Privacy in Crisis: A study of self-disclosure during the Coronavirus pandemic

Taylor Blose, * Prasanna Umar, † Anna Squicciarini ‡ and Sarah Rajtmajer §¶
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA, USA
*thb5018@psu.edu, †pxu3@psu.edu, ‡acs20@psu.edu, §¶smr48@psu.edu

Abstract—We study observed incidence of self-disclosure in a large dataset of Tweets representing user-led English-language conversation about the Coronavirus pandemic in the month between March 1 and April 3, 2020. Using an unsupervised approach to detection of voluntary disclosure of personal information, we provide early evidence that situational factors surrounding the Coronavirus pandemic may impact individuals’ privacy calculus. Text analyses reveal topical shift toward supportiveness and support-seeking in self-disclosing conversation on Twitter. We run a comparable analysis of Tweets from Hurricane Harvey to provide context for observed effects and suggest opportunities for further study.

Index Terms—Privacy, Twitter, Unsupervised learning, Social implications of technology, Text mining

I. INTRODUCTION
At the time of writing, eight in 10 Americans are directly impacted by restrictions on movement and travel, aimed at slowing the Coronavirus pandemic. Thirty two states have issued statewide directives requiring residents to remain inside with few exceptions, and more states are expected to follow in the coming days [1]. Experts suggest that these measures may need to continue for 18 months or more [2].

Unsurprisingly, there has been an unprecedented surge in online activity. A few weeks ago, the FCC granted Verizon’s request for additional spectrum to handle increased Internet usage due to the pandemic [3]. While, Netflix and YouTube are slowing down in certain areas to mitigate network congestion [4]. Much of the increased traffic extends beyond typical Internet surfing and video streaming, as people find ways to leverage online resources to stay connected with one another, personally and professionally. In a couple of months, the video-conference app Zoom has seen an increase from 10 million to more than 200 million daily users [5]. Beyond their prototypical professional and educational uses, Zoom and similar tools are increasingly supporting routine social activities – virtual happy hours, virtual yoga, blind dating and worship services [6]–[9]. Likewise, Facebook is experiencing “big surges” in usage, with voice and video calls on WhatsApp and Facebook Messenger at more than double their usual levels. The company is readying for greater increase in activity as the virus continues to spread [10].

It is to be expected that this move to life online will magnify privacy risks for individual users, most simply, because the expanded breadth and depth of online activity brings increased opportunity for privacy violations. But there are other more complex, sociological reasons as well. The emergence of virtual play dates and book clubs suggests that during a time of social distancing, people are looking for ways to stay close (e.g., “apart but together” campaigns). This may be particularly the case for the many individuals facing heightened anxiety, stress and depression due to social isolation, grief, financial insecurity, and of course health-related fears of the virus itself [11], [12].

The literature on social communication suggests that interpersonal connectedness and relationship development is fundamentally facilitated through iterative self-disclosure [13], that is, intentionally revealing personal information such as personal motives, desires, feelings, thoughts, and experiences to others [14]. In fact, there is a robust literature on (routine) self-disclosure in online social media outside the particular domain of crisis [15]–[20]. Research indicates that as users engage in discussion online, they leverage self-disclosure as a way to enhance immediate social rewards [21], increase legitimacy and likeability [22], and derive social support [23]. Despite the “upsides”, i.e., socially adaptive motivations for disclosure, we know that self-privacy violations can come at a cost, leaving users exposed to victimization from identity theft, cyber fraud and other crimes [24], discrimination in job searches, credit and visa applications [25], harassment and bullying [26]. Indeed, studies have repeatedly shown that the overwhelming majority of users have privacy concerns about their online interactions [27]. These concerns evolve over time, tied to the days events and the longer arc of shifting norms [28], [29].

Little is known about the evolution of users’ sharing practices during crisis. We know that victims of Hurricane Harvey sought assistance through social media, in some cases revealing their full names and addresses online [30]. But what we are witnessing in the case of the Coronavirus pandemic is inherently distinct from previous crises in important ways. COVID-19 is a global, relatively protracted acute threat. Unlike natural disasters or military engagements, the pandemic has left communications infrastructure intact. Digital outlets have become lifelines. We posit that self-privacy violations

*Corresponding author
during the Coronavirus pandemic can help individuals feel more socially connected during a time of anxiety and physical distance, even though the long term impact of these disclosures is unknown.

In this paper, we carry out a large scale analysis of instances of self-disclosure in a dataset of 3,785,312 Tweets representing user-led conversations on Coronavirus-related topics. We leverage an unsupervised method for identifying and labeling voluntarily disclosed personal information, both subjective and objective in nature. Our main finding reveals an expected and unprecedented uptick in instances of self-disclosure, particularly related to users’ emotional state and personal experience of the crisis. To our knowledge, this work is the first to study self-disclosure on social media during the Coronavirus pandemic.

Critically, we run comparative analyses on a dataset of Tweets representing conversations about Hurricane Harvey in late summer 2017. The Harvey study provides valuable context for these unprecedented times, suggesting similarities and differences that better inform our observations during the current crisis and the privacy and crisis literatures in general.

II. RELATED WORK

Research in the space of online interactions has sought to understand the actualization of self-disclosure in digitally-mediated social communication. Studies suggest that disclosure behaviors in online environments may be meaningfully different than their offline counterparts, e.g., anonymity and lack of nonverbal cues afforded by social media may encourage greater disclosure of sensitive information [31], [32]. Similar findings are reported in [33], where authors explore the impact of content intimacy on self-disclosure. It is well-established for face-to-face communication that people disclose less as content intimacy increases, but this effect seems to be weakened in online interactions.

Recent work has positioned online self-disclosure as strategic behavior targeting social connectedness, self-expression, relationship development, identity clarification and social control [34]–[36]. Voluntary disclosure of personal information has been associated with improved well-being, meaningfully related to increased informational and emotional support [37].

There are, however, potential costs to users’ privacy, including the unforeseen use and sharing of disclosed data. Early work by Acquisti and Gross [38] suggested that social network users were neither fully aware nor responsive to privacy risks. Over time, studies have captured a shift toward increased privacy awareness [39], [40], but there remains great variability in information sharing behaviors amongst individuals and across platforms [41]. It has been shown that culture plays a role in disclosure decisions [42]–[44], as does gender [45] and socioeconomic status (SES) [46]. Overarching, the cost-benefit analyses underlying an individual’s decision to share in the presence of privacy risk is postulated by social exchange theory [47] and re-framed in the context of online social networks as the so-called privacy calculus [48], [49].

Critically, work in a number of domains suggests that contextual and situational factors, e.g., trust, anonymity, financial incentives, are embedded within the privacy calculus [50]–[52]. Amongst these factors, emotion has also been suggested to play a meaningful role in privacy behaviors [53]–[55]. This finding is in keeping with the general theory of feeling-as-information [56], whereby emotions serve as information cues directly invoking adaptive behaviors [57]. Studies linking emotion to disclosure online have thus far limited scope to considering the emotional impact of a particular website. The ways in which an individual’s mood and general emotional state impact individual privacy calculus are unknown.

To our knowledge, there is no literature examining changes to patterns of self-disclosure in crisis through the lens of privacy risk. The crisis community is interested in the related but fundamentally distinct problem of mining self-disclosed information for the purposes of identifying and deploying assistance and relief to impacted individuals and communities (see [58] for review).

This work also dovetails with the literature on detection and tagging of self-disclosure in text, e.g., [15], [59]–[63]. Chow et al. [62] developed an association rules-based inference model that identified sensitive keywords which could be used to infer a private topic. Similarly, multiple studies utilized pattern or rule based methods to detect specific types of disclosures [16], [61].

Past work has attempted to classify self-disclosure by levels, or degree of disclosure. Caliskan et al. [59] used AdaBoost with Naive Bayes classifier to detect privacy scores for Twitter users’ timelines. Bak et al. [15] applied modified Latent Dirichlet Allocation (LDA) topic models for semi-supervised classification of Twitter conversations into three self-disclosure levels: general, medium and high. Wang et al. [60] used regression models with extensive feature sets to detect degree of self-disclosure. Because the notion of sensitive information is based on user perception and context, studies on detection of self-disclosure levels are often difficult to generalize beyond their original context.

III. PRIMARY DATASET

Our primary dataset is a repository of Tweet IDs corresponding to content posted on Twitter related to the Coronavirus pandemic [64]. At the time of writing, the repository contains 87,209,465 Tweet IDs for the period of activity from January 21, 2020 through April 3, 2020. Tweets were compiled utilizing a combination of Twitter’s Search API (for activity January 21 through January 28) and Streaming API (for activity January 28 through April 3). The repository represents the full activity of selected accounts as well as topically-relevant Tweets across the platform, canvassed based on designated keywords (See Tables 1 and 2). We focused our analysis on the period March 1 through April 3 (latest available data at time of writing). In doing so, we aim to highlight an important transitional period in the pandemic for most users in the United States. In analyses that follow, we consider two sub-periods – prior to and after March 11th, the date of the
World Health Organization’s pandemic declaration and just two days preceding U.S. President Trump’s declared national emergency.

The period March 1 through April 3 is represented by 42,943,973 unique Tweet IDs in the dataset. Text and metadata corresponding to these IDs were obtained through rehydration using the Twarc Python library. Of the 42,943,973 Tweet IDs passed for rehydration, 39,389,715 were successfully rehydrated. The 8% loss represents deleted content, therefore unretrievable through Twitter’s API.

For the purpose of measuring and studying self-disclosure, we filtered the corpus to capture Tweets that represent original content posted by individual users. Specifically, we removed quoted Tweets, retweets as well as all Tweets associated with verified accounts and the specific organizational accounts listed in Table I. We narrow our analysis to English-language content in order to reduce situational heterogeneity and maintain confidence in our labeling approach, which has been developed and validated on English-language text. The resulting corpus, which forms the basis of our analyses, consists of 3,785,312 unique Tweets.

TABLE I: Accounts followed, by start date

| Account Followed | Start Date |
|------------------|------------|
| PneumoniaWuhan, CoronavirusInfo,V2019N, CDCemergency, CDCgov, WHO, HHSGov, NIAIDNews | 1/28/2020 |
| drtedios | 3/15/2020 |

IV. AUTOMATED DETECTION OF SELF-DISCLOSURE

We use an unsupervised method [16] to detect instances of self-disclosure in our dataset. Consistent with the literature on detection of self-disclosure in text (see, e.g., [15], [19], [36], [60]), we consider the presence of first-person pronouns. Specifically, we consider sentences containing self-reference as the subject, a category-related verb and associated named entity. Consider the example of location self-disclosure shown in Figure 1. A first person pronoun “I” is the self-referent subject of the sentence. It is used with a location-related verb “live” in the vicinity of the associated location entity “Pennsylvania”. Notably, subjective categories of self-disclosure such as interests and feelings do not have associated named entities. These are differentiated through rule-based schemas based on subject-verb pairs. The approach described is implemented in three phases: 1) subject, verb and object triplet extraction with awareness to voice (active or passive) in the sentence; 2) named entity recognition; 3) rule-based matching to established dictionaries. Our dictionaries are adopted from [16].

As the proposed approach is based on sentence structure and syntactic resources (subject, verb, object and entities), it can be applied to any textual content. However, we acknowledge that Tweets present unique characteristics. Due to character limits and consequent emerging norms of the platform, users more frequently engage acronyms and abbreviations [65]. Relatedly, sentence structure and syntax are noisier when compared to more verbose platforms [66]. User mentions, hashtags, and graphic symbols are embedded within text. Considering these differences, we pre-processed Tweets as follows. All Unicode encoding errors were corrected. We removed markers associated with retweets (e.g., “RT”) and filtered user mentions and hashtags. Additionally, symbols like “&” and “$” were replaced with their respective word representations. Email addresses and phone numbers were replaced with placeholders “emailid” and “phonenumber”, while URLs were filtered. We also replaced contractions in the Tweets like “I’m” to “I am” and corrected incorrect use of spacing between words. These pre-processing steps enabled cleaner input to the detection algorithm.

We validated our approach on a baseline dataset of 3,708 annotated Tweets [59]. The dataset was manually labelled for presence or absence of self-disclosed personal information such as location, medical information, demographic information and so on. For our purposes, category-specific labels of self-disclosure were binarized, resulting in a total of 3,188 self-disclosing and 520 non-self-disclosing Tweets. Using the unsupervised method described, we classified the baseline data and compared against the binarized manual labels. Our approach demonstrated acceptable performance yielding precision, recall and F scores of 92.5%, 50.3% and 65.1% respectively. Recall is lower than reported in [16], attributable to differences between the taxonomy used to manually label the baseline Twitter dataset and the taxonomy used by

TABLE II: Keywords followed, by start date

| Keyword Followed | Start Date |
|------------------|------------|
| Coronavirus, Koronavirus, Corona, CDC, Wuhan-coronavirus, Wuhanlockdown, Ncov, Wuhan, N95, Kungflu, Epidemic, outbreak, Sinophobia, China | 1/28/2020 |
| covid-19 | 2/16/2020 |
| corona virus | 3/2/2020 |
| covid, Covid19, sars-cov-2 | 3/6/2020 |
| COVID-19 | 3/8/2020 |
| COVID, pandemic | 3/12/2020 |
| coronavirus, canceleverything, Coronavirus, SocialDistancingNow, Social Distancing, Social Distancing | 3/13/2020 |
| panicbuy, panic buy, panicbuying, panic buying, 14DayQuarantine, During14My14DayQuarantine, panic shop, panic shopping, panic-shop, InMyQuarantineSurvivalKit, panic-buy, panic-shop | 3/14/2020 |
| quarantine, Chinese virus, chinesevirus, stayhomechallenge, stay home challenge, stayhomechallenge, dontbeaspreader, lockdown, lock down | 3/18/2020 |
| shelteringinplace, sheltering in place, stayastsafe, stay home, trumpdemic, trump pandemic, flattenthecurve, flatten the curve, china virus, chinavirus_quarantine, PPEshortage, safelystayhome, stayathome, stay at home, stayhome | 3/19/2020 |
| GetMEPPE | 3/21/2020 |
| covidiot | 3/25/2020 |
| epiphwitter | 3/28/2020 |
| pandemon | 3/31/2020 |

1https://github.com/DocNow/twarc
2Retweets were identified through the existence of the “retweetedstatus” field in the Tweet object returned by the API. Tweets beginning with the string 'RT @' were also treated as retweeted records.
to create the unsupervised detection method (e.g., our unsupervised approach was not developed to detect categories like drug use and personal attacks, whereas these categories were explicitly offered to manual labelers). Given these differences, the performance of our automated approach is acceptable and therefore, applicable to detecting self-disclosure in Tweets.

V. TOPIC MODELING

Latent Dirichlet Allocation (LDA) [67] was used for topic modeling of Tweets in our dataset. In this approach, each document (a single Tweet) in the corpus (set of Tweets) is considered to be generated as a mixture of latent topics and each topic is a distribution over words. For a document, each word is assigned topics according to a Dirichlet distribution. Iteratively cycling through each word in each document and all documents in the corpus, topic assignments are updated based on the prevalence of words across topics and the prevalence of topics in the document. Based on this process, final topic distributions for documents and word distributions for topics are generated.

We ran topic analyses over subsets of interest within the complete Coronavirus dataset. Namely, we explored unique topic models for Tweets pre- and post-March 11, as well as Tweets with presence or absence of detected instances of self-disclosure. In addition to the cleaning steps described in Section IV, we pre-processed Tweets using tokenization, conversion to lower case, lemmatization and removal of punctuation and stopwords. Topics were generated from the resulting corpora using the LDA model within the Gensim Python library.

VI. FINDINGS

We analysed 3,785,312 Coronavirus-related Tweets, and identified that approximately 18% (681,556 Tweets) contain elements of self-disclosure. Looking more closely at daily variance, we identify a transition point in activity around March 11, 2020, as shown in Figure 2.

In the period March 1 through March 11, the average daily percentage of self-disclosing Tweets is 16.23%; from March 12 through the end of collection, April 3, the average daily percentage is 19.78%. This change in activity coincides with an escalation in severity and increased global awareness of the crisis, with the World Health Organization (WHO) officially classifying Coronavirus as a pandemic (March 11). Current events coincident with observed changes in the rate of self-disclosure are noted on March 13 and March 19 when the United States officially declared a national state of emergency, and when the governor of California issued the first statewide ‘stay-at-home’ order, respectively. These observed parallels suggest that situational context, in particular during crisis, may meaningfully influence disclosure behaviors.

We further examined messaging around the Coronavirus pandemic through topic modeling, as outlined in Section V. We compare topical variance between disclosing and non-disclosing Coronavirus-related Tweets (see Figure 3). Generally, extracted topics reflect terminology pervasive in mainstream media at the onset of the crisis, including but not limited to expected impacts of COVID-19 (health, crisis, cancel, social distance), recommendations (stay home, wash, mask), and political impact (China, Trump, country). There is no noticeable distinction between topics extracted from the subset of Tweets containing self-disclosure and those without.

An interesting distinction is noticeable in the topical breakdown before and after March 11, reported in Figure 4. Leading up to March 11 (Figure 4a), self-disclosing conversation focused on general statistics (Topic 3) and global impact (Topic

---

3https://radimrehurek.com/gensim/models/ldamodel.html
VII. COMPARATIVE ANALYSIS

For context, we compared observed self-disclosure during the Coronavirus pandemic to observed self-disclosure during Hurricane Harvey (2017). Although hurricanes are an annual expectation, the landfall duration and subsequent impact of Hurricane Harvey created a crisis throughout communities in the South Central region of the United States. This overwhelmed traditional emergency response infrastructure and affected citizens took to social media to seek emergency assistance [68], [69].

We consider a collection of 6,732,546 Tweet IDs representing posted content inclusive of keywords Hurricane Harvey, Harvey, and/or HurricaneHarvey during the 12-day period August 25, when Harvey first made landfall, through September 5, 2017 [70]. Similar to the Coronavirus dataset, we passed the set of Hurricane Harvey-related Tweet IDs through the Hydrator Tweet Retrieval Tool (v2.0)4, a sister desktop application to Twarc. We experienced a 33% loss during data reconstitution (compare to 8% loss in the Coronavirus-related dataset), attributable to Tweet and account deletion during the nearly 3 years which have passed. We filtered the resulting 4,379,462 Tweets for original content in the same fashion as we handled the Coronavirus data – removing all quoted Tweets, retweets, content from verified accounts, and non-English content as identified by Twitter. The resulting 551,061 Tweets were passed through pre-processing and unsupervised labeling to detect instances of self-disclosure, and through topic analysis as detailed in Sections IV and V.

Across all 551,061 Tweets in this subset, we observed an average 9% self-disclosure rate (49,595 Tweets) over the 12-day collection period – substantially lower than the 18%
observed in the Coronavirus-related dataset. Several factors might account for the difference. The current pandemic has been marked by efforts to maintain social distance and, as we have proposed, increased levels of disclosure may be related to relationship-building with online cohorts. But what we may also be seeing is a reflection of a general trend toward greater self-disclosure in online social media over the course of nearly three years separating the two events.

With respect to topical focus, we see important parallels between the two crises. As illustrated in Figure 5a, self-disclosing Tweets revealed emotional messaging centered on seeking immediate spiritual, physical, and monetary support (Topics 1, 2, 3) with top terms including “help”, “donate”, “victim”, “red cross”, and “relief”. These themes are much less prominent in the non-self-disclosing data. This finding across both datasets suggests that support-seeking during crisis might be a driver of self-disclosure and play a meaningful role in users’ sharing practices.

Also mirroring the Coronavirus dataset, we observe one topic representing politically-motivated conversation (Topic 4). Messages related to public news announcements and welfare updates are represented in Topics 5 and 6. While there is topical overlap between the two classes of Harvey-related Tweets, non-disclosing Tweets (Figure 5b) presented a focus on contextual information related to the crisis with relevant keywords “flood”, “Texas”, and “Houston” (Topic 4).

VIII. DISCUSSION

Perhaps the most striking observation in the analyses we have described is early evidence of heightened self-disclosure during the ongoing Coronavirus pandemic, as compared to observed disclosure during Hurricane Harvey. This global crisis is unprecedented in a number of ways, one being the scale and scope of human interaction through social media. Concerns about privacy have been at the center of discussion in popular press (see, e.g., [71]–[73]), but most of this conversation has been about privacy tradeoffs related to cellphone tracking and similar approaches to location surveillance and individual health monitoring in service to public health. Google and Apple have just announced their plans to enable opt-in Bluetooth-based COVID-19 contact tracing through Android and iOS [74]. Many are willing to sacrifice some privacy in hopes of stemming the spread of the disease and helping to accelerate the return to normalcy; others are not. Scholarly work has begun to propose “privacy first” decentralized approaches for COVID-related tracking and notification (see, e.g., [25]–[27]).

We aim, with this work, to engage the research community in the work of better understanding more subtle, voluntary self-privacy violations emergent in the Coronavirus pandemic and in crisis more generally. We know that, during Hurricane Harvey, individuals took to social media in search of immediate aid. Today, individuals are leaning on their online social communities for engagement and support. Isolation, economic uncertainty, and health-related anxiety pose serious threat to mental health and well-being [78], [79], but the potential manifestations of psychological impact in the domain of voluntary self-disclosure are unknown. The existing literature hints at the role of mood and emotion in the privacy calculus, but these relationships have not be well-established. Our analyses suggest that the current pandemic and its effects may impact self-disclosure behaviors. As the crisis continues to unfold in the next several months, additional work should aim to refine and validate these hypotheses.

REFERENCES

[1] T. Lorenz, E. Griffith, and M. Isaac, “We live in zoom now,” New York Times, 2020. [Online]. Available: https://www.nytimes.com/2020/03/17/style/zoom-parties-coronavirus-memes.html

[2] N. Ferguson et al., “Impact of non-pharmaceutical interventions (npis) to reduce covid-19 mortality and healthcare demand,” Imperial College COVID-19 Response Team Report, 2020. [Online]. Available: https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf

[3] M. James, “Fcc ok for verizon request for more capacity as network use spikes,” Government and Technology, 2020. [Online]. Available: https://www.govtech.com/news/FCC-OKs-Verizon-Request-For-More-Capacity-as-Network-Use-Spikes.html

[4] H. Gold, “Netflix and youtube are slowing down in europe to keep the internet from breaking,” CNN Business, 2020. [Online]. Available: https://www.cnn.com/2020/03/19/tech/netflix-internet-overload-eu/index.html
