Goals, Risks, and Safety Practices in Online Labor Abuse Disclosures

Veronica A. Rivera[§†*], Tracy Li[†], Alex Ozdemir[§*],
Catherine Han[†], Zakir Durumeric[†], Elissa M. Redmiles[‡]

[§]*Max Planck Institute for Security and Privacy* [†]*Stanford University*
[*]*Georgia Institute of Technology* [‡]*Georgetown University*

Abstract

Survivors of workplace labor abuse disclose sensitive experiences online to build collective power, seek support, and warn others. Doing this without triggering retaliation or further abuse requires achieving targeted visibility: reaching supportive audiences while remaining hidden from abusers. Existing social media tools offer limited support for navigating the tensions between visibility, trust-building, and obscurity, and current threat models for at-risk groups do not account for these dynamics. We report findings from interviews with 17 survivors of labor abuse who disclosed their experiences online. Survivors disclose to different platforms and audiences to manage risk and evade adversaries. Yet, these strategies often fail to meaningfully reduce risk and can also undermine their goals. We extend existing frameworks of at-risk technology use by foregrounding targeted visibility and conclude with implications for designing systems that better support survivors of labor abuse and other interpersonal harms.

1 Introduction

Workplace labor abuse (LA) is widespread and costly in the United States [14, 115]. It includes economic exploitation (e.g., wage theft, nonpayment, breach of contract, financial scams) [7], discrimination [115], health and safety violations (e.g., resulting in injury or illness) [62], and sexual misconduct [76]. Yet, these harms are often underreported through formal grievance procedures, such as those administered by organizations' Human Resources groups (HR), government agencies, and other regulatory bodies [45,65]. Although these procedures purport to support victims of LA, they too perpetuate harm. Many forms of LA are embedded within these structures themselves. Distrust in institutions, employer retaliation, workplace consequences, and incentive misalignments between victims and organizations, further silence victims [26,65,102].

Thus, LA survivors turn to covert communication channels called *whisper networks*, to covertly disclose workplace LA [45]. Historically, these networks have existed *offline* in physical workplaces, where LA survivors share details (e.g., perpetrator names), while trying to evade detection by adversaries, such as individual abusers and complicit institutional actors [10,45,66,105]. This allows LA survivors to broadcast warnings, seek support, and collectively organize [45].

To transcend the limits of physical communities in building collective power and obtaining support, LA survivors also use social media and communication technologies to conduct these communications *at scale*. The scale and reach of the internet have enhanced grassroots organizing [30,54,88,90], warnings about perpetrators [28,37,111], and support-seeking [35,121]. Yet, this scale is not without risk. Survivors also experience retaliation and harassment from individuals and organizations while navigating the reach of their disclosures in digital channels. These consequences were especially salient following the #MeToo movement in 2017 [1,50].

LA survivors seek *targeted visibility*: to have their disclosures widely believed and acted upon by those who could provide support, participate in collective action, or otherwise benefit from the information, while avoiding individual and institutional adversaries. To vet confidants and build trust in physically-located whisper networks, survivors have historically used coded signals delivered via indirect references and body language [45]. Yet, these approaches are near-infeasible in large online channels. In these channels, LA survivors make sensitive disclosures to broad audiences, sometimes on the order of thousands of people, where vetting and trust-building are challenging. Furthermore, the technologies they use to make these disclosures (e.g., social media, email, messaging applications, and collaborative documents) have limited secu-

USENIX Symposium on Usable Privacy and Security (SOUPS) 2026.
August 23–26, 2026, Hannover, Germany.

rity and safety guarantees that balance the competing needs between visibility, trust-building, and obscurity [35].

Targeted visibility is a contextual risk factor that puts online disclosers of LA at risk of digital safety threats. Within Warford et al.'s framework of at-risk users, targeted visibility is most closely related to prominence [119]. However, the idea of prominence—that individuals (e.g., celebrities, content creators, activists) are targeted for their notability in a population—is framed as a static and inherent attribute of the at-risk individual. In contrast, targeted visibility is dynamic and tied to the spread of information. It emerges from an active safety calculus where survivors weigh disclosure goals against potential harms. The choices they make shape both the reach of their disclosures and their exposure to risk.

Similar disclosure tensions have been previously studied among other groups, including online daters, content creators, and disability activists [15, 21, 101]. This work identifies how such groups navigate these tensions within single platforms, often by adapting existing platform features for their goals. We build on this by considering how LA survivors *construct* audiences *across* multiple platforms, revealing design requirements for new tools, rather than adaptations of existing ones.

The targeted visibility sought by LA survivors also resembles the goals and risks that activists balance, but key differences necessitate a distinct threat model. Two key differences are the adversary capabilities and their goals for their online communications. While activists, especially political activists, face nation-state adversaries with extensive technical, political, and social capabilities, LA survivors face locally powerful adversaries (e.g., managers or employers) who have economic and reputational leverage, but lack the resources and technical expertise of a nation-state. Activists often seek broad reach to drive collective action [3, 16, 44]. As we find, LA survivors sometimes also seek to mobilize, but also have other disclosure goals. These span influencing specific individuals, reforming institutions, and obtaining personal resolution. This combination of goals is not documented in prior work. The tensions that arise from pursuing individual and collective goals simultaneously, in shared groups, and under adversarial conditions, reveal a novel design space.

To characterize this design space, we study the threat model introduced by targeted visibility in the context of LA online disclosures, responding to the call of Usman and Zappala for human-centered threat modeling that considers "other factors of context" [116]. We conduct semi-structured interviews with 17 labor abuse survivors who have disclosed their experiences online, focusing on four research questions. We refer to the communication networks our participants use as **digital whisper networks (DWNs)**.

RQ1. **Structure:** What types of technologies (e.g., their affordances and governance structure) do participants use to disclose their experiences? (Section 4)

RQ2. **Goals:** What goals do participants have when sharing their experiences? How do their goals shape which technologies they use? (Section 4)

RQ3. **Risks:** What risks do participants experience or perceive when using DWNs? How do risks relate to the disclosure and/or the technology? (Section 5)

RQ4. **Mitigations:** What decisions do participants make during disclosure? How do these decisions relate to their goals and risks? (Section 6)

We find that LA survivors have three goals when disclosing their experiences in DWNs and that the vast majority seek to broadcast their experiences to as many people as possible. Participants fear, and try to self-protect against, four risks: uncontrolled-resharing of their disclosure beyond its intended audience, retaliation by an abuser, reputational damage, and networked harassment. Targeted visibility amplifies tensions between participants' goals, risks, and the intended impact of their disclosure. Participants' self-protective strategies often require them to prioritize one objective over others despite uncertainty about the consequences of these tradeoffs. For instance, some participants obfuscate details about their abusive experience (e.g., name of perpetrator) and their own identity. Yet, such information is often necessary for the community to trust them and act on their disclosure.

We conclude with a set of recommendations for sociotechnical interventions that could help LA survivors decide what, to whom, and how to disclose that draw on human factors and cryptographic techniques. While our results and implications focus on LA survivors, we discuss how our work can be extended to other at-risk groups that may also seek targeted visibility when making online disclosures.

2 Related Work

In this section, we discuss related work on (1) secure communication among at-risk groups and (2) online disclosures.

2.1 Secure Communication

Participants in a DWN hope to communicate securely by having their online disclosures evade adversaries. At first glance, this suggests a connection to the huge literature on secure communication via encryption [23, 31, 75, 87, 104].

Some of this literature studies encryption for at-risk groups, including activists, journalists, lawyers, and politicians [13, 73, 98, 119]. For instance, Dafalla et al. found that activists in the Sudanese revolution favored E2EE messaging apps for communication [16]. And Albrecht et al. found similar results for protesters in Hong Kong [3]. Other work has developed private communication tools for specific groups [78]. For instance, McGregor et al. [73] surfaced limitations of encryption for journalists, which Lerner et al. [64] built on to design an encrypted email client for journalists and lawyers.

More recently, Di Salvo analyzed an encrypted system for journalists and whistleblowers to communicate [22].

This literature focuses on a computationally powerful network adversary and it assumes that secret messages are **not** sent to the adversary as plaintexts. We consider an abuser (e.g., a manager or employer) that is computationally weaker, but might attempt to infiltrate the online communities where disclosures are made and observe messages as plaintexts.

2.2 Online disclosures

Support-seeking and collective organizing. Prior work has studied support-seeking in online communities, including around sensitive topics like health [80, 81, 125], sexual abuse [4, 47, 79, 121], and precarious work [6, 100, 124]. This work finds that people turn to online communities for support when institutional resources are inaccessible or fail to meet their needs. For instance, Gui et al. [33] found that limited access to healthcare providers and offline social support led women to seek medical support from online communities during pregnancy. Similarly, Barakat et al. [6] found that sex workers turn to online communities to share experiences with exploitation and seek advice. While some of this work investigates how individual people seek support online, other work has studied how communication technologies support broader social movements around sensitive topics and abuse [24].

Online community and action have been studied among gig workers [103, 114, 124]. This work finds that gig work's distributed nature, low pay, isolation, and precarity drive workers to seek community online [34, 67, 95, 124]. They use both mainstream social media and community-owned technologies to combat power imbalances [97], wage theft [40], privacy concerns [100] and algorithmic opacity [42]. Yet, little is known about online disclosures of labor abuse in more traditional work, which is our focus. We also consider how workers navigate both the goals and risks of disclosures.

Anonymity online. Online support-seeking in sensitive/stigmatized contexts has pronounced social risks, which create barriers to disclosure and help-seeking. Prior work has studied anonymity [17, 46, 122] and pseudonymity [25, 61] as protections in these contexts. Obscuring one's identity can protect against socially-induced risks [17] by creating greater social distance between people and their network [2]. People manage these boundaries by participating in anonymous social media sites, creating throwaway accounts [5, 61] and maintaining distinct online personas [36, 41, 70, 113]. Kang et al. [48] found that some of the adversaries people try to evade online through anonymity are organizations and employers, the two primary adversaries we consider in our work.

Prior work also identifies the limitations of online anonymity, from both security and behavioral perspectives. Regarding security, it is easy to re-identify someone by linking information like gender, age, diagnoses, zip code, and birth date [57]. Yet, people may underestimate these risks [57, 59]. Thus, recent work supports more privacy-informed disclosures [56, 57]. We extend this by identifying *goals* for online abuse disclosures that must be considered in addition to the risks. On the behavioral side, online anonymity weakens social ties and increases harassment and toxicity [49, 68, 82, 118]. This has sparked new design ideas that balance credibility and privacy, such as meronymity [107]. Our work suggests concrete problems these ideas might apply to.

3 Methods

We interviewed 17 adults in the US who have shared about labor abuse in a DWN. In this section, we describe our participant selection process, interview procedure, data analysis, and the limitations and ethical considerations of our study.

3.1 Participant Recruitment

We recruited participants in the US who have shared a labor abuse experience in a DWN. Because DWNs often use private channels, recruitment is challenging. Thus, we used broad recruitment strategies. We shared a recruitment flier (Appendix D) through our social media accounts (X, LinkedIn, Facebook, and Reddit) and asked moderators of closed labor-related online groups to share it in their communities. We also asked professional contacts to advertise within their professional networks and at relevant events. We employed snowball sampling, asking participants to share information about our study with others in their communities. Finally, we recruited some participants via Prolific. In total, we recruited 14 participants through social media, networking, and snowball sampling and 3 participants through Prolific.

Potential participants did a brief sign-up survey, which we used to filter out those who did not meet our criteria. We invited participants to the study if they indicated having experienced some type of labor abuse, per our definition (Section 1) *and* having shared it with others via social media and/or messaging apps. To ensure we did not exclude users of DWNs who do not know or associate with the term "DWN," we did not use the term in our recruitment materials and interviews.

When selecting interviewees we tried to balance across self-reported gender, race, and work industry. In the end, our participants represent nine industries: healthcare (4), service (3), the arts (2), technology (2), academia (2), industrial (1), hospitality (1), law (1), and journalism (1). Appendix C summarizes their demographic and professional backgrounds.

We interviewed participants until saturation. We did not seek saturation in each industry, but rather across the whole dataset since cross-industry comparisons were not our goal.

3.2 Data Collection

We conducted 18 interviews on Zoom from April to August 2024, excluding one that was topically irrelevant from the dataset. The lead researcher conducted all interviews while 1-2 members of our team took notes. Interviews were in English and lasted 30-60 minutes. All interviews except one were audio-recorded with participant consent. For the other, the participant (P6) consented to note-taking, so two team members took detailed notes for our data analysis. We compensated participants with a $30 Amazon gift card.

In the interviews, we first asked participants to describe what they do for work and to describe the negative experience they had indicated sharing with a DWN in the sign-up survey. We did this (1) for more details on participants' abuse experience, and (2) to confirm that participants met the study criteria. We then asked participants questions on the following topics.

Membership, goals, and community structure. We asked participants about the types of communities where they disclose and their reasons for sharing with those groups. We then asked about the composition and structure of those communities: their relationships with other members, the digital platforms that support the group, and the norms that govern how members share and engage with content.

Perceived risks of sharing. Next, we focused on participants' perceived risks in sharing via DWNs. We asked about both *hypothetical concerns* and *previously experienced harm*.

Trust. We asked participants what factors they use to determine whether another member in the network (and the content they share) is trustworthy. We also probed about community guidelines for handling bad actors within the network to further understand community structure and governance.

Security. Finally, we asked participants questions about theirs and the community's defenses against perceived risks. We also asked about their overall feelings of safety in DWNs.

The full interview protocol is in Appendix A.

3.3 Data Analysis

We qualitatively analyzed the transcripts using codebook thematic analysis [8]. We began with a few deductive codes based on our research questions, interview questions, and prior theoretical affinity for the goal-risk-behavior paradigm. Then, three researchers independently coded two transcripts inductively, met to discuss differences, wrote a codebook, collaboratively coded a third transcript, and updated the codebook. Then, two researchers independently coded the remaining transcripts, met to resolve all differences, inductively updated the codebook, discussed emerging themes, and ultimately agreed on the codes for each transcript. We do not report inter-rater reliability, following the guidance of McDonald et al. [72]. In our study, codes were a "process" to surface themes; they were not the "product", and all coders coded all transcripts. Agreement was reached for each transcript.

We iteratively sorted codes into thematic categories by using memos, revisiting data, and engaging all researchers in regular discussion [83]. The themes describe participants' *goals* for joining and sharing with DWNs, their perceived *threats*, the *mitigations* they use for those threats, and how their goals and perceived threats influence *platform selection*. See Appendix B for the final codebook.

3.4 Limitations

Our study has several limitations. First, our sample skews toward participants who self-selected into the study and engage in more visible DWNs. Our results do not extend to people who do not disclose to DWNs. And, within those that do disclose, our participants may represent a less risk-averse group. We hypothesize that non-disclosers perceive greater risks, fewer benefits, and reason about them differently. People who do not disclose in DWNs may also choose other venues, on or offline, to disclose. Identifying these in future work would be valuable for designing technology-based interventions.

While we reached thematic saturation, our sample size is too small to generalize, as with all interview studies, or to draw meaningful comparisons across specific dimensions. Thus, we do not differentiate our results by industry, platform type, abuse category, or participant background. Instead, we characterize the unified goal/risk space across several labor industries. Given the sensitive nature of the interview topics, participants may have misrepresented some details. To reduce social desirability bias, we framed participants as experts of their own experience at the start of each interview. Finally, our study does not include people who are discussed in DWNs, who may also experience harm (e.g., defamation). This perspective may be important for future work.

3.5 Harms

To contextualize our findings, here we explain the kinds of abuse that our participants disclosed in DWNs: economic/professional, psychological/social, and sexual/physical. The perpetrators were individual and/or institutional adversaries with varying power and capacity for retaliation.

Eight participants discussed economic and professional harms, in which colleagues or supervisors appropriated their labor (P4), withheld pay (P1, P2, P5, P8), or took credit for their work (P7, P14). Eight participants also reported psychological and social harms by the same types of actors, including psychological abuse (P6, P12), non-sexual harassment (P13, P14, P16, P17), and race (P1, P8, P13) and gender-based discrimination (P17). Finally, seven participants described sexual and physical harms from individuals or systemic organizational failures, like unwanted touching (P2, P3, P5, P11, P15), sexual harassment (P1, P11), and workplace hazards (P9).

Behavior and Goal	FBG	MG	BAF	PVT
Broadcasting:				
preventing harm	✓	✓	✓	
documentation	✓			
relieving frustration			✓	✓
Support-seeking:				
validation and empathy	✓	✓		
advice	✓			
Organizing:				
mobilizing action	✓	✓	✓	

Table 1: Participants find that different goals align better with different platform modalities. The modalities are: feed-based group (FBG), messaging group (MG), broad-audience feed (BAF), and publicly-visible threads (PVT).

4 Goals of Digital Whisper Network Behaviors

Formal LA grievance procedures are often ineffective and harmful [37,60,67]. Our participants feared formal reporting's pitfalls, such as social shame, lack of documentation with survivor attribution, and physical or financial retaliation; 13 participants found formal reporting to be nonviable. Thus, they turned to DWNs, strategically constructing audiences across social and messaging platforms to achieve the targeted visibility they believed necessary to meet their goals.

In this section, we systematize DWN behaviors into three classes: (1) broadcasting abuse experiences, (2) soliciting support, and (3) organizing community. We also characterize the underlying *goals* that motivate each of these behaviors, and connect these goals to participants' use of four platform modalities: feed-based groups, broad-audience feeds, messaging groups, and publicly-visible threads. Table 1 summarizes the connection between behaviors, goals, and platform choice.

To protect privacy, we did not require participants to name specific platforms. Thus, we do not name specific platforms, except as examples.

4.1 Broadcasting Abuse Experiences

Traditional whisper networks create an informal, "safe" environment in which to disclose abuse [45, 89] to inform other vulnerable individuals. In line with this, we find that 15 participants used DWNs to *broadcast* their experiences. We identify three underlying goals that motivate this.

Preventing harm to network members. Twelve participants sought to warn others about an attacker that they might interact with in the future, similar to the goals of non-digital whisper networks [45]. These warnings were particularly important to participants whose professional networks were highly connected and therefore largely reputational. For instance, P15, who works in entertainment, shared their experience and their attacker's name via a social media feed where they had thousands of followers. *"[I was] super connected with a lot of other models and photographers online...if people were to work with this person in the future, at least [they would be] aware [...] and could then make a judgment call whether or not to work with that person again."* Participants varied in how broadly they hoped their warning would reach, especially as they weighed their disclosure goals against potential harms and this shaped which platform types they used to construct their audiences. Yet, regardless of the platforms they used to disclose, participants all shared a desire to "save" others by sharing information they would have liked to know themselves before the abuse (P4, P6), which was viewed as an *"obligation"* (P4) or duty to the broader community.

Seven participants sought to warn as many people as possible, thus leveraging platform structures that facilitate connection to large audiences via feeds. This enabled participants to *"make a public statement about a perpetrator of violence and harm in a way that otherwise someone might not know about them"* (P8). The majority of these participants leveraged *feed-based groups*: open or closed communities where members can access or contribute posts to a shared feed (e.g., a closed Facebook group where moderator approval is required to join). Participants chose affinity-based groups organized around shared characteristics such as gender, race, religion, or interest, reasoning that these groups would allow them to reach those who would be likely to encounter the perpetrator and experience similar harms. *"I think that was why I could trust the group [of cleaning staff], because I felt like someone in the group. We're like family, we just share ideas."* (P4)

Others prioritized broader reach by sharing via platforms that aggregate posts (e.g., X, Meta). This allowed participants to warn people who might not be in their immediate social networks, since unlike feed-based groups, *broad-audience feeds* do not have a clearly defined membership boundary. Participants thus found them useful for *"mak[ing] a public statement about a perpetrator of violence and harm in a way that otherwise someone might not know about them"* (P8).

Five participants wanted to warn specific people and chose *messaging groups* for direct or small group communication (e.g., text, WhatsApp, Facebook Messenger, or direct messages on Slack and Instagram). *"I used slack and whatsapp to communicate to the people I work with that I don't feel comfortable working with this person."* (P7) This goal was especially salient among people who worked in physical workplaces with close-knit communities. *"We should not withhold anything that might be of concern regarding work. We should not cease to make comments or posts...that will make the group, the department, and the organization as a whole, a better and safer place for all of us."* (P9)

Documenting firsthand harms. Three participants used features of feeds (such as wide-reaching visibility, re-posting functions, and archived content), to semi-formally document their abuse. P10 explained that they posted in a feed-based

group in case their attacker pursued legal action: *"Say like a huge case comes up...I felt like I needed to share to have a witness."* Even if a post did not explicitly request a witness or discuss legal action, participants felt that – by sharing with a large audience – they could better find a witness in the future.

A few participants saw documentation as an explicit step towards legal action. One hoped that by participating in a DWN, they could better understand the feasibility of composing a class as a basis for legal action (P7). The financial cost of formal legal support made participants turn to DWNs for this kind of advice. *"[DWNs] can help us get a real court case...because most of the people that I know on that group are not actually rich. They're just comfortable...it's not easy to pay $100 per hour."* (P2)

Relieving emotional frustration. Sharing negative experiences is known to be therapeutic and key in survivors' recovery, especially through writing [5, 86]. Corroborating this in the context of labor abuse, we find that seven participants disclosed via DWNs seeking to relieve frustration. For them, the disclosure was cathartic, irrespective of others' responses. *"I was just ready to let it out."* (P11) These participants chose platforms with large audiences to *"tell anybody that would listen."* (P4) These included feed-based groups and broad-audience feeds, as well as *publicly-visible threads*: discussion spaces that can be viewed by any internet user without platform login, such as a public subreddit. Those who sought responses from their audience in the form of emotional support and guidance prioritized feed-based groups where they could better connect with their audience over shared characteristics. We discuss these in Section 4.2.

4.2 Soliciting Support

All 17 participants disclosed abuse to DWNs for emotional support (e.g., validation) or for specific advice on managing abuse in their workplace. This corroborates prior work on the goals of online, post-abuse support-seeking [5, 6, 29, 35, 121]. In addition to corroborating this prior result for *labor abuse*, we extend prior work by connecting people's support-seeking goals to their decisions about which platforms to use.

Validation and empathy. Thirteen participants sought emotional support. Eight sought others with similar experiences, to know they are not alone: *"[I was] looking for, has anybody else [had] these experiences, like am I crazy? Just kind of looking for validation"* (P16). Similarly, P8 explained that others' experiences helped them see that they were *"not the only one...suffering like this."* Participants explained that similar experiences helped them feel less hurt (P13) and *"get clarity"* (P3). *"We might not share the same story, but we can understand what other people are going through."* (P2) Five participants sought sympathy: *"[DWNs] are like a small group of family that really want to help each other...[when] someone just went through a really big bullshit."* (P5)

Participants seeking validation and empathy disclosed via platforms where they could connect with others over shared experiences. Thus, they shared in small messaging groups with known contacts or in feed-based groups based on a shared identity characteristic or interest that could facilitate support-giving. *"I wanted a community that would actually resonate well with me, a community that makes me feel that sense of belonging. So I just went to a black support group."* (P12) Small messaging groups connected participants with close online or in-person friends. *"Hey, just wanted to share if you have time. This thing happened to me...I just want to talk it out."* (P6) Larger feed-based groups balanced the desire for connection with the desire to obtain raw, unfiltered empathy from strangers. *"I love when I get to hear of honest reaction, honest sympathy"* (P2).

Job and Situational Advice. Fourteen participants disclosed abuse in the hopes of obtaining advice on how to handle particular situations, even post-facto. *"Most of our stories is asking questions. "Okay, here's what went down. Here's what I did. Do you think what I did was right? Do you think I could have done something else? Did I handle this correctly?""* (P17) Several reported feeling guilty for what they had experienced and wanted others to tell them *"what [they] could have done to remove this whole situation"* (P3) so that they could *"be more careful next time."* (P9)

Participants sought advice mostly from feed-based groups that began as online spaces to share job advice, like leads and/or resources, and over time, people came to see them as relevant spaces for asking for advice regarding abuse. P6, who created such a community, explained that over time, as employees started leaving the company, group members opened up to discuss negative experiences with their manager. Similarly, P16 explained that one community was first about classes, but later served a dual purpose to *"communicate like, "Hey, this is happening to me, is anybody experiencing this?""*

Like those seeking validation and empathy, participants seeking advice also appreciated the diversity of opinions that feed-based groups afford. *"That's the thing about the platform. You get to see different people's opinions to different things...not everybody will tell you what you want to hear...You have the positive replies, you have the negative replies"* (P3). While the majority of participants seeking advice shared in feed-based groups, P12 and P17 felt these groups were too restrictive in the audiences they enabled. Thus, they shared more widely, in publicly-visible threads. P17 wanted to support passive observers who might feel that membership in a DWN is too risky (P17): *"[The group] being public does help for those people who are just trying to get their feet wet, and they don't want to share things necessarily about themselves. But they're looking for advice."* (P17) P12 only wanted opinions from strangers because *"I wanted a genuine concern"*. They worried that people they knew might give them a *"biased outlook of the whole situation."*

4.3 Organizing Community

All but two participants disclosed abuse to DWNs to build community, take collective action, and practice agency over their situation. In particular, participants wanted to mobilize action to improve the safety of their own and others' work.

Mobilizing Action. Participants hoped their sharing would drive systemic change. They sought more than just retribution against their abuser; they sought to lead and be part of a movement addressing systemic issues in their physical communities. *"It's never a one-off thing...I'm trying to show that we're all having this experience, and we need to mobilize as a community about this experience."* (P8) Formal reporting systems are not designed for this. Thus, to make *"the department and organization a better place and a safer place for all of us"* (P9), participants turned to DWNs.

Feed-based groups, messaging groups, and broad-audience feeds were all used to mobilize action. Participants explained that within their immediate physical communities, it is unclear whether abuse is happening to one or multiple individuals (P6). Thus, large-audience platforms, such as feed-based groups and broad-audience feeds, were useful in connecting with others with similar experiences. *"We're all spread across the country, but we know that as a group we're better, we act as one voice...we all know if we act individually, things are not going to change."* (P17) Smaller messaging groups served to coordinate offline meetups to organize in more intimate settings, often after participants have first identified potential connections through larger online communities. For instance, P2 explained that some people in a DWN they are a part of had been meeting in person *"hanging out and just comforting each other."* These offline meetings can help participants open up, even if they are reluctant to share details online (P6).

5 Disclosure Risks

DWNs also expose participants to harm. Now, we show how DWNs *reproduce* some risks of formal reporting, and *introduce* new risks exacerbated by their digital platforms. We find four key risks, and we highlight how resulting harms arise from both participants' *goals* and the *platform types* they use to disclose to DWNs.

While participants primarily discussed hypothetical risks/concerns, some of these were also *experienced harms*, either by participants or someone they know in a DWN. For each risk, Table 2 reports the number of participants who fear it, have experienced it, or know someone who has.

Harm	Feared	Heard-of	Experienced
re-sharing	9	2	0
retaliation	8	5	2
reputational damage	6	1	0
harassment	10	5	4

Table 2: For each risk, the number of participants who fear it (hypothetically), know someone who has experienced it, and have personally experienced it. These counts are cumulative.

5.1 Uncontrolled re-sharing

Attack vector. Prior work has shown that technologies can harm users' autonomy in both digital and physical settings [12, 95]. Such harm arises both from platforms that limit or distort the information available to users [43, 112] and from devices that enable direct control over users' physical actions [27, 110]. Nine participants feared harms to their autonomy stemming from uncontrolled re-sharing of their disclosures in DWNs. Sharing information digitally allows anyone with access to the data, including someone "authenticated" within the network, to aid an attacker. *"They might take the whole text, screenshots, and everything, and send them back to whoever you had a disagreement with."* (P10) Here, the scale and reach of digital platforms (especially feed-based groups, broad-audience feeds, and publicly-visible threads) enabled information to surpass its intended bounds. Even when disclosures were targeted via messaging groups, participants noted that digital permanence made it difficult to control spread. *"You may know who you're talking to, but you never know who they're going to talk to."* (P1) Goals to broadcast disclosures and/or collectively organize exacerbated risks of uncontrolled resharing.

Harm. Participants worried that if their disclosures went viral, their intended message could be harmfully misinterpreted, either by their abuser or by another adversary. *"When people start sharing screenshots, they're always telling their version of what has happened...If [the context of the story] is not shared completely, it can influence people."* (P7) P8 agreed, worrying that *"someone will literally take what I'm saying and manifest violence out of it."* Uncontrolled re-sharing threatens autonomy and also contributes to fears of retaliation (Section 5.2) and reputational damage (Section 5.3).

5.2 Retaliation

Attack vector. Eight participants feared that the original abuser might discover their participation in a DWN, either through uncontrolled re-sharing or otherwise, and retaliate physically or financially. Similar to other types of interpersonal abuse, like intimate partner violence [11], participants feared that the platforms that facilitate DWNs may serve as an attack vector for perpetrators. The same platform features that enable the benefits of DWNs (Section 4) also create this risk. Participants explained that sharing identifying details about their experience (like their workplace, profession, or

location) often increases their credibility within the network and builds trust, making others more likely to act on their disclosure: *"We share information about where we work...we share everything that has happened...most actually use their real name...[Otherwise] you might be talking to a pedophile, someone who has committed some really sick crime."* (P5)

Harm. Participants noted the risks inherent to sharing this information. *"If there's enough information in those posts, they know that you're talking about them and they might come and hunt you down in real life."* (P17) Some participants faced direct physical threats for disclosing. *"I was warned by the person who assaulted me, that I shouldn't report to anyone or he would get me killed."* (P2) Others feared they could lose their job, or face challenges finding a new one, compounding existing financial insecurity: *"The ultimate reason they chose not to hire me was...traced back to me expressing myself on socials about things that happened to me."* (P8)

5.3 Reputational damage

Attack vector. Prior work has found that survivors may hesitate to disclose interpersonal abuse to family, friends, or support services due to fears of reputational harm [85]. Our participants also expressed this concern; furthermore, we find that disclosing labor abuse *via DWNs* heightens these fears. Six participants described concerns about reputational damage from sharing their experiences digitally, especially if close friends, family, or colleagues were unknowingly in the same online communities, or followed their social media accounts. For instance, P15 worried her family might learn about the sexual abuse she experienced at work through social media: *"[I shared] NSFW stuff that I wouldn't talk about with my family...I wasn't as open with them in general at the time."*

Harm. Participants feared lasting damage to their relationships with others. Some worried that those close to them *"might see me in a different way."* (P3) Others worried that if work colleagues found out, especially if they disclosed on workplace platforms like Slack and LinkedIn, it would be *"an uncomfortable situation...to be acting like that in a public professional space."* (P7) Some who identified with a marginalized group felt an increased risk of reputational damage, credibility loss, and diminished impact. *"I'm disabled, I'm queer, and Black. When I say these things, I look angry, and I'm perceived as coming at this from anger."* (P8)

5.4 Harassment

Attack vector. Prior work has shown that despite their benefits for support-seeking, online communities can also expose members to harassment [52, 69]. Ten participants saw this risk in DWNs: *"You want a support system. You don't want people telling you, "Oh, you're nasty! You could have avoided this"."* (P3) The scale and reach of digital platforms, and the challenges of moderating large networks, make it difficult to establish community norms shared by all members and for prospective members to discern what those are. Therefore, participants were often unsure how others would respond to their experiences and if they would align with the community's values. For example, participants wondered whether *"people will use your story to make jestery"* (P2) or whether *"half of the people are going to say this is ridiculous."* (P1) In Section 6.1, we discuss how some of participants' safety mitigations can exacerbate the risk of networked harassment.

Harm. This risk illustrates Marwick's *morally motivated networked harassment*: when a member of an online community accuses a target of violating community norms, ultimately resulting in coordinated harassment at scale within the entire community [69]. In DWNs, networked harassment threatens further harming those who have the most to gain from these communities – the people marginalized by formal reporting systems, including women, people of color, and those facing work precarity [55, 71, 109]. For instance, P4 explained that after her disclosure, *"someone said I was lying, that the admins should take me out [of the group] and that maybe I am the scammer"*. Sometimes participants hesitated to disclose via DWNs, fearing *"cancel culture"* (P1), and *"being attacked by white supremacists"* (P8) or *"chastised by [co-workers]"* (P9) who disagreed with their stories.

6 Mitigations and Goal/Risk Tensions

Despite the risks, participants still found value in disclosing via DWNs: *"I had a very unique voice that people wanted to hear and needed to hear. And so that's what motivated me to stay, even when the online space became a bit more negative and elitist."* (P9) Participants shared various ways they try to mitigate the risks to use DWNs more safely.

In this section, we describe participants' mitigations and highlight where their strategies are in tension with their goals and/or risks. We summarize these relationships in Table 3. Ultimately, we find that many mitigations can make goals (Section 4) harder and/or exacerbate risks (Section 5). We present design recommendations to address some of these tensions in Section 7.1.

6.1 Navigating Disclosure and Obfuscation

We find that the degree of information disclosure differed greatly across participants. They must navigate complex tradeoffs when choosing whether to disclose personally-revealing details, the identity of their attacker, and their own identity.

Fourteen participants said that when they disclose, they obfuscate details about their experience, such as the name of the perpetrator, workplace, and others involved. *"I was very careful not to include certain details that can be traced back to me.*

		Obfusc.	Anon.	Dist.	Deleg.
goals	broadcast	✗	✗	✗	✓
	support			✓	
	mobilize	✗	✗	✗	✓
risks	re-sharing				✗
	retaliation	✓	✓		✗
	reputation	✓	✓	✓	
	harassment	✗	✓/✗		

Table 3: A summary of the synergies and tensions in our findings between participants' mitigations (columns) and goals/risks (rows). A ✓ indicates a synergy (supports a goal or reduces a risk), while an ✗ indicates a tension (impedes a goal or amplifies a risk). A blank indicates that we found neither. Mitigations include: obfuscation (Obfusc.), anonymity (Anon.), sharing with an audience at greater social distance (Dist.), and delegating audience management to group administrators (Deleg.).

Not saying the name of my clinic, not saying what city I'm in, especially about the situation." (P16) Participants believed that these omissions could protect them from retaliation and reputational harm by making it harder for attackers to identify them. While most participants obfuscated to protect against a perpetrator within the group or an insider who might leak details, some also sought to protect themselves from search engines: *"I try not to share geographically locating information, identifiable information, or names and places...I get so many views, my socials are the first thing that comes up when you Google my name. So I just don't want to create that search engine optimization link."* (P8)

Six participants shared anonymously. Participants believed that separating their identity from their account would protect them not only from retaliation and reputational harm, but also from networked harassment. For instance, P2, who has experienced networked harassment, explained that they share anonymously because *"people can really hurt you if they actually know you"*. Similarly, P3 explained that sharing anonymously *"is the only way you can be yourself...I couldn't share the whole thing because people would have done the blaming thing of 'this could have been avoided'."*

While obfuscation and anonymity have clear benefits, they are also in tension with other goals and risks. Prior work has found that obfuscation and anonymity can increase harassment, limit accountability, and decrease community participation [5, 6, 58, 118, 121]. Among our participants, we find that anonymization and obfuscation exacerbate networked harassment. Anonymity may protect participants from offline risks, such as retaliation and reputational damage. Yet some experienced greater online harassment in anonymous spaces, echoing prior work on online flaming [59]. Participants thus faced a choice between two kinds of exposure: offline risks (retaliation, reputational harm) versus online risks (networked harassment). For instance, P2 described hurtful comments they received when disclosing LA anonymously via a publicly-visible thread: *"Sometimes it's depressing to see most of the hateful things you see. Reddit is harsh...it's the reason I like the [non-anonymous] private group because any hateful comment would definitely be out of the group."*

We also find that anonymity and obfuscation are in tension with participants' broadcasting, support-seeking, and organizing goals. Anonymity and obfuscation undermine credibility and trust, which are requirements for warnings and collective action efforts to succeed. For instance, P16 decided which stories to trust and react to based on whether *"[the other person] shares things that most people would find confidential."* These details are used as a trustworthiness metric in many DWNs, in lieu of the coded signals used in traditional offline whisper networks [45], because *"you don't have that level of detail about these things unless you are truly going through them."* (P17) Anonymous or seemingly unreliable posts impede the trust that sustains community safety: *"I made sure not to miss anything that might want someone else to like doubt, whatever it is, that I posted or commented."* (P9) We discuss some ideas to support privacy tradeoffs in Section 7.1.

6.2 Managing Audience Access

6.2.1 Audience Social Distance

Twelve participants engaged in audience management, using characteristics about their audience to to try to control who had access to their story initially, mostly in feed-based groups.

Participants considered the social distance between themselves and their audience. In addition to avoiding LA perpetrators, participants also avoided offline contacts. Seven avoided sharing with networks with close contacts, including friends, family, or immediate coworkers. They did this for two main reasons. First, they feared sharing with those closest to them and *"being seen in a different light"* (P3) especially when their experiences do not fall into societal norms of abuse: *"I didn't want to share with friends and family because I felt like a disgrace. If friends and family actually heard what had happened it would be like, oh, he isn't man enough"* (P2).

Second, some participants believed that connections that were too close might not offer relevant advice either because their closeness could bias the advice, or because they might not have the right context: *"I didn't want to share this with my family members. Like what would they tell me about it? They are not even working in this space?"* (P7) When sharing in feed-based groups, participants sought DWNs that felt like *"neutral ground,"* (P3) where others had enough shared context to understand and for the information to be relevant and actioned upon, but not enough proximity to affect their offline relationships. In the absence of formal vetting mechanisms, participants used shared identity or loose ties as proxies for

trustworthiness, disclosing in groups built around common interests, or where they knew someone distantly but had no close contacts: *"I wanted a community that would actually resonate well with me, a community that makes me feel that sense of belonging."* (P12) This creates a core tension. The ideal audience excludes perpetrators and sometimes close contacts, while still reaching relevant people.

6.2.2 Delegation

The scale and reach of feed-based groups, broad-audience feeds, and publicly-visible threads make audience management challenging. While participants can make decisions based on characteristics of the community as a whole, they are unable to vet individuals, especially in communities with hundreds to thousands of members. This makes it challenging for participants to both protect their safety and construct audiences that are large enough for broadcasting and mobilizing. Thus, DWNs often delegate admissions decisions to a centralized moderator that operates on behalf of members using some criteria or group norm (e.g., personal attributes of a prospective member).

In DWNs, moderators implement strict safety controls, such as vetting stories before posting them on members' behalf, or collecting significant personal information from potential members. P3 noted that in one such group moderators post on behalf of members as a safety measure because *"everybody knows all the stories do not belong to the admin."* (P3) P4 described the strict verification requirements of some of the groups they are in: *"you have to drop your name, your number, your home address...you have to submit a 360 degree picture to show you're a real person. [And] a video introducing yourself and telling them what you do."* (P4)

The presence of a central admin made participants feel safer, especially in large spaces where they might not know, or trust, individual members. *"I don't know that I can trust the group [members] alone. But I know that I can trust the process of meeting people in the group. I know how I joined, and I know the rules about joining. And I know everyone in the group is invited by the admin."* (P13) Even when participants did not trust the admin or understand the governance procedures, they still trusted the existence of a process: *"I didn't know if I could trust [the admin]. All I know is I didn't have a choice. I had to tell him everything because that's part of the process. I needed to trust the process even if I didn't trust him."* (P14)

Yet audience delegation also heightens risks for individual members. Moderators make group-level safety decisions that may not account for individual circumstances (e.g., what protects one member may expose another). Human moderation is also inherently error-prone [77]. One participant described how despite collecting detailed personal data to authenticate members, moderators in a DWN they participate in were still unable to keep bad actors out: *"Someone's account got hacked, like a member of the group and we were not informed [by the moderators]."* (P4) This left members vulnerable to exactly the uncontrolled resharing and retaliation that delegation was meant to prevent. In Section 7.1 we discuss approaches for supporting moderators in authenticating members and giving individuals greater agency over their own safety, even within large online communities.

6.3 Platform Selection Considerations

Lastly, 10 participants chose DWNs based on platform characteristics they believed made those platforms safer. These were primarily interface- or marketing-level features rather than formal security guarantees.

Some participants shared via platforms and/or accounts they associated with professionalism, equating this with greater security and a more relevant, trustworthy audience. *"I found Slack more professional because there are a lot of professional groups that I'm part of and I work with, as compared to, let's say, WhatsApp, which is very much personal"* (P7). Others shared via platforms whose affordances they equated with stronger security, such as those that facilitate private groups or ephemeral posts. For example, P15 shared via Instagram stories[1] that allowed them to *"not have [the perpetrator's] name and the situation permanently on, like a regular post."*

While these affordances made participants feel safer, it is not clear that they actually reduce risk. Participants' mental models of platform security were driven by perceptions of security, amounting to a form of security theater that may increase risk by obscuring the actual tradeoffs involved. As a result, platform choice was driven more by what a platform enabled than by what it protected against.

Formal security features were rarely considered in selecting platforms to disclose to DWNs. Only P6, who works in information security, deliberately chose a platform (e.g., Signal) for its encryption properties. Others used end-to-end encrypted (E2EE) messaging apps incidentally, because they happened to suit their goals: notifying specific individuals, seeking support from a small group, or coordinating in-person meetings (Section 4). While E2EE messaging apps secure communication for many at-risk groups [16, 63, 98], our findings indicate that these apps may not be relevant to the needs of LA DWNs. These apps can protect against nation-state surveillance, which is relevant to some at-risk groups (e.g., activists, journalists, and politicians) [120], but was not a concern of the LA survivors in our study. Our participants faced two different threats: (1) harm arising when their stories spread in semi-trusted spaces and might reach an abuser or a known contact, and (2) harm arising from networked harassment and non-action when their disclosure is not deemed to be trustworthy. This necessitates distinct security requirements and design directions, which we discuss in the next section.

[1] Instagram stories are short-form content and disappear after 24 hours.

7 Discussion

Through semi-structured interviews with 17 survivors of LA, we examine the goals and risks of disclosing such abuse online, in digital whisper networks (DWNs). Protecting against these risks is challenging; disclosure goals, risks, and intended impact are often in tension and participants have few ways to evaluate the tradeoffs of their decision-making. Here we present recommendations for sociotechnical tools to better support these disclosures (Section 7.1). We also discuss how our conclusions may apply to other at-risk groups who engage in similar communications (Section 7.2).

7.1 Design Recommendations

Navigating privacy tradeoffs. DWN participants navigate the goals, risk, and impact of disclosure. Yet, the value of different decisions is hard to compare, potentially increasing risks or reducing impact. We suggest two research directions: a tool to help people make informed privacy assessments in context and a technology that could mitigate these tradeoffs.

Recent work has proposed tools to aid users in assessing privacy risks in social media self-disclosures. Such work uses LLMs to quantify privacy risk based on the size of the population matching the given information [127] and studies how this information should be presented [56,57]. Krsek et al. find that to avoid dis-empowering users, privacy risk metrics must consider the communicative intent of the disclosure and guide users on how to balance these competing constraints [57].

To address the needs surfaced by our DWN participants, we propose extending the work of Krsek et al. [57] to study how to identify quantifiable metrics related to communicative goals (e.g., goals and intended impact) in DWNs and how to guide users in making tradeoffs that balance goals and risks. For instance, future work could use engagement data from similar posts within a DWN as a proxy for a disclosure's potential impact and adapt based on user feedback. Future work could also examine how users want tradeoffs measured and visualized, and how these preferences may change depending on the disclosure goal (e.g., broadcast vs. support-seeking).

Another direction for mitigating privacy tradeoffs involving credibility and trust involves a cryptographic tool called a *ring signature* [96]. A ring signature allows a member of a public organization to sign a message in a special way that proves the message comes from someone in the organization, without revealing whom. With ring signatures, an individual could anonymously post a story about their employer or a perpetrator, while still proving to the network that they are employed at the organization. Future work could build a system for this and experiment with its usefulness in DWNs.

Threshold cryptography [19] is a different tool that might help participants in a DWN initiate collective action while reducing privacy risk. The idea of threshold cryptography is to reveal some object (like a message [18] or signature [20]) only once at least n (the threshold) users approve it. Perhaps a suitable threshold cryptosystem could allow individuals to post "encrypted" allegations of abuse to a group, in such a way that their allegations automatically decrypt only once there are n similar allegations posted. When implemented into a tool for online disclosures, this might help individuals find safety in numbers, without revealing their identity or allegations until they have formed a sufficiently large group.

Socially-informed audience management. Tensions in DWNs often stem from a desire to reach the "right" audience, particularly when users have little control over who joins the groups that define DWNs. Perhaps we could give individuals greater control over how their content spreads and to whom.

Many participants disclosed on existing social media platforms with algorithmically-curated feeds. In one direction, we propose studying new social media patterns to better support LA disclosures. For example, recent work introduces a system in which users share posts to a small group of trusted contacts, who collaboratively decide whether and how to route that content to broader audiences [126]. Future work could study whether this design pattern improves DWNs, helping members achieve their disclosure goals more safely.

In another direction, we propose systems for collective audience management in private groups (e.g., some feed-based groups). Currently, moderators individually make admissions decisions with low-quality signals about the trustworthiness of prospective members. Future work could leverage the rich social graphs underlying DWNs to study how groups' collective social capital might inform admissions decisions. For instance, a system might help moderators find existing members who could provide insight on an admission decision, leveraging social knowledge. Or, it might passively use a group's social graph to give moderators information about how "trusted" or "connected" a prospective member is to the current group, leveraging classical work on trust in graphs [53,84] or more recent work on private reputation systems [39].

Mitigating screenshot attacks. Our participants feared that screenshots of their disclosure might reach an adversary. Prior work, primarily on sex work [108] and non-consensual intimate imagery [91], has also identified taking and sharing screenshots as a challenge on digital platforms. Our findings extend this prior work by showing how they can create specific harms for LA survivors who disclose abuse online.

In this section, we discuss two kinds of technology – digital rights management and deniability – which might be relevant.

One idea is to make it harder to take screenshots or copy data. This is a form of *digital rights management* (DRM), which has been studied for decades [99], and recently applied to non-consensual intimate media [92]. DRM speaks to a core participant concern, so further research could study whether DRM could support DWNs. However, DRM also has significant limitations – workarounds can always be found [38], for example, by physically photographing the screen.

Another approach (paradoxically) is to make *fake* screenshots *easier* to forge. This is related to the cryptographic and legal concept of *deniability*. Consider some data (e.g., a screenshot) that could be used as evidence to convince someone of something (e.g., that a survivor has disclosed their abuse). Deniability aims to undermine evidentiary value, by making it easy to forge the data. If data is forgeable, it may be less convincing.

Deniability has been well-studied [9, 32, 74] and deployed to billions [106, 117] at the *cryptographic layer*, where it undermines the evidentiary value of cryptographic objects like ciphertexts and keys. Yet, this has been found to have little effect on users in social settings [123], likely because cryptographic objects are inaccessible to normal users, so these objects are not used as social evidence anyway.

A recent direction in deniability, which we call **deniable UIs**, seems more relevant. The idea is to build a chat app with a UI that enables undetectable modifications to the chat log. This has been shown to effectively undermine evidence in legal [94] and some social [93] settings. It would be interesting to prototype a deniable UI for DWNs, and experiment with it. *We caution that this would be ethically delicate.* Deniability can be beneficial, but also harmful, depending on the context [123]. The goal would be to determine whether deniability causes more benefit than harm in a DWN.

7.2 Comparisons with Other At-Risk Groups

We study the goals, risks, and protective strategies of targeted visibility-seeking when disclosing LA in DWNs. We focus on LA survivors due to their long history with whisper networks. Yet other at-risk groups who seek post-abuse support online, including sex workers, survivors of intimate partner violence or scams, online daters, and individuals who identify with a vulnerable group (e.g., LGBTQ+, people with a disability, etc.), may also participate in DWNs by actively trying to evade adversaries in their disclosures. Future work may consider how these groups navigate threats while balancing disclosure goals and community impact.

While some tensions we find in the LA context may generalize across groups, differences in adversary capabilities likely shape disclosure strategies. For instance, prior work on online daters and content creators documents similar disclosure tensions but studies these within single platforms [15, 21]. LA survivors face targeted visibility tensions across multiple platforms simultaneously, and under the additional constraint of an adversary, which shapes how they construct audiences using multiple social and messaging platforms.

Activists are perhaps the most similar to LA survivors but are themselves diverse in their goals and adversaries. Some activists face local adversaries including friends, family, and community members [51, 63], similar to LA survivors. However, activists primarily seek collective action, whereas LA survivors simultaneously seek personal support, legal documentation, and collective organizing within the same disclosure act. This combination of goals produces tensions that activist privacy frameworks do not address. LA survivors also differ from activists in their relationship to technology: political activists often eschew technology, potentially reflecting the scale and power of their adversaries [3, 16], whereas LA survivors actively seek out technology and focus on managing multiple disclosure channels and audience reach.

Beyond activists, other at-risk groups may also communicate via DWNs. These groups have adversaries with distinct approaches. Survivors of intimate partner violence face adversaries in close physical proximity, often with access to their devices. Political activists face nation-states with large technical and social resources. And scam survivors face adversaries who operate large-scale distributed campaigns. These differences may impact survivors' protective behaviors and the design requirements that follow.

8 Conclusion

We study how survivors of labor abuse use digital whisper networks as an alternative to formal reporting channels. Through semi-structured interviews with 17 participants, we find that survivors have several objectives they hope their disclosure to DWNs will accomplish. Some of these paradoxically require reaching a broad audience while managing visibility risks. Participants' decision-making for navigating these risks is often at odds with their goals for reporting to DWNs and with the norms of these communities. We formalize these risks in a threat model and provide design recommendations for future DWN reporting tools that better align user needs and safety.

9 Ethics Statement

Our study was reviewed and approved by our IRB, and we took several additional measures to protect our participants. First, DWNs seek covertness to reduce harm to their participants. Thus, we did not ask participants to name specific DWNs or platforms that they use. Some still disclosed this, in which case we redacted the names from our transcripts. Second, revisiting past abuse can be triggering; to minimize risk, we informed participants about our focus during consent, and emphasized that questions could be skipped or we could end the interview any time. Further, we did not ask too many questions about the original abuse, since our focus was reporting. Third, to protect privacy, we did not ask participants to name employers, and we redacted any PII from transcripts. We did use participant emails for scheduling and payment, but deleted them after. Fourth, to reduce harm to researchers, interviews had at least two researchers and researchers debriefed after each interview.

Acknowledgements

We thank Sasha Ronaghi, Makenzy Caldwell and Hannah Kim for their help with early study planning and data collection. Additionally, we thank Boya Wang, Yixin Zou, Kentrell Owens, Anne Newman, Leif Wenar, Shannon Abelson, Daniel Webber, Moya Mapps, Wanheng Hu, Emma Duncan, and Ben Mylius, for their feedback and advice on the paper. We also thank Daniel Votipka and Chloé Messdaghi for their support with participant recruitment. We are also grateful to the anonymous reviewers for their feedback, and to our participants for their time and for trusting us with their experienes. This research was funded by the Stanford University McCoy Family Center for Ethics in Society and by the US National Science Foundation under Grant Number 2206950.

References

[1] Alexis A. Adams-Clark, Aanandita Bali, Ananya Sharma, Elizabeth Tampke, Prachi H. Bhuptani, and Lindsay M. Orchowski. Characterizing Survivors' Descriptions of #MeToo Backlash. *Journal of Community Psychology*, 54(1):e70086, 2026.

[2] Ramona Alaggia and Susan Wang. "I Never Told Anyone until the #MeToo Movement": What Can We Learn from Sexual Abuse and Sexual Assault Disclosures Made through Social Media? *Child abuse & neglect*, 103:104312, 2020.

[3] Martin R Albrecht, Jorge Blasco, Rikke Bjerg Jensen, and Lenka Mareková. Collective Information Security in Large-Scale Urban Protests: The Case of Hong Kong. In *USENIX Security*, 2021.

[4] Nazanin Andalibi, Oliver L Haimson, Munmun De Choudhury, and Andrea Forte. Social Support, Reciprocity, and Anonymity in Responses to Sexual Abuse Disclosures on Social Media. *TOCHI*, 25(5):1–35, 2018.

[5] Nazanin Andalibi, Oliver L Haimson, Munmun De Choudhury, and Andrea Forte. Understanding Social Media Disclosures of Sexual Abuse through the Lenses of Support Seeking and Anonymity. In *CHI*, 2016.

[6] Hanna Barakat and Elissa M Redmiles. Community under Surveillance: Impacts of Marginalization on an Online Labor Forum. In *ICWSM*, 2022.

[7] Annette Bernhardt, Ruth Milkman, Nik Theodore, Douglas Heckathorn, Mirabai Auer, James DeFilippis, Ana Luz González, Victor Narro, Jason Perelshteyn, Diana Polson, and Michael Spiller. Broken Laws, Unprotected Workers: Violations of Employment and Labor Laws in America's Cities. Technical report, National Employment Law Project, New York, 2009.

[8] Virginia Braun and Victoria Clarke. Conceptual and Design Thinking for Thematic Analysis. *Qualitative psychology*, 9(1):3, 2022.

[9] Rein Canetti, Cynthia Dwork, Moni Naor, and Rafail Ostrovsky. Deniable Encryption. In *CRYPTO*, pages 90–104, 1997.

[10] Richard Edward Carter. *"It's the Only Thing We Have": Whisper Networks among Women Theatre Actors*. PhD thesis, University of Kentucky, 2021.

[11] Rahul Chatterjee, Periwinkle Doerfler, Hadas Orgad, Sam Havron, Jackeline Palmer, Diana Freed, Karen Levy, Nicola Dell, Damon McCoy, and Thomas Ristenpart. The Spyware Used in Intimate Partner Violence. In *IEEE S&P*, 2018.

[12] Danielle Keats Citron and Daniel J Solove. Privacy Harms. *BUL Rev.*, 102:793, 2022.

[13] Sunny Consolvo, Patrick Gage Kelley, Tara Matthews, Kurt Thomas, Lee Dunn, and Elie Bursztein. "Why Wouldn't Someone Think of Democracy as a Target?": Security Practices & Challenges of People Involved with US. Political Campaigns. In *USENIX Security*, 2021.

[14] David Cooper and Teresa Kroeger. Employers Steal Billions from Workers' Paychecks Each Year: Survey Data Show Millions of Workers Are Paid Less than the Minimum Wage, at Significant Cost to Taxpayers and State Economies. Technical report, Economic Policy Institute, May 2017.

[15] Yichao Cui, Naomi Yamashita, Mingjie Liu, and Yi-Chieh Lee. "So Close, yet So Far": Exploring Sexual-Minority Women's Relationship-Building via Online Dating in China. In *CHI*, 2022.

[16] Alaa Daffalla, Lucy Simko, Tadayoshi Kohno, and Alexandru G Bardas. Defensive Technology Use by Political Activists during the Sudanese Revolution. In *IEEE S&P*, 2021.

[17] Munmun De Choudhury and Sushovan De. Mental Health Discourse on Reddit: Self-Disclosure, Social Support, and Anonymity. In *ICWSM*, 2014.

[18] Yvo Desmedt and Yair Frankel. Threshold Cryptosystems. In *CRYPTO*, 1989.

[19] Yvo G. Desmedt. Society and Group Oriented Cryptography: A New Concept. In *CRYPTO*, 1987.

[20] Yvo G. Desmedt and Yair Frankel. Shared Generation of Authenticators and Signatures. In *CRYPTO*, 1991.

[21] Michael Ann DeVito. How Transfeminine TikTok Creators Navigate the Algorithmic Trap of Visibility via Folk Theorization. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), November 2022.

[22] Philip Di Salvo. Securing Whistleblowing in the Digital Age: SecureDrop and the Changing Journalistic Practices for Source Protection. *Digital Journalism*, 9(4):443–460, 2021.

[23] Whitfield Diffie and Martin E Hellman. New Directions in Cryptography. *IEEE Transactions On Information Theory*, 22(6), 1976.

[24] Jill P Dimond, Michaelanne Dye, Daphne LaRose, and Amy S Bruckman. Hollaback! The Role of Storytelling Online in a Social Movement Organization. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 477–490, 2013.

[25] Judith S Donath. Identity and Deception in the Virtual Community. In *Communities in cyberspace*, pages 37–68. Routledge, 2002.

[26] Chai R. Feldblum and Victoria A. Lipnic. Select Task Force on the Study of Harassment in the Workplace Report. Technical report, U.S. Equal Employment Opportunity Commission, 2016.

[27] Diana Freed, Jackeline Palmer, Diana Elizabeth Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. Digital Technologies and Intimate Partner Violence: A Qualitative Analysis with Multiple Stakeholders. In *CSCW*, 2017.

[28] Radhika Gajjala. When an Indian Whisper Network Went Digital. *Communication Culture & Critique*, 11(3):489–493, 2018.

[29] Christine Geeng, Mike Harris, Elissa Redmiles, and Franziska Roesner. "Like Lesbians Walking the Perimeter": Experiences of US LGBTQ+ Folks with Online Security, Safety, and Privacy Advice. In *USENIX Security*, 2022.

[30] Sucheta Ghoshal and Amy Bruckman. The Role of Social Computing Technologies in Grassroots Movement Building. *TOCHI*, 26(3):1–36, 2019.

[31] Shafi Goldwasser and Silvio Micali. Probabilistic Encryption & How to Play Mental Poker Keeping Secret All Partial Information. In *STOC*, 1982.

[32] Paul Grubbs, Jiahui Lu, and Thomas Ristenpart. Message Franking via Committing Authenticated Encryption. In *CRYPTO*, 2017.

[33] Xinning Gui, Yu Chen, Yubo Kou, Katie Pine, and Yunan Chen. Investigating Support Seeking from Peers for Pregnancy in Online Health Communities. In *CSCW*, 2017.

[34] Meghna Gupta, Arpita Bhattacharya, and Julie Kientz. " Being a Nanny Isn't Just Caregiving": An Analysis of How Nannies Seek Support in Online Communities like/r/Nanny. In *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*, pages 1–15, 2025.

[35] Naman Gupta, Kate Walsh, Sanchari Das, and Rahul Chatterjee. "I Really Just Leaned on My Community for Support": Barriers, Challenges, and Coping Mechanisms Used by Survivors of Technology-Facilitated Abuse to Seek Social Support. In *USENIX Security*, 2024.

[36] Oliver L Haimson and Anna Lauren Hoffmann. Constructing and Enforcing "Authentic" Identity Online: Facebook, Real Names, and Non-Normative Identities. *First Monday*, 2016.

[37] Bridget Haire, Christy E Newman, and Bianca Fileborn. Shitty Media Men. In *#MeToo and the politics of social change*. Springer, 2019.

[38] J. Alex Halderman and Edward W. Felten. Lessons from the Sony CD DRM Episode. In *USENIX Security*, 2006.

[39] Omar Hasan, Lionel Brunie, and Elisa Bertino. Privacy-Preserving Reputation Systems Based on Blockchain and Other Cryptographic Building Blocks: A Survey. *ACM Computing Surveys*, 55(2):1–37, 2022.

[40] Jane Hsieh, Angie Zhang, Sajel Surati, Sijia Xie, Yeshua Ayala, Nithila Sathiya, Tzu-Sheng Kuo, Min Kyung Lee, and Haiyi Zhu. Gig2Gether: Datasharing to Empower, Unify and Demystify Gig Work. In *CHI*, 2025.

[41] Xiaoyun Huang and Jessica Vitak. "Finsta Gets All My Bad Pictures": Instagram Users' Self-Presentation across Finsta and Rinsta Accounts. In *CSCW*, 2022.

[42] Mohammad Hossein Jarrahi and Will Sutherland. Algorithmic Management and Algorithmic Competencies: Understanding and Appropriating Algorithms in Gig Work. In *International conference on information*, pages 578–589. Springer, 2019.

[43] Mohammad Hossein Jarrahi, Will Sutherland, Sarah Beth Nelson, and Steve Sawyer. Platformic Management, Boundary Resources for Gig Work, and Worker Autonomy. In *CSCW*, 2020.

[44] Haiyan Jia and Eric PS Baumer. Birds of a Feather: Collective Privacy of Online Social Activist Groups. *Computers & Security*, 115:102614, 2022.

[45] Carrie Ann Johnson. The Purpose of Whisper Networks: A New Lens for Studying Informal Communication Channels in Organizations. *Frontiers in Communication*, 2023.

[46] Adam N Joinson, Carina B Paine, et al. Self-Disclosure, Privacy and the Internet. *The Oxford handbook of Internet psychology*, pages 237–252, 2007.

[47] Nur Shazwani Kamarudin, Vineeth Rakesh, Ghazaleh Beigi, Lydia Manikouda, and Huan Liu. A Study of Reddit-User's Response to Rape. In *IEEE/ACM ASONAM*, 2018.

[48] Ruogu Kang, Stephanie Brown, and Sara Kiesler. Why Do People Seek Anonymity on the Internet? Informing Policy and Design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2657–2666, 2013.

[49] Ruogu Kang, Laura Dabbish, and Katherine Sutton. Strangers on Your Phone: Why People Use Anonymous Communication Applications. In *CSCW*, 2016.

[50] Jodi Kantor and Megan Twohey. *She Said: Breaking the Sexual Harassment Story That Helped Ignite a Movement*. Bloomsbury Circus, 2019.

[51] Sayash Kapoor, Matthew Sun, Mona Wang, Klaudia Jazwinska, and Elizabeth Anne Watkins. Weaving Privacy and Power: On the Privacy Practices of Labor Organizers in the US Technology Industry. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–33, 2022.

[52] Haesoo Kim, Juhoon Lee, Jeong-Woo Jang, and Juho Kim. ReSPect: Enabling Active and Scalable Responses to Networked Online Harassment. In *CSCW*, 2024.

[53] Jon M Kleinberg. Authoritative Sources in a Hyperlinked Environment. *JACM*, 46(5):604–632, 1999.

[54] Sai Amulya Komarraju. Whisper Networks and Workarounds: Negotiating Urban Company's Interface. *Feminist Futures of Work*, page 87, 2023.

[55] Bertina Kreshpaj, Theo Bodin, David H Wegman, Nuria Matilla-Santander, Bo Burstrom, Katarina Kjellberg, Letitia Davis, Tomas Hemmingsson, Johanna Jonsson, Carin Håkansta, et al. Under-Reporting of Non-Fatal Occupational Injuries among Precarious and Non-Precarious Workers in Sweden. *Occupational and environmental medicine*, 79(1):3–9, 2022.

[56] Isadora Krsek, Anubha Kabra, Yao Dou, Tarek Naous, Laura A Dabbish, Alan Ritter, Wei Xu, and Sauvik Das. Measuring, Modeling, and Helping People Account for Privacy Risks in Online Self-Disclosures with AI. In *CSCW*, 2025.

[57] Isadora Krsek, Meryl Ye, Wei Xu, Alan Ritter, Laura Dabbish, and Sauvik Das. Supporting Informed Self-Disclosure: Design Recommendations for Presenting AI-Estimates of Privacy Risks to Users. *arXiv preprint arXiv:2601.20161*, 2026.

[58] Noam Lapidot-Lefler and Azy Barak. Effects of Anonymity, Invisibility, and Lack of Eye-Contact on Toxic Online Disinhibition. *Computers in human behavior*, 28(2):434–443, 2012.

[59] Noam Lapidot-Lefler and Azy Barak. The benign online disinhibition effect: Could situational factors induce self-disclosure and prosocial behaviors? *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 9(2), 2015.

[60] Wendy Larcombe. Falling Rape Conviction Rates: (Some) Feminist Aims and Measures for Rape Law. *Feminist Legal Studies*, 19, 2011.

[61] Alex Leavitt. "This Is a Throwaway Account" Temporary Technical Identities and Perceptions of Anonymity in a Massive Online Community. In *CSCW*, 2015.

[62] J Paul Leigh. Economic Burden of Occupational Injury and Illness in the United States. *The Milbank Quarterly*, 89(4):728–772, 2011.

[63] Ada Lerner, Helen Yuxun He, Anna Kawakami, Silvia Catherine Zeamer, and Roberto Hoyle. Privacy and Activism in the Transgender Community. In *CHI*, 2020.

[64] Ada Lerner, Eric Zeng, and Franziska Roesner. Confidante: Usable Encrypted Email: A Case Study with Lawyers and Journalists. In *EuroS&P*, 2017.

[65] Peiyao Liu and Norman Makoto Su. Emotional Re-Victimization in the Workplace: The Burden of Concern Reporting Systems. In *CHIWORK*, 2025.

[66] María Isabel Marqués López. Sexual Violence in the Dark Room: Reading Whisper Networks as Collective Narratives. *Moving Beyond the Pandemic: English and American Studies in Spain*, pages 201–206, 2022.

[67] Ning F. Ma, Veronica A. Rivera, Zheng Yao, and Dongwook Yoon. "Brush It Off": How Women Workers Manage and Cope with Bias and Harassment in Gender-Agnostic Gig Platforms. In *CHI*, 2022.

[68] Xiao Ma, Jeff Hancock, and Mor Naaman. Anonymity, Intimacy and Self-Disclosure in Social Media. In *CHI*, 2016.

[69] Alice E Marwick. Morally Motivated Networked Harassment as Normative Reinforcement. *Social Media+ Society*, 7(2), 2021.

[70] Alice E Marwick and Danah Boyd. I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience. *New media & society*, 13(1):114–133, 2011.

[71] Bronwyn McBride, Kate Shannon, Brittany Bingham, Melissa Braschel, Steffanie Strathdee, and Shira M Goldenberg. Underreporting of Violence to Police among Women Sex Workers in Canada: Amplified Inequities for Im/migrant and In-Call Workers Prior to and Following End-Demand Legislation. *Health and human rights*, 22(2):257, 2020.

[72] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. In *CSCW*, 2019.

[73] Susan E McGregor, Polina Charters, Tobin Holliday, and Franziska Roesner. Investigating the Computer Security Practices and Needs of Journalists. In *USENIX Security*, 2015.

[74] Sanketh Menda, Julia Len, Paul Grubbs, and Thomas Ristenpart. Context Discovery and Commitment Attacks: How to Break CCM, EAX, SIV, and More. In *EUROCRYPT*, 2023.

[75] Ralph C Merkle. Secure Communications over Insecure Channels. *CACM*, 21(4):294–299, 1978.

[76] Ruth Milkman, Ana Luz González, and Victor Narro. Wage Theft and Workplace Violations in Los Angeles: The Failure of Employment and Labor Law for Low-Wage Workers. Technical report, UCLA, 2010.

[77] Jaron Mink, Miranda Wei, Collins W Munyendo, Kurt Hugenberg, Tadayoshi Kohno, Elissa M Redmiles, and Gang Wang. It's Trying Too Hard to Look Real: Deepfake Moderation Mistakes and Identity-Based Bias. In *CHI*, 2024.

[78] Amy Mitchell, Jesse Holcomb, and Kristen Purcell. Investigative Journalists and Digital Security: Perceptions of Vulnerability and Changes in Behavior, 2015. Pew Research.

[79] Rosetta Moors and Ruth Webber. The Dance of Disclosure: Online Self-Disclosure of Sexual Assault. *Qualitative Social Work*, 12(6):799–815, 2013.

[80] Stephanie Murphy, Ava Hickey, Doireann Peelo Dennehy, Kellie Morrissey, John McCarthy, and Sarah Foley. Sharing, Support-Seeking, and Managing Safety: A Qualitative Study of Online Platform Engagement after Pregnancy Loss. In *CSCW*, 2025.

[81] Mark W Newman, Debra Lauterbach, Sean A Munson, Paul Resnick, and Margaret E Morris. It's Not That I Don't Have Problems, I'm Just Not Putting Them on Facebook: Challenges and Opportunities in Using Online Social Networks for Health. In *CSCW*, 2011.

[82] Fayika Farhat Nova, MD Rashidujjaman Rifat, Pratyasha Saha, Syed Ishtiaque Ahmed, and Shion Guha. Online Sexual Harassment over Anonymous Social Media in Bangladesh. In *International Conference on Information and Communication Technologies and Development*, 2019.

[83] Lorelli S Nowell, Jill M Norris, Deborah E White, and Nancy J Moules. Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International journal of qualitative methods*, 16(1), 2017.

[84] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford InfoLab, 1999.

[85] Stefano Pagliaro, Nicoletta Cavazza, Daniele Paolini, Manuel Teresi, James D Johnson, and Maria Giuseppina Pacilli. Adding Insult to Injury: The Effects of Intimate Partner Violence Spillover on the Victim's Reputation. *Violence against women*, 28(6-7):1523–1541, 2022.

[86] James W Pennebaker and Sandra K Beall. Confronting a traumatic event: Toward an understanding of inhibition and disease. *Journal of abnormal psychology*, 95(3):274–281, 1986.

[87] Trevor Perrin and Moxie Marlinspike. The Double Ratchet Algorithm. https://signal.org/docs/specifications/doubleratchet/doubleratchet.pdf.

[88] Adrian Petterson, Ashique Ali Thuppilikkat, Paridhi Gupta, Shamika Klassen, Margaret C Jack, Jun Liu, and Priyank Chandra. Supporting Social Movements through HCI and Design. In *CHI*, 2023.

[89] Anthony Poon, Matthew Luebke, Julia Loughman, Ann Lee, Lourdes Guerrero, Madeline Sterling, and Nicola Dell. Computer-Mediated Sharing Circles for Intersectional Peer Support with Home Care Workers. In *CSCW*, 2023.

[90] Urszula Pruchniewska. "A Group That's Just Women for Women": Feminist Affordances of Private Facebook Groups for Professionals. *New Media & Society*, 21(6), 2019.

[91] Lucy Qin, Vaughn Hamilton, Sharon Wang, Yigit Aydinalp, Marin Scarlett, and Elissa M Redmiles. " Did They F***ing Consent to That?": Safer Digital Intimacy via Proactive Protection against Image-Based Sexual Abuse. In *USENIX Security*, 2024.

[92] Li Qiwei, Francesca Lameiro, Shefali Patel, Cristi Isaula-Reyes, Eytan Adar, Eric Gilbert, and Sarita Schoenebeck. Feminist Interaction Techniques: Social Consent Signals to Deter NCIM Screenshots. In *UIST*, 2024.

[93] Anamika Rajendran, Tarun Kumar Yadav, Malek Al-Jbour, Francisco Manuel Mares Solano, Kent Seamons, and Joshua Reynolds. Deniable Encrypted Messaging: User Understanding after Hands-on Social Experience. In *EuroUSEC*, 2024.

[94] Nathan Reitinger, Nathan Malkin, Omer Akgul, Michelle L Mazurek, and Ian Miers. Is Cryptographic Deniability Sufficient? Non-Expert Perceptions of Deniability in Secure Messaging. In *IEEE S&P*, 2023.

[95] Veronica A Rivera, Daricia Wilkinson, Aurelia Augusta, Sophie Li, Elissa M Redmiles, and Angelika Strohmayer. Safer Algorithmically-Mediated Offline Introductions: Harms and Protective Behaviors. In *CSCW*, 2024.

[96] Ron Rivest, Adi Shamir, and Yael Tauman. How to Leak a Secret. In *ASIACRYPT*, 2001.

[97] Niloufar Salehi, Lilly C Irani, Michael S Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers. In *CHI*, 2015.

[98] Pedro Sanches, Vasiliki Tsaknaki, Asreen Rostami, and Barry Brown. Under Surveillance: Technology Practices of Those Monitored by the State. In *CHI*, 2020.

[99] Tomas Sander, editor. *Workshop on Security and Privacy in Digital Rights Management*, 2002.

[100] Shruti Sannon, Billie Sun, and Dan Cosley. Privacy, Surveillance, and Power in the Gig Economy. In *CHI*, 2022.

[101] Shruti Sannon, Jordyn Young, Erica Shusas, and Andrea Forte. Disability Activism on Social Media: Sociotechnical Challenges in the Pursuit of Visibility. In *CHI*, 2023.

[102] Jennifer A Scarduzio, Shawna Malvini Redden, and Jennifer Fletcher. Everyone's 'Uncomfortable' but Only Some People Report: Privacy Management, Threshold Levels, and Reporting Decisions Stemming from Coworker Online Sexual Harassment. *Journal of Applied Communication Research*, 49(1):66–85, 2021.

[103] Bhavani Seetharaman, Joyojeet Pal, and Julie Hui. Delivery Work and the Experience of Social Isolation. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–17, 2021.

[104] Claude E Shannon. Communication Theory of Secrecy Systems. *The Bell system technical journal*, 28(4):656–715, 1949.

[105] Aliza Shatzman. The Clerkships Whisper Network. *Columbia Law Review*, 123(4):110–145, 2023.

[106] Manish Singh. Signal's Brian Acton Talks about Exploding Growth, Monetization and WhatsApp Data-sharing Outrage, January 2021. Accessed: 2025-10-15.

[107] Nouran Soliman, Hyeonsu B Kang, Matthew Latzke, Jonathan Bragg, Joseph Chee Chang, Amy Xian Zhang, and David R Karger. Mitigating Barriers to Public Social Interaction with Meronymous Communication. In *CHI*, 2024.

[108] Ananta Soneji, Vaughn Hamilton, Adam Doupé, Allison McDonald, and Elissa M Redmiles. " I Feel Physically Safe but Not Politically Safe": Understanding the Digital Threats and Safety Practices of OnlyFans Creators. In *USENIX Security*, 2024.

[109] Zahra Stardust, Rosalie Gillett, and Kath Albury. Surveillance Does Not Equal Safety: Police, Data and Consent on Dating Apps. *Crime, Media, Culture*, 19(2):274–295, 2023.

[110] Sophie Stephenson, Lana Ramjit, Thomas Ristenpart, and Nicola Dell. Digital Technologies and Human Trafficking: Combating Coercive Control and Navigating Digital Autonomy. In *CHI*, 2025.

[111] Angelika Strohmayer, Jenn Clamen, and Mary Laing. Technologies for Social Justice: Lessons from Sex Workers on the Front Lines. In *CHI*, 2019.

[112] Daniel Susser, Beate Roessler, and Helen Nissenbaum. Technology, Autonomy, and Manipulation. *Internet policy review*, 8(2):1–22, 2019.

[113] Lee Taber and Steve Whittaker. "On Finsta, I Can Say 'Hail Satan' ": Being Authentic but Disagreeable on Instagram. In *CHI*, 2020.

[114] Julia Ticona. Red Flags, Sob Stories, and Scams: The Contested Meaning of Governance on Carework Labor Platforms. *New Media & Society*, 24(7):1548–1566, 2022.

[115] U.S. Equal Employment Opportunity Commission. Select Task Force on the Study of Harassment in the Workplace: Report of the Co-Chairs. Technical report, U.S. Equal Employment Opportunity Commission, June 2016.

[116] Warda Usman and Daniel Zappala. SoK: A Framework and Guide for Human-Centered Threat Modeling in Security and Privacy Research. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 2697–2715. IEEE, 2025.

[117] Nihal Vatandas, Rosario Gennaro, Bertrand Ithurburn, and Hugo Krawczyk. On the Cryptographic Deniability of the Signal Protocol. In *ACNS*, 2020.

[118] Gang Wang, Bolun Wang, Tianyi Wang, Ana Nika, Haitao Zheng, and Ben Y Zhao. Whispers in the Dark: Analysis of an Anonymous Social Network. In *IMC*, 2014.

[119] Noel Warford, Tara Matthews, Kaitlyn Yang, Omer Akgul, Sunny Consolvo, Patrick Gage Kelley, Nathan Malkin, Michelle L Mazurek, Manya Sleeper, and Kurt Thomas. SoK: A Framework for Unifying At-Risk User Research. In *IEEE S&P*, 2022.

[120] Noel Warford, Collins W Munyendo, Ashna Mediratta, Adam J Aviv, and Michelle L Mazurek. Strategies and Perceived Risks of Sending Sensitive Documents. In *USENIX Security*, 2021.

[121] Miranda Wei, Sunny Consolvo, Patrick Gage Kelley, Tadayoshi Kohno, Tara Matthews, Sarah Meiklejohn, Franziska Roesner, Renee Shelby, Kurt Thomas, and Rebecca Umbach. Understanding Help-Seeking and Help-Giving on Social Media for Image-Based Sexual Abuse. In *USENIX Security*, 2024.

[122] Kevin Wright. Perceptions of On-Line Support Providers: An Examination of Perceived Homophily, Source Credibility, Communication and Social Support within On-Line Support Groups. *Communication Quarterly*, 48(1):44–59, 2000.

[123] Tarun Kumar Yadav, Devashish Gosain, and Kent Seamons. Cryptographic Deniability: A Multi-Perspective Study of User Perceptions and Expectations. In *USENIX Security*, 2023.

[124] Zheng Yao, Silas Weden, Lea Emerlyn, Haiyi Zhu, and Robert E Kraut. Together but Alone: Atomization and Peer Support among Gig Workers. In *CSCW*, 2021.

[125] Zheng Yao, Diyi Yang, John M Levine, Carissa A Low, Tenbroeck Smith, Haiyi Zhu, and Robert E Kraut. Join, Stay or Go? A Closer Look at Members' Life Cycles in Online Health Communities. In *CSCW*, 2021.

[126] Yutong Zhang, Taeuk Kang, Sydney Yeh, Anavi Baddepudi, Lindsay Popowski, Tiziano Piccardi, and Michael S Bernstein. Burst: Collaborative Curation in Connected Social Media Communities. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–29, 2025.

[127] Jonathan Zheng, Sauvik Das, Alan Ritter, and Wei Xu. Probabilistic Reasoning with LLMs for K-Anonymity Estimation. *arXiv preprint arXiv:2503.09674v5*, 2025.

A Interview Script

Reporting Background and Work Context.

1. You mentioned that you had shared a negative experience online. Could you tell me a bit more about what you shared?
2. Who did you share that with?
3. Why did you choose to share that information with that group?
4. Could you describe what you do for work?
    a. Was the negative experience you shared something you experienced at work or because of the work you do?

Membership and Structure.

1. How did you find the groups with which you shared your negative experience?
    a. How does posting work?
    b. What digital tools or online spaces are involved?
    c. Are there particular roles members serve within these networks? What are those roles and responsibilities?
        i. Are there admins?
        ii. How are they selected?
2. Who are the other people in the group?
    a. Did you know them before joining the group?
    b. Approximately how many people are in the group?
    c. What kinds of things do people discuss or share in the group?
    d. Do you find any of this information helpful? Why or why not?
    e. Is there a way of knowing that people are who they say they are? How are people vetted?
    f. How do you know that you can trust the people in the group?
3. Why did you choose to share your experience with that (those) group(s) in particular?
    a. Were there parts of your experience that you decided to not share?
        i. If yes, how did you decide what to share or not?
    b. Did you share your experience with any other people beyond those networks?
        i. If yes, why those groups/individuals?
        ii. If no, why did you specifically decide to share with X network(s) as opposed to other people you might know online or face-to-face?

iii. Do you choose to share different parts of your experience with different groups?

4. How do people in the network engage with shared content?
5. What is the purpose of the groups you are a part of?

Threats to Sharing.

1. Thinking back to the times you have shared stories and information about your negative experiences online, were you worried about anything bad happening when you posted?
   a. If yes:
      i. What bad things were you worried about?
      ii. What decisions and tradeoffs did you consider when sharing with this group?
      iii. What things about the group made you feel safe/unsafe when sharing?
2. Have you, or someone you know from the groups you are in, ever experienced negative consequences as a result of sharing your story with these groups?
   a. If yes: could you tell me what happened?
3. What do you see as the threats to the goals/purpose/existence of the digital reporting networks you're a part of?

Trust.

1. What are specific things you do to figure out if you can trust someone in the group?
2. How do you know you can trust the information someone shares with you or with the whole network?
   a. What specific things do you do to make decisions about the integrity of the information shared?
3. Have you ever been wrong about whether or not someone or some piece of information shared was trustworthy?
   a. If yes: Could you tell me more? How was your intuition wrong?
4. If you were to believe that someone in the network wasn't trustworthy, or that some piece of information shared could not be trusted, what if anything would you do?

Security.

1. Are there particular ways in which the members of the network work together to protect themselves against the concerns you described earlier?
   a. If yes:
      i. What do they do?
      ii. How effective do you think those approaches are?
      iii. Who decides what concerns to prioritize and ultimately try to address?
   b. If no:
      i. Do you think there is a reason for that?
      ii. What do you think would make it easier for the members of the network to protect themselves from the concerns you described earlier?
2. How safe have you felt overall in sharing your experience with the digital reporting networks you're a part of?
3. Do you feel that technology makes discussing abuse and other sensitive topics easier or harder? Explain.
4. Is there something you would like to see improved or changed about these groups?

B Codebook

Our codebook is located here: https://osf.io/we87h/overview?view_only=b1db57f8e0d545e890039997815bf4ce

C Participant Demographics

This is some demographic information on our participants:

PID	Industry	Sector/Role	Gender	Race
P1	Arts	Acting	Woman	White
P2	Industrial	Construction	Man	Black
P3	Service	Domestic work	Woman	Black
P4	Service	Domestic work	Woman	Black
P5	Service	Restaurant worker	Man	Black
P6	Engineering	Computer security	Man	Asian
P7	Academia	Programming	Man	Native American
P8	Academia	Science	Woman	Black
P9	Engineering	Electrical engineering	Man	Black
P10	Law	Paralegal	Man	Black
P11	Hospitality	Hotel worker	Woman	White
P12	Healthcare	Resident	Man	Black
P13	Journalism	Magazine writing	Man	Asian
P14	Healthcare	Sports physio	Man	Latino
P15	Arts	Photography/modeling	Woman	White; Latina
P16	Healthcare	Psychology	Woman	White
P17	Healthcare	Physician's assistant	Woman	White

D Recruitment Materials

This is our recruitment flyer: