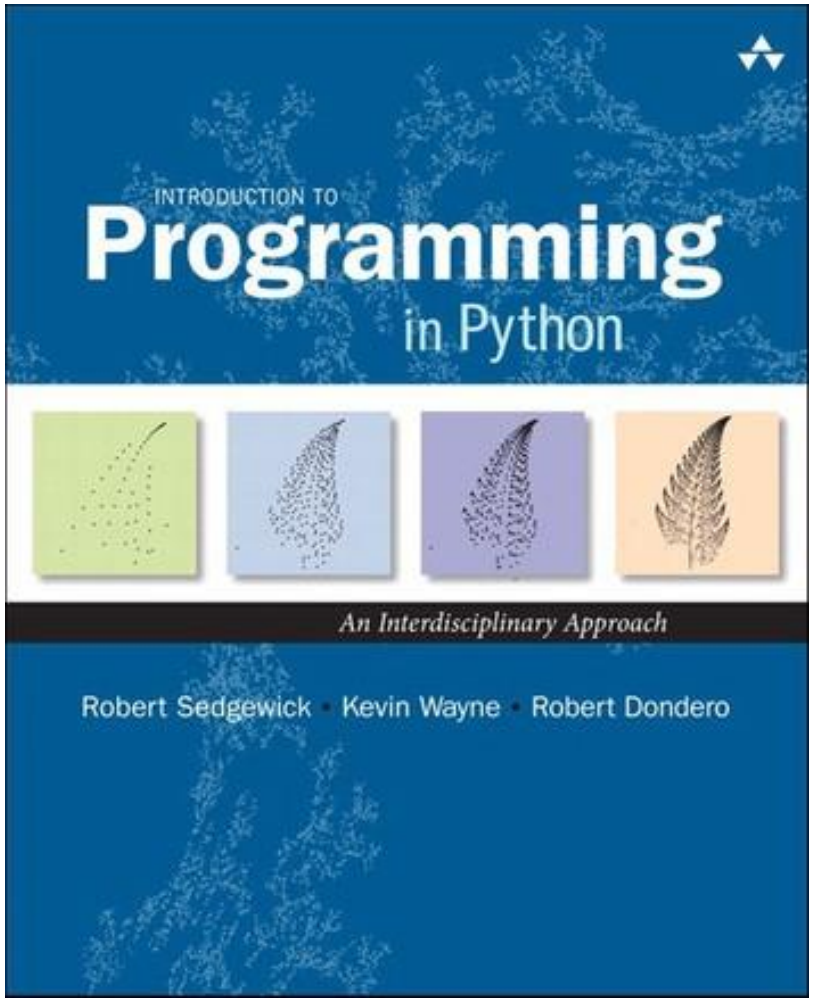
عناصر برنامه‌نویسی: کاوشگر تصادفی وب

سید ناصر رضوی **www.snrazavi.ir**
۱۳۹۸

۱-۶ کاوشگر وب

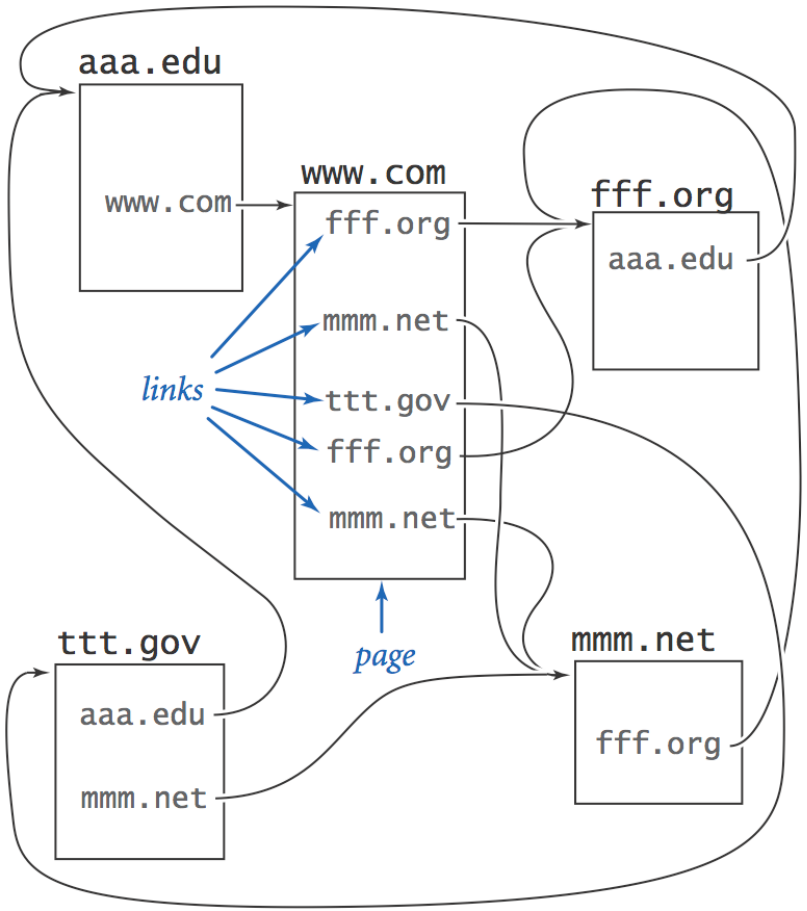

INTRODUCTION TO
Programming
in Python
An Interdisciplinary Approach
Robert Sedgewick · Kevin Wayne · Robert Dondero

جستجوی وب

- مرتبط بودن. آیا سند مفروض شبیه واژه‌های پرس و جو است؟
- اهمیت. آیا سند مفروض برای کاربرهای گوناگون مفید است؟

الگوریتم PageRank

- الگوریتم PageRank گوگل. [سرجی برین و لری پیج، ۱۹۹۸]
  - اندازه‌گیری محبوبیت صفحات بر اساس ساختار پیوندی وب.
  - یک انقلاب در زمینه دسترسی به اطلاعات.

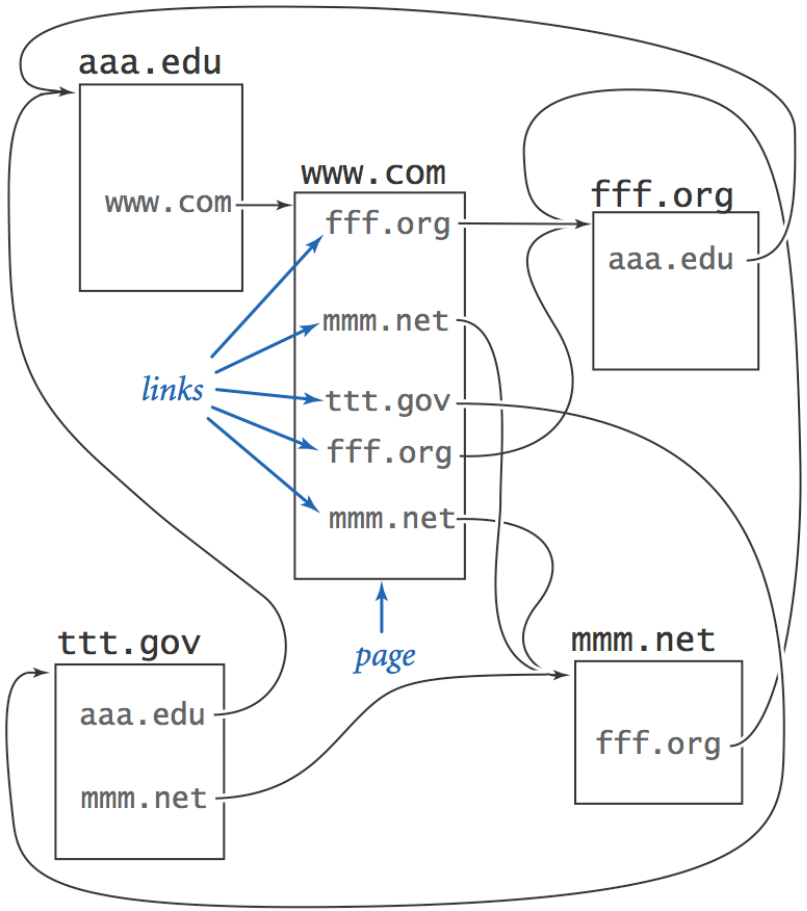


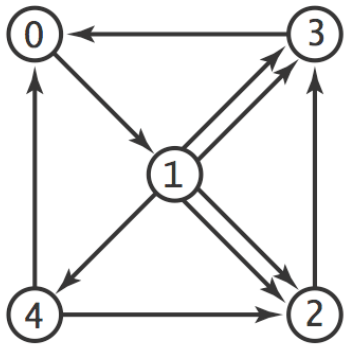
قاعده ۹۰-۱۰

- مدل. انتخاب صفحه بعدی به وسیله کاوشگر وب:
  - ۹۰ درصد مواقع، کاوشگر بر روی یک پیوند تصادفی کلیک می‌کند.
  - ۱۰ درصد مواقع، کاوشگر به یک صفحه کاملاً تصادفی می‌رود.
- هشدارها. یک مدل خام اما مفید برای کاوش صفحات وب.
  - هیچ‌کس پیوندها را با احتمال تصادفی انتخاب نمی‌کند.
  - قاعده ۹۰-۱۰ تنها یک حدس است.
  - این مدل دکمه بازگشت به صفحه قبل و همچنین صفحات نشانه‌گذاری شده را در نظر نمی‌گیرد.
  - تنها می‌تواند از عهده یک نمونه بسیار کوچک از وب برآید.
  - ...

قالب ورودی برای گراف وب

- قالب ورودی.
  - N صفحه با شماره‌های 0 تا N - 1
  - نمایش هر پیوند به وسیله یک زوج از اعداد صحیح





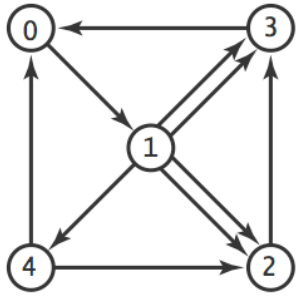

```
% more tiny.txt
5          ← N
0 1
1 2   1 2
1 3   1 3   1 4
2 3
3 0
4 0   4 2
```

links

ماتریس انتقال

- ماتریس انتقال. `p[i][j]` = احتمال رفتن از صفحه `i` به صفحه `j`

input graph




```
5
0 1
1 2   1 2
1 3   1 3   1 4
2 3
3 0
4 0   4 2
```

link counts

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 2 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$

degrees

$$\begin{bmatrix} 1 \\ 5 \\ 1 \\ 1 \\ 2 \end{bmatrix}$$

leap probabilities

$$\begin{bmatrix} .02 & .02 & .02 & .02 & .02 \\ .02 & .02 & .02 & .02 & .02 \\ .02 & .02 & .02 & .02 & .02 \\ .02 & .02 & .02 & .02 & .02 \\ .02 & .02 & .02 & .02 & .02 \end{bmatrix} +$$

link probabilities

$$\begin{bmatrix} 0 & .90 & 0 & 0 & 0 \\ 0 & 0 & .36 & .36 & .18 \\ 0 & 0 & 0 & .90 & 0 \\ .90 & 0 & 0 & 0 & 0 \\ .45 & 0 & .45 & 0 & 0 \end{bmatrix} =$$

transition matrix

$$\begin{bmatrix} .02 & .92 & .02 & .02 & .02 \\ .02 & .02 & .38 & .38 & .20 \\ .02 & .02 & .02 & .92 & .02 \\ .92 & .02 & .02 & .02 & .02 \\ .47 & .02 & .47 & .02 & .02 \end{bmatrix}$$

احتمال رفتن از صفحه ۱ به صفحه ۲

محاسبه ماتریس انتقال از روی گراف وب

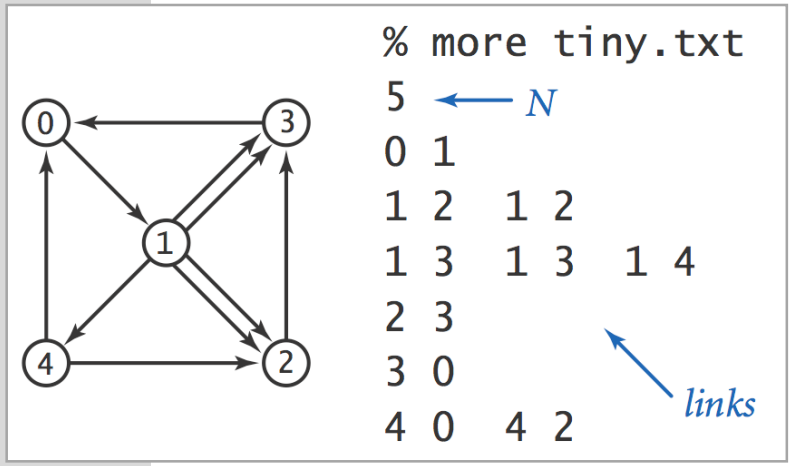

```
import stdio

n = stdio.readInt()
link_counts = [[0] * n for i in range(n)]
out_degrees = [0] * n

while not stdio.isEmpty():
    i = stdio.readInt()
    j = stdio.readInt()
    out_degrees[i] += 1
    link_counts[i][j] += 1

print('%d %d' % (n, n))

for i in range(n):
    for j in range(n):
        p = .9 * (link_counts[i][j] / out_degrees[i]) + .1 / n
        print('%8.5f' % p, end='')
    print()
```




```
% python transition.py < tiny.txt
5 5
 0.02000 0.92000 0.02000 0.02000 0.02000
 0.02000 0.02000 0.38000 0.38000 0.20000
 0.02000 0.02000 0.02000 0.92000 0.02000
 0.92000 0.02000 0.02000 0.02000 0.02000
 0.47000 0.02000 0.47000 0.02000 0.02000
```

قاعده ۹۰-۱۰

شبیه‌سازی مونت کارلو

- شبیه‌سازی مونت کارلو.
  - کاوشگر از صفحه صفر شروع می‌کند.
  - به صورت تکراری و بر اساس ماتریس انتقال صفحه بعدی را انتخاب می‌کند.
  - سپس محاسبه می‌کند از هر صفحه به چه میزان بازدید شده است.

چگونه؟ اسلاید بعد را ببینید

$$\begin{bmatrix} .02 & .92 & .02 & .02 & .02 \\ .02 & .02 & .38 & .38 & .20 \\ .02 & .02 & .02 & 92 & .02 \\ .92 & .02 & .02 & .02 & .02 \\ .47 & .02 & .47 & .02 & .02 \end{bmatrix}$$

ماتریس انتقال

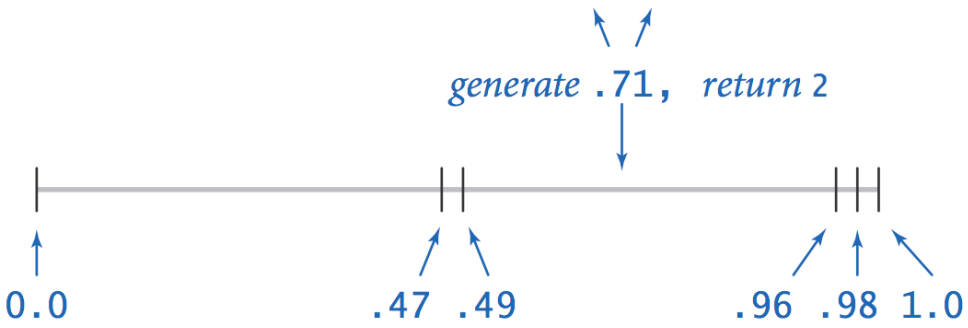
کاوشگر تصادفی

- حرکت تصادفی. فرض کنید کاوشگر در صفحه page است؟ صفحه بعدی j چگونه انتخاب می‌شود؟
  - سطر page از ماتریس انتقال شامل احتمالات مربوط به این صفحه است.
  - احتمال‌های تجمعی برای این صفحه محاسبه می‌شود.
  - یک عدد تصادفی مانند r در بازه 0.0 و 1.0 تولید می‌شود.
  - صفحه بعدی j بر اساس بازه‌ای که r در آن قرار گرفته انتخاب می‌شود.

page	.47	.02	.47	.02	.02

ماتریس انتقال

j	0	1	2	3	4
probabilities p[page][j]	.47	.02	.47	.02	.02
cumulated sum *values*	.47	.49	.96	.98	1.0

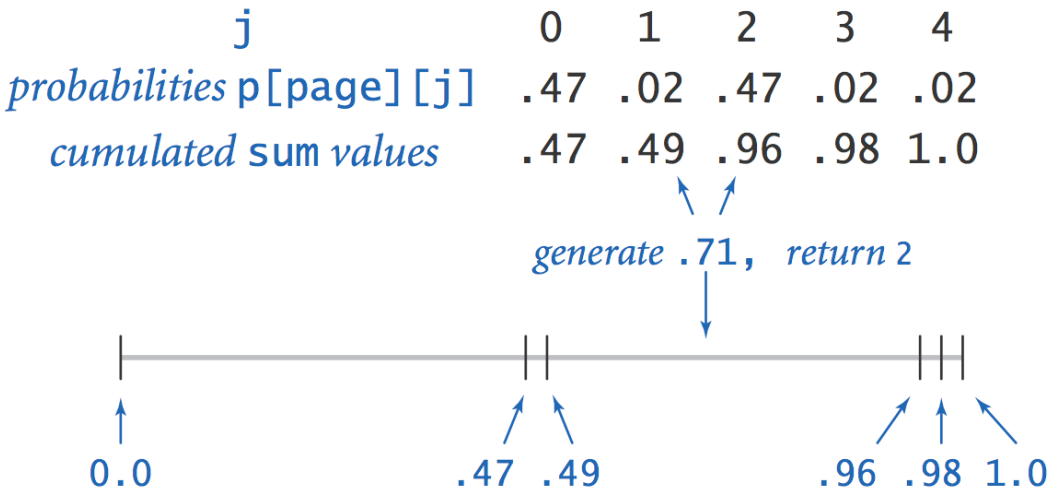


کاوشگر تصادفی

- حرکت تصادفی. فرض کنید کاوشگر در صفحه page است؟ صفحه بعدی j چگونه انتخاب می‌شود؟
  - سطر page از ماتریس انتقال شامل احتمالات مربوط به این صفحه است.
  - احتمال‌های تجمعی برای این صفحه محاسبه می‌شود.
  - یک عدد تصادفی مانند r در بازه 0.0 و 1.0 تولید می‌شود.
  - صفحه بعدی j بر اساس بازه‌ای که r در آن قرار گرفته انتخاب می‌شود.

```
r = random.random()
total = 0.0

for j in range(n):
    total += p[page][j]
    if r < total:
        page = j
        break
```



کاوشگر وب: شبیه‌سازی مونت کارلو (۱)


```
import sys
import random
import stdio

moves = int(sys.argv[1])

n = stdio.readInt()
stdio.readInt()

p = [[0.0] * n for i in range(n)]

for i in range(n):
    for j in range(n):
        p[i][j] = stdio.readFloat()
...
```

```
% python transition.py < tiny.txt
5 5
 0.02000 0.92000 0.02000 0.02000 0.02000
 0.02000 0.02000 0.38000 0.38000 0.20000
 0.02000 0.02000 0.02000 0.92000 0.02000
 0.92000 0.02000 0.02000 0.02000 0.02000
 0.47000 0.02000 0.47000 0.02000 0.02000

% python transition.py < tiny.txt | randomsurfer.py 1000000
```

کاوشگر وب: شبیه‌سازی مونت کارلو (۲)


```
hits = [0] * n
page = 0
for i in range(moves):
    r = random.random()
    total = 0.0
    for j in range(n):
        total += p[page][j]
        if r < total:
            page = j
            break                    حرکت از صفحه page به صفحه j
    hits[page] += 1
for v in hits:
    print('%8.5f' % (1. * v / moves), end='')
```

```
% python transition.py < tiny.txt
5 5
 0.02000 0.92000 0.02000 0.02000 0.02000
 0.02000 0.02000 0.38000 0.38000 0.20000
 0.02000 0.02000 0.02000 0.92000 0.02000
 0.92000 0.02000 0.02000 0.02000 0.02000
 0.47000 0.02000 0.47000 0.02000 0.02000

% python transition.py < tiny.txt | randomsurfer.py 1000000
0.27480 0.26180 0.14860 0.25030 0.06450
```

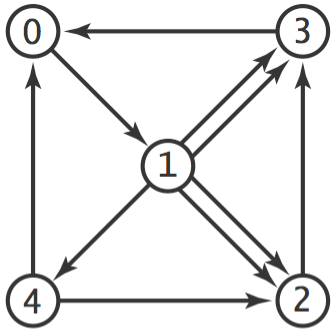
همگرایی: کمی ریاضیات

- همگرایی. برای مدل کاوشگر تصادفی، مدت زمانی که کاوشگر بر روی هر صفحه صرف می‌کند، به یک توزیع منحصربه‌فرد همگرا می‌شود که مستقل از صفحه شروع است.

رتبه صفحه

توزیع ایستای زنجیره مارکوف

بردارهای ویژه اصلی ماتریس انتقال



0 .27
1 .27
3 .25
2 .15
4 .07

$$\left[\frac{428{,}671}{1{,}570{,}055} \quad \frac{417{,}205}{1{,}570{,}055} \quad \frac{229{,}519}{1{,}570{,}055} \quad \frac{388{,}162}{1{,}570{,}055} \quad \frac{106{,}498}{1{,}570{,}055}\right]$$

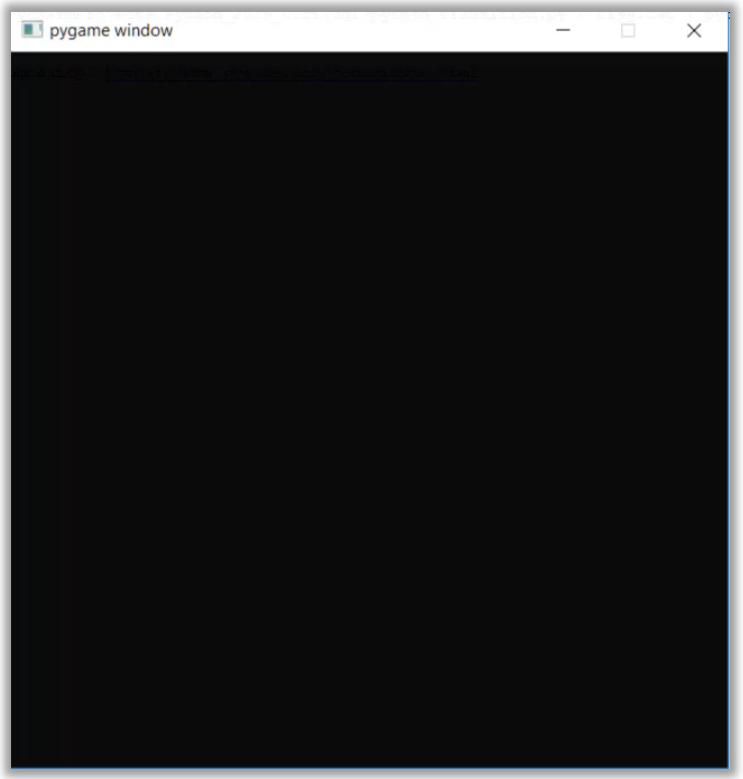
همگرایی: کمی ریاضیات

- همگرایی. برای مدل کاوشگر تصادفی، مدت زمانی که کاوشگر بر روی هر صفحه صرف می‌کند، به یک توزیع منحصربه‌فرد همگرا می‌شود که مستقل از صفحه شروع است.

رتبه صفحه

توزیع ایستای زنجیره مارکوف

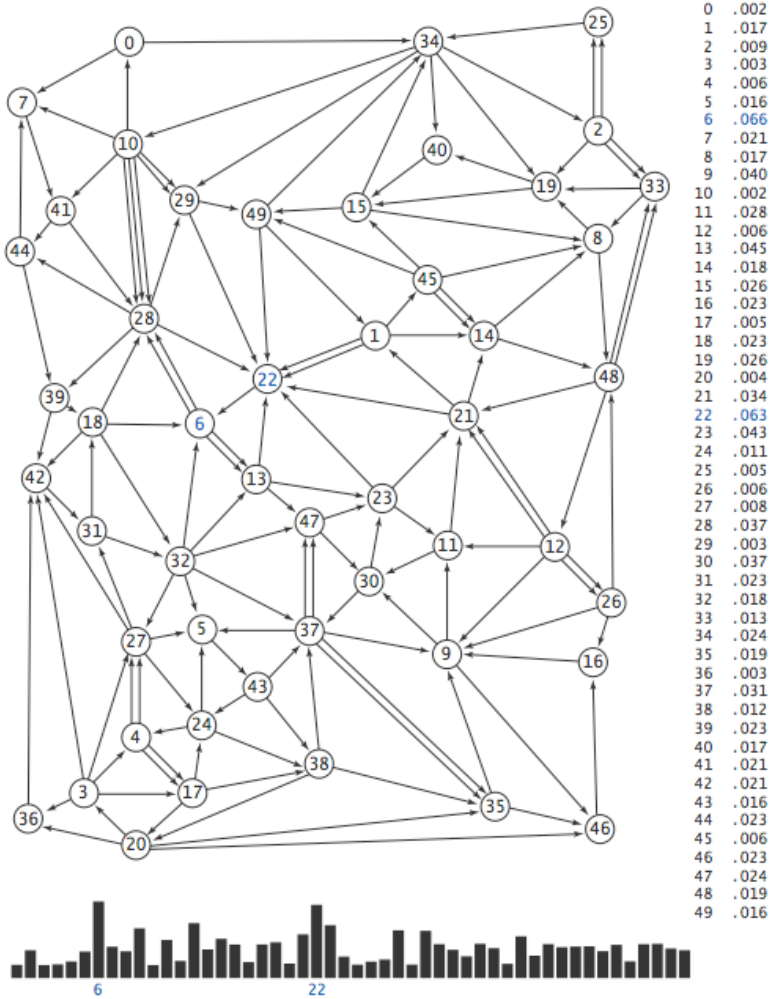
بردارهای ویژه اصلی ماتریس انتقال



$$\left[\frac{428{,}671}{1{,}570{,}055} \quad \frac{417{,}205}{1{,}570{,}055} \quad \frac{229{,}519}{1{,}570{,}055} \quad \frac{388{,}162}{1{,}570{,}055} \quad \frac{106{,}498}{1{,}570{,}055}\right]$$

یک گراف بزرگ‌تر


```
% more medium.txt
50
 0  7  0 34
 1 14  1 22  1 22  1 45
 2 19  2 25  2 33
 3  4  3 17  3 27  3 36  3 42
 4 17  4 17  4 27  4 27
 5 43
 6 13  6 13  6 28
 7 41
 8 19  8 48
 9 11  9 30  9 46
 ...
```



Page ranks with histogram for a larger example