# RoundTripRank

## Graph-based Proximity with Importance and Specificity

Yuan Fang Univ. of Illinois at Urbana-Champaign

Kevin C.-C. Chang Univ. of Illinois at Urbana-Champaign

Hady W. Lauw Singapore Management University

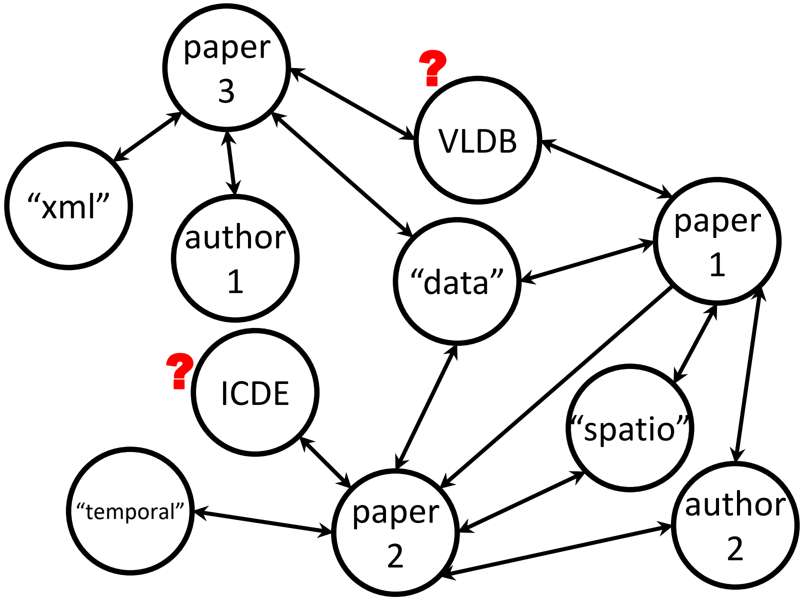
ICDE 2013 @ Brisbane, Australia
April 10, 2013

ARISE
arise.adsc.com.sg

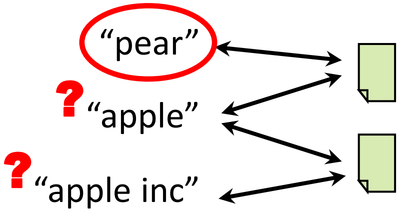
DAIS The Data and Information Systems Laboratories
at The University of Illinois at Urbana-Champaign
*Large Scale Information Management and Mining*

# Graph-based proximity has many applications with different ranking needs

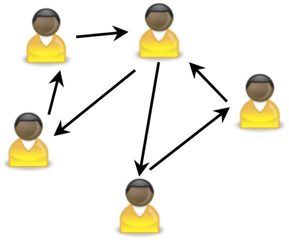
*Citation graph*



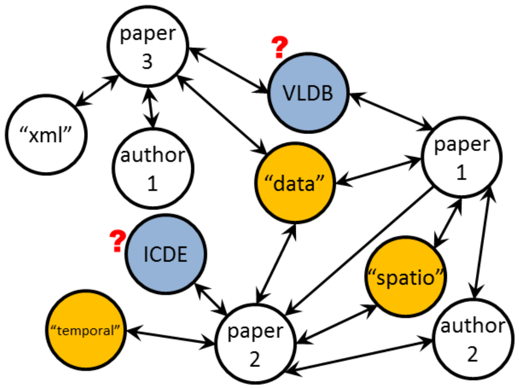
*Query log graph*



*Social network*



# Although various applications involve different needs, ranking by existing graph proximity is limited



**Query**

"spatio", "temporal", "data"

**Matching venues by P-PageRank**

| | |
|---|---|
| SIGMOD | Intl Conference |
| VLDB | Intl Conference |
| ICDE | Intl Conference |

Looks reasonable?
What's missing?

favor very **popular** or **important** venues

only **categorically related** as data topics

**"schema", "matching"**?

# Other venues are needed for different purposes

**Query**

“spatio”, “temporal”, “data”

**More *specific* venues?**

| | | |
|---|---|---|
| quick background study | Spatio-Temporal Databases | Springer Book |
| report preliminary results | Spatio-Temporal Data Mining | Intl Workshop |
| | Temporal Aspects in Information Systems | Working Conference |

**A *balanced* mixture of venues?**

| | | |
|---|---|---|
| important | VLDB | Intl Conference |
| specific | Spatio-Temporal Databases | Springer Book |
| balanced | ACM SIGSPATIAL/GIS | Intl Conference |

# **Specificity** has been traditionally ignored

*Semantics*

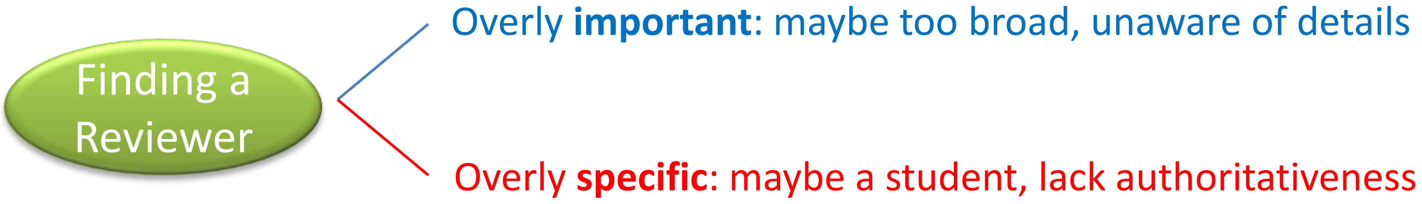
*Methodology*

| | Closeness | Importance | Specificity |
|---|---|---|---|
| Common neighbor | Jaccard coefficient<br>[Jaccard1901]<br>AdamicAdar<br>[Adamic2003] | | |
| Hitting time | Escape probability<br>[Koren2006, Tong2007]<br>SimRank<br>[Jeh2002] | | |
| Reachability | | P-PageRank<br>[Page1999]<br>ObjectRank<br>[Balmin2004]<br>PopRank<br>[Nie2005] | |
| Ad-hoc | Katz<br>[Katz1953] | | InvObjectRank<br>Inverse global ObjectRank<br>Inverse node degree<br>[Hristidis2008] |

# Applications require **varying degrees of trade-off** between importance and specificity

**Observation 1**

Most Tasks Require Both Importance and Specificity.



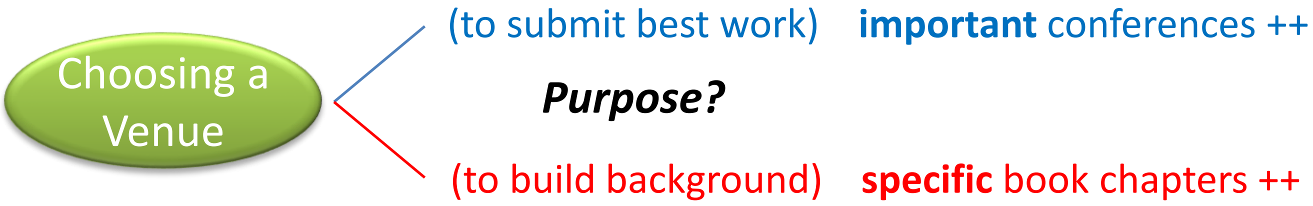
Finding a Reviewer

Overly **important**: maybe too broad, unaware of details

Overly **specific**: maybe a student, lack authoritativeness

**Observation 2**

The Desirable Trade-off Varies from Task to Task.



Choosing a Venue

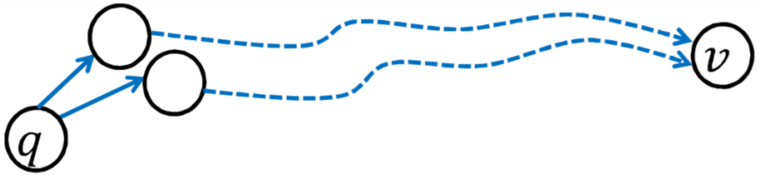
(to submit best work) **important** conferences ++

***Purpose?***

(to build background) **specific** book chapters ++

# Addressing the two observations is challenging

**Challenge 1:** How do we unify importance & specificity into a single proximity measure?



Generalize random walk based importance to integrate specificity.

**Challenge 2:** How do we generalize our unified model to accommodate flexible trade-offs?

more importance

more specificity

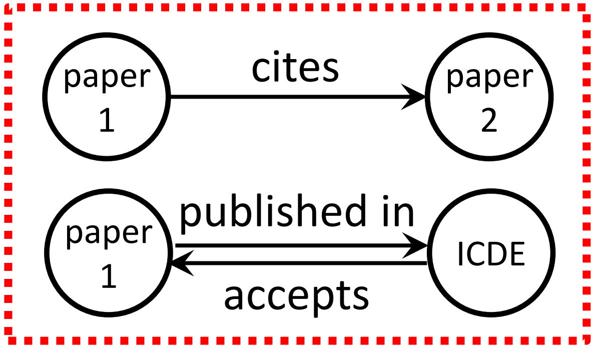
**Challenge 3:** How do we efficiently compute the proximity measure?

Real-time search is indispensable.

# Challenge 1

How do we *unify importance & specificity* into a single proximity measure?

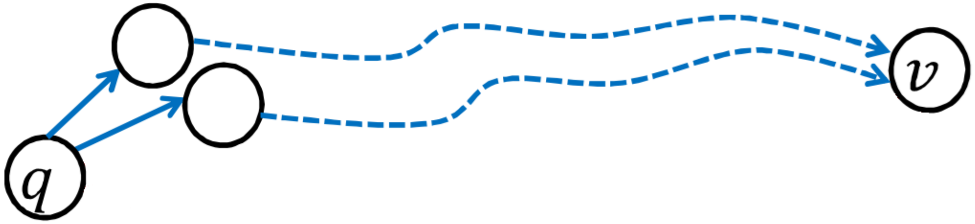
# Let’s first review **reachability-based importance** for generalization to specificity



“citations” or “endorsements”

**If <u>node $v$</u> is important to <u>query $q$</u>…**

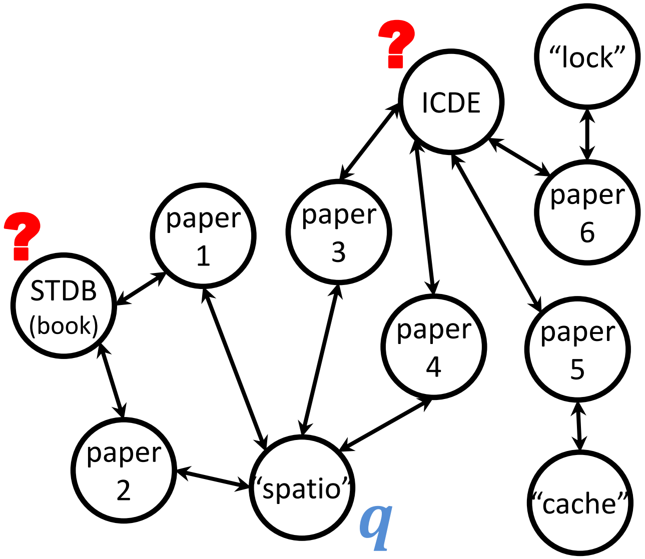
- $q$ is likely to **cite** $v$, directly or indirectly
- Reachability from $q$ to $v$



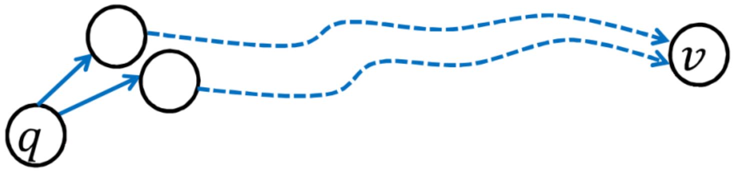
# **Generalize importance to specificity**
# based on the same citation analogy

## If <u>node $v$</u> is specific to <u>query $q$</u>…

- $v$ tends to cite nodes more tailored to $q$
- $q$ is likely to **be cited by** $v$, directly or indirectly
- Reachability from $v$ to $q$

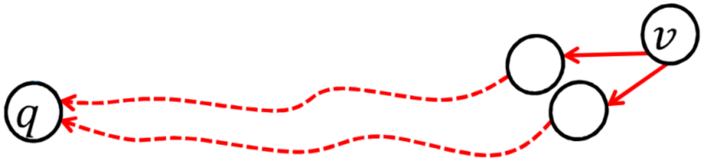


### Importance of $v$ to $q$



forward walk $q \rightarrow v$

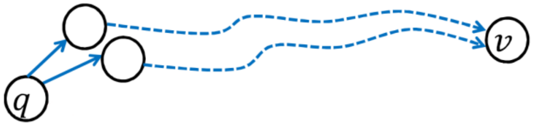
### Specificity of $v$ to $q$



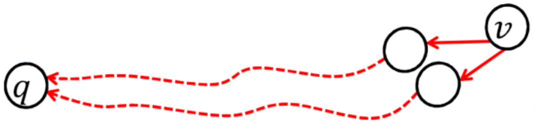
backward walk $v \rightarrow q$

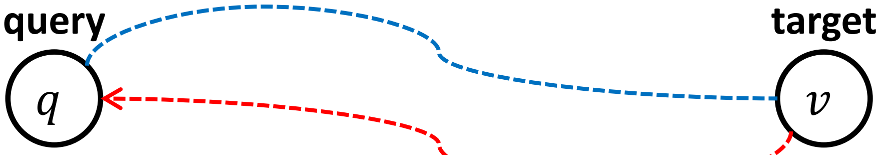
# Unify forward and backward walks into a **round trip** for both importance & specificity

**Importance of $v$ w.r.t. $q$**



**Specificity of $v$ w.r.t. $q$**




*Walk forward in some $L$ steps*

**query**

**target**

*Walk backward in some $L'$ steps*

**Random walk:** $W_0, W_1, \dots, W_L, W_{L+1}, \dots, W_{L+L'}$

**Round trip:** $W_0 = W_{L+L'}$

**Target node:** $W_L$

**RoundTripRank:** $r(q,v) \triangleq p(W_L = v | W_0 = W_{L+L'}, W_0 = q)$

## Challenge 2

How do we *generalize our unified model* to accommodate flexible trade-offs?

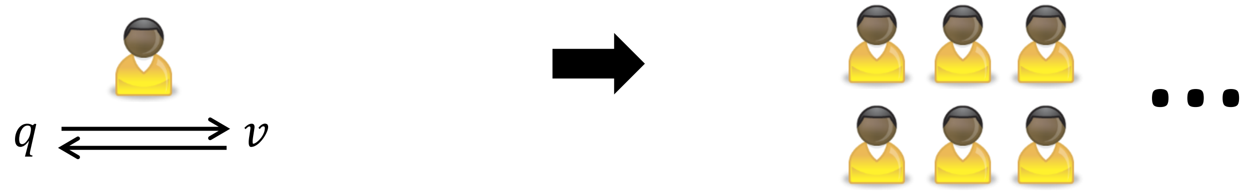
⬇

Based on the same principle of random walk in a round trip.

# Further generalize RoundTripRank using **hybrid random surfers** of different goals

Single random surfer $\omega$

Hybrid random surfer $\Omega$

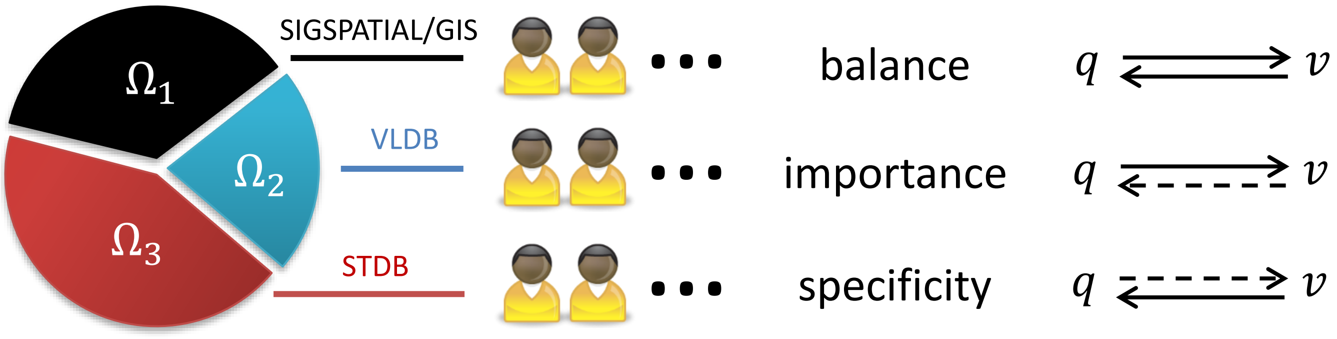


Goal: balance b/w importance and specificity

Different surfers $\omega \in \Omega$ may have different goals!

# Generalize the behaviors of hybrid random surfers for **customizable trade-offs**

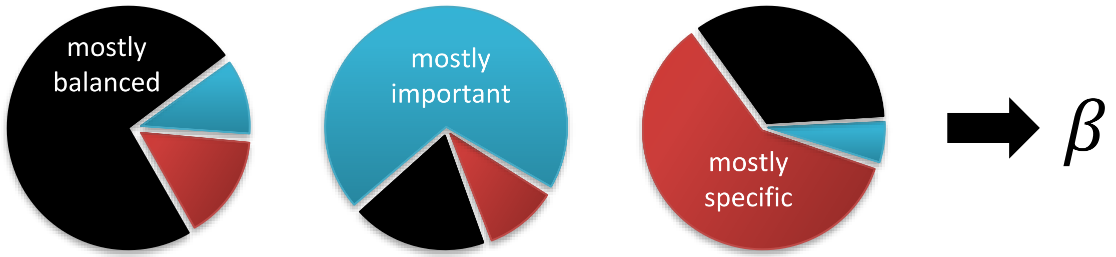
**Hybrid Surfers**



**RoundTripRank+**

$$r_{\Omega}(q,v) \triangleq p(x = v | \forall \omega \in \Omega: W_0^{\omega} = W_{L+L'}^{\omega} = q, W_L^{\omega} = x)$$

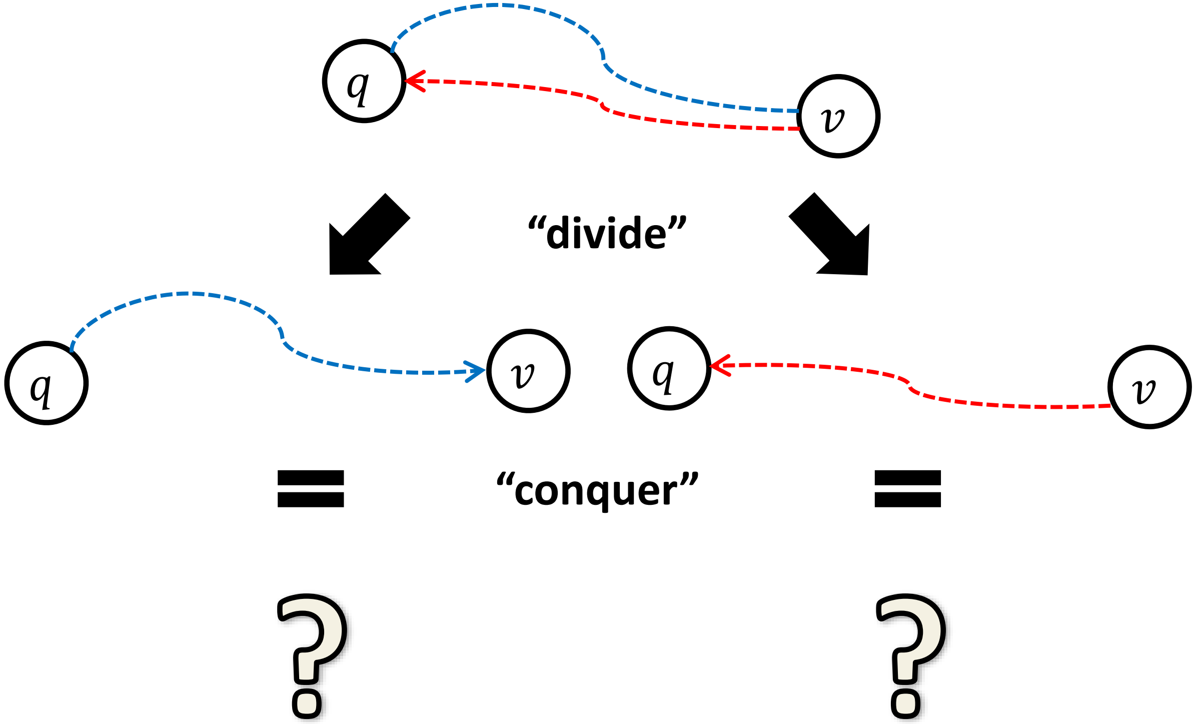
**Adjusting Composition**



## Challenge 3

How do we efficiently compute the proximity measure?

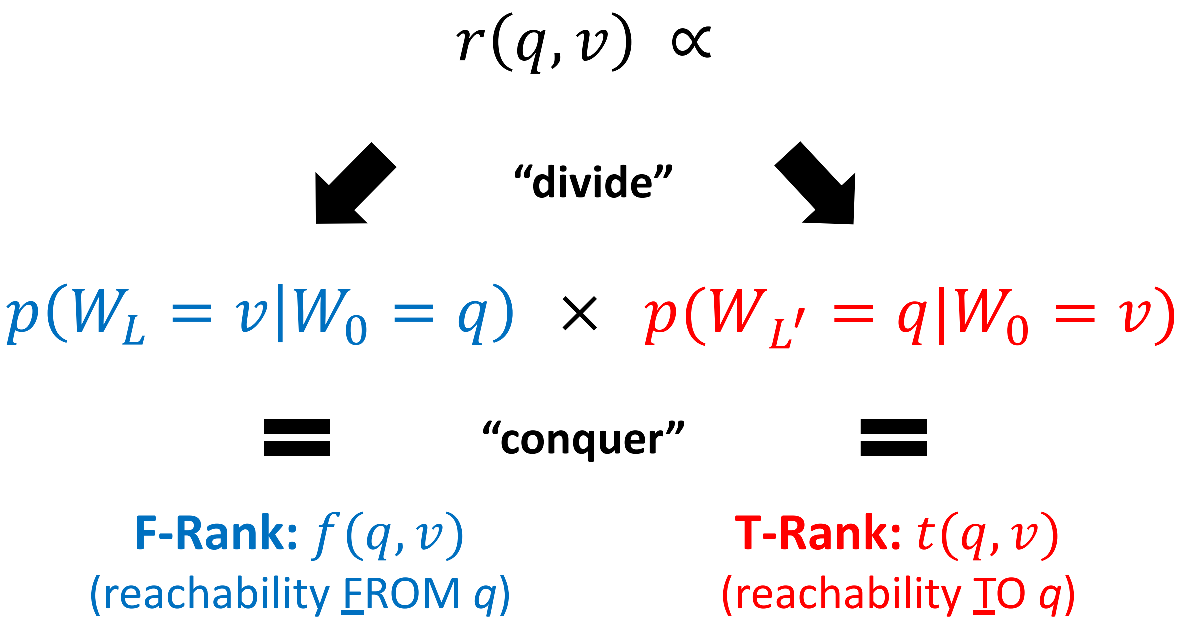
# Compute RoundTripRank by “**divide & conquer**”



**“divide”**

**“conquer”**

# Compute RoundTripRank by “**divide & conquer**”



$r(q,v) \propto$

“divide”

$p(W_L = v | W_0 = q) \times p(W_{L'} = q | W_0 = v)$

“conquer”

**F-Rank:** $f(q,v)$
(reachability FROM $q$)

**T-Rank:** $t(q,v)$
(reachability TO $q$)

**RoundTripRank**: $r(q,v) \propto f(q,v)t(q,v)$

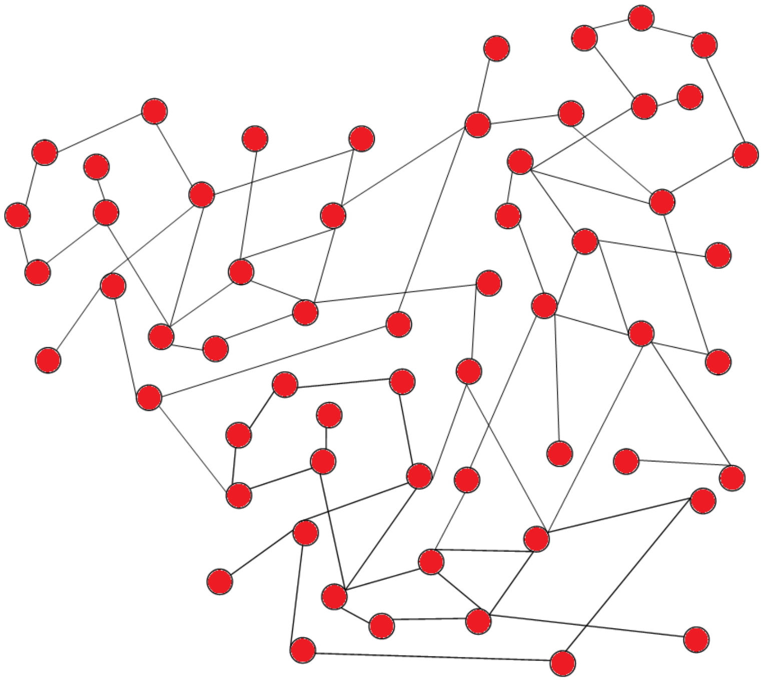
**RoundTripRank+**: $r_\Omega(q,v) \propto f(q,v)^{1-\beta} t(q,v)^\beta$

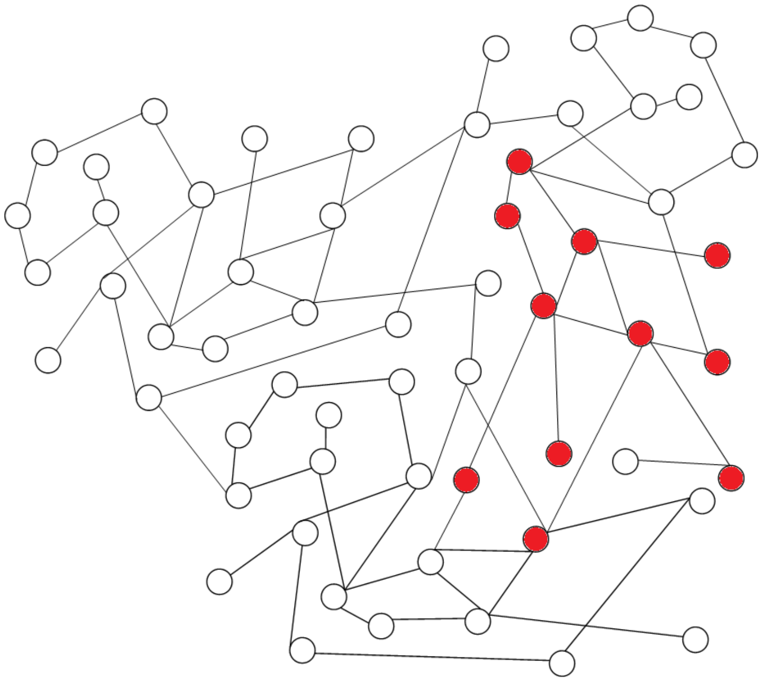
Specificity bias: $\beta = \frac{|\Omega_1|+|\Omega_3|}{2|\Omega_1|+|\Omega_2|+|\Omega_3|} \in [0,1]$

# Top-*K* ranking is more practical & scalable

**Full ranking**
[over the entire graph]

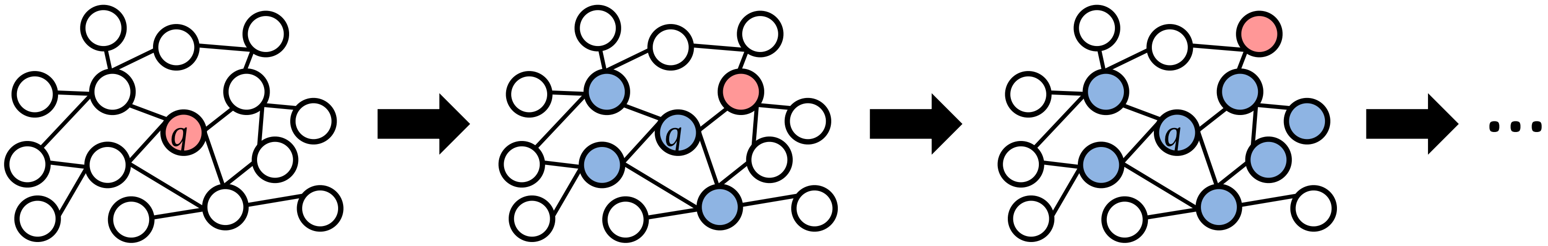
**Top-*K* ranking**
[based on a neighborhood]





# Branch-and-bound algorithm

**Neighborhood expansion**

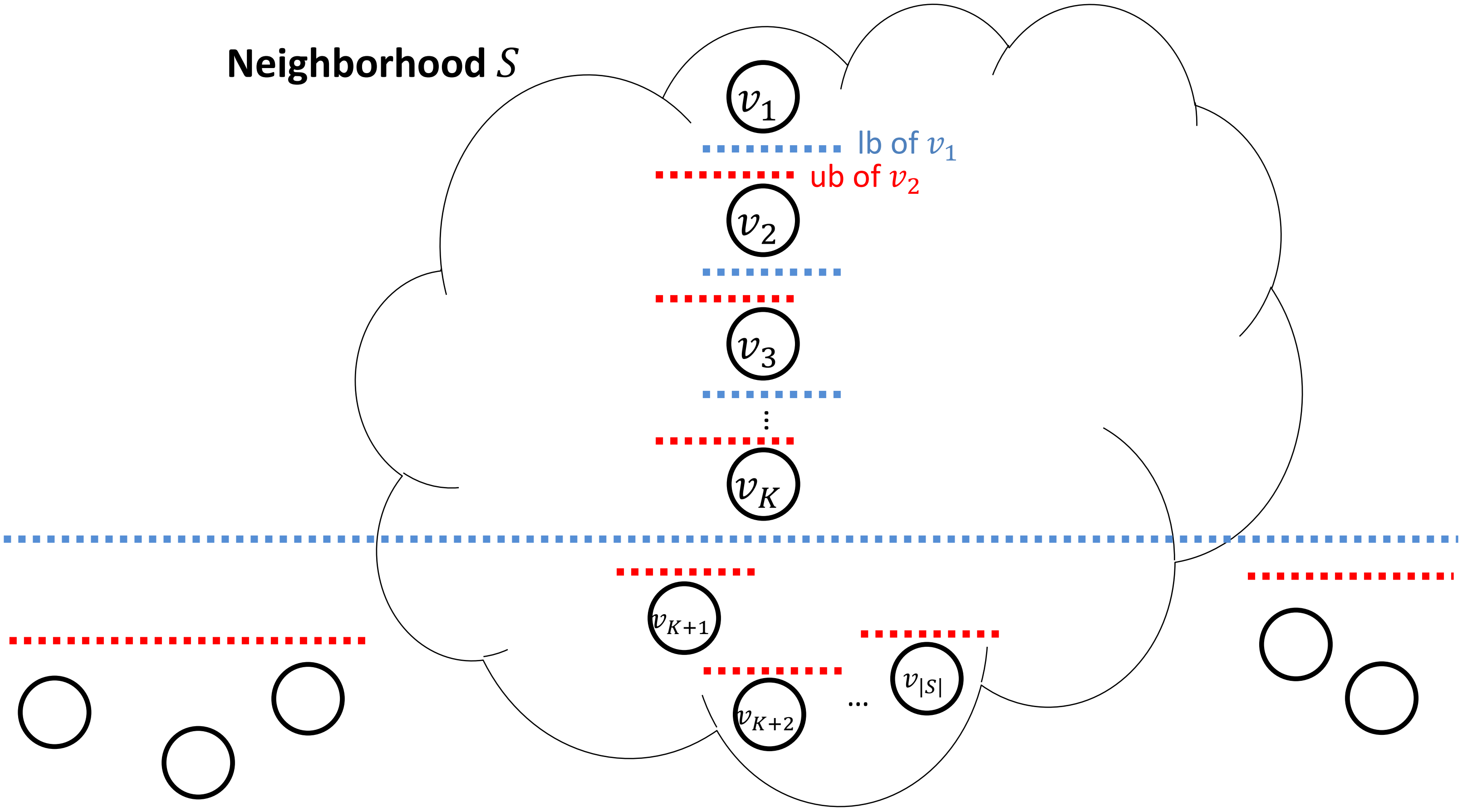


**Bounds**

Given the current neighborhood $S$:

$$\left.\begin{aligned}\check{r}(q,v) \leq r(q,v) \leq \hat{r}(q,v), \ \forall v \in S\\ r(q,v) \leq \hat{r}(q), \ \forall v \notin S\end{aligned}\right\} \text{determine top-}K\text{ nodes}$$

# Is a candidate top-$K$ ranking $v_1, \ldots, v_K$ correct?



$$\check{r}(q, v_i) > \hat{r}(q, v_{i+1}) - \epsilon \qquad \forall i \in \{1, \ldots, K-1\}$$

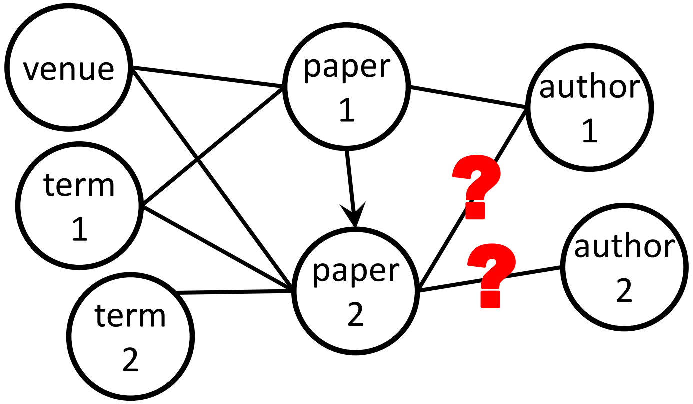
$$\check{r}(q, v_K) > \max\left\{\hat{r}(q, v_{K+1}), \ldots, \hat{r}(q, v_{|S|}), \hat{r}(q)\right\} - \epsilon$$

# Experiments

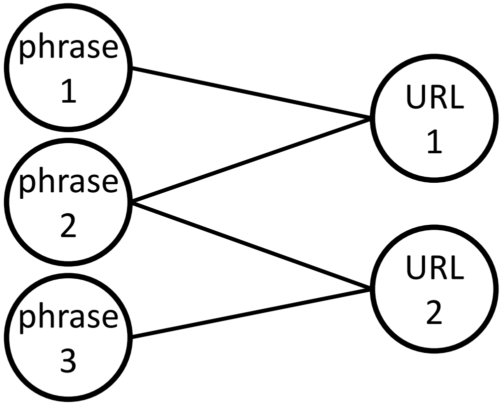
# Experimental Setup

**Graphs**

Bibliographic network (BibNet)



Query log graph (QLog)

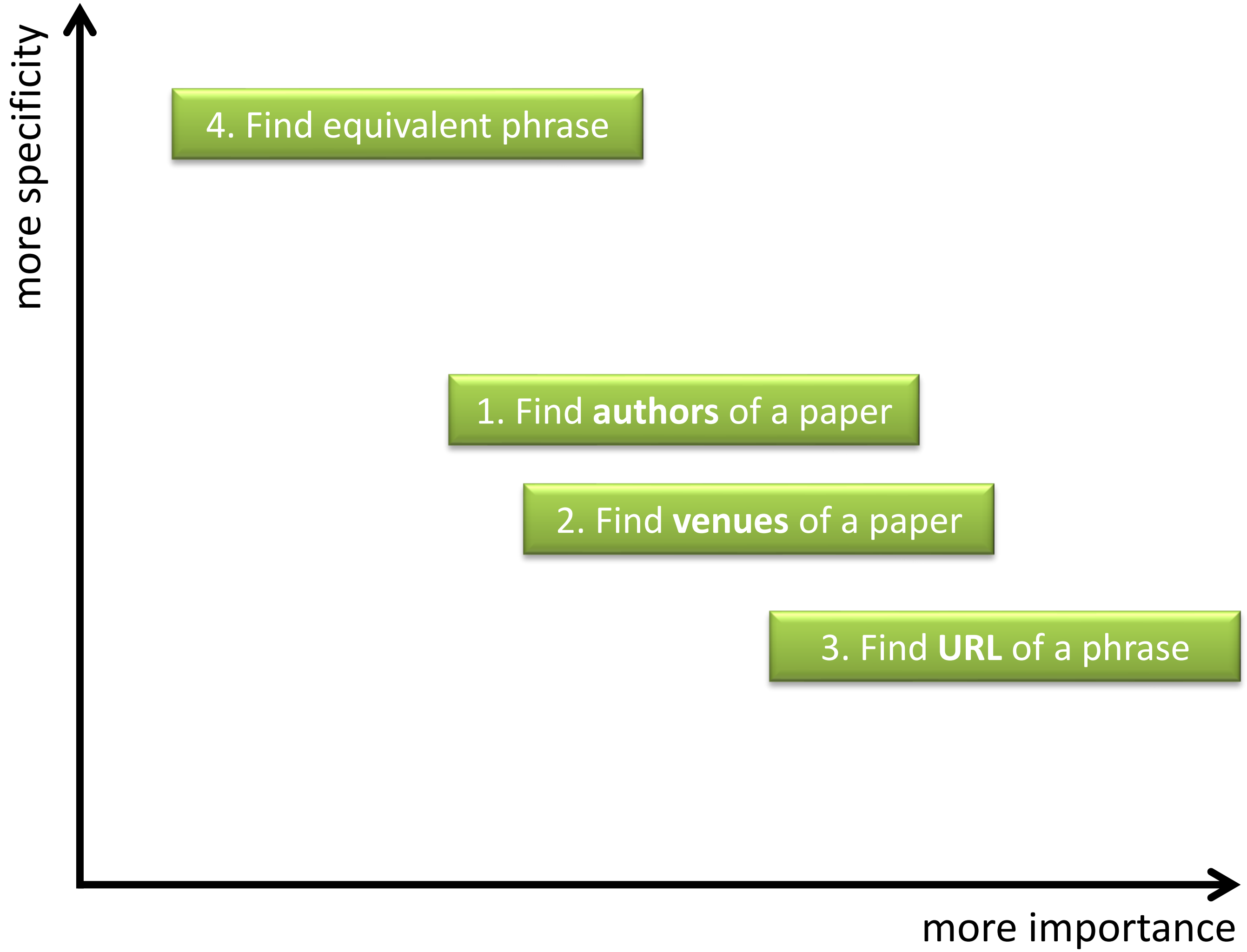


**Evaluation methodology**

**Hide-and-rediscover**

- Reserve nodes with known associations to query
- Remove those associations from the graph
- Can a proximity measure still rank those nodes highly?

# Evaluation Tasks



more specificity

4. Find equivalent phrase

1. Find **authors** of a paper

2. Find **venues** of a paper

3. Find **URL** of a phrase

more importance

Finding 1

# Both **importance & specificity** are needed

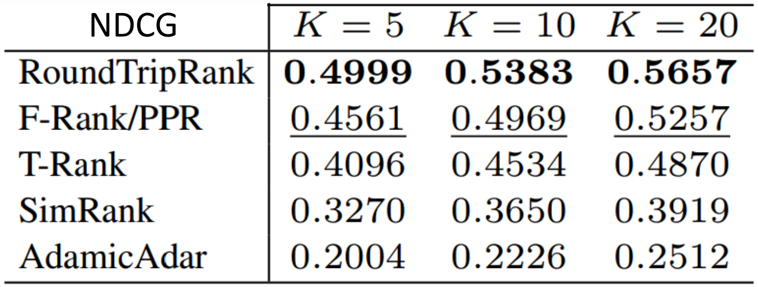
## Venues matching "spatio temporal data"

| F-Rank/PPR | T-Rank | RoundTripRank |
|---|---|---|
| dell | dell c1295 | dell battery |
| dell com | battery for dell inspiron 8000 | battery for dell inspiron 8000 |
| dell computers | 312 0068 | dell |
| *important* | *specific* | *balanced* |

## Phrases similar to "dell notebook"

| F-Rank/PPR | T-Rank | RoundTripRank |
|---|---|---|
| dell | dell c1295 | dell battery |
| dell com | battery for dell inspiron 8000 | battery for dell inspiron 8000 |
| dell computers | 312 0068 | dell |
| *important* | *specific* | *balanced* |

## Quantitative evaluation (hide-and-rediscover)

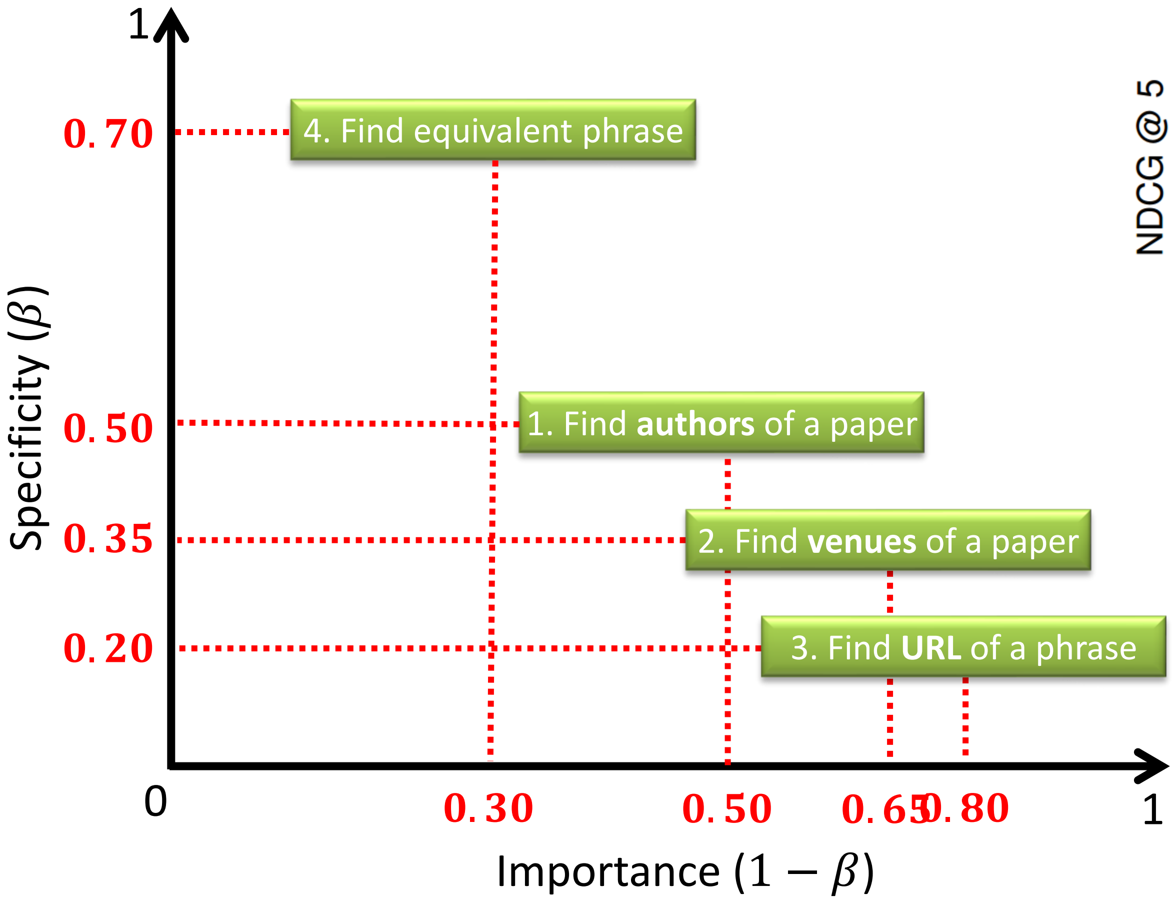


| NDCG | $K = 5$ | $K = 10$ | $K = 20$ |
|---|---|---|---|
| RoundTripRank | **0.4999** | **0.5383** | **0.5657** |
| F-Rank/PPR | $\underline{0.4561}$ | $\underline{0.4969}$ | $\underline{0.5257}$ |
| T-Rank | 0.4096 | 0.4534 | 0.4870 |
| SimRank | 0.3270 | 0.3650 | 0.3919 |
| AdamicAdar | 0.2004 | 0.2226 | 0.2512 |

+ 8% ~ 10%

Finding 2

# Optimal trade-offs $\beta^*$ **vary** task by task

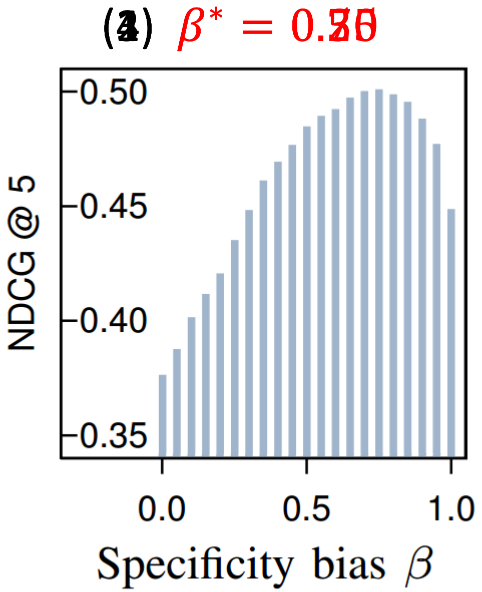


4. Find equivalent phrase

1. Find **authors** of a paper

2. Find **venues** of a paper

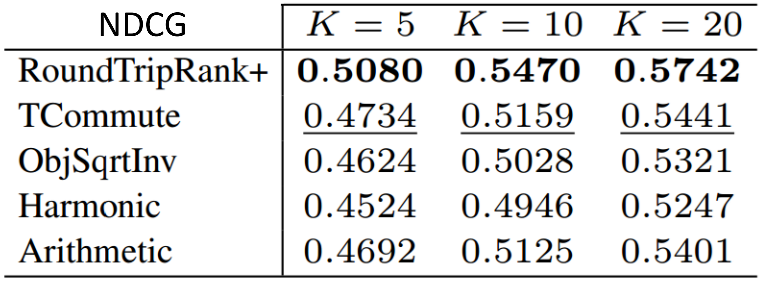
3. Find **URL** of a phrase



Finding 2

# **Optimal trade-offs $\beta^*$ vary** task by task

**Comparison to non-customizable dual-sensed proximity**



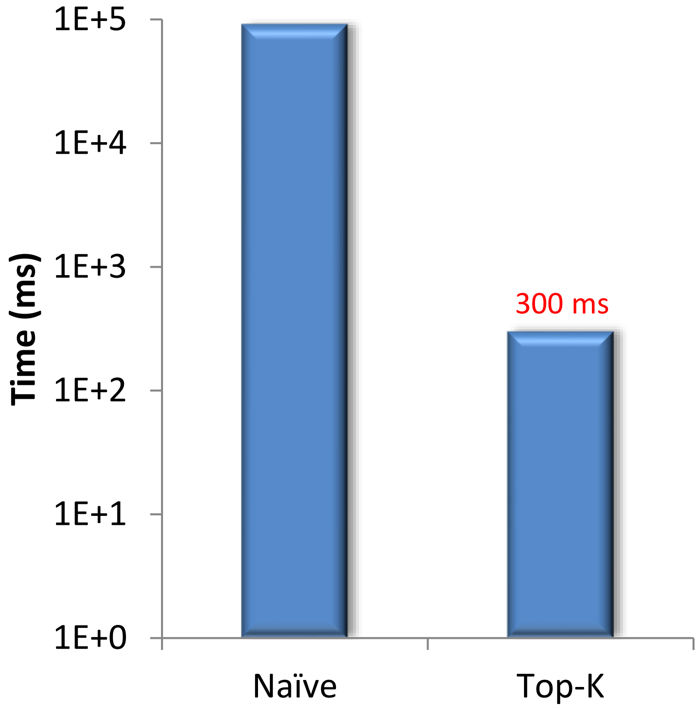
| NDCG | $K=5$ | $K=10$ | $K=20$ |
|---|---|---|---|
| RoundTripRank+ | **0.5080** | **0.5470** | **0.5742** |
| TCommute | 0.4734 | 0.5159 | 0.5441 |
| ObjSqrtInv | 0.4624 | 0.5028 | 0.5321 |
| Harmonic | 0.4524 | 0.4946 | 0.5247 |
| Arithmetic | 0.4692 | 0.5125 | 0.5401 |

+ 6% ~ 7%

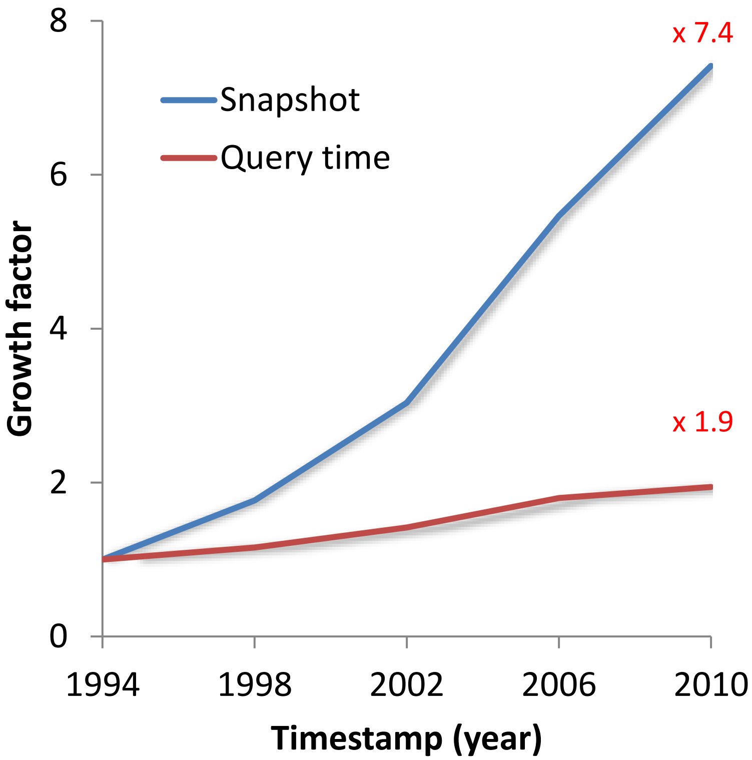
Finding 3

# Our top-*K* method is efficient & scalable

**Efficiency**



**Scalability**



# Conclusion

Importance as “Reachability” → Specificity as “Returnability”

“Reachability” + “Returnability” → a Round Trip