My Favorite Algorithm

Fabian Hadiji

Lehrstuhl für Künstliche Intelligenz, TU Dortmund

DSDay7, Berlin, October 30, 2014

Joint work with Kristian Kersting

My Favorite Algorithm?

Desirable Properties:

My Favorite Algorithm?

Desirable Properties:

- Should be widely applicable!

My Favorite Algorithm?

Desirable Properties:

- Should be widely applicable!
- Should scale well!

My Favorite Algorithm?

Desirable Properties:

- Should be widely applicable!
- Should scale well!
- Should have theoretic guarantees!

My Favorite Algorithm?

Desirable Properties:

- Should be widely applicable!
- Should scale well!
- Should have theoretic guarantees!
- Should need only a few lines of code!

My Favorite Algorithm?

Desirable Properties:

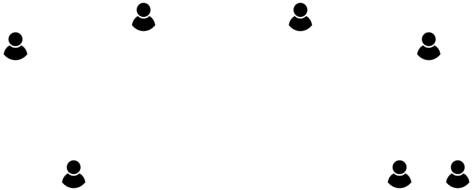
- Should be widely applicable!
- Should scale well!
- Should have theoretic guarantees!
- Should need only a few lines of code!

Label Propagation

[Zhu and Ghahramani, 2002, Zhu et al., 2003]

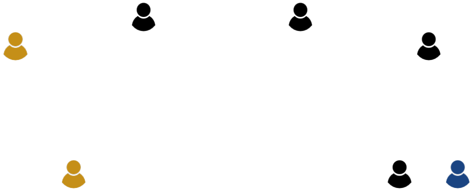
Label Propagation — Intuition

- Set of nodes



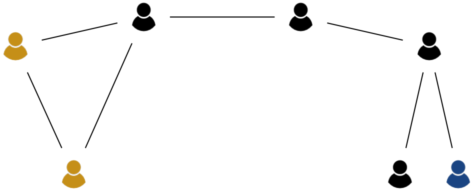
Label Propagation — Intuition

- Set of nodes
- Set of known labels



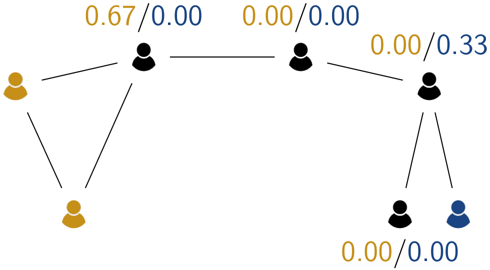
Label Propagation — Intuition

- Set of nodes
- Set of known labels
- Similarity function
  - e.g. $\exp\left(-\sum_d \frac{(x_{id}-x_{jd})^2}{\sigma_d^2}\right)$



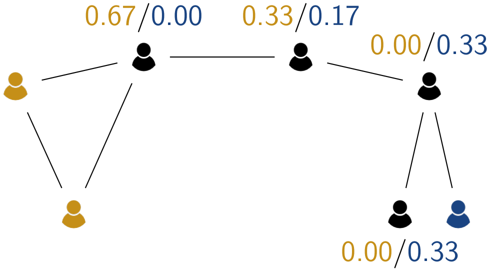
Label Propagation — Intuition

- Set of nodes
- Set of known labels
- Similarity function
  - e.g. $\exp\left(-\sum_d \frac{(x_{id}-x_{jd})^2}{\sigma_d^2}\right)$
- Iteratively propagate labels



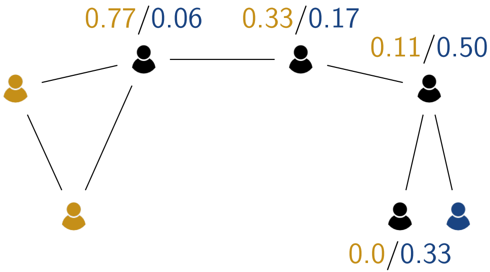
Label Propagation — Intuition

- Set of nodes
- Set of known labels
- Similarity function
  - e.g. $\exp\left(-\sum_d \frac{(x_{id}-x_{jd})^2}{\sigma_d^2}\right)$
- Iteratively propagate labels



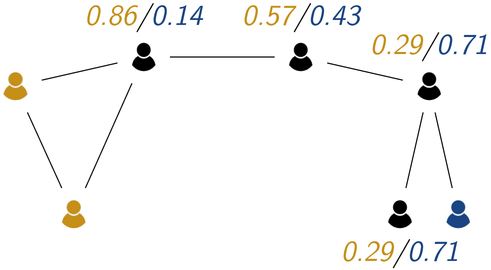
Label Propagation — Intuition

- Set of nodes
- Set of known labels
- Similarity function
  - e.g. $\exp\left(-\sum_d \frac{(x_{id}-x_{jd})^2}{\sigma_d^2}\right)$
- Iteratively propagate labels



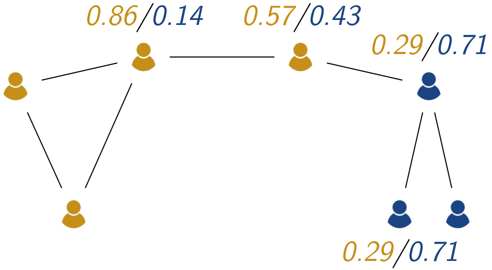
Label Propagation — Intuition

- Set of nodes
- Set of known labels
- Similarity function
  - e.g. $\exp\left(-\sum_d \frac{(x_{id} - x_{jd})^2}{\sigma_d^2}\right)$
- Iteratively propagate labels



Label Propagation — Intuition

- Set of nodes
- Set of known labels
- Similarity function
  - e.g. $\exp\left(-\sum_d \frac{(x_{id}-x_{jd})^2}{\sigma_d^2}\right)$
- Iteratively propagate labels
- Read off labels



Python Code

```
# W is similarity matrix
# Y is label matrix
W = preprocess(W, Y)
Y_old = Y.copy()
iters = 0
while True:
    Y = W * Y
    max_diff = np.abs(Y-Y_old).max()
    iters += 1
    if max_diff < th:
        break
    Y_old = Y.copy()
```

GeoDBLP

DBLP enriched with geo-locations

- DBLP[1] is a bibliography database with $\approx$ 1.5M authors and $\approx$ 2.8M papers

[1] http://dblp.uni-trier.de/

GeoDBLP

DBLP enriched with geo-locations

- DBLP[1] is a bibliography database with $\approx$ 1.5M authors and $\approx$ 2.8M papers
- Unfortunately, DBLP does not contain affiliations/geo-locations

[1] `http://dblp.uni-trier.de/`

GeoDBLP

DBLP enriched with geo-locations

- DBLP[1] is a bibliography database with $\approx$ 1.5M authors and $\approx$ 2.8M papers
- Unfortunately, DBLP does not contain affiliations/geo-locations
- Obtaining seed affiliations/geo-locations is possible but challenging

[1] `http://dblp.uni-trier.de/`

GeoDBLP

DBLP enriched with geo-locations

- DBLP[1] is a bibliography database with $\approx$ 1.5M authors and $\approx$ 2.8M papers
- Unfortunately, DBLP does not contain affiliations/geo-locations
- Obtaining seed affiliations/geo-locations is possible but challenging
- Application: label $\approx$ 5 million author-paper-pairs form DBLP with one of $\approx$ 4.5 thousand cities

[1] `http://dblp.uni-trier.de/`

GeoDBLP

DBLP enriched with geo-locations

- DBLP[1] is a bibliography database with $\approx$ 1.5M authors and $\approx$ 2.8M papers
- Unfortunately, DBLP does not contain affiliations/geo-locations
- Obtaining seed affiliations/geo-locations is possible but challenging
- Application: label $\approx$ 5 million author-paper-pairs form DBLP with one of $\approx$ 4.5 thousand cities

$$W \cdot Y = \underbrace{\begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{pmatrix}}_{5M \times 5M} \cdot \underbrace{\begin{pmatrix} y_{11} & \cdots & y_{1k} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nk} \end{pmatrix}}_{5M \times 4.5k}$$

[1] http://dblp.uni-trier.de/

Large-scale Label Propagation

- **Problem**: Impossible to store dense affinity matrix in RAM.
- **Solution**: Use similarity function based on relational formulas [Hadiji et al., 2013]. E.g.:

$$w_{ij} \mathrel{+}= \lambda_d \text{ if } \texttt{author}(i) = \texttt{author}(j) \wedge \texttt{year}(i) = \texttt{year}(j)$$

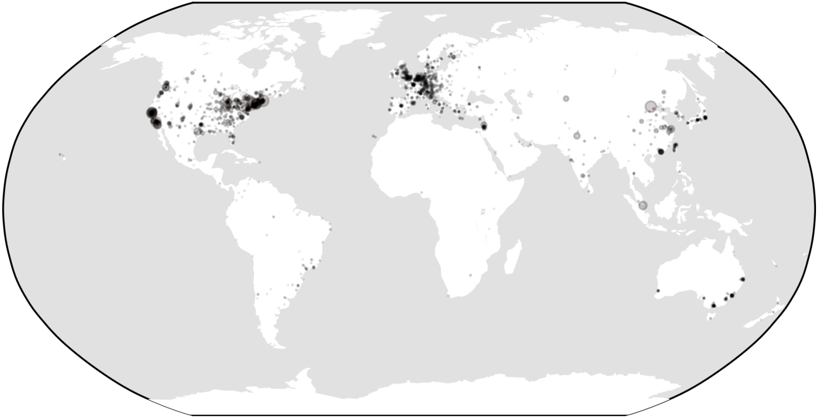
Large-scale Label Propagation

- **Problem**: Impossible to store dense affinity matrix in RAM.
- **Solution**: Use similarity function based on relational formulas [Hadiji et al., 2013]. E.g.:

$$w_{ij} += \lambda_d \text{ if } \texttt{author}(i) = \texttt{author}(j) \land \texttt{year}(i) = \texttt{year}(j)$$

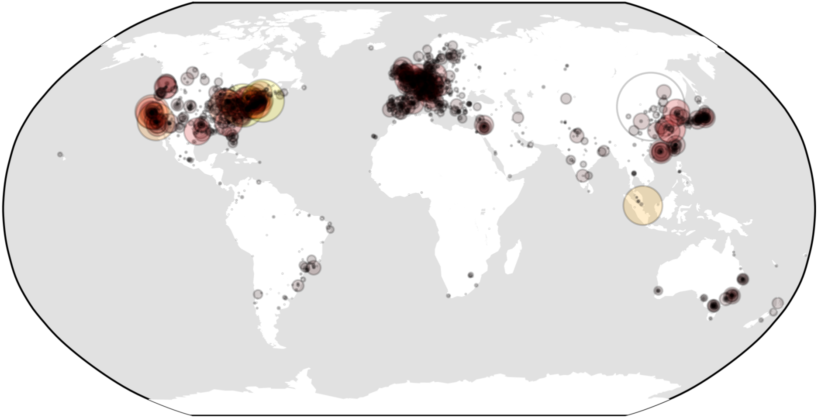
- **Problem**: LP often suffers from slow convergence
- **Solution**: Bootstrapping to speed up convergence [Hadiji and Kersting, 2013]

Propagated Data



Initial Data

Propagated Data



Completed Data

Thank You

Questions ?

www.hadiji.com | @fabianhadiji

References I

Hadiji, F. and Kersting, K. (2013).
Reduce and re-lift: Bootstrapped lifted likelihood maximization for map.
In *AAAI*.

Hadiji, F., Kersting, K., Bauckhage, C., and Ahmadi, B. (2013).
Geodblp: Geo-tagging dblp for mining the sociology of computer science.
arXiv preprint arXiv:1304.7984.

Zhu, X. and Ghahramani, Z. (2002).
Learning from labeled and unlabeled data with label propagation.
Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University.

References II

Zhu, X., Ghahramani, Z., and Lafferty, J. (2003).
Semi-supervised learning using gaussian fields and harmonic functions.
In *ICML*, volume 3, pages 912–919.