

Network Analysis of Cyberbullying Interactions on Instagram

Satyaki Sikdar¹, Manuel Sandoval¹, Taylor Hales¹, Chloe Kilroy¹, Maddie Juarez¹, Tyler Rosario¹,
Juan J. Rosendo¹, Deborah L. Hall², Yasin N. Silva¹

¹Loyola University Chicago

²Arizona State University

{ssikdar, msandovalmadrigal, thales, ckilroy, mjuarez4, trosario, jrosendo, ysilva1}@luc.edu, d.hall@asu.edu

Abstract

Cyberbullying continues to grow in prevalence and its impact is felt by thousands worldwide. This study seeks a network science perspective on cyberbullying interaction patterns on the popular photo and video-sharing platform, Instagram. Using an annotated cyberbullying dataset containing over 400 Instagram posts (sessions), we outline a set of heuristics for building Session Graphs, where nodes represent users and their cyberbullying role, and edges represent their exchanged communications via comments. Over these graphs, we compute the Bully Score, a measure of the net malice introduced by bullies as they attack victims (attacks minus pushback), and the Victim Score, a measure of the net support victims receive from their defenders (support minus attacks). Utilizing small subgraph (motif) enumeration, our analysis uncovers the most common interaction patterns over all cyberbullying sessions. We also explore the prevalence of specific motif patterns across different ranges of Bully and Victim Scores. We find that a majority of cyberbullying sessions have negative Victim Scores (attacks outweighing support), while the Bully Score distribution has a slight positive skew (attacks outweighing pushback). We also observe that while bullies are the most common role in motifs, defenders are also consistently present. This suggests that bullying mitigation is a recurring structural feature of many interactions. To the best of our knowledge, this is the first study to explore this granular scale of network interactions using human annotations at the session and comment levels on Instagram.

Code & Dataset — <https://github.com/ysilva/cb-motifs>

Introduction

Social interactions define the human experience. Positive interactions help foster connections, build communities, and enrich our day-to-day lives. There is also a darker underbelly of communication—bullying, hate speech, name calling, and other means of verbal abuse—that are rife with harmful intent. These behaviors present differently based on the communication medium and the user demographics. In this work, we present insight into the negative and often harmful interactions found on online social media platforms from *cyberbullying*. The definition of cyberbullying

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

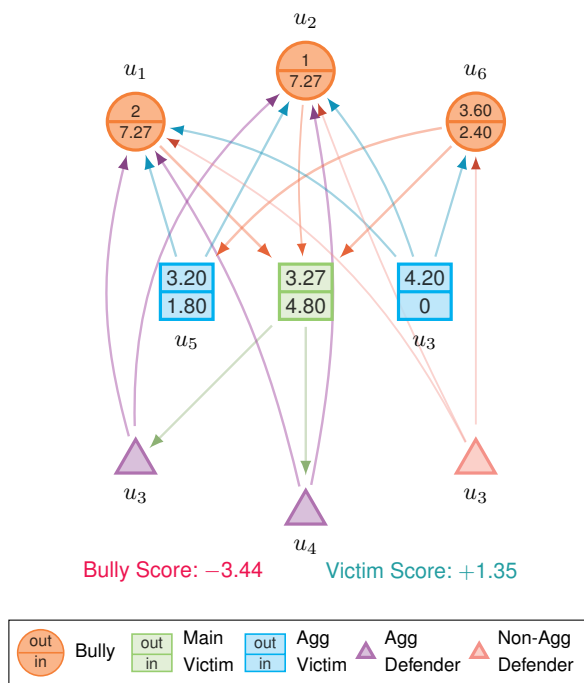


Figure 1: Session Graph constructed from comment annotations in Tab. 2. Usernames are shown adjacent to each node except for the main victim. Certain nodes are presented with their weighted out (top) and in (bottom) degrees. Victim nodes are denoted as rectangles, and Bully nodes as circles. The Bully and Victim scores are the difference between the weighted out- and in-degrees (top – bottom) averaged over all Bully and Victim nodes, respectively.

continues to be debated, though it is often recognized as an aggressive, intentional act carried out by an individual or a group of individuals using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend themselves (Zych and Farrington 2026). For example, a 2022 survey conducted by the Pew Research Center found that 46% of teens ages 13 to 17 in the US experienced at least one form of cyberbullying (Vogels 2022).

The diversity of user demographics along with the multifaceted nature of social media platforms adds complex-

ity when attempting to understand and analyze cyberbullying (Festl, Scharkow, and Quandt 2015; Singh et al. 2024; Chen and Zolkepli 2025). Furthermore, finding large quantities of cyberbullying data for analysis is challenging due to the limited availability of labeled datasets, the high cost of labeling at both session- and comment-level, and strict query limitations imposed by most social media platforms¹. Given these constraints, we worked with a previously collected Instagram dataset with cyberbullying annotations at the session level and integrated granular comment-level annotations. These annotations including topics, severity, and user role.

We employ network science methods to investigate cyberbullying patterns on Instagram. Our objectives can be summarized as follows:

- Construct localized user interaction networks (graphs) using labeled comments from Instagram sessions with prior evidence of cyberbullying. The proposed network construct is a novel representation of Instagram user interactions based on their cyberbullying roles. Fig. 1 provides an example Session Graph from the dataset.
- Quantify and characterize the different behavioral tendencies of the users across Instagram sessions based on metrics derived from the aforementioned networks. Specifically, we define the *Bully Score* and the *Victim Score*. Conceptually, these scores track the residual flow of malice and peer support accumulated on the Bully and the Victim nodes, respectively.
- Investigate the different *motifs* (graphlets)—connected sub-graphs with three or four nodes—present in the Session Graphs. We tabulate the frequently occurring motifs to reveal interesting mesoscale interaction patterns between groups of actors.

Related Work

Cyberbullying & Social Science. Cyberbullying is a complex interpersonal behavior with features and characteristics that are distinct from its offline counterpart (Baldry, Farrington, and Sorrentino 2017; Modecki et al. 2014). For instance, with traditional bullying, the bullying interactions are confined to a physical location. With cyberbullying, interactions escape their physical bounds and can manifest in whichever way the bully can contact or make a statement about the victim using electronic means, including sessions and private direct messages (Ali et al. 2022). Cyberbullying is also highly context dependent, manifesting differently based on the specific characteristics of the participants involved, developmental stages, and location (Ryoo, Wang, and Swearer 2015; Festl et al. 2017). While cyberbullying is often characterized as occurring primarily between two participants, it is shaped by individual, family, peer, community, and cultural factors (Swearer and Hymel 2015; Ma et al. 2024). Prior research has found cyberbullying typically involves different behavioral roles from its participants. Five

¹In this work, we use the terms *session* and *post* interchangeably. Each session consists of an image, a text caption, and a list of user posted comments.

commonly explored roles in the literature are: *Bully*, *Victim*, *Bully Supporter*, *Defender of the Victim*, and *Bystander*. We provide a detailed description of various cyberbullying roles relevant to this work in Tab. 1.

Incorporating user roles alongside severity is key to better grasp the persistent and pervasive nature of cyberbullying (Baumann et al. 2022; Bansal et al. 2024). An important nuance is that user roles are seldom *static* within a session. This is especially true if session participants make multiple comments within a session. Behavioral actions also permeate across sessions wherein past role incarnations leave downstream effects. Specially, the act of cyberbullying is a known predictor of prior victimization and vice-versa (Povedano et al. 2015; Lozano-Blasco, Cortes, and Latorre 2020).

Cyberbullying & Network Analysis. Networks, also referred to as graphs, are a flexible construct to model complex, interconnected systems where entities have relationships. Networks are widely used by sociologists to model complicated phenomena, *e.g.*, political polarization on social media and homophily in social networks (Conover et al. 2011; Christakis and Fowler 2013). Relevant to our discussion is the application of networks in the context of cyberbullying. Users within a session are a natural choice for nodes in a network with edges documenting their interactions. The specific characteristics of the edges, such as their directionality and weight, depend on specific context and nuances of the interactions. For example, directed edges can be used to model one-sided, directional relationships. Edge weights, on the other hand, can be used to signify the frequency of interactions, or encode the intensity associated with these interactions (Squicciarini et al. 2015; Soleimani, Pourshahbaz, and Shackelford 2026). Networks have also been used to draw connections between an individual’s aggressive behavior in relation to someone’s social status (Faris and Felmler 2011). In prior work, researchers have also turned to networks as input to cyberbullying classifiers. For example, Wang and Potika (2021) constructed follower networks that captured the strength of relationships between community members. The work by Kao et al. (2019) is particularly relevant for its consideration of cyberbullying roles (Victim, Bully, and Supporter). They identified victim-bully and victim-supporter pairs as found in user-comment ego networks. These pairs were then utilized for classifying individual user comments.

Data

The dataset at the core of our analysis is an extended version of the Instagram cyberbullying dataset collected by Hosseinmardi et al. (2015). The original dataset is a historical snapshot of sessions (posts) from 2013 to 2015, totaling 158,201 comments across 2,219 sessions. We opted to supplement this existing dataset, as after the 2018 Cambridge Analytica scandal, collecting data from Instagram became very difficult (Bruns 2021; Trezza 2023). Despite the EU’s Digital Services Act (DSA) requiring very large online platforms (VLOPs) to provide publicly accessible data to researchers, data collection still remains a significant challenge. Numerous researchers have demonstrated that DSA compliance by

Table 1: Definition and prevalence of cyberbullying roles used in this work. Count: the frequency (and percentage) of the 414 total sessions each role appears in. The roles associated with cyberbullying are italicized.

Role	Count	Definition
<i>Bully</i>	406 (98%)	Someone who attacks, harasses, humiliates, or threatens other people.
<i>Bully Assistant</i>	106 (26%)	Someone who sees bullying and begins attacking others.
<i>Aggressive Victim</i>	118 (29%)	A person being attacked who responds aggressively to their attackers.
Non-Aggressive Victim	228 (55%)	A person being attacked who either ignores the attack or responds non-aggressively to it.
<i>Aggressive Defender of Victim</i>	298 (72%)	A person who attempts to help someone being attacked by responding aggressively towards the attackers.
Non-Aggressive Defender of Victim	(A) 204 (49%) (B) 203 (49%)	A person who attempts to help someone being attacked through non-aggressive support. Exclusive to this role are two sub-types based on the intended target of the comment: (A) <i>Directly Confronting a Bully</i> or (B) <i>Supporting the Victim</i> .
Passive Bystander	411 (99%)	A person who either has not seen any bullying or chooses to ignore any bullying-related comments.

VLOPs has been both surface-level and antithetical to the needs of researchers (Mimizuka et al. 2025).

Each session in our dataset has an accompanying image and caption, a timestamp of when the session was posted, the session author’s username, the number of likes (at time of collection) and other metadata. Because the original dataset included cyberbullying annotations at only the session-level, it was unsuitable for capturing user interactions at the *comment-level*. Thus, we selected a subset of 438 sessions (35,364 comments in total) based on the criteria that a majority of the five original annotators agreed that the session constituted cyberbullying. With these 438 sessions, we used Amazon MTurk—a crowd-sourcing platform—to hire experienced *Master* annotators to label and categorize all 35,364 comments. All personally identifiable information was anonymized prior to being presented to the MTurk workers. Every comment was annotated by five independent annotators to ensure consistency. In addition, the deployed survey had built-in attention checks to improve the annotation reliability. Failing the attention checks lead to the disposal of the corresponding annotations.

In the survey, annotators were presented with a near replica of the original Instagram session page, where each comment had drop-down menus for annotating. When annotating a comment, the annotators were required to identify, (1) if the comment constituted bullying, (2) the comment author’s role, (3) the severity of the bullying, and (4) the topics of the bullying (*e.g.*, gender identity, race, social status). Severity is measured at three levels: *mild*, *moderate*, and *severe*. The cyberbullying roles and their definitions are present in Tab. 1. We present these roles in title case throughout the manuscript.

Crucial to our analyses, annotators were prompted at the end of the survey to pick the overall *Main Victim* of the session. This session-level annotation determines who, after considering all comments in the session, is the primarily-

impacted victim. The choices were, (a) the user who created the post, designated OP for Original Poster, (b) people depicted in the picture, designated Picture, (c) the participants in the comments, designated Participants, or (d) Other.

Preprocessing. As noted earlier, each comment has five independent annotations. To arrive at aggregated role labels, we opt for a two-step approach. We first segregate annotators into two groups, based on the boolean option ‘is bullying’ and ‘is not bullying’. We then take the annotators who are in the majority and disregard the minority annotators. For example, if three annotators said ‘is bullying’, but two annotators said ‘is not bullying’, we discard the annotations from those who said not bullying.

We then take another majority vote to decide the commenter’s role. It is possible for each of the remaining annotators to select a different role, hence, the need for a role tie-breaking heuristic. Of the 35,364 comments, 14,117 comments (40%) have ties, mostly between two and three roles. For cyberbullying identified comments, the role heuristic chooses a role with the preference: *Aggressive Defender* > *Aggressive Victim* > *Bully Assistant* > *Bully*. For non-cyberbullying identified comments, the heuristic chooses a role with the preference: *Non-Aggressive Defender: Support of the Victim* > *Non-Aggressive Defender: Direct to the Bully* > *Non-Aggressive Victim* > *Passive Bystander*. Notice that the heuristic is designed to prefer Defenders and Victims over Bullies and Passive Bystanders, as the latter two are the most common in the dataset. In addition, after arriving at the final consensus label for each comment, we discard all comments with the Passive Bystander role. This role communicates non-participation in sessions, *i.e.*, the Passive Bystander comments do not actively shape the discourse. Although bystander presence likely influences user behavior, measuring this impact accurately is challenging; therefore, we omitted them from the current analyses.

To quantify a comment’s annotated severity numerically,

Table 2: Example of preprocessed comments for a session. The consensus role and average severity score are computed across the majority annotations for each comment. These comments are used to construct the Session Graph in Fig 1. The annotated Main Victim is the *Poster*. Users can take on different roles as a comment thread progresses, e.g., u_3 is an Aggressive Defender, an Aggressive Victim, and a Non-Aggressive Victim within the same session. Seq: comment sequence id ordered by comment creation timestamp, Avg. Sev.: average comment severity score.

Seq.	User	Consensus Role	Avg. Sev.
1	u_1	Bully	2.00
2	u_2	Bully	1.00
3	u_3	Agg Defender	1.60
4	u_4	Agg Defender	1.67
5	u_5	Agg Victim	1.60
6	u_6	Bully	1.80
7	u_3	Agg Victim	1.40
8	u_3	Non-Agg Defender (Type A)	1.00

we use the following conversion scale. Non-cyberbullying and Mild annotations map to 1, Moderate maps to 2, and Severe to 3. Then we take the arithmetic mean across the majority annotators. Across all sessions, the average comment severity is 1.19.

Finally, to reach a consensus for the Main Victim of the session, first the Picture and the OP annotations are combined into a single category, *Poster*, and then the overall victim is determined through another majority vote. In the 36 (8.2%) sessions that contain a tie in their Main Victim majority, Poster is given preference over Participants. Because modeling the interactions between bullies and the victims is at the heart of this analysis, we subsequently filtered 24 sessions where the vote determined the Main Victim to be Other. Tab. 2 provides an illustration of the comment-level annotations of a session after applying the preprocessing step. From this point forward in the paper, the analysis focuses on the 414 sessions containing valuable comment-level labels.

Methods & Results

In this section, we describe our proposed frameworks and formalisms that substantiate our key contributions.

Session Graph

We construct user interaction networks based on the processed comment-level annotations. These networks are henceforth referred to as *Session Graphs*. Tab. 3 summarizes key statistics across all the constructed Session Graphs.

Formalism. A Session Graph \mathcal{G} is a directed graph represented by a 4-tuple $\mathcal{G} = \langle V, E, \kappa, \rho \rangle$, where V is a set of nodes; $E \subseteq V \times V$ is the set of directed edges; $\kappa : E \mapsto \mathbb{R}^+$ is a function assigning weights to edges; and $\rho : V \mapsto L$ is a function assigning nodes different labels, i.e., cyberbullying roles. L is the set of all possible cyberbullying roles (see Tab. 1 for definitions). Edges have a default weight of one

and the weights additively accumulate over repeated interactions between the same pair of nodes. An example Session Graph can be found in Fig. 1 with 9 nodes and 18 edges. Roles are differentiated by both node shape and color.

For consistency and convenience, we define three special subsets of nodes that convey related roles. Let \mathcal{B} represent the combined set of Bully and Bully Assistant nodes. Let \mathcal{V} represent the combined set of Main Victim, Aggressive Victim, and Non-Aggressive Victim nodes. And finally, let \mathcal{D} represent the aggregated set of Aggressive and Non-Aggressive Defenders. Formally, we define the sets as:

$$\mathcal{B} = \{v \in V \mid \rho(v) \in \{\text{Bully, Bully Asst}\}\}$$

$$\mathcal{V} = \{v \in V \mid \rho(v) \in \{\text{Main Victim, Agg Victim, Non-Agg Victim}\}\}$$

$$\mathcal{D} = \{v \in V \mid \rho(v) \in \{\text{Agg Def, Non-Agg Def}\}\}$$

We use the terms *Victims* and *Bullies* interchangeably with \mathcal{V} and \mathcal{B} , respectively. We now detail how Session Graphs are constructed from the annotated comment data.

Graph Construction. Consider a session S with k annotated comments where each comment C has a timestamp t , the commenter’s username u , the commenter’s role r , and severity score w . Every Session Graph \mathcal{G} is initialized with a single node representing the session’s Main Victim (either Poster, Participants, or Other). Subsequent nodes are distinct tuples of the commenters’ usernames and roles. Comments are processed sequentially, starting from the oldest to the newest, with each triggering the following topological changes in \mathcal{G} .

1. Add a new node in \mathcal{G} corresponding to the pair of username and role values (u, r) only if the current comment is the pair’s first appearance in the session. Note, this allows for multiple appearances of the same user in \mathcal{G} , provided they embody distinct roles across the comments.
2. Add new edges between the node (u, r) and *already existing* nodes following the guidelines described below. Instead of creating multiple edges between an already connected node pair, we simply increase the weight of the existing edge by the comment’s severity score w . This is a similar strategy to what Kao et al. (2019) proposed to track multiple interactions between users. This construction reflects the backwards looking perspective a user has when authoring a comment, wherein they respond to participants who have already written comments.

Guidelines for establishing edges from the current node (\circ) are presented below and are also illustrated in Fig. 2.

- Bully and Bully Assistants attack the Victims. We add separate directed edges of weight w , originating from the current node targeting Victims, i.e., $\circ \rightarrow \mathcal{V}$.
- Aggressive Defenders *simultaneously* pacify the Victims and attack the Bullies. We establish two sets of edges, all with the same edge weight w . The first set originate from the Victims and terminate at the current node, i.e., $\mathcal{V} \rightarrow \circ$. The second set of edges originate from the current node and are targeted towards the Bullies, i.e., $\circ \rightarrow \mathcal{B}$.
- Non-Aggressive Defenders are unique in that they have a subtype. Either they confront the Bullies (Type A) or they support the Victims (Type B). For Type A, we add

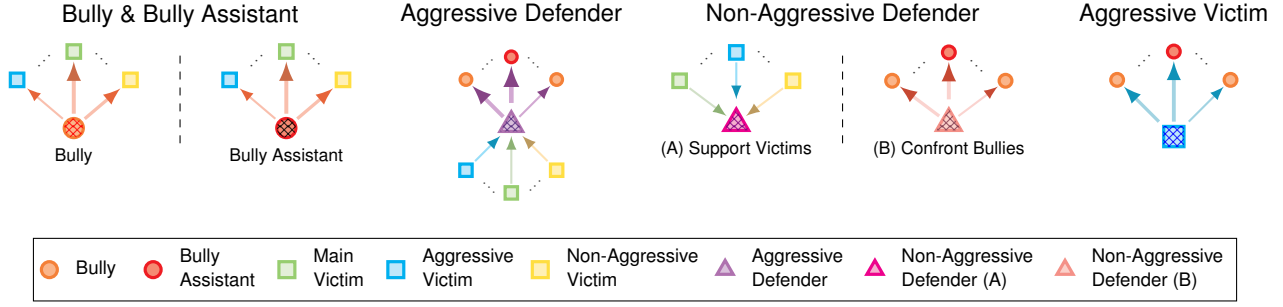


Figure 2: Schematic description of the different edges induced by a new comment during the Session Graph creation process. The focal node in each panel is shaded. The Main Victim and Non-Aggressive Victim nodes do not result in the addition of new edges and are therefore absent in this figure.

directed edges, $\circ \rightarrow \mathcal{B}$, with weight $w = 1$. Whereas Type B corresponds to directed edges, $\mathcal{V} \rightarrow \circ$, also with $w = 1$.

- Aggressive Victims confront the Bullies. This is represented by separate directed edges of weight w resembling $\circ \rightarrow \mathcal{B}$.
- Non-Aggressive Victim and Main Victim nodes do not automatically establish any edges of their own. They only receive edges as the result of actions taken by other nodes as they are introduced into the graph.

We illustrate the graph building process in Fig. 3 by utilizing the first five comments found in Tab. 2. After processing the remaining three comments, we arrive at the partially completed Session Graph shown in Fig. 1. Walking through the process, starting with comments one and two, as there are no bullies currently present in the graph, the only edges that are added are between the Main Victim node to the two Bullies (users u_1 and u_2) with weights 2.00 and 1.00, respectively. For the third comment, since u_3 is Aggressive Defender, this node will establish a directed edge with weight 1.60 to the Main Victim node and two directed edges with weight 1.60 to the Bullies. For comment four, since u_4 is also an Aggressive Defender, this node establishes a directed edge with weight 1.67 to the Main Victim and two directed edges with the same weight to the two Bully nodes. Finally, for comment five, as u_5 is an Aggressive Victim, it establishes two edges with the Bullies, one with u_1 and the other with u_2 , each of weight 1.60.

Victim & Bully Scores

To both quantify and differentiate the activity of Victims and Bullies within a session, we leverage network metrics that focus their attention on the subsets of nodes \mathcal{B} and \mathcal{V} that we defined earlier.

For each subset, we define a numeric *score* that is the average of the differences between the *weighted out-* and the *weighted in-degrees* of each set’s member nodes. Specially, we define these as the *Victim Score* and the *Bully Score*:

$$\text{Victim Score} = \frac{\sum_{v \in \mathcal{V}} (d_w^{\text{out}}(v) - d_w^{\text{in}}(v))}{|\mathcal{V}|}$$

$$\text{Bully Score} = \frac{\sum_{v \in \mathcal{B}} (d_w^{\text{out}}(v) - d_w^{\text{in}}(v))}{|\mathcal{B}|},$$

where d_w^{out} and d_w^{in} represent the weighted in- and out-degree of a node in \mathcal{G} .

These scores, when considered in unison, communicate the balance of power present between the Victims and the Bullies of a session. A high Victim Score reflects that Victims receive outweighed support from Defenders, a lack of Bullies, or both, whereas a high Bully Score indicates that the bullying occurred without significant pushback from Defenders or Aggressive Victims.

We now look at the trends for the scores across the dataset. Fig. 4(A) presents the distribution of the Victim Score and Tab. 3 offers the median of said distribution at -8.98. The Victim Score distribution is visibly left-skewed, with 345 (84%) (highlighted in red) of the scores being negative. Only 69 (16%) (highlighted in blue) are positive. Similarly, Fig. 4(B) presents the distribution of the Bully Score. Per Tab. 3, the median Bully Score is 1.33. This distribution has a slight right skew, with 141 (34%) (shown in red) negative scores compared to 273 (66%) (shown in blue) positive scores. The Victim Score distribution shows considerably greater variability, as evidenced by its long, drawn-out tail of negative scores. This is also reflected in the width of the confidence intervals for the mean (Tab. 3). Considering the average number of Victim nodes is 3.68, this suggests Victim nodes have substantially more incoming edges from Bullies, as opposed to outgoing edges from the presence of Defenders. The distribution of the Bully Scores, on the other hand, contains less variability.

Fig. 5 presents the joint distribution of the Bully and Victim scores. Each quadrant represents one of four possible combinations of the different signs assumed by the Bully Score and the Victim Score, respectively. Therefore, these quadrants communicate different outcomes in the bully-victim dynamic. For instance, the session depicted in Fig. 1 would be located in Quadrant IV where Bullies dominate the discourse. Both Scores are simultaneously positive in the 16 sessions (3.9%) present in Quadrant I. These sessions represent an evenly matched scenario where the amount of aggression from the Bullies is displaced by nearly equal support by Defenders and Aggressive Victims. Quadrant II, with 53 sessions (12.8%), demonstrates when Victims and Defenders dominate the interaction. They either vastly outnumber the Bullies or overwhelm them by flooding sup-

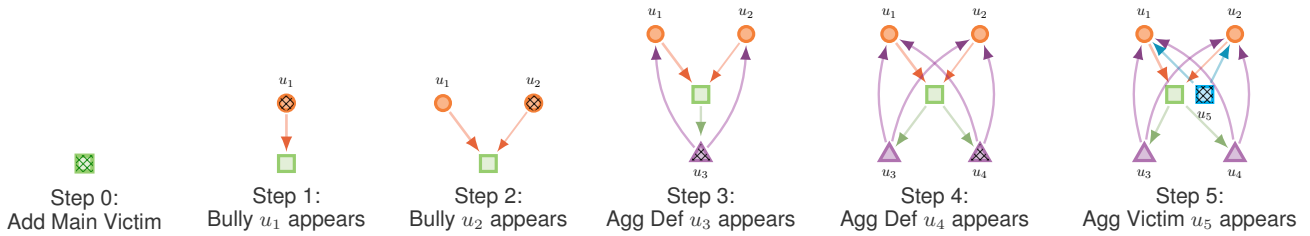


Figure 3: Step-by-step construction of the Session Graph \mathcal{G} based on the first five comments in Tab. 2. The newly introduced node in each step is shaded. Roles are abbreviated to save space.

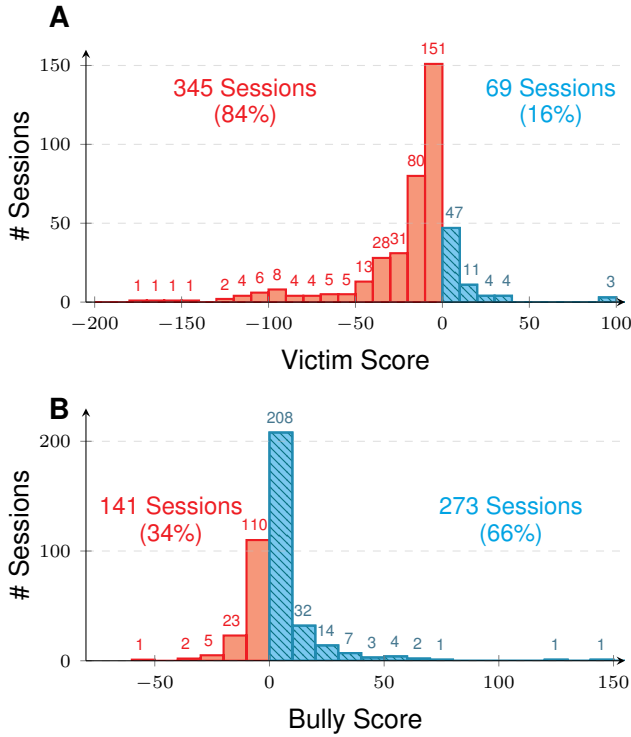


Figure 4: Victim Score (panel A) and Bully Score (panel B) distributions for $N = 414$ sessions. Positive scores are marked in blue, while the negative scores are in red. Bin width in both panels is 10. Victim Scores are strongly skewed to the left, whereas the Bully Scores have a slight right skew.

port to the Victims. In Quadrant III, which contains 88 sessions (21.2%), Bullies face pushback from the Aggressive Victims and Defenders. The simultaneously negative Victim and Bully scores suggest repeated back and forth between the two factions, with the Bullies maintaining a slight advantage. Finally, Quadrant IV is where most sessions reside (257 sessions, 62.1%). It is also the quadrant where Bullies clearly dominate the interactions. Despite the wide range of scores, we find a dense concentration of sessions around the origin. This implies that in many sessions, the positive and the negative interactions mutually neutralize each other.

Table 3: Session Graph statistics for 414 cyberbullying sessions. The Mean CI column represents the 95% confidence intervals around the mean. On average, graphs typically have 24 nodes and 85 edges, with Bullies outnumbering the Victims by a factor of 4. Victims display a wide weight disparity across their incoming and outgoing links, manifesting in their highly negative score. Bullies have a more balanced incoming and outgoing connection intensities.

	Median	Mean	Mean CI
Nodes	17	23.65	[21.74, 25.74]
Edges	41	85.14	[74.49, 99.04]
Victims			
Count	2	3.68	[3.33, 4.11]
Wtd. Out-degree	4.44	8.53	[7.45, 9.97]
Wtd. In-degree	14.72	26.04	[23.31, 29.25]
Score	-8.98	-17.51	[-20.79, -14.72]
Bullies			
Count	8	13.19	[11.86, 14.90]
Wtd. Out-degree	3.32	9.03	[7.74, 10.83]
Wtd. In-degree	2.63	5.39	[4.74, 6.14]
Score	1.33	3.64	[2.32, 5.35]

Motifs

As indicated earlier, we seek to understand the prevalence of certain interaction patterns between nodes of different roles. We achieve this by tracking distributions of substructures called *motifs* within the Session Graphs. Motifs can be considered the building blocks that define a graph. Therefore, they also provide a mesoscale view of connectivity patterns in a larger graph. For our purposes, we consider a motif \mathcal{M} to be a connected, induced subgraph of a larger graph containing either three or four nodes. We obtain motifs by using the iGraph library’s implementation of the FANMOD algorithm on our Session Graphs \mathcal{G} (Wernicke and Rasche 2006; Csárdi and Nepusz 2006). It is important to note that a specific motif \mathcal{M} may appear multiple times within a graph, and similarly a graph may contain several distinct motifs.

Our motif analysis strategy applies a few simplifications to the Session Graph structure. We aggregate roles corresponding to \mathcal{V} , \mathcal{B} , and \mathcal{D} before running the motif search algorithm. So, for example, the Main Victim, Non-Aggressive Victims, and Aggressive Victims are mapped to a singular role *Victim*. We also group the edges by weight into two buckets, *light* (weight < 2) and *heavy* (weight ≥ 2). These

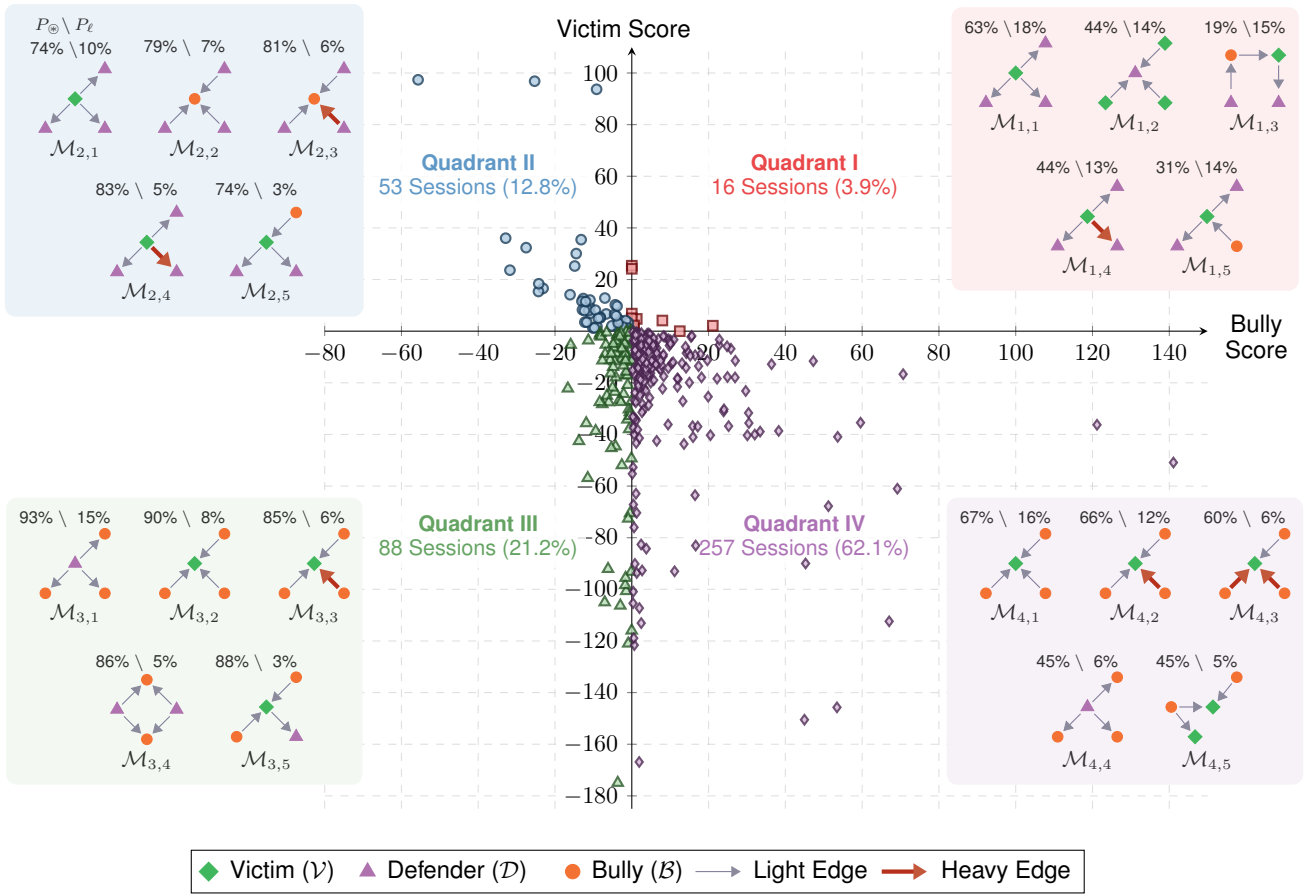


Figure 5: Joint distribution plot of the Bully and Victim Scores for 414 sessions. The main plot is divided into four quadrants. Quadrant I (red squares) represents evenly matched sessions; In Quadrant II (blue circles) Victims and Defenders Dominate; Bullies face pushback in Quadrant III (green triangles), and Bullies dominate in Quadrant IV (purple diamonds). Within each quadrant, we enumerate the top 5 motifs based on a combined Local and Global Prevalence ranking. Motifs are indexed $\mathcal{M}_{i,j}$, where $i \in \{1, 2, 3, 4\}$ is the quadrant and $j \in \{1, \dots, 5\}$ is the within-quadrant rank. Node shapes denote roles: Victim (green diamond), Defender (purple triangle), and Bully (orange circle). Directed, weighted edges indicate the flow and intensity of interactions within each motif.

simplifications reduce the number of unique motif patterns to analyze, simplify the presentation of the most frequently occurring structures while preserving the most frequently occurring patterns, and let us differentiate between one-off or mild interactions and the repeated, more intense kinds. Figures 5 and 6 provide examples of motifs that were found as part of the enumeration process.

Session Graphs typically contain a wide variety of motifs. Therefore, we rely on two complementary statistical measures to quantify the significance of particular motifs. These measures are designed to capture a motif \mathcal{M} 's prevalence or relative importance. The first measure, *Local Prevalence*, is the prevalence of a motif localized to a specific session. While the second measure, *Global Prevalence*, considers the prevalence of a motif across multiple sessions. We provide their formal definitions below.

Let $\mathcal{S} = \{S_1, \dots, S_m\}$ represent the set of m sessions currently under consideration. We define a function $f(\mathcal{M}, S)$ that counts the frequency of occurrence of a given

motif \mathcal{M} in a session $S \in \mathcal{S}$.

Global Prevalence. $P_{\otimes}(\mathcal{M}, \mathcal{S})$ measures the proportion of sessions within \mathcal{S} where the motif \mathcal{M} appears at least once.

$$P_{\otimes}(\mathcal{M}, \mathcal{S}) = \frac{|\{S \mid f(\mathcal{M}, S) > 0, S \in \mathcal{S}\}|}{|\mathcal{S}|}$$

This measure captures how widespread a motif is across a set of sessions, regardless of how frequently it appears within any individual session. Even if a motif is rare within each session, a motif with high P_{\otimes} score shapes the background dynamics through its frequent reappearance.

Local Prevalence. $P_{\ell}(\mathcal{M}, \mathcal{S})$ is the average probability of observing the motif \mathcal{M} within the sessions it appears in.

$$P_{\ell}(\mathcal{M}, \mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{S \in \mathcal{S}} \left(\frac{f(\mathcal{M}, S)}{\sum_{m \in \mu(S)} f(m, S)} \right)$$

The Local Prevalence calculation has the additional dependency on the set $\mu(S) = \{\mathcal{M}_1, \dots, \mathcal{M}_n\}$, the set of n dis-

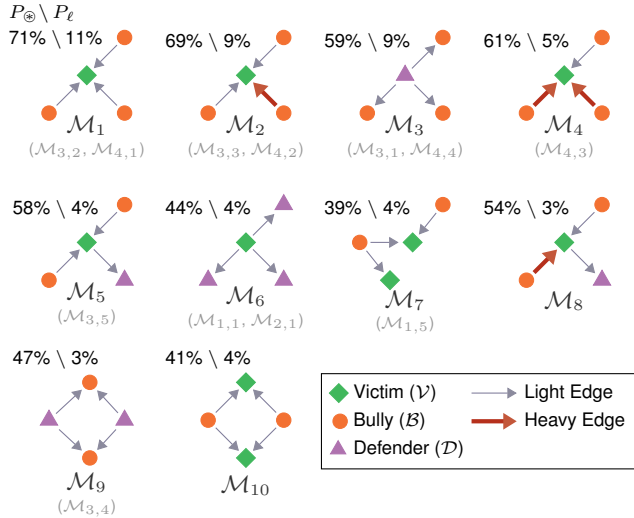


Figure 6: Overall top 10 motifs with high Global Prevalence (P_g) and high Local Prevalence (P_l) scores. Motifs are indexed \mathcal{M}_1 through \mathcal{M}_{10} in ascending order of the rank product. Most motifs also appear in Fig. 5, so we also list those indices in gray. Each motif is annotated with its $P_g \setminus P_l$ scores. Node shapes represent roles: Victim (green diamond), Defender (purple triangle), and Bully (orange circle). Directed, weighted edges indicate the flow and intensity of interactions within each motif.

tinct motifs that appear in a given session S . The inner fraction measures the relative frequency of \mathcal{M} compared to the rest of the motifs appearing in session S . The outer summation serves to compute the average.

When a motif manifests in a session, Local Prevalence (P_l) is a indicator of its intensity. For example, motifs that appear in many sessions but at low frequencies would have a low P_l score. In contrast, motifs that appear in fewer sessions overall but recur frequently within those sessions would have a high P_l score. By focusing our analysis on motifs with high P_l , we avoid overemphasizing those that are scarcely present and therefore do not meaningfully shape the sessions' dynamics.

Motif Ranking Framework. To identify the most prominent motifs, we apply a ranking framework based on the two measures P_g and P_l . Depending on the context, \mathcal{S} is either set as the entire set of sessions, or the sessions within an individual quadrant found in Fig. 5. Motifs are first ranked separately by P_g and by P_l using dense ranking, with the highest values assigned rank 1. We then compute a combined rank as the product of these two ranks. This formulation emphasizes motifs with high Global and Local Prevalence scores since only motifs that perform strongly on both measures yield a low rank product. Finally, motifs are ordered by increasing rank product, with the lowest values indicating motifs that score strongly on both measures. Fig. 6 shows the top ranked motifs across the entire dataset while Fig. 5 displays the top performers within each quadrant in separate boxes. We will describe these results next.

Global Motif Patterns. In Fig. 6, we show the overall top 10 motifs identified across all 414 sessions in the dataset. We set \mathcal{S} to span all sessions while calculating the motif prevalence scores. Following this process, three notable structural patterns emerge. (1) Mild, repeated low-impact bullying. In this pattern, Bullies repeatedly attack the Victims. This pattern is the most prevalent motif structure across all sessions, though with some variation introduced through the intensity of the edge weights. In short, we find that repeated low-intensity or one-off attacks are more widespread than severe exchanges of hostility. We can also consider this structure to be a defining characteristic of cyberbullying sessions. (2) Intense bullying and heavy intervention occur intermittently. While heavy edges appear in several top-ranking motifs, they do not dominate the structures. They tend to be localized, highlighting specific repeated, intense attacks within otherwise light interactions. This indicates that intense harassment is present, but not pervasive. (3) Defender involvement is common, but is of low-intensity. The presence of Defenders in high-ranking motifs shows that intervention is a regular feature of the sessions, but the edges they contribute are predominantly light. This reflects quick attempts to counteract bullying, though Defenders do not engage in prolonged interventions. Throughout the top motifs, bullying, victimization, and defending behaviors frequently co-occur. The edge weights reveal that these multi-actor structures are often held together by streams of light weight interactions.

Quadrant-specific Motif Patterns. Next, we zoom in on the top 5 motifs in the four quadrants in Fig. 5. Here, we set \mathcal{S} to the sessions present in a given quadrant before computing the motif prevalence scores. For example, the Local Prevalence score is calculated only across the sessions in that quadrant. This means this value can differ significantly from the overall Local Prevalence score, as the latter is calculated across all sessions. We summarize our findings below.

Quadrant I is the smallest quadrant by far, with only 16 sessions. While the quadrant accounts for a tiny portion of the dataset, it still has influence. Notably, the $\mathcal{M}_{1,1}$ in Fig. 5 that depicts a Victim reaching out to three Defenders for support with light edges also appears among the overall top 10 motifs (\mathcal{M}_6). The motifs $\mathcal{M}_{1,1}$ and $\mathcal{M}_{1,5}$ also appear in the top 5 for Quadrant II. The motifs in Quadrant I are characterized by Defender-Victim dynamics, with Defenders playing a central and structurally consistent role across the top five motifs. Another key feature of these motifs is the predominance of light interactions. Overall, these motifs represent situations in which Victims are shielded by multiple Defenders. This configuration reflects scenarios where several users step in to support a Victim, each contributing support or counter-attacking the Bullies. These patterns in unison reveal that in Quadrant I, cyberbullying interactions are defined less by aggression and more by low-intensity defensive behavior.

Quadrant II contains 53 sessions and is defined by strong Defender activity. Across the top motifs, Defenders consistently cluster around either a Victim or a Bully, intervening through light, low-intensity interactions. Heavy edges appear sporadically ($\mathcal{M}_{2,3}$ and $\mathcal{M}_{2,4}$ in Fig. 5), indicating iso-

lated moments of escalation rather than sustained conflict. Overall, this quadrant reflects situations where Defenders frequently intervene, shaping the interaction pattern through repeated but brief supportive or punitive actions. The quadrant also contributes to the global landscape, with $\mathcal{M}_{2,1}$ appearing among the overall top 10 motifs (\mathcal{M}_6).

Quadrant III is the second largest quadrant with 88 sessions and is characterized by high Bully activity, with Bullies comprising the majority of nodes in the top motifs. All but one interaction is light, which reflects the low-intensity attacks directed at either a Victim or a Defender. Only one heavy edge appears in ($\mathcal{M}_{3,3}$ in Fig. 5), which shows that there is occasional escalation, but it is very rare. Across the motifs, Victims and Defenders are outnumbered by Bullies. In three of the five motifs, $\mathcal{M}_{3,2}$, $\mathcal{M}_{3,3}$, and $\mathcal{M}_{3,5}$, Bullies are actively targeting a Victim, reflecting persistent and coordinated bullying behavior. Some resistance does emerge, most notably in $\mathcal{M}_{3,1}$, where a single Defender confronts three Bullies. Although less common, the Defender-driven motifs does have influence, with a Global Prevalence of 93% and a Local Prevalence of 15%. Overall, Quadrant III captures environments characterized by widespread bullying activity alongside selective defensive intervention. The motifs identified here have strong overall influence with $\mathcal{M}_{3,1}$ through $\mathcal{M}_{3,5}$ appearing in the overall top 10 motifs, highlighting their importance across the dataset.

Bullies dominate in Quadrant IV. This is by far the largest quadrant, with 257 sessions (62% of the dataset), and the most influential quadrant, shaping the motif landscape across the entire dataset. $\mathcal{M}_{4,1}$ through $\mathcal{M}_{4,4}$ are also the top four motifs overall (\mathcal{M}_1 – \mathcal{M}_4 in Fig. 6), emphasizing the dominance of the structural patterns in this quadrant. In every top motif, Bullies appear at substantially higher frequencies than Defenders and Victims. This quadrant exhibits the greatest concentration of heavy edges among all quadrants, appearing in $\mathcal{M}_{4,2}$ and $\mathcal{M}_{4,3}$. Although motifs $\mathcal{M}_{4,1}$ through $\mathcal{M}_{4,3}$ share the same underlying structure of three Bullies targeting a single Victim, differences in edge weights reveal meaningful variation in the intensity of these interactions. The first motif, $\mathcal{M}_{4,1}$, contains only light edges; however, as the ranking progresses, heavy edges begin to appear. This progression provides a clear illustration of escalation. When one Bully increases the severity of their attack, it appears to signal to others that similar escalation is acceptable, leading additional Bullies to intensify their behavior. While Defender activity is present in this quadrant, particularly in $\mathcal{M}_{4,4}$ with a local prevalence of 6% and a global prevalence of 45%, such motifs remain limited, and the quadrant exhibits little meaningful resistance to bullying behavior. Overall, Quadrant IV reflects environments where bullying occurs repeatedly and with little interruption, shaping much of the global motif landscape through widespread, largely unopposed aggression.

Broader Behavioral Dynamics. We now zoom out further to get a broader perspective and highlight a few significant motif structures.

Coordinated Behavior: These manifest as *star* structures, embodying two primary variants: (1) The *Mobbing* star structure is characterized by three Bullies targeting a single

Victim. It emerges as the most distinct and dominant structure across the dataset (\mathcal{M}_1 , \mathcal{M}_2 , and \mathcal{M}_4 in Fig. 6), particularly within Quadrant III ($\mathcal{M}_{3,2}$ and $\mathcal{M}_{3,3}$ in Fig. 5) and Quadrant IV ($\mathcal{M}_{4,1}$, $\mathcal{M}_{4,2}$, and $\mathcal{M}_{4,3}$). The primary variation of this motif, \mathcal{M}_1 , is exceptionally common, exhibiting Local and Global Prevalence Scores of 11% and 71% respectively. Structurally, the Victim functions as a terminal sink node, absorbing directed aggression from three separate Bullies. This isolation suggests a contagion effect: once a Victim is marked by one Bully, they become a target for others, accelerating the harassment. While the most prevalent form (\mathcal{M}_1 in Fig. 6) involves low-intensity edges, the structure evolves into heavier variants (\mathcal{M}_2 and \mathcal{M}_4), indicating that as the mobbing normalizes within the group, the severity of the attacks escalates.

(2) The *Mobilization* star structure represents the inverse of the mobbing dynamic. Instead of multiple Bullies, there is a single Victim being pacified by multiple Defenders within the wider network. This structure is a defining feature of Quadrant II, particularly the motif $\mathcal{M}_{2,1}$, which appears in nearly 74% of the quadrant’s sessions and accounts for over 10% of the local interaction volume. The pattern is also visible in the global baseline (\mathcal{M}_6) and in Quadrant I ($\mathcal{M}_{1,1}$ and $\mathcal{M}_{1,4}$).

Distributed Defense: These are the structural inverses of the Mobbing star. Rather than collective aggression, these motifs represent collective resistance. This resistance manifests in three distinct configurations: a single Defender managing multiple threats (\mathcal{M}_3 , $\mathcal{M}_{3,1}$, and $\mathcal{M}_{4,4}$), a balanced standoff (\mathcal{M}_9 and $\mathcal{M}_{3,4}$), or an overwhelming defense ($\mathcal{M}_{2,2}$ and $\mathcal{M}_{2,3}$). Among these, motif $\mathcal{M}_{3,1}$ is particularly dominant, appearing in 93% of Quadrant III sessions and accounting for 15% of the local interaction volume. Crucially, with the exception of one heavy edge in $\mathcal{M}_{2,3}$, these interactions are characterized by low-intensity edges. This suggests that Defenders are utilizing a deterrence-oriented strategy. Rather than engaging in high-conflict arguments, they act as informal moderators, spreading their intervention across the group to signal behavioral boundaries and regulate aggression without causing escalation.

The Lack of Cycles: While our analysis prioritizes significant structures, the absence of specific patterns is equally telling. Notably, our results show a complete lack of cycle structures. In network science, an acyclic graph implies a strict hierarchy or a unidirectional flow of energy from a source node to a sink node. In a cyberbullying context, we would typically expect to observe reciprocity, a feedback loop where a Defender pushes back against a Bully, and the Bully retaliates. The absence of this cycle suggests that Bullies operate mainly as hit-and-run aggressors: they attack a Victim and disengage. Even when a Defender intervenes, the Bully does not fight back, implying they either ignore the intervention or shift targets. Consequently, the network functions as a cascading waterfall where aggression flows from the Source (Bully) to the Target (Victim) and finally to the Sink (Defender), as seen in linear motifs such as $\mathcal{M}_{1,3}$. This acyclic nature reveals a topology of unresolved conflict; interactions are linear and fleeting, preventing the formation of the resilient, reciprocal community structures found in

healthier social networks (Coleman 1988).

Discussion

Broader Impact. The identification of distinct motifs in the present work has important implications. For instance, the findings can facilitate the development of more effective moderation techniques on social media platforms. To illustrate, the inclusion of Bully and Victim scores in moderation dashboards could help moderators identify instances in which a Bully score is increasing rapidly in comparison to a Victim score and more quickly intervene. Moreover, because motifs, themselves, encode rich, nuanced, mesoscale user interactions that are seldom captured by conventional text-based cyberbullying detection algorithms, integrating them into training datasets can help existing models make better predictions. Our findings can also contribute new insights about the dynamics of social interaction in online environments that may be particularly informative for social scientists. A deeper understanding of factors that give rise to motifs involving retaliatory aggression by a victim and/or defender versus the provision of more prosocial support can, for example, advance empirical work in the areas of conflict management and mitigation (Zeitzoff 2017), shed light on distinct attributes of cyberbystander behavior (Rudnicki et al. 2023) within the context of a social media session, and add to emerging research on the multidirectional relationship between cyberbullying perpetration and victimization (Jin et al. 2025).

Limitations & Future Work. We recognize two main limitations of our work. First, this work was conducted on a relatively old and very small Instagram dataset which is also limited to cyberbullying sessions. This increases the likelihood that our findings may have resulted from confounding factors and selection bias. Furthermore, due to the age of this dataset, our framework for interactions lacks a critical feature that is found on most contemporary social media platforms: reply-threads. We made a deliberate choice not to utilize any supervised or unsupervised learning techniques for annotating the dataset, but this has limited our ability to scale up our annotation process due to the costs involved in using high-quality human annotators. Since machine learning models are increasingly considered viable tools for annotating datasets (Tan et al. 2024), we plan to collect new, larger datasets from multiple platforms and deploy machine learning models to annotate them. This addition will allow us to perform cross-platform analysis of cyberbullying interactions. We believe that our interaction framework can easily be extended to incorporate the reply-thread feature found on most current day platforms. The second significant limitation is the lack of statistical validation. Every design decision is an opportunity to validate our findings via a *null model*. For example, one could randomly rewire the edges of a Session Graph to assess the significance of the observed Bully and Victim scores. The systematic application of this methodology at every design decision point would increase the robustness of this approach by providing greater contextualization and facilitate a better understanding of the impact of each choice. In addition, this work focuses on analyzing interactions in *static*, completed sessions. To address this,

our future work will focus on studying cyberbullying interactions as they unfold over time. A temporal analysis would ideally provide insight on the tendency for Bully and Victim scores to gravitate towards the origin, as seen in Fig. 5. Future work can also explore alternative mappings of comment severity scores to the edge weights, which, by design, contribute significantly to both the Victim and Bully scores and how different motifs manifest.

Conclusion

Using 414 Instagram sessions annotated for severity and cyberbullying roles, we utilize a network science framework to construct cyberbullying interaction graphs. To quantify the balance of power between the Victims and Bullies within sessions, we proposed two metrics, the Victim Score and Bully Score. We found that most sessions exhibit a negative Victim Score and a slightly positive Bully Score. Moreover, the joint distribution of the Bully and Victim Scores reveals four quadrants to explore. These embody qualitatively distinct dynamics between Bullies and Victims—evenly-matched sessions in which Bullies and Victims engage in comparable levels of activity; sessions where the Victims and Defenders dominate; sessions in which Bullies face substantial pushback; and sessions in which Bullies dominate. To further understand the more granular cyberbullying dynamics, we enumerated motifs—small subgraphs that provide a mesoscale view of connectivity patterns. Through the Global (across session) and Local (within session) Prevalence scores, we identified the most frequently-occurring motifs. The analysis of prevalent motifs highlighted emergent patterns, such as coordinated behavior (stars) and distributed defense. To our knowledge, this is the first study that leverages human annotations at the session and comment level to explore network-level granular cyberbullying interactions on Instagram.

Acknowledgment

This work was supported by National Science Foundation Awards No. 2435164 and No. 2435165, a Google Award for Inclusion Research, and a Lambda Research Grant.

The authors declare no competing interests.

References

- Ali, S.; Razi, A.; Kim, S.; Alsoubai, A.; Gracie, J.; De Choudhury, M.; Wisniewski, P. J.; and Stringhini, G. 2022. Understanding the digital lives of youth: Analyzing media shared within safe versus unsafe private conversations on Instagram. In *CHI 2022*.
- Baldry, A. C.; Farrington, D. P.; and Sorrentino, A. 2017. School bullying and cyberbullying among boys and girls: Roles and overlap. *Journal of Aggression, Maltreatment & Trauma*.
- Bansal, S.; Garg, N.; Singh, J.; and Van Der Walt, F. 2024. Cyberbullying and mental health: past, present and future. *Frontiers in Psychology*.
- Baumann, S.; Bernhard, A.; Martinelli, A.; Ackermann, K.; Herpertz-Dahlmann, B.; Freitag, C.; Konrad, K.; and Kohls,

- G. 2022. Perpetrators and victims of cyberbullying among youth with conduct disorder. *European Child & Adolescent Psychiatry*.
- Bruns, A. 2021. After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Dis-information and data lockdown on social platforms*.
- Chen, H.; and Zolkepli, I. A. 2025. Cyberbullying and bystanders: A bibliometric analysis. *Online Journal of Communication and Media Technologies*.
- Christakis, N. A.; and Fowler, J. H. 2013. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in medicine*.
- Coleman, J. S. 1988. Social capital in the creation of human capital. *American Journal of Sociology*.
- Conover, M.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Menczer, F.; and Flammini, A. 2011. Political polarization on twitter. In *ICWSM 2011*.
- Csárdi, G.; and Nepusz, T. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems*.
- Faris, R.; and Felmlee, D. 2011. Status struggles: Network centrality and gender segregation in same-and cross-gender aggression. *American Sociological Review*.
- Festl, R.; Scharkow, M.; and Quandt, T. 2015. The Individual or the Group: A Multilevel Analysis of Cyberbullying in School Classes. *Human Communication Research*.
- Festl, R.; Vogelgesang, J.; Scharkow, M.; and Quandt, T. 2017. Longitudinal patterns of involvement in cyberbullying: Results from a Latent Transition Analysis. *Computers in Human Behavior*.
- FORCE11. 2020. The FAIR Data principles. <https://force11.org/info/the-fair-data-principles/>.
- Geburu, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*.
- Hosseinmardi, H.; Mattson, S. A.; Rafiq, R. I.; Han, R.; Lv, Q.; and Mishra, S. 2015. Detection of cyberbullying incidents on the instagram social network. *arXiv:1503.03909*.
- Jin, S.; Li, B.; Hu, G.; Lian, Z.; and Liu, Z. 2025. A Self-Perpetuating Loop: Bivariate Dynamic Bidirectional Relationships Linking Cyberbullying Perpetration, Cybervictimization, and Social Networking Use Intensity. *Journal of Interpersonal Violence*.
- Kao, H.-T.; Yan, S.; Huang, D.; Bartley, N.; Hosseinmardi, H.; and Ferrara, E. 2019. Understanding cyberbullying on Instagram and Ask. Fm via social role detection. In *WWW 2019*.
- Lozano-Blasco, R.; Cortes, A.; and Latorre, P. 2020. Being a cybervictim and a cyberbully – The duality of cyberbullying: A meta-analysis. *Computers in Human Behavior*.
- Ma, J.; Su, L.; Li, M.; Sheng, J.; Liu, F.; Zhang, X.; Yang, Y.; and Xiao, Y. 2024. Analysis of prevalence and related factors of cyberbullying–victimization among adolescents. *Children*.
- Mimizuka, K.; Brown, M. A.; Yang, K.-C.; and Lukito, J. 2025. Post-post-API age: Studying digital platforms in scant data access times. *arXiv:2505.09877*.
- Modecki, K. L.; Minchin, J.; Harbaugh, A. G.; Guerra, N. G.; and Runions, K. C. 2014. Bullying prevalence across contexts: A meta-analysis measuring cyber and traditional bullying. *Journal of Adolescent Health*.
- Povedano, A.; Cava, M.-J.; Monreal, M.-C.; Varela, R.; and Musitu, G. 2015. Victimization, loneliness, overt and relational violence at the school from a gender perspective. *International Journal of Clinical and Health Psychology*.
- Rudnicki, K.; Vandebosch, H.; Voué, P.; and Poels, K. 2023. Systematic review of determinants and consequences of bystander interventions in online hate and cyberbullying among adults. *Behaviour & Information Technology*.
- Ryoo, J. H.; Wang, C.; and Swearer, S. M. 2015. Examination of the change in latent statuses in bullying behaviors across time. *School Psychology Quarterly*.
- Singh, A.; Rejeb, A.; Nangru, H.; and Pathak, S. 2024. Global research trends on cyberbullying: A bibliometric study. *Computers in Human Behavior Reports*.
- Soleimanian, M.; Pourshahbaz, A.; and Shackelford, T. K. 2026. The personality architecture of online aggression: A network analysis integrating HEXACO, the Dark Tetrad, and Life History Strategy in Iranian adolescents. *Personality and Individual Differences*.
- Squicciarini, A.; Rajtmajer, S.; Liu, Y.; and Griffin, C. 2015. Identification and characterization of cyberbullying dynamics in an online social network. In *ASONAM 2015*.
- Swearer, S. M.; and Hymel, S. 2015. Understanding the psychology of bullying: Moving toward a social-ecological diathesis–stress model. *American Psychologist*.
- Tan, Z.; Li, D.; Wang, S.; Beigi, A.; Jiang, B.; Bhattacharjee, A.; Karami, M.; Li, J.; Cheng, L.; and Liu, H. 2024. Large Language Models for Data Annotation and Synthesis: A Survey. In *EMNLP 2024*.
- Trezza, D. 2023. To scrape or not to scrape, this is dilemma. The post-API scenario and implications on digital research. *Frontiers in sociology*.
- Vogels, E. A. 2022. Teens and Cyberbullying 2022. <https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/>.
- Wang, A.; and Potika, K. 2021. Cyberbullying classification based on social network analysis. In *IEEE BigDataService 2021*.
- Wernicke, S.; and Rasche, F. 2006. FANMOD: a tool for fast network motif detection. *Bioinformatics*.
- Zeitsoff, T. 2017. How social media is changing conflict. *Journal of Conflict Resolution*.
- Zych, I.; and Farrington, D. P. 2026. Defining Bullying and Cyberbullying According to the Scientific Community: A Mixed Method Study. *International Journal of Bullying Prevention*.

Paper Checklist to be included in your paper

1. For most authors...
 - (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes, our research investigates cyberbullying by using network analysis without violating privacy norms. Our model uses anonymized data and aims at decreasing unfair profiling, not disrespecting societies or cultures, and discouraging unfair profiling online.**
 - (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? **Yes, our abstract and introduction state our main contributions of (1) constructing networks from a rich annotated dataset, (2) defining network centric metrics, and (3) using motifs to find higher-order patterns in cyberbullying.**
 - (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes, see the Methods section. We explain our methodology, goals, dataset characteristics, and model architecture, which all enhance our claims made about network analysis on cyberbullying.**
 - (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **Yes, see the Limitations & Future Work section. We recognize temporal limitations and random factors that could impact motifs.**
 - (e) Did you describe the limitations of your work? **Yes, see the Limitations & Future Work sections.**
 - (f) Did you discuss any potential negative societal impacts of your work? **Yes, see the Introduction section. We acknowledge that cyberbullying research is multi-faceted and complex, making it difficult to understand and analyze problems.**
 - (g) Did you discuss any potential misuse of your work? **See the Introduction section where we explain the importance of utilizing research to foster positive online interactions rather than enabling cyberbullying on social media.**
 - (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **Yes, see the Dataset section and, specifically, our discussion about attention checks, data anonymization, data splitting, and secure and responsible access to data to ensure integrity.**
 - (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes, our research focuses on examining cyberbullying by calculating bullying and victim scores to protect against harmful interactions on social media. We use anonymized data during collection to reduce harm and conform to ethical guidelines for research.**
2. Additionally, if your study involves hypotheses testing...
 - (a) Did you clearly state the assumptions underlying all theoretical results? **NA**
 - (b) Have you provided justifications for all theoretical results? **NA**
 - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **NA**
 - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **NA**
 - (e) Did you address potential biases or limitations in your theoretical framework? **NA**
 - (f) Have you related your theoretical results to the existing literature in social science? **NA**
 - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **NA**
3. Additionally, if you are including theoretical proofs...
 - (a) Did you state the full set of assumptions of all theoretical results? **NA**
 - (b) Did you include complete proofs of all theoretical results? **NA**
4. Additionally, if you ran machine learning experiments...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **NA**
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **NA**
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **NA**
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **NA**
 - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **NA**
 - (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? **NA**
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
 - (a) If your work uses existing assets, did you cite the creators? **Yes, a citation to the original Instagram dataset is included in the Data section. We have also uploaded the dataset in the GitHub repository (<https://github.com/ysilva/cb-motifs>).**
 - (b) Did you mention the license of the assets? **No, but the original Instagram dataset we used is publicly available for research purposes through the cited publication (Hosseinmardi et al. 2015).**
 - (c) Did you include any new assets in the supplemental material or as a URL? **Yes, we made our code available through a GitHub repository.**

- (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes, see the Dataset section.](#)
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes, see the Dataset section.](#)
 - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? [NA.](#) Please note that FAIR-related aspects of the used dataset are described in the dataset paper undergoing parallel review. If the current submission is accepted, we will make the extended dataset, metadata, and survey-annotation code available either referencing the dataset paper (if also accepted), or directly.
 - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Geburu et al. (2021))? [NA.](#) Please observe that FAIR-related aspects of the used dataset are described in the Dataset paper undergoing parallel review.
6. Additionally, if you used crowdsourcing or conducted research with human subjects...
- (a) Did you include the full text of instructions given to participants and screenshots? [NA.](#) See answers to Sec 5.
 - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? [NA](#)
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [NA](#)
 - (d) Did you discuss how data is stored, shared, and de-identified? [NA](#)