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
{schlecdk,jonas.kuhn,schulte}@ims.uni-stuttgart.de, kicasta@gmail.com

Abstract

We suggest to model human-annotated Word Usage Graphs capturing fine-grained semantic proximity distinctions between word uses with a Bayesian formulation of the Weighted Stochastic Block Model, a generative model for random graphs popular in biology, physics and social sciences. By providing a probabilistic model of graded word meaning we aim to approach the slippery and yet widely used notion of word sense in a novel way. The proposed framework enables us to rigorously compare models of word senses with respect to their fit to the data. We perform extensive experiments and select the empirically most adequate model.

1 Introduction

Word Usage Graphs (WUGs) are a relatively new model of graded word meaning in context (Erk et al., 2013; McCarthy et al., 2016; Schlechtweg et al., 2021). They represent word uses (i.e., words in context) within a weighted undirected graph, with edge weights reflecting the semantic proximity between uses. WUGs may be obtained via human annotation by presenting annotators with pairs of words uses and asking them for proximity judgments. The WUGs may then be clustered into sets of uses exhibiting high semantic proximity, in order to reflect traditional word sense distinctions (McCarthy et al., 2016), and to provide insight into key aspects of word meaning such as polysemy, vagueness, and lexical semantic change (Schlechtweg et al., 2020, 2021).

We suggest to model WUGs with a Bayesian formulation of the Weighted Stochastic Block Model (WSBM), a generative model for random graphs popular in biology, physics and social sciences (Aicher et al., 2014; Peixoto, 2017). The basic assumption of WSBMs is that vertices belong to latent blocks (clusters), and that vertices in the same block are stochastically equivalent (i.e., they have edges drawn from the same distribution). Fitting the model is equivalent to determining the optimal latent block structure providing a **clustering** of word uses.

By using a Bayesian **probabilistic model** of WUG data we aim to approach graded word meaning in a rigorous scientific way: We perform **model selection**, i.e., different models are compared according to their fit to the data, and the model which explains the data best is chosen as most adequate representation of the semantic structure behind human-annotated word uses. If blocks are equated with **word senses**, this allows us to approach this slippery and yet widely used concept in a novel way. We may test long-standing hypotheses such as whether a graded model allowing sense overlap is a better model than a discrete one (Kilgarriff, 1997; Erk et al., 2013; McCarthy et al., 2016).

As a probabilistic model, the WSBM allows to **generate data** from a fitted model, which is useful for simulating realistic WUGs, e.g. when planning annotation studies. A fitted WSBM may also be used to **predict** values of unobserved edge weights, which is helpful for enhancing annotations.

Our contributions can be summarized as follows:

- Introducing a rigorous scientific way to infer the number and the nature of word senses.
- Improving WSBM with marginalizing over edge probabilities.
- Model selection: inferring the most likely number of discrete word senses for words in DWUG DE/EN data sets (Schlechtweg et al., 2021).
- Model checking: validating WSBMs as a reasonable model of WUGs and word senses with respect to external criteria.
- Publication of fitted WSBM models which can be used for simulating realistic data.

Proceedings of the 10th Conference on Lexical and Computational Semantics, pages 241–251
August 5–6, 2021, Bangkok, Thailand (online)

- Analysis: identifying shortcomings of WSBMs (such as edge probabilities, hub effect).

2 Related Work

Our approach generally falls within the area of Bayesian probabilistic modeling (Koch, 2007). More specifically, it is related to model-based graph clustering techniques, e.g., Latent Space models such as Gaussian Mixture Models (Hoff et al., 2002; Duda and Hart, 1973). These methods are common in the field of *community detection* (Abbe, 2017). Within computational linguistics our approach is most strongly related to generative probabilistic topic models, where words in documents are modeled as being drawn from a latent topic distribution (Steyvers and Griffiths, 2007). Topics are often interpreted as senses (Frermann and Lapata, 2016; Perrone et al., 2019). Another common, yet non-probabilistic, modeling approach for word senses is to group word uses expressing similar meanings into clusters based on contextual features (Schütze, 1998; Biemann, 2006).

As to our knowledge, only a small set of studies is concerned with the modeling of *human-annotated* WUGs (McCarthy et al., 2016; Schlechtweg et al., 2020, 2021). This research line is motivated by insights from lexical semantics that word senses are no discrete objects (Kilgarriff, 1997; Erk et al., 2013). Most important to note is the pioneering work of McCarthy et al. (2016) as the first to represent human-annotated word uses within graphs and then clustering the uses based on heuristics such as connected components and cliques. McCarthy et al. derived edge weights from human lexical substitution judgments for the respective target words and binarized them according to a threshold. This idea was recently modified and extended by Schlechtweg et al. (2020, 2021). Schlechtweg et al. used semantic proximity judgments to annotate edges. They applied correlation clustering (Bansal et al., 2004) in connection with a global threshold to group vertices with high edge weights and developed an efficient iterative sampling strategy for edges to reduce annotation load. However, these approaches are ad-hoc clustering methods which do not provide a probabilistic model for WUGs.

3 Data

A Word Usage Graph $\mathbf{G} = (\mathbf{U}, \mathbf{E}, \mathbf{W})$ is a weighted, undirected graph, where vertices $u \in U$ represent word uses and weights $w \in W$ represent the semantic proximity of a pair of uses $(u_1, u_2) \in E$ (Schlechtweg and Schulte im Walde, submitted). In practice, semantic proximity can be measured by human annotator judgments on a scale of relatedness (Brown, 2008; Schlechtweg et al., 2018) or similarity (Erk et al., 2013). Human-annotated WUGs are often sparsely observed and noisy, i.e., only a small percentage of edges from the full graph are annotated, and annotators often show disagreements, e.g. for ambiguous uses, as can be seen in Figure 1.

↑	4: Identical
	3: Closely Related
	2: Distantly Related
	1: Unrelated

Table 1: DURel relatedness scale (Schlechtweg et al., 2018).

Recently, Schlechtweg et al. (2020, 2021) developed a large-scale multi-lingual resource of WUGs. Annotators were asked to judge the semantic relatedness of pairs of word uses (such as the two uses of *grasp* in (1) and (2)) according to the scale in Table 1.[1]

(1) He continued to **grasp**, between forefinger and thumb, the edge of the cloth I had been sewing.

(2) For just a moment he didn't **grasp** the import of what the old man had said.

The uses were sampled from diachronic corpora of four languages (English, German, Latin, Swedish). The data was annotated in four rounds. After each round the accumulated annotations from the previous rounds wer represented in a WUG, which was then clustered with correlation clustering (Bansal et al., 2004), and then further use pairs were chosen according to heuristics aiming to compare uses to clusters to which they had not yet been compared. Annotators showed high agreement, and comparable to previous studies. The final resource consists of WUGs for 168 words with a total of 100,000 judgments including nouns, verbs and adjectives as well as monosemous and polysemous words. In our experiments we use the German and English subparts of the data set comprising 88 WUGs.

While for some WUGs a clustering structure grouping vertices with high edge weights together

[1] https://www.ims.uni-stuttgart.de/data/wugs

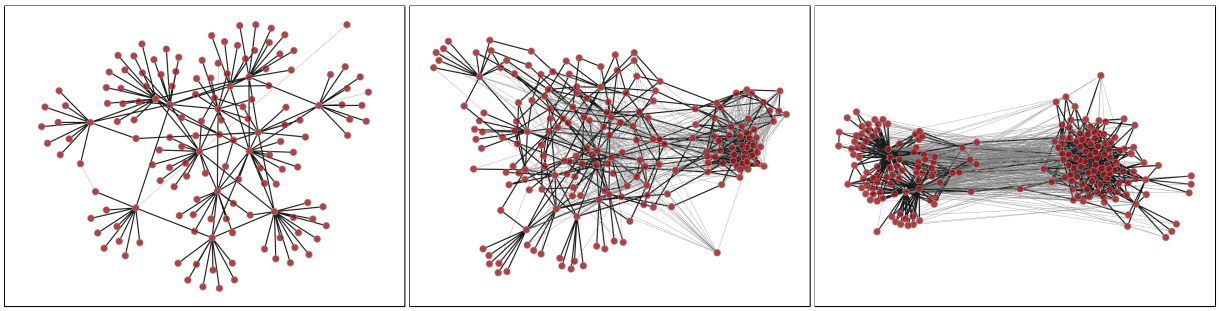

Figure 1: Word Usage Graphs of German *Festspiel* (left), *Abgesang* (middle) and *zersetzen* (right). Vertices represent uses of the respective target word. Edge weights represent the median of relatedness judgments between uses (**black**/gray lines for **high**/low edge weights, i.e., weights ≥ 2.5/weights < 2.5).

is obvious, for others this is not the case (cf. McCarthy et al., 2016). For example, see Figure 1 showing the annotated uses for three words from Schlechtweg et al. (2021).

The uses of the word *Festspiel* on the left and *zersetzen* on the right can be clearly partitioned into one/two main clusters, while the uses of *Abgesang* in the middle have a less clearly clusterable structure. Hence, it is unclear how many senses *Abgesang* has and what the assignment of uses to senses should be. We approach these two questions by searching for the model which best explains the data. The block structure inferred by this model will then give us a number of blocks and an assignment of uses to blocks.

4 Stochastic Block Model

The Stochastic Block Model (SBM) (Holland et al., 1983) is a simple generative process of random graphs based on the notion of groups of vertices. It assumes that each vertex of an observed graph G is member of a latent block (group) and that G was generated by first sampling vertices and then sampling edges between these vertices where the probability of observing an edge between two vertices is only determined by the block membership. Once this process is formulated mathematically, the optimal latent block structure can be inferred from G. For this, given the partition $b = \{b_i\}$ of G into B blocks, where $b_i \in [0, B-1]$ is the block membership of vertex i, we define a model that generates a graph A with a probability

$$P(A|\theta, b)$$

where θ are additional model (edge bundle) parameters that govern how the vertex partition affects the placing of edges (Peixoto, 2014a). Therefore, if we observe a graph A, the likelihood that it was generated by a given partition b is given by the Bayesian posterior probability

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

where $P(\theta, b)$ is the prior probability of the model parameters, and $P(A)$ is called the *evidence*, and corresponds to the total probability of the data summed over all model parameters (Peixoto, 2014a). The standard SBM takes as parameters the partition of the vertices into blocks b and a $B \times B$ matrix of edge counts e, where e_{rs} is the number of edges between groups r and s.

4.1 Edge weights

The Weighted Stochastic Block Model (WSBM) is an extension of the standard SBM to weighted graphs (Aicher et al., 2014; Peixoto, 2017). In the WSBM the inference of the latent block structure is driven by both edge existence and edge weights. This is achieved by treating edge weights as covariates that are sampled from some distribution (e.g. binomial) conditioned on the vertex partition (Peixoto, 2014a), i.e.,

$$P(A, x|\theta, \gamma, b) = P(x|A, \gamma, b)P(A|\theta, b)$$

with the covariates being sampled only on existing edges, and where $\boldsymbol{\gamma}_{rs}$ is a set of parameters that govern the sampling of the weights between groups r and s. The posterior partition distribution is then

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

omitting the parameters θ, γ as in the non-parametric WSBM through the use of marginal likelihoods (Peixoto, 2017). In our experiments we use the non-parametric, micro-canonical implementation of the WSBM which avoids explicitly

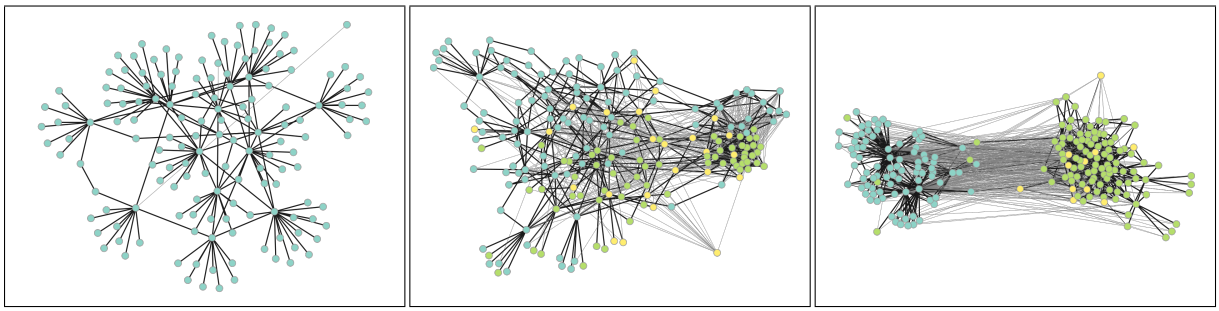

Figure 2: Word Usage Graphs of German *Festspiel* (left), *Abgesang* (middle) and *zersetzen* (right) with inferred block structure.

encoding distribution parameters for edge weights by replacing them with hard quantities (Peixoto, 2014c).[2] The non-parametric model avoids overfitting, and micro-canonical distributions are easier to compute while approaching their canonical counterparts asymptotically (Peixoto, 2017).[3]

4.2 Marginalizing over edge probabilities

The basic assumption of the WSBM is that vertices in the same block are stochastically equivalent. This should hold with respect to edge weights x and edge probabilities A. However, the distribution of edge probabilities in our case is guided exclusively by Schlechtweg et al.'s sampling procedure. Hence, the assumption of stochastic equivalence of edge probabilities does not hold for WUGs. Thus, we aim to make the block structure independent from the observed edge probability distribution between blocks as far as possible.[4] We reach this by marginalizing over edge probabilities, while keeping their number the same between groups. The latter is needed as the edge probabilities build the support of the edge weights. The posterior partition distribution is then

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

where

$$P(x|b) = \sum_{A \in \Lambda} P(x|A,b)P(A|b)$$

and Λ is the set of all networks A that have the same number of edges between groups as the observed network A' under block assignment b. We sum over all possible edge assignments with the same number of edges between groups. In this way edge probabilities are marginalized and the posterior distribution $P(x|b)$ is mainly driven by edge weights.

4.3 Inference

Finding the maximum of the posterior distribution of the WSBM is NP-hard (Peixoto, 2015). Hence, we infer the optimal partitioning of vertices $P(b|x)$ asymptotically with multilevel agglomerative Markov chain Monte Carlo Peixoto (2014b). The central idea is to sample from $P(b|x)$ by first starting from some initial state and making move proposals depending on the current state such that, ultimately, the Markov Chain converges to $P(b|x)$. In order to alleviate the problem of metastable states the chain is first equilibrated for a larger number of blocks, which are then merged. (Find a discussion of the problem of metastable states in Peixoto (2014b).)

4.4 A Model for Word Senses?

The basic assumption of the WSBM with marginalized edge probabilities is that vertices in the same block are stochastically equivalent with respect to edge weights. We argue that this assumption is reasonable for word senses: From previous work we inherit the insight that graded proximity judgments reflect single-sense judgments (Erk et al., 2013; McCarthy et al., 2016). This is to say that use pairs expressing the same sense receive high values on

[2] We recover the non-microcanonical versions of the distributions by fitting these to the observed edge weights between blocks after fitting the WSBM.

[3] All experiments were done with graph-tool: `https://graph-tool.skewed.de/`. Additional code is provided at `https://github.com/kicasta/Modeling_WUGS_WSBM`.

[4] Note that degree-correction relaxes the homogeneity assumption and would thus serve as a first modeling approach (Karrer and Newman, 2011; Peixoto, 2019). However, the degree-corrected model still suffers from the hub effect, i.e., vertices with many edges tend to be assigned to the same block (Peixoto, 2020). This effect could be avoided with Latent Poisson models (Peixoto, 2020). However, we want the inferred block structure to be largely independent from edge probabilities, which neither of the models fully guarantees.

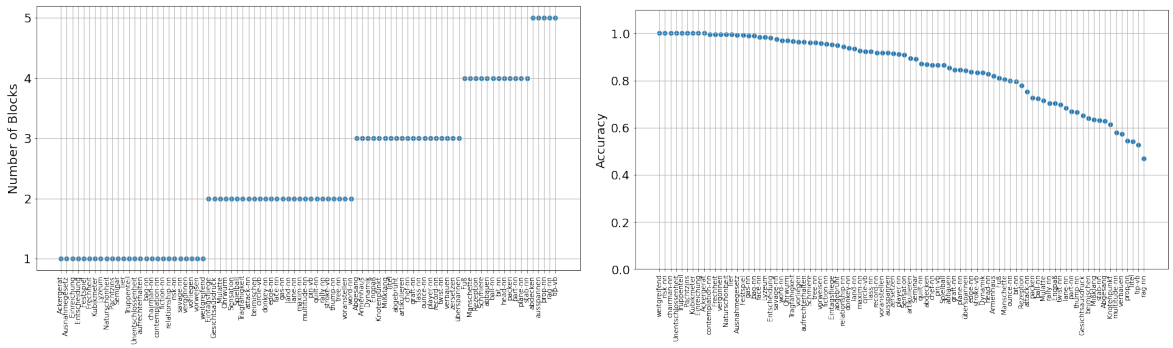

Figure 3: Inferred number of blocks with best-fitting models (left). Correspondence to clustering result from Schlechtweg et al. (2021) (right).

the annotation scale, while use pairs expressing different senses receive low values. This behavior can be modeled by assuming that same-sense pairs receive edge weights chosen from a common distribution with a high mean, while the same holds for different-sense pairs with a distribution with a low mean. The interesting question is, though, how well the WSBM (or any other model) can model the unclear cases, i.e., use pairs receiving intermediate judgments on the annotation scale. The WSBM is a very general model that can learn many different structures. It can handle heterogenous and overlapping edge weight distributions and also allows blocks to be more or less related to each other. In principle, it also allows mixed membership of vertices in blocks (Peixoto, 2015). The advantage of our approach is that we do not have to define senses in any further way. As latent variables, they can be found by themselves, guided by the independent criterion of how well they explain the data.

Note that this approach does not in any way depend on the concept of sense. In principle, any other probabilistically formulated model aiming to explain WUG data can be introduced. Such a model does not have to rely on the idea of stochastically equivalent blocks. If this model were to explain the data better, the WSBM could be neglected.

5 Model Selection[5]

Following Peixoto (2015) we select the best model according to the Minimum Description Length Principle (Grünwald and Grunwald, 2007). The description length of a graph measures the amount of information required to describe the data, if we encode it using a particular parametrization of the model being tested. This approach corresponds to an implementation of Occam's razor, where the simplest model is selected, among all possibilities with the same explanatory power (Peixoto, 2014a).

5.1 Number of blocks

88 WUGs were fitted using three different distributions for edge weights (see below). The optimal number of blocks is found during fitting (Peixoto, 2014b). We start fitting by choosing an initial number of blocks $1 \leq b \leq 30$. Peixoto's algorithm then tries to find a partition of the Graph into $1 \leq b \leq 30$ blocks with minimum description length. It does so by choosing some $b' > b$, finding the best partition of the graph into b' blocks and then greedily merging these b' into b blocks. Then, it repeats this step for a b_1 and b_2, such that $b_1 < b < b_2$ and decides whether it should increase the number of blocks or decrease it depending on whether it results in a decrease in description length. This is done until convergence. Figure 3 (left) shows the optimal number of blocks obtained for each WUG in the above-described way. We see a tendency to favor simpler structures over more complex ones. That is, most WUGs are modeled best with one or two blocks. The highest number of blocks found is 5. The inferred block structure for the three graphs in Figure 1 is displayed in Figure 2 with 1/3/3 blocks respectively.[6]

5.2 Edge Weight Distribution

Each WUG was fitted using three different distributions for edge weights: (micro-canonical versions of) binomial, poisson and geometric. Figure 4 (left) shows the number of graphs for which each dis-

[5]We provide all fitted models as well as our code at https://github.com/kicasta/Modeling_WUGS_WSBM.

[6]For further example plots see Figures 7 and 8 in Appendix A.

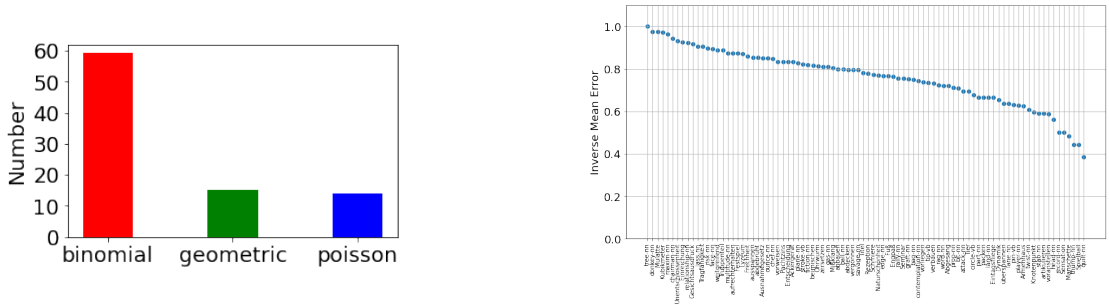

Figure 4: Comparison of model fit wrt. edge weight distributions (left), where Y-axis gives number of graphs for which the respective distribution had minimum description length. Evaluation result of link prediction (right).

tribution type yielded the best fit. The binomial distribution shows the best fits in the large majority of cases. This makes sense, because it is also the most general and flexible of the three distributions. Figure 5 shows the observed edge weight distributions of the graphs between blocks after fitting (red) and within blocks (blue), as well as the respective inferred distributions (curve). Despite the fact that edge weight distributions may be heterogenous (middle), there is a clear tendency for negative edges between blocks and positive edges within blocks. The inferred distributions reflect this pattern.[7]

5.3 Analysis

We now take a closer look at the WUGs from Figures 1 and 2 and their block-related edge weight distributions in Figure 5. For example, the following two use pairs of *Festspiel* homogeneously received high ratings of 3 and 4. The best fit is reached with one block and a binomial distribution.

(3) ...war die DDR bei den Wiener Festwochen, den Salzburger **Festspielen** und...
'*...the GDR was represented at the Wiener Festwochen, the Salzburg **Festival** and...*'

(4) ...im Rahmen der Wettbewerbe und **Festspiele** der Volkskunst...
'*...as part of the competitions and **festivals** of folk art...*'

Abgesang is a different case: It received heterogenous judgments across the scale from 1–4. No clear block structure is visible at first in Figure 1. The best fit is obtained with three blocks and a binomial distribution. The three blocks reflect meaningful fine-grained sense differences as displayed by the following three examples:

(5) In den ersten Strophen der Klage der Ceres findet sich ein [...] 4 zeiliger Aufgesang mit einem 8 zeiligen **Abgesang**.
'*In the first stanzas of Ceres' lament there is a [...] 4-line stance start with an 8-line **stance end**.*'

(6) ...und radelte unter dem **Abgesang** schmutziger Lieder davon.
'*...and cycled off while **singing** dirty songs.*'

(7) ...daß dieser Vorgriff auf den Sommer nicht schon den **Abgesang** des Wintersports bedeutet...
'*...that this anticipation of summer doesn't mean the **swan song** of winter sports...*'

We observe that the sparsity of the annotation has a strong influence: if a word use is richly annotated with several edges, then the model has information on its relation to other blocks and can infer a reasonable block assignment, even if there are annotation errors. If, however, the use is only annotated with e.g. one low-valued edge, the model is likely to assign it to a block with semantically very different uses which also tend to have low judgments with other uses. That is, unrelated uses may appear homogeneous to the model, because they have similar (as sparsely observed) relations to third uses. This effect disappears with richer annotation.

Just as *Abgesang*, *zersetzen* yields the best fit with three blocks and a binomial distribution. As can be seen in Figure 5, the weights also cover the whole scale. However, in this case they are more homogeneously distributed within and between blocks. This is because *zersetzen* has two main and clearly distinguishable senses, as illustrated by (8) and (9):

[7] Note that Figure 5 shows only the combined distributions within and between blocks across all combinations. The per-block distributions are very similar though, as can be seen in Figure 9 in Appendix A for *zersetzen*.

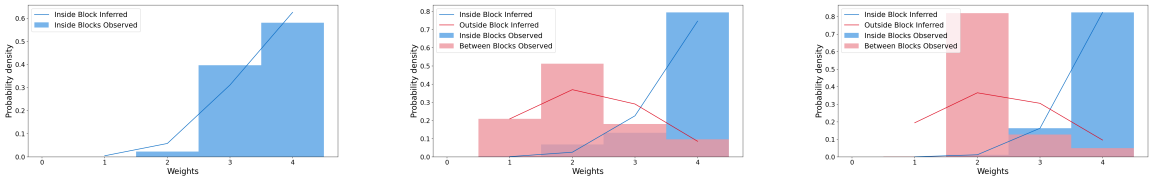

Figure 5: Combined weight distributions of German *Festspiel* (left), *Abgesang* (middle) and *zersetzen* (right). The curves show the inferred distributions.

(8) ...dass die Pflanzen das kohlensaure Gas beym Sonnenlichte **zersetzen**...
*'...that the plants **decompose** the carbonic acid gas in the sunlight...'*

(9) Das System des Frontstadtsenats hat die westberliner Schule bedrohlich **zersetzt**.
*'The system of the front city senate has **destroyed** the West Berlin school.'*

There is a third block where uses are mostly variations of the sense in (9), e.g. referring to a rather physical than chemical decomposition, or uses where the meaning is unclear. We made a similar observation for other graphs (e.g. *rag*): There are separate and semantically heterogeneous blocks for unclear and sparsely annotated uses. German *zersetzen* also illustrates an effect that we observe across many graphs: While the inferred binomial distribution within blocks (see Figure 5) can be closely fit, the distribution between blocks has a considerable error. This is mostly because weights of 1 are rare, while weights of 4 are common. The probability mass of weights between blocks is concentrated at 2, not 1. The binomial distribution has considerable problems modeling this behavior. Other distributions are also deficient, however: the geometric distribution cannot model right-skewed distributions at all, and thus has high errors for within-block distributions. Consequently, the cases where it yields the best fits, are the ones with a high number of low edge weights (e.g. *tip*) which lead to strongly left-skewed distributions. The poisson distribution suffers from the problem that it cannot model steep and peeked distributions. An important challenge for future modeling approaches will be to find appropriate distributions to model the behavior of edge weights. We believe that a signed (invertible) geometric distribution will yield good fits in many cases. Another important challenge will be to avoid sparsity of annotation, e.g. by developing efficient and iterative sampling techniques for edges. It also should be examined how much the inferred block structure is influenced by the difference in the way a particular annotator interprets uses, yielding homogeneous judgments for edges annotated exclusively by this annotator. This could be modeled by multi-graph models (Peixoto, 2017) where the information from each annotator can be represented individually.

6 Model Checking

In order to validate the fitted models externally we test whether the inferred clustering corresponds to a clustering obtained with an independent algorithm. Additionally, we use two internal validation criteria which test how well the structural properties of the observed graphs are recoverable from the inferred models. For this we apply two strategies: (i) Posterior Predictive Checking (Gelman et al., 2013) and (ii) Link Prediction (Liben-Nowell and Kleinberg, 2007).

6.1 Correspondence to Independent Clustering Algorithm

Figure 3 (right) shows the correspondence (accuracy) of the inferred block structures to those found by Schlechtweg et al. (2021) with correlation clustering and a global threshold on edge weights. The results often show strong correspondence ($> .9$) to Schlechtweg et al., although they were obtained with a completely different approach. For a number of graphs with one inferred block the structures are exactly the same. However, there are also clear differences: Especially for graphs with complex block structures (e.g. *tip* or *rag*) the correspondence to Schlechtweg et al. is very low. This also holds for some cases with simpler block structure (e.g. *Gesichtsausdruck* or *multitude*). Our three graphs from Figure 1 nicely display this pattern: *Festspiel* has a simple one-block structure and high accuracy (≈ 1.0), while *Abgesang* has a complex structure and low accuracy (< 0.6). *zersetzen* has three blocks (as *Abgesang*), but two main and clearly separated blocks and rather high accuracy.

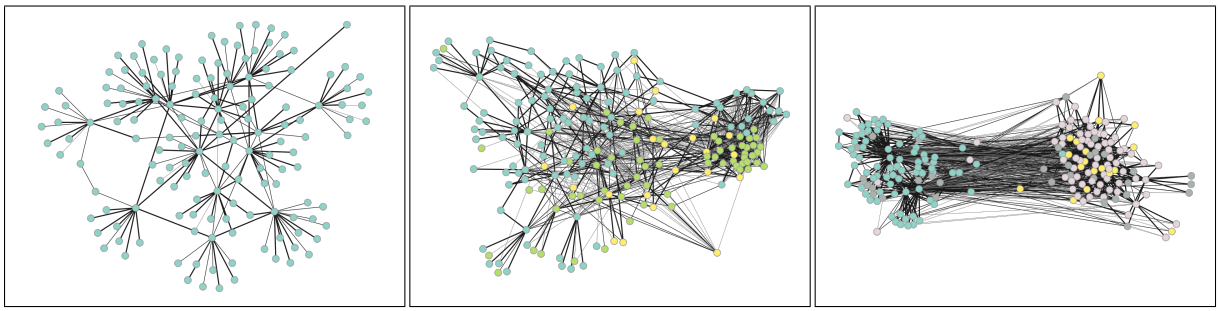

Figure 6: Sampled Word Usage Graphs of German *Festspiel* (left), German *Abgesang* (middle) and *zersetzen* (right).

In summary, the two clustering algorithms often make similar decisions, but different decisions especially where the clustering structure is complex and unclear.

6.2 Posterior Predictive Checking

We now test how well a model $P(b|A, x)$ fitted to a WUG $G = (U, E, W_G)$ retains the structural properties of G. For this we create a new graph $H = (U, E, W_H = \{\})$ with the same vertices and edges as G, but without the weights. This means $U_G = U_H$ and $E_G = E_H$. Then for all edges $e \in E_H$ we sample from the inferred distribution $\mathcal{D}$ with parameters $\mathcal{P}$ that best describes the weight distribution for the respective block combination (b_{e_1}, b_{e_2}) of e. For a model with a very close fit to the data the drawn edge weights will resemble the observed weights. We then visually compare the observed and the sampled graphs (Figures 6 vs. 2).

In Figure 6 we can see that in simple graphs like *Festspiel* the inferred structure coincides completely with the observed one. However, in graphs with a more complex structure like *Abgesang* and *zersetzen* (see Figure 5, middle and right) no distribution is flexible enough to fully describe the observed weight distributions, reflecting the observations from above. This is clearly manifested in the amount of high weights (black edges) inferred between the different blocks which are not present in the observed graph (see Figure 2).

6.3 Link Prediction

With link prediction we test how well a fitted model $P(b|A, x)$ from a WUG $G = (U, E, W_G)$ can predict unobserved annotations, i.e., missing edges in the graph. For this we randomly delete 5% of the edges of G and predict them by drawing from the distribution $\mathcal{D}$ as described above. We then quantify the difference between each predicted w_p and the corresponding observed edge weight w_o and define the Inverse Mean Error

$$\text{IME} = 1 - \frac{|w_o - w_p|}{4 - 1}$$

as a measure of how well a model structure predicts the observed graph structure (Figure 4, right). For about half of the graphs this score is quite high (IME $> .8$), i.e., the sampled weights are close to the observed values. Again, simpler block structures are easier to fit and are thus better predictable. For half of the graphs the predictability is lower though, for some even $< .5$. These results quantitatively confirm our observations from above, i.e., the fitted distributions often do not model the observed graphs sufficiently well.

7 Conclusion

We suggested to model human-annotated Word Usage Graphs with a Bayesian formulation of the Weighted Stochastic Block Model, compared several variations of the model and chose the best-fitting model in a principled way. In addition, we demonstrated how to interpret the inferred model as a model of word senses, but also that this interpretation is in no way necessary. The inferred models provide a stochastically-driven clustering and can be used to simulate realistic WUGs. An analysis of the model fits illustrated that more flexible distributions for edge weights are needed to yield good fits for a range of graphs.

We would like to emphasize that we do *not* claim that the WSBM is the *best* model for WUGs. Rather, we propose WSBMs as a reasonable probabilistic model for our data that can be rigorously compared against competing models in a Bayesian probabilistic framework, and potentially be neglected.

In the future, we aim to test more flexible edge weight distributions and to compare WSBMs to further probabilistic models, such as Gaussian Mixture models (Duda and Hart, 1973) and Latent Space models (Hoff et al., 2002). These models are interesting because they explicitly enforce the triangular property on graphs, which certain types of proximity judgments are known to obey (Erk et al., 2013). We also aim to explore Mixed Membership SBMs (Airoldi et al., 2008; Peixoto, 2015) and multi-graph models (Peixoto, 2017) where the information from each annotator can be represented individually.

Acknowledgments

We thank Tiago de Paula Peixoto for advice and for providing the idea and an implementation of the marginalization over edge probabilities. We further thank the reviewers for their constructive feedback. 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

Emmanuel Abbe. 2017. Community detection and stochastic block models: recent developments.

Christopher Aicher, Abigail Z. Jacobs, and Aaron Clauset. 2014. Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2):221—-248.

Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. 2008. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2004. Correlation clustering. *Machine Learning*, 56(1-3):89–113.

Chris Biemann. 2006. Chinese whispers: An efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, TextGraphs-1, page 73–80, USA. Association for Computational Linguistics.

Susan Windisch Brown. 2008. Choosing sense distinctions for WSD: Psycholinguistic evidence. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 249–252, Stroudsburg, PA, USA.

Richard O. Duda and Peter E. Hart. 1973. *Pattern classification and scene analysis*.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554.

Lea Frermann and Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. 2013. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Peter D. Grünwald and Abhijit Grunwald. 2007. *The Minimum Description Length Principle*. Adaptive computation and machine learning. MIT Press.

Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. 2002. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. 1983. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109 – 137.

Brian Karrer and M. E. J. Newman. 2011. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83:016107.

Adam Kilgarriff. 1997. "I don't believe in word senses". *Computers and the Humanities*, 31(2).

Karl-Rudolf Koch. 2007. *Introduction to Bayesian Statistics*, 2nd edition. Springer Publishing Company, Incorporated.

David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031.

Diana McCarthy, Marianna Apidianaki, and Katrin Erk. 2016. Word sense clustering and clusterability. *Computational Linguistics*, 42(2):245–275.

Tiago Peixoto. 2017. Nonparametric weighted stochastic block models. *Physical Review E*, 97.

Tiago P. Peixoto. 2014a. Documentation of the graph-tool python library. (last checked july 17, 2020).

Tiago P. Peixoto. 2014b. Efficient monte carlo and greedy heuristic for the inference of stochastic block models. *Physical Review E*, 89(1).

Tiago P. Peixoto. 2014c. The graph-tool python library. *figshare*.

Tiago P. Peixoto. 2015. Model selection and hypothesis testing for large-scale network models with overlapping groups. *Physical Review X*, 5:011033.

Tiago P. Peixoto. 2019. Bayesian stochastic blockmodeling. In *Advances in Network Clustering and Blockmodeling*, chapter 11, pages 289–332. John Wiley & Sons, Ltd.

Tiago P. Peixoto. 2020. Latent poisson models for networks with heterogeneous density. *Physical Review E*, 102:012309.

Valerio Perrone, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray. 2019. GASC: Genre-aware semantic change for ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 56–66, Florence, Italy. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Dominik Schlechtweg and Sabine Schulte im Walde. submitted. Clustering Word Usage Graphs: A Flexible Framework to Measure Changes in Contextual Word Meaning.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 169–174, New Orleans, Louisiana.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages. *CoRR*, abs/2104.08540.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Mark Steyvers and Tom Griffiths. 2007. *Probabilistic Topic Models*. Lawrence Erlbaum Associates.

A Additional Plots

Find additional plots in Figures 7, 8 and 9.

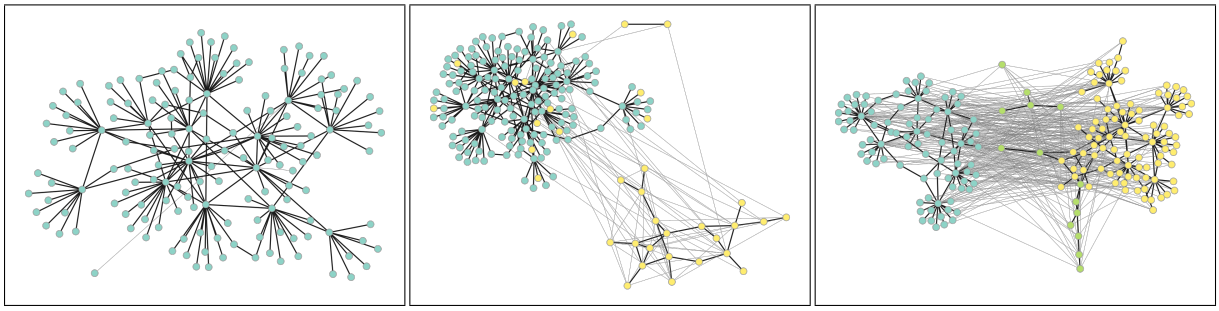

Figure 7: Word Usage Graphs of German *Ausnahmegesetz* (left), *stroke* (middle) and *plane* (right).

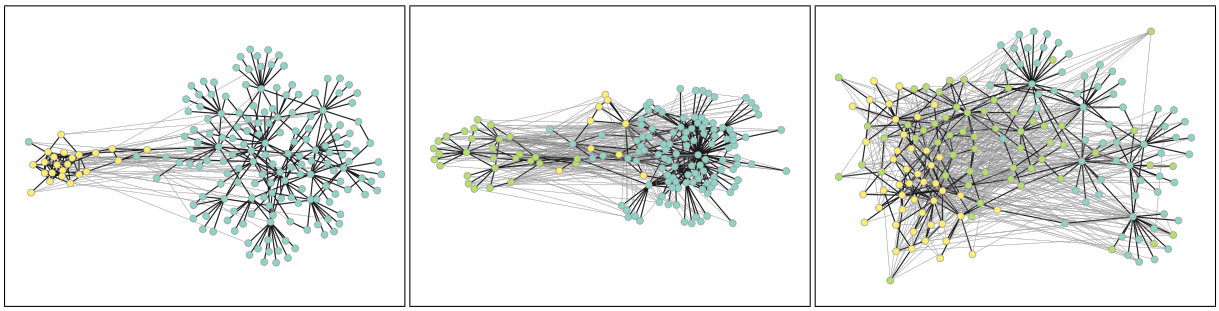

Figure 8: Word Usage Graphs of German *Sensation* (left), German *artikulieren* (middle) and *verbauen* (right).

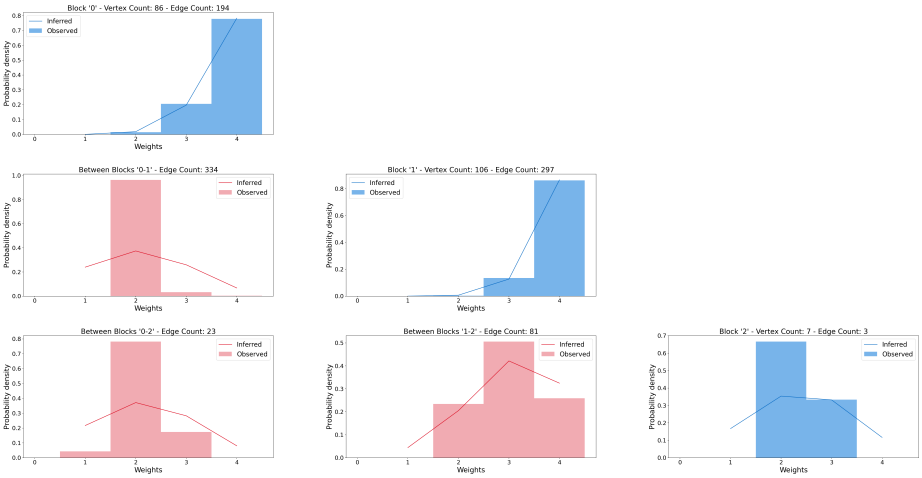

Figure 9: Detailed weight distribution of German *zersetzen*. Distribution within blocks in the diagonal and between blocks outside. Block ‘0’ maps the cyan cluster, Block ‘1’ the green cluster and Block ‘2’ the yellow one in Figure 2. The bars represent the observed values while the curves represent the inferred binomial distribution.