scientific reports

OPEN

Breaking the silence: leveraging social interaction data to identify high-risk suicide users online using network analysis and machine learning

Damien Lekkas[1,2] & Nicholas C. Jacobson[1,2,3,4]

Suicidal thought and behavior (STB) is highly stigmatized and taboo. Prone to censorship, yet pervasive online, STB risk detection may be improved through development of uniquely insightful digital markers. Focusing on Sanctioned Suicide, an online pro-choice suicide forum, this work derived 17 egocentric network features to capture dynamics of social interaction and engagement within this uniquely uncensored community. Using network data generated from over 3.2 million unique interactions of $N = 192$ individuals, $n = 48$ of which were determined to be highest risk users (HRUs), a machine learning classification model was trained, validated, and tested to predict HRU status. Model prediction dynamics were analyzed using introspection techniques to uncover patterns in feature influence and highlight social phenomena. The model achieved a test AUC = 0.73 ([0.61, 0.85], 95% CI), suggesting that network-based socio-behavioral patterns of online interaction can signal for heightened suicide risk. Transitivity, density, and in-degree centrality were among the most important features driving this performance. Moreover, predicted HRUs tended to be targets of social exchanges with lesser frequency and possessed egocentric networks with "small world" network properties. Through the implementation of an underutilized method on an unlikely data source, findings support future incorporation of network-based social interaction features in descriptive, predictive, and preventative STB research.

The World Health Organization (WHO) reports an estimated 703,000 annual deaths worldwide due to suicide, making it the fourth leading cause of death among individuals aged 15–29 years[1]. In the United States alone, 2020 saw approximately 1.2 million reported suicide attempts[2]. Suitably identified as a global health threat[1], suicide is etiologically heterogeneous and phenomenologically complex. Its inherently sensitive, taboo, and, in some cases, illegal nature makes studying suicide and its associated mosaic of pre-motivational, motivational, and volitional factors[3] particularly challenging. However, major shifts in the *modus operandi* of day-to-day communication, specifically the increased reliance on technology-driven social interactions, have provided an inherent ability to historically preserve details of social engagement. This has begun to surmount traditional barriers associated with studying a relatively rare, phenotypically variable, and socially undesirable suite of thoughts and behaviors. Such "digital footprints" of online activity generate an unprecedented amount of information on individual thought, action, and reaction which can be leveraged to study suicidal thought and behavior (STB) with greater contextual and temporal granularity.

The field of suicidology has rapidly begun to take advantage of this modern digital zeitgeist, leveraging data collected from online communities[4] and social media platforms[5,6] to describe[7], detect[8], and predict[9] STB. Applying a combination of natural language processing[10], network analysis[11], and machine learning techniques[12], among others, research has demonstrated the ability to extract meaningful signals of STB from the content and behaviors reflected in online activity. Works have applied and tested prominent suicide theories[13], highlighted

[1]Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College, 46 Centerra Parkway, Suite 300, Office #313S, Lebanon, NH 03766, USA. [2]Quantitative Biomedical Sciences Program, Dartmouth College, Hanover, NH, USA. [3]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA. [4]Department of Psychiatry, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA. ✉email: Damien.Lekkas.GR@dartmouth.edu

important semantic features[14], characterized conversational topics[15], profiled emotion[16], studied information flow[11], and probed one or more specific concepts such as stigma[17], anti-mattering[18], and negative social comparison[19]. The richness of the data that drive these endeavors is owed in part to source ubiquity and ease of access, the potential for interactive anonymity, and a temporal density of sampling that typically characterizes a "passively collected" record of internet activity.

Despite the impressive and invaluable collection of scholarly efforts that have developed around this data, the literature has largely overlooked *patterns in social interaction* as potential digital markers of STB. Indeed, efforts within network analysis have almost exclusively focused on modeling STB as a complex system of construct associations,[20] providing insights, for example, into the relative importance (centrality) of different theory-guided risk and protective factors.[21] This is especially surprising considering the large body of work surrounding the applications of social network analysis to the study of related disorders such as depression,[22] anxiety,[23] eating disorders,[24] and borderline personality disorder.[25].

To the authors' knowledge, only two studies have explicitly looked at the structure of online social network interactions to better understand STB. The first focused on modeling suicidal ideation on Twitter and found a higher degree of reciprocal connectivity among users with suicidal content than previously reported in other studies on non-suicidal Twitter users.[26] In addition, the identification of bridge and hub nodes within constructed retweet networks provided evidence for a potential contagion effect, propagating suicidal content to a broader audience of non-suicidal individuals.[26] A second work looked at a popular Chinese social media platform, Sina Weibo.[27] Herein, the authors built a series of temporal "moment" networks to graphically describe the interactional patterns of users with and without suicidal ideation. Akin to the previous Twitter work, a contagion effect was observed where non-suicidal users expressed suicidal ideation only after interacting with suicidal individuals on the platform.[27] Together, these efforts underline a clear potential for the ability of online social network structures to provide useful insight into the social dynamics of STB.

It is important to recognize that, although there have been tremendous benefits in the utilization of online data, the taboo and socially undesirable properties of suicide heavily censor its expression,[28] ultimately imposing a limit in the ability to capture naturalistic displays of STB. This censorship can be human, algorithmic,[29,30] self-imposed, or community-driven.[31,32] To circumvent these concerns, an ideal scenario would involve the collection of behavioral data from a community where interactions are not tempered by the fear of social reproach. Sanctioned Suicide, a self-described pro-choice suicide forum providing an anonymous "safe space to discuss the topic of suicide without the censorship of other places," presents a unique example of an unfiltered, naturalistic medium of online STB communication—a platform that effectively chronicles the social interactions of suicidal individuals as they interact with like-minded others to vent, empathize, and share experiences. Probing the dynamics of such a community offers an exceptionally rare opportunity to fully leverage many of the aforementioned benefits of online "big data" without needing to contend with the artificiality borne from censorship behaviors and practices on mainstream media. Moreover, recent work has indicated that there are novel STB-related insights to be gleaned from studying the communication of users within this forum.[33].

To highlight risk-signaling patterns of social interaction in online STB, the current work marries the methodological promise of social network analysis with the unusual level of behavioral transparency afforded from data collected within a fringe, pro-suicide, online community of like-minded individuals. Specifically, activity across $N = 192$ representative users on the Sanctioned Suicide forum was modeled using network analysis techniques to derive social network features that capture interaction and engagement within and across threads. These features were then leveraged within a machine learning framework to evaluate and introspect the value of social network-based interaction patterns as predictors of heightened suicide risk, operationalized in terms of user-expressed and community-confirmed suicidal attempts. Accordingly, this research was guided by the following aims:

1. Provide a repeatable, social network-based operationalization schema to capture social interaction across an online forum.
2. Derive network structural features that broadly quantify and summarize community engagement.
3. Build, validate, and evaluate a binary classification machine learning model with social network attributes as the sole predictors of a (completed) suicide attempt.
4. Use model introspection techniques to report on the relative importance and marginal directional association of network attributes for outcome prediction.
5. Contextualize key network features against the backdrop of suicidology and highlight high-risk social phenomena for further STB-related research.

Methods

Data source characteristics

The pro-choice suicide forum, "Sanctioned Suicide," served as the digital environment of interest in the current study. Sanctioned Suicide presents as an international stigma-free haven of free speech for those who hold socially undesirable opinions and attitudes regarding their desire and right to end their lives. The forum provides a rare platform on the surface web where users can communicate and share suicide-related thoughts, behaviors, and associated negative experiences while remaining anonymous. Although the forum is publicly accessible by anyone browsing the internet, there are strict rules and moderation efforts in place to prevent the disclosure of personally identifying information. Anyone can read the content presented across the forum, but the ability to post and interact with members of the community is granted only to those who complete a screened registration process. Registration requires acknowledgement of being at least 18 years of age and includes a free-form response section where individuals must explain in detail why they want to join. Importantly, no account creation or interaction with users on Sanctioned Suicide was carried out by the researchers of this work.

All methods concerning the acquisition and processing of data utilized in the current study were carried out in accordance with relevant guidelines and regulations. This study (#00,032,141) and all associated protocols were approved by the Committee for the Protection of Human Subjects (CPHS) at Dartmouth College. CPHS deemed this study and all associated protocols to present no greater than minimal risk to subjects. Accordingly, written informed consent was waived and this study "exempt" from further review.

Data collection

A complete record of posting activity within the "Suicide Discussion" subforum of Sanctioned Suicide, from inception on March 17, 2018 to February 5, 2021, was programmatically collected and organized as tabular data using a custom Python (v3.8) script that primarily leveraged the *BeautifulSoup* package to parse the site's HTML and XML information.[34] This effort resulted in a dataset containing more than 600,000 time-stamped posts across nearly 40,000 threads and over 11,000 users. This posting activity information consisted of (i) thread title, (ii) thread author, (iii) post author, (iv) post date, (v) post text content, and (vi) direct mentions and references to other user comments within the post text. All information, except for post text, was used in this study. To impose an added layer of user anonymity, each username was automatically assigned a randomly generated, 32-character hashed ID. These de-identifying IDs were automatically replaced with all instances of users' online handles within the data prior to subsequent preprocessing and analysis.

Data preprocessing

Cohort sampling and outcome labeling

Aims 3 and 4 of this study concerned the development and introspection of a binary classification predictive model that could identify the highest-risk users (HRUs) on Sanctioned Suicide from network-based patterns of their social interactions. Adopting the framework of modern ideation-to-action models, the highest risk was equated to a current state of active volition,[3] herein specifically defined by a written account of a suicide attempt while active on the forum. A structured approach to select a subset of appropriate users and identify HRUs was devised based on the findings discussed through the New York Times investigation into Sanctioned Suicide[35] as well as the authors' thorough review of the forum content. Moreover, this strategy was described and utilized in a previously published analysis of users on the Sanctioned Suicide forum.[33] To reiterate herein, data was first filtered by searching for thread titles with the following keywords/phrases: "bus is here," "catch the bus," "fare well," "farewell," "final day," "good bye," "goodbye," "leaving," "my time," "my turn," "so long," and "took SN." Of note, "catch the bus" is a euphemism adopted by the community to symbolize suicide,[33] while "SN" is short for sodium nitrate, an increasingly popular chemical used in suicide-related methodology. These terms were used to identify "goodbye threads" on Sanctioned Suicide, and thus have the highest probability of signaling for an impending attempt. From this scheme, threads were flagged, and each thread was manually read in its entirety. The thread author was determined to be an HRU if and only if the following criteria were met: (i) no record of post or engagement activity by the user after the date of the last post within the goodbye thread, (ii) no other users mentioned seeing the user "online" in their profile status after the date of the last post in their goodbye thread, and (iii) the thread contained a "confirmation" or mention of suspected completion as stated by users who either allegedly knew the attempter personally in real life or who directly interacted with the user during their attempt. These conditions fall closely in line with findings from the New York Times' investigation which linked behaviors on Sanctioned Suicide with real-life suicide incidents.[35] As in the previous work,[33] this strategy resulted in the identification of $n = 48$ users as attempting, and due to the nature of these threads, the community dialog surrounding these attempts, and this study's filtering criteria, likely completing suicide. Given the anonymity of this platform and the nature of online interactions, it is not possible to have validate the truth of any claims; however, this manually and contextually selected group of $n = 48$ users flags a likely subset of individuals for whom this outcome came to pass, ultimately representing among the highest risk users on the platform.

Unsurprisingly, user activity across the forum was highly variable due to a myriad of factors, including the amount of time spent as a forum member, total number of posts, posting frequency, and the size or word count of posts. For the purposes of the current analysis, it was important to obtain a cohort that was both holistic in its representation of activity across the forum and consisted of suitable behavioral controls for the HRUs. To accomplish this, each HRU was matched to three control users. Suitable matches were determined based on first selecting users with an equivalent duration of time since first posting on the forum (within 2 days and determined by the difference in days from a user's most recent post and their first post on Sanctioned Suicide) and then selecting users with the closest total number of posts on record. With few exceptions due to three HRUs of unusually high activity, this step resulted in controls with a total number of posts that were within one order of magnitude of their respective HRU match. Accordingly, $n = 144$ users were selected as representative controls. This yielded $N = 192$ users for this study's cohort. Details on the activity of each user in the cohort are available as Supplementary File 1. This file also contains results from Wilcoxon rank-sum tests (non-parametric comparison of medians, M) which ensured non-significant differences between HRU and control groups based on total number of posts ($M_{HRU} = 135.5$, $M_{control} = 105.5$, $P = 0.254$), total number of words ($M_{HRU} = 7391$, $M_{control} = 5456.5$, $P = 0.238$), and the range (in days) of post activity on the forum ($M_{HRU} = 57$, $M_{control} = 57$, $P = 0.996$).

Network-based quantification of forum activity

In line with Aim 1 of this research endeavor, this work derived a consistent and repeatable network-based operationalization of forum posting behavior. While there are undoubtedly many ways to justifiably quantify forum interaction, the current approach sought to model engagement and exposure where the "thread" was the fundamental reference point for quantification. A thread represents an independent and self-contained conversation within the broader scope of a forum, initiated and defined on the part of the thread author and open to response from other users who are drawn to its premise and content. A thread can be thought of as a microcosm with its own dynamic and evolution separate from other threads and the broader forum universe. To holistically capture forum activity for the selected cohort, all threads where at least one of the 192 users initiated a thread (thread author), posted in a thread (post author), or was directly mentioned (@[username]) or directly quoted ([username] said:) in a thread were first filtered out of the broader dataset. For each of these resulting 13,796 threads, values were then assigned to interactions among all thread participants based on the nature and temporal order of their posting activity within the thread. Inspired from previously operationalizations of social capital[36] and information flow[37,38] within online communities, three rules defined the quantification scheme:

1. Self-directed interactions captured the creation of a new thread (and the associated first post in the thread), where a value of 1.0 was given to link the thread author to themself.
2. If a post contained a mention, that mention was used to assign a value of 2.0 to link from the post author to the mentioned user.
3. Additionally, each post represented a set of ties which linked from that post's author to all other users who had previously posted in the thread. Moreover, the value of each connection from the post's author to all other previous thread participants varied and was based on the time, in days, between the post of interest and the last post made by the target thread participant. The exact value for a user-user dyad based solely on making a new post in a pre-existing thread was based on the following decay formula:

$$interaction_{A\rightarrow B} = 1.0 * e^{-0.1*(tA-tB)}$$

where *tA* is the date of the post of interest authored by User A, and *tB* is the date of the last post authored by User B in the thread. Under this calibration, the magnitude of an interaction drops by half after seven days. This means, for example, that User A posting in a thread the same day as when User B last posted holds double the weight (1.0) compared to a situation where User A posted one week after User B last posted in the thread (~0.5). Essentially, the greater the time elapsed between a new post and the last post of another user, the lower the value that new post carries as an interaction between the new post's author and the other user in the thread. This formula is repeated between User A and all users who have previously posted in the thread. Naturally, this does not include users who have records of posting in the thread after User A's post date as they have not interacted with the thread up to this point.

Following the application of this scheme, all values belonging to each unique directional dyadic interaction (e.g., User A → User B, User B → User A, User C → User A, *et cetera*) were then summed to represent the total magnitude of that dyadic interaction within the thread. Accordingly, these values comprised a weighted edgelist with which to ultimately construct directed weighted graphs for network feature extraction. For brevity and intuition, this process has been summarized and illustrated in panel A of Fig. 1.

Construction of global interaction networks

Using the *networkX* package (v2.4) in Python,[39] two interaction network types of differing scope were built: (i) thread-specific and (ii) thread-agnostic. While the thread-specific networks focused on the summation of interactivity *within* each thread and were built directly from the edgelists derived in "Data Preprocessing" above, the thread-agnostic network focused on the summation of interactivity *across* threads and was built by first collapsing all thread-specific edgelists into a single representative edgelist. This was done in the same way as handling repeat dyadic interactions within any single thread as described previously. The resulting thread-agnostic network was the same for all users in the cohort as it included all interactions for all users in one graph. However, from the derived series of thread-specific networks, only a subset of the 13,796 networks comprised each user's collection, namely those networks represented by threads in which the user was involved. For reference, this step is summarized and illustrated in panel B of Fig. 1.

Extraction of egocentric networks

Egocentric networks are subnetworks within weighted directional networks that include a focal node (ego) and all the nodes (alters) that fall within a specified distance (radius) from the ego. The scope and inclusivity of the egocentric network can be altered by modulating the radius of interest and by considering both inward-directed and outward-directed edges. In the current analysis, egocentric networks were constructed with a radius of 1 and considered both inward-directed and outward-directed edges simultaneously. Importantly, based on the quantification scheme described above, edge weights and distances were inversely related. Therefore, any two nodes were closer in distance to each other if their corresponding edge weight was higher. Egocentric networks for each user (ego) were extracted using *networkX* with built-in functionalities for this task. As a result, each user had $n + 1$ egocentric networks, where n is the number of thread-specific egocentric networks (each from a thread in which the user was involved), and the "1" accounts for the thread-agnostic egocentric network derived from its associated global network precursor. For reference, this step is summarized and illustrated in panel C of Fig. 1.

Derivation of network structural features

To fulfill Aim 2 of this work, seven core network features were calculated for each egocentric network extracted. These features were selected because of their common implementation within the applied social network science literature and because of their relative interpretive ease. The features broadly fell into two types: those that summarized the overall egocentric network structure (i.e., order, density, transitivity) and those that summarized the ego node's position within the egocentric network (i.e., betweenness centrality, in-degree centrality, out-degree centrality, out-degree to in-degree ratio). The *networkX* package was used to programmatically extract these

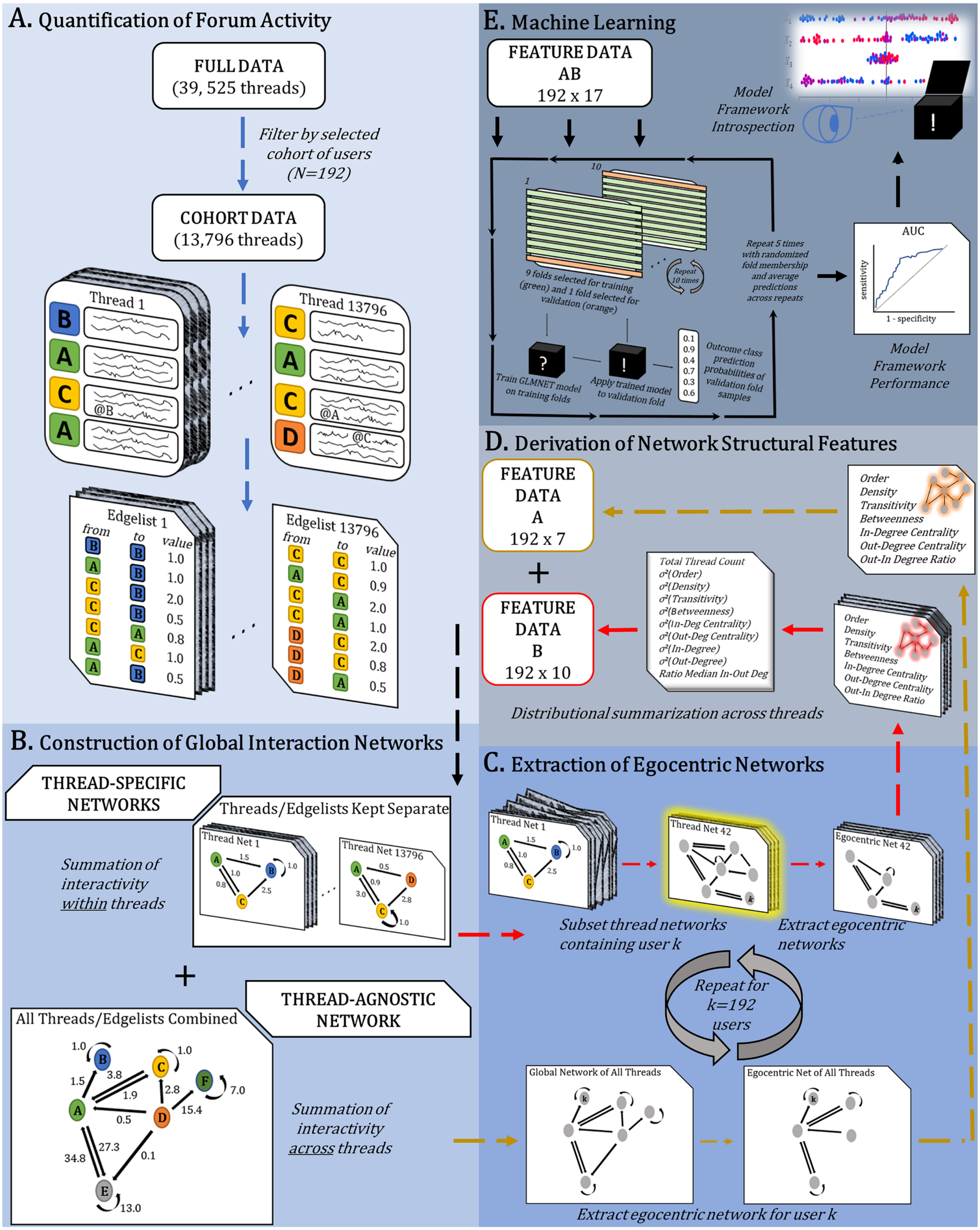
A. Quantification of Forum Activity
FULL DATA
(39, 525 threads)
Filter by selected cohort of users (N=192)
COHORT DATA
(13,796 threads)
Thread 1
B
A
C
@B
A
Thread 13796
C
A
C
@A
D
@C
Edgelist 1
from
to
value
Edgelist 13796
B. Construction of Global Interaction Networks
THREAD-SPECIFIC NETWORKS
Threads/Edgelists Kept Separate
Summation of interactivity within threads
Thread Net 1
Thread Net 13796
+
All Threads/Edgelists Combined
THREAD-AGNOSTIC NETWORK
Summation of interactivity across threads
C. Extraction of Egocentric Networks
Thread Net 1
Thread Net 42
Egocentric Net 42
Subset thread networks containing user k
Extract egocentric networks
Repeat for k=192 users
Global Network of All Threads
Egocentric Net of All Threads
Extract egocentric network for user k
D. Derivation of Network Structural Features
FEATURE DATA A 192 x 7
+
FEATURE DATA B 192 x 10
Order
Density
Transitivity
Betweenness
In-Degree Centrality
Out-Degree Centrality
Out-In Degree Ratio
Total Thread Count
σ²(Order)
σ²(Density)
σ²(Transitivity)
σ²(Betweenness)
σ²(In-Deg Centrality)
σ²(Out-Deg Centrality)
σ²(In-Degree)
σ²(Out-Degree)
Ratio Median In-Out Deg
Distributional summarization across threads
E. Machine Learning
FEATURE DATA AB 192 x 17
Model Framework Introspection
1
10
9 folds selected for training (green) and 1 fold selected for validation (orange)
Repeat 10 times
Repeat 5 times with randomized fold membership and average predictions across repeats
Train GLMNET model on training folds
Apply trained model to validation fold
0.1
0.9
0.4
0.7
0.3
0.6
Outcome class prediction probabilities of validation fold samples
AUC
sensitivity
1 - specificity
Model Framework Performance

Figure 1. Data Preprocessing and Modeling Scheme. *Note*. All presented values and networks are hypothetical simplifications of the actual data utilized in analysis and are illustrated here for intuition and summarization purposes.

features from each egocentric network. For a user's thread-agnostic egocentric network, these features directly described the local interaction dynamics of the user within the greater context of the cohort's activity across Sanctioned Suicide. However, to further characterize a user's engagement and capture variability in a user's local embedding within the community, these seven features were statistically summarized across the user's respective thread-specific series of egocentric networks. This yielded 10 additional features for a total of 17 to be used in downstream predictive modeling. Table 1 provides an enumeration and definition of each feature, with contextual illustration provided in panel D of Fig. 1.

Machine learning modeling and introspection

The *caret* package[40] in R (v4.0.2) was used to carry out machine learning modeling. Addressing Aims 3 and 4, the 17 egocentric network-based features of user interaction on Sanctioned Suicide were leveraged to train and validate a binary classification model. This model was built for, and evaluated in, its ability to flag highest risk users among $N = 192$ representative users on the platform using only network-based quantifications of forum engagement behavior. To avoid an over-engineered and intractably complex solution for this task, a logistic regression model with elastic net penalization[41] was used as the core model algorithm. With this strategy in mind, a held-out test set of $n = 72$ users was first derived by randomly sampling 18 HRUs and 54 controls. Using data from the remaining $n = 120$ users (30 HRUs and 90 controls), the model was trained and validated within a five-times-repeated, k-fold cross-validation framework with $k = 10$. To mitigate biases due to class imbalance, oversampling of the minority class was performed using the popular and effective Synthetic Minority Oversampling TEchnique (SMOTE).[42] Model hyperparameters (alpha and lambda) were tuned to maximize Cohen's kappa using *caret*'s built-in grid search methodology. The resultant best-fit model was then applied to the held-out test, and performance was assessed using metrics of sensitivity, specificity, F1 score, Cohen's kappa, and the area under

		Mean [Range]		
Feature	**Definition**	**Cohort ($N = 192$)**	**HRUs ($n = 48$)**	**Controls ($n = 144$)**
order	total number of nodes	439.99 [19, 2118]	540.85 [44, 2118]	406.37 [19, 1251]
density	ratio of existing ties to total possible ties	0.32 [0.07, 0.73]	0.32 [0.07, 0.73]	0.32 [0.13, 0.71]
transitivity	ratio of existing triads to total possible triads	0.56 [0.32, 0.80]	0.55 [0.32, 0.80]	0.56 [0.38, 0.80]
betweenness centrality (BetwCent)	sum of the fraction of all paths in which the ego is on the shortest path to all pairs of alters	0.06 [0.01, 0.45]	0.06 [0.02, 0.26]	0.06 [0.01, 0.45]
in-degree centrality (InDegCent)	the fraction of alters the ego is connected to, taking only edges *to* the ego into account	0.73 [0.41, 0.97]	0.76 [0.57, 0.97]	0.72 [0.41, 0.95]
out-degree centrality (OutDegCent)	the fraction of alters the ego is connected to, taking only edges *from* the ego into account	0.71 [0.39, 0.89]	0.71 [0.43, 0.89]	0.71 [0.39, 0.87]
out-in degree ratio (OutInDegRatio)	ratio of the sum of all weighted edges directed *from* the ego to the sum of all weighted edges directed *to* the ego	1.07 [0.31, 3.15]	1.20 [0.51, 3.15]	1.03 [0.31, 2.91]
variance order (varOrder)	variance in order across all user's thread-specific ego networks	144.53 [2, 1751.62]	180.25 [17.43, 1751.62]	132.62 [2, 1180.52]
variance density (varDensity)	variance in density across all user's thread-specific ego networks	0.03 [0.01, 0.13]	0.03 [0.01, 0.10]	0.03 [0.01, 0.13]
variance transitivity (varTransitivity)	variance in transitivity across all ego's thread-specific ego networks	0.04 [0.01, 0.15]	0.04 [0.01, 0.10]	0.04 [0.01, 0.15]
variance betweenness centrality (varBetwCent)	variance in betweenness centrality across all egor's thread-specific ego networks	0.02 [0.00, 0.12]	0.01 [0.00, 0.07]	0.02 [0.00, 0.12]
variance in-degree (varInDeg)	variance in the sum of all weighted edges directed *to* the ego across all ego's thread-specific ego networks	819.60 [10.75, 22,979.64]	964.12 [83.69, 7243.07]	771.43 [10.75, 22,979.64]
variance out-degree (varOutDeg)	variance in the sum of all weighted edges directed *from* the ego across all ego's thread-specific ego networks	3033.14 [12.10, 86,751.27]	8274.66 [64.25, 86,751.27]	1285.97 [12.10, 30,066.46]
variance in-degree centrality (varInDegCent)	variance in in-degree centrality across all egor's thread-specific ego networks	0.14 [0.04, 0.45]	0.14 [0.04, 0.40]	0.14 [0.06, 0.45]
variance out-degree centrality (varOutDegCent)	variance in out-degree centrality across all egor's thread-specific ego networks	0.11 [0.04, 0.37]	0.11 [0.04, 0.29]	0.11 [0.04, 0.37]
median out-in degree ratio (medOutInDegRatio)	median of the out-degree to in-degree ratio across all ego's thread-specific ego networks	0.78 [0.14, 2.5]	0.74 [0.14, 1.30]	0.79 [0.22, 2.5]
total threads	the total number of threads/thread-specific ego networks comprising the ego's data	106.68 [3, 1101]	158.81 [3, 1101]	89.30 [4, 416]

Table 1. Egocentric network features for modeling. *Note.* All features listed were used as predictors for modeling. Seven features were derived by extracting egocentric networks from the thread-agnostic interaction network (order, density, transitivity, BetwCent, InDegCent, OutDegCent, OutInDegRatio) and 10 features were derived by extracting egocentric networks from user-based series of thread-specific interaction networks (varOrder, varDensity, varTransitivity, varBetwCent, varInDeg, varOutDeg, varInDegCent, varOutDegCent, medOutInDegRatio, total threads). InDegCent, OutDegCent, and BetwCent values were normalized. Feature definitions reflect *networkX* implementations where appropriate. For reference, the mean, minimum, and maximum values for each of these features are provided on the entire cohort, as well as HRU and control stratifications.

the receiver operating characteristic (AUC) curve. Significance in performance was assessed through calculation of the 95% confidence area around the AUC as implemented in the *MLeval* package.[43].

Following performance assessment, the model's prediction behaviors were explored using SHapley Additive exPlanations (SHAP) which are based on the Shapley values of game theory.[44,45] Where the original Shapley values represent relative payouts to players in a cooperative game based on their relative contribution, SHAP equates players to features in a prediction task game. As such, SHAP aims to explain the prediction outcome of each sample in the dataset by calculating each feature's marginal contribution to that prediction. The resulting values are therefore understood as the relative magnitudes by which features influence prediction outcomes. The *iBreakdown* and *SHAPforxgboost* packages in R were used to estimate and visualize the SHAP values for the model, respectively. The emergence of the most prominent features and their associated model prediction trends served as the basis with which to address Aim 5. For reference, these steps are illustrated in Fig. 1E.

Data and code availability

Aligning with Aim 1, transparency, accessibility, and reproducibility are important aspects of this work. All Python code used to carry out data preprocessing is available as a commented Jupyter Notebook script in Supplemental File 2. All R code used to carry out machine learning and model introspection is available as a commented R Markdown file in Supplemental File 3. The raw data used to construct all networks in this analysis are available as Supplemental File 4. Egocentric network feature data used to train and validate the machine learning framework are available as Supplemental Files 5 and 6.

Results

Model predictive performance

Using only network-based features to capture online social interaction patterns, the machine learning model attained an AUC of 0.73 ([0.64, 0.82], 95% C.I.) with a corresponding sensitivity of 0.70, specificity of 0.70, F1 score of 0.78, and Cohen's kappa of 0.29 in cross-validation. On the held-out test set, the model attained an AUC of 0.73 ([0.61, 0.85], 95% C.I.) with a corresponding sensitivity of 0.67, specificity of 0.78, F1 score of 0.77, and Cohen's kappa of 0.31. The receiver operating characteristic (ROC) curve for the model's test performance is shown in Fig. 2A. This significant, above-chance ability to predict HRUs from controls represents a model whose underlying predictive tendencies may highlight phenomenologically important links between user behavior and suicidal risk. In other words, the performance of the model is not only generally promising for the utility of social network-based features for suicide risk prediction but also justifies introspection of its underlying decision-making.

Prediction dynamics and influential features

Figure 2B displays the SHAP results from the model and implicates egocentric (i) thread-agnostic network density, (ii) thread-agnostic network transitivity, (iii) variation in thread-specific network in-degree centrality, and (iv) thread-agnostic network in-degree centrality among the most influential features in the prediction of HRUs. Moreover, noticeable trends within these features suggest that users with egocentric networks that are (i) more sparsely connected, (ii) contain a higher proportion of triadic interactions, or reflect a user with (iv) lower and/or (v) less variable in-degree centrality are more likely to be flagged as an HRU by the model. Other notable but less influential trends include a tendency toward HRU prediction with decreasing order and a higher median ratio of out-degree to in-degree. While not highly influential to the model overall, higher counts of thread involvement along with higher betweenness centrality presented as slightly protective of HRU designation.

Discussion

To study STB as it manifests within a modern modality of communication, free from the limitations of censorship found on mainstream platforms, this work leveraged data from an unconventional and unique community—the pro-choice forum, "Sanctioned Suicide." Specifically, the interactive qualities of users were operationalized for the prediction of heightened suicide risk. Presented in detail, the social network-based approach to quantifying patterns of social engagement served to highlight both the overall predictive merit of social network features as well as the potential importance of specific socio-behavioral phenomena as precursors to incredibly dangerous or fatal suicidal behavior. Through the application of derived social network-based features, the study trained, validated, and tested a machine learning model on the data of $N = 192$ forum users which included over 3.2 million unique interactions across three years. The model demonstrated a statistically significant ability to predict highest-risk suicidal behaviors on a held-out test set with an $\text{AUC} = 0.73$ ([0.61, 0.85], 95% C.I.). Additionally, introspection of the model's prediction behavior revealed a few key patterns among the social network-based feature predictors, placing emphasis on graph density and transitivity, as well as the in-degree centrality of users within their local network of interactions.

A well-documented and ubiquitous configuration among widely differing real-world phenomena, including social networks, the "small world" network is a graph characterized by a high degree of local clustering with cliques that are sparsely connected by a small number of edges.[46] These edges link members of the network (and their cliques) through a relatively small number of intermediaries, leading to the network's characteristic short path lengths.[46] The social network feature of transitivity, which is precisely a measure of local clustering, cannot solely be used to satisfy the definition of a small world network; however, high transitivity can suggest dynamics that approach a small world scenario. The SHAP results indicated that thread-agnostic egocentric networks exhibiting higher degrees of transitivity were predicted by the model to belong to HRUs. Relatedly, the structural feature of density, which is the ratio of existing connections to total possible connections, was found to signal for a HRU as its value decreased. Taken together, this suggests that users whose local interaction networks are

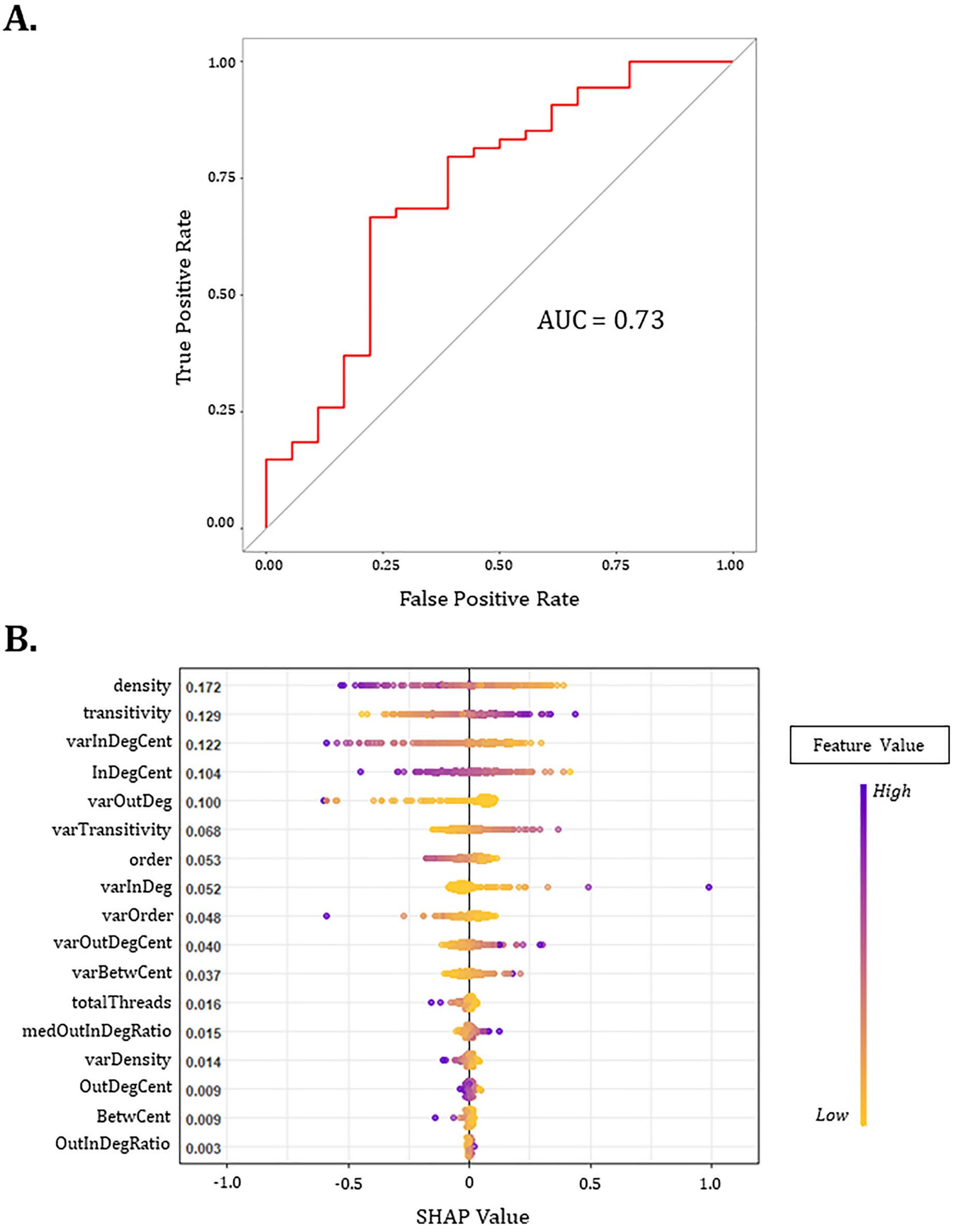

Figure 2. Model Framework Performance and SHAP Introspection. *Note*. (**A**) The Receiver Operating Characteristic (ROC) curve reflects above-chance performance in the binary classification task. The AUC of 0.73 corresponds to an optimal sensitivity of 0.70 and an optimal specificity of 0.70. (**B**) The SHapley Additive ExPlanations (SHAP) plot illustrates the ranked relative importance of features utilized for the prediction of a (completed) suicidal attempt in the machine learning model framework. The average SHAP value for each feature is listed next to its respective variable name. The absolute magnitude of any one average SHAP value does not hold any meaning; however, the relative magnitude of average SHAP values can be compared directly. For example, InDegCent (0.104) was found to be about twice as influential as order (0.053). Each point on the plot represents a sample (192 for each feature row). The color of the point denotes the value of that feature, and the point's position on the x-axis represents the degree to which the value of that feature marginally influenced the model's prediction. Points falling above 0 (right of center) indicate influence toward the model predicting that the outcome will occur, while points falling below 0 (left of center) indicate influence toward the model predicting that the outcome will not occur. Accordingly, model prediction dynamics can be ascertained by observing the patterns of how feature values are distributed along the x-axis.

characterized by small subgroups of users that are not highly integrated with each other (i.e., sparsely distributed) present as among the highest risk in the community. Future research may benefit from studying users on other online communities to see if their STB content or severity is associated with the degree to which their egocentric networks mirror small world architecture. This emphasis on user-centric community embeddings and their consequences for localized information flow may lead to more robust predictors of suicide risk.

Also germane to information flow, the directionality of interactions was found to hold predictive value in the current investigation. The prominence of in-degree, both in terms of centrality and its relative magnitude to out-degree, suggests further focus on the degree to which an individual is the target or receiver of social communication/information. SHAP-based prediction dynamics indicate that users with an overall lower in-degree centrality and a higher ratio of out-degree to in-degree across their thread-based egocentric networks are more likely to be classified as HRUs by the model. This suggests that being a frequent target of social exchanges, posting in threads with higher user engagement, or receiving more information for a lesser relative effort of generation/solicitation, are protective against making a suicidal attempt while being a user on Sanctioned Suicide. The unique moral duality of Sanctioned Suicide—existing simultaneously as both an unconventional "therapeutic" resource that discourages the act of suicide and a place to obtain all information necessary to successfully attempt it—makes it difficult to hypothesize if these protective trends would be echoed across other mainstream online communities. It will, of course, be beneficial for future works to consider the impact of relative "social directionality" on STB risk. However, the question of why inwardly-directed social dialog flow was found to be protective specifically for Sanctioned Suicide may hint at an imbalanced nature within its duality.

Upholding a pro-choice, censorship-free philosophy, Sanctioned Suicide has deservedly come under fire due to public outcry regarding suicides that were believed to have been facilitated through the content hosted within its virtual walls.[35] The forum's content covers a variety of topics and themes, with the most concerning in regards to the solicitation and sharing of suicide methods-related advice.[33] Although this obviates any potentially positive impact this community has on its highly vulnerable members, it is still important to note that individuals come to Sanctioned Suicide for different reasons. While some individuals seek aid in carrying out their plans to end their own lives, others join the community to be heard, understood, and validated due to thoughts, feelings, and opinions that make them pariahs in their day-to-day lives. This allows them to achieve a sense of belonging. Therefore, these users do not necessarily represent imminent attempters as they interact with others, share life histories, and bond through a mutual understanding of ideology. From a social network perspective, this means that there are users in search of camaraderie who may engage more broadly with the community, and users in search of expertise who may engage more selectively with groups of individuals who can provide them with the specific resources they need.

Neither archetype is mutually exclusive nor immune to risk. However, the results of the current analysis suggest that individuals who are involved with more highly connected users (higher egocentric network density), interact directly or indirectly with a larger number of users (higher egocentric network order), post in a larger array of threads (higher total threads), or are more central receivers of information/attention (higher in-degree centrality) are less likely to be HRUs than their peers. Conversely, the association and involvement with smaller cliques of users (high transitivity and low density) or interaction with a less connected or smaller number of users was shown by the model's prediction tendencies to be HRU signatures. This high-risk labeling of social behavior tracks with more topically focused users who engage less with the community at large and interact only with individuals whose expertise or interests are relevant to their plans. Echoing a main pillar of the Interpersonal Psychological Theory of Suicide,[47] the results suggest that thwarted belongingness, expressed here as a social network-based positioning among the fringes of a community, which interestingly *itself* exists at the fringes of broader Internet society, is a key component that drives suicidal action. Moreover, other modern ideation-to-action models of suicide also place emphasis on the importance of social connectedness. For example, the parsimonious Three-Step Theory (3ST) of suicide cites connectedness as one of four factors (alongside pain, hopelessness, and suicide capability) that determine the development of suicidal ideation and the progression to suicidal attempts.[48] More specifically, given the presence of both pain and hopelessness, disrupted connectedness is theorized to play a major role in the escalation from a passive to an active ideation state.[48] Another prominent example is the Integrated Motivational-Volitional (IMV) model of suicide which places connectedness/thwarted belongingness as a key motivational moderator that influences the progression from entrapment to suicidal ideation.[3]

Other (limited) research on the users of suicide-focused forums has yielded complementary findings. One study employed cultural discourse analysis on a pro-recovery suicide forum and found discursive themes relating to issues of placelessness and entrapment as well as to notions of safe places that are defined in part by the presence of empathetic others.[49] This pattern reflects the importance of finding community, both physically and emotionally, and further supports the centrality of social connectedness as a protective factor. Another work focusing on investigating the determinants of loneliness through posts on Reddit, including those contained within the suicide subforum, found that temporal trends in loneliness were associated with future posts in the suicide subforum, particularly that decreased loneliness led to lower likelihoods of suicide posting.[50] Bolstered by both theory and these initial empirical findings, a focus toward a more robust quantification of connectedness may be fruitful toward improving the detection and prediction capabilities of suicide risk models. More broadly, future work will also benefit from a close examination of network features within more mainstream social platforms to see if interactivity shifts which influence an individual's social positioning within their network may indicate a heightened STB risk profile.

Despite the novel insights gleaned from this work, there are several important limitations. First, the analysis only considered the interactive qualities of STB and did not account for the context of these interactions. Integration of text-based features found in forum posts ("what is said") alongside the patterns in how posts are communicated ("to whom it is said") will offer a more complete ability to study how themes of STB evolve and are transmitted as they move through the online community. These considerations should be foci for future work that seeks to build on the preliminary findings of this study. Second, while an online community like Sanctioned Suicide has the capacity to offer novel insight into the nature of STB, it is not entirely representative of other modern communication platforms with much stricter content policies and mechanisms of censorship (e.g., Instagram or Reddit subforums for suicidal ideation). Even though there is value in broadening the collective understanding of STB heterogeneity by studying an understudied public face of STB manifestation, it is important to recognize that the results cannot strictly serve as a generalization of how STB presents online. Third, SHAP is a method that quantifies the marginal influence of a feature on model prediction. Accordingly, multi-feature interaction trends were not considered. Fourth, it is important to note that SHAP results reflect the relative importance of features for a model's decision-making processes and thus cannot directly speak to real-world significance. Thus, SHAP-related findings, while incredibly useful for hypothesis generation, underlining potentially promising emphases in future research, nevertheless cannot be leveraged for hypothesis testing on natural phenomena. Lastly, the strictly anonymous nature of Sanctioned Suicide, while importantly protective of users, precluded the ability to assess the sociodemographic profile, heterogeneity, and global representativeness of the cohort.

STB is pervasive across the World Wide Web, and there is a concerted effort among suicidology and data science researchers to understand, detect, and prevent it. The research presented herein contributes to this effort through the implementation and assessment of an underutilized quantitative approach on a rare, naturalistic corner of the broader internet ecosystem. Taken together, the results promote the idea that future research efforts should incorporate network-based features of social interaction into their exploration and analysis pipelines. Pairing these network-based features with other proven digital markers of STB risk may improve data-driven suicide prevention efforts. As the world's social fora become larger, more integrated, and increasingly virtual, a rich and variable analytical toolkit for deconstructing their many digital incarnations will be a powerful weapon to combat STB and offer timely aid to an ailing and highly vulnerable population.

Data availability

The raw data used to construct all networks in this analysis, the egocentric network feature data used to train and validate the machine learning framework, as well as all (commented) code used to carry out data preprocessing and modeling is available in the supplementary files as labeled.

Received: 2 February 2024; Accepted: 14 August 2024
Published online: 20 August 2024

References

1. World Health Organization. Suicide Fact Sheet. https://www.who.int/news-room/fact-sheets/detail/suicide (2022).
2. American Foundation for Suicide Prevention. Suicide Statistics. https://afsp.org/suicide-statistics/ (2023).
3. O'Connor, R. C. & Kirtley, O. J. The integrated motivational–volitional model of suicidal behaviour. *Phil. Trans. R. Soc. B* **373**, 20170268 (2018).
4. De Choudhury, M. & Kıcıman, E. The Language of Social Support in Social Media and its Effect on Suicidal Ideation Risk. *Proc Int AAAI Conf Weblogs Soc Media* **2017**, 32–41 (2017).
5. Carlyle, K. E., Guidry, J. P. D., Williams, K., Tabaac, A. & Perrin, P. B. Suicide conversations on Instagram™: contagion or caring?. *J. Commun. Healthc.* **11**, 12–18 (2018).
6. Jashinsky, J. *et al.* Tracking suicide risk factors through Twitter in the US. *Crisis* **35**, 51–59 (2014).
7. O'Dea, B., Larsen, M. E., Batterham, P. J., Calear, A. L. & Christensen, H. A linguistic analysis of suicide-related Twitter posts. *Crisis J. Crisis Intervent. Suicide Prevent.* **38**, 319–329 (2017).
8. Renjith, S., Abraham, A., Jyothi, S. B., Chandran, L. & Thomson, J. An ensemble deep learning technique for detecting suicidal ideation from posts in social media platforms. *Journal of King Saud University - Computer and Information Sciences* **34**, 9564–9575 (2022).
9. Roy, A. *et al.* A machine learning approach predicts future risk to suicidal ideation from social media data. *npj Digital Medicine* **3**, 1–12 (2020).
10. Brown, R. C. *et al.* Can acute suicidality be predicted by Instagram data? Results from qualitative and quantitative language analyses. *PLoS ONE* **14**, e0220623 (2019).
11. Kemp, C. G. & Collings, S. C. Hyperlinked suicide: Assessing the prominence and accessibility of suicide websites. *Crisis* **32**, 143–151 (2011).
12. Tadesse, M. M., Lin, H., Xu, B. & Yang, L. Detection of suicide ideation in social media forums using deep learning. *Algorithms* **13**, 7 (2020).
13. Unruh-Dawes, E. L., Smith, L. M., Krug Marks, C. P. & Wells, T. T. Differing relationships between Instagram and Twitter on suicidal thinking: The importance of interpersonal factors. *Soc. Media Soc.* **8**, 20563051221077028 (2022).
14. Sierra, G. *et al.* Suicide risk factors: A language analysis approach in social media. *J. Lang. Soc. Psychol.* **41**, 312–330 (2022).
15. Grant, R. N. *et al.* Automatic extraction of informal topics from online suicidal ideation. *BMC Bioinform.* **19**, 211 (2018).
16. Ren, F., Kang, X. & Quan, C. Examining accumulated emotional traits in suicide blogs with an emotion topic model. *IEEE J. Biomed. Health Inform.* **20**, 1384–1396 (2016).
17. Li, A. *et al.* An analysis of stigma and suicide literacy in responses to suicides broadcast on social media. *Asia-Pacific Psych.* **10**, e12314 (2018).
18. Deas, N. *et al.* I just want to matter: Examining the role of anti-mattering in online suicide support communities using natural language processing. *Comput. Human Behav.* **139**, 107499 (2023).
19. Spitzer, E. G., Crosby, E. S. & Witte, T. K. Looking through a filtered lens: Negative social comparison on social media and suicidal ideation among young adults. *Psychol. Popular Media* **12**, 69–76 (2023).
20. de Beurs, D. *et al.* A network perspective on suicidal behavior: understanding suicidality as a complex system. *Suicide Life-Threatening Behav.* **51**, 115–126 (2021).
21. Holman, M. S. & Williams, M. N. Suicide risk and protective factors: a network approach. *Arch. Suicide Res.* **26**, 137–154 (2022).
22. De Choudhury, M., Gamon, M., Counts, S. & Horvitz, E. Predicting depression via social media. *Proc. Int. AAAI Conf. Web Social Med.* **7**, 128–137 (2013).
23. Dutta, S. & De Choudhury, M. Characterizing Anxiety Disorders with Online Social and Interactional Networks. in *HCI International 2020 – Late Breaking Papers: Interaction, Knowledge and Social Media* (eds. Stephanidis, C. et al.) 249–264 (Springer International Publishing, Cham, 2020). https://doi.org/10.1007/978-3-030-60152-2_20.

24. Abuhassan, M. *et al.* Classification of Twitter users with eating disorder engagement: learning from the biographies. *Comput. Human Behav.* **140**, 107519 (2023).
25. Beeney, J. E., Hallquist, M. N., Clifton, A. D., Lazarus, S. A. & Pilkonis, P. A. Social disadvantage and borderline personality disorder: A study of social networks. *Person. Disorders: Theory, Res. Treat.* **9**, 62–72 (2018).
26. Colombo, G. B., Burnap, P., Hodorog, A. & Scourfield, J. Analysing the connectivity and communication of suicidal users on twitter. *Comput. Commun.* **73**, 291–300 (2016).
27. Wang, Z., Yu, G. & Tian, X. Exploring behavior of people with suicidal ideation in a Chinese online suicidal community. *Int. J. Environ. Res. Public Health* **16**, 54 (2019).
28. Myers West, S. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media Soc.* **20**, 4366–4383 (2018).
29. Cobbe, J. Algorithmic censorship by social platforms: Power and resistance. *Philosophy Technol.* **34**, 739–766 (2021).
30. Gillespie, T. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* (Yale University Press, 2018).
31. Das, S. & Kramer, A. Self-censorship on facebook. *Proc. Int. AAAI Conf. Web Social Media* **7**, 120–127 (2021).
32. Powers, E., Koliska, M. & Guha, P. "Shouting matches and echo chambers": perceived identity threats and political self-censorship on social media. *Int. J. Commun.* **13**, 20 (2019).
33. Lekkas, D. & Jacobson, N. C. The hidden depths of suicidal discourse: Network analysis and natural language processing unmask uncensored expression. *Digital Health* **9**, 20552076231210710 (2023).
34. Richardson, L. Beautiful soup documentation. (2007).
35. Barbaro, M. Kids Are Dying. How Are These Sites Still Allowed?
36. Rafaeli, S., Ravid, G. & Soroka, V. De-lurking in Virtual Communities: A Social Communication Network Approach to Measuring the Effects of Social and Cultural Capital. in *Proceedings of the Hawaii International Conference on System Sciences* (Honolulu, HI, 2004). https://doi.org/10.1109/HICSS.2004.1265478.
37. da Silva, L. F. C., Barbosa, M. W. & Gomes, R. R. Measuring participation in distance education online discussion forums using social network analysis. *J. Assoc. Inform. Sci. Technol.* **70**, 140–150 (2019).
38. Hecking, T., Chounta, I.-A. & Hoppe, H. U. Investigating social and semantic user roles in MOOC discussion forums. in *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16* 198–207 (ACM Press, Edinburgh, United Kingdom, 2016). https://doi.org/10.1145/2883851.2883924.
39. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. in *Proceedings of the 7th Python in Science Conference (SciPy2008)* 11–15 (Pasadena, CA USA, 2008).
40. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **28**, 1–26 (2008).
41. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
42. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *Jair* **16**, 321–357 (2002).
43. John, C. R. MLeval: Machine Learning Model Evaluation. (2020).
44. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874 [cs, stat]* (2017).
45. Shapley, L. S. A value for n-person games. *Contribut. Theory Games* **2**, 307–317 (1953).
46. Humphries, M. D. & Gurney, K. Network 'small-world-ness': a quantitative method for determining canonical network equivalence. *PLOS ONE* **3**, e0002051 (2008).
47. Joiner, T. E. *Why People Die By Suicide*. (Harvard University Press, 2005).
48. Klonsky, E. D. & May, A. M. The Three-Step Theory (3ST): A new theory of suicide rooted in the "ideation-to-action" framework. *Int. J. Cognitive Therapy* **8**, 114–129 (2015).
49. Alvarez, M. "Life is about trying to find a better place to live": Discourses of dwelling in a pro-recovery suicide forum. *Qual. Res. Med. Healthc.* **6**, 10437 (2022).
50. Mazuz, K. & Yom-Tov, E. Analyzing trends of loneliness through large-scale analysis of social media postings: observational study. *JMIR Ment. Health* **7**, e17188 (2020).

Acknowledgements

This work was supported by an institutional grant from the National Institute on Drug Abuse (NIDA-5P30DA02992610).

Author contributions

The following reflects individual contributions per the nomenclature of the Contributor Roles Taxonomy (CRediT): DL: Conceptualization, Data Curation, Methodology, Software, Investigation, Formal analysis, Visualization, Writing—original draft, Writing—review & editing. NCJ: Methodology, Writing—review & editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-70282-0.

Correspondence and requests for materials should be addressed to D.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.