Graph similarity algorithms

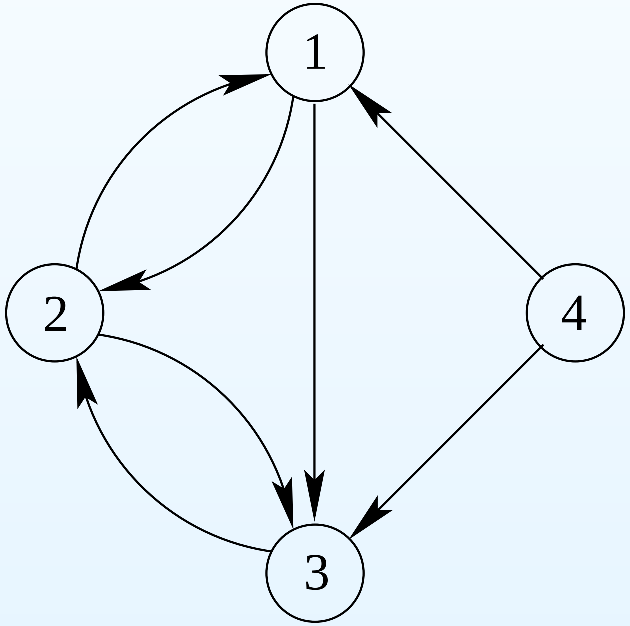
Laure Ninove

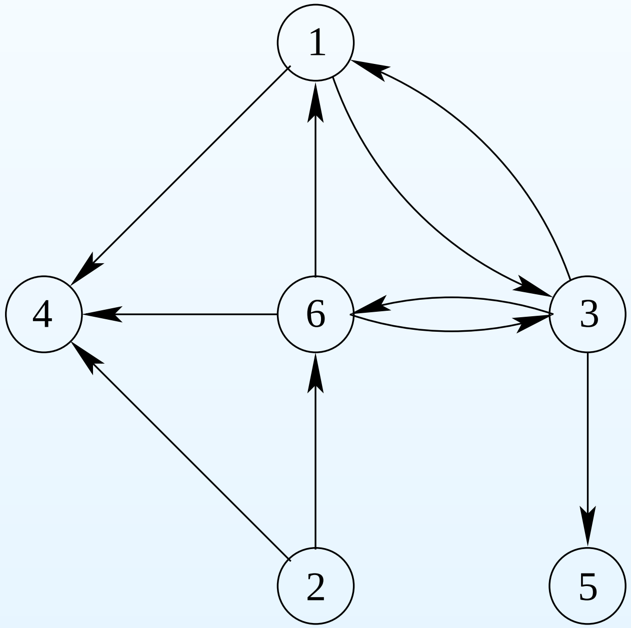
Department of mathematical engineering

Université catholique de Louvain

Similarity scores between nodes of graphs

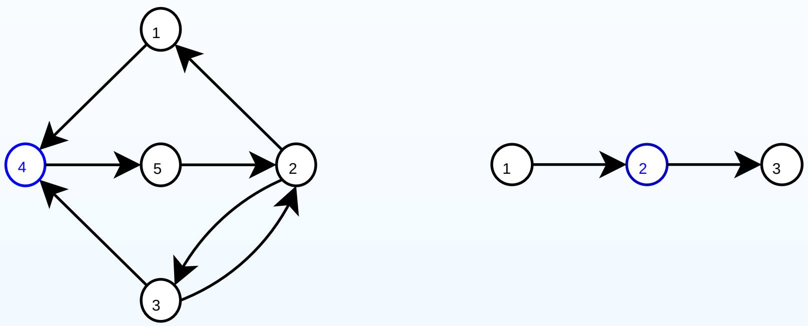
How can we compare the nodes of two graphs?





Similarity of nodes

Two nodes will be similar if they have similar in/out neighbors



- Similarity score between 4 of G_A and 2 of G_B:

$$s(a_4, b_2) \leftarrow s(a_1, b_1) + s(a_3, b_1) + s(a_5, b_3).$$

- Simultaneous iterative computation of the scores of all the pairs.
- The score of a pair is reinforced by the scores of its “neighbors pairs”.

[Blondel, Gajardo, Heymans, Senellart, Van Dooren 2004]
[Melnik,Garcia-Molina,Rahm 2002]

Computation of the similarity scores

- Let A and B be the adjacency matrices of G_A and G_B.
- Let S be the similarity matrix:

$$S = \begin{pmatrix} s(a_1, b_1) & \cdots & s(a_n, b_1) \\ \vdots & & \vdots \\ s(a_1, b_m) & \cdots & s(a_n, b_m) \end{pmatrix}.$$

- S is computed iteratively:

$$S \leftarrow \frac{BSA^T + B^T SA}{\|BSA^T + B^T SA\|}.$$

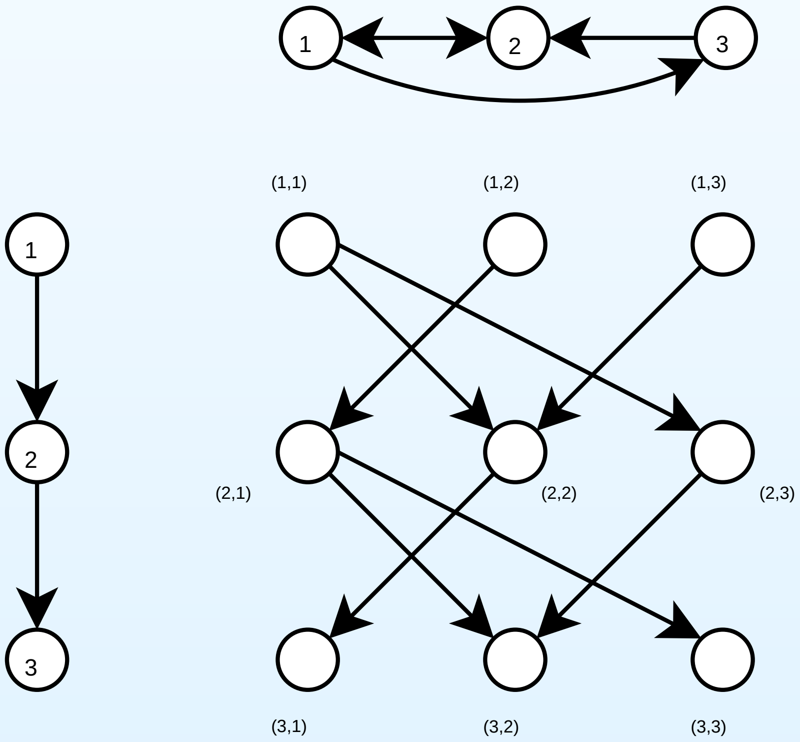
- Convergence concerns: see later.

Computation of the similarity scores

Propagation of scores in the product graph

- Let $G_A \times G_B$ be the product graph and $A \otimes B$ its adjacency matrix.
- The similarity scores are propagated from pair to pair in $G_A \times G_B$.

$$s \leftarrow \frac{(A \otimes B + A^T \otimes B^T)s}{\|(A \otimes B + A^T \otimes B^T)s\|}$$



Computation of the similarity scores

Convergence concerns

The iteration $s_{k+1} = \frac{(A \otimes B + A^T \otimes B^T) s_k}{\|(A \otimes B + A^T \otimes B^T) s_k\|}$ does not always converge!

- **One solution** [Blondel et al. 2002]
  - $A \otimes B + A^T \otimes B^T$ is symmetric
    $\Rightarrow$ each of the subsequences $\{s_{2k}\}_k$ and $\{s_{2k+1}\}_k$ converges.
  - Let $s_{\text{even}}(s_0)$ and $s_{\text{odd}}(s_0)$ these limits.
  - The limit $s_{\text{even}}(\mathbf{1})$ has a nice maximizing property.
  - $s_{\text{even}}(\mathbf{1}) = \lim_{k \to \infty} \dfrac{(A \otimes B + A^T \otimes B^T)^{2k} \mathbf{1}}{\|(A \otimes B + A^T \otimes B^T)^{2k} \mathbf{1}\|}$
    is chosen as the similarity scores vector.

Computation of the similarity scores

Convergence concerns

- **Another solution** [Melnik et al. 2004]
  - Change the iteration formula for

$$s_{k+1} = \frac{(A \otimes B + A^T \otimes B^T)s_k + d}{\|(A \otimes B + A^T \otimes B^T)s_k + d\|}.$$

  - Convergence OK for $d > 0$. [Krause U. 1986]
  - If $d = \varepsilon \mathbf{1}$ then $s_* \approx \dfrac{s_{\text{even}}(\mathbf{1}) + s_{\text{odd}}(\mathbf{1})}{2}$.

Computation of the similarity scores

Convergence concerns

- **Another solution** [Melnik et al. 2002]
  - Change the iteration formula for

$$s_{k+1} = \frac{(A \otimes B + A^T \otimes B^T)s_k + d}{\|(A \otimes B + A^T \otimes B^T)s_k + d\|}.$$

  - The similarity vector s_* is the solution of

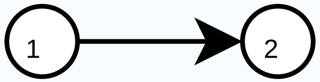
$$\rho(A + dc_*^T)s_* = (A + dc_*^T)s_*$$

with $c_* = \arg\max \rho(A + dc^T)$ on $\{c \geq 0 : \|c\|^D = 1\}$.

[Blondel, N., Van Dooren]

Applications

Hub and authority scores for web searching

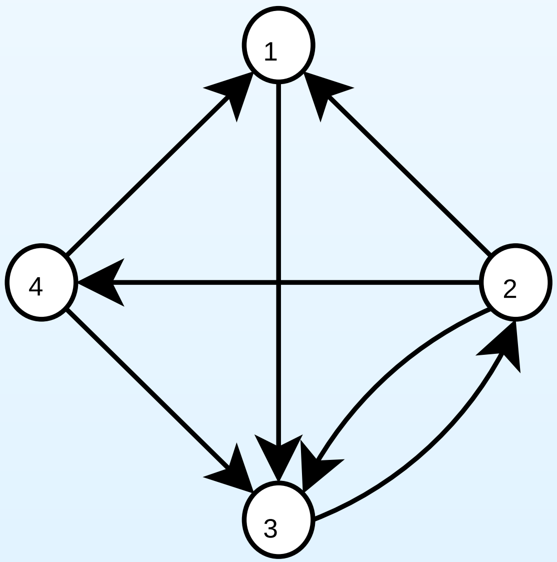
- If G_A is the graph 

the similarity scores give the hub and authority scores.

[Kleinberg 1999]

[Blondel et al. 2004]

- Hub score of a node of G_B = similarity score with node 1 of G_A
- Authority sc. of a node of G_B = similarity sc. with node 2 of G_A

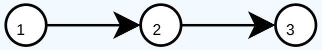


$$S = \begin{pmatrix} (hub) & (auth) \\ 0.2319 & 0.4179 \\ 0.5211 & 0.0000 \\ 0.0000 & 0.5211 \\ 0.4179 & 0.2319 \end{pmatrix}$$

Applications

Synonym extraction and matching of two relational schemas

Some applications of the similarity score:

- Automatic extraction of synonyms:
  - G_A is the graph 
  - G_B a graph constructed from a dictionary.

[Senellart, Blondel 2003]
[Blondel et al. 2004]

- Matching elements of two data schemas:
  - transform the databases in graphs,
  - compute the similarity scores,
  - try to find a good matching.

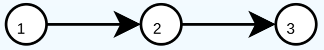
[Melnik et al. 2002]

Examples

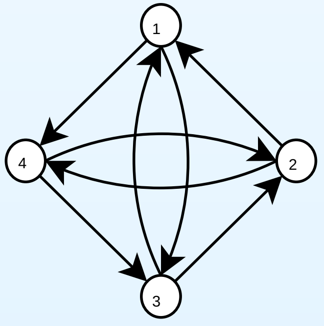
Self similarity

Compare nodes of a graph with nodes of the same graph:

- Path graph: S is diagonal



- Cycle or regular graph: all entries of S are equal



Some limitations

This definition of similarity is still not totally satisfactory

- *Self similarity:*
  the similarity matrix is not always diagonally dominant.

- Similarity matrix does not allow *global comparison* of two graphs.

References

- V. D. Blondel, A. Gajardo, M. Heymans , P. Senellart, and P. Van Dooren, *A measure of similarity between graph vertices: Applications to synonym extraction and web searching*, SIAM Rev. **46** (2004), no. 4, 647–666.
- V. D. Blondel, L. Ninove, and P. Van Dooren, *An affine eigenvalue problem on the nonnegative orthant*, to appear in Linear Algebra and its Applications.
- U. Krause, *A nonlinear extension of the Birkhoff-Jentzsch theorem*, J. Math. Anal. Appl. **114** (1986), no. 2, 552–568.
- S. Melnik, H. Garcia-Molina, and E. Rahm, *Similarity flooding: A versatile graph matching algorithm and its application to schema matching*, Proc. 18th ICDE Conf., 2002.
- P. Senellart and V. D. Blondel, *Automatic discovery of similar words*, ch. 2 in Survey of Text Mining. Clustering, classification, and retrieval, Michael Berry (Ed), pp. 25–44, Springer-Verlag, 2003.

Acknowledgements to FNRS

Large Graphs and Networks
UCL Université catholique de Louvain