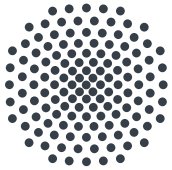


University of Stuttgart
Germany

UNIVERSITY OF CAMBRIDGE

DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages

October 14, 2021

Dominik Schlechtweg,[♣] Nina Tahmasebi,[♠] Simon Hengchen,[♠]
Haim Dubossarsky,[♦] Barbara McGillivray[◇,♡]

[♣]University of Stuttgart, [♠]University of Gothenburg, [♦]University of Cambridge,
[◇]King's College London [♡]The Alan Turing Institute

The Alan Turing Institute

SPRÅKBANKENTEXT

Introduction

- traditional approach to annotate word senses are binary assignments to sense descriptions (Kilgarriff, 1998)
  - ignores gradedness of word meaning

    (Erk, McCarthy, & Gaylord, 2013)
- two alternatives proposed by Erk et al. (2013):
  - (i) graded judgments of word usage pairs (usage-usage)
  - (ii) graded assignments of word usages to sense descriptions (usage-sense)
- judgments populate weighted graph

  (McCarthy, Apidianaki, & Erk, 2016)
- senses are not annotated directly, but **inferred** on the graph
- problems: applicability, scalability
- data available at:
  `https://www.ims.uni-stuttgart.de/data/wugs`

Data

	C_1	C_2
English	CCOHA 1810–1860	CCOHA 1960–2010
German	DTA 1800–1899	BZ+ND 1946–1990
Swedish	Kubhist 1790–1830	Kubhist 1895–1903
Latin	LatinISE -200–0	LatinISE 0–2000

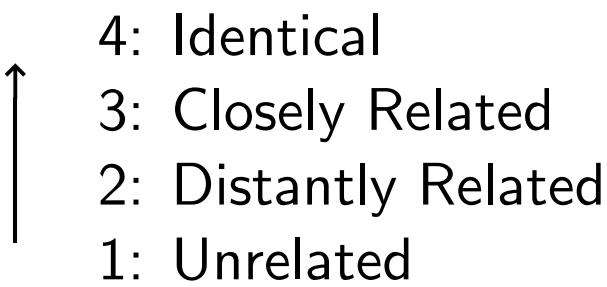
Time-defined subcorpora (Schlechtweg et al., 2020).

Procedure (i): Usage-Usage Graphs

(Usage) Von Hassel replied that he had such faith in the **plane** that he had no hesitation about allowing his only son to become a Starfighter pilot.

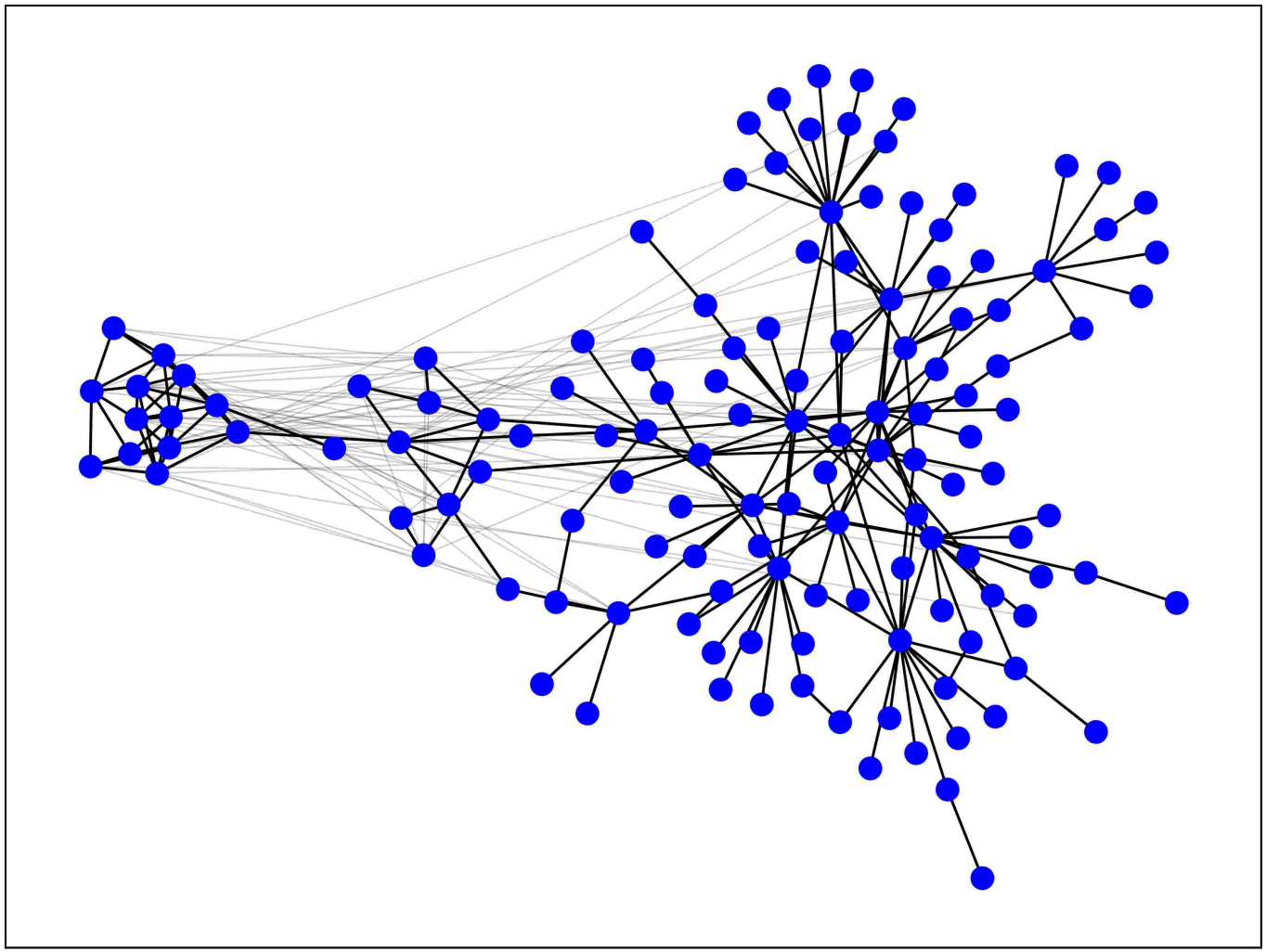
(Usage) This point, where the rays pass through the perspective **plane**, is called the seat of their representation.

Scale



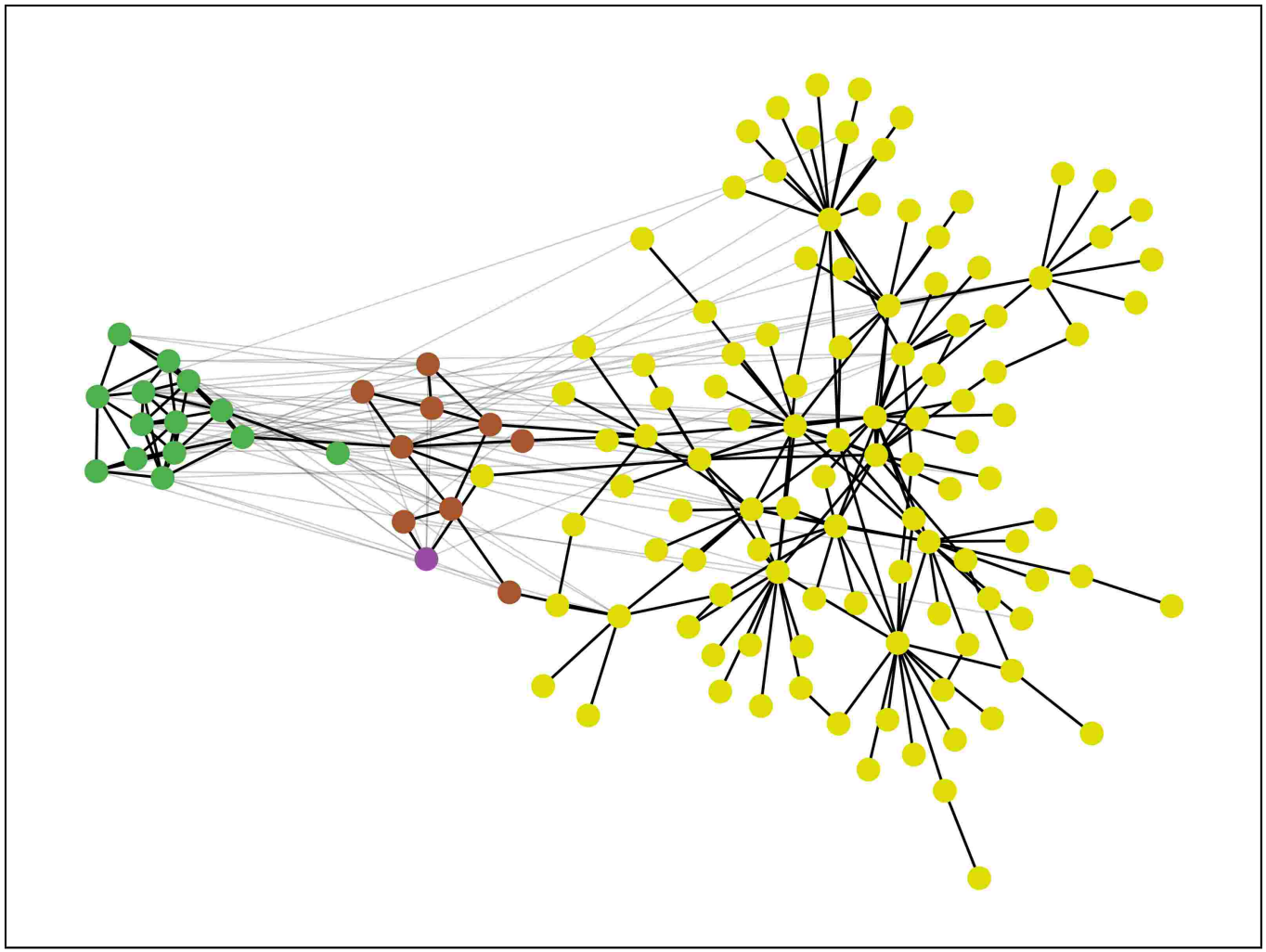
DURel relatedness scale (Schlechtweg et al., 2018).

Graph representation



Usage-usage graph of Swedish *ledning*. Nodes represent usages of the respective target word. Edge weights represent the median of relatedness judgments between usages (**black**/gray lines for **high**/low edge weights, i.e., weights ≥ 2.5/weights < 2.5).

Clustering



Usage-usage graph of Swedish *ledning*.

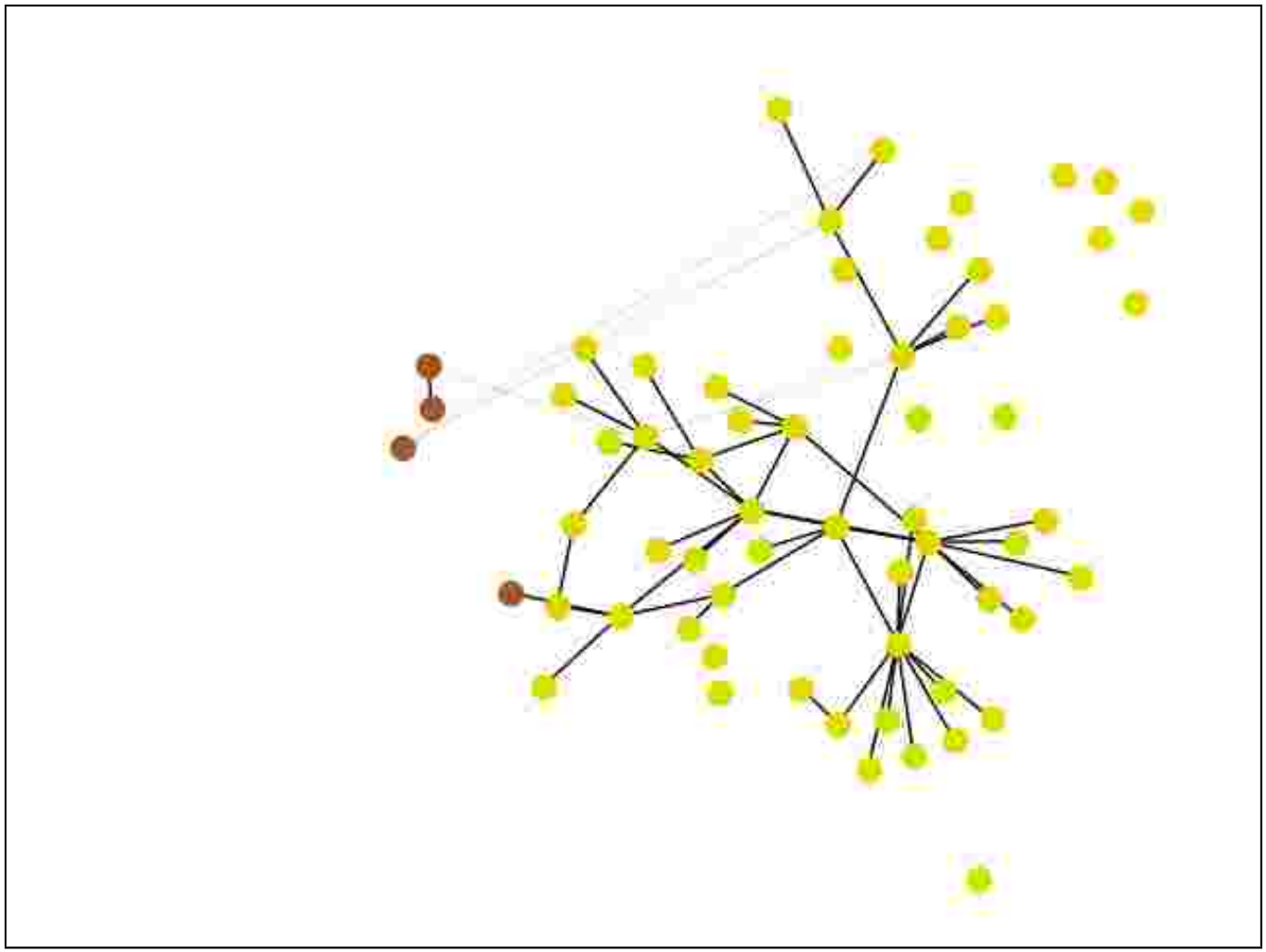
Clustering

- correlation clustering (Bansal, Blum, & Chawla, 2004)
- optimization criterion: **minimize (weighted) number of cluster-edge conflicts** (Schlechtweg et al., 2020)

$$\arg\min_{C} L(C) = \sum_{e \in \phi_{E,C}} W'(e) + \sum_{e \in \psi_{E,C}} |W'(e)|$$

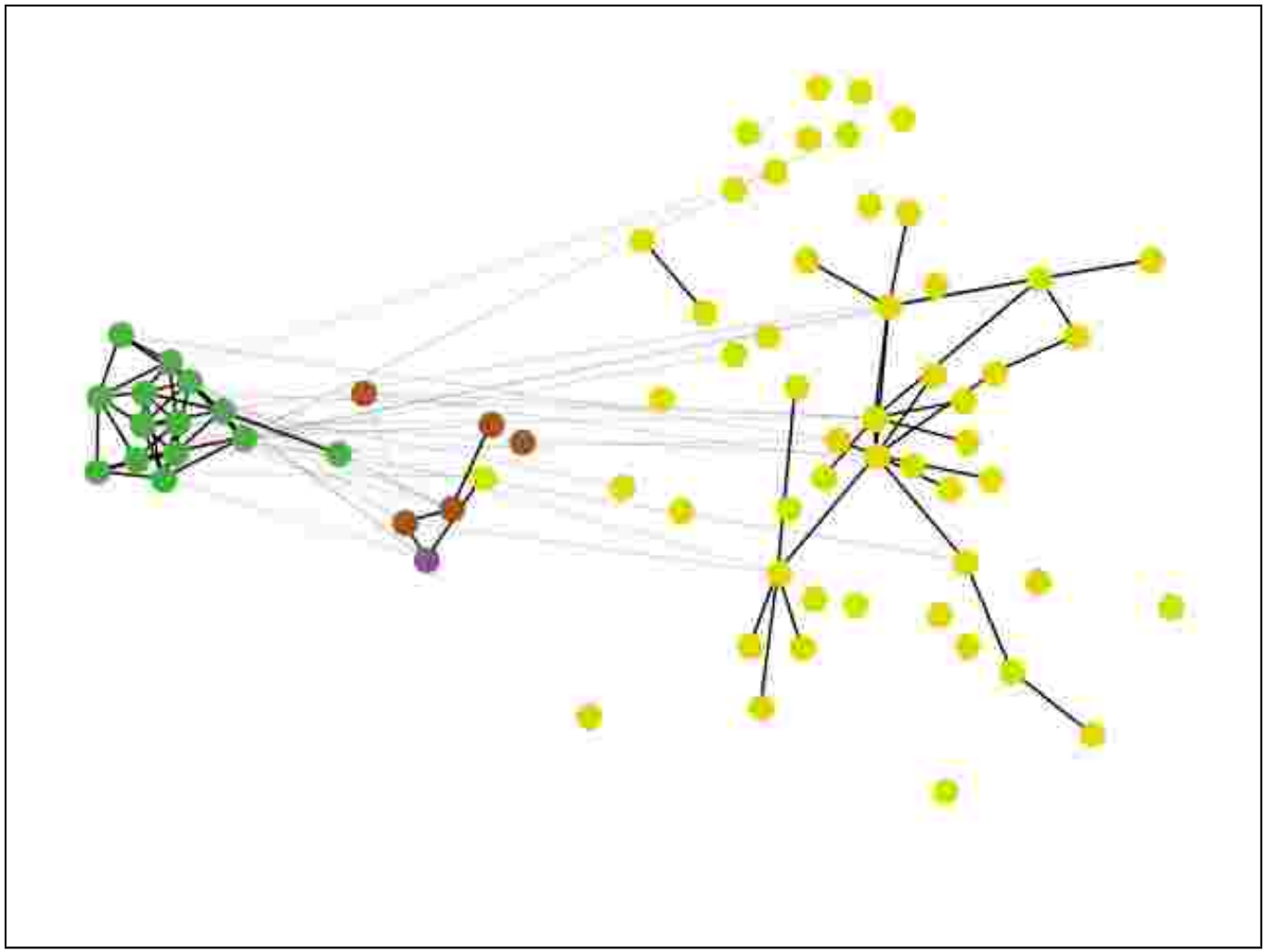
(i) finds the optimal number of clusters on its own
(ii) handles missing information (non-observed edges)
(iii) robust to errors by using the global information
(iv) respects the gradedness of word meaning
(v) dominated in simulation study

Time-specific subgraphs



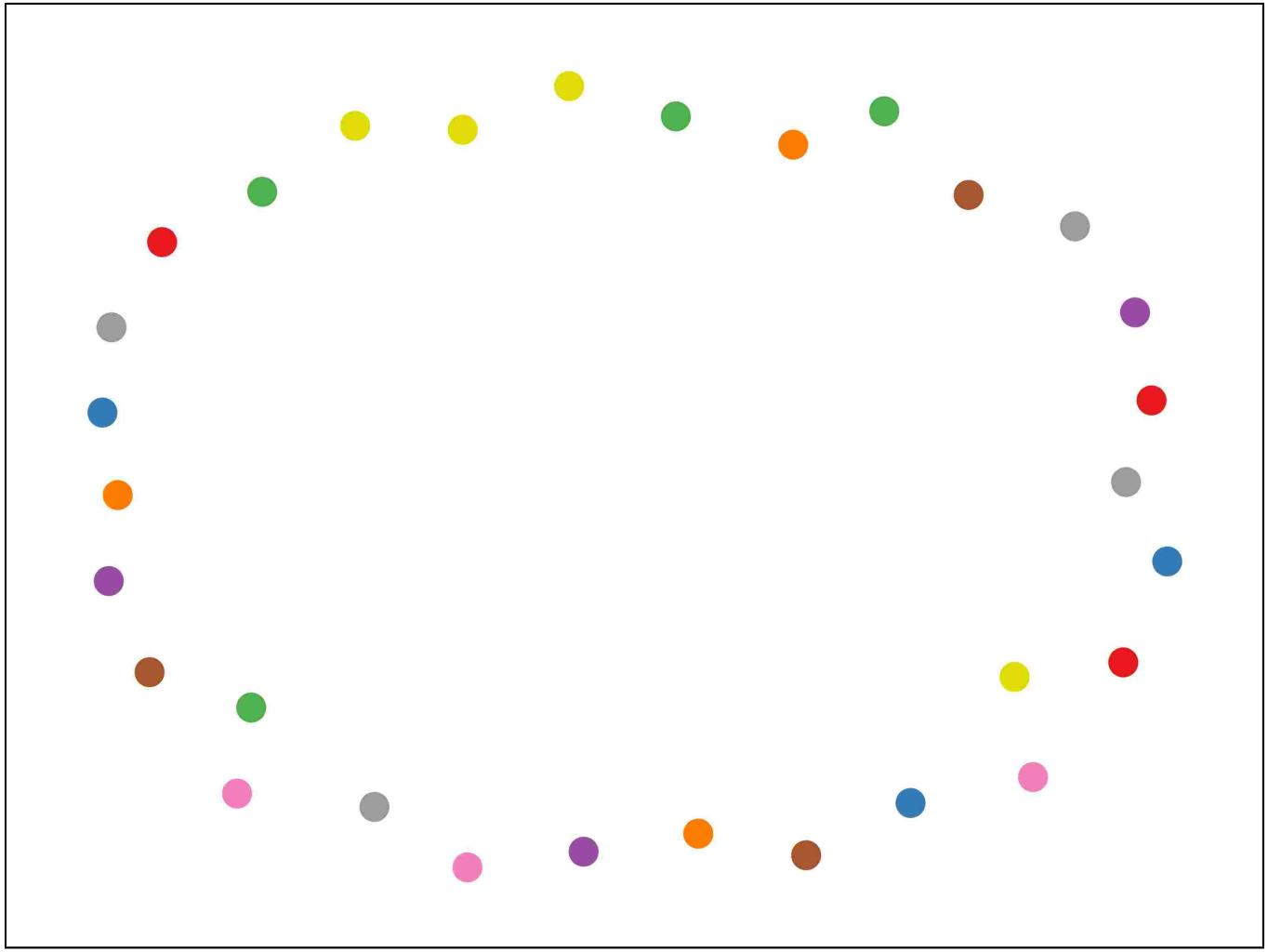
Subgraph of Swedish *ledning* for **old** subcorpus.

Time-specific subgraphs



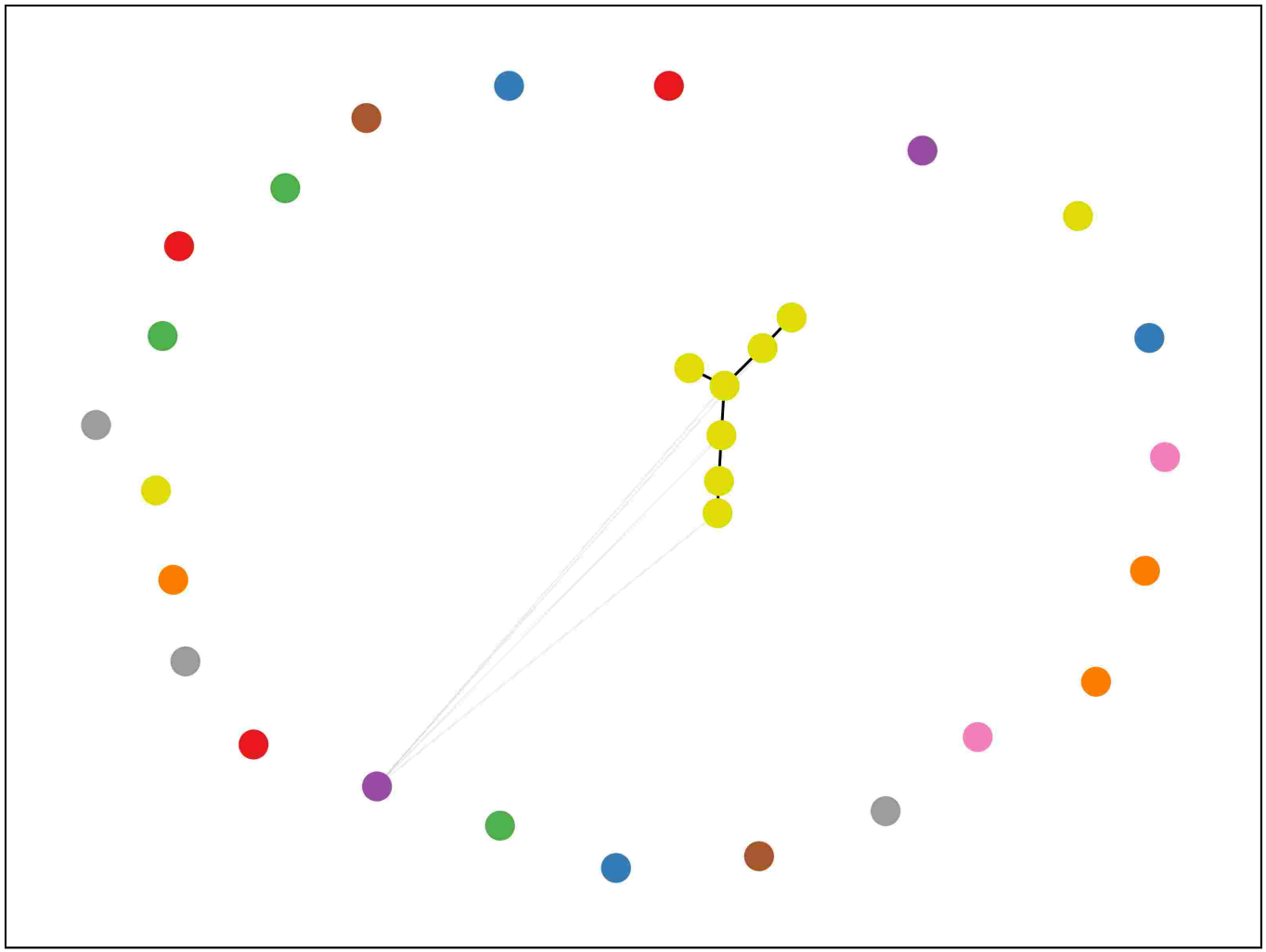
Subgraph of Swedish *ledning* for **new** subcorpus.

Edge Sampling



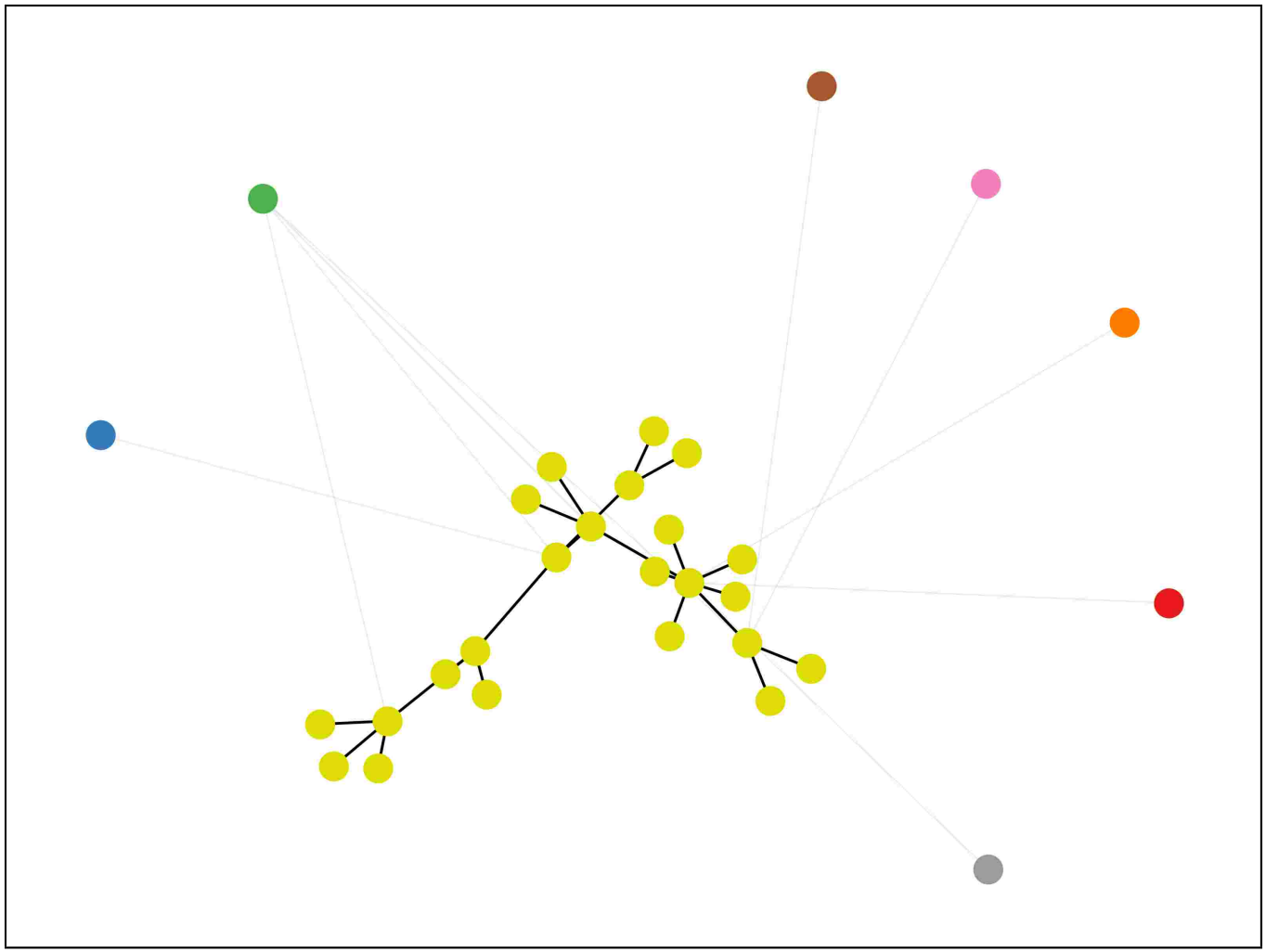
Round 0: No information.

Edge Sampling



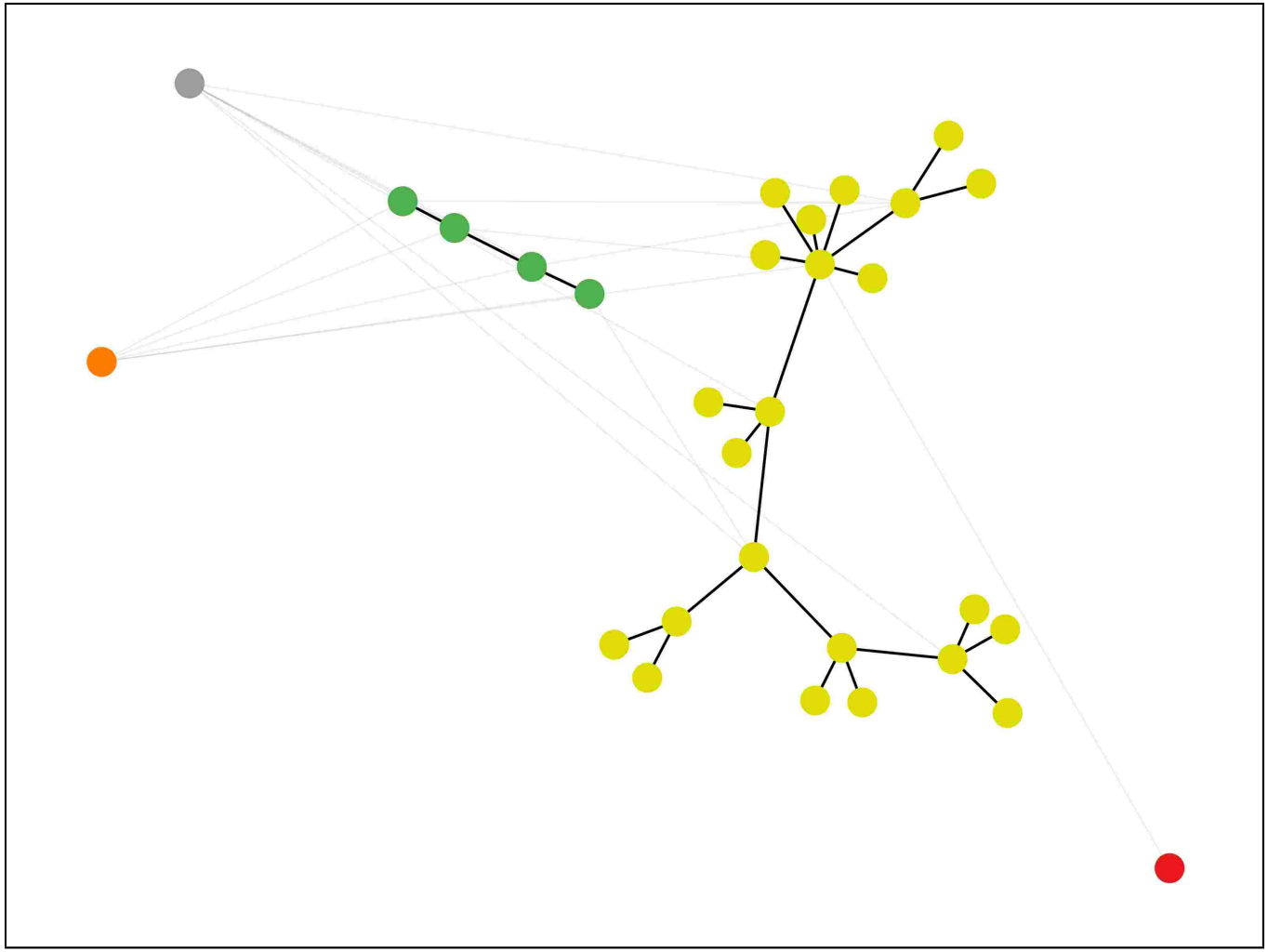
Round 1: Initial clustering (exploration).

Edge Sampling



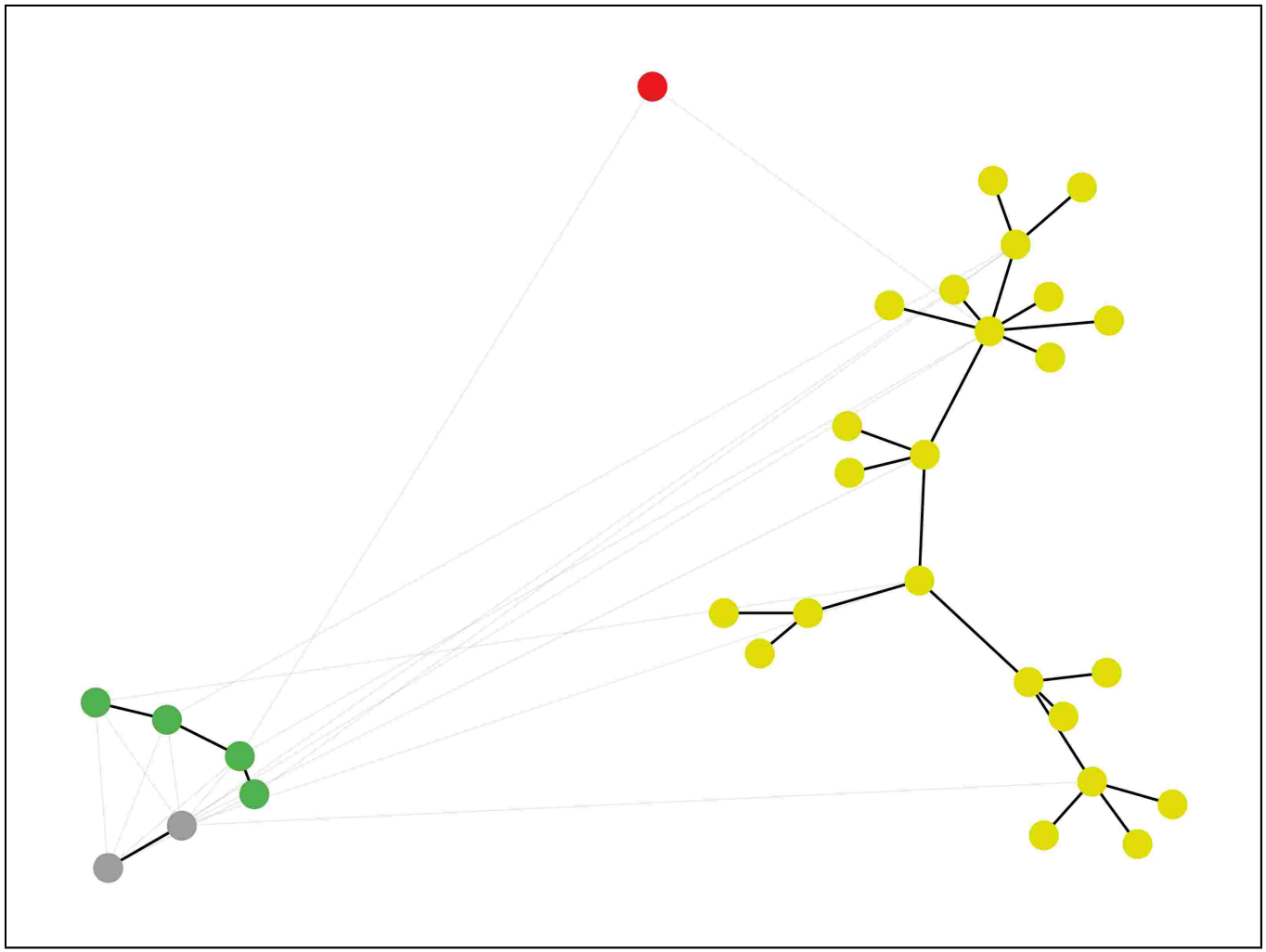
Round 2: Cluster comparison (combination).

Edge Sampling



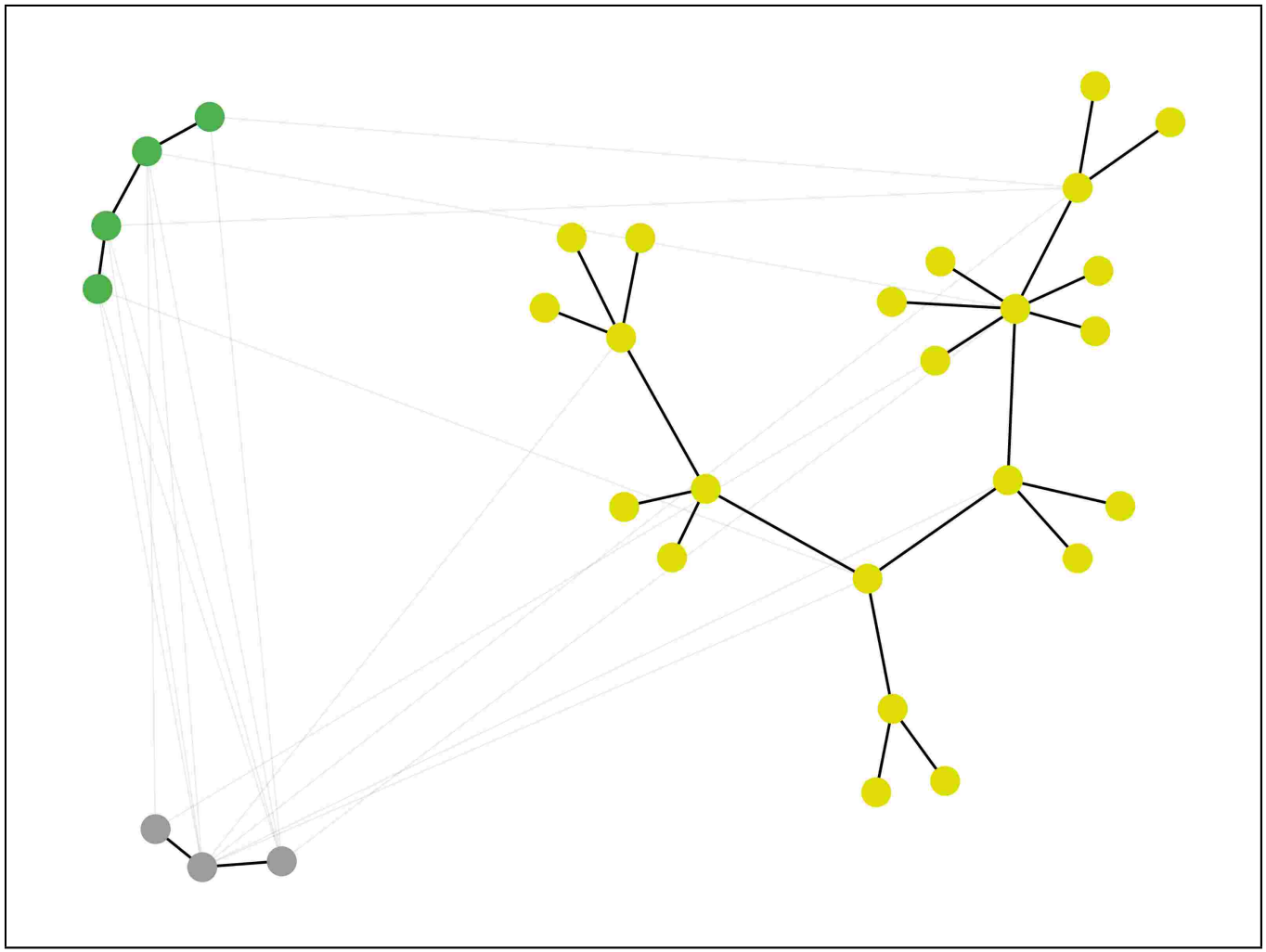
Round 3: Compare non-assignable uses (exploration).

Edge Sampling



Round 4: Combination and exploration.

Edge Sampling



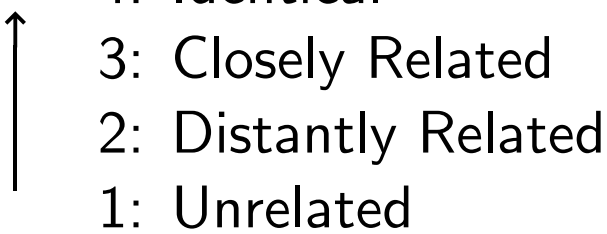
Round 5: Combination.

Procedure (ii): Usage-Sense Graphs

(Usage) Cum Arretinae mulieris libertatem defenderem et Cotta xviris religionem iniecisset non posse nostrum **sacramentum** iustum iudicari, [. . .]
'When I was defending the liberty of a woman of Arretium, and when Cotta had suggested a scruple to the decemvirs that our **action** *was not a regular one, [. . .]'*

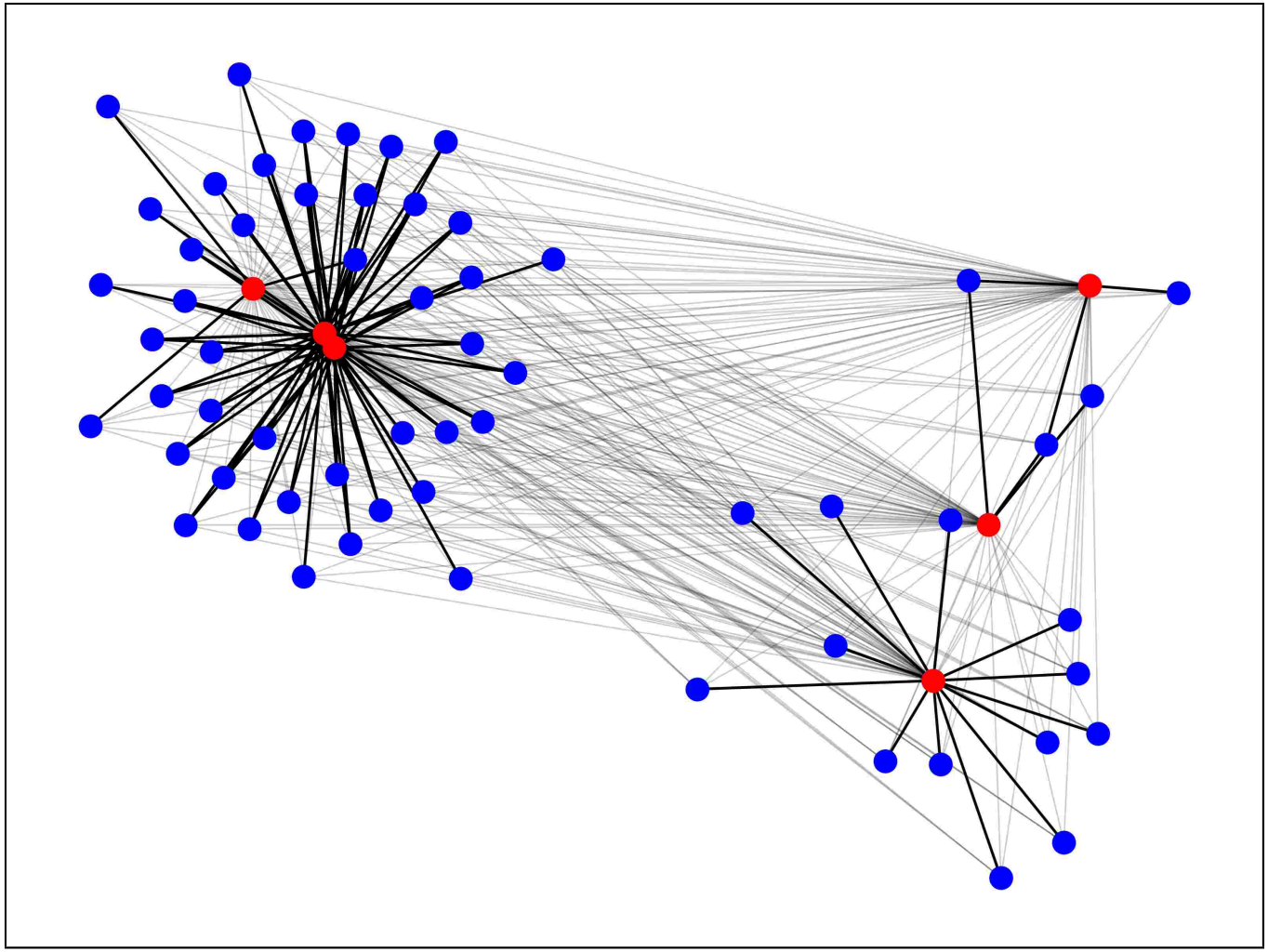
(Sense) "a cause, a civil suit or process"

Scale



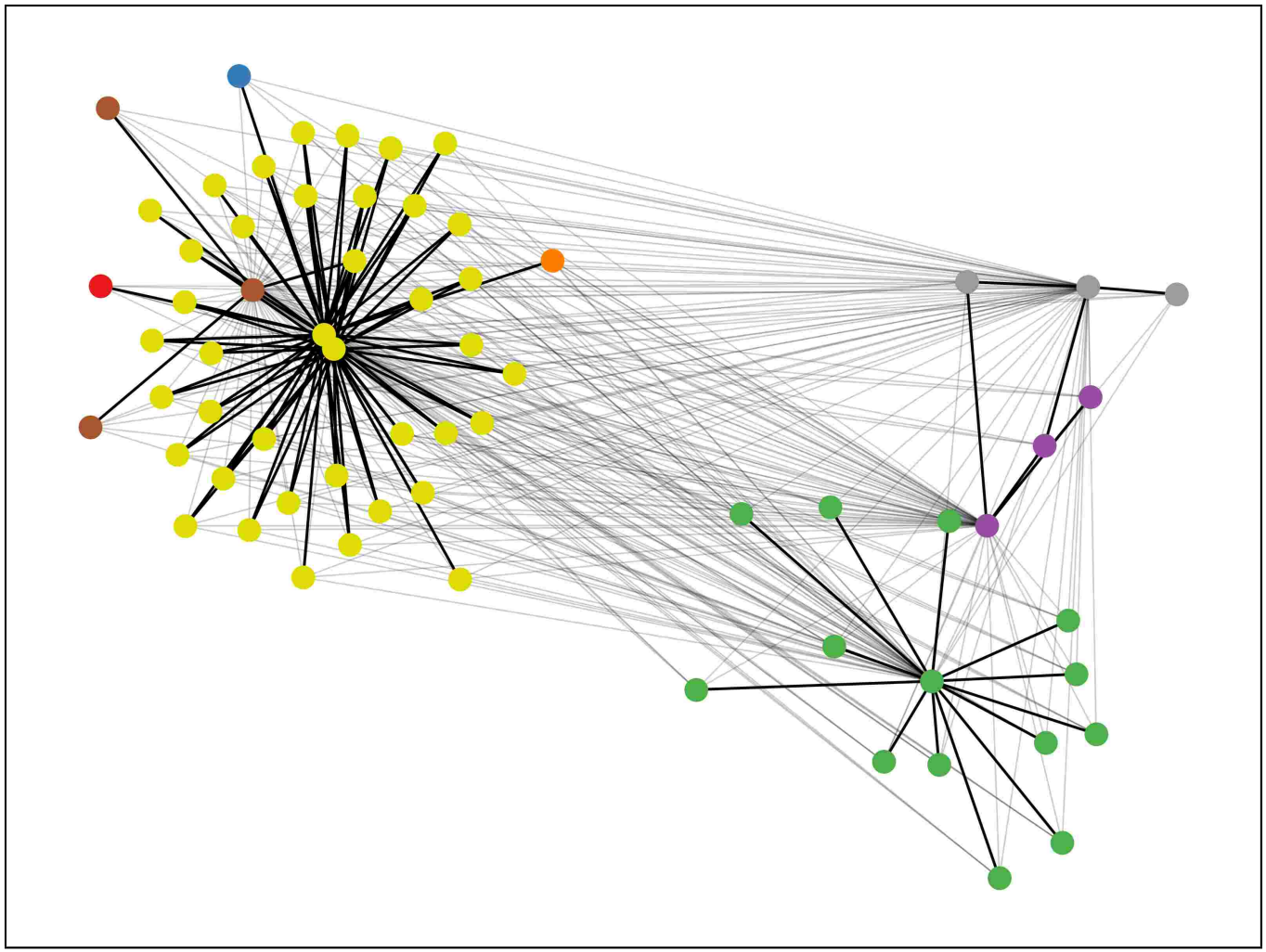
DURel relatedness scale (Schlechtweg et al., 2018).

Graph representation



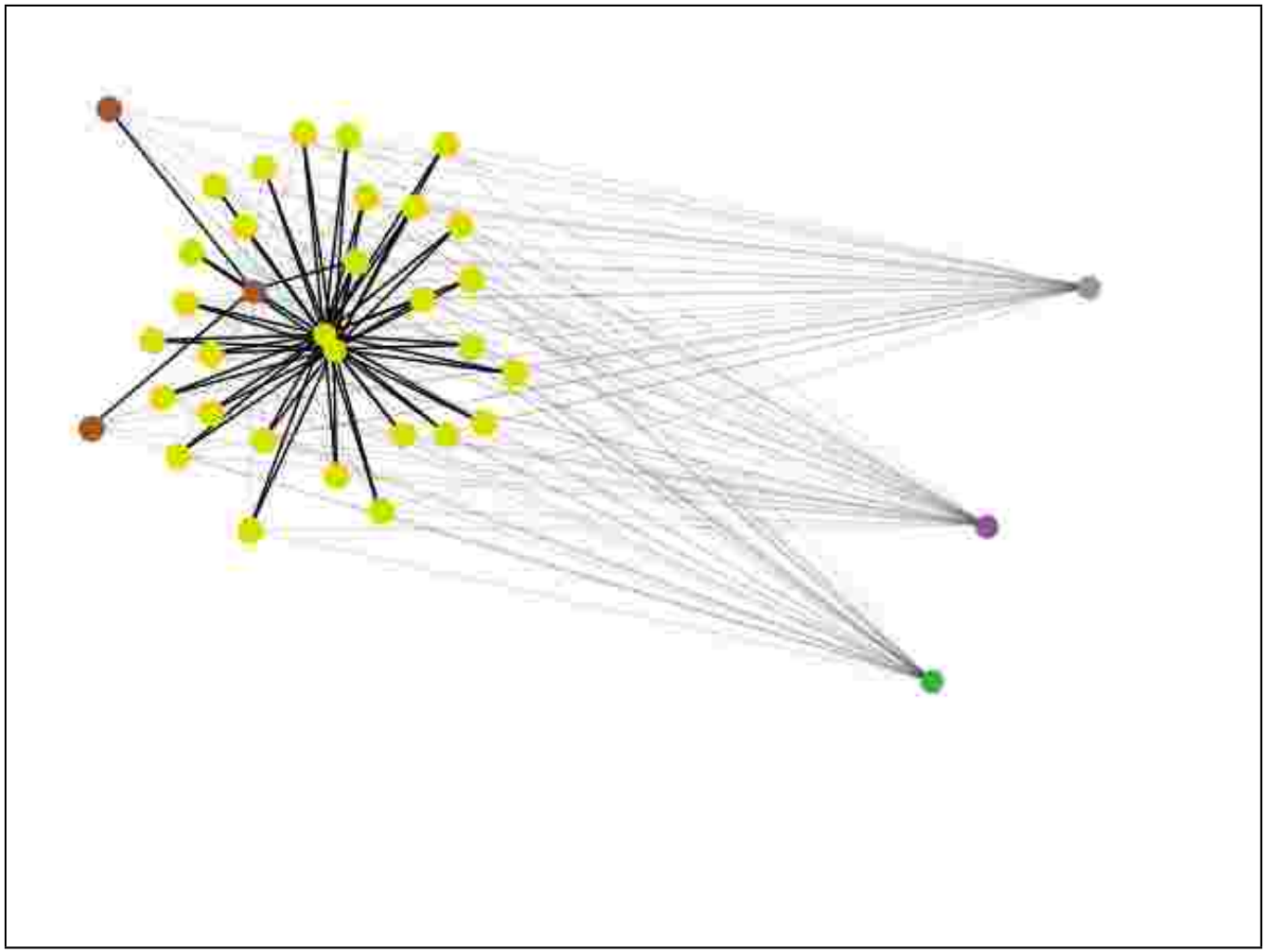
Usage-sense graph of Latin *sacramentum*. Nodes in blue/red represent usages/senses respectively.

Clustering



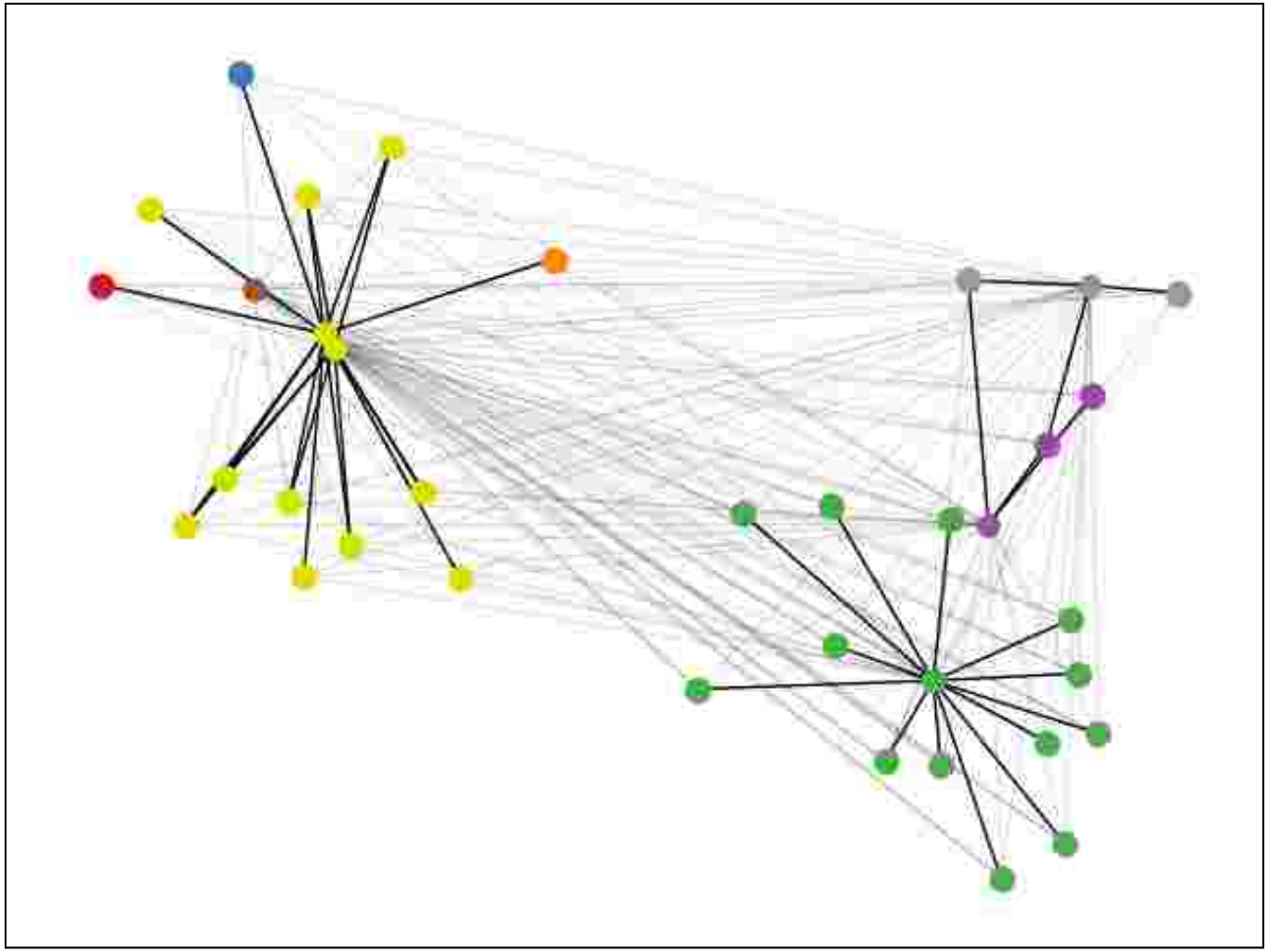
Usage-sense graph of Latin *sacramentum*.

Time-specific subgraphs



Subgraph of Latin *sacramentum* for **old** subcorpus.

Time-specific subgraphs



Subgraph of Latin *sacramentum* for **new** subcorpus.

Overview

LGS	n	N/V/A	$\lvert U \rvert$	AN	JUD	SPR	KRI
EN	40	36/4/0	189	9	29k	.69	.61
DE	48	32/14/2	178	8	37k	.59	.53
SV	40	31/6/3	168	5	20k	.57	.56
LA	40	27/5/8	59	1	9k	.64	.62

Dataset overview.

Possible Uses

- as large sets (thousands) of pairwise **semantic proximity judgments** to evaluate contextualized embeddings in multiple languages;
- the inferred change scores can be used to evaluate **semantic change detection** models;
- as **word sense disambiguation/discrimination** resources with additional aspects such as variation over time;
- graphs may be treated as research objects in their own right

Conclusion

- largest existing resource of word usage graphs and graded semantic proximity judgments
- usage-usage graphs **avoid the need for a priori sense descriptions**
- usage-sense graphs naturally **reduce the number of necessary judgments**
- senses are not annotated directly, but **inferred** on the annotated data with a robust clustering procedure
- future:
  - evaluate inferred clusterings and optimize clustering procedure
  - compare **probabilistic models** of the annotated data

    (Schlechtweg, Castaneda, Kuhn, & Schulte im Walde, 2021)
- we openly release the data, clusterings, visualizations, statistics and code:

  `https://www.ims.uni-stuttgart.de/data/wugs`

Bibliography

Bansal, N., Blum, A., & Chawla, S. (2004). Correlation clustering. *Machine Learning*, *56*(1-3), 89–113. doi: 10.1023/B:MACH.0000033116.57574.95

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.

Schlechtweg, D., Castaneda, E., Kuhn, J., & Schulte im Walde, S. (2021). Modeling sense structure in word usage graphs with the weighted stochastic block model. In *Proceedings of the 10th Joint Conference on Lexical and Computational Semantics.*

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (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. Retrieved from `https://www.aclweb.org/anthology/2020.semeval-1.1/`

Schlechtweg, D., Schulte im Walde, S., & Eckmann, S. (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* (pp. 169–174). New Orleans, Louisiana. Retrieved from `https://www.aclweb.org/anthology/N18-2027/`