

# 重要情報を見つけよ！ Googleの革新をもたらした数学とは

お茶の水女子大学  
情報科学科

郡 宏（こおり ひろし）

お相手：北畑裕之（千葉大・物理学科）

# ランキングの重要性

- ▶ 現代は情報があふれている
- ▶ その中から有用な情報を獲得したい！
  - 重要なウェブページ, 重要(危険?)人物
  - しらみつぶしはもちろん無理
- ▶ 重要度をランキングする技術が必要

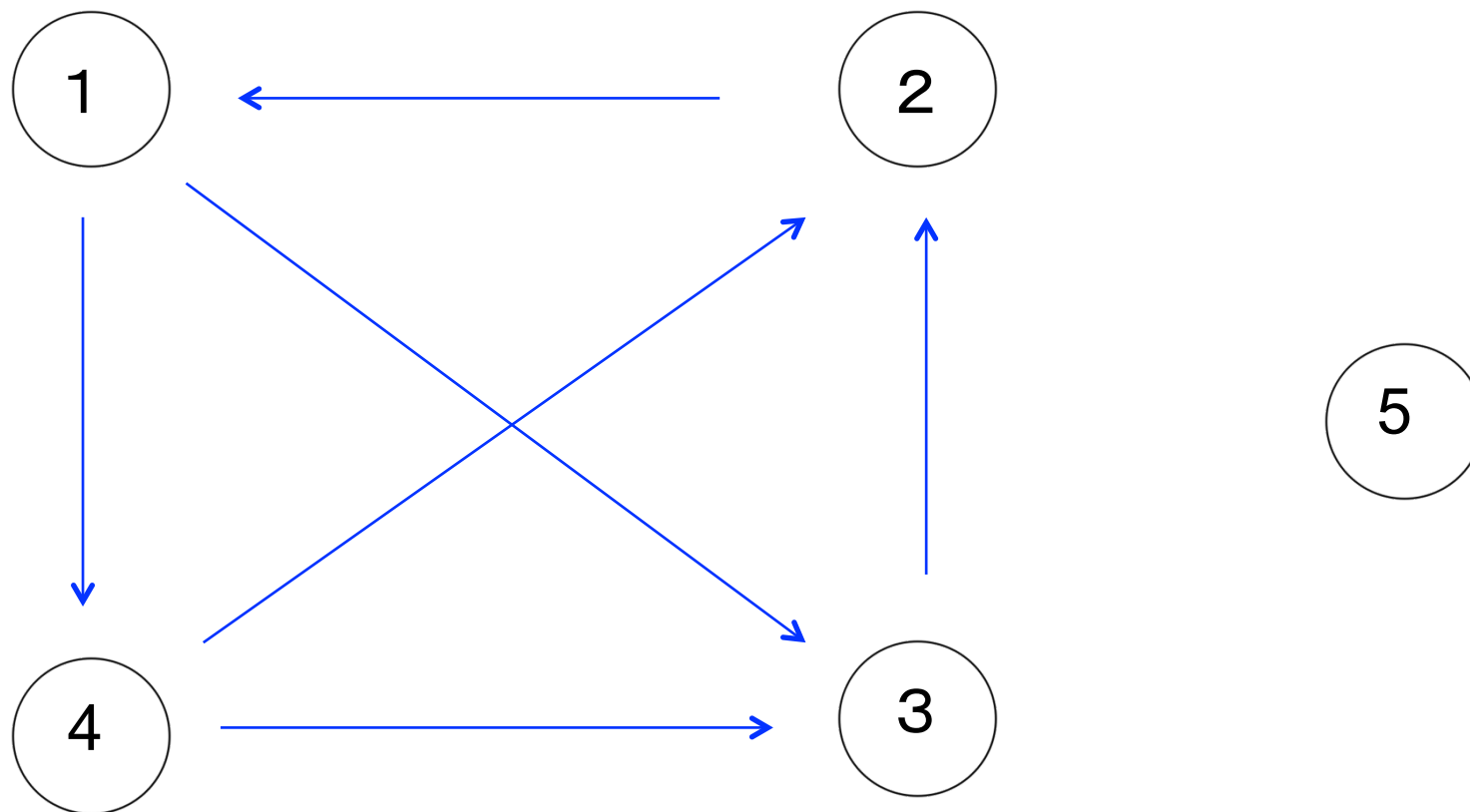
Googleの革新:

ページの内容ではなく,

ページとページのリンク関係に基づいて

ページの重要度をランキング

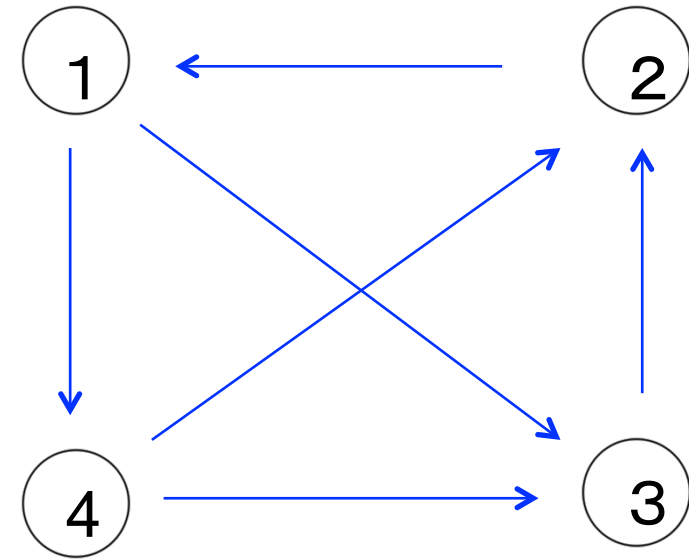
例題：以下のウェブページのネットワークで重要なページはどれ??



①, ②, ③, ④, ⑤: ウェブページ  
矢印: リンク

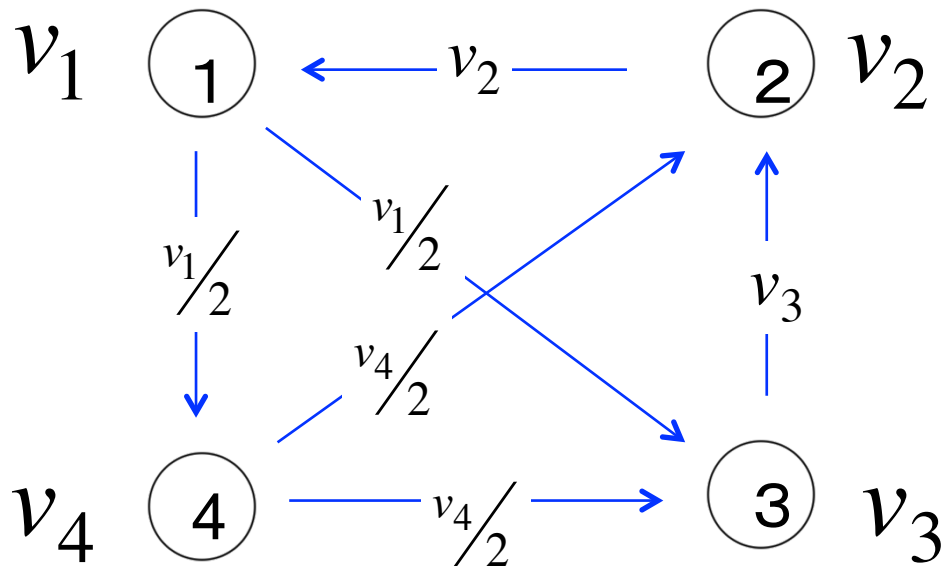
# 重要なページとは？

- ▶ 多くのページからリンクされている  
(多くの支持が大事)
- ▶ 良質なページからリンクされている  
(信頼できるページからの支持が大事)
- ▶ でも、むやみにリンクしているサイトからの  
リンクはあてにならない  
(厳選された支持が大事)



# ページをランキングしよう！

$v_i$  : ページの重要度



1. 矢印に沿って重要度を「送る」.
2. ただし, 複数のリンクがあれば重要度を等分して送る.
3. 受け取った重要度の和がそのページの重要度であるとする.

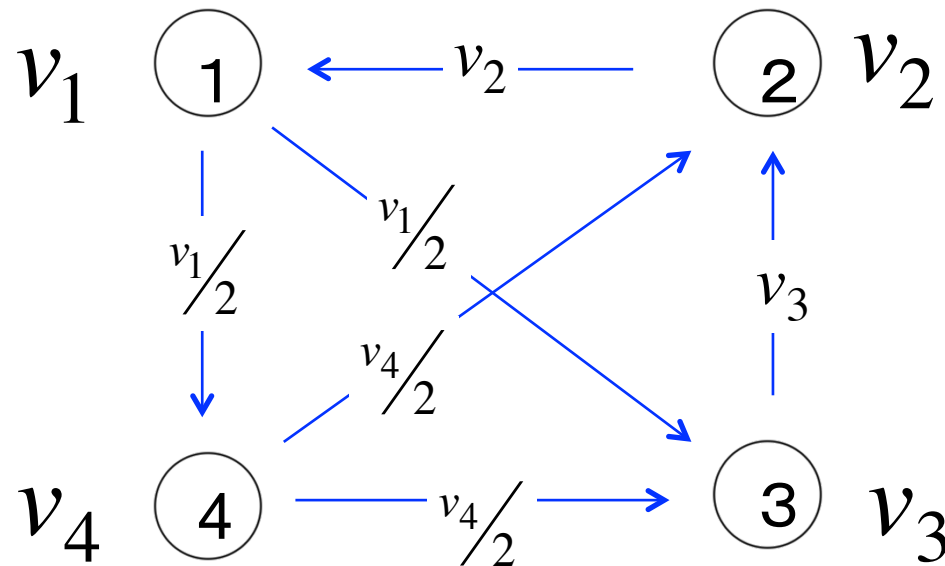
$$\begin{cases} v_1 = v_2 \\ v_2 = v_3 + v_4/2 \\ v_3 = v_1/2 + v_4/2 \\ v_4 = v_1/2 \end{cases}$$

単なる連立方程式. 解ける?!

$$v_1 : v_2 : v_3 : v_4 = 4 : 4 : 3 : 2$$

# ページをランキングしよう！

$v_i$  : ページの重要度 (総和を1とする)



$$v_1 : v_2 : v_3 : v_4 = 4 : 4 : 3 : 2$$

$$(v_1, v_2, v_3, v_4) = \left( \frac{4}{13}, \frac{4}{13}, \frac{3}{13}, \frac{2}{13} \right)$$

重要度が求まり, ランキングできた!

この数値は PageRank とよばれる. ページの内容を見ない!  
これがネットの検索技術のイノベーションをもたらした.

ちなみにPageさんはgoogleの創始者の一人.

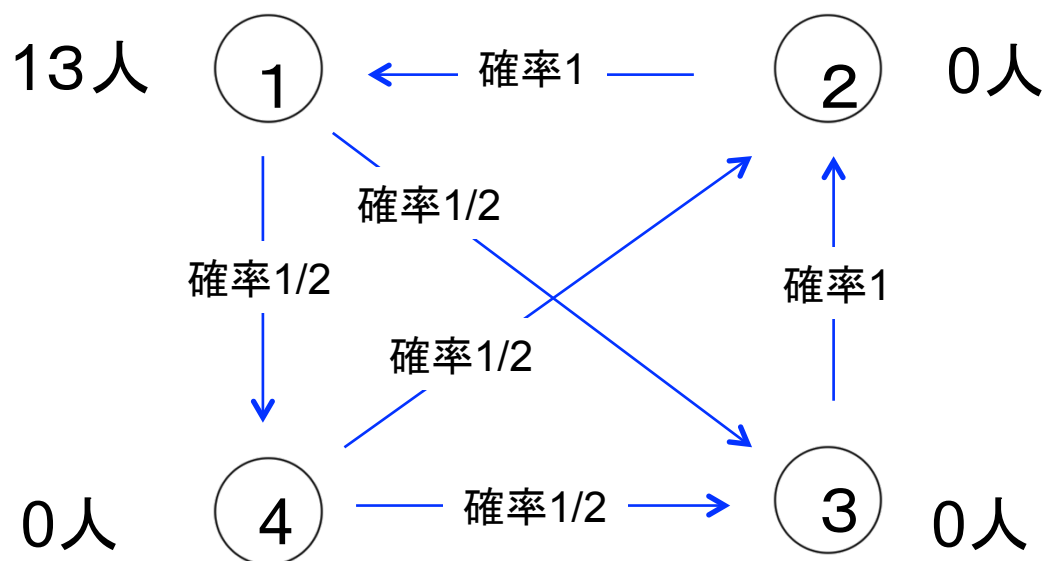
# 疑問

- ▶ ページランクは、本当にいい指標？
- ▶ どういった意味で妥当な指標？

# PageRankは本当にいいランキング法？

- ▶ ぼーっとウェブ・サーフィンしてみよう

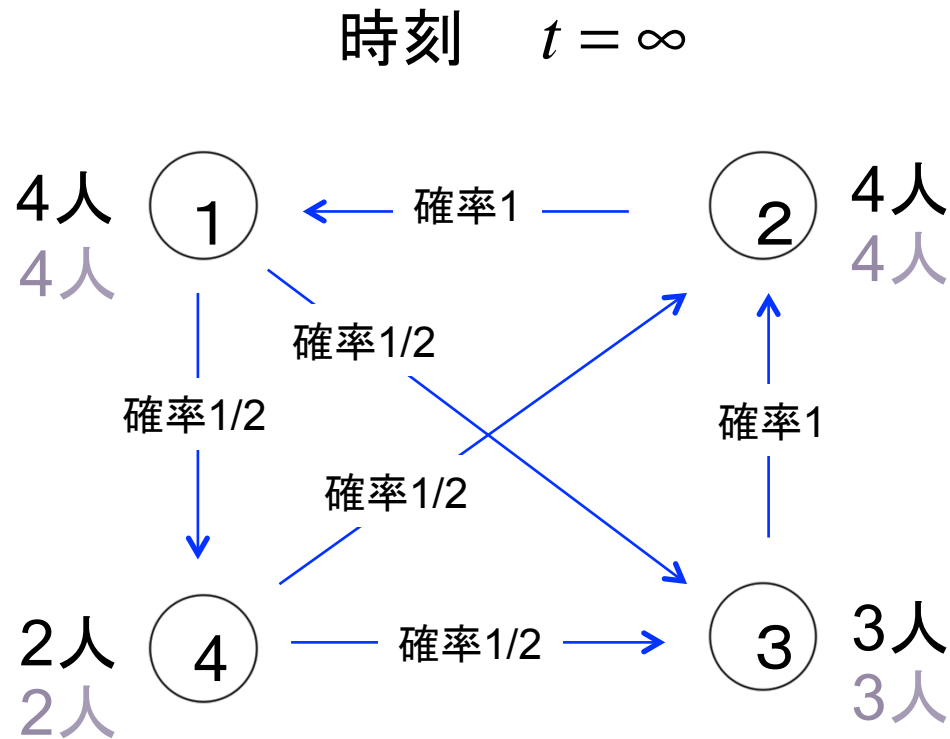
時刻  $t = 1$





# PageRankは本当にいいランキング法？

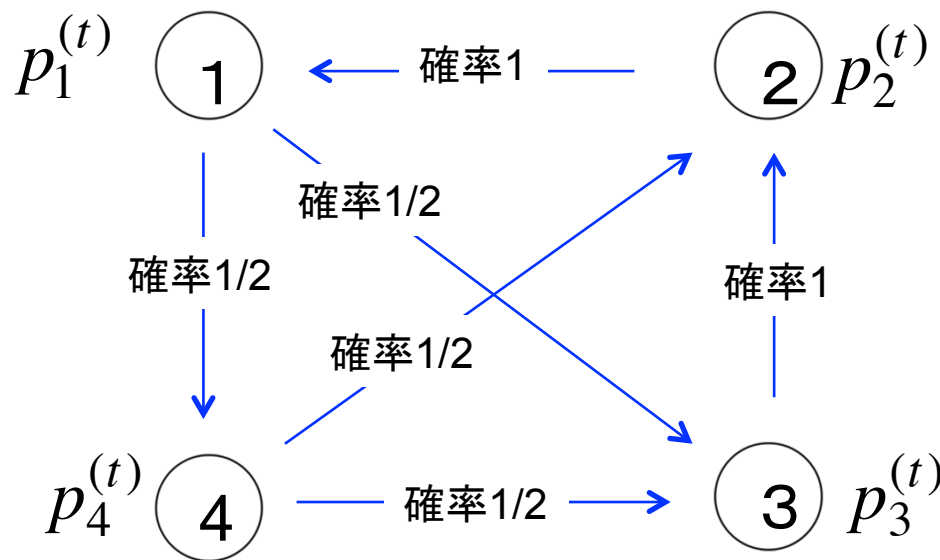
- ▶ ぼーっとウェブ・サーフィンしてみよう



繰り返していくと  
一つ前の時刻も今も  
同じ人数分布になる。  
(収束)

# PageRankは本当にいいランキング法？

$p_i^{(t)}$  : 時刻tにサイトiを訪ねている人数



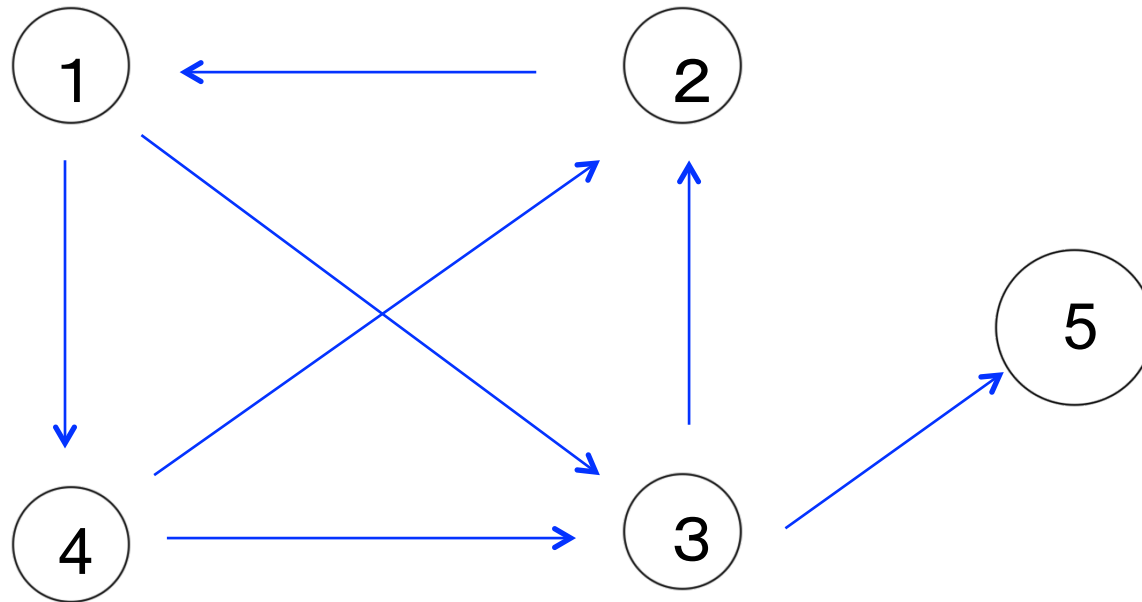
$$\begin{cases} p_1^{(t)} = p_2^{(t-1)} \\ p_2^{(t)} = p_3^{(t-1)} + \frac{1}{2} p_4^{(t-1)} \\ p_3^{(t)} = \frac{1}{2} p_1^{(t-1)} + \frac{1}{2} p_4^{(t-1)} \\ p_4^{(t)} = \frac{1}{2} p_1^{(t-1)} \end{cases}$$

連立漸化式

収束, つまり  $p_i^{(t)} = p_i^{(t-1)}$  とすると, さきほどのvの式とまったく同じ!!

ページランクはネットワーク上で「ランダムウオーク」をしたときの, 各ページの訪問確率を表している.

もし行き止まりがあったら？



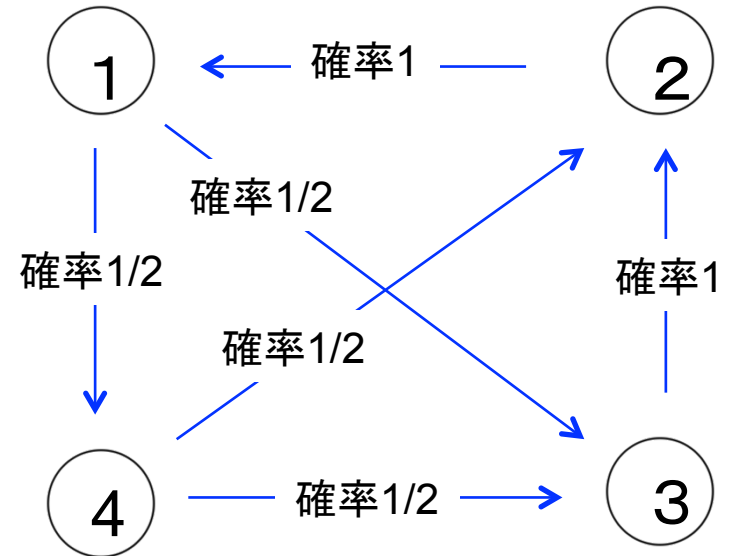
重要度  $(v_1, v_2, v_3, v_4, v_5) = (0, 0, 0, 0, 1)$

行き止まりが独り占めしてしまう！ こまった.

ページランクはこれを防ぐため「テレポーション」を導入.

# 大学の数学との関係

まず、ネットワークを  
行列をつかって表現  
(線形代数)



つながりを以下のように表現

$$A = \begin{matrix} & \textcircled{1} & \textcircled{2} & \textcircled{3} & \textcircled{4} \\ \textcircled{1} & 0 & 1 & 0 & 0 \\ \textcircled{2} & 0 & 0 & 1 & 1 \\ \textcircled{3} & 1 & 0 & 0 & 1 \\ \textcircled{4} & 1 & 0 & 0 & 0 \end{matrix}$$

「隣接行列」といいます  
(グラフ理論)

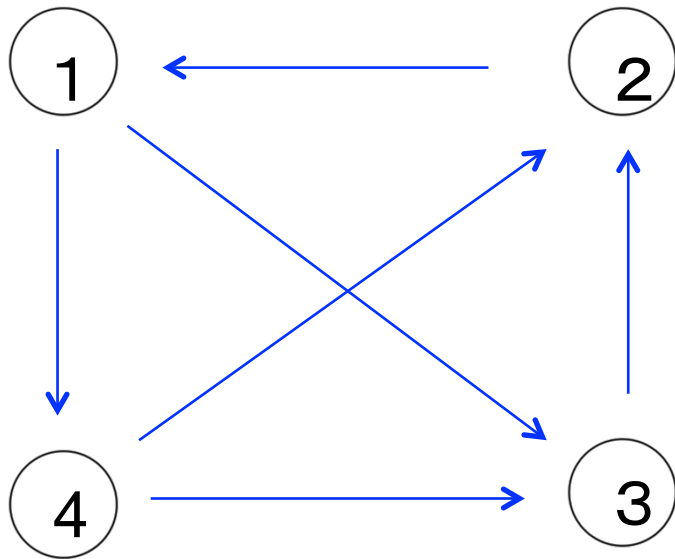
各列の総和を1にする

$$B = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 \end{pmatrix}$$

「確率行列」といいます  
(確率過程)

# 大学の数学との関係: グラフ理論

まず, ネットワークを行列をつかって表現  
(線形代数)

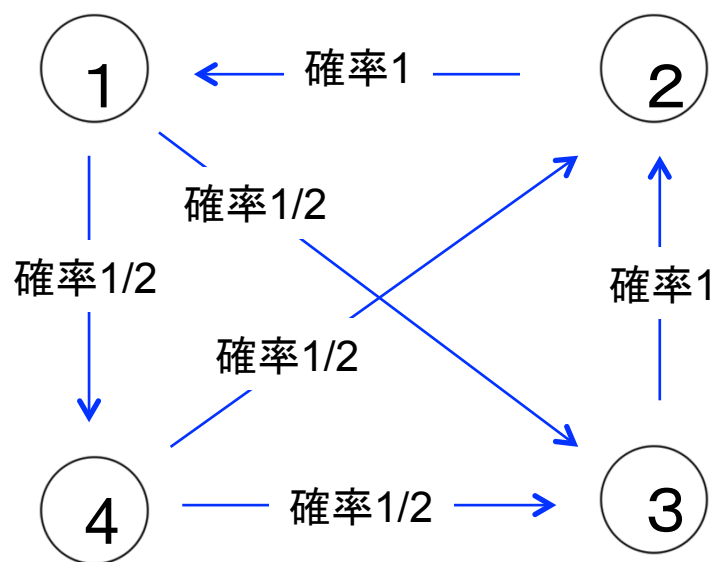


つながりを以下のように表現

$$\begin{array}{c} \textcircled{1} \quad \textcircled{2} \quad \textcircled{3} \quad \textcircled{4} \\ \textcircled{1} \\ \textcircled{2} \\ \textcircled{3} \\ \textcircled{4} \end{array} \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} = A$$

「隣接行列」といいます  
(グラフ理論)

# 大学の数学との関係:確率論



各列の総和を1にする

$$B = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 \end{pmatrix}$$

「遷移確率行列」といいます

# 大学の数学との関係：線形代数

ページランクをベクトルで  
つけて表現

$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix}$$

ページランクを求めた式を次の形式で書き表す

$$\begin{cases} v_1 = v_2 \\ v_2 = v_3 + v_4/2 \\ v_3 = v_1/2 + v_4/2 \\ v_4 = v_1/2 \end{cases} \Rightarrow \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix}$$

$$\vec{v} = B\vec{v} \quad \text{「固有方程式」といいます}$$

# 大学の数学との関係:ランダムウォーク

$$\begin{cases} p_1^{(t)} = p_2^{(t-1)} \\ p_2^{(t)} = p_3^{(t-1)} + \frac{1}{2} p_4^{(t-1)} \\ p_3^{(t)} = \frac{1}{2} p_1^{(t-1)} + \frac{1}{2} p_4^{(t-1)} \\ p_4^{(t)} = \frac{1}{2} p_1^{(t-1)} \end{cases} \quad \begin{pmatrix} p_1^{(t)} \\ p_2^{(t)} \\ p_3^{(t)} \\ p_4^{(t)} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} p_1^{(t-1)} \\ p_2^{(t-1)} \\ p_3^{(t-1)} \\ p_4^{(t-1)} \end{pmatrix}$$

$$\vec{p}^{(t)} = B\vec{p}^{(t-1)} = B^2\vec{p}^{(t-2)} = \dots = B^t\vec{p}^{(0)}$$

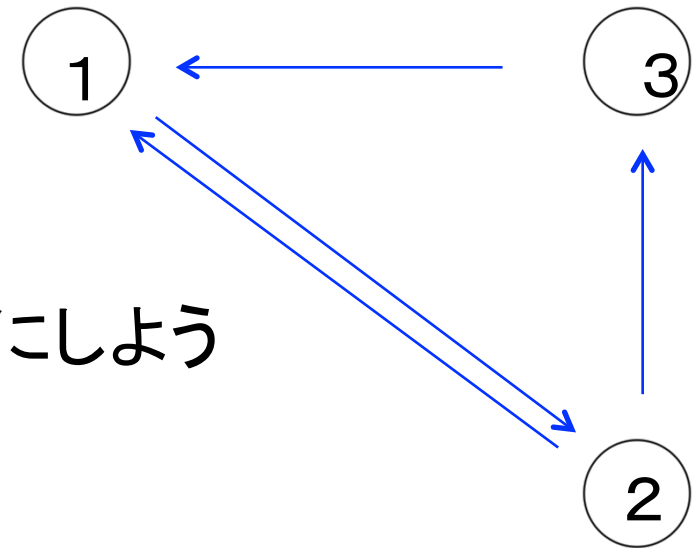
初期ベクトルを適当に選び、tを十分大きな数にすれば、PageRankが近似的に求まる。

実際Googleはこの方法をつかってPageRankを求めている(t=50程度らしい)。ちなみにページ数は10兆個を超える！！



# 課題

- (1) 何かのネットワークを考えて図にしよう
- (2) 遷移確率を書き込もう
- (3) ページランクを算出しよう
- (4) 結果について考察しよう



# まとめ

- ▶ ページランク: 膨大な情報があふれる世界から、重要な情報を探し出す数学的技術
- ▶ 今はビックデータの時代。ページランク以外にも、さまざまな数学的手法が、重要情報の探索に用いられている。
- ▶ 線形代数, 確率論, グラフ理論などの, 大学で学ぶたのしーい数学がいろいろと関係しています。  
うわー、大学に行くのが楽しみだー！！