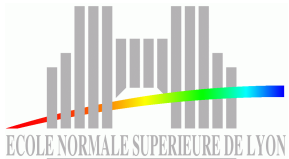
Finding Related Pages Using Green Measures: The Example of Wikipedia

Yann Ollivier[1,2] Pierre Senellart[3,4]

1  2 

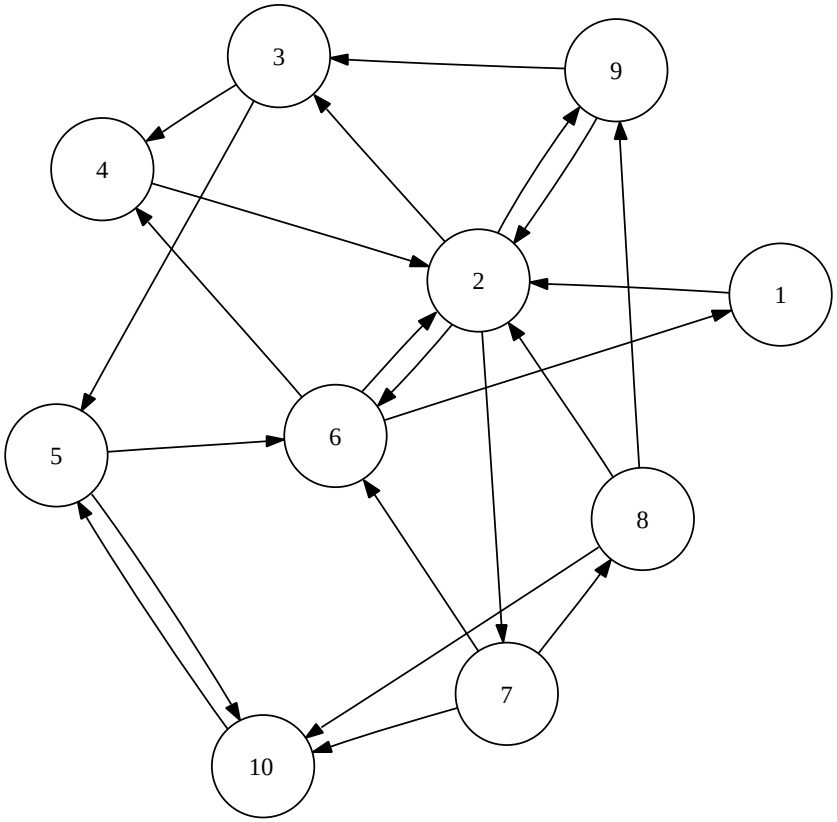
3 

4 

Gemo Seminar
November 10th, 2006

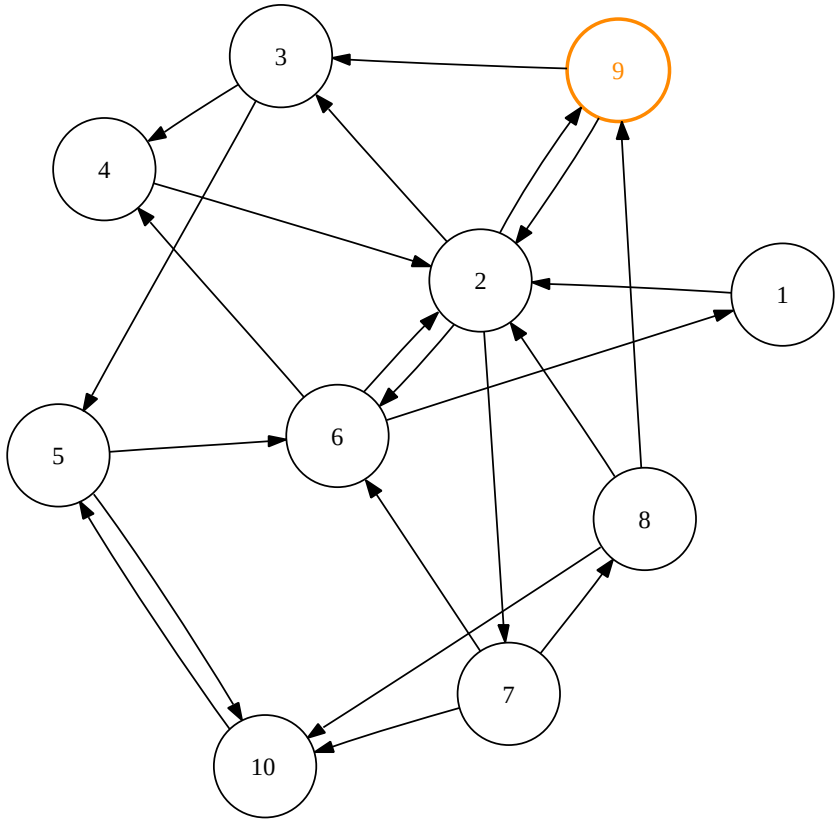
Related nodes in a graph

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



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 of related nodes

Example (World Wide Web)

Nodes: Web pages

Edges: hyperlinks

Related nodes: similar/related pages (cf Google)

Example of related nodes

Example (World Wide Web)

Nodes: Web pages

Edges: hyperlinks

Related nodes: similar/related pages (cf Google)

Example (Publications)

Nodes: publications

Edges: citations

Related nodes: related articles (cf Google Scholar)

Example of related nodes

Example (World Wide Web)

Nodes: Web pages

Edges: hyperlinks

Related nodes: similar/related pages (cf Google)

Example (Publications)

Nodes: publications

Edges: citations

Related nodes: related articles (cf Google Scholar)

Example (Wikipedia)

Nodes: articles

Edges: hyperlinks

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

Classical approaches

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

Classical approaches

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

- Based on the use of variants of PageRank on local subgraphs.

Classical approaches

Classical approached 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...

Classical approaches

Classical approached 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:

Contributions

Our contributions:

1. A novel use of Green measures for extracting semantic information in a graph.

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.

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

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.

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:

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

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.

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:

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

Equilibrium measure

Definition (Measure)

Assignments of real numbers to the state set.

Definition (Propagation operator)

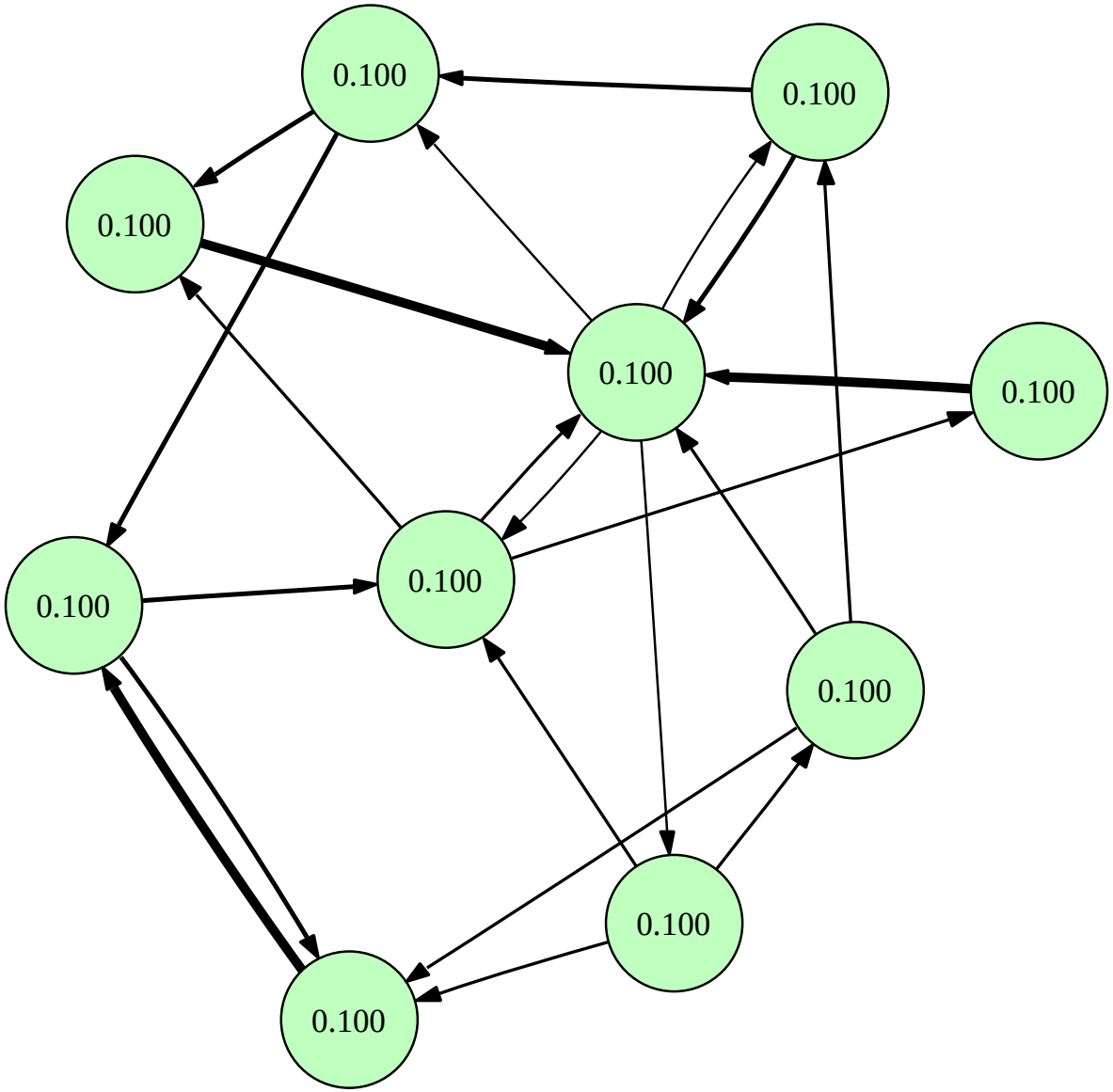
Operator which maps a measure μ to a measure μ' computed as:

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

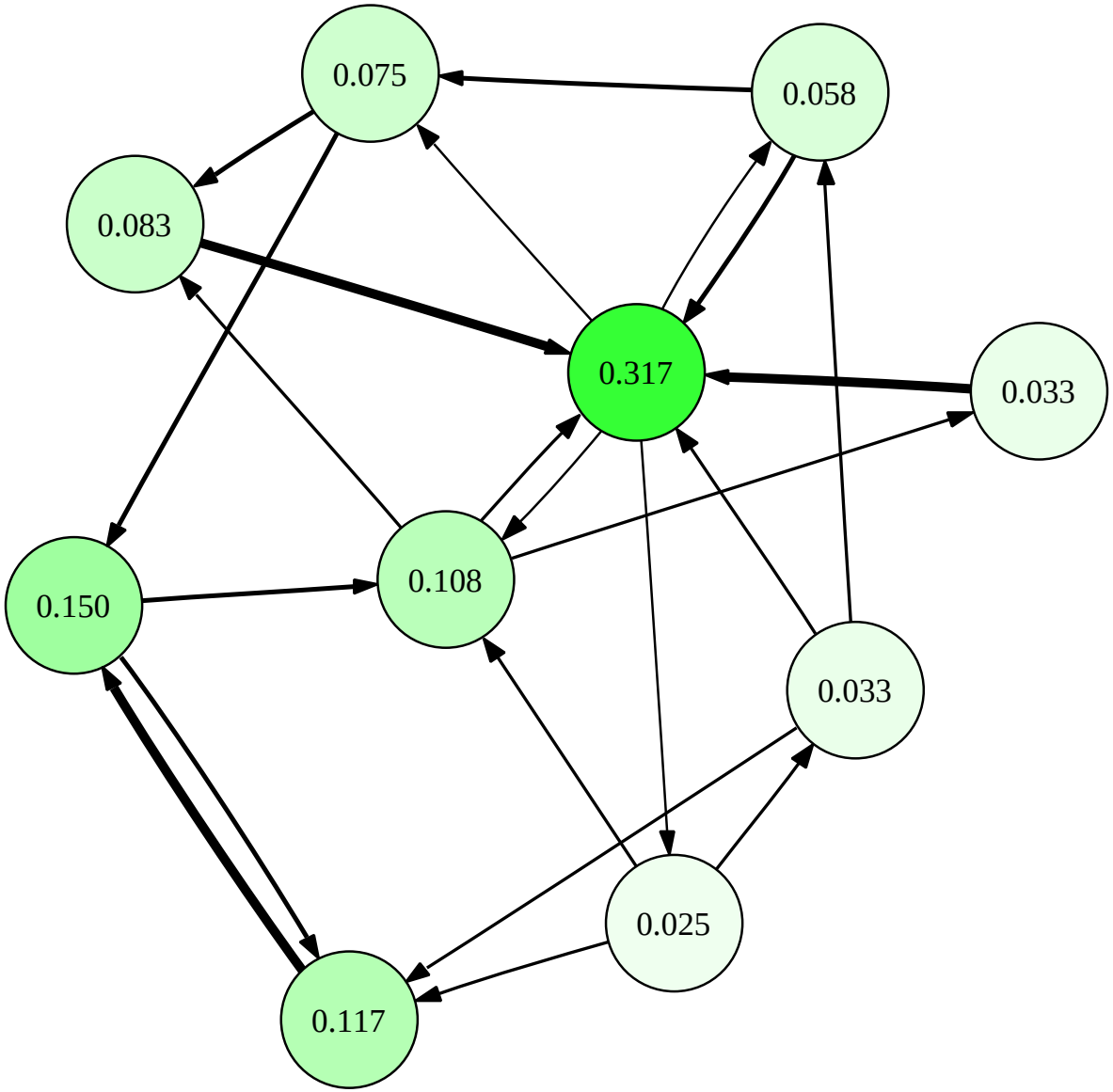
Result

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

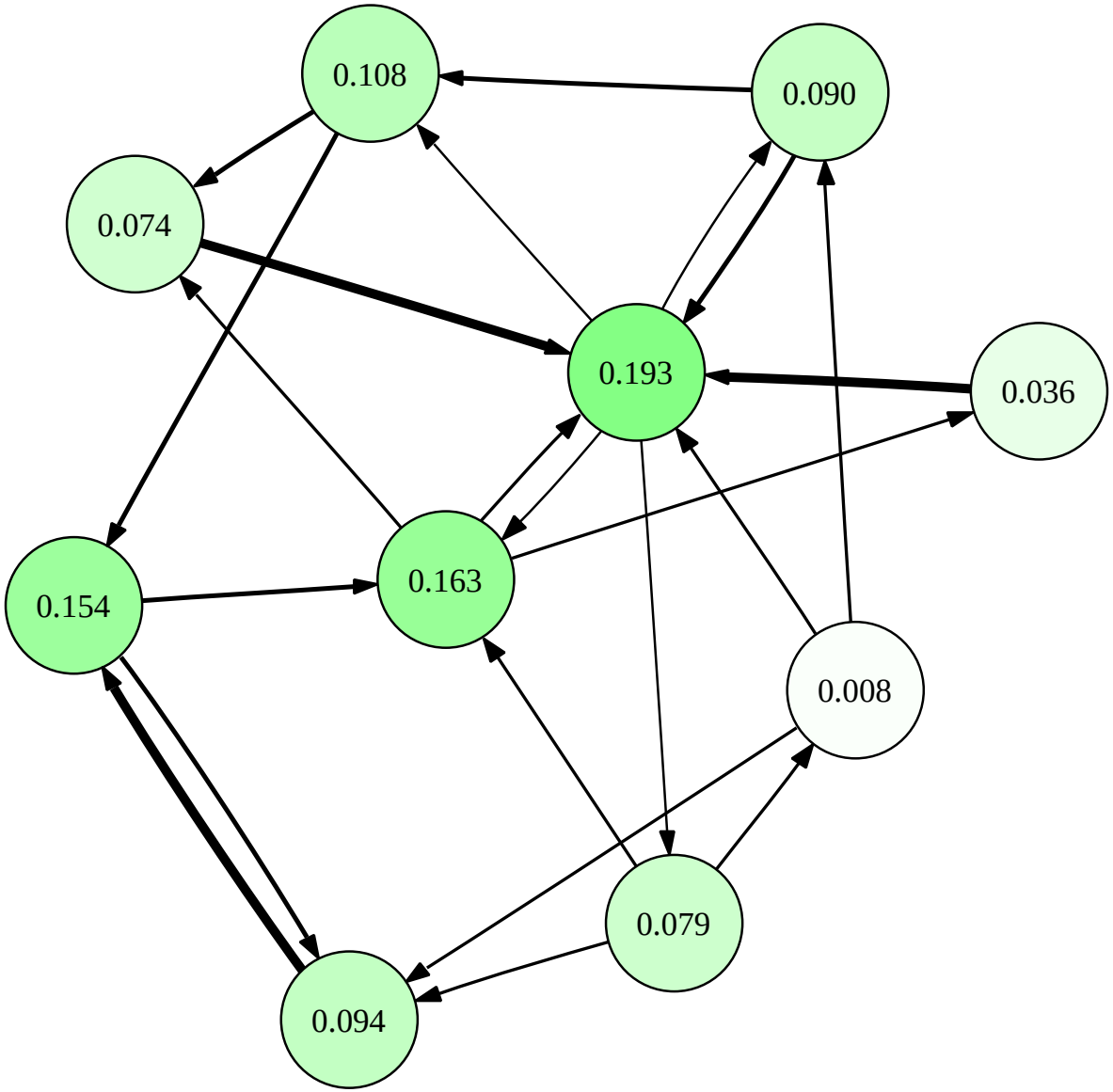
PageRank — Iteration ♯1



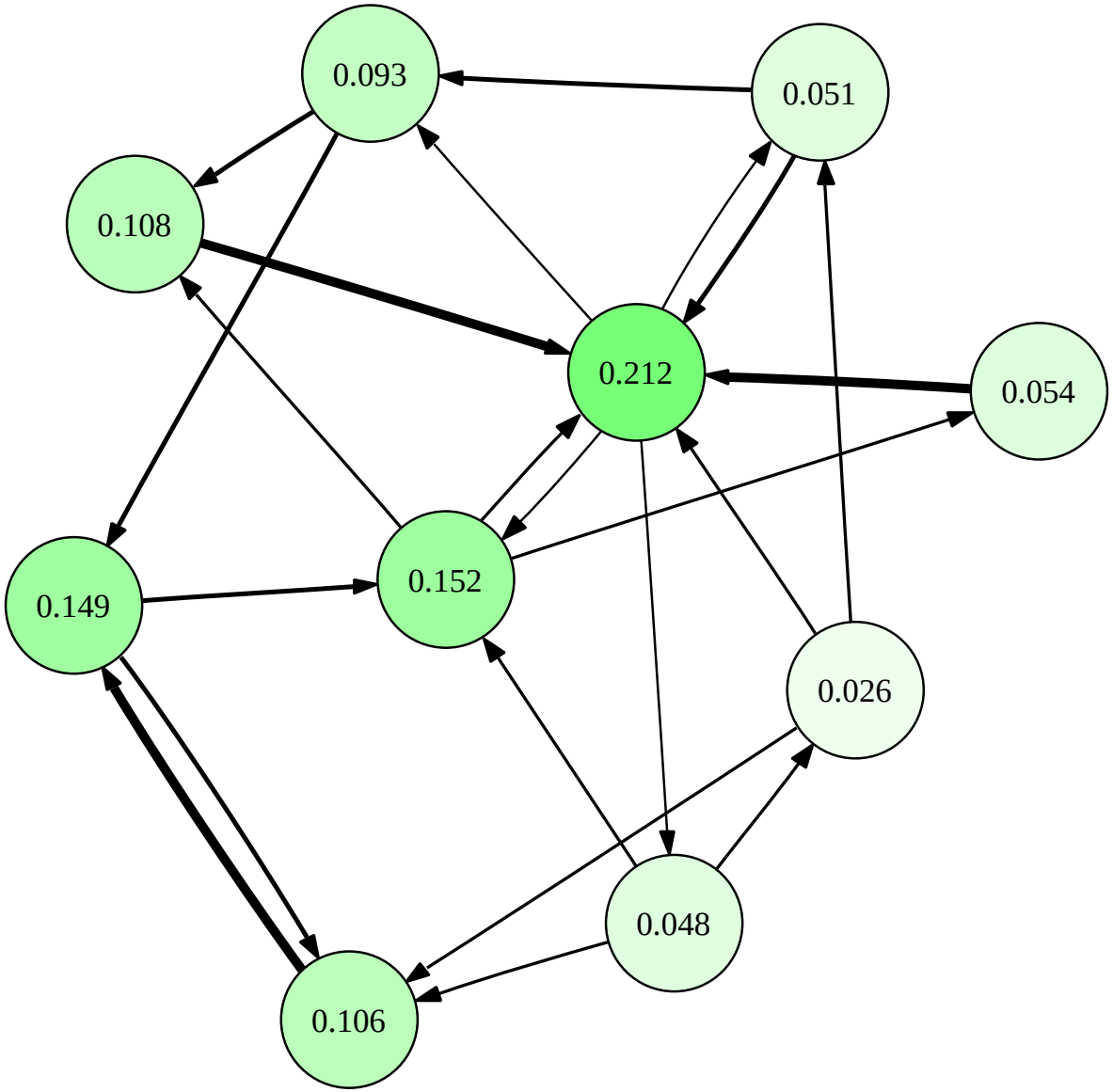
PageRank — Iteration ♯2



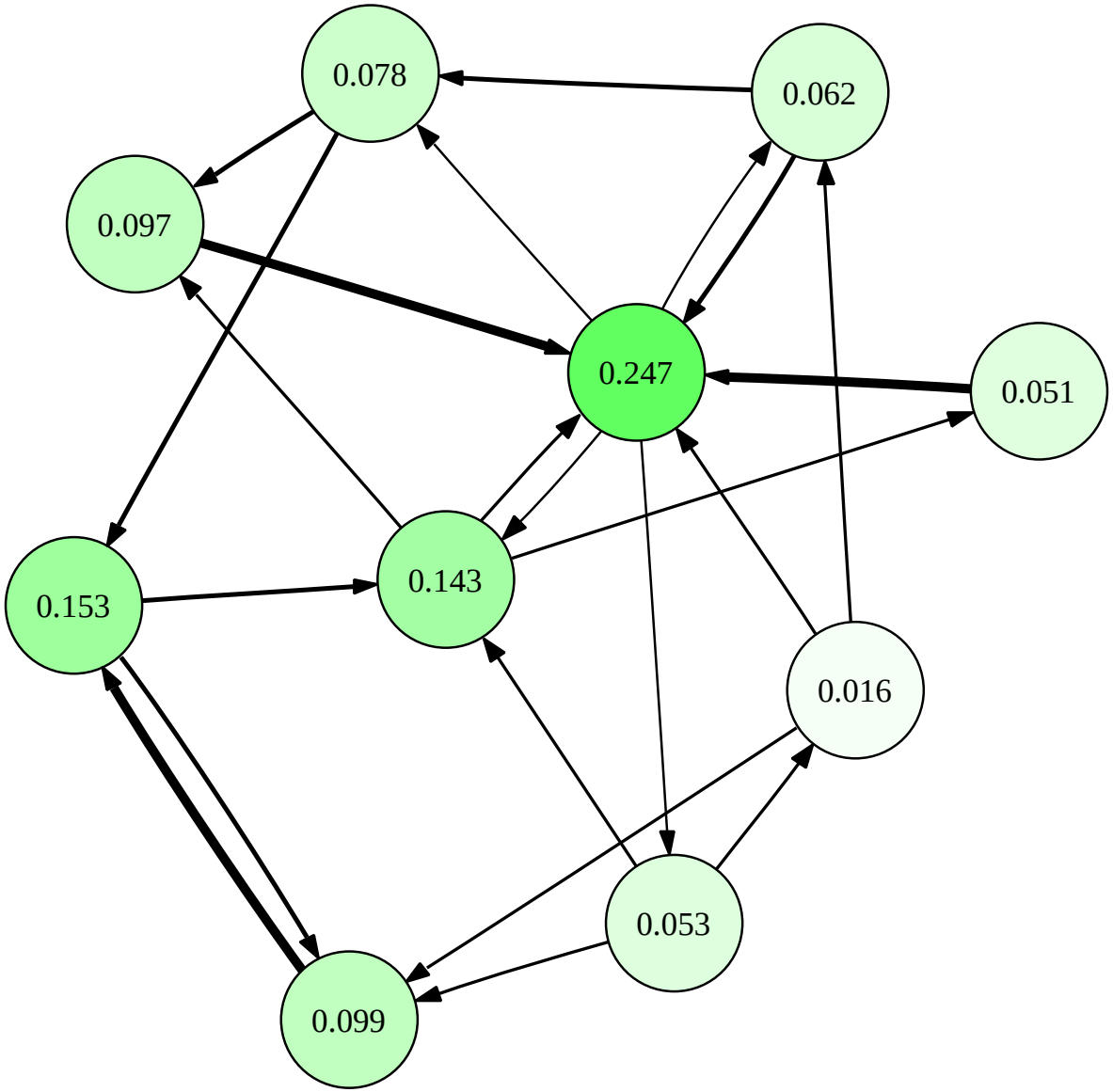
PageRank — Iteration ♯3



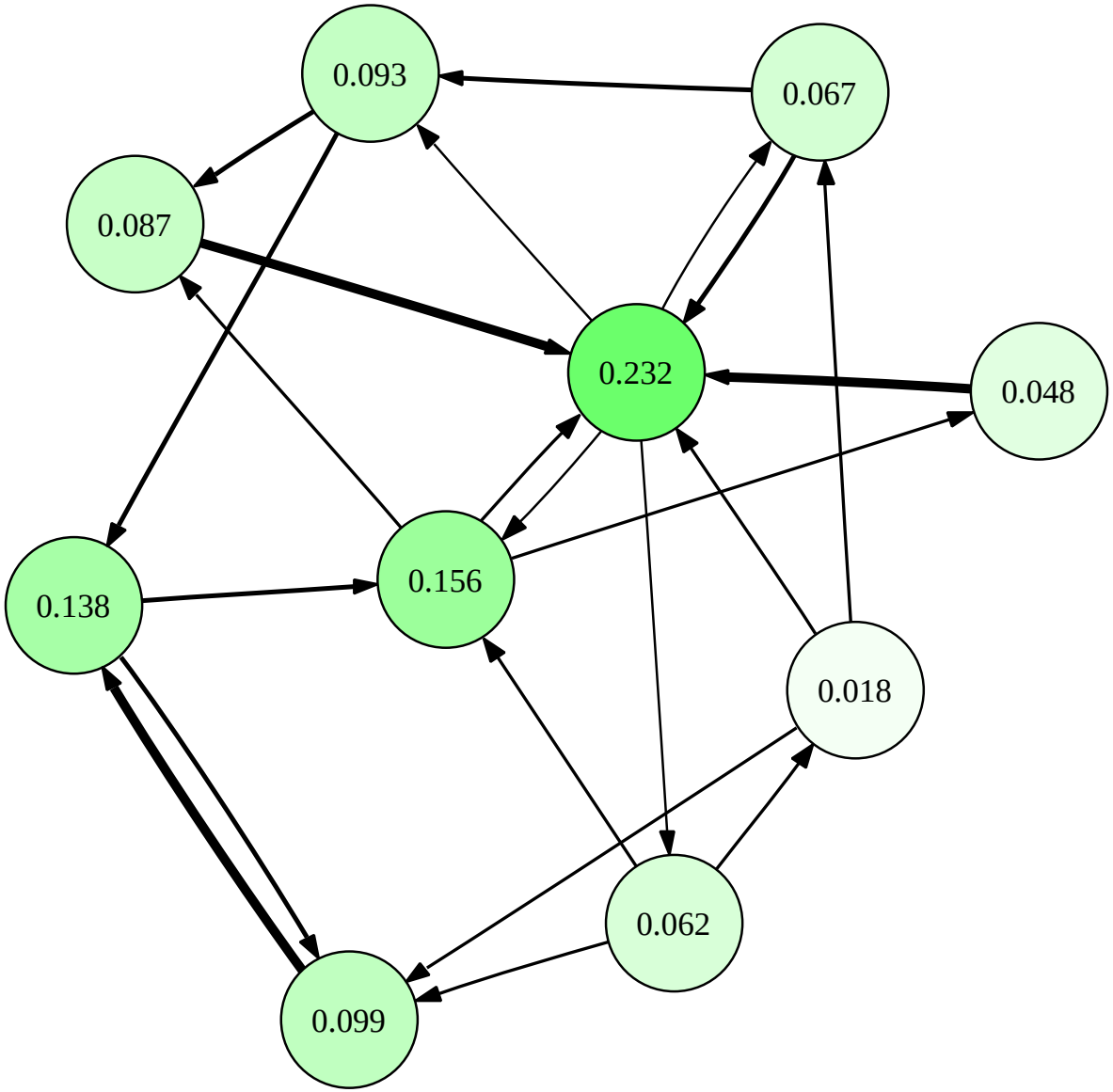
PageRank — Iteration ♯4



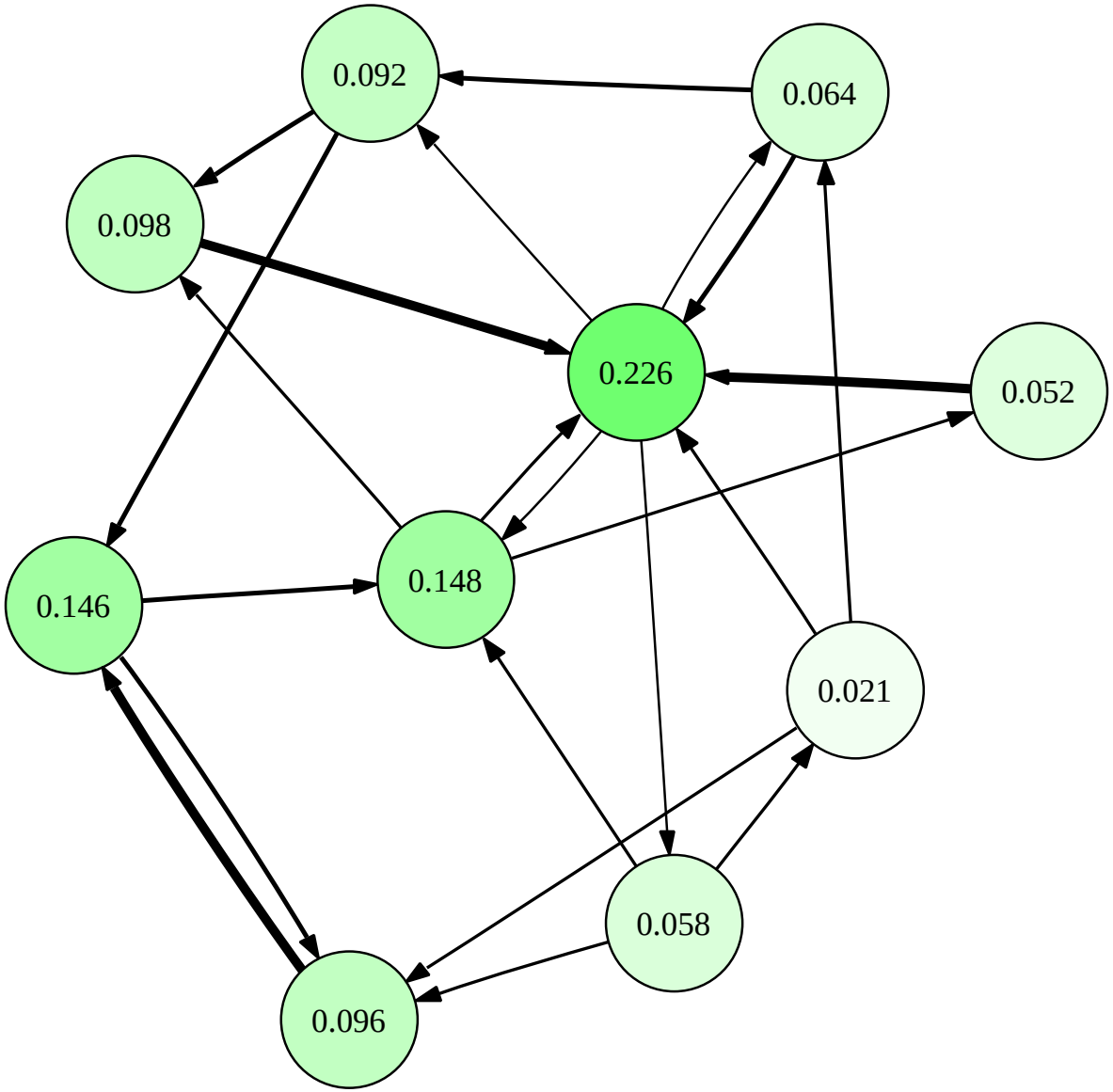
PageRank — Iteration ♯5



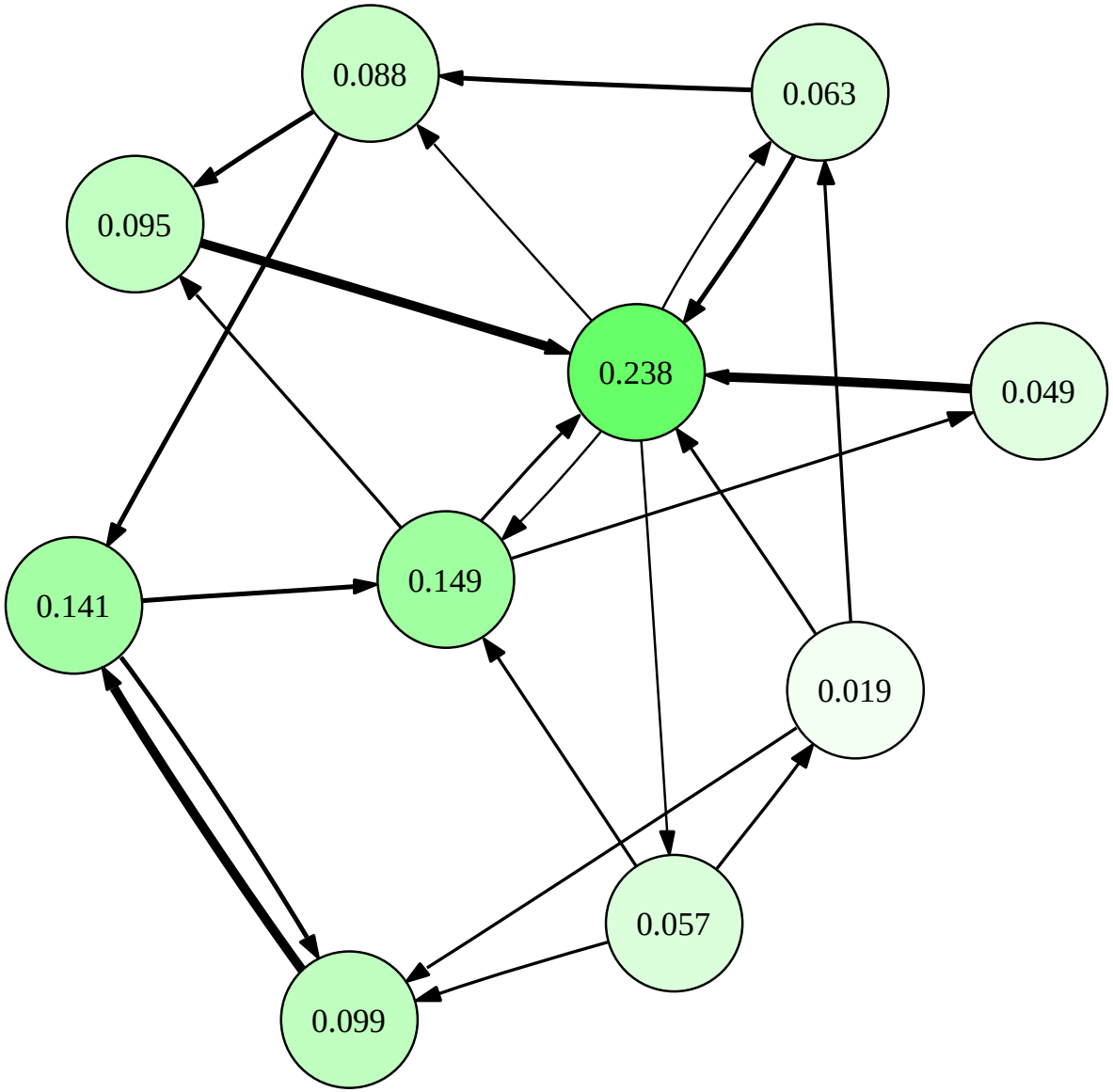
PageRank — Iteration ♯6



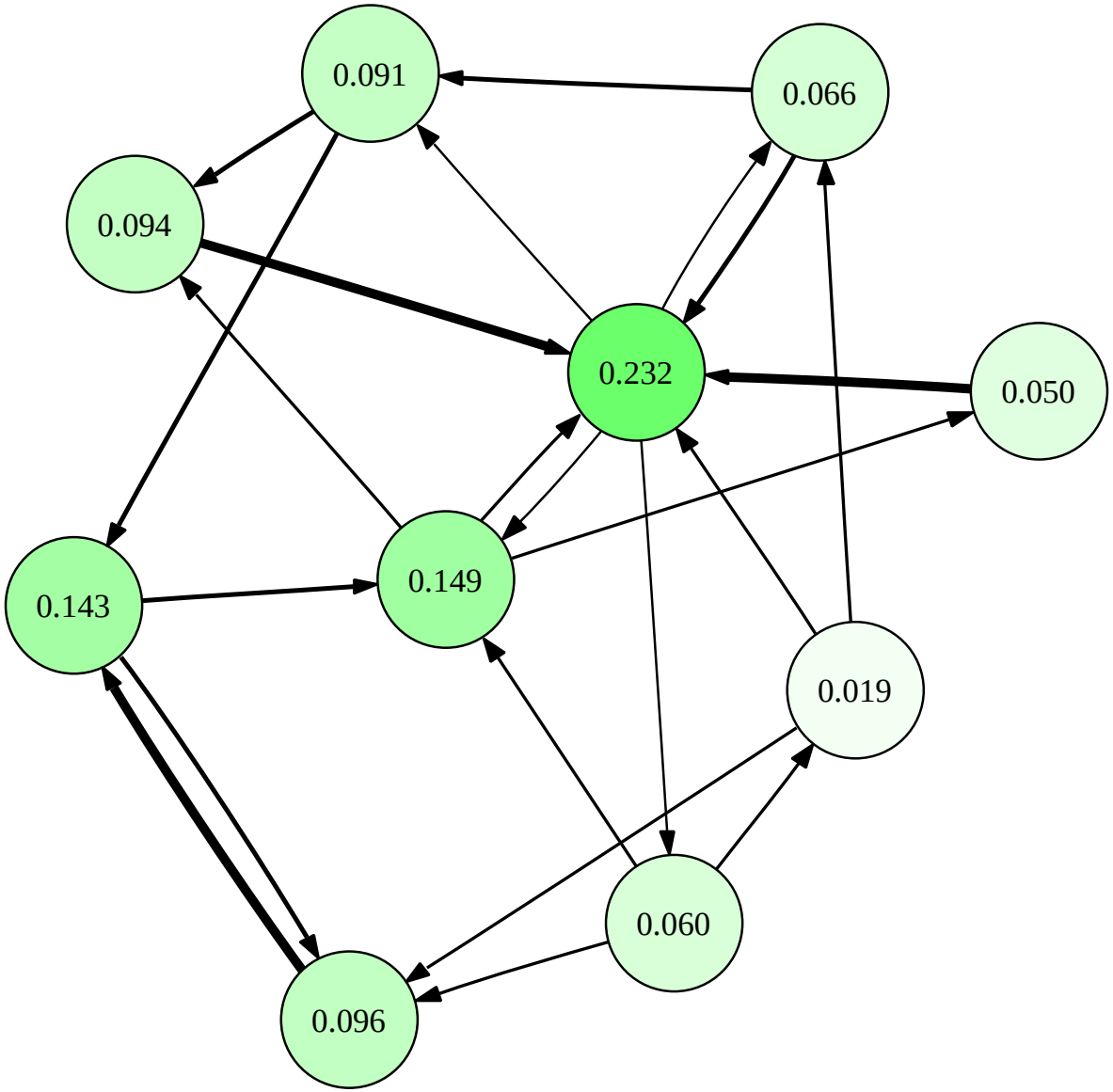
PageRank — Iteration ♯7



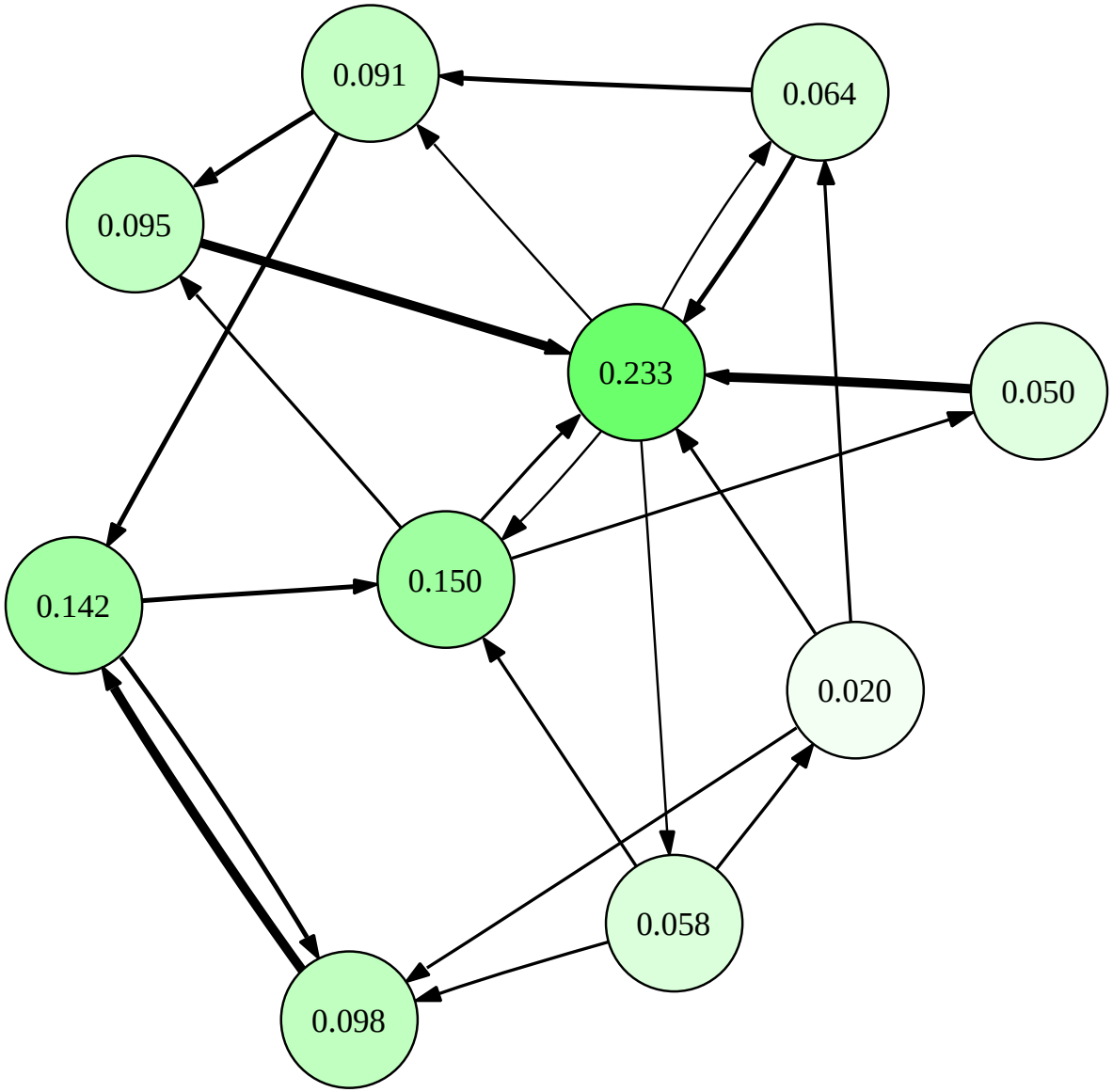
PageRank — Iteration ♯8



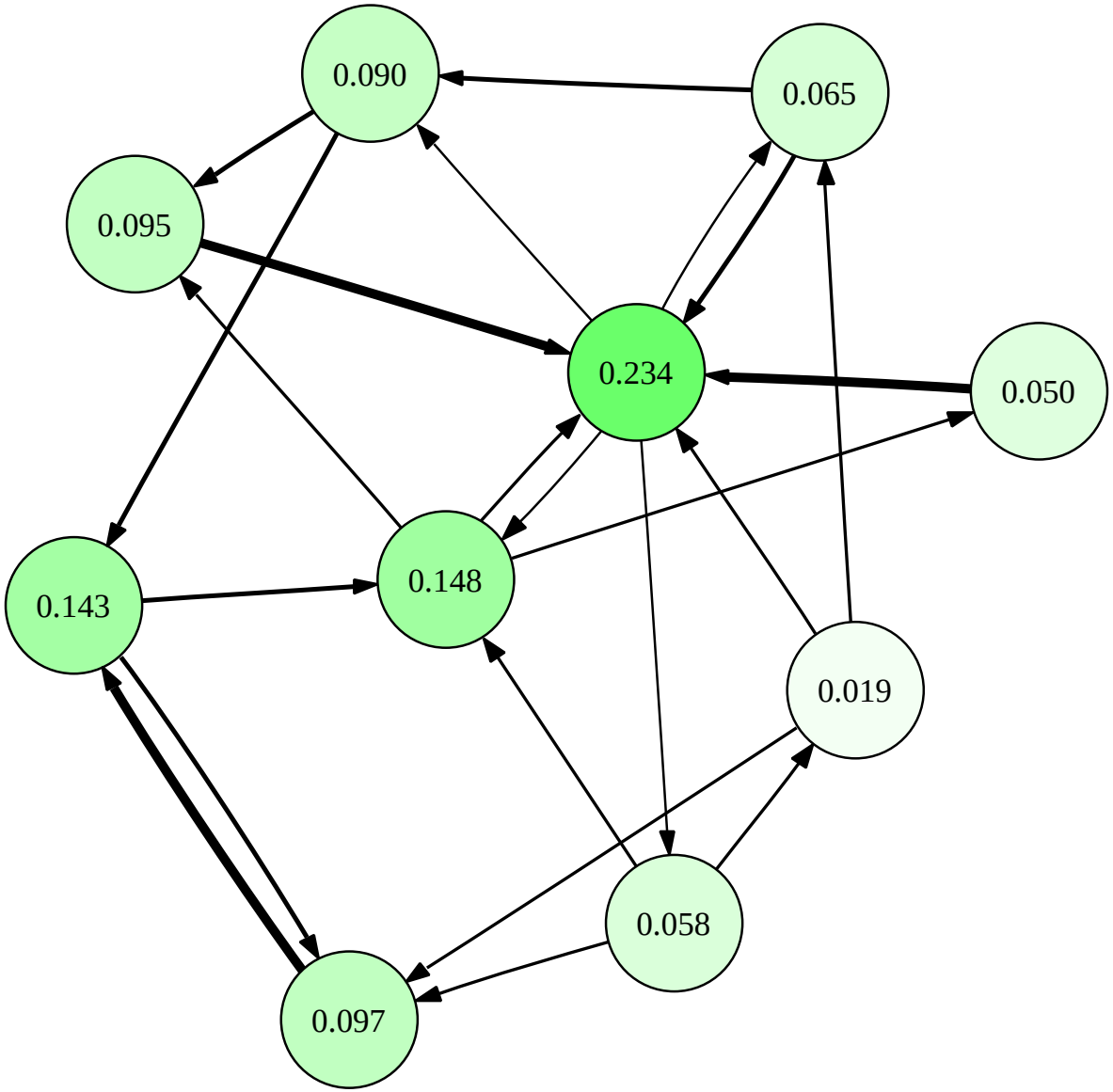
PageRank — Iteration ♯9



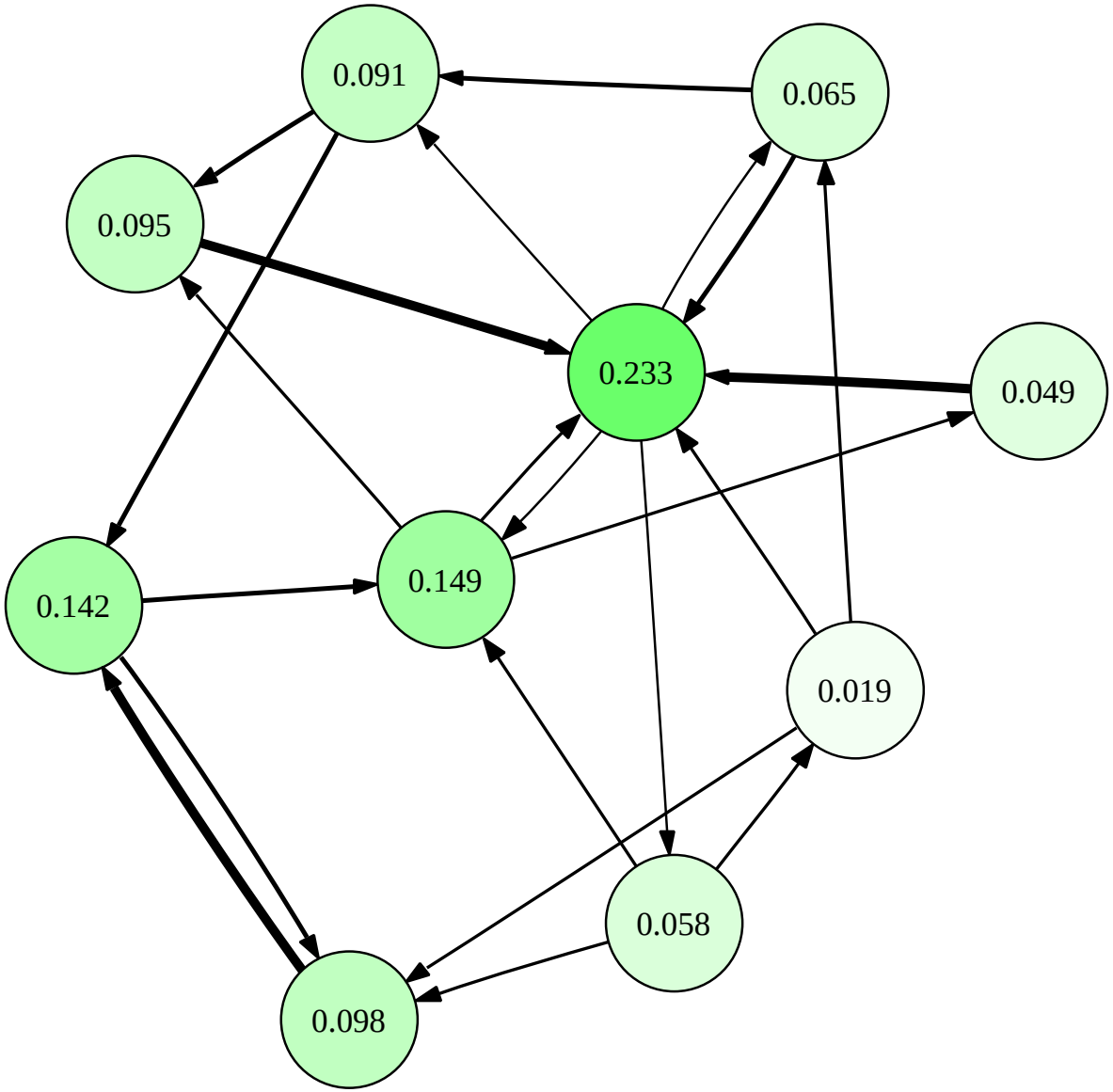
PageRank — Iteration ♯10



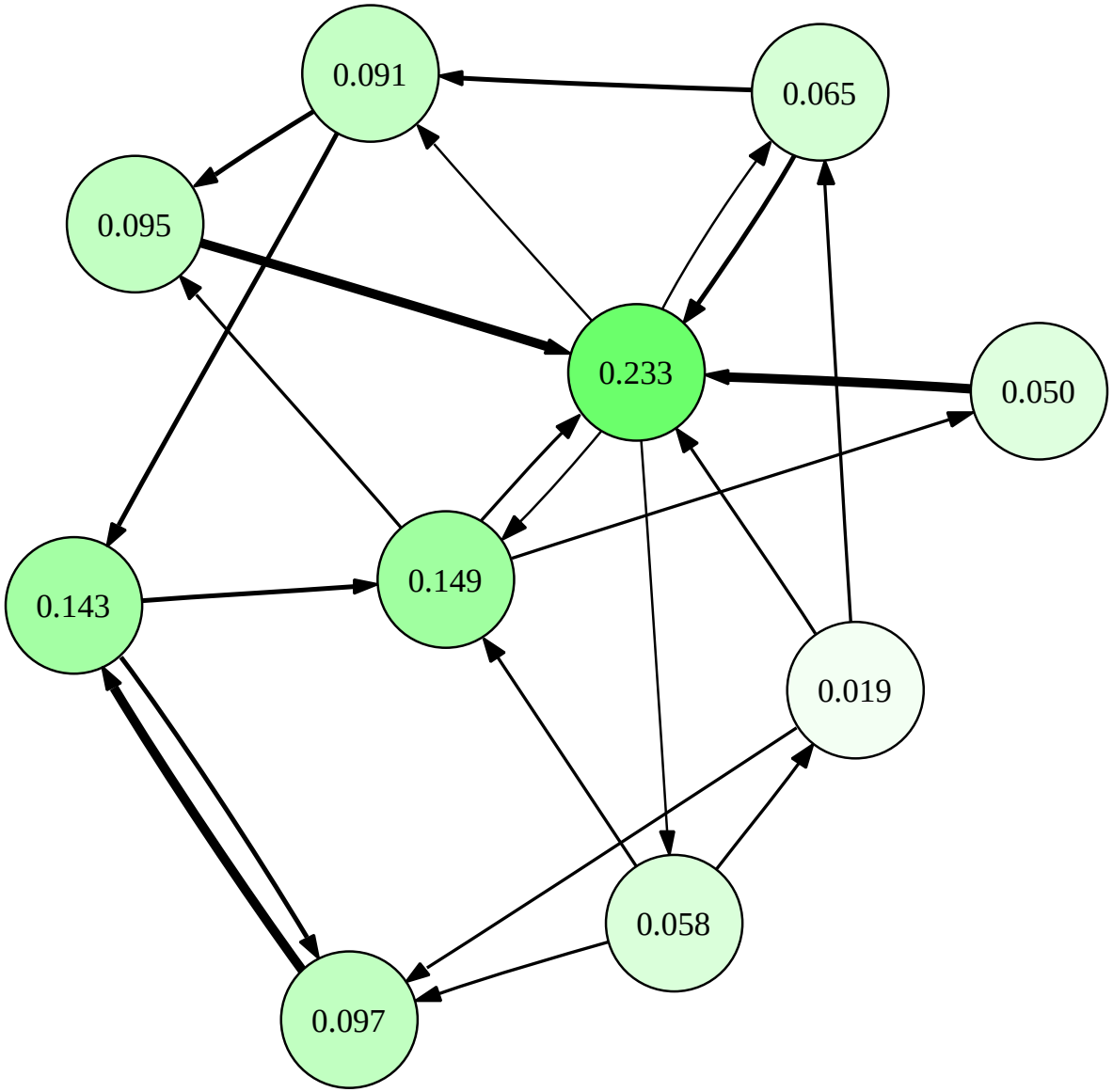
PageRank — Iteration ♯11



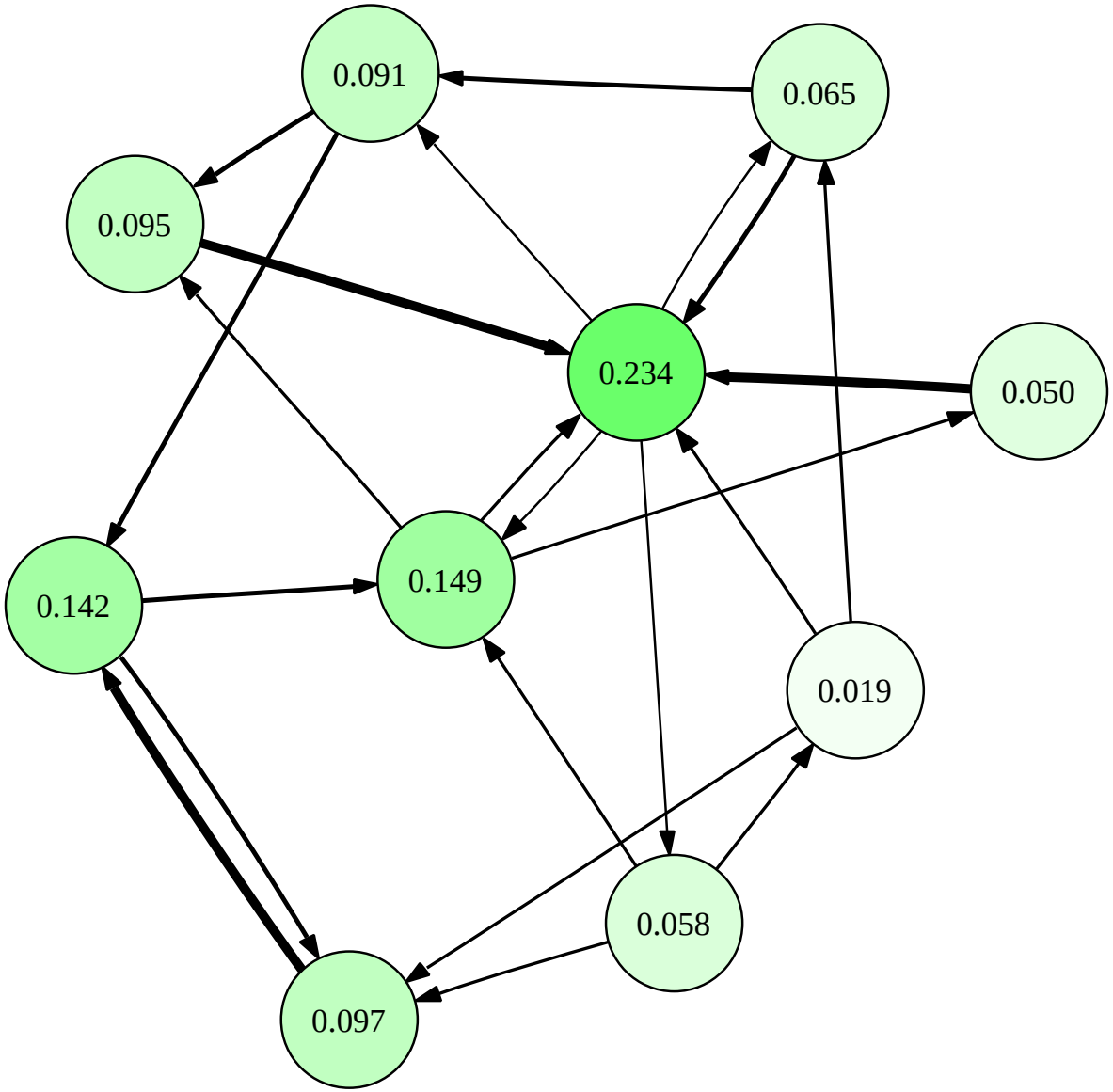
PageRank — Iteration ♯12



PageRank — Iteration ♯13



PageRank — Iteration ♯14



Background on Green measures

Green functions

- Come from electrostatic theory (potential created by a charge distribution).

Background on Green measures

Green functions

- Come from electrostatic theory (potential created by a charge distribution).
- Analogy between electrostatic potential theory and Markov chains.

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:

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

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

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.

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

Purpose

- Finding nodes in the graph related to i.

Purpose

- Finding nodes in the graph related to i.
- For each method, output an ordered list of nodes related 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.

Green — Method Description

Method Description

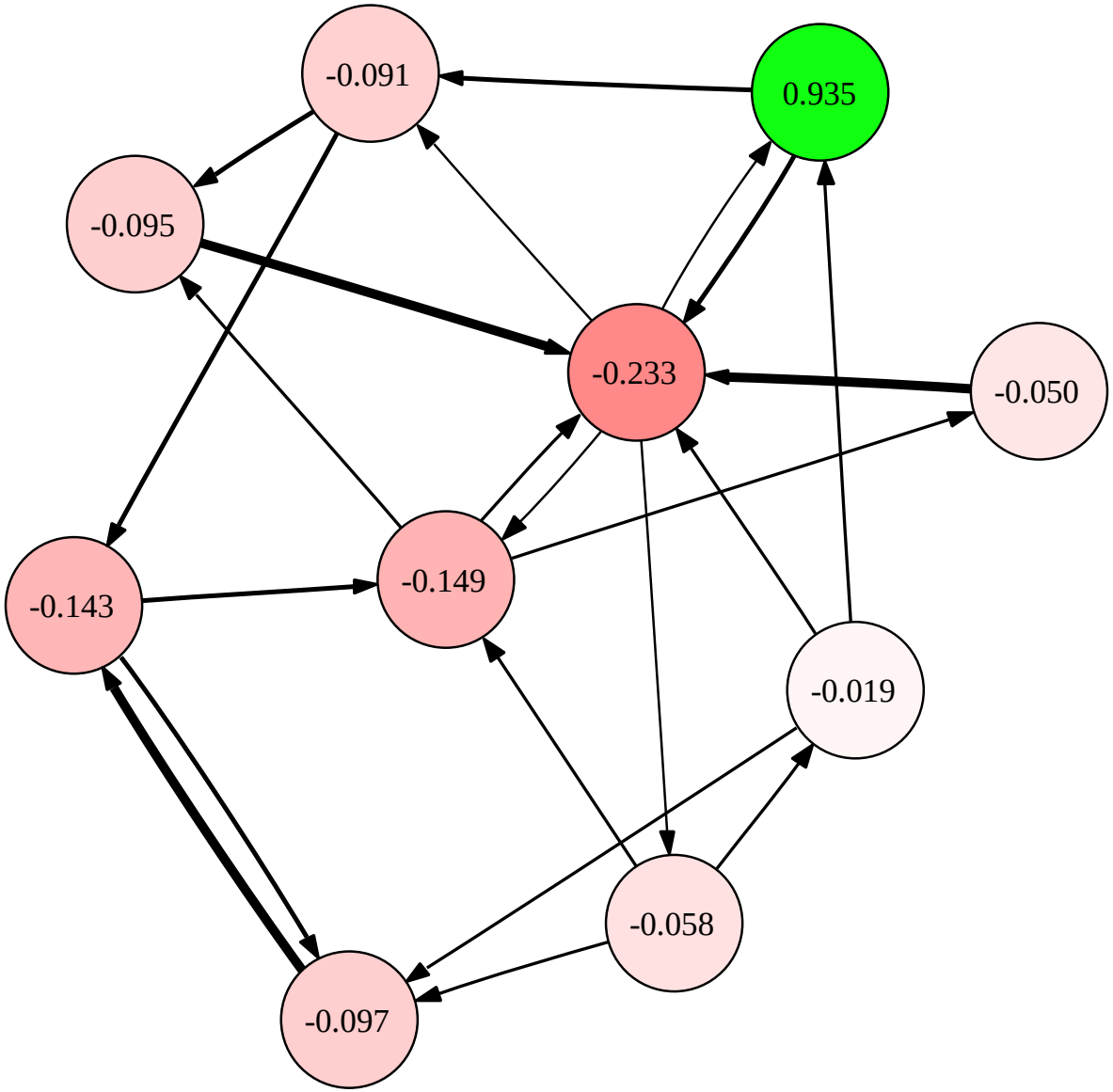
- Direct application of the theory of Green measures.
- Improvement: multiplication by a term favoring uncommon nodes $\log(1/\nu_j)$ (quantity of information).

Green — Method Description

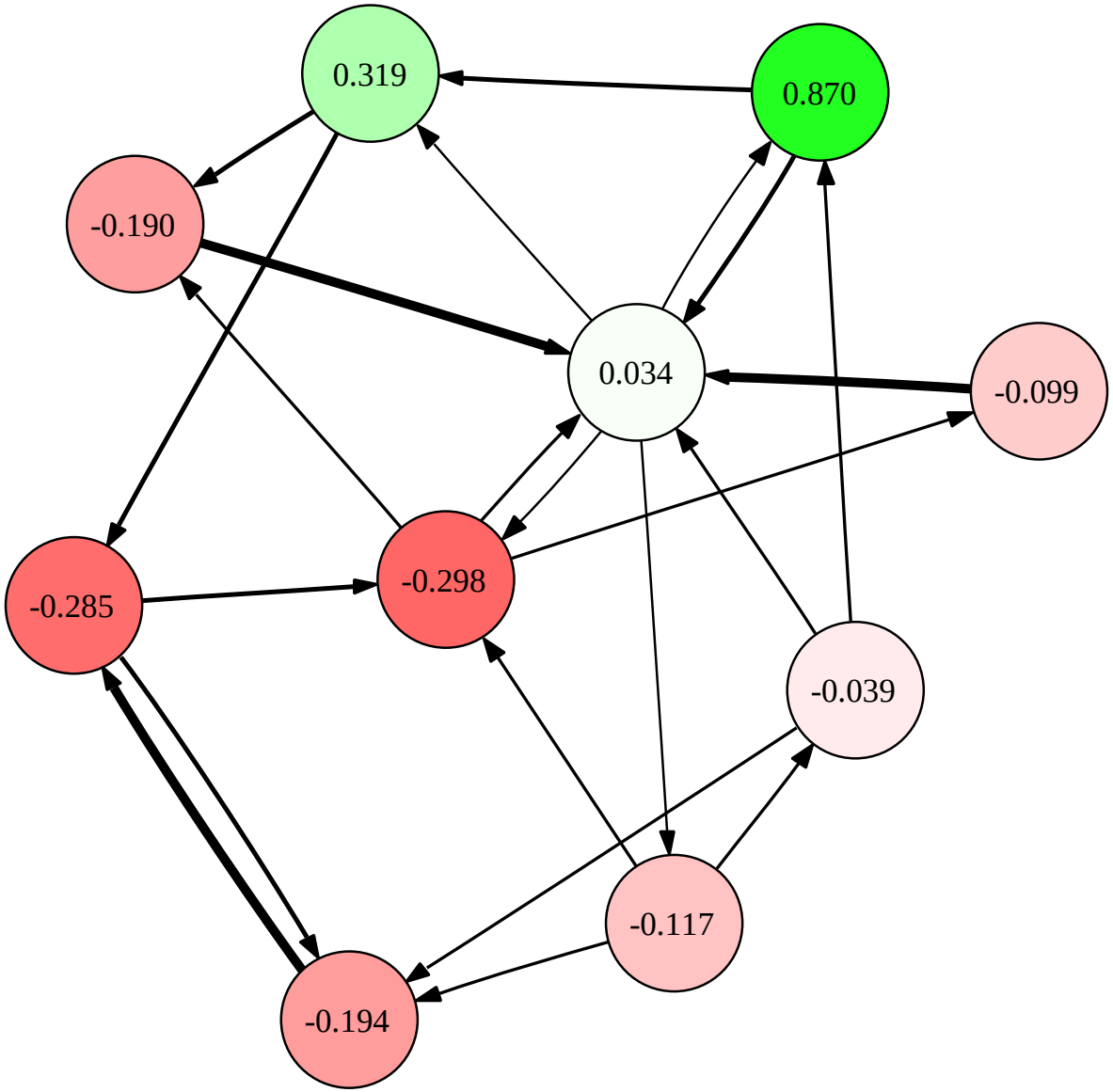
Method Description

- 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.

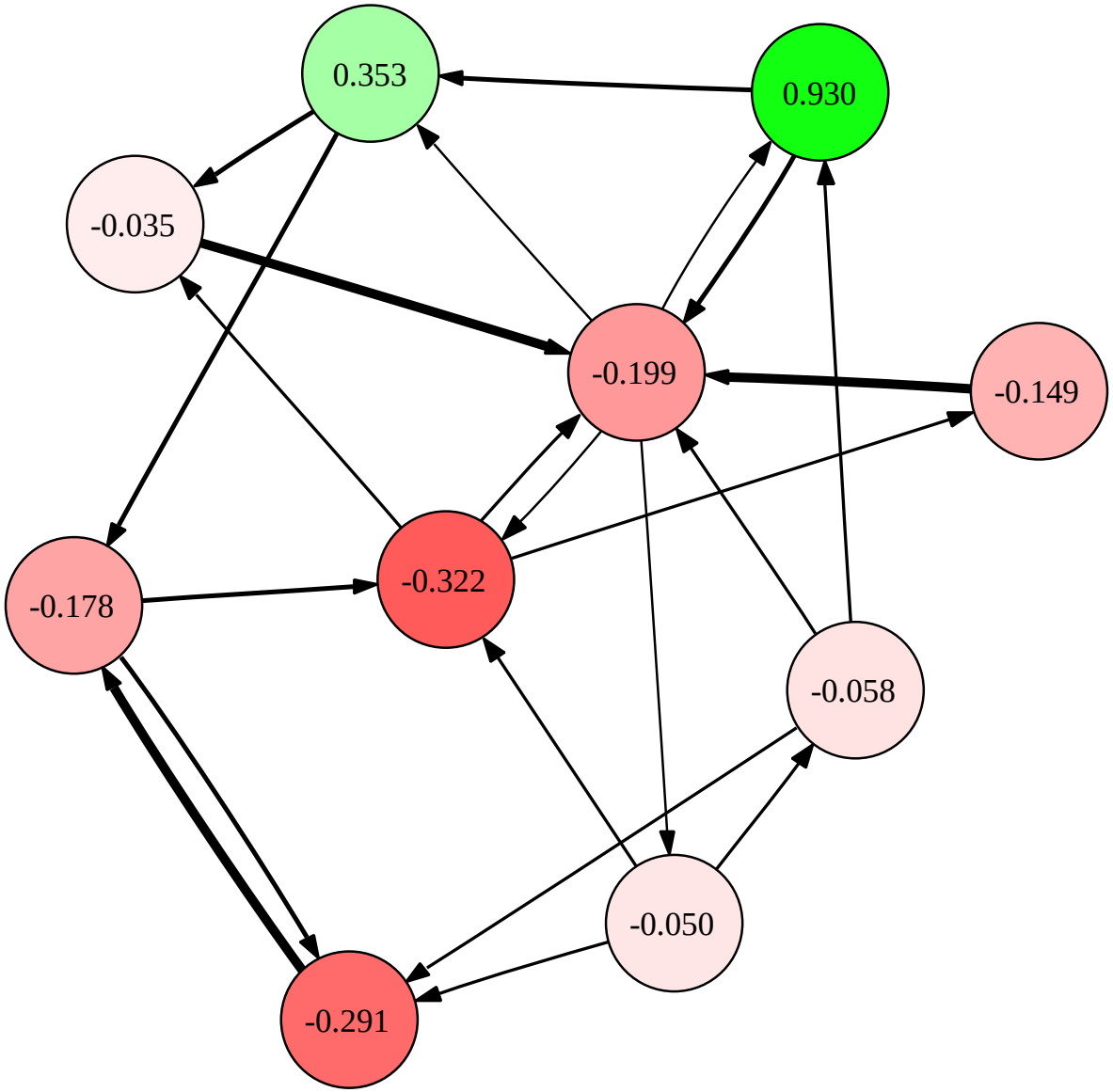
Green — Iteration ♯1



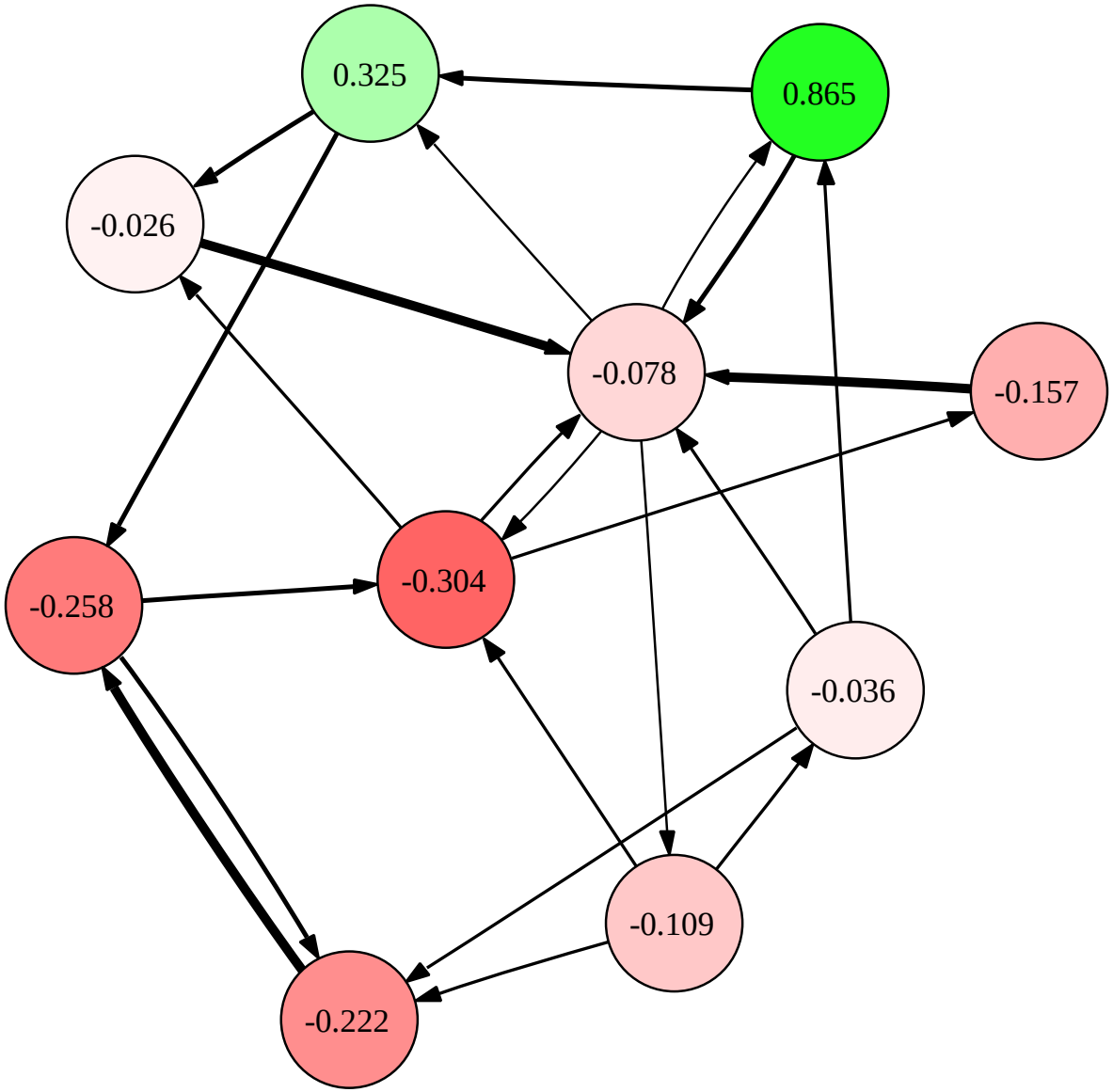
Green — Iteration ♯2



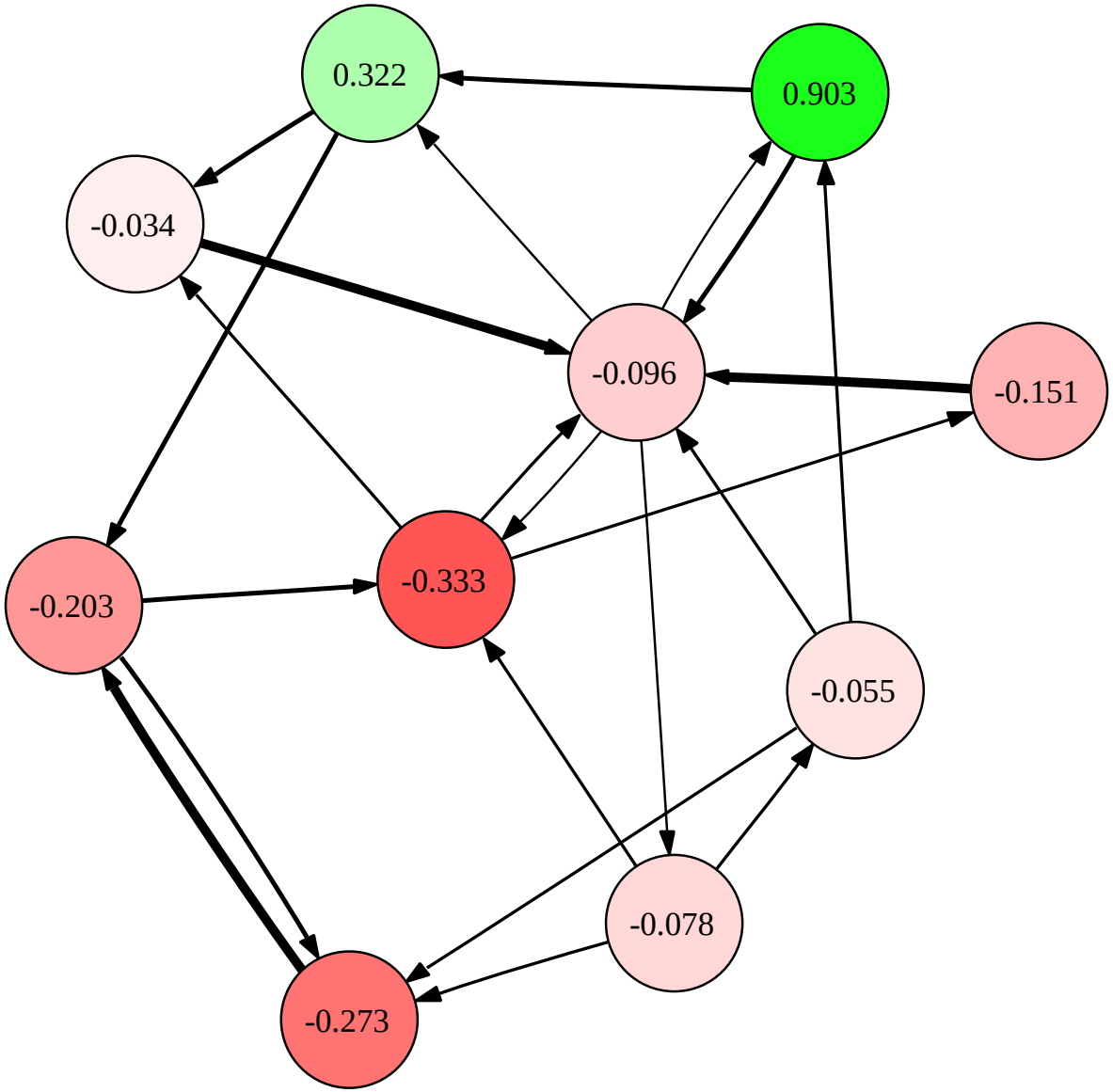
Green — Iteration ♯3



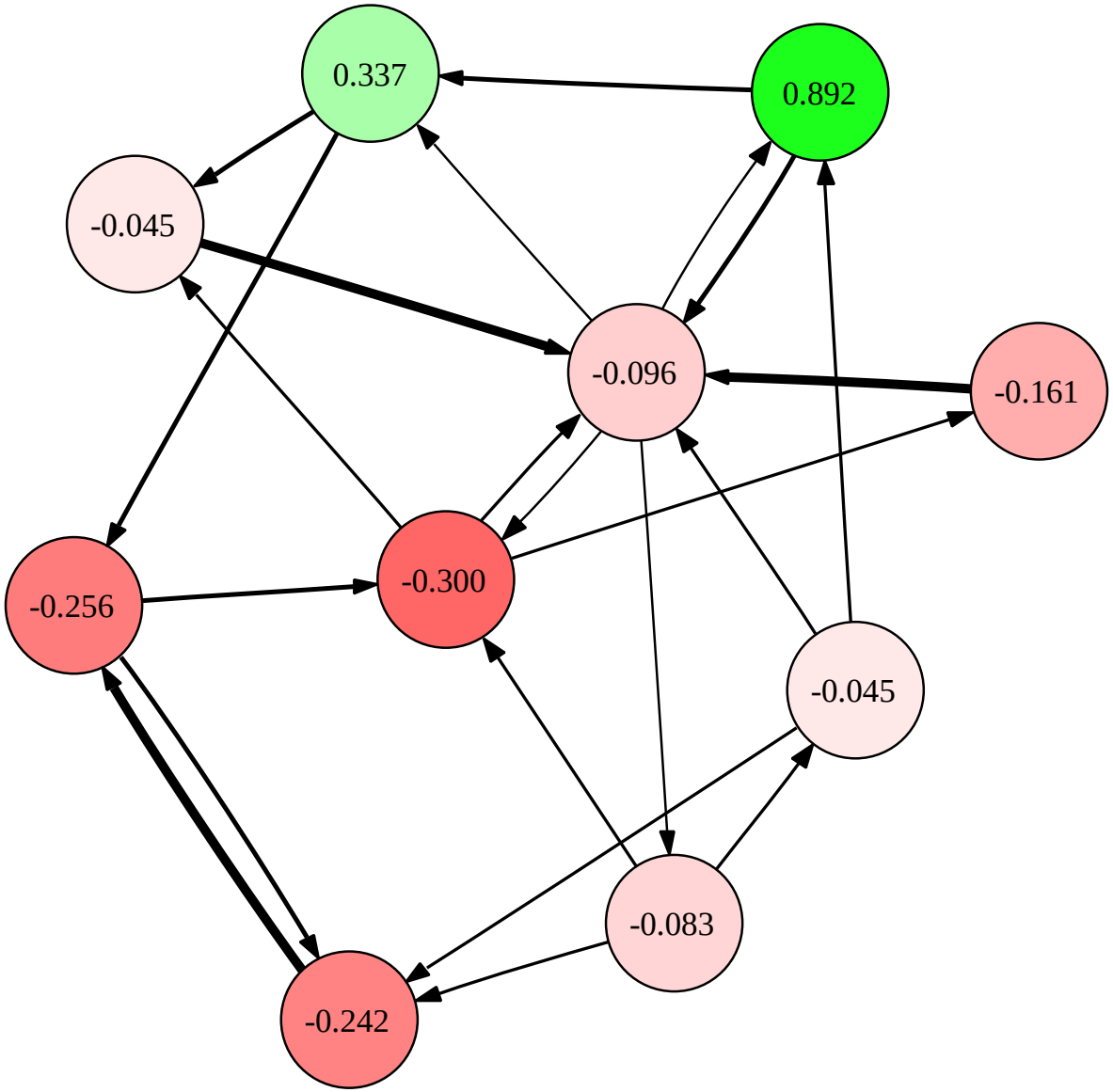
Green — Iteration ♯4



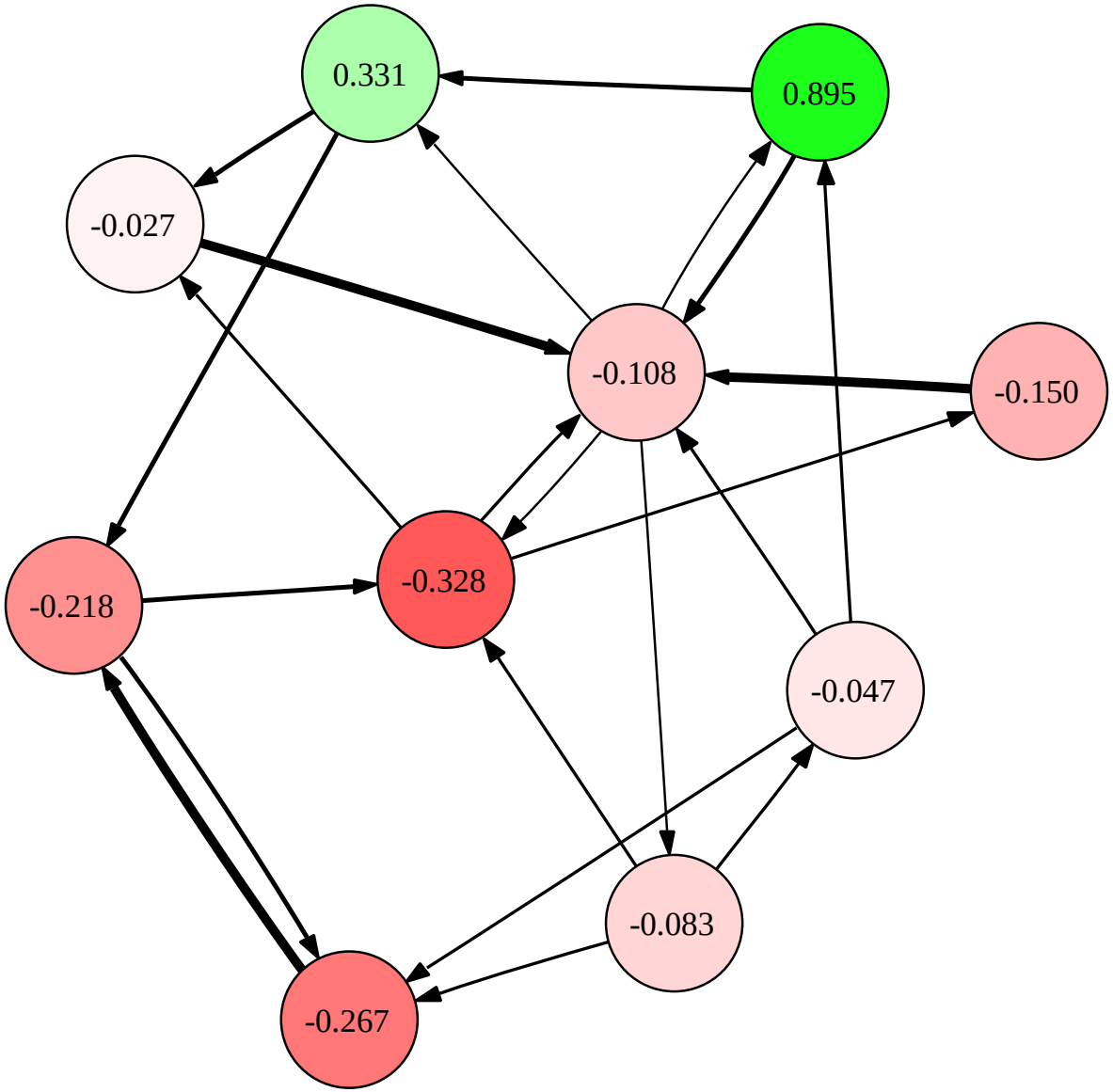
Green — Iteration ♯5



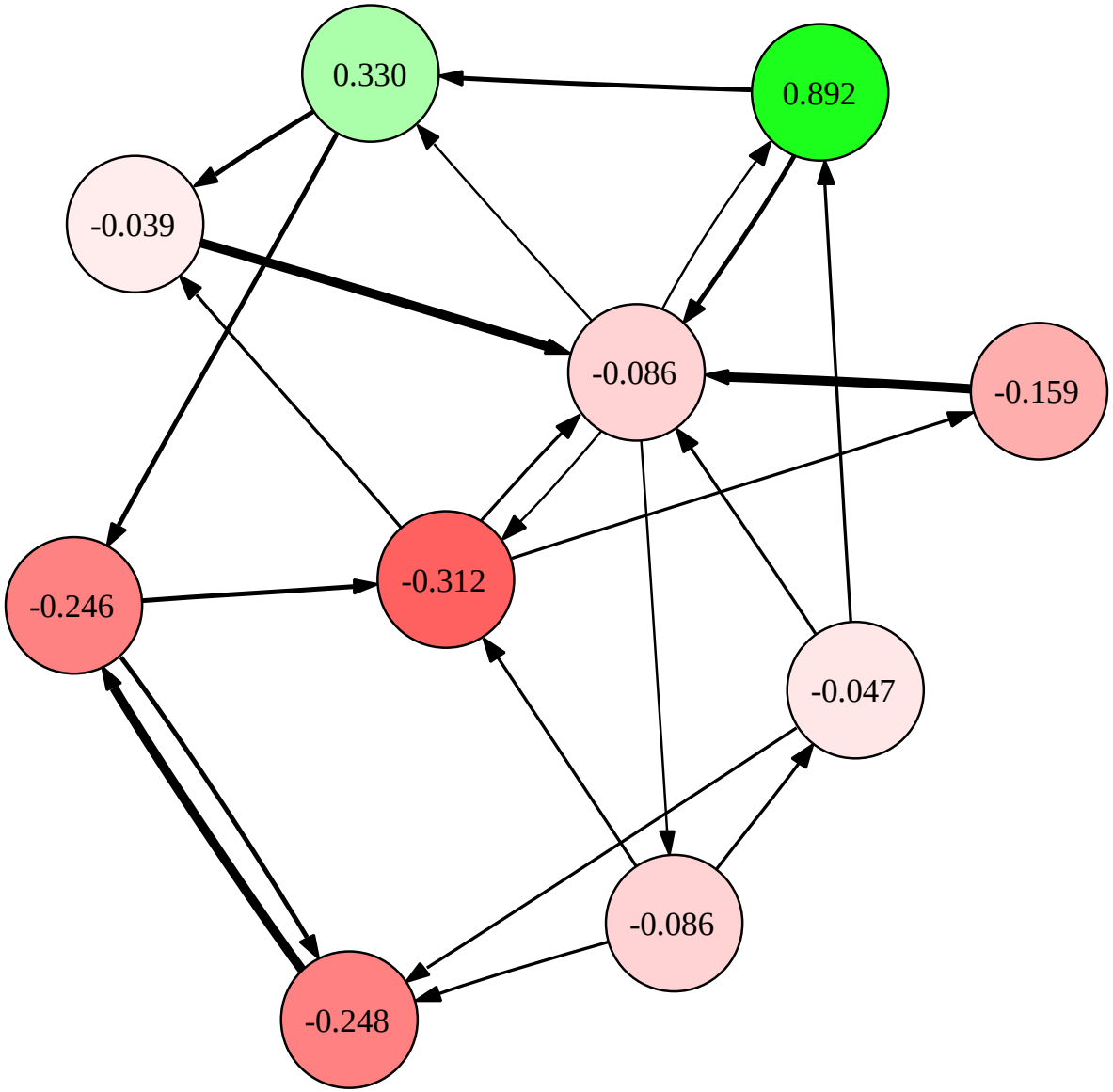
Green — Iteration ♯6



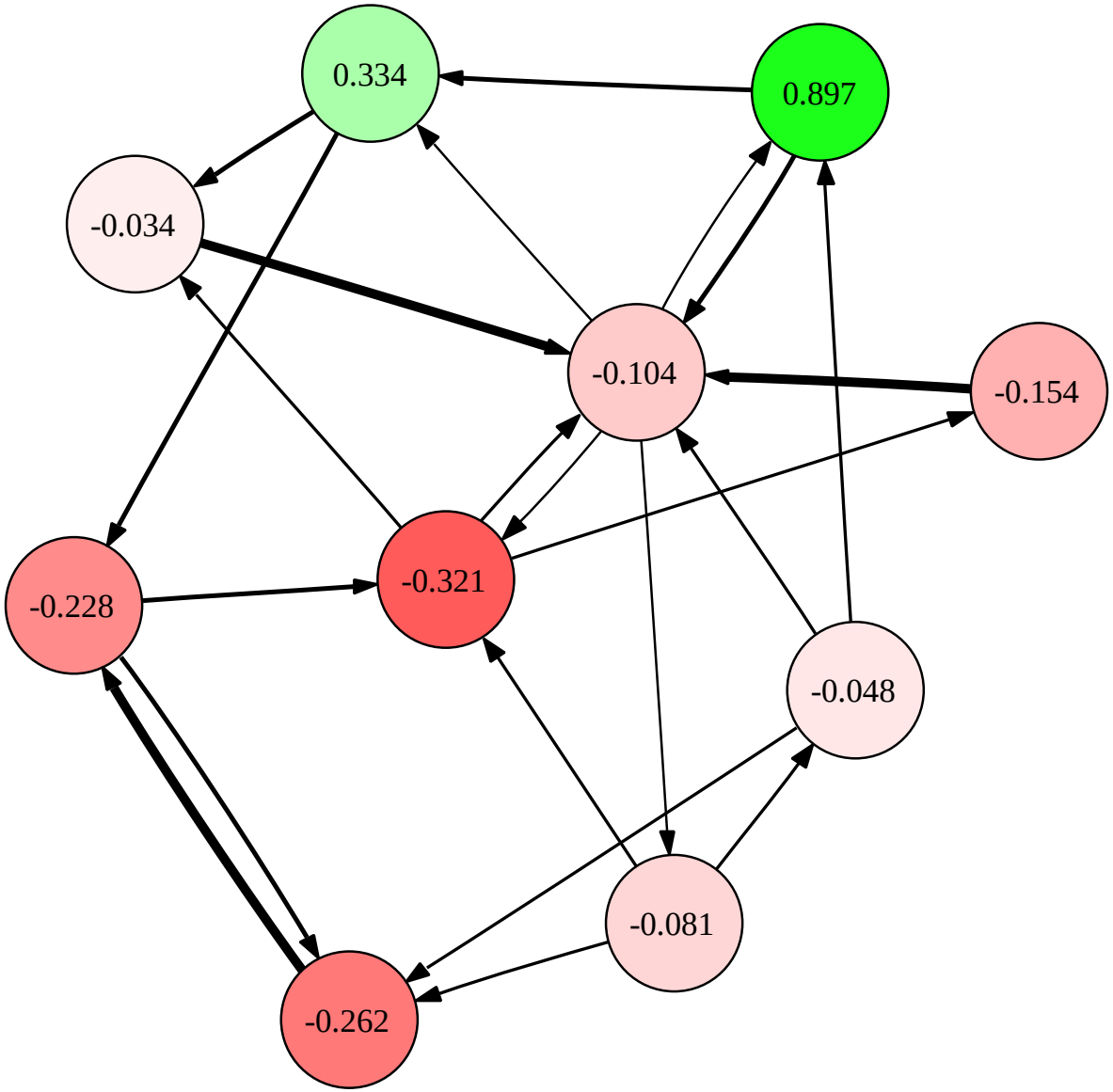
Green — Iteration ♯7



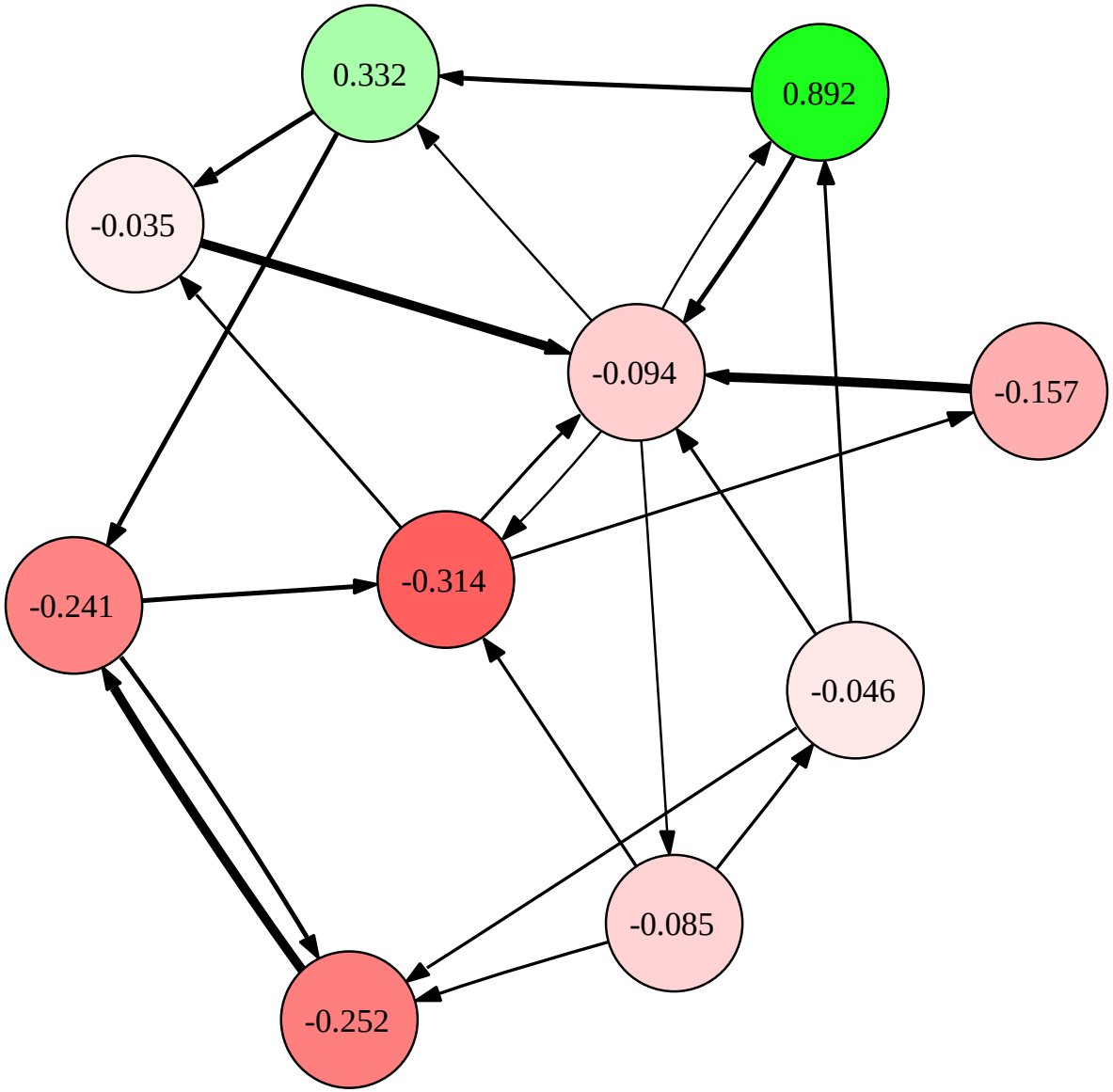
Green — Iteration ♯8



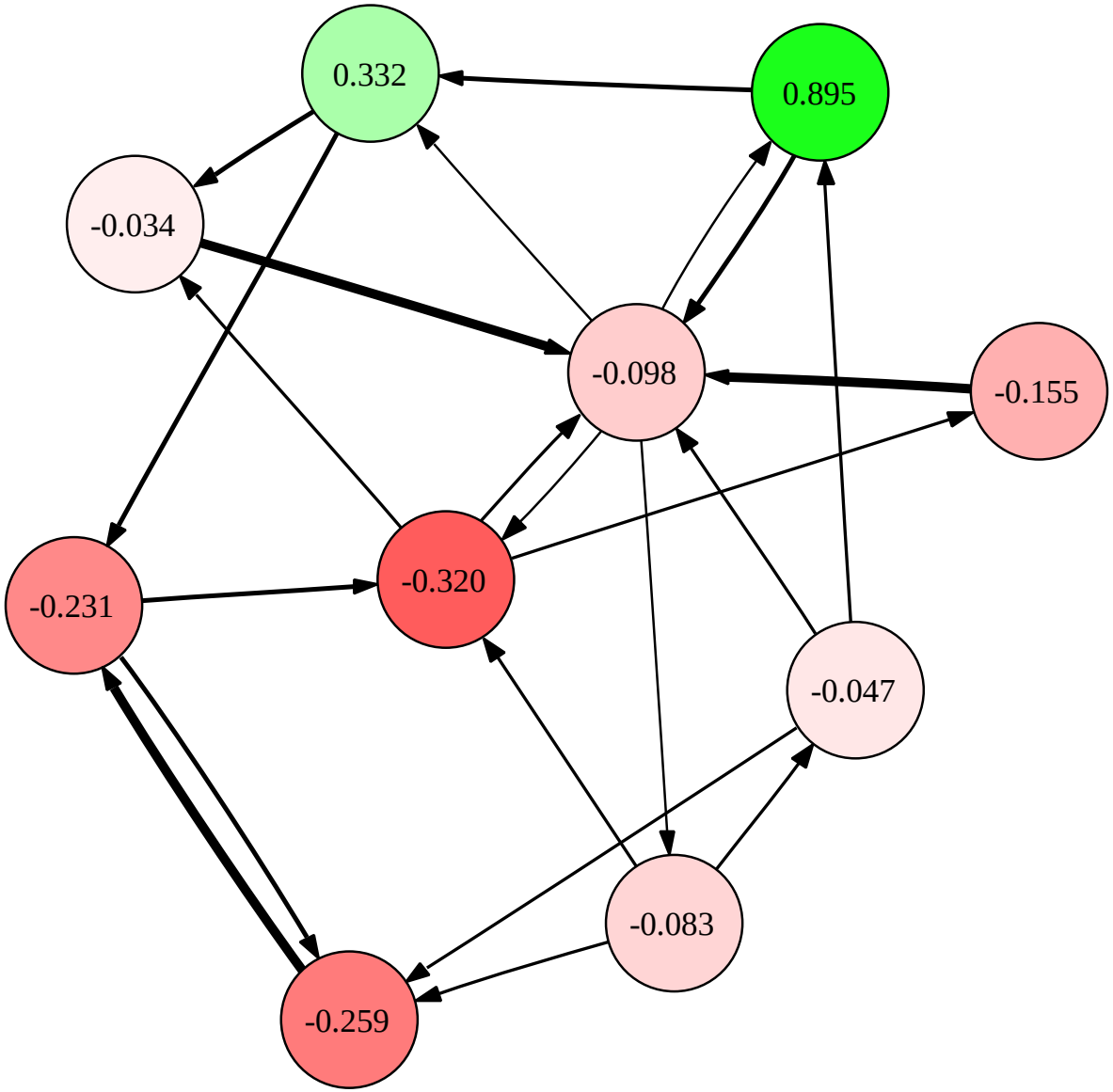
Green — Iteration ♯9



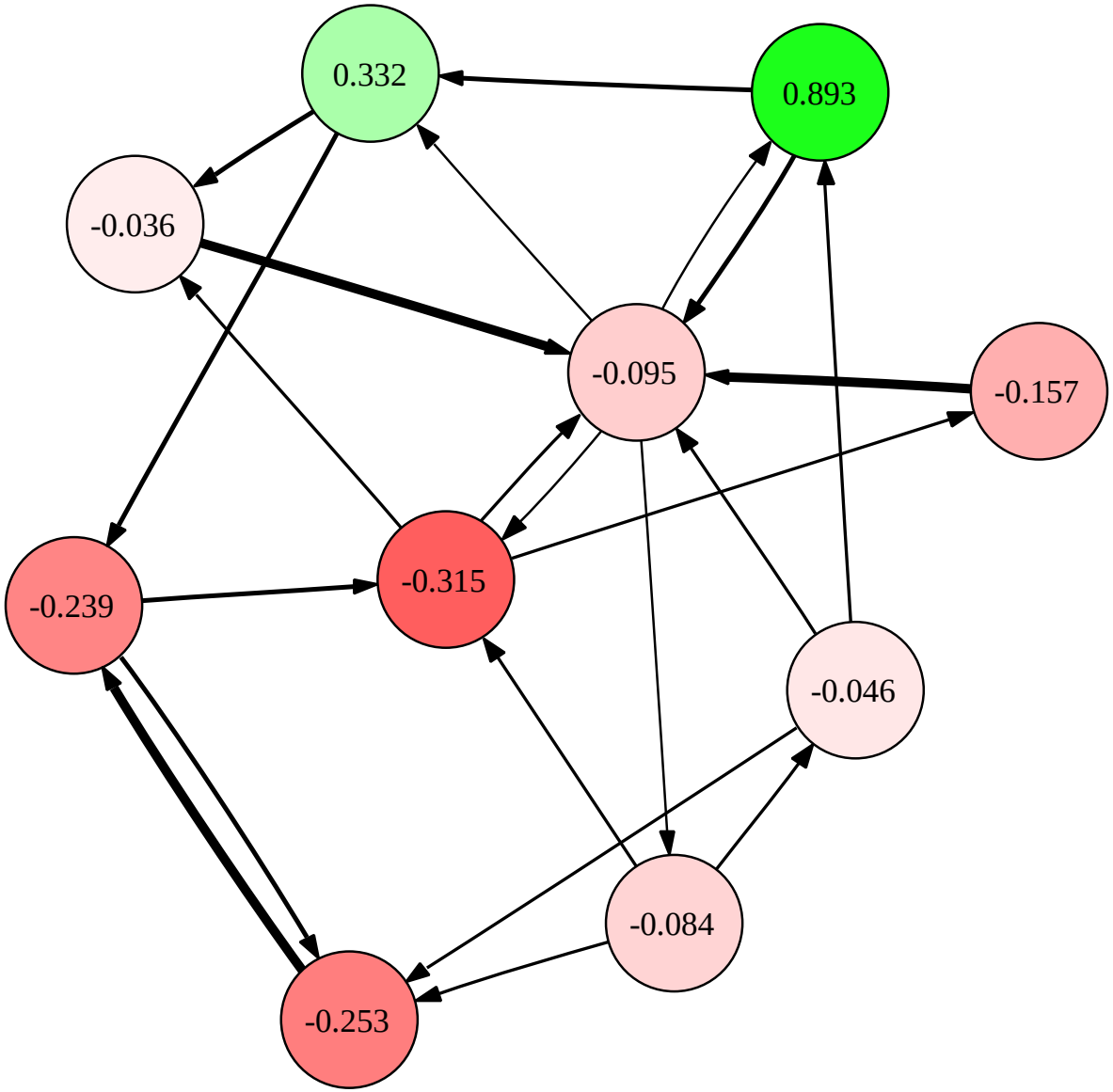
Green — Iteration ♯10



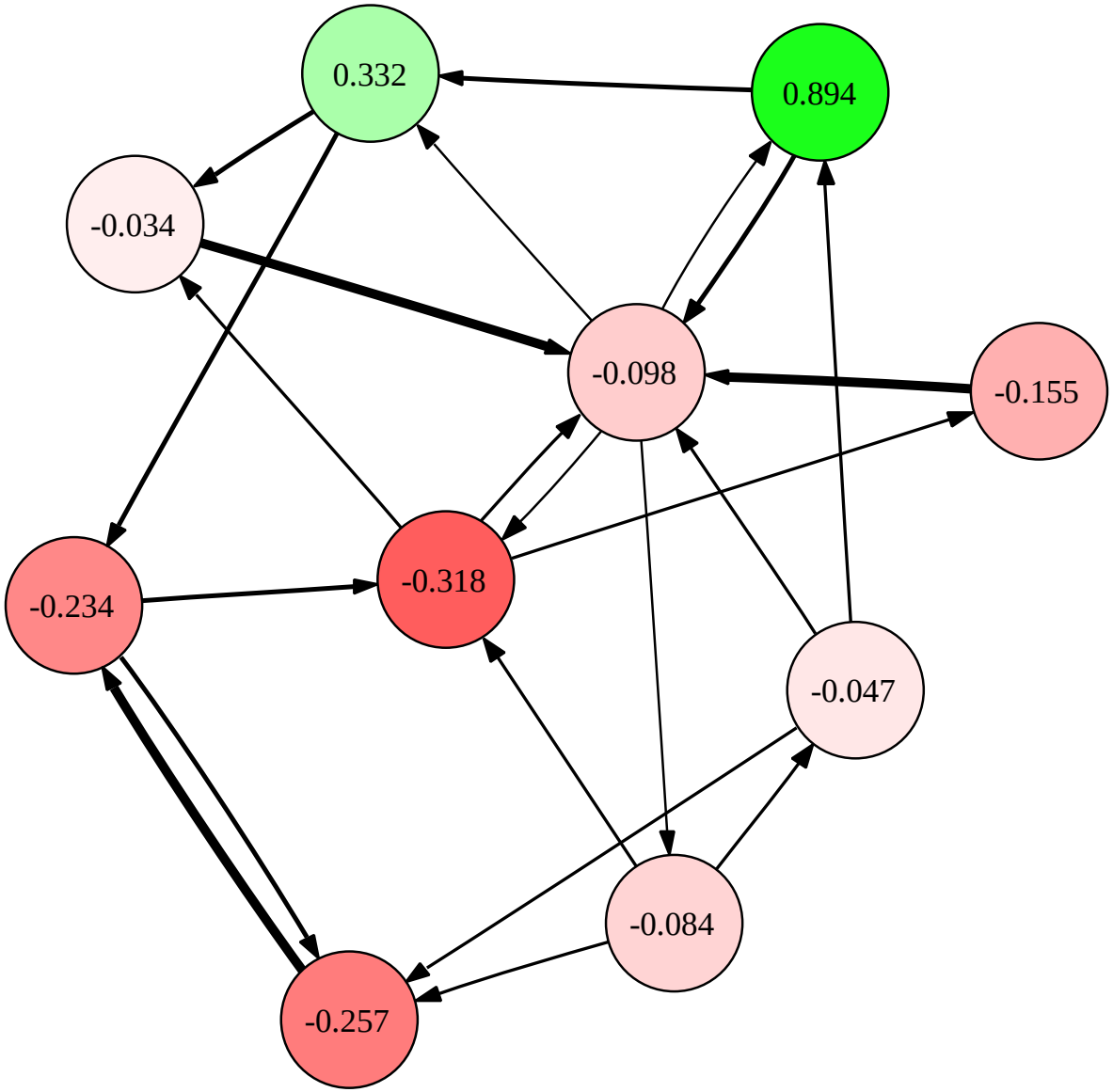
Green — Iteration ♯11



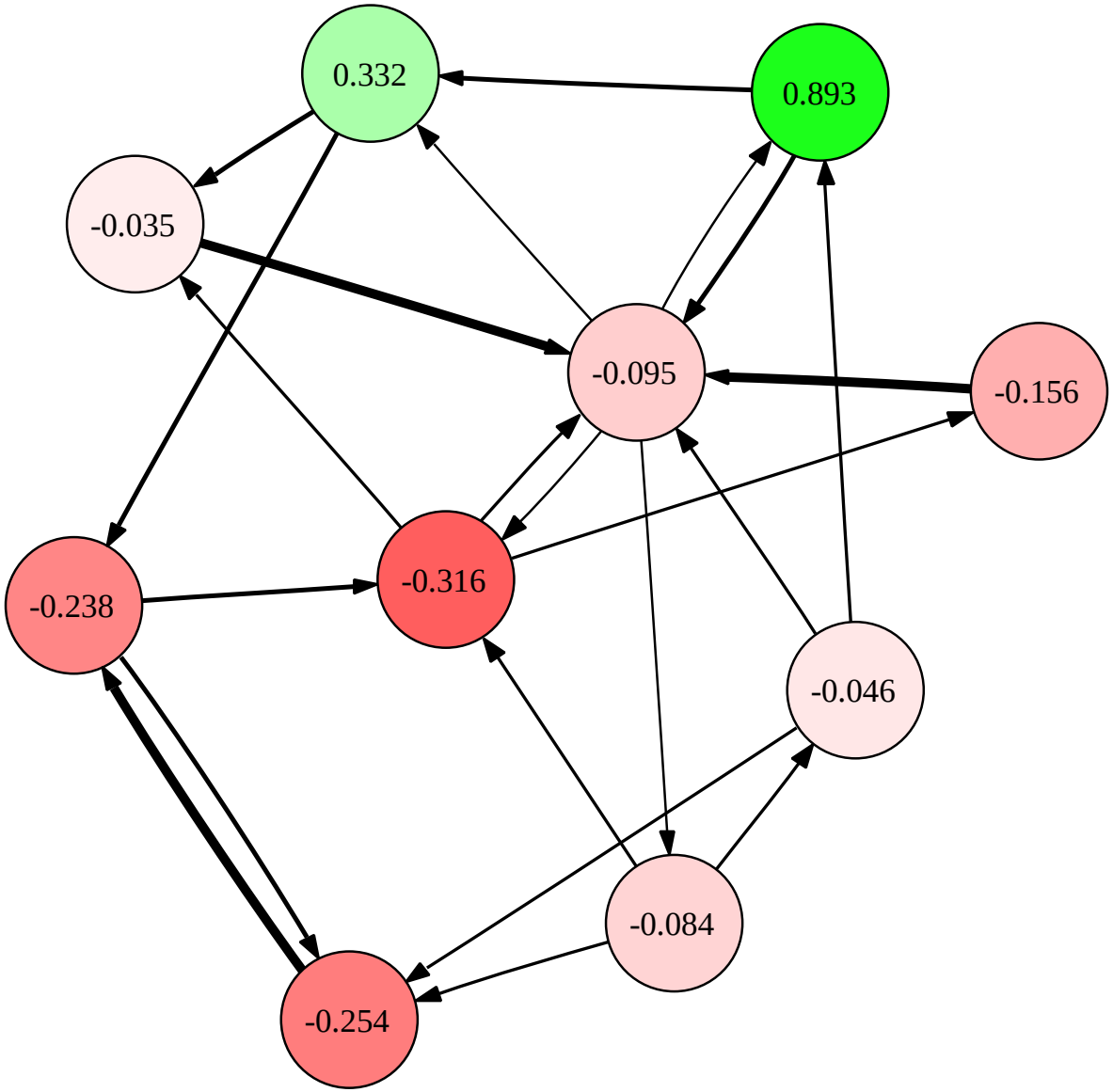
Green — Iteration ♯12



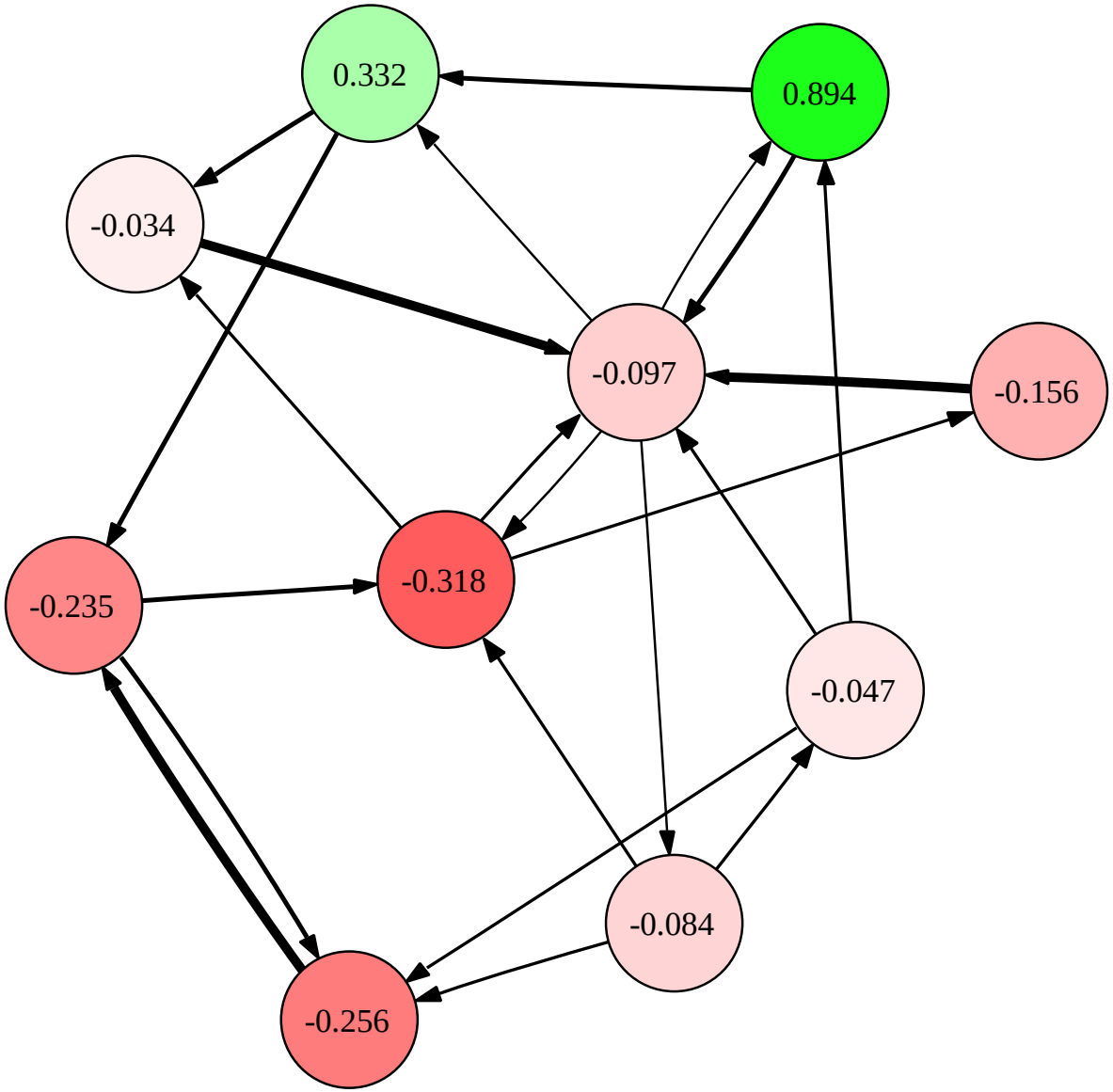
Green — Iteration ♯13



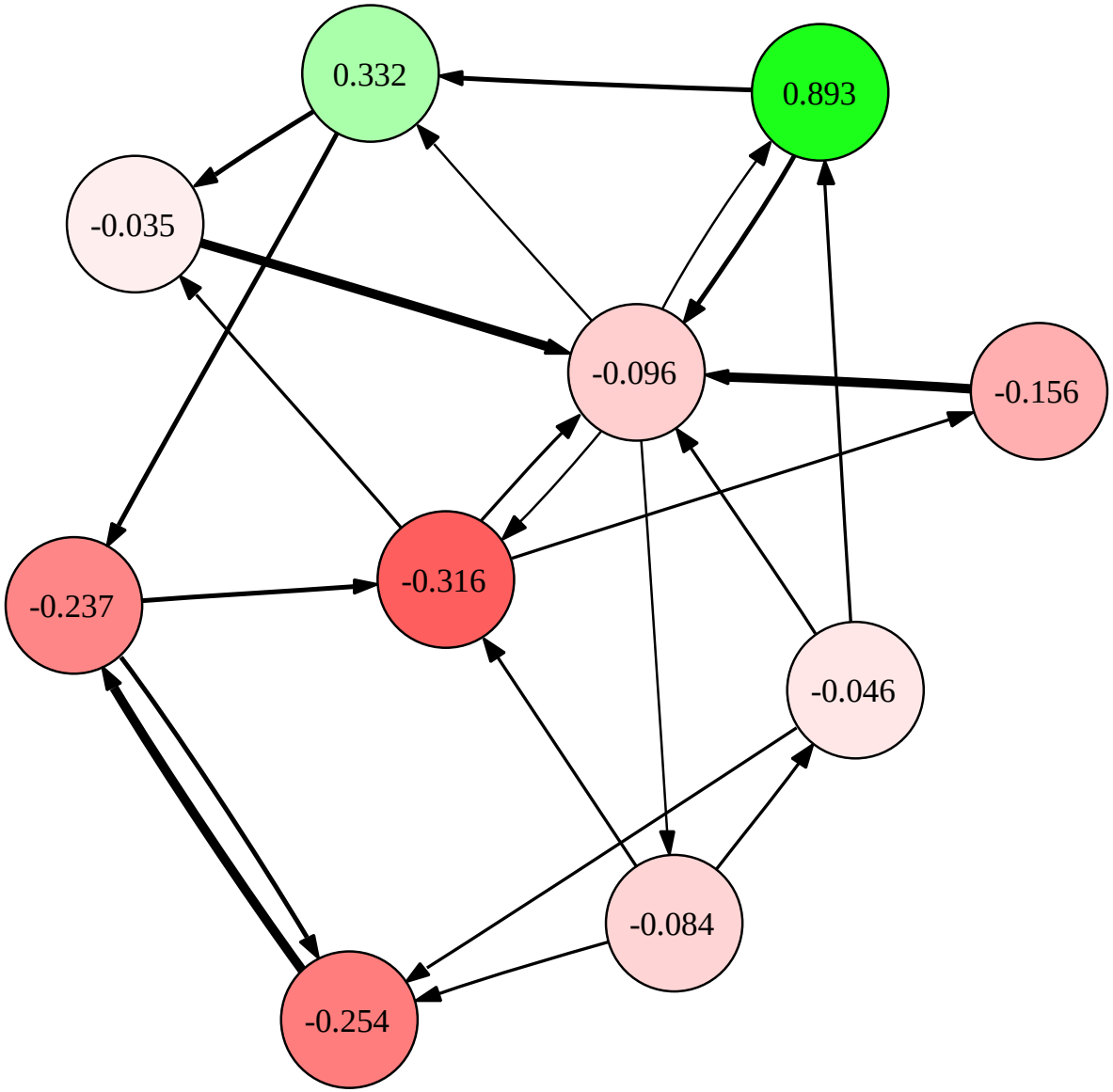
Green — Iteration ♯14



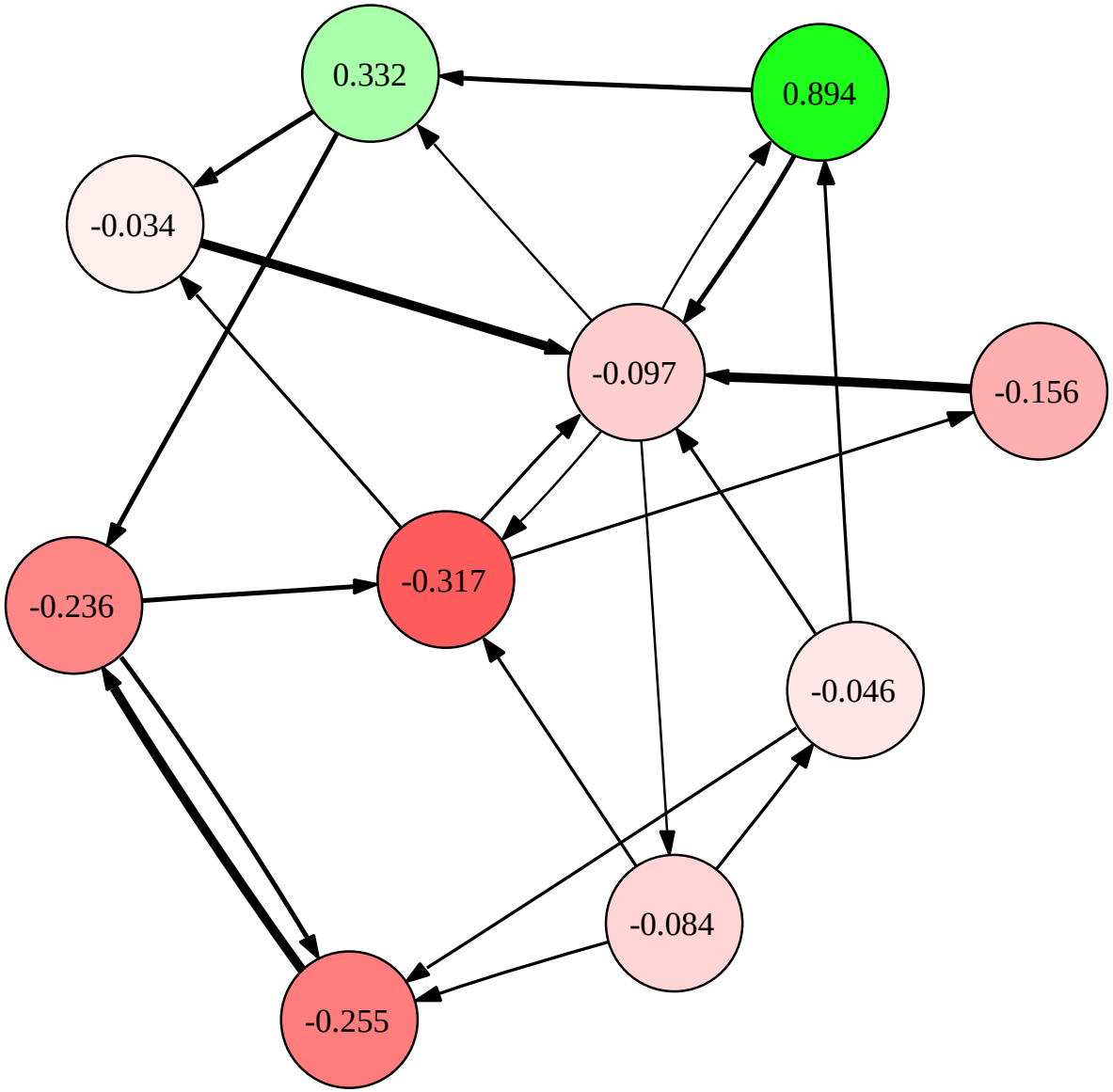
Green — Iteration ♯15



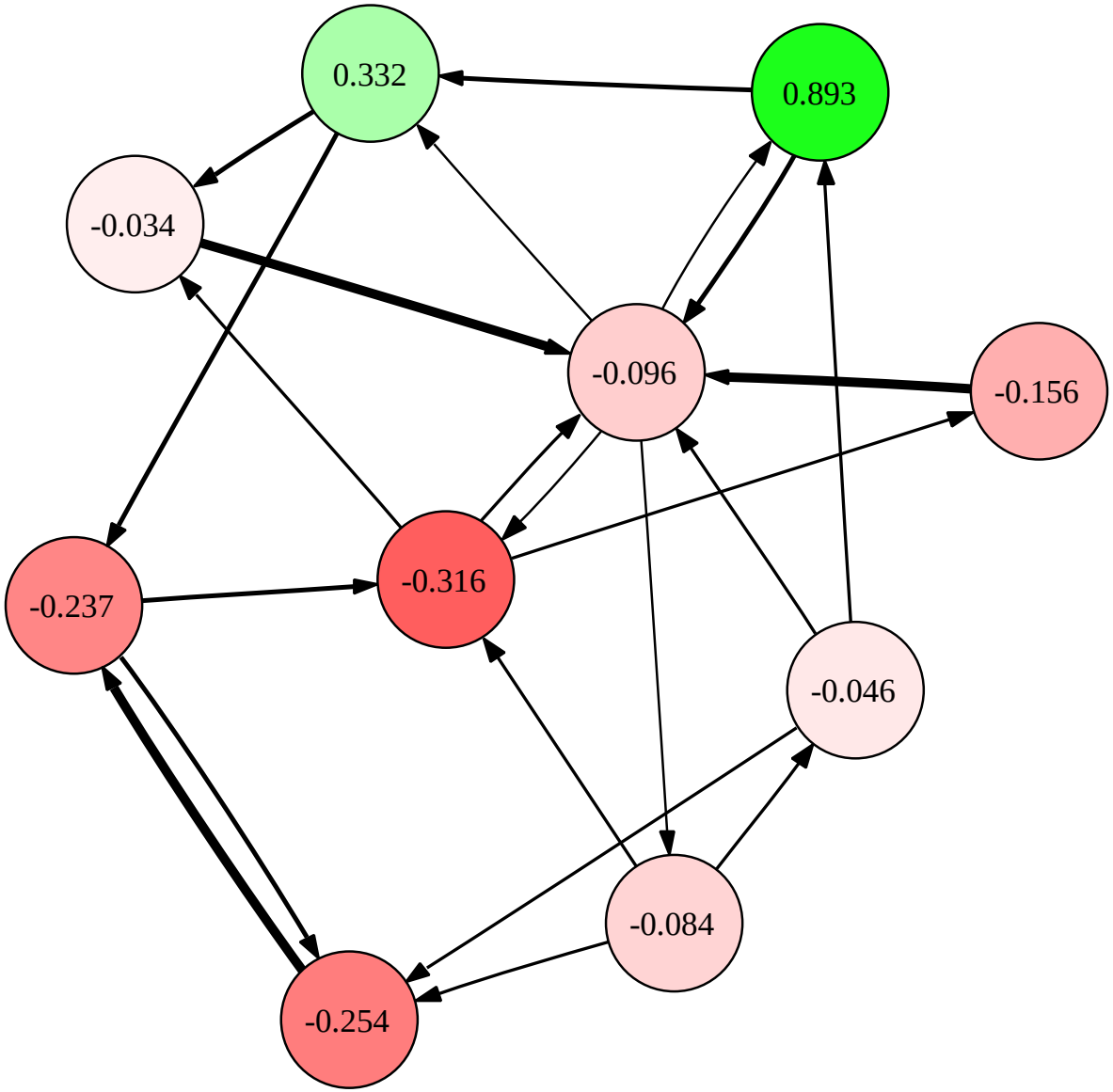
Green — Iteration ♯16



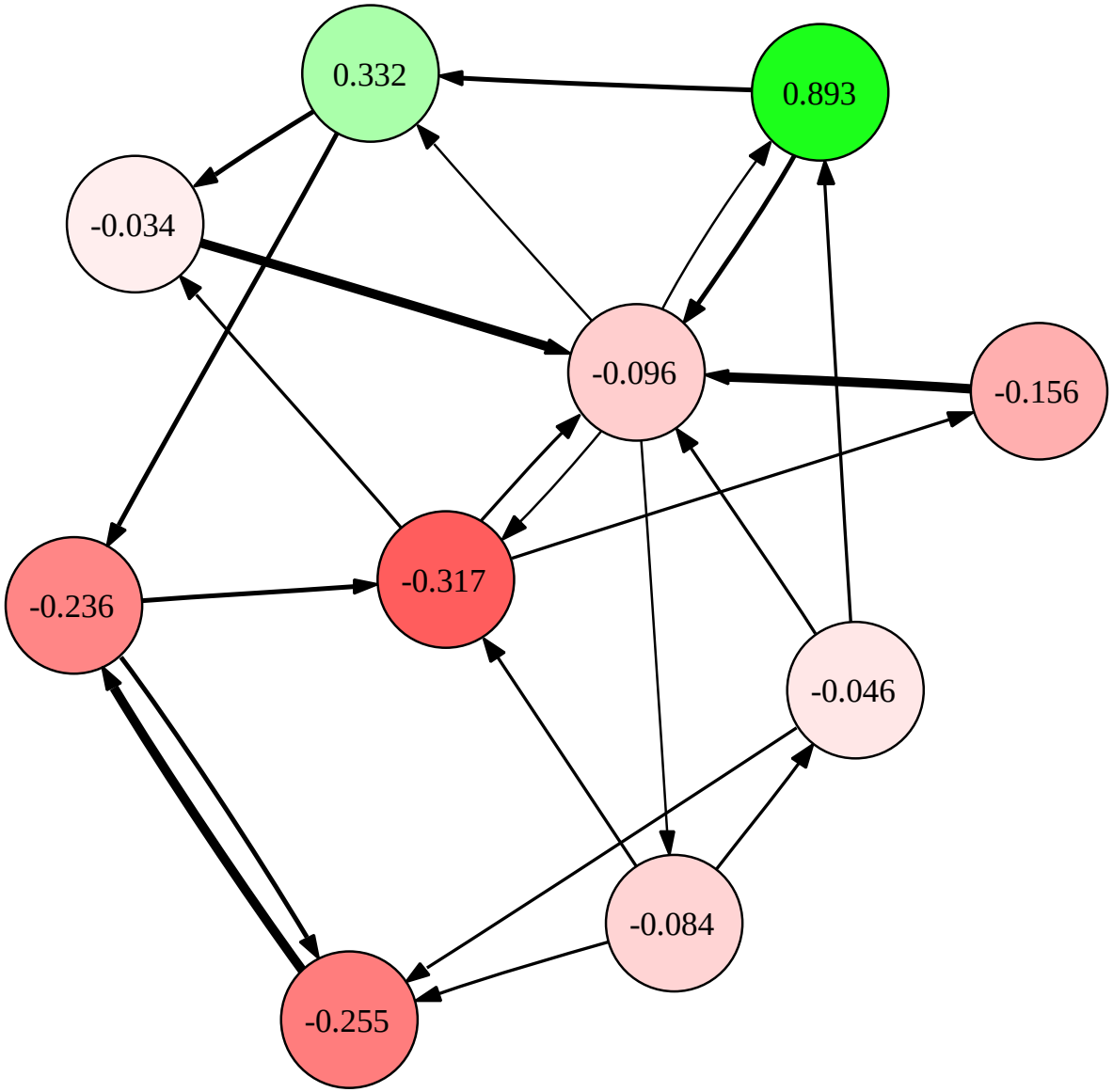
Green — Iteration ♯17



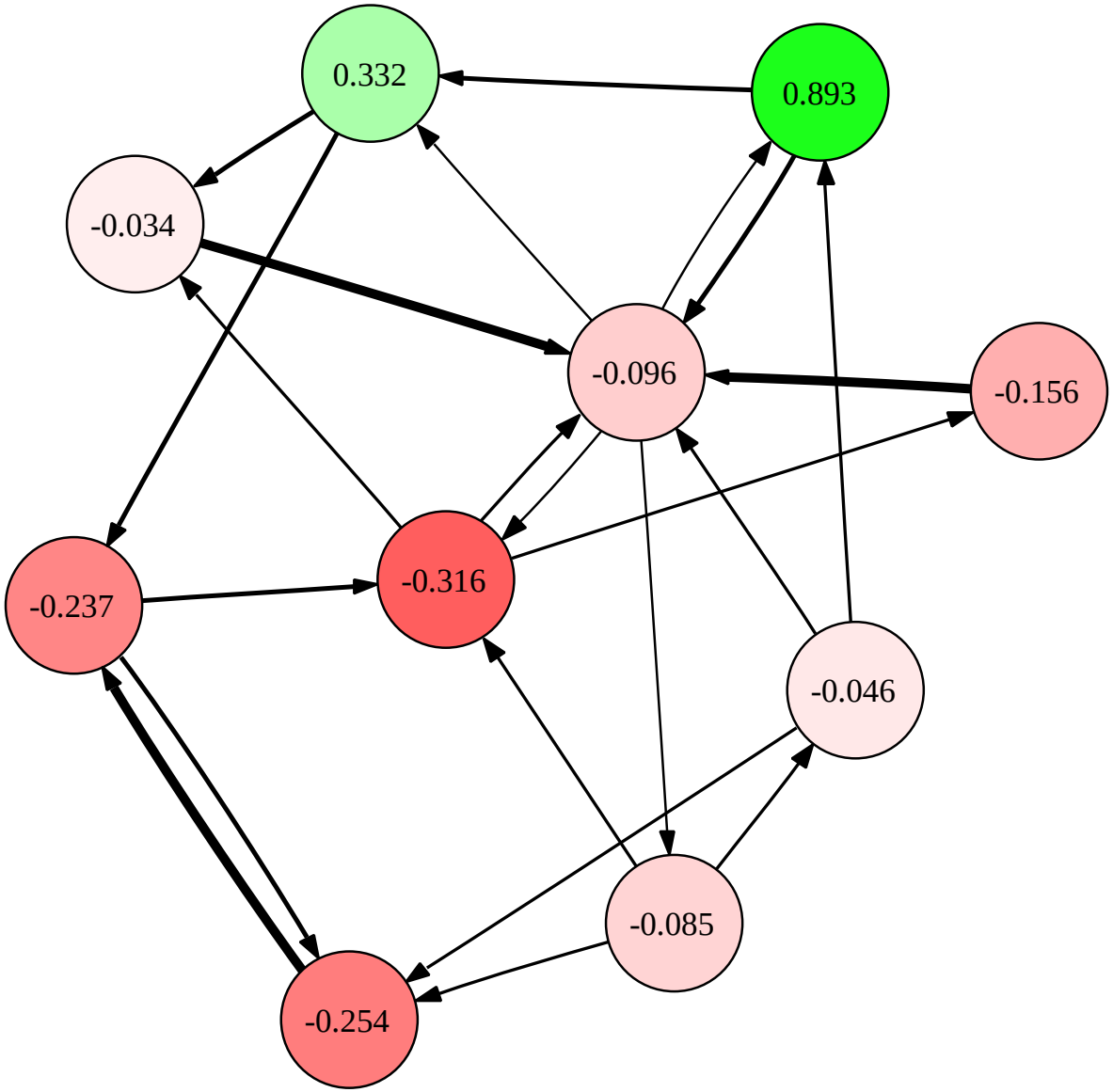
Green — Iteration ♯18



Green — Iteration ♯19



Green — Iteration ♯20



SymGreen — Method Description

Method Description

- *Green* goes only forward, may be a limitation.

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.

SymGreen — Method Description

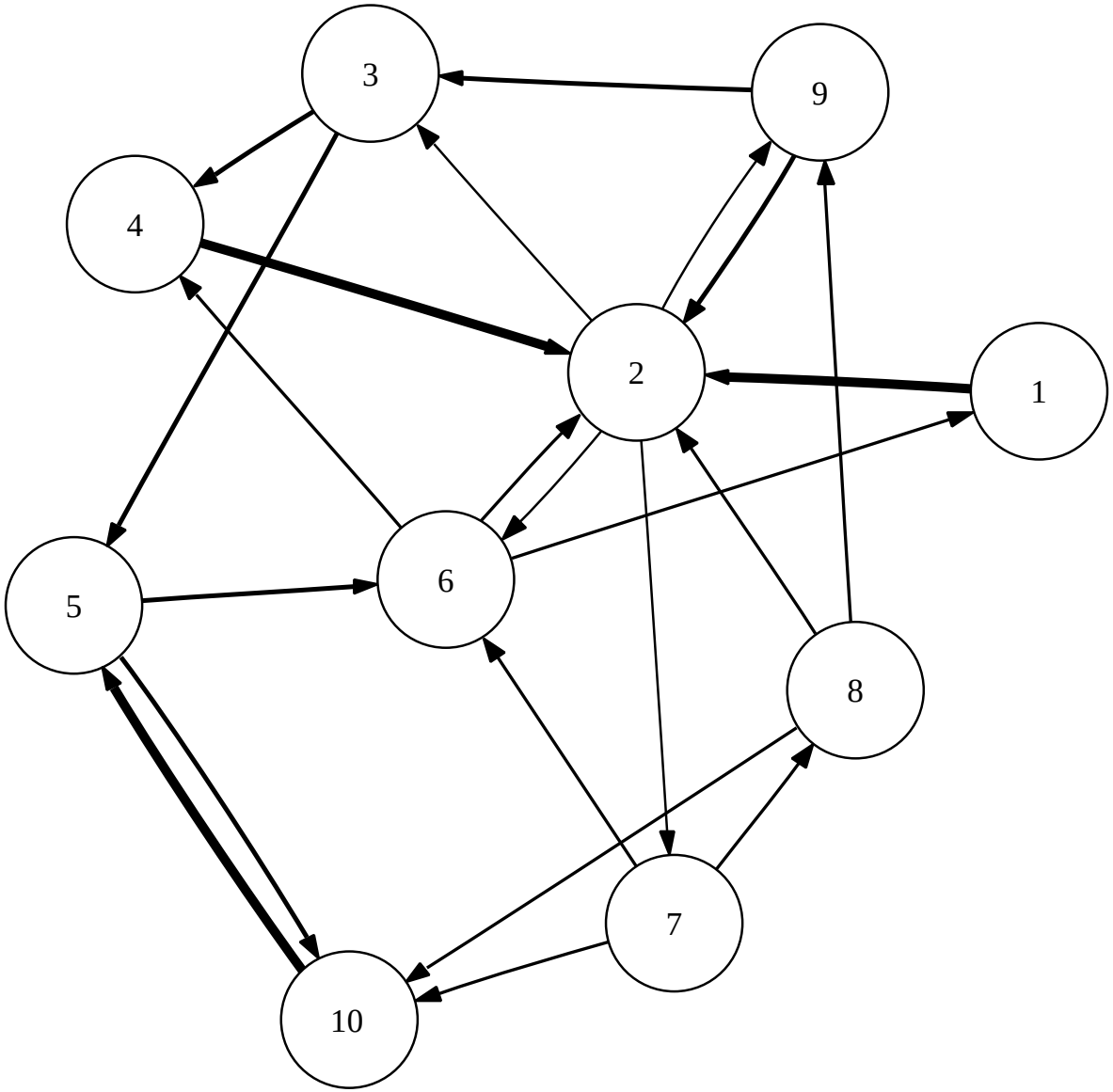
Method Description

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

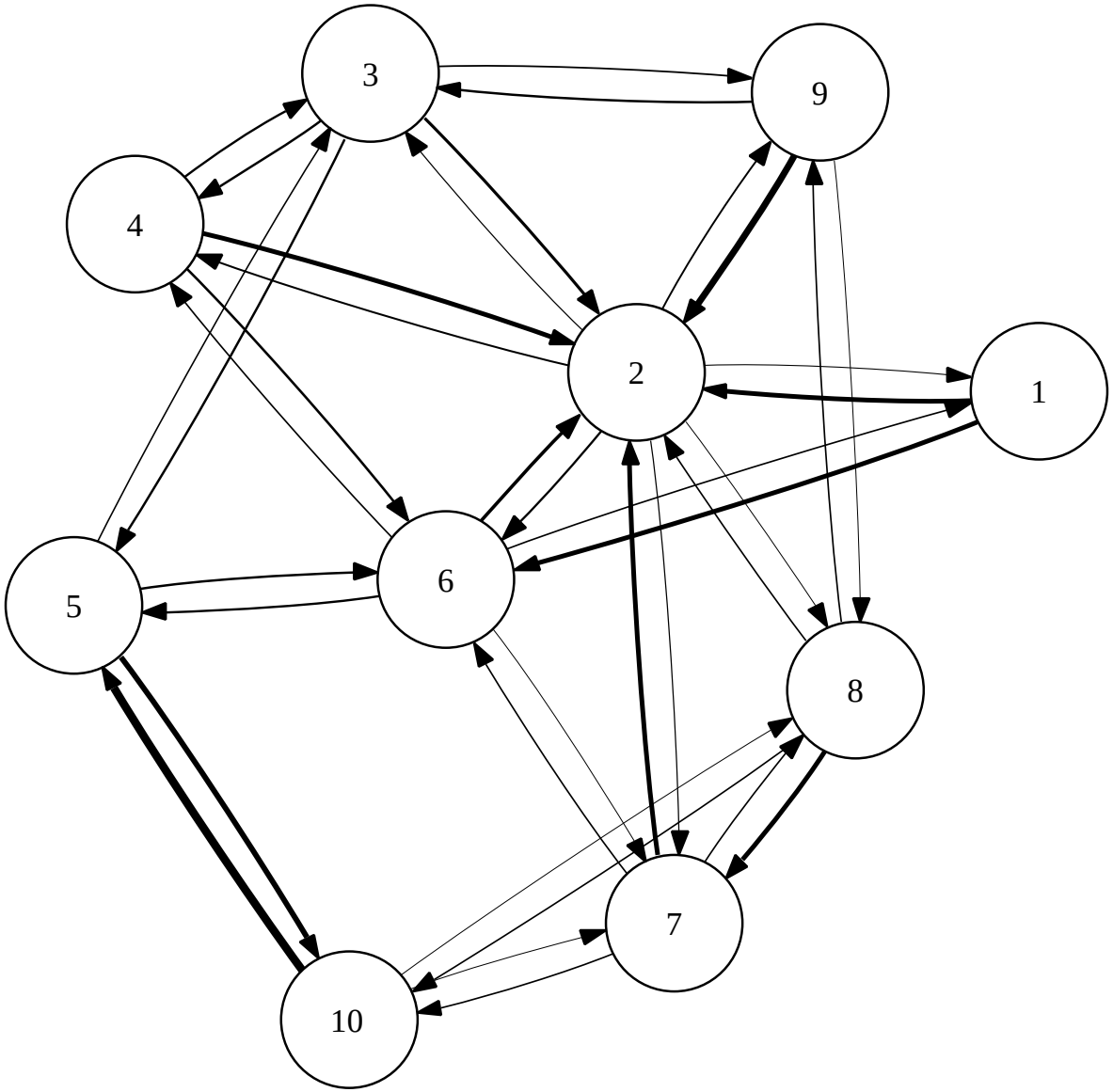
$$\tilde{p}_{ij} = \frac{1}{2}\left(p_{ij} + p_{ji}\frac{\nu_j}{\nu_i}\right)$$

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

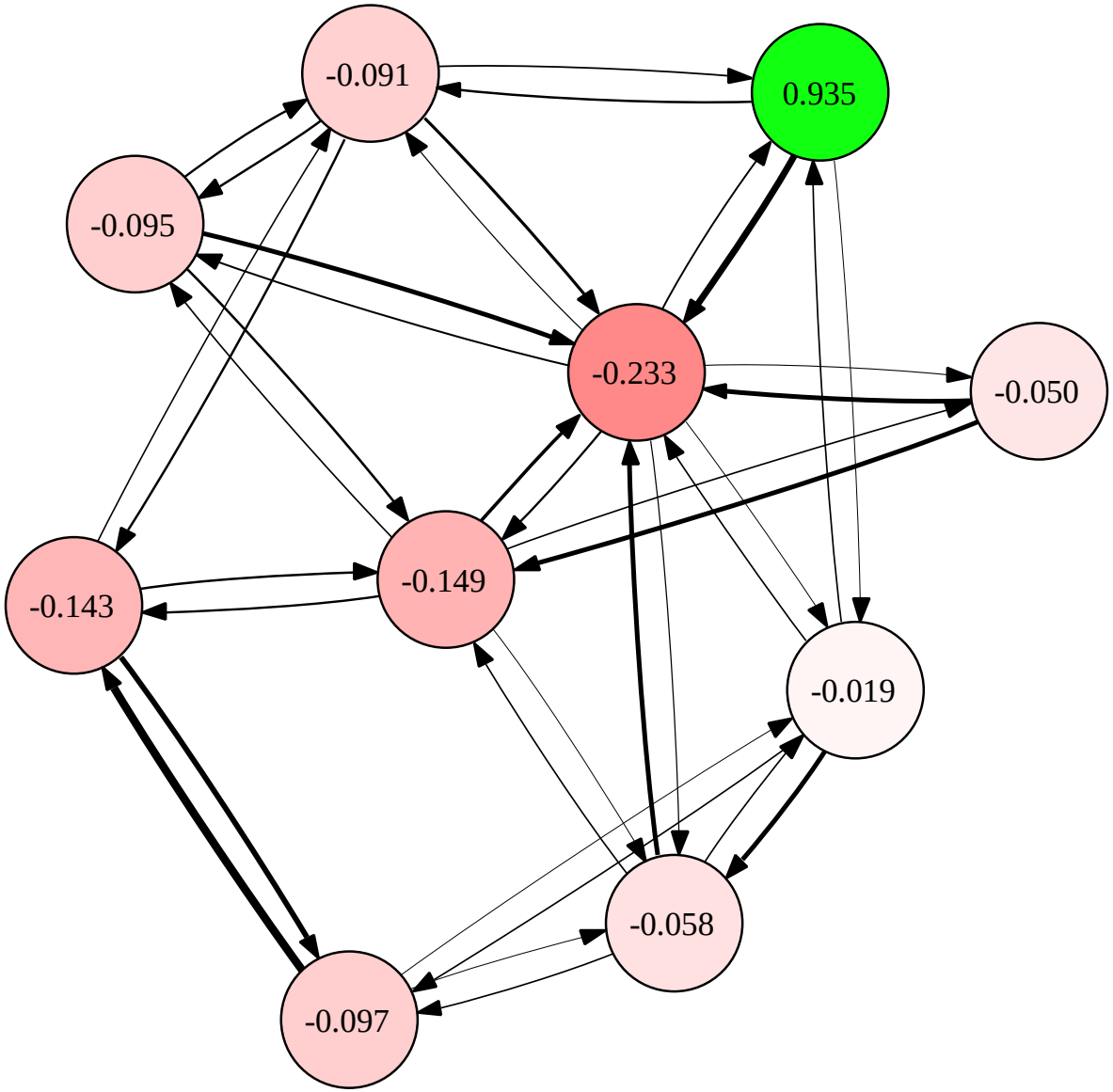
SymGreen — Original Graph



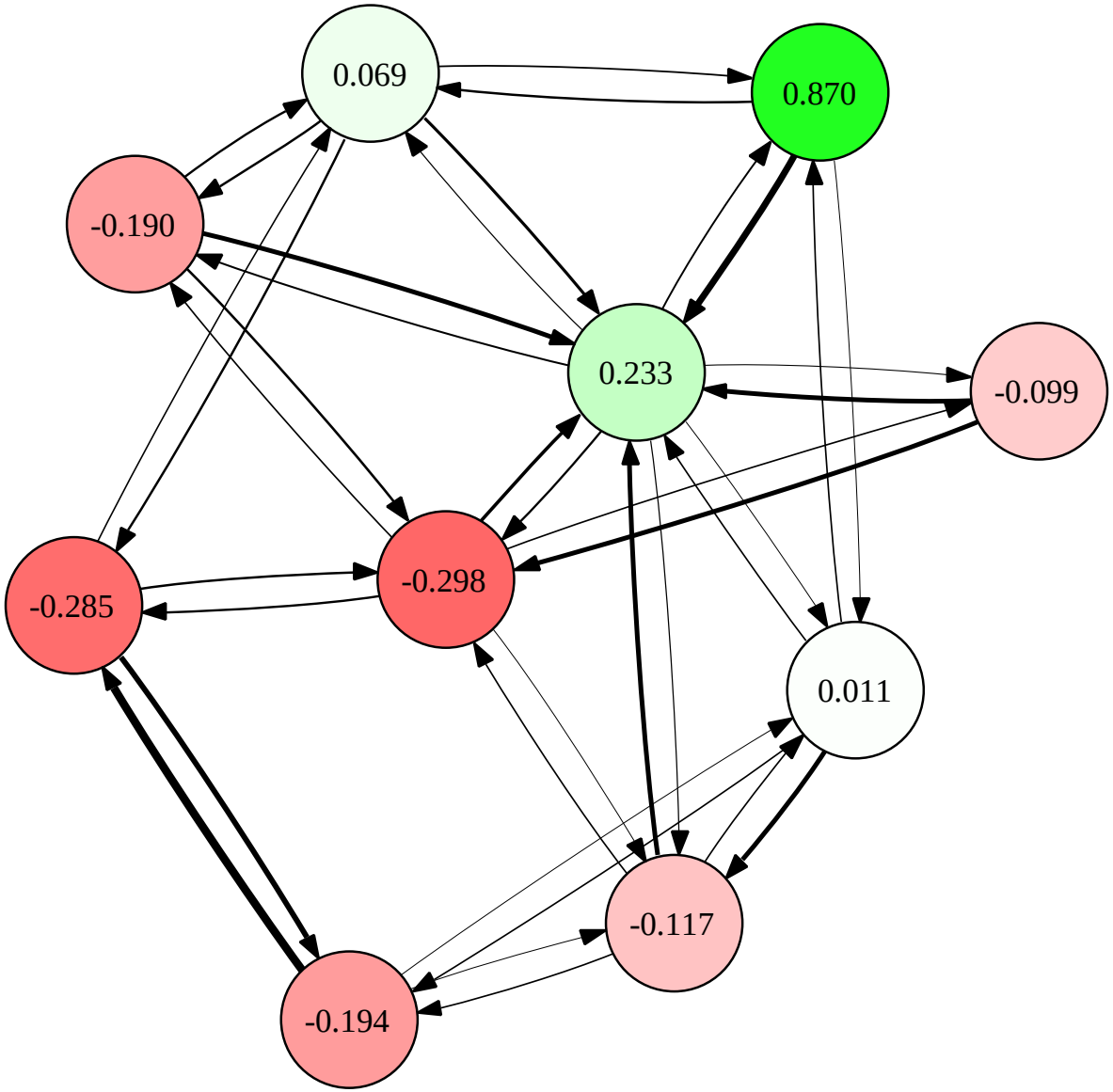
SymGreen — Symmetrized Graph



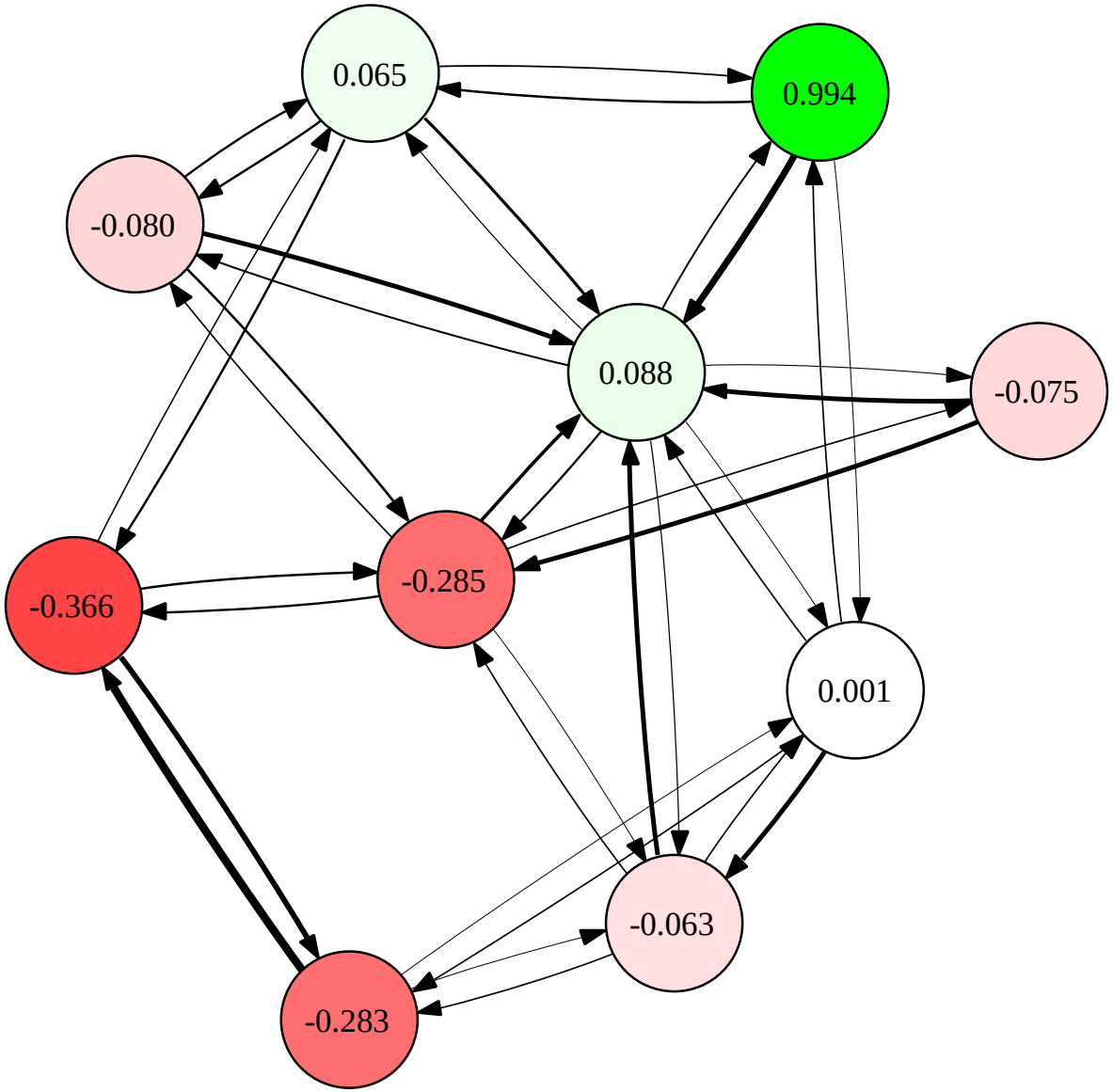
SymGreen — Iteration ♯1



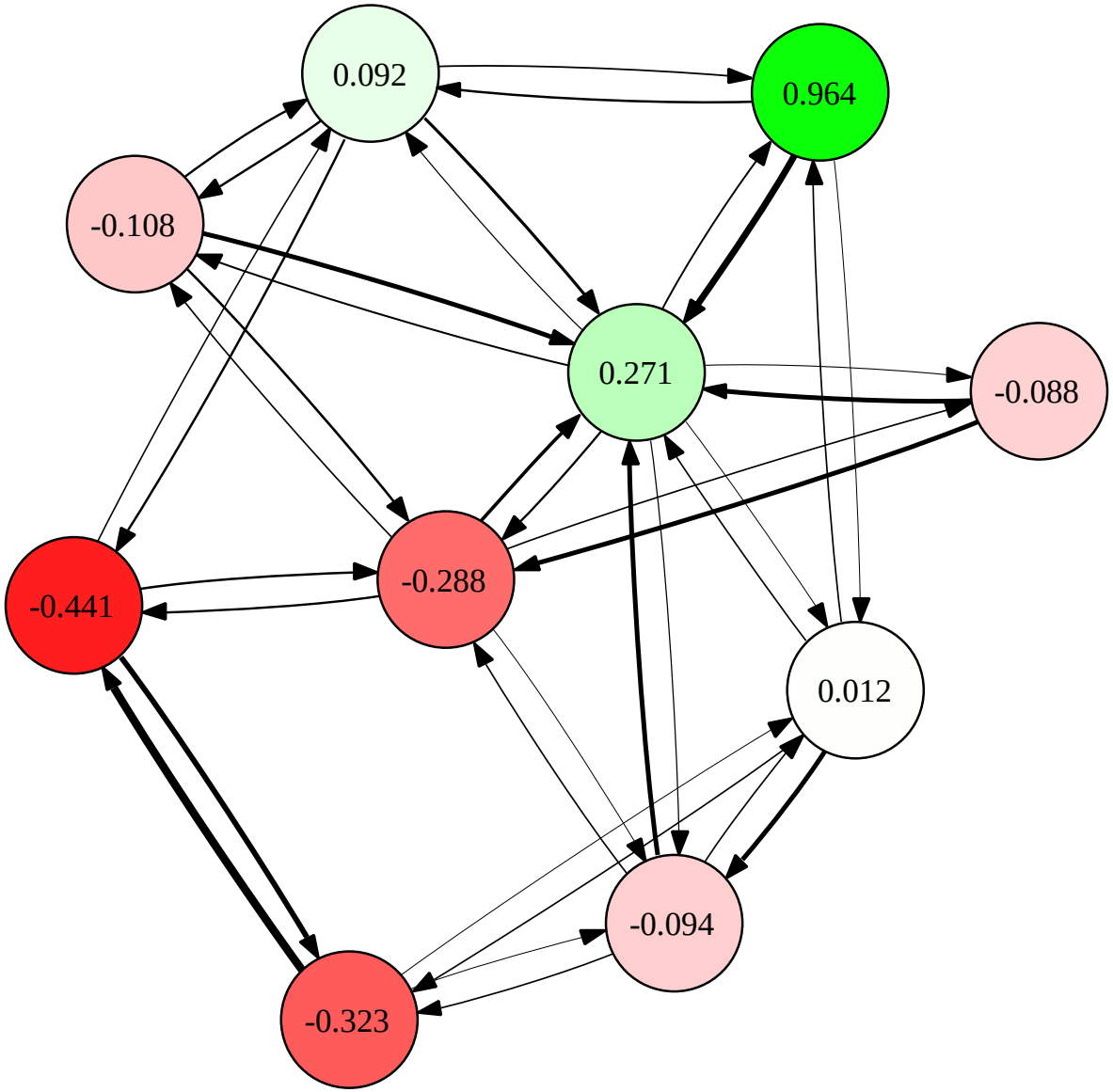
SymGreen — Iteration ♯2



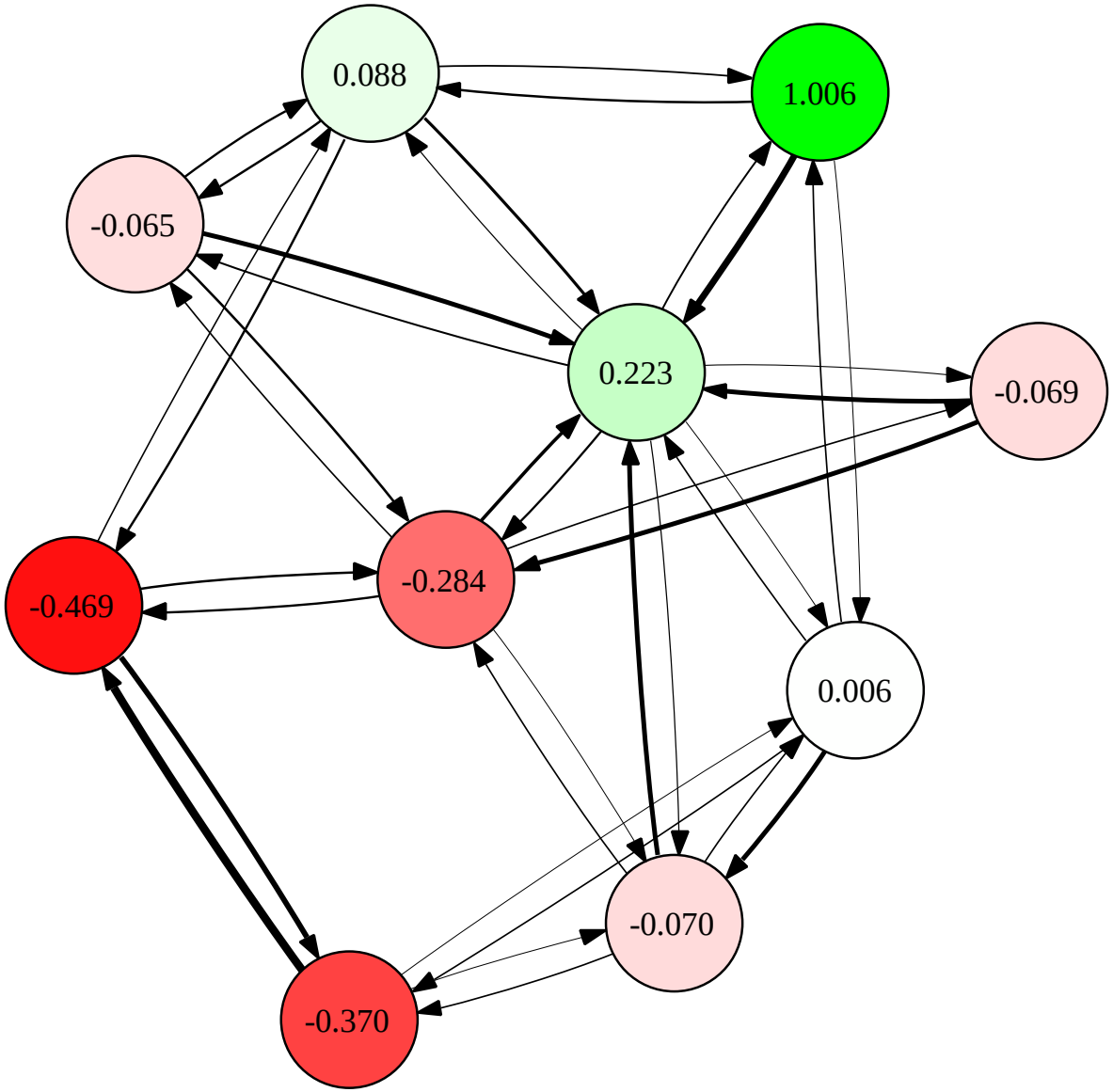
SymGreen — Iteration ♯3



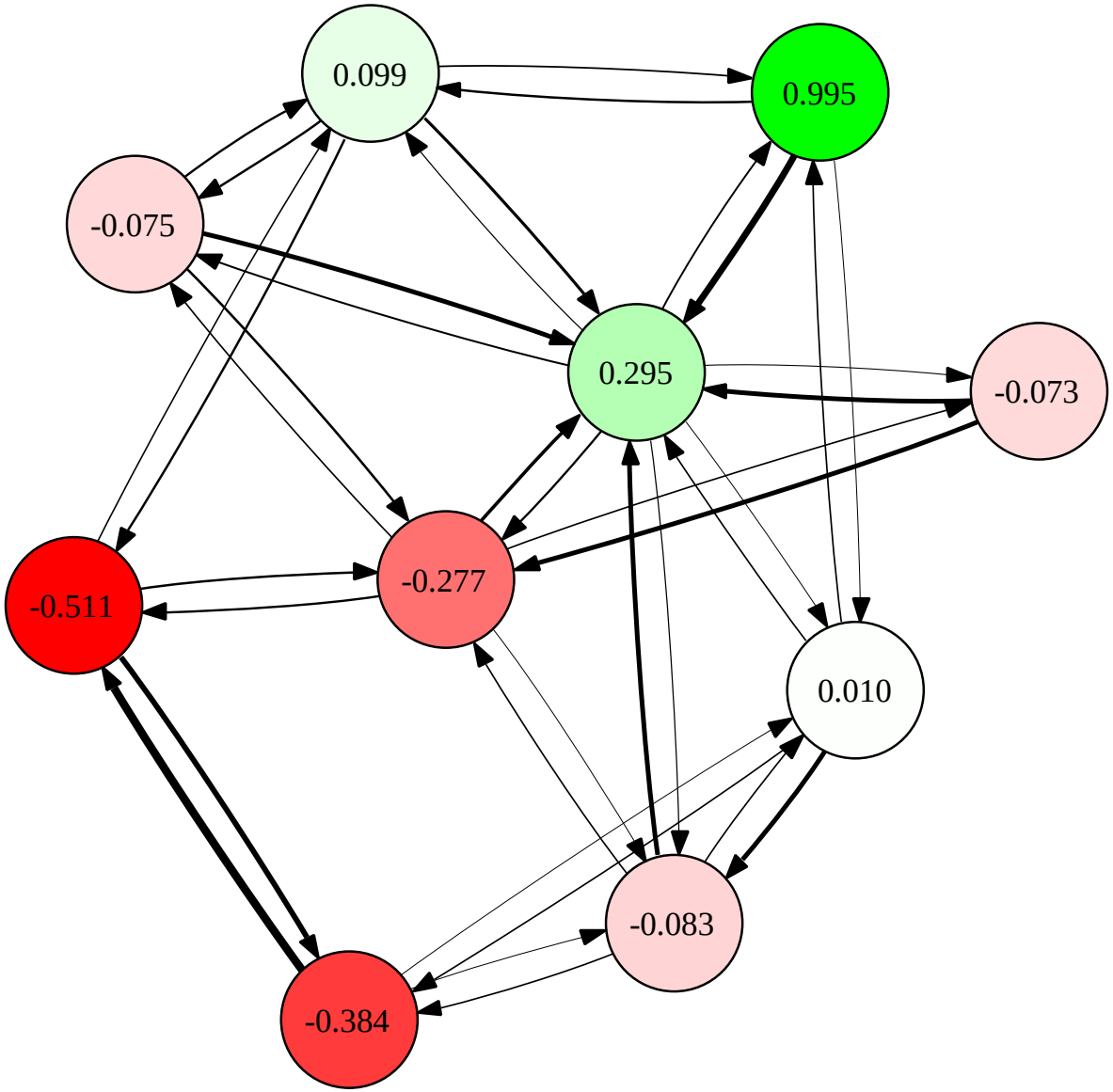
SymGreen — Iteration ♯4



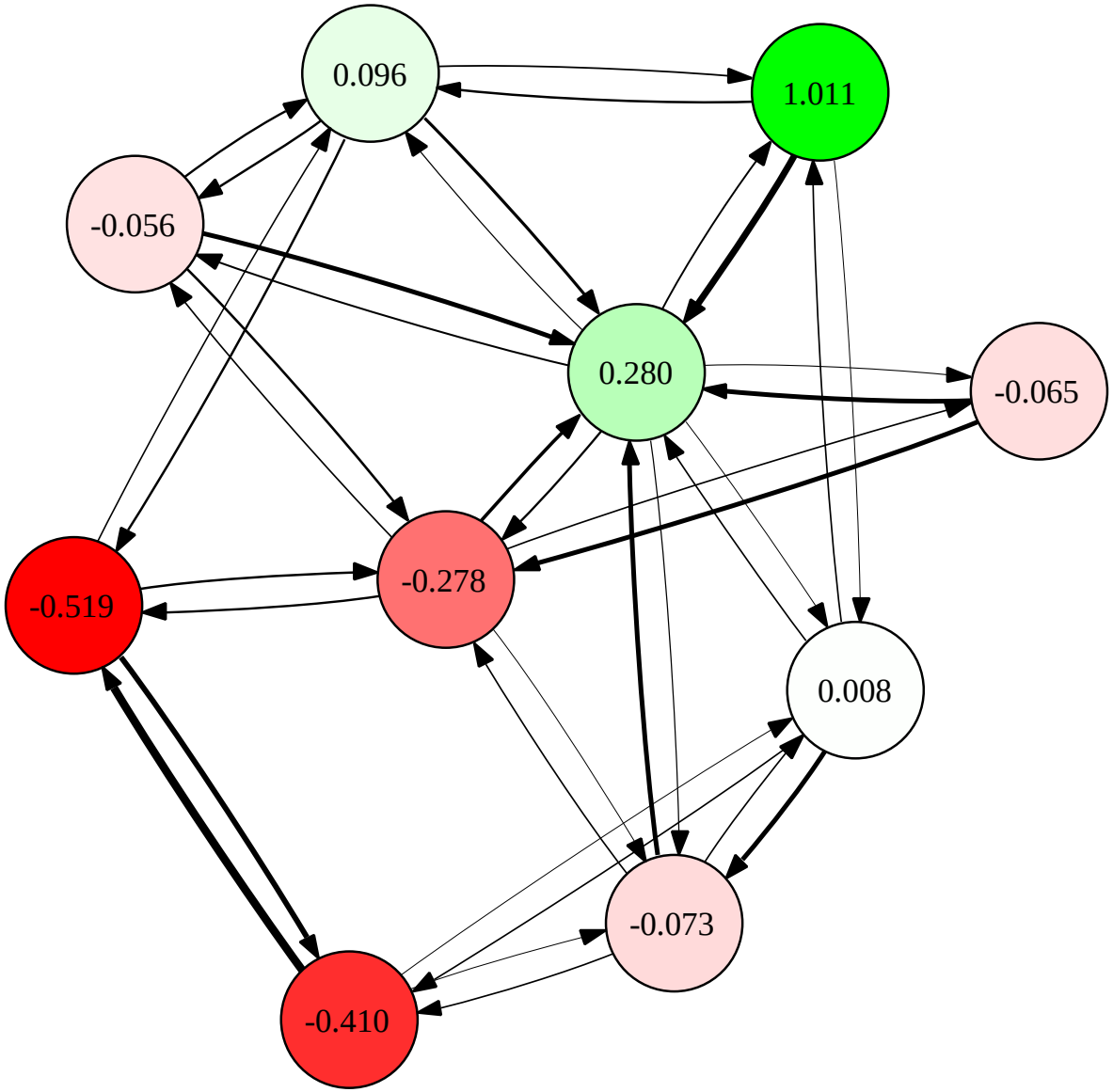
SymGreen — Iteration ♯5



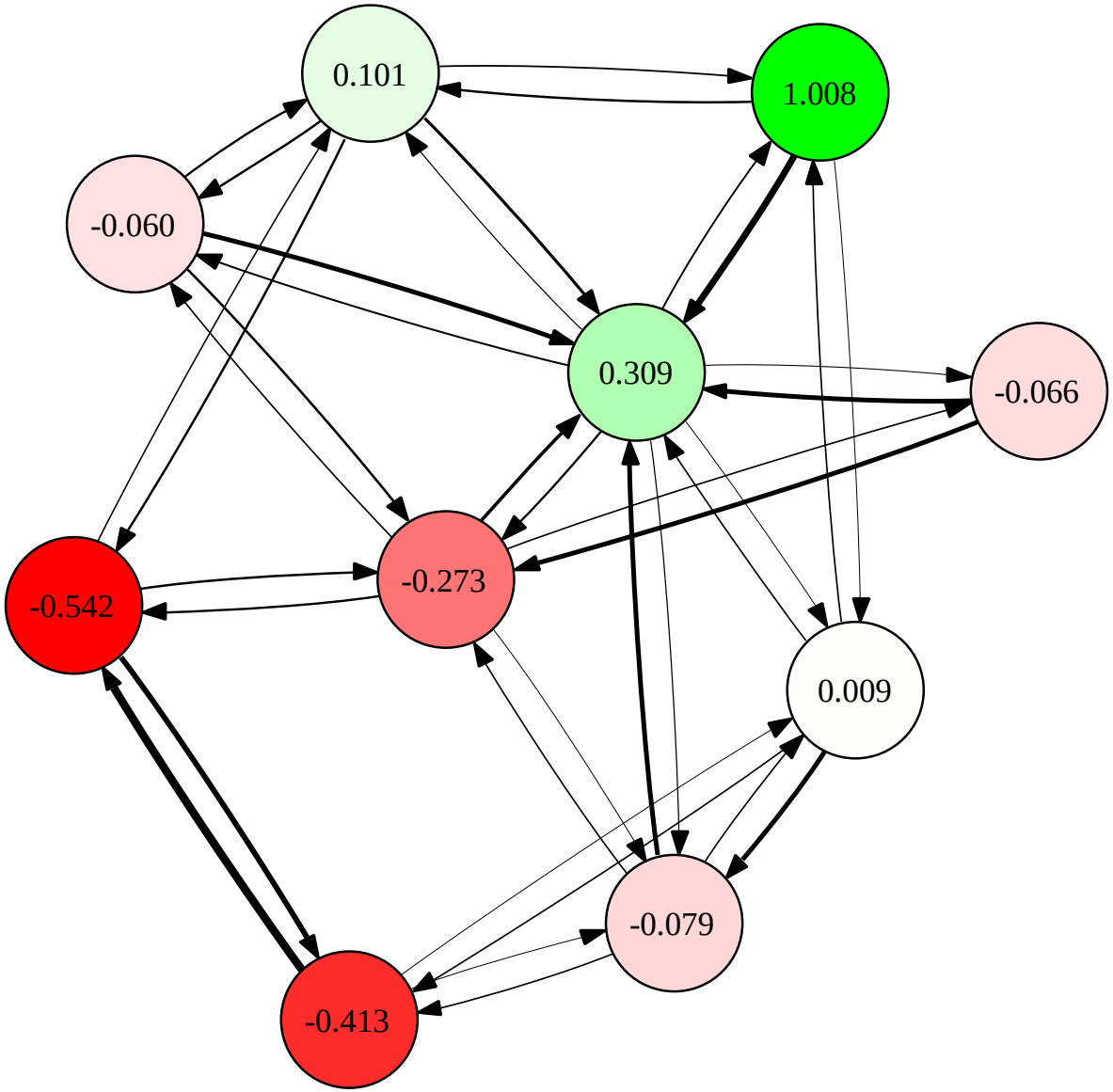
SymGreen — Iteration ♯6



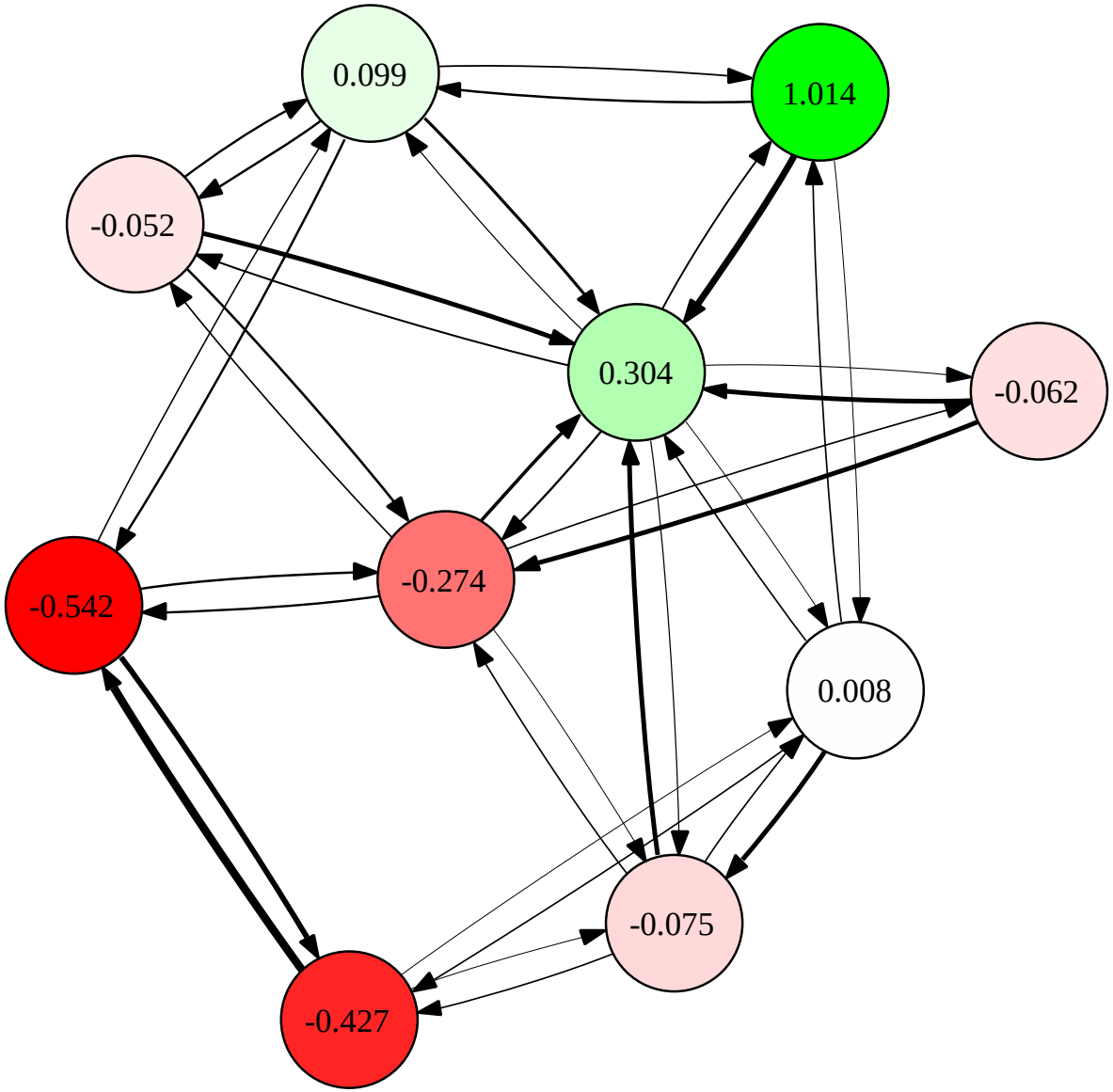
SymGreen — Iteration ♯7



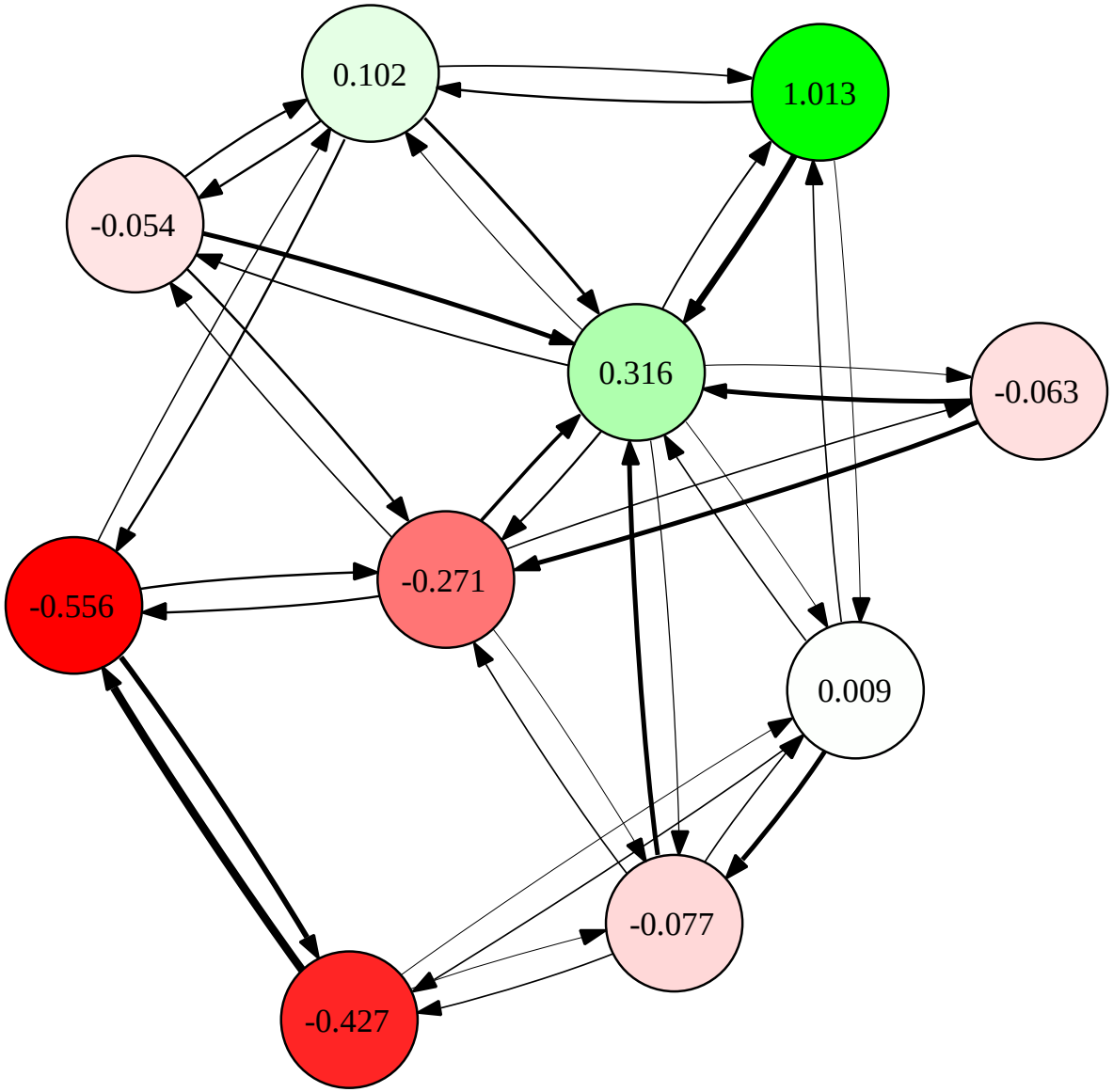
SymGreen — Iteration ♯8



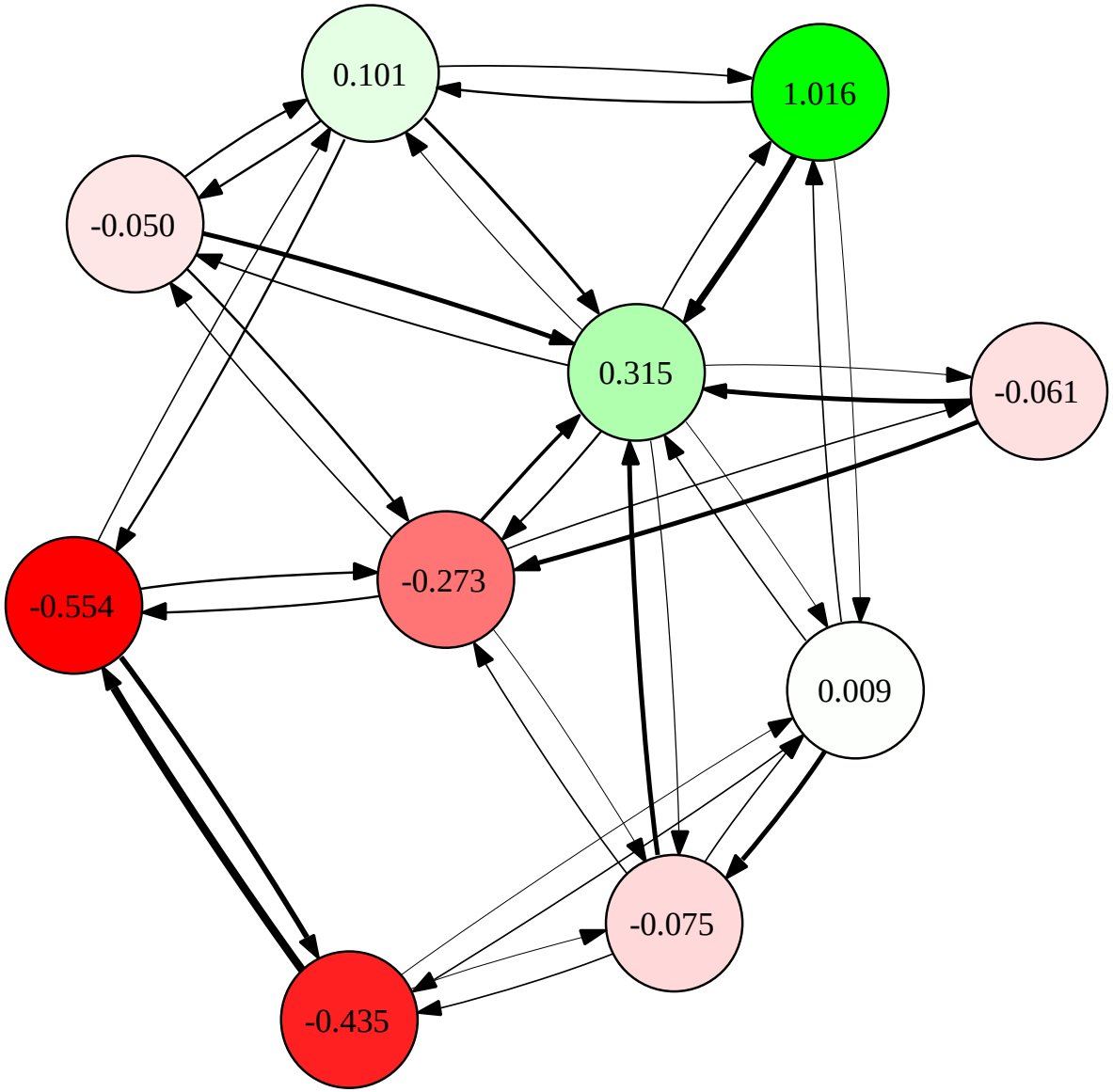
SymGreen — Iteration ♯9



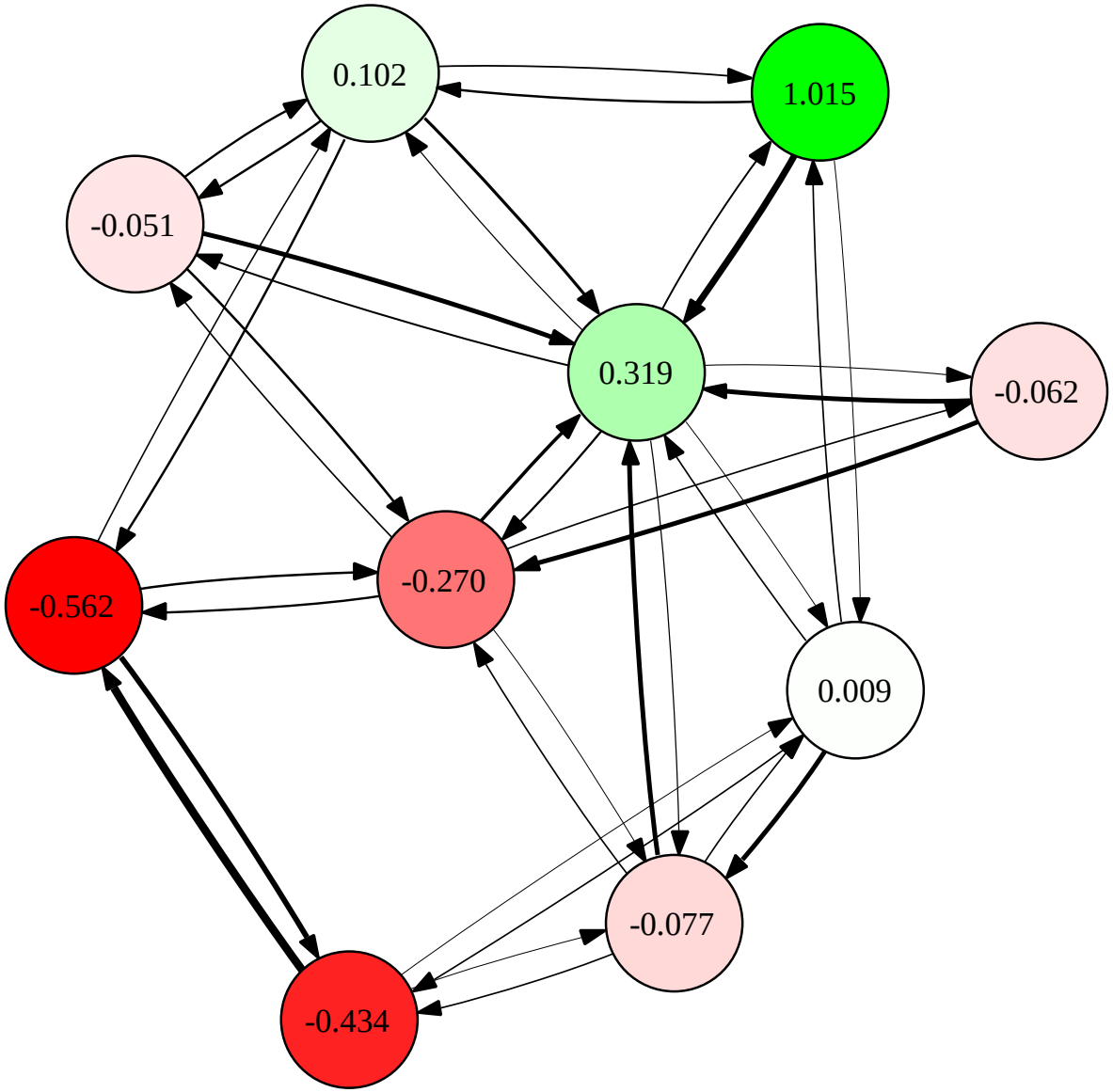
SymGreen — Iteration ♯10



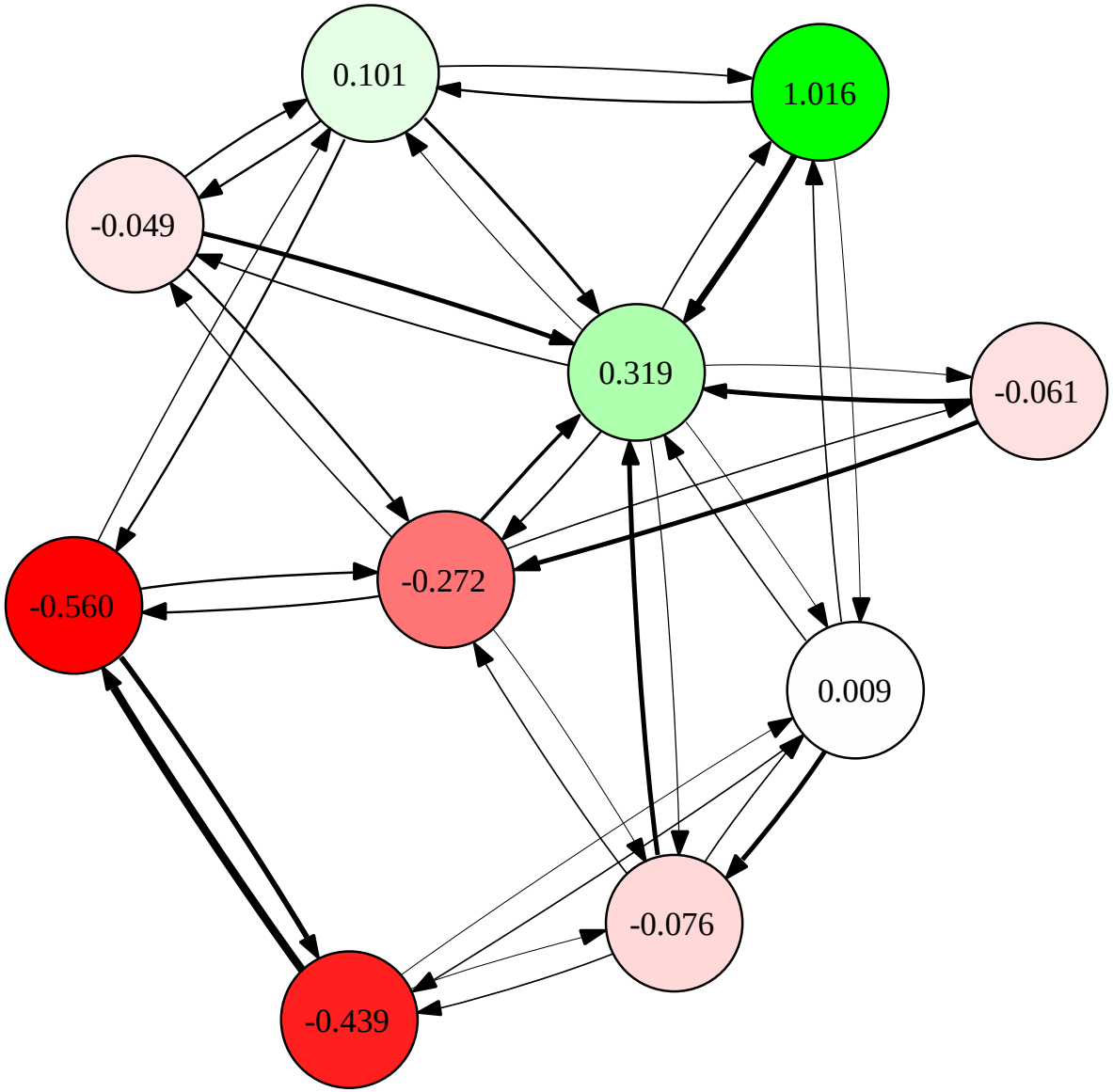
SymGreen — Iteration ♯11



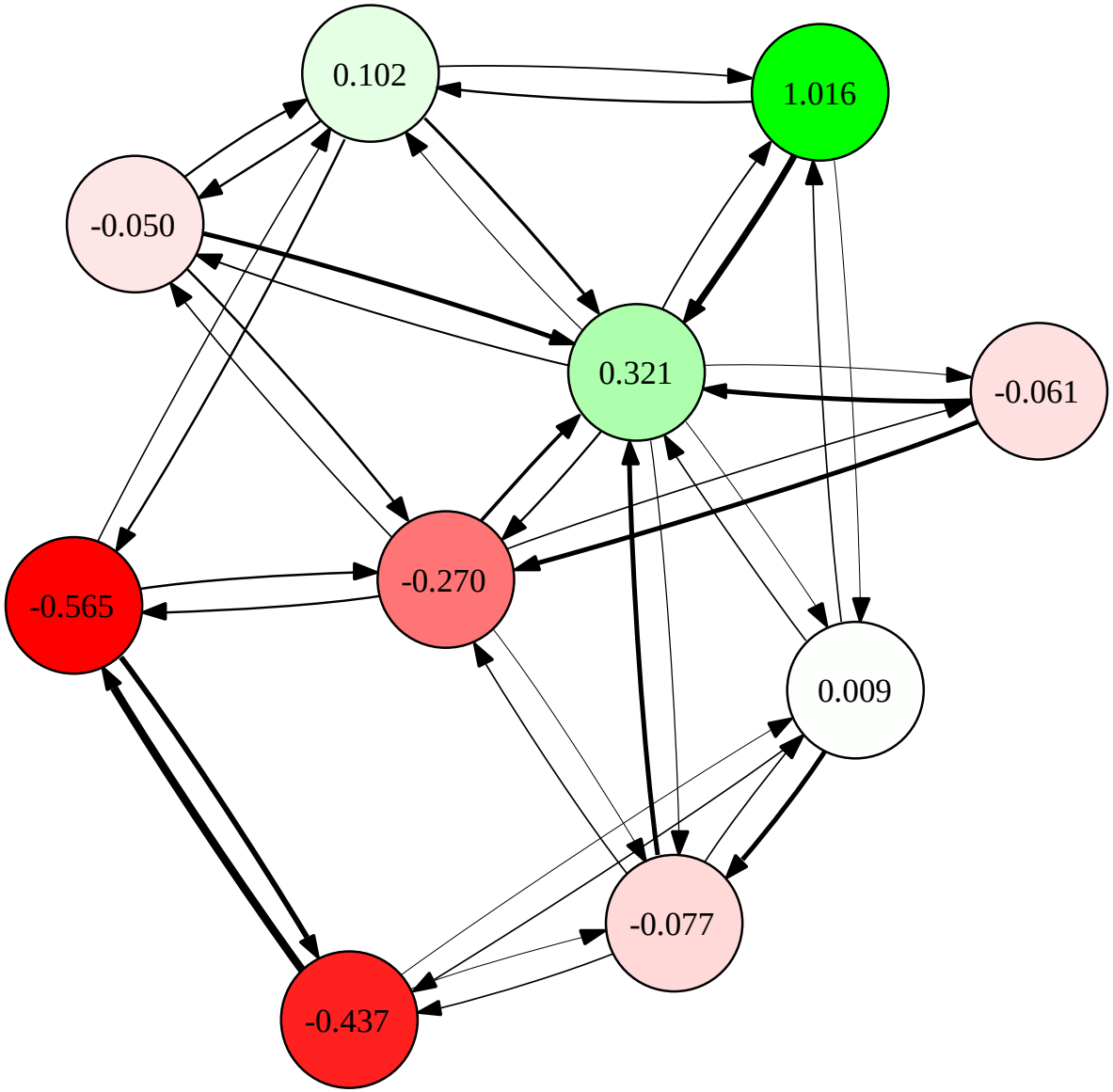
SymGreen — Iteration ♯12



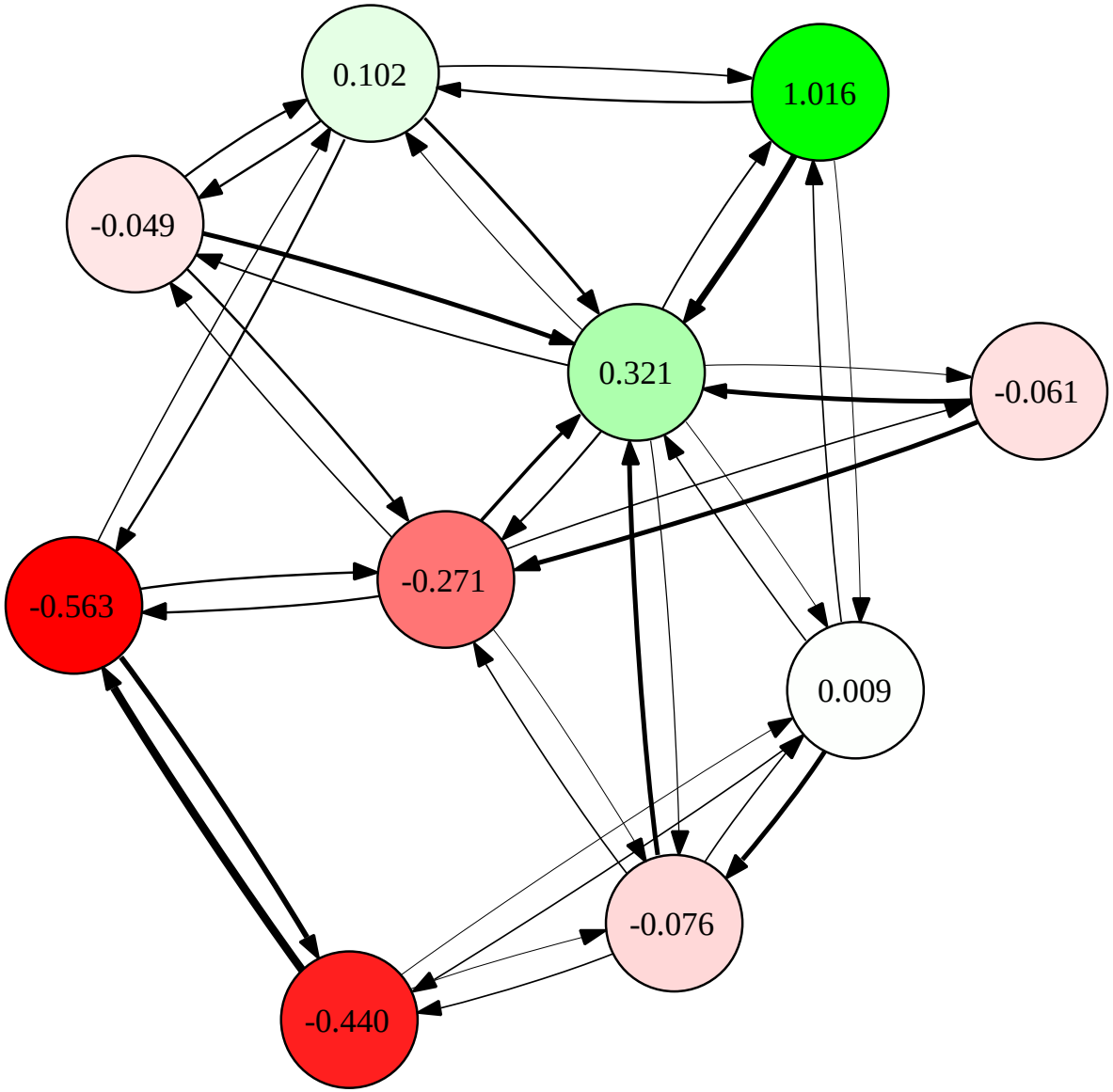
SymGreen — Iteration ♯13



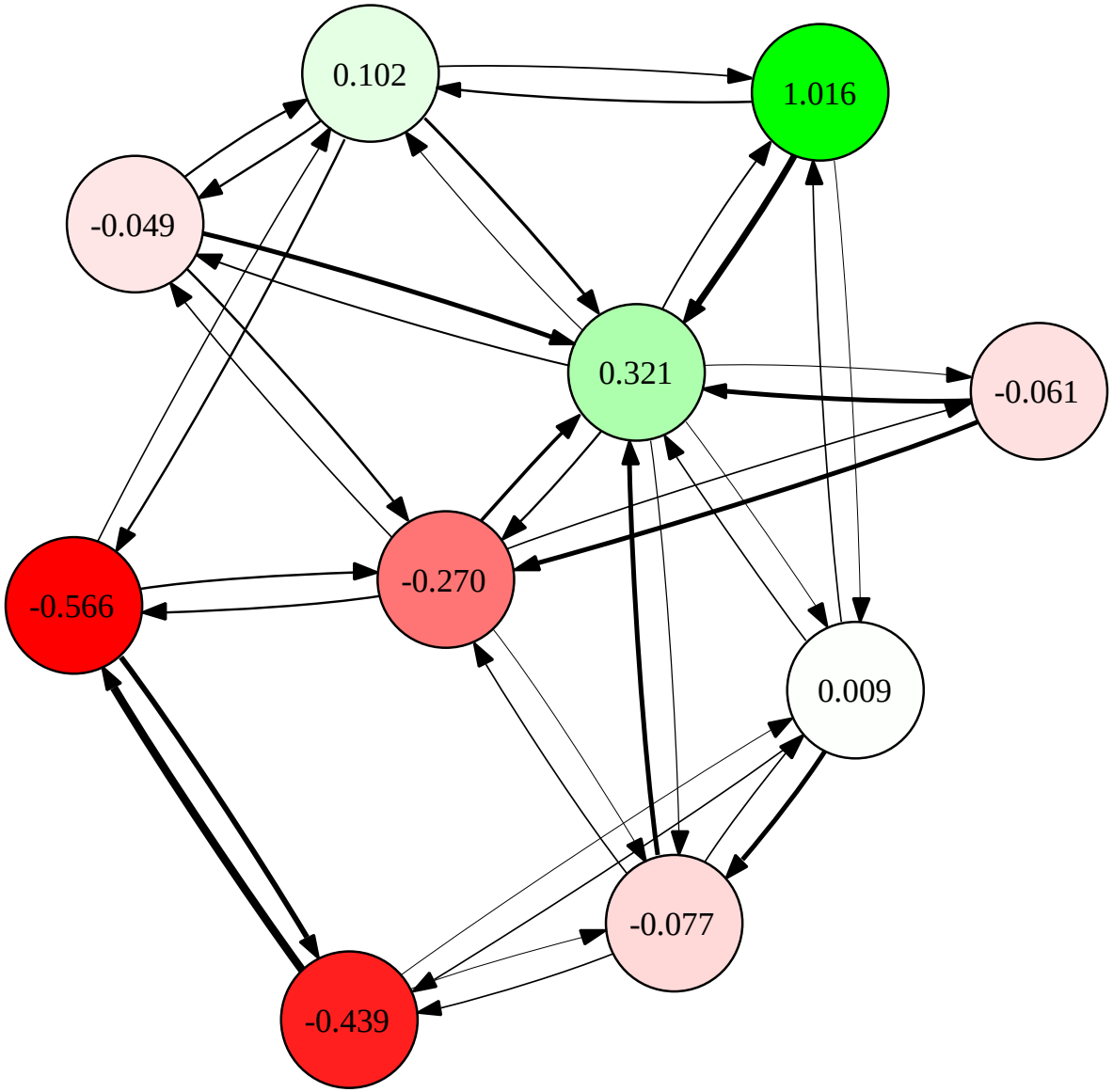
SymGreen — Iteration ♯14



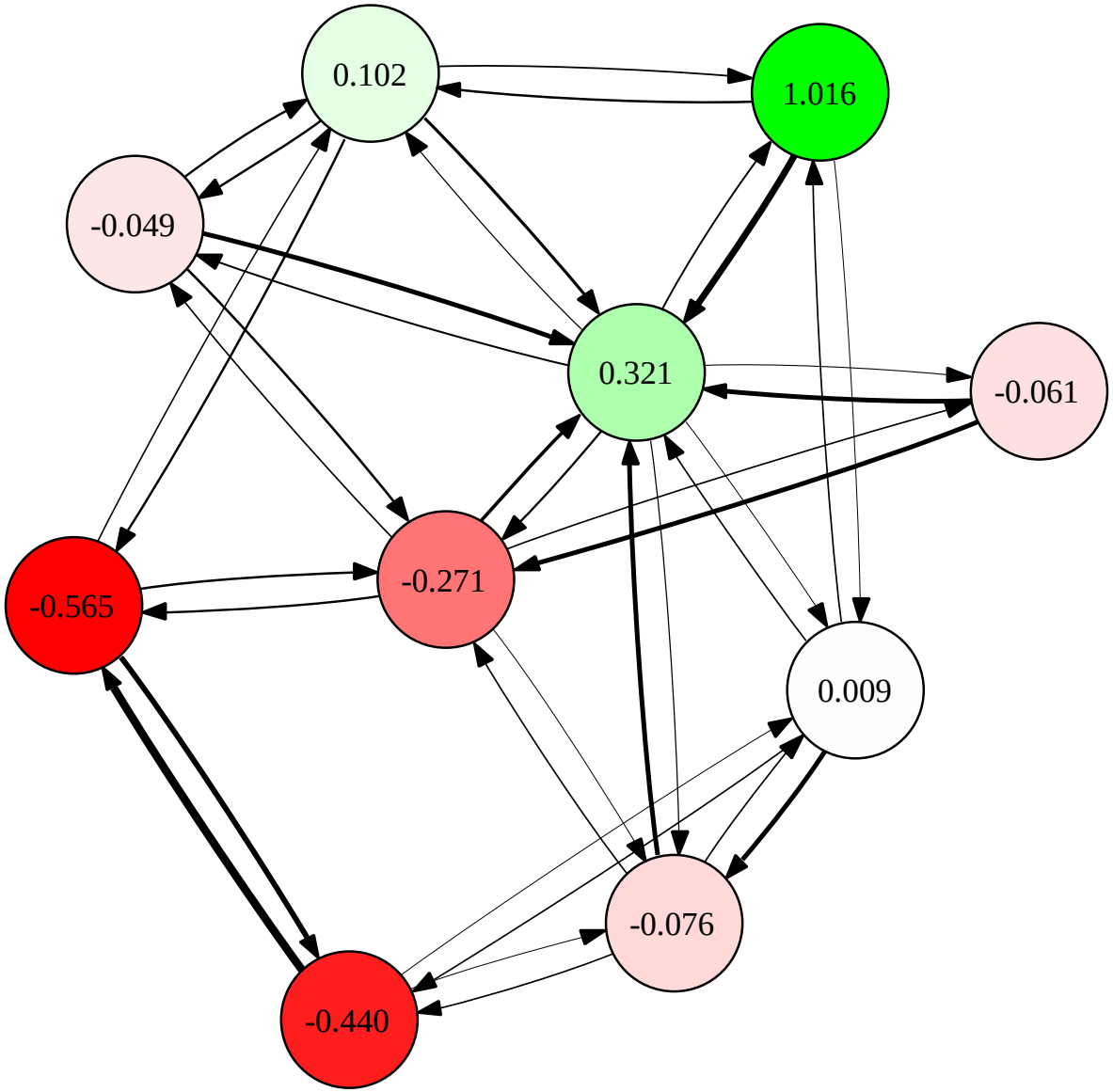
SymGreen — Iteration ♯15



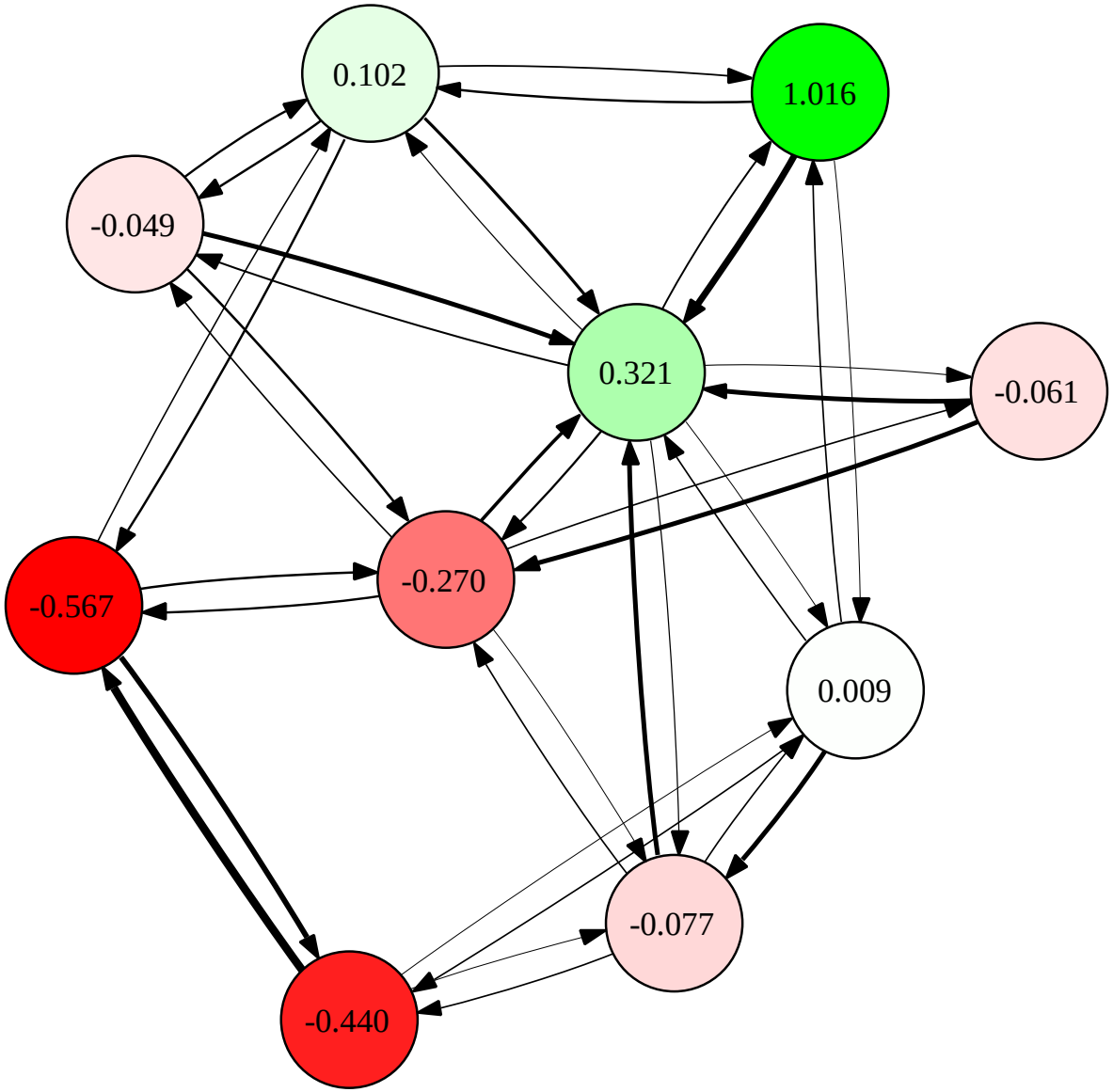
SymGreen — Iteration ♯16



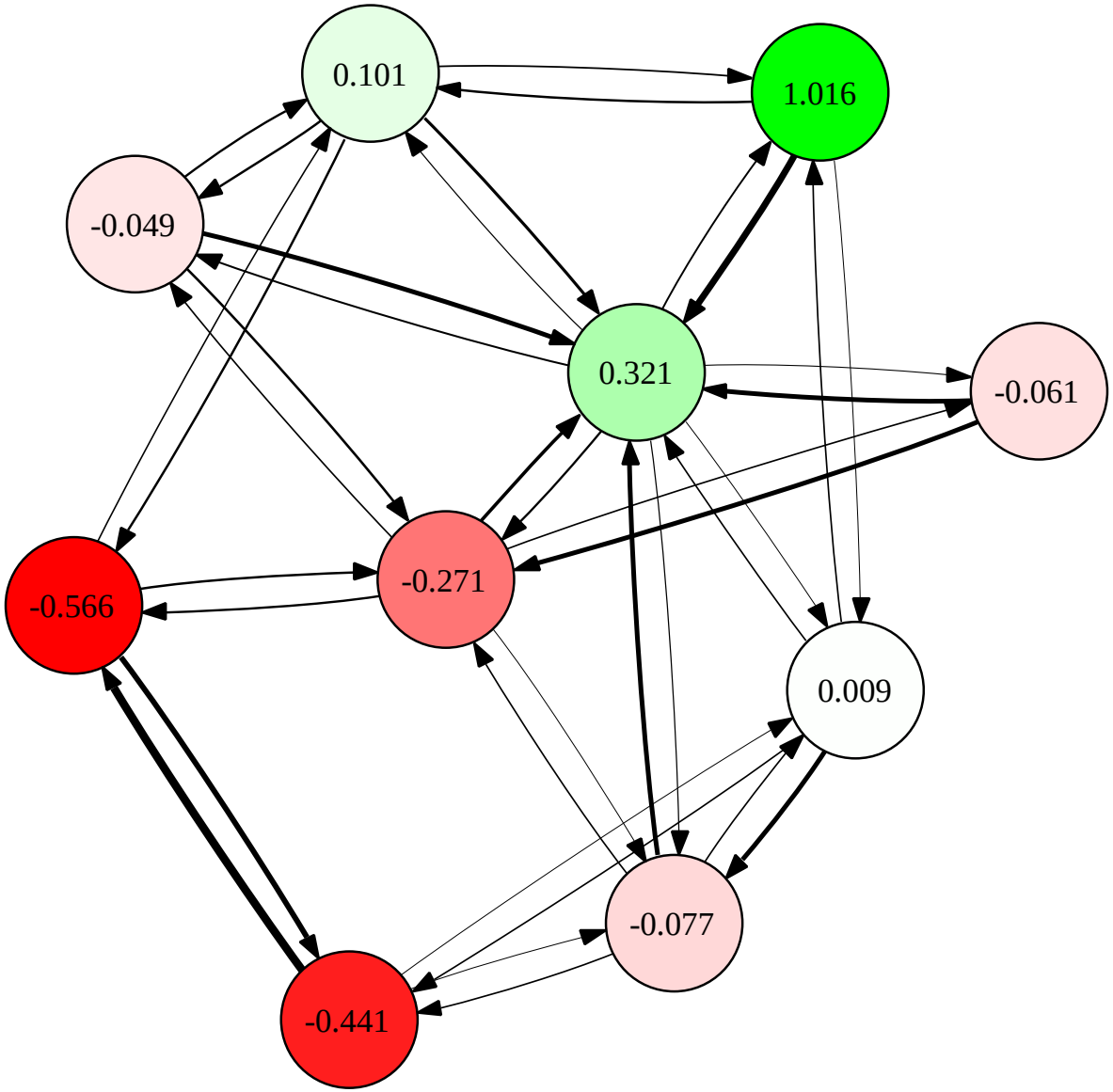
SymGreen — Iteration ♯17



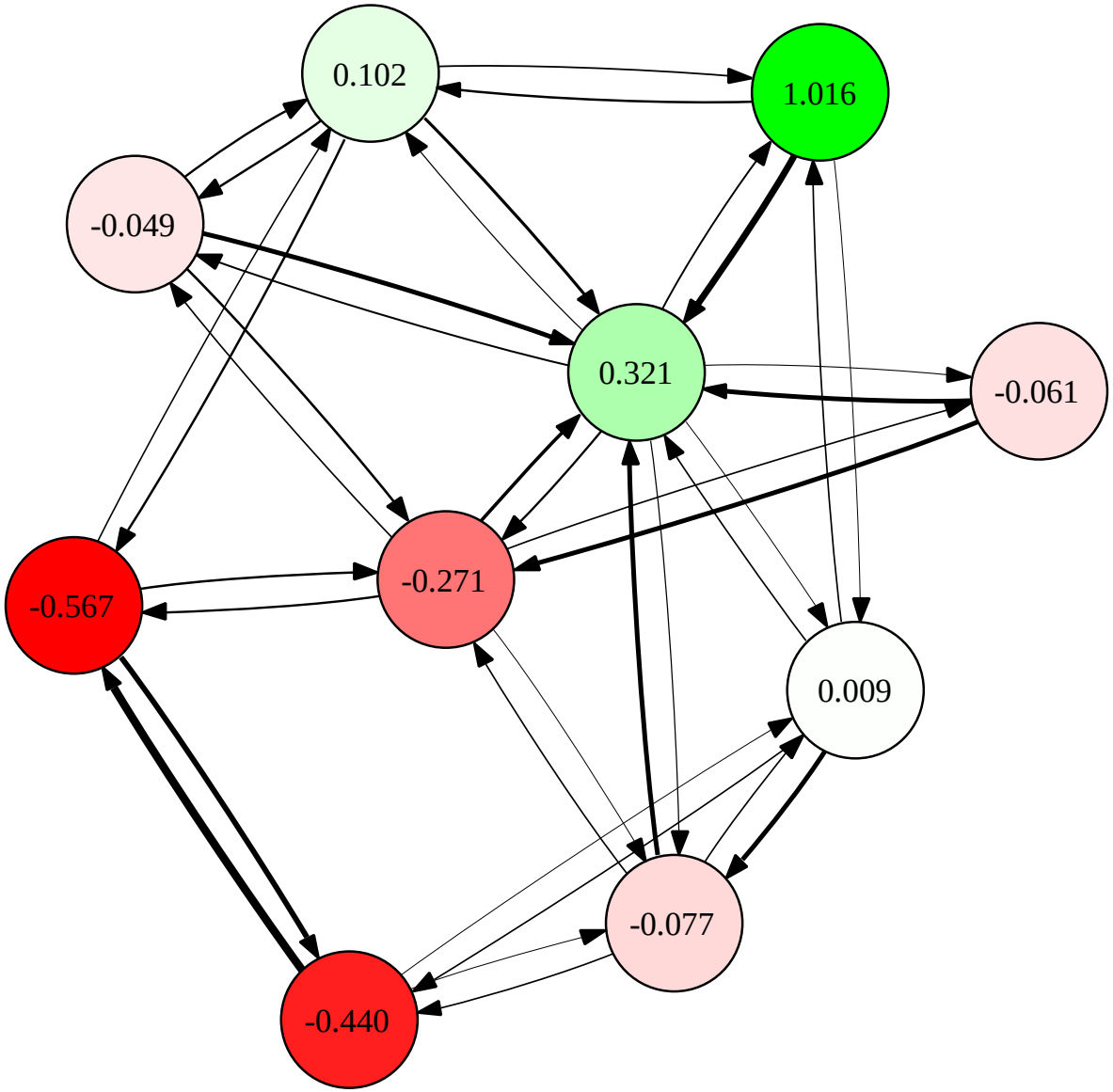
SymGreen — Iteration ♯18



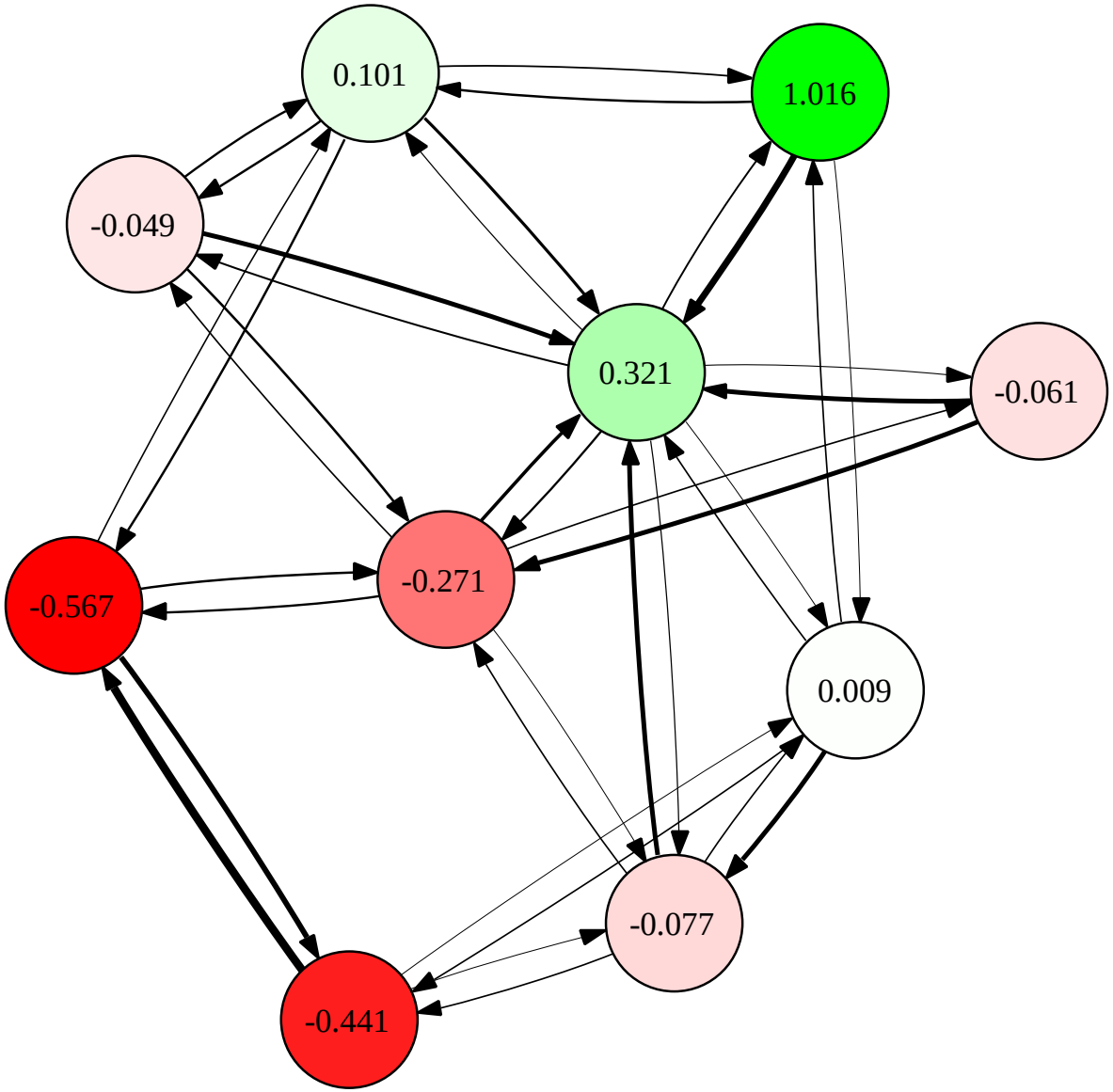
SymGreen — Iteration ♯19



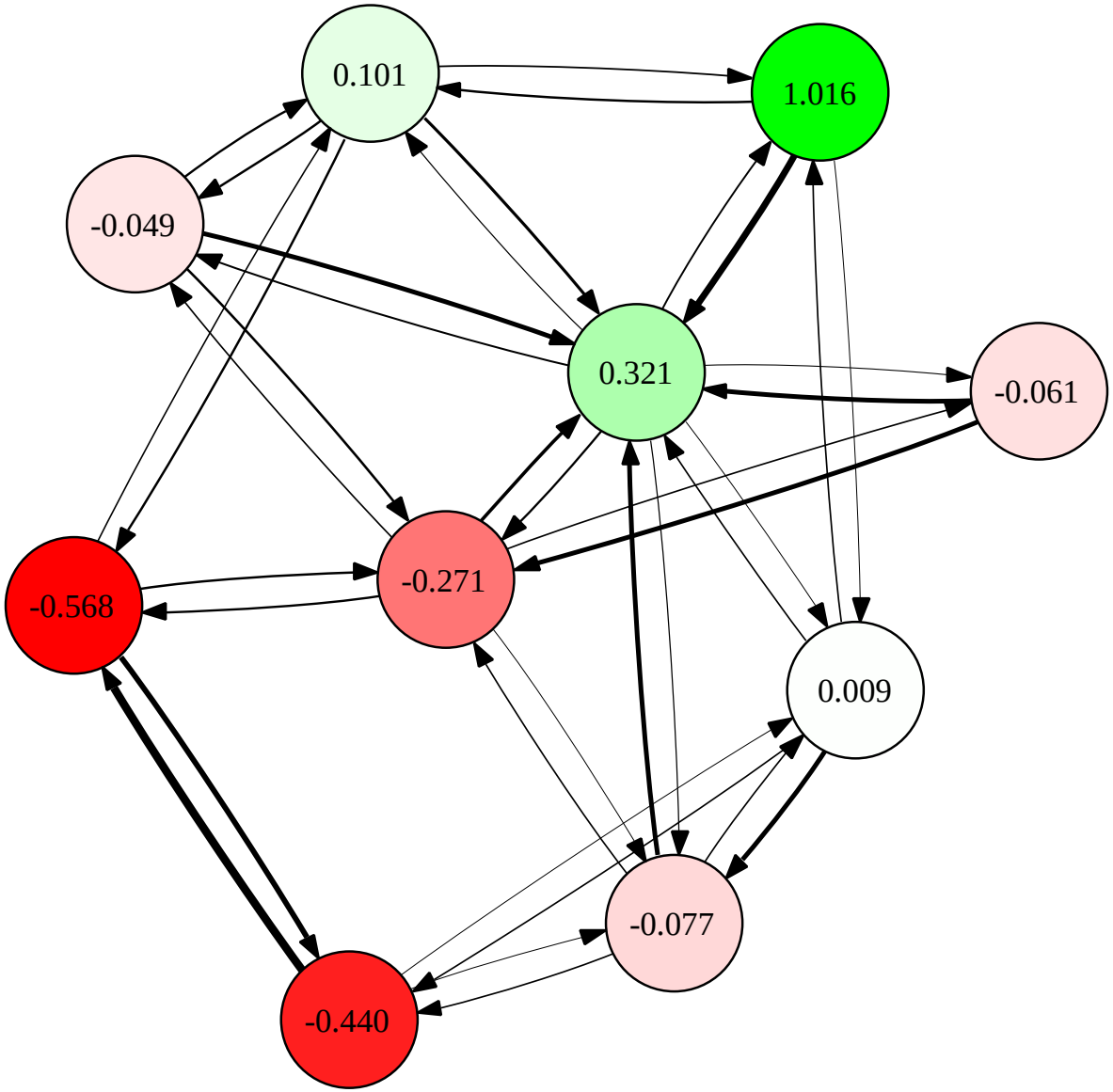
SymGreen — Iteration ♯20



SymGreen — Iteration ♯21



SymGreen — Iteration ♯22



PageRankOfLinks — Method Description

Method Description

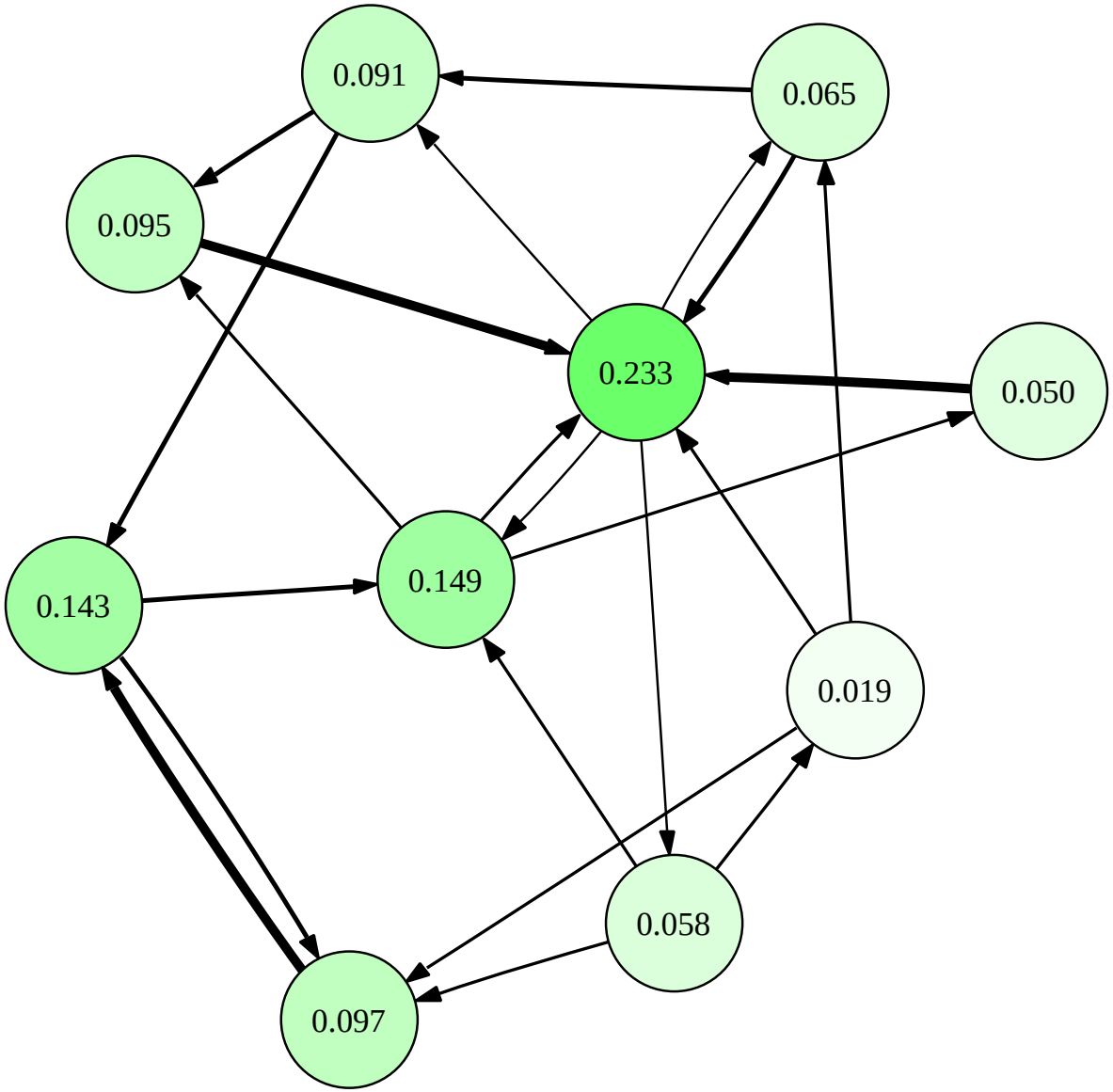
- Related nodes: nodes pointed by i.

PageRankOfLinks — Method Description

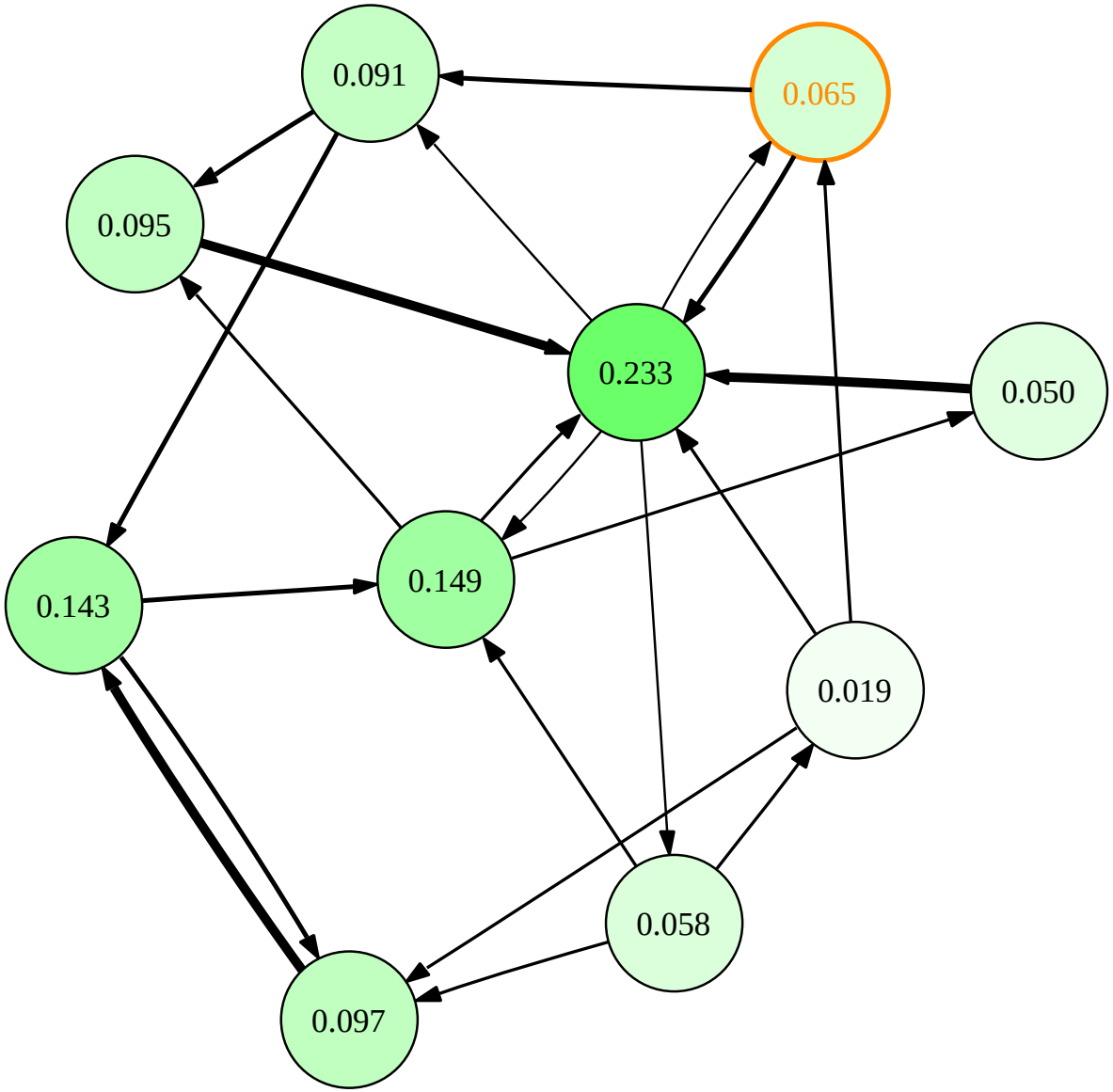
Method Description

- Related nodes: nodes pointed by i.
- Sort the nodes pointed by i by their equilibrium measure value.

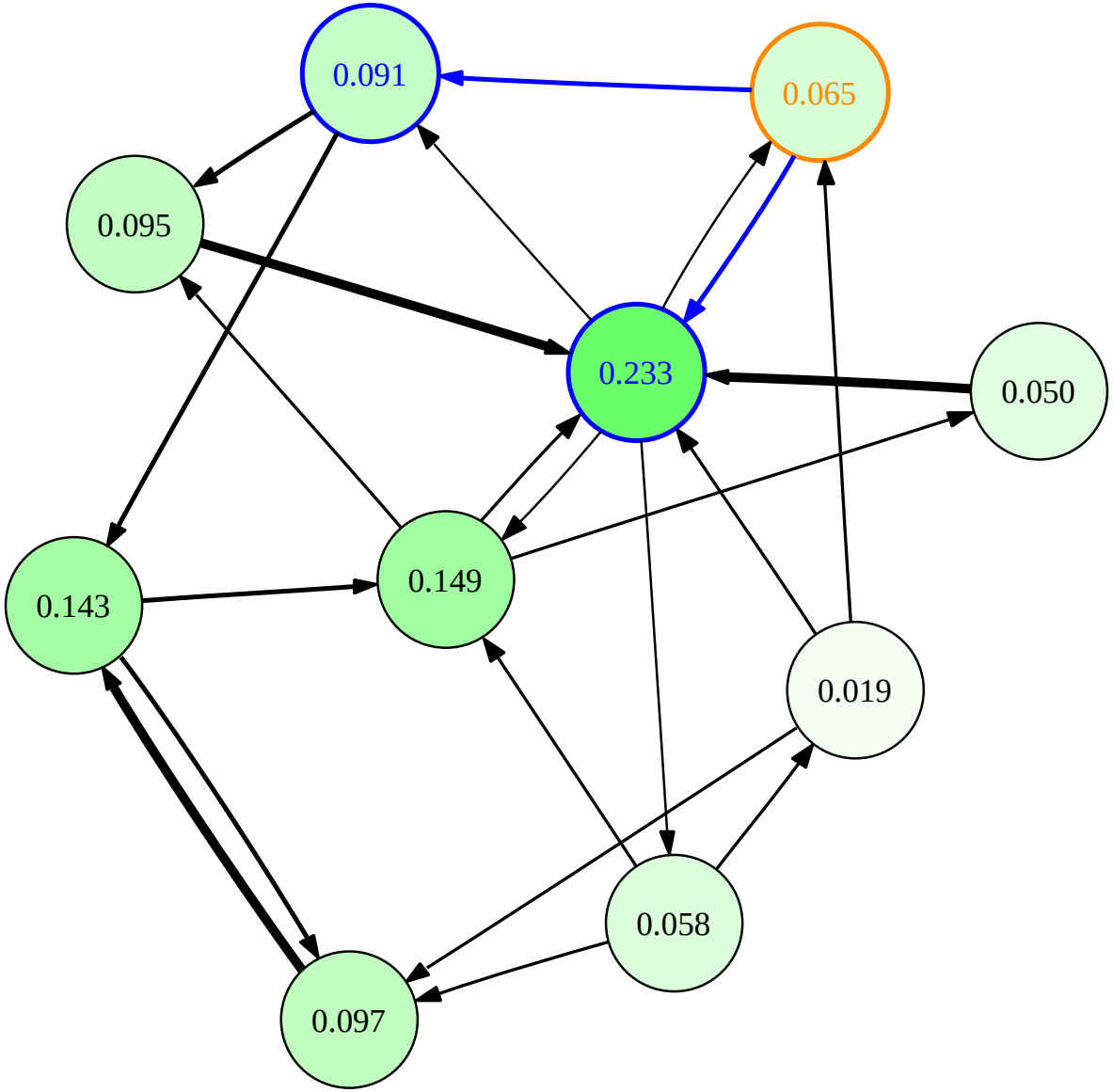
PageRankOfLinks



PageRankOfLinks



PageRankOfLinks



LocalPageRank — Method Description

Method Description

- Compute a neighborhood subgraph of i.

LocalPageRank — Method Description

Method Description

- Compute a neighborhood subgraph of i.
- Compute the strongly connected component of this subgraph containing i.

LocalPageRank — Method Description

Method Description

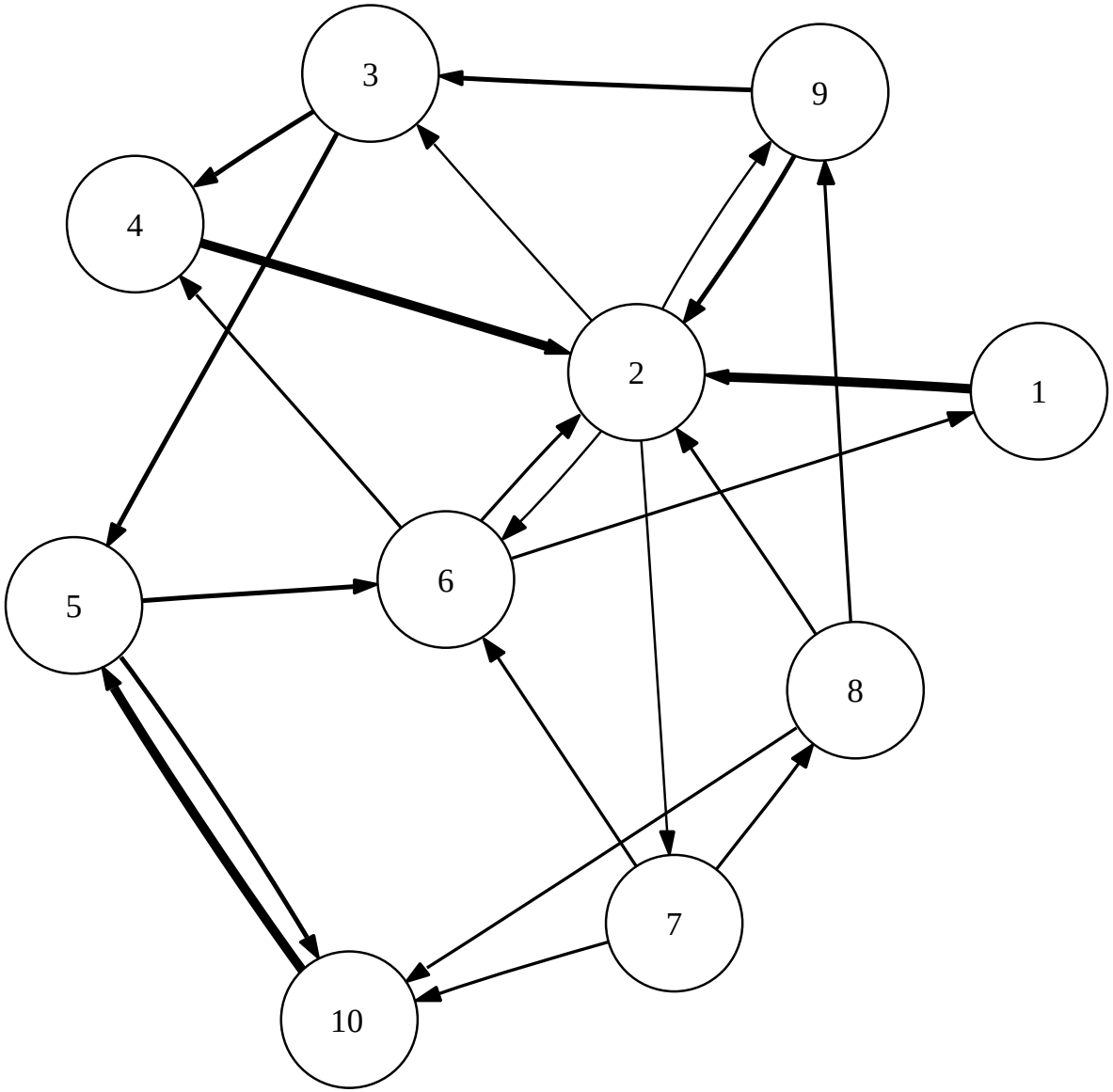
- Compute a neighborhood subgraph of i.
- Compute the strongly connected component of this subgraph containing i.
- Order the nodes by their equilibrium measure in this component.

LocalPageRank — Method Description

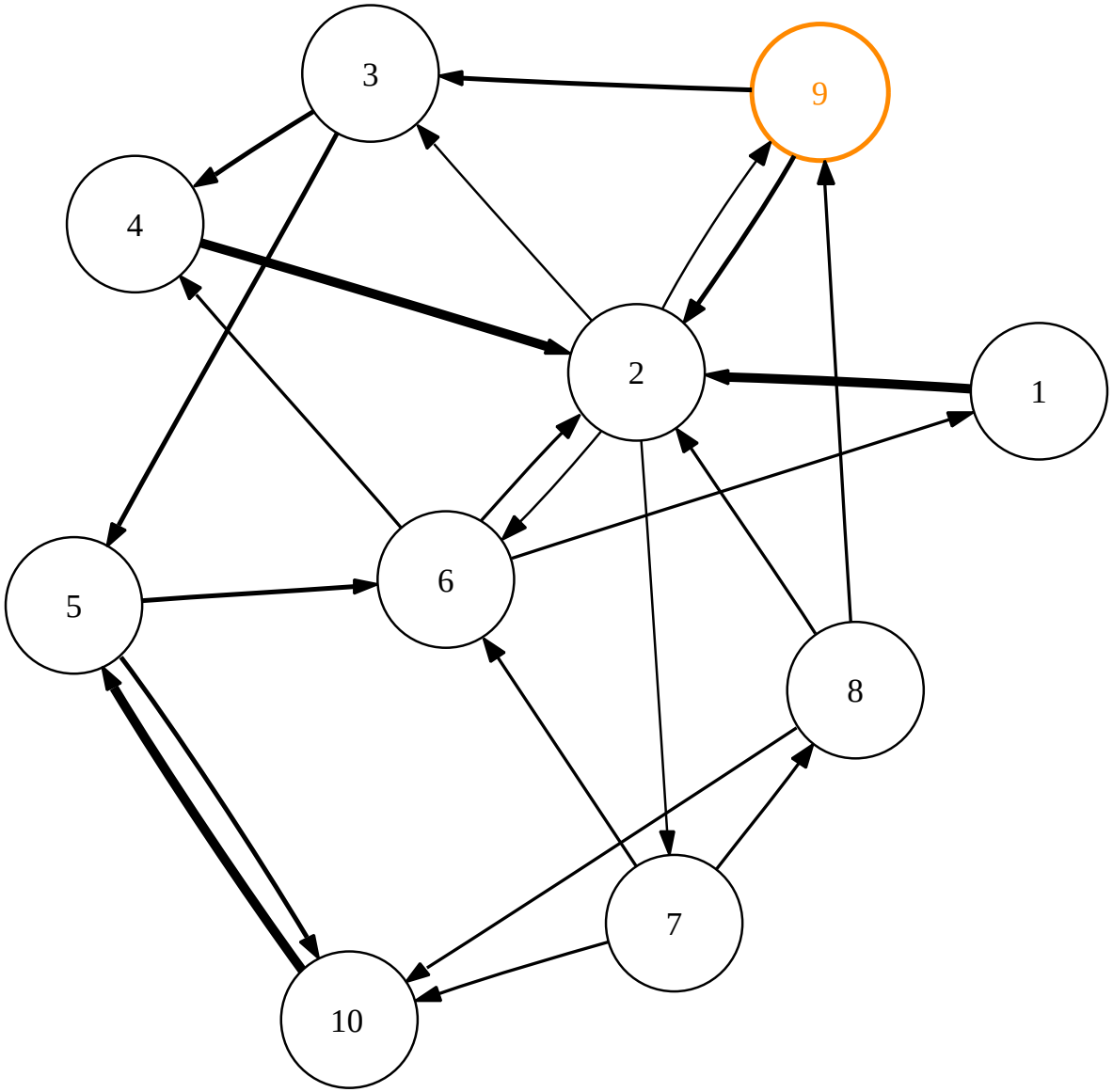
Method Description

- Compute a neighborhood subgraph of i.
- Compute the strongly connected component of this subgraph containing i.
- Order the nodes by their equilibrium measure in this component.
- (Approach widely used on the World Wide Web).

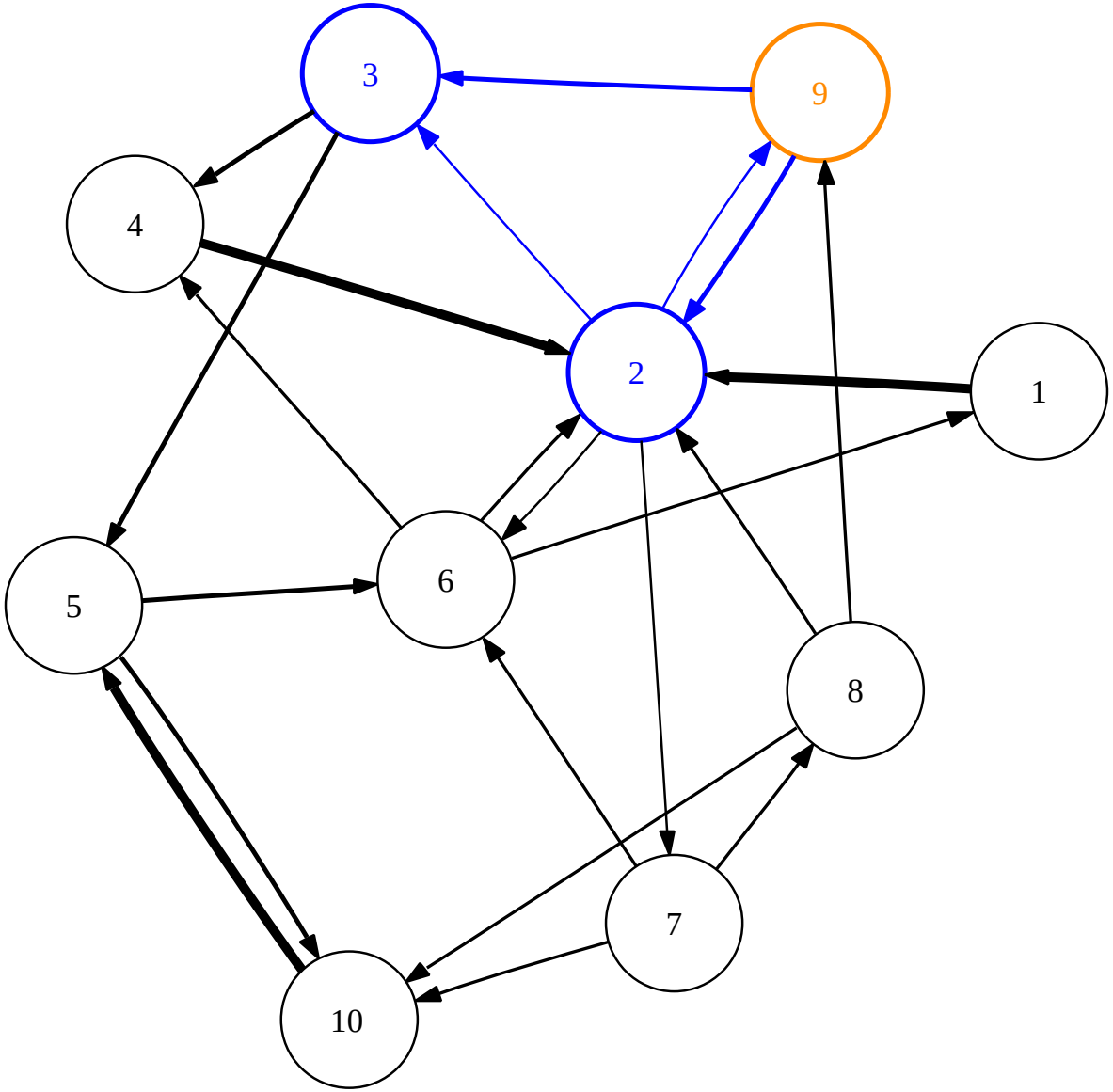
LocalPageRank — Original Graph



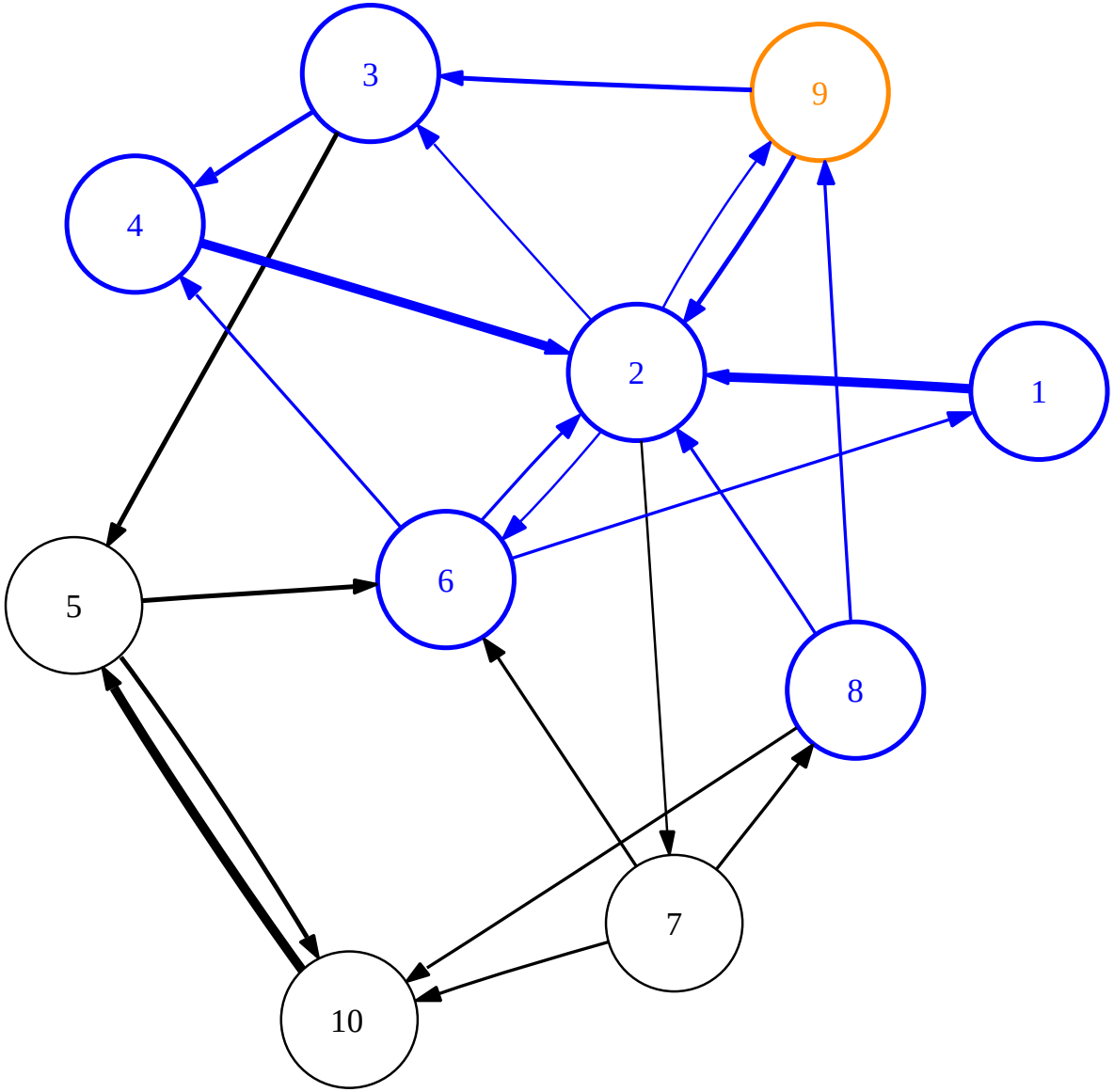
LocalPageRank — Local Subgraph



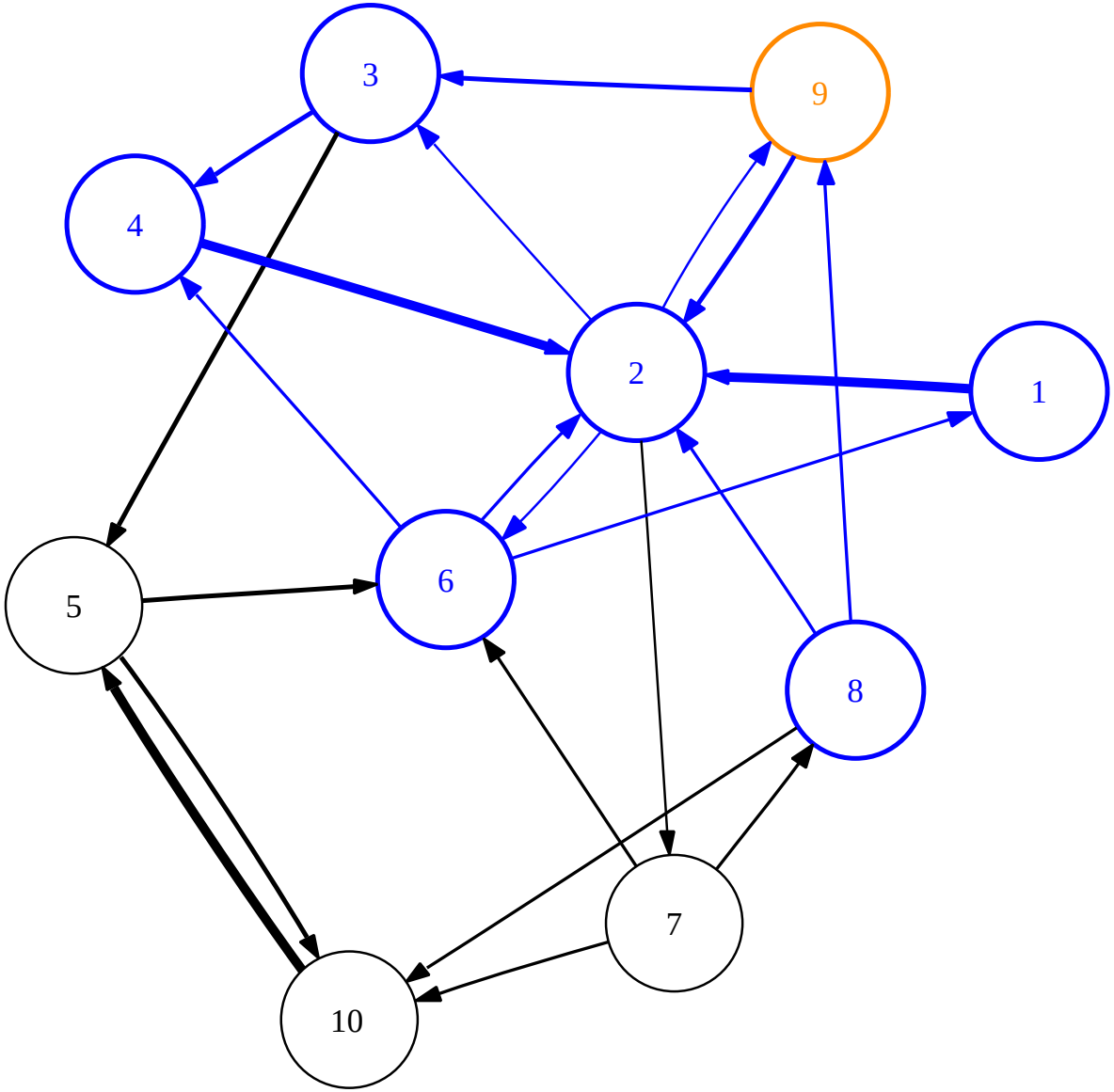
LocalPageRank — Local Subgraph



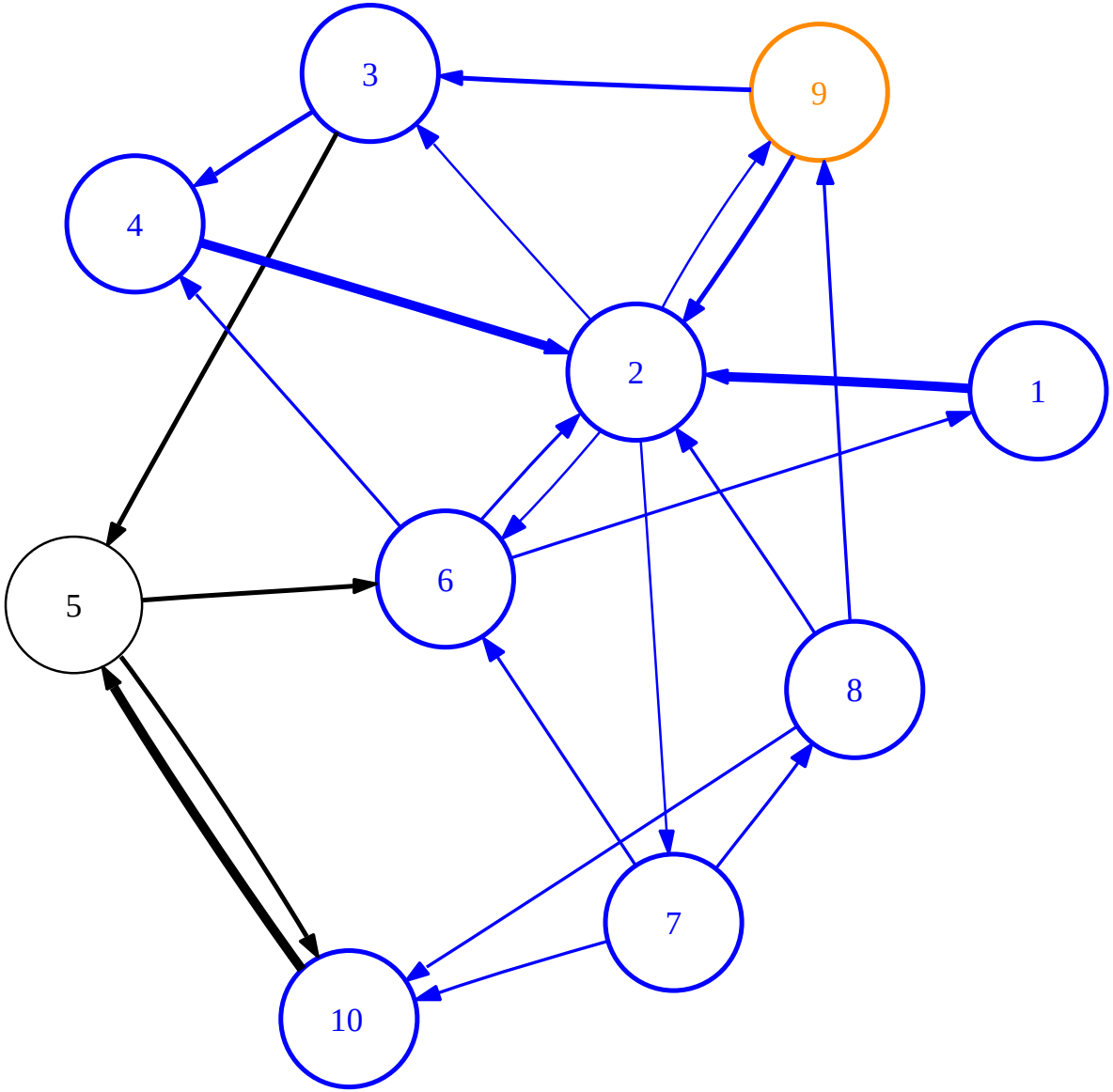
LocalPageRank — Local Subgraph



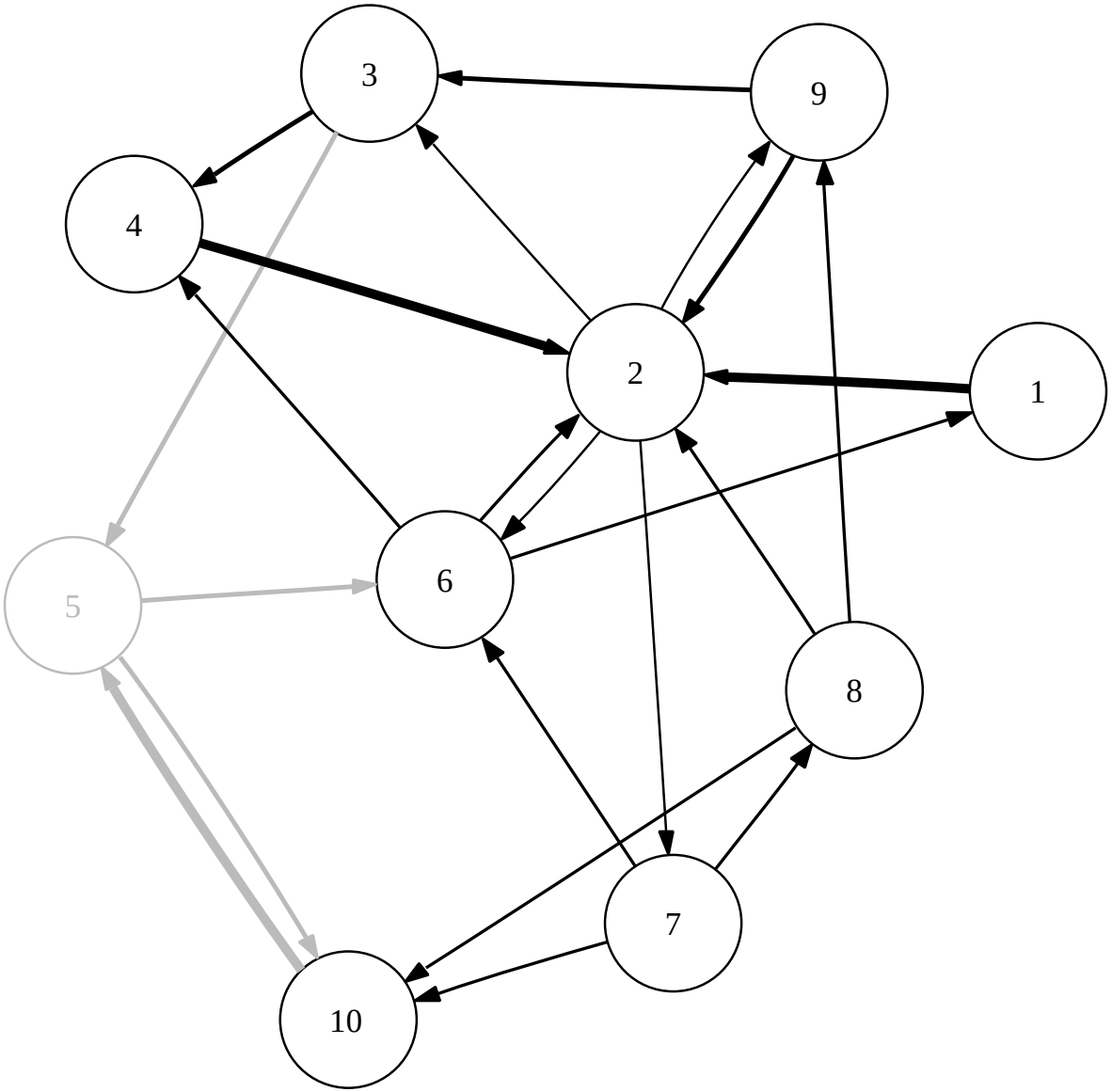
LocalPageRank — Local Subgraph



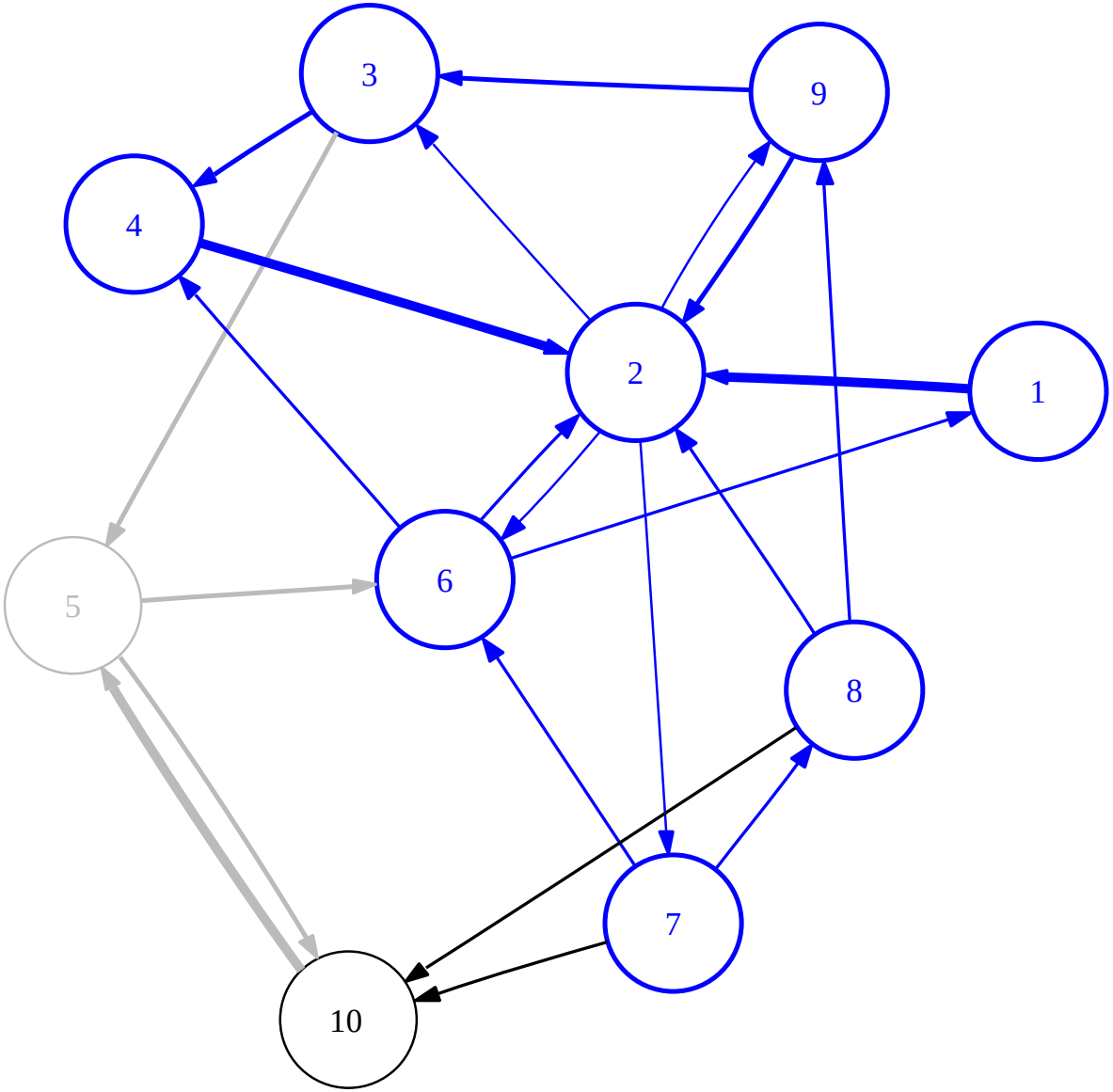
LocalPageRank — Local Subgraph



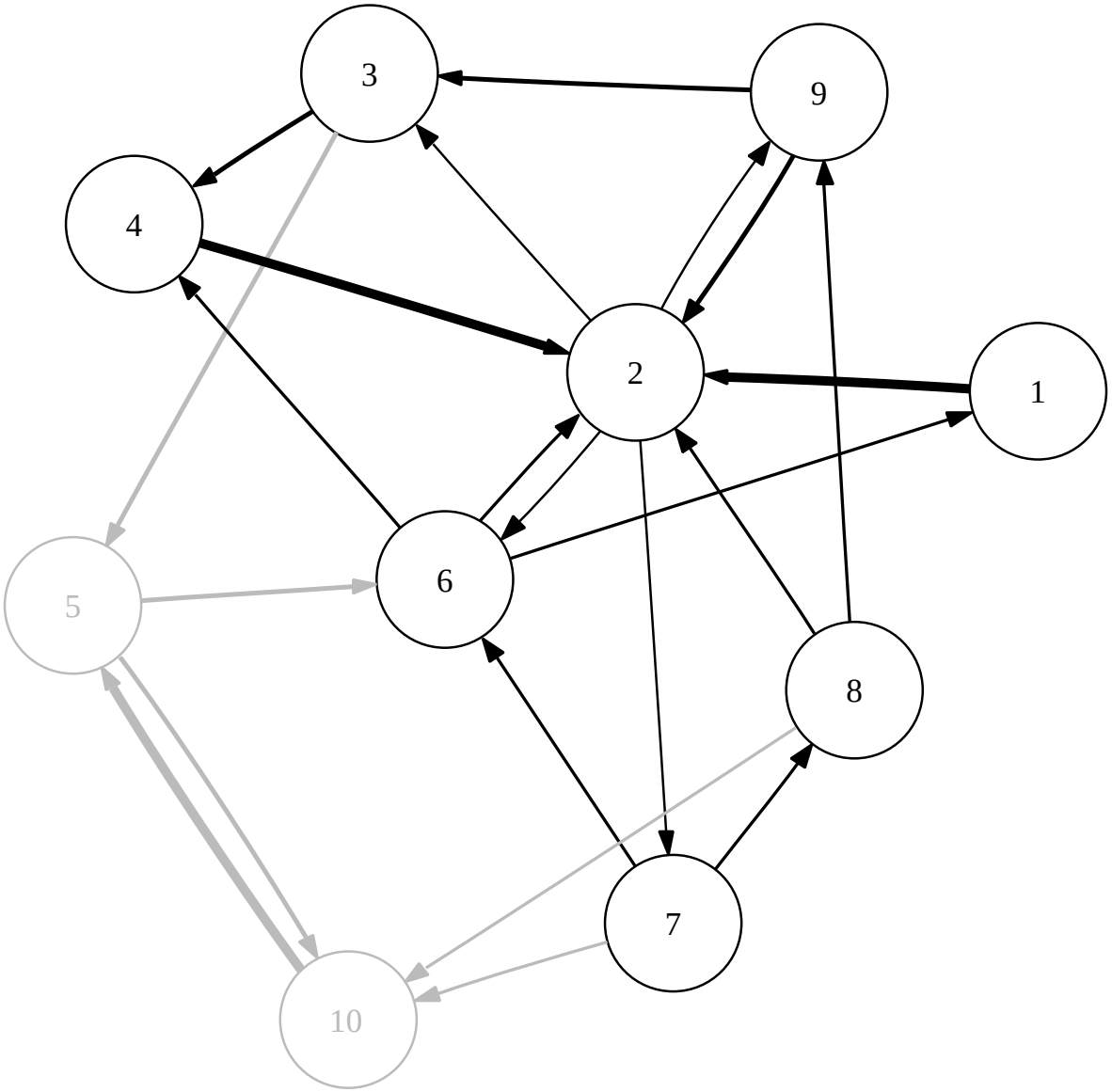
LocalPageRank — Local Subgraph



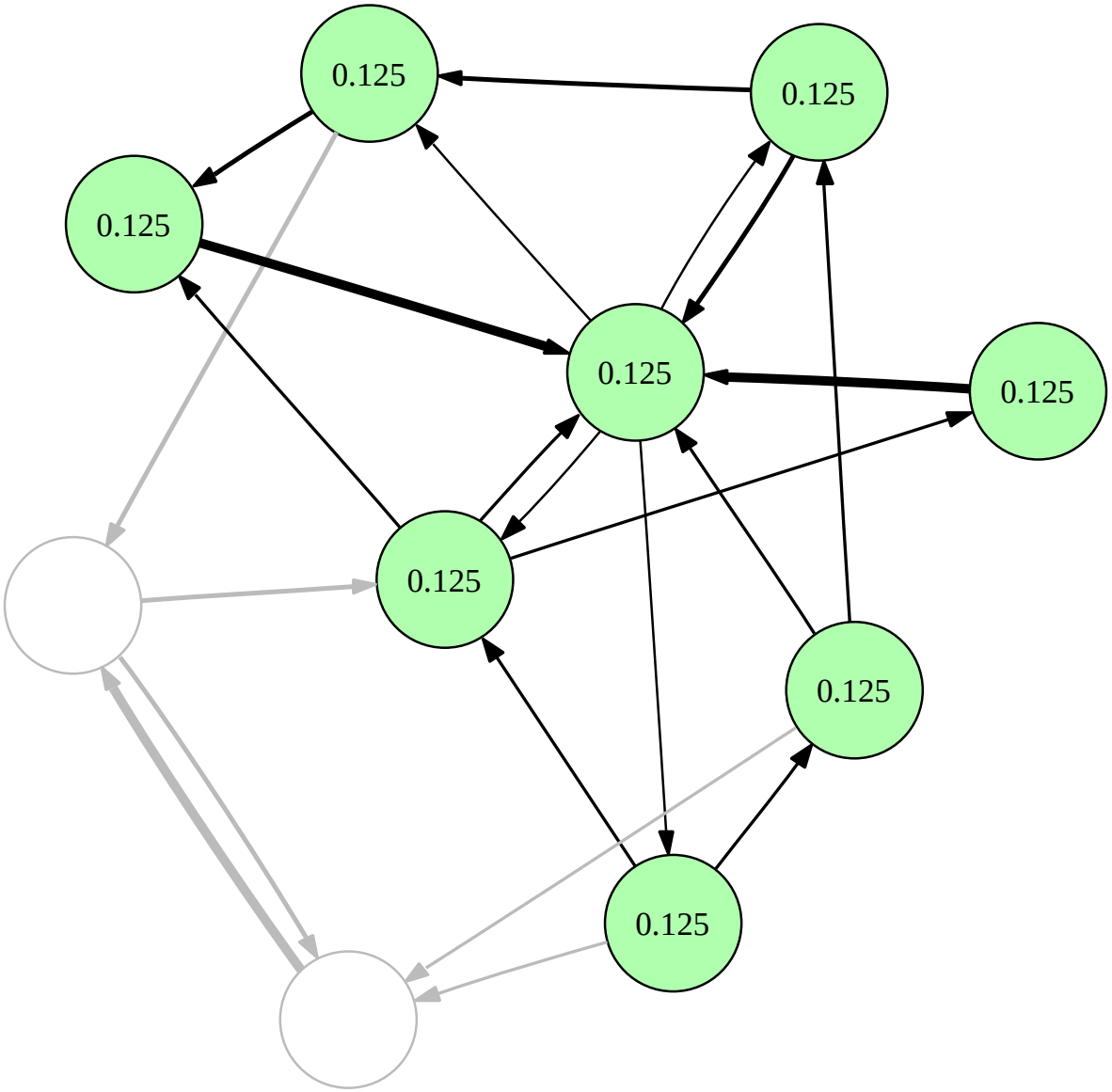
LocalPageRank — Local Strongly Connected Subgraph



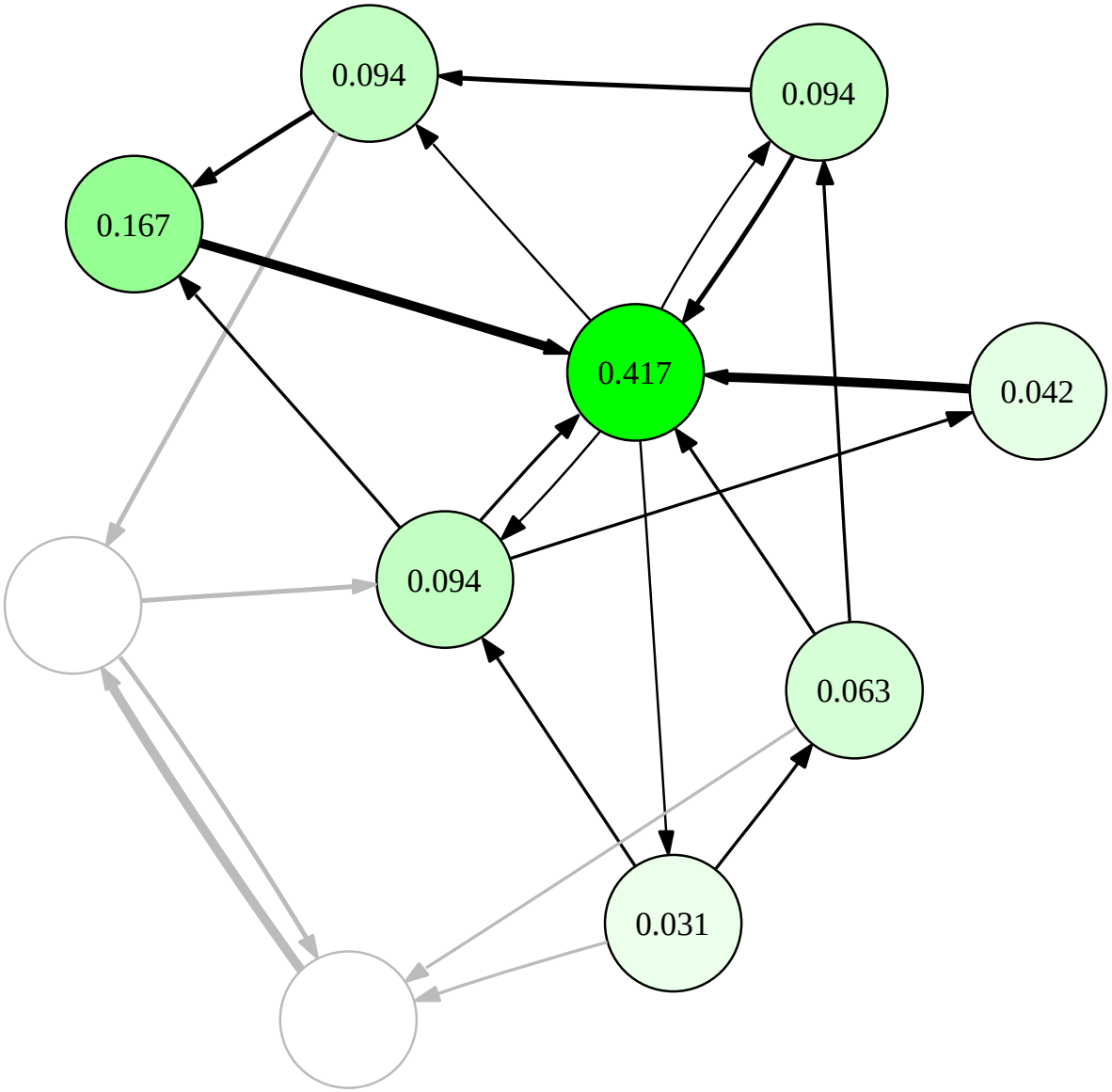
LocalPageRank — Local Strongly Connected Subgraph



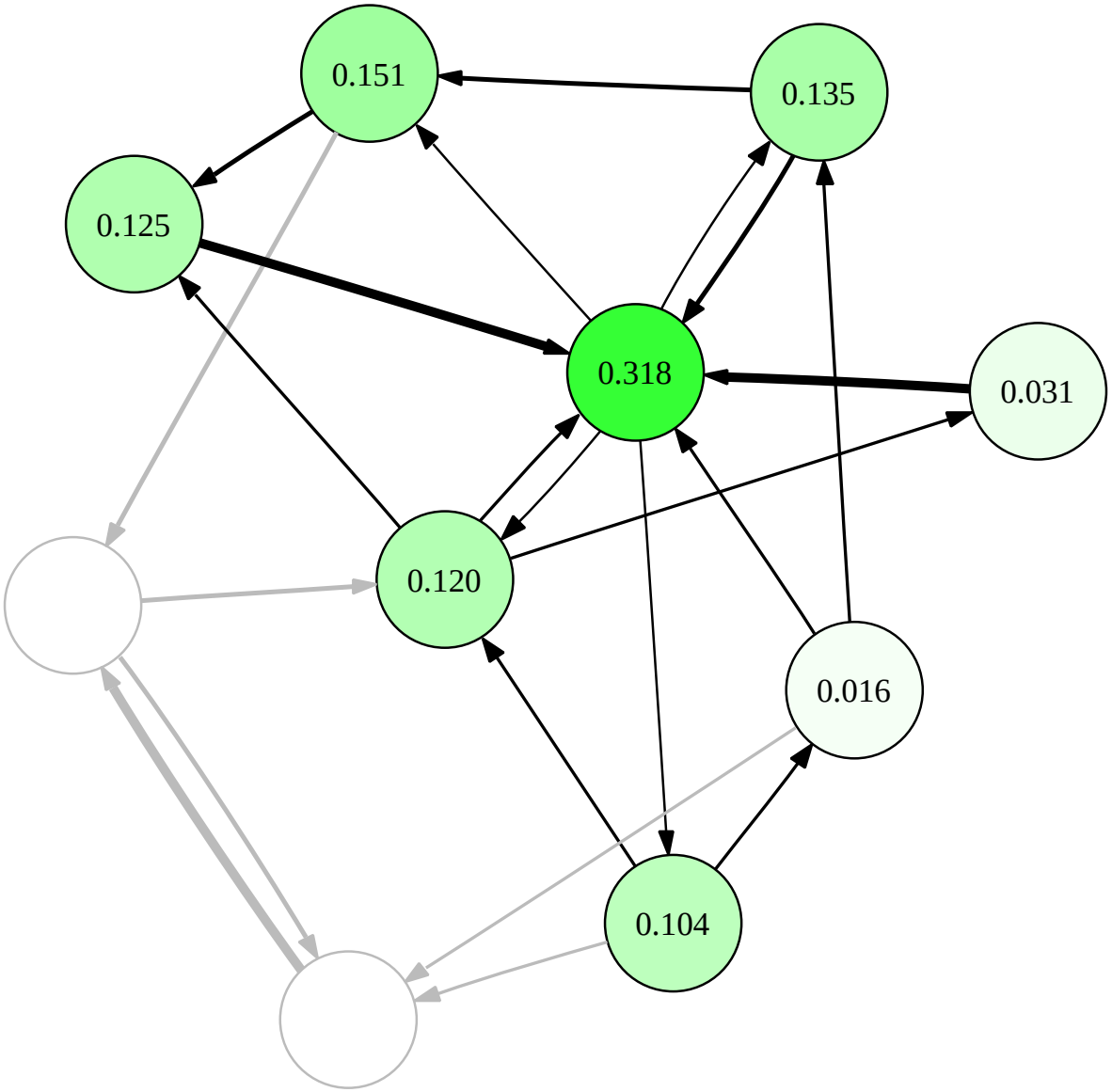
LocalPageRank — Iteration ♯1



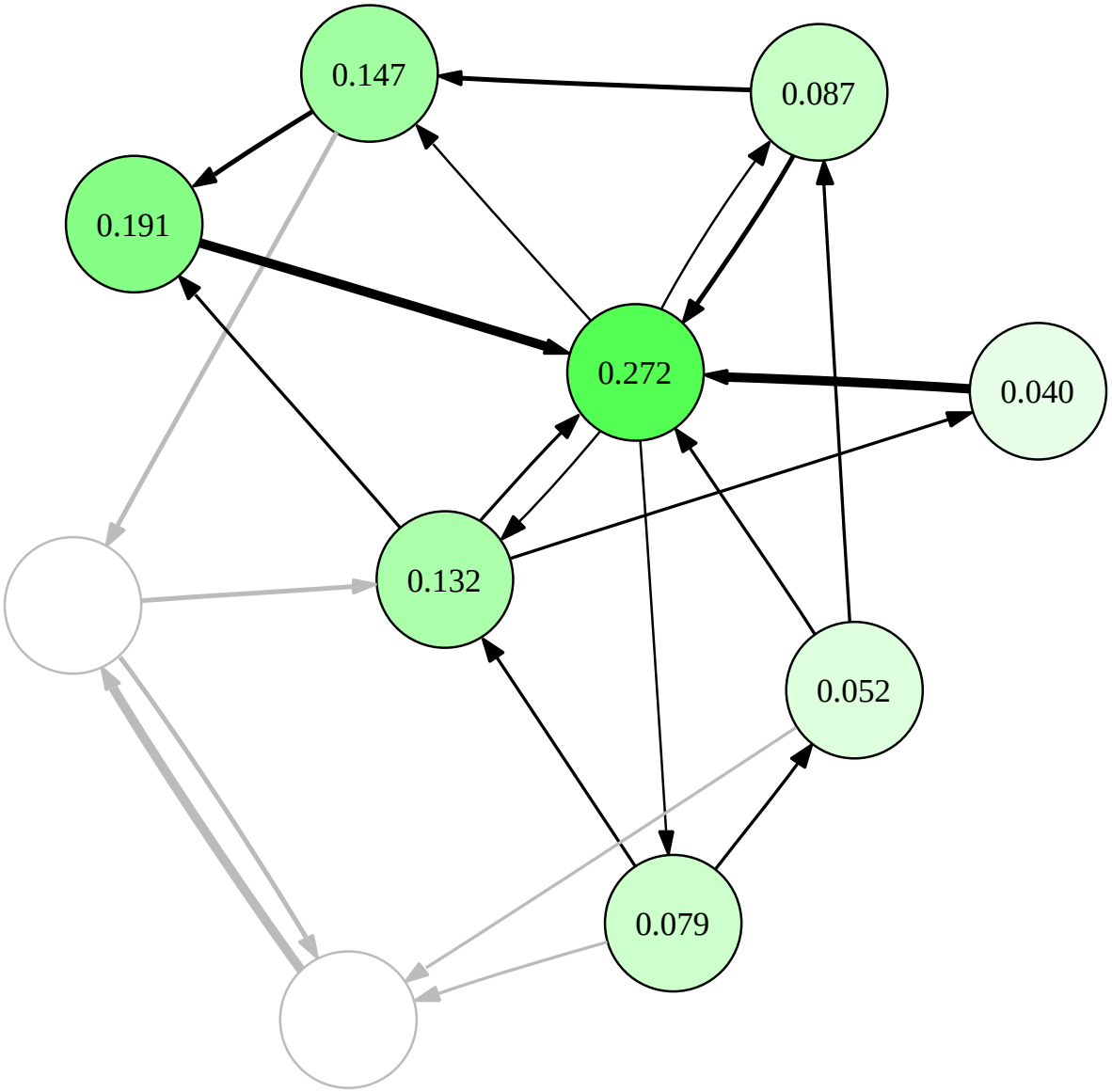
LocalPageRank — Iteration ♯2



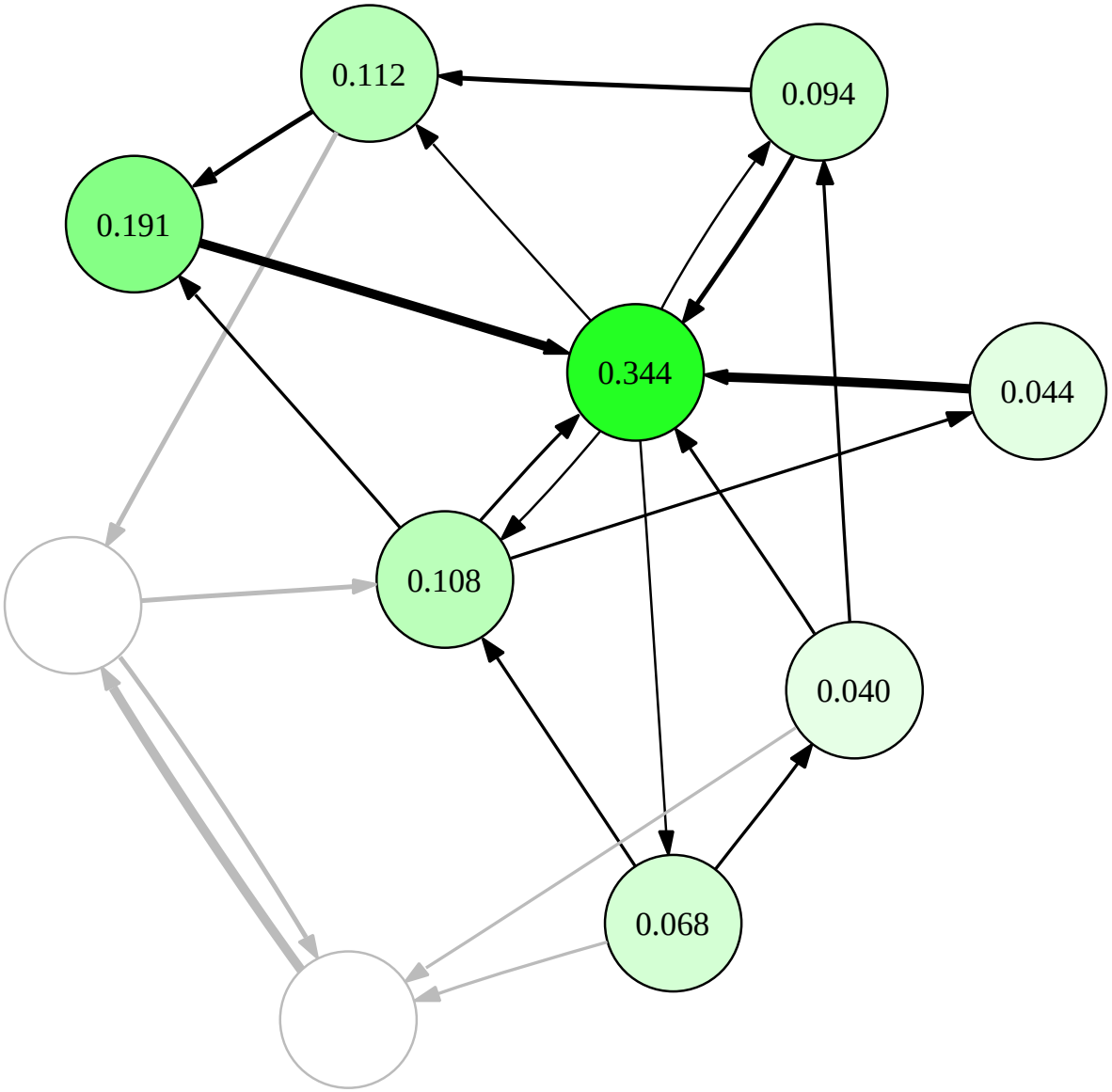
LocalPageRank — Iteration ♯3



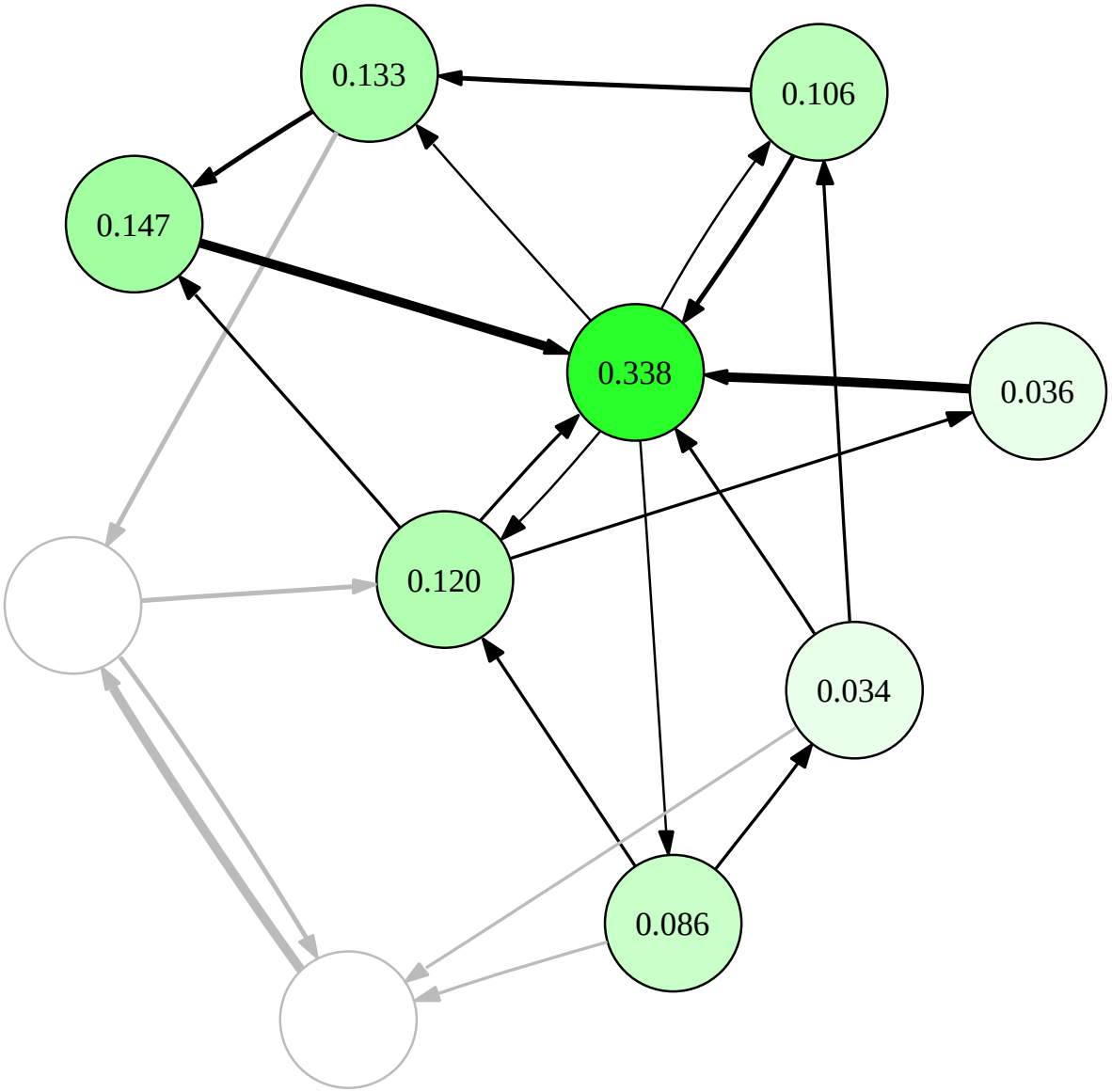
LocalPageRank — Iteration ♯4



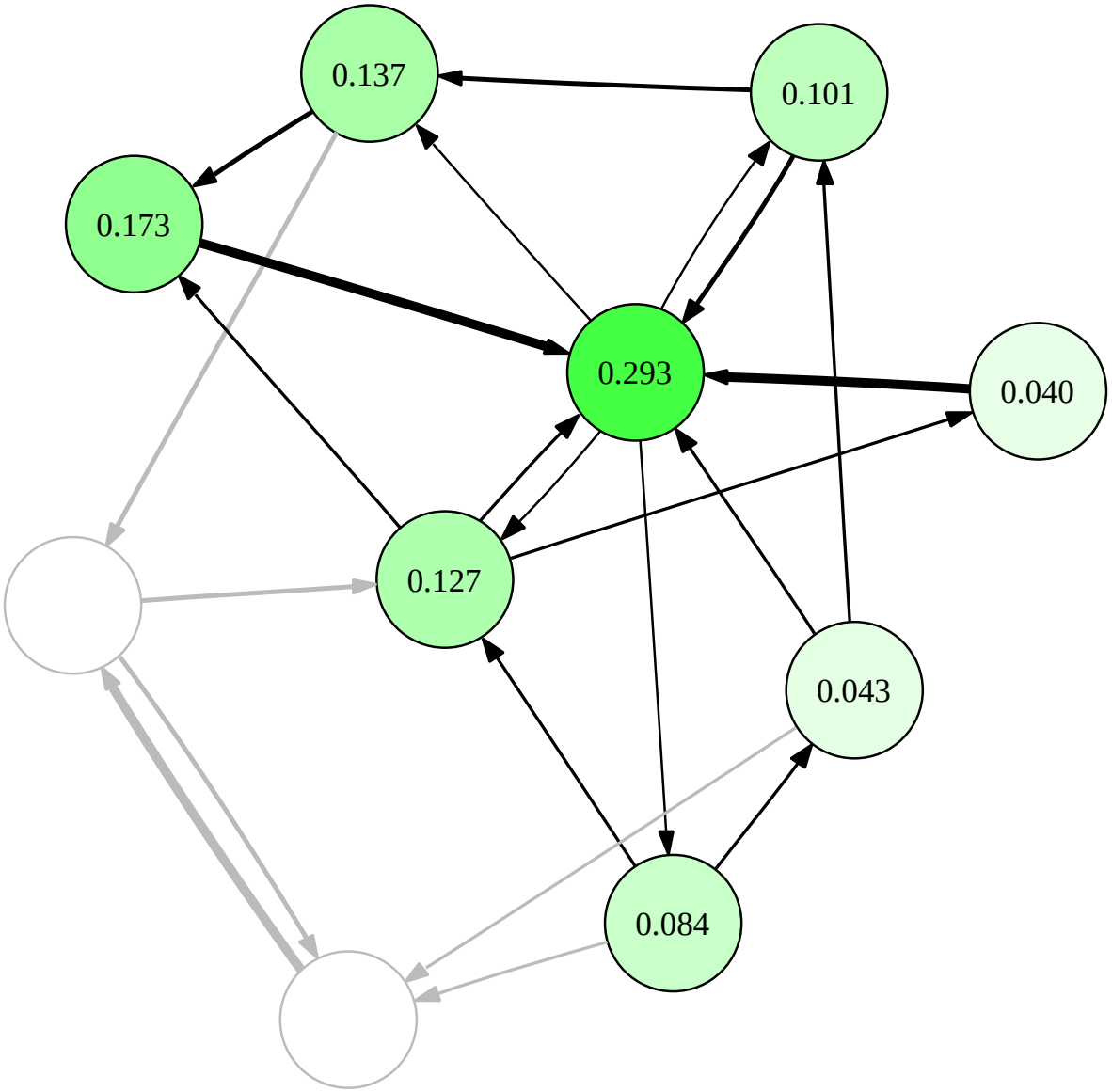
LocalPageRank — Iteration ♯5



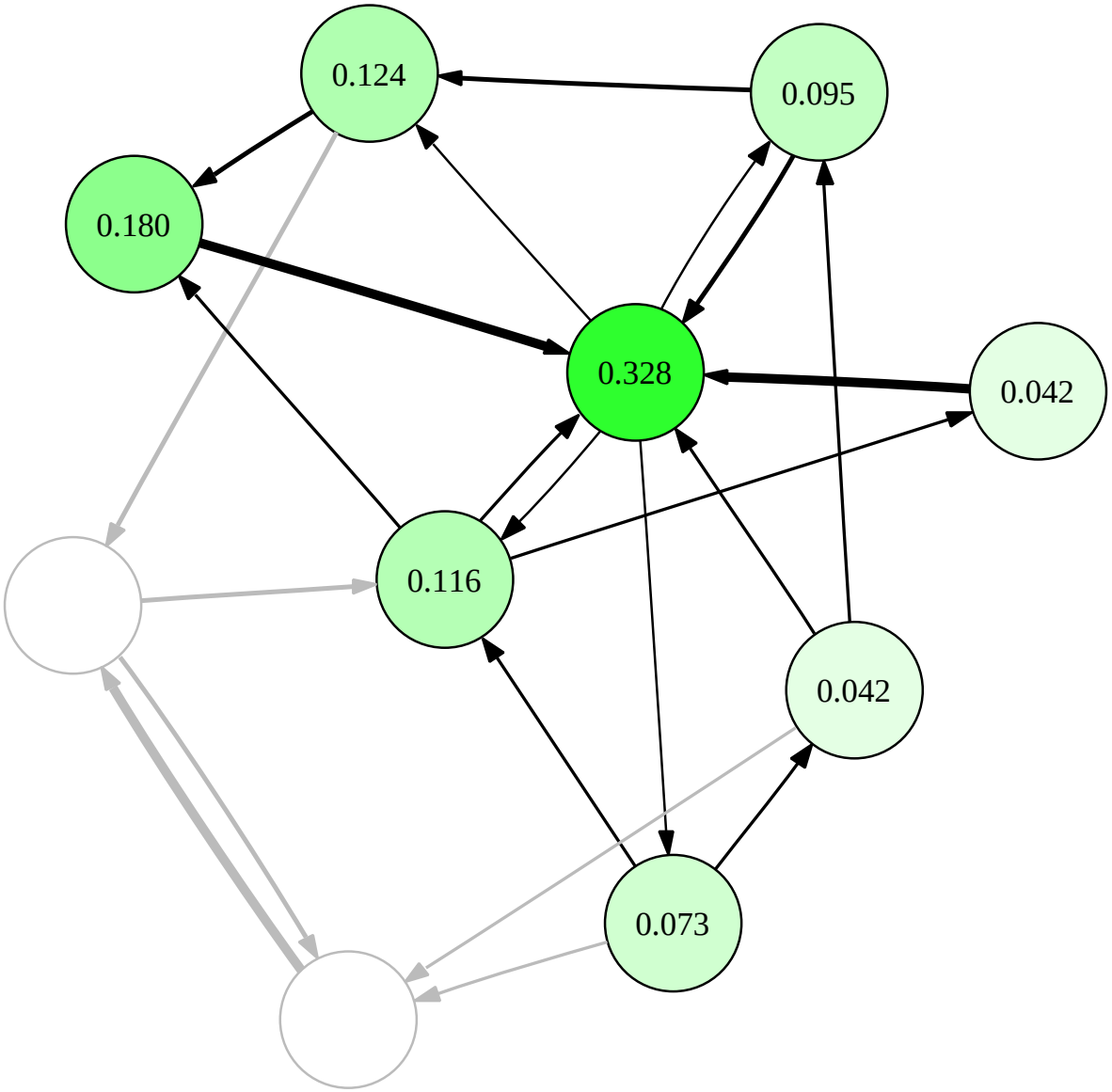
LocalPageRank — Iteration ♯6



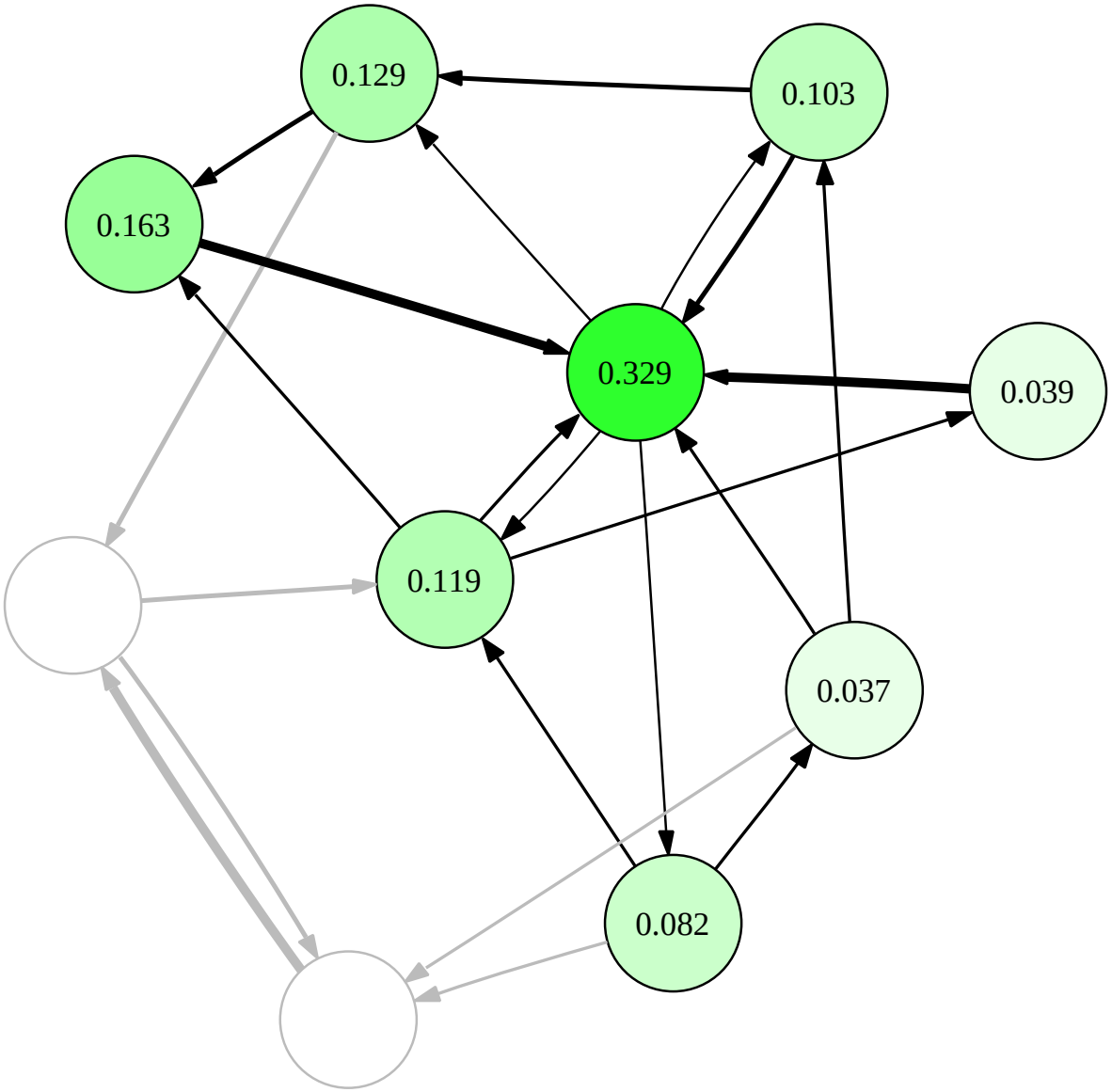
LocalPageRank — Iteration ♯7



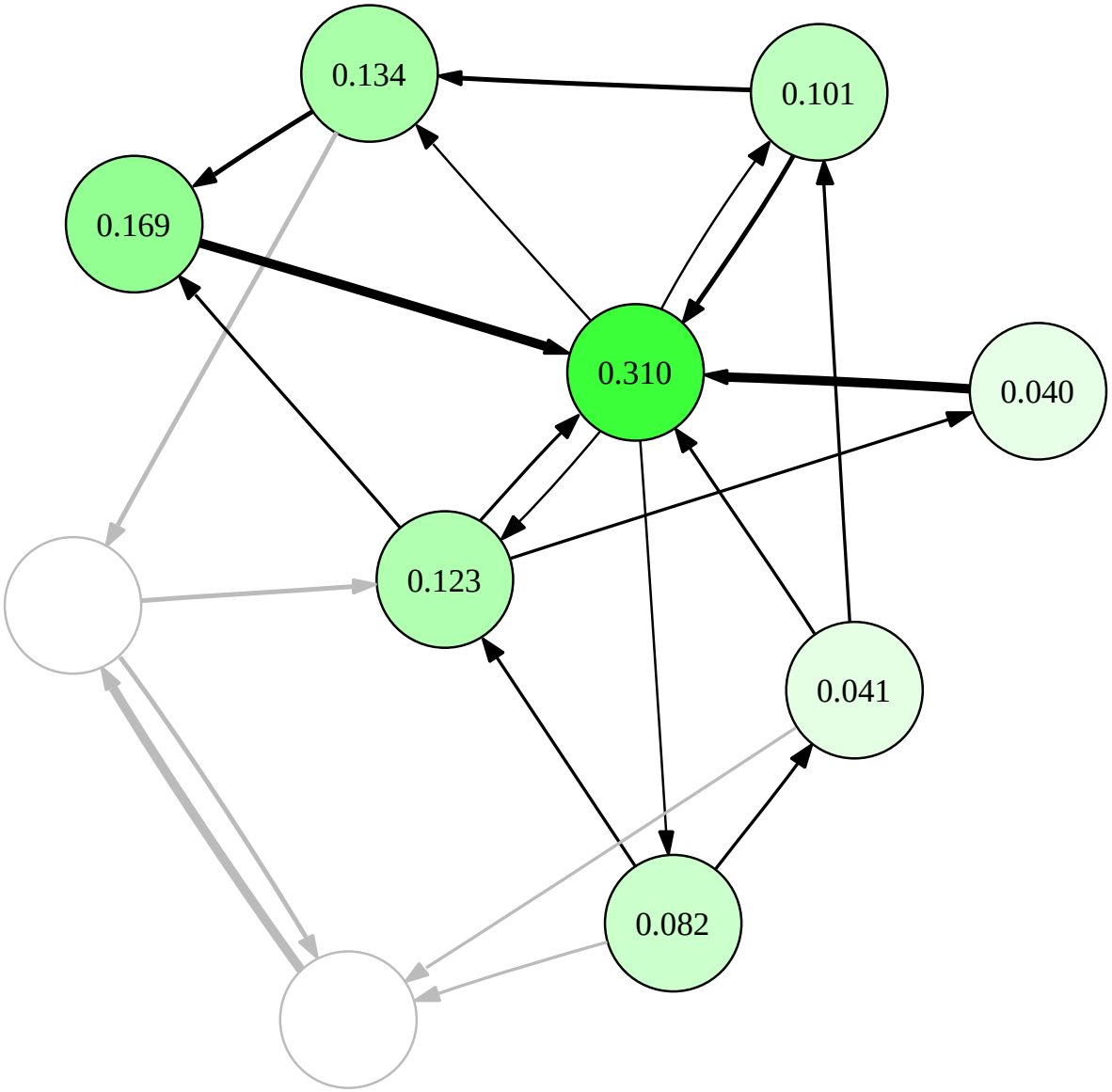
LocalPageRank — Iteration ♯8



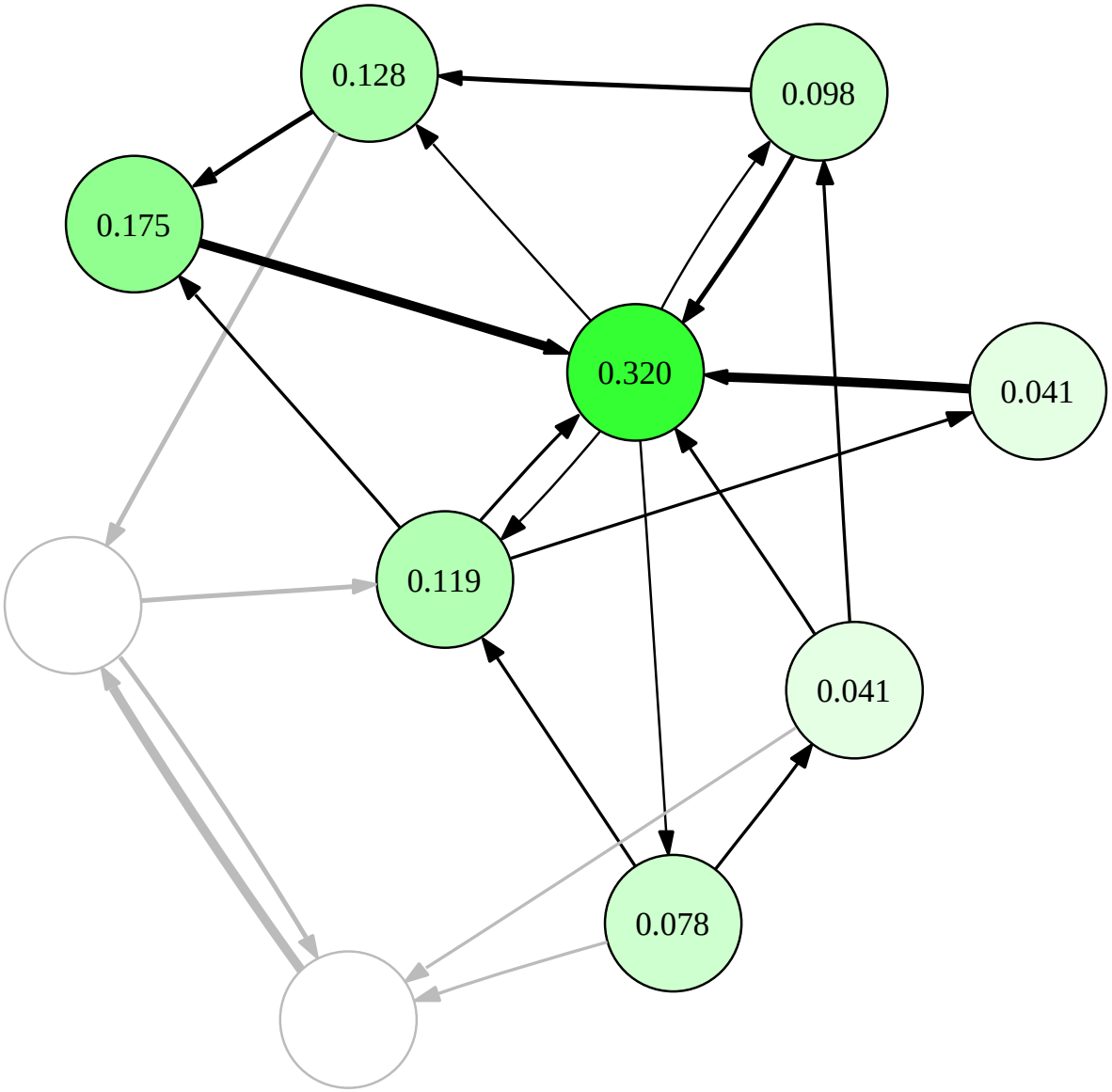
LocalPageRank — Iteration ♯9



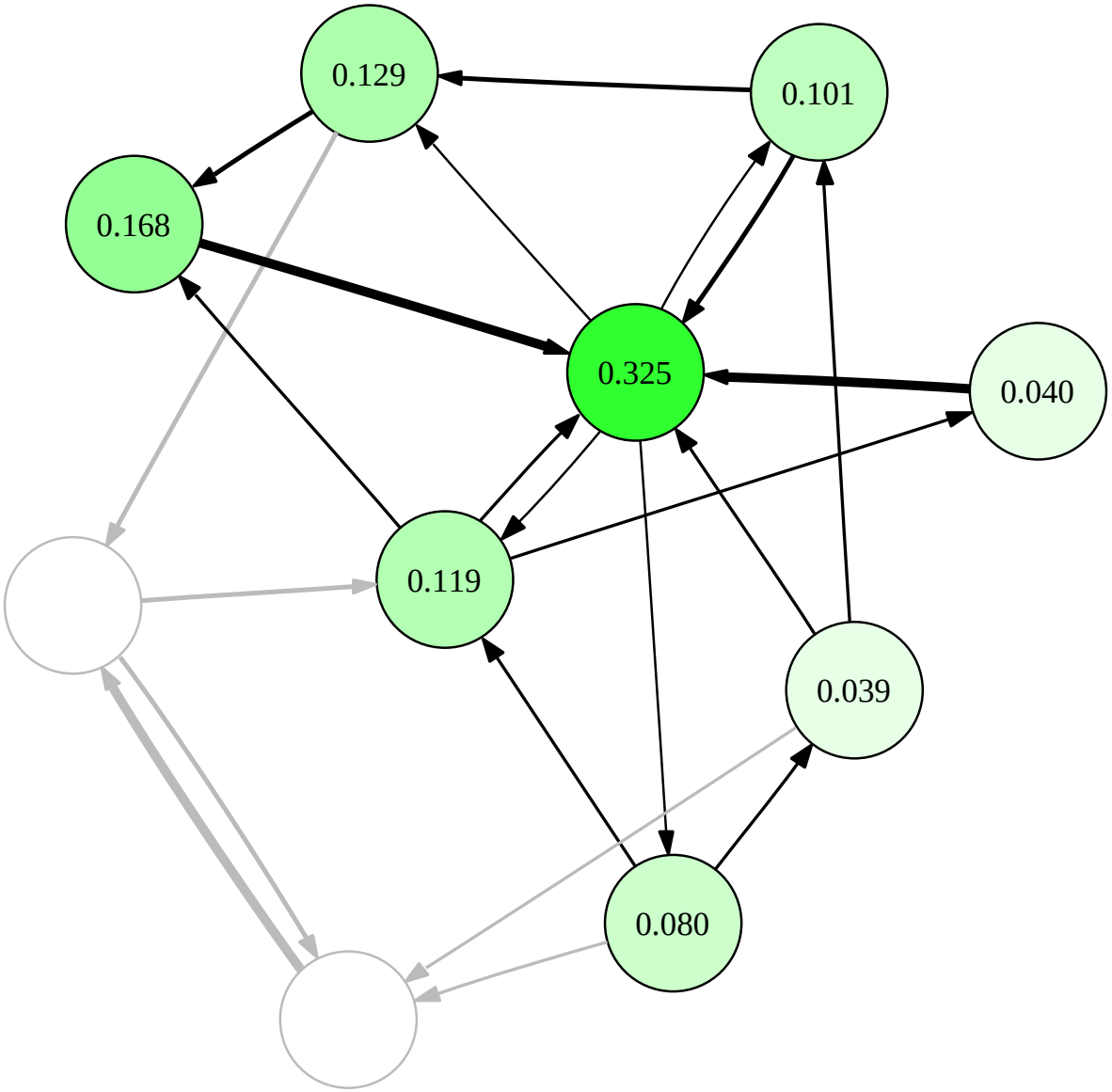
LocalPageRank — Iteration ♯10



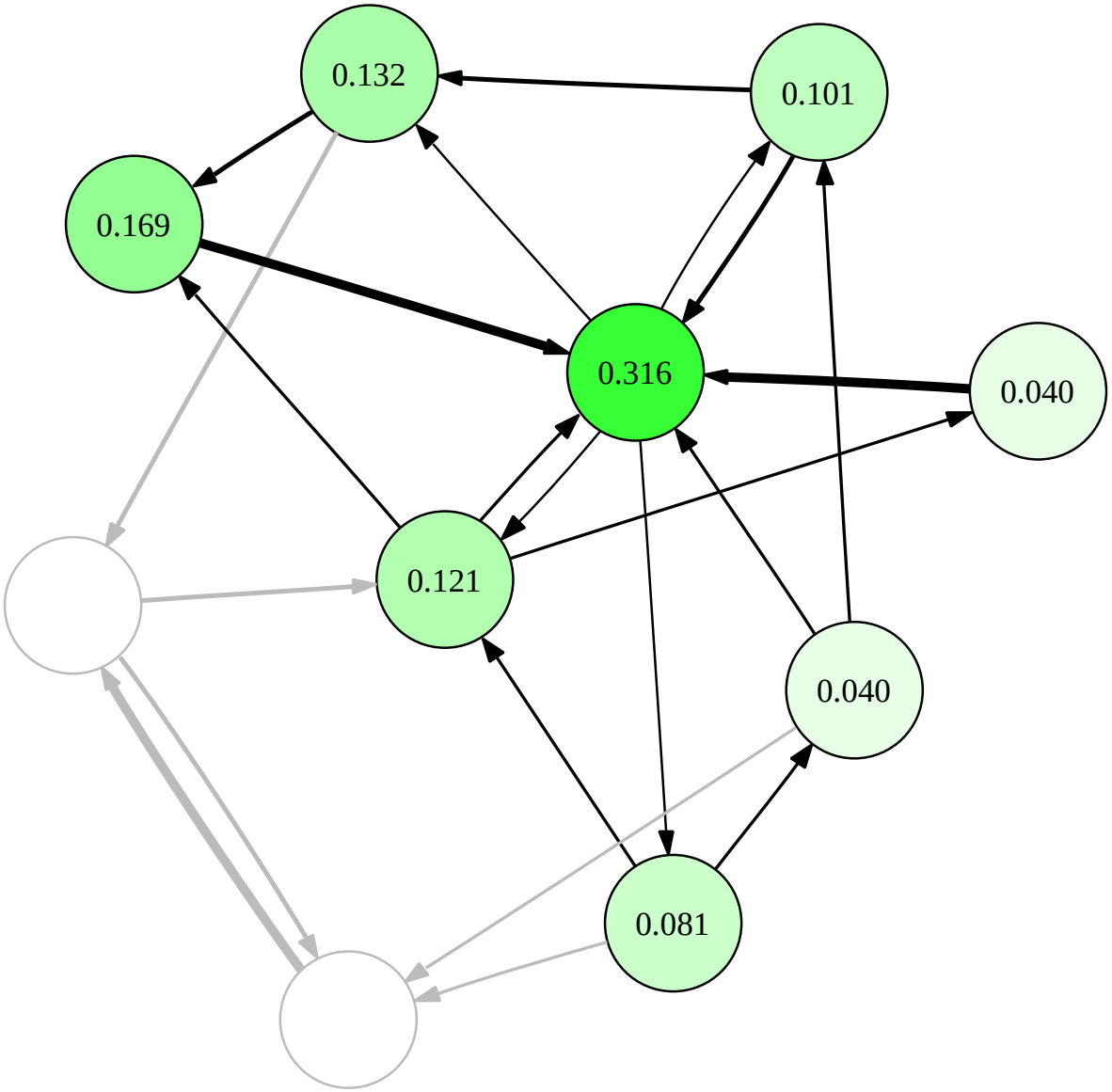
LocalPageRank — Iteration ♯11



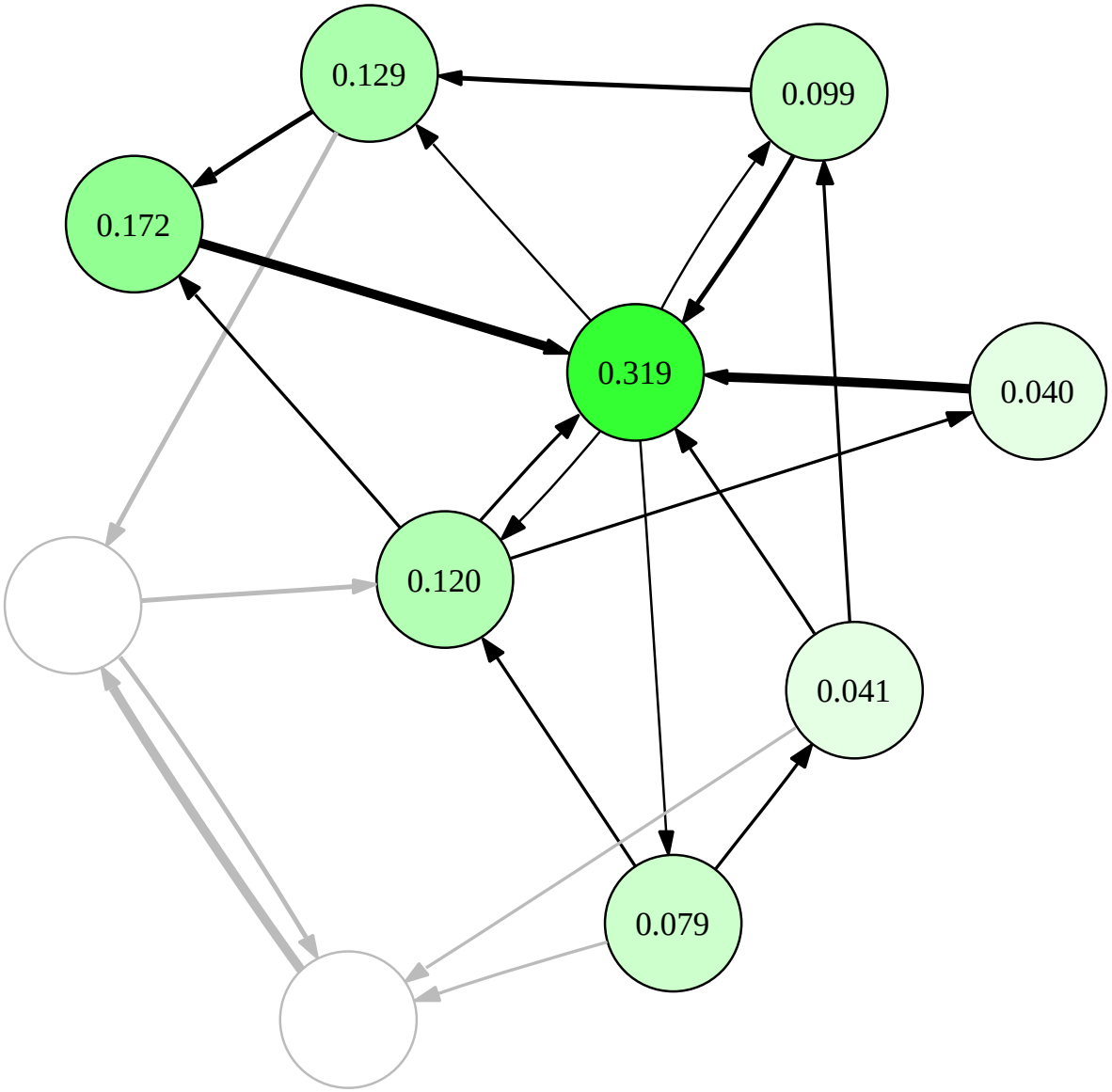
LocalPageRank — Iteration ♯12



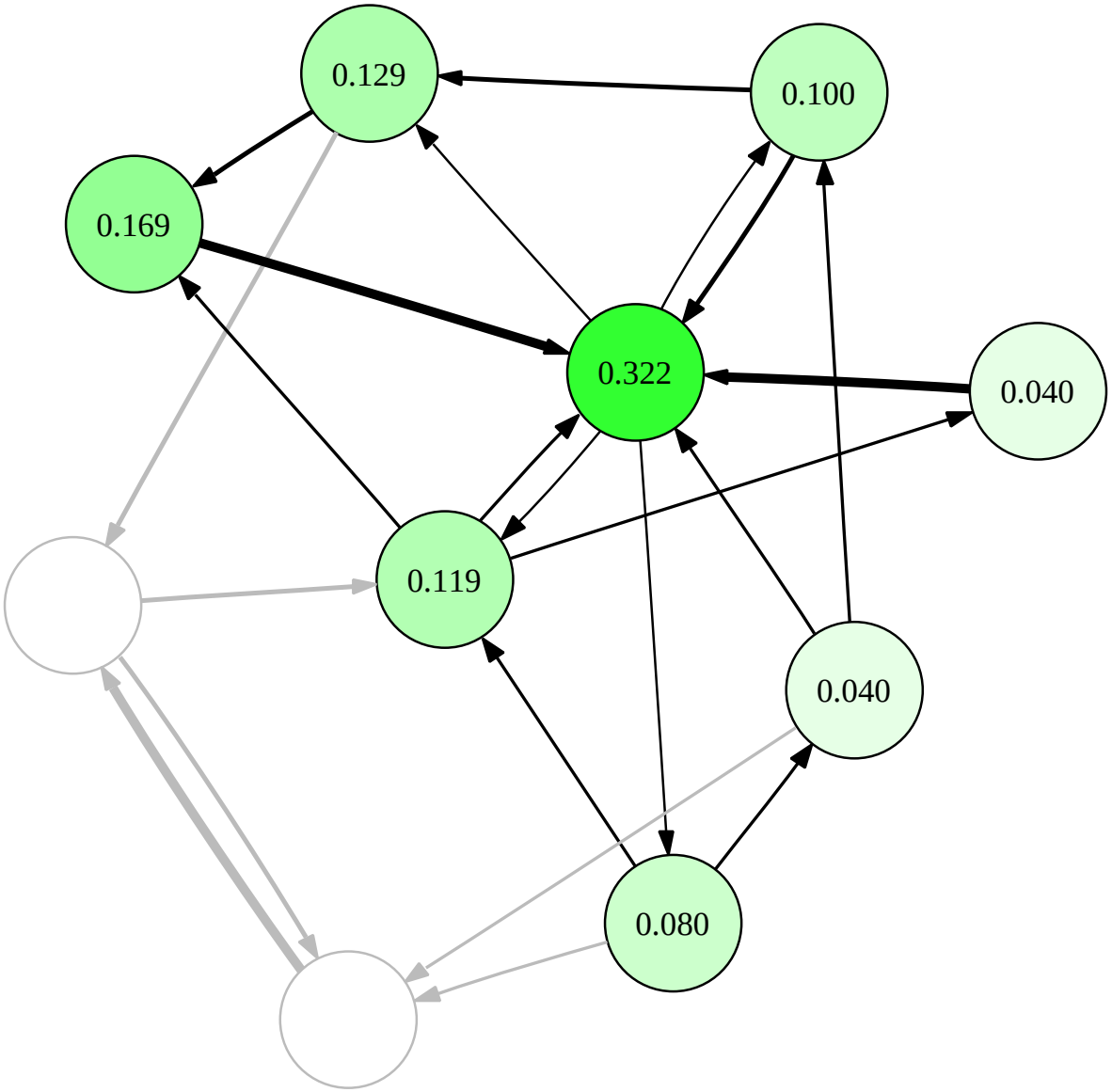
LocalPageRank — Iteration ♯13



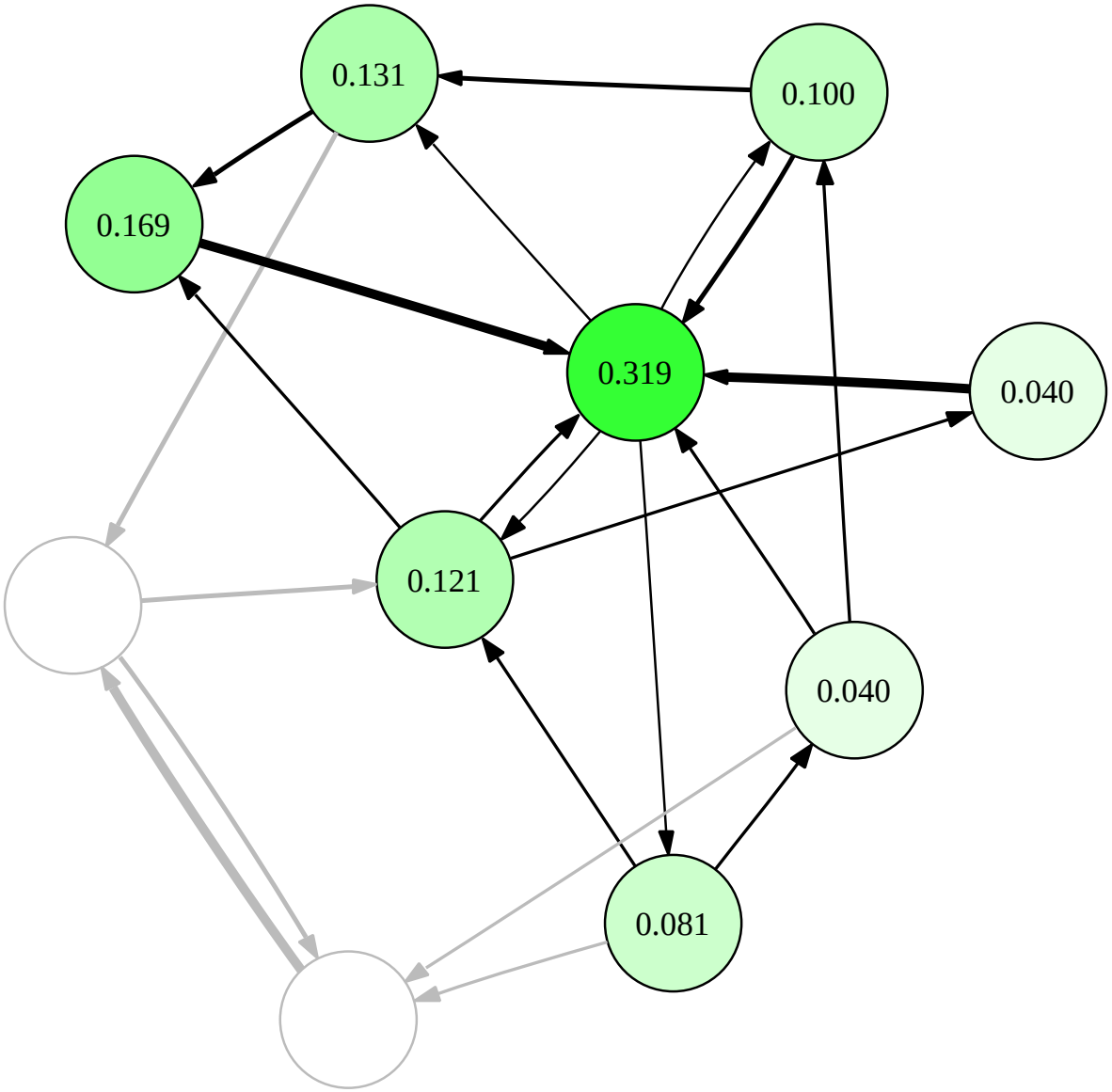
LocalPageRank — Iteration ♯14



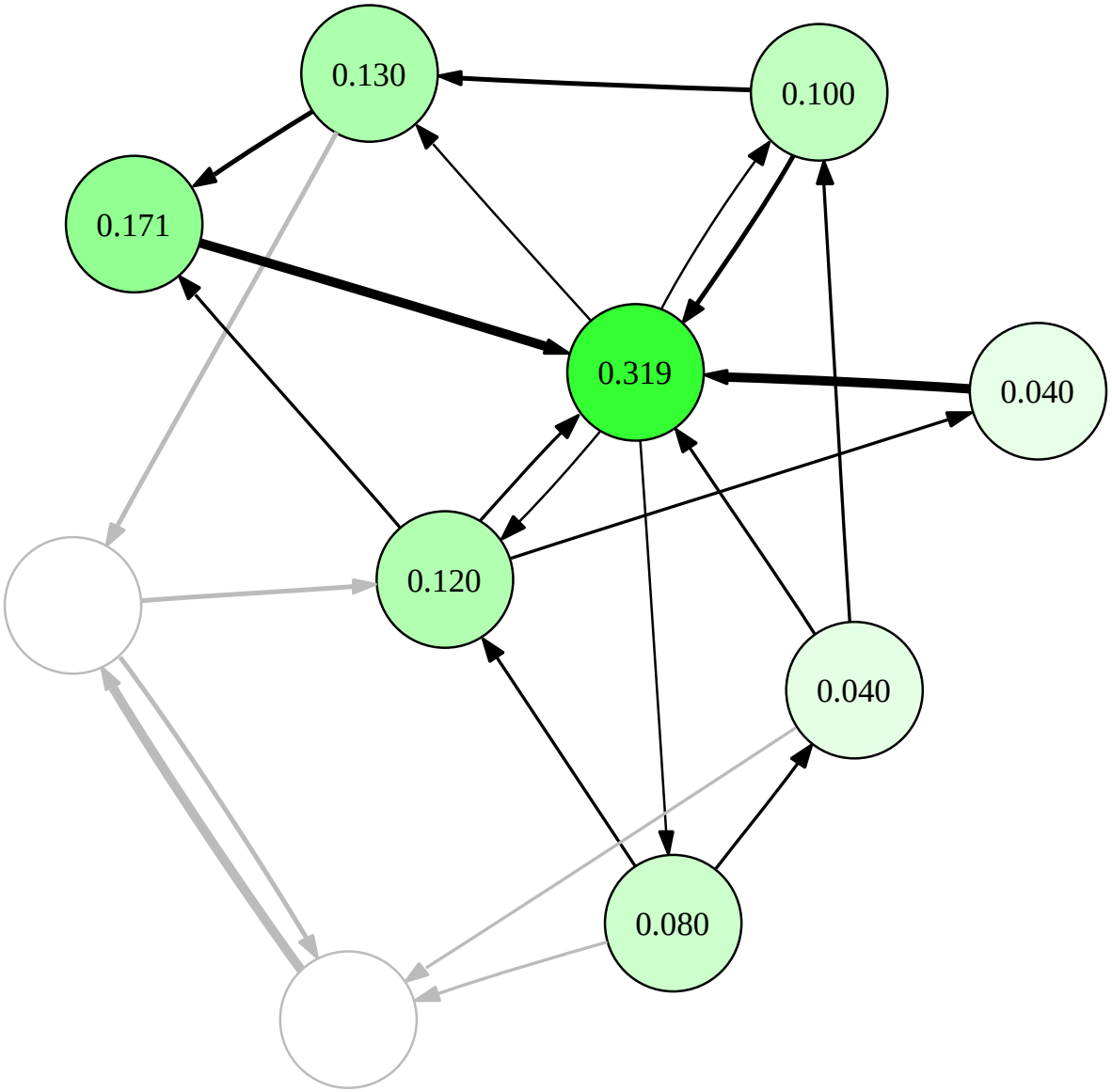
LocalPageRank — Iteration ♯15



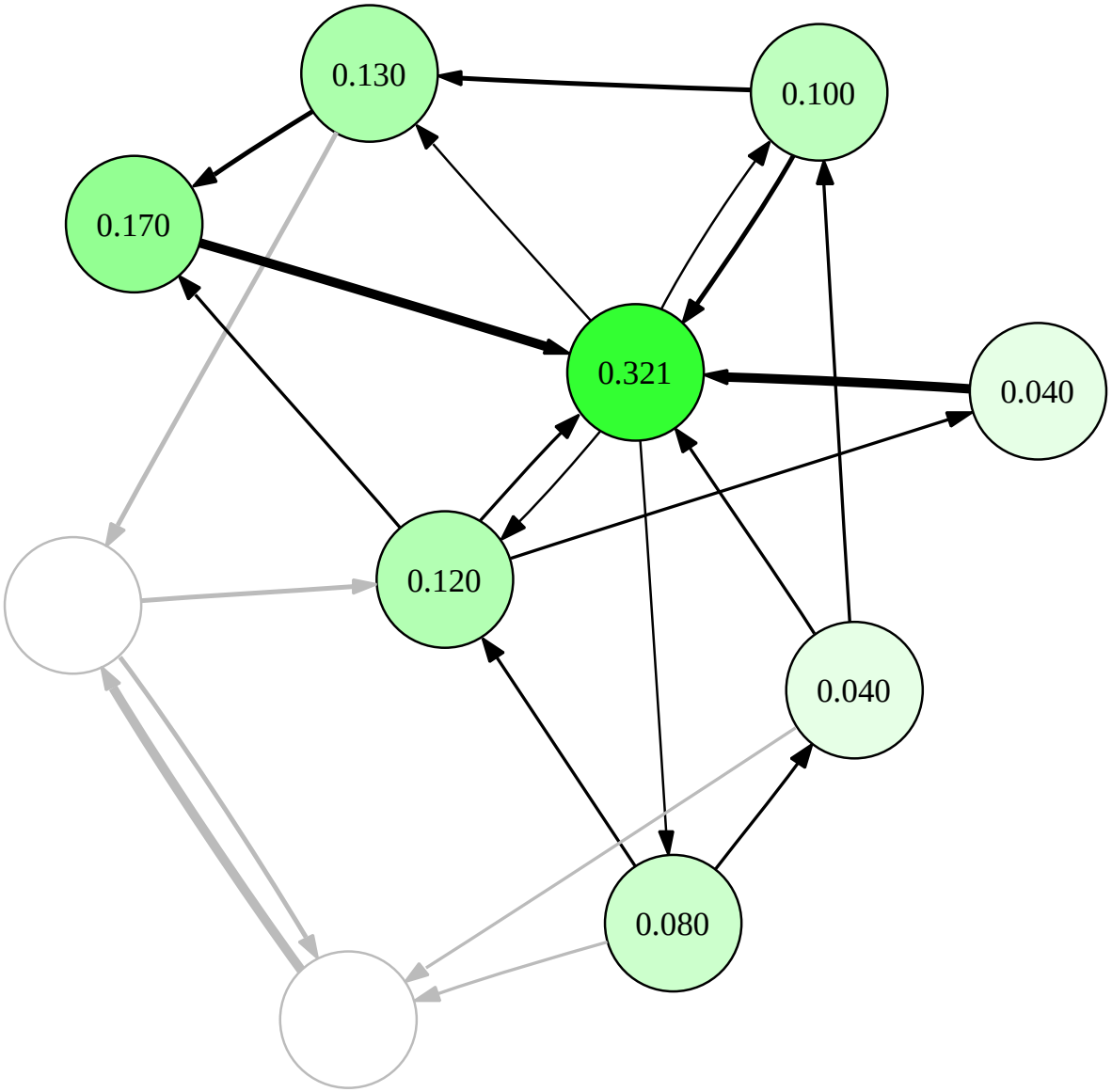
LocalPageRank — Iteration ♯16



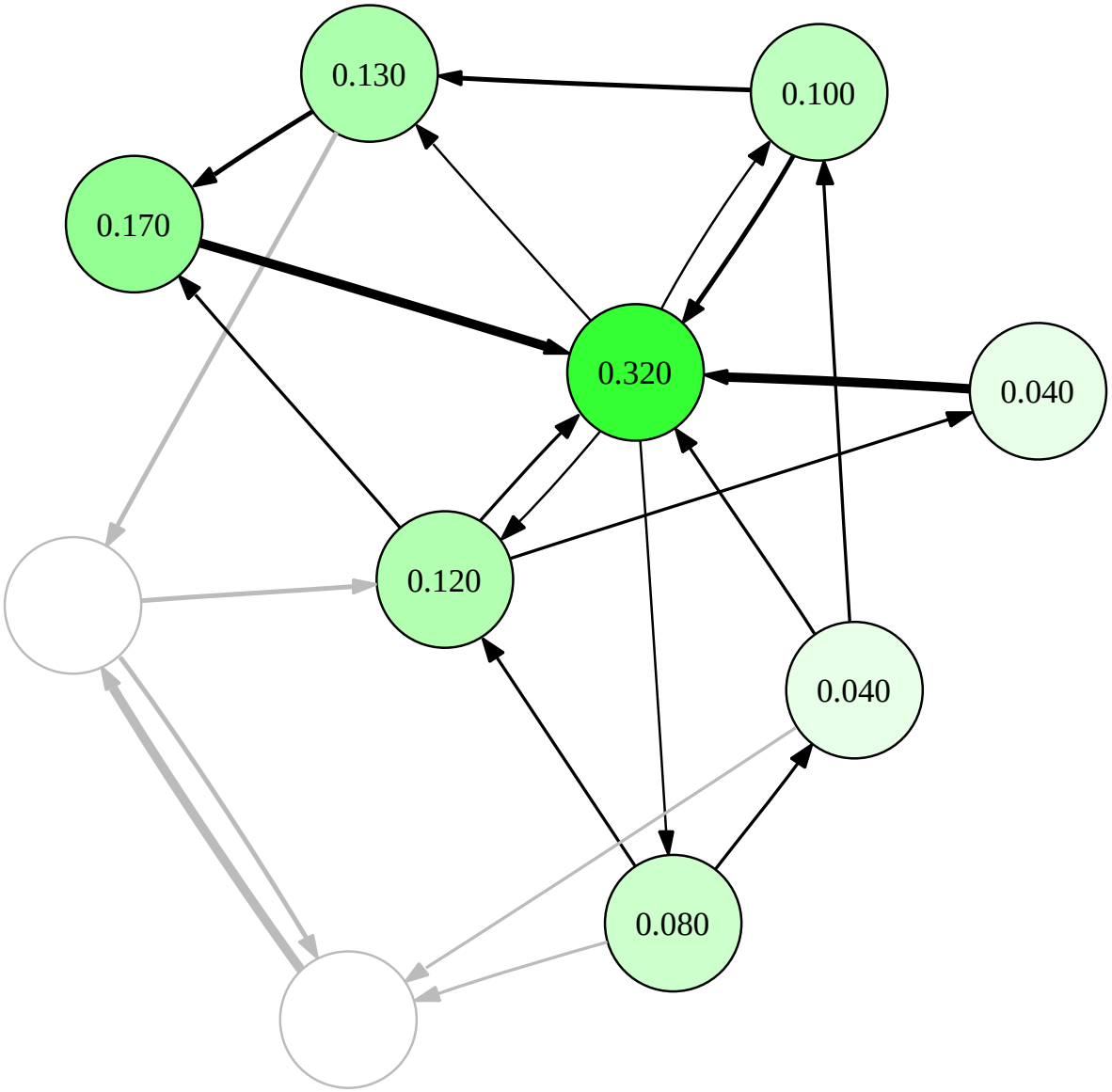
LocalPageRank — Iteration ♯17



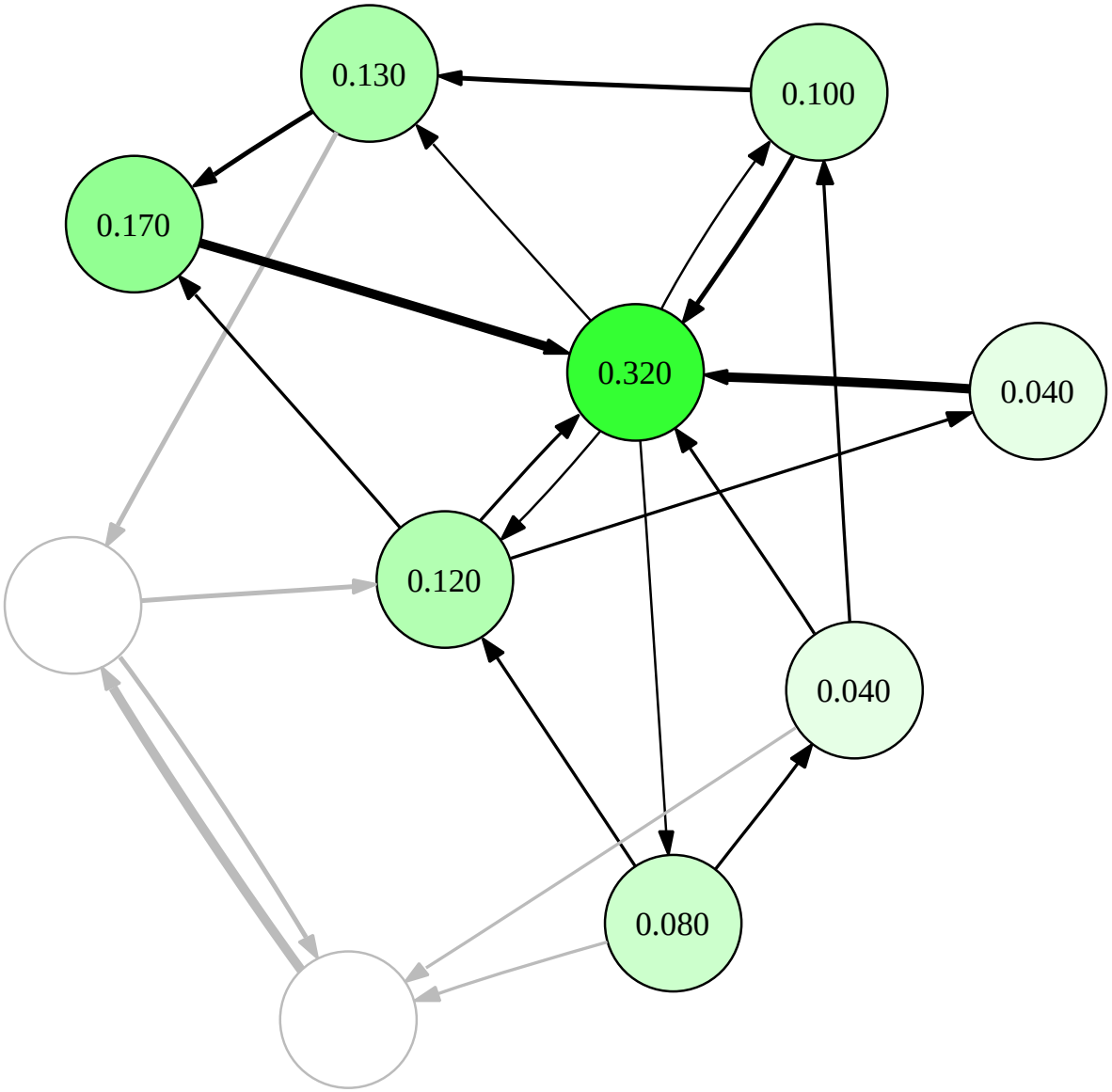
LocalPageRank — Iteration ♯18



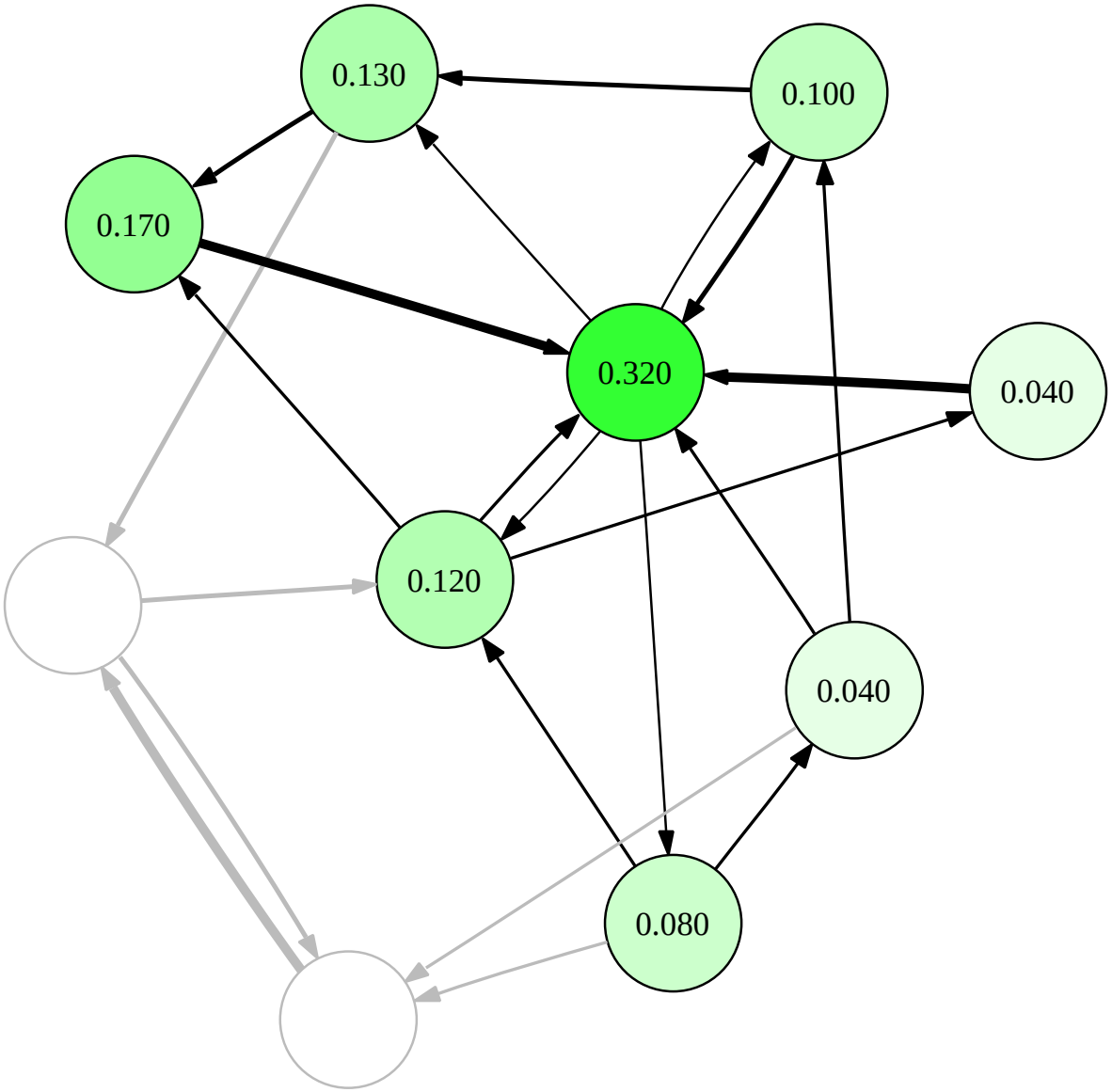
LocalPageRank — Iteration ♯19



LocalPageRank — Iteration ♯20



LocalPageRank — Iteration ♯21



Cosine — Method Description

Method Description

- Vector-space model: nodes are both documents and dimensions.

Cosine — Method Description

Method Description

- Vector-space model: nodes are both documents and dimensions.
- Coordinate of node i along dimension j: term frequency/inverse document frequency weight of the edge between i and j.

Cosine — Method Description

Method Description

- Vector-space model: nodes are both documents and dimensions.
- Coordinate of node i along dimension j: term frequency/inverse document frequency weight of the edge between i and j.
- Score of similarity between i and j: cosine of the corresponding vectors:

$$cos(x^{(i)}, x^{(j)}) = \frac{\sum x_k^{(i)} x_k^{(j)}}{\sqrt{\sum {x_k^{(i)}}^2} \sqrt{\sum {x_k^{(j)}}^2}}$$

Cosine — Method Description

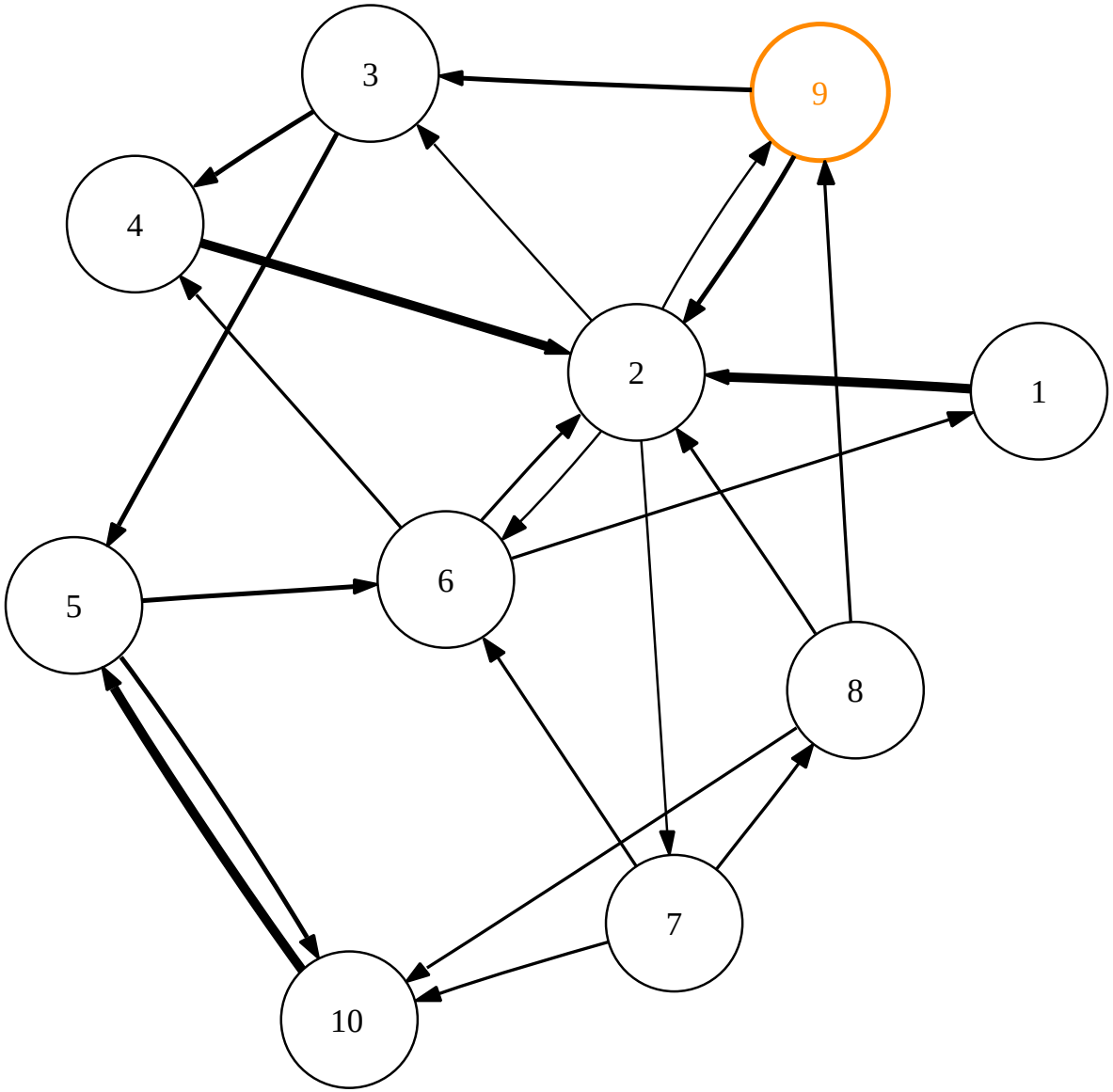
Method Description

- Vector-space model: nodes are both documents and dimensions.
- Coordinate of node i along dimension j: term frequency/inverse document frequency weight of the edge between i and j.
- Score of similarity between i and j: cosine of the corresponding vectors:

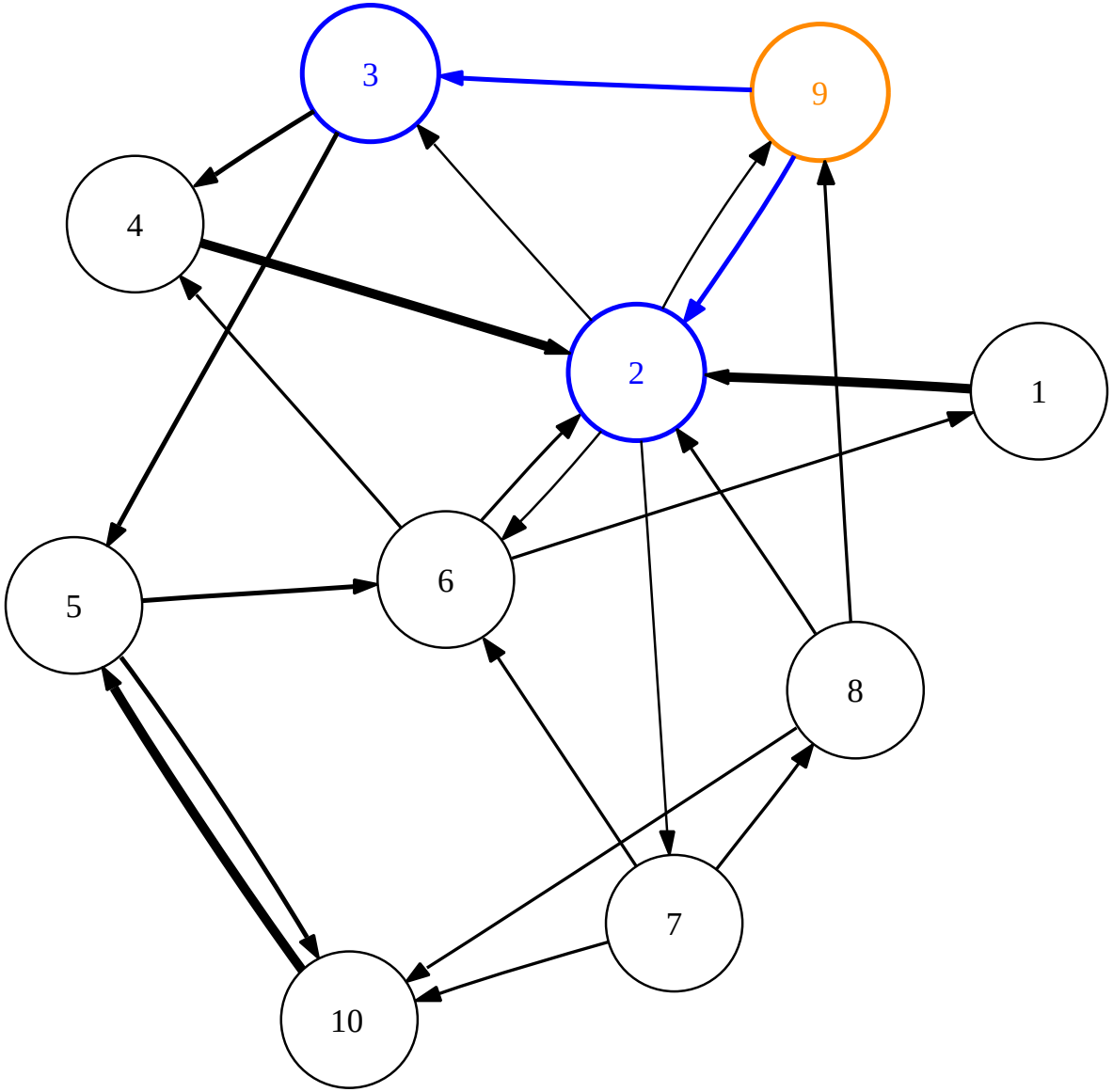
$$cos(x^{(i)}, x^{(j)}) = \frac{\sum x_k^{(i)} x_k^{(j)}}{\sqrt{\sum {x_k^{(i)}}^2} \sqrt{\sum {x_k^{(j)}}^2}}$$

- Classical Text Mining technique, more often used with textual words as dimensions.

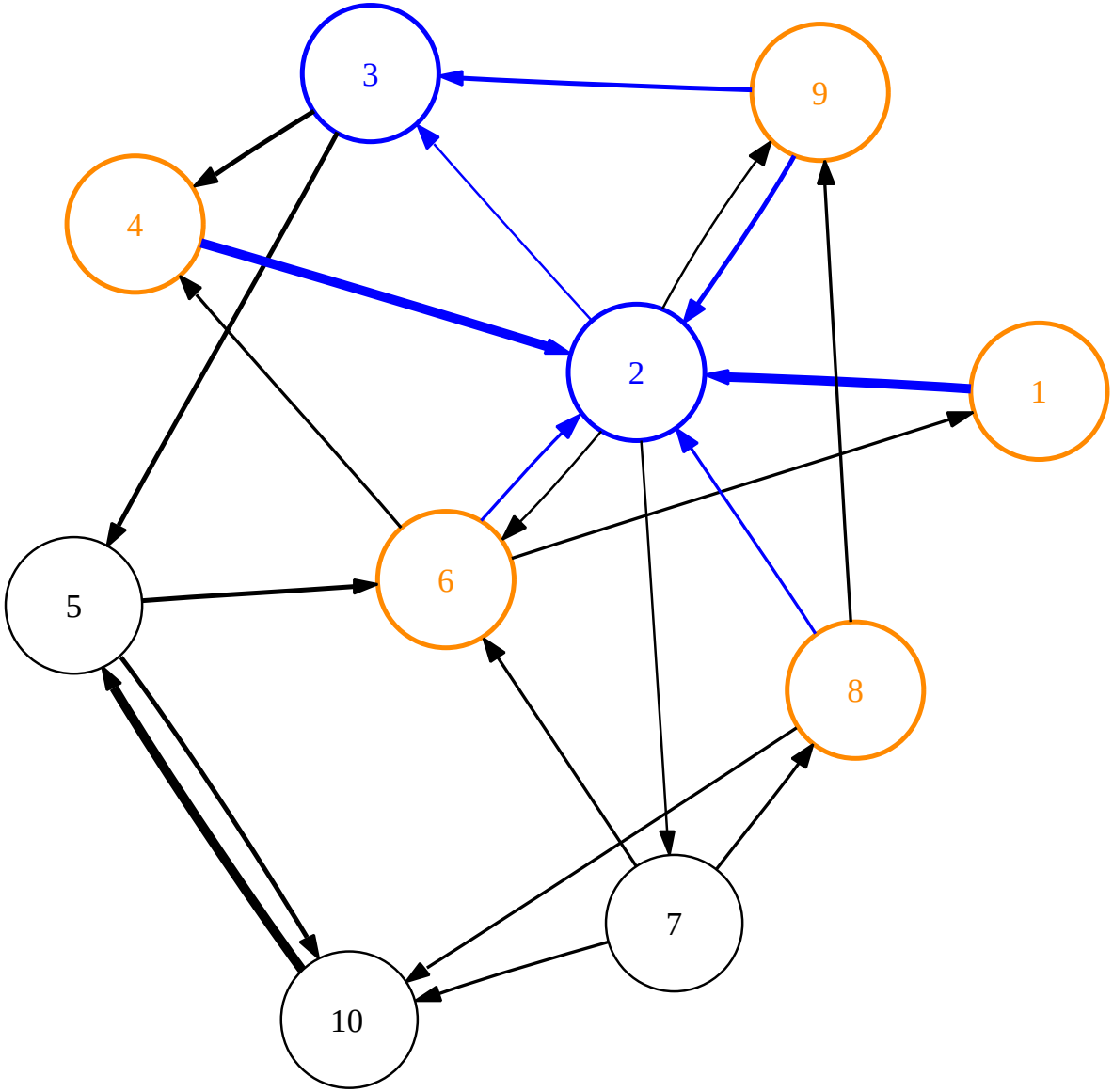
Cosine



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 — Method Description

Method Description

- Related nodes: nodes that are pointed to by the same nodes.

Cocitations — Method Description

Method Description

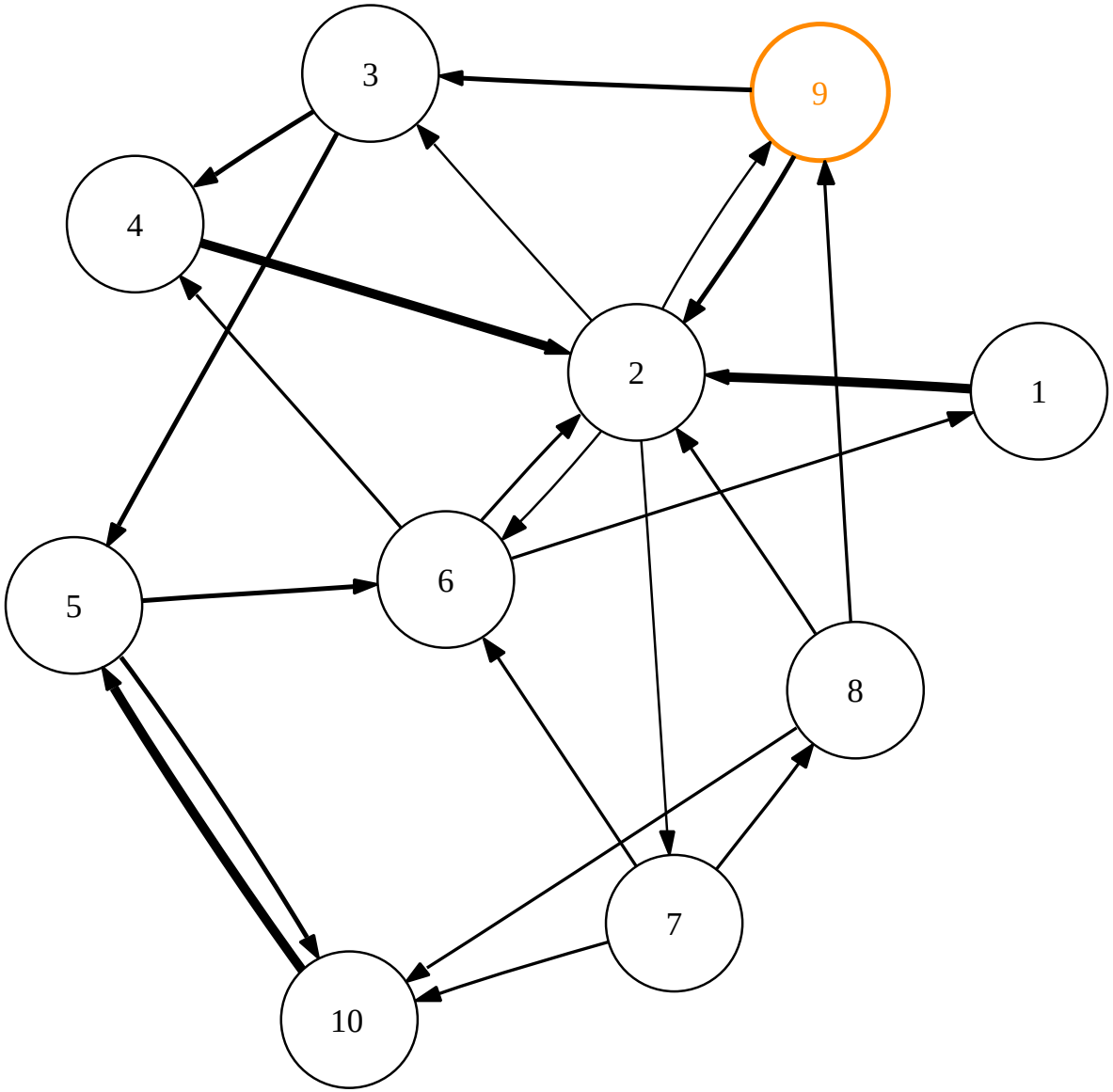
- Related nodes: nodes that are pointed to by the same nodes.
- Score of similarity between i and j: number of nodes pointing to both.

Cocitations — Method Description

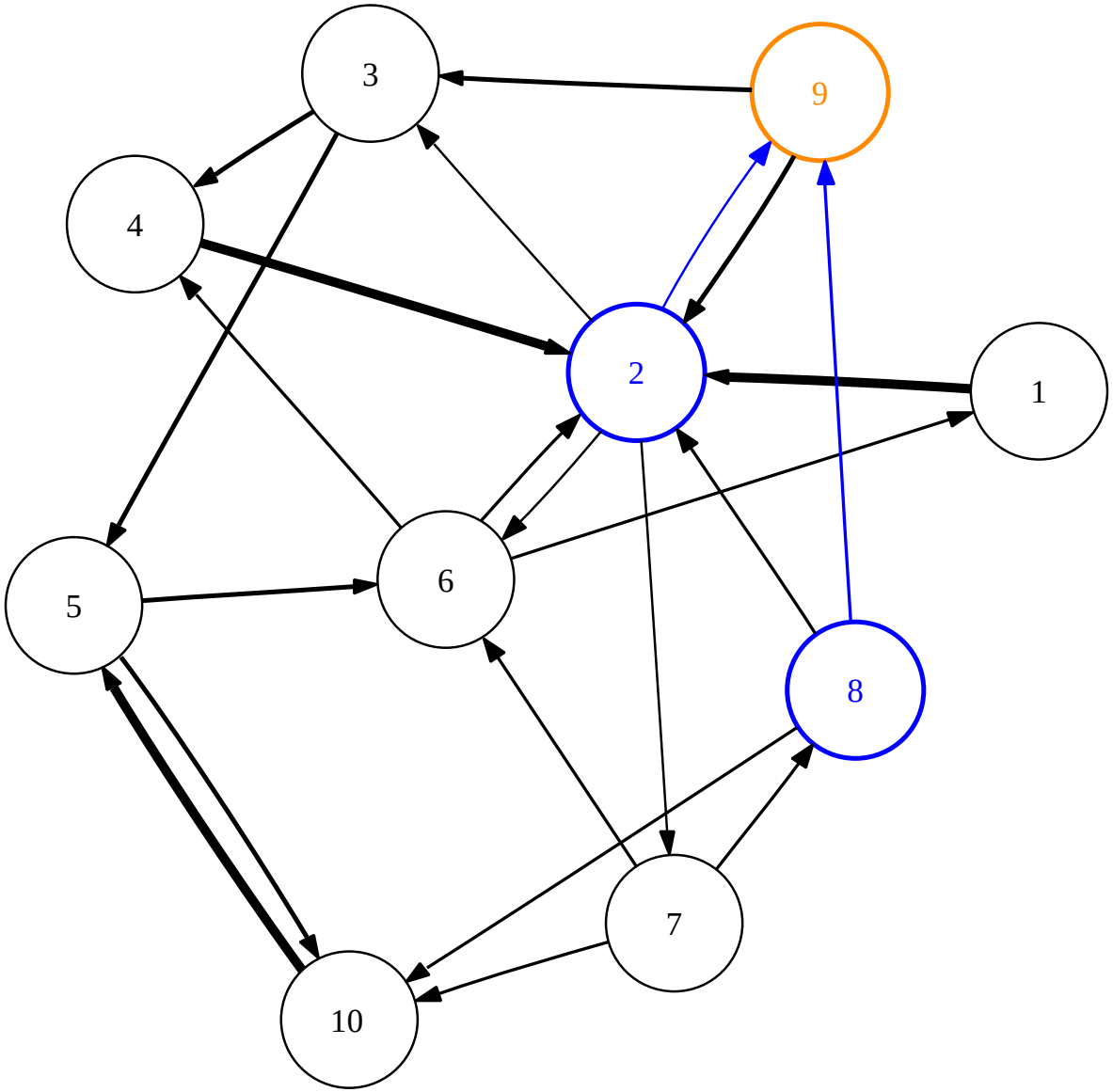
Method Description

- Related nodes: nodes that are pointed to by the same nodes.
- Score of similarity between i and j: number of nodes pointing to both.
- Classically used in bibliometrics.

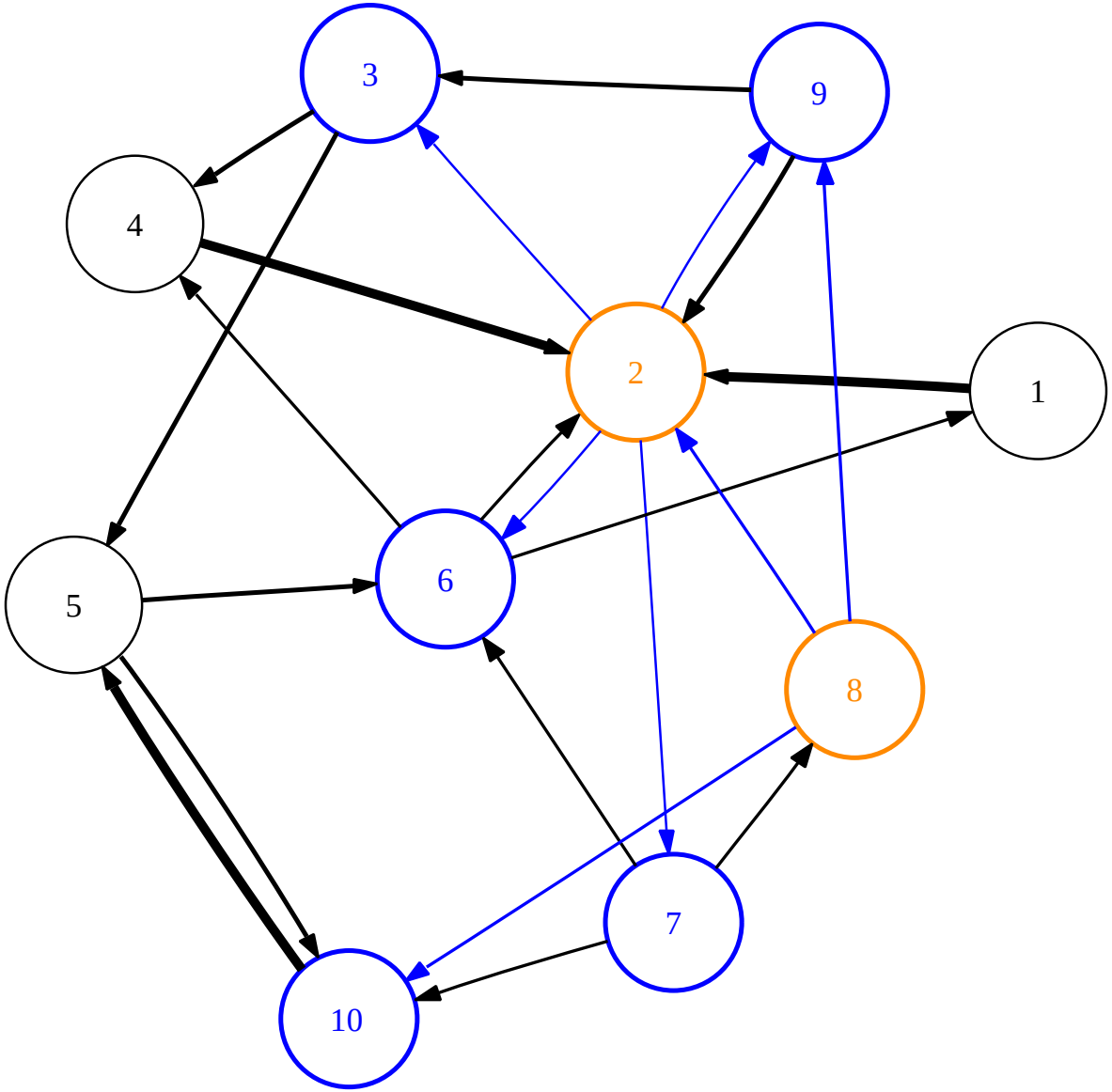
Cocitations



Cocitations



Cocitations



Outline

The graph of Wikipedia

Extraction

- Extraction of the graph from a full XML dump of Wikipedia.

The graph of Wikipedia

Extraction

- Extraction of the graph from a full XML dump of Wikipedia.
- Multiplicity of links used as weight of edges.

The graph of Wikipedia

Extraction

- Extraction of the graph from a full XML dump of Wikipedia.
- Multiplicity of links used as weight of edges.
- Redirections and Categories dealt with consistently with the Web interface.

The graph of Wikipedia

Extraction

- Extraction of the graph from a full XML dump of Wikipedia.
- Multiplicity of links used as weight of edges.
- Redirections and Categories dealt with consistently with the Web interface.
- Only most used Templates are expanded, in order not to "pollute" too much the graph.

The graph of Wikipedia

Extraction

- Extraction of the graph from a full XML dump of Wikipedia.
- Multiplicity of links used as weight of edges.
- Redirections and Categories dealt with consistently with the Web interface.
- Only most used Templates are expanded, in order not to "pollute" too much the graph.

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.

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

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.

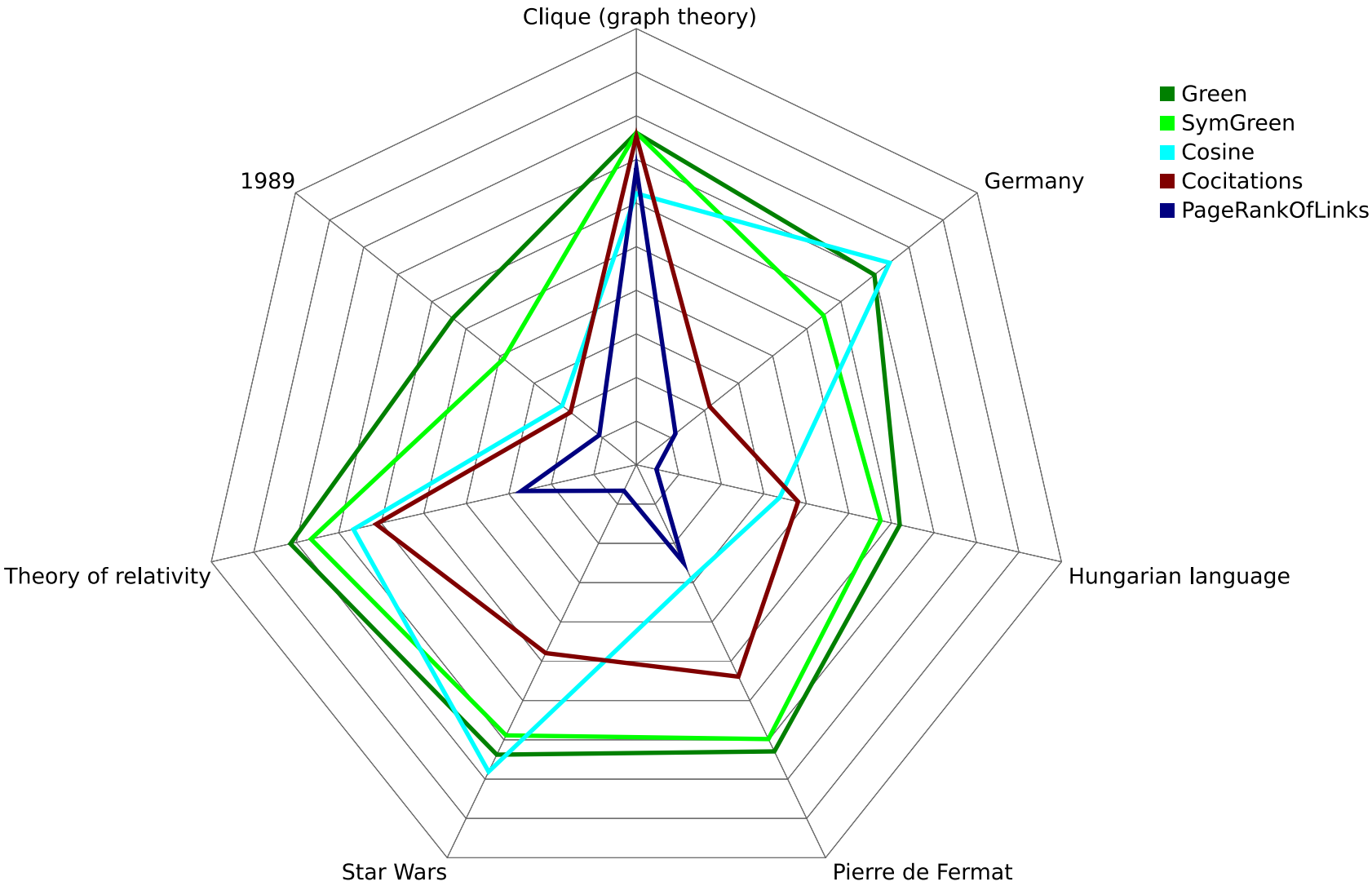
Output on Pierre de Fermat

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. Community of Agglomeration of Greater Toulouse	8. Isaac Newton
9. France	9. Fermat's little theorem	9. Number theory	9. Lilhac	9. Blaise Pascal
10. Pseudoprime	10. Calculus	10. René Descartes	10. Institut d'études politiques de Toulouse	10. Carl Friedrich Gauss

Output on Germany

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

Discussion of the results

- *PageRankOfLinks* (like *LocalPageRank*): usually very bad.

Discussion of the results

- *PageRankOfLinks* (like *LocalPageRank*): usually very bad.
- *Cosine*: very variable, good performance on some articles, poor on other, a lot of false positives in all cases.

Discussion of the results

- *PageRankOfLinks* (like *LocalPageRank*): usually very bad.
- *Cosine*: very variable, good performance on some articles, poor on other, a lot of false positives in all cases.
- *Green* & *SymGreen*: similar performances, a notable advantage for *Green* on these examples (may change).

Discussion of the results

- *PageRankOfLinks* (like *LocalPageRank*): usually very bad.
- *Cosine*: very variable, good performance on some articles, poor on other, a lot of false positives in all cases.
- *Green* & *SymGreen*: similar performances, a notable advantage for *Green* on these examples (may change).
- *Green*: good robustness, very few false positives.

Discussion of the results

- *PageRankOfLinks* (like *LocalPageRank*): usually very bad.
- *Cosine*: very variable, good performance on some articles, poor on other, a lot of false positives in all cases.
- *Green* & *SymGreen*: similar performances, a notable advantage for *Green* on these examples (may change).
- *Green*: good robustness, very few false positives.
- *Green* & *SymGreen*: only methods to find some important semantic relations: Finnish language for Hungarian language, Berlin Wall or The Satanic Verses (novel) for 1989...

Outline

Summary

- Green measures: a tool for extracting semantic information in a graph.

Summary

- Green measures: a tool for extracting semantic information in a graph.
- In comparison to other methods, in the case of Wikipedia:

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.

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.

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.

Perspectives

- Application to the Web graph.
- Interpolation between *Green* and *SymGreen*.

Perspectives

- Application to the Web graph.
- Interpolation between *Green* and *SymGreen*.
- Clustering using *Green* measures: unpractical now because of computation times.

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.