DEMON

Detect and Evaluate Manipulation of ONline information

Deliverable number: **WP3**

Network-based analysis of manipulation of sub-narratives (M6–M18)

POLITECNICO
MILANO 1863

Project Acronym:	DEMON
Project Title:	Detect and Evaluate Manipulation of ONline information
Project Type:	PRIN 2022
Project Code:	2022BAXSPY
Project Website:	`https://demon.unica.it/`
Workpackage:	**WP3:** Network-based analysis of manipulation of sub-narratives
Deliverable Type:	Report (R)
Dissemination Level:	Public (PU)
Authors:	Stefano Ceri, Francesco Pierri (POLIMI)

Contents

1 Introduction to WP3

This work package focuses on the analysis of information diffusion, the characterization of harmful and manipulative content, and the study of the structural and behavioural properties of online social platforms. In particular, WP3 addresses how narratives emerge and spread in digital environments, how platform design and moderation affect such processes, and how methodological choices in computational analysis may alter our understanding of diffusion dynamics.

A first line of work concerns the study of decentralized social media platforms, which have recently attracted growing interest as alternatives to mainstream social networks. These environments provide an opportunity to observe the emergence of new information ecosystems, the evolution of user activity, and the dissemination of both trustworthy and low-credibility content. The opening of Bluesky to the public represented a particularly relevant case study, allowing us to investigate how a platform changes when rapidly exposed to a large inflow of new users.

A second line of work examines the relationship between platform growth and the spread of harmful or manipulative phenomena, including misinformation, political polarization, toxicity, and suspicious coordinated behaviour. Understanding these aspects is crucial for DEMON, as they are directly related to the broader goal of detecting and evaluating manipulation of online information. The analysis of user behaviour, source credibility, resharing practices, and moderation outcomes contributes to a more precise understanding of how online narratives are amplified or constrained.

Finally, this work package also includes a methodological contribution on the reconstruction of diffusion cascades. Platform-provided interaction data often do not reveal the true pathways through which content spreads. As a consequence, many analyses rely either on raw platform traces or on reconstructed cascades based on simplifying assumptions. This introduces a major source of uncertainty, since the identification of influential actors and the estimation of network-level diffusion properties may depend strongly on the assumptions adopted. Addressing this issue is essential to ensure that the study of online manipulation rests on reliable and interpretable computational methods.

Overall, WP3 contributes to the DEMON project by combining empirical analysis of emerging social platforms with methodological advances for studying information spread. The findings provide both substantive insights on the circulation of news and harmful content and practical guidance for future analyses of online narratives, misinformation, and manipulation.

1.1 Contribution and Organization

The following contains the contributions of the DEMON project falling into the area of Work Package 3 described above. In particular:

- an empirical study of Bluesky during the two months surrounding its public opening, aimed at characterizing user activity, content production, toxicity, and suspicious behaviour (Section 2);
- an extended longitudinal analysis of Bluesky investigating misinformation, political polarization, source credibility, and moderation outcomes after the platform's public launch (Section 3);

- a methodological study on information diffusion showing how cascade reconstruction assumptions can significantly distort the identification of influential actors and the structural properties of diffusion networks (Section 4);
- a discussion of how these results contribute to the broader goals of DEMON, namely the detection, interpretation, and evaluation of manipulation dynamics across online platforms (Section 5).

The deliverable is organized as follows. Section 2 presents the initial study on Bluesky around its public opening. Section 3 describes the journal extension, which broadens the analysis to include misinformation, polarization, and harmful content. Section 4 discusses the methodological contribution on diffusion reconstruction and its implications for the study of online social network dynamics. Finally, Section 5 summarizes the main outcomes and outlines their relevance for the project.

2 The dawn of decentralized social media: an exploration of Bluesky's public opening

We conducted a longitudinal analysis of user activity on Bluesky, a decentralized social media platform that experienced rapid growth after opening to the public on February 6th, 2024 [2]. The study focused on the two months surrounding the opening, with the objective of understanding how the platform evolved as it transitioned from an invite-only environment to a publicly accessible social network.

The analysis revealed that Bluesky exhibits several characteristics that are comparable to more established platforms, including a broad and highly skewed distribution of user activity. However, the platform also showed distinctive traits. In particular, the amount of original content was substantially higher than the amount of reshared content, suggesting a communication environment less dominated by reposting dynamics than those typically observed on more mature platforms. At the same time, toxicity levels were found to be very low, indicating that the early growth phase of the platform did not immediately translate into a substantial increase in harmful language.

A major effect of the public opening was the sharp increase in both new users and overall activity. This growth was especially visible in the production of content in English and Japanese, pointing to a geographically and linguistically diverse expansion of the user base. Alongside this rapid growth, the study also detected the emergence of suspicious behaviours. Some accounts were characterized by atypical activity patterns, such as following very large numbers of users and sharing content from low-credibility news sources. These signals are particularly relevant for the goals of DEMON, as they suggest that newly expanding platforms can quickly become targets for manipulation attempts, opportunistic influence operations, or spam-like activity.

Importantly, the analysis also observed that some of the suspicious accounts identified during the study had already been flagged as spam or suspended. This suggests that Bluesky's moderation mechanisms were active during the early public phase and may have contributed to mitigating some of the most visible manipulative behaviours. From the perspective of the project, this study provides a first characterization of a decentralized platform as a potential arena for information manipulation, while also highlighting the importance of timely moderation and platform governance in shaping the health of the information ecosystem.

3 A longitudinal analysis of misinformation, polarization and toxicity on Bluesky after its public launch

Building on the initial conference study, we produced an extended version, currently under consideration in Online Social Networks and Media that broadens the scope of the analysis and provides a more comprehensive picture of Bluesky after its public opening. While the earlier work primarily documented the growth dynamics and general activity patterns of the platform, this extended version explicitly investigates misinformation, political polarization, source credibility, and harmful interactions.

The results confirm some of the main observations of the earlier study while deepening the interpretation of the platform's evolution. Bluesky still appears to be characterized by a higher proportion of original content than reshared content and by generally low toxicity. However, the extended analysis adds an important political dimension: the user base was found to lean predominantly to the left, and resharing communities also displayed a prevalence of left-leaning users. This aspect is relevant because the ideological composition of a platform may shape the types of narratives that gain visibility, the range of media sources that circulate, and the forms of contestation or reinforcement that emerge within public debate.

Another important result concerns the credibility of shared information sources. The analysis found that Bluesky users tend, overall, to share high-credibility sources. This suggests that, in its early public phase, the platform was not dominated by low-quality or overtly misleading content. Nevertheless, suspicious activity was still present. As in the conference version, several accounts that entered the platform after the public launch displayed behaviours associated with manipulation attempts, such as mass-following practices and the sharing of content from low-credibility outlets. Thus, although Bluesky may currently provide a comparatively healthier environment than some mainstream platforms, it is not immune to attempts at exploiting the platform for low-credibility information dissemination.

A further contribution of the journal version lies in its discussion of moderation. The observation that several suspicious accounts were later flagged as spam or removed provides evidence that moderation interventions were taking place and may have played a role in limiting the spread of harmful or manipulative content. This finding is especially important in the context of decentralized social media, where content governance is often assumed to be weaker, more fragmented, or harder to enforce. The results instead suggest that decentralization does not automatically imply the absence of effective moderation, and that platform design, governance choices, and community norms all interact in shaping the quality of online discourse.

Taken together, the conference and journal versions form a coherent line of research within WP3. The first paper offers a timely empirical snapshot of a rapidly expanding platform; the second extends that contribution by connecting user growth to broader questions about misinformation, ideological structure, harmful content, and moderation. For the DEMON project, this line of work is valuable because it provides a real-world case study of how platform architecture and user migration can affect the vulnerability of an online environment to manipulation and abuse.

4 Information diffusion assumptions and their impact on the analysis of social network dynamics

In addition to the empirical study of Bluesky, WP3 also includes a methodological contribution addressing a central issue in the analysis of online information spread: the reconstruction of diffusion cascades [1]. In many social media datasets, the information provided by platforms links reshared content directly to the original post, without revealing the actual path through which users encountered and retransmitted the content. This creates a serious limitation, because the observed data do not necessarily correspond to the true diffusion process.

The study investigates the implications of this problem by comparing analyses based on raw platform data with analyses based on reconstructed cascades. It shows that the common practice of either ignoring reconstruction entirely or adopting simplified assumptions can substantially alter the resulting picture of diffusion. In particular, the identification of influential users changes significantly depending on how cascade reconstruction is handled. Since the detection of influential actors is a key component in the analysis of manipulation, amplification, and coordinated activity, this result has important consequences for the methodological foundations of the project.

The work also proposes a novel reconstruction approach designed to evaluate the effects of different assumptions made during the inference procedure. Through the analysis of more than 40,000 true and false news stories on Twitter, complemented by case studies on Twitter and Bluesky, the paper shows that reconstruction assumptions can drastically distort both microscopic and macroscopic properties of cascade networks. At the microscopic level, this affects which users appear central or influential; at the macroscopic level, it affects the estimated shape, depth, and breadth of diffusion structures. In other words, different assumptions can lead researchers to different conclusions not only about who matters in the spread of information, but also about how information spreads in the first place.

This methodological contribution is directly aligned with the objectives of DEMON. Studies of misinformation, propaganda, manipulation, and coordinated amplification often rely on diffusion analyses to identify influential actors, estimate narrative reach, and understand how content propagates through a network. If the underlying diffusion structure is inaccurately represented, downstream analyses may inherit substantial biases. The paper therefore serves as an important warning against overly simplistic interpretations of platform traces and highlights the need for careful reconstruction strategies, sensitivity analyses, and methodological transparency.

Within WP3, this contribution complements the empirical findings on Bluesky. While the Bluesky papers show how misinformation-related and suspicious behaviours can emerge on a decentralized social platform, the diffusion paper reminds us that even when such behaviours are observed, the computational tools used to interpret their spread must be applied with caution. This dual perspective – empirical and methodological – strengthens the overall contribution of the work package.

5 Discussion and conclusions

The activities presented in this deliverable contribute to WP3 by combining the analysis of an emerging social platform with a methodological reflection on the study of information

diffusion. Together, these works improve our understanding of how online environments evolve, how harmful or manipulative behaviour may appear during phases of rapid growth, and how analytical assumptions can affect the interpretation of social network dynamics.

The studies on Bluesky show that decentralized social media platforms can develop communication patterns that differ from those of more established platforms. In the period immediately following the public launch, Bluesky displayed a high prevalence of original content, low toxicity, and a tendency to circulate high-credibility sources. At the same time, the platform also exhibited signs of vulnerability, with suspicious accounts attempting to exploit the rapid influx of users by engaging in mass-following practices and sharing low-credibility content. The evidence that several of these accounts were later flagged or suspended suggests that moderation can play a meaningful role even in decentralized settings.

The methodological study on diffusion reconstruction adds an essential layer of rigor to this picture. Since the analysis of online manipulation often depends on the identification of influential actors and on the characterization of diffusion cascades, the finding that reconstruction assumptions can significantly distort results is of central importance. It implies that future work within DEMON should not only continue to analyze the content and behaviour associated with manipulation, but also carefully assess the validity of the computational procedures used to infer diffusion pathways and network influence.

Overall, the contributions described here advance the project along two complementary dimensions. On the one hand, they provide empirical evidence about the emergence of misinformation-related and suspicious behaviours on a growing online platform. On the other hand, they offer methodological guidance for ensuring that the analysis of such behaviours is robust and interpretable. These results lay the groundwork for future research in DEMON on narrative spread, influence estimation, harmful content monitoring, and the detection of manipulation strategies across different online ecosystems.

References

Publications Acknowledging the DEMON Project

[1] Matthew R DeVerna, Francesco Pierri, Rachith Aiyappa, Diogo Pacheco, John Bryden, and Filippo Menczer. "Information diffusion assumptions can distort our understanding of social network dynamics". In: *arXiv preprint arXiv:2410.21554* (2024).

[2] Erfan Samieyan Sahneh, Gianluca Nogara, R DeVerna Matthew, Nick Liu, Luca Luceri, Filippo Menczer, Francesco Pierri, Silvia Giordano, et al. "The Dawn of Decentralized Social Media: An Exploration of Bluesky's Public Opening". In: *LECTURE NOTES IN COMPUTER SCIENCE* (2024), pp. 406–421.