Dynamic graph generation to simulate spreading for large online social platforms evaluation

supervision: Mehdi Naima and Lionel Tabourier

1 Context and objectives

In this project, we aim to **model large social platforms and the processes that take place there**, in particular how information spreads and how this affects public exposure to various opinions. For that purpose, we need to describe accurately what makes the interactions on these social platforms particular, in terms of structure and in terms of dynamics. This leads us to generate synthetic structures that mimic realistically their most salient characteristics in order to predict future behaviours. In this way, we aim to contribute to the evaluation of online social platforms, in particular how their practices in terms of moderation and recommendation affect the spreading and exposure to information. It comes within the priority of **building more resilient societies** fostered by PostGenAI@Paris.

Through the lens of complex networks analysis, social networks are seen as large sparse graphs which exhibit interesting structural properties, such as a heterogeneous degree distribution or locally dense regions. Moreover, the description of the dynamical behaviour of these systems reveals another level of complexity: cascading bursty activities, correlated with the structure of the network. Standard graph models and dynamic processes largely fail to reproduce these complex behaviours. However, to simulate the complex phenomena that occur in online social networks, we **need to reproduce realistically the characteristics of these interactions**. We propose to build on the data mining and graph generation methods that have been achieved these last few years to propose scalable, realistic models of these platforms, therefore allowing to simulate these complex processes.

2 Data

The adoption of the Digital Services Act (DSA) by the European Union in November 2022 **aims to create the conditions for a safer and sounder numerical environment** by allowing the public audit of the practices of the very large online platforms (VLOPs) and very large online search engines (VLOSEs). The DSA validates that this public audit goes through making data available for researchers to evaluate if contents are recommended or moderated in agreement with the European legislation. This regulation creates a brand new context for the evaluation of the VLOPs and VLOSEs practices, as in the last few years, they have largely limited and even hindered the access to their data to favour their commercial gains. Data is now available on the European platform and regularly enriched to permit this audit: `https://transparency.dsa.ec.europa.eu/explore-data/overview`.

We believe that an essential role of the academics involved in the field is to take up this issue by **proposing new methods to achieve this public audit**. It would create the conditions to ensure the regulation and ethical use of the AI technologies underpinning the functioning of the VLOPs and VLOSEs, one of the core mission of PostGenAI@Paris. Having access to data through the DSA is an excellent way to investigate the complex entanglement between the structure and dynamics of information diffusion on social platforms. Our first target would be micro-blogging platforms (e.g., X/Twitter), where the spreading of misinformation has been largely documented [2, 9] and the techniques to detect it well-improved. However, the detection is made afterwards, making these systems inoperative as alert warnings.

Our view is that **the simulation of spreading processes on synthetic platforms allow for a better understanding of their dynamics and allow the definition of efficient warnings**. Indeed, having access to the interaction data on these platforms (follower/followee links, re-twitting, citing, etc.) we have the elements to make realistic artificial models and simulate various spreading scenarios and actions that could efficiently limit their negative effects.

3 Challenges and scientific approach

In this PhD project, there are two main methodological aspects to investigate. The first one concerns **mining patterns in the data that allow to characterise their structure and dynamics**. Indeed, while there is

a rich toolbox to identify relevant structural features on graphs, there is still a need to develop their equivalent to characterise synthetically the properties of interaction data through time. Our team acquired an expertise in the design and development of tools to analyse temporal data from real world interaction networks [3, 1, 4, 6]. We plan to use this knowledge for the identification and measurement of the features which are specific to the data collected.

The second aspect concerns generation processes. Simple graph models -that emerged 25 years ago- helped describe some coarse-grain properties observed in real world, such as the heavy-tailed degree distribution or high clustering, but they proved limited when trying to account for more elaborate properties. Consequently, powerful techniques were proposed to generate graph models with stronger properties [5, 8, 7]. But improved access to data uncovered the important role of the temporality of interactions, especially when considering user behaviours online. There is now a **need for realistic models that also account for the dynamics of the interactions**. We think that Monte Carlo Markov Chain methods are probably the best lead to explore because they offer a possibility of unbiased generation even of complex structures. The main challenge in this area is to adapt efficiently these methods to temporal data.

4 Profile required and environment

The PhD would take place at Sorbonne University (Jussieu Campus, Paris 5), under the supervision of Mehdi Naima and Lionel Tabourier. Mehdi Naima (Associate Professor, LIP6, Sorbonne-University/CNRS) His field of research focuses on algorithm analysis as well as temporal graph algorithms (more information at `https://busyweaver.github.io/`). Lionel Tabourier is an Associate Professor (HDR) at LIP6. His field of research is complex networks analysis. During his previous works, he contributed to realistic graph generation methods, as well as the analysis of information spreading on social platforms (more information at `https://lioneltabourier.github.io/`).

We are looking for students whose primary interests are algorithmics and programming for data mining. A taste for interdisciplinary questions is also important.

References

[1] Alexis Baudin, Lionel Tabourier, and Clémence Magnien. Lscpm: Communities in massive real-world link streams by clique percolation method. In *30th International Symposium on Temporal Representation and Reasoning*, 2023.

[2] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3):554–559, 2016.

[3] Matthieu Latapy, Tiphaine Viard, and Clémence Magnien. Stream graphs and link streams for the modeling of interactions over time. *Social Network Analysis and Mining*, 8:1–29, 2018.

[4] Mehdi Naima. Temporal betweenness centrality on shortest walks variants. *(accepted for publication at Applied Network Science), arXiv preprint arXiv:2305.01080*, 2023.

[5] Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2):173–191, 2007.

[6] Felix I Stamm, Mehdi Naima, and Michael T Schaub. Efficient sampling of temporal networks with preserved causality structure. *(accepted for publication at SIAM International Conference on Data Mining (SDM25)), arXiv preprint arXiv:2501.09856*, 2025.

[7] Lionel Tabourier and Julien Karadayi. Probabilistic k-swap method for uniform graph generation beyond the configuration model. *Journal of Complex Networks*, 12(1):cnae002, 2024.

[8] Katherine Van Koevering, Austin Benson, and Jon Kleinberg. Random graphs with prescribed k-core sequences: A new null model for network analysis. In *Proceedings of the Web Conference 2021*, pages 367–378, 2021.

[9] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.