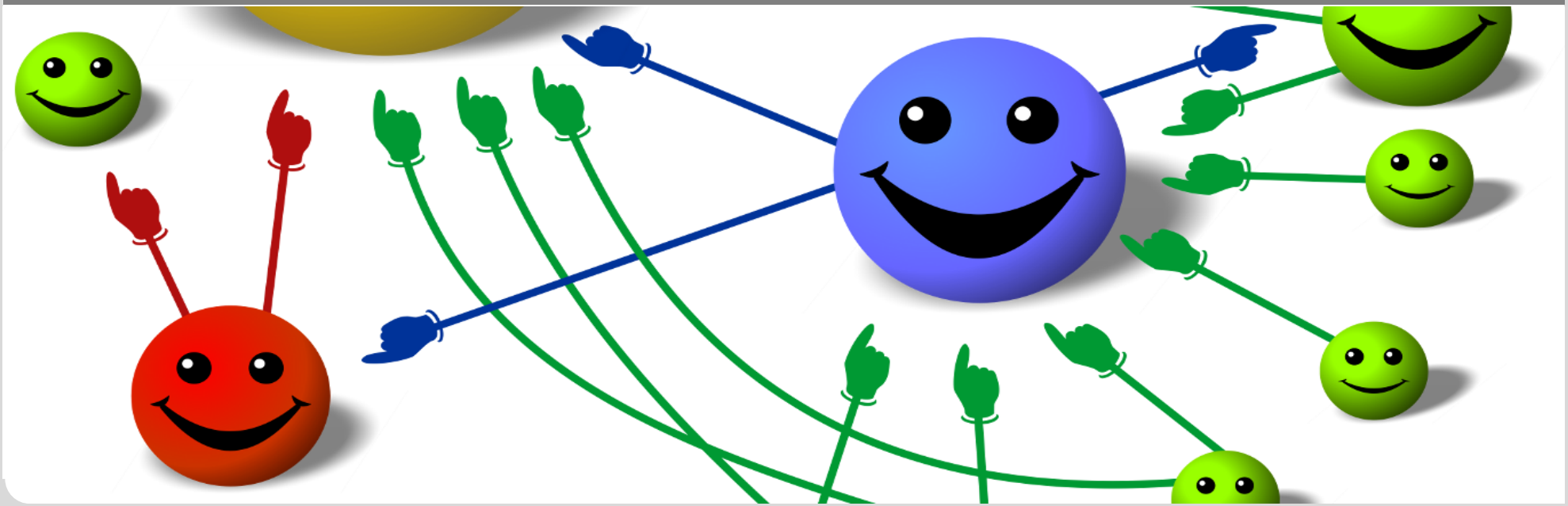
Karlsruhe Institute of Technology

How Google Searches Work

History of searching and the PageRank algorithm

Martin Thoma, Benjamin Lipp | 7th of February, 2013

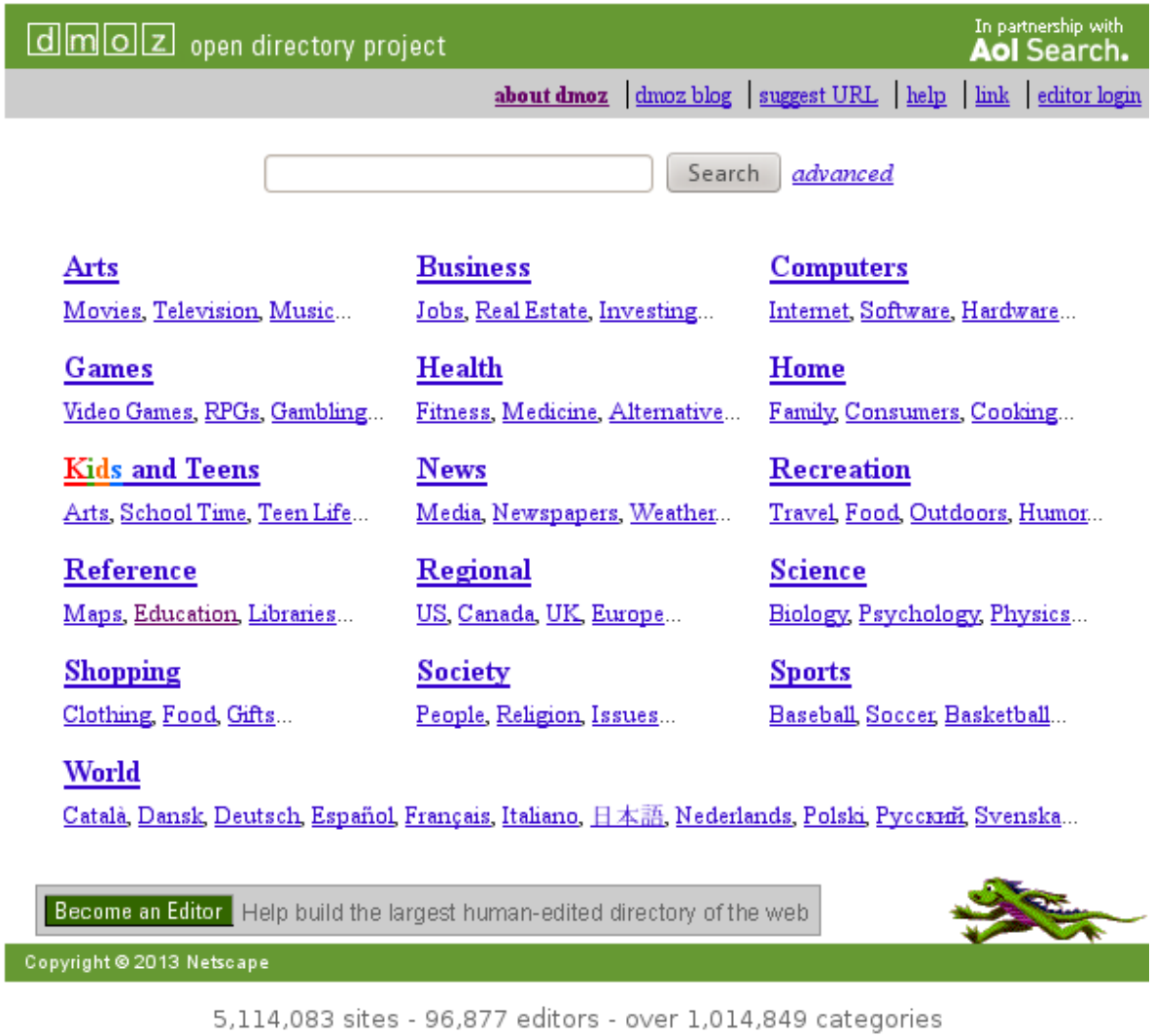
SPRACHENZENTRUM



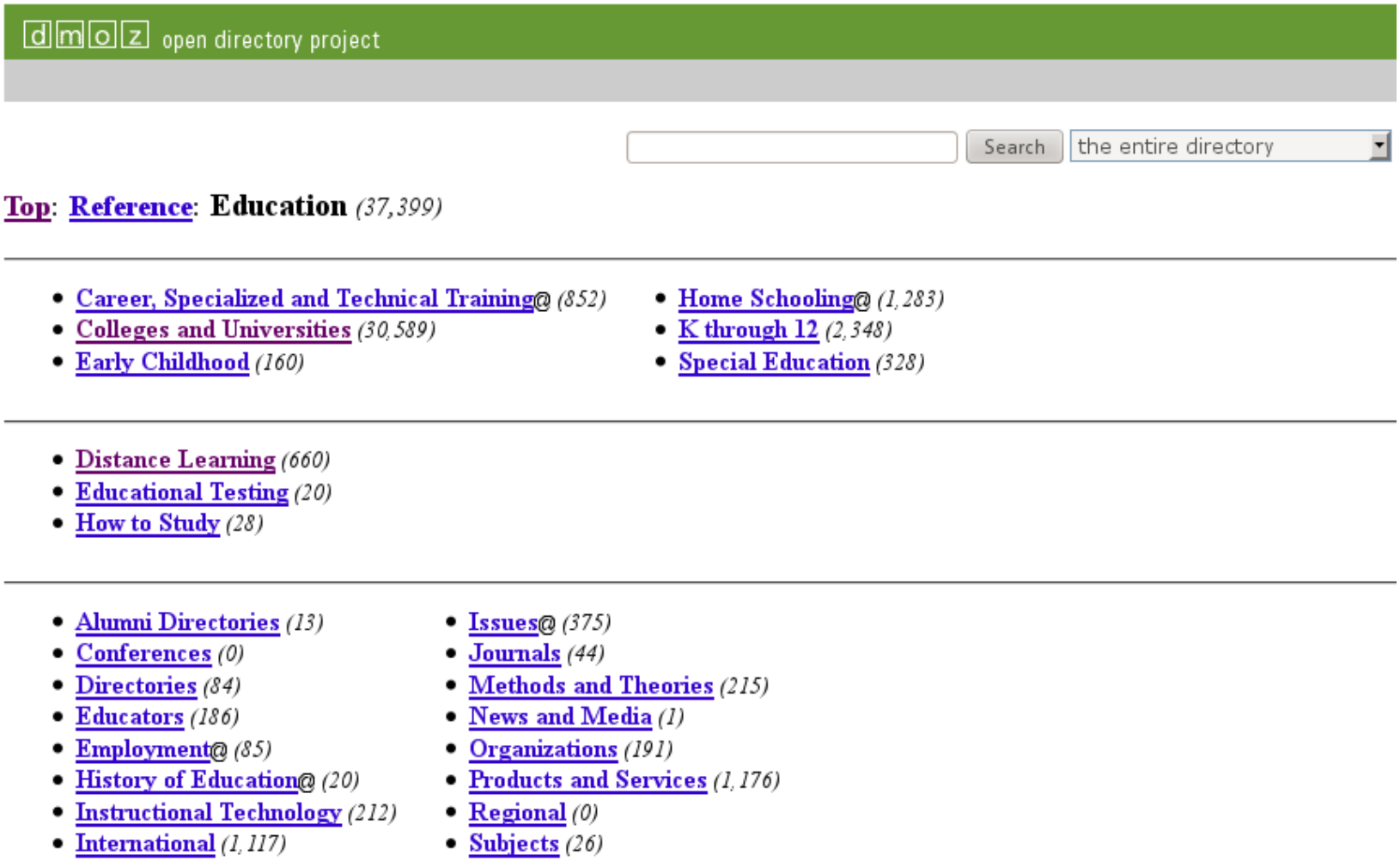
Contents

1. Introduction
2. PageRank
3. End

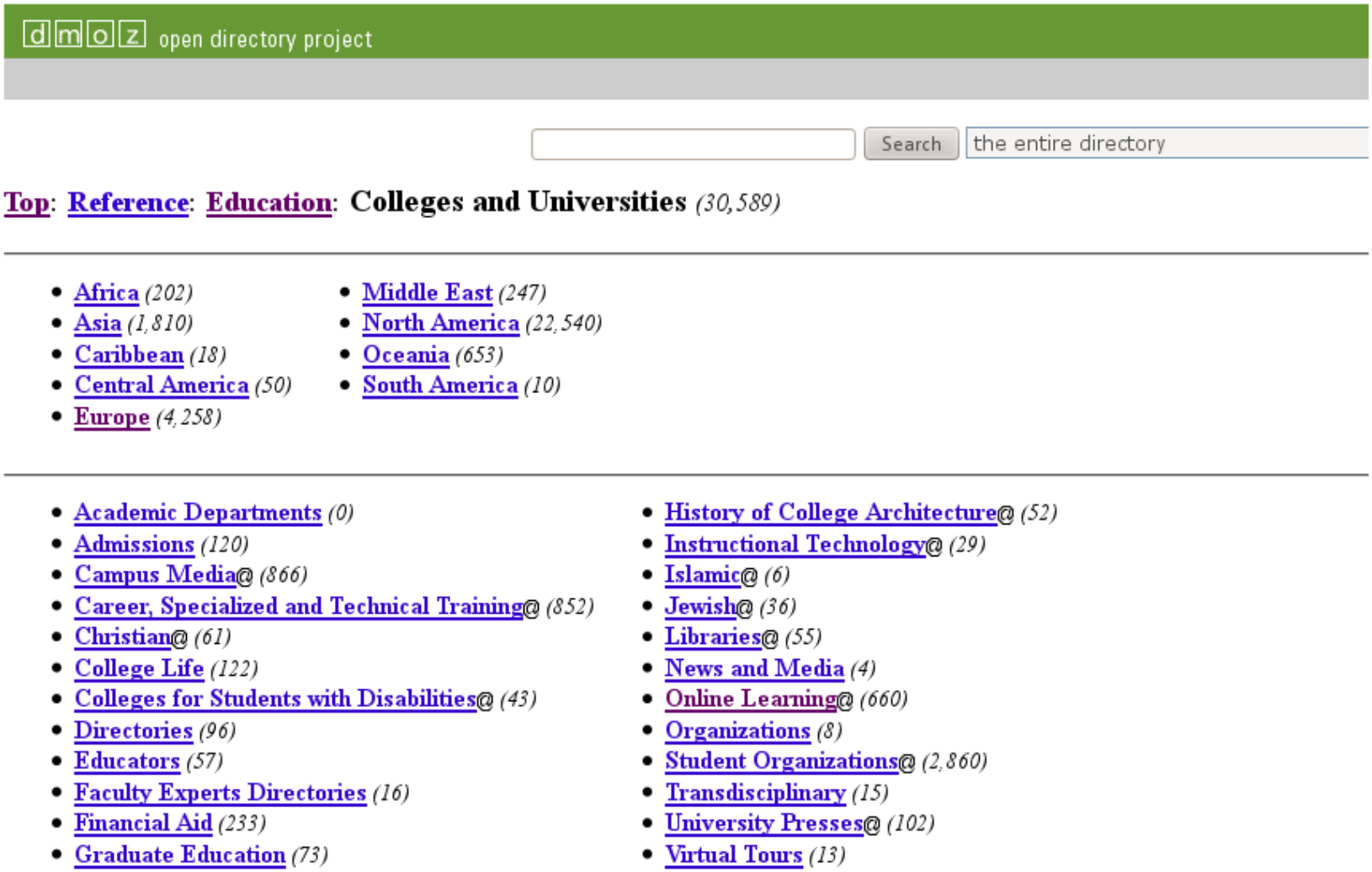
The Early Days: Web Directories



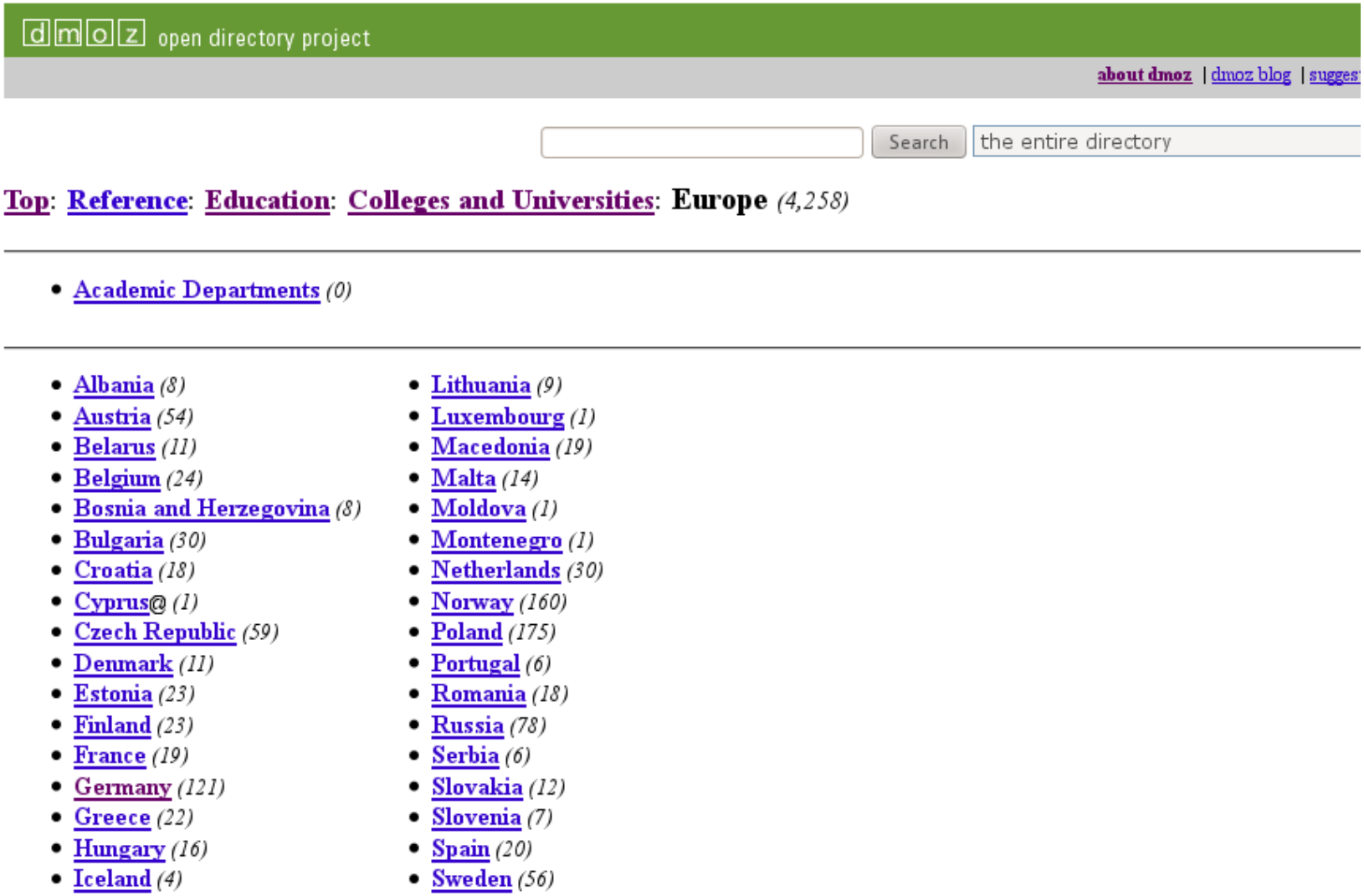
The Early Days: Web Directories



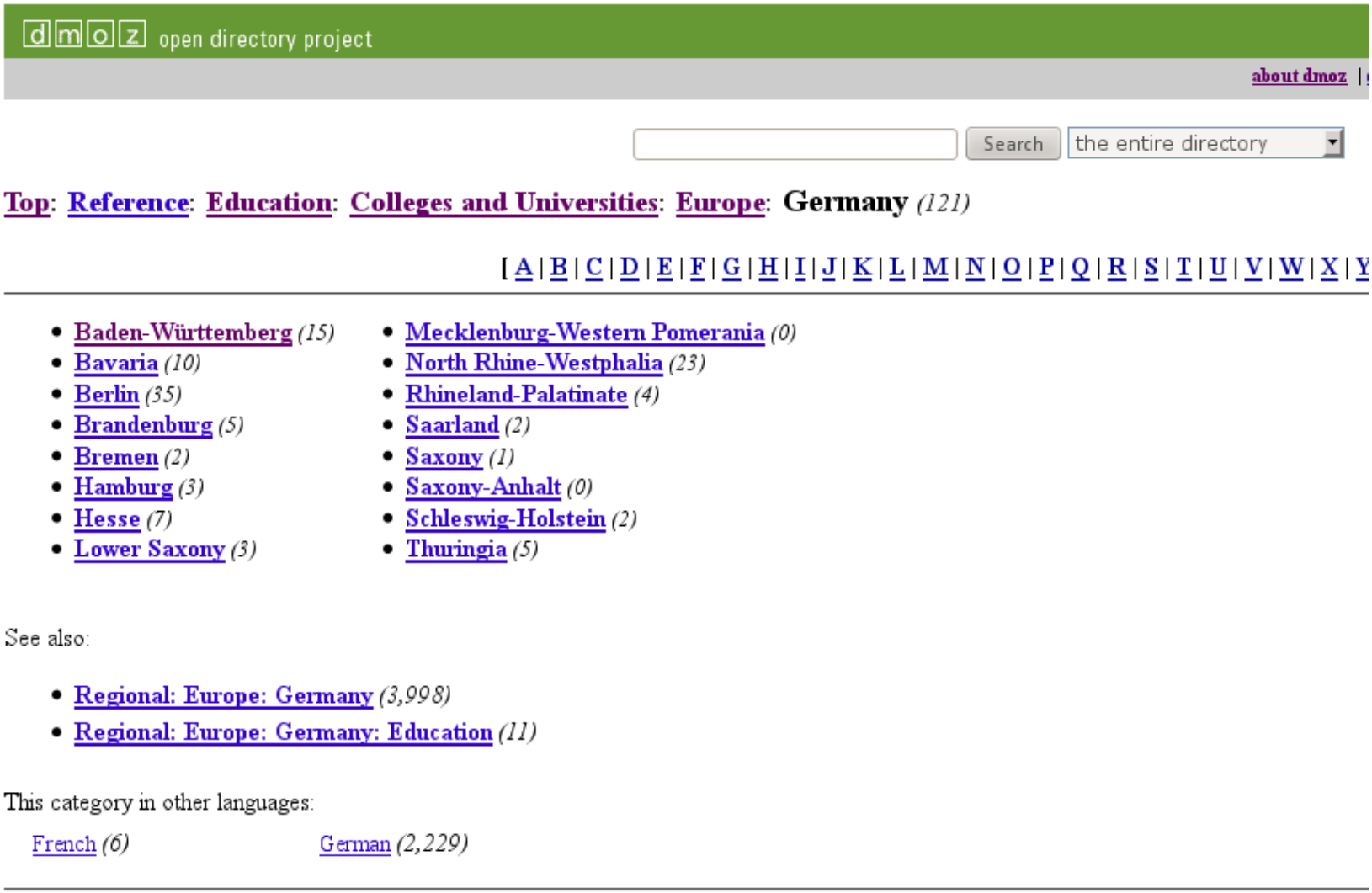
The Early Days: Web Directories



The Early Days: Web Directories



The Early Days: Web Directories



dmoz open directory project

about dmoz

Search | the entire directory

Top: Reference: Education: Colleges and Universities: Europe: Germany (121)

[A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y

- Baden-Württemberg (15)
- Bavaria (10)
- Berlin (35)
- Brandenburg (5)
- Bremen (2)
- Hamburg (3)
- Hesse (7)
- Lower Saxony (3)
- Mecklenburg-Western Pomerania (0)
- North Rhine-Westphalia (23)
- Rhineland-Palatinate (4)
- Saarland (2)
- Saxony (1)
- Saxony-Anhalt (0)
- Schleswig-Holstein (2)
- Thuringia (5)

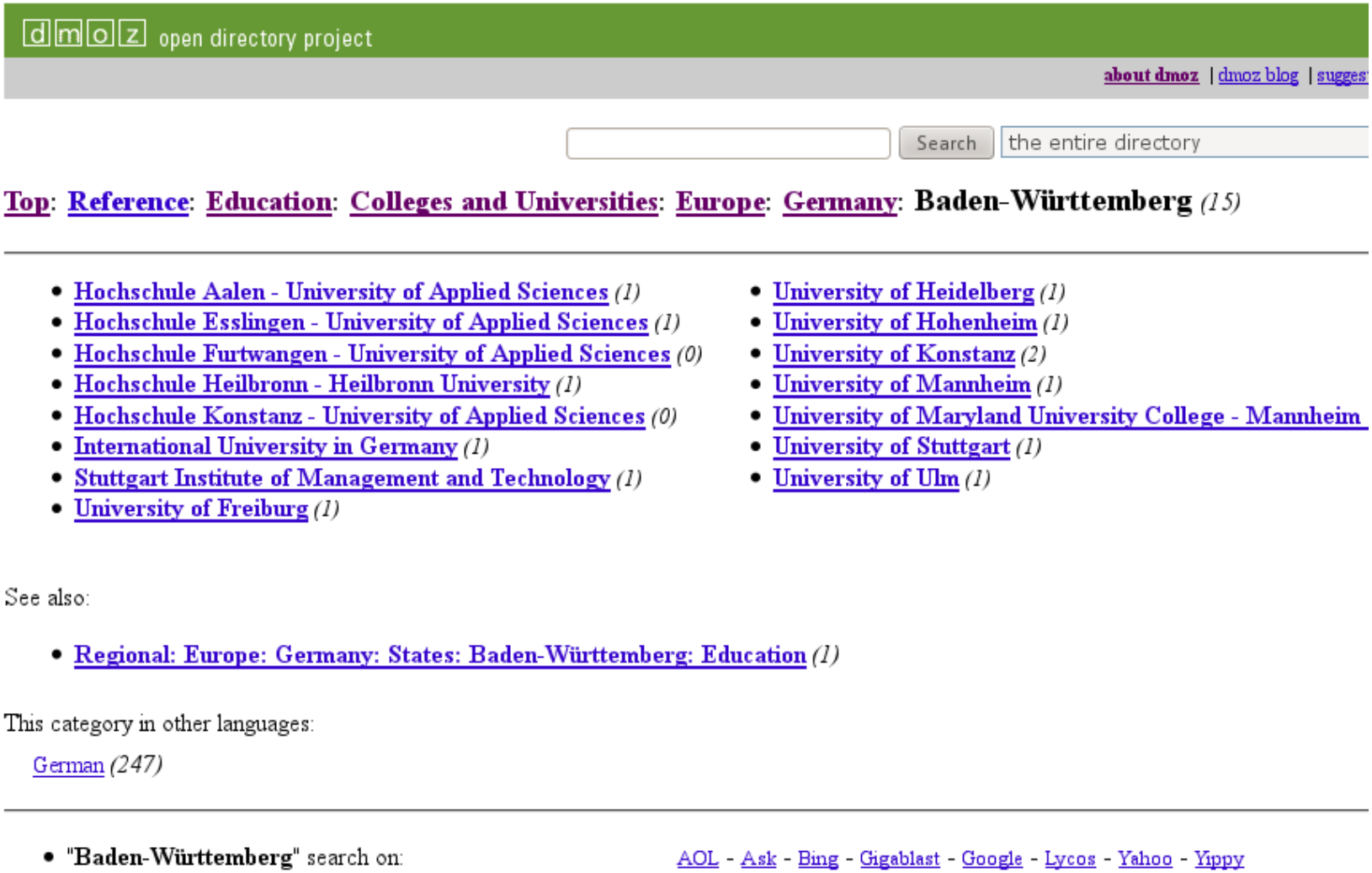
See also:

- Regional: Europe: Germany (3,998)
- Regional: Europe: Germany: Education (11)

This category in other languages:

French (6) German (2,229)

The Early Days: Web Directories



dmoz open directory project

about dmoz | dmoz blog | sugges

Search | the entire directory

Top: Reference: Education: Colleges and Universities: Europe: Germany: **Baden-Württemberg** *(15)*

- Hochschule Aalen - University of Applied Sciences *(1)*
- Hochschule Esslingen - University of Applied Sciences *(1)*
- Hochschule Furtwangen - University of Applied Sciences *(0)*
- Hochschule Heilbronn - Heilbronn University *(1)*
- Hochschule Konstanz - University of Applied Sciences *(0)*
- International University in Germany *(1)*
- Stuttgart Institute of Management and Technology *(1)*
- University of Freiburg *(1)*
- University of Heidelberg *(1)*
- University of Hohenheim *(1)*
- University of Konstanz *(2)*
- University of Mannheim *(1)*
- University of Maryland University College - Mannheim
- University of Stuttgart *(1)*
- University of Ulm *(1)*

See also:

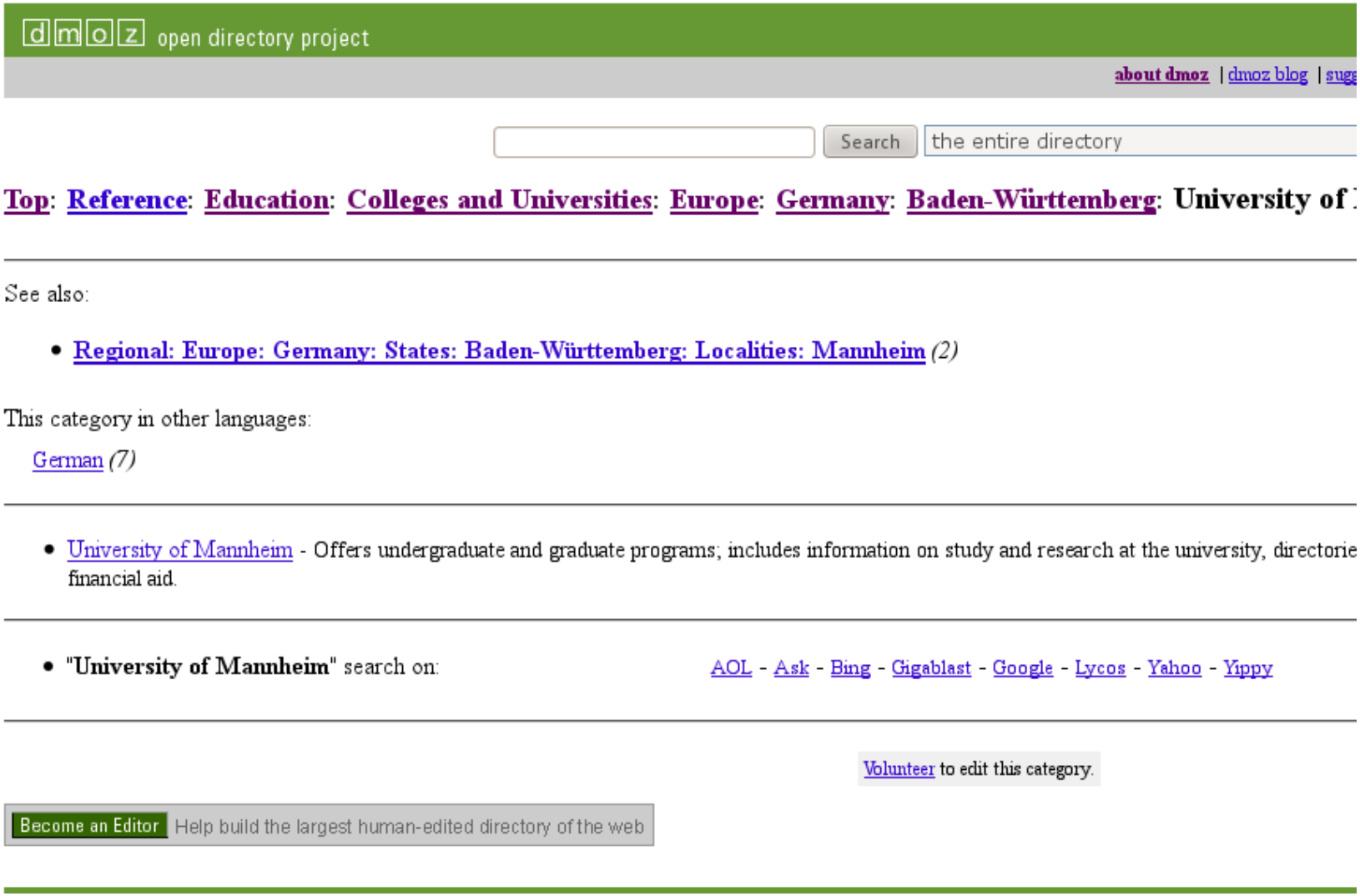
- Regional: Europe: Germany: States: Baden-Württemberg: Education *(1)*

This category in other languages:

German *(247)*

- "**Baden-Württemberg**" search on: AOL - Ask - Bing - Gigablast - Google - Lycos - Yahoo - Yippy

The Early Days: Web Directories



The Early Days: Web Directories

Karlsruhe Institute of Technology

dmoz open directory project

Search: **uni mannheim**

Open Directory Categories (1-5 of 100)

1. **World: Français: Régional: Amérique: Etats-Unis: Etat et politique: Ambassades et consulats: Pays représentés** (39)
2. **World: Deutsch: Wissen: Bibliotheken: Universitäts- und Hochschulbibliotheken: Deutschland** (33)
3. **World: Français: Régional: Amérique: Etats-Unis: Etat et politique: Ambassades et consulats: Représentations à l'étranger** (24)
4. **World: Nederlands: Maatschappij: Overheid: Europese Unie** (22)
5. **World: Deutsch: Gesundheit: Krankenhäuser und Kliniken: Universitätskliniken: Deutschland** (21)

more...

Open Directory Sites (1-20 of 15698)

1. amnesty international Hochschulgruppe Mannheim - Die amnesty-Gruppe an der Uni Mannheim setzt sich unter anderem mit Petitionen, Appellbrief ein und stellt aktuelle Aktivitäten vor.
   -- *http://www.amnesty-uni-mannheim.de/* *World: Deutsch: Regional: Europa: Deutschland: Baden-Württemberg: Städte und Gemeinden: M: Mannheim: Gesellschaft (23)*

2. LHG - Liberale Hochschulgruppe Mannheim - Die liberale Studierenden an der Uni Mannheim stellen sich vor.
   -- *http://www.uni-mannheim.de/studorg/liberale/* *World: Deutsch: Wissen: Bildung: Hochschulen: Europa: Deutschland: Baden-Württemberg: Universität Mannheim ()*

3. Uni Mannheim - Marktübersicht deutscher Anbieter für Online-Preisvergleiche.
   -- *http://projekt.wifo.uni-mannheim.de/preisvergleich/* *World: Deutsch: Zuhause: Verbraucherinformationen: Preisagenturen: Online-Preisvergleiche ()*

4. University of Mannheim - Offers undergraduate and graduate programs; includes information on study and research at the university, directories of financial aid.
   -- *http://www.uni-mannheim.de/* *Reference: Education: Colleges and Universities: Europe: Germany: Baden-Württemberg: University of Mannheim ()*

The Early Days: Web Crawlers

- crawls through the web using hyperlinks
- makes an index of the words contained in a page
- ranks pages for a search query according to number of occurences of keywords

The Early Days: Web Crawlers

- crawls through the web using hyperlinks
- makes an index of the words contained in a page
- ranks pages for a search query according to number of occurences of keywords

The Early Days: Web Crawlers

- crawls through the web using hyperlinks
- makes an index of the words contained in a page
- ranks pages for a search query according to number of occurences of keywords

How can we use this massive amout of information?

- 625.3 million websites
- Wikipedia is one website and has several millions of pages

$\Rightarrow$ we need to rank websites!

How can we use this massive amout of information?

- 625.3 million websites
- Wikipedia is one website and has several millions of pages

$\Rightarrow$ we need to rank websites!

How can we use this massive amout of information?

- 625.3 million websites
- Wikipedia is one website and has several millions of pages

$\Rightarrow$ we need to rank websites!

Basics of PageRank

We all know that:

- humans know what is good for them
- ✗ machines don't know what's good for humans
- humans create websites
- humans will only link to websites they like
- ⇒ hyperlinks are a quality indicator

Basics of PageRank

We all know that:

- humans know what is good for them
- ✗ machines don't know what's good for humans
- humans create websites
- humans will only link to websites they like

⇒ hyperlinks are a quality indicator

Basics of PageRank

We all know that:

- humans know what is good for them
- ✗ machines don't know what's good for humans
- humans create websites
- humans will only link to websites they like

⇒ hyperlinks are a quality indicator

Basics of PageRank

We all know that:

- humans know what is good for them
- ✗ machines don't know what's good for humans
- humans create websites
- humans will only link to websites they like

⇒ hyperlinks are a quality indicator

Basics of PageRank

We all know that:

- humans know what is good for them
- ✗ machines don't know what's good for humans
- humans create websites
- humans will only link to websites they like

⇒ hyperlinks are a quality indicator

How Could We Use That?

- simply count number of links to a website
- ✗ 10,000 links from only one page
- count number of websites that link to a website
- ✗ quality of the linking website matters
- ✗ total number of links on the source page matters

How Could We Use That?

- simply count number of links to a website
- ✗ 10,000 links from only one page
- count number of websites that link to a website
- ✗ quality of the linking website matters
- ✗ total number of links on the source page matters

How Could We Use That?

- simply count number of links to a website
- ✗ 10,000 links from only one page
- count number of websites that link to a website
- ✗ quality of the linking website matters
- ✗ total number of links on the source page matters

How Could We Use That?

- simply count number of links to a website
- ✗ 10,000 links from only one page
- count number of websites that link to a website
- ✗ quality of the linking website matters
- ✗ total number of links on the source page matters

How Could We Use That?

- simply count number of links to a website
- ✗ 10,000 links from only one page
- count number of websites that link to a website
- ✗ quality of the linking website matters
- ✗ total number of links on the source page matters

A Brilliant Idea

Sergey Brin

Larry Page

Ideas of PageRank

- decisions of humans are complicated
- a lot of webpages get visited

$\Rightarrow$ modellize clicks on links as random behaviour

- links are important
  - links of page A get less important, if A has many links
  - links of page A get more important, if many link to A

$\Rightarrow$ if B has a link from A, the rank of B increases by $\frac{Rank(A)}{Links(A)}$

if A links to B **then**

$$Rank(B) \mathrel{+}= \frac{Rank(A)}{Links(A)}$$

Ideas of PageRank

- decisions of humans are complicated
- a lot of webpages get visited

$\Rightarrow$ modellize clicks on links as random behaviour

- links are important
  - links of page A get less important, if A has many links
  - links of page A get more important, if many link to A

$\Rightarrow$ if B has a link from A, the rank of B increases by $\frac{Rank(A)}{Links(A)}$

if A links to B **then**

$$Rank(B) \mathrel{+}= \frac{Rank(A)}{Links(A)}$$

Ideas of PageRank

- decisions of humans are complicated
- a lot of webpages get visited

$\Rightarrow$ modellize clicks on links as random behaviour

- links are important
  - links of page A get less important, if A has many links
  - links of page A get more important, if many link to A

$\Rightarrow$ if B has a link from A, the rank of B increases by $\frac{Rank(A)}{Links(A)}$

if A links to B **then**

$$Rank(B) \mathrel{+}= \frac{Rank(A)}{Links(A)}$$

Ideas of PageRank

- decisions of humans are complicated
- a lot of webpages get visited

$\Rightarrow$ modellize clicks on links as random behaviour

- links are important
  - links of page A get less important, if A has many links
  - links of page A get more important, if many link to A

$\Rightarrow$ if B has a link from A, the rank of B increases by $\frac{Rank(A)}{Links(A)}$

if A links to B **then**

$$Rank(B) \mathrel{+}= \frac{Rank(A)}{Links(A)}$$

Ideas of PageRank

- decisions of humans are complicated
- a lot of webpages get visited

$\Rightarrow$ modellize clicks on links as random behaviour

- links are important
  - links of page A get less important, if A has many links
  - links of page A get more important, if many link to A

$\Rightarrow$ if B has a link from A, the rank of B increases by $\frac{Rank(A)}{Links(A)}$

if A links to B **then**

$$Rank(B) \mathrel{+}= \frac{Rank(A)}{Links(A)}$$

Ideas of PageRank

- decisions of humans are complicated
- a lot of webpages get visited

$\Rightarrow$ modellize clicks on links as random behaviour

- links are important
  - links of page A get less important, if A has many links
  - links of page A get more important, if many link to A

$\Rightarrow$ if B has a link from A, the rank of B increases by $\frac{Rank(A)}{Links(A)}$

if A links to B **then**

$$Rank(B) \mathrel{+}= \frac{Rank(A)}{Links(A)}$$

Ideas of PageRank

- decisions of humans are complicated
- a lot of webpages get visited

$\Rightarrow$ modellize clicks on links as random behaviour

- links are important
  - links of page A get less important, if A has many links
  - links of page A get more important, if many link to A

$\Rightarrow$ if B has a link from A, the rank of B increases by $\frac{Rank(A)}{Links(A)}$

if A links to B **then**

$$Rank(B) \mathrel{+}= \frac{Rank(A)}{Links(A)}$$

Ideas of PageRank

- decisions of humans are complicated
- a lot of webpages get visited

$\Rightarrow$ modellize clicks on links as random behaviour

- links are important
  - links of page A get less important, if A has many links
  - links of page A get more important, if many link to A

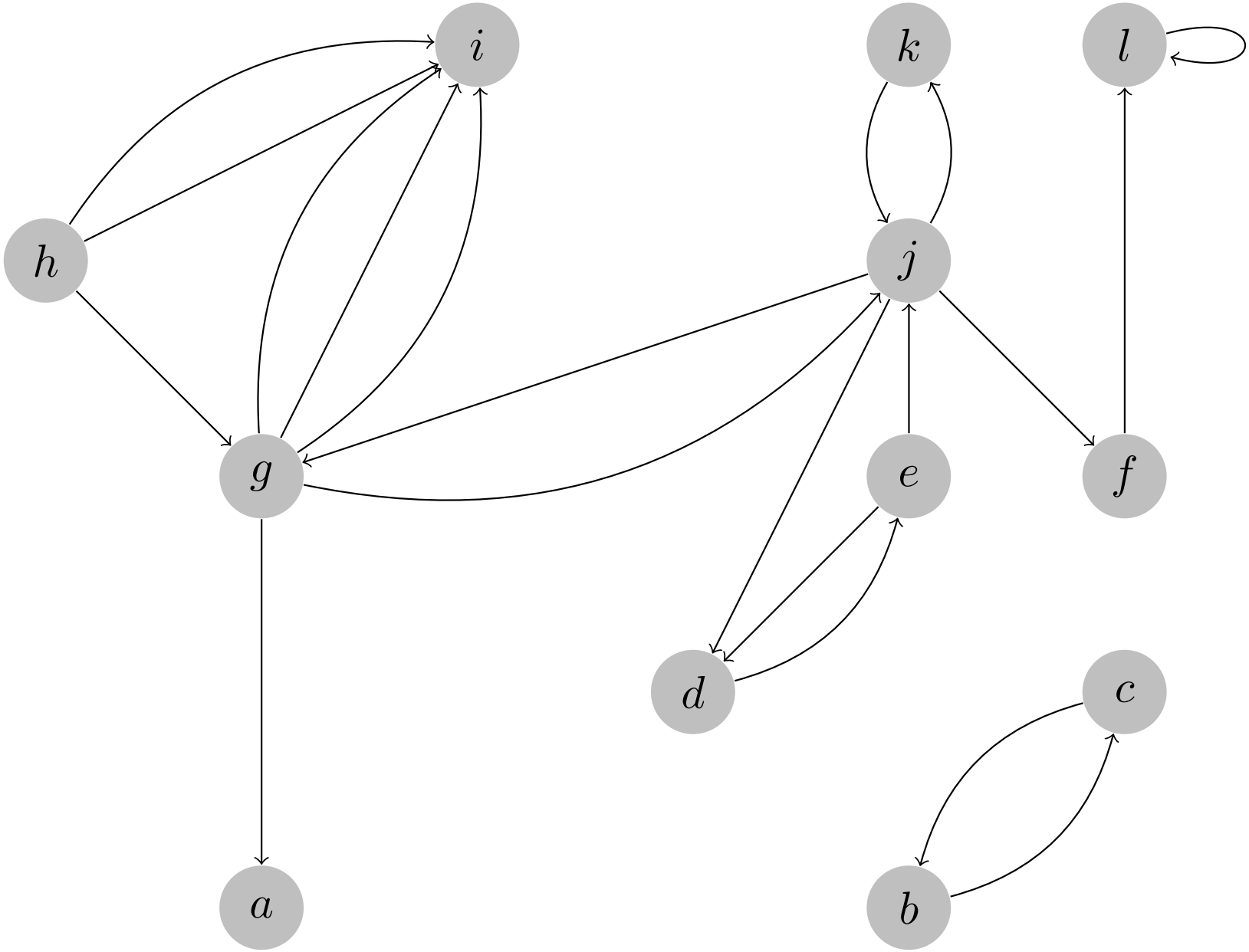
$\Rightarrow$ if B has a link from A, the rank of B increases by $\frac{Rank(A)}{Links(A)}$

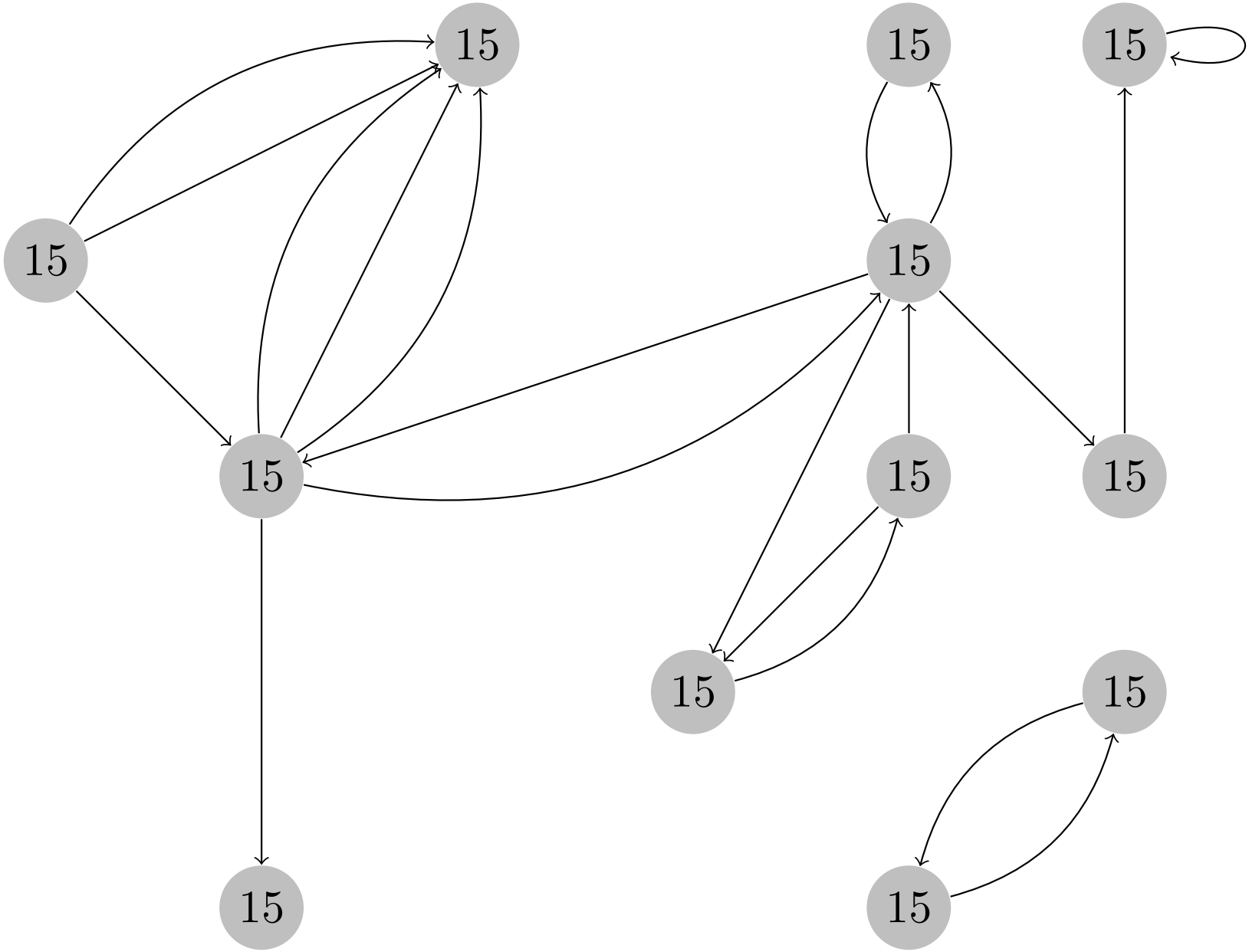
if A links to B **then**

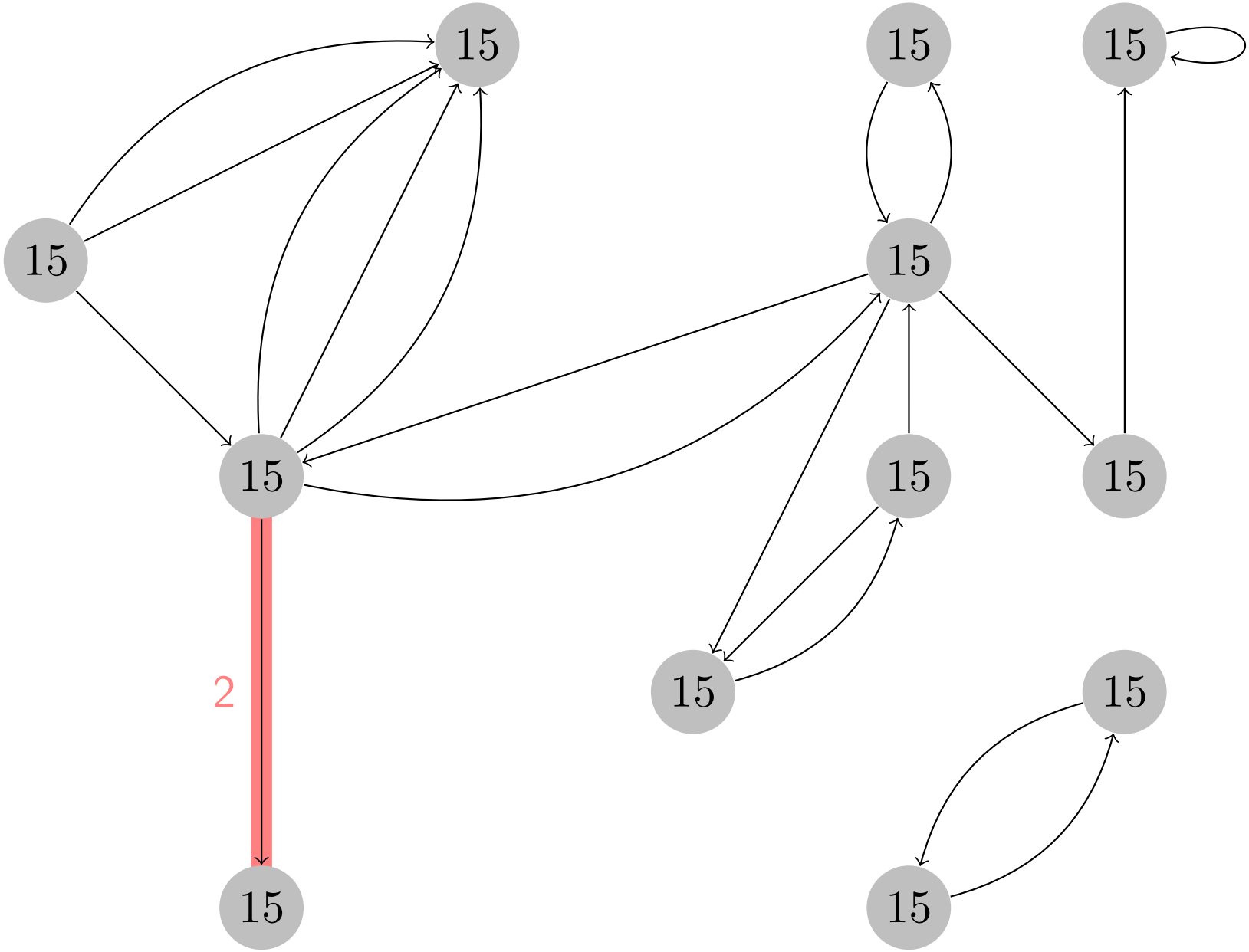
$$Rank(B) \mathrel{+}= \frac{Rank(A)}{Links(A)}$$

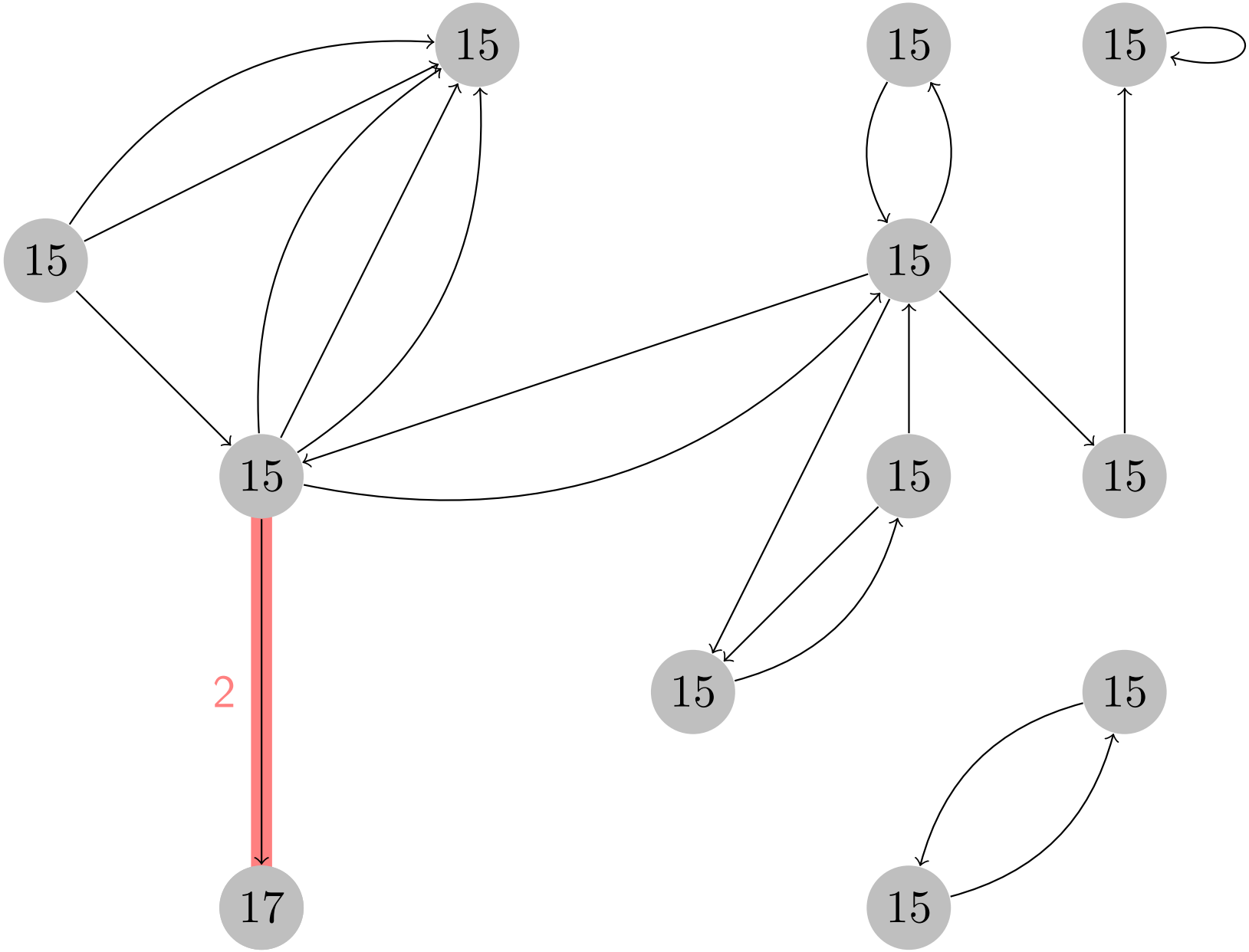
What is PageRank?

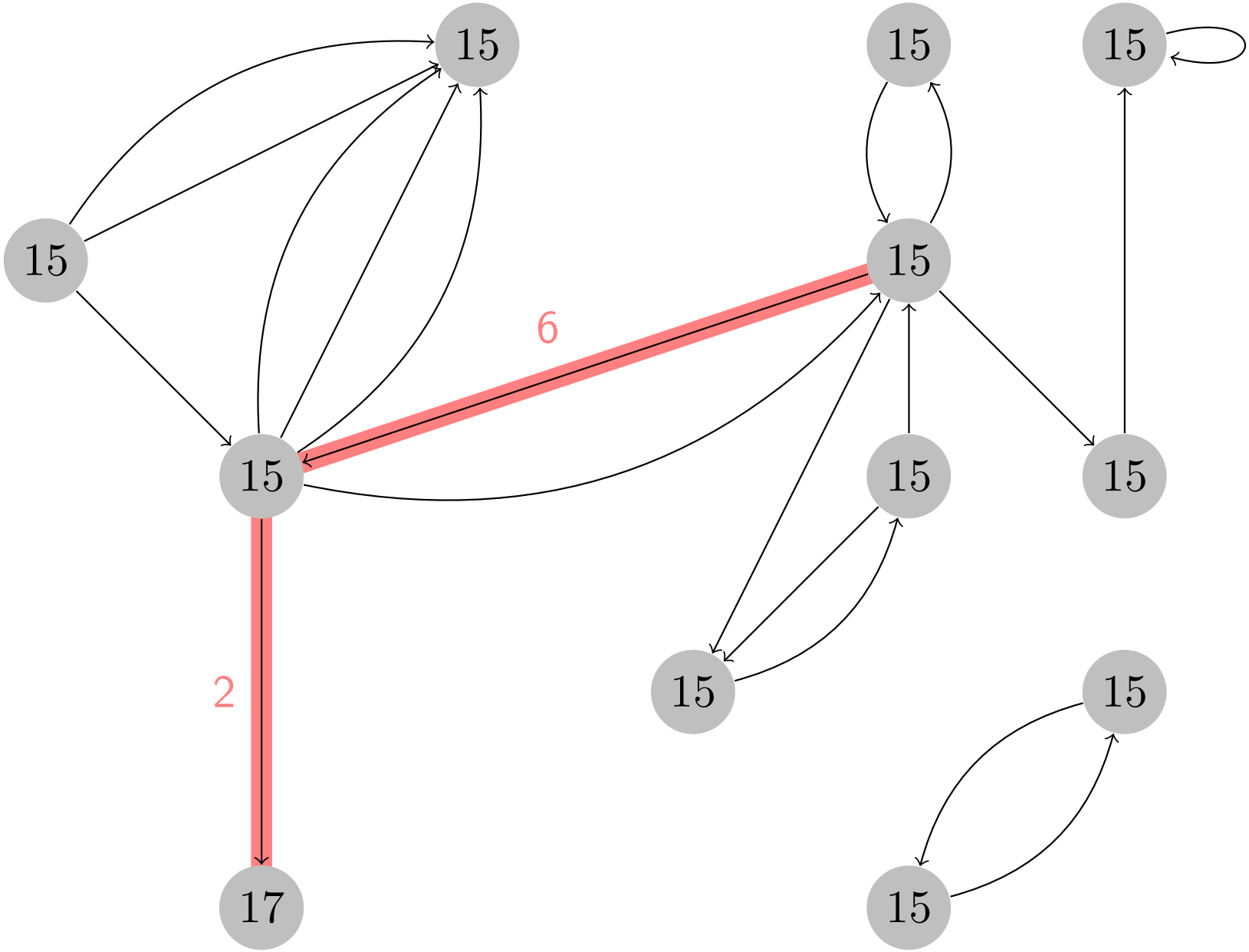
The PageRank algorithm calculates the probability of a randomly clicking user ending up on a given page.

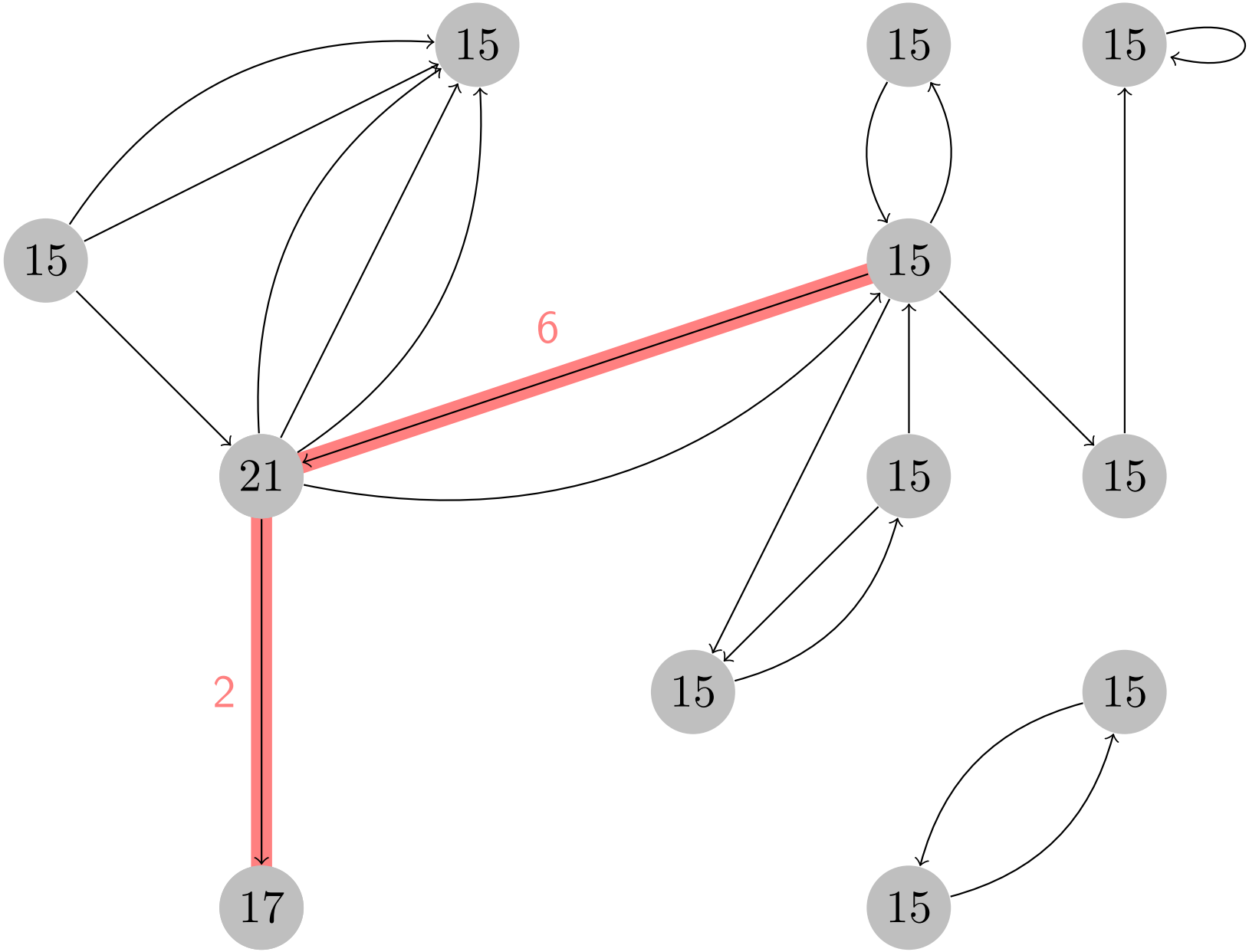

i
k
l
h
j
g
e
f
d
c
a
b

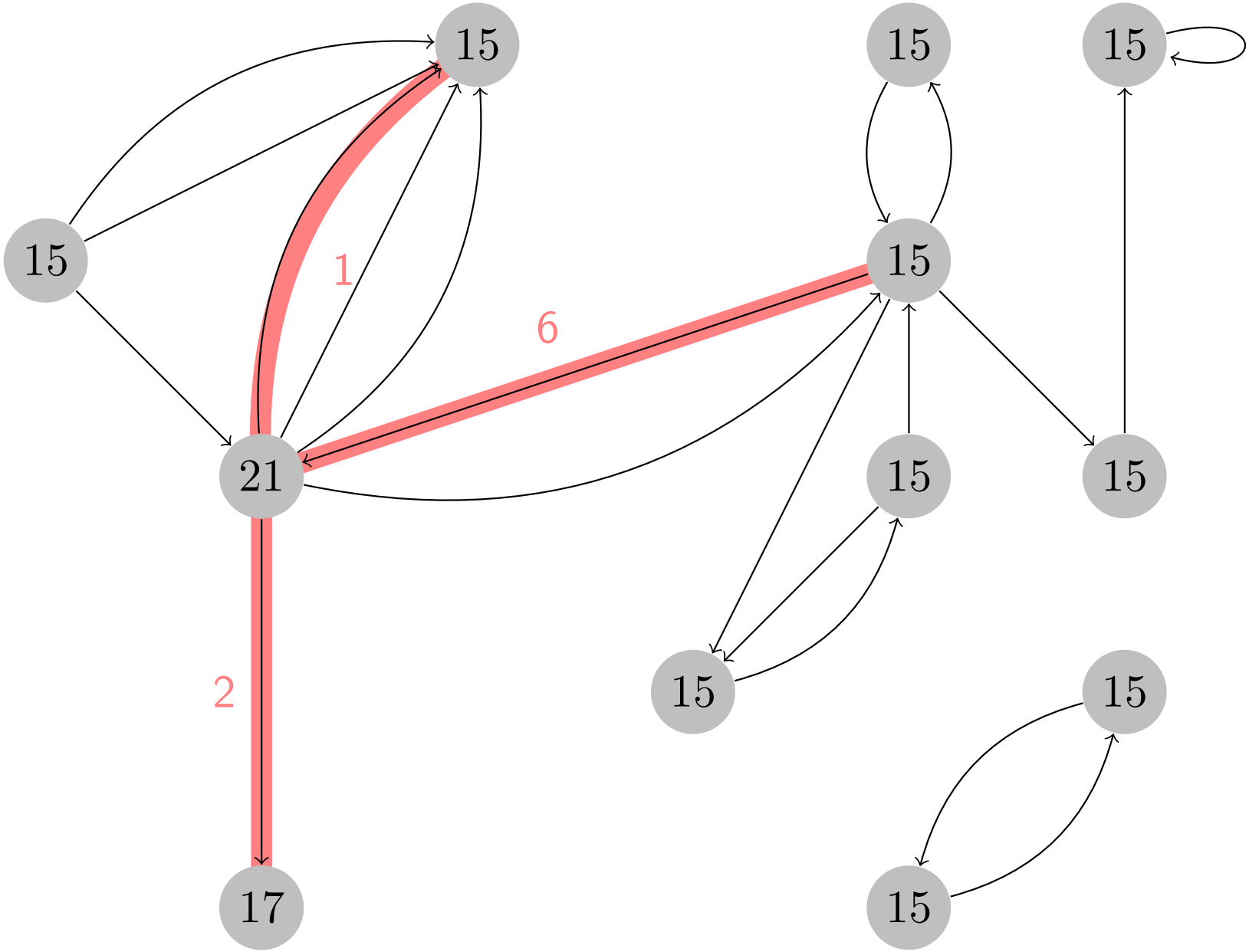

15
15
15
15
15
15
15
15
15
15
15
15

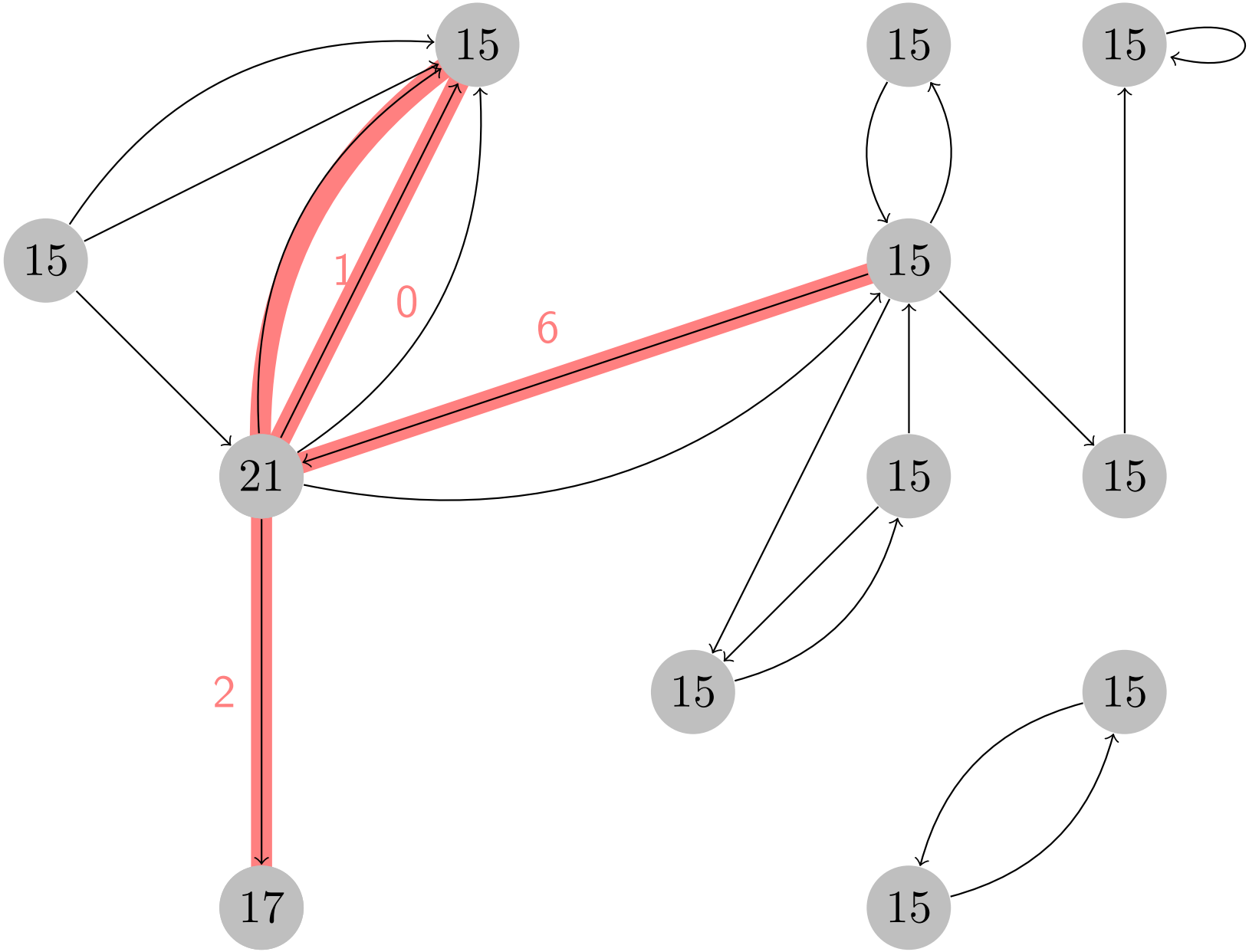

15
15
15
15
15
15
15
15
15
15
15
2
15

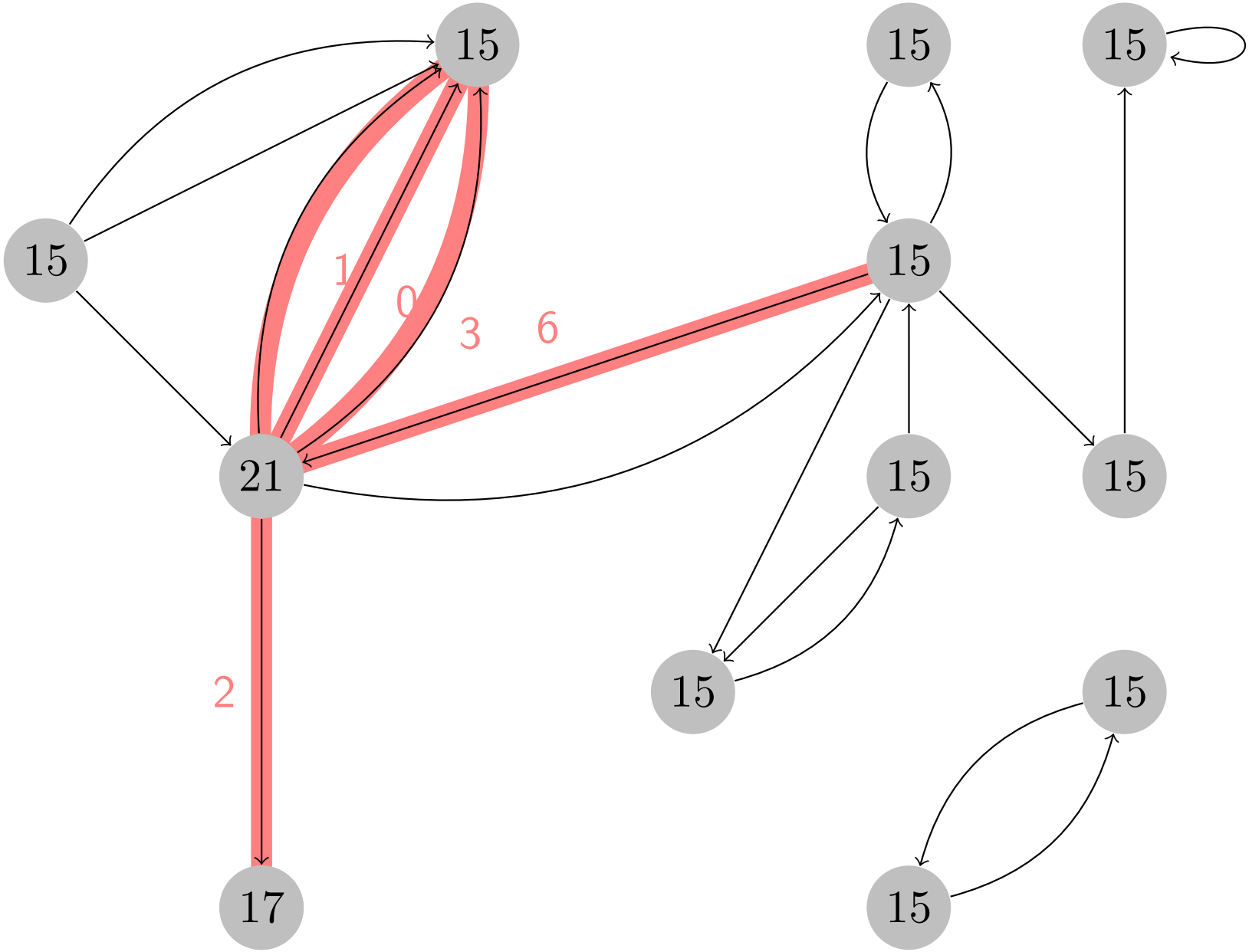

15
15
15
15
15
15
15
15
15
15
2
17

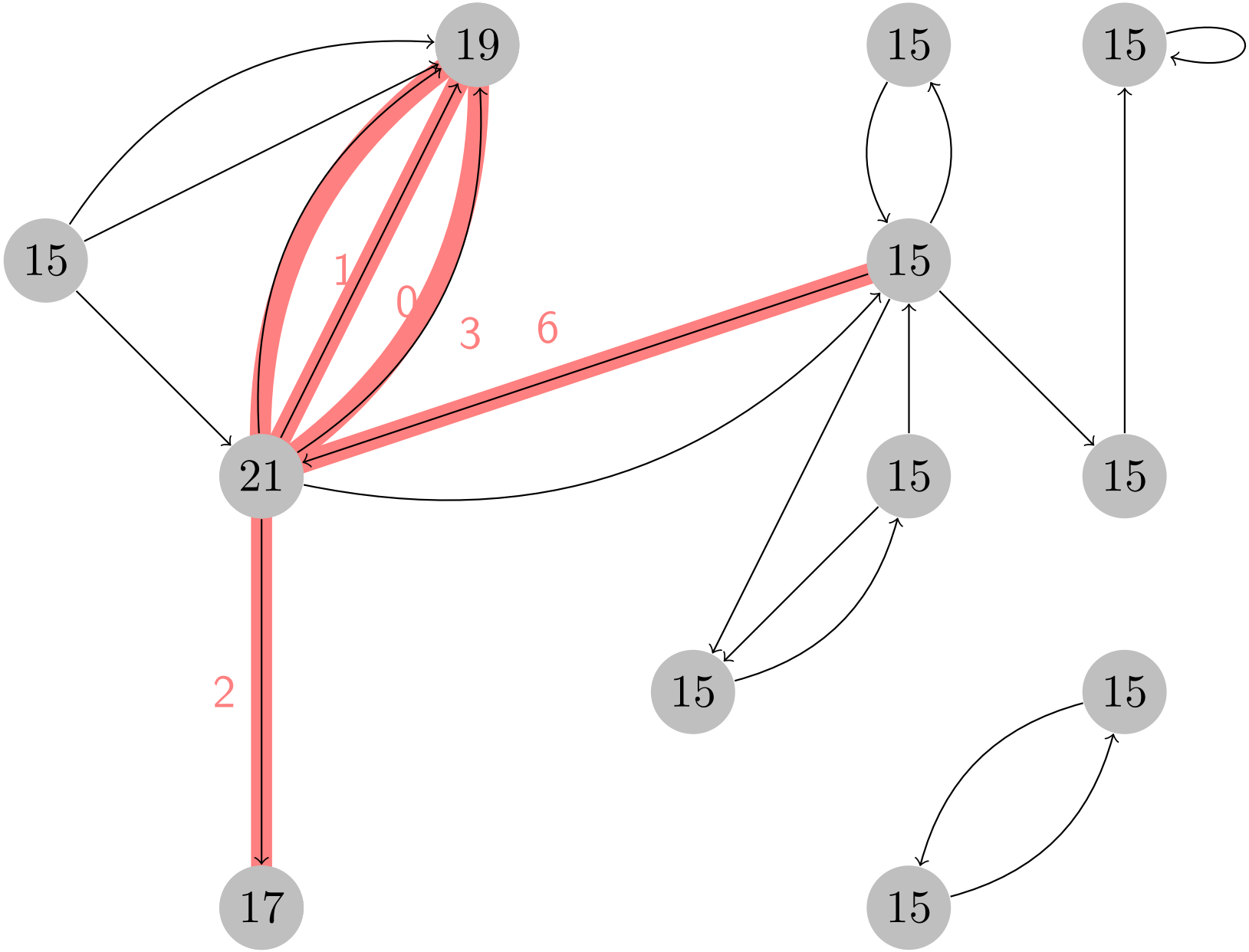

15
15
15
15
15
6
15
15
15
2
15
15
17
15


15
15
15
15
15
6
21
15
15
2
15
15
17
15


15
15
15
15
15
1
6
21
15
15
2
15
15
17
15


15
15
15
15
15
21
15
15
15
15
15
17
1
0
6
2


15
15
15
15
15
21
15
15
15
15
15
17
1
0
3
6
2


19
15
15
15
15
1
0
3
6
21
15
15
2
15
15
17
15

Mathematics

Let x be a web page. Then

- $L(x)$ is the set of websites that link to x
- $C(y)$ is the out-degree of page y
- α is probability of random jump
- N is the total number of websites

$$PR(x) := \alpha \left(\frac{1}{N}\right) + (1-\alpha) \sum_{y \in L(x)} \frac{PR(y)}{C(y)}$$

Pseudocode

function PageRank(Graph web, double $q = 0.15$, int $iterations$)
    while $iterations > 0$ **do**
        for all $page \in web$ **do** ▷ calculate pageRank of $page$
            $page.pageRank = q$
            for all $y \in L(page)$ **do**
                $page.pageRank \mathrel{+}= \frac{y.pageRank}{C(y)}$
            end for
        end for
        $iterations \mathrel{-}= 1$

Pseudocode

function PageRank(Graph web, double $q = 0.15$, int $iterations$)
  while $iterations > 0$ **do**
    for all $page \in web$ **do** ▷ calculate pageRank of $page$
      $page.pageRank = q$
      for all $y \in L(page)$ **do**
        $page.pageRank \mathrel{+}= \frac{y.pageRank}{C(y)}$
      end for
    end for
    $iterations$ -= 1

Pseudocode

function PageRank(Graph web, double $q = 0.15$, int $iterations$)
  while $iterations > 0$ **do**
    for all $page \in web$ **do** ▷ calculate pageRank of $page$
      $page.pageRank = q$
      for all $y \in L(page)$ **do**
        $page.pageRank \mathrel{+}= \frac{y.pageRank}{C(y)}$
      end for
    end for
    $iterations$ -= 1

Pseudocode

function PageRank(Graph web, double $q = 0.15$, int $iterations$)
    while $iterations > 0$ **do**
        for all $page \in web$ **do** ▷ calculate pageRank of $page$
            $page.pageRank = q$
            for all $y \in L(page)$ **do**
                $page.pageRank \mathrel{+}= \frac{y.pageRank}{C(y)}$
            end for
        end for
        $iterations$ -= 1

Pseudocode

function PageRank(Graph web, double $q = 0.15$, int $iterations$)
    while $iterations > 0$ **do**
        for all $page \in web$ **do** ▷ calculate pageRank of $page$
            $page.pageRank = q$
            for all $y \in L(page)$ **do**
                $page.pageRank \mathrel{+}= \frac{y.pageRank}{C(y)}$
            end for
        end for
        $iterations \mathrel{-}= 1$

Pseudocode

function PageRank(Graph web, double $q = 0.15$, int $iterations$)
  while $iterations > 0$ **do**
    for all $page \in web$ **do** ▷ calculate pageRank of $page$
      $page.pageRank = q$
      for all $y \in L(page)$ **do**
        $page.pageRank \mathrel{+}= \frac{y.pageRank}{C(y)}$
      end for
    end for
    $iterations$ -= 1

Factors of Influence

- language
- location
- social information: what your friends like

Factors of Influence

- language
- location
- social information: what your friends like

Factors of Influence

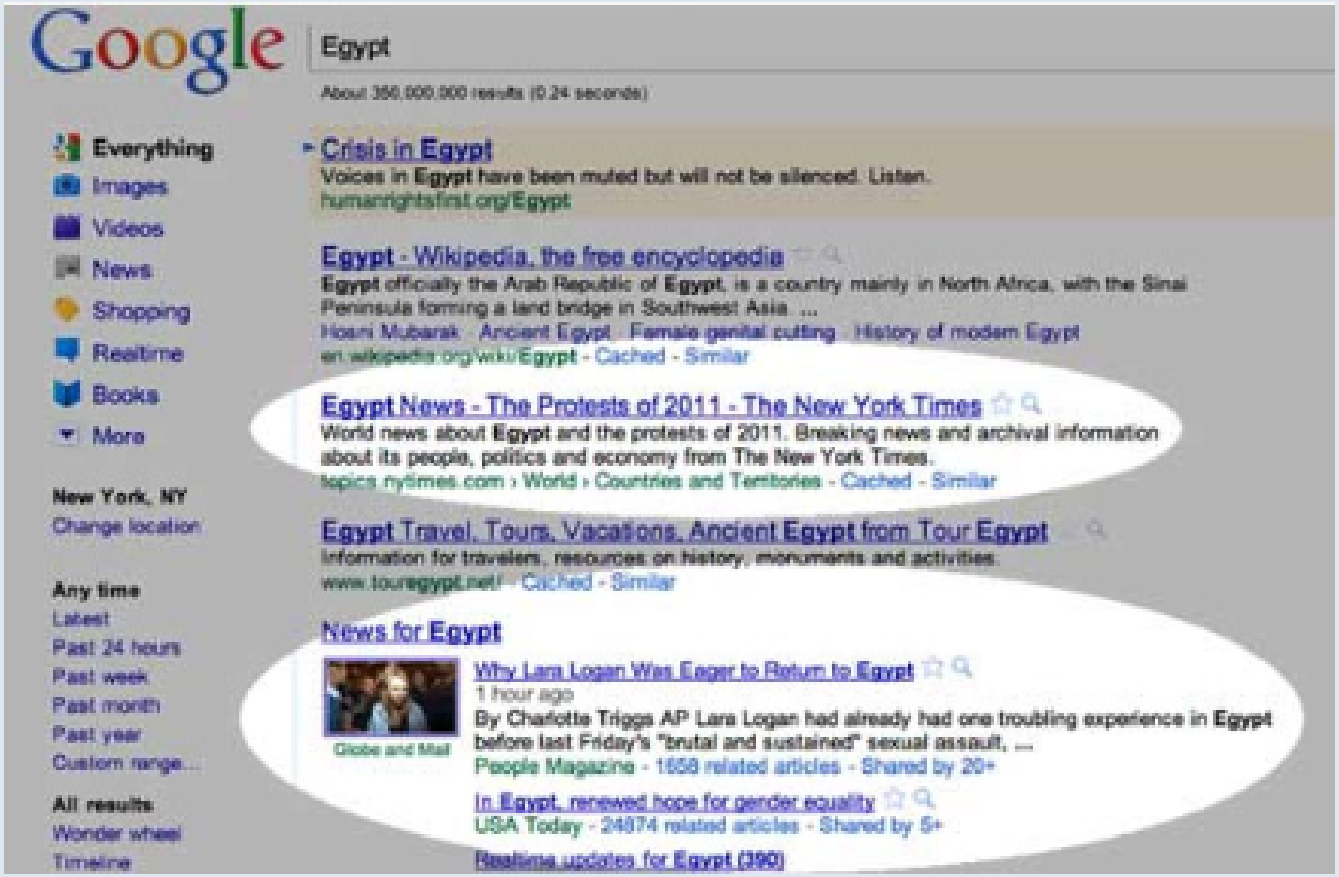
- language
- location
- social information: what your friends like

The Filter Bubble

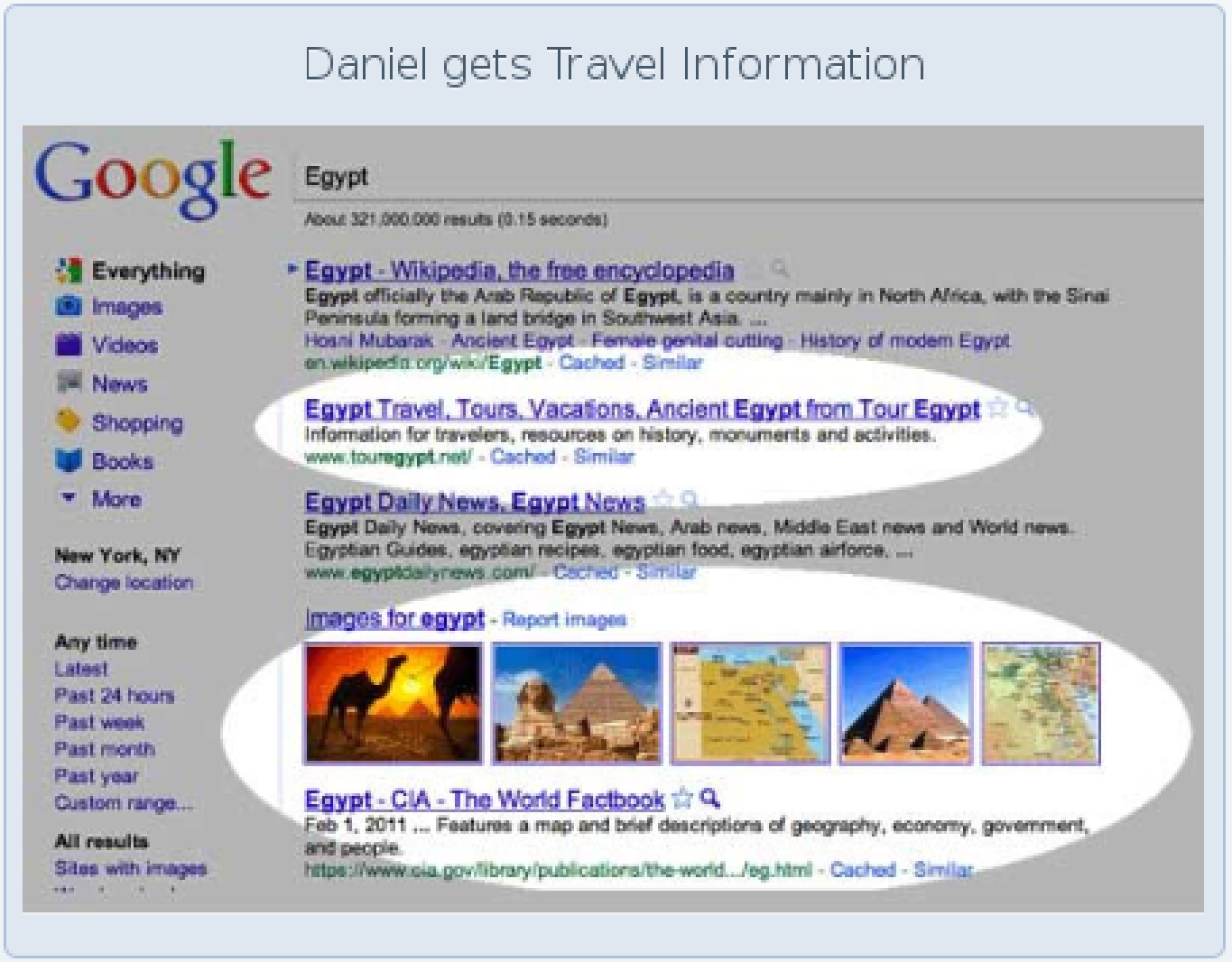
dontbubble.us
www.thefilterbubble.com

The Filter Bubble

Scott gets Egyptian Protests



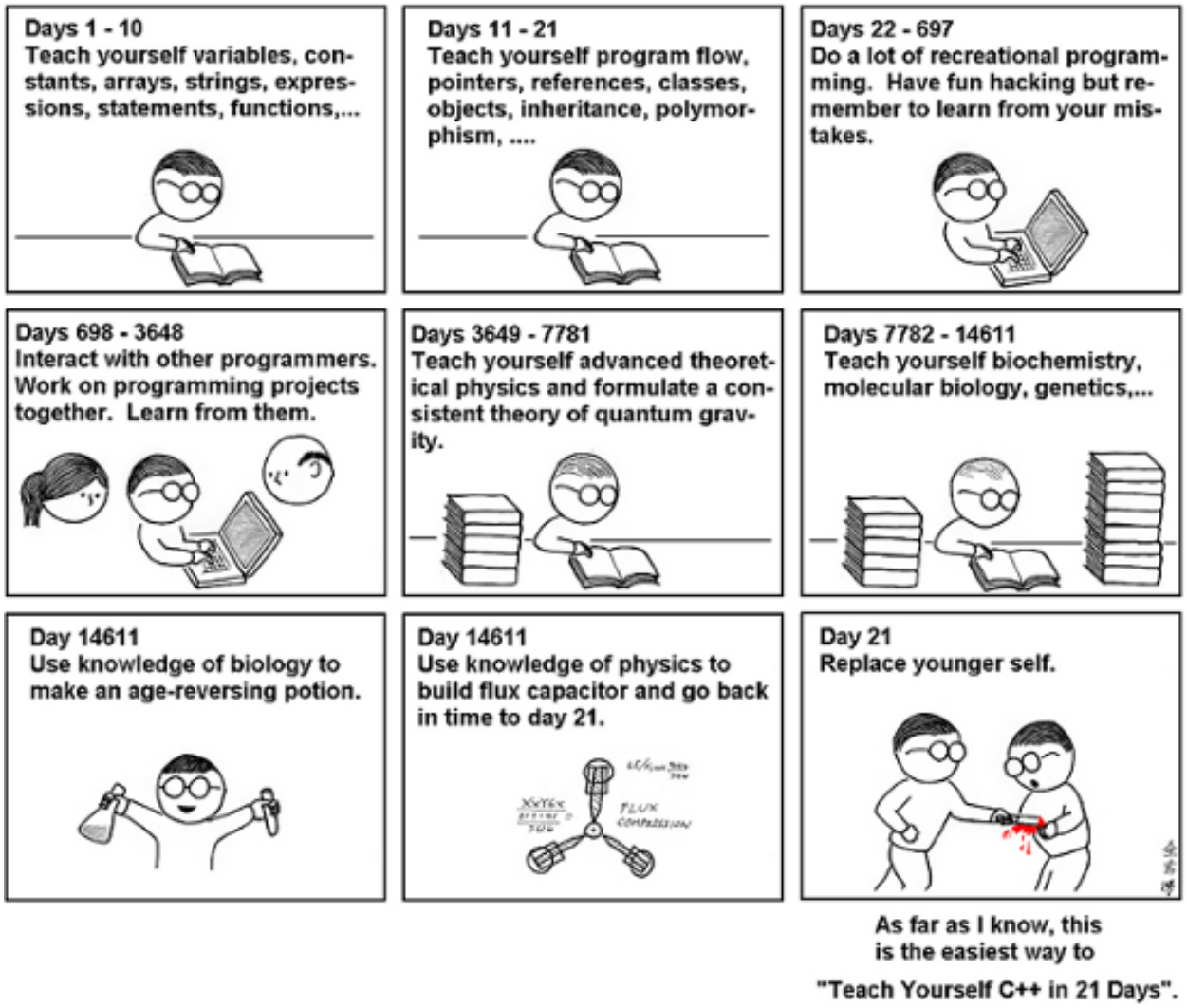
The Filter Bubble



What You've Learned

- web directories
- web crawler
- graph (nodes, eges)
- random walk (ants)
- PageRank
- read pseudocode
- filter bubble

Image Sources

- PageRank by Felipe Micaroni Lalli
- screenshots of www.dmoz.org
- Hyperlink by Bernard Ladenthin
- screenshots of dontbubble.us
- Sergey Brin by enlewof
- Larry Page by aweigend

Thanks for Your Attention!