中心性と Google の PageRank

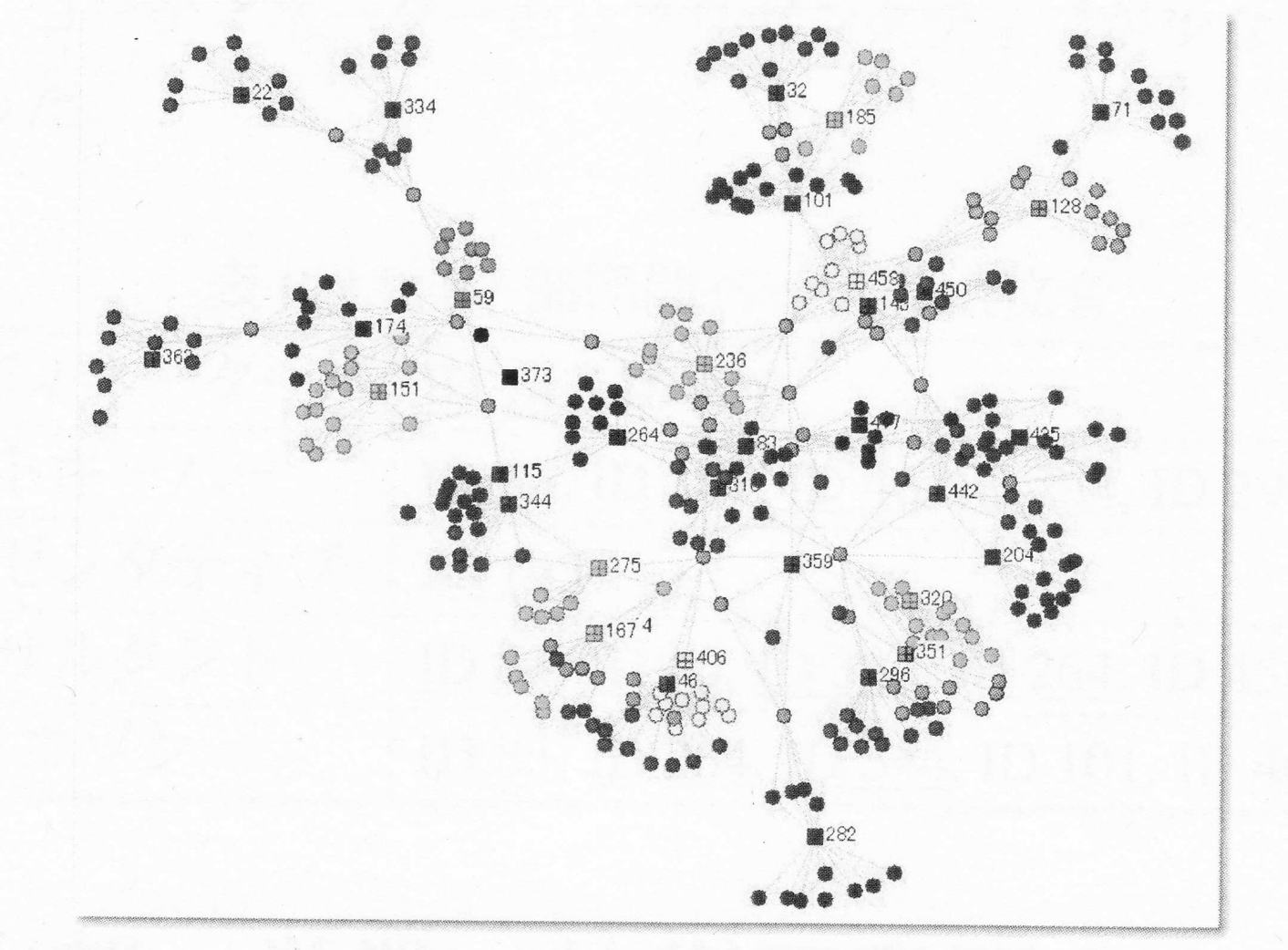
林　幸雄

北陸先端科学技術大学院大学

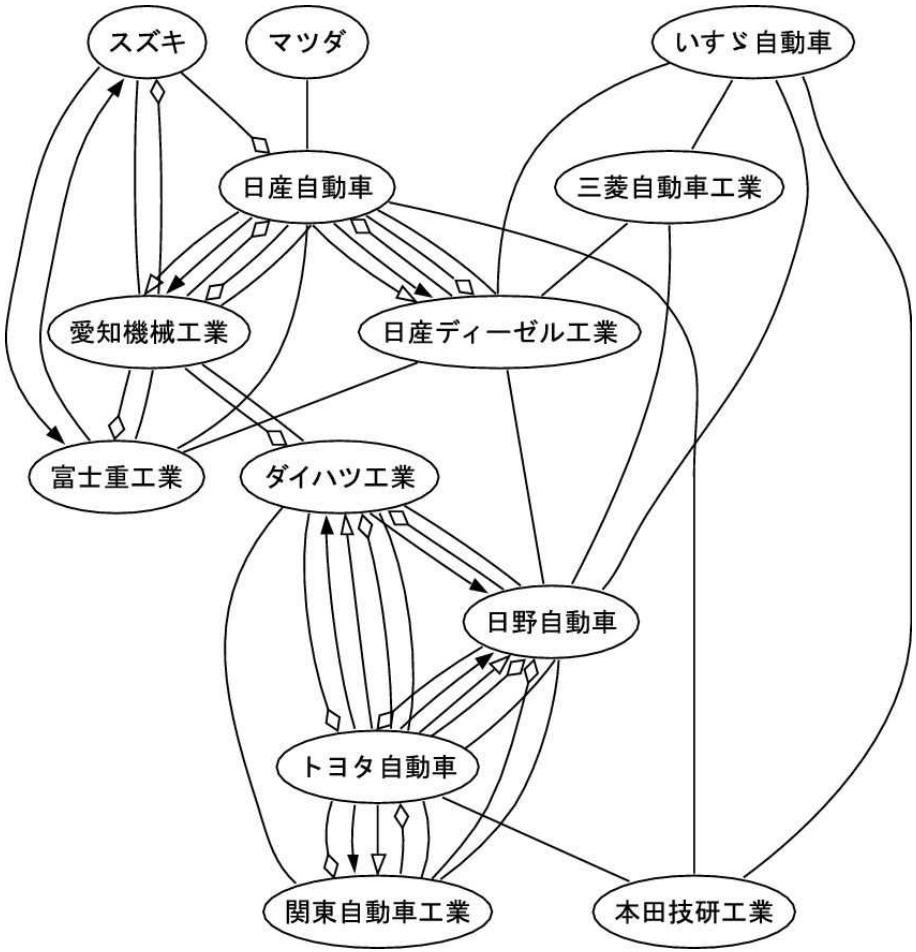
改訂第 2 版 2019

1. 中心人物は誰だ

営業, 研究開発, 本社スタッフ等の異なる部門組織に属する人々の繋がり



自動車産業における多重ネット



黒矢印：株所有, 白矢印：役員の兼任や前歴, 白菱形：取引の関係

林編著　ネットワーク科学の道具箱, 第 6 章, 近代科学社, 2007

2. 代表的な中心性指標

中心的な人物や企業を数量化して抽出する為の分析手法

次数中心性 (Degree Centrality): ノード i の次数 k_i に比例した $k_i/(N-1)$ で定義される. 分母の $N-1$ は, ノード i が自分以外の他の全ての $N-1$ 個のノードと結合する最大次数の場合で正規化する為. これは, 各ノード 自身からのエゴセントリックな局所的な影響力を表す.

情報中心性 (Information Centrality): ノード i と j をつなぐ複数の経路に伝達したい情報が (ホップ数で測った) その経路長の逆数に比例して分散して流れるとした情報量の ▸ 調和平均値 で定義される.

代表的な中心性指標：その2

近接中心性 (Closeness Centrality)：　ノード i から j に到達する経路上で最小のホップ数:仲介数+1 を H_{ij} として，

$$\left(\frac{\sum_j H_{ij}}{N-1}\right)^{-1} = \frac{N-1}{\sum_j H_{ij}}.$$

で定義される．分子の $N-1$ は星形ネットワークの場合の最小値で正規化する為．この指標は，あるノードから他の全てのノードに到達するのに要する広がりの伝搬時間の和に関する量と考えられる．

フロー中心性 (Flow Centrality)：　グラフの最大フローがあるノード i を媒介する頻度で定義される．

代表的な中心性指標：その3

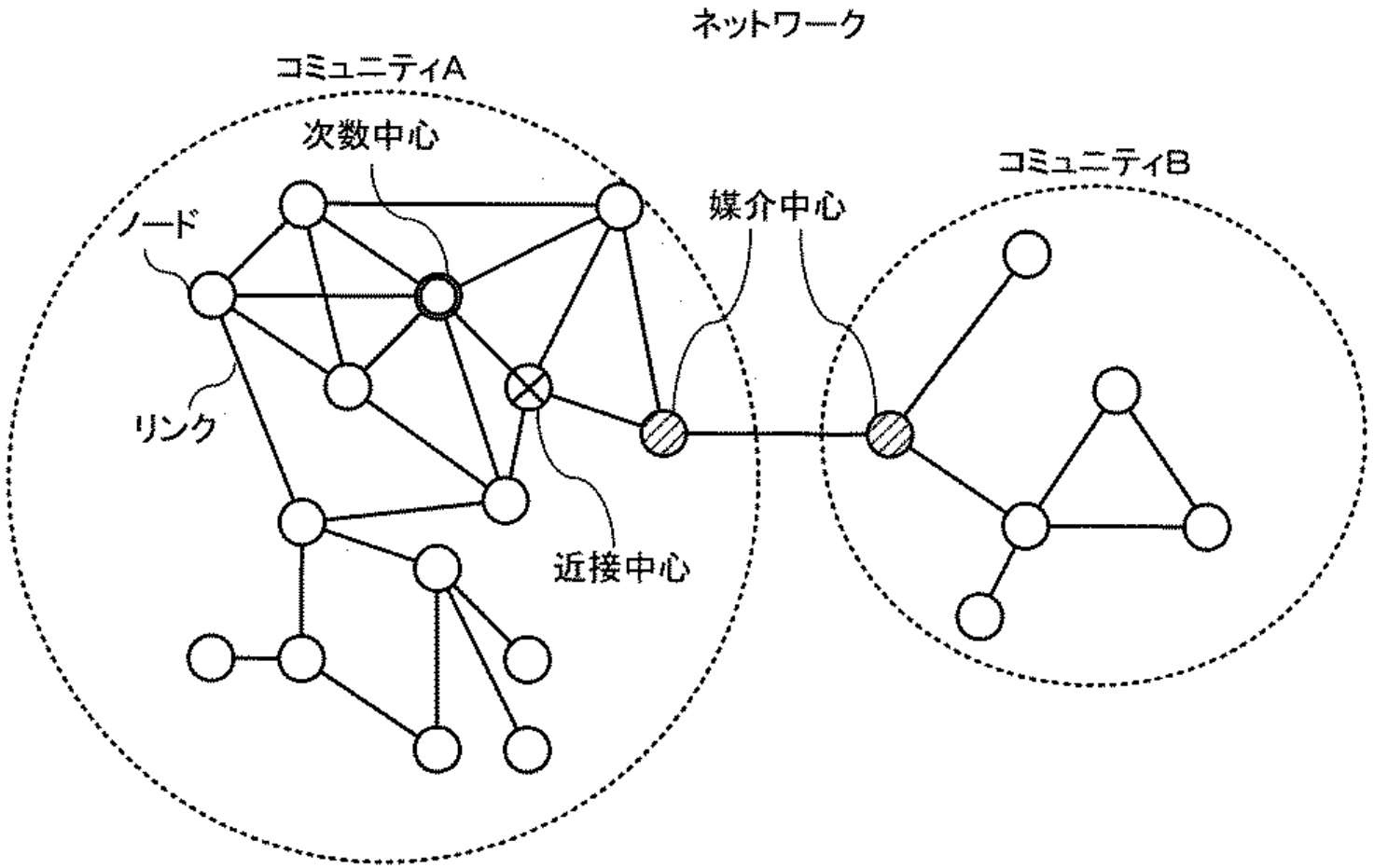
媒介中心性 (betweenness centrality): ノード v の媒介中心性 $\sum_{s,t} \delta_{st}(v)$, σ_{st} は始終点ペア (s,t) 間の最短経路数, $\sigma_{st}(v)$ はそのうち v を通る数

$$\delta_{st}(v) \stackrel{\text{def}}{=} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

L.C. Freeman, Sociometry 40, 1977

http://www.geocities.jp/woodone3831/kanntou/c-4-11-sekisyo-HAKONE.html

BC は橋渡し役を重視



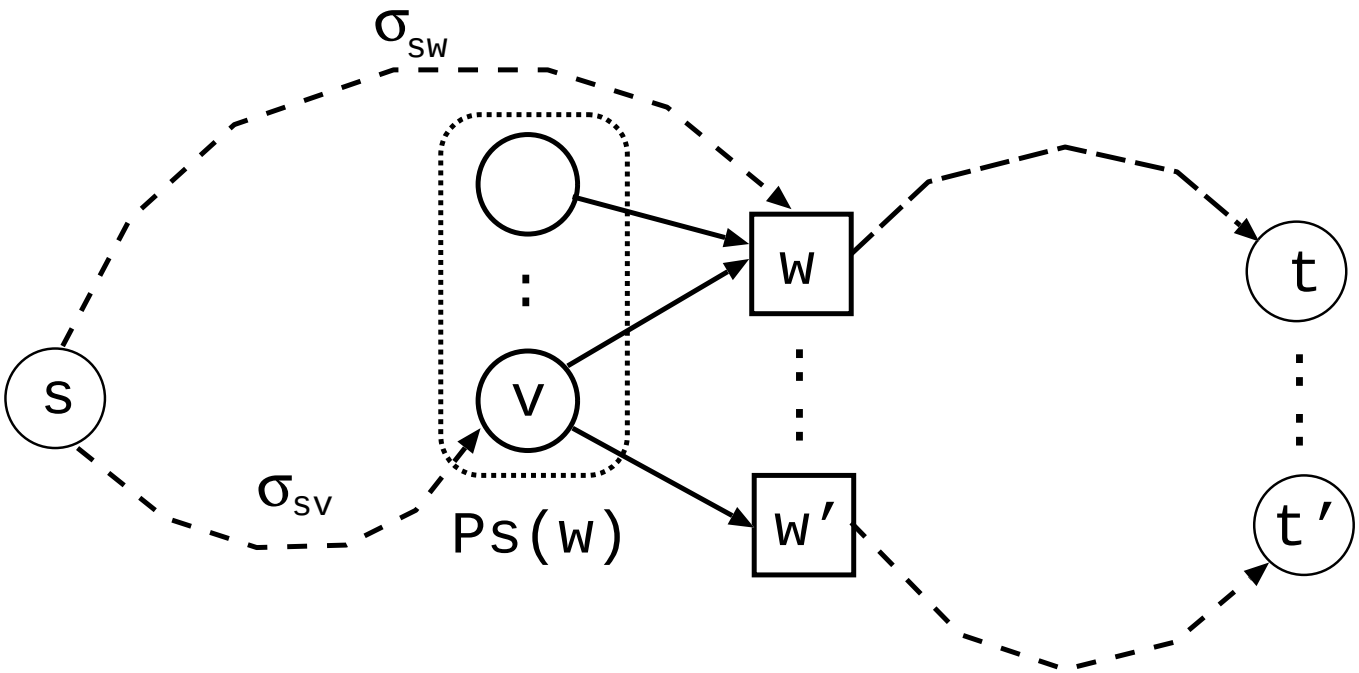
http://astamuse.com/ja/published/JP/No/2010141442

Brandesの高速BCアルゴリズム

始点 s から各ノード u への最短経路数 σ_{su} を全て求めた後に，ホップ数で一番遠いノードから順に $\delta_{s,\bullet}(w)$ を確定していく
$w \sim w'$ が終点 t の時と，$w \sim w'$ を経由した先に t がある時の和

$$\delta_{s,\bullet}(v) = \sum_{\{w|v \in P_s(w)\}} \frac{\sigma_{sv}}{\sigma_{sw}}(1 + \delta_{s,\bullet}(w)),$$

$$P_s(w) \stackrel{\text{def}}{=} \{v \in V | (v,w) \in E, d(s,w) = d(s,v) + 1\},\ w \in \partial v$$



U.Brandes, Journal of Math. Sociology 25, 2001

ルーティング中心性への拡張

ノードの管轄領域内の人口量を考慮するなど, ノード間の通信要求や輸送要求の発生頻度の重みを付与して拡張

$$\delta_{st}(v) = \sum_{u \in Pred_{s,t}(v)} \delta_{st}(u) \times R(s,u,v,t),$$

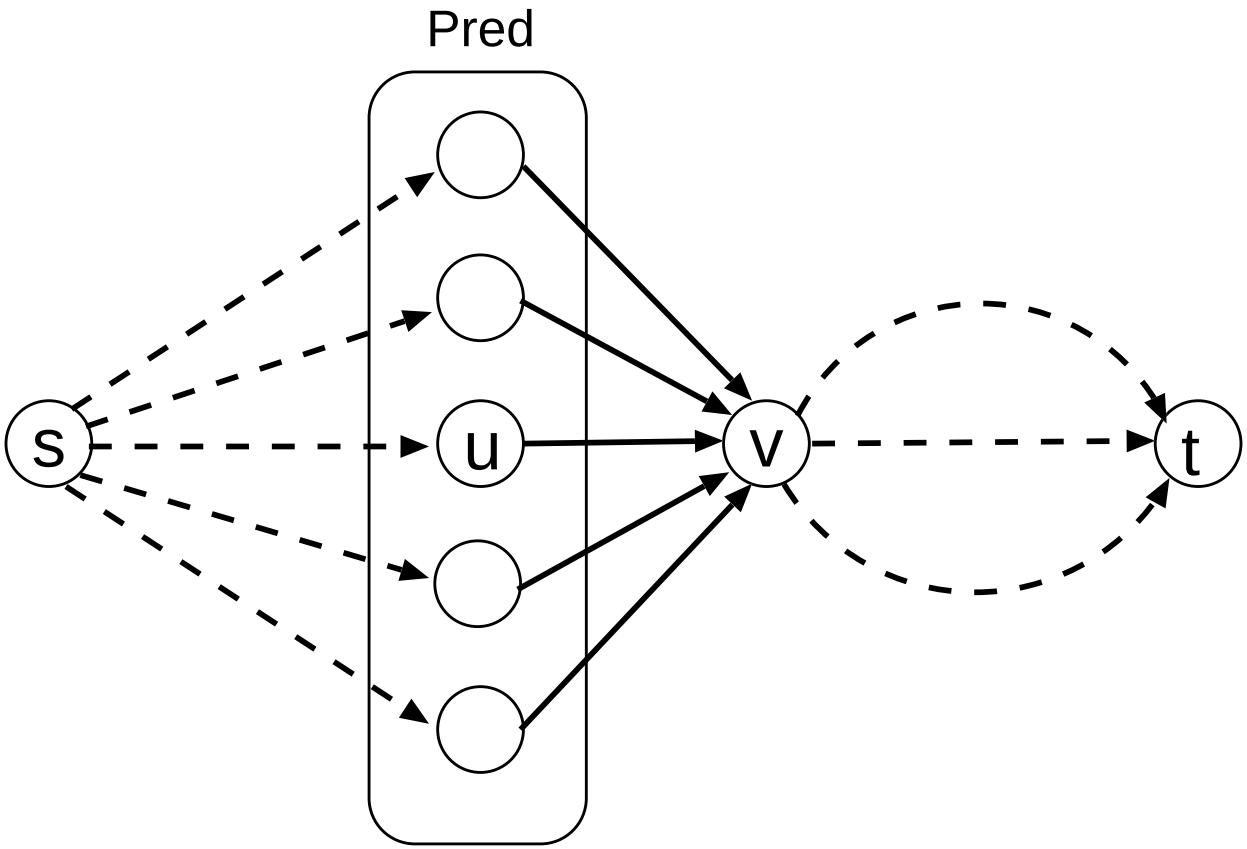
$Pred_{s,t}(v)$ は s-t 間の経路上でノード v の 1 つ前に経由するノードの集合 $\{u\}$, $R(s,u,v,t)$ は s から t への経路がリンク (u,v) を経由する確率で, ルーティング選択に依存.
s-t 間の送受信要求の発生頻度の期待値を $T(s,t)$ と表記して, ノード $v \in V$ の媒介中心性が求められる.

$$\delta_{\bullet,\bullet}(v) = \sum_{s,t \in V} \delta_{st}(v) \times T(s,t).$$

S.Dolev et al., Journal of the ACM 57, 2010

$R(s, u, v, t)$と $T(s, t)$

始点 s から辺 u-v を経由して終点 t に至る経路



通信要求の発生頻度の重み $T(s, t)$ は，例えば ▸ グラビティ モデル で定義

代表的な中心性指標：その4

ボナチッチ中心性 (Bonacichi Centrality)：　中心性が高いノードと数多く結合すれば自身の中心性も高くなるとして定義.

$$\mathbf{x} = \alpha A\mathbf{1} + \beta A\mathbf{x},$$

を $\mathbf{x}$ について解くと, $\frac{1}{1-z} = 1 + z + z^2 + z^3 + \ldots$ に気づけば

$$\begin{aligned}\mathbf{x} = (I - \beta A)^{-1}(\alpha A\mathbf{1}) &= \alpha \sum_{k=0}^{\infty} \beta^k A^{k+1}\mathbf{1}, \\ &= \alpha(A + \beta A^2 + \beta^2 A^3 + \ldots)\mathbf{1}.\end{aligned}$$

$\Rightarrow$ 各ノード i から j へ, 1 ホップの直の隣接, 2 ホップ経由先, 3 ホップ経由先,　$\ldots$ の（減衰係数 β 付きで）影響力の和を表す.
$\alpha = 0$, $\beta = 1/\lambda$ の時, 隣接行列 $[a_{ij}]$ に対する最大固有値の優固有ベクトルに帰着.

注意点

他にも, ノードに入る方だけでなく出る方の分配の影響も考慮して, 隣接行列の代りに各 i 行ベクトルの要素 a_{ij} を $1/k_i$ に置き換えた確率遷移行列を考えると, その優固有ベクトルはネットワーク上を乱歩した時のノードの滞在頻度と一致し, PageRank 値に対応.

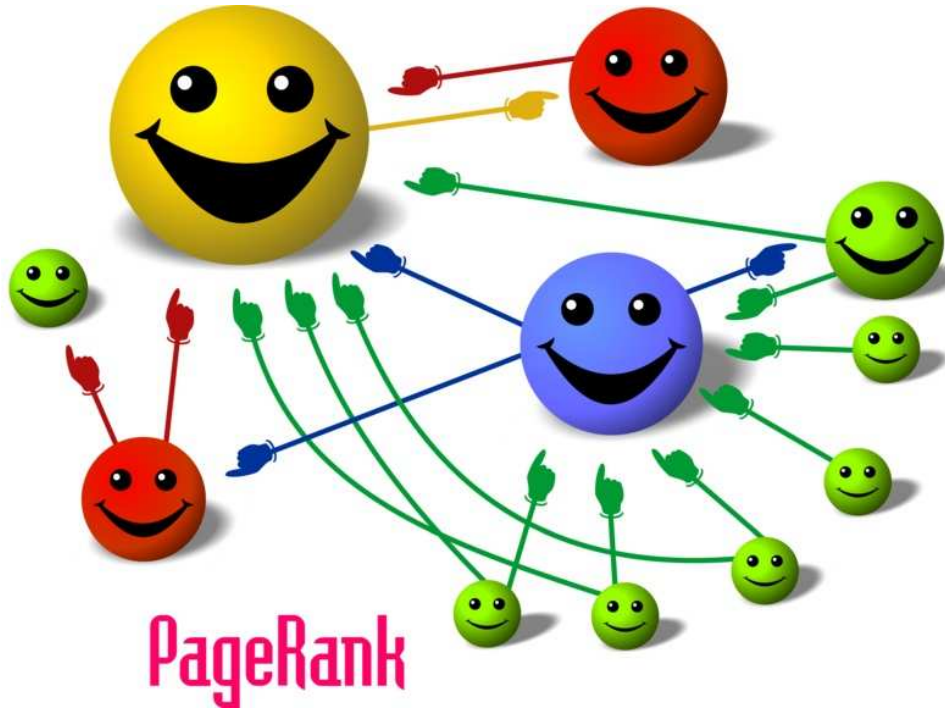
他にも, 権威付けの地位の高い人と結合するほど値が高くなる Katz 中心性や Hubbel 中心性など.

⇒ 但し, それぞれの指標の特徴や応用面での分析目的に応じて使い分ける必要がある.

3. PageRank 中心性

WWW の頁における中心性を測る指標として

- 入: 多くから参照されるリンクを持つ頁の価値は高くなる
- 入: ランク値が高い頁からの数少ない貴重なリンクの価値は高い
- 出 : 多くを参照するリンクの価値は分配される



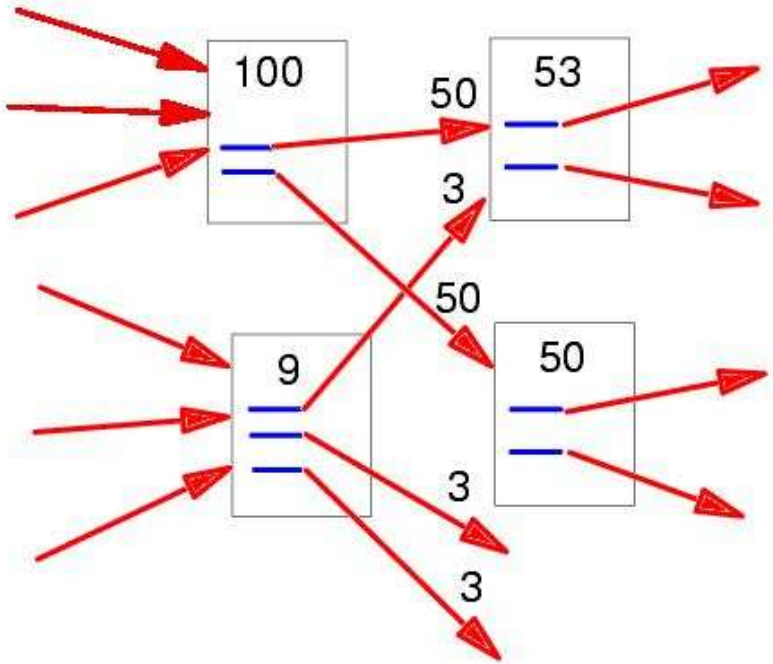
直感的で割と自然な定義に感じられるが, 理論的な裏付けがある!

メッセージ伝搬に基づく優れた計算法

基本形：$\mathbf{r} \leftarrow P\mathbf{r}$, $\mathbf{r}$ は▸確率行列 Pの優固有ベクトルに収束

$$r_v \leftarrow d \sum_{v' \in \mathcal{N}_v} \frac{r_{v'}}{k_{v'}} + \frac{1-d}{N}, \quad d \approx 0.85,$$

割合 $1-d$ だけランダムサーフを入れて強連結にする $\Rightarrow$ マルコフ連鎖



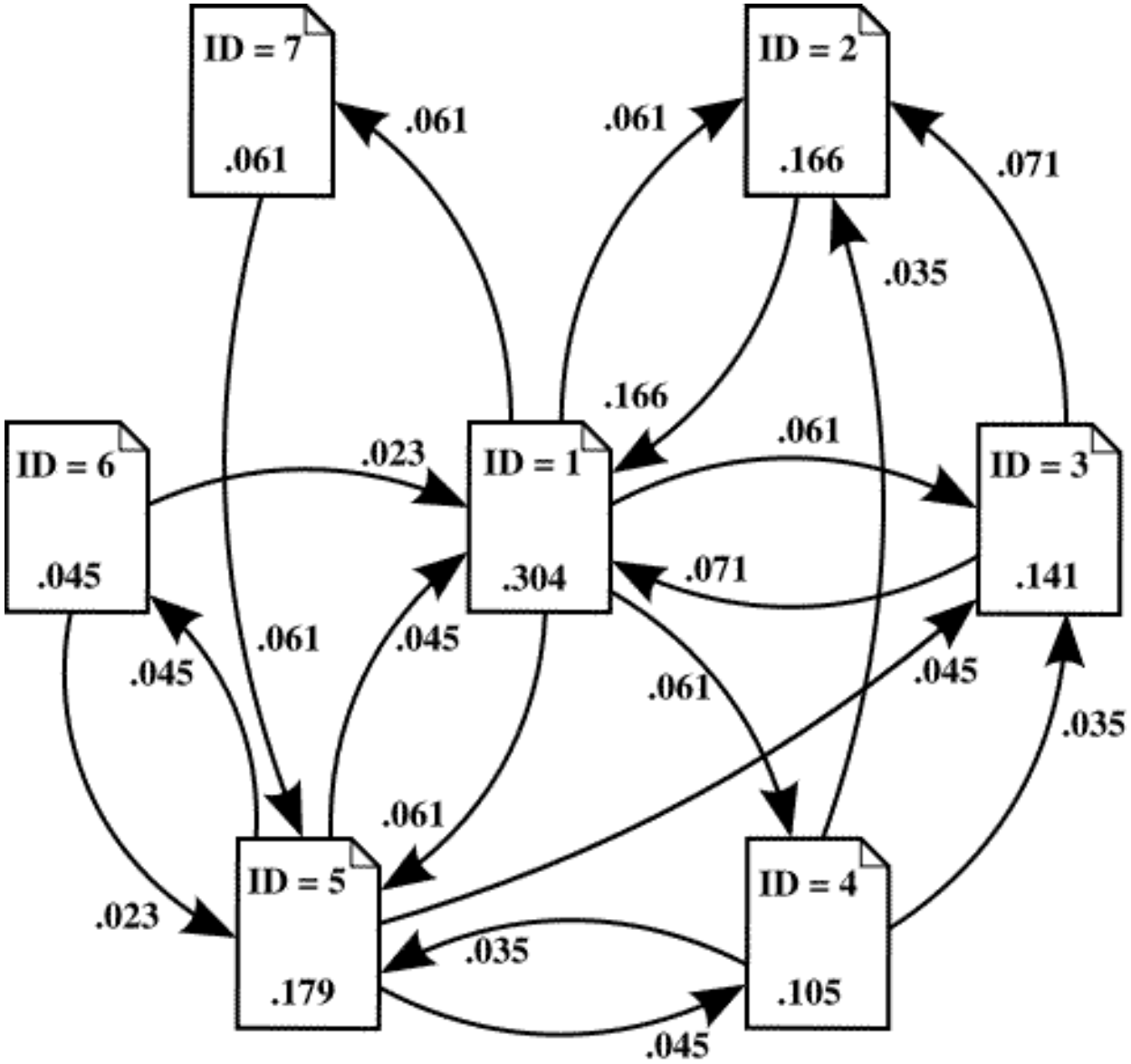
Google行列の優固有ベクトル

ノード i の滞在確率 $r_i \Leftrightarrow \pi_i$, $\sum_i \pi_i = 1$ に注意して, 更新式 $\mathbf{r} \leftarrow d \times P\mathbf{r} + (1-d)/N$ を書き直すと

$$\begin{pmatrix} \pi_1 \\ \vdots \\ \pi_i \\ \vdots \\ \pi_j \\ \vdots \\ \pi_n \end{pmatrix}^T \begin{pmatrix} \frac{1-d}{N} & \cdots & \cdots & \cdots & \cdots & \cdots & \frac{1-d}{N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \cdots & \frac{1-d}{N} & \cdots & \frac{d}{k_i} + \frac{1-d}{N} & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \cdots & \frac{1-d}{N} & \cdots & \frac{1-d}{N} & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1-d}{N} & \cdots & \cdots & \cdots & \cdots & \cdots & \frac{1-d}{N} \end{pmatrix} = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_i \\ \vdots \\ \pi_j \\ \vdots \\ \pi_n \end{pmatrix}^T$$

A.N.Langville and C.D.Meyer, Google's Page Rank and Beyond, Princeton Univ. Press, 2006

数学的意味：WWW 上の乱歩との等価性



固有ベクトル中心性として

$$\pi_i = d\sum_j A_{ij}\frac{\pi_j}{k_j^{out}} + (1-d)\beta\sum_j \pi_j.$$

$$\pi = dAD^{-1}\pi + (1-d)\beta\mathbf{1} \quad \rightarrow \quad \pi = (1-d)\beta(I - dAD^{-1})^{-1}\mathbf{1}.$$

Katz 中心性

$$x_i = \alpha\sum_j A_{ij}x_i + \beta'.$$

$$\mathbf{x} = \alpha A\mathbf{x} + \beta'\mathbf{1} \quad \rightarrow \quad \mathbf{x} = \beta'(I - \alpha A)^{-1}\mathbf{1} = \beta'\sum_{k=0}^{\infty}(\alpha A)^k\mathbf{1}.$$

隣接行列 A の優固有ベクトル中心性: $\mathbf{x} = \frac{1}{\lambda_1}A\mathbf{x}$
次数中心性: 無向グラフ上の乱歩 $\neq$ PageRank とは似て非なるもの!

$$x_i(t) = \sum_j \frac{A_{ij}}{k_j}x_j(t-1), \quad \text{平衡解: } x_i = \frac{k_i}{\sum_j k_j}.$$

$$\mathbf{x} = AD^{-1}\mathbf{x} \rightarrow (I - AD^{-1})\mathbf{x} = (D-A)D^{-1}\mathbf{x} = \mathbf{0}, D^{-1}\mathbf{x} = \mathbf{1}.$$

M.E.J.Newman, Networks -An Introduction-, OXFORD Univ. Press, 2010.

大規模行列の固有値の数値解法

長い研究の蓄積もあるが, スケーラビリティは特に問題となる

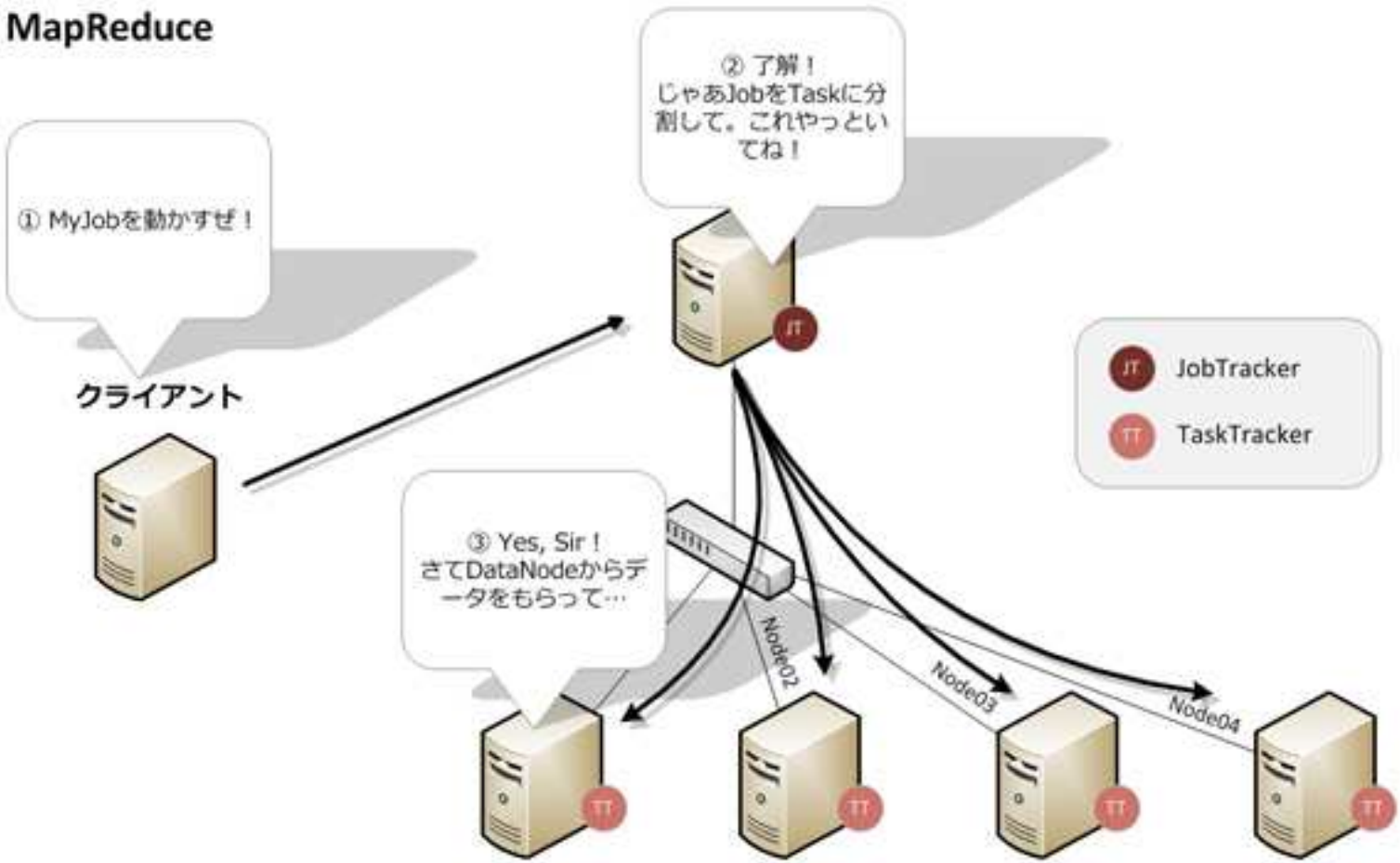
- べき乗法(基本形, 逆反復, シフト , 減次)
- 同時反復法
- Gram-Schmidt の直交化や Householder 変換により 上三角行列 R と 直交行列 Q に分解した QR 法
- Rayleigh-Ritz の技法
- Krylov 列： x, Ax, A^2x, A^3x, . . . を用いた Arnoldi 法, Lanczos 法
- 対称行列専用の二分法, Jacobi 法

森, 杉原, 室田　著, 線形計算, 岩波講座　応用数学 [方法 2], 1994, 岩波書店

Googleの創設者L.E.Pageの先見性

- 誰も重視しなかったリンク構造に着目.
  直感的に妥当そうな意味付けだけでなく, 理論的な裏付けもある.
- リンクの方向性で行き止まりが発生し得るWWW探索において, 少量の割合のランダムサーフによるジャンプを考慮することでリンク構造に基づくランク値と計算の収束性を両立.
- 数十年の研究歴史がある大規模行列の固有値問題の数値解法としては扱わず, 日々更新されるWWWでは厳密解は不要な一方でスケーラビリティが求められることから, 自分らの得意な単純なメッセージ伝搬による分散処理として大量の廉価PCで実現.

Indexing 計算を 分散処理で

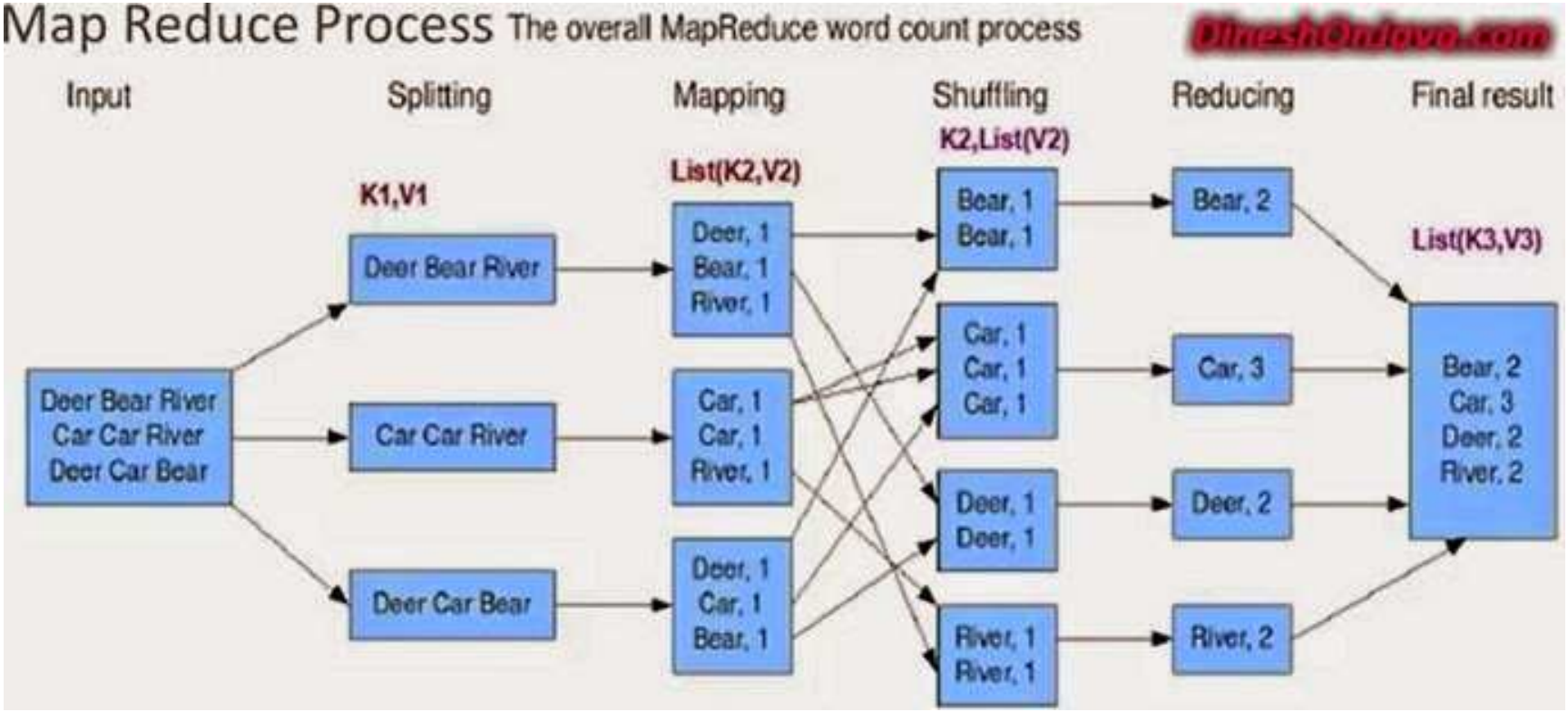


導入編～Hadoop クラスタを 構築するまで

https://enterprisezine.jp/dbonline/detail/4254

MapReduceミドルウエアの開発に

Map：各行で Key と Value の組, Reduce：Key 毎に集計



http://www.dineshonjava.com/2014/11/mapreduce-flow-chart-sample-example.html

⇒ ▸ より詳しい分散処理の技術動向

4. HITS：秀才 J.M.Kleinberg の敗北

HITS (Hyperlink-induced topic search)

Authority: 頻繁に参照される頁の度合い

$$\mathbf{x} \leftarrow A^T\mathbf{y} \leftarrow A^TA\mathbf{x},$$

$\mathbf{x}$ は A^TA の優固有ベクトルに収束

Hub: リンク集的に参照先を紹介する頁の度合い

$$\mathbf{y} \leftarrow A\mathbf{x} \leftarrow AA^T\mathbf{y}.$$

$\mathbf{y}$ は AA^T の優固有ベクトルに収束.

A はネットワークのノード間の接続関係を表す隣接行列, A^T はその転置行列.

J.M.Kleinberg, ▸ Journal of the ACM 46, 1999

5. 分散アルゴリズムの例

メッセージ伝搬による負荷均一化:
ジャンクションツリー アルゴリズム

- 葉から順に各ノード i は親ノードに負荷量 L_i を伝達, その和を上位の親ノードに転送.
- 根ノードに達すると総負荷量 $\sum_i L_i$ が求まる
  総ノード数 N も総負荷量と同様に伝搬で求まる.
- 均一負荷量 $\bar{L} \stackrel{\text{def}}{=} \sum_i L_i/N$ を根ノードから下位の各ノードに発信.
- 葉ノード i に $\bar{L}$ が届いたら, その親ノードに負荷 $\Delta L_i = L_i - \bar{L}$ 分を移動. $\Delta L_i < 0$ なら負荷をもらう.
  葉ノードから順次上位へ負荷移動を行う.

補足）根は未伝達ノードで動的に決めることも可, 但し一旦決まったら役割が付く