Finding Related Pages Using Green Measures: The Example of Wikipedia

Yann Ollivier

Pierre Senellart

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

ECOLE NORMALE SUPERIEURE DE LYON

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

INRIA FUTURS

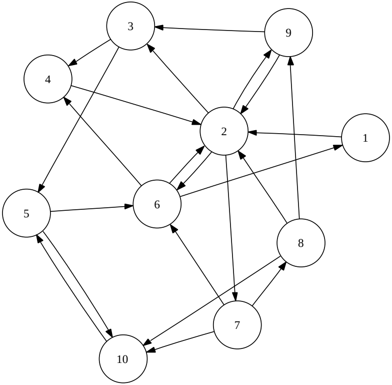
UNIVERSITÉ PARIS-SUD 11

AAAI

July 24th, 2007

Related nodes in a graph

Given a hyperlinked environment (= a graph)...

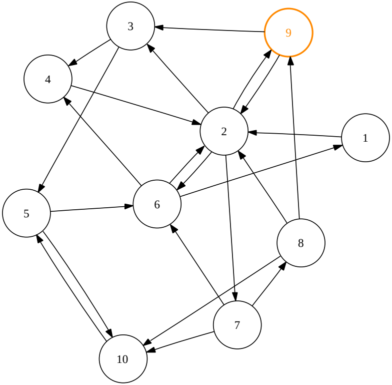


Problem

Finding nodes semantically related to some given node.

Related nodes in a graph

Given a hyperlinked environment (= a graph)...



Problem

Finding nodes semantically related to some given node.

Example of related nodes

Example (World Wide Web)

Nodes: Web pages

Edges: hyperlinks

Related nodes: similar/related pages (cf Google)

Example (Wikipedia)

Nodes: articles

Edges: hyperlinks

Related nodes: related articles (= articles on semantically related topics)

Example of related nodes

Example (World Wide Web)

Nodes: Web pages

Edges: hyperlinks

Related nodes: similar/related pages (cf Google)

Example (Wikipedia)

Nodes: articles

Edges: hyperlinks

Related nodes: related articles (= articles on semantically related topics)

Example of related nodes

Example (World Wide Web)

Nodes: Web pages

Edges: hyperlinks

Related nodes: similar/related pages (cf Google)

Example (Wikipedia)

Nodes: articles

Edges: hyperlinks

Related nodes: related articles (= articles on semantically related topics)

Classical approaches

Classical approaches for finding related nodes (e.g. on the World Wide Web):

- Based on the use of variants of PageRank on local subgraphs.
- Text Mining techniques : cocitations, vector-space model...

Our approach

Use of a classical Markov chain tool: Green measures.

Classical approaches

Classical approaches for finding related nodes (e.g. on the World Wide Web):

- Based on the use of variants of PageRank on local subgraphs.
- Text Mining techniques : cocitations, vector-space model...

Our approach

Use of a classical Markov chain tool: Green measures.

Classical approaches

Classical approaches for finding related nodes (e.g. on the World Wide Web):

- Based on the use of variants of PageRank on local subgraphs.
- Text Mining techniques : cocitations, vector-space model...

Our approach

Use of a classical Markov chain tool: Green measures.

Classical approaches

Classical approaches for finding related nodes (e.g. on the World Wide Web):

- Based on the use of variants of PageRank on local subgraphs.
- Text Mining techniques : cocitations, vector-space model...

Our approach

Use of a classical Markov chain tool: Green measures.

Contributions

Our contributions:

1. A novel use of Green measures for extracting semantic information in a graph.
2. An extensive comparative study with classical approaches, on the English version of Wikipedia.

Remark

Only pure mathematical methods, no Wikipedia-specific tricks included.

Contributions

Our contributions:

1. A novel use of Green measures for extracting semantic information in a graph.
2. An extensive comparative study with classical approaches, on the English version of Wikipedia.

Remark

Only pure mathematical methods, no Wikipedia-specific tricks included.

Contributions

Our contributions:

1. A novel use of Green measures for extracting semantic information in a graph.
2. An extensive comparative study with classical approaches, on the English version of Wikipedia.

Remark

Only pure mathematical methods, no Wikipedia-specific tricks included.

Contributions

Our contributions:

1. A novel use of Green measures for extracting semantic information in a graph.
2. An extensive comparative study with classical approaches, on the English version of Wikipedia.

Remark

Only pure mathematical methods, no Wikipedia-specific tricks included.

Outline

1. Introduction
2. Green measures
   - Graphs as Markov chains
   - Green measures
3. Methods Compared
4. Experiment on Wikipedia
5. Conclusion

Graph = Markov chain

Definition (Markov chain)

Probabilistic process on a state space, defined by transition probabilities p_{ij} from each state i to each state j.

For a directed graph:

State space: set of nodes

Transition probabilities: based on existence (and weight) of edges

Remark

All graphs will be supposed strongly connected and with gcd of length of all cycles equal to 1.

Graph = Markov chain

Definition (Markov chain)

Probabilistic process on a state space, defined by transition probabilities p_{ij} from each state i to each state j.

For a directed graph:

State space: set of nodes

Transition probabilities: based on existence (and weight) of edges

Remark

All graphs will be supposed strongly connected and with gcd of length of all cycles equal to 1.

Graph = Markov chain

Definition (Markov chain)

Probabilistic process on a state space, defined by transition probabilities p_{ij} from each state i to each state j.

For a directed graph:

State space: set of nodes

Transition probabilities: based on existence (and weight) of edges

Remark

All graphs will be supposed strongly connected and with gcd of length of all cycles equal to 1.

Graph = Markov chain

Definition (Markov chain)

Probabilistic process on a state space, defined by transition probabilities p_{ij} from each state i to each state j.

For a directed graph:

State space: set of nodes

Transition probabilities: based on existence (and weight) of edges

Remark

All graphs will be supposed strongly connected and with gcd of length of all cycles equal to 1.

Equilibrium measure

Definition (Measure)

Assignments of real numbers to the state set.

Definition (Propagation operator)

Operator which maps a measure μ to a measure μ' computed as:

$$\mu'_j = \sum_i (\mu_i p_{ij})$$

Result

If we iterate the propagation operator from any measure summing to 1, we converge to a unique equilibrium measure. (PageRank with no random jumps).

Equilibrium measure

Definition (Measure)

Assignments of real numbers to the state set.

Definition (Propagation operator)

Operator which maps a measure μ to a measure μ' computed as:

$$\mu'_j = \sum_i (\mu_i p_{ij})$$

Result

If we iterate the propagation operator from any measure summing to 1, we converge to a unique equilibrium measure. (PageRank with no random jumps).

Equilibrium measure

Definition (Measure)

Assignments of real numbers to the state set.

Definition (Propagation operator)

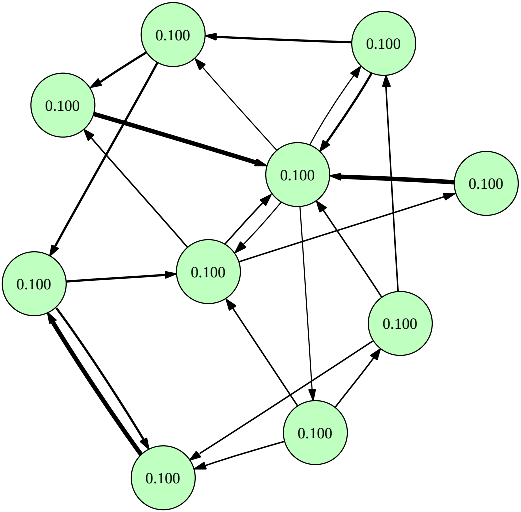
Operator which maps a measure μ to a measure μ' computed as:

$$\mu'_j = \sum_i (\mu_i p_{ij})$$

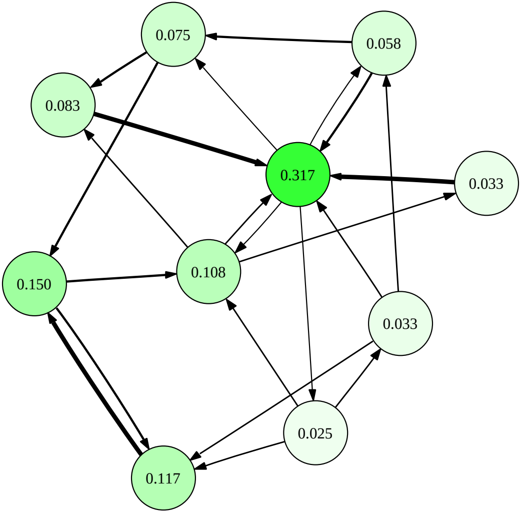
Result

If we iterate the propagation operator from any measure summing to 1, we converge to a unique equilibrium measure. (PageRank with no random jumps).

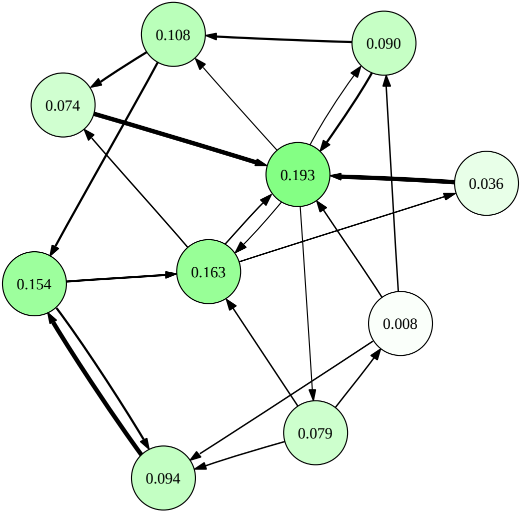
PageRank — Iteration ♯1



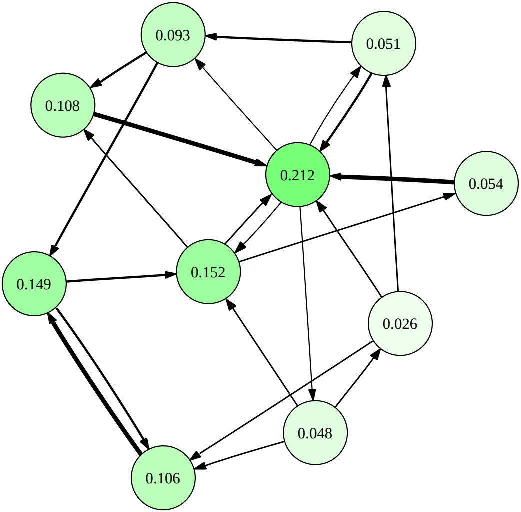
PageRank — Iteration ♯2



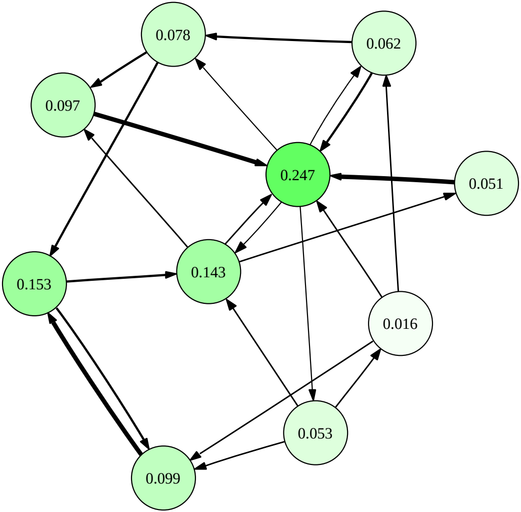
PageRank — Iteration ♯3



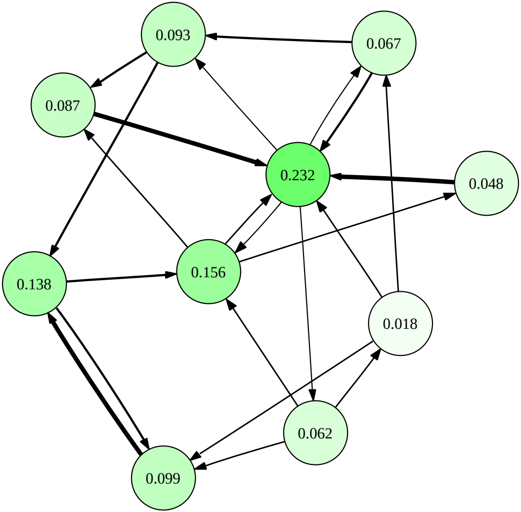
PageRank — Iteration ♯4



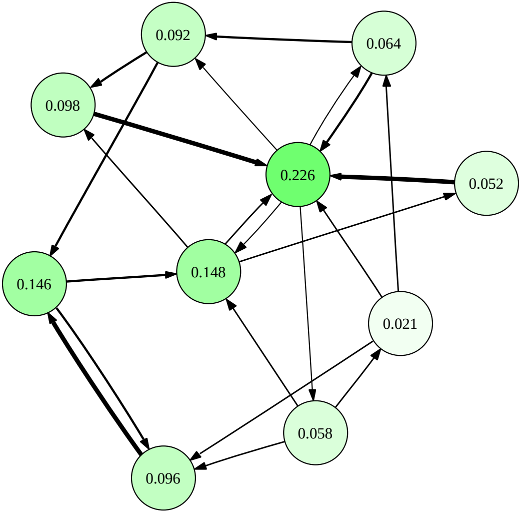
PageRank — Iteration ♯5



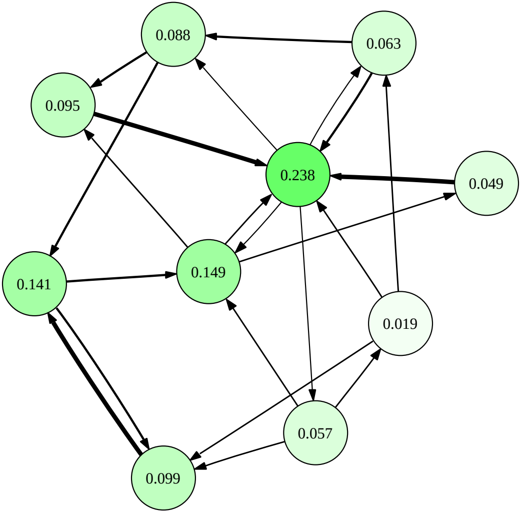
PageRank — Iteration ♯6



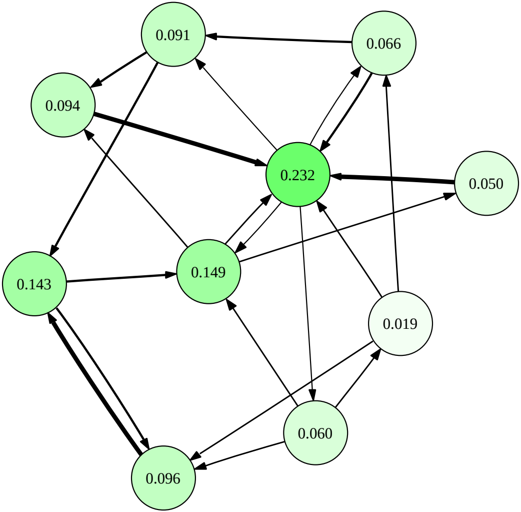
PageRank — Iteration ♯7



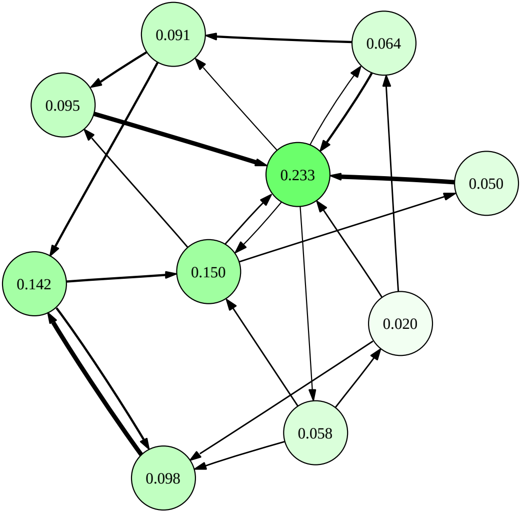
PageRank — Iteration ♯8



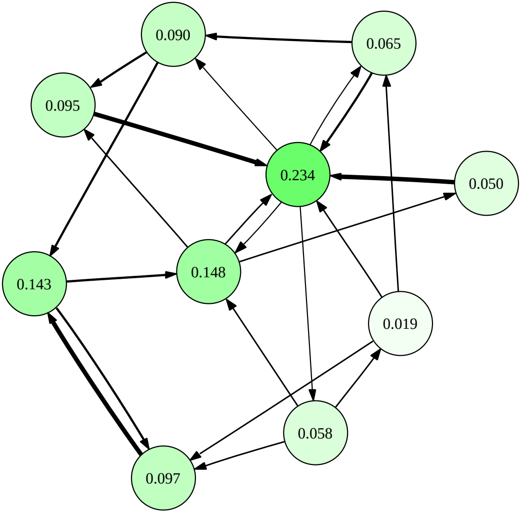
PageRank — Iteration ♯9



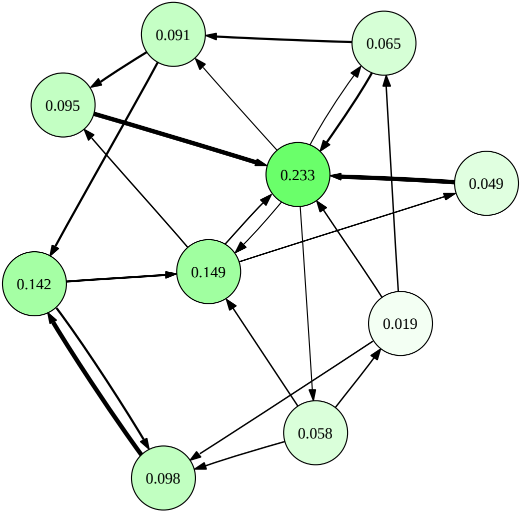
PageRank — Iteration ♯10



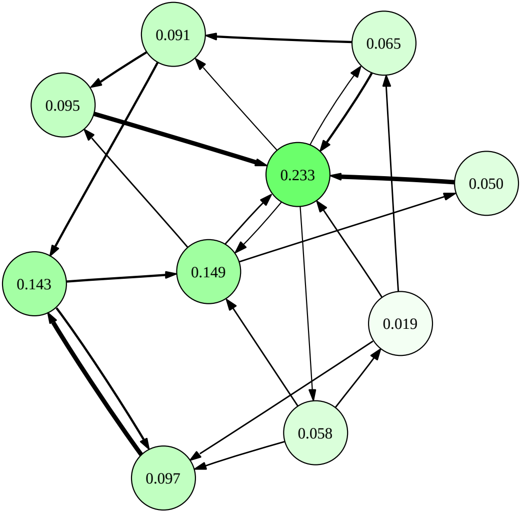
PageRank — Iteration ♯11



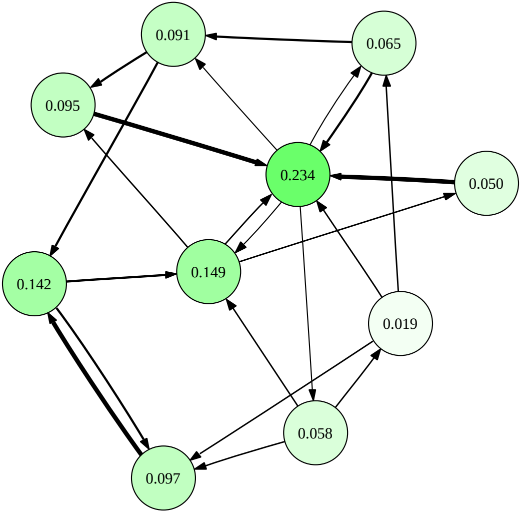
PageRank — Iteration ♯12



PageRank — Iteration ♯13



PageRank — Iteration ♯14



Background on Green measures

Green functions

- Come from electrostatic theory (potential created by a charge distribution).
- Analogy between electrostatic potential theory and Markov chains.
- Green measures: discrete analogues of Green functions.

Background on Green measures

Green functions

- Come from electrostatic theory (potential created by a charge distribution).
- Analogy between electrostatic potential theory and Markov chains.
- Green measures: discrete analogues of Green functions.

Background on Green measures

Green functions

- Come from electrostatic theory (potential created by a charge distribution).
- Analogy between electrostatic potential theory and Markov chains.
- Green measures: discrete analogues of Green functions.

Definition of Green measures

Definition (Green measure centered at node i)

Only fixed point of the operator on measures defined by:

$$\mu_j \mapsto \sum_k (\mu_k p_{kj}) + (\delta_{ij} - \nu_j) \quad \text{where} \quad \delta_{ij} = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ otherwise} \end{cases}$$

Interpretations

- PageRank with source at i: standard PageRank computation while, at each iteration, adding 1 to the measure of i, and subtracting ν_j to every node j.
- Time spent at a node knowing the initial node is i.

Definition of Green measures

Definition (Green measure centered at node i)

Only fixed point of the operator on measures defined by:

$$\mu_j \mapsto \sum_k (\mu_k p_{kj}) + (\delta_{ij} - \nu_j) \quad \text{where} \quad \delta_{ij} = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ otherwise} \end{cases}$$

Interpretations

- PageRank with source at i: standard PageRank computation while, at each iteration, adding 1 to the measure of i, and subtracting ν_j to every node j.
- Time spent at a node knowing the initial node is i.

Definition of Green measures

Definition (Green measure centered at node i)

Only fixed point of the operator on measures defined by:

$$\mu_j \mapsto \sum_k (\mu_k p_{kj}) + (\delta_{ij} - \nu_j) \quad \text{where} \quad \delta_{ij} = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ otherwise} \end{cases}$$

Interpretations

- PageRank with source at i: standard PageRank computation while, at each iteration, adding 1 to the measure of i, and subtracting ν_j to every node j.
- Time spent at a node knowing the initial node is i.

Definition of Green measures

Definition (Green measure centered at node i)

Only fixed point of the operator on measures defined by:

$$\mu_j \mapsto \sum_k (\mu_k p_{kj}) + (\delta_{ij} - \nu_j) \quad \text{where} \quad \delta_{ij} = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ otherwise} \end{cases}$$

Interpretations

- PageRank with source at i: standard PageRank computation while, at each iteration, adding 1 to the measure of i, and subtracting ν_j to every node j.
- Time spent at a node knowing the initial node is i.

Outline

1 Introduction

2 Green measures

3 Methods Compared

- *Green* and *SymGreen*
- *PageRankOfLinks*
- *Cosine*
- *Cocitations*

4 Experiment on Wikipedia

5 Conclusion

Purpose

- Finding nodes in the graph related to i.
- For each method, output an ordered list of nodes related to i.
- Each method provides a similarity score to i.

Purpose

- Finding nodes in the graph related to i.
- For each method, output an ordered list of nodes related to i.
- Each method provides a similarity score to i.

Purpose

- Finding nodes in the graph related to i.
- For each method, output an ordered list of nodes related to i.
- Each method provides a similarity score to i.

Green — Method Description

Method Description

- Direct application of the theory of Green measures.
- Improvement: multiplication by a term favoring uncommon nodes $\log(1/\nu_j)$ (quantity of information).
- Iteration until reasonable convergence on the top results.

Green — Method Description

Method Description

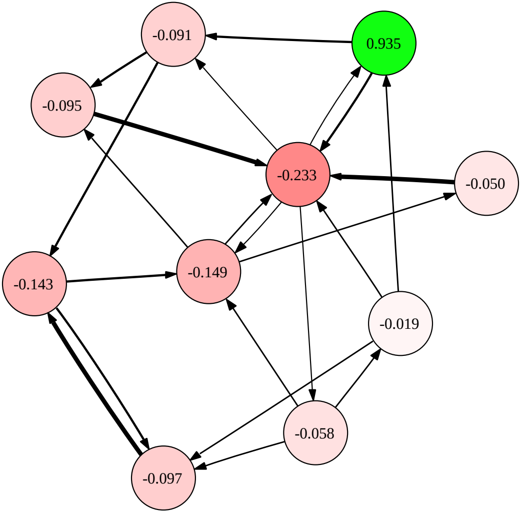
- Direct application of the theory of Green measures.
- Improvement: multiplication by a term favoring uncommon nodes $\log(1/\nu_j)$ (quantity of information).
- Iteration until reasonable convergence on the top results.

Green — Method Description

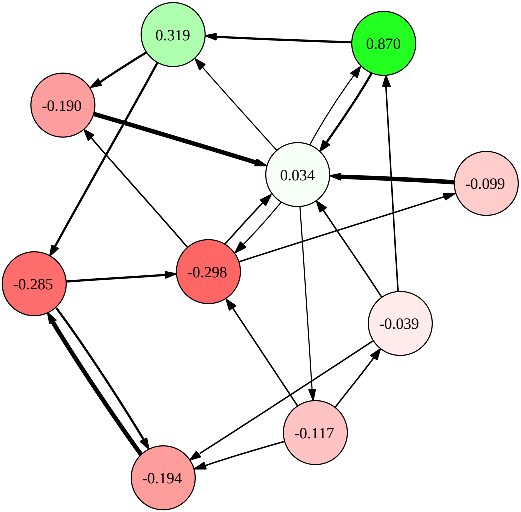
Method Description

- Direct application of the theory of Green measures.
- Improvement: multiplication by a term favoring uncommon nodes $\log(1/\nu_j)$ (quantity of information).
- Iteration until reasonable convergence on the top results.

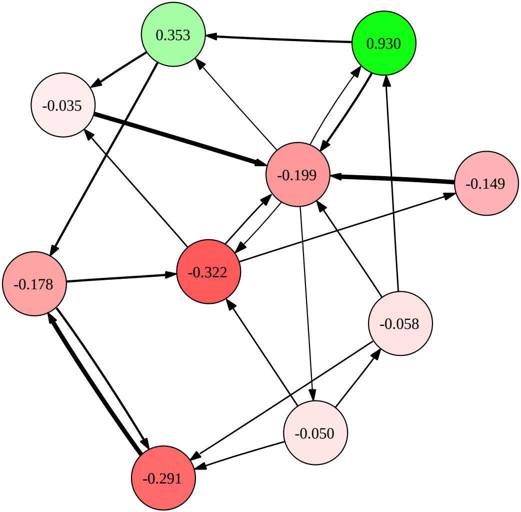
Green — Iteration ♯1



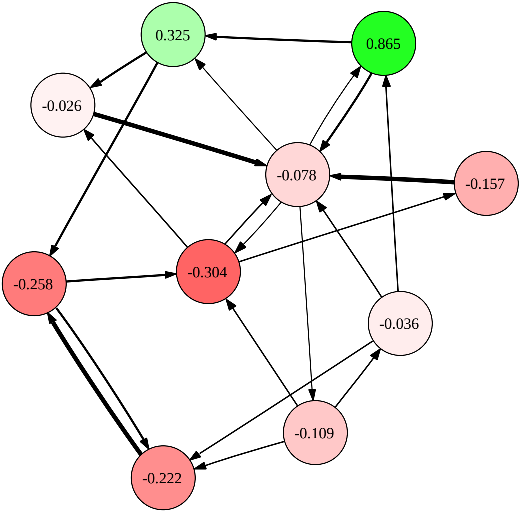
Green — Iteration ♯2



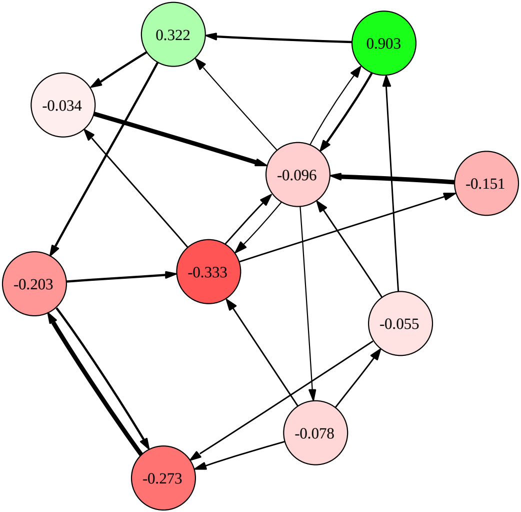
Green — Iteration ♯3



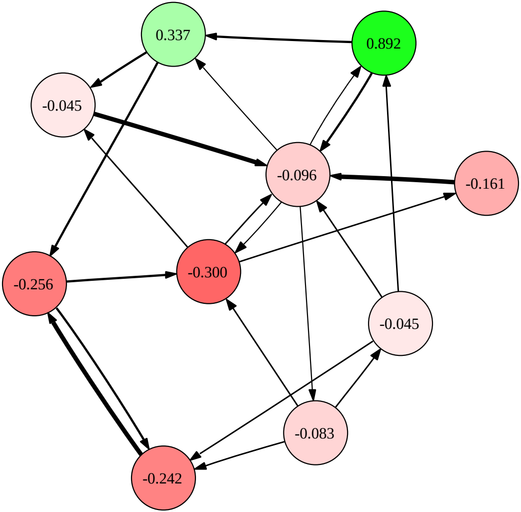
Green — Iteration ♯4



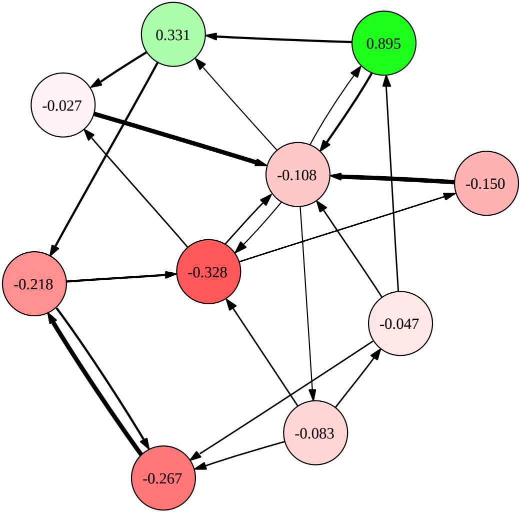
Green — Iteration ♯5



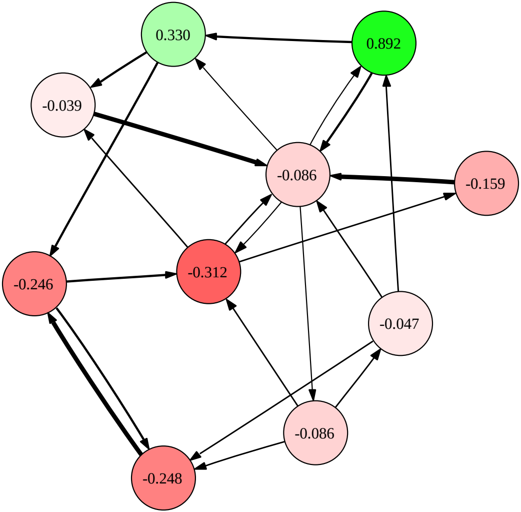
Green — Iteration ♯6



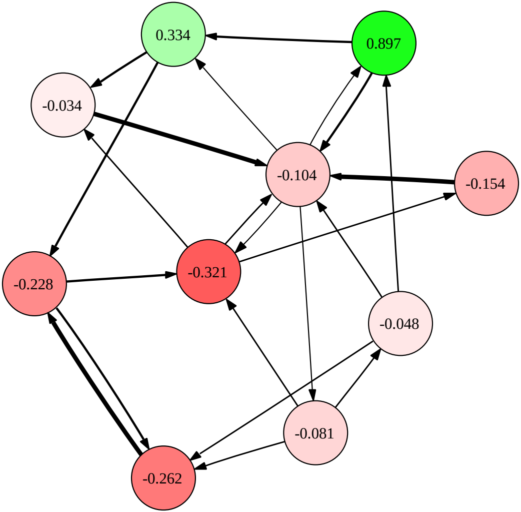
Green — Iteration ♯7



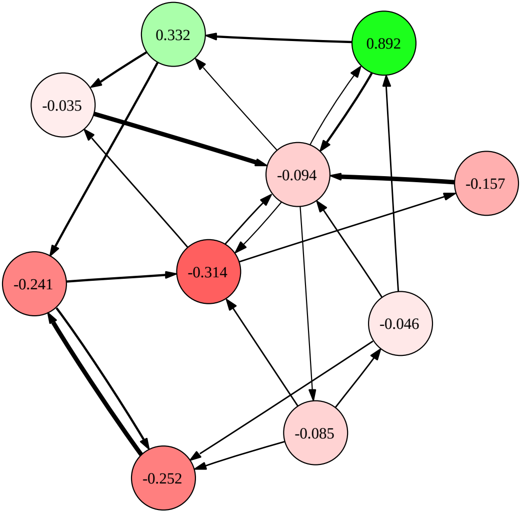
Green — Iteration ♯8



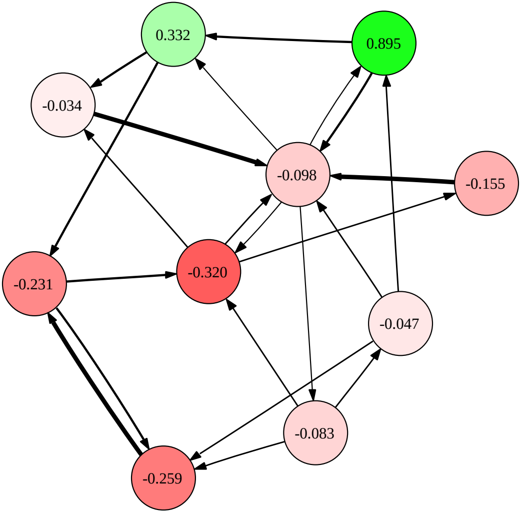
Green — Iteration ♯9



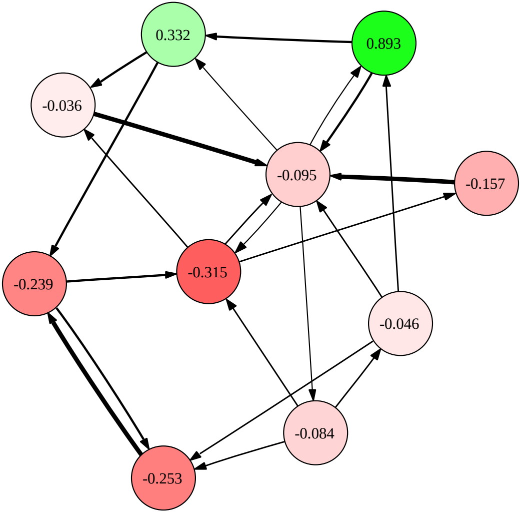
Green — Iteration ♯10



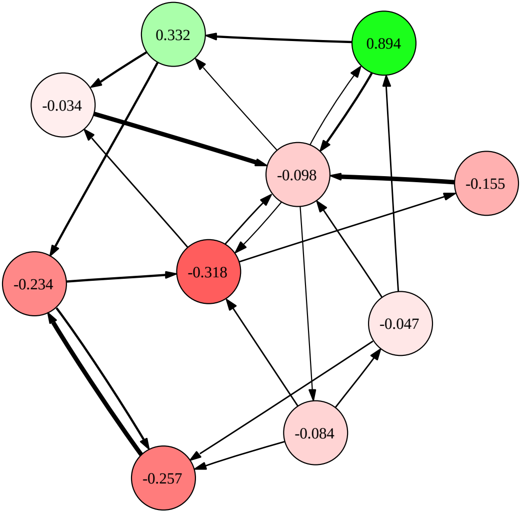
Green — Iteration ♯11



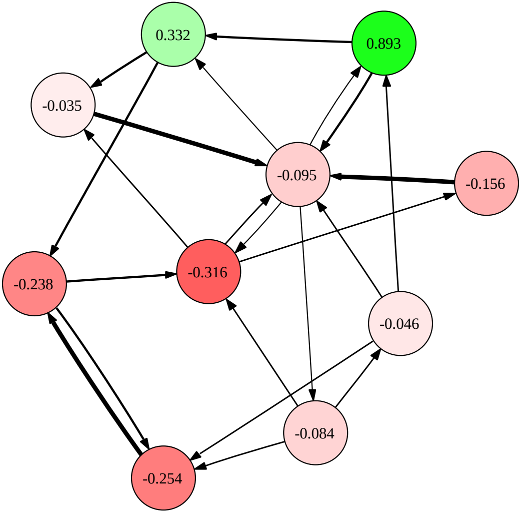
Green — Iteration ♯12



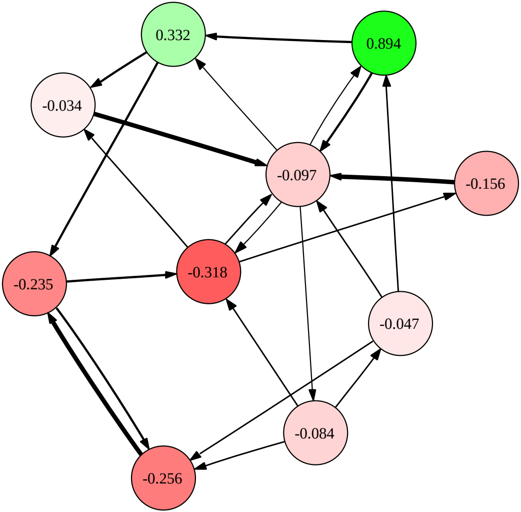
Green — Iteration ♯13



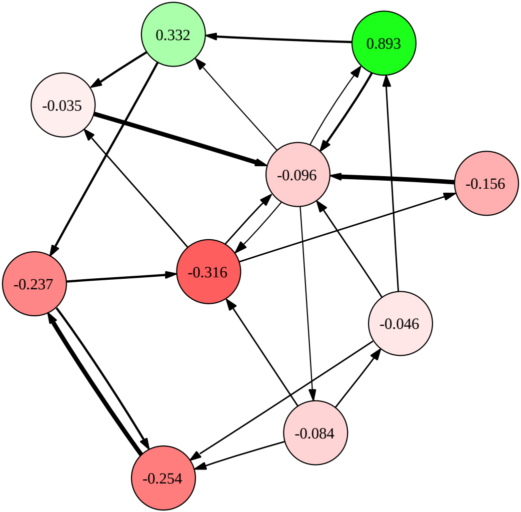
Green — Iteration ♯14



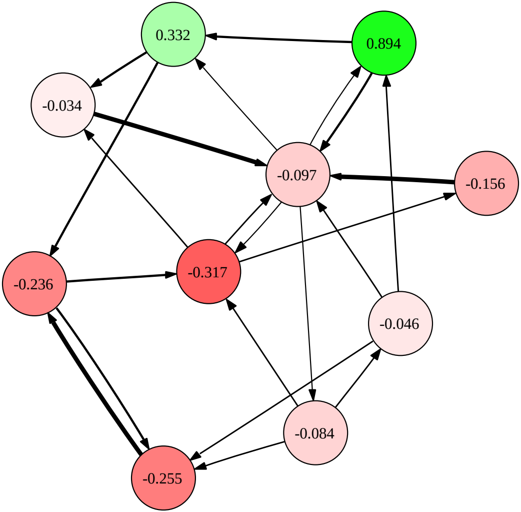
Green — Iteration ♯15



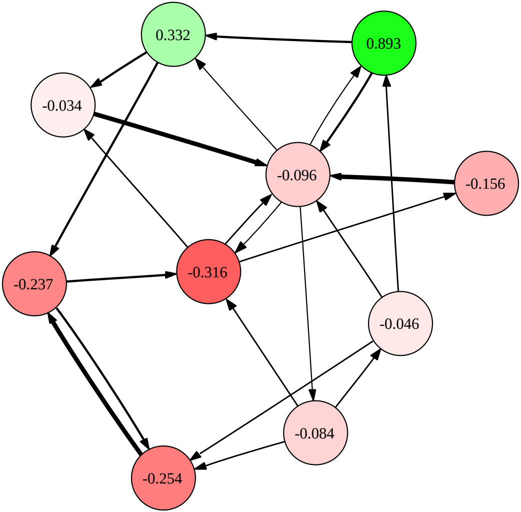
Green — Iteration ♯16



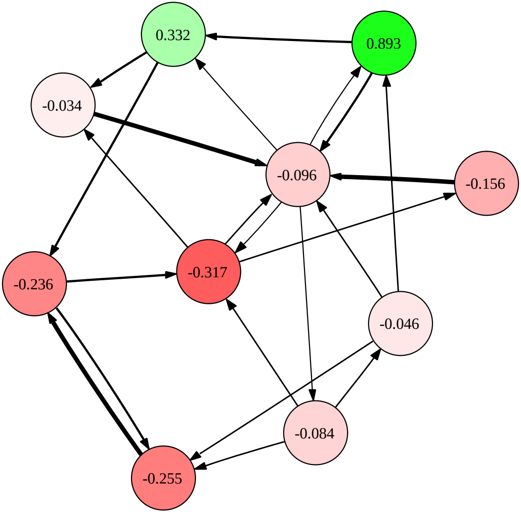
Green — Iteration ♯17



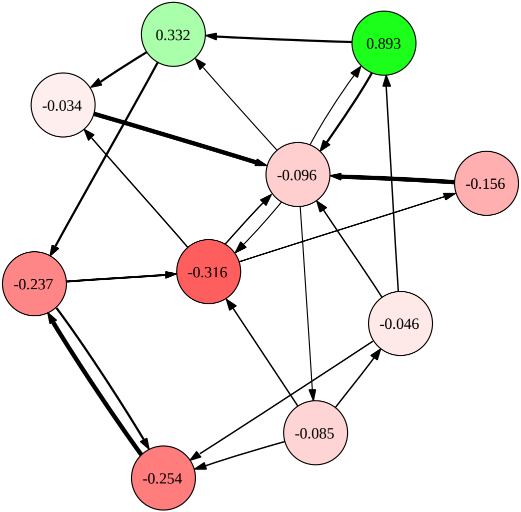
Green — Iteration ♯18



Green — Iteration ♯19



Green — Iteration ♯20



SymGreen — Method Description

Method Description

- *Green* goes only forward, may be a limitation.
- Symmetrize the graph, in a canonical sense in relation to the equilibrium measure:

$$\tilde{p}_{ij} = \frac{1}{2}\left(p_{ij} + p_{ji}\frac{\nu_j}{\nu_i}\right)$$

  The resulting graph has the same equilibrium measure.
- Same as *Green* on this symmetrized graph.

SymGreen — Method Description

Method Description

- *Green* goes only forward, may be a limitation.
- Symmetrize the graph, in a canonical sense in relation to the equilibrium measure:

$$\tilde{p}_{ij} = \frac{1}{2}(p_{ij} + p_{ji}\frac{\nu_j}{\nu_i})$$

The resulting graph has the same equilibrium measure.

- Same as *Green* on this symmetrized graph.

SymGreen — Method Description

Method Description

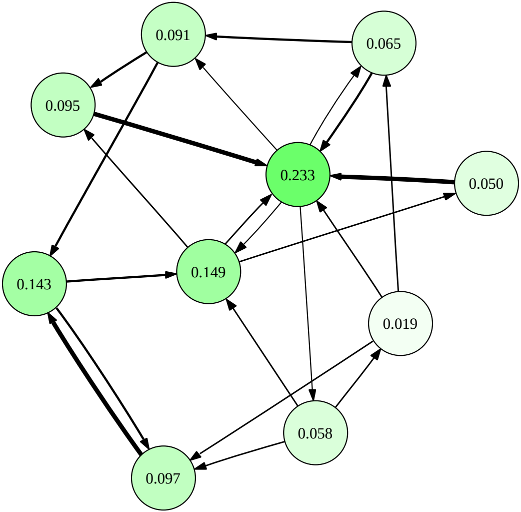
- *Green* goes only forward, may be a limitation.
- Symmetrize the graph, in a canonical sense in relation to the equilibrium measure:

$$\tilde{p}_{ij} = \frac{1}{2}(p_{ij} + p_{ji}\frac{\nu_j}{\nu_i})$$

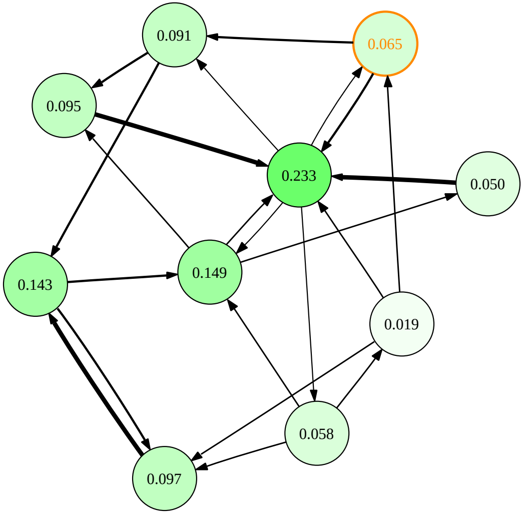
The resulting graph has the same equilibrium measure.

- Same as *Green* on this symmetrized graph.

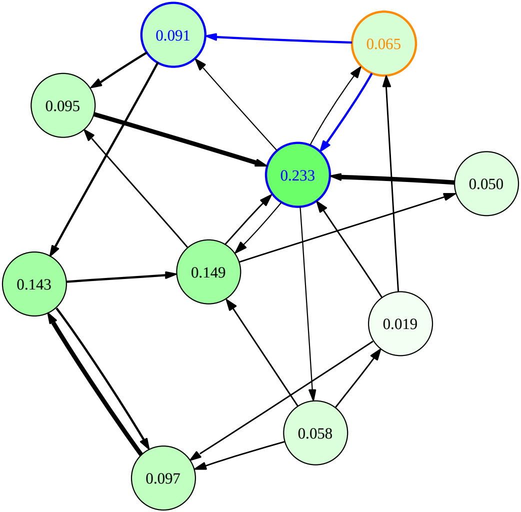
PageRankOfLinks



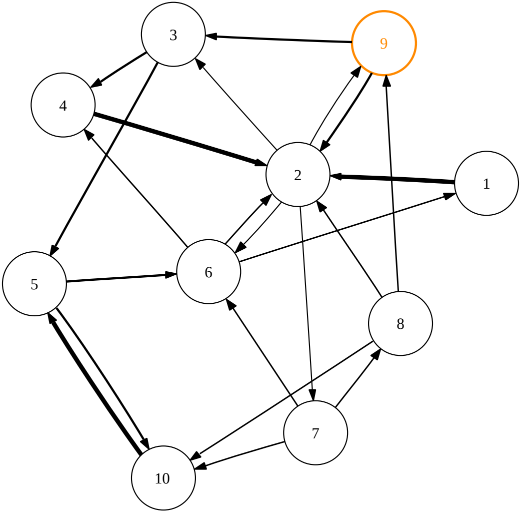
PageRankOfLinks



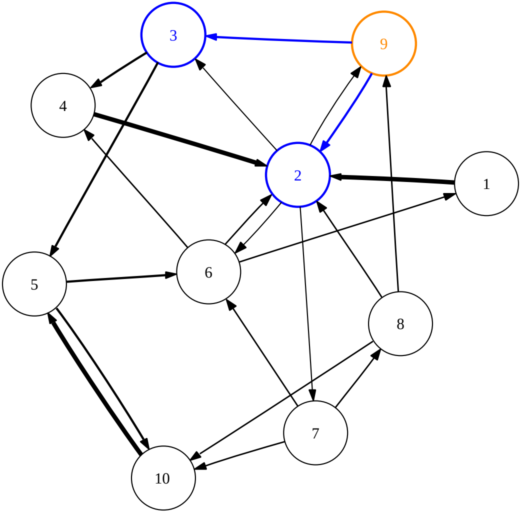
PageRankOfLinks



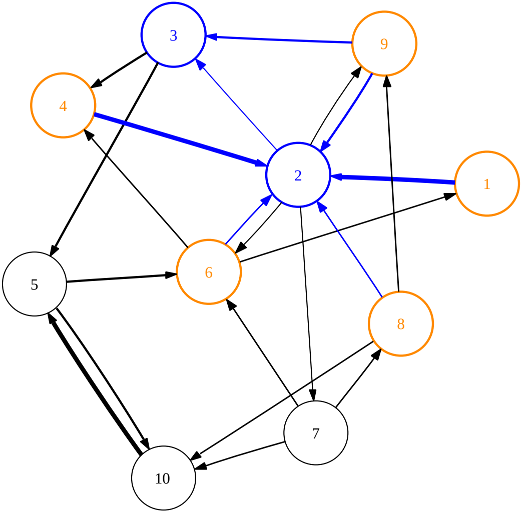
Cosine



Cosine



Cosine



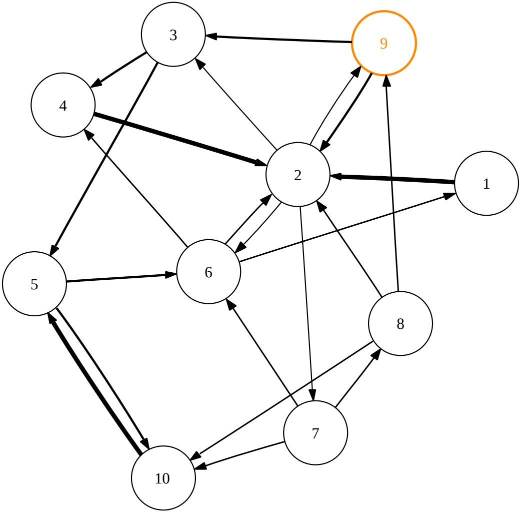
Cosine

Dimensions

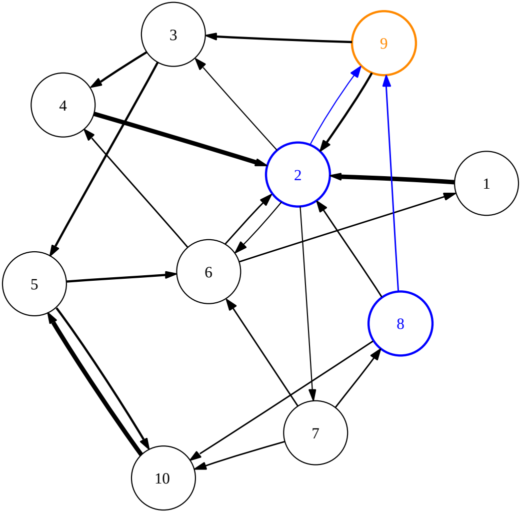
Documents

	1	2	3	4	6	7	9	10	Cosine with 9
1		✓							0.40
2			✓		✓	✓	✓		0.43
4		✓							0.40
6	✓	✓		✓					0.09
8		✓					✓	✓	0.13
9		✓	✓						1.00

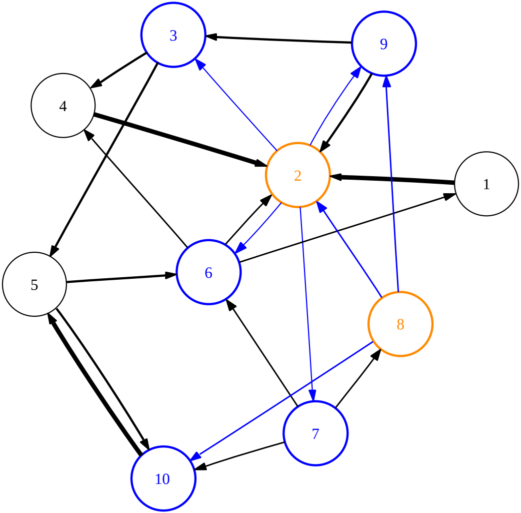
Cocitations



Cocitations



Cocitations



Outline

1. Introduction
2. Green measures
3. Methods Compared
4. Experiment on Wikipedia
   - Wikipedia graph
   - Evaluation
   - Results
5. Conclusion

The graph of Wikipedia

Statistics

- $1,606,896$ nodes (as of September 25th, 2006).
- $38,896,462$ edges.
- 95% of the nodes belong to the largest strongly connected component.

Evaluation methodology

- Blind evaluation of the methods.
- Articles selected for their diversity:
  - Clique (graph theory)
  - Germany
  - Hungarian language
  - Pierre de Fermat
  - Star Wars
  - Theory of relativity
  - 1989
- 66 evaluators asked to give a mark to each list of words.

Evaluation methodology

- Blind evaluation of the methods.
- Articles selected for their diversity:
  - Clique (graph theory)
  - Germany
  - Hungarian language
  - Pierre de Fermat
  - Star Wars
  - Theory of relativity
  - 1989
- 66 evaluators asked to give a mark to each list of words.

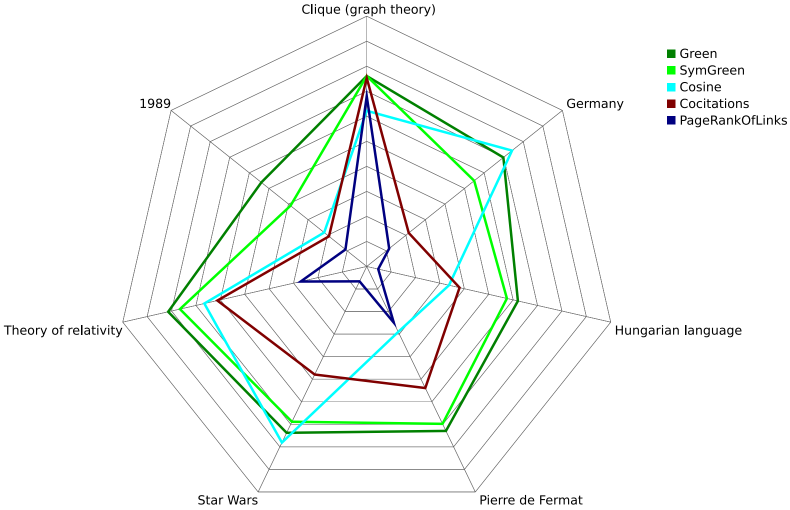
Evaluation methodology

- Blind evaluation of the methods.
- Articles selected for their diversity:
  - Clique (graph theory)
  - Germany
  - Hungarian language
  - Pierre de Fermat
  - Star Wars
  - Theory of relativity
  - 1989
- 66 evaluators asked to give a mark to each list of words.

Output on Germany

Green	*SymGreen*	*PageRankOfLinks*	*Cosine*	*Cocitations*
1. Germany	1. Germany	1. United States	1. Germany	1. Germany
2. Berlin	2. Berlin	2. United Kingdom	2. History of Germany since 1945	2. United States
3. German language	3. France	3. France	3. History of Germany	3. France
4. Christian Democratic Union (Germany)	4. Austria	4. 2005	4. Timeline of German history	4. United Kingdom
5. Austria	5. German language	5. Germany	5. States of Germany	5. World War II
6. Hamburg	6. Bavaria	6. World War II	6. Politics of Germany	6. Italy
7. German reunification	7. World War II	7. Canada	7. List of Germany-related topics	7. Netherlands
8. Social Democratic Party of Germany	8. German Democratic Republic	8. English language	8. Hildesheimer Rabbinical Seminary	8. Japan
9. German Empire	9. European Union	9. Japan	9. Pleasure Victim	9. 2005
10. German Democratic Republic	10. Hamburg	10. Italy	10. German Unity Day	10. Category:Living people

Results


Clique (graph theory)
Germany
Hungarian language
Pierre de Fermat
Star Wars
Theory of relativity
1989
🟩 Green
🟩 SymGreen
🟦 Cosine
🟥 Cocitations
🟦 PageRankOfLinks

Outline

1 Introduction

2 Green measures

3 Methods Compared

4 Experiment on Wikipedia

5 Conclusion
- Summary
- Perspectives

Summary

- Green measures: a tool for extracting semantic information in a graph.
- In comparison to other methods, in the case of Wikipedia:
    - Better overall performance.
    - Robustness.
    - Discovery of relevant semantic relations.

Summary

- Green measures: a tool for extracting semantic information in a graph.
- In comparison to other methods, in the case of Wikipedia:
  - Better overall performance.
  - Robustness.
  - Discovery of relevant semantic relations.

Summary

- Green measures: a tool for extracting semantic information in a graph.
- In comparison to other methods, in the case of Wikipedia:
  - Better overall performance.
  - Robustness.
  - Discovery of relevant semantic relations.

Summary

- Green measures: a tool for extracting semantic information in a graph.
- In comparison to other methods, in the case of Wikipedia:
  - Better overall performance.
  - Robustness.
  - Discovery of relevant semantic relations.

Summary

- Green measures: a tool for extracting semantic information in a graph.
- In comparison to other methods, in the case of Wikipedia:
  - Better overall performance.
  - Robustness.
  - Discovery of relevant semantic relations.

Perspectives

- Application to the Web graph.
- Interpolation between *Green* and *SymGreen*.
- Clustering using *Green* measures: unpractical now because of computation times.
- Use of *Green* measures on other Markov chains, e.g. for computing authority scores.

Perspectives

- Application to the Web graph.
- Interpolation between *Green* and *SymGreen*.
- Clustering using *Green* measures: unpractical now because of computation times.
- Use of *Green* measures on other Markov chains, e.g. for computing authority scores.

Perspectives

- Application to the Web graph.
- Interpolation between *Green* and *SymGreen*.
- Clustering using *Green* measures: unpractical now because of computation times.
- Use of *Green* measures on other Markov chains, e.g. for computing authority scores.

Perspectives

- Application to the Web graph.
- Interpolation between *Green* and *SymGreen*.
- Clustering using *Green* measures: unpractical now because of computation times.
- Use of *Green* measures on other Markov chains, e.g. for computing authority scores.