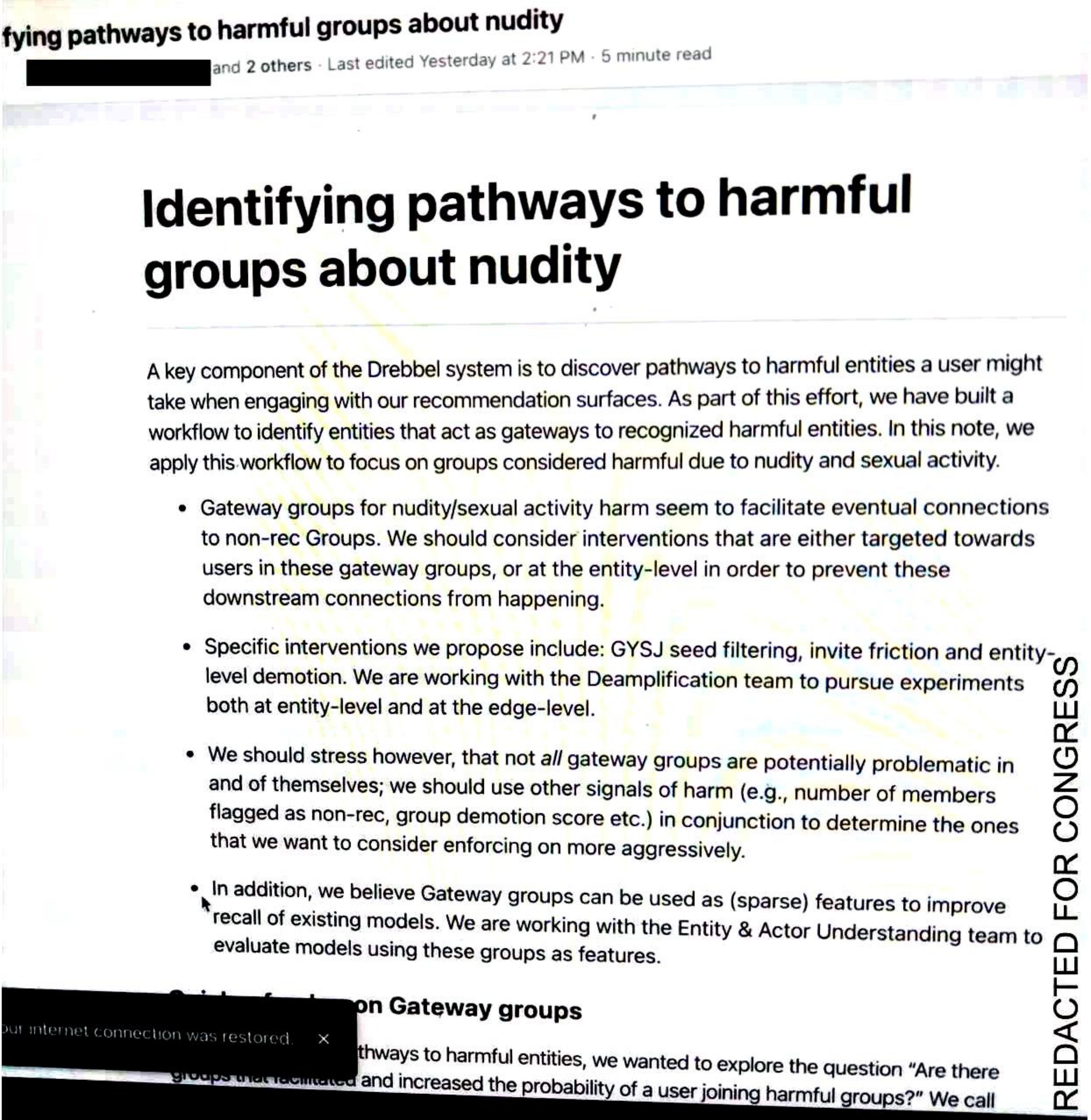


fying pathways to harmful groups about nudity

[redacted] and 2 others · Last edited Yesterday at 2:21 PM · 5 minute read

Identifying pathways to harmful groups about nudity

A key component of the Drebbel system is to discover pathways to harmful entities a user might take when engaging with our recommendation surfaces. As part of this effort, we have built a workflow to identify entities that act as gateways to recognized harmful entities. In this note, we apply this workflow to focus on groups considered harmful due to nudity and sexual activity.

- Gateway groups for nudity/sexual activity harm seem to facilitate eventual connections to non-rec Groups. We should consider interventions that are either targeted towards users in these gateway groups, or at the entity-level in order to prevent these downstream connections from happening.
- Specific interventions we propose include: GYSJ seed filtering, invite friction and entity-level demotion. We are working with the Deamplification team to pursue experiments both at entity-level and at the edge-level.
- We should stress however, that not *all* gateway groups are potentially problematic in and of themselves; we should use other signals of harm (e.g., number of members flagged as non-rec, group demotion score etc.) in conjunction to determine the ones that we want to consider enforcing on more aggressively.
- In addition, we believe Gateway groups can be used as (sparse) features to improve recall of existing models. We are working with the Entity & Actor Understanding team to evaluate models using these groups as features.

on Gateway groups

our internet connection was restored. ×

thways to harmful entities, we wanted to explore the question "Are there groups that facilitated and increased the probability of a user joining harmful groups?" We call

REDACTED FOR CONGRESS

Identifying pathways to harmful groups about nudity

A key component of the Drebbel system is to discover pathways to harmful entities a user might take when engaging with our recommendation surfaces. As part of this effort, we have built a workflow to identify entities that act as gateways to recognized harmful entities. In this note, we apply this workflow to focus on groups considered harmful due to nudity and sexual activity.

- Gateway groups for nudity/sexual activity harm seem to facilitate eventual connections to non-rec Groups. We should consider interventions that are either targeted towards users in these gateway groups, or at the entity-level in order to prevent these downstream connections from happening.
- Specific interventions we propose include: GYSJ seed filtering, invite friction and entity-level demotion. We are working with the Deamplification team to pursue experiments both at entity-level and at the edge-level.
- We should stress however, that not *all* gateway groups are potentially problematic in and of themselves; we should use other signals of harm (e.g., number of members flagged as non-rec, group demotion score etc.) in conjunction to determine the ones that we want to consider enforcing on more aggressively.
- In addition, we believe Gateway groups can be used as (sparse) features to improve recall of existing models. We are working with the Entity & Actor Understanding team to evaluate models using these groups as features.

on Gateway groups

nnection was restored ×

thways to harmful entities, we wanted to explore the question "Are there groups that facilitated and increased the probability of a user joining harmful groups?" We call

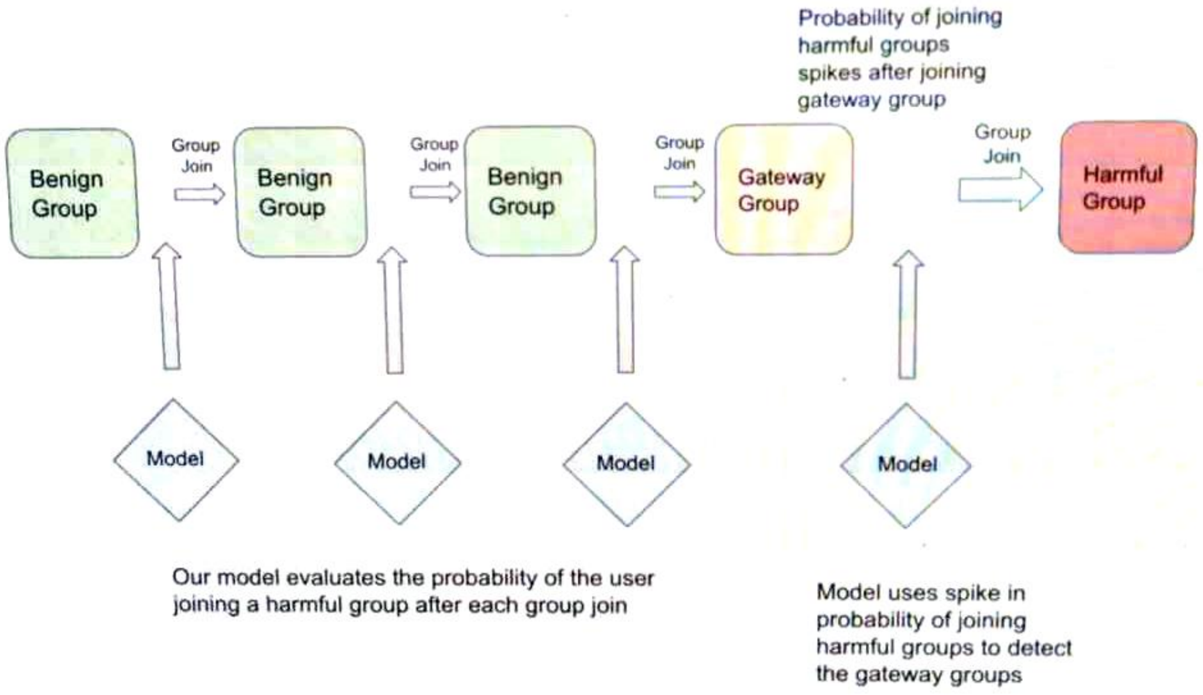
REDACTED FOR CONGRESS

- In addition, we believe Gateway groups can be used as (sparse) features to improve recall of existing models. We are working with the Entity & Actor Understanding team to evaluate models using these groups as features.

Quick refresher on Gateway groups

As part of studying pathways to harmful entities, we wanted to explore the question "Are there groups that facilitated and increased the probability of a user joining harmful groups?" We call such groups gateway groups as they often lead people to join harmful groups.

Here, we provide a brief overview of how we detect gateway groups. For thorough details see this note.



Probability of joining harmful groups spikes after joining gateway group

Our model evaluates the probability of the user joining a harmful group after each group join

Model uses spike in probability of joining harmful groups to detect the gateway groups

To answer the question, we first build a classifier that, given a list of groups joined by an user, can predict with high accuracy whether the user will end up joining a given target harmful group. For a

To answer the question, we first build a classifier that, given a list of groups joined by an user, can predict with high accuracy whether the user will end up joining a given target harmful group. For a particular user, after every group they join, we evaluate the probability of them joining a harmful group in the future. If this probability spikes after a group join, that is a sign that the group just joined might be a gateway. If this spike happens for multiple users, after joining the same group, we identify it as a gateway group.

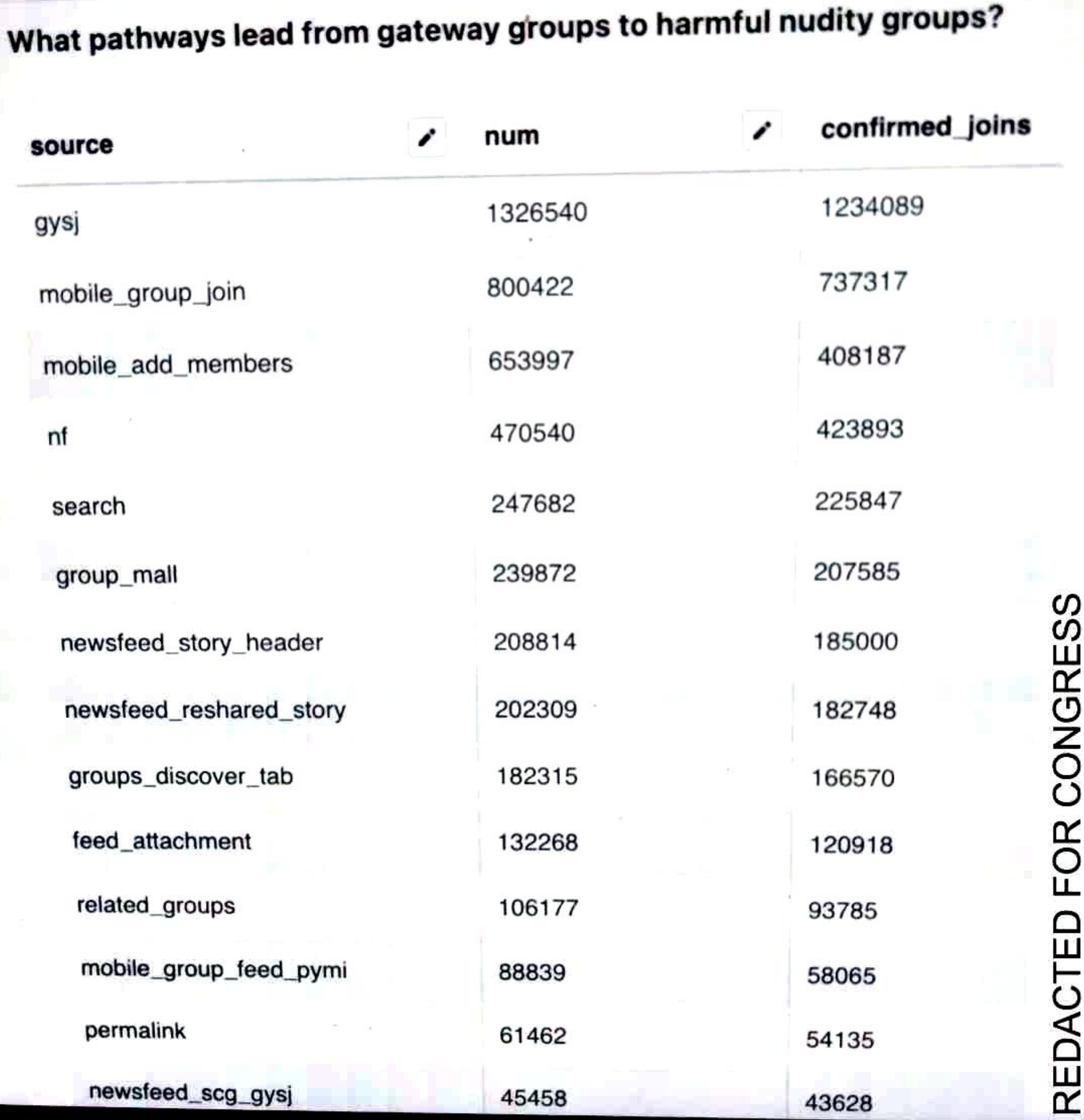
For this note, we used as the set of target groups those based in US with at least 60 content-level strikes for nudity and sexual activity in the month of March (source table: `eau_nudity_sexual_activity_strike_harm_source:integrity`).

What pathways lead from gateway groups to harmful nudity groups?

source	num	confirmed_joins
gysj	1326540	1234089
mobile_group_join	800422	737317
mobile_add_members	653997	408187
nf	470540	423893
search	247682	225847
group_mall	239872	207585
newsfeed_story_header	208814	185000
newsfeed_reshared_story	202309	182748
groups_discover_tab	182215	166570

REDACTED FOR CONGRESS

[redacted]



What pathways lead from gateway groups to harmful nudity groups?

source	num	confirmed_joins
gysj	1326540	1234089
mobile_group_join	800422	737317
mobile_add_members	653997	408187
nf	470540	423893
search	247682	225847
group_mall	239872	207585
newsfeed_story_header	208814	185000
newsfeed_reshared_story	202309	182748
groups_discover_tab	182315	166570
feed_attachment	132268	120918
related_groups	106177	93785
mobile_group_feed_pymi	88839	58065
permalink	61462	54135
newsfeed_scg_gysj	45458	43628

REDACTED FOR CONGRESS

feed_attachment	132268	120918
related_groups	106177	93785
mobile_group_feed_pymi	88839	58065
permalink	61462	54135
newsfeed_scg_gysj	45458	43628
messenger_group_attachment	38879	35208

These are sources of joins of gateway group members to target harmful groups over all time. We see that GYSJ is the top vector here.

source	num	confirmed_joins
mobile_group_join	351341	320524
gysj	313822	268211
nf	273377	251610
group_mall	149788	131753
newsfeed_story_header	148850	134951
newsfeed_reshared_story	142128	127599
mobile_add_members	118133	63896

REDACTED FOR CONGRESS

[redacted] and 2 others · Last edited Yesterday at

group_mall	149788	131753
newsfeed_story_header	148850	134951
newsfeed_reshared_story	142128	127599
mobile_add_members	118133	63896
feed_attachment	62775	55977
groups_discover_tab	45399	38031
permalink	40290	35186
search	35605	29506
related_groups	22375	18304
messenger_group_attachment	21895	19170
groups_tab_reshared_story	16014	14232
mobile_group_feed_pymi	10827	5444

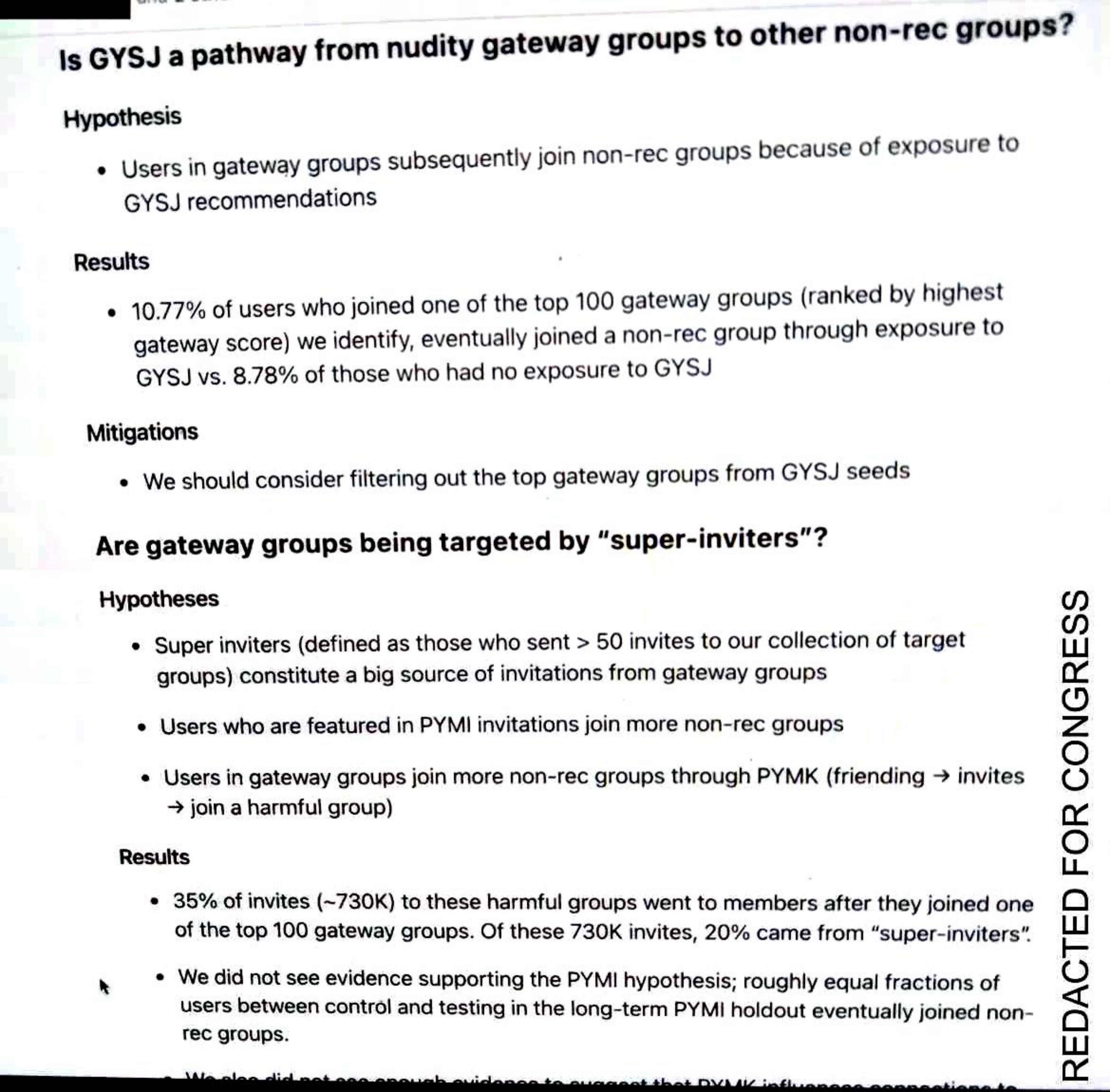
These are sources of joins of gateway group members to target harmful groups after March 01. We see GYSJ in second place here because some groups have been flagged as non_rec - but it is still a big vector.

Is GYSJ a pathway from nudity gateway groups to other non-rec groups?

Hypothesis

- Users in gateway groups subsequently join non-rec groups because of exposure to GYSJ recommendations

[redacted]



Is GYSJ a pathway from nudity gateway groups to other non-rec groups?

Hypothesis

- Users in gateway groups subsequently join non-rec groups because of exposure to GYSJ recommendations

Results

- 10.77% of users who joined one of the top 100 gateway groups (ranked by highest gateway score) we identify, eventually joined a non-rec group through exposure to GYSJ vs. 8.78% of those who had no exposure to GYSJ

Mitigations

- We should consider filtering out the top gateway groups from GYSJ seeds

Are gateway groups being targeted by "super-inviters"?

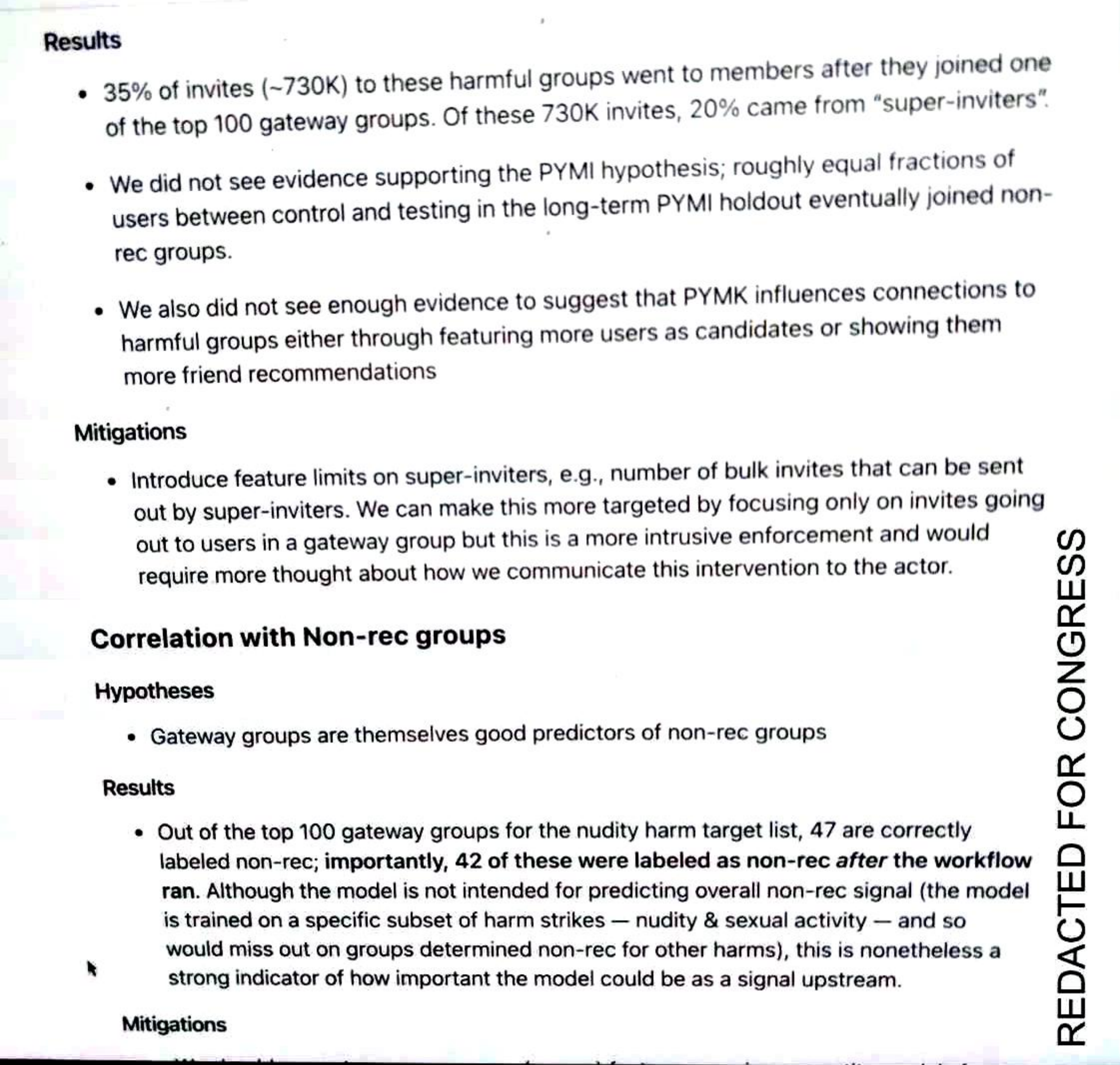
Hypotheses

- Super inviters (defined as those who sent > 50 invites to our collection of target groups) constitute a big source of invitations from gateway groups
- Users who are featured in PYMI invitations join more non-rec groups
- Users in gateway groups join more non-rec groups through PYMK (friending → invites → join a harmful group)

Results

- 35% of invites (~730K) to these harmful groups went to members after they joined one of the top 100 gateway groups. Of these 730K invites, 20% came from "super-inviters".
- We did not see evidence supporting the PYMI hypothesis; roughly equal fractions of users between control and testing in the long-term PYMI holdout eventually joined non-rec groups.
- We also did not see enough evidence to suggest that PYMK influences connections to

REDACTED FOR CONGRESS



Results

- 35% of invites (~730K) to these harmful groups went to members after they joined one of the top 100 gateway groups. Of these 730K invites, 20% came from "super-inviters".
- We did not see evidence supporting the PYMI hypothesis; roughly equal fractions of users between control and testing in the long-term PYMI holdout eventually joined non-rec groups.
- We also did not see enough evidence to suggest that PYMK influences connections to harmful groups either through featuring more users as candidates or showing them more friend recommendations

Mitigations

- Introduce feature limits on super-inviters, e.g., number of bulk invites that can be sent out by super-inviters. We can make this more targeted by focusing only on invites going out to users in a gateway group but this is a more intrusive enforcement and would require more thought about how we communicate this intervention to the actor.

Correlation with Non-rec groups

Hypotheses

- Gateway groups are themselves good predictors of non-rec groups

Results

- Out of the top 100 gateway groups for the nudity harm target list, 47 are correctly labeled non-rec; **importantly, 42 of these were labeled as non-rec *after* the workflow ran**. Although the model is not intended for predicting overall non-rec signal (the model is trained on a specific subset of harm strikes — nudity & sexual activity — and so would miss out on groups determined non-rec for other harms), this is nonetheless a strong indicator of how important the model could be as a signal upstream.

Mitigations

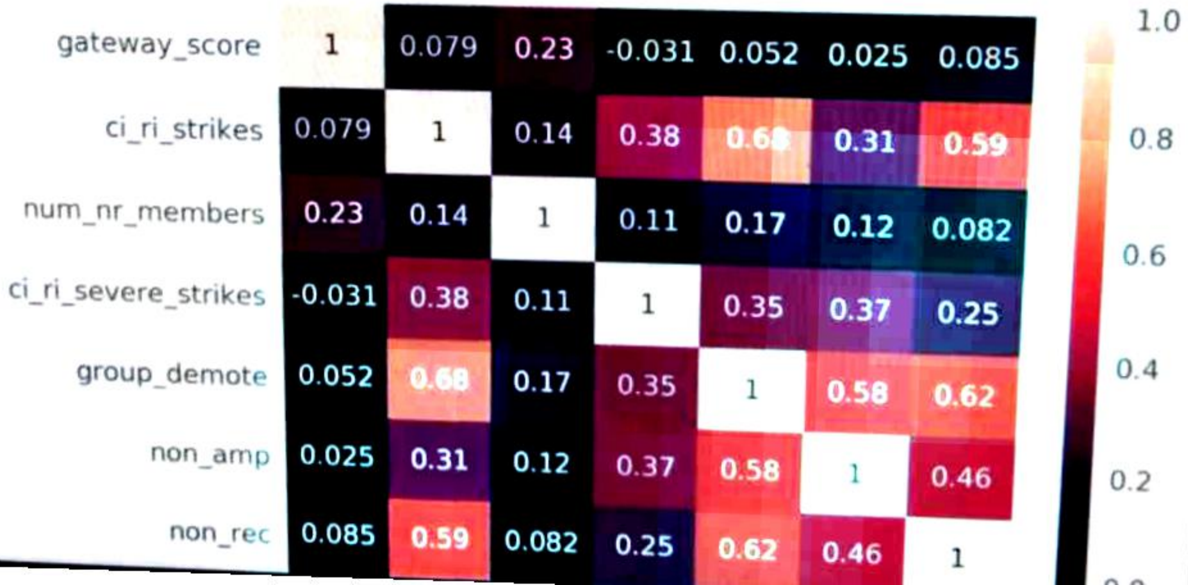
Results

- Out of the top 100 gateway groups for the nudity harm target list, 47 are correctly labeled non-rec; **importantly, 42 of these were labeled as non-rec *after* the workflow ran**. Although the model is not intended for predicting overall non-rec signal (the model is trained on a specific subset of harm strikes — nudity & sexual activity — and so would miss out on groups determined non-rec for other harms), this is nonetheless a strong indicator of how important the model could be as a signal upstream.

Mitigations

- We should use gateway groups as a (sparse) feature powering our entity models for determining non-amplifiable and non-rec entities.

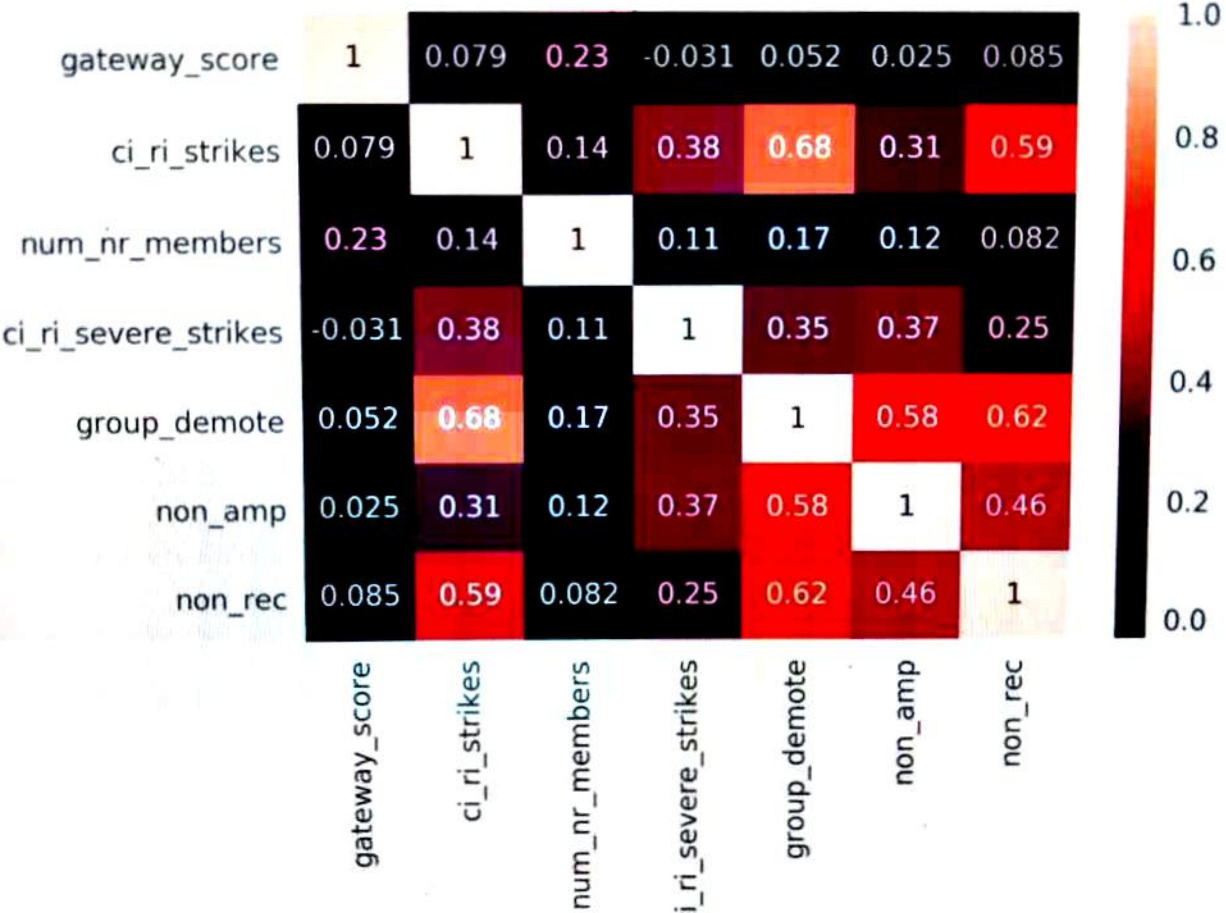
- In conjunction with other signals, such as content strike roll-ups, number of non-rec members, entity strikes, we can pursue entity-level demotions. Our signal has high correlation with the number of group members considered non-rec and has positive correlation with other signals such as strikes and the CPI non-amplifiable flag.



gateway_score	1	0.079	0.23	-0.031	0.052	0.025	0.085
ci_ri_strikes	0.079	1	0.14	0.38	0.68	0.31	0.59
num_nr_members	0.23	0.14	1	0.11	0.17	0.12	0.082
ci_ri_severe_strikes	-0.031	0.38	0.11	1	0.35	0.37	0.25
group_demote	0.052	0.68	0.17	0.35	1	0.58	0.62
non_amp	0.025	0.31	0.12	0.37	0.58	1	0.46
non_rec	0.085	0.59	0.082	0.25	0.62	0.46	1

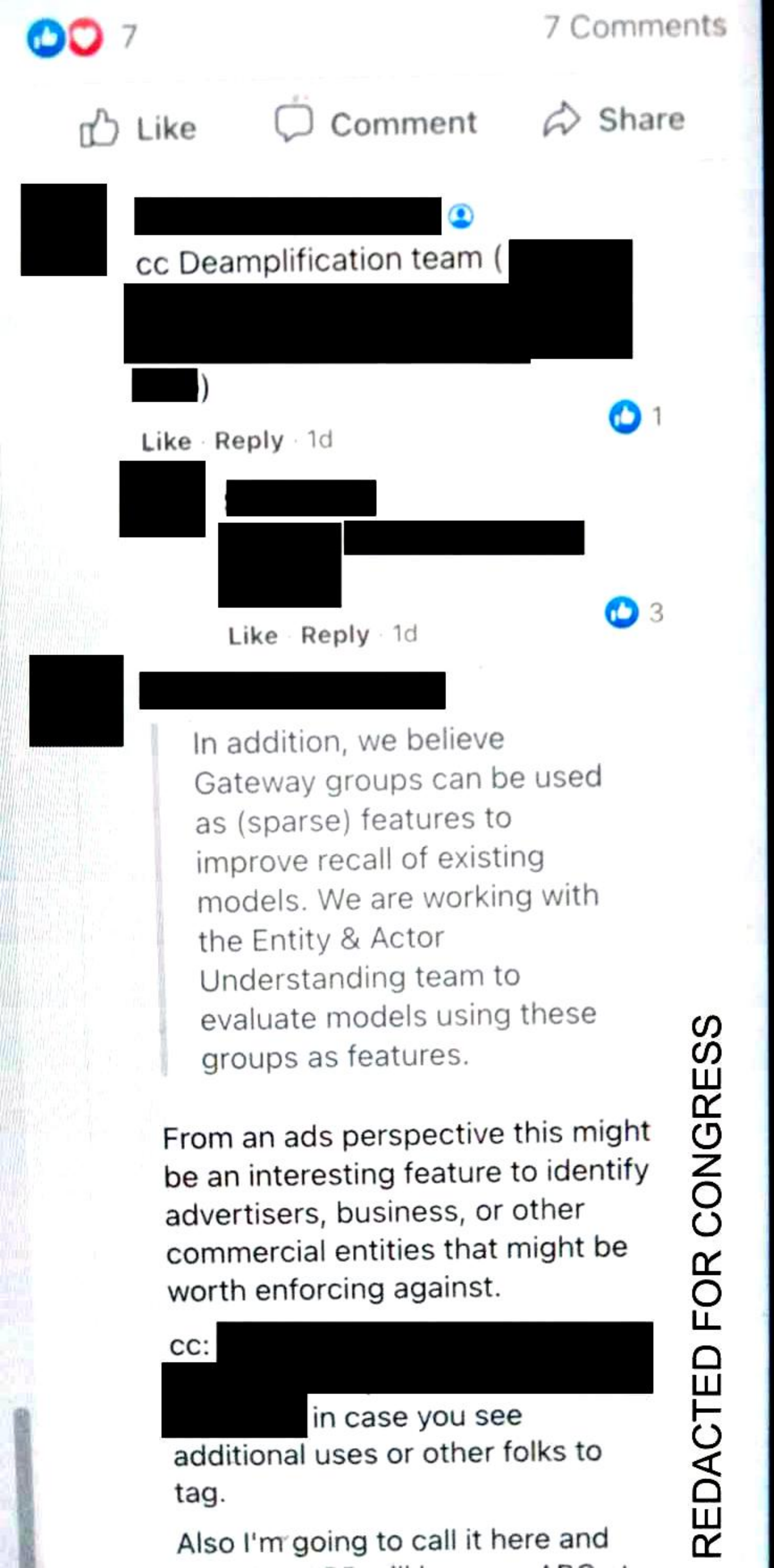
REDACTED FOR CONGRESS

members, entity strikes, we can pursue entity-level demotions. Our signal has high correlation with the number of group members considered non-rec and has positive correlation with other signals such as strikes and the CPI non-amplifiable flag.



	gateway_score	ci_ri_strikes	num_nr_members	ci_ri_severe_strikes	group_demote	non_amp	non_rec
gateway_score	1	0.079	0.23	-0.031	0.052	0.025	0.085
ci_ri_strikes	0.079	1	0.14	0.38	0.68	0.31	0.59
num_nr_members	0.23	0.14	1	0.11	0.17	0.12	0.082
ci_ri_severe_strikes	-0.031	0.38	0.11	1	0.35	0.37	0.25
group_demote	0.052	0.68	0.17	0.35	1	0.58	0.62
non_amp	0.025	0.31	0.12	0.37	0.58	1	0.46
non_rec	0.085	0.59	0.082	0.25	0.62	0.46	1

REDACTED FOR CONGRESS



[redacted] [redacted]

cc Deamplification team ([redacted] [redacted])

[redacted] [redacted]

[redacted] [redacted]

[redacted] [redacted]

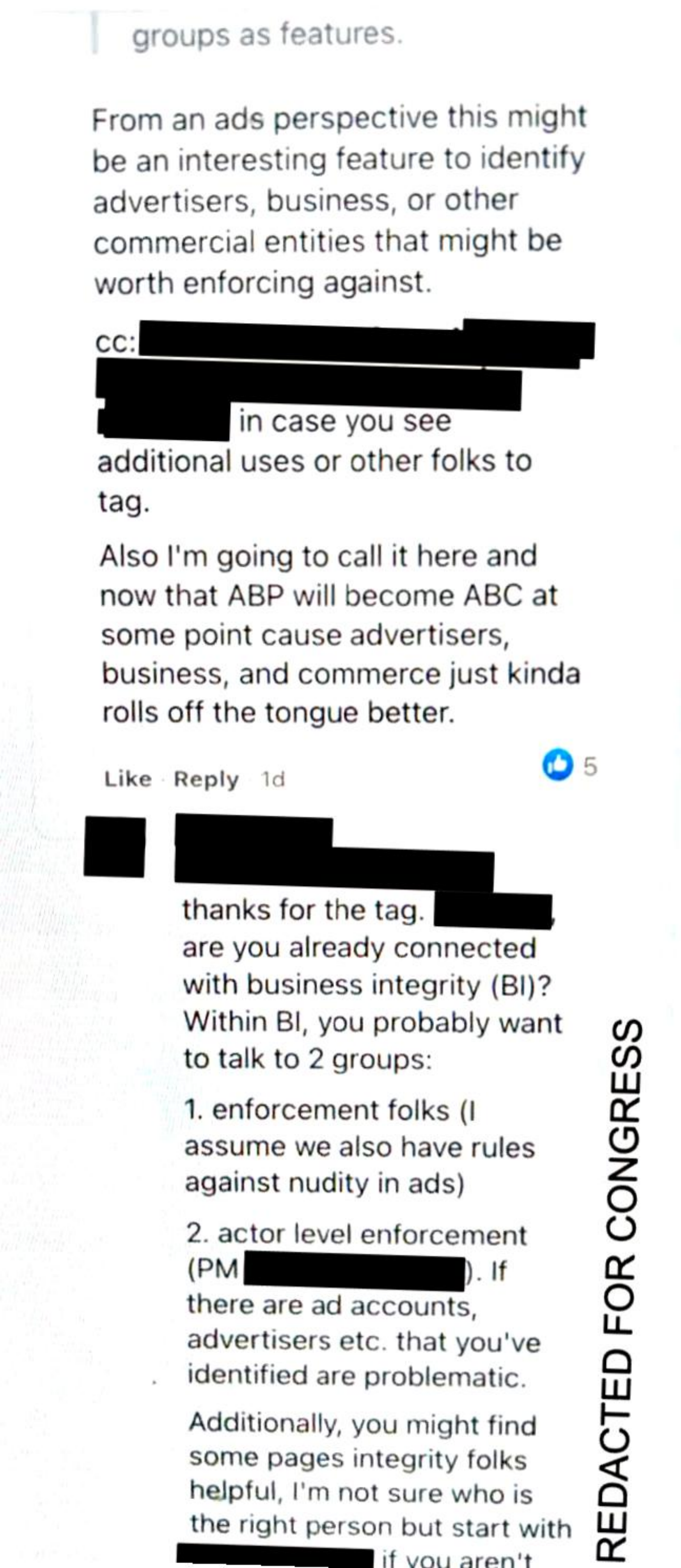
> In addition, we believe Gateway groups can be used as (sparse) features to improve recall of existing models. We are working with the Entity & Actor Understanding team to evaluate models using these groups as features.

From an ads perspective this might be an interesting feature to identify advertisers, business, or other commercial entities that might be worth enforcing against.

cc: [redacted] in case you see additional uses or other folks to tag.

Also I'm going to call it here and

REDACTED FOR CONGRESS



groups as features.

From an ads perspective this might be an interesting feature to identify advertisers, business, or other commercial entities that might be worth enforcing against.

cc: [redacted] in case you see additional uses or other folks to tag.

Also I'm going to call it here and now that ABP will become ABC at some point cause advertisers, business, and commerce just kinda rolls off the tongue better.

[redacted] [redacted]

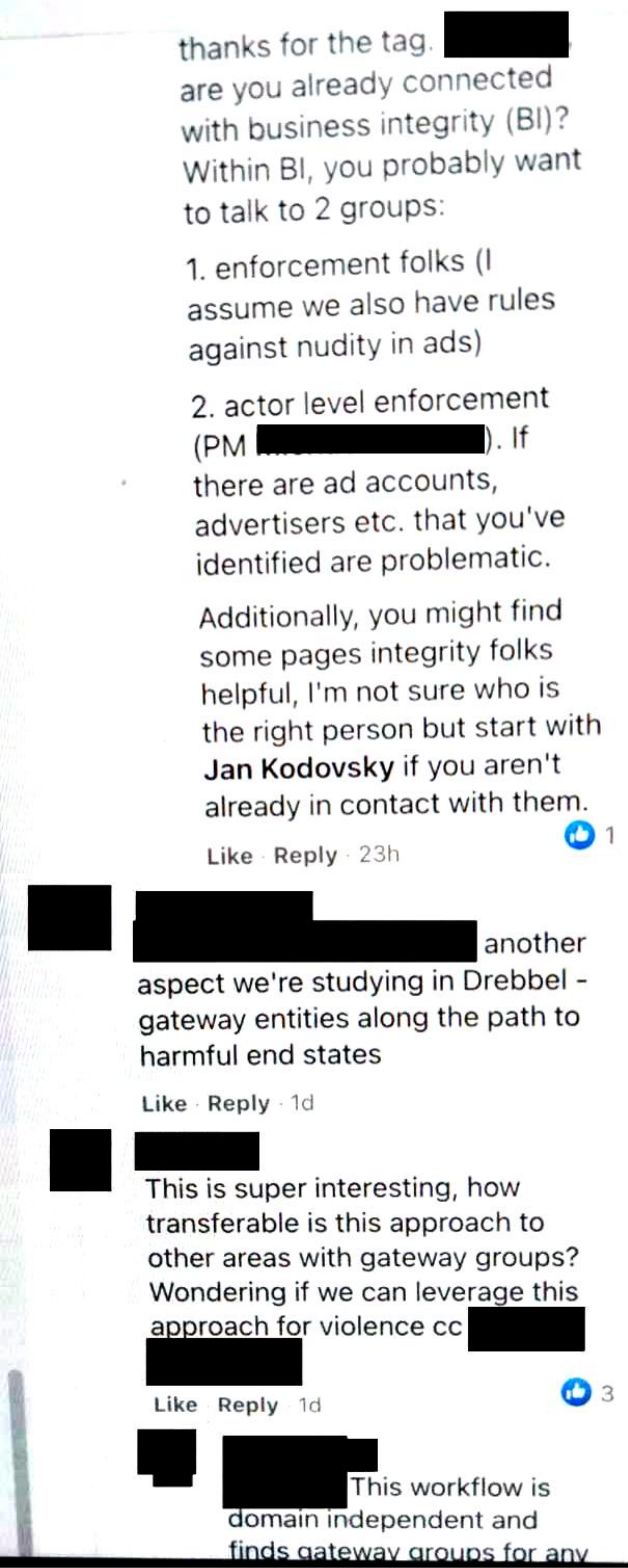
thanks for the tag. [redacted], are you already connected with business integrity (BI)? Within BI, you probably want to talk to 2 groups:

1. enforcement folks (I assume we also have rules against nudity in ads)

2. actor level enforcement (PM [redacted]). If there are ad accounts, advertisers etc. that you've identified are problematic.

Additionally, you might find some pages integrity folks helpful, I'm not sure who is the right person but start with [redacted] if you aren't

REDACTED FOR CONGRESS



thanks for the tag. [redacted]
are you already connected with business integrity (BI)? Within BI, you probably want to talk to 2 groups:

1. enforcement folks (I assume we also have rules against nudity in ads)

2. actor level enforcement (PM [redacted]). If there are ad accounts, advertisers etc. that you've identified are problematic.

Additionally, you might find some pages integrity folks helpful, I'm not sure who is the right person but start with **Jan Kodovsky** if you aren't already in contact with them.

👍 1

Like · Reply · 23h

[redacted]

[redacted] another aspect we're studying in Drebbel - gateway entities along the path to harmful end states

Like · Reply · 1d

[redacted]

[redacted]

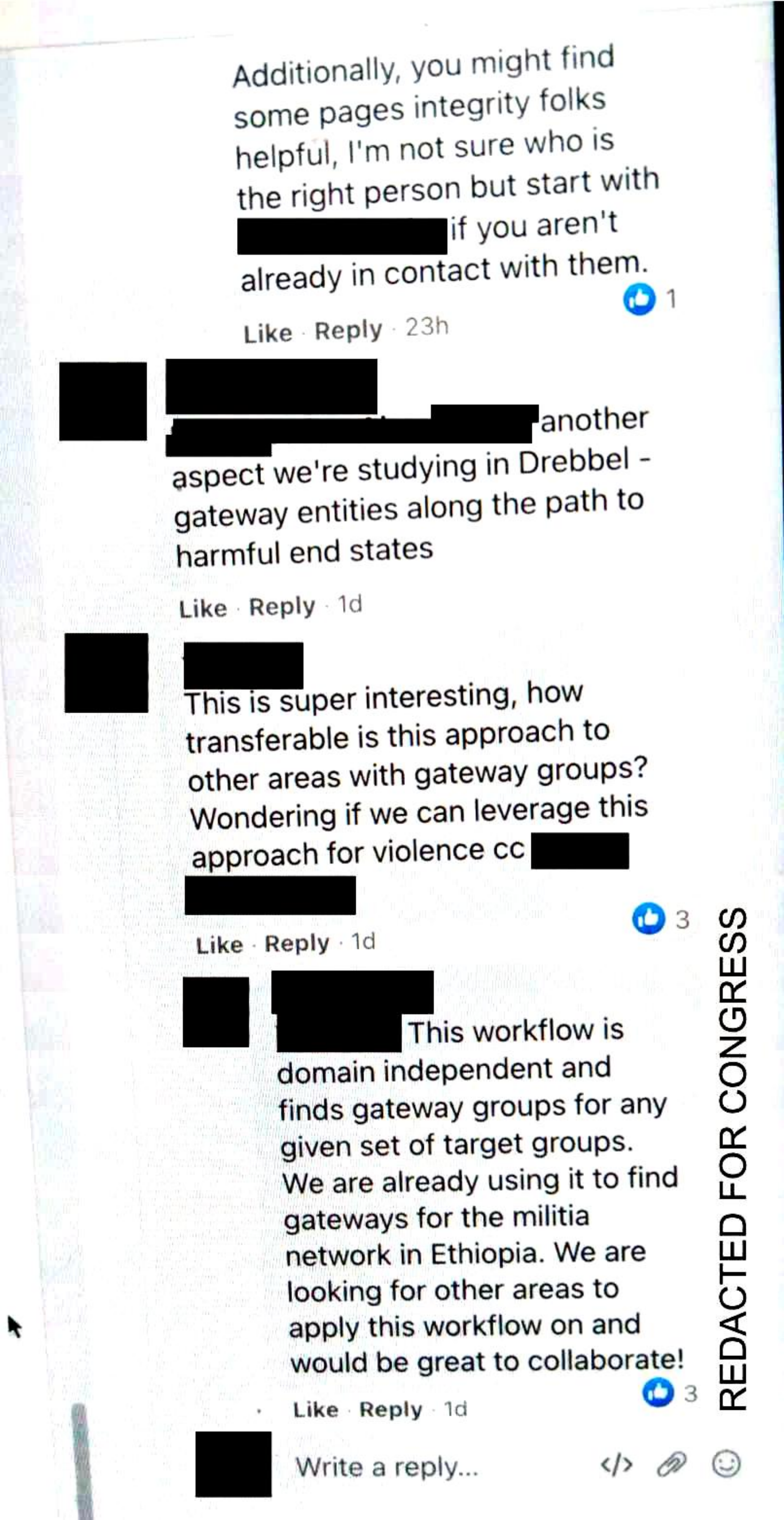
This is super interesting, how transferable is this approach to other areas with gateway groups? Wondering if we can leverage this approach for violence cc [redacted]

👍 3

Like · Reply · 1d

[redacted]

[redacted] This workflow is domain independent and finds gateway groups for any



Additionally, you might find some pages integrity folks helpful, I'm not sure who is the right person but start with [redacted] if you aren't already in contact with them.

👍 1

Like · Reply · 23h

[redacted]

[redacted]

[redacted] another aspect we're studying in Drebbel - gateway entities along the path to harmful end states

Like · Reply · 1d

[redacted]

[redacted]

This is super interesting, how transferable is this approach to other areas with gateway groups? Wondering if we can leverage this approach for violence cc [redacted]

👍 3

Like · Reply · 1d

[redacted]

[redacted]

[redacted] This workflow is domain independent and finds gateway groups for any given set of target groups. We are already using it to find gateways for the militia network in Ethiopia. We are looking for other areas to apply this workflow on and would be great to collaborate!

👍 3

Like · Reply · 1d

[redacted]

Write a reply...

REDACTED FOR CONGRESS