used by users to post a picture of themselves[7]. Other tags that highly co-occur with "#picslip" are "#bodyslip", "#fat", "#selfharm" and "#failure", indicating a theme of body image and body dissatisfaction, thereby labeled *body*. As shown in Figure 1(e), topic 22 is mainly about thinspiration (or pro-ED) content (e.g., "#thinspo" and "#proana") which is designed to inspire people to lose weight and stay extremely thin [18, 19]. We label this topic *thinspo*. Moreover, to illustrate the relationships of these popular topics, we visualize a co-occurrence network of popular topics in Figure 1(f).

To verify our results, we check (i) if the topics found above cover real-world events in ED communities, and (ii) if the relationships of these topics align with findings in prior qualitative studies on online ED content [18, 19]. Results of these checks strongly confirm the reliability of our content analysis (see Appendix, Section 4).

3.2 Network analysis

We proceed to explore how different types of content flow through interpersonal interactions using network analysis methods. To do this, we first categorize users' conversations based on the topics of hashtags found above. Given a conversation, we track the sequence of hashtags used in the conversation and annotate the topics of this conversation with the topic labels of these hashtags. To avoid ambiguous annotations, we only consider conversations that are labeled with only one unique topic; those with multiple topics or without a hashtag are excluded in our analysis[8]. This results in 102,554 conversations consisting of 201,155 unique tweets. Then, we

preliminary analysis. In this work, we use the Louvain method due to its efficiency of processing large-scale networks [41].

[5] An annual campaign on social media to break the silence around mental illness and support mental health: https://letstalk.bell.ca/en/

[6] https://www.urbandictionary.com/define.php?term=followfriday

[7] https://www.urbandictionary.com/define.php?term=picslip

[8] While this process reduces the size of our raw data, it can avoid biased results. For example, one could build classifiers based on content features (e.g., bag-of-words

represent information about who interacts with whom and on which topic in users' conversations via a multilayer network with $N = 55,164$ nodes representing users and $M = 5$ layers representing topics. The multilayer network can be described by a set of M adjacency matrices, one for each layer, $G = [A^{[1]}, A^{[2]}, ..., A^{[M]}] \in \mathbb{R}^{N\times N\times M}$. Each layer $A^{[\alpha]}$ is a directed, weighted network, in which a link $a_{ij}^{[\alpha]}$ runs from a node representing user i to a node representing user j if i mentions or replies to j in the conversations on topic $\alpha = 1, 2, ..., M$, weighted by the frequency of these mentions and replies.

Based on this multilayer representation, we characterize communication patterns in online ED communities by quantifying structural properties of the multilayer network. First, we measure structural properties of single-layer networks (i.e., each layer is considered as a separated network) to examine organizational features of each type of communication. Second, we measure inter-layer dependencies in the multilayer network (i.e., structural correlations between inter-layer networks) to explore associations of different types of communication.

3.2.1 Structures of single-layer networks

We first examine structures of single-layer networks to explore the organization of an online ED community in a type of communication. Figure 2 shows cumulative in- and out-strength distributions of each single-layer network, and Table 1 gives details about structural properties of these networks. The key results are as follows.

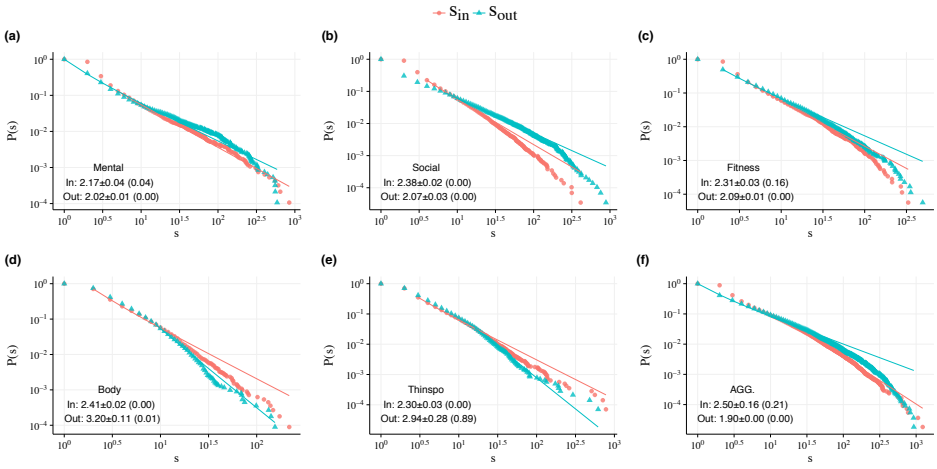

Figure 2 Distributions of in-/out-strength (s_{in}/s_{out}) at each layer and the full aggregated network (AGG.). All values of s are shifted by 1 to account for nodes with zero values on the log-log plots. Lines fit a power-law distribution $P(s) = s^{-\lambda}$ using the maximum likelihood estimator and a p-value for the goodness of fit is obtained using a bootstrapping procedure [45]. The mean values and standard deviations of exponents λ are shown in the legends, and p-values obtained via 1,000 bootstrap replications are reported in parenthesis.

Users' engagement levels in posting harmful content have skewed distributions. Figures 2(a-e) show two distinct behaviors in in- and out-strength (s_{in}

[44]) in conversations labeled with a single topic to predict the most likely topic for conversations labeled with multiple topics and conversations without a hashtag. This however can introduce classification errors and noise data.

Table 1 Statistics of single-layer networks and the aggregated (AGG.) network, including 1. the number of active nodes $N^{[\alpha]}$, i.e., nodes that are connected by at least one in-/out-link [30]; 2. the total number of edges $E^{[\alpha]}$; 3. the average strength $\langle s^{[\alpha]} \rangle$; 4. density $D^{[\alpha]}$ measuring the ratio of the number of edges to maximum possible number of edges; 5. fraction of nodes in the giant weakly connected component $\%G^{[\alpha]}$; 6. reciprocity $r^{[\alpha]}$ quantifying the likelihood of nodes with mutual links; 7. the Kendall's τ correlation between in- and out-strengths $\tau(s_{in}^{[\alpha]}, s_{out}^{[\alpha]})$. 8. global clustering coefficient $C^{[\alpha]}$ which measures the extent that two neighbors of a node are connected; 9. assortativity coefficient by strength $A_s^{[\alpha]}$, i.e., the correlation between the out-strengths of source nodes and the in-strengths of destination nodes [46]. Values of $z(x)$ are z-scores for the empirical results based on null models. For each property x of a network, we generate 1,000 randomized networks via the configuration model [43] and measure the property in these randomized networks. Then, the deviation of x from randomness is quantified by a z-score: $z(x) = (x - \langle x \rangle)/\sigma_x$, where $\langle x \rangle$ is the mean value of the property in randomized networks and σ_x is the standard deviation.

Network	Mental	Social	Fitness	Body	Thinspo	AGG. (all αs)
$N^{[\alpha]}$	9,381	28,959	17,689	11,199	14,156	55,164
$E^{[\alpha]}$	17,306	54,609	34,040	17,881	27,807	140,330
$\langle s^{[\alpha]} \rangle$	3.55	2.89	2.94	2.46	2.96	4.32
$D^{[\alpha]}$	1.97×10^{-4}	6.51×10^{-5}	1.09×10^{-4}	1.43×10^{-4}	1.39×10^{-4}	4.61×10^{-5}
$\%G^{[\alpha]}$	76.55%	89.17%	83.84%	73.87%	88.87%	95.67%
$r^{[\alpha]}$	0.24	0.19	0.36	0.45	0.33	0.29
$\tau(s_{in}^{[\alpha]}, s_{out}^{[\alpha]})$	-0.06	-0.04	0.09	0.21	0.13	0.11
$C^{[\alpha]}$	0.06	0.04	0.03	0.03	0.01	0.03
$z(C^{[\alpha]})$	40.70	198.96	61.25	109.21	6.29	160.67
$A_s^{[\alpha]}$	-0.08	-0.07	-0.1	-0.02	-0.08	-0.1
$z(A_s^{[\alpha]})$	-10.64	-17.24	-19.05	-4.60	-14.04	-37.80

and s_{out}) distributions of single-layer networks. Specifically, the distributions of s_{in} are more skewed than those of s_{out} in the *mental*, *social* and *fitness* layers, while the distributions of s_{out} are more skewed in the *body* and *thinspo* layers. These behaviors can be quantified by fitting a power-law function $P(s) = s^{-\lambda}$. We find that all networks have comparable values of λ in s_{in} distributions, indicating similar patterns of popularity ranking for actors in different interactions. However, the s_{out} distributions in the *body* and *thinspo* layers ($\lambda \approx 3$) have a larger value of λ than those in the *mental*, *social* and *fitness* layers ($\lambda \approx 2$). As exposure to thin-ideal content (*thinspo* and *body*) is associated with higher risks of ED [47, 48], this implies that the fractions of users who actively post harmful content are relatively small.

Private communication takes place in small groups. As shown in Table 1, *mental* and *body* layers have lower fractions of nodes in the giant weakly connected component $\%G^{[\alpha]}$ than other layers, revealing that users tend to form smaller communities when discussing mental health and body image. This may be related to the private nature of these topics—due to fear of rejection and feelings of shame [49, 50], people are more likely to talk about their illnesses and body image to someone they can trust rather than any friends online.

Interactions related to body image are reciprocal. Table 1 shows that interactions on body image and appearance management (*fitness*, *body* and *thinspo*) have higher degrees of reciprocity $r^{[\alpha]}$ than those on other types of content (*mental* and *social*). High reciprocity indicates a tendency to reciprocate the interactions received from others, which can reward and reinforce these interactions [51]. The high degrees of reciprocity of interactions in the *fitness*, *body* and *thinspo* layers are confirmed by positive correlations between in- and out-strengths ($\tau > 0$), while $\tau < 0$ implies a suppression of reciprocity in the *mental* and *social* layers.

Users in general communication cluster. While the clustering coefficients $C^{[\alpha]}$ are low in each network (Table 1), the value of $z(C^{[\alpha]})$ in general commu-

nication (*social*) is larger than those in communication on specific topics (*thinspo* and *mental*). Higher values of $z(C^{[\alpha]})$ indicate that users are more likely to cluster together, compared to a baseline of random clustering. Such high value of $z(C^{[\alpha]})$ in general communication may be due to the fact that more general topics tend to be of interest to a wider variety of individuals, and a higher level of individuals sharing common interests leads to a more cohesive social community [52].

Private communication forms a weakly disassortative network. As Table 1 shows, all networks are characterized by disassortative mixing by strength ($A_s^{[\alpha]} < 0$), i.e., hubs tend to be attached to peripheral nodes, which aligns with prior evidence on online social networks [53]. Compared to null models, the disassortative strengths in private communication (*body* and *mental*) are relatively weaker than those in other communication (*social*, *fitness* and *thinspo*). This implies that people tend to discuss private topics with others who have similar social-status characteristics in a community.

The independent analysis of single-layer networks described above shows different organizational structures in different types of communication, highlighting the multiplex nature of human interactions [54, 55]. To demonstrate the disadvantage of not distinguishing types of communication, we include the statistics for the aggregated network (i.e., aggregating all single-layer networks in a single network) in Figure 2 and Table 1. We see that ignoring the differences of interactions leads to the loss of essential information and a misrepresentation of the system, e.g., losing information on differential network structures between harmful and healthy communication.

3.2.2 Dependencies between inter-layer networks

We next extend the independent analysis of single-layer networks to analysis of interdependencies between these networks. The aim of this interdependency analysis is to examine the correlations of individuals' activities and their functional roles in different types of communication. We consider the following measures.

Activity correlation: the tendency of users to be involved in one type of communication if they are involved in another type of communication. This can be measured by multiplexity, i.e., the fraction of nodes that are active at both layers α and β in all nodes of a multilayer network [30].

Role correlation: the extent to which hubs (e.g., those users who have high popularity or active engagement) in one type of communication are also hubs in another type of communication. We measure this by the Kendall's τ rank correlations of nodes' in-/out-strengths between two layers of the multilayer communication network. To avoid bias due to a low degree of multiplexity in real-world networks [30], we only consider nodes that are active in both layers.

Link overlap: the tendency that user i connects to user j in both types of communication. This can be measured by the Jaccard coefficient between two sets of links (binary links) at two layers [54].

Link-strength correlation: the extent to which user i has frequent interactions with user j in two types of communication. We measure this by Kendall's τ correlation of link strengths between two layers. Due to the sparseness of connections in real-world networks (see Table 1), we only consider links between two nodes that are present in the two layers.

These measures alone, however, are not adequate for evaluating inter-layer correlations. This is because the values of these measures are influenced by the size and connectivity of each single-layer network, which can be related to the processes of data collection and content categorization discussed in the previous sections. For a reliable evaluation, we need to assess the statistical significance of a correlation result. A standard statistical approach for distinguishing patterns of networks from those generated by chance is null models [56, 43]. A null model generates patterns by randomizing an observed network many times under proper constraints; an observed pattern that differs from the distribution of randomly generated patterns is potentially derived from meaningful processes rather than chance [57, 58]. According to the null hypothesis in question, null models can have different constraints and randomization processes. Here, we consider four null models for testing hypotheses of interest (see Table 2). In each model, randomized networks in each layer have the same sizes (i.e., the numbers of active nodes and edges) as the original ones, so as to control for the effects of data collection and content categorization on inter-layer correlations. Details of these null models are introduced in Appendix, Section 5.

Table 2 Null hypotheses on correlations of individuals' activities and roles in types of communication.

Correlation	Null hypothesis	Null model
Activity correlation	The activities of users in a type of communication are unrelated to those in other types of communication.	Hypergeometric model [30]
Role correlation	The roles of users in a type of communication (i.e., their positions in a type of communication network) are unrelated to those in other communication.	Independent multilayer node-permutation model [59]
Link overlap	Users' interconnections in one type of communication are unrelated to those in other communication.	Independent multilayer configuration model [58]
Link-strength correlation	The strength/frequency of interactions between two users in one type of communication is unrelated to those in other communication.	Independent directed-weight reshuffling model [60]

We generate 1,000 randomized multilayer networks for each null model, and measure a z-score for the empirical value of an inter-layer correlation measured in the original network x as $z(x) = (x - \langle x \rangle)/\sigma_x$, where $\langle x \rangle$ and σ_x are the mean and standard deviation of the values of x measured in randomized networks respectively. The results are shown in Figure 3, which can be summarized as follows.

Social networks in the *body* layer bridge those in the *mental* and *thinspo* layers. Figure 3(a) shows z-scores of inter-layer multiplexity compared to a hypergeometric model [30]. The largest z-score occurs between *body* and *thinspo* layers, indicating that the correlation of users' activities in sharing *thinspo* and *body* topics is much stronger than expected at random. On the other hand, while the overlap of actors in *mental* and *thinspo* layers is not significantly different from randomness, actors in the *mental* layer have a pronounced overlap with those in the *body* layer. These results imply that the group of users who engage in sharing *body* may bridge two groups who engage in sharing *mental* health and *thinspo* content.

Actors play different roles in healthy and harmful communication. Figures 3(b-d) show z-scores for in- and out-strength correlations of nodes in pairwise layers, as compared to an independent multilayer node-permutation model [59]. In most pairwise layers, nodes with higher in-/out-strengths in a layer tend to have higher in-/out-strengths in the other layer (Figures 3(b-c)), which indicates that popular/active users in a field are likely to be popular/active in the other field.

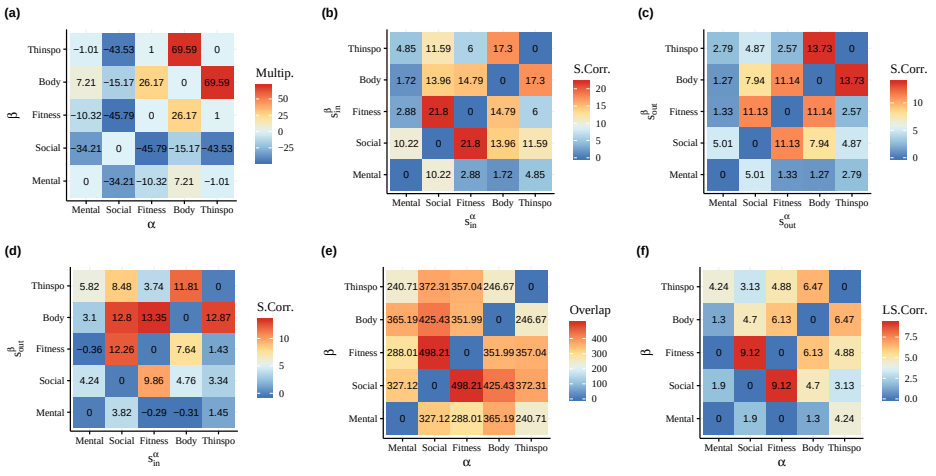

Figure 3 Deviations of empirical pairwise correlations of inter-layer networks to null models. (a) Multiplexity, (b-d) in-/out-strength correlations, (e) link overlaps and (f) correlations of link strengths, where $p < 0.05$ when $z < -1.96$ or $z > 1.96$ with assumptions of normality.

However, nodes' positions in the *mental* layer are not significantly correlated with those in the *body* layer, implying that actors may play a different role in these types of communication. Surprisingly, this pattern is absent between the *mental* and *thinspo* layers, i.e., nodes' positions in these layers are significantly correlated. A possible reason for such correlations is that pro-recovery users who actively post *mental* health may send healthy information to pro-ED users who post *thinspo* content as interventions [25]. This can be illustrated by the results in Figure 3(d) that nodes with higher out-strengths in the *thinspo* layer are likely to have higher in-strengths in the *mental* layer. That is, users who post more *thinspo* content tend to receive more content on *mental* health. Figure 3(d) also reveals users' responses when receiving different content. For example, nodes with higher in-strengths in the *fitness* and *body* layers tend to have higher out-strengths in the *thinspo* layer and lower out-strengths in the *mental* layer. This indicates that receiving more *fitness* and *body* content may reinforce users' engagement in posting *thinspo* content and reduce their engagement in posting *mental* health.

People often connect to the same friends in different types of communication. Figure 3(e) shows z-scores for overlaps of links in pairwise layers, as compared to an independent multilayer configuration model [58]. We see that high z-scores show in each pair of layers, indicating that users generally tend to connect to the same friends when discussing different topics. This aligns with prior evidence that people are often surrounded by a relatively stable social network [61].

Strengths of interactions on *mental* health generally have no significant correlations with those on other content. Figure 3(f) shows z-scores for correlations of link strengths, compared to an independent directed-weight reshuffling model [60]. A notable pattern is that users who often exchange content of *mental* health have no significant tendencies to frequently discuss other topics such as *social*, *fitness* and *body*. This can arise from two different processes: (i) actors in the *mental* layer exclusively focus on discussing *mental* health, while largely ignoring

interactions on other topics; and (ii) actors who previously engaged in other topics are less likely to engage in discussing *mental* health later. Distinguishing the two processes requires detailed time information on different interactions, which will be discussed in the next section.

3.3 Analysis of temporal patterns

To better understand the relationships among different types of communication, we further consider the time dimension of Twitter conversations and examine the dynamics of communication networks over time. Compared to the above analysis on static networks, dynamic analysis on temporal networks allows to explore how users start/stop to engage in a topic and change interests from one topic to others, yielding further insights into the correlation patterns of different types of communication.

To this aim, we represent temporal information about who interacts with whom on which topic and when in Twitter conversations using temporal multilayer networks. Specifically, we divide users' conversations into multiple sub-sets over time periods $1, ..., T$, based on the posting timestamp of a tweet. To reduce potential bias due to intermittent posting activities of users and temporal popularity of topics online[9], we build temporal networks by a fixed number of tweets instead of a fixed time interval. We rank all tweets by a chronological ordering and partition the tweets into subsets with a fixed number of tweets. The number of subsets is estimated by the Freedman-Diaconis rule [62], resulting in 55 subsets. For conversations in a subset at period $t \in [1, ..., T]$, we build a temporal multilayer network $G_t = [A_t^{[1]}, A_t^{[2]}, ..., A_t^{[M]}] \in \mathbb{R}^{N \times N \times M}$ in the same way that we build the static multilayer network, where M layers representing M topics and N nodes representing N users are fixed over time. Detailed statistics for these temporal multilayer networks are reported in Appendix, Section 6.

Based on these temporal networks, we study the dynamics for users' communication in two ways. First, we measure the likelihood of users engaging in a type of communication given that they have engaged in other types of communication. Clarifying such likelihood is not only useful to understand how the above correlation patterns appear among different types of communication, but also helps to identify signs suggestive of engagement in a type of communication, e.g., risk factors for engaging in harmful communication. Second, we examine the stability of a community of users who engage in a type of communication over time, particularly on investigating the presence of hardcore actors who have long-standing involvement in a type of communication. Evidence from this investigation can give insights into what strategies are likely to achieve quality, cost-effective outcomes in interventions. For example, if a type of communication is mainly carried out by a fixed set of hardcore actors, banning a small number of these actors can lead to serious damage to the connectivity of the communication network [63] and reduce the efficacy of the network in shaping individual cognition and behavior [64], while banning a larger number of actors at random may have limited influence on the network [65].

[9] As shown in Figure S2(d) of Appendix, users are highly active in posting tweets at some time periods, e.g., in 2013. This can be related to several factors, e.g., users in our sample might have high levels of engagement at these periods (i.e., sampling bias), or some topics were popular at these periods (i.e., environmental factors).

3.3.1 Transition of engagement activities

We first examine how users change their engagement between types of communication by measuring transitions of nodes' activities across layers in temporal multilayer networks. As users can engage in discussing multiple topics at the same time period, following prior work [30], we represent the activity state of node i across layers at time t by a node-activity vector $b_{i,t} = (b_{i,t}^{[1]}, ..., b_{i,t}^{[M]})$, where $b_{i,t}^{[\alpha]} = 1$ if node i is active at layer α of G_t (i.e., user i engages in topic α at time t) and $b_{i,t}^{[\alpha]} = 0$ otherwise. For computational efficiency, each binary vector $b_{i,t} = (b_{i,t}^{[1]}, ..., b_{i,t}^{[M]})$ is encoded as a decimal integer $R_{i,t} = \sum_{m=1}^{M} b_{i,t}^{[m]} \cdot 2^{M-m}$, where $R_{i,t} = 0$ indicates that node i has no interaction with others at time t and $R_{i,t} = 2^M - 1$ indicates that node i interacts with others in discussing all topics at time t. Then, we measure the transitions of users' engagement from a set of topics to another set by the period-to-period transition probability of node i from state $R_t = x$ to state $R_{t+1} = y$[10] as:

$$P(R_{t+1} = y | R_t = x) = \frac{\sum_{t=1}^{T-1} \sum_{i=1}^{N} I(R_{i,t} = x, R_{i,t+1} = y)}{\sum_{t=1}^{T-1} \sum_{i=1}^{N} I(R_{i,t} = x)}. \quad (1)$$

where $I(R_{i,t} = x, R_{i,t+1} = y)$ is an indicator function denoting whether node i has both an activity state $R_{i,t} = x$ at time t and a state $R_{i,t+1} = y$ at $t+1$, defined as:

$$I(R_{i,t} = x, R_{i,t+1} = y) = \begin{cases} 1 & \text{if } R_{i,t} = x \text{ and } R_{i,t+1} = y \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Similarly, $I(R_{i,t} = x) = 1$ if node i has an activity state $R_{i,t} = x$ at time t and $I(R_{i,t} = x) = 0$ otherwise.

Figure 4(a) shows results of transition probabilities $P(R_{t+1}|R_t)$ in our data, where we only consider nodes that are active in at least one of the two successive periods, i.e., $R_{i,t} + R_{i,t+1} > 0$. These results reveal the following patterns.

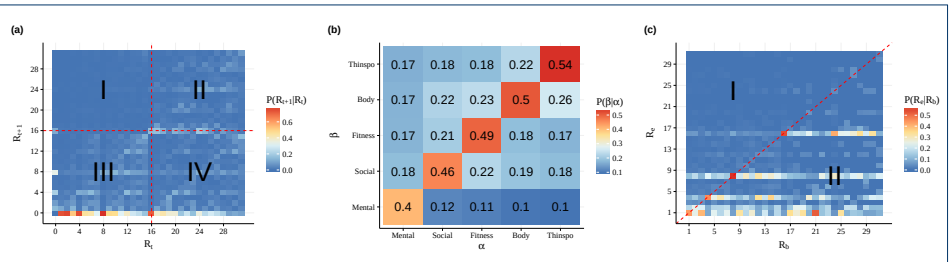

Figure 4 Transitions of engagement in different topics. (a) Transition probabilities of topic engagement between two subsequent observations R_t and R_{t+1}; (b) Fractions of users who posted content on topic α earlier will post β later (note that $\sum_\beta P(\beta|\alpha)$ is not necessarily equal to 1 as a user can post multiple different topics β after posting α); (c) Transition probabilities of topic engagement between the beginning R_b and the end R_e of participation.

Users tend to shift engagement from healthy communication to other communication. One notable pattern in Figure 4(a) is that the probability values

[10]For simplicity, we assume that the conditional probability for engagement at the next period depends only on the current state of engagement and not on the states of engagement at previous periods.

in region I, namely $P(R_{t+1} > 16|R_t < 16)$, are smaller than those in other regions. Since $R_t \geq 16$ and $R_t < 16$ denote whether nodes are active in the *mental* layer or not respectively, this result indicates that users who previously engaged in other topics are less likely to discuss *mental* health subsequently. In contrast, the values of $P(R_{t+1} < 16|R_t > 16)$ in region IV are relatively high, showing that users who previously talked about *mental* health tend to change to talk about other topics like *thinspo* (i.e., $R_{t+1} = 1$). Together, these results imply that users are more likely to shift engagement from pro-recovery to pro-ED communication than vice versa.

To reinforce the above argument on users' engagement between pro-recovery and pro-ED communication, we inspect users' historical tweets and compute the probability that users post content on topic β after posting content on topic α. The results are shown in Figure 4(b). We see that 17% of users who posted *mental* earlier will post *thinspo* later, while only 10% of users who posted *thinspo* earlier will post *mental* later, which confirms that users are more likely to shift engagement from pro-recovery to pro-ED communication. Also, the probabilities in the last row of Figure 4(b) are relatively low ($P(\beta|\alpha) \leq 0.12$ when $\alpha \neq \beta$), indicating that users previously engaged in posting other content are less likely to engage in posting *mental* health. This explains why the link strengths in the *mental* layer are less correlated with those at other layers (Figure 3(f)). Moreover, the highest value of $P(\beta|\alpha) = 0.26$ occurs when users post *body* content after posting *thinspo*. This explains the significant inter-layer correlations between *body* and *thinspo* (Section 3.2.2), and also confirms that individuals are likely to engage in comparison of body image after viewing thinspo content [66].

Users interested in a specific topic earlier tend to engage in the same topic later. Another notable pattern in Figure 4(a) is the relatively high values of $P(R_{t+1}|R_t)$ when $R_{t+1} = 1, 2, 4, 8, 16$ and $R_t = 0$. Since $R_t = 0$ denotes users having no engagement in the communication system at time t and $R_{t+1} = 1, 2, 4, 8, 16$ denotes users engaging in a single topic at $t+1$, this result suggests that a relatively large number of new users (and those who restore to active state after an inactive period) join the communication system by discussing a single topic. Similarly, the results of $P(R_{t+1}|R_t)$ when $R_{t+1} = 0$ show the statuses of users' engagement in topics before they leave the system. We see that the values of $P(R_{t+1}|R_t)$ when $R_{t+1} = 0$ and $R_t = 1, 2, 4, 8, 16$ are generally high, indicating that users have high dropout rates when discussing only a single topic. Thus, a natural question is whether users have constant interests in the same topics at the beginning and the end of participation in the communication system.

To explore this question, we use the same method described above to measure the beginning-to-end transition probabilities $P(R_e|R_b)$, where R_b and R_e are nodes' activities across layers at the beginning and the end of participation, respectively. To avoid overestimation of $P(R_e|R_b)$ for users who are observed only in one time period[11], we only consider nodes that are active at least in two different temporal networks (i.e., $1 \leq b < e \leq T$). The results are shown in Figure 4(c). As expected, high probabilities of $P(R_e|R_b)$ appear when $R_e = R_b$ (highlighted in a red line) and $R_b = 1, 2, 4, 8, 16$, indicating that users who engage in a single topic earlier

[11]For a user i who is observed once, the initial state of participation $R_{i,b}$ and the final state of participation $R_{i,e}$ are the same, leading to $P(R_{i,e}|R_{i,b}) = 1$.

are more likely to engage in the same topic later. However, this pattern is absent when users engage in more than one topic at an early stage, i.e., $R_e = R_b$ but $R_b \neq 1, 2, 4, 8, 16$. Measuring Cohen's κ between $R_{i,b}$ and $R_{i,e}$ for each user i confirms that the consistency between the beginning and end of participation for users with $R_{i,b} = 1, 2, 4, 8, 16$ ($\kappa = 0.34$) is higher than that with $R_{i,b} \neq 1, 2, 4, 8, 16$ ($\kappa = 0.01$).

The diversity of users' interests decreases over time. We also notice that the probabilities in region II of Figure 4(c) are higher than those in region I. In region II, $R_b > R_e$, meaning that users who engage in a wide range of topics at the beginning of participation tend to focus on a small number of specific topics at the end of participation. To verify this pattern, we measure the diversity of users' interests in tweets over sliding windows. Again, we set sliding windows by a fixed number of tweets rather than a fixed time interval. This is to avoid bias from intermittent activities of users, e.g., as a user becomes less active in posting content, the number of tweets posted in a fixed time interval decreases and the diversity of topics in these tweets will also decrease over time. Given user u posting n distinct tweets on topics $T_1, ..., T_n$ (with repetition), the diversity of posting interests of u in window $i \in [1, n-k+1]$ is measured by the entropy of topics $T_i, ..., T_{i+k}$:

$$H_i(u) = -\sum_{T_j \in \mathbb{T}_{u,i}} P(T_j) \log P(T_j), \tag{3}$$

where $\mathbb{T}_{u,i}$ is the set of distinct topics among $T_i, ..., T_{i+k}$. $P(T_j) = C(T_j)/k$ in which $C(T_j)$ counts the frequency of T_j in $T_i, ..., T_{i+k}$. A larger value of $H_i(u)$ indicates a higher degree of diversity in users' interests. In a similar way, we also measure the diversity of user interests based on tweets that are received from other users.

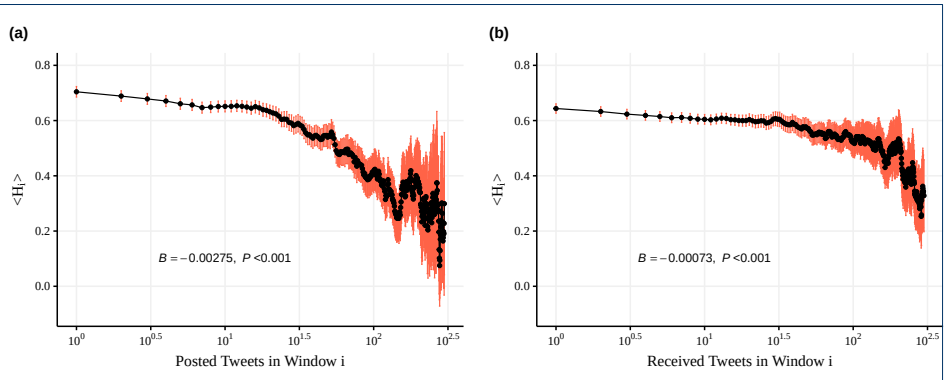

Figure 5 Mean entropy $\langle H_i \rangle$ and 95% CI of topics in (a) posted and (b) received tweets over sliding windows with a size of $k = 10$. CI become wider due to the decreased sample sizes of users having a large number of tweets. Legends report estimated coefficients B and p-values in a linear regression model: $\langle H_i \rangle = Bi + \epsilon$, where estimations are based on windows over $i \in [1, 150]$ due to relatively small mean errors in this range.

Figure 5 shows the mean entropy $\langle H_i \rangle$ and 95% confidence intervals (CI) of topics in tweets posted and received by users over sliding windows, where a window size of $k = 10$ is used. Inactive users who have posted or received less than 20 tweets are excluded to avoid noise. Both plots show that the diversity of user interests has a decreasing trend over time. Results of linear regression models that relate $\langle H_i \rangle$

to a function of i confirm negative correlations between $\langle H_i \rangle$ and i, with $p < 0.001$ in both models. Robustness checks using other window sizes and thresholds for excluding inactive users produce similar results. These findings strongly support the hypothesis that users tend to focus on a small number of specific topics as they engage more online. Moreover, the diversity of interests in received tweets declines more slowly, as compared to that in posted tweets. This hints a time-lag between the two trends, likely because a user may continue to receive information on a topic from other users even when the user loses interests in posting the topic.

3.3.2 Stability of communities

While "dynamics" typically imply changes, another important direction in studying the evolution of social networks is to examine the stability or sustainability of a network over time [64, 67, 68]. The core concept of such research is that members of a social network retain and remember historical ties to former members of the network. This ongoing strengthening of relationships will affect the extent to which the interpersonal network can shape cognition and behavior [64]. As such, we now turn our focus from analyzing changes in topics of conversations to studying stability of a community of users involved in a type of communication. We measure the stability of a community by overlaps of users who engage in the same type of communication over time, i.e., the overlaps of active nodes in the same layer α in different pairs of temporal multilayer networks G_t and $G_{t+\Delta t}$. This can be computed by the Jaccard similarity of nodes that are active in $G_t^{[\alpha]}$ and $G_{t+\Delta t}^{[\alpha]}$ as:

$$J^{[\alpha]}(t, t+\Delta t) = \frac{N_{11}^{[\alpha]}}{N_{01}^{[\alpha]} + N_{11}^{[\alpha]} + N_{10}^{[\alpha]}}, \tag{4}$$

where $\Delta t \in [1, T-1]$ is the time interval between two networks G_t and $G_{t+\Delta t}$. $N_{11}^{[\alpha]}$ is the number of nodes that are active at both $G_t^{[\alpha]}$ and $G_{t+\Delta t}^{[\alpha]}$, $N_{01}^{[\alpha]}$ is the number of nodes that are active at $G_{t+\Delta t}^{[\alpha]}$ but not in $G_t^{[\alpha]}$, and $N_{10}^{[\alpha]}$ is the number of nodes that are active at $G_t^{[\alpha]}$ but not in $G_{t+\Delta t}^{[\alpha]}$. Then, we calculate the mean similarity across intervals Δt, and can obtain the overlaps of actors as a function of Δt:

$$O^{[\alpha]}(\Delta t) = \frac{1}{T-\Delta t} \sum_{t=1}^{T-\Delta t} J^{[\alpha]}(t, t+\Delta t). \tag{5}$$

Figure 6(a) shows results of $O^{[\alpha]}(\Delta t)$ in each layer of temporal networks. As only a small number of observations are available for large values of Δt, we only consider results of $\Delta t \in [1, 40]$ to reduce noise. The key findings are summarized as follows.

Limited numbers of hardcore members engage in harmful communication. As shown in Figure 6(a), users engaged in discussing *mental* health have the largest overlaps over time, indicating strong stability of pro-recovery communities. Moreover, the overlaps of actors in the *mental*, *social* and *fitness* layers tend to be relatively stable (see the highlighted region in Figure 6(a)), suggesting the presence of a large set of hardcore users who have a constantly high level of engagement in exchanging these types of content. In contrast, the overlaps in the *thinspo* and *body*

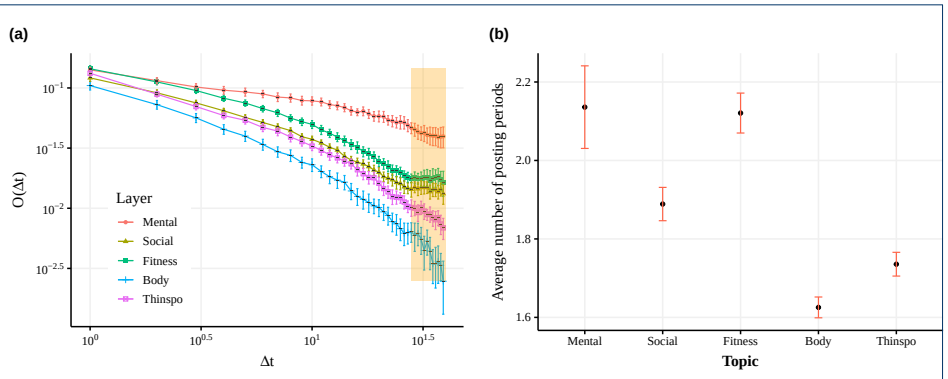

Figure 6 (a) Mean overlaps of active users in temporal networks across interval Δt. Error bars give 95% CI and yellow region highlights the stable stages in the *mental*, *social* and *fitness* layers. (b) Average numbers of periods that users post content on each topic. Error bars show 95% CI.

layers continue to decline as the interval Δt increases. This indicates that members of pro-ED communities have frequent entries into and exits from the system, revealing a high level of fluctuation in these communities.

Individuals engage in harmful communication while organizations engage in healthy communication. To better understand the results in Figure 6(a), we examine users' posting activities in more detail and compute the number of time periods that a user posts a topic. Figure 6(b) shows the average number of posting time periods of users on each topic. We see that users on average share *body* and *thinspo* content in 1.63 and 1.74 time periods respectively, less frequently than sharing other content. This aligns with our results that active nodes in the *body* and *thinspo* layers are highly fluctuating in Figure 6(a). Inspecting the most active users in sharing each topic, we find that active users in sharing *mental* health are often charities and organizations that aim to prevent ED and mental illnesses, such as *@HealingFromBPD*, *@beatED* and *@NEDAstaff*. Similarly, active users in sharing *social* and *fitness* often show a brand-promoting or marketing purpose, e.g., *@WWE* for *social* and *@Reebok* for *fitness*. In contrast, most active users in sharing *thinspo* and *body* content tend to be personal users. Compared to professional organizations and marketing accounts, personal users are less likely to keep continuously active engagement online due to their limited time and attention. Thus, it is not surprising that the *thinspo* and *body* layers have less overlaps of active nodes over time than other layers. This may also explain why the *thinspo* and *body* layers have a more skewed distribution of nodes' out-strengths than other layers in Figures 2(d-e).

4 Discussion

In this study, we have investigated patterns of communication revolved around topics in online ED communities through a large set of conversations among users who self-identified with ED and their friends on Twitter. Applying clustering algorithms to textual content of these Twitter conversations, we find that members of online ED communities are interested in discussing specific topics. By projecting interpersonal interactions in exchanging different topics into a multilayer communication network, we show that different types of communication have distinct network structures and

people play different roles in different types of communication. We further incorporate an additional dimension, namely time, into the multilayer network and reveal dynamic characteristics of multiplex communication in online ED communities.

We show that online ED communities largely focus on discussing mental health, general social activities, fitness, body image and thinspo content, which aligns with previous qualitative studies on the content in these communities [18, 19, 28]. Beyond such content analysis, we further find that different types of content are diffused in different ways, e.g., conversations on private content often take place within small groups and actors in sharing general topics tend to cluster. This multiplex feature of communication cannot be observed through a single-layer network obtained by aggregating all different types of communication, highlighting the importance of considering multiplex patterns in studying human interactions [30, 54].

In line with evidence on other social media platforms [12, 19, 13, 25], we find the presence of two communities with distinct stances on Twitter: (i) a pro-recovery community in which members discuss their health problems and support sufferers to recover from ED and (ii) a pro-ED community in which members often encourage people to lose weight and stay thin. We observe that a small number of users engage in exchanging both pro-ED and pro-recovery content, as indicated by the low value of multiplexity between *mental* and *thinspo* layers. This aligns with prior evidence that social networks of pro-ED communities have small overlaps with those of pro-recovery communities on Flickr [25] and YouTube [11]. Despite these small direct overlaps, our results suggest that both pro-ED and pro-recovery communities have pronounced overlaps with communities of users who engage in exchanging content on *body* image, revealing an indirect connection of social networks between pro-ED and pro-recovery communities. Moreover, we find that users who receive more content on *body* image are likely to post more *thinspo* content and less content on *mental* health. This confirms a conceptual model based on social comparison theory [48] where people who are exposed to images of others' bodies tend to compare their appearance with others, which can lead to a negative view of their own bodies and social pressures to have a thin body that can promote the development of ED.

Our results show that users are more likely to engage in pro-ED communication after pro-recovery communication than vice versa. A possible reason for this is that pro-recovery communities tend to post comments on pro-ED content as an intervention for pro-ED communities [25]. We also find that people tend to focus their communication on narrow, specific topics over time. This can be explained as follows: an individual's time and attention are finite resources, and hence each individual must make a choice about how best to use them given the priority of personal preferences, interests and needs [69]. Prior studies have shown that focusing on a single topic and posting creative or insightful content on the topic can help people to gain influence online [70], and the perception of being valued and respected by others can further motivate people to do so [71]. Moreover, our results suggest that pro-ED communities have a limited number of hardcore actors, with strongly fluctuating membership in the periphery of the communities. A recent study has also shown that individuals of ED communities have short periods of activity in posting content on Twitter [72]. This unstable community structure aligns with views of the pro-ED communities as hidden, secretive groups with frequent migrations [17],

which can make it hard to monitor and track the positions/roles of individuals (e.g., influential cores) in these communities [73]. Such fluctuating characteristics is likely to be reinforced by the banning actions of pro-ED content [17, 28], making pro-ED communities less reachable by health care professionals on social media sites.

Our findings have implications for public health. To prevent ED and minimize the negative impact of pro-ED content online, many social media sites have begun to ban *thinspo* content. Our results show that pro-ED communities may engage in disseminating other content that is related to *thinspo* but has not been banned online, e.g., *body* image. Exposure to such content can potentially reinforce individuals' engagement in pro-ED communication and weaken their engagement in pro-recovery communication, which may be used as alternatives to the *thinspo* content to avoid censorship [74, 15]. Thus, to enhance health outcomes, content-based interventions should account for the relationships of types of content which can be extracted automatically as we did in this study. Another common intervention strategy in public health is network-based intervention which focuses on using social network data to promote organizational well-being [75]. One typical approach in this strategy is identifying community opinion leaders to accelerate behavior change [76, 77]. We show that people can have different roles in different types of interactions and these roles can change over time. Thus, network-based interventions should account for the multiplex and dynamic nature of social interactions for identifying appropriate opinion leaders for a targeted community.

Future work will focus on exploring effective intervention strategies, tailored to the structural and dynamic characteristics of different interactions. For example, one could identify important/central actors in disseminating harmful content, such that removing or isolating these actors can cause maximum damage to the communication of harmful content but with minimum impact on the communication of other content in online ED communities. However, traditional measures of centrality in monolayer networks may not be able to identify these actors, since actors that are not central in each single-layer network might be important for flows of information across layers and have a high centrality score in a multilayer network [78]. It is thus necessary to consider the multilayer structure in measuring actors' roles in diffusive processes. Another direction of future work would be to identify individuals' traits, e.g., personality, emotions, positions in social networks, that can predict the interaction behaviors and dynamics in online ED communities, as well as proper models for characterizing and evaluating dynamic patterns in multilayer networks, so as to enhance our understanding of these communities. In particular, our results indicate that actors' activities and positions across different layers of the multilayer communication network are correlated in different ways. Thus, to build models that predict a user's engagement in sharing a type of information or topic, we should not only consider the user's historical engagement in this topic but also her/his engagement in other topics. Also, we will explore whether our findings are applicable to other online communities based on different social media platforms (e.g., Facebook and Instagram), types of multimedia content (e.g., images and videos), and a wider range of health problems.

In conclusion, our investigation of communication behaviors in online ED communities has uncovered distinct patterns in different types of communication on

Twitter. The rich information in our data allows us to explore the effects of multi-dimensional interactions on the structure and evolution of a large-scale social network, thereby establishing the first empirical basis for modeling multiplex and dynamic communication in online health communities. Our findings can guide public health officials to design advanced online interventions to prevent harmful content from reaching risky individuals and promote organizational well-being.

Availability of data and materials
The datasets generated and analysed during the current study are not publicly available due to Twitter Privacy Policy but are available from the corresponding author on reasonable request.

Ethics approval and consent to participate
The protocol of this study was approved by the Ethics Committee at the University of Southampton. All data analyzed in this study was *public* information on Twitter and available through the Twitter APIs. No personally identifiable information was used in this research.

Funding
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. This work was done when T.W. was with The Alan Turing Institute.

Authors' contributions
T.W., M.B., A.I. and E.M. designed research; T.W. performed data collection and measurements; T.W. and M.B. analyzed results; T.W. wrote the manuscript. All authors reviewed and revised the manuscript.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]ESRC Doctoral Training Centre, University of Southampton, Southampton, UK. [2]The Alan Turing Institute, London, UK. [3]Department of Electronics and Computer Science, University of Southampton, Southampton, UK. [4]Department of Economics, University of Southampton, Southampton, UK. [5]Department of Economics, Università Cà Foscari, Venice, Italy.

References

1. Association, A.P.: Diagnostic and statistical manual of mental disorders (dsm-5®) (2013)
2. Arcelus, J., Mitchell, A.J., Wales, J., Nielsen, S.: Mortality rates in patients with anorexia nervosa and other eating disorders: a meta-analysis of 36 studies. Archives of general psychiatry **68**(7), 724–731 (2011)
3. Wilson, J.L., Peebles, R., Hardy, K.K., Litt, I.F.: Surfing for thinness: A pilot study of pro–eating disorder web site usage in adolescents with eating disorders. Pediatrics **118**(6), 1635–1643 (2006)
4. Overbeke, G.: Pro-anorexia websites: Content, impact, and explanations of popularity. Mind Matters: The Wesleyan Journal of Psychology **3**(1), 49–62 (2008)
5. Mulveen, R., Hepworth, J.: An interpretative phenomenological analysis of participation in a pro-anorexia internet site and its relationship with disordered eating. Journal of health psychology **11**(2), 283–296 (2006)
6. Harper, K., Sperry, S., Thompson, J.K.: Viewership of pro-eating disorder websites: Association with body image and eating disturbances. International Journal of Eating Disorders **41**(1), 92–95 (2008)
7. Mabe, A.G., Forney, K.J., Keel, P.K.: Do you "like" my photo? facebook use maintains eating disorder risk. International Journal of Eating Disorders **47**(5), 516–523 (2014)
8. Rodgers, R.F., Lowy, A.S., Halperin, D.M., Franko, D.L.: A meta-analysis examining the influence of pro-eating disorder websites on body image and eating pathology. European Eating Disorders Review **24**(1), 3–8 (2016)
9. Syed-Abdul, S., Fernandez-Luque, L., Jian, W.-S., Li, Y.-C., Crain, S., Hsu, M.-H., Wang, Y.-C., Khandregzen, D., Chuluunbaatar, E., Nguyen, P.A., *et al.*: Misleading health-related information promoted through video-based social media: anorexia on youtube. Journal of medical Internet research **15**(2), 30 (2013)
10. Lyons, E.J., Mehl, M.R., Pennebaker, J.W.: Pro-anorexics and recovering anorexics differ in their linguistic internet self-presentation. Journal of psychosomatic research **60**(3), 253–256 (2006)
11. Oksanen, A., Garcia, D., Sirola, A., Näsi, M., Kaakinen, M., Keipi, T., Räsänen, P.: Pro-anorexia and anti-pro-anorexia videos on youtube: Sentiment analysis of user responses. Journal of medical Internet research **17**(11) (2015)
12. De Choudhury, M.: Anorexia on tumblr: A characterization study. In: Proceedings of the 5th International Conference on Digital Health 2015, pp. 43–50 (2015). ACM
13. Chancellor, S., Mitra, T., De Choudhury, M.: Recovery amid pro-anorexia: Analysis of recovery in social media. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 2111–2123 (2016). ACM
14. Wang, T., Brede, M., Ianni, A., Mentzakis, E.: Social interactions in online eating disorder communities: A network perspective. PloS one **13**(7), 0200800 (2018)
15. Chancellor, S., Pater, J.A., Clear, T., Gilbert, E., De Choudhury, M.: # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW), pp. 1201–1213 (2016). ACM

16. Chancellor, S., Kalantidis, Y., Pater, J.A., De Choudhury, M., Shamma, D.A.: Multimodal classification of moderated online pro-eating disorder content. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 3213–3226 (2017). ACM
17. Casilli, A.A., Pailler, F., Tubaro, P., *et al.*: Online networks of eating-disorder websites: why censoring pro-ana might be a bad idea. Perspectives in public health **133**(2), 94–95 (2013)
18. Juarascio, A.S., Shoaib, A., Timko, C.A.: Pro-eating disorder communities on social networking sites: a content analysis. Eating disorders **18**(5), 393–407 (2010)
19. Borzekowski, D.L., Schenk, S., Wilson, J.L., Peebles, R.: e-ana and e-mia: A content analysis of pro–eating disorder web sites. American journal of public health **100**(8), 1526–1534 (2010)
20. Teufel, M., Hofer, E., Junne, F., Sauer, H., Zipfel, S., Giel, K.E.: A comparative analysis of anorexia nervosa groups on facebook. Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity **18**(4), 413–420 (2013)
21. Sowles, S.J., McLeary, M., Optican, A., Cahn, E., Krauss, M.J., Fitzsimmons-Craft, E.E., Wilfley, D.E., Cavazos-Rehg, P.A.: A content analysis of an online pro-eating disorder community on reddit. Body image **24**, 137–144 (2018)
22. Wick, M., Harriger, J.: A content analysis of thinspiration images and text posts on tumblr. Body image **24**, 13–16 (2018)
23. Branley, D.B., Covey, J.: Pro-ana versus pro-recovery: A content analytic comparison of social media users' communication about eating disorders on twitter and tumblr. Frontiers in psychology **8**, 1356 (2017)
24. Chancellor, S., Lin, Z., Goodman, E.L., Zerwas, S., De Choudhury, M.: Quantifying and predicting mental illness severity in online pro-eating disorder communities. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW), pp. 1171–1184 (2016). ACM
25. Yom-Tov, E., Fernandez-Luque, L., Weber, I., Crain, S.P.: Pro-anorexia and pro-recovery photo sharing: a tale of two warring tribes. Journal of medical Internet research **14**(6), 151 (2012)
26. Arseniev-Koehler, A., Lee, H., McCormick, T., Moreno, M.A.: # proana: Pro-eating disorder socialization on twitter. Journal of Adolescent Health **58**(6), 659–664 (2016)
27. Moessner, M., Feldhege, J., Wolf, M., Bauer, S.: Analyzing big data in social media: Text and network analyses of an eating disorder forum. International Journal of Eating Disorders (2018)
28. Tiggemann, M., Churches, O., Mitchell, L., Brown, Z.: Tweeting weight loss: A comparison of# thinspiration and# fitspiration communities on twitter. Body image **25**, 133–138 (2018)
29. Newman, M.: Networks: an introduction (2010)
30. Nicosia, V., Latora, V.: Measuring and modeling correlations in multiplex networks. Physical Review E **92**(3), 032805 (2015)
31. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer networks. Journal of complex networks **2**(3), 203–271 (2014)
32. Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C.I., Gómez-Gardenes, J., Romance, M., Sendina-Nadal, I., Wang, Z., Zanin, M.: The structure and dynamics of multilayer networks. Physics Reports **544**(1), 1–122 (2014)
33. Wang, T., Brede, M., Ianni, A., Mentzakis, E.: Detecting and characterizing eating-disorder communities on social media. In: Proceedings of the Tenth International Conference on Web Search and Data Mining (WSDM) 2017, pp. 91–100 (2017). ACM
34. Weng, L., Menczer, F.: Topicality and impact in social media: diverse messages, focused messengers. PloS one **10**(2), 0118410 (2015)
35. Alvarez-Melis, D., Saveski, M.: Topic modeling in twitter: Aggregating tweets by conversations. ICWSM **2016**, 519–522 (2016)
36. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
37. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
38. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp. 1188–1196 (2014)
39. Wang, Y., Liu, J., Qu, J., Huang, Y., Chen, J., Feng, X.: Hashtag graph based topic model for tweet mining. In: Data Mining (ICDM), 2014 IEEE International Conference On, pp. 1025–1030 (2014). IEEE
40. Steinskog, A.O., Therkelsen, J.F., Gambäck, B.: Twitter topic modeling by tweet aggregation. In: Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden, pp. 77–86 (2017). Linköping University Electronic Press
41. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), 10008 (2008)
42. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences **105**(4), 1118–1123 (2008)
43. Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. Physical review E **69**(2), 026113 (2004)
44. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short text classification in twitter to improve information filtering. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 841–842 (2010). ACM
45. Clauset, A., Shalizi, C.R., Newman, M.E.: Power-law distributions in empirical data. SIAM review **51**(4), 661–703 (2009)
46. Newman, M.E.: Mixing patterns in networks. Physical Review E **67**(2), 026126 (2003)
47. Hargreaves, D., Tiggemann, M.: Longer-term implications of responsiveness to 'thin-ideal'television: support for a cumulative hypothesis of body image disturbance? European Eating Disorders Review: The Professional Journal of the Eating Disorders Association **11**(6), 465–477 (2003)
48. Yu, U.-J.: Deconstructing college students' perceptions of thin-idealized versus nonidealized media images on

body dissatisfaction and advertising effectiveness. Clothing and Textiles Research Journal **32**(3), 153–169 (2014)
49. Becker, A.E., Hadley Arrindell, A., Perloe, A., Fay, K., Striegel-Moore, R.H.: A qualitative study of perceived social barriers to care for eating disorders: perspectives from ethnically diverse health care consumers. International Journal of Eating Disorders **43**(7), 633–647 (2010)
50. Swanson, S.A., Crow, S.J., Le Grange, D., Swendsen, J., Merikangas, K.R.: Prevalence and correlates of eating disorders in adolescents: Results from the national comorbidity survey replication adolescent supplement. Archives of general psychiatry **68**(7), 714–723 (2011)
51. Fehr, E., Gächter, S.: Fairness and retaliation: The economics of reciprocity. Journal of economic perspectives **14**(3), 159–181 (2000)
52. Lim, K.H., Datta, A.: A topological approach for detecting twitter communities with common interests, 23–43 (2013)
53. Hu, H.-B., Wang, X.-F.: Disassortative mixing in online social networks. EPL (Europhysics Letters) **86**(1), 18003 (2009)
54. Szell, M., Lambiotte, R., Thurner, S.: Multirelational organization of large-scale social networks in an online world. Proceedings of the National Academy of Sciences **107**(31), 13636–13641 (2010)
55. Lewis, K., Gonzalez, M., Kaufman, J.: Social selection and peer influence in an online social network. Proceedings of the National Academy of Sciences **109**(1), 68–72 (2012)
56. Connor, N., Barberán, A., Clauset, A.: Using null models to infer microbial co-occurrence networks. PloS one **12**(5), 0176751 (2017)
57. Gotelli, N.J., Ulrich, W.: Statistical challenges in null model analysis. Oikos **121**(2), 171–180 (2012)
58. Paul, S., Chen, Y.: Null models and modularity based community detection in multi-layer networks. arXiv preprint arXiv:1608.00623 (2016)
59. Croft, D.P., Madden, J.R., Franks, D.W., James, R.: Hypothesis testing in animal social networks. Trends in Ecology & Evolution **26**(10), 502–507 (2011)
60. Opsahl, T., Colizza, V., Panzarasa, P., Ramasco, J.J.: Prominence and control: the weighted rich-club effect. Physical review letters **101**(16), 168702 (2008)
61. Viswanath, B., Mislove, A., Cha, M., Gummadi, K.P.: On the evolution of user interaction in facebook. In: Proceedings of the 2nd ACM Workshop on Online Social Networks, pp. 37–42 (2009). ACM
62. Freedman, D., Diaconis, P.: On the histogram as a density estimator: L 2 theory. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **57**(4), 453–476 (1981)
63. Kirman, B., Lawson, S.: Hardcore classification: Identifying play styles in social games using network analysis. In: International Conference on Entertainment Computing, pp. 246–251 (2009). Springer
64. Kilduff, M., Tsai, W., Hanke, R.: A paradigm too far? a dynamic stability reconsideration of the social network research program. Academy of Management Review **31**(4), 1031–1048 (2006)
65. Albert, R., Jeong, H., Barabási, A.-L.: Error and attack tolerance of complex networks. nature **406**(6794), 378 (2000)
66. Bardone-Cone, A.M., Cass, K.M.: What does viewing a pro-anorexia website do? an experimental examination of website exposure and moderating effects. International Journal of Eating Disorders **40**(6), 537–548 (2007)
67. Becke, G.: Mindful change in times of permanent reorganization. Berlin, Germany (2014)
68. Onnela, J.-P., O'Malley, A.J., Keating, N.L., Landon, B.E.: Comparison of physician networks constructed from thresholded ties versus shared clinical episodes. Applied Network Science **3**(1), 28 (2018)
69. Gonçalves, B., Perra, N., Vespignani, A.: Modeling users' activity on twitter networks: Validation of dunbar's number. PloS one **6**(8), 22656 (2011)
70. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K.: Measuring user influence in twitter: The million follower fallacy. ICWSM **10**(10-17), 30 (2010)
71. Ryan, R.M., Deci, E.L.: Intrinsic and extrinsic motivations: Classic definitions and new directions. Contemporary educational psychology **25**(1), 54–67 (2000)
72. Wang, T., Mentzakis, E., Brede, M., Ianni, A.: Estimating determinants of attrition in eating disorder communities on twitter: an instrumental variables approach. Journal of medical Internet research (forthcoming)
73. Sekara, V., Stopczynski, A., Lehmann, S.: Fundamental structures of dynamic social networks. Proceedings of the national academy of sciences **113**(36), 9977–9982 (2016)
74. Boepple, L., Ata, R.N., Rum, R., Thompson, J.K.: Strong is the new skinny: A content analysis of fitspiration websites. Body image **17**, 132–135 (2016)
75. Valente, T.W.: Network interventions. Science **337**(6090), 49–53 (2012)
76. Valente, T.W., Pumpuang, P.: Identifying opinion leaders to promote behavior change. Health Education & Behavior **34**(6), 881–896 (2007)
77. Laranjo, L., Arguel, A., Neves, A.L., Gallagher, A.M., Kaplan, R., Mortimer, N., Mendes, G.A., Lau, A.Y.: The influence of social networking sites on health behavior change: a systematic review and meta-analysis. Journal of the American Medical Informatics Association **22**(1), 243–256 (2014)
78. De Domenico, M., Solé-Ribalta, A., Omodei, E., Gómez, S., Arenas, A.: Ranking in interconnected multilayer networks reveals versatile nodes. Nature communications **6**, 6868 (2015)