Confidential - DSS-3

Workiva updated on	1/17/2024
MAP Due Date	11/30/2023
IA Validation Date	12/18/2023
Validation Performed By	Yojita Ialpuria
Validation Reviewed By	Yanan Li
Review Date	01/10/2024
MAP Remediation Status	Remediated

I. Audit Observation:

IGD Private Messaging Restrictions: Group Chats

The Messenger Wellbeing team does not always prevent unconnected adults from initiating group threads with teens on IGD. For example, on 6/10/2023, IA observed that out of 315K teens added to group threads by adults globally, 35K (11%) were by

II. Management Action Plan:

The Messenger Wellbeing team will:
1.Collaborate with XFN teams to determine required risk coverage and acceptable tolerance for the unconnected adult - teen intervention in group threads. Document the decision.
2.Implement mitigation to limit the unconnected adult - teen interactions in group threads.
3.Establish monitoring to track adult side group messaging effectiveness and set up automated alerts to notify oncall in case of anomalies.

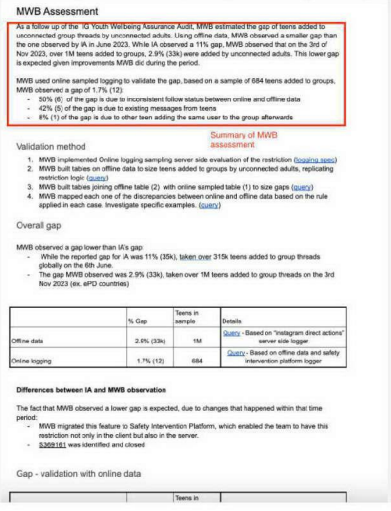
III. Management Remediation Approach:

1. MWB performed detailed analysis to determine the effectiveness of unconnected adult - teen restriction in group threads.
2. MWB implemented Online logging sampling server side evaluation of the restriction.
3. MWB built tables on offline data to size teens added to groups by unconnected adults replicating restriction logic and IA testing logic.
4. MWB built tables joining offline table with online sampled table to size gaps.
5 Using online logging, MWB could validate a sample of the discrepancies identified by offline data based on the rule applied in each case and was able to validate that:
- 50% of the gap is due to inconsistent follow status between online and offline data
- 42% of the gap is due to existing messages from teens
- 8% of the gap is due to other teen adding the same user to the group afterwards
6. MWB established Dashboards for monitoring as follows: Audit dashboard , Server side logging . Information from the following tables was used to extract information for dashboard agg_mwb_unami_evaluations for dashboards and monitoring, Fact table for case by case debugging
7. MWB set up automated alerts to notify oncall in case of anomalies as follows: Audit alerts , Server side alerts

IA Validation Approach:

Step 1: Based on the explanation provided by MWB, IA reviewed the analysis document and noted that 100% of the leakage was explained as false positives.

Screenshot from the analysis document showing summary of assessment performed by MWB team



MWB Assessment

As a follow up of the 'IG Youth Wellbeing Assurance Audit, MWB estimated the gap of teens added to unconnected group threads by unconnected adults. Using offline data, MWB observed a smaller gap than the one observed by IA in June 2023. While IA observed a 11% gap, MWB observed that on the 3rd of Nov 2023, over 1M teens added to groups, 2.9% (33k) were added by unconnected adults. This lower gap is expected given improvements MWB did during the period.

MWB used online sampled logging to validate the gap, based on a sample of 684 teens added to groups, MWB observed a gap of 1.7% (12)
- 50% (6) of the gap is due to inconsistent follow status between online and offline data
- 42% (5) of the gap is due to existing messages from teens
- 8% (1) of the gap is due to other teen adding the same user to the group afterwards

Summary of MWB assessment

Validation method

1. MWB implemented Online logging sampling server side evaluation of the restriction (logging spec)
2. MWB built tables on offline data to size teens added to groups by unconnected adults, replicating restriction logic (query)
3. MWB built tables joining offline table (2) with online sampled table (1) to size gaps (query)
4. MWB mapped each one of the discrepancies between online and offline data based on the rule applied in each case. Investigate specific examples. (query)

Overall gap

MWB observed a gap lower than IA's gap
- While the reported gap for IA was 11% (35k), taken over 315k teens added to group threads globally on the 6th June.
- The gap MWB observed was 2.9% (33k), taken over 1M teens added to group threads on the 3rd Nov 2023 (ex. ePD countries)

	% Gap	Teens in sample	Details
Offline data	2.9% (33k)	1M	Query - Based on "instagram direct actions" server side logger
Online logging	1.7% (12)	684	Query - Based on offline data and safety intervention platform logger

Differences between IA and MWB observation

The fact that MWB observed a lower gap is expected, due to changes that happened within that time period:
- MWB migrated this feature to Safety Intervention Platform, which enabled the team to have this restriction not only in the client but also in the server.
- 5369161 was identified and closed

Gap - validation with online data

		Teens in

Step 2: IA evaluated the Logging spec for implementation done around online logging sampling server side evaluation of the restriction.

Screenshot reflecting the logging specifications of the online table

Name	Data type	Description
actor_id	int	User_id
actor_struct	**int**	**Optional - Acting profile id (for SOAP)**
other_user_id	int	Subject id (Recipient_id)
thread_id	string	thread_id
~~other_id_type~~	~~string~~	~~Values: User / thread / content~~
conformed_interface	string	Indicates it's coming from msgr android / IG iOS / ....
session_id	string	Identifier of execution
use_case	string	Example: IG_CLIENT_INTERVENTIONS
spec	string	Spec under evaluation Example: SafetyInterventionMinorToHighRiskAdultMediaShareEducationSpec
spec_version	int	Optional - Version number of spec
evaluators_seq	vec<string>	"SafetyInterventionHighRiskAdultEvaluator, SafetyInterventionTeenEvaluator, SafetyInterventionThreadAgeNewerThanEvaluator,.."
evaluators_result	vec<int>	1 = true, 0 = false, 2 = not evaluated
evaluators_exception	vec<int>	1 = exception, 0 = no exception

Step 3: IA validated the queries for tables built on offline data to size teens added to groups by unconnected adults, replicating restriction logic which is as follows:
-Restriction is applied if ALL the following conditions are TRUE:
1. Thread is private group chat
2. The target is teen (17 years old or less)
3. Actor is adult (18 years old or more)
4. Teen is NOT following the adult
5. Actor is trying to add teen to group or actor is trying to create new group thread with teen
6. Restriction is not applied if adult stated age is 18 or 19 AND age difference with teen stated age is 1 or 2 years (e.g. users aged 19 years can message users aged 17 years but not users aged 16)

Screenshot reflecting the details of the restrictions implemented

```
SELECT
    ds,
    spec_result IS NOT NULL has_spec,
    COUNT_IF(
        actor_adult AND NOT is_following_starter AND NOT is_2y_gap_exception
    ) not_covered_attempts,
    COUNT(*) add_attempts,
    COUNT(
        DISTINCT IF(
            actor_adult AND NOT is_following_starter AND NOT is_2y_gap_exception,
                target_id,
                NULL
            )
    ) not_covered_teens,
    COUNT(DISTINCT target_id) unc_added_teens
```

Actor is adult — teen is not following actor — 2 year exception

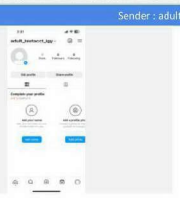
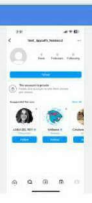
```
FROM tmp_isla_group_threads_sample_ya_online_offline_v5   is group thread
WHERE
    target_teen                                           Target is teen
GROUP BY
    1,2
```

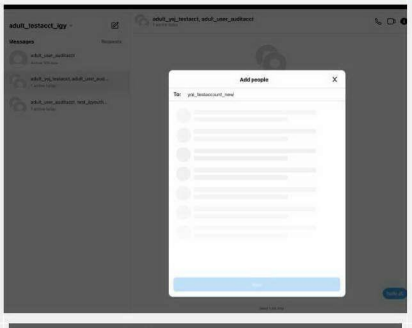
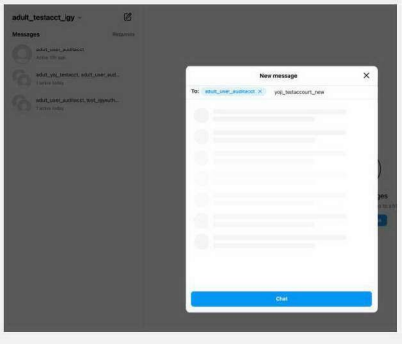
Step 4: IA validated the query for tables built by joining offline table with online sampled table to size gaps.

Screenshot reflecting the details of the join established

```
CREATE TABLE tmp_isla_group_threads_sample_ya_online_offline_v5
WITH (
    oncall = 'mwb_ds',
    partitioned_by = ARRAY['ds'],
    retention_days = 7,
    uii = TRUE
) AS (
    SELECT
        s.spec,
        s.spec_result,
        s.spec_version,
        s.use_case,
        s.evaluators_exception,
        s.evaluators_result,
        s.evaluators_seq,
        f.is_following_starter,
        f.is_reciprocal,
        f.ts,
        f.platform,
        f.actor_id,
        f.users,
        f.num_users_added,
        f.participant_ids,
        f.thread_subtype,
        f.event_data,
        f.target_id,
        f.recommended_adult_minor_flag_target,
        f.stated_age_target,
        f.recommended_adult_minor_flag_actor,
        f.stated_age_actor,
        f.target_teen,
        f.actor_adult,
        f.is_2y_gap_exception,
        f.ds
    FROM safety_intervention_spec_evaluations:messages s
    RIGHT JOIN tmp_isla_group_threads_sample_ya_with_follows_v3 f
        ON f.actor_ic = FB_FBID_TO_IGID_V2(s.actor_id)
        AND f.target_id = FB_FBID_TO_IGID_V2(s.other_user_id)
        AND s.ds = '2023-11-03'
        AND s.use_case IN (13,7)
    WHERE f.ds = '2023-11-03'
)
```

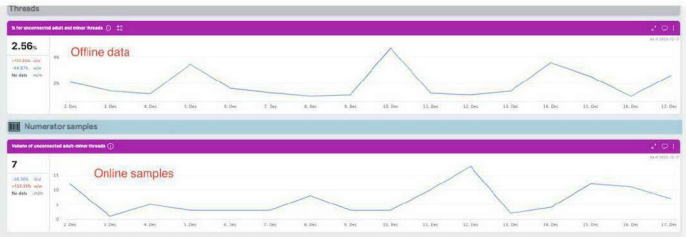
Step 5: IA performed transaction testing to validate if the Unconnected adult <> Minor restriction was enforced. Results of the sample testing are below.

	Sender : adult_testacct_IGY		Target teen: yoji_testaccount_new	
TEST USER DOB				

S.N.	Actor | Target	Actor Side Screershot	Teen Side Screenshot	IA Comment
1	IGID:adult_testacct_IGY (unconnected adult DOB: Nov 2000) was unable to add yoji_testaccount_new (14 yr old) to a group chat .		NA	[NO EXCEPTION NOTED] The teen was not following the adult and so the adult was "unconnected". Unconnected adult <> Minor restriction was enforced through composer block or server-side block. Actor was restricted from adding target teen to group chat
2	IGID:adult_testacct_IGY (unconnected adult DOB: Nov 2000) was unable to create a new group chat with yoji_testaccount_new (14 yr old).		NA	[NO EXCEPTION NOTED] The teen was not following the adult and so the adult was "unconnected". Unconnected adult <> Minor restriction was enforced through composer block or server-side block. Actor was restricted from creating new group with target teen

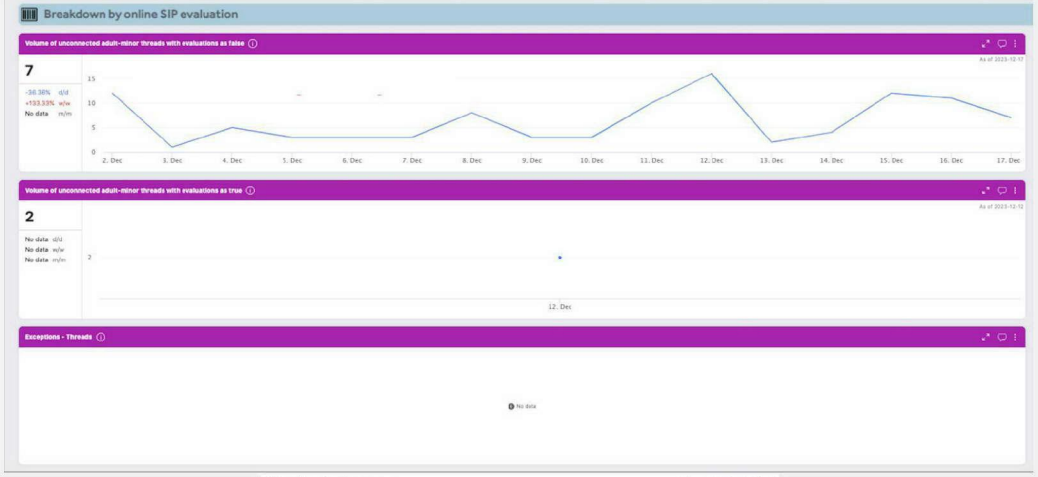
6. IA validated the following Dashboards established for monitoring: Audit dashboard , Server side logging.

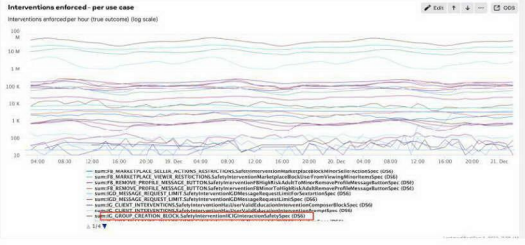




As can be seen from the dashboards below, MWB is monitoring from the samples selected, how many were passed or restricted by the evaluator and corresponding exceptions and false positives. When a message exchange is initiated an evaluator helps determine whether the adult should be able to add teen to the group or not. An evaluator returns a evaluation of "True" if the criteria to implement UNAMI restriction is fulfilled which means teen should not be added. If the evaluator returns a evaluation of ""False"" it means the criteria to implement UNAMI restriction is not fulfilled which means teen can be added. The key dashboards being monitored are as follows:

1. "Volume of Urconnected adult-minor threads with evaluations as False" (Where evaluations resulted in false which means adult should be able to add teen to the group and the teen was actually added due to reasons such as the teen was already following the adult or the group had an existing message from teen).
2. "Volume of Unconnected adult-minor threads with evaluations as True" (Where evaluations resulted in true which means adult should not be able to add teen to the group but the teen was actually added and it was found that teen was added to the group by another teen.
3. "Exceptions" (actual exceptions found by joining online and offline tables which are not explained by above scenarios)



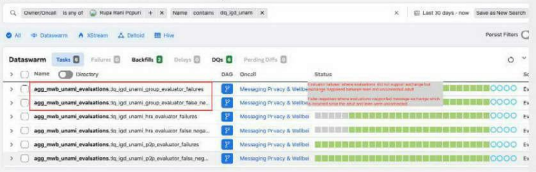


Step 7. IA validated the following automated alerts set up to notify oncall in case of anomalies: Audit alerts , Server side alerts

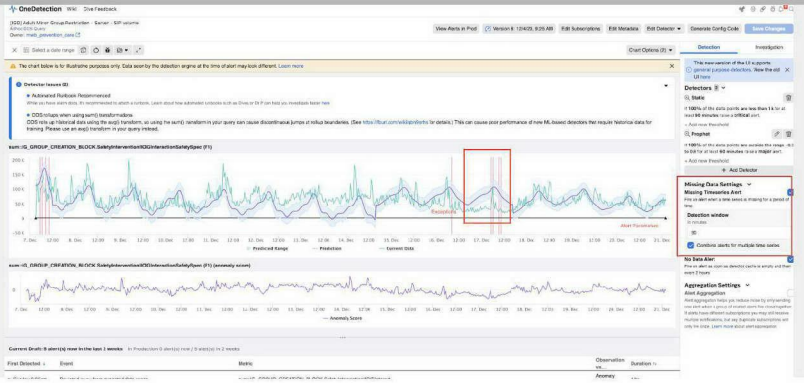
Screenshot reflecting the alerts setup to measure Failed interventions and False negatives for group chats

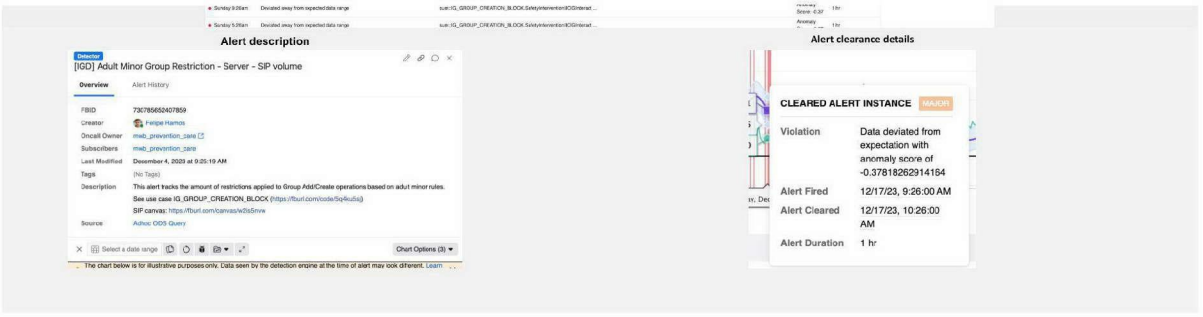
The Key alerts are as follows:

1. Failed interventions (where evaluations resulted in true but message exchange happened between teen and unconnected adult)
2. False negatives(where evaluations resulted in false which shouldn't happen since the adult and teen were unconnected)



Screenshot reflecting the alerts setup on composer block side for group chats. This alert tracks the amount of restrictions applied to Group Add/Create operations based on adult minor rules.




Alert description
Detector
[IGD] Adult Minor Group Restriction - Server - SIP volume
Overview
Alert History
FBID
7307856524078S9
Creator
Felipe Ramos
Oncall Owner
meb_prevention_care
Subscribers
meb_prevention_care
Last Modified
December 4, 2023 at 9:25:19 AM
Tags
(No Tags)
Description
This alert tracks the amount of restrictions applied to Group Add/Create operations based on adult minor rules.
See use case IG_GROUP_CREATION_BLOCK (https://fburl.com/code/5q4cu5ij)
SIP canvas: https://fburl.com/canvas/w3s5nnw
Source
Adhoc ODS Query
Select a date range
Chart Options (3)
The chart below is for illustrative purposes only. Data seen by the detection engine at the time of alert may look different. Learn
Alert clearance details
CLEARED ALERT INSTANCE
MAJOR
Violation
Data deviated from expectation with anomaly score of -0.37818262914164
Alert Fired
12/17/23, 9:26:00 AM
Alert Cleared
12/17/23, 10:26:00 AM
Alert Duration
1 hr