Universitat de Girona
Escola Politècnica Superior

Treball Final de Màster

Estudi: Màster en Ciència de Dades

Títol: Eïnes d'aprenentatge no supervisat
per millorar la robustesa de xarxes

Document: Memòria

Alumne: Sergi Bergillos Pedraza

Tutor: Eusebi Calle Ortega
Departament: Arquitectura i Tecnologia de Computadors
Àrea: Arquitectura i Tecnologia de Computadors

Convocatòria (mes/any): Setembre 2022

Universitat de Girona
Escola Politècnica Superior

TREBALL FINAL DE MÀSTER

Eïnes d'aprenentatge no supervisat per millorar la robustesa de xarxes

Autor:
Sergi BERGILLOS PEDRAZA

Setembre 2022

Màster en Ciència de Dades

Tutor:
Eusebi CALLE ORTEGA

Resum

La robustesa es defineix com la característica d'una xarxa que reflecteix la capacitat d'aquesta a l'hora de continuar funcionant correctament davant de fallades o atacs [Marzo 2019] i és un aspecte fonamental que s'ha de tenir en compte a l'hora de dissenyar noves xarxes perquè la fallada d'un o més elements pot suposar pèrdues milionàries i deixar sense servei una part significativa de la xarxa.

En aquesta tesi no es proposa una nova forma de calcular la robustesa, sinó que es busca clusteritzar xarxes de telecomunicacions per la seva robustesa i veure com es diferencien certes mètriques segons la seva etiqueta realitzant aquest estudi amb més de dos-cents xarxes del *Topology Zoo*.

Els resultats, encara que amb un èxit limitat per la clusterització amb UMAP i HDBSCAN, han estat molt interessants perquè ha permès aprofundir en el coneixement de què representa cada mètrica i quina informació dona a l'hora de definir la robustesa de les xarxes.

Agraïments

Per començar vull agrair molt especialment al meu tutor, Eusebi Calle Ortega, pel seu interès i la seva disponibilitat els mesos d'estiu. També al grup BCDS per la formació continuada en el tema, per la seva comprensió i flexibilitat d'horaris que m'ha permès dedicar-me aquestes últimes setmanes per complet a la tesi. Finalment, a la meva família pel seu suport i la seva paciència.

Índex

1 Introducció **1**

2 Preliminars **3**
2.1 Domini . . . 3
2.2 Notació i terminologia . . . 4
2.3 Mètodes utilitzats . . . 4
2.3.1 UMAP . . . 5
2.3.2 HDBSCAN . . . 5
2.3.3 PCA . . . 6

3 Estat de l'Art **7**

4 Simulador i Dades **11**
4.1 Network Research Simulator . . . 11
4.1.1 Experiment de robustesa . . . 12
4.2 Dades . . . 13

5 Planificació i Metodologia **19**
5.1 Metodologia . . . 19
5.2 Planificació . . . 21
5.2.1 Planificació del projecte . . . 21
5.2.2 Estudi de l'estat de l'art . . . 21
5.2.3 Estudi previ a la generació de les dades . . . 21
5.2.4 Generació de les dades . . . 21
5.2.5 Preparació i modelització . . . 21
5.2.6 Redacció de la memòria . . . 22
5.2.7 Implementació de noves millores . . . 22

6 Contribució Metodològica **25**

7 Resultats **27**
7.1 Estudi Previ . . . 27
7.1.1 Precisió del càlcul de la robustesa . . . 27
7.1.2 Profunditat de l'atac . . . 29
7.2 Anàlisi exploratòria de dades . . . 32
7.2.1 Exploració inicial i etiquetatge de les dades . . . 32
7.2.2 Mètriques rellevants . . . 36
7.2.3 Mètriques recomanades . . . 39

7.3 Selecció de les mètriques 43
7.4 Enginyeria de característiques 46
7.5 Modelització 50
7.6 Generació de noves xarxes 55
7.7 Nova Mètrica 58

8 Conclusions i Treball Futur **63**
8.1 Conclusions 63
8.2 Treball Futur 64

Bibliografia **65**

Índex de figures

2.1 Visualització de la xarxa Arpanet el desembre del 1969, la precursora d'Internet. 3

4.1 Iteracions: a) seqüencial i b) condicional. 11
4.2 Arquitectura del Network Research Simulator 12

5.1 Diagrama de la metodologia CRISP-DM. 19
5.2 Diagrama de les tasques (negreta) i sortides (cursiva) de la metodologia CRISP-DM. 20
5.3 Diagrama de Gantt amb la temporització de la tesi. 23

7.1 Diagrama de caixes de la precisió en les diferents *M* 27
7.2 Relació entre la precisió i el temps de cómput dels experiments. . 29
7.3 Evolució de la robustesa en deu xarxes de telecomunicacions. . . . 30
7.4 Evolució de la robustesa en dos xarxes de clavegueram. 31
7.5 Evolució de la robustesa en dos xarxes sintètiques Erdös-Rènyi. . . 31
7.6 Diagrama de caixes de la robustesa mitjana respecte els diferents atacs. 33
7.7 Diagrama de caixes de la robustesa mitjana respecte els diferents atacs. 34
7.8 Visualització de la xarxa Ulaknet amb el NRS. 34
7.9 Visualització de la xarxa Pern amb el NRS. 35
7.10 Comparació de la robustesa per *Betweennes Centrality* per les xarxes Ulaknet i Pern. 36
7.12 Visualització de la xarxa Belnet2005 amb el NRS. 39
7.13 Visualització de la xarxa Cernet amb el NRS. 40
7.14 Diagrama de caixes del *Number of Nodes* sense *outliers*. 41
7.16 Matriu de correlacions amb les mètriques del NRS després del primer filtre. 45
7.19 Visualització de la xarxa Syringa amb el NRS. 48
7.20 Visualització de la xarxa VtlWavenet2011 amb el NRS. 48
7.22 Reducció de la dimensionalitat amb UMAP no supervisat. 50
7.23 Reducció de la dimensionalitat amb les dades de *train* i UMAP semisupervisat. 51
7.24 Reducció de la dimensionalitat amb les dades de *test* i UMAP semisupervisat. 52
7.25 Clusterització de les dades reduïdes de *train* sense tractar. 52

7.26 Clusterització de les dades reduïdes de *train* tractades. 53
7.27 Clusterització de les dades reduïdes de *test* sense tractar. 54
7.28 Clusterització de les dades reduïdes de *test* tractades. 54
7.29 Visualització del graf generat amb l'histograma [0,2,1,2,1]. 55
7.30 Robustesa respecte la severitat de l'atac per deu variacions de la xarxa Surfnet. 56
7.31 Robustesa respecte la severitat de l'atac per Surfnet i una variació sense modificacions. 57
7.33 Gràfic de dispersió de l'*Adjusted Heterogenity* respecte la robustesa per P=10. 60
7.34 Visualització de la xarxa Uninett2011 amb el NRS. 61
7.35 Visualització de la xarxa Uunet amb el NRS. 61
7.36 Visualització d'una part de la xarxa Uninett2011 amb el NRS. . . . 62

Índex de taules

4.1 Mètriques del NRS 14
4.2 Notació de les fòrmules de les mètriques 16

7.1 Taula amb els temps d'execució per cada M. 28
7.2 Diferència entre la robustesa mitjana per *Betweenness Centrality* i l'atac més devastador. 32
7.3 Diferència entre la robustesa a $P = 10$ per *Betweenness Centrality* i l'atac més devastador. 33
7.4 Mètriques de les dos xarxes amb pitjor robustesa mitjana per *Betweenness Centrality* 33
7.5 Les cinc xarxes menys robustes amb millor heterogeneïtat 37
7.6 Les cinc xarxes menys robustes amb millor *Clustering Coefficient* . 37
7.7 Les cinc xarxes menys robustes amb pitjor *Scaled Effective Resistance*. 46
7.8 Les cinc xarxes amb millor *Scaled Number of Spanning Trees*. . . . 49
7.9 Precisió de l'algorisme HDBSCAN per les dades reduides de *test*. . 53
7.10 Les cinc pitjors xarxes per Adjusted Heterogenity i etiqueta 0 . . . 59
7.11 Les cinc pitjors xarxes per Adjusted Heterogenity i etiqueta 3 . . . 59

Capítol 1

Introducció

El tema general d'aquesta tesi és la robustesa de les xarxes de telecomunicacions. La robustesa es defineix com la característica d'una xarxa que reflecteix la capacitat d'aquesta a l'hora de continuar funcionant correctament davant de fallades o atacs [Marzo 2019].

La robustesa és un aspecte fonamental que s'ha de considerar durant el disseny de qualsevol classe de xarxa perquè donat l'alt cost de construcció i de manteniment, una fallada crítica, sigui per un mal funcionament o per un atac, pot provocar pèrdues milionàries a l'empresa propietària en deixar una part significativa de la xarxa sense servei.

Tot i la seva importància, el càlcul de la robustesa d'una xarxa no és trivial perquè s'han de simular centenars de variacions de la xarxa original, a la que s'han eliminat un o més elements, i milers de mètriques complexes de teoria de grafs, ja que s'estudia la xarxa des d'un punt de vista teòric, és a dir, com una simple combinació de nodes i enllaços.

La literatura relacionada i l'estat de l'art s'ha concentrat en trobar aquest valor de la robustesa, ja sigui a partir de mètriques de connectivitat, com el *Largest Connected Componenent* o el *Average Two Terminal Reliability*, o amb aproximacions més complexes combinant diferents mètriques.

Aquesta segona estratègia, originada a partir de l'article *Robustness envelopes of networks* [Trajanovski 2013], és la que ha seguit i ampliat el grup de recerca *Broadband Communications and Distributed Systems* (BCDS) amb els articles *Robustness surfaces of complex networks* [Manzano 2014], *On selecting the relevant metrics of network robustness* [Marzo 2018] i *A study of the robustness of optical networks under massive failures* [Marzo 2019] i que ha culminat amb el disseny i la implementació del simulador Network Research Simulator (NRS) [Marzo 2022].

Al capítol 3 es pot trobar una descripció detallada de com es calcula la robustesa d'una xarxa i a la subsecció 4.1, una descripció del simulador, NRS, utilitzat per calcular-la.

La contribució d'aquesta tesi a l'estat de l'art no és una nova forma de calcular la robustesa, sinó un estudi sobre les xarxes de telecomunicacions publicades al *dataset* Topology Zoo, que tenen almenys quinze nodes i que estan completament connectades inicialment, per descobrir quines són les mètriques de la xarxa original més rellevants a l'hora d'explicar la seva robustesa.

El segon objectiu és classificar i clusteritzar aquestes xarxes per quin és l'atac més devastador, entre el *Betweenness Centrality*, el *Closeness Centrality*, l'*Eigenvector Centrality* i el *Nodal Degree*.

Tanmateix, aquesta estratègia no ha estat viable perquè, al tractar-se de xarxes amb una tipologia molt similar, l'atac més fort ha estat sempre el mateix, el *Betweenness Centrality* en el que es prioritza els elements a eliminar per aquells pel que passen un nombre més gran de camins mínims entre parelles de nodes. Per aquest motiu, s'ha acabat utilitzant com etiqueta quan de robusta és la xarxa per aquest atac discriminant per a quin quartil pertany.

Els resultats han estat limitats per a la clusterització de les xarxes utilitzant UMAP i HDBSCAN, essent excel·lents per la meitat de les mostres de *test*, però caient en picat quan la confiança de l'HDBSCAN és baixa.

Tot i això, quan la predicció de l'HDBSCAN és errònia, rarament s'equivoca per un factor més gran de dos; és a dir, per poques mostres del *dataset* de *test* l'HDBSCAN diu que té una bona robustesa quan és molt dolenta o que té molt bona robustesa quan és regular.

Finalment, l'exploració de les xarxes del Topology Zoo ha resultat molt interessant, ja que s'ha vist com les diferents mètriques expliquen les diferents debilitats i fortaleses de les xarxes i, en molts casos, s'ha pogut veure reflectida aquesta mètrica amb la visualització de la xarxa utilitzant el NRS.

CAPÍTOL 2

Preliminars

2.1 Domini

L'objecte d'estudi d'aquesta tesi són les xarxes de telecomunicacions, encara que pot abstreure's a qualsevol classe de xarxa que pugui ser representada com un graf.

Un graf és un conjunt de nodes, o vèrtexs, units per enllaços, o arestes. A la figura 2.1 es pot veure la representació d'un graf amb quatre nodes i quatre arestes.

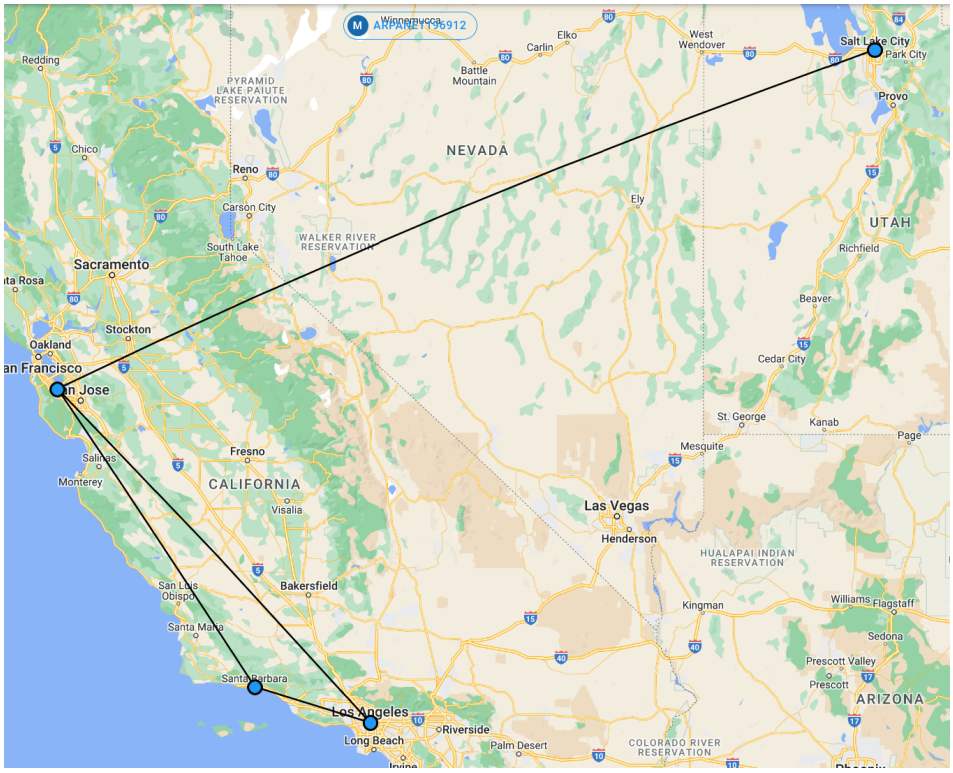


Figura 2.1: Visualització de la xarxa Arpanet el desembre del 1969, la precursora d'Internet.

Les xarxes de telecomunicacions es caracteritzen per ser relativament petites, solen tenir entre trenta i seixanta nodes i un grau nodal mitjà entre dos i

tres. Això és perquè la construcció i el manteniment d'aquestes xarxes té un cost molt elevat per les empreses, tant en diners com en temps. A més, són grafs no dirigits, és a dir, els enllaços no tenen un sentit definit.

Altres tipologies de xarxes físiques poden ser: de transport, per exemple les línies de metro, o de clavegueram; xarxes socials: d'amistats de Facebook o de piulades de Twitter; o simplement xarxes sintètiques: *Scale Free*, *Exponential*...

2.2 Notació i terminologia

Alguns conceptes importants que s'han de tenir clar des d'un bon principi són:

- Atac (a una xarxa): fet d'eliminar un o més nodes i/o arestes d'una xarxa, considerant que una xarxa no és res més que un graf. Un atac genera sempre un subconjunt de la xarxa original, aquesta xarxa resultant tindrà igual o menys connectivitat.

- Component: és un subconjunt dels nodes originals que estan connectats i que no formen part d'un component més gran.

- Mètrica: característica d'una xarxa que es pot calcular a partir del seu graf. Poden ser tan simples com el nombre de nodes o d'enllaços del graf o tan complicades com l'assortivitat. Les mètriques que es poden calcular amb el NRS estan explicades a la secció 4.2, però hi ha moltes més com la constant de Kemeny.

- NRS (Network Research Simulator): simulador de xarxes propietat del grup de recerca BCDS, descrit a la secció 4.1.

- Robustesa: Característica d'una xarxa que reflecteix la capacitat d'aquesta a l'hora de continuar funcionant correctament davant de fallades o atacs. La robustesa es calcula usant conceptes de la teoria de grafs, més concretament, de connectivitat de grafs [Marzo 2019]. La robustesa està explicada en més detall al capítol 3.

2.3 Mètodes utilitzats

En aquesta tesi s'han utilitzat principalment dos mètodes: l'UMAP i l'HDBSCAN, però també s'ha descrit el PCA perquè és una part fonamental del càlcul de la robustesa al NRS.

2.3.1 UMAP

L'UMAP (*Uniform Manifold Approximation and Projection*) [McInnes 2018b] és una tècnica de reducció de la dimensionalitat utilitzada principalment per la visualització de les dades i per l'aprenentatge automàtic. Una diferència amb altres algorismes d'aquest tipus, com el PCA (*Principal Component Analysis*), és que permet fer una reducció semisupervisada de les dades. És a dir, es poden fer servir etiquetes per guiar a l'algorisme a l'hora de fer la reducció.

Els paràmetres d'entrada bàsics d'UMAP per Python [McInnes 2018a] són:

- **n_neighbors**: aquest paràmetre controla com UMAP equilibra l'estructura local i global de les dades. Un nombre de veïns baix forçarà l'UMAP a concentrar-se en estructures locals, mentre un nombre de veïns més alt obligarà a l'UMAP a mantenir l'estructura general de les dades, ja que tindrà en compte un veïnat més gran a l'hora de fer l'estimació de cada punt.
- **min_dist**: és la distància mínima en què poden estar els punts en la representació reduïda.
- **n_components**: el nombre de dimensions que ha de tenir l'espai reduït.
- **metric**: la mètrica que s'utilitza per calcular la distància entre els punts.

2.3.2 HDBSCAN

Per la seva banda, l'HDBSCAN (*Hierarchical Density Based Clustering of Applications with Noise)* [McInnes 2017] és un algorisme d'aprenentatge automàtic no supervisat que identifica clústers en les dades d'acord amb la idea en què un clúster és una regió d'alta densitat de punts separats per altres regions de menor densitat. Una millora davant de DBSCAN [Ester 1996] és que permet trobar clústers de diferents densitats.

A més, a diferència de la majoria d'algorismes de clusterització habituals, l'HDBSCAN disposa de dos avantatges més: permet detectar mostres com a soroll, és a dir, que no pertanyen a cap clúster; i, juntament amb la predicció del clúster al qual pertany una mostra, també retorna la confiança de la seva predicció.

Els paràmetres d'entrada més importants de la implementació d'HDBSCAN feta per Python a *The hdbscan Clustering Library* [McInnes 2016] i basada en *Density-Based Clustering Based on Hierarchical Density Estimates* [Campello 2013] són:

- **min_cluster_size**: La mida mínima dels clústers.
- **min_samples**: Les mostres mínimes que ha de tenir cada element dintre de la seva *core distance*[1]. Quan més gran és el valor, més mostres de les dades seran considerades soroll.
- **cluster_selection_epsilon**: Sel·lecció de clústers ε. L'ε diu quina és la distància màxima entre dos punts per considerar-los del mateix clúster.
- **alpha**: L'α és una altre aproximació per regular la mida dels clústers. El seu valor per defecte és 1, però incrementar-lo fa que l'algorisme sigui més conservador. No és un paràmetre recomanable de modificar i menys per principiants.

Les tres mesures de similitud entre clústers que s'han utilitzat per avaluar el correcte funcionament de l'HDBSCAN són: l'índex Rand, l'índex Rand ajustat i la informació mútua ajustada.

2.3.3 PCA

El PCA (*Principal Component Analysis*) és, igual que l'UMAP, un algorisme estadístic de reducció de la dimensionalitat, però ho fa d'una forma diferent.

L'anàlisi de components principals (PCA) és una tècnica per augmentar la interpretabilitat de grans *datasets*, però alhora minimitzant la pèrdua d'informació. Ho fa creant noves variables no correlacionades que successivament maximitzen la variància. Trobar aquestes variables noves, els components principals, es redueix a resoldre un problema de valors propis/vectors propis, i les noves variables es defineixen pel conjunt de dades disponible, no a priori, per tant, convertint PCA en una tècnica d'anàlisi de dades adaptativa [Joliffe 2016].

Així, per cada component principal s'obté una nova variable a través de la transformació lineal de les variables originals multiplicades per un coeficient, diferent per cada variable del *dataset* original. Aquests coeficients són més coneguts amb el nom de pesos.

[1] La *core distance* és la distància des del punt x als seus n veïns més propers.

CAPÍTOL 3

Estat de l'Art

La robustesa d'una xarxa reflecteix la capacitat d'aquesta alhora continuar funcionant correctament davant de fallades o atacs [Marzo 2019], però és una característica difícil de mesurar.

Tradicionalment, la robustesa d'una xarxa és una mesura que pot ser aproximada amb mètriques de connectivitat com *Fractional Size Largest Component*, la fracció de nodes al component connectat més gran de la xarxa, o *Average Two Terminal Reliability*, la probabilitat de què dos nodes qualssevol de la xarxa estiguin connectats.

Tanmateix, l'any 2013, Trajanovski i altres van proposar una forma original de calcular-la com el sumatori de diverses mètriques [Trajanovski 2013]: $R = \sum_{n}^{k=1} s_k t_k$ on s_k representa el pes de la mètrica i t_k el valor de la mètrica k. Aquest mètode, però, presenta dos problemes: quines mètriques es fan servir i quin pes té cada mètrica en el còmput de la robustesa.

L'any següent, Manzano i altres membres del grup de recerca BCDS van proposar la idea d'utilitzar un algorisme de reducció de la dimensionalitat, el PCA (*Principal Component Analysis*) per trobar els pesos de les mètriques [Manzano 2014].

El grup BCDS va continuar treballant en el tema fins que el 2018 van recomanar una llista de deu mètriques per calcular la robustesa d'una xarxa [Marzo 2018]. Les mètriques recomanades, de les que es pot trobar una descripció més detallada a la secció 4.2, van ser:

- *Average Nodal Degree*: el grau nodal mitjà.
- *Efficency*: l'eficiència.
- *Largest Eigenvalue*: el valor propi més gran.
- *Largest Connected Component*: el component més gran connectat.
- *Average Two Terminal Reliability*: la fiabilitat mitjana entre dos terminals.
- *Algebraic Connectivity*: la connectivitat algebraica.
- *Natural Connectivity*: la connectivitat natural.

- *Edge Betweenness Centrality*: la centralitat entre arestes.
- *Closeness Centrality*: la centralitat de proximitat.
- *Eigenvector Centrality*: la centralitat del vector propi.

Tanmateix, l'última mètrica de la llista, l'*Eigenvector Centrality*, ja no s'utilitza per calcular la robustesa en la versió actual del NRS, perquè s'ha modificat perquè retorni el vector propi en comptes del valor de la centralitat.

Aquest procés d'eliminació de les mètriques es va fer de forma iterativa, tenint en compte l'escalabilitat, algunes de les mètriques són molt costoses de calcular en xarxes grans perquè s'han de trobar tots els camins mínims entre totes les parelles de nodes, la seva variància i la seva correlació.

Així doncs, per calcular el valor de la robustesa d'una xarxa s'han de fer els passos següents:

1. S'ataca la xarxa original, eliminant els elements de forma seqüencial fins a un valor P ($p = [1, P)$). Això es repeteix M vegades el que genera MxP xarxes noves a partir de l'original.
2. Per cada xarxa generada en el punt anterior es calcula el valor de les deu mètriques recomanades generant una matriu de $(MxP)x10$.
3. Sobre la matriu de mètriques s'aplica la reducció de la dimensionalitat amb el PCA i s'escull la primera component principal, que explica la la variància més gran.
4. Els pesos de la primera component principal es normalitzen perquè la robustesa de la xarxa original sigui u.
5. Es multipliquen totes les files de la matriu de mètriques pels pesos normalitzats obtinguts en el punt anterior.
6. Es calcula la robustesa mitjana de la xarxa a partir del vector de robusteses.

El grup BCDS va seguir treballant amb el tema i l'any 2019 va publicar un article [Marzo 2019] sobre com evoluciona la robustesa amb atacs molt profunds i va determinar que a partir del trenta per cent d'elements atacats, la xarxa ja està completament destrossada pel que no és necessari calcular la robustesa amb una P superior a aquest número, paràmetre dels atacs explicat a la subsecció 4.1.1. Aquest fenomen es pot observar a les figures 7.3 de la secció 7.1 en el que s'ha repetit aquest experiment per trobar un valor de la P en què es vegi una diferència clara entre les xarxes més robustes i les menys.

Finalment, també s'ha de mencionar el treball de final de grau d'en Martí Madrenys, company del grup BCDS, perquè segueix amb la línia proposada anteriorment d'utilitzar la intel·ligència artificial per enriquir el treball previ del grup i que ha modelitzat amb molt d'èxit un algorisme per predir el valor de la robustesa per una xarxa i una P concretes [Madrenys 2022].

CAPÍTOL 4

Simulador i Dades

4.1 Network Research Simulator

El Network Research Simulator (NRS) [Marzo 2022] és una eina de desenvolupament pròpia del grup d'investigació BCDS que permet tant la visualització de xarxes en 2D, en un mapa o en 3D, com la generació de noves xarxes sintètiques. Tanmateix, la tasca principal del NRS és calcular les diferents mètriques d'una xarxa, explicades a la secció següent, i la robustesa, explicada al capítol 3.

Una de les principals característiques del NRS davant del seu predecessor, el NRS-1, és que permet un alt grau de flexibilitat a l'hora d'afegir nous experiments al simulador definint simplement dos conjunts diferents:

- Conjunt d'objectes $\mathscr{O}$ que poden contenir els següents elements: xarxes, mètriques i valors.
- Conjunt d'iteracions $\mathscr{I}$ que transforma un conjunt d'objectes $\mathscr{O}_n$ en un de nou $\mathscr{O}_{n+1}$.

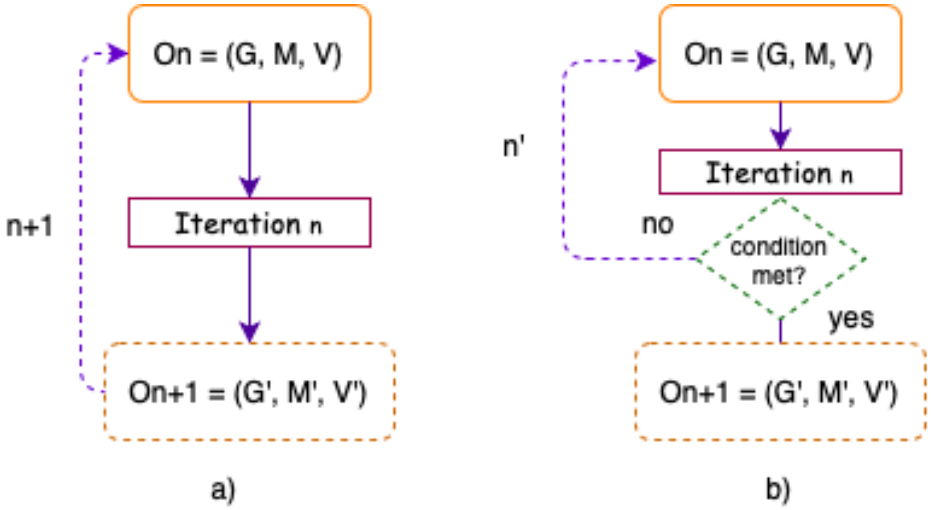


Figura 4.1: Iteracions: a) seqüencial i b) condicional.

A més, l'arquitectura, que es pot veure a la figura 4.2, és molt escalable perquè permet la replicació dels servidors de computació que s'encarreguen de la generació de les xarxes, simulació dels atacs i càlcul de les mètriques i de la robustesa o de qualsevol classe d'experiment que s'afegeixi.

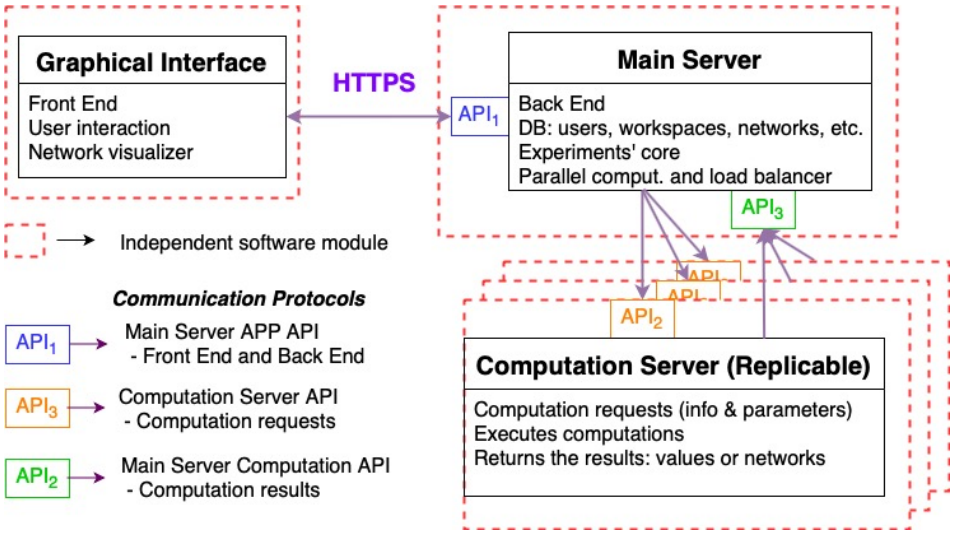


Figura 4.2: Arquitectura del Network Research Simulator

Tot i això, en aquest punt del desenvolupament hi ha un coll d'ampolla al servidor principal del NRS en quan eficiència perquè, com s'ha implementat en JavaScript, el servidor principal només té un nucli. Això fa que quan el nombre de computacions a executar és molt gran, el servidor principal està tan ocupat gestionant la paral·lelització que no és capaç de respondre amb un temps raonable les peticions de la interfície gràfica o d'enviar noves computacions als servidors de computació.

4.1.1 Experiment de robustesa

L'experiment de robustesa, explicat al capítol 3, encara és el més típic del nou simulador i per poder executar-lo s'han de definir els següents elements:

1. Xarxes: una o més xarxes a les que s'ha d'executar l'experiment.

2. Atacs: un o més atacs que s'han d'aplicar a la xarxa. Cada atac té els paràmetres següents:

   - Tipus d'atac: aleatori o dirigit per l'estratègia que s'utilitza a l'hora d'eliminar els elements de la xarxa.
   - Elements a atacar: nodes o enllaços.

- Política de l'atac: simultània o progressiva. Aquest paràmetre només és rellevant amb els atacs dirigits. Quan la política és simultània, la mètrica per prioritzar els elements atacats es calcula només al principi; quan la política és progressiva, la mètrica es calcula cada vegada que s'elimina un element.
- Nombre d'atacs (M): quants atacs diferents s'han de fer sobre la xarxa original.
- Rang de d'elements atacats (P): quants elements s'ataquen de la xarxa. Pot ser en percentatge o en discret.
- Mètrica: mètrica que s'utilitza a l'hora de calcular la probabilitat d'eliminar els elements de la xarxa. Actualment hi ha cinc: *Betweenness Centrality*, *Closeness Centrality*, *Eigenvector Centrality*, *Nodal Degree* i *Critical Nodes*.

3. Mètriques: llista de mètriques que es vol calcular per cada xarxa atacada. Si es vol calcular també la robustesa, es recomana fer servir la llista de mètriques explicada al capítol 3.

4. Computacions opcionals: per si es vol calcular les estadístiques de cada mètrica i P; o la robustesa de les xarxes.

El nombre de computacions que s'han de realitzar per calcular la robustesa és molt alt: $n(X * M + X * M * P + 2)$ on n és el nombre de mètriques d'atac seleccionades més u si també s'ha seleccionat *Random*, X és el nombre de xarxes, M és el nombre d'atacs i P és el nombre d'elements dintre del interval.

Com es pot veure, el nombre de computacions escala principalment amb la M pel que el temps d'execució incrementa considerablement amb valors alts de M i és l'únic paràmetre, juntament amb la *P*, que el seu valor no està predefinit per la naturalesa de l'experiment. L'estudi de com diferents M afecten el temps d'execució i la precisió del càlcul de robustesa es troba a la secció 7.1

4.2 Dades

El *dataset* que s'ha decidit fer servir és el Topology Zoo [Kinght 2011], una col·lecció de xarxes de telecomunicacions molt utilitzada en la literatura relacionada. Un altre repositori interessant de xarxes de moltes tipologies diferents és el Network Data Repository [Rossi 2015] que a més compta amb un visualitzador molt potent.

En concret, s'ha escollit un subconjunt del Topology Zoo amb aquelles xarxes que tenen almenys quinze nodes i que estan inicialment connectades. Només

dues-centes cinc de les dues-centes seixanta-una xarxes compleixen aquestes condicions.

Cada una d'aquestes xarxes té associada els valors de trenta mètriques que es calculen amb el NRS. Aquestes mètriques es poden trobar a la taula 4.1, una simplificació d'un document intern del BCDS. Algunes de les mètriques, com *Number of nodes* o *Number of links*, no tenen descripció o no tenen com es calculen perquè són trivials. Altres, com *Edge Connectivity* o *Node Connectivity*, perquè el càlcul és massa complicat per posar en una cel·la de la taula. La notació de les fòrmules es pot trobar a la taula 4.2.

Taula 4.1: Mètriques del NRS

Name	*Symbol Units*	*Description*	*Computation*
ND Nodal degree			
Number of nodes	natural		
Number of links (edges)	natural		
Nodal Degree	array (natural)	The degree of a node is the number of incident edges	
Average Nodal Degree	real+	The degree of a node is the number of incident edges. Thus, AND is the average of all node degrees	1) From adjacency matrix A 2) $AND = \frac{2N}{\#L}$ 3) Needs to compute Nodal Degree: AND = mean(ND)
Maximum Nodal Degree	natural		Needs to copmute Nodal Degree: max(ND) Special case: complete graph. The maximum number of links is: $L_{max} = \frac{N(N-1)}{2}$ then its AND $= N-1$
Histogram Nodal Degree	array (natural)	Distribution of nodal degrees 0, 1, 2, ...	Needs to compute Nodal Degree
Heterogeneity	real	It is the standard deviation of the nodal degree divided by the average nodal degree	Needs to compute ND $HET = \frac{sd(deg(G))}{mean(deg(G))} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - x^{mean})^2}$ where x_i is the ND of node i
Shortest Path Length			
Average Shortest Path Length	real	Average shortest path length between every node pairs of the network	Needs to compute shortest distance between all pair of nodes $ASPL = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}$ where d_{ij} is the distance between node i and node j
Diameter	natural	Maximum shortest path length between every node pairs of the network	Needs to compute shortest distance between all pair of nodes
Efficiency	real+	This is the averaged sum of the inverse of all the distances (length of the shortest path) between all pairs of nodes (conductance)	Needs to compute shortest distance between all pair of nodes $E = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\frac{1}{d_{ij}}$
Laplacian			
Effective (Graph) Resistance	real+	This considers the graph as an electrical circuit, where an edge is a resistor. The effective resistance between two nodes can be calculated by series and parallel manipulations (Kirchhoff's Current Law) Laplacian. ER is the sum of the effective resistances over all pairs of vertices	Needs to compute Laplacian Eigenvalues $R = \sum_{1\le i=j\ge n} Rij = n\sum_{i=2}^{n}\frac{1}{\lambda_i}$
Laplacian Eigenvalues	array (real)		

Name	*Symbol Units*	*Description*	*Computation*
Number of Spanning Trees	ξ natural (large)	A spanning tree is a subgraph containing $N-1$ edges, all N nodes and no cycles ξ is a function of the unweighted Laplacian eigenvalue	Needs to compute Laplacian Eigenvalues $\xi = \frac{1}{n}\prod_{i=2}^{n}\lambda_i$
Clustering Coefficient	real	CC measures the probability that the adjacent nodes of a node are connected (among them). CC captures the presence of triangles and compares it to the number of connected triplets	$CC = \sum_{i=1}^{N} \frac{2e_i}{k_i(k_i-1)}$
Assortativity	real $[-1,1]$	r is the preference for a network's node to attach to others which have a similar nodal degree	$r = \frac{M^{-1}\sum_i j_i k_i - [M^1 \sum_i \frac{1}{2}(j_i+k_i)]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2+k_i^2) - [M^1 \sum_i \frac{1}{2}(j_i+k_i)]^2}$
Eigenvalues			
EigenValues	array (real)	Eigenvalues of the graph (obtained from the adjacency matrix)	$Ax = \lambda x$ the number of eigenvalues is the dimension of the adjacency matrix (number of nodes N)
Symmetry Ratio	real	SR is the quotient between the number of distinct eigenvalues of the graph (obtained from the adjacency matrix) and the diameter	Needs to compute shortest distance between all pair of nodes Needs to compute EigenValues
Largest Eigenvalue	real	Largest Eigenvalue of the adjacency matrix)The higher the more robust. This is also associated in defining the epidemic threshold	Needs to compute EigenValues $LE = \lambda_1$
Connectivity			
Largest Connected Component	natural	LCC is the number of nodes of the largest connected component of the network. This metric is a measure of the global connectivity of the network	Needs to compute the graph's components
Analytical LCC	real	Estimation based on the histogram of nodal degrees	Needs to compute Nodal Degree
Fractional Size Largest Component	real	Fraction of nodes of the largest connected component of the network. This metric is a measure of the global connectivity of the network	Needs to compute the graph's components
Average Two Terminal Reliability	real	ATTR is the probability that a randomly chosen pair of nodes remains connected after a failure. It can be computed as the sum over the number of node pairs in the connected component divided by the total number of node pairs (according to a ratio of removed nodes)	Needs to compute the graph's components
Lower Bound ATTR	real	The ATTR considering only the LCC is connected	Needs to compute the graph's components
Upper Bound ATTR	real	The ATTR considering there are only two components: the LCC and all the other nodes	Needs to compute the graph's components
Degree of Fragmentation	natural	Number of connected components (islands)	Needs to compute the graph's components
Edge (Link) Connectivity	natural	The minimal number of edges which has to be removed to disconnect the graph (aka Resilience Factor)	
(Vertex) Nodes Connectivity	natural	The minimal number of nodes which has to be removed to disconnect the graph	

Name	*Symbol Units*	*Description*	*Computation*
Algebraic Connectivity	real $[0,2]$	It measures how difficult it is to break the network into disconnected components. Higher values indicates better robustness against both node and link removal. Graphs with identical algebraic connectivity can be compared using natural connectivity	Needs to compute Laplacian Eigenvalues It is defined as the second smallest Laplacian eigenvalue $AC = \mu_{n-1}$
Natural Connectivity	real	It is based on the number of closed walks (cw) in a graph, that can be related to the sum of eigenvalues. A walk of length k is a path over the nodes and links of the graph, starting at v0, passing k - 1 nodes and k links to end in vk. If v0 = vk this path is called a closed walk	Needs to compute EigenValues $\lambda = ln(\frac{S}{n}) = ln(\frac{\sum_{i=1}^{n} e^{\lambda_i}}{n})$
Centrality			
Degree Centrality	real		
Node Betweenness Centrality	real	This measures the fraction of shortest paths that pass through a given node, averaged over all pairs of node in a network	
Edge Betweenness Centrality	real	This measures the fraction of shortest paths that pass through a given link	
Maximum Edge Betweenness	natural	The maximum of shortest path that pass through a given link	
Closeness Centrality	real [0,2]	This measures the degree to which a node is close to other nodes on average considering shortest paths (per node).	$CLC = \sum_i \frac{1}{\sum_j shortestpath(i,j) \forall i \neq j}$
Eigenvector Centrality	array	It is based on the idea that if an important node is connected to important neighbors. The EC of a node is equal to its component of the eigenvector corresponding to the largest eigenvalue of the adjacency matrix	It can be calculated by iterations. $x_i(t+1) = \sum_{j=1} A_{ij} x_j(t)$
End of Table			

Taula 4.2: Notació de les fòrmules de les mètriques

Symbol	*Name*	*Units*	*Description*
N	Number of nodes	natural	Order of the graph
L	Number of links	natural	In some cases this is denoted as M (size of the network) but M is used in NRS for the number of trial of an experiment
G	Network (Graph)	$G(N,V)$	N set of N nodes, V set of L links
A(G)	Adjacency matrix of G	matrix $(N \times N)$	A specifies the interconnection pattern of the graph. The element of the adjacency matrix $a_{ik} = 1$ only if the pair of nodes i and k are connected by a direct link; otherwise $a_{ik} = 0$
λ_i	i_{th} largest eigenvalue of $A(G)$	array (real)	
μ_i	Eigenvalue of Laplacian for node$_i$	real	
End of Table			

A més a més, cada xarxa té associat els resultats obtinguts dels experiments de robustesa en un fitxer JSON amb el contingut explicat a continuació:

1. Identificador de la xarxa.
2. Paràmetres de l'atac que s'ha realitzat:
    - Codi de l'atac: és una cadena de caràcter que codifica els paràmetres de l'atac explicats a la subsecció 4.1.1. El primer caràcter pot ser 'R' o 'T' per *Random* o *Targeted*; el segon caràcter pot ser 'N' o 'E' per nodes o arestes; el tercer caràcter pot ser 'S' o 'P' per simultani o progressiu; i els dos últims caràcters poden ser 'BC', 'CC', 'CN', 'EV' o 'ND' per *Betweenness Centrality*, *Closeness Centrality*, *Critical Nodes*, *Eigenvalues Centrality* o *Nodal Degree*.
    - M: el nombre d'atacs.
    - P: la profunditat dels atacs. Un vector amb els percentatges dels elements atacats que s'han de simular.
3. Llista de mètriques que s'han de computar per cada xarxa simulada després de l'atac. Cada mètrica conté l'identificador únic i el valor per la xarxa original.
4. Un objecte amb tres valors *booleans* per si s'han de calcular les computacions opcionals de:
    - Estadístiques: si és cert, es calculen els màxims, mínims i mitjanes per cada P i cada mètrica.
    - PCA: si és cert, es calculen els pesos del *Principal Component Analysis*. És un requisit si es vol calcular la robustesa.
    - Robustesa: si és cert, es calcula la robustesa per cada xarxa atacada i la mitjana. A més, normalitza els pesos del PCA perquè la xarxa original tingui una robustesa igual a u.
5. El resultat dels atacs: és una matriu de mida MxP amb els elements eliminats de la xarxa original.
6. El resultat de les mètriques: és una matriu de mida MxP (nombre de xarxes generades amb els atacs) per N (nombre de mètriques).
7. El resultat del càlcul de les estadístiques: és un vector de mida N on cada cel·la conté el mínim, màxim i mitjana de cada mètrica.
8. El resultat del PCA: és un vector de mida N amb els pesos de cada mètrica.

9. El resultat de la robustesa: és un objecte amb la robustesa mitjana, els pesos anteriors normalitzats i una matriu de mida M per N amb la robustesa de cada xarxa generada durant els atacs.

Tal com es pot veure, el fitxer resultant de l'experiment de robustesa del NRS és francament complicat i gran donat el gran nombre de valors intermedis que s'han de calcular. Finalment, com s'ha de poder fer l'associació entre els identificadors de les mètriques i de les xarxes amb un nom que un humà entengui hi ha dos fitxers JSON extres que tenen aquesta relació, però aquests no s'expliquen aquí perquè no tenen massa interès.

CAPÍTOL 5

Planificació i Metodologia

5.1 Metodologia

La metodologia de desenvolupament que s'ha escollit per gestionar la tesi és la *Cross-Industry Standard Process for Data Mining* (CRISP-DM) perquè és la més habitual en la indústria per projectes de minat de dades, anàlisi i ciència de dades [Saltz 022].

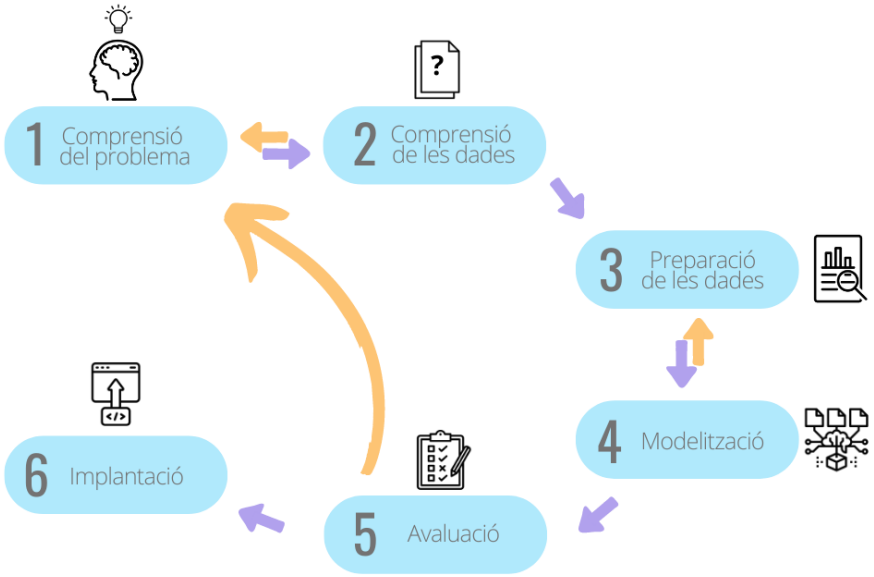


Figura 5.1: Diagrama de la metodologia CRISP-DM.

El cicle de vida de la metodologia, representat en el diagrama 5.1, consta de sis fases:

1. **Comprensió del problema**: S'han d'entendre els objectius i els requisits del projecte.
2. **Comprensió de les dades**: S'han de recol·lectar les dades inicials i elaborar una descripció, exploració i verificació de la qualitat de les dades.

3. **Preparació de les dades**: S'ha de netejar i formatar les dades, seleccionar les característiques rellevants i generar-ne de noves.

4. **Modelització**: S'ha de seleccionar el model d'intel·ligència artificial que s'implementarà. Finalment, s'avaluarà el model i el seu rendiment.

5. **Avaluació**: S'ha d'avaluar el procediment i els resultats del projecte. També s'han de determinar quins són els següents passos.

6. **Implantació**: S'ha d'elaborar un pla d'estratègia tant pel desplegament del model a producció això com el seu monitoratge i manteniment. També s'ha de realitzar la documentació final i revisió del projecte.

Una descripció més detallada de les tasques i de les sortides de cada etapa de la metodologia es pot veure a la figura 5.2 [Chapman 2000]. Tanmateix, al tractar-se d'una tesi en el que la seva component principal és la recerca, algunes de les tasques i sortides no s'han realitzat.

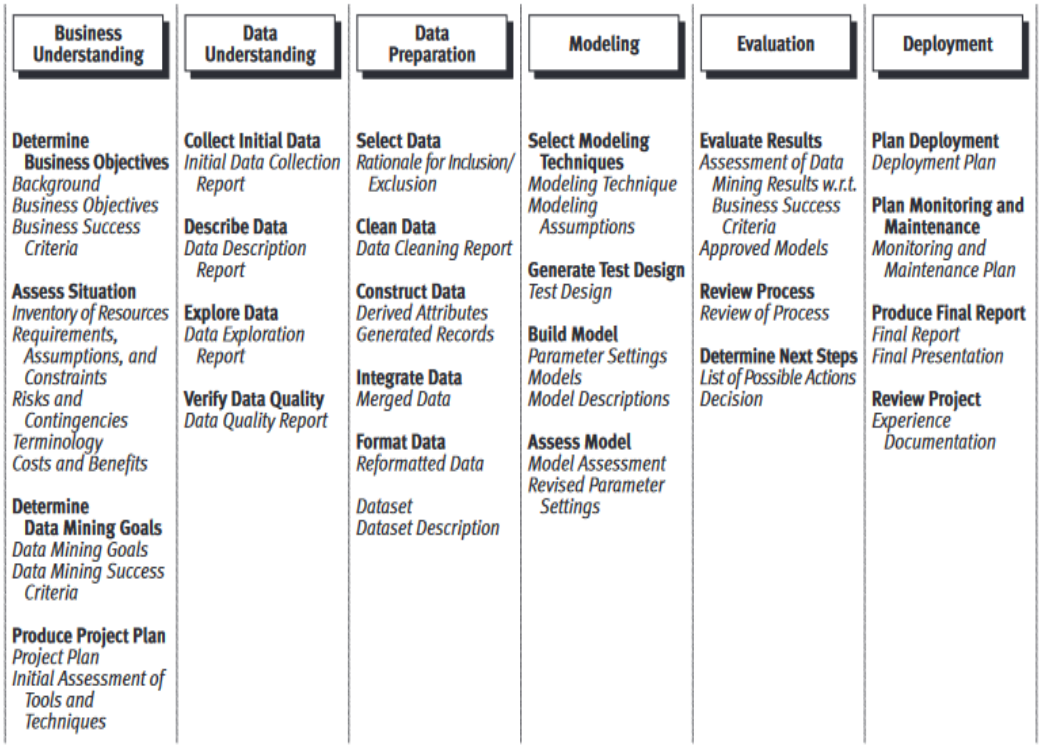


Figura 5.2: Diagrama de les tasques (negreta) i sortides (cursiva) de la metodologia CRISP-DM.

5.2 Planificació

D'acord amb la metodologia explicada a la secció anterior, s'han planificat les següents tasques:

5.2.1 Planificació del projecte

En aquesta primera tasca s'ha decidit els objectius i els algorismes d'intel·ligència artificial que s'utilitzaran juntament amb el tutor de la tesi, l'Eusebi Calle Ortega, i en Marc Comas Cufi, professor del màster en Ciència de Dades, i s'ha realitzat la temporització del projecte que es pot veure al diagrama de Gantt de la figura 5.3.

5.2.2 Estudi de l'estat de l'art

Després de la planificació del projecte, s'ha de llegir la literatura relacionada amb el tema que ha recomanat el tutor de la tesi: *Robustness envelopes of networks* [Trajanovski 2013], *Robustness surfaces of complex networks* [Manzano 2014], *On selecting the relevant metrics of network robustness* [Marzo 2018] i *A study of the robustness of optical networks under massive failures* [Marzo 2019]. També s'ha de llegir la documentació bàsica dels algorismes d'intel·ligència artificial que es volen utilitzar: UMAP i HDBSCAN.

5.2.3 Estudi previ a la generació de les dades

Una de les característiques més importants a l'hora de generar el *dataset* és l'elecció dels paràmetres dels diferents atacs que es realitzen a les xarxes. En aquesta tasca s'ha d'estudiar com afecten diverses M i P en un subconjunt de les xarxes del Topology Zoo.

5.2.4 Generació de les dades

Una vegada escollida la M i la P dels atacs d'acord amb l'estudi anterior, s'han de llançar els experiments necessaris al NRS2 pel *dataset* complet de Topology Zoo que compleixen les restriccions següents: tenen almenys 15 nodes i estan connectades inicialment.

5.2.5 Preparació i modelització

Aquesta tasca inclou les fases de **Preparació de les dades** i **Modelització** de la metodologia explicada perquè són dues etapes molt lligades i que a la pràctica

costen de separar. Això ja es contempla en la CRISP-DM perquè, tal com es pot veure en el diagrama 5.1, hi ha la possibilitat de tornar enrere una vegada s'està fent la modelització del problema. Per aquest motiu, les tasques de Tractament de les dades, Modelització dels algorismes d'intel·ligència artificial i Avaluació del model són concurrents al diagrama 5.3.

5.2.6 Redacció de la memòria

A principis d'agost s'ha de començar a escriure la memòria per tenir prou temps d'acabar-la.

5.2.7 Implementació de noves millores

Si després d'acabar la redacció de la memòria encara hi ha prou temps abans d'entregar la tesi final de màster, s'explorarà altres formes de millorar el model final.

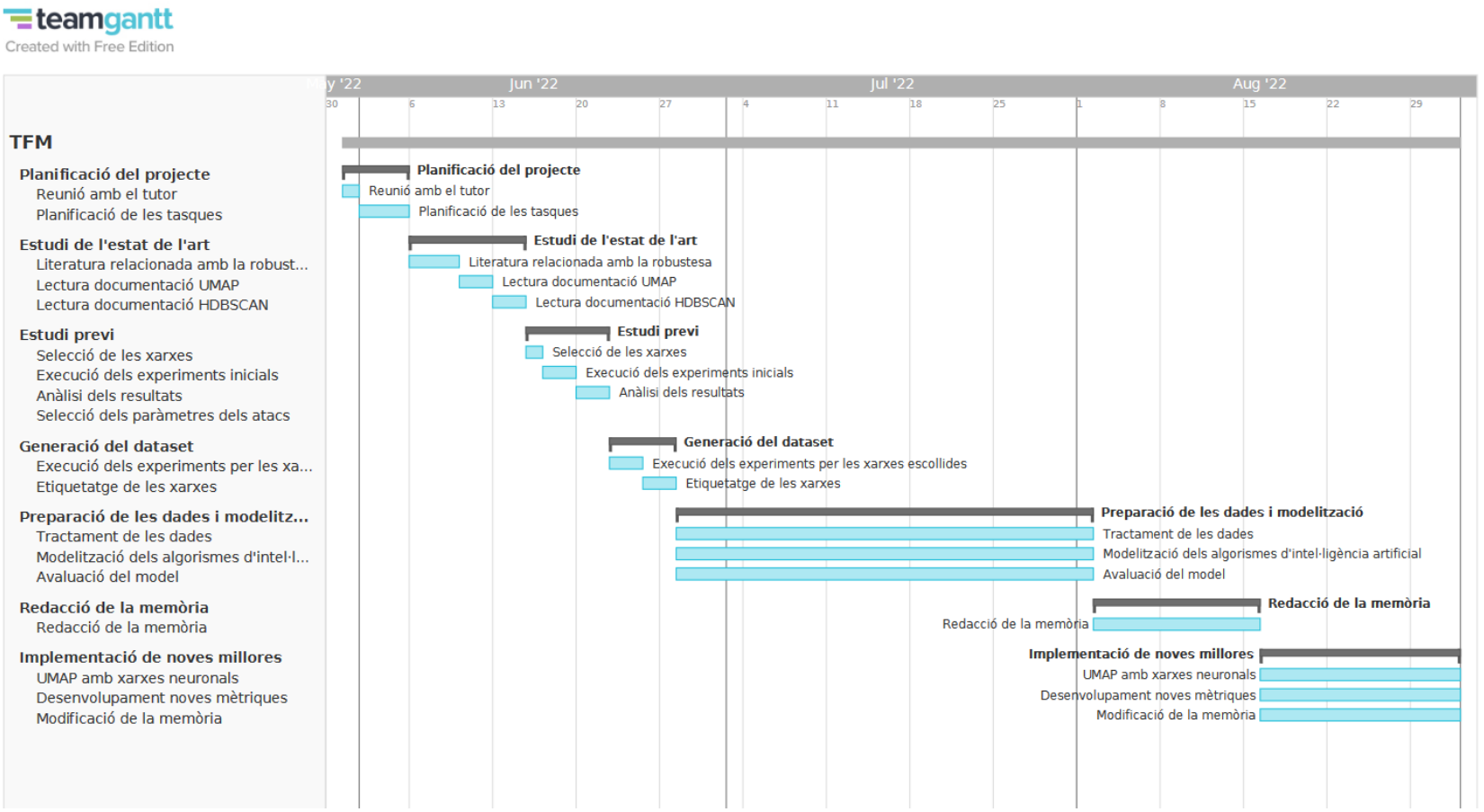


Figura 5.3: Diagrama de Gantt amb la temporització de la tesi.

CAPÍTOL 6

Contribució Metodològica

El mètode que s'ha proposat per la tesi és el següent:

1. Realitzar un estudi previ de com rellevant és l'elecció de la M per la precisió i la consistència dels atacs.

2. Trobar per quina P la variància de la robustesa de les xarxes és màxima; és a dir, per quina P la diferència entre les xarxes més robustes i les menys robustes és més clara.

3. Generar el *dataset* amb la robustesa de les xarxes seleccionades del Topology Zoo.

4. Categoritzar les xarxes segons la seva robustesa, ja sigui per quin atac és més devastador o, com s'ha fet al final, per la seva robustesa respecte de l'atac *Betweenness Centrality*.

5. Explorar quines mètriques són les més rellevants per explicar la robustesa de la xarxa i com es diferencien entre les etiquetes.

6. Clusteritzar les xarxes utilitzant l'UMAP i l'HDBSCAN perquè a partir de les mètriques originals de la xarxa sense atacar es pugui saber si una xarxa és robusta o no.

7. Millorar la precisió de l'UMAP i l'HDBSCAN aplicant tècniques de *feature selection*, *feature engineering* i *data augmentation*.

Un dels primers problemes que poden sorgir és que totes les xarxes del *dataset*, en tractar-se d'una tipologia molt concreta: són xarxes de telecomunicacions, siguin vulnerables al mateix atac. Això es pot solucionar afegint noves tipologies al *dataset*, com xarxes d'aigua o xarxes socials; o modificant l'estratègia que s'utilitza a l'hora d'etiquetar-les.

Una segona dificultat que es pot trobar és que els clústers que es trobin amb el HDBSCAN poden no ser reals, sinó artefactes de l'UMAP [Shchubert 022]. Això pot passar perquè, com el t-SNE, l'UMAP no preserva les distàncies ni la densitat.

Tanmateix, a vegades pot resultar útil reduir la dimensionalitat de grans *datasets* abans de la clusterització. Una estratègia similar ha funcionat bé amb dades molt més complexes [Shekhar 2016].

La innovació que s'aporta amb aquesta tesi és tant l'estudi de com les mètriques expliquen que una xarxa sigui poc robusta com la utilització de tècniques d'intel·ligència artificial per classificar xarxes segons la seva robustesa perquè, com s'ha vist al capítol 3, la literatura relacionada i l'estat de l'art s'ha concentrat a trobar aquest valor de la robustesa, no a explicar-lo.

CAPÍTOL 7

Resultats

7.1 Estudi Previ

7.1.1 Precisió del càlcul de la robustesa

Un dels paràmetres més rellevants a l'hora de definir l'atac és la M, tal com s'ha vist en el capítol 3 i en la subsecció 4.1.1, però cap de la literatura llegida adreça de quin és el seu valor idoni.

Tradicionalment, al grup BCDS, s'ha fet servir una $M = 100$, però és aquest valor prou alt per assegurar que el valor de la robustesa és estable? Per respondre a aquesta pregunta prèvia a la generació del *dataset* s'ha realitzat l'estudi actual.

Primer, s'han escollit deu xarxes del Topology Zoo amb més de trenta nodes. Aquestes són: BtNorthAmerica, Surfnet, Dfn, Garr201201, Esnet, Uninett2011, VtlWavenet2011, Deltacom, GtsCe i Cogentco.

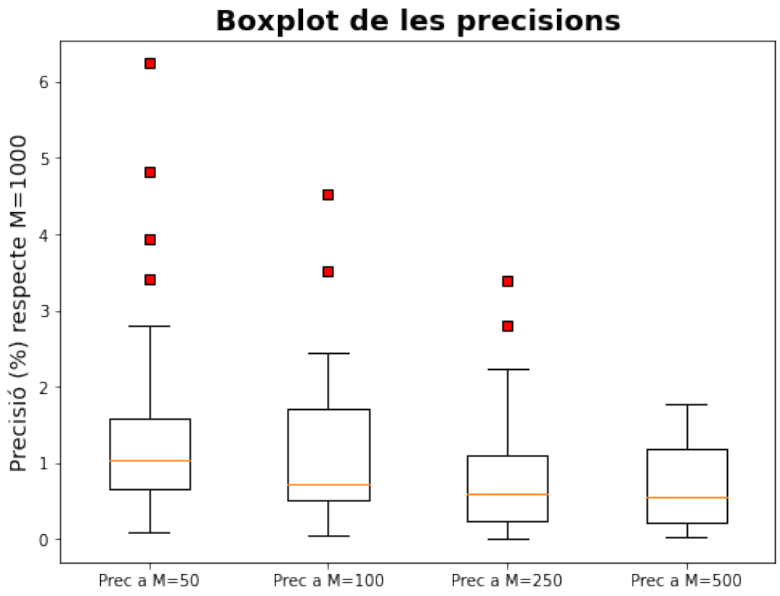


Figura 7.1: Diagrama de caixes de la precisió en les diferents M

Després, s'ha fet una primera iteració en el que s'ha fixat la P d'1 al 50% i s'han executat els quaranta experiments de robustesa, un per cada xarxa i per

Xarxa	Arestes	M=50	M=100	M=250	M=500	M=1000
BtNorthAmerica	76	462	923	2375	4945	10143
Surfnet	73	461	921	2321	4747	10324
Dfn	87	462	921	2351	4876	10480
Garr201201	89	462	922	2362	4877	10087
Esnet	92	462	920	2355	4868	10120
Uninett2011	98	463	929	2356	4865	10213
VtlWavenet2011	96	462	923	2354	4863	10617
Deltacom	183	463	926	2355	4867	10629
GtsCe	193	463	927	2366	4887	10343
Cogentco	245	464	930	2373	4887	10465

Taula 7.1: Taula amb els temps d'execució per cada M.

cada atac dirigit, *Betweeness Centrality*, *Closeness Centrality*, *Eigenvector Centrality* i *Nodal Degree*, amb diferents M. En concret s'han provat 5 M diferents, $M = 50$, $M = 100$, $M = 250$, $M = 500$ i $M = 1000$, executant dos-cents experiments al NRS.

Una vegada s'han acabat els dos-cents experiments anteriors, s'ha calculat l'error relatiu per cada P respecte del valor de $M = 1000$, que al tractar-se del valor més gran de la M s'ha considerat el més exacte. El resultat obtingut es pot veure en el diagrama de caixes de la figura 7.1.

Tal com es podia esperar, la precisió de la robustesa creix amb la M: hi ha menys *outliers* i menys variància, però fins i tot amb les M més petites, l'error relatiu mitjà està al voltant de l'u per cent. Tanmateix, també s'ha de tenir en compte el temps de còmput per fer una valoració objectiva.

El temps d'execució en segons de cada xarxa i pels quatre atacs es pot trobar a la taula 7.1. En aquesta taula ja es pot començar a veure el *bug* del NRS, explicat a la secció 4.1 perquè el temps d'execució entre $M = 500$ i $M = 1000$ és més que el doble.

La relació entre l'error relatiu de la precisió i el temps mitjà d'execució es pot veure a la figura 7.2, al que no s'ha afegit $M = 1000$ perquè en cap moment es va contemplar de fer servir una M tan alta: el temps teòric, sense tenir ja en compte el *bug* mencionat anteriorment, seria de gairebé vint-i-cinc dies. Els únics valors realistes de M són entre cent i dos-cents cinquanta i, tanmateix, el temps d'execució de les dos-centes sinc xarxes serà un i dos dies respectivament.

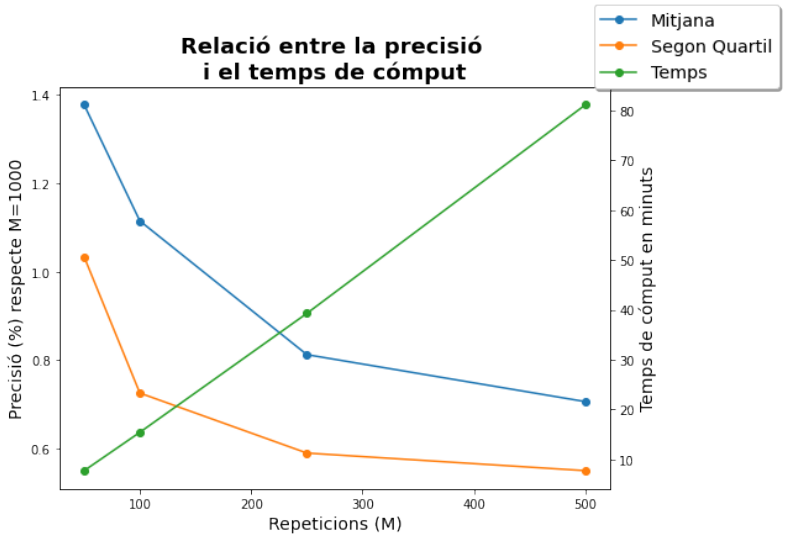


Figura 7.2: Relació entre la precisió i el temps de cómput dels experiments.

7.1.2 Profunditat de l'atac

Per la segona part d'aquest estudi previ, s'ha fixat la M a cent i s'ha analitzat com afecta la P a les diferents xarxes escollides. Això permet veure fins quina P la diferència entre atacs és significativa, ja que, com ja s'ha vist al capítol 3, és que per P relativament altes arriba un moment en què no importa l'atac perquè la xarxa està completament destrossada.

A més, si es redueix la P, també es redueix el nombre de computacions que s'han de calcular i, per tant, el temps d'execució dels experiments serà menor el que pot significar alhora incrementar també la M.

El primer que es pot observar de la figura 7.3 és que el pitjor atac per aquest tipus de xarxes sempre és el *Betweenness Centrality*. En aquest atac els nodes a eliminar es prioritzen pel nombre de camins més curts que passen per ells.

Això es pot explicar per què les xarxes del Topology Zoo són xarxes de telecomunicacions que es caracteritzen per tenir un grau nodal mitjà baix i, com els nodes a vegades poden estar separats per centenars de quilòmetres, normalment no existeixen enllaços que uneixin els extrems de la xarxa.

Tanmateix, si aquesta tendència es repeteix en el *dataset* complet, s'haurà d'estudiar la possibilitat d'afegir noves topologies de xarxes a la investigació perquè el més probable que passi és que pertanyin totes al mateix clúster o canviar la idea original de la tesis per considerar diferents graus de robustesa.

Finalment, també es veu que es pot reduir la severitat de l'atac perquè en la majoria dels casos, a partir del vint per cent, ja es pot considerar que les xarxes, en general, estan completament destrossades. Per aquest motiu, s'ha escollit que la P màxima sigui deu quan es generi el *dataset*.

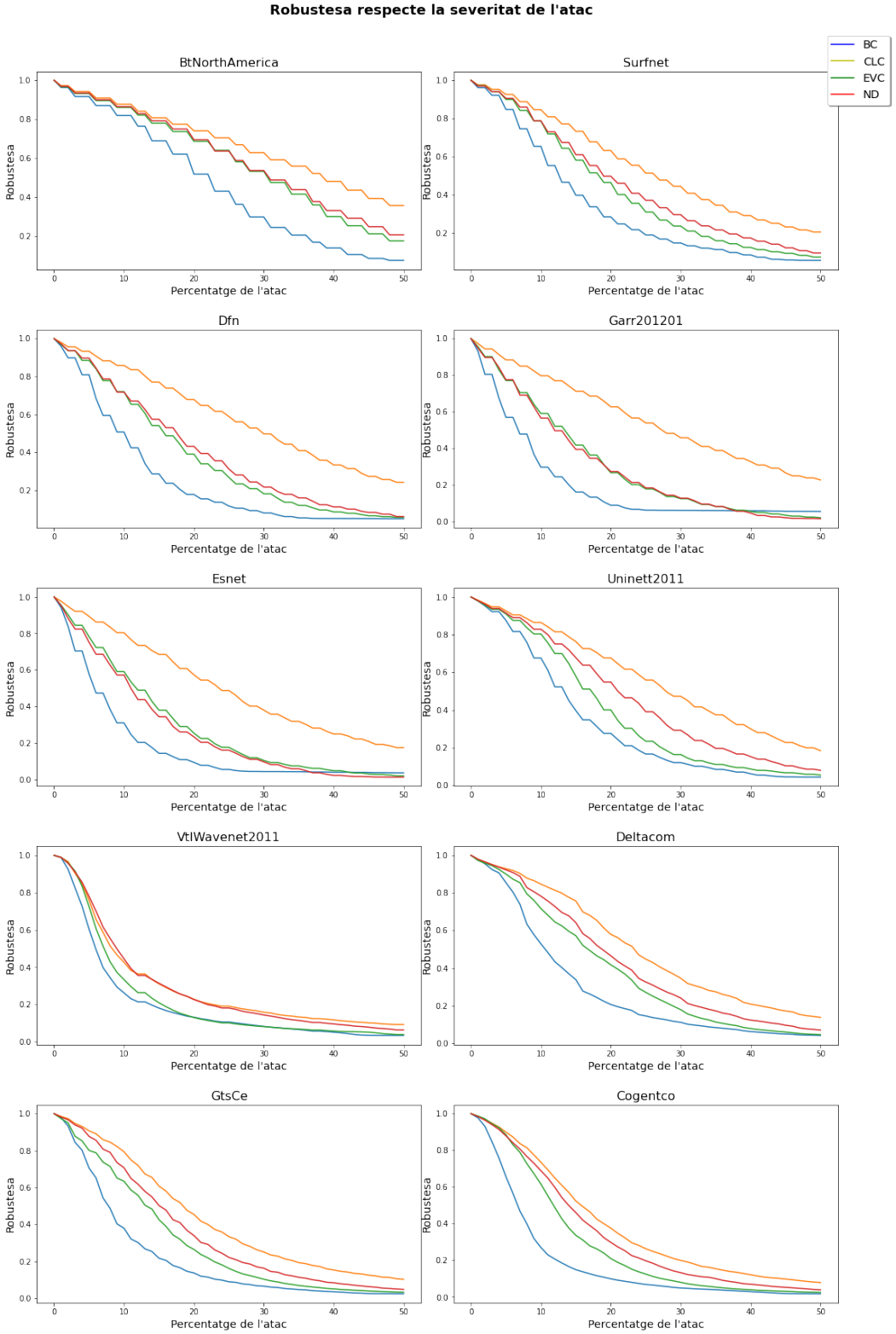


Figura 7.3: Evolució de la robustesa en deu xarxes de telecomunicacions.

S'ha repetit aquest experiment per dues tipologies diferents, però tal com es pot observar a les figures 7.4 i 7.5, la tendència és la mateixa: l'atac més devastador és el *Betweenness Centrality*.

En el primer cas, es pot veure que les xarxes de clavegueram són completament vulnerables a tots els atacs. Això es deu perquè normalment són arbres en els quals hi ha un sol node, la depuradora, al que l'aigua de tots els altres nodes ha d'arribar. No és exactament el cas en aquestes dues xarxes, ja que són només parcials en tractar-se de dos barris de Barcelona.

En el segon cas, la robustesa és molt bona fins que arriba a una profunditat considerable, al voltant del quaranta per cent, però les xarxes Erdös-Rènyi estan molt lluny de semblar-se a una xarxa de telecomunicacions perquè estan molt ben connectades.

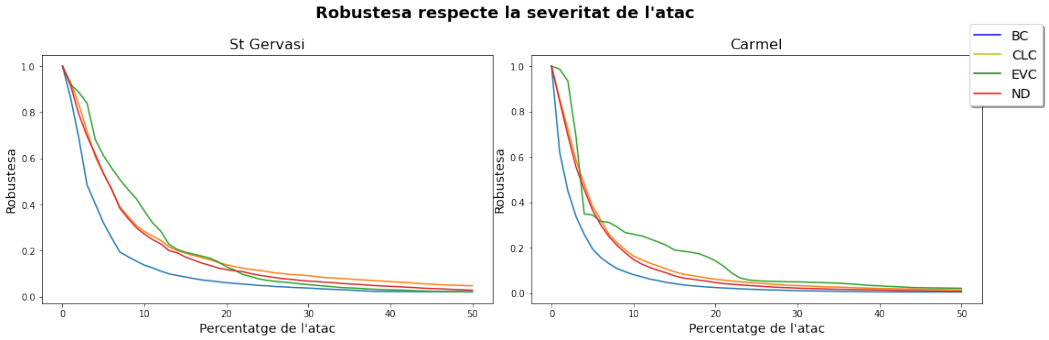


Figura 7.4: Evolució de la robustesa en dos xarxes de clavegueram.

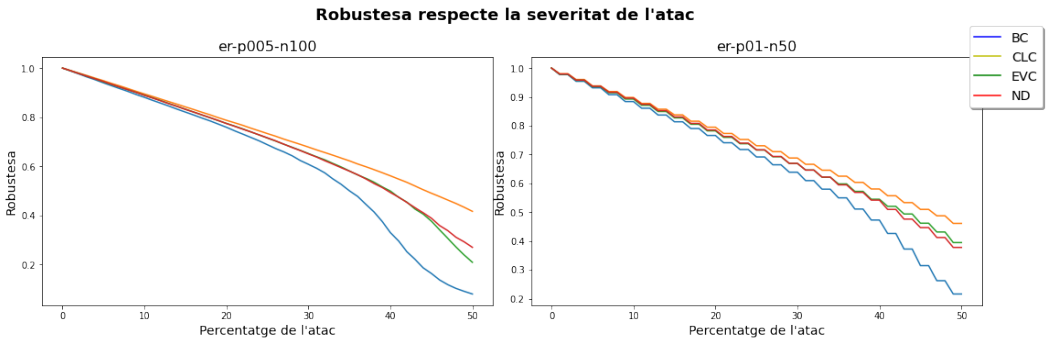


Figura 7.5: Evolució de la robustesa en dos xarxes sintètiques Erdös-Rènyi.

7.2 Anàlisi exploratòria de dades

Després de l'Estudi Previ, s'ha generat el *dataset* amb el NRS. Com a recordatori, s'ha calculat la robustesa de cada xarxa i cada atac escollit, *Betweenness Centrality*, *Closeness Centrality*, *Eigenvector Centrality* i *Nodal Degree*, amb els següents paràmetres: $M = 100$; $P = 10$; i les nou mètriques recomanades llistades al capítol 3.

El *dataset*, doncs, en aquest punt del treball conté les dades següents: nom de la xarxa; codi de l'atac; els valors de totes les mètriques del NRS, excepte aquelles que són vectors; el valor de la robustesa mitjana; i el valor mitjà per cada P.

7.2.1 Exploració inicial i etiquetatge de les dades

Un possible problema que s'ha detectat durant l'Estudi Previ a la generació del *dataset*, subsecció 7.1, és que totes les xarxes del Topology Zoo, en tractar-se de xarxes de la mateixa tipologia, siguin més vulnerables a l'atac *Betweenness Centrality*.

De fet, només per tres xarxes, dos si es considera la robustesa mitjana i una, per $P = 10$, la robustesa és pitjor per un atac que no sigui el *Betweenness Centrality*, però com es pot veure a les taules 7.2 i 7.3, la diferència és insignificant. Per tant, es pot assumir que si una xarxa és resistent a l'atac per *Betweenness Centrality*, llavors és una xarxa robusta per tots els atacs.

Per aquest motiu, la idea original de la tesi, fer servir l'UMAP per reduir la dimensionalitat de les dades com un algorisme semisupervisat en què l'etiqueta és l'atac més vulnerable, no sembla viable i s'ha de buscar una alternativa, ja sigui ampliant el *dataset* amb altres tipologies de xarxa o utilitzar una estratègia diferent a l'hora d'etiquetar.

Primer, s'ha decidit explorar la segona opció i utilitzar el valor de la robustesa per *Betweenness Centrality*, degudament categoritzada, per etiquetar les dades. A la figura 7.6 es pot veure un diagrama de caixes per la robustesa mitjana i a la figura 7.7, per $P = 10$.

En els dos casos, els possibles valors de la robustesa per l'atac *Betweenness Centrality* ocupa gairebé tot el rang de valors possibles, de zero a u, però sorprèn

Xarxa	Robustesa mitjana per BC	Diferència amb l'atac més devastador
Belnet2009	0.802240	0.000724
Nextgen	0.829055	0.007802

Taula 7.2: Diferència entre la robustesa mitjana per *Betweenness Centrality* i l'atac més devastador.

Xarxa	Robustesa a P=10 per BC	Diferència amb l'atac més devastador
Belnet2010	0.598901	0.003045

Taula 7.3: Diferència entre la robustesa a $P = 10$ per *Betweenness Centrality* i l'atac més devastador.

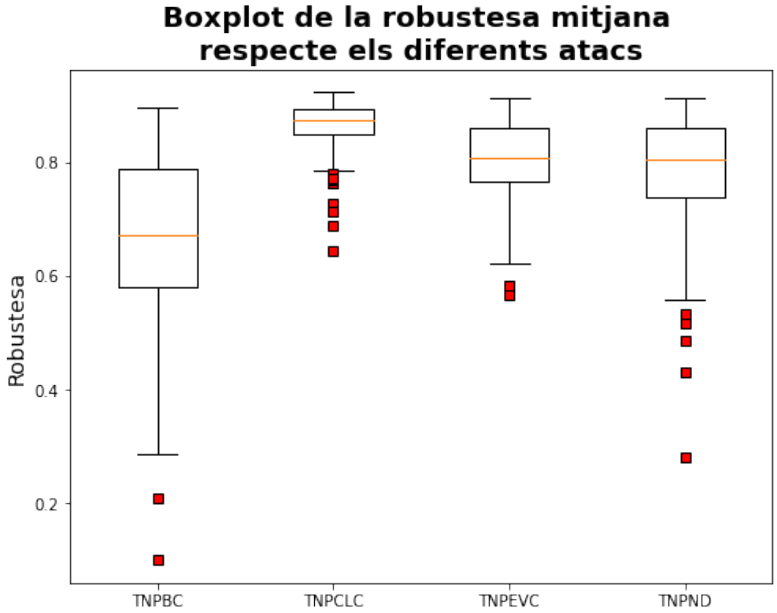


Figura 7.6: Diagrama de caixes de la robustesa mitjana respecte els diferents atacs.

una mica l'existència de dos *outliers* tan baixos a la figura 7.6. Tanmateix, amb només les mètriques més senzilles del NRS, a la taula 7.4, ja es pot començar a intuir quin és el problema: són xarxes poc connectades, però amb nodes molt centrals.

Si es visualitzen les dues xarxes, Ulaknet i Pern, amb el visualitzador del NRS es pot tenir una idea més clara de la seva estructura. Aquestes visualitzacions es troben a les figures 7.8 i 7.9.

Xarxa	N. of Nodes	N. of Edges	Avg. Nodal Degree	Max. Nodal Degree
Ulaknet	82	82	2.0	58
Pern	127	129	2.03	31

Taula 7.4: Mètriques de les dos xarxes amb pitjor robustesa mitjana per *Betweenness Centrality*

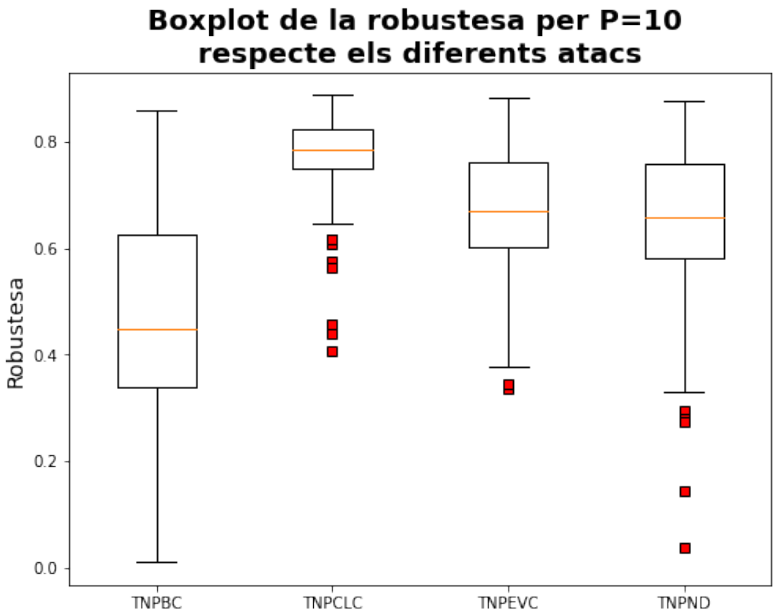


Figura 7.7: Diagrama de caixes de la robustesa mitjana respecte els diferents atacs.

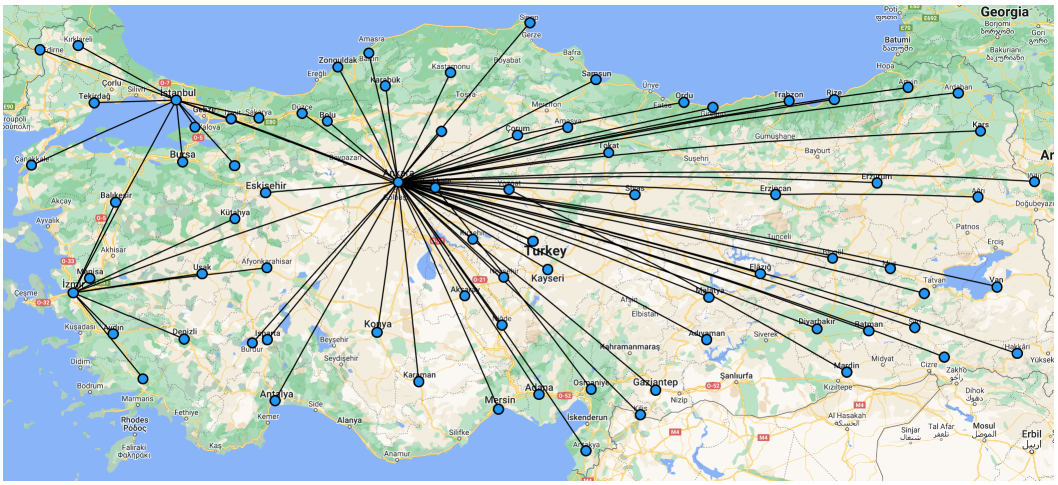


Figura 7.8: Visualització de la xarxa Ulaknet amb el NRS.

La primera xarxa, Ulaknet, és una xarxa de telecomunicacions turca on es pot veure clarament que la gran majoria de camins més curts passen per un node central a Ankara i que només que s'elimini aquest node la xarxa queda completament destrossada.

La segona xarxa, per la seva banda, es tracta d'una xarxa de telecomunicacions pakistanesa, però no s'ha mostrat amb el Google Maps per sota perquè alguns dels nodes, al tractar-se d'universitats o institucions similars associades

a una ciutat, no tenen la informació de la latitud i la longitud i no es veia correctament.

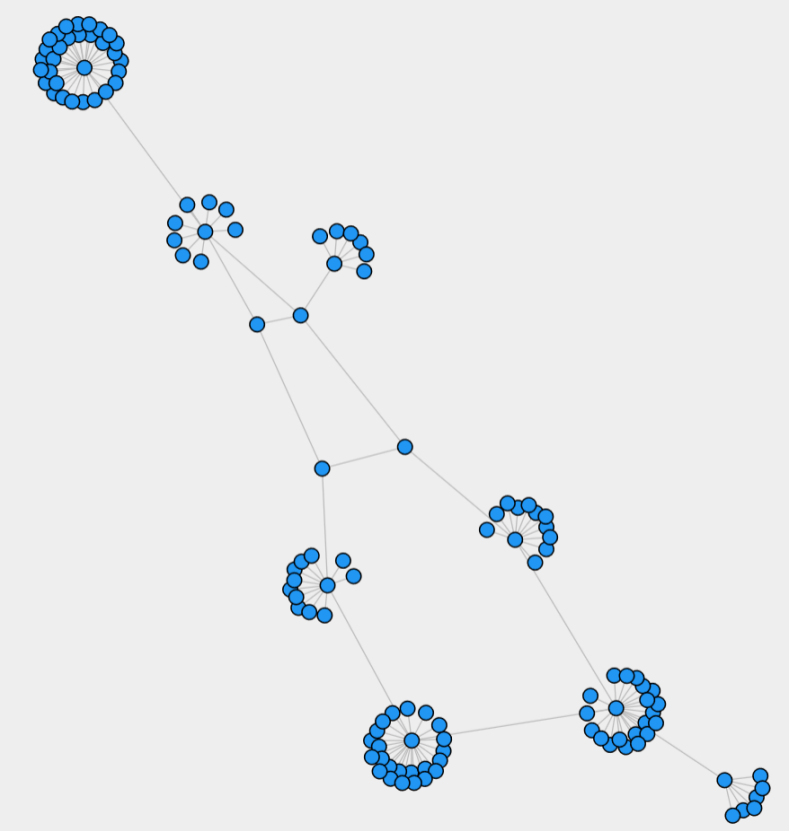


Figura 7.9: Visualització de la xarxa Pern amb el NRS.

Donada la tipologia de les dues xarxes també es pot esperar que la robustesa de la primera caigui de seguida, però que la segona ho faci més lentament, almenys comparada amb Ulaknet, fins que s'han eliminat tots els nodes centrals. Si es repeteixen els gràfics per robustesa respecte al percentatge d'atac que ja s'han vist a l'Estudi Previ, es pot comprovar que és així a la figura 7.10. Tanmateix, no deixa de ser una curiositat.

Finalment, s'ha decidit etiquetar les xarxes utilitzant la robustesa a $P = 10$ i discriminant amb els diferents quartils perquè, com s'ha vist a la figura 7.7, el rang dels valors ocupa gairebé tot el possible i no es detecten *outliers*. Per tant, les xarxes menys robustes tenen una etiqueta 0 i les més, una etiqueta 3.

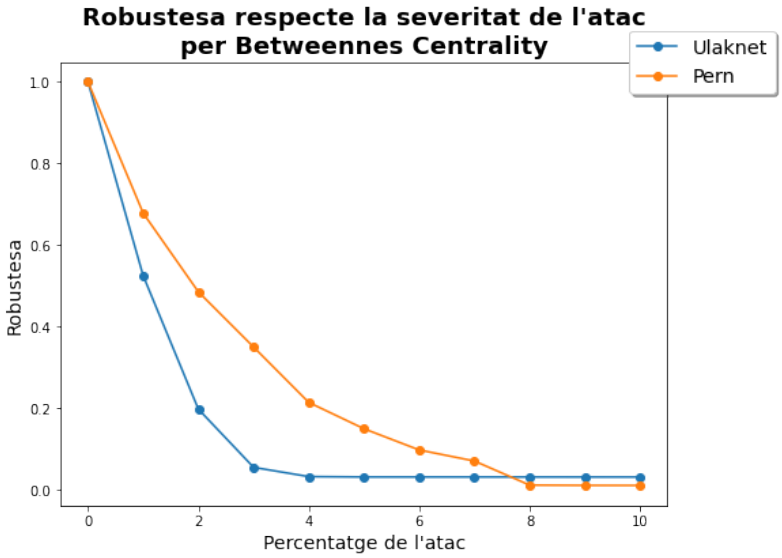


Figura 7.10: Comparació de la robustesa per *Betweennes Centrality* per les xarxes Ulaknet i Pern.

7.2.2 Mètriques rellevants

Hi ha certes mètriques que, a l'autor, li resulten particularment interessants a l'hora d'estudiar la robustesa de les xarxes davant d'atacs de *Betweennes Centrality*. Aquestes són:

- *Heterogenity*: indica quan de regular és una xarxa. Un valor més petit significa que la xarxa és més regular. Un graf en què tots els seus nodes tenen el mateix grau té una heterogeneïtat de 0. Aquesta mètrica també és la que té una correlació més forta amb l'etiqueta.
- *Effective Resistance*: en aquesta mètrica es considera el graf com un circuit elèctric en el qual cada aresta és una resistència d'1Ω. Per tant, una ER més petita és millor perquè significa que hi ha més camins entre parelles de nodes i que aquests són més curts. Un problema d'aquesta mètrica és que escala amb el nombre de nodes pel que s'hauria de normalitzar abans.
- *Clustering Coefficient*: la probabilitat de què els nodes adjacents a un node particular estiguin connectats entre ells. Rang: [0, 1].
- *Assortativity Coefficient*: la preferència dels nodes de connectar-se amb nodes que tenen un grau nodal similar. Rang: [-1, 1].
- *Node Betweenness Centrality*: aquesta mètrica potser és la més directa amb l'etiqueta, ja que diu quina és la mitjana dels camins mínims que passa per cada node.

Xarxa	N. of Nodes	N. of Edges	Avg. Nodal Degree	Heterogenity
VtlWavenet2008	88	92	2.09	0.156
VtlWavenet2011	92	96	2.08	0.168
UsCarrier	158	189	2.39	0.342
Kdl	754	899	2.38	0.358
Cogentco	197	245	2.48	0.427

Taula 7.5: Les cinc xarxes menys robustes amb millor heterogeneïtat

Xarxa	N. of Nodes	N. of Edges	Avg. Nodal Degree	Clustering Coefficient
Belnet2005	23	44	3.82	0.685
Belnet2006	23	44	3.82	0.685
Belnet2003	23	43	3.73	0.592
Belnet2004	23	43	3.73	0.592
Cernet	41	59	2.87	0.442

Taula 7.6: Les cinc xarxes menys robustes amb millor *Clustering Coefficient*

- *Closeness Centrality*: mesura com és de proper un node als altres en mitjana considerant els camins més curts. Rang: [0, 2].

Si es realitza els diagrames de caixa de les mètriques anteriors respecte de l'etiqueta, excepte per l'*Effective Resistance* perquè s'ha d'escalar, es pot observar a la figura 7.11 que encara que sí que sembla que hi ha una certa correlació entre les mètriques destacades i la robustesa categoritzada, aquesta està lluny de ser conclusiva per si soles.

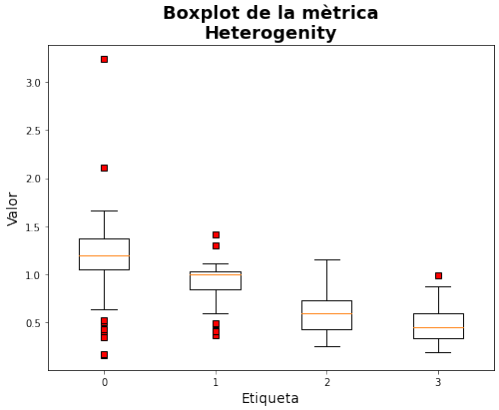
A més a més, tant a la figura 7.11a com a la figura 7.11b, per les mètriques d'*Hetoregenity* i *Clustering Coefficient* respectivament, es poden identificar nombrosos *outliers* en les etiquetes més baixes. Si s'explora el *dataset* per descobrir quines són aquestes s'obtenen les taules 7.5 i 7.6.

Sembla que els *outliers*, tal com es pot veure a la taula 7.5, que destaquen per la seva bona *Heterogenity* i la seva mala robustesa són les xarxes més grans del *dataset*.

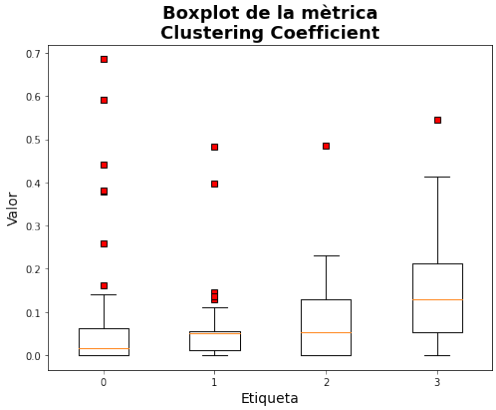
Una possible explicació a aquest fenomen és que encara que són xarxes força regulars, cada node està connectat a una part molt petita de la xarxa, és a dir, el seu *Average Nodal Degree* és minúscul comparat amb l'ordre de la xarxa, pel que són relativament fàcils de trencar.

Pels *outliers* del *Clustering Coefficient* costa de trobar una explicació, almenys amb les mètriques que s'ha escollit mostrar a la taula 7.6, però si s'utilitza el NRS per visualitzar la xarxa Belnet2005 la causa de la seva mala robustesa es veu de forma gairebé immediata.

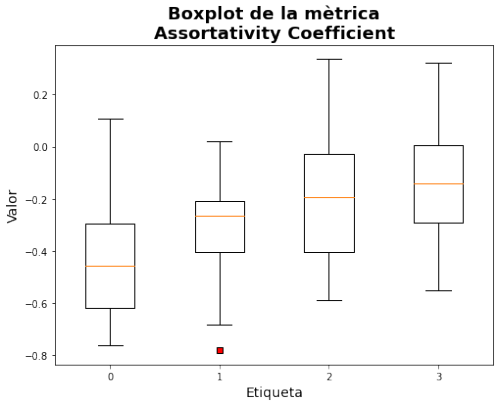
Tal com es pot observar a la figura 7.12, l'estructura de la xarxa Belnet2005



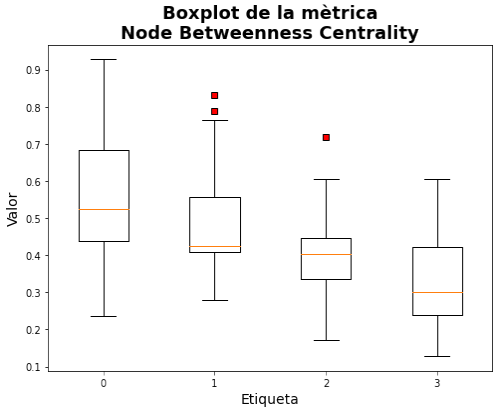
(a) Heterogenity



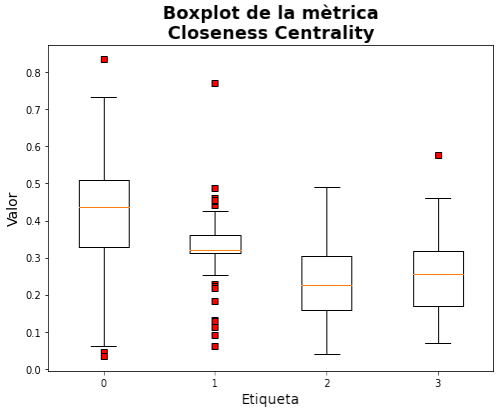
(b) Clustering Coefficient



(c) Assortavity Coefficient



(d) Node Betweenness Centrality



(e) Closeness Centrality

Figura 7.11: Diagrames de caixa de les mètriques rellevants respecte l'etiqueta de les xarxes.

és bàsicament una estrella amb dos nodes centrals pel que la probabilitat de què els nodes adjacents a un node particular estiguin connectats entre ells és molt alta, però, si s'eliminen aquests dos nodes, la xarxa queda completament destrossada.

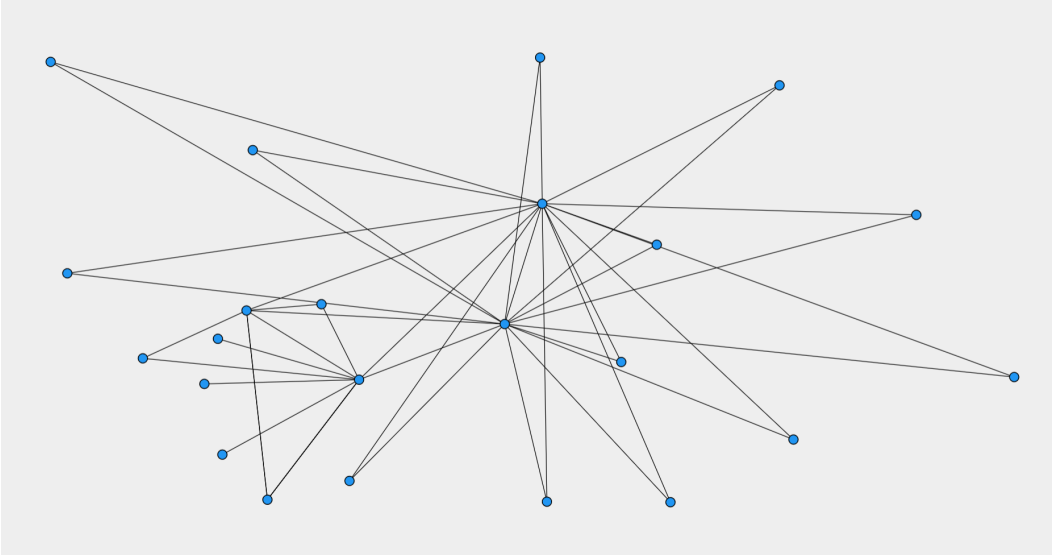


Figura 7.12: Visualització de la xarxa Belnet2005 amb el NRS.

De les xarxes restants només s'ha mostrat Cernet, a la figura 7.13, perquè les altres xarxes són versions lleugerament diferents de la Belnet2005 pel que la tipologia ha de ser molt similar.

La Cernet pateix en menor mesura el mateix problema: té diversos nodes centrals a Beijing, Xi'An i Shanghai, però el valor del *Clustering Coefficient* és alt perquè els nodes dels extrems tenen la tendència a estar connectats entre ells.

7.2.3 Mètriques recomanades

Si ara es comparen amb les mètriques recomanades per calcular la robustesa [Marzo 2018], explicades al capítol 3:

- *Average Nodal Degree*: el grau nodal mitjà dels nodes. Com més gran millor.
- *Efficiency*: la mitjana de la inversa de la llargada dels camins mínims entre tots els nodes.
- *Largest Eigenvalue*: el valor propi més gran de la matriu d'adjacència. Com més gran més robusta és la xarxa.
- *Algebraic Connectivity*: mesura com de difícil és de trencar un graf. Valors més grans indiquen una millor robustesa.

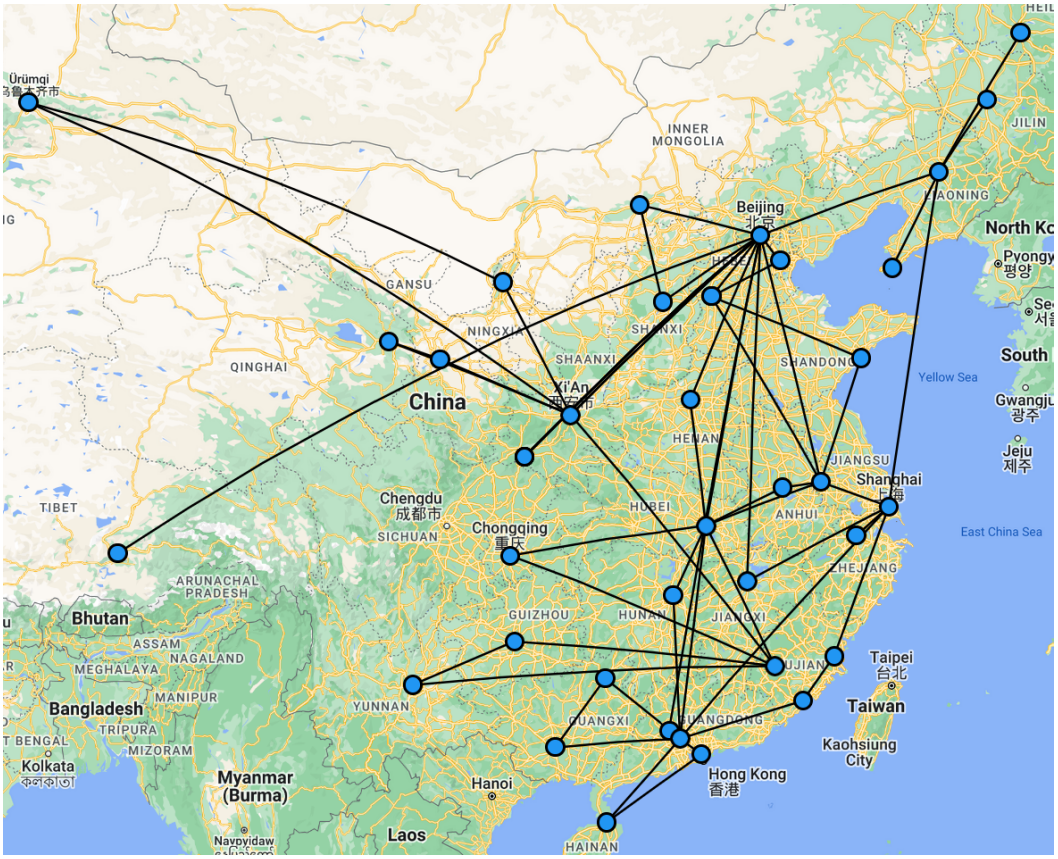


Figura 7.13: Visualització de la xarxa Cernet amb el NRS.

- *Natural Connectivity*: aquesta mètrica està basada en el nombre de camins tancats, *closed walks* en anglès, d'un graf. Només es pot fer servir per comparar entre xarxes quan tenen el mateix *Algebraic Connectivity*.
- *Edge Betweenness Centrality*: el nombre de camins mínims mitjans que passen per cada enllaç.

Les tres mètriques recomanades que falten: *Clustering Coefficient* ja s'ha vist a la subsecció anterior com a rellevant; i *Largest Connected Component* i *Average Two Terminal Reliability* s'eliminen a la secció següent perquè la primera és igual al nombre de nodes i la segona sempre és 1.

A la figura 7.15 es pot veure que encara que aquestes mètriques són molt útils a l'hora de calcular la robustesa d'una xarxa, per la majoria no hi ha una diferència significativa entre etiquetes.

Només al diagrama de caixes de l'*Average Nodal Degree*, 7.15a, es veu una diferència important entre etiquetes: les xarxes més robustes tenen un grau nodal mitjà més gran com s'esperaria. Encara que hi ha diversos *outliers* per l'etiqueta 0 i 1 amb un bon grau nodal mitjà, aquestes xarxes corresponen a variacions

de la Belnet2005 que ja s'ha vist a la figura 7.12 i per la que ja s'ha trobat una explicació de la seva mala robustesa.

Una de les mètriques recomanades, fins i tot, té un comportament completament contrari a l'esperat d'ella: els valors mitjans del *Largest Eigenvalue*, 7.15c, tenen un valor inferior per les xarxes més robustes, però això segurament es deu que són xarxes més petites.

A la figura 7.14 es pot veure un diagrama de caixes del *Number of Nodes* de les xarxes sense *outliers* perquè sinó la xarxa Kdl domina completament el gràfic.

Encara que la majoria d'aquestes mètriques no donen informació suficient per distingir com de bona és la robustesa de la xarxa s'ha decidit deixar-las perquè el més probable és que es complementin amb altres.

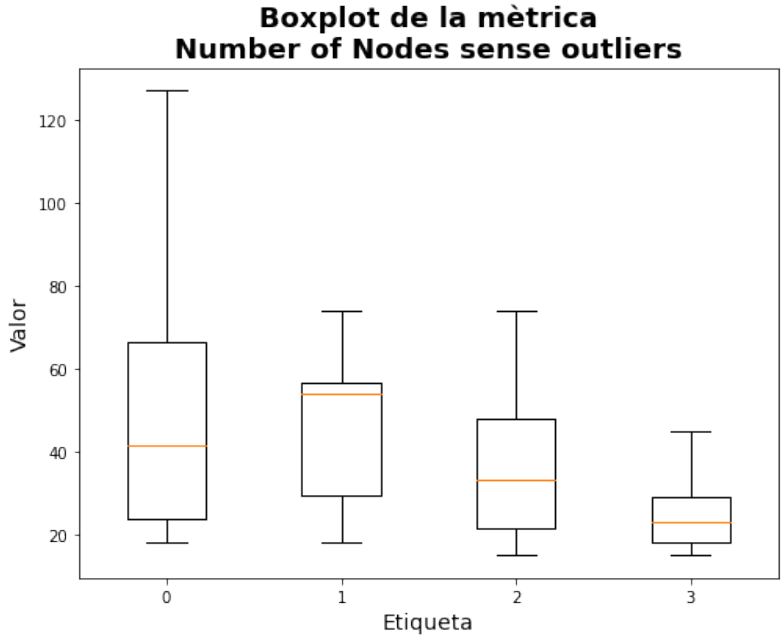
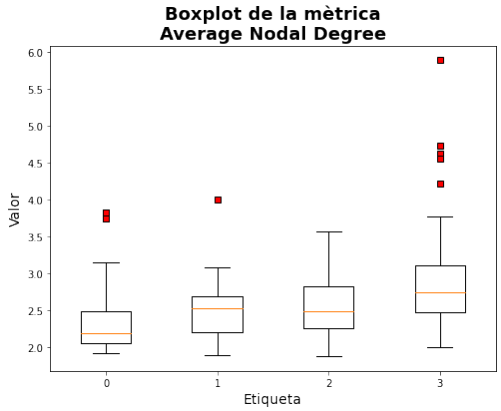
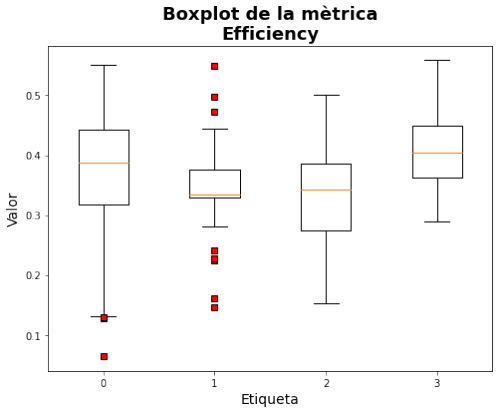


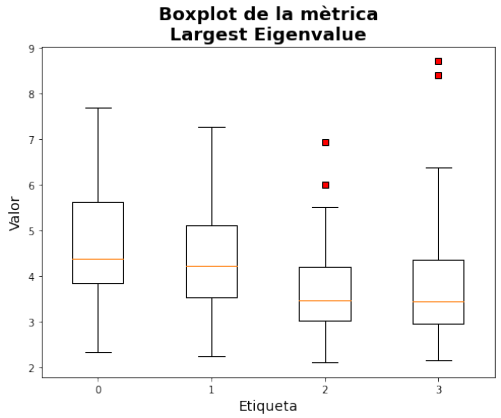
Figura 7.14: Diagrama de caixes del *Number of Nodes* sense *outliers*.



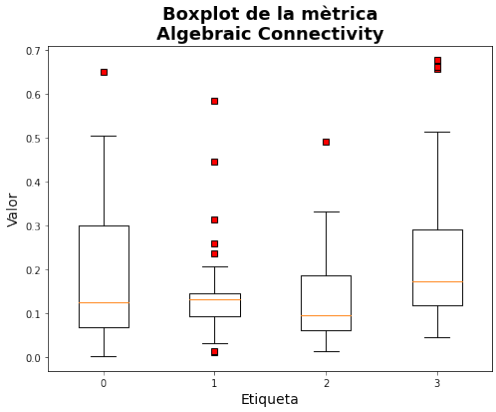
(a) Average Nodal Degree



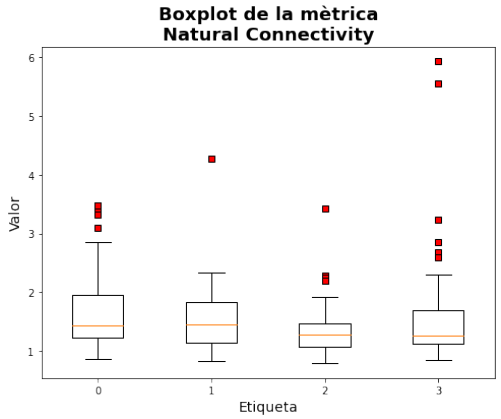
(b) Efficiency



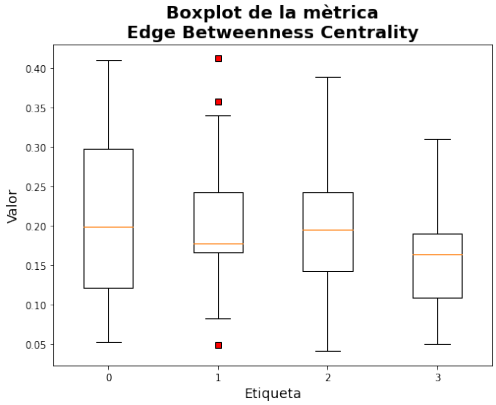
(c) Largest Eigenvalue



(d) Algebraic Connectivity



(e) Natural Connectivity



(f) Edge Betweenness Centrality

Figura 7.15: Diagrames de caixa de les mètriques rellevants respecte l'etiqueta de les xarxes.

7.3 Selecció de les mètriques

Un primer filtre sobre el conjunt de mètriques inicials és eliminar aquelles que són constants per la imposició que s'ha fet sobre la selecció de les xarxes del Topology Zoo: han de ser xarxes completament connectades.

Les columnes que són constants per totes les xarxes són: *Fractional Size Largest Component*, *Average Two Terminal Reliability*, *Lower bound Average Two Terminal Reliability*, *Upper Bound Average Two Terminal Reliability* i *Degree of Fragmentation*. Una explicació més detallada d'aquestes mètriques es pot trobar a la subsecció 4.2.

Dues altres mètriques que també poden ser eliminades són *Largest Connected Component* perquè el seu valor sempre és igual a *Number of Nodes* i *Analytical Approach Largest Connected Component* perquè, tal com el nom diu, és una aproximació analítica al component més gran connectat a partir del vector de graus nodals del graf i que encara està en una etapa experimental.

Encara que algunes d'aquestes mètriques han estat essencials per calcular la robustesa de la xarxa, com ara el *Average Two Terminal Reliability*, la probabilitat de què dos nodes qualssevol estiguin connectades, o *Largest Connected Component*, la mida del component més gran, en aquest punt de la tesi no donen cap informació significativa.

Després d'aquestes set primeres eliminacions s'ha fet la matriu de correlacions de la figura 7.16 amb les vint-i-tres mètriques restants.

El primer que es pot observar és que hi ha certes mètriques positivament correlacionades amb el *Number of Nodes*, és a dir, que escalen amb la mida de la xarxa: el *Number of Edges*, *Average Shortest Path Length*, *Diameter*, *Effective Resistance*, *Number of Spanning Trees* i *Maximum Edge Betweenness*.

D'aquestes s'ha decidit eliminar en primera instància el *Number of Edges*, és una dada recuperable a partir del *Number of Nodes* i l'*Average Nodal Degree*, i el *Diameter*, ja que està correlacionada amb *Average Shortest Path Length*.

Tanmateix, si es vol aprofitar les mètriques d'*Effective Resistance*, *Number of Spanning Trees* i *Maximum Edge Betweenness* primer s'han de tractar, però això es fa a la subsecció 7.4

Altres parelles de variables que destaquen per la seva forta correlació són: *Maximum Nodal Degree* i *Heterogenity*; *Vertex Connectivity* i *Edge Connectivity*; *Largest Eigenvalue* i *Natural Connectivity*; i *Degree Centrality* i *Closeness Centrality*.

D'aquestes parelles s'ha eliminat les mètriques següents:

- *Maximum Nodal Degree*: *Heterogenity* dona una visió més global de la xarxa i és una de les mètriques considerades rellevants per l'autor, explicat a la subsecció 7.2.2.

- *Edge Connectivity*: els atacs que s'han realitzat són per nodes.
- *Degree Centrality*: *Closeness Centrality* és també una de les mètriques considerades rellevants.
- *Natural Connectivity*: també està fortament correlacionada amb *Average Nodal Degree*.

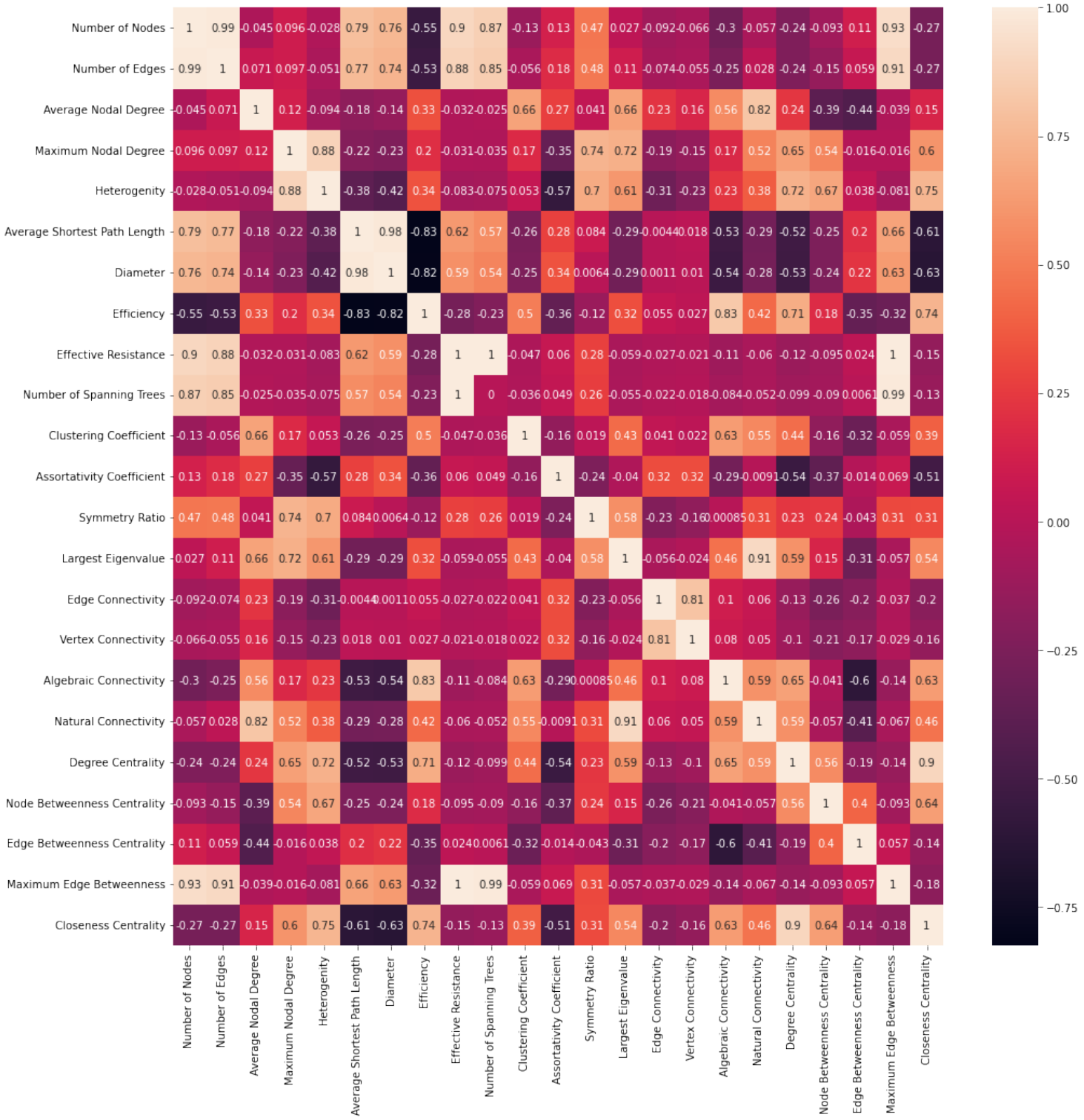


Figura 7.16: Matriu de correlacions amb les mètriques del NRS després del primer filtre.

7.4 Enginyeria de característiques

Després de la selecció de característiques de la secció anterior queden les disset mètriques següents: *Number of Nodes*, *Average Nodal Degree*, *Hetoregenity*, *Average Shorest Path Length*, *Efficiency*, *Effective Resistance*, *Number of Spanning Trees*, *Clustering Coefficient*, *Assortavity Coefficient*, *Symmetry Ratio*, *Largest Eigenvalue*, *Vertex Connectivity*, *Algebraic Connectivity*, *Node Betweenness Centrality*, *Edge Betweenness Centrality*, *Maximum Edge Betweenness* i *Closeness Centrality*.

D'aquestes, però, hi ha tres que tenen obvis problemes, com ja s'ha mencionat: *Effective Resistance*, *Number of Spanning Trees* i *Maximum Edge Betweenness*. A la figura 7.17 es pot veure que hi ha un *outlier* que domina les tres mètriques. Aquest *outlier* és la xarxa Kdl, la més gran del *dataset* amb gairebé vuit-cents nodes.

Tanmateix, per dues d'aquestes tres mètriques la solució és relativament senzilla, ja que el valor d'*Effective Resistance* i de *Maximum Edge Betweenness* pot ser escalat pel nombre de parelles del graf. És a dir, per $n(n-1)/2$ on n és el nombre de nodes.

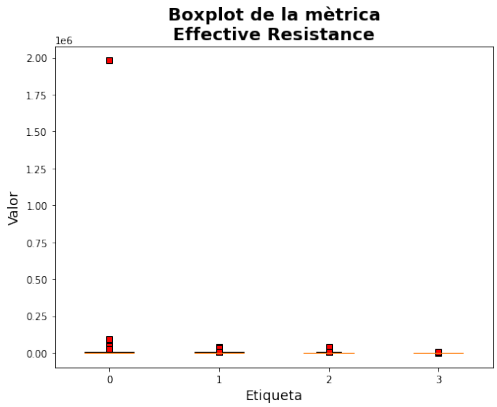
Això és així perquè el nombre de parelles d'un graf també és el nombre total de camins mínims possibles i perquè l'*Effective Resistance*, tal com s'ha explicat a la taula 4.1, és la suma de la resistència efectiva de tots les parelles de vèrtexs.

Per tant, aquestes dos noves mètriques escalades signifiquen: 1) la resistència efectiva mitjana per cada enllaç; i 2) la fracció màxima de camins mínims que passa per un enllaç. Els nous diagrames de caixes d'aquestes dues mètriques es poden veure a la figura 7.18 i les mètriques de les xarxes que destaquen com *outliers* a la figura 7.18a, a la taula 7.7: la xarxa Kdl continua essent un *outlier*, però no d'una forma tan extrema.

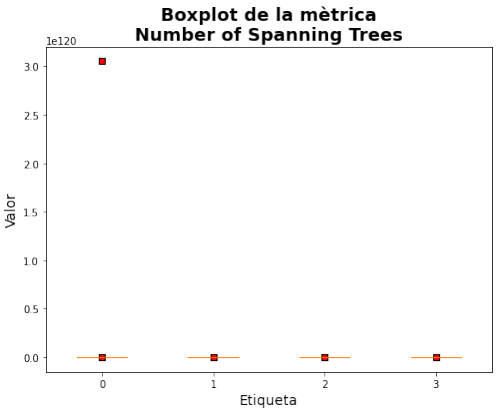
Si s'utilitza el NRS per visualitzar, figura 7.19 la xarxa Syringa es pot entendre perfectament perquè té la resistència efectiva escalada més gran del *dataset*: és bàsicament un bus i un anell units. Per la seva banda, la xarxa VtlWavenet2011, figura 7.20, són diversos anells units.

Xarxa	N. of Nodes	N. of Edges	Avg. Nodal Degree	S. Eff. Resistance
Syringa	74	74	2.0	2.242
VtlWavenet2011	92	96	2.08	1.883
VtlWavenet2008	88	92	2.09	1.839
Kdl	754	899	2.38	1.744
RedBestel	84	101	2.40	1.693

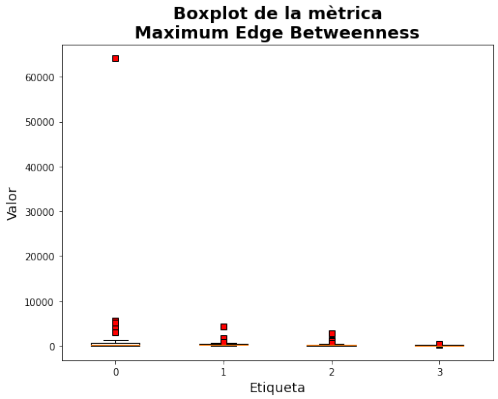
Taula 7.7: Les cinc xarxes menys robustes amb pitjor *Scaled Effective Resistance*.



(a) Effective Resistance

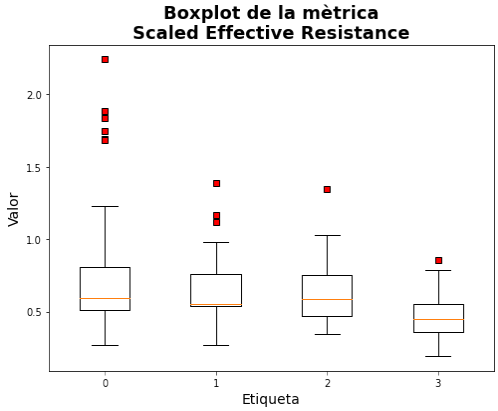


(b) Number of Spanning Trees

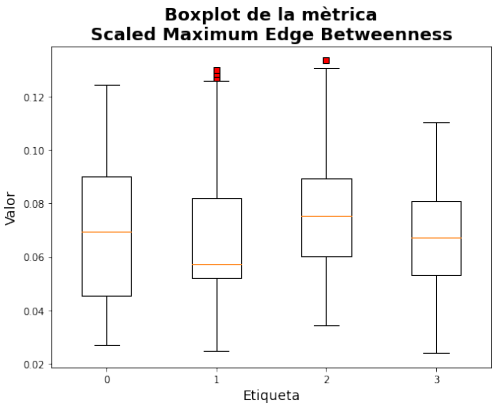


(c) Maximum Edge Betweenness

Figura 7.17: Diagrames de caixa de les tres mètriques problemàtiques.



(a) Scaled Effective Resistance



(b) Scaled Maximum Edge Betweenness

Figura 7.18: Diagrames de caixa de les dues mètriques arreglades.

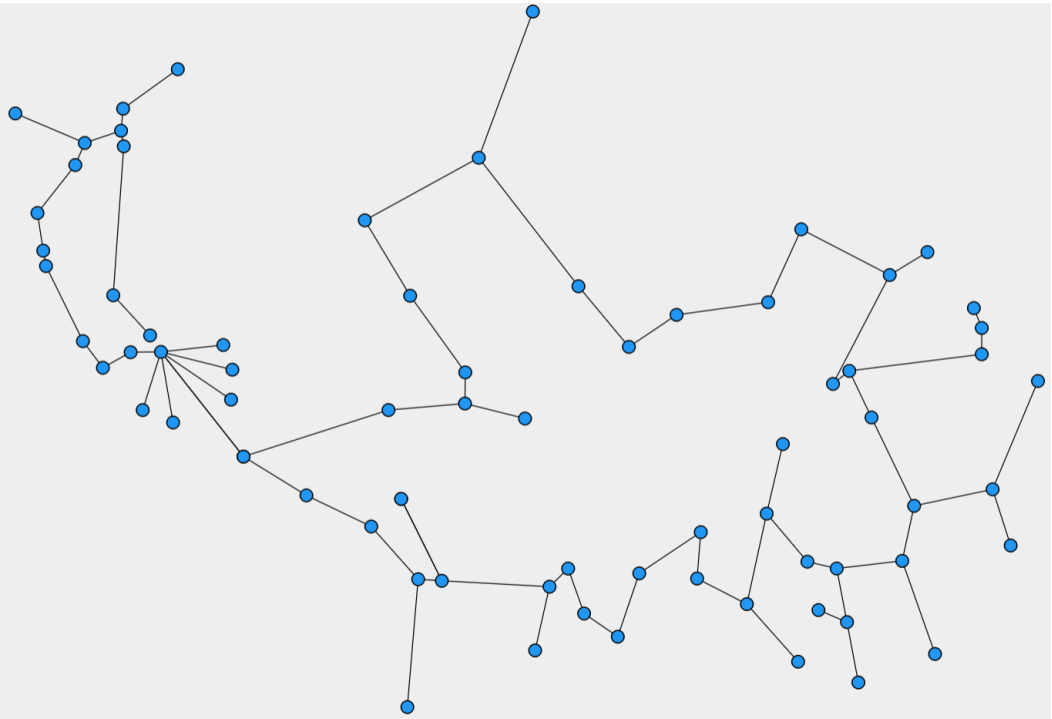
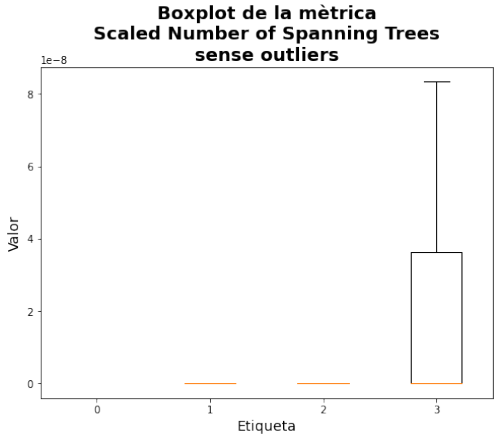


Figura 7.19: Visualització de la xarxa Syringa amb el NRS.

Figura 7.20: Visualització de la xarxa VtlWavenet2011 amb el NRS.



(a) Diagrama de caixes sense outliers

label	0	1	2	3
count	5.200000e+01	5.100000e+01	5.100000e+01	5.100000e+01
mean	4.106003e-19	2.653849e-19	1.140370e-15	3.383268e-13
std	2.959854e-18	1.894047e-18	7.983776e-15	2.179066e-12
min	0.000000e+00	0.000000e+00	0.000000e+00	1.869468e-107
25%	4.489128e-111	5.290069e-87	3.776935e-68	3.073071e-36
50%	6.331411e-57	5.448612e-83	1.926899e-41	3.141349e-23
75%	3.051712e-29	7.130951e-38	4.787532e-23	7.200189e-17
max	2.134396e-17	1.352637e-17	5.703437e-14	1.553579e-11

(b) Descripció estadística de la mètrica

Figura 7.21: Figures relacionades amb la mètrica *Scaled Number of Spanning Trees*.

Xarxa	N. of Nodes	N. of Edges	Avg. Nodal Degree	S. N. of Spanning Trees
Highwinds	18	53	5.88	1.55e-11
Airtel	16	37	4.62	1.359e-12
Internetmci	19	45	4.73	1.790e-13
Goodnet	17	31	3.64	1.076e-13
Claranet	15	18	2.4	5.703e-14

Taula 7.8: Les cinc xarxes amb millor *Scaled Number of Spanning Trees*.

El *Number of Spanning Trees* és una mètrica molt interessant també perquè, com ja s'ha explicat a la taula 4.1, un *spanning tree* és un subgraf que conté $N-1$ enllaços, tots els nodes i sense cicles. Tanmateix, no és una mètrica tan fàcil d'escalar com les altres.

Una opció és utilitzar el nombre de *spanning trees* d'un graf complet de N nodes com el valor màxim possible per cada xarxa. Aquesta aproximació, però, resulta en valors molt propers al zero perquè el *Number of Spanning Trees* d'un graf complet és: N^{N-2}. A més, els valors de la *Scaled Number of Spanning Trees* es dispersen molt pel que cap dels escalats de scikit-learn funcionen correctament.

Això és una llàstima perquè com es pot veure a les figures 7.21a i 7.21b amb la descripció estadística de la mètrica sembla que hi ha una diferència significativa entre les etiquetes de la robustesa.

Si s'ordenen les xarxes per aquesta nova mètrica escalada i es mira la taula resultant, 7.8, de seguida es pot veure que les xarxes que destaquen són relativament petites, però molt ben connectades.

Tanmateix, no hi ha hagut un altra opció que eliminar les tres mètriques originals i la mètrica escalada per *Number of Spanning Trees* deixant el *dataset* amb només setze mètriques.

7.5 Modelització

Per a la modelització, primer s'ha provat l'algorisme UMAP de forma no supervisada, és a dir, sense utilitzar les etiquetes de robustesa de la xarxa, però amb les dades ja seleccionades i processades d'acord amb les seccions 7.3 i 7.4.

Aquest primer intent es pot veure a la figura 7.22, però encara que es comencen a formar clústers ben definits amb predominança de certs colors, per exemple el clúster inferior amb el taronja, el central amb el vermell o el superior amb el blau, el resultat està lluny de ser el desitjat.

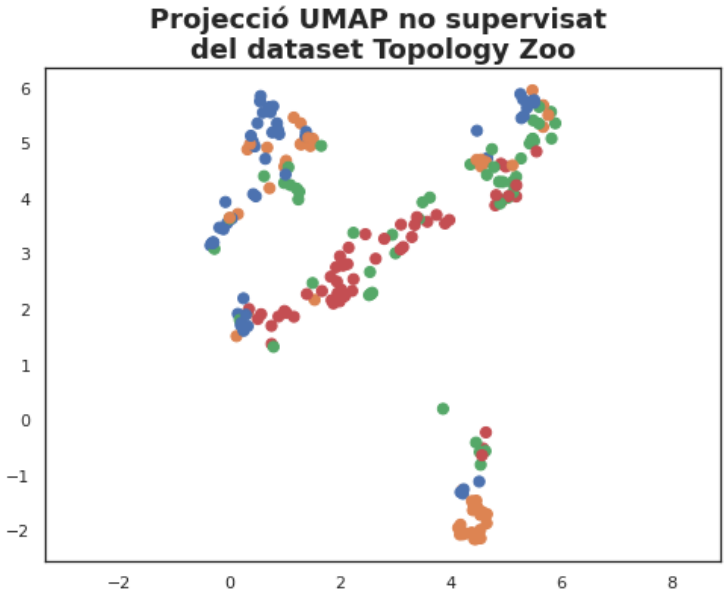


Figura 7.22: Reducció de la dimensionalitat amb UMAP no supervisat.

Per aquest motiu, el següent pas ha estat utilitzar l'UMAP semisupervisat amb les dades separades en *train* i *test* en una proporció del 80% - 20%. El resultat d'aquesta nova reducció del *dataset* ha resultat més exitosa amb una clara distinció de les mostres de *train*, figura 7.23, encara que les dades amb una etiqueta d'1 s'han separat en dos clústers el que posteriorment ha donat problemes a l'hora de fer servir l'HDBSCAN.

Aquesta separació no és tan clara amb les dades de *test*, figura 7.24, però sí que es pot identificar que la majoria de mostres estan ben agrupades i, aquelles que no, rarament es barregen amb les dades en el que la seva etiqueta és

diferent per un factor de dos o més; és a dir, que una xarxa molt poc robusta, etiqueta 0, no es confon amb una xarxa robusta o molt robusta, etiqueta 2 o 3 respectivament.

Per trobar els millors valors dels paràmetres de l'UMAP, explicats a la subsecció 2.3, s'ha realitzat una cerca exhaustiva i els candidats finals han estat:

- *n_neighbors*= 25. Així es manté una visió més local de les dades.
- *min_dist*= 0. Perquè ens interessa formar clústers ben definits.
- *metric*= *canberra*.

Com es pot esperar després de veure la reducció de la dimensionalitat a la figura 7.23, l'HDBSCAN no ha tingut cap problema a l'hora de clusteritzar les dades que s'han vist a la figura 7.25, excepte que ha detectat una com a soroll i que ha separat els dos clústers de l'etiqueta 1 en dos.

Per aquest motiu, s'ha realitzat el posttractament següent pels clústers obtinguts:

- S'han modificat aquelles mostres que pertanyen al clúster 3 perquè pertanyin al 0.
- S'ha canviat el clúster 4 perquè sigui el 3.

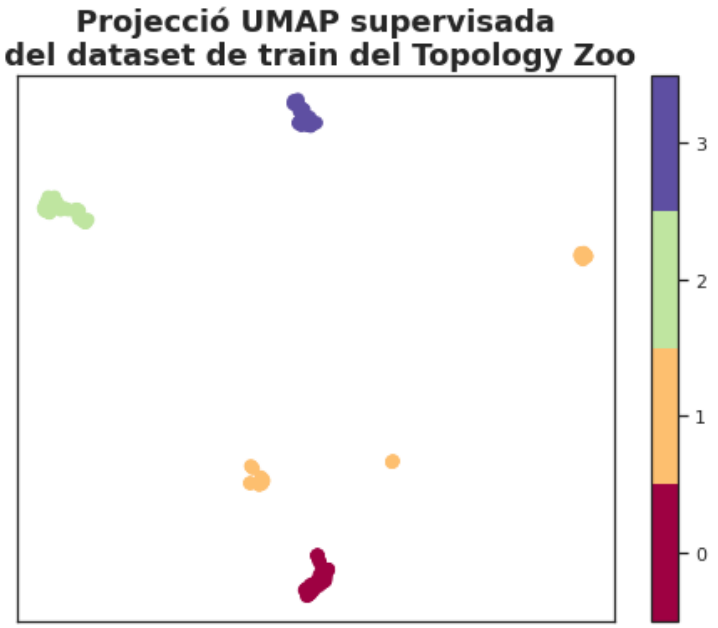


Figura 7.23: Reducció de la dimensionalitat amb les dades de *train* i UMAP semisupervisat.

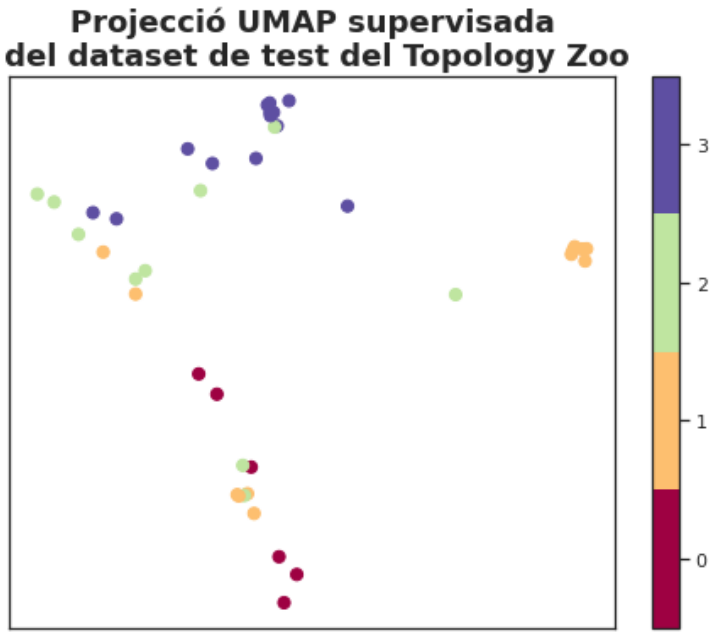


Figura 7.24: Reducció de la dimensionalitat amb les dades de *test* i UMAP semi-supervisat.

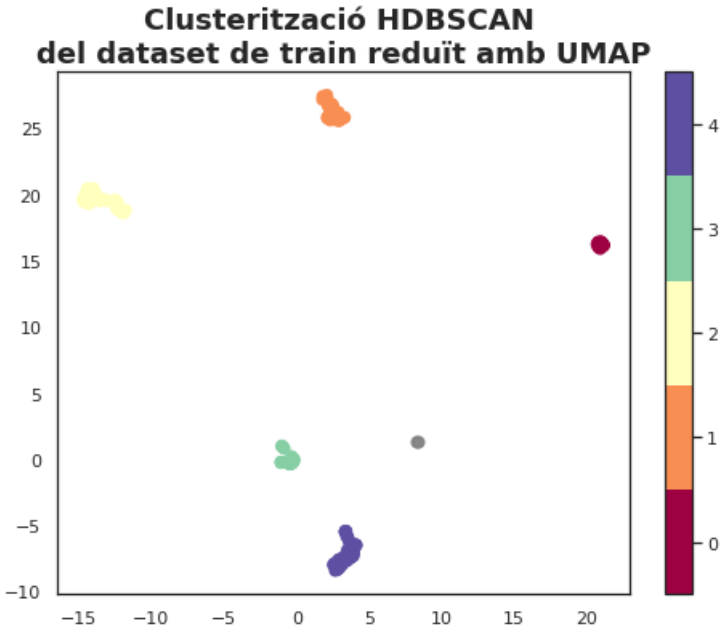


Figura 7.25: Clusterització de les dades reduïdes de *train* sense tractar.

Confiança	% Mostres	Rand Score	Adj. Rand Score	Adj. Mutual Info Score
0.5	26.8	1.0	1.0	1.0
0.45	51.2	0.95	0.88	0.87
0.4	53.7	0.91	0.78	0.76
0.3	63.4	0.81	0.55	0.54
0	95.1	0.78	0.44	0.46

Taula 7.9: Precisió de l'algorisme HDBSCAN per les dades reduides de *test*.

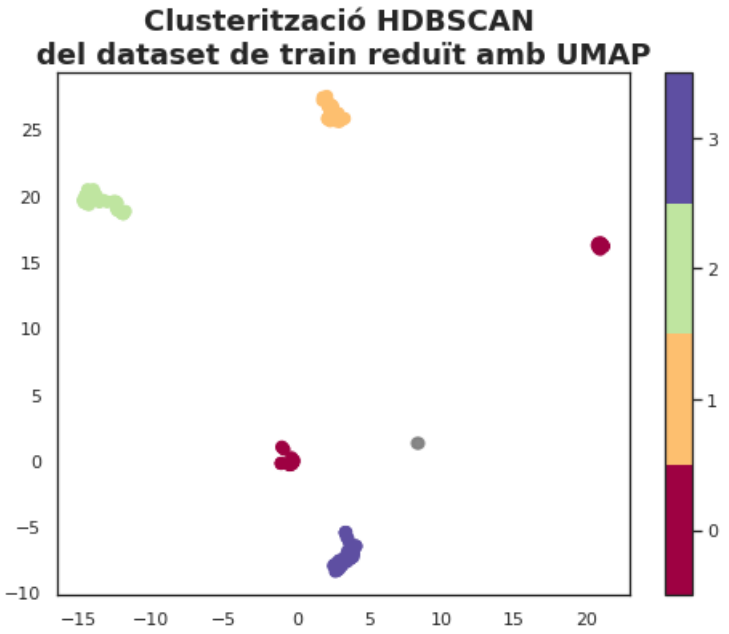


Figura 7.26: Clusterització de les dades reduïdes de *train* tractades.

Una cosa que s'ha de tenir en compte, però, és que l'HDBSCAN ha recodificat també els clústers pel que l'etiqueta 0 de l'UMAP no correspon amb l'etiqueta 0 de l'HDBSCAN sinó amb la 3 de la figura 7.26 ja tractada.

Aquest procés s'ha repetit per les dades de *test* i a les figures 7.27 i 7.28 es pot veure el resultat obtingut sense tractar i una vegada tractades respectivament.

Una fortalesa, que ja s'ha mencionat a la subsecció 2.3, de l'HDBSCAN davant de mètodes similars és que juntament amb la predicció de quin clúster pertany la dada retorna també quina és la seva confiança en la predicció.

A la taula 7.9 es pot veure com la precisió de l'algorisme millora significativament amb les confiances més altes, però que a mesura que disminueix la confiança de l'HDBSCAN les mesures de similitud *Adjusted Rand Score* i *Adjusted Mutual Information Score* penalitzen molt la incorrecta classificació de les mostres. Tanmateix, per un 51,2% de les mostres, el resultat és excel·lent.

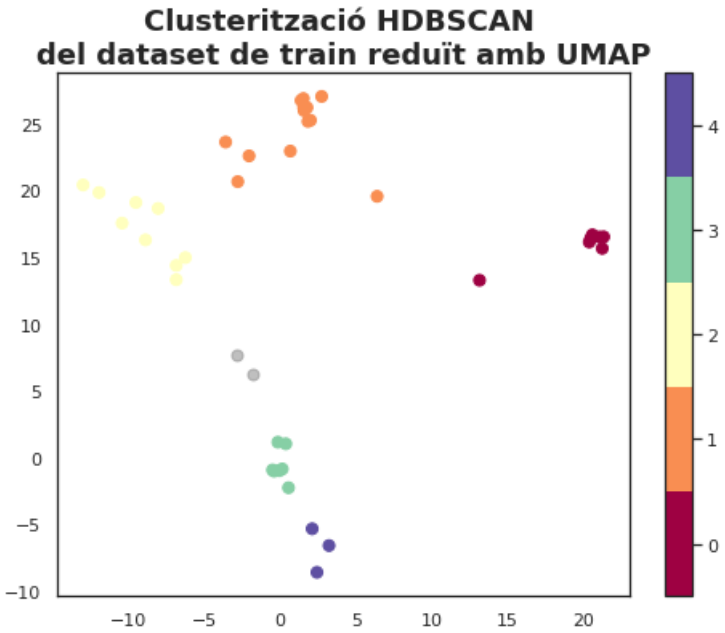


Figura 7.27: Clusterització de les dades reduïdes de *test* sense tractar.

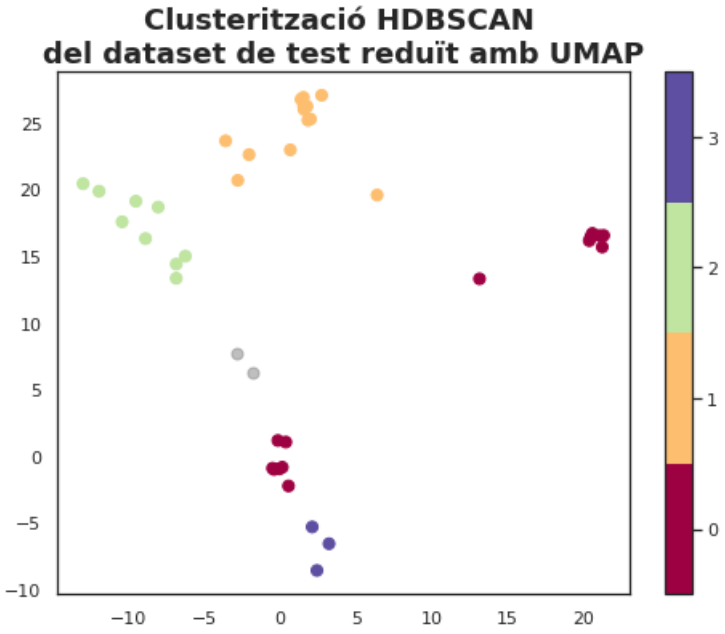


Figura 7.28: Clusterització de les dades reduïdes de *test* tractades.

7.6 Generació de noves xarxes

Una forma amb la qual encara es pot intentar millorar el resultat de la clusterització amb UMAP i HDBSCAN és augmentant les dades de les quals es disposen. Això es pot fer de diverses formes: trobant noves xarxes de telecomunicacions en línia, afegint noves tipologies al *dataset* o generant xarxes sintètiques que tinguin unes característiques similars.

S'ha decidit per aquesta última opció perquè el NRS2 ja disposa d'un generador de xarxes: concretament, s'utilitzarà l'histograma de graus nodals, al qual se li aplicaran diverses modificacions, de xarxes ja existents al *dataset* per crear-ne de noves.

Aquestes modificacions tant poden ser afegir o treure nodes com afegir o treure arestes. Fins i tot si es generen dues xarxes amb el mateix histograma del grau nodal és més que probable que el resultat tant de l'experiment com de les mètriques que es calculen de la xarxa inicial siguin diferent, ja que el nombre de possibles permutacions és molt gran.

Per exemple, en l'histograma del grau nodal següent [0,2,1,1,1] hi ha 0 nodes amb grau nodal 0, 2 nodes amb grau nodal 1, 1 node amb grau 2, 1 node amb grau nodal 3 i, finalment, 1 node amb grau nodal 4. Aquest histograma, però, no és correcte perquè la suma dels graus nodals ha de ser parell i $2*1+1*2+1*3+1*4=11$. Si s'afegeix un altre node amb un grau nodal senar, o s'afegeix una nova aresta, es podrà generar un graf a partir d'aquest histograma. Un segon problema que pot sorgir a l'hora de crear nous grafs és que siguin inconnexes pel que també s'ha de tenir en compte durant la generació.

Pel nou histograma, [0,2,1,2,1], a la que s'ha afegit un node amb grau nodal 3, el graf obtingut amb el generador del NRS es pot veure a la figura 7.29.

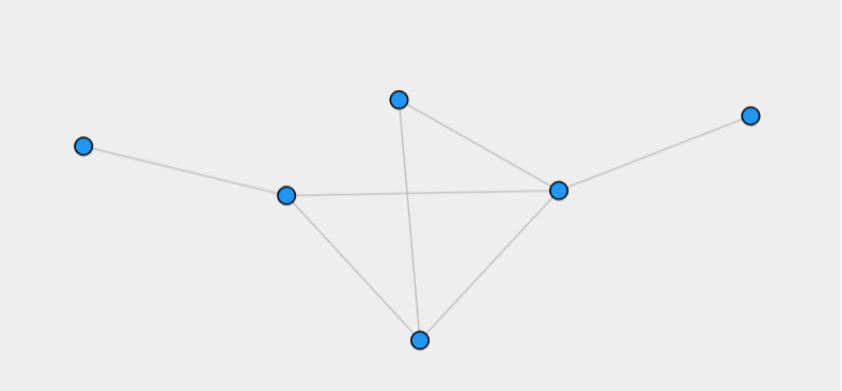


Figura 7.29: Visualització del graf generat amb l'histograma [0,2,1,2,1].

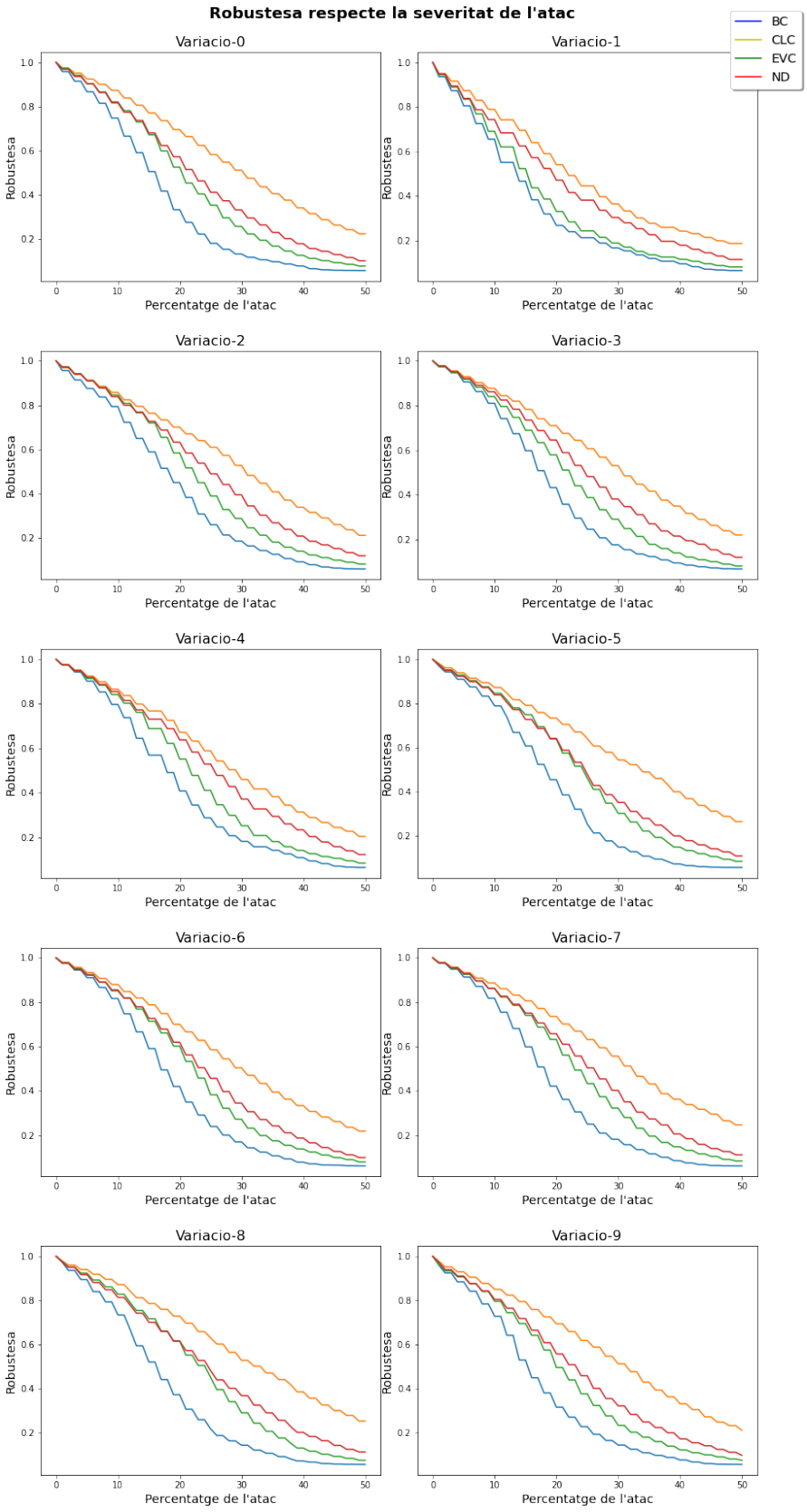


Figura 7.30: Robustesa respecte la severitat de l'atac per deu variacions de la xarxa Surfnet.

Primer s'ha realitzat una prova de concepte en la que s'han generat deu xarxes a partir de la mateixa per demostrar que petits canvis en la tipologia poden modificar la robustesa de la xarxa.

La xarxa que s'ha escollit és Surfnet perquè ja s'ha vist a l'Estudi previ, secció 7.1, és relativament gran, té cinquanta nodes, i una robustesa a $P = 10$ al voltant del 0,7. El seu histograma del grau nodal és [0, 2, 28, 8, 7, 2, 1, 0, 0, 0, 2].

L'evolució de la robustesa pels quatre atacs, *Betweenness Centrality*, *Closeness Centrality*, *Eigenvector Centrality* i *Nodal Degree*, i les deu variacions de la xarxa es pot veure a la figura 7.30.

Encara que la diferència és significativa per *Eigenvector Centrality* i per *Nodal Degree*, no és així per *Betweenness Centrality*, que és l'atac que interessa per l'experiment, i per *Closeness Centrality*.

Això és degut, probablement, que la generació de noves xarxes està lluny de ser perfecte. Alguns dels seus defectes són: no es poden afegir nodes amb un grau nodal superior al màxim original; la probabilitat de què s'afegeixi un node de grau nodal X és la mateixa per tots els graus nodals, excepte per grau nodal igual a zero que la probabilitat també és zero perquè la xarxa estigui inicialment connectada; i el màxim d'elements que es poden afegir o treure són un deu per cent dels originals, però el nombre actual d'elements es decideix amb un aleatori amb el rang $[-0.1N, 0.1N]$, inclòs el zero perquè, com es pot veure a la figura 7.31, fins i tot així el resultat de la robustesa és diferent.

Finalment, s'han generat noranta-cinc xarxes noves a partir de histogrames aleatoris del *dataset* amb entre trenta i seixanta nodes, però l'execució posterior de la *pipeline* amb les dades originals no només no ha millorat el resultat del HDBSCAN, explicat a la secció anterior, sinó que l'ha fet empitjorar. Per aquest motiu s'han descartat les noves xarxes creades.

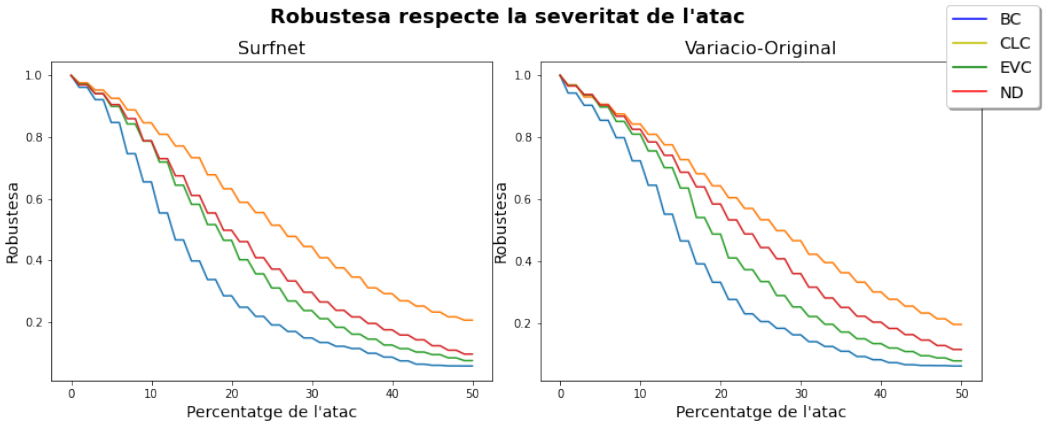


Figura 7.31: Robustesa respecte la severitat de l'atac per Surfnet i una variació sense modificacions.

7.7 Nova Mètrica

Una de les mètriques més interessants que s'ha trobat durant l'exploració inicial ha estat l'*Heterogenity*. Com es pot veure a la figura 7.11a, la diferència entre les etiquetes és significativa i, no només per una o dues caixes, sinó que hi ha una clara distinció entre totes.

Tanmateix, el valor d'aquesta mètrica es veu enfosquida per nombrosos *outliers* per les etiquetes 0 i 1 en el que xarxes amb una molt bona *Heterogenity* tenen una robustesa molt dolenta. A la taula 7.5 es pot veure que aquests *outliers* corresponen a les xarxes més grans del *dataset* pel que encara que són xarxes molt regulars, la seva connectivitat no és prou bona.

Per aquest motiu, s'ha decidit provar una nova mètrica en les que es combinen aquests dos factors per descobrir si són suficients per classificar la robustesa d'una xarxa.

Per calcular aquesta heterogeneïtat ajustada s'ha fet servir la fórmula següent:

$ADJ.HET = HET/(AND/N)$

on HET és el valor de l'*Heterogenity*, AND és l'*Average Nodal Degree* i N és el valor del *Number of Nodes*.

A la figura 7.32 es poden veure els dos diagrames de caixes d'aquesta mètrica nova, amb *outliers* i sense respectivament, i observar una clara diferència entre les mitjanes i la variància dels valors de l'*Adjusted Heterogenity* per cada nivell de la robustesa.

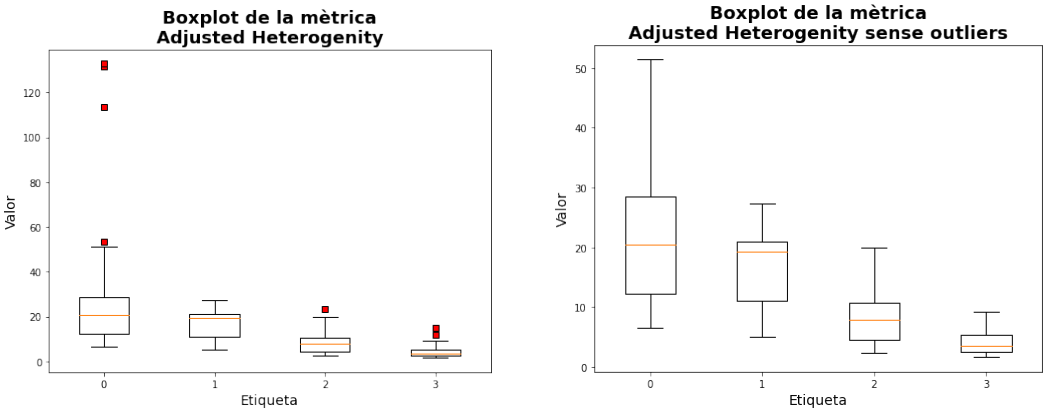
Tanmateix, encara es poden identificar nombrosos *outliers* per sobre del tercer quartil, tant per l'etiqueta 0 com per la 3. D'aquests només són preocupants els segons perquè quan més petita és aquesta nova mètrica, millor hauria de ser la robustesa.

A la figura 7.33, a la que s'han eliminat els valors superiors a cinquanta per millorar la visualització, es pot comprovar que en general és així, però que hi ha certs punts que destaquen per l'etiqueta 3.

A la taula 7.10 es pot veure les cinc pitjors xarxes per *Adjusted Heterogenity* i amb l'etiqueta 0. Com s'espera, aquesta nova mètrica penalitza per igual aquelles xarxes que tenen una *Heterogenity* alta, mentre més propera a zero millor, però també aquelles que, com Kdl i Colt, que tenen una bona heterogeneïtat i un grau nodal mitjà, comparat amb la seva mida, molt baix.

Per la seva banda, a la taula 7.11 es pot veure les cinc pitjors xarxes per *Adjusted Heterogenity* amb l'etiqueta 3 i aquests són els punts grocs que destaquen per la figura 7.33, especialment Uninett2011 i Uunet. Per trobar una millor explicació de la seva robustesa, s'han visualitzat les dos xarxes anteriors amb el NRS, figures 7.34 i 7.35.

Per la xarxa Uninett2011, figura 7.34, costa de veure el detall central, per



(a) Diagrama de caixes amb *outliers* (b) Diagrama de caixes sense *outliers*

Figura 7.32: Figures relacionades amb la mètrica *Adjusted Heterogenity*.

Xarxa	N. of Nodes	Avg. Nodal Degree	Heterogenity	Adj. Heterogenity
Ulaknet	82	2	3.23	132.73
Pern	127	2.03	2.10	131.64
Kdl	754	2.38	0.35	113.37
Latnet	69	2.14	1.66	53.53
Colt	153	2.49	0.83	51.38

Taula 7.10: Les cinc pitjors xarxes per Adjusted Heterogenity i etiqueta 0

Xarxa	N. of Nodes	Avg. Nodal Degree	Heterogenity	Adj. Heterogenity
Uninett2011	69	2.84	0.61	14.85
Uunet	49	3.42	0.83	11.94
PionierL3	38	2.73	0.66	9.18
Geant2012	40	3.05	0.63	8.37
Geant2010	37	3.13	0.59	6.99

Taula 7.11: Les cinc pitjors xarxes per Adjusted Heterogenity i etiqueta 3

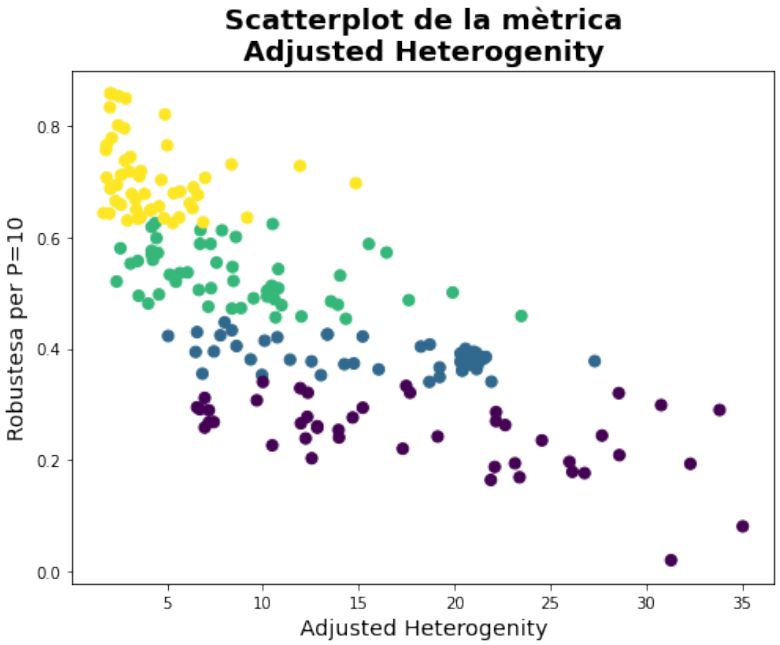


Figura 7.33: Gràfic de dispersió de l'*Adjusted Heterogenity* respecte la robustesa per P=10.

aquest motiu, s'ha ampliat i marcat els punts importants a la figura 7.36. En aquesta figura es pot veure que hi ha diversos punts centrals connectats entre ells, tant a la part superior com a la inferior, per la qual cosa en cas d'atac en què un d'aquests nodes sigui eliminat els altres dos continuen mantenint la xarxa sencera.

La xarxa Uunet, en canvi, figura 7.35, és una xarxa molt més connectada amb nombrosos nodes adjacents, com Seattle i San Francisco, que s'enllacen amb els mateixos nodes, com Dallas, a pesar de l'enorme distància. A més, tot i l'existència de nodes amb un grau nodal gran al centre del país, els nodes als extrems de la xarxa estan connectats entre ells. Això fa que la robustesa de la xarxa sigui molt bona, però també significa que el cost de construir-la va ser molt alt.

Aquesta nova mètrica, encara que soluciona alguns dels problemes de l'*Heterogenity* no és capaç de detectar aquest tipus de situacions.

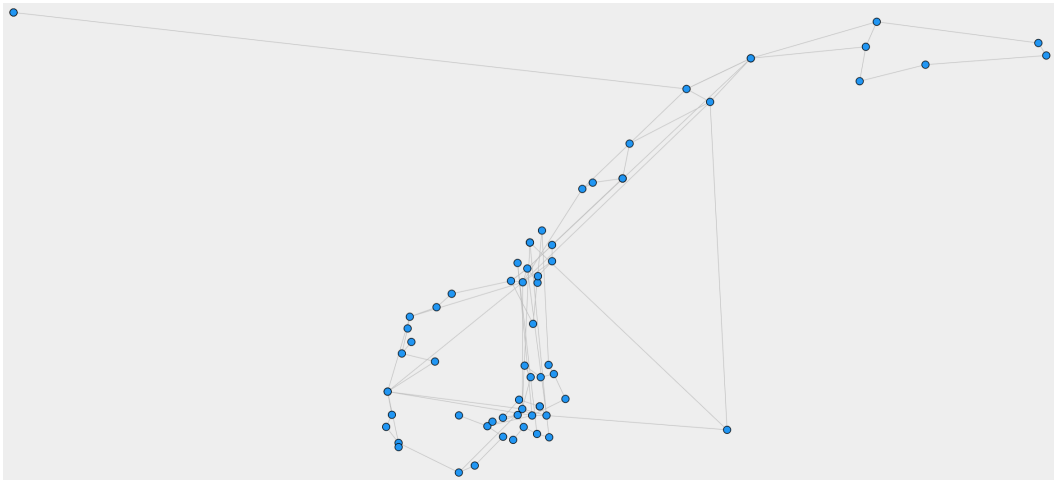


Figura 7.34: Visualització de la xarxa Uninett2011 amb el NRS.

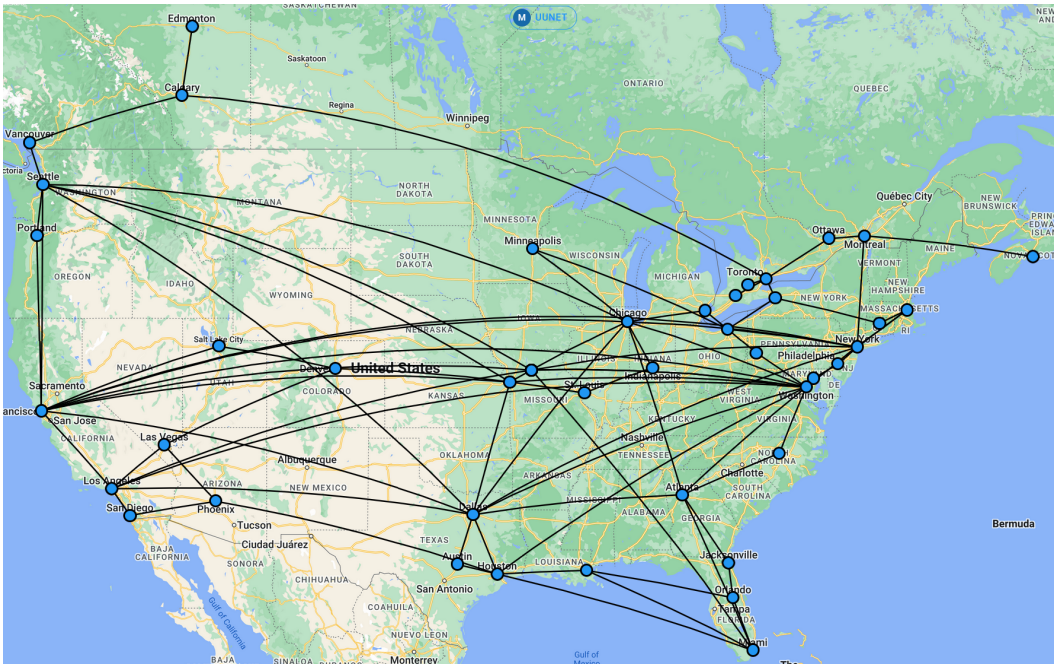


Figura 7.35: Visualització de la xarxa Uunet amb el NRS.

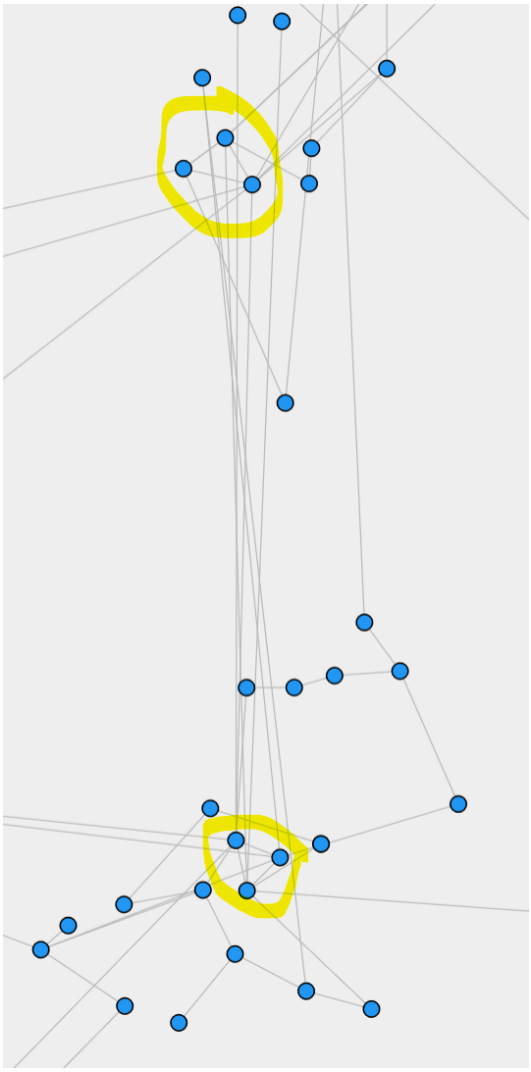


Figura 7.36: Visualització d'una part de la xarxa Uninett2011 amb el NRS.

CAPÍTOL 8

Conclusions i Treball Futur

8.1 Conclusions

El càlcul de la robustesa no és trivial, però tampoc explicar-la només amb els valors de les mètriques descrites perquè la majoria d'elles són difícils de comparar entre xarxes, com per exemple el *Largest Eigenvalue* o l'*Effective Resistance*, encara que aquesta última s'ha pogut normalitzar i explorar amb èxit com descriu la tipologia de la xarxa.

D'altres, en canvi, la comparació és més directe. D'aquestes la més interessant ha estat l'*Hetoregenity*, com de regular és una xarxa. Tanmateix, també pateix de certs problemes i a raó d'aquests s'ha desenvolupat al final una nova mètrica que intenta compensar l'heterogeneïtat per aquelles xarxes amb un grau nodal mitjà baix comparat amb la mida de la xarxa.

Tot i que aquesta nova mètrica soluciona algunes d'aquestes qüestions, hi ha altres característiques de la xarxa que no és capaç d'explicar, ja que s'ha de conèixer en més detall la tipologia.

Aquesta exploració de les mètriques i del Topology Zoo ha donat resultats molt interessants, però la precisió de la clusterització amb l'UMAP i l'HDBSCAN ha deixat una mica a desitjar perquè a partir del 50% de les mostres amb més confiança la predicció cau en picat.

Encara que s'ha intentat nombroses tècniques d'intel·ligència artificial per millorar els resultats de la clusterització, com *feature selection*, *feature engineering* i *data augmentation*, els resultats han estat limitats.

Tanmateix, també s'ha de tenir en compte que quan el mètode s'equivoca rarament ho fa per un factor superior a dos; és a dir, rarament clusteritza una xarxa molt poc robusta com robusta,... Això pot permetre donar una primera aproximació a quina és la robustesa esperada de la xarxa i, si la confiança de l'HDBSCAN és baixa, després es pot calcular la robustesa exacta amb el NRS.

Finalment, s'ha de comentar que no s'ha explorat ni treballat en l'últim punt descrit al full de la tesi en el que es deia que es desenvoluparien algorismes per millorar la robustesa de la xarxa afegint nodes i/o enllaços. Això s'ha degut a l'enorme complexitat del tema i de tots els possibles factors que s'han de tenir en compte a l'hora de trobar la robustesa d'una xarxa, explicar-la i, per últim, millorar-la. Així i tot, aquesta és una feina que queda pendent.

8.2 Treball Futur

Algunes de les feines que encara es poden fer en aquesta tesi són:

- Aprofundir en l'estudi de la robustesa i com les diferents mètriques l'expliquen.
- Implementar noves mètriques de robustesa al NRS com la constant de Kemeny.
- Millorar l'algorisme de generació de noves xarxes a partir d'un histograma.
- Millorar la clusterització de l'UMAP i l'HDBSCAN afegint noves xarxes al *dataset* tot i que siguin de tipologies diferents.
- Implementar l'algorisme d'intel·ligència artificial al NRS.
- Provar l'UMAP Paramètric on la reducció de la dimensionalitat es fa a través de xarxes neuronals.
- Provar una Graph Neural Network per classificar les xarxes directament en comptes d'utilitzar les diferents mètriques.
- Desenvolupar algorismes que permetin millorar la robustesa de les xarxes afegint nodes i/o enllaços.

Bibliografia

[Campello 2013] Ricardo J. G. B. Campello, Davoud Moulavi and Joerg Sander. *Density-Based Clustering Based on Hierarchical Density Estimates*. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda and Guandong Xu, editeurs, Advances in Knowledge Discovery and Data Mining, pages 160–172, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. (Cited on page 5.)

[Chapman 2000] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer and Rüdiger Wirth. CRISP-DM 1.0. 2000. (Cited on page 20.)

[Ester 1996] Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu. *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, page 226–231. AAAI Press, 1996. (Cited on page 5.)

[Joliffe 2016] Ian T. Joliffe and Jorge Cadima. *Principal component analysis: a review and recent developments*. vol. 374, apr 2016. (Cited on page 6.)

[Kinght 2011] Simon Kinght, Hung X. Nguyen, Nickolas Falkner, Rhys Bowden and Matthew Roughan. *The Internet Topology Zoo*. IEEE Journal on Selected Areas in Communications, vol. 29, no. 9, oct 2011. (Cited on page 13.)

[Madrenys 2022] Martí Madrenys. *Tècniques d'Intel·ligència artificial per calcular la robustesa de xarxes*, 2022. (Cited on page 9.)

[Manzano 2014] Marc Manzano, Faryad Sahneh, Caterina Scoglio, Eusebi Calle and Jose Luis Marzo. *Robustness surfaces of complex networks*. Optical Switching and Networking, vol. 4, 2014. (Cited on pages 1, 7 and 21.)

[Marzo 2018] Jose L Marzo, Eusebi Calle, Sergio G Cosgaya, Diego Rueda and Andreu Mañosa. *On selecting the relevant metrics of network robustness*. In 2018 10th International Workshop on Resilient Networks Design and Modeling (RNDM), pages 1–7. IEEE, 2018. (Cited on pages 1, 7, 21 and 39.)

[Marzo 2019] Jose L Marzo, Sergio G Cosgaya, Nina Skorin-Kapov, Caterina Scoglio and Heman Shakeri. *A study of the robustness of optical networks*

under massive failures. Optical Switching and Networking, vol. 31, pages 1–7, 2019. (Cited on pages i, 1, 4, 7, 8 and 21.)

[Marzo 2022] Jose L Marzo, David Martinez, Sergi Bergillos and Eusebi Calle. *Network Research Simulator. An abstract model formulation*. In 2022 18th International Conference on the Design of Reliable Communication Networks (DRCN), pages 1–4, 2022. (Cited on pages 1 and 11.)

[McInnes 2016] Leland McInnes, John Healy and Steve Astels. *The hdbscan Clustering Library*, 2016. (Cited on page 5.)

[McInnes 2017] Leland McInnes, John Healy and Steve Astels. *hdbscan: Hierarchical density based clustering*. The Journal of Open Source Software, vol. 2, no. 11, mar 2017. (Cited on page 5.)

[McInnes 2018a] Leland McInnes. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction - umap 0.5 documentation*, 2018. (Cited on page 5.)

[McInnes 2018b] Leland McInnes, John Healy and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, 2018. (Cited on page 5.)

[Rossi 2015] Ryan A. Rossi and Nesreen K. Ahmed. *The Network Data Repository with Interactive Graph Analytics and Visualization*. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015. (Cited on page 13.)

[Saltz 022] Jeff Saltz. *CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects*, (Accedit: Agost 2022). Disponible a https://www.datascience-pm.com/crisp-dm-still-most-popular/. (Cited on page 19.)

[Shchubert 022] Erich Shchubert. *Resposta a interpretation - Clustering on the output of t-SNE - Cross Validated*, (Accedit: Agost 2022). Disponible a https://stats.stackexchange.com/questions/263539/clustering-on-the-output-of-t-sne. (Cited on page 25.)

[Shekhar 2016] Karthik Shekhar, Sylvain W. Lapan, Irene E. Whitney, Nicholas M. Tran, Evan Z. Macosko, Monika Kowalczyk, Xian Adiconis, Joshua Z. Levin, James Nemesh, Melissa Goldman, Steven A. McCarroll, Constance L. Cepko, Aviv Regev and Joshua R. Sanes. *Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics*. Cell, vol. 166, no. 5, pages 1308–1323.e30, 2016. (Cited on page 26.)

[Trajanovski 2013] Stojan Trajanovski, Javier Martín-Hernández, Wynand Winterbach and Piet Van Mieghem. *Robustness envelopes of networks*. Journal of Complex Networks, vol. 1, no. 1, pages 44–62, 2013. (Cited on pages 1, 7 and 21.)