Putting the Count Back Into Accountability

An Audit of Transparency Disclosures About the Sexual Exploitation of Minors

ROBERT GRIMM, Independent Investigator, United States

We have data on online child sexual exploitation.

CCS Concepts: • **Social and professional topics → Computing / technology policy**; • **Information systems → Social networking sites**.

Additional Key Words and Phrases: social media, content moderation, transparency reporting, audit, minor safety, sexual exploitation of minors, child sexual exploitation, child sexual abuse material, CSAM, CyberTipline, National Center for Missing and Exploited Children, NCMEC, teenage sexting

1 INTRODUCTION

Child sexual exploitation (CSE) has become a hotly debated topic over the last few years. There are the headlines in national newspapers announcing that “the internet is overrun with images of child sexual abuse” [4], that “Instagram connects vast pedophile network” [3], and that “AI is about to make the online child sex abuse problem much worse”. There is the hearing by the Senate Judiciary Committee on “big tech and the online child sexual exploitation crisis” featuring the CEOs of Discord, Meta, Snap, TikTok, and X née Twitter [2]. There also are the Qanon adherents and Moms for Liberty mobilizing to “#SaveTheChildren” while maligning pretty much everyone else, particularly LGBT folk, as “groomers” [1, 5].

Despite all their substantial differences, the examples share the urgent emotional appeal and a strong othering effect—though they cannot agree on the identity of that other, alternately pointing at a vast pedophile network, AI, big tech, or groomers. Even if we are inclined to dismiss Qanon and Moms for Liberty for compounding the exploitation of children by exploiting the issue for their own gains, journalists’ and politicians’ sensationalism raises uncomfortable questions about what is fact and what is, maybe not fiction, but at least exaggeration.

Unfortunately, many experts seem to

exploiting the issue of child sexual exploitation

abusing the issue of child exp for their own gain

2 METHODS

This article provides a critical review of transparency disclosures by technology firms about child sexual exploitation and by the United States’ clearinghouse for such incidents, the National Center for Missing and Exploited Children. Notably, it focuses on disclosed data, i.e., reported metrics and their quantities, and seeks to answer the following research questions:

(1) What is the extent of technology firms’ transparency disclosures?
(2) How accurate are disclosed transparency data?
(3) What does the data imply about online child sexual exploitation?

A comparative review requires that reporting organizations, reported metrics, and reported time periods are coherent and in fact comparable. As a result, this article considers only *pieces*, that is, images, videos, and other documents

Author’s address: Robert Grimm, Independent Investigator, Brooklyn, New York, United States, rgrimm@alum.mit.edu.

featuring CSAM, and *reports*, that is, incidents reported to NCMEC, for entire corporations at yearly granularity from 2019 through 2023. All technology firms that do make transparency disclosures do so more frequently, either quarterly or semiannually. However, NCMEC makes only yearly transparency disclosures. Several technology firms as well as NCMEC also break down their statistics by services or brands. For example, Meta accounts separately for Facebook and Instagram. NCMEC does the same, but only started doing so in 2021. Similarly, Google distinguishes between YouTube and the rest of Google. But NCMEC does not. While NCMEC discloses only report counts, many technology firms disclose additional metrics.

Out of the surveyed firms, Meta was the first to make regular disclosures about CSE, starting in Q3 2018. Like other technology firms that make disclosures, it makes several

NCMEC started making detailed transparency disclosures in 2019. Out of the surveyed firms, Meta was the first to make transparency disclosures about content moderation, starting in Q3 2018.

The firms that do make regular transparency disclosures do so more frequently than NCMEC, i.e., either semiannually or quarterly, and do so more granularly, i.e., for each social media brand. While NCMEC also began breaking down its disclosures by social media brands in 2021,

started to break down its report counts by social media brands in

Since NCMEC is the only organization to have a comprehensive, overall view of reported CSE, this article does rely on

focuses on disclosed data, i.e., reported metrics and their quantities

are mostly based on textual analysis, this article focuses on the disclosed data.

have largely focused on qualitative and textual analysis,

Whereas previous investigations of technology firms' transparency disclosures about content moderation have largely focused on qualitative aspects

transparency disclosures by transparency disclosures about content moderation

Unlike previous investigations on transparency disclosures by technology firms including social networks, this article focuses on the disclosed data

It seeks to answer the following research questions:

- What does disclosed does disclosed data tell us about the kinds and magnitude of child sexual exploitation?
- How reliable is the disclosed data?

  This article starts out by reviewing the legal requirements and technological means for detecting and reporting CSE by electronic service providers.

3 US LAW

3.1 Minimal and Reactive Reporting

While US law does require that organizations report any and all child sexual exploitation to NCMEC, it only mandates the reporting of the incident type and date as well as time of the incident. That minimum requirement, however, does not include nearly sufficient information to be actionable. Notably, given the global reach of many US-based technology firms, the location of the perpetrator is critical for routing a report to an appropriate law enforcement agency, whether that is local, federal, or international. In practice, the consequences of this minimal reporting burden are a large volume of "informative," that is, useless, rather than "actionable" reports.

Furthermore, the reporting requirement is purely reactive. Technology firms are *not* required to seek out CSAM and CSE. They only need to report incidents once they become aware of such materials or behaviors. Many technology firms do go beyond that requirement and proactively scan images and videos as they are uploaded. Typically, scans are performed asynchronously and hence there is a short window of time during which such material may be accessible.

3.2 Expansive Definition

For comparison, Germany's penal code distinguishes between "child pornography" and "youth pornography," with the former covering children aged 0 through 13 and the latter 14 through 17. In fact, §665 covering youth pornography was introduced only in 2009. Similar to US law, it outlaws possession of pornographic materials; though it relies on a far broader definition of what such materials might be, including text and audio content as well. Unlike US law, §665 does include an explicit exemption for materials "produced exclusively for personal use with consent of the depicted people." While the statutory language describes somebody producing such materials for their own use, it is commonly understood to also apply to sexting teenagers. However, the lack of suitable language also results in legal uncertainty in Germany. Sentence durations are an order of magnitude shorter:

Both US and German laws do criminalize behaviors, such as parents collecting evidence when their child receives unwanted explicit content, that are designed to support the victim of CSE.

Meanwhile about 20 US states treat consensual sexting as a misdemeanor instead of a felony. However, that may have the perverse effect of prosecutors persuing teenagers more aggressively. While the legal jeopardy can be eliminated, it is important to acknowledge that a teenager's frequency does not influence the probability of becoming the target of abuse, but the critical difference is between not sexting and sexting.

4 TECHNOLOGY

Detection via Cryptographic hashes, perceptive hashes, increasingly machine learning models

4.1 Based on Trusted Authority

Gaining access to NCMEC's hash database requires an application

4.2 Susceptible to Adversarial Manipulation

MD5 is susceptible to known-prefix attacks, whereas hash collision for SHA1 has been documented. The latter may not sound like much but when combined with a suitable document format, notably PDF, suffices for targeted attacks.

Tables 1–3 show yearly *CSAM pieces* and *CyberTipline reports* as disclosed by technology firms (left column triplet) and NCMEC (right column triplet) for the years 2019–2023. CSAM pieces are photos, videos, and other documents. CyberTipline reports describe incidents involving online CSE. Figure 3 plots entries above 1,800 reports from the first column of NCMEC's disclosures across two equally dimensioned coordinate grids. Curves for Meta and Google are repeated across both grids, while the nonzero minimum for the y-axis obscures the curves for Apple and Wikimedia. The three tables present firms in decreasing order of their CyberTipline report counts for 2023. That also is the labelling order for curves in the figure.

Table 1. CSAM pieces and CyberTipline reports disclosed by technology firms and NCMEC: 2019 to 2023 (pt. 1)

	Disclosed by Corporation				Disclosed by NCMEC		
Year	Pieces	per	Reports	Δ%	Reports	of (%)	Total
	META (Q 🖼 📹)						
2019	39,368,400	2.48			15,884,511	93.508	16,987,361
2020	38,890,800	1.92			20,307,216	93.362	21,751,085
2021	78,012,400	2.90			26,885,302	91.454	29,397,681
2022	105,800,000	3.89			27,190,665	84.814	32,059,029
2023	63,300,000	2.07			30,658,047	84.666	36,210,368
	GOOGLE (H 🖼 📹)						
2019					449,283	2.645	16,987,361
2020	4,437,853	8.10	547,875	-0.21	546,704	2.513	21,751,085
2021	6,696,497	7.69	870,319	+0.64	875,783	2.979	29,397,681
2022	13,402,885	6.16	2,174,319	+0.01	2,174,548	6.783	32,059,029
2023	7,955,169	5.40	1,472,221	-0.09	1,470,958	4.062	36,210,368
	X NÉE TWITTER (H 🖼 📹)						
2019					45,726	0.269	16,987,361
2020					65,062	0.299	21,751,085
2021					86,666	0.295	29,397,681
2022					98,050	0.306	32,059,029
2023					870,503	2.404	36,210,368
	SNAP (H 🖼 📹)						
2019					82,030	0.483	16,987,361
2020					144,095	0.662	21,751,085
2021					512,522	1.743	29,397,681
2022	1,273,838	2.31	550,755	+0.06	551,086	1.719	32,059,029
2023	1,594,805	2.31	691,225	+3.16	713,055	1.969	36,210,368
	TIKTOK (Q 🖼 📹)						
2019					596	0.004	16,987,361
2020					22,692	0.104	21,751,085
2021					154,618	0.526	29,397,681
2022					288,125	0.899	32,059,029
2023	107,418,328	181.95			590,376	1.630	36,210,368
	DISCORD (Q 🖼)						
2019					19,480	0.115	16,987,361
2020					15,324	0.071	21,751,085
2021			24,623	+20.24	29,606	0.101	29,397,681
2022			**58,179**	**+191.86**	**169,800**	0.530	32,059,029
2023			**164,478**	**+106.36**	**339,412**	0.937	36,210,368
	REDDIT (H 🖼 📹)						
2019			724	≡	724	0.004	16,987,361
2020			2,233	≡	2,233	0.010	21,751,085
2021	9,258	0.92	10,059	≡	10,059	0.034	29,397,681
2022	80,888	1.54	52,592	≡	52,592	0.164	32,059,029
2023			290,121	+0.01	290,141	0.801	36,210,368

Table 2. CSAM pieces and CyberTipline reports disclosed by technology firms and NCMEC: 2019 to 2023 (pt. 2)

	Disclosed by Corporation				Disclosed by NCMEC		
Year	Pieces	per	Reports	Δ%	Reports	of (%)	Total
	OMEGLE						
2019					3,470	0.020	16,987,361
2020					20,265	0.093	21,751,085
2021					46,924	0.160	29,397,681
2022					608,601	1.898	32,059,029
2023					188,102	0.520	36,210,368
	MICROSOFT (H 🖼 📹)						
2019					123,927	0.730	16,987,361
2020	1,256,652	13.03	96,435	+0.42	96,836	0.445	21,751,085
2021	564,383	7.12	78,926	-0.05	78,883	0.268	29,397,681
2022	452,384	4.22	107,599	+1.11	108,798	0.339	32,059,029
2023	??		??		141,236	0.390	36,210,368
	PINTEREST (Q/H 🖼 📹)						
2019					7,360	0.043	16,987,361
2020			3,432	≡	3,432	0.016	21,751,085
2021	1,608	0.60	2,684	-14.94	2,283	0.008	29,397,681
2022	37,136	1.13	32,964	+4.08	34,310	0.107	32,059,029
2023	57,774	1.15	50$_4$37	+3.81	52,356	0.145	36,210,368
	AMAZON (Y)						
2019					549	0.003	16,987,361
2020			2,235	≡	2,235	0.010	21,751,085
2021	27,244	0.81	33,848	-0.04	33,833	0.116	29,397,681
2022	52,656	0.79	67,073	+4.60	70,157	0.220	32,059,029
2023	24,756	0.79	31,281	+3.45	32,359	0.090	36,210,368
	AUTOMATTIC						
2019					10,443	0.062	16,987,361
2020					9,130	0.042	21,751,085
2021					4,821	0.016	29,397,681
2022					5,035	0.016	32,059,029
2023					19,591	0.054	36,210,368
	QUORA						
2019					10,443	0.062	16,987,361
2020					9,130	0.042	21,751,085
2021					4,821	0.016	29,397,681
2022					5,035	0.016	32,059,029
2023					19,591	0.054	36,210,368
	AYLO NÉE MINDGEEK (H 🖼 📹)						
2019							16,987,361
2020			**4,171**	**+217.17**	**13,229**	0.062	21,751,085
2021	20,401	2.26	9,029	+0.82	9,103	0.031	29,397,681
2022	9,588	4.80	1,996	+4.96	2,095	0.007	32,059,029
2023	7,313	2.92	2,503	+3.72	2,596	0.007	36,210,368

Table 3. CSAM pieces and CyberTipline reports disclosed by technology firms and NCMEC: 2019 to 2023 (pt. 3)

	Disclosed by Corporation				Disclosed by NCMEC		
Year	Pieces	per	Reports	Δ%	Reports	of (%)	Total
		APPLE					
2019					205	1.2e-3	16,987,361
2020					265	1.2e-3	21,751,085
2021					160	5.5e-4	29,397,681
2022					234	7.4e-4	32,059,029
2023					267	7.4e-4	36,210,368
		WIKIMEDIA (🖼)					
2019					13	7.7e-5	16,987,361
2020					11	5.1e-5	21,751,085
2021					8	2.7e-5	29,397,681
2022					29	9.1e-5	32,059,029
2023					34	9.4e-5	36,210,368

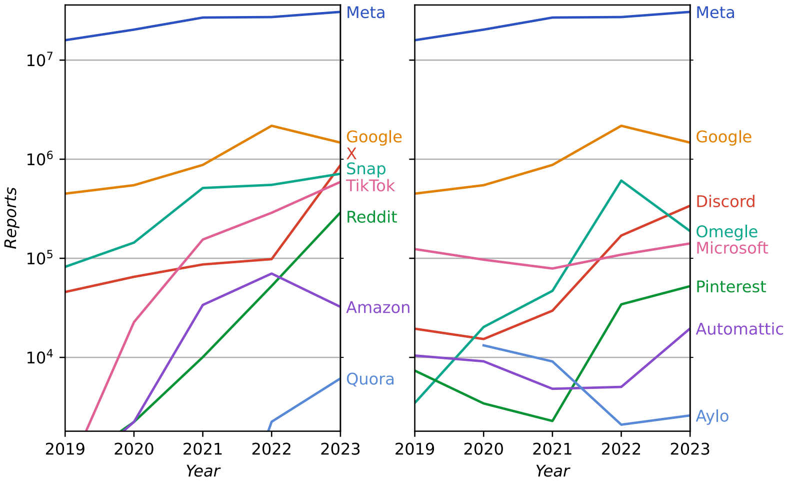

Fig. 1. CyberTipline reports per corporation and year above a threshold of 1,800 on a log scale (NCMEC)

The derived *pieces per report* metric in the middle column for technology firm uses corporations' own piece counts as numerators and report counts as denominators, with the latter falling back onto NCMEC's report counts otherwise. The derived *percentage fraction* metric in the middle column for NCMEC uses NCMEC's corporate report counts as numerators and their totals as denominators. Empty table cells indicate the unavailability of the attendant values, whereas cells with two question marks indicate expected future disclosures for 2023. The parenthesized letters behind firm names indicate the original granularity of their disclosures. Pinterest releases data at quarterly granularity every six months, hence the "Q/H" designation.

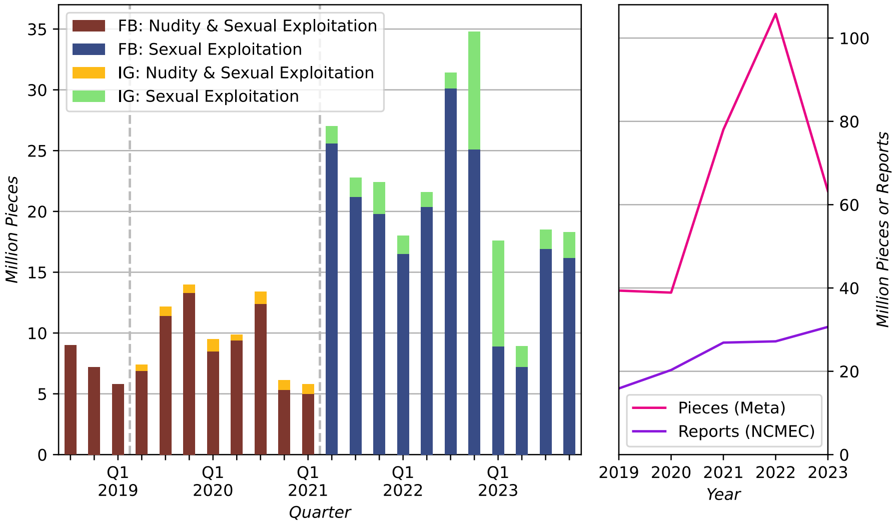

Fig. 2. (a) Quarterly pieces for Facebook and Instagram (Meta); (b) Yearly pieces and reports for Meta (Meta & NCMEC)

Since the tables organize the available data by technology firm, the five years 2019–2023 and five yearly total report counts are repeated for each firm, i.e., twelve times. To reduce visual clutter, they have been greyed out. The total report counts are true totals, reflecting *all* CyberTipline received by NCMEC during a given year. Hence, they are not the same as those included in NCMEC's tabulations of reports received from electronic service providers, which omit reports received from the public. However, the latter subtotals also amount to 98.6%–99.3% of totals, i.e., almost all of them. The total report counts are true totals, accounting for reports submitted by technology firms and public alike. Hence, they are not the same as those included in NCMEC's tabulations of CyberTipline reports received from "electronic service providers," which omit reports submitted by the public. At the same time, those subtotals amount to 98.6%–99.3% of all CyberTipline reports, i.e., almost all of them. Out of technology firms, in turn, Meta is responsible for the vast majority of all reports, at a minimum of 84.7% of total or 85.3% of the subtotal for technology firms. Meta also is the only organization to round its statistics. Meanwhile, Discord and Pinterest stand out for repeatedly disclosing report counts that are substantially different from NCMEC's, in case of Discord's disclosures for 2022 and 2023 severely so.

Figure 3 plots entries above 1,800 from the first column of NCMEC's disclosures in Tables 1–3, i.e., the number of CyberTipline reports the organization received from each corporation, on a log scale across two subfigures, with curves for Meta and Google present in both. The combination of non-zero origin, log scale, and side-by-side subplots ensures that fourteen curves fit into the same figure while still remaining readable.

firms into one chart and the nonzero origin for the y-axis spreads out the cluster from Google to Microsoft, thus facilitating labelling. Discord, Microsoft, Pinterest, and Automattic show significant declines for 2020, which continue into 2021 for Discord, Microsoft, and Automattic. That is surprising because the pandemic is usually credited with a marked increase in CSAM. Alas, in the absence of additional information, it is impossible to explain this phenomenon. Likewise, we cannot explain the decrease in report counts for Google in 2023. We *can*, however, explain the sharp decrease for Omegle: The video chat provider shut down that year.

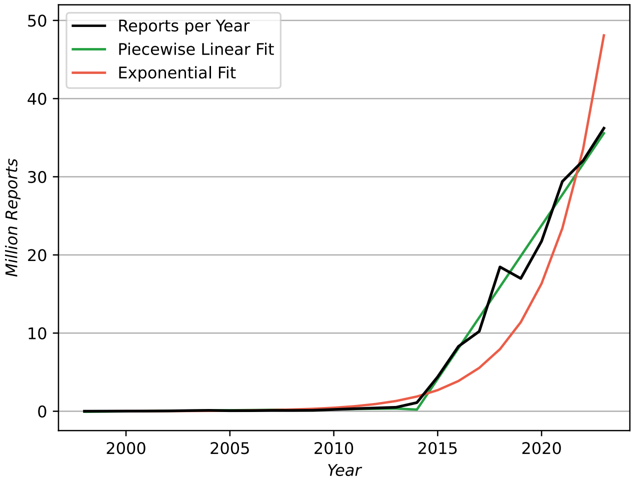

Fig. 3. Yearly reports vs piecewise-linear and exponential curves (NCMEC)

We now discuss each corporation in more detail, in the same order as in the two tables.

REFERENCES

[1] Cody Buntain, Monique Deal Barlow, Mia Bloom, and Mila A. Johns. 2022. Paved with Bad Intentions: QAnon's Save the Children Campaign. *Journal of Online Trust and Safety* 1, 2 (Feb. 2022). https://doi.org/10.54501/jots.v1i2.51

[2] Dick Durbin. 2024. Big Tech and the Online Child Sexual Exploitation Crisis. https://www.judiciary.senate.gov/committee-activity/hearings/big-tech-and-the-online-child-sexual-exploitation-crisis

[3] Jeff Horwitz and Katherine Blunt. 2023. Instagram Connects Vast Pedophile Network. *Wall Street Journal* (June 2023). https://www.wsj.com/articles/instagram-vast-pedophile-network-4ab7189

[4] Michael H. Keller and Gabriel J. X. Dance. 2019. The Internet Is Overrun With Images of Child Sexual Abuse. What Went Wrong? *The New York Times* (Sept. 2019). https://www.nytimes.com/interactive/2019/09/28/us/child-sex-abuse.html

[5] Clara Martiny and Sabine Lawrence. 2023. *A Year of Hate: Anti-Drag Mobilization Efforts Targeting LGBTQ+ People in the US.* Technical Report. Institute for Strategic Dialogue, London, United Kingdom. https://www.isdglobal.org/isd-publications/a-year-of-hate-anti-drag-mobilization-efforts-targeting-lgbtq-people-in-the-us/