UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Clemente Arroyo, Joel. *Privadesa diferencial a nivell d'aresta en xarxes neuronals de grafs : compromís entre privadesa i utilitat.* Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Enginyeria de Dades)

This version is available at https://ddd.uab.cat/record/337339

under the terms of the CC BY-SA license.

Privadesa diferencial a nivell d'aresta en xarxes neuronals de grafs: compromís entre privadesa i utilitat

Joel Clemente Arroyo

Resum– Les xarxes neuronals de grafs (GNN) exploten l'estructura relacional de les dades, però aquesta dependència pot filtrar informació privada continguda a les arestes. Aquest treball estudia empíricament el compromís entre privadesa i utilitat d'aplicar privadesa diferencial a nivell d'aresta (edge-DP) sobre l'entrada d'una GNN. S'implementen dos mecanismes edge-DP — Randomized Response asimètrica i Laplace amb selecció top-k, calibrats contra una estimació de densitat privada — i s'avalua la utilitat (classificació de nodes) distingint tres escenaris de desplegament, i la protecció efectiva mitjançant un atac d'inferència d'arestes de caixa negra. Els resultats sobre Cora, Citeseer i Pubmed mostren que el cost de la privadesa depèn críticament de l'escenari: amb el graf real disponible en inferència s'obté una AUC d'atac propera a l'atzar amb poca pèrdua d'utilitat, mentre que en l'escenari estrictament privat utilitat i protecció entren en conflicte.

Paraules clau– Atac d'inferència d'arestes, compromís privadesa-utilitat, edge-DP, privadesa diferencial, randomized response, xarxes neuronals de grafs

Abstract– Graph neural networks (GNNs) exploit the relational structure of data, but this dependency can leak private information held in the graph's edges. This work empirically studies the privacy-utility trade-off of applying edge-level differential privacy (edge-DP) to the input of a GNN. Two edge-DP mechanisms — asymmetric Randomized Response and a Laplace mechanism with top-k selection, calibrated against a private density estimate — are implemented; utility (node classification) is evaluated across three deployment scenarios, and effective protection through a black-box edge-inference attack. Results on Cora, Citeseer and Pubmed show that the cost of privacy critically depends on the scenario: when the real graph is available at inference the attack AUC drops near random with little utility loss, whereas in the strictly private scenario utility and protection directly conflict.

Keywords– Differential privacy, edge-DP, edge-inference attack, graph neural networks, privacy-utility trade-off, randomized response

1 INTRODUCCIÓ

LES xarxes neuronals sobre grafs (Graph Neural Networks, GNN) s'han consolidat com l'eina de referència per a tasques d'aprenentatge automàtic sobre dades estructurades en forma de graf, com ara xarxes socials, xarxes de citacions o grafs de relacions biològiques. La seva potència rau precisament en l'explotació de l'estructura relacional: a diferència d'un model tabular, una GNN propaga informació entre nodes veïns mitjançant el mecanisme de *message passing* (pas de missatges, en què cada node intercanvia i agrega informació amb els seus veïns), de manera que la predicció sobre un node incorpora el senyal dels nodes amb què està connectat. Un exemple paradigmàtic és la classificació de nodes en una xarxa de citacions, on es vol predir l'àrea temàtica d'un article (un node) aprofitant no només el seu contingut, sinó també les publicacions que cita o que el citen (els seus veïns); és precisament la tasca sobre la qual s'articula aquest treball.

Aquesta dependència de l'estructura, però, comporta un risc de privadesa sovint subestimat. En molts dominis sensibles, les **arestes** del graf són la informació confidencial: en una xarxa social, una aresta pot revelar una amistat o un contacte; en un graf mèdic, una relació entre pacients; en una xarxa de comunicacions, qui parla amb qui. Encara que les característiques dels nodes es publiquin de manera agregada o anonimitzada, l'estructura de connexions pot filtrar informació privada. S'ha demostrat a la literatura que un adversari amb accés a un model entrenat, o fins i tot només a representacions derivades del graf, pot reconstruir parcialment les arestes originals [1].

La **privadesa diferencial** (Differential Privacy, DP) pro-

• E-mail de contacte: Joel.Clemente@autonoma.cat
• Treball tutoritzat per: Guillermo Navarro Arribas i Julián Salas Piñón (Àrea de Ciències de la Computació i Intel·ligència Artificial)
• Curs 2025/26

porciona un marc formal i rigorós per quantificar i limitar aquesta fuita d'informació. Concretament, la **privadesa diferencial a nivell d'aresta** (edge-DP) garanteix que la presència o absència d'una aresta individual tingui un efecte acotat sobre la sortida del mecanisme, fet que ofereix una protecció matemàticament demostrable a cada relació del graf. El paràmetre ε controla aquest compromís: valors petits impliquen més soroll i més privadesa; valors grans, menys soroll i més utilitat.

Aquest projecte aborda la pregunta central que sorgeix d'aquesta tensió: quin és el compromís real entre privadesa i utilitat quan apliquem edge-DP sobre l'entrada d'una GNN, i quina protecció efectiva ofereix davant d'un atac de reconstrucció d'arestes? Per respondre-la, es pertorba el graf d'entrada amb dos mecanismes edge-DP (Randomized Response i un mecanisme basat en soroll de Laplace), s'entrena una GNN sobre el graf protegit i se'n mesura tant la utilitat (precisió en la classificació de nodes) com la resistència a un atac d'inferència d'arestes.

1.1 Objectius

L'**objectiu general** del treball és avaluar empíricament el compromís entre privadesa i utilitat que comporta protegir l'estructura d'un graf mitjançant privadesa diferencial a nivell d'aresta abans d'entrenar-hi una xarxa neuronal de grafs, i quantificar la protecció real que aquesta tècnica ofereix davant d'un atac de reconstrucció d'arestes.

Aquest objectiu general es concreta en cinc objectius específics (OE):

- **OE1 — Formalitzar el marc d'edge-DP sobre grafs.** Definir rigorosament la noció de grafs veïns a nivell d'aresta, la sensibilitat de les consultes estructurals implicades i la composició del pressupost de privadesa, de manera que tot mecanisme implementat tingui una garantia formal ε-edge-DP demostrable.

- **OE2 — Establir un model de referència sense privadesa.** Implementar i validar una arquitectura GNN estàndard sobre la tasca de classificació de nodes, que serveixi de cota superior d'utilitat contra la qual mesurar la degradació introduïda pels mecanismes de privadesa.

- **OE3 — Dissenyar i implementar dos mecanismes ε-edge-DP.** Desenvolupar un mecanisme de Randomized Response i un mecanisme de Laplace que pertorbin l'estructura del graf, amb un calibratge del pressupost correcta per a grafs dispersos i verificable experimentalment.

- **OE4 — Quantificar el compromís privadesa–utilitat.** Mesurar la pèrdua d'utilitat de la GNN en funció de ε, distingint explícitament tres escenaris de desplegament segons quin graf — pertorbat o real — s'utilitza per entrenar, validar i avaluar, per tal d'aïllar l'efecte del soroll en cada fase.

- **OE5 — Avaluar la protecció efectiva mitjançant un atac d'inferència.** Dissenyar un atac de reconstrucció d'arestes amb un model d'adversari realista (caixa negra, sense coneixement de cap aresta real) i mesurar, amb una mètrica invariant a la prevalença com a referència principal, fins a quin punt cada mecanisme impedeix recuperar l'estructura original.

1.2 Planificació i desenvolupament del treball

El treball s'ha dut a terme seguint la planificació temporal prevista per l'assignatura de Treball de Fi de Grau, amb una dedicació aproximada de 300 hores distribuïdes al llarg de les 18 setmanes del semestre. El desenvolupament s'ha organitzat en fases consecutives que cobreixen des de la fonamentació teòrica i la formalització del problema fins a la implementació dels mecanismes, l'execució dels experiments i la redacció del document.

El projecte s'ha desenvolupat de manera autònoma, amb reunions periòdiques amb els directors per consensuar les decisions de disseny, validar els resultats parcials i ajustar l'abast del treball. Fruit d'aquest seguiment, l'abast es va concretar al voltant dels dos mecanismes edge-DP que actuen directament sobre l'estructura del graf (Randomized Response i Laplace), que constitueixen el nucli conceptual de l'estudi. El calendari detallat de fases, amb la dedicació estimada de cadascuna, es recull a l'Annex A.

1.3 Contingut

La resta del document s'organitza de la manera següent. La Secció 2 presenta el context teòric i l'estat de l'art. La Secció 3 descriu la metodologia: conjunt de dades, mecanismes, model i disseny experimental. La Secció 4 detalla l'atac d'inferència. La Secció 5 presenta i discuteix els resultats, i la Secció 6 en recull les conclusions i línies de continuació.

2 Context i estat de l'art

2.1 Definicions

Xarxes neuronals sobre grafs. Una GNN opera sobre un graf $G = (V, E)$, on V és el conjunt de nodes i $E \subseteq V \times V$ el conjunt d'arestes, amb una matriu de característiques de nodes $X \in \mathbb{R}^{|V| \times d}$. En aquest treball es consideren **grafs no dirigits**, de manera que una aresta $\{i, j\}$ no té orientació i la matriu d'adjacència A és simètrica. Convé introduir aquí el concepte de **no-aresta**: quan entre dos nodes i i j no hi ha cap aresta, direm que aquests formen una no-aresta, és a dir, una no-aresta es defineix com l'absència d'aresta entre un parell de nodes. Aquesta noció serà central en la resta del treball, ja que els mecanismes de privadesa actuen sobre el conjunt complet de parells de nodes, distingint arestes de no-arestes. El paradigma dominant és el *message passing*: a cada capa, la representació d'un node s'actualitza agregant les representacions dels seus veïns. Concretament, la Graph Convolutional Network (GCN) de Kipf i Welling [2] defineix la propagació com

$$H^{(l+1)} = \sigma\Big(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\Big), \qquad (1)$$

on $\tilde{A} = A + I$ és la matriu d'adjacència amb autoconnexions (*self-loops*, és a dir, una aresta de cada node a si mateix, afegida perquè cada node també tingui en compte la seva pròpia representació en l'agregació), $\tilde{D}$ la seva matriu de graus, $W^{(l)}$ els pesos entrenables de la capa l i σ una no-linealitat. La conseqüència clau per a aquest treball és que la predicció sobre cada node depèn explícitament de A, és a dir, de l'estructura d'arestes: si l'estructura és informació sensible, la sortida del model pot filtrar-la.

Privadesa diferencial. Un mecanisme aleatori $\mathcal{M}$ satisfà ε-privadesa diferencial si, per a tota parella de bases

de dades veïnes D, D' que difereixen en un únic registre i per a tot conjunt de sortides S, es compleix [3]

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S]. \quad (2)$$

El paràmetre $\varepsilon \geq 0$ (el *pressupost de privadesa*) acota la influència d'un sol registre sobre la distribució de sortida: com més petit és ε, més indistingibles són les sortides $\mathcal{M}(D)$ i $\mathcal{M}(D')$, i per tant més forta és la garantia. La DP és immune al postprocessament (qualsevol funció aplicada a una sortida ε-DP segueix sent ε-DP) i és composable (l'aplicació seqüencial de mecanismes ε_1- i ε_2-DP és $(\varepsilon_1+\varepsilon_2)$-DP), propietats que aquest treball explota.

Privadesa diferencial a nivell d'aresta. En dades de graf, la noció de "registre veí" admet dues interpretacions [4]. Sota privadesa diferencial a nivell d'aresta (edge-DP), dos grafs es consideren veïns si difereixen exactament en una aresta; sota privadesa a nivell de node (node-DP), si difereixen en un node i totes les seves arestes incidents. La node-DP és una garantia substancialment més forta, ja que protegeix simultàniament la presència d'un node i totes les seves arestes incidents, però sovint és inviable en grafs dispersos perquè exigeix afegir uns nivells de soroll que destrueixen la utilitat. Aquest treball adopta l'edge-DP perquè el seu objectiu és precisament protegir la confidencialitat de les relacions (arestes) del graf, mantenint alhora una utilitat acceptable.

Formalment, un mecanisme $\mathcal{M}$ és ε-edge-DP si per a tot parell de grafs G, G' que difereixen en una sola aresta i per a tot subconjunt de sortides possibles $S \subseteq \mathrm{Im}(\mathcal{M})$ es compleix

$$\Pr[\mathcal{M}(G) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(G') \in S]. \quad (3)$$

Sensibilitat i mecanismes base. La quantitat de soroll necessària per garantir DP es calibra segons la *sensibilitat* de la consulta, és a dir, el màxim canvi en la seva sortida en passar a un graf veí. Per a una consulta (o funció) f, denotem la seva sensibilitat com Δ_f. El mecanisme de Laplace [3] afegeix a una consulta numèrica f de sensibilitat Δ_f un soroll que segueix una distribució de Laplace de mitjana 0 i escala Δ_f/ε, que denotem $\mathrm{Lap}(\Delta_f/\varepsilon)$. La Randomized Response [5], originada en enquestes sobre temes delicats, aleatoritza una resposta binària de manera que es preservi la negació plausible; aplicada a cada entrada de la matriu d'adjacència, fer un *flip* de la presència de cada aresta amb una probabilitat calibrada produeix un mecanisme edge-DP.

2.2 Estat de l'art

La distinció entre edge- i node-DP va ser introduïda per Hay et al. [4], mentre que els primers algorismes edge-DP es deuen a Nissim et al. [6]. A partir d'aquests fonaments, l'aplicació de la privadesa diferencial a les GNN ha cristal·litzat en diverses famílies d'enfocaments, que es poden ordenar segons en quin punt del pipeline s'introdueix la garantia.

Pertorbació de l'entrada (graf protegit). La línia més directa, i la que adopta aquest treball, consisteix a protegir el graf d'entrada abans de qualsevol entrenament, de manera que la garantia edge-DP es propaga per immunitat al postprocessament a tot el que se'n derivi. La Randomized Response sobre la matriu d'adjacència és l'instrument canònic d'aquesta família i s'ha estudiat extensament en el context de la DP local sobre grafs. El seu avantatge és la simplicitat i la garantia neta sobre l'artefacte publicat; el seu desavantatge, que en grafs dispersos el soroll pot dominar el senyal estructural.

Pertorbació de l'agregació. GAP, de Sajadmanesh et al. [7], introdueix soroll directament en la funció d'agregació de la GNN, de manera que es difumina estadísticament la presència d'una aresta (edge-level) o d'un node i les seves arestes (node-level), oferint garanties formals amb una degradació de precisió competitiva. Aquesta línia tendeix a preservar més utilitat que la pertorbació de l'entrada, a canvi de modificar l'arquitectura del model. Treballs posteriors n'han refinat el compromís: ProGAP [8] substitueix el desacoblament de GAP per un entrenament progressiu per etapes que recupera capacitat expressiva sense incórrer en costos de privadesa excessius, i Chien et al. [9] proposen convolucions desacoblades amb garanties de protecció topològica a diferents granularitats.

Pertorbació del gradient. L'adaptació de DP-SGD (*Differentially Private Stochastic Gradient Descent*, descens estocàstic del gradient amb privadesa diferencial) a les GNN aplica soroll i retalla els gradients durant l'entrenament. El repte específic en grafs és que el *message passing* fa que un sol node o aresta influeixi en múltiples termes de la pèrdua, complicant la comptabilitat de privadesa; els treballs a nivell de node aborden aquesta dependència mitjançant mostreig de veïns acotat [10]. Aproximacions més recents acoten la sensibilitat amb esquemes desacoblats o de propagació restringida per assolir privadesa a nivell de node amb millors compromisos d'utilitat [11], [12].

Atacs d'inferència sobre grafs. En paral·lel a les defenses, s'ha demostrat que les GNN filtren informació estructural [13], [14]. LinkTeller, de Wu et al. [1], recupera arestes privades d'una GNN mitjançant anàlisi d'influència, mostrant que un adversari amb accés a consultes al model pot reconstruir connexions amb alta precisió. En la mateixa línia, He et al. [15] formulen atacs de robatori d'enllaços que, amb accés de caixa negra a les sortides del model, infereixen si existeix una aresta entre qualsevol parell de nodes, i altres treballs estenen aquesta amenaça a la inferència de pertinença a nivell de node [16], [17]. Aquesta mena d'atacs motiven la necessitat de quantificar empíricament no només la pèrdua d'utilitat, sinó també la protecció efectiva que un mecanisme ofereix.

Posicionament d'aquest treball. Davant d'aquest panorama, la contribució del projecte no és un mecanisme edge-DP nou, sinó l'adaptació i l'avaluació controlada de mecanismes de pertorbació de l'entrada pensats per a grafs dispersos. En aquest sentit, el treball aporta tres elements: (i) l'adaptació de dos mecanismes edge-DP — una formulació *asimètrica* de Randomized Response i un mecanisme de Laplace amb selecció top-k, tots dos calibrats contra una estimació de densitat privada — que preserven la densitat del graf i n'eviten el col·lapse cap a un graf dens, a diferència de les versions simètriques clàssiques, que en grafs dispersos generen centenars de milers de falses arestes i destrueixen la utilitat; (ii) la distinció explícita de tres escenaris de desplegament segons quin graf — pertorbat o real — s'utilitza per entrenar, validar i avaluar, que aïlla l'efecte del soroll en cada fase; i (iii) l'avaluació de la protecció efectiva mitjançant un atac d'inferència d'arestes amb un model d'adversari realista de caixa negra, sense cap coneixement d'arestes reals.

3 METODOLOGIA

3.1 Conjunt de dades

Els experiments s'han dut a terme sobre tres grafs de citacions científiques estàndard en l'avaluació de GNN: Cora,

Citeseer i Pubmed [18]. Cora és el conjunt principal sobre el qual es desenvolupa l'anàlisi completa; Citeseer i Pubmed s'incorporen com a test de robustesa, per verificar que les conclusions es generalitzen més enllà d'un únic graf. La Taula 1 en resumeix les característiques.

TAULA 1: CARACTERÍSTIQUES DELS TRES CONJUNTS DE DADES (CORA, CITESEER I PUBMED).

Conjunt	Nodes	Arestes	Classes	Densitat
Cora	2.708	5.278	7	$1{,}44 \times 10^{-3}$
Citeseer	3.327	4.552	6	$0{,}82 \times 10^{-3}$
Pubmed	19.717	44.324	3	$0{,}23 \times 10^{-3}$

Tots tres comparteixen el caràcter dispers i homofílic propi de les xarxes de citacions, però difereixen en escala (Pubmed és un ordre de magnitud més gran), en nombre de classes (de 3 a 7) i en densitat. Aquesta diversitat permet avaluar fins a quin punt les conclusions depenen de les característiques concretes del graf.

La densitat d'un graf $G = (V, E)$ es defineix com $d_G = \frac{|E|}{\binom{|V|}{2}}$ i representa la proporció d'arestes presents respecte a totes les possibles parelles de nodes. En les xarxes de citacions aquesta proporció és minúscula (Taula 1): la immensa majoria de parelles són no-arestes. Aquest fet condiciona el disseny dels mecanismes (Secció 3.2), perquè tant la Randomized Response com el soroll de Laplace, aplicats directament sobre totes les entrades de la matriu d'adjacència, tendeixen a afegir més arestes falses que no pas a preservarne d'autèntiques, densificant el graf protegit i destruint-ne l'estructura.

Sobre el conjunt de nodes etiquetats de cada graf s'ha definit una partició fixa 60/20/20 (entrenament/validació/test) amb llavor controlada, garantint la mateixa partició en totes les execucions per assegurar comparabilitat. Cal precisar que aquesta partició s'aplica únicament sobre els nodes que ja apareixien marcats com a etiquetats en la divisió original del conjunt de dades (és a dir, els nodes que la versió estàndard inclou en alguna de les màscares d'entrenament, validació o test); la resta de nodes intervenen en la propagació del *message passing* però no en el càlcul de la pèrdua ni de les mètriques.

3.2 Mecanismes Edge-DP

Tots dos mecanismes operen sobre el triangle superior de la matriu d'adjacència (atès que el graf és no dirigit, $\binom{|V|}{2}$ entrades) i comparteixen una mateixa estratègia en dues etapes. La idea general és la següent: abans de pertorbar l'estructura, cal saber quantes arestes ha de tenir aproximadament el graf resultant, perquè l'objectiu no és només protegir les arestes individuals, sinó també que el graf protegit conservi una densitat semblant a la de l'original. Per això, en una primera etapa s'estima la densitat del graf de manera privada i, en una segona etapa, s'aplica el mecanisme de pertorbació estructural (Randomized Response o Laplace) calibrat contra aquesta estimació.

Aquesta preocupació per preservar la densitat és especialment crítica en grafs dispersos com Cora. Quan la immensa majoria de parells de nodes són no-arestes, els mètodes edge-DP més clàssics —que tracten igual totes les entrades de la matriu d'adjacència— tendeixen a generar moltes més arestes de les reals i, a la pràctica, no preserven la densitat original: el graf protegit acaba sent molt més dens que el de partida, fet que en destrueix l'estructura. Estimar primer la densitat i forçar després que el graf resultant s'hi ajusti és precisament el que permet evitar aquest col·lapse.

Com que la densitat és una quantitat sensible (depèn directament del nombre real d'arestes), la seva estimació també ha de consumir pressupost de privadesa. Per això el pressupost total ε es reparteix en dues fraccions, ε_1 per a la consulta de densitat i ε_2 per a la pertorbació estructural, de manera que pel teorema de composició seqüencial el mecanisme complet satisfà ε-edge-DP. La proporció concreta d'aquest repartiment es justifica al final de la secció; per ara n'hi ha prou amb saber que les dues etapes consumeixen, sumades, el pressupost total disponible ($\varepsilon = \varepsilon_1 + \varepsilon_2$).

Consulta de densitat privada. Tots dos mecanismes estimen primer la densitat del graf de forma privada. La densitat d_G és una consulta de sensibilitat $\Delta = 1/\binom{|V|}{2}$ (afegir o treure una aresta modifica d_G en exactament aquesta quantitat), de manera que s'hi aplica el mecanisme de Laplace amb escala Δ/ε_1:

$$\tilde{d} = \mathrm{clip}\big(d_G + \mathrm{Lap}(\Delta/\varepsilon_1),\ 10^{-6},\ 1\big)\,, \tag{4}$$

on $\mathrm{clip}(x, a, b) = \min(\max(x, a), b)$ retalla el resultat a l'interval $[10^{-6}, 1]$ per garantir que $\tilde{d}$ sigui una densitat vàlida. Aquesta densitat privada $\tilde{d}$ és l'única quantitat sensible que s'utilitza per calibrar la fase estructural; és deliberadament una quantitat DP, no el recompte real d'arestes, de manera que el pressupost ε_1 es consumeix efectivament i els paràmetres de la fase següent no depenen de cap valor sensible exacte. Aquesta és una correcció metodològica important respecte a un calibratge ingenu que utilitzés el valor real de $|E|$, cosa que violaria la garantia de DP.

Mecanisme 1: Randomized Response asimètrica. En la Randomized Response, cada entrada de la matriu d'adjacència es conserva amb una certa probabilitat (l'anomenada *probabilitat de retenció*, la probabilitat de no invertir el valor de l'entrada) i s'inverteix amb la probabilitat complementària. La formulació simètrica clàssica de RR fa servir una única probabilitat de retenció $p = e^{\varepsilon_2}/(e^{\varepsilon_2} + 1)$, igual per a arestes i no-arestes, i és inadequada en grafs dispersos: aplicada a les $\sim$3,66 milions de parelles de nodes possibles —de les quals pràcticament totes són no-arestes— generaria centenars de milers de falses arestes i destruiria l'estructura. Per evitar-ho, s'adopta una formulació asimètrica amb dues probabilitats — p_1 (retenir una aresta) i p_0 (retenir una no-aresta) — determinades per dues condicions: (i) que la ràtio de DP es mantingui exacta, $p_1/(1 - p_0) = e^{\varepsilon_2}$, i (ii) que el nombre esperat d'arestes es preservi respecte a l'estimació privada $\tilde{E} = \tilde{d} \cdot \binom{|V|}{2}$. Resolent el sistema:

$$1 - p_0 = \frac{\tilde{E}}{e^{\varepsilon_2}\tilde{E} + \left(\binom{|V|}{2} - \tilde{E}\right)}, \quad p_1 = e^{\varepsilon_2}(1 - p_0). \tag{5}$$

Cada entrada de la matriu d'adjacència rep llavors un *flip* de manera independent segons aquestes probabilitats. Com que p_0 i p_1 depenen únicament de $\tilde{d}$ (quantitat DP obtinguda amb pressupost ε_1) i de ε_2, el mecanisme complet és ε-edge-DP per composició seqüencial de les dues etapes.

Mecanisme 2: Laplace amb selecció top-k. Aquest mecanisme afegeix soroll de Laplace directament a cada entrada del triangle superior de la matriu d'adjacència. La sensibilitat d'una entrada és $\Delta = 1$ (un canvi d'aresta altera una entrada en exactament 1), de manera que l'escala del soroll és $1/\varepsilon_2$:

$$\tilde{A}_{ij} = A_{ij} + \mathrm{Lap}(1/\varepsilon_2). \tag{6}$$

Com que la matriu de soroll és contínua, cal binaritzar-la. En lloc d'un llindar fix, es conserven les k entrades amb el valor sorollós més alt, on $k = \mathrm{round}\big(\tilde{d} \cdot \binom{|V|}{2}\big)$ ve fixat per la densitat privada. Aquesta tria té dues virtuts: (i) garanteix que la densitat final coincideixi amb l'objectiu privat $\tilde{d}$ en comptes de dependre de la cua de la distribució del soroll, i (ii) és vàlida des del punt de vista de DP, ja que k depèn només de $\tilde{d}$ (consulta ε_1) i la selecció és funció exclusivament de la matriu sorollosa (sortida del mecanisme de Laplace, ε_2); el graf binari final és, doncs, un postprocessament del parell $(\tilde{A}, \tilde{d})$, que per composició és ε-edge-DP.

És oportú observar que tots dos mecanismes materialitzen la matriu d'adjacència densa, amb cost $O(|V|^2)$ en temps i memòria. Per a Cora aquest cost és perfectament assumible; en grafs substancialment més grans constituiria una limitació pràctica que requeriria una implementació dispersa.

Repartiment del pressupost de privadesa. Establerts els dos mecanismes, queda per justificar en quina proporció es divideix el pressupost ε entre la consulta de densitat (ε_1) i la pertorbació estructural (ε_2). Aquesta proporció no és arbitrària, sinó el resultat d'una exploració empírica. Es van assajar diversos repartiments (50/50, 80/20, 90/10 i 99/1, en fraccions estructura/densitat) i es va observar que la utilitat del model millorava de manera consistent a mesura que es destinava una fracció més gran del pressupost a la pertorbació estructural. La raó és la disparitat de sensibilitats entre les dues consultes: la consulta de densitat té una sensibilitat de $1/\binom{|V|}{2} \approx 2{,}7 \times 10^{-7}$, de manera que una fracció mínima del pressupost ($\varepsilon_1 = 0{,}01\,\varepsilon$) ja proporciona una estimació de densitat $\tilde{d}$ prou precisa en el rang de ε d'interès; en canvi, la pertorbació estructural opera sobre una consulta de sensibilitat 1, molt més exigent en soroll, de manera que cada unitat addicional de pressupost que se li assigna ($\varepsilon_2 = 0{,}99\,\varepsilon$) es tradueix directament en més senyal topològic preservat. El repartiment adoptat en tots els experiments és, doncs, el que destina el 99% del pressupost a l'estructura i l'1% a la densitat.

3.3 Model i procediment d'entrenament

Com a model s'utilitza una GCN de dues capes [2] amb 64 dimensions ocultes i activació ReLU (*rectified linear unit*), entrenada amb l'optimitzador Adam (*adaptive moment estimation*, taxa d'aprenentatge 0,01) i pèrdua d'entropia creuada. L'entrenament incorpora *early stopping* (aturada anticipada quan la mètrica de validació deixa de millorar) sobre la precisió de validació amb una paciència de 20 èpoques (màxim 500), restaurant els pesos del millor punt de control abans de l'avaluació final. S'ha optat per una arquitectura sense *dropout* (tècnica de regularització que desactiva aleatòriament una fracció de neurones durant l'entrenament), atès que en el règim experimental considerat no s'observa sobreajustament rellevant; aquesta decisió manté el model deliberadament simple i facilita l'atribució de la degradació de rendiment a la pertorbació de privadesa i no a la regularització.

3.4 Disseny experimental: tres escenaris de desplegament

L'aportació metodològica central d'aquest treball és la distinció de tres escenaris segons quin graf — el pertorbat $\tilde{G}$, sortida de qualsevol dels dos mecanismes edge-DP, o el real G — intervé en cada fase del cicle de vida del model. En tots tres, l'entrenament es fa sempre sobre el graf pertorbat (la GNN mai veu el graf real durant l'aprenentatge dels pesos); el que canvia és el graf emprat per a la selecció de model (validació) i per a l'avaluació final:

- **S1 (privadesa total):** entrenament, validació i avaluació sobre $\tilde{G}$. Representa el desplegament estrictament privat, en què el graf real no s'utilitza mai en cap fase.
- **S2 (validació privada, inferència real):** entrenament i validació sobre $\tilde{G}$, però avaluació sobre G. Modela el cas en què el desenvolupament del model és privat, però en producció el model s'aplica sobre el graf real. És, de fet, l'escenari més comú i realista en la pràctica: el que es comparteix o es publica amb garanties de privadesa és el graf emprat per desenvolupar el model, mentre que en explotació cada usuari aplica el model sobre les dades reals de què disposa.
- **S3 (selecció amb graf real):** entrenament sobre $\tilde{G}$, però validació i avaluació sobre G. La selecció de model utilitza ja el graf real, cosa que proporciona una cota superior del rendiment que els pesos apresos sobre el graf pertorbat poden assolir sobre el graf real.

Aquesta separació permet aïllar dos efectes que sovint es confonen: la degradació deguda a *aprendre* sobre estructura sorollosa (present en tots tres) i la deguda a *avaluar* sobre estructura sorollosa (només a S1). Comparar S1 amb S2 i S3 quantifica fins a quin punt la utilitat es recupera quan el graf real torna a estar disponible en inferència.

Per fer la implementació eficient, els tres escenaris es deriven d'un únic bucle d'entrenament: com que la validació no influeix en els gradients sinó només en la selecció del punt de control, n'hi ha prou amb mantenir dos rastrejadors d'*early stopping* independents (un sobre la validació de $\tilde{G}$, per a S1/S2, i un sobre la validació de G, per a S3), cadascun congelat en exhaurir la seva paciència. Això produeix exactament els mateixos punts de control que tres entrenaments separats, amb un terç del cost.

3.5 Protocol d'avaluació

Utilitat. Es reporten la precisió (*accuracy*) i la F_1 ponderada sobre el conjunt de test, com a mitjana i desviació estàndard mostral sobre 10 execucions independents amb llavors diferents. Com a referència s'inclou un *baseline* sense privadesa, entrenat i avaluat íntegrament sobre G. S'escombra un rang ampli de pressupostos, $\varepsilon \in \{0,1;\ 1;\ 2;\ 3;\ 4;\ 5;\ 6;\ 7;\ 8;\ 9;\ 10;\ 11;\ 12;\ 13;\ 14;\ 15\}$, per caracteritzar la corba privadesa-utilitat. Cal notar que les execucions comparteixen la mateixa seqüència de llavors entre valors de ε i entre mecanismes; es tracta, doncs, d'un disseny aparellat que redueix la variància de les comparacions a costa de correlacionar-les, cosa que cal tenir present en interpretar les barres d'error.

Privadesa. La protecció efectiva es mesura amb l'atac que es detalla a la Secció 4. La mètrica principal és l'AUC-ROC, invariant a la prevalença, amb interpretació directa: 0,5 indica protecció perfecta (l'adversari no supera l'atzar) i 1 indica filtració total. Es reporta també l'*Average Precision* (AP), que s'ha d'interpretar sempre contra la prevalença del conjunt avaluat.

4 ATAC D'INFERÈNCIA D'ARESTES

L'atac avalua el grau de protecció que ofereixen els mecanismes davant d'un adversari que intenta inferir les arestes originals a partir del graf publicat. El model d'adversari és de **caixa negra sense coneixement de cap aresta real**: l'atacant observa únicament el graf publicat $\tilde{G}$, les característiques (públiques) dels nodes i els *embeddings* (representacions vectorials internes que la xarxa aprèn per a cada node) d'una GCN entrenada sobre $\tilde{G}$. No coneix el graf original ni cap aresta real.

La conseqüència d'aquest model és que l'adversari **no entrena cap classificador supervisat**, perquè fer-ho exigiria conèixer la *ground truth* de quines parelles són arestes. En lloc d'això, calcula una *puntuació de versemblança d'aresta* per a cada parella de nodes utilitzant només $\tilde{G}$, i ordena totes les parelles per aquesta puntuació. Cada puntuació és, doncs, un atac independent. Es consideren cinc heurístiques estructurals clàssiques de predicció d'enllaços [19] — Common Neighbors, Jaccard, Adamic-Adar, Resource Allocation i Preferential Attachment — calculades sobre $\tilde{G}$, més una puntuació basada en la similitud cosinus entre els *embeddings* dels dos nodes.

És important distingir qui sap què. L'adversari només toca $\tilde{G}$ i produeix puntuacions. L'avaluador (l'investigador) coneix G i l'utilitza només per mesurar l'èxit de l'adversari — calcular l'AUC-ROC i l'AP de cada puntuació contra la *ground truth* —; l'adversari mai accedeix a aquestes etiquetes. Aquesta separació és el que fa el model d'amenaça consistent.

L'atac aquí considerat és deliberadament no supervisat i de caixa negra, coherent amb un model d'amenaça que nega a l'adversari tota aresta real. Per construcció, doncs, les mètriques de filtració reportades constitueixen una **cota inferior** de la informació estructural recuperable: un adversari amb accés a consultes al model entrenat, com LinkTeller [1] o els atacs de robatori d'enllaços de He et al. [15], podria assolir una reconstrucció superior. Les conclusions de protecció s'han d'entendre, per tant, referides a la classe d'adversaris considerada i no a l'òptim teòric assolible.

El conjunt d'avaluació es construeix amb totes les arestes reals com a positius i una mostra de no-arestes com a negatius, amb una ràtio d'1:100 (sense reemplaçament i sense duplicats). Aquesta ràtio reflecteix deliberadament el fort desbalanceig real del problema: en un graf dispers com Cora, les no-arestes superen les arestes en un factor de l'ordre de 10^3, de manera que un conjunt equilibrat 1:1 sobreestimaria la capacitat de l'adversari en un escenari artificialment favorable. Amb la ràtio 1:100, la prevalença del conjunt és de $\sim$0,0099, i és contra aquest valor que s'ha d'interpretar l'*Average Precision*. L'AUC-ROC, en canvi, és invariant a la prevalença, propietat que la converteix en la mètrica principal i comparable entre configuracions.

5 RESULTATS I DISCUSSIÓ

L'anàlisi detallada que segueix —utilitat per escenaris, atac d'inferència i índex de compromís— es desenvolupa sobre Cora, el conjunt principal del treball; la generalització a Citeseer i Pubmed es tracta a part (Fig. 1 i 2) i els valors complets dels tres conjunts de dades són als annexos.

5.1 Referència sense privadesa

El model GCN sense cap pertorbació assoleix una precisió de test de $0{,}905 \pm 0{,}005$ sobre Cora, $0{,}748 \pm 0{,}015$ sobre Citeseer i $0{,}823 \pm 0{,}003$ sobre Pubmed (mitjana sobre 10 execucions). Aquests resultats constitueixen la cota superior d'utilitat de cada conjunt —contra la qual es mesura la degradació introduïda per la privadesa— i coincideixen amb els que reporta la literatura per a una GCN amb particions de fracció d'entrenament elevada (60/20/20), fet que confirma que l'experimentació parteix d'una referència sòlida i competitiva.

5.2 Compromís privadesa-utilitat segons l'escenari

La Taula 2 resumeix la precisió de test dels tres escenaris al llarg de tot l'escombrat de ε, per als dos mecanismes, sobre Cora. La diferència de precisió entre tots dos mecanismes és de com a màxim 0,022 (0,0223) en tot el rang i per a tots tres escenaris (el màxim s'assoleix a $\varepsilon = 8$ en S1). El resultat central és que els tres escenaris es comporten de manera radicalment diferent, cosa que valida l'interès de distingir-los.

TAULA 2: PRECISIÓ DE TEST SOBRE CORA (MITJANA SOBRE 10 EXECUCIONS) PER ESCENARI I MECANISME, PER A UNA SELECCIÓ DE VALORS DE ε.

	S1		S2		S3	
ε	RR	Lap	RR	Lap	RR	Lap
0,1	**0,407**	0,406	**0,874**	0,868	0,875	**0,877**
1	**0,398**	0,392	**0,870**	0,855	0,876	**0,877**
3	0,401	**0,404**	0,867	**0,880**	0,876	**0,880**
5	0,481	**0,485**	**0,888**	0,887	0,886	**0,887**
7	**0,751**	0,739	**0,902**	0,899	**0,902**	0,898
9	**0,879**	0,859	**0,910**	**0,910**	**0,910**	0,904
11	**0,903**	0,893	0,905	**0,909**	**0,907**	0,906
13	**0,906**	0,894	**0,906**	0,901	0,905	**0,906**
15	**0,905**	0,902	**0,905**	**0,905**	**0,905**	**0,905**

Baseline sense DP (Cora): $0{,}905 \pm 0{,}005$

RR = Randomized Response asimètrica; Lap = Laplace amb selecció top-k. En **negreta**, el mecanisme amb millor precisió dins de cada escenari per a cada ε (tots dos en cas d'empat).

Escenari S1 (privadesa total). Quan tant l'entrenament com l'avaluació es fan sobre el graf pertorbat, la utilitat es desploma dràsticament en el rang de privadesa forta: per a ε entre 0,1 i 4, la precisió es manté al voltant de $\sim$0,39–0,42, per sota del rendiment que les característiques dels nodes sostenen per si soles. Aquesta cota la proporciona el mateix experiment: els escenaris S2 i S3 comparteixen exactament els pesos apresos en S1 —l'entrenament és idèntic en els tres— i, en avaluar-se sobre l'estructura real, assoleixen $\sim$0,87, una precisió que prové majoritàriament de la transformació de característiques que la GCN ha après. Si aquesta mateixa transformació sosté $\sim$0,87 quan opera sobre una estructura informativa, el fet que en S1 el resultat caigui a $\sim$0,40 indica que el graf pertorbat no només elimina senyal estructural útil, sinó que el substitueix per veïns espuris que perjudiquen activament la predicció, deixant el model per sota del que assoliria ignorant l'estructura i emprant només les característiques. La recuperació és gradual a partir de $\varepsilon \approx 5$ (0,481), passa per $\sim$0,75 a $\varepsilon = 7$, i no és fins a $\varepsilon \approx 11$ que pràcticament iguala el baseline (0,903).

Escenaris S2 i S3 (graf real en inferència). En contrast, quan el graf real torna a estar disponible per a l'avaluació — encara que el model s'hagi entrenat íntegrament sobre el graf pertorbat — la utilitat es manté notablement alta en tot el rang de ε: a $\varepsilon = 0{,}1$, S2 i S3 ja assoleixen $\sim$0,87, a

només ∼3–4 punts del baseline. Els pesos apresos sobre el graf pertorbat capturen majoritàriament la transformació de característiques (que no es pertorba mai), de manera que en aplicar el model sobre l'estructura real, el *message passing* recupera gairebé tota la utilitat. La diferència entre S2 i S3 és mínima: seleccionar el model sobre el graf real (S3) aporta poc respecte a fer-ho sobre el pertorbat (S2).

La conclusió pràctica és clara: el cost de la privadesa recau gairebé exclusivament sobre l'ús del graf pertorbat en inferència, no sobre l'entrenament. Si el desplegament permet aplicar el model sobre el graf real (S2/S3) —el cas més habitual—, es pot entrenar amb garanties edge-DP fortes amb una pèrdua d'utilitat molt continguda.

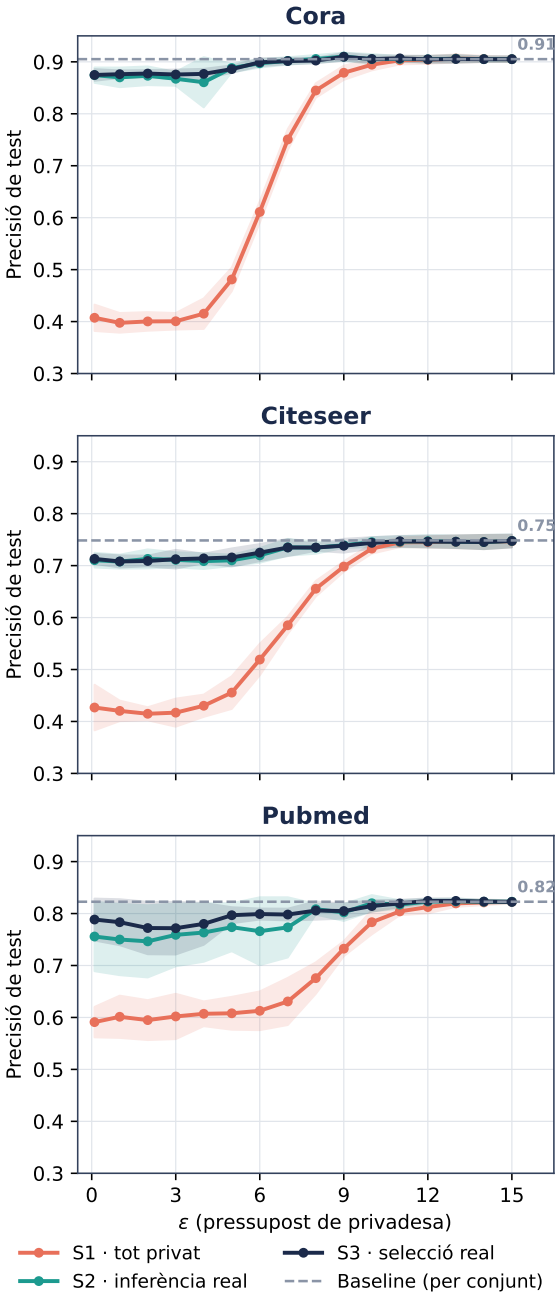


Fig. 1: Precisió de test en funció de ε per als tres escenaris (mecanisme RR) a Cora, Citeseer i Pubmed. Les bandes indiquen $\pm$ una desviació estàndard; la línia puntejada és el baseline de cada conjunt. La jerarquia S1 $\ll$ S2 $\approx$ S3 es manté en els tres grafs.

Generalització a Citeseer i Pubmed. El patró qualitatiu identificat a Cora es reprodueix als altres dos conjunts de dades (Fig. 1). En els tres, la jerarquia entre escenaris és idèntica i estable: S1 (tot privat) queda molt per sota mentre S2 i S3 (graf real en inferència) es mantenen propers al baseline al llarg de tot l'escombrat. A $\varepsilon = 1$, la dissociació S1→S2 és de $+47$ punts a Cora (0,40 → 0,87), $+29$ a Citeseer (0,42 → 0,71) i $+15$ a Pubmed (0,60 → 0,75). Que el mateix model, amb idèntics pesos, guanyi desenes de punts només per avaluar-se sobre el graf real confirma que el cost de la privadesa recau sobre l'ús del graf pertorbat en inferència, i no és un artefacte de Cora.

La magnitud de l'efecte, en canvi, sí depèn del graf. A Pubmed, S1 no s'enfonsa tant (es manté al voltant de 0,60 en privadesa forta, enfront del ∼0,40 de Cora i Citeseer): amb només 3 classes, la predicció basada únicament en característiques parteix d'una base més alta i el dany dels veïns espuris és proporcionalment menor. Aquesta dependència de la magnitud amb el nombre de classes i l'estructura del graf delimita l'abast de les conclusions: el patró és robust, però la seva intensitat és específica de cada conjunt.

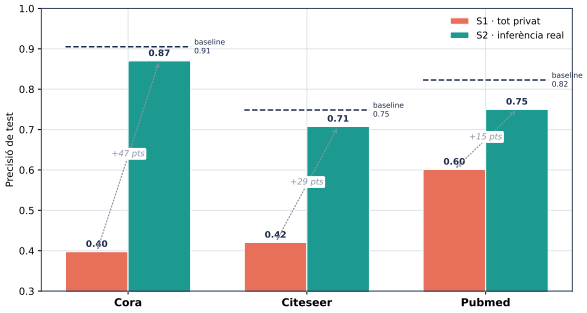


Fig. 2: Dissociació entre escenaris a $\varepsilon = 1$ a Cora, Citeseer i Pubmed. El mateix model, amb idèntics pesos, guanya entre 15 i 47 punts de precisió només per avaluar-se sobre el graf real (S2) en lloc del pertorbat (S1).

5.3 Protecció efectiva: atac d'inferència

La Fig. 3 i la Taula 3 resumeixen l'AUC-ROC de l'atac per a les puntuacions més informatives. Sense privadesa, l'adversari reconstrueix l'estructura amb gran eficàcia: la similitud d'*embeddings* assoleix AUC 0,975 i les heurístiques estructurals ∼0,76–0,79. Això confirma que una GCN entrenada sobre el graf real filtra de manera severa la seva topologia.

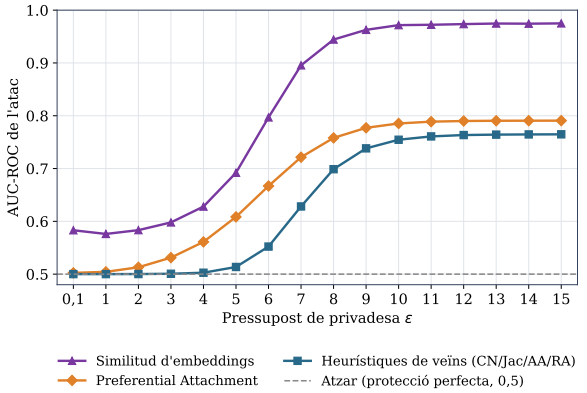


Fig. 3: AUC-ROC de l'atac d'inferència sobre Cora en funció de ε (mecanisme RR, mitjana sobre 10 execucions), per a les puntuacions més representatives (heurístiques de veïns CN/Jac/AA/RA —que coincideixen numèricament—, Preferential Attachment i similitud cosinus d'*embeddings*). La línia discontínua a 0,5 indica protecció perfecta (atzar).

TAULA 3: AUC-ROC DE L'ATAC D'INFERÈNCIA SOBRE CORA (MECANISME RR, MITJANA SOBRE 10 EXECUCIONS), ON 0,5 INDICA PROTECCIÓ PERFECTA (ATZAR).

ε	Heur. veïns	Pref. Attach.	Emb. cosine
Baseline	0,765	0,791	0,975
0,1	0,500	0,502	0,583
1	0,500	0,504	0,576
3	0,501	0,531	0,598
5	0,514	0,608	0,692
7	0,628	0,721	0,895
9	0,738	0,777	0,963
11	0,761	0,789	0,972
13	0,764	0,790	0,975
15	0,765	0,791	0,975

La columna ≪Heur. veïns≫ és la mitjana de les quatre heurístiques de veïns (CN/Jac/AA/RA), que coincideixen numèricament.

Sota privadesa, la capacitat de l'atac cau de forma ordenada i monòtona fins a nivells propers a l'atzar. Per a $\varepsilon \leq 3$, totes les heurístiques estructurals cauen a AUC $\approx$ 0,50 (atzar pur): el soroll destrueix la coincidència de veïns que aquestes mètriques exploten, fins al punt que la majoria de parelles tenen zero veïns comuns i la puntuació esdevé constant. La similitud d'*embeddings*, la puntuació més resistent, queda continguda a ∼0,57–0,60 en aquest rang. A mesura que ε creix, l'atac es recupera progressivament: a $\varepsilon = 7$ l'*embedding* ja arriba a 0,895 i a $\varepsilon = 15$ recupera el nivell del baseline (0,975). L'*Average Precision* confirma la mateixa lectura (Taula 5 de l'apèndix): amb una prevalença de 0,0099, sota privadesa forta l'AP de totes les puntuacions iguala la prevalença (cap capacitat de prioritzar arestes reals), mentre que al baseline s'enfila fins a 0,608 per a l'*embedding*. Els dos mecanismes ofereixen una protecció gairebé indistingible (diferències d'AUC inferiors a 0,022 en tot el rang).

L'*Average Precision*, que quantifica la capacitat real de prioritzar arestes en el conjunt fortament desbalancejat, no només confirma la lectura de l'AUC-ROC sinó que en matisa la interpretació. Sota privadesa forta ($\varepsilon \leq 3$) l'AP de totes les puntuacions s'enganxa a la prevalença, i no és fins a $\varepsilon \approx 6$ que comença a créixer de manera apreciable. Més enllà d'aquesta coincidència, l'AP revela un comportament que l'AUC amaga: la puntuació de *Preferential Attachment*, tot i assolir una AUC elevada al baseline (0,791), presenta una AP molt baixa (0,073), molt per sota de la d'heurístiques amb una AUC pràcticament idèntica com Adamic-Adar (AP 0,395). La causa és que *Preferential Attachment* ordena les parelles segons el producte de graus, de manera que prioritza nodes molt connectats que dominen el rànquing però no es corresponen necessàriament amb arestes reals; en un règim de desbalanceig extrem com el de Cora, és doncs un atac molt menys efectiu del que la seva AUC-ROC suggeriria. Aquesta dissociació justifica reportar ambdues mètriques i no confiar únicament en l'AUC-ROC a l'hora de valorar la protecció. La Taula 5 de l'apèndix recull els valors d'AP de totes les puntuacions al llarg de l'escombrat de ε.

Generalització a Citeseer i Pubmed. El patró de protecció identificat a Cora es reprodueix als altres dos conjunts de dades (Taules 7 i 8). En tots tres, sota privadesa forta ($\varepsilon \leq 3$) les heurístiques estructurals cauen a l'atzar pur (AUC $\approx$ 0,50) i la similitud d'*embeddings* —sempre la puntuació més resistent— queda continguda per sota de ∼0,63, per recuperar després de manera monòtona el nivell del *baseline* a mesura que ε creix. La intensitat, però, depèn del graf: a Pubmed la similitud d'*embeddings* filtra força menys (*baseline* 0,85, enfront de ∼0,98 a Cora i Citeseer), però el *Preferential Attachment* hi esdevé la puntuació més informativa (*baseline* 0,88, i AUC ja arriba a 0,83 a $\varepsilon = 7$). Aquesta diferència —coherent amb la menor dependència estructural d'un graf amb només 3 classes, que ja s'havia observat en la utilitat— confirma que la conclusió de protecció és robusta entre conjunts, si bé la puntuació concreta més vulnerable pot variar.

5.4 Síntesi del compromís

La Fig. 4 sintetitza les dues anàlisis en un sol diagrama privadesa-utilitat, i la Fig. 5 les superposa sobre el mateix eix de ε. En el règim de privadesa forta ($\varepsilon \leq 3$), l'atac queda reduït a nivells indistingibles de l'atzar (AUC $\approx$ 0,5–0,6) per a l'adversari considerat, mentre que la utilitat en inferència real (S2/S3) es manté al voltant del 87% — és a dir, existeix una finestra de pressupost on l'AUC de l'atac és indistingible de l'atzar amb una pèrdua d'utilitat moderada, sempre que el desplegament permeti utilitzar el graf real en inferència. En l'escenari de privadesa total (S1), en canvi, aquesta finestra es tanca: la utilitat només esdevé acceptable per a $\varepsilon \geq 9$, just on l'atac comença a ser molt efectiu (AUC $> 0{,}96$ per a l'*embedding*). El compromís privadesa-utilitat, doncs, no és absolut sinó que depèn críticament de l'escenari de desplegament.

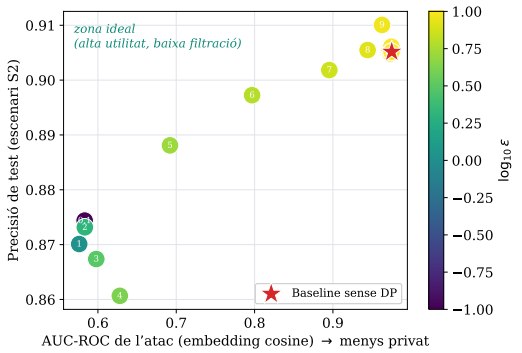


Fig. 4: Mapa privadesa-utilitat sobre Cora (mecanisme RR): utilitat (escenari S2) enfront de la filtració de l'atac (AUC-ROC de la similitud d'*embeddings*). Cada punt és un valor de ε (color); l'estrella és el *baseline* sense DP. La cantonada superior esquerra és la regió ideal.

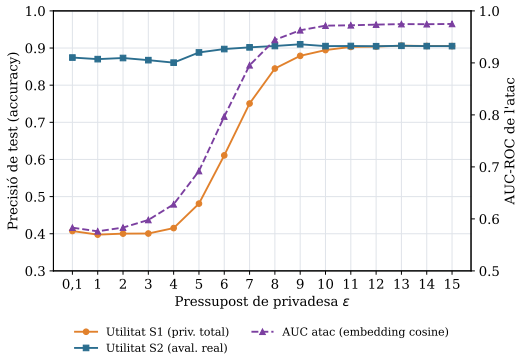


Fig. 5: Utilitat (escenaris S1 i S2, eix esquerre) i filtració de l'atac (AUC-ROC de la similitud d'*embeddings*, eix dret) sobre Cora en funció de ε (mecanisme RR). Permet visualitzar la finestra on protecció i utilitat conviuen.

Per situar de manera objectiva aquesta finestra es pot definir un índex de compromís que combini les dues mètriques en una mateixa escala $[0, 1]$. Per a cada ε es pren la utilitat en l'escenari S2, $U(\varepsilon)$, ja expressada en l'interval $[0, 1]$, i es transforma la filtració de l'atac — mesurada amb l'AUC-ROC de la similitud d'*embeddings*, la puntuació més resistent — en un terme de protecció

$$P(\varepsilon) = 2\,\big(1 - \mathrm{AUC}(\varepsilon)\big), \tag{7}$$

de manera que $P = 1$ correspon a protecció perfecta ($\mathrm{AUC} = 0{,}5$) i $P = 0$ a filtració total ($\mathrm{AUC} = 1$); el factor 2 redimensiona el rang útil de l'AUC, que es mou entre 0,5 i 1, perquè P i U comparteixin la mateixa amplitud. L'índex de compromís es defineix llavors com la mitjana dels dos termes,

$$J(\varepsilon) = \tfrac{1}{2}\big[\,U(\varepsilon) + P(\varepsilon)\,\big], \tag{8}$$

de manera que, en ser la mitjana de dues quantitats de l'interval $[0, 1]$, J queda també acotat a $[0, 1]$: $J = 1$ representa el cas ideal (utilitat màxima i protecció perfecta alhora) i $J = 0$ el pitjor cas (utilitat nul·la i filtració total). Cal subratllar que J no és una mètrica formal, sinó un resum que pondera igual els dos objectius per localitzar el millor punt operatiu; un criteri que en prioritzés un només caldria substituir la mitjana simple de (8) per una mitjana ponderada $J = w\,U + (1 - w)\,P$.

La Fig. 6 representa els tres termes al llarg de l'escombrat. L'índex J es manté pràcticament pla i màxim en tot el règim de privadesa forta i cau de manera pronunciada a partir de $\varepsilon \approx 5$, quan la protecció s'esfondra sense que la utilitat de S2 millori de forma apreciable. El màxim s'assoleix a $\varepsilon = 1$ ($J \approx 0{,}86$) i tota la regió $\varepsilon \in [0{,}1; 3]$ es manté a menys d'un 3% del màxim. Aquesta finestra és més estreta del que faria pensar la corba d'utilitat per si sola: tot i que la utilitat de S2 es manté alta fins a $\varepsilon \approx 9$, la protecció ja s'ha perdut en bona part a $\varepsilon = 5$ ($P = 0{,}62$; AUC 0,69) i és gairebé nul·la a $\varepsilon = 7$ ($P = 0{,}21$; AUC 0,90). El punt operatiu recomanat per a un desplegament que prioritzi la confidencialitat de les relacions és, doncs, $\varepsilon \in [1; 3]$, on s'obté una utilitat de S2 d'entre 0,86 i 0,88 — a 2–5 punts del baseline — amb una AUC de l'atac per sota de 0,60.

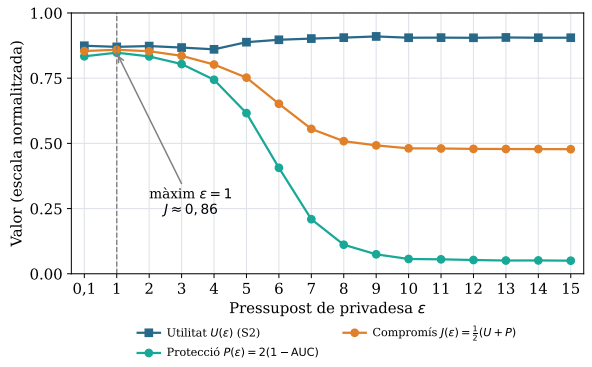


Fig. 6: Índex de compromís $J(\varepsilon) = \frac{1}{2}[U(\varepsilon) + P(\varepsilon)]$ i els seus dos termes: la utilitat U de l'escenari S2 i la protecció $P = 2(1 - \mathrm{AUC})$ derivada de la similitud d'*embeddings*. El màxim a $\varepsilon = 1$ ($J \approx 0{,}86$) marca el punt operatiu òptim. Mecanisme RR, dataset Cora.

5.5 Significança estadística

Per determinar si les diferències de precisió entre Randomized Response i Laplace són atribuïbles a l'atzar, s'ha aplicat un test de Wilcoxon de rangs amb signe aparellat sobre les execucions per llavor, seguint la pràctica recomanada per a la comparació d'algorismes d'aprenentatge automàtic [20]; els p-valors complets per a tots els valors de ε, escenaris i conjunts es recullen a la Taula 11 de l'apèndix. En la regió de privadesa forta ($\varepsilon \leq 3$) —on se situa el punt operatiu d'interès— la diferència entre mecanismes no és estadísticament significativa en cap dels tres conjunts de dades: cap comparació assoleix $p < 0{,}05$ a Citeseer ni a Pubmed, i només una de quinze a Cora, concretament a $\varepsilon = 3$ en l'escenari S2 ($p = 0{,}012$). Com que el test és de dues cues, el sentit de la diferència es determina a partir de les mitjanes (Taula 9): en aquest punt és favorable a Laplace, amb una magnitud inferior a 1,5 punts (0,880 enfront de 0,867). A pressupostos més grans apareixen algunes diferències detectables, però de magnitud pràctica reduïda (com a màxim uns 2,7 punts). Es conclou que tots dos mecanismes són equivalents en la regió rellevant. Amb 10 execucions per parella, el test té prou potència per detectar diferències sistemàtiques (el p-valor mínim assolible se situa entre 0,002 i 0,031 segons el nombre de parelles no nul·les, en tot cas inferior a 0,05), de manera que l'absència de significança a $\varepsilon \leq 3$ no és un artefacte de manca de potència, sinó una equivalència real en aquesta regió.

6 Conclusions

Aquest treball ha caracteritzat empíricament el compromís entre privadesa i utilitat d'aplicar privadesa diferencial a nivell d'aresta sobre l'entrada d'una GNN, i la protecció efectiva davant d'un atac de reconstrucció d'arestes. Els objectius plantejats s'han assolit i en deriven diverses conclusions.

Sobre la implementació dels mecanismes. S'ha formalitzat el marc d'edge-DP sobre grafs —grafs veïns a nivell d'aresta, sensibilitat de les consultes estructurals i composició del pressupost—, de manera que tot mecanisme implementat disposa d'una garantia ε-edge-DP demostrable (OE1). Sobre aquesta base s'ha implementat i validat una arquitectura GCN de referència sense privadesa, que assoleix valors de precisió equiparables als que reporta la literatura per a Cora, Citeseer i Pubmed i que actua com a cota superior d'utilitat contra la qual mesurar la degradació (OE2). Finalment, s'han desenvolupat dos mecanismes edge-DP —Randomized Response asimètrica i Laplace amb selecció top-k— amb un calibratge del pressupost correcte per a grafs dispersos (OE3). L'element clau és calibrar la pertorbació estructural contra una estimació de densitat privada $\tilde{d}$, i no contra el recompte real d'arestes, cosa que assegura la garantia ε-edge-DP per composició i postprocessament; la formulació asimètrica de RR s'ha mostrat imprescindible en grafs dispersos, on la versió simètrica clàssica generaria centenars de milers de falses arestes.

Sobre el compromís privadesa-utilitat. La conclusió més rellevant és que aquest compromís no és absolut, sinó que depèn críticament de l'escenari de desplegament (OE4). Quan el graf real està disponible en inferència (S2/S3), la utilitat es manté alta fins i tot amb privadesa forta ($\varepsilon \leq 3$), a pocs punts del baseline; quan el model opera només sobre el graf pertorbat (S1), només esdevé acceptable per a ε gran. El cost de la privadesa recau, doncs, sobre l'ús del graf pertorbat en inferència i no sobre l'entrenament: els pesos apresos capturen sobretot la transformació de característiques, que es recupera quan el *message passing* torna a operar sobre l'estructura real.

Sobre la protecció efectiva. L'atac d'inferència, amb un

adversari realista de caixa negra sense coneixement de cap aresta real, confirma que els mecanismes redueixen substancialment la capacitat de l'adversari considerat en privadesa forta (OE5): per a $\varepsilon \leq 3$, l'AUC-ROC es manté propera a 0,5 per a les puntuacions estructurals i continguda per als *embeddings*. La protecció es degrada de manera monòtona fins a recuperar la filtració del baseline ($\varepsilon = 15$), amb tots dos mecanismes a nivells gairebé indistingibles.

Síntesi. Es perfila una finestra de pressupost ($\varepsilon \leq 3$) on la capacitat de l'atac cau a nivells indistingibles de l'atzar alhora que la utilitat en inferència real es manté elevada, condicionat al fet que el desplegament permeti utilitzar el graf real. En privadesa total, aquesta finestra es tanca i utilitat i protecció entren en conflicte directe. En conjunt, els resultats es consideren satisfactoris i competitius: la referència sense privadesa iguala els valors que reporta la literatura per a una GCN sobre Cora, Citeseer i Pubmed, i el compromís privadesa-utilitat —col·lapse en privadesa forta, recuperació monòtona i convergència cap al baseline— reprodueix els patrons descrits en treballs previs. Dins de la perturbació de l'entrada, els dos mecanismes assoleixen un punt operatiu ($\varepsilon \in [1, 3]$) on l'atac queda reduït a nivells propers a l'atzar amb una pèrdua d'utilitat continguda quan el graf real és disponible en inferència.

Limitacions i treball futur. L'avaluació s'ha circumscrit a tres conjunts de citacions i a una arquitectura GCN de dues capes. Validar els resultats sobre tres grafs de mida, densitat i nombre de classes diferents aporta robustesa al patró central —la jerarquia S1 $\ll$ S2 $\approx$ S3 es manté en tots tres—, si bé la seva intensitat depèn del graf: a Pubmed, amb només tres classes, la degradació de S1 és menys pronunciada perquè la predicció basada únicament en característiques parteix d'una base més alta. Les conclusions quantitatives s'entenen, doncs, referides al règim experimental considerat, i no com a valors universals.

D'aquí es deriven diverses línies de continuació. La primera és estendre l'estudi a grafs no homofílics —on els nodes connectats no tendeixen a compartir classe i, per tant, l'estructura pot ser menys informativa per a la tasca— i a altres arquitectures de GNN més enllà de la GCN, per comprovar fins a quin punt el patró observat es generalitza. Aquesta extensió, però, requeriria reescriure els mecanismes en forma dispersa, ja que la implementació densa actual té un cost $O(|V|^2)$ que es dispara amb la mida del graf: de Cora a Pubmed, el temps de pertorbació es multiplica per $\sim$70 i la memòria arriba a 8,2 GB, mentre que l'entrenament a penes varia (Annex D). La segona línia és estabilitzar el règim de privadesa extrema: a $\varepsilon = 0{,}01$ (Cora), S1 repunta a $\sim$0,51 però amb una desviació estàndard de $\pm$0,18 (coeficient de variació $\approx$35%), perquè amb $\varepsilon_1 = 10^{-4}$ el soroll de la consulta de densitat supera la densitat real i el graf protegit varia erràticament entre execucions; un calibratge més robust per a pressupostos petits —un mínim fix per a ε_1 o una noció de sensibilitat suavitzada (*smooth sensitivity* [6])— ho mitigaria. Finalment, una tercera línia natural seria comparar de manera directa la pertorbació de l'entrada amb les altres dues grans famílies de defenses —la pertorbació de l'agregació i la del gradient—, cosa que permetria situar el compromís privadesa-utilitat obtingut dins del ventall complet d'estratègies disponibles i valorar-ne el cost relatiu en utilitat i en complexitat d'implementació.

Ús de la IA generativa

Per donar suport al desenvolupament del projecte, s'ha utilitzat la IA generativa per revisar el codi dels mecanismes i dels experiments, i per millorar el format i la redacció del document. Tot el contingut generat ha estat revisat i validat per l'autor.

Agraïments

Vull expressar el meu agraïment als meus tutors, Guillermo Navarro Arribas i Julián Salas Piñón, per la seva guia constant, les seves observacions rigoroses i la confiança que han dipositat en aquest treball al llarg de tot el procés. Les seves aportacions han estat decisives per orientar la recerca i millorar-ne el resultat.

Referències

[1] F. Wu, Y. Long, C. Zhang, and B. Li, "LinkTeller: Recovering Private Edges from Graph Neural Networks via Influence Analysis," in *Proc. IEEE Symposium on Security and Privacy (S&P)*, 2022, pp. 522–541.

[2] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *Proc. ICLR*, 2017.

[3] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3–4, pp. 211–407, 2014.

[4] M. Hay, C. Li, G. Miklau, and D. Jensen, "Accurate Estimation of the Degree Distribution of Private Networks," in *Proc. IEEE ICDM*, 2009, pp. 169–178.

[5] S. L. Warner, "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," *J. Amer. Statist. Assoc.*, vol. 60, no. 309, pp. 63–69, 1965.

[6] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth Sensitivity and Sampling in Private Data Analysis," in *Proc. ACM STOC*, 2007, pp. 75–84.

[7] S. Sajadmanesh, A. S. Shamsabadi, A. Bellet, and D. Gatica-Perez, "GAP: Differentially Private Graph Neural Networks with Aggregation Perturbation," in *Proc. 32nd USENIX Security Symposium*, 2023, pp. 3223–3240.

[8] S. Sajadmanesh and D. Gatica-Perez, "ProGAP: Progressive Graph Neural Networks with Differential Privacy Guarantees," in *Proc. 17th ACM Int. Conf. on Web Search and Data Mining (WSDM)*, 2024, pp. 596–605.

[9] E. Chien, W.-N. Chen, C. Pan, P. Li, A. Ozgur, and O. Milenkovic, "Differentially Private Decoupled Graph Convolutions for Multigranular Topology Protection," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2024.

[10] A. Daigavane, G. Madan, A. Sinha, A. Guha Thakurta, G. Aggarwal, and P. Jain, "Node-Level Differentially Private Graph Neural Networks," arXiv:2111.15521, 2021.

[11] Q. Zhang, H. K. Lee, J. Ma, J. Lou, C. Yang, and L. Xiong, "DPAR: Decoupled Graph Neural Networks with Node-Level Differential Privacy," in *Proc. ACM Web Conference (WWW)*, 2024, pp. 1170–1181.

[12] Z. Xiang, T. Wang, and D. Wang, "Preserving Node-Level Privacy in Graph Neural Networks," in *Proc. IEEE Symposium on Security and Privacy (S&P)*, 2024, pp. 4714–4732.

[13] Y. Zhang, Y. Zhao, Z. Li, X. Cheng, Y. Wang, O. Kotevska, P. S. Yu, and T. Derr, "A Survey on Privacy in Graph Neural Networks: Attacks, Preservation, and Applications," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 12, pp. 7497–7515, 2024.

[14] F. Guan, T. Zhu, W. Zhou, and K.-K. R. Choo, "Graph Neural Networks: A Survey on the Links between Privacy and Security," *Artificial Intelligence Review*, vol. 57, no. 2, art. 40, 2024.

[15] X. He, J. Jia, M. Backes, N. Z. Gong, and Y. Zhang, "Stealing Links from Graph Neural Networks," in *Proc. 30th USENIX Security Symposium*, 2021, pp. 2669–2686.

[16] X. He, R. Wen, Y. Wu, M. Backes, Y. Shen, and Y. Zhang, "Node-Level Membership Inference Attacks Against Graph Neural Networks," arXiv:2102.05429, 2021.

[17] Z. Zhang, M. Chen, M. Backes, Y. Shen, and Y. Zhang, "Inference Attacks Against Graph Neural Networks," in *Proc. 31st USENIX Security Symposium*, 2022, pp. 4543–4560.

[18] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective Classification in Network Data," *AI Magazine*, vol. 29, no. 3, pp. 93–106, 2008.

[19] D. Liben-Nowell and J. Kleinberg, "The Link-Prediction Problem for Social Networks," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031, 2007.

[20] J. Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.

A PLANIFICACIÓ TEMPORAL DEL TREBALL

Aquest apèndix recull la planificació temporal seguida durant el desenvolupament del treball, esmentada a la subsecció de planificació i desenvolupament de la Introducció. El projecte s'ha estructurat en set fases consecutives al llarg de les 18 setmanes del semestre, amb una dedicació total aproximada de 300 hores. La Fig. 7 en mostra la distribució en forma de diagrama de Gantt, i la Taula 4 en detalla la dedicació estimada per fase.

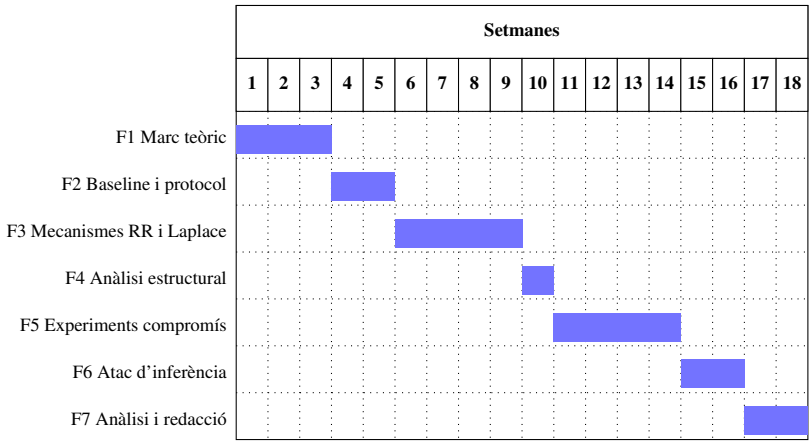


Fig. 7: Diagrama de Gantt de la planificació temporal del treball, distribuïda en set fases al llarg de les 18 setmanes del semestre.

TAULA 4: PLANIFICACIÓ PER FASES, AMB INTERVAL DE SETMANES I DEDICACIÓ ESTIMADA.

Fase	Activitat principal	Setmanes	Dedicació
F1	Marc teòric i formalització	1–3	30 h
F2	Baseline i protocol experimental	4–5	30 h
F3	Implementació dels mecanismes RR i Laplace	6–9	80 h
F4	Anàlisi estructural del graf	10	20 h
F5	Experiments del compromís privadesa–utilitat	11–14	70 h
F6	Atac d'inferència d'arestes	15–16	30 h
F7	Anàlisi de resultats i redacció	17–18	40 h
		Total	**300 h**

B REPRODUCTIBILITAT DELS EXPERIMENTS

Tots els resultats d'aquest treball són reproduïbles de manera exacta quan el codi s'executa sobre la mateixa màquina i el mateix entorn. Aquest apèndix resumeix les decisions de disseny que ho garanteixen. El codi font complet, juntament amb un document `REPRODUCIBILITAT.md` que documenta aquests mateixos mecanismes, acompanya la memòria.

Control de l'aleatorietat. L'aleatorietat es fixa a dos nivells. Al començament de cada script (`main.py` per als experiments d'utilitat i `inference_attack.py` per a l'atac) s'estableix una llavor global de valor 42 que es propaga als generadors de PyTorch i de NumPy i, en cas d'executar-se sobre GPU, al de CUDA; addicionalment s'activa el mode determinista de cuDNN i se'n desactiva l'optimització de *benchmark*, de manera que les operacions convolucionals no introdueixin variabilitat entre execucions. Per sobre d'aquesta llavor global, cada execució individual (run) reinicialitza les llavors de forma determinista a partir de la base, generant la seqüència 43, 44, …, 52 per a les deu execucions, tant d'utilitat com de l'atac). Aquesta llavor per execució es fixa just abans tant de la inicialització del model com de la pertorbació del graf, de manera que cada graf protegit és una realització reproduïble i no depèn de l'ordre en què s'executen les configuracions.

Disseny aparellat. La mateixa seqüència de llavors es reutilitza deliberadament entre valors de ε i entre els dos mecanismes (Randomized Response i Laplace). Això fa que les comparacions entre configuracions quedin aparellades — comparteixen la mateixa font d'aleatorietat— i en redueix la variància, a canvi que les mesures de configuracions diferents estiguin correlacionades, fet que es té present en interpretar les barres d'error.

Dades i partició. S'empra el conjunt Cora a través de la classe `Planetoid` de PyTorch Geometric, que en fixa el contingut. La partició 60/20/20 dels nodes etiquetats es genera amb una funció pròpia (`create_custom_split`) amb llavor controlada, de manera que les màscares d'entrenament, validació i test són idèntiques en totes les execucions i en tots dos scripts; l'atac reutilitza exactament la mateixa partició que els experiments d'utilitat.

Font única de configuració. El repartiment del pressupost de privadesa (1% per a la consulta de densitat i 99% per a la pertorbació estructural) es defineix en un únic mòdul (`dp_mechanisms/config.py`), del qual el llegeixen tots els components. Així s'evita que el valor que s'executa, el que s'imprimeix i el que es desa als fitxers de resultats puguin divergir, un problema que es va detectar i corregir durant el desenvolupament. La resta d'hiperparàmetres rellevants (llavor global 42; deu execucions tant per a utilitat com per a l'atac; màxim de 500 èpoques amb aturada anticipada de paciència 20; dimensió oculta 64; optimitzador Adam amb taxa d'aprenentatge 0,01; proporció de mostreig 1:100 entre arestes i no-arestes a l'atac) queden fixats al codi i documentats.

Persistència i verificació. Cada execució desa els resultats en un directori amb marca de temps que conté un resum llegible, les metadades de la configuració i els valors detallats per execució en format CSV. Dues execucions consecutives

sobre la mateixa màquina produeixen valors numèrics idèntics (l'única diferència és el nom del directori, que incorpora la marca de temps). Sobre màquines, versions de PyTorch/CUDA o sistemes operatius diferents poden aparèixer divergències en els darrers decimals, atribuïbles a l'ordre de les operacions en coma flotant, que no afecten cap de les conclusions del treball.

C MATERIAL COMPLEMENTARI: TAULES COMPLETES

Aquest apèndix recull els valors numèrics complets de l'escombrat de ε per als dos mecanismes, com a material de suport a les anàlisis de la Secció 5. Les Taules 5 i 6 amplien la Taula 3 amb les sis puntuacions de l'atac sobre Cora, recollint conjuntament l'AUC-ROC i l'*Average Precision* per a Randomized Response i per a Laplace, respectivament; les Taules 7 i 8 estenen aquests valors (mecanisme RR) a Citeseer i Pubmed. Les Taules 9 i 10 reprodueixen la precisió de test de tots els escenaris i mecanismes per als tres conjunts, i la Taula 11 recull els resultats complets del test de Wilcoxon aparellat RR vs Laplace. A les taules de l'atac, els acrònims corresponen a Common Neighbors (CN), Jaccard (Jac), Adamic-Adar (AA), Resource Allocation (RA), Preferential Attachment (PA) i similitud cosinus d'*embeddings* (Emb).

TAULA 5: AUC-ROC I AP DE L'ATAC (RANDOMIZED RESPONSE) SOBRE CORA, TOTES LES PUNTUACIONS (MITJANA SOBRE 10 EXECUCIONS).

	AUC-ROC (0,5 = protecció perfecta)						Average Precision (prev. 0,0099)					
ε	CN	Jac	AA	RA	PA	Emb	CN	Jac	AA	RA	PA	Emb
Base	0,765	0,763	0,766	0,766	0,791	0,975	0,280	0,167	0,395	0,392	0,073	0,608
0,1	0,500	0,500	0,500	0,500	0,502	0,583	0,010	0,010	0,010	0,010	0,010	0,013
1	0,500	0,500	0,500	0,500	0,504	0,576	0,010	0,010	0,010	0,010	0,010	0,012
2	0,500	0,500	0,500	0,500	0,513	0,583	0,010	0,010	0,010	0,010	0,010	0,013
3	0,501	0,501	0,501	0,501	0,531	0,598	0,010	0,010	0,010	0,010	0,012	0,014
4	0,503	0,503	0,503	0,503	0,561	0,628	0,010	0,010	0,010	0,010	0,017	0,017
5	0,514	0,514	0,514	0,514	0,608	0,692	0,012	0,011	0,011	0,011	0,031	0,026
6	0,552	0,552	0,552	0,552	0,667	0,797	0,031	0,022	0,030	0,029	0,047	0,068
7	0,628	0,628	0,628	0,628	0,721	0,895	0,101	0,064	0,124	0,118	0,060	0,196
8	0,699	0,698	0,699	0,699	0,758	0,944	0,186	0,115	0,252	0,244	0,067	0,361
9	0,738	0,737	0,739	0,739	0,777	0,963	0,241	0,146	0,336	0,330	0,071	0,485
10	0,754	0,753	0,755	0,755	0,785	0,972	0,264	0,158	0,372	0,367	0,072	0,569
11	0,761	0,759	0,762	0,762	0,789	0,972	0,274	0,163	0,386	0,383	0,073	0,574
12	0,763	0,762	0,764	0,764	0,790	0,974	0,278	0,166	0,392	0,389	0,073	0,589
13	0,764	0,763	0,765	0,765	0,790	0,975	0,279	0,166	0,394	0,391	0,073	0,602
14	0,764	0,763	0,766	0,766	0,791	0,974	0,279	0,166	0,394	0,392	0,073	0,599
15	0,765	0,763	0,766	0,766	0,791	0,975	0,280	0,167	0,395	0,392	0,073	0,606

TAULA 6: AUC-ROC I AP DE L'ATAC (LAPLACE AMB SELECCIÓ TOP-k) SOBRE CORA, TOTES LES PUNTUACIONS (MITJANA SOBRE 10 EXECUCIONS).

	AUC-ROC (0,5 = protecció perfecta)						Average Precision (prev. 0,0099)					
ε	CN	Jac	AA	RA	PA	Emb	CN	Jac	AA	RA	PA	Emb
Base	0,765	0,763	0,766	0,766	0,791	0,975	0,280	0,167	0,395	0,392	0,073	0,608
0,1	0,500	0,500	0,500	0,500	0,501	0,582	0,010	0,010	0,010	0,010	0,010	0,013
1	0,500	0,500	0,500	0,500	0,502	0,576	0,010	0,010	0,010	0,010	0,010	0,012
2	0,500	0,500	0,500	0,500	0,510	0,584	0,010	0,010	0,010	0,010	0,010	0,013
3	0,501	0,501	0,501	0,501	0,526	0,599	0,010	0,010	0,010	0,010	0,011	0,014
4	0,503	0,503	0,503	0,503	0,560	0,629	0,010	0,010	0,010	0,010	0,015	0,017
5	0,514	0,514	0,514	0,514	0,610	0,696	0,012	0,011	0,012	0,012	0,031	0,028
6	0,553	0,553	0,553	0,553	0,668	0,799	0,032	0,022	0,031	0,029	0,047	0,069
7	0,627	0,626	0,627	0,627	0,720	0,895	0,100	0,063	0,121	0,114	0,059	0,200
8	0,683	0,682	0,683	0,683	0,750	0,935	0,167	0,104	0,221	0,212	0,065	0,308
9	0,717	0,716	0,717	0,717	0,767	0,954	0,211	0,129	0,290	0,282	0,068	0,415
10	0,736	0,734	0,736	0,736	0,776	0,962	0,237	0,143	0,330	0,324	0,070	0,481
11	0,747	0,746	0,748	0,748	0,782	0,966	0,254	0,152	0,356	0,351	0,071	0,514
12	0,754	0,753	0,755	0,755	0,785	0,971	0,264	0,159	0,372	0,368	0,072	0,564
13	0,759	0,757	0,760	0,760	0,787	0,971	0,271	0,162	0,382	0,378	0,073	0,565
14	0,761	0,760	0,762	0,762	0,789	0,974	0,275	0,164	0,387	0,384	0,073	0,591
15	0,762	0,761	0,763	0,763	0,790	0,973	0,276	0,165	0,390	0,387	0,073	0,582

TAULA 7: AUC-ROC DE L'ATAC (RANDOMIZED RESPONSE) SOBRE CITESEER I PUBMED, TOTES LES PUNTUACIONS (MITJANA SOBRE 10 EXECUCIONS).

	Citeseer						Pubmed					
ε	CN	Jac	AA	RA	PA	Emb	CN	Jac	AA	RA	PA	Emb
Base	0,703	0,702	0,703	0,703	0,778	0,990	0,714	0,712	0,715	0,715	0,878	0,846
0,1	0,500	0,500	0,500	0,500	0,497	0,632	0,500	0,500	0,500	0,500	0,500	0,538
1	0,500	0,500	0,500	0,500	0,498	0,618	0,500	0,500	0,500	0,500	0,508	0,539
2	0,500	0,500	0,500	0,500	0,502	0,623	0,501	0,501	0,501	0,501	0,518	0,542
3	0,500	0,500	0,500	0,500	0,511	0,630	0,501	0,501	0,501	0,501	0,548	0,550
4	0,501	0,501	0,501	0,501	0,530	0,655	0,504	0,504	0,504	0,504	0,604	0,566
5	0,504	0,504	0,504	0,504	0,571	0,702	0,517	0,517	0,517	0,517	0,688	0,598
6	0,521	0,521	0,521	0,521	0,630	0,787	0,557	0,557	0,557	0,557	0,774	0,664
7	0,568	0,568	0,568	0,568	0,691	0,884	0,623	0,622	0,623	0,623	0,827	0,734
8	0,628	0,627	0,628	0,628	0,732	0,948	0,673	0,671	0,674	0,673	0,851	0,799
9	0,668	0,668	0,668	0,668	0,758	0,975	0,698	0,696	0,699	0,699	0,865	0,822
10	0,689	0,689	0,689	0,689	0,770	0,984	0,708	0,705	0,709	0,709	0,873	0,838
11	0,697	0,697	0,697	0,697	0,775	0,988	0,712	0,709	0,713	0,713	0,876	0,839
12	0,700	0,700	0,701	0,701	0,777	0,989	0,714	0,711	0,714	0,714	0,877	0,844
13	0,702	0,702	0,702	0,702	0,778	0,989	0,714	0,711	0,715	0,715	0,877	0,846
14	0,702	0,702	0,703	0,703	0,778	0,990	0,714	0,711	0,715	0,715	0,878	0,846
15	0,703	0,702	0,703	0,703	0,778	0,990	0,714	0,712	0,715	0,715	0,878	0,846

TAULA 8: AVERAGE PRECISION (AP) DE L'ATAC (RANDOMIZED RESPONSE) SOBRE CITESEER I PUBMED, TOTES LES PUNTUACIONS (MITJANA SOBRE 10 EXECUCIONS).

	Citeseer						Pubmed					
ε	CN	Jac	AA	RA	PA	Emb	CN	Jac	AA	RA	PA	Emb
Base	0,276	0,223	0,332	0,330	0,105	0,776	0,186	0,069	0,244	0,236	0,091	0,109
0,1	0,010	0,010	0,010	0,010	0,010	0,018	0,010	0,010	0,010	0,010	0,010	0,011
1	0,010	0,010	0,010	0,010	0,010	0,015	0,010	0,010	0,010	0,010	0,010	0,011
2	0,010	0,010	0,010	0,010	0,010	0,015	0,010	0,010	0,010	0,010	0,011	0,011
3	0,010	0,010	0,010	0,010	0,010	0,016	0,010	0,010	0,010	0,010	0,012	0,012
4	0,010	0,010	0,010	0,010	0,011	0,021	0,010	0,010	0,010	0,010	0,020	0,013
5	0,010	0,010	0,010	0,010	0,017	0,040	0,012	0,011	0,011	0,011	0,041	0,015
6	0,018	0,015	0,016	0,016	0,035	0,108	0,034	0,018	0,030	0,025	0,066	0,023
7	0,068	0,047	0,064	0,061	0,062	0,274	0,093	0,040	0,101	0,083	0,081	0,046
8	0,155	0,115	0,169	0,161	0,084	0,473	0,142	0,057	0,174	0,155	0,087	0,075
9	0,220	0,173	0,255	0,248	0,096	0,623	0,169	0,065	0,215	0,201	0,089	0,089
10	0,254	0,203	0,302	0,298	0,102	0,710	0,179	0,067	0,233	0,222	0,091	0,112
11	0,267	0,215	0,320	0,317	0,104	0,747	0,183	0,068	0,240	0,231	0,091	0,103
12	0,273	0,220	0,328	0,325	0,105	0,757	0,186	0,069	0,243	0,234	0,091	0,108
13	0,275	0,222	0,331	0,329	0,105	0,764	0,186	0,069	0,244	0,235	0,091	0,114
14	0,275	0,222	0,332	0,329	0,105	0,775	0,186	0,069	0,244	0,236	0,091	0,111
15	0,276	0,223	0,332	0,330	0,105	0,768	0,186	0,069	0,244	0,236	0,091	0,111

TAULA 9: PRECISIÓ DE TEST SOBRE CORA (MITJANA SOBRE 10 EXECUCIONS) PER ESCENARI I MECANISME, PER A TOT L'ESCOMBRAT DE ε.

ε	RR-S1	RR-S2	RR-S3	Lap-S1	Lap-S2	Lap-S3
0,1	0,407	0,874	0,875	0,406	0,868	0,877
1	0,398	0,870	0,876	0,392	0,855	0,877
2	0,400	0,873	0,877	0,397	0,866	0,877
3	0,401	0,867	0,876	0,404	0,880	0,880
4	0,415	0,861	0,877	0,419	0,877	0,879
5	0,481	0,888	0,886	0,485	0,887	0,887
6	0,611	0,897	0,899	0,607	0,894	0,896
7	0,751	0,902	0,902	0,739	0,899	0,898
8	0,845	0,905	0,902	0,823	0,902	0,907
9	0,879	0,910	0,910	0,859	0,910	0,904
10	0,894	0,905	0,905	0,878	0,909	0,903
11	0,903	0,905	0,907	0,893	0,909	0,906
12	0,904	0,905	0,905	0,894	0,909	0,907
13	0,906	0,906	0,905	0,894	0,901	0,906
14	0,905	0,905	0,905	0,900	0,904	0,905
15	0,905	0,905	0,905	0,902	0,905	0,905

RR = Randomized Response asimètrica; Lap = Laplace amb selecció top-k.

TAULA 10: PRECISIÓ DE TEST SOBRE CITESEER I PUBMED (MITJANA SOBRE 10 EXECUCIONS) PER ESCENARI I MECANISME, PER A TOT L'ESCOMBRAT DE ε.

	Citeseer						Pubmed					
ε	RR-S1	RR-S2	RR-S3	Lap-S1	Lap-S2	Lap-S3	RR-S1	RR-S2	RR-S3	Lap-S1	Lap-S2	Lap-S3
0,1	0,427	0,710	0,713	0,424	0,710	0,715	0,591	0,756	0,788	0,596	0,771	0,773
1	0,420	0,708	0,708	0,412	0,708	0,714	0,601	0,750	0,783	0,592	0,748	0,771
2	0,415	0,713	0,709	0,413	0,710	0,713	0,595	0,746	0,772	0,596	0,762	0,784
3	0,417	0,711	0,712	0,421	0,709	0,711	0,602	0,759	0,772	0,600	0,760	0,771
4	0,430	0,709	0,714	0,426	0,716	0,721	0,607	0,763	0,780	0,594	0,747	0,782
5	0,456	0,710	0,716	0,456	0,714	0,720	0,608	0,774	0,797	0,597	0,770	0,796
6	0,519	0,720	0,725	0,505	0,719	0,719	0,613	0,766	0,799	0,606	0,757	0,800
7	0,585	0,735	0,735	0,581	0,731	0,732	0,631	0,773	0,798	0,640	0,793	0,804
8	0,656	0,735	0,735	0,648	0,738	0,740	0,676	0,808	0,806	0,689	0,810	0,799
9	0,698	0,739	0,738	0,688	0,737	0,733	0,732	0,802	0,805	0,729	0,816	0,804
10	0,733	0,745	0,744	0,713	0,747	0,742	0,783	0,820	0,814	0,756	0,815	0,820
11	0,745	0,746	0,747	0,724	0,744	0,746	0,804	0,817	0,820	0,781	0,819	0,819
12	0,745	0,747	0,746	0,738	0,747	0,749	0,812	0,823	0,824	0,793	0,817	0,821
13	0,746	0,746	0,746	0,744	0,748	0,748	0,820	0,823	0,824	0,803	0,817	0,818
14	0,745	0,745	0,745	0,746	0,748	0,749	0,821	0,823	0,823	0,811	0,818	0,821
15	0,747	0,748	0,748	0,744	0,745	0,746	0,822	0,823	0,822	0,816	0,820	0,821

RR = Randomized Response asimètrica; Lap = Laplace amb selecció top-k.

TAULA 11: TEST DE WILCOXON DE RANGS AMB SIGNE APARELLAT ENTRE RR I LAPLACE SOBRE LA PRECISIÓ DE TEST (MITJANA SOBRE 10 EXECUCIONS), AMB EL p-VALOR DE DUES CUES PER CONJUNT.

ε	Esc.	Cora	Cite.	Pub.	ε	Esc.	Cora	Cite.	Pub.
0,01	S1	0,562	0,602	0,910	7	S3	0,227	0,846	0,092
0,01	S2	0,234	0,859	0,250	8	S1	**0,004**	0,238	0,154
0,01	S3	0,484	0,312	0,992	8	S2	0,572	0,314	0,762
0,1	S1	1,000	0,516	0,570	8	S3	0,094	0,273	0,225
0,1	S2	0,156	0,758	0,719	9	S1	**0,004**	0,082	0,844
0,1	S3	0,281	0,781	0,062	9	S2	0,859	0,633	0,094
1	S1	0,156	0,086	0,361	9	S3	**0,023**	0,707	0,992
1	S2	1,000	0,775	1,000	10	S1	**0,008**	**0,002**	**0,002**
1	S3	1,000	0,219	0,242	10	S2	0,252	0,480	0,715
2	S1	0,141	0,941	0,914	10	S3	0,314	0,637	0,195
2	S2	0,406	0,502	0,211	11	S1	**0,008**	**0,008**	**0,002**
2	S3	1,000	0,156	0,750	11	S2	0,062	0,562	0,484
3	S1	0,594	0,605	0,434	11	S3	0,656	0,828	0,777
3	S2	**0,012**	0,406	0,172	12	S1	**0,008**	**0,020**	**0,002**
3	S3	0,299	0,906	0,438	12	S2	0,264	0,797	0,273
4	S1	0,547	0,539	0,219	12	S3	0,766	0,447	0,254
4	S2	0,535	0,074	0,562	13	S1	**0,020**	0,438	**0,006**
4	S3	0,562	0,059	0,578	13	S2	0,293	0,555	0,094
5	S1	0,473	0,934	0,211	13	S3	0,812	0,555	0,068
5	S2	0,887	0,783	0,891	14	S1	**0,023**	0,688	**0,008**
5	S3	0,922	0,506	1,000	14	S2	0,750	0,168	0,104
6	S1	0,711	0,152	0,334	14	S3	1,000	0,062	0,438
6	S2	0,172	0,770	0,238	15	S1	0,090	0,328	0,070
6	S3	0,211	0,252	0,984	15	S2	1,000	0,344	0,125
7	S1	0,191	0,752	0,625	15	S3	1,000	0,812	0,438
7	S2	0,281	0,707	0,416					

En **negreta**, $p < 0{,}05$. La taula es disposa en dos blocs per compacitat.

D COST COMPUTACIONAL

La Taula 12 recull el cost computacional mesurat per als tres conjunts (mitjana sobre 10 execucions, en CPU). El temps de pertorbació i la memòria creixen de manera molt pronunciada amb la mida del graf, fet coherent amb el cost $O(|V|^2)$ de materialitzar l'adjacència densa: de Cora a Pubmed ($\sim$7$\times$ més nodes), la pertorbació passa de 0,23 s a 16,3 s ($\sim$70$\times$) i la RAM de 0,65 GB a 8,2 GB, mentre que l'entrenament escala de forma molt més moderada. El coll d'ampolla del mètode és, doncs, la pertorbació densa i no l'entrenament de la GNN.

TAULA 12: COST COMPUTACIONAL MESURAT SOBRE ELS TRES CONJUNTS (MITJANA SOBRE 10 EXECUCIONS, EN CPU; TEMPS EN SEGONS PER EXECUCIÓ I RAM EN PIC DE MEMÒRIA RESIDENT).

Mètrica	Cora	Citeseer	Pubmed
Entrenament baseline (s)	1,23	1,57	7,07
Entrenament sota DP (s)	2,22	2,45	11,99
Pertorbació DP (s)	0,23	0,33	16,27
Overhead privadesa (s)	1,22	1,21	21,19
Pic de RAM (GB)	0,65	0,93	8,21