TFG: Privacy in dynamic graphs
Informe de progrés (I)

Guillem Garcia [NIU: 1636279]

Data d'entrega: 20/04/2025

1 Revisió d'objectius

Dintre dels objectius que es van plantejar en el primer informe, no s'ha realitzat cap canvi destacable respecte als que es van mencionar en aquell moment. Per rememorar, les propostes eren:

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.

De moment no s'ha trobat cap problema amb la definició d'aquests objectius. Tant les tasques que s'han de resoldre com la priorització d'aquestes són les adients per aquest projecte. No obstant, durant els darrers dies s'han obert noves possibilitats que poden ser interessants d'estudiar, i ens poden complementar o substituir objectius en cas de no assolir-los. Un nou repte opcional seria anar a un tema més concret com poden ser les *Lightning Networks*, per veure com es comporta si es protegeixen aquestes xarxes sobre *Bitcoin*.

2 Revisió de la metodologia

En el cas de la metodologia, les eines i forma de fer les tasques són les mateixes a les quals es van proposar. Però sí que s'ha hagut de modificar la llista de *datasets* que s'havia escollit en l'inici. Es van seleccionar aquests conjunts de dades:

Name	**‖V‖**	**‖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* inicials, amb paràmetres bàsics dels seus grafs temporals

En comprovar que els processos de protecció eren molt costosos pel *dataset* de *IA-Facebook*, ja que tenia molts nodes, arestes i una quantitat exagerada de grafs temporals, de moment s'ha optat no utilitzar-lo. S'havia intentat fer que els *datasets* que tenen *Unix Timestamps* (com aquest i *CollegeMsg*), es puguin agrupar per hores, dies i setmanes, per tal de facilitar el procediment i també estudiar com canvien entre ells, però encara així *IA-Facebook* no és viable amb els recursos que es tenen.

Per tant, s'ha incorporat un altre *dataset*, que conté comunicacions entre empleats de l'excompanyia *Enron* [2]. El motiu d'aquesta selecció és que es volia un *dataset* amb una gran quantitat de *timestamps*, però amb un nombre de nodes i connexions reduïdes. Llavors, la llista de *datasets* corregida és la de a continuació:

Name	**‖V‖**	**‖E‖**	**Is directed?**	**Is weighted?**	**#Snapshots**	**#Hour**	**#Day**	**#Week**
Aves-sparrow	52	516	False	True	2	-	-	-
Mammalia-voles	1480	4569	False	False	61	-	-	-
Insecta-ant	152	194K	False	True	41	-	-	-
Enron-Employees	**151**	**50.5K**	**True**	**True**	**16067**	**6440**	**867**	**161**
CollegeMsg	1899	59.8K	True	False	58911	**3320**	**193**	**29**

Taula 2: Llista de *datasets* actuals, amb les columnes *#Hour*, *#Day* i *#Week* afegides, que representa quants *snapshots* té el *dataset* si s'agrupa per hores, dies o setmanes respectivament.

3 Revisió de la planificació

Pel que fa a la planificació, s'està seguint pas a pas el cronograma i diagrama de tasques que es va fer, com es pot observar en la *Figura 1* i *Figura 2*. Les tasques estan durant el que s'imaginava en un primer moment. Per tant, no s'està tenint problemes de temps. Fins ara s'ha desenvolupat els primers dos objectius, que era el que es tenia previst.

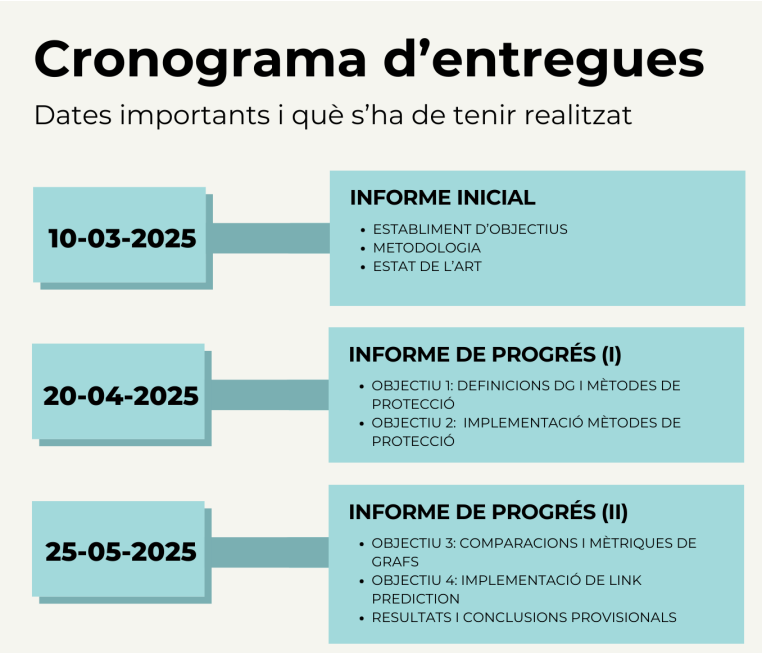

Figura 1: Cronograma d'objectius que s'han de desenvolupar en les dates determinades

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 Desenvolupament

A continuació s'explicarà pas a pas què i com s'han desenvolupat les primeres dues tasques del projecte.

1. Com a començament, es va fer recerca de totes les definicions necessàries que comporten els grafs temporals i els mètodes de privacitat, enfocat tot en termes conceptuals. Sobretot, ha estat d'utilitat [3] i [4], que expliquen amb detall què són els grafs temporals, i dissenyen models de protecció utilitzant *K-degree Anonimity* i *Epsilon Edge-Local Differential Privacy*.

2. Seguidament, es va realitzar un disseny de la implementació dels mètodes de privacitat mitjançant un diagrama de classes. El diagrama correspon a la *Figura 3*, que conté tres mòduls principals: *Reader*, *GraphProtection* i *ModuleManager*.

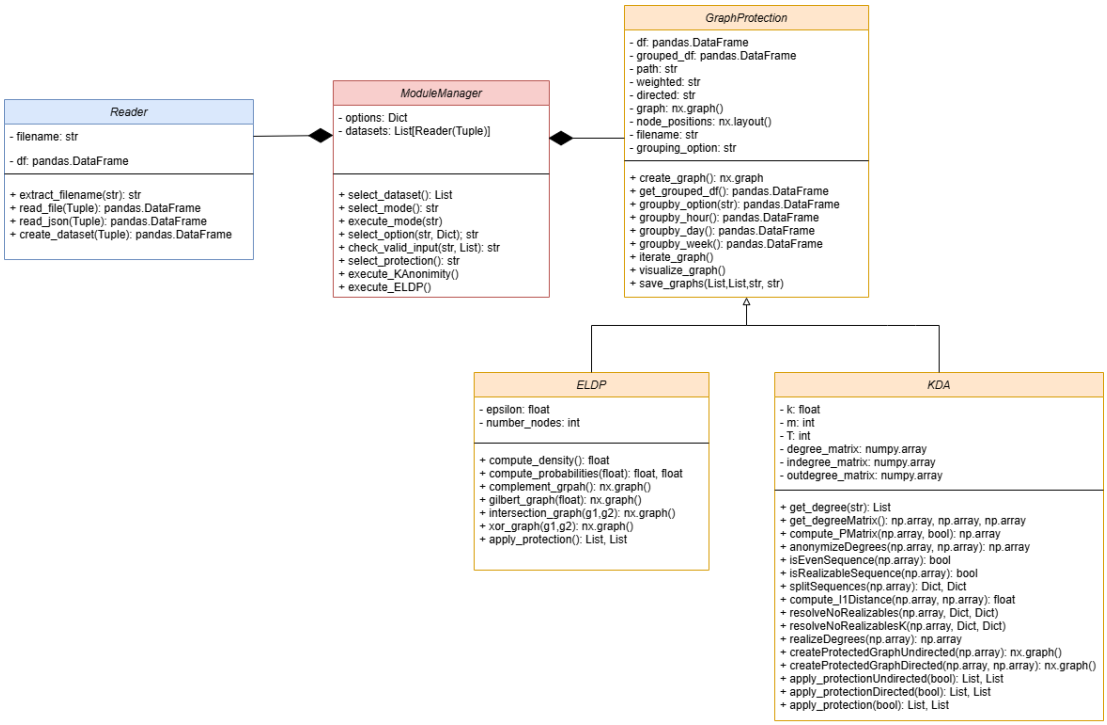

Figura 3: Diagrama de classes per a la implementació dels mètodes de privacitat.

Resumidament, el mòdul *Reader* obté un fitxer i el llegeix, per tal de convertir-lo en un *dataset* de *pandas*. *GraphProtection* agafa els *datasets*, els converteix en grafs temporals, i se'ls hi aplica un mètode de protecció, que pot ser o bé *K-degree Anonimity* (KDA) o bé *Epsilon Edge-Local Differential Privacy* (ELDP). Per últim, es té *ModuleManager*, que serveix per entrellaçar mòduls i executar-los.

3. Per acabar, s'ha implementat el que s'ha dissenyat en *Python*. Per comprovar que tot està funcionant, tenim diversos arxius que fan *Unittesting* de les funcions principals de cada mòdul, i si compleixen les condicions principals per assegurar privacitat en cada graf.

Figura 4: *Unittest* de les funcionalitats dels mòduls implementats.

Els resultats de les proteccions s'han guardat en fitxers utilitzant la llibreria *pickle*, que aquests contenen una llista dels grafs originals, com també els grafs protegits per cert mètode i paràmetre. S'ha fet d'aquesta forma, perquè en el moment de calcular mètriques sigui de forma directa, sense passar novament per l'execució dels mètodes de protecció.

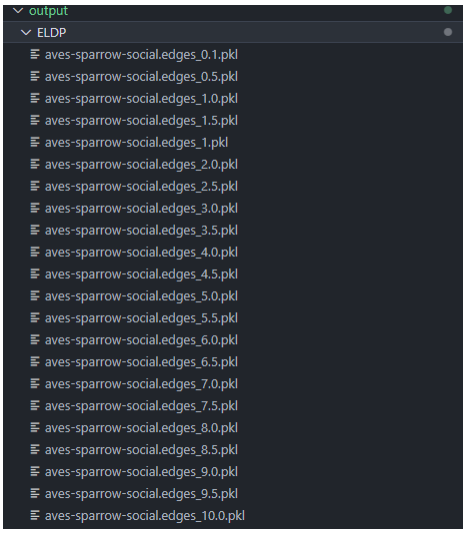

Figura 5: Exemples d'arxius *output* amb format *".pkl"*. En aquest cas, són del mètode *Epsilon Edge-Local Diferential Privacy*, on cada arxiu són els resultats del *dataset aves-sparrow* provant diverses *epsilons*.

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]

[11] "Havel–Hakimi algorithm", Wikipedia, l'enciclopèdia lliure. Disponible en: https://en.wikipedia.org/wiki/Havel–Hakimi_algorithm [Darrer accés: 13-abr-2025].

[12] "Erdős–Gallai theorem", Wikipedia, l'enciclopèdia lliure. Disponible en: https://en.wikipedia.org/wiki/Erdos–Gallai_theorem [Darrer accés: 14-abr-2025].