Participatory Action Research, Polarisation, and Social Media: Ongoing lessons from the Digital MAPS program

Digital Media Arts for an inclusive Public Sphere (Digital MAPS) is a program that supports young leaders from creative and media arts organisations across eight countries to conduct social media mapping in order to understand how issues of identity, social cohesion and inclusion manifest on and are shaped by social media. This participatory action research is then used by these organisations to design pilot interventions that counter polarization and promote inclusivity and openness in the networked public sphere. The program is implemented by Build Up, Sheffield University, Manchester University, the American University of Kurdistan and datavaluepeople, in collaboration with the British Council.

The DMAPS program is of interest to both practitioners in the peacebuilding and creative arts spaces and to academics researching how conflict dynamics intersect with online space and communication practices because it pilots a distinct approach to affective polarisation. The program is premised on a recognition that affective polarisation is something that is happening to users of social media – it is built into the very logic of the social media platforms – and is arguably exacerbating existing fault lines. Thus, the emerging norms of digital culture and their effect on interests, attitudes, affiliations and interactions of social media users present a risk to the type of inclusive public sphere required to build strong open and democratic cultures. This framing is an important shift from mainstream interventions in the network public sphere that focus on content moderation (including identifying and taking down hate speech) and digital literacy (understood as identifying and avoiding misinformation), missing the full complexity of how affective polarisation operates on social media.

In order to share the thinking that has informed this innovative approach and our evolving learnings from this program, we are producing two papers for discussion. This paper outlines the participatory action research process and provides examples of how the resulting social media analysis identified signals of polarisation. It also briefly discusses challenges and lessons from this practitioner-led social media analysis process. The other paper outlines the four areas of theoretical research that support the program, and how these translate into a research-design journey for participants.

Participatory action research applied to social media mapping

The DMAPS program applies participatory action research to social media mapping in order to explore the details of polarizing dynamics that unfold on social media. The participatory action research process comprises four participant-led steps:

1. Defining problem statements
2. Collecting data

3. Coding data
4. Analysing and designing action

Throughout the four steps, participants are in the lead, making this a distinct approach that ensures research findings are contextually grounded and can directly connect to the design of relevant actions.

1. Defining problem statements

Participants start by formulating problem statements that define the scope of the social media mapping. Problem statements comprise a conflict / polarisation thematic area and a series of research questions, and list the country (or countries) and social media platform(s) of interest. Partners worked within regional clusters in a facilitated process to identify their most pressing issue areas and research questions. For example, in Yemen the problem statement focused on expressions of religious and ethnic identities on Facebook and Twitter, addressing the following research questions:

- What identities are Yemenis expressing online?
    - What terms and expressions do Yemenis use to describe their identity online?
    - What other topics are discussed along with identity?
- Do conflict fault lines show up in the expression of identities online? If so, how?
    - What actors make statements about Yemenis? Are these actors “grassroots” or in positions of power?
- Are there identifiable networks of identity on social media?
    - How do social media accounts cluster into networks using similar terms and expressions?
    - Do people who express different identities interact with each other? If so, how?
    - Do polarising conversations about identity attract a certain type of actors?

2. Collecting data

In order to collect data relevant to their problem statements, teams decide on what concrete sources of social media data they want to monitor, producing a detailed list of accounts (public Facebook pages & groups, Twitter handles, YouTube channels) and keywords (to search on public posts on Facebook, tweets or public YouTube videos). These lists are a way of narrowing the social media sphere to a scope that is deemed relevant to the problem statement. Data from these lists was then scraped (i.e. automatically collected), with new data added to a database every three days.

3. Data coding

Data collected (scraped) comes with some useful information about the date it was posted, who posted it, and how many reactions it had. However, with thousands of pieces of data, the research process required additional data coding to organise the information. Classification models were built for topics, actors and sentiment.

Topic labels indicate the general topic (or sub-topic), such as “arrests” or “economic collapse” or “covid / vaccination”. Topics might also include incidents or behaviors related to the conflict / polarisation theme identified in the problem statement, such as “hate crime” or “self-censorship” or “harassment” or “kidnapping”. Topic tagging was also used to group posts / tweets / videos / comments into relevant and irrelevant groups: those with a topic label are relevant since they are about a topic of relevance to the problem statement, while those with no topic label are irrelevant. Topic labeling starts with a manual labeling of approximately 2000 posts / tweets / videos, where team members looked at the text to assign a label, and then identified what features of the text (words and phrases) indicate that this label should be applied. This logic is then repeated by an automated model and applied to all other data collected.

Actor labels indicate what group(s) the author of a post, tweet or video belongs to. Actor labels might include affiliation to a group that is a party to the conflict, country where they are located, profession, or type of audience they attract. Actor labels were manually applied to the data source list and then populated automatically to all data collected.

Sentiment labels indicated the semantic sentiment, that is whether the language used in a post / tweet / video is positive, negative or neutral. Sentiment labels were automatically applied using AWS comprehend, a trained sentiment analysis model.

4. Analysing and designing action

Once the data coding was finalized, all the data collected was presented in a dashboard designed to support participant-led analysis. The dashboard comprises tables and graphs that support topic, actor and sentiment analysis. For topic analysis, graphs and tables show the main topics discussed, who discusses them, what words are most commonly used to discuss it, and what topics are discussed together. For actor analysis, graphs and tables show the main actors, what topics they discuss and how they are connected in the network public sphere (where two actors are connected if they comment on or retweet the same content). For sentiment analysis, graphs and tables show what sentiment is most common for any topic or actor, using both semantic sentiment labels and (for Facebook) emojis. Finally, the dashboard makes the entire dataset searchable, both using a series of filters (actor, topic, sentiment, date) and through a keyword search (applied to the text or account name).

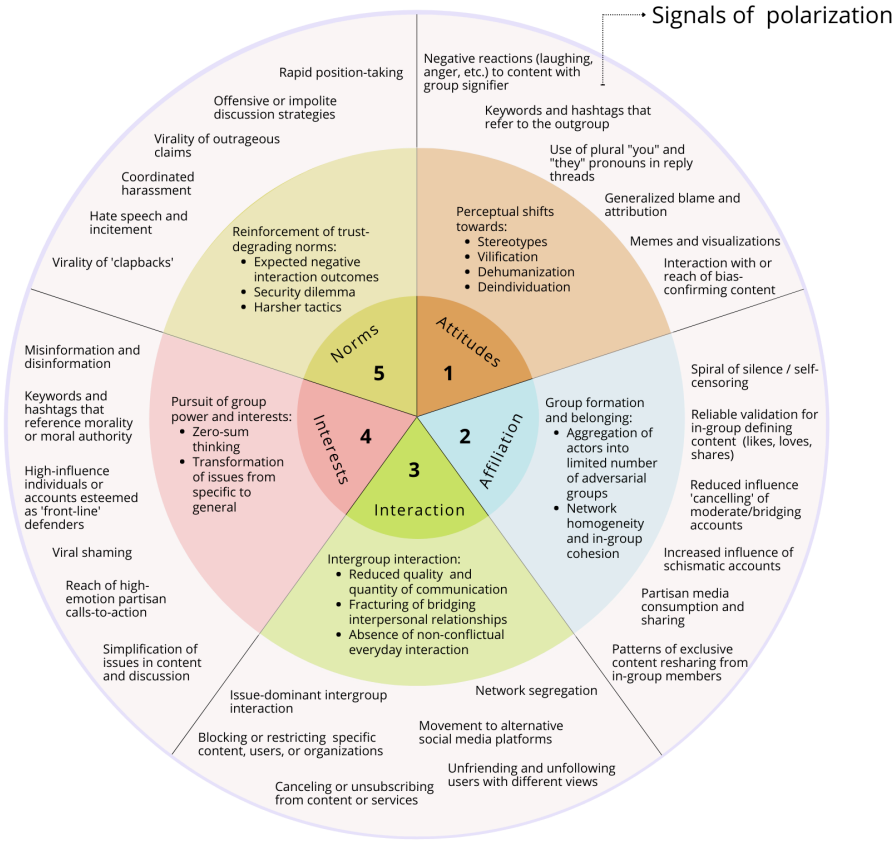
For topics, actors and sentiment, the dashboard enables participants to conduct quantitative, qualitative and network analysis. Quantitative analysis, conducted using the bar and pie charts provided, tells us what people are talking most about, who is talking about this, and how are people reacting to it. Qualitative analysis, conducted using the tables provided to search and read content, tells us how people are talking, and what the tone and emotion are in a narrative. Network analysis, conducted using the interactive network graphs provided, tells us who is talking about what with whom, what voices and topics are on the margins, and which ones are bridges. Overall, the dashboards allow participants to not only answer their specific research questions, but also to come up with key findings that outline what the data tells us about what

conversations selected actors are having on social media, how they are connected and where participants' own social media accounts are placed within these conversations.

Identifying signals of polarisation

This paper does not summarise all the findings that each team arrived at, nor the general finds for each country. Instead, it aims to provide an overview of the type of findings that participants arrived at, in order to offer an understanding of the opportunities of participatory action research applied to social media mapping. Concretely, this section provides examples of how the resulting social media analysis identified signals of polarisation.

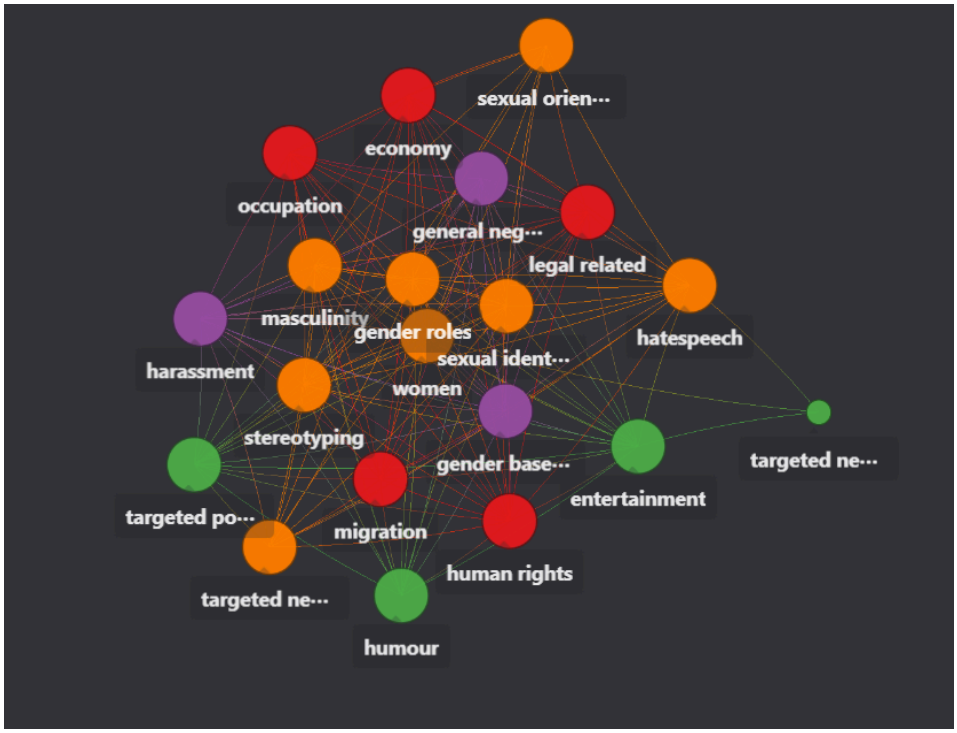
The Digital MAPS approach recognises that the attention-for-profit platform logic of social media leads to segregated networks and spreadable spectacle, with emotive content rewarded by algorithms and therefore more likely to be viral. To concretise this into an approach to analysis, participants were introduced to a framework for problem analysis that specifies the roles of social media as a digital conflict driver that powers affective polarization through influencing identity construction, incentives, and discourse, and maps the induced behaviors of social media users into observable archetypes of polarization within the network public sphere.



In conducting the social media mapping, participants found concrete signals in their data that pointed to each of these archetypes of polarization. Where there are other ways to organise the findings of the social media mapping process, we find that this categorization points most directly to how research findings eventually informed the design of pilot projects.

Attitudes & Norms

Participants from Syria found evidence of polarizing attitudes on social media posts discussing the roles women can play in society, where there is clear evidence of a perceptual shift towards stereotypes. This is clear not just in a qualitative analysis of posts, but also in observing the topic network graph, which shows that the issue of the role of women is largely addressed in terms of masculinity and gender identity. Digging into these posts showed that conversations revolved around the appropriate jobs for women based on physical condition or customs and traditions. On the other hand, the role of women is not often discussed in relation to human rights or (even less) economic issues including material needs.



Looking at the same data, participants in Lebanon and Palestine observed a polarization of attitudes towards women that has resulted in the perpetuation of polarized norms in the network public sphere. Men have a greater tendency to stereotype women, as was clear in qualitative analysis showing frequent comments that suggest emotions are a sign of weakness in women, or sentiment analysis showing negative reactions towards posts indicating that women have financial and educational independence. On posts addressing issues of gender on Facebook, women use mostly love reactions where men use angry and laughing reactions, suggesting that men tend to ridicule this content. The wordclouds show a pattern where the word “woman” or

نساء/مرأة is always joined with a male figure (زوج، شيخ، رجل). It was evident from wordclouds and qualitative analysis that men are commenting on topics predominantly about women (such as economic independence, custodial rights, marriage and divorce, gender-based violence) all the time, whereas women tend to be sidelined when it comes to topics predominantly about men. In general, participants observed that men either react negatively to content about women (quoting religious ideas or arguments) or approach it with ridicule (stereotyping women by using humor).

The analysis results on polarised attitudes and norms around gender in Lebanon, Syria and Palestine informed the content strategies of three teams. Participants in Syria are running a campaign to challenge stereotypes about women in scientific positions. Participants in Palestine have set out to influence a change in the discourse and attitudes towards gender roles and stereotypes by producing content that tackles these issues and that calls for action and discussion of such topics, more specifically, women assuming certain jobs and positions. Participants in Lebanon have designed an approach that uses humor and gender conscious content to mainstream gender equality and break gender stereotypes.

Attitudes & Interests

In Kurdistan, a qualitative analysis of language used on Facebook posts revealed a toxic and polarizing atmosphere against religious minorities (especially Yazidis and Christians), fueled by misinformation and dehumanization, signaling both attitude and interest polarization. This was contrary to the initial hypothesis of participants that hateful narratives would center around ethnic identity. Posts that are related to religious ceremonies (especially during religious festivities) draw discriminatory comments and mockery against religions symbols. In addition to discriminatory comments, Islamic verses and sayings of Prophet Mohammad are posted in the comment sections; these verses are not inherently problematic, but are being instrumentalised to support hateful narratives. The network of Facebook commenters shows that the commenters are normal individuals with their original identities, with no remarkable connection that could be drawn between the accounts. A unifying characteristic of comments is the use of plural “we” and “you” – in reference to superiority. As a result of insights on attitude and interest polarization, participants in Kurdistan are changing the strategy on their own page, by creating a dedicated group of moderators who intervene in comment threads where misinformation and dehumanising narratives towards religious minorities is shared.

In Jordan, the language used across Facebook and Twitter shows signs of the use of morality and moral authority to polarize interests. Participants observed that social media users use words that show moral anger, such as (shame, taboo, homosexuality) and were more likely to respond to emotional words about religion (such as heaven, hell, haram, wrath of God). They also noted the influence of religious influencers on people’s opinions. For example, the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW) was published by Dima Tahboub, who posted her opinion that some of the provisions of the agreement contradict the Islamic religion, and attracted many comments repeating this moral stance. More generally, mainstream media pages often publish questions about other religions or share videos of people

speaking against other religions. Where these posts are disguised as open exchanges, they attribute moral authority to one position, attracting comments only from those who agree, and often leading to hateful or violent comments.

Identifying this tactic to fuel polarization informed the content strategy of program interventions by two participating teams. One team is intervening to raise awareness of the importance of creating safe comments on Facebook, especially those related to the posts of media actors and religious actors that drive polarization. Another team noticed that commenters on social media invoke freedom of opinion and expression to justify hatespeech and incitement to violence, and is targeting content addressing this issue in relation to the topics where comments most fueled polarization, namely religion, customs, traditions, and women's issues.

Norms

In Yemen, the prevalence of coordinated harassment and back-attacks (exclusion from the group based on identity) on social media was an evident signal of polarised norms. In the Yemeni retweet network, there are three clear poles representing the three conflicting political parties. These relatively homogeneous sub-networks present discourse directed exclusively to the same group, with repeated patterns of content sharing by the same group, including lots of negative comments and hate speech on accounts, especially on posts that talk about the group. The rate of adoption of hate speech on Twitter is high, and that its actors are individuals / influences and not external platforms or institutions.

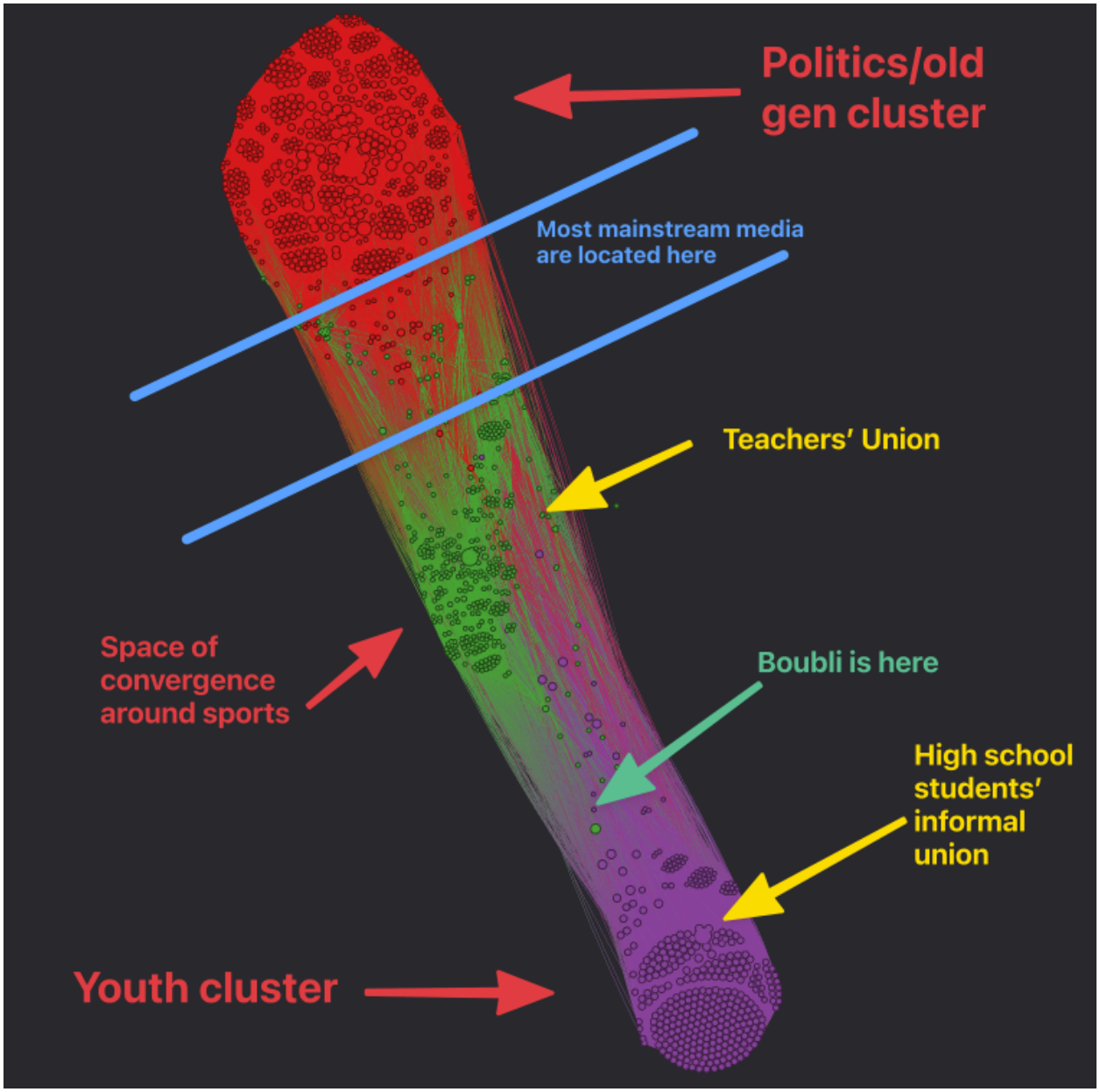
Identifying the pervasiveness of hate speech resulted in all three participating teams from Yemen focusing on this. One team is running a narrative change campaign to counter hate speech by offering a view that highlights the peaceful coexistence between North and South before the conflict started. Another team is running an awareness raising campaign on hate speech and its impact on society, while a third team is focusing its content on sharing people's experience of hate speech.

Affiliation & Interactions

The Libyan social media map (focused on conversations about gender) and the Tunisia social media map (focused on understanding differences between older and younger generations) both provide evidence of the difference between polarized affiliation (where aggregation of actors, homogeneity, self-censorship, echo chambers) and polarized interactions (where network segregation, limited bridging).

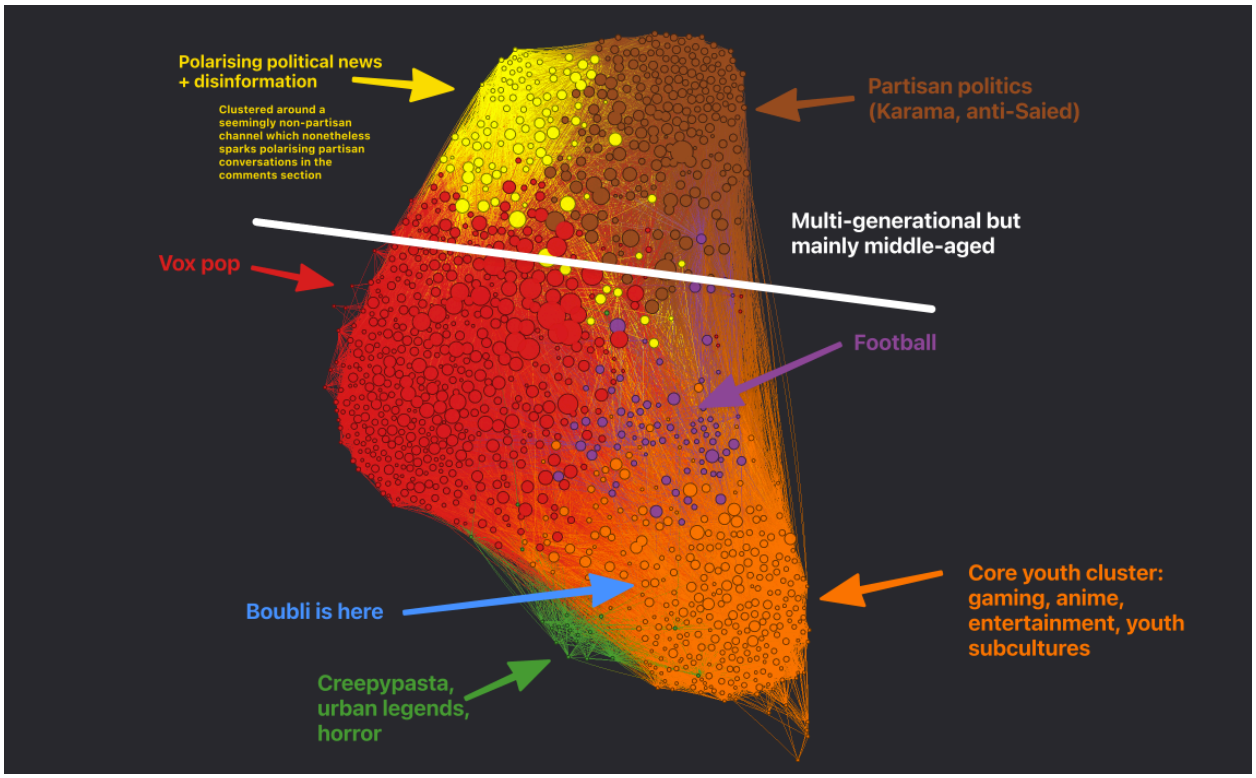
The Facebook commenter network for monitored accounts in Tunisia shows clear signs of network segregation (a signal of polarized interactions) with some bridging (a possible space for intervention). There is a clear boundary between a youth cluster (with content mostly from alternative media and activist accounts) and a cluster dominated by the older generation (with most content coming from mainstream media and partisan political accounts). The most influential account in the youth cluster is such as Lyceena, a page which represents the voice of high

school students and is engaged in a struggle against the education establishment which is dominated by older people. There is a space of convergence around football (centred on Foot24), with commenters on this content bridging between the two poles.



The YouTube commenter network of monitored accounts in Tunisia shows signs of an aggregation of homogenous actors into distinct clusters, a signal of a network that is likely to be dominated by a polarised affiliation. On one end of the graph, there is a youth subcultures cluster which is dominated by entertainment. On the other end of the graph, there are two political clusters that tend to attract middle-aged people whose content is highly polarising and often spreads disinformation. Between them, there is a large cluster around vox pop content which

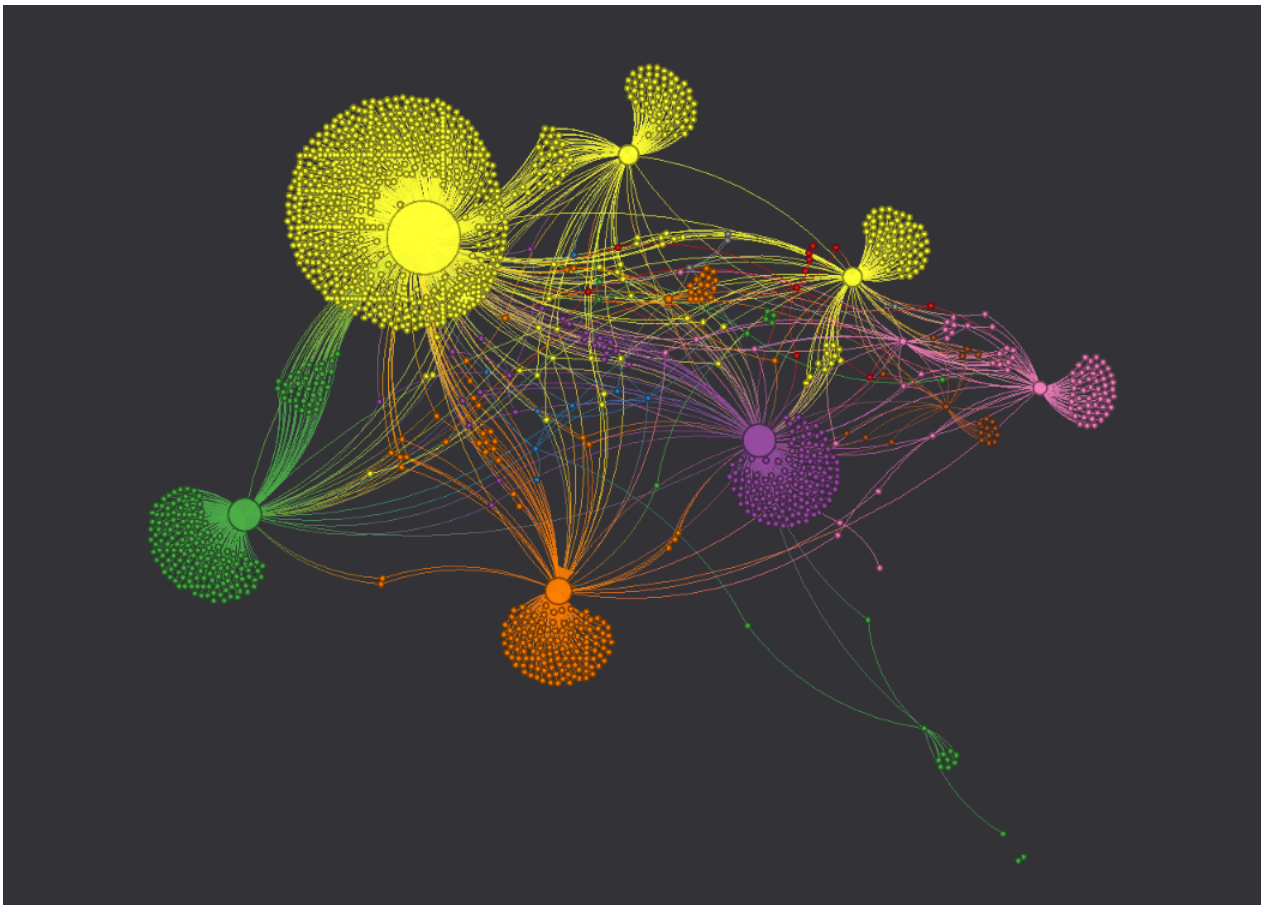
suggests a strong appetite for content which features the voices and perspectives of ordinary people.



The analysis results on polarised affiliation in Tunisia have led participants to design an intervention that centers the voices of young and old people sharing their perspectives in a video to start a conversation, in an attempt to approach the “vox pop” style that appears to act as a bridge between the youth and older generation cluster.

The Facebook commenter network of accounts monitored in Libya shows clear signs of network segregations with limited bridging, signals of polarized interactions. Overall, most accounts monitored are not connected regularly by commenters, including all of the participants’ accounts. For those that are connected, there is a small pole (orange) made up exclusively of Jaafar Tok DW and his tight network of commenters, who are not engaged in the rest of the conversation.

DW is the account on Facebook with the most engagement, and he deliberately creates a space for controversial posts in the way he writes captions, he invites people to disagree in the comments section and encourages engagement. There are very few bridging accounts who comment on both DW and pages in the other pole (six concrete accounts were identified).



The network of retweets by accounts monitored in Libya shows clear signs of an aggregation of homogenous actors, a signal of a network that is likely to be an echo chamber dominated by a polarised affiliation. All the main nodes are activists working on gender. Interestingly, one of the participant accounts sits in the centre of this echo chamber.

The analysis of polarised aggregation and affiliation in Libya has defined the targeting strategy used by some participants, whose focus is now to bridge the gap between its circle of influence on social media platforms and where the major discussions on gender roles are happening.

Challenges & lessons from practitioner-led social media analysis process, and a position for future work in this area

Two key challenges have emerged in the participatory action research process described above. First, the research only looks at a sub-space within the broader network public sphere, and analysis is therefore limited by the problem statements defined by participants and by selection bias in relation to data sources. Second, it takes time to set up the data collection and coding needed to perform this kind of in-depth analysis, and to acquire relevant social media analysis

skills. Some participants more focused on intervention / activism do not see the value in performing participatory action research and would prefer an external research report that provides evidence and supporting inputs to their programming.

These challenges point to the main lesson learned from the analysis process of the Digital MAPS program: that there is a clear opportunity for conducting research that informs evidence-based interventions that address different aspects of polarization in the network public sphere, but sustained participatory action research may only be worthwhile for actors who expect a sustained engagement in digital interventions to shift polarization in the network public sphere. Actors who would want a one-off input may benefit most from externally produced research on a snapshot of the network public sphere, rather than an ongoing participatory analysis process.

Despite these challenges, the examples above illustrate that the participatory action research process resulted in participants identifying concrete signals of polarization in a sub-space of the network public sphere relevant to their problem statements. The clearest opportunity arising from this process is that these insights, because they are concrete, can be directly translated into program design. Concretely, participants used research insights to define content production, change targeting strategies, and refocus their own social media activity. Perhaps most interestingly, the analysis process in some contexts resulted in participating teams designing interventions that used different tactics to address the same identified signals of polarization, which makes a shift in the network public sphere as a whole more likely. We believe this pilot program therefore demonstrates the value of participatory action research on social media, and its potential to inform the design of interventions to address polarisation.