Middleware 2015

Dec. 7-11, 2015, Vancouver, Canada

Scaling Out Link Prediction with SNAPLE: 1 Billion Edges and Beyond

Anne-Marie Kermarrec, François Taïani, Juan M. Tirado

UMR IRISA

UNIVERSITÉ DE RENNES 1

Inria INVENTEURS DU MONDE NUMÉRIQUE

esir ECOLE SUPERIEURE D'INGENIEURS DE RENNES

Algorithms and Middleware

ars technica

ALL APPLE ASK ARS BUSINESS GADGETS GAMING MICROSOFT OPEN SOURCE SCIE

TECH POLICY

NEWS GUIDES REVIEWS FUTURE OF CARS

› Gear & Gadgets

Essential toys, tools, and hardware

Netflix never used its $1 million algorithm due to engineering costs

By Casey Johnston | Published April 13, 2012 4:25 PM

Netflix awarded a $1 million prize to a developer team in 2009 for an algorithm that increased the accuracy of the company's recommendation engine by 10 percent. But today it doesn't use the million-dollar code, and has no plans to implement it in the future, Netflix announced on its blog Friday. The post goes on to explain why: a combination of too much engineering effort for the results, and a shift from movie recommendations to the "next level" of personalization caused by the transition of the business from mailed DVDs to video streaming.

LATEST TO

Algorithms and Middleware

ars technica

ALL APPLE ASK ARS BUSINESS GADGETS GAMING MICROSOFT OPEN SOURCE SCIE

TECH POLICY

NEWS GUIDES REVIEWS FUTURE OF CARS

› Gear & Gadgets Essential toys, tools, and hardware

Netflix never used its $1 million algorithm due to engineering costs

By Casey Johnston | Published April 13, 2012 4:25 PM

Netflix awa **$1 million prize** veloper team in 2009 for an algorithm that increased the accuracy of the compar **recommendation** engine by 10 percent. But today it doesn't use the million-dollar code, and has no plans to implement it in the future, Netflix announced on its blog Friday. The post goes on to explain why: a combination of too much **too much engineering effort** n movie recommendations to the "next level" of personalization caused by the transition of the business from mailed DVDs to video streaming.

LATEST TO

NETFLIX

amazon webservices™

Apache

Large **Distributed** **Complex**

› Gear & Gadgets | Essential toys, tools, and hardware

Netflix never used its $1 million algorithm due to engineering costs

By Casey Johnston | Published April 13, 2012 4:25 PM

Netflix awa **$1 million prize** veloper team in 2009 for an algorithm that increased the accuracy of the compar **recommendation** engine by 10 percent. But today it doesn't use the million-dollar code, and has no plans to implement it in the future, Netflix announced on its blog Friday. The post goes on to explain why: a combination of too much **too much engineering effort** n movie recommendations to the "next level" of personalization caused by the transition of the business from mailed DVDs to video streaming.

LATEST TO

Message and Plan

Tools impose what tasks you can solve

This talk

- The Tool: **Gather Apply Scatter Graph Engines**
- The Task: **Link Prediction** (LP) **in Large Graphs**

Message and Plan

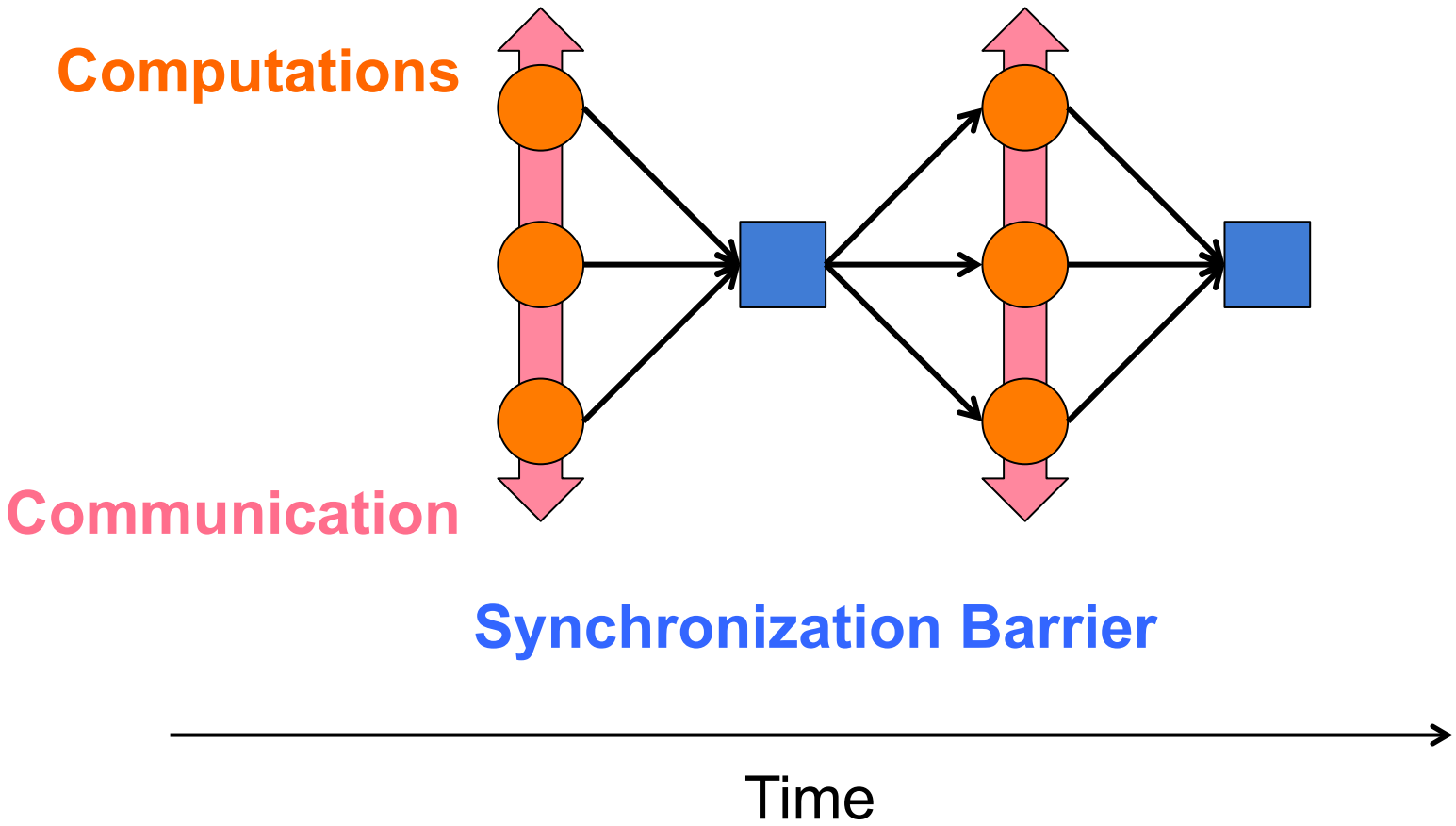
Tools impose what tasks you can solve

This talk

- The Tool: **Gather Apply Scatter Graph Engines**
- The Task: **Link Prediction** (LP) **in Large Graphs**

The GA(S) Model

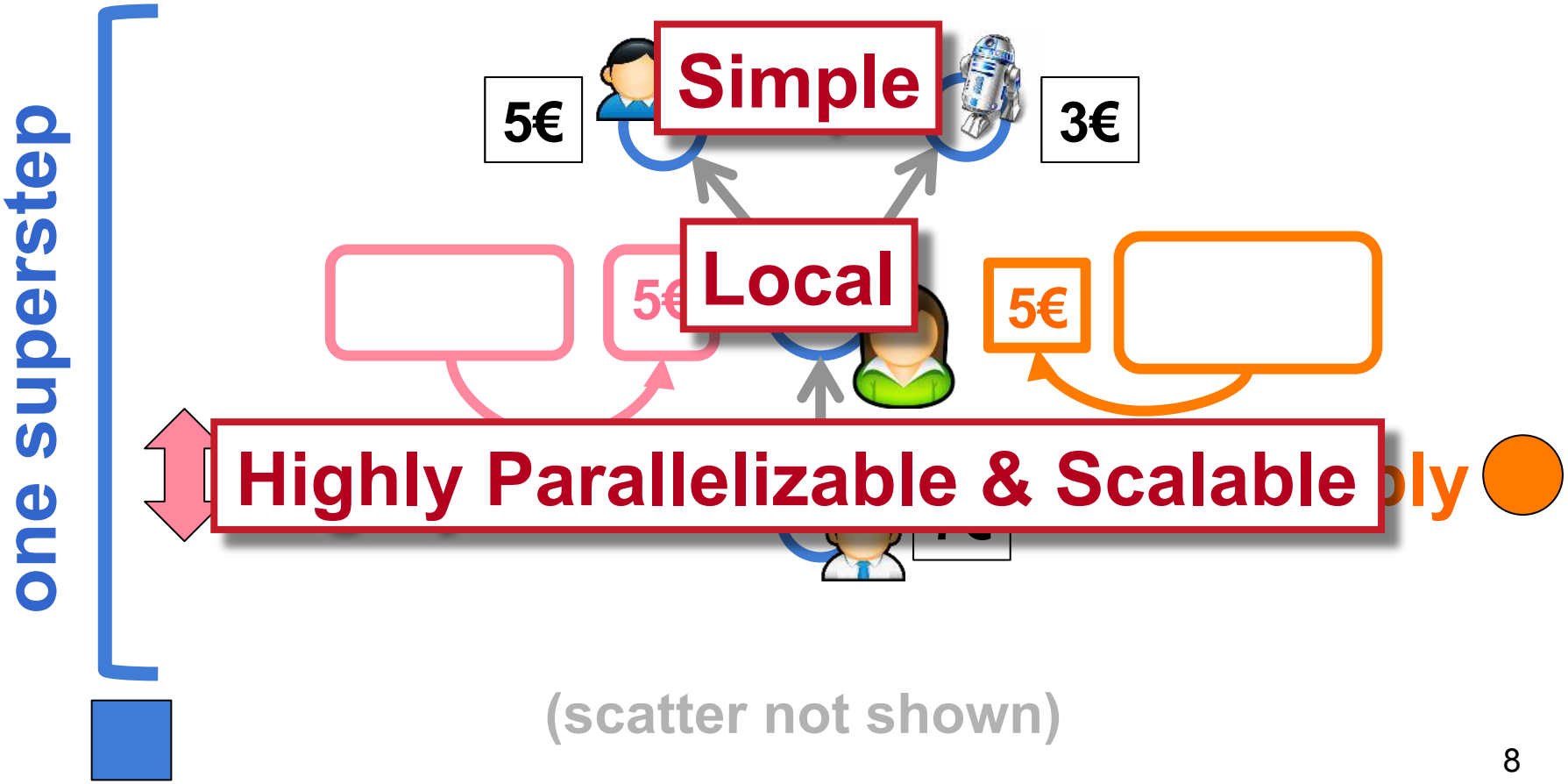
- **Gather Apply Scatter**
  - ➔ Adaptation of **B**ulk **S**ynchronous **P**arallel (BSP) to graphs



The GA(S) Model

Think like a vertex (Pregel, Giraph, GraphLab, ...)

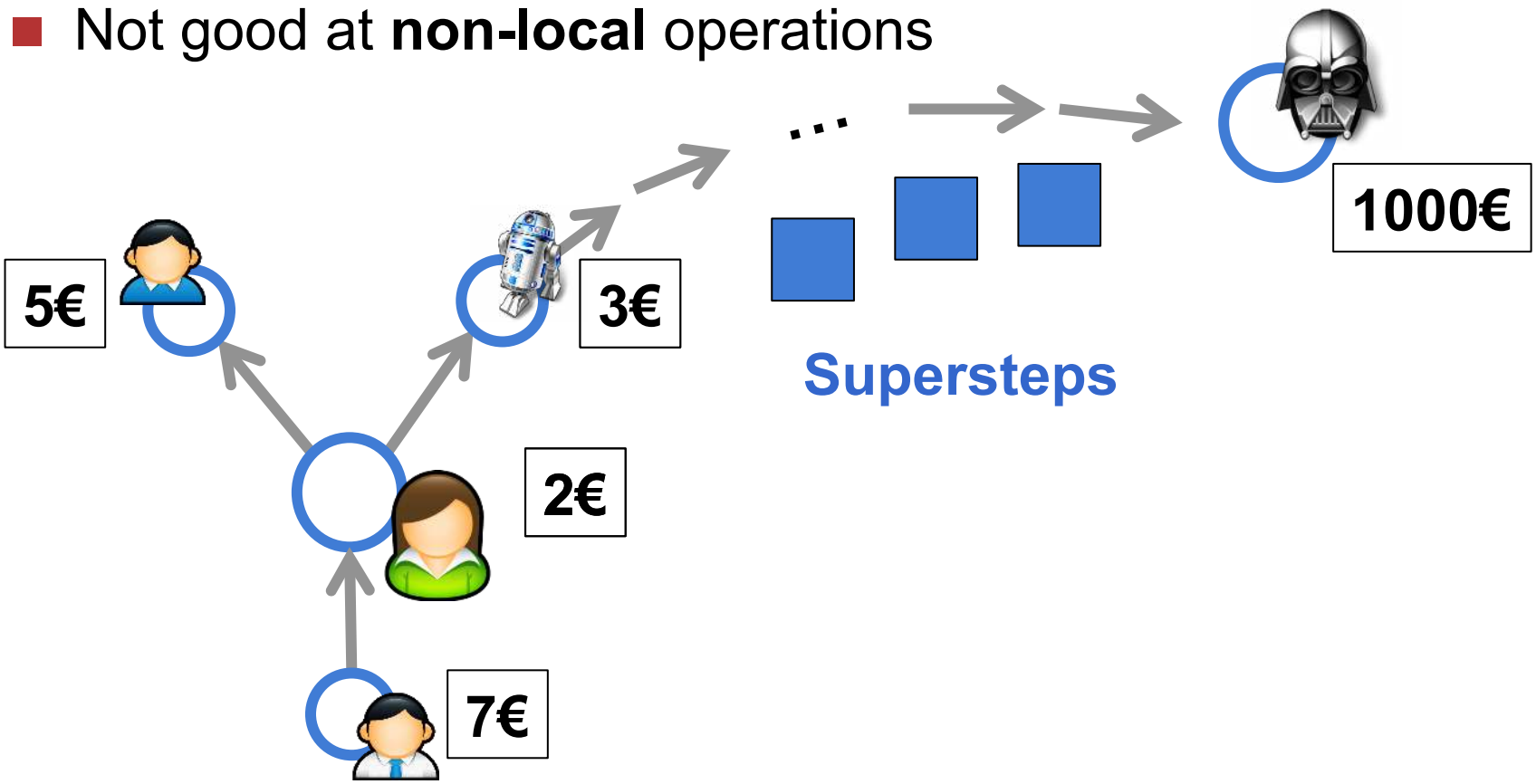
- Example: "Top net worth in my 1-hop network."



The GA(S) Model

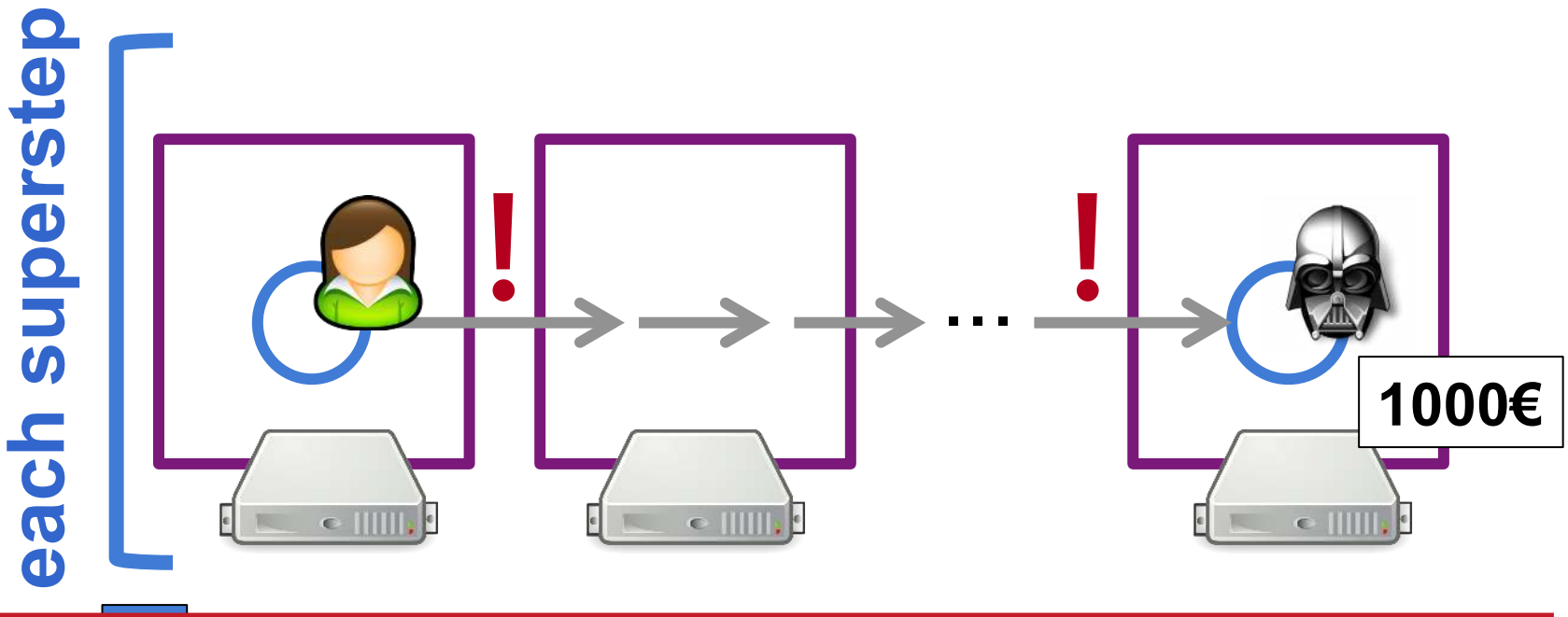
Limitations

- Not good at **non-local** operations



Non Local Computations

- GAS engines provide **horizontal scalability**



Non-locality: Network and memory costs

Too much non-locality: scalability hit

Message and Plan

Tools impose what tasks you can solve

This talk

- The Tool: **Gather Apply Scatter Graph Engines**
- The Task: **Link Prediction** (LP) **in Large Graphs**

Home | 18 Notifications | Messages | Search Twitter

Who to follow · Refresh · View all

UBO - Univ. Brest @UBO_U...
Followed by Université Renn...
Follow

Jay Kreps @jaykreps
Follow

Find friends

Link prediction

Jobs Advertise Businesses Media
Developers

While you were away...

Yehia Elkhatib @yelkhatib · Dec 9
This one's for you @hejinb and @mrowebot
youtube.com/watch?v=wGieq0...
1 2 View media

Collège Doctoral @CDoctoralULNF · 19h
Ouverture des inscriptions au concours #MT180FR #Doctorant #Lille
formadoc.net

Le challenge « Ma thèse en 180 secondes », c'est quoi ?

1 formation à la communication dans votre région + 1 concours national ET international avec de nombreux prix à gagner.

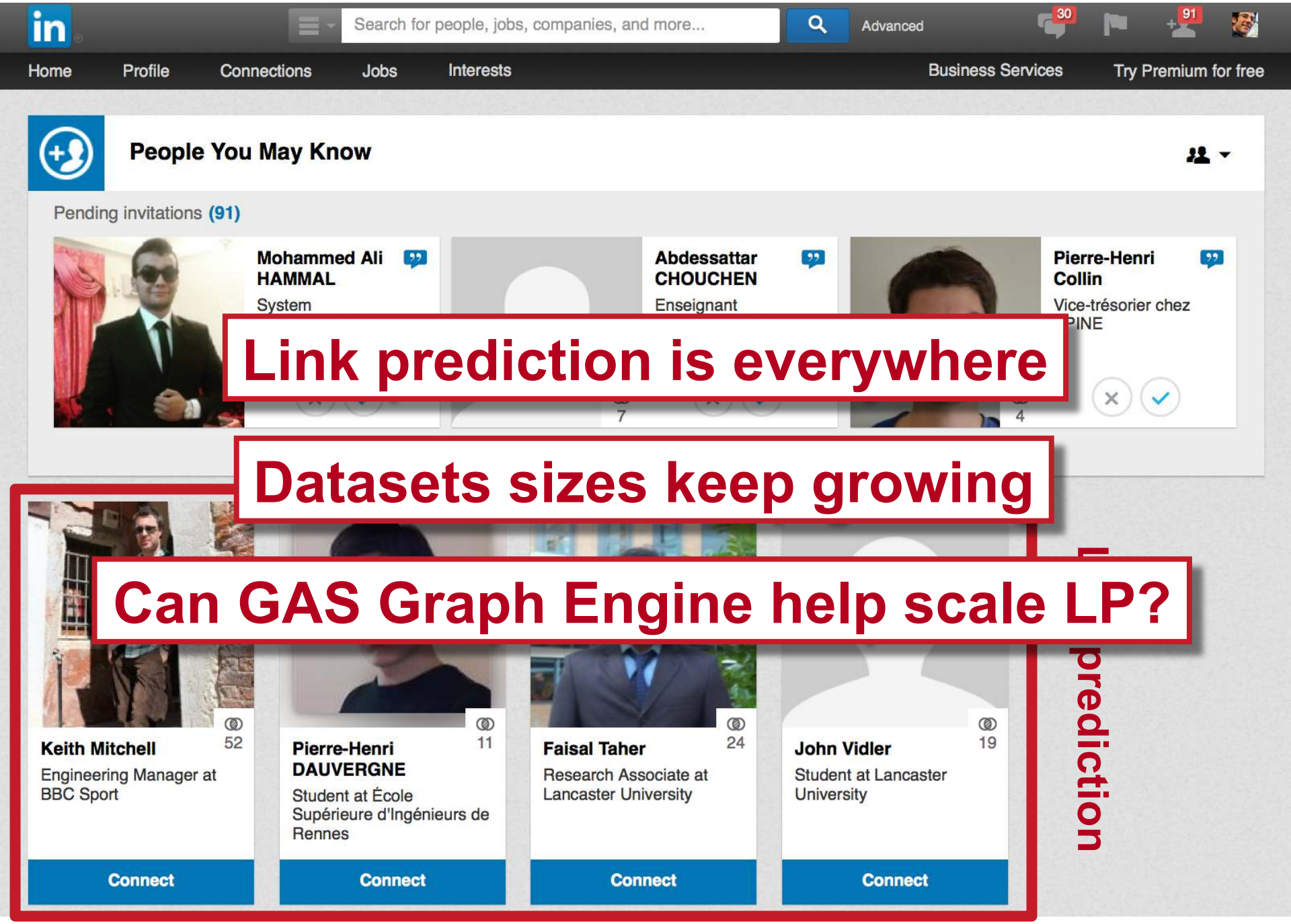
Le principe

180 secondes et un visuel pour présenter clairement, rapidement et simplement son travail de recherche, exercice qui vous sera utile dans votre futur environnement professionnel. Une occasion aussi pour vous de promouvoir la Recherche et le Doctorat auprès du grand public.

Ils nous soutiennent

Cet évènement est soutenu par les Universités de la région NPDC, des entreprises, des journalistes, les collectivités territoriales, des professionnels de la culture scientifique.

Calendrier



Link prediction is everywhere

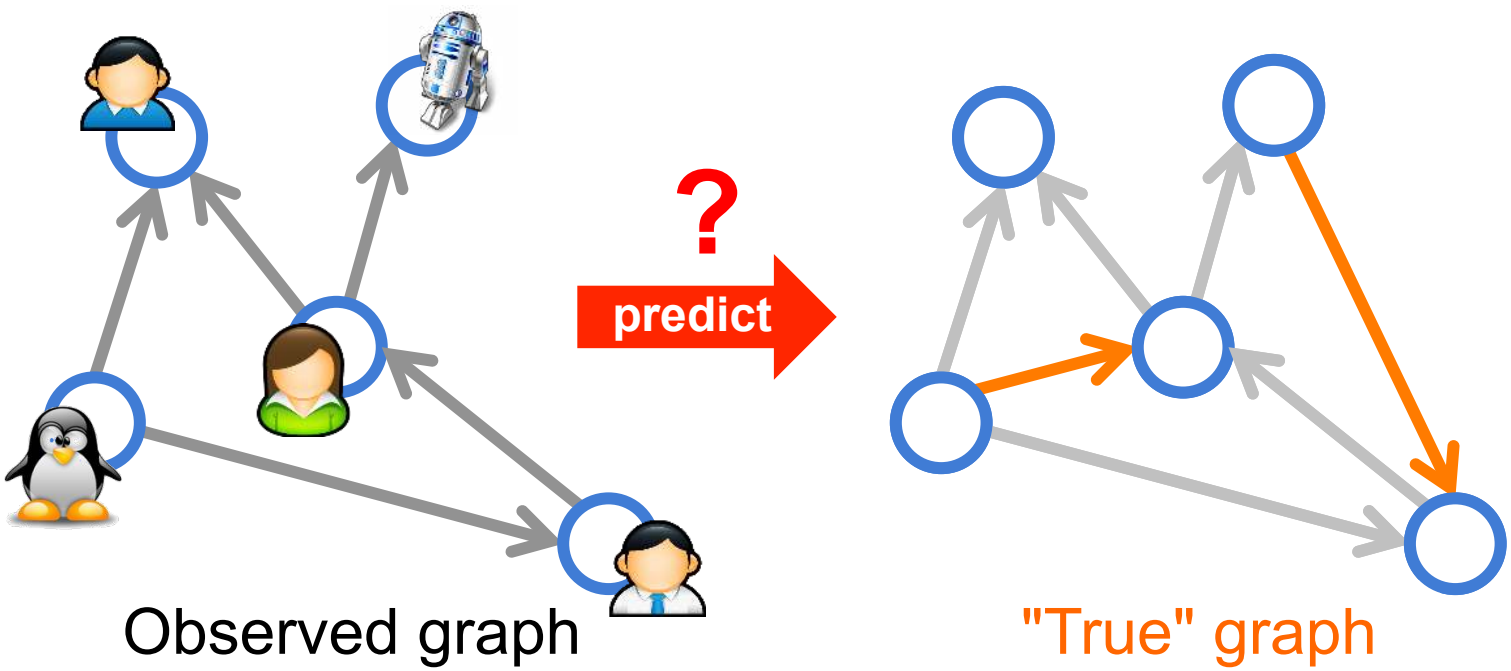
Datasets sizes keep growing

Can GAS Graph Engine help scale LP?

Link prediction

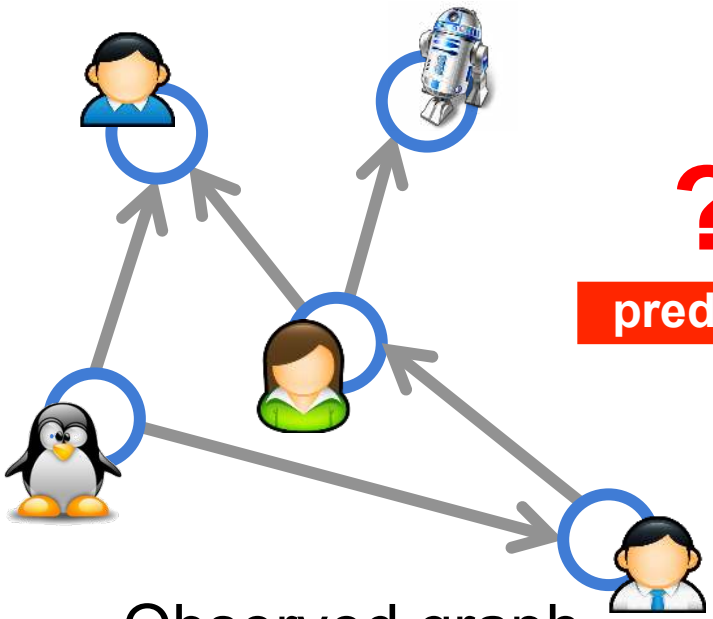
Link Prediction

1/3

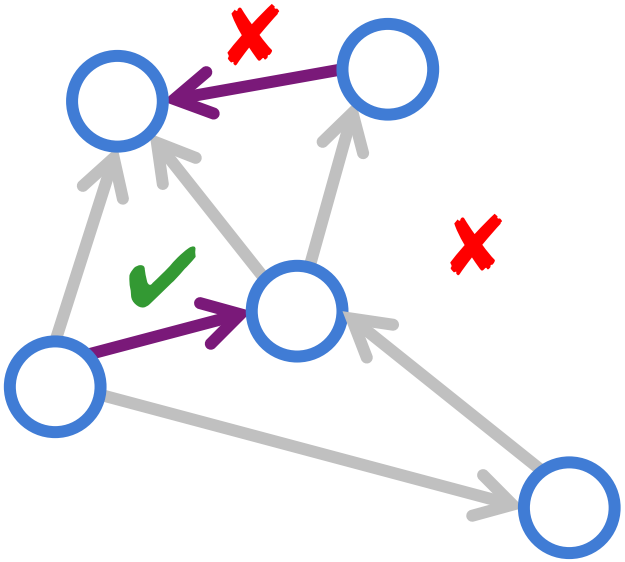


Link Prediction

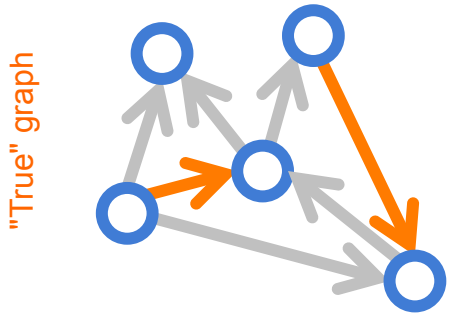
- Twin Goals
  - ➔ Good predictions (✔)
  - ➔ Fast ( )



Observed graph

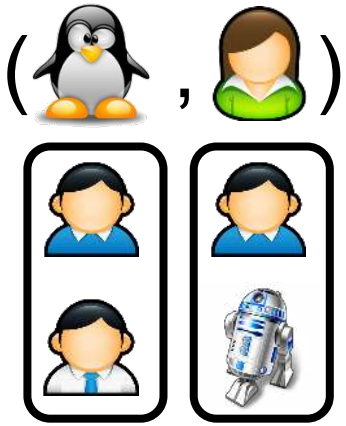


Predicted graph

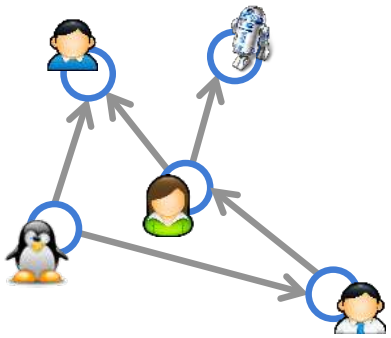


Link Prediction

- A classification problem
  - ➔ Does e = ( , ) belong to True Graph ?
  - ➔ Typical : Scoring exploiting topological information



→ 0.5

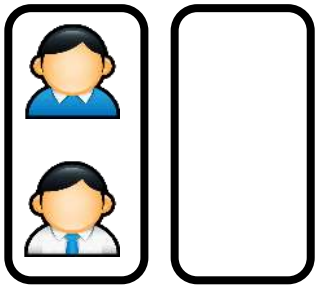


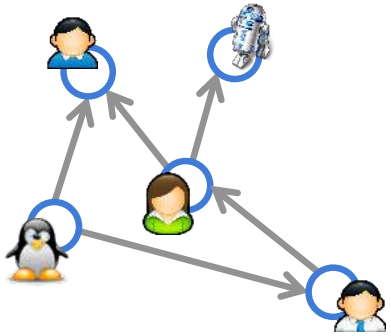
Link Prediction

- A classification problem
  - ➔ Does e = ( , ) belong to True Graph ?
  - ➔ Typical : Scoring exploiting topological information

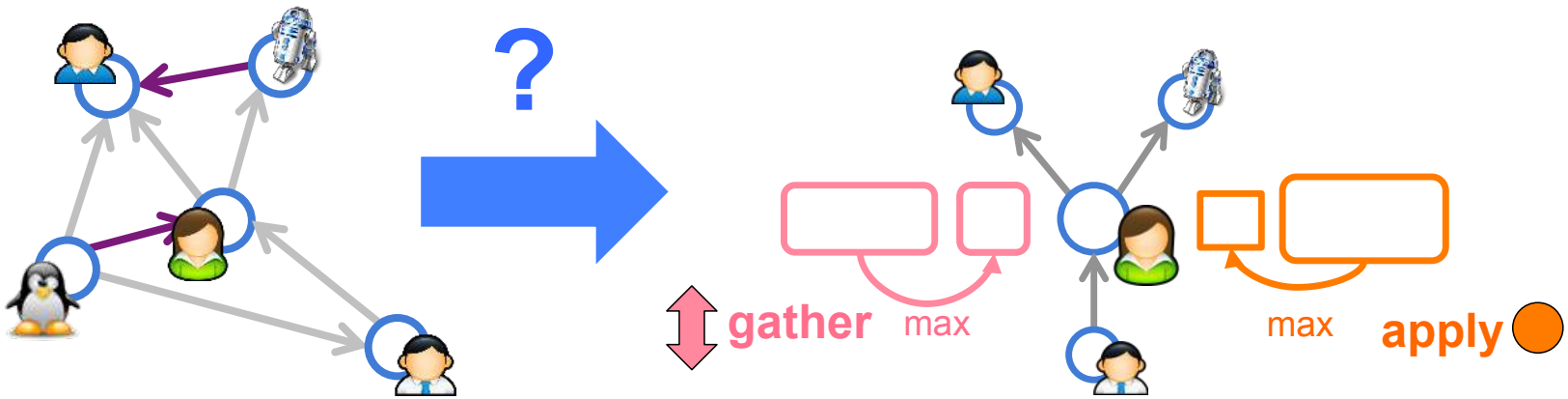
( , ) ⟶ 0.5 ✔

( , ) ⟶ 0.0 ✘



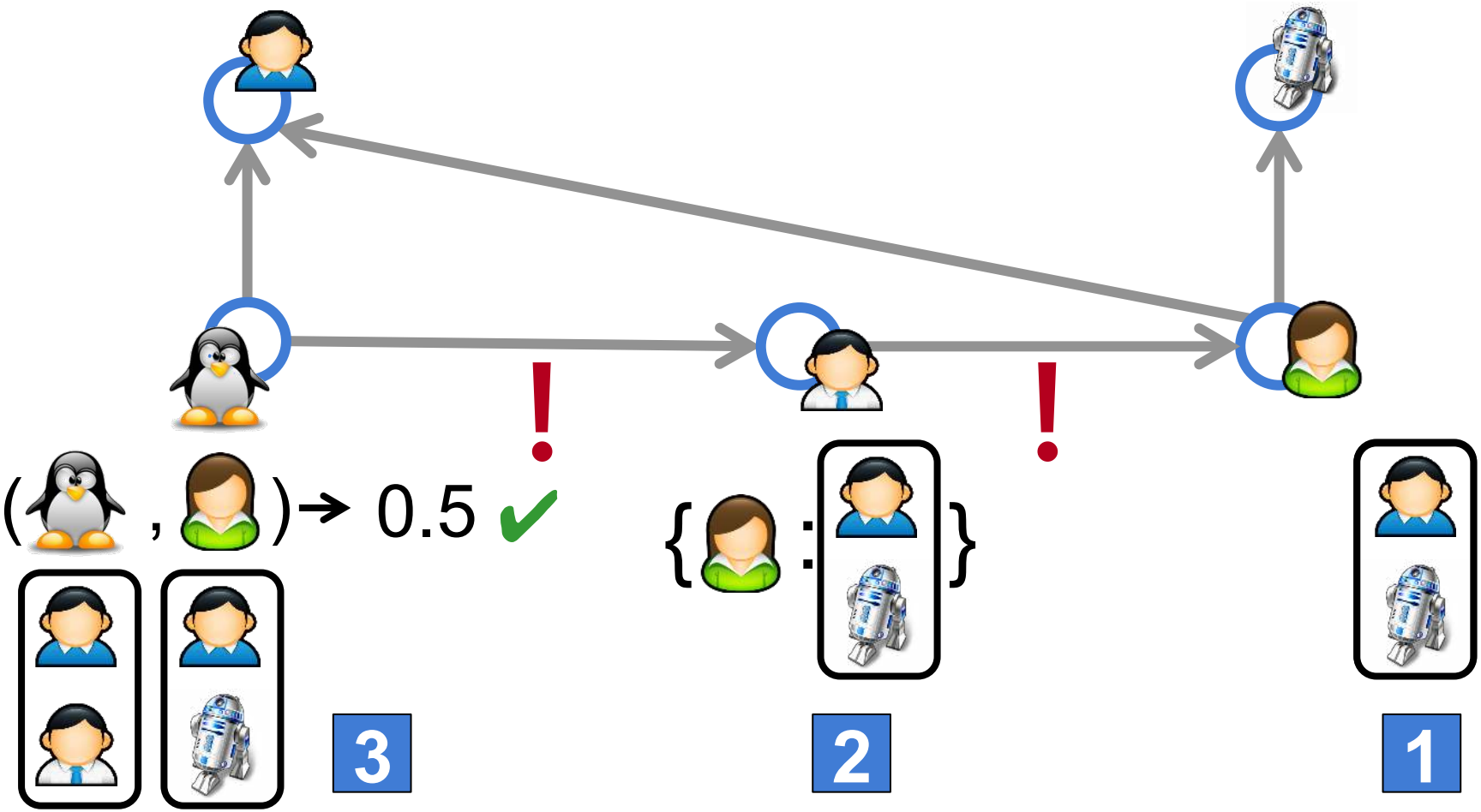


Porting Link Prediction to GAS

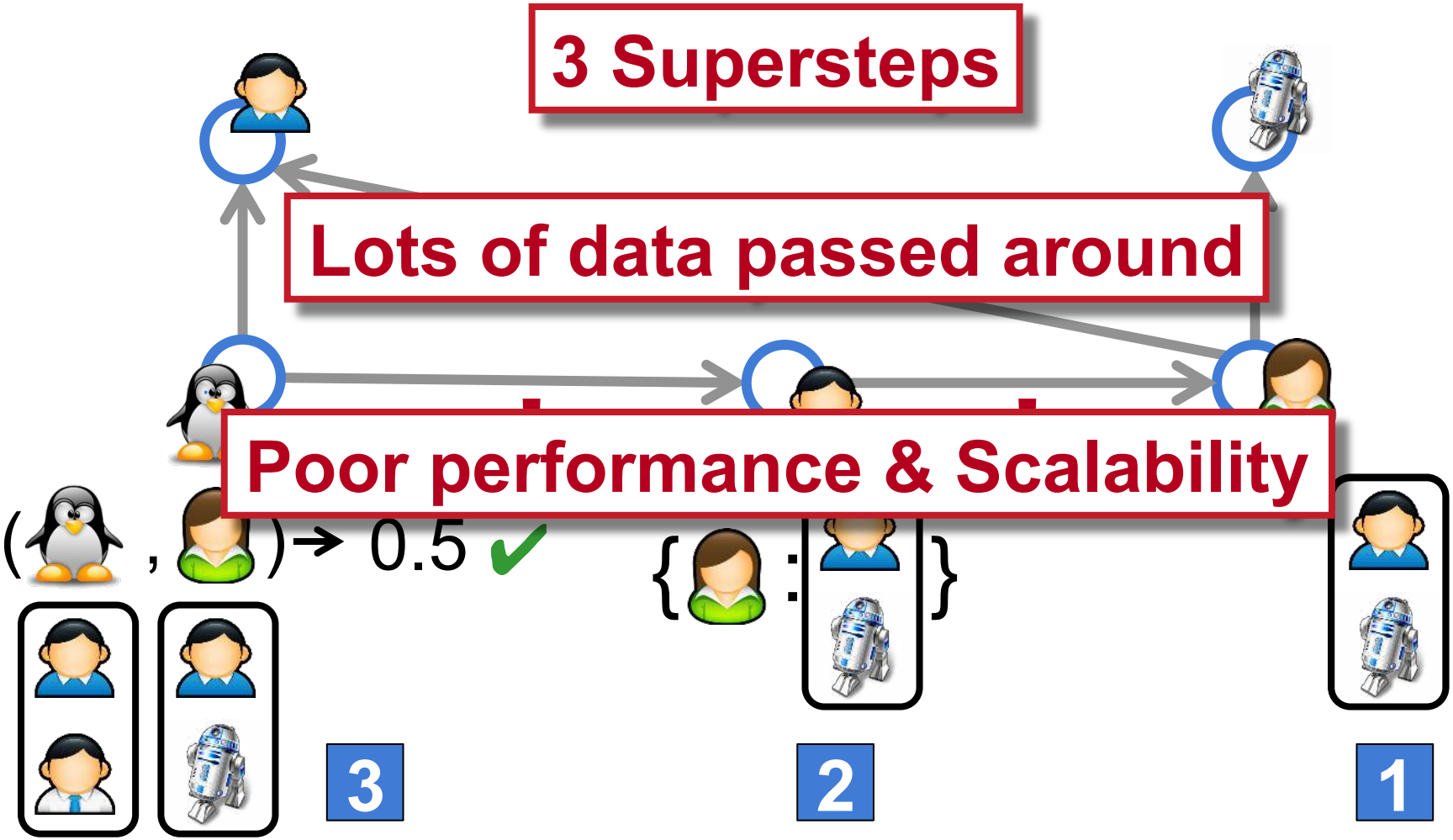


- **Challenge 1**: Random graph access impossible
  - ➔ Strategy: reduce search space to Friends-of-Friends
  - ➔ Avoid ~ N^2 score computation
- **Challenge 2**: Obtaining topological Information is costly
  - ➔ It must be propagated.

The Propagation Problem

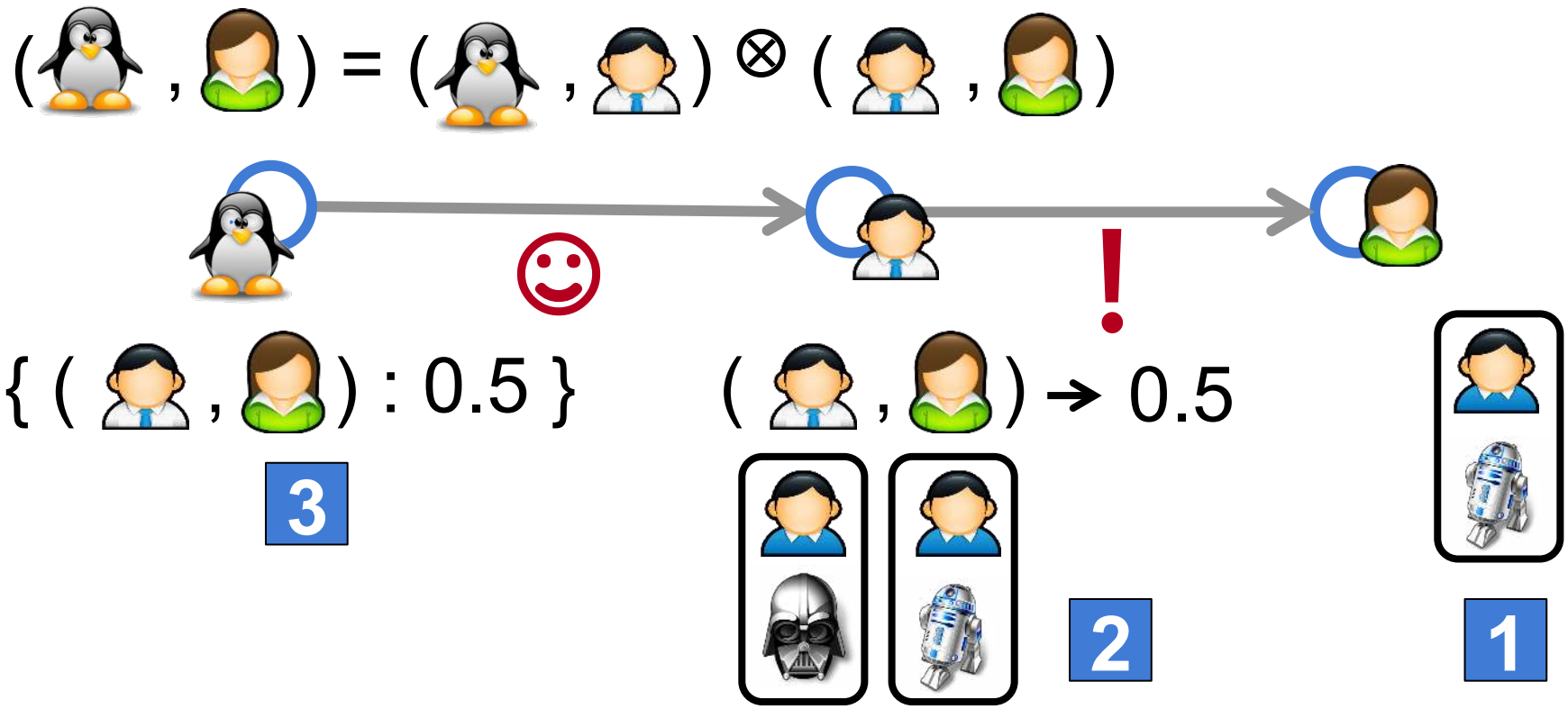


The Propagation Problem



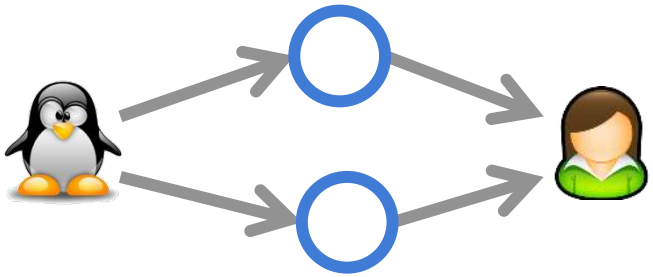
Snaple

- Intuition
  - ➔ **Propagate similarities**, not neighborhoods



Additional Complexities

- Which combination operator $\otimes$?
  - ➔ **Linear combination** is best
- Multiple paths
  - ➔ Path **aggregation**. **Sum** is best.



- Very large neighborhoods (e.g. up to 10^4 neighbors)
  - ➔ **Sampling** above threshold size **1**
  - ➔ **Selection** of best neighbors **2**

- Details are in the paper

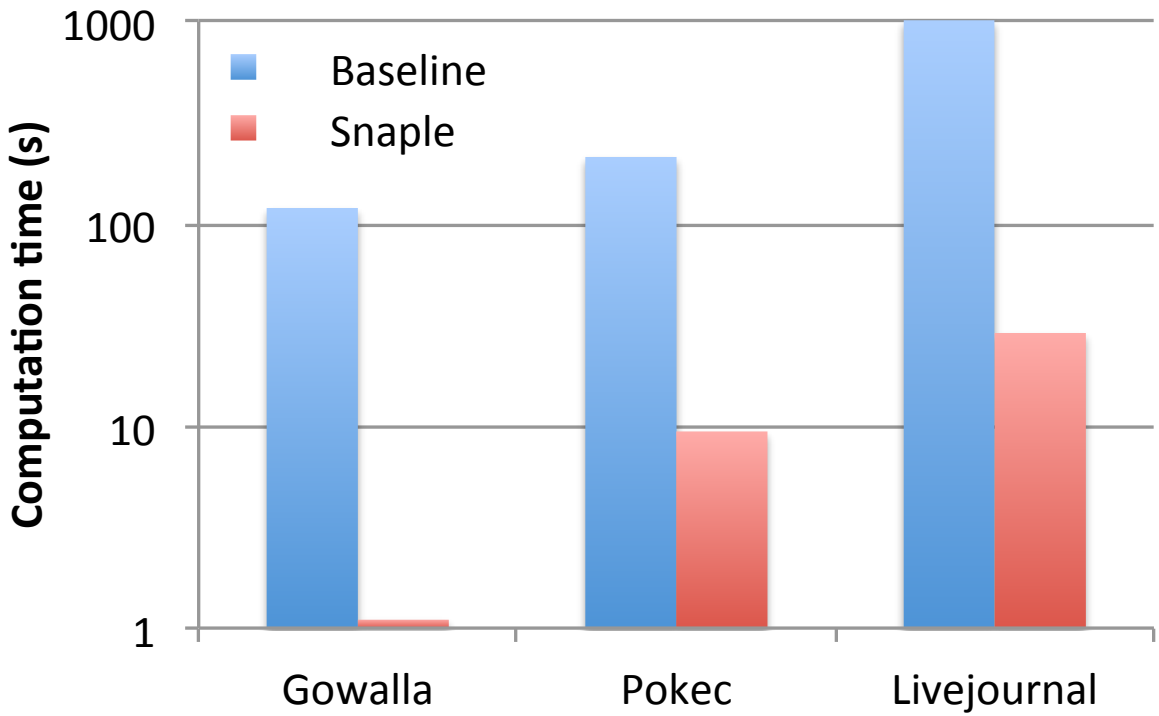
Evaluation

- **GraphLab** Implementation
- 5 datasets:
  - ➔ 1 outgoing edge removed per vertex

dataset	$\|V\|$	$\|E\|$	domain
gowalla [8]	196,591	0.95M	social network
pokec [40]	1.6M	30.6M	social network
orkut [28]	3M	223M	social network
livejournal [2]	4.8M	68.9M	co-authorship
twitter-rv [18]	41M	1.4B	microblogging

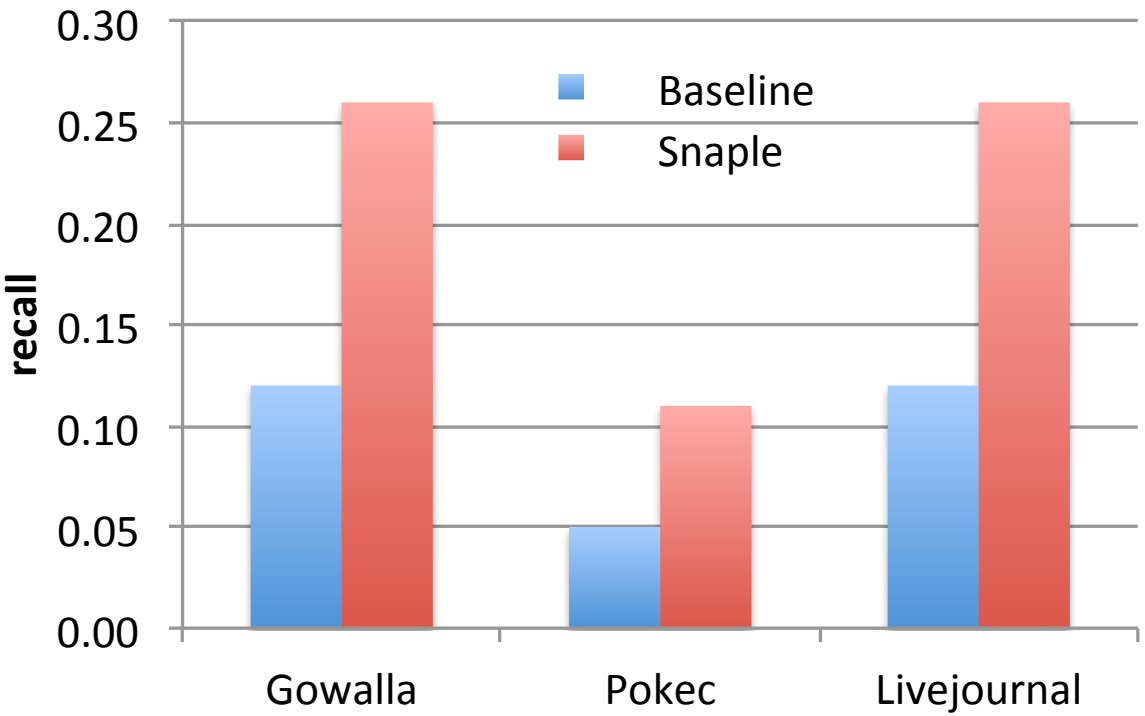
- Similarity: **Jaccard** (+ linear comb. and sum)
- Baseline: Direct implementation in GraphLab

Computation Time



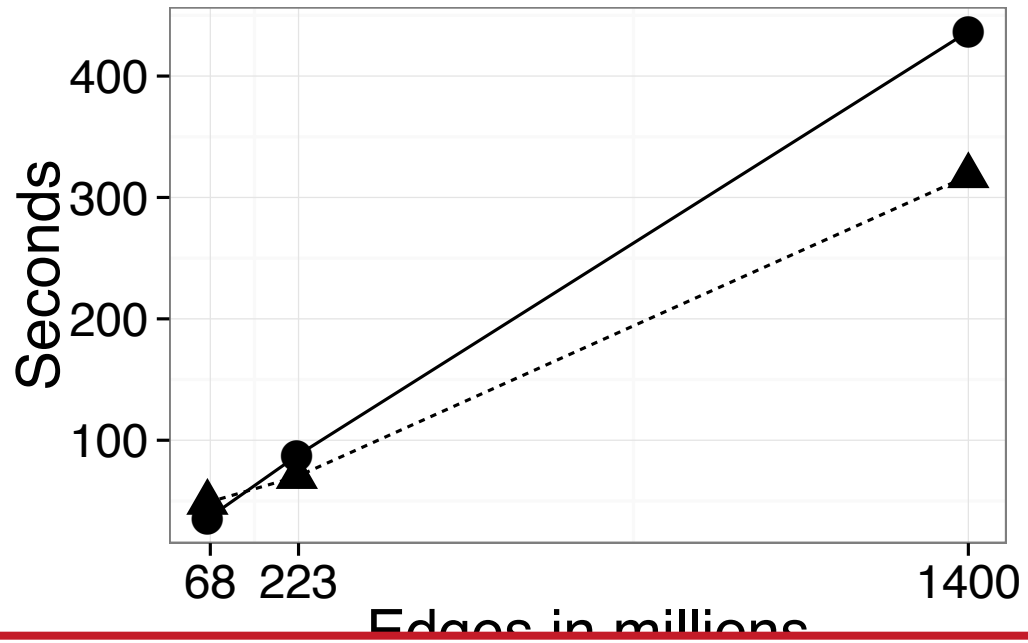
4 nodes, 80 **Average speedup: × 34.3** node (20 cores per node), 128 GBytes of memory, 10-Gigabit ethernet

Recall



Average recall gain: × 2.17

Scalability



1.4 billion edges processed in 2m57s on 160 cores

4 or 8 nodes, 2 Intel Xeon E5-2660v2 (2.2 Ghz) per node (20 cores per n

Scales linearly w.r.t. graph size

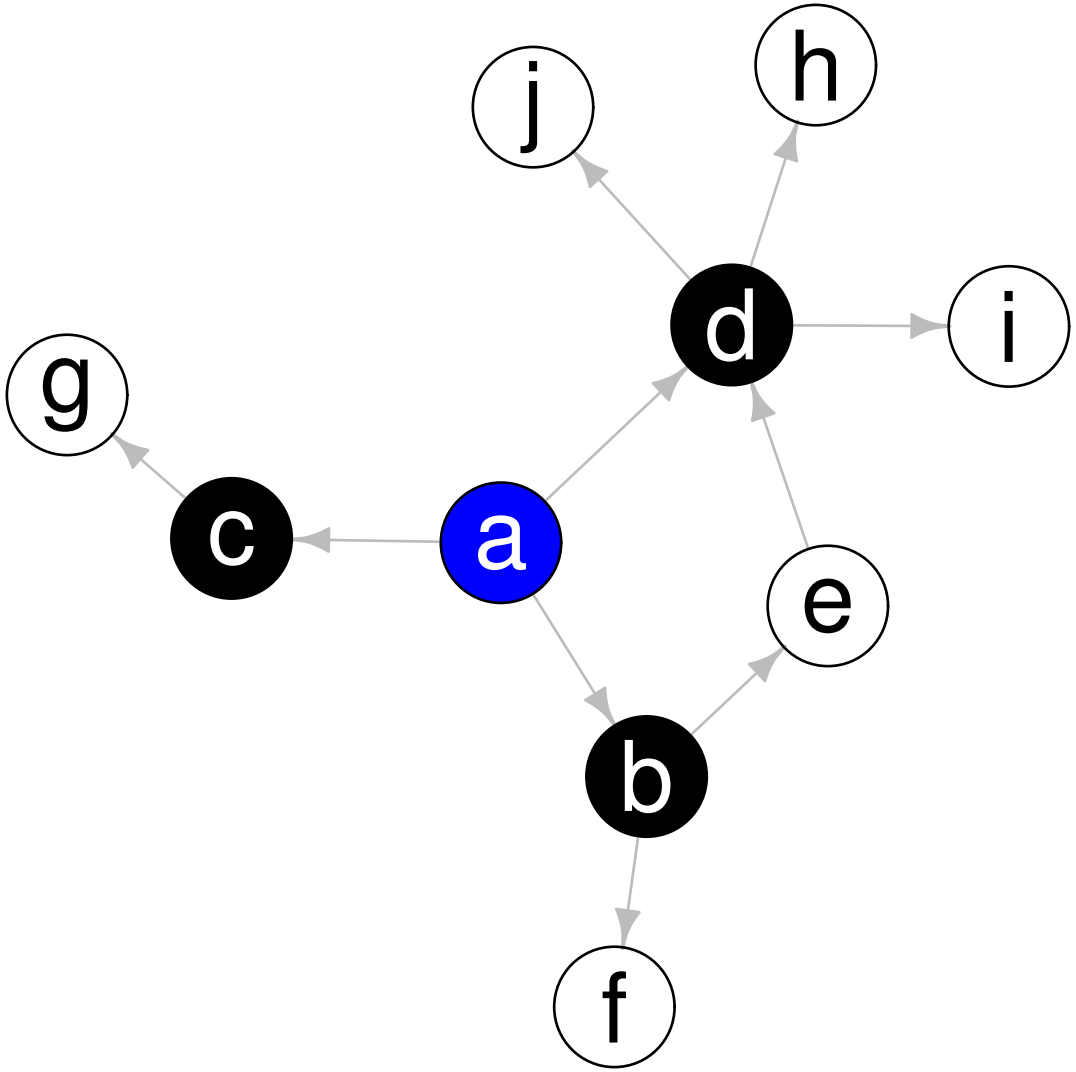
Also in the paper

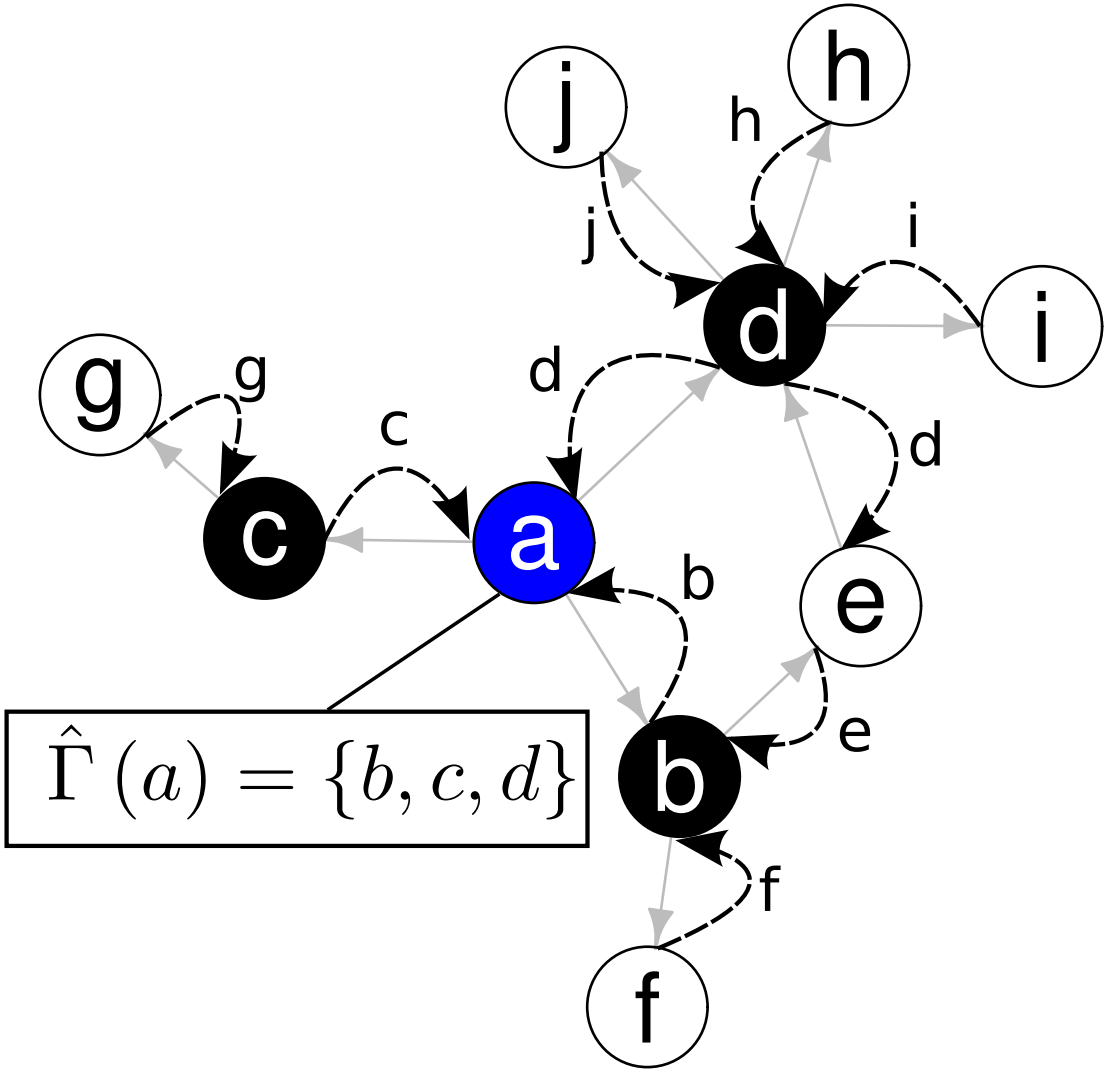
- More combinator / aggregator combinations
- Effect of sampling and selection
- Recall – speed tradeoff analysis
- Single-machine comparison with Cassovary

Conclusion

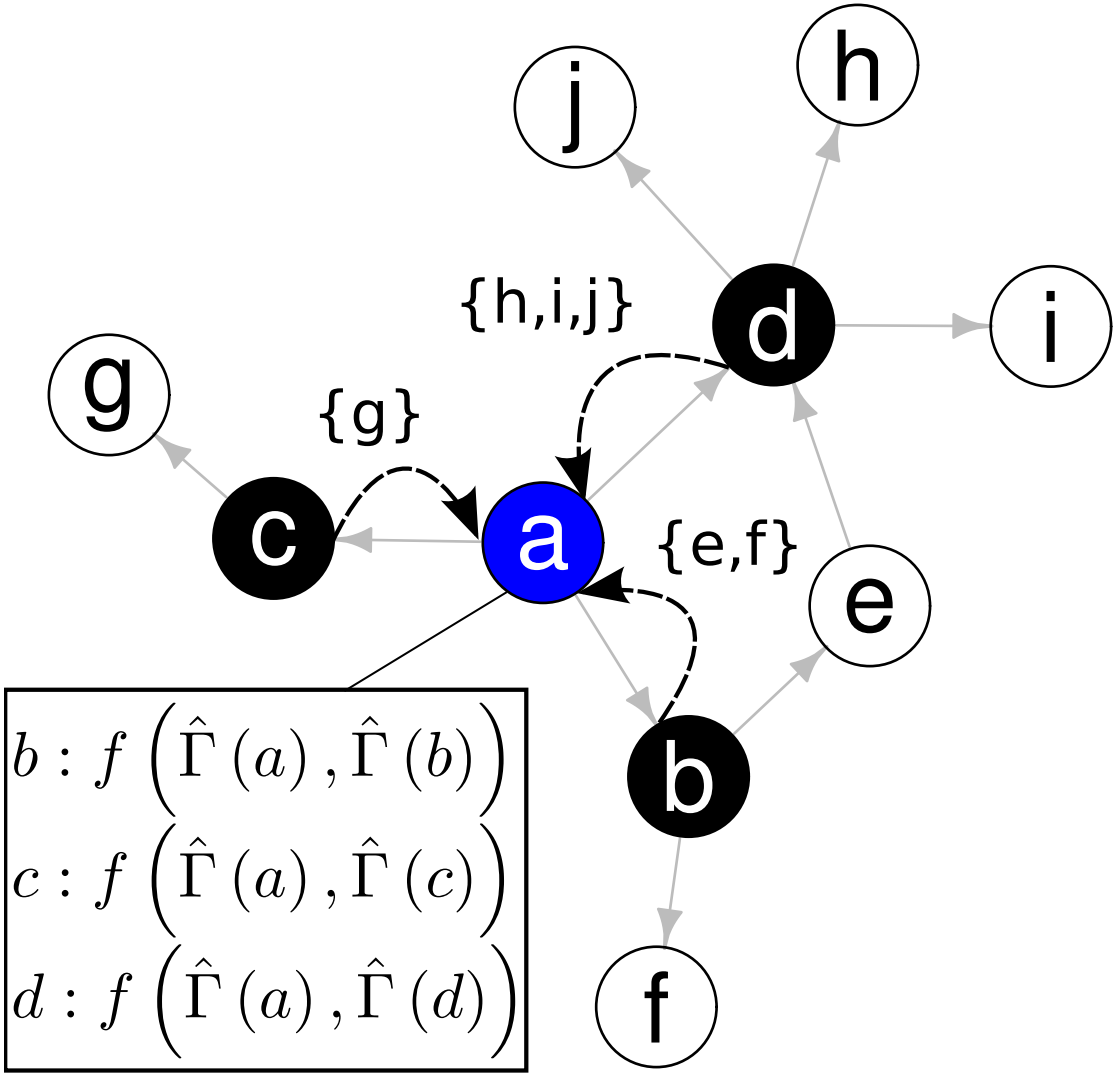
- Graph engines not well adapted to all algorithms
  - ➔ Substantial adaptation might be required
- Snaple: a framework for GAS-based link-prediction
- Large design space
  - ➔ We have only explored a small part
- Future work
  - ➔ Supervised similarities
  - ➔ Supervised combinator / aggregator design
- Ports to other GAS engines
  - ➔ Apache Giraph, Spark GraphX

Thank you
Questions?

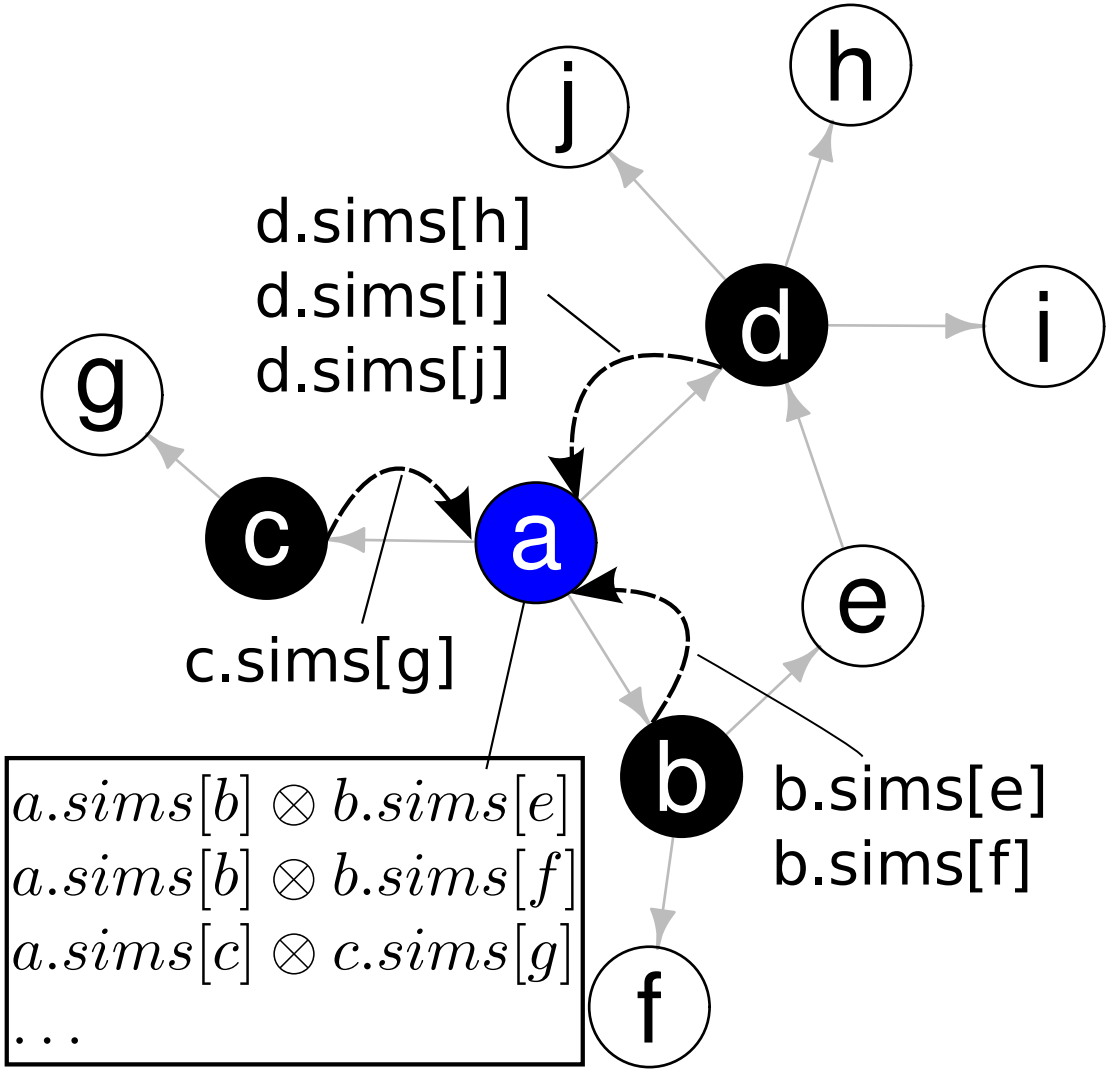

j
h
d
i
g
c
a
e
b
f

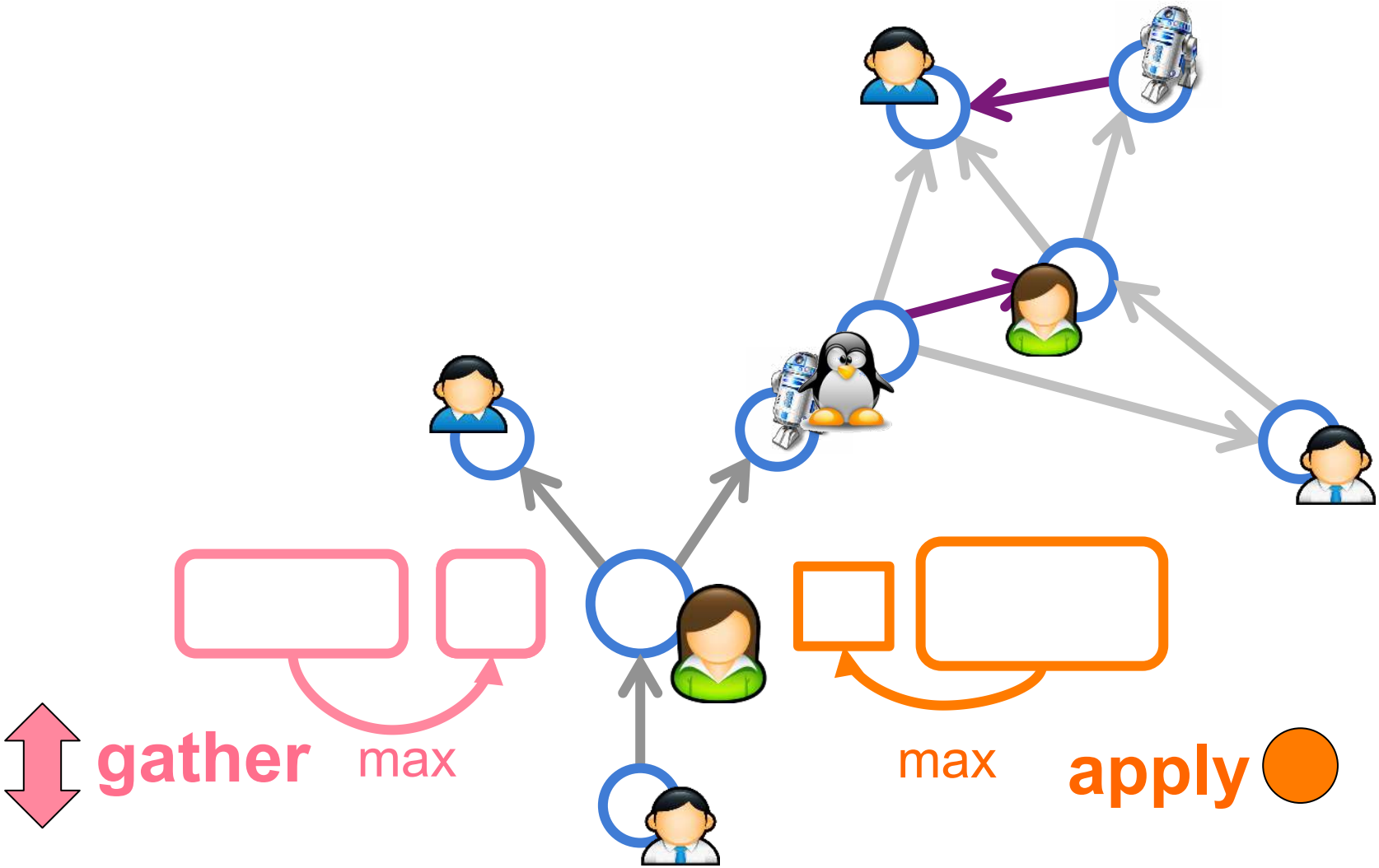


$$\hat{\Gamma}(a) = \{b, c, d\}$$



$$b : f\left(\hat{\Gamma}(a), \hat{\Gamma}(b)\right)$$
$$c : f\left(\hat{\Gamma}(a), \hat{\Gamma}(c)\right)$$
$$d : f\left(\hat{\Gamma}(a), \hat{\Gamma}(d)\right)$$


j
h
d.sims[h]
d.sims[i]
d.sims[j]
d
i
g
c
a
e
c.sims[g]
b
b.sims[e]
b.sims[f]
f
$a.sims[b] \otimes b.sims[e]$
$a.sims[b] \otimes b.sims[f]$
$a.sims[c] \otimes c.sims[g]$
$\ldots$


gather
max
max
apply