University of Stuttgart
Germany

Modeling Sense Structure in Word Usage Graphs with the Weighted Stochastic Block Model

Dominik Schlechtweg Enrique Castaneda Jonas Kuhn Sabine Schulte im Walde

Institute for Natural Language Processing, University of Stuttgart, Germany

Introduction

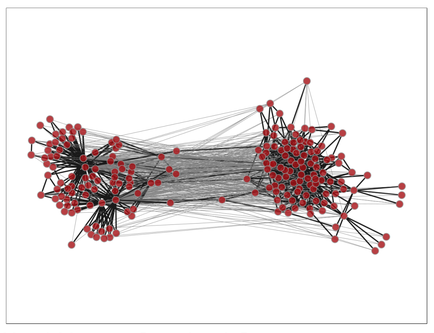
- traditional approach to annotate word senses are binary assignments to sense descriptions [Kilgarriff 1998]
  - manual effort to create sense descriptions
  - ignores gradedness of word meaning
- alternative: pairwise semantic proximity judgments of word use pairs [Erk et al. 2013]
  - use pair judgments populate weighted graph [McCarthy et al. 2016]
  - senses are not annotated directly, but **inferred** on the graph

  → clustering procedure is needed
  - we use the weighted stochastic block model

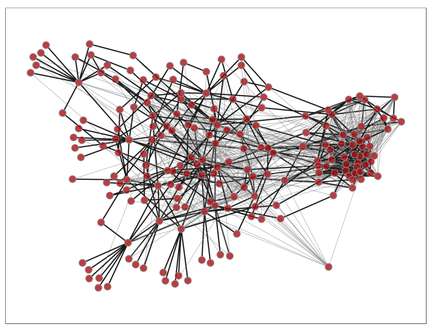
Weighted Stochastic Block Model (WSBM)

- a generative probabilistic model for random graphs [Aicher et al. 2014, Peixoto 2019]
- popular in biology, physics and social sciences
- models nodes as part of blocks (clusters)
- assumes that nodes in the same block are stochastically equivalent
- advantages:
  - allows model selection in absence of ground truth senses
  - captures gradedness by flexible distributions between blocks
  - allows simulation from fitted models
  - extensions allow block (sense) overlap

Data

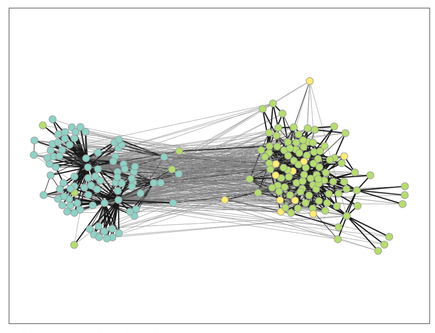


Word Usage Graph of German *zersetzen*.

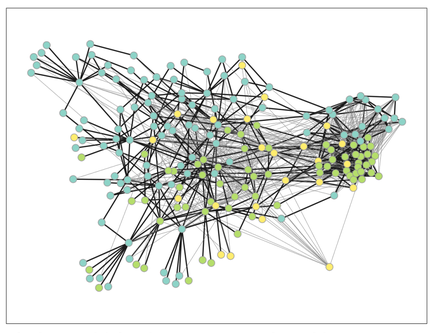


Word Usage Graph of German *Abgesang*.

Inference →



Inferred block structure for *zersetzen*.



Inferred block structure for *Abgesang*.

Find Schlechtweg et al. [2021]'s data at: `https://www.ims.uni-stuttgart.de/data/wugs`

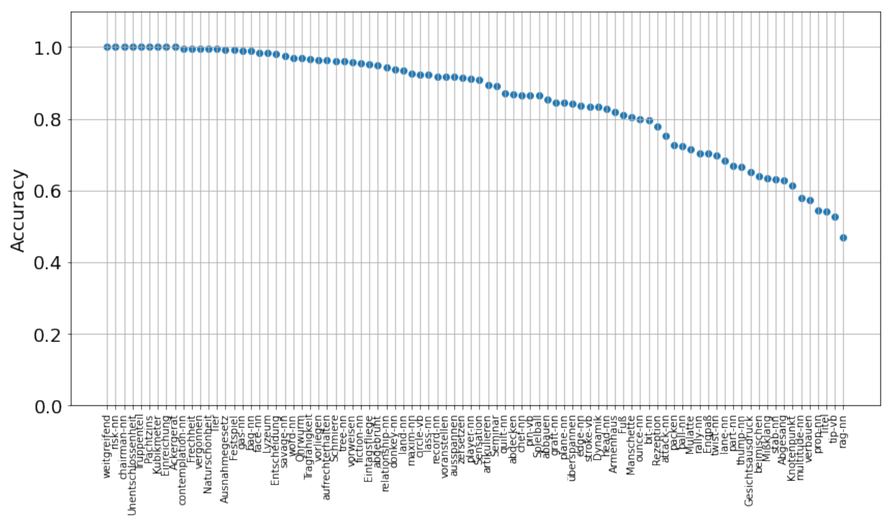
Inference of Block Structure

- we maximize the Bayesian posterior probability

$$P(b|A,x) = \frac{P(x|A,b)P(A|b)P(b)}{P(A,x)}$$

  where b is the inferred block structure, A is the (unweighted) observed graph, and x are the observed edge weights
- approximation: multilevel agglomerative Markov chain Monte Carlo [Peixoto 2014]
- All experiments were done with graph-tool: `https://graph-tool.skewed.de/` [Peixoto 2017]

Correspondence to Independent Clustering



Fitted Edge Weight Distributions

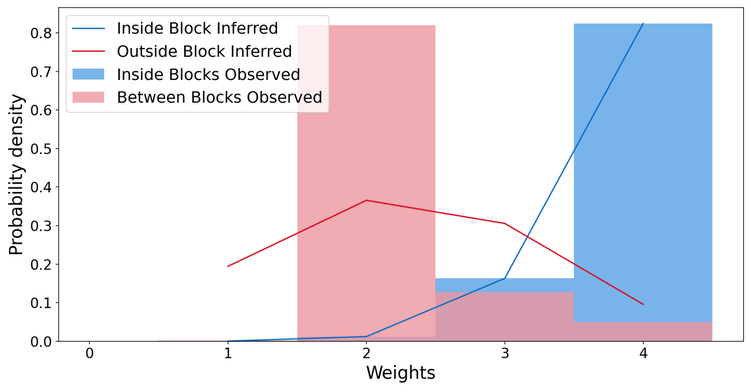


Figure: Fitted (line) and observed (bars) edge weight distributions for *zersetzen*.

Conclusion

- we inferred sense structure on WUGs exploiting patterns of semantic proximity
- model selection allows principled inference of sense structures
- the model can be rigorously compared to other probabilistic models
- the inferred structures mostly reflect intuitive sense distinctions
- structural properties of observed graphs are often not very well preserved

  → more flexible distributions for edge weights are needed
- inferred models can be used for simulation of realistic WUGs: `https://www.ims.uni-stuttgart.de/data/wugs`
- future: do senses overlap? Which model best describes the data?

Acknowledgments

We thank Tiago de Paula Peixoto for advice and for providing the idea and an implementation of the marginalization over edge probabilities. Dominik Schlechtweg was supported by the Konrad Adenauer Foundation and the CRETA center funded by the German Ministry for Education and Research (BMBF) during the conduct of this study.

References

Aicher, C., Jacobs, A. Z., & Clauset, A. (2014, Jun). Learning latent block structure in weighted networks. *Journal of Complex Networks*, *3*(2), 221—248. Retrieved from `http://dx.doi.org/10.1093/comnet/cnu026` doi: 10.1093/comnet/cnu026

Erk, K., McCarthy, D., & Gaylord, N. (2013). Measuring word meaning in context. *Computational Linguistics*, *39*(3), 511–554.

Kilgarriff, A. (1998, aug). Senseval: An exercise in evaluating word sense disambiguation programs. In A. M. A. M. S. T. Thierry Fontenelle Philippe Hiligsmann (Ed.), *Proceedings of the 8th euralex international congress* (pp. 167–174). Liège, Belgium: Euralex.

McCarthy, D., Apidianaki, M., & Erk, K. (2016). Word sense clustering and clusterability. *Computational Linguistics*, *42*(2), 245–275.

Peixoto, T. P. (2014, Jan). Efficient monte carlo and greedy heuristic for the inference of stochastic block models. *Physical Review E*, *89*(1). Retrieved from `http://dx.doi.org/10.1103/PhysRevE.89.012804` doi: 10.1103/physreve.89.012804

Peixoto, T. P. (2017, 08). Nonparametric weighted stochastic block models. *Physical Review E*, *97*. doi: 10.1103/PhysRevE.97.012306

Peixoto, T. P. (2019). Bayesian stochastic blockmodeling. In *Advances in network clustering and blockmodeling* (p. 289-332). John Wiley & Sons, Ltd. doi: 10.1002/9781119483298.ch11

Schlechtweg, D., Tahmasebi, N., Hengchen, S., Dubossarsky, H., & McGillivray, B. (2021). DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages. *CoRR*, *abs/2104.08540*. Retrieved from `https://arxiv.org/abs/2104.08540`