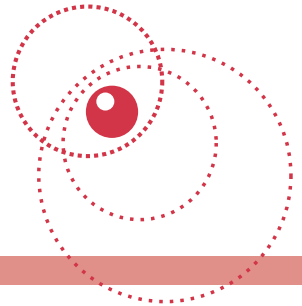


Web Mining

Web search: The Web Graph

8 October 2010

Pierre Senellart

TELECOM ParisTech

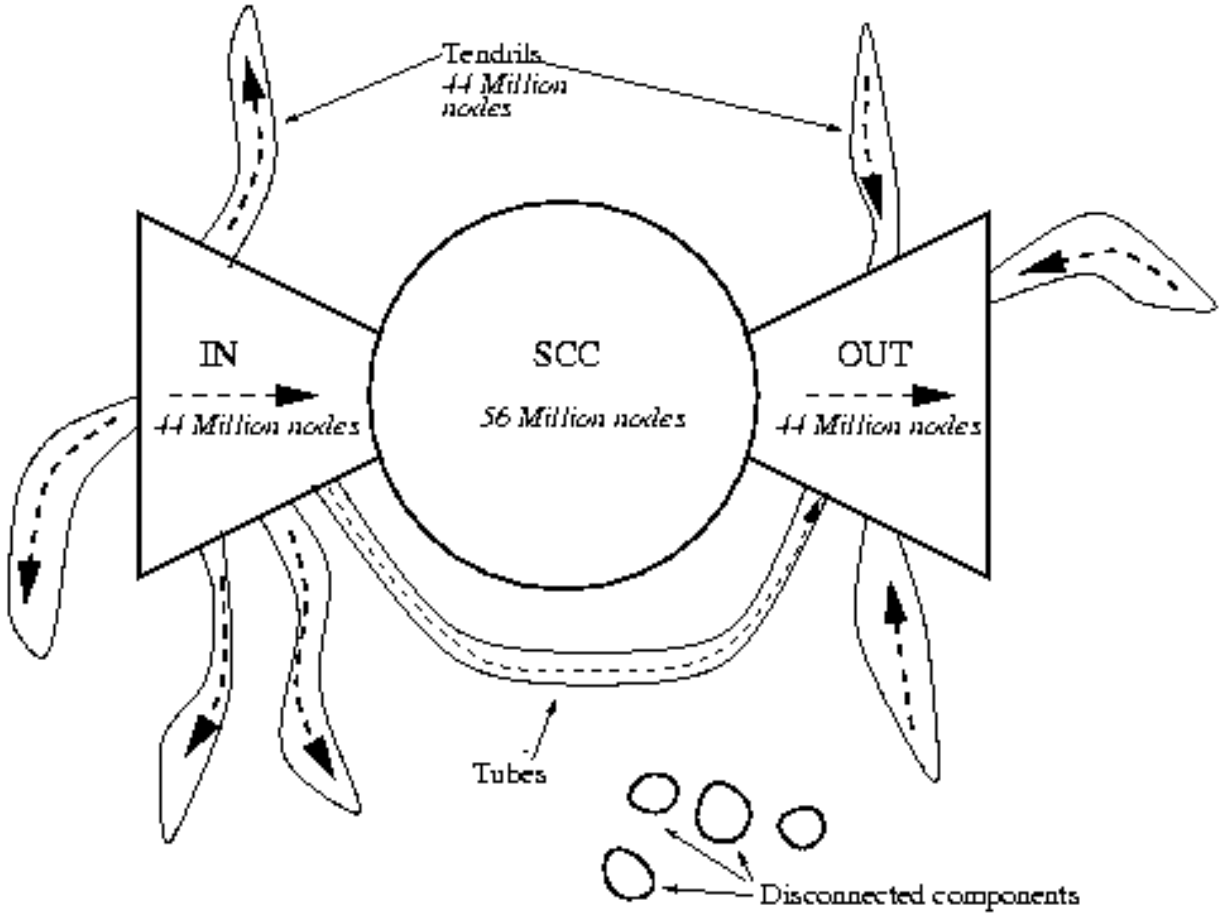
The World Wide Web seen as a (directed) graph:

Vertices: Web pages

Edges: hyperlinks

Same for other interlinked environments:

- dictionaries
- encyclopedias
- scientific publications
- social networks



[Broder et al., 2000]

Sparsity. Is the graph sparse ($|A| \ll |S|^2$)?

Sparsity. Is the graph sparse ($|A| \ll |S|^2$)?

Yes, the World Wide Web is sparse

Sparsity. Is the graph sparse ($|A| \ll |S|^2$)?
Yes, the World Wide Web is sparse

Typical distance. What is the mean distance between any pairs of vertices?

Sparsity. Is the graph sparse ($|A| \ll |S|^2$)?
Yes, the World Wide Web is sparse

Typical distance. What is the mean distance between any pairs of vertices? Logarithmic typical distance

Sparsity. Is the graph sparse ($|A| \ll |S|^2$)?
Yes, the World Wide Web is sparse

Typical distance. What is the mean distance between any pairs of vertices? Logarithmic typical distance

Local clustering. If a is connected to both b and c, is the probability that b is connected to c significantly greater than the probability any two nodes are connected?

Sparsity. Is the graph sparse ($|A| \ll |S|^2$)? Yes, the World Wide Web is sparse

Typical distance. What is the mean distance between any pairs of vertices? Logarithmic typical distance

Local clustering. If a is connected to both b and c, is the probability that b is connected to c significantly greater than the probability any two nodes are connected? Strong local clustering

Sparsity. Is the graph sparse ($|A| \ll |S|^2$)? Yes, the World Wide Web is sparse

Typical distance. What is the mean distance between any pairs of vertices? Logarithmic typical distance

Local clustering. If a is connected to both b and c, is the probability that b is connected to c significantly greater than the probability any two nodes are connected? Strong local clustering

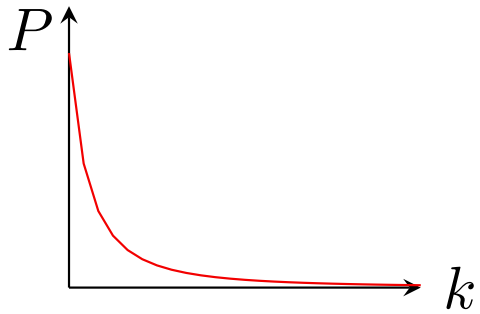
Degree distribution. What is the distribution of the degree of vertices?

Sparsity. Is the graph sparse ($|A| \ll |S|^2$)? Yes, the World Wide Web is sparse

Typical distance. What is the mean distance between any pairs of vertices? Logarithmic typical distance

Local clustering. If a is connected to both b and c, is the probability that b is connected to c significantly greater than the probability any two nodes are connected? Strong local clustering

Degree distribution. What is the distribution of the degree of vertices? Power-law indegree and outdegree distribution ($2 \leq \gamma \leq 3$) [Broder et al., 2000]



Important pages are pages pointed to by important pages.

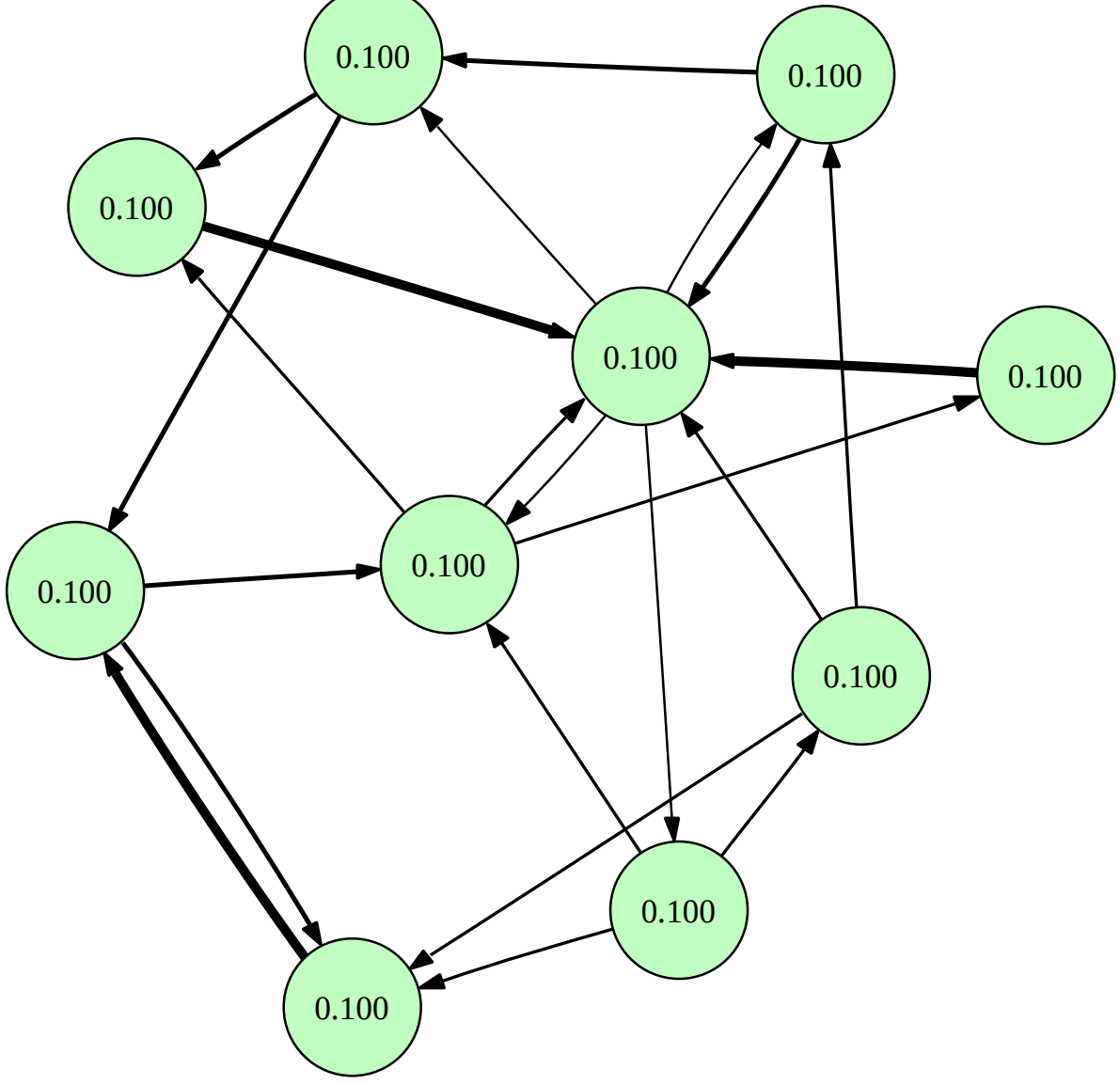
$$\begin{cases} g_{ij} = 0 & \text{if there is no link between page } i \text{ and } j; \\ g_{ij} = \frac{1}{n_i} & \text{otherwise, with } n_i \text{ the number of outgoing links of page } i. \end{cases}$$

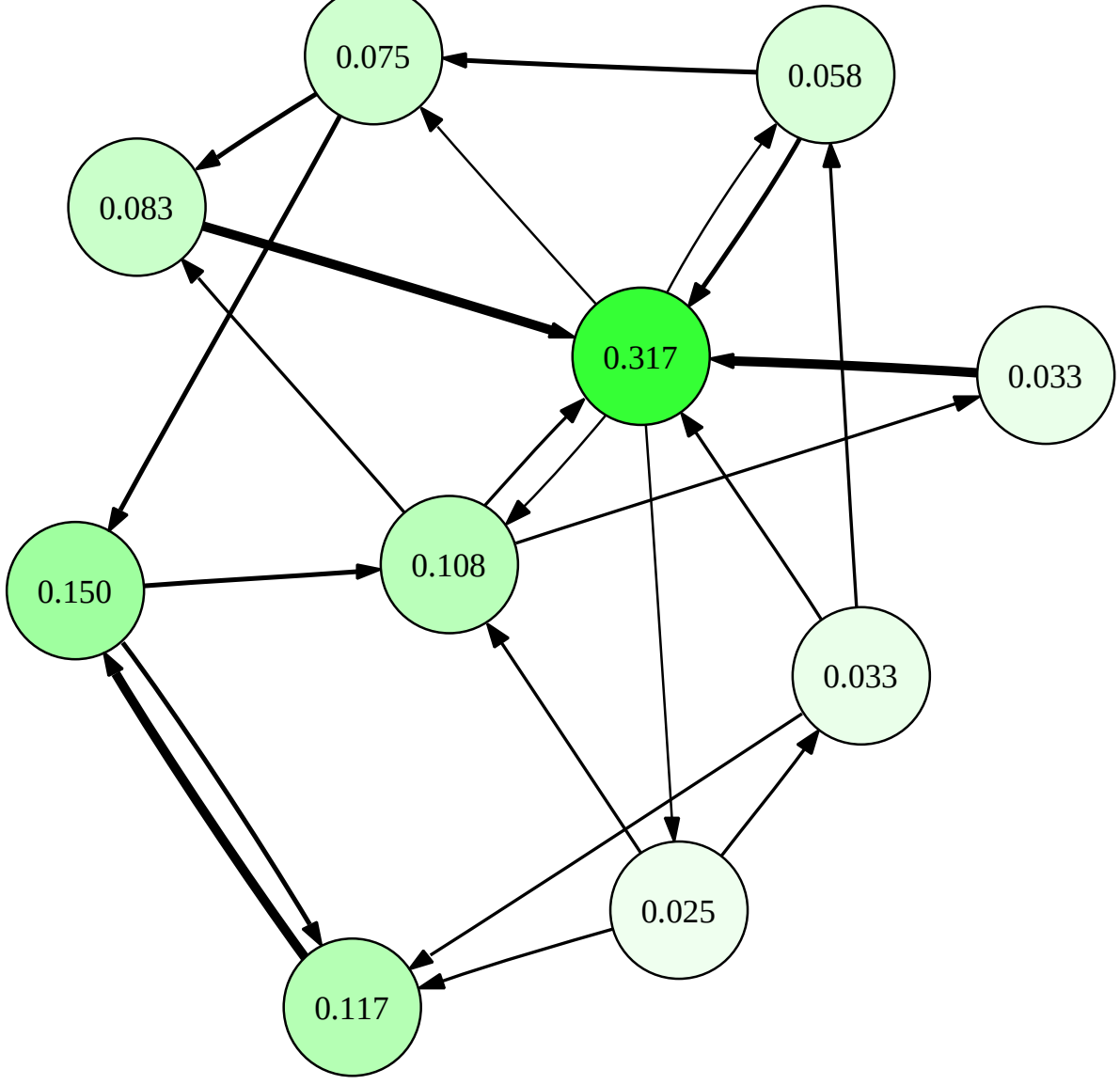
Definition (Tentative)

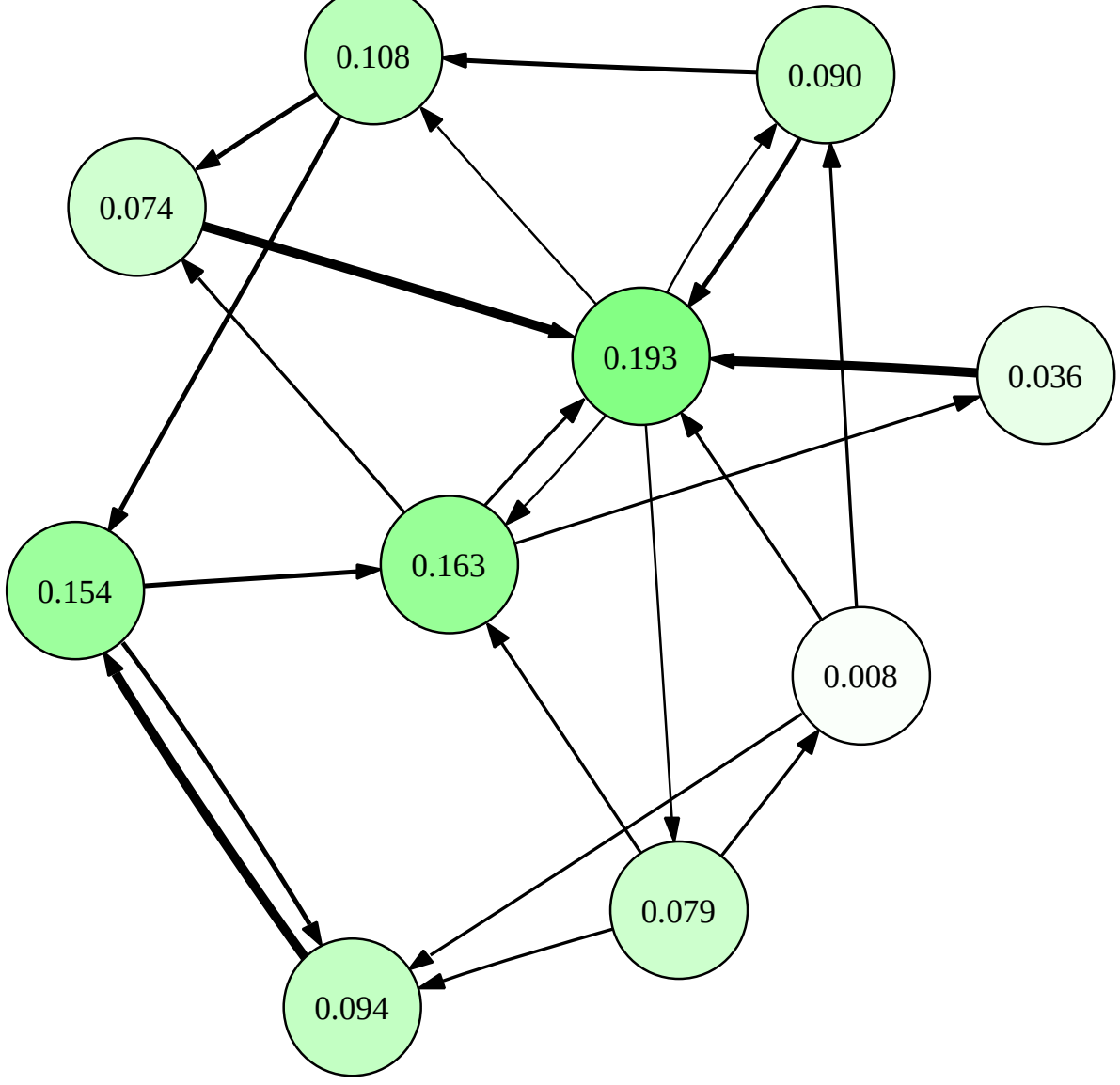
Probability that the surfer following the random walk in G has arrived on page i at some distant given point in the future.

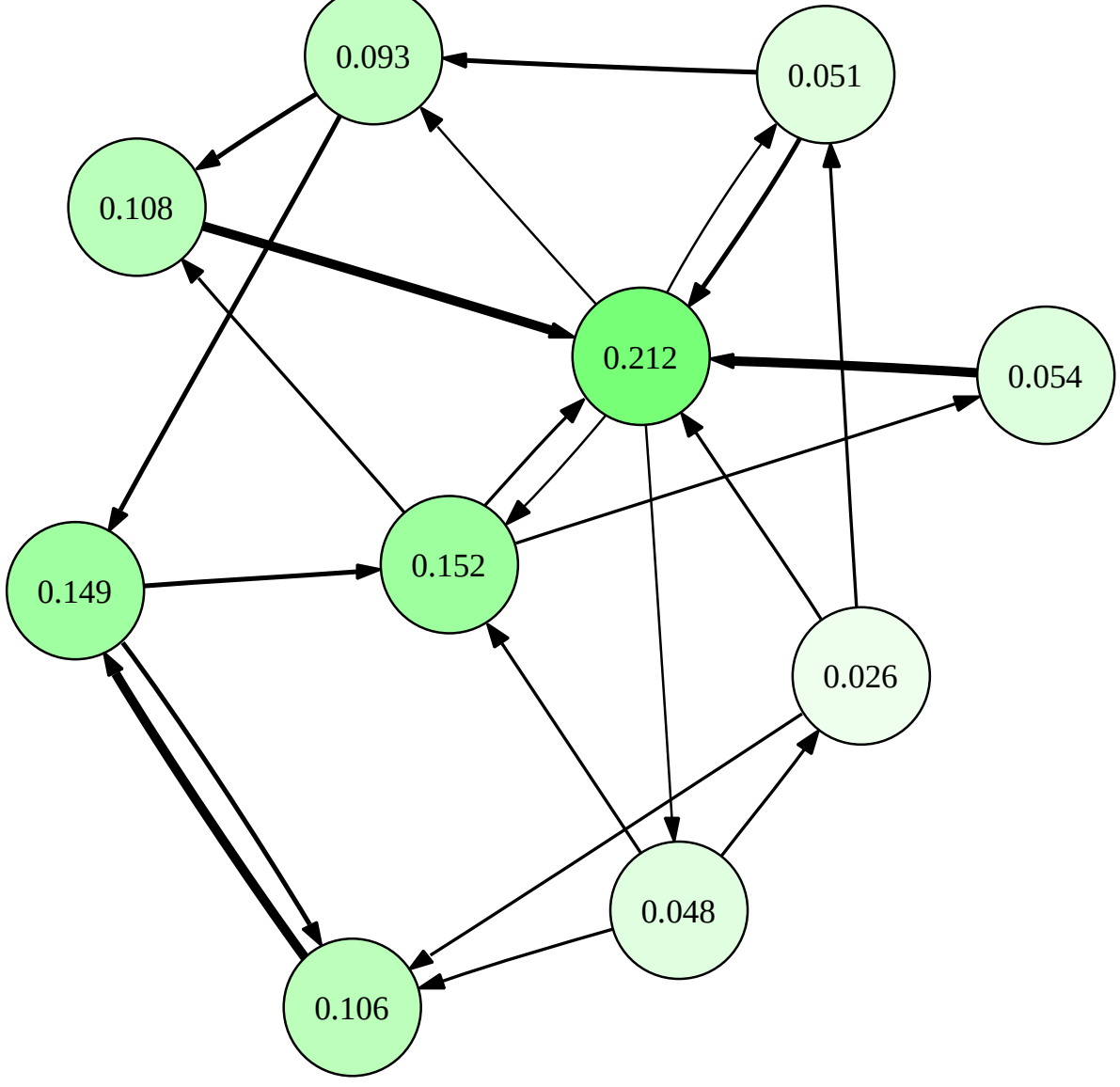
$$\mathrm{pr}(i) = \left(\lim_{k \to +\infty} (G^T)^k v \right)_i$$

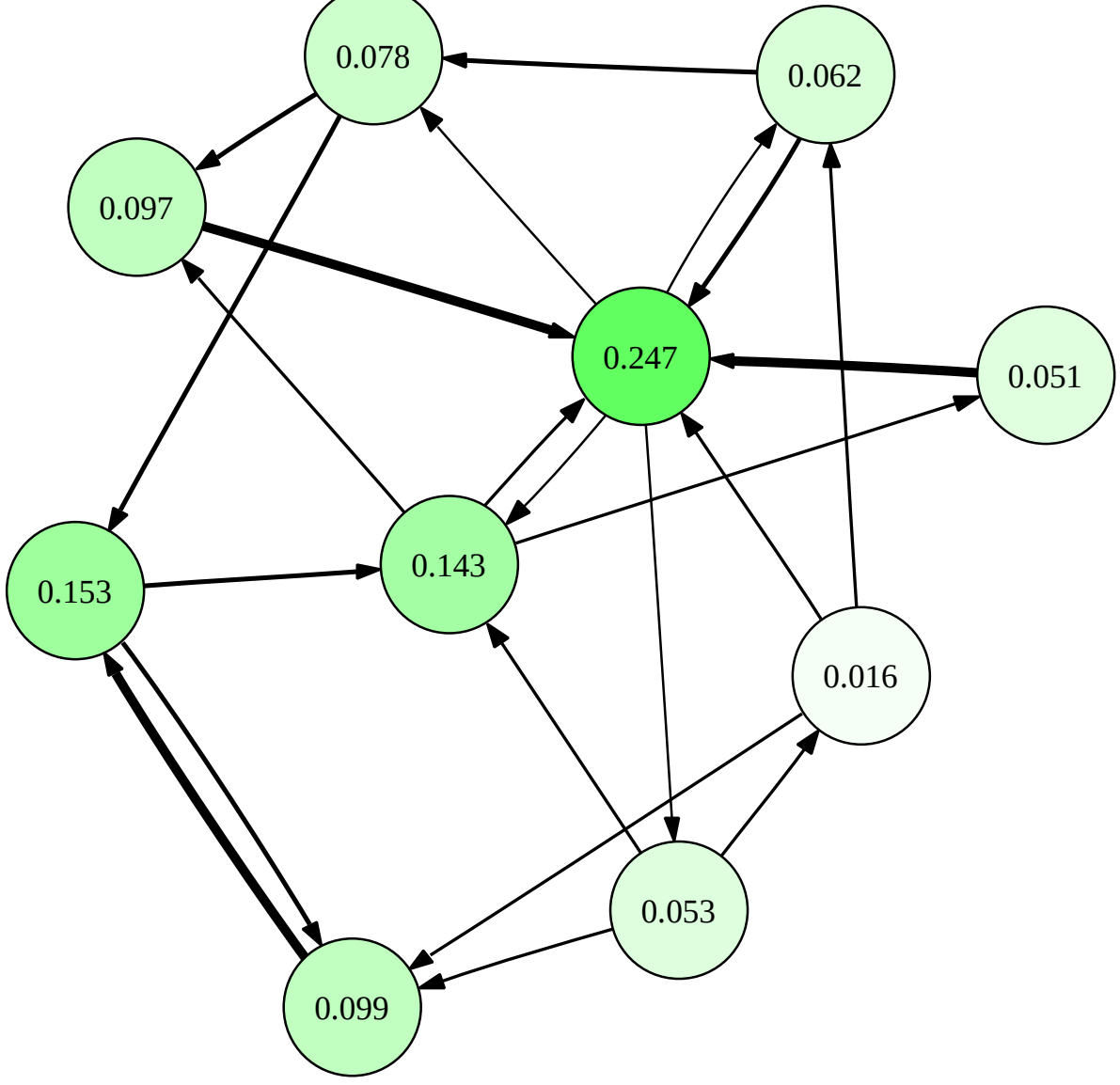
where v is some initial column vector.

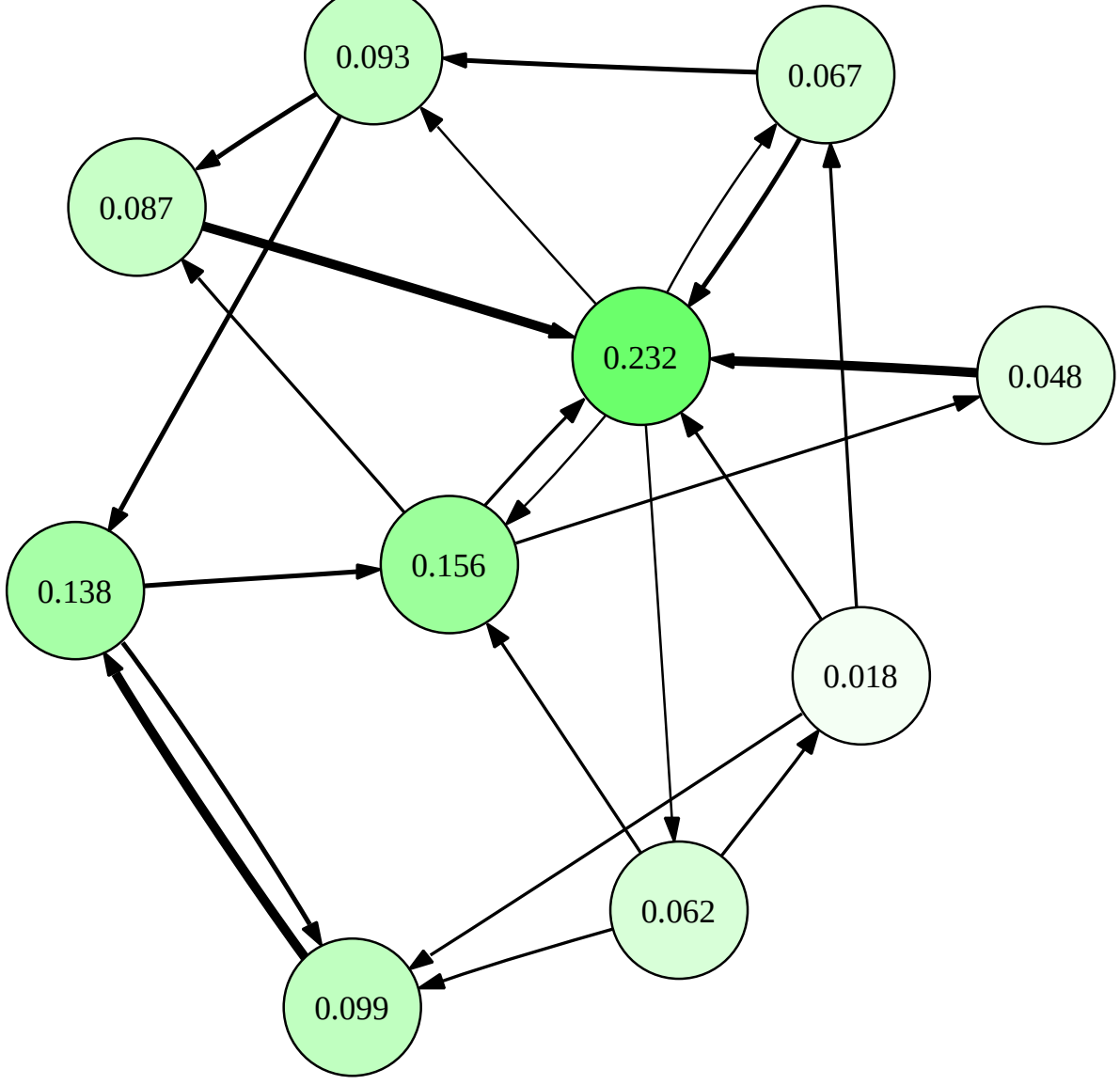


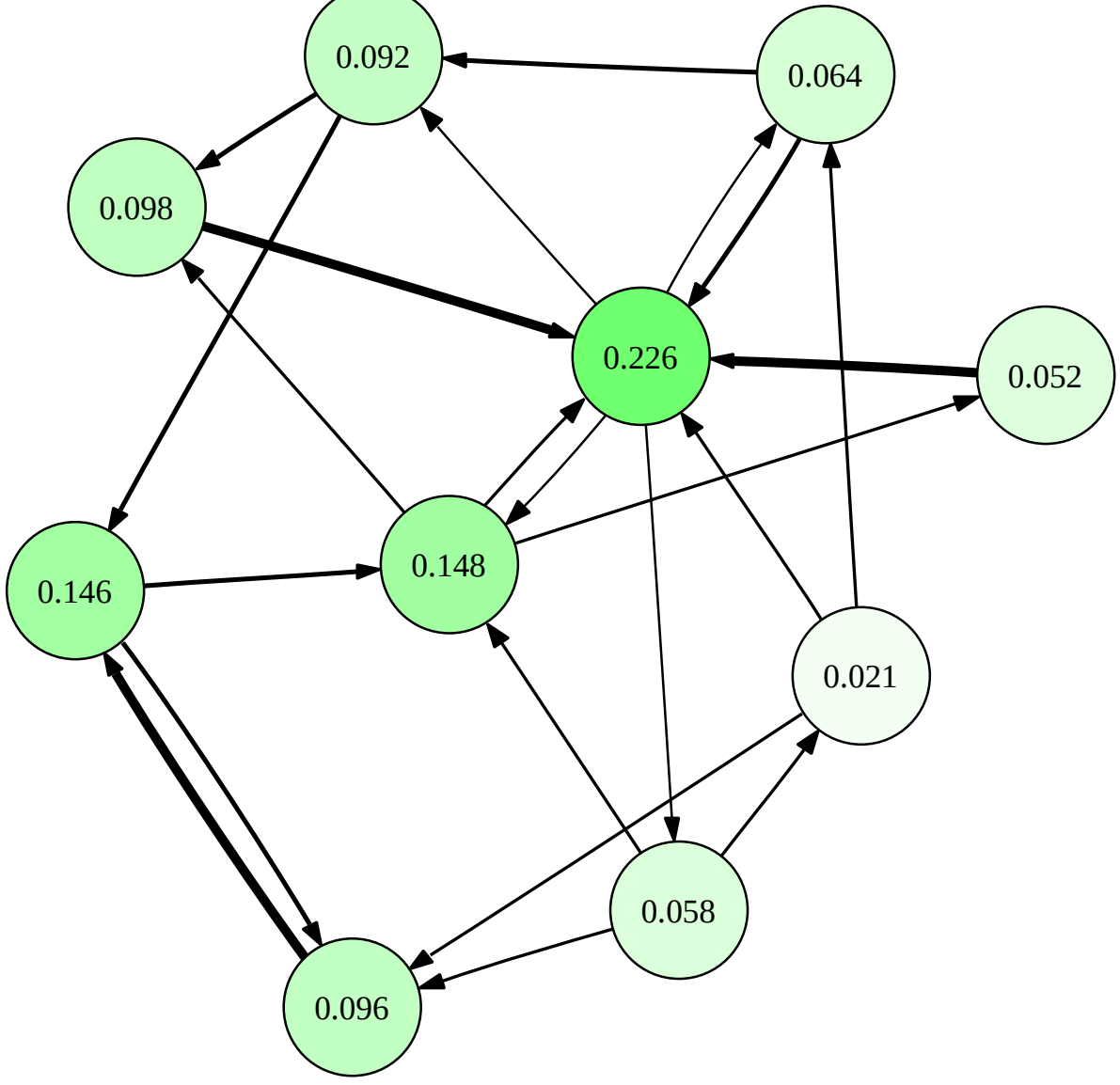


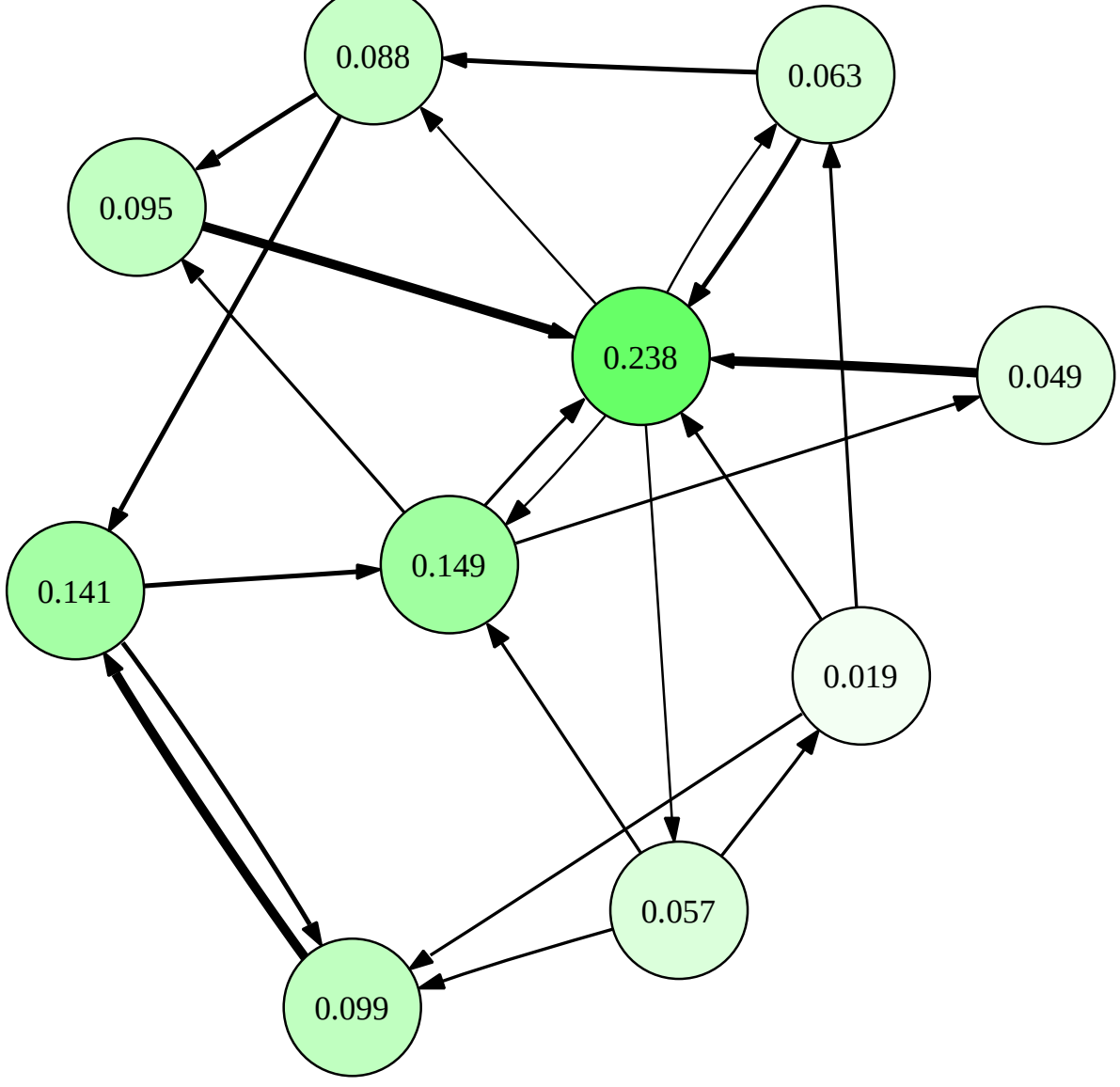


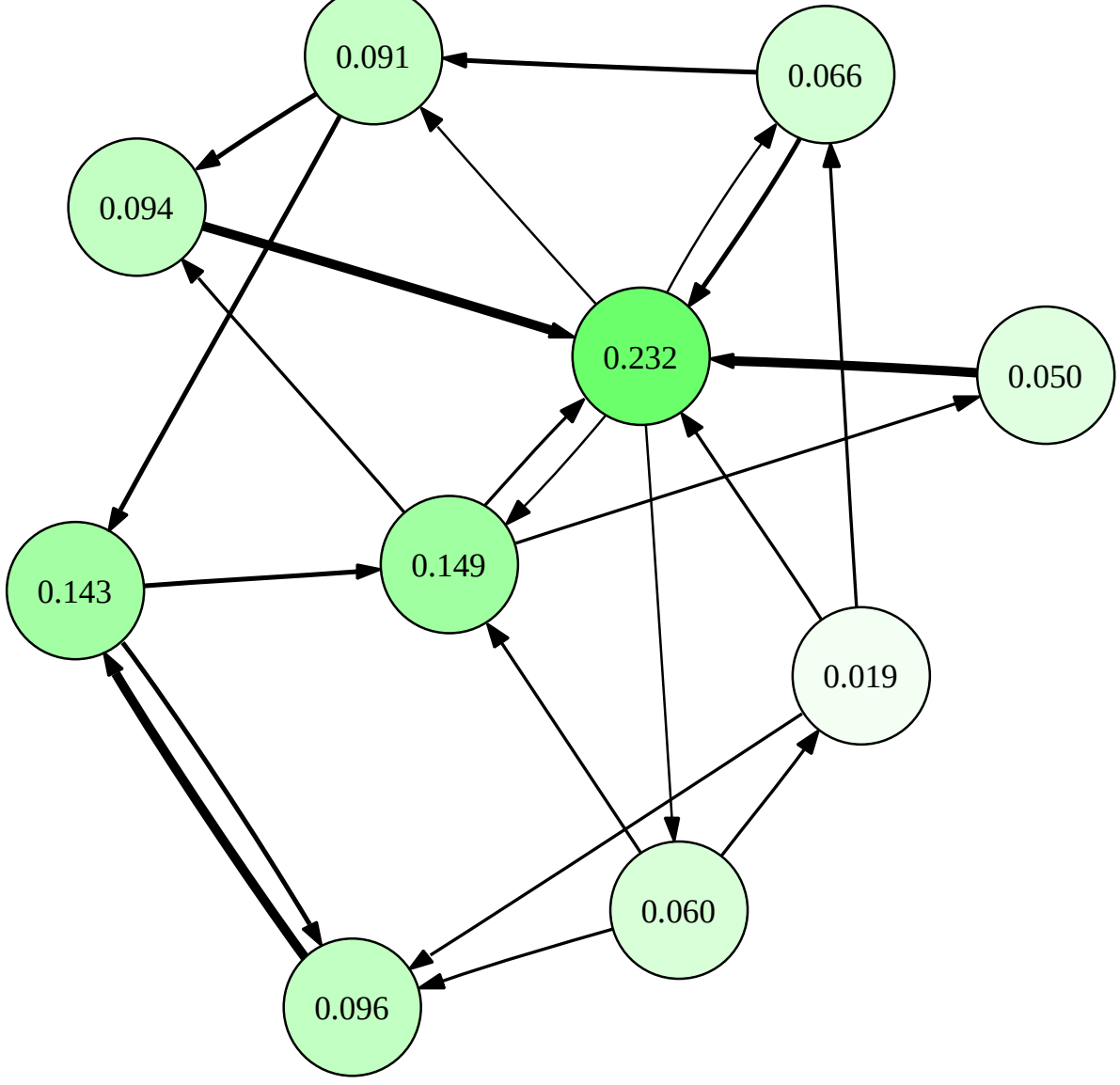


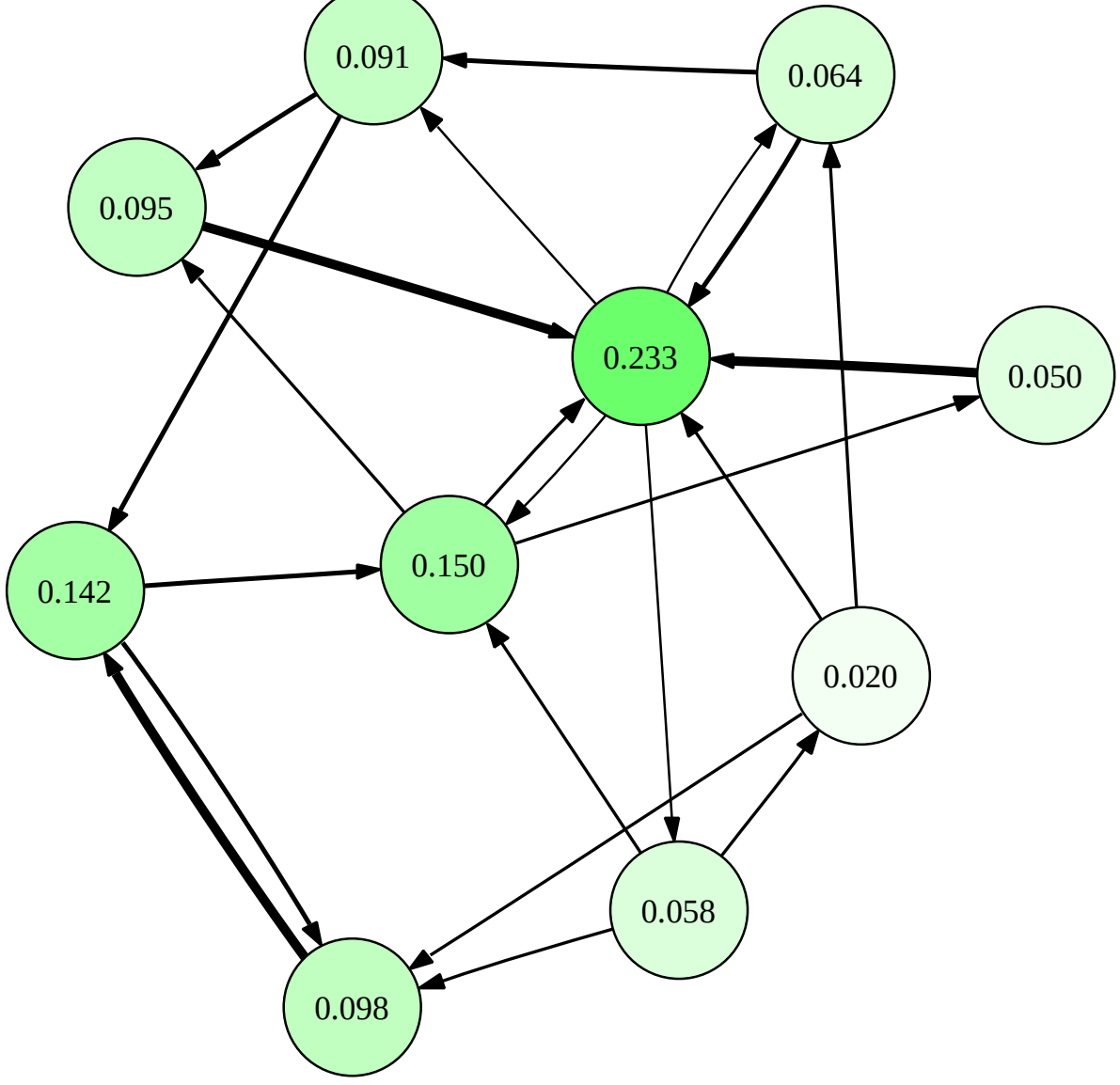


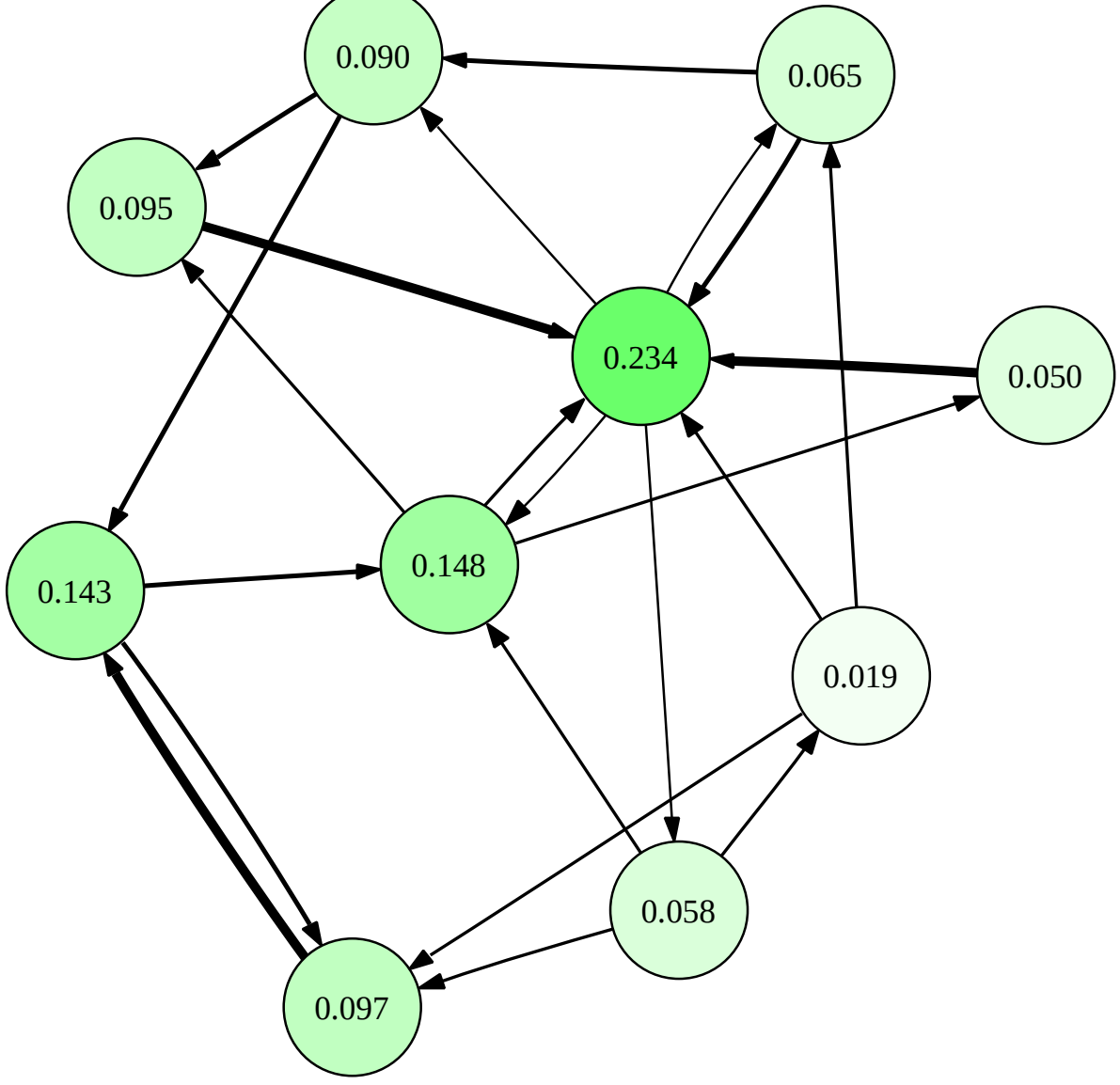


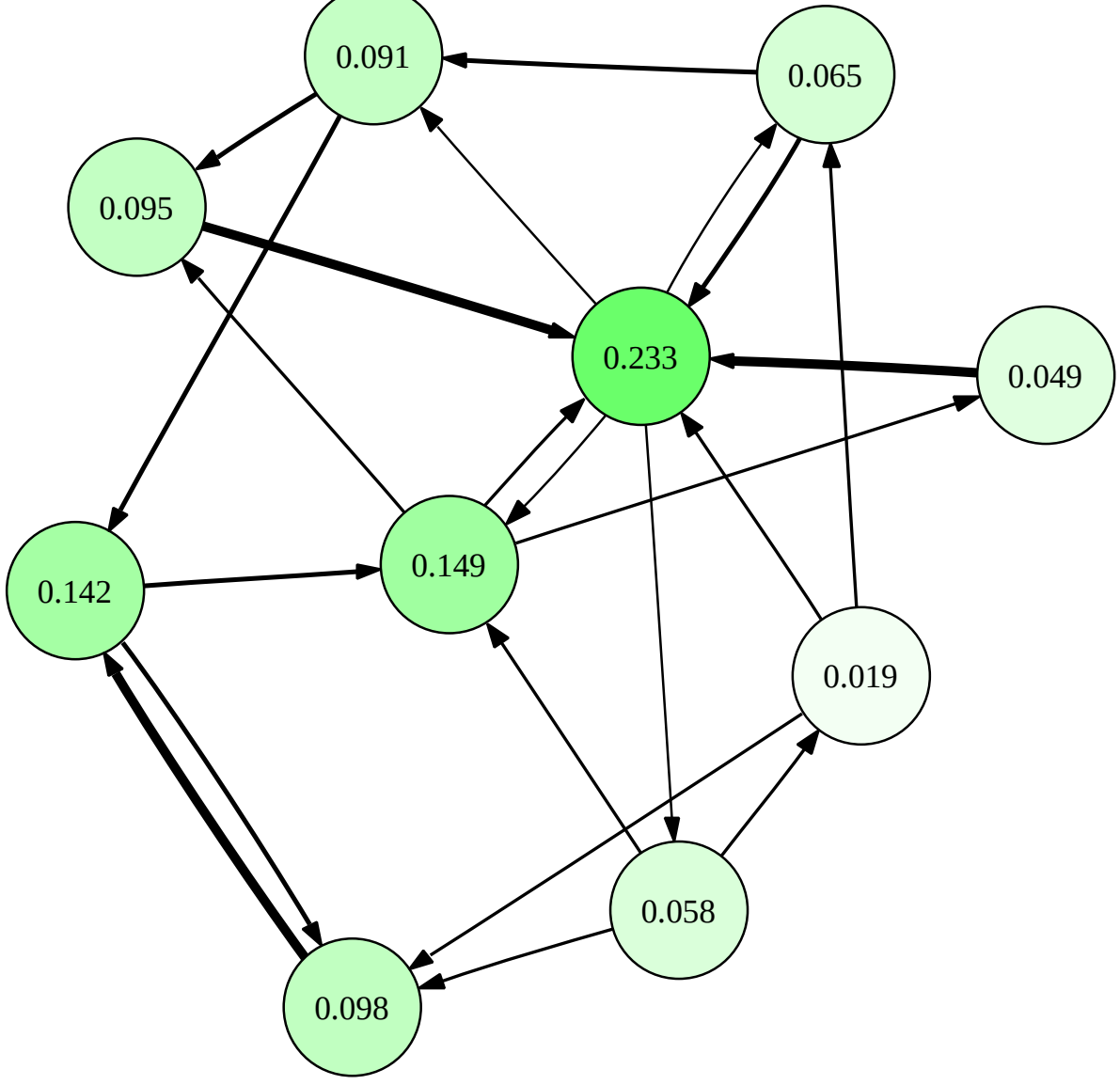

0.092
0.064
0.098
0.226
0.052
0.146
0.148
0.021
0.058
0.096

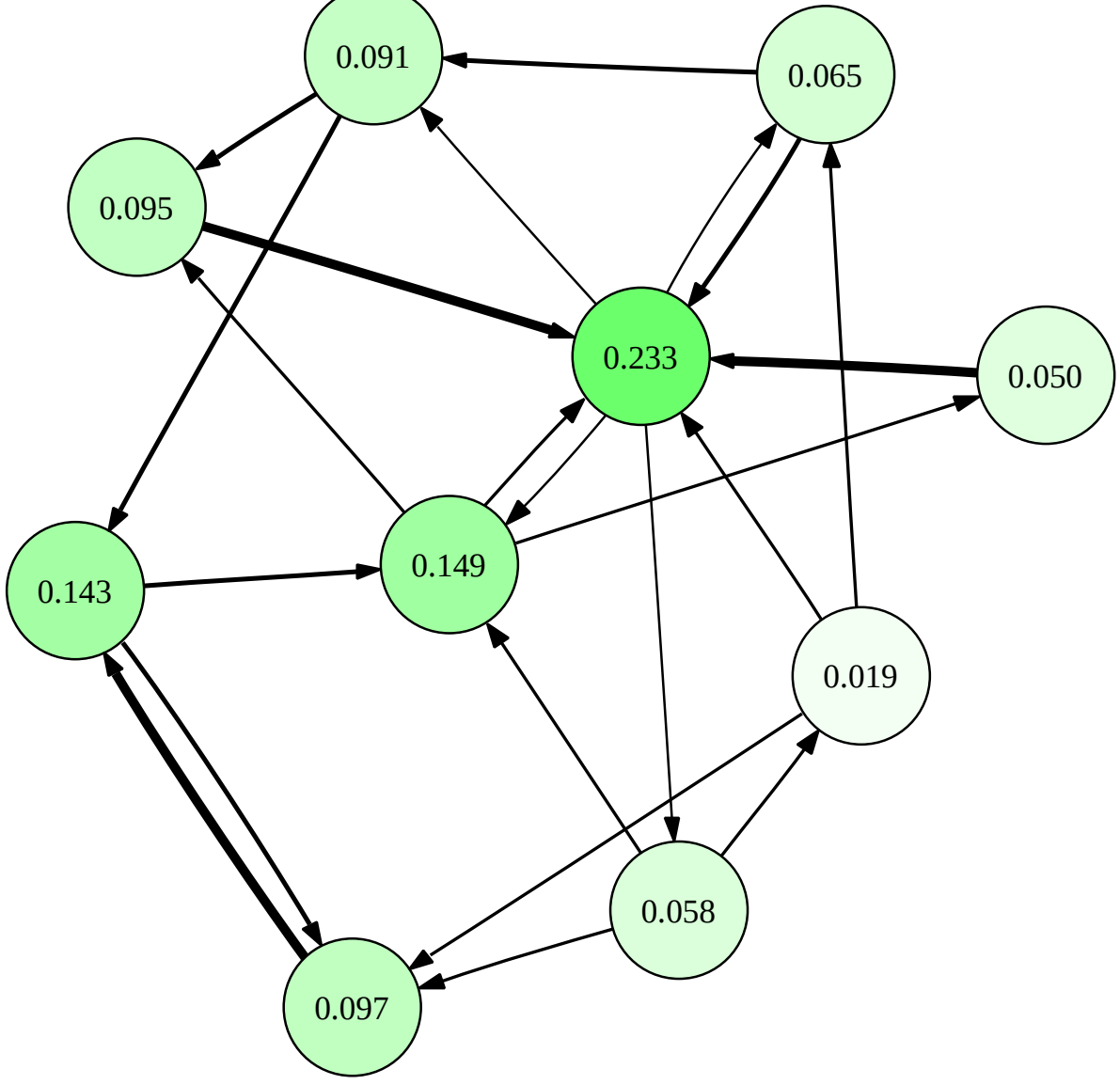


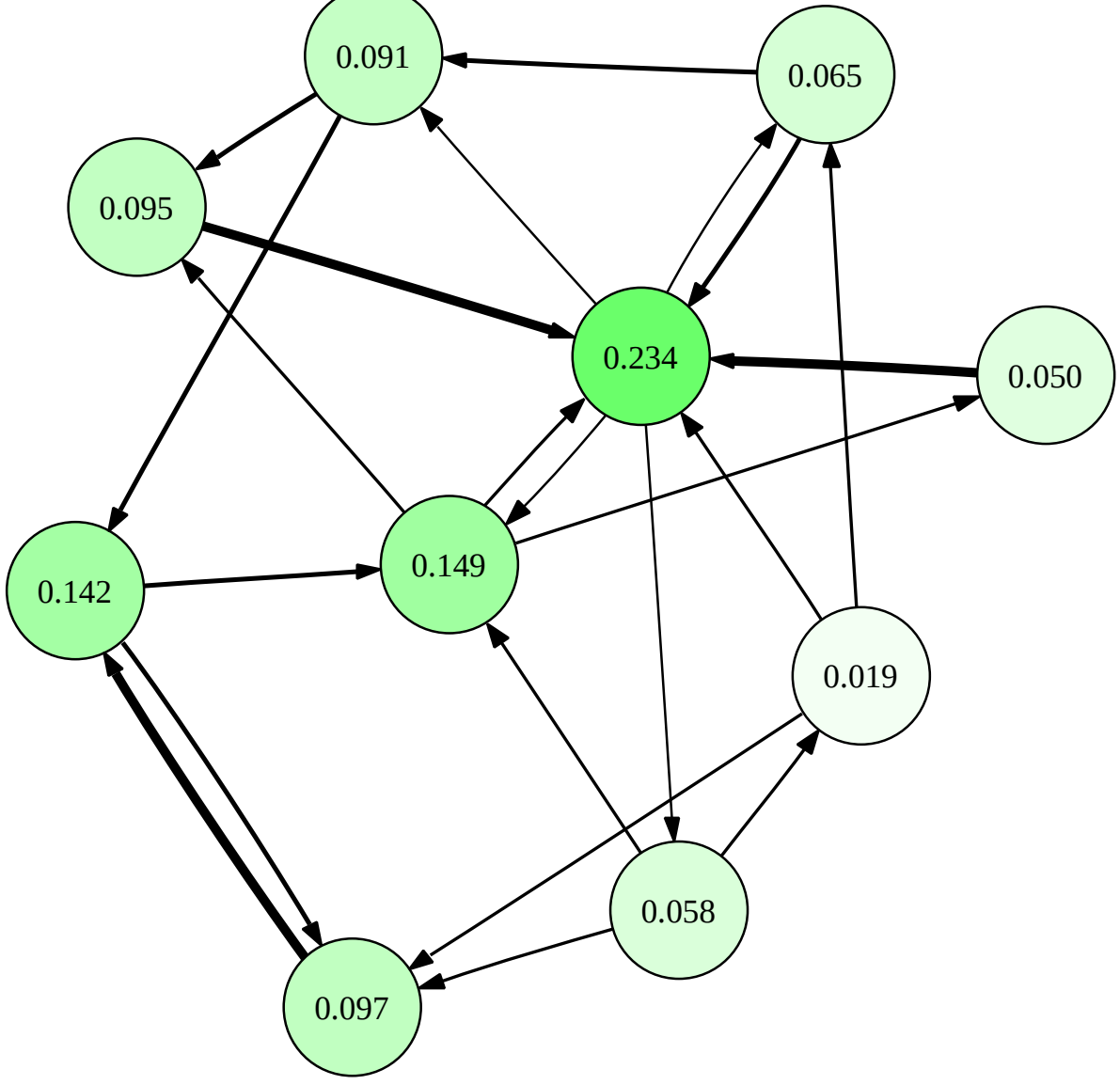











0.091
0.065
0.095
0.234
0.050
0.142
0.149
0.019
0.058
0.097

May not always converge, or convergence may not be unique.
To fix this, the random surfer can at each step randomly jump to any page of the Web with some probability d ($1 - d$: damping factor).

$$\mathrm{pr}(i) = \left(\lim_{k \to +\infty} ((1-d)G^T + dU)^k v \right)_i$$

where U is the matrix with all $\frac{1}{N}$ values with N the number of vertices.

- PageRank: global score, independent of the query
- Can be used to raise the weight of important pages:

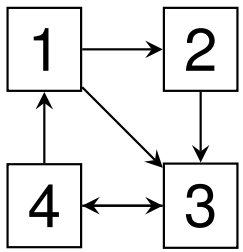
$$\text{weight}(t, d) = \text{tfidf}(t, d) \times \text{pr}(d),$$

- This can be directly incorporated in the index.

Compute G (often stored as its adjacency list). Make sure lines sum to 1.

2. Let u be the uniform vector of sum 1, $v = u$, w the *zero* vector.
3. While v is different enough from w:
   - Set $w = v$.
   - Set $v = (1 - d)G^T v + du$.

Exercise



Run the first iteration of the PageRank computation.

Online Page Importance Computation

- Allows computing PageRank without storing the Web graph
- Remember for each page its cash and its history
- Initially, each page has some initial cash
- Each time a page is accessed, distribute uniformly its cash to outgoing links
- PageRank is approximated as its cashflow: history divided by number of visits
- Converge whatever the strategy for crawling the Web is (as soon as each page is accessed repeatedly)
- Good strategy: crawl page with largest cash first

Two kinds of important pages: hubs and authorities. Hubs are pages that point to good authorities, whereas authorities are pages that are pointed to by good hubs.

G' transition matrix (with 0 and 1 values) of a subgraph of the Web. We use the following iterative process (starting with a and h vectors of norm 1):

$$\begin{cases} a := \frac{1}{\|G'^T h\|} G'^T h \\ h := \frac{1}{\|G' a\|} G' a \end{cases}$$

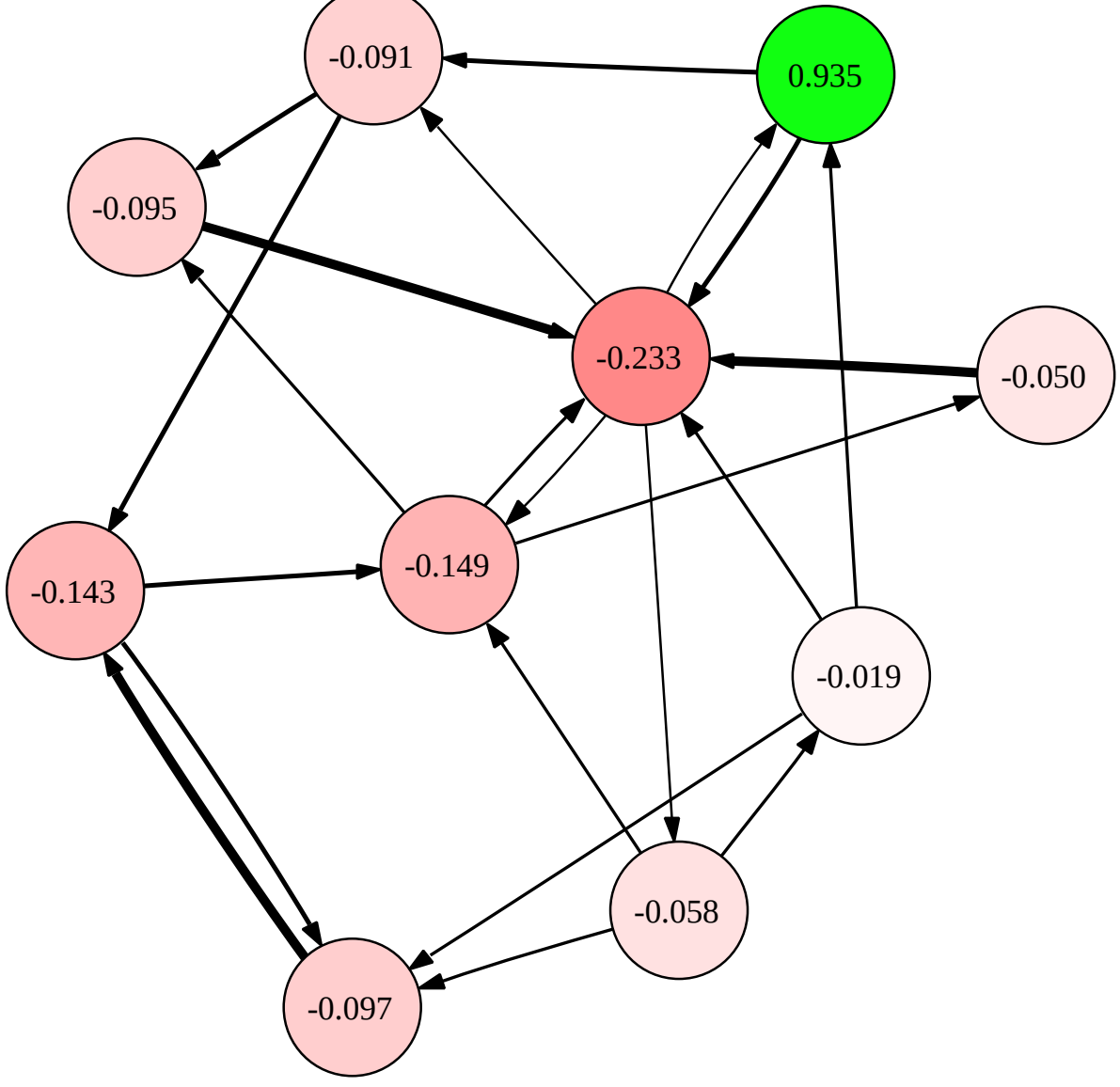
Converges under some technical assumptions to authority and hub scores.

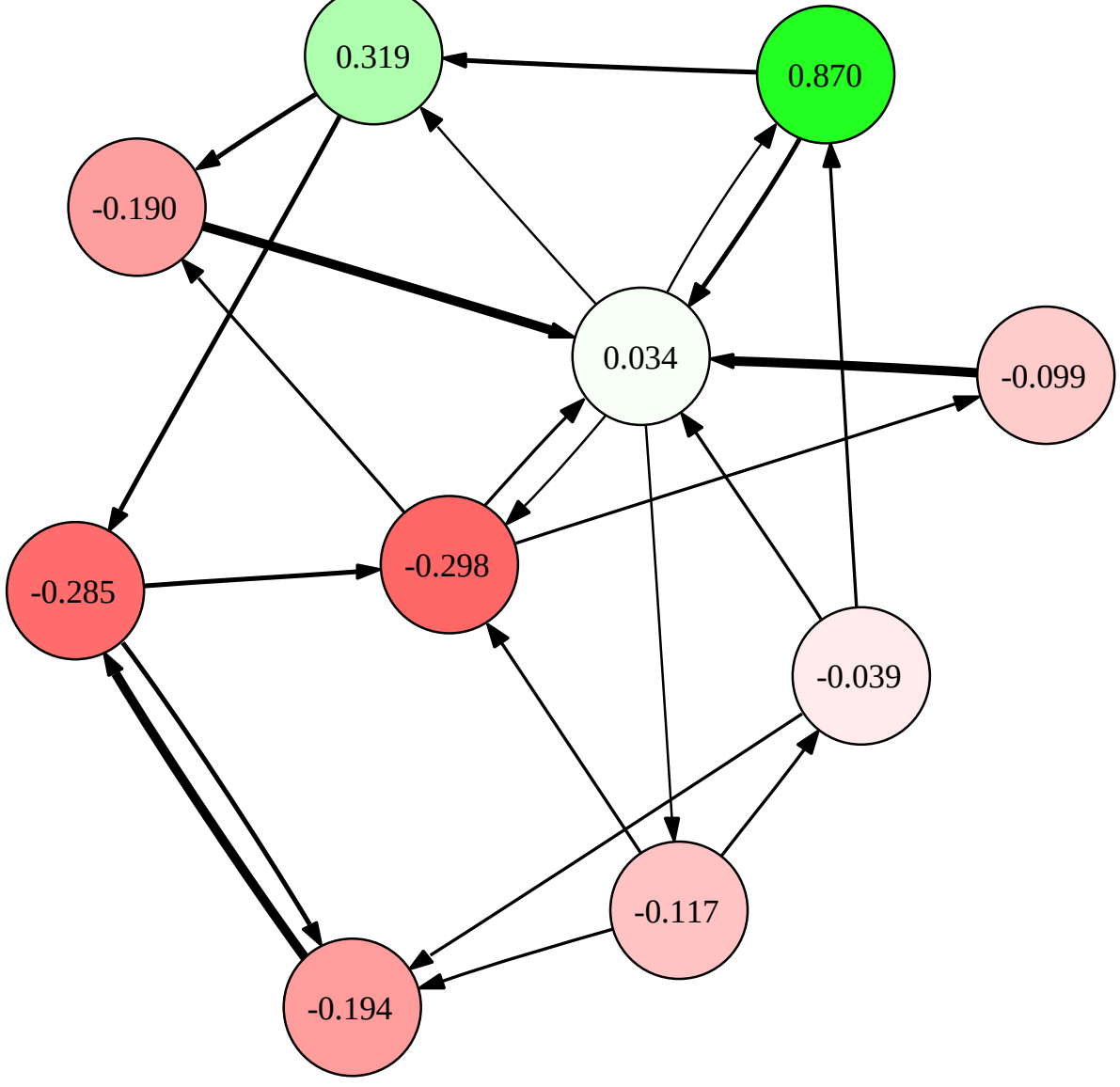
1. Retrieve the set D of Web pages matching a keyword query.
2. Retrieve the set D^* of Web pages obtained from D by adding all linked pages, as well as all pages linking to pages of D.
3. Build from D^* the corresponding subgraph G' of the Web graph.
4. Compute iteratively hubs and authority scores.
5. Sort documents from D by authority scores.

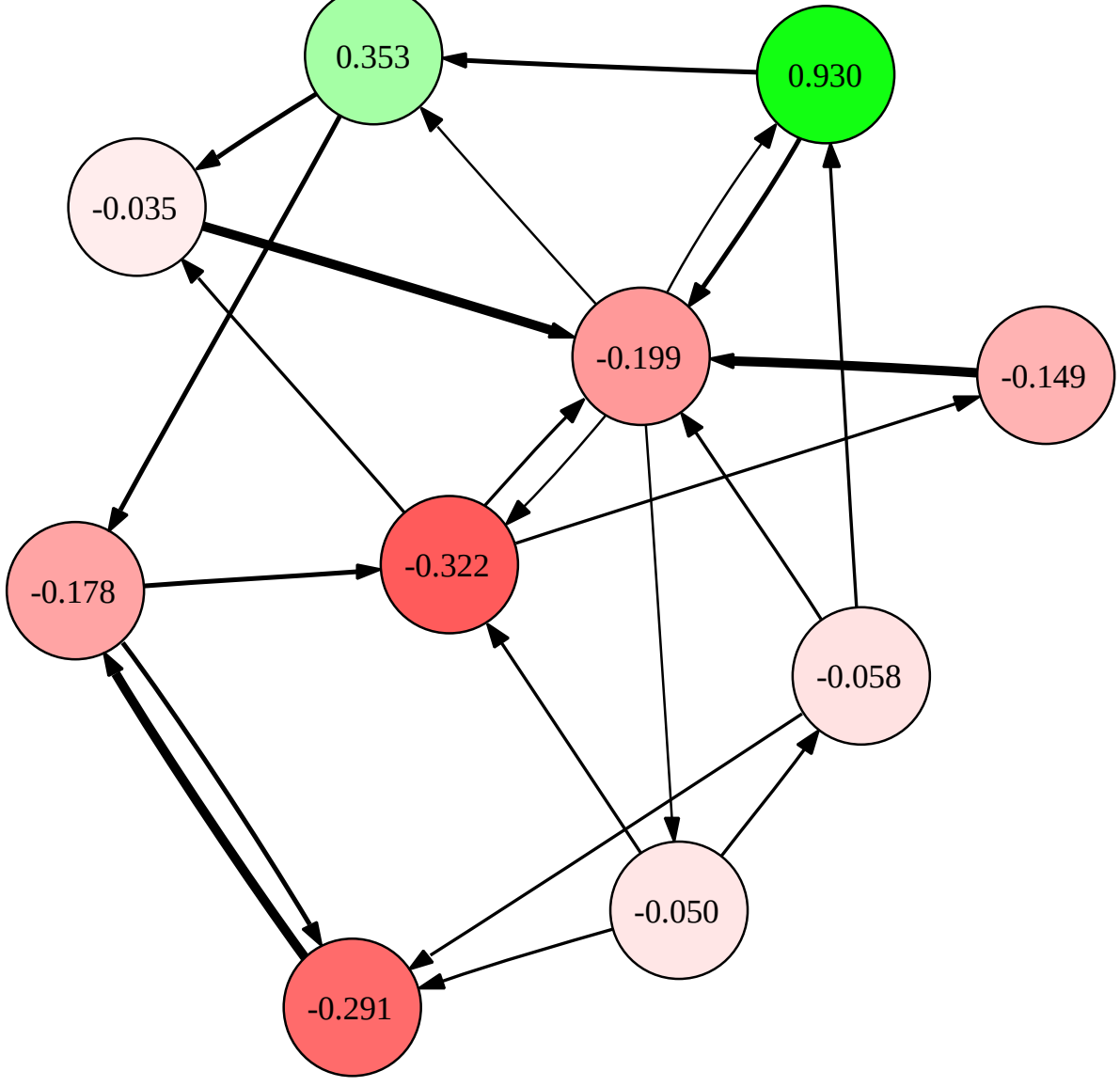
Less efficient than PageRank, because local scores.

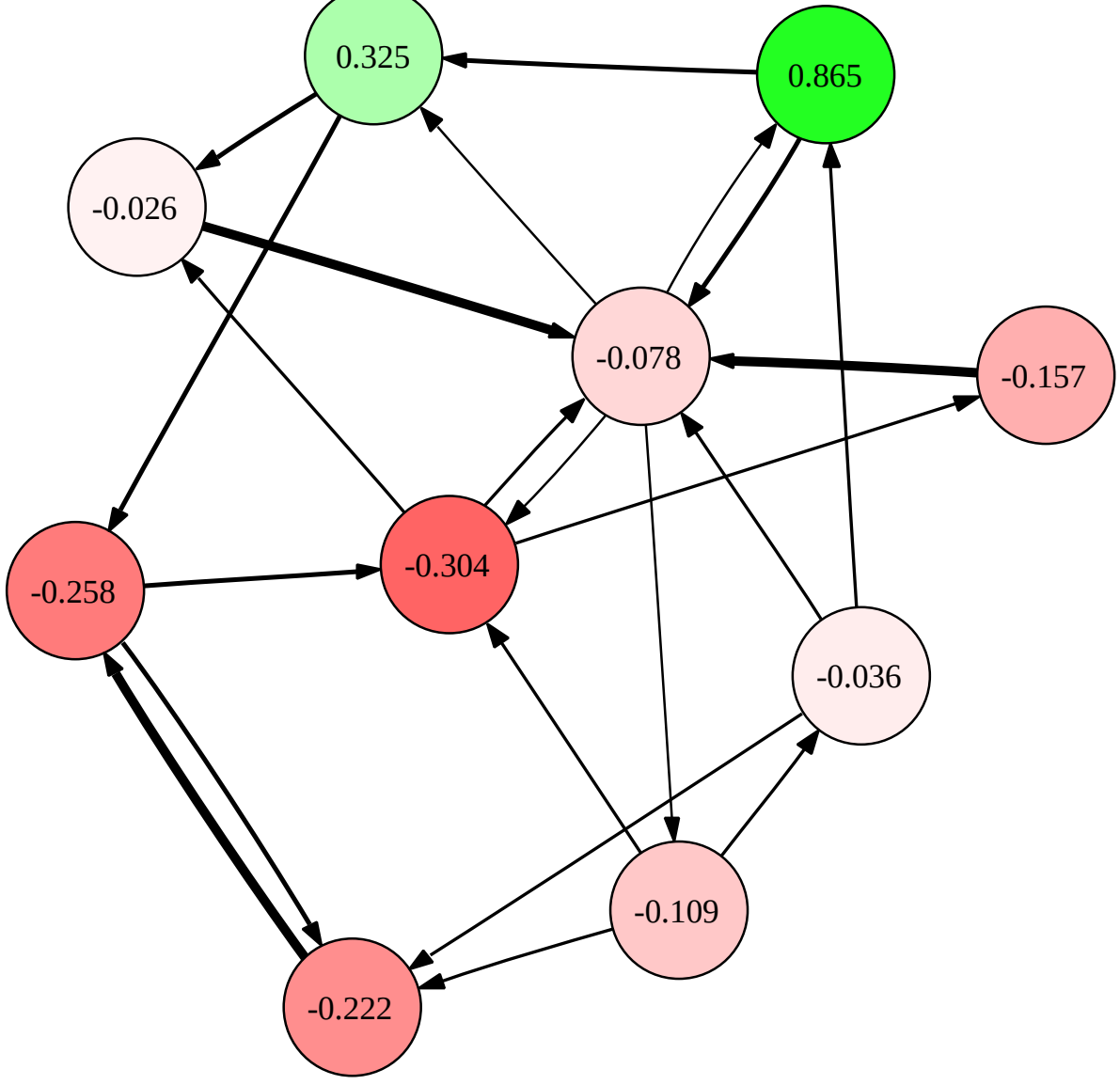
Equivalent definitions of Green measure centered at node i

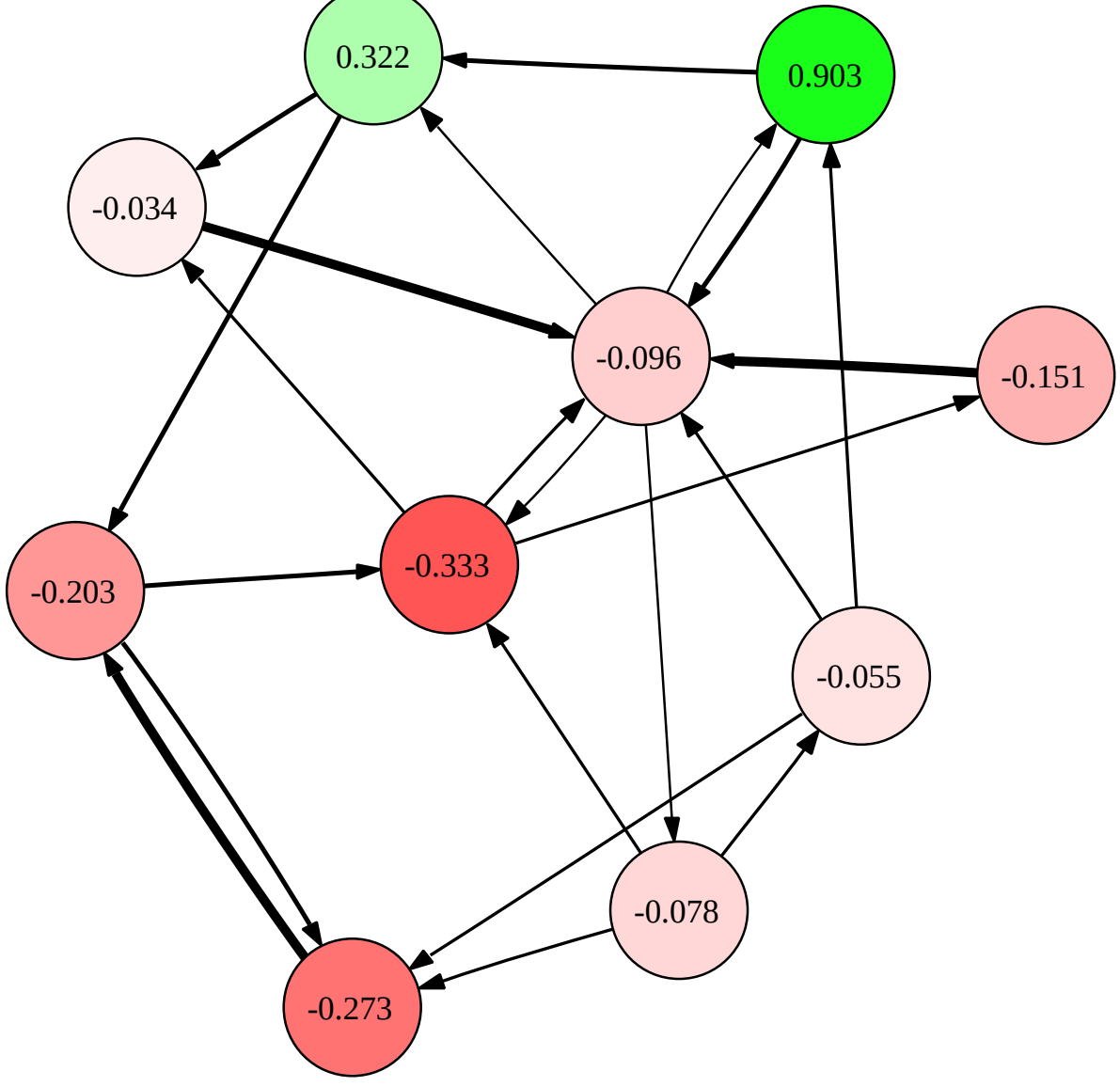
- PageRank with source at i: standard PageRank computation while, at each iteration, adding 1 to the measure of i, and subtracting $pr(j)$ to every node j.
- Time spent at a node knowing the initial node is i.

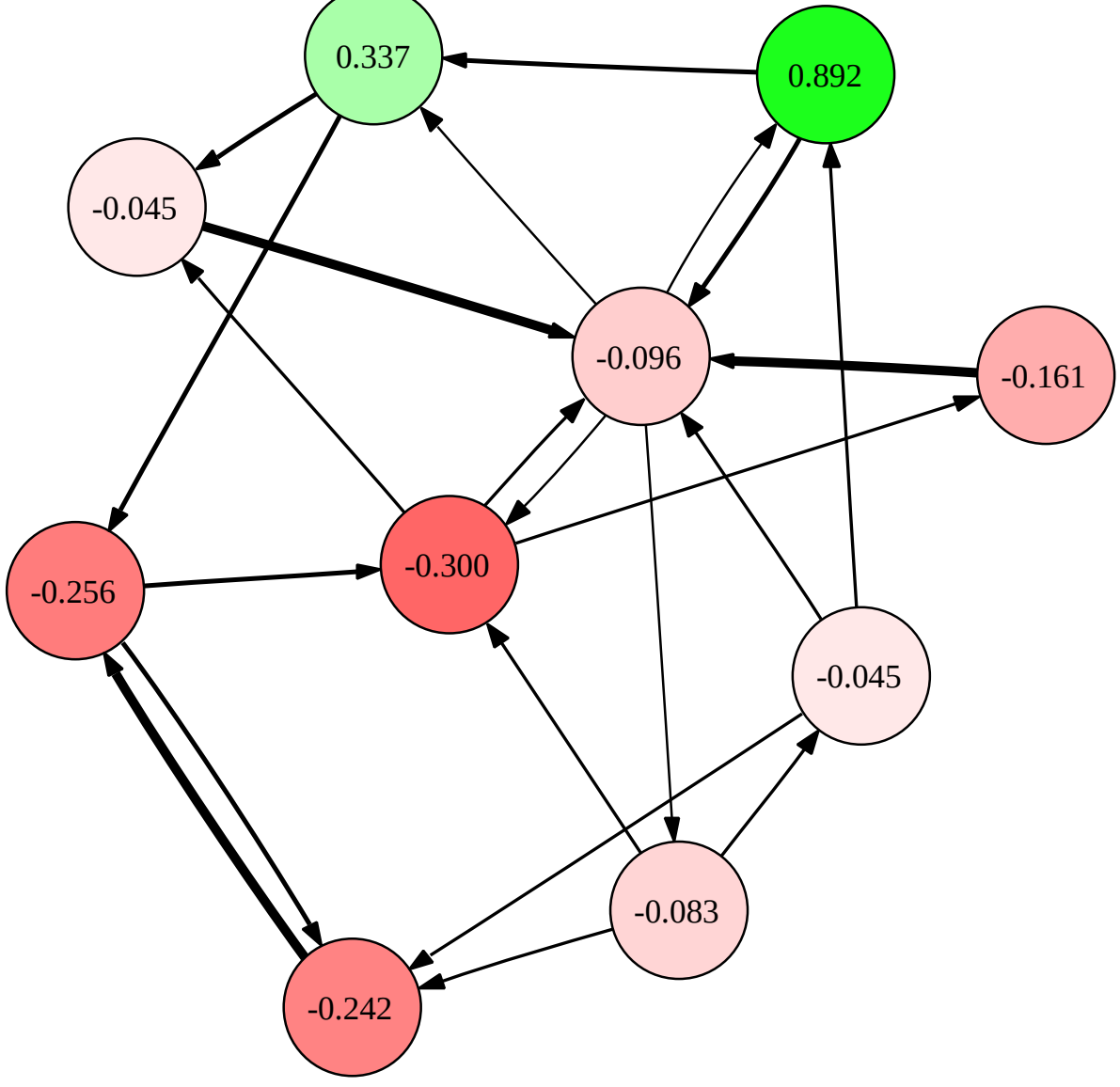


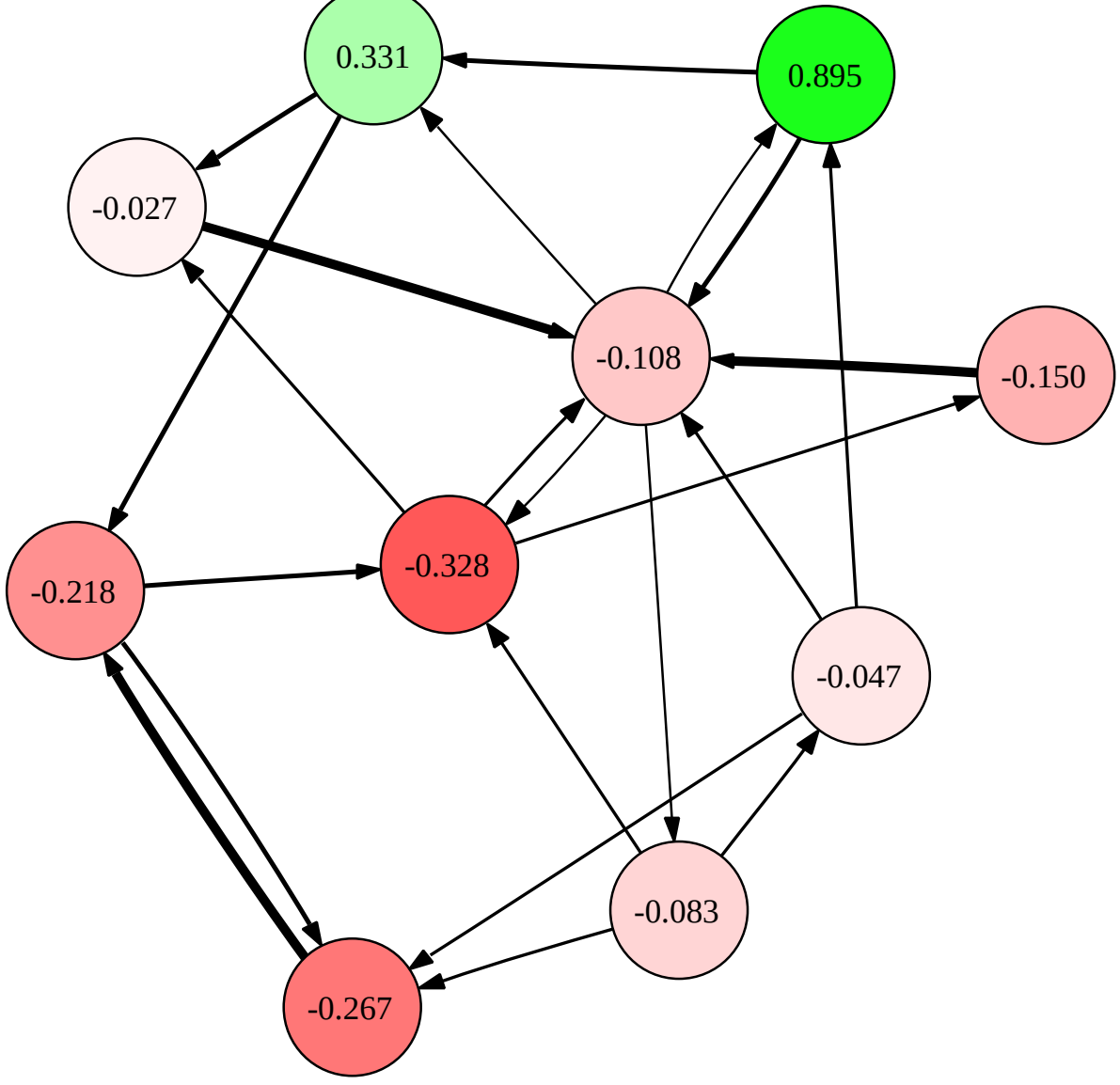

0.319
0.870
-0.190
0.034
-0.099
-0.285
-0.298
-0.039
-0.117
-0.194

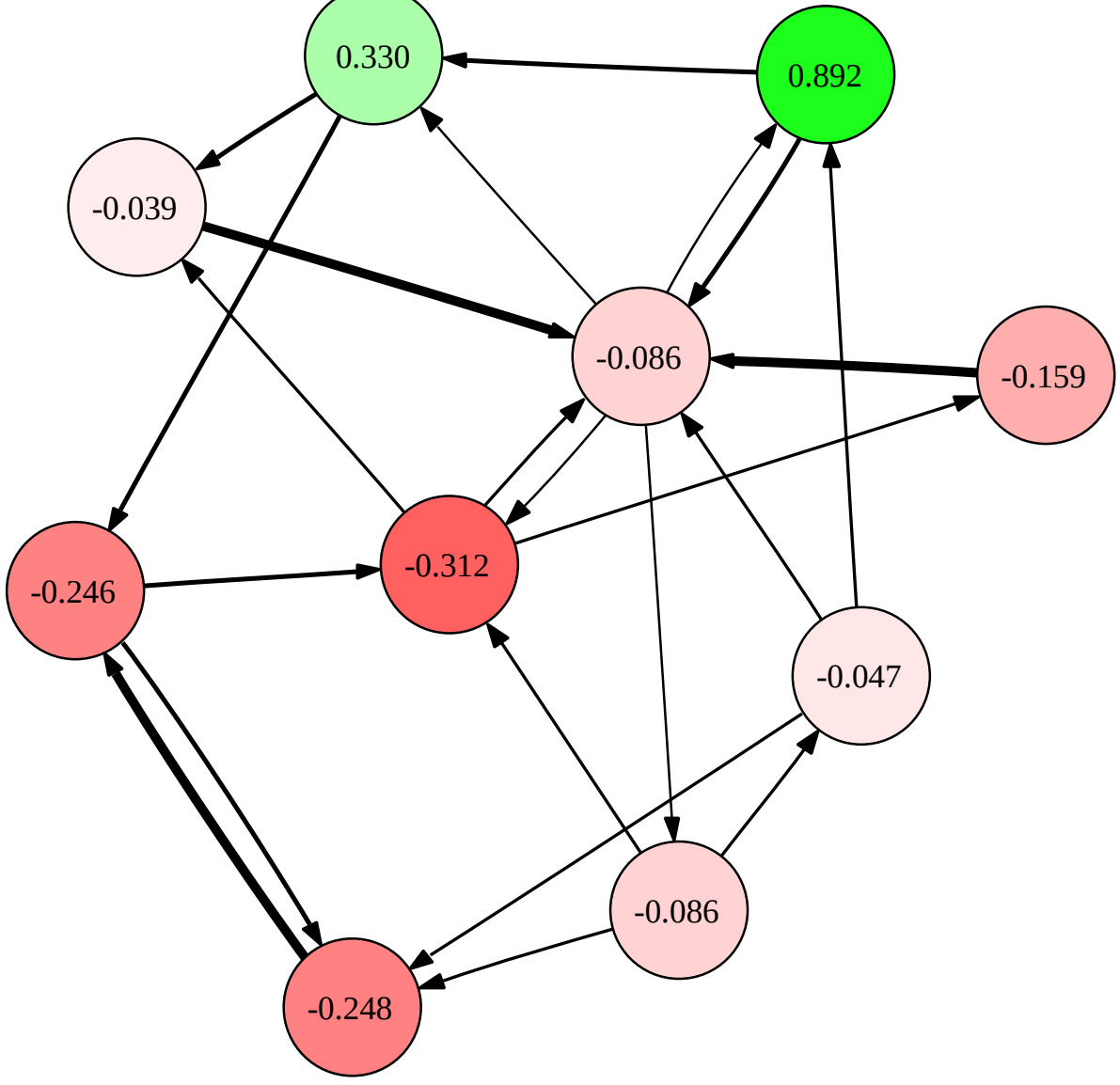


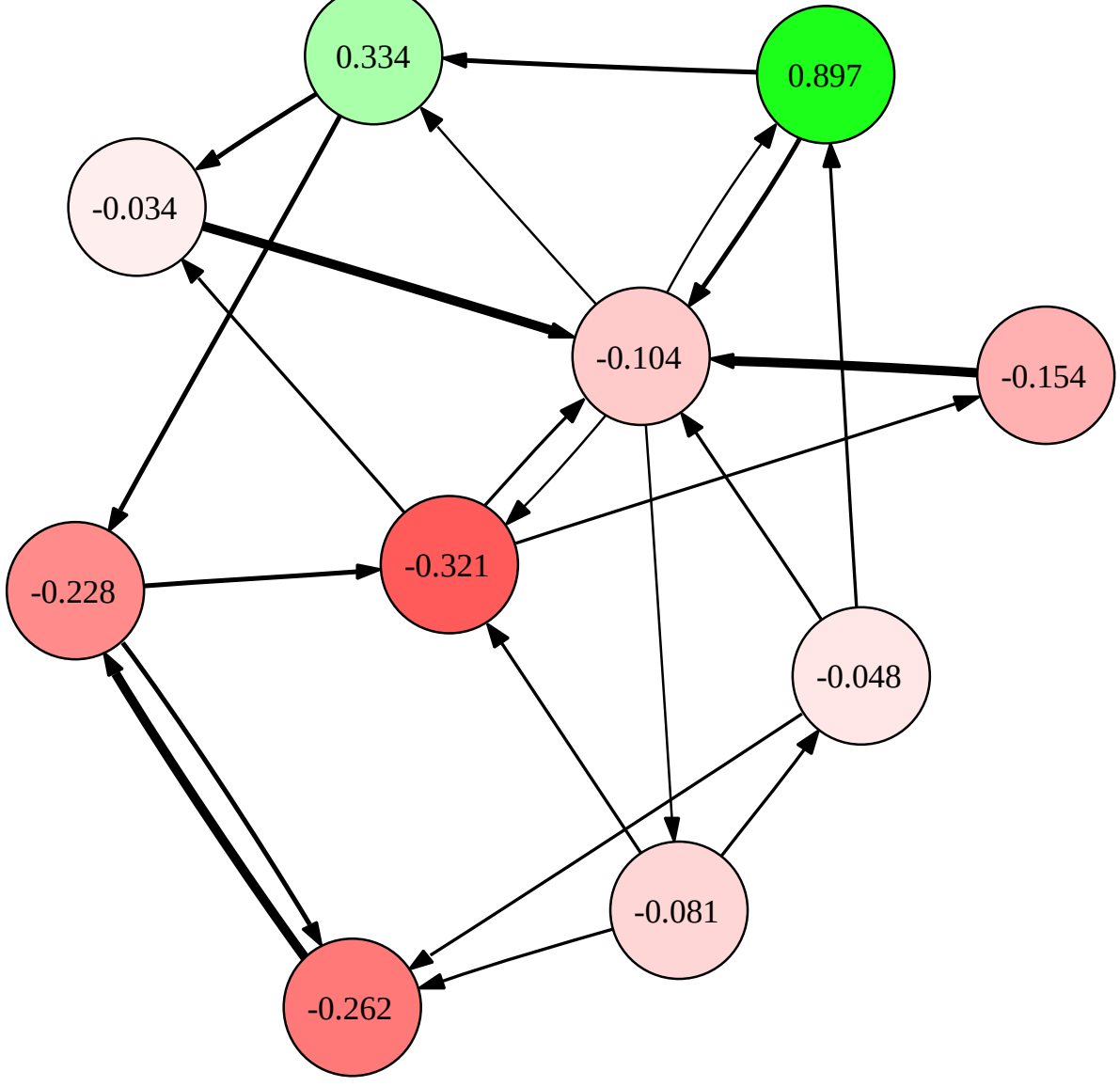

0.325
0.865
-0.026
-0.078
-0.157
-0.258
-0.304
-0.036
-0.109
-0.222

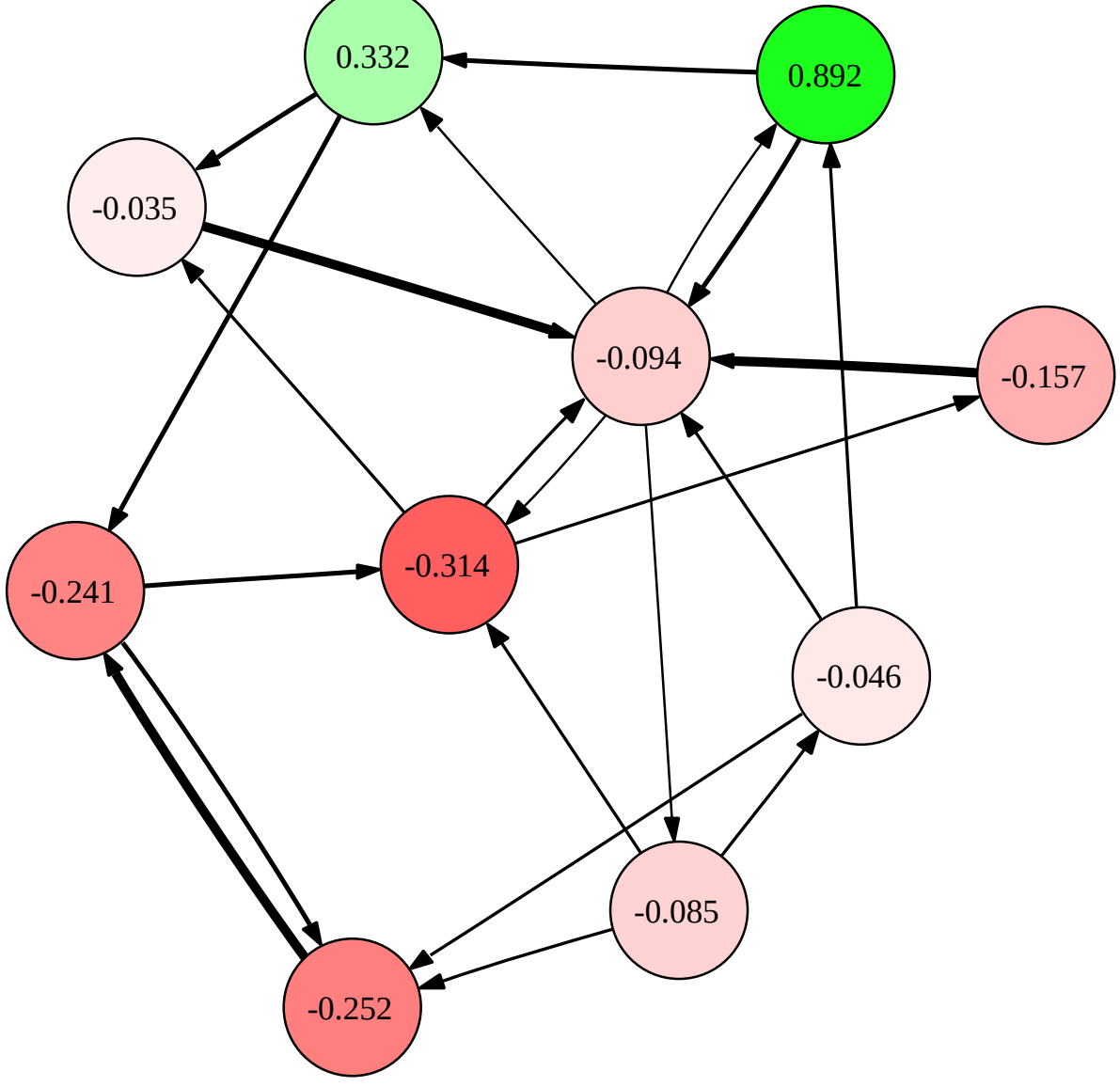


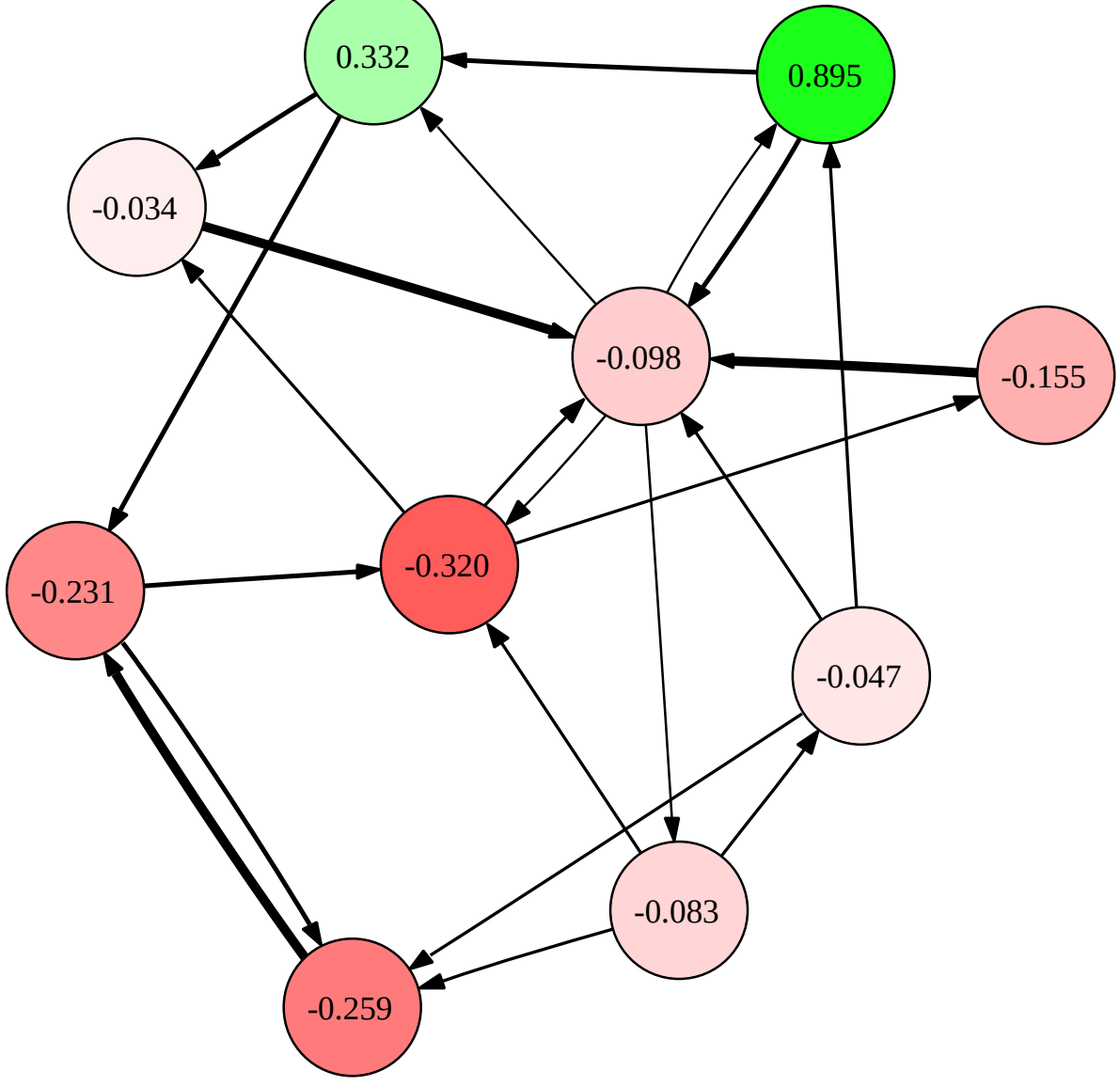


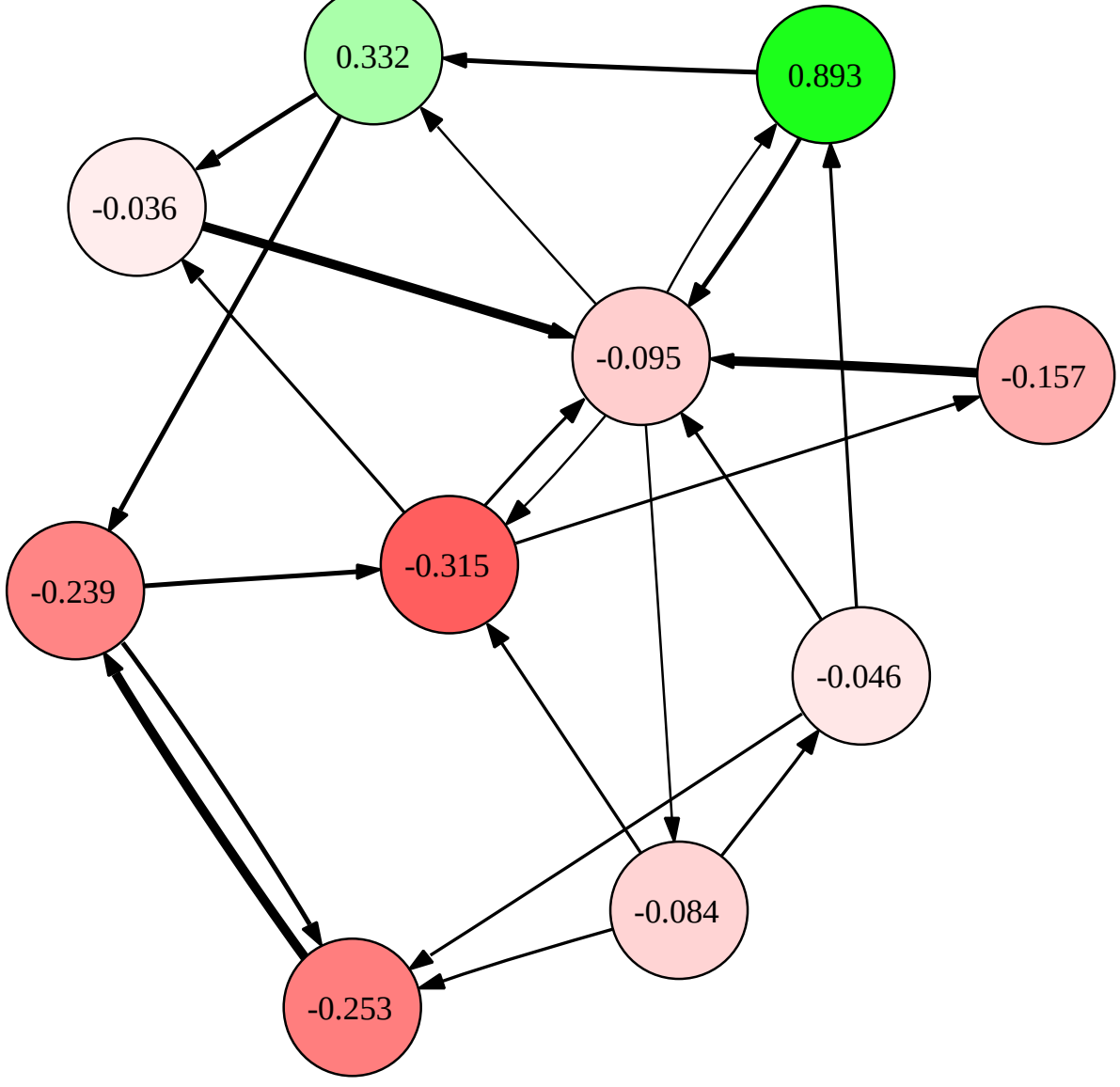


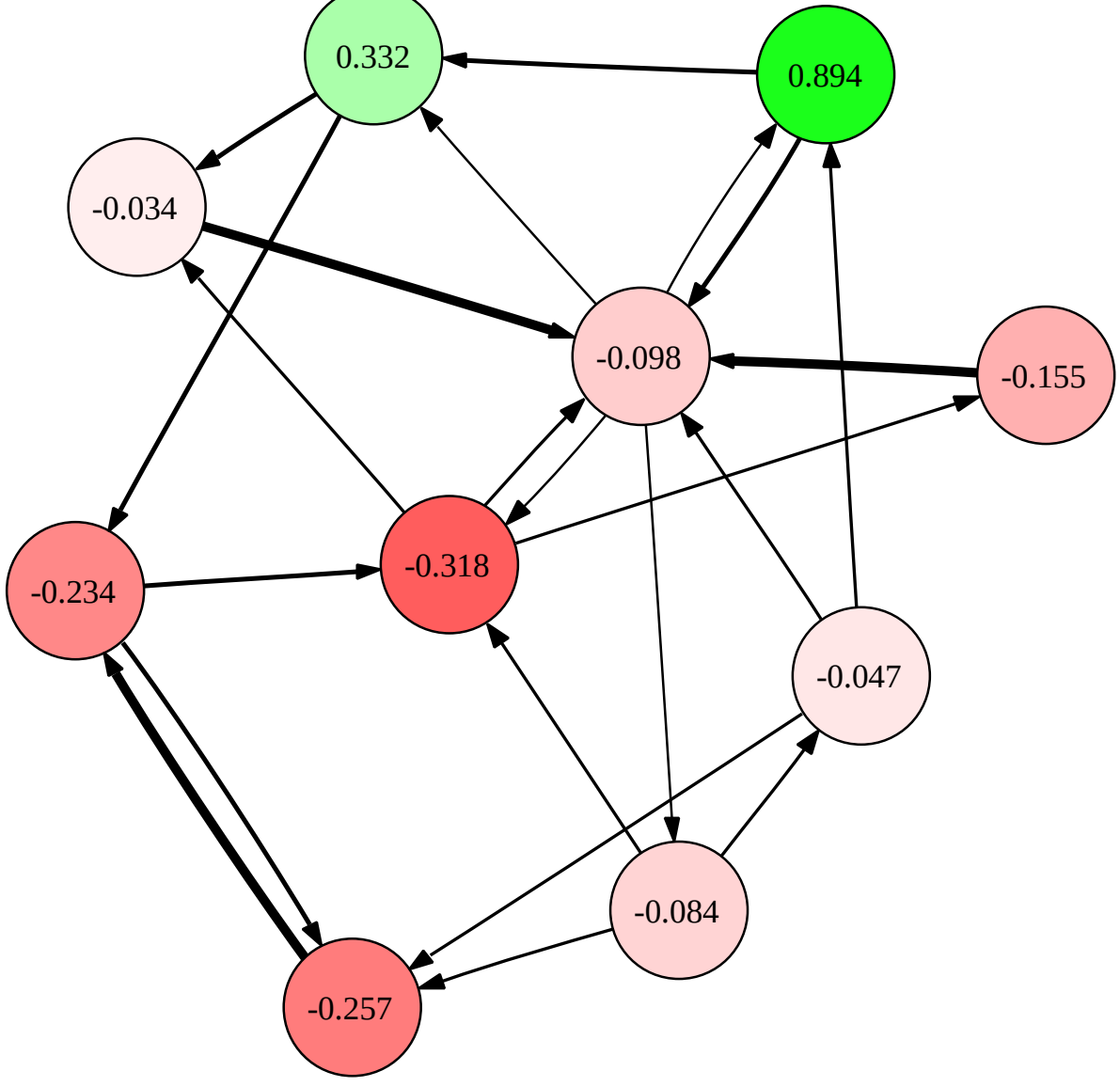


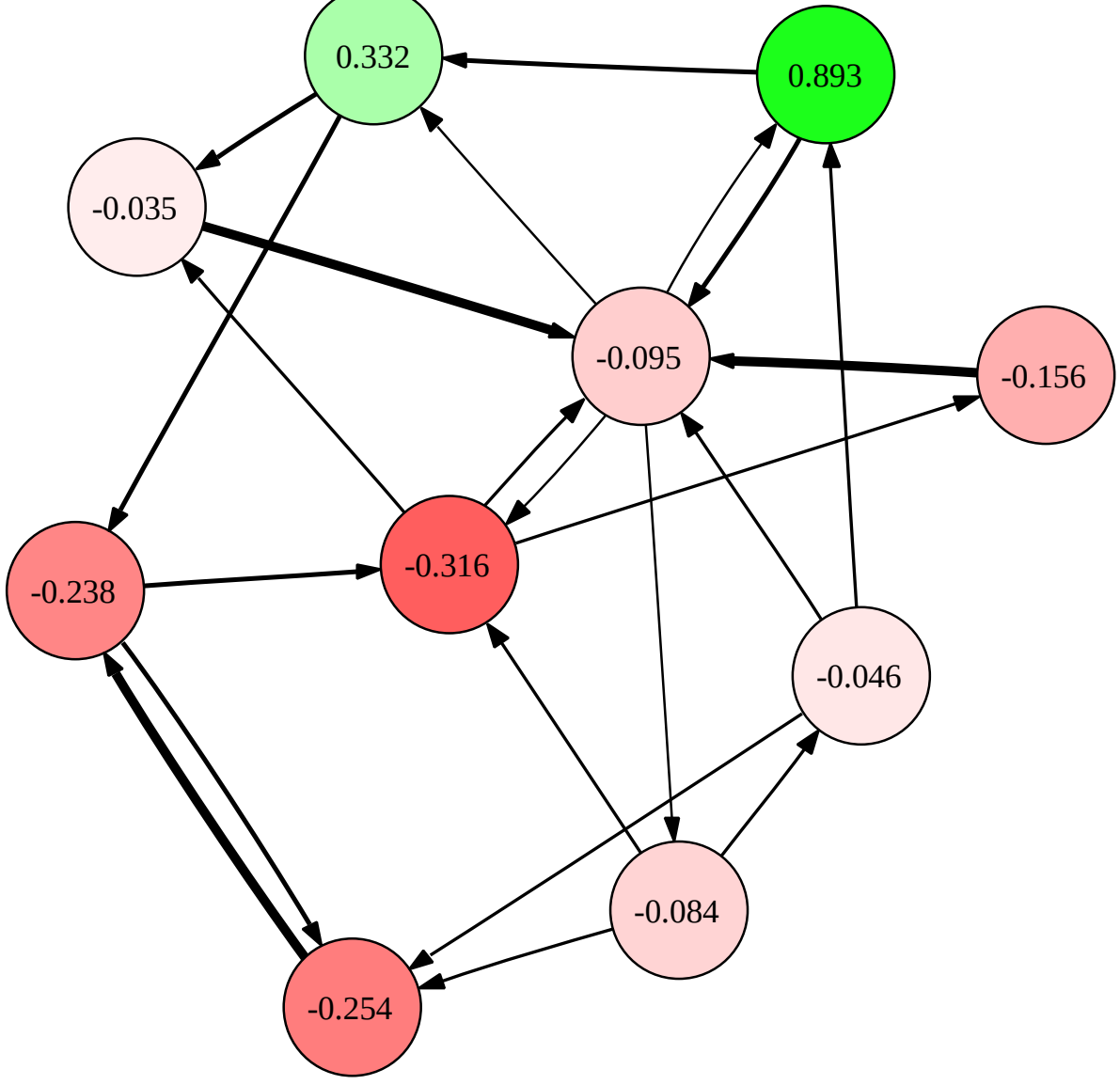


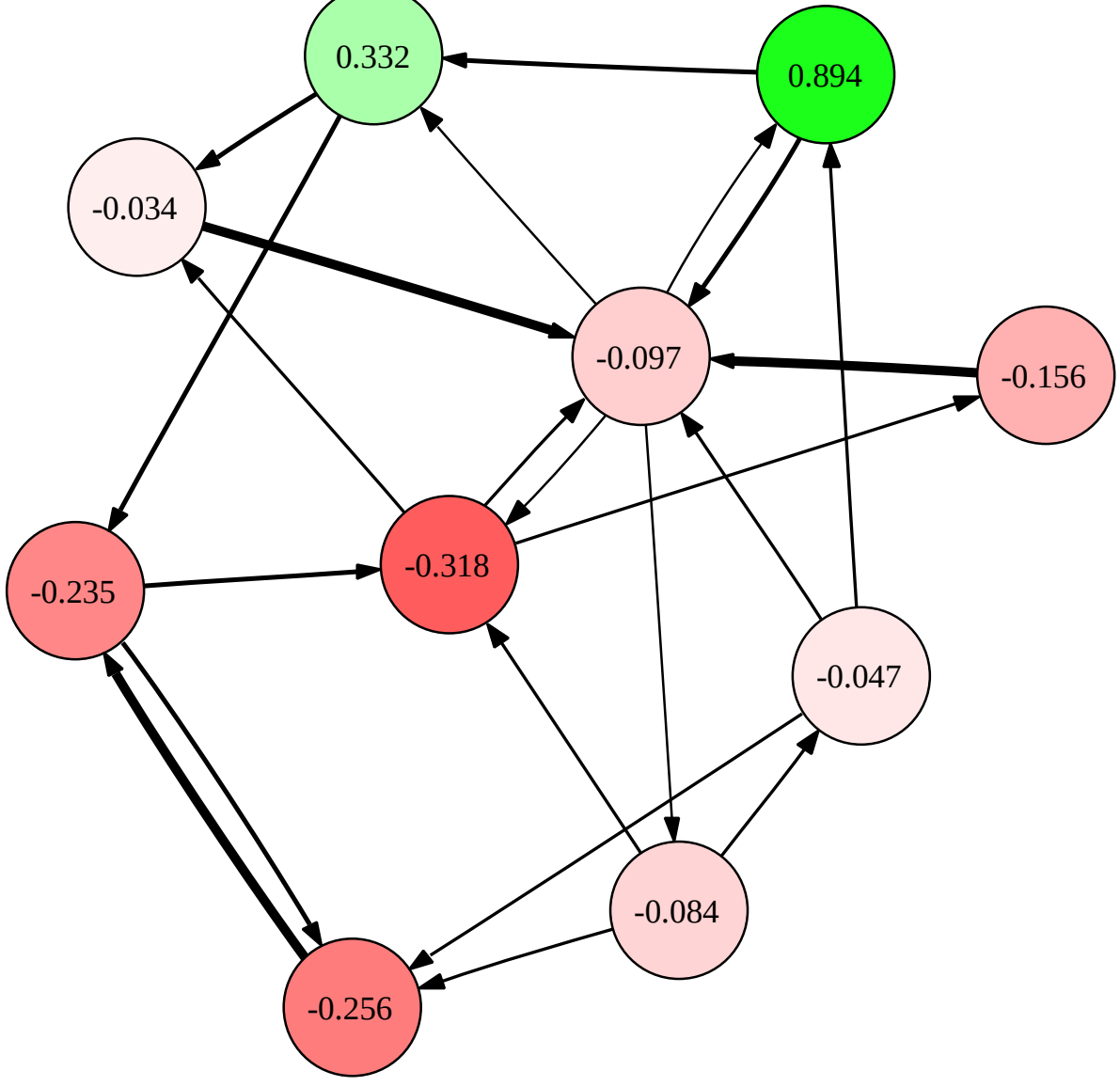


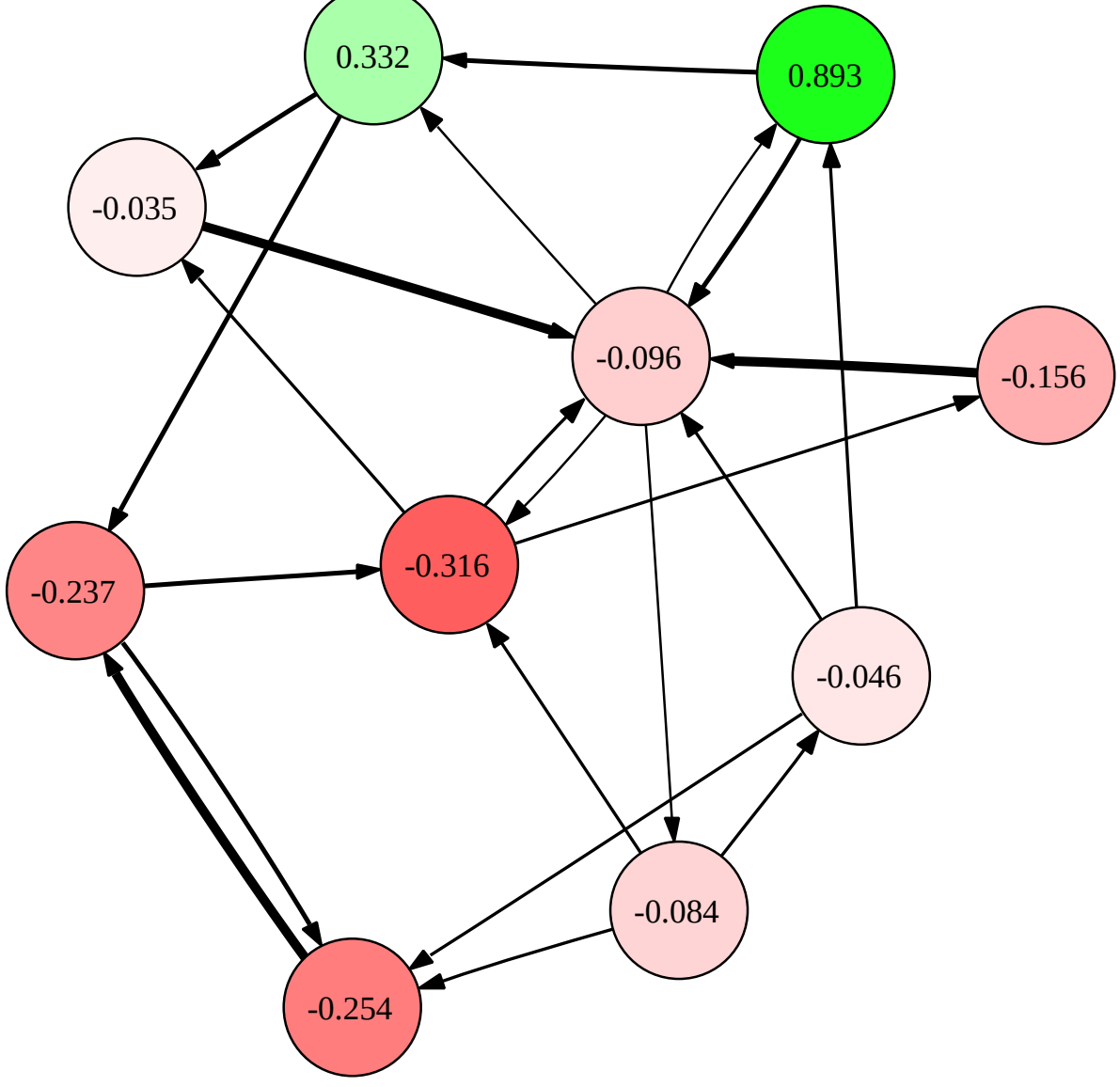


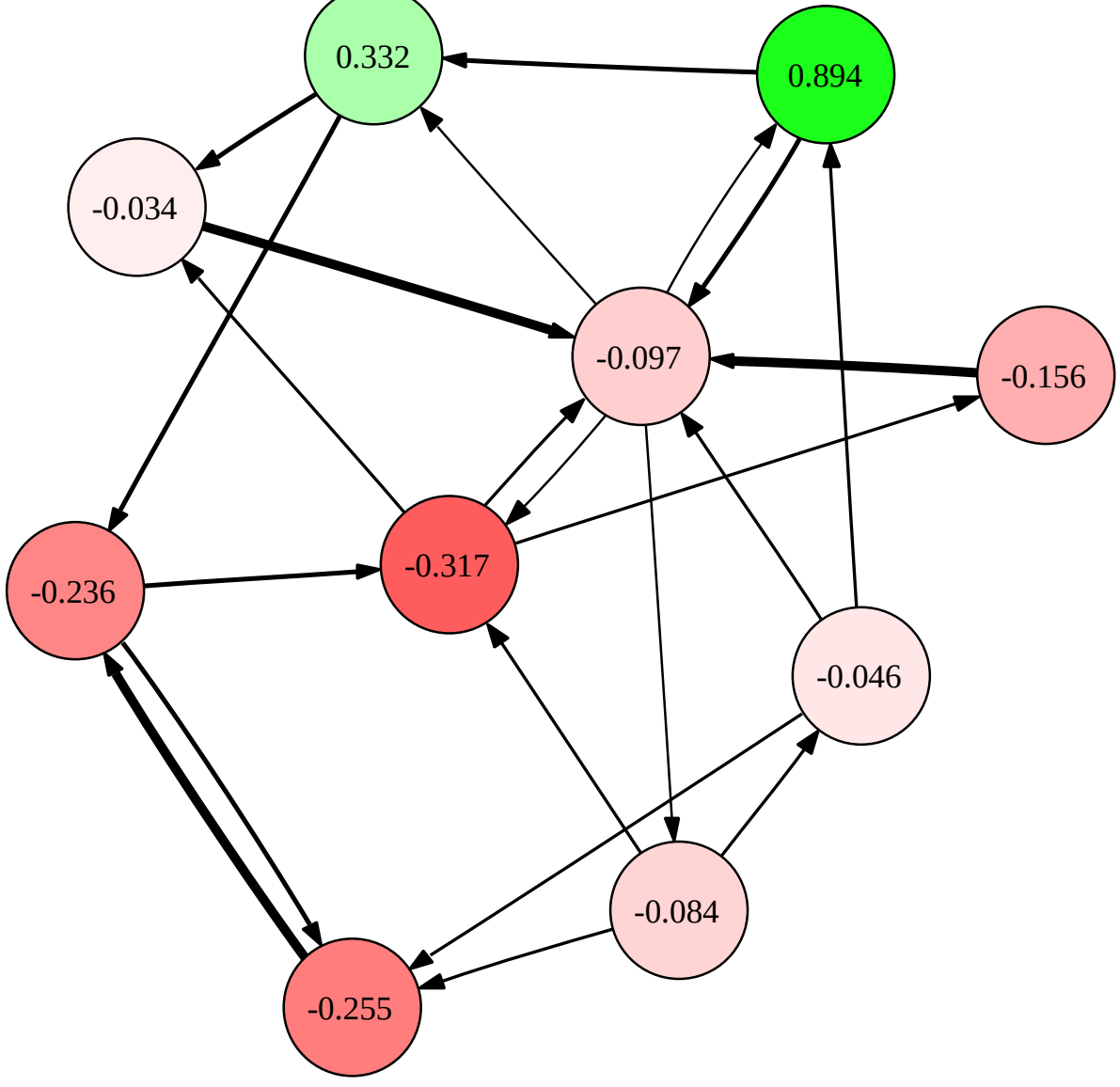


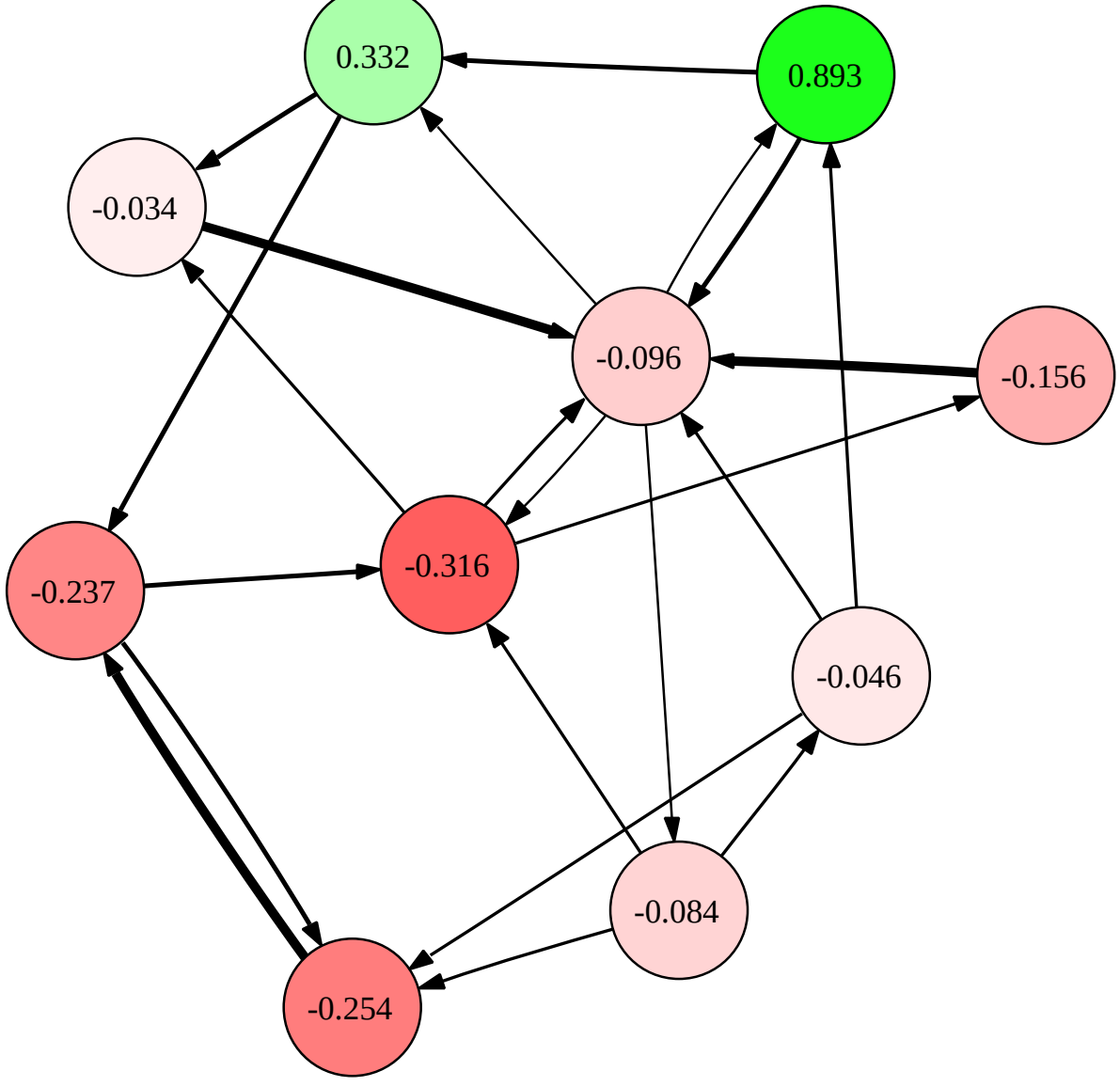


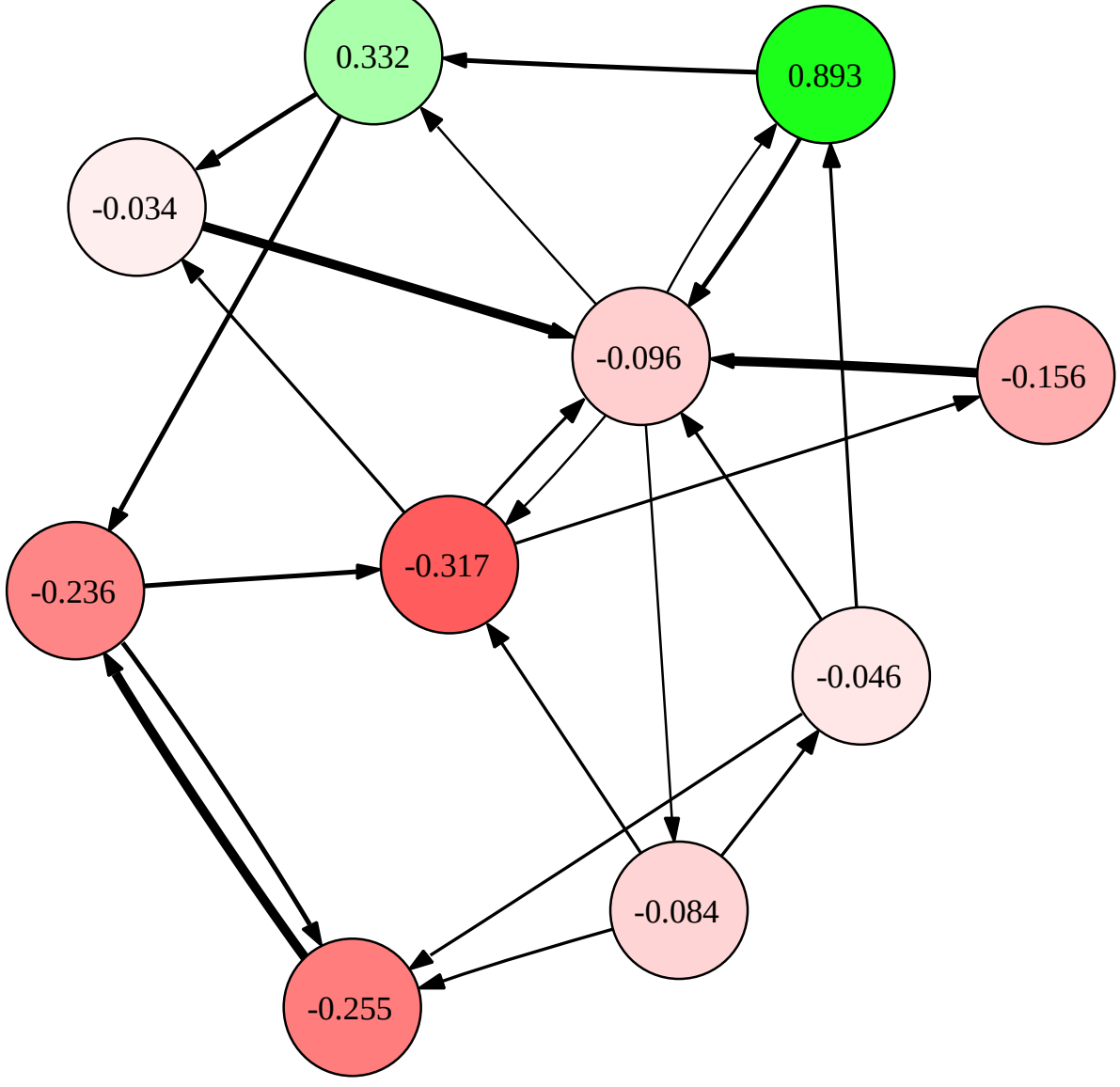


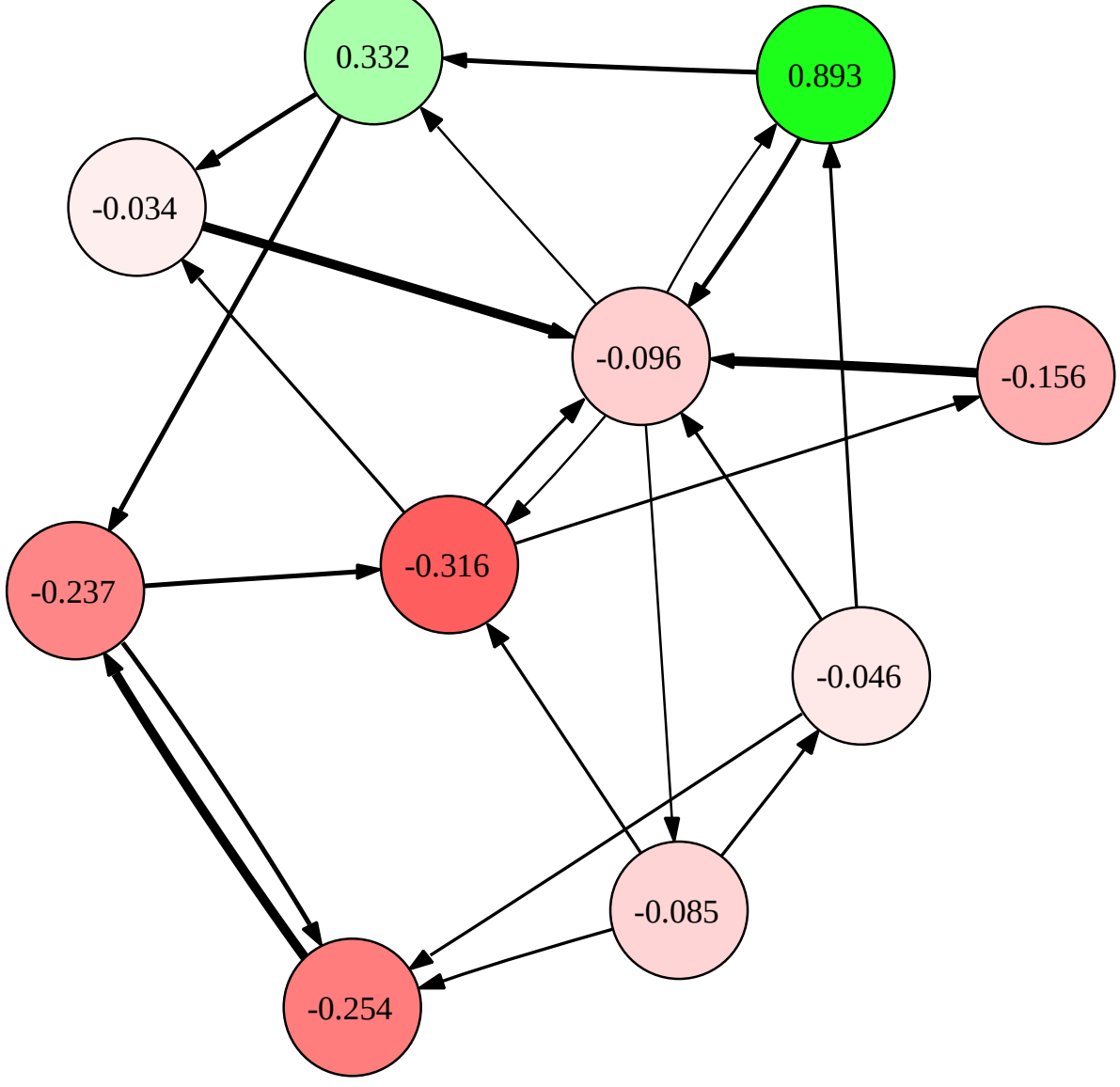












Definition

Fraudulent techniques that are used by unscrupulous webmasters to artificially raise the visibility of their website to users of search engines

Purpose: attracting visitors to websites to make profit.

Unceasing war between spamdexers and search engines

Put unrelated terms in:

- meta-information (`<meta name="description">`, `<meta name="keywords">`)
- text content hidden to the user with JavaScript, CSS, or HTML presentational elements

Countertechnique

- Ignore meta-information
- Try and detect invisible text

Technique

Huge number of hosts on the Internet used for the sole purpose of referencing each other, without any content in themselves, to raise the importance of a given website or set of websites.

Countertechnique

- Detection of websites with empty or duplicate content
- Use of heuristics to discover subgraphs that look like link farms

Technique

Pollute user-editable websites (blogs, wikis) or exploit security bugs to add artificial links to websites, in order to raise its importance.

Countertechnique

`rel="nofollow"` attribute to `<a>` links not validated by a page's owner

Social Networking Sites

- Typology
- Social Network Graphs
- Algorithms on social networks

Conclusion

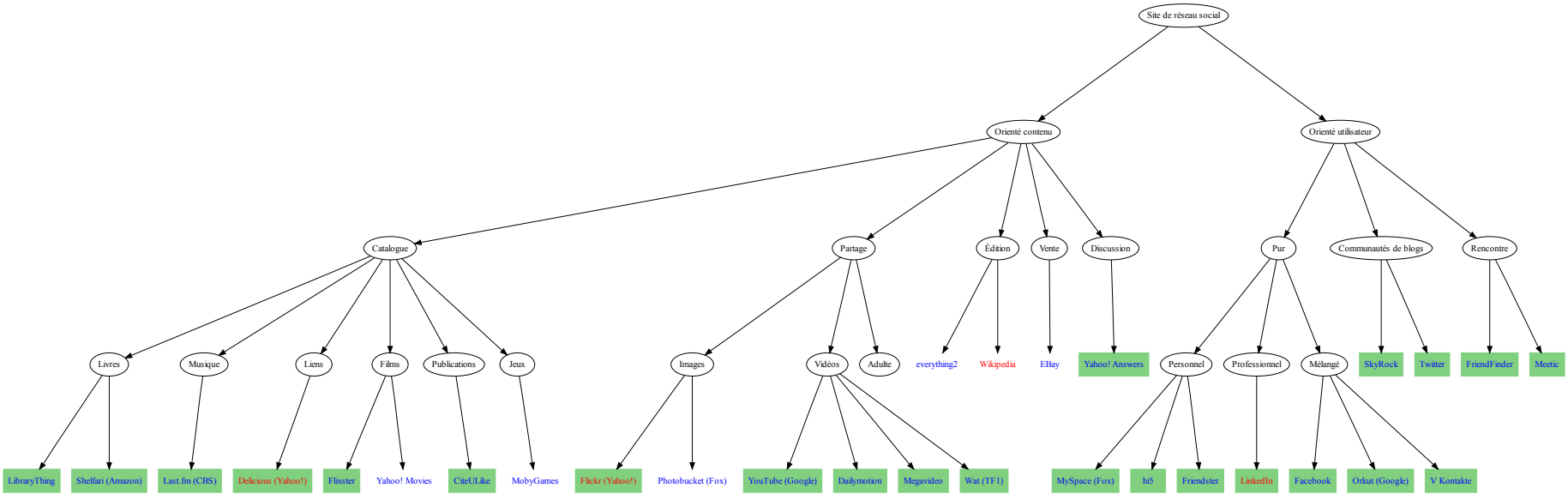
- Graph mining tools used for the Web: also applicable to a variety of different graphs:
  - dictionary
  - encyclopedia
  - citation graph
  - ...
- Among these: social networking Web sites
- Attention: on undirected graphs, PageRank is proportional to degree! Not very useful.

Social Networking Sites

Typology

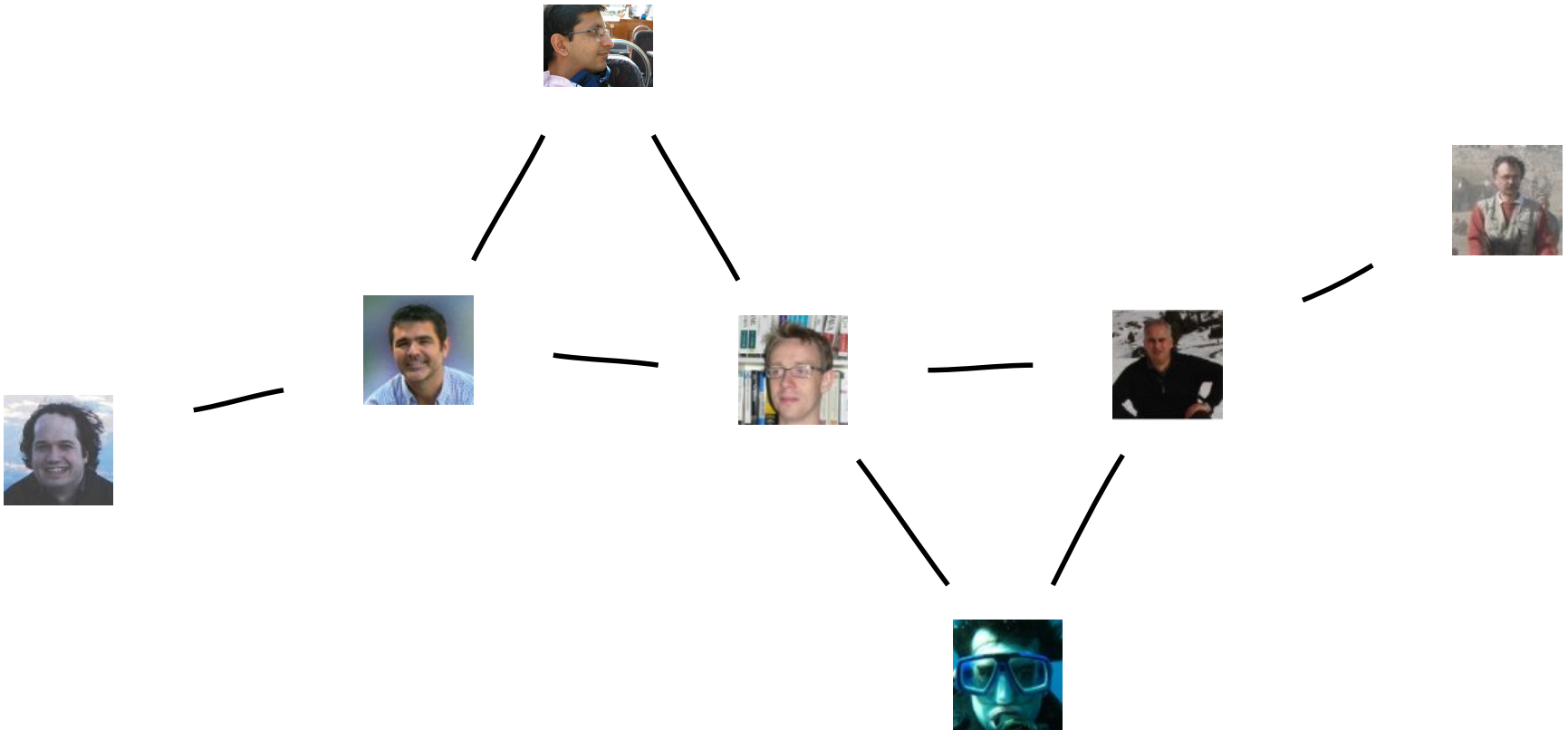
Social networking sites the most popular in the world and in France (Web site ranked with the most traffic, according to Alexa)

	Monde	France
SkyRock	51	3
YouTube	3	4
MySpace	17	7
Facebook	5	8
Dailymotion	61	11
EBay	18	12
Wikipedia	8	13
Meetic	565	27
ImageShack	47	53
hi5	15	59
Megavideo	133	80
Adult Friendfinder	55	82
Wat.tv	1568	88
Flickr	33	94
Orkut	19	>100
V Kontakte	28	>100
Friendster	39	>100

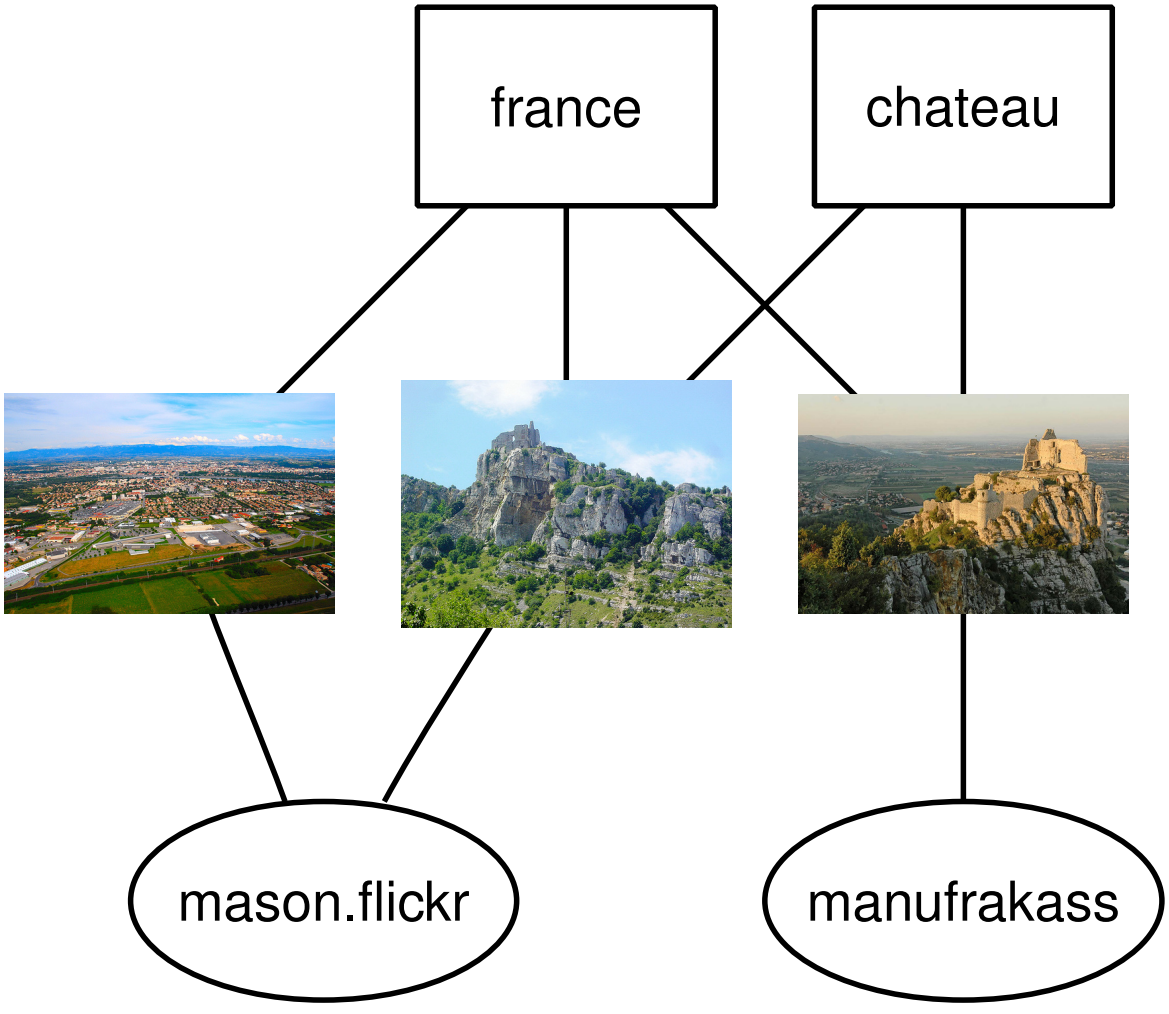


- Natural model: social network = graph
- Entities = nodes, Relations = edges
- Depending on the specific case:
  - mostly undirected graphs
  - bipartite, *n*-partite
  - annotated edges, weighted edges

Adapted for pure social networking sites with symmetrical relations (e.g. LinkedIn)



Adapted to most social networks for sharing content, with annotations, [illegible]ent, etc. (e.g., Flickr)



Idea that any two persons on Earth are separated by a chain of six persons such that any two successive links know each other

- Demonstrated by an experiment by Stanley Milgram [Travers and Milgram, 1969] (package to transmit to an unknown person, using acquaintances)
- Popularized in numerous media
- Number 6 is not too be taken too seriously! But main idea validated in more recent experiments
- In other fields:
  - Erdős number for scientific publications
  - Kevin Bacon for Hollywood movies

Common Characteristics of (most) social networks !

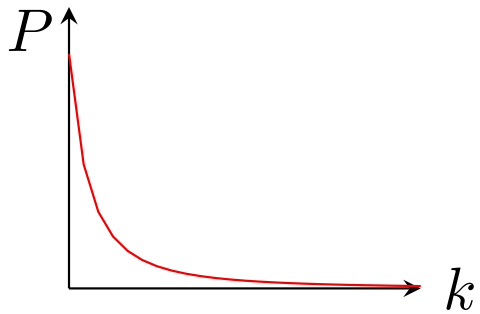
Four important characteristics [Newman et al., 2006]:

Sparse graphs: much more edges than a complete graph

Small typical distance: shortest path between two nodes logarithmic with respect to the size of the graph

High clustering: if *a* is connected to *b* and *b* to *c*, then *b* has good probability of being linked to *c*

Power-law degree distribution: the number of nodes with degree *k* of the order of $k^{-\gamma}$ (γ constant)



More on this in Athens course on Collective Intelligence (TPT09)

- Context: Multipartite graph, e.g., Flickr
- Purpose: bias results with one's social network
- Social weighting:
  - Given a friendship relation $F(u, u')$ (explicit or implicit) between two users, an extended friendship relation can be computed

$$\tilde{F}(u, u') = \frac{\alpha}{|U|} + (1 - \alpha) \max_{\text{chemin } u = u_0 \ldots u_k = u'} \prod_{i=0}^{k-1} F(u_i, u_{i+1})$$

    ($0 < \alpha < 1$ constant; $|U|$: number of users)
  - Instead of a global weighting

$$\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t, d)$$

    we choose a social weighting dependent of u:

$$\text{tf-idf}_u(t, d) = \left(\sum_{u' \in U} F(u, u') \cdot \text{tf}_{u'}(t, d) \right) \times \text{idf}(t, d)$$

- Possible to adapt Fagin's trheshold algorithm...
- ... but impossible to precompute the scores tf-idf$_u(t, d)$ for each user
- To avoid too much computation time:
  1. Cluster the graph of users into strongly similar components
  2. Use the scores inside these components as estimates of the threshold in Fagin's algorithm
  3. $\Rightarrow$ gives approximate results, but good quality

The Web as a graph

Social Networking Sites

Conclusion

What you should remember

- The Web seen as a graph
- PageRank, and its iterative computation
- Graph mining techniques applicable to a wide range of graphs
- Social networks share important common characteristics with the Web graph

To go further

- A good textbook [Chakrabarti, 2003]
- Two accessible influential research papers:
  - On HITS [Kleinberg, 1999]
  - On PageRank [Brin and Page, 1998]

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, 30(1–7): 107–117, April 1998.

Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer Networks*, 33(1-6): 309–320, 2000.

Soumen Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann, San Fransisco, USA, 2003.

Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, 1999.

Mark Newman, Albert-László Barabási, and Duncan J. Watts. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 34(4), December 1969.