Le Web, un graphe à fouiller

Pierre Senellart

TELECOM ParisTech

24 juillet 2008

Le graphe du Web

Le World Wide Web vu comme un graphe (orienté) :

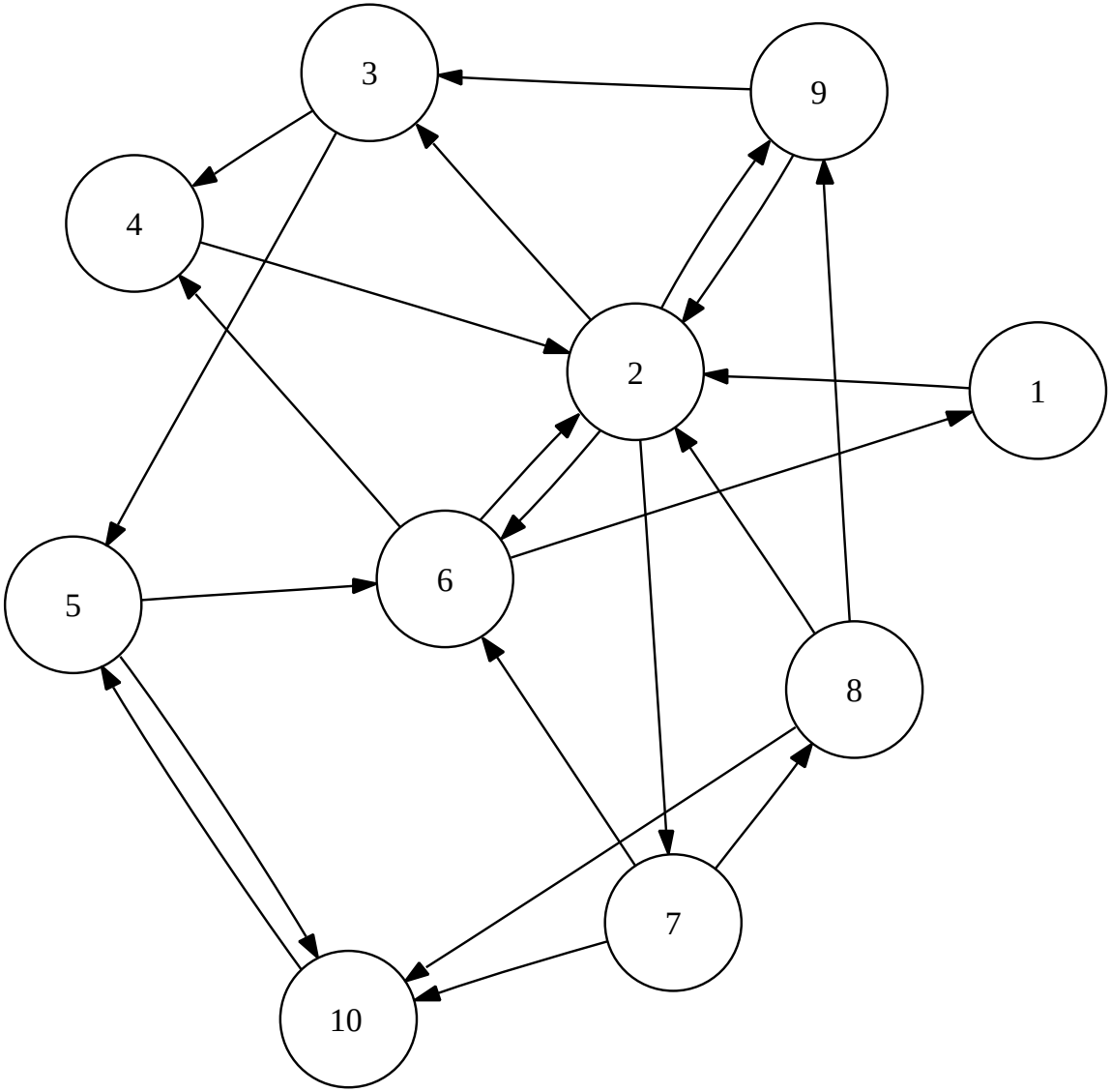
Nœuds : pages Web ;

Arêtes : liens entre pages.

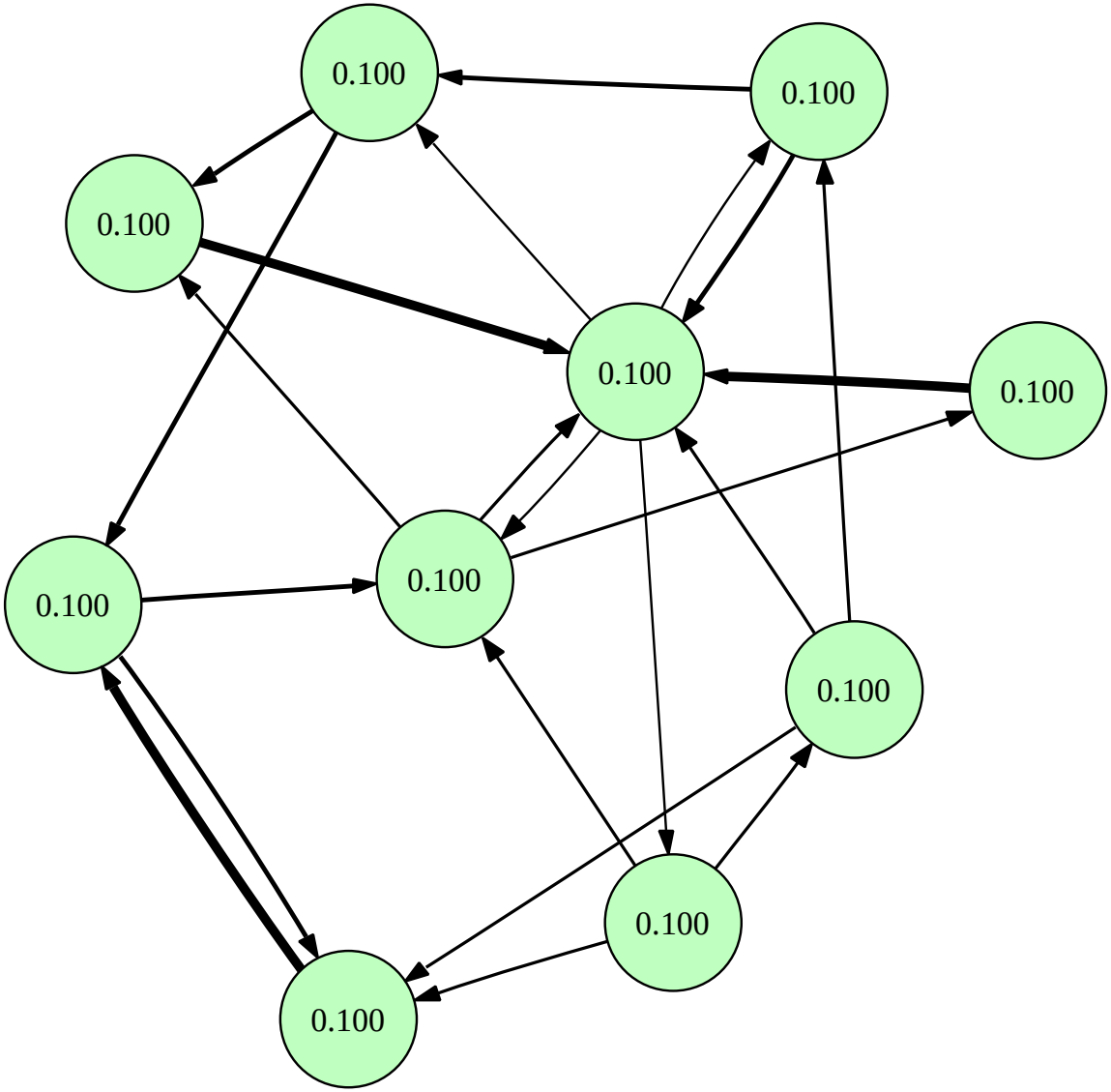
Idem pour d'autres environnements hypertextes :

- dictionnaire monolingue ;
- encyclopédie ;
- publications scientifiques, avec liens de citation.

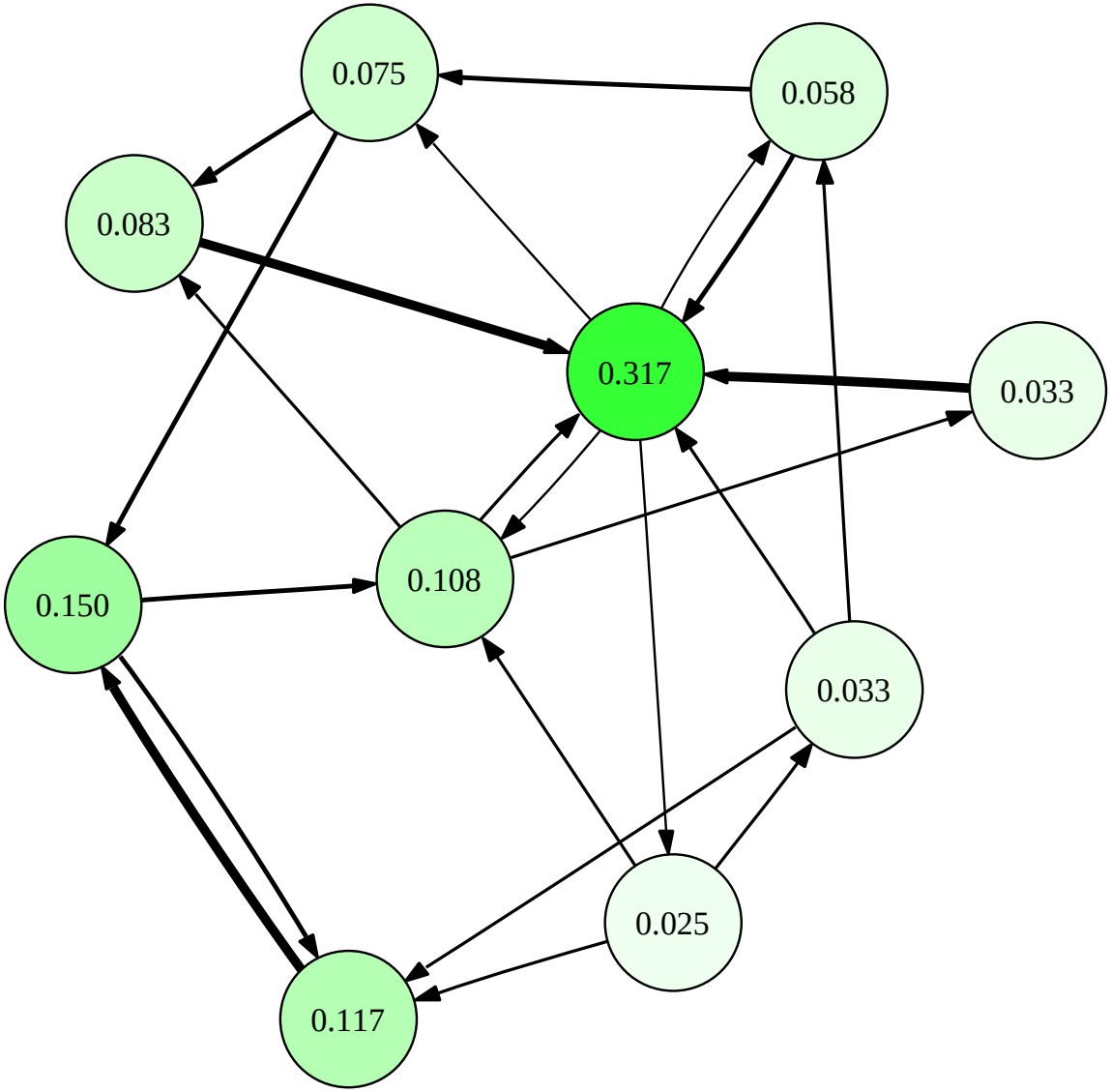
Introduction à PageRank



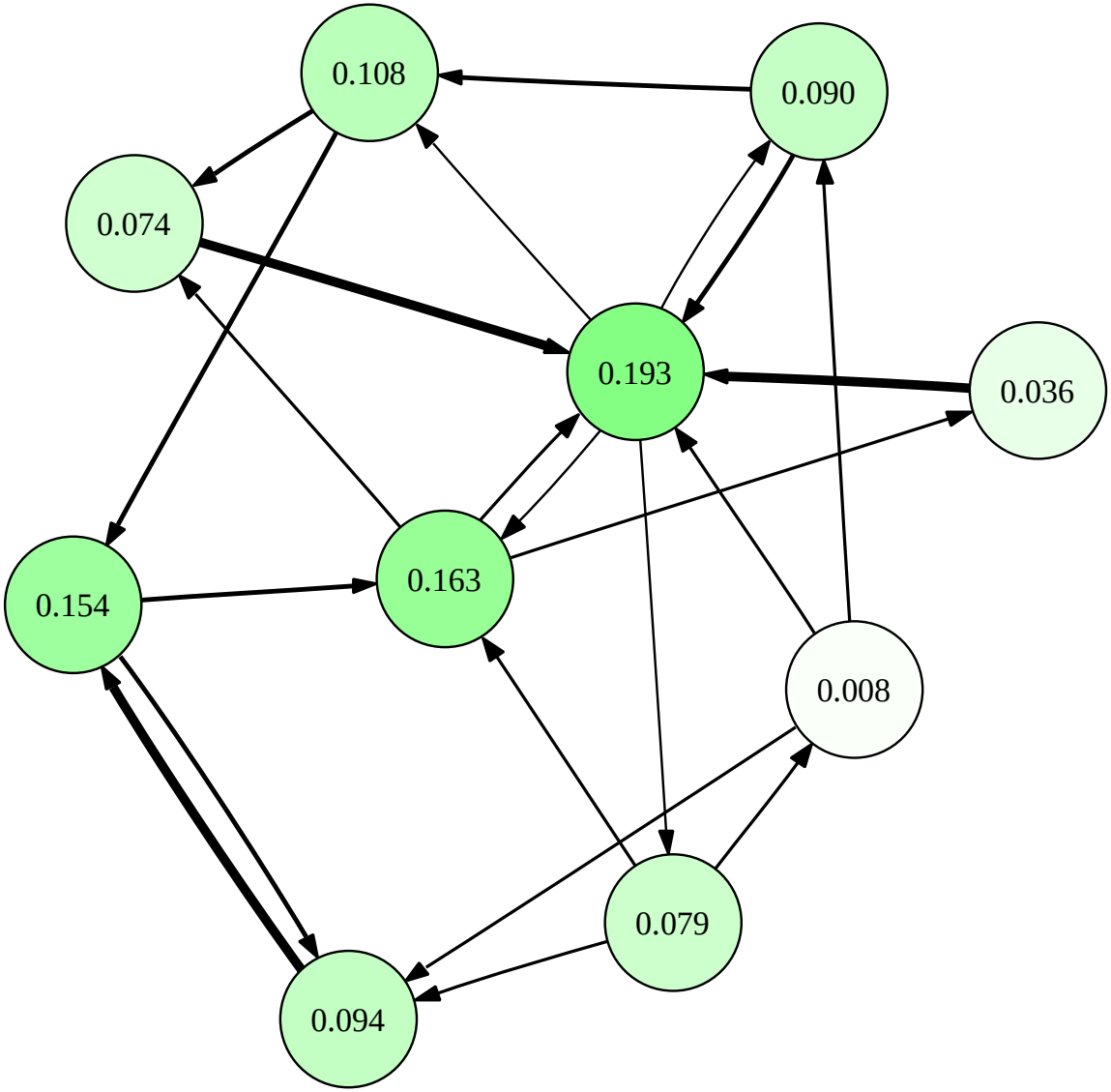
Introduction à PageRank



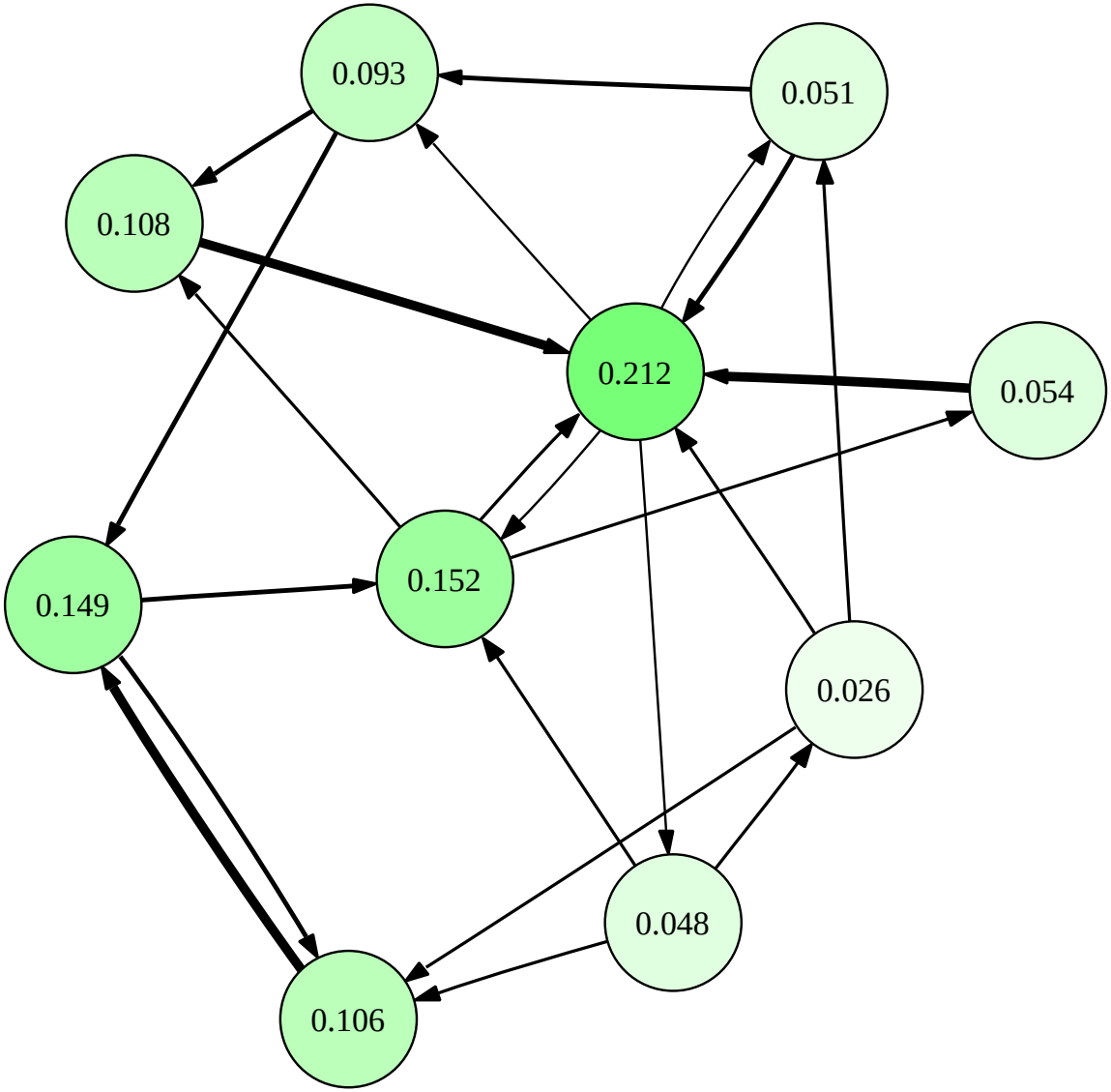
Introduction à PageRank



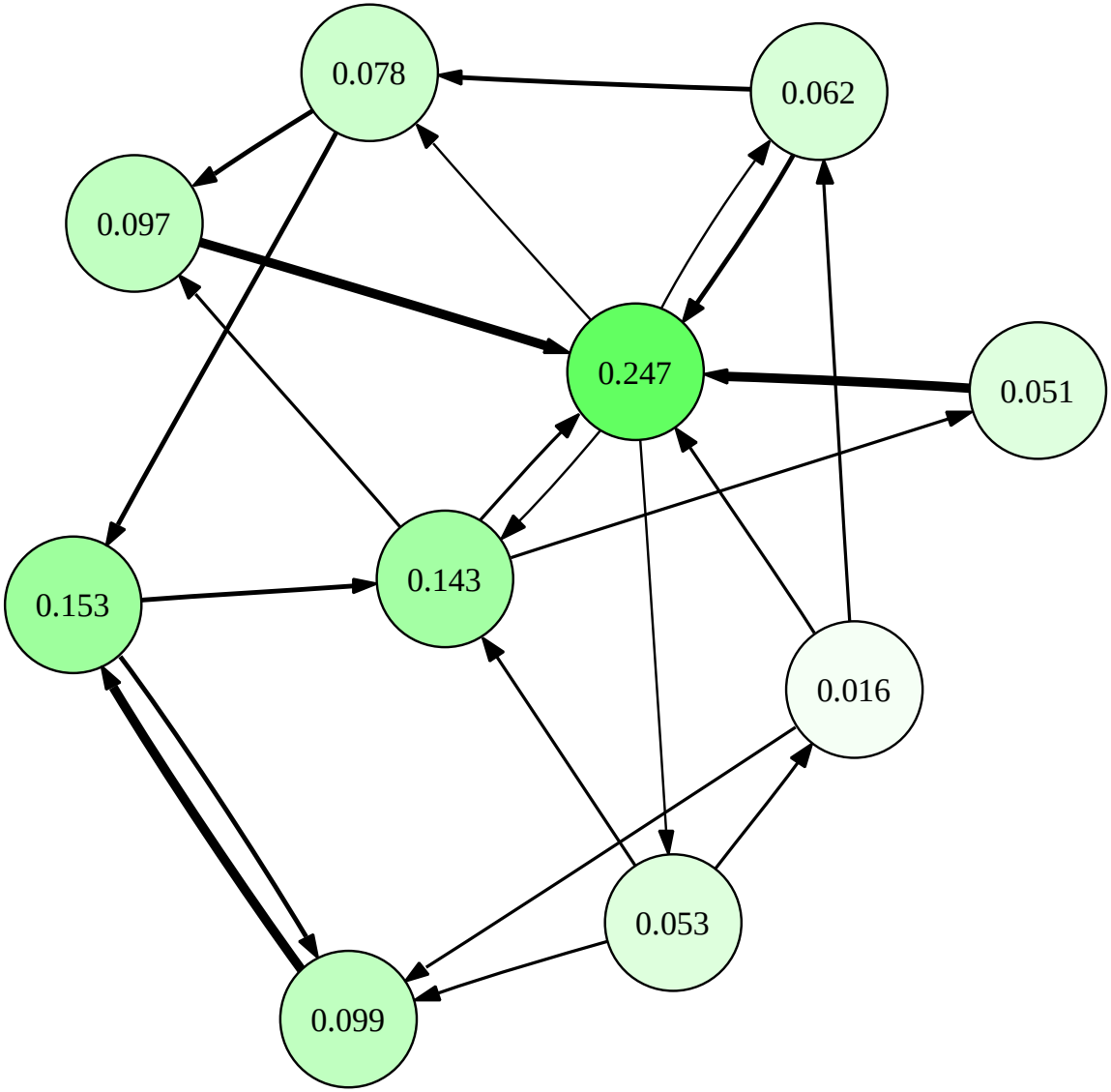
Introduction à PageRank



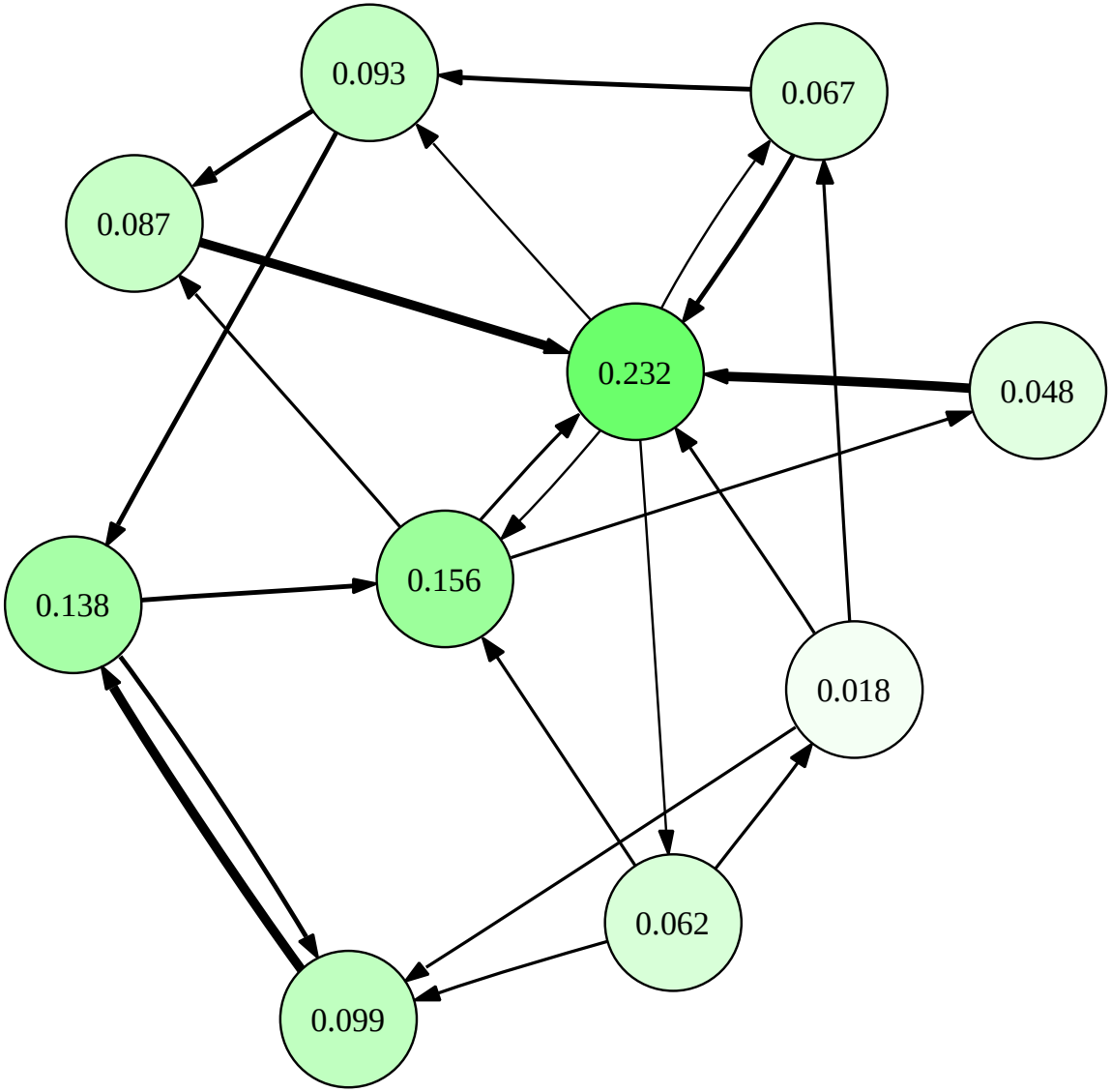
Introduction à PageRank



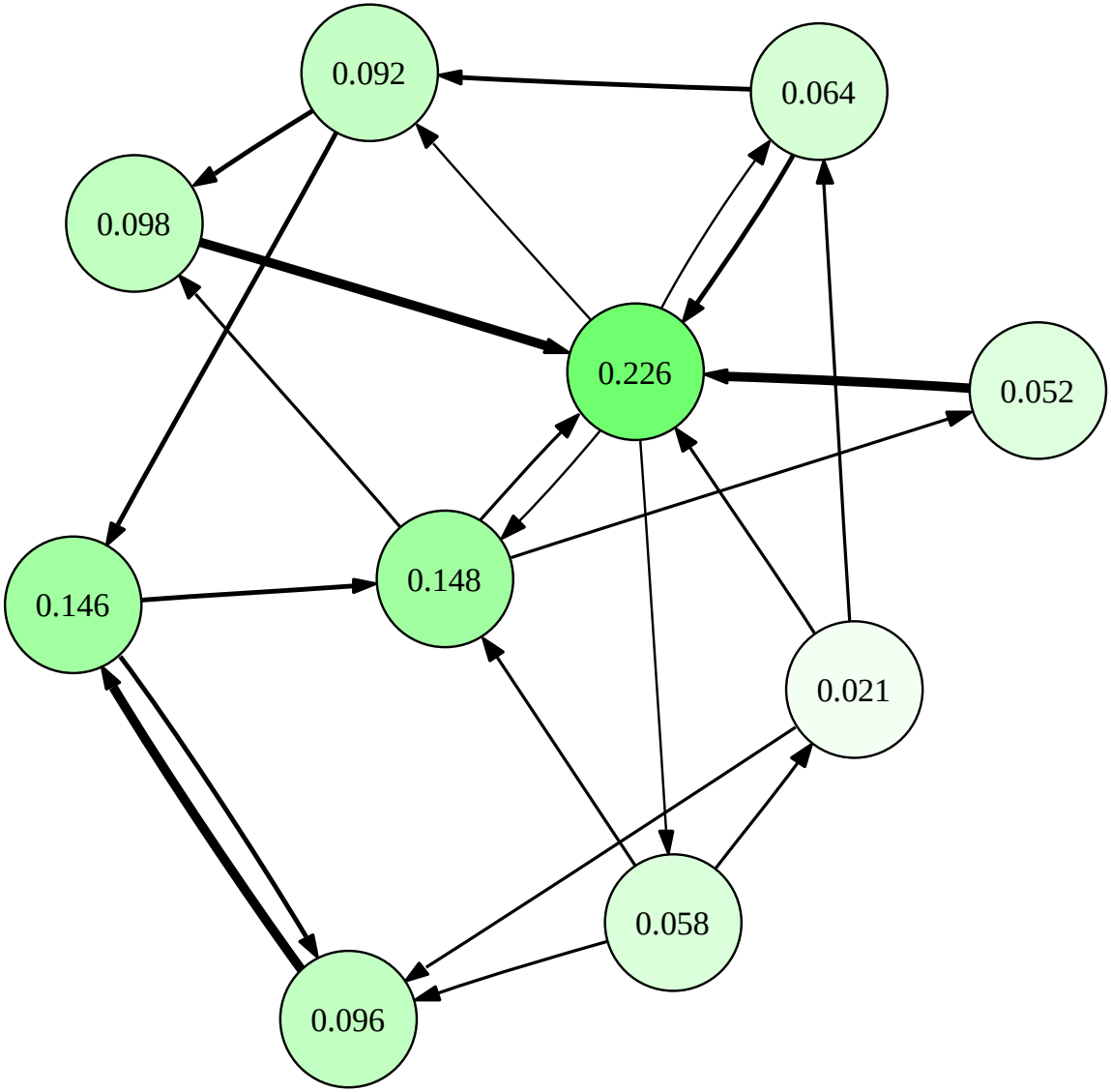
Introduction à PageRank



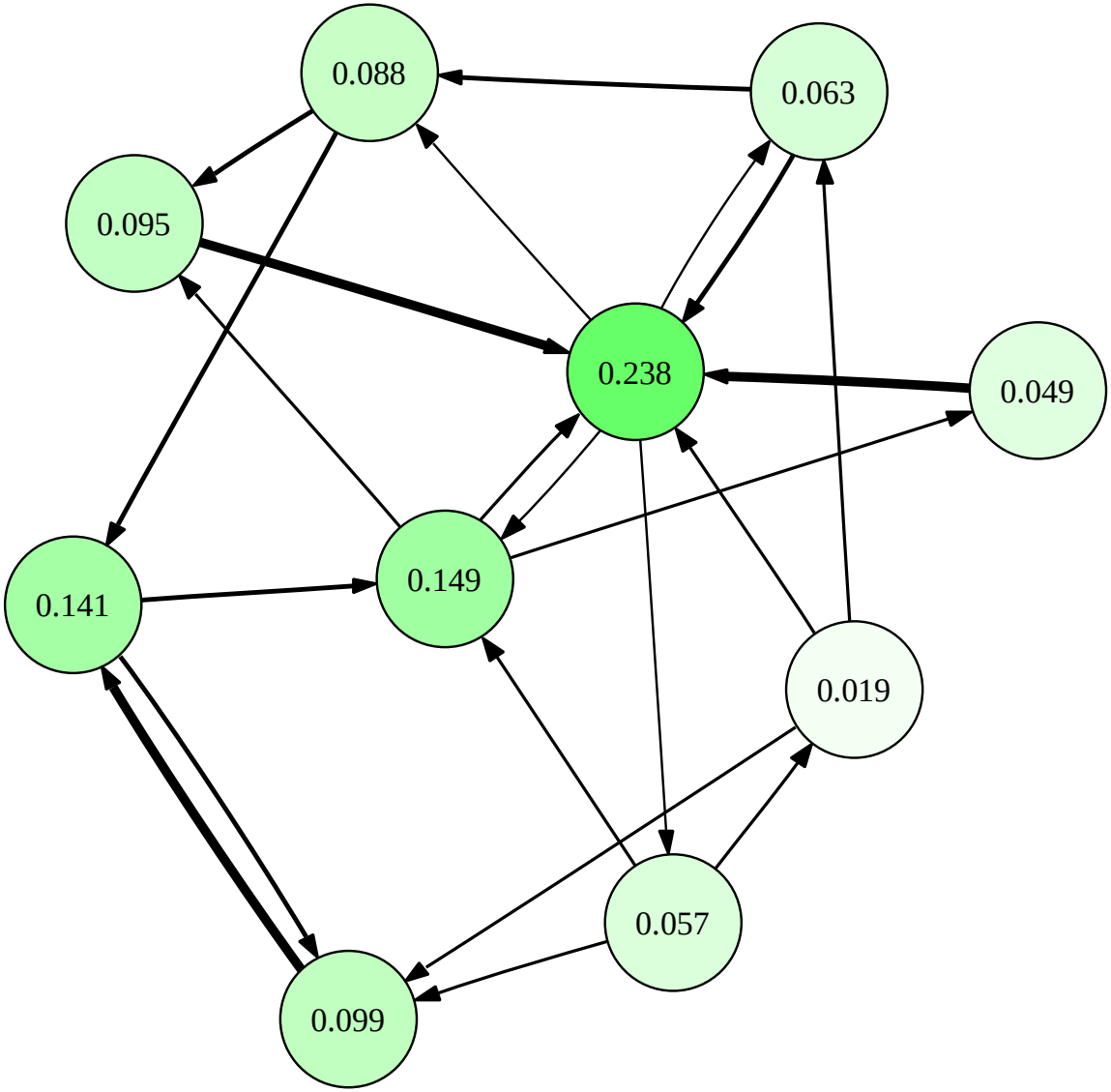
Introduction à PageRank



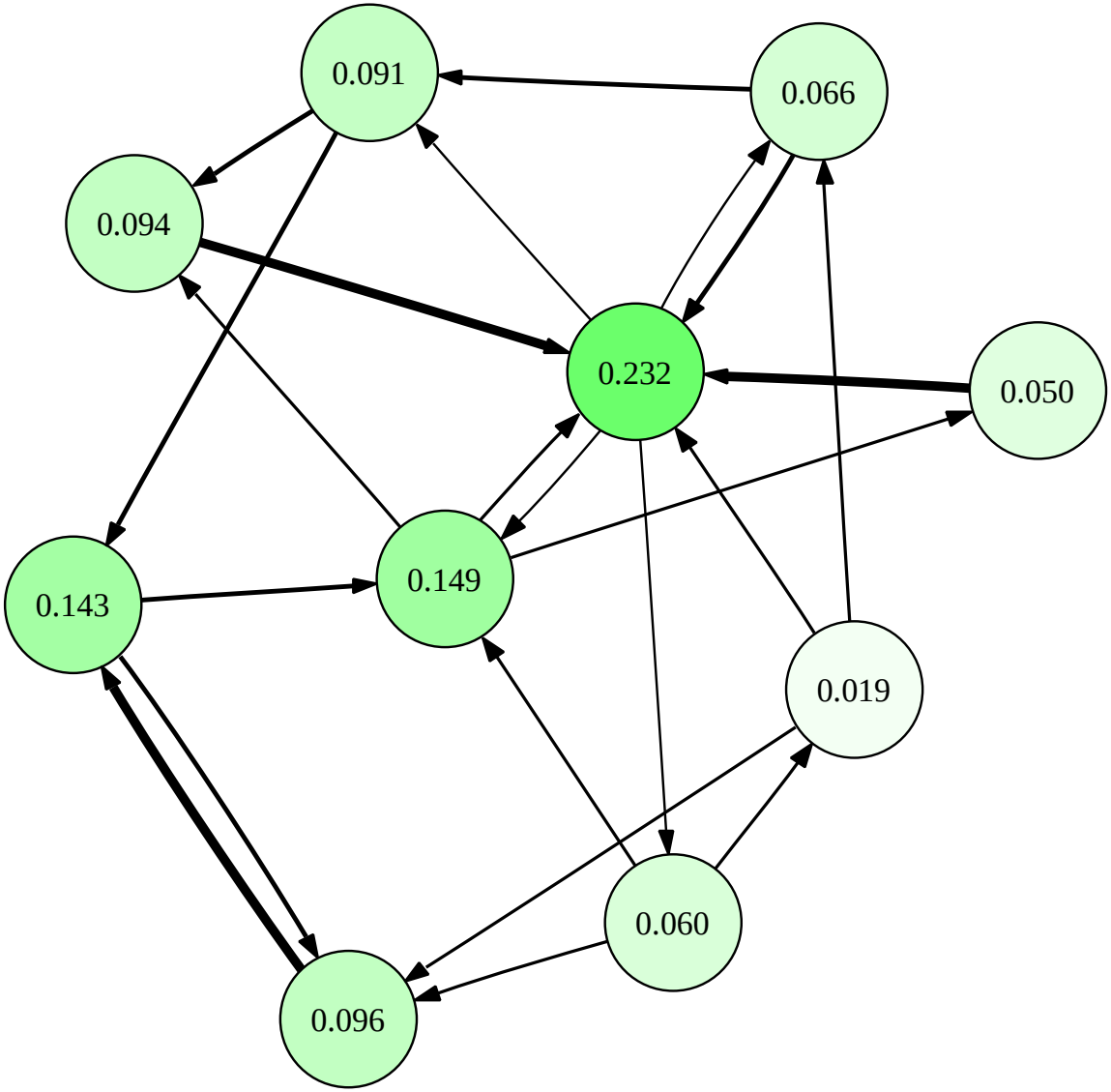
Introduction à PageRank



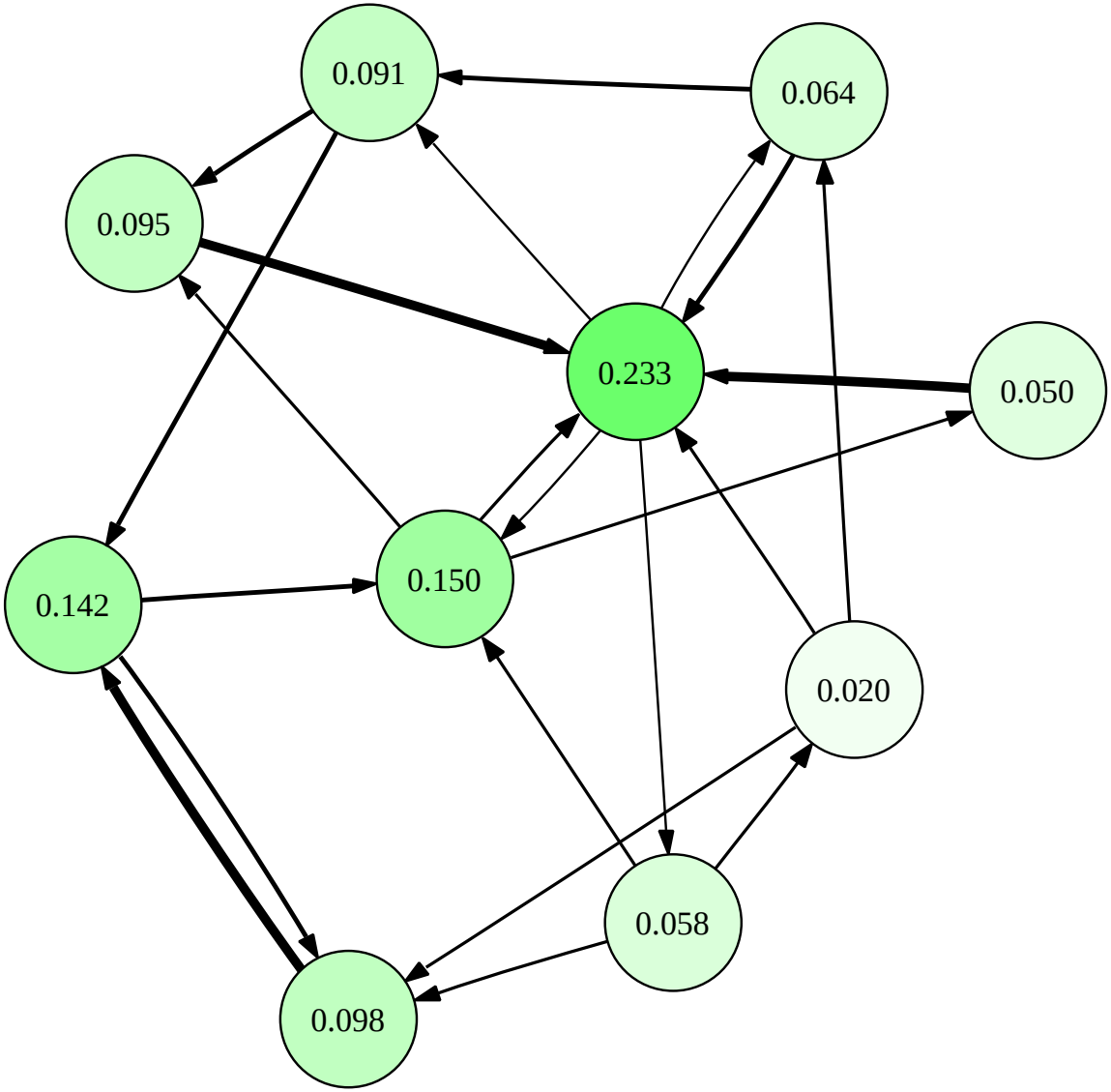
Introduction à PageRank



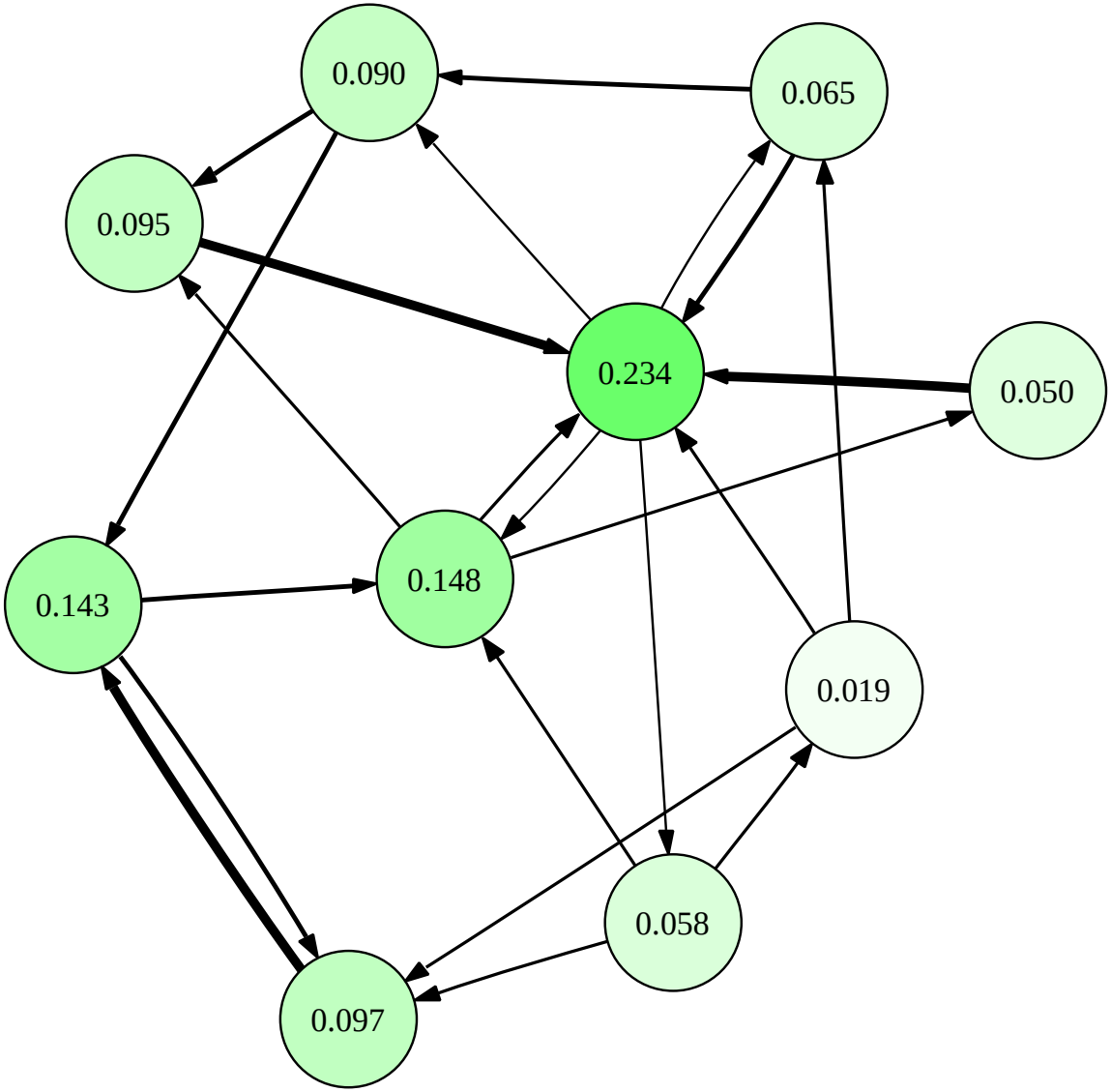
Introduction à PageRank



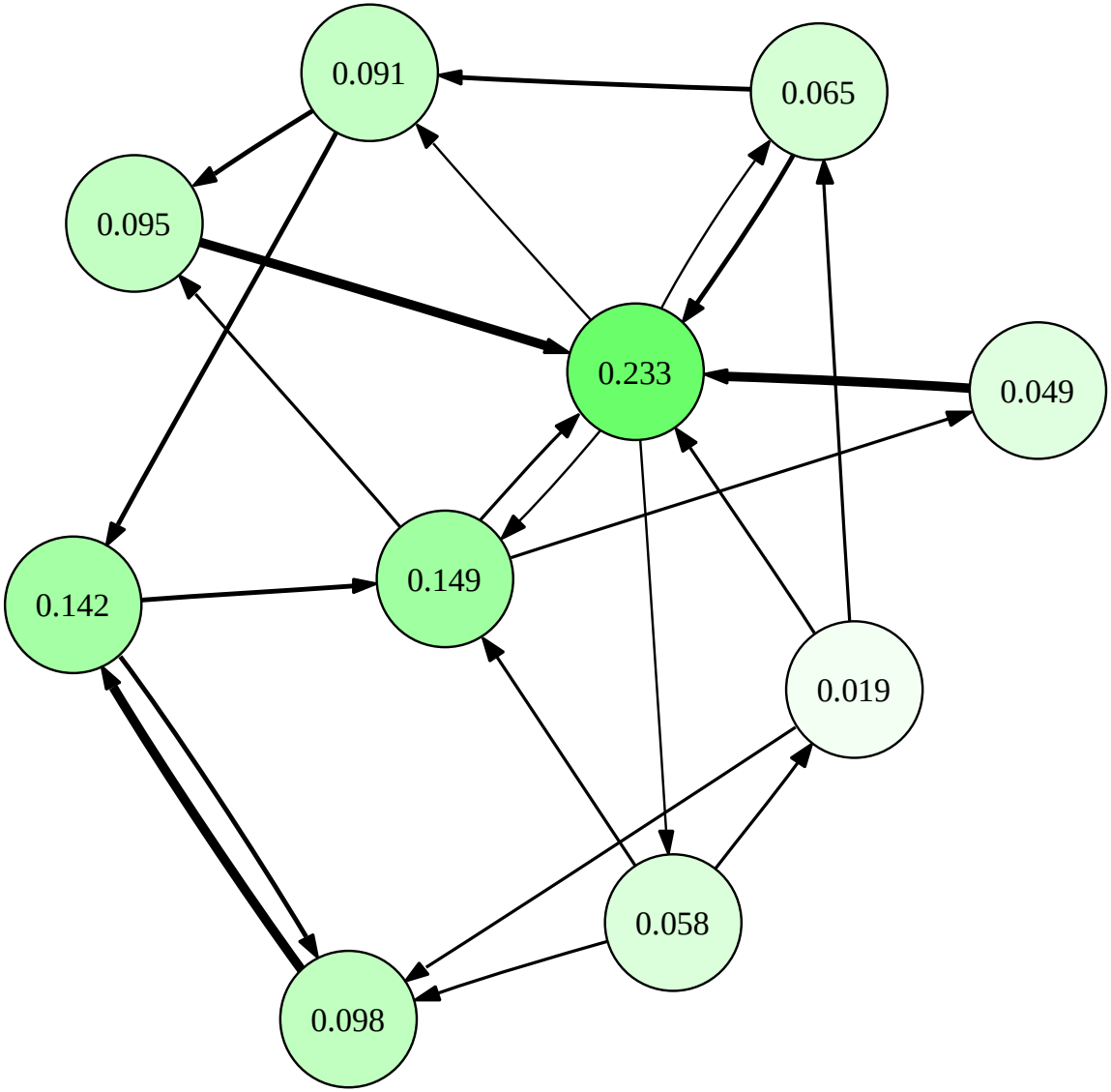
Introduction à PageRank



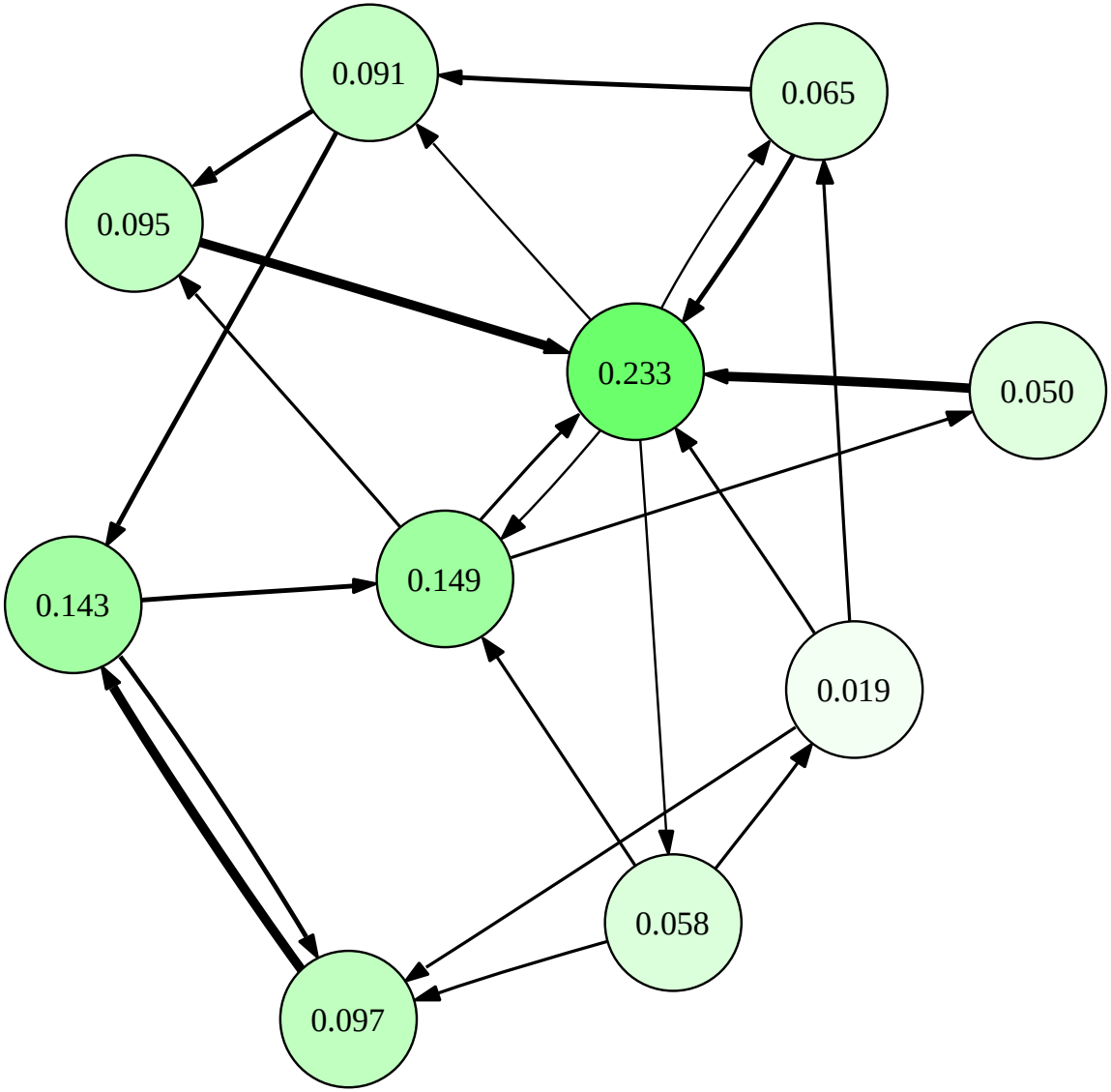
Introduction à PageRank



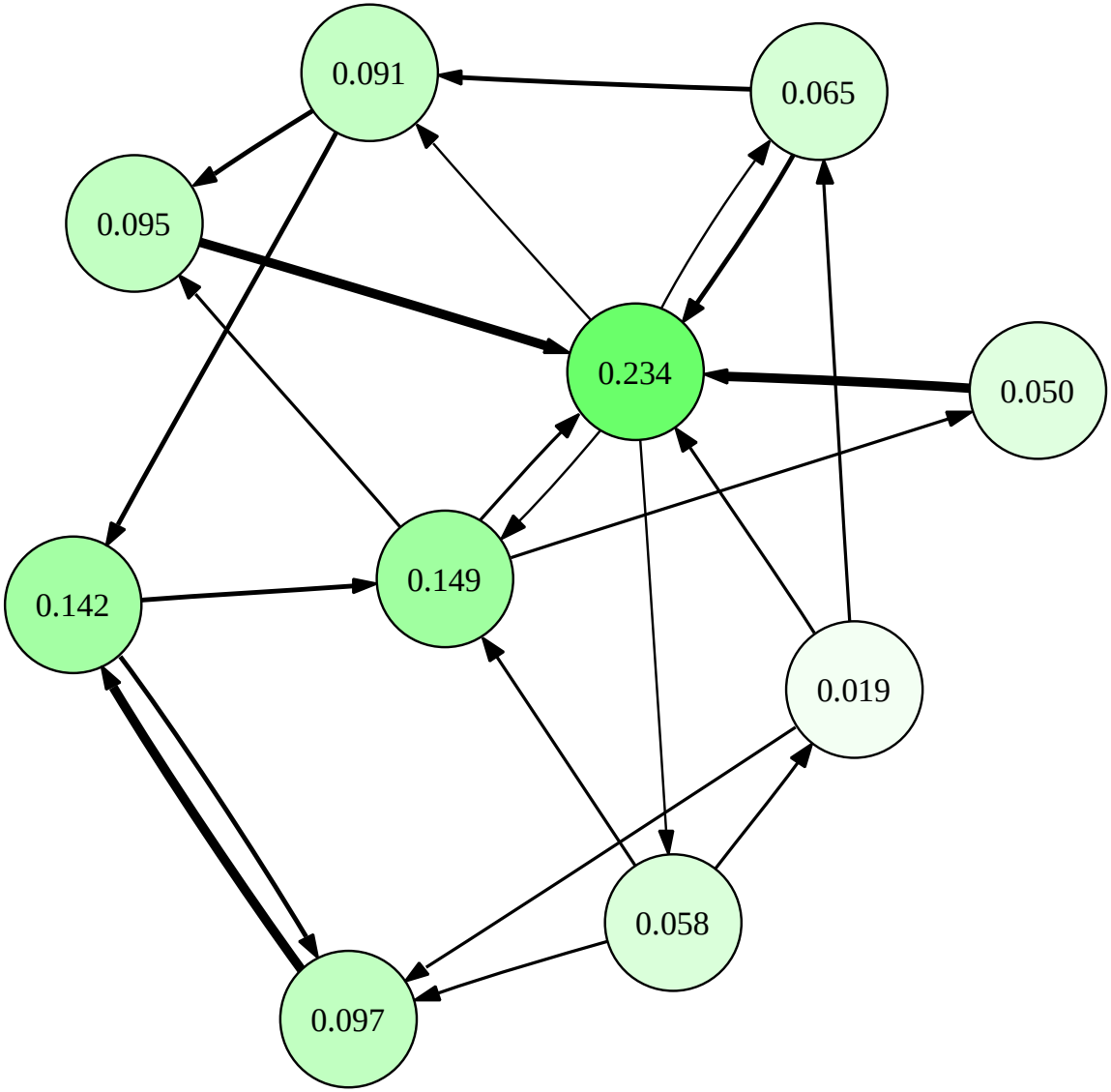
Introduction à PageRank



Introduction à PageRank

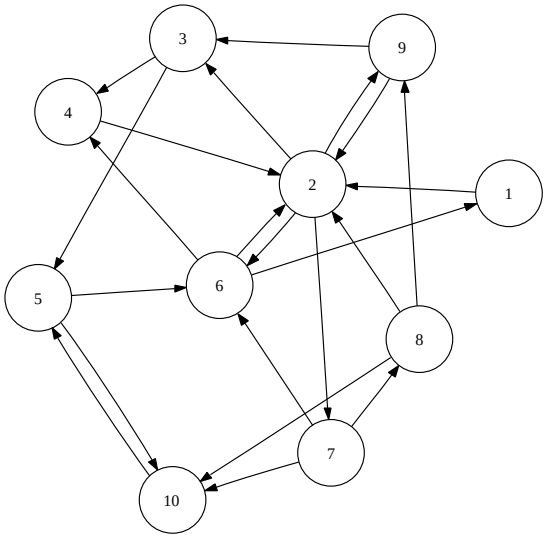


Introduction à PageRank



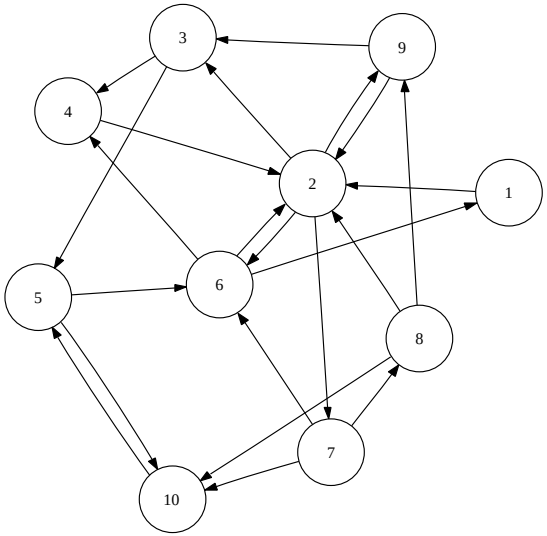
Comparaison de graphes

- HITS (Kleinberg)



Portail $\longrightarrow$ Autorité

- Généralisation

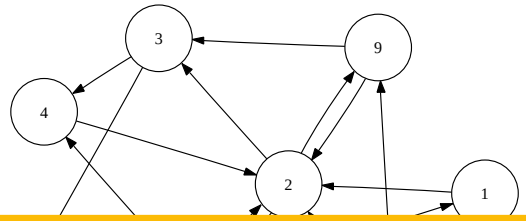


Graphe arbitraire

Collaboration avec Vincent Blondel, UCL.

Comparaison de graphes

- HITS (Kleinberg)

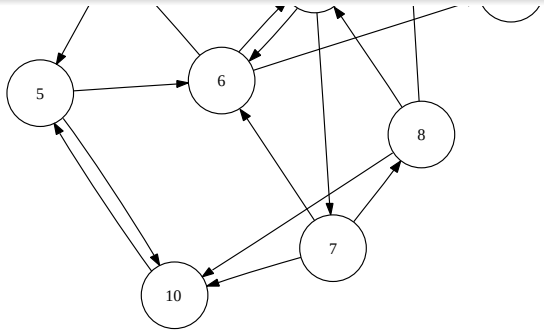


Travaux

- Modèle mathématique pour généraliser HITS à la comparaison de nœuds entre deux graphes arbitraires, avec processus itératif de calcul ;
- Application au graphe $1 \to 2 \to 3$ pour extraire des synonymes dans un dictionnaire monolingue.

rité

- C



Graphe arbitraire

Collaboration avec Vincent Blondel, UCL.

Frontières de sites Web

Site Web : ensemble des pages d'un même serveur Web ? Mais :

- De nombreux serveurs hébergent des sites différents.
- Certains sites se répartissent sur des serveurs multiples (p.ex. pages statiques/dynamiques).
- Notion hiérarchique de ce qu'est un site Web.

Collaboration avec Serge Abiteboul, INRIA Futurs.

Frontières de sites Web

Site Web : ensemble des pages d'un même serveur Web ? Mais :

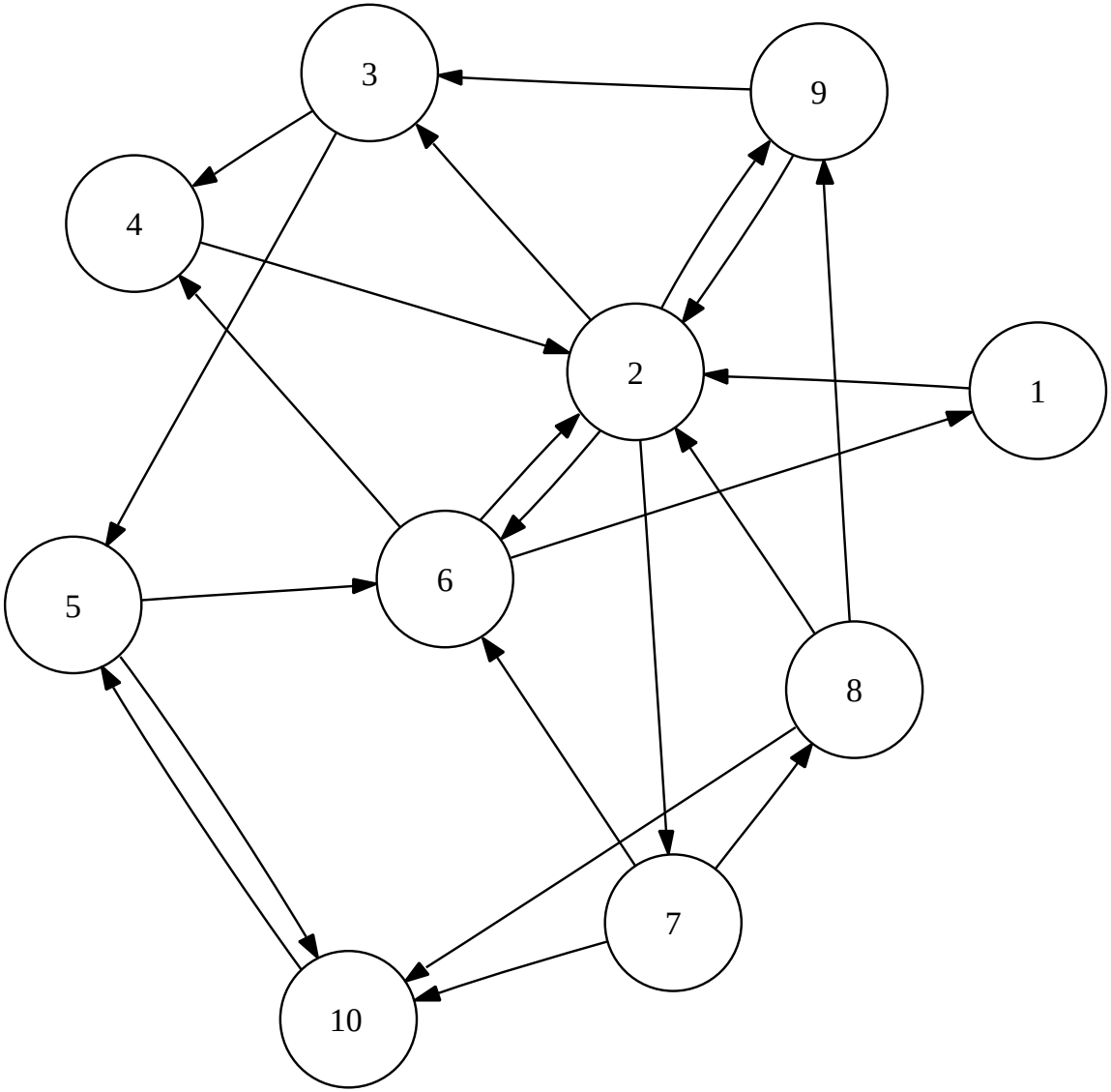
- D
- C
  (
- N

Travaux

- Méthode automatique, utilisant le flot maximal dans un réseau de transports, pour trouver les frontières d'un site Web.
- Basée essentiellement sur le graphe du site, pas sur le contenu des pages.

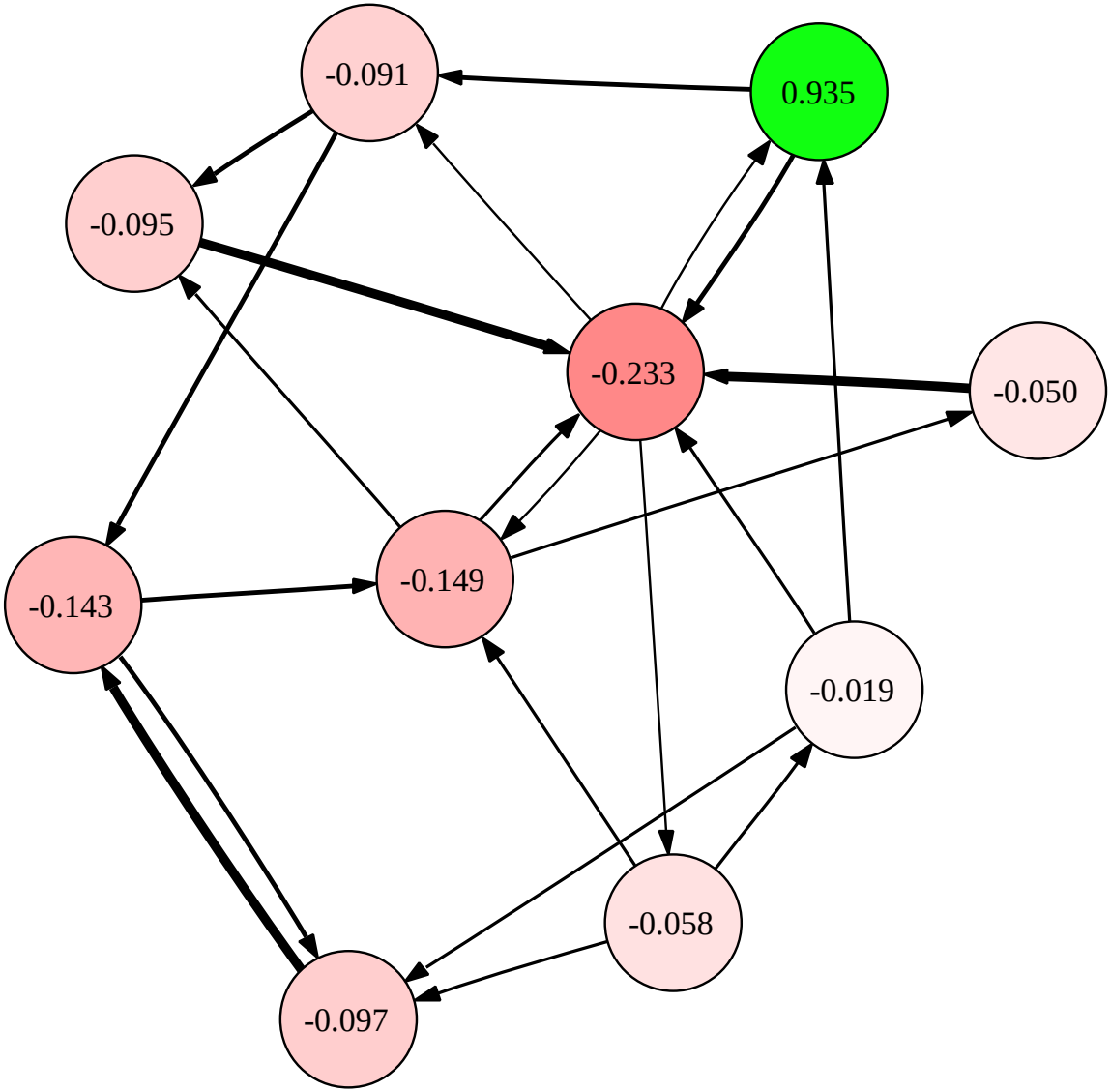
Collaboration avec Serge Abiteboul, INRIA Futurs.

Découverte de pages similaires



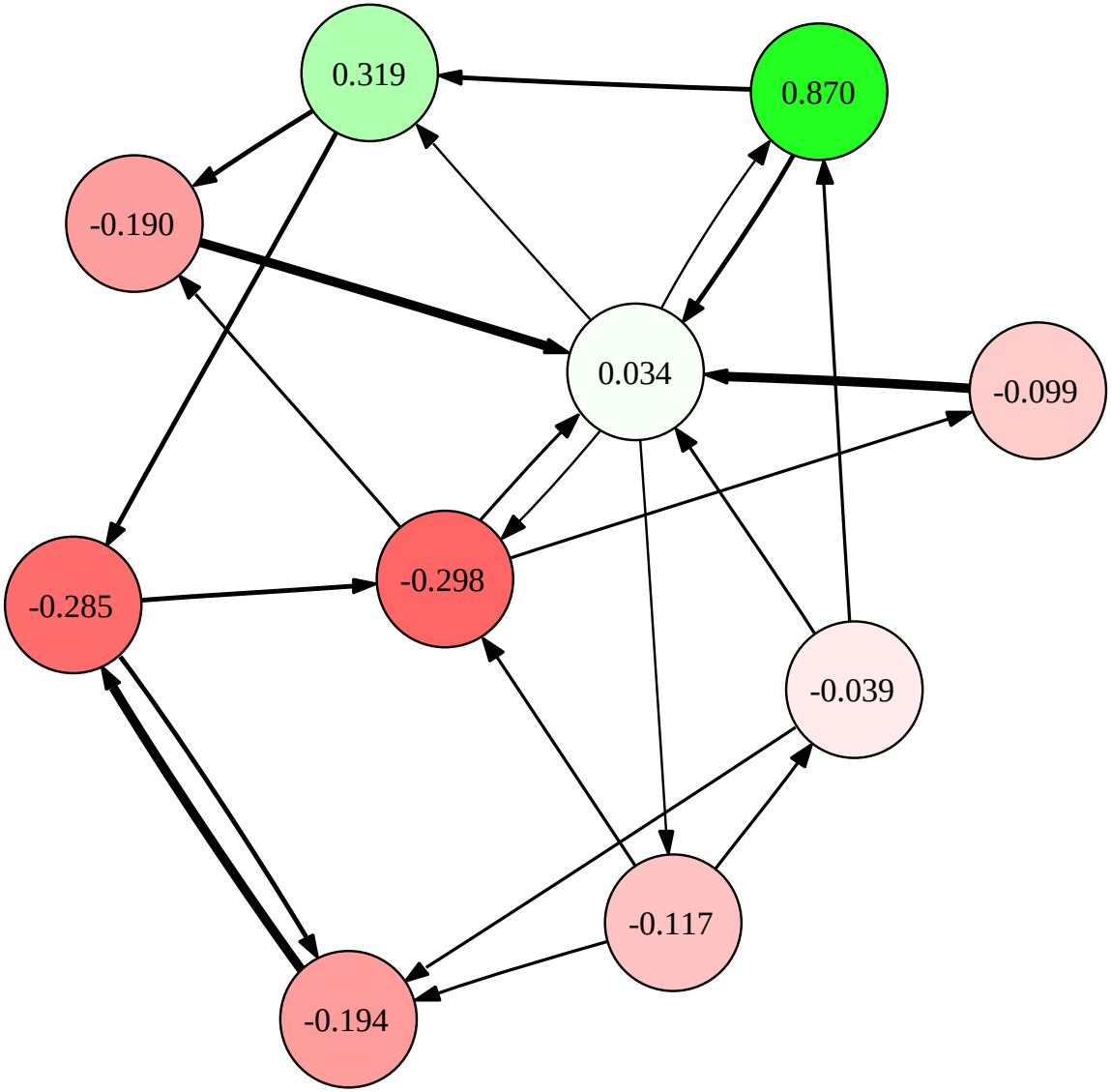
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



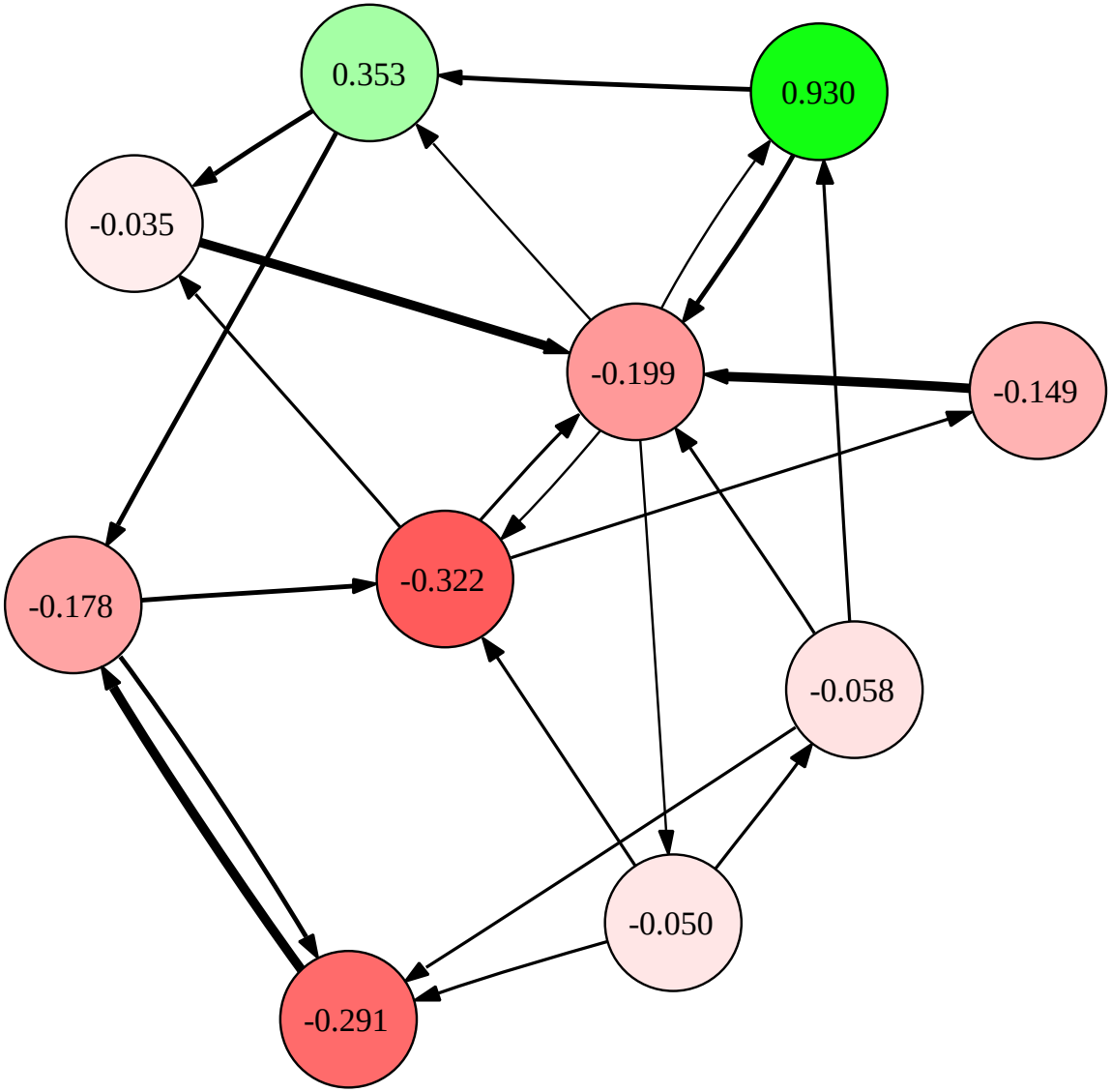
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



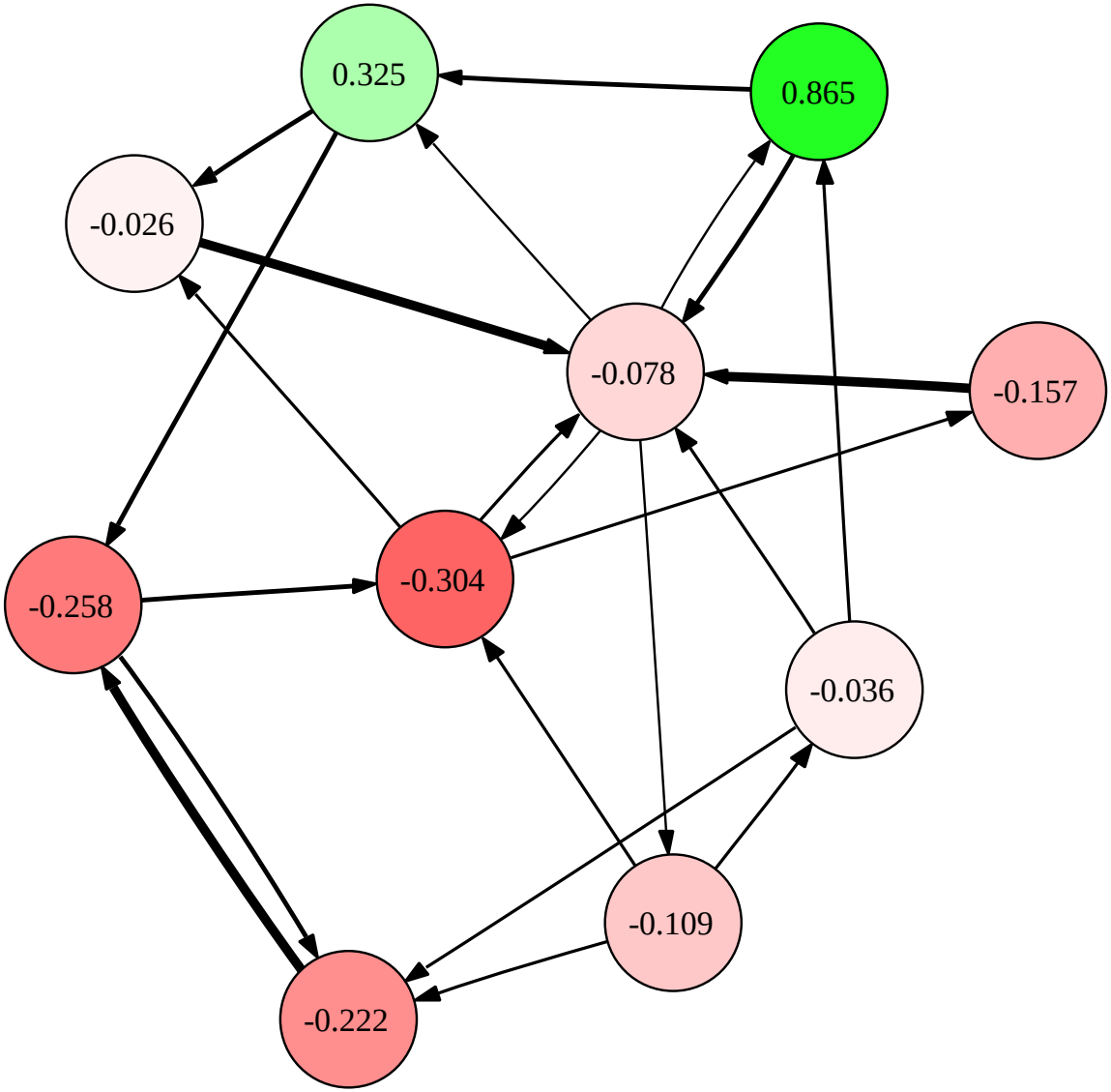
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



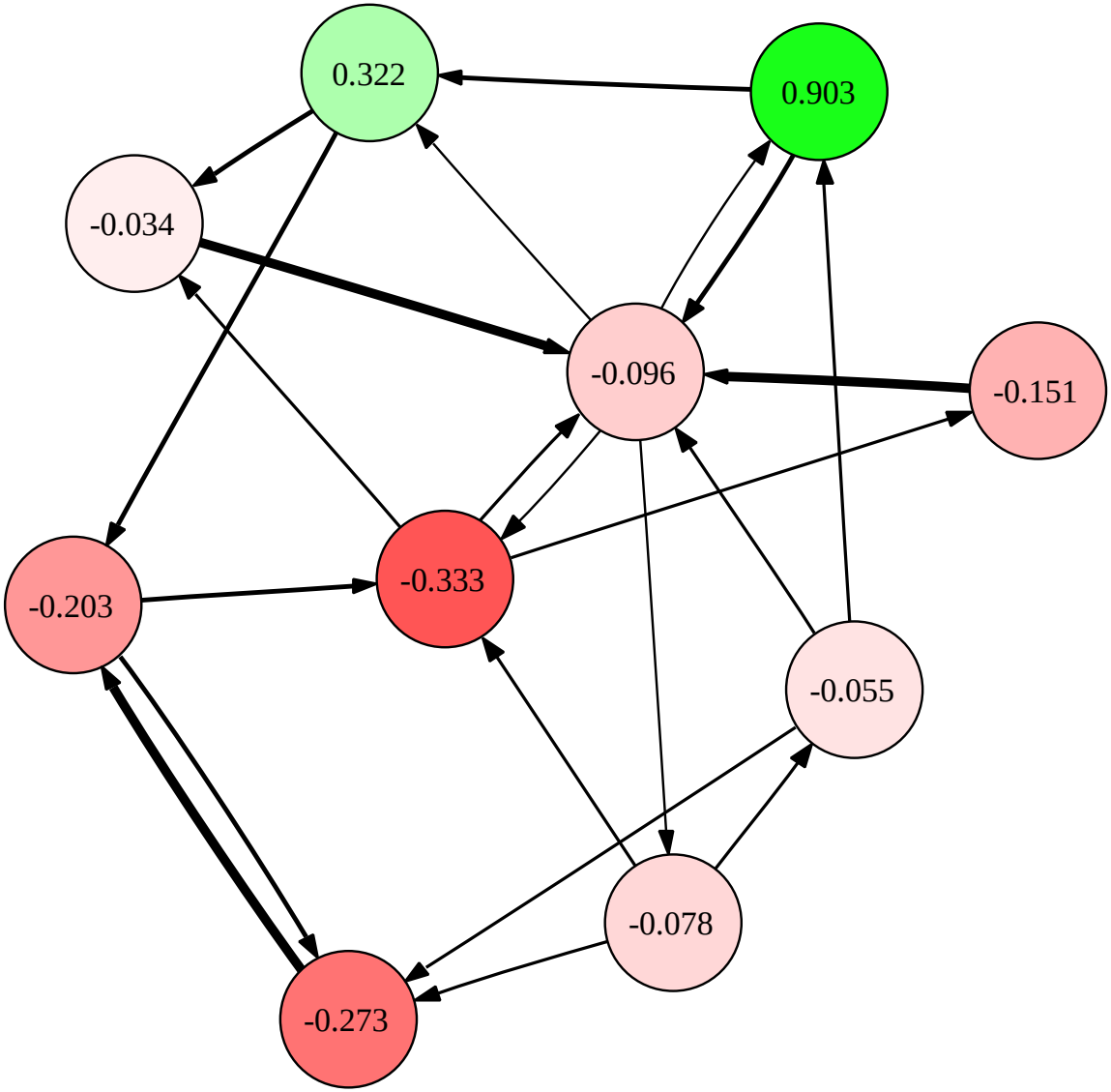
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



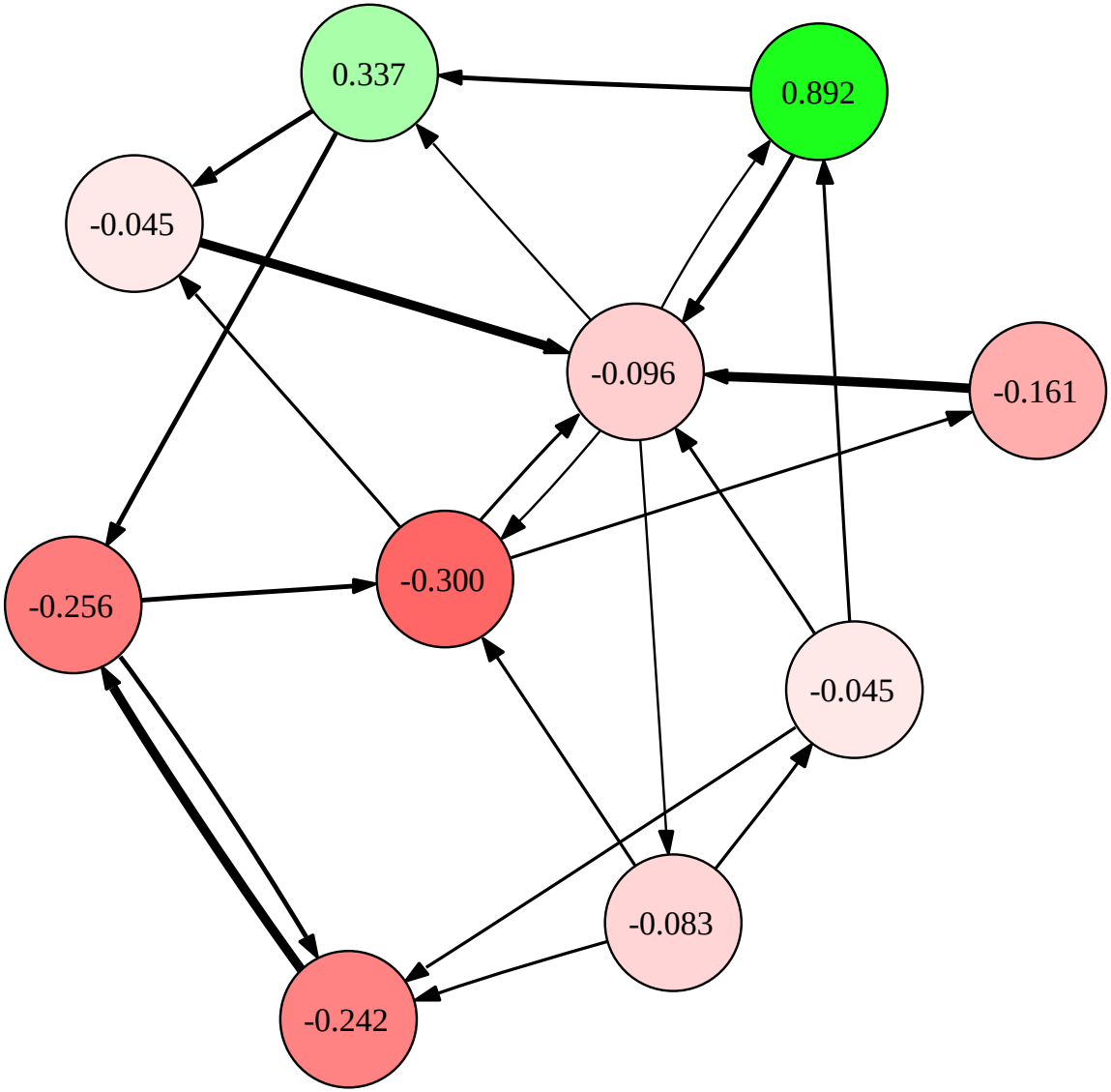
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



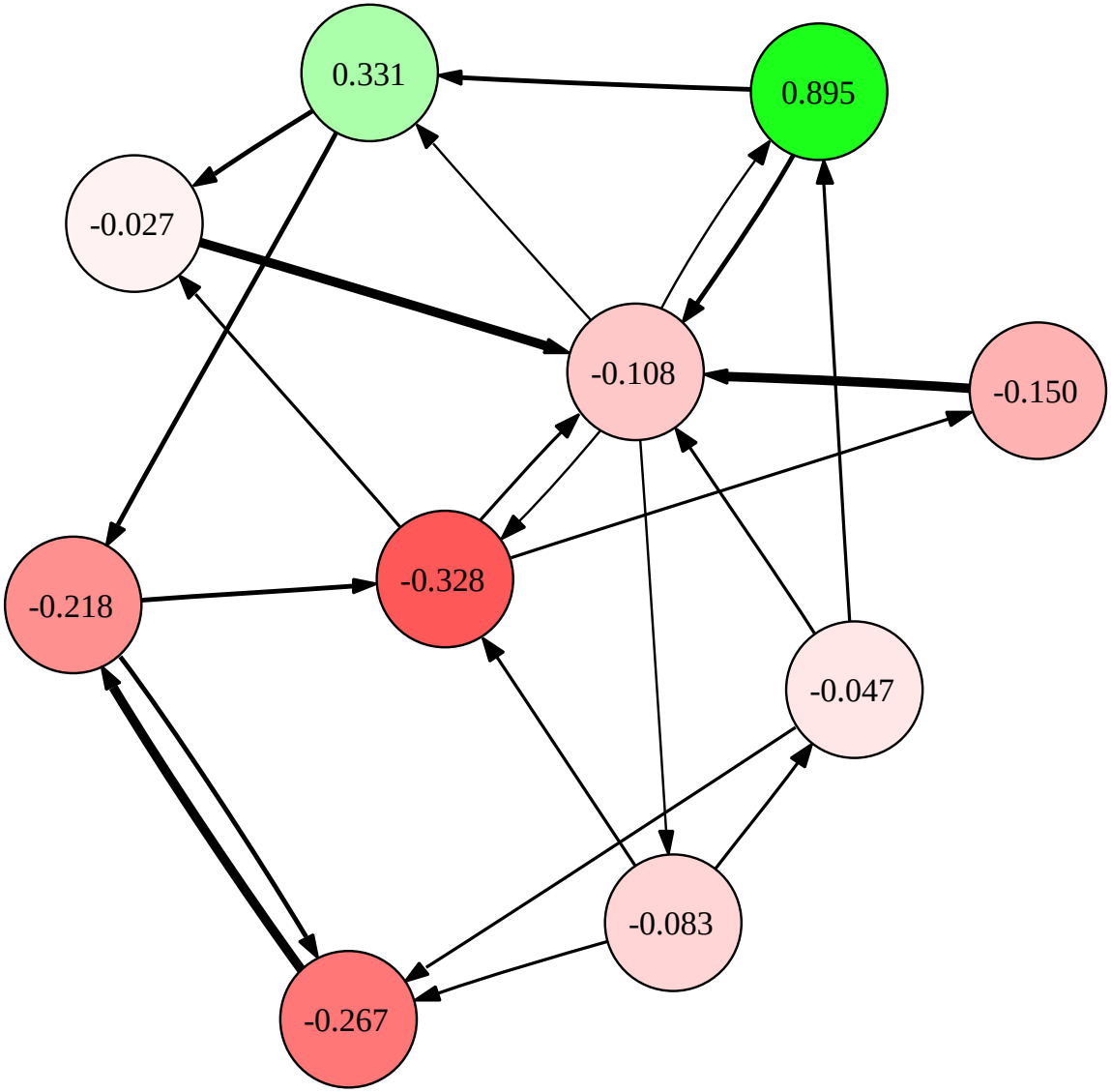
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



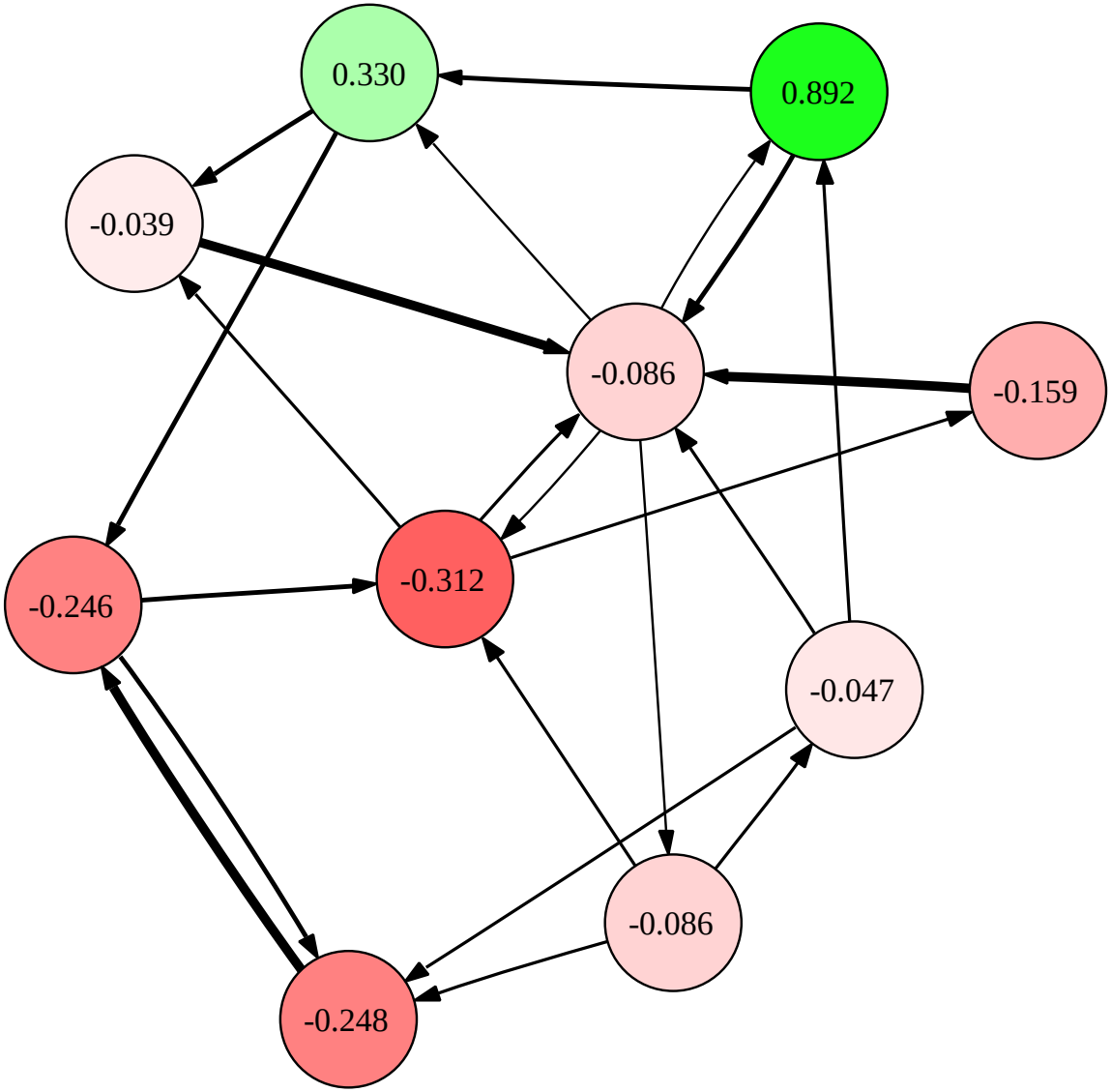
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



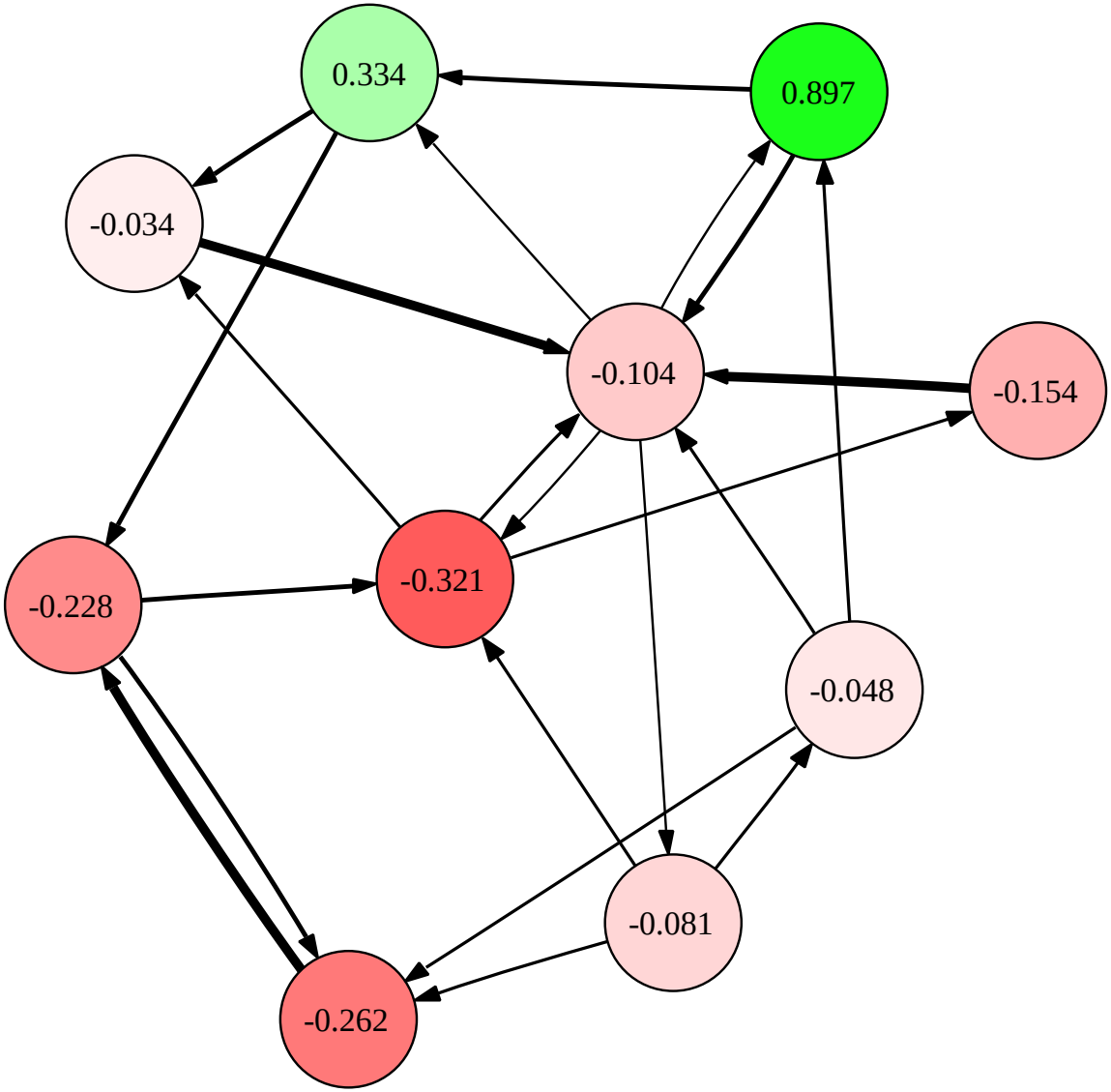
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



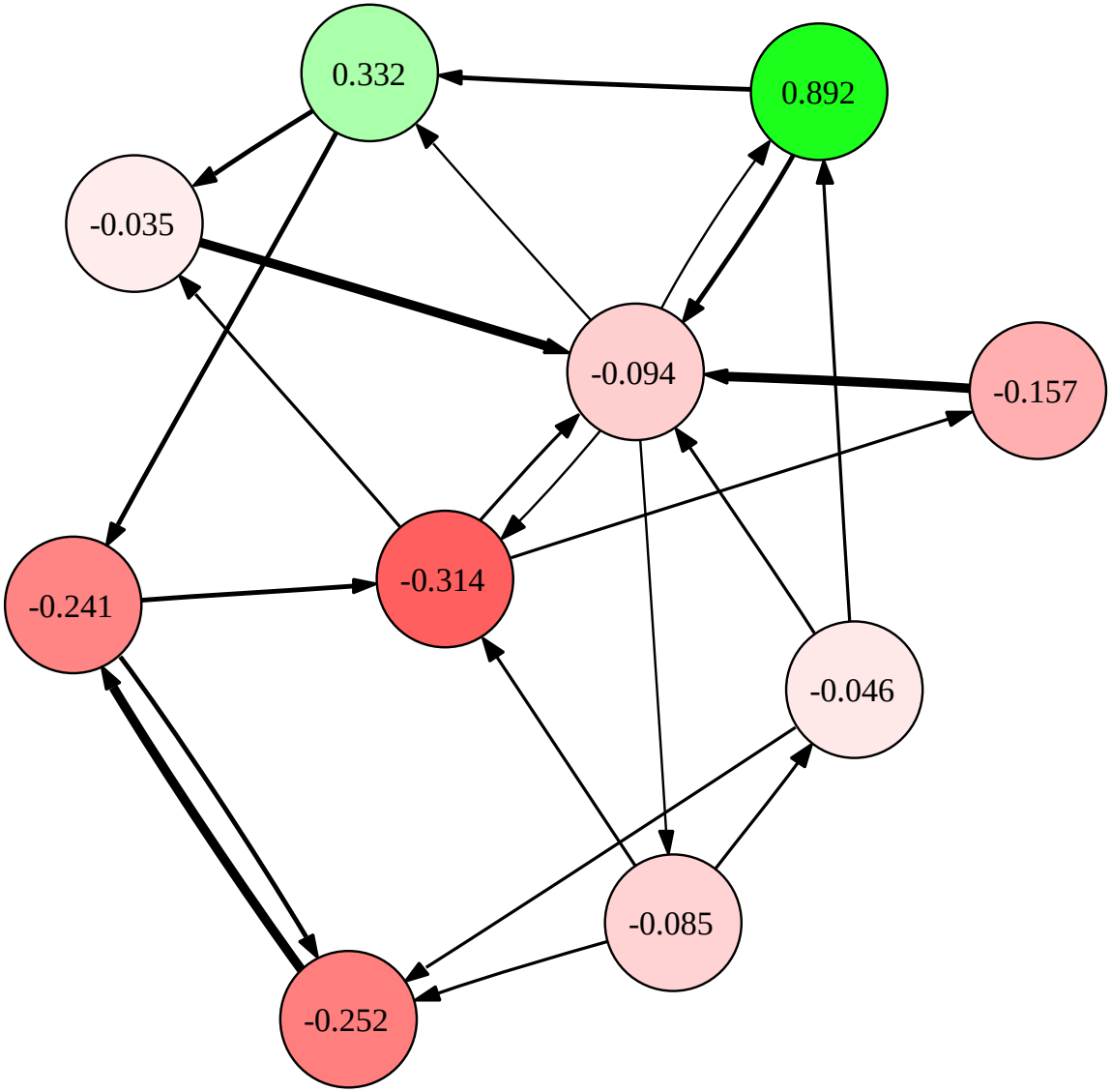
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



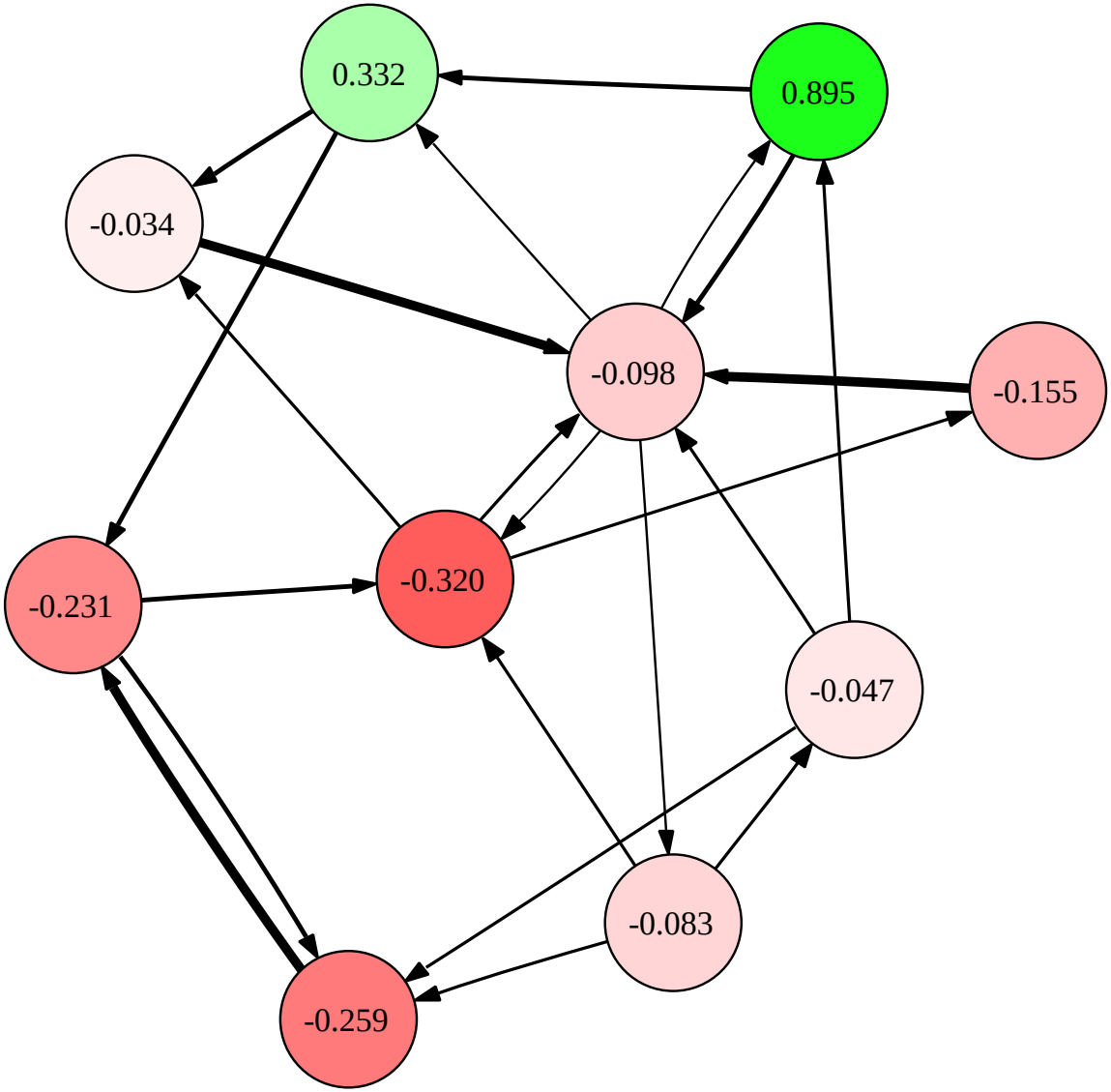
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



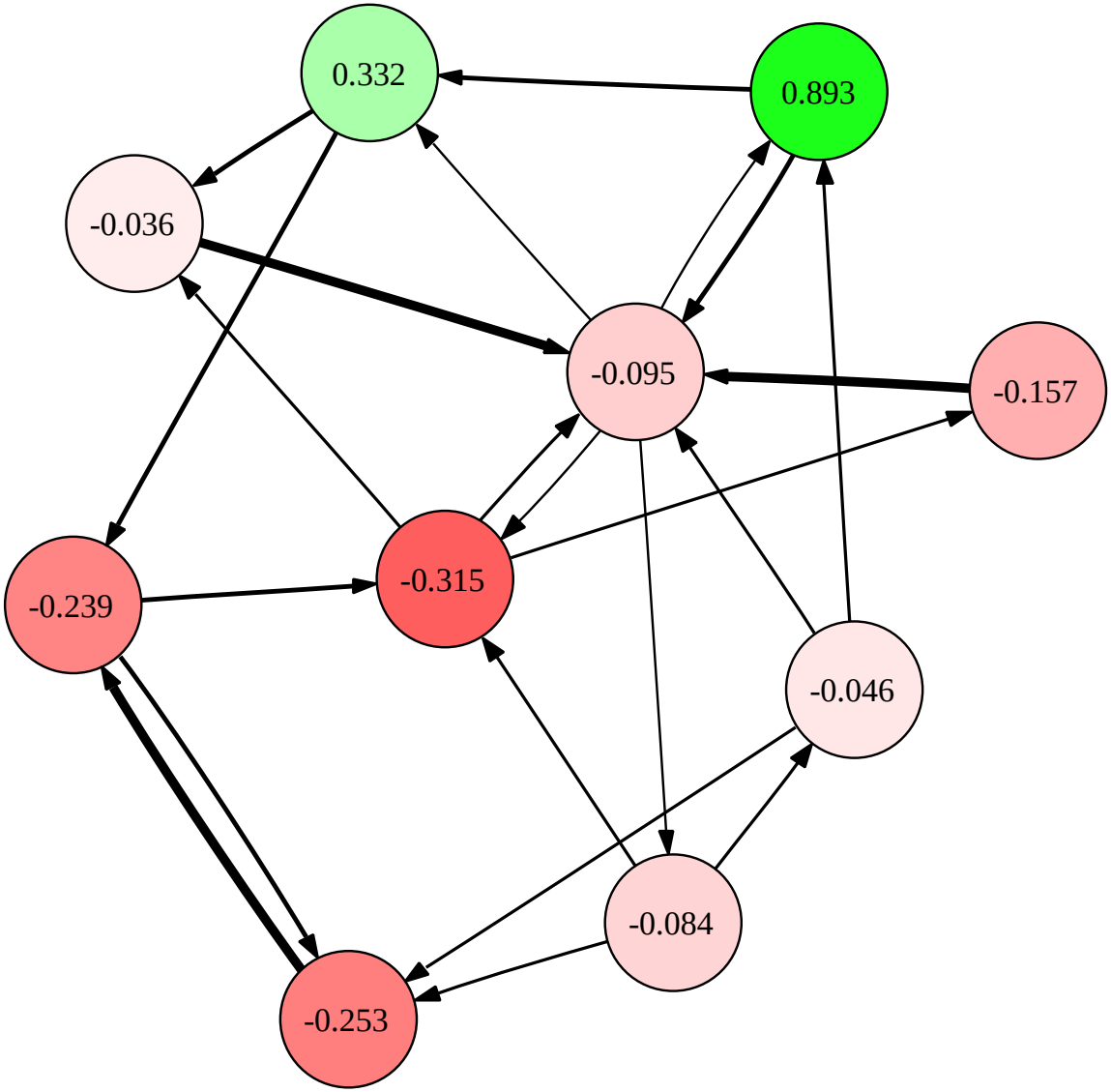
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



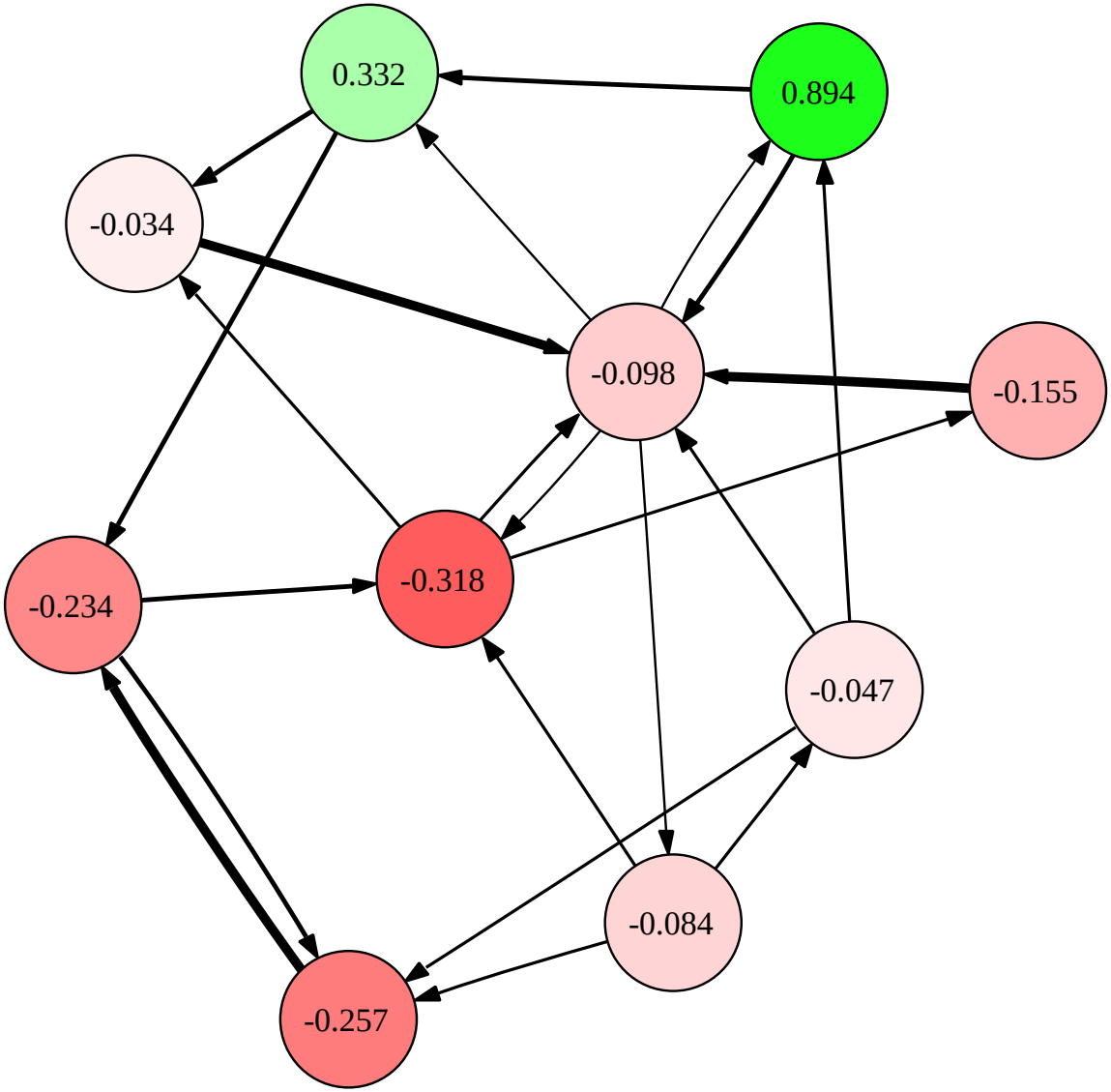
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



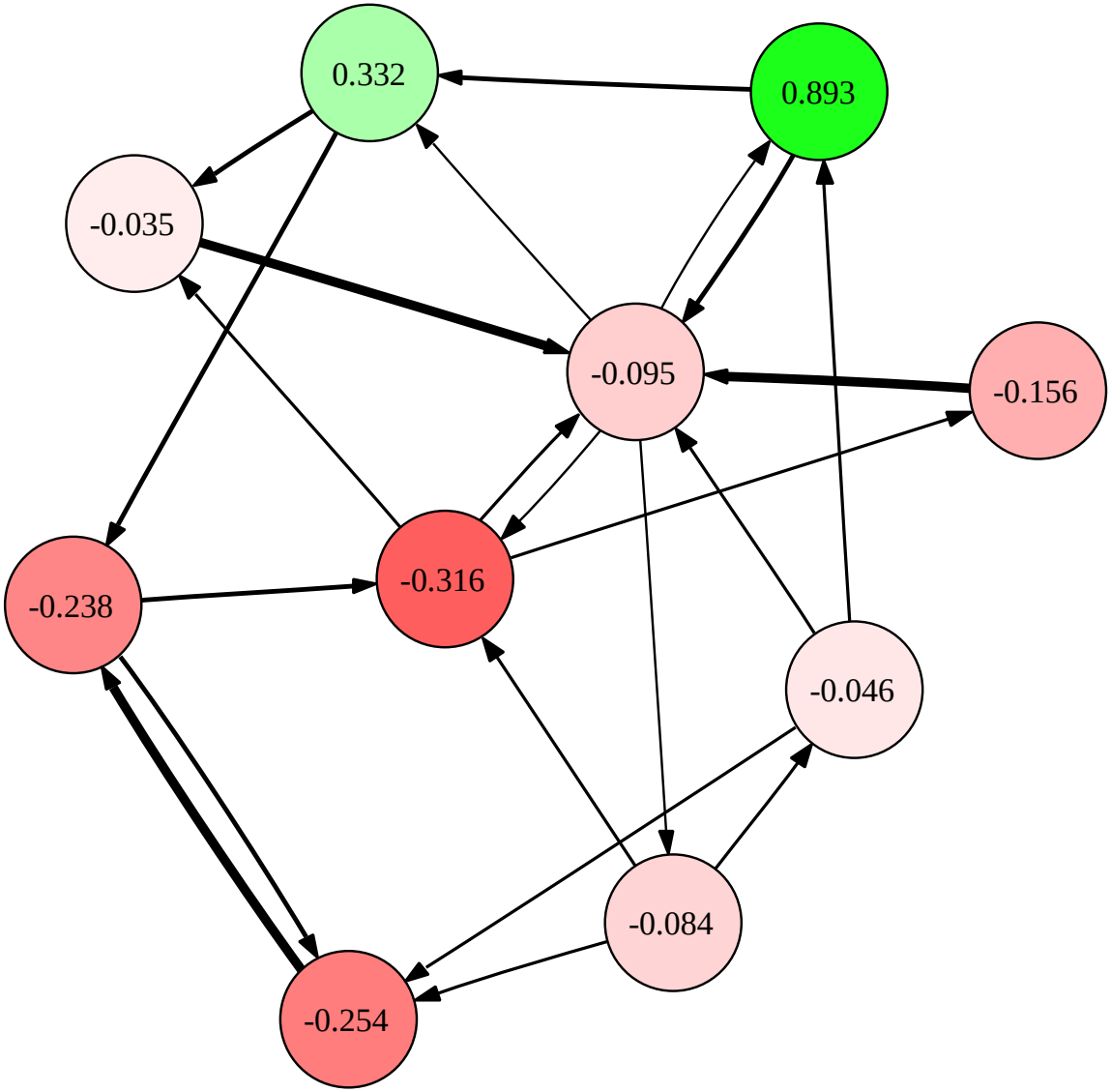
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



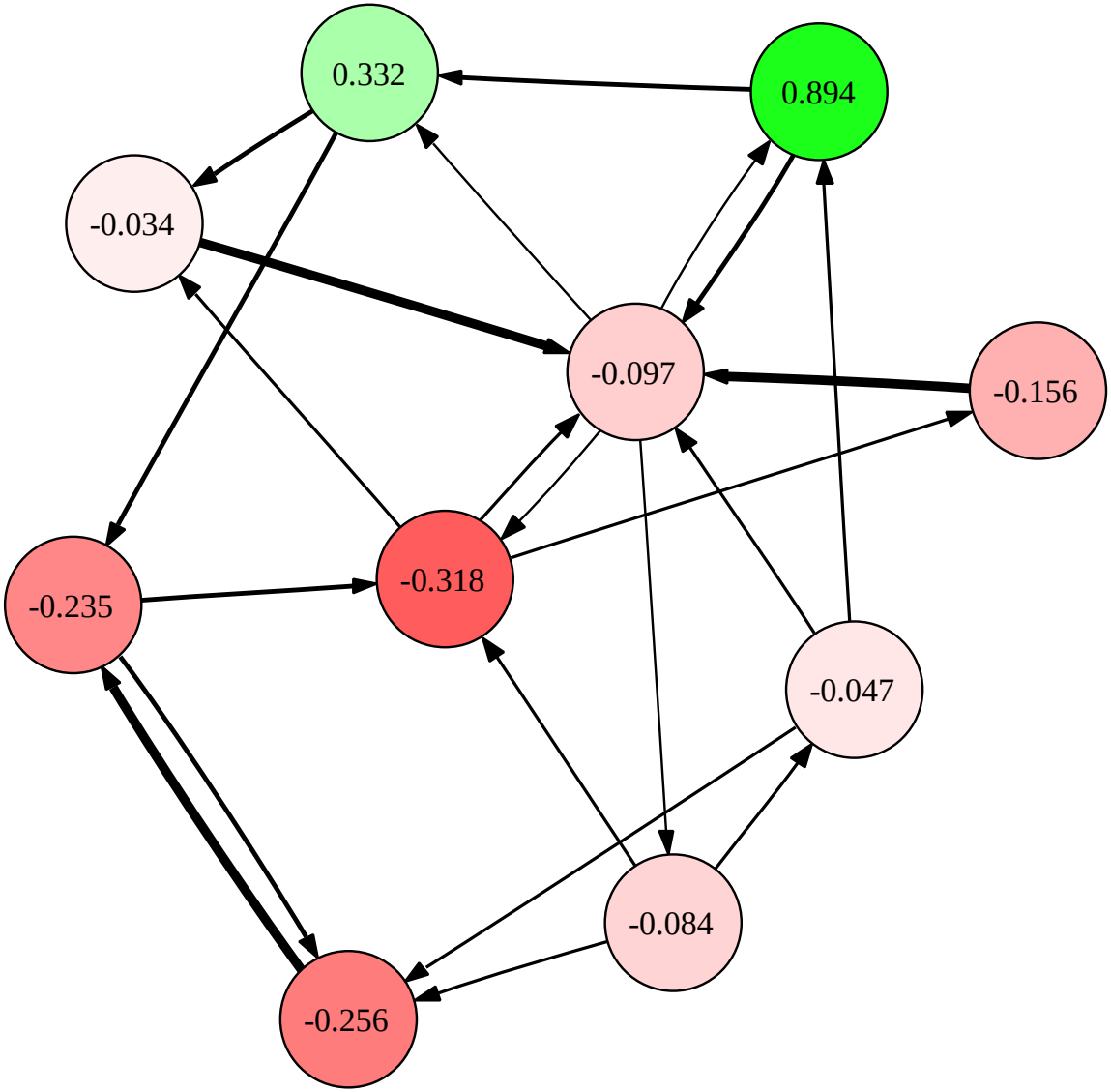
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



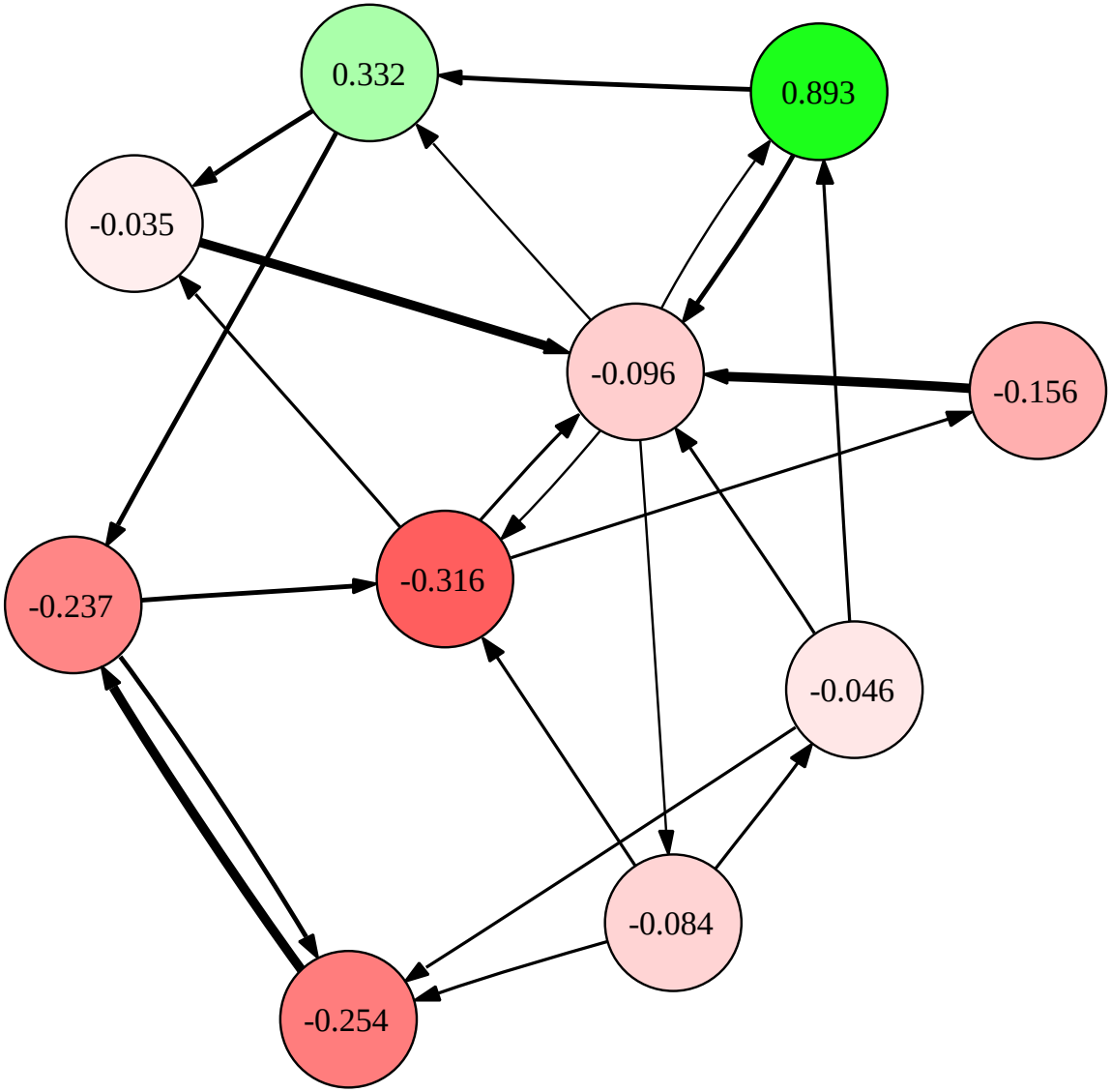
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



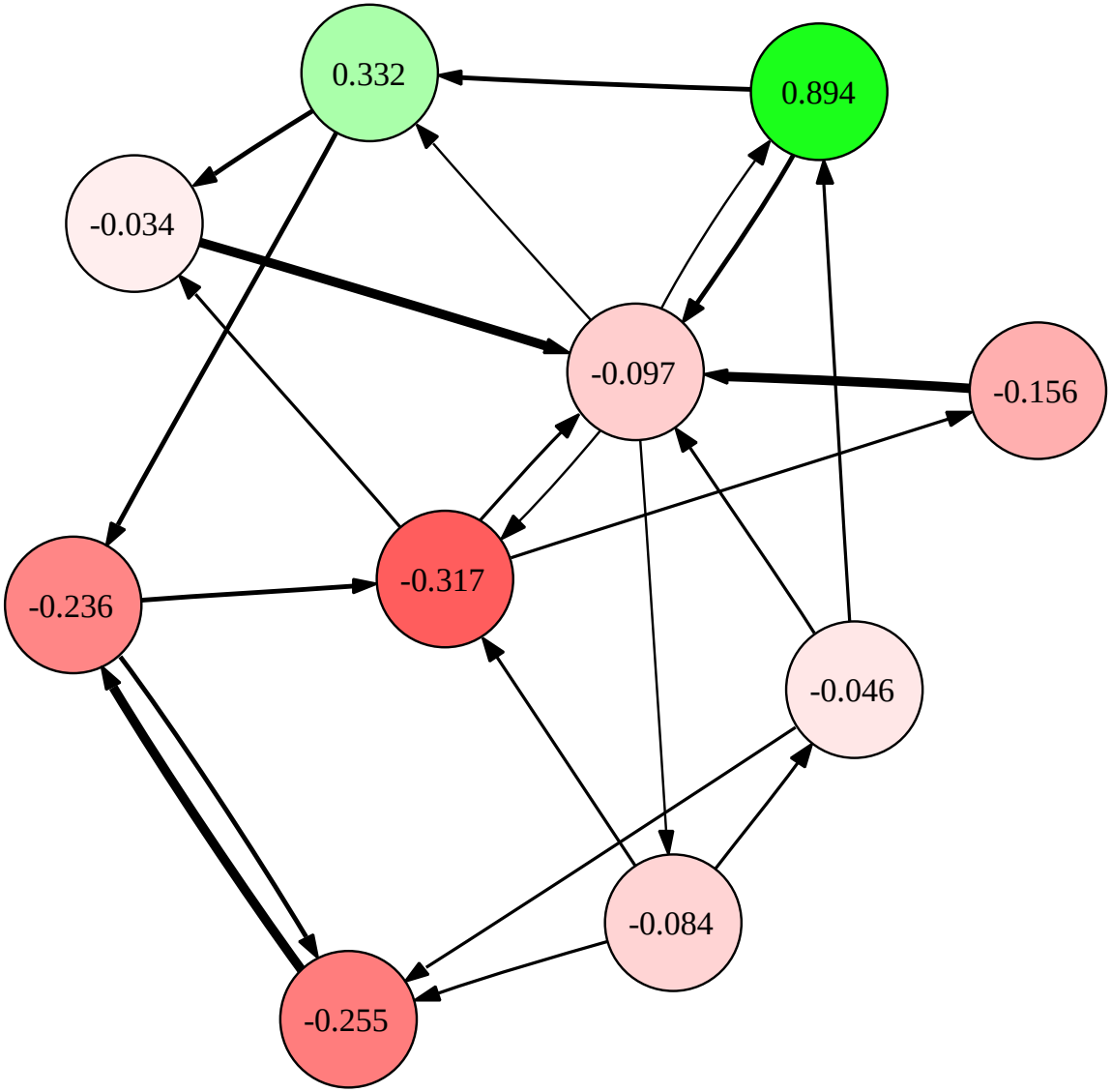
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



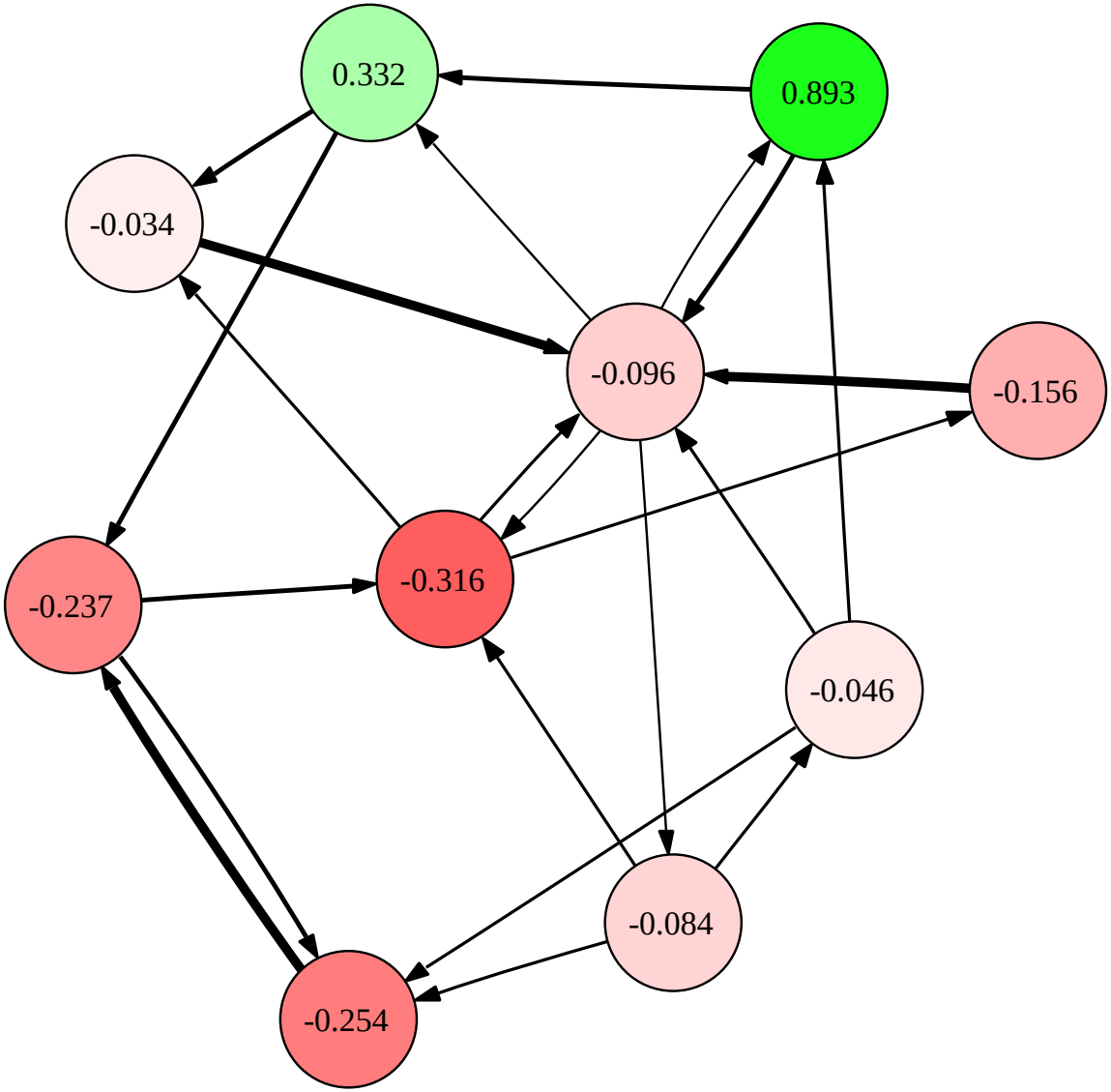
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



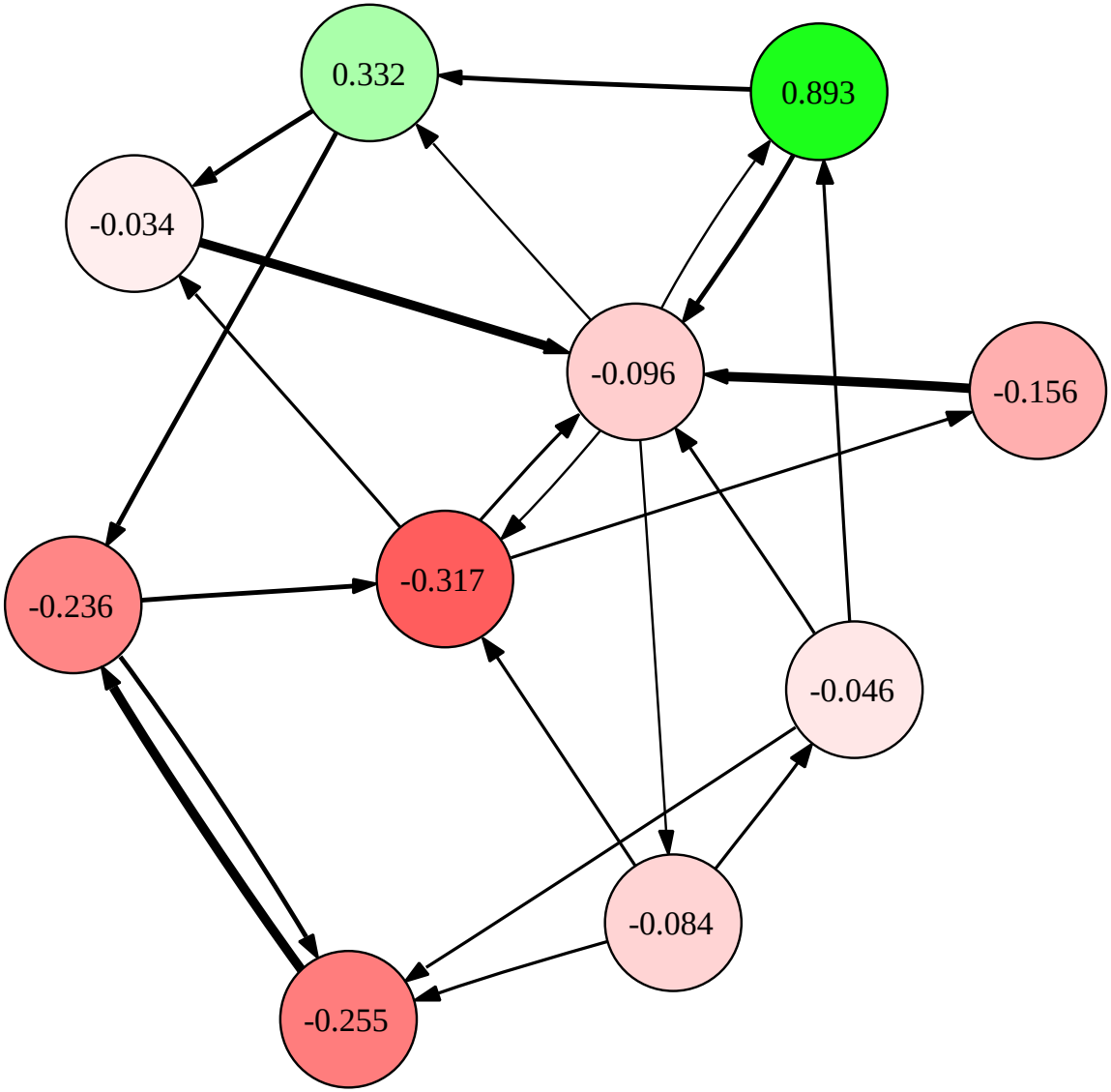
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



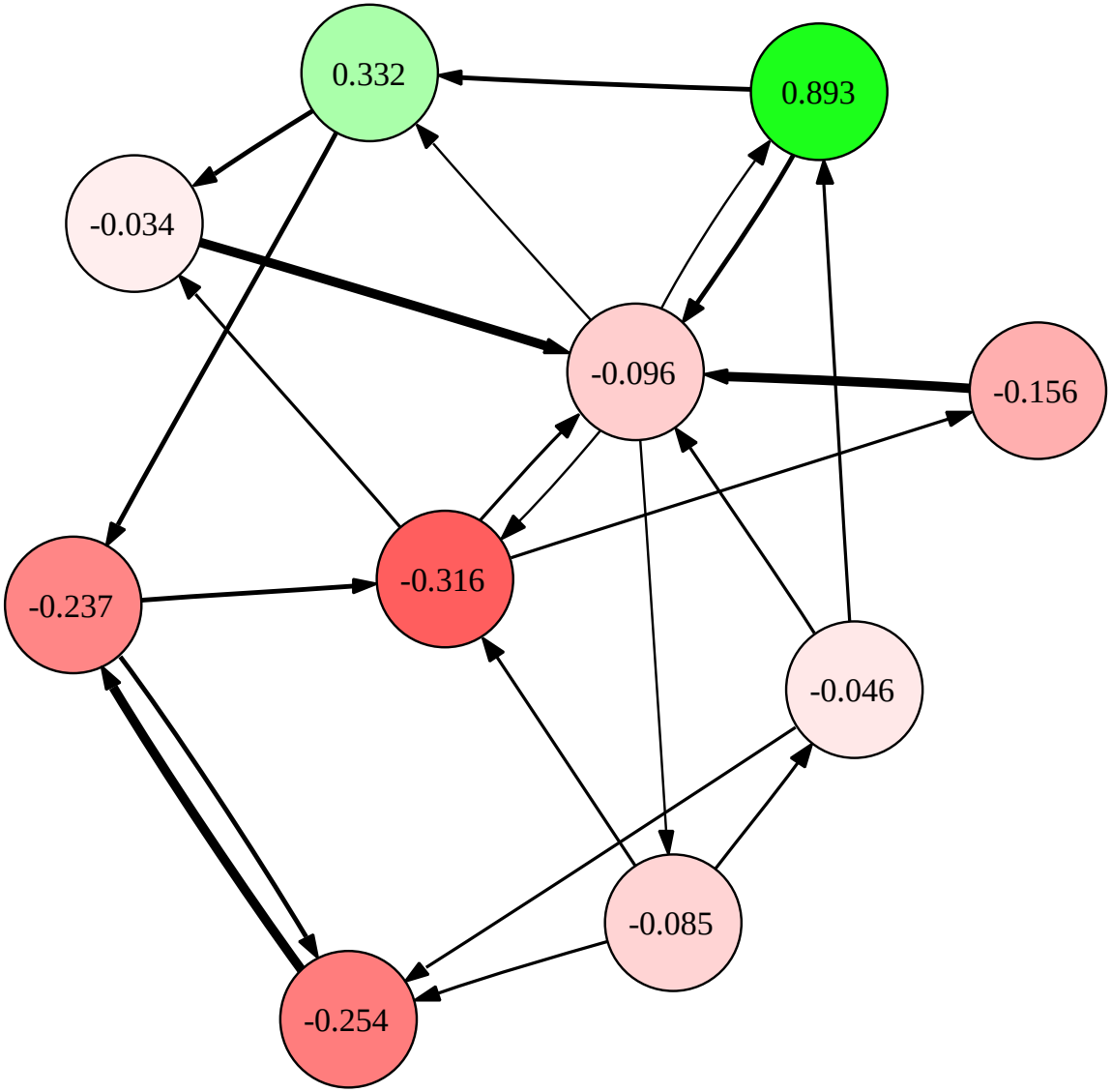
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



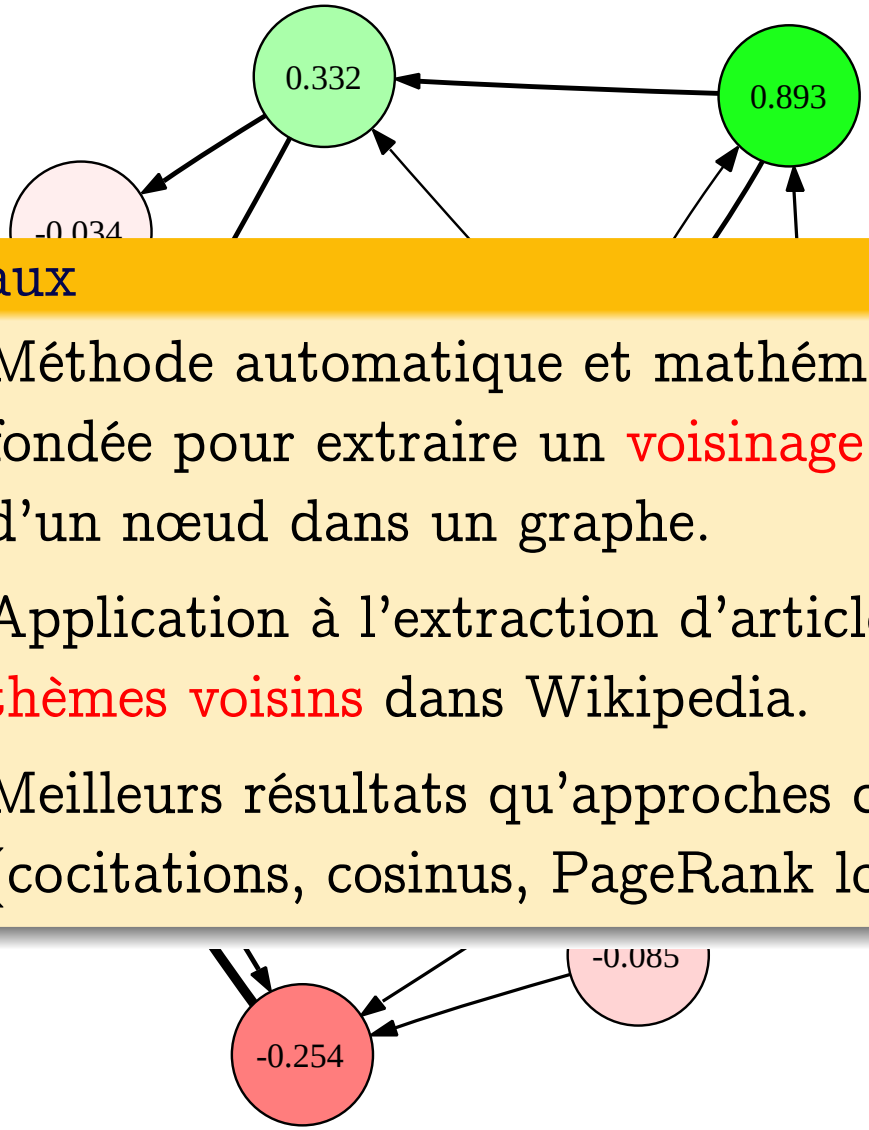
Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires



Travaux

- Méthode automatique et mathématiquement fondée pour extraire un voisinage conceptuel d'un nœud dans un graphe.
- Application à l'extraction d'articles sur des thèmes voisins dans Wikipedia.
- Meilleurs résultats qu'approches classiques (cocitations, cosinus, PageRank local...).

Collaboration avec Yann Ollivier, ENS Lyon.

Découverte de pages similaires

Example of related articles for Pierre de Fermat :

Green	SymGreen	PageRankOfLinks	Cosine	Cocitations
1. Pierre de Fermat	1. Pierre de Fermat	1. France	1. Pierre de Fermat	1. Pierre de Fermat
2. Toulouse	2. Mathematics	2. 17th century	2. ENSICA	2. Leonhard Euler
3. Fermat's Last Theorem	3. Probability theory	3. March 4	3. Fermat's theorem	3. Mathematics
4. Diophantine equation	4. Fermat's Last Theorem	4. January 12	4. International School of Toulouse	4. René Descartes
5. Fermat's little theorem	5. Number theory	5. August 17	5. École Nationale Supérieure d'Électronique, d'Électrotechnique, d'Informatique, d'Hydraulique, et de Télécommunications	5. Mathematician
6. Fermat number	6. Toulouse	6. Calculus	6. Languedoc	6. Gottfried Leibniz
7. Grandes écoles	7. Diophantine equation	7. Lawyer	7. Hélène Pince	7. Calculus
8. Blaise Pascal	8. Blaise Pascal	8. 1660	...	8. Isaac Newton
9. France	9. Fermat's little theorem	9. Number theory		9. Blaise Pascal
10. Pseudo-prime	10. Calculus	10. René Descartes		10. Carl Friedrich Gauss

Collaboration avec Yann Ollivier, ENS Lyon.

Prédiction de PageRank

- Le score (PageRank) d'une page Web évolue au cours du temps.
- Coûteux de parcourir tout le Web afin de mettre à jour ces scores.
- Possibilité de prévoir ces évolutions, en identifiant des tendances récurrentes ?

Collaboration avec Michalis Vazirgiannis, Athens Univ. of Economics & Business.

Prédiction de PageRank

- Le score (PageRank) d'une page Web évolue au cours du t
- C r ces s
- F t

Travaux

- Prédiction de PageRank avec un modèle de Markov caché ;
- Mesures de similarités adaptées pour comparer les scores effectifs et prédits.

Collaboration avec Michalis Vazirgiannis, Athens Univ. of Economics & Business.

Le Web caché

Définition (Web caché, Web profond, Web invisible)

La partie du Web qui n'est pas accessible via des hyperliens : formulaires HTML, services Web. Donc cachée aux moteurs de recherche classiques.

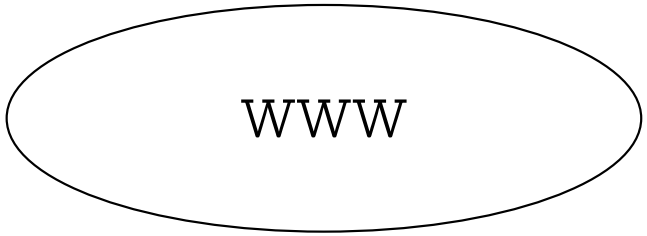
Taille (estimation) : 500 fois plus grand que le Web de surface.

But

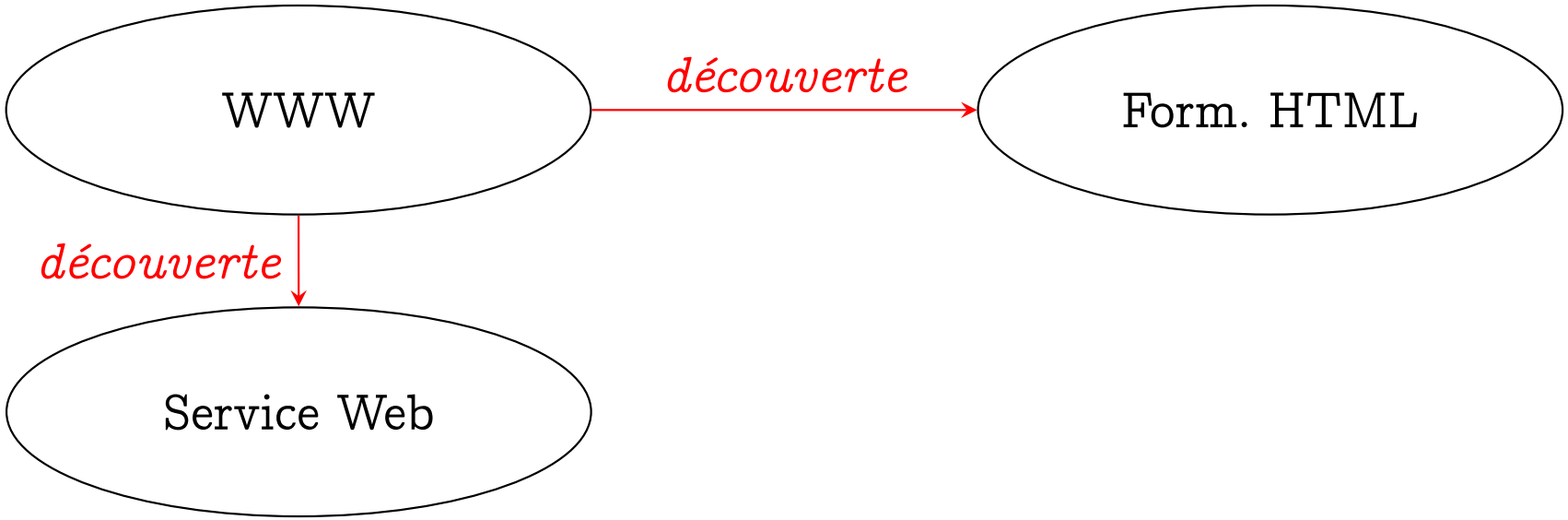
- Indexation en compréhension (c.à.d., pas en extension) du Web caché.
- De manière complètement automatique et non supervisée !

Problème très large ⇒ restriction à un domaine d'intérêt particulier.

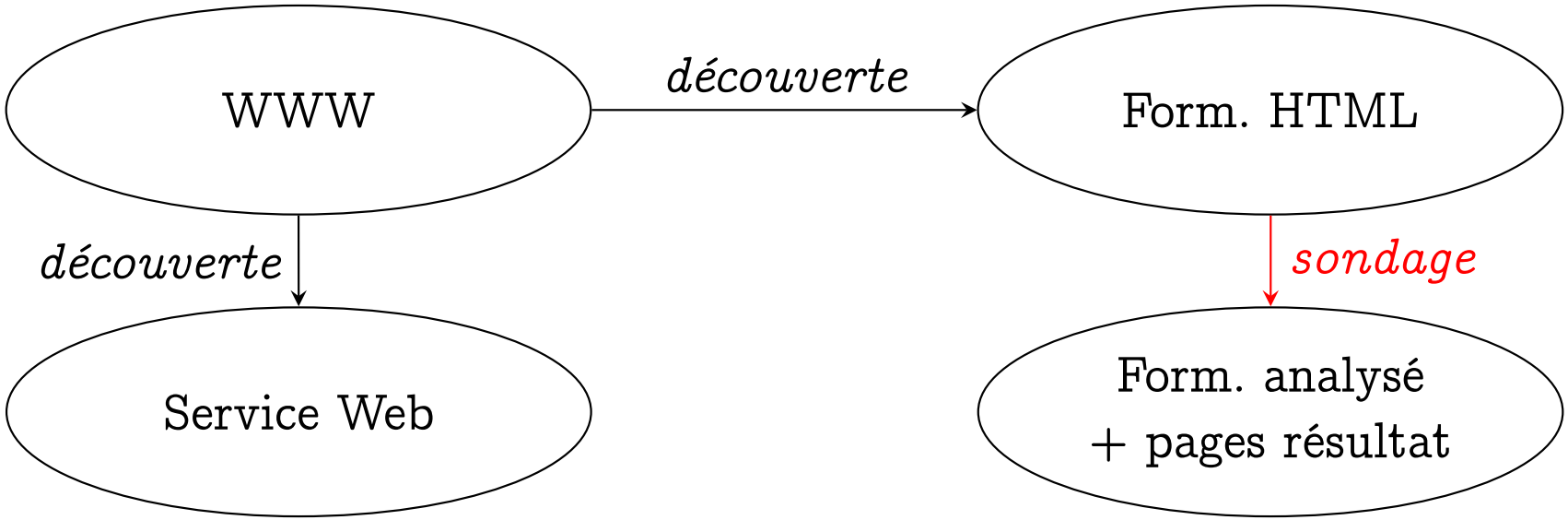
Processus automatique pour comprendre le Web caché



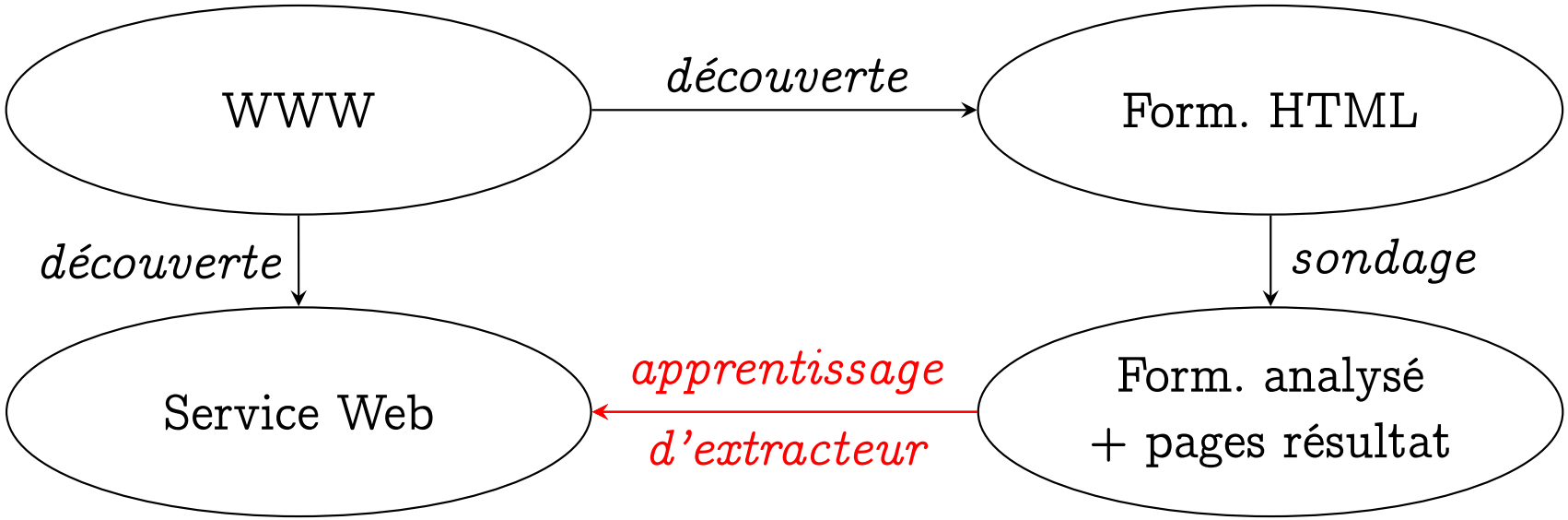
Processus automatique pour comprendre le Web caché



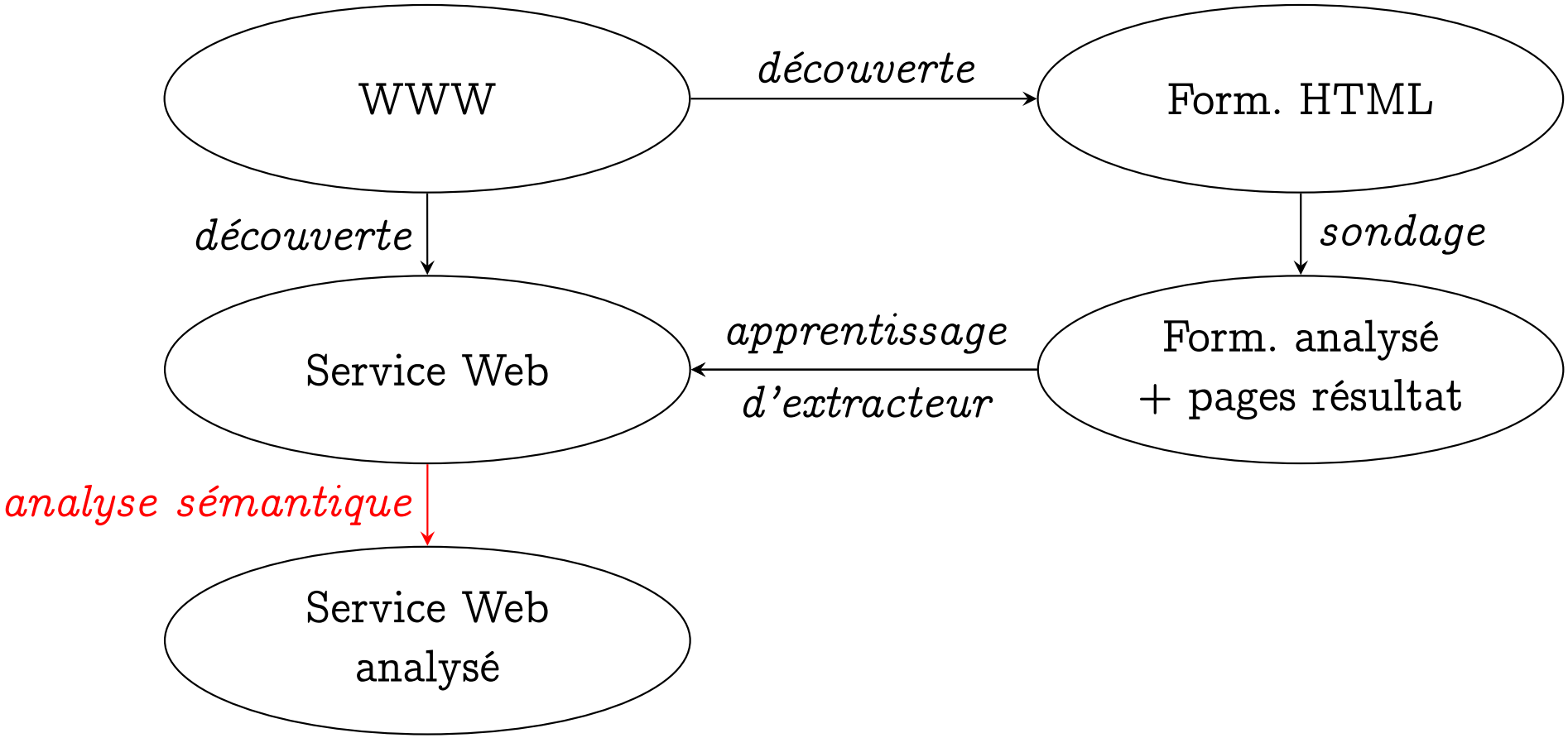
Processus automatique pour comprendre le Web caché



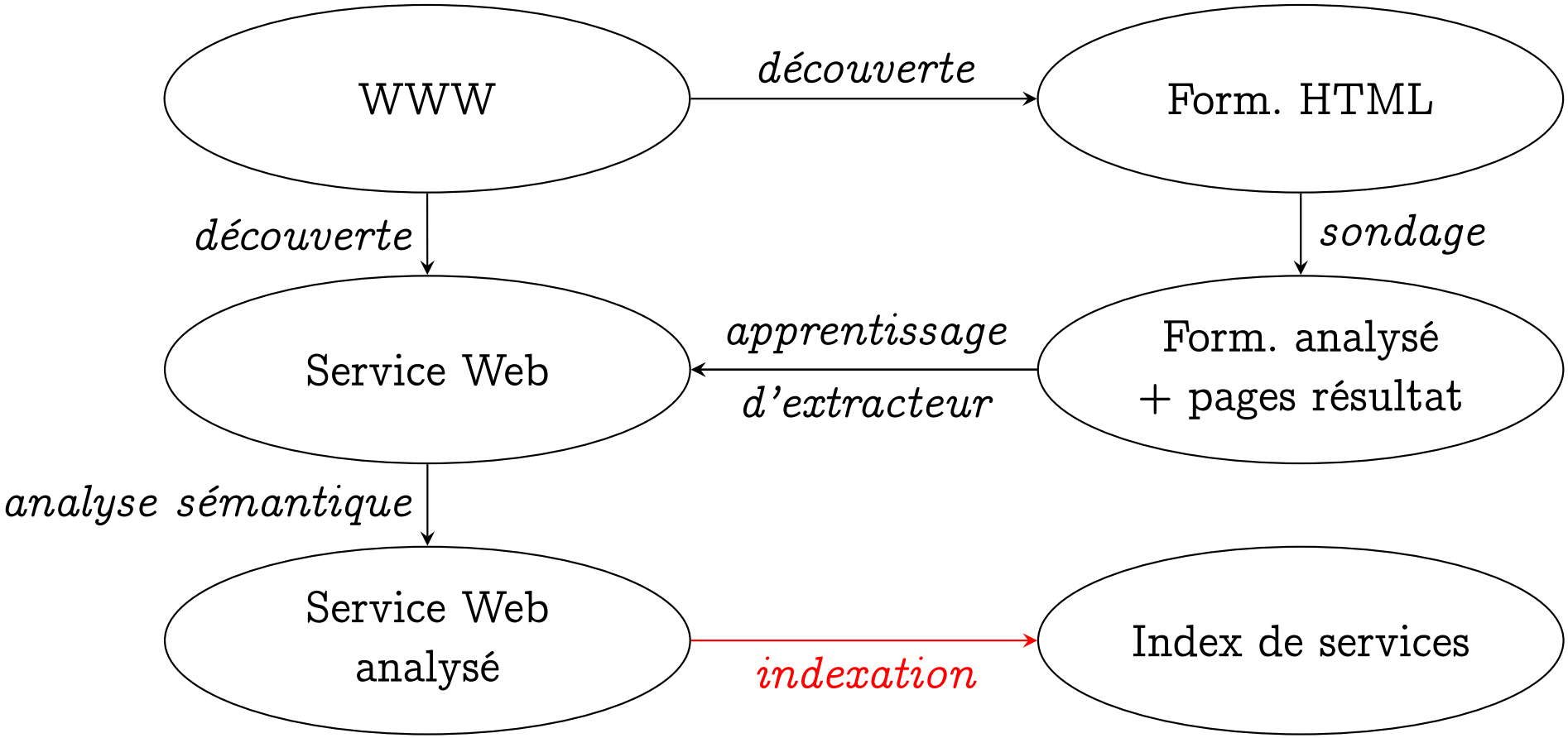
Processus automatique pour comprendre le Web caché



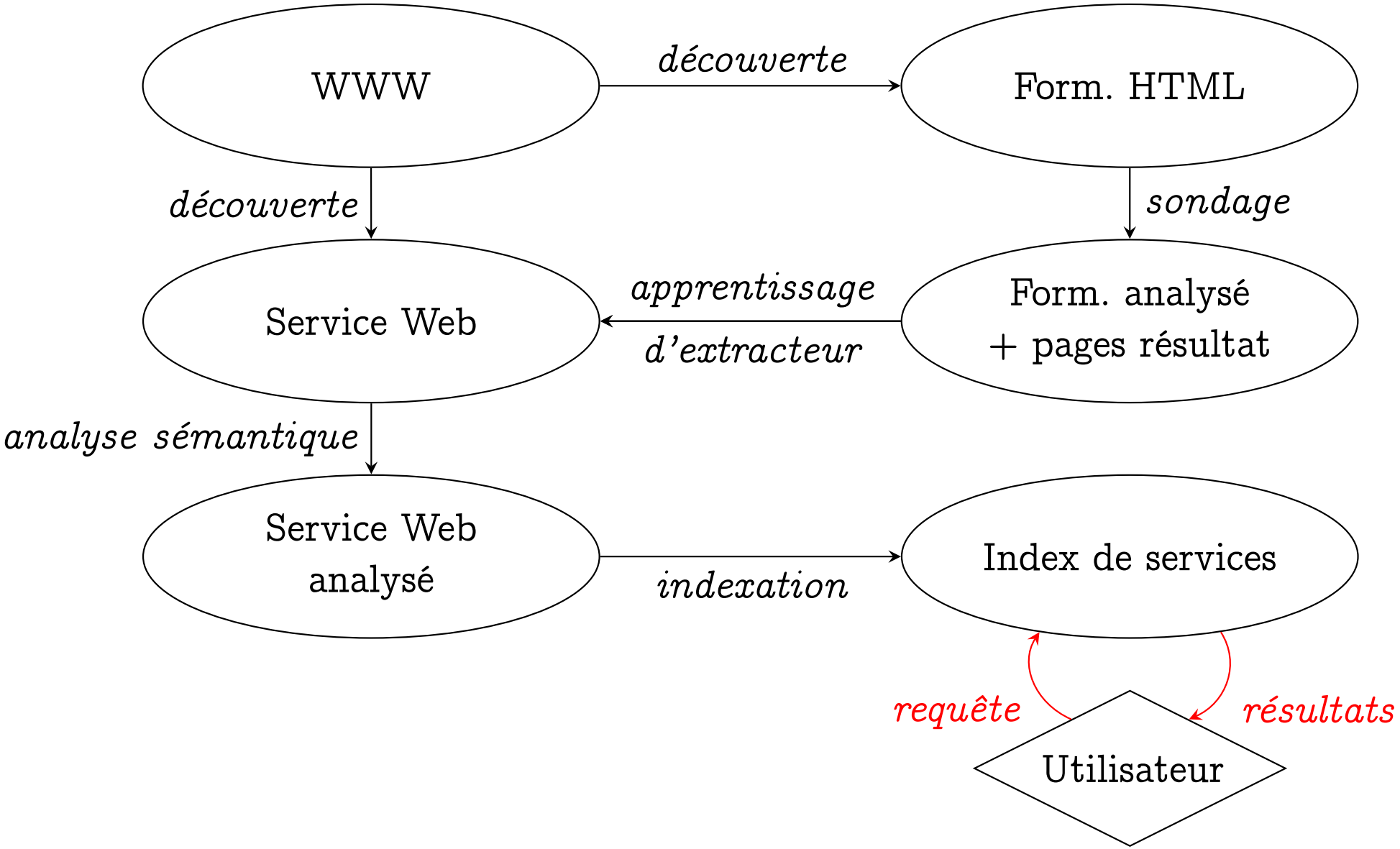
Processus automatique pour comprendre le Web caché



Processus automatique pour comprendre le Web caché

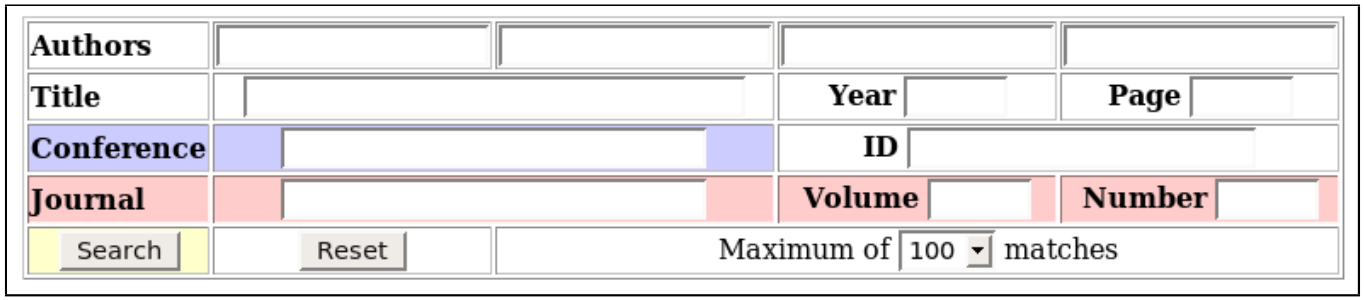


Processus automatique pour comprendre le Web caché



Sondage du Web caché

Analyse de la structure de formulaires HTML.



Problème

Associer à chaque champ de formulaire le concept du domaine approprié.

Collaboration avec Avin Mittal, IIT Bombay.

Sondage du Web caché

Analyse de la structure de formulaires HTML.

Travaux

- Heuristiques pour une première annotation.
- Sondage du formulaire par des mots du domaine pour infirmer ou confirmer.
- Naïf, mais bonne performance en pratique !

Probl

Assoc. …maine approprié.

Collaboration avec Avin Mittal, IIT Bombay.

Extraction automatique à partir de pages de résultats

Showing results 1 through 25 (of 94 total) for all:xml

1. cs.LO/0601085 [abs, ps, pdf, other] :
Title: **A Formal Foundation for ODRL**
Authors: Riccardo Pucella, Vicky Weissman
Comments: 30 pgs, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security), 2004
Subj-class: Logic in Computer Science; Cryptography and Security
ACM-class: H.2.7; K.4.4

2. astro-ph/0512493 [abs, pdf] :
Title: **VOFilter, Bridging Virtual Observatory and Industrial Office Applications**
Authors: Chen-zhou Cui (1), Markus Dolensky (2), Peter Quinn (2), Yong-heng Zhao (1), Francoise Genova (3) ((1)NAO China, (2) ESO, (3) CDS)
Comments: Accepted for publication in ChJAA (9 pages, 2 figures, 185KB)

3. cs.DS/0512061 [abs, ps, pdf, other] :
Title: **Matching Subsequences in Trees**
Authors: Philip Bille, Inge Li Goertz
Subj-class: Data Structures and Algorithms

4. cs.IR/0510025 [abs, ps, pdf, other] :
Title: **Practical Semantic Analysis of Web Sites and Documents**
Authors: Thierry Despeyroux (INRIA Rocquencourt / INRIA Sophia Antipolis)
Subj-class: Information Retrieval

5. cs.CR/0510013 [abs, pdf] :
Title: **Safe Data Sharing and Data Dissemination on Smart Devices**
Authors: Luc Bouganim (INRIA Rocquencourt), Cosmin Cremarenco (INRIA Rocquencourt), François Dang Ngoc (INRIA Rocquencourt, PRISM - UVSQ), Nicolas Dieu (INRIA Rocquencourt), Philippe Pucheral (INRIA Rocquencourt, PRISM - UVSQ)
Subj-class: Cryptography and Security; Databases

Collaboration avec équipe-projet MOSTRARE & Daniel Muschick, INRIA Futurs & Technische Univ. Graz.

Extraction automatique à partir de pages de résultats

Showing results 1 through 25 (of 94 total) for all:xml

1. cs.LO/0601085 [abs, ps, pdf, other] :
Title: **A Formal Foundation for ODRL**
Authors: Riccardo Pucella, Vicky Weissman
Comments: 30 pgs, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security), 2004
Subj-class: Logic in Computer Science; Cryptography and Security
ACM-class: H.2.7; K.4.4

2. astro-ph/0512493 [abs, pdf] :
Title: **VOFilter, Bridging Virtual Observatory and Industrial Office Applications**
Authors: Chen-zhou Cui (1), Markus Dolensky (2), Peter Quinn (2), Yong-heng Zhao (1), Francoise Genova (3) ((1)NAO China, (2) ESO, (3) CDS)
Comments: Accepted for publication in ChJAA (9 pages, 2 figures, 185KB)

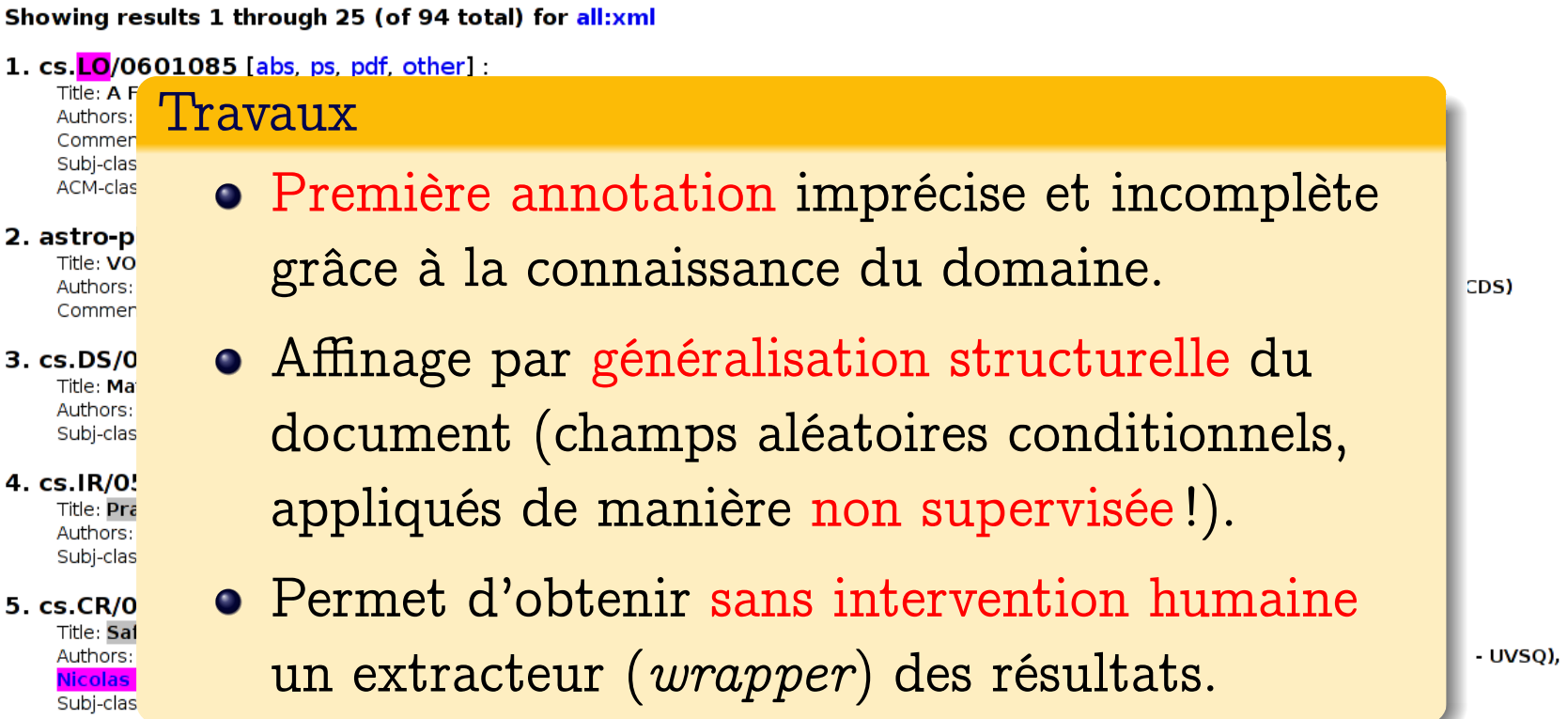
3. cs.DS/0512061 [abs, ps, pdf, other] :
Title: **Matching Subsequences in Trees**
Authors: Philip Bille, Inge Li Goertz
Subj-class: Data Structures and Algorithms

4. cs.IR/0510025 [abs, ps, pdf, other] :
Title: **Practical Semantic Analysis of Web Sites and Documents**
Authors: Thierry Despeyroux (INRIA Rocquencourt / INRIA Sophia Antipolis)
Subj-class: Information Retrieval

5. cs.CR/0510013 [abs, pdf] :
Title: **Safe Data Sharing and Data Dissemination on Smart Devices**
Authors: Luc Bouganim (INRIA Rocquencourt), Cosmin Cremarenco (INRIA Rocquencourt), François Dang Ngoc (INRIA Rocquencourt, PRISM - UVSQ), Nicolas Dieu (INRIA Rocquencourt), Philippe Pucheral (INRIA Rocquencourt, PRISM - UVSQ)
Subj-class: Cryptography and Security; Databases

Collaboration avec équipe-projet MOSTRARE & Daniel Muschick, INRIA Futurs & Technische Univ. Graz.

Extraction automatique à partir de pages de résultats



Travaux

- Première annotation imprécise et incomplète grâce à la connaissance du domaine.
- Affinage par généralisation structurelle du document (champs aléatoires conditionnels, appliqués de manière non supervisée !).
- Permet d'obtenir sans intervention humaine un extracteur (*wrapper*) des résultats.

Collaboration avec équipe-projet Mostrare & Daniel Muschick, INRIA Futurs & Technische Univ. Graz.

Perspectives

Découvrir et structurer données et informations

. . . du monde réel (p. ex. du Web)

Quelques perspectives sur des thèmes proches

- Utilisation de sources multiples pour corroborer les informations qu'elles produisent.
- Apprentissage sans overfitting en essayant de réduire la longueur de description de l'extracteur appris.
- Résolution de doublons (ou déduplication).