



UNIVERSITÀ DEGLI STUDI DI PISA

Facoltà di Scienze Matematiche, Fisiche e Naturali

Facoltà di Informatica

CORSO DI LAUREA SPECIALISTICA IN INFORMATICA

TESI DI LAUREA

Link Prediction su reti Multidimensionali

Candidato:

Giulio Rossetti

Relatore:

Dott.ssa Fosca Giannotti

Controrelatore:

Dott.ssa Chiara Bodei

Correlatore:

Dott. Michele Berlingario

Anno Accademico 2009-2010

Giulio Rossetti: "*Link Prediction su reti Multidimensionali*"
Facoltà di Scienze Matematiche, Fisiche e Naturali
Pisa, Anno Accademico 2009-2010

*"Quelli che s'innamora di pratica senza scienza son come l'nocchier ch'entra in naviglio senza timone
ne bussola, che mai certezza ha dove si vada"*

— Leonardo da Vinci

ABSTRACT

L'analisi di reti sociali (SNA) è un campo di ricerca interdisciplinare, che vede coinvolti fisici, sociologi, matematici, economisti e informatici, e che studia modelli e tecniche atti alla compresione dei fenomeni sociali all'interno di gruppi di persone.

Il Link Prediction, ossia la predizione di collegamenti futuri fra individui, rappresenta uno dei temi più caldi della Social Network Analysis. In questa tesi si estende lo scenario classico del Link Prediction al contesto delle reti multidimensionali, ossia quelle reti che annoverano molteplici connessioni fra coppie di individui. Partendo da tale modello si propone una nuova definizione per il problema di Link Prediction che tenga conto delle informazioni multidimensionali in esame: si presenta quindi una vasta tassonomia di approcci algoritmici studiati appositamente per sfruttare tali informazioni per la risoluzione del problema.

Vengono quindi introdotti nuovi predittori su reti multidimensionali, la cui validità è confermata da un'estensivo lavoro sperimentale effettuato su reti provenienti dal mondo reale.

*“Due cose riempiono l’animo di ammirazione
e venerazione sempre nuova e crescente,
quanto più spesso e più a lungo
la riflessione si occupa di esse:
il cielo stellato sopra di me, e la legge morale in me.”*

— Immanuel Kant, “Critica della Ragion Pratica”

RINGRAZIAMENTI

Una volta giunti alla fine di un percorso è naturale voltarsi ad osservare come questo sia stato favorito dalla presenza dei giusti compagni di viaggio.

Questo lavoro di tesi non sarebbe stato possibile senza la disponibilità dei miei relatori e del dott. Michele Coscia che mi hanno fornito i mezzi, e le basi, per affrontare le problematiche trattate.

Un sentito ringraziamento è dovuto ai miei familiari, i miei genitori e mia nonna che mi hanno sempre supportato durante il cammino insegnandomi ad aver fiducia in me stesso e nei miei mezzi. Un ringraziamento speciale va a mia sorella, la mia metà “irrazionale” che, fortunatamente, mi ricorda quanto sia necessario di tanto in tanto lasciare che i propri passi si discostino dai cammini tracciati dalla sola logica.

Tra i più importanti compagni di viaggio è impossibile dimenticare gli amici di sempre, compagni fraterni con cui da anni ormai incrocio il mio percorso: Elisa Del Santo, Roberto E. Bussi, Alessandro Carpi e tutti coloro che sono sempre stati presenti nelle tappe per me importanti.

E' doveroso ringraziare i colleghi, gli amici e i compagni di corso conosciuti, in questi anni, tra i libri universitari: Andrea Pinucci, Giulia Anticaglia e tutti coloro che hanno orbitato attorno al “Lab 5” e con cui ho diviso ore sui banchi in modo spensierato.

Molti altri nomi sarebbe doveroso aggiungere a questa lista: anche se non sono qui riportati a loro vanno comunque i miei più sinceri ringraziamenti.

Un ultimo ringraziamento va alle due persone che mi diedero, ormai anni fa e ciascuno a suo modo, il miglior consiglio mai avuto: “Figliolo, nella vita tutto quello che serve è la grinta, non rinunciare mai”. Grazie nonni, grazie.

INDICE

1	Introduzione	1
I STATO DELL'ARTE		5
2	Stato dell'arte	7
2.1	Reti e Grafi	7
2.1.1	Le reti nel mondo reale	9
2.1.2	Modello Multidimensionale	11
2.1.3	Reti Multidimensionali nel mondo reale	12
2.1.4	Informazioni Temporali	14
2.2	Graph Mining e Link Mining	16
2.2.1	Graph Mining	16
2.2.1.1	Scale Free Networks	16
2.2.1.2	Small Diameters	18
2.2.1.3	Community Structure	18
2.2.1.4	Resilience	19
2.2.2	Link Mining	19
2.2.2.1	Link-based Object Ranking	20
2.2.2.2	Link-based Object Classification	20
2.2.2.3	Community Discovery	20
2.2.2.4	Entity Resolution	21
2.2.2.5	Frequent Subgraph Discovery	21
2.2.2.6	Graph Classification	21
II LINK PREDICTION		23
3	Link Prediction	25
3.1	Formulazione del problema	25
3.1.1	Misure di base	26

3.1.2	Modelli monodimensionali	27
3.1.2.1	Modelli Unsupervised	28
3.1.2.2	Modelli Supervised	32
3.1.3	Limiti degli approcci classici	34
3.2	Link Prediction Multidimensionale	35
3.2.1	Formulazione del problema	35
3.2.2	Misure multidimensionali	37
3.2.3	Approcci adottati	43
3.3	Modelli derivati da algoritmi monodimensionali	44
3.3.1	Modelli derivati di base	44
3.3.2	Moltiplicatori locali	45
3.3.3	Moltiplicatori globali	47
3.3.4	Moltiplicatori temporali	49
3.4	Modelli AdHoc	52
III ANALISI Sperimentale		55
4	Metodologia di analisi	57
4.1	Presentazione dei Dataset	57
4.2	Metodologia di analisi dei risultati	67
4.2.1	Matrice di confusione	67
4.2.2	Curve ROC	69
4.2.3	Curve Precision/Recall	70
4.2.4	Analisi dell'andamento della Precision	72
5	Risultati sperimentali	75
5.1	AOL Query Log	76
5.1.1	Predittori derivati dai modelli Monodimensionali	76
5.1.2	Predittori Ad-Hoc	86
5.1.3	Analisi Riassuntiva	87
5.2	DBLP: Computer Science Bibliography	89
5.2.1	Predittori derivati dai modelli Monodimensionali	89
5.2.2	Predittori Ad-Hoc	99

5.2.3 Analisi Riassuntiva	100
5.3 IMDB: Internet Movie Database	101
5.3.1 Predittori derivati dai modelli Monodimensionali	101
5.3.2 Predittori Ad-Hoc	111
5.3.3 Analisi Riassuntiva	112
5.4 GTD: Global Terrorism Database	113
5.4.1 Predittori derivati dai modelli Monodimensionali	113
5.4.2 Predittori Ad-Hoc	123
5.4.3 Analisi Riassuntiva	124
5.5 GCD: Grand Comics Database	125
5.5.1 Predittori derivati dai modelli Monodimensionali	125
5.5.2 Predittori Ad-Hoc	135
5.5.3 Analisi Riassuntiva	136
5.6 VDC: International Dyadic Events	137
5.6.1 Predittori derivati dai modelli Monodimensionali	137
5.6.2 Predittori Ad-Hoc	147
5.6.3 Analisi Riassuntiva	148
IV CONCLUSIONI	149
6 Conclusioni	151
6.1 Valutazione dei risultati ottenuti	151
6.2 Ulteriori sviluppi	155
V APPENDICI	157
A Appendice	159
A.1 Specifiche implementative	159
A.2 Cytoscape Plugin	161
bibliografia	169

INTRODUZIONE

Durante gli ultimi anni si è assistito al sorgere, nella comunità scientifica, di un grande interesse per l'analisi e l'estrazione di conoscenza dalle reti complesse che oggi dominano il mondo reale.

Tale interesse, cresciuto in particolar modo grazie dall'avvento dei Social Network Online, aventi un significativo bacino di utenza e dai quali è possibile estrarre una gran mole di dati interessanti, è da considerarsi trasversale a molteplici campi di ricerca: fisici, matematici, biologi, informatici, sociologi ed economisti hanno rivolto la loro attenzione a problemi derivanti dallo studio delle reti sociali.

Il mondo reale, nell'esperienza quotidiana, fornisce innumerevoli esempi di interazioni e fenomeni che possano essere modellati per mezzo dell'astrazione fornita dal concetto di "rete". La possibilità di applicare tale descrizione ad una vasta tassonomia di fenomeni e di poterla formalizzare tramite un modello rigoroso, i grafi, ha consentito la nascita di svariate tecniche di analisi aventi solide basi matematiche.

Spesso le reti da cui si ha interesse ad estrarre informazioni non banali evolvono con il passare del tempo, espandendosi a causa dell'introduzione di nuove entità e dell'instaurarsi di nuove relazioni tra gli attori che ne fanno parte. Il tempo nelle reti può giocare un duplice ruolo: nel primo la struttura della rete evolve, nel secondo al passare del tempo si compiono delle azioni tra i nodi facenti parte della rete (gli utenti di un social network si scambiano messaggi, scienziati collaborano alla stesura di articoli...). Data la grande ricchezza di informazione che questo comporta, è comprensibile che l'interesse di numerosi studi si sia incentrato sulla ricerca di pattern evolutivi che siano in grado di predire, con una buona approssimazione, come una rete evolva nel tempo.

Riprendendo alcune delle idee presentate da Liben-Nowell e Kleinberg nell'articolo "The Link Prediction Problem for Social Network", in questa tesi si propongono alcuni modelli predittivi aventi come obiettivo l'analisi di una nuova e più complessa tipologia di reti: le reti multidimensionali. Chiamiamo multidimensionali le reti in cui due attori possono essere connessi tramite più di un link, ciascuno dei quali chiamiamo *dimensione*. I diversi link possono esprimere, alternativamente, o diversi tipi di relazione fra i due attori (sono amici, sono

colleghi, parlano via email, via telefono, ...), oppure possono esprimere valori differenti dello stesso tipo di relazione fra i due (ipotizzando una relazione di collaborazione scientifica le diverse conferenze di pubblicazione sono valori differenti della dimensione “collaborazione”). Le reti multidimensionali, modellate tramite i multigrafi, consentono di rappresentare le differenti tipologie di relazioni espresse nella realtà in modo da garantire una minore perdita di informazione durante la fase di analisi.

L’analisi multidimensionale è un utile strumento per la comprensione delle reti reali: la topologia multidimensionale è, infatti, comune a moltissime reti del mondo reale e la possibilità di sfruttare le informazioni da questa fornite consente di ottenere dei risultati aventi una affidabilità maggiore rispetto a quelli ottenuti a seguito dell’analisi di un modello semplificato. Reti sociali, tecnologiche, biologiche e di collaborazione possono presentare molteplici dimensioni: si pensi alle tipologie di relazioni interpersonali, ai differenti mezzi di trasporto, alle diverse tipologie di interazione in una rete neurale e, come precedentemente proposto, alle diverse conferenze di pubblicazione di un articolo.

In questo contesto, molti dei problemi affrontati dall’analisi classica devono essere rivisti nelle loro definizioni così come negli approcci proposti per la loro risoluzione. Nel caso specifico del problema di Link Prediction, in cui l’attenzione è focalizzata sulla predizione delle nuove possibili interazioni tra gli attori di una rete, l’introduzione della multidimensionalità rappresenta un nuovo ed interessante spunto di analisi. Il maggior livello di dettaglio topologico della rete costringe a rivedere gli approcci già noti in letteratura poiché, nella fase di predizione, è necessario ottenere anche una stima della dimensione in cui ciasun arco è previsto comparire.

Dati i molti approcci proposti in letteratura per il problema trattato si è deciso, in questo lavoro, di restringere l’ambito di ricerca alla classe dei modelli non supervisionati, modelli che sfruttano esclusivamente le informazioni topologiche inferite dalla rete, e in tale ambito sono state tracciate diverse tipologie di analisi.

Un primo approccio proposto ha visto l’estensione di modelli predittivi già noti in ambito monodimensionale al particolare caso studiato e il loro successivo arricchimento con informazioni multidimensionali sia globali alla rete sia locali ai singoli nodi. Tali modelli predittivi, successivamente chiamati “predittori base”, sono poi stati estesi per sfruttare anche le informazioni temporali associate agli archi della rete introducendo l’analisi della storia evolutiva come supporto al task affrontato.

Il secondo approccio proposto, invece, si basa sull’introduzione di una nuova classe di predittori (chiamati “Ad Hoc”) che, esulando dai modelli già noti in letteratura, sfruttano

esclusivamente informazioni di tipo multidimensionale e temporale per predire l’evoluzione della rete.

Queste due tipologie di approcci, che possiamo considerare complementari, sono dovute all’introduzione di nuove misure atte ad analizzare l’aspetto multidimensionale delle reti: l’introduzione di queste misure ha consentito l’adattamento dei modelli preesistenti e la creazione di nuovi modelli predittivi.

Come si mostra durante la trattazione, l’analisi eseguita, supportata da vasti risultati sperimentali, non è riuscita da individuare un predittore “universale” tra i 58 modelli proposti: i risultati positivi ottenuti devono essere considerati come linee guida per una successiva indagine.

Nella Parte I, in cui si analizza lo Stato dell’Arte, sono presentati i modelli matematici utilizzati durante la trattazione (grafi e multigrafi) e si discute l’ambito di ricerca in cui si colloca il problema trattato: il Graph Mining (e nello specifico i problemi affrontati dal Link Mining).

Una volta fissato l’ambito generale di indagine, nella Parte II si introduce il problema del Link Prediction, dandone una formalizzazione e presentando gli approcci noti in letteratura per il caso monodimensionale. Dopo aver presentato nel dettaglio il task analizzato se ne propone un’estensione all’ambito multidimensionale e si introducono i modelli predittivi elaborati per sfruttare al meglio le nuove informazioni topologiche fornite dal modello.

Nella Parte III, in cui viene descritta l’estensiva analisi sperimentale effettuata (prima presentando, e giustificando, la metodologia scelta, poi introducento i risultati ottenuti), si riporta sia graficamente che analiticamente il risultato dei test eseguiti su sei diverse reti multidimensionali tracciando, per ciascuna di esse un profilo topologico e un’analisi complessiva dell’andamento dei predittori. Si presentano inoltre, a supporto dei risultati forniti, i sei multigrafi appositamente costruiti, a partire da reti reali, per l’analisi sperimentale dei predittori precedentemente introdotti.

Nelle conclusioni, Parte IV, si riprendono i risultati ottenuti sulle singole reti e si dà un’analisi complessiva dell’andamento delle diverse tipologie di predittori adottati.

A concludere il presente lavoro di tesi si riporta in Appendice una breve presentazione dell’implementazione opensource dei multigrafi prodotta per la fase di analisi sperimentale e del plugin per Cytoscape realizzato per fornire un supporto visuale all’analisi di Link Prediction su reti multidimensionali.

Parte I

STATO DELL'ARTE

2

STATO DELL'ARTE

In questo capitolo si introduce la teoria dei grafi, utilizzata nel corso della trattazione, presentandone le basi per la modellazione di reti monodimensionali (2.1) ed estendendo successivamente i concetti mostrati all'ambito delle reti multidimensionali (2.1.2) arricchite da informazioni evolutive (2.1.4). Fissato il modello utilizzato per l'analisi si introduce il contesto di ricerca in cui questa tesi si colloca presentando le problematiche affrontate dal Graph Mining (2.2.1) e dal Link Mining (2.2.2).

2.1 RETI E GRAFI

Molteplici modelli matematici sono stati proposti per l'analisi delle reti nel corso dei secoli. Attualmente gli approcci principalmente utilizzati risultano essere due: le matrici di adiacenza (o loro generalizzazioni n-dimensionali chiamate tensori) e i grafi. Nel corso della trattazione faremo uso di questa seconda modalità semantico/sintattica sia per presentare le reti che andremo ad analizzare, sia per introdurre gli algoritmi di volta in volta presentati.

La scelta di preferire l'adozione dei grafi a quella dei tensori, seppure i modelli possano considerarsi equivalentemente espressivi, è stata maturata a seguito dell'osservazione che - oltre a rappresentare un interfaccia più familiare e facilmente interpretabile - buona parte della letteratura analizzata utilizza tale modello. Inoltre introducendo notevoli modifiche al modello di base (tramite l'adozione di informazioni temporali e multidimensionali) l'impiego delle matrici di tensori sarebbe risultata scarsamente utilizzabile per fornire esempi pratici necessari ad una corretta presentazione del problema affrontato.

Il modello matematico noto con il nome di "Grafo" è largamente utilizzato per modellare problemi di variegati ambiti di indagine sia di ambito teorico che pratico. La sua prima apparizione documentata è del 1735 ad opera di Eulero che introducesse tale rappresentazione nella sua pubblicazione sul problema dei "Sette ponti di Konigsberg". Altri noti problemi affrontati durante il XIX e XX secolo per mezzo di tale approccio topologico sono stati ad esempio il "Problema dei quattro colori" e quello di "Ciclo Hamiltoniano". Nella seconda metà

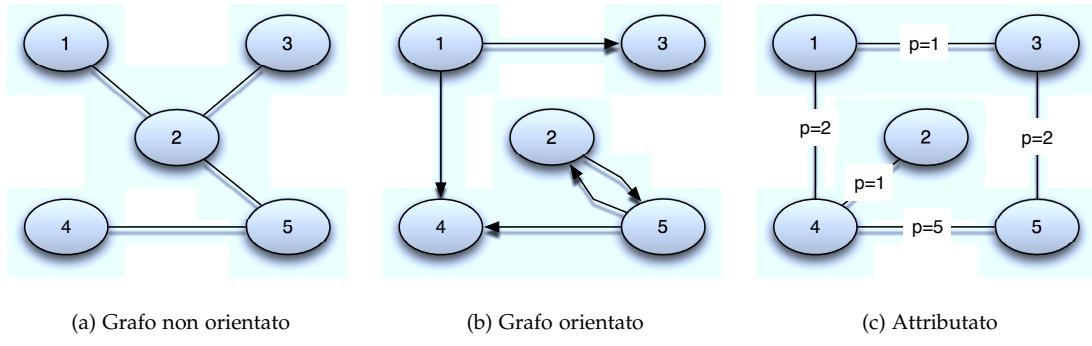


Figura 1: Grafi

del XX secolo, a seguito del rapido sviluppo dell'analisi combinatoria e della computer science, la teoria dei grafi è stata formalizzata e ampliata in modo dettagliato; numerosi problemi, tutt'oggi, vengono espressi per mezzo di tale impostazione metodologica.

Si da di seguito una definizione rigorosa di grafo introducendo la simbologia di base che verrà ampliata ed utilizzata nel corso della trattazione.

Definition 1. (Grafo)

Si definisce un grafo G come una coppia (V, E) di insiemi finiti dove: V è l'insieme dei nodi e E è un insieme di coppie di nodi. Nel caso esista un ordinamento tra i nodi componenti ogni coppia in E il grafo si dice orientato (Figura 1b), altrimenti è detto non orientato (Figura 1a). Scriveremo quindi $G = (V, E)$ con $v \in V$ e $(x, y) \in E$ $x, y \in V$.

Nella seguente trattazione ogni riferimento a grafi deve essere inteso, se non diversamente specificato, relativo a grafi non orientati.

La definizione appena data può essere estesa in molti modi: è infatti possibile introdurre nel modello pesi ed etichette (sia relativi agli archi che ad i nodi - Figura 1c) per rappresentare informazioni semantiche che meglio dettaglino la rete da analizzare.

Come abbiamo detto il modello proposto consente di rappresentare compiutamente molteplici tipologie di reti: questa flessibilità consente l'impiego dei grafi in molti ambiti di ricerca dall'analisi sociologica, al data mining, alla bioinformatica sino all'analisi di reti di trasporto e di infrastrutture.

Si mostrano, nella prossima sezione, alcune tipologie di reti del mondo reale che possono essere oggetto di studio per mezzo dei grafi in modo da dare al lettore un'idea dell'utilizzo del

modello introdotto per rappresentare varie tipologie di problemi.

2.1.1 *Le reti nel mondo reale*

Attribuendo accuratamente il corretto significato ai componenti di un grafo è possibile definire diverse tipologie di reti: modellare un contesto applicativo su di un grafo comporta scelte oculate che identifichino univocamente le entità rappresentate (i nodi) e le interazioni tra loro espresse (gli archi).

In letteratura sono state proposte diverse suddivisioni in categorie per le varie tipologie di reti (ad esempio in [24]): si presentano di seguito quattro macro-categorie che rendono evidente l'interdisciplinarità della tipologia di indagine che verrà introdotta nei prossimi capitoli.

RETI SOCIALI

Una rete sociale è definibile come un insieme di persone (di gruppi di persone, di associazioni...) che interagiscono tra di loro per mezzo di una stessa relazione.

Le interazioni possono essere di varia tipologia; relazioni sentimentali, di parentela, di lavoro sono tutte tipologie di connessioni che possono essere modellate con un grafo. Molteplici reti di questa tipologia possono essere costruite e, grazie alla diffusione capillare avvenuta negli ultimi anni dei Social Network online, è possibile recuperare grandi dataset per effettuare numerose tipologie di test su dati aventi una buona affidabilità.

RETI BIOLOGICHE

Numerosi sistemi biologici possono essere rappresentati per mezzo di una rete: interazioni tra proteine, reazioni metaboliche, la mappatura genetica, le reti neurali sono solo alcuni esempi.

Per sottolineare quanto la bioinformatica sia fortemente indirizzata sulla modellazione per mezzo di reti dei problemi affrontati è opportuno ricordare che uno dei principali progetti opensource per l'analisi di reti (Cytoscape) è nato proprio per soddisfare le necessità di questo ambito di ricerca¹.

¹ I predittori multidimensionali proposti in questa tesi sono stati implementati anche come plugin per l'analisi visiva di Link Prediction per Cytoscape (si veda Appendice A).

RETI TECNOLOGICHE

In questa categoria rientrano tutte quelle reti, create dall'uomo, aventi come obiettivo la distribuzione di beni o servizi.

Alcuni esempi di reti tecnologiche sono: la rete di approvvigionamento idrico, la rete telefonica, le reti di trasporti, Internet [34, 26], la rete elettrica. In questi specifici casi è abbastanza semplice individuare la semantica da applicare a nodi ed archi poiché spesso tali tipologie di reti sono strutturate, già prima di essere realizzate, proprio per mezzo di tale modello in modo da risolvere problemi di ottimizzazione sul traffico dei beni (o servizi) forniti.

RETI DI INFORMAZIONE

Questa categoria contiene diverse tipologie di reti: esempi sono le reti costruite sulle citazioni accademiche e la rete del World Wide Web.

Topologicamente le due reti prese in esempio sono molto diverse: le reti costruite sulle citazioni accademiche sono necessariamente acicliche (un articolo può citare solo articoli pubblicati precedentemente la sua stesura) mentre quella costruita sul WWW non ha restrizioni sui link data la modificabilità dei testi una volta pubblicati.

Le reti di informazione hanno la pregevole caratteristica di essere molto accurate e di fornire una ingente mole di dati da analizzare.

2.1.2 Modello Multidimensionale

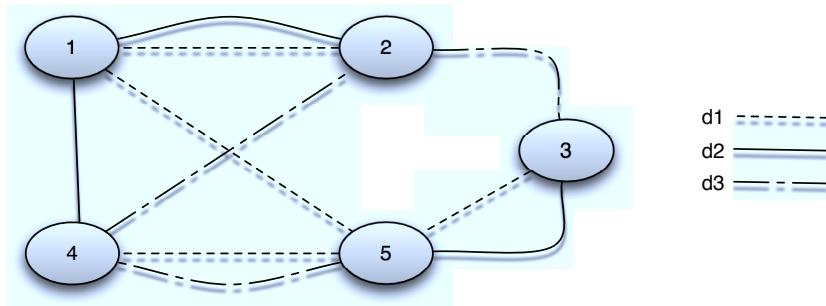


Figura 2: Multigrafo

Una volta introdotti alcuni esempi relativi alle tipologie di rete oggetto di analisi, è possibile comprendere come sia estremamente vincolante la restrizione imposta dal considerare, per ogni rete, una singola tipologia di interazione tra due nodi (Figura 2).

Perndiamo come esempio la classe delle reti sociali. Nella definizione proposta nel paragrafo precedente, per ogni rete sociale che possiamo costruire, deve essere considerata una ed una sola tipologia di relazione interpersonale tra le persone facenti parte del gruppo analizzato: tale restrizione comporta una semplificazione del modello fornito dal mondo reale a discapito di informazioni spesso rilevanti per l'analisi dei dati. Se consideriamo possibile, contrariamente a quanto fatto sino ad ora, la costruzione di una rete in cui due nodi possano essere messi in relazione tramite diverse tipologie di interazione, il modello che saremo in grado di analizzare rispecchierà più fedelmente le complessità presenti nel mondo reale.

Le informazioni topologiche introdotte, che chiameremo *dimensioni*, ci consentono di rappresentare un quadro più dettagliato della rete: diversificando le tipologie di archi che uniscono i nodi è possibile far emergere informazioni nascoste nella struttura della rete che possono essere sfruttate per studiare il modello costruito in modo più sofisticato.

Introduciamo ora la definizione di multigrafo.

Definition 2. (Multigrafo)

Si definisce un multigrafo G come una tripla (V, E, D) di insiemi finiti dove: V è l'insieme dei nodi, E è un insieme di coppie di nodi e D è un insieme di etichette, associate agli archi, rappresentanti le dimensioni del grafo. Scriveremo quindi $G = (V, E, D)$ con $v \in V$ $d \in D$

e $(x, y, d) \in E$ $x, y \in V$ $d \in D$. Nel caso in cui si presentino due triple del tipo (x, y, d') e (x, y, d'') necessariamente si ha che $d' \neq d''$.

Le informazioni che si riveleranno cruciali, per la corretta analisi di una rete, sono quelle di correlazione e anticorrelazione tra le dimensioni.

Queste informazioni, non rappresentate esplicitamente dal modello ma inferibili da esso, ci consentono di stimare per tutte le possibili coppie di insiemi di dimensioni appartenenti ad una rete quale sia il grado di correlazione, ridondanza ed esclusività dell'informazione fornita: sfruttando questa particolare tipologia di informazioni la modellazione della realtà analizzata assume una maggiore specificità e ricchezza di dettaglio.

Si presentano, come fatto precedentemente per le reti monodimensionali, alcuni esempi di reti multidimensionali appartenenti alle macro categorie presentate in [24].

2.1.3 Reti Multidimensionali nel mondo reale

Non essendo, ancora, le reti multidimensionali oggetto di analisi approfondita, in letteratura non sono presenti articoli che ne classifichino le varie tipologie in categorie (contrariamente a quanto accade per le reti monodimensionali). Per coerenza utilizzeremo le categorie portate come esempio a pagina 9.

RETI SOCIALI

Abbiamo già proposto un primo esempio di reti sociali multidimensionali per introdurre il modello espresso dai multigrafi.

Per fornire un esempio concreto possiamo pensare di dover modellare la struttura costruita sull'insieme di relazioni {amicizia, parentela, odio} per un determinato insieme di individui ben noti. Come facilmente intuibile alcune coppie di relazioni sono mutuamente esclusive mentre altre hanno diverso grado di correlazione.

RETI BIOLOGICHE

In molte reti biologiche è possibile discernere tra le tipologie di interazioni esistenti tra i nodi della rete. Nelle reti neurali, ad esempio, è possibile, a seconda dell'intensità dell'impulso elettrico instauratosi tra due neuroni stabilire la tipologia di informazione da essi trasportata: per

questi particolari casi è facile pensare di modellare tali reti come multidimensionali in modo da cogliere informazioni non altrimenti analizzabili mediante la visione classica monodimensionale delle interazioni tra i nodi.

RETI TECNOLOGICHE

Un esempio di rete tecnologica multidimensionale è quella costruita considerando differenti tipologie di trasporti.

E' possibile definire una rete di distribuzione differenziando le varie tratte percorribili in tipologie discrete (ad esempio: trasporti su strada, rotaia, nave, aereo): una volta introdotta tale informazione multidimensionale sulla rete è interessante, tra le altre cose, poter trovare il percorso ottimo in base a specifici vincoli sulla tipologia di trasporto da usare, sul costo o sul numero massimo di volte in cui è consentito cambiare dimensione durante la consegna.

RETI DI INFORMAZIONE

Se consideriamo una rete di co-authorship (due autori sono collegati se hanno contribuito alla stesura di uno stesso testo) è possibile introdurre la multidimensionalità associando ad ogni arco informazioni relative alla tipologia del testo in cui compare la collaborazione, o la conferenza a cui è stato presentato.

Secondo la definizione data di multigrafo anche informazioni di tipo temporale possono essere utilizzate come dimensioni. E' quindi possibile etichettare le collaborazioni, ad esempio in base all'anno in cui si sono sviluppate, e lavorare sul grafo ottenuto sfruttando tale informazione.

2.1.4 Informazioni Temporali

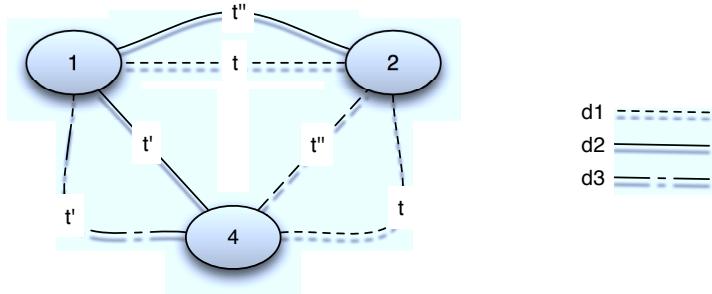


Figura 3: Multigrafo con informazioni temporali

Nell'ultimo esempio di rete multidimensionale proposta (rete di co-authorship con dimensioni costruite sugli anni di pubblicazione degli articoli) abbiamo introdotto una nuova informazione di tipo semantico: l'informazione temporale (Figura 3).

Tramite le informazioni temporali (associate agli archi nel caso proposto ma, sempre tramite etichette, estendibili anche ai nodi) è possibile effettuare un'analisi evolutiva della rete. Tale analisi è molto utile in diversi campi di indagine (e come vedremo in dettaglio anche nel Link Prediction) poiché ci consente di ottenere informazioni sul modello con cui la rete si evolve nel tempo.

Ovviamente, per come sono state introdotte sino ad ora, le informazioni temporali possono essere rappresentate con la struttura del multigrafo etichettato e rappresentano in un caso particolare di modellazione di reti multidimensionali.

Spesso questa modellazione, seppur più ricca semanticamente di quella offerta dai grafi monodimensionali, soffre delle stesse carenze informative riscontrate da questi ultimi. Scegliere di rappresentare le informazioni temporali come componente dimensionale di una rete esclude la possibilità di introdurre le informazioni sulla tipologia di interazioni che intercorrono tra gli attori della rete. Per ovviare a questo problema si introduce un ulteriore estensione del multigrafo presentato a pagina 11.

Definition 3. (Multigrafo con informazioni evolutive)

Si definisce un multigrafo G come una quadrupla (V, E, D, T) di insiemi finiti dove: V è l'insieme dei nodi, E è un insieme di coppie di nodi, D e T sono insiemi di etichette, associate agli archi, rappresentanti rispettivamente le dimensioni del grafo e gli istanti temporali di comparsa.

Scriveremo quindi $G = (V, E, D, T)$ con $v \in V$ $d \in D$ $t \in T$ e $(x, y, d, t) \in E$ $x, y \in V$ $d \in D$ $t \in T$. Nel caso in cui si presentino due quadruple del tipo (x, y, d', t') e (x, y, d'', t'') necessariamente si ha che $t' \neq t''$ oppure $d' \neq d''$.

Come risulta chiaro dalla definizione data, non esiste un informazione di tipo gerarchico tra le informazioni multidimensionali e quelle temporali. Rispetto alla definizione proposta di multigrafo, viene rilasciato il vincolo che imponeva la presenza, tra due nodi, di un singolo arco per ogni tipologia di interazione in modo da permettere, per ciascuna dimensione, l'eventualità della ripetizione in più intervalli temporali distinti.

Buona parte delle reti precedentemente presentate possono essere analizzate mediante informazioni di tipo multidimensionale ed evolutivo: gli unici casi in cui ciò non risulta possibile si verificano, nel momento in cui i dati, di cui si dispone per la creazione della rete, rappresentino informazioni aggregate. In tali casi, spesso, l'informazione temporale subisce una perdita di significato o risulta parzialmente/totalmente non disponibile.

Riprenderemo il concetto di reti multidimensionali ed evolutive quando presenteremo i dataset presi in considerazione come base per l'analisi dei modelli di Link Prediction proposti (Capitolo 4.1).

2.2 GRAPH MINING E LINK MINING

In questa sezione si presentano alcuni degli ambiti di indagine, appartenenti al Data Mining, che, utilizzando la modellazione proposta, hanno come fine quello di far emergere informazioni non banali dall'analisi delle reti. Come prima cosa si presentano alcune peculiarità relative alla struttura delle reti analizzate (Graph Mining [5]); successivamente si introduce lo specifico ambito di ricerca che racchiude in se il problema affrontato in questa tesi (Link Mining [8]) presentando al lettore le differenti tipologie di informazioni su cui si articola la ricerca in questo settore.

2.2.1 Graph Mining

Durante gli anni diverse tipologie di studi hanno posto l'analisi della struttura topologica delle reti come loro punto di interesse principale. Un obiettivo comune alle ricerche presenti nella letteratura analizzata è quello di riuscire a riconoscere, e codificare, caratteristiche peculiari comuni ai grafi realizzati per modellare reti appartenenti al mondo reale.

La ricerca dei pattern comuni a varie tipologie di reti acquisisce una notevole importanza non solo per l'impatto che questi hanno sulle metodologie di analisi (i cui risultati possono essere meglio interpretati utilizzando tali informazioni) ma anche per lo studio di generatori di grafi.

Un generatore di grafi è uno strumento che, mantenendo degli invarianti topologici, cerca di costruire un grafo sintetico rappresentante una predeterminata tipologia di rete da analizzare. Vari modelli evolutivi sono stati studiati per generare reti con particolari, desiderabili, proprietà: tali modelli sono spesso sfruttati nel task di Link Prediction per tentare di cogliere le informazioni latenti della rete. Ovviamente non esiste un modello che riesca a descrivere tutte le possibili tipologie di reti analizzabili: proprio per questo motivo, a seconda della rete analizzata, ciascun modello proposto può descriverne più o meno bene la topologia.

Alcuni pattern riconosciuti come comuni nelle reti ottenute dal mondo reale sono presentati di seguito.

2.2.1.1 Scale Free Networks

In molte reti è possibile osservare una particolare distribuzione del *degree* dei nodi: tale distribuzione va sotto il nome di Power Law e, nella sua forma classica può essere espressa nel

seguente modo:

Definition 4. (Power Law classica)

Due variabili x e y sono correlate da una power law quando:

$$y(x) = Ax^{-\gamma} \quad (2.2.1)$$

dove A e γ sono costanti positive. La costante γ è definita “esponente della power law”.

Una volta definita la Power Law è possibile definire analogamente la distribuzione probabilistica da questa definita.

Definition 5. (Distribuzione Power Law)

Una variabile aleatoria si distribuisce secondo una Power Law quando la funzione di densità probabilistica è data da:

$$p(x) = Ax^{-\gamma}, \gamma > 1, x \geq x_{\min} \quad (2.2.2)$$

Il vincolo presente su $\gamma > 1$ assicura che $p(x)$ possa essere normalizzata. Power Law con esponente $\gamma < 1$ occorrono raramente in natura.

L'importanza della distribuzione presentata è evidenziata dal fatto che questa stabilisce una decadenza polinomiale per $x \rightarrow \infty$

contrariamente a quanto avviene per la distribuzione Gaussiana (dove il decadimento è esponenziale). Per questo motivo una distribuzione del degree dei nodi che si conformi alla Power Law ci consente di sostenere che in tale grafo esistono pochi nodi con alto degree e molti nodi con bassissimo degree.

Grafi che presentano tali peculiarità sono definiti scale-free poiché $y(x)$ rimane immutato, a meno di una costante moltiplicativa, quando la variabile x sia moltiplicata per un valore scalare.

2.2.1.2 Small Diameters

Travers e Milgram [30] hanno condotto un famoso esperimento in cui ai partecipanti era richiesto di raggiungere un destinatario, scelto casualmente, creando una “catena di lettere” costruita in modo da avvicinarsi ad ogni passo tramite conoscenti ad una persona non nota.

L'esperimento mostrò che, mediamente, la lunghezza della catena di lettere si assesta intorno a sei. Tale valore (che ha reso noto l'esperimento come “I sei gradi di separazione”) è molto ridotto se prendiamo in considerazione l'estesa popolazione di partecipanti potenziali all'esperimento: successivi test hanno portato ad concludere che, in molte reti costruite basandosi su dati forniti dal mondo reale, il diametro mantiene valori molto bassi se relazionati alla dimensione della rete analizzata [20].

2.2.1.3 Community Structure

Una community è generalmente descritta come un insieme di nodi in cui ciascun nodo è “più vicino” a nodi della community stessa che non a nodi esterni ad essa.

Questa peculiare struttura è riscontrabile in molte reti derivate dall'osservazione di fenomeni del mondo reale: in particolare tale pattern è frequente nelle reti sociali.

Varie misure sono state introdotte per descrivere questo fenomeno: la principale è il coefficiente di clustering.

Definition 6. (Coefficiente di Clustering - locale)

Supponiamo che il nodo i abbia k_i vicini e che esistano n_i archi tra di essi: il coefficiente di clustering del nodo i è definito come:

$$C_i = \begin{cases} \frac{n_i}{k_i} & k_i > 1 \\ 0 & k_i = 0 \text{ o } 1 \end{cases} \quad (2.2.3)$$

Nella teoria dei grafi, il coefficiente di clustering rappresenta la misura del grado in cui i nodi di una rete tendono a raggrupparsi insieme.

Sono state presentate due diverse versioni del coefficiente di clustering: una globale ed una locale. La versione globale tende a dare informazioni globali relativamente al grado di

Clustering della rete, la versione locale invece mostra quanto un singolo nodo sia interconnesso ai propri vicini.

Il coefficiente globale analizza triple di nodi. Una tripla è definita da tre nodi connessi da due (tripla aperta) o da tre (triangolo, o clique di tre) archi non orientati. Il coefficiente globale di clustering può quindi essere computato secondo la formula:

$$C = \frac{3 * \text{numero di triangoli}}{\text{numero di triple di vertici connesse}} \quad (2.2.4)$$

La versione locale indica per ogni nodo i quanto sia prossimo il formarsi di clique con i suoi vicini. Tali misure sono spesso utilizzate per valutare se una rete esprime le caratteristiche di small-world e per analizzarne la struttura gerarchica [6, 11].

2.2.1.4 Resilience

Un'ultima misura, tra le tante che potrebbero essere introdotte, è quella di resilience. In un grafo la resilience misura la robustezza della struttura in caso venissero a mancare determinati archi e/o nodi.

Molti grafi ottenuti da reti del mondo reale dimostrano un alta resilience relativamente a fallimenti randomici di nodi e/o archi ma bassa resilience in caso di attacchi mirati.

Questi risultati si spiegano alla luce della distribuzione che abbiamo detto accomunare molte reti del mondo reale: la Power Law. Attacchi mirati ai pochi nodi aventi altissimo degree, infatti, possono causare una frammentazione del grafo in più componenti (spesso di piccole dimensioni) completamente sconnesse tra di loro. Fallimenti casuali, d'altro canto dato il tipo di distribuzione, procurano pochi danni alla struttura globale della rete.

2.2.2 Link Mining

Con il termine Link Mining [8] si raggruppano tutte le tecniche di Data Mining che considerano esplicitamente le informazioni fornite dai link appartenenti ad una rete per costruire modelli descrittivi, predittivi e generativi atti ad analizzare i dati rappresentati.

Il link mining rappresenta un area di ricerca relativamente nuova e nasce dall'intersezione di vari lavori sul link analysis, hypertext e web mining, relational learning, programmazione induttiva e graph mining.

Di seguito si presentano alcuni dei temi affrontati in questo campo di ricerca in modo da presentare il contesto generale in cui si pone l'argomento centrale della trattazione e, al contempo, di consentire al lettore una miglior comprensione dei settori a cui è possibile estendere l'utilizzo delle informazioni multidimensionali ed evolutive.

2.2.2.1 Link-based Object Ranking

Uno dei più conosciuti task di mining è probabilmente il Link-based Object Ranking (LBR). Il fine di questa analisi è quello di sfruttare la struttura di un grafo per di ordinare (o assegnare una priorità) ad un'insieme di oggetti appartenenti al grafo stesso.

Alcuni esempi ben noti sono ritrovabili nell'ambito del web information retrieval: gli algoritmi di HITS e PageRank sono a tutti gli effetti approcci di tipo LBR.

Nella Social Network Analysis spesso si introducono misure di centrality per definire l'importanza relativa dei nodi appartenenti alla rete; un'altra misura comunemente usata è quella di similarity tra due oggetti.

2.2.2.2 Link-based Object Classification

Dato un grafo e un insieme finito di etichette, gli approcci di Object Classification sfruttano la tipologia dei link incidenti in ciascun nodo per assegnare a questo una specifica categoria (tra quelle definite nell'insieme di etichette). La tipologia di analisi intrapresa discosta LBC dal classico problema di classificazione poiché, come accade in molti casi, i valori delle etichette di oggetti collegati tra loro tendono ad essere correlati.

LBC viene spesso utilizzato per eseguire task di machine learning che abbiano come dati di input grafi.

2.2.2.3 Community Discovery

L'obiettivo del community discovery è quello di clusterizzare i nodi appartenenti ad un grafo in gruppi che condividono determinate caratteristiche.

Consideriamo, come esempio, un grafo che contenga oggetti e link di una singola tipologia, senza attributi. Molte tecniche sono state sviluppate per identificare gruppi in uno scenario simile: approcci di clustering sia di tipo agglomerativo che divisivo. Alcuni approcci segnalati in letteratura sono: spectral graph partitioning, stochastic blockmodeling e l'uso della misura di edge betweenness.

2.2.2.4 Entity Resolution

Gli approcci di entity resolution sono mirati ad identificare quali oggetti di un grafo rappresentino la stessa entità del mondo reale. Esempi di applicazione possono essere trovati nell'elaborazione di database, nel natural language processing (risoluzione di co-reference, consolidamento dei dati), nella gestione delle informazioni personali e in molti altri campi.

Il problema può essere definito in molte varianti; nella forma più generale si assumono sconosciute sia le entità del dominio sia il numero di tali entità.

L'idea centrale che sta alla base degli approcci di entity resolution consiste nel considerare, non solo gli attributi associati alle entità da risolvere, ma anche quelli delle entità ad esse linkate. Questi link possono rappresentare, ad esempio, relazioni di co-authorship tra le reference di autori in una rete bibliografica o link gerarchici tra reference spaziali in dati geografici.

2.2.2.5 Frequent Subgraph Discovery

Un problema di data mining strettamente correlato al link mining è il subgraph discovery.

Le ricerche in questo ambito hanno come fine quello di individuare, in un insieme di grafi, strutture comuni o interessanti secondo alcuni criteri. La ricerca di questi pattern può essere intesa non solo come task a se stante ma anche come prima fase per effettuare un processo di classificazione.

Molti degli approcci di subgraph discovery sfruttano la proprietà nota con il nome di *Apriori* per effettuare il mining di insiemi di oggetti frequenti. Tipicamente il processo prevede due fasi: una prima fase di generazione dei candidati seguita da una in cui viene eseguito il match delle strutture individuate.

Esempi algoritmici a questo problema sono quelli forniti da gSpan[33], GERM[2] e LFR[15].

2.2.2.6 Graph Classification

Diversamente dai procedimenti di classificazione link-based, il cui scopo è quello di etichettare i nodi di un grafo, graph classification è un problema di apprendimento supervisionato il cui fine è quello di categorizzare un intero grafo come istanza positiva o negativa di un determinato concetto. Questo task si presenta come uno dei primi studiati per l'analisi di grafi sia dal data mining che dal machine learning.

Tre principali tipologie di approcci sono state proposti per il problema: analisi tramite feature mining, definizione del kernel di un grafo e programmazione logica induttiva (ILP).

Parte II

LINK PREDICTION

3

LINK PREDICTION

In questo capitolo si introduce l'argomento centrale affrontato nel lavoro di tesi.

Utilizzando i modelli presentati in 2.1 si formula in modo rigoroso il problema di Link Prediction per reti monodimensionali (3.2.1) quindi, dopo aver presentato le necessarie misure (3.1.1), si analizzano nel dettaglio alcuni approcci proposti in letteratura e le differenti tipologie di analisi che possono essere effettuate su tale modello (3.1.2).

Una volta esaurita tale trattazione si estende la formulazione del problema ai modelli multidimensionali con informazioni temporali (3.2.1) introdotti nel Capitolo 2, si presentano nuove funzioni di misura per gestire le più ricche informazioni topologiche (3.2.2) e si mostrano le tipologie di predittori proposte (3.2.3).

Nelle Sezioni 3.5 e 3.6 si analizzano i predittori appartenenti alle tipologie introdotte: una valutazione delle performance sarà data nei Capitoli 4 e 5.

3.1 FORMULAZIONE DEL PROBLEMA

Le reti, normalmente oggetto di analisi, sono dinamiche. Durante la propria vita in una rete appaiono nuovi link, indicanti nuove interazioni tra oggetti.

Nel problema di link prediction, usualmente, partiamo dalla conoscenza fornita da uno snapshot della rete, ad un prestabilito istante t , e siamo interessati a predire quali saranno gli archi che si aggiungeranno alla rete in un intervallo di tempo che va dall'istante della nostra osservazione ad un dato istante futuro t' . L'obiettivo che ci poniamo è quello di predire come evolverà la struttura della rete analizzata sfruttando le informazioni topologiche da essa stessa fornite.

Questo approccio, che come precedentemente osservato rientra nell'ambito del Link Mining (presentato in 2.2.2), è strettamente legato alla ricerca del modello evolutivo della rete stessa: la predizione sarà tanto più accurata quanto più, il predittore usato, riuscirà a sfruttare le informazioni topologiche a sua disposizione per inferire (o applicare) un modello evolutivo il più possibile prossimo a quello che regola la rete di volta in volta analizzata.

Come primo esempio si consideri una rete sociale modellata sulla relazione di co-authorship tra scienziati: intuitivamente possiamo predire che due scienziati che sono “vicini” nella rete costruita molto probabilmente collaboreranno in futuro alla stesura di un articolo.

Numerosi approcci al problema di Link Prediction hanno proposto varie misure per valutare la “vicinanza” tra i nodi appartenenti ad una rete: molte di queste misure sono originate da tecniche mutuate dalla teoria dei grafi e dalla Social Network Analysis.

Definiamo in modo formale il problema classico di Link Prediction.

Definition 7. (Link Prediction monodimensionale)

Sia $G_0 = (V, E_0)$ un grafo non orientato definito nelle sue componenti di nodi ed archi osservato ad un dato istante t_0 . Il task di Link Prediction, dato un istante $t_1 > t_0$ e il relativo insieme degli archi E_1 , consiste nel prevedere gli archi che entreranno a far parte del grafo originario nell’istante futuro t_1 (archi appartenenti all’insieme $E_{\text{new}} = E_1 - E_0$). Per ogni arco nell’insieme dei risultati deve inoltre essere presente uno *score* di confidenza.

Dalla definizione risulta chiaro che, per ogni arco predetto, deve essere calcolato un valore che indichi la “bontà”, stimata dal predittore, che questo sia un possibile candidato all’entrata nel grafo nell’intervallo di tempo analizzato. Il risultato dell’applicazione di un predittore è quindi un insieme di triple $(\text{nodo}, \text{nodo}, \text{score})$ con $\text{score} \neq 0$.

Il problema affrontato può presentare alcune varianti nella sua formulazione come avviene in [19] in cui si considera il grafo di partenza privo di archi.

Introduciamo alcune misure di base utilizzate dai predittori analizzati nel seguito in modo da rendere esplicativi alcuni concetti che saranno ripresi al momento dell’estensione all’ambito multidimensionale del problema trattato.

3.1.1 Misure di base

Molte informazioni topologiche possono essere sfruttate per calcolare uno score, per ogni coppia di nodi appartenenti al grafo, al fine di decidere quale sia la “probabilità” con cui si verrà a creare un arco tra le due entità prese in considerazione. Si introducono di seguito alcuni concetti basilari usati spesso per “misurare” valori di interesse relativi al singolo nodo (Neighbours e Degree) o a coppie di nodi.

NEIGHBOURS

Il valore restituito dalla funzione Neighbours (che nella trattazione successiva verrà spesso rappresentato simbolicamente con la scrittura $\Gamma(x)$ dove $x \in V$) identifica l'insieme dei nodi "vicini" al nodo a cui essa è applicata. In un grafo non diretto tale insieme corrisponde esattamente ai nodi collegati tramite un arco al nodo a cui è applicata la funzione; in un grafo diretto, al contrario, i vicini ad un nodo sono tutti e soli i nodi raggiungibili dal nodo in questione tramite un arco uscente da esso.

L'insieme dei Neighbours di un nodo è utilizzato spesso sia per computare score sia per ridurre il numero delle possibili coppie di nodi candidate all'analisi da parte dei predittori.

DEGREE

Il valore della funzione Degree (indicata simbolicamente con $|\Gamma(x)|$ dove $x \in V$) identifica la cardinalità dell'insieme dei Neighbours del nodo a cui è applicata. Simmetricamente a quanto accade per la funzione Neighbours, nel caso di grafi non diretti, il valore coincide con il numero dei nodi collegati tramite un arco al nodo a cui è applicata; nel caso invece di grafi diretti si distinguono due funzioni InDegree e OutDegree che tengono traccia, rispettivamente, del numero di archi uscenti dal nodo (equivalente alla cardinalità dell'insieme dei Neighbours per il grafo diretto) e del numero di archi entranti.

SHORTEST PATH

Nella teoria dei grafi si definisce lo Shortest Path come il cammino minimo tra due vertici, ovvero - nel caso di grafo con archi non pesati - il percorso più breve che congiunge i due vertici dati.

Per risolvere tale problema esistono numerosi approcci tra cui si ricordano: l'algoritmo di Dijkstra e l'algoritmo di Bellman-Ford. Un applicazione già citata (2.2.1.2) è quella dell'esperimento noto con il nome di "sei gradi di separazione".

3.1.2 Modelli monodimensionali

In letteratura sono stati analizzati molteplici modelli predittivi sino ad ora. Tutti gli approcci proposti possono essere classificati in due ben precise categorie: predittori supervised e non supervised (come mostrato in Figura 4).

Nelle seguenti sotto sezioni si analizzano le caratteristiche principali di questi due macro

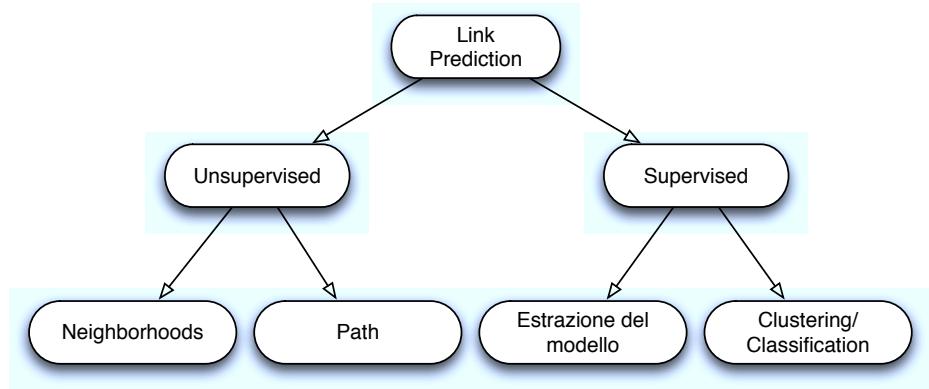


Figura 4: Gerarchia Modelli Link Prediction

gruppi riportando riferimenti bibliografici per consentire al lettore di approfondire gli argomenti introdotti: si dettagliano, inoltre, i modelli utilizzati come base di analisi nella trattazione successiva (3.5) per l'estensione all'ambito multidimensionale e temporale.

3.1.2.1 Modelli Unsupervised

Nella categoria dei predittori non supervisionati rientrano tutti gli approcci algoritmici che assumono un modello evolutivo per il grafo analizzato al fine di computare gli score per gli archi candidati.

Questa tipologia di approccio è stata ampiamente analizzata in letteratura e i modelli predittivi che vi appartengono possono, a loro volta, essere raggruppati in due sotto-insiemi: quelli basati su neighborhoods e quelli basati sull'analisi dei path.

Inoltre sono stati proposti alcuni “meta-approcci”, ovvero modelli che possono essere utilizzati in modo congiunto con gli altri due sottoinsiemi[21].

METODI BASATI SULLA NEIGHBORHOODS

Numerosi approcci si basano sull'idea che due nodi, x e y , hanno maggiori probabilità di essere connessi da un arco nel futuro se i loro insiemi di vicini, $\Gamma(x)$ e $\Gamma(y)$, presentano una larga sovrapposizione; questi modelli seguono l'idea intuitiva che, prendendo come esempio una rete di co-authorship, due autori aventi molti colleghi in comune abbiano buona probabilità di venire in contatto e di contribuire alla stesura di un articolo. I principali approcci che rientrano in questa categoria sono presentati di seguito.

Common Neighbours L'implementazione più diretta dell'idea introdotta è certamente rappresentata dal algoritmo che va sotto il nome di Common Neighbours. Tale predittore assegna lo score agli archi secondo la formula:

$$\text{score}(x, y) = |\Gamma(x) \cap \Gamma(y)| \quad (3.1.1)$$

ovvero tenendo conto esattamente del numero dei vicini in comune ai due nodi. Tale misura è stata utilizzata, ad esempio, da Newman [23] su di una rete di collaborazioni, verificando una correlazione tra il numero dei vicini in comune ai nodi x e y all'istante t , e la probabilità che questi collaborino in futuro.

Jaccard Il coefficiente di Jaccard è una misura spesso utilizzata per calcolare la similarity nell'ambito dell'information retrieval [9]: la misura fornita è un indice della probabilità, data una *feature* f appartenente ad x o y , che questa si presenti sia in x che in y . Nel caso del problema di Link Prediction la “feature” considerata sono i vicini nel grafo all'istante t analizzato: lo score viene quindi assegnato tramite la formula:

$$\text{score}(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \quad (3.1.2)$$

Poiché questa misura è diversa da zero solo per le coppie di nodi per cui il valore di Common Neighbours è diverso da zero si può osservare che l'insieme dei risultati dei due predittori coincide a meno dell'ordine degli score.

Adamic Adar La misura di similarity proposta da Adamic e Adar [1] è una ulteriore tipologia di valutazione degli score per coppie appartenenti allo stesso insieme di risultati proposto dai due modelli precedenti.

La formula per che esprime il modello è la seguente:

$$\text{score}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|} \quad (3.1.3)$$

Tale misura nasce per valutare la misura di correlazione tra due pagine web in base ad una determinata feature (nel caso del link prediction la feature utilizzata è nuovamente l'insieme dei vicini in comune ai due nodi).

Preferential Attachment Preferential Attachment ha ricevuto una considerevole attenzione come modello di sviluppo delle reti. La premessa basilare, sostenuta da tale modello, è che la probabilità che un nuovo arco abbia il nodo x come suo estremo è proporzionale a $|\Gamma(x)|$, ovvero al numero corrente dei vicini al nodo stesso.

Newman [23] e Barabasi, a seguito di analisi empiriche, hanno osservato che in reti di co-authorship la probabilità che due nodi x e y siano collegati da un arco è correlata al numero di pubblicazioni associate ai due nodi.

Tale modello è esprimibile nel seguente modo:

$$\text{score}(x, y) = |\Gamma(x)| \times |\Gamma(y)| \quad (3.1.4)$$

Numerose pubblicazioni hanno analizzato nel dettaglio tale modello di crescita delle reti e valutato il suo impatto in task di link prediction: in particolare [5, 27, 32, 13].

Altri modelli Altri modelli, presenti in letteratura, fanno delle misure di Neighbourhood e Degree gli strumenti principali per calcolare la “prossimità” tra due nodi.

Esempi sono forniti dai modelli evolutivi di Forest Fire e DMC [22] che, seppur non ideati per task di Link Prediction, sono stati adattati per applicazioni in questo ambito (ed anche per affrontare al problema di Network Archeology¹ che possiamo considerare correlato al tema trattato).

METODI BASATI SULL'ANALISI DEI PATH

Alcuni metodi di Link Prediction sfruttano la nozione di shortest-path tra due nodi considerando l'insieme di tutti i cammini esistenti tra di essi. Vediamo alcuni metodi di analisi unsupervised che fanno uso di questa misura.

¹ La Network Archeology si occupa, dato uno snapshot della rete, di determinare in che modo questa si è evoluta nel tempo ricostruendone il passato. Modelli evolutivi (e quindi approcci di link prediction come PA, CN, AA e JC) possono, in molti casi, essere adattati per effettuare questo particolare tipo di analisi.

Katz Questo predittore definisce lo score sommando tutti i path, tra i due nodi candidati alla formazione un arco, dando maggior peso ai path più corti tramite un fattore esponenziale. La misura è definita dalla formula:

$$\text{score}(x, y) = \sum_{\ell=1}^{\infty} \beta^\ell \cdot |\text{paths}_{x,y}^{(\ell)}| \quad (3.1.5)$$

dove $\text{paths}_{x,y}^{(\ell)}$ è l'insieme di tutti i cammini di lunghezza ℓ tra i nodi x e y e $\beta > 0$ è un parametro del predittore.

Hitting Time, Page Rank Un “cammino casuale” sul grafo G inizia da un nodo x e iterativamente prosegue su un vicino al nodo scelto con probabilità uniforme. Hitting Time ($H_{x,y}$) è il numero atteso di passi richiesti da un cammino casuale per arrivare dal nodo x a y . Poiché tale misura non è simmisura solitamente si usa la misura definita “Commute Time” definita come $C_{x,y} = H_{x,y} + H_{y,x}$.

PageRank apporta delle modifiche al modello proposto da Hitting Time introducendo la possibilità che, durante il cammino casuale, avvengano a random dei “reset”: ad ogni passo esiste la probabilità non nulla α di ritornare al nodo di origine del cammino. Questo modello è un adattamento della versione base dell’algoritmo di PageRank usato per task di LBR (come abbiamo visto in 2.2.2.1).

SimRank Lo score calcolato da SimRank è definito come il punto fisso della seguente definizione: due nodi sono simili nella misura in cui questi sono collegati a vicini simili. Numericamente possiamo specificare tale quantità ponendo $\text{Similarity}(x, x) = 1$ e

$$\text{Similarity}(x, y) = \gamma \cdot \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} \text{Similarity}(a, b)}{|\Gamma(x)| \cdot |\Gamma(y)|} \quad (3.1.6)$$

fissato un parametro $\gamma \in [0, 1]$. Possiamo quindi definire $\text{score}(x, y) = \text{Similarity}(x, y)$.

METODI DI ALTO LIVELLO: “META-APPROCCI”

Oltre ai modelli di predizione discussi, rappresentanti solo una parte - seppur significativa - di quelli proposti in letteratura, si introducono due degli approcci di più alto livello

che possono essere utilizzati, in congiunzione con tutti quelli analizzati precedentemente, e che mirano a migliorarne le performance o a semplificare la complessità di calcolo.

Low-rank approximation Sino ad ora abbiamo considerato come modello matematico di riferimento quello del grafo (introdotto in 1). Tale modello può essere espresso anche per mezzo di matrici di adiacenza: tutti gli algoritmi di Link Prediction, mostrati in precedenza, hanno una formulazione equivalente per tale modellazione.

La scelta di utilizzare matrici di adiacenza è spesso dettata da motivi, inerenti alla riduzione di complessità computazionale, che alcune operazioni su matrici consentono.

Nel caso d'uso di matrici di grandi dimensioni è tecnica comune quella di "ridurre" l'analisi su una matrice di rango k (con k sufficientemente piccolo) in modo da ridurre l'occupazione della matrice in memoria considerando solamente un sottoinsieme significativo della matrice di partenza. Questa riduzione può essere effettuata in modo efficiente tramite SVD (singular value decomposition).

Unseen bigrams Il problema di Link Prediction è simile a quello di predire la frequenza di "coppie non presenti" nella modellazione di un linguaggio - coppie di parole che occorrono contemporaneamente all'interno di un testo, ma non nel testo utilizzato per effettuare il training.

Supponiamo di avere una misura di score (x, y) già computata da uno dei predittori introdotti precedentemente: l'idea è quella di utilizzare tale score per predire score (y, z) per un nodo z simile ad x . Nello specifico definiamo $S_x^{(\delta)}$ come l'insieme dei δ nodi più simili ad x secondo score (x, \cdot) , per $\delta \in \mathbb{Z}^+$. Possiamo quindi definire il nuovo score (nel caso di grafo con archi non pesati) come:

$$\text{score}(x, y) = |\{z : a \in \Gamma(y) \cap S_x^{(\delta)}\}| \quad (3.1.7)$$

3.1.2.2 Modelli Supervised

Nella categoria dei modelli supervised (vari approcci sono stati presentati in [14, 35, 28, 31, 37, 25, 12, 18]) possiamo riconoscere due particolari sottoclassi di algoritmi:

- quelli che estraggono il modello evolutivo dalla topologia della rete e lo utilizzano come informazione per effettuare le predizioni;

- e quelli che affiancano al task di Link Prediction altri approcci (solitamente di *Classification* o di *Clustering*) in modo da sfruttare informazioni contestuali alla rete ulteriori a quelle topologiche.

Introduciamo queste due tipologie di analisi presentando le idee che stanno alla base di tali approcci.

ESTRARRE IL MODELLO DALLA RETE: GERM E LFR

I predittori sino ad ora introdotti assumono, in larga parte, un unico proprio modello evolutivo per tutte le reti analizzate: le performance che questi ottengono sono chiaramente legate al grado in cui tale modello rispecchia quello effettivo seguito dalle singole reti.

Per riuscire a meglio sfruttare le informazioni fornite dalla rete è possibile tentare di inferire, nel modo più preciso possibile, le regole che ne fissano il reale modello evolutivo e utilizzarle per guidare il task di Link Prediction. In questa ottica possiamo far riferimento a due algoritmi (GERM e LFR [2, 15]) nati, non specificatamente per affrontare il problema trattato, ma che presentano interessanti spunti di contatto con la tipologia di approccio appena introdotta.

GERM (Graph Evolution Rule Miner), e la sua variante etichettata (LFR), rappresentano due strumenti nati per studiare l’evoluzione di una rete partendo dalle regole evolutive che possono essere estratte dalla sua analisi topologica. Il procedimento predittivo risulta quindi essere guidato da informazioni fornite dalla rete e non da assunzioni effettuate su di essa dallo specifico algoritmo.

CLASSIFICATION E CLUSTERING

Alcuni articoli hanno rilevato che, diverse tipologie di task affrontati nell’ambito del Data Mining, possono trarre benefici se applicate in modo congiunto. Nel caso specifico del Link Prediction alcune ricerche [4, 10, 36] hanno rilevato che affiancare in modo iterativo applicazioni di algoritmi di classificazione (o, analogamente, di clustering) e algoritmi di Link Prediction (di tipo unsupervised) può migliorare in modo sensibile la performance finale rispetto all’applicazione dei soli modelli predittivi.

L’intuizione che si pone alla base di questo particolare approccio è semplice. Come abbiamo visto (2.2.1.3), spesso, molte delle reti analizzate presentano un pattern che abbiamo definito Community Structure: l’applicazione di algoritmi di clustering\classificazione fanno emergere informazioni topologiche, relative alle community presenti nel grafo analizzato, che possono essere utilizzate come guida agli approcci di link prediction visti precedentemente.

Questo scambio di informazioni tra i due algoritmi utilizzati può essere esteso in modo iterativo così da raffinare i risultati di entrambi sfruttando, in modo reciproco, i risultati intermedi computati ad ogni passo.

3.1.3 *Limiti degli approcci classici*

Come illustrato esistono diverse tipologie di approcci al problema trattato: ciascuna di esse sfrutta le informazioni fornite dalla topologia del grafo analizzato (prediligendone alcune a seconda del modello evolutivo che assume o che inferisce) in modo da stimare quali link andranno ad aggiungersi a quest'ultimo in un ben determinato arco temporale.

Alcuni articoli (ad esempio [29]) hanno rilevato che tali approcci sono sensibili ad un analisi arricchita da informazioni temporali. I predittori introdotti, infatti, hanno il grosso limite di non riuscire a cogliere in pieno la storia evolutiva della rete analizzata poiché basano la loro analisi su uno snapshot temporale rappresentante il passato “appiattito” su di un unico istante temporale t .

Un altro limite compare nel caso si decida di arricchire di informazioni la struttura della rete analizzata (introducendo la multidimensionalità): nessuno dei modelli introdotti è abbastanza flessibile da poter sfruttare le informazioni topologiche aggiuntive rendendo, in molti casi, impossibile il task di predizione se non in modo approssimato.

Per questi motivi, nel caso in cui si decida di modellare la realtà mantenendo un insieme maggiore di informazioni (sfruttando, nello specifico, le nozioni di multidimensionalità e temporalità quando applicabili), gli approcci di Link Prediction devono essere rivisti per affrontare, compiutamente, la nuova definizione del task in modo da superare i limiti che caratterizzano l'approccio classico sino ad ora presentato.

3.2 LINK PREDICTION MULTIDIMENSIONALE

Nella precedente sezione abbiamo introdotto il problema di Link Prediction e si sono mostrate le principali tipologie di approcci proposti, sino ad ora, in letteratura evidenziandone i maggiori limiti.

In questa sezione si estende la formulazione proposta a pagina 26 in modo da dettagliare, in modo soddisfacente, il problema specifico per cui verranno introdotti gli approcci analizzati in 3.2.3. Per meglio comprendere le soluzioni proposte si introducono inoltre (3.2.2) le misure multidimensionali adottate estensivamente nei modelli presentati.

3.2.1 *Formulazione del problema*

Abbiamo già visto nel Capitolo 2 come sia possibile modellare reti reali, mantenendo un maggiore livello di dettaglio sulle interazioni rappresentate, tramite il modello fornito dai multigrafi: enunciamo adesso una formulazione del problema di Link Prediction che tenga conto di tali informazioni.

Definition 8. (Link Prediction multidimensionale)

Sia $G_0 = (V, E_0, D)$ un grafo multidimensionale, non orientato, definito nelle sue componenti di nodi, archi e dimensioni osservato ad un dato istante t_0 . Il task di Link Prediction multidimensionale, dato un istante $t_1 > t_0$ e il relativo insieme degli archi E_1 , consiste nel prevedere gli archi che entreranno a far parte del grafo originario nell'istante futuro t_1 (archi appartenenti all'insieme $E_{\text{new}} = E_1 - E_0$) e la specifica dimensione a cui essi apparterranno. Per ogni arco - identificato dalla tripla $(\text{nodo}, \text{nodo}, \text{dimensione})$ - nell'insieme dei risultati deve inoltre essere presente uno "score" di confidenza.

L'estensione proposta evidenzia la necessità, perché il task sia ultimato in modo corretto, di includere nei modelli di Link Prediction multidimensionali un modo per discriminare la dimensione di appartenenza di ciascun arco predetto. Per riuscire ad ottenere tale informazione è necessario definire un insieme di funzioni di misura atte a stimare quale sia la correlazione tra le dimensioni del grafo in modo da valutare la "probabilità" che, per ogni coppia di nodi predetta, una determinata dimensione sia preferibile rispetto alle altre per il formarsi di un arco.

Data la natura del problema, il termine “probabilità” deve essere inteso in senso lato: la somma dei valori delle stime proposte per ciascuna dimensione, data una determinata coppia di nodi tra cui debba instaurarsi una arco, non è infatti necessariamente uguale ad 1. Tale risultato è accettabile poiché le informazioni fornite dalle singole dimensioni raramente si rivelano mutuamente esclusive: spesso alcune dimensioni forniscono ridondanza di informazione poiché si ha la possibilità di *overlapping* tra gli insiemi degli archi appartenenti a ciascuna di esse.

Come vedremo nei prossimi paragrafi, abbiamo proposto diverse metodologie per introdurre tali informazioni di correlazione ed anticorrelazione tra le dimensioni nel processo di Link Prediction.

Il secondo limite evidenziato dai modelli predittivi presentati in letteratura è, come sottolineato in 3.1.3, l’insufficiente rappresentazione (ed uso) nel modello delle informazioni descriventi la storia evolutiva della rete utilizzata come test set. Sino ad ora abbiamo considerato il nostro grafo di partenza definito in un singolo istante temporale t : per fare questo ci siamo posti in una situazione, semplificata, in cui la rete è definita come una rappresentazione istantanea che non tiene conto dell’ordine in cui gli archi presenti sono venuti a formarsi.

Perdere questo tipo di informazione temporale può, per alcune reti, rendere il modello troppo semplice per un’analisi dettagliata di Link Prediction non cogliendo, ad esempio, interazioni tra nodi che avvengono con una specifica cadenza, oppure, quale sia il peso di un’interazione in relazione alla sua data di apparizione.

Per capire meglio quale sia l’importanza dell’informazione temporale in un’analisi di tipo predittivo si pensi di dover prevedere la posizione di una pallina da tennis, ad un secondo da adesso, avendo come unica informazione una foto che la mostri nell’istante in cui questa ha iniziato ad allontanarsi dalla racchetta. E’ chiaro che la posizione della pallina può essere in qualche modo predetta con una certa approssimazione ma, senza dubbio, avere un maggior numero di informazioni su come questa si è spostata nella sua traiettoria sino ad ora può rendere tale predizione ancora più accurata.

In altre parole è necessario conoscere la “velocità”, e il modo, in cui una rete si sviluppa e non solo la sua posizione finale: introducendo queste informazioni (ed utilizzando quindi il modello presentato a pagina 14) il problema originario può essere esteso come segue:

Definition 9. (Link Prediction multidimensionale con informazioni temporali)

Sia $G = (V, E, D, T)$ un grafo multidimensionale ed evolutivo non orientato definito nelle sue componenti di nodi, archi, dimensioni e istanti temporali. Il task di Link Prediction, dato

un istante $t_1 > \max\{t : t \in T\}$ e il relativo insieme degli archi E_1 , consiste nel prevedere gli archi che entreranno a far parte del grafo originario nell'istante futuro t_1 (archi appartenenti all'insieme $E_{new} = E_1 - E$) e la specifica dimensione in cui essi appariranno tenendo conto della storia evolutiva della rete espressa dalle informazioni temporali associate agli archi in E . Per ogni arco - identificato dalla tripla (nodo, nodo, dimensione) - nell'insieme dei risultati deve inoltre essere presente uno "score" di confidenza.

Come già sottolineato, una volta introdotto il modello fornito dal multigrafo con informazioni temporali, non esiste una gerarchia informativa tra le informazioni di tipo temporale e multidimensionale.

Due diversi approcci al problema di Link Prediction possono essere valutati (sia per la definizione classica del problema che per quella esclusivamente multidimensionale): può essere sensato, infatti, decidere di escludere o meno dai risultati tutti gli archi già appartenenti al grafo di partenza. La prima scelta è giustificabile poiché, nella formulazione classica, non avendo informazioni temporali associate agli archi, è possibile pensare di restringere la predizione ai soli archi non presenti nel training set: si può assumere che l'informazione data da un arco, già facente parte della rete, che venga predetto come candidato ad una nuova comparsa sia solo ridondante poiché non modifica la topologia della rete.

Nel caso specifico in cui ci poniamo questa fonte di indecisione sull'insieme dei risultati da ritornare non si presenta: le informazioni di ricorrenza di un arco sono infatti un elemento significativo, che influenza sulla probabilità che questo si ripresenti in futuro, perciò tale arco deve essere necessariamente considerato nell'insieme dei possibili risultati (inoltre la topologia della rete subirebbe modifiche tramite tale aggiunta a differenza dei casi proposti precedentemente).

3.2.2 *Misure multidimensionali*

Si presentano alcuni adattamenti delle funzioni di misura monodimensionali proposte in 3.1.1 per l'analisi dei grafi multidimensionali; si introducono inoltre ulteriori misure, esclusivamente multidimensionali, per valutare la correlazione/anticorrelazione tra insiemi di dimensioni appartenenti alla rete.

Tutte la funzioni di misura proposte sono tratte da [3] in cui viene presentato un framework per l'analisi di reti multidimensionali.

DEGREE

Come presentato in 3.1.1 la funzione di Degree restituisce il numero di archi incidenti al nodo su cui è applicata. Nel caso monodimensionale tale valore è equivalente al numero dei vicini al nodo stesso: nel caso multidimensionale questo non risulta più essere vero.

Nel caso multidimensionale infatti il degree di un nodo, dati n vicini, è un valore compreso tra $[0, n \cdot |D|]$ dove D è l'insieme delle dimensioni del grafo. Per tale motivo può aver senso introdurre due ulteriori misure:

Degree Set La funzione che computa il numero di vicini in uno specifico insieme di dimensioni

$$\text{Degree}_{\text{set}}(v, D) = |\{(u, v, d) \in E \text{ s.t. } u \in V \wedge d \in D\}| \quad (3.2.1)$$

Average Degree La funzione che computa il Degree medio di uno specifico insieme di dimensioni rispetto al totale delle dimensioni

$$\text{AvgDegree}(v, D) = \frac{\text{Degree}_{\text{set}}(v, D)}{|D|} \quad (3.2.2)$$

NEIGHBOURS

La funzione Neighbours applicata ad un nodo restituisce, come abbiamo visto, l'insieme dei vicini al nodo stesso. Introducendo una discriminante sulla tipologia degli archi che possono connettere due nodi è facile riuscire ad immaginare varianti di tale funzione che esprimano una maggiore informazione sulla topologia della rete.

Neighbours Set Una prima variante è quella che definiamo come:

$$\text{Neighbours}_{\text{Set}}(v, D) = \{u \in V \text{ s.t. } \exists (u, v, d) \in E \wedge d \in D\} \quad (3.2.3)$$

Il risultato dell'applicazione di tale funzione è una restrizione dell'insieme dei vicini al nodo v connessi tramite archi appartenenti alle dimensioni dell'insieme D .

Neighbours Xor Una funzione più utile, poiché tiene conto la correlazione tra le dimensioni degli archi incidenti al nodo analizzato, è quella fornita da Neighbours_{Xor}:

$$\text{Neighbours}_{\text{Xor}}(v, D) = \sum_{u \in V} k_{uv}(D) \quad (3.2.4)$$

dove:

$$k_{uv}(D) = \begin{cases} 1 & \text{se } \forall (u, v, d) \in E : d \in D \\ 0 & \text{altrimenti} \end{cases}$$

Con questa funzione si individua l'insieme dei vicini raggiungibili dal nodo v esclusivamente tramite archi appartenenti all'insieme delle dimensioni specificato con D e non raggiungibili tramite archi etichettati con altre dimensioni.

Questa funzione, come vedremo, sarà usata estensivamente per rendere multidimensionali alcuni dei modelli presentati per il caso monodimensionale.

DIMENSION RELEVANCE

Dovendo lavorare con reti multidimensionali è interessante analizzare l'importanza che una dimensione (o un insieme di dimensioni) assume per la connettività di un nodo. Per fare questo si fa uso della misura di DimensionRelevance: questa analizza in che grado il nodo a cui è applicata viene disconnesso dalla rete se vengono eliminate da questa le dimensioni specificate.

Si propongono nel seguito tre varianti di tale funzione definita in $V \times D \rightarrow [0, 1]$.

Dimension Relevance Una prima definizione è la seguente:

$$\text{DimensionRelevance}(v, D) = \frac{\text{Neighboresset}(v, D)}{\text{Neighbours}(v)} \quad (3.2.5)$$

Si calcola in questo caso il rapporto tra i vicini al nodo appartenenti al sottoinsieme delle dimensioni D rispetto al totale dei vicini su tutte le dimensioni.

Dimension Relevance Xor Questa variante fa uso della funzione $\text{Neighbours}_{\text{Xor}}$, già introdotta, per sfruttare le informazioni di correlazione tra le dimensioni.

$$\text{DimensionRelevance}_{\text{Xor}}(v, D) = \frac{\text{Neighbours}_{\text{Xor}}(v, D)}{\text{Neighbours}(v)} \quad (3.2.6)$$

Si calcola il rapporto tra il numero dei vicini raggiungibili esclusivamente tramite archi appartenenti al sottoinsieme delle dimensioni D e il totale dei vicini al nodo.

Dimension Relevance Weighted Quest'ultima variante è una versione pesata che tiene conto del numero di alternative (il numero degli archi appartenenti alle dimensioni non appartenenti all'insieme specificato) per raggiungere un nodo.

$$\text{DimensionRelevance}_W(v, D) = \frac{\sum_{u \in \text{Neighbours}_{\text{Set}}(v, D)} \frac{n_{uvd}}{n_{uv}}}{\text{Neighbours}(v)} \quad (3.2.7)$$

dove:

- n_{uvd} è il numero di dimensioni che etichettano gli archi tra due nodi u e v e che appartengono a D;
- n_{uv} è il numero di dimensioni che etichettano gli archi tra due nodi u e v .

TOTALLY MIXED E TOTALLY SPLITTED

Le misure di Totally Splitted e Totally Mixed definiscono la struttura e la densità degli archi incidenti al dato nodo analizzato.

Tali nozioni, derivate dalla combinazione delle funzioni Degree e Neighbours, definiscono rispettivamente un nodo come Totally Mixed se la struttura degli archi incidenti è densa e ridondante e Totally Splitted se, al contrario, questa si presenta sparsa.

Entrando nel dettaglio:

- Un nodo si definisce Totally Splitted se ciascuno dei suoi vicini è raggiungibile esclusivamente tramite una singola dimensione

$$\forall u \in \text{Neighbours}_{\text{Set}}(v, D) : \exists! d \in D \ (u, v, d) \in E \quad (3.2.8)$$

- Un nodo si definisce **Totally Mixed** se ciascuno dei suoi vicini è raggiungibile tramite tutte le dimensioni presenti nella rete

$$\forall u \in \text{Neighbours}_{\text{Set}}(v, D) : \forall d \in D \ (u, v, d) \in E \quad (3.2.9)$$

Si può notare che se un nodo è **Totally Splitted** si ha $\text{Degree}(v, D) = \text{Neighbours}(v, D)$ mentre nel caso questo sia **Totally Mixed** al contrario si verifica che $\text{Degree}(v, D) = \text{Neighbours}(v, D) \times |D|$.

DIMENSION DEGREE

Un'analisi interessante sulla struttura di una rete multidimensionale è quella che ci consente di capire quale sia la percentuale di nodi e archi contenuti in una specifica dimensione. Per questo motivo si introducono due funzioni chiamate Dimension Degree, definite in $D \rightarrow [0, 1]$.

Edge Dimension Degree La funzione di Edge Dimension Degree calcola il rapporto tra gli archi della rete etichettati con la dimensione d rispetto al totale degli archi.

$$\text{EDD}(d) = \frac{|\{(u, v, d) \in E : u, v \in V\}|}{|E|} \quad (3.2.10)$$

Node Dimension Degree Analogamente la funzione di Node Dimension Degree calcola la frazione dei nodi della rete appartenenti alla dimensione d .

$$\text{NDD}(d) = \frac{|\{u \in V : \exists v \in V \ (u, v, d) \in E\}|}{|V|} \quad (3.2.11)$$

DIMENSION PARENT

Un'importante tipologia di relazione tra due dimensioni che è necessario tenere in considerazione è quella che esprime quanto una di esse "includa" l'altra.

Questa relazione che stabilisce implicitamente una gerarchia tra le dimensioni è chiamata Parent ed è definita in $D \times D \rightarrow [0, 1]$.

Edge Parent La funzione Edge Parent computa la percentuale degli archi appartenenti ad una dimensione d_1 che appartengono anche alla dimensione d_2 .

$$\text{EdgeParent}(d_1, d_2) = \frac{|E_{d_1} \cap E_{d_2}|}{|E_{d_1}|} \quad (3.2.12)$$

Node Parent La funzione Node Parent computa la percentuale dei nodi appartenenti ad una dimensione d_1 che appartengono anche alla dimensione d_2 .

$$\text{NodeParent}(d_1, d_2) = \frac{|V_{d_1} \cap V_{d_2}|}{|V_{d_1}|} \quad (3.2.13)$$

DIMENSION CORRELATION

Le seguenti misure sono importanti per cogliere la correlazione tra due dimensioni: l'idea che viene colta è la ridondanza espressa². L'ambito di definizione della funzione è ancora $D \times D \longrightarrow [0, 1]$.

Edge Correlation Questa funzione calcola il rapporto tra il numero di archi appartenenti alle due dimensioni rispetto al totale degli archi appartenenti ad almeno una di esse.

$$\text{EdgeCorrelation}(d_1, d_2) = \frac{|E_{d_1} \cap E_{d_2}|}{|E_{d_1} \cup E_{d_2}|} \quad (3.2.14)$$

Node Correlation Questa funzione calcola il rapporto tra il numero di nodi appartenenti ad entrambe le dimensioni prese in esame rispetto al totale dei nodi appartenenti ad almeno una di esse.

$$\text{NodeCorrelation}(d_1, d_2) = \frac{|V_{d_1} \cap V_{d_2}|}{|V_{d_1} \cup V_{d_2}|} \quad (3.2.15)$$

² La correlazione viene calcolata tramite un Jaccard tra le due dimensioni.

3.2.3 Approcci adottati

Una volta fissate le misure applicabili ad un modello multidimensionale è possibile definire alcune tipologie di approcci al problema di Link Prediction (per come questo è stato definito a pagina 36).

In questa trattazione si è deciso di applicare due diverse tipologie di approcci:

1. Introdurre modificatori multidimensionali (e temporali) in predittori già noti in letteratura ed appartenenti alla classe dei modelli unsupervised basati su informazioni di neighborhoods (3.1.2.1);
2. Creare nuovi predittori basati esclusivamente su informazioni di tipo multidimensionale e temporale escludendo gli approcci presenti in letteratura.

Nelle prossime sezioni si presentano nel dettaglio i modelli proposti e valutati nei capitoli dedicati all'analisi sperimentale in modo da dare al lettore una visione d'insieme il più dettagliata possibile.

3.3 MODELLI DERIVATI DA ALGORITMI MONODIMENSIONALI

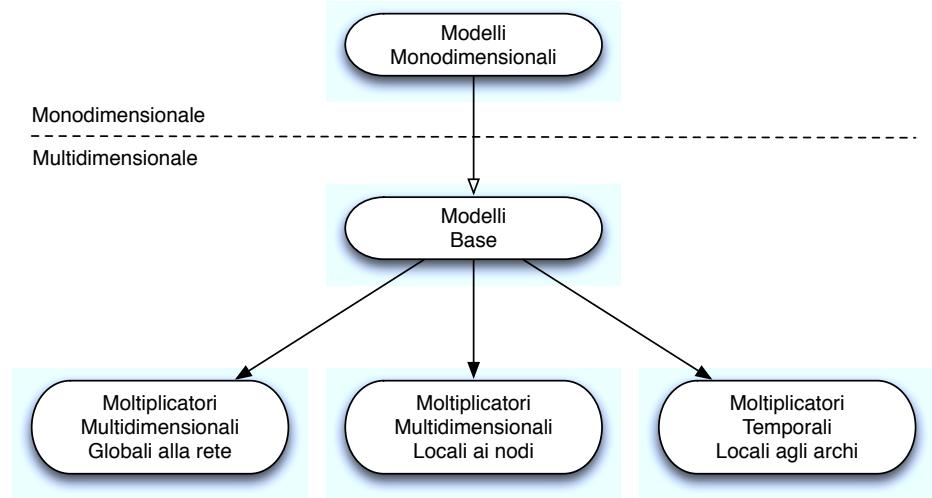


Figura 5: Modelli derivati da approcci monodimensionali

In questa sezione si introducono i modelli predittivi derivati dalla teoria nota per il caso monodimensionale: gli approcci proposti sono stati suddivisi, in modo gerarchico, in quattro insiemi (come mostrato in Figura 5). Si presenteranno nuovamente, in breve, i quattro modelli monodimensionali che sono stati utilizzati come base di analisi (3.3.1): si vedrà quindi come possono essere introdotte, durante la fase predittiva, le informazioni multidimensionali localmente ai nodi/archi (3.3.2) e globalmente alla rete (3.3.3). Inoltre sarà mostrato un insieme di moltiplicatori utilizzati per arricchire, con informazioni temporali, l'analisi dei modelli presi come base d'analisi (3.3.4).

Il totale dei modelli predittivi analizzati consta di: 4 modelli base, 6 predittori multidimensionali locali, 6 moltiplicatori multidimensionali globali, 6 moltiplicatori temporali. Il totale dei predittori sale quindi a 58 (i moltiplicatori globali e temporali possono essere applicati a ogni modello base portando quindi il numero di tali sottocategorie di predittori a 24 elementi).

3.3.1 *Modelli derivati di base*

I modelli predittivi unsupervised basati su neighborhood presi in considerazione per essere resi utilizzabili in ambito multidimensionale sono: Common Neighbours, Jaccard, Adamic Adar e

Preferential Attachment.

Tutti i modelli proposti fanno uso, nella loro formulazione e nella loro implementazione, delle misure di Neighbours e Degree: come abbiamo visto in 3.2.2 tali misure possono essere adattate per un'analisi di tipo multidimensionale. Nello specifico si è deciso di sostituire alla funzione Neighbours (v) la versione Neighbours_{XOR} (v, d) in modo da poter calcolare, dati due nodi ed una dimensione, lo score per l'arco definito da questa tripla tenendo implicitamente conto, seppur in modo basilare, dell'importanza relativa di tale dimensione per la coppia di nodi presi in esame.

Nel caso del Preferential Attachment tale informazione è sfruttata a livello implementativo non comparendo esplicitamente nella formula che esprime il modello predittivo.

3.3.2 *Moltiplicatori locali*

Una prima strada, per introdurre più specifiche informazioni di tipo multidimensionale all'interno del processo di calcolo degli score dei modelli base, (che, come abbiamo visto, sono già stati estesi per un' analisi multidimensionale tramite la sostituzione della funzione Neighbours (v) con la versione Neighbours_{XOR} (v, d)) è quella di introdurre informazioni multidimensionali implicite (coefficiente Mix/Split, rapporto multi-monodimensionale, rapporto Neighbours_{XOR} /Neighbours_{SET}) o esplicite (Dimension Relevance) che possono essere calcolate localmente alla coppia di nodi e alla dimensione di volta in volta prese in analisi.

Si presentano, nei successivi paragrafi, i sei modelli multidimensionali locali analizzati.

ADAMIC ADAR CON DIMENSION RELEVANCE XOR

In questo modello si attribuisce un maggior peso sullo score finale ad i vicini comuni ai nodi x e y che sono soliti stabilire legami nella dimensione d .

$$\text{AADR}(x, y, d) = \sum_{z \in \Gamma_{\text{Xor}}(x, d) \cap \Gamma_{\text{Xor}}(y, d)} \frac{1}{\log(|\Gamma_{\text{Xor}}(z, d)| * \text{DR}_{\text{Xor}}(z, d))} \quad (3.3.1)$$

ADAMIC ADAR CON COEFFICIENTE MIX/SPLIT

In questa variante di Adamic Adar si sommano i coefficienti di Split/Mixed dei vicini comuni (in scala logaritmica) in modo da ricavare un infromazione sulla struttura della rete

nell'intorno dei nodi x ed y .

$$\text{AAMixSplit}(x, y, d) = \sum_{z \in \Gamma_{\text{Xor}}(x, d) \cap \Gamma_{\text{Xor}}(y, d)} \frac{\log(\text{Degree}_{\text{Set}}(z, d))}{\log(|\Gamma_{\text{Xor}}(z, d)|)} \quad (3.3.2)$$

Il valore minimo della sommatoria (è equivalente a $|\Gamma(x, d) \cap \Gamma(y, d)|$ ovvero al numero dei vicini comuni) delinea una conformazione sparsa della rete in comune a x ed y .

COMMON NEIGHBOURS RAPPORTO MULTI-MONODIMENSIONALE

Rapporto tra lo score di Common Neighbours calcolato utilizzando il relativo modello base multidimensionale e quello calcolato sul grafo ottenuto dall'appiattimento del multigrafo.

$$\text{CNMM}(x, y, d) = \frac{|\Gamma_{\text{Xor}}(x, d) \cap \Gamma_{\text{Xor}}(y, d)|}{|\Gamma(x) \cap \Gamma(y)|} \quad (3.3.3)$$

JACCARD RAPPORTO NEIGHBOURS XOR/NEIGHBOURS SET

In questo modello lo score per il predittore Jaccard è calcolato come rapporto tra i vicini comuni ai nodi x ed y raggiungibili esclusivamente tramite archi in d e l'insieme complessivo dei vicini di x ed y (sempre relativamente alla dimensione d).

$$\text{JaccardNX}(x, y, d) = \frac{|\Gamma_{\text{Xor}}(x, d) \cap \Gamma_{\text{Xor}}(y, d)|}{|\Gamma(x, d) \cup \Gamma(y, d)|} \quad (3.3.4)$$

PREFERENTIAL ATTACHMENT CON DIMENSION RELEVANCE XOR

In questo modello si utilizza la funzione $\text{DimensionRelevance}_{\text{Xor}}$ per dare un peso al numero dei vicini di ciascuno dei due nodi, valutando, per la dimensione di interesse, un valore di correlazione calcolato per il singolo nodo.

$$\text{PADR}(x, y, d) = (|\Gamma_{\text{Xor}}(x, d)| * \text{DR}_{\text{Xor}}(x, d)) * (|\Gamma_{\text{Xor}}(y, d)| * \text{DR}(y, d)) \quad (3.3.5)$$

PREFERENTIAL ATTACHMENT CON COEFFICIENTE MIX/SPLIT

L'informazione aggiuntiva che si ottiene considerando per ogni nodo il rapporto tra degree,

relativo alla dimensione d, e numero di vicini sulla stessa dimensione calcolati in Xor, è il coefficiente di Split\Mixed di un nodo.

Tale coefficiente è minimo (con valore uguale a 0) se le reti centrate sui nodi³ sono costituite solo da archi appartenenti alla dimensione d, è massimo (pari a $((|\Gamma_{Set}(x, d)| + 1) * (|\Gamma_{Set}(y, d)| + 1)) - 1$) nel caso in cui la dimensione d presa in considerazione sia totalmente ridondante rispetto alle altre dimensioni su cui i nodi hanno archi.

$$PAMS(x, y, d) = - \left[\left(\frac{\text{Degree}_{Set}(x, d) + 1}{|\Gamma_{Xor}(x, d)| + 1} * \frac{\text{Degree}_{Set}(y, d) + 1}{|\Gamma_{Xor}(y, d)| + 1} \right) - 1 \right] \quad (3.3.6)$$

Il segno meno davanti alla formula consente di mantenere una scala dei valori crescente per l'ordinamento dei risultati: i valori prossimi allo 0 corrispondono agli score preferibili.

3.3.3 Moltiplicatori globali

Dopo aver introdotto informazioni inerenti la multidimensionalità nel calcolo dello score, agendo in modo locale sui nodi di volta in volta candidati alla formazione di un arco, è interessante analizzare se, alternativamente, ha senso sfruttare misure globali alla rete per ottenere risultati simili o migliori partendo dagli stessi modelli predittivi di base.

Non esistendo in letteratura materiale che affronti il problema di Link Prediction per reti multidimensionali le tipologie di analisi che sono state proposte hanno come obiettivo quello di coprire, seppur in modo parziale, più tipologie di approcci in modo da fornire una base di partenza per valutare quale di questi sia preferibile, sia per le performance ottenute, sia per la complessità algoritmica.

Quelli a cui ci riferiamo con il nome di "Moltiplicatori Globali" sono da intendersi come particolari coefficienti calcolati, uno per ciascuna dimensione della rete, in base a criteri prestabiliti e utilizzati come fattore moltiplicativo per tutti gli score degli archi predetti appartenenti a ciascuna specifica dimensione del grafo.

Nei prossimi paragrafi si introducono sei moltiplicatori derivati dalle funzioni di misura multidimensionale introdotte in 3.2.2: dove specificato è stato calcolata una sommatoria, pesata sul numero totale di dimensioni della rete, della funzione multidimensionale presa in esame.

³ La rete centrata su di un nodo è chiamata ego-network.

NODE DIMENSION DEGREE

$$Ndd(d) = \frac{|\{(x \in V | \exists y \in V : (x, y, d) \in E\}|}{|V|} \quad (3.3.7)$$

Il moltiplicatore Ndd tiene conto del rapporto relativo al numero di nodi appartenenti a ciascuna dimensione rispetto al totale dei nodi del grafo per valutare l'importanza relativa tra le dimensioni.

EDGE DIMENSION DEGREE

$$Edd(d) = \frac{|\{(x, y, d) \in E | x, y \in V\}|}{|E|} \quad (3.3.8)$$

Il moltiplicatore Edd , analogamente a quanto fatto dal precedentemente introdotto Ndd , valuta l'importanza di ogni dimensione in base ad un rapporto: in questo caso si considerano gli archi appartenenti alla specifica dimensione sul totale degli archi appartenenti alla rete.

AVERAGE NODE PARENT

$$NP(d) = \frac{\sum_{s \in D} NodeParent(d, s)}{|D|} \quad (3.3.9)$$

La funzione $NodeParent(d_1, d_2)$ esplicita, come illustrato, il grado di inclusività della dimensione d_1 rispetto alla dimensione d_2 calcolata sugli insiemi dei nodi: questa funzione aggregata calcola una media che esprime il grado di inclusione tra la dimensione presa in esame d ed il resto delle dimensioni appartenenti alla rete. Tale moltiplicatore (come quelli presentati successivamente) assume valori nel dominio $\left[\frac{1}{|D|}, 1\right]$.

AVERAGE EDGE PARENT

$$EP(d) = \frac{\sum_{s \in D} EdgeParent(d, s)}{|D|} \quad (3.3.10)$$

La funzione $EdgeParent(d_1, d_2)$ esprime il grado di inclusività della dimensione d_1 rispetto alla dimensione d_2 calcolata sugli insiemi degli archi. Analogamente a quanto visto nel paragrafo precedente il moltiplicatore proposto non è altro che una media che valuta l'inclusività della dimensione considerata rispetto a tutte le altre.

AVERAGE NODE CORRELATION

$$NC(d) = \frac{\sum_{s \in D} NodeCorrelation(d, s)}{|D|} \quad (3.3.11)$$

Le funzioni di correlazione tra dimensioni calcolano un Jaccard tra gli insiemi di volta in volta presi in esame (nel caso di `NodeCorrelation` l'insieme dei nodi) appartenenti alle dimensioni passate come argomenti. Utilizzando un procedimento analogo a quello proposto per le funzioni `Parent` si calcola per ogni dimensione la media della correlazione tra questa e le altre dimensioni della rete.

AVERAGE EDGE CORRELATION

$$EC(d) = \frac{\sum_{s \in D} EdgeCorrelation(d, s)}{|D|} \quad (3.3.12)$$

Per calcolare tale coefficiente moltiplicativo si applicano i concetti visti precedentemente ottenendo una media della correlazione tra la dimensione passata in input e le altre dimensioni della rete calcolata sugli insiemi degli archi aventi tali etichette.

3.3.4 *Moltiplicatori temporali*

L'analisi sino ad ora proposta si è focalizzata, esclusivamente, su modelli di tipo multidimensionale senza tenere in considerazione le informazioni di tipo evolutivo fornite dalla rete. Per valutare l'importanza di tali informazioni, si è deciso di introdurre alcuni moltiplicatori che quantificassero la ricorrenza, per ciascuna coppia di nodi, delle interazioni presenti nella rete da analizzare. Tali moltiplicatori sono applicati agli archi restituiti come risultati, dell'analisi predittiva, condotta tramite i modelli base.

Diversamente da i moltiplicatori introdotti in 3.3.3 quelli di seguito presentati, non sono da considerarsi globali alla rete ma locali agli archi predetti. Alcuni approcci all'analisi temporale per il problema di Link Prediction sono stati proposti in passato per reti monodimensionali [29]: per mantenere semplice il modello si è deciso di calcolare misure temporali non complesse e rappresentanti ben distinguibili pattern di ricorsività.

Di seguito si presentano tre funzioni di misura: tutte presentano due varianti, una in cui la ricorsione, per ogni arco, è valutata sulla singola dimensione e l'altra in cui è considerata sul

totale delle dimensioni in cui questo compare. Per completezza si riportano entrambi i modelli per ogni misura.

FREQUENCY

$$\text{Freq}(x, y, d) = |\{(x, y, d, t) : (x, y, d, t) \in E\}| \quad (3.3.13)$$

Il moltiplicatore Frequency calcola il totale degli istanti temporali in cui un determinato arco è comparso nella rete in una specifica dimensione.

OVER ALL FREQUENCY

$$\text{OAFreq}(x, y) = |\{(x, y, d, t) : (x, y, d, t) \in E\}| \quad (3.3.14)$$

Il moltiplicatore Over All Frequency calcola il totale degli istanti temporali in cui un determinato arco è comparso nella rete.

WEIGHTED PRESENCE

$$W\text{pres}(x, y, d) = \sum_{e \in \{(x, y, d, t) : (x, y, d, t) \in E\}} \Pi_t(e) \quad (3.3.15)$$

Il moltiplicatore Weighted Presence calcola il totale degli istanti - pesato in base all'ordine temporale - in cui un determinato arco è comparso nella rete per la dimensione specificata.

Archi comparsi nella rete in istanti recenti hanno peso maggiore. La funzione $\Pi_t(e)$ ($[E \rightarrow \mathbb{N}]$) definisce la proiezione del valore di t (rappresentante l'intervallo temporale) rispetto all'arco e a cui è applicata.

OVER ALL WEIGHTED PRESENCE

$$\text{OAW}\text{pres}(x, y) = \sum_{e \in \{(x, y, d, t) : (x, y, d, t) \in E\}} \Pi_t(e) \quad (3.3.16)$$

Il moltiplicatore Over All Weighted Presence calcola il totale degli istanti - pesato in base all'ordine temporale - in cui un determinato arco è comparso nella rete.

MAX DOUBLE OCCURRENCE

Il moltiplicatore MDO è una variante di Weighted Presence in cui è calcolata una sommatoria, sempre pesata in base all'ordine temporale di occorrenza delle interazioni, che tiene conto solo delle coppie di archi occorsi in istanti temporali adiacenti.

Consideriamo il vettore $v = [0, 1, 1, 0, 1, 1, 0]$ come rappresentante la cronologia delle interazioni tra due nodi in una specifica dimensione d (dove 0 indica l'assenza di interazione e 1 la presenza): MDO in questo caso valuta nella sommatoria la presenza di due valori diversi da 0, uno per gli istanti ($v[1], v[2]$) e uno per ($v[4], v[5]$). I valori di presenza calcolati sono pesati in base al valore dell'istante temporale rappresentante il secondo componente di ciascuna coppia (nel caso dell'esempio MDO $(x, y, d) = 2 + 5 = 7$).

OVER ALL MAX DOUBLE OCCURRENCE

In questa variante, come accade per OAWpres e OAFreq, il valore di MDO è calcolato senza effettuare discriminazioni sulle dimensioni di appartenenza degli archi.

3.4 MODELLI ADHOC

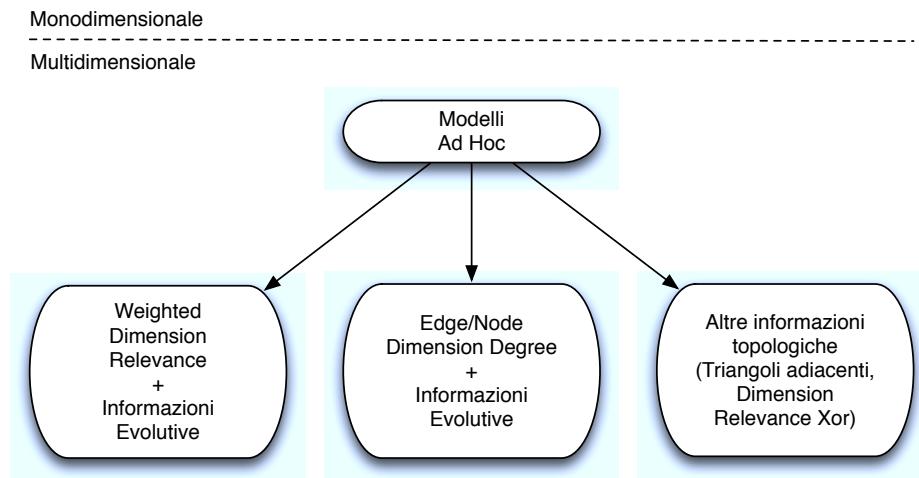


Figura 6: Modelli AdHoc

Nella sezione precedente sono stati introdotti i predittori sviluppati partendo da modelli noti in letteratura per il problema monodimensionale: da tali modelli si sono ottenuti, tramite varie tipologie di interventi, algoritmi capaci di sfruttare informazioni multidimensionali e temporali fornite dalla rete.

Nei successivi paragrafi si vuole presentare un approccio alternativo al problema affrontato: si propongono modelli che utilizzino solo, ed esclusivamente, funzioni di misura definite per catturare informazioni multidimensionali e temporali non considerando, quindi, approcci già noti in letteratura.

Ovviamente, come per le categorie di predittori precedentemente presentate, non si considera la trattazione presentata esaustiva di tutti i possibili modelli ideabili a partire dalle misure presentate in 3.2.2.

Come mostrato in Figura 6, si sono raggruppati i sei modelli proposti di seguito in tre gruppi: due predittori fanno uso della variante Weighted della funzione Dimension Relevance (arricchendone la definizione con le informazioni temporali rispettivamente di Frequency e di Weighted Presence), due sfruttano una versione locale dei coefficienti di Edd e Ndd (anche essi arricchiti con informazione temporale di Frequency) ed infine gli ultimi due sfruttano rispettivamente le funzioni di Dimension Relevance Xor e una misura che calcola il numero di triangoli adiacenti ai nodi candidati alla formazione dell'arco.

LOCAL EDGE DIMENSION DEGREE & FREQUENCY

La funzione di Local Edge Dimension Degree è definibile analogamente a quanto visto in 3.2.2 se, al posto di considerare l'insieme degli archi appartenenti al grafo analizzato, si considerano esclusivamente gli archi incidenti al nodo a cui questa è applicata.

$$\text{LEdd}(x, y, d) = \frac{|\{(x, z, d, t) : z \in \Gamma(x)\}|}{|\{(x, z, s, t) : z \in \Gamma(x) \wedge s \in D\}|} + \left(\frac{|\{(x, z, d, t) : z \in \Gamma(x)\}|}{|\{(x, z, s, t) : z \in \Gamma(x) \wedge s \in D\}|} * \text{Freq}(x, y, d) \right) \quad (3.4.1)$$

LOCAL NODE DIMENSION DEGREE & FREQUENCY

Analogamente alla definizione data per LEdd si può definire LNdd nel seguente modo:

$$\text{LNdd}(x, y, d) = \frac{|\Gamma_{\text{Set}}(x, d)|}{|\Gamma(x)|} + \left(\frac{|\Gamma_{\text{Set}}(x, d)|}{|\Gamma(x)|} * \text{Freq}(x, y, d) \right) \quad (3.4.2)$$

DIMENSION RELEVANCE WEIGHTED & WEIGHTED PRESENCE

Questo modello fa uso della misura multidimensionale di Weighted Dimension Relevance e delle informazioni temporali fornite da Weighted Presence,

$$\text{WDR}_w(x, y, d) = \text{DR}_W(x, d) * \text{DR}_W(y, d) + (\text{DR}_W(x, d) * \text{DR}_W(y, d) * \text{Wpres}(x, y, d)) \quad (3.4.3)$$

DIMENSION RELEVANCE WEIGHTED & FREQUENCY

Questo modello fa uso della misura multidimensionale di Weighted Dimension Relevance e delle informazioni temporali fornite da Frequency,

$$\text{WDR}_f(x, y, d) = \text{DR}_W(x, d) * \text{DR}_W(y, d) + (\text{DR}_W(x, d) * \text{DR}_W(y, d) * \text{Freq}(x, y, d)) \quad (3.4.4)$$

DIMENSION RELEVANCE XOR

$$\text{AM}(x, y, d) = \left(1 - \frac{\sum_{i \in D} \text{DR}_{\text{Xor}}(x, i) + \sum_{i \in D} \text{DR}_{\text{Xor}}(y, i)}{2} \right) * \text{DR}(x, d) * \text{DR}(y, d) \quad (3.4.5)$$

Usare i valori di DimensionRelevance_{X_{or}} permette di individuare i casi in cui una dimensione, già presente tra i due nodi x e y , influenzi, positivamente o negativamente, la probabilità che si stabilisca un arco in un'altra dimensione attualmente non presente. L'1 - DimensionRelevance_{X_{or}} serve a valutare, dati x e y , quanto questi nodi tendano ad avere dimensioni alternative per i loro vicini (1 tutti i loro DimensionRelevance_{X_{or}} sono 0 quindi hanno sempre un'alternativa, 0 la somma dei loro DimensionRelevance_{X_{or}} è 1 quindi sono Totally Split). Nella formula si considera anche quanto la dimensione d è effettivamente importante per i due nodi.

ADJACENT TRIANGLE

$$\text{Triangle}(x, y, d) = \text{Clique}_3(x, d) * \text{Clique}_3(y, d) \quad (3.4.6)$$

Questo modello, l'ultimo introdotto, calcola lo score da attribuire a ciascuna tripla in base al prodotto del numero di clique di tre (triangoli) che, nella dimensione specificata, hanno tra i vertici uno dei due nodi. Questa misura è di facile computazione e riesce a sfruttare informazioni topologiche (seppur non temporali) non ancora analizzate dagli altri modelli proposti.

Parte III

ANALISI SPERIMENTALE

4

METODOLOGIA DI ANALISI

Dopo aver introdotto il problema di Link Prediction (a pagina 26) ed esteso lo stesso per un'analisi multidimensionale (a pagina 35) arricchita da informazioni temporali (a pagina 36) si introducono in questo capitolo i dataset utilizzati per valutare le performance dei modelli predittivi introdotti (4.1).

Successivamente (4.2) si illustra la metodologia di analisi e presentazione dei risultati utilizzata nel Capitolo 5 per comparare i modelli valutati.

4.1 PRESENTAZIONE DEI DATASET

Per effettuare l'analisi delle performance ottenibili dai predittori introdotti nel capitolo precedente, è stato necessario reperire in rete alcuni dataset con caratteristiche tali da consentire la costruzione (utilizzando tutti o solo una parte dei dati rappresentati) di reti multidimensionali che presentassero informazioni dettagliate sull'ordine temporale di comparsa degli archi.

Data la particolare complessità della modellazione introdotta non è stato possibile utilizzare reti già costruite: molte delle reti reperibili, infatti, denotano l'assenza delle informazioni topologiche relative alla multidimensionalità e, quand'anche presentino tali informazioni, non forniscono la componente temporale associata ai singoli archi rendendo impraticabile la tipologia di analisi precedentemente introdotta.

Nei seguenti paragrafi si introducono i sei dataset utilizzati per costruire le reti analizzate nel seguito: per ogni dataset sono fornite alcune specifiche della rete costruita e, laddove possibile¹, sono fornite le informazioni topologiche inerenti il coefficiente di clustering della rete e altre statistiche su di essa calcolate.

¹ Nel caso della rete costruita sui query log trimestrali di AOL le statistiche sulla rete non sono state calcolate a causa della dimensione, e densità, della rete stessa. Tutte le statistiche sono ottenute dall'analisi delle reti monodimensionali, costruite a partire dalle rispettive multidimensionali, tramite il plugin NetworkAnalyzer di Cytoscape.

AMERICA ONLINE QUERY LOG

I dati di partenza per la costruzione del dataset utilizzato sono un insieme di, approssimativamente, 20 milioni di queries effettuate da 650 000 utenti nel periodo di Marzo-Maggio del 2006 sul motore di ricerca “America On Line”.

<i>Proprietà</i>	<i>Valore</i>
$ V $	15 949
$ E_{old} $	1 369 533
$ E_{new} $	100 280
$ D $	6
PPV_{rand}	1,316 538 456 52 e-04

Tabella 1: Query Log Statistiche

Un record appartenente a tale log rappresenta la visita ad un sito effettuata in seguito alla ricerca sul motore. Ciascun record mantiene un Id anonimo che consente di raggruppare queries di uno stesso utente senza rivelarne l’identità, la query effettuata dall’utente, la data e l’ora in cui è avvenuta la richiesta, il rank della posizione del risultato scelto dall’utente ed il dominio del risultato scelto.

La rete multidimensionale costruita su tale dataset è una rete parola-parola (due nodi, rappresentanti due parole, sono collegati da un arco se compaiono entrambe in una stessa query) avente come dimensioni i rank del risultato scelto per la query (informazioni convenientemente raggruppate) e come informazioni temporali una partizione del lasso di tempo analizzato (esempio in Figura 7).

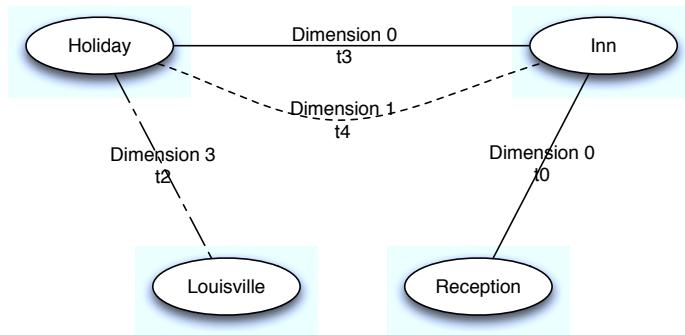


Figura 7: Esempio Query Log

Il dataset di training su cui è stata costruita la rete oggetto di analisi - Tabella 1 - comprende le parole accorse in ricerche effettuate nel periodo che va dal 1/04/2006 al 15/04/2006 (da cui sono state filtrate le stop-words e le parole con una frequenza inferiore alle 20 presenze).

Il dataset di test - Tabella 1 - è costituito dalle relazioni parola-parola accorse nel periodo che va dal 16/04/2006 al 18/04/2006. Gli intervalli temporali considerati per il training set sono 5 formati raggruppando le query ogni 3 giorni in uno stesso intervallo.

Le dimensioni utilizzate (ottenute a partire dal rank delle posizioni dei risultati selezionati per ogni query) sono 6 e sono state costruite sfruttando i seguenti insiemi equamente popolati: dimensione 1 per rank uguale a 1, dimensione 2 per i rank in [2 – 3], dimensione 3 per i rank in [4 – 6,] dimensione 4 per i rank in [7 – 10] , dimensione 5 per i rank in [11 – 58] e dimensione 6 per i rank in [59 – 500].

DBLP: COMPUTER SCIENCE BIBLIOGRAPHY

DBLP² è un database bibliografico che tiene traccia di tutte le pubblicazioni nell'ambito della Computer Science includendo conferenze, libri e articoli.

<i>Proprietà</i>	<i>Valore</i>	<i>Proprietà</i>	<i>Valore</i>
$ V $	30 178	Connected Components	1 630
$ E_{old} $	78 965	Network Diameter	22
$ E_{new} $	14 650	Clustering Coefficient	0,666
$ D $	28	Shortest Path	7, 241
PPV_{rand}	1,149 068 083 29 e-06		

Tabella 2: DBLP Statistiche

La rete multidimensionale costruita su tale database è una rete di co-authorship (due nodi, rappresentanti due autori, sono collegati da un arco se hanno partecipato entrambi alla stesura di uno stesso testo) avente come dimensioni alcune conferenze scelte tra le presenti e come informazioni temporali gli anni di pubblicazione dei testi.

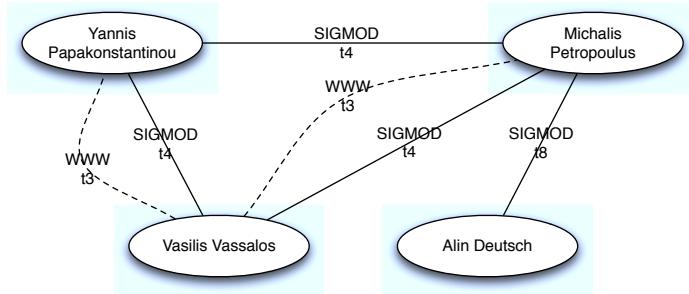


Figura 8: Esempio DBLP

Il dataset di training su cui è stata costruita la rete oggetto di analisi comprende 30 177 autori, un periodo temporale di 10 anni (1998-2007) e una base di 78 965 archi (su 28 dimensioni) rappresentanti le compartecipazioni a testi presentati nelle conferenze scelte.

Il dataset di test, invece, comprende tutte le compartecipazioni avvenute, tra gli autori già presenti nella rete, nel corso dell'anno successivo al periodo preso in esame per costruire la rete di partenza (2008). Tale insieme è costituito da 14 650 archi.

² <http://www.informatik.uni-trier.de/~ley/db/welcome.html>

IMDB: INTERNET MOVIE DATABASE

IMDB³ è un database online che tiene traccia di tutte le produzioni cinematografiche a partire dal 1888.

Proprietà	Valore	Proprietà	Valore
$ V $	43 869	Connected Components	335
$ E_{old} $	216 544	Network Diameter	18
$ E_{new} $	54 749	Clustering Coefficient	0,609
$ D $	28	Shortest Path	5,395
PPV_{rand}	3,032 104 682 23 e-06		

Tabella 3: IMDB Statistiche

La rete multidimensionale costruita su tale database è una rete di co-partecipazione (due nodi, rappresentanti due attori, sono collegati da un arco se hanno recitato entrambi in nello stesso film) avente come dimensioni i generi dei film (informazione contenuta nel database) e come informazioni temporali gli anni di produzione degli stessi.

Il dataset di training su cui è stata costruita la rete oggetto di analisi comprende 43 869 attori, un periodo temporale di 10 anni (1999-2008) e una base di 216 544 archi (su 28 dimensioni) rappresentanti le co-partecipazioni a film categorizzati nei generi scelti.

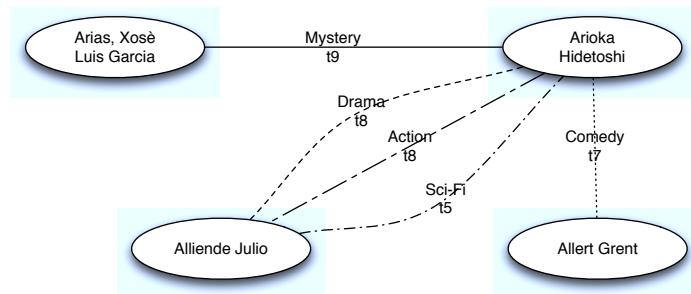


Figura 9: Esempio IMDB

Il dataset di test, invece, comprende tutte le co-partecipazioni avvenute, tra gli attori già presenti nella rete, nel corso dell'anno successivo al periodo preso in esame per costruire la rete di partenza (2009). Tale insieme è costituito da 54 749 archi (come mostrato in Tabella 3).

³ <http://www.imdb.com/>

GTD: GLOBAL TERRORISM DATABASE

In letteratura si trovano molti articoli in cui il task di Link Prediction è analizzato per la finalità di prevenzione di possibili attacchi di tipo terroristico (ad esempio in [17]).

<i>Proprietà</i>	<i>Valore</i>	<i>Proprietà</i>	<i>Valore</i>
$ V $	2 756	Connected Components	46
$ E_{old} $	32 279	Network Diameter	9
$ E_{new} $	2 572	Clustering Coefficient	0,778
$ D $	209	Shortest Path	3,082
PPV_{rand}	3,241 690 405 27 e-06		

Tabella 4: GTD Statistiche

GTD⁴ nasce da un progetto open source dell'università del Maryland. Tale progetto ha come fine la creazione e il mantenimento di un database che tenga traccia delle informazioni relative agli eventi terroristici.

I dati raccolti coprono l'arco temporale che va dal 1970 al 2008 e comprendono più di 87 000 attentati terroristici o presunti tali (caso quest'ultimo che si verifica quando non avvengono rivendicazioni da parte di gruppi ufficialmente riconosciuti). Le informazioni fruibili, per ogni evento, sono molteplici tra cui: data, luogo (informazioni gerarchiche che vanno da area geografica sino alla singola città), armi usate, tipologia di attacco, obiettivo, responsabili dell'attacco e statistiche sull'impatto finale.

Il Consorzio Nazionale per gli Studi sul Terrorismo e in Risposta al Terrorismo degli Stati Uniti (START) rende GTD disponibile online in modo da fornire un facile accesso, per chi ne fosse interessato, ai dati per lo studio e l'analisi della minaccia terroristica.

Caratteristiche di GTD:

- 87 000 azioni terroristiche in totale;
- 38 000 attacchi esplosivi ;
- 13 000 assassinii;
- 4 000 rapimenti ;
- 45 variabili per ciascun evento (per i più recenti si arriva a 120 variabili).

⁴ <http://www.start.umd.edu/gtd/>

Il grafo che andiamo ad analizzare è costruito su tale dataset (informazioni dettagliate in Tabella 4):

- i vertici identificano i gruppi terroristici conosciuti;
- le dimensioni rappresentano gli stati in cui sono avvenuti gli attacchi terroristici;
- esiste un arco tra due gruppi terroristici se, nello stesso periodo temporale, hanno portato uno, o più, attacchi ad uno stesso stato.

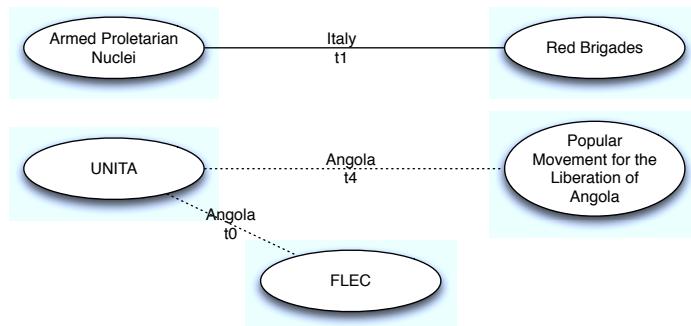


Figura 10: Esempio GTD

Lo scopo dell'applicazione di un predittore a tale rete è quello di prevedere quali sono le coppie di gruppi terroristici che, con maggior probabilità, progetteranno un attacco in futuro in un determinato stato.

GCD: GRAND COMICS DATABASE

GCD⁵ è un progetto che si propone di costruire un archivio dettagliato di informazioni relative ai fumetti di ambito internazionale. Le informazioni contenute in tale database coprono un arco temporale che va dagli inizi del 1800 sino al 2010 e sono relative alle singole pubblicazioni di ogni testata fumettistica, includendo i dati relativi ai disegnatori che hanno contribuito alla realizzazione di ogni singolo numero.

<i>Proprietà</i>	<i>Valore</i>	<i>Proprietà</i>	<i>Valore</i>
$ V $	10 000	Connected Components	1
$ E_{old} $	140 546	Network Diameter	4
$ E_{new} $	4 945	Clustering Coefficient	0,817
$ D $	7	Shortest Path	1,677
PPV_{rand}	1, 413 566 131 09 e-05		

Tabella 5: GCD Statistiche

Per costruire la rete è stato applicato il seguente criterio:

- i nodi identificano i fumettisti;
- le dimensioni rappresentano la tipologia di lavoro svolto per il singolo albo (copertina, storia, rubrica, introduzione..);
- un arco tra due fumettisti esiste se, nello stesso anno, entrambi hanno collaborato almeno una volta ad uno stesso albo nella stessa stessa sezione.

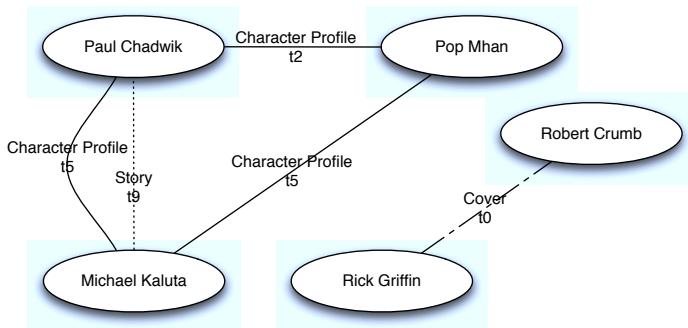


Figura 11: Esempio GCD

⁵ <http://www.comics.org/>

Per limitare la dimensione della rete si è considerato un numero ristretto di disegnatori e si è limitato il numero delle sezioni considerate per la multidimensionalità della rete a 7 (su un totale di 21).

Il training set comprende 10 anni (dal 1990 al 1999) mentre il test set è formato dall'anno successivo (informazioni dettagliate in Tabella 5).

VDC: INTERNATIONAL DYADIC EVENTS

VDC⁶ è un dataset creato da Gary King ed il suo team di ricerca: contiene informazioni relative a 10 milioni di eventi diadi di raccolti giornalmente nell'arco temporale che va dal 1990 al 2004.

<i>Proprietà</i>	<i>Valore</i>	<i>Proprietà</i>	<i>Valore</i>
$ V $	321	Connected Components	1
$ E_{old} $	413 766	Network Diameter	4
$ E_{new} $	43 741	Clustering Coefficient	0,705
$ D $	248	Shortest Path	1,851
PPV_{rand}	3,549 393 346 034 e-03		

Tabella 6: VDC Statistiche

Per ciascun evento sono specificati gli agenti (attore A e attore B) oltre alla tipologia di azione. Gli attori rappresentano, circa, 450 tra stati e gruppi inter-statali mentre, le informazioni di correlazione fornite, rappresentano eventi endogeni ed esogeni tra i nodi.

Nella modellazione fornita si è scelto di considerare solo le relazioni interstatali e di utilizzare come dimensioni le differenti tipologie di rapporti che possono intercorrere tra gli stati.

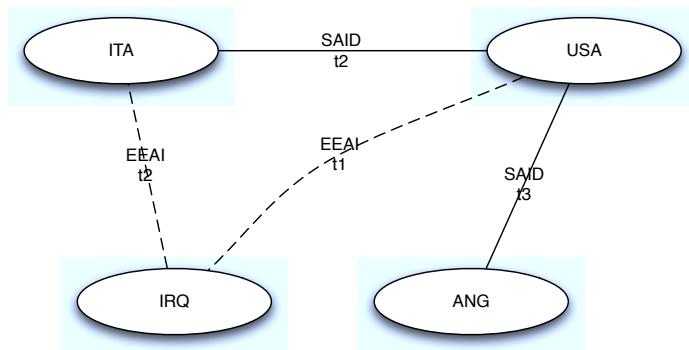


Figura 12: Esempio VDC

La presenza di un ristretto numero di nodi e di una relativamente alta dimensionalità della rete rende questo dataset un esempio di particolare interesse per gli approcci che utilizzino come supporto l'analisi temporale delle interazioni (come si intuisce dai valori in Tabella 6).

⁶ <http://hdl.handle.net/1902.1/FYXLAWZRIA>

4.2 METODOLOGIA DI ANALISI DEI RISULTATI

A seguito dell'applicazione di un predittore ad una rete si ottiene un insieme di archi, ciascuno definito dalla relativa quadrupla (*nodo, nodo, dimensione, score*), rappresentante il risultato del processo predittivo. Una volta ordinati gli archi appartenenti a tale insieme per score decrescente è possibile effettuare un'analisi complessiva delle performance ottenute dal predittore sulla data rete.

Per effettuare tale valutazione sono state proposte, in letteratura, diverse tipologie di analisi: di seguito si introducono le tre metodologie utilizzate nel prossimo capitolo per presentare i risultati ottenuti. Nel dettaglio si presenta brevemente in 4.2.1 la teoria che sta dietro alla formulazione della “Matrice di confusione” (dai cui tutte le tipologie di analisi effettuate sono dipendenti) quindi si introducono le curve ROC (4.2.2), le curve di Precision/Recall (4.2.3) e l'analisi grafico/numerica dell'andamento della Precision (4.2.4).

4.2.1 Matrice di confusione

True Positive (TP)	False Positive (FP)
True Negative (TN)	False Negative (FN)

Tabella 7: Matrice di confusione

La matrice di confusione (rappresentata in Tabella 7) è uno strumento utile per l'analisi dei risultati ottenuti da un generico task di classificazione.

Il problema di classificazione consiste, dato un dataset di partenza e un insieme finito di etichette, nell'assegnare, secondo regole predefinite costruite su di un insieme di training, a ciascun oggetto facente parte del dataset una etichetta. Questo procedimento di apprendimento (di tipo supervised) da origine, una volta completato, ad una classificazione degli oggetti presenti nel dataset analizzato (test set).

Dato l'insieme dei risultati fornito da un processo di classificazione su un insieme di m etichette (classi) è possibile costruire una matrice di confusione di dimensione $m \times m$: nel nostro caso consideriamo un processo di classificazione binaria.

Per leggere la Tabella 7 si assumano due classi (True e False) come indici sia di riga che di colonna:

- il valore contenuto nel campo True Positive (TP) rappresenta il totale degli oggetti predetti appartenere alla classe True ed effettivamente appartenenti alla stessa;
- il valore contenuto nel campo False Positive (FP) rappresenta il totale degli oggetti predetti appartenere alla classe True ma appartenenti nella realtà alla classe False;
- il valore contenuto nel campo True Negative (TN) rappresenta il totale degli oggetti predetti appartenere alla classe False ed effettivamente appartenenti alla stessa;
- il valore contenuto nel campo False Negative (FN) rappresenta il totale degli oggetti predetti appartenere alla classe False ma appartenenti nella realtà alla classe True.

La costruzione della matrice di confusione è descrivibile tramite un processo incrementale: è possibile, infatti, costruire una successione di matrici in cui, ad ogni passo, l'ultima differisce dalla precedente solo per il valore di classificazione di uno degli oggetti analizzati. Per costruzione quindi, data la successione di matrici di confusione definita per un generico processo di classificazione, le successioni dei valori appartenenti rispettivamente ai campi TP, FP, TN e FN sono positive e crescenti.

Abbiamo detto che tale strumento è stato introdotto per l'analisi dei risultati di processi di classificazione: in un processo di tipo predittivo è ancora possibile l'uso di tale rappresentazione tabulare (così come la sua costruzione incrementale) ma il vincolo costruttivo sulla crescenza delle successioni dei valori viene a cadere per le categorie di TN e FN.

Nel task di Link Prediction possiamo immaginare che la classe True sia da attribuire agli archi predetti mentre la False sia attribuita, implicitamente a quelli non predetti. Nello specifico definiamo E_{new} come l'insieme degli archi entrati a far parte del grafo nell'istante da predire (insieme dei risultati attesi):

- True Positive: archi predetti che appartengono all'insieme E_{new} ;
- False Positive: archi predetti che non appartengono all'insieme E_{new} ;
- True Negative: archi non predetti che non appartengono all'insieme E_{new} ;
- False Negative: archi non predetti che appartengono all'insieme E_{new} .

Il valore di TN è quindi dato per differenza ad ogni passo tra il totale degli archi che possono appartenere al grafo $\frac{|V| \times (|V|-1)}{2} \times |D|$ e gli archi sino ad ora analizzati dell'insieme dei risultati che appartengono a TP, mentre FN è ottenibile dalla differenza tra il valore totale degli archi appartenenti ad E_{new} e il valore TP calcolato allo stesso passo.

Questa particolarità costruttiva non inficia le informazioni fornite dalla matrice di confusione: l'unica differenza che è riscontrabile rispetto all'analisi di un problema di classificazione è quella fornita dalla diversa pendenza che assumono le curve ROC e Precision/Recall.

4.2.2 Curve ROC

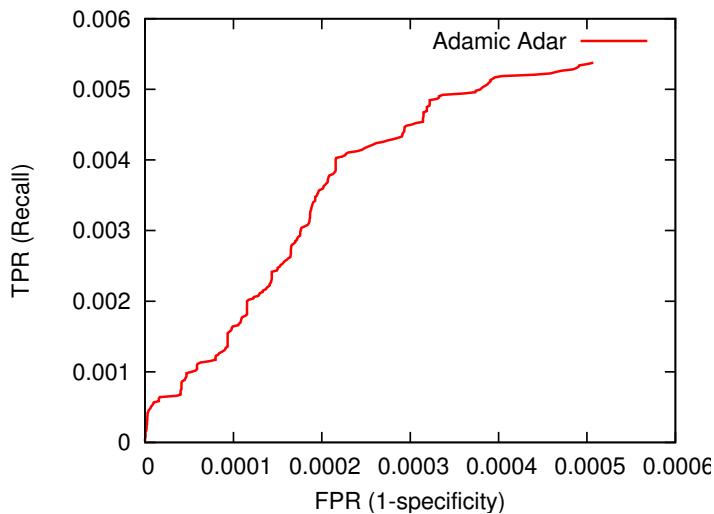


Figura 13: ROC curve

Le curve Receiver Operating Characteristic, o più semplicemente ROC, sono un plot dei valori di Sensitivity, o True Positive Rate (TPR), e False Positive Rate (FPR), o 1-Specificity.

Le curve ROC furono utilizzate per la prima volta da alcuni ingegneri elettrici che, durante la seconda guerra mondiale, avevano come compito la localizzazione dei nemici mediante l'uso del radar durante le battaglie. Recentemente, invece, le curve ROC sono utilizzate in machine learning e data mining, medicina, radiologia, psicologia, veterinaria e molti altri ambiti.

A partire dalla matrice di confusione, precedentemente introdotta, possono essere definite alcune misure: nel caso delle curve ROC siamo interessati nello specifico a TPR e FPR.

True Positive Rate definisce la performance raggiunta dal predittore nel determinare correttamente gli archi appartenenti all'insieme E_{new} rispetto al totale degli archi predetti:

$$TPR = \frac{TP}{P} = \frac{TP}{(TP + FN)} \quad (4.2.1)$$

False Positive Rate definisce quanti errori di misclassificazione occorrono tra gli archi predetti positivi rispetto al totale degli archi non appartenenti ad E_{new} :

$$FPR = \frac{FP}{N} = \frac{FP}{(FP + TN)} \quad (4.2.2)$$

Lo spazio cartesiano definito per rappresentare le curve ROC è costruito ponendo rispettivamente i valori di FPR e TPR sull'asse delle ascisse e su quello delle ordinate. Si rappresenta in questo modo il trade-off relativo tra il True Positive e i False Positive.

Ciascun punto appartenente al grafo rappresenta una matrice di confusione: per questo motivo, per ottenere una curva di ROC, (esempio in Figura 13) si utilizzano le successioni delle matrici di confusione calcolate incrementalmente (come discusso precedentemente).

4.2.3 Curve Precision/Recall

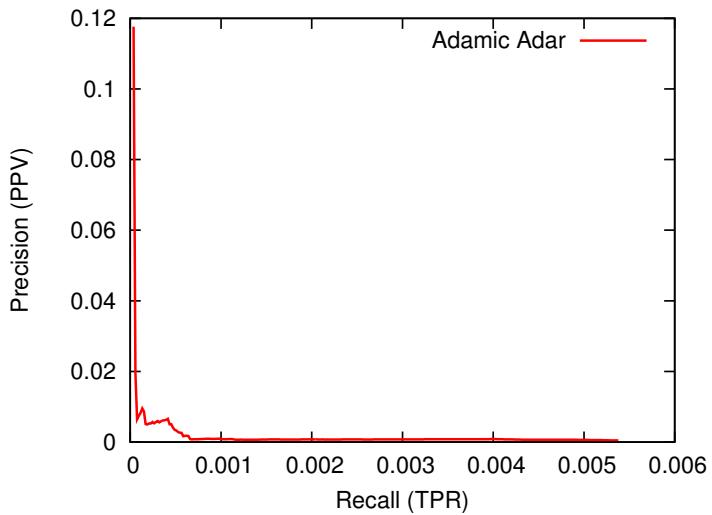


Figura 14: PR curve

Precision e Recall sono due misure ampiamente utilizzate per valutare la correttezza di algoritmi applicabili al problema di pattern recognition. Tali misure, ricavabili anch'esse dalla matrice di confusione, possono essere interpretate come una versione estesa del concetto di accuracy (una semplice misura che calcola la frazione delle istanze per cui è ritornata la classe corretta).

Utilizzando Precision e Recall, implicitamente, si suddivide l'insieme delle istanze appartenenti al risultato del predittore in due sottoinsiemi uno dei quali considerato “rilevante” per lo scopo delle misure.

Il valore di Recall è calcolato in modo equivalente a TPR (espresso dalla equazione 4.2.1) mentre Precision, anche chiamato Precision Predictive Value (o Performance), è definito come:

$$\text{PPV} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (4.2.3)$$

e rappresenta il rapporto tra le istanze correttamente predette ed il totale delle predizioni effettuate.

Lo spazio cartesiano definito per rappresentare le curve di Precision/Recall è costruito ponendo, rispettivamente, i valori di TPR e PPV sull'asse delle ascisse e quello delle ordinate.

Poiché lo spazio definito per le curve di Precision/Recall è isomorfo a quello definito per le curve ROC i plot delle due tipologie sono da considerarsi equivalenti [7].

4.2.4 Analisi dell'andamento della Precision

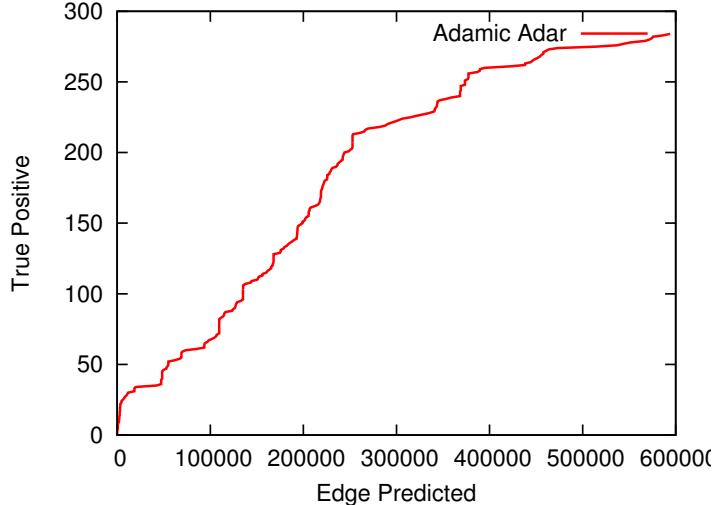


Figura 15: PT curve

L'ultima tipologia di analisi presentata è relativa all'andamento della Precision. Il valore della Precision, definito dall'equazione riportata in 4.2.3, come le altre misure già introdotte, essendo calcolato sulla matrice di confusione può essere ottenuto incrementalmente per i risultati forniti dal predittore una volta ordinati per score decrescente.

La rappresentazione grafica proponibile per tale analisi è quella che vede sull'asse delle ordinate il valore corrente di TP e su quello delle ascisse il valore di TP+FP: definito in tale modo le informazioni da associare agli assi, si ottiene un'equivalenza stretta (relativamente all'andamento) tra le curve appartenenti a questo spazio cartesiano e quelle appartenenti allo spazio definito per le curve di ROC.

L'analisi ulteriore consentita dallo studio dell'andamento della Precision è quella, introdotta da Kleinberg e Nowell in [21], relativa al rapporto tra le performance del predittore analizzato ed il predittore randomico applicato allo stesso grafo. Possiamo definire per ciascun grafo G la performance del predittore randomico tramite la formula:

$$\text{PPV}_{\text{rand}} = \frac{|\mathcal{E}_{\text{new}}|}{|\mathcal{D}| \times \frac{|V| \times (|V|-1)}{2} - |\mathcal{E}_{\text{old}}|} \quad (4.2.4)$$

rappresentante un predittore che predice, correttamente, tutti gli archi appartenenti all'insieme

me E_{new} con probabilità uniforme data dal rapporto tra la cardinalità di tale insieme ed il totale degli archi che possono essere generati per il grafo (su tutte le dimensioni) meno gli archi già appartenenti allo stesso. Come abbiamo già osservato i predittori introdotti, dovendo sfruttare le informazioni temporali, possono predire anche archi già appartenenti all'insieme E_{old} : essendo il valore definito da $|D| \times \frac{|V| \times (|V|-1)}{2} \gg |E_{old}|$ si utilizza la formula precedente per mantenere continuità con la definizione data da Kleinberg e Nowell (già lievemente modificata per tenere in conto la multidimensionalità della rete), assumendo lo scarto computato irrilevante.

Il rapporto tra la performance del predittore analizzato e quella del predittore randomico

$$p = \frac{PPV_{pred}}{PPV_{rand}} \quad (4.2.5)$$

rappresenta numericamente l'incremento di precisione introdotto dal modello applicato alla rete rispetto a quella ottenuta dal predittore randomico.

Nell'analisi successiva sono state introdotte 3 soglie per la valutazione del rapporto di performance tra il predittore di volta in volta analizzato e quello randomico: tali soglie sono state definite per $\frac{1}{3}, \frac{2}{3}, \frac{3}{3}$ dei risultati classificati come TP dal predittore preso in esame.

Tale scelta comporta per i modelli predittivi di Common Neighbours, Adamic Adar, Jaccard e tutti i derivati la possibilità di una comparazione diretta del rapporto di performance ad ogni soglia: infatti, come già discusso, l'insieme dei risultati proposti da tali predittori è lo stesso a meno di ordinamento sullo score, quindi anche gli insiemi dei TP coincidono.

5

RISULTATI Sperimentali

In questo capitolo si riportano i risultati sperimentali ottenuti a seguito dei test effettuati sui dataset, presentati in 4.1, per i predittori multidimensionali ed evolutivi proposti (sezioni 3.3 - 3.4).

In ogni sezione si analizzano i risultati ottenuti su di un singolo dataset raggruppando i predittori tramite la gerarchia introdotta nelle Figure 5 e 6 presentate nel capitolo 3. Al termine dell’analisi dei modelli predittivi (effettuata tramite curve di ROC, di Precision/Recall e studio dell’andamento della precision) si presenta, per ogni dataset, una valutazione complessiva degli approcci adottati.

Laddove è riportata, in modo tabellare, l’analisi numerica delle performance per ogni singolo predittore sono stati specificati:

- tramite i valori p_1 , p_2 ed p_3 il rapporto tra la performance del predittore analizzato e quella espressa dal predittore randomico¹ (il cui valore è specificato nei dettagli delle singole reti nel Capitolo 4) calcolato rispettivamente ad $\frac{1}{3}$, $\frac{2}{3}$ e $\frac{3}{3}$ dei True Positive predetti dallo specifico modello;
- tramite i valori g_1 , g_2 e g_3 il fattore di gain rispetto al rispettivo modello base (valori non presenti per i modelli AdHoc).

Inoltre si è evidenziato, in corsivo, tutti i valori di p_1 , p_2 ed p_3 che migliorino la precision del predittore base di riferimento e, in grassetto, i valori massimi di gain per ogni classe di moltiplicatori applicata ad ogni singolo modello base.

¹ Si veda definizione (4.2.5).

5.1 AOL QUERY LOG

5.1.1 Preditori derivati dai modelli Monodimensionali

MODELLO BASE

Dai grafici riportati in Figura 16 si osserva che la curva descritta da Preferential Attachment sul grafo in oggetto domina nettamente quella degli altri predittori utilizzati.

Tale osservazione preliminare ci porta ad ipotizzare che i predittori derivati da tale modello base saranno più performanti rispetto a quelli derivati dagli altri predittori analizzati.

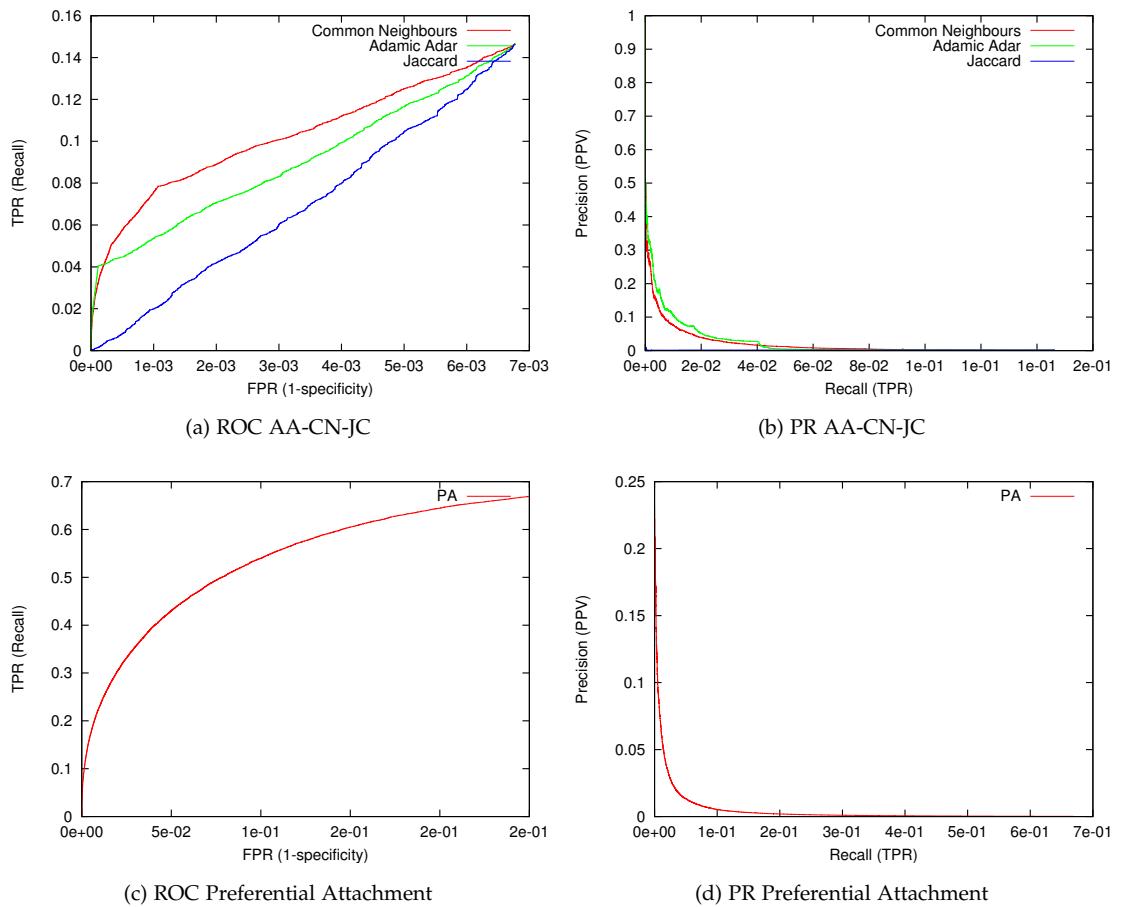


Figura 16: LOG Modelli Base

PREDITTORE	p ₁	p ₂	p ₃
AA	56,5	14,6	11,6
CN	103,8	23,2	11,6
JC	12,4	11,7	11,6
PA	22,1	5,1	1,0

Table 8: Query Log - Performance modelli Base

Inoltre analizzando le Figure 16a e 16b si nota che Adamic Adar riesce ad ottenere performance elevate nella prima fase predittiva (come sottolinea il grafico di PR) seppure complessivamente è sovrastato da Common Neighbours.

In Tabella 8 si mostra l'andamento della performance dei quattro modelli base. L'informazione, che a prima vista può apparire contrastante, relativa alle performance di Preferential Attachment è da interpretare ricordando che l'insieme dei risultati forniti da tale predittore è molto più esteso di quello proposto dagli altri presi in esame: per questo motivo seppure le curve ROC e PR risultino migliori, come si desume dai grafici precedentemente riportati, complessivamente, per le soglie stabilite, la performance del predittore è inferiore a quella degli altri analizzati.

Questa osservazione vale per tutti i modelli derivati da Preferential Attachment, per modelli multidimensionali locali (in alcuni casi) e per i modelli Ad Hoc.

MULTIDIMENSIONALI LOCALI

Dai grafici riportati per i predittori multidimensionali locali risulta un sostanziale peggioramento nelle performance rispetto ai modelli base. In particolare si ottiene una riduzione nel numero dei risultati restituiti dai derivati di Adamic Adar e performance inferiori per i predittori derivati da Jaccard e Common Neighbours. I predittori derivati da Preferential Attachment riescono appena a eguagliarne la performance nel caso di PAMS (coefficiente Mix/Split).

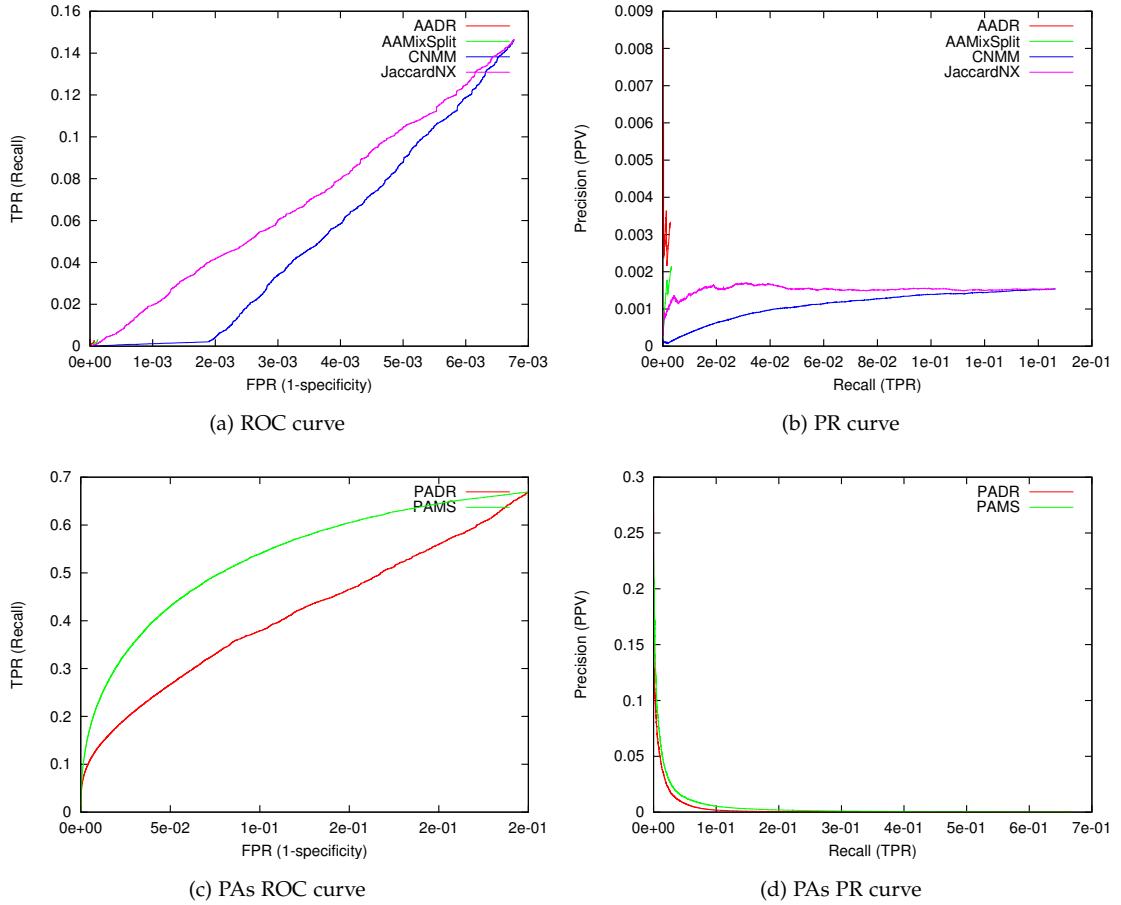


Figura 17: LOG Modelli Multidimensionali Locali

PREDITTORE	p ₁	p ₂	p ₃
AAMixSplit	10.0	11.8	16.3
AADR	24.0	19.0	25.3
CNMM	7.7	10.3	11.6
JaccardNX	12.1	11.7	11.6
PAMixSplit	22.1	5.1	0.9
PADR	5.9	1.8	0.9

Table 9: Query Log - Performance Modelli Multidimensionali Locali

In Tabella 9 non si riporta il fattore di gain poiché l'insieme dei risultati forniti dai modelli analizzati non coincide con quello fornito dai predittori presi come base di analisi: per lo stesso motivo le soglie di performance riportate - pur essendo rappresentative dell'andamento di questi modelli predittivi - non devono essere comparate direttamente con quelle riportate precedentemente per i modelli base se non alla luce dell'analisi grafica fornita dalle curve di ROC e PR.

MOLTIPLICATORI MULTIDIMENSIONALI GLOBALI

Dai grafici riportati in Figura 18 e 19 si denota un miglioramento per Adamic Adar tramite l’impiego dei moltiplicatori che sfruttano le informazioni di Parent e Correlation, una sostanziale stabilità per quanto riguarda Common Neighbours, Jaccard e Preferential Attachment.

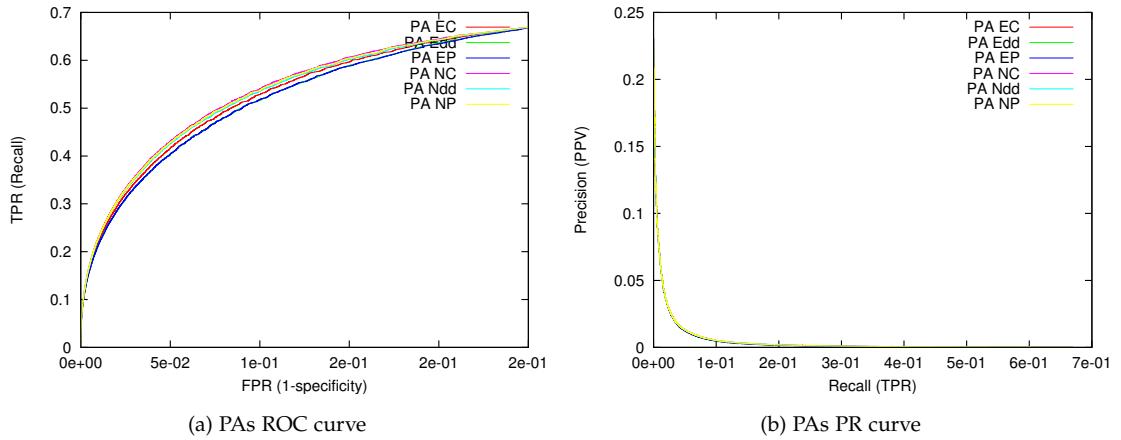


Figura 18: LOG Moltiplicatori Multidimensionali Globali I

In Tabella 10 si mostra l’incremento di performance relativo al predittore randomico e il fattore di gain, ove positivo, che ciascun moltiplicatore introduce se applicato allo specifico modello Base.

Da notare, come già evidenziato in precedenza, il fattore di gain introdotto dal moltiplicatore Edge Correlation ad Adamic Adar: complessivamente infatti tale combinazione porta alle migliori performance predittive ottenute su questo dataset utilizzando i moltiplicatori globali.

Per quanto riguarda i derivati da Preferential Attachment non si registrano particolari variazioni nella performance rispetto al modello base.

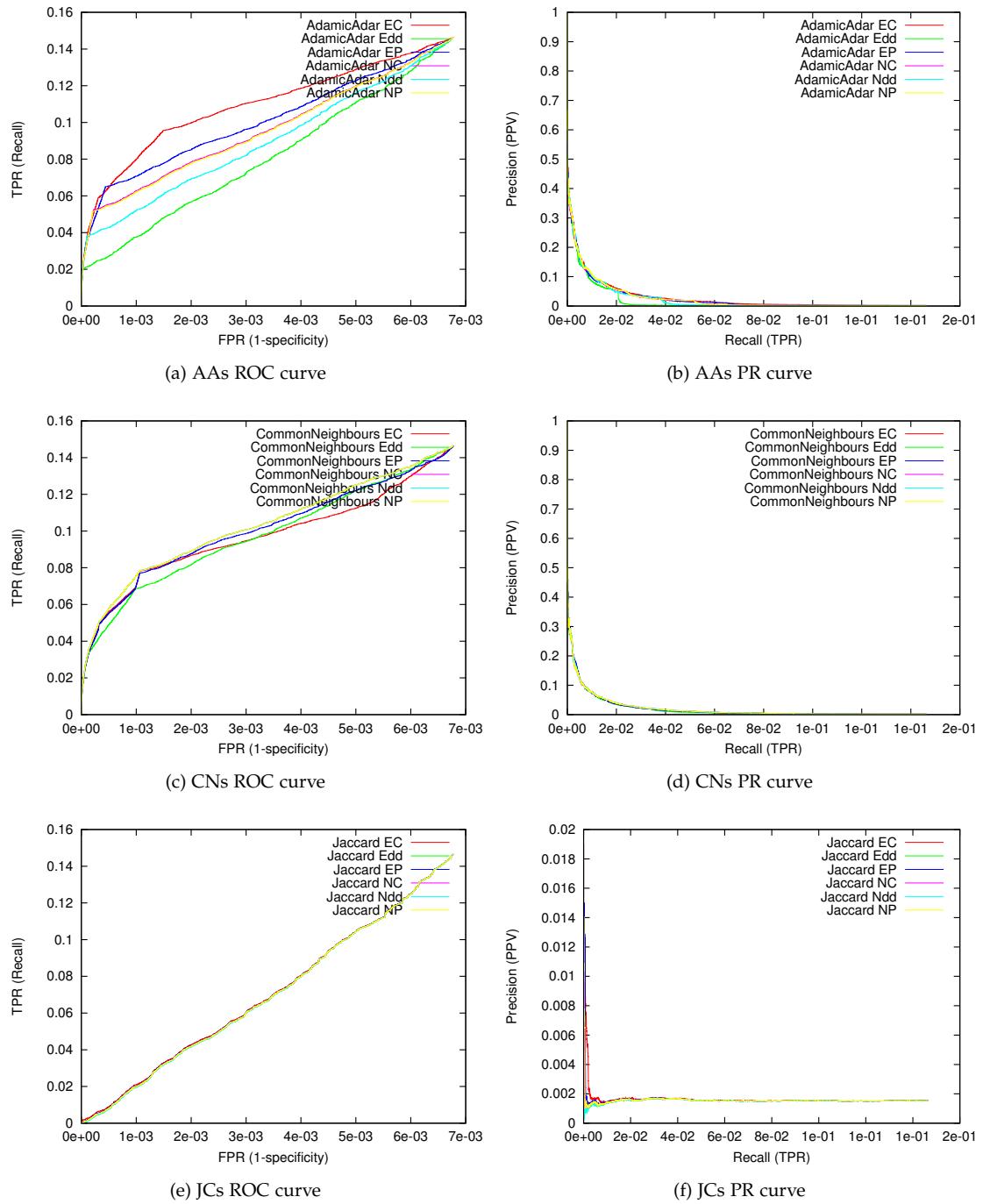


Figura 19: LOG Moltiplicatori Multidimensionali Globali II

PREDITTORE	p ₁	p ₂	p ₃	g ₁	g ₂	g ₃
AA EC	161,8	37,0	11,6	+105,3	+22,4	
AA Edd	19,7	12,6	11,6			
AA EP	125,6	19,3	11,6	+69,1	+4,7	
AA NC	151,3	16,5	11,6	+98,4	+1,9	
AA Ndd	43,2	14,3	11,6			
AA NP	155,2	16,2	11,6	+98,7	+1,6	
CN EC	93,2	18,7	11,6			
CN Edd	69,3	18,5	11,6			
CN EP	90,8	22,5	11,6			
CN NC	103,9	23,2	11,6	+0,1		
CN Ndd	104,0	23,2	11,6	+0,2		
CN NP	103,8	23,2	11,6			
JC EC	12,4	11,7	11,6			
JC Edd	12,4	11,7	11,6			
JC EP	12,3	11,7	11,6			
JC NC	12,2	11,7	11,6			
JC Ndd	12,1	11,7	11,6			
JC NP	12,1	11,7	11,6			
PA EC	20,0	4,6	1,0			
PA Edd	18,7	4,3	1,0			
PA EP	18,7	4,3	1,0			
PA NC	22,0	5,1	1,0			
PA Ndd	21,6	4,9	1,0	+0,5		
PA NP	21,8	5,1	1,0	+0,7		

Table 10: Query Log - Performance Moltiplicatori Multidimensionali Globali

MOLTIPLICATORI TEMPORALI

L'impiego di moltiplicatori temporali migliora le performance di Adamic Adar in modo significativo (in particolare quando si utilizzino informazioni di tipo Wpres); anche nel caso di Common Neighbours, Jaccard e Preferential Attachment si denotano incrementi nelle performance (seppure più significativi con la versione Over All del moltiplicatore Wpres).

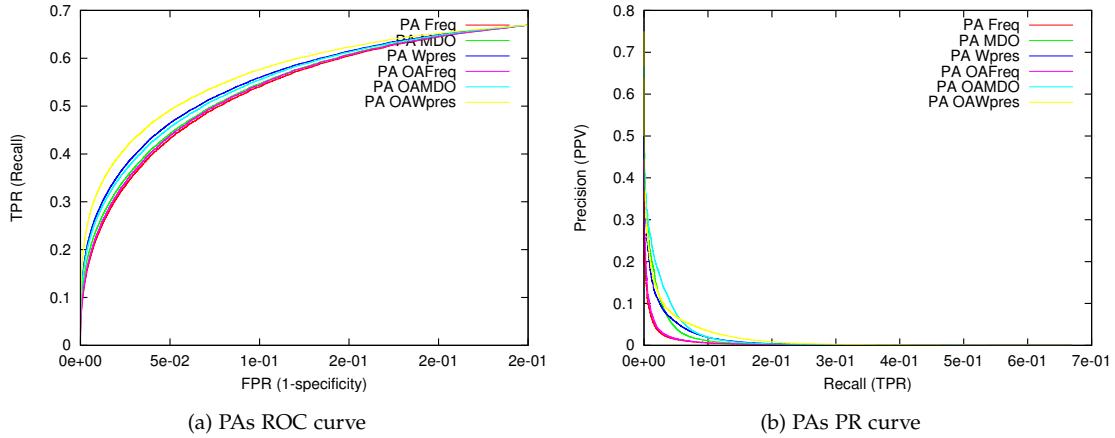


Figura 20: LOG Moltiplicatori Temporali I

In Tabella 11 si mostrano l'incremento di performance relativo al predittore randomico e il fattore di gain, ove positivo, che ciascun moltiplicatore introduce se applicato allo specifico modello Base.

A differenza di quanto accade con i moltiplicatori multidimensionali globali, analizzati precedentemente, si evidenziano netti miglioramenti, a livello di performance, ottenuti tramite le modifiche introdotte ai modelli base. Appare chiaro che, in questa rete, le informazioni temporali associate agli archi sono un fattore chiave per l'analisi predittiva.

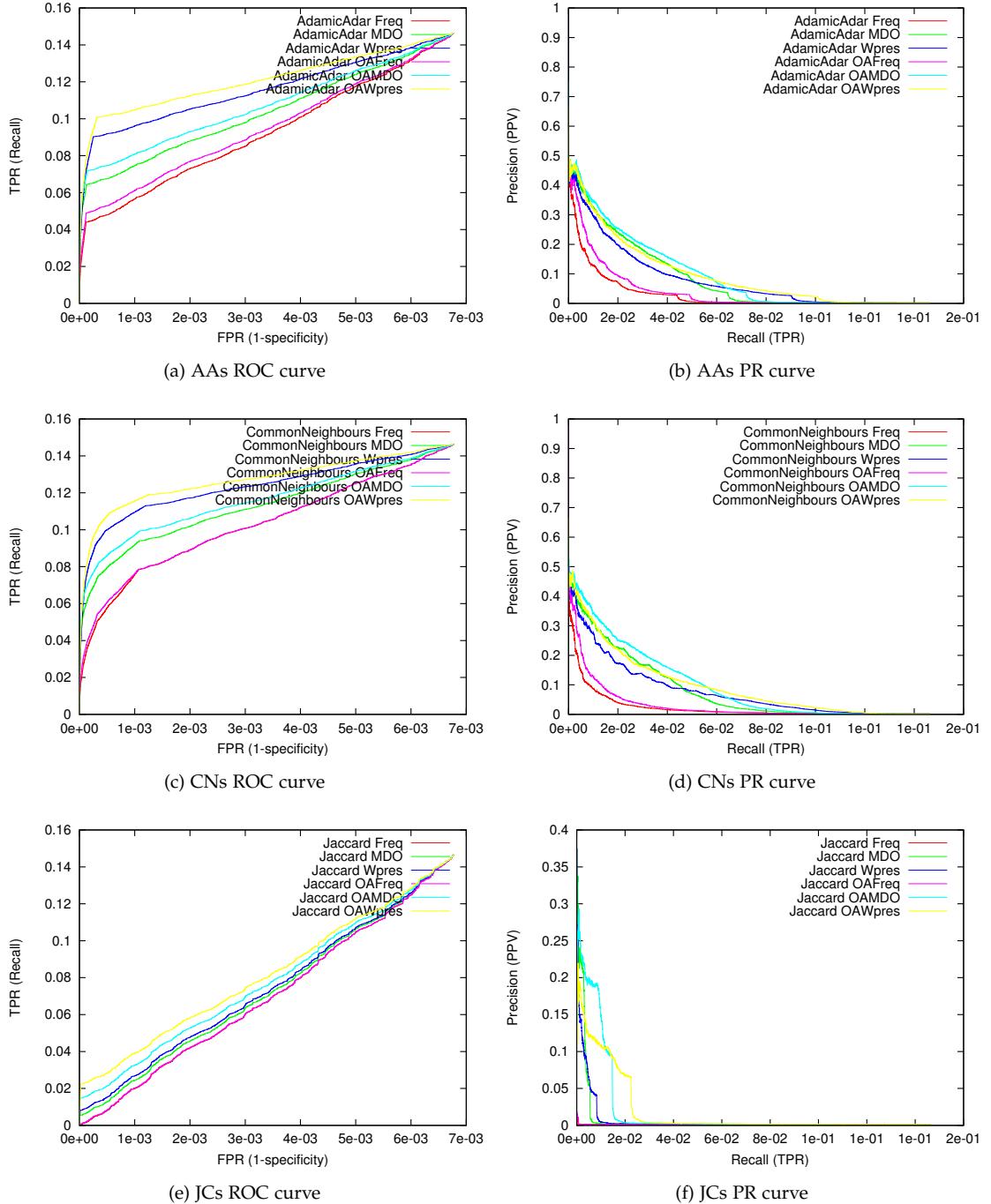


Figura 21: LOG Moltiplicatori Temporali II

PREDITTORE	p ₁	p ₂	p ₃	g ₁	g ₂	g ₃
AA Freq	148,1	15,0	11,6	+91,6	+0,4	
AA Wpres	665,8	75,8	11,6	+609,3	+61,2	
AA MDO	819,2	20,6	11,6	+762,7	+6,0	
AA OAFreq	242,4	15,9	11,6	+185,9	+1,3	
AA OAWpres	877,8	214,9	11,6	+821,3	+200,3	
AA OAMDO	1 047,4	26,1	11,6	+990,9	+11,5	
CN Freq	103,8	23,2	11,6			
CN Wpres	670,5	163,3	11,6	+566,7	+140,1	
CN MDO	763,8	50,1	11,6	+660,0	+26,9	
CN OAFreq	124,6	23,4	11,6	+20,8	+0,2	
CN OAWpres	871,8	233,8	11,6	+768,0	+210,6	
CN OAMDO	1 081,8	64,3	11,6	+978,0	+41,1	
JC Freq	12,1	11,7	11,6			
JC Wpres	14,4	12,1	11,6	+2,0		
JC MDO	13,7	11,8	11,6	+1,3	+0,1	
JC OAFreq	12,2	11,7	11,6			
JC OAWpres	20,3	12,9	11,6	+5,9	+1,2	
JC OAMDO	17,1	12,4	11,6	+4,7	+0,7	
PA Freq	22,9	5,2	1,0	+0,8	+0,1	
PA Wpres	121,4	10,6	1,0	+100,3	+5,5	
PA MDO	33,2	5,7	1,0	+12,2	+0,6	
PA OAFreq	25,1	5,4	1,0	+4,0	+0,3	
PA OAWpres	121,4	10,6	1,0	+100,3	+5,5	
PA OAMDO	33,2	5,7	1,0	+12,2	+0,6	

Table 11: Query Log - Performance Moltiplicatori Temporali

5.1.2 Predittori Ad-Hoc

I predittori AdHoc² applicati alla rete word-word analizzata presentano delle performance interessanti se comparate con quelle ottenute dai modelli analizzati in precedenza. Come si può notare in Tabella 12 i predittori che basano la loro analisi sulle informazioni di Weighted Dimension Relevance segnano le performance migliori (sia localmente al gruppo sia globalmente al totale dei predittori analizzati sino ad ora): tale risultato non tiene però conto del ristretto insieme dei risultati presentati dai due modelli. Per questo motivo, concordando con i grafici proposti in Figura 22 possiamo considerare Triangle e LNdd i predittori del gruppo che meglio performano su questo dataset.

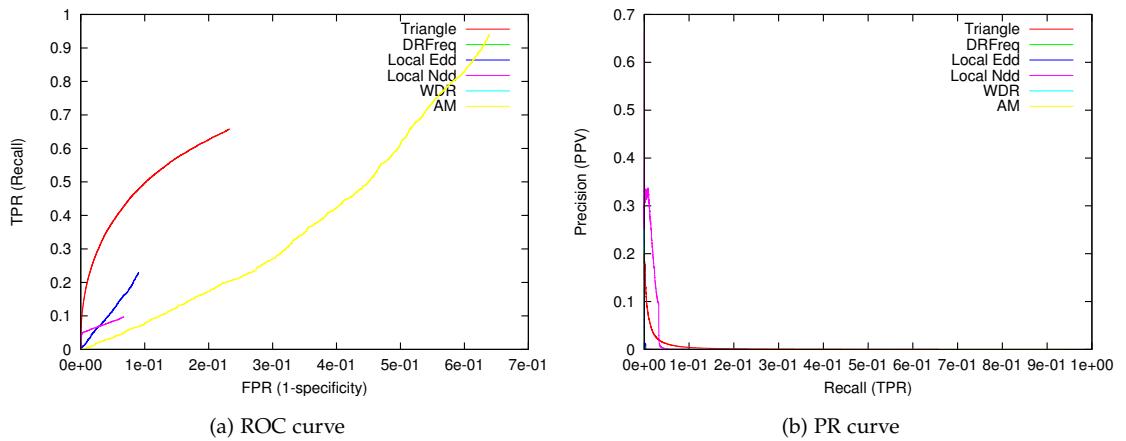


Figura 22: LOG Modelli AdHoc

² A causa della densità di archi della rete analizzata è stato necessario impostare una soglia minima agli score dei risultati proposti dai predittori AdHoc: tale soglia (il cui valore è stato fissato a 0,4) ha causato la riduzione del numero dei risultati forniti dai modelli ma non ha inficiato l'accuratezza predittiva.

PREDITTORE	p ₁	p ₂	p ₃
LEdd	1.4	1.3	1.2
LNdd	750.2	1.8	0.8
WDRw	1898.9	1898.9	1309.5
WDRf	1898.9	2278.7	30.9
AM	0.4	0.4	0.5
Triangle	15.7	3.8	1.0

Table 12: Query Log - Performance Modelli AdHoc

5.1.3 Analisi Riassuntiva

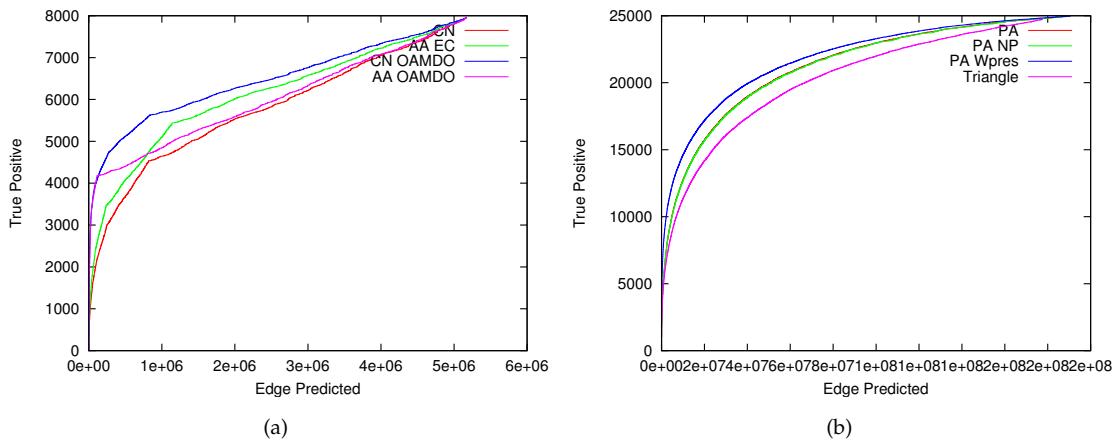


Figura 23: LOG Analisi riassuntiva

Analizzando i risultati ottenuti dai predittori sulla rete in esame si possono riscontrare i seguenti risultati:

- la migliore performance è ottenuta tramite l'impiego dell'informazione temporale di Weighted Presence usata in congiunzione con Preferential Attachment;
- tra ai predittori derivati da Adamic Adar, Common Neighbours e Jaccard, aventi un insieme di risultati ridotto rispetto a Preferential Attachment e modelli AdHoc, la precision più alta è raggiunta tramite l'introduzione del moltiplicatore temporale Over All Max Double Occurrence nel modello Common Neighbours.

Come si desume dai grafici riportati in Figura 23 per questa particolare rete l’impiego di informazioni temporali risulta cruciale per ottenere l’innalzamento delle performance predittive dei modelli base; le informazioni multidimensionali sono in grado di fornire incrementi di precisione se di tipo globale alla rete ma non se di tipo locale.

I modelli AdHoc, complice anche la soglia imposta che ne ha limitato l’insieme dei risultati ottenuti, non riescono ad esprimere ottimi risultati predittivi attestandosi, nel caso di Triangle, all’andamento medio dei predittori derivati da Preferential Attachment.

5.2 DBLP: COMPUTER SCIENCE BIBLIOGRAPHY

5.2.1 Predittori derivati dai modelli Monodimensionali

MODELLO BASE

Dai dati riportati in Tabella 13 e mostrati nei grafici di ROC e PR in Figura 24 si nota che le performance migliori su questa rete, relativamente ai modelli base, sono raggiunte tramite Common Neighbours (che ha una maggiore precisione nei primi risultati proposti) e Adamic Adar (che denota un andamento di performance crescente).

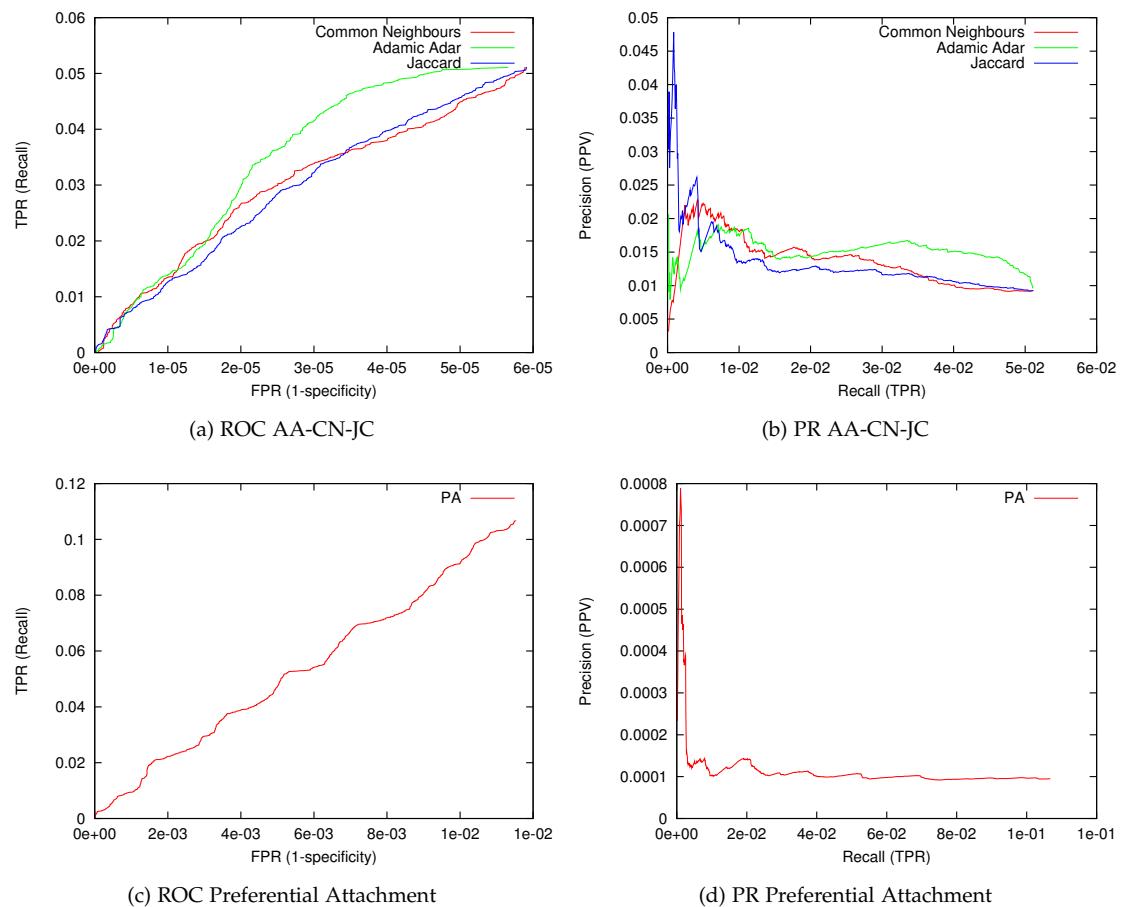


Figura 24: DBLP Modelli Base

PREDITTORE	p ₁	p ₂	p ₃
AA	12 360,5	14 580,6	8 397,4
CN	13 251,7	10 750,3	8 067,1
JC	10 548,5	10 238,5	8 040,1
PA	111,7	95,7	92,8

Table 13: DBLP- Performance modelli Base

Inoltre, contrariamente a quanto accade nel dataset dei Query Log di AOL, le curve di Preferential Attachment non dominano mai (né localmente né globalmente) quelle degli altri predittori. Questa particolarità evita, in questo caso come nei prossimi dataset, possibili incomprensioni sull'interpretazione dei valori riportati dalle tabelle e i grafici riportati in figura.

MULTIDIMENSIONALI LOCALI

In questa particolare rete i modelli predittivi AAMixSplit e AADR non riescono a fornire alcun risultato quindi sono presentati i grafici dei rimanenti quattro predittori analizzati.

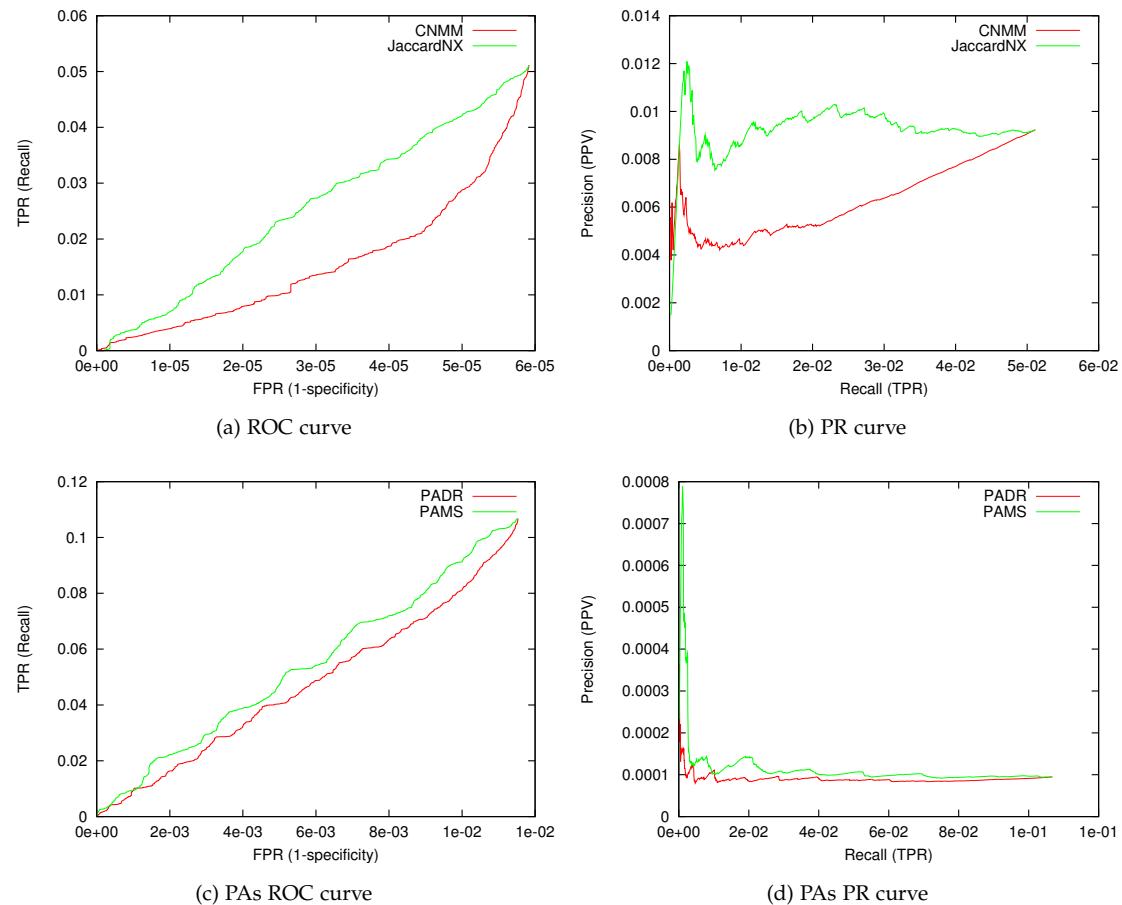


Figura 25: DBLP Modelli Multidimensionali Locali

PREDITTORE	p ₁	p ₂	p ₃
AAMixSplit	N.A.	N.A.	N.A.
AADR	N.A.	N.A.	N.A.
CNMM	4495.9	5925.4	8037.5
JaccardNX	8427.8	8158.8	8039.9
PAMixSplit	78.5	73.8	82.4
PADR	95.3	89.0	82.5

Table 14: DBLP - Performance Modelli Multidimensionali Locali

Anche per questo dataset le performance prodotte dai modelli multidimensionali locali sono inferiori a quelle dei rispettivi modelli base.

MOLTIPLICATORI MULTIDIMENSIONALI GLOBALI

Come mostrato in Figura 26 e Figura 27 i moltiplicatori multidimensionali globali portano ad una discreta varianza rispetto alle curve che descrivono il comportamento dei modelli predittivi di base.

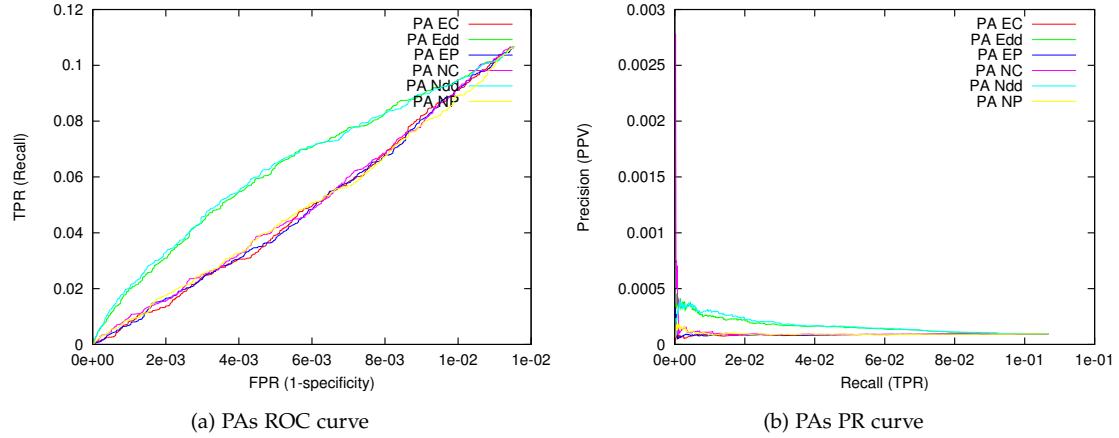


Figura 26: DBLP Moltiplicatori Multidimensionali Globali I

I risultati dell’analisi delle performance riportati in Tabella 15 mostrano che, complessivamente, il moltiplicatore Node Dimension Degree è quello che influisce maggiormente sui modelli analizzati. Nello specifico Preferential Attachment risulta essere il predittore più sensibile, in positivo ed in modo costante, all’introduzione delle informazioni multidimensionale di carattere globale alla rete mentre Adamic Adar riesce ad incrementare le sue performance solo ed esclusivamente sulle predizioni relative ai risultati con alto score.

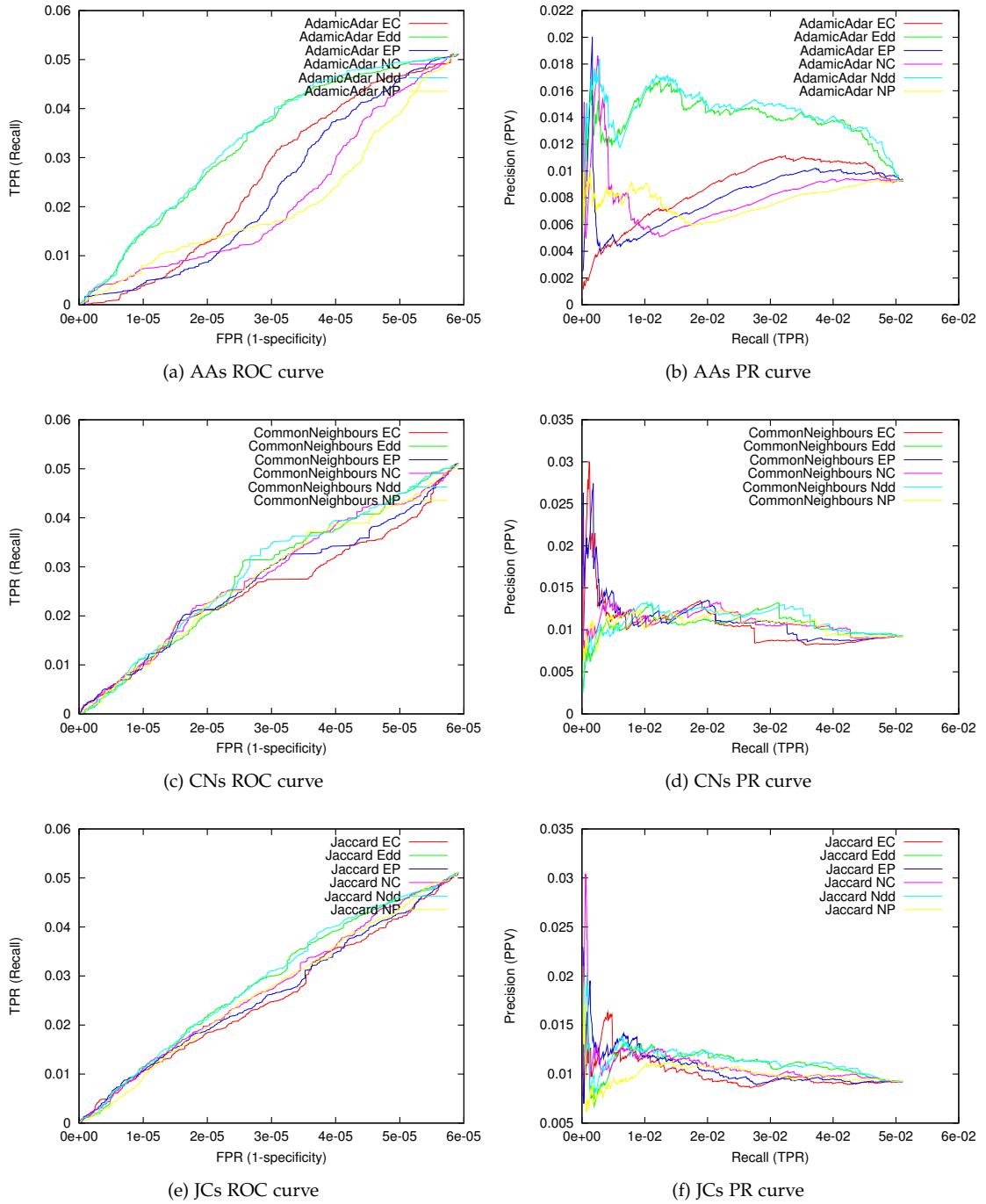


Figura 27: DBLP Moltiplicatori Multidimensionali Globali II

PREDITTORE	p ₁	p ₂	p ₃	g ₁	g ₂	g ₃
AA EC	6 932,3	9 468,7	8 130,5			
AA Edd	12 767,4	12 057,7	8 036,4	+406,9		
AA EP	5 975,2	8 472,1	8 048,9			
AA NC	5 099,6	7 514,5	8 035,6			
AA Ndd	13 548,4	12 834,0	8 050,6	+1187,9		
AA NP	5 292,1	6 993,4	8 080,3			
CN EC	10 810,0	7 630,4	8 067,8		+9,7	
CN Edd	9 270,9	9 774,4	8 036,4			
CN EP	10 502,3	8 381,1	8 036,8			
CN NC	10 944,8	9 308,5	8 036,6			
CN Ndd	10 777,4	11 169,2	8 036,4		+418,9	
CN NP	9 472,8	9 421,5	8 045,9			
JC EC	8 875,1	8 438,3	8 040,0			
JC Edd	10 212,7	9 633,1	8 042,3		+2,2	
JC EP	9 652,3	8 176,1	8 041,2		+1,1	
JC NC	10 106,1	8 901,4	8 040,7		+0,6	
JC Ndd	10 455,9	9 430,1	8 041,9		+1,8	
JC NP	9 605,6	8 261,3	8 040,5		+0,4	
PA EC	70,3	70,4	77,6			
PA Edd	198,2	148,3	131,8	+86,6	+52,6	+39
PA EP	80,0	73,7	77,9			
PA NC	74,7	76,4	76,5			
PA Ndd	217,9	157,7	137,4	+106,2	+62,0	+44,6
PA NP	89,4	77,2	78,1			

Table 15: DBLP - Performance Moltiplicatori Multidimensionali Globali

MOLTIPLICATORI TEMPORALI

Come avviene sulla rete costruita sui Query Log di AOL anche in IMDB le performance dei predittori base possono essere sensibilmente migliorate mediante l'adozione di moltiplicatori che sfruttino le informazioni temporali associate agli archi presenti nella rete.

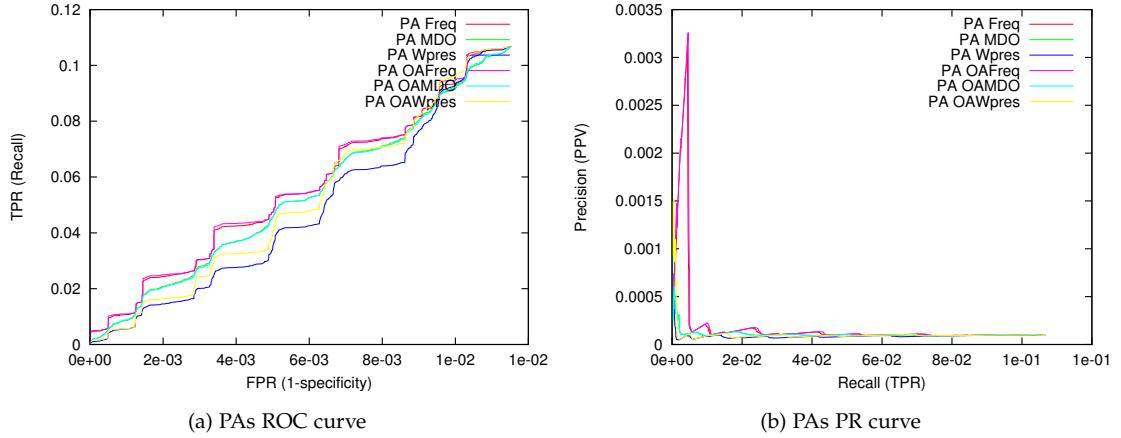


Figura 28: DBLP Multiplicatori Temporali I

In particolare possiamo osservare (Tabella 16) che il moltiplicatore Max Double Occurrence, nella sua versione Over All, è quello che fa segnare incrementi maggiori nelle performance predittive di Adamic Adar (+4 272,8\+3 555,1 rispettivamente sulla 1° e 2° soglia), Common Neighbours (+7 987,7\+5 485,5) e Jaccard (+5 085,1 sulla 2° soglia). Sempre su Jaccard è interessante notare l'incremento fornito dalla variante Over All di Weighted Presence sulla 1° soglia analizzata (+12 930,0).

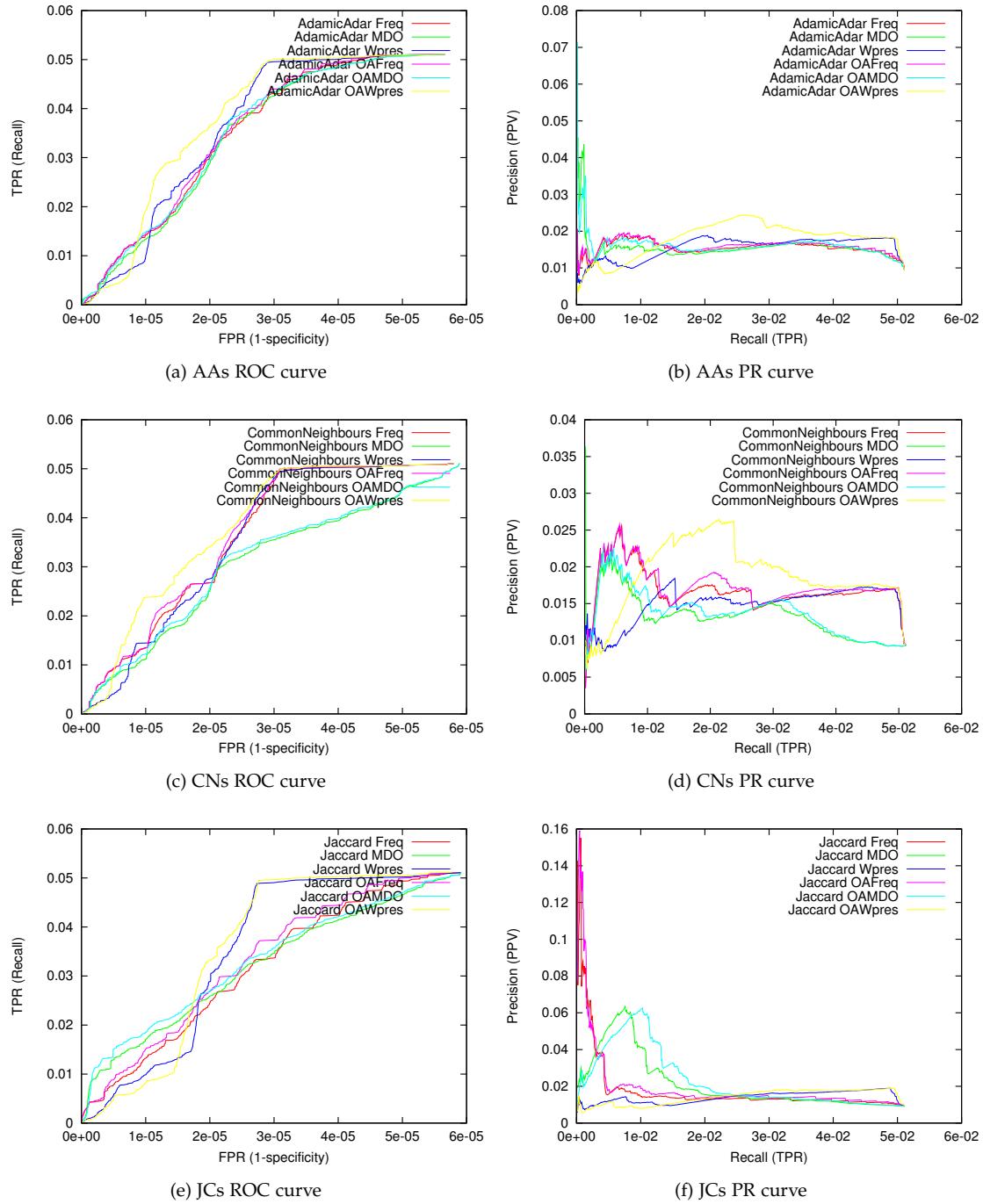


Figura 29: DBLP Moltiplicatori Temporali II

PREDITTORE	p ₁	p ₂	p ₃	g ₁	g ₂	g ₃
AA Freq	12 415,9	14 596,8	8 391,8	+55,4	+16,2	
AA Wpres	11 927,9	14 436,1	8 396,5			
AA MDO	14 400,7	15 056,1	8 388,5	+2 040,2	+475,5	
AA OAFreq	12 628,0	14 613,6	8 389,8	+267,5	+33	
AA OAWpres	12 722,7	14 545,0	8 396,3	+362,2		
AA OAMDO	16 633,3	17 935,7	8 387,5	+4 272,8	+3 555,1	
CN Freq	14 326,2	13 878,8	8 188,5	+1 074,5	+3 128,5	+121,4
CN Wpres	12 299,4	11 707,8	8 066,9		+957,5	
CN OAFreq	13 028,7	13 754,5	8 188,5		+3 004,2	+121,4
CN OAFreq	14 724,1	14 204,6	8 086,5	+1 022,4	+3 454,3	
CN OAWpres	13 059,9	12 463,1	8 066,5		+1 712,8	
CN OAMDO	21 239,4	16 235,8	8 187,5	+7 987,7	+5 485,5	+120,4
JC Freq	12 070,7	14 938,3	8 039,1	+1 522,2	+4 699,8	
JC Wpres	16 266,0	10 945,7	8 040,0	+5 717,5	+707,2	
JC MDO	9 089,7	14 246,8	8 039,1		+4 008,3	
JC OAFreq	13 063,3	12 189,3	8 238,6	+2 514,8	+1 950,8	
JC OAWpres	23 478,5	12 075,5	8 040,0	+12 930,0	+1837	
JC OAMDO	16 919,5	15 323,6	8 238,6	+6 371	+5 085,1	
PA Freq	111,9	95,8	63,0	+0,2		
PA Wpres	111,6	92,4	58,6			
PA MDO	56,24	64,2	49,4			
PA OAFreq	111,9	95,8	63,0	+0,2		
PA OAWpres	111,3	93,1	60,0			
PA OAMDO	77,2	66,9	50,7			

Table 16: DBLP - Performance Moltiplicatori Temporali

5.2.2 Predittori Ad-Hoc

Per quanto riguarda i modelli predittivi Ad Hoc possiamo notare che le performance registrate da Weighted Dimension Relevance (nella versione arricchita da informazioni di Weighted Presence) sono, per quanto riguarda la 1° soglia le migliori sino ad ora ottenute su questo dataset.

Un analisi più approfondita, che tenga conto anche della dimensione dell'insieme delle soluzioni fornito dai modelli Ad Hoc rispetto a quelli forniti dai modelli base, porta a constatare che globalmente, tutti i modelli facenti parte di questa categoria, dominano nei grafici di ROC e PR le curve dei modelli predittivi derivati dai modelli base.

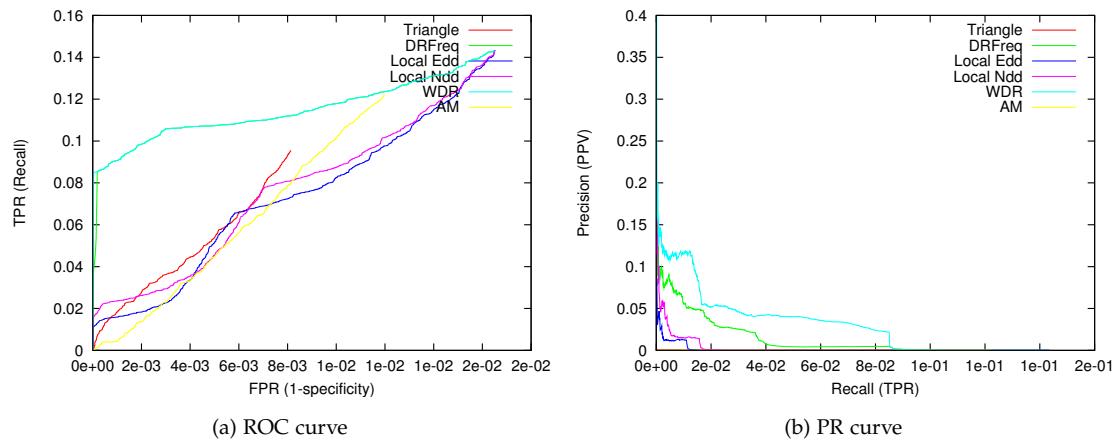


Figura 30: DBLP Modelli AdHoc

PREDITTORE	p ₁	p ₂	p ₃
LEdd	76,9	100,3	73,0
LNdd	84,7	77,4	74,8
WDRw	36 541,1	729,9	75,0
WDRf	4611,8	729,9	75,0
AM	80,0	80,0	89,6
Triangle	131,0	99,6	105,6

Table 17: DBLP - Performance Modelli AdHoc

5.2.3 Analisi Riassuntiva

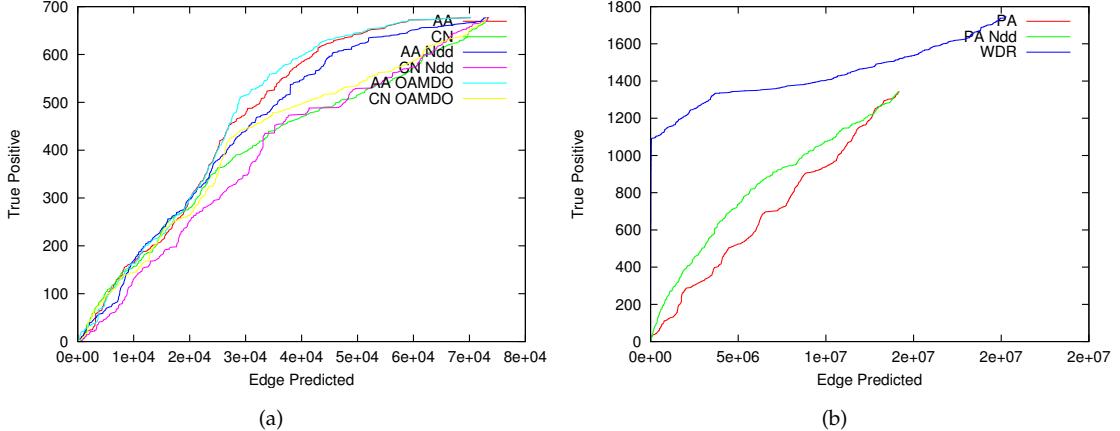


Figura 31: DBLP Analisi riassuntiva

Come mostrato in Figura 31a, dove sono riportati i migliori modelli derivati da Adamic Adar e Common Neighbours, per la rete DBLP l'informazione temporale introdotta da Over All Max Double Occurrence è quella che riesce ad innalzare maggiormente le performance predittive. Sempre per tali modelli il moltiplicatore multidimensionale che fa registrare, su alcune soglie, i maggiori incrementi di precisione è Node Dimension Degree.

Analizzando i risultati proposti in Figura 31b si nota, come già evidenziato quando sono stati introdotti i risultati della classe dei predittori AdHoc, che i migliori risultati - complessivamente a tutti i predittori proposti - sono raggiunti da Weighted Dimension Relevance (con informazioni di Weighted Presence).

5.3 IMDB: INTERNET MOVIE DATABASE

5.3.1 *Predittori derivati dai modelli Monodimensionali*

MODELLO BASE

Come illustrato in Figura 32 e Tabella 18 nel caso di IMDB il modello base con le performance migliori risulta essere Adamic Adar.

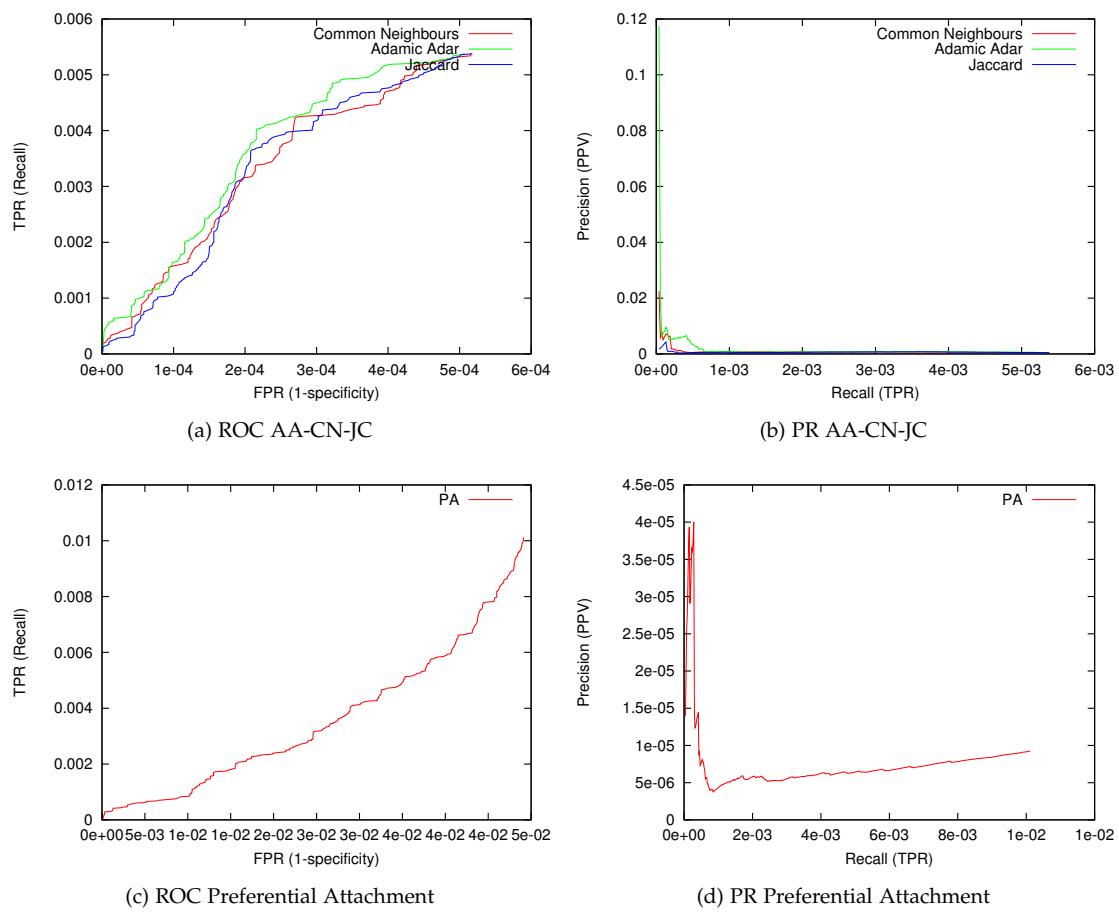


Figura 32: IMDB Modelli Base

PREDITTORE	p ₁	p ₂	p ₃
AA	358,8	402,2	235,2
CN	319,5	322,9	230,5
JC	264,8	379,7	230,3
PA	2,6	2,8	3,1

Table 18: IMDB - Performance modelli Base

Analizzando i dati presentati si nota che, sostanzialmente, le performance complessive (3° soglia) di Common Neighbours e Jaccard raggiungono quelle di Adamic Adar: ci è dovuto all'insieme dei risultati fornito dai tre predittori che risulta essere equivalente a meno di ordinamenti sugli score. Le curve prodotte dai risultati di Preferential Attachment non riescono mai a dominare quelle degli altri predittori (come è facile intuire osservando la scala degli assi dei grafici presentati).

MULTIDIMENSIONALI LOCALI

Analizzando i dati riportati si può notare come solo i modelli derivati da Preferential Attachment (PADR e PAMS) presentino lievi miglioramenti nelle performance.

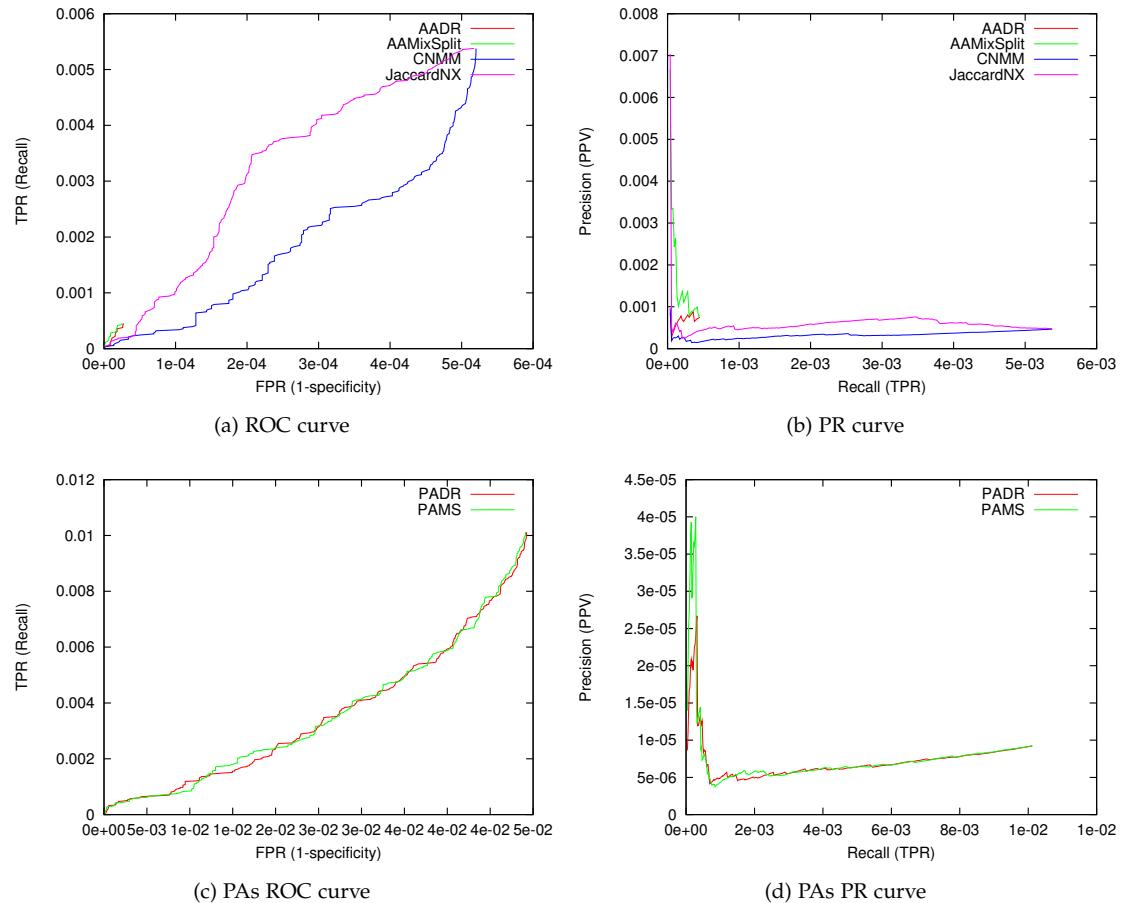


Figura 33: IMDB Modelli Multidimensionali Locali

PREDITTORE	p ₁	p ₂	p ₃
AAMixSplit	497.4	391.5	391.1
AADR	335.0	368.4	372.8
CNMM	151.4	167.0	229.3
JaccardNX	261.0	349.1	230.5
PAMixSplit	2.8	3.4	4.5
PADR	2.9	3.5	4.5

Table 19: IMDB - Performance Modelli Multidimensionali Locali

I predittori basati su Adamic Adar (AADR e AAMixSplit) segnano alte performance ma ottengono un insieme di risultati molto ristretto rispetto al loro modello base quindi i valori riportati non debbono essere paragonati se non alla luce dell'analisi effettuata sui grafici riportati.

MOLTIPLICATORI MULTIDIMENSIONALI GLOBALI

Dai grafici in Figura 34 e Figura 35 si nota che l'impatto dei fattori moltiplicativi multidimensionali globali alla rete porta ad una nutevole varianza, sia in positivo che in negativo, sulle curve descritte dai modelli predittivi di base.

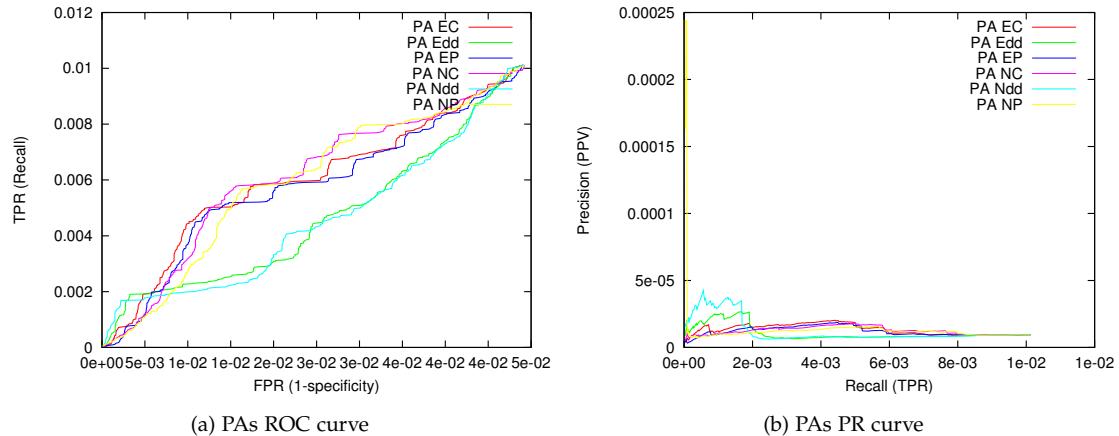


Figura 34: IMDB Moltiplicatori Multidimensionali Globali I

Come mostrato nel dettaglio in Tabella 20 le informazioni multidimensionali che fruttano gli incrementi di performance più sostanziali sono: Node Dimension Relevance (Adamic Adar e Common Neighbours) e Node Correlation (Jaccard e Preferential Attachment).

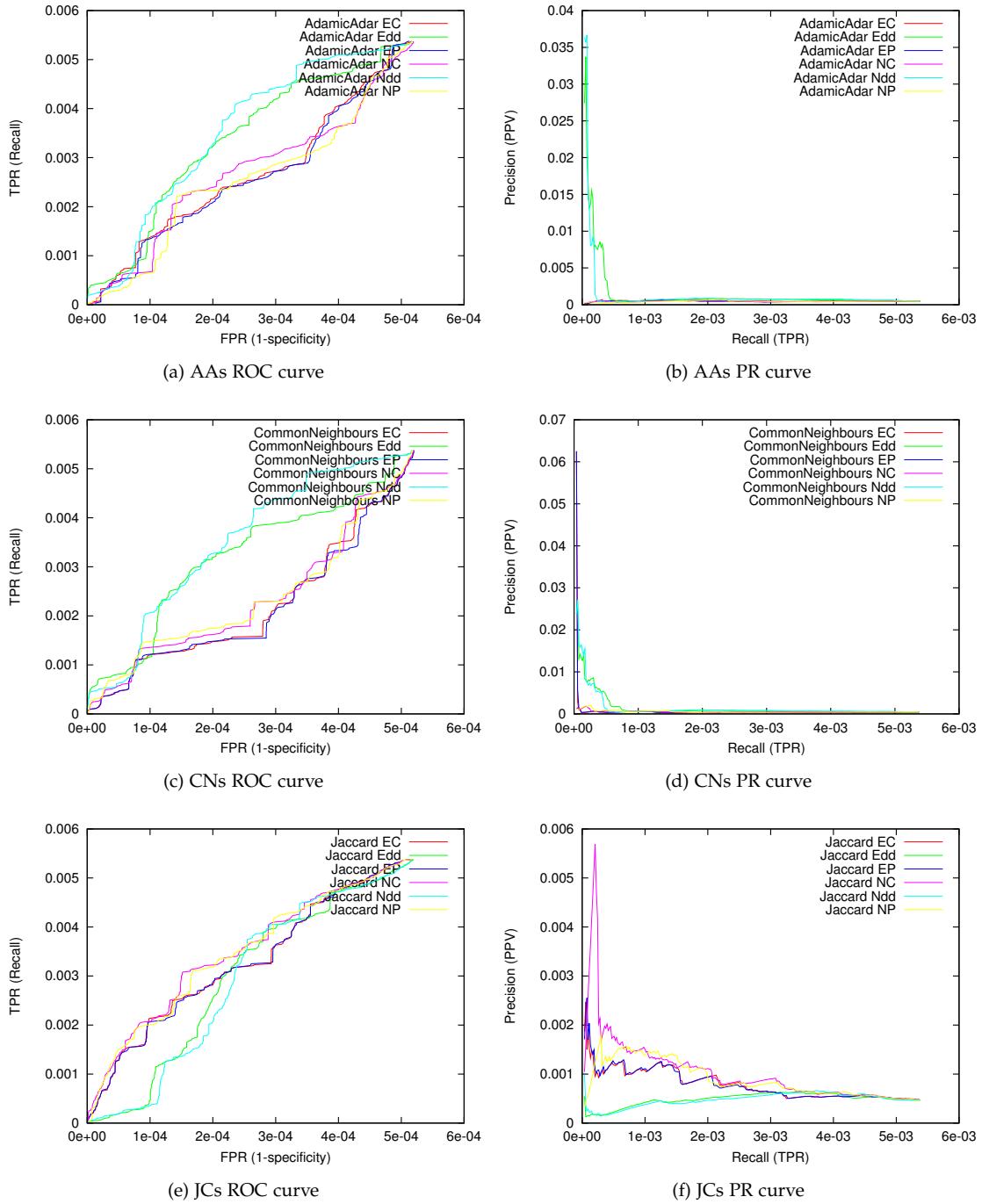


Figura 35: IMDB Moltiplicatori Multidimensionali Globali II

PREDITTORE	p ₁	p ₂	p ₃	g ₁	g ₂	g ₃
AA EC	284,1	214,5	231,8			
AA Edd	369,4	333,6	230,6	+10,6		
AA EP	259,6	210,9	232,6			
AA NC	295,9	205,7	229,6			
AA Ndd	427,5	366,9	230,5	+68,7		
AA NP	281,8	198,6	239,4			
CN EC	141,1	187,7	229,5			
CN Edd	356,0	311,6	231,0	+36,5		+0,5
CN EP	138,4	183,0	229,3			
CN NC	171,3	193,5	229,5			
CN Ndd	445,8	352,7	230,9	+126,3	+29,8	+0,4
CN NP	182,1	219,1	229,8			
JC EC	415,2	268,8	230,3	+150,4		
JC Edd	225,1	293,5	231,0			+0,7
JC EP	420,1	266,9	230,3	+155,3		
JC NC	562,1	320,0	230,3	+297,3		
JC Ndd	213,3	309,9	230,1			
JC NP	526,6	315,8	230,3	+261,8		
PA EC	8,3	9,0	7,0	+5,7	+6,2	+3,9
PA Edd	12,5	3,4	3,7	+9,9	+0,6	+0,3
PA EP	7,0	8,1	6,0	+4,4	+3,2	+5
PA NC	5,8	7,2	8,3	+3,2	+4,4	+5,2
PA Ndd	8,0	3,7	3,7	+5,4	+0,9	+0,6
PA NP	5,1	6,2	7,6	+2,5	+3,4	+4,5

Table 20: IMDB - Performance Moltiplicatori Multidimensionali Globali

MOLTIPLICATORI TEMPORALI

Le informazioni di carattere temporale possono essere sfruttate per incrementare le performance dei modelli predittivi proposti su IMDB. Come si desume sia dai grafici riportati sia dalla tabella riassuntiva gli incrementi di performance si registrano esclusivamente sui valori di 1° e 2° soglia: seppure complessivamente non si riesce a migliorare la capacità predittiva si ottiene lo sperabile risultato di assegnare in modo più accurato gli score alti di confidenza alle predizioni effettuate.

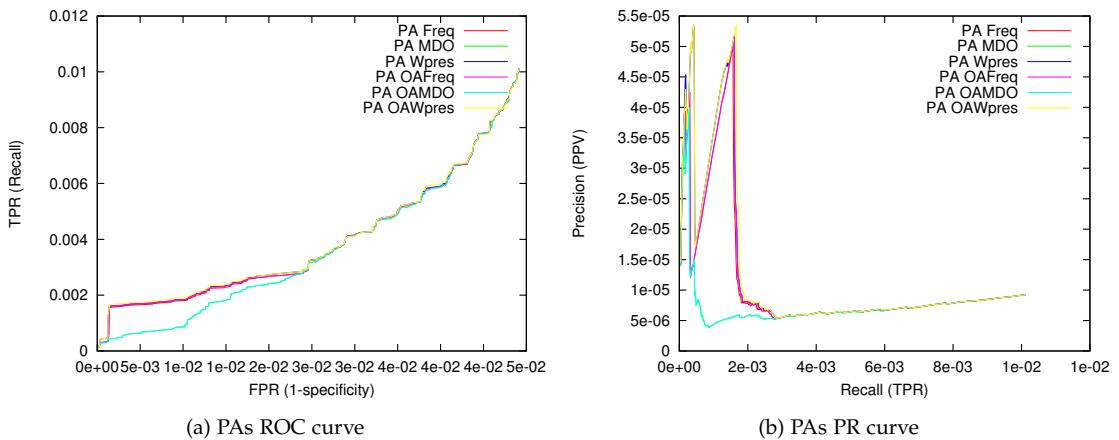


Figura 36: IMDB Moltiplicatori Temporali I

In particolare, come mostrato in Tabella 21, i modelli predittivi che riescono a trarre maggior beneficio dall'analisi delle informazioni temporali sono Common Neighbours e Jaccard (in particolare se utilizzati in congiunzione con le informazioni di Over All Max Double Occurrence), rispettivamente sulla 2° e 1° soglia.

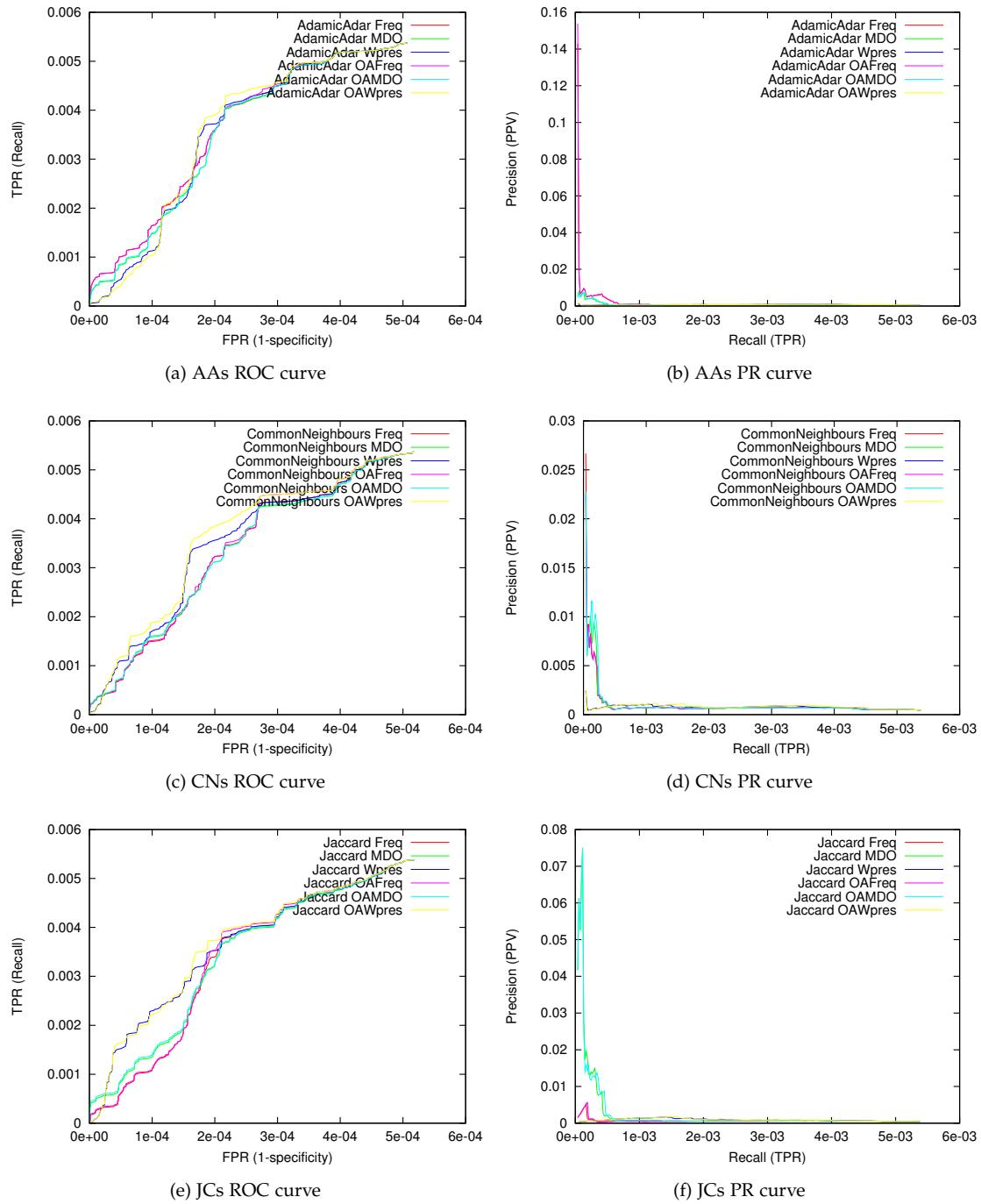


Figura 37: IMDB Moltiplicatori Temporali II

PREDITTORE	p ₁	p ₂	p ₃	g ₁	g ₂	g ₃
AA Freq	347,2	402,1	235,2			
AA Wpres	343,1	394,5	235,2			
AA MDO	341,7	437,5	235,2		+35,3	
AA OAFreq	347,1	402,1	235,2			
AA OAWpres	342,7	400,4	235,2			
AA OAMDO	341,1	452,4	235,2		+50,2	
CN Freq	310,4	327,8	230,5		+4,9	
CN Wpres	319,7	328,7	230,5	+0,2	+5,8	
CN OAFreq	351,0	399,5	230,5	+31,5	+76,6	
CN OAFreq	312,7	339,4	230,5		+16,5	
CN OAWpres	321,2	328,8	230,5	+1,7	+5,9	
CN OAMDO	347,5	479,0	230,5	+28,0	+156,1	
JC Freq	266,0	379,6	230,3	+1,2		
JC Wpres	291,3	379,7	230,2	+26,5		
JC MDO	659,2	385,2	230,3	+394,4	+5,5	
JC OAFreq	265,8	389,5	230,3	+1	+9,8	
JC OAWpres	297,8	379,9	230,3	+33	+0,2	
JC OAMDO	663,0	421,1	230,3	+398,2	+41,4	
PA Freq	4,5	2,9	3,1	+1,9	+01	
PA Wpres	2,7	2,9	3,1	+0,1	+0,1	
PA MDO	4,9	2,9	3,1	+2,3	+0,1	
PA OAFreq	4,9	2,9	3,1	+2,3	+0,1	
PA OAWpres	2,7	2,8	3,1	+0,1		
PA OAMDO	6,1	2,9	3,1	+3,5	+0,1	

Table 21: IMDB - Performance Moltiplicatori Temporali

5.3.2 Predittori Ad-Hoc

I predittori Ad-Hoc ottengono degli ottimi risultati sulla rete analizzata: tutti i modelli presentati, ad eccezione di AM, esprimono performance migliori dei modelli predittivi analizzati sino ad ora. Come riportato in Tabella 22 i modelli predittivi con performance migliori di questo gruppo sono quelli basati su Weighted Dimension Relevance, in particolare la versione affiancata da informazioni temporali di Weighted Presence. Un altro modello che ottiene ottime performance (seppur presentando un insieme più ridotto di risultati) è Triangle che, come mostrato in Figura 38b, riesce a dominare le curve degli altri predittori se analizzato nello spazio Precision-Recall.

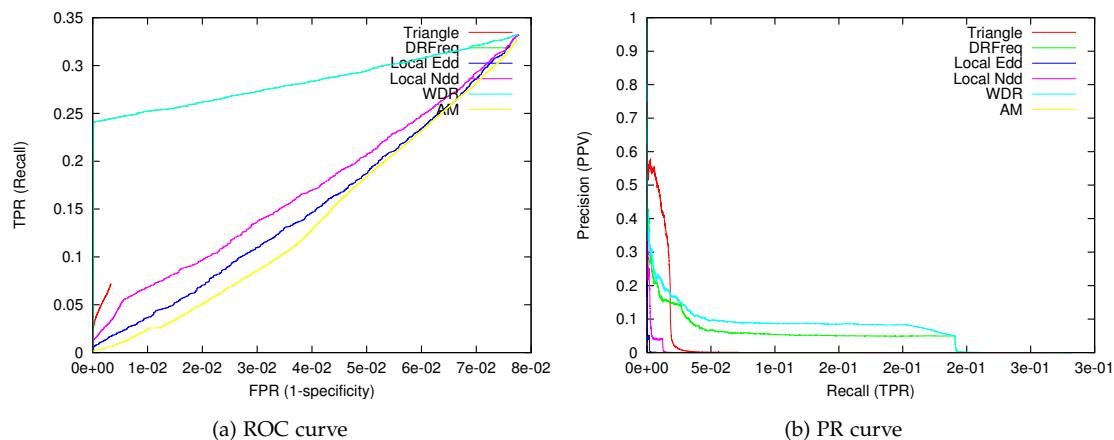


Figura 38: IMDB Modelli AdHoc

PREDITTORE	p ₁	p ₂	p ₃
LEdd	74,4	71,0	71,4
LNdd	103,1	76,9	71,5
WDRw	42 950,5	40 910,6	71,6
WDRf	27 636,5	24 216,9	71,6
AM	58,7	69,7	60,0
Triangle	7 715,9	755,3	450,5

Table 22: IMDB - Performance Modelli AdHoc

5.3.3 Analisi Riassuntiva

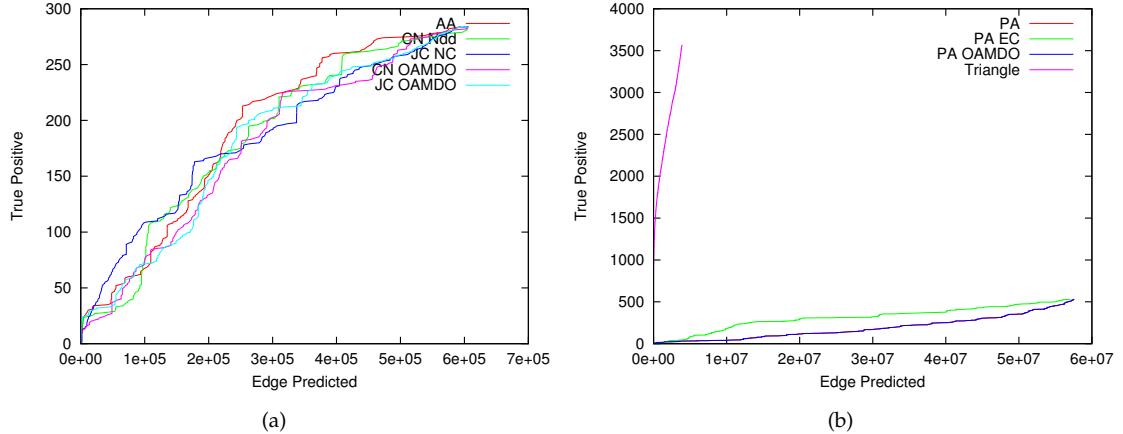


Figura 39: IMDB Analisi riassuntiva

Nella presentazione dei risultati ottenuti abbiamo già evidenziato come il predittore Weighted Dimension Relevance (con informazioni di Weighted Presence) riesca ad ottenere i migliori risultati su questa rete: per tale motivo in Figura 39 si sono riportati solo i migliori risultati ottenuti dai modelli derivati, rispettivamente, da Adamic Adar, Common Neighbours, Jaccard (grafico (a)) e Preferential Attachment (grafico (b)), in cui si è inserito anche il predittore AdHoc Triangle per completare l'analisi). Complessivamente si può osservare dai grafici che:

- nel caso dei derivati dai modelli Adamic Adar, Common Neighbours e Jaccard il solo coefficiente Node Correlation (in congiunzione di Jaccard) riesce a migliorare la precision nell'assegnamento degli score più alti agli archi predetti relativamente al migliore dei modelli base;
- analizzando i derivati da Preferential Attachment si ottengono i migliori risultati tramite il moltiplicatore Edge Correlation;
- il predittore Triangle ottiene ottimi risultati se comparato con i modelli base.

Nel caso della Figura 39a è interessante far notare che i predittori proposti migliorano ciascuno il proprio modello base: Adamic Adar risulta essere il modello meno sensibile all'introduzione di informazioni multidimensionali/temporali mentre gli altri riescono a sfruttarle per innalzare le loro performance tendendo al raggiungimento di quelle del migliore modello predittivo di base.

5.4 GTD: GLOBAL TERRORISM DATABASE

5.4.1 Predittori derivati dai modelli Monodimensionali

MODELLO BASE

Nel caso di GTD gli insiemi dei risultati forniti dai quattro modelli predittivi base risultano di dimensioni comparabili pertanto, con buona approssimazione, possiamo considerare equivalenti i valori di soglia stabiliti per tutti i modelli.

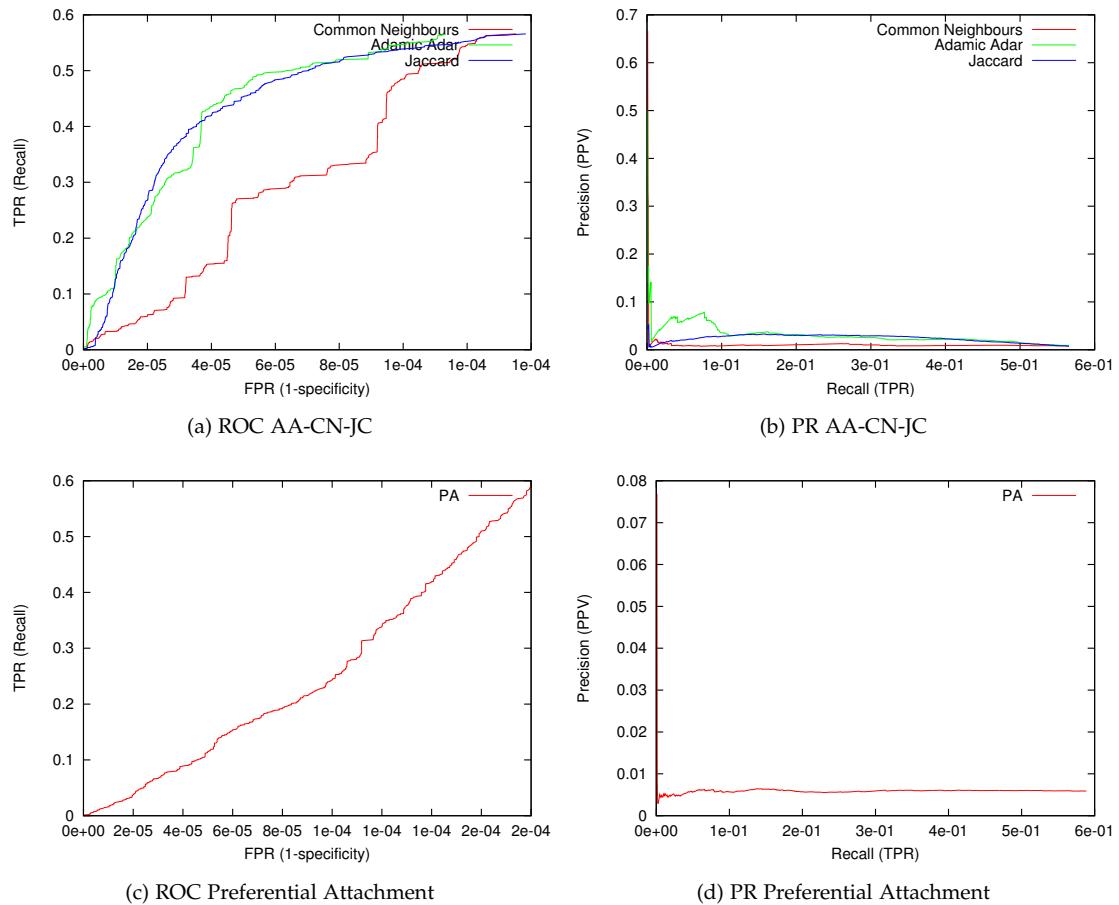


Figura 40: GTD Modelli Base

PREDITTORE	p ₁	p ₂	p ₃
AA	10 335,3	7 269,2	2 812,1
CN	2 866,3	2 784,2	2 357,5
JC	9 853,9	8 908,6	2 305,5
PA	1 966,3	1 856,7	1 824,2

Table 23: GTD - Performance modelli Base

Come mostrato nei grafici e espresso numericamente in Tabella 23 sono due i predittori che segnano le performance migliori su questo dataset: Adamic Adar (1° e 3° soglia) e Jaccard (2° soglia).

MULTIDIMENSIONALI LOCALI

I modelli che sfruttano le informazioni multidimensionali di ambito locale ai nodi non riescono ad ottenere performance migliori di quelle fornite dai modelli base.

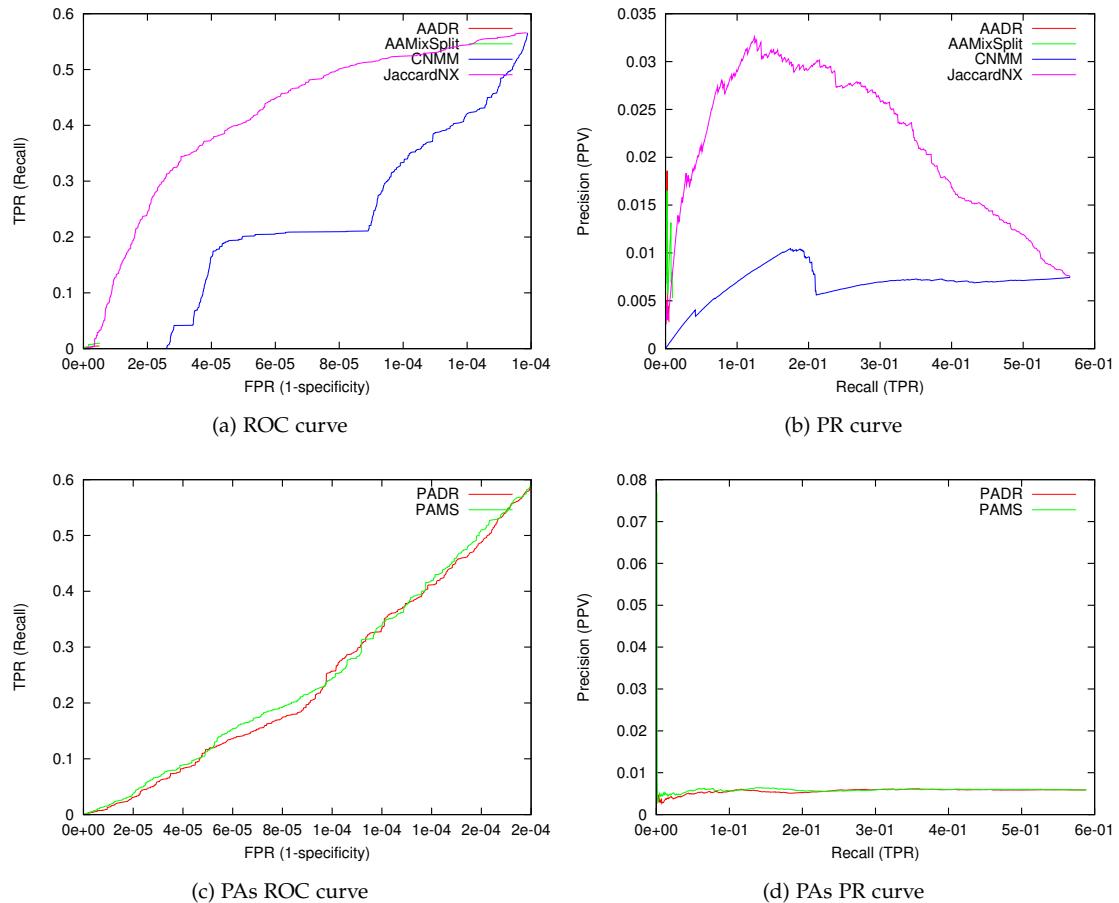


Figura 41: GTD Modelli Multidimensionali Locali

PREDITTORE	p ₁	p ₂	p ₃
AAMixSplit	3 137,0	5 090,4	5 090,4
AADR	3 545,7	5 733,8	5 733,8
CNMM	2 715,9	2 193,2	2 193,2
JaccardNX	9 392,4	7 719,0	7 719,0
PAMixSplit	1 975,5	1 826,0	1 826,0
PADR	1 767,5	1 873,5	1 873,5

Table 24: GTD - Performance Modelli Multidimensionali Locali

L'unico preditore che riesce ad avvicinarsi alle performance ottenute dal rispettivo modello base è JaccardNX, come mostrato in Tabella 24.

MOLTIPLICATORI MULTIDIMENSIONALI GLOBALI

Analizzando i risultati ottenuti tramite l'introduzione dei moltiplicatori multidimensionali globali si rileva una forte incidenza, sia in positivo che in negativo, sulle performance predittive. Tale informazione, esplicitata chiaramente dai grafici in figura 42 e 43, è evidenziata dai risultati proposti in Tabella 25.

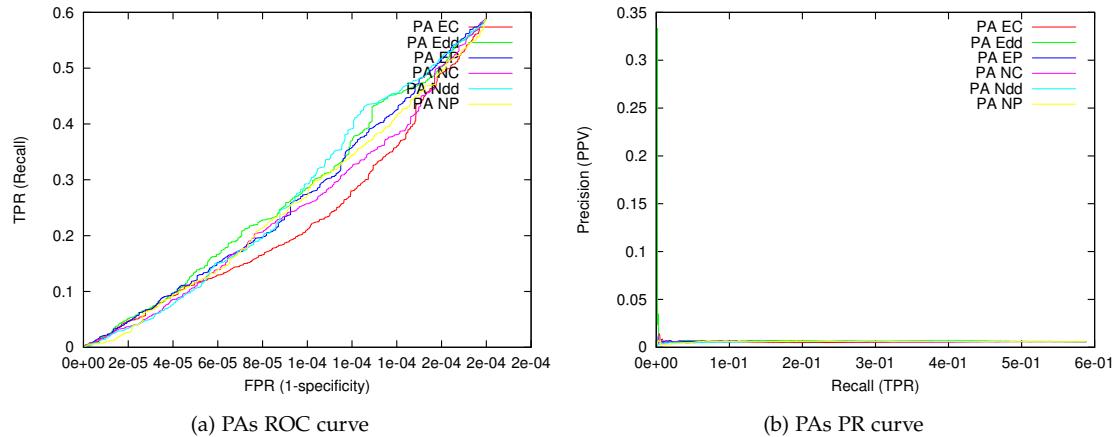


Figura 42: GTD Moltiplicatori Multidimensionali Globali I

Come riportato i coefficienti multidimensionali che riescono a migliorare (localmente) le performance dei predittori analizzati sono principalmente quelli facenti uso delle misure di Local Edge\Node Dimension Degree. In particolare Common Neighbours risulta essere il modello che beneficia in modo più accentuato dell'introduzione di tali moltiplicatori.

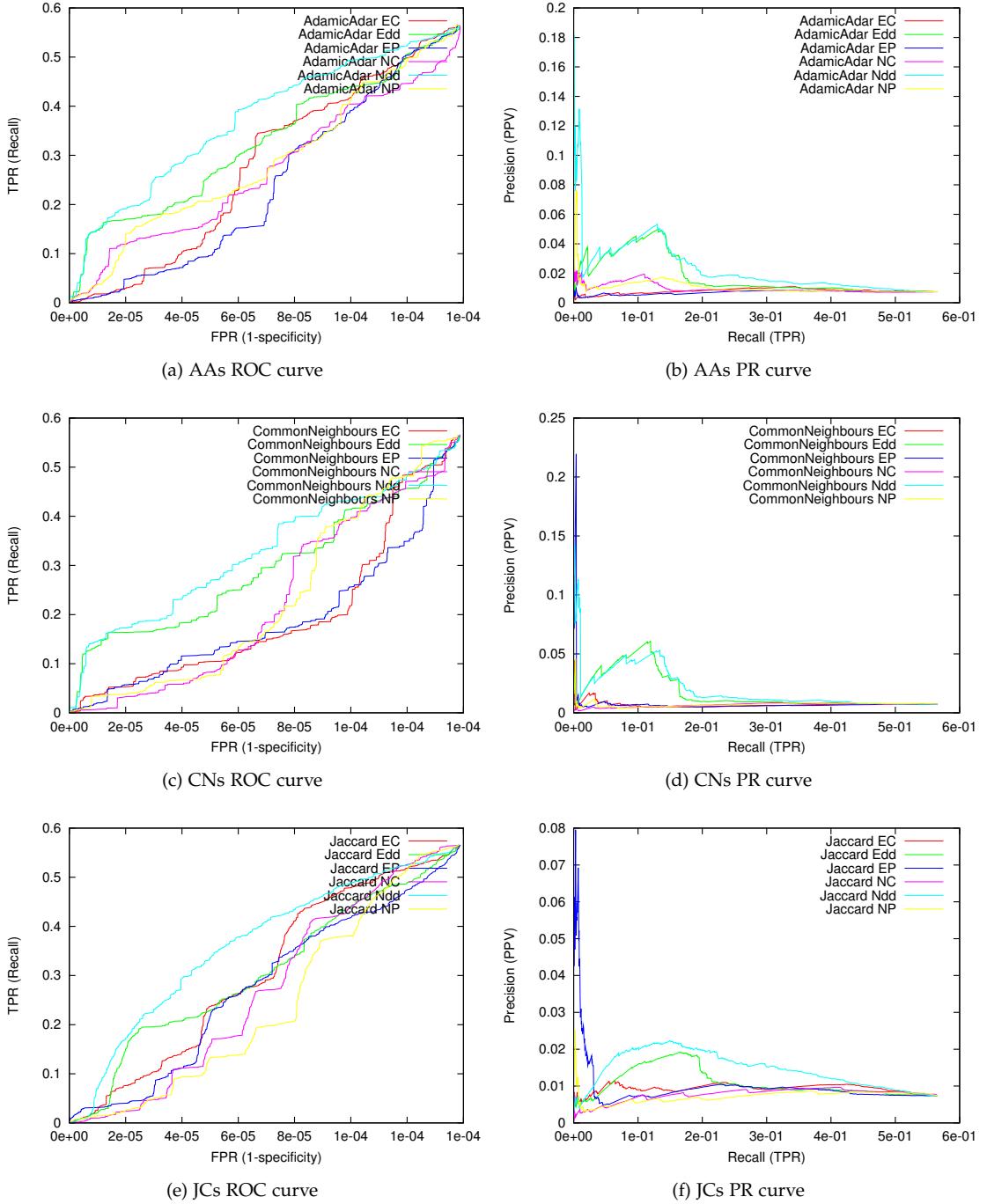


Figura 43: GTD Moltiplicatori Multidimensionali Globali II

PREDITTORE	p ₁	p ₂	p ₃	g ₁	g ₂	g ₃
AA EC	2 209,0	3 204,1	2 314,8			
AA Edd	14 961,6	3 246,9	2 300,7	+4 626,3		
AA EP	1 931,7	2 611,2	2 302,5			
AA NC	3 435,4	2 611,2	2 302,5			
AA Ndd	15 189,3	4 443,4	2 301,7	+4 854,0		
AA NP	5 263,8	2 558,1	2 298,1			
CN EC	1 649,5	1 949,8	2 305,1			
CN Edd	10 218,2	2 815,5	2 299,7	+7 351,9	+31,3	
CN EP	1 861,0	1 884,3	2 299,6			
CN NC	1 651,3	2 665,2	2 298,4			
CN Ndd	14 966,8	3 268,1	2 300,7	+12 100,5	+438,9	
CN NP	1 746,5	2 416,8	2 304,2			
JC EC	2 854,8	2 876,2	2 300,5			
JC Edd	5 465,6	2 887,5	2 296,5			
JC EP	2 333,5	2 941,5	2 297,5			
JC NC	2 201,0	2 737,8	2 301,8			
JC Ndd	6 604,8	4 658,9	2 307,5		+2,0	
JC NP	1 894,7	2 515,1	2 301,9			
PA EC	1 642,6	1 670,8	1 823,7			
PA Edd	2 094,5	1 918,3	1 825,1	+128,2	+61,6	+0,9
PA EP	1 934,2	1 896,6	1 822,2		+39,9	
PA NC	1 779,5	1 811,8	1 824,4			+0,2
PA Ndd	1 913,1	2 023,2	1 822,3		+166,5	
PA NP	1 792,7	1 890,9	1 819,4		+34,2	

Table 25: GTD - Performance Moltiplicatori Multidimensionali Globali

MOLTIPLICATORI TEMPORALI

Le informazioni temporali introducono, come mostrato in Tabella 26, diffusi innalzamenti nelle performance (sia locali che globali) rispetto ai modelli predittivi presi in esame.

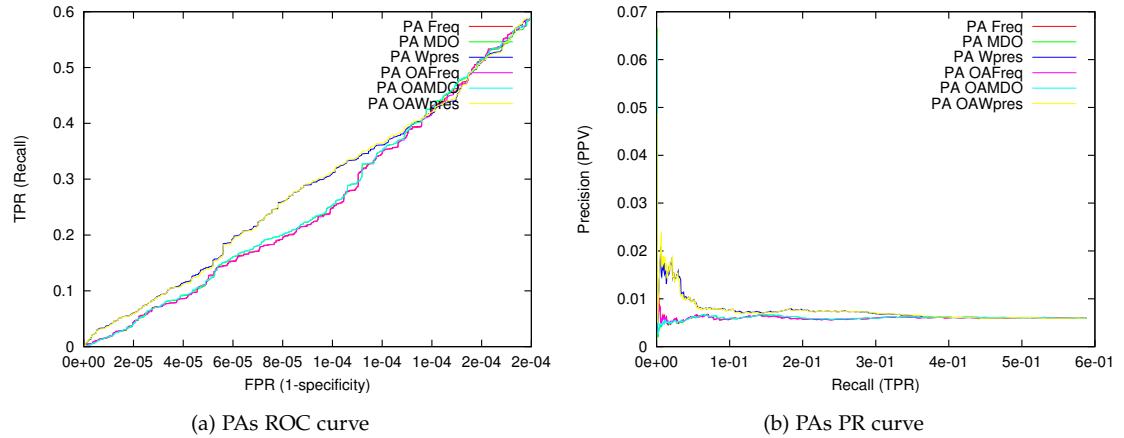


Figura 44: GTD Moltiplicatori Temporali I

Le informazioni relative all Weighted Presence (sia calcolata su singola dimensione che nella versione Over All) fanno registrare gli incrementi più significativi su tutti i modelli analizzati ad eccezione di Jaccard dove sembrano ottenere maggiori performance i modelli derivati estesi con i moltiplicatori facenti uso di Frequency.

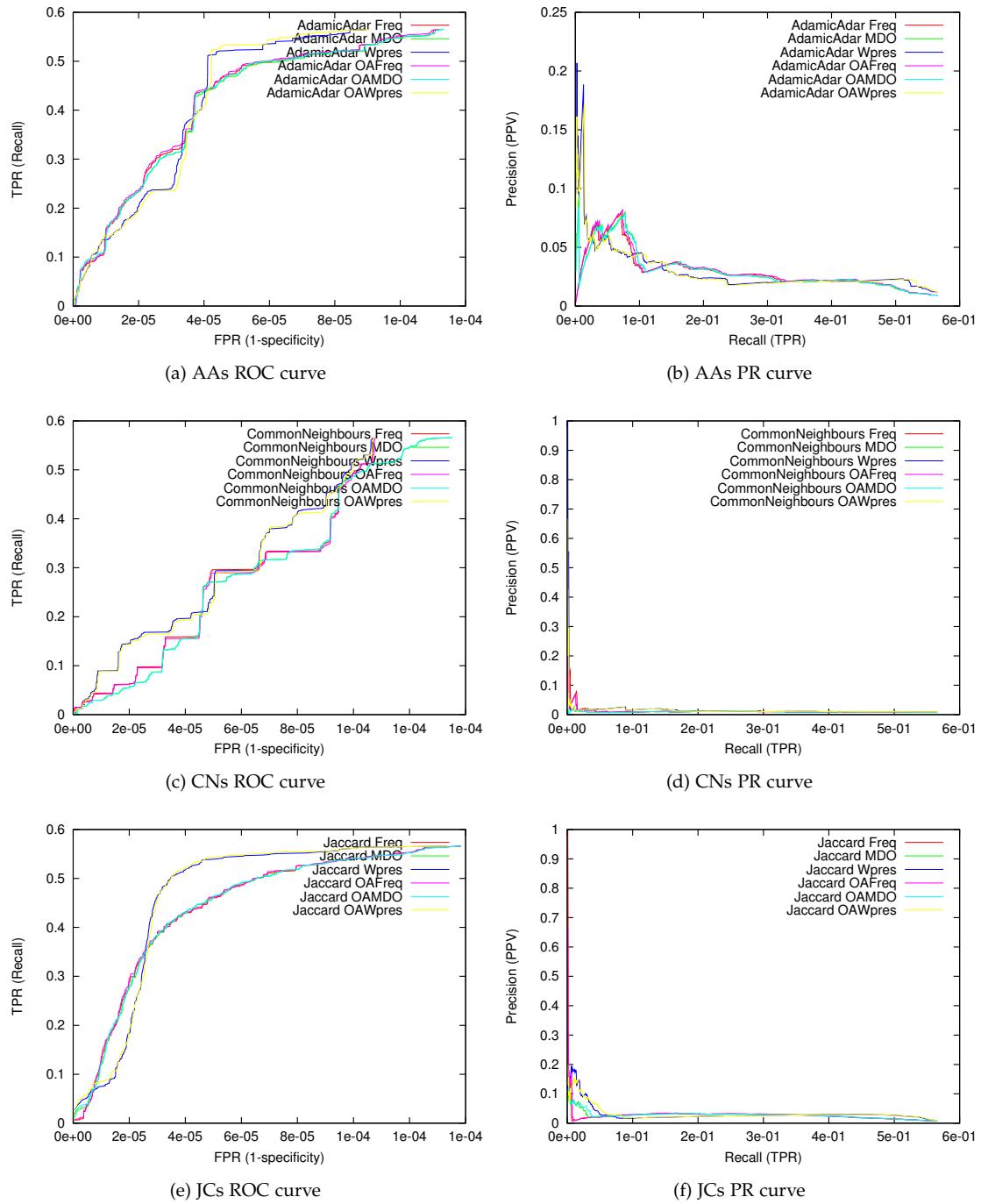


Figura 45: GTD Moltiplicatori Temporali II

PREDITTORE	p ₁	p ₂	p ₃	g ₁	g ₂	g ₃
AA Freq	10 520,3	6 916,6	2 852,7	+185,0		+40,6
AA Wpres	9 725,1	6 297,4	3 565,2			+753,1
AA MDO	10 446,0	6 376,4	2 812,0	+110,7		
AA OAFreq	10 597,4	7 233,6	2 851,0	+262,1		+38,9
AA OAWpres	11 303,4	6 084,4	3 529,3	+968,1		+717,2
AA OAMDO	10 283,9	6 335,7	2 809,6			
CN Freq	3 273,2	3 097,5	2 965,3	+406,9	+313,3	+607,8
CN Wpres	6 148,1	3 188,8	2 982,7	+3 281,8	+404,6	+625,2
CN MDO	2 907,2	2 923,4	2 357,4	+40,9	+139,2	
CN OAFreq	3 262,7	3 096,8	2 947,8	+396,4	+312,6	+570,3
CN OAWpres	5 995,6	3 179,4	2 964,7	+3 129,3	+395,2	+607,2
CN OAMDO	2 900,3	2 915,5	2 357,0	+34,0	+131,3	
JC Freq	9 973,2	9 241,9	2 306,4	+119,3	+333,3	+0,9
JC Wpres	5 988,0	8 343,0	2 392,6			+87,1
JC MDO	9 385,1	9 023,7	2 305,4		+151,1	
JC OAFreq	10 305,1	9 329,9	2 306,3	+451,2		+0,8
JC OAWpres	6 300,2	8 196,3	2 391,4			+85,9
JC OAMDO	9 368,5	8 955,1	2 305,4		+46,5	
PA Freq	2 046,6	1 909,1	1 828,1	+80,3	+52,4	+3,9
PA Wpres	2 226,8	2 074,1	1 833,7	+260,5	+217,4	+9,5
PA MDO	2 026,8	1 870,8	1 824,2	+60,5	+14,1	
PA OAFreq	2 029,6	1 901,2	1 827,5	+63,3	+44,5	+3,3
PA OAWpres	2 119,9	2 088,6	1 833,0	+153,6	+231,9	+8,8
PA OAMDO	2 024,1	1 947,6	1 824,2	+57,8	+90,9	

Table 26: GTD - Performance Moltiplicatori Temporali

5.4.2 Preditori Ad-Hoc

Il gruppo costituito dai predittori Ad-Hoc ottiene, complessivamente, risultati altalenanti se paragonato a quelli forniti dal migliore modello predittivo base analizzato per questa rete (Adamic Adar).

I predittori che mostrano i risultati più interessanti, configurandosi globalmente come i migliori per questa rete, sono quelli basati su Weighted Dimension Relevance che come mostrato in Tabella 27 e Figura 46a riescono a predire efficientemente un alto numero di archi assegnandogli un corretto score di confidenza.

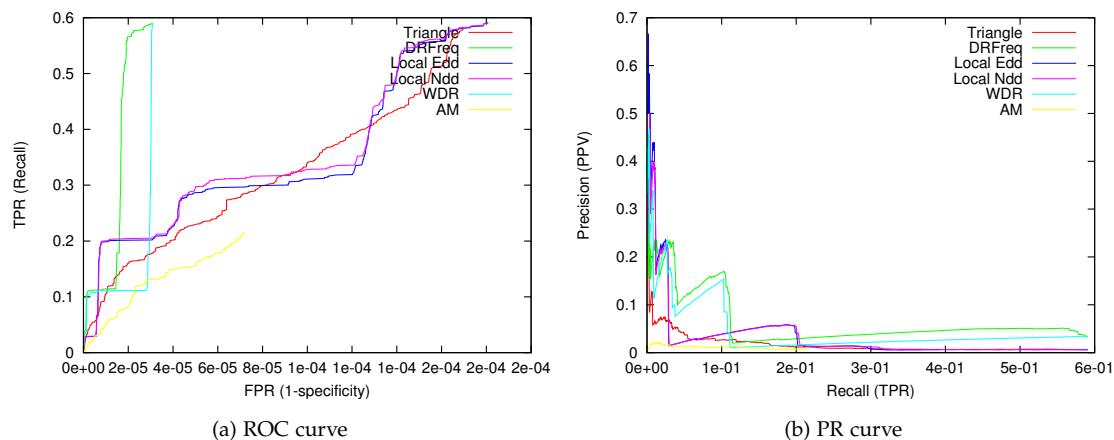


Figura 46: GTD Modelli Ad Hoc

PREDITTORE	p ₁	p ₂	p ₃
LEdd	15 410,2	1 842,7	1 935,3
LNdd	15 160,5	18 51,6	1 933,8
WDRw	3 736,8	75 61,4	10 303,5
WDRf	7 208,2	13 335,7	11 740,8
AM	3 558,0	2 891,6	2 183,0
Triangle	6 866,6	21 95,3	1 910,5

Table 27: GTD - Performance Modelli AdHoc

5.4.3 Analisi Riassuntiva

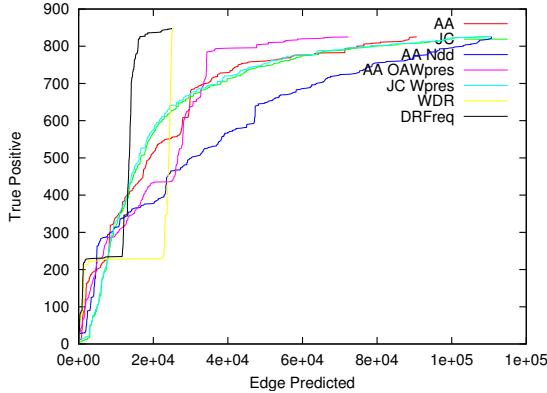


Figura 47: GTD Analisi riassuntiva

In Figura 47 si riporta il grafico dell’andamento della Precision dei predittori derivati, che segnano i maggiori incrementi rispetto al relativo modello base, e dei migliori predittori AdHoc.

Ad una prima analisi risulta evidente che, per quanto riguarda i predittori derivati, gli incrementi di performance non sono globali all’intero insieme dei risultati ma, al contrario, sono molto spesso locali a particolari fasce di assegnazione di score: questo risultato, già evidenziato nell’analisi precedente, è portato all’estremo se si considerano nel complesso anche i due predittori AdHoc le cui curve compaiono nel grafico riportato.

I predittori basati su Weighted Dimension Relevance riescono ad ottenere una performance complessiva migliore di qualsiasi altro modello predittivo analizzato ma, al contempo, se valutati sulla 2°soglia di risultati ottengono prestazioni inferiori a quelle degli altri predittori derivati dai modelli di base per mezzo di moltiplicatori multidimensionali globali o temporali.

5.5 GCD: GRAND COMICS DATABASE

5.5.1 Predittori derivati dai modelli Monodimensionali

MODELLO BASE

I modelli predittivi di base, se applicati alla rete costruita sul dataset GCD, restituiscono insiemi di risultati aventi lo stesso numero di elementi ed è quindi possibile comparare direttamente le performance ottenute sulle soglie presentate.

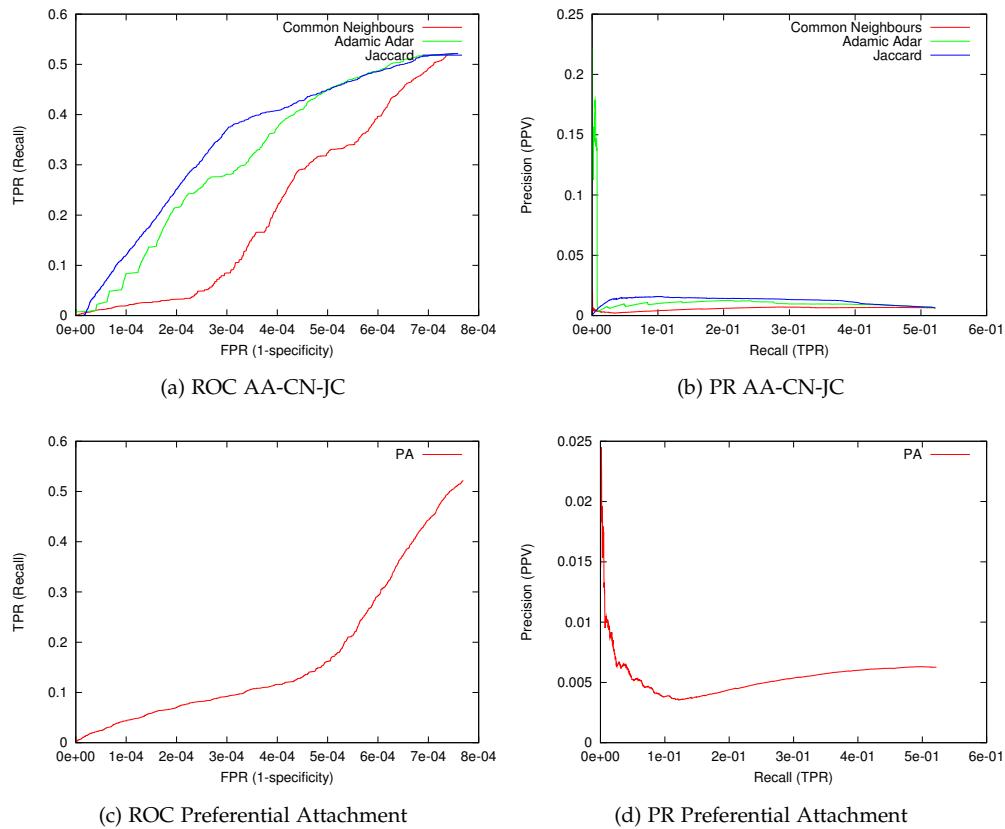


Figura 48: GCD Modelli Base

PREDITTORE	p ₁	p ₂	p ₃
AA	797,7	695,1	452,1
CN	336,2	495,6	453,9
JC	1 063,7	941,4	448,9
PA	254,6	376,4	442,9

Table 28: GCD - Performance modelli Base

Come espresso in Tabella 28 i predittori che performano, localmente, meglio su questa rete sono Jaccard (1° e 2° soglia) e Common Neighbours (3° soglia). Valutando complessivamente gli andamenti dei predittori è possibile affermare che il migliore dei modelli base risulta quindi essere Jaccard.

MULTIDIMENSIONALI LOCALI

JaccardNX è l'unico, tra i modelli aventi un insieme di risultati comparabile a quelli dei predittori base, che riesce a presentare un innalzamento delle performance rispetto a quelle riportate in Tabella 28.

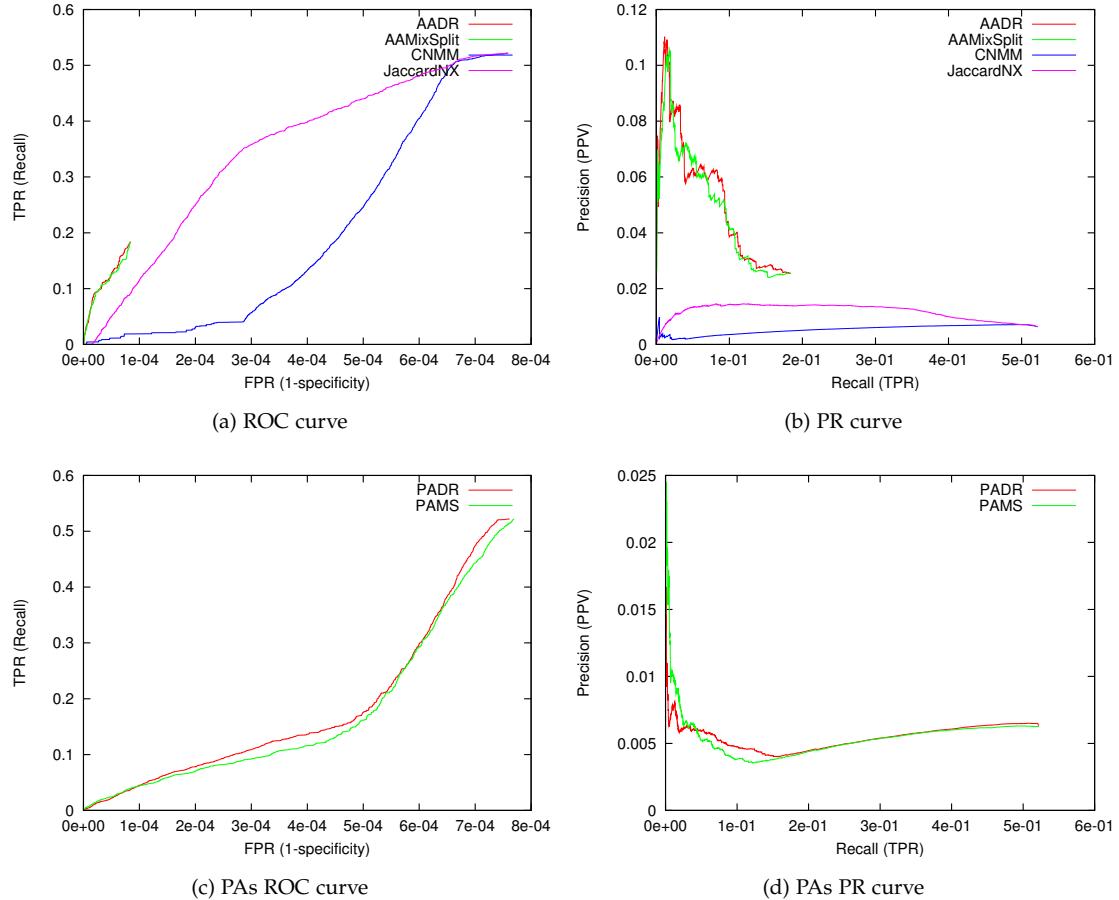


Figura 49: GCD Modelli Multidimensionali Locali

PREDITTORE	p ₁	p ₂	p ₃
AAMixSplit	4 608,1	2 160,8	1 800,3
AADR	4 287,4	2 255,3	1 799,7
CNMM	287,2	425,8	449,4
JaccardNX	1 019,3	960,2	448,7
PAMixSplit	254,6	376,4	442,9
PADR	310,1	379,3	447,4

Table 29: GCD - Performance Modelli Multidimensionali Locali

I predittori derivati da Adamic Adar, seppur restituendo un numero inferiore di risultati rispetto agli altri modelli, riescono a catturare un maggior numero di risultati TP sul totale delle predizioni effettuate garantendo una performance molto alta (come mostrato in Tabella 29, e evidenziato dalle curve di ROC e PR che dominano quelle dei gli altri modelli in Figura 49).

MOLTIPLICATORI MULTIDIMENSIONALI GLOBALI

I moltiplicatori multidimensionali globali agiscono sensibilmente sulla precisione dei predittori analizzati (come si osserva nei grafici in figura 50 e 51).

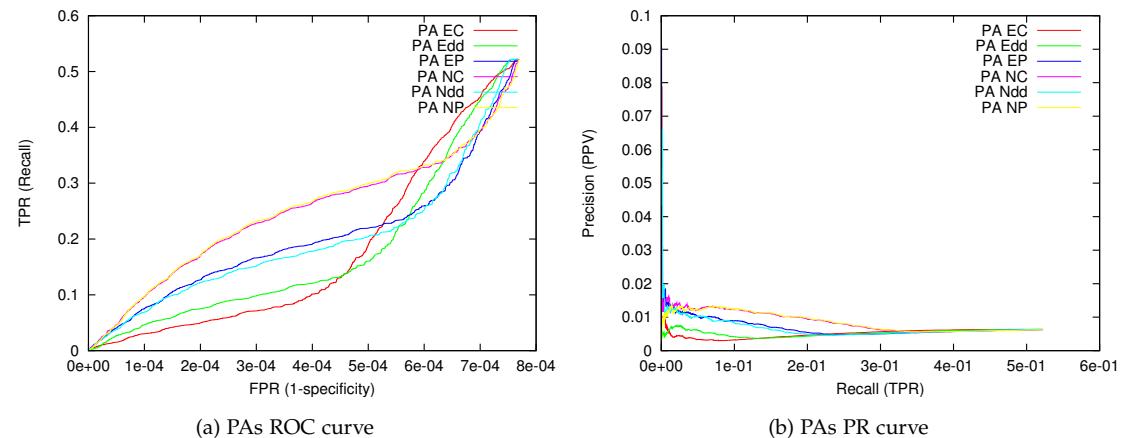


Figura 50: GCD Moltiplicatori Multidimensionali Globali I

I coefficienti multidimensionali che migliorano le performance in modo più consistente per questa rete sono: Node Dimension Relevance (Adamic Adar), Node Parent (Common Neighbours e Preferential Attachment), Edge Correlation (Jaccard) e Node Correlation (Preferential Attachment).

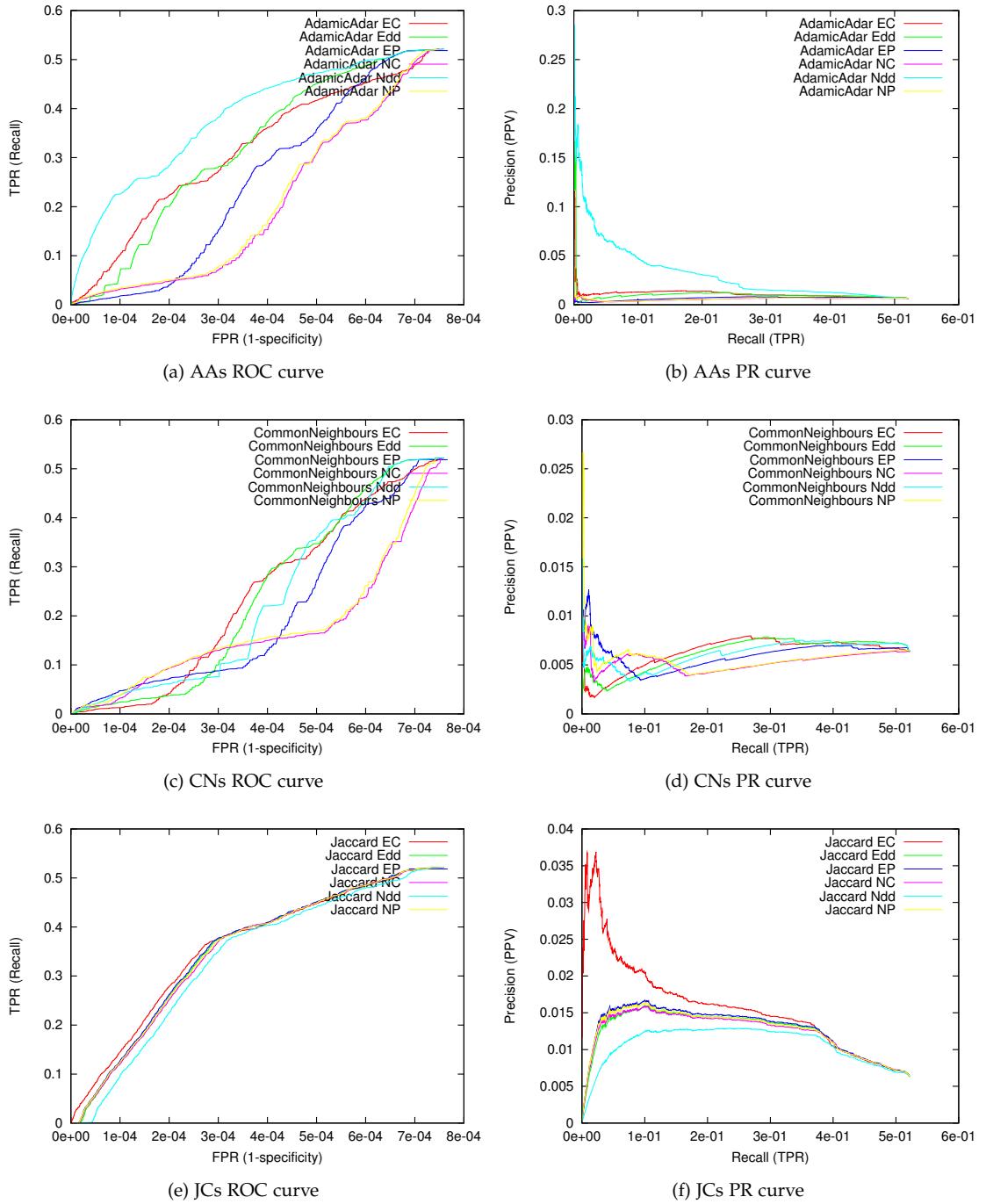


Figura 51: GCD Moltiplicatori Multidimensionali Globali II

PREDITTORE	p ₁	p ₂	p ₃	g ₁	g ₂	g ₃
AA EC	959,1	702,6	452,6	+153,4	+7,5	+0,5
AA Edd	703,7	686,2	454,9			+2,8
AA EP	398,8	559,6	448,9			
AA NC	307,8	460,6	452,1			
AA Ndd	2 825,3	1 075,8	448,9	+2 027,6	+380,7	
AA NP	317,5	461,9	452,1			
CN EC	394,8	543,1	449,0	+58,6	+47,5	
CN Edd	356,8	554,7	448,9	+20,6	+59,1	
CN EP	284,9	443,5	449,0			
CN NC	385,7	361,0	450,6	+49,5		
CN Ndd	312,4	495,2	449,0			
CN NP	399,5	362,5	456,9	+63,3		+4,8
JC EC	1 270,5	1 036,6	448,9	+206,8	+95,2	
JC Edd	1 074,1	967,7	448,7	+10,4	+26,2	
JC EP	1 110,4	981,1	448,8		+39,7	
JC NC	1 063,2	942,5	448,9		+1,1	
JC Ndd	887,4	880,6	449,0			
JC NP	1 091,3	958,7	448,9	+27,6	+17,3	
PA EC	255,7	397,8	442,9		+21,4	
PA Edd	262,7	375,1	443,5	+8,1		+0,6
PA EP	560,1	354,5	443,5	+305,5		+0,6
PA NC	795,3	450,5	443,0	+540,7	+74,1	
PA Ndd	495,8	355,6	443,7	+241,2		+0,8
PA NP	799,8	464,6	442,8	+545,2	+88,2	

Table 30: GCD - Performance Moltiplicatori Multidimensionali Globali

MOLTIPLICATORI TEMPORALI

Analizzando i grafici riportati e i risultati proposti in Tabella 31 si nota che le informazioni temporali associate alla rete GCD non sono sfruttate in modo proficuo da tutti i modelli base presi in analisi.

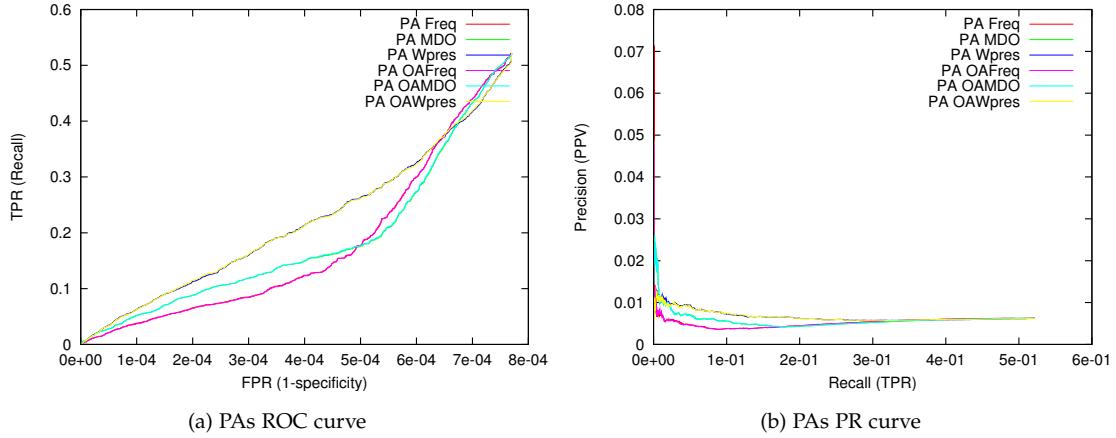


Figura 52: GCD Moltiplicatori Temporali I

In particolare si evidenzia la quasi completa assenza di miglioramento delle performance per i predittori derivati da Adamic Adar e Common Neighbours; meglio si comportano i modelli basati su Jaccard e Preferential Attachment se accompagnati da informazioni di Weighted Presence (sia calcolate per singola dimensione sia nella versione Over All).

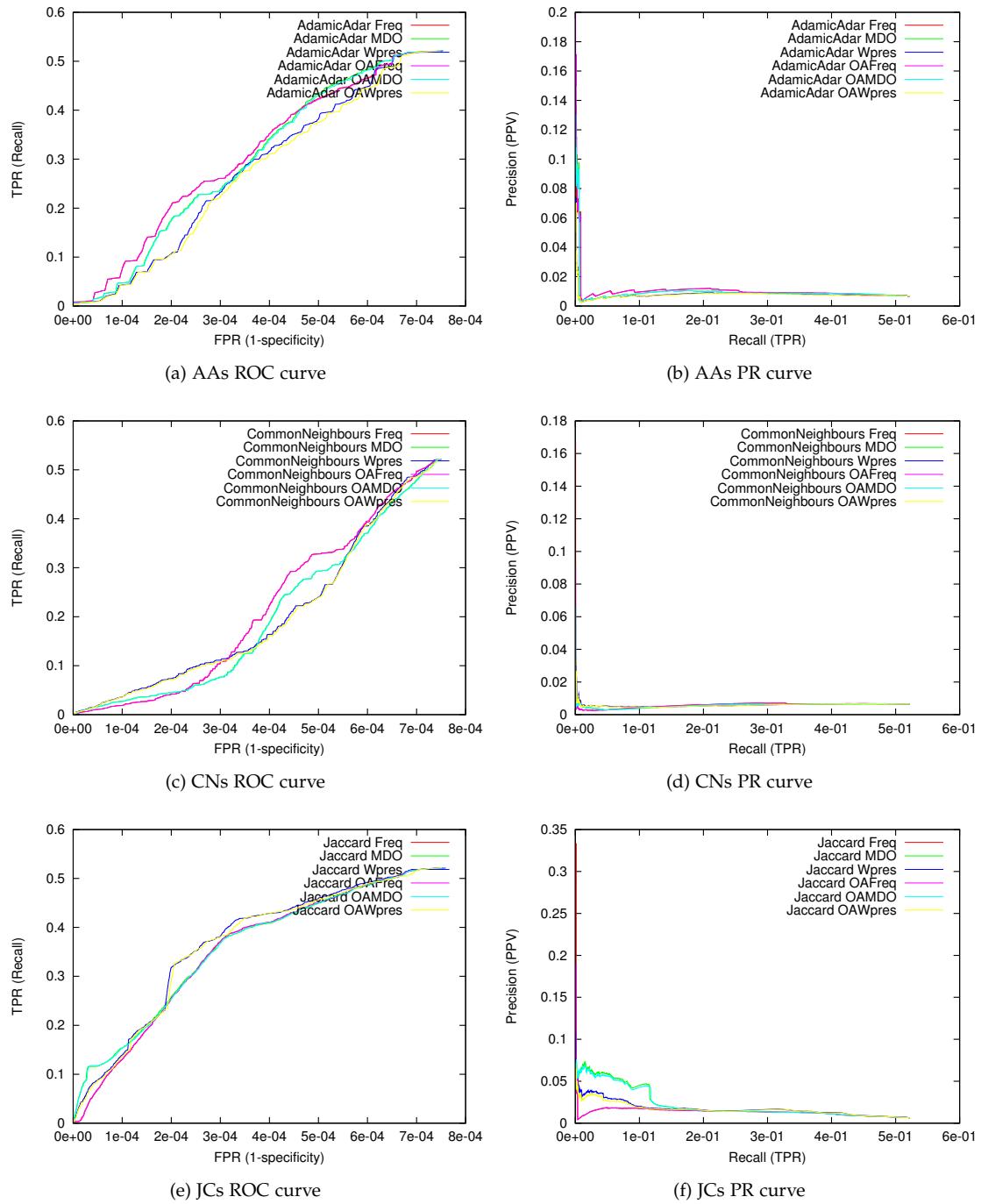


Figura 53: GCD Moltiplicatori Temporali II

PREDITTORE	p ₁	p ₂	p ₃	g ₁	g ₂	g ₃
AA Freq	778,7	649,1	452,1			
AA Wpres	517,3	619,3	452,1			
AA MDO	691,5	626,4	452,1			
AA OAFreq	776,8	649,6	452,8		+0,7	
AA OAWpres	499,9	585,8	454,3		+2,1	
AA OAMDO	685,8	622,3	451,6			
CN Freq	345,0	496,5	453,8	+8,8	+0,9	
CN Wpres	326,5	417,7	453,8			
CN MDO	310,3	436,9	453,9			
CN OAFreq	345,2	496,5	455,6	+9,0	+0,9	+1,7
CN OAWpres	316,5	416,7	455,4			+1,5
CN OAMDO	309,6	436,6	453,3			
JC Freq	1 141,7	955,9	448,9	+78,0	+14,5	
JC Wpres	1 236,8	1 147,9	448,9	+173,1	+206,5	
JC MDO	1 520,3	971,8	448,9	+456,6	+30,4	
JC OAFreq	1 150,3	958,5	449,0	+86,6	+17,1	+0,1
JC OAWpres	1 206,6	1 119,3	453,3	+142,9	+177,9	+4,4
JC OAMDO	1 507,9	960,3	448,9	+442,2	+18,9	
PA Freq	269,0	382,0	442,9	+14,4	+5,6	
PA Wpres	463,4	406,6	442,5	+208,8	+30,2	
PA MDO	337,2	369,0	442,5	+82,6	+7,4	
PA OAFreq	264,5	381,9	442,9	+9,9	+5,5	
PA OAWpres	474,0	207,3	442,4	+219,4		
PA OAMDO	339,7	369,7	442,5	+85,1		

Table 31: GCD - Performance Moltiplicatori Temporali

5.5.2 Predittori Ad-Hoc

I predittori facenti parte di questo gruppo riescono, localmente, ad ottenere performance migliori di quelle segnalate per i modelli base.

In particolare le curve dei predittori Triangle e Weighted Dimension Relevance con informazioni di Frequency dominano (quasi) costantemente quelle dei modelli predittivi di base.

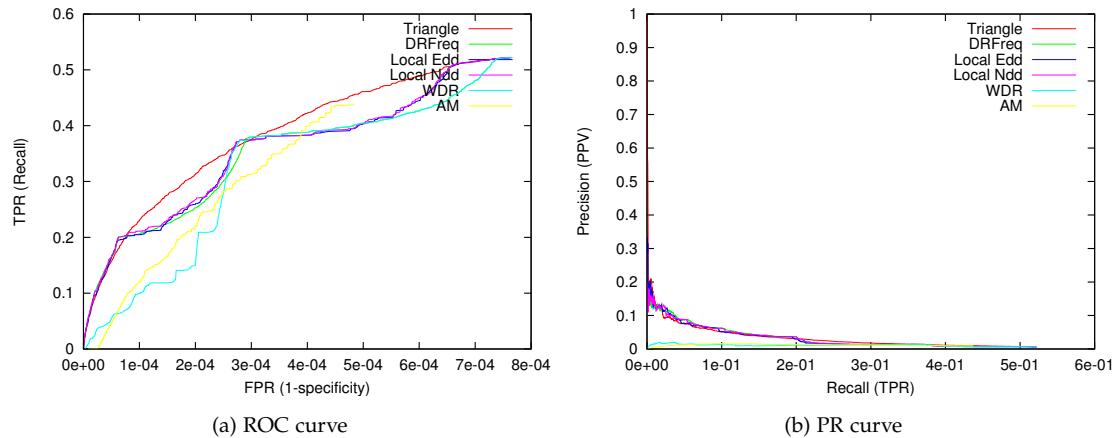


Figura 54: GCD Modelli AdHoc

PREDITTORE	p ₁	p ₂	p ₃
LEdd	3 071,7	939,7	445,8
LNdd	3 247,7	939,4	446,3
WDRw	686,3	891,7	444,6
WDRf	3 336,1	912,9	444,6
AM	1 084,8	855,6	623,2
Triangle	2 985,4	1 262,1	447,2

Table 32: GCD - Performance Modelli AdHoc

5.5.3 Analisi Riassuntiva

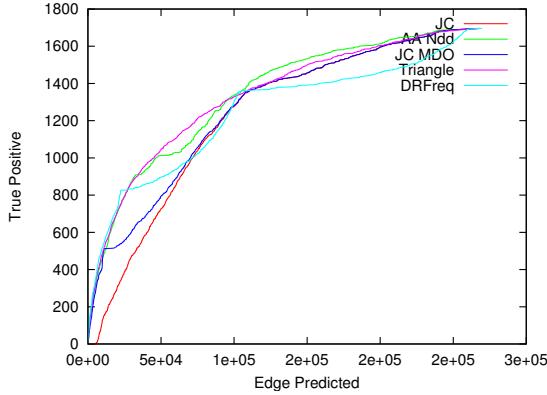


Figura 55: GCD Analisi riassuntiva

Il grafico presentato in Figura 55 mostra l'andamento dei modelli predittivi che ottengono le migliori performance sulla rete analizzata.

Tra le curve riportate in figura non esiste una che domini in modo definitivo le altre; possiamo riconoscere, per ogni soglia, un predittore che riesce a sfruttare al meglio le informazioni topologiche per assegnare gli score ai risultati ottenuti:

- 1° soglia: Weighted Dimension Relevance (con informazioni di Frequency)
- 2° soglia: Triangle
- 3° soglia: Adamic Adar con Node Dimension Degree

Possiamo notare, date le tipologie di modelli riportati in questa analisi riassuntiva, che le informazioni di tipo temporale e multidimensionale globale (e loro combinazioni) riescono ad influire in modo significativo sulle performance espresse dai modelli base.

5.6 VDC: INTERNATIONAL DYADIC EVENTS

5.6.1 Predittori derivati dai modelli Monodimensionali

MODELLO BASE

Analizzando separatamente i modelli predittivi appartenenti a questo gruppo in base alle dimensioni degli insiemi di risultati forniti possono essere fatte alcune osservazioni che rendano più chiaro il significato assunto dai valori numerici riportati in Tabella 33.

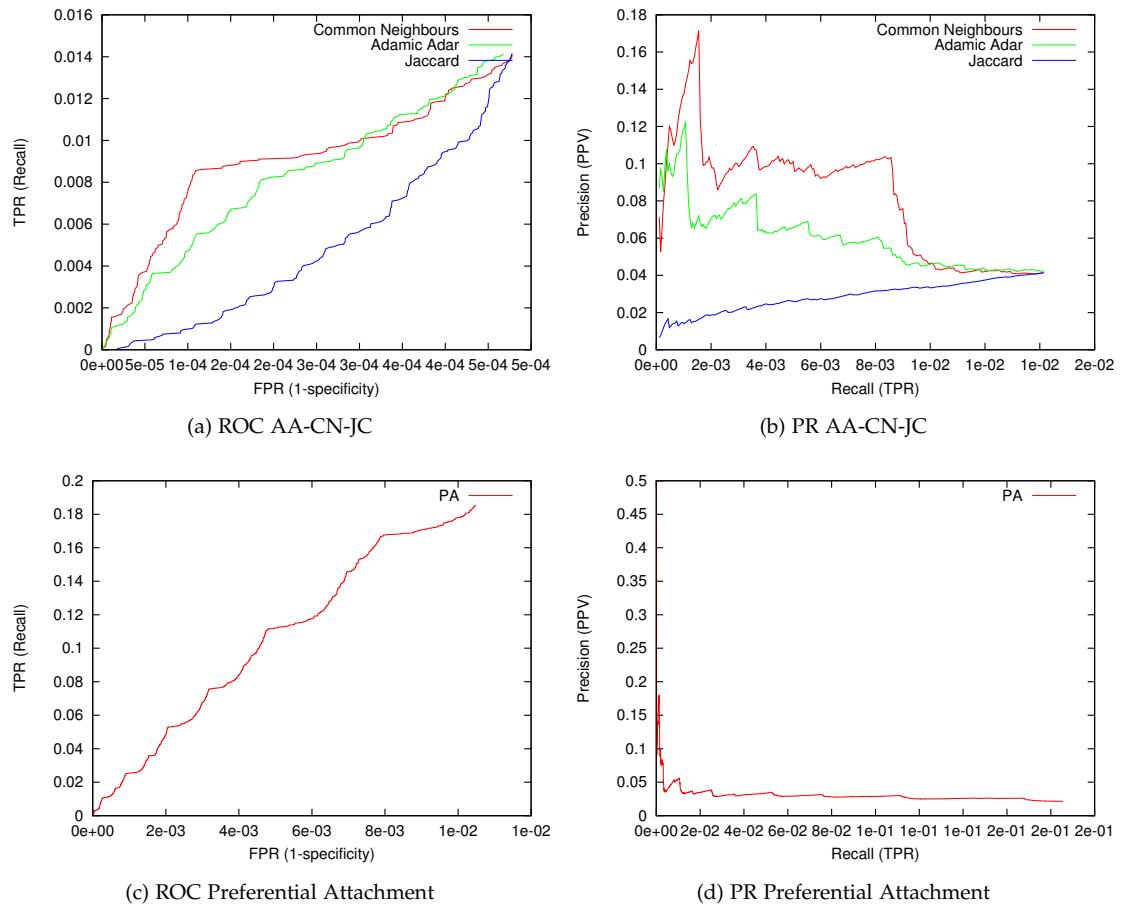


Figura 56: VDC Modelli Base

PREDITTORE	p ₁	p ₂	p ₃
AA	18,6	13,0	11,9
CN	28,1	14,4	11,6
JC	7,2	9,4	11,6
PA	8,4	7,2	6,0

Table 33: VDC - Performance modelli Base

Dai valori riportati appare infatti evidente una preponderanza nelle performance a carico di Common Neighbours (informazione confermata anche dagli andamenti delle relative curve ROC e PR mostrate nei grafici in figura 56a e 56b): considerando però anche la dimensione dell'insieme dei risultati si può osservare che la curva descritta da Preferential Attachment riesce a dominare parzialmente quella degli altri modelli all'altezza della 2° soglia dei risultati di questi ultimi.

MULTIDIMENSIONALI LOCALI

I modelli che introducono le informazioni multidimensionale localmente ai nodi non riescono ad ottenere performance che consentano alle curve descritte dai grafici di ROC e di PR di sovrastare quelle relative ai modelli base da cui sono derivati.

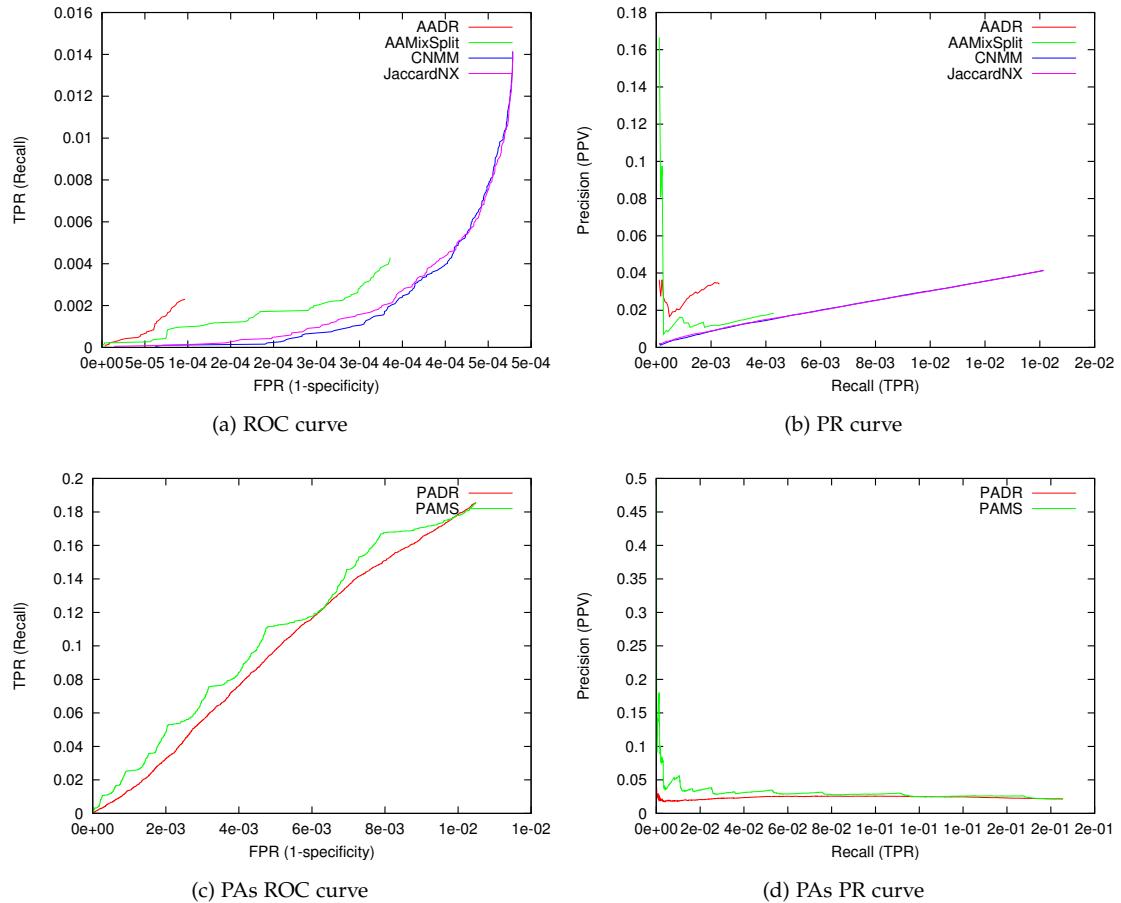


Figura 57: VDC Modelli Multidimensionali Locali

PREDITTORE	p ₁	p ₂	p ₃
AAMixSplit	3,3	3,8	5,2
AADR	5,8	8,2	9,5
CNMM	4,6	8,1	11,6
JaccardNX	4,6	8,0	11,6
PAMixSplit	8,4	7,2	6,0
PADR	7,1	1,0	6,0

Table 34: VDC - Performance Modelli Multidimensionali Locali

PAMixSplit risulta essere l'unico caso in cui la precision di un predittore di questo gruppo riesce ad eguagliare quella del suo modello base.

MOLTIPLICATORI MULTIDIMENSIONALI GLOBALI

Come mostrato in Tabella 35 la precision dei modelli base subisce solo lievi variazioni se questi vengono estesi tramite l'introduzione informazioni di tipo multidimensionale globali alla rete.

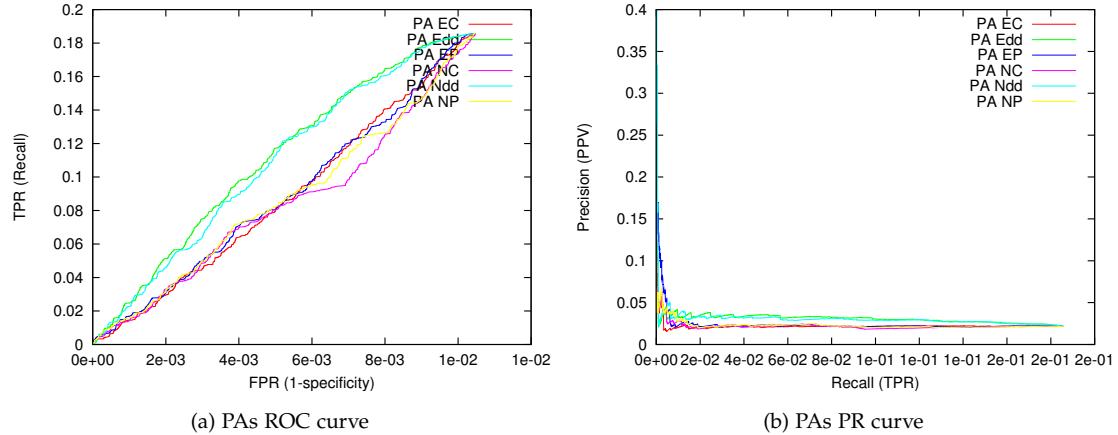


Figura 58: VDC Moltiplicatori Multidimensionali Globali I

In particolare gli incrementi più rilevanti sono segnati da Node/Edge Dimension Degree che riescono, in maggior misura con Adamic Adar (+9,5) e Jaccard (+5,9), ad innalzare le performance in modo significativo.

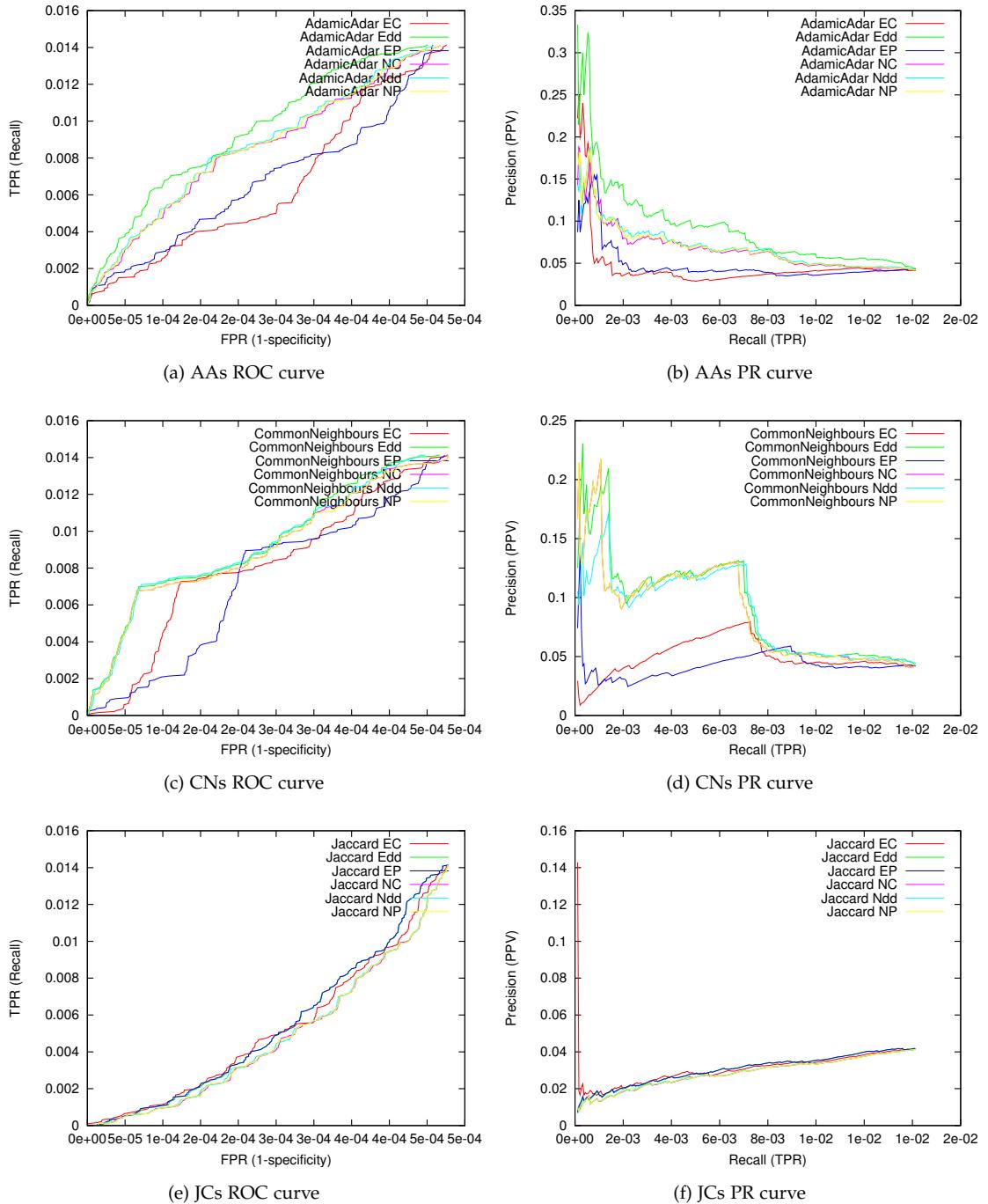


Figura 59: VDC Moltiplicatori Multidimensionali Globali II

PREDITTORE	p ₁	p ₂	p ₃	g ₁	g ₂	g ₃
AA EC	8,4	11,0	11,7			
AA Edd	28,1	17,0	12,3	+9,5	+4,0	+0,4
AA EP	12,4	10,3	12,1			+0,2
AA NC	19,9	13,9	11,9	+1,3	+0,9	
AA Ndd	19,6	14,8	12,1	+1,3	+1,8	+0,2
AA NP	19,5	14,3	11,9	+1,4	+1,3	
CN EC	17,5	12,5	11,6			
CN Edd	34,0	14,5	11,9	+5,9	+0,1	+0,3
CN EP	10,6	14,2	11,7			+0,1
CN NC	33,5	14,5	11,6	+5,4	+0,1	
CN Ndd	32,4	15,0	12,4	+4,3	+0,6	+0,8
CN NP	33,5	14,5	11,6	+5,4	+0,1	
JC EC	8,2	9,5	11,6	+1,0	+0,1	
JC Edd	7,7	9,8	11,6	+0,5	+0,4	
JC EP	7,7	9,8	11,6	+0,5	+0,4	
JC NC	7,3	9,4	11,6	+0,1		
JC Ndd	7,0	9,5	11,6		+0,1	
JC NP	7,3	9,4	11,6	+0,1		
PA EC	6,0	6,2	6,0			
PA Edd	9,7	8,5	6,0	+1,3	+1,3	
PA EP	6,3	6,2	6,0			
PA NC	6,4	5,4	6,0			
PA Ndd	9,2	8,2	6,1	+0,8	+1,0	+0,1
PA NP	6,3	5,9	6,0			

Table 35: VDC - Performance Moltiplicatori Multidimensionali Globali

MOLTIPLICATORI TEMPORALI

Le informazioni temporali sulla ricorrenza degli archi appartenenti alla rete risulta essere un'informazione molto utile per l'analisi di Link Prediction per la rete in esame: come si può notare infatti la Tabella 36 esplicita un sostanziale incremento della precisione di tutti i modelli analizzati se applicato uno qualsiasi dei moltiplicatori temporali proposti.

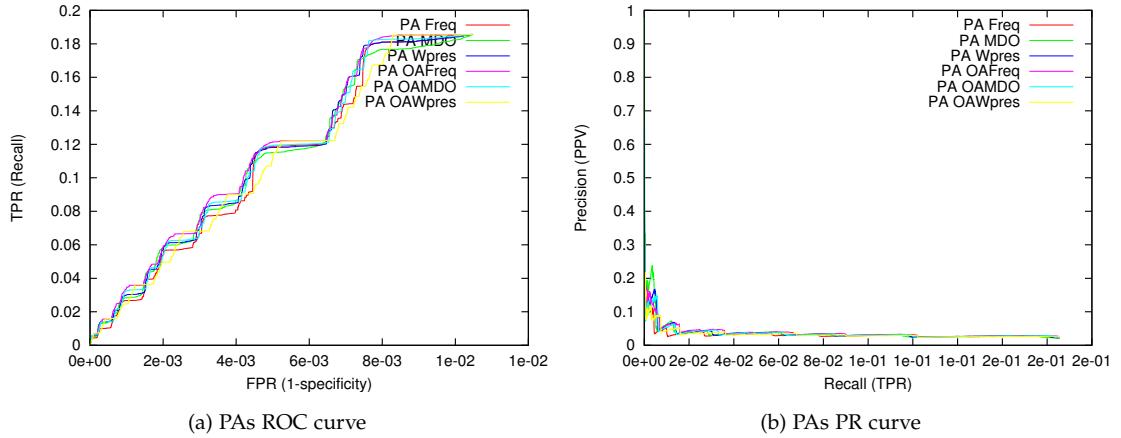


Figura 60: VDC Moltiplicatori Temporali I

Come si può notare nei grafici riportati in Figura 60 e in Figura 61 esiste una discreta varianza nelle curve raffiguranti l'andamento dei predittori nei sistemi di riferimento ROC e PR: non si delineano però moltiplicatori preferenziali, comuni a tutti i modelli tra quelli proposti, per l'analisi della rete in questione.

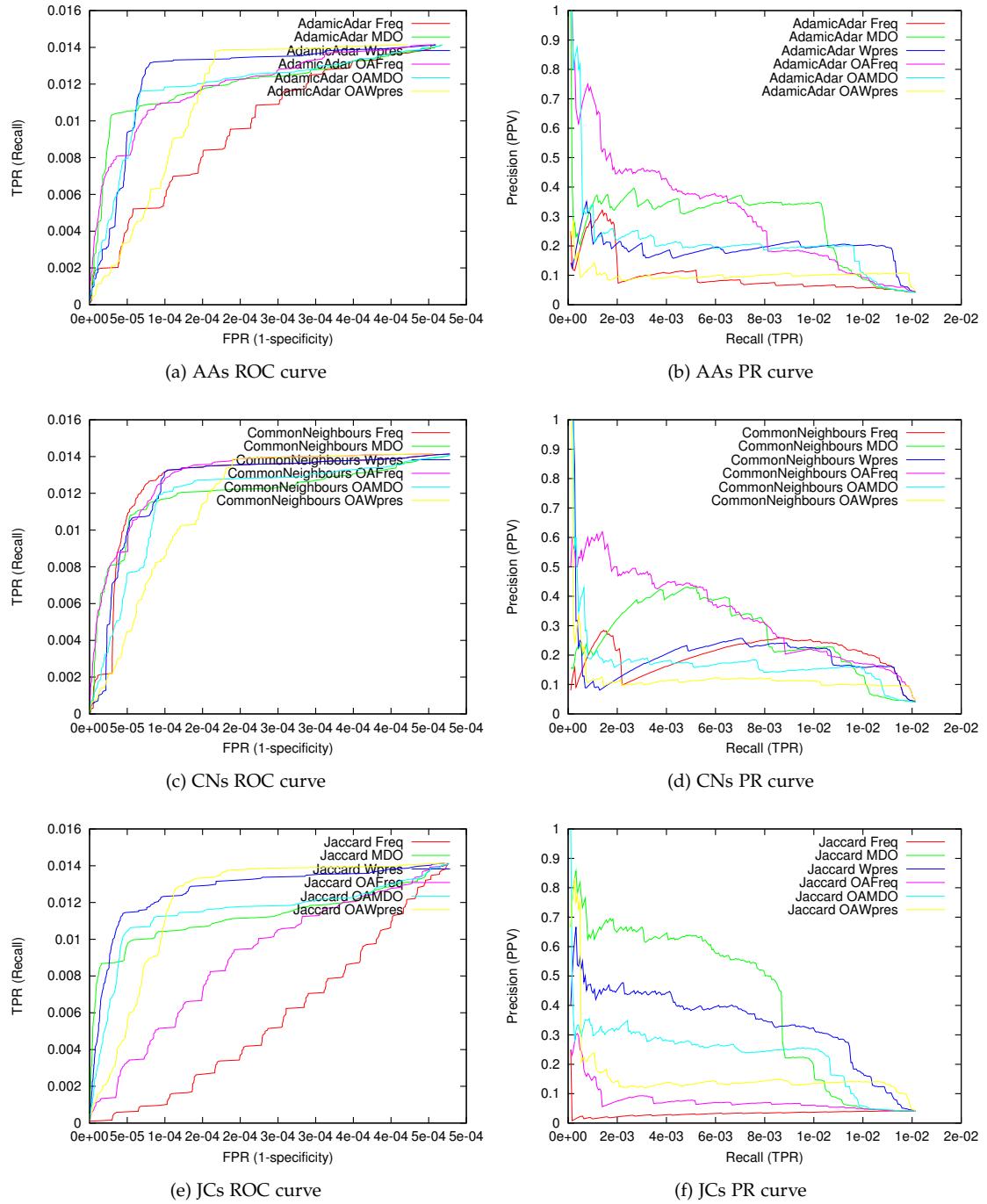


Figura 61: VDC Moltiplicatori Temporali II

PREDITTORE	p ₁	p ₂	p ₃	g ₁	g ₂	g ₃
AA Freq	32,3	19,3	12,1	+4,7	+6,3	+0,2
AA Wpres	47,4	60,6	12,1	+28,8	+47,6	+0,2
AA MDO	87,2	96,8	11,9	+68,6	+83,8	
AA OAFreq	106,5	52,2	12,2	+87,9	+39,2	+0,3
AA OAWpres	25,2	27,8	13,1	+6,6	+14,8	+1,2
AA OAMDO	54,5	55,2	11,9	+35,9	+42,2	
CN Freq	51,1	69,8	11,6	+23,0	+55,4	
CN Wpres	64,5	62,9	11,6	+36,4	+48,5	
CN MDO	119,5	61,6	11,6	+91,4	+47,2	
CN OAFreq	123,1	59,9	12,2	+95,0	+45,5	+0,6
CN OAWpres	30,6	31,8	12,2	+2,5	+17,4	+0,6
CN OAMDO	48,7	42,1	11,6	+20,6	+27,7	
JC Freq	8,2	10,3	11,7	+1,0	+0,9	
JC Wpres	111,9	93,0	11,8	+104,7	+83,6	+0,2
JC MDO	178,9	63,0	11,6	+171,7	+53,6	
JC OAFreq	21,1	18,8	11,7	+23,9	+9,4	+0,1
JC OAWpres	35,7	36,5	11,7	+28,5	+27,1	+0,1
JC OAMDO	77,0	71,9	11,6	+69,8	+62,5	
PA Freq	10,3	9,0	6,1	+1,9	+1,8	+0,1
PA Wpres	10,6	8,9	6,1	+2,2	+1,7	+0,1
PA MDO	11,0	7,6	6,0	+2,6	+0,4	
PA OAFreq	10,7	9,0	6,1	+2,3	+1,8	+0,1
PA OAWpres	9,1	8,2	6,1	+0,7		+0,1
PA OAMDO	10,3	9,0	6,1	+1,9	+1,8	+0,1

Table 36: VDC - Performance Moltiplicatori Temporali

5.6.2 Predittori Ad-Hoc

Esclusi i modelli LEdd e AM tutte le curve descritte dai modelli Ad-Hoc negli spazi ROC e PR dominano in modo definitivo quelle dei predittori analizzati sino ad ora.

In particolare le migliori performance sono raggiunte di due predittori facenti uso delle informazioni di Weighted Dimension Relevance.

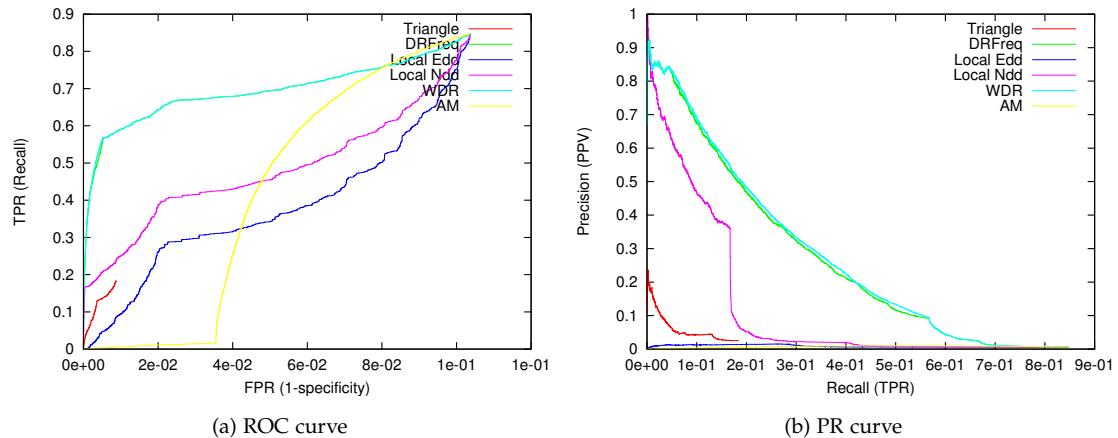


Figura 62: VDC Modelli AdHoc

PREDITTORE	p ₁	p ₂	p ₃
LEdd	3,8	1,8	1,8
LNdd	22,2	2,8	1,8
WDRw	145,1	50,3	1,8
WDRf	141,9	48,4	1,8
AM	1,6	2,6	1,8
Triangle	15,3	12,0	7,2

Table 37: VDC - Performance Modelli AdHoc

5.6.3 Analisi Riassuntiva

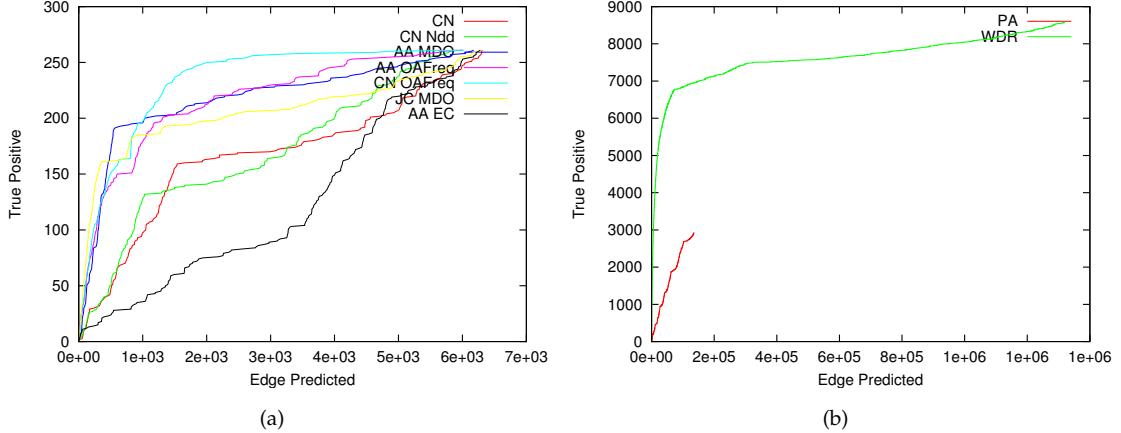


Figura 63: VDC Analisi riassuntiva

Data la particolare struttura della rete, costituita da un numero ridotto di nodi (dello stesso ordine di grandezza delle dimensioni), è interessante valutare i risultati ottenuti complessivamente per stabilire se - come auspicabile - le informazioni di tipo temporale possono essere efficacemente utilizzate a supporto del task di Link Prediction.

Come mostrato in Figura 63b, e già riportato precedentemente, il miglior modello predittivo per la rete risulta essere basato su Weighted Dimension Relevance (con informazioni di Weighted Presence): si è deciso di riportare solo questo modello predittivo AdHoc anche se altri riescono a performare meglio dei modelli base e derivati perché già analizzati precedentemente.

Passando ad analizzare i predittori aventi un insieme di risultati più contenuto (Figura 63a) si può notare che gli incrementi rispetto al migliore modello di base (Common Neighbours, anch'esso riportato in figura) sono molto marcati: in special modo le informazioni temporali di Max Double Occurrence e Frequency (come visto anche per i modelli AdHoc) riescono a far registrare i maggiori incrementi di performance complessivi per questo insieme di predittori.

Parte IV
CONCLUSIONI

6

CONCLUSIONI

In questo ultimo capitolo, alla luce della definizione del problema affrontato proposta nella Parte II, si riassumono i risultati ottenuti dall’analisi sperimentale affrontata nella Parte III del presente lavoro di tesi.

Dopo aver riassunto quanto è stato presentato nel corso della trattazione si dà una valutazione complessiva dei risultati: si propone, altresì, un’analisi delle possibili strade da intraprendere per sviluppare ulteriormente l’argomento affrontato alla luce di quanto emerso dai dati sperimentali.

6.1 VALUTAZIONE DEI RISULTATI OTTENUTI

In questo lavoro di tesi è stato introdotto il problema di Link Prediction multidimensionale (def. 9 a pagina 36): la definizione proposta arricchisce il task, analizzato sino ad ora in letteratura, tramite l’introduzione delle informazioni topologiche di dimensionalità e temporalità associate agli archi costituenti la rete da analizzare.

L’introduzione di tali informazioni causa la necessità di uno studio più accurato della rete analizzata al fine di consentire il successo di analisi predittive: in particolare la multidimensionalità impone l’introduzione di modifiche agli approcci già esistenti al problema di Link Prediction in modo da renderli in grado di garantire la completezza dei risultati ottenuti. Predire la specifica dimensione in cui un nuovo arco verrà a formarsi tra due nodi è il fine di un percorso di analisi, della topologia della rete, che nasconde al suo interno un accurato studio delle correlazioni presenti tra le dimensioni su cui la rete stessa si sviluppa.

Come abbiamo visto, catturare informazioni di correlazione, o anticorrelazione, tra le dimensioni non è un’operazione consentita dalle misure sino ad ora proposte in letteratura: è stato quindi necessario introdurre nuove funzioni di misura, studiate appositamente per l’ambito multidimensionale, che consentano di sfruttare le informazioni necessarie all’analisi di questa nuova particolare tipologia di reti.

Analogamente alle informazioni inerenti alla multidimensionalità sono state introdotte durante la trattazione, nel modello da analizzare, informazioni di carattere temporale che

consentissero di avere una più chiara visione dell'andamento evolutivo della rete: anche in questo caso il problema affrontato ha, necessariamente, dovuto subire una nuova lettura e definizione. Poder fare riferimento alla storia evolutiva di una rete consente di analizzare i pattern ricorrenti nella formazione degli archi e di sfruttarli per successive analisi predittive.

Dopo aver fissato il problema in modo univoco, e analizzato i modelli noti in letteratura per il problema di partenza, sono state definite alcune classi di soluzioni applicabili, alla nuova definizione proposta, così raggruppabili:

- Approcci derivanti la loro origine ad algoritmi già conosciuti per il problema monodimensionale
 - Modelli Base
 - Multidimensionali Locali
 - Multidimensionali Globali
 - Temporali
- Approcci basati esclusivamente sulle nuove misure multidimensionali e sull'analisi evolutiva.

Per valutare la bontà dei modelli predittivi proposti sono stati effettuati test estensivi su sei reti multidimensionali costruite appositamente per questo lavoro di tesi. Tali reti, ciascuna avendo caratteristiche topologiche diverse dalle altre, sono state modellate utilizzando dei dataset, costruiti a partire da diverse tipologie di servizi online, contenenti le informazioni di multidimensionalità e temporalità necessarie per una analisi completa tramite i mezzi proposti.

Al termine della fase di test sono state quindi comparate le classi di predittori analizzati in base alle performance riscontrate su ogni singola rete.

Da tale analisi sono emersi alcuni fattori comuni che possono fornire un'immagine di massima relativa all'efficacia degli approcci adottati. Innanzitutto è bene sottolineare che non è stato trovato un modello predittivo che possa essere definito "universale" - ovvero un modello che performi costantemente meglio di tutti quelli analizzati indipendentemente dal dataset a cui questi siano applicati - ma, al contrario, che i risultati ottenuti mettono in evidenza come un ristretto numero di predittori, appartenenti alla categoria Ad Hoc, riescano mediamente ad ottenere buone performance se paragonati a quelli derivati da approcci monodimensionali.

Molt.	P ₁	P ₂	P ₃	Molt.	P ₁	P ₂	P ₃
EC	29%	29%	12%	Freq	79%	70%	29%
Edd	62%	41%	41%	Wpres	66%	62%	29%
EP	16%	20%	25%	MDO	70%	83%	4%
NC	37%	25%	12%	OAFreq	83%	75%	41%
Ndd	54%	41%	33%	OAWpres	75%	58%	45%
NP	41%	29%	12%	OAMDO	75%	83%	8%

(a) Moltiplicatori Multidimensionali Globali
(b) Moltiplicatori Temporali

Tabella 38: Statistiche Globali

In particolare i modelli basati su Weighted Dimension Relevance (misura arricchita con informazioni temporali) e sul numero di triangoli adiacenti riescono spesso ad essere riconosciuti come i migliori approcci predittivi localmente al singolo dataset analizzato.

Una valutazione diametralmente opposta è attribuibile ai modelli derivati, da quelli che sono stati individuati come modelli base, per mezzo dell'introduzione dell'informazione multidimensionale in modo locale ai nodi. Questo particolare insieme di predittori non riesce mai ad ottenere risultati che riescano a migliorare le performance del modello di cui rappresentano una modifica.

Ovviamente tale risultato è da considerarsi una semplice indicazione parziale poiché i modelli analizzati, in questo lavoro di tesi, non rappresentano che una minima parte di quelli ideabili utilizzando questo principio costruttivo.

Analizzare i predittori appartenenti alle classi derivate dai modelli base per mezzo di fattori moltiplicativi è un'operazione che può essere effettuata in modo più dettagliato. Per poter dare un'analisi complessiva dell'impatto che ciascun moltiplicatore (sia questo recante informazioni di tipo multidimensionale globali alla rete, o informazioni inerenti alla temporalità di comparsa degli archi) introduce sulle performance dei modelli base a cui è applicato, si ritiene necessario riportare una statistica tabellare che evidenzi quale sia la percentuale dei casi, su tutte le applicazioni, in cui questo abbia contribuito a migliorare le performance predittive.

In Tabella 38 si riportano i valori, raggruppati rispetto alle soglie di $\frac{1}{3}$, $\frac{2}{3}$ e $\frac{3}{3}$ dei risultati calcolati sui vari dataset, che evidenziano quale sia complessivamente la percentuale dei casi in cui ciascun fattore moltiplicativo si è dimostrato utile a innalzare le performance dei modelli

base.

Come si può notare, percentualmente, le informazioni temporali riescono a innalzare le performance in più casi rispetto a quelle multidimensionali di carattere globale: in particolare Max Double Occurrency (in entrambe le sue varianti) e Frequency riescono ad essere utili all'innalzamento della precisione dei modelli predittivi di base nell'85% dei casi analizzati (relativamente a determinate soglie). Per i moltiplicatori multidimensionali, invece, sono da segnalare Edge/Node Dimension Degree che registrano un impatto positivo sul 40-60% dei casi analizzati.

I maggiori contributi portati da questo lavoro di tesi sono stati:

- l'introduzione e la definizione del problema del problema di Link Prediction Multidimensionale;
- la definizione formale, ed implementazione, di due diverse classi di predittori che sfruttano le più ricche informazioni topologiche offerte dal modello proposto;
- la costruzione di sei reti multidimensionali a partire da altrettanti dataset reperiti in rete;
- una vasta sperimentazione atta a dimostrare la bontà degli approcci proposti e la necessità della formulazione multidimensionale per il problema affrontato.

I principali risultati ottenuti dalla sperimentazione condotta sono:

- identificazione dell'inesistenza di un approccio predittivo universale, tra quelli analizzati, che offre performance migliori indipendentemente dalla tipologia del dataset analizzato;
- riconoscimento della scarsa utilità dell'introduzione, in modo locale ai singoli nodi, delle informazioni multidimensionali durante la fase di predizione (nel caso dei modelli derivati da predittori della classe base);
- osservazione di come sia possibile incrementare le performance predittive tramite l'utilizzo di informazioni multidimensionali ed evolutive di livello globale alla rete;
- definizione di modelli predittivi, non derivati da approcci studiati per il problema monodimensionale ma basati esclusivamente su misure multidimensionali e temporali, capaci, in molti casi di garantire notevoli incrementi di performance se paragonati alle altre classi predittive analizzate.

6.2 ULTERIORI SVILUPPI

Il lavoro effettuato non può ritenersi autoconclusivo: il problema di Link Prediction multidimensionale è attualmente oggetto di indagine e gli approcci presentati vogliono porsi come una guida iniziale all'analisi di tale ambito di ricerca.

Dai risultati ottenuti emergono due possibili strade percorribili per ottenere modelli predittivi aventi alte performance su reti multidimensionali con informazioni temporali:

1. Arricchire i modelli già noti in letteratura tramite moltiplicatori globali di carattere multidimensionale o temporale (o un combinazione dei due approcci);
2. Studiare nuovi modelli predittivi che sfruttino esclusivamente le più ricche informazioni topologiche offerte dalla rete esulando da modelli già noti (come fatto nella classe dei predittori Ad Hoc).

Altre tecniche, non coperte dall'analisi riportata in questa tesi, possono essere studiate per un'estensione allo specifico ambito trattato: ad esempio approcci supervisionati, come quello proposto da GERM [2], possono essere analizzati per un adattamento al problema di Link Prediction.

Part V

APPENDICI

A

APPENDICE

In questo capitolo si propone una panoramica sugli strumenti implementati durante il periodo di tesi: sono compresi in tale insieme i parser dei dataset utilizzati per costruire le reti, un implementazione di grafo multidimensionale ed evolutivo, una (parziale) implementazione della libreria presentata in [3], i predittori descritti nel capitolo 3, gli strumenti di analisi dei risultati proposti nel capitolo 4 ed un plugin per l’analisi visuale del task affrontato per Cytoscape.

Tutto il codice prodotto è rilasciato con licenza GPLv3.

A.1 SPECIFICHE IMPLEMENTATIVE

Il lavoro di tesi, data la sua natura sperimentale, ha portato alla produzione degli strumenti necessari alla rappresentazione delle reti e alla loro analisi.

Ottenuti i dataset delle reti da analizzare, a seguito di una estensiva ricerca online, si è proceduto ad una fase di trasformazione e consolidamento dei dati. Fissata una sintassi standard per la descrizione della rete è stato quindi necessario interpretare, ridurre, e modellare i dati ottenuti (eterogeneamente rappresentati: database, querylog, file strutturati, file testuali, XML) tramite strumenti scritti appositamente per ogni specifico dataset.

Una volta compilate le descrizioni delle reti (costituite da tre file per ciascuna rete: descrizione delle dimensioni, dei vertici e degli archi) si è passati alla costruzione del multigrafo.

A seguito di numerose ricerche, si è optato per la scrittura ex-novo di una libreria in Java per la gestione di multigrafi con informazioni evolutive. Tale decisione è giustificata da l’analisi, effettuata su alcune delle principali librerie per l’analisi di grafi esistenti per tale linguaggio, che ha evidenziato l’insufficienza degli strumenti, da queste forniti, per la specifica tipologia di studio previsto dal lavoro di tesi.

Le implementazioni analizzate presentavano, in rari casi, la possibilità di gestire multigrafi arricchiti da informazioni temporali: anche quando tale tipo di modellazione fosse reso possibile, si è riscontrato, a seguito di alcuni test, un ingombro eccessivo in memoria per grafi di

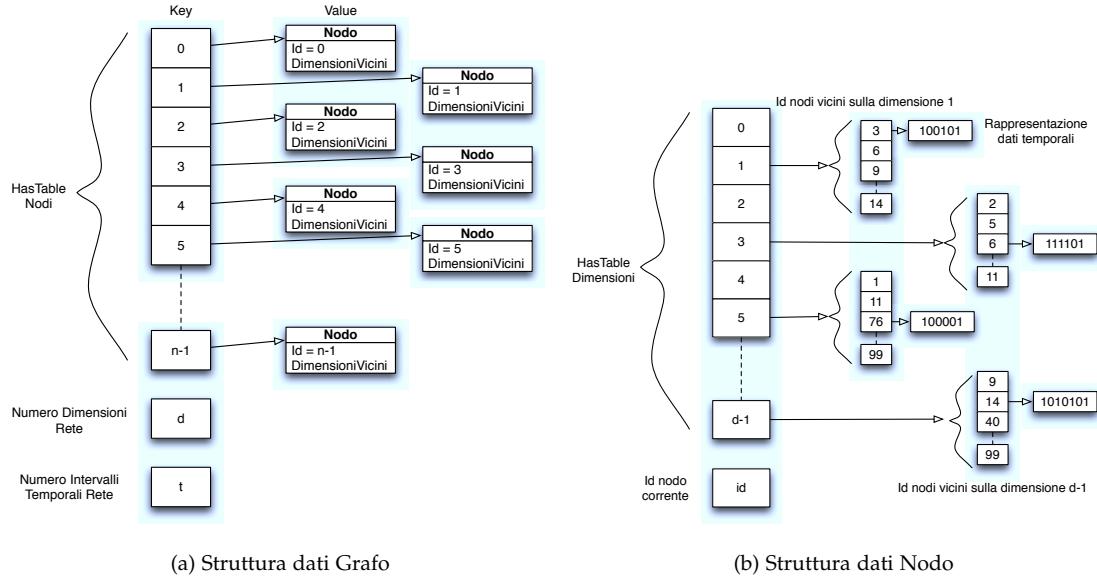


Figura 64: Struttura Grafo

dimensioni comparabili a quelli da analizzare. Si è deciso quindi di tentare un nuovo approccio alla pianificazione della struttura dati descrivente il multigrafo evolutivo che ha portato alla modellazione illustrata nel seguito.

Due classi gestiscono la struttura della rete in modo gerarchico: la classe Grafo e la classe Nodo.

- La classe Grafo (Figura 64a) mantiene al suo interno le seguenti informazioni:
 - numero di dimensioni della rete
 - numero degli istanti temporali della rete di training
 - un hashtable indirizzata tramite l'id numerico del nodo contenente il riferimento al relativo oggetto della classe Nodo
- La classe Nodo (Figura 64b) gestisce le seguenti informazioni
 - tramite un hashtable (indirizzata dagli id delle dimensioni della rete) si ottiene una seconda struttura dati;
 - tale struttura dati (anche essa una hashtable) è indirizzata dagli id dei nodo con cui esistono archi nella specifica dimensione e fornisce, come valore, una rappresentazione binaria degli istanti temporali in cui si sviluppa l'interazione stessa.

La struttura, che può apparire complessa a prima vista, consente di effettuare molte delle operazioni più frequenti mantenendo la complessità temporale superiormente limitata da:

- $O(1)$: per il calcolo di $\text{Neighbours}_{\text{Set}}$ e $\text{Degree}_{\text{Set}}$ oltre che nei casi di verifica di esistenza di un arco in una specifica dimensione;
- $O(n)$: per tutte le operazioni base nella loro versione Xor (dove n è il numero massimo tra le chiavi appartenenti a due hashtable da confrontare).

Tramite questi accorgimenti si riescono a garantire le effettive complessità degli algoritmi predittivi di base:

- $O(\log |E|)$: per Common Neighbours, Jaccard e derivati multidimensionali locali;
- $O(|V|^2)$: per Adamic Adar, Preferential Attachment e derivati multidimensionali locali.

Per i modelli che fanno uso di moltiplicatori multidimensionali globali, la complessità computazionale risulta essere quella associata al calcolo dei valori dei singoli moltiplicatori (solitamente dell'ordine di $O(|E|)$ per quelli basati sugli archi e $O(|V|)$ per quelli basati sui nodi) poiché sono riutilizzati gli insiemi dei risultati prodotti dai modelli base.

Per i modelli Ad Hoc la complessità è dell'ordine di $O(|V|^2)$: in alcuni casi, tramite attività di prefetching di alcune misure utilizzate nel calcolo, è possibile ridurre in modo sensibile la costante moltiplicativa.

Ottenuti i risultati del task di Link Prediction si è quindi fornito una libreria per l'analisi degli stessi: tale libreria consente (con complessità pari a $O(|E_{\text{new}}|)$) di valutare i valori necessari alla composizione dei grafici di Precision-Recall, di ROC e all'analisi numerica sull'andamento della Precision.

A.2 CYTOSCAPE PLUGIN

Cytoscape¹ è un programma libero (rilasciato sotto licenza GPL) utilizzato in bioinformatica per la visualizzazione delle reti di interazione molecolare. Molte caratteristiche aggiuntive sono state rese disponibili mediante una serie di plugin sviluppati dalla comunità rendendolo un ottimo strumento per l'analisi visuale di molteplici tipologie di reti.

¹ <http://www.cytoscape.org/>

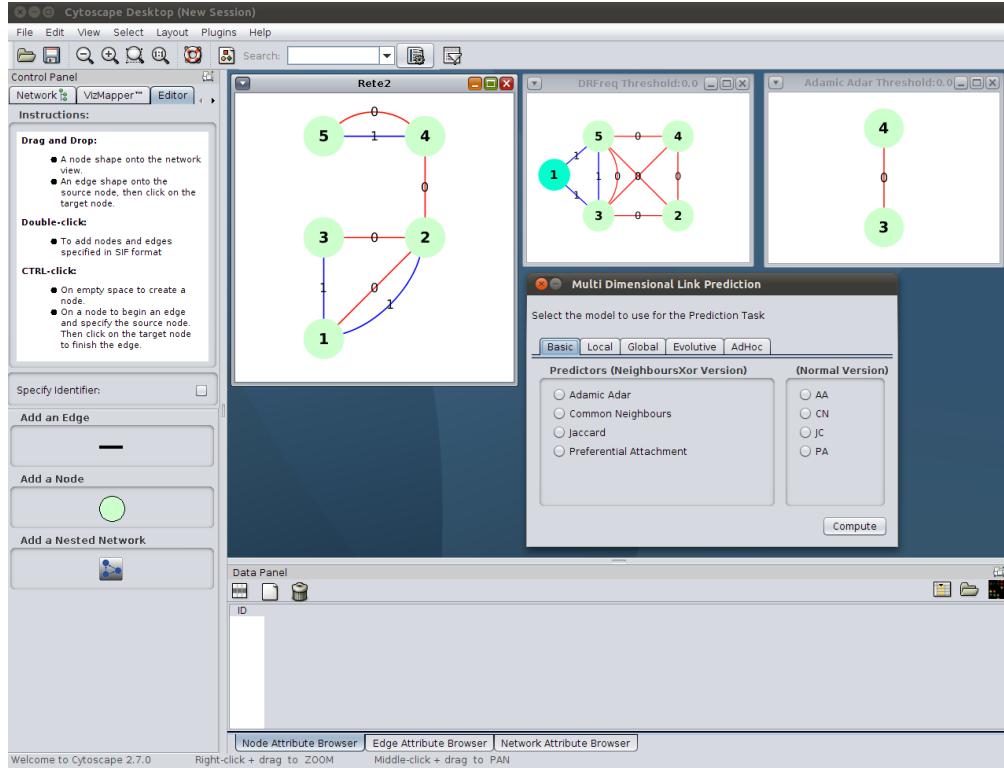


Figura 65: Plugin Cytoscape

Originariamente realizzato dall’Institute for System Biology, nel 2002 a Seattle, tale software attualmente è sviluppato da un consorzio di sviluppatori opensource. Il programma è stato reso pubblico per la prima volta nel luglio 2002.

Essendo Cytoscape sviluppato interamente in Java, si è deciso di trasportare le funzionalità implementate per l’analisi di Link Prediction su tale piattaforma in modo da consentire la visualizzazione di reti multidimensionali ed evolutive, ed i risultati dei vari processi predittivi proposti in modo grafico.

Il plugin (mostrato in Figura 65) consente di selezionare il modello predittivo desiderato tra i 58 analizzati in questa tesi e, impostando una soglia sulla bontà dello score dei risultati attesi, di visualizzare gli archi che, potenzialmente, entreranno a far parte della rete a seguito di un’evoluzione della stessa. La visualizzazione dei risultati è implementata in modo incrementale su finestre diverse in modo da poter valutare parallelamente i risultati di più predittori se applicati ad uno stesso grafo di partenza.

ELENCO DELLE FIGURE

1	Grafi	8
2	Multigrafo	11
3	Multigrafo con informazioni temporali	14
4	Gerarchia Modelli Link Prediction	28
5	Modelli derivati da approcci monodimensionali	44
6	Modelli AdHoc	52
7	Esempio Query Log	58
8	Esempio DBLP	60
9	Esempio IMDB	61
10	Esempio GTD	63
11	Esempio GCD	64
12	Esempio VDC	66
13	ROC curve	69
14	PR curve	70
15	PT curve	72
16	LOG Modelli Base	76
17	LOG Modelli Multidimensionali Locali	78
18	LOG Moltiplicatori Multidimensionali Globali I	80
19	LOG Moltiplicatori Multidimensionali Globali II	81
20	LOG Moltiplicatori Temporali I	83
21	LOG Moltiplicatori Temporali II	84
22	LOG Modelli AdHoc	86
23	LOG Analisi riassuntiva	87
24	DBLP Modelli Base	89
25	DBLP Modelli Multidimensionali Locali	91
26	DBLP Moltiplicatori Multidimensionali Globali I	93
27	DBLP Moltiplicatori Multidimensionali Globali II	94
28	DBLP Moltiplicatori Temporali I	96

29	DBLP Moltiplicatori Temporali II	97
30	DBLP Modelli AdHoc	99
31	DBLP Analisi riassuntiva	100
32	IMDB Modelli Base	101
33	IMDB Modelli Multidimensionali Locali	103
34	IMDB Moltiplicatori Multidimensionali Globali I	105
35	IMDB Moltiplicatori Multidimensionali Globali II	106
36	IMDB Moltiplicatori Temporali I	108
37	IMDB Moltiplicatori Temporali II	109
38	IMDB Modelli AdHoc	111
39	IMDB Analisi riassuntiva	112
40	GTD Modelli Base	113
41	GTD Modelli Multidimensionali Locali	115
42	GTD Moltiplicatori Multidimensionali Globali I	117
43	GTD Moltiplicatori Multidimensionali Globali II	118
44	GTD Moltiplicatori Temporali I	120
45	GTD Moltiplicatori Temporali II	121
46	GTD Modelli Ad Hoc	123
47	GTD Analisi riassuntiva	124
48	GCD Modelli Base	125
49	GCD Modelli Multidimensionali Locali	127
50	GCD Moltiplicatori Multidimensionali Globali I	129
51	GCD Moltiplicatori Multidimensionali Globali II	130
52	GCD Moltiplicatori Temporali I	132
53	GCD Moltiplicatori Temporali II	133
54	GCD Modelli AdHoc	135
55	GCD Analisi riassuntiva	136
56	VDC Modelli Base	137
57	VDC Modelli Multidimensionali Locali	139
58	VDC Moltiplicatori Multidimensionali Globali I	141
59	VDC Moltiplicatori Multidimensionali Globali II	142
60	VDC Moltiplicatori Temporali I	144
61	VDC Moltiplicatori Temporali II	145

62	VDC Modelli AdHoc	147
63	VDC Analisi riassuntiva	148
64	Struttura Grafo	160
65	Plugin Cytoscape	162

ELENCO DELLE TABELLE

1	Query Log Statistiche	58
2	DBLP Statistiche	60
3	IMDB Statistiche	61
4	GTD Statistiche	62
5	GCD Statistiche	64
6	VDC Statistiche	66
7	Matrice di confusione	67
8	Query Log - Performance modelli Base	77
9	Query Log - Performance Modelli Multidimensionali Locali	79
10	Query Log - Performance Moltiplicatori Multidimensionali Globali	82
11	Query Log - Performance Moltiplicatori Temporali	85
12	Query Log - Performance Modelli AdHoc	87
13	DBLP- Performance modelli Base	90
14	DBLP - Performance Modelli Multidimensionali Locali	92
15	DBLP - Performance Moltiplicatori Multidimensionali Globali	95
16	DBLP - Performance Moltiplicatori Temporali	98
17	DBLP - Performance Modelli AdHoc	99
18	IMDB - Performance modelli Base	102
19	IMDB - Performance Modelli Multidimensionali Locali	104
20	IMDB - Performance Moltiplicatori Multidimensionali Globali	107
21	IMDB - Performance Moltiplicatori Temporali	110
22	IMDB - Performance Modelli AdHoc	111
23	GTD - Performance modelli Base	114
24	GTD - Performance Modelli Multidimensionali Locali	116
25	GTD - Performance Moltiplicatori Multidimensionali Globali	119
26	GTD - Performance Moltiplicatori Temporali	122
27	GTD - Performance Modelli AdHoc	123
28	GCD - Performance modelli Base	126

29	GCD - Performance Modelli Multidimensionali Locali	128
30	GCD - Performance Moltiplicatori Multidimensionali Globali	131
31	GCD - Performance Moltiplicatori Temporali	134
32	GCD - Performance Modelli AdHoc	135
33	VDC - Performance modelli Base	138
34	VDC - Performance Modelli Multidimensionali Locali	140
35	VDC - Performance Moltiplicatori Multidimensionali Globali	143
36	VDC - Performance Moltiplicatori Temporali	146
37	VDC - Performance Modelli AdHoc	147
38	Statistiche Globali	153

ACRONIMI

AA	Adamic Adar
JM	Jaccard Measure
CN	Common Neighbours
PA	Preferential Attachment
DR	Dimension Relevance
DRXor	Dimension Relevance Xor
WDR	Weighted Dimension Relevance
Edd	Edge Dimension Degree
Ndd	Node Dimension Degree
NP	Node Parent
EP	Edge Parent
NC	Node Correlation
EC	Edge Correlation
MDO	Max Double Occurrency
ROC	Receiver Operating Characteristic
PR	Precision - Recall

BIBLIOGRAFIA

- [1] Lada A. Adamic and Eytan Adar. Friends and neighbours on the web. *Social Networks*, 25(3):211–230, July 2003. (Citato a pagina 29.)
- [2] Michele Berlingero, Francesco Bonchi, Bjorn Bringmann, and Aristides Gionis. Mining graph evolution rules. (Citato alle pagine 21, 33 e 155.)
- [3] Michele Berlingero, Michele Coscia, Fosca Giannotti, Anna Monreale, and Dino Pedreschi. Foundations of multidimensional network analysis. 2010. (Citato alle pagine 37 e 159.)
- [4] Mustafa Bilgic, Galileo Mark Namata, and Lise Getoor. Combining collective classification and link prediction. (Citato a pagina 33.)
- [5] Deepayan Chakrabarti and Christos Faloutsos. Graph mining: Laws, generators and algorithms. *ACM Computing Surveys*, 38, March 2006. (Citato alle pagine 16 e 30.)
- [6] Aaron Clauset, Cristopher Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, pages 98 – 101, 2008. (Citato a pagina 19.)
- [7] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. *Proceedings of the 23rd International Conference on Machine Learning*, 2006. (Citato a pagina 71.)
- [8] Lise Getoor and Christopher P. Diehl. Link mining: A survey. *SIGKDD Explorations*, 7(2). (Citato alle pagine 16 e 19.)
- [9] Jiawei Han and Micheline Kamber. *Data Mining Concepts and Techniques*. Morgan Kaufmann, 2006. (Citato a pagina 29.)
- [10] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. Link prediction using supervised learning. (Citato a pagina 33.)
- [11] Zan Huang. Link prediction based on graph topology: The predictive value of the generalized clustering coefficient. *Workshop on Link Analysis, the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (20–23), August 2006. (Citato a pagina 19.)

- [12] Zan Huang, Xin Li, and Hsinchun Chen. Link prediction approach to collaborative filtering. *JCDL*, 2005. (Citato a pagina 32.)
- [13] J. H. Jones and M. S. Handcock. An assessment of preferential attachment as a mechanism for human sexual network formation. *The Royal Society*, 2003. (Citato a pagina 30.)
- [14] Indika Kahanda and Jennifer Neville. Using transactional information to predict link strength in online social networks. 2009. (Citato a pagina 32.)
- [15] Cane Wing ki Leung, Ee-Peng Lim, David Lo, and Jianshu Weng. Mining interesting link formation rules in social networks. *CIKM'10*, October 2010. (Citato alle pagine 21 e 33.)
- [16] Gary King. 10 million international dyadic events. URL <http://hdl.handle.net/1902.1/FYXLAWZRIA>.
- [17] Valdis E. Krebs. Mapping networks of terrorist cells. 2002. (Citato a pagina 62.)
- [18] Jeremy Kubica, Andrew Moore, David Cohn, and Jeff Schneider. cgraph: A fast graph-based method for link analysis and querues. (Citato a pagina 32.)
- [19] Vincent Leroy, B. Barla Cambazoglu, and Francesco Bonchi. Cold start link prediction. *KDD'10*, pages 25–28, July 2010. (Citato a pagina 26.)
- [20] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1), March 2007. (Citato a pagina 18.)
- [21] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, May 2007. (Citato alle pagine 28 e 72.)
- [22] Saket Navlakha and Carl Kingsford. Network archaeology: Uncovering ancient networks from present-day interactions. September 2010. (Citato a pagina 30.)
- [23] M. E. J. Newman. Clustering and preferential attachment in growing networks. April 2001. (Citato alle pagine 29 e 30.)
- [24] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 2(45):167, 2003. (Citato alle pagine 9 e 12.)

- [25] Joshua O'Madadhain, Jon Hutchins, and Pedhraic Smyth. Prediction and ranking algorithms for event-based network data. *SIGKDD Explorations*. (Citato a pagina 32.)
- [26] Christopher R. Palmer, Philip B. Gibbons, and Christos Faloutsos. Anf: A fast and scalable tool for data mining in massive graphs. *SIGKDD*, 2002. (Citato a pagina 10.)
- [27] David M. Pennock, Gary W. Flake, Steve Lawrence, Eric J. Glover, and C. Lee Giles. Winners don't take all: Characterizing the competition for links on the web. 2002. (Citato a pagina 30.)
- [28] Alexandrin Popescul and Lyle H. Ungar. Statistical relational learning for link prediction. *IJCAI*, 2003. (Citato a pagina 32.)
- [29] A. Potgieter, Kurt April, R.J.E. Cooke, and I.O. Osunmakinde. Temporality in link prediction: Understanding social complexity. *Sprouts: Working Papers on Information Systems*, 2007. (Citato alle pagine 34 e 49.)
- [30] Travers and Milgram. An experimental study of the small world problem. *Sociometry*, 1969. (Citato a pagina 18.)
- [31] Chao Wang, Venu Satuluri, and Srinivasan Parthasarathy. Local probabilistic models for link prediction. (Citato a pagina 32.)
- [32] Kazuko Yamasaki, Kaushik Matia, Sergey V. Buldyrev, Dongfeng Fu, Fabio Pammolli, Massimo Riccaboni, and H. Eugene Stanley. Preferential attachment and growth dynamics in complex systems. *PHYSICAL REVIEW E* 74, 2006. (Citato a pagina 30.)
- [33] Xifeng Yan and Jiawei Han. gspan: Graph-based substructure pattern mining. *International Conference on Data Mining (ICDM)*, 2002. (Citato a pagina 21.)
- [34] Soon-Hyung Yook, Hawoong Jeong, and Albert-Lazlo Barabasi. Modeling the internet's large-scale topology. *PNAS*, 2002. (Citato a pagina 10.)
- [35] Kai Yu and Wei Chu. Stochastic relational models for discriminative link prediction. (Citato a pagina 32.)
- [36] Elena Zheleva, Lise Getoor, Jennifer Golbeck, and Ugur Kuter. Using friendship ties and family circles for link prediction. *The 2nd SNA-KDD Workshop '08*, 2008. (Citato a pagina 33.)

- [37] Jianhan Zhu, Jun Hong, and John G. Hughes. Using markov chains for link prediction in adaptive web sites. *Soft-Ware*, 2002. (Citato a pagina 32.)