

# Wikipedia Networks

Big Data Course – Project Presentation

Nina Varchavsky-Bergin

18.12.2019



# Wikipedia is community-driven



# Wikipedia is community-driven



*Is there an impact of  
language and culture on  
wikipedia article  
networks structure ?*

# The dataset:

- French

*Apprentissage automatique*

- English

*Machine Learning*

## **Id**

Title

Content of the page

URL

Length of the article

## **Links to other articles**

Last modification date

Wikibase number

Wikidata URL

Aliases

**Nodes:** articles

**Edges:** hyperlinks

*Scraped thanks to wptools module*

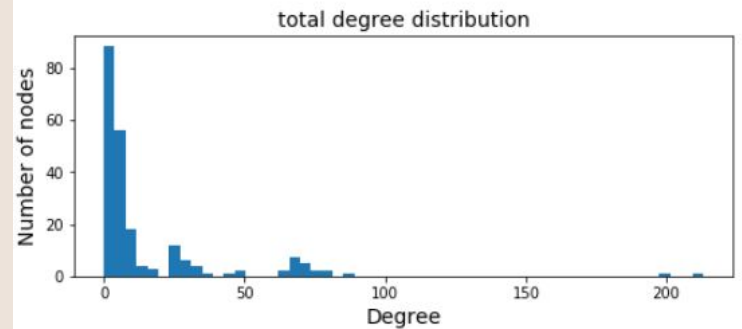
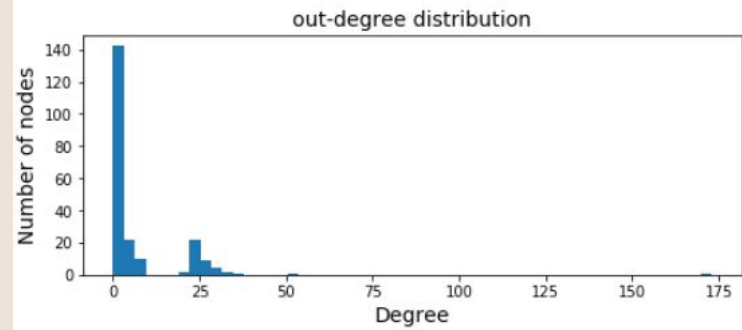
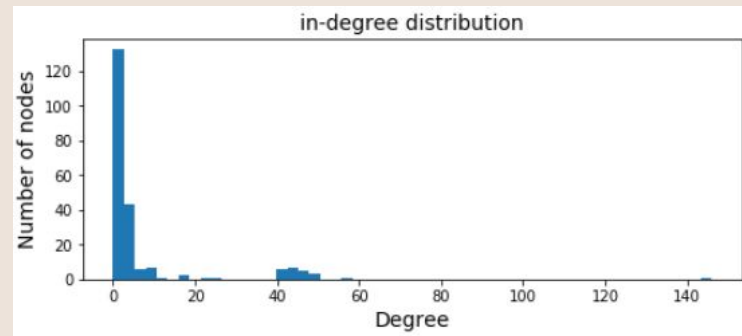
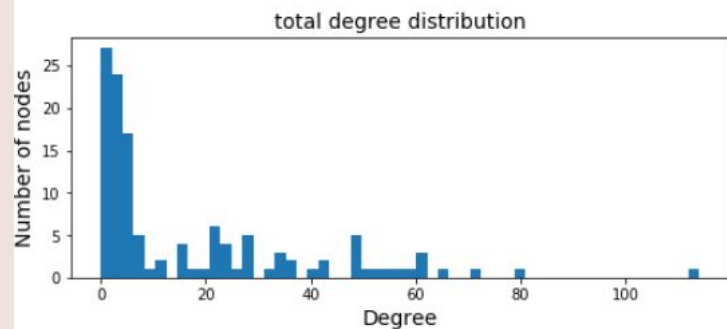
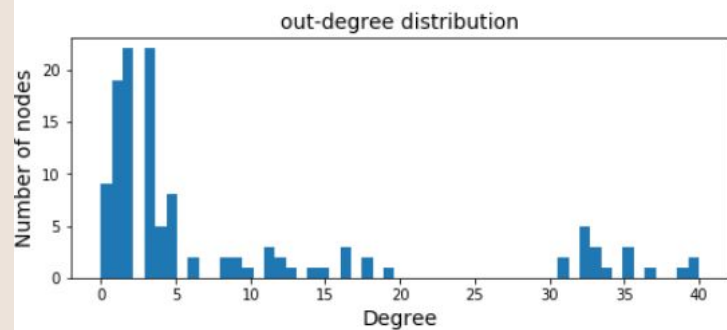
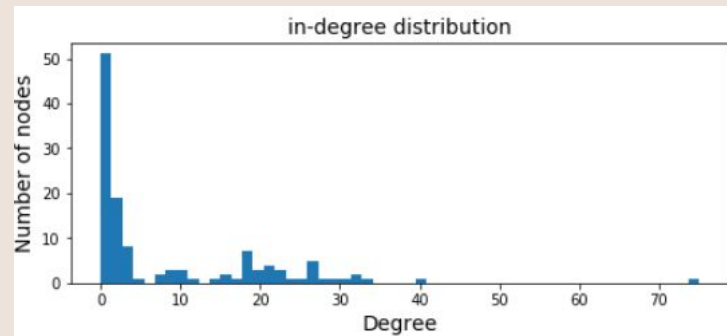
# French

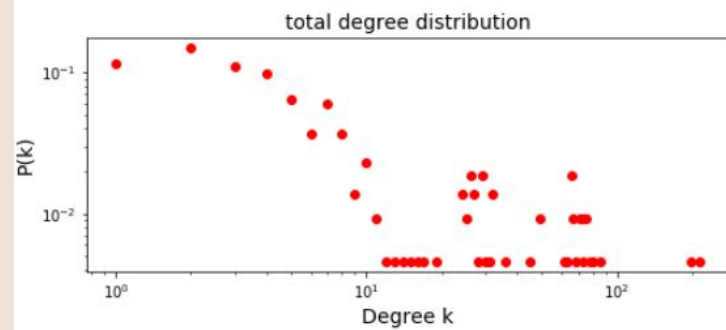
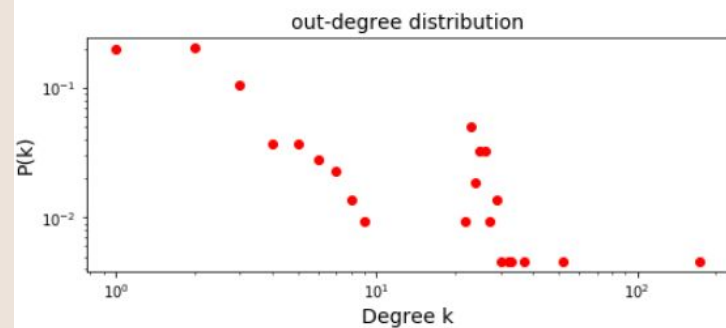
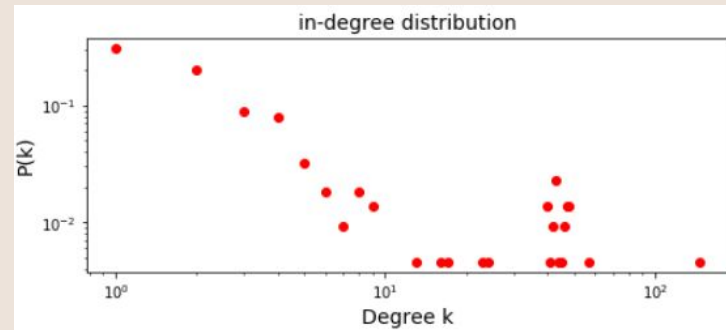
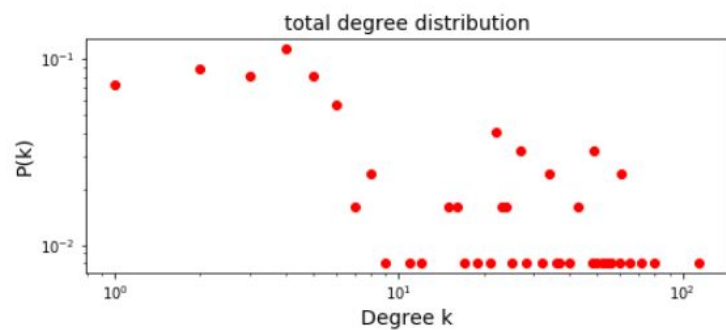
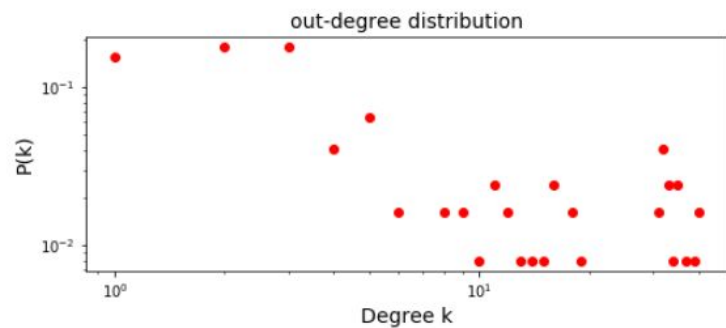
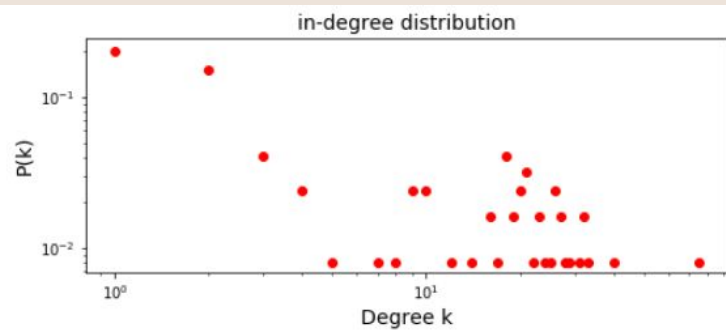
# English

<p> <b>Apprentissage par transfert</b>          Réduction du biais dimensionnalité          Mécanisme temporel et sélectif          Transferts on Neural Machine          Extrême learning machine          BERT (traitement automatique du langage)          Apprentissage par transfert          Recherche des piliers proches voisins          Astuce du noyau (méthode de la fibre de jonction)          Méthode des k plus proches voisins          Charni-Jaelle          Analyse formelle de concepts          Apprentissage semi-supervisé          Apprentissage automatique          Machine à vecteurs de support          Fashion MNIST          Réseaux bayésiens          Classification bayésienne          Dimension du vecteur          Complexité algorithmique          Théorie de la complexité          Classification des classes multiples          Théorème de Cox-Jaynes          Réseaux de neurones       </p>	<p>         Réseaux de neurones artificiels          Apprentissage ensembliste          Apprentissage basé sur l'explication          Algorithmes de Markov          Algorithme I03          Naïve Bayes et Learning Systems          Perceptron linéaire          Réseaux de neurones          Apprentissage automatique appliqué aux systèmes de détection d'intrusion réseau          Propagation des cas convectifs          Algorithmes de la grille          QUEST (arbre de décision)          Les Learning Systems          Algorithmes de la grille          Analyse formelle de concepts          Machine Learning          Apprentissage automatique          Machine à vecteurs de support          Réseaux bayésiens          Réseaux de neurones          Classification bayésienne          Dimension du vecteur          Complexité algorithmique          Théorie de la complexité          Classification des classes multiples          Théorème de Cox-Jaynes          Réseaux de neurones       </p>	<p>         Développement de nouvelles routines          Pattern language (Routledge)          Confusion matrix          Instantaneously trained neural network          Association Classification          Dictionary of artificial intelligence          Vanishing gradient problem          Sparse dictionary learning          Connectionist learning          Sparse matrix learning          Cross-entropy          Pattern recognition          Preference learning          Dimensionality reduction          Local case-control sampling          Machine learning in bioinformatics          Coupled pattern learning          Structured sparsity regularization          Transduction (machine learning)          Stochastic block model          Statistical classification          Documenting bias          Anomaly detection          Bayesian structural time series          Confusion matrix          Confusion matrix analysis          Conventional control          Large margin (nearest neighbor)       </p>	<p>         Développement de nouvelles routines          Pattern language (Routledge)          Confusion matrix          Instantaneously trained neural network          Association Classification          Dictionary of artificial intelligence          Vanishing gradient problem          Sparse dictionary learning          Connectionist learning          Sparse matrix learning          Cross-entropy          Pattern recognition          Preference learning          Dimensionality reduction          Local case-control sampling          Machine learning in bioinformatics          Coupled pattern learning          Structured sparsity regularization          Transduction (machine learning)          Stochastic block model          Statistical classification          Documenting bias          Anomaly detection          Bayesian structural time series          Confusion matrix          Confusion matrix analysis          Conventional control          Large margin (nearest neighbor)       </p>	<p>         Kernel embeddings of expert opinions          Multitask portfolio algorithm          Multitask portfolio algorithm          Hyperparameter optimization          Sparse matrix learning          Predictive state representation          Matrix regularization          Sparse dictionary learning          Connectionist learning          Sparse matrix learning          Cross-entropy          Pattern recognition          Preference learning          Dimensionality reduction          Local case-control sampling          Machine learning in bioinformatics          Coupled pattern learning          Structured sparsity regularization          Transduction (machine learning)          Stochastic block model          Statistical classification          Documenting bias          Anomaly detection          Bayesian structural time series          Confusion matrix          Confusion matrix analysis          Conventional control          Large margin (nearest neighbor)       </p>	<p>         Action model learning          Kernel embeddings of expert opinions          Multitask portfolio algorithm          Multitask portfolio algorithm          Hyperparameter optimization          Sparse matrix learning          Predictive state representation          Matrix regularization          Sparse dictionary learning          Connectionist learning          Sparse matrix learning          Cross-entropy          Pattern recognition          Preference learning          Dimensionality reduction          Local case-control sampling          Machine learning in bioinformatics          Coupled pattern learning          Structured sparsity regularization          Transduction (machine learning)          Stochastic block model          Statistical classification          Documenting bias          Anomaly detection          Bayesian structural time series          Confusion matrix          Confusion matrix analysis          Conventional control          Large margin (nearest neighbor)       </p>
---	---	--	--	---	--

# Basic characteristics

	French	English
<b>Nodes</b>	124	216
<b>Edges</b>	1063	1633
<b>Ratio nodes/edges</b>	0.12	0.13
<b>In degree</b>	8.57	7.56
<b>Out degree</b>	8.57	7.56
<b>Total degree</b>	17.14	15.12







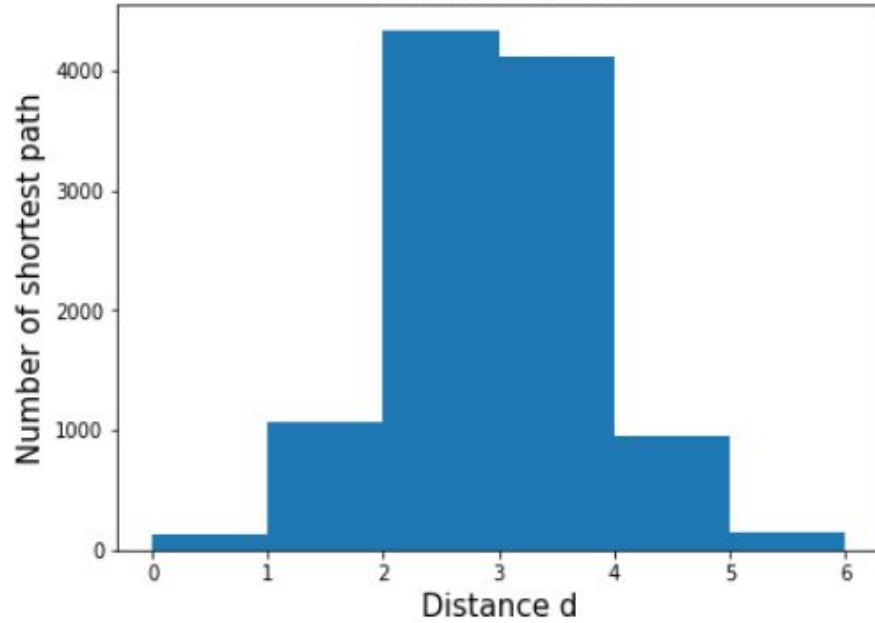
# Top 10 total degree nodes

	title	degree
0	Apprentissage automatique	75
1	Réseau de neurones artificiels	40
2	TensorFlow	33
3	Keras	32
4	Apprentissage supervisé	32
5	Theano (logiciel)	31
6	Méthode des k plus proches voisins	29
7	Apprentissage non supervisé	28
8	Microsoft Cognitive Toolkit	27
9	Scikit-learn	27



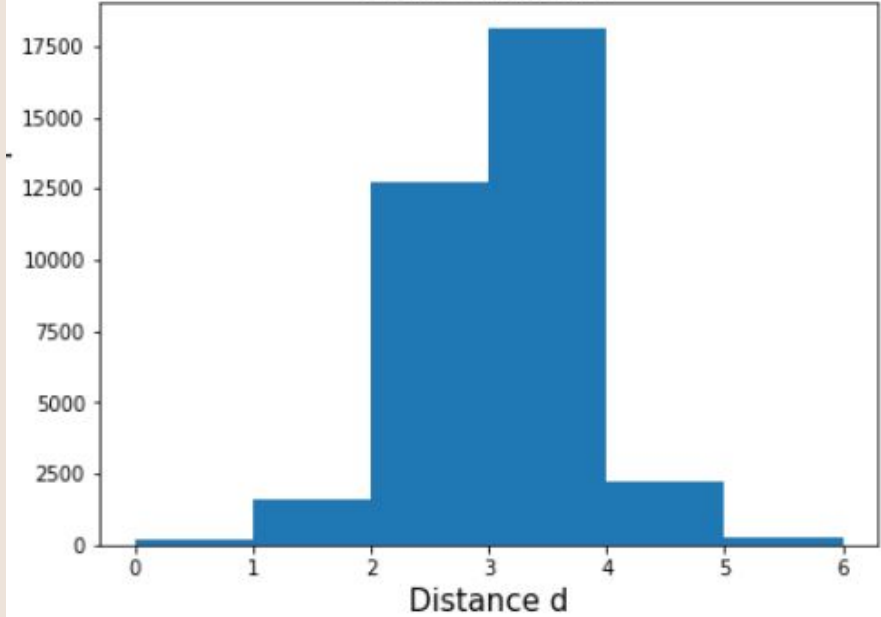
	title	degree
0	Machine learning	146
1	Statistical classification	57
2	Convolutional neural network	48
3	Statistical learning theory	48
4	Computational learning theory	48
5	Machine Learning (journal)	47
6	Empirical risk minimization	47
7	Unsupervised learning	47
8	Semi-supervised learning	46
9	Dimensionality reduction	46

Distance distribution



Mean: 2.48  
Variance: 0.78

Distance distribution



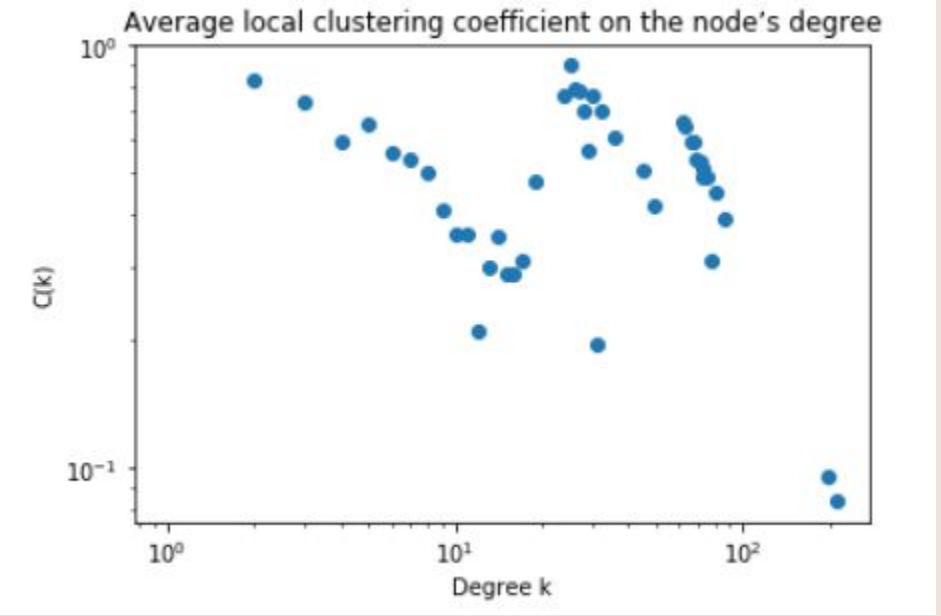
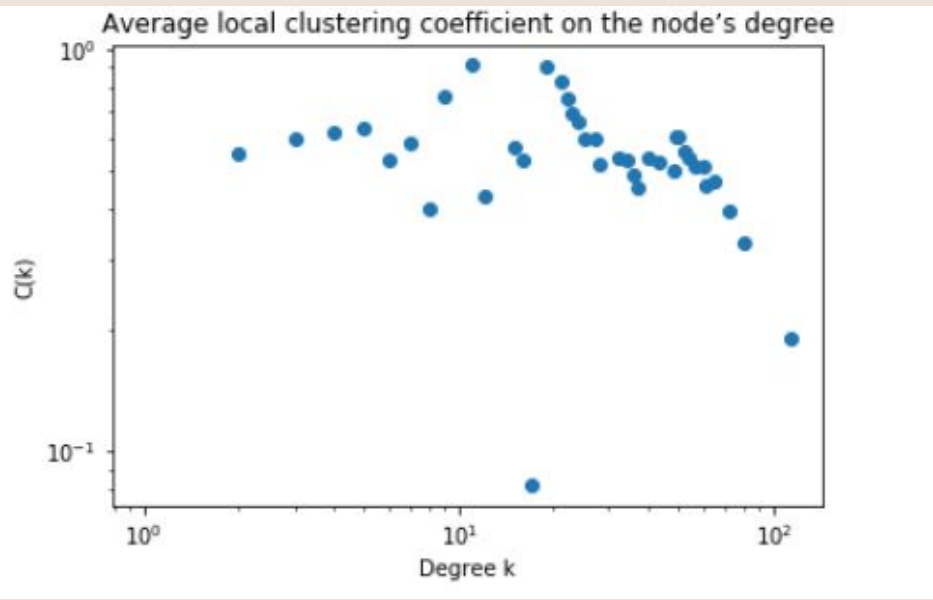
Mean: 2.61  
Variance: 0.55

# Connected components

	French	English
<b>Weak</b>	8	8
<b>Multi-node, weak</b>	1 (117 nodes)	1 (209 nodes)
<b>Strong</b>	31	51
<b>Multi-node, strong</b>	3 (2-3-91 nodes)	1 (166 nodes)

# Main connected components characteristics

		French	English
Average shortest path length	Weak cc	1.96	2.11
	Strong cc	2.46	2.58
Diameter	Strong cc	6	6



Average clustering coeff: 0.5

Average clustering coeff: 0.53

# Conclusion

## Similarities:

- Directed and disconnected
- Scale free networks
- Small-world structure

## Differences:

- More nodes in English than in French

# Next steps

Metadata exploration

Community detection

Bipartite En-Fr Network  
analysis

# Notebook and wpnetwork module on my GitHub

[github.com/Ninanouchka/  
wikipedia-article-network](https://github.com/Ninanouchka/wikipedia-article-network)



# Bibliography

1. Barabási, Albert-László. *Network Science*. Consulté le 17 décembre 2019. <http://networksciencebook.com/>.
2. Arenas, A., L. Danon, A. Díaz-Guilera, P. M. Gleiser, et R. Guimerá. « Community Analysis in Social Networks ». *The European Physical Journal B* 38, n° 2 (1 mars 2004): 373-80. <https://doi.org/10.1140/epjb/e2004-00130-1>.
3. Fujiwara, Yuya, Yu Suzuki, Yukio Konishi, et Akiyo Nadamoto. « Extracting Difference Information from Multilingual Wikipedia ». In *Web Technologies and Applications*, édité par Quan Z. Sheng, Guoren Wang, Christian S. Jensen, et Guandong Xu, 496-503. *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer, 2012. [https://doi.org/10.1007/978-3-642-29253-8\\_42](https://doi.org/10.1007/978-3-642-29253-8_42).
4. Massa, Paolo, et Federico Scrinzi. « Manypedia:ComparingLanguagePointsof Viewof Wikipedia Communities », s. d., 9. <https://www.opensym.org/ws2012/p13wikisym2012.pdf>
5. Pfeil, Ulrike, Panayiotis Zaphiris, et Chee Siang Ang. « Cultural Differences in Collaborative Authoring of Wikipedia ». *Journal of Computer-Mediated Communication* 12, n° 1 (1 octobre 2006): 88-113. <https://doi.org/10.1111/j.1083-6101.2006.00316.x>.
6. Nemoto, Keiichi, et Peter A. Gloor. « Analyzing Cultural Differences in Collaborative Innovation Networks by Analyzing Editing Behavior in Different-Language Wikipedias ». *Procedia - Social and Behavioral Sciences*, The 2nd Collaborative Innovation Networks Conference - COINs2010, 26 (1 janvier 2011): 180-90. <https://doi.org/10.1016/j.sbspro.2011.10.574>.