



SAPIENZA
UNIVERSITÀ DI ROMA

Similarity Ranking in Large-Scale Bipartite Graphs

Alessandro Epasto

Brown University - 20th March 2014



Google



THE PORT OF NEW YORK AUTHORITY

BANANA REPUBLIC

SAVE
40%

Joint work with J. Feldman, S. Lattanzi,
S. Leonardi, V. Mirrokni [WWW, 2014]

AdWords



[Web](#) [Images](#) [Shopping](#) [Maps](#) [Videos](#) [More ▾](#) [Search tools](#)

About 74,700,000 results (0.35 seconds)

Soccer Shoes at Zappos - Zappos.com

Ad www.zappos.com/Soccer ▾

Free Shipping, Free 365 Day Returns on **Soccer Shoes**

Zappos.com has 44,873 followers on Google+

Soccer Shoes - Soccer.com

Ad www.soccer.com/Cleats ▾ (800) 950-1994

Huge Selection of Top **Soccer Shoes**. \$3.99 Shipping On Orders Over \$99!

SOCCER.COM has 48,289 followers on Google+

Turf Soccer Shoes - \$3.99 Shipping on \$99+ - [Soccer Shoes-Free Returns](#)

Ads ⓘ

Indoor Soccer Shoes - World Soccer Shop

www.worldsoccershop.com ▸ [Departments](#) ▸ [Indoor](#) ▾

140+ items - Indoor **Soccer Shoes** - World Soccer Shop's selection of indoor ...

Puma evoPower 3 IT (Fluo Peach/Ombre Blue). \$74.99

adidas F10 IN Messi (Running White/Black). \$64.99

SOCCER.COM - Soccer Shoes, Soccer Jerseys, Soccer Ball...

www.soccer.com/ ▾

Shop for all your soccer equipment and apparel needs from EUROSPOORT including **soccer shoes**, replica soccer jerseys, soccer balls, team uniforms, ...

[Firm Ground Shoes](#) - [Footwear](#) - [Team](#) - [Indoor Shoes](#)

Pro-Direct Soccer US - Soccer Shoes, Soccer Cleats, Nike, ...

www.prodirectsoccer.com/us/soccer-shoes.aspx ▾

Pro-Direct Soccer US - The Professionals Choice for **Soccer Shoes**. Next Day Shipping on **Soccer Cleats** from adidas, Nike, Puma, Umbro, Under Armour ...





Amazon.com: Soccer - Athletic: Shoes

www.amazon.com ▸ [Shoes](#) ▸ [Men](#) ▸ [Athletic](#) ▾ Amazon.com ▾

Results 1 - 24 of 1737 - Online shopping for **Soccer** Athletic from a great selection of **Shoes**: Men & more at everyday low prices.

Shop for soccer shoes on Google

Sponsored ⓘ

			
Under Armour Men's UA Blu...	Green / Silver Mens Nike M...	Nike Jr Cr7 Mercurial Vict...	Adidas Soccer Cleats - Adid...
\$79.99	\$69.99	\$59.99	\$149.99
Under Armour	ShopWSS	eBay	Soccer.com

Shop by brand

[Nike](#) [adidas](#) [PUMA](#) [Diadora](#) [Under Armour](#)

Save on Soccer Shoes

www.soccersavings.com/Shoes ▾

Order Now - Low Flat Rate Shipping!

Great Selection of Top **Shoe** Brands.

Boys Soccer Cleat Sale

www.6pm.com/Boys-Soccer-Cleats ▾

4.6 ★★★★★ rating for 6pm.com

All Boys **Soccer Cleats** on Sale!

Save up to 70% and Free Shipping.

Soccer Cleats at Sears®

www.sears.com/Free-Store-Pickup ▾

Save Big on **Shoes** and Get them Today with Free Store Pickup!

Ads

Our Goal

- Tackling AdWords data to identify *automatically*, for each advertiser, its main competitors and suggest relevant queries to each advertiser.
- Goals:
 - Useful business information.
 - Improve advertisement.
 - More relevant performance benchmarks.

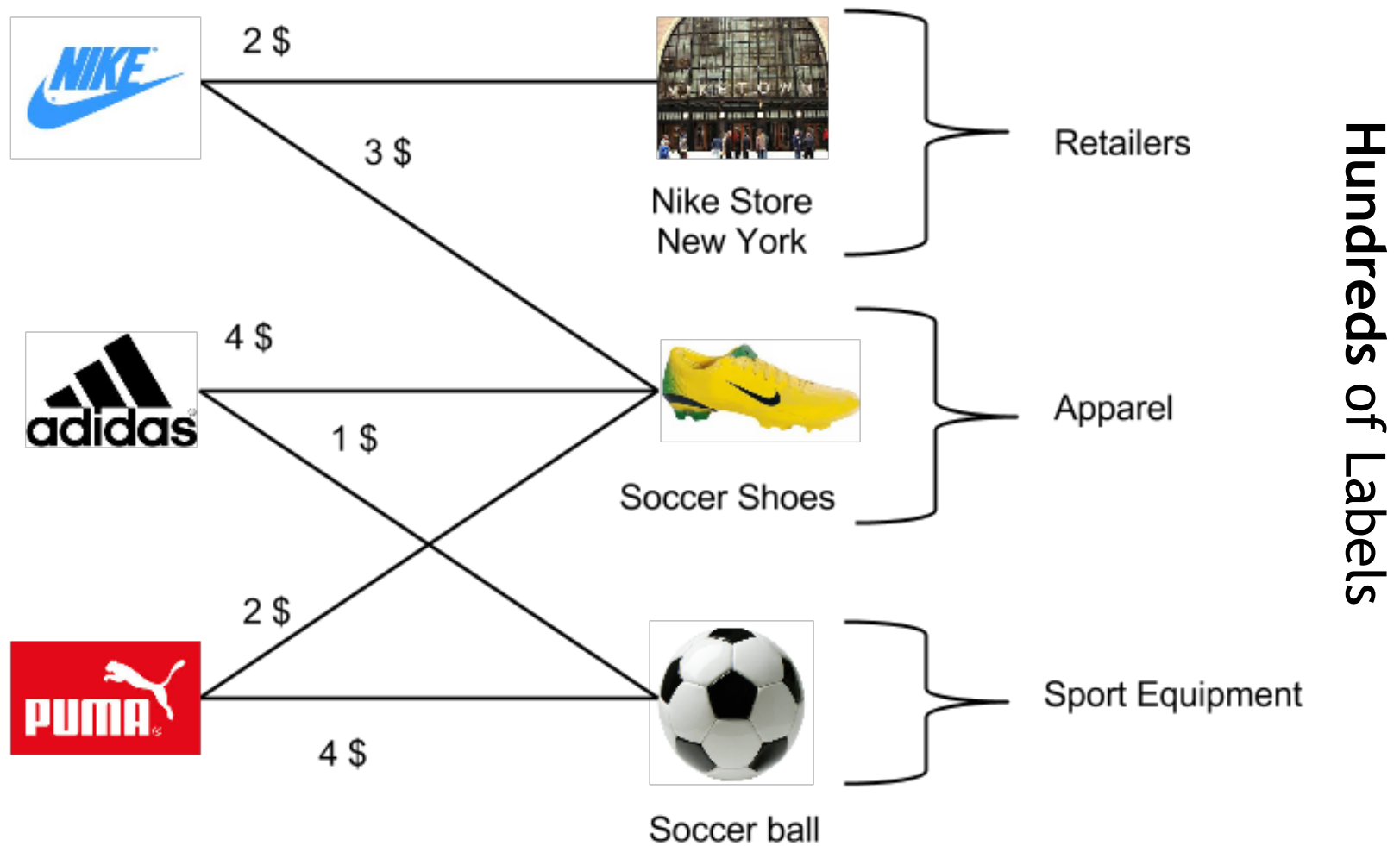
The Data

Query	Information
Nike store New York	Market Segment: Retailer , Geo: NY (USA), Stats: 10 clicks
Soccer shoes	Market Segment: Apparel , Geo: London, UK, Stats: 4 clicks
Soccer ball	Market Segment: Equipment Geo: San Francisco (USA), Stats: 5 clicks
.... millions of other queries	



Large advertisers (e.g., Amazon, Ask.com, etc) compete in several **market segments** with very different advertisers.

Modeling the Data as a Bipartite Graph



Millions of Advertisers

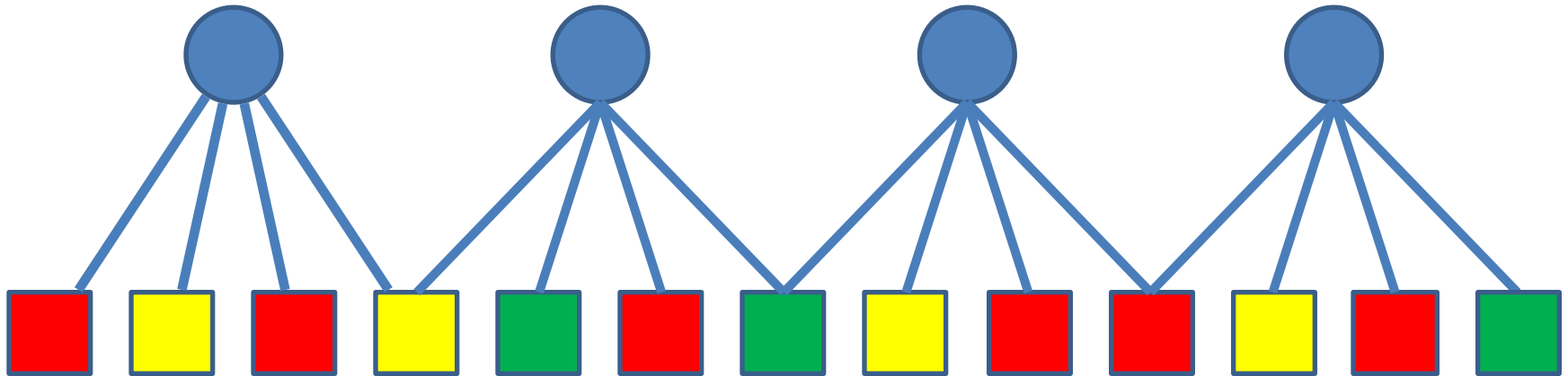
Billions of Queries

Other Applications

- General approach applicable to several contexts:
 - **User, Movies, Categories:** find similar users and suggest movies.
 - **Authors, Papers, Conferences:** find related authors and suggest papers to read.
- Generally this bipartite graphs are **lopsided**: we want algorithms with complexity depending on the smaller side.

Semi-Formal Problem Definition

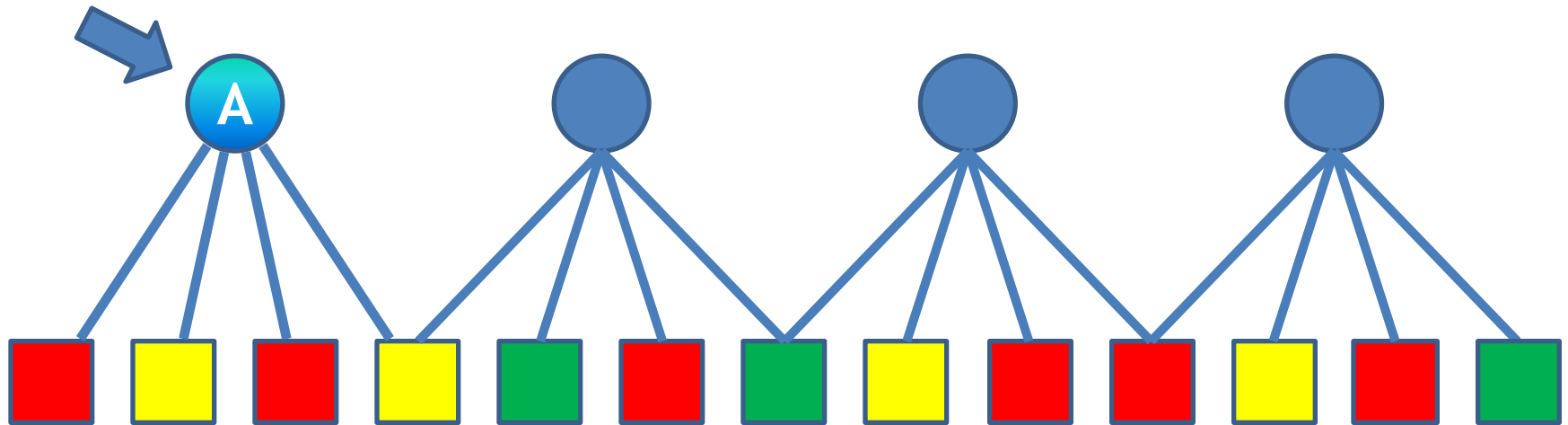
Advertisers



Queries

Semi-Formal Problem Definition

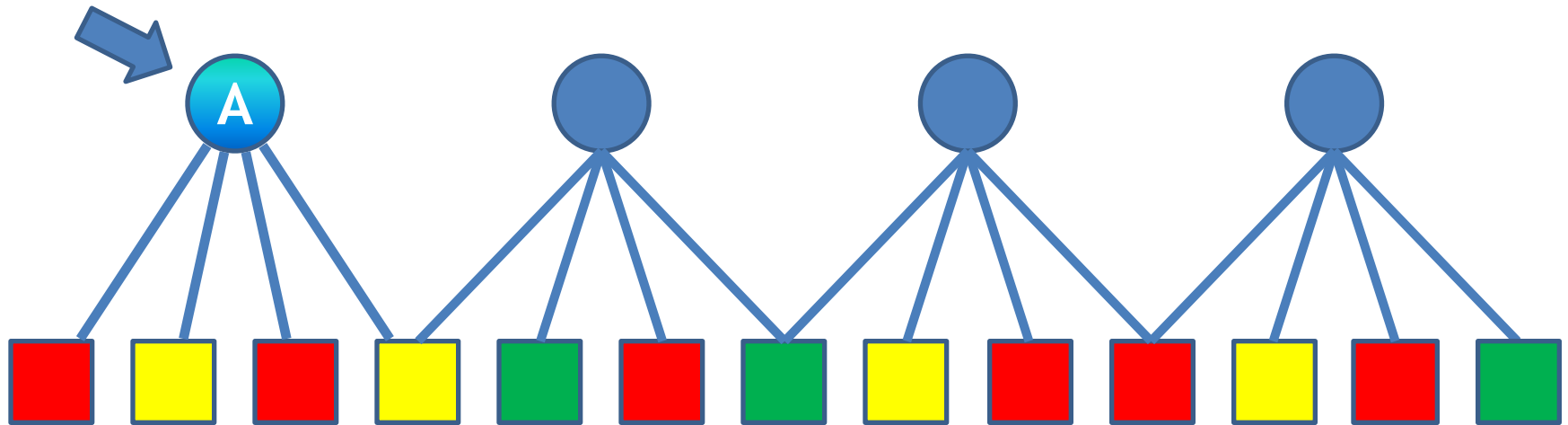
Advertisers



Queries

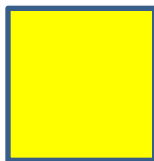
Semi-Formal Problem Definition

Advertisers



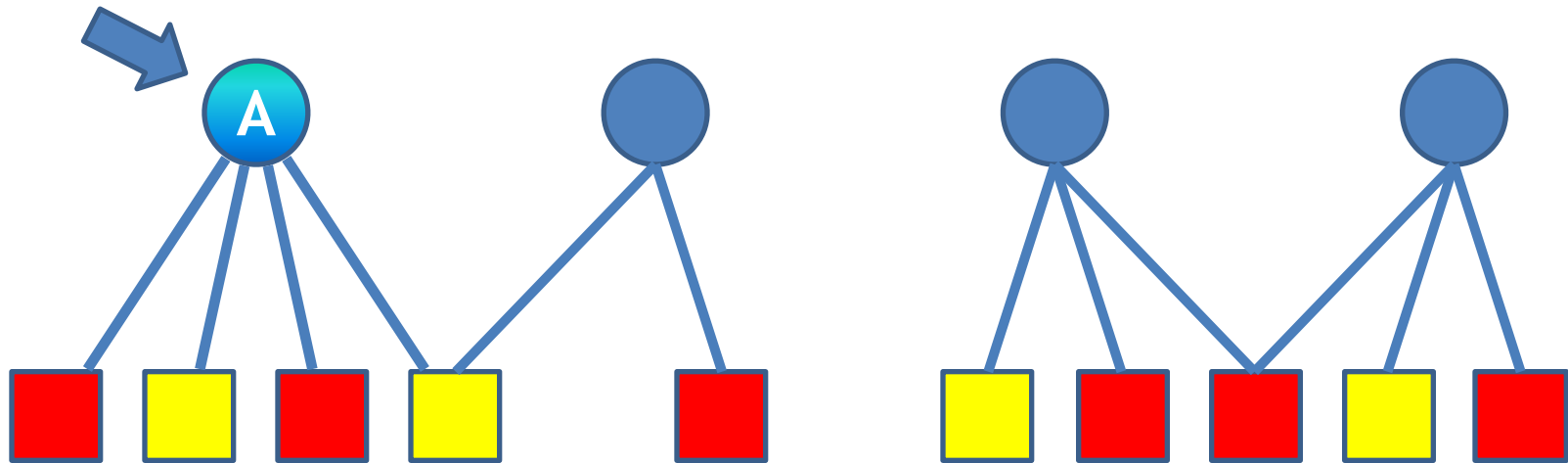
Queries

Labels:



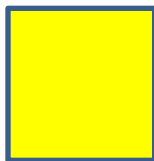
Semi-Formal Problem Definition

Advertisers



Queries

Labels:



Goal:
Find the nodes most
“similar” to A.

How to Define Similarity?

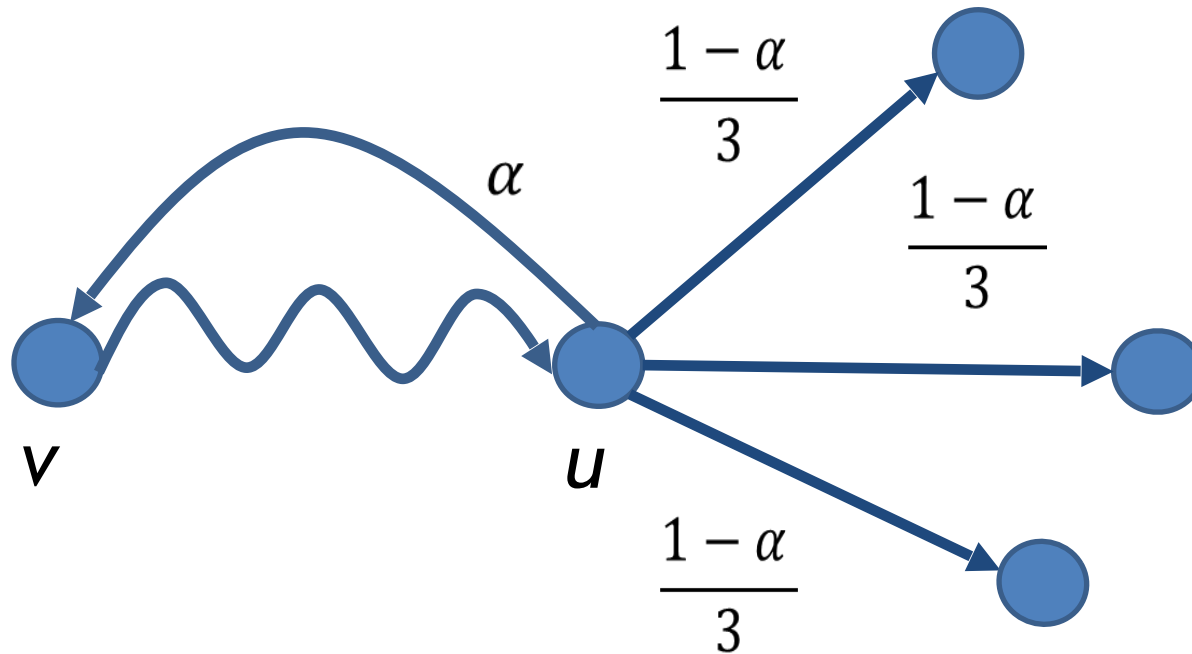
- We address the computation of several node similarity measures:
 - Neighborhood based: Common neighbors, Jaccard Coefficient, Adamic-Adar.
 - Paths based: Katz.
 - Random Walk based: Personalized PageRank.
- What is the accuracy?
- Can it scale to huge graphs?
- Can be computed in **real-time**?

Our Contribution

- **Reduce and Aggregate:** general approach to induce real-time similarity rankings in multi-categorical bipartite graphs, that we apply to several similarity measures.
- Theoretical guarantees for the precision of the algorithms.
- Experimental evaluation with real world data.

Personalized PageRank

For a node v (the seed) and a probability α



The stationary distribution assigns a similarity score to each node in the graph w.r.t. node v .

Personalized PageRank

- Extensive algorithmic literature.
- Very good accuracy in our experimental evaluation compared to other similarities (Jaccard, Intersection, etc.).
- Efficient MapReduce algorithm scaling to *large* graphs (hundred of millions of nodes).

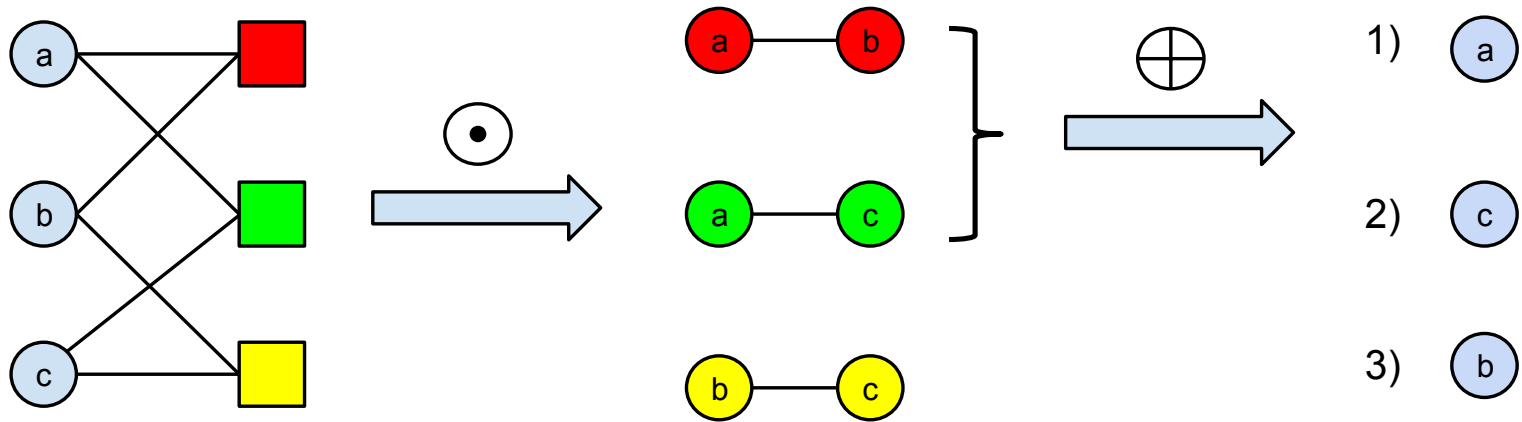
However...

Personalized PageRank



- Our graphs are too big (**billions** of nodes) even for large-scale systems.
- MapReduce is not real-time.
- We cannot pre-compute the rankings for each subset of labels.

Reduce and Aggregate



Reduce: Given the bipartite and a category construct a graph with only A nodes that preserves the ranking on the entire graph.

Aggregate: Given a node v in A and the reduced graphs of the subset of categories interested determine the ranking for v .

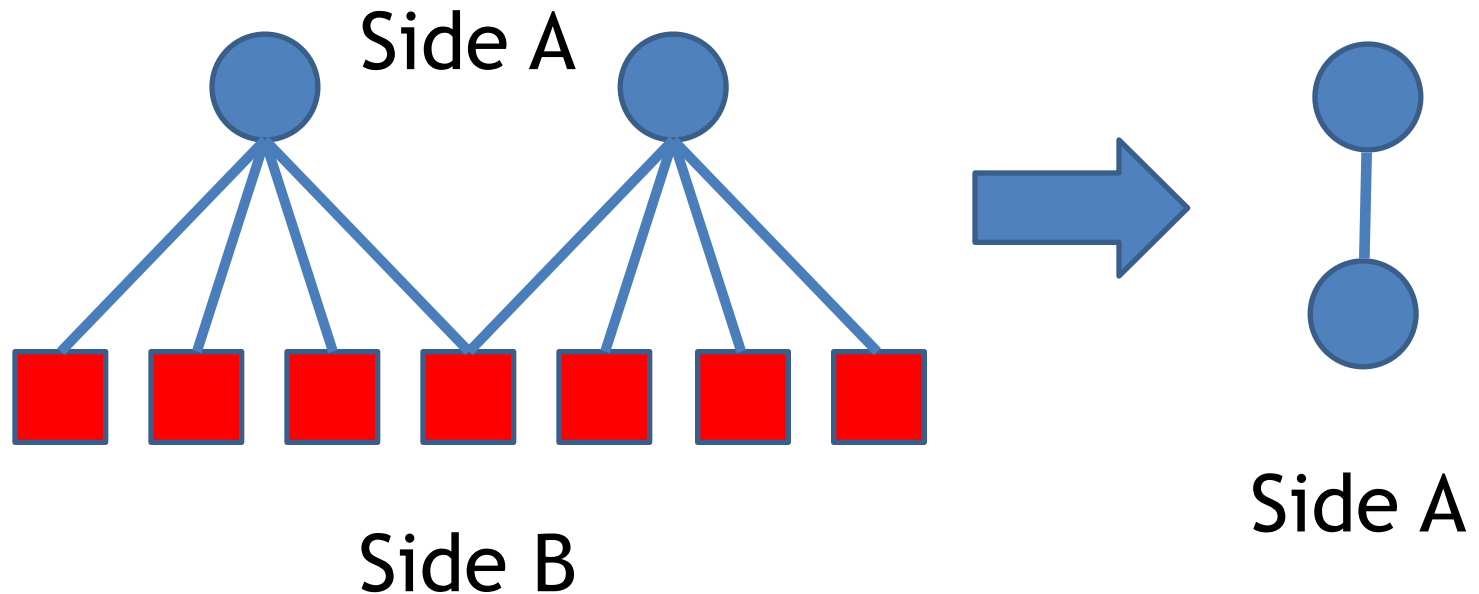
In practice



First stage: Large-scale (but feasible) MapReduce pre-computation of the individual category reduced graphs.

Second Stage: Fast real-time algorithm aggregation algorithm.

Reduce for Personalized PageRank



- Markov Chain state aggregation theory (Simon and Ado, '61; Meyer '89, etc.).
- **750x** reduction in the number of node while preserving correctly the PPR distribution on the **entire graph**.

Stochastic Complementation

$$\begin{vmatrix} P_{11} & \dots & P_{1i} & \dots & P_{1k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{i1} & \dots & P_{ii} & \dots & P_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{k1} & \dots & P_{ki} & \dots & P_{kk} \end{vmatrix}$$

- The stochastic complement of C_i is the following $|C_i| \times |C_i|$ matrix

$$S_i = P_{ii} + P_{i*}(1 - P_i^*)^{-1}P_{*i}$$

Stochastic Complementation

$$\begin{vmatrix} P_{11} & \dots & P_{1i} & \dots & P_{1k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{i1} & \dots & P_{ii} & \dots & P_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{k1} & \dots & P_{ki} & \dots & P_{kk} \end{vmatrix}$$

- The stochastic complement of C_i is the following $|C_i| \times |C_i|$ matrix

$$S_i = P_{ii} + P_{i*} (1 - P_i^*)^{-1} P_{*i}$$

Stochastic Complementation

$$\begin{bmatrix} P_{11} & \dots & P_{1i} & \dots & P_{1k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{i1} & \dots & P_{ii} & \dots & P_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{k1} & \dots & P_{ki} & \dots & P_{kk} \end{bmatrix}$$

- The stochastic complement of C_i is the following $|C_i| \times |C_i|$ matrix

$$S_i = P_{ii} + P_{i*}(1 - P_i^*)^{-1}P_{*i}$$

Stochastic Complementation

$$\begin{pmatrix} \boxed{P_{11} \quad \dots} & P_{1i} & \boxed{\dots \quad P_{1k}} \\ \vdots & \vdots & \vdots \\ P_{i1} \quad \dots & P_{ii} & \dots \quad P_{ik} \\ \vdots & \vdots & \vdots \\ \boxed{P_{k1} \quad \dots} & P_{ki} & \boxed{\dots \quad P_{kk}} \end{pmatrix}$$

- The stochastic complement of C_i is the following $|C_i| \times |C_i|$ matrix

$$S_i = P_{ii} + P_{i*} (1 - \boxed{P_i^*})^{-1} P_{*i}$$

Stochastic Complementation

$$\left| \begin{array}{cc|cc} P_{11} & \dots & P_{1i} & \dots & P_{1k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{i1} & \dots & P_{ii} & \dots & P_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{k1} & \dots & P_{ki} & \dots & P_{kk} \end{array} \right|$$

- The stochastic complement of C_i is the following $|C_i| \times |C_i|$ matrix

$$S_i = P_{ii} + P_{i*} (1 - P_i^*)^{-1} P_{*i}$$

Stochastic Complementation

Theorem [Meyer '89] For every irreducible aperiodic Markov Chain,

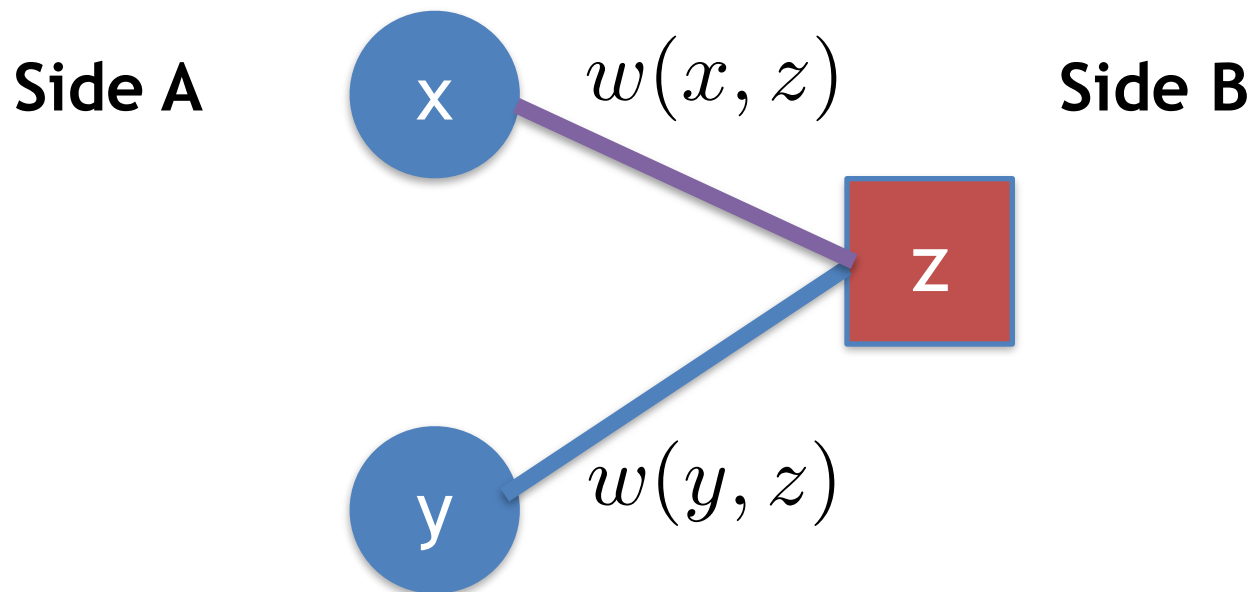
$$\pi_i = t_i s_i$$

where π_i is the stationary distribution of the nodes in C_i and s_i is the stationary distribution of S_i

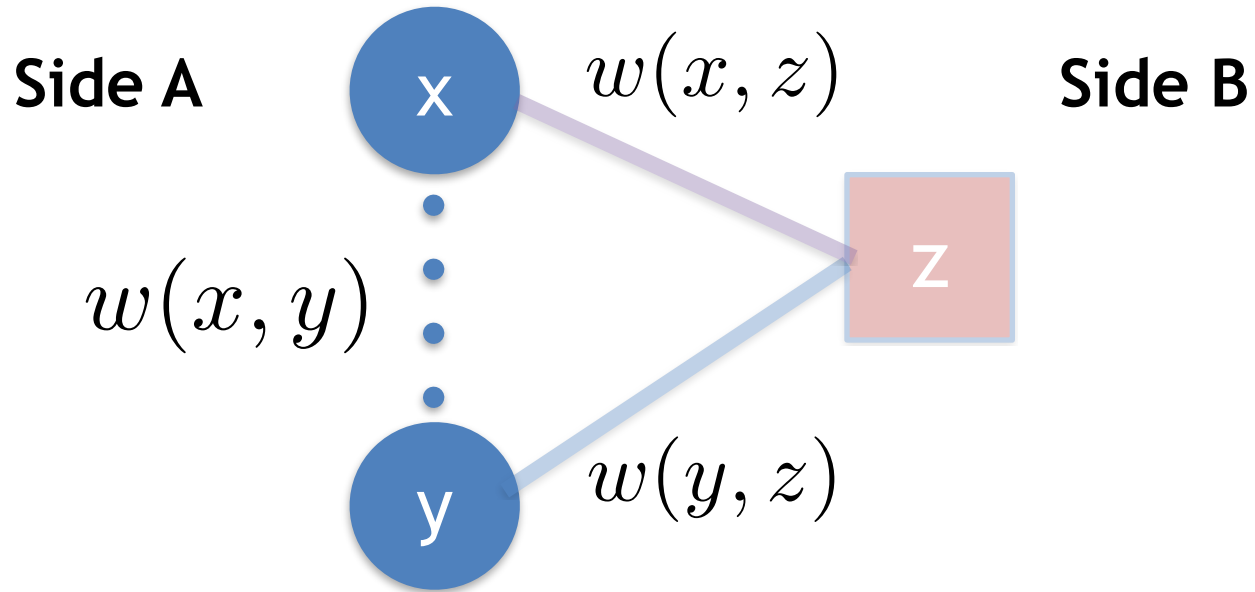
Stochastic Complementation

- Computing the stochastic complements is unfeasible in general for large matrices (matrix inversion).
- In our case we can exploit the properties of random walks on Bipartite graphs to invert the matrix analytically.

Reduce for PPR

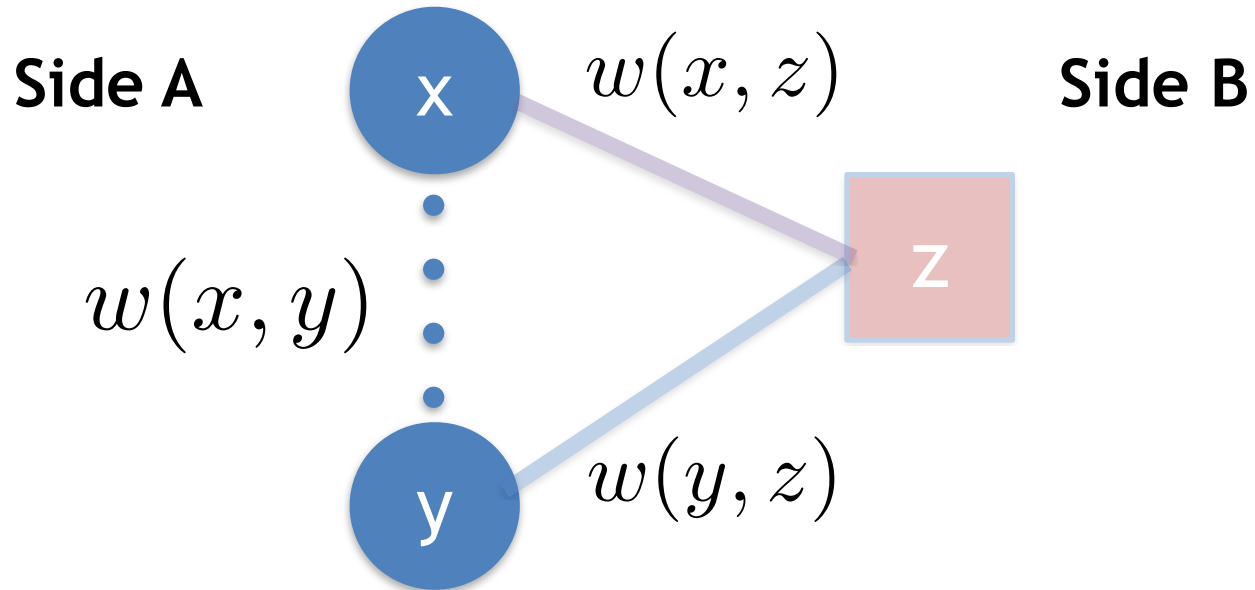


Reduce for PPR



$$w(x, y) = \sum_{z \in N(x) \cup N(y)} \frac{w(x, z)w(y, z)}{\sum_{h \in N(z)} w(z, h)}$$

Reduce for PPR



One step in the reduced graph is equivalent to **two steps** in the bipartite graph.

Properties of the Reduced Graph

Lemma 1: $\text{PPR}(G, \alpha, a)[A] = \frac{1}{2-\alpha} \text{PPR}(\hat{G}, 2\alpha - \alpha^2, a)$

Proof Sketch:

- Every path between nodes in A is even.
- Probability of not jumping for two steps.
- The probability of being in the A-Side at stationarity does not depend on the graph.

Properties of the Reduced Graph

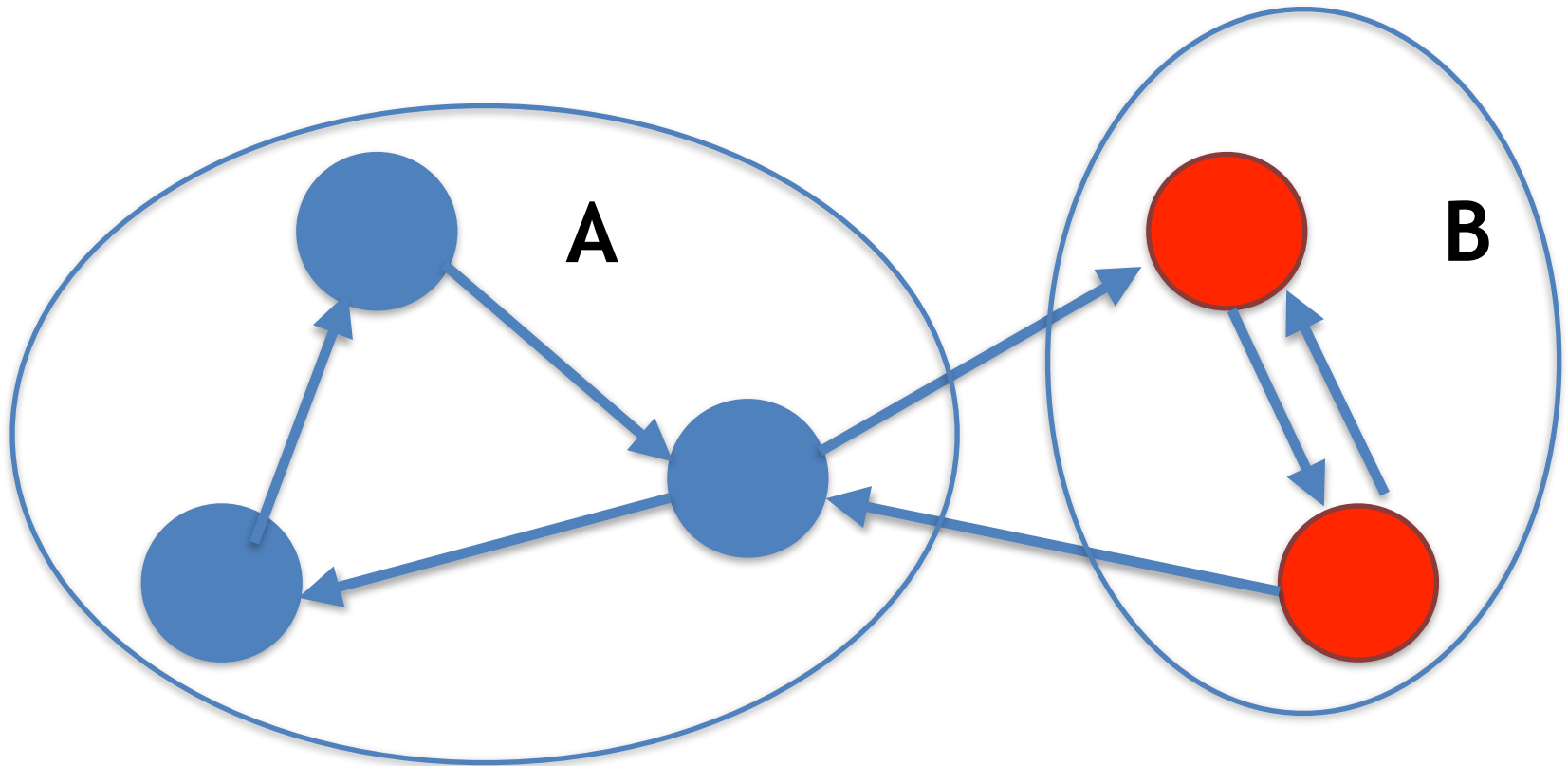
Similarly, we can reduce the process to a graph with B-Side nodes only.

Lemma 2:

$$\text{PPR}(G, \alpha, a)[B] = \frac{1-\alpha}{2-\alpha} \sum_{b \in N(a)} w(a, b) \text{PPR}(\hat{G}_B, 2\alpha - \alpha^2, b)$$

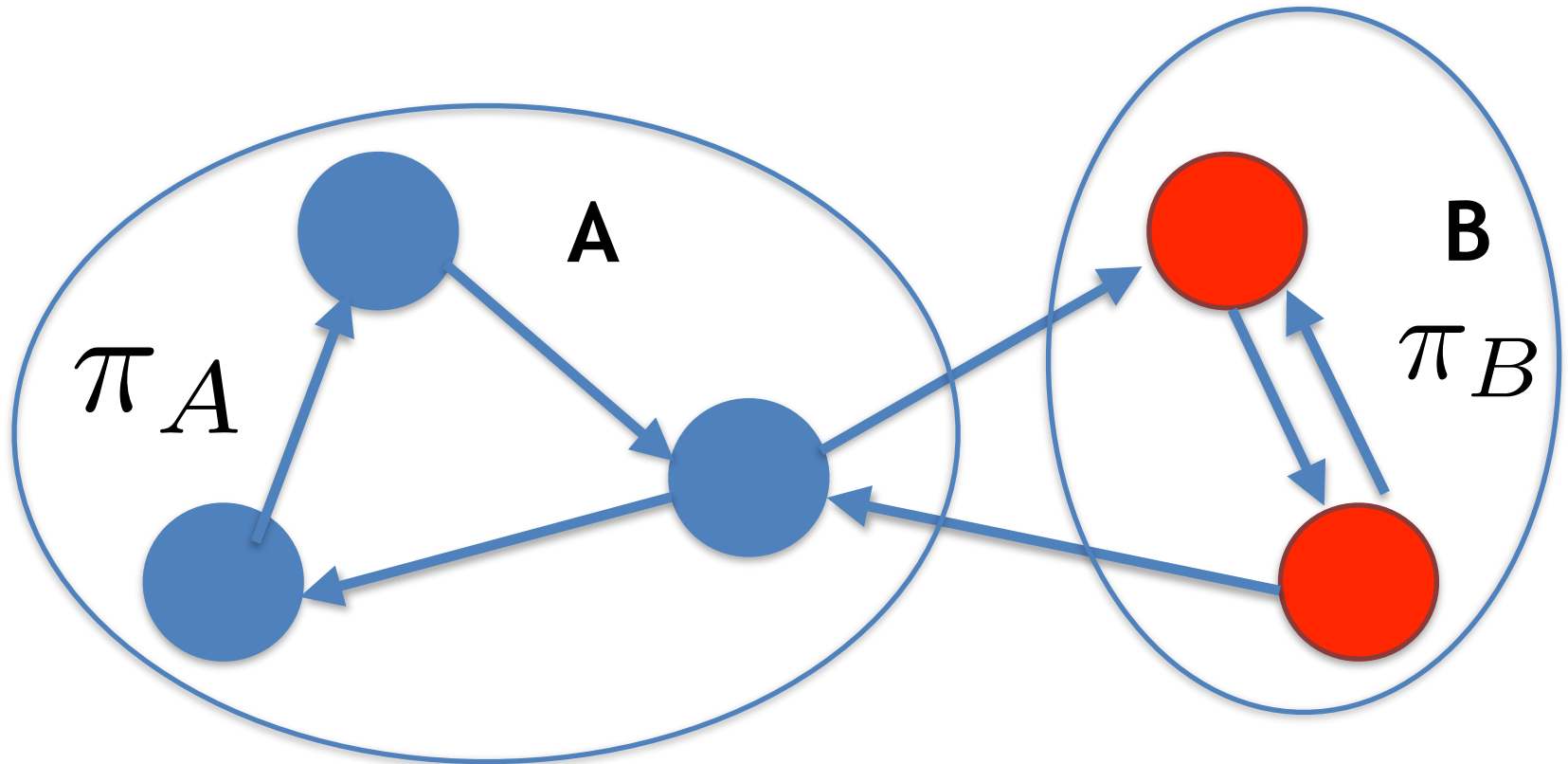
Finally, the stationary distribution of either side **uniquely determines** that of the other side.

Koury et al. Aggregation-Disaggregation Algorithm



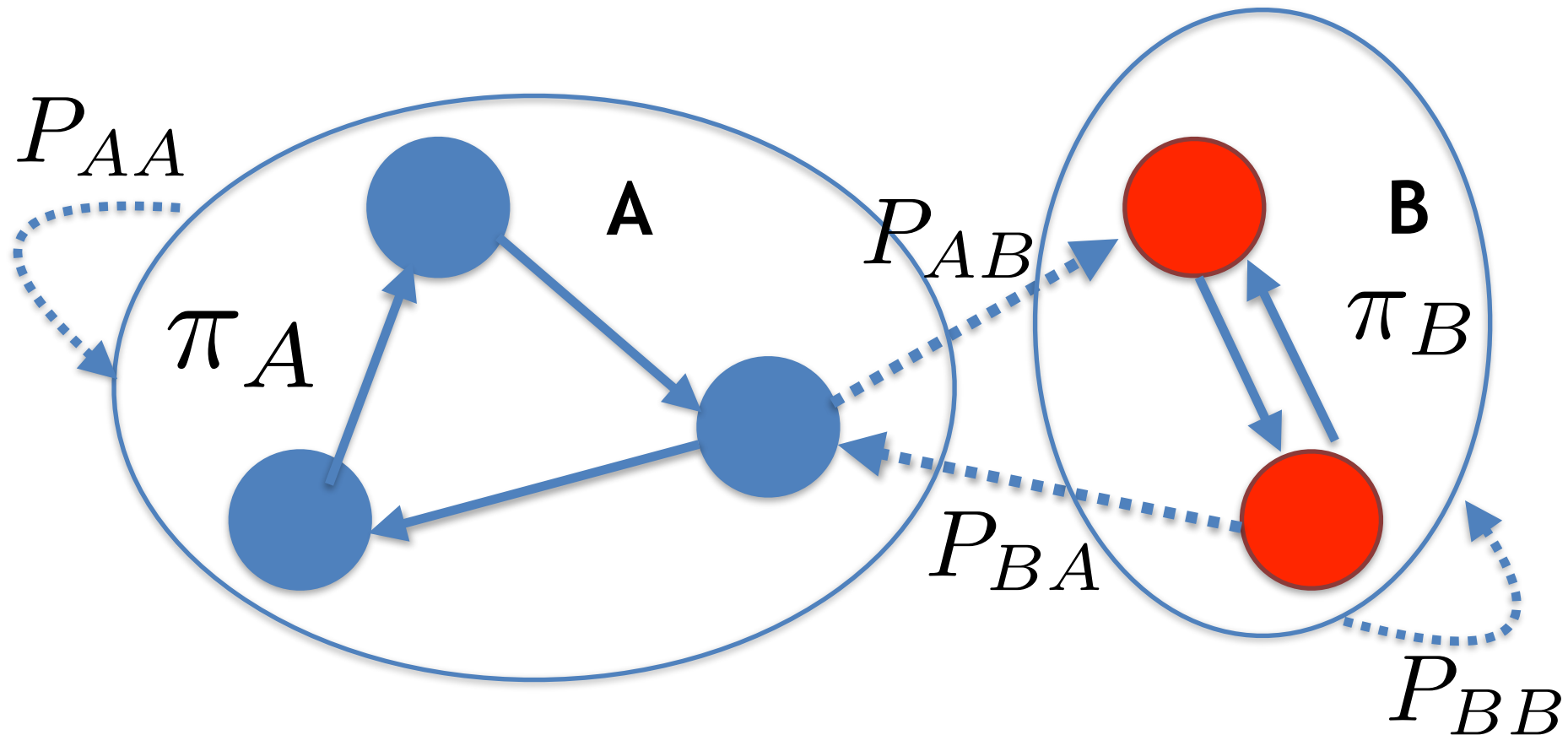
Step 1: Partition the Markov chain into disjoint subsets

Koury et al. Aggregation-Disaggregation Algorithm



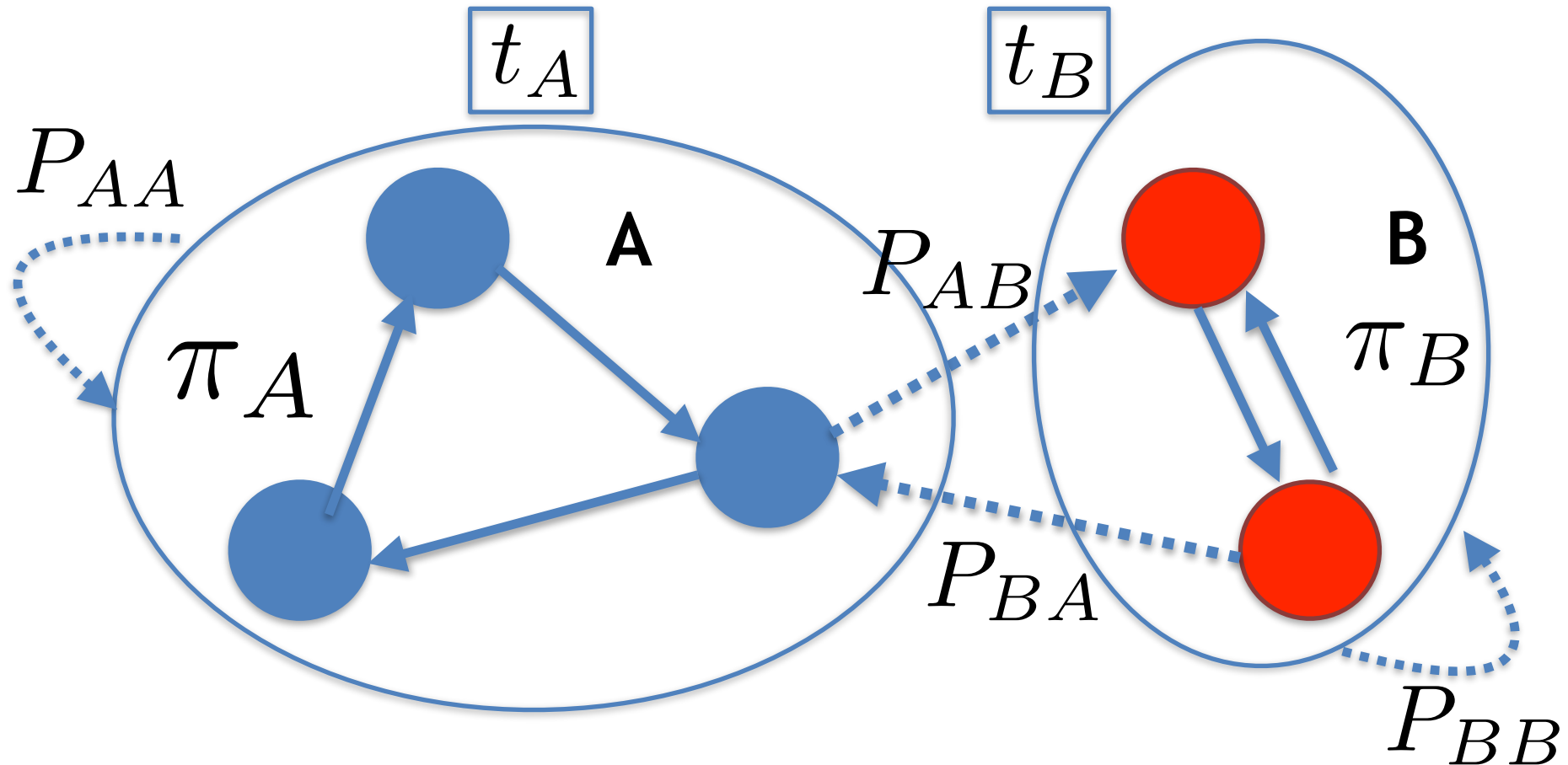
Step 2: Approximate the stationary distribution on each subset independently.

Koury et al. Aggregation-Disaggregation Algorithm



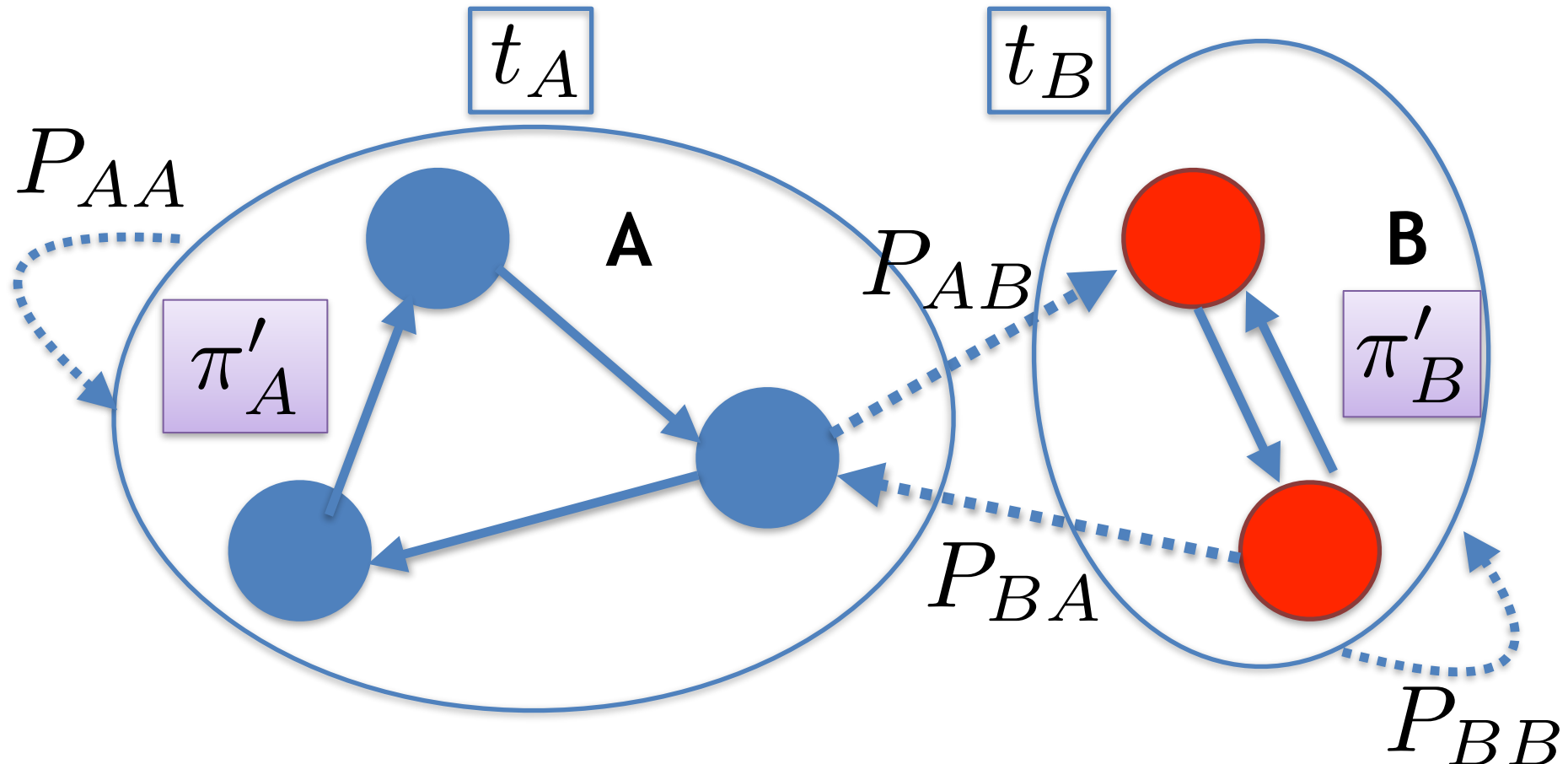
Step 3: Compute the $k \times k$ approximated transition matrix T between the subsets.

Koury et al. Aggregation-Disaggregation Algorithm



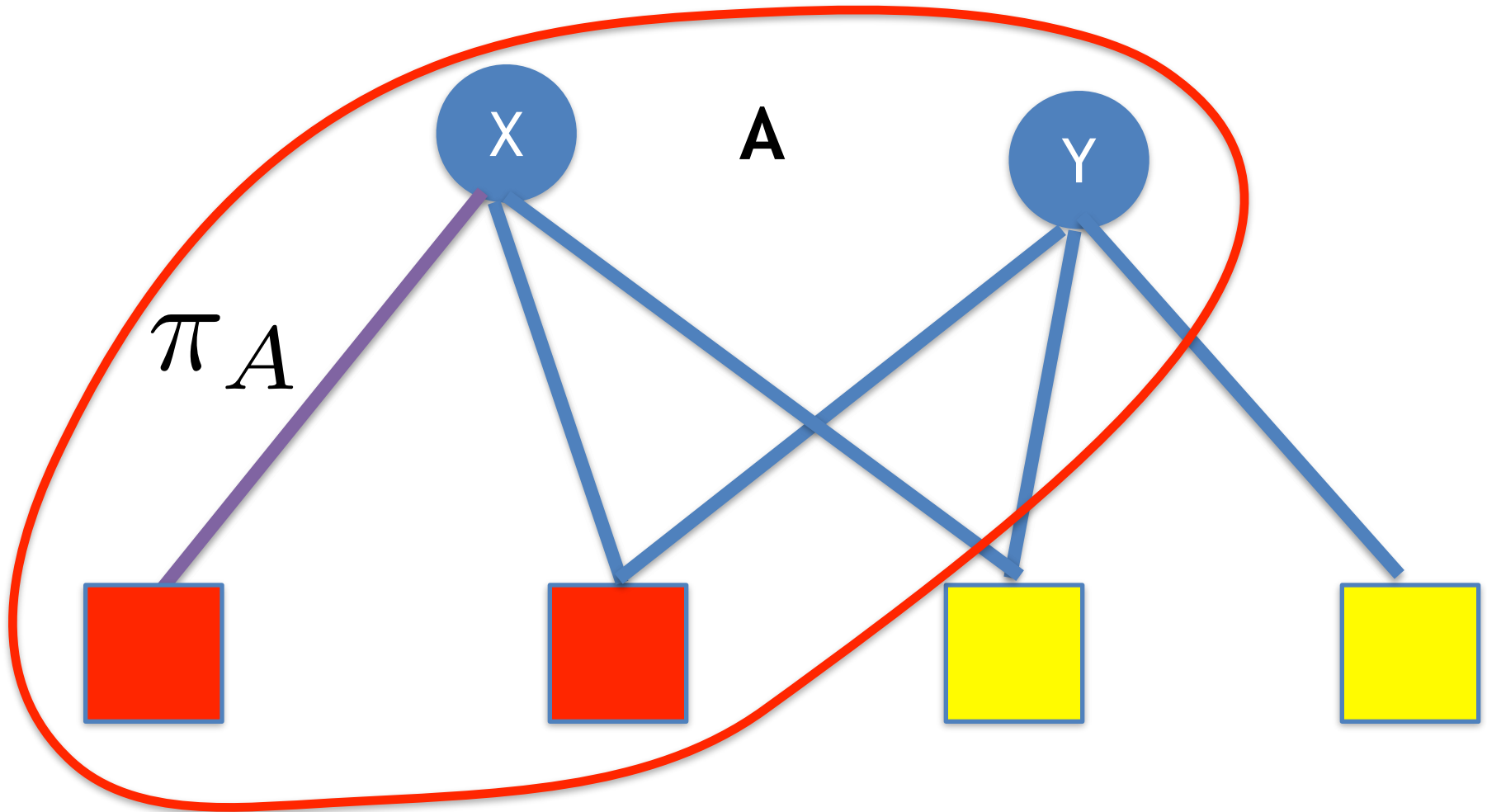
Step 4: Compute the stationary distribution of \mathbf{T} .

Koury et al. Aggregation-Disaggregation Algorithm



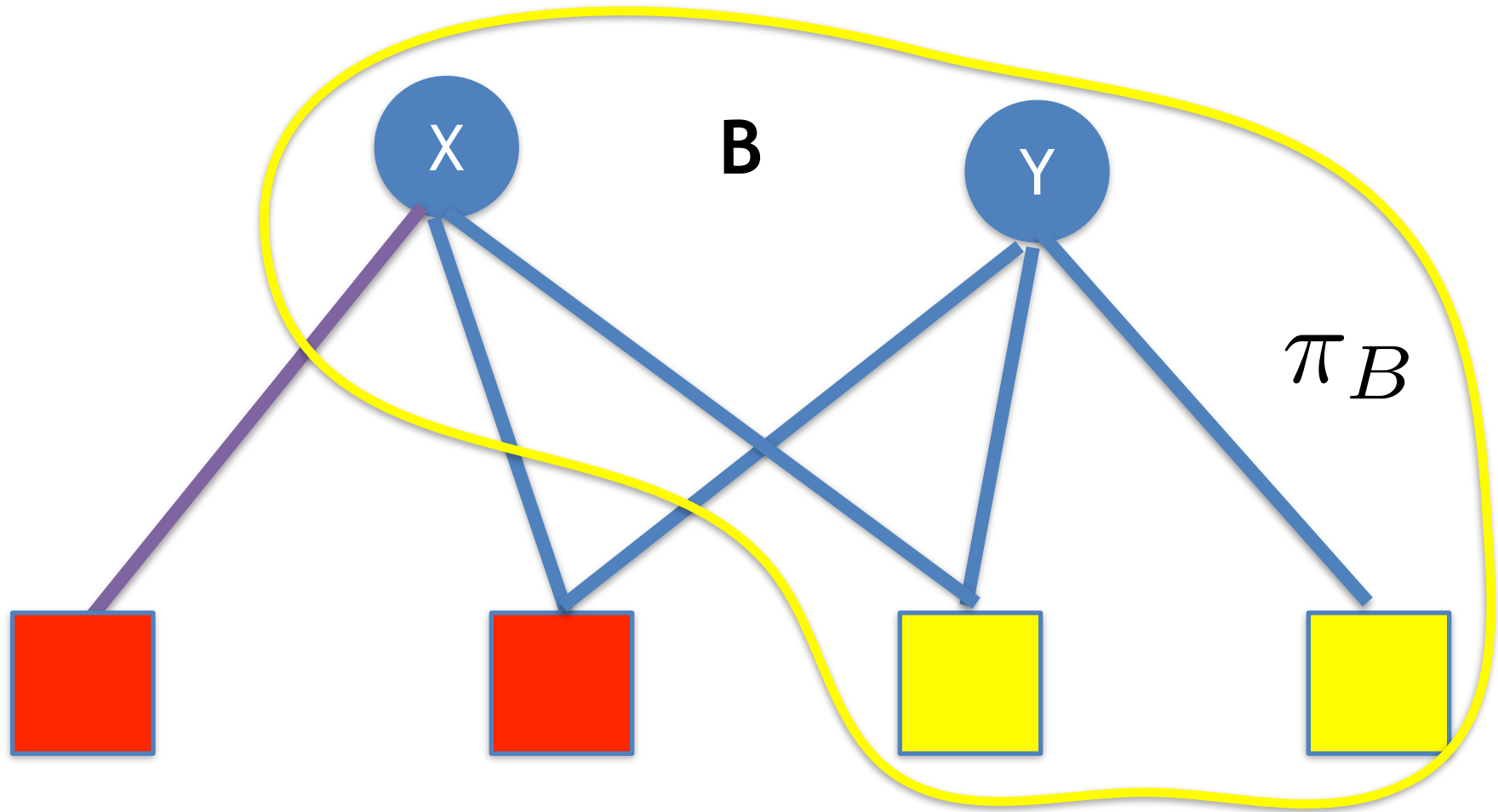
Step 5: Based on the stationary distribution improve the estimation of π_A and π_B . **Repeat** until convergence.

Aggregation in PPR



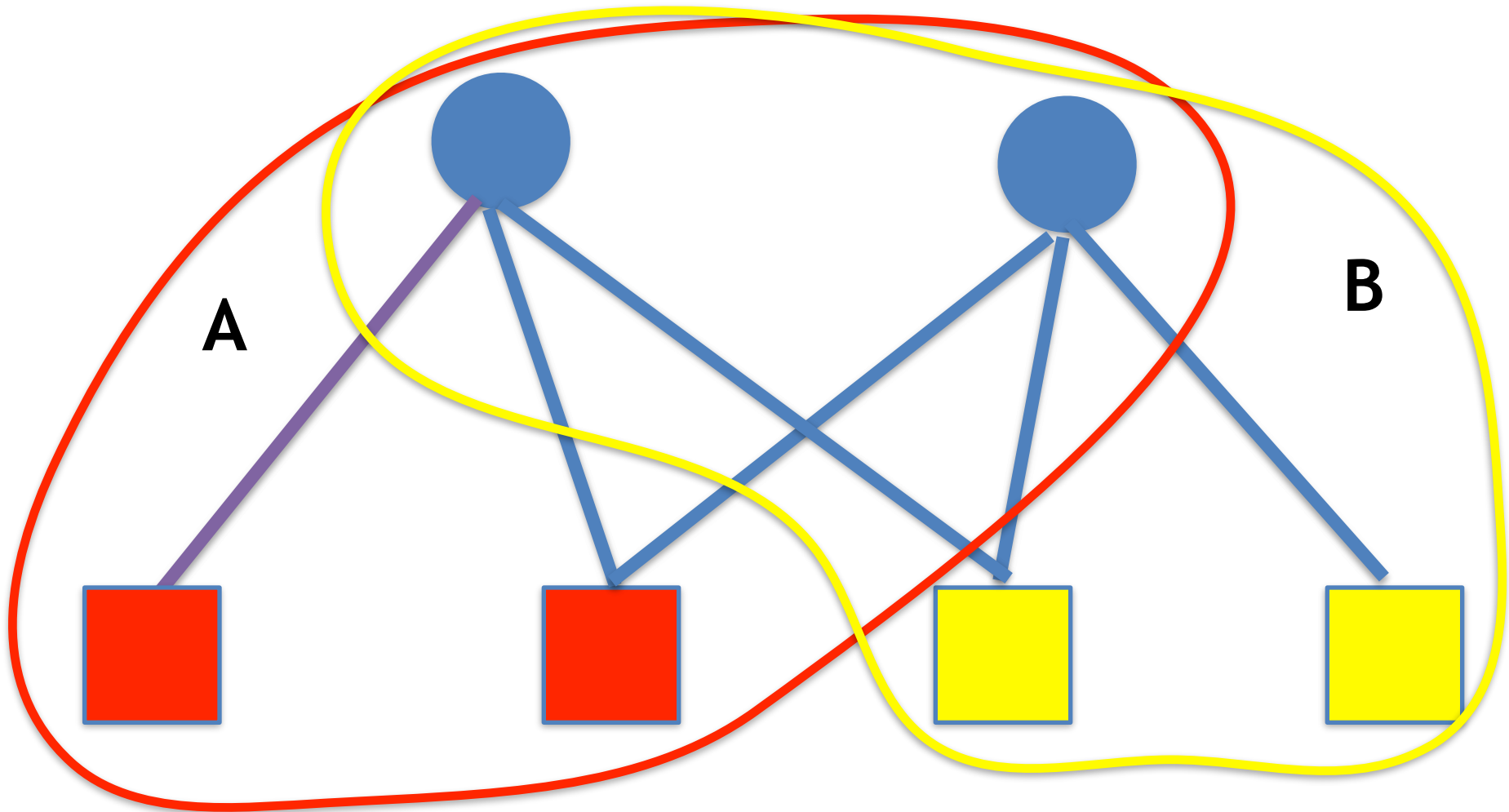
Precompute the stationary distributions individually

Aggregation in PPR



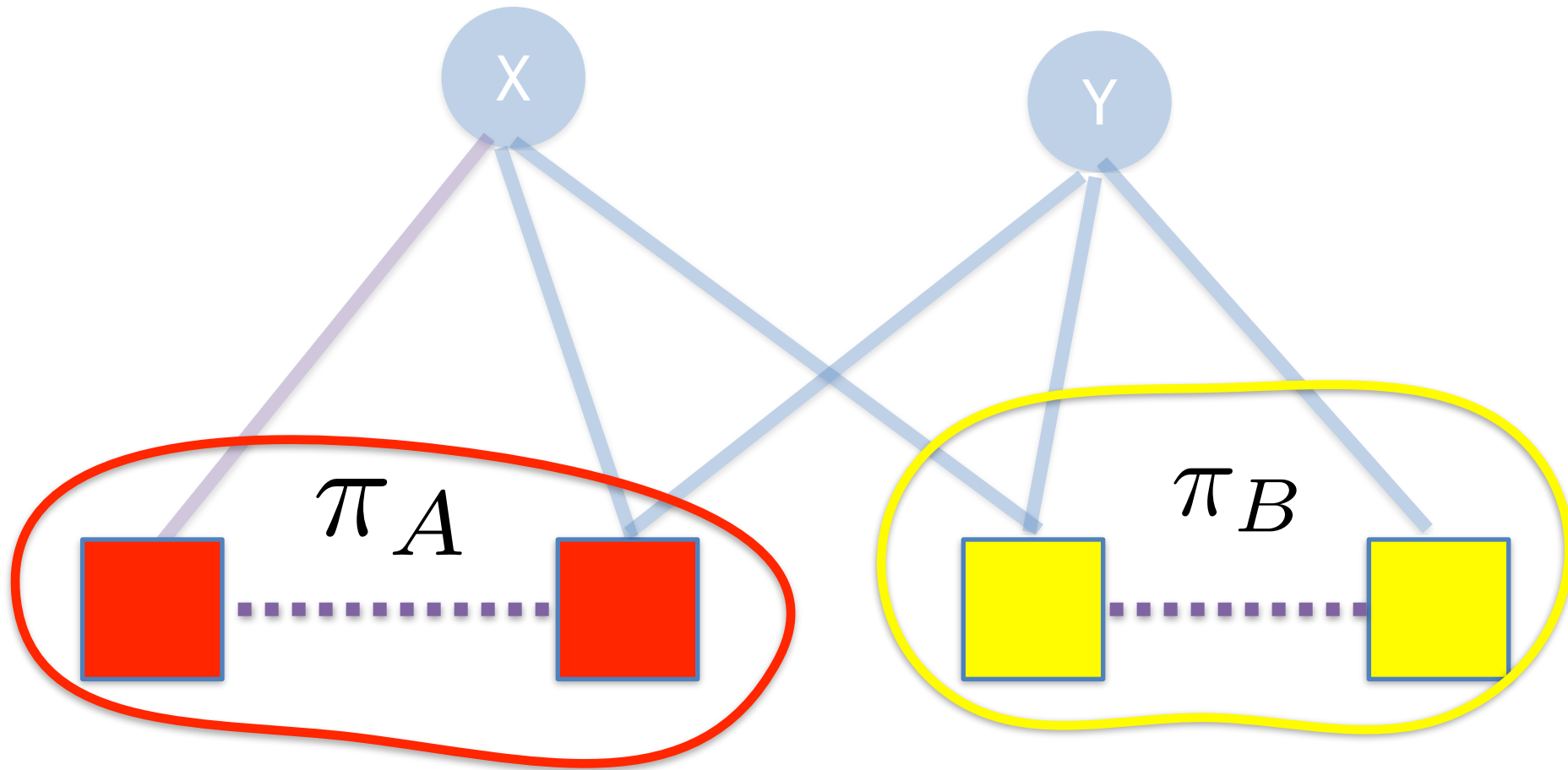
Precompute the stationary distributions individually

Aggregation in PPR

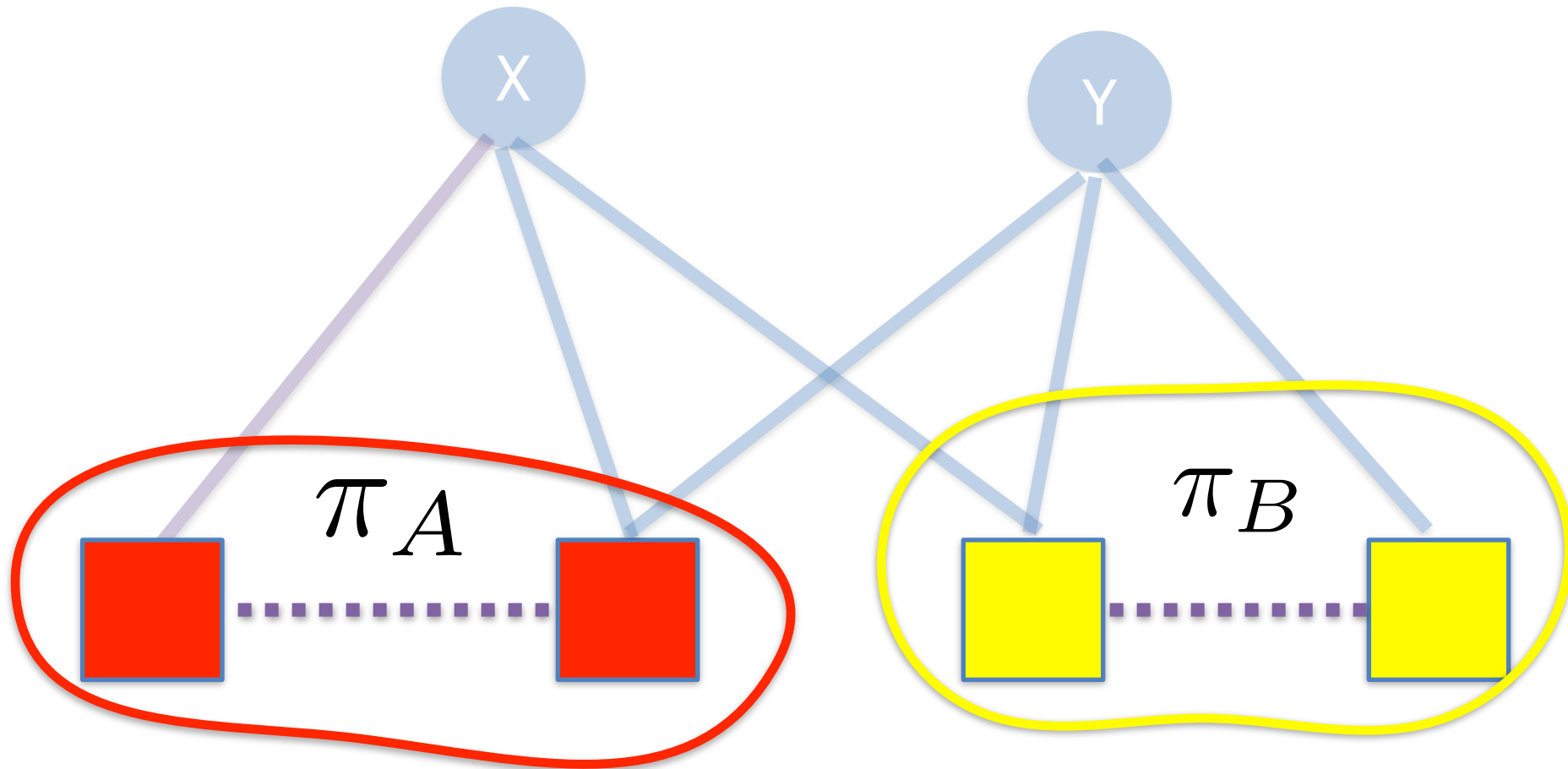


The two subsets are not disjoint!

Reduction to the Query Side

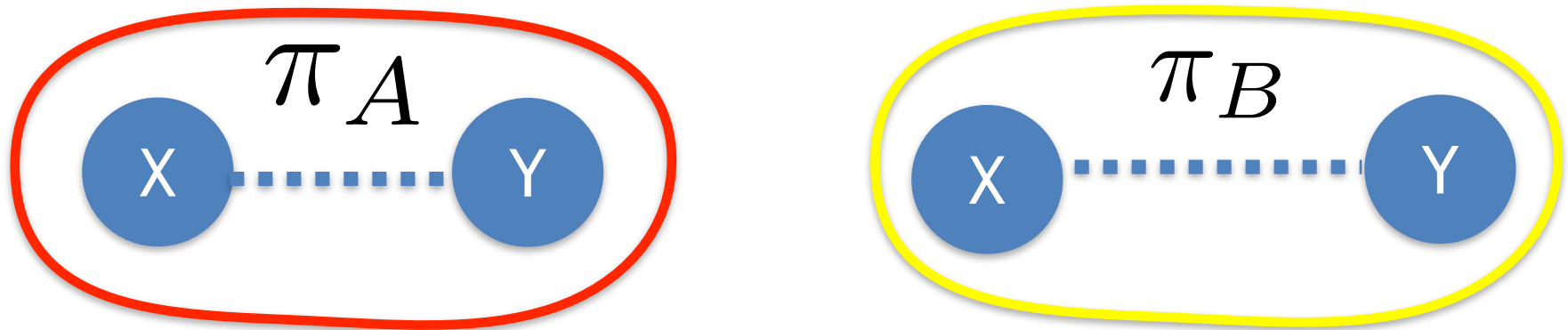


Reduction to the Query Side



This is the larger side of the graph.

Our Approach

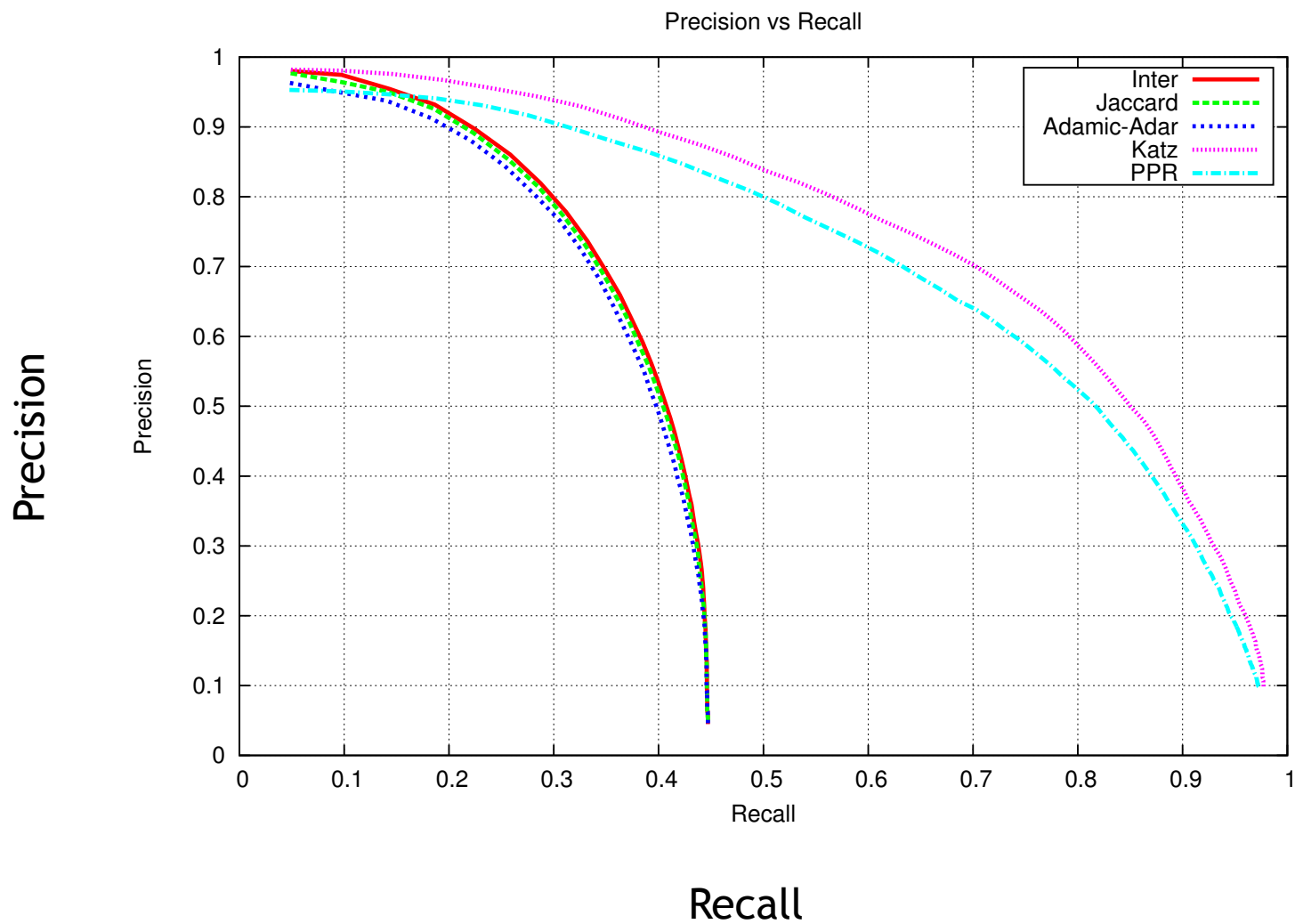


- We tackle the bijective relationships between the stationary distributions of the two sides.
- The algorithm is based **only** on the reduced graphs with Advertiser-Side nodes.
- The aggregation algorithm is scalable and converges to the correct distribution.

Experimental Evaluation

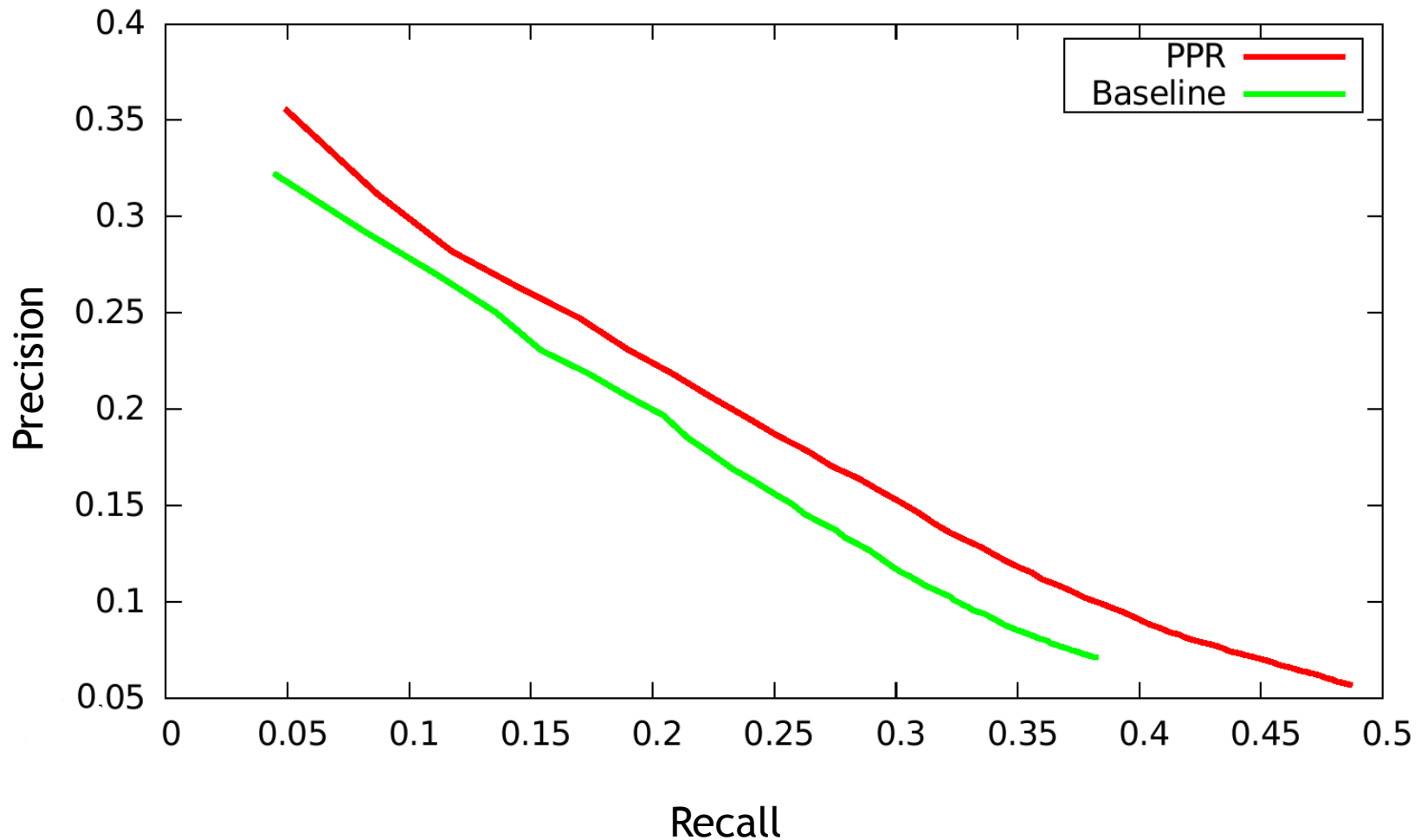
- We experimented with publicly available and proprietary datasets:
 - Query-Ads graph from Google AdWords > 1.5 billions nodes, > 5 billions edges.
 - DBLP Author-Papers and Patent Inventor-Inventions graphs.
- Ground-Truth clusters of competitors in Google AdWords.

Patent Graph



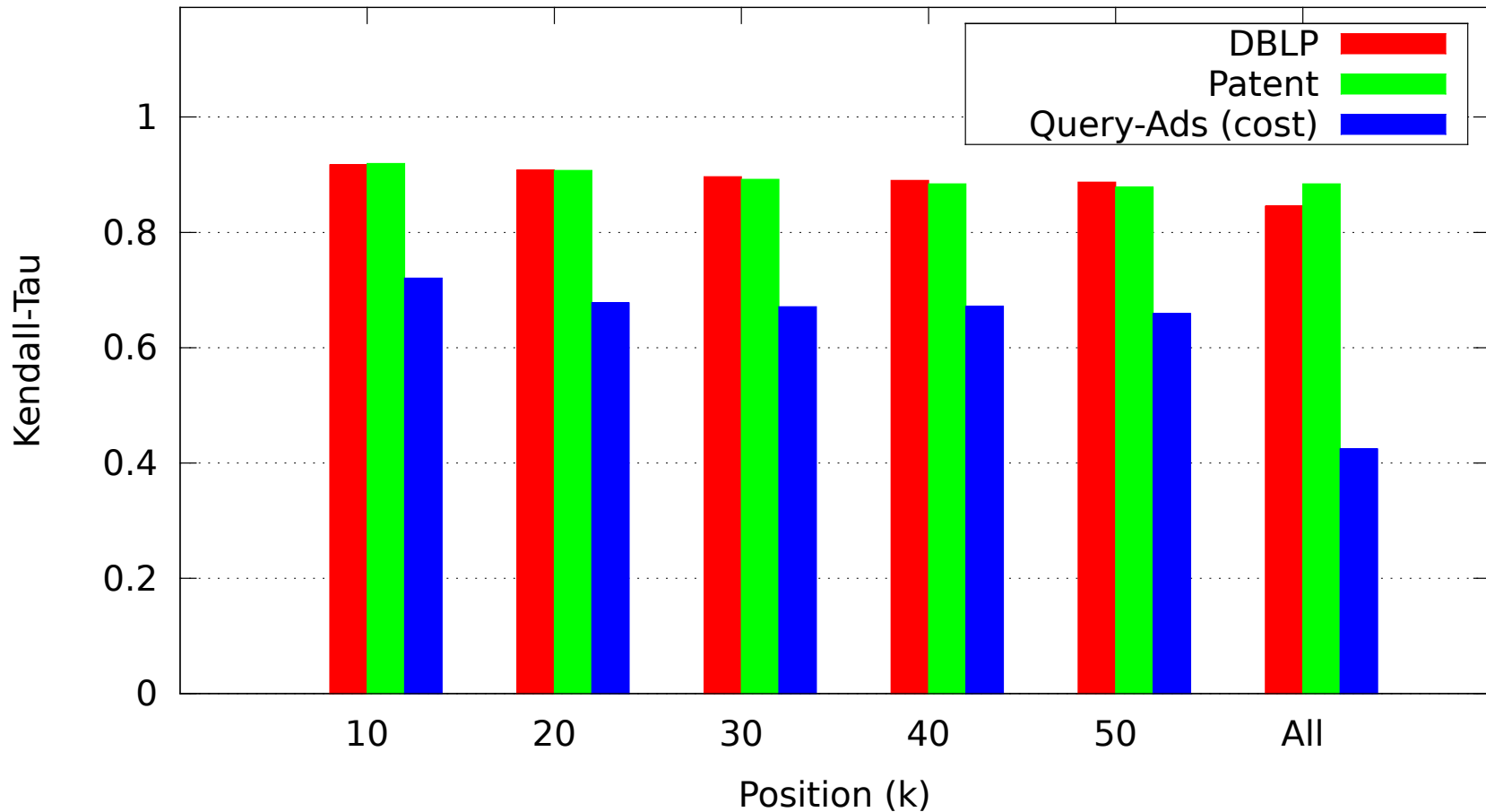
Google AdWords

Precision vs Recall



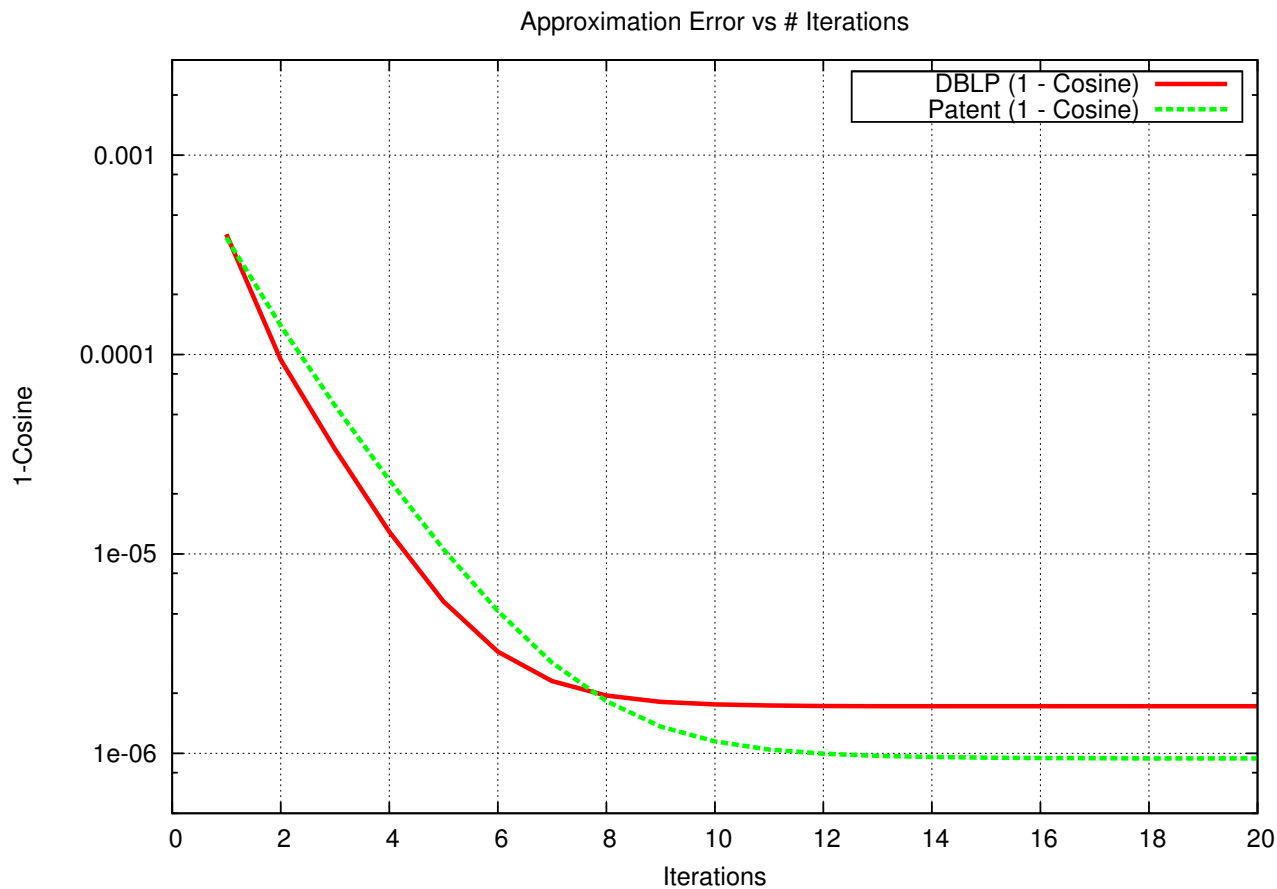
Convergence after One Iteration

Kendall-Tau Correlation



Convergence

1-Cosine Similarity



Iterations

Conclusions and Future Work

- Good accuracy and fast convergence.
- The framework can be applied to other problems and similarity measures.
- Future work could focus on the case where categories are not disjoint is relevant.

Thank you for your attention