# POLITECNICO DI TORINO

Corso di Laurea Magistrale

in INGEGNERIA INFORMATICA (COMPUTER ENGINEERING)

Tesi di Laurea Magistrale

Analisi del traffico di una rete wireless mediante algoritmi
di estrazione di Itemset generalizzati


**Relatori:**
prof. Tania Cerquitelli
prof. Paolo Garza

**Candidato:**
Andreino Garibaldi

Anno Accademico 2018/2019

# Sommario

L'attività di estrazione di *itemset* generalizzati è una tecnica di *data mining* ben consolidata che ha per obiettivo la scoperta della conoscenza nascosta nei dati analizzati a diversi livelli di astrazione, che è come dire aggregati secondo tassonomie costruite sui dati stessi. In tale modo è possibile recuperare a un più alto grado di astrazione quegli *itemset* che altrimenti, per la loro scarsa frequenza osservata, sarebbero eliminati nel processo di *data mining*. Tuttavia, i tradizionali *itemset* generalizzati non sempre e non tutti possiedono un vero valore aggiunto in quanto essi sovente esprimono semplicemente la medesima conoscenza già espressa dai loro discendenti estratti a più basso livello, costituendone pertanto una mera ridondanza. Tra i diversi metodi proposti per superare tale limite, uno è rappresentato dall'estrazione dei cosiddetti *Maximal Expressive Generalized Itemset* (Max-EGI), ovvero l'estrazione dei soli *itemset* ad alto livello che rappresentino dati non già ricompresi in alcuno dei loro discendenti frequenti a più basso livello. In questo lavoro sono stati analizzati e comparati i risultati dell'applicazione di questi metodi (tradizionale e Max-EGI) a dati reali sperimentali, costituiti dalla raccolta del traffico wireless della rete di ateneo e, inoltre, anche introdotte nuove misure di interesse per gli *itemset*. L'analisi ha mostrato la netta superiorità dell'innovativo *Max-EGI extraction algorithm* rispetto al tradizionale *GENIO Algorithm* per quanto riguarda la riduzione delle ridondanze, mentre non è apparso qualitativamente ragguardevole il suo apporto nella generazione di nuovi *itemset* espressivi. Per contro, il lavoro ha evidenziato le criticità insite nella definizione della tassonomia e dei criteri di discretizzazione di talune *feature* e i loro effetti sui risultati generati dagli algoritmi, proponendone gli opportuni emendamenti.

## Indice

# Capitolo 1

# Introduzione

Uno dei compiti centrali nell'attività di *data mining* è il trovare correlazioni tra dati, soprattutto quando queste correlazioni non sono né esplicitate né immediatamente percepibili. L'esempio più tipico è costituto dalla c.d. *market basket analysis*, dove gli oggetti (*item*) che vengono venduti insieme sono organizzati in un insieme di transazioni (*market basket*) e dove la frequente co-occorrenza dei medesimi gruppi di *item* (*itemset*), con rispetto all'insieme di tutte le transazioni, costituisce la base per l'identificazione di possibili correlazioni. Peraltro, e più in generale, il concetto di "venduto insieme" può essere ed è stato esteso ai più vari ambiti — per esempio: parole "usate insieme", alimenti "assunti insieme", sintomi "presenti insieme" — e pertanto l'estrazione di *itemset* frequenti [1] è una tecnica esplorativa largamente usata per costruire un insieme limitato dei dati maggiormente suscettibili di contenere interessanti correlazioni tra gli oggetti (qualsiasi cosa essi rappresentino) di un *set* di dati.

Un generico *itemset* frequente $X$ è pertanto un insieme di *item* la cui frequenza relativa osservata di co-occorrenza (detta comunemente supporto o, brevemente, *supp*) in un insieme di transazioni $\mathcal{D}$ è superiore o uguale a una soglia prefissata $\varepsilon$, ovvero

$$\text{supp}(X) = \frac{|\{t \in \mathcal{D} : X \subseteq t\}|}{|\mathcal{D}|} \geq \varepsilon, \qquad X = \{x_1, x_2, \ldots, x_k\} \tag{1.1}$$

La determinazione della soglia $\varepsilon$ è in un certo qual modo arbitraria giacché, da un lato, la si vorrebbe sufficientemente alta per limitare — anche per ragioni computazionali — quanto più è possibile l'insieme risultante ai più significativi tra gli *itemset* frequenti e, dall'altro lato, sufficientemente bassa per catturare anche la conoscenza eventualmente contenuta in *itemset* non necessariamente ad altissima frequenza. È chiaro che comunque essa venga prefissata, raramente il valore così scelto soddisferà entrambe le aspettative. Tra i rimedi proposti per ovviare a tale inconveniente vi è la tecnica di introdurre una generalizzazione [2] [3] degli *item* costituenti gli *itemset*, anche con diversi livelli gerarchici, al fine di recuperare, a un più alto livello di astrazione, la conoscenza insita negli *itemset* non soddisfacenti la (1.1) e considerati quindi infrequenti. Si considerino infatti, per esempio, i seguenti *itemset* infrequenti per una generica soglia $\varepsilon_0$: {whisky, shampoo}, {vodka, shampoo} e {brandy, shampoo}. Ove gli *item* «whisky», «vodka» e «brandy» fossero aggregati in un *item* generico «liquore» potrebbe ben accadere che il supporto dello *itemset* così generalizzato {liquore, shampoo} — che sarebbe ovviamente uguale alla somma di quelli dei suoi originanti — possa soddisfare, per $\varepsilon = \varepsilon_0$, la (1.1) e quindi la conoscenza insita negli *itemset* scartati sarebbe

recuperata al livello di astrazione immediatamente superiore.

Conseguentemente, l'estrazione di *itemset* generalizzati frequenti è divenuta anch'essa una tecnica ben consolidata che si concentra sulla scoperta delle conoscenze nascoste nei dati analizzati a

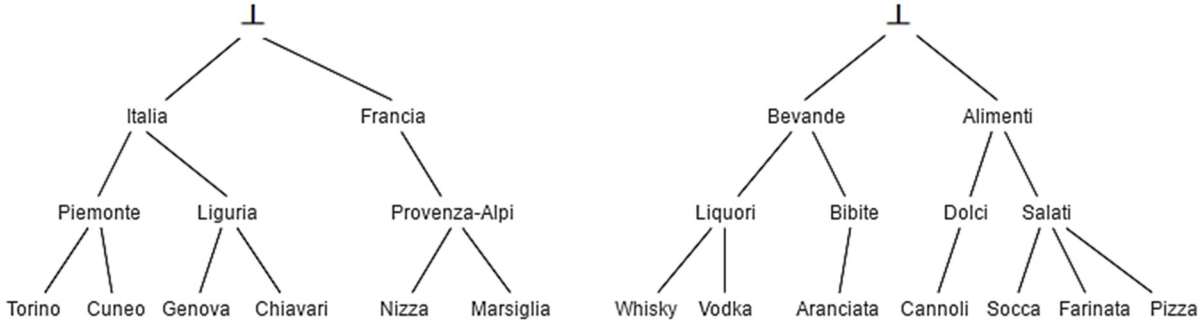



**Figura 1** — *Esempi di alberi di generalizzazione.*

diversi livelli di granularità sfruttando una tassonomia (ovvero un insieme di gerarchie, di tipo[1] $x \lhd y$, costruite sui dati) e superando così la perdita di informazione derivante dall'applicazione di soglie troppo severe nella tecnica tradizionale.

Formalmente, ampliando il concetto intuitivo di *market basket* e inquadrandolo nel contesto dei dati strutturati, un *set* di dati è costituito da un insieme di record, ciascuno dei quali rappresenta un insieme di *item*. Ciascun *item* è invero una coppia $\langle$*caratteristica, valore*$\rangle$ dove *caratteristica* rappresenta la descrizione di un tratto distintivo (*feature*) dei dati e *valore* l'informazione ad essa associata, tratta da un dominio $\Omega$ corrispondente. Con questa premessa risulta possibile procedere ad alcune definizioni.

**Definizione 1** (*Albero di generalizzazione*) Sia $t$ una caratteristica (*feature*) dei dati e $\Omega$ ne sia il corrispondente dominio. Un albero di generalizzazione $GT$ è una gerarchia di generalizzazione costruita sui valori in $\Omega$ e rappresentata da un albero radicato ed etichettato $\langle r, N, L, E \rangle$, dove

- l'insieme delle etichette $L$ è un soprainsieme di $\Omega$ $(\Omega \subseteq L)$ e contiene sia i valori nel dominio sia le loro generalizzazioni,
- i nodi foglia sono etichettati con i valori in $\Omega$,
- i nodi non foglia sono aggregazioni dei valori in $\Omega$ e sono etichettati con i valori in $L \backslash \Omega$,
- il nodo radice $r$ è etichettato con il simbolo speciale $\perp$,
- per ciascuna etichetta $l \in L$ esiste uno e un solo nodo etichettato con $l$.

In figura 1 sono rappresentati due esempi di alberi di generalizzazione, l'uno relativo ad una generalizzazione geografica, l'altro merceologica.

**Definizione 2** (*Tassonomia*) Sia $\mathcal{T}$ un insieme di caratteristiche (*feature*) di un *set* di dati $\mathcal{D}$. Una tassonomia $\Gamma = \{GT_1, GT_2, \ldots, GT_n\}$ è una foresta di alberi di generalizzazione dove $GT_i$ è l'albero di generalizzazione della caratteristica $t_i \in \mathcal{T}$.

In presenza di una tassonomia, gli *itemset* sono costituiti allora da insiemi di *item*, ciascuno associato a una distinta caratteristica del *set* di dati e il cui valore è uno dei nodi del corrispondente

[1] Il simbolo $\lhd$ viene qui e nel seguito impiegato per indicare genericamente una relazione di inclusione transitiva del tipo: «*è-un*», «*è-incluso-in*», «*pertiene-a*». Si consideri, per esempio, «whisky»$\lhd$«liquore», «vodka»$\lhd$«liquore» o, ancora, «liquore»$\lhd$«bevanda» e «aranciata»$\lhd$«bibita»$\lhd$«bevanda.

albero di generalizzazione, fatta eccezione pel nodo radice.

**Definizione 3** (*Itemset generalizzato*) Sia $\mathcal{D}$ un *set* di dati strutturato, $\Gamma$ una tassonomia costruita sulle caratteristiche di $\mathcal{D}$ e $X = \{x_1, x_2, \dots, x_k\}$ un *itemset*. $X$ è un *itemset* generalizzato se almeno uno dei suoi *item* $x_i$ ha un valore associato a uno dei nodi non foglia della tassonomia $\Gamma$.

**Definizione 4** (*Relazione di discendenza*) Sia $\mathcal{D}$ un *set* di dati strutturato, $\Gamma$ una tassonomia costruita sulle caratteristiche di $\mathcal{D}$ e $X = \{x_1, x_2, \dots, x_k\}$, $Y = \{y_1, y_2, \dots, y_k\}$ due *itemset* (generalizzati). $X$ è un discendente di $Y$ in $\Gamma$ se $\forall\, x_i \in X,\ (x_i = y_i) \veebar (x_i \lhd y_i) \in \Gamma$.

Se l'estensione della nozione di *itemset* a quella di *itemset* generalizzato, e la sua conseguente applicazione nell'estrazione dei più frequenti (generalizzati e non) apporta i benefici testé considerati, nondimeno non tutti gli *itemset* generalizzati apportano un reale valore aggiunto in termini di conoscenza né tutti forniscono una interpretazione agevole in termini di espressività. Accade infatti che vengano generalizzati anche gli *itemset* già caratterizzati *per se* d'una frequenza relativa osservata superiore o uguale alla soglia prefissata, determinando:

a) la generazione di *itemset* che replicano esattamente la medesima conoscenza di *itemset* frequenti a un più basso livello della gerarchia tassonomica e che risultano perciò ridondanti;
b) la generazione di *itemset* generalizzati che aggregano contemporaneamente *itemset* frequenti e infrequenti a più basso livello nella tassonomia, per i quali la valutazione dell'interesse risulta ardua giacché da un lato è vero che essi incorporano una conoscenza non contenuta in alcuno dei loro discendenti frequenti, ma, dall'altro lato, questa conoscenza non è esplicitamente disgiunta e separabile da quella già espressa da questi ultimi.

Sono state pertanto proposte ulteriori estensioni alla nozione di *itemset* generalizzato, tra le quali la nozione di *itemset* espressivo generalizzato [3] (*expressive generalized itemset* o *EGI*) e la sua restrizione denominata *itemset* espressivo generalizzato massimale (*maximal expressive generalized itemset* o *Max-EGI*). Gli *EGI* sono rappresentati nella forma $X \wr S$, dove $X$ è un *itemset* generalizzato e $S$ è un insieme di *itemset*[2] (generalizzati) che contiene solo discendenti frequenti di $X$. Il simbolo $\wr$ rappresenta la trasposizione, nell'ambito degli *itemset*, dell'operatore di complemento insiemistico e ne riproduce, *mutatis mutandis*, la nozione. Per esempio, facendo riferimento alla tassonomia di figura 1, il seguente *EGI* {(città, Liguria)}$\wr${{(città, Genova)}} rappresenterebbe la generalizzazione di tutte le città della Liguria, esclusa però Genova. Si noti inoltre che un *itemset* generalizzato $X$ non è che un caso particolare di *EGI* per il quale $S = \emptyset$.

**Definizione 5** (*Itemset espressivo generalizzato* o *EGI*) Sia $\mathcal{D}$ un *set* di dati strutturato, $\Gamma$ una tassonomia costruita sulle caratteristiche di $\mathcal{D}$, $X$ un *itemset* generalizzato e $S$ un insieme di *itemset*[3] (generalizzati). $X \wr S$ è un *EGI* se e solo se

a) $S = \emptyset$, oppure
b) $\forall\, Y \in S$, $Y$ è un discendente di $X$ in $\mathcal{D}$ rispetto a $\Gamma$.

L'introduzione della generalizzazione e dell'espressività alla nozione standard di *itemset* pone al-

[2] Invero, sarebbe possibile e più generale definire $S$ come un insieme a sua volta di EGI, ma, per semplicità, nel seguito si farà riferimento solo al caso non ricorsivo.

[3] Vedi nota 2

cune questioni in merito alla definizione di supporto, che evidentemente deve essere convenientemente, e rigorosamente, riformulata, rispetto alla (1.1) per ricomprendere anche questi nuovi oggetti. È allora innanzitutto necessario definire la nozione di copertura di un *itemset* (generalizzato (espressivo)) rispetto ai *record* di un *dataset* e, successivamente derivare da essa le definizioni di frequenza assoluta e di supporto.

**Definizione 6** (*Copertura*) Sia $\mathcal{D}$ un *set* di dati strutturato, $r \in \mathcal{D}$ un generico *record*, $\Gamma$ una tassonomia costruita sulle caratteristiche di $\mathcal{D}$ e $X \wr S$ un *EGI*. Allora $X \wr S \xrightarrow{\text{copre}} r$ se e solo se valgono alternativamente e ricorsivamente le seguenti:

- $\Gamma = \emptyset \wedge S = \emptyset \wedge X \subseteq r$, oppure
- $\Gamma \neq \emptyset \wedge S = \emptyset \wedge \forall\, x \in X \backslash r\ \exists\, y \in r \backslash X : (y \lhd x) \in \Gamma$, oppure
- $\Gamma \neq \emptyset \wedge S \neq \emptyset \wedge X \xrightarrow{\text{copre}} r \wedge \nexists\, Y \in S : Y \xrightarrow{\text{copre}} r$.

**Definizione 7** (*Frequenza assoluta e supporto*) Sia $\mathcal{D}$ un *set* di dati strutturato, $\Gamma$ una tassonomia costruita sulle caratteristiche di $\mathcal{D}$ e $X \wr S$ un *EGI*. La frequenza assoluta $\#(X \wr S)$ e il supporto $\mathsf{supp}(X \wr S)$ sono definiti rispettivamente come:

$$\#(X \wr S) = \left|\{r \in \mathcal{D} :\ X \wr S \xrightarrow{\text{copre}} r\}\right| \tag{1.2}$$

$$\mathsf{supp}(X \wr S) = \frac{\#(X \wr S)}{|\mathcal{D}|} \tag{1.3}$$

Gli *itemset* espressivi generalizzati massimali (*Max-EGI*) sono un sottoinsieme degli *EGI* per i quali valgono almeno le seguenti restrizioni: (i) devono essere frequenti rispetto a una soglia prefissata $\varepsilon$ nonché (ii) il loro insieme complementare $S$ deve contenere *tutti* i relativi discendenti frequenti (rispetto a $\varepsilon$). Tali restrizioni assicurano che il sottoinsieme così definito offra il più alto livello di espressività e di facilità nell'interpretazione dei risultati del processo di *data mining*, riducendo nel contempo la cardinalità dei risultati senza sacrificio della completezza.

**Definizione 8** (*Itemset espressivo generalizzato massimale* (*Max-EGI*)) Sia $\mathcal{D}$ un *set* di dati strutturato, $\Gamma$ una tassonomia costruita sulle caratteristiche di $\mathcal{D}$, $X \wr S$ un *EGI* ed $\varepsilon$ una soglia prefissata. $X \wr S$ è un *Max-EGI* se e solo se

a) $\mathsf{supp}(X \wr S) \geq \varepsilon$, e inoltre,
b) è vera almeno una delle seguenti:
   - $X \wr S$ non ha discendenti, oppure
   - $\nexists\, Y$ tale che, a un tempo, $\mathsf{supp}(Y) \geq \varepsilon$ e $Y$ sia un discendente di $X$ rispetto a $\Gamma$ e $Y \notin S$.

Con le estensioni della nozione di *itemset* frequente che sono state descritte sopra, la soglia di supporto minimo di cui alla (1.1) cessa di essere l'unico parametro discriminante e, in qualche modo, cede parte di tale ruolo alla tassonomia, che pertanto dovrà essere il più possibile adeguata e aderente al contesto applicativo.

Tra tutti i contesti applicativi ai quali possono essere applicate le metodologie sopra descritte, quello dell'analisi del traffico di rete è certamente uno di quelli che appaiono maggiormente significativi [4]. Da un lato il grande sviluppo in termini di velocità, diffusione e dimensioni delle reti di calcolatori, che ha determinato e determina un enorme e monotonicamente crescente flusso di

dati veicolato da esse, è accompagnato da una ugualmente crescente necessità di strumenti atti ad affrontare le inerenti problematiche in termini di prestazioni e di sicurezza. Dall'altro, la varietà e quantità dei dati veicolati ha assunto dimensioni tali da renderne oltremodo ardua l'analisi da parte degli esperti del dominio, e conveniente l'applicazione sui dati grezzi di tecniche tipiche della *KDD*. A tutto questo va aggiunta la sempre più crescente rilevanza del traffico generato dalle reti *wireless*, in ragione della sempre maggiore diffusione da un lato dei dispositivi mobili e dei laptop, e dall'altro degli *hot-spot* pubblici e delle reti *wireless* aziendali e scolastiche che offrono connettività rispettivamente ai propri dipendenti o studenti.

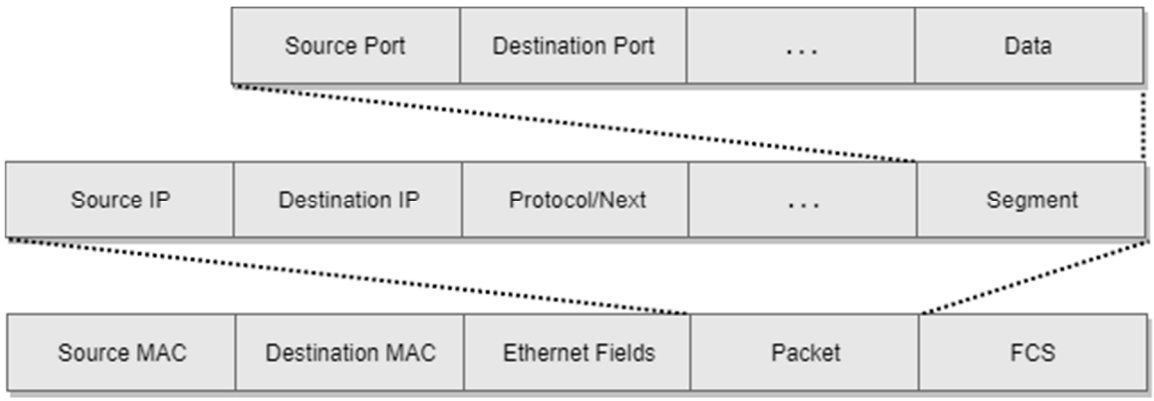



**Figura 2** — *Rappresentazione sintetica dei primi livelli dello stack di rete.*

Il traffico di rete ha peculiarità che lo rendono un buon candidato all'applicazione delle metodologie sopra descritte. Nei primi livelli dello *stack*, come si vede nell'esempio in figura 2, la strutturazione dei frame, dei datagrammi e dei segmenti — decodificabile con semplici macchine a stati — permette di estrarre naturalmente un primo insieme di caratteristiche (*feature*) sulle quali condurre analisi di co-occorrenza, quali per esempio gli indirizzi sorgente e destinazione di livello 2 e 3, le porte sorgente e destinazione, il protocollo di livello 4. Riconducendo poi le sequenze di segmenti e datagrammi utente a flussi e messaggi si ottiene da un lato la riduzione del volume dei dati e dall'altro l'estrazione di ulteriori caratteristiche quali la quantità di dati scambiata nei due versi tra gli *endpoint* e la durata delle connessioni e delle sessioni. Analogamente, l'ispezione dei *payload* di più alto livello, con macchine a stati finiti di livello applicativo o con ricerca di *signature*, può identificare e classificare gran parte del traffico applicativo più comune e aggiungere ulteriori *feature*. Inoltre, le *feature* così identificabili sono suscettibili d'esser facilmente organizzate in tassonomie. Per esempio, gli indirizzi di livello 2 possono essere generalizzati in base ai primi tre ottetti (OUI) e successivamente in globali, locali e *multicast*. Gli indirizzi IP possono essere generalizzati per esempio variando a passi opportuni la lunghezza in bit della maschera di sottorete e ancora aggregati in pubblici, privati, *multicast*, *anycast*. Analoghe gerarchie possono essere facilmente costruite per altre caratteristiche quali le porte, i protocolli, i servizi, i dati di flusso.

Per queste ragioni appare particolarmente interessante analizzare i risultati dell'applicazione delle metodologie di estrazione di *itemset* frequenti sopra descritte al traffico di rete reale. Questa tesi persegue come obiettivi l'analisi e la comparazione — sia con metriche oggettive sia con strumenti di valutazione *domain-driven* — dei risultati di un esperimento di estrazione di *itemset* generalizzati frequenti e di *Max-EGI* da un insieme di *set* di dati ottenuti dalla cattura del traffico *wireless* nelle aree studenti dell'Ateneo, riallacciandosi in questo a un precedente lavoro di tesi [6] che è consistito per l'appunto nella cattura e preparazione dei dati e nella successiva estrazione dei

risultati che sono alla base del seguente lavoro di analisi e comparazione. Nel capitolo 2 verranno quindi presentati e descritti i risultati di questo precedente lavoro. Nel capitolo 3 verranno descritte le metodologie utilizzate per traslare questi risultati in un framework di lavoro; nel capitolo 4 saranno presentate e illustrate le metriche impiegate; nei capitoli 5 e 6 verrà sviluppata l'analisi e la comparazione dei risultati e, infine, nel capitolo 7, tratte le conclusioni.

# Capitolo 2

# Descrizione dei dati sperimentali

I risultati che sono alla base del seguente lavoro di analisi e comparazione sono il frutto di un precedente lavoro di tesi [6] che ha riguardato la cattura, il pre-processamento e l'applicazione delle tecniche di *data mining*, descritte nell'introduzione, ad un insieme di catture di traffico di rete *wireless* eseguite presso l'Ateneo. Le catture, come meglio descritto in seguito, sono state eseguite per più giorni in due differenti zone del campus utilizzando direttamente degli *sniffer* passivi *wireless* e non il traffico *wired* a valle degli access point.

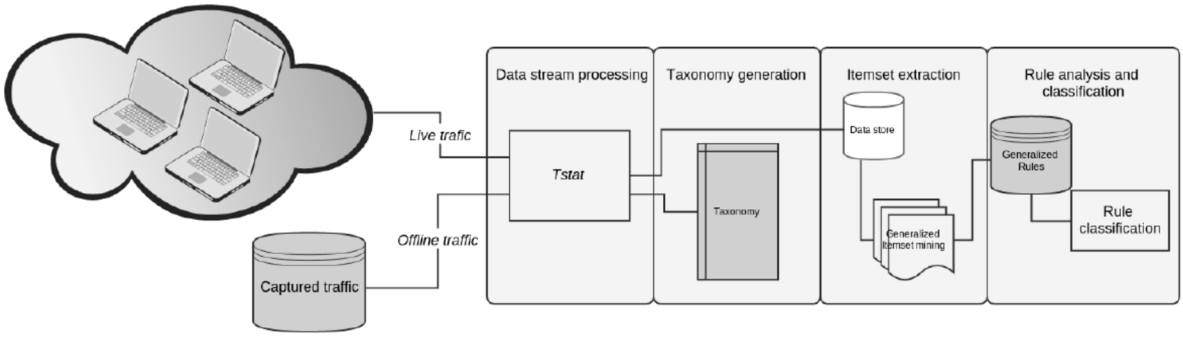



**Figura 3** — *Schema a blocchi del framework NetMine 2.0.*
(fonte: E. Osmani, "Analysis of wireless network data by means of generalized association rules" [6])

I dati di traffico grezzi così catturati sono stati sottoposti ad un pre-processamento attraverso un framework (*NetMine* 2.0) sviluppato dall'autore originale del lavoro [6], che a sua volta è basato su di uno strumento (*Tstat*) sviluppato dal *Telecommunication Network Group* dell'Ateneo.

Lo strumento *Tstat* è, nelle prime fasi, utilizzato per il filtraggio e la riduzione delle sequenze di datagrammi e segmenti a flussi bidirezionali e, successivamente, per la classificazione di questi in termini di protocolli e/o servizi mediante sia una analisi *port-based* sia una ispezione dei *payload*. L'esito è, a valle, trasformato in dataset strutturati testuali, suddivisi per giorno e luogo di cattura e organizzati in record costituiti da coppie <*caratteristica*, *valore*>, i quali sono poi sottoposti a un processo di estrazione degli *itemset* generalizzati frequenti e dei *Max-EGI* con differenti valori della soglia di supporto minimo, previa definizione di una tassonomia di generalizzazione.

Per ciascun insieme di *itemset* generalizzati frequenti sono infine ricavate, come ultimo processo di *data mining*, delle regole di associazione. I dettagli sul procedimento di cattura e sul funzionamento generale del framework *NetMine* sono stati interamente tratti dal lavoro di tesi [6]; invece,

le informazioni concrete sulla tassonomia impiegata, sulla struttura dei set di dati e sulle attività esperite di *data-mining* sono state desunte direttamente per analisi dei risultati, memorizzati su *file,* messi a disposizione dall'Ateneo.

## 2.1 Procedimento di cattura

La rete oggetto della cattura è stata la rete *wireless* dell'Ateneo, che è una rete *open* con un *captive-portal* per l'autenticazione. Le catture sono state eseguite in due distinte zone dell'Ateneo, scelte tra i luoghi maggiormente affollati dagli studenti al fine di ottenere la maggior quantità di traffico possibile: l'aula mensa e l'aula segreteria. Il monitoraggio è avvenuto nell'arco di 10 giorni ricompresi nelle prime due settimane di novembre 2013, nella fascia oraria dalle 10.00 alle 16.00, per un volume complessivo di dati giornaliero compreso tra i 10.0 e i 14.0 GiB. Nell'aula mensa la cattura si è estesa per 6 giorni, mentre nell'aula segreteria per 3 giorni.

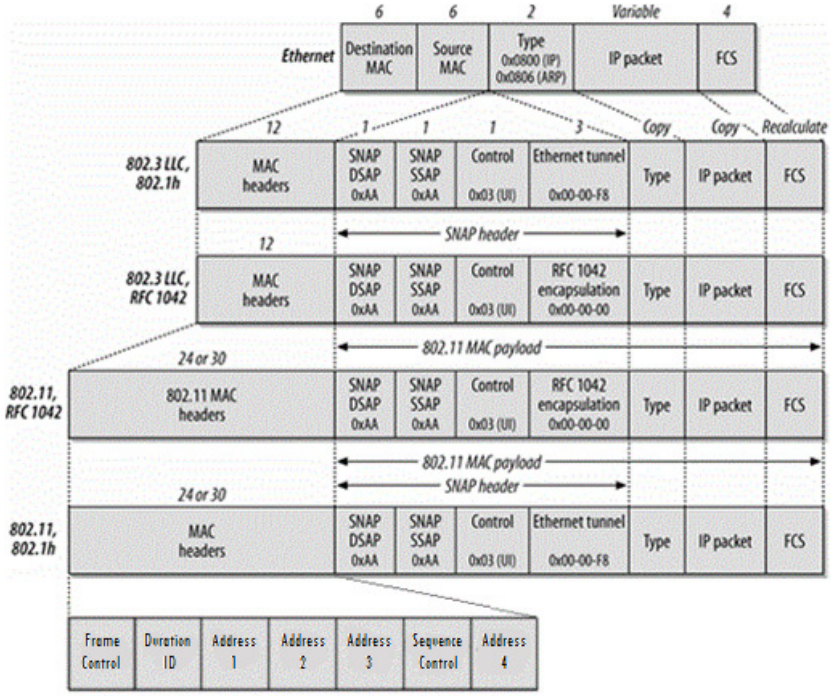



**Figura 4** — *Struttura comparata dei frame 802.11 e 802.3.*
(fonte: https://flylib.com/books/en/2.519.1/encapsulation_of_higher_layer_protocols_within_80211.html)

Operativamente, sono stati impiegati degli *sniffer* passivi «*AirPcap Classic USB 802.11b/g Adapter (capture only)*», in modo da catturare il traffico direttamente dai *frame* IEEE 802.11. Tale scelta ha peraltro cagionato la necessità di eseguire successivamente una trasformazione *off-line* delle catture utilizzando il tool *tcprewrite,* al fine di realizzare l'adattamento, come da figura 4, degli *headers* IEEE 802.11 al formato *wired* IEEE 802.3, l'unico utilizzabile dai *tool* successivi.

## 2.2 Processamento preliminare

Il processamento eseguito con *Tstat* ha permesso di filtrare, classificare e ricondurre il traffico a un elenco di flussi bidirezionali trasportati da `TCP` e un analogo elenco di messaggi coordinati trasportati da `UDP`. La direzionalità del flusso è stata inferita avendo riguardo a quale degli *endpoint* ha

inviato il primo pacchetto, talché questo è stato contrassegnato come *client*.

Per ciascun elenco sono state elaborate e memorizzate le seguenti informazioni:

- per i flussi trasportati da `TCP`:
  a) l'indirizzo `IPv4`, la porta e il numero di pacchetti inviati dall'*endpoint* client;
  b) l'indirizzo `IPv4`, la porta e il numero di pacchetti inviati dall'*endpoint* server;
  c) tre codici di classificazione relativi al tipo di connessione, al tipo di traffico `P2P` e al tipo di traffico `http`;
- per gli scambi di messaggi trasportati da `UDP`:
  a) l'indirizzo `IPv4`, la porta e il numero di pacchetti inviati dall'*endpoint* client, unitamente a un codice di classificazione;
  b) l'indirizzo `IPv4`, la porta e il numero di pacchetti inviati dall'*endpoint* server, unitamente a un codice di classificazione;

I due elenchi sono stati quindi ulteriormente processati da *NetMine* per derivare dalla pluralità dei codici originari di classificazione un'unica caratterizzazione del traffico[4] etichettata come «*protocol*» e infine fusi a formare un unico set di dati strutturato, esemplificato, per estrazione casuale, nella tabella 1.

**Tabella 1** — *Estratto esemplificativo dal dataset prodotto da NetMine.*

| ipsource | portsource | c2s_packets | ipdest | portdest | s2c_packets | protocol |
|---|---|---|---|---|---|---|
| … | … | … | … | … | … | … |
| 172.20.90.11 | 54651 | (0-140] | 195.248.250.100 | 80 | (0-45] | HTTP_GET |
| 172.20.91.15 | 50328 | (0-140] | 213.254.17.111 | 80 | (0-45] | HTTP_GET |
| 172.20.91.28 | 49165 | (0-140] | 74.201.86.29 | 443 | (45-90] | SSL/TLS |
| 172.20.91.28 | 49213 | (0-140] | 108.160.162.111 | 80 | (45-90] | HTTP_DROPBOX |
| 172.20.91.118 | 64349 | (0-140] | 63.111.29.135 | 443 | (45-90] | SSL/TLS |
| 172.20.91.23 | 34338 | (0-140] | 184.173.136.74 | 443 | (180-225] | Bittorrent_MSE/PE |
| 172.20.91.202 | 49294 | (0-140] | 108.160.163.49 | 80 | (45-90] | HTTP_DROPBOX |
| … | … | … | … | … | … | … |

Si può osservare come sia stata anche eseguita una discretizzazione, in *bucket* equidimensionati, del numero di pacchetti scambiati nelle due direzioni: «*c2s_packets*»[5] e «*s2c_packets*»[6].

## 2.3 Dettagli sulla tassonomia

Come illustrato nell'introduzione, alla base delle tecniche di estrazione fin qui analizzate vi è la definizione di una tassonomia fondata su opportune relazioni gerarchiche di tipo $x \lhd y$ che possono essere prefissate con criteri *domain-driven* o anche generate in modo automatico e adattivo a seguito di una analisi dei domini delle *feature*. Il *framework* NetMine permette di utilizzare entrambe le modalità e, in particolare, avendo riguardo ai *dataset* generati dal pre-processamento dei dati sperimentali, risulta che:

a) sono state prefissate con criteri *domain-driven* le gerarchie di «*portsource*», «*portdest*» e di «*protocol*»;

[4] Ovvero un tag testuale del tipo: `http_get`, `udp_dns`, `http_facebook` *et sim.*

[5] Ovvero il numero di pacchetti inviati dall'*endpoint client*

[6] Ovvero il numero di pacchetti inviati dall'*endpoint server*

b) sono state generate automaticamente e adattivamente le gerarchie di «*c2s_packets*» e di «*s2c_packets*»;
c) le gerarchie di «*ipsource*» e «*ipdest*» sono state prefissate per quanto riguarda il numero di livelli, ma con un automatismo non *domain-driven* per quanto riguarda la struttura.

Da tali gerarchie, sostanziate nei relativi alberi di generalizzazione, è conseguita la tassonomia impiegata per la generazione dei dati sperimentali delineata sinteticamente in figura 5. Si deve osservare che l'albero di generalizzazione della caratteristica «*protocol*» di figura 5 (b) desunto dai *dataset* effettivi è risultato diverso da quello descritto, nel lavoro di tesi [6], dall'autore del campionamento e del processamento delle catture.

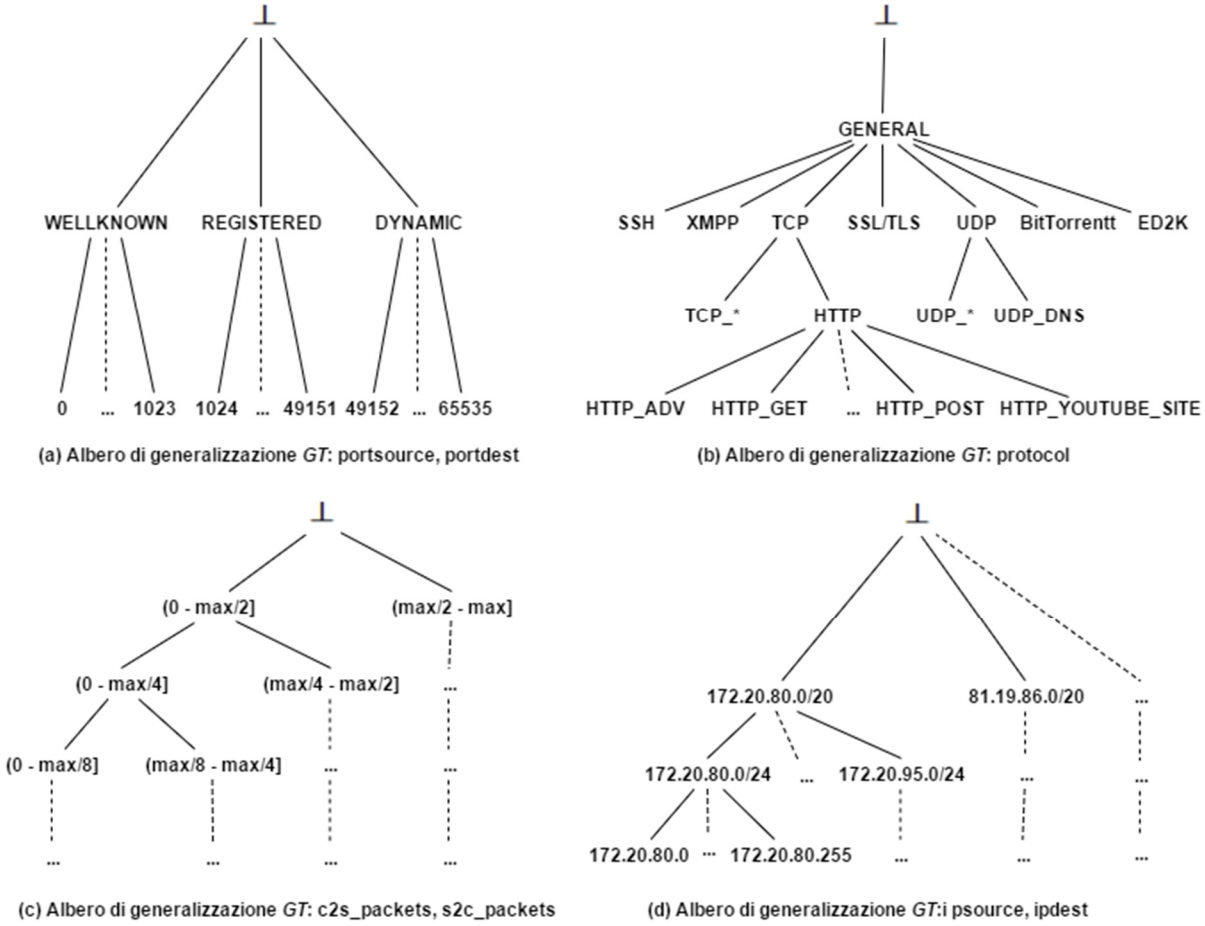



**Figura 5** — *Tassonomia impiegata per i dati sperimentali.*

## 2.4 Dettagli sui processi di data-mining sperimentalmente eseguiti

*NetMine* implementa due differenti algoritmi: (i) il *Genio Algorithm*, descritto in [3], per l'estrazione di *itemset* generalizzati classica e (ii) il *Max-EGI extraction algorithm*, descritto in [4], per l'estrazione dei *Max-EGI*.

Dall'analisi del complesso dei risultati presi a base del presente lavoro, risulta che le elaborazioni eseguite sono state le seguenti:

- sono stati estratti 9 distinti *set* di dati, nel formato di cui alla tabella 1, ciascuno riferentesi ad un singolo sito e giorno di cattura: ovvero 6 *dataset* per il sito «aula mensa» e 3 *dataset* per il

sito «aula segreteria»;

- su ciascun *dataset* del passo precedente sono stati eseguiti sia il *Genio Algorithm* sia il *Max-EGI extraction algorithm* distintamente e ciascuno con le seguenti soglie di supporto minimo percentuali: {0,1%, 0,2%, 0,4%, 0,8%, 1%, 1,5%, 2%, 3%, 4%, 5%} utilizzando la tassonomia schematizzata in figura 5, ottenendo quindi 90 insiemi di *itemset* generalizzati frequenti e 90 insiemi di *Max-EGI*;
- per ciascun insieme di *itemset* generalizzati frequenti sono state ricavate le inerenti regole di associazione, ristrette a quelle con confidenza maggiore di ½, ottenendo 90 insiemi di regole di associazione.

## 2.5 Dettagli sull'organizzazione fisica dei risultati sperimentali

Tutti i dati di quello che è, a un tempo, il risultato dell'esperimento descritto in questo capitolo e il punto di partenza dell'analisi e comparazione descritta in questa tesi nei capitoli successivi, sono contenuti in un insieme di file di testo organizzati in direttorî e sotto-direttorî.

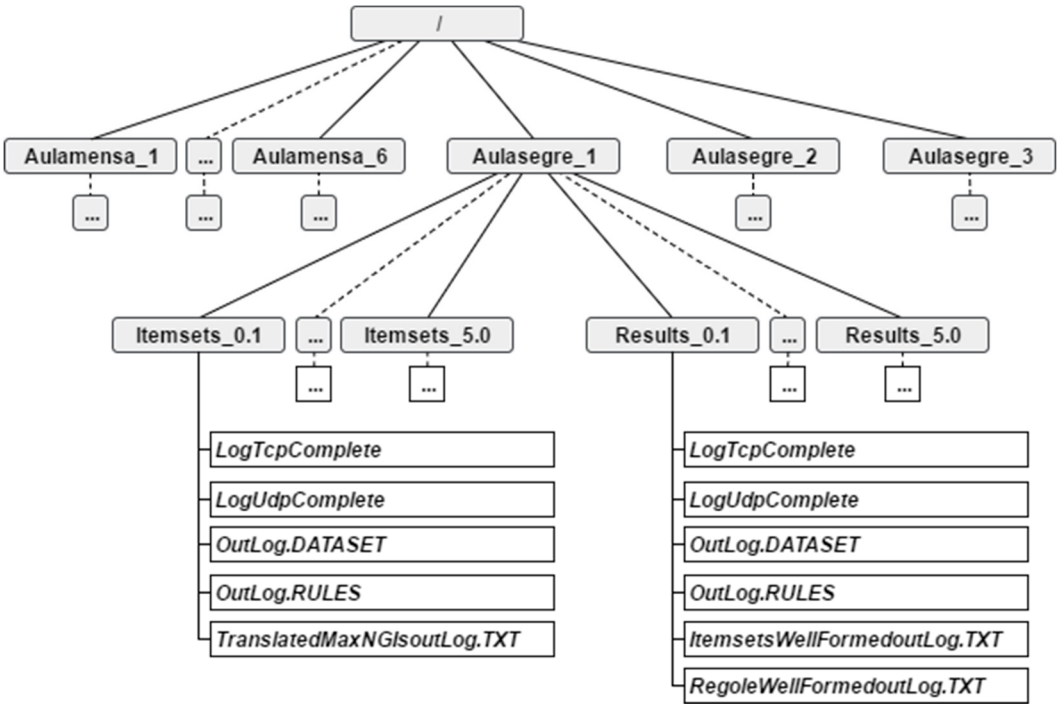



**Figura 6** — *Organizzazione in file e direttorî dei risultati sperimentali.*

Tale organizzazione è mostrata schematicamente nella figura 6, con solo qualche abbreviazione e adattamento nei nomi e con l'esclusione dei file intermedi e temporanei. Nella figura si distinguono 9 direttorî principali corrispondenti alle coppie sito/giorno, ciascuna delle quali contiene 10 sotto-direttorî — corrispondenti alle diverse soglie di supporto minimo — contenenti i *file* inerenti all'estrazione dei *Max-EGI* e altrettanti sotto direttorî contenenti i *file* inerenti all'estrazione degli *itemset* generalizzati frequenti con le relative regole d'associazione. In particolare:

- `LogTcpComplete` contiene l'elenco dei flussi trasportati da TCP con i relativi codici di classificazione;

- `LogUdpComplete` contiene l'elenco degli scambi di messaggi UDP con i relativi codici di classificazione;
- `OutLog.DATASET` contiene la rappresentazione del dataset, strutturato come in tabella 1;
- `OutLog.RULES` contiene la rappresentazione testuale della tassonomia generata e utilizzata;
- `ItemsetWellformedoutLog.TXT` contiene la rappresentazione testuale degli *itemset* generalizzati frequenti estratti ($X$) e la loro frequenza assoluta $\#X$;
- `RegoleWellformedoutLog.TXT` contiene la rappresentazione testuale delle regole di associazione $X \to Y$ ricavate dagli *itemset* frequenti generalizzati e i valori di $\mathsf{supp}(X \to Y)$ $\mathsf{conf}(X \to Y)$ e $\mathsf{lift}(X \to Y)$ di ciascuna;
- `TranslatedMaxNGIsoutLog.TXT` contiene la rappresentazione testuale dei *Max-EGI* ($X \wr S$) e le loro frequenze assolute $\#(X \wr S)$ e $\#S$.

# Capitolo 3

# Preparazione dei dati per l'analisi

La suddivisione dei risultati sperimentali descritti nel capitolo 2 in numerosi e distinti file di natura testuale e di vario formato avrebbe reso ardua e prona ad errori l'attività di analisi e comparazione oggetto di questo lavoro. Conseguentemente, si è ritenuto preliminarmente opportuno caricare tutti questi risultati in un'unica base di dati relazionale, in modo tale da poter utilizzare tutta la potenza espressiva del linguaggio SQL per l'esecuzione delle attività di analisi e di calcolo sulla generalità dei dati.

Tale scelta ha inoltre consentito di poter calcolare con relativa facilità indici e misure che sono state ampiamente utilizzate sia per la valutazione oggettiva sia per l'analisi *domain-driven*. Infatti, mentre per la valutazione dell'attendibilità e del grado di interesse delle regole di associazione esistono consolidate misure in letteratura, tra le quali sono ben note la *confidenza* e il *lift*, altrettanto non può dirsi per quanto riguarda gli *itemset* e quindi, per valutare e comparare direttamente questi ultimi (siano essi generalizzati o *Max-EGI*) è stato necessario formulare e calcolare per essi delle misure di correlazione e interesse che fossero equivalenti, in termini di efficacia, con quelle comunemente utilizzate con le regole di associazione, senza tuttavia far venir meno il rigore nel sottostante significato statistico. Nei successivi paragrafi verrà pertanto descritto l'ambiente di lavoro utilizzato, lo schema della base di dati relazionali e il processo di caricamento.

## 3.1 Dettagli sull'ambiente di lavoro

Tutte le attività sono state eseguite utilizzando una workstation dual processor Intel Xeon™, 3.20 GHz x 2, 4.0 GiB RAM, con S.O. Windows 10 Ultimate Education. Gli applicativi e *tool* software utilizzati sono stati esclusivamente:

- *Microsoft SQL Server 2014 Express* come DBMS relazionale;
- *Microsoft SQL Server Management Studio Express* come strumento GUI per l'esecuzione delle query e l'estrazione dei risultati di analisi;
- *Microsoft Log Parser 2.2*, come strumento versatile in grado di fornire accesso universale con query SQL-*like* ai dati basati su testo e con il quale è possibile indicare le informazioni desiderate, le modalità di elaborazione e reindirizzare i risultati delle query a un DMBS.

Tutti gli applicativi sopra indicati sono liberamente disponibili in formato eseguibile e scaricabili

dal sito web di Microsoft Corporation. Peraltro, applicativi analoghi sono disponibili sia su piattaforma Windows sia su altri S.O. e l'intero processo di caricamento e gestione avrebbe potuto essere eseguito su altre piattaforme con adattamenti minimi.

## 3.2 Caricamento della base di dati

### *3.2.1 Schema della base di dati*

Come accennato precedentemente, tutti i risultati alla base di questo lavoro, distribuiti su numerosi file testuali, sono stati caricati in una unica base di dati, il cui schema è mostrato in figura 7. Al fine di mantenere l'informazione sui *dataset* di origine e sui parametri dell'algoritmo eseguito, in tutte le relazioni sono presenti gli attributi «*site*» e «*day#*» che rappresentano il sito e il giorno della cattura; nelle relazioni contenenti *itemset* e regole d'associazione è inoltre presente l'attributo «*minsupp*» che rappresenta, in percentuale, la soglia minima di supporto utilizzata dall'algoritmo che li ha generati.

### *3.2.2 Processo di caricamento*

Tutto il processo di caricamento della base di dati è guidato da uno script della *shell*, che pone in esecuzione diversi moduli costituiti o da script di *Log Parser* o da script *Transact-SQL*. Il flusso esecutivo è schematizzato in figura 8 mentre il listato dello script principale e dei moduli più significativi è riportato nella Appendice A.

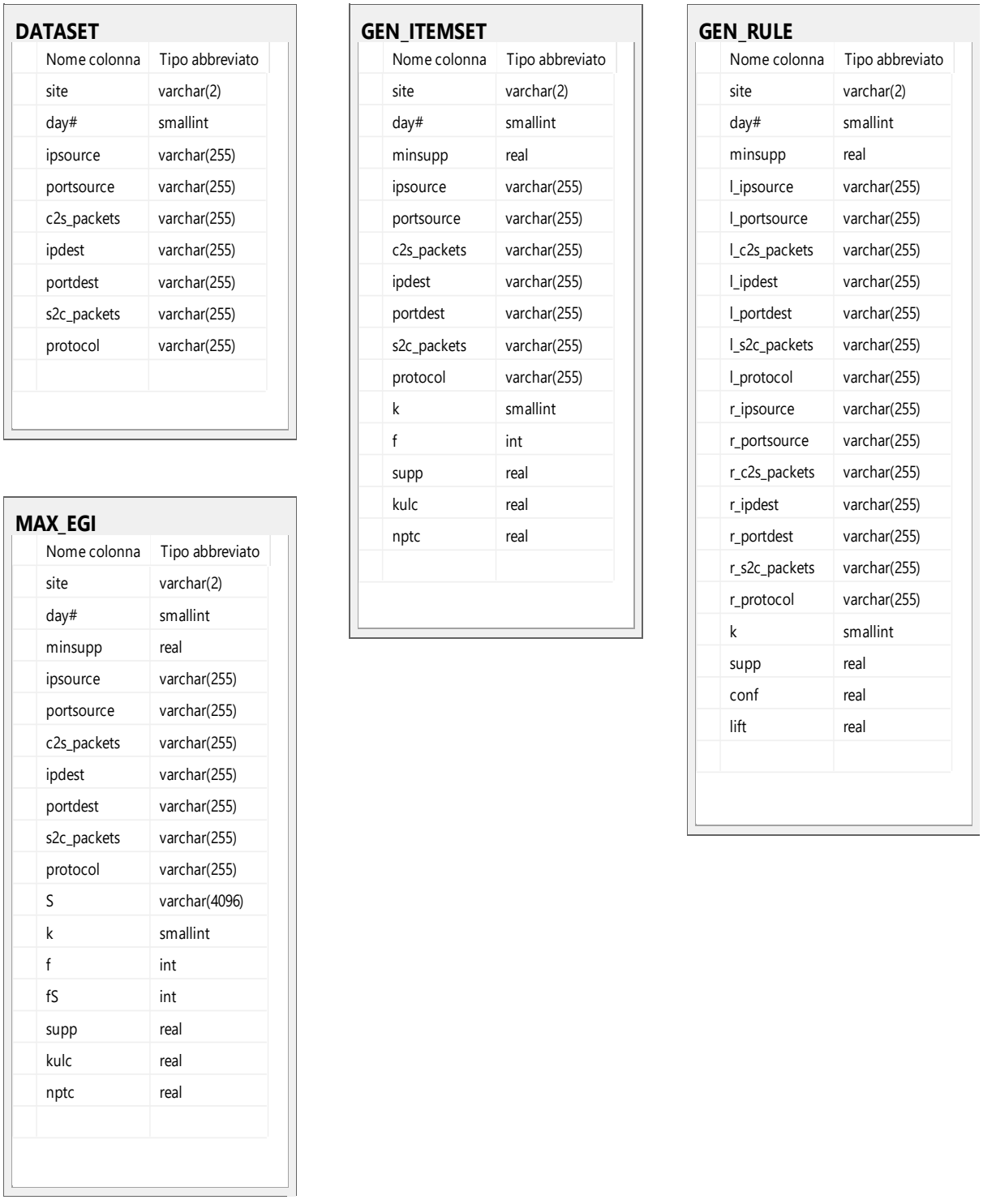
Sommariamente, lo script principale, dopo aver creato le tabelle nella base dati, esplora tutti i direttorî e sotto-direttorî illustrati nella figura 6 (provvedendo anche a gestire alcune irregolarità di fatto nella struttura) e invoca ripetutamente *Log Parser* che provvede a reperire e interpretare i file testuali `OutLog.DATASET`, `ItemsetWellformedoutLog.TXT`, `RegoleWellformedoutLog.TXT` e `TranslatedMaxNGIsoutLog.TXT` e a convertirli in *tuple* delle corrispondenti tabelle. Durante questo processo sono acquisite tutte le informazioni contenute nei predetti file e, inoltre, in base alle informazioni contenute nel nome dei direttorî e sotto-direttorî vengono generati gli attributi «*site*», «*day#*», «*minsupp*» e calcolata la lunghezza $k$ degli *itemset* o delle regole associative. Infine, a caricamento concluso, lo script provvede a eseguire ulteriori tre moduli, che provvedono a:

1. Organizzare le tabelle con indici *clustered* per migliorare le performance delle *query*.
2. Eliminare alcuni *itemset* e regole spuri, presenti già nei dati originali, che si sono rivelati durante una prima analisi e che sono verosimilmente riconducibili a un errore nell'implementazione degli algoritmi in *NetMine* che ha cagionato, per alcuni *itemset* e *Max-EGI* generalizzati sulla *feature* «*protocol*», la presenza nei *file* di più copie di essi, delle quali una corretta e le altre palesemente errate, avendo valori di $\#X$ o $\#(X \wr S)$ molto inferiori al reale.
3. Calcolare tutte le misure sui dati propedeutiche alle successive analisi, nonché definire le viste e le funzioni utilizzate nel corso di esse.

### *3.2.3 Consistenza numerica dei dati caricati*

Al termine del caricamento in base dati, le consistenze numeriche rilevate dei *dataset*, in termini

di numero di *record*, sono riportate nella tabella 2.



**DATASET**

| Nome colonna | Tipo abbreviato |
|---|---|
| site | varchar(2) |
| day# | smallint |
| ipsource | varchar(255) |
| portsource | varchar(255) |
| c2s_packets | varchar(255) |
| ipdest | varchar(255) |
| portdest | varchar(255) |
| s2c_packets | varchar(255) |
| protocol | varchar(255) |

**GEN_ITEMSET**

| Nome colonna | Tipo abbreviato |
|---|---|
| site | varchar(2) |
| day# | smallint |
| minsupp | real |
| ipsource | varchar(255) |
| portsource | varchar(255) |
| c2s_packets | varchar(255) |
| ipdest | varchar(255) |
| portdest | varchar(255) |
| s2c_packets | varchar(255) |
| protocol | varchar(255) |
| k | smallint |
| f | int |
| supp | real |
| kulc | real |
| nptc | real |

**GEN_RULE**

| Nome colonna | Tipo abbreviato |
|---|---|
| site | varchar(2) |
| day# | smallint |
| minsupp | real |
| l_ipsource | varchar(255) |
| l_portsource | varchar(255) |
| l_c2s_packets | varchar(255) |
| l_ipdest | varchar(255) |
| l_portdest | varchar(255) |
| l_s2c_packets | varchar(255) |
| l_protocol | varchar(255) |
| r_ipsource | varchar(255) |
| r_portsource | varchar(255) |
| r_c2s_packets | varchar(255) |
| r_ipdest | varchar(255) |
| r_portdest | varchar(255) |
| r_s2c_packets | varchar(255) |
| r_protocol | varchar(255) |
| k | smallint |
| supp | real |
| conf | real |
| lift | real |

**MAX_EGI**

| Nome colonna | Tipo abbreviato |
|---|---|
| site | varchar(2) |
| day# | smallint |
| minsupp | real |
| ipsource | varchar(255) |
| portsource | varchar(255) |
| c2s_packets | varchar(255) |
| ipdest | varchar(255) |
| portdest | varchar(255) |
| s2c_packets | varchar(255) |
| protocol | varchar(255) |
| S | varchar(4096) |
| k | smallint |
| f | int |
| fS | int |
| supp | real |
| kulc | real |
| nptc | real |

**Figura 7** — *Schema della base di dati.*

Per quanto riguarda gli *itemset* generalizzati, i *Max-EGI* e le regole di associazione estratti dai *dataset*, le consistenze numeriche sono suddivise in base alle 10 soglie di supporto minimo utilizzate dagli algoritmi. Tali consistenze sono mostrate rispettivamente nella tabella 3, nella tabella 4 e nella tabella 5.

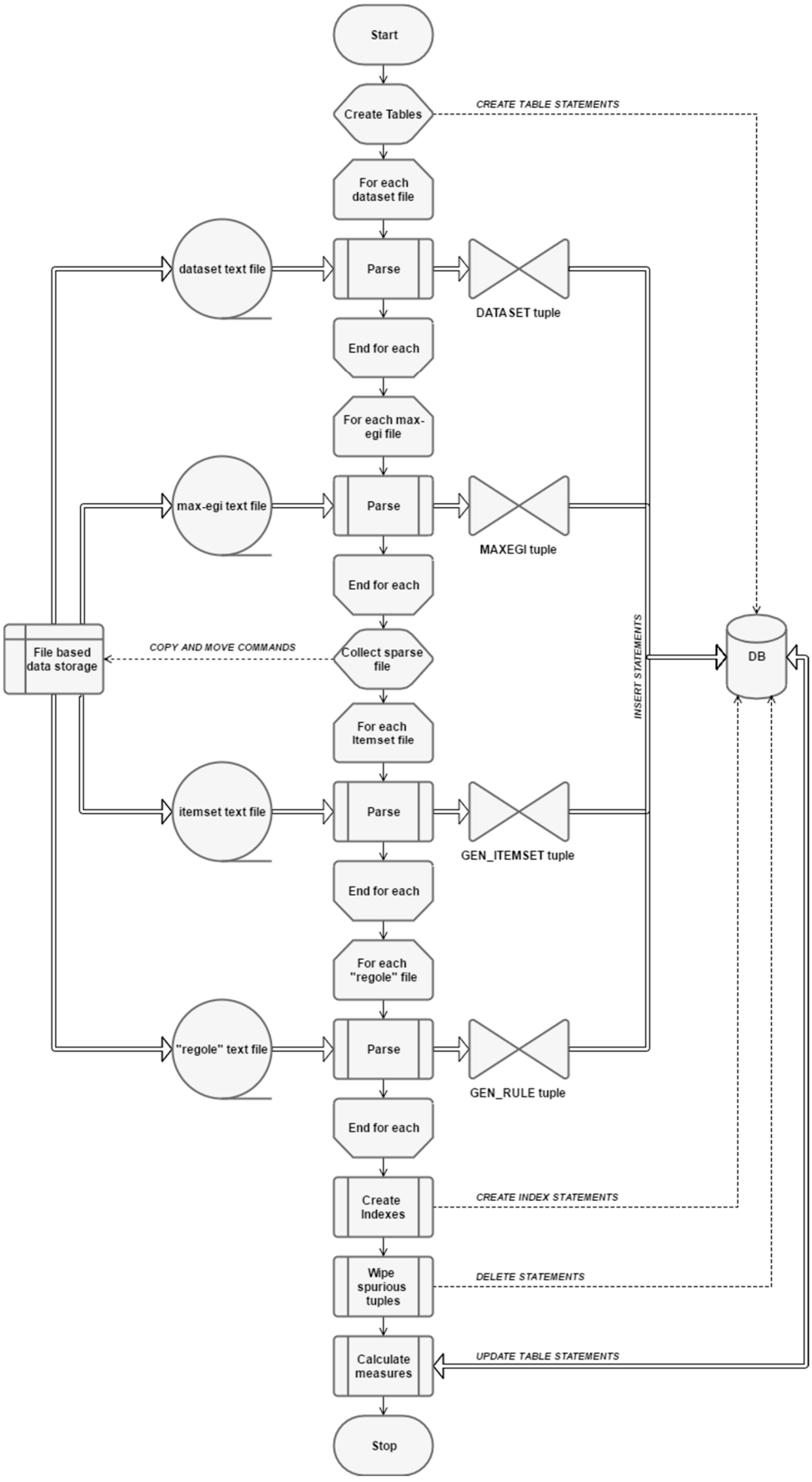



**Figura 8** — *Diagramma di flusso del processo di caricamento.*

**Tabella 2** — *Consistenza numerica dei dataset.*

| **Dataset** | **Sito** | **Giorno** | **Record** |
|---|---|---|---|
| M1 | Aula Mensa | 1 | 97.117 |
| M2 | Aula Mensa | 2 | 151.097 |
| M3 | Aula Mensa | 3 | 324.145 |
| M4 | Aula Mensa | 4 | 50.771 |
| M5 | Aula Mensa | 5 | 39.586 |
| M6 | Aula Mensa | 6 | 242.634 |
| S1 | Aula Segreteria | 1 | 160.463 |
| S2 | Aula Segreteria | 2 | 153.399 |
| S3 | Aula Segreteria | 3 | 77.590 |

**Tabella 3** — *Consistenza numerica degli itemset generalizzati.*

| Dataset | Soglia di supporto minimo (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0,1 | 0,2 | 0,4 | 0,8 | 1,0 | 1,5 | 2,0 | 3,0 | 4,0 | 5,0 |
| M1 | 34.191 | 20.158 | 9.272 | 3.511 | 2.362 | 1.593 | 1.279 | 871 | 652 | 501 |
| M2 | 30.749 | 16.948 | 7.281 | 3.017 | 2.404 | 1.693 | 1.203 | 807 | 719 | 621 |
| M3 | 33.912 | 14.842 | 6.362 | 2.365 | 1.741 | 1.109 | 952 | 743 | 604 | 491 |
| M4 | 28.912 | 16.519 | 9.417 | 4.811 | 3.592 | 2.241 | 1.706 | 926 | 679 | 539 |
| M5 | 30.660 | 16.346 | 8.890 | 4.411 | 3.784 | 2.475 | 1.700 | 932 | 758 | 606 |
| M6 | 34.313 | 15.516 | 6.795 | 3.104 | 2.436 | 1.540 | 1.042 | 815 | 570 | 456 |
| S1 | 29.173 | 16.270 | 7.615 | 3.846 | 3.020 | 1.863 | 1.313 | 871 | 722 | 493 |
| S2 | 35.249 | 19.909 | 8.190 | 3.113 | 2.253 | 1.499 | 1.168 | 897 | 795 | 690 |
| S3 | 25.451 | 15.627 | 9.093 | 4.244 | 3.255 | 2.111 | 1.837 | 1.041 | 738 | 458 |

**Tabella 4** — *Consistenza numerica dei Max-EGI.*

| Dataset | Soglia di supporto minimo (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0,1 | 0,2 | 0,4 | 0,8 | 1,0 | 1,5 | 2,0 | 3,0 | 4,0 | 5,0 |
| M1 | 16.232 | 9.557 | 4.447 | 1.727 | 1.175 | 799 | 627 | 449 | 343 | 263 |
| M2 | 14.855 | 8.303 | 2.962 | 1.511 | 1.179 | 823 | 575 | 403 | 359 | 311 |
| M3 | 16.494 | 7.355 | 3.195 | 1.223 | 963 | 567 | 483 | 379 | 311 | 247 |
| M4 | 13.851 | 7.919 | 4.479 | 2.271 | 1.711 | 1.087 | 827 | 479 | 347 | 271 |
| M5 | 14.558 | 7.945 | 4.361 | 2.171 | 1.839 | 1.203 | 815 | 467 | 379 | 307 |
| M6 | 16.011 | 7.339 | 3.263 | 1.519 | 1.219 | 763 | 507 | 395 | 295 | 231 |
| S1 | 14.283 | 8.095 | 3.855 | 1.979 | 1.555 | 975 | 639 | 443 | 371 | 255 |
| S2 | 16.819 | 9.519 | 3.995 | 1.559 | 1.107 | 747 | 571 | 443 | 391 | 339 |
| S3 | 11.943 | 7.323 | 4.299 | 2.091 | 1.615 | 1.043 | 883 | 535 | 387 | 227 |

**Tabella 5** — *Consistenza numerica delle regole d'associazione estratte dagli itemset generalizzati.*

| Dataset | Soglia di supporto minimo (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0,1 | 0,2 | 0,4 | 0,8 | 1,0 | 1,5 | 2,0 | 3,0 | 4,0 | 5,0 |
| M1 | 259.854 | 155.480 | 65.473 | 20.149 | 13.105 | 8.724 | 7.150 | 4.722 | 3.301 | 2.495 |
| M2 | 207.618 | 111.326 | 43.710 | 17.514 | 14.701 | 10.697 | 6.752 | 4.797 | 4.042 | 3.494 |
| M3 | 213.677 | 90.146 | 37.484 | 12.701 | 9.158 | 6.005 | 5.383 | 4.171 | 3.443 | 2.768 |
| M4 | 215.407 | 120.691 | 67.579 | 31.294 | 22.096 | 12.884 | 9.914 | 5.098 | 3.728 | 3.262 |
| M5 | 214.767 | 105.589 | 57.438 | 28.976 | 25.301 | 15.492 | 10.715 | 5.022 | 4.018 | 3.304 |
| M6 | 241.912 | 105.683 | 41.533 | 18.473 | 15.214 | 8.652 | 5.716 | 4.618 | 3.302 | 2.698 |
| S1 | 185.351 | 101.589 | 47.022 | 23.270 | 18.250 | 11.062 | 7.681 | 5.019 | 3.974 | 2.560 |
| S2 | 249.912 | 140.602 | 54.629 | 19.613 | 14.386 | 9.356 | 7.955 | 6.358 | 5.973 | 5.397 |
| S3 | 171.537 | 101.122 | 58.199 | 26.177 | 19.468 | 12.710 | 10.918 | 5.969 | 4.246 | 2.538 |

## 3.3 Calcolo delle misure sui dati.

Al fine di condurre le successive analisi in modo efficiente, le principali misure impiegate sono state calcolate *una tantum* durante il processo di caricamento e preparazione. Pur potendosi, in tale sede, procedere al calcolo di numerose misure, si è scelto, per motivi che saranno chiari più avanti, di calcolarne due, che fossero, a un tempo, complementari tra loro e ben fondate in termini teorici; inoltre, se ne è introdotta e calcolata una terza in grado di esprimere la riunione delle caratteristiche peculiari delle prime due in un unico indicatore di interesse congiunto.

In particolare, sono state direttamente pre-calcolate, sia per gli *itemset* generalizzati sia per i *Max-EGI*: (i) una misura *null-invariant* [7], rappresentata dalla ben nota misura di Kulczynsky (o *kulc*), (ii) una misura *expectation-based,* derivata da [8], [9] e [10] e di seguito riferita come *normalized pointwise total correlation* (o *nptc*) e, infine, (iii) una misura di interesse congiunto di seguito riferita come *zeta*. Tutte queste misure sono analizzate criticamente e presentate nel successivo capitolo 4, al quale si rimanda per la loro descrizione e giustificazione, la loro applicabilità e i loro fondamenti teorici.

In questa sede basti per ora rammentare che nei dati sperimentali erano già disponibili la cardinalità $\#X$ per gli *itemset* generalizzati e le cardinalità $\#(X \wr S)$ e $\#S$ per i *Max-EGI*, così come la lunghezza $k$ di entrambi, rilevata durante la generazione delle *tuple*. I calcoli computazionalmente più significativi, pertanto, sono stati quelli necessari per ricavare da esse la misura di Kulczynsky, ovvero

$$\mathrm{kulc}(X) = \frac{1}{k}\sum_{i=1}^{k}\frac{\#X}{\#x_i} \tag{3.1}$$

$$\mathrm{kulc}(X \wr S) = \frac{1}{k}\sum_{i=1}^{k}\frac{\#(X \wr S)}{\#x_i - \#S} \tag{3.2}$$

e il precursore della *nptc*, ovvero la *pointwise total correlation* neperiana e non standardizzata

$$\mathrm{ptc}(X) = (k-1)\ln|\mathcal{D}| + \ln \#X - \sum_{i=1}^{k}\ln \#x_i \tag{3.3}$$

$$\mathrm{ptc}(X \wr S) = (k-1)\ln|\mathcal{D}| + \ln \#(X \wr S) - \sum_{i=1}^{k}\ln(\#x_i - \#S) \tag{3.4}$$

e ciò in ragione soprattutto della necessità di procedere all'associazione tra le *tuple* recanti gli $\#X$ o gli $\#(X \wr S)$ con quelle loro corrispondenti recanti gli $\#x_i$ (ovvero quel che tecnicamente viene detto un *self-join*).

Si noti che la (3.2) e la (3.4) sono così formulate in conseguenza dell'applicazione di un teorema sui *Max-EGI* anch'esso illustrato e dimostrato nel successivo capitolo 4.

## 3.4 Partizionamento logico dei dati

Per meglio realizzare l'obiettivo di confrontare gli esiti dell'applicazione del *Genio algorithm* con quelli del *Max-EGI extraction algorithm* si è ritenuto opportuno separare i risultati peculiari

dell'uno e dell'altro dai risultati che essi forniscono in comune. Si è pertanto proceduto in un primo tempo all'unione di tutti gli *itemset* provenienti dai diversi *dataset* in modo tale da ottenere degli insiemi complessivi, distinti e caratterizzati ciascuno dal riunire tutti e soli gli *itemset* provenienti da *dataset* dotati del medesimo supporto minimo. La liceità di tale unione giace sulla considerazione che i diversi *dataset* di origine sono in realtà partizioni, disgiunte nel tempo e nello spazio, delle osservazioni di un medesimo fenomeno e, come tali, riunibili. Ciascuno di questi insiemi omogenei è stato quindi suddiviso in tre partizioni distinte e disgiunte rappresentanti:

1. l'insieme degli *itemset* generalizzati espressivi prodotti esclusivamente dal *Max-EGI extraction algorithm* e indicato nel seguito brevemente come $\mathsf{M \backslash G}$,
2. l'insieme degli *itemset* generalizzati prodotti esclusivamente dal *Genio algorithm* e indicato come $\mathsf{G \backslash M}$ e, infine,
3. l'insieme degli *itemset* (generalizzati) prodotti in comune da ambedue gli algoritmi, indicato come $\mathsf{G \cap M}$.

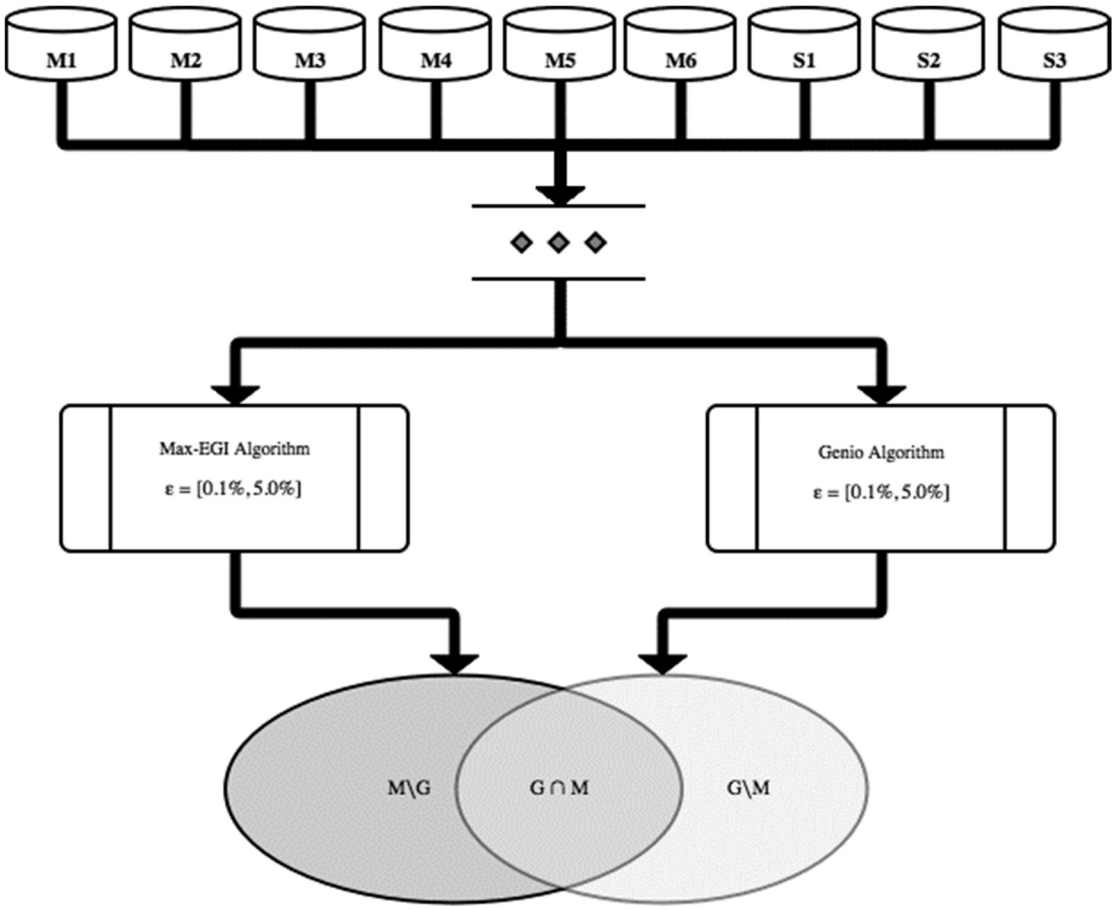



**Figura 9** — *Partizionamento dei risultati sperimentali.*

Un riepilogo di come è avvenuto il partizionamento sopra descritto è schematizzato in figura 9 e le relative cardinalità sono riportate nella tabella 6. Si può ivi osservare che il numero di *itemset* generalizzati espressivi confluiti in $\mathsf{M \backslash G}$ è decisamente basso e, decrescendo all'aumentare della soglia di supporto minimo, giunge sino ad azzerarsi, per la soglia del 5%, in alcuni *dataset*. Ragion per cui, in talune analisi, non si è andati oltre la soglia del 3% per mantenersi entro una ragionevole significatività statistica.

**Tabella 6** — *Cardinalità delle partizioni M\G, G∩M e G\M (numero di k-itemset con $k \geq 2$).*

| Dataset di origine | Partizione di destinazione | Soglia di supporto minimo (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0,1 | 0,2 | 0,4 | 0,8 | 1,0 | 1,5 | 2,0 | 3,0 | 4,0 | 5,0 |
| M1 | M\G | 474 | 277 | 125 | 41 | 37 | 37 | 18 | 20 | 14 | 10 |
| | M∩G | 15.465 | 9.070 | 4.205 | 1.615 | 1.083 | 721 | 576 | 403 | 306 | 233 |
| | G\M | 18.426 | 10.870 | 4.943 | 1.818 | 1.217 | 824 | 662 | 434 | 315 | 240 |
| M2 | M\G | 399 | 228 | 61 | 67 | 37 | 25 | 18 | 18 | 11 | 11 |
| | M∩G | 14.184 | 7.888 | 2.786 | 1.382 | 1.095 | 762 | 526 | 362 | 326 | 279 |
| | G\M | 16.284 | 8.864 | 4.371 | 1.566 | 1.255 | 888 | 638 | 414 | 362 | 312 |
| M3 | M\G | 449 | 278 | 92 | 56 | 56 | 25 | 25 | 21 | 14 | 0 |
| | M∩G | 15.707 | 6.870 | 2.982 | 1.108 | 856 | 510 | 432 | 336 | 278 | 231 |
| | G\M | 17.860 | 7.759 | 3.252 | 1.191 | 827 | 559 | 486 | 377 | 299 | 235 |
| M4 | M\G | 454 | 215 | 155 | 67 | 51 | 40 | 22 | 11 | 15 | 18 |
| | M∩G | 13.159 | 7.547 | 4.220 | 2.127 | 1.597 | 1.001 | 766 | 440 | 308 | 234 |
| | G\M | 15.507 | 8.807 | 5.087 | 2.601 | 1.926 | 1.187 | 893 | 449 | 338 | 278 |
| M5 | M\G | 298 | 211 | 170 | 92 | 37 | 41 | 22 | 14 | 10 | 7 |
| | M∩G | 14.007 | 7.575 | 4.093 | 2.008 | 1.739 | 1.112 | 757 | 427 | 345 | 281 |
| | G\M | 16.392 | 8.603 | 4.691 | 2.326 | 1.975 | 1.306 | 899 | 471 | 381 | 297 |
| M6 | M\G | 328 | 215 | 80 | 48 | 29 | 25 | 18 | 18 | 18 | 18 |
| | M∩G | 15.350 | 6.931 | 3.071 | 1.406 | 1.140 | 699 | 456 | 351 | 256 | 196 |
| | G\M | 18.621 | 8.385 | 3.605 | 1.626 | 1.239 | 795 | 545 | 430 | 285 | 235 |
| S1 | M\G | 392 | 235 | 147 | 53 | 49 | 30 | 22 | 14 | 7 | 0 |
| | M∩G | 13.628 | 7.687 | 3.619 | 1.866 | 1.451 | 902 | 585 | 405 | 343 | 238 |
| | G\M | 15.274 | 8.403 | 3.900 | 1.913 | 1.507 | 910 | 688 | 434 | 349 | 229 |
| S2 | M\G | 379 | 255 | 116 | 44 | 44 | 25 | 18 | 18 | 14 | 14 |
| | M∩G | 16.128 | 9.037 | 3.748 | 1.449 | 1.009 | 680 | 521 | 399 | 355 | 305 |
| | G\M | 18.800 | 10.638 | 4.304 | 1.591 | 1.183 | 770 | 607 | 464 | 410 | 357 |
| S3 | M\G | 211 | 128 | 78 | 54 | 66 | 54 | 30 | 3 | 7 | 0 |
| | M∩G | 11.534 | 7.065 | 4.137 | 1.987 | 1.505 | 950 | 816 | 504 | 359 | 212 |
| | G\M | 13.711 | 8.425 | 4.865 | 2.200 | 1.699 | 1.115 | 977 | 501 | 350 | 223 |

# Capitolo 4

# Giustificazione delle misure sui dati

Riprendendo l'affermazione espressa preliminarmente nell'introduzione e cioè che uno dei compiti centrali delle attività di *data-mining* è lo scoprire correlazioni nascoste tra i dati, non è difficile convincersi che la frequenza di co-occorrenza non è, di per sé, sufficiente per inferirle. È stato pertanto storicamente necessario formulare ulteriori misure per indicare il grado di significatività e di interesse delle co-occorrenze, fondandole sulla considerazione che sebbene la "*correlation does not imply causation*", essa ne costituisce comunque un indizio. In questo capitolo verranno pertanto introdotte, dopo una sistemazione teorica, alcune misure di questa natura e giustificate quelle che si è scelto di calcolare sui dati sperimentali.

## 4.1 Interpretazione probabilistica delle misure sugli itemset

Dal punto di vista statistico e probabilistico, un *dataset* strutturato $\mathcal{D}$ può essere visto come una sequenza finita di esperimenti ripetuti dei quali ciascun *record* $r \in \mathcal{D}$ rappresenta l'esito. Posto che ciascun *record* sia rappresentato da un insieme $r = \{r_1, r_2, \dots, r_n\}$ allora lo spazio campionario del singolo esperimento sarà dato dal prodotto cartesiano $\Omega = \Omega_1 \times \Omega_2 \times \dots \times \Omega_n$, dove ciascun singolo spazio $\Omega_i$ rappresenta un insieme finito o numerabile degli eventi elementari (ovvero di tutti i possibili valori assumibili) dallo $i$-esimo attributo di $r$. Si può a questo punto prendere come spazio degli eventi l'insieme delle parti $\wp(\Omega)$ che costituisce senz'altro, per la finitezza o numerabilità di $\Omega$, una $\sigma$-algebra e assegnare a ciascun evento elementare $\omega = \{\omega_1, \omega_2, \dots, \omega_n\} \in \Omega$ una probabilità osservata, in ossequio alla interpretazione classica, così determinata

$$\hat{p}(\{\omega_1, \omega_2, \dots, \omega_n\}) = \frac{\left|\{r \in \mathcal{D} : r = \{\omega_1, \omega_2, \dots, \omega_n\}\}\right|}{|\mathcal{D}|} \tag{4.1}$$

Si consideri ora l'insieme $\mathfrak{X}$ di tutti gli *itemset* (generalizzati (espressivi)) che possono essere costruiti sul *dataset* $\mathcal{D}$ usando una tassonomia $\Gamma$ (che può anche essere $\Gamma = \emptyset$). Si può allora definire quanto segue.

**Definizione 9** (*Evento associato*) Sia $\omega = \{\omega_1, \omega_2, \dots, \omega_n\} \in \Omega$ un evento elementare dello spazio campionario e $X \wr S \in \mathfrak{X}$ un *itemset* (generalizzato (espressivo)). Allora $(X \wr S) \xrightarrow{\text{associa}} \omega$ se e solo se valgono alternativamente e ricorsivamente le seguenti:

- $(\Gamma = \emptyset) \wedge (S = \emptyset) \wedge (X \subseteq \omega)$, oppure
- $(\Gamma \neq \emptyset) \wedge (S = \emptyset) \wedge (\forall\, x \in X \backslash \omega\ \exists\ \omega_i \in \omega \backslash X : (\omega_i \lhd x) \in \Gamma)$, oppure
- $(\Gamma \neq \emptyset) \wedge (S \neq \emptyset) \wedge (X \wr \emptyset \xrightarrow{\text{associa}} \omega)\ \wedge\ (\nexists\, Y \in S : Y \xrightarrow{\text{associa}} \omega)$.

Si può allora costruire formalmente una funzione suriettiva di proiezione $\pi\colon \mathfrak{X} \longmapsto \wp(\Omega)$ che associa ad ogni $X \wr S \in \mathfrak{X}$ un elemento dell'insieme delle parti e spazio degli eventi $\wp(\Omega)$

$$\pi(X \wr S) = \left\{\omega \in \Omega : (X \wr S) \xrightarrow{\text{associa}} \omega\right\} \tag{4.2}$$

ovvero un formale evento aleatorio dotato di probabilità $\hat{p}\big(\pi(X \wr S)\big)$.

Si noti che confrontando la definizione 6 con la definizione 9 e le (1.1) e (1.3) con le definizioni sopra indicate di evento associato a un *itemset* (generalizzato (espressivo)) si può facilmente concludere che, con gli assunti fatti, vale l'equivalenza

$$\operatorname{supp}(X \wr S) = \hat{p}\big(\pi(X \wr S)\big) \tag{4.3}$$

della quale ci si avvarrà nel seguito della trattazione, così come di tutte le altre nozioni e formule legate al concetto di probabilità di eventi. Talvolta, per brevità, l'indicazione della funzione di proiezione sullo spazio degli eventi $\pi$ sarà omessa oppure sarà sostituita semplicemente dalle parentesi biquadre $[\![\cdot]\!]$ in tutti i casi in cui potrebbero sorgere delle ambiguità[7].

## 4.2 Misure di correlazione e interesse

Operativamente, l'estrazione degli *itemset* (generalizzati) frequenti è un primo passo propedeutico all'estrazione da questi ultimi di regole d'associazione suscettibili di mostrare un grado di significatività minimo — ovvero superiore a una soglia prefissata $\gamma_0$ — che viene indicato come confidenza [1]. Tuttavia, pur introducendo una valutazione statisticamente ragionevole e più accurata della semplice frequenza di co-occorrenza, anche tale grado di confidenza non è di per sé adeguato a misurare l'effettivo interesse della regola, anzi talvolta risulta addirittura fuorviante, talché altre misure, e segnatamente tra queste il c.d. *interest* o *lift* [11], sono state introdotte per conseguire una valutazione maggiormente attendibile. Formalmente, le due misure testé introdotte sono così formulate:

$$\operatorname{conf}(X \to Y) = \frac{\operatorname{supp}(X \cup Y)}{\operatorname{supp}(X)} = \frac{\hat{p}([\![X]\!] \cap [\![Y]\!])}{\hat{p}([\![X]\!])} = \hat{p}([\![Y]\!] \mid [\![X]\!]) \tag{4.4}$$

$$\operatorname{lift}(X \to Y) = \operatorname{lift}(Y \to X) = \frac{\operatorname{conf}(X \to Y)}{\operatorname{supp}(Y)} = \frac{\hat{p}([\![Y]\!] \mid [\![X]\!])}{\hat{p}([\![Y]\!])} = \frac{\hat{p}([\![X]\!] \cap [\![Y]\!])}{\hat{p}([\![X]\!]) \cdot \hat{p}([\![Y]\!])} \tag{4.5}$$

Si noti il significato statistico sottostante, che mostra come la confidenza misuri la probabilità osservata di trovare gli *item* rappresentati dal conseguente $Y$, sotto la condizione che nella transazione o *record* siano già presenti gli *item* rappresentati dall'antecedente $X$; mentre il *lift* misura

[7] Poiché un itemset $X$ è un insieme e anche la sua proiezione nello spazio degli eventi $\pi(X)$ è, a sua volta, un insieme, una scrittura $X \cup Y$ potrebbe risultare ambigua, non essendo chiaro se si intende l'unione di due *itemset* o l'unione degli eventi associati a due *itemset*, che avrebbe significato opposto. Quindi, nel secondo caso si scriverà $\pi(X) \cup \pi(Y)$ oppure, più frequentemente, $[\![X]\!] \cup [\![Y]\!]$.

direttamente il rapporto tra la probabilità congiunta degli eventi rappresentati da $X$ e $Y$ e la probabilità attesa qualora gli eventi rappresentati da $X$ e $Y$ fossero statisticamente indipendenti[8].

Si deve osservare che la confidenza risente sensibilmente della frequenza relativa osservata $\hat{p}([\![Y]\!])$, talché, in presenza di conseguenti molto frequenti, essa tende ad assumere valori molto elevati anche in assenza di un effettivo legame. D'altro canto, anche il *lift* è suscettibile di esprimere valori fuorvianti soprattutto in presenza di *dataset* dove la probabilità attesa stimata risulti estremamente piccola e pertanto la casuale e sporadica co-occorrenza di $X$ e $Y$ pur non avendo un reale valore statistico è tuttavia suscettibile di produrre elevati (e ingiustificati) valori della misura. Nondimeno, se si prendono le due misure insieme, si può notare che mentre il *lift* in qualche modo dipende dalla cardinalità del *dataset*, la confidenza non ne è invece influenzata e soprattutto non è influenzata dai *record* che non concorrono al suo calcolo[9], mentre d'altro lato, in quei casi ove la confidenza diviene inattendibile per l'elevata frequenza del conseguente, ciò invece non influenza affatto il *lift*. Da ciò si può congetturare che l'osservazione congiunta di due misure di questo tipo possa mitigare il *misleading* che esse possono cagionare se osservate separatamente.

Poiché questo lavoro si propone soprattutto di valutare e comparare il grado di interesse degli *itemset* generalizzati prodotti da un algoritmo standard confrontandolo con quello dei *Max-EGI* tratti dai medesimi *dataset*, il primo ostacolo da superare è rappresentato dal fatto che le misure testé introdotte sono ben definite solo per le regole d'associazione e non per gli *itemset* che le originano. Pertanto, vi è la necessità di formulare delle misure sugli *itemset* che siano in qualche modo riconducibili a quelle definite per le regole e quindi, considerandone le caratteristiche, costituite da almeno una misura *null-invariant* e da almeno una misura *expectation-based*.

## 4.3 Misure di correlazione e interesse per gli itemset

Per gli *itemset* sono ben note in letteratura [7] numerose misure *null-invariant*, alcune delle quali sono mostrate nella tabella 7.

**Tabella 7** — *Misure null-invariant definite sugli itemset*

| Nome | Misura | Significato |
|---|---|---|
| *All-Confidence* | $\min_{i=1}^{k} \dfrac{p(\xi_1, \xi_2, \dots, \xi_k)}{p(\xi_i)}$ | minimo |
| *Coherence* | $\left( \dfrac{1}{k} \displaystyle\sum_{i=1}^{k} \dfrac{p(\xi_i)}{p(\xi_1, \xi_2, \dots, \xi_k)} \right)^{-1}$ | media armonica |
| *Cosine* | $\sqrt[k]{\displaystyle\prod_{i=1}^{k} \dfrac{p(\xi_1, \xi_2, \dots, \xi_k)}{p(\xi_i)}}$ | media geometrica |
| *Kulczynsky* | $\dfrac{1}{k} \displaystyle\sum_{i=1}^{k} \dfrac{p(\xi_1, \xi_2, \dots, \xi_k)}{p(\xi_i)}$ | media aritmetica |

[8] Si tratta pertanto di una misura cosiddetta *expectation-based*

[9] Il che ne fa, per l'appunto, una misura cosiddetta *null (transaction) invariant*

Si noti che tutte quelle mostrate condividono un termine comune che rappresenta una probabilità condizionata $p(\xi_1, \xi_2, \dots, \xi_k)/p(\xi_i)$ la quale, a sua volta, può essere interpretata, posto che sia $\xi_i = [\![x_i]\!]\ \forall x_i \in X$, come la confidenza di una generica regola di forma $\{x_i\} \to X \backslash \{x_i\}$ ottenibile da un generico *itemset* $X$. Con questa interpretazione appare chiaro come queste misure rappresentino delle stime o del valor minimo o più sovente del valor medio della confidenza delle più semplici regole d'associazione generabili dall'*itemset* in esame, che condensano nell'intervallo $[0, 1]$ i possibili esiti quali indicatori di correlazione negativa, neutralità o correlazione positiva.

Una misura *expectation-based* si può invece derivare da una delle generalizzazioni a $k > 2$ variabili aleatorie della ben nota *mutual information*, proposta come *total correlation* [8] o *multi-information* [9] e definita come la seguente divergenza di Kullback-Leibler

$$C(X_1, X_2, \cdots, X_k) = \sum_{x_1 \in X_1} \sum_{x_2 \in X_2} \cdots \sum_{x_k \in X_k} p(x_1, x_2, \cdots, x_k) \log \frac{p(x_1, x_2, \dots, x_k)}{\prod_{i=1}^{k} p(x_i)} \tag{4.6}$$

dalla quale, in analogia alla *pointwise mutual information,* si deriva la correlazione per specifiche istanze delle variabili aleatorie, addivenendo alla correlazione puntuale definita come

$$\mathrm{ptc}(x_1, x_2, \dots, x_k) = \log \frac{p(x_1, x_2, \dots, x_k)}{\prod_{i=1}^{k} p(x_i)} \tag{4.7}$$

e dove si può notare come l'argomento del logaritmo costituisca, in effetti, una generalizzazione del *lift* definito per le regole d'associazione nella (4.5).

Sempre in analogia con la *pointwise mutual information* si fa derivare [10] ancora la misura normalizzata entropicamente

$$\mathrm{hptc}(x_1, x_2, \dots, x_k) = \frac{\mathrm{ptc}(x_1, x_2, \dots, x_k)}{I(x_1, x_2, \dots, x_k)} = \frac{\mathrm{ptc}(x_1, x_2, \dots, x_k)}{-\log p(x_1, x_2, \dots, x_k)} \tag{4.8}$$

dove $I(x_1, x_2, \dots, x_k)$ è la autoinformazione della probabilità congiunta. I pregi di quest'ultima misura consistono nel fatto che essa possiede la caratteristica di mitigare la dipendenza dalla cardinalità e dal supporto minimo del *dataset*, rendendola particolarmente appropriata per le attività di confronto di dati provenienti da partizionamenti disomogenei, come si vedrà nei successivi paragrafi. Essa è inoltre neutrale rispetto alla base scelta per il logaritmo.

## 4.4 Applicabilità delle misure alle generalizzazioni degli itemset

L'estensione della nozione di *itemset* a quella di *itemset* generalizzato e, soprattutto, a quella di *EGI* introduce alcune questioni che debbono essere discusse prima di un'applicazione indiscriminata delle formule introdotte nel § 4.3.

In primo luogo, la divergenza (4.6) richiede almeno che $p(x_1, x_2, \dots, x_k)$ sia una distribuzione di probabilità, ovvero che

$$\sum_{x_1 \in X_1} \sum_{x_2 \in X_2} \cdots \sum_{x_k \in X_k} p(x_1, x_2, \cdots, x_k) = 1 \tag{4.9}$$

e questo è vero, per esempio, se ciascuna $X_1, X_2, \dots, X_k$ è definita su tutti e solo gli elementi appartenenti a un medesimo livello del corrispondente albero $GT_i$ nella tassonomia $\Gamma$ oppure, qualora sia $GT_i = \emptyset$, se è definita su tutti e solo i valori appartenenti al corrispondente spazio campionario $\Omega_i$,

ma in generale non è vero negli altri casi. Quindi l'applicabilità a livello globale della divergenza sembrerebbe esclusa per i dati sperimentali in oggetto, che possono contenere *itemset* definiti su livelli plurimi delle gerarchie. Si noti tuttavia che tale condizione globale può essere lecitamente rimossa a livello puntuale e quindi per la (4.7) e la (4.8), in quanto esse sono implicitamente definite sul medesimo spazio di probabilità per effetto della corrispondenza dei valori assunti dalle variabili indipendenti.

In secondo luogo, occorre stabilire come un *EGI* $X \wr S$ (ovvero costituito da due parti $X$ e $S$) vada correttamente ricondotto a una forma standardizzata nelle formulazioni in cui compaiono sia l'*EGI* stesso sia i singoli elementi che lo costituiscono. Ci si chiede cioè come vada applicata una formula $\hat{p}(\xi_1, \xi_2)/\hat{p}(\xi_i)$ in presenza di un generico *EGI* $X \wr S$. Se è evidente che deve essere $\hat{p}(\xi_1, \xi_2) = \hat{p}([\![X \wr S]\!])$ non è invece ben chiaro e definito cosa si intenda per le singole probabilità $\hat{p}(\xi_i)$ — e in generale cosa rappresentino gli $\xi_i$ — in presenza di elementi contenuti in $S$.

**Teorema 1** Sia $X \wr S$ un EGI di lunghezza $k$, dove $X = \{x_1, x_2, \dots, x_k\}$ e $S = \{s_1, s_2, \dots, s_n\}$ e sia $C = \bigcup_{i=1}^{n} [\![s_i]\!]$ l'insieme unione degli eventi associati a $S$. Allora la probabilità congiunta associata a $\hat{p}([\![X \wr S]\!])$ si può sempre scrivere come

$$\hat{p}([\![X \wr S]\!]) = \hat{p}(\xi_1, \xi_2, \dots, \xi_k) = \hat{p}([\![x_1]\!] \backslash C, [\![x_2]\!] \backslash C, \dots, [\![x_k]\!] \backslash C) \tag{4.10}$$

*Dim.* Dalla definizione 9 si ha che se un evento elementare $\omega$ è associato a $X$ ma nello stesso tempo anche ad almeno uno degli elementi di $S$ allora esso non è associato a $X \wr S$, ovvero vale l'implicazione: $\forall \omega \in \Omega, (\omega \in [\![X]\!] \wedge \omega \in C) \Longrightarrow \omega \notin [\![X \wr S]\!]$. Questo equivale a dire che $[\![X \wr S]\!] = [\![X]\!] \backslash C$ e quindi essendo, per definizione stessa di *itemset*, $[\![X]\!] = [\![x_1]\!] \cap [\![x_2]\!] \cap \dots \cap [\![x_k]\!]$, ne consegue necessariamente che $[\![X \wr S]\!] = ([\![x_1]\!] \cap [\![x_2]\!] \cap \dots \cap [\![x_k]\!]) \setminus C$ e che quindi, per le proprietà della sottrazione, si possa scrivere $[\![X \wr S]\!] = [[\![x_1]\!] \backslash (C \cap [\![X]\!])] \cap [[\![x_2]\!] \backslash (C \cap [\![X]\!])] \cap \dots \cap [[\![x_k]\!] \backslash (C \cap [\![X]\!])]$. Ma dalla definizione 5 si sa che $S$ contiene solo discendenti di $X$ per cui $C \subset [\![X]\!]$ e quindi necessariamente $C \cap [\![X]\!] = C$. Allora $[\![X \wr S]\!] = ([\![x_1]\!] \backslash C) \cap ([\![x_2]\!] \backslash C) \cap \dots \cap ([\![x_k]\!] \backslash C)$, da cui segue la tesi *c.v.d.*

Dal teorema 1 scaturisce pertanto una fondata interpretazione probabilistica di $\hat{p}(\xi_i)$ per la quale, ove fosse $X \wr S =$ `{(città, Liguria), (merce, liquori)}≀{{(città, Genova), (merce, vodka)}}`, allora $\hat{p}(\xi_1)$ sarebbe null'altro che la probabilità osservata di una vendita in Liguria, escluse le vendite a Genova di vodka e $\hat{p}(\xi_2)$ la probabilità osservata di una vendita di liquori, escluse le vendite di vodka a Genova.

## 4.5 Relazioni tra la misura di Kulczynsky e la confidenza

Le misure *null transaction invariant* introdotte per gli *itemset*, ancorché siano tutte basate sul concetto di confidenza delle regole generabili, non costituiscono una semplice trasposizione di tale concetto e della sua applicazione dalle regole agli *itemset*. Se si considerano infatti le misure mediate, cercando di valutarne il potere discriminante, si palesano significative differenze. La confidenza, infatti, operando sull'elemento ultimo di interesse (ovvero la regola), permette la fissazione una volta per tutte[10] d'una soglia $\gamma_0$ discrezionale per discriminare tra regole significative e non

[10] È appena il caso di ricordare che questo uso della confidenza nasce da esigenze di complessità computazionale e non tiene conto delle probabilità a priori, ignorando volutamente che il punto di indipendenza

significative. Invece, la fissazione di una soglia $\vartheta_0$ per una generica misura mediata ha una corrispondenza più rilassata con la confidenza delle regole generabili dagli *itemset* che si situano al di sopra (o al di sotto) di essa, come ovvia conseguenza dell'utilizzo un indice di tendenza centrale senza tener conto della dispersione.

Tra le misure mediate, quella che offre la maggiore semplicità di analisi, interpretazione e aderenza ai modelli probabilistici è la misura di Kulczynsky [7] e pertanto soprattutto di essa verrà fatto uso nella successiva analisi e in tutto il presente lavoro. La rilassatezza della misura di Kulczynsky rispetto alla confidenza si può verificare facilmente. Infatti, fissando una soglia $\vartheta_0 = ½$ e considerando due *itemset* (generalizzati) $X_1 = \{x_1, y_1\}$ e $X_2 = \{x_2, y_2\}$, con valori di supporto opportunamente scelti[11], può ben accadere che si abbia:

a) $\text{kulc}(X_1) = \vartheta_a = \mu\{\text{conf}(x_1 \to y_1);\ \text{conf}(y_1 \to x_1)\} = \ \mu\{0{,}61;\ 0{,}37\} = 0{,}49 \ < \vartheta_0$
b) $\text{kulc}(X_2) = \vartheta_b = \mu\{\text{conf}(x_2 \to y_2);\ \text{conf}(y_2 \to x_2)\} = \ \mu\{0{,}55;\ 0{,}47\} = 0{,}51 \ > \vartheta_0$

ovvero, in entrambi i casi: (i) che solo una delle confidenze sia superiore alla soglia e (ii) che, inoltre, la confidenza d'una regola estraibile da un *itemset* sotto la soglia $\vartheta_0$ (caso a) sia superiore alla più grande tra quelle estraibili da un *itemset* sopra la soglia $\vartheta_0$ (caso b). A ben guardare, si comprende che (i) e (ii) sono dovuti al fatto che è sì vero che $\vartheta_b > \vartheta_0 > \vartheta_a$ ma, d'altra parte, $0 < \sigma_b^2 < \sigma_a^2$ e quindi, a meno di non poter imporre delle restrizioni sulla varianza, non parrebbe lecito utilizzare la relazione d'ordine $\succcurlyeq_\vartheta$ come fosse una mera trasposizione agli *itemset* della relazione $\succcurlyeq_\gamma$. Nondimeno una dipendenza tra le due esiste e può essere indagata nel modo seguente.

**Osservazione 1** *(Intervalli di generabilità)* Si osservi innanzitutto che per ogni generico $k$-*itemset* (generalizzato) $X$ o espressivo $X \wr S$ esistono e gli sono sempre associabili una probabilità congiunta $s = \hat{p}(\xi_1,\ \xi_2, \dots, \xi_k)$ e un insieme di probabilità elementari $\{\hat{p}(\xi_1), \hat{p}(\xi_2), \cdots, \hat{p}(\xi_k)\}$, dalle quali, per combinazione, è generabile una ennupla ordinata di confidenze minime[12] $\langle \gamma_1, \gamma_2, \cdots, \gamma_k \rangle$, dove $\gamma_n = \min_i^{(n)}\{\, s/\hat{p}(\xi_i)\}$ e $s \le \gamma_n \le 1$. Se si fissa una soglia di confidenza pari a $\gamma_0$ si può dimostrare che esiste una sola ennupla tale che la sua media $\vartheta = \mu\{\gamma_1, \gamma_2, \cdots, \gamma_k\}$ sia massima senza che alcuno dei suoi elementi sia però maggiore di $\gamma_0$ e che tale ennupla necessariamente deve essere $\langle \gamma_0\ , \cdots, \gamma_0 \rangle$. Infatti, qualunque alterazione di tale ennupla che non comporti diminuzione della media comporta necessariamente l'incremento di almeno uno dei suoi elementi, contraddicendo quindi l'ipotesi che nessuno di essi superi $\gamma_0$. Analogamente, e per gli stessi motivi, la ennupla che massimizza la media avendo al più un solo elemento che supera $\gamma_0$ sarà necessariamente costituita da $k-1$ elementi di valore pari a $\gamma_0$ e da un elemento pari a 1. In generale, allora, le ennuple che massimizzano la media avendo al più $n$ elementi che la superano saranno costituite da $k-n$ elementi di valore pari a $\gamma_0$ e da $n$ elementi pari a 1. Ciò permette di definire una successione finita di punti

$$\{\vartheta_n\} = \left\{\frac{(k-n)\gamma_0 + n}{k}\right\}_n, \quad 0 \le n \le k \tag{4.11}$$

e conseguentemente di intervalli $\{(0, \vartheta_0], (\vartheta_0, \vartheta_1], \cdots, (\vartheta_{k-1}, \vartheta_k]\}$ dove, nell'ordine, non è assicurata

---

statistica oltre il quale vi è significatività non è fisso ma è $p(E_y|E_x) = \ p(E_y)$.

[11] Per esempio, $\text{supp}(X_1) = \text{supp}(X_2) = 0.11, \text{supp}(x_1) = 0.18, \text{supp}(y_1) = 0.30, \text{supp}(x_2) = 0.20, \text{supp}(y_2) = 0.23$.

[12] Le misure mediate considerano solo le potenziali regole aventi parte sinistra di lunghezza unitaria e, quindi, per la anti-monotonicità del supporto, formulano una media di minimi

la generabilità di alcuna regola con confidenza maggiore di $\gamma_0$, ne è assicurata almeno 1, ne sono assicurate almeno 2, *et cetera*. La successione di punti nella (4.11) individua degli intervalli dove, per ciascuno, è definito il minimo numero di elementi delle ennuple superiori alla soglia fissata. È tuttavia necessario, per completezza di analisi, individuare una analoga successione di intervalli che definiscano ciascuno il numero massimo di elementi ivi possibili.

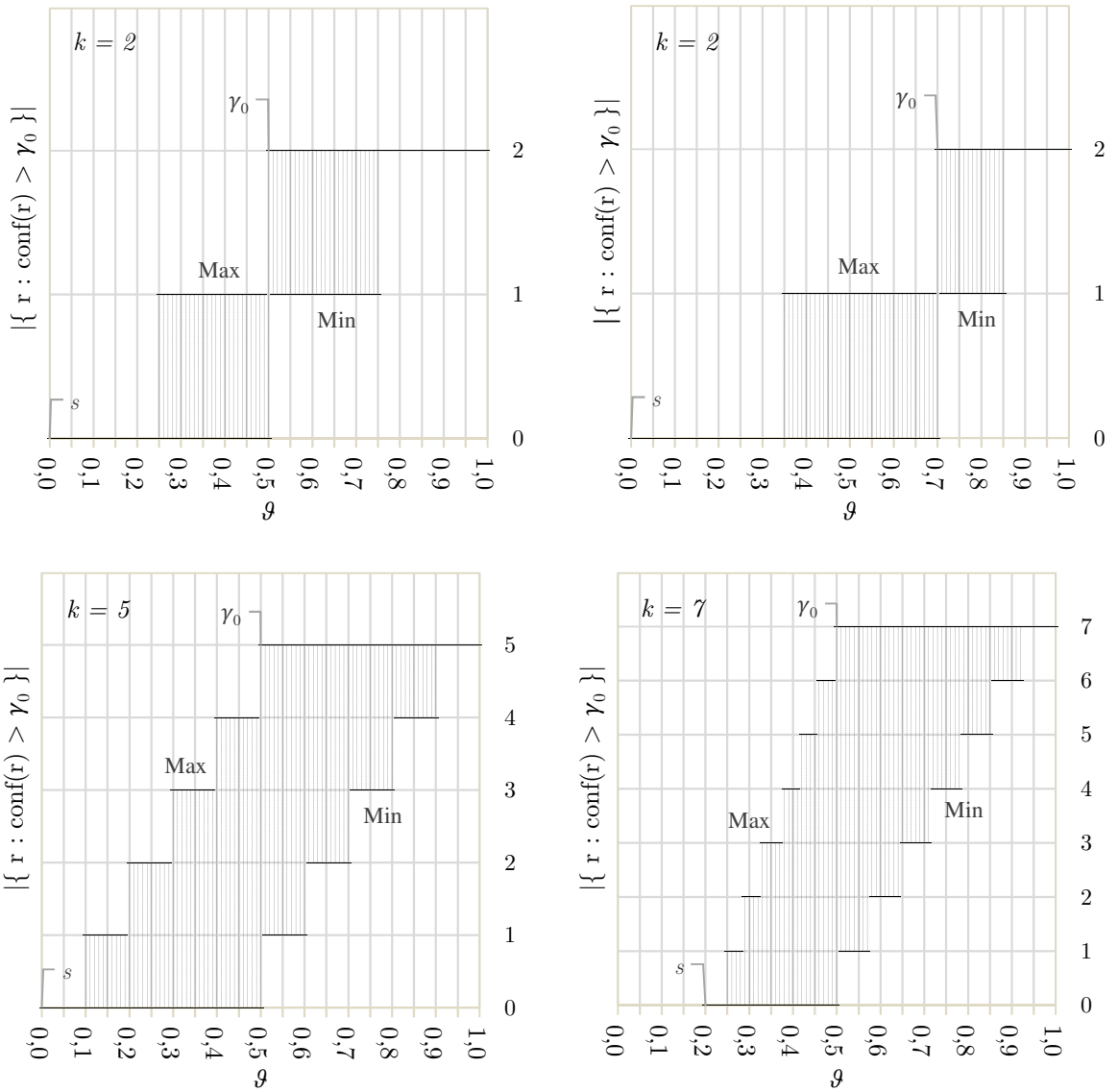



**Figura 10** — *Regole a confidenza minima generabili in funzione di* $\boldsymbol{\vartheta}$*, per alcune combinazioni di parametri* $\boldsymbol{k, s}$ *e* $\boldsymbol{\gamma_0}$

Senza ripetere l'analisi già svolta, si può facilmente mostrare che per valori di $\vartheta$ superiori a $\gamma_0$ esiste sempre almeno una ennupla che ha tutti gli elementi superiori a $\gamma_0$ (e per convincersene basta prendere le ennuple di forma $\langle \vartheta, \cdots, \vartheta \rangle$), mentre nell'intervallo $[s, \gamma_0]$, sempreché sia $s < \gamma_0$, sarà sufficiente considerare le ennuple di forma $\langle s, \cdots, s \rangle, \langle s, \cdots, s, \gamma_0 \rangle, \cdots, \langle s, \gamma_0, \cdots, \gamma_0 \rangle, \langle \gamma_0, \cdots, \gamma_0 \rangle$.

Come nel caso precedente, ciò conduce, quando l'intervallo esiste, alla definizione di una successione finita di punti

$$\{\tau_n\} = \left\{ \frac{(k-n)s + n\gamma_0}{k} \right\}_n \qquad 0 \le n \le k, \;\; s < \gamma_0 \tag{4.12}$$

e conseguentemente di intervalli $\{\,[\tau_0,\tau_1],(\tau_1,\tau_2],\cdots,(\tau_{k-1},\tau_k],(\tau_k,1]\,\}$ dove, nell'ordine, non è possibile generare alcuna regola con confidenza maggiore di $\gamma_0$, se ne può generare al più 1, se ne possono generare al più 2, *et cetera*.

Dall'esame combinato di questi *upper* e *lower bound*, la conclusione che si trae è che la relazione d'ordine $\succcurlyeq_\vartheta$, rispetto a una arbitrariamente fissata soglia di confidenza $\gamma_0$, ha valenza discriminante e deterministica solo nei due intervalli estremi $[\tau_0,\tau_1]$ e $(\vartheta_{k-1},\vartheta_k]$, mentre nell'intervallo $(\tau_1,\vartheta_{k-1}]$ ha una mera funzione indicatrice (con incertezza massima in $\vartheta=\gamma_0$ e nulla negli estremi) della completezza delle regole generabili con riguardo a $\gamma_0$.

## 4.6 Relazioni tra la misura di Kulczynsky e il cross-support ratio

La comparazione della misura di Kulczynsky con la confidenza è in un certo qual modo naturale, ma è lecito chiedersi se tale misura non possa essere messa in relazione anche con altri indici o misure utilizzate per la valutazione degli *itemset*. Uno di questi indici, a prima vista non direttamente collegato con le misure di questo capitolo, è il cosiddetto *cross-support ratio* che è definito in letteratura [12] come il rapporto

$$\chi=\frac{\min_{i=1}^{k}\hat{p}([\![x_i]\!])}{\max_{j=1}^{k}\hat{p}([\![x_j]\!])},\qquad x_i,x_j\in X=\{x_1,x_2,\cdots,x_k\} \tag{4.13}$$

e che indica, con valori vieppiù piccoli, la presenza, nel medesimo *itemset*, di *item* molto rari congiunti a *item* molto frequenti. Questo tipo di *pattern* caratterizza gli *itemset* che, a dispetto dell'alta confidenza di una o più delle regole estraibili, sono spesso privi di interesse. Sono infatti i casi dove si ha, a un tempo, che $\exists\, x_i, x_j\in X:\ \hat{p}([\![X]\!])\approx\hat{p}([\![x_i]\!])\ll\hat{p}([\![x_j]\!])\approx 1$, il che conduce all'estrazione di regole che hanno più o meno il medesimo valore inferenziale di «*chi trova un quadrifoglio → vuol bene alla mamma*».

Ponendo $\hat{p}(\xi_i)=\hat{p}([\![x_i]\!])\ \forall\, x_i\in X$ e considerando che per gli *itemset* espressivi le probabilità $\hat{p}(\xi_i)$ possono essere ricavate semplicemente applicando il teorema 1, se si moltiplicano nel rapporto (4.13) sia il numeratore sia il denominatore per la probabilità congiunta, che è sempre strettamente positiva, si ricava che

$$\begin{aligned}\chi\ &=\frac{\min_{i=1}^{k}\hat{p}(\xi_i)}{\max_{j=1}^{k}\hat{p}(\xi_j)}\cdot\frac{\hat{p}(\xi_1,\,\xi_2,\ldots,\xi_k)}{\hat{p}(\xi_1,\,\xi_2,\ldots,\xi_k)}=\frac{\hat{p}(\xi_1,\,\xi_2,\ldots,\xi_k)}{\max_{j=1}^{k}\hat{p}(\xi_j)}\cdot\frac{\min_{i=1}^{k}\hat{p}(\xi_i)}{\hat{p}(\xi_1,\,\xi_2,\ldots,\xi_k)}=\\ &=\frac{\dfrac{\hat{p}(\xi_1,\,\xi_2,\ldots,\xi_k)}{\max_{j=1}^{k}\hat{p}(\xi_j)}}{\dfrac{\hat{p}(\xi_1,\,\xi_2,\ldots,\xi_k)}{\min_{i=1}^{k}\hat{p}(\xi_i)}}=\frac{\min_{j=1}^{k}\dfrac{\hat{p}(\xi_1,\,\xi_2,\ldots,\xi_k)}{\hat{p}(\xi_j)}}{\max_{i=1}^{k}\dfrac{\hat{p}(\xi_1,\,\xi_2,\ldots,\xi_k)}{\hat{p}(\xi_i)}}\end{aligned} \tag{4.14}$$

ovvero che il *cross-support ratio* è un rapporto tra probabilità condizionate e quindi, estensivamente, un rapporto tra confidenze. Si scorge pertanto la comune base con la misura di Kulczynsky e la liceità dell'indagine sulle connessioni tra le due misure.

**Osservazione 2** *(Valenza del cross-support ratio come indicatore di dispersione)* Intuitivamente, sulla base della definizione stessa di $\chi$, si può congetturare che il valore del rapporto tra l'elemento minimo e l'elemento massimo di un insieme di valori sia idoneo a rappresentarne, in qualche modo, la dispersione. Invero si può mostrare, con un certo grado di rigore, come esso, almeno nella sua

applicazione agli *itemset*, conduca direttamente a un approssimante della varianza. Per verificarlo, si cominci col considerare una ennupla ordinata crescente di $k$ valori reali positivi $\langle \gamma_1, \gamma_2, \cdots, \gamma_k \rangle$ tale da rappresentare le $k$ confidenze minime ricavabili da un generico *itemset* (generalizzato (espressivo)) come nell'osservazione 1. Ci si ponga quindi il problema di approssimarne la varianza conoscendo solo i valori dei due estremi $\gamma_1, \gamma_k$, che, per comodità, saranno nel seguito riferiti, rispettivamente, come $m, M$. Un modo per maggiorare tale approssimazione è costituito dal prendere in considerazione le ennuple dove gli elementi, noti e ignoti, abbiano tutti distanza massima da un comune baricentro, secondo la definizione stessa di varianza. Si può dimostrare che tali ennuple sono necessariamente le seguenti:

$$\langle \underbrace{m, \cdots, m}_{\lceil k/2 \rceil}, \underbrace{M, \cdots, M}_{\lfloor k/2 \rfloor} \rangle, \langle \underbrace{m, \cdots, m}_{\lfloor k/2 \rfloor}, \underbrace{M, \cdots, M}_{\lceil k/2 \rceil} \rangle \tag{4.15}$$

che naturalmente coincidono quando $k$ è pari. Per tali ennuple il calcolo della varianza è particolarmente semplice e conduce alla seguente espressione:

$$\max \sigma^2 = \begin{cases} \left(\frac{M-m}{2}\right)^2; & \text{per } k \text{ pari e } \hat{\mu} = \frac{M+m}{2} \\ \left(\frac{M-m}{2}\right)^2 \left(1 - \frac{1}{k^2}\right); & \text{per } k \text{ dispari e } \hat{\mu} = \frac{\lfloor k/2 \rfloor M + \lceil k/2 \rceil m}{k} \end{cases} \tag{4.16}$$

dove con $\hat{\mu}$ è indicato il baricentro utilizzato. Se di un *itemset* (generico (espressivo)) $X$ o $X \wr S$ è noto sia il valore della misura di Kulczynsky $\vartheta$, che per definizione è la media delle confidenze, sia il valore del *cross-support ratio* $\chi$, che per la (4.14) è il rapporto $m/M$ tra confidenza minima e massima, allora, a meno dell'errore commesso considerando il baricentro $\hat{\mu}$ approssimativamente uguale alla media effettiva $\vartheta$, si ricava che l'espressione

$$\vartheta^2 \left(\frac{\chi - 1}{\chi + 1}\right)^2 = \left(\vartheta \cdot \frac{2}{m+M} \cdot \frac{m-M}{2}\right)^2 = \left(\frac{\vartheta}{\hat{\mu}} \cdot \frac{m-M}{2}\right)^2 \approx \left(\frac{m-M}{2}\right)^2 = \max \sigma^2 \tag{4.17}$$

costituisce un approssimante per eccesso, seppur distorto da $\left(\vartheta/\hat{\mu}\right)^2$ per $k > 2$, della varianza.

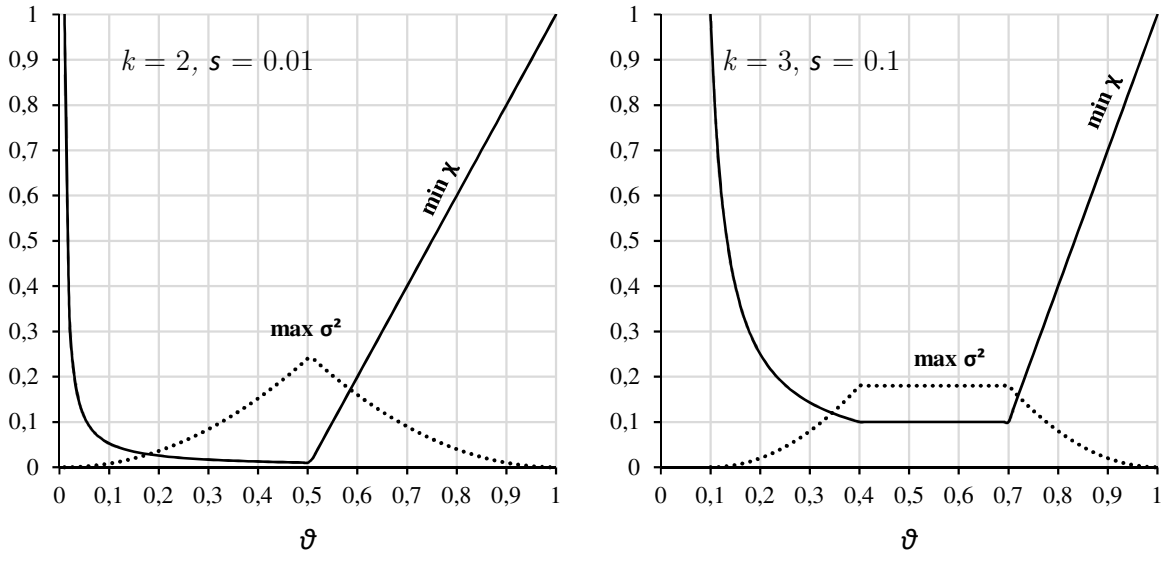



**Figura 11** — *Dipendenza del* min–$\chi$–ratio *e della massima varianza da* $\vartheta$ *e da* $k$

L'espressione (4.17) potrebbe ingenerare l'impressione che $\vartheta$ e $\chi$ possano essere fatti variare indipendentemente, così come se ci si proponesse di costruire un sottoinsieme di $\mathbb{R}$ con $\mu$ e $\sigma^2$ arbitrariamente prefissati. Tale impressione è fallace e ciò può essere dimostrato per assurdo scegliendo $\vartheta = 1$, un qualsiasi valore $\chi \neq 1$ e mostrando che non può esistere alcuna ennupla di confidenze soddisfacente i valori scelti, mentre tale ennupla esiste (ed è unica) per $\vartheta = \chi = 1$.

Per indagare allora l'interdipendenza tra *cross-support ratio* e $\vartheta$, si consideri dapprima la ennupla ordinata di forma $\langle s, 1, \cdots, 1\rangle$, dove $s = \hat{p}(\xi_1, \xi_2, \ldots, \xi_k)$. Tale ennupla ha la caratteristica di avere $\chi_0 = s$, $\vartheta_0 = \mu = \frac{k-1+s}{k}$ e $\sigma_0^2 = (k-1)(\vartheta_0 - 1)^2$ e si può dimostrare che ivi il $\chi$ è minimo e la $\sigma^2$ è massima rispetto a tutte le possibili alterazioni della ennupla che ne lascino invariata la media $\vartheta_0$ e che inoltre $\chi_0$ è anche minimo assoluto. Se ora si vuole alterare tale ennupla in modo da incrementarne il valore della media di una quantità $0 < \Delta\vartheta \leq \frac{1-s}{k}$, così che sia $\vartheta = \vartheta_0 + \Delta\vartheta$, allora non vi è altro modo di farlo se non incrementandone il primo elemento di una quantità $k\Delta\vartheta$, sicché la nuova ennupla diviene $\langle s + k\Delta\vartheta, 1, \cdots, 1\rangle$ ed essa, come la sua originante, ha ancora $\chi$ minimo e $\sigma^2$ massima rispetto a tutte le possibili alterazioni di essa che siano invarianti per la media, ma, soprattutto, per questa ennupla $\chi$ e $\sigma^2$ sono funzionalmente dipendenti da $\vartheta$, talché i valori di interesse sono $\chi = k(\vartheta - 1) + 1$ e $\sigma^2 = (k-1)(\vartheta - 1)^2$. Esiste pertanto un valore limite (rappresentato da $\vartheta_0$) oltre il quale sussiste necessariamente la dipendenza funzionale

$$\begin{cases} \min \chi = k(\vartheta - 1) + 1 \\ \max \sigma^2 = (k-1)(\vartheta - 1)^2 \end{cases} \qquad \forall\, \vartheta \geq \frac{(k-1)+s}{k} \tag{4.18}$$

Analogamente, se si considera la ennupla di forma $\langle s, \cdots, s, 1\rangle$, si può dimostrare, come nel caso precedente, che ivi $\chi$ è minimo e $\sigma^2$ è massima rispetto a tutte le possibili alterazioni della ennupla che ne lascino invariata la media. Ripetendo le considerazioni già fatte e applicandole però all'alterazione della ennupla finalizzato a un arbitrario decremento $0 < \Delta\vartheta \leq \frac{1-s}{k}$, si giunge facilmente a trovare anche in questo caso il valore limite al di sotto del quale sussiste

$$\begin{cases} \min \chi = \dfrac{s}{k(\vartheta - s) + s} \\ \max \sigma^2 = (k-1)(\vartheta - s)^2 \end{cases} \qquad \forall\, \vartheta \leq \frac{(k-1)s + 1}{k} \tag{4.19}$$

È allora possibile caratterizzare completamente la dipendenza nel minimo del *cross-support ratio* riunendo le (4.18) e (4.19), ottenendo

$$\min \chi = \begin{cases} \dfrac{s}{k(\vartheta - s) + s}; & \text{se } \vartheta \leq \dfrac{(k-1)s + 1}{k} \\ s & \text{se } \dfrac{(k-1)s + 1}{k} < \vartheta < \dfrac{(k-1)+s}{k} \\ k(\vartheta - 1) + 1; & \text{se } \vartheta \geq \dfrac{(k-1)+s}{k} \end{cases} \tag{4.20}$$

La (4.20) è interessante nella misura in cui rivela che il valore del *min–$\chi$–ratio* non è influenzato in modo simmetrico da $\vartheta$, come la varianza, ma che a partire dall'estremo sinistro, ovvero $s$, esso attraversa un tratto iperbolico strettamente decrescente, quindi un tratto costante (se $k > 2$) e, infine, un tratto lineare strettamente crescente, come si può vedere in figura 11.

## 4.7 Relazioni tra la misura di Kulczynsky e il supporto

Nell'analisi condotta nei paragrafi precedenti si è avuto modo di osservare che la probabilità congiunta o supporto $s = \hat{p}(\xi_1, \xi_2, \dots, \xi_k)$ dei generici *itemset* (generalizzati (espressivi)) presi in considerazione veniva a essere sovente un parametro delle espressioni ivi considerate. Sembra lecito allora chiedersi in che modo il valore di $s$ influenzi la misura di Kulczynsky e, nel modo più generale possibile, se e come si possa valutare il valore atteso di quest'ultima con riguardo al solo valore del primo.

Si osservi innanzitutto che la misura di Kulczynsky può essere scritta come il rapporto tra la probabilità congiunta $s$ e la media armonica delle singole probabilità costituenti. Infatti,

$$\vartheta = \frac{1}{k}\sum_{i=1}^{k}\frac{\hat{p}(\xi_1, \xi_2, \dots, \xi_k)}{\hat{p}(\xi_i)} = \hat{p}(\xi_1, \xi_2, \dots, \xi_k)\cdot\frac{1}{k}\sum_{i=1}^{k}\frac{1}{\hat{p}(\xi_i)} = \frac{\hat{p}(\xi_1, \xi_2, \dots, \xi_k)}{\mu_H\{\hat{p}(\xi_1), \hat{p}(\xi_2), \dots, \hat{p}(\xi_k)\}} \tag{4.21}$$

e quindi, essendo le $\hat{p}(\xi_i) \leq 1$, per il principio di internalità di Cauchy anche la loro media $\mu_H$ sarà necessariamente tale e quindi sarà anche $\vartheta \geq \hat{p}(\xi_1, \xi_2, \dots, \xi_k)$. Sin qui le conclusioni di ordine deterministico che possono farsi su $\vartheta$ con riguardo a $s$, ma se si vuole uscire dal determinismo e calcolarne il valore atteso $\mathbb{E}[\Theta|S = s]$ sarà necessario trovare prima il valore atteso di $\mu_H$ condizionatamente a $s$ e, ancor prima, formulare un'ipotesi sulla distribuzione delle $\hat{p}(\xi_i)$, a partire dalla considerazione che ciascuna $\hat{p}(\xi_i) \geq s$ per la anti-monotonicità di $s$.

**Osservazione 3** (*Funzione di densità di probabilità pel supporto d'un generico 1-itemset*) Sia $\mathfrak{X}$ una popolazione di *itemset* e $X \in \mathfrak{X}$ un generico *itemset* di lunghezza $k = 1$ ad essa appartenente; sia anche $\varepsilon$ il supporto minimo degli *itemset* in $\mathfrak{X}$ e $n$ la loro lunghezza massima. Se consideriamo la probabilità $p([\![X]\!]) \geq \varepsilon$ del verificarsi dell'evento associato, allora è chiaro che in $\mathfrak{X}$ potranno *al più* verificarsi $\binom{n}{k}\frac{1}{p([\![X]\!])}$ eventi indipendenti e distinti aventi la medesima esatta probabilità, per cui, in assenza di ogni altra condizione o ipotesi, è ragionevole ipotizzare che gli *itemset* di lunghezza unitaria si distribuiscano in $\mathfrak{X}$ secondo il loro supporto $x = p([\![X]\!])$ con una funzione di densità continua $\varphi(x) \sim P\left(\frac{1}{x}\right)$. Si noti che, essendo l'intervallo di interesse limitato a $[\varepsilon, 1]$, $\varphi$ è ivi localmente sommabile ovunque e non v'è mai necessità di valutarne il valore principale.

Volendo allora calcolare il valore atteso condizionato di $\vartheta$ dato $s$ si potrà finalmente scrivere[13], sulla base della (4.21) e delle osservazioni precedenti,

$$\mathbb{E}[\Theta|S = s] = s\cdot\left(\int_s^1 \frac{1}{x}\varphi(x)dx\right) = s\cdot\left(\frac{1}{-\ln s}\int_s^1 \frac{1}{x^2}dx\right) = \frac{s-1}{\ln s} \tag{4.22}$$

tenendo in considerazione che, agli estremi, sono calcolabili e valgono i seguenti limiti:

$$\lim_{s\to 0}\frac{s-1}{\ln s} = 0 \tag{4.23}$$

e

$$\lim_{s\to 1}\frac{s-1}{\ln s} = 1 \tag{4.24}$$

[13] Si sono omesse le semplificazioni degli integrali di convoluzione, a rigore implicati nelle $k$ somme.

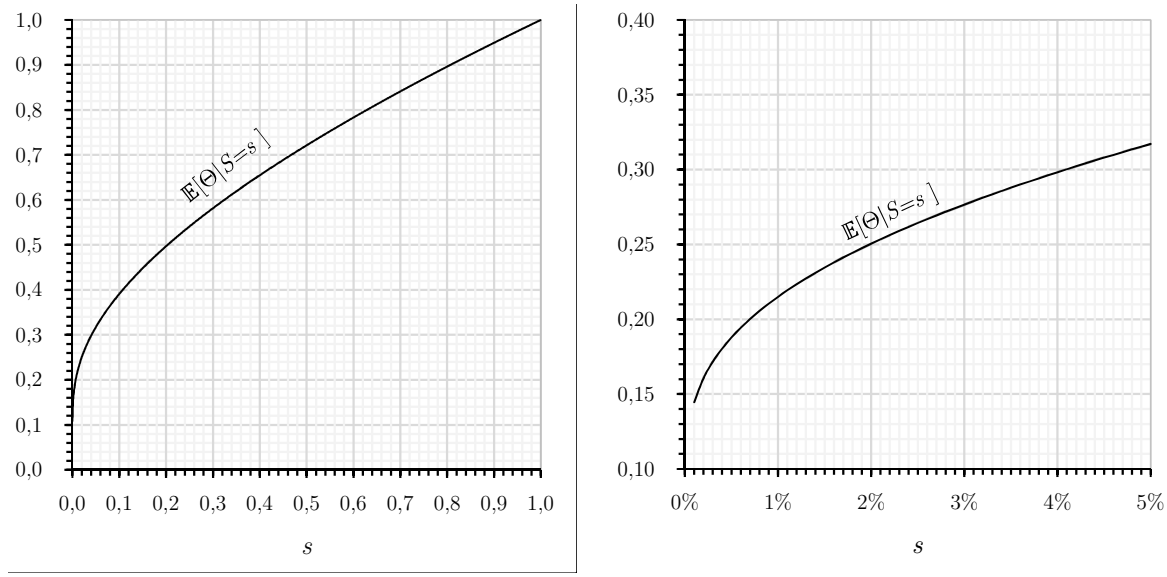



**Figura 12** — *Valore atteso condizionato da s della misura di Kulczynsky, in generale (a sinistra) e limitatamente all'insieme dei valori ove si situano i supporti minimi dei dati sperimentali (a destra)*

## 4.8 Interpretazione e limiti delle misure basate sulla confidenza

Dall'analisi condotta nei paragrafi precedenti si è dato prova di come la misura di Kulczynsky costituisca una efficace funzione indicatrice della suscettibilità di un *itemset* di generare (o di non generare) delle regole di interesse con riguardo sia a una ipotetica soglia di confidenza $\gamma_0$ arbitrariamente fissata sia al *cross-support ratio* $\chi$. In entrambi i casi la misura aumenta di precisione predittiva quanto più il suo valore si avvicina agli estremi del suo intervallo di definizione $[s, 1]$ o, equivalentemente, essa rivela la massima incertezza quanto più si avvicina un un suo valore centrale determinato nei due casi, rispettivamente e indipendentemente, da $\gamma_0$ e da quanto vale $s$.

Tuttavia, nonostante quanto testé osservato, vi sono dei limiti di natura logica e probabilistica che debbono essere considerati.

Il primo limite è di natura interpretativa ed è strettamente legato al confronto tra *itemset* di diversa lunghezza. Se si confrontano, per esempio, due *itemset* $X$ e $Y$ tali che per essi si abbia che $\vartheta_X = \vartheta_Y = 0{,}71, k_X = 2$ e $k_Y = 5$, basta consultare la figura 10 per rendersi conto che il numero di regole utili generabili da $Y$ sopravanza di molto quelle generabili da $X$ in termini assoluti: da un minimo di tre fino a un massimo di cinque nel primo caso, contro un minimo di una a un massimo di due nel secondo, senza tener conto che, per $k > 2$, quindi per $Y$, si tratterebbe in realtà non di regole singole ma di interi sottoalberi di regole. È chiaro che sotto questa prospettiva, *itemset* di diversa lunghezza diverrebbero inconfrontabili e qualunque pretesa di ordinamento verrebbe meno in presenza di insiemi costituenti delle *mixture.*

Per superare tale limite, è purtroppo necessario rinunciare all'interezza del contenuto informativo portato dalla misura di Kulczynsky e ricondursi a una interpretazione minimalista, che minimizzi tuttavia la dipendenza da $k$, consentendo la confrontabilità e quindi l'ordinamento. Sia dunque $\mathfrak{X}$ una popolazione di *itemset* (generalizzati (espressivi)) (ovvero formata, a un tempo, da *itemset* $X = X \wr S, S = \emptyset$ e *itemset* $X \wr S, S \neq \emptyset$) e $\gamma_0$ una arbitraria soglia di confidenza. Allora, $\forall\, X \wr S$, posto $R_{X \wr S} = \{r \mid (X \wr S) \xrightarrow{\text{genera}} r\}$, si potrà scrivere:

$$\begin{cases} p(\exists\, r \in R_{X\wr C} : \text{conf}(r) > \gamma_0) = 0 & \text{se } \vartheta \leq \tau_1 = \dfrac{(k-1)s + \gamma_0}{k} \\ p(\exists\, r \in R_{X\wr C} : \text{conf}(r) > \gamma_0) = 1 & \text{se } \vartheta > \gamma_0 \\ p(\exists\, r \in R_{X\wr C} : \text{conf}(r) > \gamma_0) = \varphi(\vartheta, s, \gamma_0, k), & \text{altrimenti} \end{cases} \tag{4.25}$$

dove $\varphi(\vartheta, s, \gamma_0, k)$ è una funzione strettamente crescente a valori in $(0,1)$ per la quale valgono i limiti $\lim_{\vartheta \to \tau_1} \varphi(\vartheta, s, \gamma_0, k) = 0$ e $\lim_{\vartheta \to \gamma_0} \varphi(\vartheta, s, \gamma_0, k) = 1$, $\forall\, \gamma_0, k, s < \gamma_0$ e che, valutando la sola probabilità a priori, è ben approssimata dalla seguente espressione:

$$\varphi(\vartheta, s, \gamma, k) \approx \frac{1 - (\gamma_0 - s)^{\frac{k(\vartheta - s)}{\gamma_0 - s} - 1}}{1 - (\gamma_0 - s)^{k-1}} \tag{4.26}$$

La (4.25) è giustificata dalle (4.11) e (4.12), mentre alla (4.26) si perviene attraverso il rapporto dei casi favorevoli rispetto ai possibili di estrarre una regola con confidenza $\gamma > \gamma_0$ in $k - 1$ estrazioni[14]. Si noti che l'espressione analitica di $\varphi(\vartheta, s, \gamma_0, k)$ rappresenta una curva interpolante che congiunge, nella figura 10, i punti $\tau_1$ con i punti $\gamma_0$.

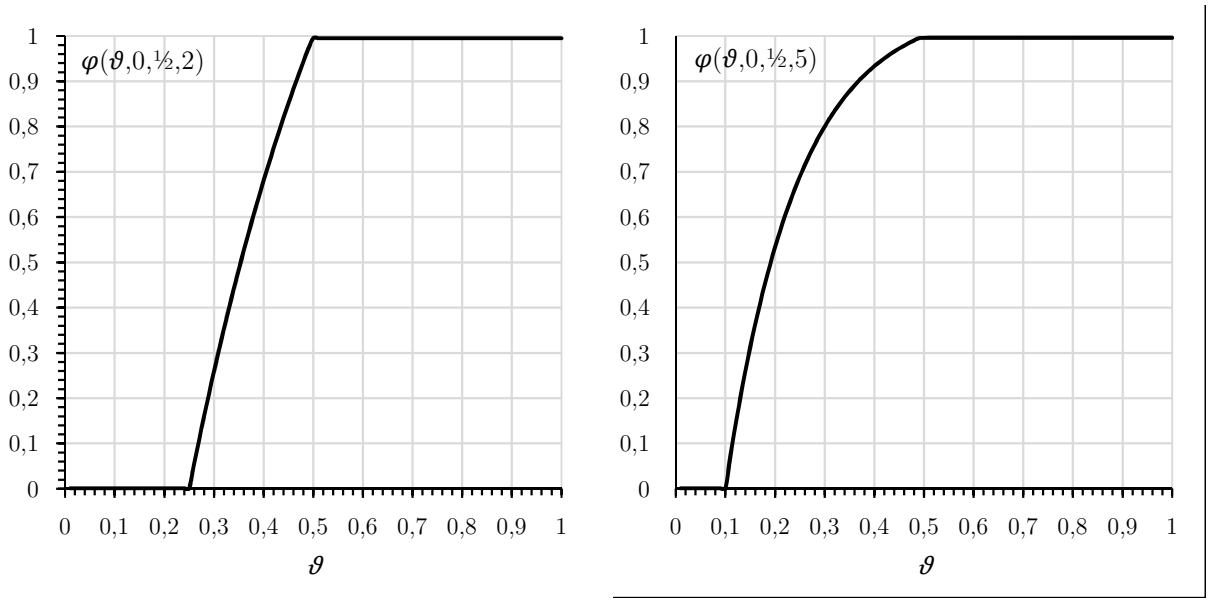



**Figura 13** — *Curve della probabilità a priori di generare almeno una regola con confidenza* $\boldsymbol{\gamma > \gamma_0}$

Il fatto cruciale da porre in rilievo è che quando si pone $\gamma_0 = \vartheta_X$ il campo di variazione viene diviso in quel punto tra una certezza nell'intervallo destro e una probabilità (qualunque sia $\varphi$) decrescente fino a zero nell'intervallo sinistro. Quindi, preso un *itemset* qualunque, tutti quelli che hanno un valore della misura di Kulczynsky a esso inferiore non potranno mai avere probabilità maggiore del primo di generare una regola *con confidenza uguale o maggiore della sua misura di Kulczynsky*, in ragione del fatto che la sua probabilità è ivi la più alta possibile. Ma allora è proprio sotto questa interpretazione che si potrà dunque affermare che $\forall\, X, Y \in \mathfrak{X}, \vartheta_X \geq \vartheta_Y \Longrightarrow X \succcurlyeq Y$ con

[14] Le estrazioni sono $k - 1$ in quanto l'ultima estrazione è deterministicamente imposta dalle precedenti. Il problema equivale alla probabilità di fallire nell'evitare di far traboccare uno qualunque di $k$ vasi di capacità $\gamma_0 - s$ versando in ciascuno di essi una quantità casuale da un recipiente che contiene in tutto $k(\vartheta - s)$ unità di liquido fino ad esaurirlo.

riguardo a un ordinamento basato sulla confidenza, giacché, per la (4.25), sarà possibile confrontare le probabilità e trovare che $p(\exists\, r \in R_X : \mathrm{conf}(r) \geq \vartheta_X\}) \geq p(\exists\, r \in R_Y : \mathrm{conf}(r) \geq \vartheta_X\})$, indipendentemente da $k$ e dall'espressione analitica di $\varphi$, superando così il primo dei limiti precedentemente accennati.

Il secondo limite, ben più significativo, è invece di natura probabilistica, ma non dipende dalla misura di Kulczynsky (o da qualunque altra misura di confidenza mediata) in sé, bensì dalla definizione stessa di confidenza. Il valore infatti che si attribuisce a questa misura è determinato dal fatto che, dati due eventi $E_x, E_y$, se $p(E_y|E_x) \gg ½$, allora sembra ragionevole ritenere, poiché l'evento $E_x$ appare condizionare favorevolmente l'evento $E_y$, che ci si possa aspettare che $E_x \Rightarrow E_y$. Questo ragionamento è ben noto per la sua fallacia, almeno *in parte qua* si applichi senza chiedersi quanto sia $p(E_y)$. Infatti, perché si possa realmente dire che $E_x$ condiziona l'evento $E_y$ nel verso dell'implicazione, non basta osservare il valore, per grande che sia, di $p(E_y|E_x)$, ma anche verificare che sia $p(E_y|E_x) > p(E_y)$. Invero, con qualche manipolazione, è immediato rendersi conto che:

$$p(E_y|E_x) = \frac{p(E_x, E_y)}{p(E_x)} = \frac{p(E_x, E_y)}{p(E_x)p(E_y)} p(E_y) \tag{4.27}$$

e che quindi, trattandosi di quantità tutte strettamente positive,

$$p(E_y|E_x) > p(E_y) \Leftrightarrow \frac{p(E_x, E_y)}{p(E_x)p(E_y)} > 1 \tag{4.28}$$

dove, se i due eventi sono riconducibili all'antecedente e al conseguente d'una regola associativa, si riconosce, nel membro co-implicato di destra, il già citato *lift* della (4.5). Operando in sede di generazione di regole, sarebbe pertanto possibile, ancorché computazionalmente gravoso, rimuovere le regole spurie. Essendo tuttavia interessati in questa sede agli *itemset* ed essendo le (4.27) e (4.28) applicabili direttamente solo alle regole, allora per conseguire il medesimo obiettivo in sede di analisi dei primi, sarà necessario esaminare se e in che misura sia possibile impiegare valutazioni sintetiche su di essi, introducendo le misure appropriate.

## 4.9 Relazioni tra la correlazione totale puntuale e la misura di Kulczynsky

Partendo dalla definizione di misura di Kulczynsky, con qualche manipolazione, è possibile ricavare una espressione che la scompone in due parti di interesse:

$$\vartheta = \frac{1}{k}\sum_{i=1}^{k} \frac{\hat{p}(x_1, x_2, \ldots, x_k)}{\hat{p}(x_i)} = \underbrace{\frac{\hat{p}(x_1, x_2, \ldots, x_k)}{\prod_{i=1}^{k} \hat{p}(x_i)}} \cdot \underbrace{\frac{1}{k}\sum_{i=1}^{k} \left( \prod_{j=1, j\neq i}^{k} \hat{p}(x_j) \right)} \tag{4.29}$$

ovvero un termine che costituisce l'argomento della correlazione totale puntuale che moltiplica una media di prodotti. Ricordando che la misura di Kulczynsky valuta solo le regole che hanno antecedenti di lunghezza unitaria, si può constatare che questi prodotti coinvolgono le probabilità delle variabili destinate a costituire i conseguenti di queste regole. Parrebbe quindi, *prima facie*, di aver ricondotto la (4.27) a un confronto tra medie superando quindi, nel senso della sintesi, il limite evidenziato nel § 4.8, ma disgraziatamente questa interpretazione può dirsi corretta solo quando

$k = 2$, mentre per valori di $k$ superiori essa è viziata da un errore ineliminabile. Per convincersene è sufficiente sviluppare la (4.29) oltre che per $k = 2$ anche per $k = 3$, ottenendo:

$$\frac{1}{k}\sum_{i=1}^{k}\frac{\hat{p}(x_1, x_2, \dots, x_k)}{\hat{p}(x_i)} = \begin{cases} \dfrac{\hat{p}(x_1, x_2)}{\hat{p}(x_1)\hat{p}(x_2)} \cdot \dfrac{\hat{p}(x_2) + \hat{p}(x_1)}{2}; & k = 2 \\ \dfrac{\hat{p}(x_1, x_2, x_3)}{\hat{p}(x_1)\hat{p}(x_2)\hat{p}(x_3)} \cdot \dfrac{\hat{p}(x_2)\hat{p}(x_3) + \hat{p}(x_1)\hat{p}(x_3) + \hat{p}(x_1)\hat{p}(x_2)}{3}; & k = 3 \end{cases} \tag{4.30}$$

Nel primo caso è immediatamente verificato, essendo le possibili regole generabili e valutabili solamente $E_1 \to E_2$ e $E_2 \to E_1$, che la (4.29) è la trasposizione in forma sintetica mediata della (4.27) e che quindi effettivamente il fattore comune a secondo membro assolve per entrambe le regole la sua funzione verificatrice in modo sintetico; nel secondo caso si osserva invece che la media a secondo membro non è calcolata sugli appropriati conseguenti, che dovrebbero essere di forma $\hat{p}(x_i, x_j)$, bensì essa è la loro media solo nel particolare e infrequente caso in cui essi rappresentino congiunzioni di eventi statisticamente indipendenti.

Per approfondire converrà ancora prendere in esame il caso $k = 3$ come punto di partenza. Si cominci col notare che l'argomento della correlazione totale puntuale osservata può essere sempre scritto, per ognuna delle $k$ variabili, come un prodotto $L = \ell_i \eta_i$, ossia, per $k = 3$:

$$L = \frac{\hat{p}(x_1, x_2, x_3)}{\hat{p}(x_1)\hat{p}(x_2)\hat{p}(x_3)} = \begin{cases} \dfrac{\hat{p}(x_1, x_2, x_3)}{\hat{p}(x_2, x_3)\hat{p}(x_1)} \cdot \dfrac{\hat{p}(x_2, x_3)}{\hat{p}(x_2)\hat{p}(x_3)} = \ell_1 \eta_1; & (x_1 \to x_2, x_3) \\ \dfrac{\hat{p}(x_1, x_2, x_3)}{\hat{p}(x_1, x_3)\hat{p}(x_2)} \cdot \dfrac{\hat{p}(x_1, x_3)}{\hat{p}(x_1)\hat{p}(x_3)} = \ell_2 \eta_2; & (x_2 \to x_1, x_3) \\ \dfrac{\hat{p}(x_1, x_2, x_3)}{\hat{p}(x_1, x_2)\hat{p}(x_3)} \cdot \dfrac{\hat{p}(x_1, x_2)}{\hat{p}(x_1)\hat{p}(x_2)} = \ell_3 \eta_3; & (x_3 \to x_1, x_2) \end{cases} \tag{4.31}$$

ed è banale passare da $k$ a $k + 1$, come si può vedere, per esempio, nel caso $k = 4$:

$$\frac{\hat{p}(x_1, x_2, x_3, x_4)}{\hat{p}(x_1)\hat{p}(x_2)\hat{p}(x_3)\hat{p}(x_4)} = \begin{cases} \dfrac{\hat{p}(x_1, x_2, x_3, x_4)}{\hat{p}(x_2, x_3, x_4)\hat{p}(x_1)} \cdot \dfrac{\hat{p}(x_2, x_3, x_4)}{\hat{p}(x_2)\hat{p}(x_3)\hat{p}(x_4)}; & (x_1 \to x_2, x_3, x_4) \\ \dfrac{\hat{p}(x_1, x_2, x_3, x_4)}{\hat{p}(x_1, x_3, x_4)\hat{p}(x_2)} \cdot \dfrac{\hat{p}(x_1, x_3, x_4)}{\hat{p}(x_1)\hat{p}(x_3)\hat{p}(x_4)}; & (x_2 \to x_1, x_3, x_4) \\ \dfrac{\hat{p}(x_1, x_2, x_3, x_4)}{\hat{p}(x_1, x_2, x_4)\hat{p}(x_3)} \cdot \dfrac{\hat{p}(x_1, x_2, x_4)}{\hat{p}(x_1)\hat{p}(x_2)\hat{p}(x_4)}; & (x_3 \to x_1, x_2, x_4) \\ \dfrac{\hat{p}(x_1, x_2, x_3, x_4)}{\hat{p}(x_1, x_2, x_3)\hat{p}(x_4)} \cdot \dfrac{\hat{p}(x_1, x_2, x_3)}{\hat{p}(x_1)\hat{p}(x_2)\hat{p}(x_3)}; & (x_4 \to x_1, x_2, x_3) \end{cases} \tag{4.32}$$

Se si sviluppa il prodotto a secondo membro della (4.30) e ivi si sostituisce il fattore comune con l'appropriato ed equivalente termine a secondo membro della (4.31), si può osservare che ciò che resta dopo le opportune semplificazioni sono dei termini $\ell_i$ che moltiplicano ciascuno il loro corretto e appropriato conseguente nella regola relativa. In altre parole, essi sono proprio quelle correlazioni puntuali che dovrebbero essere correttamente considerate, nella (4.29), per ciascuna singola regola, mentre $\eta_i$ è l'errore che si commetterebbe utilizzando $L$ in vece di $\ell_i$. Se dunque per $k = 2$ l'argomento della correlazione puntuale $L$ rappresenta il corretto indicatore sintetico di verifica ricercato nel § 4.8, ciò non può affatto dirsi quando $k > 2$, giacché in quel caso può ben

essere che $\exists\, \ell_i \mid L \neq \ell_i$ e, anzi, quest'ultima disuguaglianza è non solo possibile, ma anche inevitabile per determinati valori di $L$, come si vedrà nel seguito.

Per proseguire l'esposizione è conveniente utilizzare una notazione vettoriale per talune delle grandezze sopra richiamate, per cui nel seguito verranno utilizzati dei vettori colonna, salvo diversa specificazione, definiti come

$$\mathbf{x} = \begin{pmatrix} \hat{p}(x_1) \\ \vdots \\ \hat{p}(x_k) \end{pmatrix}, \quad \boldsymbol{\ell} = \begin{pmatrix} \ell_1 \\ \vdots \\ \ell_k \end{pmatrix}, \quad \boldsymbol{\eta} = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_k \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} \frac{\prod_{j=1}^{k} \hat{p}(x_j)}{\hat{p}(x_1)} \\ \vdots \\ \frac{\prod_{j=1}^{k} \hat{p}(x_j)}{\hat{\hat{p}}(x_k)} \end{pmatrix} \tag{4.33}$$

e dove $\mathbf{x}$ denota il vettore delle probabilità degli antecedenti e $\mathbf{c}$ il vettore delle probabilità dei conseguenti se indipendenti. Si utilizzeranno altresì, con un lieve abuso di notazione, le forme $\mu(\mathbf{v})$, $\mu_g(\mathbf{v})$ e $\mu_h(\mathbf{v})$ per indicare rispettivamente la media aritmetica, geometrica e armonica degli elementi dell'argomento inteso come vettore. Si osservi che con questa notazione le (4.31), (4.32) e in generale le rappresentazioni di $L$ per ogni $k$, possono essere tutte ricondotte a una forma sintetica, che si ottiene attraverso il prodotto membro a membro dei $k$ termini equivalenti e la successiva estrazione della radice $k$-esima, talché

$$L = \left|\sqrt[k]{L^k}\right| = \left|\sqrt[k]{\prod_{i=1}^{k} \eta_i \cdot \ell_i}\right| = \left|\sqrt[k]{\prod_{i=1}^{k} \ell_i}\right| \cdot \left|\sqrt[k]{\prod_{i=1}^{k} \eta_i}\right| = \mu_g(\boldsymbol{\ell}) \cdot \mu_g(\boldsymbol{\eta}) \tag{4.34}$$

e si riconoscono facilmente le medie geometriche dei singoli termini $\ell_i$ e $\eta_i$, la cui stretta positività ne assicura la consistenza. Inoltre, posto $s = \hat{p}(x_1, x_2, \dots, x_k)$, la (4.29) può essere convenientemente riscritta anche come

$$\vartheta = \frac{s}{\mu_h(\mathbf{x})} = \mu_g(\boldsymbol{\ell}) \cdot \mu_g(\boldsymbol{\eta}) \cdot \mu(\mathbf{c}) = L \cdot \mu(\mathbf{c}) \tag{4.35}$$

e confrontata con l'espressione alla quale è necessario ricondursi per ricavare il corretto moltiplicatore in forma di media aritmetica della (4.27), ovvero

$$\vartheta = \frac{s}{\mu_h(\mathbf{x})} = L_a \cdot \frac{1}{k} \boldsymbol{\eta}^{\mathrm{T}} \mathbf{c} = L_a \cdot \mu(\mathbf{y}) \tag{4.36}$$

dove $L_a$ è il moltiplicatore cercato e $\mathbf{y}$ è il vettore, ignoto, dei corretti e appropriati conseguenti per le regole ad antecedente unitario prese in esame dalla misura di Kulczynsky. Evidentemente, può dirsi che $L_a = L$ solo quando $k = 2$, in quanto, in quel caso e solo in quel caso, $\mathbf{y}$ e $\mathbf{c}$ coincidono necessariamente, mentre al di fuori di tale caso singolare, il problema può essere risolto esattamente solo attraverso la conoscenza o del vettore $\boldsymbol{\ell}$ o del vettore $\boldsymbol{\eta}$ o del vettore $\mathbf{y}$, alternativamente.

Pur considerando la forte riduzione dei gradi di libertà che viene indotta dall'essere interessati solo alle medie dei vettori coinvolti, nondimeno la soluzione, anche in questi termini, non può essere trovata esattamente con la sola conoscenza di $\vartheta, L$ e $s$, ma può essere al più stimata. Ancor più

sfortunatamente, per stimare $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{c}$ non è possibile avvalersi solo delle medie disponibili, ma è necessario anche poter eseguire una stima puntuale per ogni elemento di almeno uno dei vettori ignoti, in quanto è necessario farli corrispondere esattamente per condurre il prodotto scalare.

Si noti tuttavia che tale ulteriore conoscenza non sarebbe affatto richiesta qualora si volesse ricavare la media geometrica di $\boldsymbol{\ell}$, giacché in quel caso, basandosi l'intero calcolo su prodotti, la corrispondenza puntuale non sarebbe più necessaria, in virtù della libertà di scambio dei fattori. La quantità di informazione disponibile determina pertanto anche il tipo di media che si può eseguire. Nel caso $k = 2$, ancorché non sarebbe ivi necessario stimare alcunché, nondimeno la scelta dell'una o dell'altra media sarebbe indifferente, giacché entrambe restituirebbero il medesimo valore; negli altri casi la scelta determinerà la sensibilità della misura in caso di regole squilibrate: e cioè verso le regole più correlate qualora si usi la media aritmetica ($L_a$) oppure verso le regole meno correlate ove si usi la media geometrica ($L_g$). Per valutare la scelta, si deve rammentare il più volte ribadito valore di media di minimi rappresentato dalla misura di Kulczynsky e allora, per restare in quel quadro di *lower bound*, sembra del tutto ragionevole orientarsi, nella ricerca di una misura verificatrice di $\vartheta$, proprio sulla media geometrica, che risulta pertanto non solo quella stimabile col minor numero di informazioni, ma anche la più coerente.

Tutto ciò premesso, per trovare la formulazione della misura cercata sotto forma di media geometrica è innanzitutto necessario ricavare alcune relazioni fondamentali. Si cominci con l'osservare che, in base alle (4.31) e (4.32) ed eseguendo i prodotti in colonna, è possibile esprimere la media $\mu_g(\boldsymbol{\ell})$ anche nel modo seguente

$$L_g = \mu_g(\boldsymbol{\ell}) = \frac{s}{\mu_g(\mathbf{y}) \cdot \mu_g(\mathbf{x})} \tag{4.37}$$

e quindi per stimare $L_g$ si potrà maggiorare e minorare $\mu_g(\mathbf{y})$ utilizzando opportunamente a tale scopo le proprietà di anti-monotonicità della probabilità congiunta e il criterio di internalità di Cauchy. Essendo gli elementi di $\mathbf{y}$ rappresentati ciascuno dalle probabilità congiunte di $k-1$ eventi, ottenuti combinandone $k$, necessariamente tali probabilità non possono essere né inferiori alla probabilità congiunta di tutti i $k$ eventi né ciascuna superiore alla più piccola delle probabilità dei $k-1$ eventi che essa in particolare congiunge. Passando alle medie, ciò si traduce in

$$s \leq \mu_g(\mathbf{y}) \leq \nu(\mathbf{x}) \leq \mu_g(\mathbf{x}) \tag{4.38}$$

dove $\nu(\mathbf{x})$ è la media geometrica dei minimi delle $k$ combinazioni in gruppi di $k-1$ degli elementi di $\mathbf{x}$**.** Si può facilmente dimostrare che $\nu(\mathbf{x})$ è la media pesata tra i due più piccoli valori in $\mathbf{x}$ e si potrebbe anche mostrare, ma la trattazione è piuttosto complessa ed esula dagli scopi di questo lavoro, che nella generalità dei casi $\nu(\mathbf{x}) \lesssim \mu_h(\mathbf{x})$ e solo in alcuni casi — e in presenza di elevate differenze tra il primo e il secondo minimo — $\mu_h(\mathbf{x}) \lesssim \nu(\mathbf{x}) \leq \mu_g(\mathbf{x})$. Peraltro, il problema dell'approssimazione si pone solamente *a posteriori*, giacché in sede di calcolo quello di $\nu(\mathbf{x})$ non richiede che sia noto alcun valore aggiuntivo rispetto a quelli occorrenti per calcolare $\vartheta$ o $L$. In questo senso la stima di $L_g$ può svolgersi in due modi: o utilizzando solo $\vartheta$ e $L$ oppure aggiungendovi la conoscenza del vettore $\mathbf{x}$, conoscenza che è comunque sempre necessaria in sede di calcolo delle prime due misure. Naturalmente, è appena il caso di far rilevare che quando $k = 2$, allora $L_g = L$.

### *4.9.1 Stima con la sola conoscenza di $\vartheta$ e $L$*

Con le informazioni ricavate esclusivamente da $\vartheta$ e $L$ la stima di $L_g$ potrà coinvolgere solo funzioni lisce e ciò, in particolare, non permette di ricavare $\nu(\mathbf{x})$. Tuttavia, considerandone la natura quasi equivalente a quella di minimo, si commetterà un errore del tutto accettabile utilizzando a tal fine $\mu_h(\mathbf{x})$, talché, in base alla relazione (4.37) e sfruttando opportunamente le relazioni $\mu_h(\mathbf{x}) = \frac{s}{\vartheta}$ e $\mu_g(\mathbf{x}) = \sqrt[k]{s/L}$ si giunge alla conclusione che, per $k > 2$,

$$\frac{s}{\mu_h(\mathbf{x})\mu_g(\mathbf{x})} \leq L_g \leq \frac{1}{\mu_g(\mathbf{x})} \implies s \leq \frac{\vartheta}{s}\left(\sqrt[k]{\frac{L}{s}}\right) \leq L_g \leq \sqrt[k]{\frac{L}{s}} \leq s^{-1} \tag{4.39}$$

da cui, ipotizzando una distribuzione log-normale (o anche solo log-uniforme) di $L_g$, si può trarre sinteticamente una stima del valore cercato dal valore log-centrale

$$L_g \cong \sqrt{\vartheta\left(\sqrt[k]{\frac{L}{s}}\right)\left(\sqrt[k]{\frac{L}{s}}\right)} = \sqrt{\vartheta}\left(\sqrt[k]{\frac{L}{s}}\right), \quad \text{se } k > 2 \tag{4.40}$$

È evidente che l'accuratezza della previsione non può andare oltre un certo limite, in virtù delle approssimazioni introdotte e della mera ipotesi fatta sulla distribuzione dei valori; nondimeno, quella che era la sostanziale inattendibilità di $L$ per $k > 2$ viene ora ricondotta entro i limiti inferiori e superiori di $L_g$, che delimitano un campo di variazione più che soddisfacente ai fini dell'analisi condotta in questo lavoro.

### *4.9.2 Stima con la conoscenza del vettore di probabilità degli antecedenti unitari*

Se alla conoscenza delle misure di base si aggiunge — o vi si sostituisce — quella del vettore $\mathbf{x}$, si ottiene innanzitutto la facoltà di poter razionalmente stimare anche $\mu(\mathbf{y})$, ma in ragione delle scelte fatte ciò non rappresenta un immediato vantaggio. Di ben maggior pregio, almeno per gli scopi di questo lavoro, è invece la possibilità di aumentare la precisione della (4.40) calcolando esattamente la funzione non lineare $\nu(\mathbf{x})$, grazie alla quale si potrà scrivere

$$L_g \cong \sqrt{\frac{s}{\nu(\mathbf{x})\mu_g^2(\mathbf{x})}} = \sqrt{\frac{s}{\nu(\mathbf{x})}}\left(\sqrt[k]{\frac{L}{s}}\right), \quad \text{se } k > 2 \tag{4.41}$$

che riproduce la (4.40) con un grado però maggiore di precisione.

## 4.10 Valenza della correlazione totale puntuale come misura autonoma

La riconduzione della correlazione totale puntuale $L$, quando essa è inattendibile, a un suo valore probante $L_g$ per la misura di Kulczynsky $\vartheta$, che è stata sviluppata nel paragrafo precedente, è limitata al solo sottoinsieme di regole che quest'ultima misura prende in considerazione, ovvero quelle ad antecedente unitario o, se si vuole, a confidenza minima. Ci si aspetterebbe allora che il valore originale di $L$ possa rappresentare un indicatore probante con riguardo *a tutte le regole*

generabili dall'*itemset* sulla quale essa viene calcolata, ma disgraziatamente, anche in questo caso, tale aspettativa risulta immediatamente soddisfatta solo quando $k = 2$.

Infatti, se ci si propone, utilizzando le notazioni del paragrafo precedente, di costruire il vettore $\mathbf{y}$ delle probabilità dei conseguenti di tutte le possibili regole generabili da un generico *itemset* di lunghezza $k$, si vede facilmente che esso sarà costituito necessariamente dalle probabilità di tutte le congiunzioni ottenute per combinazione di $k$ eventi in gruppi di $1,2,\dots,k-1$ elementi, ovvero $\binom{k}{1}$ elementi di forma $p(x_i)$, $\binom{k}{2}$ elementi di forma $p(x_i, x_j)$ *et cetera*; analogamente accadrebbe volendo costruire il vettore $\mathbf{x}$ delle probabilità di tutti gli antecedenti, il quale avrebbe i medesimi elementi di $\mathbf{y}$ semplicemente ordinati in modo diverso per farli corrispondere. Ad ogni modo, quali che siano i valori degli elementi di $\mathbf{x}$ e di $\mathbf{y}$, quel che è certo è che a nessuno di essi potrà mai essere, in virtù della proprietà di anti-monotonicità, né minore di $s = \hat{p}(x_1, x_2, \dots, x_k)$ né maggiore di $M_x = \max\{\hat{p}(x_1), \hat{p}(x_2), \dots, \hat{p}(x_k)\}$ e quindi, per ciascuna regola $i$-esima, in base alla (4.27), dovrà valere necessariamente

$$\frac{s}{[\mathbf{x}]_i} = [\boldsymbol{\ell}]_i [\mathbf{y}]_i \Longrightarrow [\boldsymbol{\ell}]_i = \frac{s}{[\mathbf{x}]_i [\mathbf{y}]_i} \Longrightarrow \frac{s}{M_x^2} \leq [\boldsymbol{\ell}]_i \leq \frac{1}{s} \tag{4.42}$$

ma, d'altro canto, se ci si propone di minorare e maggiorare $L$ in stretta analogia, ci si avvede immediatamente che

$$L = \frac{s}{\prod_{i=1}^{k} \hat{p}(x_i)} \Longrightarrow \frac{s}{M_x^k} \leq L \leq \frac{1}{s^{k-1}} \tag{4.43}$$

e che, conseguentemente, per $k > 2$ e valori dei $\hat{p}(x_i)$ sufficientemente piccoli, il valore assunto da $L$ può superare di molti ordini di grandezza l'analogo indicatore di qualsivoglia regola, per ben correlata che essa sia, generabile dall'*itemset* preso in considerazione. Ciò suggerisce che $L$, quando è maggiore dell'unità e in particolare di $s^{-1}$, trasporti e riunisca in sé un'informazione di tipo complesso, che deve essere scomposta per poterne estrarre la parte di interesse.

È appena il caso di dire che la parte di interesse è naturalmente rappresentata dal valore log-centrale $\mu_g(\boldsymbol{\ell})$ che da un lato è strettamente limitato tra $s$ e $s^{-1}$, dall'altro rappresenta una misura uniforme e coerente con quella analoga, seppur limitata alla verifica di $\vartheta$, rappresentata da $L_g$, che si è ampiamente analizzata nel paragrafo precedente.

La ricerca della possibile relazione tra $L$ e $\mu_g(\boldsymbol{\ell})$ richiede innanzitutto che sia chiarita la forma matematica della seconda. Il numero $n$ di regole generabili dipende solo da $k$ ed è ovviamente, rappresentato dalla somma delle combinazioni già sopra citate, ovvero

$$n = \sum_{i=2}^{k-1} \binom{k}{i} = 2^k - 2 \tag{4.44}$$

e conseguentemente il valore $\mu_g(\boldsymbol{\ell})$ sarà allora

$$\mu_g(\boldsymbol{\ell}) = \sqrt[n]{\prod_{i=1}^{n} \frac{s}{[\mathbf{x}]_i [\mathbf{y}]_i}} = \frac{s}{\mu_g(\mathbf{x}) \mu_g(\mathbf{y})} = \frac{s}{\mu_g^2(\mathbf{x})} \tag{4.45}$$

Confrontando la (4.45) con la (4.43) ci si avvede che la ricerca della relazione tra $L$ e $\mu_g(\boldsymbol{\ell})$ si riduce in sostanza alla ricerca della relazione tra $\prod_{i=1}^{k}\hat{p}(x_i)$ e $\mu_g^2(\mathbf{x})$, trovata la quale verrebbe risolto immediatamente anche il problema principale. Sfortunatamente il secondo termine, ancorché dipenda da congiunzioni rigidamente combinate degli eventi le cui probabilità sono i fattori del primo, nondimeno è di natura prettamente aleatoria, per cui il problema non sembra, *prima facie*, aver soluzioni determinabili.

A ben guardare, tuttavia, si può pensare di mettersi nelle condizioni di rimuovere l'aleatorietà e di trovare almeno le soluzioni deterministiche, cercando di trovare da quelle una relazione che possa applicarsi analogicamente anche quando entri in gioco la casualità. Si ipotizzi per esempio di limitare l'indagine a quando è $L \geq 1$ e di consentire per le probabilità elementari $\hat{p}(x_i)$ i soli due valori estremi $s$ e $1$. Con l'imposizione di tale enumerazione finita, da un lato le congiunzioni del tipo $p(\Omega, \Omega, \ldots, \Omega)$ avrebbero deterministicamente probabilità $1$ e, dall'altro, altrettanto deterministicamente qualunque altra congiunzione di eventi non tutti certi avrebbe probabilità $s$. Si noti anche che in conseguenza di ciò i valori di $L$ e $\prod_{i=1}^{n}[\boldsymbol{\ell}]_i$ sarebbero condizionati rispettivamente ad assumere solo valori di forma $s^p$ e $s^q$, con $p$ e $q$ interi relativi. Rimossa in questo modo l'aleatorietà, la soluzione del problema in questi termini è un semplice esercizio di calcolo combinatorio orientato a trovare $q$ dato $p$, che può essere svolto facilmente e che fornisce il seguente risultato

$$q = \log_s\left(\prod_{i=1}^{n}[\boldsymbol{\ell}]_i\right) = \log_s\left(\prod_{i=1}^{2^k-2}\frac{s}{[\mathbf{x}]_i[\mathbf{y}]_i}\right) = 2^k\left(2^{-p}-1\right) = 2^k\left(2^{\log_s L}-1\right) \tag{4.46}$$

il quale vale per $L = s^{-p},\ p \in \mathbb{N},\ p \leq k,\ s < 1$.

Se a questo punto si rimuovono i vincoli imposti sulle $\hat{p}(x_i)$ e si consente così che possa essere $p \in \mathbb{R}, 0 \leq p \leq k$, la (4.46) diviene l'espressione di una curva interpolante i valori esatti quando $p$ è intero. Lungi dall'aver ovviato all'aleatorietà di $\mathbf{x}$, ciò che si è in realtà ottenuto è la riconduzione di un fenomeno aleatorio dotato di una misura nota (ovvero $L$) ad almeno un *equivalente* modello deterministico — che con riguardo ai $\hat{p}(x_i)$ si potrebbe definire asintotico — che esibisce ovunque la medesima misura e che fornisce altresì un calcolo, seppure di applicabilità forzatamente analogica, su come estrarre dalla misura nota la parte ignota di interesse.

Chiarito l'approccio seguito, che nella sua limitatezza ha comunque una base razionale, a differenza delle soluzioni empiriche al problema costituite dal saturare superiormente $L$ con $s^{-1}$ o dal prendere sistematicamente $\sqrt[k-1]{L}$, l'espressione (4.46) conduce, eseguendo l'estrazione di radice e invertendo il logaritmo, alla formulazione della seguente ipotesi:

$$\mu_g(\boldsymbol{\ell}) \approx s^{\frac{2^k\left(2^{\log_s L}-1\right)}{2^k-2}} = s^{-\left(\frac{2^{\log_s L}-1}{2^{1-k}-1}\right)} = s^{-\lambda},\ \text{se}\ k > 2 \wedge s < 1 \wedge L \geq 1 \tag{4.47}$$

che, con riguardo all'esponente, può essere anche più efficacemente formulata in termini di logaritmi naturali come

$$\lambda = \frac{2^{\frac{\ln L}{\ln s}}-1}{2^{1-k}-1} \tag{4.48}$$

dove si noti che quando $s \to 1$, allora $L \to \frac{s}{s^k}$, per cui $\lim_{s\to 1}\frac{\ln L}{\ln s} = (1-k)$ e quindi $\lim_{s\to 1}\lambda = 1$.

A rigore, resterebbe da considerare quando si ha, a un tempo, $L < 1$ e $k > 2$, ma in tali casi se da un lato l'approccio seguito per ottenere la (4.46) non potrebbe essere utilizzato, dall'altro non vi sarebbe neppure la necessità di scomporre $L$ giacché esso in base alla (4.42) e (4.43) assumerebbe sempre valori in $[s, 1)$ qual che fosse il valore di $k$, suggerendo che lo stesso $L$ possa essere preso direttamente come indicatore o, equivalentemente, possa esser preso $\lambda = -\log_s L$.

## 4.11 Impiego applicativo delle misure

Una volta esaminate le caratteristiche, i limiti e l'interpretazione delle misure proposte in letteratura oppure qui direttamente ricavate, è necessario discuterne l'impiego in relazione ai dati sperimentali che costituiscono l'oggetto di questo lavoro.

Il presupposto, già accennato in principio di capitolo, è che l'osservazione congiunta di una misura *null-invariant* e di una misura *expectation-based* sia apportatrice di benefici in termini di mutua attenuazione dei valori eventualmente fuorvianti forniti dall'una o dall'altra prese separatamente. Per quanto riguarda la misura *null-invariant*, è già stata anticipata e discussa la scelta della misura di Kulczynsky, della quale si è data ampia analisi nei paragrafi precedenti, mentre per quanto riguarda invece la misura *expectation-based,* si è già mostrato come la scelta d'elezione dovrebbe essere una derivazione della *pointwise total correlation,* una volta che da essa siano rimossi gli inconvenienti, anch'essi ampiamente analizzati, che non ne consentono la diretta e immediata applicabilità agli *itemset* di lunghezza $k > 2$.

Per la misura di Kulczynsky $\vartheta$ vi è poco da aggiungere, in quanto essa esprime una probabilità ed è pertanto naturalmente limitata all'intervallo $[0,1]$ e inoltre la sua confrontabilità tra *itemset* di diversa lunghezza è già stata chiarita nel § 4.8. Affatto diverso è il caso delle misure derivate dalla *pointwise total correlation*, per le quali è ancora necessario affrontare il problema della loro confrontabilità con riguardo sia alla lunghezza $k$ sia al supporto $s$. Per quanto riguarda $k$ si è visto invero che $L$ è limitato a $[s, s^{-1}]$ quando $k = 2$ e che quando $k > 2$, in vece di $L$ si dovrebbero impiegare le misure $L_g$ e $s^{-\lambda}$, le quali sono anch'esse rigorosamente limitate all'intervallo $[s, s^{-1}]$, per cui occorre solo discutere la normalizzazione rispetto al supporto. È allora appena il caso di osservare che sarà sufficiente prendere di esse il logaritmo in base $s$ col segno opportuno per ottenere misure rigorosamente limitate all'intervallo $[-1, +1]$ e quindi sarà sufficiente estendere la (4.48) e ricondurre alla medesima misura $\lambda$ (quindi al logaritmo) anche tutti gli altri casi qui considerati, ovvero

$$\lambda = \begin{cases} 0 & \text{se } k = 1 \\ 1 & \text{se } s = 1 \wedge k \neq 1 \\ -\dfrac{\ln L}{\ln s} & \text{se } k = 2 \vee L < 1 \\ \dfrac{2^{\frac{\ln L}{\ln s}} - 1}{2^{1-k} - 1} & \text{se } k > 2 \ \wedge L \geq 1 \end{cases} \tag{4.49}$$

Si noti che il logaritmo in base $s$ di $L$ non è nient'altro (a meno del segno) che una sua normalizzazione entropica, se si considera che $-\ln s = -\ln \hat{p}(x_1, x_2, \dots, x_k) = I(x_1, x_2, \dots, x_k)$ è proprio

l'autoinformazione in base neperiana recata dalla probabilità congiunta. Si noti altresì che nel caso $k = 3$, se si ha a disposizione $\vartheta$ o il vettore di probabilità degli antecedenti, è possibile ottenere un valore molto più accurato prendendo $\lambda = \log_s L_g$, in quanto tutte le regole generabili da un siffatto *itemset* sono anche tutte prese in considerazione dalla sua misura di Kulczynsky. Si può dunque concludere che $\lambda$, così come definita nella (4.49) costituisca una normalizzazione coerente di tutte le derivazioni dalla *pointwise total correlation* e quindi ad essa, nel seguito, ci si riferirà anche come *nptc*.

Le due misure autonome alle quali si è così infine pervenuti, ovvero $\vartheta$ e $\lambda$, sono quindi entrambe limitate superiormente e inferiormente e sono ortogonali[15] almeno dal punto di vista del significato loro proprio, talché, in un certo senso, è possibile pensarle come la parte rispettivamente reale e immaginaria di un numero complesso $z = \vartheta + i\lambda$, dotato di modulo e fase e in grado di esprimere, a un tempo, la potenzialità in termini di confidenza e la validità statistica delle regole generabili da un *itemset*. Disgraziatamente, se tale lettura è consistente dal punto di vista strettamente analitico, lo stesso non può dirsi in termini di applicabilità numerica e di impiego operativo della misura complessa così ottenuta, se non fosse altro per l'ottimo motivo che l'insieme di appartenenza non è un campo ordinato né una struttura algebrica ordinabile rispetto alla metrica indotta da una norma. Peraltro, i tentativi di generare, tramite combinazione delle due misure, un unico valore reale cumulativo e ordinabile finirebbe facilmente o per essere un mero empirismo — e quindi discutibile — o per cagionare una perdita di informazione, anche consistente, rispetto all'una o all'altra delle due componenti.

In realtà, esiste una espressione in cui la dualità delle misure *null-invariant* ed *expectation-based* coesiste rigorosamente ed è del tutto consistente col sottostante significato statistico e probabilistico. Tale espressione è la (4.27), ed è proprio da questa che si dovrà partire per ottenere la misura combinata cercata. Nello stesso modo in cui si può valutare complessivamente uno studente sia in base ai risultati raggiunti sia all'impegno profuso nello studio, considerandone sia la differenza tra i suoi voti finali rispetto alla sufficienza sia la differenza tra i suoi voti finali rispetto ai voti iniziali e in entrambi i casi usando la medesima grandezza, ossia il voto, così con la medesima uniformità di grandezze dovrebbe essere ricercata una misurazione complessiva sugli *itemset*, ciascuno dei quali, peraltro, più che un singolo studente rappresenterebbe una classe scolastica più o meno numerosa. Allora si potrebbe ben considerare la misura complessivamente cercata come la somma, eventualmente normalizzata, di due valori aggiunti: l'uno rappresentato dalla differenza di probabilità rispetto alla probabilità *a priori* dell'evento casuale e l'altro rappresentato dalla differenza di probabilità rispetto alla probabilità *a posteriori* dell'evento dipendente, ossia, con riferimento alla (4.27),

$$\left(p(E_y|E_x) - \frac{1}{2}\right) + \left(p(E_y|E_x) - p(E_y)\right) \tag{4.50}$$

Ma nell'ambito degli *itemset* si è già veduto che le probabilità riferite nella (4.50) sono tutte ricavabili sinteticamente, o perlomeno stimabili, conoscendo i valori di $L$ e di $\vartheta$ o del vettore delle

[15] Naturalmente $\vartheta$ e $\lambda$ non sono ortogonali in senso rigorosamente matematico. Posto che entrambe sono funzioni che mappano $P \subset (0,1]^{k+1} \longmapsto \mathbb{R}$, essendo $\vartheta = f_\vartheta(s, \mathbf{x})$ e $\lambda = f_\lambda(s, \mathbf{x})$, si potrebbe dimostrare che $\int_P f_\vartheta(s, \mathbf{x}) \cdot f_\lambda(s, \mathbf{x}) \mathrm{d}s\mathrm{d}X \neq 0$

probabilità elementari, talché usando i risultati del § 4.9 e normalizzando si potrà dunque scrivere

$$\zeta = \frac{2}{3}\left[\left(\vartheta - \frac{1}{2}\right) + \left(\vartheta - \frac{\vartheta}{L_g}\right)\right] = \frac{1}{3}\left[4\vartheta\left(1 - \frac{1}{2L_g}\right) - 1\right] \tag{4.51}$$

dove $\zeta$ rappresenta una misura complessiva, uniforme e consistente in termini di grandezze; inoltre, con la normalizzazione introdotta essa ha valori rigorosamente nell'intervallo $[-1, +1]$ e la presenza di possibili valori tanto negativi quanto positivi della misura costituisce un notevole vantaggio operativo, in quanto permette di individuare in modulo, ai due estremi, gli *itemset* di maggior interesse, ovvero non solo quelli ad alta correlazione positiva, ma anche, all'altro estremo, quelli ad elevata anti-dipendenza, ovvero suscettibili di fornire significative regole negative, ove di interesse.

Si è così pervenuti a individuare e definire, in modo specifico per gli *itemset*:

- un equivalente sintetico di ciò che è la confidenza per le regole associative, rappresentato dalla misura di Kulczynsky ($\vartheta$);
- degli equivalenti sintetici del *lift*, ossia $s^{-\lambda}$ e $L_g$, con valori rispettivamente riferiti il primo a *tutte* le regole generabili da un *itemset* e il secondo alle *sole* regole prese per esso in considerazione dalla misura di Kulczynsky, con le loro normalizzazioni entropiche (*nptc*);
- una nuova misura di interesse combinato, ossia *zeta* ($\zeta$), che sintetizza in un'unica misura operativamente utilizzabile, ordinabile e limitata, il rigoroso significato statistico sia della confidenza sia del *lift* congiuntamente riferiti alle regole associative elementari ad antecedente unitario generabili da un *itemset* (generalizzato (espressivo))

Si deve sottolineare che lo scopo delle misure sopra richiamate è in via principale quello di ordinare gli *itemset* sulla base delle loro caratteristiche oggettive nella misura in cui esse possono trasmutarsi in un rapido giudizio di qualità e interpretabilità delle relative regole e non quello di fornire un criterio per selezionare gli *itemset* maggiormente suscettibili di generarne. Sotto quest'ultimo profilo, tanto la misura di Kulczynsky quanto la misura *zeta*, sebbene in minor grado, sarebbero eccessivamente selettive con riguardo ai loro punti di neutralità e, d'altro canto, la lunghezza $k$ degli *itemset* rappresenterebbe invece un elemento cruciale in ragione della dipendenza esponenziale da esso del numero di regole generabili. Esse sarebbero tuttavia ben utilizzabili, in un approccio a *dataset* con cardinalità poco trattabili, per estrarre un nucleo fondamentale di *itemset* sui quali condurre le indagini e le calibrazioni preventive ad un processamento integrale dei dati. Per richiamare l'esempio sulla valutazione degli studenti, si potrebbe affermare che queste misure ignorerebbero lo studente geniale malcapitato in una classe di *minus habentes*, cercandolo solo nelle classi con profitti sopra la media. Ma, del resto, quell'uomo che, smarrite le chiavi di casa in una notte buia, le andava cercando sotto il lampione dall'altro lato della strada così rispose a chi gli faceva notare che più verosimilmente egli le avesse perdute dal lato opposto: "*È ben vero, ma è solo qui sotto che posso nutrire una speranza di ritrovarle.*"

# Capitolo 5

# Analisi e comparazione oggettiva

Al fine di ottenere una prima valutazione degli effetti derivati dall'applicazione degli algoritmi esaminati in questo lavoro ai risultati sperimentali presentati nel capitolo 2, si è ritenuto opportuno elaborare un quadro oggettivo delle caratteristiche statistiche di questi effetti. Tale quadro è ottenibile, nel modo più semplice, osservando le medie e le distribuzioni delle misure discusse nel capitolo 4 quando applicate agli *itemset* generati dall'applicazione degli algoritmi con diversi valori del supporto minimo. L'analisi fatta in questo capitolo consisterà pertanto nell'esame delle caratteristiche statistiche più salienti per ciascuna delle tre misure (Kulczynsky, *nptc* e *zeta*), valutando distintamente ciascuna delle partizioni ($\mathsf{M}\backslash\mathsf{G}$, $\mathsf{G}\backslash\mathsf{M}$ e $\mathsf{G}\cap\mathsf{M}$) per le soglie di supporto minimo più significative utilizzate dagli algoritmi, utilizzando, in ogni caso, solo gli *itemset* con $k \geq 2$ per eliminare l'effetto distorcente degli *itemset* unitari, privi di valore informativo.

## 5.1 Analisi in base alla misura di Kulczynsky

Osservando l'andamento della misura di Kulczynsky al crescere del supporto minimo $\varepsilon$, rappresentato nella parte sinistra della figura 14, si ha dapprima l'impressione che tutti gli algoritmi producano *itemset* vieppiù confidenti e che quindi l'azione di generalizzazione e recupero informativo da essi condotta sia efficace sui dati in esame. Purtroppo, tale impressione non è del tutto veritiera e infatti, richiamando quando osservato nel § 4.7 e depurando $\vartheta$ del suo valore minimo atteso a priori, si può vedere che dai dati in esame si osserva in realtà solo un modesto incremento per le partizioni $\mathsf{G}\backslash\mathsf{M}$ e $\mathsf{G}\cap\mathsf{M}$ e una sostanziale assenza di dinamismo per la partizione $\mathsf{M}\backslash\mathsf{G}$, come mostrato nella parte destra di figura 14.

Un ulteriore contributo alla comprensione dell'azione degli algoritmi in ragione dell'aumentare del supporto minimo proviene dall'esame delle distribuzioni in quattro classi della misura di Kulczynsky, mostrate in figura 15. Le quattro classi di interesse sono gli intervalli di $\vartheta$ corrispondenti a $(0, ¼]$, $(¼, ½]$, $(½, ¾]$ e $(¾, 1]$ e sono rappresentate graficamente con toni di grigio crescente dal basso verso l'alto. Per tutte e tre le partizioni appare evidente la progressiva riduzione della percentuale di *itemset* con $\vartheta \in (0, ¼]$, che è nettamente più marcata tuttavia in $\mathsf{G}\backslash\mathsf{M}$ e $\mathsf{G}\cap\mathsf{M}$. Appare interessante anche il differente dinamismo della classe $(¾, 1]$, che dapprima significativa in $\mathsf{M}\backslash\mathsf{G}$ va ivi decrescendo fino ad annullarsi al crescere di $\varepsilon$, in modo esattamente opposto alle altre due partizioni.

Approfondendo ancora il livello di dettaglio, le distribuzioni ad alta granularità delle classi, tale

da renderle quasi continue, mostrate in figura 16 e figura 17 rivelano ulteriori e significative caratteristiche.

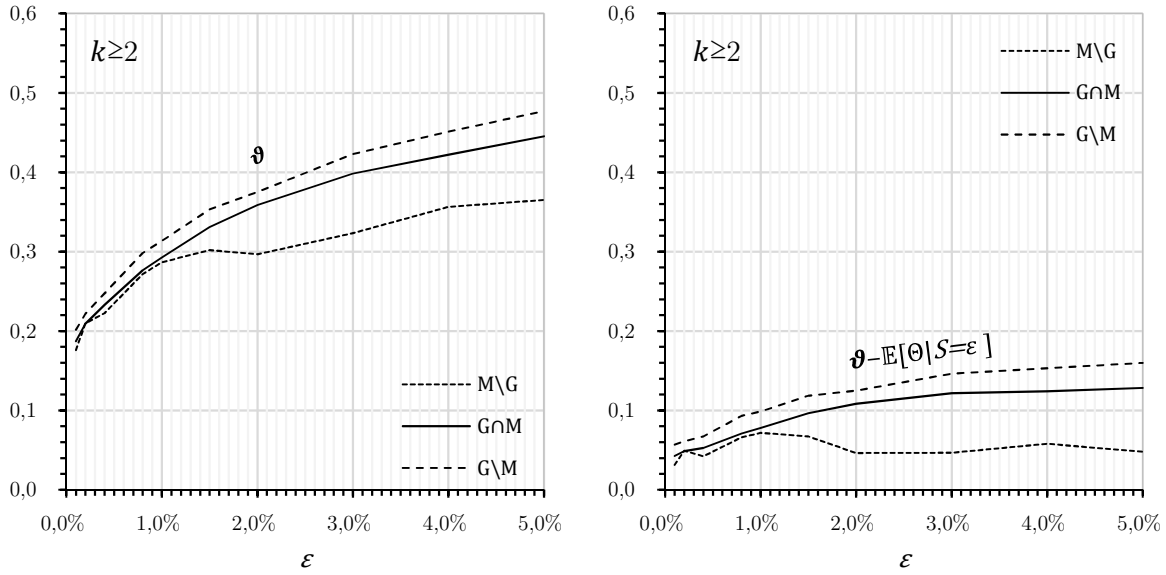



**Figura 14** — *Misura di Kulczynsky: valor medio ponderato osservato (a sinistra) e depurato del valore minimo atteso (a destra) in dipendenza dalla soglia percentuale di supporto minimo* $\varepsilon$.

La più significativa di esse è che tali distribuzioni, lungi dall'essere regolari, si palesano come se fossero costituite da una distribuzione di base positivamente asimmetrica e ad ampio supporto alla quale si sovrappongono delle distribuzioni a stretto supporto ed elevata curtosi, che conferiscono un carattere di distribuzione multimodale molto spiccato, soprattutto alla partizione $\mathsf{G\backslash M}$.

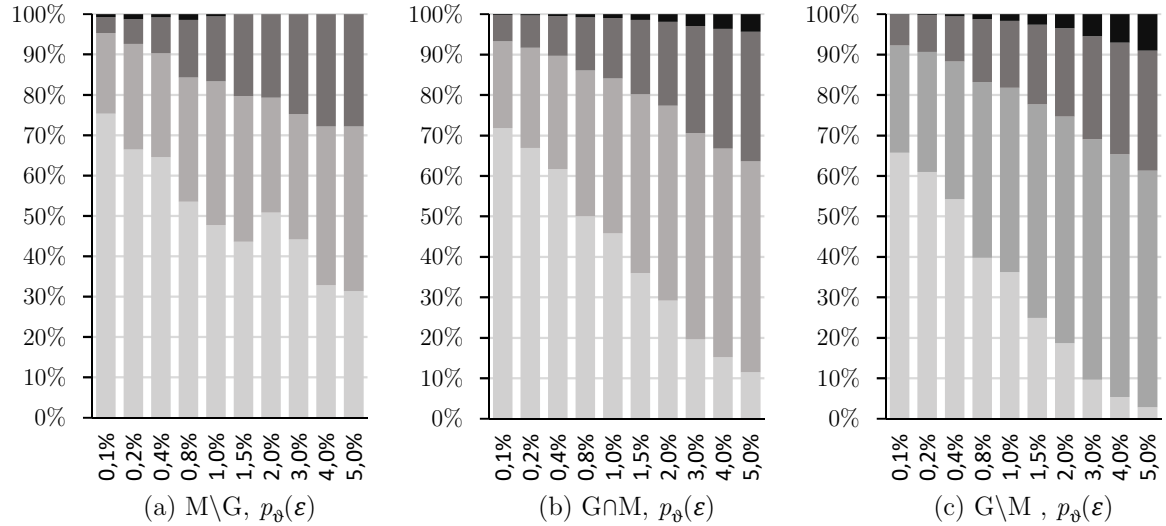
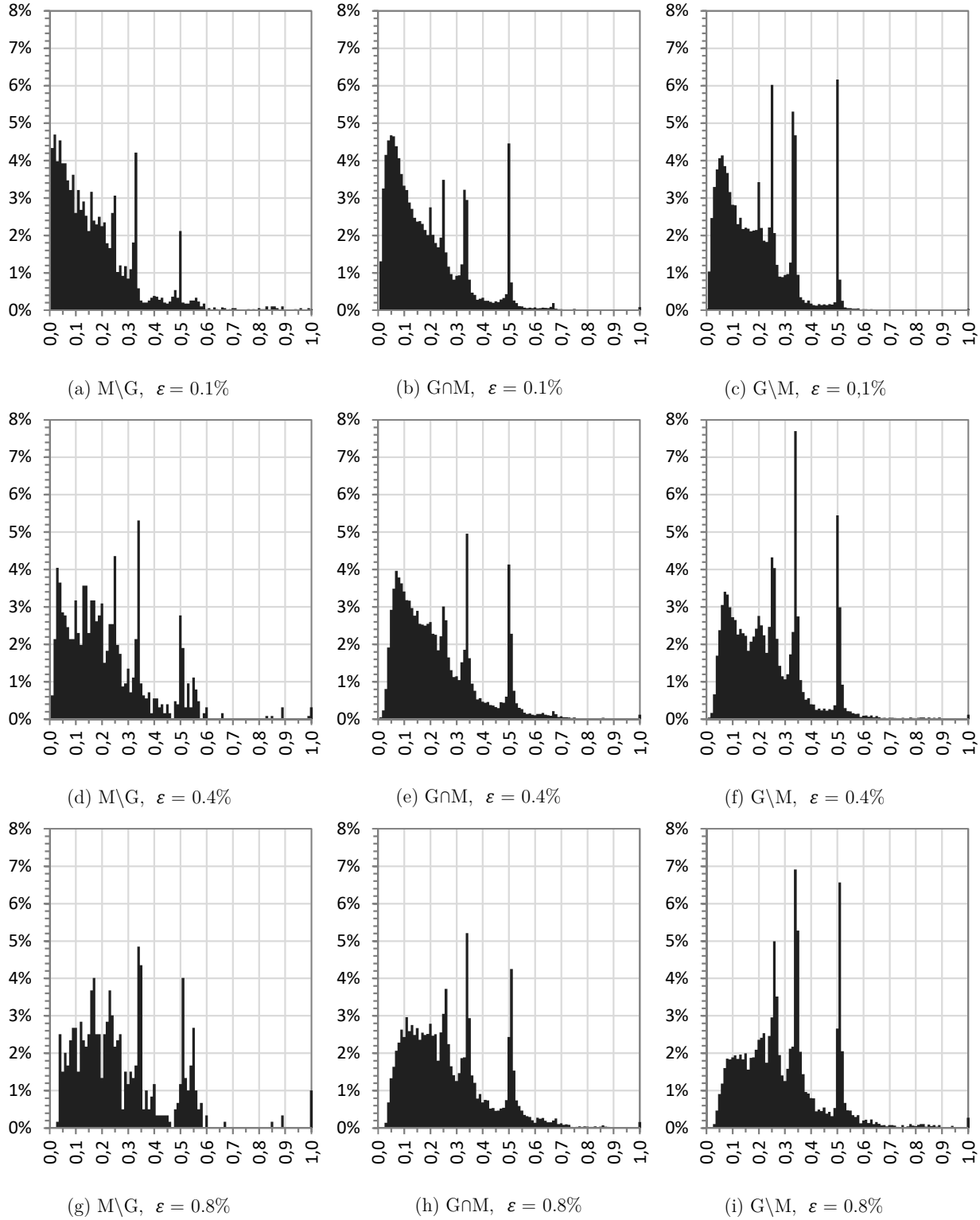



**Figura 15** — *Misura di Kulczynsky: distribuzione delle frequenze percentuali nelle quattro classi* $(\mathbf{0}, ¼]$, $(¼, ½]$, $(½, ¾]$ e $(¾, \mathbf{1}]$*, partendo dal basso, in dipendenza dalla soglia percentuale di supporto minimo* $\varepsilon$

La distribuzione di base, al crescere di $\varepsilon$, appare vieppiù perdere la sua positiva asimmetria e la sua rilevanza, soprattutto in $\mathsf{G\backslash M}$, mentre la rilevanza delle distribuzioni sovrapposte ad essa e ad elevata curtosi si accresce corrispondentemente. Tali distribuzioni a elevata curtosi hanno approssimativamente mode (almeno quelle facilmente distinguibili nei grafici) nei punti $½, ⅓, ¼$ e $⅕$, segno che esse provengono da ricorrenti e limitate combinazioni delle confidenze prese in esame

dalla misura di Kulczynsky. Il fatto che esse si manifestino in misura nettamente più significativa nella partizione G\M parrebbe indicare per essa un contenuto più tipicamente caratterizzato da *itemset* con tali limitate combinazioni.

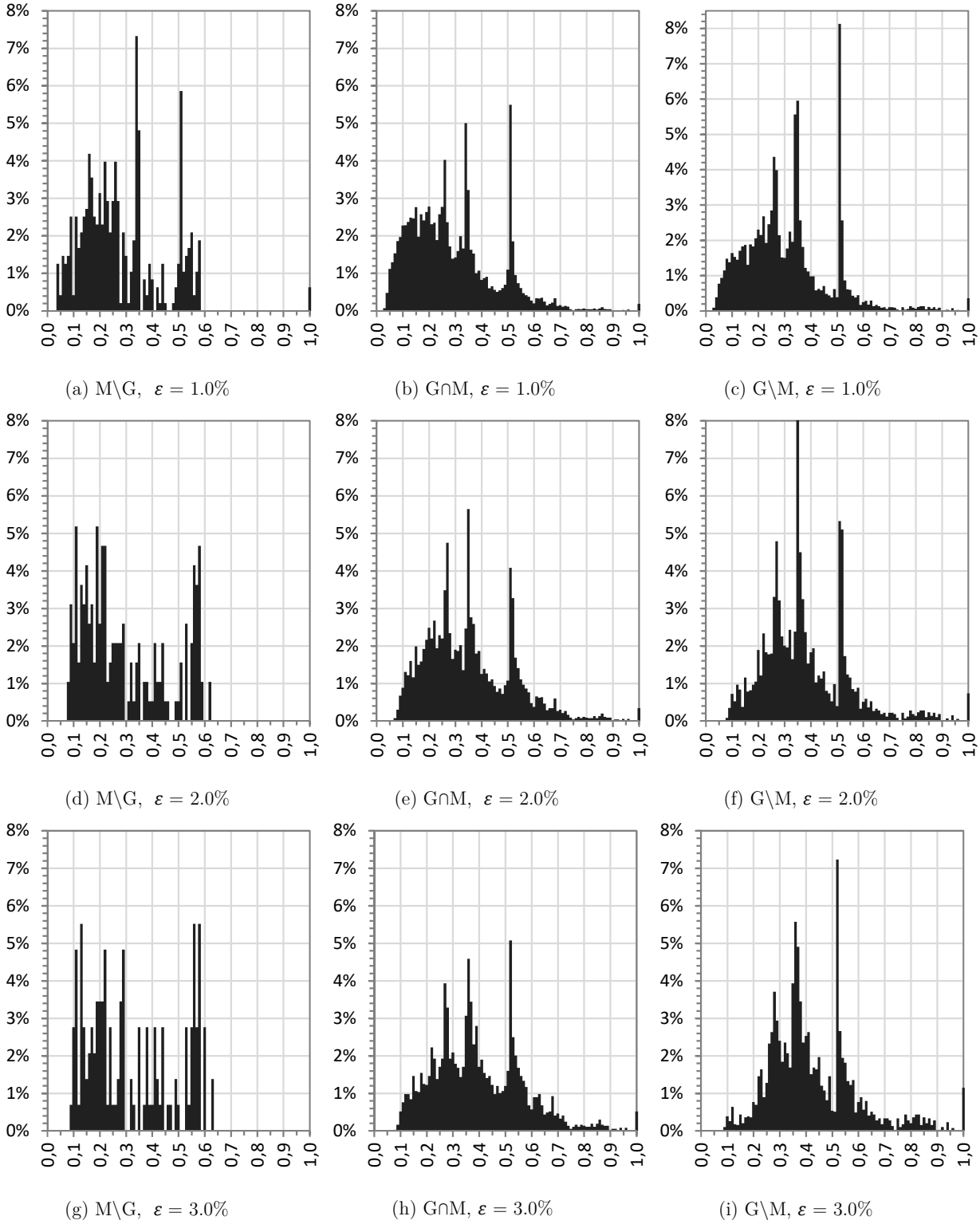



(a) M\G, $\varepsilon = 0.1\%$ (b) G∩M, $\varepsilon = 0.1\%$ (c) G\M, $\varepsilon = 0{,}1\%$

(d) M\G, $\varepsilon = 0.4\%$ (e) G∩M, $\varepsilon = 0.4\%$ (f) G\M, $\varepsilon = 0.4\%$

(g) M\G, $\varepsilon = 0.8\%$ (h) G∩M, $\varepsilon = 0.8\%$ (i) G\M, $\varepsilon = 0.8\%$

**Figura 16** — *Misura di Kulczynsky: distribuzione dei valori di* $\vartheta$ *(in ascissa) per soglie di supporto minimo* $\varepsilon < 1.0\%$



(a) M\G, $\varepsilon = 1.0\%$ (b) G∩M, $\varepsilon = 1.0\%$ (c) G\M, $\varepsilon = 1.0\%$

(d) M\G, $\varepsilon = 2.0\%$ (e) G∩M, $\varepsilon = 2.0\%$ (f) G\M, $\varepsilon = 2.0\%$

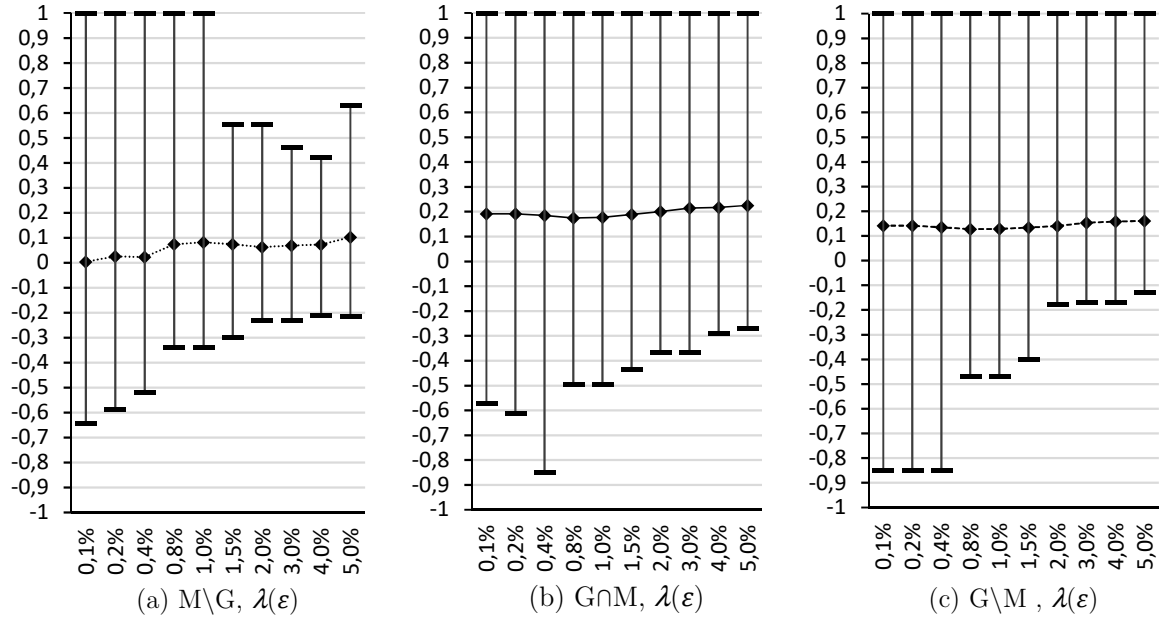
(g) M\G, $\varepsilon = 3.0\%$ (h) G∩M, $\varepsilon = 3.0\%$ (i) G\M, $\varepsilon = 3.0\%$

**Figura 17** — *Misura di Kulczynsky: distribuzione dei valori di $\vartheta$ (in ascissa) per soglie di supporto minimo $\varepsilon \geq 1.0\%$*

Per quanto riguarda la partizione $\mathsf{M \backslash G}$, si può notare come la sua bassa cardinalità rende vieppiù meno significativa la lettura della distribuzione, che da un certo punto in poi neppure può essere resa ad un tale alto livello di granularità. Nondimeno, fin dove è possibile, si scorgono in essa le medesime caratteristiche di sovrapposizione nettamente visibili nelle altre due.

## 5.2 Analisi in base alla correlazione totale puntuale normalizzata

L'analisi dei valori minimi, medi e massimi della correlazione totale puntuale normalizzata, mostrati in figura 18, rivela un profilo migliore per la partizione G∩M rispetto alle altre due, sia per la media maggiormente orientata verso la correlazione, sia per il supporto di $\lambda$ che arriva ovunque al valore positivo massimo. Si può infatti constatare che il valore medio della correlazione puntuale osservata in G∩M è sempre superiore, sebbene per frazioni decimali, a quella osservata nelle altre partizioni e questo per ogni valore del supporto minimo $\varepsilon$ considerato.



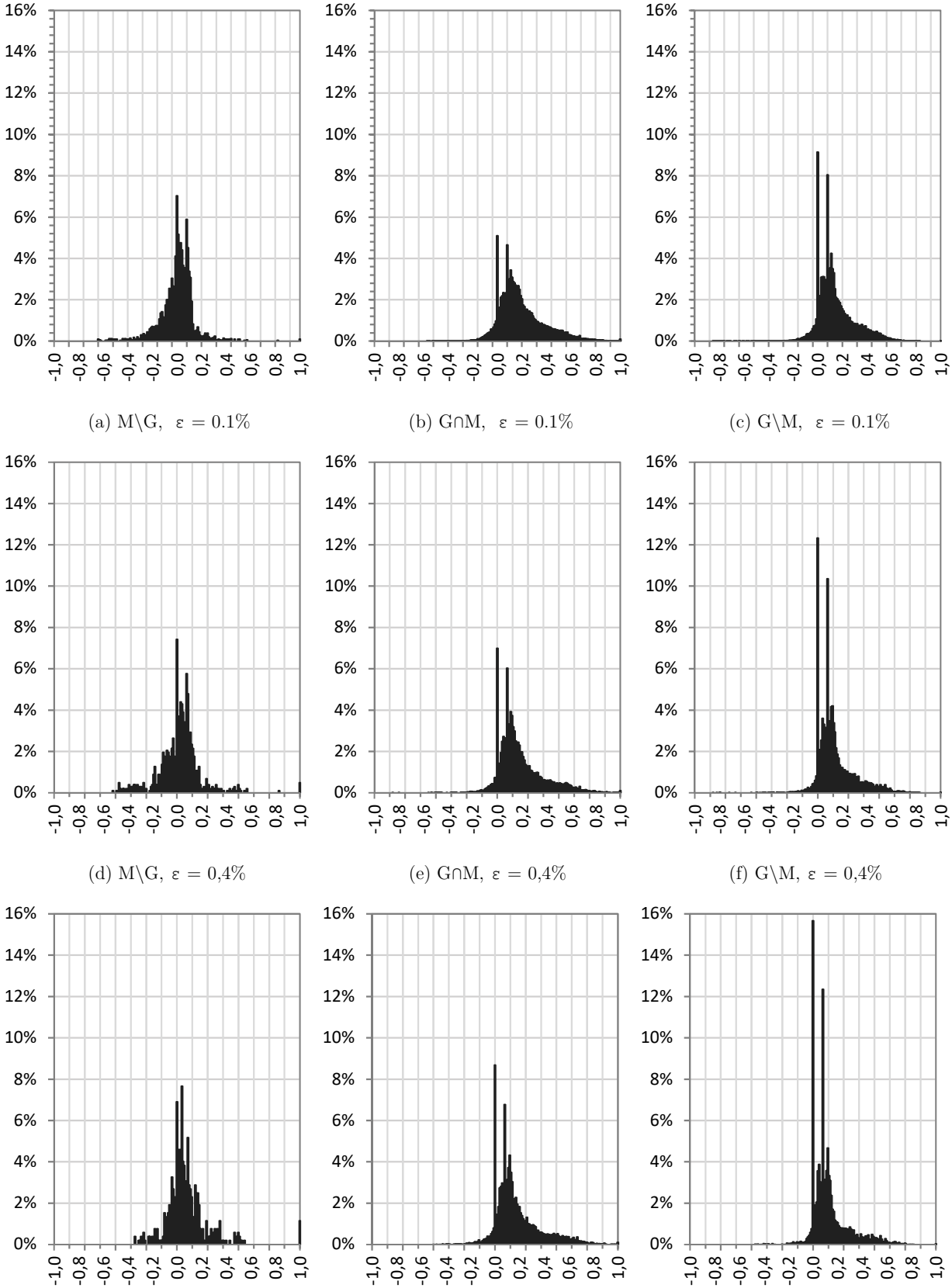

(a) M\G, $\lambda(\varepsilon)$ (b) G∩M, $\lambda(\varepsilon)$ (c) G\M , $\lambda(\varepsilon)$

**Figura 18** — *Correlazione totale normalizzata puntuale: dipendenza dei valori minimi, medi e massimi dalla soglia percentuale di supporto minimo* $\varepsilon$

Questa prima impressione è confermata dall'esame delle distribuzioni ad alta granularità delle classi rappresentate in figura 19 e figura 20, dalle quali si evincono delle ulteriori e significative caratterizzazioni e differenze tra le tre partizioni di interesse. Tutte le distribuzioni appaiono almeno bimodali, con la prima moda in un intorno del punto zero e la seconda moda in un intorno del valore 0,1 positivo.

La partizione M\G appare caratterizzata da una spiccata asimmetria negativa, con una significativa coda a sinistra del punto modale nell'intorno di zero. Mentre le altre due partizioni, per converso, sono ambedue caratterizzate da apprezzabile asimmetria positiva, con code significative (soprattutto per G∩M) a destra del punto modale positivo. Tutte e tre le partizioni manifestano una spiccata curtosi positiva, ma tra esse, quella con una curtosi nettamente più significativa è la partizione G\M, talché in questa partizione la massiccia presenza di *itemset* caratterizzati da (quasi) indipendenza statistica ne rappresenta il tratto più significativo.

Al crescere di $\varepsilon$ la partizione M\G appare vieppiù perdere la sua asimmetria, al contrario delle altre due partizioni, mentre G\M diviene sempre più leptocurtica, tanto da richiedere talvolta un cambiamento di scala dell'asse delle ordinate.

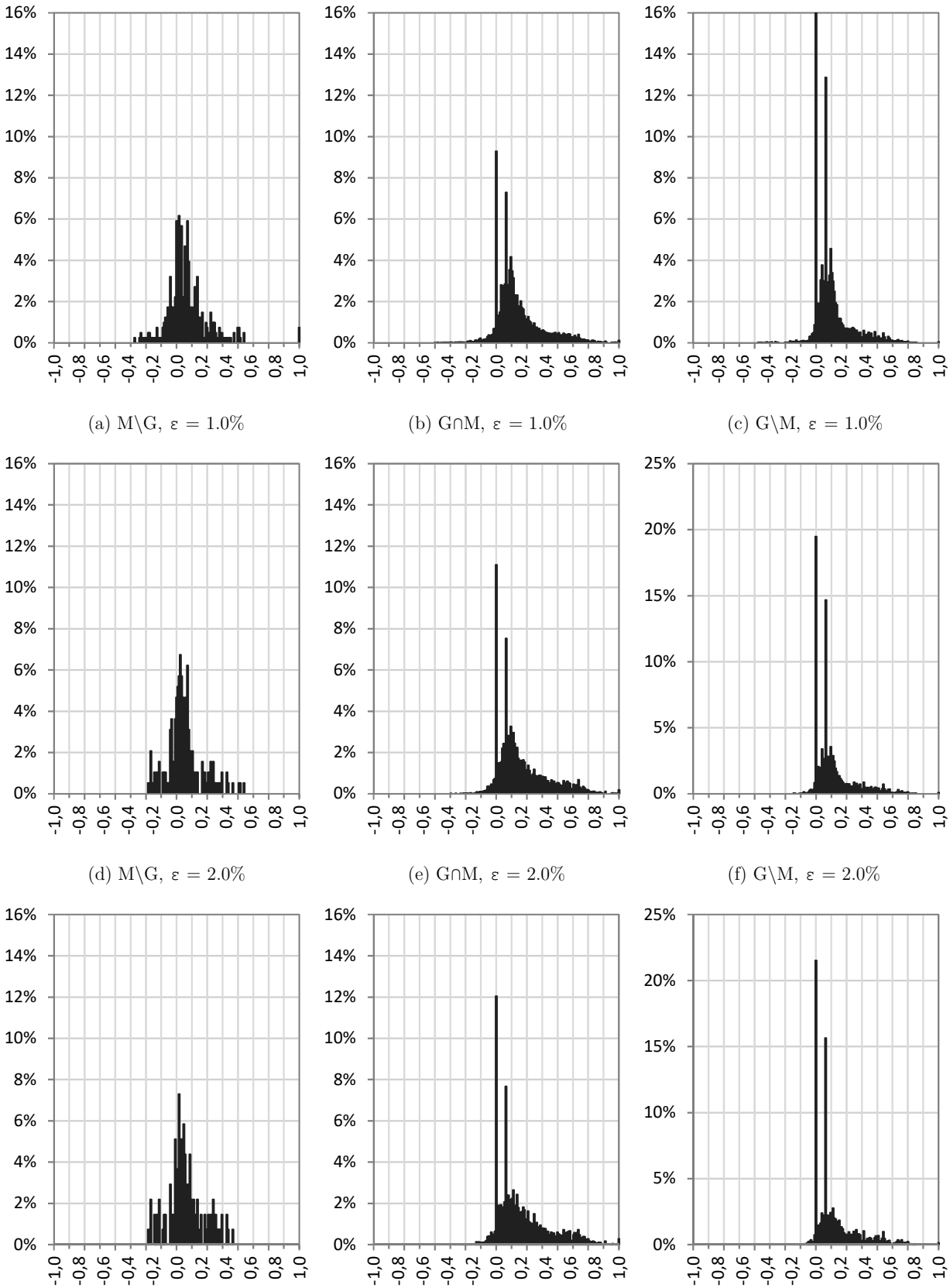



(a) M\G, $\varepsilon = 0.1\%$ (b) G∩M, $\varepsilon = 0.1\%$ (c) G\M, $\varepsilon = 0.1\%$

(d) M\G, $\varepsilon = 0{,}4\%$ (e) G∩M, $\varepsilon = 0{,}4\%$ (f) G\M, $\varepsilon = 0{,}4\%$

(g) M\G, $\varepsilon = 0{,}8\%$ (h) G∩M, $\varepsilon = 0{,}8\%$ (i) G\M, $\varepsilon = 0{,}8\%$

**Figura 19** — *Correlazione totale puntuale normalizzata: distribuzione dei valori* $\boldsymbol{\lambda}$ *per supporto minimo* $\boldsymbol{\varepsilon} < 1.0\%$




(a) M\G, $\varepsilon = 1.0\%$ (b) G∩M, $\varepsilon = 1.0\%$ (c) G\M, $\varepsilon = 1.0\%$

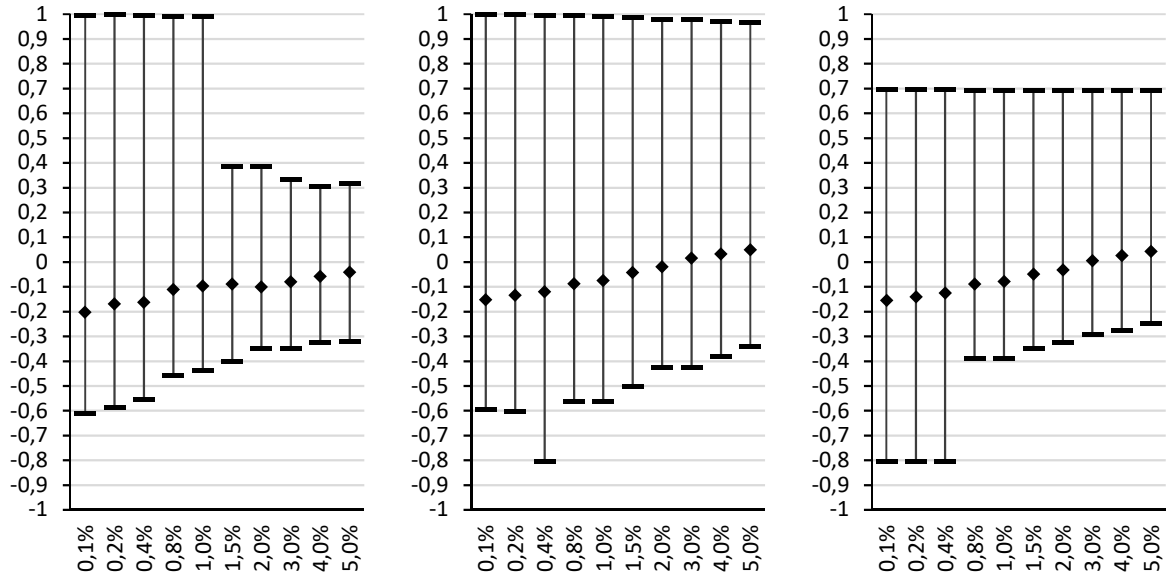
(d) M\G, $\varepsilon = 2.0\%$ (e) G∩M, $\varepsilon = 2.0\%$ (f) G\M, $\varepsilon = 2.0\%$

(g) M\G, $\varepsilon = 3.0\%$ (h) G∩M, $\varepsilon = 3.0\%$ (i) G\M, $\varepsilon = 3.0\%$

**Figura 20** — *Correlazione totale puntuale normalizzata: distribuzione dei valori* $\boldsymbol{\lambda}$ *per supporto minimo* $\boldsymbol{\varepsilon} \geq 1.0\%$

## 5.3 Analisi in base alla misura zeta

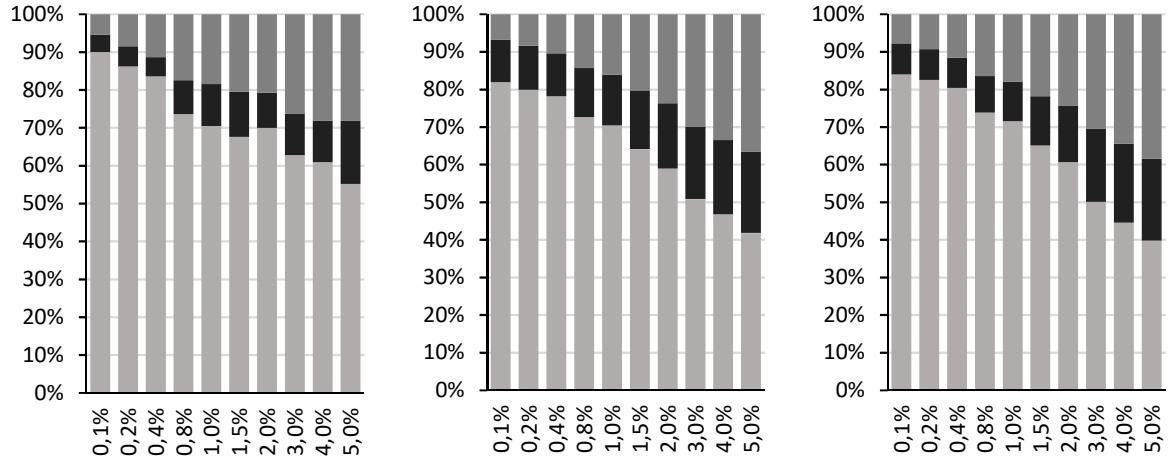
Le caratteristiche che sono state osservate nei due paragrafi precedenti si combinano significativamente nella misura *zeta*. L'analisi dei valori minimi, medi e massimi, rappresentati graficamente in figura 21, mostra, per esempio, come l'andamento del valor medio tenda a rispecchiare l'andamento della misura di Kulczynsky, mentre il supporto della misura tenda a rispecchiare invece l'andamento del supporto della *nptc*, con l'importante eccezione dei valori massimi della partizione G\M.




(a) M\G, $\zeta(\varepsilon)$ (b) G∩M, $\zeta(\varepsilon)$ (c) G\M, $\zeta(\varepsilon)$

**Figura 21** — *Misura zeta: dipendenza dei valori minimi, medi e massimi dalla soglia di supporto minimo* $\varepsilon$

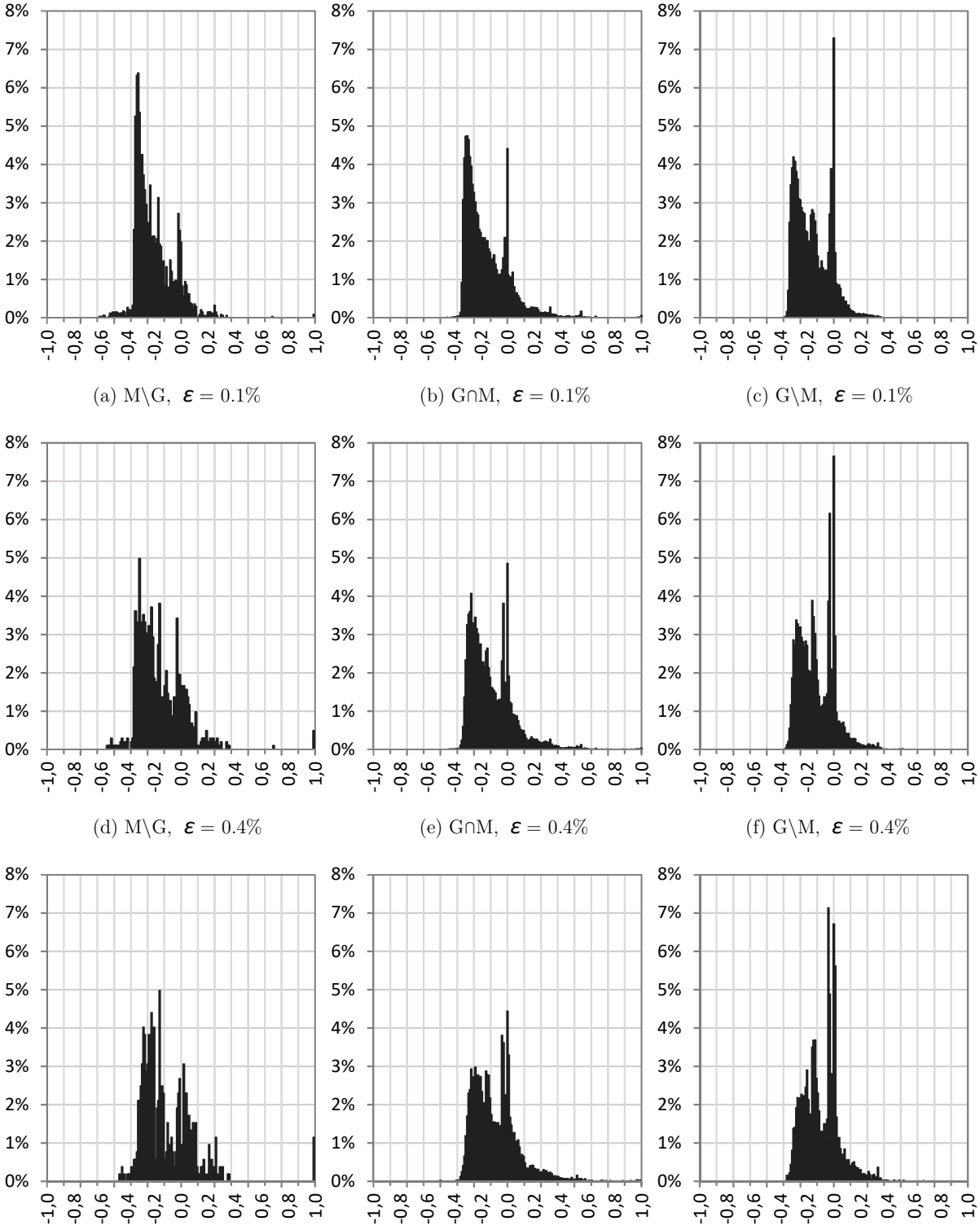
Un aspetto interessante è poi rappresentato dall'osservazione dell'eventuale passaggio di *itemset* da sopra la neutralità a sotto la neutralità e viceversa, nel confronto tra la misura di Kulczynsky e la misura *zeta*. Di ciò dà contezza la figura 22, che mostra la distribuzione degli *itemset* nelle tre classi $(\zeta \leq 0 \wedge \vartheta \leq ½)$, $(\zeta > 0 \wedge \vartheta \leq ½)$ e $(\zeta > 0 \wedge \vartheta > ½)$.




(a) M\G, $\zeta(\varepsilon) \leftrightarrow \vartheta(\varepsilon)$ (b) G∩M, $\zeta(\varepsilon) \leftrightarrow \vartheta(\varepsilon)$ (c) G\M , $\zeta(\varepsilon) \leftrightarrow \vartheta(\varepsilon)$

**Figura 22** — *Misura zeta: distribuzione degli itemset nelle tre classi, a partire dal basso:* $(\boldsymbol{\zeta \leq 0 \wedge \vartheta \leq ½})$, $(\boldsymbol{\zeta > 0 \wedge \vartheta \leq ½})$ e $(\boldsymbol{\zeta > 0 \wedge \vartheta > ½})$*, in dipendenza dalla soglia percentuale di supporto minimo* $\boldsymbol{\varepsilon}$.

Si può notare che, soprattutto in G∩M, vi è una significativa presenza di *itemset* che vengono positivamente ricollocati oltre la soglia di neutralità di $\zeta$, partendo da valori di $\vartheta$ al di sotto di essa. Sono invece rari e non significativi i casi della classe $(\zeta \leq 0 \wedge \vartheta > ½)$ che quindi non è stata neppure rappresentata.

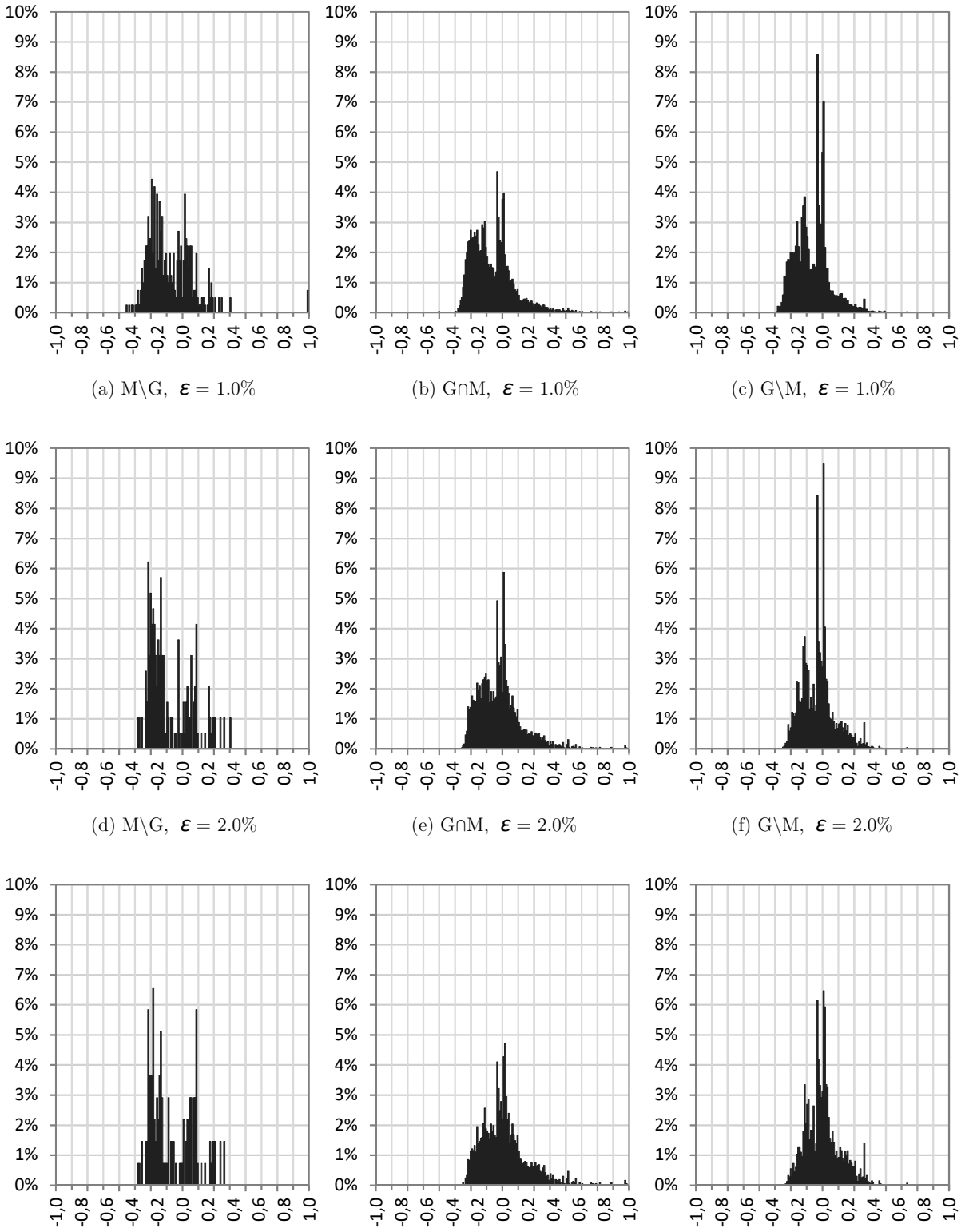



(a) M\G, $\varepsilon = 0.1\%$ (b) G∩M, $\varepsilon = 0.1\%$ (c) G\M, $\varepsilon = 0.1\%$

(d) M\G, $\varepsilon = 0.4\%$ (e) G∩M, $\varepsilon = 0.4\%$ (f) G\M, $\varepsilon = 0.4\%$

(g) M\G, $\varepsilon = 0.8\%$ (h) G∩M, $\varepsilon = 0.8\%$ (i) G\M, $\varepsilon = 0.8\%$

**Figura 23** — *Misura zeta: distribuzione dei valori per soglie di supporto minimo percentuale $\varepsilon < 1.0\%$*




(a) M\G, $\varepsilon = 1.0\%$ (b) G∩M, $\varepsilon = 1.0\%$ (c) G\M, $\varepsilon = 1.0\%$

(d) M\G, $\varepsilon = 2.0\%$ (e) G∩M, $\varepsilon = 2.0\%$ (f) G\M, $\varepsilon = 2.0\%$

(g) M\G, $\varepsilon = 3.0\%$ (h) G∩M, $\varepsilon = 3.0\%$ (i) G\M, $\varepsilon = 3.0\%$

**Figura 24** — *Misura zeta: distribuzione dei valori per soglie di supporto minimo percentuale* $\varepsilon \geq 1.0\%$

## 5.4 Caratteristiche delle regole generate

Per concludere l'analisi è certamente di interesse analizzare le caratteristiche oggettive delle regole generate dagli *itemset* presi in esame nei paragrafi precedenti, in particolare le regole generate con una soglia di confidenza $\gamma_0 = \frac{1}{2}$ a partire dagli *itemset* prodotti dal *Genio Algorithm*.

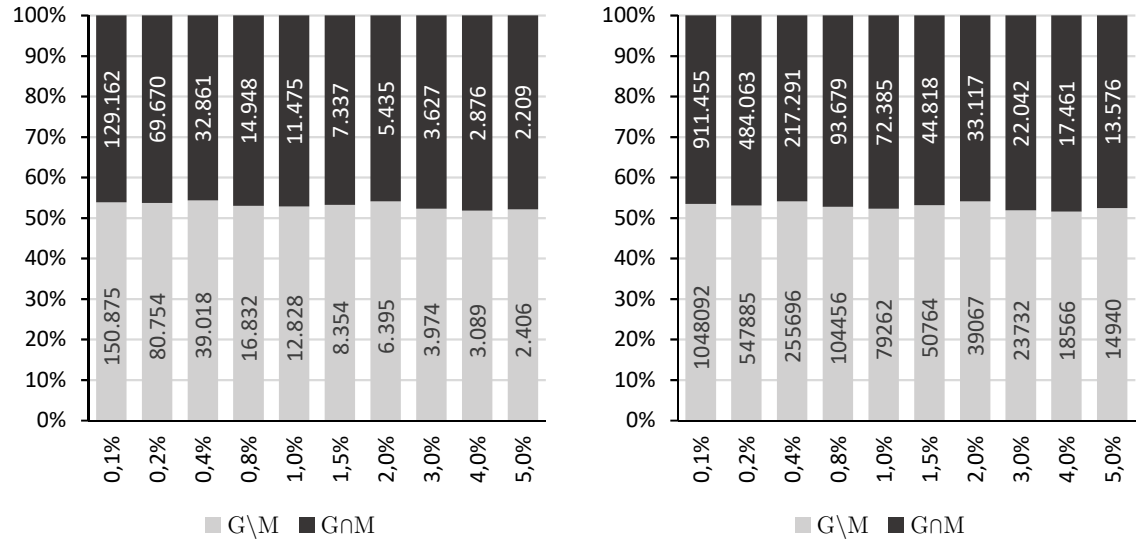
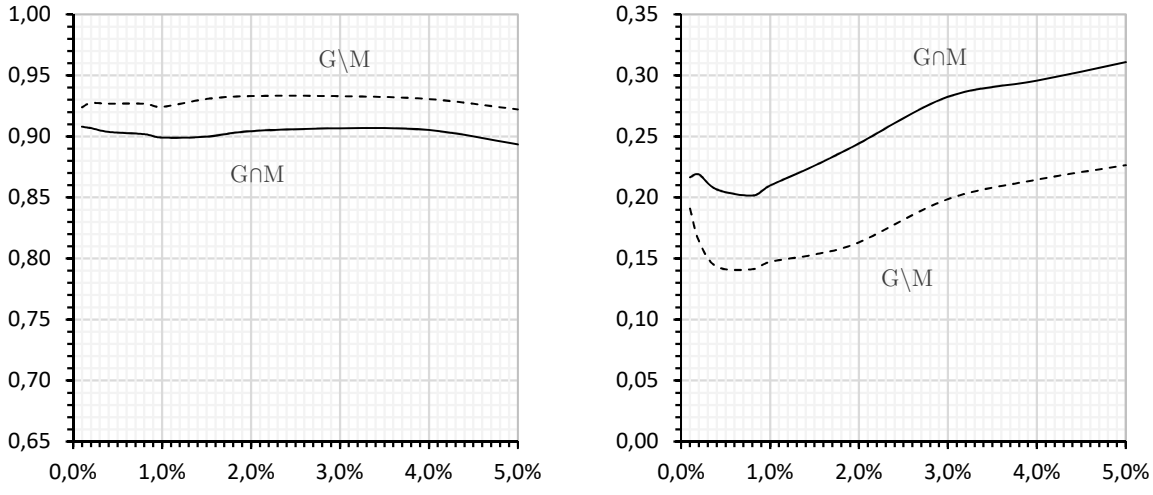



**Figura 25** — *Ripartizione degli itemset (sinistra) e delle corrispondenti regole (destra) tra le partizioni*

Come si può osservare dalla figura 25, nell'ambito delle produzioni del *Genio Algorithm* tanto gli *itemset* quanto le regole da essi generate appaiono pressoché equamente ripartiti tra le partizioni $\mathsf{G}\cap\mathsf{M}$ e $\mathsf{G}\setminus\mathsf{M}$, indipendentemente dal supporto minimo usato dall'algoritmo. Appare inoltre uniforme anche il rapporto tra regole e *itemset* che, indipendentemente dalla partizione e dal supporto minimo, risulta approssimativamente compreso tra le 6 e 7 regole per ciascun *itemset*.




(a) Confidenza media, $\gamma(\varepsilon)$ (b) Lift normalizzato medio, $\mathrm{nLift}(\varepsilon)$

**Figura 26** — *Confidenza media (a sinistra) e lift normalizzato (a destra) delle regole generate*

Invece, per quanto riguarda le caratteristiche più specificatamente inerenti alla correlazione, nella figura 26 si possono osservare differenze interessanti tra i valori della confidenza media e del

*lift* medio tra le due partizioni. Al di là dei valori assoluti in sé, quello che appare rilevante è che le regole generate dagli *itemset* in $\mathsf{G}\cap\mathsf{M}$ hanno valori della confidenza media appena lievemente più bassi rispetto a quelli generati dagli *itemset* in $\mathsf{G}\setminus\mathsf{M}$, ma per converso valori medi del *lift* sempre superiori. Tale comportamento peraltro riproduce l'analogo comportamento degli *itemset* di origine delle regole, come si può verificare rispettivamente in figura 14 e in figura 18, laddove si associ la misura di Kulczynsky alla confidenza e la correlazione puntuale normalizzata al *lift*.

Naturalmente, per rendere possibile il confronto tra insiemi con diverse soglie di supporto minimo, il *lift* rappresentato graficamente in figura 26 è la sua versione normalizzata, ovvero ristretta all'intervallo chiuso $[-1, +1]$ mediante la formula

$$\mathrm{nLift}(X \to Y) = -\frac{\ln \mathrm{lift}(X \to Y)}{\ln \mathrm{supp}(X \to Y)} \tag{5.1}$$

che rappresenta, come si è già visto nel capitolo 4, la normalizzazione per mezzo dell'autoinformazione della regola.

Un quadro molto interessante è infine delineato da un'analisi che combina da un lato il potere discriminante delle misure sugli *itemset* esposte nel § 4.11 con il loro effetto sulle regole estratte dai medesimi *itemset* una volta discriminati. L'analisi è stata svolta selezionando per ognuna delle tre misure $\vartheta$, $\lambda$ e $\zeta$ solo gli *itemset* superiori alla soglia di indifferenza di ciascuna e verificando poi la percentuale di regole utili, ovvero con *lift* superiore all'unità, corrispondentemente generate da quei soli *itemset*. In ciascuna figura risultante è pertanto possibile leggere nella scala di sinistra a che percentuale di regole utili corrisponde la percentuale di *itemset* selezionati e nella scala di destra qual è il *lift* normalizzato medio di queste regole così catturate.

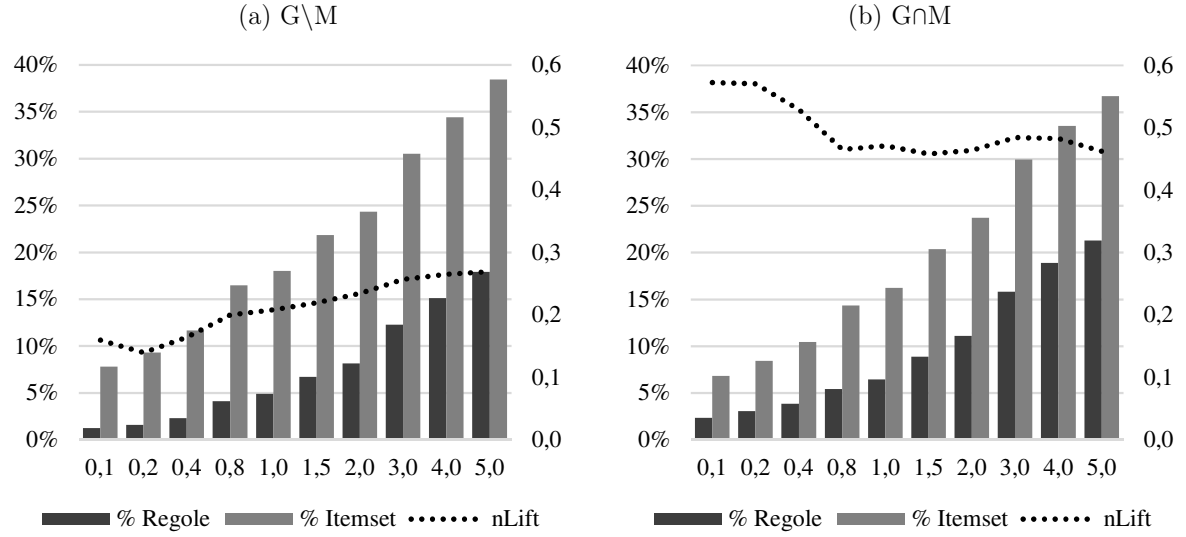



**Figura 27** — *Percentuali di itemset e regole corrispondenti, rispettivamente, a $\vartheta > ½$ e lift sopra l'unità*

Per esempio, nella figura 27(a), in corrispondenza della soglia di supporto minimo $\varepsilon = 5\%$, si legge che gli *itemset* con $\vartheta > ½$ sono circa il 38% del totale e questi *itemset* generano circa il 18% delle regole utili, le quali hanno un *lift* normalizzato medio pari a circa 0,27.

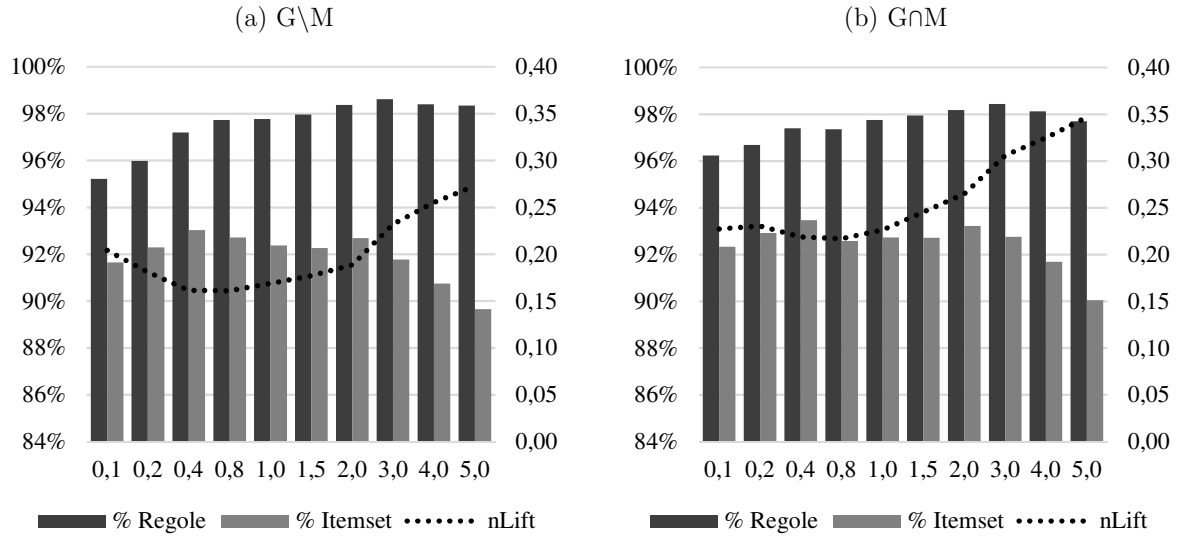



**Figura 28** — *Percentuali di itemset e regole corrispondenti, rispettivamente, a* $\lambda > 0$ *e lift sopra l'unità*

Nel mentre tutte e tre le figure confermano in modo ancor più accurato e accentuato la superiorità oggettiva degli *itemset* in $\mathsf{G} \cap \mathsf{M}$, allo stesso tempo si può osservare come la misura *zeta* sia in grado di selezionare molto più accuratamente delle altre due gli *itemset* suscettibili di generare regole fortemente correlate. Infatti, il sottoinsieme che ne viene estratto è maggiore in termini percentuali e pur tuttavia dotato di quasi ovunque migliore *lift* medio *per entrambe* le partizioni rispetto a quanto estrarrebbe la sola misura di Kulczynsky; inoltre, anche rispetto a $\boldsymbol{\lambda}$, se è vero che quest'ultima misura avrebbe il vantaggio di poter essere impiegata come filtro preselettivo in virtù del suo rapporto favorevole tra percentuale di *itemset* e percentuale di regole da questi generate, tuttavia al di fuori di tale, pur importante, applicazione il valore medio del *lift* delle regole da essa selezionate non si distanzia in modo significativo dal valore medio complessivo di tutte le regole mostrato nella figura 26(b).

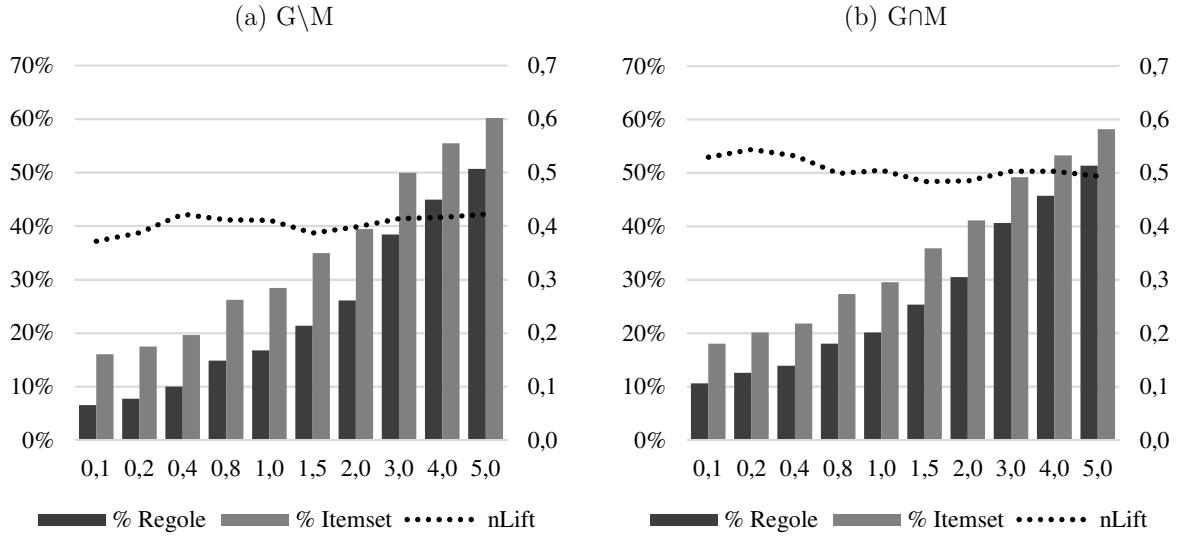



**Figura 29** — *Percentuali di itemset e regole corrispondenti, rispettivamente, a* $\zeta > 0$ *e lift sopra l'unità*

# Capitolo 6

# Analisi in base al dominio applicativo

L'analisi svolta e descritta in questo capitolo è fondata sulla conoscenza del dominio applicativo e, conseguentemente, sull'interesse qualitativo e soggettivo degli *itemset* (generalizzati (espressivi)) già analizzati e comparati con metriche quantitative nel capitolo 5. L'impiego di una metrica qualitativa, da un lato e la grande ampiezza dei dati sperimentali dall'altro, ha richiesto la definizione tanto di un criterio uniforme a priori per la valutazione, quanto di una metodologia esplorativa che, non potendo essere esaustiva, potesse perlomeno percorrere con una certa uniformità sistematica l'intero spazio dei risultati sperimentali. Per quanto riguarda il primo punto si è seguito il consolidato metodo di giudicare un *pattern* tanto più interessante quanto più inatteso (*unexpected*) oppure fungibile per trarre da esso delle conclusioni o delle azioni (*actionable*) e, nello stesso tempo, tanto meno interessante, ancorché inatteso o fungibile, quanto meno specializzato nel fornire un'informazione già resa, allo stesso livello di soglia minima di supporto, da *pattern* con minor grado di generalizzazione. Per il secondo punto si è ritenuto che il modo più efficace e meno distorsivo per esplorare i dati potesse essere mutuato dalla tecnica della stratificazione ben nota in statistica inferenziale, ovvero l'estrazione da strati omogenei di un numero limitato e finito di *itemset*. A differenza del campionamento stratificato, tuttavia, qui la scelta all'interno d'ogni strato non è avvenuta casualmente ma è stata compiuta in base a una relazione d'ordine.

## 6.1 Considerazioni metodologiche

### *6.1.1 Classi di interpretazione della co-occorrenza degli itemset*

Nell'ottica di valutare qualitativamente gli *itemset* (generalizzati (espressivi)) costituenti i risultati sperimentali, è apparso opportuno definire un criterio di interpretazione della co-occorrenza, ovvero un modello di traduzione dell'informazione resa dall'*itemset* in una qualche forma più vicina agli schemi inferenziali dell'osservatore. Le regole associative sono un esempio di traduzione di questo tipo, in quanto riorganizzano la struttura dell'*itemset* ripartendolo variamente, ma sempre dicotomicamente, e simulando così uno schema inferenziale *premesse* $\rightarrow$ *conclusioni*. Un altro modo, più rilassato ma più semplice, è quello che si potrebbe definire *pivotale*, ovvero basato sulla scelta di una caratteristica di interesse, o *pivot*, e nella valutazione della sua associazione con le altre caratteristiche, con un grado di specializzazione vieppiù crescente con la lunghezza $k$ dell'*itemset*. Anche

questo risponde ad uno schema inferenziale del potenziale osservatore che si interroga se e come, per esempio, la porta sorgente o il protocollo o l'`IP` di destinazione *et cetera*, si correlino in particolare con una o più delle rimanenti caratteristiche. In questo schema le classi d'interpretazione sono essenzialmente legate in via principale al *pivot* d'interesse scelto e, in subordine, al grado di specializzazione desiderato ovvero alla lunghezza $k$ dell'*itemset,* in uno schema bidirezionale del tipo *pivot* ⇆ *resto degli item.*

Un pericolo insito in questo modello interpretativo è quello legato alla possibile scelta di un *pivot* interessante qualitativamente a prima vista, ma suscettibile di rivelarsi in seguito privo dalla capacità di generare regole effettivamente interessanti. Non bisogna dimenticare infatti che le misure sull'*itemset* lo valutano sinteticamente e non vi è modo di controllare se ciò che consegue dal *pivot* scelto sia al di sopra o sia al di sotto di questa valutazione media; in ciò vi è una sostanziale differenza con le regole associative, che consentono invece una misurazione puntuale.

Ad ogni modo, si può facilmente ricavare che, se la lunghezza massima degli *itemset* è uguale a $n$, il numero massimo di classi ottenibili è $n^2$, asintoticamente ben minore rispetto alle possibili regole che si sanno essere uguali a $2^n - 2$. Per di più, se si considera che i casi rappresentati da $k = 1$ non sono di interesse e che il caso $k = n$ definisce $n$ sottoclassi però tutte coincidenti, le classi effettivamente rilevanti sono al più $n^2 - 2n + 1$. È tuttavia il caso di notare che le classi così definite non rappresentano delle partizioni e pertanto un *itemset* può appartenere contemporaneamente a più d'una classe. Si veda, nella tabella 8, l'esemplificazione di alcune di queste classi nel caso in cui si consideri l'ennupla di *feature* oggetto del presente lavoro.

**Tabella 8** — *Esempi di classi d'interpretazione*

| Itemset | Note |
|---|---|
| {ipsource:1.2.3.4 portdest:53}<br>{ipsource:2.3.3.4 protocol:http}<br>{ipsource:internal portsource:wellknown}≀{{ipsource:10.1.1.1 portsource:53}} | Pivot: `ipsource`, specializzazione: 2, interpretazione: dato in indirizzo IP sorgente, quale altro item lo caratterizza? |
| {portdest:53 protocol:udp_dns}<br>{portdest:wellknown ipsource:2.3.3.4}<br>{portdest:wellknown portsource:wellknown}≀{{portdest:53 portsource:53}} | Pivot: `portdest`, specializzazione: 2, interpretazione: data una porta destinazione, quale altro item la caratterizza? |
| {ipdest:1.2.3.4 portdest:53 protocol:udp_dns} | Pivot: `ipdest`, specializzazione: 3, interpretazione: dato un IP di destinazione, da quale altra coppia di item è caratterizzato? |
| {ipsource:2.3.3.4 c2s_packets: (0-50] ipdest:1.2.3.4 portdest:53 protocol:http}<br>{ipsource:2.3.3.4 c2s_packets: (0-50] ipdest:1.2.3.4 portdest:80 s2c_packets: (0-50]} | Pivot: `ipsource`, specializzazione: 5 |

Nell'analisi qualitativa condotta in questo capitolo si è fatto uso in via principale, sebbene non sempre lo si sia reso esplicito, di questa classificazione così introdotta al fine di valutare se e quanto un *itemset* appartenente ad una determinata classe fosse *unexpected* o *actionable* in senso assoluto oppure relativamente ad altri elementi della medesima classe.

### *6.1.2 Stratificazione dei dati sperimentali*

Come già accennato nelle premesse, al fine di consentire una maggiore uniformità e sistematicità nell'esplorazione dei risultati sperimentali si è mutuata la tecnica, *mutatis mutandis*, del campionamento stratificato. Gli strati sono stati ottenuti per ricombinazione e ripartizione a più stadi dei

risultati sperimentali, cosi che ogni strato fosse univocamente determinato dalla terna $\langle P, \varepsilon, k \rangle$, dove $P \in \{\mathsf{M}\backslash\mathsf{G}, \mathsf{G}\cap\mathsf{M}, \mathsf{G}\backslash\mathsf{M}\}$ è una delle partizioni definite nel § 3.4 e già ampiamente utilizzate nel capitolo 5, $k$ rappresenta la lunghezza propria dell'*itemset* e $\varepsilon$ la soglia di supporto minimo utilizzata dall'algoritmo generante. Tale scelta, oltre a garantire una significativa omogeneità, consente anche di ricondursi facilmente alle analisi oggettive condotte nel già citato capitolo 5.

È appena il caso di far rilevare che all'interno di ciascuno strato sono contemporaneamente presenti *itemset* derivanti da più *dataset* originanti, ciascuno mantenendo questa informazione di provenienza sotto forma di *itemset* virtualmente aumentato degli *item* $\langle site, value \rangle$ e $\langle day\#, value \rangle$. Complessivamente, quindi, l'insieme unione di tutti gli *itemset*, comunque generati e d'ovunque provenienti, è stato partizionato in $3 \times 10 \times 7 = 210$ strati omogenei ai fini dell'analisi.

### *6.1.3 Esplorazione degli strati*

Per ciascun strato sono stati valutati un numero finito di *itemset*, scelti in base all'ordinamento fornito dalla misura *zeta*. Il processo di valutazione si è poi svolto in questo ordine: (i) per ogni soglia $\varepsilon$ e lunghezza $k$ si sono estratti e valutati qualitativamente prima gli *itemset* comuni ai due algoritmi, ovvero quelli appartenenti allo strato identificato da $\langle \mathsf{G}\cap\mathsf{M}, \varepsilon, k \rangle$ e ciò al fine di valutare la bontà dell'applicazione di entrambe le metodologie al caso del traffico reale di rete, e quindi, (ii) si sono cercati nei due strati omologhi identificati da $\langle \mathsf{G}\backslash\mathsf{M}, \varepsilon, k \rangle$ e $\langle \mathsf{M}\backslash\mathsf{G}, \varepsilon, k \rangle$ gli eventuali *itemset* rappresentanti un valore aggiunto fornito esclusivamente dall'uno o dall'altro dei due algoritmi, e ciò al fine di valutare comparativamente i contributi originali dell'uno e dell'altro.

## 6.2 Risultati di interesse dell'analisi qualitativa

I risultati sono stati organizzati come una successione di esempi di *itemset actionable* o *unexpected*, per la maggior parte dei quali si è fornita anche la relativa interpretazione. Occasionalmente e *passim* si sono anche mostrati taluni esempi di *itemset* triviali o ridondanti o non fungibili, al solo fine di fornire elementi per comparare le produzioni dei diversi algoritmi. Le produzioni per ciascuna soglia di supporto minimo sono state trattate in paragrafi distinti, iniziando dalle soglie più alte, e ciò al fine di poter omettere quando possibile, nei paragrafi successivi, gli esempi già mostrati nei precedenti, sulla scorta della considerazione che, *in generale*, gli *itemset* presenti per una generica soglia $\varepsilon_0$ o sono presenti anche nelle soglie $\varepsilon < \varepsilon_0$ o sono ivi presenti direttamente i loro discendenti. In ogni paragrafo sono poi stati mostrati talvolta insieme talvolta separatamente i casi caratterizzati da diverse lunghezze $k$ degli *itemset*, in ordine crescente a partire da $k = 2$. Per ogni *itemset* è stato indicato il numero di *dataset* nei quali è rinvenibile, nonché i valori delle tre misure calcolate, prese dal *dataset* dove hanno il valore massimo.

### *6.2.1 Itemset con supporto maggiore o uguale a 5%*

A questo livello di supporto minimo si riesce facilmente a identificare il traffico relativo alle attività più significative della rete. Si tratta in generale di *itemset* di tipo *actionable* e solo in minima parte di itemset *unexpected*. Gli *itemset* di lunghezza $k = 2$ offrono la panoramica più completa ma

spesso necessitano di essere variamente riuniti da parte dell'osservatore per comprenderne compiutamente il significato; gli *itemset* di lunghezza $k = 3$ sono in generale autoconsistenti per quanto riguarda la loro interpretazione, mentre, per le lunghezze $k \geq 4$, solo le più brevi forniscono informazioni ancora utili.

**Tabella 9** — *Alcuni itemset di interesse negli strati di* G∩M *con* $\varepsilon = 5.0\%$

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {ipdest:224.0.0.252 portdest:5355} | 3 | 1,00 | 1,00 | 0,97 |
| 2 | {ipdest:239.255.255.253 portdest:6005} | 1 | 1,00 | 1,00 | 0,97 |
| 3 | {ipsource:172.20.90.87 ipdest:239.255.255.253 portdest:6005} | 1 | 0,94 | 0,98 | 0,89 |
| 4 | {ipsource:172.20.90.87 ipdest:239.255.255.253} | 1 | 0,92 | 0,95 | 0,86 |
| 5 | {ipsource:172.20.90.87 portdest:6005} | 1 | 0,92 | 0,95 | 0,86 |
| 6 | {ipsource:130.192.3.103 portsource:53} | 5 | 0,87 | 0,87 | 0,74 |
| 7 | {ipsource:130.192.3.103 portdest:DYNAMIC} | 5 | 0,81 | 0,81 | 0,66 |
| 8 | {ipsource:172.20.90.87 ipdest:239.255.255.253 portdest:6005 protocol:UDP_*} | 1 | 0,73 | 0,89 | 0,60 |
| 9 | {ipdest:130.192.3.103 portdest:53} | 9 | 0,83 | 0,72 | 0,58 |
| 10 | {ipdest:224.0.0.252 portdest:5355 protocol:UDP_*} | 3 | 0,72 | 0,80 | 0,58 |
| 11 | {ipdest:239.255.255.253 portdest:6005 protocol:UDP_*} | 1 | 0,70 | 0,75 | 0,55 |
| 12 | {ipdest:130.192.3.103 protocol:UDP_DNS} | 8 | 0,79 | 0,70 | 0,55 |
| 13 | {ipsource:172.20.90.87 portdest:6005 protocol:UDP_*} | 1 | 0,65 | 0,72 | 0,48 |
| 14 | {ipsource:172.20.90.87 ipdest:239.255.255.253 protocol:UDP_*} | 1 | 0,65 | 0,72 | 0,48 |
| 15 | {ipdest:130.192.3.103 portdest:53 protocol:UDP_DNS} | 8 | 0,72 | 0,71 | 0,48 |
| 16 | {ipsource:172.20.90.87 portsource:DYNAMIC ipdest:239.255.255.253} | 1 | 0,64 | 0,67 | 0,46 |
| 17 | {ipsource:130.192.3.103 portsource:53 protocol:UDP_*} | 5 | 0,66 | 0,70 | 0,45 |
| 18 | {ipdest:130.192.3.24 portdest:53} | 9 | 0,65 | 0,45 | 0,37 |
| 19 | {ipsource:130.192.3.103 portsource:WELLKNOWN portdest:DYNAMIC protocol:UDP} | 5 | 0,59 | 0,80 | 0,36 |
| 20 | {ipsource:130.192.3.103 portdest:DYNAMIC protocol:UDP} | 5 | 0,58 | 0,60 | 0,35 |
| 21 | {ipsource:130.192.3.103 protocol:UDP_*} | 6 | 0,62 | 0,43 | 0,34 |
| 22 | {portdest:5355 protocol:UDP_*} | 3 | 0,58 | 0,39 | 0,32 |
| 23 | {portdest:5223 protocol:TCP_*} | 2 | 0,59 | 0,39 | 0,32 |
| 24 | {ipsource:172.20.90.87 portsource:DYNAMIC ipdest:239.255.255.253 portdest:REGISTERED} | 1 | 0,53 | 0,73 | 0,31 |
| 25 | {ipdest:130.192.3.21 portdest:53} | 3 | 0,62 | 0,33 | 0,28 |
| 26 | {ipdest:130.192.3.24 protocol:UDP_DNS} | 8 | 0,55 | 0,41 | 0,28 |
| 27 | {ipsource:172.20.90.87 portsource:DYNAMIC ipdest:239.255.255.253 protocol:UDP} | 1 | 0,50 | 0,63 | 0,26 |
| 28 | {portdest:6005 protocol:UDP_*} | 1 | 0,55 | 0,25 | 0,23 |
| 29 | {ipdest:239.255.255.253 protocol:UDP_*} | 1 | 0,55 | 0,25 | 0,23 |
| 30 | {ipdest:130.192.3.103 protocol:UDP_*} | 6 | 0,62 | 0,27 | 0,22 |
| 31 | {ipdest:130.192.3.103 portdest:53 protocol:UDP_*} | 6 | 0,55 | 0,38 | 0,20 |
| 32 | {ipdest:130.192.3.24 protocol:UDP_*} | 1 | 0,58 | 0,23 | 0,19 |
| 33 | {ipsource:172.20.90.87 portsource:DYNAMIC ipdest:239.255.255.253 portdest:REGISTERED protocol:UDP} | 1 | 0,43 | 0,71 | 0,18 |
| 34 | {ipdest:130.192.3.21 protocol:UDP} | 2 | 0,56 | 0,18 | 0,18 |
| 35 | {ipdest:130.192.3.21 protocol:UDP_*} | 1 | 0,56 | 0,21 | 0,17 |
| 36 | {ipdest:130.192.3.24 portdest:53 protocol:UDP_DNS} | 8 | 0,43 | 0,46 | 0,16 |
| 37 | {ipdest:130.192.3.24 portdest:53 protocol:UDP_*} | 1 | 0,47 | 0,33 | 0,13 |

**Esempio 1** (*Individuazione di server DNS locali*) La prima immediata informazione che si può trarre dai risultati mostrati in tabella 9 è la presenza di tre DNS di ateneo a servizio della rete:

130.192.3.103, 130.192.3.24 e 130.192.3.21. Questa informazione, a diversi livelli di granularità e di *dataset* d'origine, è catturata dalla maggior parte degli *itemset*, i quali, fra l'altro, con un differente valore di *zeta* loro associato, evidenziano altresì un chiaro sbilanciamento nella distribuzione del carico, che appare gravare soprattutto sull'*host* 130.192.3.103.

**Esempio 2** (*Individuazione di traffico anomalo*) Un primo e molto significativo esempio di traffico anomalo rilevato è quello originato da un singolo *host* in un luogo e giorno specifico, che appare anomalo in quanto utilizza una porta e un protocollo (6005/UDP) solitamente associati al servizio grafico «*X Window System*» e, nello stesso tempo, un indirizzo *multicast* (239.255.255.253) ufficialmente associato al «*Service Location Protocol (SLP)*» che tuttavia di norma utilizza altre porte. Si tratta quindi d'una combinazione inusuale che dovrebbe essere indagata ulteriormente da un amministratore di rete.

**Esempio 3** (*Evidenze di protocolli e servizi specifici e diffusi*) Un esempio ben caratterizzato è il rilevamento di attività del protocollo «*Link-Local Multicast Name Resolution*» sulla porta 5355 che viene utilizzato soprattutto dai sistemi operativi Windows desktop e server per la risoluzione dei nomi locali. Analogamente, la rilevazione della porta (5223) e del protocollo del servizio notifiche *push* di Apple Inc. fa palese l'attività sulla rete di apparati di questo *vendor*.

Tutto quanto sopra esposto è derivato dall'analisi degli strati comuni ai due algoritmi. Se si analizzano invece gli strati omologhi complementari, si rileva che negli strati tratti da G\M non vi è alcun *itemset* di maggior interesse rispetto alla parte comune già esplorata, giacché ivi si trovano solo casi ridondanti ovvero caratterizzati da una generalizzazione inutilmente più ampia; mentre negli strati tratti da M\G vi sono bensì *itemset* aventi un contenuto informativo originale, ma solo pochi di essi offrono una interpretabilità realmente fungibile. Tali conclusioni, che valgono anche per i livelli di supporto minimo inferiori, verranno discussi più approfonditamente nel successivo § 6.3, al quale si rimanda.

### *6.2.2 Itemset con supporto maggiore o uguale a 4% e minore di 5%*

Anche a questo livello di supporto minimo valgono per intero le considerazioni già espresse per il livello immediatamente superiore. Rispetto a quest'ultimo e al netto degli *itemset* presenti in entrambi e qui omessi, emergono solo poche informazioni aggiuntive, ma nondimeno interessanti.

**Tabella 10** *Alcuni itemset di interesse negli strati di* G∩M *con* $\boldsymbol{\varepsilon = 4.0\%}$

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {ipsource:172.20.91.243 portsource:18012} | 1 | 0,96 | 0,97 | 0,92 |
| 2 | {ipsource:172.20.90.167 portdest:82} | 1 | 0,91 | 0,94 | 0,86 |
| 3 | {ipsource:172.20.91.243 portsource:18012 protocol:UDP_*} | 1 | 0,67 | 0,74 | 0,52 |
| 4 | {ipsource:172.20.90.167 portdest:82 protocol:TCP_*} | 1 | 0,64 | 0,72 | 0,48 |
| 5 | {portdest:7 protocol:UDP_*} | 2 | 0,56 | 0,30 | 0,26 |
| 6 | {ipsource:172.20.91.243 portdest:REGISTERED} | 1 | 0,52 | 0,37 | 0,26 |
| 7 | {portsource:18012 protocol:UDP_*} | 1 | 0,55 | 0,28 | 0,25 |
| 8 | {portdest:82 protocol:TCP_*} | 1 | 0,55 | 0,26 | 0,24 |
| 9 | {ipsource:172.20.91.243 protocol:UDP_*} | 1 | 0,54 | 0,27 | 0,23 |
| 10 | {ipsource:172.20.90.167 protocol:TCP_*} | 1 | 0,50 | 0,23 | 0,17 |
| 11 | {ipsource:172.20.91.243 portsource:REGISTERED portdest:REGISTERED} | 1 | 0,42 | 0,50 | 0,17 |

**Esempio 4** (*Individuazione di traffico riferibile a malware ben caratterizzati*) In tabella 10 ci sono esempi di *itemset* a un tempo *unexpected* e *actionable*. Esse mostrano la possibile presenza di un *host* infettato dal *malware W32.Netsky* che tipicamente utilizza la porta `82/TCP` per ricevere occasionalmente comandi dall'attaccante. Questa porta ufficialmente sarebbe assegnata al servizio «*XFER utility*» per il trasferimento di zone `DNS` e talvolta, sebbene raramente, potrebbe essere utilizzata anche come porta alternativa pel protocollo `http`. Tuttavia, la marginalità di questi ultimi due casi, l'elevata misura di interesse riscontrata e l'individuazione d'uno specifico *host* dovrebbero orientare un amministratore di rete verso l'intraprendimento di misure opportune.

**Esempio 5** (*Individuazione di traffico sconosciuto*) Altre righe in tabella 10 mostrano un traffico rilevante da parte di un singolo *host* su di una propria porta ufficialmente non assegnata (`18012/UDP`) e non nota per servizi o protocolli conosciuti e diffusi. A questo livello non vi sono altri elementi che permettano di caratterizzare meglio questo traffico sconosciuto, se non l'indicazione dell'opportunità di ulteriori indagini.

### *6.2.3 Itemset con supporto maggiore o uguale a 3% e minore di 4%*

A questo livello di supporto minimo — e poi in misura vieppiù crescente nei livelli sottostanti — inizia a manifestarsi il passaggio dalle semplici *evidenze* di protocolli o servizi specifici e diffusi alla *individuazione* degli *host* locali relati a questi protocolli o servizi specifici e diffusi e ciò si vede facilmente confrontando la tabella 11 con l'esempio 3 *supra*.

**Tabella 11** — *Alcuni itemset di interesse negli strati* $\langle \mathtt{G}\cap\mathtt{M}, \mathbf{3\%}, \mathbf{2}\rangle$ e $\langle \mathtt{G}\cap\mathtt{M}, \mathbf{3\%}, \mathbf{3}\rangle$

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {ipsource:172.20.91.202 portdest:5223} | 1 | 0,58 | 0,66 | 0,39 |
| 2 | {ipsource:172.20.90.95 portdest:5223} | 1 | 0,57 | 0,66 | 0,38 |
| 3 | {ipsource:172.20.90.95 portdest:5223 protocol:TCP_*} | 1 | 0,41 | 0,57 | 0,18 |
| 4 | {ipsource:172.20.91.202 portdest:5223 protocol:TCP_*} | 1 | 0,41 | 0,55 | 0,17 |

Si deve notare che, in realtà, ci si sarebbe attesa una progressione in tre fasi: *evidenze* → *individuazione di sottoreti* → *individuazione di host*, ma questo non è avvenuto né in questo caso né in altri. In altre parole, la generalizzazione sugli indirizzi `IPv4` che avrebbe dovuto aversi in virtù della tassonomia non ha avuto occasione di esprimersi in alcuna occasione, con conseguenze non irrilevanti sulla qualità dei risultati.

### *6.2.4 Itemset con supporto maggiore o uguale a 2% e minore di 3%*

A questo livello, oltre all'intensificazione delle individuazioni già citate *supra*, e, *more solito*, qui e nel seguito non riportate, cominciano ad emergere evidenze relative a `DNS` non locali e a traffico cagionato da *malware* debolmente caratterizzato, ovvero facente uso di porte e protocolli standard e riconoscibile solo per via dell'indirizzo `IP` dell'*host* remoto.

**Esempio 6** (*Individuazione di server DNS non locali*) Le righe della tabella 12 mostrano, anche se debolmente, le evidenze dell'utilizzo di `DNS` non d'ateneo da parte d'uno o più *host* locali. Si tratta in particolare del `DNS` server `google-public-dns-a.google.com` che è ben noto e diffusamente utilizzato. In questo caso non vi sono ragioni di allarme, ma la possibilità di rilevare `DNS` spuri è certamente di ausilio per il contrasto di numerosi attacchi informatici.

**Esempio 7** (*Evidenze di traffico riferibile a malware debolmente caratterizzati*) Le righe della tabella 12 mostrano l'evidenza, ma senza individuazione, di uno o più *host* che verosimilmente eseguono il processo nascosto «*Pandora*», ben noto come *malware* e installato di solito con il riproduttore video *open-source* «*KMPlayer*». Si noti che il traffico in sé non si distinguerebbe dal normale traffico `http` se non fosse per il particolare indirizzo `IP` di destinazione (`111.111.111.111`) che lo caratterizza.

**Tabella 12** — *Alcuni itemset di interesse negli strati* ⟨G∩M, **2%**, **2**⟩ e ⟨G∩M, **2%**, **3**⟩

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {ipdest:111.111.111.111 portdest:80} | 1 | 0,61 | 0,62 | 0,45 |
| 2 | {ipsource:172.20.90.20 portdest:7} | 1 | 0,59 | 0,70 | 0,42 |
| 3 | {ipdest:8.8.8.8 protocol:UDP_*} | 1 | 0,53 | 0,25 | 0,23 |
| 4 | {ipdest:8.8.8.8 portdest:53} | 1 | 0,53 | 0,24 | 0,23 |
| 5 | {ipdest:172.20.91.255 protocol:UDP_*} | 2 | 0,53 | 0,22 | 0,21 |
| 6 | {ipdest:111.111.111.111 protocol:TCP_*} | 1 | 0,52 | 0,21 | 0,21 |
| 7 | {ipdest:111.111.111.111 portdest:80 protocol:TCP_*} | 1 | 0,43 | 0,52 | 0,20 |
| 8 | {portsource:DYNAMIC ipdest:111.111.111.111} | 1 | 0,51 | 0,09 | 0,11 |
| 9 | {portsource:DYNAMIC ipdest:8.8.8.8} | 1 | 0,52 | 0,07 | 0,09 |
| 10 | {ipdest:8.8.8.8 portdest:53 protocol:UDP_*} | 1 | 0,37 | 0,31 | 0,09 |

### *6.2.5 Itemset con supporto maggiore o uguale a 1.5% e minore di 2%*

Come in tutti i casi di passaggio da un livello di supporto minimo a uno inferiore, anche qui si manifesta il passaggio dalle semplici evidenze alle individuazioni più specifiche. È il caso per esempio dell'individuazione dei *client* che fanno uso di `DNS` non locali. Iniziano inoltre a manifestarsi le evidenze di applicativi e processi software ben caratterizzati.

**Tabella 13** — *Alcuni itemset di interesse negli strati* ⟨G∩M, 1.5%, **2**⟩, ⟨G∩M, 1.5%, **3**⟩ e ⟨G∩M, 1.5%, **4**⟩

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {portsource:17500 portdest:17500} | 1 | 1,00 | 1,00 | 0,99 |
| 2 | {portsource:17500 ipdest:172.20.91.255 portdest:17500} | 1 | 0,71 | 0,89 | 0,60 |
| 3 | {portsource:17500 portdest:17500 protocol:UDP_*} | 1 | 0,68 | 0,73 | 0,55 |
| 4 | {portsource:17500 ipdest:172.20.91.255} | 1 | 0,66 | 0,79 | 0,53 |
| 5 | {ipdest:172.20.91.255 portdest:17500} | 1 | 0,66 | 0,79 | 0,53 |
| 6 | {ipsource:172.20.91.37 portdest:123} | 1 | 0,66 | 0,78 | 0,53 |
| 7 | {ipsource:172.20.91.24 ipdest:8.8.4.4} | 1 | 0,62 | 0,68 | 0,46 |
| 8 | {portsource:17500 ipdest:172.20.91.255 portdest:17500 protocol:UDP_*} | 1 | 0,54 | 0,84 | 0,36 |
| 9 | {portdest:1900 protocol:UDP_*} | 1 | 0,53 | 0,31 | 0,27 |
| 10 | {ipsource:172.20.91.24 ipdest:8.8.8.8} | 1 | 0,47 | 0,59 | 0,26 |
| 11 | {ipdest:172.20.91.255 portdest:17500 protocol:UDP_*} | 1 | 0,45 | 0,61 | 0,24 |
| 12 | {portsource:17500 ipdest:172.20.91.255 protocol:UDP_*} | 1 | 0,45 | 0,61 | 0,24 |
| 13 | {ipsource:172.20.91.37 portdest:123 protocol:UDP_*} | 1 | 0,45 | 0,60 | 0,24 |
| 14 | {portdest:389 protocol:UDP_*} | 1 | 0,52 | 0,24 | 0,24 |
| 15 | {ipdest:8.8.4.4 protocol:UDP_*} | 1 | 0,52 | 0,22 | 0,22 |
| 16 | {ipdest:8.8.4.4 portdest:53} | 1 | 0,52 | 0,22 | 0,22 |
| 17 | {ipsource:172.20.91.24 ipdest:8.8.4.4 protocol:UDP_*} | 1 | 0,42 | 0,56 | 0,20 |
| 18 | {ipsource:172.20.91.24 ipdest:8.8.4.4 portdest:53} | 1 | 0,42 | 0,55 | 0,20 |
| 19 | {portdest:17500 protocol:UDP_*} | 1 | 0,52 | 0,20 | 0,20 |
| 20 | {portsource:17500 protocol:UDP_*} | 1 | 0,52 | 0,20 | 0,20 |

**Esempio 8** (*Individuazione di client di DNS non locali*) Numerosi *itemset* nella tabella 13 individuano ora almeno un *host* che utilizza server `DNS` non d'ateneo. Tra questi vi è il `DNS` server di cui vi erano evidenze già al livello precedente e, inoltre, anche il suo omologo secondario `8.8.4.4`, ovvero `google-public-dns-b.google.com`.

**Esempio 9** (*Evidenze di applicativi e processi software ben caratterizzati*) La riga 8 della tabella 13 offre la completa caratterizzazione, unitamente a numerosi *itemset* di lunghezza inferiore, dell'attività del *Dropbox LanSync Protocol (db-lsp)* che viene utilizzato dall'omonima applicazione *Dropbox* per sincronizzare i cataloghi dei file tra i *client* presenti sulla medesima rete locale. Analogamente, le righe 9 e 14 mostrano le prime evidenze, rispettivamente, dell'attività dello *SSDP (Simple Service Discovery Protocol)*, ovvero il protocollo utilizzato dai sistemi Windows per la segnalazione e la scoperta delle periferiche *UPnP*, e dell'attività dello *LDAP (Lightweight Directory Access Protocol)* massicciamente utilizzato dai sistemi Windows per l'accesso ad Active Directory, ma anche, in generale e da qualunque sistema, per l'accesso a elenchi di e-mail, contatti e rubriche.

### *6.2.6 Itemset con supporto maggiore o uguale a 1.0% e minore di 1.5%*

Si manifestano a questo livello sia alcune identificazioni di applicativi e processi software di cui erano comparse le evidenze nei livelli precedenti sia la scoperta di nuove evidenze. È inoltre interessante la comparsa di *itemset* utili a caratterizzare ciò che in precedenza appariva semplicemente come traffico anomalo o sconosciuto.

**Tabella 14** — *Alcuni itemset di interesse nello strato* ⟨G∩M, **1%**, **2**⟩

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {ipsource:172.20.90.76 ipdest:239.255.255.253} | 1 | 0,95 | 0,97 | 0,92 |
| 2 | {ipsource:172.20.90.76 portdest:6005} | 1 | 0,95 | 0,97 | 0,92 |
| 3 | {ipdest:239.255.255.177 portdest:1900} | 1 | 0,82 | 0,90 | 0,75 |
| 4 | {ipdest:255.255.255.255 portdest:10505} | 1 | 0,79 | 0,88 | 0,71 |
| 5 | {ipsource:172.20.91.127 portdest:123} | 1 | 0,78 | 0,89 | 0,70 |
| 6 | {ipsource:172.20.33.121 ipdest:111.111.111.111} | 1 | 0,76 | 0,87 | 0,67 |
| 7 | {ipsource:172.20.91.147 portdest:10505} | 1 | 0,73 | 0,83 | 0,63 |
| 8 | {ipsource:172.20.91.129 portdest:8612} | 1 | 0,70 | 0,79 | 0,59 |
| 9 | {portsource:18012 portdest:4672} | 1 | 0,63 | 0,69 | 0,48 |
| 10 | {ipsource:172.20.91.243 portdest:4672} | 1 | 0,60 | 0,68 | 0,45 |
| 11 | {ipsource:172.20.91.202 portdest:8612} | 1 | 0,60 | 0,61 | 0,43 |
| 12 | {ipsource:172.20.91.147 ipdest:255.255.255.255} | 1 | 0,53 | 0,72 | 0,36 |
| 13 | {ipsource:172.20.91.78 ipdest:111.111.111.111} | 1 | 0,48 | 0,68 | 0,29 |
| 14 | {portdest:8612 protocol:UDP_*} | 3 | 0,53 | 0,30 | 0,27 |
| 15 | {ipdest:239.255.255.177 protocol:UDP_*} | 1 | 0,52 | 0,28 | 0,26 |
| 16 | {portdest:5228 protocol:TCP_*} | 1 | 0,52 | 0,24 | 0,24 |
| 17 | {portdest:5242 protocol:TCP_*} | 1 | 0,52 | 0,24 | 0,24 |
| 18 | {portdest:4672 protocol:UDP_*} | 1 | 0,51 | 0,19 | 0,21 |
| 19 | {portdest:10505 protocol:UDP_*} | 1 | 0,51 | 0,17 | 0,19 |
| 20 | {ipsource:172.20.68.127 portdest:33033} | 1 | 0,41 | 0,50 | 0,19 |
| 21 | {ipsource:172.20.91.212 portdest:2987} | 1 | 0,40 | 0,56 | 0,18 |

Come di consueto, gli *itemset* di lunghezza due, mostrati in tabella 14, colgono con maggiore articolazione il contenuto informativo del livello, ma richiedono un'attività di ricomposizione da

parte dell'osservatore. Si può notare come vengano qui identificati alcuni *host* verosimilmente affetti dal *malware* «*Pandora*», di cui vi erano palesate le mere evidenze nei livelli superiori.

**Tabella 15** — *Alcuni itemset di interesse negli strati* ⟨G∩M, **1%**, **3**⟩ e ⟨G∩M, **1%**, **4**⟩

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {ipsource:172.20.90.76 ipdest:239.255.255.253 portdest:6005} | 1 | 0,96 | 0,99 | 0,94 |
| 2 | {ipsource:172.20.90.76 ipdest:239.255.255.253 portdest:6005 protocol:UDP_*} | 1 | 0,73 | 0,89 | 0,63 |
| 3 | {ipsource:172.20.91.147 ipdest:255.255.255.255 portdest:10505} | 1 | 0,68 | 0,87 | 0,56 |
| 4 | {ipsource:172.20.90.76 ipdest:239.255.255.253 protocol:UDP_*} | 1 | 0,64 | 0,72 | 0,50 |
| 5 | {ipsource:172.20.90.76 portdest:6005 protocol:UDP_*} | 1 | 0,64 | 0,72 | 0,50 |
| 6 | {ipsource:172.20.90.76 portsource:DYNAMIC ipdest:239.255.255.253} | 1 | 0,64 | 0,68 | 0,49 |
| 7 | {ipdest:239.255.255.177 portdest:1900 protocol:UDP_*} | 1 | 0,56 | 0,70 | 0,40 |
| 8 | {ipsource:172.20.33.121 ipdest:111.111.111.111 portdest:80} | 1 | 0,53 | 0,73 | 0,36 |
| 9 | {ipdest:255.255.255.255 portdest:10505 protocol:UDP_*} | 1 | 0,53 | 0,65 | 0,36 |
| 10 | {ipsource:172.20.91.147 ipdest:255.255.255.255 portdest:10505 protocol:UDP_*} | 1 | 0,51 | 0,83 | 0,34 |
| 11 | {ipsource:172.20.33.121 ipdest:111.111.111.111 protocol:TCP_*} | 1 | 0,52 | 0,66 | 0,34 |
| 12 | {ipsource:172.20.33.121 portsource:DYNAMIC ipdest:111.111.111.111} | 1 | 0,51 | 0,61 | 0,32 |
| 13 | {ipsource:172.20.91.147 portdest:10505 protocol:UDP_*} | 1 | 0,49 | 0,62 | 0,30 |
| 14 | {ipsource:172.20.91.243 portsource:18012 portdest:4672} | 1 | 0,48 | 0,71 | 0,30 |
| 15 | {ipsource:172.20.91.129 portdest:8612 protocol:UDP_*} | 1 | 0,48 | 0,60 | 0,28 |
| 16 | {ipsource:172.20.90.76 portsource:DYNAMIC ipdest:239.255.255.253 protocol:UDP} | 1 | 0,48 | 0,62 | 0,28 |
| 17 | {portsource:18012 portdest:4672 protocol:UDP_*} | 1 | 0,43 | 0,55 | 0,21 |
| 18 | {ipsource:172.20.91.202 portdest:8612 protocol:UDP_*} | 1 | 0,42 | 0,54 | 0,19 |
| 19 | {ipsource:172.20.91.243 portdest:4672 protocol:UDP_*} | 1 | 0,41 | 0,54 | 0,19 |
| 20 | {ipsource:172.20.33.121 ipdest:111.111.111.111 portdest:80 protocol:TCP_*} | 1 | 0,41 | 0,74 | 0,19 |
| 21 | {ipsource:172.20.91.243 portsource:18012 portdest:4672 protocol:UDP_*} | 1 | 0,37 | 0,75 | 0,14 |
| 22 | {ipsource:172.20.91.147 ipdest:255.255.255.255 protocol:UDP_*} | 1 | 0,36 | 0,56 | 0,13 |
| 23 | {ipsource:172.20.91.78 ipdest:111.111.111.111 portdest:80} | 1 | 0,36 | 0,65 | 0,13 |
| 24 | {ipsource:172.20.91.147 portsource:DYNAMIC ipdest:255.255.255.255} | 1 | 0,35 | 0,52 | 0,11 |
| 25 | {ipsource:172.20.91.78 ipdest:111.111.111.111 protocol:TCP_*} | 1 | 0,33 | 0,54 | 0,09 |
| 26 | {ipsource:172.20.91.78 portsource:DYNAMIC ipdest:111.111.111.111} | 1 | 0,33 | 0,51 | 0,08 |
| 27 | {ipsource:172.20.91.212 portdest:2987 protocol:TCP_*} | 1 | 0,29 | 0,51 | 0,03 |
| 28 | {ipsource:172.20.91.78 ipdest:111.111.111.111 portdest:80 protocol:TCP_*} | 1 | 0,27 | 0,70 | 0,02 |

**Esempio 10** (*Identificazione di applicativi e processi software ben caratterizzati*) Le righe 3, 9 e 13 della tabella 15, per esempio, offrono la completa caratterizzazione dell'attività su di un *host* locale dell'applicativo software *BlueStacks*, un emulatore del S.O. Android per piattaforme PC e Mac. Si può identificare anche l'attività (righe 8 e 11 di tabella 14 e 18 di tabella 15) di un processo legato alla connettività in rete locale di talune stampanti, «*Canon BJNP Port 2*» e, infine, anche l'identificazione di un *host* relato all'attività dell'applicativo *Skype*. Vi sono inoltre le evidenze del traffico verso il *Google Playstore (Android market).*

**Esempio 11** (*Caratterizzazione di traffico precedentemente rilevato come sconosciuto*) Nell'esempio 5 si era individuato un significativo traffico di natura non ben definita. A questo livello di supporto minimo si ha invece una più completa caratterizzazione di quel traffico: le righe 9, 10 e 18 di tabella 14 e 14, 19 e 21 di tabella 15 permettono ora di asserire con ragionevole sicurezza che si trattava di traffico legato all'applicativo di file sharing *eMule*, in ragione della concordante associazione tra l'indirizzo `IP`, le porte `4672` e `18012` e il protocollo `UDP_*`

**Esempio 12** (*Individuazione di host affetti da malware debolmente caratterizzati*) Numerose righe della tabella 14 e della tabella 15 mostrano l'individuazione specifica di *host* che verosimilmente eseguono il processo nascosto «*Pandora*», caratterizzato dall'`IP` di destinazione `111.111.111.111`, le cui evidenze si erano già palesate nel precedente esempio 7.

### *6.2.7 Itemset con supporto maggiore o uguale a 0,8% e minore di 1%*

Per questo livello di supporto minimo non v'è emersione di nuovi *itemset* interessanti per l'analisi qualitativa, ma semplicemente l'intensificazione dei passaggi da evidenze a identificazioni di *host* e la progressiva estensione a ulteriori *dataset* delle rilevazioni precedenti, come accade per ogni discesa nel livello di supporto minimo. È da notare, peraltro, la relativamente piccola differenza tra il valore di supporto minimo di questo livello e quello precedente.

### *6.2.8 Itemset con supporto maggiore o uguale a 0,4% e minore di 0,8%*

A partire da questo livello di supporto minimo la granularità degli *itemset* interessanti sotto il profilo qualitativo diviene molto elevata, talché verranno discussi come esempi, da qui in avanti, solamente quelli maggiormente significativi, per non appesantire inutilmente la trattazione. Ciò che appare interessante è soprattutto il fatto che qui iniziano a manifestarsi le evidenze di un protocollo di servizio e di una interfaccia di programmazione suscettibili entrambi di notevoli implicazioni sulla funzionalità e vulnerabilità della rete: ovvero il `DHCP` e il `NetBIOS`.

Si può osservare, comparando le dimensioni della tabella 16 con quelle della tabella 17, che la significatività degli *itemset* di lunghezza maggiore di due inizia ad essere più marcata rispetto a quanto accadeva nei livelli di supporto minimo superiore. Da questo livello di supporto minimo a scendere, infatti, l'esame dei 2-*itemset* risulta molto dispersivo per un esperto del dominio, giacché l'informazione di interesse risulta eccessivamente frammentata. Vi è inoltre da rilevare che da questo livello di supporto minimo la gran parte degli *itemset* contiene *item* rappresentanti gli indirizzi sorgente e destinazione, ragion per cui, sul piano interpretativo, le caratterizzazioni includenti lo `IP` sorgente e lo `IP` di destinazione vengono vieppiù a prevalere sulle caratterizzazioni basate in via principale sulle porte e sui protocolli.

**Esempio 13** (*Individuazione di server DHCP*) La maggior parte degli *itemset* in tabella 16 e tabella 17 mostra le evidenze della presenza di almeno un server `DHCP` e, in particolare, ne individua uno nel server `1.1.1.1`, il cui indirizzo, del tutto incongruente con le reti dell'ateneo, potrebbe far sospettare la presenza di un *rogue DHCP*. In realtà l'anomalia è frutto della configurazione di *default* di taluni apparati *Wireless Lan Controller* (WLC), quali i Cisco e gli Airespace, che utilizzavano una *virtual interface* con indirizzo `IP` di *default*, suggerito dal *vendor*, `1.1.1.1` e il servizio di `DHCP Proxy`. Tali WLC, infatti, intercettano i *client DHCP discovery packets*, inseriscono il proprio indirizzo `IP` nella *egress interface* nel campo *relay agent* prima di inoltrarlo al vero `DHCP` server e, alla ricezione della risposta, nuovamente intercettano e sostituiscono il *DHCP server ID* con l'indirizzo `IP` della propria interfaccia virtuale. Da qui l'apparente anomalia.

**Esempio 14** (*Evidenze di alcuni servizi NetBIOS*) Numerose righe della tabella 16 e della tabella 17 manifestano le prime evidenze del *`NetBIOS` su `TCP/IP`* (o *`NBT`*), standardizzato dalle RFC 1001

e 1002, e che era utilizzato soprattutto dai sistemi Windows, prima della sua obsolescenza. In particolare, ciò che si evidenzia è il servizio *NetBios Name Service*, attivo sulla porta `137/UDP`, per la registrazione e risoluzione dei nomi nelle reti locali.

**Esempio 15** (*Individuazione di traffico anomalo o sconosciuto*) Le righe 6, 24 e 29 della tabella 16 e la 9 della tabella 17 mostrano la presenza di un server di qualche tipo, che si può congetturare essere una Apple *AirPort Base Station*. Vi sono inoltre numerosi *itemset* che mostrano le evidenze di traffico non facilmente identificabile quali, per esempio, il traffico broadcast sulla porta `4445/UDP` e il traffico sulle porte `2987`, `6861/TCP` e `8000/TCP`.

**Tabella 16** — *Alcuni itemset di interesse nello strato* ⟨G∩M, 0.4%, **2**⟩

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {portsource:68 portdest:67} | 4 | 1,00 | 1,00 | 1,00 |
| 2 | {portsource:67 portdest:68} | 2 | 1,00 | 1,00 | 1,00 |
| 3 | {ipsource:1.1.1.1 portdest:68} | 2 | 1,00 | 1,00 | 1,00 |
| 4 | {ipsource:1.1.1.1 portsource:67} | 2 | 1,00 | 1,00 | 1,00 |
| 5 | {portsource:137 portdest:137} | 3 | 0,99 | 1,00 | 0,99 |
| 6 | {ipdest:172.20.90.1 portdest:192} | 1 | 0,95 | 0,98 | 0,93 |
| 7 | {ipdest:255.255.255.255 portdest:4445} | 1 | 0,74 | 0,86 | 0,64 |
| 8 | {ipsource:172.20.90.163 portdest:6861} | 1 | 0,64 | 0,73 | 0,50 |
| 9 | {ipdest:172.20.91.255 portdest:137} | 1 | 0,60 | 0,71 | 0,46 |
| 10 | {portsource:137 ipdest:172.20.91.255} | 1 | 0,60 | 0,70 | 0,45 |
| 11 | {ipsource:172.20.91.75 portdest:8000} | 1 | 0,58 | 0,78 | 0,44 |
| 12 | {ipsource:172.20.60.192 ipdest:172.20.61.255} | 1 | 0,55 | 0,63 | 0,39 |
| 13 | {ipsource:172.20.90.160 portdest:4445} | 1 | 0,55 | 0,56 | 0,38 |
| 14 | {ipsource:172.20.60.80 portdest:8000} | 1 | 0,50 | 0,61 | 0,32 |
| 15 | {portdest:4445 protocol:UDP_*} | 1 | 0,51 | 0,23 | 0,25 |
| 16 | {portdest:6861 protocol:TCP_*} | 1 | 0,51 | 0,24 | 0,24 |
| 17 | {portdest:137 protocol:UDP_*} | 4 | 0,51 | 0,21 | 0,24 |
| 18 | {portdest:67 protocol:UDP_*} | 5 | 0,51 | 0,21 | 0,24 |
| 19 | {portsource:68 protocol:UDP_*} | 4 | 0,51 | 0,21 | 0,24 |
| 20 | {portsource:137 protocol:UDP_*} | 3 | 0,51 | 0,21 | 0,23 |
| 21 | {portdest:68 protocol:UDP_*} | 2 | 0,51 | 0,17 | 0,20 |
| 22 | {portsource:67 protocol:UDP_*} | 2 | 0,51 | 0,17 | 0,20 |
| 23 | {ipsource:1.1.1.1 protocol:UDP_*} | 2 | 0,51 | 0,17 | 0,20 |
| 24 | {portdest:192 protocol:UDP_*} | 1 | 0,50 | 0,13 | 0,18 |
| 25 | {portdest:8000 protocol:TCP_*} | 4 | 0,44 | 0,24 | 0,16 |
| 26 | {ipsource:172.20.91.162 portdest:2987} | 1 | 0,38 | 0,60 | 0,16 |
| 27 | {ipsource:172.20.91.27 portdest:2987} | 1 | 0,36 | 0,59 | 0,13 |
| 28 | {ipsource:172.20.90.182 portdest:8000} | 1 | 0,34 | 0,54 | 0,11 |
| 29 | {portsource:DYNAMIC ipdest:172.20.90.1} | 1 | 0,50 | 0,05 | 0,08 |

**Esempio 16** (*Evidenze di protocolli e servizi specifici e diffusi*) A questo livello vengono individuati uno o più *host* interessati dall'attività dello *SSDP (Simple Service Discovery Protocol)*, ossia il protocollo utilizzato dai sistemi Windows per la segnalazione e la scoperta delle periferiche *UPnP*, e le evidenze dell'attività del *tunneling* Teredo, sovente utilizzato come tecnologia di transizione tra `IPv4` e `IPv6`.

**Tabella 17** — *Alcuni itemset di interesse negli strati* ⟨G∩M, 0.4%, **3**⟩ *e* ⟨G∩M, 0.4%, **4**⟩

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {ipsource:1.1.1.1 portsource:67 portdest:68} | 2 | 1,00 | 1,00 | 1,00 |
| 2 | {ipsource:1.1.1.1 portsource:67 portdest:68 protocol:UDP_*} | 2 | 0,75 | 0,89 | 0,66 |
| 3 | {portsource:137 ipdest:172.20.91.255 portdest:137} | 1 | 0,71 | 0,89 | 0,60 |
| 4 | {portsource:68 portdest:67 protocol:UDP_*} | 4 | 0,67 | 0,74 | 0,55 |
| 5 | {portsource:67 portdest:68 protocol:UDP_*} | 2 | 0,67 | 0,72 | 0,55 |
| 6 | {ipsource:1.1.1.1 portsource:67 protocol:UDP_*} | 2 | 0,67 | 0,72 | 0,55 |
| 7 | {ipsource:1.1.1.1 portdest:68 protocol:UDP_*} | 2 | 0,67 | 0,72 | 0,55 |
| 8 | {portsource:137 portdest:137 protocol:UDP_*} | 3 | 0,67 | 0,73 | 0,55 |
| 9 | {ipdest:172.20.90.1 portdest:192 protocol:UDP_*} | 1 | 0,63 | 0,70 | 0,50 |
| 10 | {ipdest:239.255.255.250 portdest:1900 protocol:UDP_*} | 9 | 0,63 | 0,70 | 0,49 |
| 11 | {ipsource:172.20.91.46 ipdest:239.255.255.177 portdest:1900} | 1 | 0,56 | 0,83 | 0,40 |
| 12 | {portsource:137 ipdest:172.20.91.255 portdest:137 protocol:UDP_*} | 1 | 0,53 | 0,82 | 0,37 |
| 13 | {ipsource:172.20.90.160 ipdest:255.255.255.255 portdest:4445} | 1 | 0,52 | 0,77 | 0,36 |
| 14 | {ipdest:255.255.255.255 portdest:4445 protocol:UDP_*} | 1 | 0,50 | 0,66 | 0,32 |
| 15 | {ipsource:172.20.90.163 portdest:6861 protocol:TCP_*} | 1 | 0,43 | 0,59 | 0,23 |
| 16 | {ipsource:172.20.91.46 ipdest:239.255.255.177 portdest:1900 protocol:UDP_*} | 1 | 0,42 | 0,83 | 0,22 |
| 17 | {ipdest:172.20.91.255 portdest:137 protocol:UDP_*} | 1 | 0,40 | 0,54 | 0,19 |
| 18 | {portsource:137 ipdest:172.20.91.255 protocol:UDP_*} | 1 | 0,40 | 0,54 | 0,19 |
| 19 | {ipsource:172.20.90.160 ipdest:255.255.255.255 portdest:4445 protocol:UDP_*} | 1 | 0,39 | 0,78 | 0,18 |
| 20 | {ipsource:172.20.60.192 ipdest:172.20.61.255 protocol:UDP_*} | 1 | 0,37 | 0,49 | 0,14 |
| 21 | {ipsource:172.20.90.160 portdest:4445 protocol:UDP_*} | 1 | 0,37 | 0,49 | 0,14 |
| 22 | {ipsource:172.20.91.46 portdest:1900 protocol:UDP_*} | 1 | 0,36 | 0,59 | 0,13 |

### *6.2.9 Itemset con supporto maggiore o uguale a 0,2% e minore di 0,4%*

In questo livello si manifestano ulteriori individuazioni di *host* fruitori di servizi e applicativi dei quali si già avute le evidenze nei livelli precedenti oppure fruitori di taluni servizi Google, quali *Google Playstore, Google Talk, Google Chrome Sync et cetera*, dei quali non verrà fatta ulteriore analisi in questo paragrafo. Compaiono inoltre le evidenze di ulteriori servizi `NetBIOS`, l'individuazione di ulteriori *host* configurati per accedere a `DNS` non locali e l'individuazione di server d'ateneo eroganti altri servizi oltre ai server `DNS` e `DHCP` individuati sin nei livelli più alti di supporto. Infine, si manifesta la completa caratterizzazione di parte del traffico anomalo già individuato in precedenza e che ora può essere associato con certezza a *malware* operante su di alcuni *host* locali.

**Esempio 17** (*Evidenze di ulteriori servizi NetBIOS*) Sia in tabella 18 sia in tabella 19 si manifestano le chiare evidenze del *NetBios Datagram Service* (`138/UDP`), per lo scambio di messaggi tra *host,* e, in via marginale, del *NetBios Session Service* (`139/TCP`), per l'accesso a file system e risorse di stampa condivisi, oltre al *NetBios Name Service* (`137/TCP&UDP`) già rilevato nel livello precedente. Il rilievo che assume la presenza di questi servizi ormai obsoleti è determinato dal fatto che essi sono sovente utilizzati per condurre attacchi informatici. Si noti che almeno un *endpoint* (`131.246.125.186, drucker.uni-kl.de`) si trova al di fuori della rete di ateneo.

**Tabella 18** — *Alcuni itemset di interesse nello strato* ⟨G∩M, 0.2%, 2⟩

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {portsource:138 portdest:138} | 7 | 1,00 | 1,00 | 1,00 |
| 2 | {portsource:20180 portdest:17788} | 2 | 1,00 | 1,00 | 1,00 |
| 3 | {ipdest:59.151.12.98 portdest:2060} | 1 | 1,00 | 1,00 | 1,00 |
| 4 | {ipdest:130.192.55.110 portdest:993} | 1 | 0,90 | 0,96 | 0,86 |
| 5 | {ipdest:61.135.185.18 portdest:5287} | 1 | 0,84 | 0,93 | 0,78 |
| 6 | {ipdest:131.246.125.186 portdest:139} | 1 | 0,75 | 0,89 | 0,66 |
| 7 | {ipdest:130.192.41.249 portdest:445} | 1 | 0,74 | 0,88 | 0,66 |
| 8 | {ipdest:MANY portdest:8000} | 3 | 0,71 | 0,85 | 0,61 |
| 9 | {ipdest:54.235.70.232 portdest:6861} | 1 | 0,71 | 0,85 | 0,60 |
| 10 | {ipdest:239.255.255.250 portdest:3702} | 2 | 0,68 | 0,82 | 0,57 |
| 11 | {ipdest:255.255.255.255 portdest:2654} | 1 | 0,66 | 0,81 | 0,54 |
| 12 | {ipdest:54.243.75.234 portdest:5242} | 1 | 0,63 | 0,77 | 0,50 |
| 13 | {ipsource:172.20.91.34 portdest:2654} | 1 | 0,62 | 0,77 | 0,50 |
| 14 | {ipsource:172.20.91.29 portdest:5242} | 1 | 0,61 | 0,80 | 0,47 |
| 15 | {ipsource:172.20.90.37 portdest:139} | 1 | 0,60 | 0,74 | 0,46 |
| 16 | {ipsource:172.20.90.64 portdest:3702} | 1 | 0,59 | 0,78 | 0,45 |
| 17 | {ipsource:172.20.61.26 ipdest:172.20.61.255} | 1 | 0,59 | 0,80 | 0,45 |
| 18 | {ipsource:172.20.91.85 ipdest:192.168.1.1} | 1 | 0,58 | 0,71 | 0,44 |
| 19 | {ipsource:172.20.91.85 ipdest:208.67.220.220} | 1 | 0,58 | 0,71 | 0,44 |
| 20 | {ipsource:172.20.91.85 ipdest:208.67.220.222} | 1 | 0,58 | 0,71 | 0,44 |
| 21 | {ipsource:172.20.68.12 ipdest:130.192.55.110} | 1 | 0,58 | 0,70 | 0,44 |
| 22 | {ipsource:172.20.91.85 ipdest:208.67.222.222} | 1 | 0,58 | 0,71 | 0,44 |
| 23 | {portsource:138 ipdest:172.20.91.255} | 5 | 0,58 | 0,70 | 0,43 |
| 24 | {ipdest:172.20.91.255 portdest:138} | 5 | 0,58 | 0,70 | 0,43 |
| 25 | {ipdest:172.20.69.255 portdest:138} | 1 | 0,57 | 0,78 | 0,43 |
| 26 | {portsource:138 ipdest:172.20.69.255} | 1 | 0,57 | 0,78 | 0,43 |
| 27 | {ipsource:172.20.91.38 ipdest:130.192.41.249} | 1 | 0,57 | 0,67 | 0,42 |
| 28 | {ipsource:172.20.60.108 portdest:2060} | 1 | 0,54 | 0,55 | 0,37 |
| 29 | {ipdest:172.20.69.255 portdest:137} | 1 | 0,53 | 0,77 | 0,36 |
| 30 | {ipdest:172.20.63.255 portdest:137} | 1 | 0,52 | 0,78 | 0,36 |
| 31 | {ipdest:MANY portdest:82} | 1 | 0,53 | 0,54 | 0,36 |
| 32 | {ipsource:172.20.68.76 portdest:2987} | 1 | 0,53 | 0,52 | 0,36 |
| 33 | {portsource:137 ipdest:172.20.69.255} | 1 | 0,52 | 0,77 | 0,35 |
| 34 | {ipsource:172.20.69.173 portdest:3702} | 1 | 0,52 | 0,70 | 0,35 |
| 35 | {ipsource:172.20.90.27 portdest:5242} | 1 | 0,51 | 0,73 | 0,35 |
| 36 | {ipsource:172.20.68.12 portdest:993} | 1 | 0,50 | 0,67 | 0,32 |
| 37 | {ipsource:172.20.30.154 portdest:8000} | 1 | 0,47 | 0,69 | 0,29 |
| 38 | {portdest:5222 protocol:TCP_*} | 7 | 0,51 | 0,23 | 0,25 |
| 39 | {portdest:8332 protocol:TCP_*} | 1 | 0,50 | 0,23 | 0,25 |
| 40 | {portdest:139 protocol:TCP_*} | 1 | 0,50 | 0,21 | 0,25 |
| 41 | {portdest:3702 protocol:UDP_*} | 2 | 0,51 | 0,22 | 0,25 |
| 42 | {portsource:138 protocol:UDP_*} | 7 | 0,50 | 0,21 | 0,24 |
| 43 | {portdest:138 protocol:UDP_*} | 7 | 0,50 | 0,21 | 0,24 |
| 44 | {ipsource:172.20.91.177 portdest:8000} | 1 | 0,44 | 0,60 | 0,24 |
| 45 | {portsource:5353 protocol:UDP_*} | 1 | 0,50 | 0,19 | 0,23 |
| 46 | {portdest:5287 protocol:TCP_*} | 2 | 0,50 | 0,19 | 0,23 |
| 47 | {portdest:8000 protocol:TCP_*} | 3 | 0,44 | 0,24 | 0,16 |
| 48 | {ipsource:172.20.90.182 portdest:8000} | 1 | 0,34 | 0,54 | 0,11 |
| 49 | {portdest:8000 protocol:UDP_*} | 2 | 0,43 | 0,13 | 0,10 |
| 50 | {ipsource:172.20.91.168 portdest:8000} | 1 | 0,27 | 0,57 | 0,02 |

**Tabella 19** — *Alcuni itemset di interesse nello strato* ⟨G∩M, 0.2%, **3**⟩

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {ipsource:172.20.68.127 portsource:20180 portdest:17788} | 1 | 0,84 | 0,96 | 0,79 |
| 2 | {ipsource:172.20.68.149 portsource:20180 portdest:17788} | 1 | 0,83 | 0,96 | 0,77 |
| 3 | {portsource:138 ipdest:172.20.91.255 portdest:138} | 5 | 0,72 | 0,90 | 0,62 |
| 4 | {ipsource:172.20.60.108 ipdest:59.151.12.98 portdest:2060} | 1 | 0,69 | 0,85 | 0,58 |
| 5 | {portsource:138 portdest:138 protocol:UDP_*} | 7 | 0,67 | 0,74 | 0,55 |
| 6 | {portsource:20180 portdest:17788 protocol:UDP_*} | 2 | 0,67 | 0,73 | 0,55 |
| 7 | {portsource:138 ipdest:172.20.69.255 portdest:138} | 1 | 0,66 | 0,90 | 0,54 |
| 8 | {ipsource:172.20.68.12 ipdest:130.192.55.110 portdest:993} | 1 | 0,65 | 0,88 | 0,53 |
| 9 | {ipdest:130.192.55.110 portdest:993 protocol:TCP_*} | 1 | 0,60 | 0,70 | 0,46 |
| 10 | {ipsource:172.20.91.75 ipdest:MANY portdest:8000} | 1 | 0,57 | 0,84 | 0,43 |
| 11 | {ipsource:172.20.90.37 ipdest:131.246.125.186 portdest:139} | 1 | 0,57 | 0,85 | 0,42 |
| 12 | {ipdest:61.135.185.18 portdest:5287 protocol:TCP_*} | 1 | 0,56 | 0,69 | 0,41 |
| 13 | {ipsource:172.20.90.182 ipdest:61.135.185.18 portdest:5287} | 1 | 0,55 | 0,81 | 0,40 |
| 14 | {ipsource:172.20.91.38 ipdest:130.192.41.249 portdest:445} | 1 | 0,54 | 0,82 | 0,39 |
| 15 | {ipsource:172.20.91.34 ipdest:255.255.255.255 portdest:2654} | 1 | 0,52 | 0,81 | 0,36 |
| 16 | {ipsource:172.20.68.127 portdest:17788 protocol:UDP_*} | 1 | 0,52 | 0,66 | 0,34 |
| 17 | {ipsource:172.20.68.127 portsource:20180 protocol:UDP_*} | 1 | 0,52 | 0,66 | 0,34 |
| 18 | {ipdest:131.246.125.186 portdest:139 protocol:TCP_*} | 1 | 0,50 | 0,67 | 0,33 |
| 19 | {ipsource:172.20.90.64 ipdest:239.255.255.250 portdest:3702} | 1 | 0,50 | 0,80 | 0,33 |
| 20 | {ipsource:172.20.68.149 portdest:17788 protocol:UDP_*} | 1 | 0,50 | 0,63 | 0,33 |
| 21 | {ipsource:172.20.68.149 portsource:20180 protocol:UDP_*} | 1 | 0,50 | 0,63 | 0,33 |
| 22 | {ipdest:130.192.41.249 portdest:445 protocol:TCP_*} | 1 | 0,50 | 0,67 | 0,33 |
| 23 | {ipsource:172.20.60.80 ipdest:MANY portdest:8000} | 1 | 0,48 | 0,76 | 0,30 |
| 24 | {ipdest:MANY portdest:8000 protocol:TCP_*} | 3 | 0,48 | 0,66 | 0,30 |
| 25 | {ipsource:172.20.91.215 ipdest:176.31.229.25 portdest:53} | 1 | 0,46 | 0,63 | 0,28 |
| 26 | {ipsource:172.20.91.215 ipdest:176.31.229.24 portdest:53} | 1 | 0,46 | 0,63 | 0,27 |
| 27 | {ipdest:239.255.255.250 portdest:3702 protocol:UDP_*} | 2 | 0,46 | 0,64 | 0,27 |
| 28 | {ipsource:74.125.218.181 ipdest:172.20.91.71 protocol:TCP_*} | 1 | 0,45 | 0,60 | 0,25 |
| 29 | {ipdest:255.255.255.255 portdest:2654 protocol:UDP_*} | 1 | 0,44 | 0,61 | 0,25 |
| 30 | {ipsource:172.20.69.173 ipdest:239.255.255.250 portdest:3702} | 1 | 0,42 | 0,75 | 0,23 |
| 31 | {ipdest:54.243.75.234 portdest:5242 protocol:TCP_*} | 1 | 0,42 | 0,61 | 0,22 |
| 32 | {ipsource:172.20.91.34 portdest:2654 protocol:UDP_*} | 1 | 0,42 | 0,58 | 0,22 |
| 33 | {ipsource:172.20.90.182 ipdest:123.125.113.30 portdest:8000} | 1 | 0,41 | 0,74 | 0,21 |
| 34 | {ipsource:172.20.90.37 portdest:139 protocol:TCP_*} | 1 | 0,40 | 0,59 | 0,20 |
| 35 | {ipsource:172.20.90.64 portdest:3702 protocol:UDP_*} | 1 | 0,40 | 0,61 | 0,19 |
| 36 | {ipsource:172.20.61.26 ipdest:172.20.61.255 protocol:UDP_*} | 1 | 0,39 | 0,60 | 0,18 |
| 37 | {ipsource:172.20.91.85 ipdest:208.67.222.220 portdest:53} | 1 | 0,39 | 0,56 | 0,18 |
| 38 | {ipsource:172.20.91.85 ipdest:192.168.1.1 portdest:53} | 1 | 0,39 | 0,56 | 0,18 |
| 39 | {ipsource:172.20.91.85 ipdest:208.67.222.220 protocol:UDP_*} | 1 | 0,39 | 0,54 | 0,18 |
| 40 | {ipsource:172.20.91.85 ipdest:192.168.1.1 protocol:UDP_*} | 1 | 0,39 | 0,54 | 0,18 |
| 41 | {ipsource:172.20.91.85 ipdest:208.67.220.220 portdest:53} | 1 | 0,39 | 0,56 | 0,18 |
| 42 | {ipsource:172.20.91.85 ipdest:208.67.220.222 portdest:53} | 1 | 0,39 | 0,56 | 0,18 |
| 43 | {ipsource:172.20.91.85 ipdest:208.67.222.222 portdest:53} | 1 | 0,39 | 0,56 | 0,18 |
| 44 | {ipsource:172.20.68.12 ipdest:130.192.55.110 protocol:TCP_*} | 1 | 0,39 | 0,54 | 0,18 |
| 45 | {ipsource:172.20.91.85 ipdest:208.67.220.220 protocol:UDP_*} | 1 | 0,39 | 0,54 | 0,18 |
| 46 | {ipsource:172.20.91.85 ipdest:208.67.220.222 protocol:UDP_*} | 1 | 0,39 | 0,54 | 0,18 |
| 47 | {ipsource:172.20.91.85 ipdest:208.67.222.222 protocol:UDP_*} | 1 | 0,39 | 0,54 | 0,17 |
| 48 | {portsource:138 ipdest:172.20.91.255 protocol:UDP_*} | 5 | 0,39 | 0,54 | 0,17 |
| 49 | {ipdest:172.20.91.255 portdest:138 protocol:UDP_*} | 5 | 0,39 | 0,54 | 0,17 |
| 50 | {portsource:138 ipdest:172.20.69.255 protocol:UDP_*} | 1 | 0,38 | 0,60 | 0,17 |
| 51 | {ipdest:172.20.69.255 portdest:138 protocol:UDP_*} | 1 | 0,38 | 0,60 | 0,17 |
| 52 | {ipsource:172.20.91.38 ipdest:130.192.41.249 protocol:TCP_*} | 1 | 0,38 | 0,55 | 0,16 |
| 53 | {ipsource:172.20.90.167 ipdest:MANY portdest:82} | 1 | 0,37 | 0,57 | 0,16 |

**Esempio 18** (*Individuazione di ulteriori client di DNS non locali*) Numerosi *itemset* nella tabella 18 e nella tabella 19 individuano ulteriori *host* che utilizzano server DNS non d'ateneo. Si tratta dei ben noti e diffusi DNS server di opendns.com (208.67.220.220, 208.67.220.222, 208.67.222.220 e 208.67.222.222) ma anche degli altrettanto noti server (176.31.229.25 e 176.31.229.24) che sono dei falsi DNS configurati, all'insaputa dell'utente, dal *malware DNSChanger* ai fini della perpetrazione di truffe on-line e tentativi di *phising*.

**Esempio 19** (*Ulteriori caratterizzazioni di traffico precedentemente rilevato come anomalo, sconosciuto o riferibile a malware)* Nell'esempio 4 si era già rilevata l'evidenza di traffico riferibile a *malware* ben caratterizzato operante sulla porta 82/TCP; a questo livello, l'emersione degli indirizzi IP di destinazione conferma tale riferibilità, giacché la maggior parte degli indirizzi IP remoti[16] verso i quali è indirizzato tale traffico sono risultati segnalati come associati a *malware*. Analogamente, nell'esempio 15 le evidenze di traffico anomalo verso la porta 8000/TCP non avevano trovato migliore caratterizzazione a quel livello, ma la trovano in questo grazie anche in questo caso all'emersione degli indirizzi IP remoti di destinazione, ampiamente segnalati come associati a *malware* e ad attività di *phishing*.

**Tabella 20** — *Alcuni itemset di interesse nello strato* ⟨G∩M, 0.2%, **4**⟩

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {ipsource:172.20.68.127 portsource:20180 portdest:17788 protocol:UDP_*} | 1 | 0,64 | 0,87 | 0,51 |
| 2 | {ipsource:172.20.68.149 portsource:20180 portdest:17788 protocol:UDP_*} | 1 | 0,63 | 0,85 | 0,50 |
| 3 | {portsource:138 ipdest:172.20.91.255 portdest:138 protocol:UDP_*} | 5 | 0,54 | 0,83 | 0,38 |
| 4 | {ipsource:172.20.60.108 ipdest:59.151.12.98 portdest:2060 protocol:TCP_*} | 1 | 0,52 | 0,79 | 0,35 |
| 5 | {portsource:138 ipdest:172.20.69.255 portdest:138 protocol:UDP_*} | 1 | 0,50 | 0,84 | 0,32 |
| 6 | {ipsource:172.20.68.12 ipdest:130.192.55.110 portdest:993 protocol:TCP_*} | 1 | 0,49 | 0,82 | 0,32 |
| 7 | {ipsource:172.20.91.75 ipdest:MANY portdest:8000 protocol:TCP_*} | 1 | 0,43 | 0,83 | 0,24 |
| 8 | {ipsource:172.20.90.37 ipdest:131.246.125.186 portdest:139 protocol:TCP_*} | 1 | 0,43 | 0,83 | 0,23 |
| 9 | {ipsource:172.20.91.38 ipdest:130.192.41.249 portdest:445 protocol:TCP_*} | 1 | 0,41 | 0,81 | 0,21 |
| 10 | {ipsource:172.20.91.215 ipdest:176.31.229.25 portdest:53 protocol:UDP_*} | 1 | 0,35 | 0,66 | 0,12 |
| 11 | {ipsource:172.20.91.215 ipdest:176.31.229.24 portdest:53 protocol:UDP_*} | 1 | 0,35 | 0,66 | 0,12 |

**Esempio 20** (*Individuazione di ulteriori server d'ateneo*) Numerosi *itemset* individuano due ulteriori server pubblici della rete di ateneo, ossia 130.192.55.110 (compass.polito.it) e 130.192.41.249 (garnerodesk.polito.it), utilizzati il primo per l'accesso a caselle di posta elettronica via IMAP/SSL e il secondo, in modo anomalo e inatteso essendo attestato su di un IP pubblico, per l'accesso a risorse attraverso la porta 445/TCP, che corrisponde ai servizi di *Microsoft-DS Active Directory* e *Windows shares*.

**Esempio 21** (*Individuazione di traffico anomalo o sconosciuto*) Si possono individuare degli *itemset* certamente *unexpected* in quelli che rivelano un traffico verso la porta 53 (DNS) di un *host* con indirizzo IP 192.168.1.1, ma è verosimile che si tratti semplicemente di errate configurazioni. Un significativo esempio di traffico sconosciuto è quello tra le porte 20180 e 17788 con protocollo UDP, che non è riferibile ad alcun servizio o processo noto o documentato. Invece, il traffico verso la porta 2060/TCP è verosimilmente riferibile a un *malware* noto come «*Protoss*», mentre quello della porta 5287/TCP è riferibile all'attività di applicazioni di *IP Camera viewer*.

[16] Tali indirizzi sono stati riassunti manualmente con la stringa MANY, giacché, anche in questo caso, non è stata operata alcuna generalizzazione in sottoreti da parte degli algoritmi.

### *6.2.10 Itemset con supporto maggiore o uguale a 0,1% e minore di 0,2%*

A questa soglia di supporto minimo, la più bassa tra quelle applicate ai dati sperimentali, aumenta considerevolmente la quantità di *itemset* che rivelano protocolli e servizi di interesse e una porzione di essi significativa risulta ora rilevata direttamente anche dagli *itemset* di lunghezza quattro, come si può vedere in tabella 21. Tralasciando l'analisi dettagliata di essi, che, ai fini esemplificativi non muta di molto quanto già mostrato nei livelli precedenti, ciò che appare soprattutto interessante qui è l'emersione di traffico verso numerose reti a indirizzamento privato — diverse dallo spazio di indirizzamento privato della rete *wireless* — che, almeno in teoria, non avrebbe dovuto sussistere, in quanto non sembra ragionevole che da una rete aperta, seppure munita di un *captive portal* per l'autenticazione, sia stato reso possibile il *routing* verso reti private interne alla struttura d'ateneo. Inoltre, a questo livello è possibile ottenere una consistente panoramica di ulteriori server utilizzati nella rete di ateneo e raggiungibili attraverso indirizzi `IP` pubblici.

**Tabella 21** — *Alcuni itemset di interesse nello strato* ⟨G∩M, 0.1%, **4**⟩

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {ipsource:172.20.91.143 ipdest:216.34.140.195 portdest:7275 protocol:TCP_*} | 1 | 0,63 | 0,86 | 0,51 |
| 2 | {ipsource:172.20.91.40 ipdest:159.0.120.230 portdest:8822 protocol:TCP_*} | 1 | 0,60 | 0,87 | 0,47 |
| 3 | {ipsource:172.20.91.210 ipdest:64.210.203.195 portdest:7275 protocol:TCP_*} | 1 | 0,58 | 0,86 | 0,44 |
| 4 | {ipsource:172.20.91.128 ipdest:192.168.1.5 portdest:161 protocol:UDP_*} | 1 | 0,53 | 0,83 | 0,37 |
| 5 | {portsource:57621 ipdest:172.20.91.255 portdest:57621 protocol:UDP_*} | 5 | 0,52 | 0,82 | 0,35 |
| 6 | {ipsource:172.20.91.104 ipdest:174.129.38.32 portdest:1883 protocol:TCP_*} | 1 | 0,52 | 0,81 | 0,35 |
| 7 | {ipsource:172.20.60.103 ipdest:77.31.92.121 portdest:8822 protocol:TCP_*} | 1 | 0,51 | 0,81 | 0,35 |
| 8 | {ipsource:172.20.68.29 ipdest:211.152.117.201 portdest:10482 protocol:TCP_*} | 1 | 0,51 | 0,79 | 0,34 |
| 9 | {ipsource:172.20.91.19 ipdest:224.0.0.1 portdest:8612 protocol:UDP_*} | 1 | 0,45 | 0,82 | 0,26 |
| 10 | {ipsource:172.20.90.38 ipdest:84.123.192.90 portdest:8888 protocol:TCP_*} | 1 | 0,44 | 0,77 | 0,25 |
| 11 | {ipsource:172.20.90.63 ipdest:64.41.140.209 portdest:5222 protocol:TCP_*} | 1 | 0,41 | 0,82 | 0,21 |
| 12 | {ipsource:172.20.68.86 ipdest:59.151.103.20 portdest:1883 protocol:TCP_*} | 1 | 0,40 | 0,79 | 0,19 |
| 13 | {ipsource:172.20.60.45 ipdest:172.20.60.1 portdest:192 protocol:UDP_*} | 1 | 0,39 | 0,76 | 0,18 |
| 14 | {ipsource:172.20.91.212 ipdest:186.2.164.89 portdest:8332 protocol:TCP_*} | 1 | 0,39 | 0,76 | 0,18 |
| 15 | {ipsource:172.20.91.212 ipdest:186.2.164.90 portdest:8332 protocol:TCP_*} | 1 | 0,38 | 0,75 | 0,17 |
| 16 | {ipsource:172.20.91.192 ipdest:192.168.1.12 portdest:445 protocol:TCP_*} | 1 | 0,35 | 0,81 | 0,13 |
| 17 | {ipsource:172.20.91.29 ipdest:54.243.75.234 portdest:5242 protocol:TCP_*} | 1 | 0,35 | 0,81 | 0,13 |

**Esempio 22** (*Evidenze e individuazioni di protocolli e servizi specifici*) Nella tabella 21 si possono facilmente individuare le evidenze del protocollo *OMA UserPlane Location* (righe 1 e 3), utilizzato per la localizzazione via GPS, dell'attività P2P dell'app *Spotify* (riga 5) e del protocollo *Message Queuing Telemetry* (righe 6 e 12) utilizzato, fra gli altri, da *Facebook Messenger.*

**Esempio 23** (*Routing verso reti private locali*) Le reti private interessate da questo traffico risultano essere delle *subnet* della rete privata `192.168.y.x` e della rete privata `10.z.y.x,` come si rileva dal loro pieno dettaglio nella tabella 22. Il traffico più significativo appare indirizzato verso l'*endpoint* `192.168.1.5` con il protocollo `SNMP` (*Simple Network Management Protocol*) sulla porta `161/UDP`, dal che si può dedurre che tale *endpoint* ospiti un *agent* e parrebbe pertanto sotto monitoraggio o controllo remoto. Segue in ordine di importanza il traffico verso l'*endpoint* `192.168.1.12`, che in ragione della porta utilizzata (`445/TCP`) e della sua consistenza sembrerebbe trattarsi di accesso a delle condivisioni di tipo *Windows Shares*. Tutto il rimanente traffico, poi, sembrerebbe essere di tipo `LDAP` (*Lightweight Directory Access Protocol*) e presumibilmente legato alle funzionalità della *Microsoft Active Directory*, tipicamente utilizzata nei domini Windows.

**Tabella 22** — *Itemset rivelatori di unexpected routing negli strati* ⟨G∩M, 0.1%, **2**⟩, ⟨G∩M, 0.1%, **3**⟩ *e* ⟨G∩M, 0.1%, **4**⟩

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {ipdest:192.168.1.5 portdest:161} | 1 | 0,96 | 0,99 | 0,94 |
| 2 | {ipsource:172.20.91.128 ipdest:192.168.1.5 portdest:161} | 1 | 0,70 | 0,91 | 0,60 |
| 3 | {ipsource:172.20.91.192 ipdest:192.168.1.12} | 1 | 0,64 | 0,80 | 0,52 |
| 4 | {ipdest:192.168.1.5 portdest:161 protocol:UDP_*} | 1 | 0,64 | 0,70 | 0,52 |
| 5 | {c2s_packets:(0-2061] ipdest:192.168.1.5 portdest:161} | 1 | 0,64 | 0,66 | 0,51 |
| 6 | {ipdest:192.168.1.5 portdest:161 s2c_packets:(0-3960]} | 1 | 0,64 | 0,66 | 0,51 |
| 7 | {ipdest:192.168.1.12 portdest:445} | 1 | 0,61 | 0,85 | 0,48 |
| 8 | {ipsource:172.20.91.128 ipdest:192.168.1.5} | 1 | 0,60 | 0,76 | 0,46 |
| 9 | {ipsource:172.20.91.85 ipdest:192.168.1.1} | 1 | 0,58 | 0,71 | 0,44 |
| 10 | {ipsource:172.20.60.15 ipdest:192.168.1.12} | 1 | 0,58 | 0,73 | 0,43 |
| 11 | {ipdest:192.168.183.4 portdest:389} | 2 | 0,56 | 0,67 | 0,40 |
| 12 | {ipdest:192.168.159.244 portdest:389} | 2 | 0,55 | 0,67 | 0,40 |
| 13 | {ipdest:192.168.83.2 portdest:389} | 2 | 0,55 | 0,67 | 0,40 |
| 14 | {ipdest:192.168.59.1 portdest:389} | 2 | 0,55 | 0,67 | 0,40 |
| 15 | {ipdest:192.168.55.244 portdest:389} | 2 | 0,55 | 0,66 | 0,40 |
| 16 | {ipdest:192.168.3.244 portdest:389} | 2 | 0,55 | 0,66 | 0,40 |
| 17 | {ipdest:192.168.59.6 portdest:389} | 2 | 0,55 | 0,66 | 0,40 |
| 18 | {ipdest:192.168.1.4 portdest:389} | 2 | 0,55 | 0,66 | 0,40 |
| 19 | {ipdest:192.168.40.1 portdest:389} | 1 | 0,55 | 0,64 | 0,39 |
| 20 | {ipsource:172.20.91.128 ipdest:192.168.1.5 portdest:161 protocol:UDP_*} | 1 | 0,53 | 0,83 | 0,37 |
| 21 | {ipsource:172.20.91.128 c2s_packets:(0-2061] ipdest:192.168.1.5 portdest:161} | 1 | 0,53 | 0,80 | 0,37 |
| 22 | {ipsource:172.20.91.128 ipdest:192.168.1.5 portdest:161 s2c_packets:(0-3960]} | 1 | 0,53 | 0,80 | 0,37 |
| 23 | {ipsource:172.20.90.38 ipdest:10.139.56.2} | 1 | 0,52 | 0,51 | 0,35 |
| 24 | {ipdest:10.139.56.2 portdest:80} | 1 | 0,52 | 0,46 | 0,34 |
| 25 | {c2s_packets:(0-2061] ipdest:192.168.1.5 portdest:161 protocol:UDP_*} | 1 | 0,48 | 0,61 | 0,30 |
| 26 | {ipdest:192.168.1.5 portdest:161 s2c_packets:(0-3960] protocol:UDP_*} | 1 | 0,48 | 0,61 | 0,30 |
| 27 | {c2s_packets:(0-2061] ipdest:192.168.1.5 portdest:161 s2c_packets:(0-3960]} | 1 | 0,48 | 0,57 | 0,30 |
| 28 | {ipsource:172.20.91.192 ipdest:192.168.1.12 portdest:445} | 1 | 0,47 | 0,83 | 0,29 |
| 29 | {ipdest:10.139.56.2 protocol:TCP_*} | 1 | 0,50 | 0,21 | 0,25 |
| 30 | {ipsource:172.20.91.192 ipdest:192.168.1.12 protocol:TCP_*} | 1 | 0,43 | 0,61 | 0,23 |
| 31 | {ipdest:192.168.1.12 protocol:TCP_*} | 1 | 0,50 | 0,19 | 0,23 |
| 32 | {ipsource:172.20.91.192 c2s_packets:(0-3188] ipdest:192.168.1.12} | 1 | 0,43 | 0,55 | 0,23 |
| 33 | {ipsource:172.20.91.192 ipdest:192.168.1.12 s2c_packets:(0-10035]} | 1 | 0,43 | 0,55 | 0,23 |
| 34 | {ipsource:172.20.91.192 portsource:DYNAMIC ipdest:192.168.1.12} | 1 | 0,43 | 0,57 | 0,23 |
| 35 | {ipdest:192.168.183.4 protocol:UDP_*} | 2 | 0,50 | 0,17 | 0,23 |
| 36 | {ipdest:192.168.159.244 protocol:UDP_*} | 2 | 0,50 | 0,17 | 0,23 |
| 37 | {ipdest:192.168.83.2 protocol:UDP_*} | 2 | 0,50 | 0,17 | 0,23 |
| 38 | {ipdest:192.168.59.1 protocol:UDP_*} | 2 | 0,50 | 0,17 | 0,23 |
| 39 | {ipdest:192.168.59.6 protocol:UDP_*} | 2 | 0,50 | 0,17 | 0,23 |
| 40 | {ipdest:192.168.1.4 protocol:UDP_*} | 2 | 0,50 | 0,17 | 0,23 |
| 41 | {ipdest:192.168.55.244 protocol:UDP_*} | 2 | 0,50 | 0,17 | 0,23 |
| 42 | {ipdest:192.168.3.244 protocol:UDP_*} | 2 | 0,50 | 0,17 | 0,23 |
| 43 | {ipsource:172.20.91.128 c2s_packets:(0-2061] ipdest:192.168.1.5 portdest:161 protocol:UDP_*} | 1 | 0,42 | 0,78 | 0,22 |
| 44 | {ipsource:172.20.91.128 ipdest:192.168.1.5 portdest:161 s2c_packets:(0-3960] protocol:UDP_*} | 1 | 0,42 | 0,78 | 0,22 |
| 45 | {ipdest:192.168.40.1 protocol:UDP_*} | 1 | 0,50 | 0,16 | 0,22 |
| 46 | {ipdest:192.168.1.12 portdest:445 protocol:TCP_*} | 1 | 0,41 | 0,64 | 0,21 |
| 47 | {c2s_packets:(0-3188] ipdest:192.168.1.12 portdest:445} | 1 | 0,41 | 0,58 | 0,20 |
| 48 | {ipdest:192.168.1.12 portdest:445 s2c_packets:(0-10035]} | 1 | 0,41 | 0,58 | 0,20 |
| 49 | {ipdest:192.168.1.5 protocol:UDP_*} | 1 | 0,50 | 0,13 | 0,19 |
| 50 | {ipsource:172.20.91.128 ipdest:192.168.1.5 protocol:UDP_*} | 1 | 0,40 | 0,57 | 0,19 |

**Esempio 24** (*Individuazione di ulteriori server d'ateneo*) Nella tabella 23 si sono raccolti gli *itemset* che individuano ulteriori server d'ateneo che non si erano palesati nei livelli superiori. In essi sono facilmente identificabili soprattutto web server, sia utilizzanti il protocollo `http` sia il protocollo `https`; ma anche un server (`130.192.15.42, genesi.polito.it`) verso il quale è rivolto un traffico `LDAP` (*Lightweight Directory Access Protocol).*

**Tabella 23** — *Itemset rivelatori di server locali negli strati* $\langle \mathsf{G}\cap\mathsf{M}, 0.1\%, \mathbf{2}\rangle$ *e* $\langle \mathsf{G}\cap\mathsf{M}, 0.1\%, \mathbf{3}\rangle$

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {ipsource:130.192.55.221 ipdest:172.20.91.201} | 1 | 0,74 | 0,90 | 0,66 |
| 2 | {ipdest:130.192.41.249 portdest:445} | 1 | 0,74 | 0,88 | 0,66 |
| 3 | {ipsource:172.20.91.38 ipdest:130.192.41.249} | 1 | 0,57 | 0,67 | 0,42 |
| 4 | {ipdest:130.192.15.42 portdest:389} | 2 | 0,55 | 0,67 | 0,40 |
| 5 | {ipsource:172.20.91.38 ipdest:130.192.41.249 portdest:445} | 1 | 0,54 | 0,82 | 0,39 |
| 6 | {ipsource:172.20.63.98 ipdest:130.192.182.81} | 1 | 0,52 | 0,50 | 0,34 |
| 7 | {ipdest:130.192.41.249 portdest:445 protocol:TCP_*} | 1 | 0,50 | 0,67 | 0,33 |
| 8 | {ipsource:130.192.55.221 ipdest:172.20.91.201 protocol:TCP_*} | 1 | 0,50 | 0,65 | 0,32 |
| 9 | {ipsource:130.192.55.221 ipdest:172.20.91.201 s2c_packets:(0-45]} | 1 | 0,50 | 0,61 | 0,32 |
| 10 | {ipsource:130.192.55.221 c2s_packets:(0-140] ipdest:172.20.91.201} | 1 | 0,50 | 0,61 | 0,32 |
| 11 | {c2s_packets:(0-4249] ipdest:130.192.41.249 portdest:445} | 1 | 0,50 | 0,60 | 0,32 |
| 12 | {ipdest:130.192.41.249 portdest:445 s2c_packets:(0-3921]} | 1 | 0,50 | 0,60 | 0,32 |
| 13 | {ipsource:130.192.55.221 portsource:80 ipdest:172.20.91.201} | 1 | 0,49 | 0,76 | 0,31 |
| 14 | {ipsource:130.192.55.221 portsource:80} | 1 | 0,50 | 0,46 | 0,31 |
| 15 | {ipsource:130.192.55.221 ipdest:172.20.91.201 portdest:DYNAMIC} | 1 | 0,48 | 0,70 | 0,30 |
| 16 | {ipdest:130.192.55.225 portdest:443} | 2 | 0,51 | 0,30 | 0,30 |
| 17 | {ipdest:130.186.29.122 portdest:80} | 1 | 0,50 | 0,27 | 0,29 |
| 18 | {ipdest:130.192.55.240 portdest:443} | 2 | 0,50 | 0,30 | 0,28 |
| 19 | {ipdest:130.192.182.33 portdest:80} | 2 | 0,50 | 0,29 | 0,28 |
| 20 | {ipdest:130.186.29.70 portdest:80} | 1 | 0,50 | 0,28 | 0,28 |
| 21 | {ipdest:130.192.182.81 portdest:443} | 1 | 0,50 | 0,26 | 0,28 |
| 22 | {ipsource:130.192.55.221 portdest:DYNAMIC} | 1 | 0,48 | 0,31 | 0,26 |
| 23 | {ipdest:130.192.41.249 protocol:TCP_*} | 1 | 0,50 | 0,22 | 0,25 |
| 24 | {ipdest:130.192.15.42 protocol:UDP_*} | 2 | 0,50 | 0,17 | 0,23 |
| 25 | {ipdest:130.192.55.225 protocol:TCP_*} | 2 | 0,50 | 0,16 | 0,22 |
| 26 | {ipdest:130.186.29.70 protocol:TCP_*} | 1 | 0,50 | 0,16 | 0,22 |
| 27 | {ipdest:130.192.55.240 protocol:TCP_*} | 2 | 0,50 | 0,16 | 0,22 |
| 28 | {ipdest:130.186.3.24 protocol:TCP_*} | 1 | 0,50 | 0,15 | 0,22 |
| 29 | {ipdest:130.186.29.122 protocol:TCP_*} | 1 | 0,50 | 0,15 | 0,22 |
| 30 | {ipdest:130.192.182.81 protocol:TCP_*} | 1 | 0,50 | 0,15 | 0,22 |
| 31 | {ipdest:130.192.182.33 protocol:TCP_*} | 2 | 0,50 | 0,15 | 0,21 |
| 32 | {ipdest:130.186.3.24 portdest:80} | 1 | 0,44 | 0,25 | 0,20 |

## 6.3 Analisi degli itemset esclusivi dell'uno o dell'altro algoritmo

Dopo aver esemplificato i risultati di interesse per l'analisi qualitativa *domain-driven* e constatato che essi hanno riguardato *itemset* tutti contenuti nella partizione $\mathsf{G}\cap\mathsf{M}$, è necessario dare contezza degli *itemset* costituenti le due partizioni residuali $\mathsf{G}\setminus\mathsf{M}$ e $\mathsf{M}\setminus\mathsf{G}$. L'analisi di essi depone in generale per la loro irrilevanza e, più precisamente, per l'irrilevanza degli *itemset* costituenti $\mathsf{G}\setminus\mathsf{M}$ e per l'infungibilità di quelli costituenti $\mathsf{M}\setminus\mathsf{G}$. Per convincersene è sufficiente esaminare la natura e la

struttura degli *itemset* contenuti in tali partizioni anche solo negli strati rispettivamente a massimo, medio e minimo valore del supporto di soglia, prendendo ogni volta i primi dodici a maggior valore della metrica *zeta* già non presenti e mostrati negli strati superiori.

Per quanto riguarda gli *itemset* più significativi della partizione G\M, ovvero quelli generati esclusivamente dal *Genio Algorithm,* i risultati mostrati in tabella 24, tabella 25 e tabella 26 evidenziano tutte le caratteristiche peculiari a quella partizione, che bene ne descrivono il contenuto e che sono riscontrabili peraltro anche su tutti gli strati di analisi, ancorché non qui mostrati.

**Tabella 24** — *Primi dodici migliori itemset in* ⟨G\M, 5%,2⟩∪⟨G\M, 5%,3⟩∪⟨G\M, 5%,4⟩

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {ipsource:130.192.3.103 portsource:WELLKNOWN} | 5 | 0,84 | 0,82 | 0,69 |
| 2 | {ipsource:172.20.90.87 ipdest:239.255.255.253 portdest:REGISTERED} | 1 | 0,68 | 0,78 | 0,53 |
| 3 | {ipsource:172.20.90.87 ipdest:239.255.255.253 protocol:UDP} | 1 | 0,64 | 0,67 | 0,46 |
| 4 | {ipdest:224.0.0.252 portdest:REGISTERED} | 3 | 0,65 | 0,58 | 0,45 |
| 5 | {ipsource:172.20.90.87 c2s_packets:(0-8498] ipdest:239.255.255.253} | 1 | 0,63 | 0,64 | 0,45 |
| 6 | {ipsource:172.20.90.87 ipdest:239.255.255.253 s2c_packets:(0-7842]} | 1 | 0,63 | 0,64 | 0,45 |
| 7 | {ipsource:172.20.90.87 ipdest:239.255.255.253 s2c_packets:(0-62741]} | 1 | 0,63 | 0,64 | 0,45 |
| 8 | {ipsource:172.20.90.87 c2s_packets:(0-67987] ipdest:239.255.255.253} | 1 | 0,63 | 0,64 | 0,45 |
| 9 | {ipsource:172.20.90.87 ipdest:239.255.255.253 protocol:GENERAL} | 1 | 0,63 | 0,64 | 0,45 |
| 10 | {portsource:WELLKNOWN c2s_packets:(0-11177] portdest:DYNAMIC} | 1 | 0,64 | 0,60 | 0,40 |
| 11 | {portsource:WELLKNOWN portdest:DYNAMIC s2c_packets:(0-17723]} | 1 | 0,64 | 0,60 | 0,40 |
| 12 | {portsource:WELLKNOWN portdest:DYNAMIC protocol:GENERAL} | 8 | 0,64 | 0,60 | 0,40 |

**Tabella 25** — *Primi dodici migliori itemset in* ⟨G\M, 1%,2⟩∪⟨G\M, 1%,3⟩∪⟨G\M, 1%,4⟩

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {ipsource:172.20.90.76 ipdest:239.255.255.253 portdest:REGISTERED} | 1 | 0,65 | 0,75 | 0,51 |
| 2 | {ipsource:172.20.90.76 ipdest:239.255.255.253 protocol:UDP} | 1 | 0,64 | 0,68 | 0,49 |
| 3 | {ipsource:172.20.90.76 c2s_packets:(0-6376] ipdest:239.255.255.253} | 1 | 0,63 | 0,65 | 0,49 |
| 4 | {ipsource:172.20.90.76 ipdest:239.255.255.253 s2c_packets:(0-20070]} | 1 | 0,63 | 0,65 | 0,49 |
| 5 | {ipsource:172.20.90.76 ipdest:239.255.255.253 s2c_packets:(0-160562]} | 1 | 0,63 | 0,65 | 0,49 |
| 6 | {ipsource:172.20.90.76 c2s_packets:(0-51011] ipdest:239.255.255.253} | 1 | 0,63 | 0,65 | 0,49 |
| 7 | {ipsource:172.20.90.76 ipdest:239.255.255.253 protocol:GENERAL} | 1 | 0,63 | 0,65 | 0,49 |
| 8 | {ipsource:130.192.3.24 portsource:WELLKNOWN} | 8 | 0,63 | 0,63 | 0,47 |
| 9 | {portsource:WELLKNOWN ipdest:172.20.68.25} | 1 | 0,61 | 0,60 | 0,43 |
| 10 | {ipsource:130.192.3.21 portsource:WELLKNOWN} | 6 | 0,58 | 0,57 | 0,40 |
| 11 | {ipsource:172.20.91.243 portsource:REGISTERED} | 1 | 0,59 | 0,51 | 0,38 |
| 12 | {ipsource:172.20.33.121 ipdest:111.111.111.111 protocol:TCP} | 1 | 0,52 | 0,66 | 0,34 |

**Tabella 26** — *Primi dodici migliori itemset in* ⟨G\M, 0.1%,2⟩∪⟨G\M, 0.1%,3⟩∪⟨G\M, 0.1%,4⟩

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {portsource:WELLKNOWN portdest:DYNAMIC} | 1 | 0,85 | 0,83 | 0,70 |
| 2 | {ipsource:74.125.218.18 ipdest:172.20.91.77 protocol:TCP} | 1 | 0,66 | 0,72 | 0,53 |
| 3 | {ipsource:74.125.218.18 portsource:WELLKNOWN ipdest:172.20.91.77} | 1 | 0,65 | 0,78 | 0,53 |
| 4 | {ipsource:74.125.218.18 ipdest:172.20.91.77 s2c_packets:(0-90]} | 1 | 0,65 | 0,66 | 0,52 |
| 5 | {ipsource:74.125.218.18 c2s_packets:(0-281] ipdest:172.20.91.77} | 1 | 0,65 | 0,66 | 0,52 |
| 6 | {ipsource:74.125.218.18 c2s_packets:(0-2253] ipdest:172.20.91.77} | 1 | 0,65 | 0,66 | 0,52 |
| 7 | {ipsource:74.125.218.18 ipdest:172.20.91.77 s2c_packets:(0-727]} | 1 | 0,65 | 0,66 | 0,52 |
| 8 | {ipsource:74.125.218.18 ipdest:172.20.91.77 protocol:GENERAL} | 1 | 0,65 | 0,66 | 0,52 |
| 9 | {ipsource:109.201.134.238 ipdest:172.20.90.39 protocol:TCP} | 1 | 0,65 | 0,71 | 0,52 |
| 10 | {ipsource:31.216.144.40 ipdest:172.20.91.199 protocol:TCP} | 1 | 0,64 | 0,70 | 0,52 |
| 11 | {ipsource:109.201.134.238 ipdest:172.20.90.39 s2c_packets:(0-90]} | 1 | 0,65 | 0,66 | 0,51 |
| 12 | {ipsource:109.201.134.238 c2s_packets:(0-281] ipdest:172.20.90.39} | 1 | 0,65 | 0,66 | 0,51 |

A prima vista la gran parte degli *itemset* sembra essere qualitativamente interessante, ma se si approfondisce il loro esame eseguendo una comparazione con quelli della partizione G∩M si scopre subito la fallacia di questa prima impressione. Si considerino, per esempio, dapprima gli *itemset* del tipo {ipsource:130.192.3.x portsource:WELLKNOWN} e li ponga a confronto con gli omologhi {ipsource:130.192.3.x portsource:53} in G∩M. Il risultato è che per ciascuno dei nove *dataset* la coppia messa a confronto ha esattamente lo stesso supporto, a segno che quelli in G\M non sono null'altro che una generalizzazione ridondante e meno informativa dei secondi. Ancora, se si prendono gli itemset contenenti {ipsource:172.20.90.87 ipdest:239.255.255.253} oppure {ipsource:172.20.90.76 ipdest:239.255.255.253} e si esegue il confronto con i loro omologhi nell'altra partizione, si perviene allo stesso risultato, ovvero che laddove nei primi si riscontra {portdest:REGISTERED} o {protocol:UDP}, nei secondi si reperisce, rispettivamente, {portdest:6005} e {protocol:UDP_*} con esattamente lo stesso supporto e, in base alla tassonomia, con «6005» ◁ «REGISTERED»} e «UDP_*» ◁ «UDP». Si potrebbe ulteriormente mostrare con una grande varietà di esempi la costanza di questo risultato, che tuttavia non è per nulla sorprendente giacché esso mostra semplicemente l'efficacia del *Max-EGI extraction algorithm* nell'espungere le generalizzazioni ridondanti sicché in G\M rimangono proprio quegli *itemset* correttamente espunti dall'algoritmo.

Oltre a quanto sopra evidenziato, si scorgono altresì degli *itemset* che presentano delle generalizzazioni quantomeno anomale. L'esempio più evidente è rappresentato dagli *itemset* che contengono {protocol:GENERAL}. Se si va a verificare la tassonomia della figura 5(b), si vede che la semantica di questa generalizzazione equivale nientemeno a quella della radice ⊥. Ciò significa che un *itemset* del tipo {ipsource:74.125.218.18 ipdest:172.20.91.77 protocol:GENERAL} è in tutto semanticamente equivalente a {ipsource:74.125.218.18 ipdest:172.20.91.77} e quindi non solo è ridondante, ma a rigore neppure avrebbe dovuto esistere. Anche se non è immediatamente percepibile, pure *l'itemset* che contiene {s2c_packets:(0-160562]} va considerato alla medesima stregua: la semantica di quest'ultima generalizzazione è nientemeno quella di un flusso dal server al client che può andare indifferentemente da zero bytes a ≈ 225 MiB e che, proprio per l'enorme ampiezza del campo di variazione, di fatto non apporta alcuna informazione utile, tant'è vero che il supporto di {s2c_packets:(0-160562]} è sostanzialmente unitario. Il problema evidenziato da questi *itemset* allora non è semplicemente quello di esser stati iper-generalizzati, caso già esaminato *supra*, ma soprattutto quello di aver subito delle generalizzazioni a supporto unitario.

Un ultimo caso interessante è quello dell'itemset {portsource:WELLKNOWN portdest:DYNAMIC}, che compare in quasi tutti i *dataset* della partizione G∩M, salvo alcuni. In questi ultimi ciò è da attribuirsi al fatto che ne esiste una versione sotto forma di *Max-EGI* in M\G e allora la versione semplicemente generalizzata compare in G\M, come nel caso che qui si vede in tabella 26.

Riepilogando, senza pretesa di esaustività ma con la confidenza recata dal concorde risultato dell'esame degli strati, si può concludere che gli *itemset* appartenenti alla partizione G\M sono costituiti, in ordine di frequenza, da:

- *itemset* costituenti delle iper-generalizzazioni ridondanti degli *itemset* in G∩M e che come tali sono privi di interesse in quanto incapaci di apportare un contributo di conoscenza maggiore di questi ultimi;
- *itemset* costituenti delle iper-generalizzazioni su elementi della gerarchia che hanno supporto unitario e che come tali non dovrebbero neppure esistere, perché riproducono la medesima

conoscenza apportata da *itemset* di lunghezza $k$ inferiore;

- *itemset* generalizzati che possono essere posti in corrispondenza con analoghi *Max-EGI* in $\mathsf{M}\backslash\mathsf{G}$ e che rivestono interesse solo nella misura in cui i corrispondenti *Max-EGI* forniscono una conoscenza più involuta o non facilmente estraibile.

Per quanto riguarda invece gli *itemset* in $\mathsf{M}\backslash\mathsf{G}$, ovvero quelli generati esclusivamente dal *Max-EGI extraction algorithm,* occorre innanzitutto osservare questa partizione è costituita quasi ovunque da *Max-EGI* e anzi li ricomprende tutti, giacché $\mathsf{G}\backslash\mathsf{M}$ e $\mathsf{G}\cap\mathsf{M}$, com'è ovvio, sono costituite solo da *itemset* ordinari o generalizzati. Questo da un lato comporta che essi abbiano quasi sempre un contenuto informativo originale, ossia non reperibile nelle altre partizioni, ma, d'altro lato, almeno per i dati sperimentali in esame, raramente tale originalità si è tradotta in un giudizio di *actionability* o *unexpectedness,* e quindi di merito, da un punto di vista strettamente *domain-driven.*

**Tabella 27** — *Primi dodici itemset in* $\mathsf{M}\backslash\mathsf{G}$ *con* $\varepsilon = 5\%$ *e con i più alti valori di* $|\zeta|$

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {portdest:WELLKNOWN protocol:TCP} ≀ {{protocol:TCP_* portdest:443}} | 1 | 0,18 | -0,22 | -0,32 |
| 2 | {portdest:WELLKNOWN protocol:GENERAL} ≀ {{protocol:TCP_* portdest:443} {portdest:80 protocol:TCP_*} {protocol:UDP_* portdest:53} {portdest:53 protocol:UDP_DNS}} | 1 | 0,58 | 0,39 | 0,32 |
| 3 | {portdest:WELLKNOWN c2s_packets:(0-6376] s2c_packets:(0-20070] protocol:TCP} ≀ {{portdest:443 c2s_packets:(0-3188] s2c_packets:(0-10035] protocol:TCP_*}} | 1 | 0,12 | -0,12 | -0,28 |
| 4 | {portdest:WELLKNOWN c2s_packets:(0-1397] s2c_packets:(0-2215] protocol:TCP} ≀ {{portdest:443 c2s_packets:(0-698] s2c_packets:(0-1107] protocol:TCP_*}} | 1 | 0,13 | -0,10 | -0,27 |
| 5 | {portdest:WELLKNOWN c2s_packets:(0-281]} ≀ {{portdest:443 c2s_packets:(0-140]} {portdest:80 c2s_packets:(0-140]} {portdest:53 c2s_packets:(0-140]}} | 1 | 0,58 | 0,30 | 0,26 |
| 6 | {portdest:REGISTERED protocol:UDP} ≀ {{portdest:5355 protocol:UDP_*}} | 1 | 0,26 | -0,17 | -0,26 |
| 7 | {portdest:WELLKNOWN s2c_packets:(0-20070]} ≀ {{portdest:53 s2c_packets:(0-10035]} {portdest:443 s2c_packets:(0-10035]}} | 1 | 0,62 | 0,28 | 0,26 |
| 8 | {portdest:WELLKNOWN c2s_packets:(0-6376] } ≀ {{portdest:443 c2s_packets:(0-3188]} {portdest:53 c2s_packets:(0-3188]}} | 1 | 0,62 | 0,28 | 0,26 |
| 9 | {portdest:WELLKNOWN s2c_packets:(0-7842]} ≀ {{portdest:80 s2c_packets:(0-3921]} {portdest:443 s2c_packets:(0-3921]} {portdest:53 s2c_packets:(0-3921]}} | 1 | 0,59 | 0,28 | 0,25 |
| 10 | {portdest:WELLKNOWN c2s_packets:(0-8498]} ≀ {{portdest:53 c2s_packets:(0-4249]} {portdest:80 c2s_packets:(0-4249]} {portdest:443 c2s_packets:(0-4249]}} | 1 | 0,59 | 0,28 | 0,25 |
| 11 | {portdest:WELLKNOWN s2c_packets:(0-20070] protocol:TCP} ≀ {{portdest:443 s2c_packets:(0-10035] protocol:TCP_*}} | 1 | 0,14 | -0,03 | -0,25 |
| 12 | {portdest:WELLKNOWN c2s_packets:(0-6376] protocol:TCP} ≀ {{portdest:443 c2s_packets:(0-3188] protocol:TCP_*}} | 1 | 0,14 | -0,03 | -0,25 |

I risultati in tabella 27, tabella 28 e tabella 29 esemplificano la natura e la struttura degli *itemset* contenuti in $\mathsf{M}\backslash\mathsf{G}$ negli strati rispettivamente a massimo, medio e minimo valore del supporto di soglia, con valore rappresentativo d'insieme come nel caso precedente, prendendo però ogni volta, per le ragioni che saranno chiare nell'esposizione, non già i primi dodici a maggior valore, ma i primi dodici a maggior *valore assoluto* della metrica *zeta* se già non presenti e mostrati negli strati superiori, ricomprendendo quindi non solo gli *itemset* meglio correlati, ma anche quelli più significativamente anti-dipendenti.

**Tabella 28** — *Primi dodici itemset in* M\G *con* $\varepsilon = 1\%$ *e con i più alti valori di* $|\zeta|$

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {protocol:GENERAL c2s_packets:(0-11177] s2c_packets:(0-17723]} ≀ {{protocol:UDP_* c2s_packets:(0-698] s2c_packets:(0-1107]} {protocol:UDP_DNS c2s_packets:(0-698] s2c_packets:(0-1107]} {protocol:TCP_* c2s_packets:(0-698] s2c_packets:(0-1107]}} | 1 | 1,00 | 1,00 | 0,99 |
| 2 | {protocol:GENERAL s2c_packets:(0-17723]} ≀ {{protocol:TCP_* s2c_packets:(0-1107]} {protocol:UDP_* s2c_packets:(0-1107]} {protocol:UDP_DNS s2c_packets:(0-1107]}} | 1 | 1,00 | 1,00 | 0,99 |
| 3 | { protocol:GENERAL c2s_packets:(0-11177]} ≀ {{protocol:TCP_* c2s_packets:(0-698]} {protocol:UDP_DNS c2s_packets:(0-698]} {protocol:UDP_* c2s_packets:(0-698]}} | 1 | 1,00 | 1,00 | 0,99 |
| 4 | {portdest:WELLKNOWN protocol:TCP} ≀ {{portdest:80 protocol:TCP_*} {protocol:TCP_* portdest:443}} | 1 | 0,06 | -0,34 | -0,44 |
| 5 | {portdest:WELLKNOWN protocol:UDP} ≀ {{protocol:UDP_* portdest:389} {portdest:53 protocol:UDP_DNS} {portdest:53 protocol:UDP_*}} | 2 | 0,08 | -0,30 | -0,40 |
| 6 | {portdest:WELLKNOWN s2c_packets:(0-93]} ≀ {{portdest:80 s2c_packets:(0-46]} {portdest:443 s2c_packets:(0-46]} {portdest:53 s2c_packets:(0-46]}} | 1 | 0,58 | 0,52 | 0,38 |
| 7 | { portdest:WELLKNOWN c2s_packets:(0-395]} ≀ {{portdest:80 c2s_packets:(0-197]} {portdest:443 c2s_packets:(0-197]} {portdest:53 c2s_packets:(0-197]}} | 1 | 0,57 | 0,51 | 0,38 |
| 8 | {portdest:WELLKNOWN protocol:UDP} ≀ {{portdest:53 protocol:UDP_DNS} {portdest:53 protocol:UDP_*} {portdest:7 protocol:UDP_*}} | 1 | 0,11 | -0,27 | -0,37 |
| 9 | {portdest:WELLKNOWN protocol:UDP} ≀ {{portdest:53 protocol:UDP_DNS} {protocol:UDP_* portdest:123} {portdest:53 protocol:UDP_*} {portdest:7 protocol:UDP_*}} | 2 | 0,12 | -0,23 | -0,36 |
| 10 | {portdest:REGISTERED protocol:UDP} ≀ {{portdest:5355 protocol:UDP_*} {portdest:8612 protocol:UDP_*} {portdest:17500 protocol:UDP_*} {portdest:10505 protocol:UDP_*}} | 1 | 0,18 | -0,25 | -0,35 |
| 11 | {portdest:WELLKNOWN protocol:UDP} ≀ {{portdest:7 protocol:UDP_*} {portdest:53 protocol:UDP_DNS} {portdest:53 protocol:UDP_*}} | 1 | 0,11 | -0,21 | -0,35 |
| 12 | {portdest:WELLKNOWN c2s_packets:(0-8498] s2c_packets:(0-7842] protocol:TCP} ≀ {{portdest:443 protocol:TCP_* c2s_packets:(0-4249] s2c_packets:(0-3921]} {portdest:80 protocol:TCP_* c2s_packets:(0-4249] s2c_packets:(0-3921]}} | 1 | 0,04 | -0,23 | -0,33 |

Esaminando tali risultati ci si avvede di un fatto in qualche modo inatteso, ovvero che i *Max-EGI* anti-dipendenti apportano, in generale, una conoscenza più facilmente interpretabile e chiara rispetto alla maggior parte di quelli positivamente correlati. La ragione è che sovente in questi ultimi le generalizzazioni fatte sugli *item* c2s_packets e s2c_packets hanno condotto alla costruzione di *itemset* $X \wr S$ dal contenuto informativo involuto e oscuro, in ragione della difficoltà per l'osservatore di sottrarre intuitivamente da $X$ i suoi discendenti in $S$ quando l'uno e gli altri sono congiunti a generalizzazioni più o meno ampie, ma diverse, dei due *item* suddetti. Ciò non significa che tali *Max-EGI* in generale non possiedano contenuto informativo, ma che esso, ove sia presente, richiede un'analisi ulteriore della cardinalità delle generalizzazioni coinvolte e quindi uno sforzo di ricerca e comparazione all'osservatore per poterlo estrarre.

Un secondo fatto, anch'esso inatteso ma solo fino a un certo punto, è che i *Max-EGI* anti-dipendenti, quanto più si scende nella soglia di supporto minimo, tanto più vengono a dominare la scena se si guarda solo alla forza della (in)correlazione senza tener conto del suo segno. Se a una soglia di supporto $\varepsilon = 5\%$ i due segni quasi si alternano, già con $\varepsilon = 1\%$ i positivamente correlati appaiono molto più rarefatti e quando si giunge a $\varepsilon = 0{,}1\%$, essi appaiono solo nelle prime quattro

posizioni per ricomparire solo dopo una lunga tratta di *Max-EGI* anti-dipendenti.

**Tabella 29** — *Primi dodici itemset in* M\G *con* $\boldsymbol{\varepsilon = 0,1\%}$ *e con i più alti valori di* $|\boldsymbol{\zeta}|$

| № | Itemset | DS | Max | | |
|---|---|---|---|---|---|
| | | | kulc | nptc | zeta |
| 1 | {protocol:GENERAL c2s_packets:(0-2253]} ≀ {{protocol:UDP_DNS c2s_packets:(0-140]} {protocol:UDP_* c2s_packets:(0-140]} {protocol:SSL/TLS c2s_packets:(0-140]} {protocol:TCP_* c2s_packets:(0-140]} {protocol:HTTP_GET c2s_packets:(0-140]}} | 1 | 1,00 | 1,00 | 0,99 |
| 2 | {protocol:GENERAL s2c_packets:(0-17723] } ≀ {{protocol:UDP_DNS s2c_packets:(0-1107]} {protocol:UDP_* s2c_packets:(0-1107]} {protocol:SSL/TLS s2c_packets:(0-1107]} {protocol:TCP_* s2c_packets:(0-1107]} {protocol:HTTP_GET s2c_packets:(0-1107]}} | 1 | 0,99 | 1,00 | 0,99 |
| 3 | {c2s_packets:(0-2253] s2c_packets:(0-727]} ≀ {{c2s_packets:(0-140] s2c_packets:(0-45]}} | 1 | 0,99 | 1,00 | 0,99 |
| 4 | {portsource:WELLKNOWN portdest:DYNAMIC} ≀ {{portsource:443 portdest:64381} {portsource:443 portdest:57651} {portsource:80 portdest:50668} {portsource:80 portdest:57665} {portsource:80 portdest:50641}} | 1 | 0,83 | 0,82 | 0,68 |
| 5 | {portsource:WELLKNOWN portdest:WELLKNOWN} ≀ {{portsource:138 portdest:138} {portsource:68 portdest:67} {portsource:67 portdest:68} {portsource:137 portdest:137}} | 2 | 0,01 | -0,64 | -0,58 |
| 6 | {portdest:WELLKNOWN ipsource:172.20.68.25} ≀ {{portdest:80 ipsource:172.20.68.25} {portdest:53 ipsource:172.20.68.25}} | 1 | 0,03 | -0,39 | -0,54 |
| 7 | {portdest:WELLKNOWN ipsource:172.20.68.127} ≀ {{portdest:53 ipsource:172.20.68.127} {portdest:443 ipsource:172.20.68.127} {portdest:80 ipsource:172.20.68.127}} | 1 | 0,03 | -0,38 | -0,53 |
| 8 | {portdest:WELLKNOWN ipsource:172.20.90.87} ≀ {{portdest:53 ipsource:172.20.90.87} {portdest:80 ipsource:172.20.90.87}} | 1 | 0,01 | -0,49 | -0,53 |
| 9 | {portdest:WELLKNOWN ipsource:172.20.63.52} ≀ {{portdest:80 ipsource:172.20.63.52} {portdest:53 ipsource:172.20.63.52} {portdest:443 ipsource:172.20.63.52}} | 1 | 0,07 | -0,27 | -0,53 |
| 10 | {portdest:WELLKNOWN ipsource:172.20.90.160} ≀ {{portdest:443 ipsource:172.20.90.160} {portdest:53 ipsource:172.20.90.160} {portdest:143 ipsource:172.20.90.160}} | 1 | 0,02 | -0,44 | -0,53 |
| 11 | {portdest:WELLKNOWN protocol:TCP} ≀ {{protocol:TCP_* portdest:443} {protocol:TCP_* portdest:80} {protocol:TCP_* portdest:143} {protocol:TCP_* portdest:82}} | 1 | 0,01 | -0,54 | -0,51 |
| 12 | {portdest:WELLKNOWN protocol:TCP} ≀ {{portdest:443 protocol:TCP_*} {portdest:143 protocol:TCP_*} {portdest:80 protocol:TCP_* }} | 1 | 0,01 | -0,53 | -0,51 |

Con riguardo all'interesse qualitativo vi sono esempi di *unexpected itemset* molto interessanti, come per es. {portdest:WELLKNOWN protocol:TCP} ≀ {{protocol:TCP_* portdest:443}}, che rivela, in uno specifico *dataset*, che al netto del traffico https, non ci si deve affatto aspettare che il traffico verso le porte *well-known* sia trasportato da TCP, laddove invece molti avrebbero scommesso sul ruolo di primo piano per esempio di http. Oppure informazioni interessanti sono fornite per es. da {portdest:REGISTERED protocol:UDP} ≀ {{portdest:5355 protocol:UDP_*}}, laddove lo si metta a confronto anche con il suo omologo positivo alla riga 22 di tabella 9, e cioè che al netto del *Link-Local Multicast Name Resolution* non ci si deve aspettare un grande traffico UDP verso le porte *registered*; oppure ancora dal *Max-EGI* {portsource:WELLKNOWN portdest:WELLKNOWN} ≀ { {portsource:138 portdest:138} {portsource:68 portdest:67} {portsource:67

portdest:68} {portsource:137 portdest:137} } che evidenzia come, al netto del DHCP e del NetBIOS, sia raro che il rimanente traffico si svolga tra porte entrambe *well-known*.

Oltre a questi, molti altri esempi qualitativamente significativi dello stesso tipo potrebbero essere reperiti, traendoli dalla generalità degli strati, ma essi invariabilmente sarebbero caratterizzati da anti-dipendenza e quindi utili nel particolare caso in cui si volesse trarne una conoscenza in termini di regole negative $X \to \overline{Y}$.

Volendo invece restare nell'ambito delle regole positive, gli esempi di interesse si diradano o perché non sono interessanti a sufficienza o perché per valutarli occorre integrare la loro informazione con altra da essi non immediatamente recata. Per esempio il *Max-EGI* {portdest:WELLKNOWN protocol:GENERAL} ≀ {{protocol:TCP_* portdest:443} {portdest:80 protocol:TCP_*} {protocol:UDP_* portdest:53} {portdest:53 protocol:UDP_DNS}} è sì positivamente correlato e facilmente interpretabile, ma da un lato la sua correlazione è tutto sommato modesta e dall'altro non può dirsi né inatteso né particolarmente utile scoprire che il traffico verso potenziali server non si limita ai soli *web server* o *dns server*. D'altro canto, i *Max-EGI* fortemente correlati positivamente necessitano spesso di integrazioni informative, laddove, per esempio, per interpretare i primi 3 *itemset* della tabella 28 o della tabella 29 ci si deve necessariamente munire di (o costruire) un prospetto recante le distribuzioni dei flussi inerenti ai vari protocolli coinvolti.

Riepilogando quanto sopra osservato, anche senza il conforto dell'esaustività dell'analisi si può razionalmente affermare che la composizione della partizione G\M è data da:

- *Max-EGI* a significativa anti-dipendenza, suscettibili di fornire una conoscenza anche di pregio dal punto di vista qualitativo, ma che è fungibile solo nell'ipotesi di impiegarla per la generazione di regole negative;
- *Max-EGI* a correlazione positiva, ancorché raramente capaci di raggiungerne i più alti gradi, il cui contributo di conoscenza è limitato a casi con modesta o nulla *actionability* o *unexpectedness*, e che poco aggiungono all'economia complessiva dell'analisi;
- *Max-EGI* a correlazione positiva, di vario grado, che recano talvolta una conoscenza potenzialmente di interesse qualitativo in virtù della loro originalità, ma che risulta di difficile estrazione perché involuta all'interno di sottrazioni insiemistiche di significato spesso oscuro all'osservatore quando egli non abbia accesso anche a informazioni ulteriori.

## 6.4 Considerazioni emergenti dall'analisi sulla base del dominio

L'analisi condotta in questo capitolo ha permesso, in via principale, di collocare gli *itemset* a correlazione positiva più suscettibili di recare conoscenza qualitativamente di interesse dal punto di vista di un'analisi *domain-driven* nella partizione G∩M e quelli a correlazione negativa qualitativamente interessanti o in M\G o ancora nella stessa G∩M. Nondimeno, l'analisi ha anche palesato un quadro dei risultati non del tutto soddisfacente in tutte e tre le partizioni. Talune ragioni di insoddisfazione sono già state evidenziate soprattutto nell'analisi della partizione M\G, ma ne sono risultate affette in verità tutte le partizioni considerate. Altre sono state solo *passim* accennate, ma hanno verosimilmente pesato sulla qualità dei risultati che, sulla base delle analisi condotte, avrebbero potuto essere migliori. Ci si vuol riferire in particolare da un lato alle generalizzazioni che sono state compiute sulle caratteristiche c2s_packets, s2c_packets e protocol; dall'altro alle generalizzazioni che non sono state affatto compiute sulle caratteristiche ipsource e ipdest.

Nonostante dal punto di vista astrattamente numerico la differenza da un lato tra un flusso costituito per esempio di 1 pacchetto ed uno costituito da 1000 e, dall'altro, quella tra un flusso di 10.000 pacchetti ed uno costituito da 50.000 possa sembrare più marcata nel secondo caso rispetto al primo, nella realtà, con riguardo al dominio applicativo, è esattamente il contrario. Mentre si possono generalizzare e riunire, mantenendo una semantica coerente, gli ultimi due flussi in un unico intervallo $(7000, 70000]$ rappresentante i flussi di dati dell'ordine delle decine di MiB, lo stesso non può farsi per i primi due, qual che sia l'intervallo scelto, senza far perdere di vista, per esempio, le differenze tra una connessione `TCP` che non ha superato il *three-way handshake* e non si è quindi mai neppure instaurata e un flusso limitato sì ma non esiguo di dati. La ragione sottostante è che i flussi di dati andrebbero considerati in modo logaritmico e non puramente lineare. Ad ogni modo, a prescindere dalla linearità e quale che sia la semantica che si vorrebbe dare alle generalizzazioni, dal punto di vista di un esperto del dominio l'informazione che un certo flusso di dati può avere un qualsiasi valore, per esempio, nell'intervallo $[0, 70000]$ è pacifico esser priva di significato concreto e laddove si è consentito agli algoritmi di far uso di generalizzazioni del tipo $[0, N]$, con l'estremo inferiore a zero e $N$ limitato superiormente solo dalla lunghezza del più lungo flusso di dati nel *dataset* di riferimento, si è irrimediabilmente inficiato il valore informativo recato dalle caratteristiche oggetto di tali generalizzazioni.

In tutte e tre le partizioni considerate, la presenza di *itemset* con siffatte generalizzazioni — non necessariamente le più estreme — è risultata molto elevata e ciò, invariabilmente, ne ha decretato la classificazione come *uninteresting*. In $\mathsf{G} \cap \mathsf{M}$ li si è quasi ovunque ignorati nell'analisi e in $\mathsf{M} \setminus \mathsf{G}$ li si è mostrati più che altro per evidenziarne le conseguenze involutive in termini interpretativi. La ragione della diminuzione dell'interesse è dovuta al fatto che nelle generalizzazioni meno estreme l'informazione recata è semplicemente ambigua, nelle più estreme diviene una generalizzazione a supporto (quasi) unitario e quindi inutile quanto quella recata dalla generalizzazione `{protocol:GENERAL}`. Al di là poi degli effetti diretti sulle caratteristiche direttamente affette da queste generalizzazioni, si deve considerare che queste hanno rappresentato una via preferenziale, si potrebbe dire più comoda, per scalare le tassonomie a discapito di generalizzazioni più utili, quali quelle su `ipsource` e `ipdest` che conseguentemente non si sono realizzate come previsto e desiderato.

Deve essere tuttavia ben chiaro che la ragione delle anomalie sopra evidenziate non può esser fatta risalire agli algoritmi di generalizzazione, ma alla scelta inappropriata o delle discretizzazioni operate sui dati o delle tassonomie fornite agli algoritmi. È in quelle sedi, pertanto che si sono introdotti i vizi che hanno ridotto e talvolta distorto l'effetto auspicato tanto della generalizzazione in sé quanto del recupero informativo per complementazione implementato dal *Max-EGI extraction algorithm*.

# Capitolo 7

# Conclusioni

L'analisi condotta sotto il profilo *domain-driven* nel capitolo 6 ha mostrato come gli *itemset* di maggior interesse siano stati tutti reperiti nella partizione $\mathsf{G}\cap\mathsf{M}$. Pur mettendo in debito conto la soggettività dell'analisi e la sua non esaustività, ciò è avvenuto perché gli *itemset* reperibili nella partizione $\mathsf{G}\setminus\mathsf{M}$ sono risultati sostanzialmente delle generalizzazioni ridondanti e quelli nella partizione $\mathrm{M}\setminus\mathsf{G}$ si sono rivelati sostanzialmente infungibili, a parte quelli ad elevata anti-dipendenza. Questo quadro trova sostegno, inoltre, anche nell'analisi oggettiva fondata su metriche condotta nel capitolo 5, laddove essa ha mostrato che gli *itemset* in $\mathsf{G}\cap\mathsf{M}$ — così come le regole da essi generate — hanno mediamente una correlazione totale più elevata rispetto a quelli delle altre due partizioni, sia complessivamente sia in modo ancor più spiccato quando si prendano per il confronto quelli oltre le soglie di neutralità delle misure descritte nel capitolo 4.

Se si considera che la partizione $\mathsf{G}\cap\mathsf{M}$ rappresenta mediamente il 97% degli *itemset* prodotti dal *Max-EGI extraction algorithm*, ma solamente il 47% degli *itemset* prodotti dal *Genio algorithm*, si giunge alla conclusione che il *Max-EGI extraction algorithm* è una tecnica efficiente ed efficace nel selezionare gli *itemset* di maggior interesse e nell'espungere quelli ridondanti o meno specializzati. Per converso, però, la generale infungibilità degli *itemset* in $\mathrm{M}\setminus\mathsf{G}$ fa concludere altresì per il fallimento, almeno per i dati sperimentali considerati, del *Max-EGI extraction algorithm* nel recuperare conoscenza per mezzo della sua innovativa tecnica di estrazione degli *itemset* ad alto livello che rappresentano dati non già ricompresi in alcuno dei loro discendenti frequenti a più basso livello.

Ancorché appaiano sufficientemente chiari i risultati del confronto tra i due algoritmi, nondimeno occorre anche determinare le cause di questi risultati; ovvero se e in quale misura essi siano dipesi dalle caratteristiche intrinseche dei dati o da come essi sono stati preparati e modellizzati o, ancora, dalla logica stessa degli algoritmi. L'analisi dei risultati ha fornito degli utili indizi laddove essa da un lato ha permesso di rinvenire molte generalizzazioni anomale e inattese, dall'altro non ha potuto rinvenire alcune generalizzazioni che sarebbero state invece attese. È proprio da tali indizi che può trarsi la giustificazione per i risultati che si sono ottenuti e si può giungere a concludere che vi sono stati errori o inadeguatezze sia nella discretizzazione di alcune caratteristiche sia nella formulazione di alcune delle tassonomie fornite agli algoritmi.

Uno degli errori principali che sono stati commessi nella preparazione dei dati sperimentali è stato di tipo metodologico. Se da un lato è necessario ed auspicabile che l'algoritmo di *data mining* generalizzante sia neutrale rispetto al significato intrinseco dei dati che esso va a trattare, dall'altro

è altrettanto necessario e auspicabile che tutta l'informazione sul dominio gli sia fornita in modo strettamente e fortemente specifico attraverso la tassonomia. Questa dicotomia (*neutralità algoritmi* ⇔ *specificità tassonomia*) non è stata seguita nella preparazione dei dati sperimentali laddove gli alberi di generalizzazione delle caratteristiche `c2s_packets`, `s2c_packets`, `ipsource` e `ipdest` sono stati costruiti in modo più o meno automatico — in altre parole in modo neutrale — senza tener conto della specificità del dominio. Da questa neutralità è disceso anche l'errore di aver voluto trattare le caratteristiche `c2s_packets`, `s2c_packets` in modo lineare, laddove la significatività di queste informazioni è invece più correttamente quella logaritmica, come si è già avuto modo di osservare. Un secondo errore è stato di tipo logico ed è stato commesso laddove si è consentito (`c2s_packets`, `s2c_packets`) o si è materialmente imposto (`protocol`) che le generalizzazioni potessero spingersi fino a raggiungere *label* equivalenti o quasi equivalenti al grado di radice dell'albero di generalizzazione, determinando il formarsi di *itemset* del tutto spuri giacché l'*itemset* che generalizza al grado di radice una delle caratteristiche dovrebbe essere espresso o esprimibile in via esclusiva solo da quel medesimo *itemset* privato della caratteristica medesima.

Le conseguenze degli errori e delle inadeguatezze e sopra menzionate hanno avuto un impatto diversificato sui risultati dell'applicazione degli algoritmi ai dati sperimentali. In particolare:

- le generalizzazioni spinte quasi al livello di radice hanno cagionato solamente il proliferare di *itemset* spuri, soprattutto nella partizione `G\M`, giacché il *Max-EGI extraction algorithm* si è fatto carico — e ciò va ad aggiungersi ai pregi già precedentemente evidenziati — di rimuovere la gran parte di essi;
- gli alberi di generalizzazione costruiti in modo automatico e talvolta adattativo sulle caratteristiche `c2s_packets`, `s2c_packets`, `ipsource` e `ipdest` hanno cagionato, soprattutto per le prime due, una riduzione talvolta fatale della portata informativa recata da queste negli *itemset* su cui hanno agito, con effetti su entrambi gli algoritmi e sulla qualità complessiva dei risultati;
- gli alberi di generalizzazione automatici e adattativi su `c2s_packets`, `s2c_packets`, inoltre, hanno determinato il fallimento del *Max-EGI extraction algorithm* nel recupero di conoscenza mediante complementazione, giacché tale complementazione è avvenuta in gran parte proprio su generalizzazioni ambigue o semanticamente oscure proprio di queste due caratteristiche o, comunque, su *itemset* oggetto anche solo in parte di queste generalizzazioni.

Individuate nei motivi sopraddetti le cause che hanno diminuito l'efficacia delle metodologie di *data mining* nella fattispecie analizzata in questo lavoro, si può comprendere quali alberi di generalizzazione avrebbero potuto essere più convenientemente utilizzati. Se per quanto riguarda la caratteristica `protocol` sarebbe sufficiente rimuovere dall'albero di figura 5(b) la *label* «`GENERAL»`, per le altre quattro caratteristiche è necessario riformulare completamente l'albero di generalizzazione.

Per quanto riguarda le caratteristiche `ipsource` e `ipdest`, in figura 30 è riportato un esempio di albero più confacente alle caratteristiche del dominio e in grado di mantenere, al più alto livello di aggregazione, l'informazione sull'indirizzamento `IPv4` almeno nelle tre superclassi «`PRIVATE`», «`PUBLIC`» e «`NOT-UNICAST`», intendendosi l'ultima come l'unione dello spazio di indirizzamento del *multicast* con quello *broadcast*. I livelli intermedi sono basati su tre maschere di sottorete rispettivamente di 8, 16 e 24 bit, le quali, pur non potendosi adattare in ogni caso alla ampia variabilità cagionata dal *Classless Inter-Domain Routing* (`CIDR`), forniscono nondimeno delle informazioni

comunque pertinenti e di facile lettura. Vi è da dire che, a seconda della specialità del traffico di rete da analizzare e del supporto minimo che si intende utilizzare, il livello a 8 bit potrebbe anche essere superfluo e rimuovibile in taluni casi.

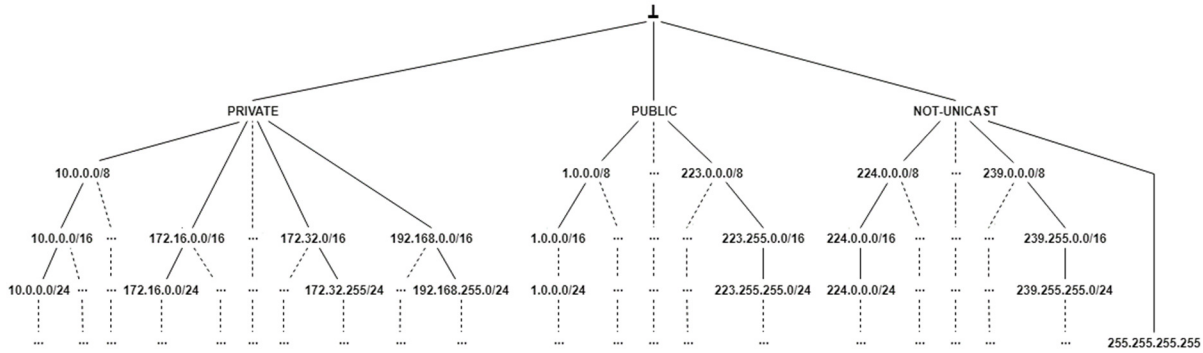



**Figura 30** — *Albero di generalizzazione GT per le caratteristiche* `ipsource` *e* `ipdest` *riformulato*

La riformulazione degli alberi relativi alle caratteristiche `c2s_packets`, `s2c_packets` è uno degli aspetti più critici nella definizione della tassonomia. Ciò dipende dal fatto che l'aderenza generica al dominio applicativo delle reti è solo una delle condizioni da rispettare; oltre a questa vi è la necessità di calibrarne l'estensione e la granularità in funzione del particolare traffico che si vuole analizzare e anche del supporto minimo che si prevede di porre quale soglia. Raramente una formulazione generica dell'albero, ancorché pienamente aderente al particolare dominio della caratteristica, potrà valere in tutte le occasioni e quindi, come avviene in generale nel *data mining*, dovranno essere valutate e testate più ipotesi prima di giungere ad una pienamente soddisfacente. Lo schema proposto nella figura 31 è pertanto solo una base di partenza semanticamente consistente, ma esemplificativa, sulla quale andranno apportate le opportune calibrazioni.

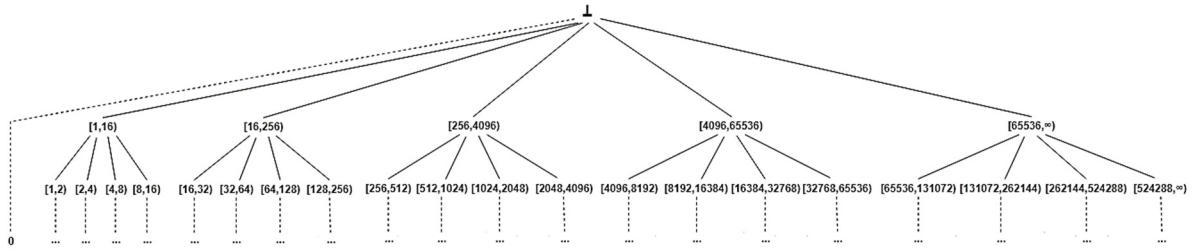



**Figura 31** — *Albero di generalizzazione GT per le caratteristiche* `c2s_packets` *e* `s2c_packets` *riformulato*

Fatta questa premessa, l'albero di generalizzazione proposto fonda la sua applicabilità almeno su questi presupposti:

- la discriminazione tra flussi di diverse dimensioni è semanticamente tanto più rilevante quanto più essi sono piccoli;
- i flussi sono tanto meno frequenti quanto più sono di elevate dimensioni, talché è necessario che l'ampiezza degli intervalli di aggregazione cresca col crescere delle dimensioni dei flussi da aggregare;
- il livello più alto dell'albero prima della radice deve conservare una associabilità semantica del traffico che etichetta alle dimensioni attese o note del traffico reale, quali per esempio quelle del traffico `DNS` o `DHCP`, del traffico Web, dello *streaming* audio e video, del `P2P`, *et cetera*.

Si noti, in figura 31, che è previsto anche un traffico di pacchetti pari a zero. Questo vale

naturalmente per il solo albero di `s2c_packets` ed etichetta tipicamente i tentativi di un *client* di raggiungere un *server* senza ricevere alcuna risposta; è appena il caso di far notare che questo traffico, tutt'altro che infrequente, non dovrebbe essere aggregato in nessun caso, pena l'irrimediabile deterioramento dell'informazione recata dalla classe aggregata.

In questa tesi si è mostrato come il *Max-EGI extraction algorithm* sia più efficiente ed efficace rispetto ai tradizionali algoritmi di generalizzazione nell'estrarre co-occorrenze e *pattern actionable* o *unexpected* dal traffico di rete *wireless*, ancorché in presenza di tassonomie non del tutto congrue con lo specifico ambito applicativo. Ulteriori *test* dovrebbero essere quindi esperiti, utilizzando alberi di generalizzazione più specifici, per una valutazione ancor più attendibile della sua efficacia in contesti determinati. Nella tesi si è inoltre introdotta, tra le altre, una nuova misura di interesse per gli *itemset*, che si è mostrata efficace per i dati sperimentali analizzati e che potrebbe essere interessante applicare anche a *dataset* provenienti da una pluralità di ambiti applicativi per testarne l'efficacia e l'utilizzabilità in senso più generale nelle attività di KDD e *data mining*.

# Bibliografia


[1] R. Agrawal, T. Imielinski and Swami, "Mining Associations rules between sets of items in large databases," *ACM SIGMOD 1993,* pp. 207-216, 1993.

[2] R. Srikant and R. Agrawal, "Mining generalized association rules," *VLDB 1995,* pp. 407-419, 1995.

[3] E. Baralis, L. Cagliero, T. Cerquitelli, V. D'Elia and P. Garza, "Expressive generalized itemsets," *Information Sciences 278,* pp. 327-343, 2014.

[4] F. Risso, E. Baralis and M. Baldi, "Data Mining Techniques For Effective and Scalable Traffic," in *9th IFIP/IEEE International Symposium on Integrated Network Management (IM05), May 2005*, Nice, France, 2005.

[5] E. Osmani, "Analysis of wireless network data by means of generalized association rules," Tesi di Laurea, Torino, 2015.

[6] L. Cagliero, "Data mining by means of generalized patterns," PhD thesis, Torino, 2012.

[7] S. Brin, R. Motwani, J. D. Ullman and S. Tsur, "Dynamic itemset counting and implication rules for market basket data," *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data,* pp. 255-264, 1997.

[8] T. Wu, Y. Chen and J. Han, "Association Mining in Large Databases: A Re-Examination of Its Measures," *Proc. 2007 Int. Conf. on Principles and Practice of Knowledge Discovery in Databases ,* p. 621–628, 2007.

[9] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of Research and Development,* pp. 66 - 82, 1960.

[10] M. Studený and J. Vejnarová, "The Multiinformation Function as a Tool for Measuring Stochastic Dependence," *Jordan M.I. (eds) Learning in Graphical Models. NATO ASI Series (Series D: Behavioural and Social Sciences),* vol. 98, 1998.

[11] G. Bouma, "Normalized (Pointwise) Mutual Information in Collocation Extraction," *Proceedings of the Biennial GSCL Conference 2009,* 2009.

[12] H. Xiong, P.-N. Tan and V. Kumar, "Mining strong affinity association patterns in data sets with skewed support distribution," in *Third IEEE International Conference on Data Mining*, Melbourne, FL, USA, 2003.

# Appendice A

# Listati dei moduli di caricamento

**Listato A.1** — *Modulo ETL principale*

```
@echo off
setlocal EnableDelayedExpansion

set Server=HP-XW6200\SQLEXPRESS
set DB=CAPTURES
set Driver={SQL Server}
set DBconn="Driver=%Driver%;server=%Server%;Database=%DB%;Trusted_Connection=yes"
set LPtext=-o:SQL -oConnString:%DBconn% -ignoreIdCols:ON
set LPtext=%LPtext% -transactionRowCount:-1 -i:TEXTLINE -iw:OFF -stats:OFF
set LPcsvs=-o:SQL -oConnString:%DBconn% -dtLines:0 -i:CSV -iw:OFF -stats:OFF

:begin
echo ETL process started at %TIME% ...

:createDB
set Status=Creating Database & echo: & echo   !Status!
sqlcmd -S %Server% -d %DB% -E -i make_DB.sql -b
if !errorlevel! neq 0 goto :error

:loadDatasets
set Status=Loading Datasets & echo: & echo   !Status!:
for /D %%c in (*) do (
  for /D %%r in (%%c\itemset*_0.1*) do (
      echo     Processing %%r\outlog.dataset
      logparser file:load_Datasets.sql?file=%%r\outlog.dataset %LPtext%
      if !errorlevel! neq 0 goto :error
  )
)

:loadMaxEGIs
set Status=Loading MaxEGIs & echo: & echo   !Status!:
for /D %%c in (*) do (
  for /D %%r in (%%c\itemset*) do (
    for %%f in (%%r\translatedMax*.txt) do (
         echo     Processing %%f
         logparser file:load_Max_EGIs.sql?file=%%f %LPtext%
         if !errorlevel! neq 0 goto :error
    )
  )
)
```

*continua*

*continua*

```
:collectSparseFiles
set Status=Collecting sparse files & echo: & echo   !Status!:
for /D %%c in (*) do (
  for %%s in (0.1 0.2 0.4 0.8 1.0 1.5 2.0 3.0 4.0 5.0) do (
    if exist %%c\*WellFormedoutLogsup%%s0AllTestuali.txt (
      if not exist %%c\results_%%s_4 (
        echo     Processing %%c
        md %%c\sparses_%%s_4
        copy %%c\*WellFormedoutLogsup%%s0AllTestuali.txt %%c\sparses_%%s_4 > NUL:
      )
    )
  )
)

:loadGenItemsets
set Status=Loading Generalized Itemsets & echo: & echo   !Status!:
for /D %%c in (*) do (
  for /D %%r in (%%c\results* %%c\sparses*) do (
    for %%f in (%%r\itemsetsWell*Testuali.txt) do (
         echo     Processing %%f
         logparser file:load_Gen_Itemsets.sql?file=%%f %LPtext%
         if !errorlevel! neq 0 goto :error
    )
  )
)

:loadGenRules
set Status=Loading rules & echo: & echo   !Status!:
for /D %%c in (*) do (
  for /D %%r in (%%c\results* %%c\sparses*) do (
    for %%f in (%%r\regoleWell*.txt) do (
         echo     Processing %%f
         logparser file:load_Gen_Rules.sql?file=%%f %LPtext%
         if !errorlevel! neq 0 goto :error
    )
  )
)

:clean
set Status=Cleaning sparse files & echo: & echo   !Status!
for /D %%c in (*) do (
  for %%s in (0.1 0.2 0.4 0.8 1.0 1.5 2.0 3.0 4.0 5.0) do (
    if exist %%c\sparses_%%s_4 rd %%c\sparses_%%s_4 /S /Q
  )
)

:indexDB
set Status=Creating Indexes & echo: & echo   !Status!
sqlcmd -S %Server% -d %DB% -E -i make_indexes.sql -b
if !errorlevel! neq 0 goto :error

:wipeSpurious
set Status=Wiping spurious itemset & echo: & echo   !Status!
sqlcmd -S %Server% -d %DB% -E -i wipe_spurious.sql -b > NUL:
if !errorlevel! neq 0 goto :error

:calculateMeasures
set Status=Calculating measures & echo: & echo   !Status!
sqlcmd -S %Server% -d %DB% -E -i make_measures.sql -b > NUL:
if !errorlevel! neq 0 goto :error
```

*continua*

*continua*

```
echo:
echo Done at %TIME% !
goto :end

:error
echo PANIC: An error raised in %Status%
pause

:end
endlocal
```

**Listato A.2** — *Modulo di creazione della base di dati (make_DB.sql)*

```
CREATE TABLE DATASET
(
   site          varchar(2),
   day#          smallint,
   ipsource      varchar(255),
   portsource    varchar(255),
   c2s_packets   varchar(255),
   ipdest        varchar(255),
   portdest      varchar(255),
   s2c_packets   varchar(255),
   protocol      varchar(255),
);

CREATE TABLE MAX_EGI
(
   site          varchar(2),
   day#          smallint,
   minsupp       real,
   ipsource      varchar(255),
   portsource    varchar(255),
   c2s_packets   varchar(255),
   ipdest        varchar(255),
   portdest      varchar(255),
   s2c_packets   varchar(255),
   protocol      varchar(255),
   S             varchar(4096),
   k             smallint,
   f             int,
   fS            int,
   supp          real,
   kulc          real,
   nptc          real,
   zeta          real
);

CREATE TABLE GEN_ITEMSET
(
   site          varchar(2),
   day#          smallint,
   minsupp       real,
   ipsource      varchar(255),
   portsource    varchar(255),
   c2s_packets   varchar(255),
   ipdest        varchar(255),
   portdest      varchar(255),
   s2c_packets   varchar(255),
   protocol      varchar(255),
   k             smallint,
   f             int,
   supp          real,
   kulc          real,
   nptc          real,
   zeta          real
);
```

*continua*

*continua*

```
CREATE TABLE GEN_RULE
(
   site          varchar(2),
   day#          smallint,
   minsupp       real,
   l_ipsource    varchar(255),
   l_portsource  varchar(255),
   l_c2s_packets varchar(255),
   l_ipdest      varchar(255),
   l_portdest    varchar(255),
   l_s2c_packets varchar(255),
   l_protocol    varchar(255),
   r_ipsource    varchar(255),
   r_portsource  varchar(255),
   r_c2s_packets varchar(255),
   r_ipdest      varchar(255),
   r_portdest    varchar(255),
   r_s2c_packets varchar(255),
   r_protocol    varchar(255),
   k             smallint,
   supp          real,
   conf          real,
   lift          real
)
```

**Listato A.3** — *Modulo di caricamento dei dataset (load_datasets.sql)*

```
SELECT
   site
  ,day#
  ,ipsource
  ,portsource
  ,c2s_packets
  ,ipdest
  ,portdest
  ,s2c_packets
  ,protocol

USING
   CASE EXTRACT_PREFIX('%file%', 0, '_')
      WHEN 'Aulamensa' THEN 'M'
      WHEN 'Aulasegre' THEN 'S'
   END AS site,

   CASE site
      WHEN 'M' THEN --Extract Day# for site #1, set it to 6 otherwise
         COALESCE(TO_INT(SUBSTR(EXTRACT_SUFFIX('%file%', 0, 'giorno_'),0,1)), 6)
      WHEN 'S' THEN --Extract Day# for site #2, set it to 4 otherwise
         COALESCE(TO_INT(SUBSTR(EXTRACT_SUFFIX('%file%', 0, 'Aulasegre_'),0,1)), 4)
   END AS day#,

   REPLACE_CHR(EXTRACT_PREFIX(Text, 0, ' \\'),':','=') AS Items,

   EXTRACT_VALUE(Items, 'ipsource', ' ')    AS ipsource,
   EXTRACT_VALUE(Items, 'portsource', ' ')  AS portsource,
   EXTRACT_VALUE(Items, 'c2s_packets', ' ') AS c2s_packets,
   EXTRACT_VALUE(Items, 'ipdest', ' ')      AS ipdest,
   EXTRACT_VALUE(Items, 'portdest', ' ')    AS portdest,
   EXTRACT_VALUE(Items, 's2c_packets', ' ') AS s2c_packets,
   EXTRACT_VALUE(Items, 'protocol', ' ')    AS protocol

INTO
   DATASET

FROM
   %file%

WHERE
  TRIM(Text) IS NOT NULL
```

**Listato A.4** — *Modulo di caricamento dei Max-Egi (load_Max_Egis.sql)*

```
SELECT
   site, day#, minsupp,
   ipsource, portsource, c2s_packets, ipdest, portdest, s2c_packets, protocol, S,
   k, f, fS, NULL, NULL, NULL

USING
   CASE EXTRACT_PREFIX('%file%', 0, '_')
      WHEN 'Aulamensa' THEN 'M'
      WHEN 'Aulasegre' THEN 'S'
   END AS site,

   CASE site
      WHEN 'M' THEN --Extract Day# for site #1, set it to 6 otherwise
         COALESCE(TO_INT(SUBSTR(EXTRACT_SUFFIX('%file%', 0, 'giorno_'),0,1)), 6)
      WHEN 'S' THEN --Extract Day# for site #2, set it to 4 otherwise
         COALESCE(TO_INT(SUBSTR(EXTRACT_SUFFIX('%file%', 0, 'Aulasegre_'),0,1)), 4)
   END AS day#,

   TO_REAL(EXTRACT_SUFFIX('%file%', 0, 'Logsup')) AS minsupp,
   EXTRACT_PREFIX(Text, 0, ' (')  AS Body,
   EXTRACT_SUFFIX(Text, 0, ' (')  AS Measures,
   REPLACE_CHR(EXTRACT_PREFIX(Body, 0, ' \\'),':','=') AS Items,
   TO_INT(TO_REAL(EXTRACT_TOKEN (Measures, 0, ','))) AS f,
   TO_INT(TO_REAL(EXTRACT_SUFFIX(Measures, 0, '='))) AS fS,

   EXTRACT_VALUE(Items, 'ipsource', ' ')    AS ipsource,
   EXTRACT_VALUE(Items, 'portsource', ' ')  AS portsource,
   EXTRACT_VALUE(Items, 'c2s_packets', ' ') AS c2s_packets,
   EXTRACT_VALUE(Items, 'ipdest', ' ')      AS ipdest,
   EXTRACT_VALUE(Items, 'portdest', ' ')    AS portdest,
   EXTRACT_VALUE(Items, 's2c_packets', ' ') AS s2c_packets,
   EXTRACT_VALUE(Items, 'protocol', ' ')    AS protocol,

   -- Extract S set
   CASE fS
      WHEN 0 THEN NULL
      WHEN NULL THEN NULL
      ELSE EXTRACT_SUFFIX(Body, 0, '\\ ')
   END AS S,

   --- Calculate Length
   ADD(CASE ipsource WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE portsource WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE c2s_packets WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE ipdest WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE portdest WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE s2c_packets WHEN NULL THEN 0 ELSE 1 END,
       CASE protocol WHEN NULL THEN 0 ELSE 1 END)))))) AS k

INTO
   MAX_EGI

FROM
   %file%

WHERE
  (k > 0)
```

**Listato A.5** — *Modulo di caricamento degli itemset generalizzati (load_Gen_Itemsets.sql)*

```
SELECT
   site, day#, minsupp,
   ipsource, portsource, c2s_packets, ipdest, portdest, s2c_packets, protocol,
   k, f, NULL, NULL, NULL

USING
   CASE EXTRACT_PREFIX('%file%', 0, '_')
      WHEN 'Aulamensa' THEN 'M'
      WHEN 'Aulasegre' THEN 'S'
   END AS site,

   CASE site
      WHEN 'M' THEN --Extract Day# for site #1, set it to 6 otherwise
         COALESCE(TO_INT(SUBSTR(EXTRACT_SUFFIX('%file%', 0, 'giorno_'),0,1)), 6)
      WHEN 'S' THEN --Extract Day# for site #2, set it to 4 otherwise
         COALESCE(TO_INT(SUBSTR(EXTRACT_SUFFIX('%file%', 0, 'Aulasegre_'),0,1)), 4)
   END AS day#,

   TO_REAL(EXTRACT_SUFFIX('%file%', 0, 'Logsup')) AS minsupp,
   EXTRACT_PREFIX(Text, 0, ' (')  AS Body,
   EXTRACT_SUFFIX(Text, 0, ' (')  AS Measures,
   REPLACE_CHR(EXTRACT_PREFIX(Body, 0, ' \\'),':','=') AS Items,
   TO_INT(TO_REAL(EXTRACT_TOKEN (Measures, 0, ','))) AS f,

   EXTRACT_VALUE(Items, 'ipsource', ' ')    AS ipsource,
   EXTRACT_VALUE(Items, 'portsource', ' ')  AS portsource,
   EXTRACT_VALUE(Items, 'c2s_packets', ' ') AS c2s_packets,
   EXTRACT_VALUE(Items, 'ipdest', ' ')      AS ipdest,
   EXTRACT_VALUE(Items, 'portdest', ' ')    AS portdest,
   EXTRACT_VALUE(Items, 's2c_packets', ' ') AS s2c_packets,
   EXTRACT_VALUE(Items, 'protocol', ' ')    AS protocol,

   --- Calculate Length
   ADD(CASE ipsource WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE portsource WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE c2s_packets WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE ipdest WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE portdest WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE s2c_packets WHEN NULL THEN 0 ELSE 1 END,
       CASE protocol WHEN NULL THEN 0 ELSE 1 END)))))) AS k

INTO
   GEN_ITEMSET

FROM
   %file%

WHERE
   (k > 0)
```

**Listato A.6** — *Modulo di caricamento delle regole associative (load_Gen_Rules.sql)*

```
SELECT site, day#, minsupp,
       l_ipsource, l_portsource, l_c2s_packets, l_ipdest, l_portdest, l_s2c_packets, l_protocol,
       r_ipsource, r_portsource, r_c2s_packets, r_ipdest, r_portdest, r_s2c_packets, r_protocol,
       k, supp, conf, lift

USING
   CASE EXTRACT_PREFIX('%file%', 0, '_')
      WHEN 'Aulamensa' THEN 'M'
      WHEN 'Aulasegre' THEN 'S'
   END AS site,

   CASE site
      WHEN 'M' THEN --Extract Day# for site #1, set it to 6 otherwise
         COALESCE(TO_INT(SUBSTR(EXTRACT_SUFFIX('%file%', 0, 'giorno_'),0,1)), 6)
      WHEN 'S' THEN --Extract Day# for site #2, set it to 4 otherwise
         COALESCE(TO_INT(SUBSTR(EXTRACT_SUFFIX('%file%', 0, 'Aulasegre_'),0,1)), 4)
   END AS day#,

   TO_REAL(EXTRACT_SUFFIX('%file%', 0, 'Logsup')) AS minsupp,
   EXTRACT_PREFIX(REPLACE_CHR(Text,':','='), 0, ' => ') AS Antecedent,
   EXTRACT_SUFFIX(REPLACE_CHR(Text,':','='), 0, ' => ') AS Body,
   EXTRACT_PREFIX(Body, 0, '  (' ) AS Consequent,
   EXTRACT_SUFFIX(Body, 0, '  (' ) AS Measures,
   EXTRACT_VALUE(Antecedent, 'ipsource', ' ')    AS l_ipsource,
   EXTRACT_VALUE(Antecedent, 'portsource', ' ')  AS l_portsource,
   EXTRACT_VALUE(Antecedent, 'c2s_packets', ' ') AS l_c2s_packets,
   EXTRACT_VALUE(Antecedent, 'ipdest', ' ')      AS l_ipdest,
   EXTRACT_VALUE(Antecedent, 'portdest', ' ')    AS l_portdest,
   EXTRACT_VALUE(Antecedent, 's2c_packets', ' ') AS l_s2c_packets,
   EXTRACT_VALUE(Antecedent, 'protocol', ' ')    AS l_protocol,
   EXTRACT_VALUE(Consequent, 'ipsource', ' ')    AS r_ipsource,
   EXTRACT_VALUE(Consequent, 'portsource', ' ')  AS r_portsource,
   EXTRACT_VALUE(Consequent, 'c2s_packets', ' ') AS r_c2s_packets,
   EXTRACT_VALUE(Consequent, 'ipdest', ' ')      AS r_ipdest,
   EXTRACT_VALUE(Consequent, 'portdest', ' ')    AS r_portdest,
   EXTRACT_VALUE(Consequent, 's2c_packets', ' ') AS r_s2c_packets,
   EXTRACT_VALUE(Consequent, 'protocol', ' ')    AS r_protocol,
   TO_REAL(EXTRACT_TOKEN(Measures, 0, ',')) AS supp,
   TO_REAL(EXTRACT_TOKEN(Measures, 1, ',')) AS conf,
   TO_REAL(EXTRACT_TOKEN(Measures, 2, ',')) AS lift,
   --- Calculate Length
   ADD(CASE l_ipsource WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE l_portsource WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE l_c2s_packets WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE l_ipdest WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE l_portdest WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE l_s2c_packets WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE l_protocol WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE r_ipsource WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE r_portsource WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE r_c2s_packets WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE r_ipdest WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE r_portdest WHEN NULL THEN 0 ELSE 1 END,
   ADD(CASE r_s2c_packets WHEN NULL THEN 0 ELSE 1 END,
       CASE r_protocol WHEN NULL THEN 0 ELSE 1 END ))))))))))))) AS k

INTO GEN_RULE FROM %file% WHERE (k > 0)
```

**Listato A.7(a)** — *Modulo di generazione delle misure (estratto da make_Measures.sql)*

```
CREATE FUNCTION Kulc(@1 INT, @2 INT, @3 INT, @4 INT, @5 INT, @6 INT, @7 INT, @f FLOAT, @fS INT, @k SMALLINT)
   RETURNS FLOAT
   BEGIN
      DECLARE @Kulc AS FLOAT = 1.0E0
      IF (@k > 1) SET @Kulc =
      (
         CASE WHEN @1 IS NULL THEN 0 ELSE @f/(@1 - @fS) END +
         CASE WHEN @2 IS NULL THEN 0 ELSE @f/(@2 - @fS) END +
         CASE WHEN @3 IS NULL THEN 0 ELSE @f/(@3 - @fS) END +
         CASE WHEN @4 IS NULL THEN 0 ELSE @f/(@4 - @fS) END +
         CASE WHEN @5 IS NULL THEN 0 ELSE @f/(@5 - @fS) END +
         CASE WHEN @6 IS NULL THEN 0 ELSE @f/(@6 - @fS) END +
         CASE WHEN @7 IS NULL THEN 0 ELSE @f/(@7 - @fS) END
      ) / @k

      RETURN @Kulc
   END;
GO

CREATE FUNCTION ElGi(@1 INT,@2 INT,@3 INT,@4 INT,@5 INT,@6 INT,@7 INT, @N INT, @f INT, @fS INT, @k SMALLINT)
   RETURNS FLOAT
   BEGIN
      DECLARE @Lg AS FLOAT
      DECLARE @Px AS FLOAT = (
         CASE WHEN @1 IS NULL THEN 0 ELSE LOG(@1 - @fS) END +
         CASE WHEN @2 IS NULL THEN 0 ELSE LOG(@2 - @fS) END +
         CASE WHEN @3 IS NULL THEN 0 ELSE LOG(@3 - @fS) END +
         CASE WHEN @4 IS NULL THEN 0 ELSE LOG(@4 - @fS) END +
         CASE WHEN @5 IS NULL THEN 0 ELSE LOG(@5 - @fS) END +
         CASE WHEN @6 IS NULL THEN 0 ELSE LOG(@6 - @fS) END +
         CASE WHEN @7 IS NULL THEN 0 ELSE LOG(@7 - @fS) END -
         @k * LOG(@N)
      )
      DECLARE @ls AS FLOAT = LOG(@f) - LOG(@N)

      IF (@k = 1)
         SET @Lg = 1.0E0
      ELSE IF (@ls = 0)
         SET @Lg = EXP(@ls)
      ELSE IF (@k = 2)
         SET @Lg = EXP(@ls - @Px)
      ELSE BEGIN
         DECLARE @m1 AS INT;
         DECLARE @m2 AS INT;
         DECLARE @ni AS FLOAT
         DECLARE @g2 AS FLOAT = EXP((2.0E0 / @k) * @Px)
         SELECT  @m1 = MIN(x) FROM (values (@1),(@2),(@3),(@4),(@5),(@6),(@7)) AS p(x)
         SELECT  @m2 = MAX(x) FROM (SELECT TOP 2 x
                              FROM (values (@1),(@2),(@3),(@4),(@5),(@6),(@7)) AS p(x)
                              WHERE x IS NOT NULL ORDER BY x ASC) AS m(x)
         SET @ni = SQRT(@m1 - @fS) * SQRT(@m2 - @fS)
         SET @Lg = SQRT(@f / (@ni * @g2))
      END

      RETURN @Lg
   END;
GO
```

*continua*

*continua*

```
CREATE FUNCTION Nptc(@1 INT,@2 INT,@3 INT,@4 INT,@5 INT,@6 INT,@7 INT, @N INT, @f INT, @fS INT, @k SMALLINT)
   RETURNS FLOAT
   BEGIN
      DECLARE @nptc AS FLOAT
      DECLARE @elle AS FLOAT
      DECLARE @ln_s AS FLOAT = LOG(@f) - LOG(@N)

      IF (@k = 1)
         SET @nptc = 0.0E0
      ELSE IF (@ln_s = 0)
         SET @nptc = 1.0E0
      ELSE IF (@k = 2 OR @k = 3)
         SET @nptc = -LOG(dbo.ElGi(@1, @2, @3, @4, @5, @6, @7, @N, @f, @fS, @k)) / @ln_s
      ELSE BEGIN
         SET @elle = LOG(@f) + (@k - 1) * LOG(@N) -
         (
            CASE WHEN @1 IS NULL THEN 0.0E0 ELSE LOG(@1 - @fS) END +
            CASE WHEN @2 IS NULL THEN 0.0E0 ELSE LOG(@2 - @fS) END +
            CASE WHEN @3 IS NULL THEN 0.0E0 ELSE LOG(@3 - @fS) END +
            CASE WHEN @4 IS NULL THEN 0.0E0 ELSE LOG(@4 - @fS) END +
            CASE WHEN @5 IS NULL THEN 0.0E0 ELSE LOG(@5 - @fS) END +
            CASE WHEN @6 IS NULL THEN 0.0E0 ELSE LOG(@6 - @fS) END +
            CASE WHEN @7 IS NULL THEN 0.0E0 ELSE LOG(@7 - @fS) END
         )
         SET @nptc = (POWER(2.0E0, @elle / @ln_s) - 1.0E0)/(POWER(2.0E0, 1 - @k) -1.0E0)
      END

      RETURN @nptc
   END;
GO

CREATE FUNCTION Zeta(@1 INT,@2 INT,@3 INT,@4 INT,@5 INT,@6 INT,@7 INT, @N INT, @f INT, @fS INT, @k SMALLINT)
   RETURNS FLOAT
   BEGIN
      DECLARE @ElGi AS FLOAT = dbo.ElGi(@1, @2, @3, @4, @5, @6, @7, @N, @f, @fS, @k)
      DECLARE @Kulc AS FLOAT = dbo.Kulc(@1, @2, @3, @4, @5, @6, @7,     @f, @fS, @k)

      RETURN (1.0E0/3.0E0) * (4.0E0 * @Kulc * (1.0E0 - 1.0E0 / (2.0E0 * @ElGi)) - 1.0E0)
   END;
GO
```

**Listato A.7(b)** — *Modulo di generazione delle viste delle partizioni (estratto da make_Measures.sql)*

```
CREATE VIEW GiM  -- GEN_ITEMSET intersect MAX_EGI
(
   site, day#, minsupp,
   ipsource, portsource, c2s_packets, ipdest, portdest, s2c_packets, protocol,
   k, f, supp, kulc, nptc, zeta
)
AS
(
   SELECT *
   FROM GEN_ITEMSET AS G
   WHERE EXISTS
   (
      SELECT *
      FROM MAX_EGI AS M
      WHERE
        (S IS NULL) AND
        (G.site = M.site) AND
        (G.day# = M.day#) AND
        (G.minsupp = m.minsupp) AND
        (G.ipsource = M.ipsource OR G.ipsource IS NULL AND M.ipsource IS NULL) AND
        (G.portsource = M.portsource OR G.portsource IS NULL AND M.portsource IS NULL) AND
        (G.c2s_packets = M.c2s_packets OR G.c2s_packets IS NULL AND M.c2s_packets IS NULL) AND
        (G.ipdest = M.ipdest OR G.ipdest IS NULL AND M.ipdest IS NULL) AND
        (G.portdest = M.portdest OR G.portdest IS NULL AND M.portdest IS NULL) AND
        (G.s2c_packets = M.s2c_packets OR G.s2c_packets IS NULL AND M.s2c_packets IS NULL) AND
        (G.protocol = M.protocol OR G.protocol IS NULL AND M.protocol IS NULL)
   )
)

CREATE VIEW GxM  -- GEN_ITEMSET except MAX_EGI
(
   site, day#, minsupp,
   ipsource, portsource, c2s_packets, ipdest, portdest, s2c_packets, protocol,
   k, f, supp, kulc, nptc, zeta
)
AS
(
   SELECT *
   FROM GEN_ITEMSET AS G
   WHERE NOT EXISTS
   (
      SELECT *
      FROM MAX_EGI AS M
      WHERE
        (S IS NULL AND
        (G.site = M.site) AND
        (G.day# = M.day#) AND
        (G.minsupp = M.minsupp) AND
        (G.ipsource = M.ipsource OR G.ipsource IS NULL AND M.ipsource IS NULL) AND
        (G.portsource = M.portsource OR G.portsource IS NULL AND M.portsource IS NULL) AND
        (G.c2s_packets = M.c2s_packets OR G.c2s_packets IS NULL AND M.c2s_packets IS NULL) AND
        (G.ipdest = M.ipdest OR G.ipdest IS NULL AND M.ipdest IS NULL) AND
        (G.portdest = M.portdest OR G.portdest IS NULL AND M.portdest IS NULL) AND
        (G.s2c_packets = M.s2c_packets OR G.s2c_packets IS NULL AND M.s2c_packets IS NULL) AND
        (G.protocol = M.protocol OR G.protocol IS NULL AND M.protocol IS NULL)
   )
)
```

*continua*

*continua*

```
CREATE VIEW MxG -- MAX_EGI except GEN_ITEMSET
(
   site, day#, minsupp,
   ipsource, portsource, c2s_packets, ipdest, portdest, s2c_packets, protocol, S,
   k, f, fS, supp, kulc, nptc, zeta
)
AS
(
   SELECT DISTINCT *
   FROM MAX_EGI
   WHERE S IS NOT NULL

   UNION

   SELECT DISTINCT *
   FROM MAX_EGI AS M
   WHERE NOT EXISTS
   (
      SELECT *
      FROM GEN_ITEMSET AS G
      WHERE
        (G.site = M.site) AND
        (G.day# = M.day#) AND
        (G.minsupp = m.minsupp) AND
        (G.ipsource = M.ipsource OR G.ipsource IS NULL AND M.ipsource IS NULL) AND
        (G.portsource = M.portsource OR G.portsource IS NULL AND M.portsource IS NULL) AND
        (G.c2s_packets = M.c2s_packets OR G.c2s_packets IS NULL AND M.c2s_packets IS NULL) AND
        (G.ipdest = M.ipdest OR G.ipdest IS NULL AND M.ipdest IS NULL) AND
        (G.portdest = M.portdest OR G.portdest IS NULL AND M.portdest IS NULL) AND
        (G.s2c_packets = M.s2c_packets OR G.s2c_packets IS NULL AND M.s2c_packets IS NULL) AND
        (G.protocol = M.protocol OR G.protocol IS NULL AND M.protocol IS NULL)
   )
)
```

**Listato A.7(c)** — *Modulo della funzione generatrice inversa di itemset (estratto da make_Measures.sql)*

```
CREATE FUNCTION STRINGIFY
(
   @L1 varchar(255)
  ,@L2 varchar(255)
  ,@L3 varchar(255)
  ,@L4 varchar(255)
  ,@L5 varchar(255)
  ,@L6 varchar(255)
  ,@L7 varchar(255)
  ,@S  varchar(4096)
  ,@p  smallint
)
RETURNS varchar(8000)

BEGIN
   RETURN
      '{' +
      CASE @p
         WHEN 1 THEN 'ipsource:' + @L1 + ' '
         WHEN 2 THEN 'portsource:' + @L2 + ' '
         WHEN 3 THEN 'c2s_packets:' + @L3 + ' '
         WHEN 4 THEN 'ipdest:' + @L4 + ' '
         WHEN 5 THEN 'portdest:' + @L5 + ' '
         WHEN 6 THEN 's2c_packets:' + @L6 + ' '
         WHEN 7 THEN 'protocol:' + @L7 + ' '
         ELSE ''
      END +
      RTRIM
      (
         CASE WHEN @L1 IS NULL OR @p = 1 THEN '' ELSE 'ipsource:' + @L1 + ' ' END +
         CASE WHEN @L2 IS NULL OR @p = 2 THEN '' ELSE 'portsource:' + @L2 + ' ' END +
         CASE WHEN @L3 IS NULL OR @p = 3 THEN '' ELSE 'c2s_packets:' + @L3 +'  ' END +
         CASE WHEN @L4 IS NULL OR @p = 4 THEN '' ELSE 'ipdest:' + @L4 + ' ' END +
         CASE WHEN @L5 IS NULL OR @p = 5 THEN '' ELSE 'portdest:' + @L5 + ' ' END +
         CASE WHEN @L6 IS NULL OR @p = 6 THEN '' ELSE 's2c_packets:' + @L6 + ' ' END +
         CASE WHEN @L7 IS NULL OR @p = 7 THEN '' ELSE 'protocol:' + @L7 + ' ' END
      ) +
      '}' +
      CASE WHEN @S IS NULL OR @S = '' THEN '' ELSE ' \ { '+ REPLACE(@S,' | ','} {')+' }' END
END
```