

***Grafi pesati e relazioni n-arie:
un approccio generale
all'organizzazione automatica di
dati secondo rapporti di rilevanza***

Marco Giunti

Università di Cagliari

giunti@unica.it

Il problema in discussione:

Le ipotesi – la base dati iniziali

Supponiamo che, a un tempo iniziale t_0 , sia fissata

- una base dati finita $DB(t_0) = \{f_1, f_2, \dots, f_m\}$;
 - pensiamo ciascun $f_i \in DB(t_0)$ come un *fatto* di tipo *relazionale*, ovvero, f_i è il fatto che una certa relazione n -aria R_i^n ($1 \leq n$) sussiste fra una certa n -upla di oggetti $O_{i,1}, O_{i,2}, \dots, O_{i,n}$;
 - non necessariamente la sussistenza o non sussistenza della relazione R_i^n è esprimibile da un valore binario 0 – 1, ma può invece essere espressa da un valore p (detto *peso*) compreso in tutto l'intervallo reale $[0, 1]$;
 - in altri termini, la relazione R_i^n può essere o una relazione classica (*crisp*) o una relazione *fuzzy*.

Il problema in discussione: Le ipotesi – la dinamica successiva

Supponiamo inoltre che, in tempi successivi t_1, t_2, t_3, \dots , la base dati $DB(t_j)$ si modifichi, mediante

- l'immissione di nuovi fatti e/o
- la cancellazione di vecchi fatti.

Il problema in discussione: La domanda

[Q] Come è possibile rappresentare l'informazione, sia quella iniziale $DB(t_0)$, sia quella di ogni stadio successivo $DB(t_j)$ ($0 < j$), in modo tale da organizzare automaticamente, via via che la base dati cresce o comunque si modifica, i *rapporti di rilevanza* fra i diversi dati immagazzinati?

La rappresentazione enunciativa non coglie i rapporti di rilevanza

- Siccome $DB(t_k)$ (per ogni $k \geq 0$) è un insieme di fatti relazionali, esso può sempre essere rappresentato come un insieme di *enunciati atomici* di un opportuno *linguaggio formale*.
- Ma tale rappresentazione non è sufficiente a mettere in luce i rapporti di rilevanza fra i fatti in $DB(t_k)$.
- Essa permette infatti di esplicitare soltanto i *rapporti logici* fra gli enunciati corrispondenti.
- Ma in questo caso, essendo tutti gli enunciati *atomici*, essi sono tutti *logicamente indipendenti*.

La soluzione proposta – linee generali

- L'alternativa qui proposta dà prima di tutto un *metodo standard* per trasformare univocamente una qualunque rappresentazione enunciativa formale di $DB(t_k)$ in un ben preciso *grafo diretto ed etichettato*.
- Ciascun enunciato formale (o il fatto da esso rappresentato) corrisponde quindi a un particolare *percorso* all'interno del grafo.
- I rapporti di rilevanza fra i fatti in $DB(t_k)$ sono infine automaticamente rappresentati dai *rapporti di connessione* fra tali percorsi all'interno del grafo.

Rapporto con il problema della riduzione di una relazione n -aria a relazioni binarie

- Come detto, la mia proposta per l'organizzazione di dati secondo rapporti di rilevanza si basa su un metodo standard per la rappresentazione di relazioni n -arie mediante grafi diretti ed etichettati.
- Da un punto di vista astratto, un grafo diretto ed etichettato può essere identificato con una famiglia di *relazioni binarie* su un dominio dato.
- Il metodo standard per la rappresentazione di relazioni n -arie mediante grafi diretti ed etichettati è quindi un metodo standard per ridurre una relazione n -aria a opportune relazioni binarie.

Le linee guida del W3C per la definizione di relazioni n -arie nel Semantic Web

- Lo stesso problema sorge anche nell'ambito di linguaggi Semantic Web quali RDF and OWL, in cui le relazioni si limitano a relazioni binarie.
- Un apposito gruppo di lavoro del W3C – *World Wide Web Consortium* ha recentemente proposto (2006) modelli di ontologie per la rappresentazione di relazioni n -arie in questi linguaggi.

Punti deboli dell'approccio del gruppo di lavoro del W3C

- L'approccio proposto dal W3C
 1. è indebitamente complicato;
 2. richiede l'introduzione di ontologie arbitrarie e immotivate;
 3. è sostanzialmente incapace di render conto dell'*ordine* in cui n oggetti stanno in una certa relazione.
- Al contrario, il metodo proposto qui non incorre in nessuno di questi problemi.

Insiemi di fatti espressi mediante enunciati soggetto-predicato

- Come detto, è possibile dare un *metodo standard* per trasformare univocamente una qualunque rappresentazione enunciativa formale di $DB(t_k)$ in un ben preciso *grafo diretto ed etichettato*.
- Tale trasformazione è particolarmente interessante quando i fatti in $DB(t_k)$ siano rappresentati da enunciati formali che rispettino il più possibile la forma *soggetto-predicato* degli enunciati del linguaggio comune che esprimono tali fatti.
- Tali enunciati formali sono di tre tipi, che corrispondono ai tre usi fondamentali del verbo essere nel linguaggio comune:
 1. identità – Il fratello maggiore di Carlo è Giovanni;
 2. predicazione sostanziale – Giovanni è un cantante lirico;
 3. predicazione accidentale – Sandra è sposata con Mario.

L'ontologia di un insieme di fatti espressi mediante enunciati soggetto-predicato

- Con questo tipo di rappresentazione formale, l'ontologia è composta da tre *tipi* di *oggetti* (rappresentati dai *nodi* del grafo):
 1. individui;
 2. classi;
 3. proprietà;
- e da tre *tipi* di relazioni (rappresentate nel grafo da *frecce consecutive opportunamente etichettate*):
 1. relazioni di *identità* fra oggetti;
 2. relazioni di *appartenenza* di oggetti a classi (predicazione sostanziale);
 3. relazioni di *inerenza* di proprietà a oggetti (predicazione accidentale).

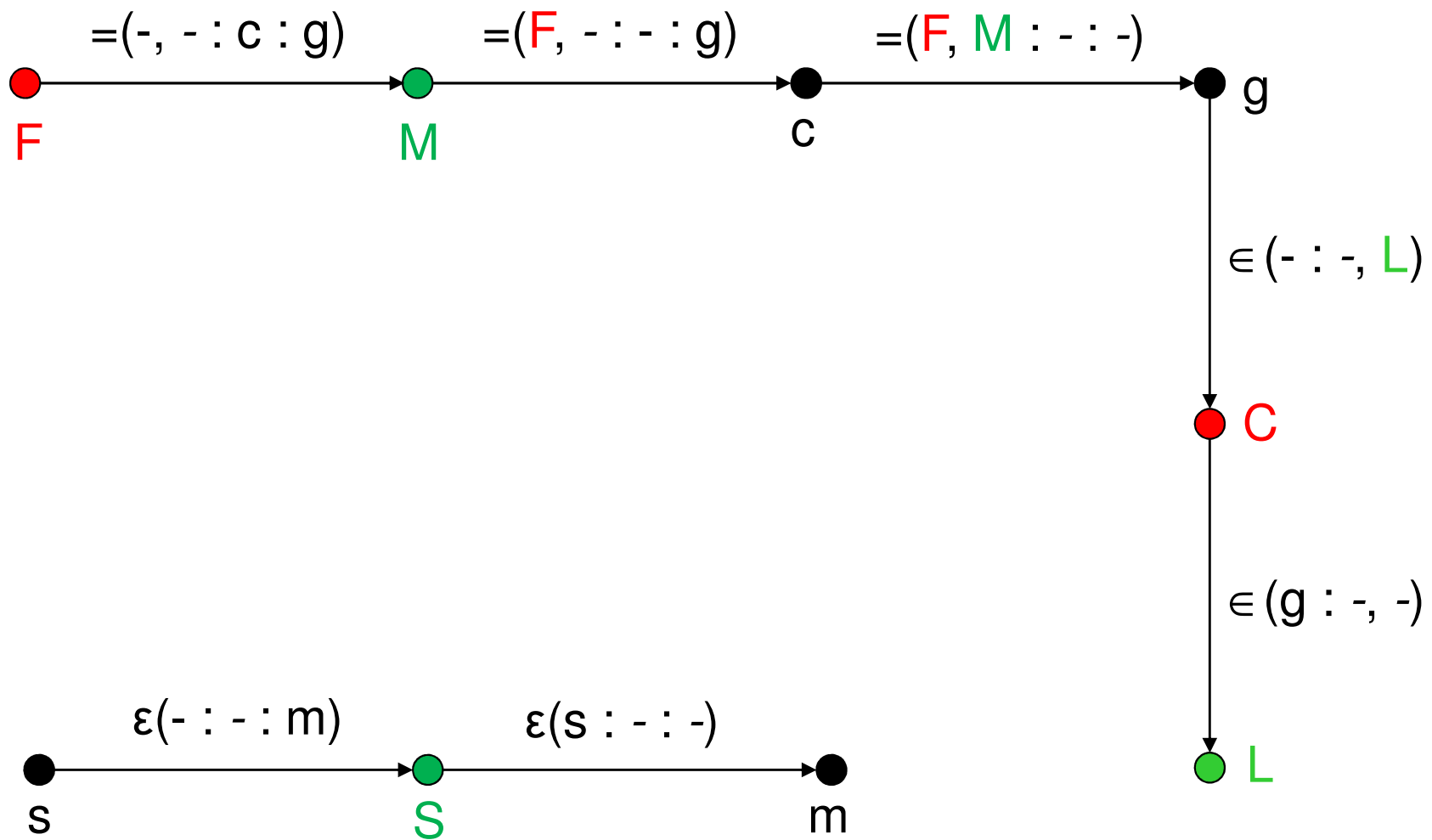
Il grafo corrispondente a un singolo fatto

1. Fatto (relazione di identità): Il fratello maggiore di Carlo è Giovanni
 - formalizzazione: $=(F, M:c:g)$
 - grafo: $F \xrightarrow{(-, -:c:g)} M \xrightarrow{(F, -: -:g)} c \xrightarrow{(F, M: -: -)} g$
2. Fatto (relazione di appartenenza): Giovanni è un cantante lirico
 - formalizzazione: $\in (g:C,L)$
 - grafo: $g \xrightarrow{(-: -, L)} C \xrightarrow{(g: -: -)} L$
3. Fatto (relazione di inerenza): Sandra è sposata con Mario
 - formalizzazione: $\varepsilon(s:S:m)$
 - grafo: $s \xrightarrow{(-: -:m)} S \xrightarrow{(s: -: -)} m$

I rapporti di rilevanza fra i fatti 1, 2 e 3 della diapositiva 12

- Intuitivamente, I fatti 1 e 2 della diapositiva precedente (dia 12) stanno in un rapporto di rilevanza reciproca, perché riguardano entrambi Giovanni; il fatto 3, al contrario, non ha alcun rapporto di rilevanza né con 1 né con 2.
- Ciò è immediatamente mostrato dai *rapporti di connessione* fra i *percorsi* corrispondenti nel grafo che rappresenta tutti e tre i fatti (si veda la diapositiva seguente – dia 14).

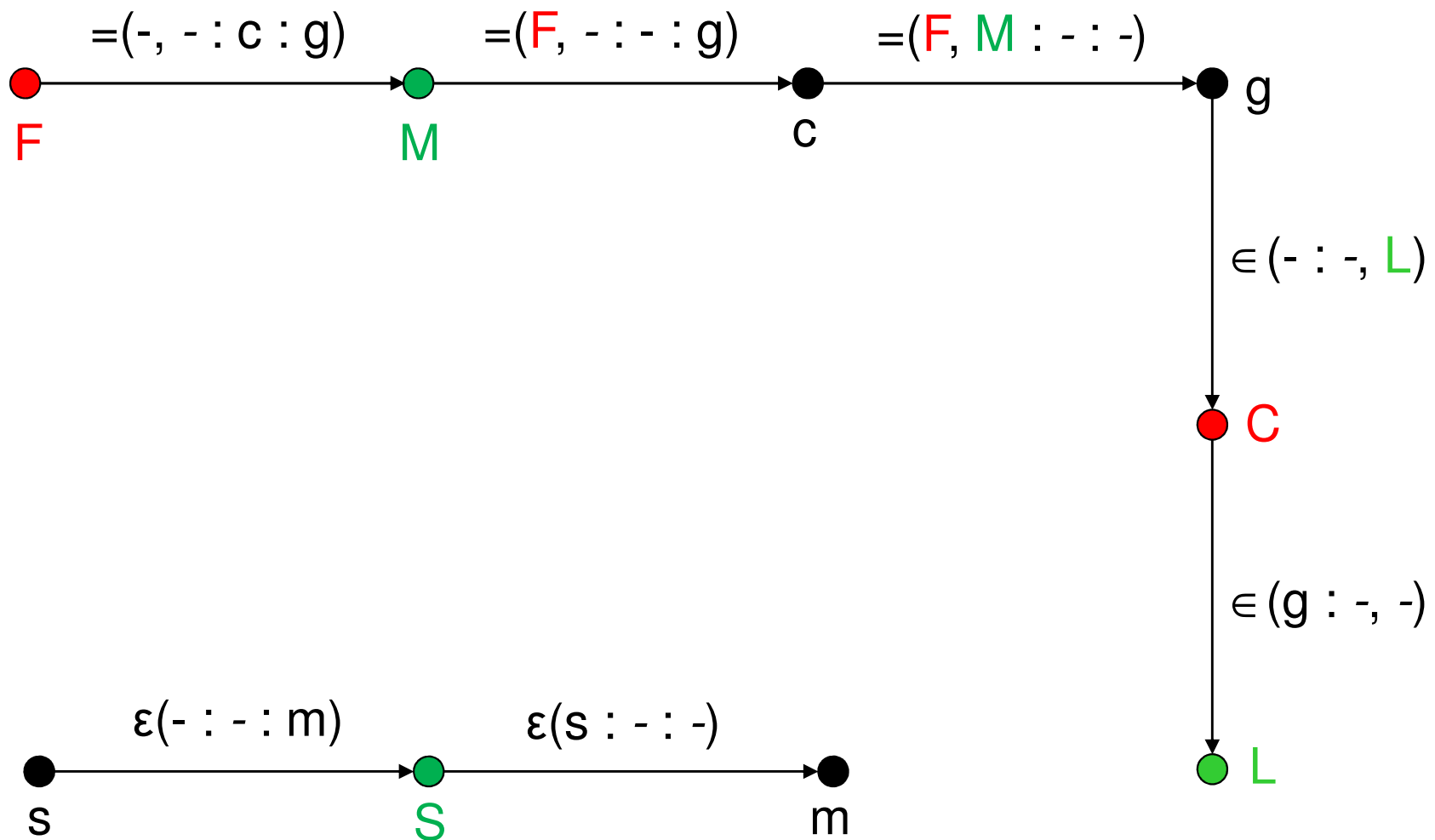
Il grafo che rappresenta i fatti 1, 2, 3 (dia 12)



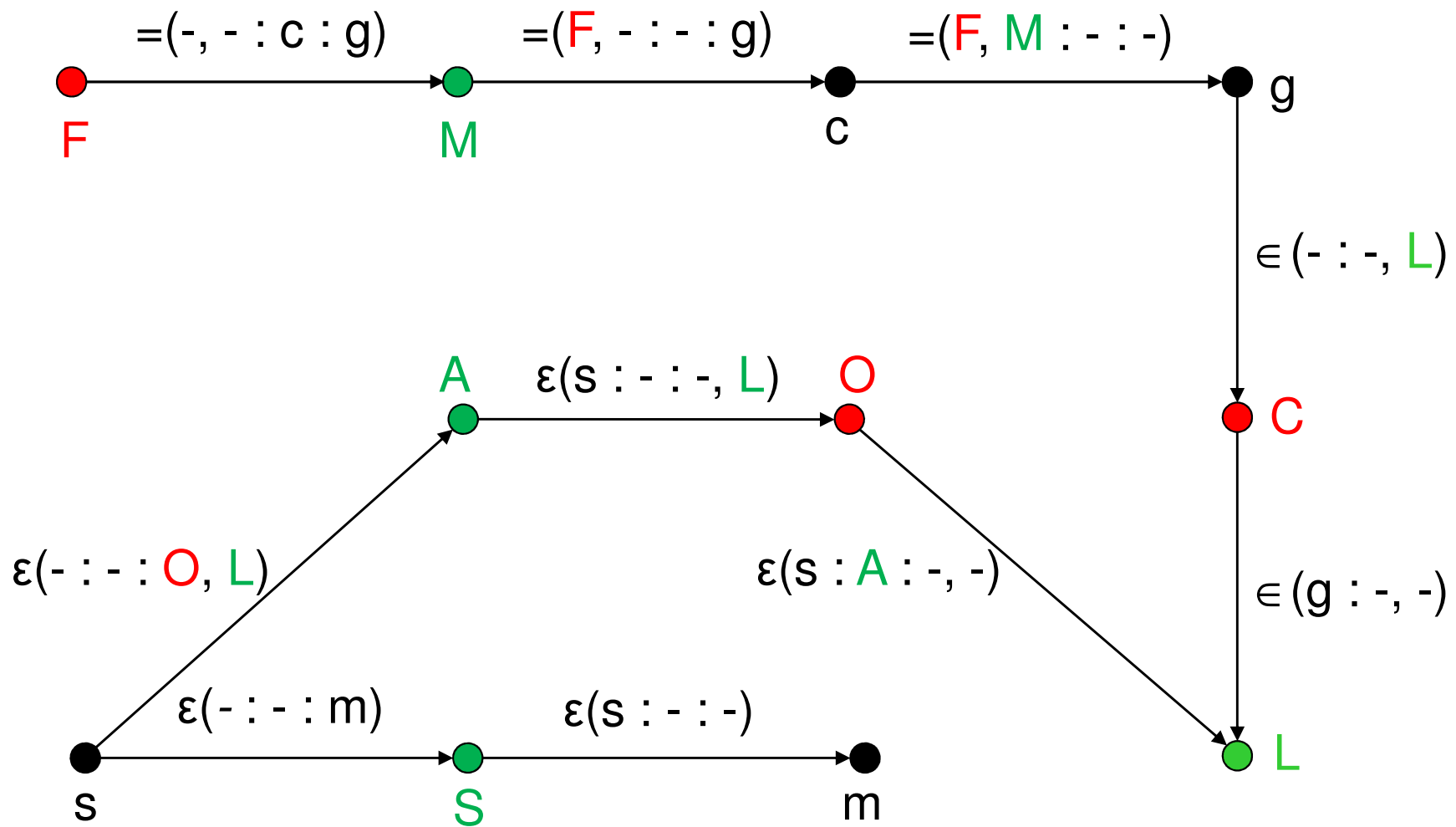
L'aggiunta di un nuovo fatto cambia i rapporti di rilevanza

- Aggiungiamo adesso
- 4. Fatto (relazione di inerenza): Sandra è appassionata di opera lirica
 - formalizzazione: $\varepsilon(s:A:O,L)$
 - grafo: $s \xrightarrow{\varepsilon(-:O,L)} A \xrightarrow{\varepsilon(s:-,L)} O \xrightarrow{\varepsilon(s:A:-)} L$
- Adesso i rapporti di rilevanza fra i fatti #1, #2 e #3 sono cambiati, perché il fatto #4 è direttamente rilevante sia per #3 che per #2 e quindi, indirettamente, #3 diventa rilevante sia per #2 che per #1.

Il grafo che rappresenta i fatti 1, 2, 3 (dia 12)



Il grafo che rappresenta i fatti 1, 2, 3 (dia 12) + 4 (dia 15)



Come quantificare i rapporti di rilevanza – (1) generalizzare la metrica standard

- La metrica standard su un grafo connesso:
 - distanza fra due nodi x e y
 - $d(x, y) :=$ lunghezza del percorso minimo fra x e y , dove la lunghezza di un percorso è il numero delle sue frecce
- Generalizzazione della metrica standard che si applica a due percorsi – distanza media fra tutti i nodi nei due percorsi:
 - distanza fra due percorsi $x_1 \dots x_m$ e $y_1 \dots y_n$
 - se $x_1 \dots x_m = y_1 \dots y_n$, $d(x_1 \dots x_m, y_1 \dots y_n) := 0$;
 - altrimenti, $d(x_1 \dots x_m, y_1 \dots y_n) :=$
$$\frac{\sum_{i=x_1 \dots x_m} \sum_{j=y_1 \dots y_n} d(x_i, y_j)}{\text{NumNod}(x_1 \dots x_m) \text{NumNod}(y_1 \dots y_n)}$$

Come quantificare i rapporti di rilevanza – (2) rilevanza come inverso della distanza

- Prendiamo come misura della rilevanza fra due percorsi l'inverso della loro distanza:
 - rilevanza fra due percorsi $x_1 \dots x_m$ e $y_1 \dots y_n$
 - $r(x_1 \dots x_m, y_1 \dots y_n) := 1 / d(x_1 \dots x_m, y_1 \dots y_n)$
- Vediamo i risultati dell'applicazione di questa misura ai percorsi che rappresentano i 4 fatti dei nostri esempi (diapositiva seguente)

Rapporti di rilevanza

fra i fatti 1, 2, 3

	1	2	3	
1	∞	0,4	0	
2	0,4	∞	0	
3	0	0	∞	

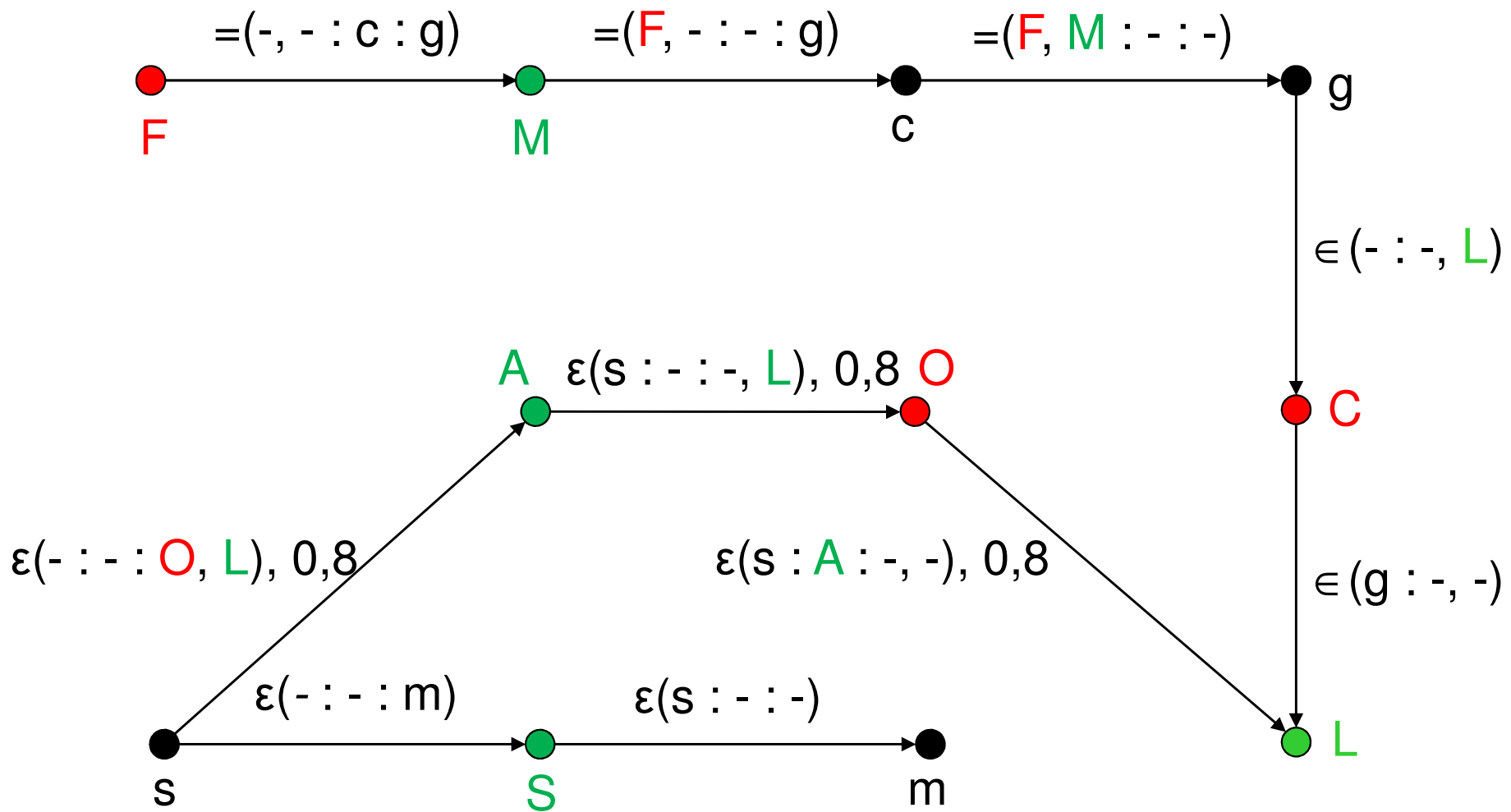
fra i fatti 1,2,3,4

	1	2	3	4
1	∞	0,4	0,13	0,25
2	0,4	∞	0,2	0,4
3	0,13	0,2	∞	0,4
4	0,25	0,4	0,4	∞

Come si trattano i dati fuzzy?

- Per fissare le idee, supponiamo adesso che il fatto #4 sia fuzzy, ovvero, assumiamo che a esso sia associato un certo peso $p \in [0, 1]$.
- Supponiamo $p = 0,8$ così che il fatto #4 diviene adesso:
 5. Fatto (relazione di inerenza): Sandra è appassionata di opera lirica, con peso 0,8
 - formalizzazione: $\varepsilon(s:A:O,L), 0,8$
 - grafo: $s \xrightarrow{\varepsilon(-:-:O,L), 0,8} A \xrightarrow{\varepsilon(s:-:-,L), 0,8} O \xrightarrow{\varepsilon(s:A:-,-), 0,8} L$

Il grafo che rappresenta i fatti 1, 2, 3 (dia 12) + 5 (dia 21)



Lunghezza di un percorso fuzzy (più bassi sono i pesi, più il percorso si allunga)

- L'introduzione dei pesi modifica il concetto di *lunghezza di un percorso $x...y$* . Essa non è più, semplicemente, il numero delle frecce nel percorso, ma piuttosto:
 - lunghezza del percorso $x...y$:= somma dell'inverso dei pesi di tutte le frecce nel percorso
 - N. B. se il peso di una freccia non è esplicitamente indicato, esso si intende uguale a 1.
- Le definizioni della metrica (sia fra nodi, che fra percorsi) e della rilevanza rimangono invariate.

L'introduzione dei pesi diminuisce la rilevanza (perché i percorsi si allungano)

**rapporti di rilevanza
fra i fatti 1,2,3,4**

	1	2	3	4
1	∞	0,4	0,13	0,25
2	0,4	∞	0,2	0,4
3	0,13	0,2	∞	0,4
4	0,25	0,4	0,4	∞

**rapporti di rilevanza
fra i fatti 1,2,3,5**

	1	2	3	5
1	∞	0,4	0,12	0,19
2	0,4	∞	0,17	0,35
3	0,12	0,17	∞	0,35
5	0,19	0,35	0,35	∞

E' tutto.

Grazie