Characterizing Audience Engagement and Assessing Its Impact on Social Media Disclosures of Mental Illnesses

Sindhu Kiranmai Ernala,[†] **Tristan Labetoulle,**[†] **Fred Bane,**[†] **Michael L. Birnbaum,**[§]
Asra F. Rizvi,[§]**John M. Kane**[§] **Munmun De Choudhury**[†]
Georgia Institute of Technology[†], Zucker Hillside Hospital, Psychiatry Research[§]
{sernala3, tristan-labetoulle, fwbane, munmund}@gatech.edu, {Mbirnbaum, ARizvi3, JKane2}@northwell.edu

Abstract

Self-disclosures of mental illnesses have been identified to yield coping and therapeutic benefits. An important construct in the self-disclosure process is the *audience* with whom the individual interacts and shares their experiences. Mental illness self-disclosures are increasingly happening online. However, unlike online support communities where the audience comprises sympathetic peers with similar experiences, what the discloser gains from an 'invisible' audience on a general purpose, public social media platform is less understood. Focusing on a highly stigmatized mental illness, schizophrenia, this paper provides the first investigation characterizing the audience of disclosures of this condition on Twitter and how the audience's engagement impacts future disclosures. Our results are based on a rich year-long temporal analysis of the data of nearly 400 disclosers and their nearly 400 thousand audiences. First, characterizing and modeling the audience engagement temporally, we find evidence of reciprocity in the disclosure process between the discloser and their audience. Then, situating our work in the Social Penetration Theory and operationalizing the disclosure process via a measure of intimacy, an auto-regressive time series model indicates that the patterns of audience engagement and content can forecast changes in the intimacy of disclosures. We discuss the implications for building socially engaging, supportive online spaces for stigmatized mental illness disclosures.

Introduction

In regard to experiences around mental illness, people are increasingly appropriating social media sites as spaces for self-expression, spreading awareness, breaking inhibitions and stigma, finding solidarity, and building communities. Self-disclosure, the "process of making the self known to others", is known to support this new and less expected use of social media platforms (Archer 1980). It is a precursor to online expression of identity, emotions, behaviors, and experiences. Particularly in individuals experiencing stigmatized conditions, like mental health challenges, self-disclosure is a frequent coping mechanism (Joinson 2001).

A variety of motivations and intents underlie people's decisions to self-disclose. One established reason is that people need 'sympathetic others', as Goffman (2009) posited: those who share the same social stigma, have had similar experiences, and those who "share with him the feeling that he is human and 'essentially' normal in spite of appearances and in spite of his own self doubt". The sympathetic others in an online social platform can, however, be varied. On platforms like Reddit, where there are dedicated support communities for mental health challenges, the others are often experts and peers with similar experiences. On social networking sites like Facebook, the others are likely social ties embedded in the offline context. Yet, recent studies have uncovered "broadcasting self-disclosures", a phenomenon that refers to sharing personal, sensitive information in a public social media context such as Twitter, to somewhat nebulous, less defined others (Bazarova and Choi 2014). Unlike online support communities, even if the disclosing individual has a mental conceptualization of their audience (Gruzd, Wellman, and Takhteyev 2011), they are likely to be 'invisible' and large, consisting not necessarily of experts or of peers undergoing similar experiences, but perhaps a wide variety of people with different backgrounds, interests, identity profiles, and purposes of social media use. Unlike social networking sites, the audience might also largely comprise weak ties (Kwak et al. 2010)—those that the individual might not know or ever encounter offline.

Initially, disclosure of sensitive, stigmatized mental illnesses to such an invisible or even imagined audience can seem puzzling. However, the prevalence of the phenomenon, as shown in prior work (De Choudhury et al. 2017; Ernala et al. 2017), suggests that the discloser might gain certain social benefits from such an audience. How can we better understand these audience, the ways they engage with stigmatized content, and the manner they impact the disclosure process on an otherwise general purpose, social media platform? Addressing these questions will help us understand the social benefits a discloser derives over time by continuing to disclose to this audience.

Building on this motivation, we present a quantitative methodology to understand audience and their engagement to stigmatized self-disclosures on Twitter. We choose the specific case of self-disclosures of schizophrenia, as it is one of the most stigmatized mental health conditions and sufferers are known to face negative stereotyping and attitudes, discriminatory and offensive behavior, and societal rejection (Dickerson et al. 2002). Specifically, we focus on the

following two research questions:
RQ1: *What are the patterns in which social media audience are engaging with the self-disclosing individuals?*
RQ2: *How does the audience engagement impact the future disclosure process? In other words, is audience engagement predictive to future intimacy of disclosures?*

Towards these research questions, employing machine learning techniques on a clinically validated dataset, we obtain a list of individuals (disclosers) who have publicly shared about their diagnosis of schizophrenia on Twitter. We define the audience of these disclosures as individuals who have interacted with the disclosers' content using the Twitter functionalities of retweets, favorites or mentions. Then, we characterize the temporal variation in audience engagement and study its alignment with respect to what the disclosers present about themselves in their postings (RQ1). Then, drawing from the Social Penetration Theory (Altman and Taylor 1973), we model the disclosers' behavior by operationalizing the notion of intimacy of disclosures. With this measure of intimacy and time series forecasting techniques, we assess if the engagement received from the audience is predictive of future intimacy of the disclosures made by the disclosers (RQ2).

Based on our characterization of audience engagement, we find evidence of temporal and topical reciprocity in the interactions between the disclosers and their audience, as would be anticipated in an online support community. In relation to the disclosers' data, the audience engagement includes major themes such as mental health resources, stigma, and emotional support. Our results from the time series forecasting model show that attributes of the audience engagement like number of mentions received, themes related to emotional support and personal, private life strongly predict patterns in future disclosure behavior. Through these findings, our work sheds new light into the role of the audience in public social media platforms toward supporting self-disclosures of stigmatized conditions and experiences.

Background & Related Work

Theoretical Framework: Social Penetration Theory One of the aims of this paper has been to study the phenomenon of broadcasting self-disclosures as an interpersonal relationship between the disclosers and their audience. The Social Penetration Theory provides a relevant theoretical framework. Introduced by Altman and Taylor (1973), the theory proposes self-disclosure as a necessary precursor and a critical component to relationship development between individuals. It describes self-disclosure as the process of sharing different levels of information, varying from superficial to intimate, about oneself to others. These varying levels of disclosure (also termed as degree of social penetration) are conceptualized in terms of two dimensions: Breadth and Depth of disclosure. Of interest here is the depth, that refers to degree of intimacy in disclosures. It relates to the extent to which one comfortably opens up about a particular aspect of their personal, private life that would otherwise not be revealed publicly.

Situating the social penetration theory in the context of broadcasting self-disclosures, as is the case in this paper, would mean understanding the process of relationship development between the disclosers and audience with respect to the varying levels of self disclosure. This necessitates examining the reciprocal behaviors between the disclosers and audience which motivates our discussion on RQ1. Further, contextualizing the dimensions of breadth and depth of disclosures to social media would require examining the topical content of these disclosures. Since we focus our attention on disclosures specific to one topic i.e. mental illness (schizophrenia in particular), we adopt the component of depth (or intimacy) to model disclosure in RQ2. Thus we refer to the framework of social penetration theory to situate our approach and inform our analysis.

Self-Disclosure on Social Media A rich body of work has studied self-disclosure in the context of computer mediated communication (Joinson 2001). The findings from this literature relating disclosure to trust and group identity (Joinson and Paine 2007), reducing uncertainty and stigma (Cozby 1973; Derlaga and Berg 2013) form the building blocks of recent work on self-disclosure on social media.

Supported by the affordances of anonymity (De Choudhury and De 2014; Andalibi et al. 2016) and social connectedness (Bazarova and Choi 2014), social media has been widely adopted as a space for self-disclosures. Specifically, in stigmatized conditions related to identity, health and wellbeing researchers have focused attention on characterizing and modeling self-disclosure behaviors (Haimson and Hayes 2017, Yang et al. 2017) and studying platform specific differences (De Choudhury et al. 2017). Several qualitative studies have augmented this line of research by examining the goals, motivations and challenges in online self-disclosures (Andalibi et al. 2017).

We note that work thus far has largely been around platforms where the others in the context of self-disclosures are sympathetic others, as Goffman (2009) posited it. However, the nature and impact of engaging with the audience of self-disclosures on public social media platforms is understudied. Our work aims to fill this gap. We focus on one of the most stigmatized conditions, schizophrenia, as the psychopathology of the condition indicates that the sufferers are particularly known to benefit therapeutically from intimate self-disclosures (Shimkunas 1972).

Online Social Capital and Support There has been a relevant line of research concerning online social capital and social support in the context of self-disclosures and wellbeing (Burke et al. 2010). Social capital allows an individual to draw on resources from other members in their social network through bonding and bridging (Coleman 1988). While online social networks have been established to support building and maintaining both kinds of social capital (Ellison et al. 2007), scholars also refer to a related concept "social support", especially in the context of self-disclosure theories and studies of stigma. A large body of work reveals the support benefits people derive from their interpersonal relationships and social networks in relation to improved health and psychological well-being, self esteem, satisfaction with life, and reciprocity (Helliwell and Putnam 2004).

Specific to our focus on stigmatized experiences around mental health, both qualitative and quantitative studies have identified social capital and social support as necessary components in self-disclosure goals and outcomes (De Choudhury and De 2014; Zhang 2017). Nevertheless, gaps still exist in our understanding of how the expectations of social support and the benefits with respect to social capital translate when the audience of self-disclosures are invisible, public, or comprise largely of weak ties. Moreover, the role that the audience of stigmatized disclosures, through support provisioning and social feedback mechanisms, plays in encouraging (or constraining) future disclosure processes, is yet to be empirically investigated. We seek to extend prior work by providing a robust data-driven study of the audience of schizophrenia disclosures on Twitter.

Data

Twitter Data on Schizophrenia Disclosures As a first step of our data collection, we obtained access to a clinician validated Twitter dataset of self-disclosures of schizophrenia from Ernala et al. (2017). This dataset included the public Twitter timelines (1,940,921 tweets) of 146 users who had self-disclosed regarding their diagnosis of schizophrenia for the first time in the year 2014. Example key-phrases that were used as seed search queries to identify these disclosures included first-person reports of schizophrenia experiences and diagnoses like “Diagnosed me with schizophrenia/ psychosis”[1]. Further, noisy data in the form of disingenuous, inappropriate statements and jokes were filtered out via manual examination and consultation with two psychiatrists who see schizophrenia patients.

Next, we adopt the supervised machine learning methodology developed by Birnbaum et al. (2017). The classifier employs clinical appraisals as ground truth and linguistic (n-grams) and psycholinguistic tokens (from the LIWC dictionary (Pennebaker, Francis, and Booth 2001)) in tweets as features to successfully recognize (with 88% area-under-curve and 80% precision) genuine self-disclosures (of schizophrenia) gathered from Twitter. Obtaining access to the clinical appraisals and adapting the technique in this classifier on a new sample of 600 Twitter users, we were able to machine label 433 of them to have genuinely disclosed their illness. Our expanded dataset (together with original 146 users) used in this paper therefore consists of 579 Twitter users who engaged in self-disclosures of schizophrenia.

Recall that for the purpose of this paper, we aim to investigate the patterns of audience engagement around the Twitter content of these users and how it impacts their disclosure behavior in the future. Therefore without loss of generality, for our analysis we focus on an year long period of Twitter activity succeeding the 579 users' self-disclosures. We found that 395 out of these 579 had an entire year's worth of Twitter data. Over the year-long period, we found these 395 users to have shared 1,491,623 tweets with an average of 3776.26 tweets per user and 17.48 tweets per day per user. We report a summary of these descriptive statistics in Table 1.

Number of disclosers	395
Total tweets of disclosers	1,491,623
Mean tweets per discloser	3776.26
Mean tweets per day per discloser	17.48
Median tweets per discloser	1338
Distinct number of retweets audience	124,630
Distinct number of favorites audience	169,041
Distinct number of mentions audience	80,090
Total number of audience	373,761
Mean distinct audience per discloser	1218.4

Table 1: Descriptive statistics of disclosers & audience data.

Definitions Next, we introduce a few definitions centering around self-disclosures of schizophrenia on Twitter and audience engagement around it. First, a ‘*discloser*’ is an individual who has self disclosed (revealed) their diagnosis of schizophrenia by publicly posting on Twitter, on day d, the day of disclosure. The ‘*audience*’ of these disclosures is the set of Twitter users who have interacted with the discloser's Twitter posts viz-a-viz the platform's functionalities—retweets, favorites or ‘likes’, mentions over the period of one year after day d. We operationalize ‘*audience engagement*’ as any instance of such an interaction between a member of the audience and the discloser. Retweets, mentions, favorites or ‘likes’ constitute the various markers of audience engagement.

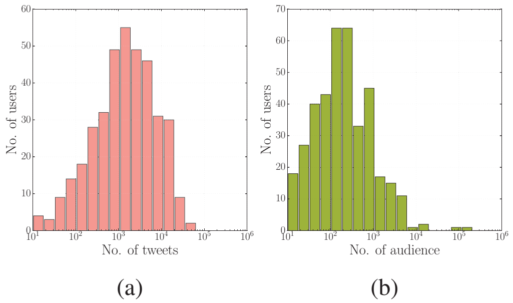
Audience and Audience Engagement Data We proceed with data collection of audience engagement, by obtaining data on the engagement markers — retweets, favorites and mentions surrounding the disclosers' data.

Retweets Data. We collected this audience engagement dataset by identifying the Twitter users who have interacted with the disclosers by retweeting their content during the one year after disclosure. First, for each tweet from the disclosers during this period of analysis, we obtained the number of retweets received by that tweet. Then, we used the official Twitter API to obtain the list of individuals who have retweeted it. Applying this method across all 395 disclosers we obtain 124,630 distinct Twitter users (retweets audience) who retweeted the disclosers' content 2,895,118 times.

Favorites Data. We identified Twitter users who interacted with the disclosers' data through favorites (liking) during the one year after disclosure. For each tweet posted by the disclosers during the period, we first obtained the number of favorites received by the tweet. Then, we parsed the JSON object of the HTML popup that shows users who have favorited a tweet. Applying this across all disclosers, we obtained a set of 169,041 Twitter users (favorites audience) who favorited the disclosers' content 4,592,890 times.

Mentions Data. Here, we collect data on those Twitter users who have interacted with the disclosers using the mentions (or @-replies) functionality. On Twitter, when an individual replies to another (say with username B), the tweet is automatically appended with the ‘@B’ string. We used this stylistic convention of tweets to compile a list of search queries by appending an ‘@’ symbol before the username of each of our disclosers. This operation provided us with

[1] Additional key-phrases included here https://tinyurl.com/yc6h38wl



(a)

(b)

Figure 1: (a) Distribution of #disclosers over #tweets. (b) Distribution of #disclosers over #distinct audience.

all tweets that were incoming mentions to the disclosers including Twitter users who mentioned them and the textual content of the mention tweets. This dataset finally consisted of 80,090 distinct users (mentions audience) who mentioned the disclosers in their 348,456 mention tweets.

Audience Data. To compile the final set of audience, we collated the list of users in the three datasets above— 124,630 retweets audience, 169,041 favorites audience and 80,090 mentions audience, and extracted overall 373,761 users. At a discloser level, on average, the audience size was 1218.4. The distribution of audience (size) and its descriptive statistics are provided in Figure 1 and Table 1.

RQ1: Characterizing Audience Engagement

Methods Per RQ1, we propose methods to characterize audience engagement around disclosers' Twitter data based on two attributes: the content of engagement and its markers.

Thematic Representation of Disclosers' Data. First, we develop a thematic representation of the data shared by the disclosers over the year-long period following their day of disclosure d. This representation is used to examine the dynamic interaction between the disclosers and their audience in terms of content sharing. We begin by employing topic modeling on Twitter timelines of our 395 disclosers. After preprocessing the tweets to remove URLs and stopwords, we run Latent Dirichlet Allocation using MALLET: MAchine Learning for LanguagE Toolkit. We perform hyperparameter optimization over the sampling iterations to extract 30 topics. Using the topic model, we compute the topic distribution via posterior probabilities for each tweet.

Next, to identify semantically interpretable, broader themes from the 30 topics, we employed qualitative labeling. Two human raters who were social media and mental health experts performed semi-open coding on the extracted topics collaboratively. Drawing from their experience studying self-disclosures of schizophrenia, the raters built a set of topical descriptors for each topic by analyzing the top contributing keywords per topic. Then, they combined the LDA topics into semantically interpretable, broader themes and also labeled whether or not each theme was related to the diagnosis and experiences of schizophrenia. Finally, to inspect the disclosers' data over time, we used the theme annotations and computed z-scores of the average probability of each theme per day across all disclosers. Since z-scores reveal relative differences in the values of a distribution, it qualifies as a suitable metric to study variation over time.

Characterizing Engagement Content. Using the same topic modeling and qualitative theme annotation approach as above, we characterized the engagement content (i.e. dataset of the mention tweets), corresponding to each discloser. First, we built an LDA model to obtain 30 topics from the linguistic content of these mention tweets. Then, we computed the topic distribution for each tweet in the mentions dataset. Next, we employed a qualitative labeling method to identify interpretable, broader themes from these LDA generated 30 topics in the engagement content. The same two raters as above were employed to analyze the top keywords per topic and come up with topical descriptors for each topic, including annotations on whether or not each theme was related to the disclosure and experiences of schizophrenia. We again used the z-scores of the average probability of each theme per day across all disclosers to identify theme-specific variation in the engagement content over time.

Characterizing Engagement Markers. To characterize the engagement markers, we use the dataset of retweets, favorites and mentions received by each discloser per day during the one year period following day of disclosure. For each day d, ranging from $d = 0$ to $d = 365$, we find the average number of retweets, favorites and mentions received by all the disclosers and transform the average values into z-scores. This transformation gives us the variation in engagement markers received by the disclosers as a function of time and allows relative comparison. We obtain three time series, one for retweets, favorites and mentions from this step.

Discovering Patterns of Audience Engagement. To study the variations in engagement indicators (markers and content) with respect to that in disclosers' data, we make the following categorization. Based on the thematic annotations over disclosers' data and their corresponding engagement content, we categorize the theme labels into: themes related to the diagnosis and experiences of schizophrenia, and those unrelated. For both theme categories, we adopt time series comparison techniques (e.g., the cross correlation measure) to understand how the z-score distributions of the engagement markers and the themes of the engagement content vary with the disclosers' theme distributions over time.

Results

Comparing Disclosers' Themes and Audience's Themes. We present results from the thematic annotations on audience's engagement content and discuss them in the context of the themes derived from disclosers' data (Ref. Table 2). This juxtaposition of themes helps us understand the audience response with respect to what the disclosers' are sharing on Twitter. First, among the engagement content themes that relate to experiences of schizophrenia, we begin by considering the theme "Mental Health Support/Stigma" (MHSS) that also surfaces in the disclosers' data. For instance, we notice the usage of words such as 'hcsmca', 'pndhour', 'awareness', 'issue' referring to online communities dedicated to exchanges around health care, mental illness and spreading awareness. The same theme

	Disclosers' Data		Engagement Content (Audience)	
Theme	n	*Top Words*	n	*Top Words*
MHSS	1	mental health depression illness pndhour anxiety mentalhealth issues submitted stigma today schizophrenia meds disorder cancer hospital support pain	2	hcsmca social support public info issue important system kids personal care health experience pndhour mental health support depression meds pain issues awareness illness anxiety story loss
Appearance	2	hair wear shirt white red clothes dress blue pants shoes fashion color back eyes head hand face softly arms neck lips smile kiss hair	2	hair wear red black dress nice clothes shirt blue body pants shoes back head eyes hand face neck smiles mouth softly cheek hugs arms lips butt
Functioning	4	love lot make time care talk anymore friends people dont life women social men thing good human work change money kids company job tax day sleep night week	3	good life hard work watch times thing love lot live make money pay food free people lot job low rich high business woman married relationship single engaged miracle divorced
Emotions	2	happy good hope today great beautiful amazing lovely year sweet make good feel bad people time life lot lol thought pretty today weird	4	care anymore worry hurt ill trust mad reason treat fuck person good bad feel life makes wrong find nice love wtf happy beautiful hope love talk fake
Sexuality	2	girl man guy hes shes sex cute love youre boy years baby dad friend mom gay woman child	1	lol girl shit girls youre fuck man ass hes funny fucking cool pretty shes guy weird guys cute
Symptoms	4	r/paranormal ufo r/creepy shit ass fuck bitch house back door night angels gods soul hell saved world	0	–
Temporal References, Planning	1	time day sleep work night today tomorrow back school home bed week hours days ill morning tonight ago gonna	3	night sleep time tomorrow week work today late hours home days morning year ago time long past day sunshine fab weekend friday
Communication	0	–	2	back text reply message lol tweet word tweets didnt haha talking forgot answer thought question funny doctor isnt meant english wrong swear correct
Others	14	cats dogg standwithrand tedcruz video football war government israel campaign police	13	law power gamergate superbowl stories club bro party school parents

Table 2: Theme descriptions obtained via topic modeling and qualitative annotations on disclosers' and audience's engagement data. n stands for number of topics per theme.

also includes overlapping words like 'depression', 'anxiety', 'meds', 'mental health', 'pain', relating to the stigma and challenges around experiences of schizophrenia. This shows that the audience, in response to the schizophrenia content of the disclosers share their experiences and resources related to mental health care, providing solidarity.

Next, we consider another common schizophrenia related theme, 'Functioning'. We observe overlapping keywords, such as 'people', 'life', 'good', 'work', 'money', 'job', 'love', 'sleep'. Relatedly, we also find the theme 'Appearance' (words: 'hair', 'wear', 'red', 'clothes', 'arms', 'softly') that surfaces in the tweets of both the disclosers and their audience. Taken together, these themes relate to the everyday experiences capturing behaviors around the social, emotional, physical, and cognitive aspects of life. Their cooccurrence as themes reflects the utility of engagement content as a mechanism to converse about everyday aspects of life, communicate, plan, and exchange thoughts and ideas.

Next, we consider the theme 'Emotions' that appears in the engagement content with words like 'love', 'happy', 'good', 'hope', 'lovely', 'miss', 'sweet', 'beautiful'. While this theme is also present in disclosers' data, we note a higher prevalence of emotional content in the engagement content than that of the disclosers based on the number of topics contributing towards the theme. This particular imbalanced overlap characterizes the emotional support provisioning nature of the engagement that the disclosers gather from their audiences; a form of support found in the literature to be key to improved mental health state and in supporting therapeutic outcomes from disclosures of stigmatized conditions (De Choudhury et al. 2014).

Lastly, we find an overlap between the audiences and the disclosers in the theme 'Sexuality' containing terms such as 'girl', 'man', 'guy', 'he's', 'she's', 'cute', 'fuck', 'sex'. This indicates a tendency of the disclosers and in response their audiences to "open up" about deeply personal aspects of their private life that are usually not revealed publicly.

Nevertheless, despite the thematic reciprocity noted above, we note a sharp distinction between the tweets of the disclosers and audience—shown by the theme 'Symptoms'. In the case of the disclosers, this theme ('r/paranormal', 'r/creepy', 'ufo') reveals a predominant occurrence of words that have symptomatic relevance to schizophrenia. We do not observe such patterns in the themes extracted from the audience's engagement content. This indicates that, although the disclosers are sharing their first person experiences of the illness, the audiences do not respond with similar personal accounts. This brings to light the distinction in broadcasting disclosures on platforms like Twitter, where, unlike support communities, the audience need not necessarily consist of peers undergoing similar experiences.

By juxtaposing the thematic annotations from the disclosers and their audiences, we find evidence of reciprocal conversations around shared themes related to experiences of schizophrenia. We situate this discussion in the social penetration theory that gives a distinctive emphasis to self-disclosing behaviors being maintained by the "gradual overlapping and exploration of their mutual selves by parties to a relationship" (Sprecher et al. 2013).

Patterns of Changes in the Engagement Content. Here, we are interested in the question—how do the above (schizophrenia related and other) themes from the disclosers and the audience co-vary over time? Inspecting Figure 2(a-d), we observe that there is a close temporal alignment between the disclosers' and the audiences' themes relating to schizophrenia experiences. Specifically, by analyzing the cross correlation between the two, we find that the highest correlation of 0.125 between the two time series occurs at a negative lag of 4. This positive correlation at a negative lag provides indications of reciprocity in the disclosure process—as the disclosers increasingly talk about their schizophrenia experiences at time $t - 4$ (in days), it correlates with the audience talking about similar themes related to these experiences at t. Reciprocity has been identified as a major norm in self-disclosure research (Jiang, Bazarova, and Hancock 2013). In contrast, we find that as the disclosers increasingly talk about their experiences, the audience begin limiting posts on other unrelated topics in the future (maximum correlation of -0.125 at a negative lag of 4).

Patterns of Changes in the Engagement Markers. We ask the question—how does the audience, with the help of various platform functionalities, respond to disclosers, and how do different engagement markers co-vary with disclosers' themes. Figure 3a shows the z-score distribution of these markers over time. We observe two findings. First, beginning at the day of disclosure, there is a peak in mentions indicating an increase in incoming engagement from the audience. However, there is lowered audience engagement during this early period through retweets and favorites. This could indicate that the audience find the disclosers' content out of place and take time to modulate their engagement around it. Second, there is a very close alignment between the temporal variation in retweets and favorites received from the audience. This may be attributable to the similar functionality between both actions i.e. they both indicate some form of acknowledgement or endorsement, and have a lower barrier for content production (at the click of a button), compared to mentions which have a higher barrier to content production, requiring consciously drafted replies.

Next, in Figure 2(e-j), we present an analysis of the temporal variation in the three engagement markers in relation to the disclosers' themes—both the schizophrenia related ones as well as the rest. Upon visual inspection, we notice that the alignment between the daily measurements of engagement markers is higher with disclosers' data related to schizophrenia experiences as compared to other unrelated content. For the time series representing thematic variation in schizophrenia related experiences, the maximum correlation with retweets and favorites is -0.09, -0.08 observed at cross correlation lags of 5, 5 respectively. The negative correlation at a positive lag denotes that as the disclosers increasingly talk about their condition and experiences, it correlates to receiving fewer retweets and favorites in the days following. This is likely explained by the perception that the actions of retweet or favorite signal information sharing intentions and do not convey an appropriate response to stigmatized disclosures. On the other hand, we observe a stronger alignment between the disclosure content related to experiences of schizophrenia and the mentions received. The correlation of disclosure related content with mentions is the strongest with a lag 0 with a positive value of 0.17. This shows that as the disclosers increasingly talk about their experiences, it correlates to receiving more mentions (on the same day). However, in the case of unrelated themes, we observe a delayed response via mentions from the audience (maximum correlation of 0.14 at lag -7). Summarily, our findings from RQ1 suggest reciprocity, temporally in the number of engagement markers received and topically, in the themes received viz-a-viz the audience engagement content.

RQ2: How Audience Engagement Predicts Future Intimacy of Disclosures

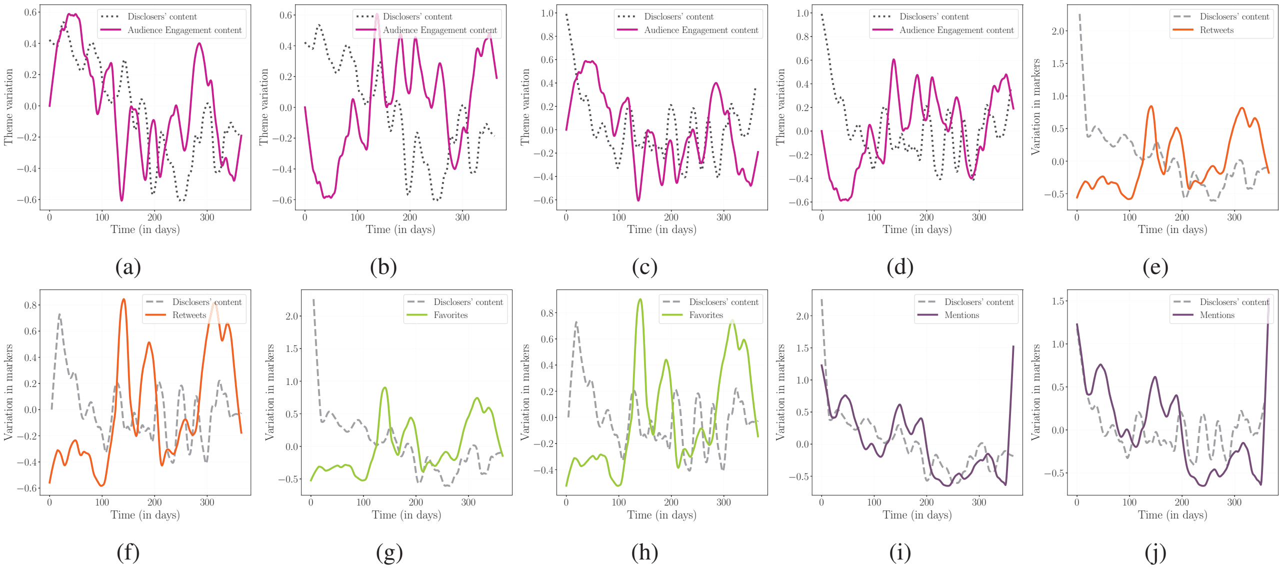
Methods For our second research question, we investigate whether the audience engagement as characterized by engagement markers and engagement content (RQ1), can predict future intimacy of disclosures. To begin, we describe how we operationalize intimacy of disclosures, and then propose and evaluate a time series forecasting model to predict these values accurately from the engagement markers and content.

Operationalizing Intimacy of Disclosures. To operationalize the disclosure process, we refer to the Social Penetration theory that models self-disclosure as a process of building intimate interpersonal relationships. We adopt one of the measures proposed by the theory i.e. depth of disclosure or *intimacy* to operationalize disclosure in our work. The depth of disclosure relates to the degree of intimacy i.e. "how open or close someone can become with another person despite their anxiety over self-disclosure". In the context of mental health related self-disclosures on Twitter, depth of disclosure would denote the extent to which the discloser continues to share information about their experiences specific to their stigmatizing condition. Given the lack of availability of ground truth data on disclosure intimacy and because discrete human judgments from a specific post may not be applicable across all users, to measure intimacy of disclosures from the textual content of disclosers' tweets, we use the following hybrid approach leveraging topic modeling and human annotations (Chancellor et al. 2016).

I. Manual annotation of disclosers' topics: Adopting the results from topic models built over disclosers' data as a thematic representation of their content (RQ1), we employed three human raters to analyze the top contributing keywords per topic and then label the level of intimacy disclosed via the topic. We defined the levels of intimacy to span a three-point Likert scale—low (1), medium (2) and high (3) motivated by prior work (Taylor and Altman 1975). First, the raters manually browsed a sample of tweets by the disclosers to familiarize themselves with the content. Then, corresponding to this rating scale, they created a set of rules to annotate each topic with one of the three levels.

High intimacy of disclosure (score of 3). This included topics specific to the experiences of schizophrenia, information that is rarely expressed on a public social media platform like Twitter. For example, topics around symptomatic expressions, social support and stigma related to mental illnesses were included in this category.

Medium intimacy of disclosure (score of 2). This category included behavioral expressions related to functioning, so-



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 2: Patterns in audience's engagement content and engagement markers with respect to Disclosers' data. We show these patterns for 10 cases: (a) disclosers' data & audience engagement content both related to schizophrenia experiences; (b) disclosers' data related to schizophrenia experiences & audience engagement content unrelated to schizophrenia experiences; (c) disclosers' data unrelated to schizophrenia experiences & audience engagement content related to schizophrenia experiences; (d) disclosers' data & audience engagement content unrelated to schizophrenia experiences; (e) disclosers' data related to schizophrenia experiences & retweets; (f) disclosers' data unrelated to schizophrenia experiences & retweets; (g) disclosers' data related to schizophrenia experiences & favorites; (h) disclosers' data unrelated to schizophrenia experiences & favorites; (i) disclosers' data related to schizophrenia experiences & mentions; (j) disclosers' data unrelated to schizophrenia experiences & mentions. The discloser' data is plotted with the lag at maximum correlation.

cial interactions, temporal planning that were not unusual to be shared on Twitter.

Low intimacy of disclosure (score of 1). This included topics that were totally unrelated to the disclosure of schizophrenia and consisted casual social media conversations such as political issues, entertainment, etc.

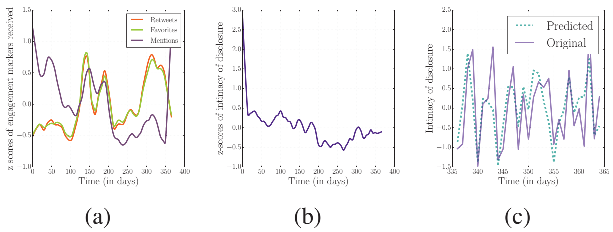
Following the manual annotation task, the raters had a high inter-rater reliability of 0.78 given by the Fleiss κ measure. Out of the 30 topics belonging to disclosers' data, this annotation task yielded 8 topics with high (3) intimacy, 7 with medium (2), and 15 with low (1) intimacy score.

II. Calculating tweet-level and time series measures of intimacy of disclosure. Given a tweet posted by the discloser, its posterior topic distribution given by the topic model (in RQ1), and the intimacy label (in RQ2) we calculate the intimacy of the tweet as a weighted sum of all topic probabilities by their intimacy labels to obtain a single score of intimacy of disclosure. We aggregate these tweet-level intimacy values per day and per discloser throughout our analysis period; we use z-scores of these aggregated values to capture their relative variation over time.

Predicting Future Intimacy of Disclosures from Audience Engagement. Given the intimacy of disclosure expressed by the disclosers and the associated engagement markers and content of the audience over time, we describe the prediction task as a time series forecasting problem. Since historical values of intimacy can also assist in predicting future intimacy values, we adopt an auto-regressive time series forecasting model. The dependent (or response) variable that is being forecasted is the time series representing daily measurements of intimacy of disclosure (obtained above). The exogenous variables (or predictors) are the engagement markers received from the audience as characterized by the following time series—number of retweets, favorites, mentions, and theme distribution of engagement content. Note that all timeseries are expressed as z-scores of average daily measurements of the variable.

Data Preparation. First, we process the data to verify stationarity assumptions of time series forecasting methods. We execute the following steps: 1) We apply a moving average transformation with a window size of 14 days to check for changes in the mean and variance over time. 2) We apply the Augmented Dickey Fuller (ADF) test, a standard test for stationarity in a series (Dickey and Fuller 1981). For the series that do not pass the ADF test, we apply a first order shift in the data and re-evaluate conditions for stationarity.

Model Fitting. We propose an Auto Regressive Integrated Moving Average with Exogenous Input (ARIMAX) model to predict the dependent variable (future intimacy) from the exogenous variables (audience engagement data). Our model is meant to forecast on day t, the intimacy of disclosure based on the exogenous variables spanning n days before t. We perform grid search over a maximum lag of 20 days for the autoregressive (p) and the moving average (q) parameters to find candidate models. Applying maximum likelihood estimation, we use log-likelihood, Akaike



(a) (b) (c)

Figure 3: (a) Engagement markers over time. (b) Intimacy of disclosure, across all 395 disclosers' data over time. (c) Predicted and original measures of intimacy over time.

& Bayesian information criterion (AIC, BIC) measures to assess goodness of fit. We validate the final model by performing in-sample rolling predictions and assessing model performance using metrics like the root mean squared error.

Results Figure 3b shows the temporal variation in intimacy of disclosure, combined across all disclosers' data. We observe a peak representing heightened levels of intimacy of disclosure on the day of disclosure ($d = 0$) and the immediately succeeding days. Since topics related to the experiences of schizophrenia were rated with an intimacy score of 3, it appears that the short period immediately following the day of disclosure continues to include high intimacy content.

With this time series of intimacy of disclosure as our response variable, we proceed to results on the forecasting model. First, on testing the stationarity assumption we find that the intimacy series, despite showing minimal changes in mean and variance over time, failed to pass the ADF test for stationarity (t = -2.68, p = 0.07). Therefore, we performed first order shift (differencing) on this series and re-evaluated for stationarity by the ADF test. If Y_t denotes the value of the time series Y at period t, then the first difference of Y at period t is equal to Y_t - Y_{t-1}. We find that the differenced series for intimacy successfully passes the ADF test (t = -9.17, p = 2 $\times 10^{-15}$). Following the same approach, we evaluate stationarity of all the exogenous variables i.e. engagement markers and content (themes). We find that all the series pass the stationarity test except for the engagement content series, specifically the following themes—"Mental Health Support/Stigma", 'Sexuality', 'Communication' and "Temporal References". We applied the same differencing technique and note that the stationarity assumptions are met.

Next, based on the grid search results for model selection and parameter tuning, we found the best lag order for the ARIMAX process i.e. the auto-regressive and moving average parameters to be p=8 and q=3. Including the differencing parameter $d = 1$ we fit an ARIMAX(8,1,3) model on the time series data (intimacy of disclosure, engagement markers and content) for forecasting. The goodness of fit of this model in terms of log-likelihood, AIC and BIC were found to be -351.9, 751.9 and 845.0 respectively.

Table 3 summarizes the ARIMAX model in terms of point mass estimates of the external variables, their 95% confidence intervals, and the corresponding p-values. We refer to this information, to examine the variables that provide the most explanatory power in the forecasting problem i.e. we ask what engagement markers and engagement content shared by the audience have high predictive power in forecasting future intimacy levels of the disclosers. We assess statistical significance here at the p=0.05 level.

Exogenous variable	estimate	P>z	95% C.I.	
mentions	**-0.0266**	**0.014**	-0.048	-0.005
retweets	-0.0197	0.748	-0.140	0.100
favorites	0.0278	0.666	-0.098	0.154
themes:appearance	0.0031	0.868	-0.033	0.039
themes:communication	0.0022	0.893	-0.030	0.035
themes:functioning	-0.0182	0.411	-0.062	0.025
themes:emotions	**0.0581**	**0.0006**	0.025	0.091
themes:mhss	0.0156	0.408	-0.021	0.052
themes:sexuality	**0.0356**	**0.0306**	0.003	0.068
themes:temporal	0.0354	0.103	-0.007	0.078
themes:other	0.002	0.918	-0.036	0.039

Table 3: Summary of point estimates of the exogenous variables in the intimacy forecasting ARIMAX model. Note that the estimates of exogenous variables in the model need to be interpreted conditional to the lags in response variable.

First, we observe that the number of mentions received is a significant predictor of future intimacy. This affirms our previous findings that mentions indicate a strong incoming engagement in ways of conversing, sharing experiences and resources with the disclosers. Next, we find two themes within the audience engagement content that are statistically significant to future intimacy levels. The first such theme is 'Emotions' with keywords such as 'care', 'worry', 'trust', 'life'. Emotional support received in cases of stigmatized conditions has been shown to help with coping and provide satisfaction in online support communities by previous studies (Vlahovic et al. 2014). Prior work has also linked intimacy to satisfaction with social support received during crisis (Hobfoll, Nadler, and Leiberman 1986). This relates with our finding that emotional content received through audience engagement can be linked to intimacy and predict future disclosure behaviors. The second significant theme is 'Sexuality'. Discussions on one's sexuality are often considered to be sensitive in nature. When they happen on a public social media platform like Twitter, they indicate the audience's intent to reciprocally converse with the disclosers about topics that are otherwise personal. This reciprocity might also motivate the disclosers to reveal more intimate aspects of their illness experiences to their audience.

Finally, to validate the model, we compute in-sample rolling predictions for the model on an out-of-sample data over the last 30 days in our year-long period of analysis. Note that the ARIMAX model forecasts the differenced intimacy of disclosure and therefore, the predicted values are compared to the original differenced values of intimacy (Ref. Figure 3c) We observe that our model is able to closely forecast the actual intimacy levels of disclosure. Assessing model performance, we find the Root Mean Square Error, Mean Absolute Error and Symmetric Mean Absolute Percentage Error measures as 0.66, 0.52 and 6.8 respectively. These values statistically establish the satisfactory performance of the model. As a final validation step, we check the residuals of the model for absence of serial correlation. We compute the Durbin-Watson statistic which tests for the null hypothesis that there is no serial correlation (Durbin 1970). We find the test statistic (Durbin-Watson's d) as 1.8, which is close to the ideal value of 2 in case of no serial correlation.

Discussion

Theoretical Implications We began this study questioning the puzzling nature of stigmatized self-disclosures made to an invisible audience on a public microblogging platform. By characterizing the audience engagement towards disclosures of schizophrenia on Twitter, we found evidence of reciprocity, both topically and temporally, in the interactions between the audience and disclosers. We also observed that using the functionalities of favorites, retweets, and mentions, the audience is able to engage with the disclosers in a variety of ways: providing support, advice, and solidarity, sharing personal experiences and online help resources, and conversing about everyday aspects of life. While these attributes are key characteristics of online support communities, their occurrence on Twitter is revealing as it lacks many critical components of an online community such as norms, moderation, roles etc. Similarly, strong social ties are considered to be the hallmark of quality support and psychological wellbeing (Burke, Marlow, and Lento 2010). However, despite lacking many aspects of a social network (Kwak et al. 2010) Twitter seems to be providing positive outcomes to individuals with a highly stigmatized condition, schizophrenia.

Further, we examined how audience engagement impacts future disclosure behavior, to understand if the disclosers gather interpersonal and social benefits through this public disclosure process. The results from our forecasting model demonstrate that key predictors, such as number of mentions, emotional support, and discussions on personal, sensitive topics can successfully forecast future intimacy of disclosures. This finding indicates that the disclosure process supports not only bridging social capital, that is, finding new acquaintances who provide access to new information and help resources, but also over time, in *bonding* social capital, in the form of reciprocity, support, and companionship (Ellison, Steinfield, and Lampe 2007). Although the nature of audience providing these social capital resources is nebulous, i.e. the disclosers may not necessarily know *who* this audience is, even if they have an imagined mental conception of who it might be (Gruzd, Wellman, and Takhteyev 2011), the reciprocal engagement that the audience provides over time confirms observations about online social platforms facilitating formation and maintenance of social capital.

Nevertheless, as argued in the literature (Steinfield, Ellison, and Lampe 2008), one might expect that disclosing about stigmatized, sensitive issues like mental illnesses to such an invisible and imagined audience might increase the likelihood of a context collapse that can hinder future disclosures. However, we find that, despite the risk of context collapse, the disclosers do not employ counteractive strategies, but rather continue to engage in schizophrenia related intimate exchanges with their audience over time.

Practical Implications Today, technology-based therapy, counseling, and intervention tools such as 7 Cups of Tea (7cups.com) and Crisis Text Line (crisistextline.org) are being increasingly adopted, where individuals in distress can talk to trained volunteers and supportive 'listeners'. There has also been an upsurge in the usage of similarly purposed artificial intelligence (AI) based conversational agents. While purported to be helpful, these services present unique opportunities and challenges. How can these tools accommodate stigmatized self-disclosures of mental illnesses and facilitate their expected social benefits? The methodology that we propose in this paper to study audience engagement towards stigmatized mental health disclosures provides a principled framework to examine the social interactions of disclosers and audiences in such contexts.

A crucial aspect of these technology-assisted therapy tools is providing the volunteers or the AI agents adequate resources, so they can successfully engage in conversations with help seekers. To do so, there is a need to capture timely feedback, in terms of the nature and quality of engagement (of the volunteer or AI agent), and their impact on future disclosure behavior of the help seekers. With our forecasting methodology (RQ2), interactive systems can be built to enable the volunteers/agents/algorithms act on the help seekers/disclosers feedback on engagement in a timely manner. Similarly, our framework for studying patterns in audience engagement with respect to what the disclosers reveal about themselves (RQ1) can be adopted to identify specific engagement patterns signaling reciprocity. Upon identification, the usage of these markers can by promoted — either manually as guidelines to volunteers and support providers or algorithmically in the case of conversational agents.

Finally, moderation efforts in online support communities and social media platforms can adopt our methodologies to similarly motivate audiences engage meaningfully with vulnerable self-disclosing individuals and to thereby create positively beneficial online therapeutic spaces.

Limitations and Future Work We acknowledge some limitations to our work. First, our findings are limited by our data acquisition capabilities. We have not probed into the nature of the audience and questions surrounding their own social media use which remains a ripe area for future work. Further, we note that the disclosers might pursue goals other than social benefits, such as trust, impression management, and social validation that we do not disentangle in our analysis. Stemming from our interest in the invisible audience, we focused our attention on finding evidence for a general form of social benefits received by disclosure. Studying the alignment between discovered patterns of audience engagement and specific disclosure goals constitutes an interesting direction for future research. Further, the social benefits that we identify in our study (such as reciprocity) need further validation using self-reported data — for example, their impact on psychological outcomes in the discloser. Qualitative data such as interviews can be powerful in complementing this line of work. Finally, in our operationalization of intimacy of disclosures, we limit our focus to studying the impact of active, incoming audience engagement. But, future work could examine how non-responsive or non-supportive audience impacts future disclosure behaviors.

Conclusion

In this paper, we provided some of the first empirical insights into the nature of audience engagement received in response to broadcasting self-disclosures of schizophrenia on Twitter. Characterizing and examining the patterns of audience engagement with respect to the disclosers' content, we find evidence of reciprocity. Further, our results from a forecasting model demonstrate that components of audience engagement such as mentions, emotional support and discussions around personal life are strong predictors of future disclosure behaviors. Our work informs the social benefits that disclosers obtain from Twitter and has implications to technology mediated support spaces on the internet.

Acknowledgments

We thank the members of the SocWeB Lab for their helpful feedback at the time of paper writing. Ernala and De Choudhury were supported by an NIH grant #R01GM112697.

References

Altman, I., and Taylor, D. 1973. Social penetration theory. *New York: Holt, Rinehart &\Mnston*.

Andalibi, N.; Haimson, O. L.; De Choudhury, M.; and Forte, A. 2016. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *CHI*.

Andalibi, N.; Ozturk, P.; and Forte, A. 2017. Sensitive self-disclosures, responses, and social support on instagram: the case of #depression. In *CSCW*.

Archer, R. L. 1980. Self-disclosure. *The self in social psychology*.

Bazarova, N. N., and Choi, Y. H. 2014. Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites. *JCM*.

Birnbaum, M. L.; Ernala, S. K.; Rizvi, A. F.; De Choudhury, M.; and Kane, J. M. 2017. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *J. Med. Internet Res*.

Burke, M.; Marlow, C.; and Lento, T. 2010. Social network activity and social well-being. In *CHI*.

Chancellor, S.; Lin, Z.; Goodman, E. L.; Zerwas, S.; and De Choudhury, M. 2016. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *CSCW*.

Coleman, J. S. 1988. Social capital in the creation of human capital.

Cozby, P. 1973. Self-disclosure: a literature review. *Psychol. Bull.*

De Choudhury, M., and De, S. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*.

De Choudhury, M.; Counts, S.; Horvitz, E. J.; and Hoff, A. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *CSCW*.

De Choudhury, M.; Sharma, S. S.; Logar, T.; Eekhout, W.; and Nielsen, R. C. 2017. Gender and cross-cultural differences in social media disclosures of mental illness. In *CSCW*.

Derlaga, V. J., and Berg, J. H. 2013. *Self-disclosure: Theory, research, and therapy*.

Dickerson, F. B.; Sommerville, J.; Origoni, A. E.; Ringel, N. B.; and Parente, F. 2002. Experiences of stigma among outpatients with schizophrenia. *Schizophrenia bulletin*.

Dickey, D. A., and Fuller, W. A. 1981. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*.

Durbin, J. 1970. Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables. *Econometrica*.

Ellison, N. B.; Steinfield, C.; and Lampe, C. 2007. The benefits of facebook "friends:" social capital and college students use of online social network sites. *JCMC*.

Ernala, S. K.; Rizvi, A. F.; Birnbaum, M. L.; Kane, J. M.; and De Choudhury, M. 2017. Linguistic markers indicating therapeutic outcomes of social media disclosures of schizophrenia. *Proc. ACM Hum.-Comput.Interact.*

Goffman, E. 2009. *Stigma: Notes on the management of spoiled identity*.

Gruzd, A.; Wellman, B.; and Takhteyev, Y. 2011. Imagining twitter as an imagined community. *Am. Behav. Sci.*

Haimson, O. L., and Hayes, G. R. 2017. Changes in social media affect, disclosure, and sociality for a sample of transgender americans in 2016's political climate. In *ICWSM*.

Helliwell, J. F., and Putnam, R. D. 2004. The social context of well-being. *Philos. Trans. Royal Soc. B*.

Hobfoll, S. E.; Nadler, A.; and Leiberman, J. 1986. Satisfaction with social support during crisis: intimacy and self-esteem as critical determinants. *J. Pers. Soc. Psychol.*

Jiang, L. C.; Bazarova, N. N.; and Hancock, J. T. 2013. From perception to behavior: Disclosure reciprocity and the intensification of intimacy in computer-mediated communication. *Commun. Res.*

Joinson, A. N., and Paine, C. B. 2007. Self-disclosure, privacy and the internet. *The Oxford handbook of Internet psychology* 237–252.

Joinson, A. N. 2001. Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, 591–600. ACM.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: Liwc 2001.

Shimkunas, A. M. 1972. Demand for intimate self-disclosure and pathological verbalization in schizophrenia. *J. Abnorm. Psychol.*

Sprecher, S.; Treger, S.; Wondra, J. D.; Hilaire, N.; and Wallpe, K. 2013. Taking turns: Reciprocal self-disclosure promotes liking in initial interactions. *J. Exp. Soc. Psychol.*

Steinfield, C.; Ellison, N. B.; and Lampe, C. 2008. Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of Applied Developmental Psychology*.

Taylor, D. A., and Altman, I. 1975. Self-disclosure as a function of reward-cost outcomes. *Sociometry*.

Vlahovic, T. A.; Wang, Y.-C.; Kraut, R. E.; and Levine, J. M. 2014. Support matching and satisfaction in an online breast cancer support community. In *CHI*.

Yang, D.; Yao, Z.; and Kraut, R. E. 2017. Self-disclosure and channel difference in online health support groups. In *ICWSM*.

Zhang, R. 2017. The stress-buffering effect of self-disclosure on facebook: An examination of stressful life events, social support, and mental health among college students. *Comput. Hum. Behav.*