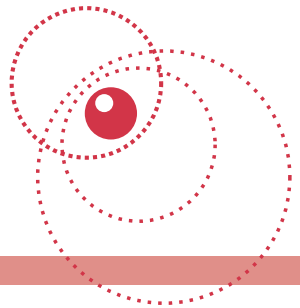


INF344: Données et algorithmes du Web

Web Ranking

19 May 2014

TELECOM ParisTech

Plan

Web search engines

- A large number of different search engines, with market shares varying a lot from country to country.
- At the world level:

  Google vastly dominating (around 80% of the market; more than 90% market share in France!)

  Yahoo!+Bing still resists to its main competitor (around 10% of the market)
- In some countries, local search engines dominate the market (Baidu and 360 Search in China, Naver in Korea, Yandex in Russia)
- Other search engines mostly either use one of these as backend (e.g., Google for AOL) or combine the results of existing search engines (e.g., DuckDuckGo, which also has a small Web index)

Yahoo!

In July 2009, Microsoft and Yahoo! announced a major agreement:

- Yahoo! stops developing its own search engine (launched in 2003, after the buyouts of Inktomi and Altavista) and will use Bing instead;
- Yahoo! will provide the advertisement services used in Bing.

Operational, but does not concern Yahoo! Japan, which on the contrary uses Google as engine.

Web search APIs

- Used to be plenty of free APIs to Web search engines. . . not the case any more
- Paid-for Web search APIs:

  Yahoo! BOSS 0.80 USD per 1,000 queries (uses Bing's index)
  `https://developer.yahoo.com/boss/search/`

  Google Custom Search Engine 100 free queries per day, 5 USD per further 1,000 queries, up to 10,000 queries per day
  `https://developers.google.com/custom-search/`

  Bing Search API free for 5,000 queries per month; $\approx$ 20 USD per further 5,000 queries
  `https://datamarket.azure.com/dataset/5BA839F1-12CE-4CCE-BF57-A49D98D29A44`

- Anything else?

Plan

PageRank (Google's Ranking [Brin and Page, 1998])

Idea

Important pages are pages pointed to by important pages.

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

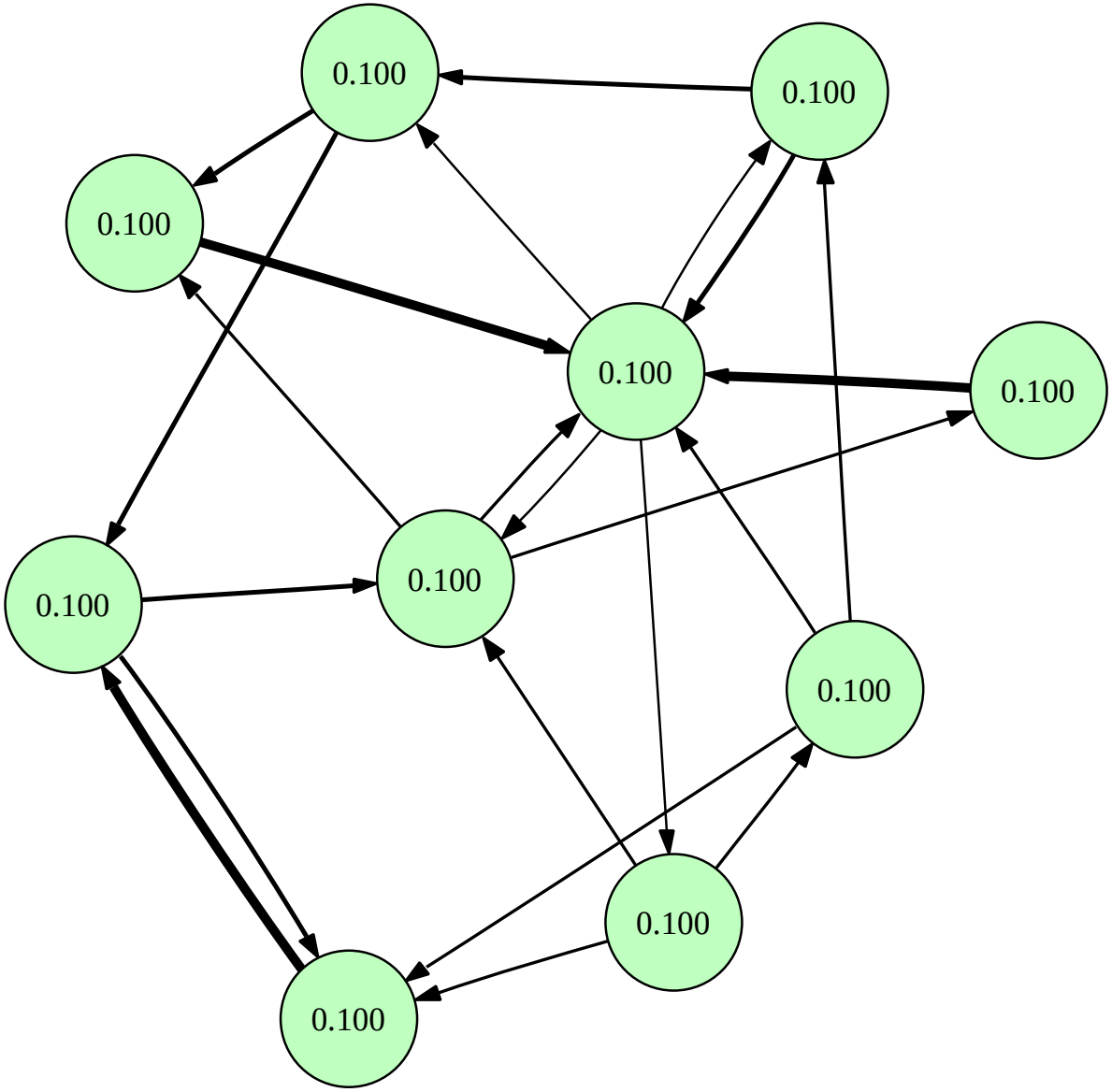
Definition (Tentative)

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

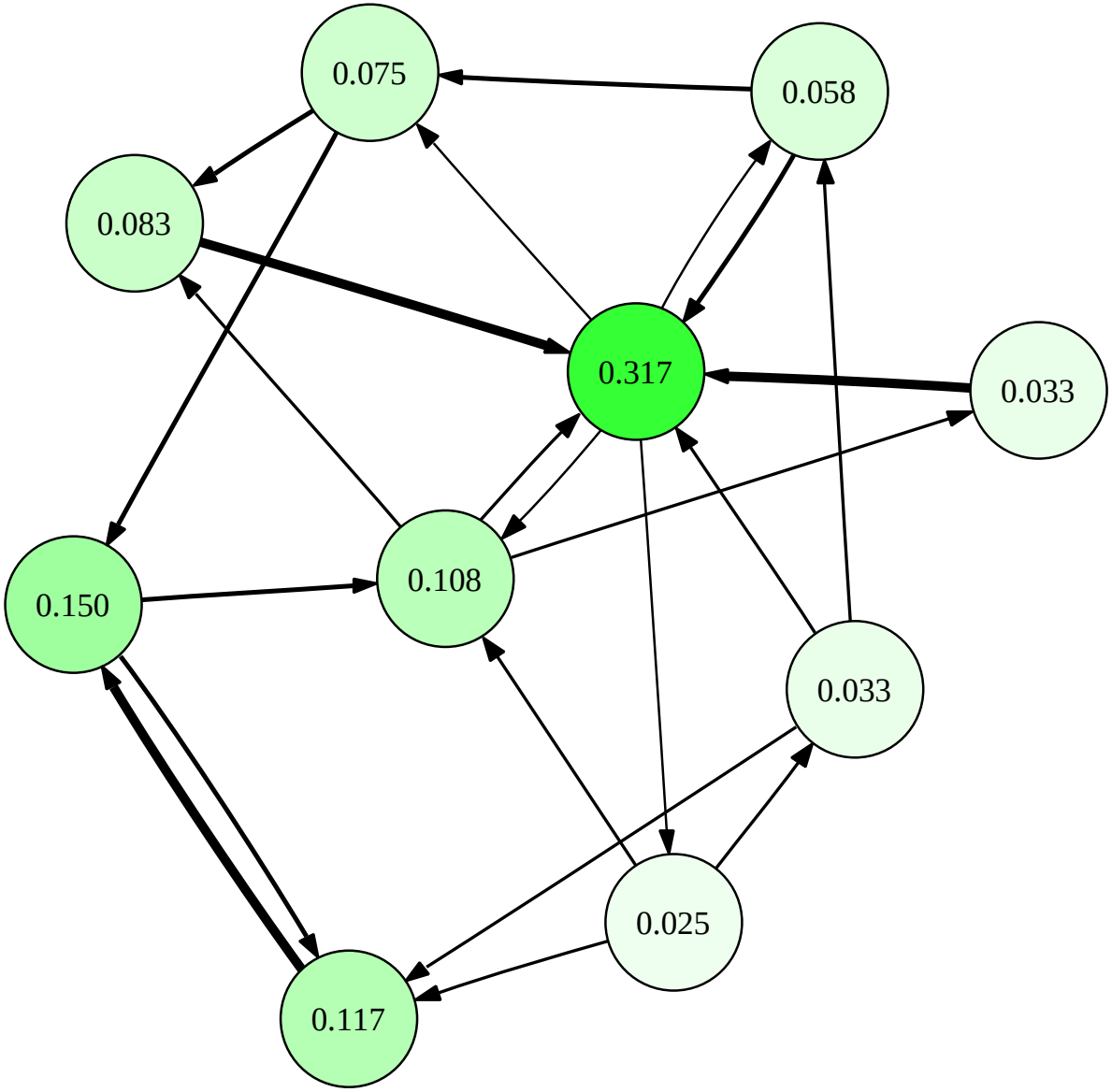
$$\mathrm{pr}(i) = \left(\lim_{k \to +\infty} (G^T)^k v \right)_i$$

where v is some initial column vector.

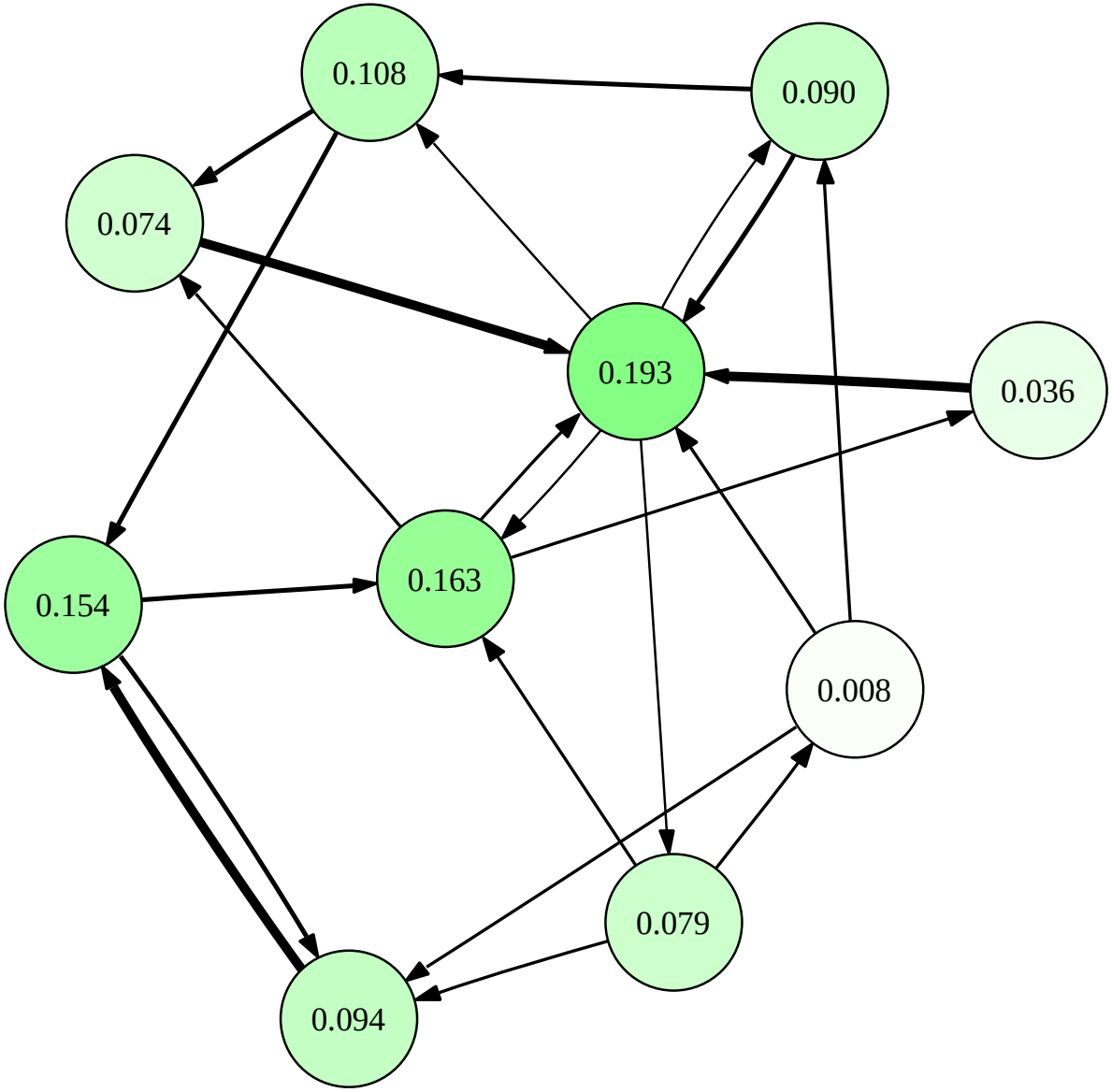
Illustrating PageRank Computation



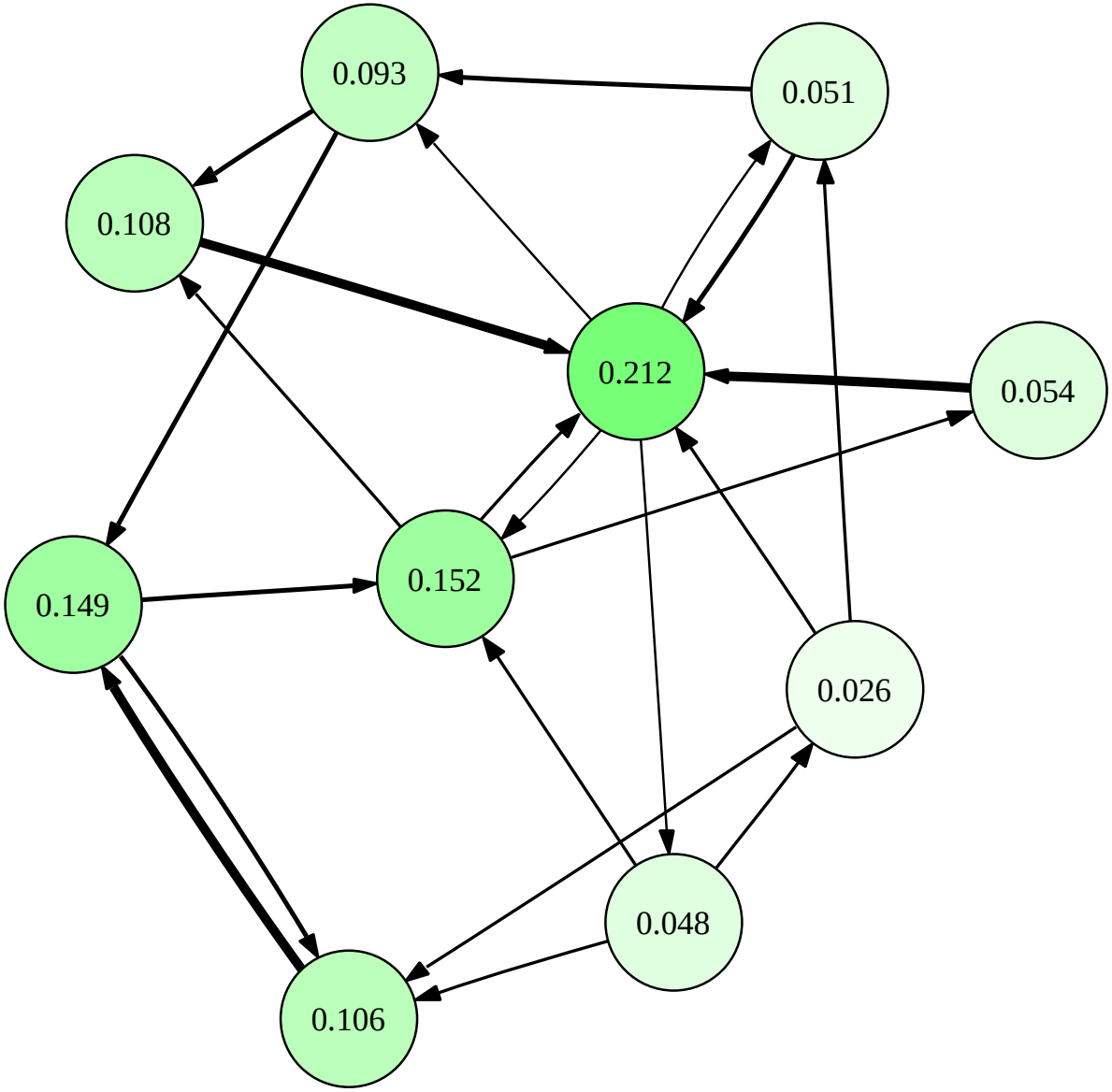
Illustrating PageRank Computation



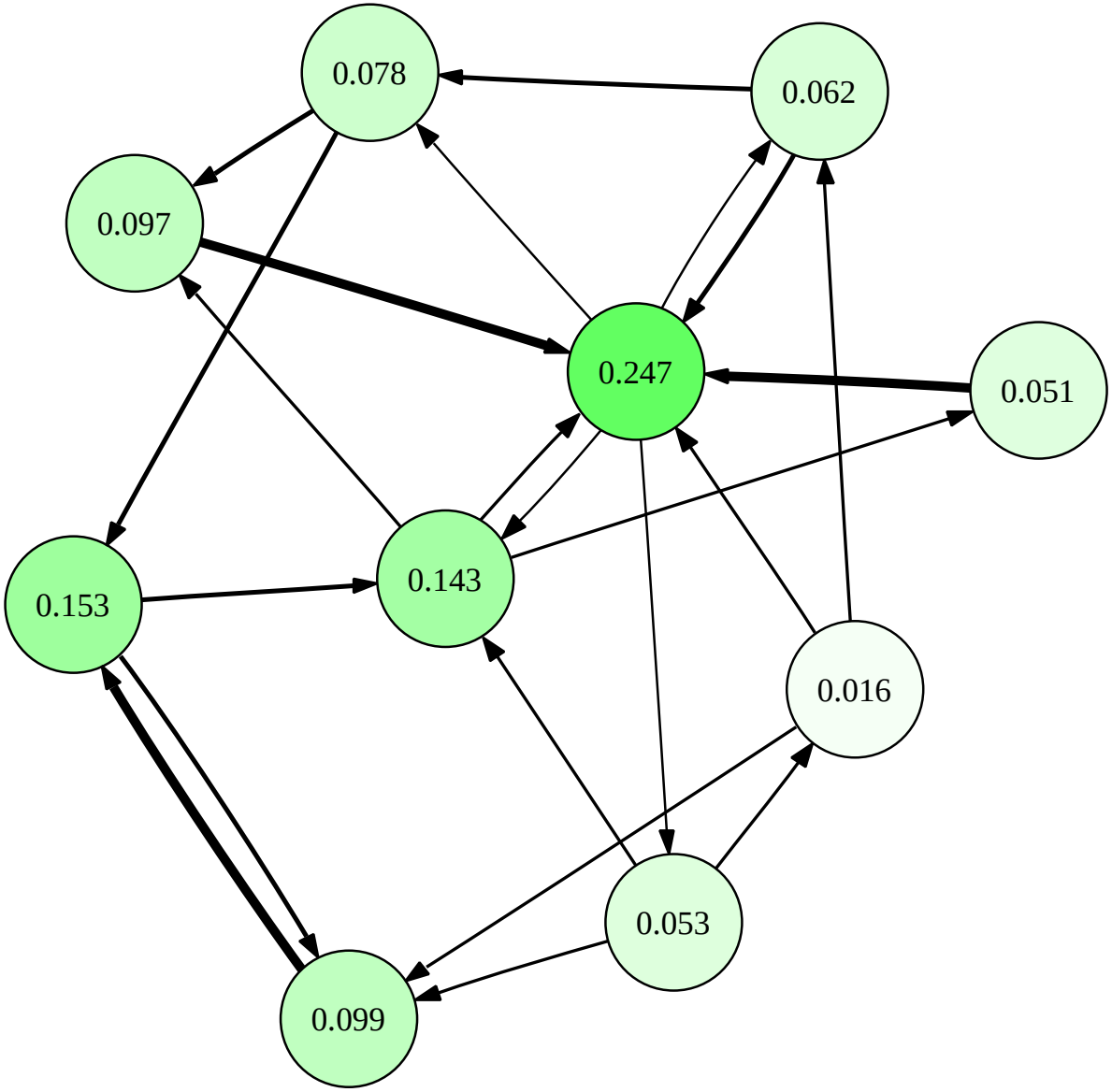
Illustrating PageRank Computation



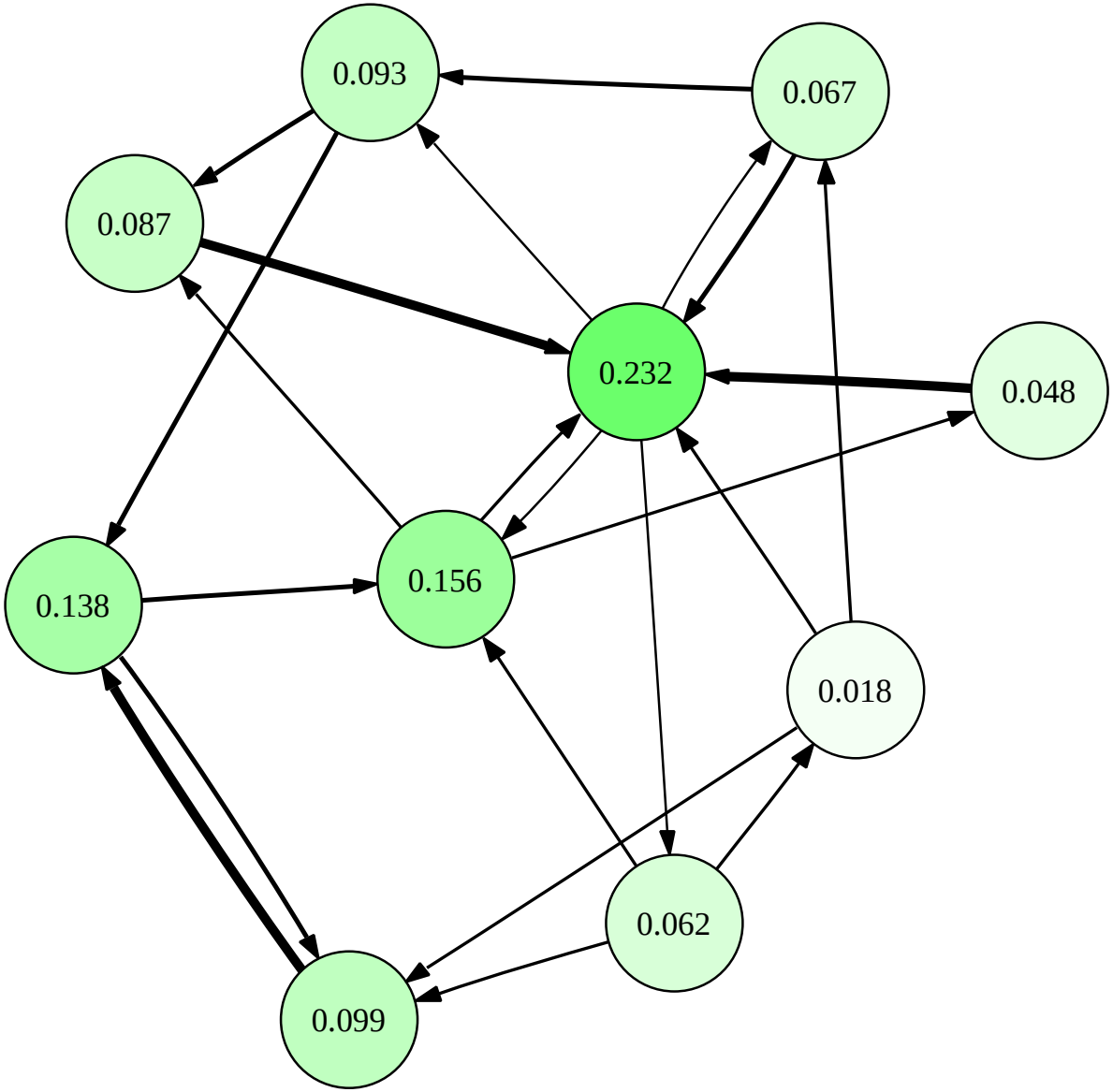
Illustrating PageRank Computation



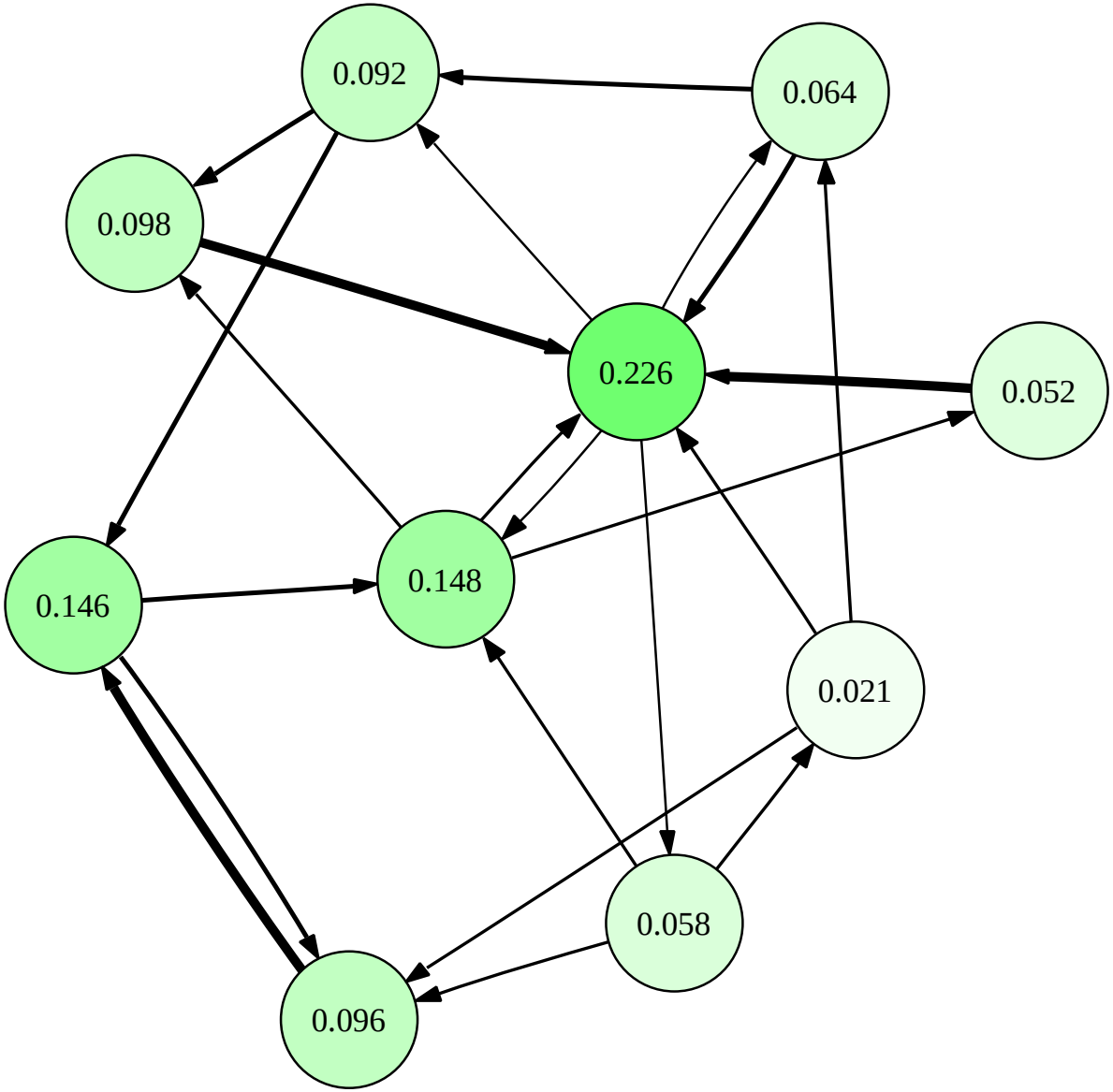
Illustrating PageRank Computation



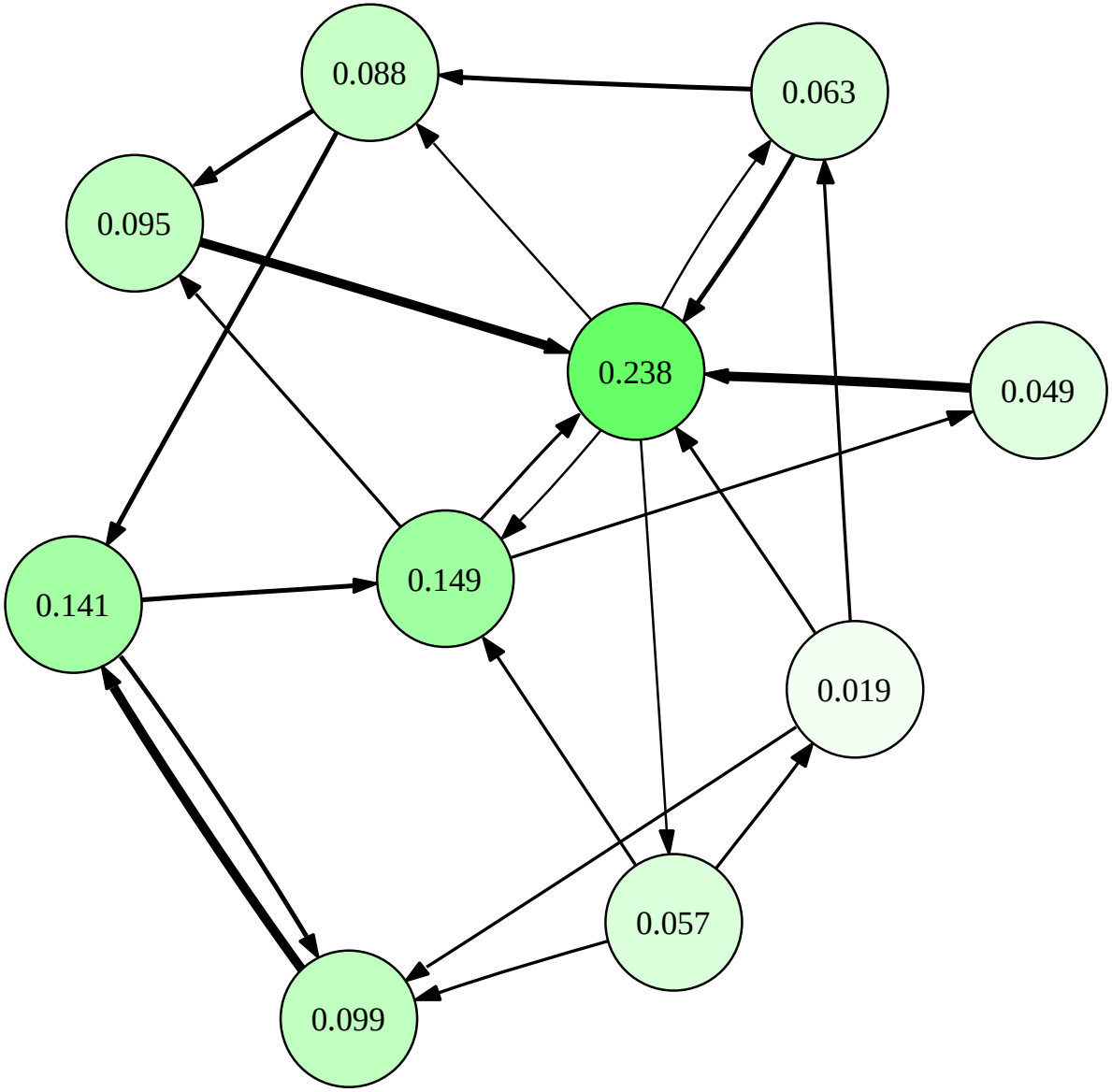
Illustrating PageRank Computation



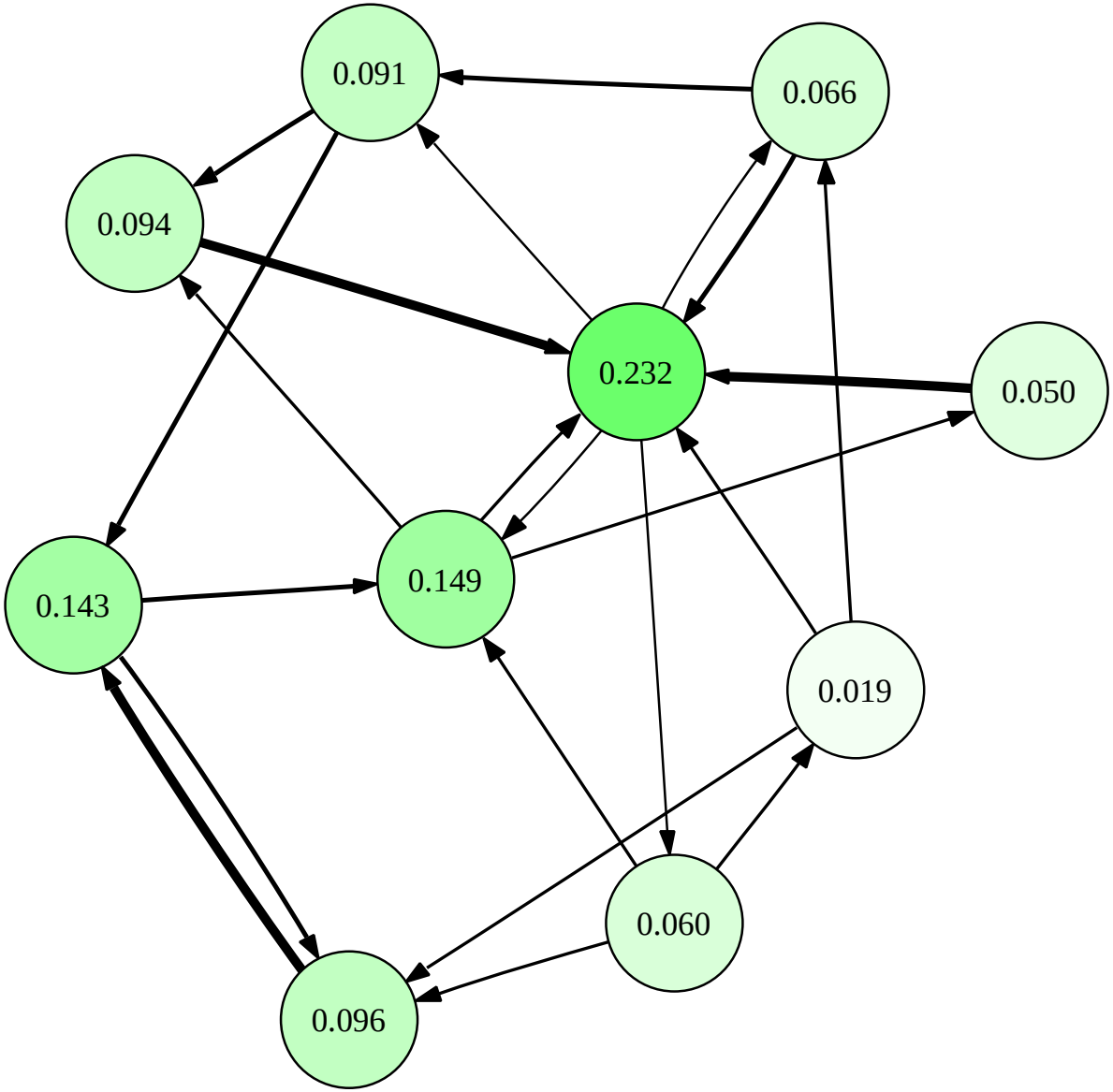
Illustrating PageRank Computation



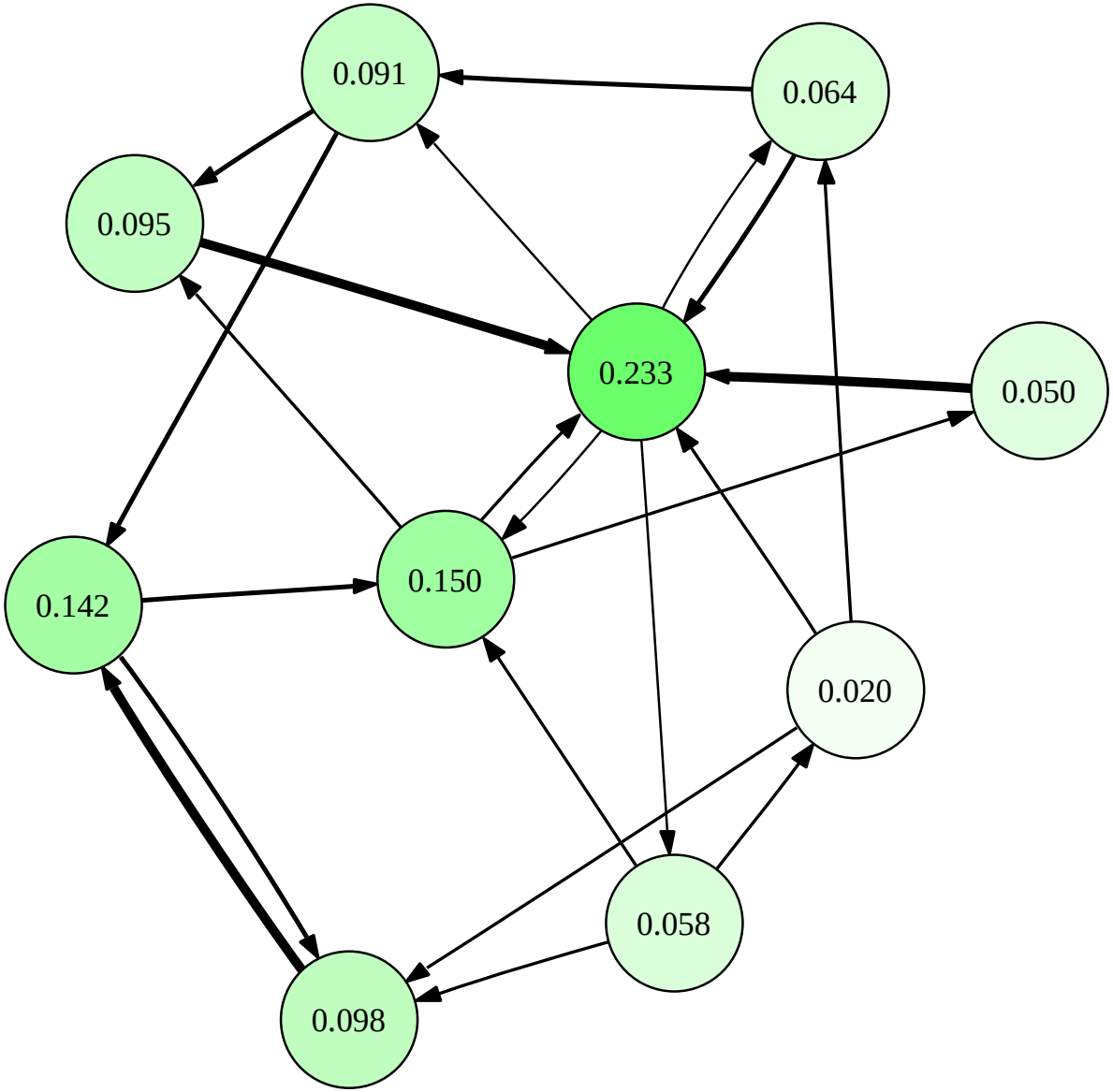
Illustrating PageRank Computation



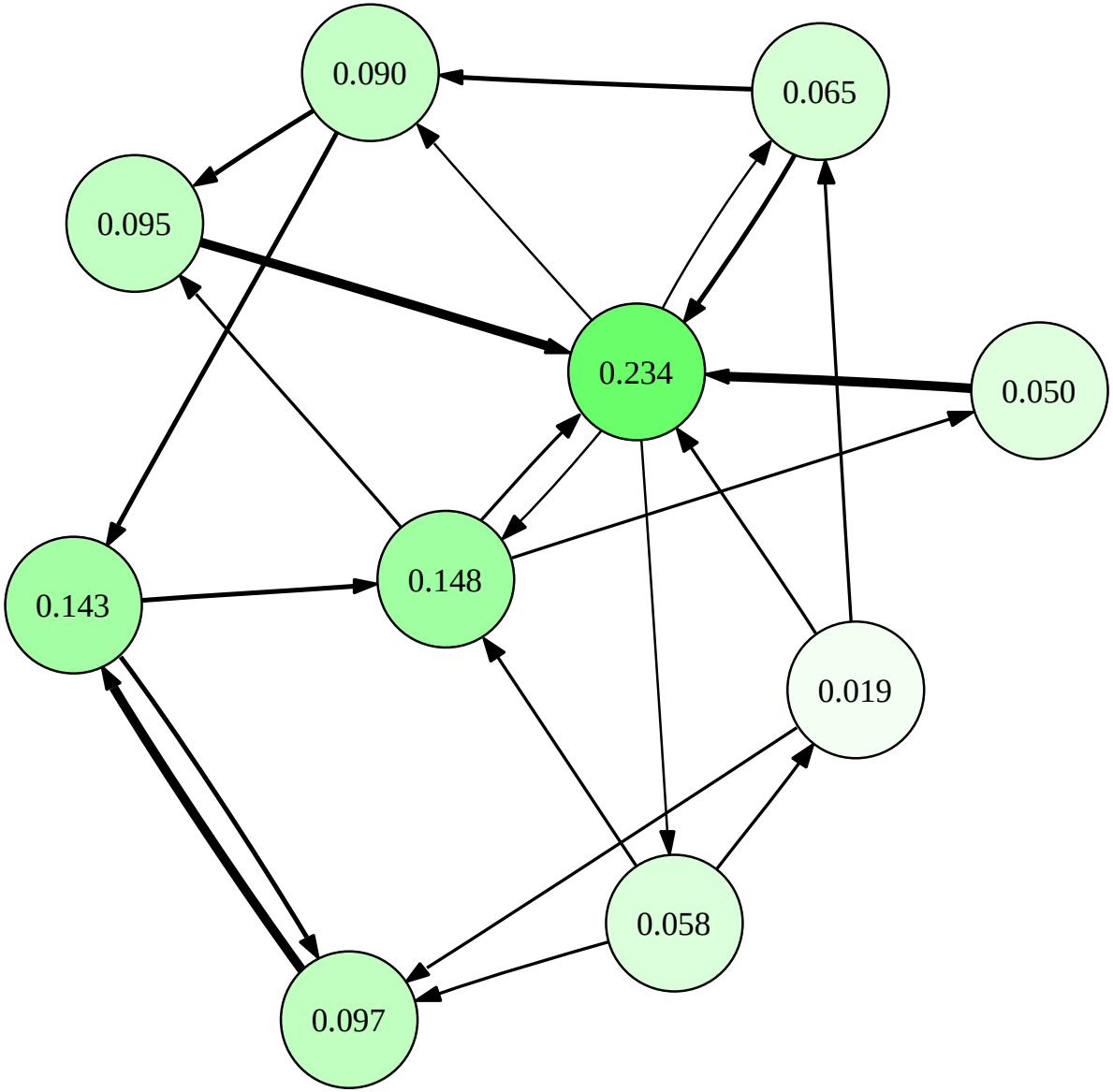
Illustrating PageRank Computation



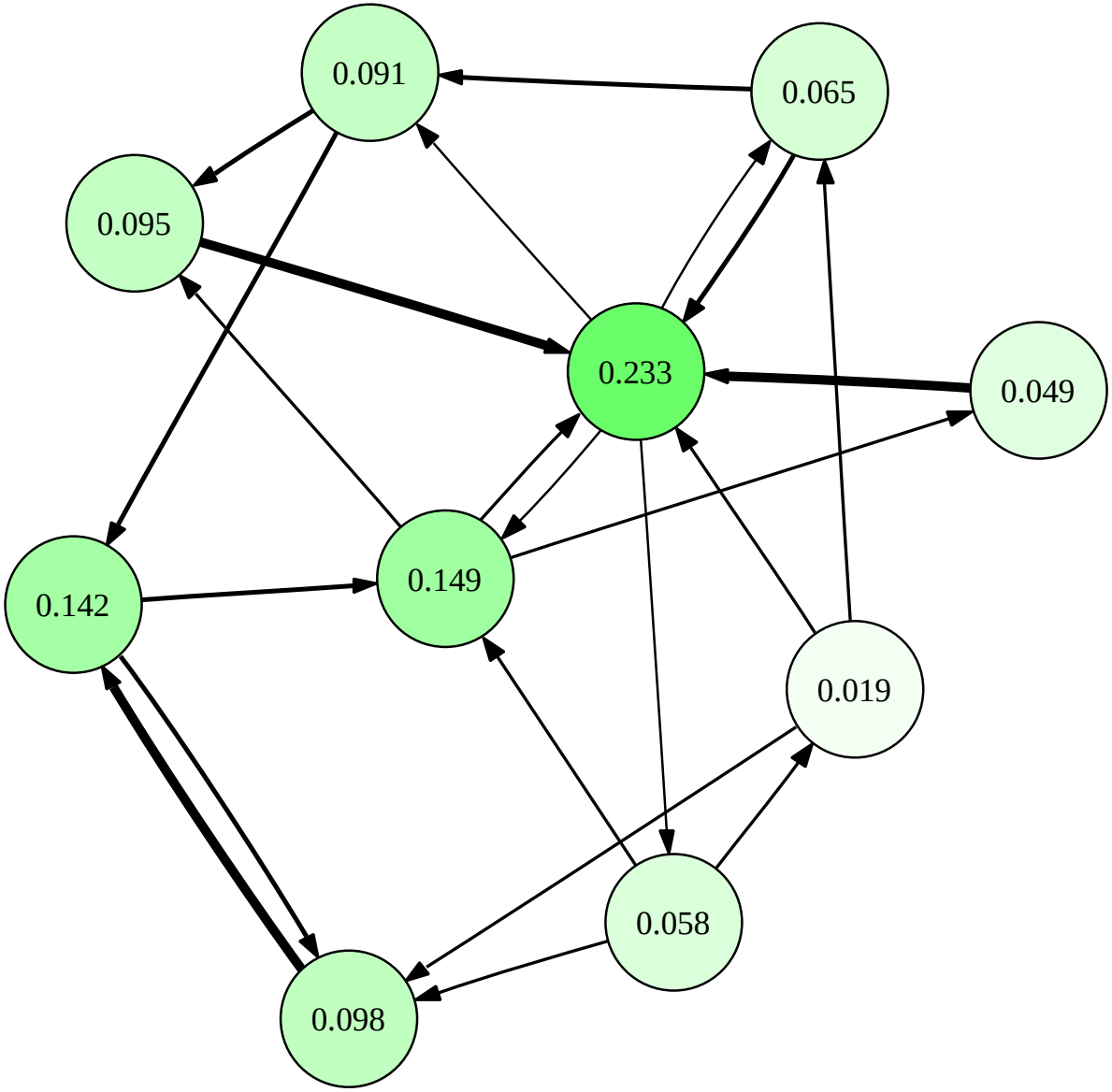
Illustrating PageRank Computation



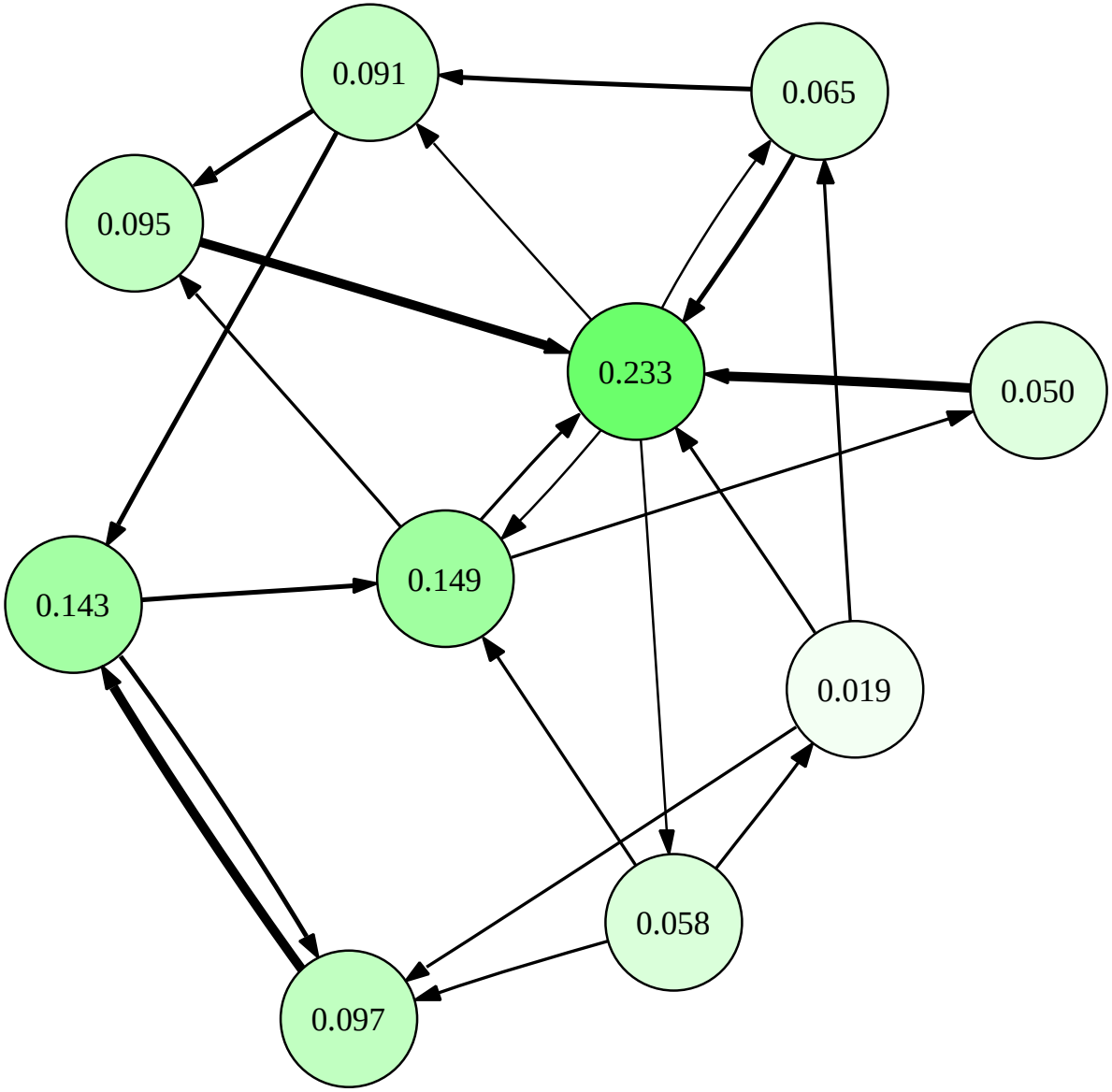
Illustrating PageRank Computation



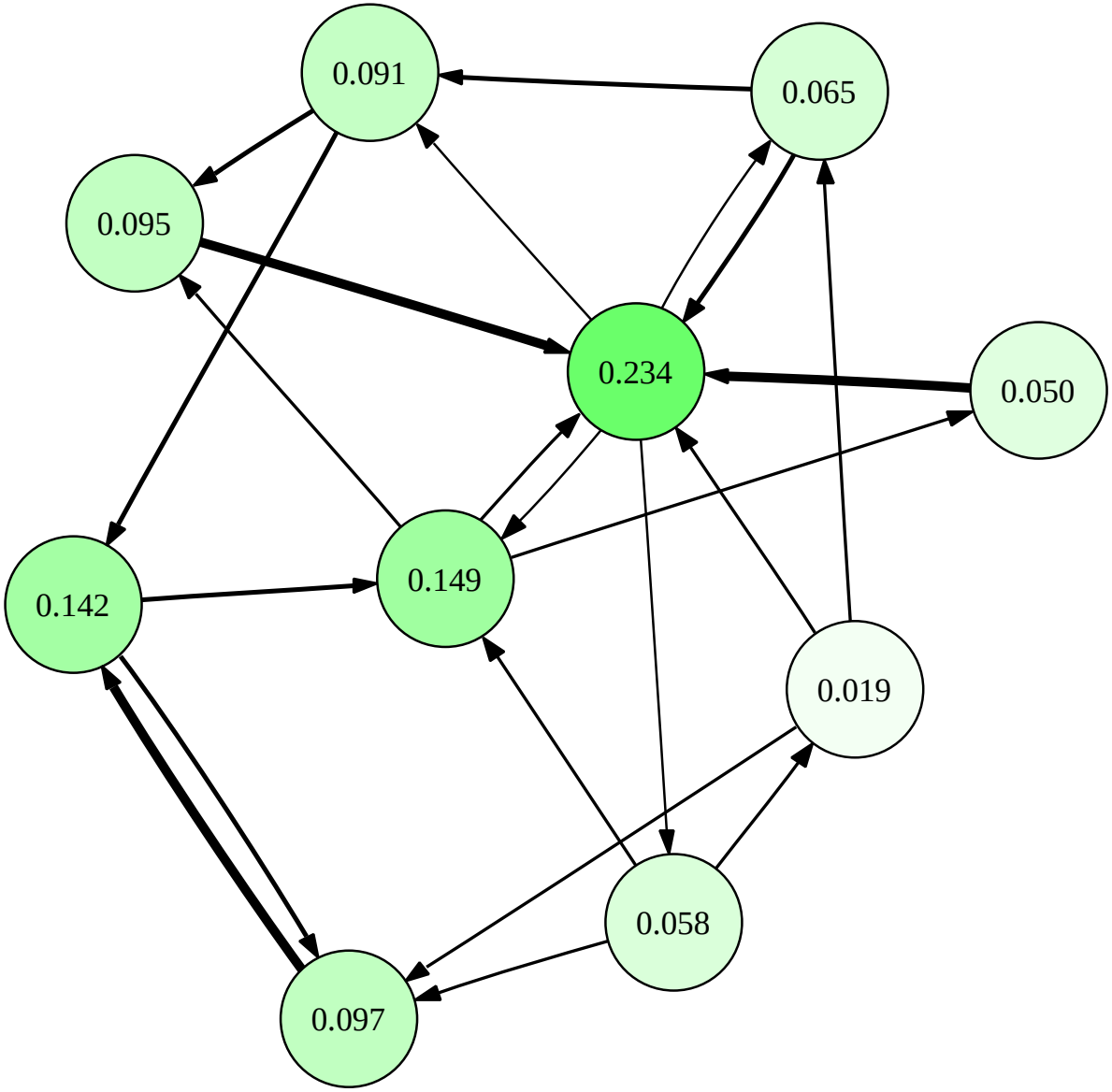
Illustrating PageRank Computation



Illustrating PageRank Computation



Illustrating PageRank Computation



PageRank with Damping

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.

Using PageRank to Score Query Results

- 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.

Plan

Ranking formula

- In modern search engines, Web query results are not just a combination of query relevance and PageRank (but these are most important)
- Instead, complex combination of dozens of components
- Can be integrated into the inverted index, or added on the fly when computing query results
- Simple way of combining: linear (or log-linear) combination of individual weights, with weights chosen in an ad-hoc manner, or, better, trained with machine learning
- Thereafter: collection of such components

Traditional IR

Relevance weighting: tf-idf, OKAPI BM25, etc.

Position-aware scoring: rank higher terms that appear closer to each other

Metadata scoring: Use information from metadata (title, keywords, etc.); Not much used any more, too much abuse

Query rewriting and spell checking: Compare to query logs or the index to issue a similar, more popular, query

Diversification: Give results as diversified as possible

Web graph mining

PageRank: important pages are pointed to by important pages

SiteRank: important sites are pointed to by important sites

TrustRank: to fight spam, assign initial trust to a seed of Web pages, and increase the score of neighboring pages

Link farm detection: lower score of subgraphs with dubious structures

Relevance feedback

Other user's feedback: use previous clicks of other users as positive examples this link is relevant (or absence of click as negative examples)

Own feedback: use user's history to rank higher previously visited pages

Manually crafted: for common, important search terms, manually design the search result pages!

Exploiting content and layout

Structural emphasis: raise score of section titles, emphasized words, etc.

Layout emphasis: render the page in a layout engine, and raise score of visually prominent items

Invisible content detection: static analysis of CSS/JS code, or layout rendering, to detect and penalize invisible content

Quality factors

Standard-compliant: raise score of valid HTML pages

Speed of access: decrease score of slow servers

Visual appearance: decrease score of gaudy-looking pages

Up-to-date character: increase score of recently modified pages

Domain names: increase score of reputable domain names vs dubious-looking, lengthy, ones

URL structure: decrease score of lengthy or convoluted URLs

User-centric factors

Social search: Bias by a users' social network (if the user is logged in)

Location search: Bias by a users' precise location (if available) or IP geolocation

Language-specific search: Bias by a user's language preferences (as reported by the browser, or as manually chosen)

History-aware search: Bias by a user's search history

Plan

Plan

Spamdexing

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

Spamdexing: Lying about the Content

Technique

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

Link Farm Attacks

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

Link Pollution

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

Plan

Optimiser le référencement d'un site Web (1/2)

- Faire attention à l'accessibilité aux robots des moteurs de recherche (Flash, JavaScript, etc.)
- Éventuellement prévoir des versions HTML pures alternatives
- Site accessible à une URL unique (par exemple, entre `http://www.toto.com/` et `http://toto.com/`, si les deux permettent d'accéder au site, l'un doit être une redirection (HTTP) vers l'autre)
- Structure cohérente (hébergement sur un seul serveur, contenu des URLs pertinentes), bonne structure de liens (de n'importe quelle page, il doit être possible d'atteindre la page principale en un clic, et n'importe quelle autre page en quelques clics)
- Liens vers les autres sites pertinents, que le site n'apparaisse pas complètement isolé

Optimiser le référencement d'un site Web (2/2)

- Pas de tentatives de spamdexing: pas de texte invisible, pas d'abus dans les mots-clefs de la balise `<meta>` (assez inutiles de toute façon), etc.
- Faire apparaître des liens vers le site (surtout vers la page principale, et éventuellement pages internes quand c'est pertinent) sur d'autres sites Web (référencement dans des annuaires, sur les pages Web des individus/sociétés en lien avec le contenu du site, etc.).
- Éventuellement s'il est difficile de faire ainsi, soumettre le site aux différents moteurs de recherche (p. ex., `http://www.google.com/addurl/` pour Google).
- Se méfier des sociétés proposant un meilleur référencement contre rémunération; leurs pratiques sont mal vues par les moteurs de recherche.
- Et enfin. . . avoir du contenu intéressant !

Plan

Conclusion

What you should remember

- A very concentrated market; hard to get into it, vast resources required
- Complex formula for ranking Web query results
- Avoid any attempt at spamdexing!

Bibliography I

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

Licence de droits d'usage

Contexte public } avec modifications

Par le téléchargement ou la consultation de ce document, l'utilisateur accepte la licence d'utilisation qui y est attachée, telle que détaillée dans les dispositions suivantes, et s'engage à la respecter intégralement.

La licence confère à l'utilisateur un droit d'usage sur le document consulté ou téléchargé, totalement ou en partie, dans les conditions définies ci-après et à l'exclusion expresse de toute utilisation commerciale.
Le droit d'usage défini par la licence autorise un usage à destination de tout public qui comprend :
– le droit de reproduire tout ou partie du document sur support informatique ou papier,
– le droit de diffuser tout ou partie du document au public sur support papier ou informatique, y compris par la mise à la disposition du public sur un réseau numérique,
– le droit de modifier la forme ou la présentation du document,
– le droit d'intégrer tout ou partie du document dans un document composite et de le diffuser dans ce nouveau document, à condition que :
  – L'auteur soit informé.

Les mentions relatives à la source du document et/ou à son auteur doivent être conservées dans leur intégralité.

Le droit d'usage défini par la licence est personnel et non exclusif.
Tout autre usage que ceux prévus par la licence est soumis à autorisation préalable et expresse de l'auteur : sitepedago@telecom-paristech.fr