This is an Accepted Manuscript of an article published by Science Direct in Computers in Human Behavior on 12/18/2024, available online: https://doi.org/10.1016/j.chb.2024.108542

To cite this document: Akar, E. (2025). Exploring the impact of social network structures on toxicity in online mental health communities. *Computers in Human Behavior*, 165, https://doi.org/10.1016/j.chb.2024.108542

Exploring the Impact of Social Network Structures on Toxicity in Online Mental Health Communities

Abstract

This study examines how structural social capital influences online toxicity within mental health communities. Using social network analysis and regression models, we analyze both direct and interaction effects of network centralities—degree, closeness, eigenvector, and betweenness—on toxicity in the r/MentalHealth subreddit. From a dataset of 90,626 posts, we constructed a network of 7,562 users interconnected through 12,699,868 relationships. Our findings highlight the nuanced relationship between network positioning and toxic behavior. Users with a higher degree centrality, reflecting broad connectivity, exhibit lower toxicity levels, indicating that well-connected individuals contribute positively to community dynamics. Conversely, higher eigenvector, closeness, and betweenness centralities are associated with increased toxicity, suggesting that influential users, those centrally located, and those acting as bridges between network segments are more likely to engage in toxic behavior. Interaction effects further reveal complexities: for instance, well-connected and influential users tend to mitigate toxicity, while those who combine influence with proximity amplify it. These insights underscore the dual role of network structures in moderating or exacerbating harmful interactions. The study offers actionable strategies for fostering healthier online environments by leveraging network centralities to design targeted interventions and reduce toxicity in online mental health communities.

Keywords: online toxicity, mental health, social capital, online communities, network centralities

1. Introduction

Toxicity, defined as language that is rude, disrespectful, or harmful, has become a pervasive issue in online communities. It can discourage participation, erode trust, and harm vulnerable users (Maleki et al., 2022). Online mental health communities, while providing essential support and resources for those grappling with mental illness, are not immune to this phenomenon. These platforms serve as safe spaces for users to share experiences and seek help, but the anonymity and accessibility that make them appeal also create an environment where toxic behaviors can flourish (De Choudhury & De, 2014). For individuals already facing challenges such as depression, anxiety, or suicidal thoughts, exposure to toxicity can exacerbate feelings of isolation, distress, and hopelessness.

The prevalence and impact of toxicity in online spaces are well-documented. Surveys indicate that 41% of Americans have personally experienced online harassment (Vogels, 2021). Transparency reports from Meta estimate that 0.14–0.15% of all views on Facebook in 2021 were of toxic posts. Twitter reports that it removed roughly two million accounts in the second half of 2020 due to hate and harassment (Kumar et al., 2023). A journalistic account by The Guardian in 2013 discussed how a 12-year-old girl committed suicide after being targeted for cyberbullying (Obadimu et al., 2021). In 2012, Charlotte Dawson, who at one time hosted the "Next Top Model" TV program in Australia, committed suicide after being targeted with malicious online comments. In mental health communities, where users seek empathy and understanding, toxic interactions can directly undermine the platform's purpose, leading to reduced engagement and the loss of diverse voices. A harmful comment can deter users from sharing their experiences or seeking advice, limiting the community's ability to provide meaningful support.

Previous research has considered predicting toxicity and developing detection models (Anjum & Katarya, 2022; Chong & Kwak, 2022; Leite et al., 2020; Urbaniak et al., 2022), exploring the role of user engagement (Aleksandric et al., 2022; Rajadesingan et al., 2020;

Saveski et al., 2021), characterizing user types (Kumar et al., 2023; Mall et al., 2020), and investigating network structures (DiCicco et al., 2023; Quattrociocchi et al., 2022). On the other hand, structural social capital is a critical dimension of social capital theory, yet its role in influencing toxic behavior in online communities has not been adequately explored. Structural social capital pertains to the patterns and configurations of relationships within a community, which enable or constrain interactions. This study specifically addresses the underexplored intersection of structural social capital and online toxicity. By examining centralities (structural positions) such as degree, closeness, eigenvector, and betweenness, this study illuminates how network positioning influences toxicity. Structural social capital, which focuses on the configuration of connections and information flow within a community, offers a framework for understanding how network dynamics contribute to or mitigate toxicity. For example, users with high connectivity or influence may play pivotal roles in either propagating toxic behavior or fostering positive interactions. This study addresses the gap by investigating whether structural social capital predicts online toxicity in the r/MentalHealth subreddit, a specialized platform where users discuss sensitive mental health topics like depression, anxiety, and suicidal ideation. The subreddit's focused and supportive environment makes exploring how network structures influence toxic behavior ideal.

2. Online Toxicity

Toxicology, the study of poisons and their effects, aims to understand the lethal doses of substances (Carillo & Marsan, 2016). It has evolved into a diverse field, encompassing chemistry, pharmacology, and other sciences. Just as toxicology examines how substances affect living organisms, this research analyzes how harmful interactions influence the health and dynamics of online communities. Online communities are "social aggregations that emerge from the Net when enough people carry on those public discussions long enough, with sufficient human feeling, to form webs of personal relationships in cyberspace" (Rheingold, 1993, pp. 6-7).

Building on Rheingold's definition of online communities as social aggregations, these digital spaces have evolved to serve as forums for general interaction and critical platforms for more specialized and impactful exchanges. Specifically, as De Choudhury and De (2014) highlight online communities have become invaluable for individuals dealing with mental health issues. On the other hand, online toxicity, including harassment and hate speech, undermines the health of online communities by eroding trust and safety, which discourages participation and dialogue. This can reduce diversity and vibrancy as members leave or avoid joining. Conversely, communities that effectively tackle toxicity through moderation and clear guidelines create more inclusive and engaging spaces, enhancing resilience and member interaction.

Prior research has extensively examined online toxicity from different aspects. A key area of focus has been the development of algorithms and models aimed at accurately identifying different types of toxicity in online environments. Anjum and Katarya (2022) tested four popular machine learning methods—logistic regression, support vector machine (SVM), multinomial naïve Bayes, and random forest—on a dataset of tweets to identify toxic content. They discovered that SVM learning performed the best in accuracy in most cases and consistently provided the most relevant results, outperforming the other classifiers. Urbaniak et al. (2022) researched whether the toxicity level of usernames on Reddit could predict the user's likelihood of exhibiting toxic behavior online. Analyzing data from 329,000 users, they used algorithms and Bayesian statistics to find that users with toxic usernames tend to post more toxic content than those with neutral usernames. The study found that the more active the user, the more significant the difference in toxicity levels, indicating a link between toxic usernames and online behavior. Chong and Kwak (2022) analyzed 643,913 comments from the Singapore subreddit using 20 text classification transformers from Hugging Face's model repository. Their research aimed to identify English toxic comments by focusing on the specific words that differentiate toxic comments from non-toxic ones. Almerekhi et al. (2020) investigated "toxicity triggers" in over 104 million Reddit comments, defining them as non-toxic comments that provoke toxic replies.

They identified specific content and textual features indicative of these triggers. They developed a neural network model to predict them, demonstrating that specific discussion content and context shifts can signal the onset of toxic responses.

Carton et al. (2020) emphasized the value of integrating quantitative methods with qualitative research to predict online toxicity, acknowledging its complex nature. This combined approach enhances the analysis of toxic behaviors by merging statistical accuracy with detailed contextual understanding. However, L. Li et al. (2023) pointed out a significant challenge in this methodology: creating these comprehensive models subjects annotators to harmful and offensive content, posing ethical concerns and necessitating considerable time and resources. For example, they investigated the role of generative AI models in overcoming this problem using ChatGPT and compared its performance with MTurker annotations related to harmful content. They showed that ChatGPT exhibited approximately 80% accuracy and provided the same response more than 90% of the time regarding reliability and consistency. Moreover, Wijesiriwardene et al. (2020) studied toxic social media interactions among high school students using the "Adolescents on Twitter" dataset, which includes 16,901 tweets, images, emojis, and metadata. They emphasized that context is critical to identifying toxicity, as individual tweets may not indicate toxic behavior. Their approach involved analyzing interactions between a source and a target to understand better the nature of their relationship, such as friendship, which could suggest sarcasm rather than toxicity. The study categorized interactions into toxic, non-toxic, or unclear and introduced the ALONE dataset, employing a lexicon to classify offensive language across various categories. This nuanced approach aimed to uncover complex toxic dynamics by considering the broader context of interactions. Leite et al. (2020) aimed to improve the detection of toxic comments on Twitter by analyzing 21,000 Brazilian-Portuguese tweets. They utilized bidirectional encoder representations from transformer models to classify tweets into various categories of toxicity, including non-toxic. To enhance model training, human annotators labeled 1,500 tweets with specific toxic categories such as LGBTQ+phobia, obscene, insult, racism,

misogyny, and xenophobia or marked them as non-toxic when appropriate. This study underscored the importance of creating models that can accurately identify and differentiate between multiple types of toxicity.

Furthermore, exploring how online toxicity, user engagement, and user profiles interact has become crucial for understanding digital interactions. This research is critical to grasping how toxic content affects user behavior and developing strategies for healthier online communities. Salehabadi et al. (2022) discovered in their analysis of over 85,300 tweets that conversations with higher visibility tend to include more toxic replies, especially if the toxicity starts with the primary tweet or its first reply. This pattern sets a precedent for the conversation's tone, suggesting that initial toxic responses can lead to a chain reaction of negativity. Aleksandric et al. (2022) further explored the behavior of toxicity victims in 79.8k Twitter conversations, noting that victims often respond with avoidance, retaliation, defensive measures, or attempts at negotiation. Notably, these victims are more active in engaging in conversations, replying toxically, and distancing themselves from instigators than other users. Saveski et al. (2021) analyzed 1.18 million Twitter conversations involving major news outlets and 2018 US midterm election candidates to study the link between conversation structure and toxicity. They looked at factors like content toxicity, the structure of replies to threads, and users' social networks to predict toxic behavior. Key findings include that toxicity often comes from users with a low to moderate level of toxicity, toxic replies are more common from users not socially connected to the original poster, and toxic conversations usually feature more extensive and more complex replies to threads but less interconnected social networks. Rajadesingan et al. (2020) explored how newcomers adapt to the norms of political communities on Reddit, particularly regarding toxicity. Their research revealed that newcomers often adopt the community's prevailing toxicity levels before posting their first comment. This suggests that individuals observe and assimilate the behavioral norms from the community's interactions right from the start instead of gradually adjusting their behavior over time after becoming active

members.

Mall et al. (2020) analyzed 4 million Reddit comments to categorize users based on their toxicity patterns: steady, fickle-minded (switching between toxic and non-toxic comments), pacified (becoming less toxic over time), and radicalized (increasingly toxic). They found that fluctuating between toxic and non-toxic comments was the most common behavior. Building on this, Kumar et al. (2023) conducted a study on 929K Reddit accounts over 18 months, identifying that these accounts produced over 14 million toxic comments, including insults and threats. They discovered that 55.2% of these accounts engage in threads with toxic comments, with 3.1% of accounts being abusive but responsible for 33.3% of all comments. They further classified users into occasional, moderate, and serial abusers based on the volume of toxic comments posted. These studies together provide insight into the prevalence and patterns of toxic behavior on Reddit, highlighting the diverse nature of online toxicity and the significant impact of a small number of highly toxic users.

A pivotal tool in previous studies has been social network analysis. This methodology allows researchers to dissect and understand the nuanced relationships and interactions that contribute to the health of digital communities. For example, Obadimu et al. (2021), Quattrociocchi et al. (2022), and DiCicco et al. (2023) used social network analysis to understand the dynamics of toxicity and misinformation within online communities, particularly during the COVID-19 pandemic. Obadimu et al. (2021) focused on identifying the top toxic users in the network, which led to the creation of experiments simulating the impact of removing these users on toxicity in the network. They found that a high degree of segregation among commenters revealed that users mostly react to toxic comments with an equal or high degree of toxicity. The most significant toxicity reduction was achieved by removing users based on the computed toxicity score of their comments. Quattrociocchi et al. (2022) delved into the spread of misinformation and the emergence of toxic conversations on Twitter by analyzing the structure and characteristics of conversation networks. They computed several network metrics such as

size, depth, assortativity, toxicity ratio, and average toxicity distance. DiCicco et al. (2023) compared toxicity and misinformation dissemination between Twitter and Parler. They identified user communities with highly toxic content, including a misinformation echo chamber within the Parler network. They identified significant bridge nodes that spread toxic COVID-19 vaccine misinformation throughout the Parler network.

3. Theoretical Background

Social capital, a concept that emerged from community research, underscores the significance of networks of strong personal relationships that evolve, fostering trust, cooperation, and collective action within communities (Sainaghi & Baggio, 2014). At the heart of social capital theory, as Bourdieu (1986) posited, is the idea that networks of relationships are valuable assets for executing social affairs. They offer members access to collectively owned capital, which grants them credit in various senses. Nahapiet and Ghoshal (1998) further elaborate, defining social capital as the accumulation of actual and potential resources accessible through and derived from the network of relationships held by an individual or a social unit. Social capital is multi-dimensional, comprising structural, cognitive, and relational aspects. Structural social capital relates to the patterns of connections, configurations, and participatory mechanisms within a community, focusing on the pathways through which information is exchanged and the way individuals are interconnected. Cognitive social capital encompasses the shared norms, values, attitudes, and beliefs that promote mutual understanding among members of the community, essentially representing the collective consciousness. Lastly, relational social capital zeroes in on the quality of the relationships within the community, highlighting the importance of trust, mutual respect, and obligations among its members. This multifaceted approach to understanding social capital reveals its critical role in facilitating effective social interactions and cooperation within various communities.

Structural social capital fundamentally concerns the configuration of a network's structure, as Sainaghi and Baggio (2014) highlighted. It encompasses the network's ties, their

arrangements, and the participatory mechanisms within a community, focusing on the patterns of information flow and connections among individuals. Borgatti and Ofem (2010) emphasize that relationships within a network are interconnected, forming a network of paths that enable indirect influences among actors. This network of paths is likened to conduits facilitating the flow of information, resources, etc., thereby shaping a structure where each actor holds a specific position. This structural positioning is crucial as it partly dictates an actor's opportunities and limitations within the network. By their position, some actors might experience a more significant influx of resources or information, whereas others might control the distribution of these flows, potentially extracting benefits in the process. Thus, an actor's location within the network's structure can significantly affect their outcomes. Additionally, the overall structure of the network is a crucial determinant of the network's behavior and dynamics, influencing what occurs within the entire network.

Researchers have established a variety of critical indicators to quantify structural social capital with network centralities. These centralities reveal the roles and impact of individuals within the network, highlighting their position and influence. The primary measures of network centrality—degree, closeness, eigenvector, and betweenness—serve as tools to evaluate the positions and influences of individuals within a network. Each centrality measure reveals different aspects of influence and connectivity within a network, highlighting how individuals can impact community dynamics positively or negatively. Their roles, whether as information disseminators, influencers, or gatekeepers, are crucial in shaping the social fabric of online communities, emphasizing the importance of responsible behavior and moderation to foster positive interactions and norms.

As described by Mislove (2009), degree centrality assesses an individual's connectivity within the network, highlighting those with extensive connections to other members. It refers to the number of direct connections a user has within a network. In other words, it counts how many other users are directly linked to the user in question. Figure 1 depicts a network of sixteen

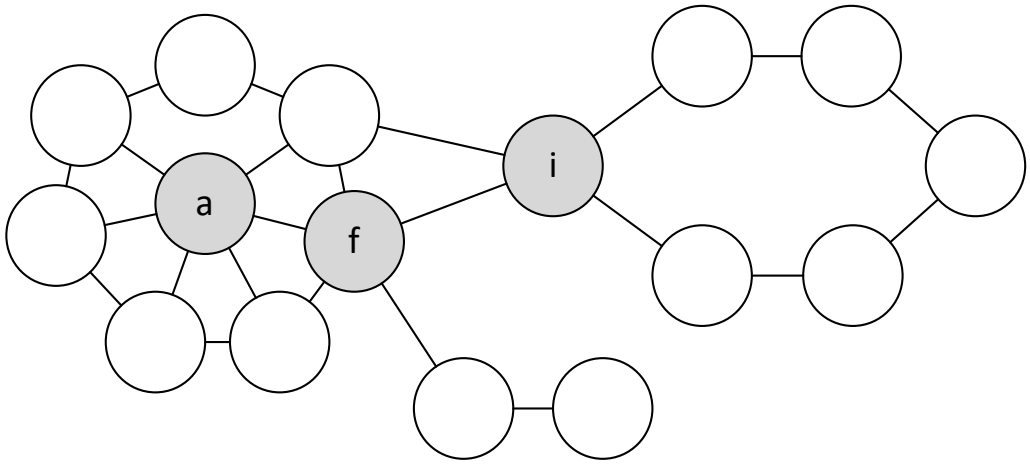
users and relationships among them. According to Figure 1, users' degree centralities in the depicted network range between one and seven. The "user a" has the highest degree centrality because seven links are connected to this user. Individuals with high centrality are central users, receiving and disseminating vast information, knowledge, and resources (E. Li et al., 2013). Users with a high degree centrality are more visible and connected to a broader set of individuals, fostering accountability and reducing their likelihood of engaging in toxic behaviors. Theoretically, this aligns with social responsibility, where individuals with extensive connections are subject to greater scrutiny and social feedback, incentivizing prosocial behavior. High degree centrality users may also receive more positive reinforcement and validation from their networks, reducing motivations to act out negatively.

Eigenvector centrality, explained by Abbasi et al. (2011), assigns users scores based on their connections to other highly scored users, indicating an individual's influence within the broader network structure. This centrality extends the concept of degree centrality by considering the number of connections a user has and the importance of those connections. For example, as shown in Figure 1, the "user a" is not just connected to many others but specifically to other influential users. Theoretically, this can be a double-edged sword. While influential users can champion positive behaviors, their elevated position amplifies their ability to disseminate toxic norms if they engage in harmful interactions. This aligns with the diffusion of innovation theory, where opinion leaders play pivotal roles in shaping community-wide behaviors, for better or worse. Thus, influential nodes with high eigenvector centrality may inadvertently or intentionally normalize toxicity within their spheres of influence.

According to Catanese et al. (2012), closeness centrality gauges how near an individual is to everyone within the network, emphasizing the efficiency of information access and distribution. Closeness centrality measures how close a user is to all other users in the network, calculated as the inverse of the average distance from the node to all other users. This metric suggests that individuals closer to the rest of the network can quickly affect the community's

culture based on their behavior, for better or worse. As displayed in Figure 1, users' closeness centralities range between 0.018 and 0.035. The "user f" has the closeness centrality with a value of 0.035. This user can quickly interact with others because they are, on average, fewer steps away from every other user in the network. Theoretically, proximity to others increases opportunities for both positive and negative influence. In the context of toxicity, high closeness centrality users can amplify harmful behaviors through swift dissemination, shaping the community's tone. This aligns with theories of contagion in social networks, where behaviors spread more effectively when intermediaries are minimized, underscoring the importance of monitoring centrally positioned users.

Betweenness centrality explores how vital a user is at bridging the gap between other users in the network (Wasserman & Faust, 1994). Users with high betweenness centrality act as bridges within the network, controlling the flow of information between separate clusters (Baek & Kim, 2015; Freeman, 1978). They are crucial for the diffusion of influence across different parts of the network, potentially gatekeeping or facilitating the spread of ideas across diverse groups. As illustrated in Figure 1, users' betweenness centrality ranges between 0 and 52. The "user i", who has the highest betweenness centrality, often facilitates communication between two groups that would otherwise be unconnected. Theoretically, users with high betweenness centrality are capable of facilitating or obstructing interactions between network segments. In online communities, such users can either diffuse toxic behaviors across clusters or act as buffers, preventing the spread of negativity. This dual role aligns with structural hole theory, which posits that intermediaries derive power from their ability to bridge gaps, enabling them to influence network-wide dynamics, including the prevalence of toxicity.



a – highest degree and eigenvector centralities

f – highest closeness centrality

i – highest betweenness centrality

Figure 1. A depiction of network centralities – degree, closeness, betweenness, and eigenvector

In summary, we examine the consequences of occupying different structural positions on contribution to online toxicity and hypothesize that:

H1: *Structural social capital - degree, closeness, betweenness, and eigenvector centralities- impacts users' contribution to online toxicity.*

On the other hand, network theory also emphasizes the interplay of centralities, where users may simultaneously occupy multiple structural roles. For example: a user with a high degree of centrality may also have high eigenvector centrality, combining broad connectivity with influential relationships. These overlapping roles can interact in ways that amplify or mitigate their effects on toxicity. This theoretical lens highlights the complexity of network dynamics, where centralities interact to shape individual and collective behaviors in nuanced ways. By examining these interactions, we can better understand users' dual or synergistic roles

in promoting or preventing harmful behaviors. In this regard, we hypothesize that:

H2: *Interactions between centrality measures influence users' contribution to online toxicity.*

4. Methodology

4.1. Data Collection

We used Communalytic to collect public social media data from Reddit (Gruzd & Mai, 2022). Communalytic is an easy-to-use social media data collector designed to collect publicly available data from Reddit, Telegram, YouTube, X (known as Twitter), Bluesky, and Mastodon. We selected the r/MentalHealth subreddit as a data source. The r/MentalHealth subreddit was chosen due to its highly focused and supportive environment, which aligns with the study's objective of examining the intersection of network structures and toxicity in sensitive online spaces. This subreddit serves as a central hub for users seeking emotional support and sharing experiences related to mental health challenges, such as depression, anxiety, and suicidal ideation. Its active user base and thematic consistency provide a controlled and rich dataset to investigate the role of structural social capital in toxicity dynamics. Additionally, the community's well-defined scope and high engagement levels allow for precise analysis of network structures, ensuring robust insights into the mechanisms driving toxic behavior in online support environments. Although the demographics of this subreddit are unknown, generally, Reddit users are mostly American (49.9%), male (67%), and young, with 22% being 18-19 years old and 14% being 30-49 years old (Low et al., 2020)

During the first quarter of 2023, we fetched 90,626 posts from the r/MentalHealth subreddit. In the dataset, each post included an ID, date, author, title, text, and type. An ID is a unique identifier assigned to each post in the dataset. It ensures that each post can be uniquely identified and referenced. Each post has a date on which it was created or published. The author is the name or identifier of the person or entity who created or published the post. The title is the heading or headline of the post, summarizing its main topic or theme. The text is the main body

of the post, containing the detailed content or message. Redditors can comment on posts and respond in a conversation tree of comments. So, the post type can be a submission, comment, or reply.

4.2. Data Pre-Processing

We focused on submissions, the initial posts that started discussions. Our process involved extracting these submissions and refining the dataset through qualitative analysis and automated preprocessing using WordStat 8.0.

We removed duplicate and system-generated submissions. The r/MentalHealth subreddit is predominantly English-speaking, with most posts written in English, so we removed non-English submissions. We excluded other languages to eliminate any variability in the analysis because the nuances of toxicity detection can vary between languages. We also removed deleted and removed posts. A post's content might be deleted or removed by the user, moderators, or Reddit's automated systems. In such cases, posts involve tags of "[deleted]" or "[removed]." We removed posts with no authors because authors were required to create a network. We deleted posts soliciting research participation or survey responses.

To ensure completeness, we merged the title and text of each submission, as they often form a continuous narrative (Chakravorti et al., 2018). We removed numbers, punctuation, extra white spaces, non-alphabetical characters, non-word emoticons, system-generated words or characters, and hyperlinks. We corrected misspellings and replaced words spelled informally with formal spelling, for example, "ive" to "I have" and "ppl" to "people" (Leung & Khalvati, 2022). We deleted common stop words such as "the," "a," and "and." We lemmatized words to analyze the inflected forms of a word as a group together (Ding et al. 2020). For example, we treated "cars," "car's," and "cars'" as "car." We converted all the words to lowercase. We eliminated words appearing less than five times to reduce noisy words. We deleted submissions exceeding 3,000 characters to fit the analysis criteria set by the Perspective API (Ding et al.,

2020; Leung & Khalvati, 2022).

After that, we assessed the toxicity score of each submission using the Perspective API, accessed via R's perspective package. This tool was developed by Google's Project Jigsaw and Counter Abuse Technology teams (Hosseini et al., 2017). The API employs machine learning models to assess the toxicity of input text, where toxicity is characterized as a comment that is rude, disrespectful, or unreasonable and likely to cause someone to withdraw from the conversation.

The Perspective API assigns each submission a score ranging from 0 to 1, where higher scores indicate a greater likelihood of toxic content. While many studies apply a standard threshold, typically 0.5-0.7, to distinguish between toxic and non-toxic content (Nogara et al., 2024), we chose not to exclude submissions with toxicity scores below this threshold. Even submissions with lower toxicity scores can contribute to a negative interaction climate and may influence network dynamics through subtle toxic behaviors. By retaining all submissions, except ones with 0 toxicity score, we aimed to capture the full spectrum of toxic interactions within the community. This approach allowed for a more nuanced analysis of the interplay between network centralities and toxicity contributions, ensuring we did not overlook the potential cumulative effects of mild toxicity. Therefore, including all posts provided a more comprehensive view of toxicity's impact on social structures in the community. All these steps resulted in a dataset of 8,580 submissions from 7,562 unique users.

4.3. Data Analysis

To construct the network for this study, we utilized author-topic relationships rather than direct user-to-user interactions. This methodological choice was guided by the premise that topic-level connections better capture collective community dynamics. In online spaces, users often engage indirectly through shared themes and topics, contributing to a collective discourse rather than forming direct interpersonal ties. By focusing on author-topic relationships, we aimed

to identify patterns of toxic behavior aggregated around shared thematic interests. This approach offers a nuanced understanding of how toxicity propagates within the broader community context, highlighting toxic hotspots that may not be visible through direct user-level interactions.

First, we identified topics using WordStat 8.0, which applied non-negative matrix factorization (NMF) to create a word-by-document frequency matrix. This method, chosen for its efficiency with large data sets, extracted topics based on words exceeding a specific frequency criterion. However, choosing the correct number of topics is challenging due to the lack of a standardized method (Németh et al., 2021). Using the elbow method on coherence scores, we identified where the gain in coherence sharply declined, indicating an optimal balance between topic granularity and comprehensiveness. We assessed the topics for interpretability qualitatively and adjusted the number of topics. We labeled the topics with the most frequent terms identified by NMF.

We categorized each submission into predefined topic(s) using QDA Miner 8.0. This process generated a matrix detailing submission IDs, authors, and their topic affiliations.

We created a weighted two-mode network linking authors to the topics they discussed. This approach enabled us to quantify the intensity of each author-topic interaction by counting the occurrences of an author engaging with a specific topic, which served as the basis for the relationship weight within the network, illustrating the strength of each connection (Haythornthwaite, 1996). This network included 8,577 submissions, 7,562 unique users, and 17,597 author-topic relationships.

We then used the R igraph package to transform this weighted two-mode network into a weighted one-mode network, focusing on authors. Authors were connected if they contributed to the same topic(s), with the relationship weights reflecting the frequency of shared topic engagement (Borgatti et al., 2009). This transformation resulted in a network of 7,562 users interconnected by 12,699,868 relationships, providing a comprehensive view of user-level dynamics within the subreddit. We calculated structural social capital by computing each

author's network centralities, including degree, eigenvector, closeness, and betweenness in this network.

We aggregated toxicity scores at the topic level to assess the general level of toxicity associated with specific topics. In addition to topic-level aggregation, we gathered toxicity scores at the author levels to evaluate an author's overall contribution to online toxicity. We calculated the average toxicity score for all of their submissions. This resulted in a single toxicity score for each author, representing the general toxicity level they contribute to the community. Finally, we gathered all dependent and independent variables with main and interaction effects, as shown in Table 1, to perform a regression analysis.

Table 1. Summary of variables: Dependent and independent factors in the regression model

Variables	**Description**
Dependent Variable	
• Online Toxicity	Harmful user behavior disrupting digital interactions.
Independent Variables	
Main Effects	
• Degree	User's connectivity (number of direct connections in the network).
• Eigenvector	User's influence through connections to other highly connected users.
• Closeness	User's proximity to all others in the network.
• Betweenness	The user's role is to be a bridge between other users in the network.
Centrality Interactions	
• Degree × Eigenvector	Explores the combined effect of connectivity and influence on toxicity.
• Degree × Closeness	Tests whether being well-connected and proximate amplifies or mitigates toxicity.
• Degree × Betweenness	Assesses whether connectivity and bridging roles jointly shape behavior.
• Eigenvector × Closeness	Examines the impact of influential users who are also close to others.
• Eigenvector × Betweenness	Investigates the effect of influential users who also act as bridges.
• Closeness × Betweenness	Tests how proximity and bridging roles combine to influence toxicity.

5. Results

According to Table 2, our analysis identified twelve distinct topics. Discussions about personal emotions and feelings were highly prevalent. General conversations about mental well-being and mental illness were also widespread. Additionally, discussions regarding symptoms, treatments, coping mechanisms, and support for depression and anxiety were common. In contrast, discussions about memory functions, memory recall, and memory improvement techniques were less frequent. Similarly, conversations about gaming experiences and their impact on mental health were less common. Discussions about sleep patterns, quality, and disorders exhibited a moderate frequency.

In terms of online toxicity, discussions about suicidal behavior exhibited the highest levels, indicating a significant presence of harmful or damaging language. Conversations about parenting also had high toxicity levels, reflecting solid emotions and differing opinions on parenting practices. Similarly, discussions on romantic relationships and friendships showed elevated toxicity. Conversations about video gaming, education, personal emotions, and sleeping exhibited moderate levels of toxicity. In contrast, discussions about depression and anxiety, panic attacks, mental well-being, and professional life demonstrated lower levels of online toxicity.

Figure 2 displays a bubble chart. The chart shows the relationship between the frequency of topics being discussed and the online toxicity levels for each topic. The size of each bubble corresponds to the frequency of topics. The chart shows that topics like suicidal behavior, parenting, and interpersonal connections not only dominate discussions but also exhibit higher toxicity, highlighting areas where more supportive and less harmful communication is needed. Conversely, topics with lower frequencies and toxicity levels, such as cognitive recall and video gaming, suggest less controversial and more positive interactions.

Table 3 shows critical centrality and online toxicity measures across 7,562 users to understand the entire community's structure and the prevalence of harmful language. Degree

centrality, with a mean of 3,358.86 (SD = 1,485.70), ranged from 292 to 7,528, indicating substantial variation in node connectivity. Eigenvector centrality values ranged from 0.0172 to 1.000, with a mean of 0.2491 (SD = 0.1586), reflecting varying levels of user influence. Closeness centrality values were tightly clustered around a mean of 0.000080343 (SD = 0.000005212), suggesting similar average shortest path lengths across users. Betweenness centrality exhibited considerable variability, ranging from 21.892 to 11,425.210, with a mean of 2,469.45 (SD = 2,299.61), highlighting diverse roles in network communication. Online toxicity levels varied widely, with a mean of 0.2318 (SD = 0.1486) and values ranging from 0.0125 to 0.9446, indicating a generally low to moderate presence of harmful language in discussions.

According to Table 4, The ANOVA results for the regression model provide insight into the significance and explanatory power of the independent variables (network centralities) in predicting online toxicity. The total sum of squares (SS = 7561.000) reflects the overall variability in toxicity levels across the dataset. In contrast, the regression sum of squares (SS = 113.000) represents the portion of this variation explained by the centrality measures included in the model. The residual sum of squares (SS = 7448.000) indicates the amount of variation that remains unexplained by the model. With 10 degrees of freedom for the regression and 7551 for the residual, the mean square values for the regression (MS = 11.300) and residual (MS = 0.986) indicate the average variation explained by the model per predictor and the average unexplained variation, respectively. The F-statistic (F = 11.456) tests whether the regression model provides a significantly better fit to the data than a model with no predictors, and the highly significant p-value ($p < 0.001$) indicates that the network centrality measures collectively explain a significant portion of the variance in online toxicity.

This is an Accepted Manuscript of an article published by Science Direct in Computers in Human Behavior on 12/18/2024, available online: https://doi.org/10.1016/j.chb.2024.108542

To cite this document: Akar, E. (2025). Exploring the impact of social network structures on toxicity in online mental health communities. *Computers in Human Behavior*, 165, https://doi.org/10.1016/j.chb.2024.108542

Table 2. Identified topics and their toxicity levels

Topic	Explanation	Top Words/Phrases	Average Online Toxicity	Frequency
Suicidal Behavior	Discussions about suicide prevention, awareness, support, and coping strategies.	Suicide, Suicidal, Attempt, Kill, Die, Death, Commit Suicide	27.69%	1,551
Parenting	Conversations focused on parenting experiences, challenges, advice, and support.	Dad, Mom, Father, Mother, Parent, Sister, Child, Brother, Abuse, Childhood, Kid, Family	26.35%	1,622
Interpersonal Connections	Discussions related to romantic relationships and friendships.	Friend, Girl, Relationship, Date, Guy, Meet, Girl Friend	26.14%	1,604
Video Gaming	Conversations about gaming experiences, recommendations, and their impact on mental health.	Game, Video, Play, Watch, Video Games, Playing Video Games	24.76%	441
Education	Conversations about memories, experiences, and challenges during the school years.	College, School, High School	24.52%	1,166
Emotional Experience	Conversations about personal emotions and feelings.	Hate, F*ck, Sh*t, Feel, Happy	24.29%	3,327
Sleeping	Discussions about sleep patterns, quality, and disorders.	Sleep, Bed, Hour, Night, Home, Wake, Room	24.01%	867
Cognitive Recall	Discussions related to memory functions, memory recall, and memory improvement techniques.	Memory, Retention, Recall, Remember, Forget	23.44%	303
Professional Life	Discussions related to work, job satisfaction, career choices, professional development, and work-life balance.	Job, Work, Life	22.74%	1,437
Mental Well-being	General conversations about mental well-being and mental illness.	Mental, Health, Mental Illness, Mental Health Issues	21.63%	2,695
Panic Attack	Discussions about symptoms, treatments, coping mechanisms, and support related to panic attacks.	Panic, Attack, Panic Attack, Panic Disorder	20.24%	1,568
Depression & Anxiety	Discussions about symptoms, treatments, coping mechanisms, and support for depression & anxiety.	Depression, Anxiety, ADHD, Disorder, Medication, OCD, Symptom, Psychiatrist	19.31%	2,134

This is an Accepted Manuscript of an article published by Science Direct in Computers in Human Behavior on 12/18/2024, available online: https://doi.org/10.1016/j.chb.2024.108542

To cite this document: Akar, E. (2025). Exploring the impact of social network structures on toxicity in online mental health communities. *Computers in Human Behavior*, 165, https://doi.org/10.1016/j.chb.2024.108542

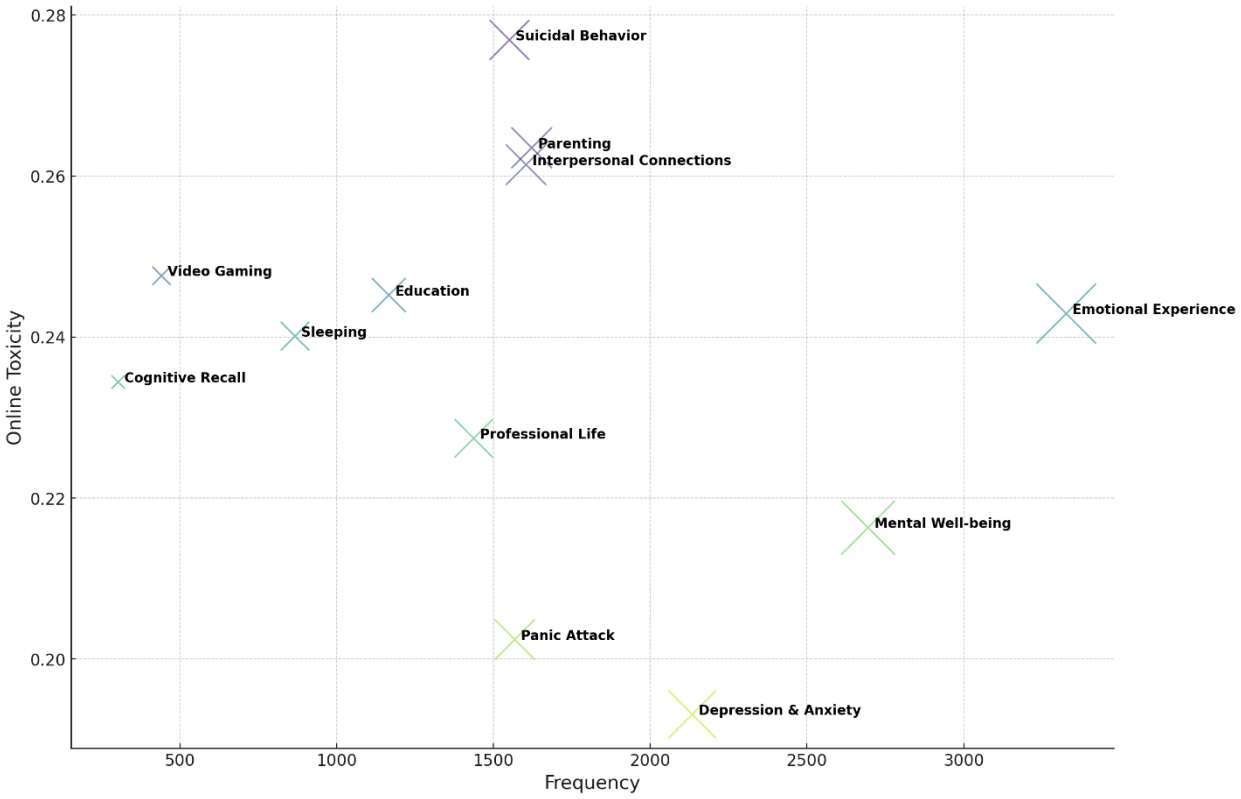


Figure 1. The relationship between the frequency of topics and their online toxicity levels

Table 3. Descriptive statistics

Descriptive Statistics					
	N	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
Degree centrality	7,562	292	7,528	3,358.860	1,485.702
Eigenvector centrality	7,562	.0172	1.000	.2491	.1586
Closeness centrality	7,562	.0000674	.0000897	.000080343	.000005212
Betweenness centrality	7,562	21.892	11,425.210	2,469.450	2,299.612
Online toxicity	7,562	.0125	.9446	.2318	.1486

Table 4. ANOVA results for the impact of network centralities on toxicity levels

ANOVA					
	Sum of Squares	**df**	**Mean Square**	**F**	**Sig.**
Regression	113.000	10	11.300	11.456	<0.001
Residual	7448.000	7551	0.986		
Total	7561.000	7561			

Table 5 shows the hypothesis testing results. Degree centrality has a strong negative impact on online toxicity ($\beta = -1.132$, $p < 0.01$), indicating that a higher degree centrality is associated with a lower contribution to online toxicity. Eigenvector centrality shows a strong positive relationship with a contribution to online toxicity ($\beta = 0.824$, $p < 0.05$), suggesting that authors with more decisive influence contributed to increased toxicity. Closeness centrality positively correlates with the contribution to online toxicity ($\beta = 0.382$, $p < 0.01$), meaning that authors closer to all other authors in the network tended to exhibit higher toxicity. Betweenness centrality negatively affected contribution to online toxicity ($\beta = 0.136$, $p < 0.01$), indicating that authors acting as bridges in the network were linked to lower levels of toxic behavior. We also explored the combined effects of multiple centrality measures on toxicity.

Users who are both well-connected and influential (degree x eigenvector, $\beta = -0.313$, $p < 0.001$) exhibit significantly lower toxicity. A significant positive interaction suggests that users who are both influential and proximate to others (eigenvector x closeness, $\beta = 0.482$, $p < 0.001$) contribute more to toxicity. A significant negative interaction indicates that users who are both proximate and act as bridges contribute less to toxicity (closeness x betweenness, $\beta = -0.350$, $p < 0.001$). On the other hand, there were not any significant interactions between degree and betweenness centralities, degree and closeness centralities, and eigenvector and betweenness centralities. As a result, we supported H1 and H2, and structural social capital plays a role in the

This is an Accepted Manuscript of an article published by Science Direct in Computers in Human Behavior on 12/18/2024, available online: https://doi.org/10.1016/j.chb.2024.108542

To cite this document: Akar, E. (2025). Exploring the impact of social network structures on toxicity in online mental health communities. *Computers in Human Behavior*, 165, https://doi.org/10.1016/j.chb.2024.108542

contribution to online toxicity.

Table 5. Impact of network centralities on online toxicity: Regression results

	Coefficients β	**t**	**Sig.**
Independent Variables			
Main Effects			
• Degree	-1.132	-2.481	<0.01
• Eigenvector	0.824	2.298	<0.05
• Closeness	0.382	2.284	<0.05
• Betweenness	0.136	3.020	<0.01
Centrality Interactions			
• Degree × Eigenvector	-0.313	-5.065	<0.001
• Degree × Closeness	0.042	0.482	Insignificant
• Degree × Betweenness	-0.054	-0.258	Insignificant
• Eigenvector × Closeness	0.482	3.866	<0.001
• Eigenvector × Betweenness	0.176	1.235	Insignificant
• Closeness × Betweenness	-0.350	-3.585	<0.001

6. Discussion

This study significantly contributes to understanding of online toxicity and structural social capital by examining how network centralities and their interactions shape toxic behavior in online mental health communities. This research highlights how users' positions within a network influence their propensity to engage in or mitigate toxicity by leveraging degree, eigenvector, closeness, and betweenness centralities. The findings underscore structural social capital's theoretical and practical importance in moderating harmful behaviors and fostering positive interactions.

Users with a high degree centrality, representing broad connectivity, are pivotal in fostering positive interactions. A significant negative relationship between degree centrality and toxicity ($\beta = -1.132$, $p < 0.01$) suggests that well-connected users tend to exhibit lower levels of harmful behavior. These findings align with social capital theory, emphasizing that robust networks enhance accountability and encourage prosocial behavior (Akar & Dalgic, 2018; Van Emmerik et al., 2011). Network theory further supports this, positing that individuals with extensive social ties are more likely to receive support and validation, reducing their likelihood of toxic engagement (Rieger et al., 2018).

In contrast, users with high eigenvector centrality, who are influential due to their connections to other influential individuals, are more likely to propagate toxic behavior ($\beta = 0.824$, $p < 0.05$). This dual nature of influence highlights a key theoretical insight: central figures within a network can either reinforce positive norms or amplify harmful interactions, depending on their behavior (Akar & Dalgic, 2018). The significant negative interaction between degree and eigenvector centralities ($\beta = -0.313$, $p < 0.001$) offers an important nuance. When users are both well-connected and influential, their combined attributes mitigate toxicity. This suggests that such users act as stabilizing forces, leveraging their network position to promote accountability and reduce harmful behaviors.

Closeness centrality, measuring proximity to others in the network, presents another layer of complexity. Users with high closeness centrality show a moderate positive relationship with toxicity ($\beta = 0.382$, $p < 0.05$), likely due to their ability to disseminate information, including toxic content, rapidly. When combined with high eigenvector centrality, this effect is amplified ($\beta = 0.482$, $p < 0.001$), as these users' proximity and influence synergistically exacerbate the spread of negative norms. This underscores the need for monitoring such users, whose central positions allow them to disproportionately shape the tone of interactions within the community.

Betweenness centrality, reflecting users' roles as bridges between disconnected groups, also impacts toxicity. A small but significant positive relationship ($\beta = 0.136$, $p < 0.01$) indicates

that these users, who control the flow of information between clusters, can propagate harmful behavior. However, the interaction between closeness and betweenness centralities reveals a mitigating effect ($\beta = -0.350$, $p < 0.001$). Users who are both close to others and act as bridges reduce toxicity, likely by facilitating positive interactions and diffusing tensions. This finding aligns with structural hole theory (Burt, 1992), which highlights the potential of intermediaries to foster cooperation and diffuse negative dynamics across groups.

These results demonstrate the multifaceted role of network centralities in shaping online toxicity. Degree centrality fosters positive engagement, while eigenvector and closeness centralities highlight the risks of influence and rapid toxic content dissemination, respectively. Betweenness centrality reveals the dual potential of bridging users to propagate or mitigate toxicity. The interactions between centralities underscore the complexity of these relationships, emphasizing that users' combined roles within a network can significantly alter their contributions to community health.

Importantly, these findings demonstrate that structural social capital is not merely an academic construct but a practical framework for understanding and addressing toxicity in digital environments. By elucidating the interplay between network structures and behavior, this study bridges a crucial gap in the literature, providing actionable insights for community management.

Users with a high degree centrality, characterized by extensive connections, play a critical role in reducing toxicity. Platforms can amplify their positive influence by designating them as "community ambassadors" or equipping them with tools to model constructive norms across their networks. Their broad connectivity positions them as key agents in fostering healthier interactions. Conversely, users with high eigenvector centrality are influential due to their connections to other prominent users, are more likely to propagate toxic. However, when users are both well-connected and influential, their combined attributes mitigate toxicity. This underscores the importance of fostering collaboration among these users to stabilize community norms. Platforms can engage influential users through incentives or recognition programs that

align their influence with positive behavior.

Users with high closeness centrality, who are centrally positioned and can disseminate information quickly, exhibit a moderate positive relationship with toxicity. Their ability to rapidly spread both positive and negative behavior highlights the need for targeted engagement. Providing these users with resources and guidance can prevent the amplification of toxicity. The interaction between closeness and eigenvector centralities shows that influential users with central positions pose heightened risks, necessitating close monitoring and proactive interventions.

Betweenness centrality, reflecting users who bridge disconnected groups, also affects toxicity dynamics. While such users show a small positive relationship with toxicity, their unique position offers opportunities to mitigate harmful behavior. The negative interaction between closeness and betweenness centralities suggests that users who bridge groups while maintaining central proximity help reduce toxicity. Platforms can strategically empower these users as mediators, leveraging their roles to diffuse tensions and promote positive exchanges.

To translate these findings into practice, platforms can integrate network centrality metrics into automated moderation tools. For example, users with high eigenvector centrality, who connect with other influential users, may serve as early indicators of emerging toxic norms. Automated moderation tools incorporating these metrics can prioritize interventions, such as flagging potentially harmful content or recommending targeted engagement for at-risk users. For instance, a system could monitor real-time network dynamics to identify and address toxicity hotspots or empower community moderators with actionable insights to focus on critical areas. Such tools can also suggest tailored community guidelines for high-toxicity topics, ensuring that moderation approaches align with the thematic and structural characteristics of the community. By combining network centrality insights with advanced machine learning algorithms for toxicity detection, platforms can create a more adaptive and responsive moderation ecosystem. This integration not only fosters healthier online communities but also supports moderators in

maintaining the balance between open discourse and respectful engagement.

Additionally, the choice to construct topic-level networks rather than direct user-to-user networks highlights the importance of considering indirect pathways of interaction in online communities. This approach recognizes that much of the toxicity in forums like r/MentalHealth emerges not from isolated interpersonal exchanges but from collective engagement around shared themes. By analyzing author-topic relationships, we capture the ripple effects of toxic discourse within thematic clusters, offering a more comprehensive understanding of how harmful behavior influences the broader community. This finding underscores the dual nature of topics as both resources for support and potential catalysts for toxicity, particularly when emotionally sensitive themes dominate the conversation. These results suggest that interventions aimed at reducing toxicity should also consider the thematic contexts in which harmful interactions occur. Moderating high-toxicity topics and fostering positive norms within these thematic spaces could significantly improve the overall health of the community.

7. Conclusion

This study explained the significant influence of structural social capital on online toxicity within mental health communities, particularly the r/MentalHealth subreddit. By analyzing the structural social capital—degree, closeness, eigenvector, and betweenness centralities—we uncovered how the structural social capital impacted the prevalence of toxic behavior. Our findings indicated that users with a higher degree centrality, who were well-connected within the community, tended to exhibit lower toxicity levels. Conversely, users with higher eigenvector centrality, closeness centrality, and betweenness centrality, who were more influential, centrally located, or acted as bridges within the network, were more likely to contribute to toxic interactions. The interaction effects highlighted the nuanced roles of network centralities in shaping toxicity. High degree and eigenvector centralities mitigate toxicity, while closeness and eigenvector centralities amplified it, necessitating targeted interventions. The interaction of closeness and betweenness also showed the potential of intermediaries to reduce

harm, emphasizing the need for nuanced moderation strategies.

These insights highlight the dual nature of network influence, where connectivity can either mitigate or exacerbate harmful behavior depending on the user's position in the network. The study underscores the importance of considering network structures in designing and managing online communities. Interventions targeting key influencers and central figures within these networks may be particularly effective in reducing toxicity and promoting healthier, more supportive environments.

However, several limitations should be acknowledged. While this study focuses on the r/MentalHealth subreddit, the insights gained have broader implications for understanding toxicity in other online communities. In communities with similar support-oriented goals, such as forums for addiction recovery or parenting advice, the dynamics of network centralities may exhibit comparable patterns, with well-connected users fostering positive interactions and influential users potentially amplifying toxicity. By focusing on these core dynamics, future research can further test and refine the generalizability of these results, providing a comprehensive understanding of how structural social capital shapes toxicity in such communities. On the other hand, in competitive or adversarial environments like gaming or political forums, the interplay of network structures and toxicity may manifest differently, influenced by the community's norms and content dynamics. Exploring these variations across diverse contexts can reveal universal and community-specific mechanisms, enriching our understanding of how structural social capital shapes toxicity in digital spaces. These broader considerations highlight the importance of tailoring interventions to the unique characteristics of each online community.

While this study primarily investigates structural social capital and toxicity, it is crucial to acknowledge the potential influence of moderation practices and community guidelines in shaping the behaviors observed. Effective moderation, which includes enforcing clear and consistent community guidelines, can mitigate toxic interactions and foster a supportive

environment. Community moderators play a pivotal role in implementing policies, identifying harmful behaviors, and mediating conflicts. Integrating moderation practices with insights from network centrality metrics can enhance their effectiveness. For instance, targeting users with high eigenvector centrality who propagate toxicity or leveraging users with high degree centrality as community ambassadors can serve as practical interventions. Additionally, moderation strategies tailored to the unique thematic contexts of discussions (e.g., suicide prevention versus gaming) can address toxicity more precisely, acknowledging the nuances of each topic. Future research could expand on this work by exploring how different moderation styles interact with network centralities to influence toxicity levels. This would provide actionable insights for platforms aiming to balance user freedom with community safety.

Additionally, our cross-sectional analysis captures a snapshot in time rather than longitudinal changes in network dynamics and toxicity levels. Future research should incorporate longitudinal data to understand better how these relationships evolve. Lastly, while the regression model is statistically significant, the residual sum of squares, as in Table 5, suggests that a large proportion of the variance remains unexplained, highlighting the potential influence of other factors on toxicity levels in the community.

8. Ethical Considerations

This study was conducted with a strong emphasis on ethical considerations to ensure the integrity and respect for the individuals involved in the online mental health community. Data collection was limited to publicly available posts on the r/MentalHealth subreddit, ensuring no private information was accessed or used without consent. The anonymity of the users was maintained throughout the research process by not collecting or analyzing any personal identifiers that could link the data back to specific individuals.

Given the sensitive nature of discussions related to mental health, special care was taken to handle the data respectfully and responsibly. Although necessary for the study, toxicity

assessments were performed using automated tools to minimize researcher exposure to potentially harmful content. This approach also helped maintain objectivity and consistency in the toxicity evaluations.

Furthermore, the study acknowledges the potential psychological impact that toxic behavior can have on individuals within online communities. As such, the research aims to provide insights that could lead to interventions and strategies to reduce toxicity, ultimately fostering a safer and more supportive online environment for those seeking mental health support.

9. Acknowledgement

While preparing this work, the author used AI-assisted technologies to improve the readability of the text. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

This study is funded through an internal grant from the University of Wisconsin–Eau Claire College of Business.

10. References

Abbasi, A., Altmann, J., & Hossain, L. (2011). Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *J Informetr, 5*(4), 594–607. https://doi.org/10.1016/j.joi.2011.05.007

Akar, E., & Dalgic, T. (2018). Understanding online consumers' purchase intentions: A contribution from social network theory. *Behav Inf Technol, 37*(5), 473–487. https://doi.org/10.1080/0144929X.2018.1456563

Aleksandric, A., Roy, S. S., & Nilizadeh, S. (2022). Twitter users' behavioral response to toxic replies. arXiv. http://arxiv.org/abs/2210.13420

Almerekhi, H., Kwak, H., Salminen, J., & Jansen, B. J. (2020). Are these comments triggering? Predicting triggers of toxicity in online discussions. *WWW '20: Proceedings of the web conference* 2020 (pp.3033–3040). Association for Computing Machinery. https://doi.org/10.1145/3366423.3380074

Anjum, A., & Katarya, R. (2022). Analysis of online toxicity detection using machine learning approaches. In G. Sanyal, C. M. Travieso-González, S. Awasthi, C. M. A. Pinto, & B. R. Purushothama (Eds.), *International conference on artificial intelligence and sustainable engineering* (Vol. 836, pp. 381–392). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-8542-2_29

Baek, S. I., & Kim, Y. M. (2015). Longitudinal analysis of online community dynamics. *Ind Manag Data Sys*, *115*(4), 661–677. https://doi.org/10.1108/IMDS-09-2014-0266

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Sci, 323*(5916), 892–895. https://doi.org/10.1126/science.1165821

Borgatti, S. P., & Ofem, B. (2010). Overview: Social network theory and analysis. In A. J. Daly (Ed.), *Social network theory and educational change*. Harvard Education Press.

Bourdieu, P. (1986). The forms of capital. In J. G. Richardson (Ed.), *Handbook theory and practice for the sociology of education* (pp. 241–258). Greenwood.

Burt, R. S. (1992). Structural holes: The social structure of competition. Cambridge, MA: Harvard University Press.

Carillo, K. D. A., & Marsan, J. (2016). "The dose makes the poison"—Exploring the toxicity phenomenon in online communities. In P. Ågerfalk, N. Levina, S. S. Kien (Eds.), *Proceedings of the international conference on information systems - Digital innovation at the crossroads*, ICIS 2016, Dublin, Ireland, December 11-14. Association for Information Systems.

Carton, S., Mei, Q., & Resnick, P. (2020). Feature-based explanations don't help people detect misclassifications of online toxicity. *Proceedings of the international AAAI conference on web and social media*, *14*(1), 95–106. https://doi.org/10.1609/icwsm.v14i1.7282

Catanese, S., De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2012). Extraction and analysis of Facebook friendship relations. In A. Abraham (Ed.), *Computational social networks* (pp. 291–324). Springer. https://doi.org/10.1007/978-1-4471-4054-2_12

Chakravorti, D., Law, K., Gemmell, J., & Raicu, D. (2018). Detecting and characterizing trends in online mental health discussions. *2018 IEEE International conference on data mining workshops (ICDMW)* (pp. 697-706). Singapore. https://doi.org/10.1109/ICDMW.2018.00107

Chong, Y. Y., & Kwak, H. (2022). Understanding toxicity triggers on Reddit in the context of Singapore. *Proceedings of the international AAAI conference on web and social media*, *16*(1), 1383–1387. https://doi.org/10.1609/icwsm.v16i1.19392

De Choudhury, M., & De, S. (2014). Mental health discourse on Reddit: Self-disclosure, social support, and anonymity. *Proceedings of the international aaai conference on web and social media*, *8*(1), 71–80. https://doi.org/10.1609/icwsm.v8i1.14526

DiCicco, K., Noor, N. B., Yousefi, N., Maleki, M., Spann, B., & Agarwal, N. (2023). Toxicity and networks of COVID-19 discourse communities: A tale of two social media platforms. *ROMCIR 2023: The 3rd workshop on reducing online misinformation through credible information retrieval*, held as part of ECIR 2023: The 45th European Conference on Information Retrieval. April 2-6, Dublin, Ireland.

This is an Accepted Manuscript of an article published by Science Direct in Computers in Human Behavior on 12/18/2024, available online: https://doi.org/10.1016/j.chb.2024.108542

To cite this document: Akar, E. (2025). Exploring the impact of social network structures on toxicity in online mental health communities. *Computers in Human Behavior*, 165, https://doi.org/10.1016/j.chb.2024.108542

Ding, K., Shu, K., Li, Y., Bhattacharjee, A., & Liu, H. (2020). Challenges in combating COVID-19 infodemic—Data, tools, and ethics. arXiv. http://arxiv.org/abs/2005.13691

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239. https://doi.org/10.1016/0378-8733(78)90021-7

Gruzd, A., & Mai, P. (2022). Communalytic: A research tool for studying online communities and online discourse. https://communalytic.com/

Haythornthwaite, C. (1996). Social network analysis: An approach and technique for the study of information exchange. *Libr Inf Sci Res, 18*(4), 323–342. https://doi.org/10.1016/S0740-8188(96)90003-1

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving Google's Perspective API built for detecting toxic comments. arXiv. http://arxiv.org/abs/1702.08138

Kumar, D., Hancock, J., Thomas, K., & Durumeric, Z. (2023). Understanding the behaviors of toxic accounts on Reddit. *Proceedings of the ACM web conference 2023* (pp. 2797–2807). Association for Computing Machinery. https://doi.org/10.1145/3543507.3583522

Leite, J. A., Silva, D. F., Bontcheva, K., & Scarton, C. (2020). Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. arXiv. http://arxiv.org/abs/2010.04543

Leung, Y. T., & Khalvati, F. (2022). Exploring COVID-19–related stressors: Topic modeling study [Preprint]. *J Med Internet Res, 24*(7), e37142. https://doi.org/10.2196/37142

Li, E. Y., Liao, C. H., & Yen, H. R. (2013). Co-authorship networks and research impact: A social capital perspective. *Res Policy, 42*(9), 1515–1530. https://doi.org/10.1016/j.respol.2013.06.012

Li, L., Fan, L., Atreja, S., & Hemphill, L. (2023). "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. https://doi.org/10.48550/ARXIV.2304.10619

Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (2020). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19: Observational study. *Journal of Medical Internet Research*, *22*(10), e22635. https://doi.org/10.2196/22635

Maleki, M., Arani, M., Mead, E., Kready, J., & Agarwal, N. (2022). Applying an epidemiological model to evaluate the propagation of toxicity related to COVID-19 on Twitter. *Proceedings of the 55th Hawaii international conference on system sciences* (pp. 3275-3284). https://doi.org/10.24251/HICSS.2022.401

Mall, R., Nagpal, M., Salminen, J., Almerekhi, H., Jung, S.-G., & Jansen, B. J. (2020). Four types of toxic people: Characterizing online users' toxicity over time. *Proceedings of the 11th Nordic conference on human-computer interaction: Shaping experiences, shaping society* (pp. 1–11). Association for Computing Machinery. https://doi.org/10.1145/3419249.3420142

Mislove, A. E. (2009). *Online social networks: Measurement, analysis, and applications to distributed information systems*. [Thesis] Rice University. https://www.khoury.northeastern.edu/home/amislove/publications/SocialNetworks-Thesis.pdf

Nahapiet, J., & Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational advantage. *Acad Manag Rev, 23*(2), 242–266. https://doi.org/10.2307/259373

Németh, R., Sik, D., & Katona, E. (2021). The asymmetries of the biopsychosocial model of depression in lay discourses—Topic modelling online depression forums. *SSM Popul Health, 14*. https://doi.org/10.1016/j.ssmph.2021.100785

Nogara, G., Francesco, P., Stefano, C., Luca, L., Petter, T., & Silvia, G. (2024). Toxic bias: perspective API misreads German as more toxic. arXiv. http://arxiv.org/abs/2312.12651

Obadimu, A., Khaund, T., Mead, E., Marcoux, T., & Agarwal, N. (2021). Developing a socio-computational approach to examine toxicity propagation and regulation in COVID-19 discourse on YouTube. *Inf Process Manag*, *58*(5). https://doi.org/10.1016/j.ipm.2021.102660

Quattrociocchi, A., Etta, G., Avalle, M., Cinelli, M., & Quattrociocchi, W. (2022). Reliability of news and toxicity in Twitter conversations. In F. Hopfgartner, K. Jaidka, P. Mayr, J. Jose, & J. Breitsohl (Eds.), *Social informatics* (Vol. 13618, pp. 245–256). Springer International Publishing. https://doi.org/10.1007/978-3-031-19097-1_15

Rajadesingan, A., Resnick, P., & Budak, C. (2020). Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. *Proceedings of the international AAAI conference on web and social media, 14(1), 557-568*. https://doi.org/10.1609/icwsm.v14i1.7323

Rheingold, H. (1993). *The virtual community: Homesteading on the electronic frontier*. Addison-Wesley.

Rieger, E., Sellbom, M., Murray, K., & Caterson, I. (2018). Measuring social support for healthy eating and physical activity in obesity. *Br J Health Psychol, 23*(4), 1021–1039. https://doi.org/10.1111/bjhp.12336

Sainaghi, R., & Baggio, R. (2014). Structural social capital and hotel performance: Is there a link? *Int J Hosp Manag, 37*, 99–110. https://doi.org/10.1016/j.ijhm.2013.11.004

Salehabadi, N., Groggel, A., Singhal, M., Roy, S. S., & Nilizadeh, S. (2022). User engagement and the toxicity of Tweets. arXiv. http://arxiv.org/abs/2211.03856

Saveski, M., Roy, B., & Roy, D. (2021). The structure of toxic conversations on Twitter. *WWW '21: Proceedings of the Web Conference 2021* (pp. 1086–1097). Association for Computing Machinery. https://doi.org/10.1145/3442381.3449861

Urbaniak, R., Tempska, P., Dowgiałło, M., Ptaszyński, M., Fortuna, M., Marcińczuk, M., Piesiewicz, J., Leliwa, G., Soliwoda, K., Dziublewska, I., Sulzhytskaya, N., Karnicka, A., Skrzek, P., Karbowska, P., Brochocki, M., & Wroczyński, M. (2022). Namespotting: Username toxicity and actual toxic behavior on Reddit. *Comput Hum Behav, 136*, 107371. https://doi.org/10.1016/j.chb.2022.107371

Van Emmerik, H., Jawahar, I. M., Schreurs, B., & De Cuyper, N. (2011). Social capital, team efficacy and team potency: The mediating role of team learning behaviors. *Career Dev Intern, 16*(1), 82–99. https://doi.org/10.1108/13620431111107829

This is an Accepted Manuscript of an article published by Science Direct in Computers in Human Behavior on 12/18/2024, available online: https://doi.org/10.1016/j.chb.2024.108542

To cite this document: Akar, E. (2025). Exploring the impact of social network structures on toxicity in online mental health communities. *Computers in Human Behavior*, 165, https://doi.org/10.1016/j.chb.2024.108542

Vogels, E. A. (2021, January 13). *The state of online harassment*. Pew Research Center. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511815478

Wijesiriwardene, T., Inan, H., Kursuncu, U., Gaur, M., Shalin, V. L., Thirunarayan, K., Sheth, A., & Arpinar, I. B. (2020). ALONE: A dataset for toxic behavior among adolescents on Twitter. In: S. Aref, K Bontcheva, M. Braghieri, F. Dignum, F. Giannotti, F. Grisolia, & D. Pedreschi (Eds.), *Social Informatics. SocInfo 2020. Lecture Notes in Computer Science* (Vol. 12467). Springer. https://doi.org/10.1007/978-3-030-60975-7_31

World Health Organization. (2022, June 16). *World mental health report: Transforming mental health for all*. World Health Organization. https://www.who.int/publications/i/item/9789240049338