Università degli studi di Modena e Reggio Emilia
Laurea triennale in Ingegneria Informatica

Entity Resolution progressiva con graph embedding

Relatore:
Prof.ssa Sonia Bergamaschi

Correlatore:
Luca Gagliardelli

Canditato:
Michele Rinaldi

Entity Resolution

Task che ha come obiettivo l'identificazione di elementi che rappresentano la stessa entità del mondo reale.

ID	Nome	Cognome	Nazionale
a	Luigi	Datome	Italia
b	**Lebron**	**James**	**USA**

ID	Name	Team	Nationality
c	Luigi Datome	Milano	Italy
d	**Lebron James**	Lakers	-
e	**James** Harden	76ers	**USA**

PROBLEMI DELL' ER

1. Grandi dataset: impossibile comparare ogni coppia di profili
   - Complessità O(n^2)

2. Datasets con schemi differenti
   - Allineamento prima di fare ER (time consuming)

SOLUZIONE: TOKEN BLOCKING

1. Ogni profilo è suddiviso in **tokens**
2. Per ogni token è prodotto un **blocco**
3. Blocco con profili che hanno il **token in comune**

PROFILI

Nome: Luigi Cognome:Datome squadra:Olimpia Milano Data nascita:Novembre 1987 Naz: Ita **P1**	Nominativo: Troy D. squadra: Olimpia Milano anno: 1991 **P4**
name surname: Troy Daniels team: Milano birth year:1991 **P2**	Player: D. **P5** Team: Milano Birth year: 1987 Nat:Ita
name surname: Luigi Datome team: Milano birth year:1987 **P3**	Giocatore: LD **P6** Squadra: Olimpia Milano Nascita:Novembre 1987

TOKENS

Luigi P1, P3	Datome P1, P3
Olimpia P1, P4, P6	Milano P1, P2, P3, P4, P5, P6
1987 P1, P3, P5, P6	Troy P2, P4
Daniels P4	1991 P2, P4
Novembre P1, P6	Ita P1, P5

VANTAGGI E SVANTAGGI DEL TOKEN BLOCKING

PRO

1. *I token non dipendono dallo schema*
2. *coppie profili senza blocchi in comune vengono scartate*

CONS

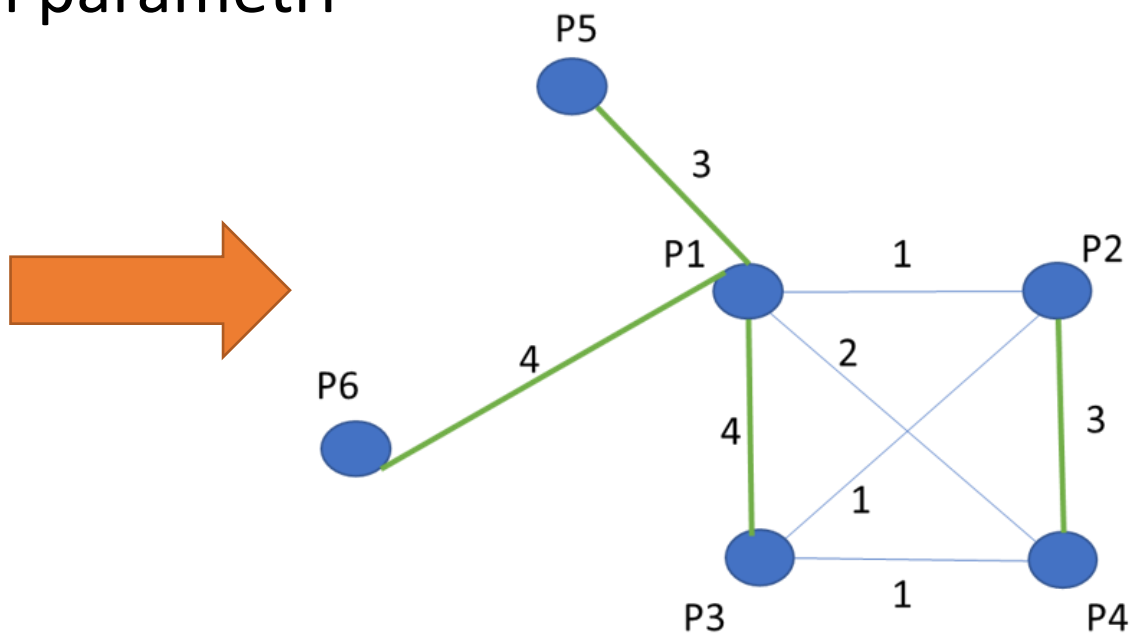
1. *Troppi blocchi (uno per token)*

SOLUZIONE: BLOCKING GRAPH[1]

Grafo in cui:

- Ogni nodo rappresenta un profilo
- Ogni arco ha un peso in funzione del numero di blocchi comuni dei profili che collega e di altri parametri

TOKENS	
Luigi P1, P3	Datome P1, P3
Olimpia P1, P4, P6	Milano P1, P2, P3, P4, P5, P6
1987 P1, P3, P5, P6	Troy P2, P4
Daniels P4	1991 P2, P4
Novembre P1, P6	Ita P1, P5



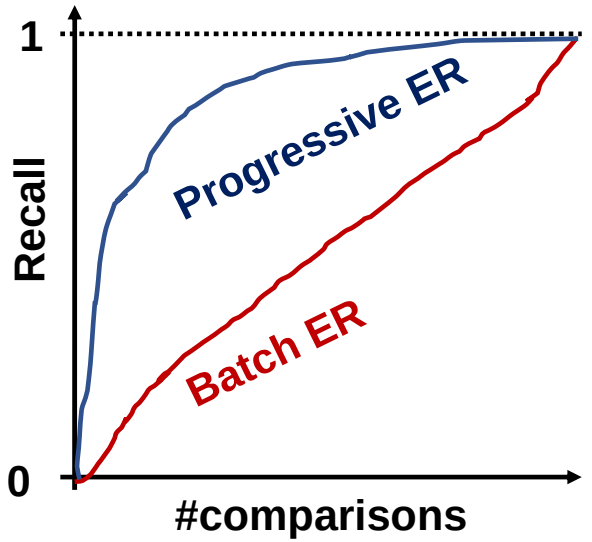
(1) Papadakis, G., Koutrika, G., Palpanas, T., & Nejdl, W. (2013). Meta-blocking: Taking entity resolutionto the next level. *IEEE Transactions on Knowledge and Data Engineering*, *26*(8), 1946-1960.

PROGRESSIVE ER

- Se il tempo è limitato (es: i dati diventano obsoleti dopo)
  - Es: Dati dei mercati azionari o sicurezza in caso di emergenza
  - Valore(ER non completa, ma veloce) > Valore(ER completa, ma lenta)

- Progressive ER cerca di massimizzare il numero di duplicati identificati in funzione del tempo

- Soluzione stato dell'arte: PPS^2(Progressive Profile Scheduling)
  - Basato su Blocking Graph



(2) Schema-agnostic Progressive Entity Resolution, G Simonini, G Papadakis, T Palpanas, S Bergamaschi IEEE ICDE2018

OBIETTIVO DELLA TESI

- Sfruttare tecniche di graph embedding per:
  - riordinare il Blocking Graph
  - trovare similarità latenti tra i nodi

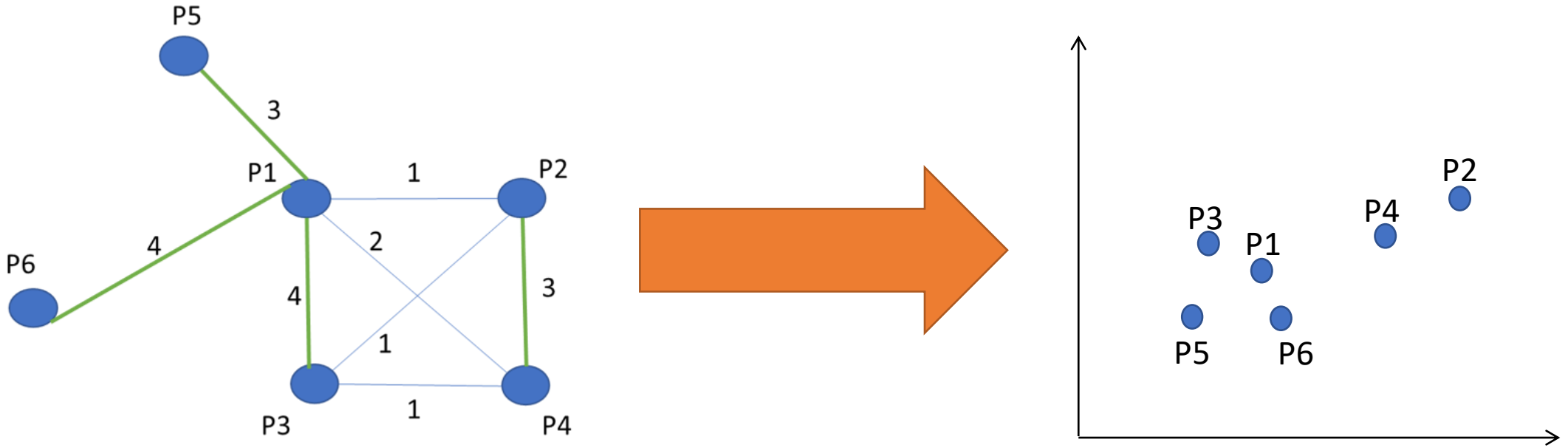
Non accade
nel PPS

GRAPH EMBEDDING

- Trasformare i nodi di un grafo in un vettore d-dimensionale
- Vantaggi:
    - Similarità latenti ‘catturate’ meglio
    - Costo comparare due profili > costo comparare due vettori

ID	Nome	Cognome	Squadra
p_i	Luigi	Datome	Olimpia Milano

(12, 3.2, ……., 0.2, 8)


P5
3
P1
1
P2
P6
4
2
4
3
1
P3
1
P4
P2
P3
P4
P1
P5
P6

Tecniche utilizzate

-HARP[3](DeepWalk[4]): Cammini casuali

-HARP(Node2Vec[5]): cammini semi-casuali

- Cammino: lista ordinata di nodi.
- Generano un certo numero di cammini

Generati gli embedding

(3) Chen, H., Perozzi, B., Hu, Y., & Skiena, S. (2018, April). Harp: Hierarchical representation learning for networks. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

(4) Perozzi, B., Al-Rfou, R., & Skiena, S. (2014, August). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 701-710).

(5) Chen, H., Perozzi, B., Hu, Y., & Skiena, S. (2018, April). Harp: Hierarchical representation learning for networks. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

ESPERIMENTI

Datasets

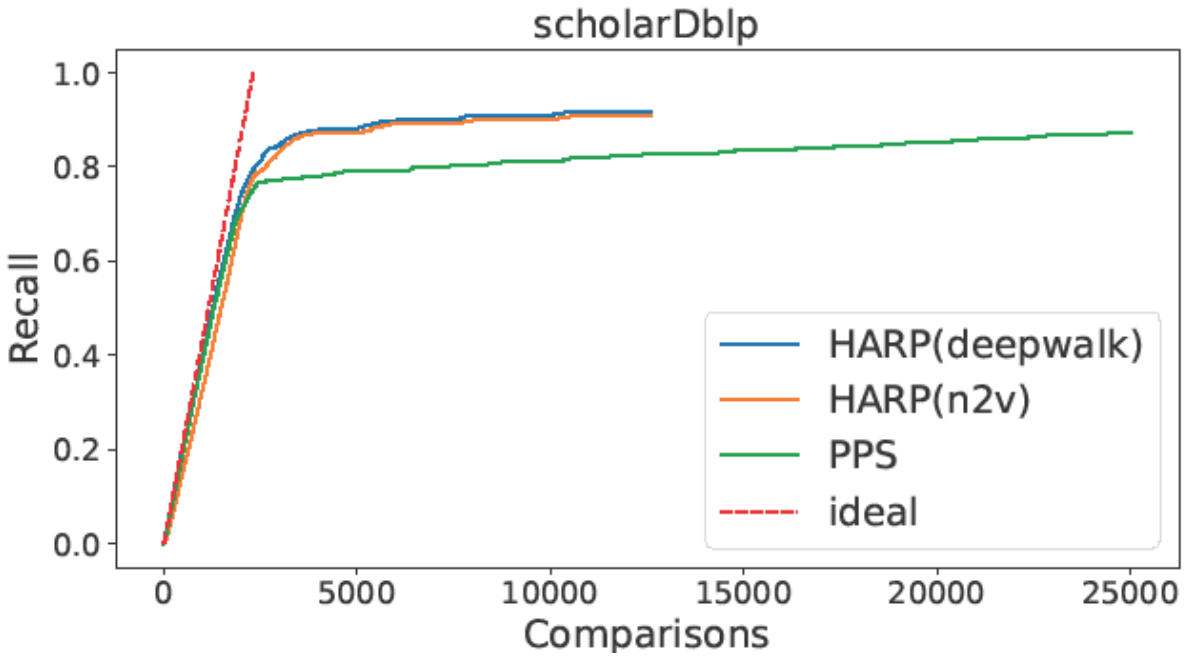
ScholarDblp

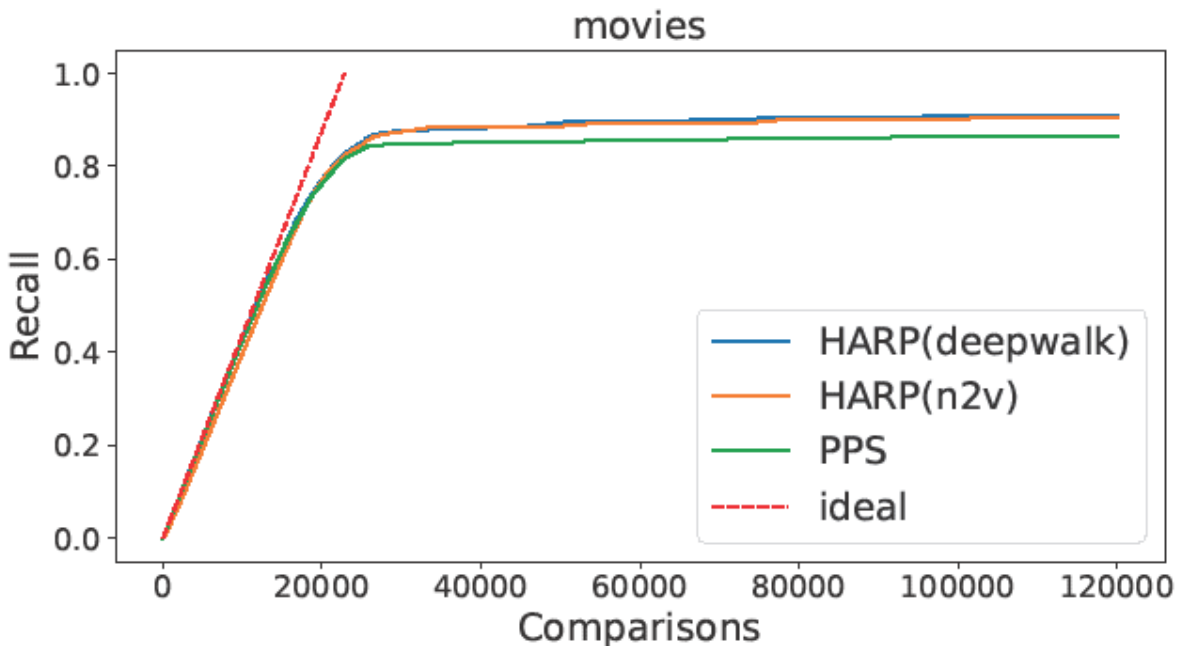
- Records Dataset1: 2516
- Records Dataset2: 61553
- Duplicati totali: 2308

Movies

- Records Dataset1: 27615
- Records Dataset2: 23182
- Duplicati totali: 22863

RESULTS





$$\boldsymbol{recall} = \frac{\boldsymbol{duplicati\ rilevati}}{\boldsymbol{duplicati\ totali}}$$

Lavoro futuro

- Articolo sottomesso a *Italian Symposium on Advanced Database Systems* 2022

- Integrare la fase di blocking ed embedding in un framework unificato

Grazie per l’attenzione!