# TFG: Privacy in dynamic graphs
# Informe incial

Guillem Garcia [NIU: 1636279]

Data d'entrega: 10/03/2025

## 1 Introducció i objectius

La protecció de dades s'ha anat convertint en un tema important en l'actualitat, on és un dilema en gairebé tot arreu, com pot ser en el sector de la salut, la tecnologia, les finances, etc. Un dels àmbits on aquest problema és preocupant són les xarxes socials i altres sistemes que generen grans volums de dades i interaccions en temps real, on és fonamental garantir la seguretat i privacitat de la informació.

Aquestes dades sovint es poden modelar mitjançant grafs, on els nodes representen objectes o usuaris, i les arestes defineixen les relacions entre ells dins la xarxa. En moltes ocasions, aquests grafs no són estàtics, sinó dinàmics, ja que els nodes i les arestes poden anar canviant al llarg del temps. Això planteja nous reptes en matèria de privacitat, pel fet que pot facilitar la re-identificació d'usuaris o l'extracció de dades personals. Per aquest motiu, és essencial desenvolupar tècniques adequades per protegir la privacitat tant en grafs dinàmics com en estàtics i mitigar possibles atacs dins la xarxa.

Llavors, l'objectiu principal d'aquest projecte és investigar sobre grafs que varien durant el temps, on es volen entendre els diversos processos per protegir-los, i quines són les propietats que canvien respecte els grafs originals quan s'apliquen els mètodes. Específicament, es volen assolir ordenadament els següents punts:

1. Establir les definicions bàsiques d'un graf temporal i quines propietats addicionals tenen en comparació dels grafs estàtics.

2. Aplicar diversos mètodes de privacitat a diferents conjunts de dades, des de volums de dades fàcils de tractar, fins a una gran quantitat. Pels mètodes de privacitat, es vol entendre com funcionen, i saber quines són les possibles situacions que poden passar si una persona ataca la xarxa.

3. Fer una comparativa a escala de privacitat i utilitat dels grafs protegits i originals. La idea és utilitzar diferents mètriques i algorismes que permetin veure les diferències que es produeixen entre grafs. Els mètodes de

protecció han de ser els més consistents i òptims possibles, on es vol fer un estudi de quins són els millors paràmetres per cada *dataset*.

4. Amb l'ús de les xarxes neuronals, intentar realitzar un model predictiu de grafs. És a dir, que el model pugui predir les següents seqüències de temps. Aquest model s'aplicaria tant en un conjunt de grafs protegits, com en el conjunt de grafs sense protegir, per tal de veure què canvia a nivell de privacitat i utilitat.

Tot i que els objectius estiguin organitzats d'aquesta manera, la màxima prioritat és desenvolupar a fons el segon i el tercer. La resta tenen una importància lleugerament més baixa, però són necessaris per completar i donar coherència al contingut en què es basen.

# 2 Metodologia

L'eina principal per arribar a realitzar els propòsits serà el llenguatge de programació *Python* (versió 3.13.0), utilitzant essencialment les llibreries següents:

- *NetworkX* (versió 3.2.1): Per a la creació dels grafs a partir de les dades dels *datasets*.
- *Pandas* (versió 2.1.4): Per fer conversió de dades en *DataFrames*. L'ús que es donarà principalment serà per fer agregacions.
- *Numpy* (versió 1.26.3): Es farà ús en el moment de fer operacions matricials de forma òptima.
- *Matplotlib* (versió 3.8.2): Utilitzat per visualització de gràfics. Eficient per observar mètriques i comparar resultats.
- *Tensorflow* (versió 2.17.0): Per a crear i entrenar una xarxa neuronal, on es farà servir els grafs temporals com a conjunt d'entrenament i test.

Per tal d'assegurar que els mètodes de privacitat funcionen en tot tipus de problemes, s'ha optat per recol·lectar diversos conjunt de dades de diferents característiques [1, 2]. En la *Taula 1* es mostra la llista de *datasets* que es contenen, i paràmetres que han estat importants en el moment d'escollir-los.

| **Name** | $\|\mathbf{V}\|$ | $\|\mathbf{E}\|$ | **Is directed?** | **Is weighted?** | **#Snapshots** |
|---|---|---|---|---|---|
| Aves-sparrow | 52 | 516 | False | True | 2 |
| Mammalia-voles | 1480 | 4569 | False | False | 61 |
| Insecta-ant | 152 | 194K | False | True | 41 |
| CollegeMsg | 1899 | 59.8K | True | False | 58911 |
| IA-Facebook | 42.4K | 877K | True | True | 867939 |

Taula 1: Llista de datasets cercats, amb paràmetres bàsics dels seus grafs temporals

A part, per comprovar que estan funcionant adequadament els mètodes implementats, l'afegiment de *Unittesting* serà imprescindible, i estarà present en tot moment que s'ha de codificar.

## 3 Planificació

Com a planificació personal, primerament s'ha fet un cronograma de les dates importants del projecte (a nivell de desenvolupament), amb els objectius principals que s'han d'assolir per a aquell moment, com es pot observar en la *Figura 1*. En principi, s'ha estimat que tots els objectius tenen una durada més o menys similar, on els hem separat en parts iguals.

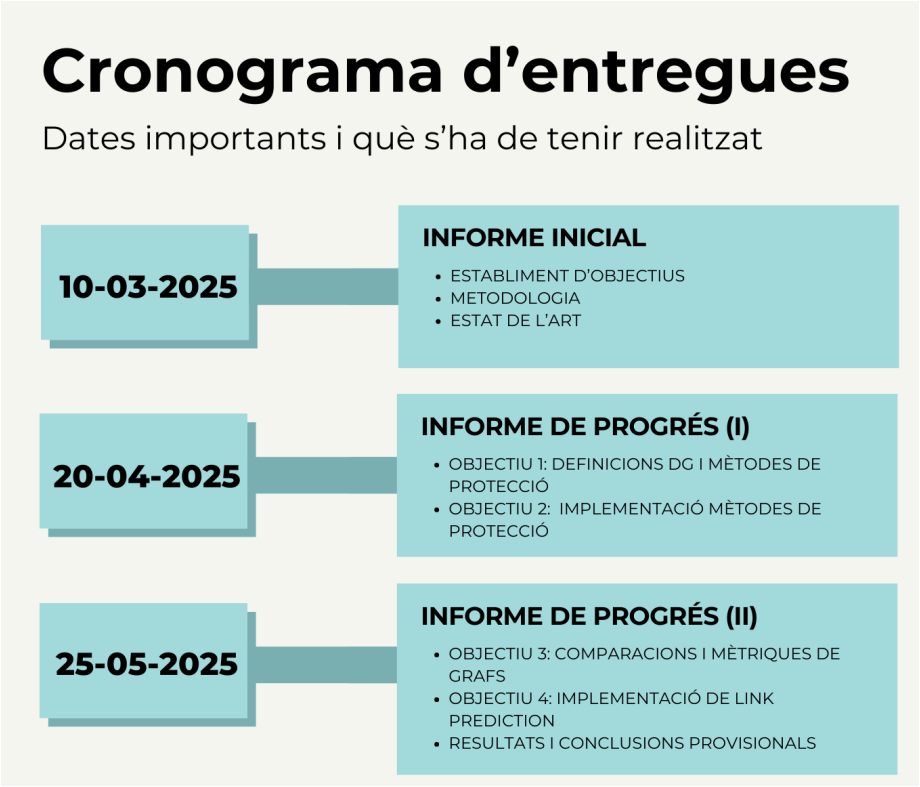


Figura 1: Cronograma d'objectius que s'han de desenvolupar en les dates determinades

A partir d'això, s'ha realitzat un diagrama de *Gantt*, estimant específicament en quin ordre es fan les tasques, i quan s'han de començar i acabar. Tot aquest procés es pot veure en la *Figura 2*.


Figura 2: Diagrama de *Gantt* amb les tasques especificades. Les tasques tenen una data de començament i d'acabament, i estan dividides segons si són a nivell conceptual ($TC$), o bé si són a nivell pràctic ($TP$).

# 4 Estat de l'art

Les maneres de protegir grafs són diverses, com per exemple fer que els nodes siguin indistingibles entre ells a partir dels seus atributs, afegir soroll, estratègies de xifratge, etc. En grafs dinàmics es troben certes semblances amb les maneres de protegir les dades si ho comparem amb grafs estàtics, però tenen una major complexitat, ja que es conté un factor temporal. Principalment, es volen implementar el *K-Anonymity* [3] i el *Edge-Local Differential Privacy* [4] aplicats en grafs temporals. Similarment, en el moment de calcular la privacitat i utilitat, si volem saber la similitud entre grafs, es poden usar diferents versions del *Coeficient de Jaccard* o la *Cosine Similarity* [5, 6], i si cal implementar algorismes de detecció de comunitats com a informació d'utilitat, es tenen diverses estratègies [7]. Per les tècniques de predicció en grafs, es tenen solucions utilitzant *Link Prediction* [8, 9, 10] en xarxes dinàmiques.

# Referències

[1] J. Leskovec, Stanford Network Analysis Project (SNAP). Disponible en: https://snap.stanford.edu/index.html [Darrer accés: 26-feb-2025].

[2] Ryan A. Rossi i Nesreen K. Ahmed, The Network Data Repository with Interactive Graph Analytics and Visualization, 2015. Disponible en: https://networkrepository.com/dynamic.php [Darrer accés: 26-feb-2025].

[3] L. Rossi, M. Musolesi i A. Torsello, "On the k-Anonymization of Time-Varying and Multi-Layer Social Graphs", Proceedings of the International AAAI Conference on Web and Social Media, 9(1), 377-386, 2021. Disponible en: https://ojs.aaai.org/index.php/ICWSM/article/view/14605 [Darrer accés: 26-feb-2025].

[4] S. Paul, J. Salas i V. Torra, "Edge Local Differential Privacy for Dynamic Graphs", In International Symposium on Security and Privacy in Social Networks and Big Data (pp. 224-238). Singapore: Springer Nature Singapore, (2023, Agost).

[5] B. Ruan, J. Gan, H. Wu, i A. Wirth, "Dynamic Structural Clustering on Graphs", arXiv preprint arXiv:2108.11549, 2021. Disponible en: https://arxiv.org/pdf/2108.11549 [Darrer accés: 1-mar-2025].

[6] E. Castrillo, E. León, i J. Gómez, "Dynamic Structural Similarity on Graphs", arXiv preprint arXiv:1805.01419, 2018. Disponible en: https://arxiv.org/pdf/1805.01419 [Darrer accés: 1-mar-2025].

[7] B. Rozemberczki, Awesome Community Detection - Temporal Networks. GitHub. Disponible en: https://github.com/benedekrozemberczki/awesome-community-detection [Darrer accés: 26-feb-2025].

[8] P. Sarkar, D. Chakrabarti i M. Jordan, "Nonparametric Link Prediction in Dynamic Networks", arXiv preprint arXiv:1206.6394, 2012. Disponible en: https://arxiv.org/pdf/1206.6394 [Darrer accés: 2-mar-2025].

[9] X. Li, N. Du, H. Li, K. Li, J. Gao i A. Zhang, "Deep Learning Approach to Link Prediction in Dynamic Networks", SIAM, 2014. Disponible en: https://epubs.siam.org/doi/pdf/10.1137/1.9781611973440.33 [Darrer accés: 2-mar-2025].

[10] J. You, T. Du, J. Leskovec "ROLAND: Graph Learning Framework for Dynamic Graphs", Conference on Knowledge Discovery and Data Mining, 2022. Disponible en: https://arxiv.org/pdf/2208.07239 [Darrer accés: 6-mar-2025]