HolOMa: Holistic Ontology Matching

Agata Barcik, Maximilian Möller

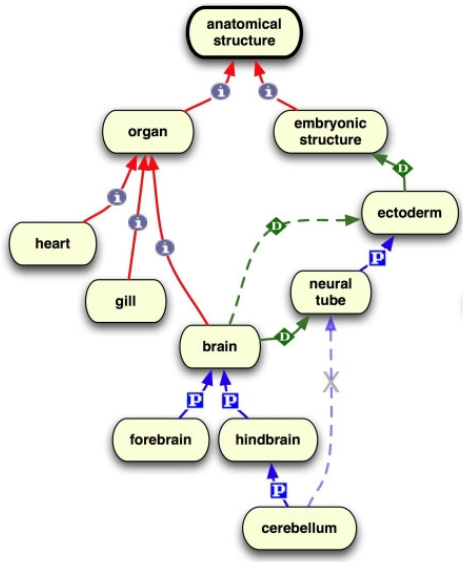
Big Data Praktikum
Universität Leipzig, Abteilung Datenbanken

Betreuer: Victor Christen

04. März 2016

Ontologie:

- konzeptionalisiert Wissen einer bestimmten Domäne $\rightsquigarrow$ *Anatomie*
- modellierbar als Graph
  - Knoten sind Konzepte $\rightsquigarrow$ *Herz*
  - Kanten sind Beziehungen zwischen Konzepten $\rightsquigarrow$ *is-a*



[WL08]

Matching:

- Finden von korrespondierenden Konzepten in zwei Ontologien

  $\rightsquigarrow$ *same-as* Beziehung zwischen diesen Konzepten

- eine Ontologie wird als Quelle, die andere als Ziel definiert
- Beispiel:

  Sei $c_1 :=$ *Leukozyt* ein Konzept in Quelle O_1.

  Sei $c_2 :=$ *Weiße Blutzellen* ein Konzept in Ziel O_2.

  Dann gilt: $\langle c_1, \textit{same-as}, c_2 \rangle$

Aufgabenstellung:

gegeben:

- Menge $\mathcal{O}$ von Ontologie $O_1, \ldots, O_n$
- Menge $\mathcal{K}$ von Korrespondenzen zwischen den Ontologien

Aufgabenstellung:

gegeben:

- Menge $\mathcal{O}$ von Ontologie $O_1, \ldots, O_n$
- Menge $\mathcal{K}$ von Korrespondenzen zwischen den Ontologien

gesucht:

- Menge $\mathcal{K}_{\text{new}}$ von Korrespondenzen, die nicht in $\mathcal{K}$ enthalten sind
- Menge $\mathcal{K}_{\text{false}}$ von Korrespondenzen, die fälschlicherweise in $\mathcal{K}$ sind

Aufgabenstellung:

gegeben:

- Menge $\mathcal{O}$ von Ontologie $O_1, \ldots, O_n$
- Menge $\mathcal{K}$ von Korrespondenzen zwischen den Ontologien

gesucht:

- Menge $\mathcal{K}_{\text{new}}$ von Korrespondenzen, die nicht in $\mathcal{K}$ enthalten sind
- Menge $\mathcal{K}_{\text{false}}$ von Korrespondenzen, die fälschlicherweise in $\mathcal{K}$ sind

Vorgehen:

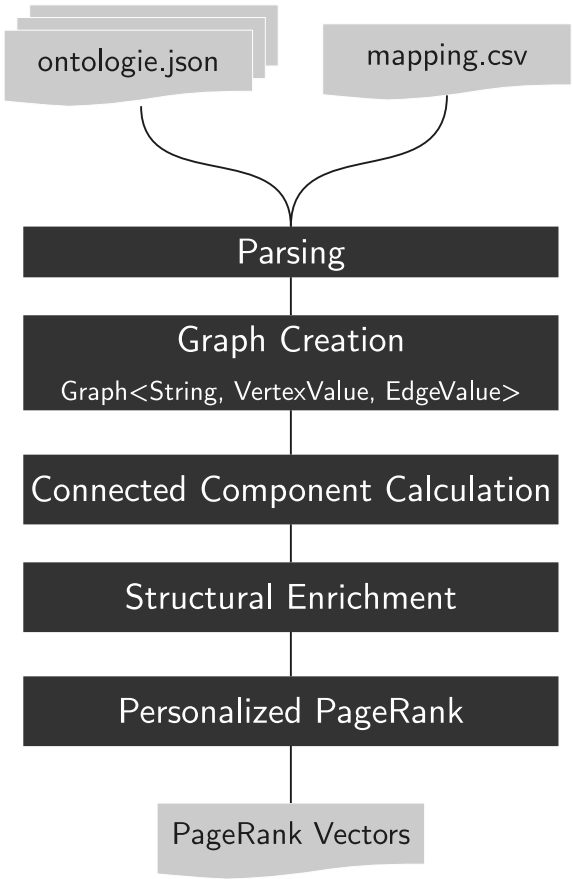
- Personalized PageRank [BCX11, PBMW99]
  - Surfen in einem Netzwerk mit einer Teleportations-wahrscheinlichkeit zum Startknoten

Technologie:

- HolOMa Prozess in Java 7 implementiert
- Gelly (Graph Processing API, Apache Flink), Release 0.10.2
  - Transformationen: *map*, *filter*, ...
  - Mutationen: *addVertex*, *removeEdge*, ...
  - Aggregation: *reduceOnEdges*, *reduceOnNeighbors*, ...
  - Iterationen: *vertex centric iteration*, ...

[Gel]

Überblick


ontologie.json
mapping.csv
Parsing
Graph Creation
Graph<String, VertexValue, EdgeValue>
Connected Component Calculation
Structural Enrichment
Personalized PageRank
PageRank Vectors

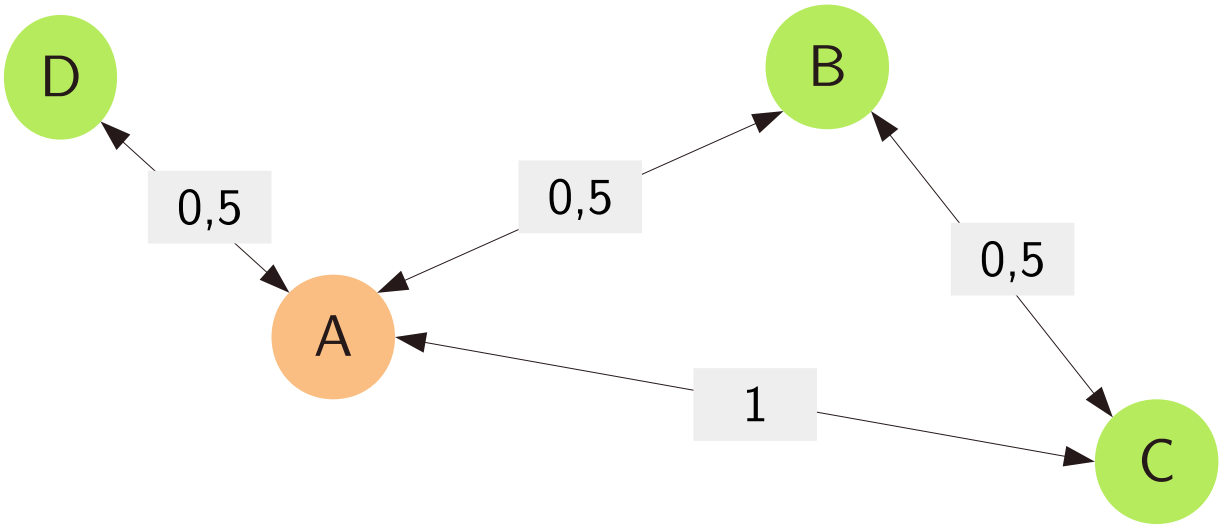
Einblick: Personalized PageRank Vertex-Centrix Iteration

- Superstep S_i aus Messaging und Updating
- **Messaging**

  PPR-Wert multipliziert mit normiertem Kantengewicht
- **Updating**

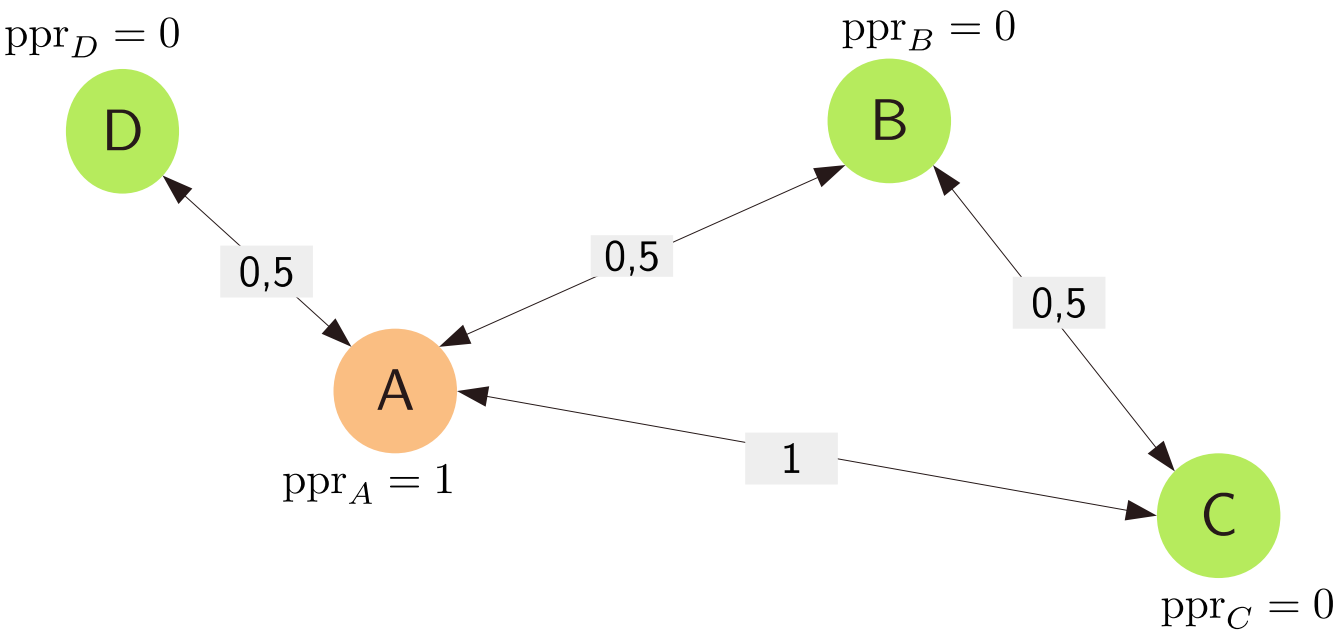
  summiere Message-Inhalt,
  Teleportationswahrscheinlichkeit, wenn Knoten gleich Startknoten
- Iteration terminiert gdw.
  - maximale Anzahl von Iterationen erreicht, oder
  - jeder Knoten bei Updating seinen Wert nicht ändert

Einblick: Personalized PageRank Vertex-Centrix Iteration



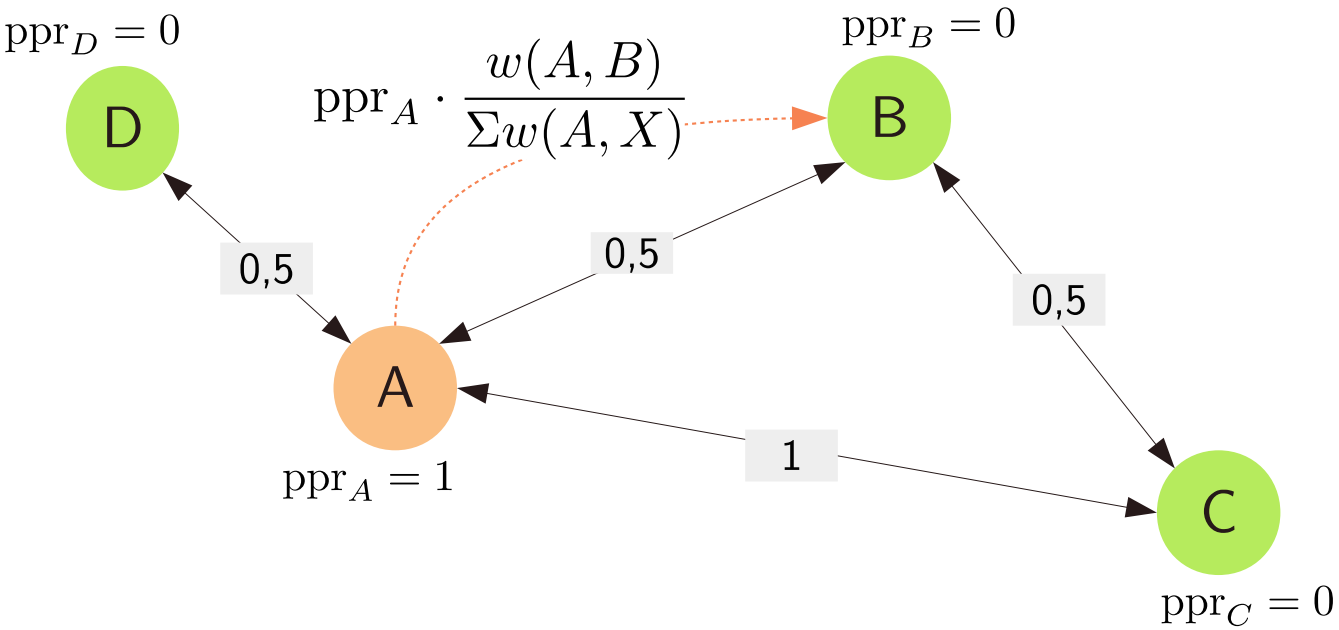
Einblick: Personalized PageRank

Initialisierung



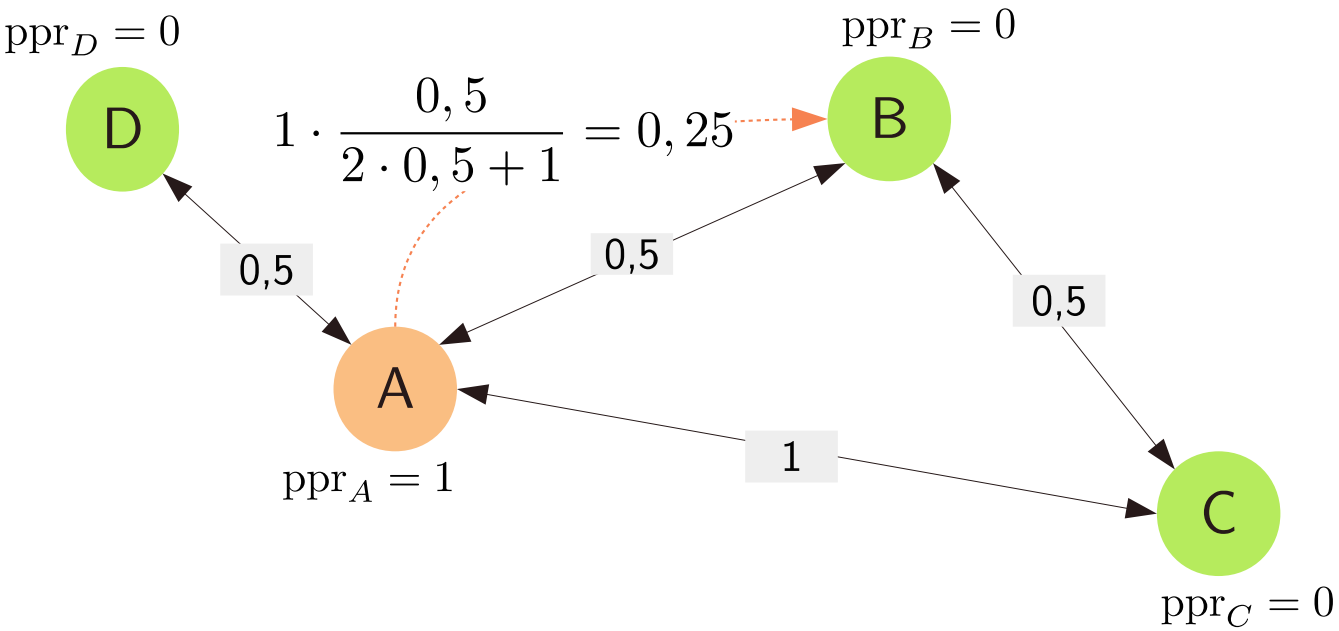
Einblick: Personalized PageRank

S1: Messaging



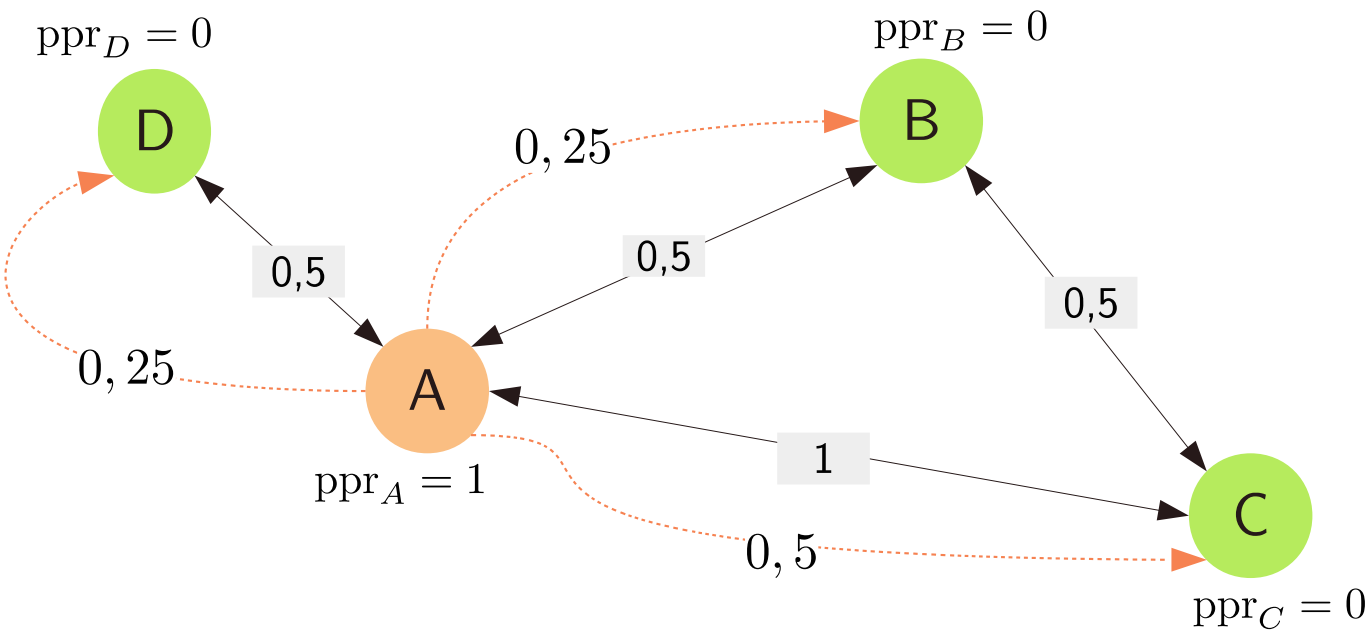
Einblick: Personalized PageRank

S1: Messaging



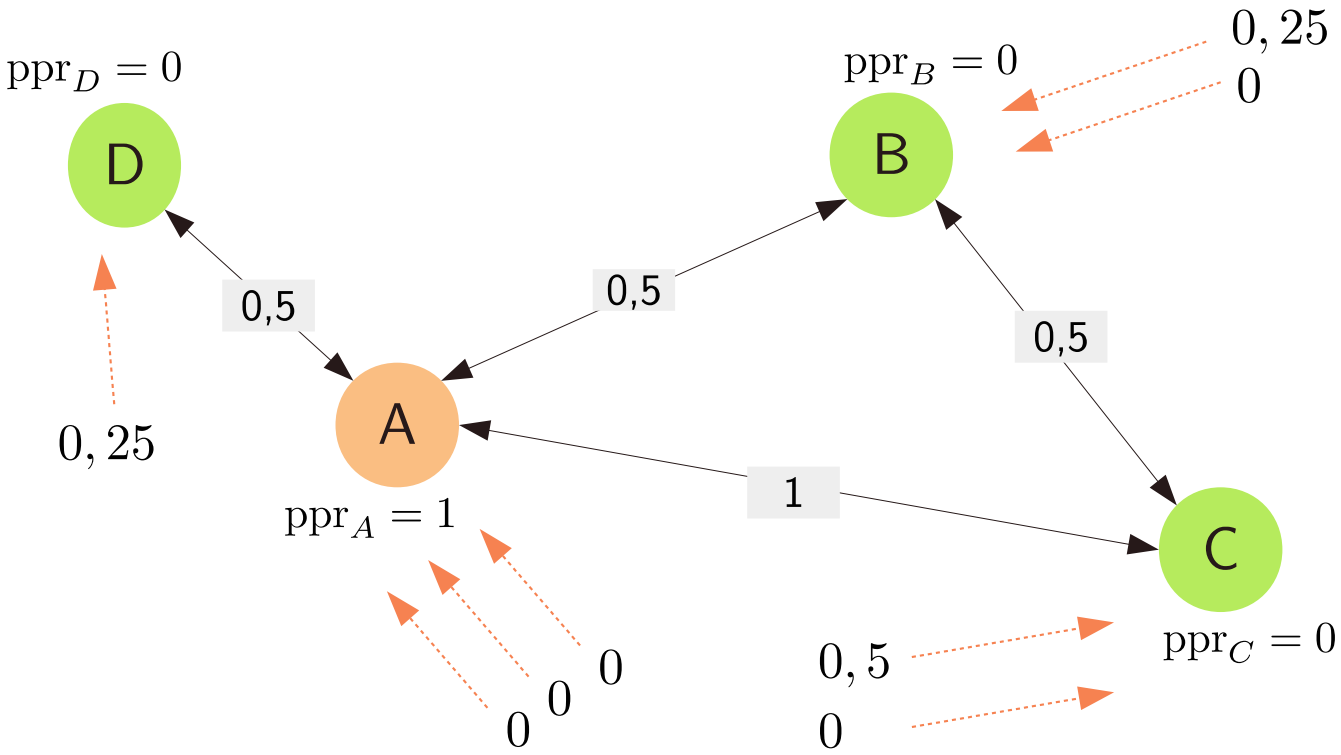
Einblick: Personalized PageRank

S1: Messaging



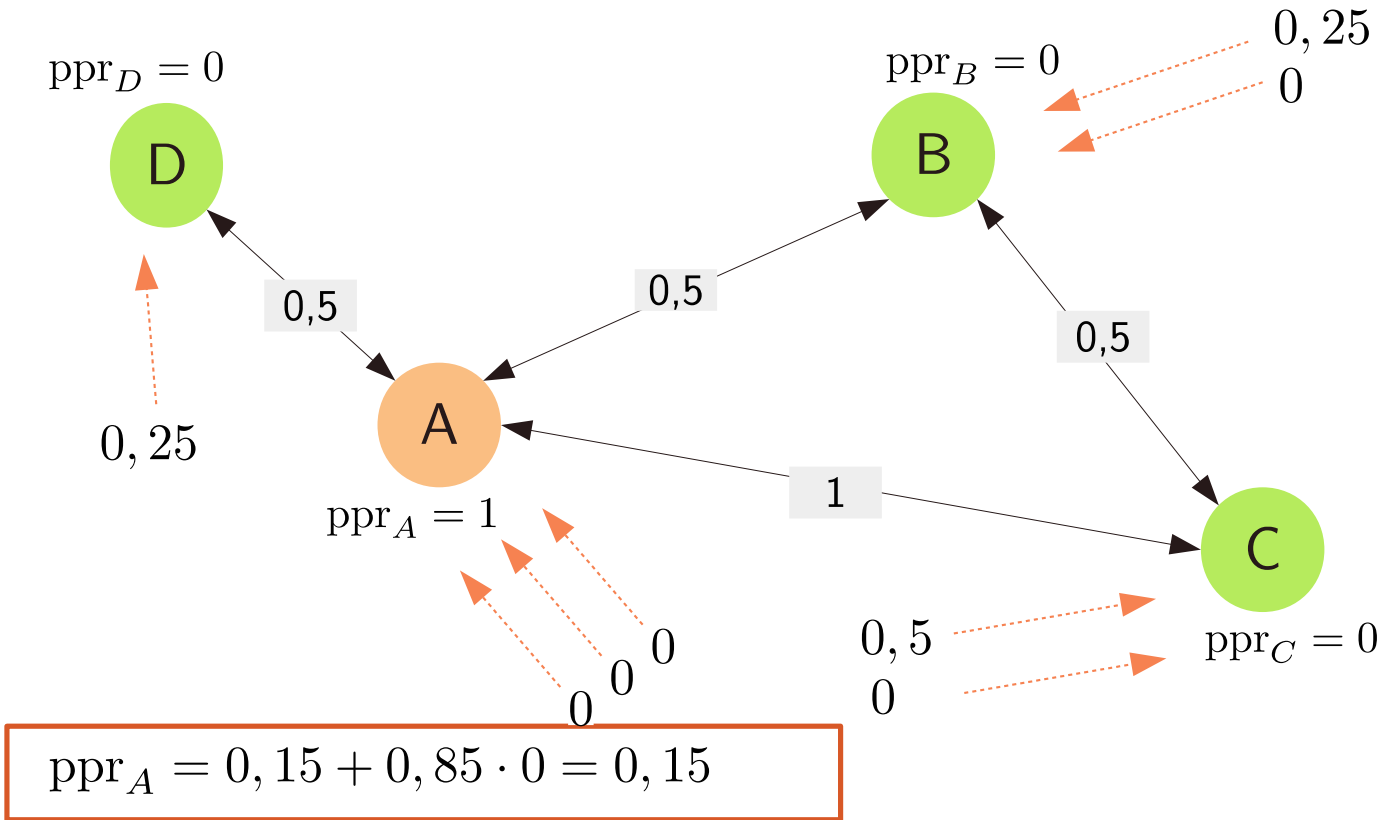
Einblick: Personalized PageRank

S1: Messaging



Einblick: Personalized PageRank

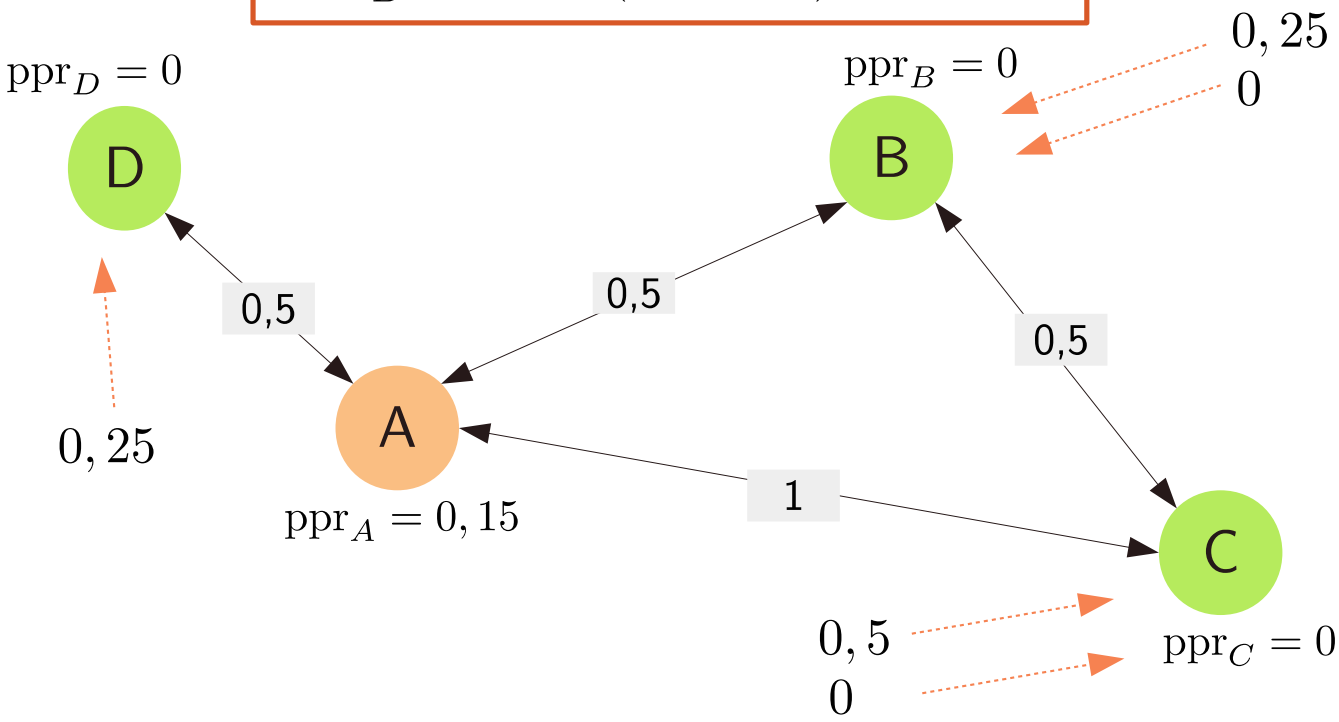
S1: Updating



Einblick: Personalized PageRank

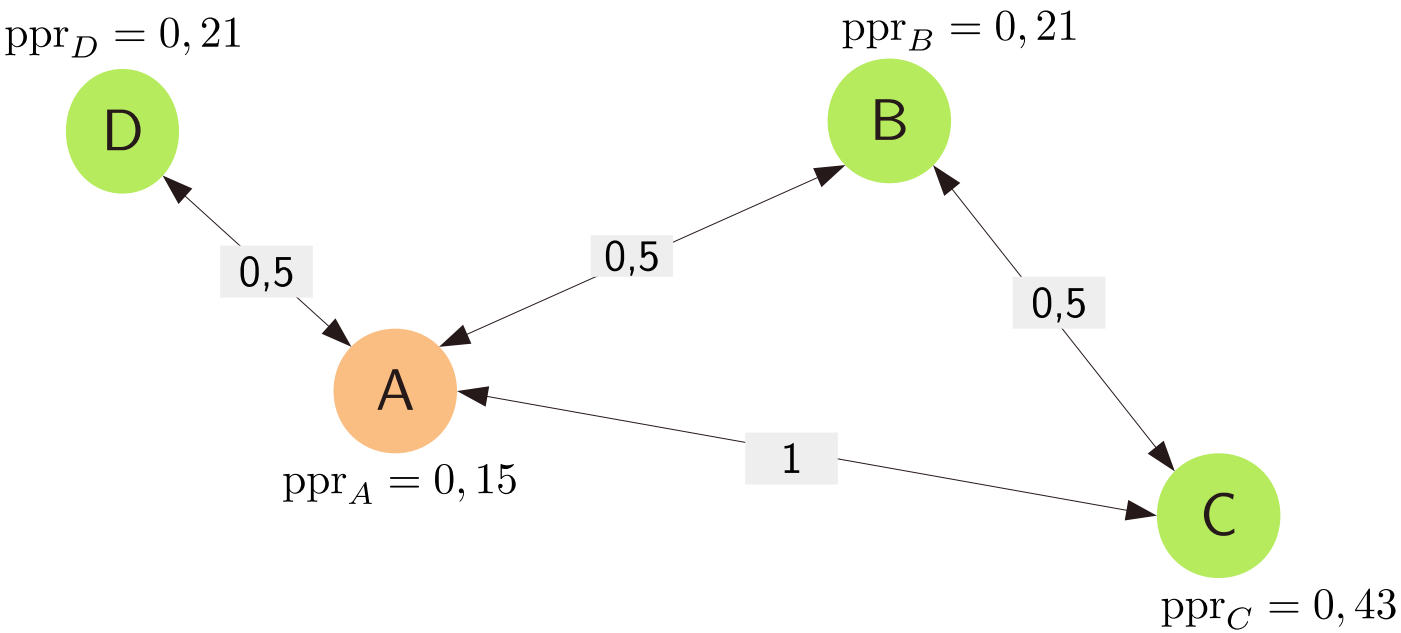
S1: Updating

$$\text{ppr}_B = 0,85 \cdot (0,25 + 0) = 0,21$$



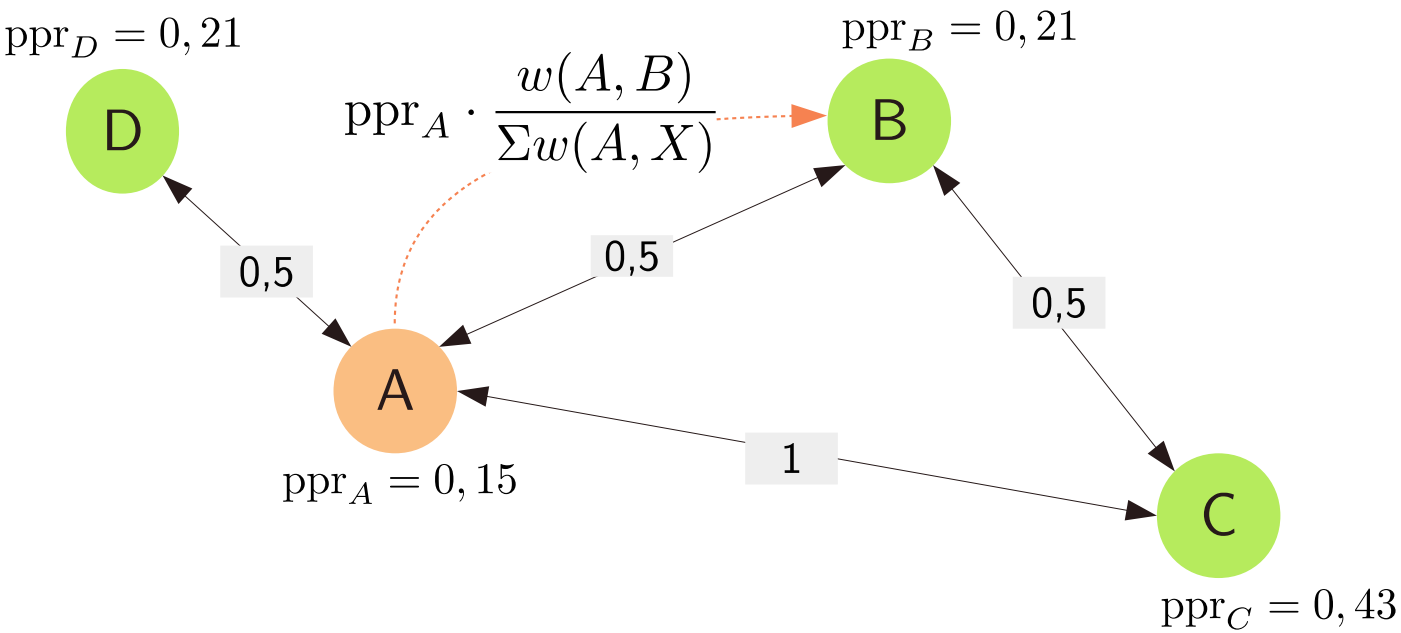
Einblick: Personalized PageRank

S1: Updating



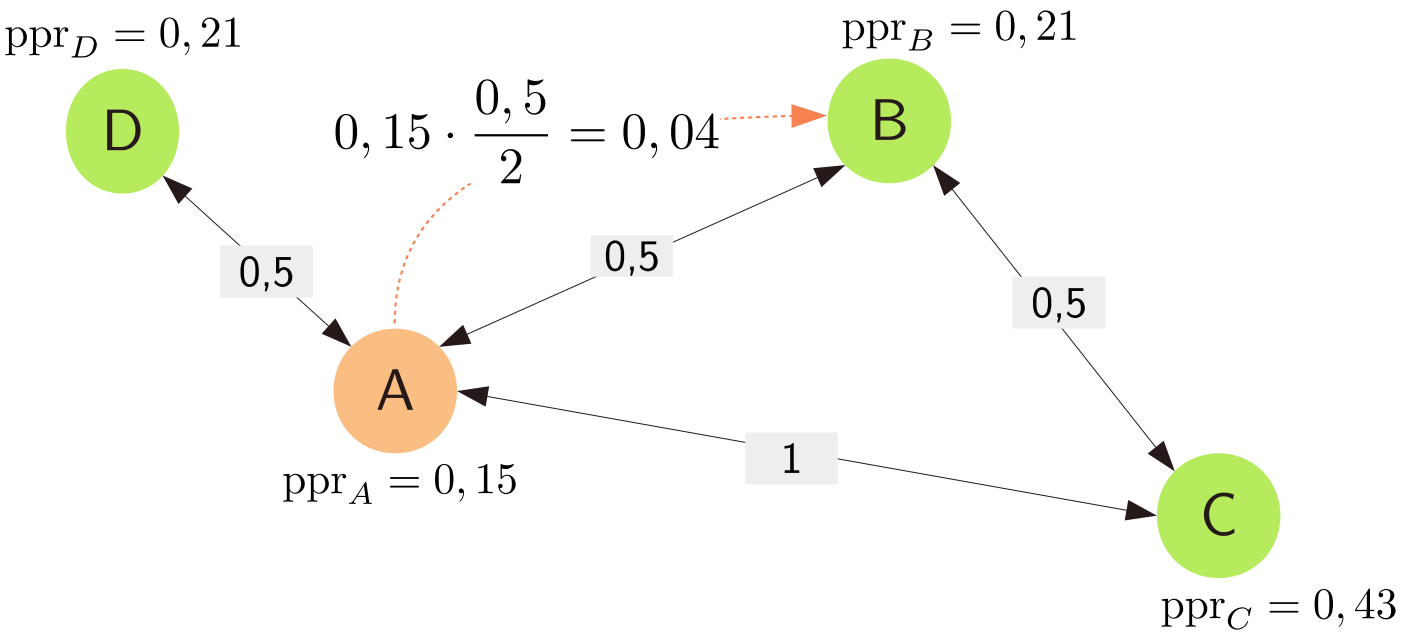
Einblick: Personalized PageRank

S2: Messaging



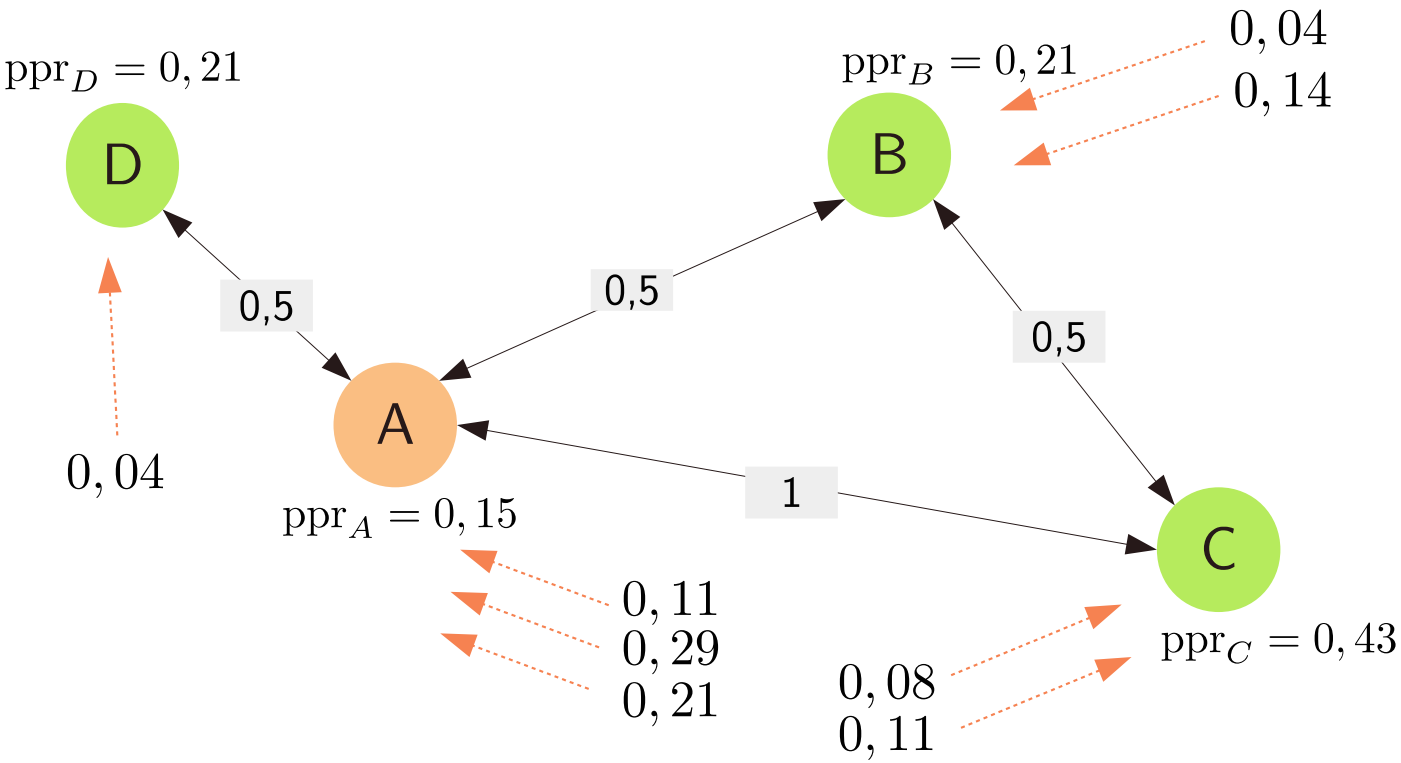
Einblick: Personalized PageRank

S2: Messaging



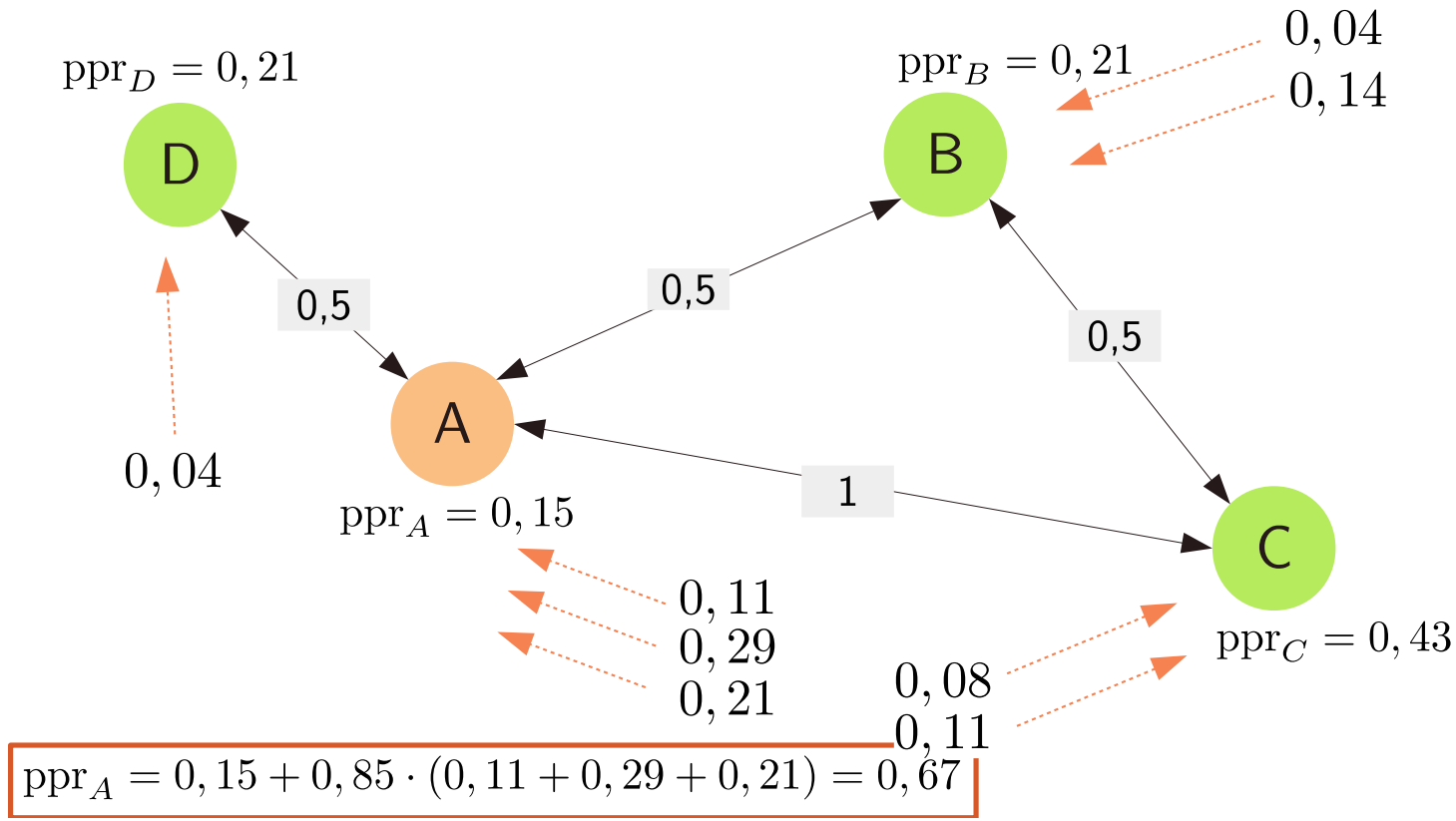
Einblick: Personalized PageRank

S2: Messaging



Einblick: Personalized PageRank

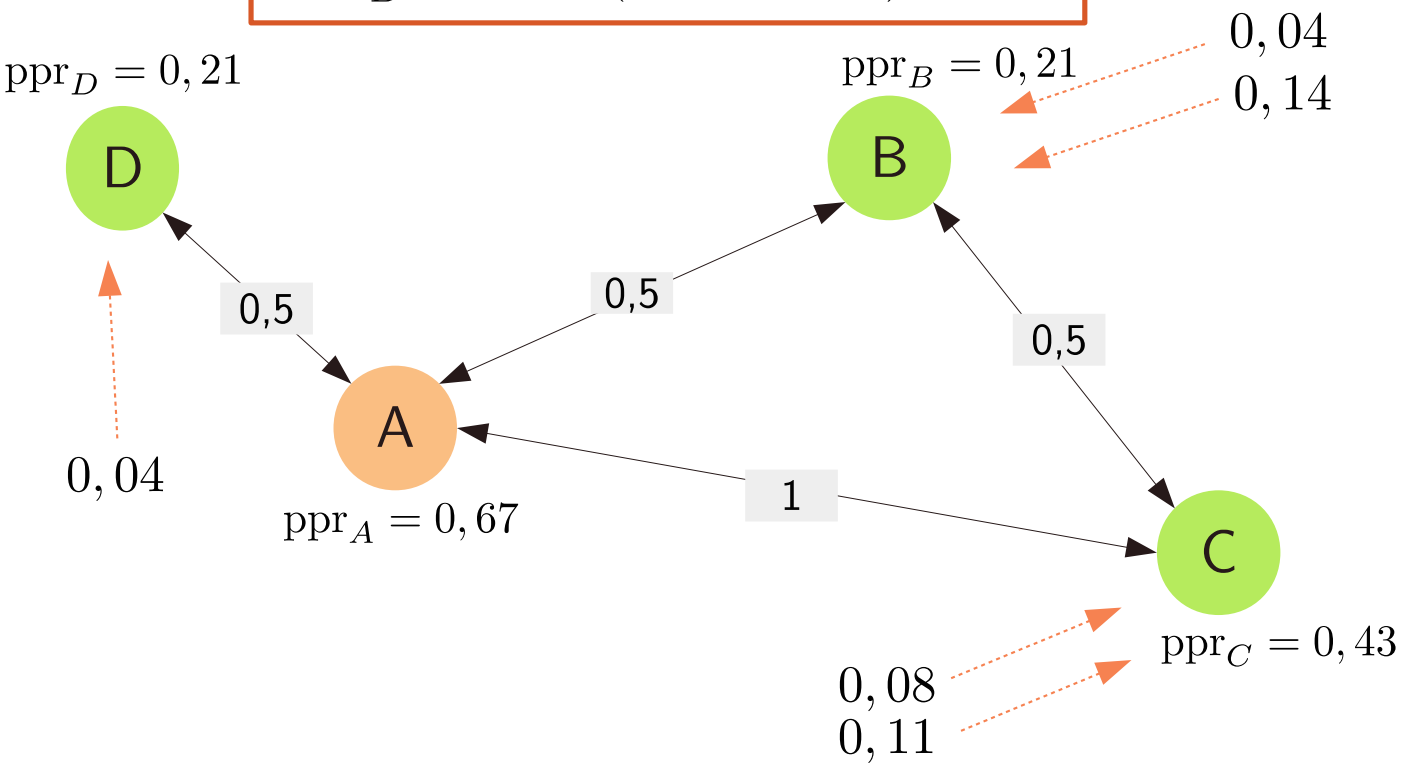
S2: Updating



Einblick: Personalized PageRank

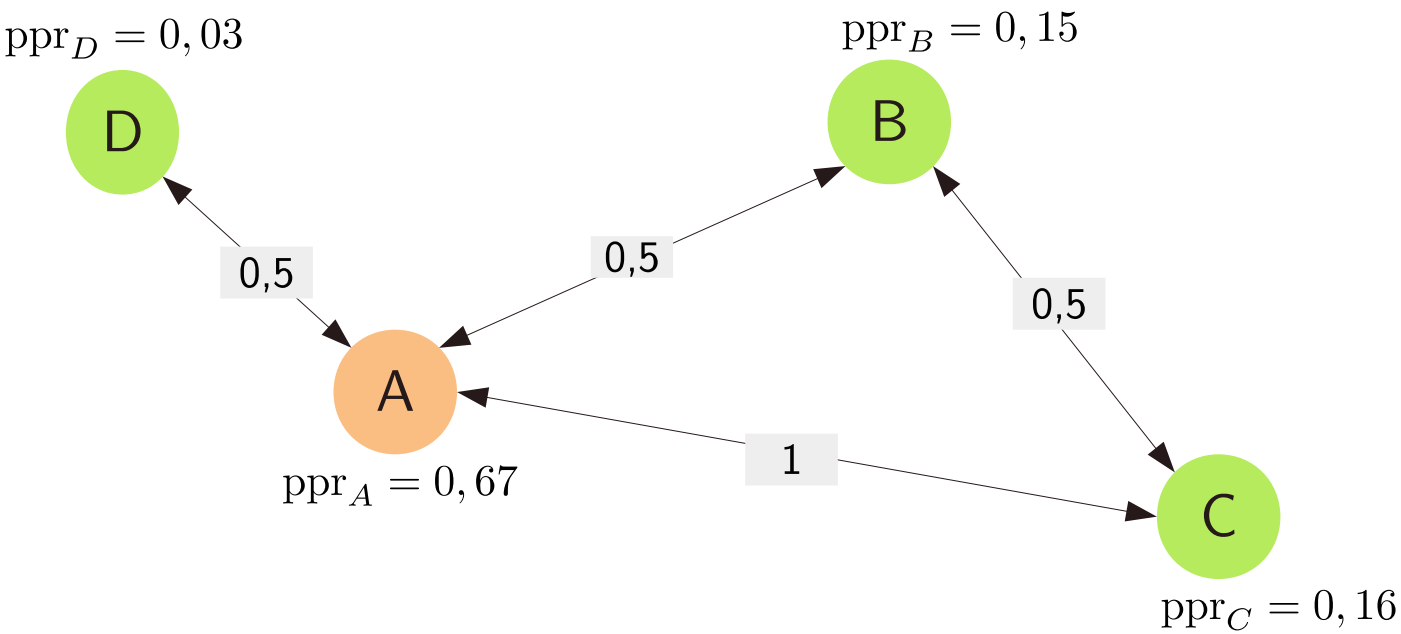
S2: Updating

$$\text{ppr}_B = 0,85 \cdot (0,04 + 0,14) = 0,15$$



Einblick: Personalized PageRank

S2: Updating

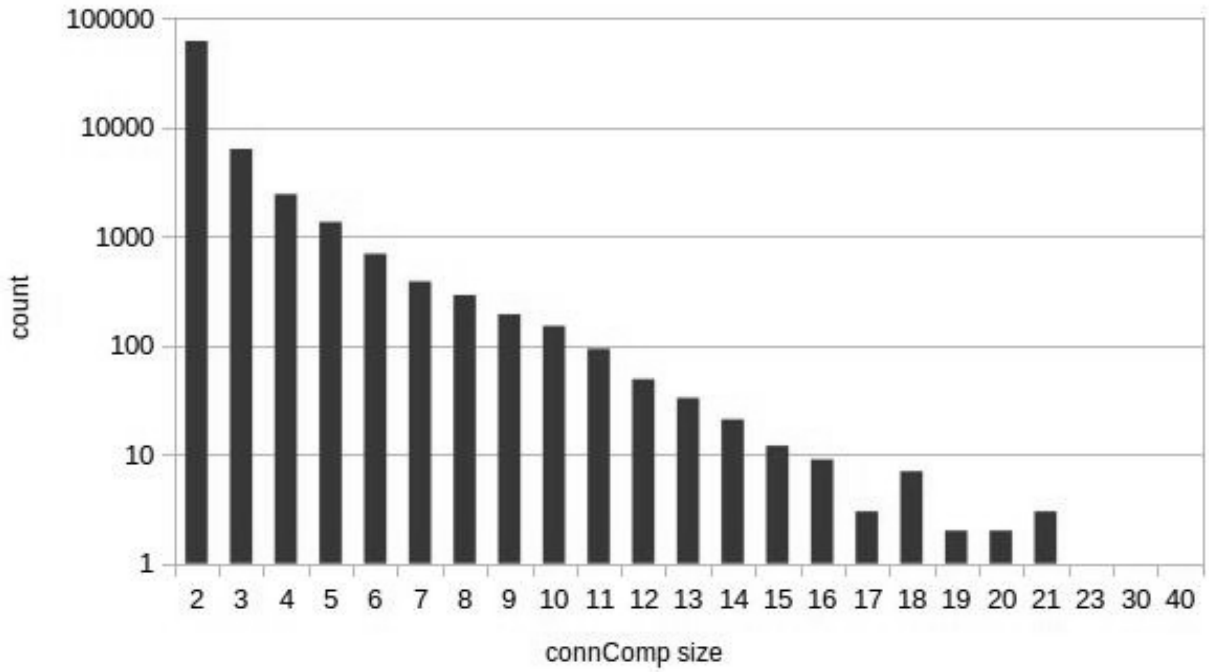


Datenbasis

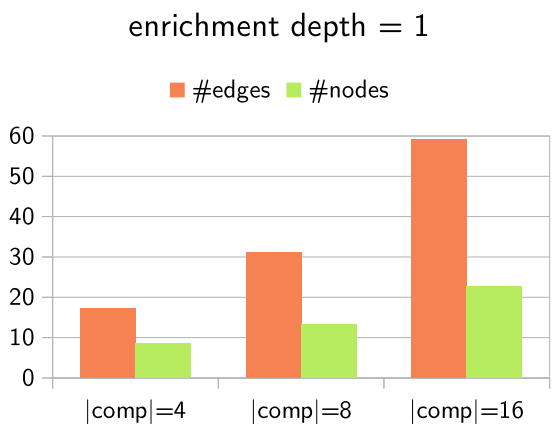
- 9 Ontologien aus der Bio-Medizin-Domäne
  - $\approx$ 360 000 Konzepten,
  - $\approx$ 596 000 Kanten
- RxNorm (Arzneimittel), PDQ (Krebs), NATPRO (Naturprodukte), Galen (Medizin), MeSH (Medizin), OMIM (Genetik), RadLex (Radiologie), ChEBI (Moleküle), FMA (Anatomie)

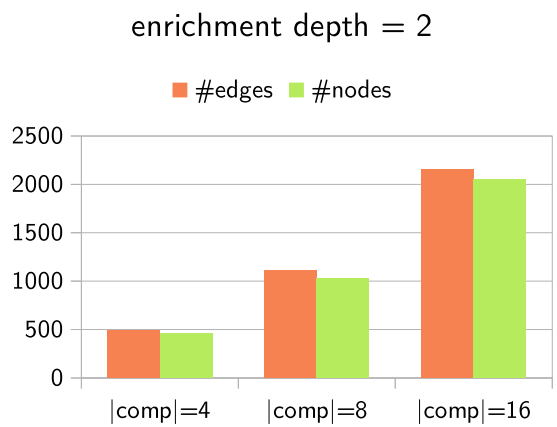
Zusammenhangskomponenten

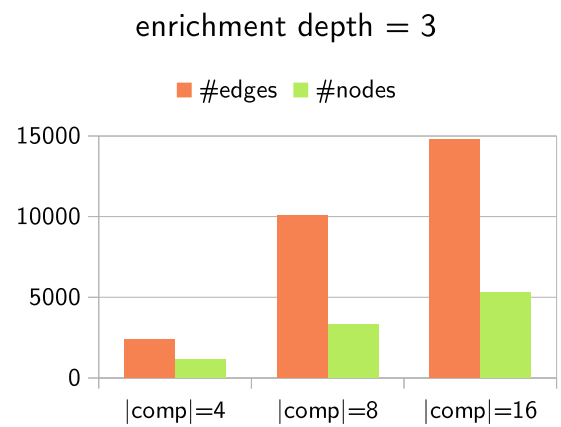
- 74 000 Komponenten mit mind. zwei Konzepten
  - AVG(|comp|): 2,36
  - MAX(|comp|): 40



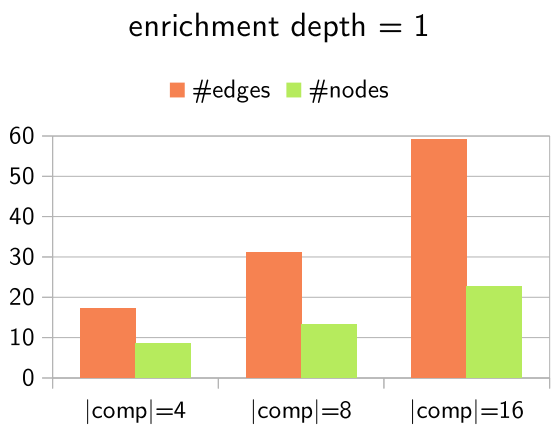
Strukturelle Anreicherung

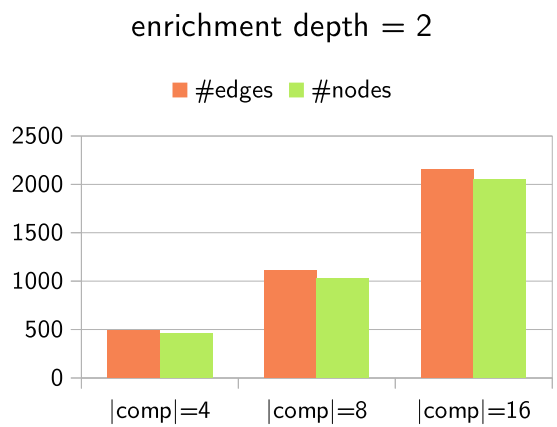


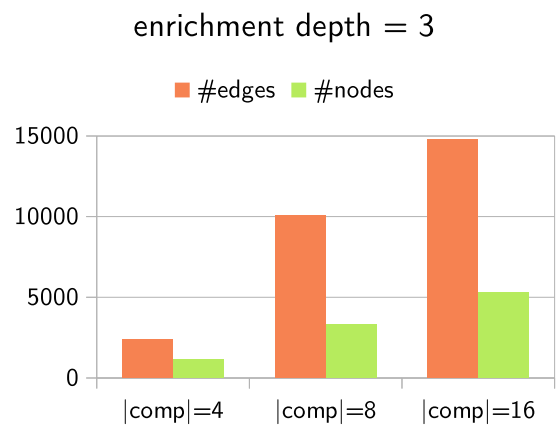


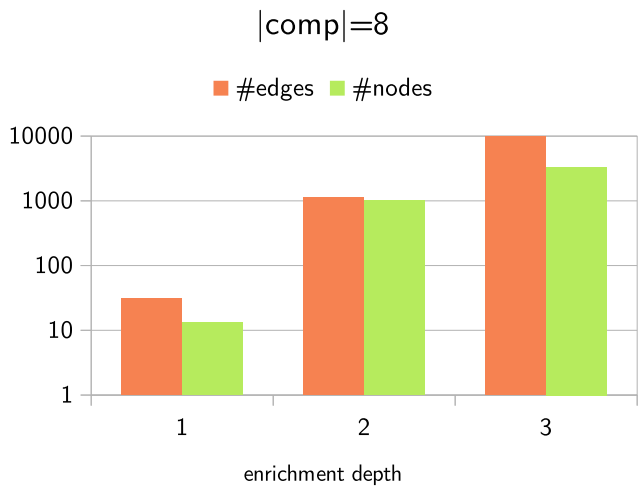


Strukturelle Anreicherung

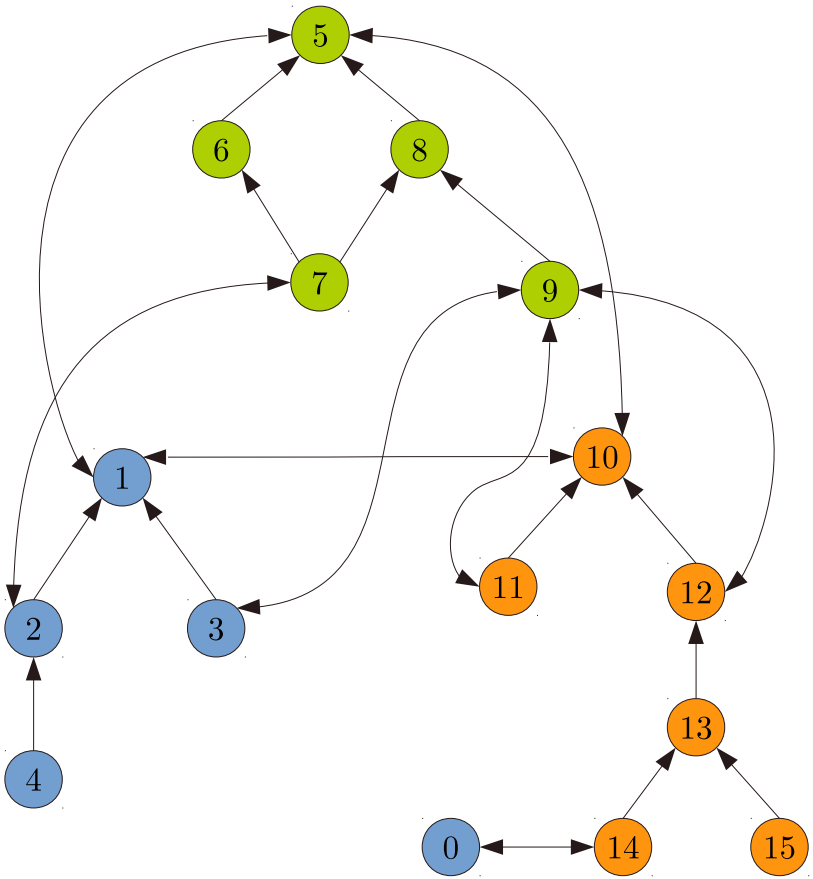




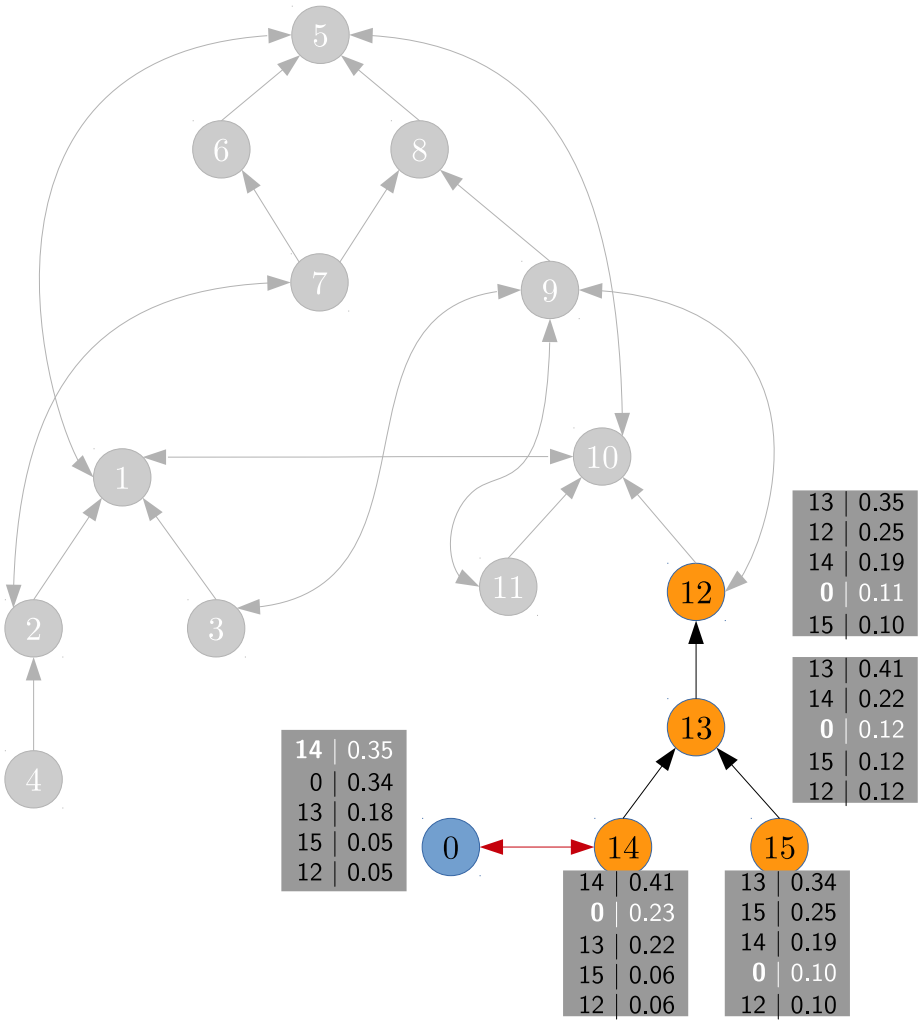




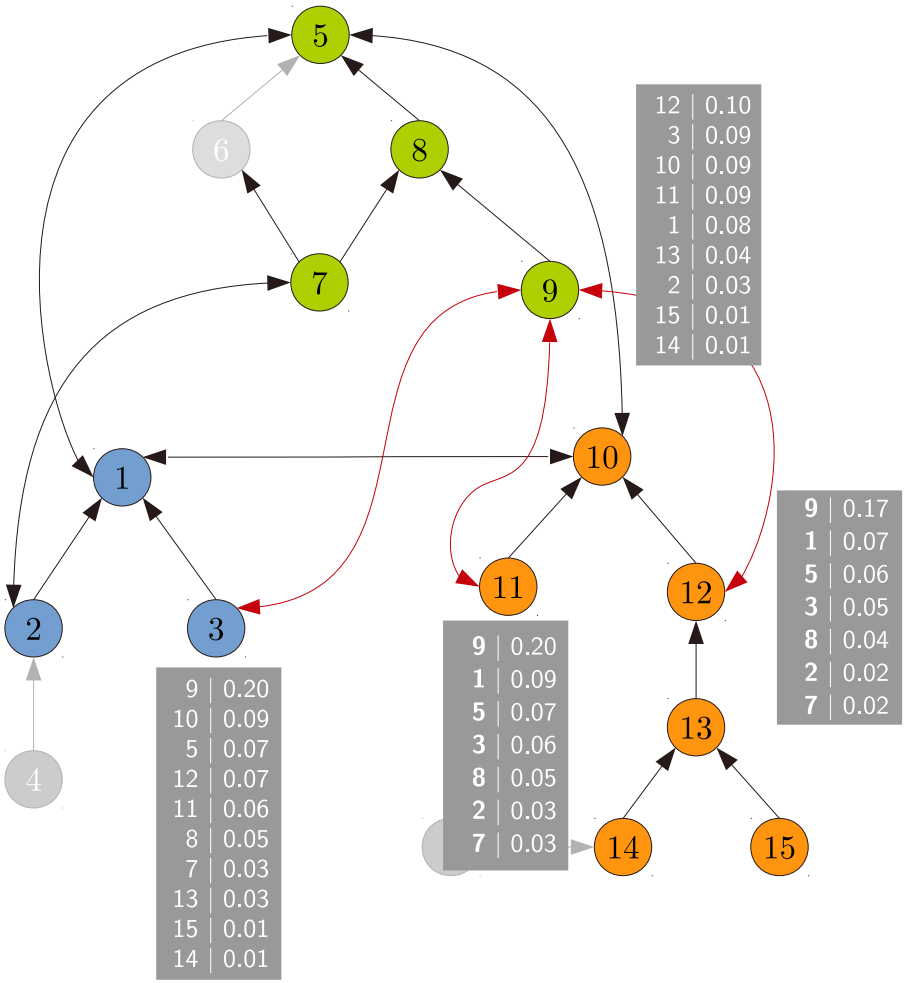
Personalized PageRank I



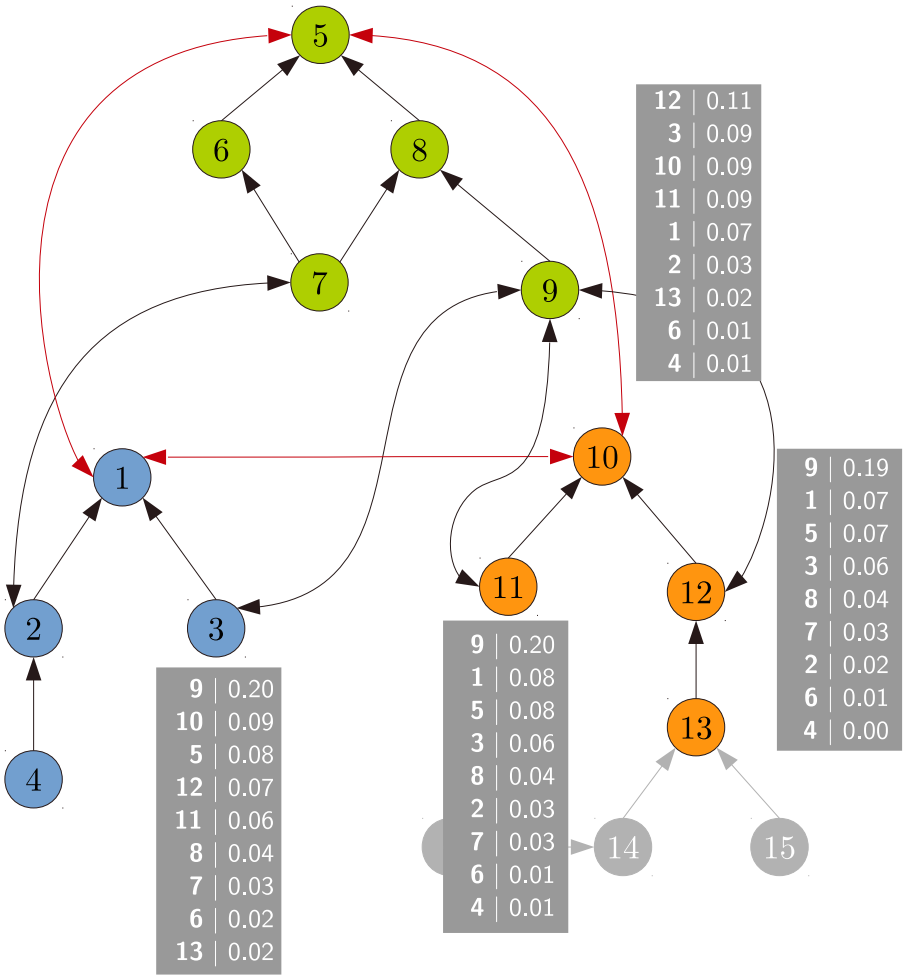
Personalized PageRank I



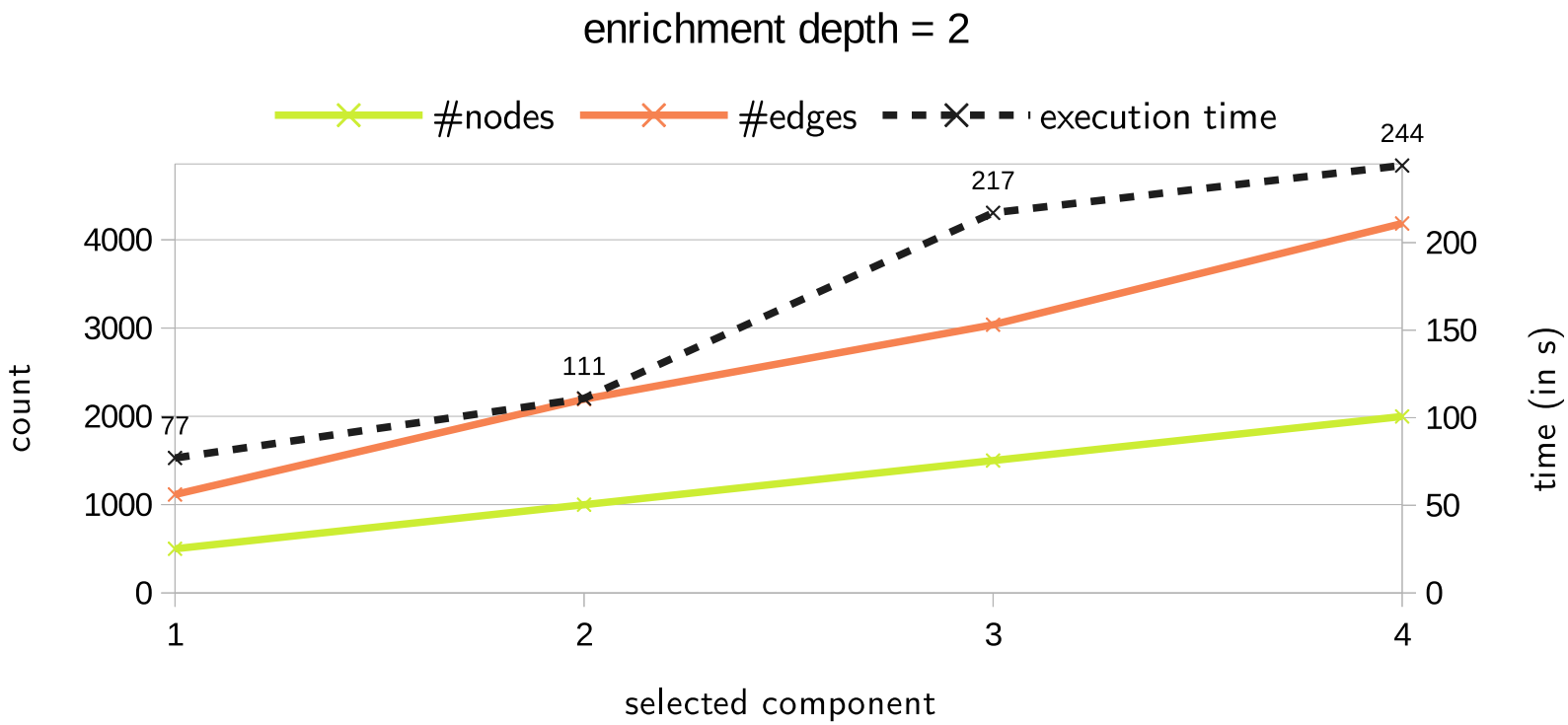
Personalized PageRank I



Personalized PageRank I



Personalized PageRank II



Zusammenfassung

- Gelly ist ein komfortables Framework
- Anreicherung wirkt sich exponentiell auf Graphgröße aus
- PPR wurde mittels *vertex centric iteration* implementiert
- noch offen:
  - Evaluation verschiedener Settings, z.B. Kantengewichte
  - Heuristiken zur Interpretation von PPR-Ergebnissen

Literatur

[BCX11] Bahman Bahmani, Kaushik Chakrabarti, and Dong Xin. Fast personalized pagerank on mapreduce. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 973–984. ACM, 2011.

[Gel] Gelly. Gellyschool. *http://gellyschool.com/index.html*. abgerufen am 26.02.16, 15:00 Uhr.

[PBMW99] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[WL08] Nicole Washington and Suzanna Lewis. Ontologies: Scientific Data Sharing Made Easy. *Nature Education*, 1(3), 2008.