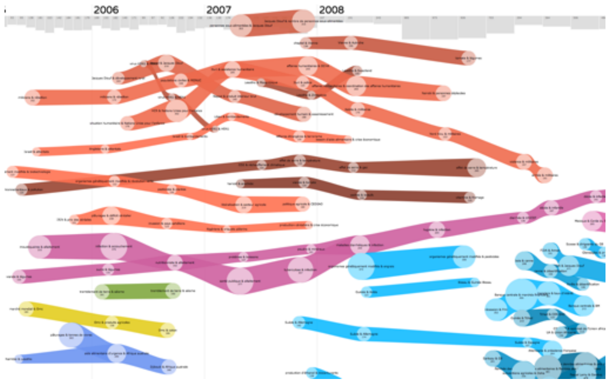
Recent advances in information network embedding @ERIC

Julien Velcin
julien.velcin@univ-lyon2.fr
https://eric.univ-lyon2.fr/~jvelcin/
Université Lumière Lyon 2 - ERIC Lab

Context

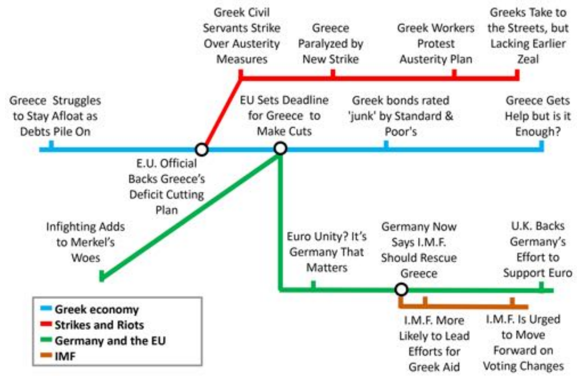
Informational landscape



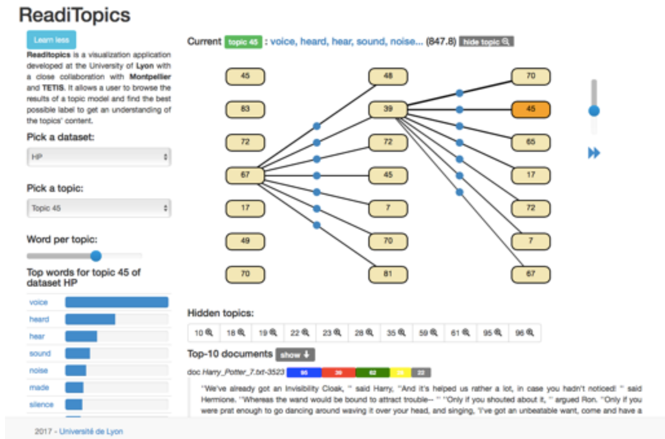
Projet Pulseweb
(Cointet, Chavalarias…)

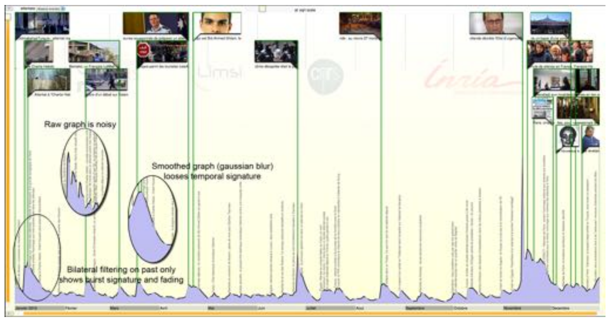
http://pulseweb.cortext.net

Metromaps
(Shahaf et al., 2015)



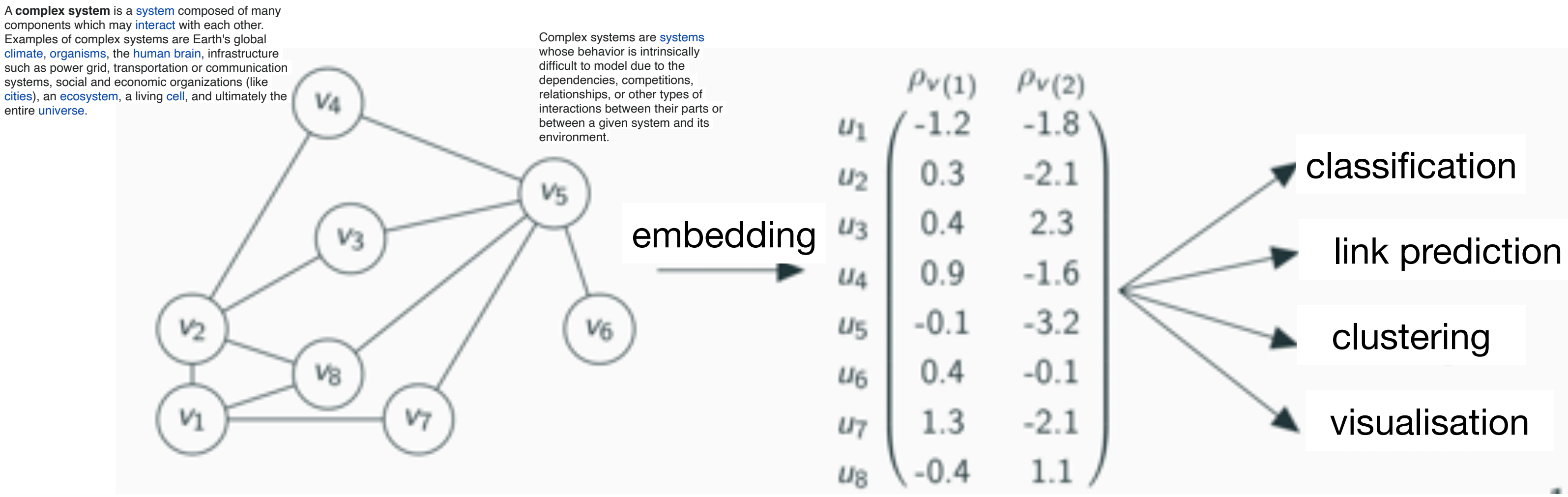
Readitopics
(Velcin et al., 2018)





Chronolines
(Nguyen et al., 2014)

Document network embedding



- **Document network**: “graph of vertices, where each vertex is associated with a text document” (Tuan et al., 2014)
  e.g.: scientific articles, newspapers, social media…

- Embedding for building a **joint** space for solving downstream tasks (e.g., link prediction, node classification, community detection)

Quick survey

- **Graph/Node embedding**
  - Laplacian Eigenmaps (Belkin and Niyogi, 2002)
  - DeepWalk (Perozzi et al., 2014), Node2vec (Grover and Leskovec, 2016)
  - Graph Neural Networks (Scarselli et al., 2009)
- **Document network embedding**
  - TADW (Yang et al., 2015)
  - Attention models and CANE (Tu et al., 2017)

Collaborators of the DMD team

Robin Brochier
Phd student
(now graduated!)

Antoine Gourru
Phd student

Adrien Guille
Associate
Professor

Julien Jacques
Professor

Contributions

Regularized Linear Embedding (RLE)

Gourru A., J. Velcin, J. Jacques and A. Guille
Document Network Projection in Pretrained Word Embedding Space. ECIR, 2020

Given:

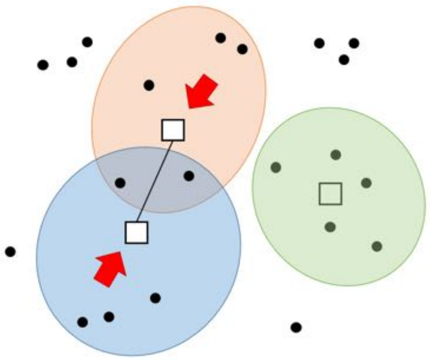
- $U \in \mathbb{R}^{v \times k}$ matrix of pretrained word embeddings
- $T \in \mathbb{R}^{n \times v}$ Document x Word matrix (*textual information*)
- $A \in \mathbb{R}^{[0,1] \times [0,1]}$ the transition matrix (*graph information*)

- **Goal:** learn the weights $p_i \in \mathbb{R}^v$ for the words composing d_i

$$d_i = p_i U$$

parameter to learn

The vector for d_i is just a weighted sum over pretrained WE



RLE (con't)

$$P = (1 - \lambda)T + \lambda B$$

with $\lambda \in [0,1]$ a tradeoff b/w textual and structural information

$$b_i = \frac{1}{\sum_j S_{i,j}} \sum_j S_{i,j} t_j$$

with $S \in \mathbb{R}^{n \times n}$ a squared matrix that reflects the pairwise similarity between nodes in the graph

(here, we use $S = \dfrac{A + A^2}{2}$)

Evaluation

Datasets:

- Cora (2,211 docs; 7 labels=topic; 5,001 citation links)
- DBLP (60,744 docs; 4 labels=topic; 52,914 links)
- New York Times (5,135 docs; 4 labels=article section; 3,050,513 links=common tag)

https://github.com/AntoineGourru/DNEmbedding

- **Task 1:** node classification
- **Task 2:** link prediction

Table 2. Comparison of AUC results on a link prediction task for different perce
hidden. The best score is in bold, second best is underlined.

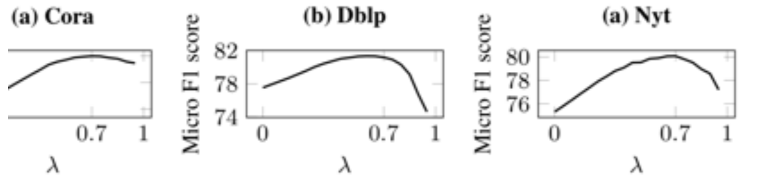
	Cora		Dblp	
% edges hidden	50%	25%	50%	25%
DeepWalk	73.2 (0.6)	80.9 (1.0)	**89.7** (0.0)	**93.2** (0.2)
LSA	87.4 (0.6)	87.2 (0.8)	54.2 (0.1)	54.8 (0.0)
Concatenation	77.9 (0.3)	83.7 (0.8)	88.8 (0.0)	92.6 (0.3)
TADW	90.1 (0.4)	93.3 (0.4)	61.2 (0.1)	65.0 (0.5)
AANE	83.1 (0.8)	86.6 (0.8)	67.4 (0.1)	66.5 (0.1)
GVNR-t	83.9 (0.9)	91.5 (1.1)	88.1 (0.3)	91.4 (0.1)
VGAE	87.1 (0.4)	88.2 (0.7)	Does not scale	
Graph2Gauss	92.0 (0.3)	93.8 (1.0)	88.0 (0.1)	92.1 (0.5)
STNE	83.1 (0.5)	90.0 (1.0)	45.6 (0.0)	53.4 (0.1)
RLE	**94.3** (0.2)	**94.8** (0.2)	89.3 (0.1)	91.2 (0.2)

Table 1. Comparison of Micro-F1 results on a classification task for different train/test ratios. The best score is in bold, second best is underlined. Execution time order is presented in seconds (Time).

	Cora				Dblp			
train/test ratio	10%	30%	50%	Time	10%	30%	50%	Time
DeepWalk	70.6 (2.0)	77.2 (0.9)	81.0 (0.7)	10^1	52.3 (0.4)	53.4 (0.1)	53.5 (0.2)	10^2
LSA	72.3 (1.9)	79.0 (0.7)	80.6 (0.7)	10^{-2}	73.5 (0.2)	74.1 (0.1)	74.2 (0.2)	10^1
Concatenation	71.4 (2.1)	80.5 (1.0)	84.0 (1.1)	10^1	77.5 (0.2)	78.0 (0.1)	78.2 (0.2)	10^2
TADW	81.9 (0.8)	86.3 (0.8)	87.4 (0.8)	10^{-1}	74.8 (0.1)	75.3 (0.2)	75.5 (0.1)	10^1
AANE	79.8 (0.9)	83.3 (1.1)	84.4 (0.7)	10^{-1}	73.3 (0.1)	73.9 (0.1)	74.2 (0.2)	10^2
GVNR-t	83.7 (1.2)	86.4 (0.7)	87.0 (0.8)	10^1	69.6 (0.1)	70.1 (0.1)	70.2 (0.2)	10^2
VGAE	72.3 (1.7)	79.2 (0.9)	81.1 (0.7)	10^1	Memory overflow			–
G2G	79.0 (1.5)	83.7 (0.8)	84.8 (0.7)	10^1	70.8 (0.1)	71.3 (0.2)	71.5 (0.2)	10^2
STNE	79.4 (1.0)	84.7 (0.7)	86.7 (0.8)	10^2	73.8 (0.2)	74.4 (0.1)	74.5 (0.1)	10^4
RLE	**84.0** (1.3)	**86.9** (0.5)	**87.7** (0.6)	10^1	**79.8** (0.2)	**80.9** (0.2)	**81.2** (0.1)	10^1

	Nyt			
train/test ratio	10%	30%	50%	Time
DeepWalk	66.9 (0.7)	68.2 (0.3)	68.7 (0.9)	10^2
LSA	71.6 (1.0)	75.7 (0.7)	76.7 (0.7)	10^{-2}
Concatenation	**77.9** (0.3)	**80.0** (0.5)	**81.1** (0.7)	10^2
TADW	75.8 (0.5)	78.4 (0.5)	79.4 (0.4)	10^1
AANE	71.7 (0.5)	75.6 (0.8)	76.9 (1.1)	10^1
GVNR-t	74.3 (0.4)	76.0 (0.6)	76.7 (0.6)	10^2
VGAE	68.1 (0.8)	69.3 (0.9)	70.1 (0.6)	10^2
G2G	69.0 (0.5)	70.5 (0.7)	71.5 (0.8)	10^2
STNE	75.1 (0.7)	77.3 (0.5)	78.1 (0.6)	10^2
RLE	77.7 (0.7)	79.3 (0.5)	80.0 (0.6)	10^1

sensitivity to λ

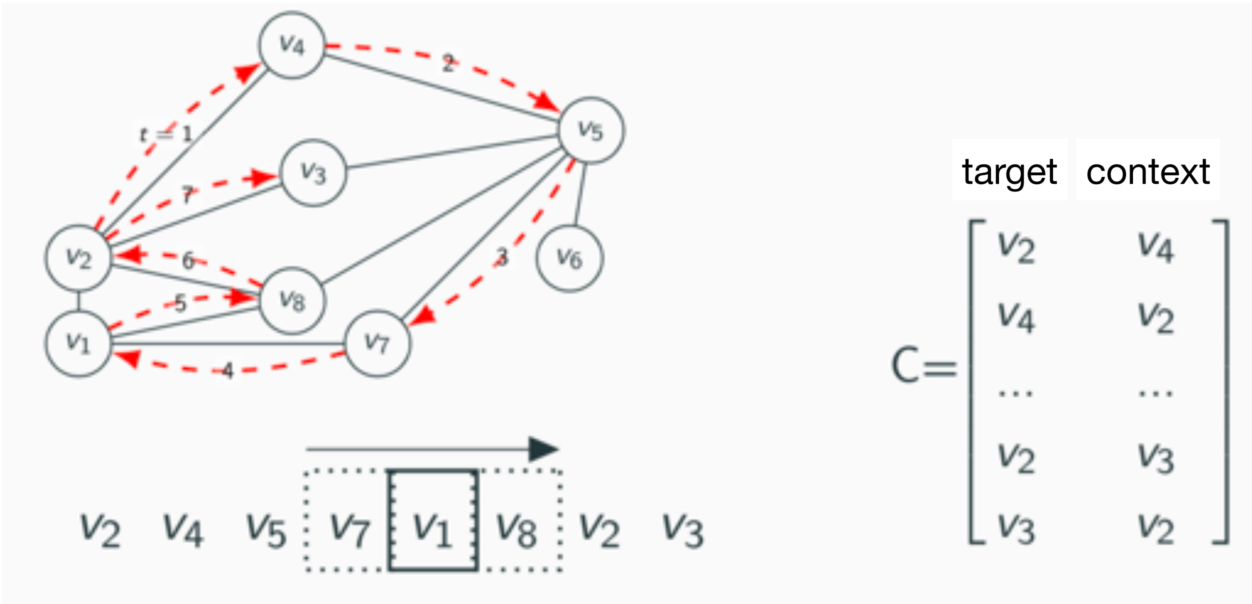


of λ on RLE in terms of document classification for $d = 160$. Optimum is achieved
each dataset (Cora, Nyt: 0.7, Dblp: 0.65).

GVNR and GVNR-t

Brochier, A., Guille and J. Velcin.
Global Vectors for Node Representation. The Web Conference (WWW), 2019

- Quick reminder of DeepWalk (Perozzi et al., 2014):
  - goal: learn vector representation of nodes
  - approach: a) make multiple random walks
    b) paths views as documents
    c) use Skip-Gram to build vectors (Mikolov et al., 2013)



GVNR and GVNR-t

Brochier, A., Guille and J. Velcin.
Global Vectors for Node Representation. The Web Conference (WWW), 2019

- Following GloVe (Pennington et al., 2014, GVNR solves **regression** task on the weighted cooccurrence matrix X where cells with small values are set to 0 (> threshold x_{min})

- We're looking for (U, b^U) and (V, b^V) s.t.:

$$\arg\min_{U,V,b^U,b^V} \sum_{i}^{n} \sum_{j}^{n} s(x_{ij})(u_i \,.\, v_j + b_i^U + b_j^V - log(1 + x_{ij}))^2$$

with $s(x_{ij}) = 1$ if $x_{ij} > 0$ and $m_i \sim B(\alpha)$ else

where α is chosen s.t. $m = k$ in average

$$X = \begin{pmatrix} 0 & 4 & 8 & 0 & 0 & 0 \\ 4 & 0 & 0 & 3 & 0 & 0 \\ 8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 6 & 1 & 0 \end{pmatrix}$$

cooccurrence b/w x_i and x_j

- GVNR-t integrates textual information by modifying v_j:

$$\arg\min_{\mathrm{U,W},b^U,b^V} \sum_{i}^{n} \sum_{j}^{n} s(x_{ij})(u_i \,.\, \boxed{\frac{\delta_j \,.\, W}{|\delta_j|_1}} + b_i^U + b_j^V - log(c + x_{ij}))^2$$

Results for GVNR-t

- Classification on two citation networks (Cora with 2,708 nodes and Citeseer with 3,312 nodes)
- Keyword recommendation on DBLP (1,397,240 documents and 3,021,489 citation relationships)

Table 8: Accuracy on the citation (1) network, considering the text features.

	% of training data				
	10%	20%	30%	40%	50%
LSA	54.7	61.0	62.4	63.0	62.8
DeepWalk+LSA	73.8	77.9	78.4	78.1	78.1
TADW	77.1	78.8	78.2	78.8	78.6
GVNR-t	**79.3**	**80.7**	**80.8**	**81.4**	**81.1**

Table 9: Accuracy on the citation (2) network, considering the text features.

	% of training data				
	10%	20%	30%	40%	50%
LSA	52.0	54.7	54.7	58.4	65.7
DeepWalk+LSA	58.3	60.7	61.1	60.0	61.2
TADW	60.6	60.1	60.1	66.2	69.3
GVNR-t	**63.3**	**62.5**	**64.9**	**68.6**	**70.4**

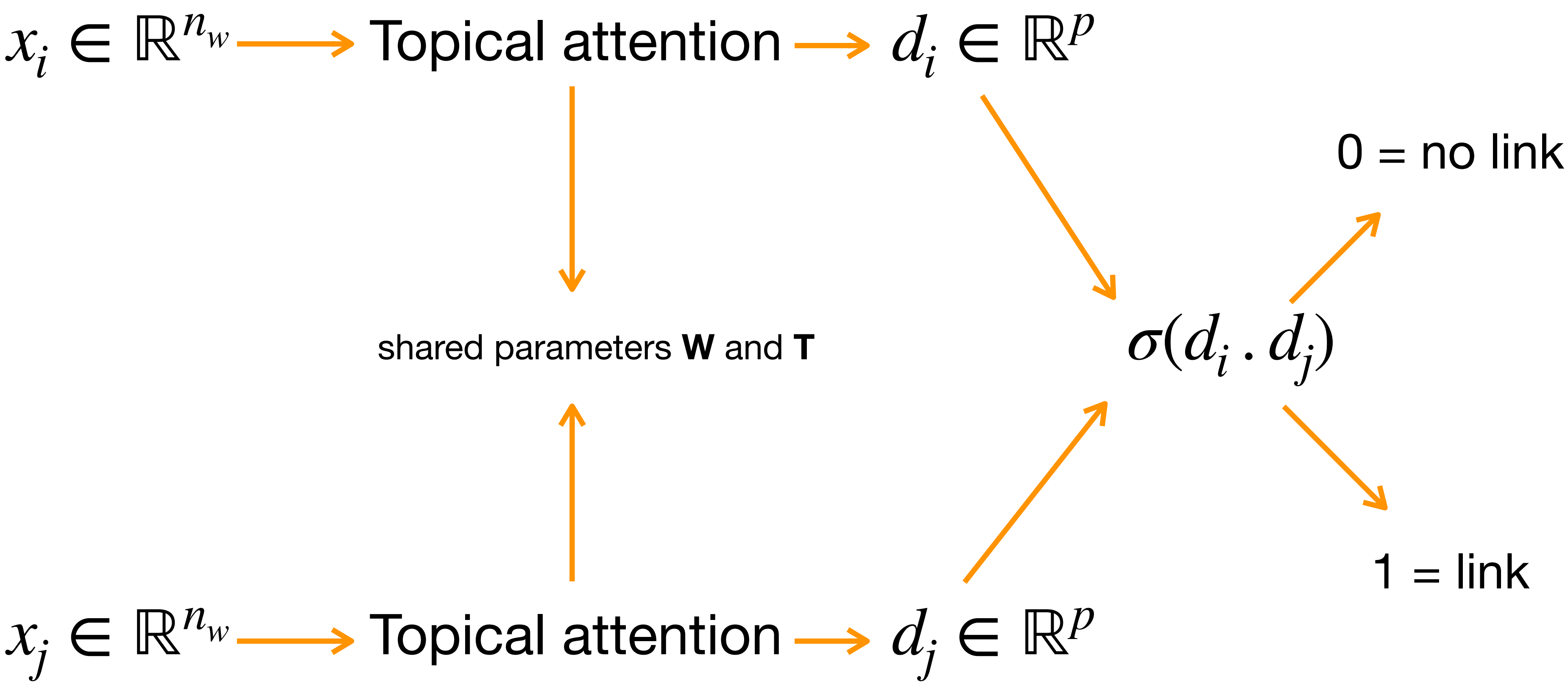
Table 10: Keyword recommendation by selecting the closest word embeddings w_k to both embeddings u (node) and v (content) of an input document (1).

Document	**A brief survey of computational approaches in social computing** Web 2.0 technologies have brought new ways of connecting people in social networks for collaboration in various on-line communities. Social Computing is a novel and emerging computing paradigm...
Closest words to u (node)	*cold start problem, storylines, document titles, movielens data, computational humor*
Closest words to v (content)	*social, social network, enron email corpus, social networks, extremely large datasets, sites blogs*

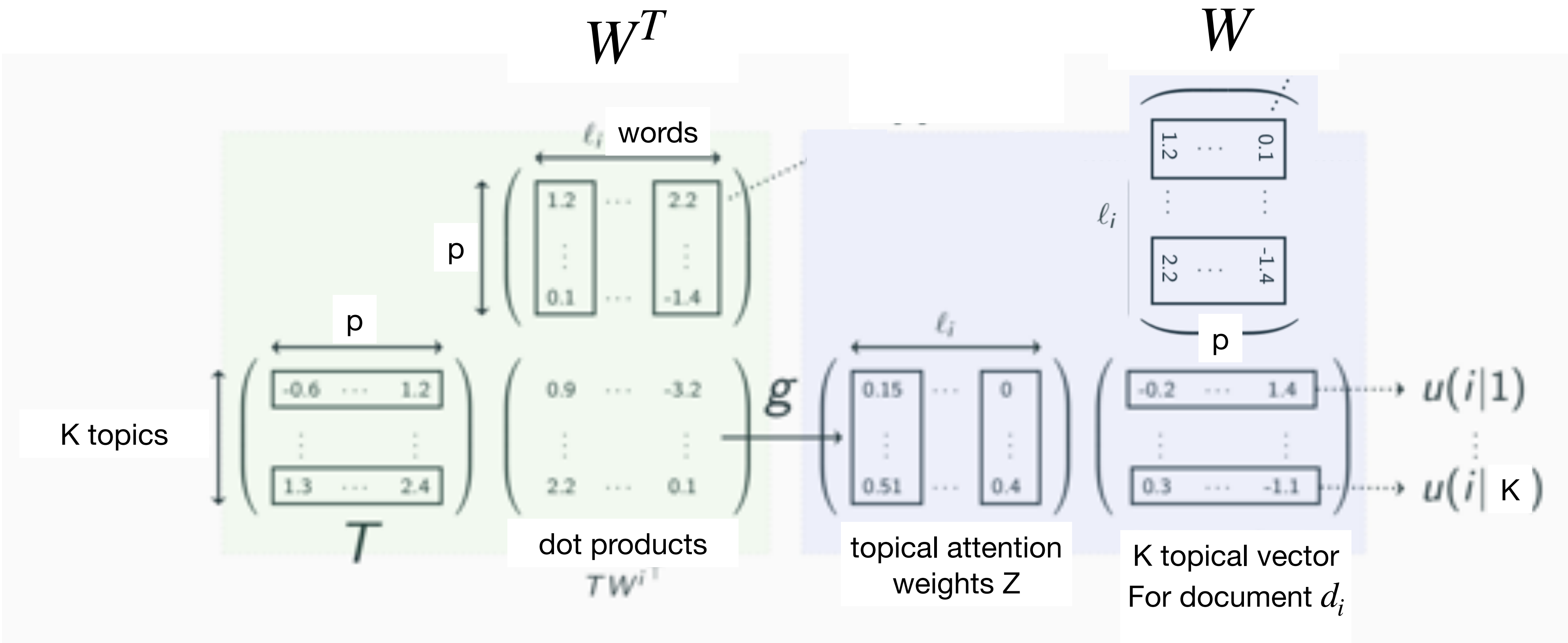
https://github.com/brochier/gvnr

Inductive Document Network Embedding (IDNE)

Brochier R., A. Guille and J. Velcin.
Inductive Document Network Embedding with Topic-Word Attention. ECIR, 2020



Topical attention



Document representation d_i is the normalized sum over the K topics: $\sum_{k=1}^{K} \frac{u(i\,|\,k)}{|x_i|_1}$

Learning IDNE

Minimize
$L(W, T) = \sum_{i=1}^{n_d} \sum_{j=1}^{n_d} s_{ij} \log \sigma(u_i \,.\, u_j) + (1 - s_{ij}) \log \sigma(-u_i \,.\, u_j)$

so that:

- S is a binary similarity matrix based on A, for instance:
  $s_{ij} = 1$ if $(A + A^2)_{ij} > 0$ else $s_{ij} = 0$

Results of IDNE on Cora

	TC										IC		TP	IP
	F1					AUC					F1	AUC	AUC	AUC
	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%	90%	90%	50%	90%
Aléatoire	13.85	13.70	12.66	12.83	12.53	51.51	51.96	51.60	51.22	51.22	11.80	48.84	49.87	50.27
TF	68.84	74.69	77.20	78.90	80.05	92.13	94.63	95.64	96.06	96.55	81.98	97.16	83.44	84.88
TF-IDF	72.56	76.99	80.13	80.66	81.84	93.11	95.24	96.38	96.74	97.20	**83.24**	**97.59**	85.17	85.02
LSA	72.46	77.22	79.92	80.56	81.16	93.65	95.47	96.63	96.95	97.17	81.26	97.35	87.17	88.63
DW+LSA	77.52	80.70	83.40	83.12	85.21	95.31	96.47	97.24	97.23	97.79	-	-	82.67	-
TADW	61.77	67.02	70.84	71.62	73.13	89.01	91.25	93.22	93.37	94.20	79.55	96.35	81.59	84.71
G2G	79.38	82.19	83.57	84.08	85.03	96.14	97.39	97.65	97.89	98.13	71.98	94.57	86.36	74.72
MATAN	75.48	76.65	77.58	79.20	78.30	95.19	95.79	96.23	96.41	96.69	76.13	94.91	82.72	71.47
GVNR-t	**82.20**	83.49	**85.26**	**85.82**	**86.67**	97.10	97.53	98.00	98.08	**98.48**	79.91	97.21	**94.31**	**92.44**
IDNE	80.41	**83.83**	84.79	84.67	86.06	**97.27**	**97.79**	**98.28**	**98.23**	98.45	82.25	97.57	92.90	88.56

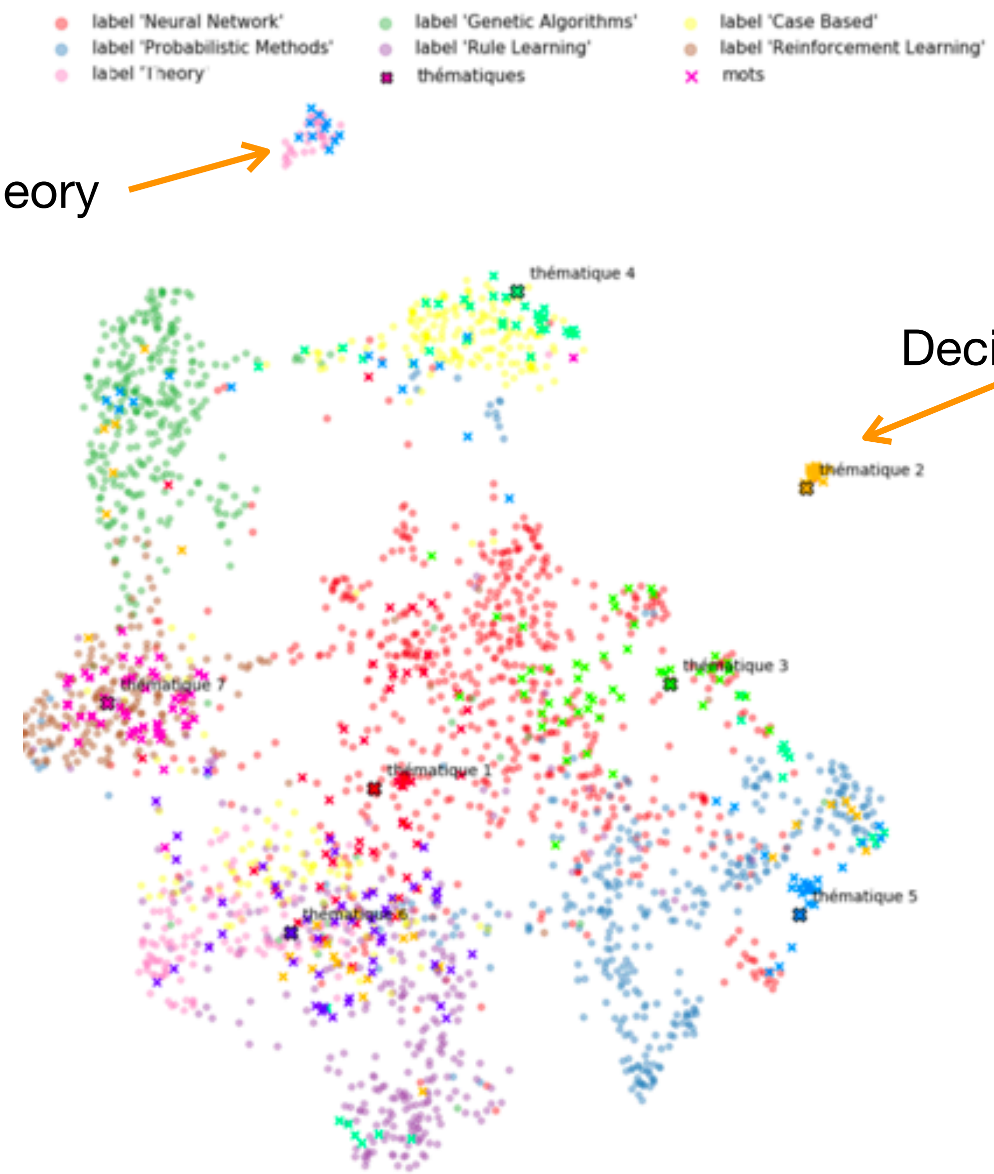
T for Transductive
I for Inductive

C = classification
P = link prediction

Observations

Thématique	Mots les plus proches
1	pattern, features, classification, regulate, machine, limitations, domains, datasets, methods, patterndirected, dataset, connectionist, backpropagation, pythia, accuracy, classifier, realworld, exemplar, symbolic, links
2	tree, decision, selection, trees, markov, radar, subsurface, moisture, cband, ers, cedar, sar, lband, polarized, minnesota, lter, ersjers, creek, raco, seasonality
3	saturation, separation, stabilization, gradient, blind, asymptotic, universal, matrix, square, signals, necessarily, sign, projection, approximated, bandpass, descent, norm, cubic, subproblems, speakers
4	design, casebased, pac, sme, analogical, mcmc, structuremapping, planning, wellunderstood, retrieval, reasoning, mistakes, case, reuse, retrieving, instance, chain, analogy, convolution, solving
5	mcmc, execution, speculation, ga, parallelism, genetic, instruction, aimed, dirichlet, instructions, chain, consensus, substantive, sequences, macroscopic, issue, processor, recipient, analogy, nerve
6	accuracy, learning, experiments, machine, ilp, datasets, comprehensibility, sufficed, sbc, inductive, accurate, decompose, warehouses, mining, induction, dataset, gratefully, attribute, assign, challenges
7	reinforcement, mdp, mdps, qlearning, policy, crosses, actions, value, reward, policies, brigade, rl, macros, relax, hinders, dynamic, satisficingoptimizing, action, functionings, barto

Observations (con't)



MCMC + theory

Decision trees

Gaussian Embedding of Linked Documents

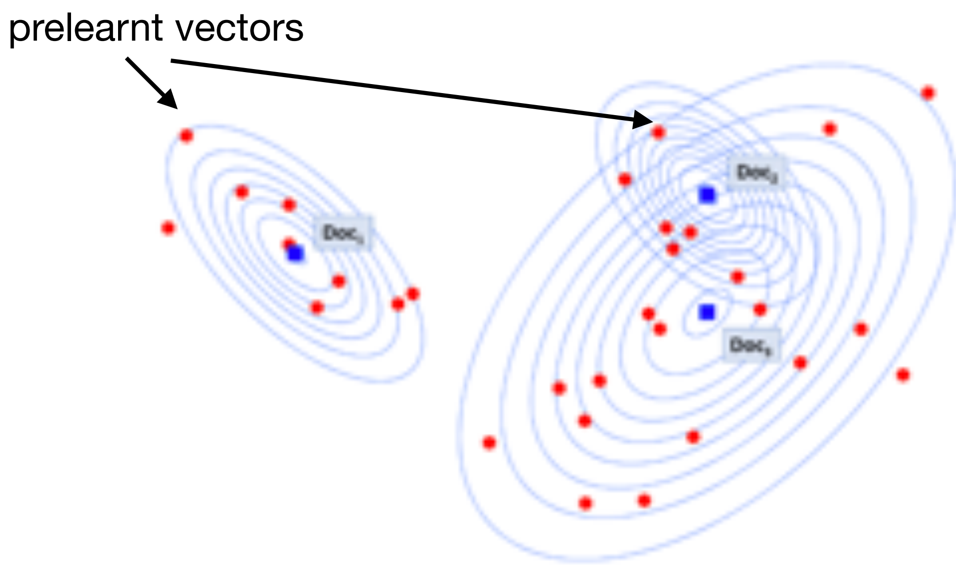
A. Gourru, J. Velcin and J. Jacques.
Gaussian Embedding of Linked Documents from a Pretrained Semantic Space, IJCAI 2020

- Each document d_i is associated to a mean and variance to capture an uncertainty degree
- $\mathscr{D}^w$ and $\mathscr{D}^l$ are bags of embedding vectors

$d_i^{(t)} \sim \mathcal{N}(\mu_i, \sigma_i^2 I)$ with $\mu_i, \sigma_i^2 \in \mathbb{R}^r$ and $\sigma_i^2 > 0$

$$\begin{array}{cccccc} & \textit{Everybody} & \textit{loves} & \textit{four} & \textit{cheese} & \textit{pizza} \\ \mathcal{D}_i^w = & \left\{ \begin{array}{c} -0.36 \\ -0.01 \\ 0.21 \end{array} \right. & \begin{array}{c} -0.38 \\ 0.68 \\ -0.54 \end{array} & \begin{array}{c} 0.34 \\ 0.45 \\ 0.55 \end{array} & \begin{array}{c} -0.05 \\ -0.31 \\ -1.32 \end{array} & \left. \begin{array}{c} 0.62 \\ -0.49 \\ -0.69 \end{array} \right\} \end{array}$$

$$\mathcal{L}(\mathcal{D}; \mu, \sigma^2) = \sum_{i=1}^{n} \sum_{f(w) \in \mathcal{D}_i^w} \log \mathcal{N}(f(w); \mu_i, \sigma_i^2 I) + \sum_{i=1}^{n} \sum_{g(d) \in \mathcal{D}_i^l} \log \mathcal{N}(g(d); \mu_i, \sigma_i^2 I)$$



Results of GELD (1)

	Cora		Dblp		Nyt	
Train/Test ratio	10%	50%	10%	50%	10%	50%
DeepWalk	70.6 (2.0)	81.0 (0.7)	52.3 (0.4)	53.5 (0.2)	66.9 (0.7)	68.7 (0.9)
LSA	72.3 (1.9)	80.6 (0.7)	73.5 (0.2)	74.2 (0.2)	71.6 (1.0)	76.7 (0.7)
Concatenation	71.4 (2.1)	84.0 (1.1)	77.5 (0.2)	78.2(0.2)	77.9 (0.3)	81.1 (0.7)
TADW	81.9 (0.8)	87.4 (0.8)	74.8 (0.1)	75.5 (0.1)	75.8 (0.5)	79.4 (0.4)
AANE	79.8 (0.9)	84.4 (0.7)	73.3 (0.1)	74.2 (0.2)	71.7 (0.5)	76.9 (1.1)
GVNR-t	83.7 (1.2)	87.0 (0.8)	69.6 (0.1)	70.2 (0.2)	74.3 (0.4)	76.7 (0.6)
RLE	84.0 (1.3)	87.7 (0.6)	79.8(0.2)	81.2 (0.1)	77.7 (0.7)	80.0 (0.6)
VGAE	72.3 (1.7)	81.1 (0.7)	Memory Overflow		68.1 (0.8)	70.1 (0.6)
G2G	79.0 (1.5)	84.8 (0.7)	70.8 (0.1)	71.5 (0.2)	69.0 (0.5)	71.5 (0.8)
STNE	79.4 (1.0)	86.7 (0.8)	73.8 (0.2)	74.5 (0.1)	75.1 (0.7)	78.1 (0.6)
GELD	**84.3** (1.1)	**88.3** (0.4)	**81.63** (0.1)	**82.3** (0.1)	**78.5** (0.8)	**81.2** (0.3)

Results of GELD (2)

simple class centroids

Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
network	reinforcement	posterior	pac	genetic	casebased	ilp
networks	rl	bayesian	schapire	ga	knowledge	clause
neural	barto	gibbs	error	mutation	reasoning	kira
feedforward	qlearning	models	queries	gp	experiences	literals
multilayer	multiagent	model	set	search	design	relational

Table 3: Class descriptions on Cora. We show the top 5 words closest to the class centroids.

Title	Variance	Class
Collective Latent Dirichlet Allocation	545	3
Spatial Latent Dirichlet Allocation	605	1
Distributed Inference for Latent Dirichlet Allocation	590	1
Fast collapsed gibbs sampling for latent dirichlet allocation	513	3
Latent Dirichlet Co-Clustering.	398	3
A perceptual hashing algorithm using latent dirichlet allocation	604	2

Table 4: Six Nearest Neighbors in the embedding for the article "Latent Dirichlet Allocation" by Blei et al., obtained on Dblp. We provide the total variance summed by axes.

Conclusion
and future works

Conclusion

- Several contributions on the embedding of documents augmented with network information:
RLE, GVNR-t, MATAN, IDNE, GELD (and more to come ;)

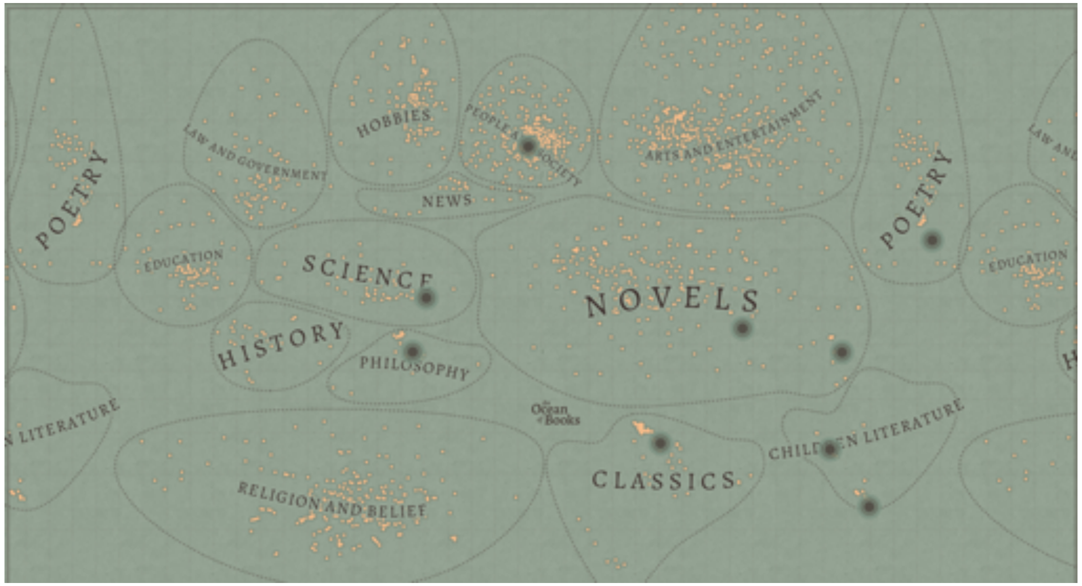
- Use of “absolute” WE leads to good results. Can they be improved using contextualized WE (Devlin et al., 2018)

- Recent advances in GNN should be considered in the future, e.g. GCN (Kipf & Welling, 2017) and GAT (Velikcovik et al., 2018)

Toward author embedding

- PhD thesis of **A. Gourru**: moving to author embedding (Ganesh et al., 2016), modeling dynamics similarly to (Balmer et al., 2017), using the Variational Information Bottleneck framework (Tishby et al., 1999)

- PhD thesis of **E. Terreau**: building a better space to represent the authors by taking their *style* into account (Yand et al. 2018), with applications to the LIFRANUM project



https://artsexperiments.withgoogle.com/ocean-of-books

References

- Brochier R., A. Guille and J. Velcin. Inductive Document Network Embedding with Topic-Word Attention. **ECIR** 2020 (virtual).
- Brochier R., A. Guille and J. Velcin. Link Prediction with Mutual Attention for Text-Attributed Networks. **Workshop on Deep Learning for Graphs and Structured Data Embedding**, colocated with WWW (Companion Volume), May 13–17, 2019, San Francisco, CA, USA.
- Brochier R., A. Guille and J. Velcin. Global Vectors for Node Representation. The Web Conference (**WWW**), May 13–17, 2019, San Francisco, CA, USA.
- Gourru A., J. Velcin, J. Jacques and A. Guille Document Network Projection in Pretrained Word Embedding Space. **ECIR** 2020 (virtual).
- Gourru A., J. Velcin and J. Jacques. Gaussian Embedding of Linked Documents from a Pretrained Semantic Space. **IJCAI** 2020 (virtual).

➡ Code for GVNR and GVNR-t: https://github.com/brochier/gvnr

➡ Code for IDNE: https://github.com/brochier/idne

➡ Code for RLE and GELD: https://github.com/AntoineGourru/DNEmbedding