

Reputation in Matrix

Matthew's notes as of Mar 2019 which otherwise haven't been written down formally anywhere other than in scrollback & emails:

Aims

- We aim to empower users to curate their **own** filters to select which content they want to see and don't want to see.
- We do not want to provide any concept of centralised or absolute filtering or reputation data: one man's freedom fighter is another man's terrorist. Instead, we rely on users aligning themselves to whatever social norms their tribe conforms to.
- We wish to let users visualise the behaviour of their filters - particularly to let them switch filters to see content from other angles, or warn them what others think (e.g. 99% of the world have downvoted this!), or warn them of filtered content (98% other rooms have been hidden!)
- We wish reputation data to be available as a general tool for Matrix (e.g. routing algorithms which consider social reputation to help avoid routing only via a single tribe which is more likely to collude together to break traffic).

Proposal

- Users rate content (messages, rooms, threads (in future)) with +1 or -1.
 - We keep it simple to avoid having to moderate the ratings themselves (e.g. is :fire: positive or negative?)
 - It's important to avoid shooting the messenger, and that content is rated based on the sender's behaviour, not the content itself. This means that emoji reactions in isolation are not a sufficient datasource for gathering rating data.
 - Also, individual ratings should be private (unlike emoji reactions) - at least from other users in the room, to avoid retaliation attacks. Ideally they would be private from server admins too, but that involves heading off into ZK-proof territory and is probably a bridge too far.
 - It should support rating arbitrary entities (e.g. mxids, per-room mxids, IP addresses, rooms, communities, server names, bridges).
 - Tracking reputation per homeserver could be particularly important for tracking whether a given server has lax signup and is emanating bots
 - Tracking reputation per mxid is obviously useful for mitigating spam.

- Users are rated by implication of their messages, primarily within a room, but reputation is then correlated cross-room if they have correlated their user ID across multiple rooms to the same mxid. (this assumes MSC1228)
- Users might choose to filter inbound invites from users based on reputation score.
- The reputation data used by a user to filter their content is derived from:
 - Their personal upvotes
 - Transitive reputation from the informal web-of-trust formed by the upvotes, as calculated by their server. If I keep upvoting someone, then I should be more likely to see content from people they have upvoted too.
 - Subscribing to published reputation data (see below).
- Users can choose to publish their reputation data, consciously trading their own privacy for the kudos of contributing their exquisite taste to the world. This could be just to a set of people, or to the wider world. It could be published on their extensible profile.
- When onboarding, users can pick reputation sources to seed their affinity - eg "50% trump, 50% doctorow".
 - One way of doing this could be to cluster analysis of tribes of users around the web of trust (and possibly reputation subscription relationships?), label them based on anyone publishing reputation data from them or some other indicator (e.g. affinity to a particular room), literally plot the resulting tribes on a map or even webgl globe. We then let the user physically stick their pin in the map to determine their current viewpoint, and in future visualise where particular pieces of content sit on the map, and give them a fun and gamified way of moving their viewpoint in future.
 - So you'd end up visualising polarisation of (e.g.) the political left & right, showing based on this upvote WoT what factions appear.
 - An earlier idea was to use hashtags as defined by users as a way to establish tribes (by using twitter firehose snapshots as experimentation), but this somewhat fails on Matrix given folks don't use hashtags and using them for reputation data requires a reputation system to manage the hashtag content itself.

Privacy?

- Rely on msc1228 for keeping mxids private?
- Need to somehow ensure the WoT is privacy protecting.
 - Could shuffle all the IDs in the WoT somehow?
 - Or rely on MSC1228 pseudo IDs?
 - Or rely on WoT data from public rooms already being on the public record, so available for use?
- Should we consider classifying based on content? Running the classifier clientside and producing privacy-preserving rep data could be hard. Perhaps federated learning (e.g. PySyft) could be one approach?
 - The best federated learning can offer is probably to give consensus of classifier results (e.g. 90 out of 100 clients in this room agree that this content has a

libnsfw result of [...]). It's a bit like their users upvoting one another if they reach the same result - you define a tribe of users who run classifiers which return the same results. Needs a lot more thought.

Details

- How do we store our personal reputation data?
- What mechanism do we use for (selectively) publishing it? Extensible profiles?

Problems?

- This inevitably encourages polarisation, by letting tribes band together and filter out other tribes they dislike. To mitigate this, we have to ensure that the flagship clients and interfaces set a precedent for visualising the wider picture too. For instance: "You can only see 20% of the traffic in the room" or "View the visualisation of what tribes exist in this room / community / server" etc.
- You could have social netsplits within a room - e.g. two separate factions talking to each other having filtered each other out in the same room. This might be fine however, if the hidden stuff is still acknowledged rather than totally removed from view.
- Need to ensure that users do not get confused between dislike and disagree when downvoting
- Our cluster analysis has to be resilient to sybils and voting rings. However, given we're trying to detect cliques of users, we should get this for free - and it should apply to both human & automated bots.

Incentivisation?

- What if you seeded earlier users in the network with more reputation to bestow, as a way to encourage early adopters to migrate their networks over?