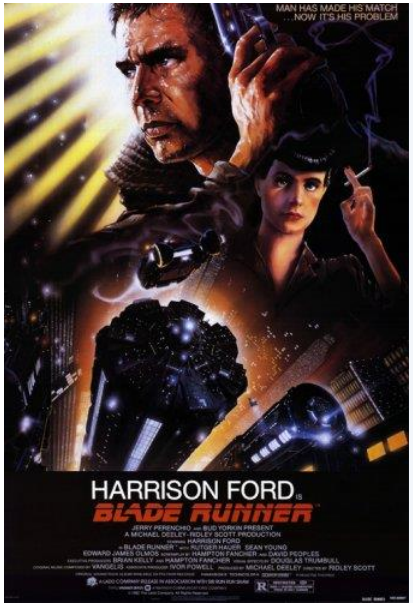
WiSP:
Weighted Shortest Paths for RDF graphs

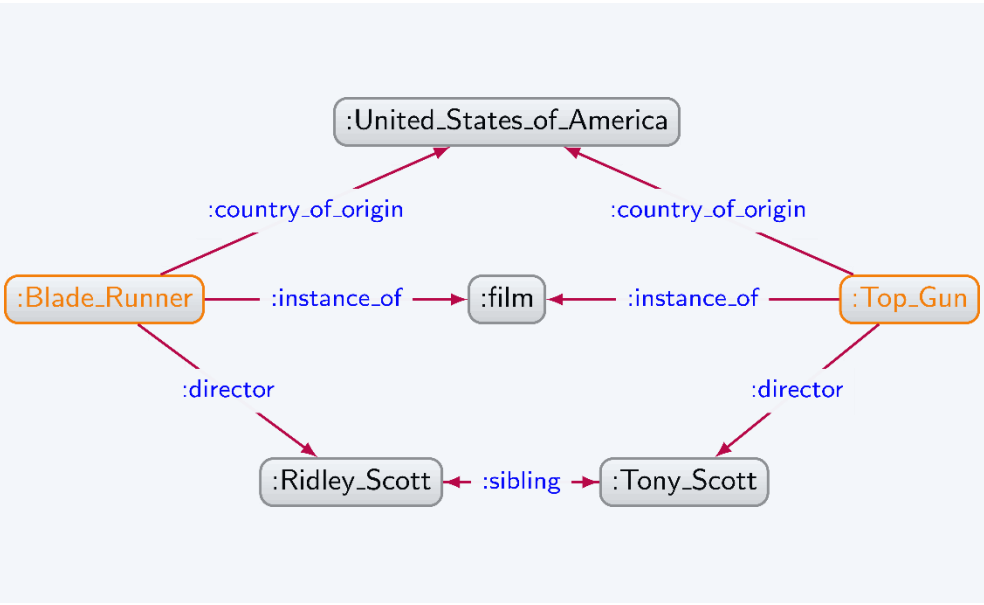
Gonzalo Tartari, Aidan Hogan

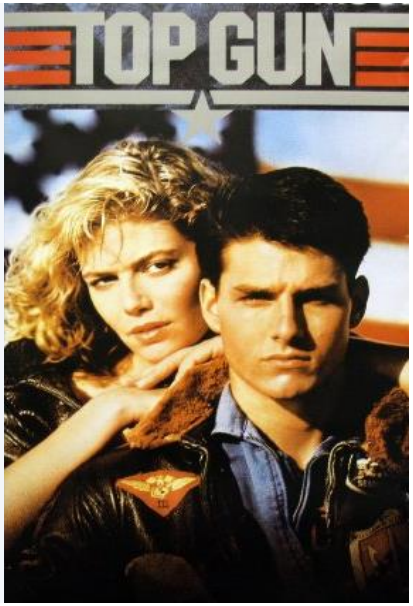
DCC, Universidad de Chile

Fundamentos de los datos

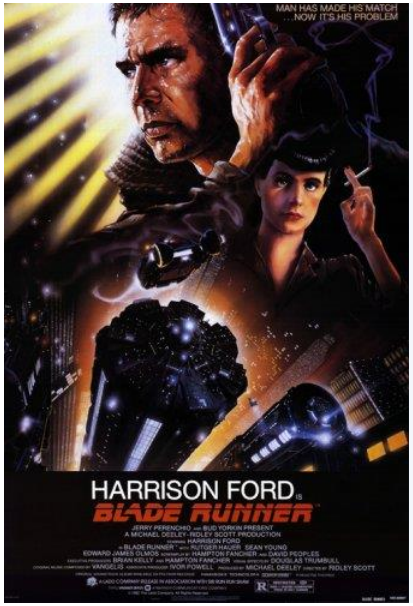
"Interesting" Paths = Shortest Paths?

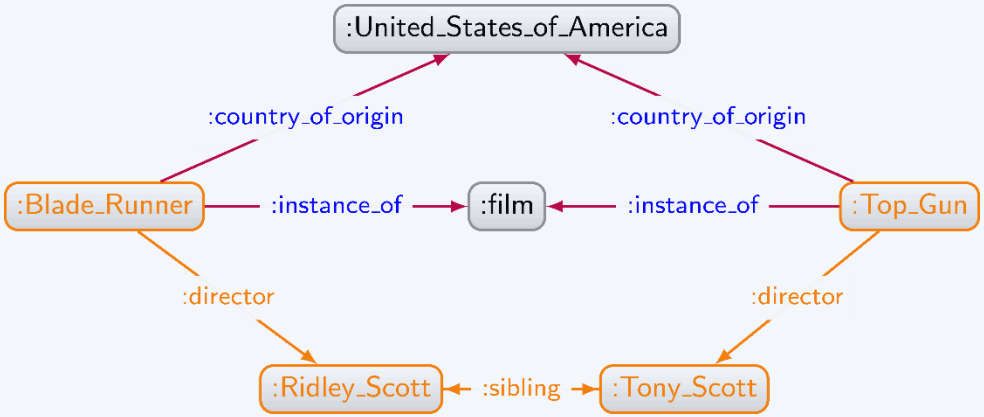


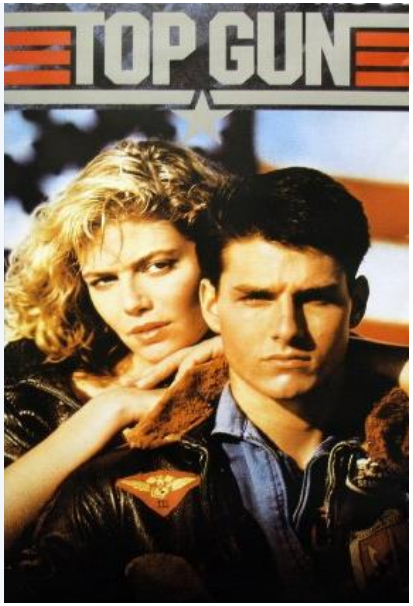




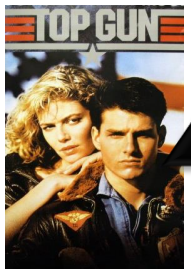
"Interesting" Paths ≠ Shortest Paths!





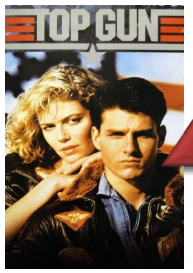


(Many of the) Existing Approaches



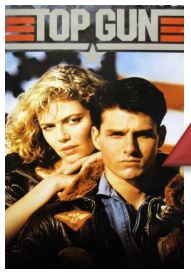
Enumerate Paths → Score Paths → Order/Filter Paths → Output

(Many of the) Existing Approaches



- $n!$ simple paths (but only a few of interest)

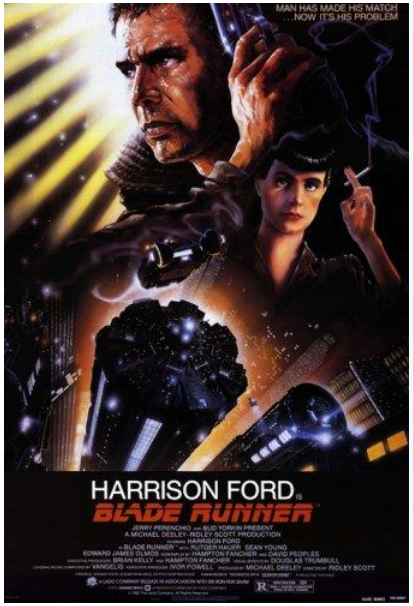
(Many of the) Existing Approaches

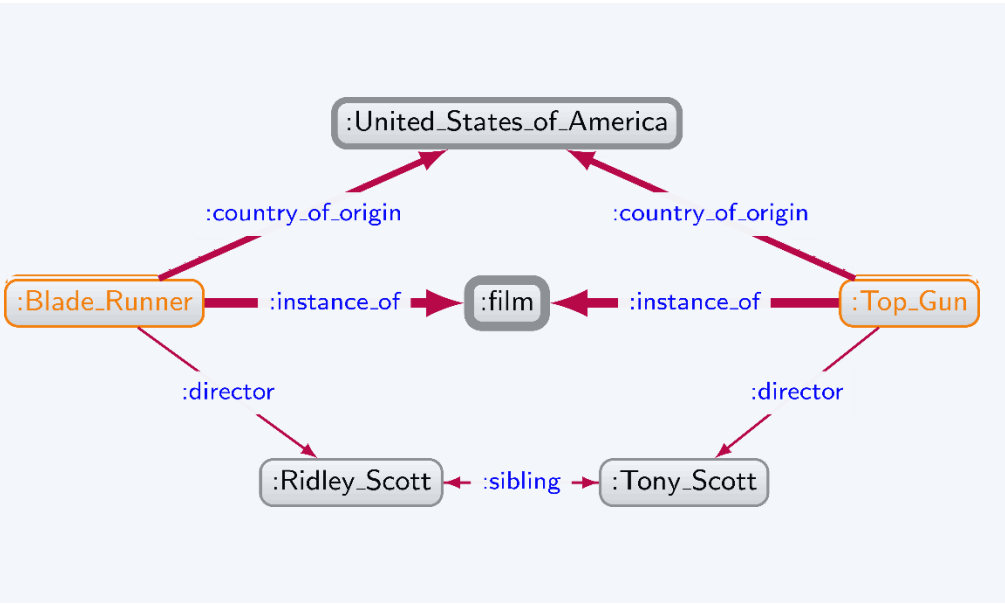


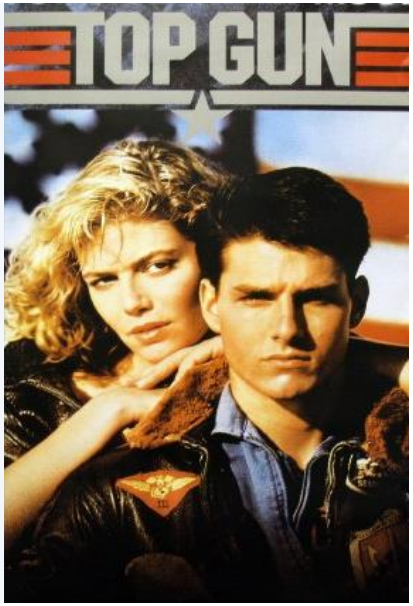
- $n!$ simple paths (but only a few of interest)
  - Paths thus often bounded in length

Our Approach: Weight Graphs

- *Idea:* "Common" nodes/edges get weighted higher



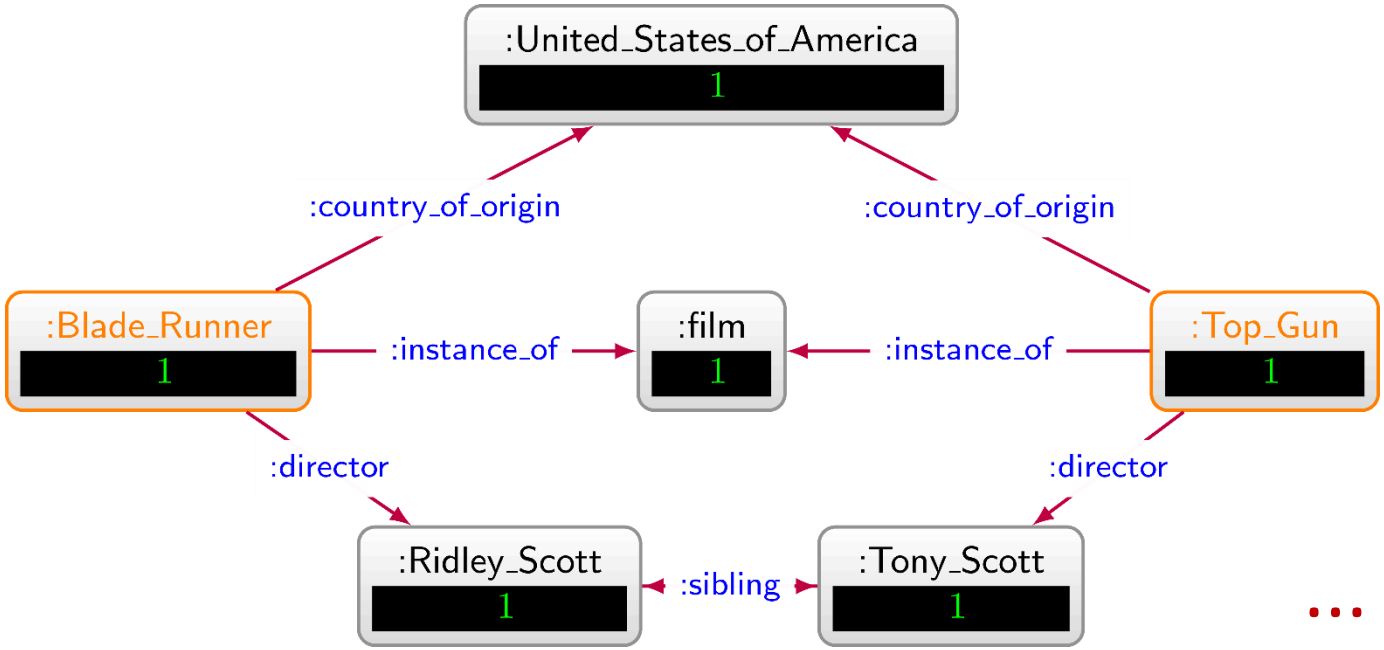




- *Benefit:* Can find "interesting" paths directly

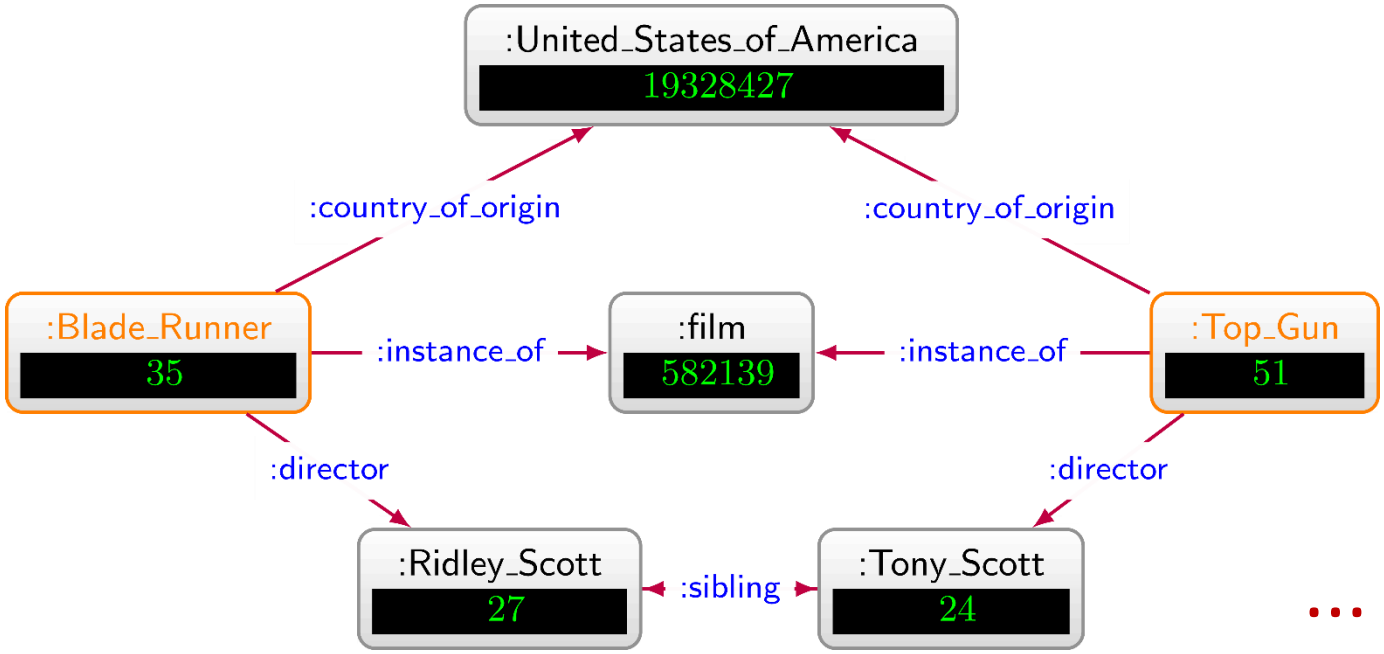
Weighting graphs: Nodes

Node Weights: Length (Baseline)



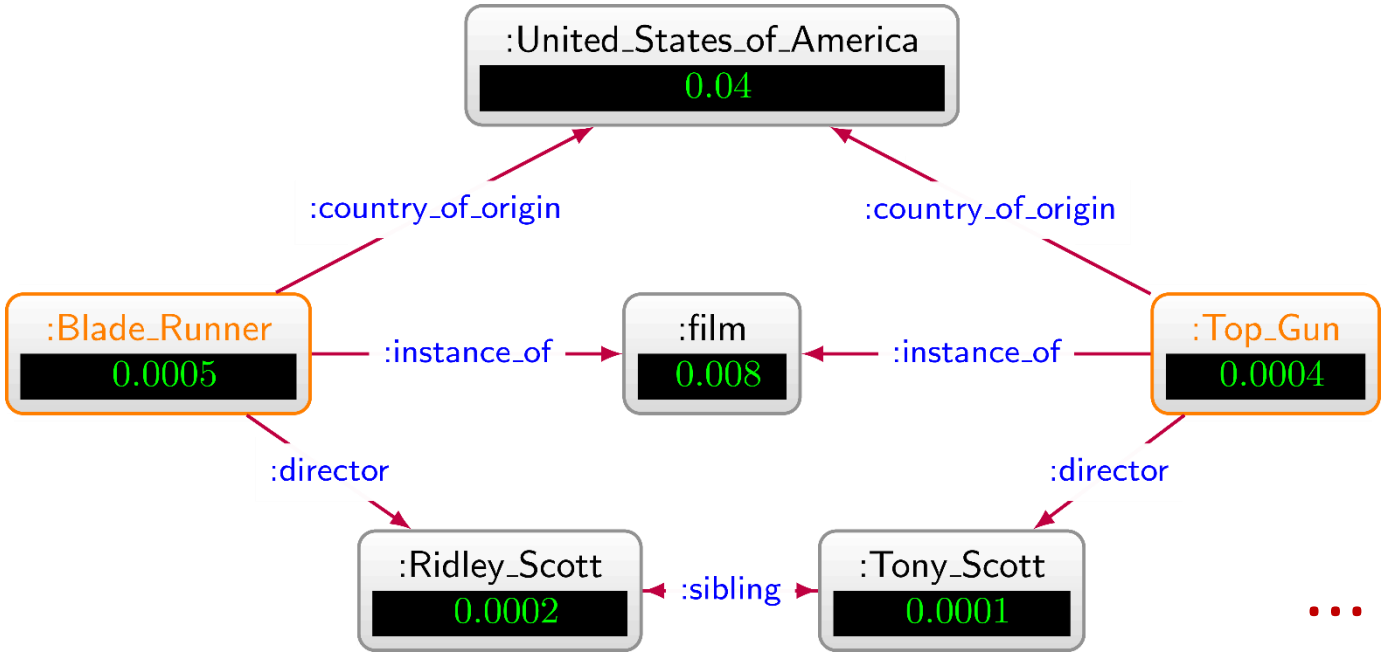
path					score
:country	:U.S.	:country$^-$			1
:instance	:film	:instance$^-$			1
:director	:RScott	:sibling	:TScott	:director$^-$	2
:director	:RScott	:sibling$^-$	:TScott	:director$^-$	2

Node Weights: Degree



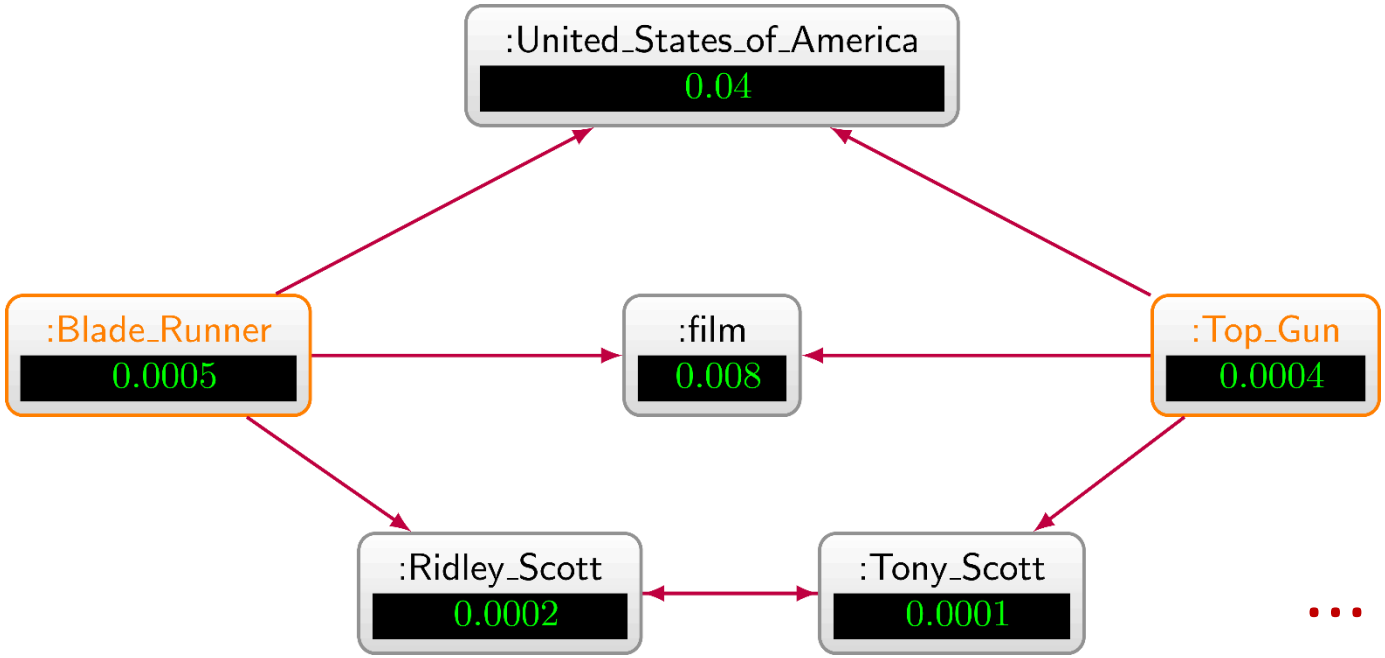
path					score
:director	:RScott	:sibling	:TScott	:director$^-$	51
:director	:RScott	:sibling$^-$	:TScott	:director$^-$	51
:instance	:film	:instance$^-$			582139
:country	:U.S.	:country$^-$			19328427

Node Weights: PageRank



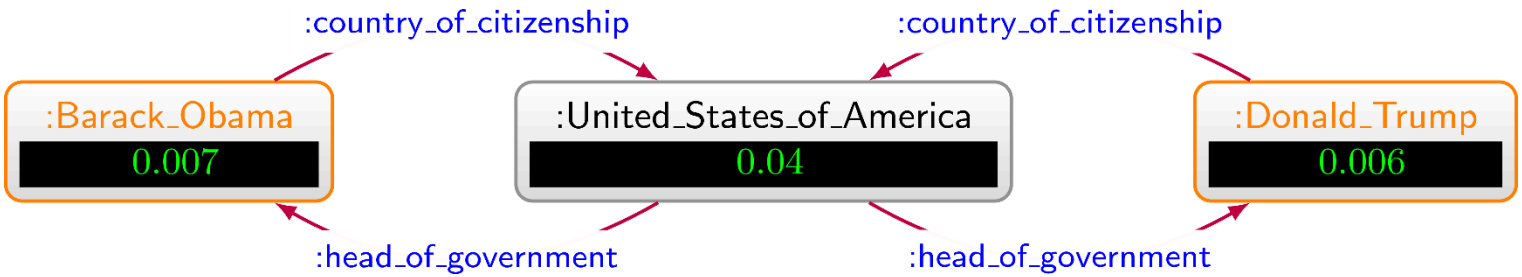
path					score
:director	:RScott	:sibling	:TScott	:director$^-$	0.0003
:director	:RScott	:sibling$^-$	:TScott	:director$^-$	0.0003
:instance	:film	:instance$^-$			0.008
:country	:U.S.	:country$^-$			0.04

Aside: PageRank / directed graph used



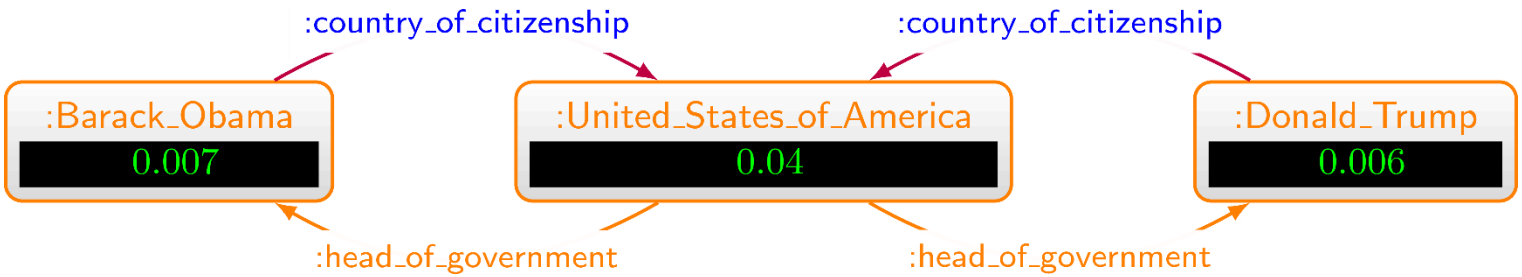
Weighting graphs: Edges

Weighting with only nodes



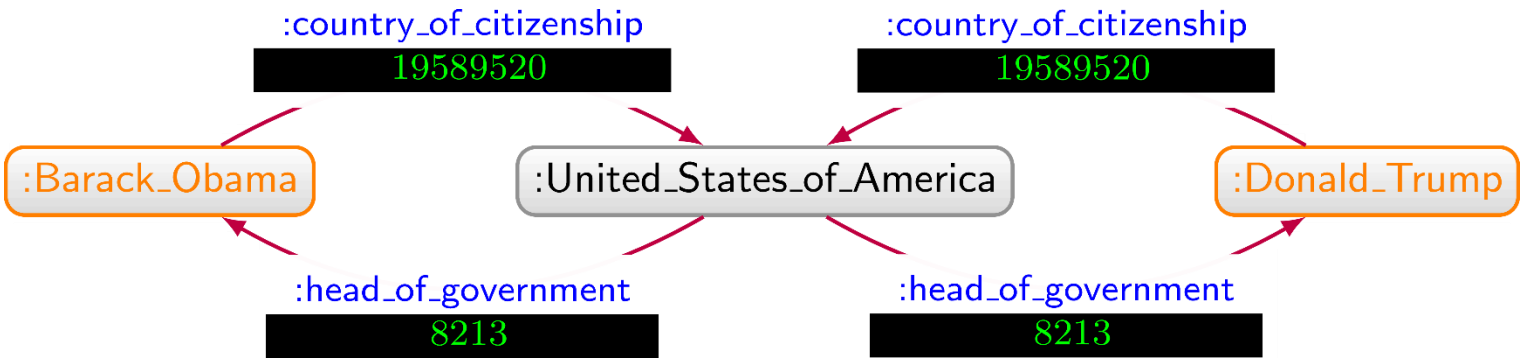
path			score
:country	:U.S.	:country$^-$	0.04
:country	:U.S.	:head$^-$	0.04
:head	:U.S.	:country$^-$	0.04
:head	:U.S.	:head$^-$	0.04

Weighting with only nodes



path			score
:country	:U.S.	:country$^-$	0.04
:country	:U.S.	:head$^-$	0.04
:head	:U.S.	:country$^-$	0.04
:head	:U.S.	:head$^-$	0.04

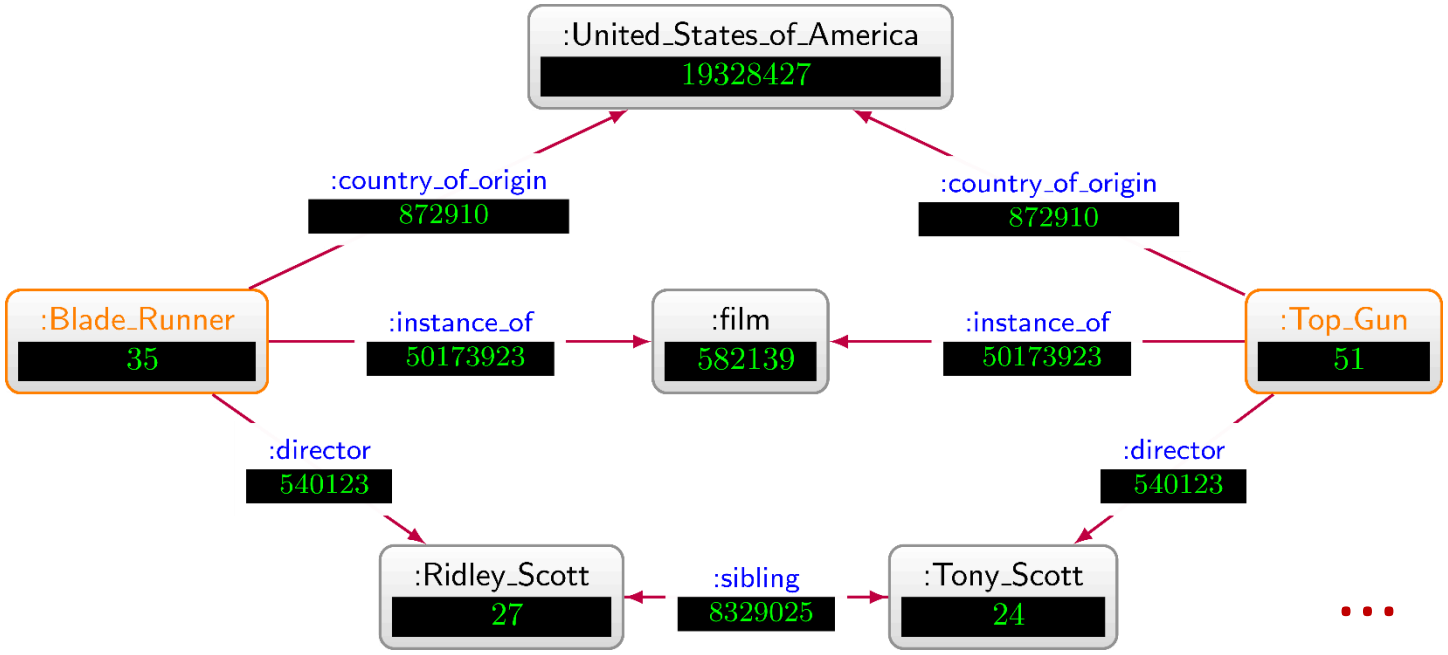
Edge Weights: Frequency



path			score
:head	:U.S.	:head$^-$	16426
:country	:U.S.	:head$^-$	19597733
:head	:U.S.	:country$^-$	19597733
:country	:U.S.	:country$^-$	39179040

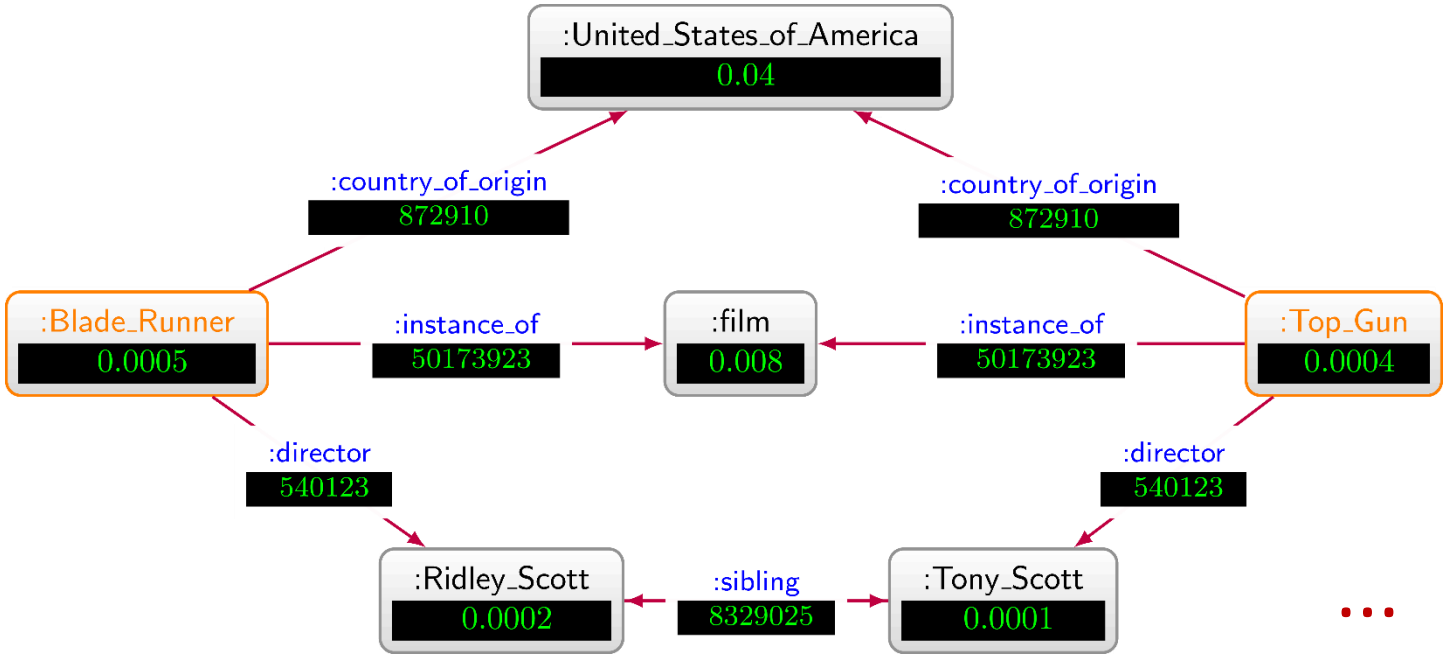
Weighting graphs: Nodes + Edges

Node + Edge Weights: Degree + Frequency



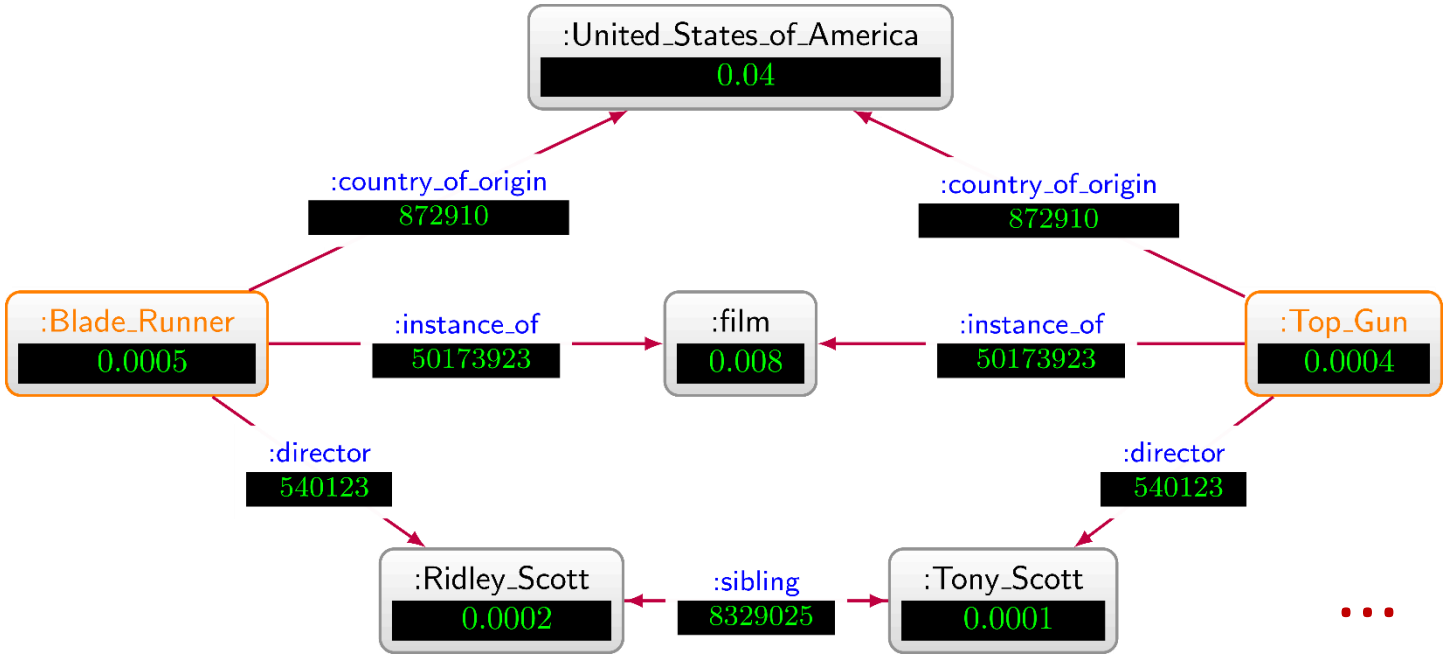
path					score
:director	:RScott	:sibling	:TScott	:director$^-$	9409322
:director	:RScott	:sibling$^-$	:TScott	:director$^-$	9409322
:country	:U.S.	:country$^-$			21074247
:instance	:film	:instance$^-$			100929984

Node + Edge Weights: PageRank + Frequency



path					score
:country	:U.S.	:country$^-$			1745820.04
:director	:RScott	:sibling	:TScott	:director$^-$	9409271.0003
:director	:RScott	:sibling$^-$	:TScott	:director$^-$	9409271.0003
:instance	:film	:instance$^-$			100347846.008

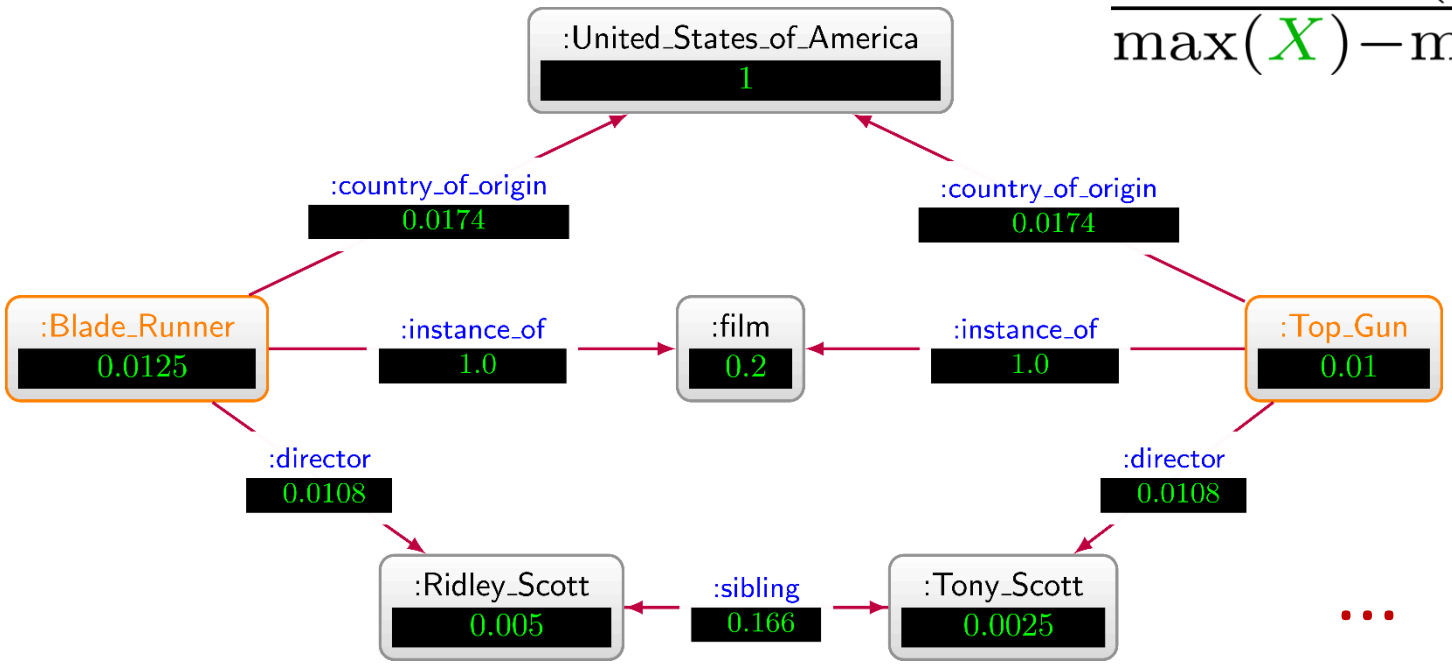
Node + Edge Weights: PageRank + Frequency



path					score
:country	:U.S.	:country$^-$			1745820.04
:director	:RScott	:sibling	:TScott	:director$^-$	9409271.0003
:director	:RScott	:sibling$^-$	:TScott	:director$^-$	9409271.0003
:instance	:film	:instance$^-$			100347846.008

Node + Edge Weights: [0,1] Normalisation

$$\frac{x - \min(X)}{\max(X) - \min(X)}$$

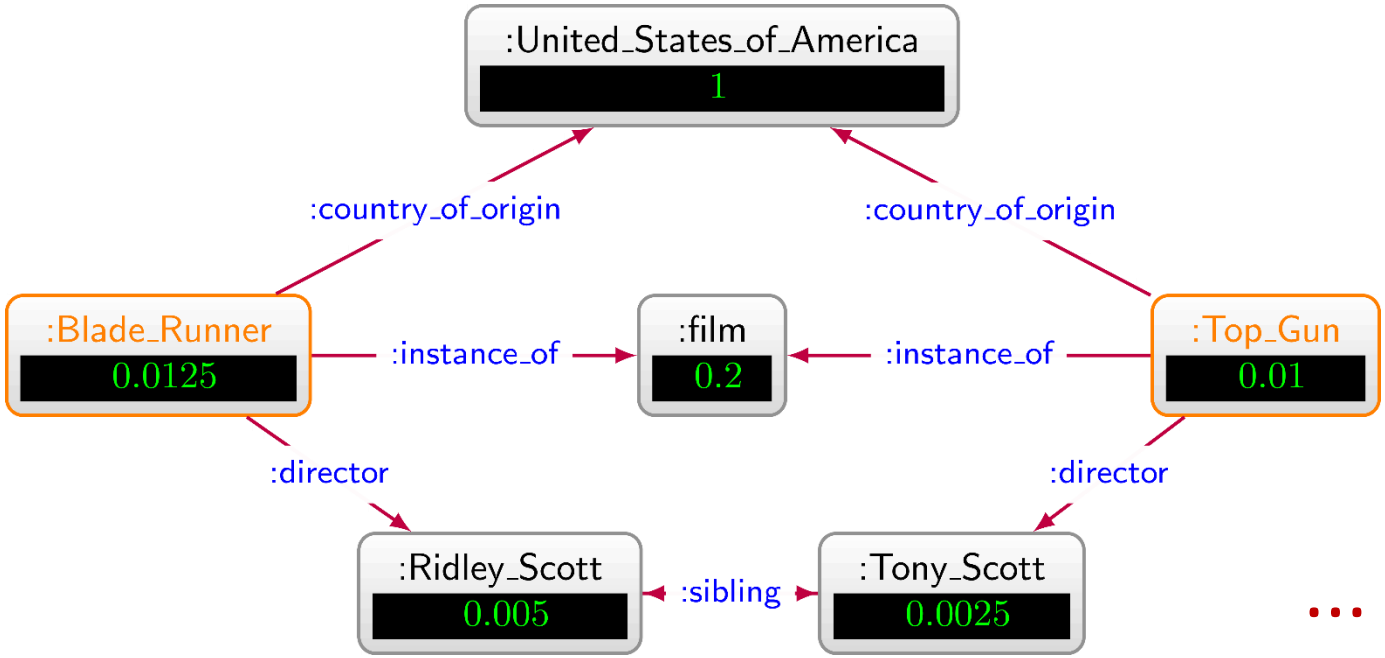


path					score
:director	:RScott	:sibling	:TScott	:director$^-$	0.195033
:director	:RScott	:sibling$^-$	:TScott	:director$^-$	0.195033
:country	:U.S.	:country$^-$			1.02153
:instance	:film	:instance$^-$			2.2

Hybrid Node Weights

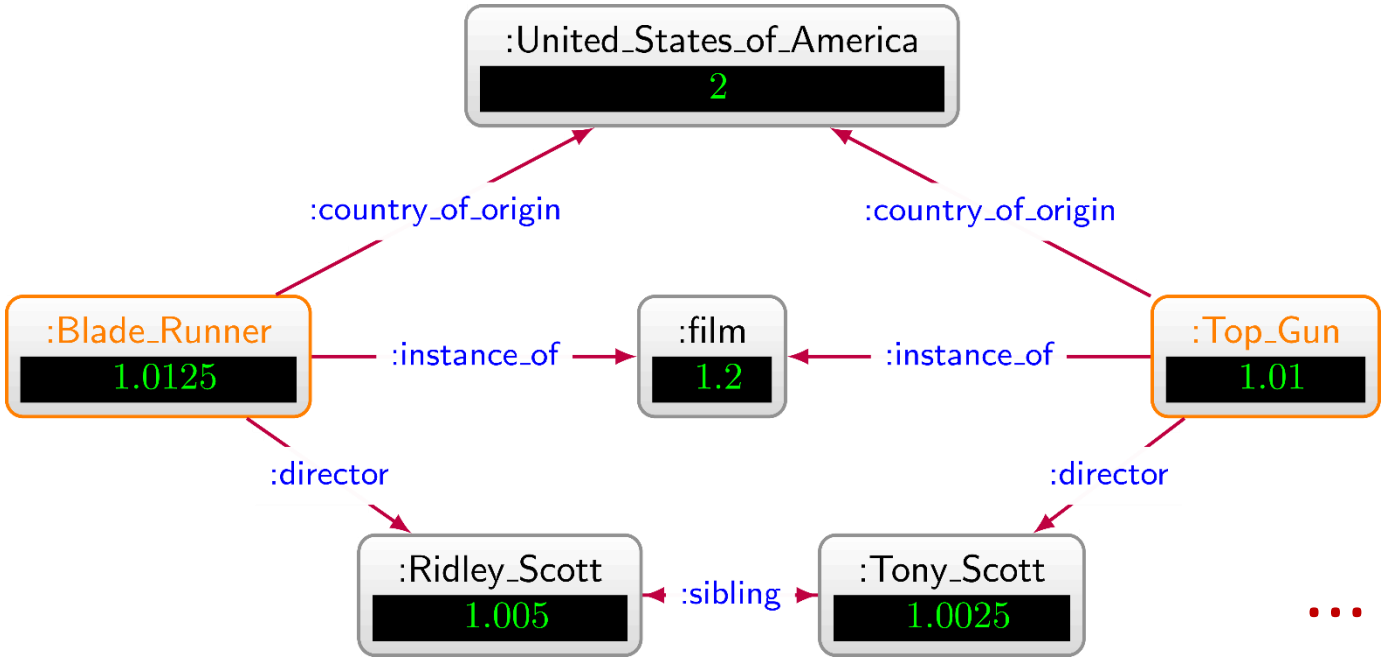
Node Weights: PageRank

Visiting one high-centrality node = Visiting thousands of low-centrality nodes



path					score
:director	:RScott	:sibling	:TScott	:director$^-$	0.0075
:director	:RScott	:sibling$^-$	:TScott	:director$^-$	0.0075
:instance	:film	:instance$^-$			0.02
:country	:U.S.	:country$^-$			1

Hybrid Node Weights: PageRank + Length



path					score
:instance	:film	:instance$^-$			1.2
:country	:U.S.	:country$^-$			2
:director	:RScott	:sibling	:TScott	:director$^-$	2.0075
:director	:RScott	:sibling$^-$	:TScott	:director$^-$	2.0075

Implementation

Weighted Shortest-Path Implementation

- Dijsktra's algorithm:
  - Worst case: $O(|E| + |V| \cdot \log|V|)$

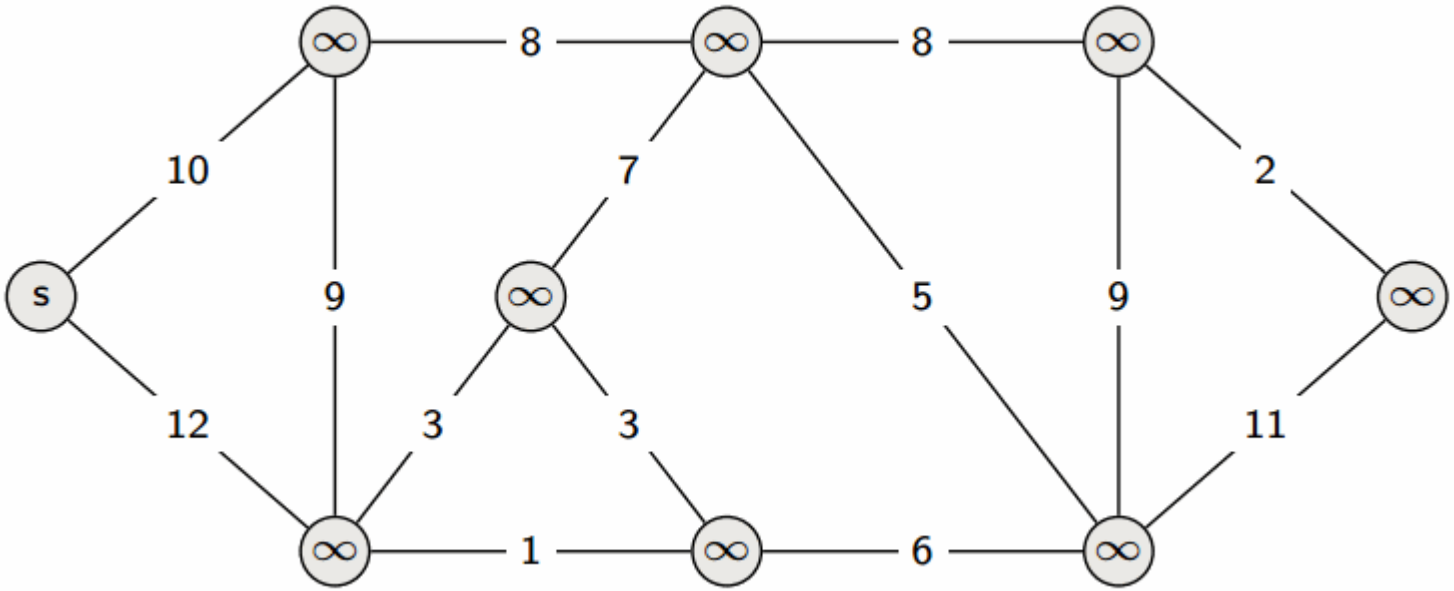


Image source: https://github.com/aakash1104/Graph-Algorithms

Experiments

Questions

- Performance:
  - How are the runtimes?
  - How is the scalability?

- Weighting schemes:
  - How similar are paths for different weightings?
  - Does weighting help find interesting paths?
  - Which weighting finds the most interesting paths?

Dataset: Wikidata

- Truthy dump: 2017-06-07
  - 25 million nodes (Q-IRIs only)
  - 90 million edges

Dataset: Wikidata Slices

Dataset	1.6 M	3.2 M	6.4 M	12.8 M	FULL
Nodes	1,227,382	2,507,582	5,303,322	10,343,129	25,081,334
Edges	6,603,412	12,160,436	22,008,446	36,404,534	89,878,092

Machine

- 2 x Intel Xeon Quad Core @1.9GHz
- 32 GB of RAM

Weighting Schemes

- Node
  - Degree (D)
  - PageRank (P)
  - Length (L)
- Node + Edge
  - Degree + Edge Frequency (DE)
  - PageRank + Edge Frequency (PE)
- Hybrid Node + Edge
  - Degree + Length + Edge Frequency (DEL)
  - PageRank + Length + Edge Frequency (PEL)

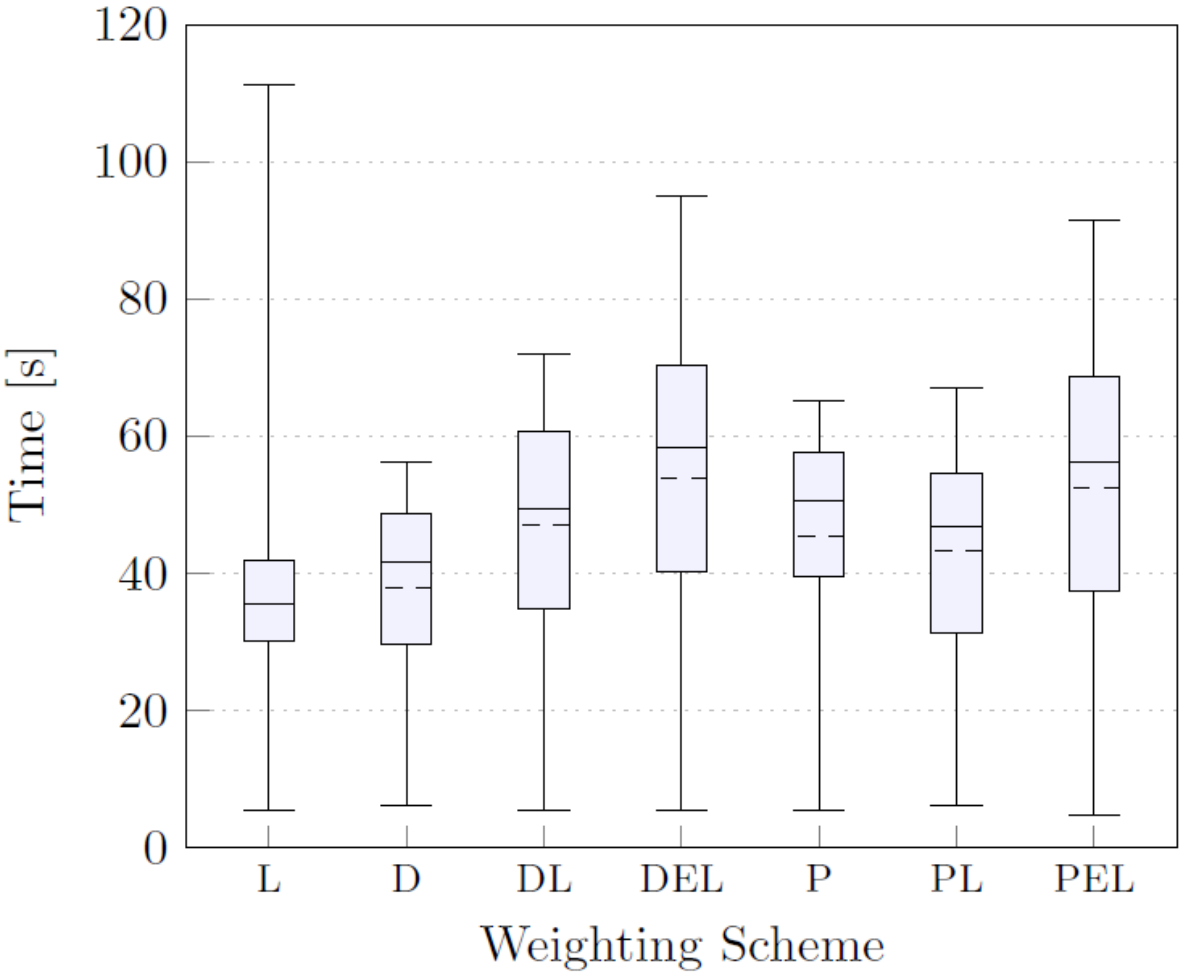
Performance

Queries (Node pairs)

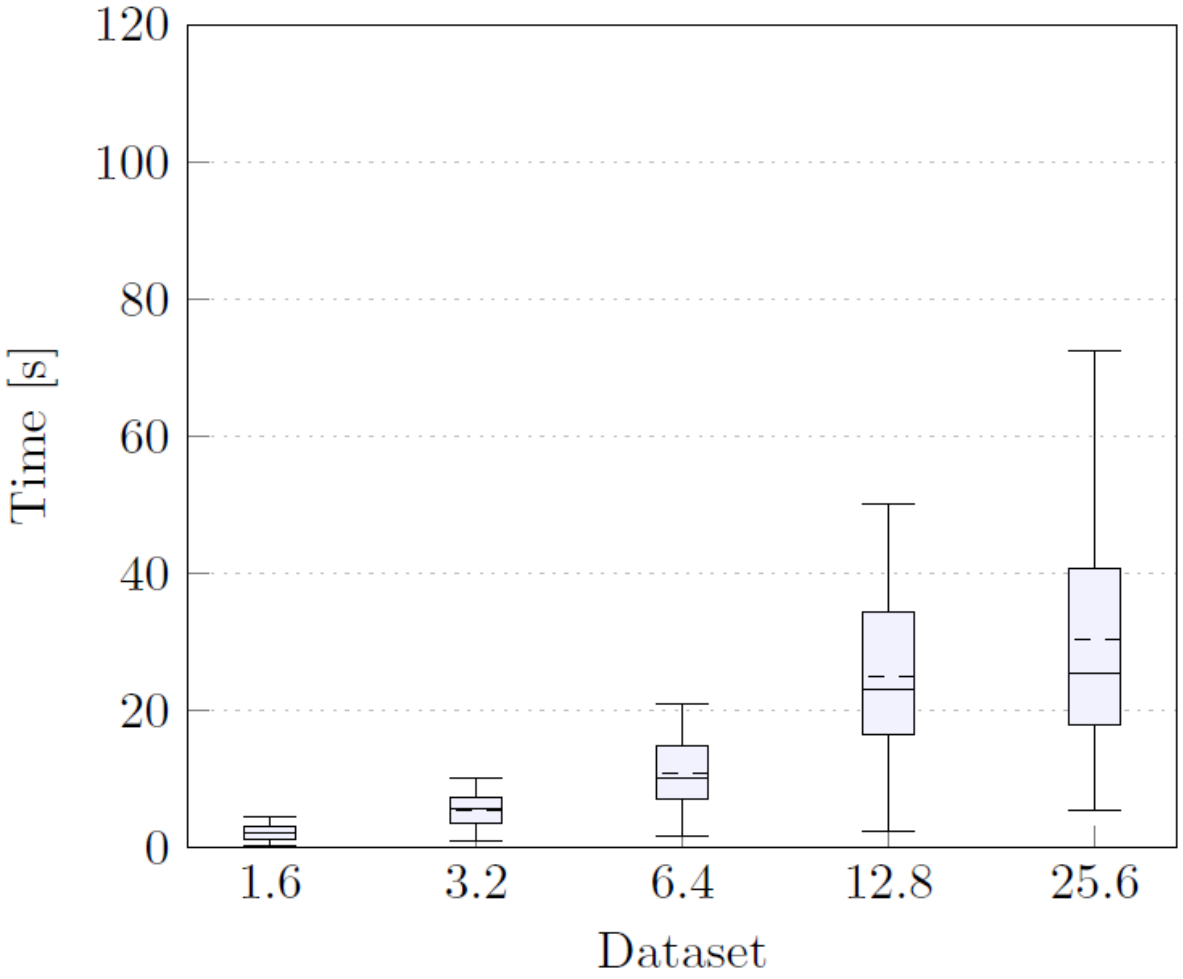
- Queries: 100 node pairs randomly sampled
  - From smallest slice (`Q` code < `100000`)
  - From each slice independently

- Task: Return one (best) path

Performance Results (Full Dataset)

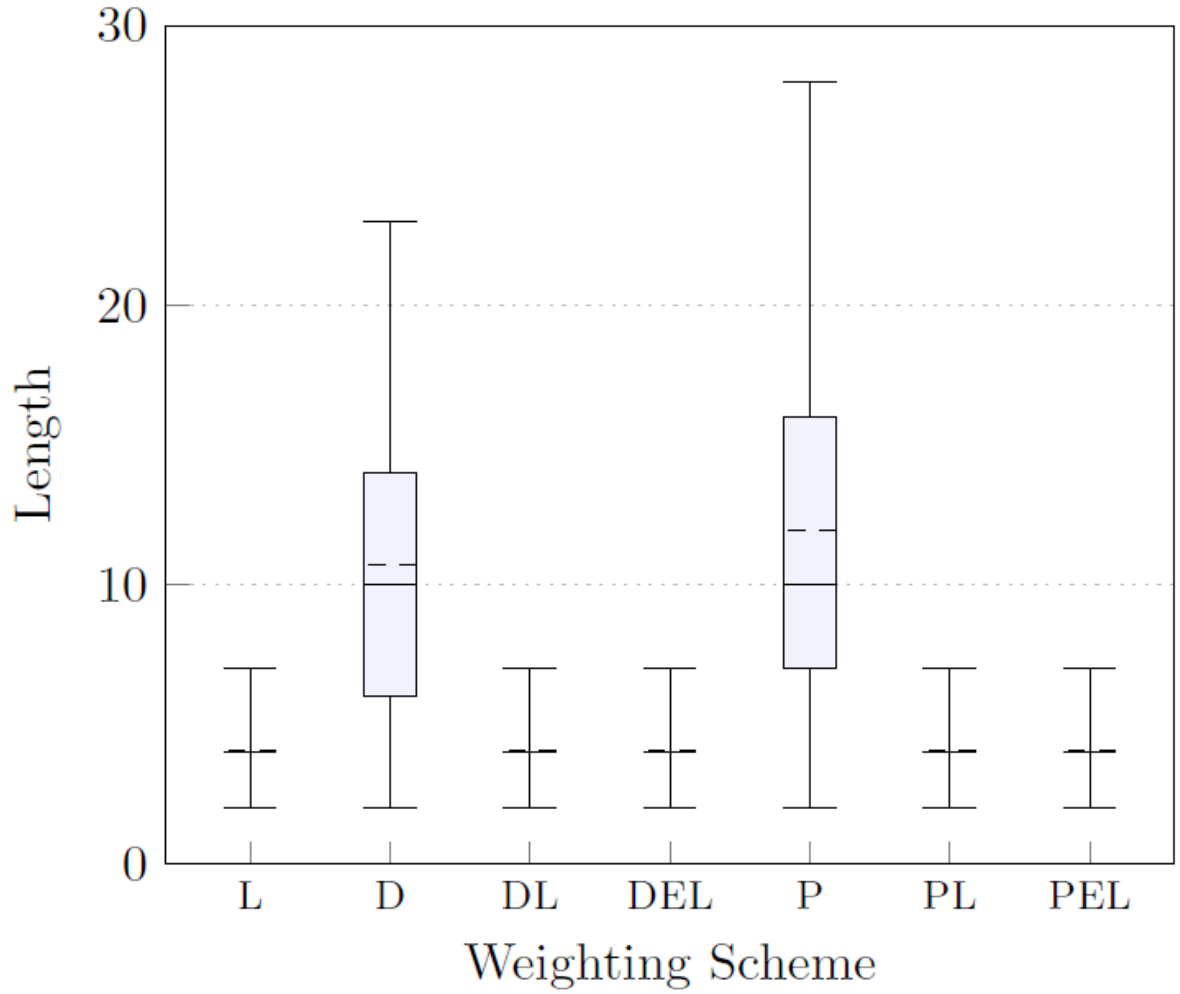


Performance Results (DEL | Various Scales)

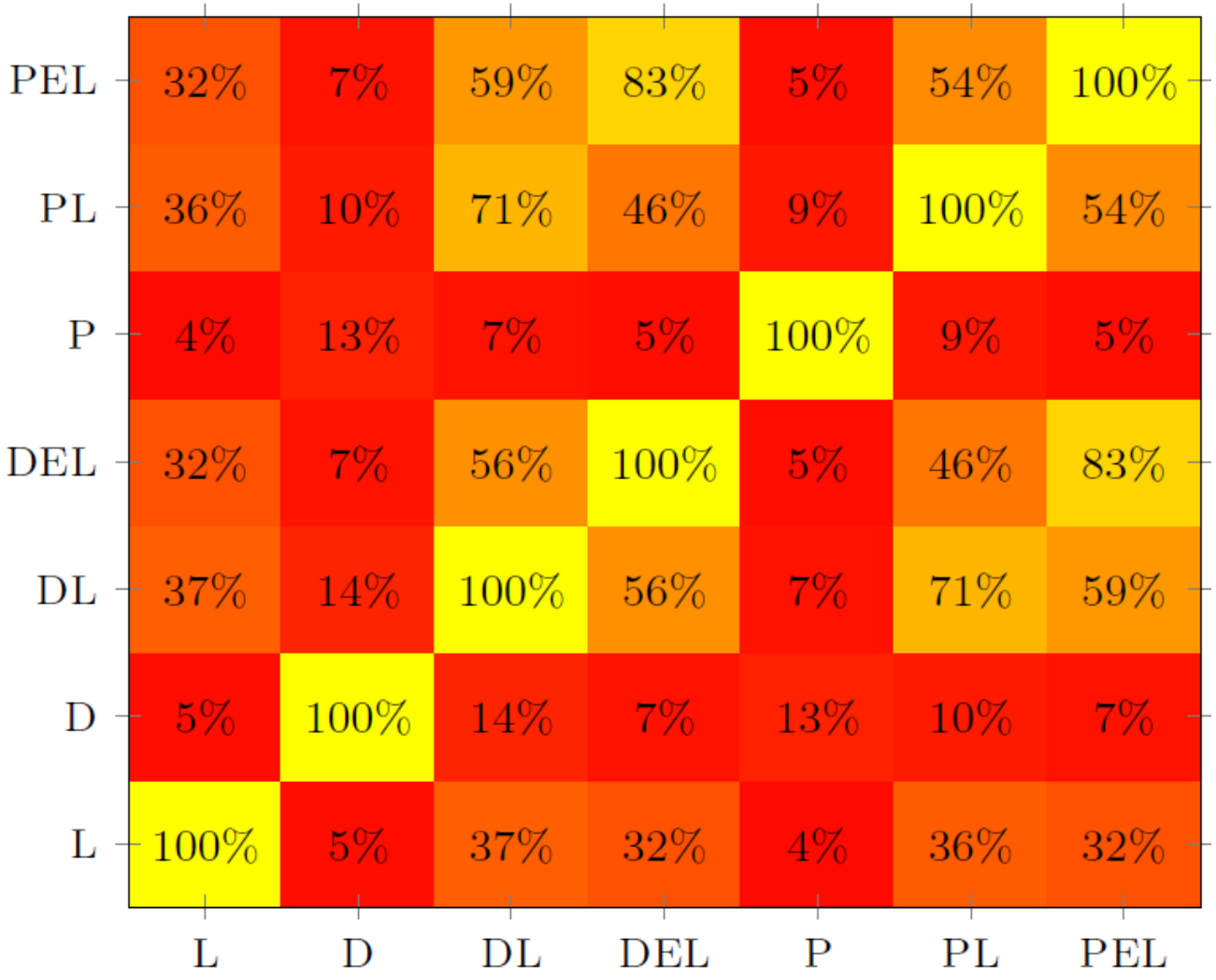


Comparison of weighting schemes

Comparison of path length (full dataset)



How many pairs give the same path? (full dataset)



User Study

Queries: Same type

Node	**Q code**	**Node**	**Q code**
George W. Bush	Q207	Barack Obama	Q76
Bolivia	Q750	Portugal	Q45
Elvis Presley	Q303	Deep Purple	Q101505
Lionel Messi	Q615	Stephen Curry	Q352159
Eiffel Tower	Q243	Pisa Tower	Q39054
Alan Turing	Q7251	Leonhard Euler	Q7604
Samsung	Q20716	Apple Inc.	Q312
Ford Motor Company	Q44294	Volkswagen	Q246
Al Pacino	Q41163	Bill Murray	Q29250
CNN	Q48340	BBC	Q9531

Queries: Different types

Node	**Q code**	**Node**	**Q code**
Donald Trump	Q22686	Pablo Neruda	Q34189
Steam (videogame platform)	Q337535	Barcelona Football Club	Q7156
Everest (mountain)	Q513	FIFA	Q253414
Lord of the Rings	Q15228	Apollo 11 (space mission)	Q43653
RMS Titanic	Q25173	Saturn	Q193
Netflix	Q907311	Napoleon Bonaparte	Q517
Second World War	Q362	Plato	Q859
The Beatles	Q1299	Star Wars	Q462
Odyssey	Q35160	Facebook (web site)	Q355
X Rays	Q34777	Amazon.com	Q3884

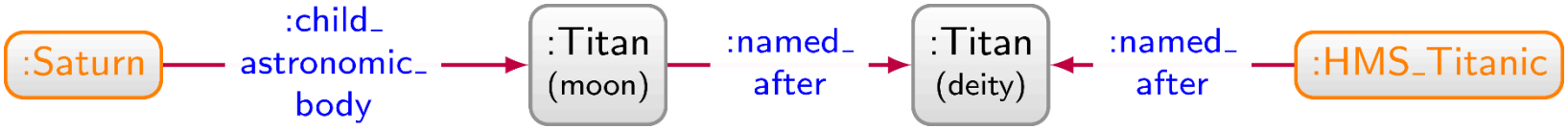
User study

- 10 students
- 1.6 M dataset
- Shown all paths for one query together
- Scores: 1 (very poor) - 7 (very good)
- 79 complete evaluations
  - 4 evaluations per query (node pair)
  - 553 scores

Lowest-rated path

mean score 1.25 ({1,1,1,2})

Highest-rated path



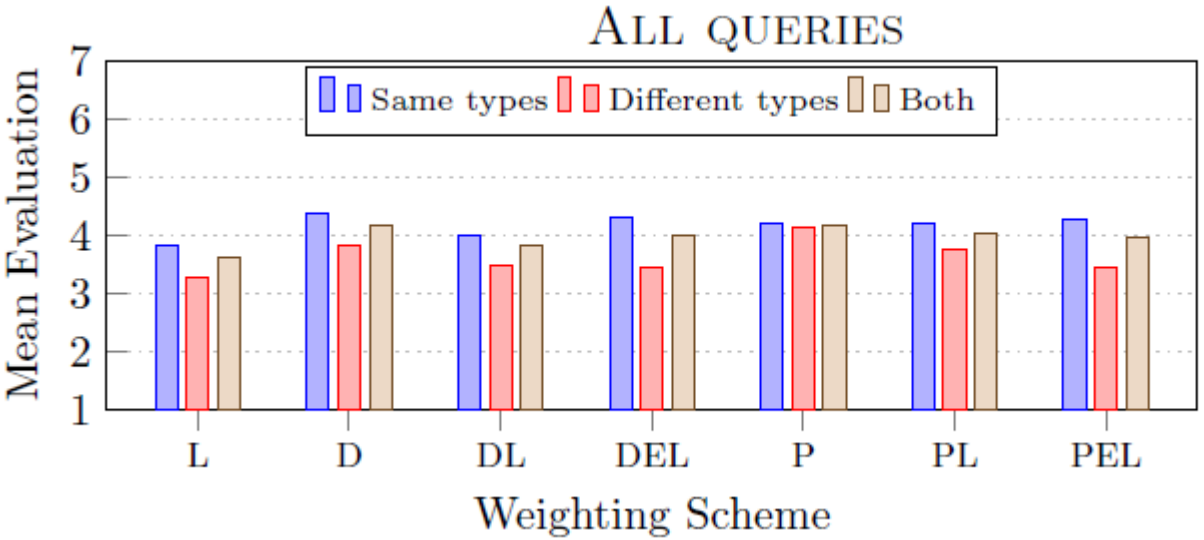
mean score 6.0 ({5,7})

Inter-rater agreement

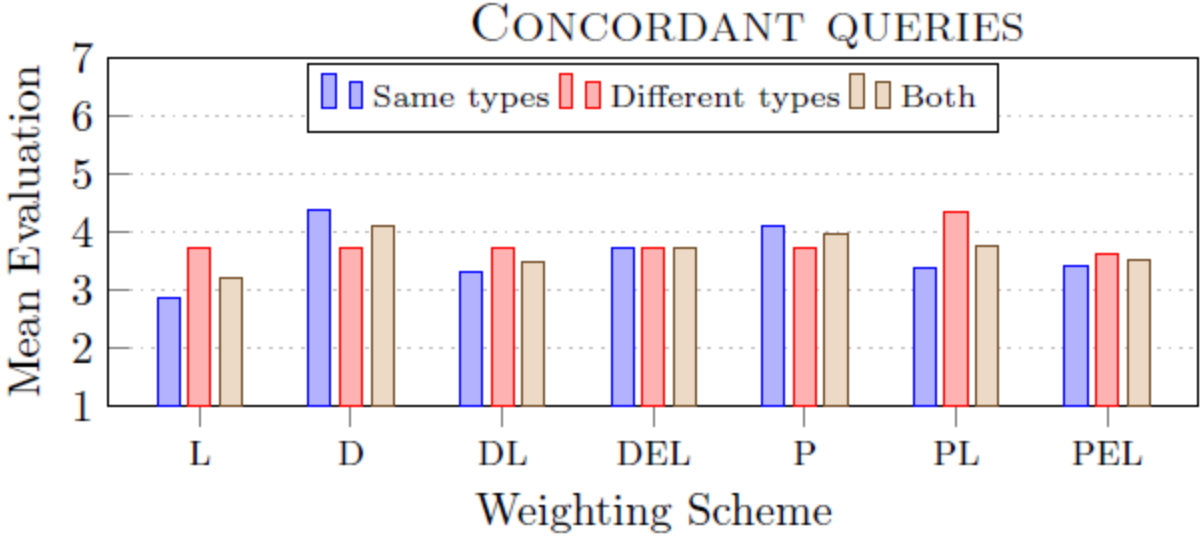
- Kendall's τ correlation (ordinal scales)
  - τ = 0.201
  - Slight, positive agreement

- Two sets of results
  - All
    - τ = 0.201, 20 queries, 79 evaluations
  - Concordant
    - Queries with positive τ correlation only
    - τ = 0.552, 8 queries, 27 evaluations

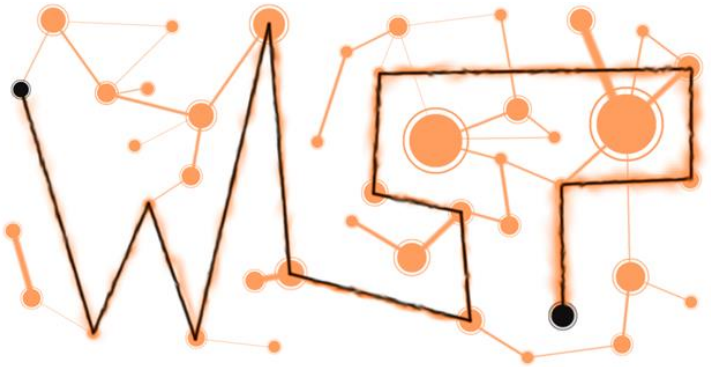
User study: Comparison of weightings



W.S.	p
D	0.022*
DL	0.230
DEL	0.026*
P	0.038*
PL	0.013*
PEL	0.034*



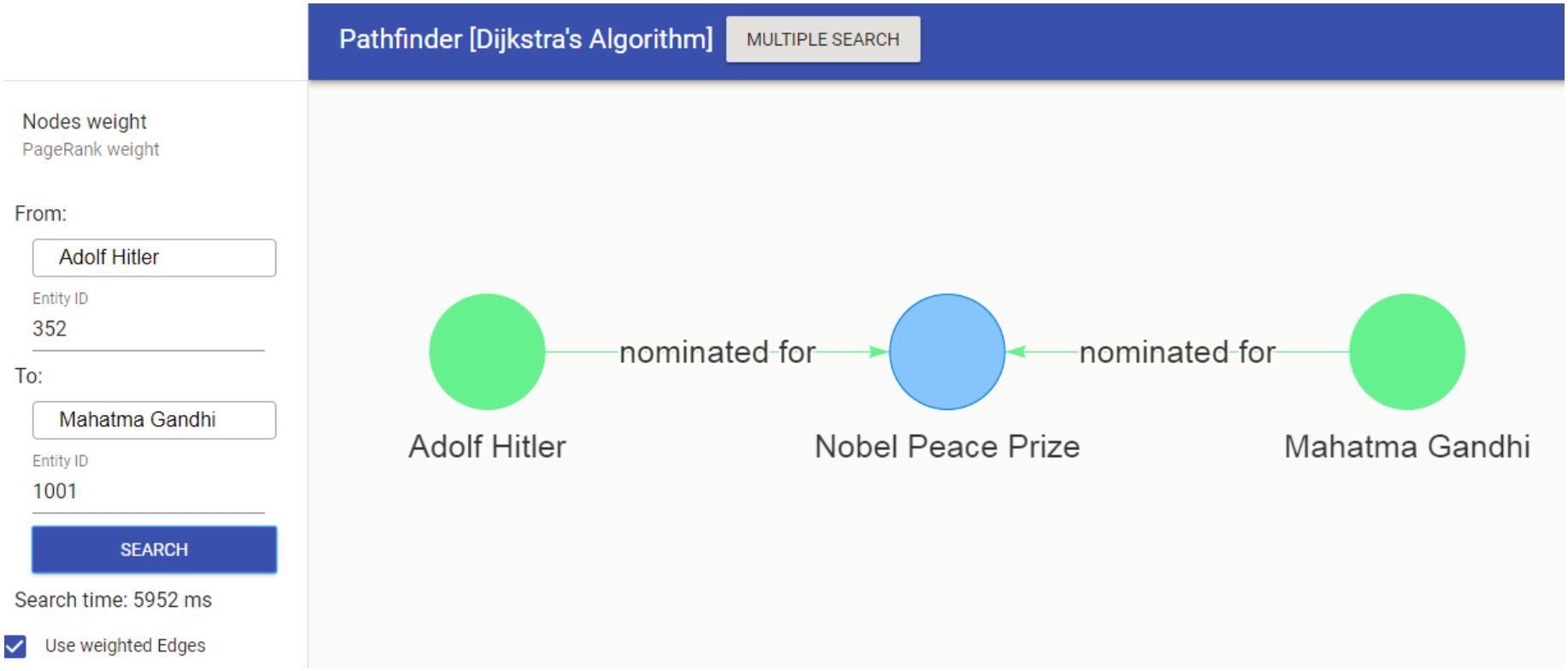
W.S.	p
D	0.006*
DL	0.182
DEL	0.020*
P	0.110
PL	0.016*
PEL	0.118



http://wisp.dcc.uchile.cl

Demo

WiSP Demo


Pathfinder [Dijkstra's Algorithm]
MULTIPLE SEARCH
Nodes weight
PageRank weight
From:
Adolf Hitler
Entity ID
352
To:
Mahatma Gandhi
Entity ID
1001
SEARCH
Search time: 5952 ms
Use weighted Edges
nominated for
nominated for
Adolf Hitler
Nobel Peace Prize
Mahatma Gandhi

Conclusions

Conclusions

- Performance:
  - How are the runtimes?
    - A few seconds (1.6 m) to a few minutes (full dataset)
  - How is the scalability?
    - Linear (roughly)
- Weighting schemes:
  - How similar are paths for different weightings?
    - **DEL** | **PEL** similar; others not so much
  - Does weighting help find interesting paths?
    - Yes!
  - Which weighting finds the most interesting paths?
    - No clear winner (**PL** best in most cases)

Future work

- Top-***k*** queries
- Explore more weightings
- Normalisation / combinations
- Performance? (Parallelism? Approximation?)
- ¡¡¡Evaluation!!!