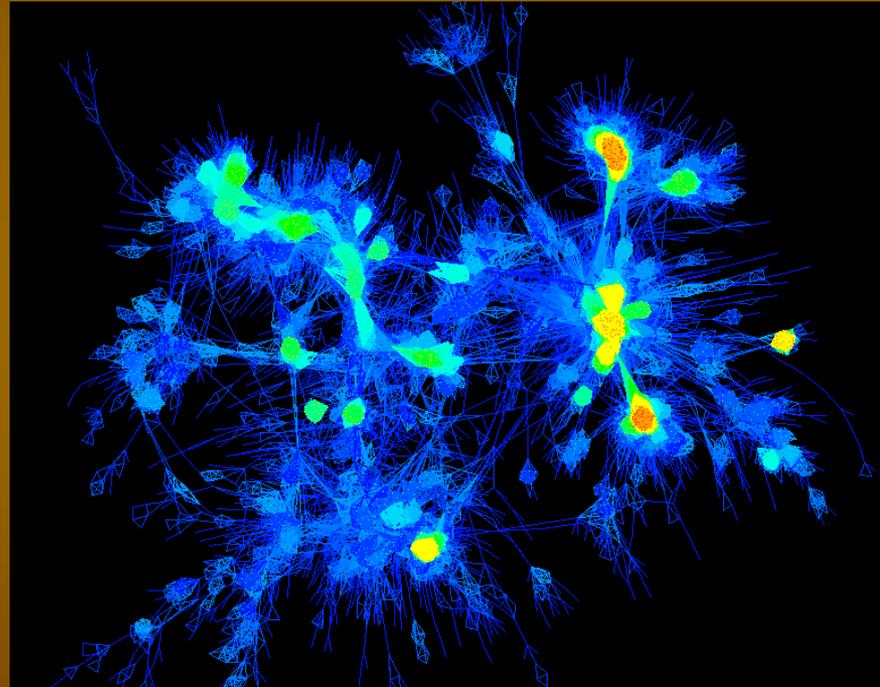


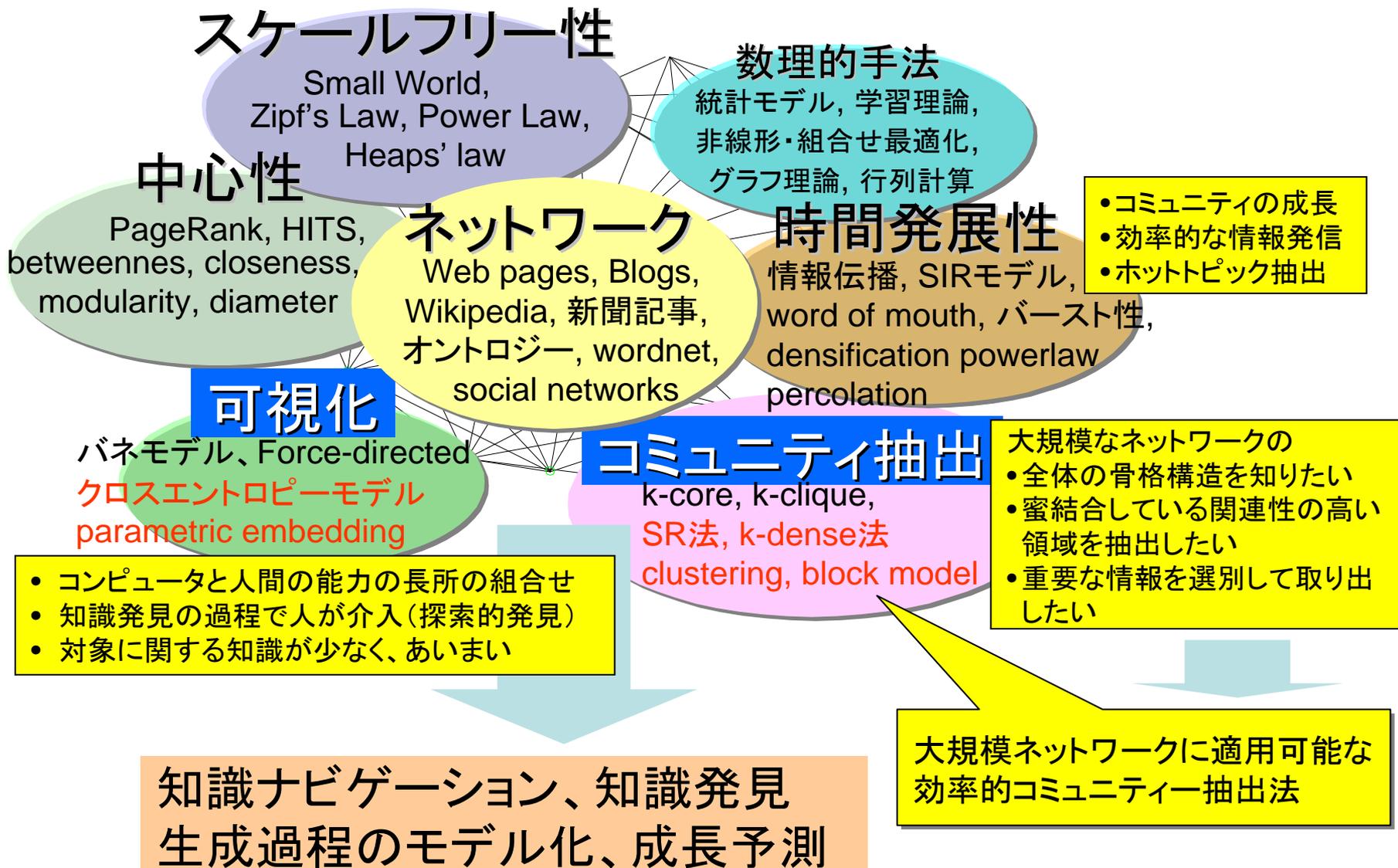
ネットワーク科学と 関係データマイニング



NTTコミュニケーション科学基礎研究所

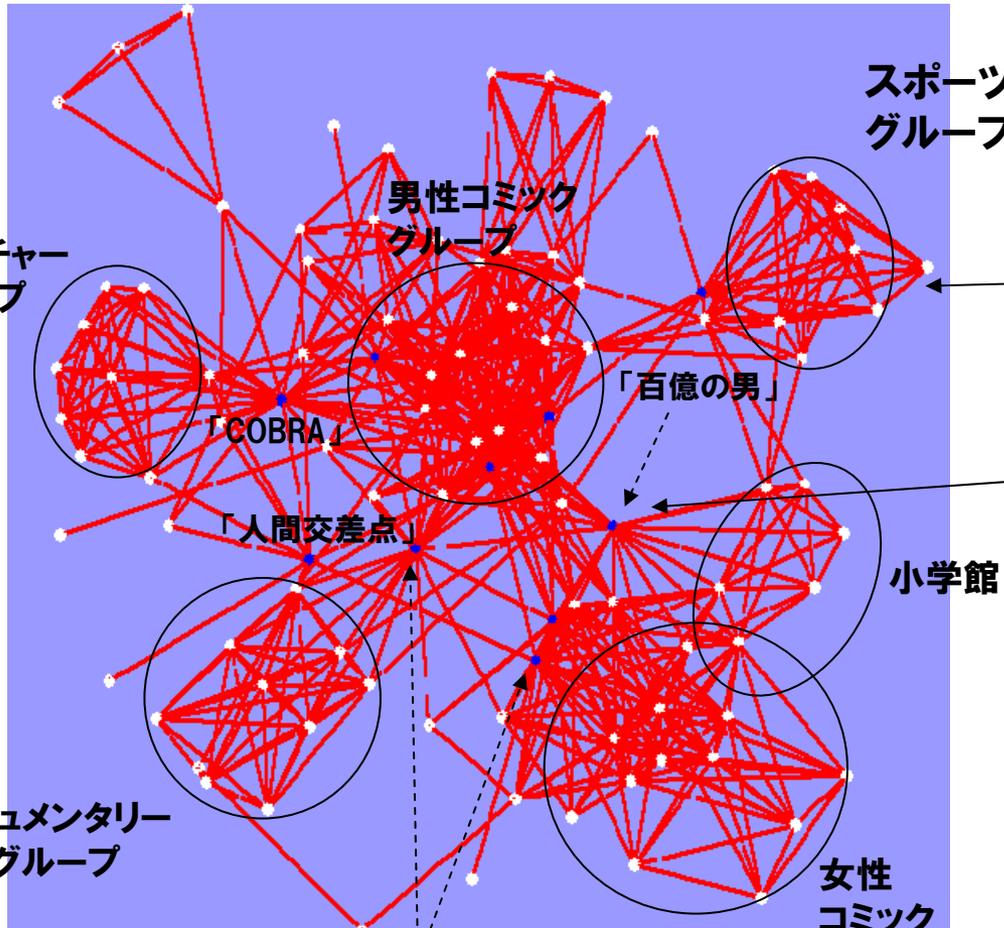
山田 武士

ネットワーク解析関連研究マップ



何故ネットワークか？

大規模なデータは、ネットワークで理解せよ
個々のデータとその関係を単純化し、マクロに解析



携帯配信コミックの
コンテンツネットワークの例

タイトルのグルーピング
に基づき類似コミックを
読者へレコメンド

「橋渡し」タイトルを読ませて
グループからグループへの
渡りを誘導

NW分析結果に基づき、
サイト上でのタイトルの
効果的露出とレコメンド
でタイトル渡りを促進

グループ間をつなぐ「橋渡し」コンテンツ(青色ノード)

目次

大規模で複雑なネットワークの主要構造を自動抽出して、その骨組みや機能を理解し、有効利用するための方法論の構築

基礎

スモールワールド
スケールフリーネットワーク

一部グラフ、無向グラフ

クラスタリング法

全ノードが重要として、それらを幾つかのコミュニティに分割し、ネットワークの全体構造を理解⇒ N&Gクラスタリング

コア抽出法

密結合するノード群(コミュニティ)に着目しネットワークの主要構造を理解⇒ k-dense コア抽出

可視化

⇒ Kamada&Kawai バネモデル
⇒ Fruchterman&Reingoldモデル

有向グラフ

ランキング

⇒ PageRank
⇒ HITS

二部グラフ、有向グラフなど、より複雑な関係データへの拡張

共起ネットワーク

⇒ 二部→一部

ブロックモデル

⇒ IRM

実世界に存在するネットワークの特徴

スモールワールド

スケールフリー

スモールワールド実験

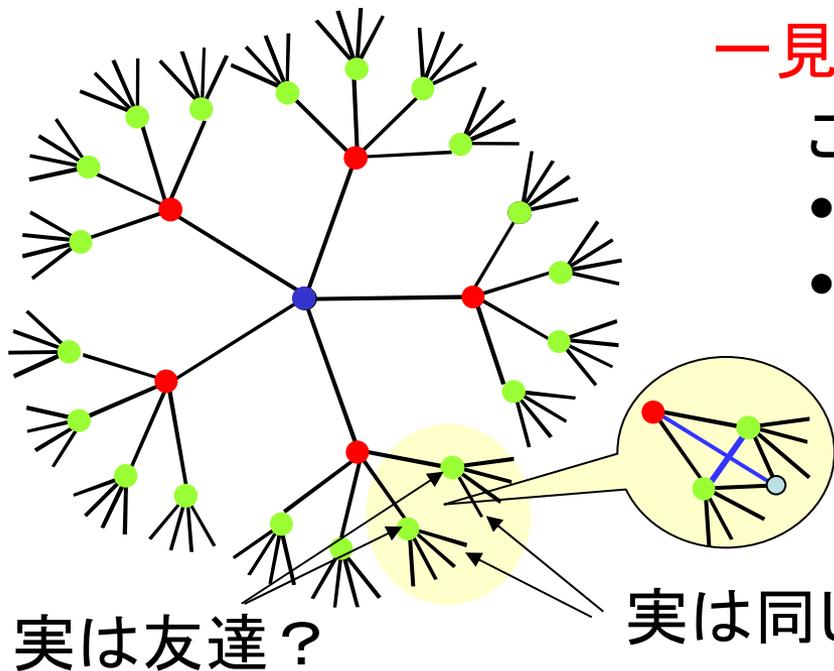
1960年代後半、Travers と Milgram が提唱。知り合い関係を次々と辿ることで短いステップ数で世界中の誰にでもたどり着く

- **ターゲット** (ボストンの株式仲買人) を設定
- ボストン (100人)、ネブラスカ (196人) から計296人の「**第一送信者**」を選出
- 各人は、自分よりターゲットに「**より近い**」人に**手紙を送信**する。
手紙を受け取った人が「**第二送信者**」となる。
- 順次これを**繰り返す**。
- 約20%がターゲットに到達。
- ターゲットに到達した繰り返しの長さの**平均値は6であった**。
⇒ Six Degrees of Separation (六次の隔たり)
⇒ It's a small world!?! (世間は狭い?)



スモールワールドの「簡単な」説明

- Aさん: 1
- Aさんの友達: 100
- Aさんの友達の友達: 10000 (=100²)
- ...
- 100⁵ = 100000000000 = 100億 > 地球の人口全体



一見もつともらしいが

これが成り立つ条件

- Aさんは「ランダム」に友達を作る
- 「Aさんの友達同士が友達」である確率は小さい

⇒現実的ではない(現実には局所的に「クラスタ」化されている)

実は友達？

実は同じ人？

クラスタ係数

C : クラスタ係数

Aさんの友達同士が友達である割合

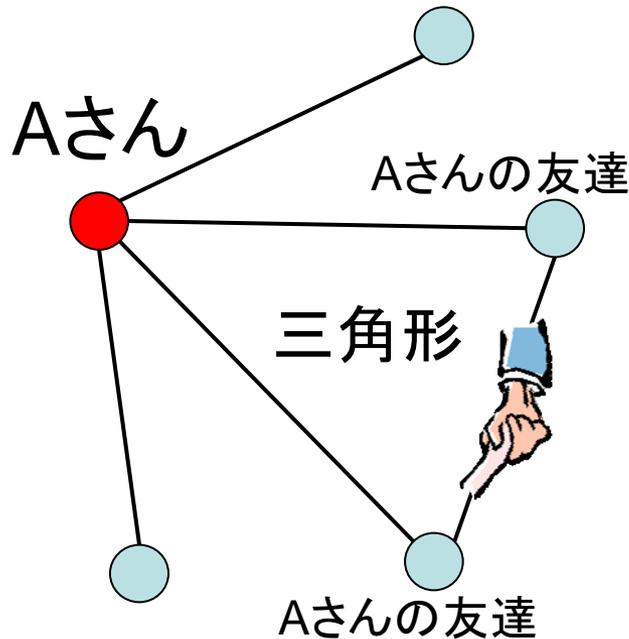
d : 次数(=リンク数)

e : 三角形の数

実際の三角形
の個数

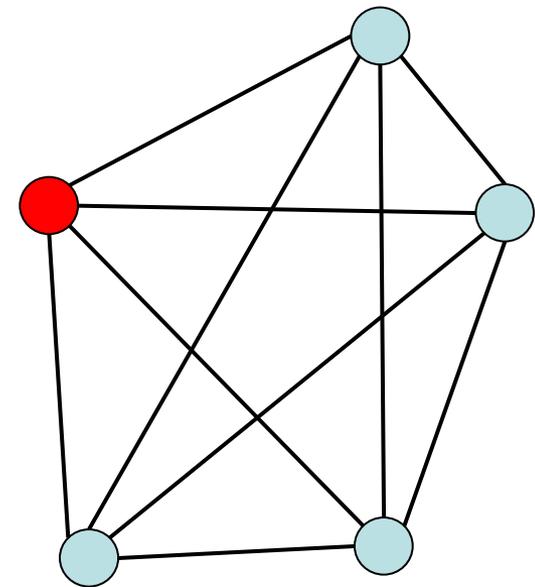
$$C = \frac{e}{d(d-1)/2}$$

可能な三角形
の個数



$$C = 1/6$$

クラスタ度が高い



$$C = 1$$

Characteristic Path Length

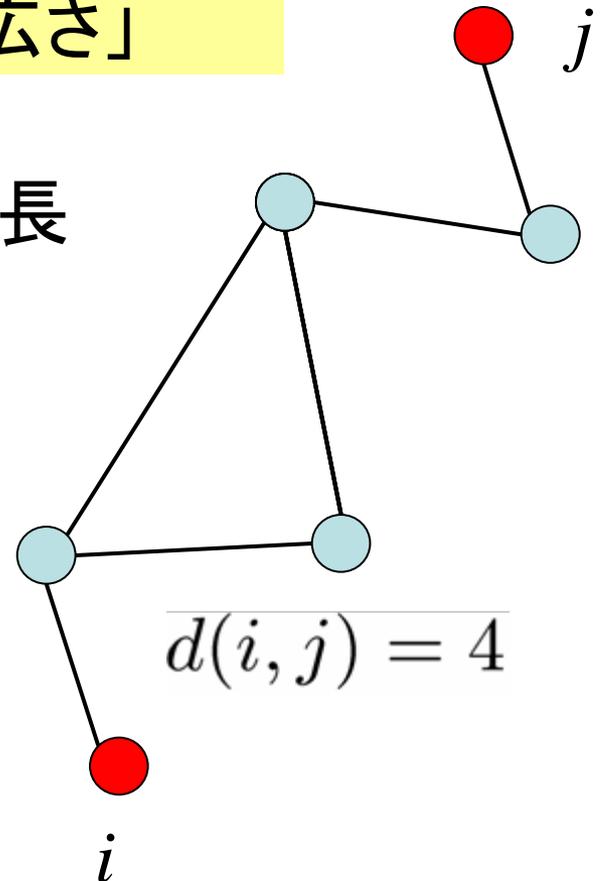
L : Characteristic Path Length (平均パス長)

任意のノードペア間のパス長の平均値
≡ ネットワークの平均的「広さ」

$d(i, j)$: i j 間のグラフ上のパス長
(最短距離)

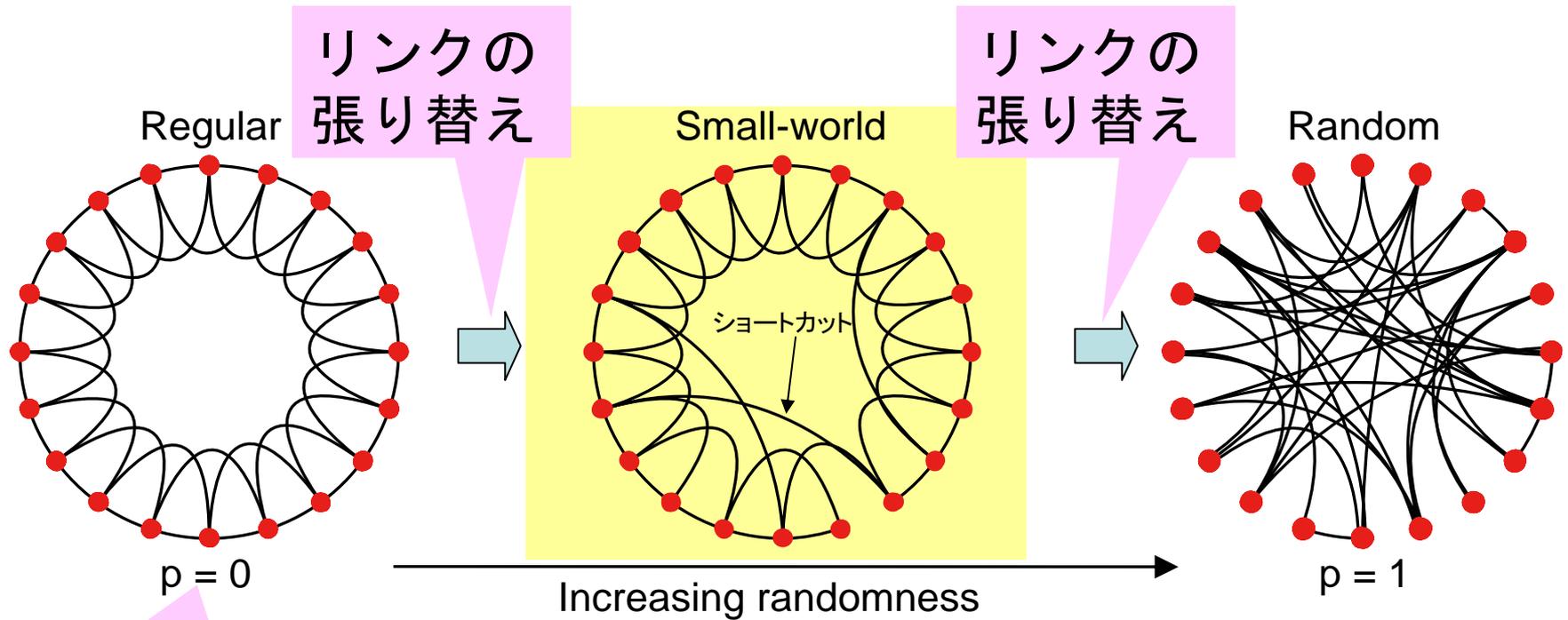
$$\bar{d}(i) = \frac{1}{n-1} \sum_{j=1}^n d(i, j)$$

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n \bar{d}(i)$$



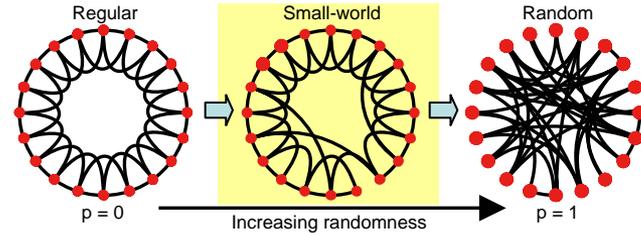
ネットワークのリンクの張替え

規則性(秩序)からランダムへ



リンクの張り替え確率

スモールワールド

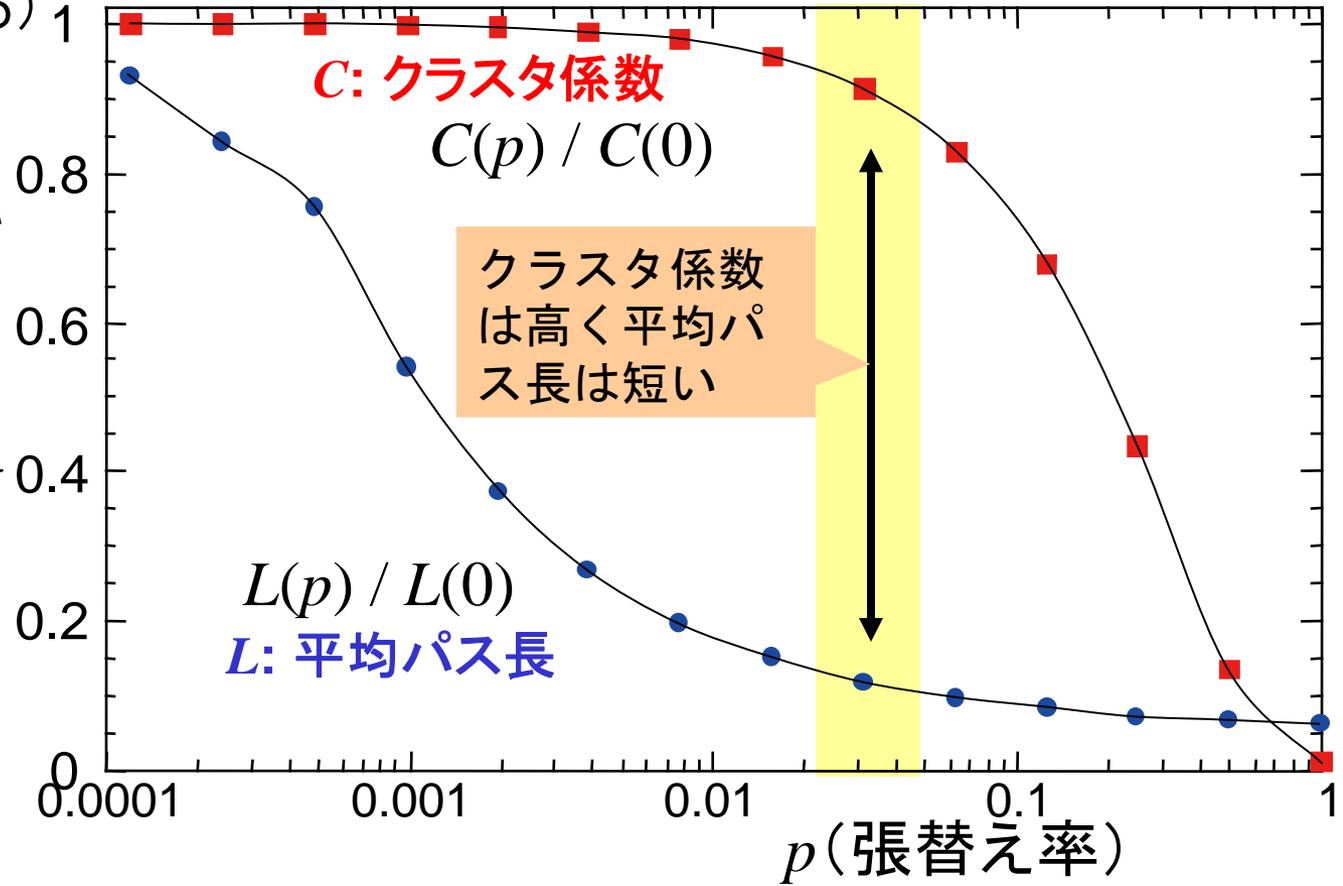


クラスタ係数が高い
=局所的に密なクラスタを
形成(構造を持っている)

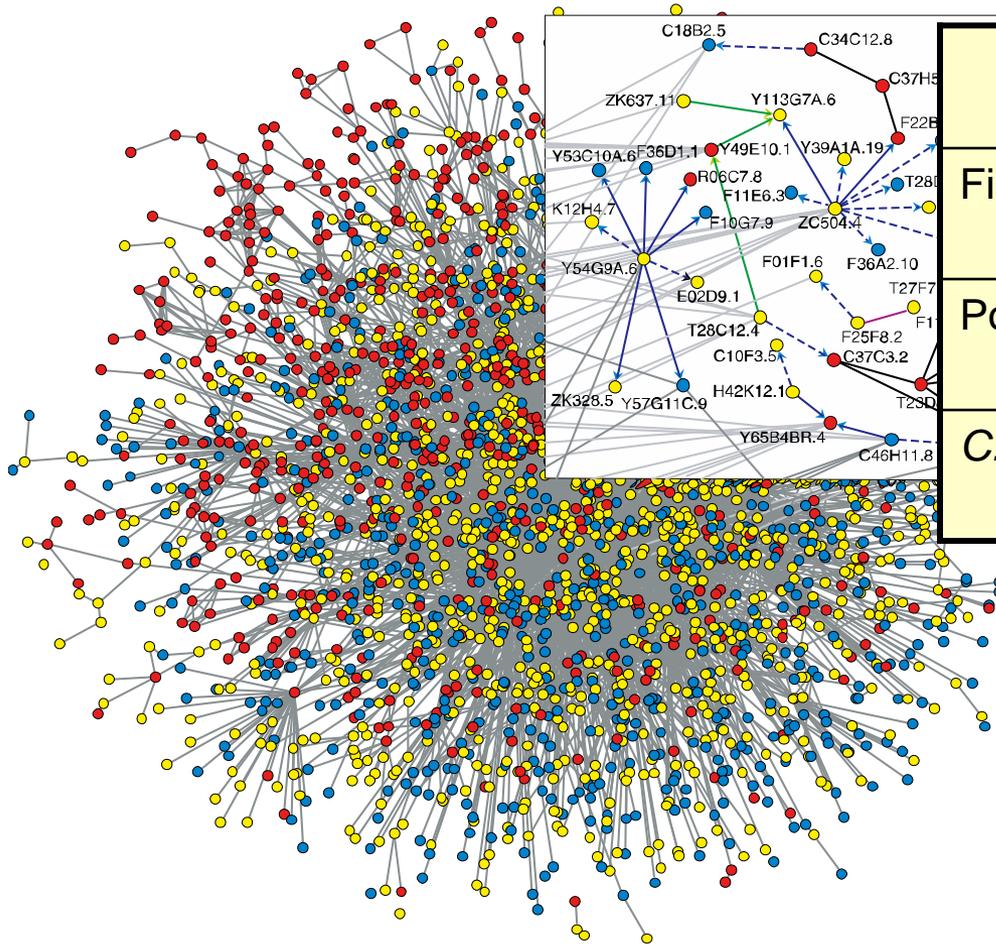
+
平均パス長が短い
=全体が離れていない

||

Small-world



スモールワールドNWの例



	L_{actual}	L_{random}	C_{actual}	C_{random}
Film Actors (俳優)	3.65	2.99	0.79	0.00027
Power grid (電力網)	18.7	12.4	0.080	0.005
<i>C. elegans</i> (桿線虫)	2.65	2.25	0.28	0.05

- Core
- - - Non-Core
- Literature
- Scaffold
- Interolog

Actors: $n = 225,226$, $k = 61$.

Power grid: $n = 4,941$, $k = 2.67$.

C. elegans: $n = 282$, $k = 14$.

n : ノード数、 k : 平均次数

類似コンテンツ高速探索技術

スモールワールドネットワークで類似コンテンツを瞬時に探索

【探索する】

概要

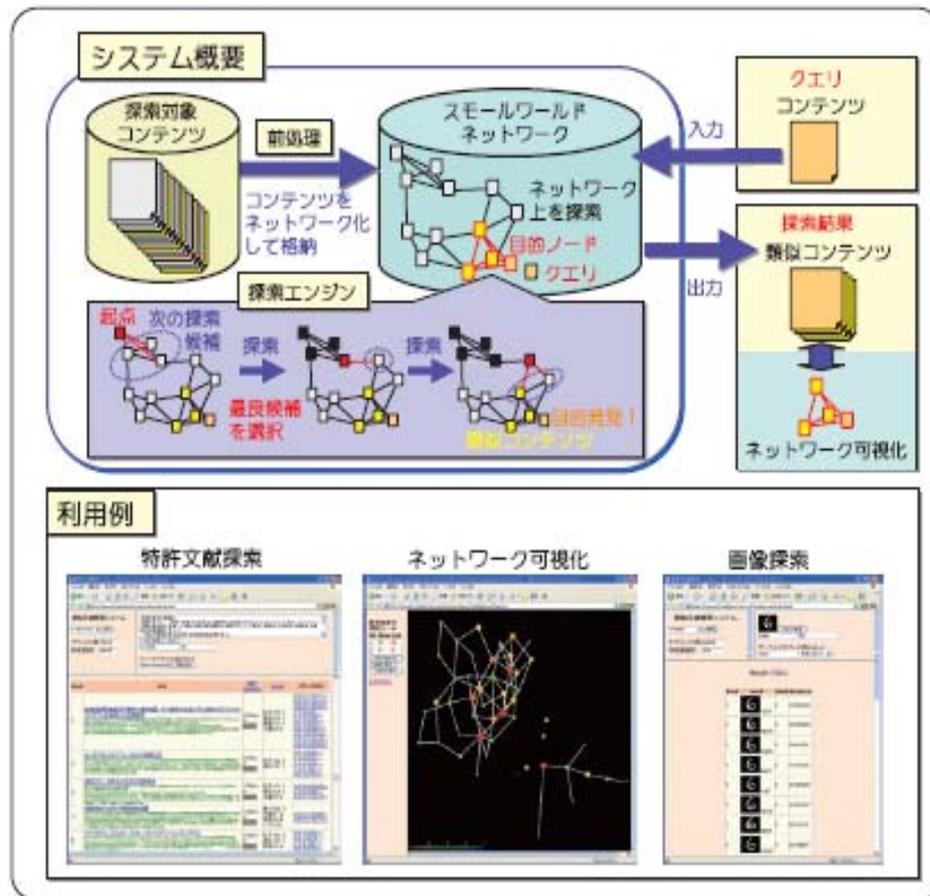
大規模コンテンツデータを、スモールワールド特性を持つ近傍ネットワークとして効率良く格納します。クエリに対し、ネットワーク上の類似ノードを順に辿り、小ステップ数で、目的ノードに到達することで、高速な類似探索を実現します。

特徴

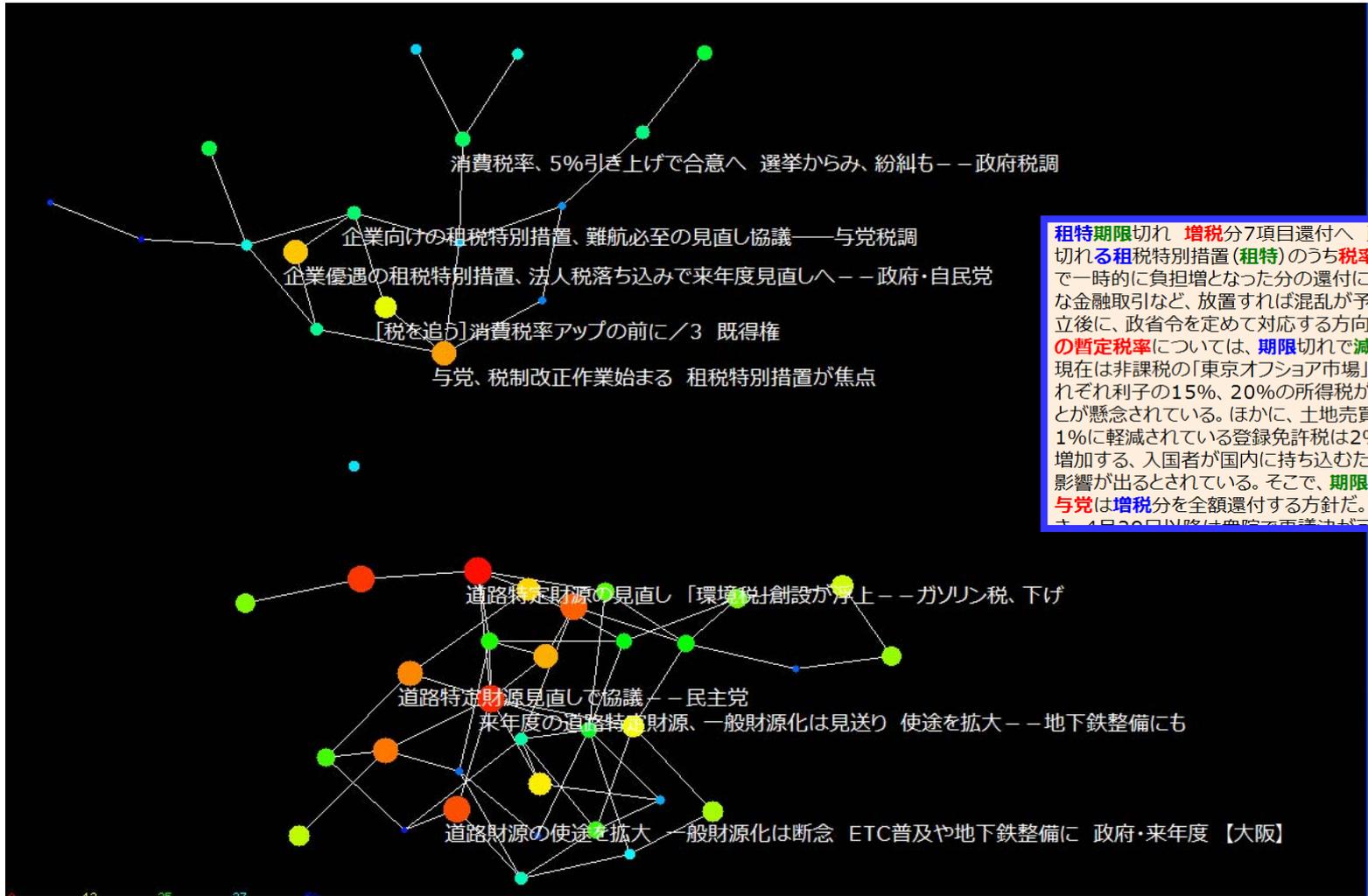
- コンテンツ自体を入力とした高速探索
- 探索結果をネットワークとして可視化
- 多様な種類のコンテンツに対応

利用分野

- 大規模データの類似探索
- マルチメディアデータ探索
- 文書自体を入力とする探索



ニュース記事探索結果



租税期限切れ 増税分7項目還付へ 政府・与
切れる租税特別措置(租特)のうち**税率**変更な
 で一時的に負担増となった分の還付に応じる方
 な金融取引など、放置すれば混乱が予想され
 立後に、政省令を定めて対応する方向で検討
の暫定税率については、**期限**切れで**減税**とな
 現在は非課税の「東京オフィス市場」取引や
 れぞれ利子の15%、20%の所得税がかかる
 とが懸念されている。ほかに、土地売買の所有
 1%に軽減されている登録免許税は2%に倍増
 増加する、入国者が国内に持ち込むたばこや
 影響が出ると思われる。そこで、**期限**切れ期
与党は**増税**分を全額還付する方針だ。**税制**関
 連、1月30日以降は衆院で再議決が可能に

4

道路財源の用途を拡大 一般財源化は断念 ETC普及や地下鉄整備に 政府・来年度【大阪】
 政府は5日、03年度予算で最大の焦点になっていた道路特定財源の見直しについて、使途を限定しない一般財源化を断念し、道路特定財源への用途拡大にとどめる方針を固めた。道路特定財源のうち揮発油税などは、道路建設の財源不足を理由に本来の税率に暫定税率を上乗せしており、一般財源化して道路に無関係な予算に使うと暫定税率分を徴収する理由がなくなる。取扱い財政事情を考慮し、財源確保の観点から見送ることとした。(3面に関連記事)小泉統一記者団は昨年4月の就任当初から、一般財源化も含めて道路特定財源を見直す方針を
 2002年11月06日記事

2.4158e-01

財源:17.0 %
定財:16.9 %
定税:10.6 %
道路:9.4 %
特定:5.4 %
税率:5.3 %

実世界に存在するネットワークの特徴

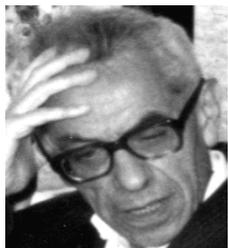
スモールワールド

スケールフリー

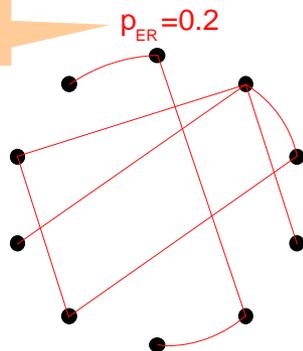
スケールフリーネットワーク

k : 次数(ノードのリンク数)、 $P(k)$: 次数 k のノード数の割合(確率)

一定の確率で
リンクを張る

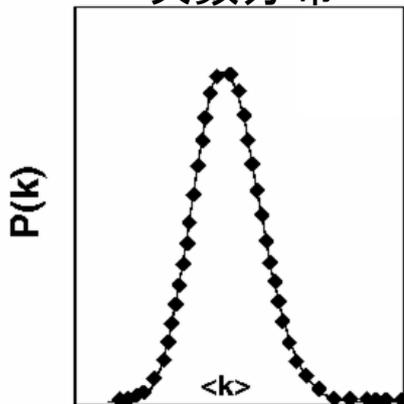


Erdős

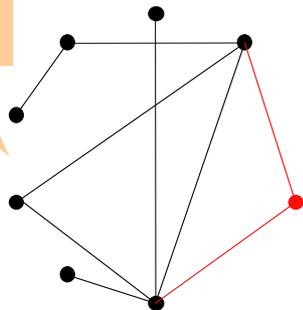


Erdős-Rényi ランダムネットワーク k

次数分布

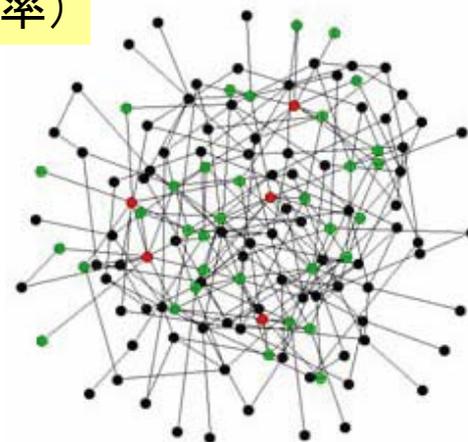
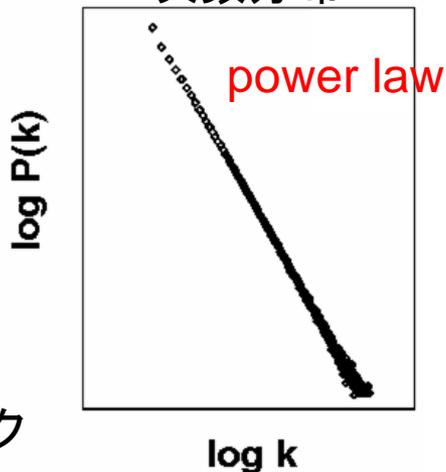


次数に比例する
確率でリンクを張る

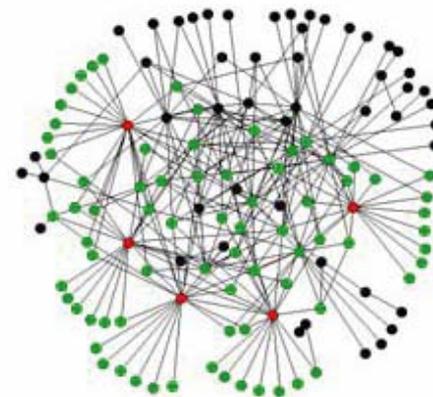


スケールフリーネットワーク

次数分布



もっとも次数の高い5ノード(赤)と繋がっている(緑)のは27%



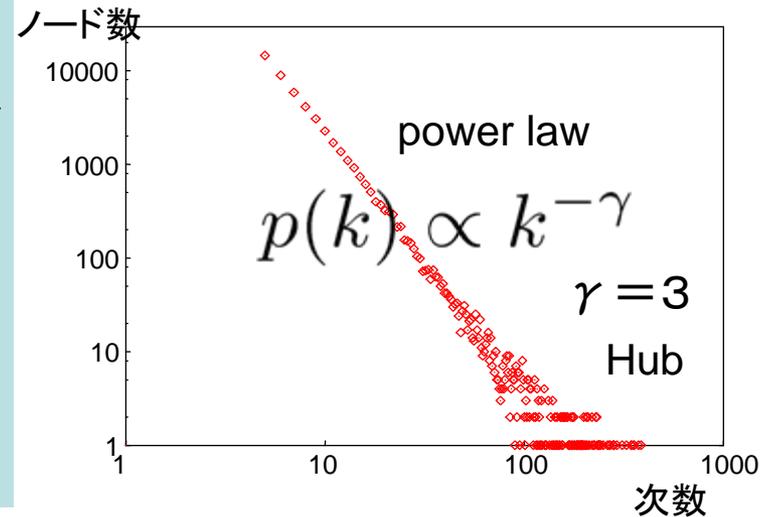
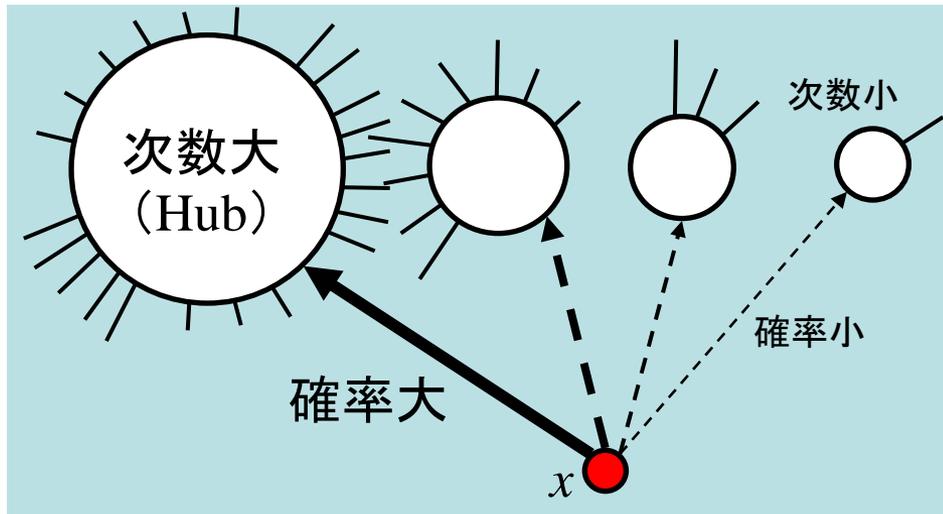
もっとも次数の高い5ノード(赤)と繋がっている(緑)のは60%

スケールフリーネットワーク構成法

優先的選択法(Preferential Attachment) [Barabasi & Albert 1999]

- 現在のネットワークに新たなノード x を追加
- 次数(リンク数)に比例する確率で、既存のノードをランダムに選択し、 x からリンクを張る、これを m 回行う(x からのリンク m 本)

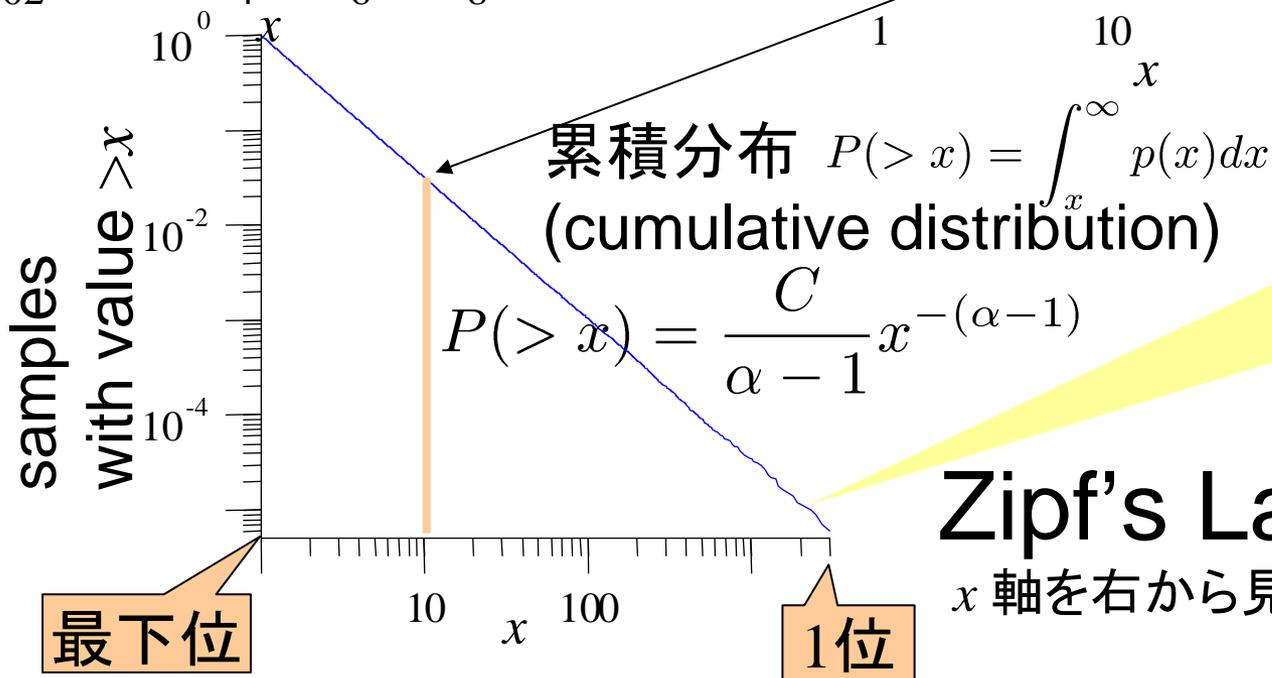
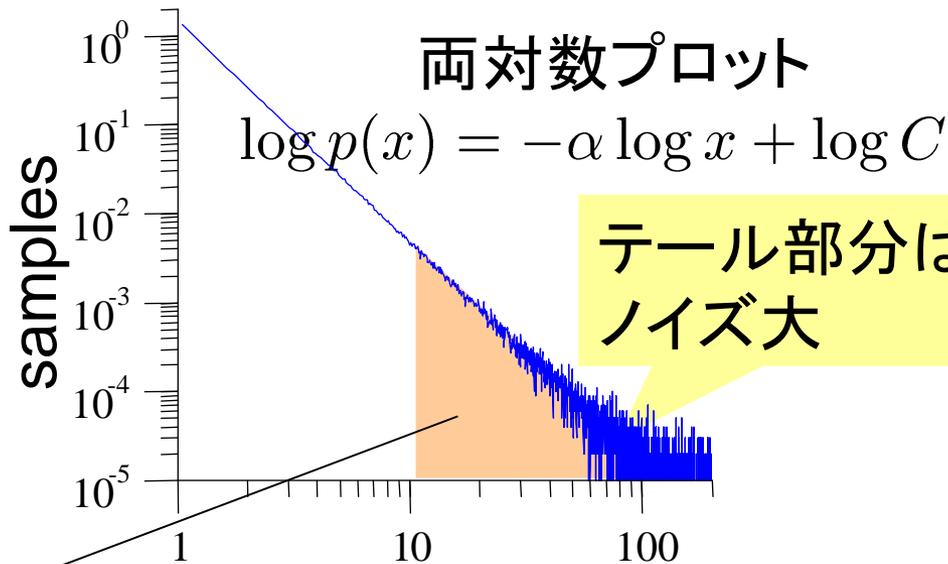
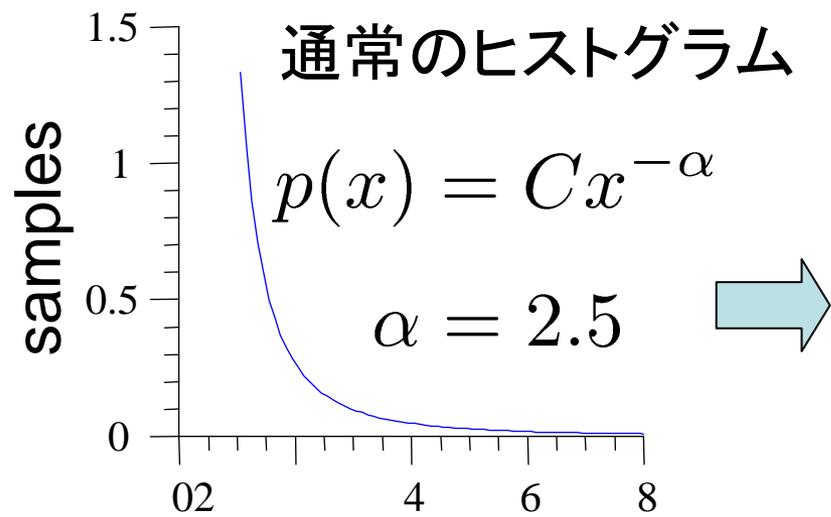
⇒ ほぼ係数 $\gamma = 3$ のべき則(power law)に従うネットワークを生成



- すでにたくさんリンクを持つノードほど、新たなリンクを張りやすい
- リンクが密な部分、疎な部分が混在⇒「構造」を持つネットワーク
- モジュール性 (Modularity)、クラスター係数、Hub = クラスタ?

「構造」抽出

べき則 (Power Law) の表現

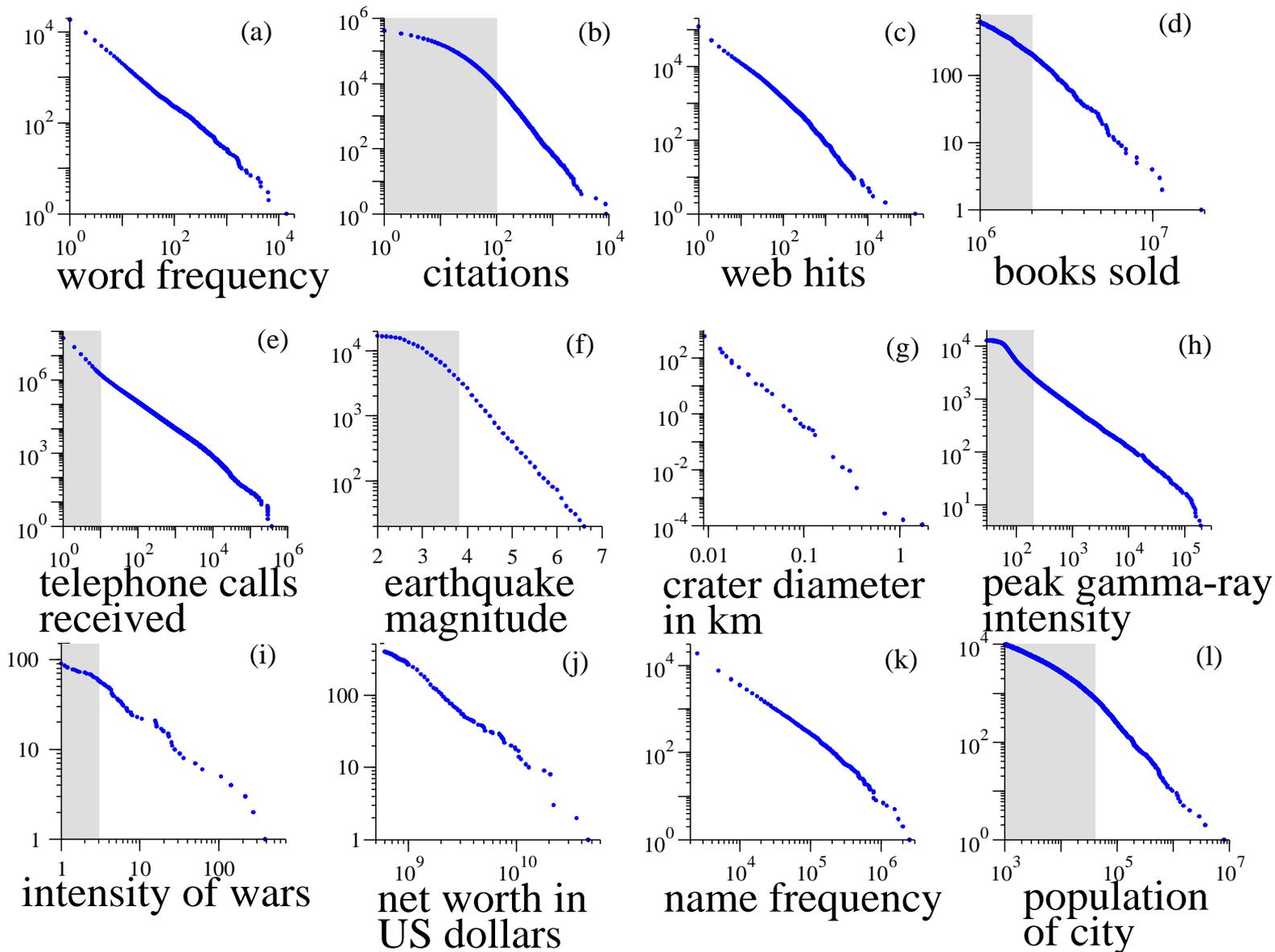


累積分布を用いることで、テール部分のノイズを吸収

Zipf's Law

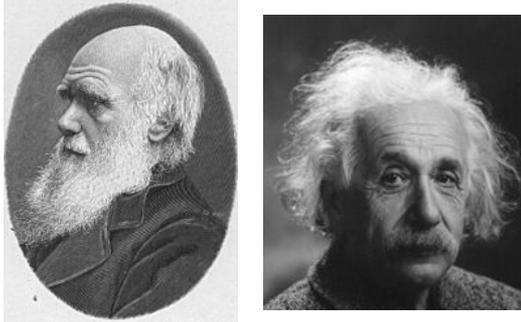
x 軸を右から見ると、「順位」

べき則 (Power Law) の例

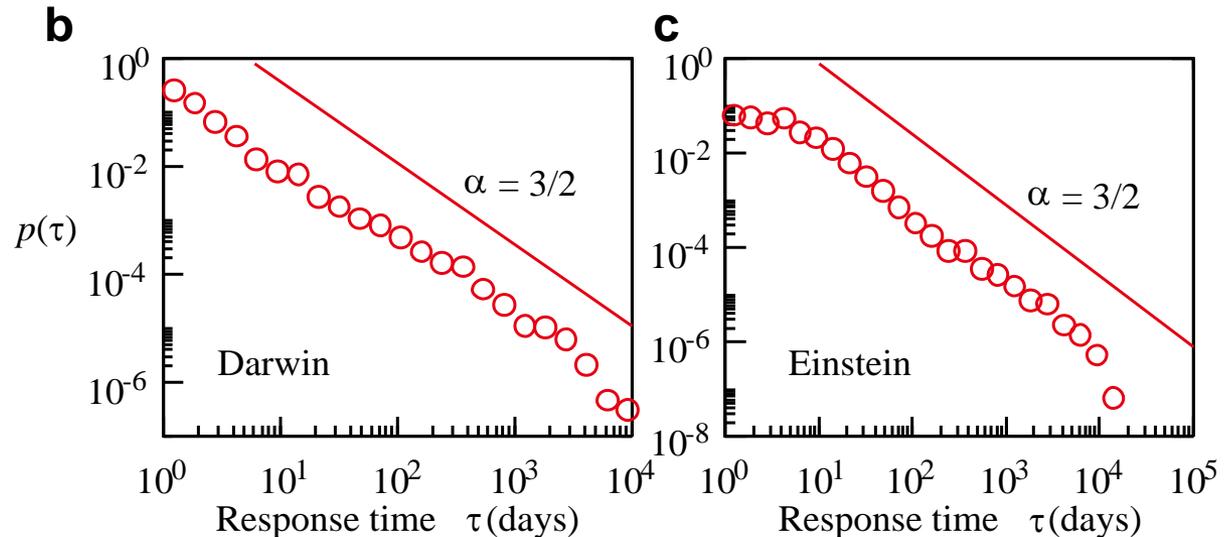
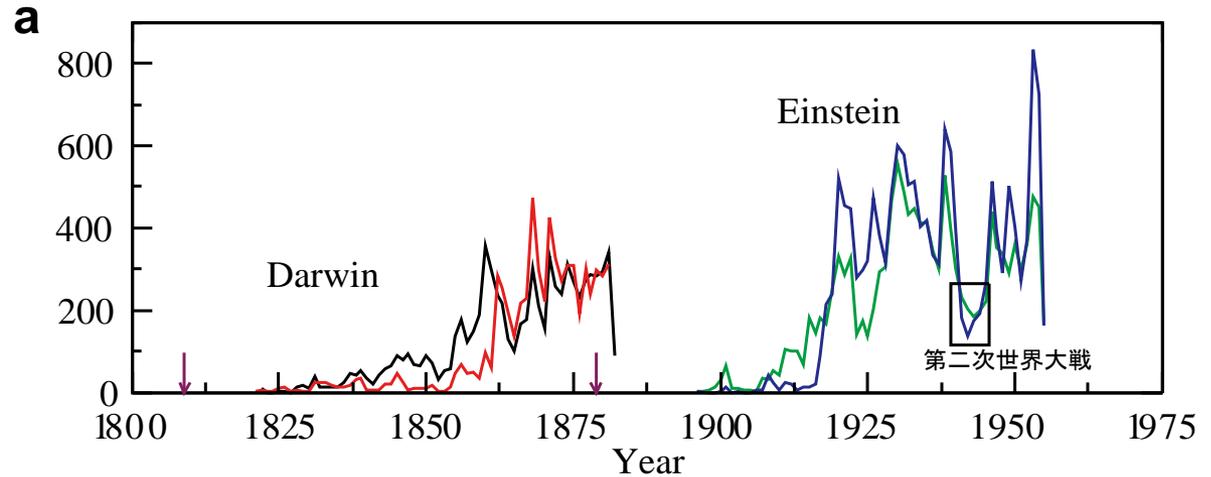


べき則の例(その2)

ダーウィンとアインシュタインに見る、手紙やりとりパターン



現代人がEメールの優先順位をつけるのと同じように、手紙の返信に優先順位をつけていた。



べき則 (Power Law) になる？ならない？

- 物理的な制約があると、「スケール」のため、べき則にはならない(「スケールフリー」⇒平均・分散無限大)
- べき則にならない例
 - 人間の身長、体重の分布(平均の周りに分布)
 - ひとり当たりの友達の数分布(付き合える友達の数には限界)
⇔ 「その人を知っている人の数」の分布
 - Webページ一枚あたりの出リンク数(書き込める量に限界)
⇔ 「入リンク数」の分布
 - 一人当たりのCD購入数の分布(購入できる数に限界)
⇔ CD一枚あたりの購入者数の分布
- “rich-gets-richer” process (富める者はますます富む)はべき則の要因 — 無制限に増えるため

スモールワールド

スケールフリー

スパース(疎)

これらの特徴
を生かして

ランキング

クラスタリング
／コア抽出

共起NW解析

可視化

スモールワールド

スケールフリー

スパース(疎)

これらの特徴
を生かして

ランキング

クラスタリング
／コア抽出

共起NW解析

可視化

GoogleのPageRankアルゴリズム

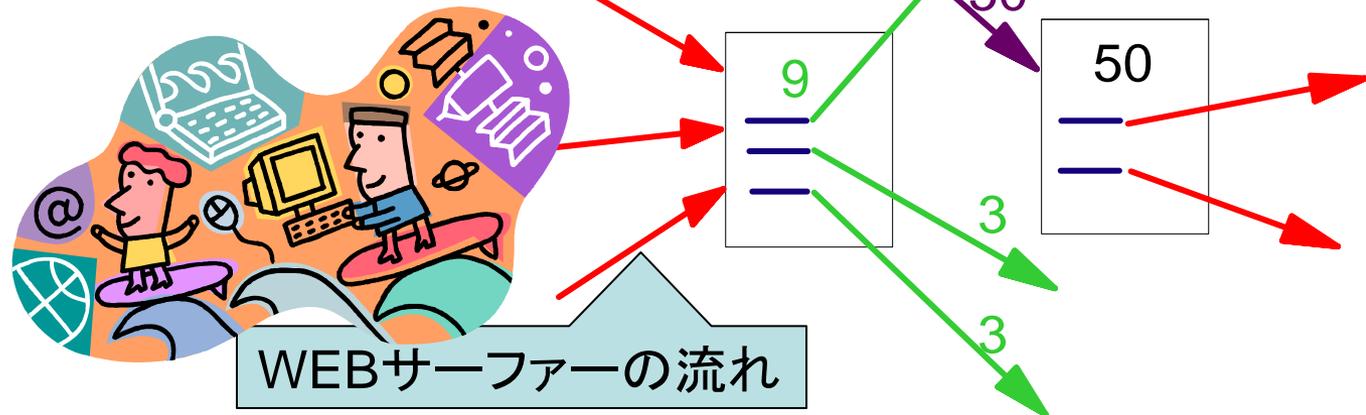
- 重要なページからリンクされているページは重要
- ランダムサーファーマデル



Larry Page

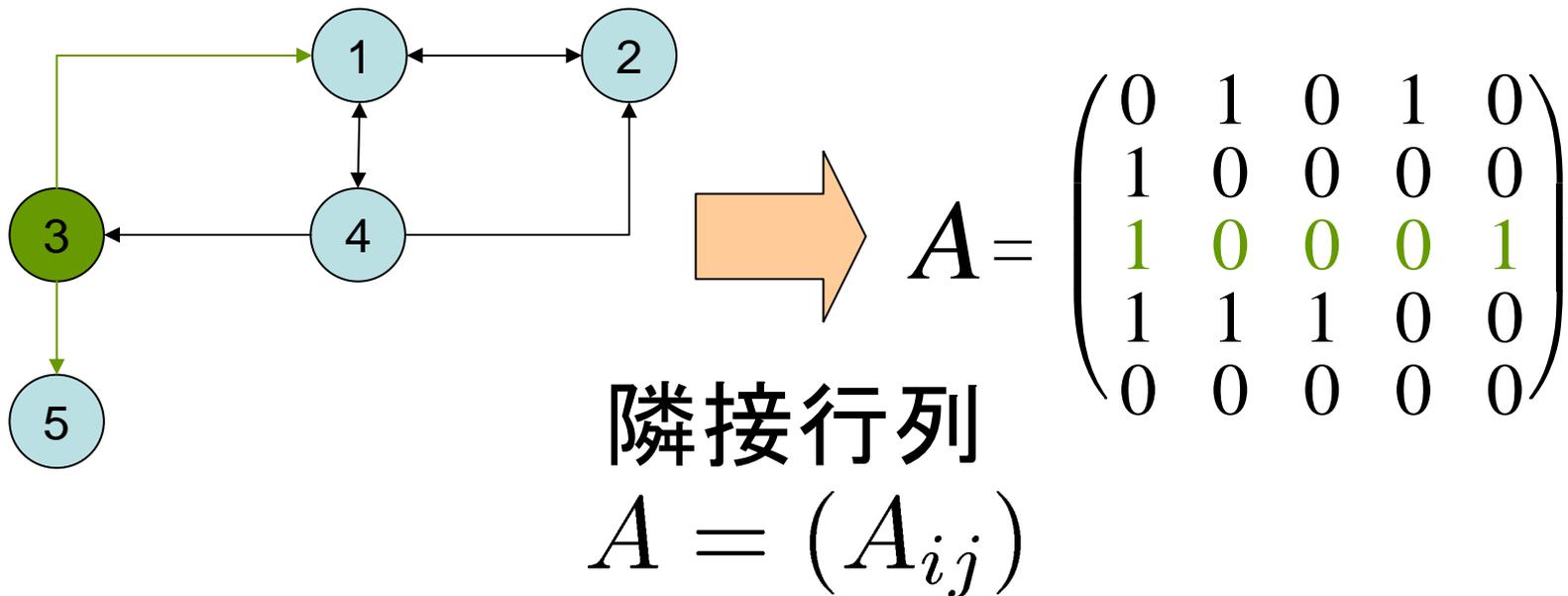


Sergey Brin

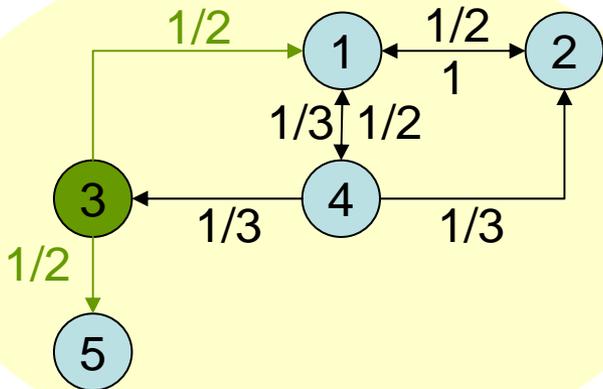


隣接行列によるネットワーク表現

隣接行列とは「あるノード」から「あるノード」へリンクがあれば「1」、なければ「0」とした行列



隣接行列から遷移行列へ



遷移行列とは、「あるノード」から「あるノード」へ遷移(移動)する確率の行列

隣接行列

$$A = (A_{ij})$$

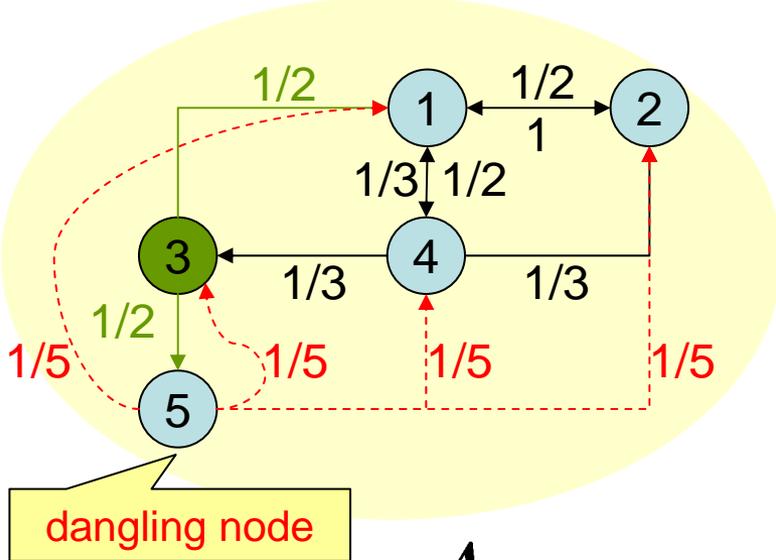
$$H_i = \frac{A_i}{\sum_k A_{ik}}$$

遷移行列

Row-Normality

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \Rightarrow H = \begin{pmatrix} 0 & 1/2 & 0 & 1/2 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

収束性を保証するため遷移行列の修正



$$\mathbf{a} = (a_i) \quad a_i = \begin{cases} 1 & \text{if } i \text{ is dangling node} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{e} = (1, \dots, 1)$$

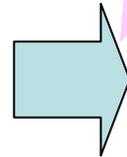
$$H_i = \frac{A_i}{\sum_k A_{ik}}$$

遷移行列

Stochasticity
収束性保証

$$S = H + \frac{\mathbf{a}\mathbf{e}^T}{n}$$

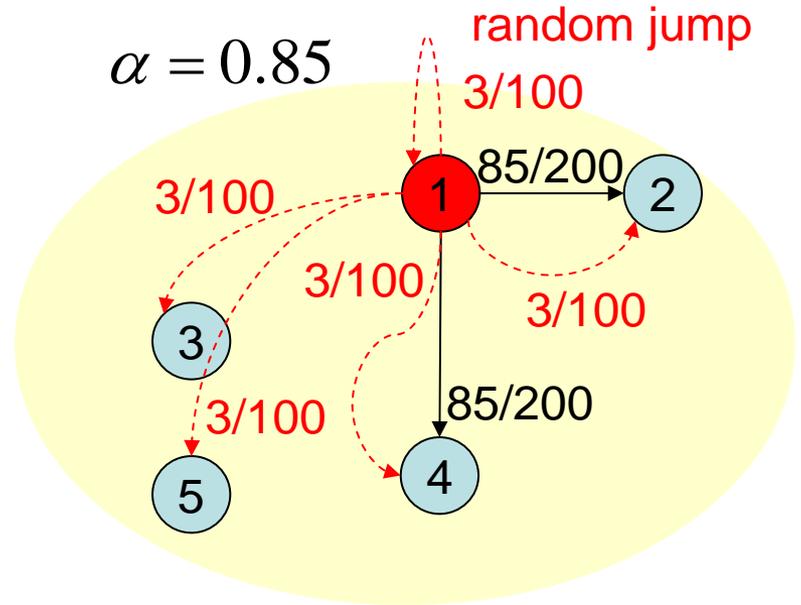
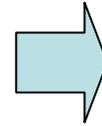
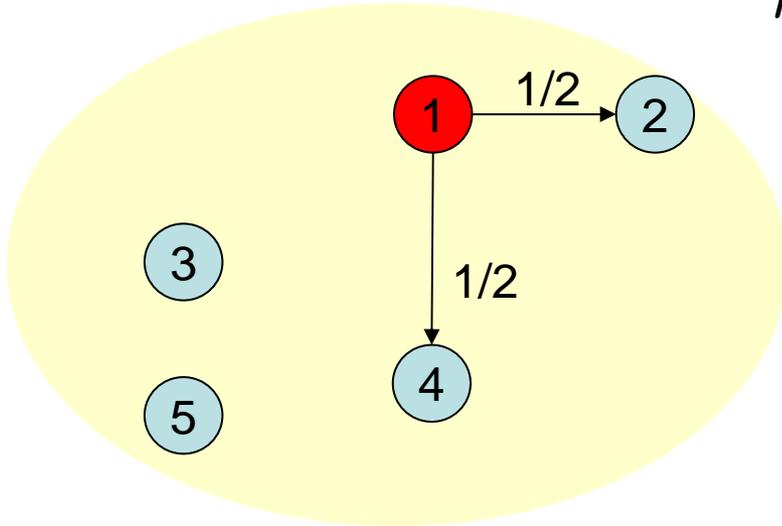
$$H = \begin{pmatrix} 0 & 1/2 & 0 & 1/2 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$



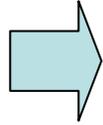
$$S = \begin{pmatrix} 0 & 1/2 & 0 & 1/2 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{pmatrix}$$

ランダムジャンプを考慮したGoogle Matrix

$$G = \alpha S + (1 - \alpha) \frac{ee^T}{n}$$



$$S = \begin{pmatrix} 0 & 1/2 & 0 & 1/2 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{pmatrix}$$



$$G = \begin{pmatrix} 3/100 & 91/200 & 3/100 & 91/200 & 3/100 \\ 22/25 & 3/100 & 3/100 & 3/100 & 3/100 \\ 91/200 & 3/100 & 3/100 & 3/100 & 91/200 \\ 47/150 & 47/150 & 47/150 & 3/100 & 3/100 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{pmatrix}$$

Irreducibility
単一の定常値保証

random jump: $0.15 \times 1/5 = 3/100$

確率的な繰り返し=パワー法による固有値計算

$$G = \alpha S + (1 - \alpha) \frac{\mathbf{e}\mathbf{e}^T}{n}$$

初期値

$$\pi^0 = \mathbf{1}/n = \begin{pmatrix} 1/n \\ \dots \\ 1/n \end{pmatrix} \xrightarrow{G \text{ で遷移}} \pi^{(1)T} = \pi^{(0)T} G$$

繰り返す

$$\xrightarrow{} \pi^{(k+1)T} = \pi^{(k)T} G \quad (= \dots = \pi^{(0)T} G^k)$$

定常状態

$$\xrightarrow{} \pi^T = \pi^T G \quad \text{固有値} = 1$$

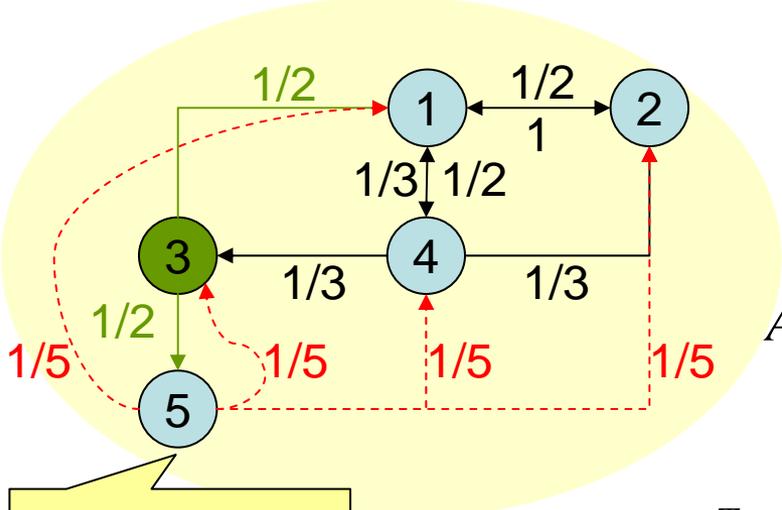
注) 固有値 > 1 だと収束しない

π : (最大) 固有値 1 に対応する固有ベクトル

$$\pi = \begin{pmatrix} 0.35961320922905 \\ 0.25380393805204 \\ 0.10096832412970 \\ 0.19776930237822 \\ 0.08784522621099 \end{pmatrix}$$

PageRank Vector

GoogleのPageRankアルゴリズム(まとめ)



隣接行列

$$A = (A_{ij})$$

$$A = \begin{pmatrix} 0 & 1/2 & 0 & 1/3 & 1/5 \\ 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Row-Normality

$$H_i = \frac{A_i}{\sum_k A_{ik}}$$

遷移行列

$$H = \begin{pmatrix} 0 & 1/2 & 0 & 1/2 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Google Matrix

random jump: $0.15 \times 1/5 = 3/100$

$$S = H + \frac{ae^T}{n}$$

Stochasticity
収束性保証

$$S = \begin{pmatrix} 0 & 1/2 & 0 & 1/2 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{pmatrix}$$

$$G = \alpha S + (1 - \alpha)E \quad \alpha = 0.85$$

Irreducibility
単一の定常値保証

$$G = \begin{pmatrix} 3/100 & 91/200 & 3/100 & 91/200 & 3/100 \\ 22/25 & 3/100 & 3/100 & 3/100 & 3/100 \\ 91/200 & 3/100 & 3/100 & 3/100 & 91/200 \\ 47/150 & 47/150 & 47/150 & 3/100 & 3/100 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{pmatrix}$$

パワー法により最大固有値(1)に対応する固有ベクトルを計算

$$\pi^T = \pi^T G$$

PageRank Vector

$$\pi^{(0)} = \mathbf{1}/n = \begin{pmatrix} 1/n \\ \vdots \\ 1/n \end{pmatrix}$$

任意

$$\pi^{(k+1)T} = \pi^{(k)T} G$$

収束するまで繰り返す

$$\pi = \begin{pmatrix} 0.35961320922905 \\ 0.25380393805204 \\ 0.10096832412970 \\ 0.19776930237822 \\ 0.08784522621099 \end{pmatrix}$$

PageRankアルゴリズムの特許

United States Patent 6,285,999

Page September 4, 2001

Method for node ranking in a linked database

Abstract

引用

ノードの重要度ランク

ある文書の順位は、それを引用する文書のランクから計算する

A method assigns importance ranks to nodes in a linked database, such as any database of documents containing citations, the world wide web or any other hypermedia database. The rank assigned to a document is calculated from the ranks of documents citing it. In addition, the rank of a document is calculated from a constant representing the probability that a browser through the database will randomly jump to the document. The method is particularly useful in enhancing the performance of search engine results for hypermedia databases, such as the world wide web, whose documents have a large variation in quality.

Inventors: Page; Lawrence (Stanford, CA)

データベースの閲覧者

閲覧者がランダムにある文書にジャンプする確率

Assignee: The Board of Trustees of the Leland Stanford Junior University

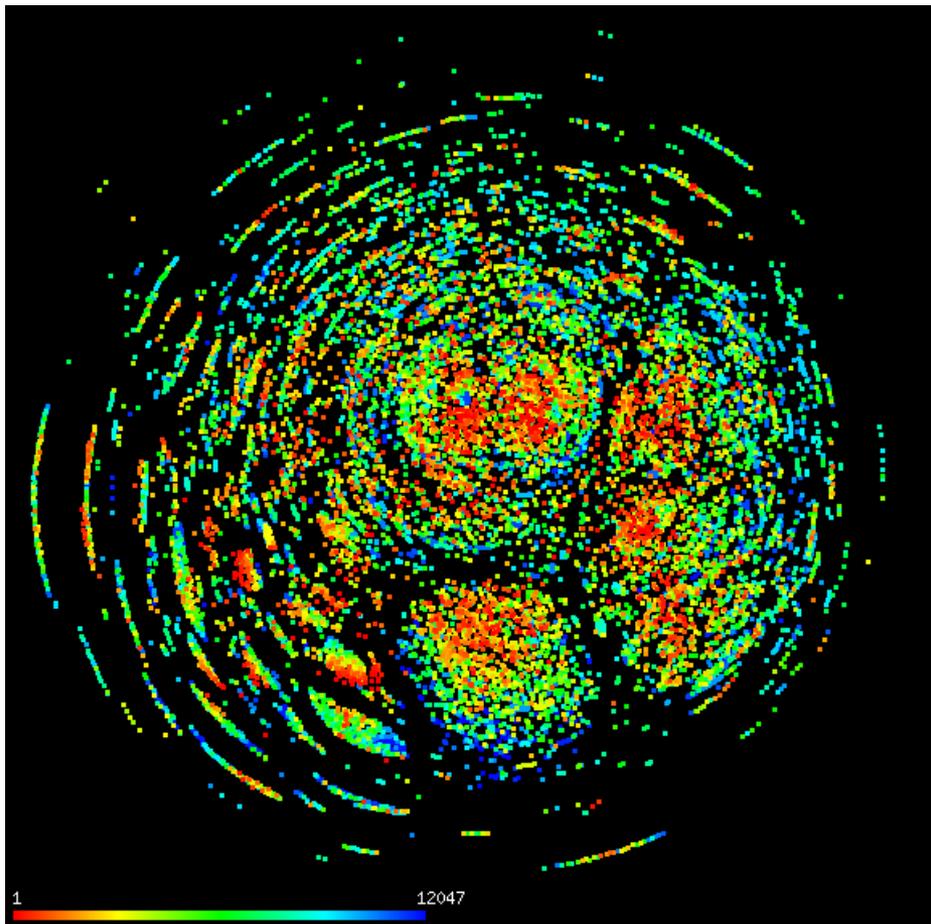
(Stanford, CA)

Appl. No.: 09/004,827

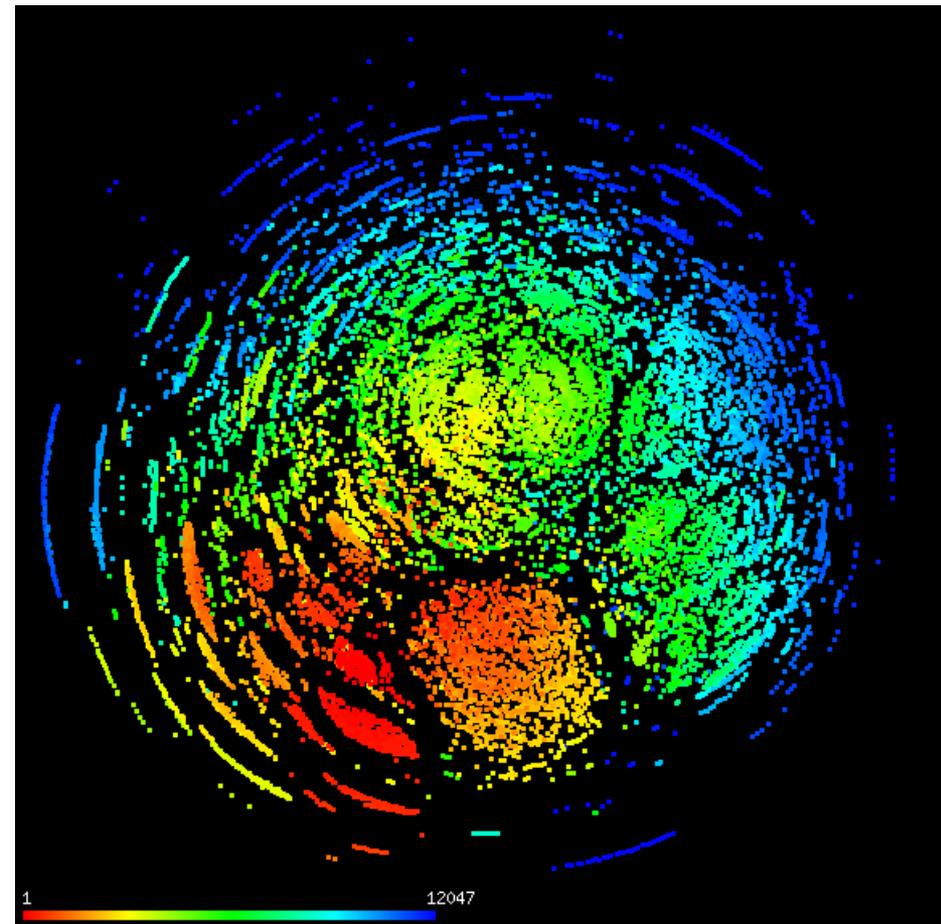
Filed: January 9, 1998

特許はスタンフォード大学が保有

PageRankとHITSの比較



PageRank: ネットワーク全体から
満遍なく重要ノードを抽出



HITS: 最大次数ノードを中心に
特定トピックに絞って抽出

スモールワールド

スケールフリー

スパース(疎)

これらの特徴
を生かして

ランキング

クラスタリング
／コア抽出

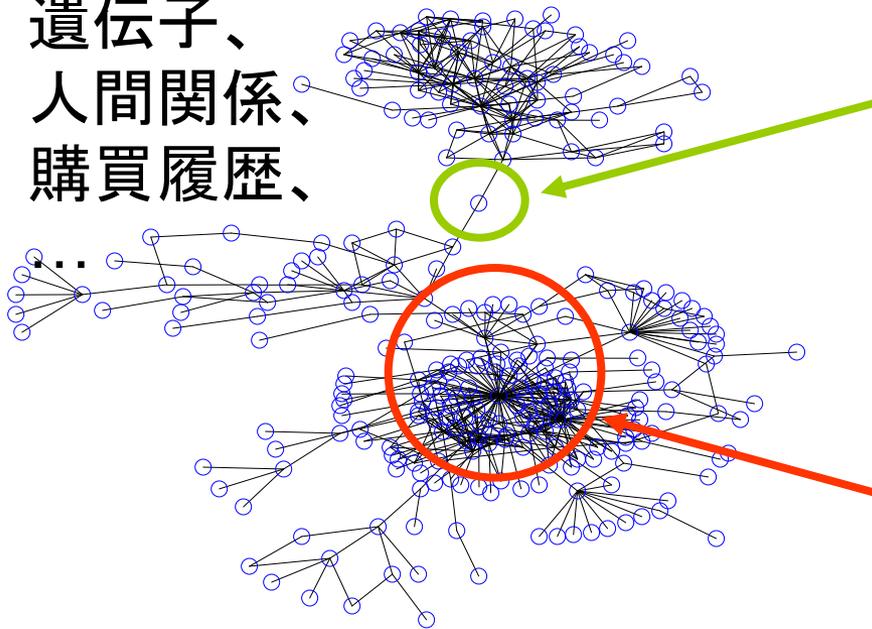
共起NW解析

可視化

ネットワークからの構造(コミュニティ)抽出

大規模で複雑なネットワークの主要構造を自動抽出して、その骨組みや機能を理解し、有効利用するための方法論の構築

ウェブ、
遺伝子、
人間関係、
購買履歴、
...



クラスタリング法

全ノードが重要として、それらを幾つかのコミュニティに分割し、ネットワークの全体構造を理解 例:GN法、IRM

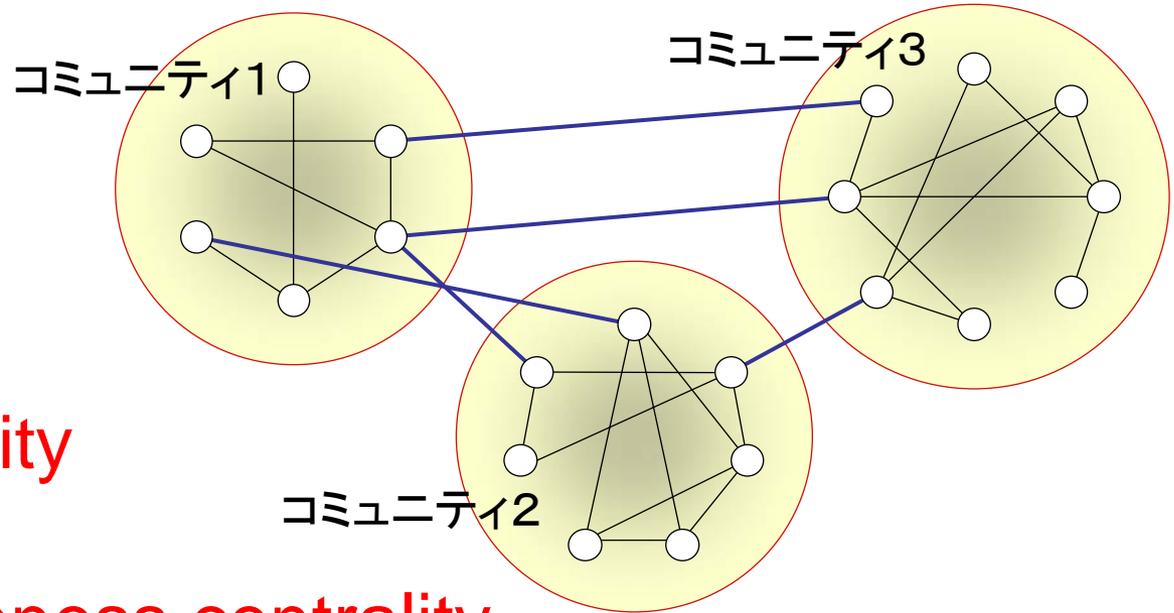
コア抽出法

密結合するノード群(コミュニティ)に着目しネットワークの主要構造を理解 例:k-core法、k-clique法、k-dense法

Newman と Girvan によるクラスタリング



M.E.J. Newman

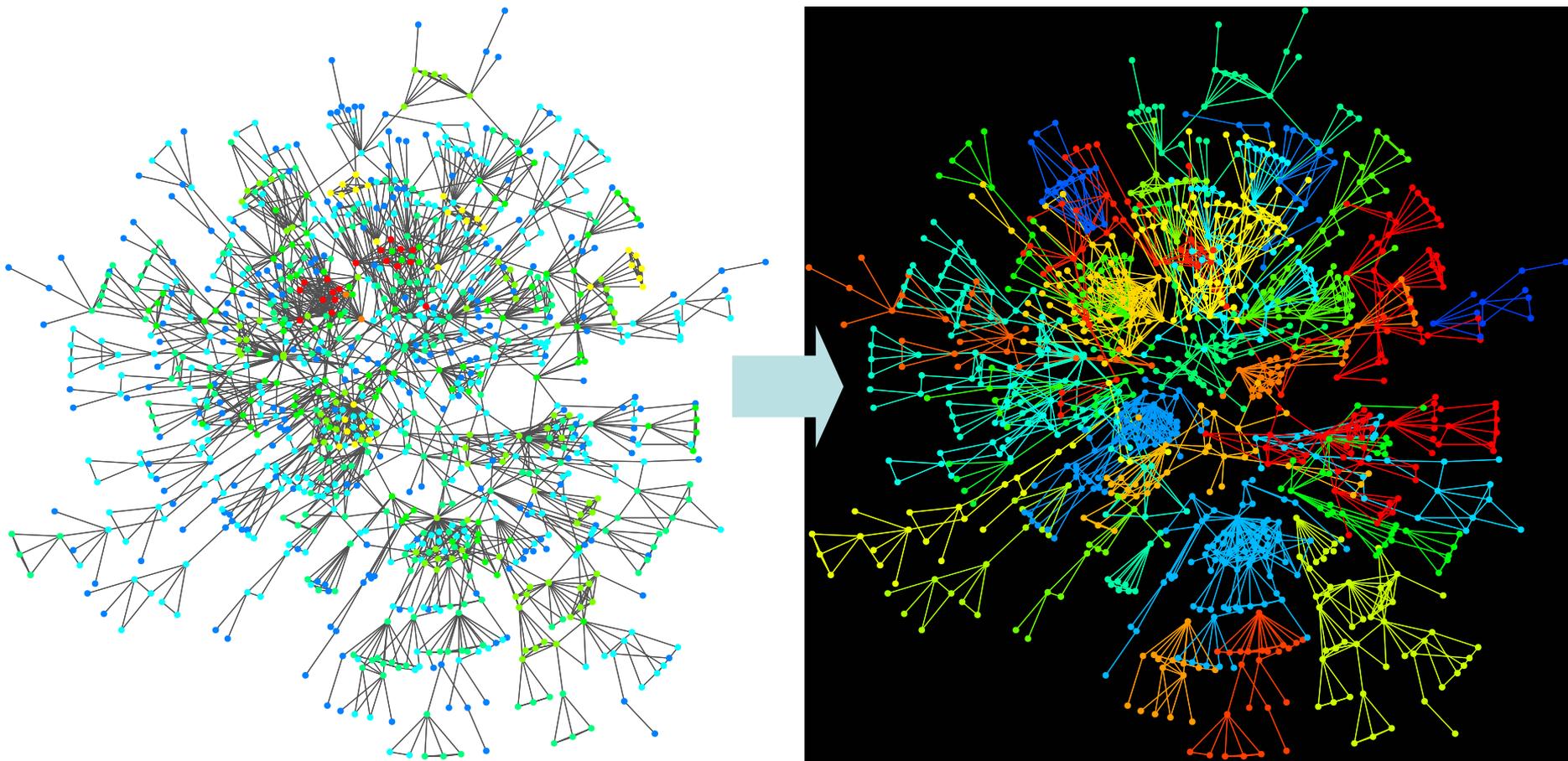


「モジュール度」
=modularity

「橋渡し度」
=betweenness centrality

- **コミュニティ**とは:「同一コミュニティ内のリンクは密、異なるコミュニティ間のリンクは疎」=「モジュール度」高い
- **クラスタリング**:「モジュール度」が高くなるようにコミュニティ間の「橋渡し度」が高いリンクを除去

NIPS論文共著者NWのクラスタリング例

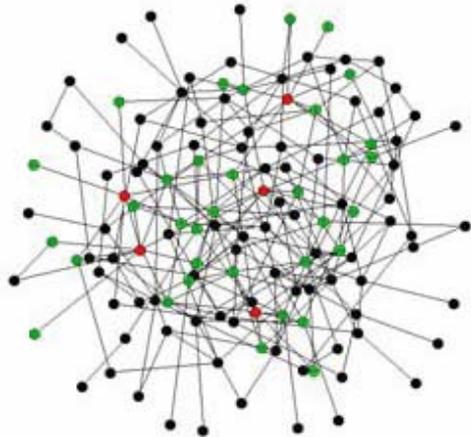


NIPS: Neural Information Processing Systems

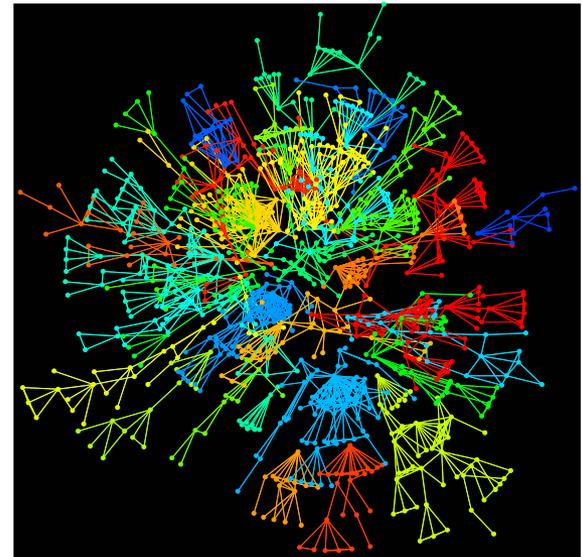
統計的機械学習やニューラルネットワークに関する国際会議における共著者ネットワーク

コミュニティ生成の考え方

相手を選ばず、ランダムにリンクを張ると、コミュニティはできない。



「好き嫌い」等で、リンクする相手を選別すると、コミュニティができる

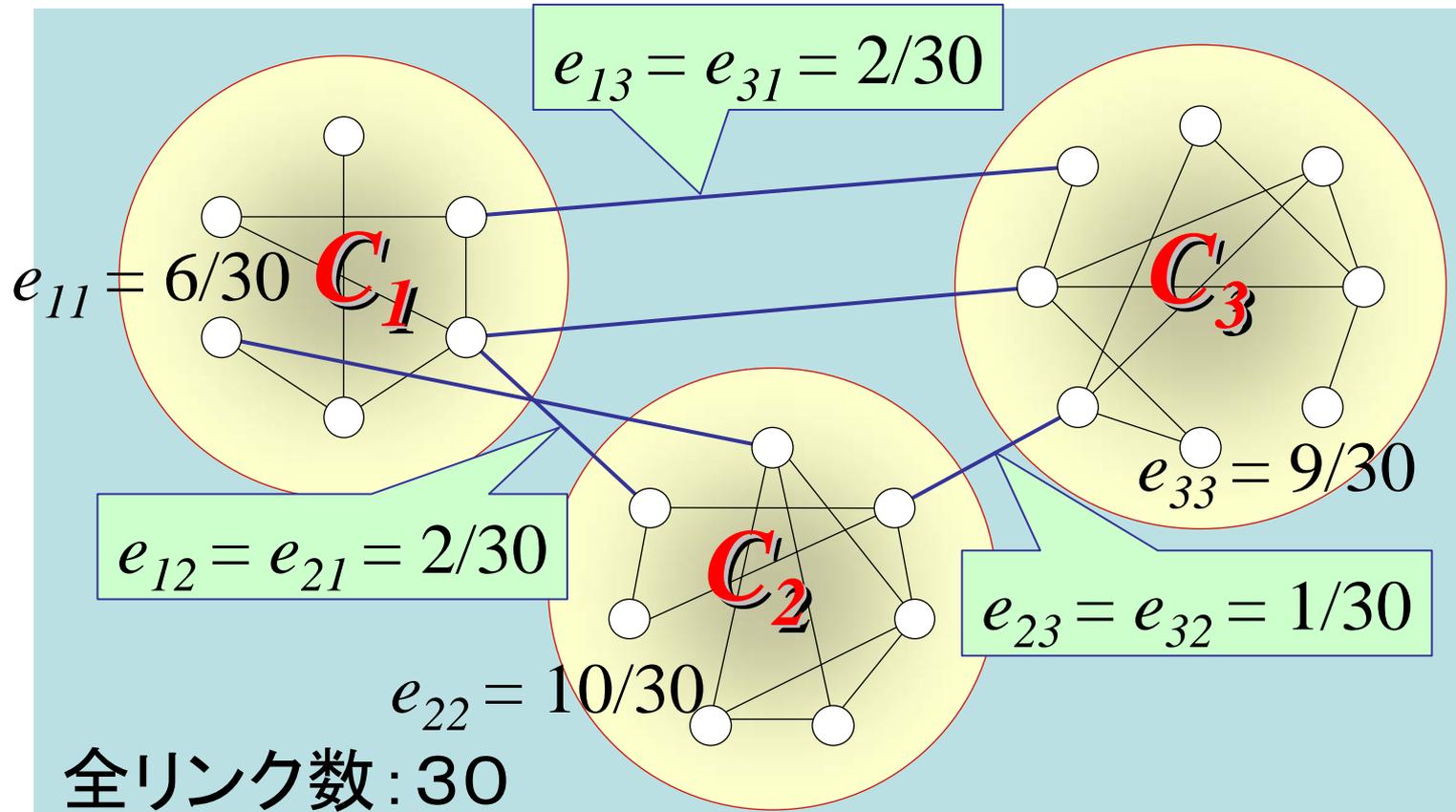


Aさん(のグループ)とは友達になるが、Bさん(のグループ)とは友達にならない、など。

コミュニティ間のリンク確率

コミュニティ i, j 間のリンク生成確率

$$e_{ij} = P(i, j) = \# \text{links}(i \leftrightarrow j) / \# \text{links}(\text{all})$$



Newman の Modularity (モジュール度)

同一コミュニティ内のリンクは密、コミュニティ間のリンクは疎

コミュニティ i, j 間のリンク生成確率

$$e_{ij} = P(i, j) = \# \text{links}(i \leftrightarrow j) / \# \text{links}(\text{all})$$

コミュニティ i のリンク生成確率

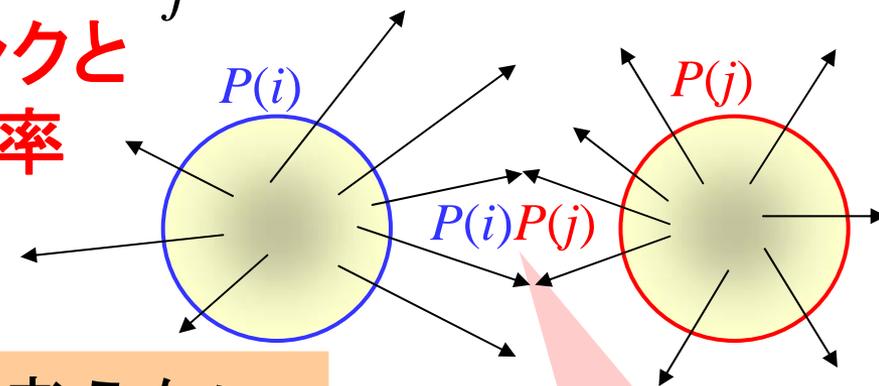
$$a_i = P(i) = \sum_j P(i, j) = \sum_j e_{ij}$$

相手への「好き嫌い」で異なる

独立に生成した C_i からのリンクと C_j からのリンクが一致する確率

$$a_i a_j = P(i) P(j)$$

相手への「好き嫌い」を考えない



たまたま一致

Newman の Modularity (モジュール度)

同一コミュニティ内のリンクは密、コミュニティ間のリンクは疎

相手の「好き嫌い」で異なる: 好き $\Rightarrow e_{ij}$ 大

相手の「好き嫌い」を考えない(たまたま)

$$Q_{ij} = e_{ij} - a_i a_j$$

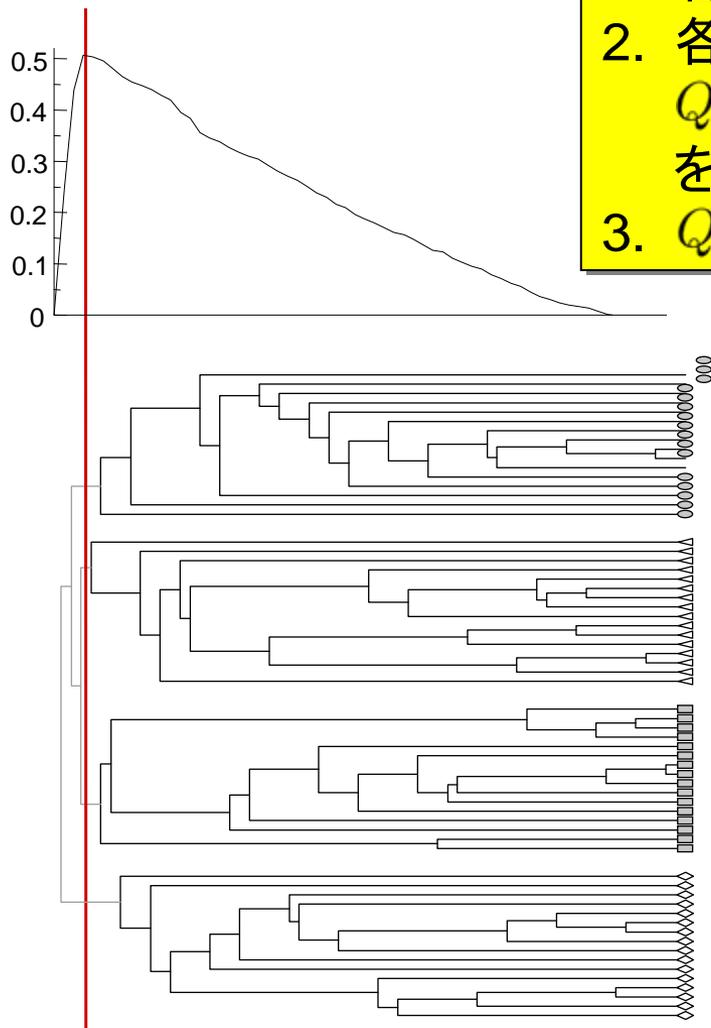
同じコミュニティには「好きな人」同士が集まっている

Modularity (モジュール度)

$$Q = \sum_i Q_{ii} = \sum_i (e_{ii} - a_i^2) \rightarrow \text{最大}$$

Newman と Girvan によるクラスタリング結果

Q: Modularity (モジュール度)



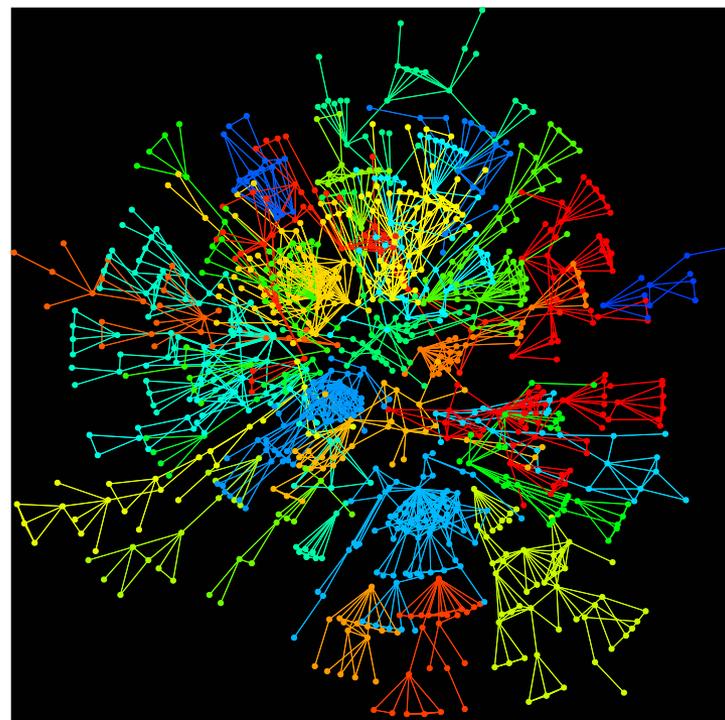
アルゴリズム (ボトムアップ方式)

1. 各ノードが独立なクラスタである状態からスタート
2. 各クラスタペアについて

$$Q(C_1, C_2) := Q(C_1 \cup C_2) - Q(C_1) - Q(C_2)$$

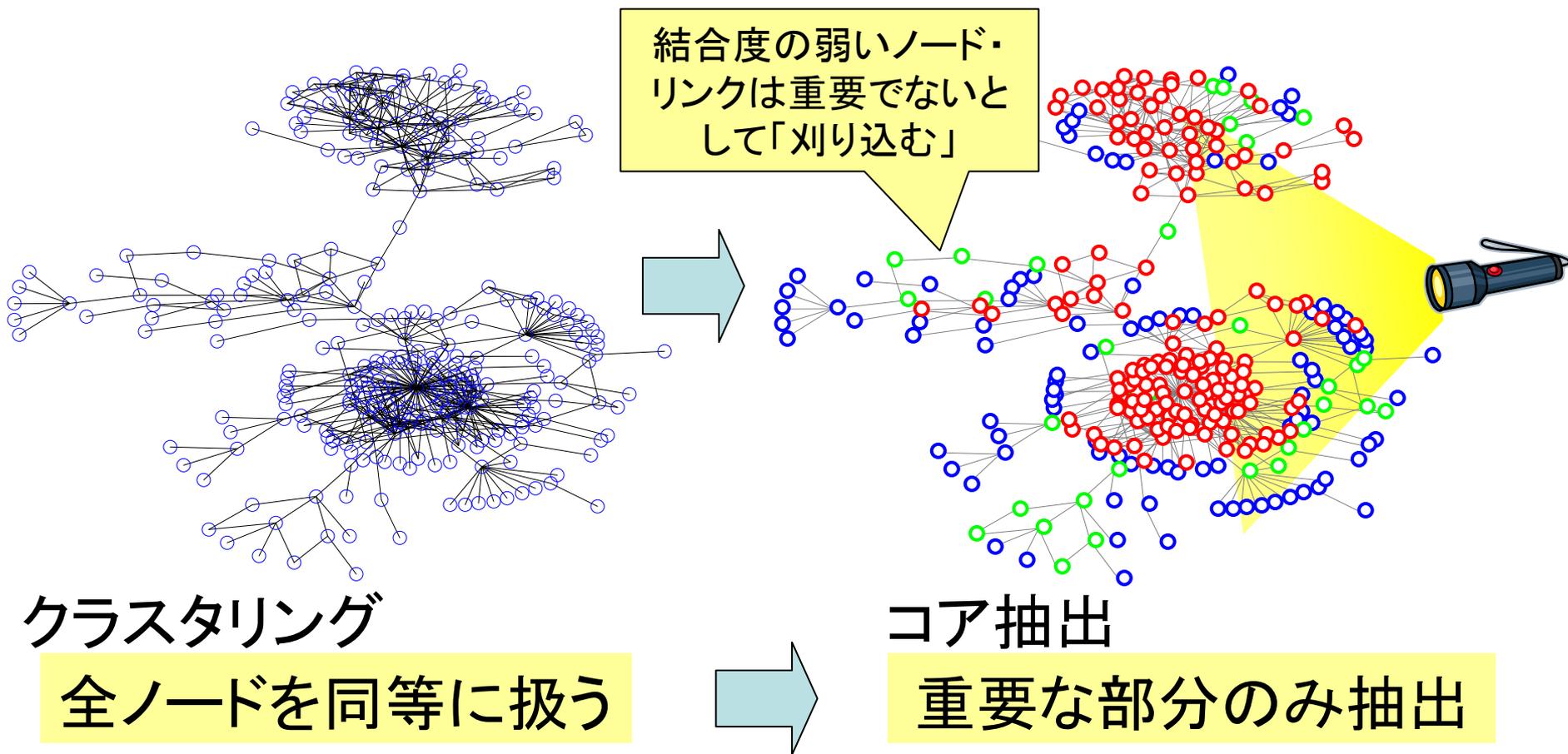
を計算

3. $Q(C_1, C_2)$ が最大となるクラスタペアをマージ



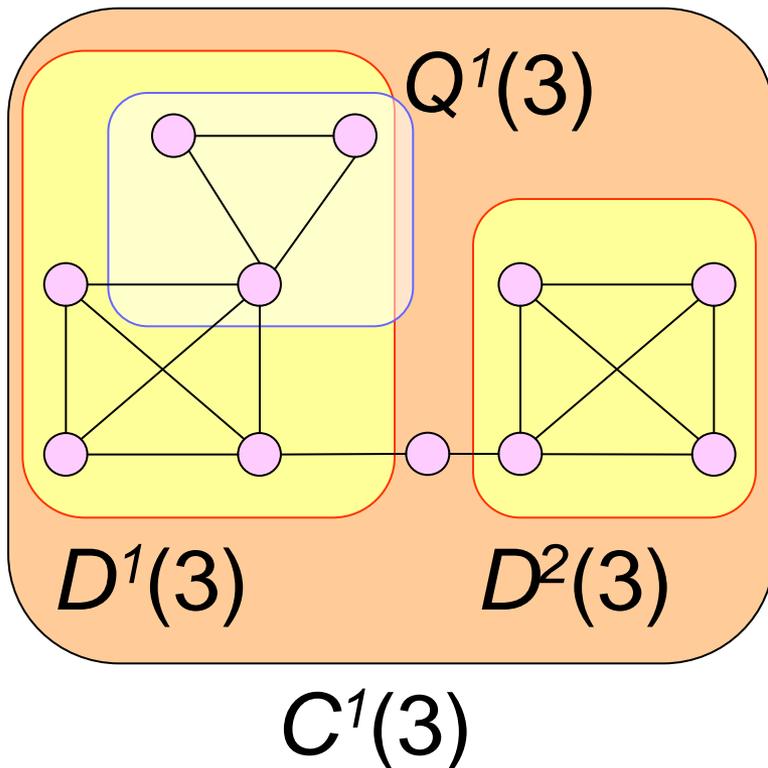
k-dense に基づくコミュニティ抽出法

密結合するノード群(コミュニティ)に着目し
ネットワークの**主要構造**を理解



k-core, k-dense, k-clique の関係

$$k\text{-clique } (Q) \subseteq k\text{-dense } (D) \subseteq k\text{-core } (C)$$



- k-core: コミュニティに所属するノードは k-1個以上の隣接ノード(知り合い)を持つもの⇒抽出精度(粒度)低い
- k-clique: 完全結合(全員が知り合い)サブネットワーク⇒計算コスト高い
- k-dense: 同一コミュニティに所属するリンクはk-2個以上の共通の知り合いを持つ⇒計算効率と精度を両立

- k-denseは(より少ない計算量で) k-clique を近似する概念
- 実は、高次k-denseまで考えるとk-dense は k-core と k-clique を自然に補間

ネットワーク解析ツールによるブログネットワークの解析

ブログトラックバックネットワーク

『ホット』な話題を抽出できる

Menu 統計グラフ表示 SR結果表示 ランキング表示 可視化表示 グラフ表示 コミュニティ表示

データ名: topic034 (ノード数: 12047, リンク数: 39960)

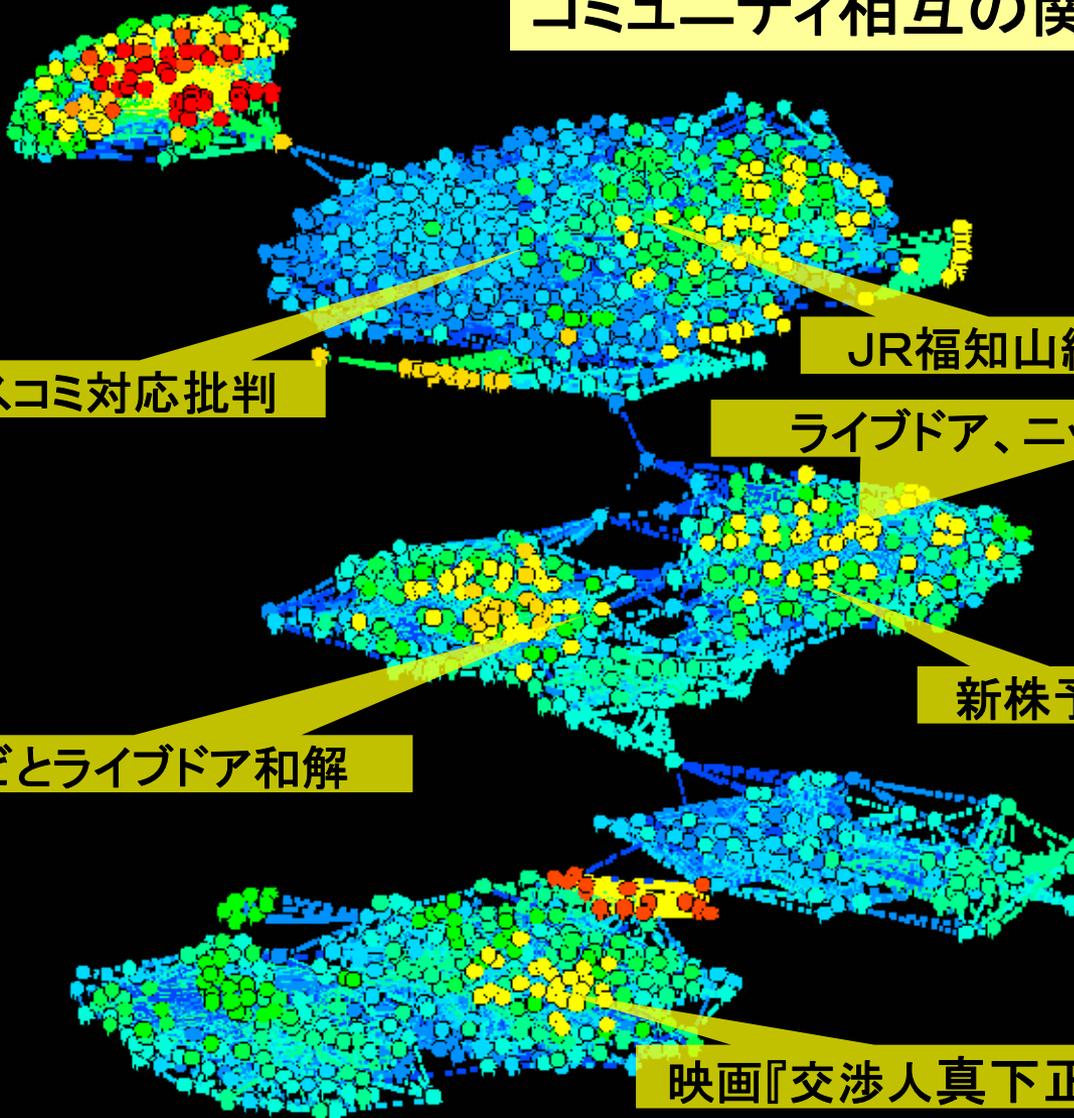
表示

グループ1 (102 ノード)	グループ3 (56 ノード)	グループ4 (43 ノード)	グループ5 (24 ノード)	グループ6 (24 ノード)	グループ7 (20 ノード)
番号	番号	番号	番号	番号	番号
2567 交渉人	2497 フジ・ライブド	63 急ぎすぎている	6389 ニッポン放送陥落	6511 仮処分	88 報道機関のつっこみどころ
3422 『交渉人 真下正義』	3215 『フジテレビと	103 JR史上最悪の脱	6390 フジ系列局 フジ	8918 今後の展開は?	89 JR東海線脱線事故にバ
3423 交渉人 真下正義	3216 フジTVとライ	136 JR福知山線で	6391 ライブドアとフジ	8925 ライブドア ニッ	93 報道機関に対する私の真
3424 交渉人 真下正義	3221 ライブドア・フ	140 電車脱線、マン	9119 大勝負	8927 新株予約権、差し	238 曲がりきれなかったカー
3425 交渉人 真下正義	3225 ライブドアとフ	144 この悲劇から何	9225 05/03/17 livedo	8932 焦土作戦?	240 曲がりきれなかったカー
3429 ★★★「交渉人	3228 フジ、ライブド	146 兵庫県尼崎市の	9230 ライブドア、フジ	9115 新株予約権、発行	271 献花に行ってきた
3430 幸薄いって友達	3233 「ライブドア	147 死亡50人、負	9235 ライブドア 堀江	9116 有力株主村上氏が	272 連帯責任の国
3431 交渉人真下正義	3235 ライブドア問題	150 JR西日本福知山	9236 ライブドア、フジ	9172 05/03/11 U・M	274 ポウリングイメージの低
3432 交渉人 真下正義	3236 祝!!ライブドア	155 JR尼崎の事故	9243 ライブドア、フジ	9177 よく分かる(?)	275 JR西日本叩き 我々は
3433 交渉人 真下正義	3237 きっかけはホリ	173 電車脱線、50	9245 ニッポン放送株5	9190 東京電力で株主代	277 バッシングで終わらせな
3435 交渉人 真下正義	3238 「ライブドア、	470 兵庫・尼崎で列	9246 ライブドアがフジ	9222 新株予約権発行の	278 JR福知山線事故(2)…マ
3436 交渉人 真下正義	3241 ライブドアとフ	476 電車脱線	9247 ライブドア、フジ	9249 livedoorとりあえ	280 JR西日本脱線事故の報道
3437 映画『交渉人 真	3242 共同会見ですっ	477 電車脱線	9248 ライブドアとフ	9473 ライブドアVSフ	282 マスコミのJR西日本叩
3439 交渉人真下正義	3243 ライブドア vs.	479 今朝	9254 4753ライブドア	9559 ライブドア・フジ	290 JR脱線事故報道の姿勢
3441 交渉人真下正義	3246 和解	480 福知	9255 ライブドア、フジ	9563 ニッポン放送新株	291 愛国無罪、報道無罪。
3442 交渉人 真下正義	3247 フジテレビとラ	481 オール	9261 今日のライブドア	9568 ライブドア勝訴で	293 鉄道(脱線事故のその後
3443 交渉人 真下正義	3248 livedoorとフジ	497 JR福知山線	9262 ひどすぎる、ニッ	9571 4753ライブドア	295 マスコミの不適切な行動
3444 安易なスピコン	3249 ライブドアとフ	490 JR福知山線の	9264 レバレッジド・パ	9573 正義無き戦い 第	297 アンケート (JR西日本)
3446 ■交渉人 真下	3250 これで決	494 JR福知山線 脱	9265 ライブドア VS	9574 きっとあなたも誤	349 JR西日本は馬鹿だが、そ
3447 交渉人 真下正義	3251 ライブ	568 尼崎での大惨事	9266 ライブドア 堀江	9578 東京地裁「新株発	1054 落ちたマスコミ
3448 交渉人 真下正義	3252 ライブ	569 【続報】兵庫	9267 ライブドア VS	9581 新株予約権発行差	
3449 交渉人 真下正義	3253 ライブ	575 「未来に夢	9268 ライブドア 堀江	9584 ホリエモン第一段	
3450 交渉人 真下正義	3254 ライブ		9269 ライブドア・フジ	9587 新	
3451 交渉人 真下正義	3255 なんですか			9620 3/	
3452 交渉人 真下正義	3256 ライブ				
3453 交渉人 真下正義	3257 ホリエモ				

⇒ ノード数: 12,047、リンク数: 39,960

ブログネットワークのコミュニティ構造の例

コミュニティ相互の関係を抽出する



マスコミ対応批判

JR福知山線事故

ライブドア、ニッポン放送、フジテレビ

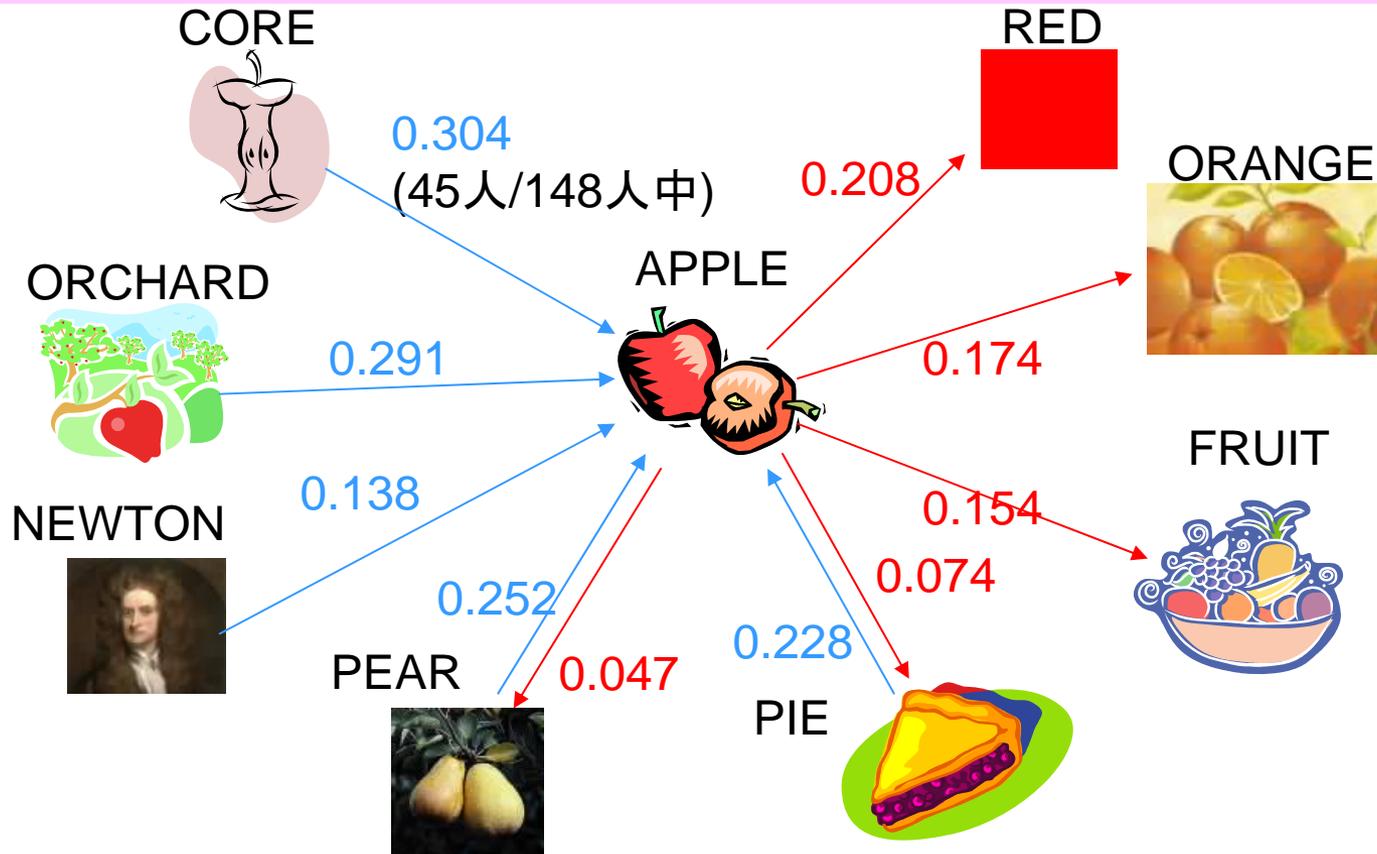
新株予約権

フジテレビとライブドア和解

映画『交渉人真下正義』

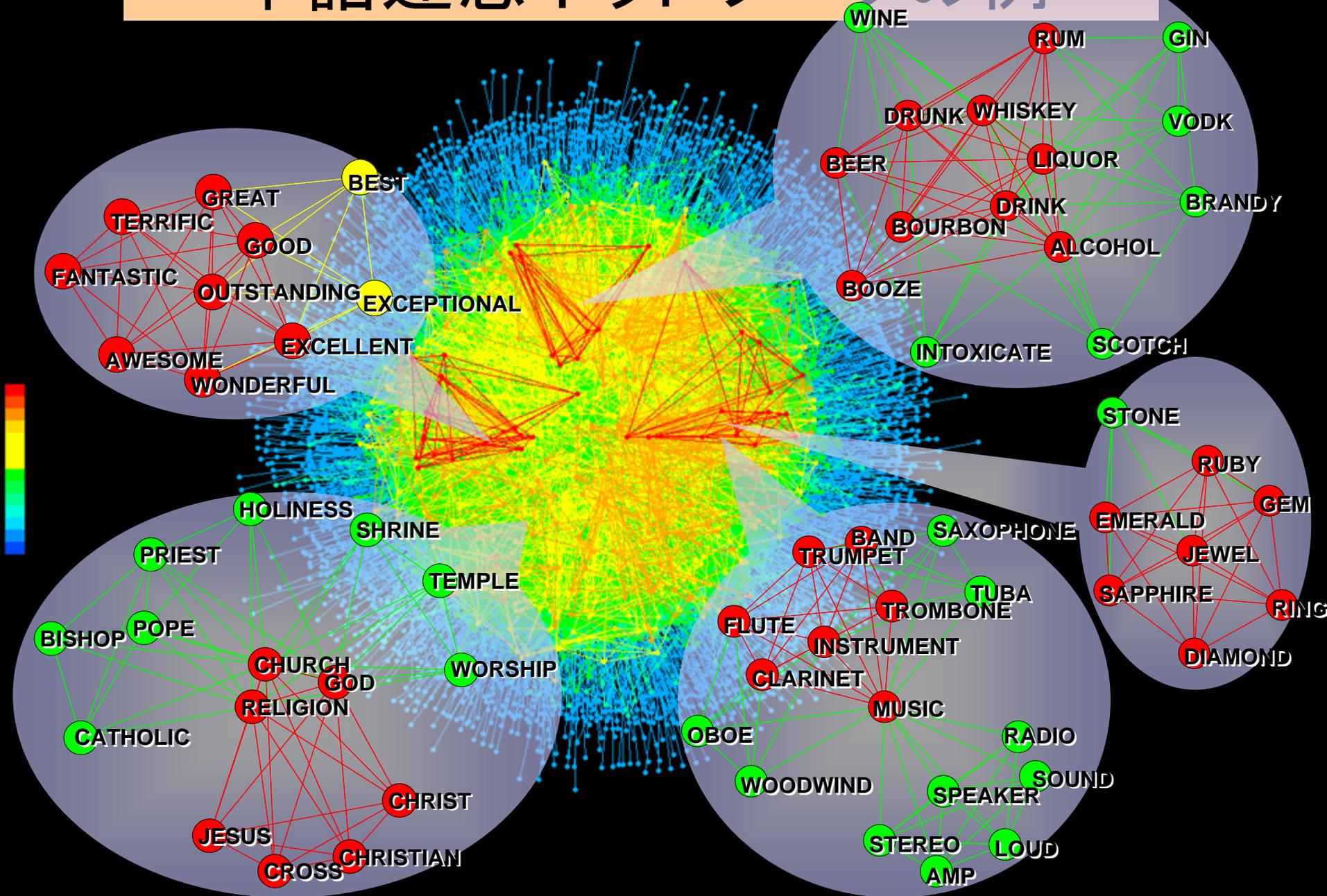
単語連想ネットワーク

元データは、南フロリダ大学心理学科のプロジェクトで作成
被験者数: 平均149人(±15程度)
被験者にキューとなる単語を与え、関連する単語を連想してもらう

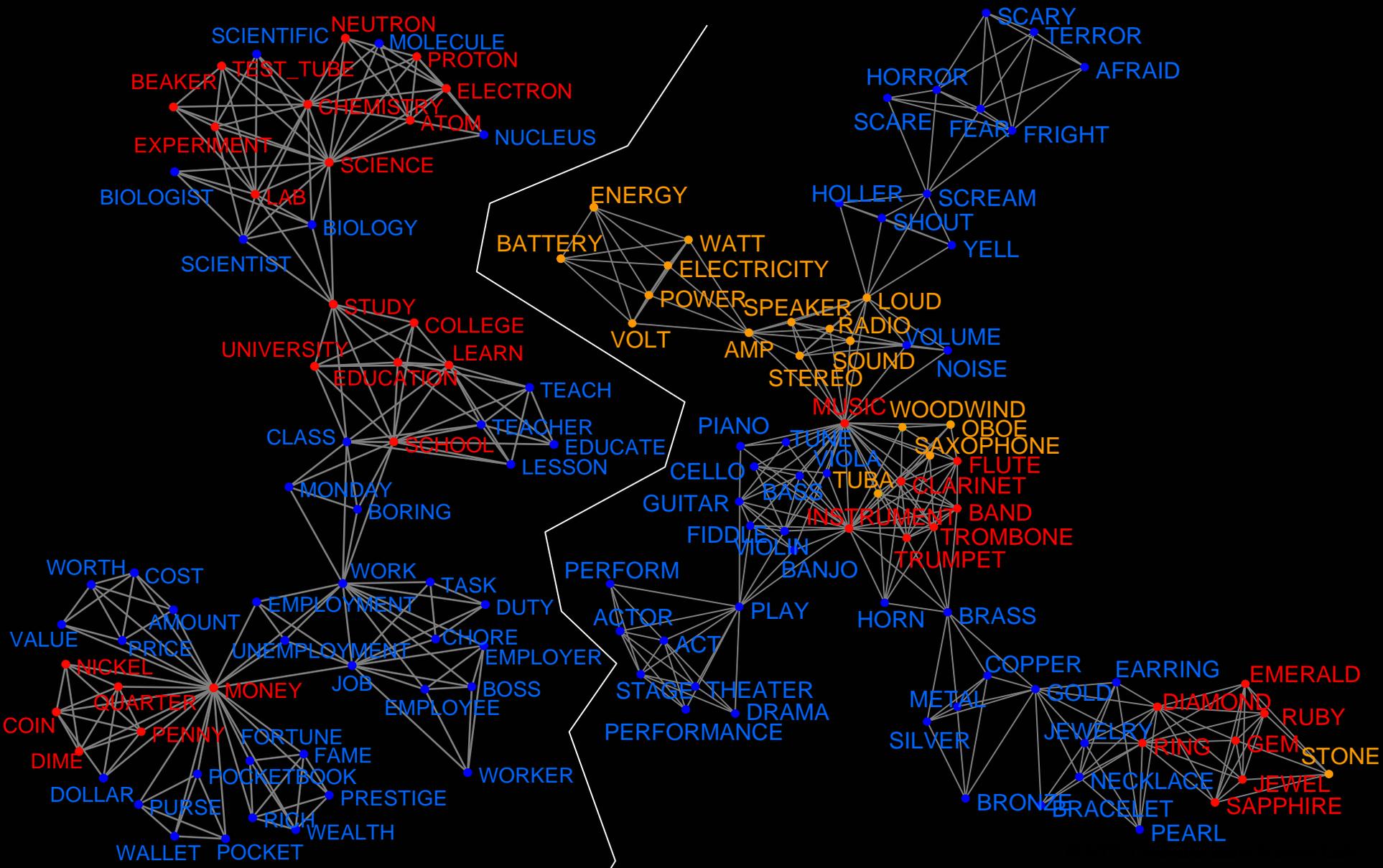


双方向のリンクの強さの和を一定の閾値(0.025)で無向化
⇒ ノード数 : 7,207、リンク数 : 31,784

単語連想ネットワークの例



単語連想ネットワークの例(2)



スモールワールド

スケールフリー

スパース(疎)

これらの特徴
を生かして

ランキング

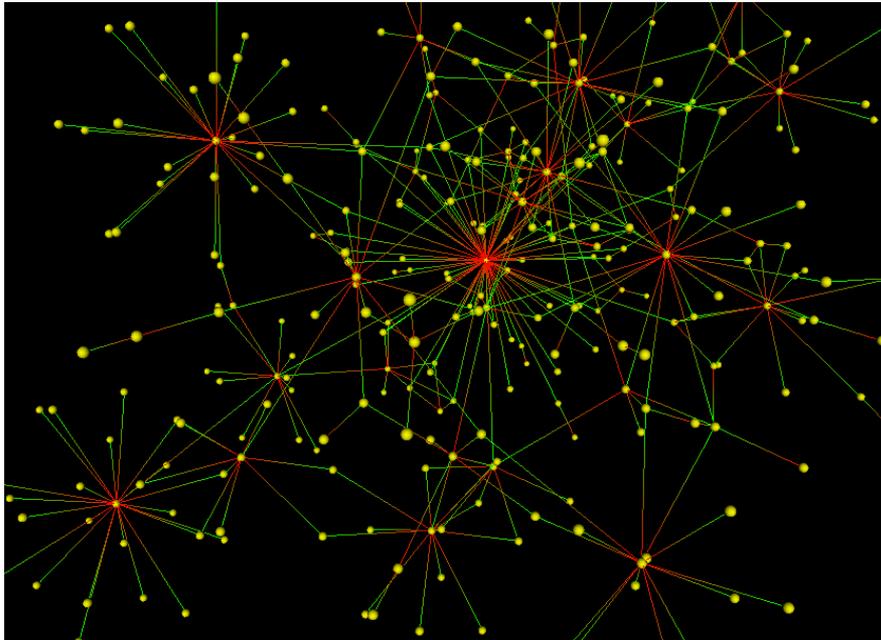
クラスタリング
／コア抽出

共起NW解析

可視化

直感的な理解：ネットワークの可視化

低次元空間への投影により、**潜在的な構造を浮き彫りにする**



- 全体像の把握
- 潜在的な構造の解明
- 体験型ネットワークブラウザ？

➤ 潜在的なNW構造の顕在化

- スケールフリー性
- クラスタ構造
- 階層構造

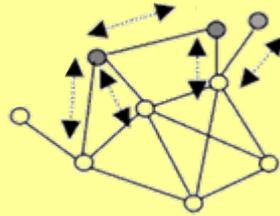
- ⇒ これらの構造を浮き彫りにする
- ⇒ 知識発見の手がかりを見つける

Graph Drawing by Force-directed Placement

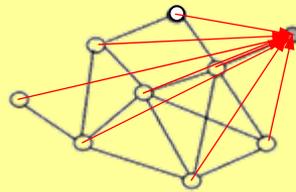
- ノード間に引力、斥力を与える
- 冷却関数を用いてアニーリング

attractive + repulsive force

リンク上のノード f_a
 引力 $O(E)$



ノード \leftrightarrow ノード f_r
 斥力 $O(N^2)$



[Fruchterman & Reingold (1991)]

$$f_a(d) = \frac{d^2}{K}$$

$f_a + f_r$

0

K

$$f_r(d) = -\frac{K^2}{d}$$

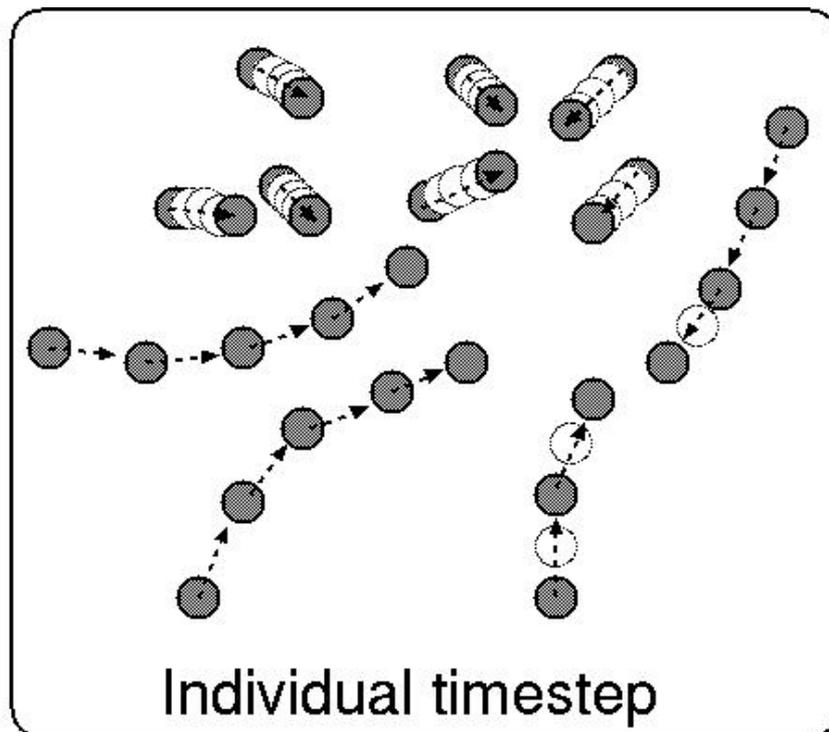
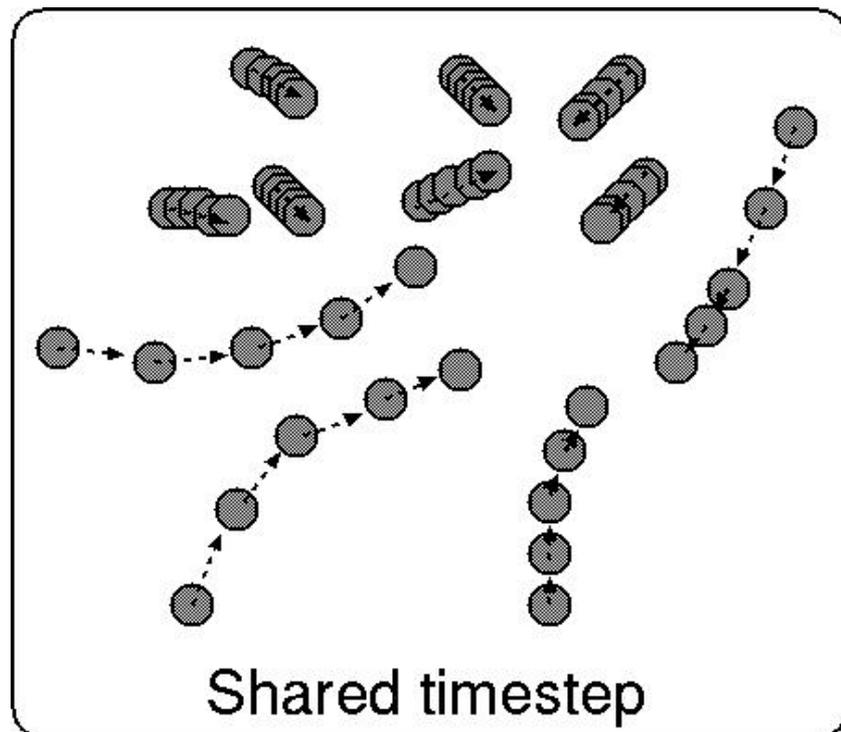
$$\vec{a}_i = \frac{1}{K} \sum_j^{E_i} \|\vec{x}_{ij}\| \times \vec{x}_{ij} - K^2 \sum_{j \neq i}^N \frac{\vec{x}_{ij}}{(\|\vec{x}_{ij}\|^2 + \varepsilon^2)}$$

$$x_i(t+1) = x_i(t) + \vec{a}_i$$

$$E_{FR} = \frac{1}{3K} \sum_{i,j \in E} \|\vec{x}_{ij}\|^3 - \frac{K^2}{2} \sum_i^{N-1} \sum_{j \neq i}^N \log(\|\vec{x}_{ij}\|^2 + \varepsilon^2)^{\frac{1}{2}}$$

階層的独立固有時間刻み法による高速化

天体力学の手法を可視化計算に拡張



従来法

全ノードを常に更新

提案法

- 加速度の大きいものを優先的に更新
- 加速度の小さいものは時間刻みを大きく

階層的時間更新方法

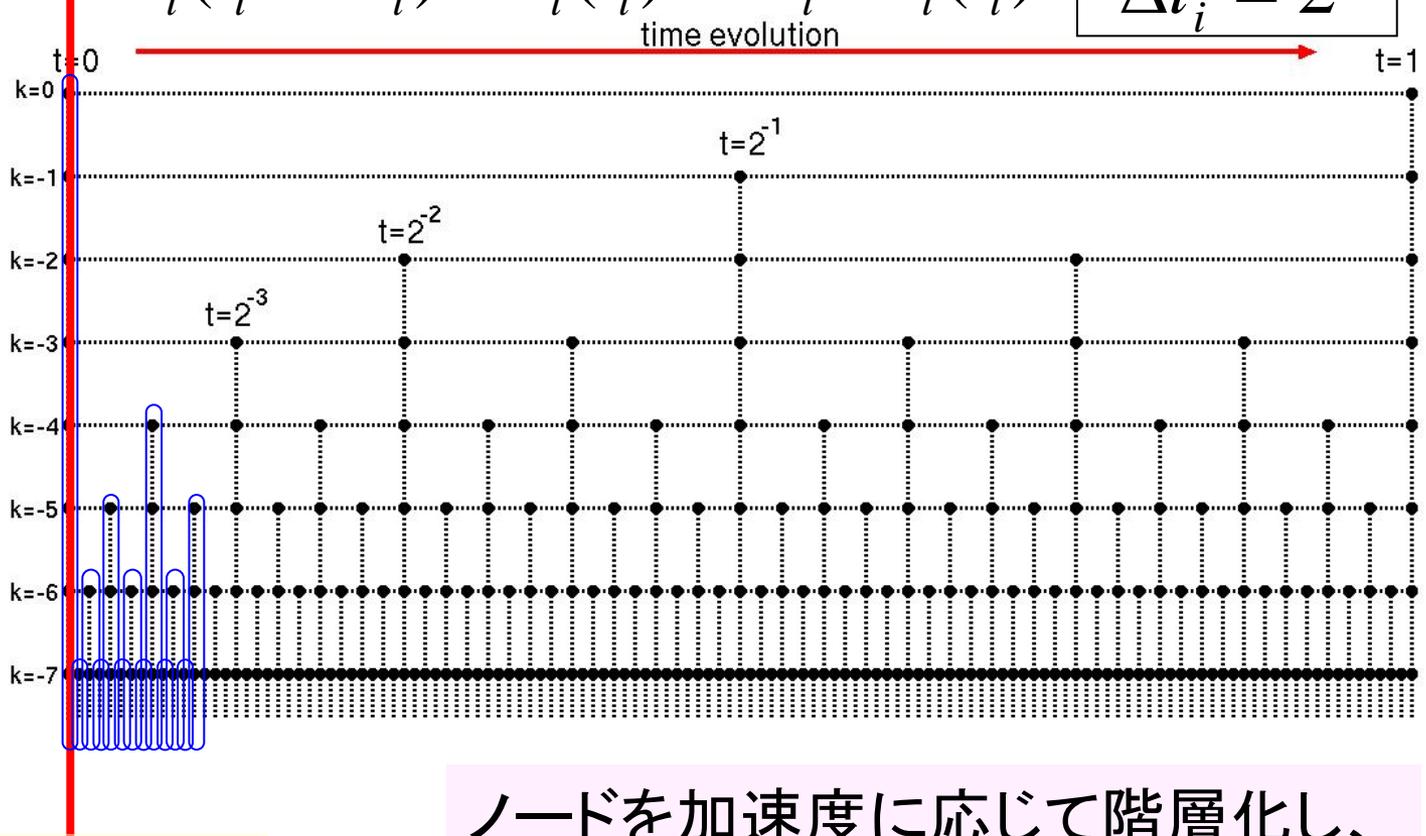
$$x_i(t_i + \Delta t_i) = x_i(t_i) + \Delta t_i \times a_i(t_i)$$

$$\Delta t_i = 2^k$$

time evolution

a が小さい
頻度が少ない

a が大きい
頻度が多い



ノードを加速度に応じて階層化し、階層ごとに更新頻度を設定する

時刻 t で同期するノード数 n_s のみ加速度の計算が行われる

計算量

従来法 $O(N^2)$
提案法 $O(n_s \times N)$

並列化による大規模ネットワークへの対応

リンク

さらに

ノード↔ノード

加速度

$\mathbf{a}_i =$

$$\frac{1}{K} \sum_j^{E_i} |x_{ij}| \times \vec{x}_{ij}$$

—

$$K^2 \sum_{j \neq i}^N \frac{\vec{x}_{ij}}{(x_{ij}^2 + \varepsilon^2)}$$

エネルギー

$E =$

$$\frac{1}{3K} \sum_{(i,j) \in E} |x_{ij}|^3$$

—

$$\frac{K^2}{2} \sum_i^N \sum_{j \neq i}^N \log |x_{ij}^2 + \varepsilon^2|^{1/2}$$



Position etc.

Force etc.

Data Transfer

Attractive force
Position Integration etc.

$O(E)$ or $O(N)$ Calculations



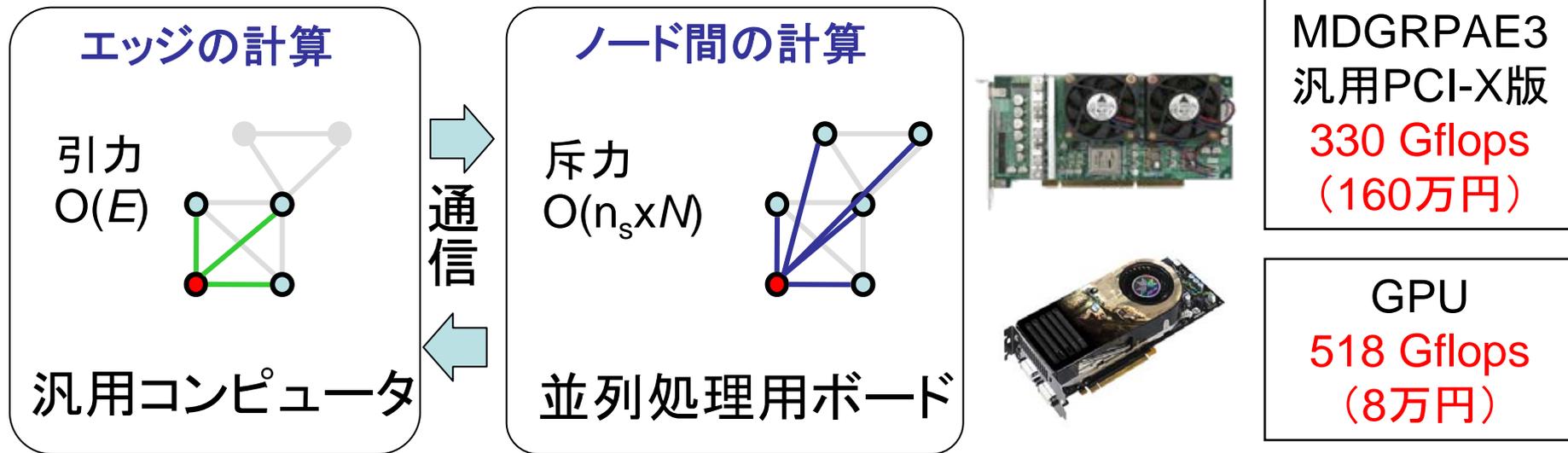
Repulsive Force Calculation

$O(N^2)$ Calculations

並列処理のポイント

固有時間の量子化により、同期するノードを並列処理

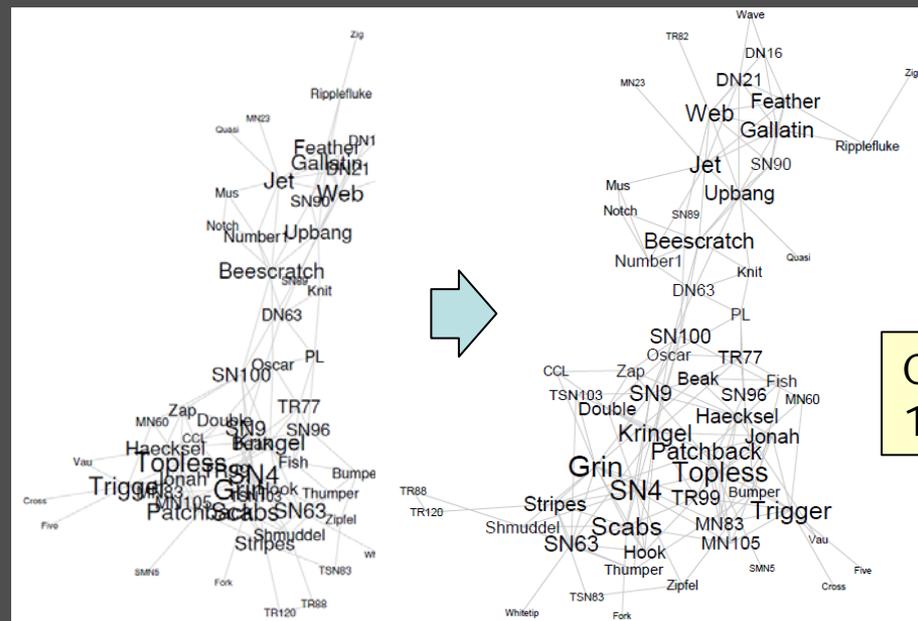
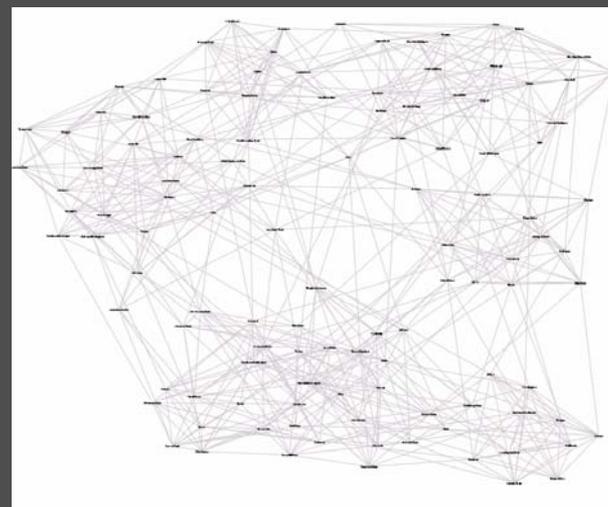
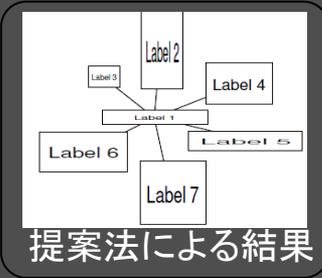
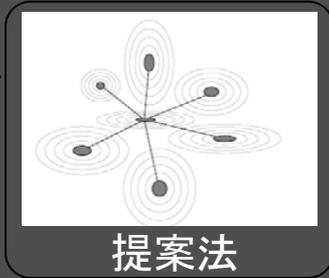
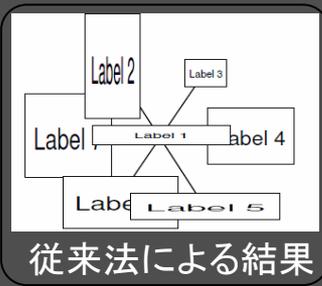
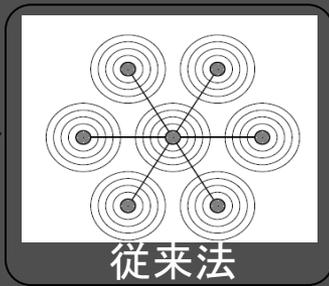
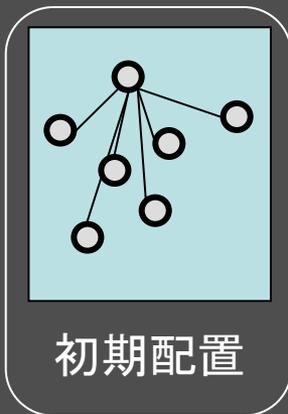
加速度計算の特性を生かし、ホストと専用ボードに処理を分割



計算コストの99%以上がノード間の斥力計算 → ノード間の処理を並列化

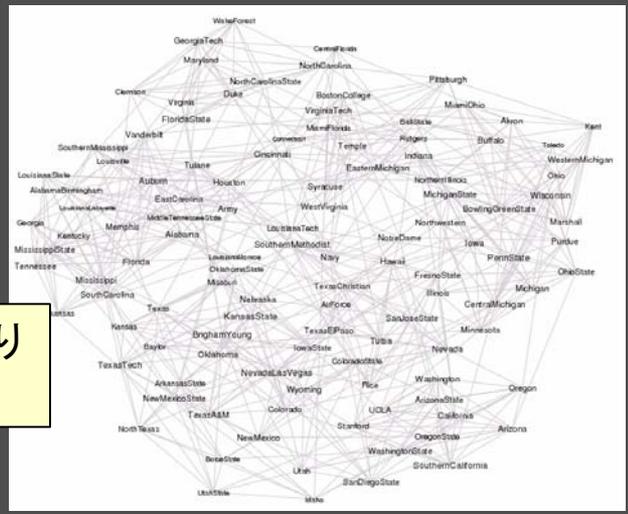
並列処理に特化したボードへの実装により、数百倍の高速化

楕円ポテンシャルによるラベル付きグラフの可視化



GPGPUによる並列化

CPUによる処理より
100倍高速に



BLOGRANGER TG

巨大な「タグクラウド」をスクローラブルマップとして提供

【従来型】

検索結果が1次元的

- タグが多く読めない
- 類似したものが離れる

【NTT新技術】

類似したタグの2次元配置

- スクローラブルマップ
- 新しい興味の発見
- 新しいコンテンツサービスの形態

<http://ranger.labs.goo.ne.jp/>



キーワードを入力して探検ボタンをクリックしてください。そのキーワードを含むブログ記事からタグを探します。

キーワード: 探検

または、次のタグをクリックして、仮想大陸上のそのタグの地点に移動してください。

タグ: [スポーツ](#) [サッカー](#) [野球](#) [芸能](#) [テレビ](#) [音楽](#) [ライブ](#) [政治](#) [パソコン](#) [グルメ](#) [料理](#) [健康](#) [育児](#) [学](#)

仮想大陸上のタグをクリックすると、そのジャンルのブログ記事を表示します。

タグ: [料理](#) の検索結果: 1107
キーワード: なし



1. [レストラン](#) [食べる](#) [料理](#)
2. [サラダ油は必須](#)
◎ [気ままに旅日記](#) 15
人間もそうだけど油が足りず、近所中につけを見つけた。体に湿疹が出来てカサカサ。タグ: [レシピ](#) [食糧](#) [料理](#)
3. [離乳食とか歯とか子育て](#)
◎ [パンダモバイルアーマー](#)
こんにちわ！久しぶりの更新です。息子ハルやん早え。もう半年か。今月にタグ: [離乳食](#) [育児](#) [料理](#)
4. [最近ハマってるもの](#)
◎ [シングルママ生活](#) 21
最近ぬれせんべいにはポリポリするのが好きだったぬれせんべいを食しまつてタグ: [食べ物](#) [グルメ](#) [料理](#)
5. [「わかさぎ」の唐揚げ](#)

[この地図のリンク](#) [ブックマーク](#) [メール送信](#)

<http://ranger.labs.goo.ne.jp/TG/?MT=&tag=%E6%96%99%E7%90%86>

タグの大きさ: 記事数 (多い) 例 スポーツ ↑ ↓ (少ない) スポーツ スポーツ	タグ間の距離: 類似度 (近い) 例 スポーツ-野球 ↑ ↓ (遠い) スポーツ-----政治	タグの高さ: 話題の広さ (広い) 例 スポーツ ↑ ↓ (狭い) 野球 中日ドラゴンズ
--	---	---



まとめ

大規模で複雑なネットワークの主要構造を自動抽出して、その骨組みや機能を**理解し、有効利用するための方法論の構築**

基礎

スモールワールド
スケールフリーネットワーク

一部グラフ、無向グラフ

クラスタリング法

全ノードが重要として、それらを幾つかのコミュニティに分割し、ネットワークの全体構造を理解⇒ N&Gクラスタリング

コア抽出法

密結合するノード群(コミュニティ)に着目しネットワークの主要構造を理解⇒ k-dense コア抽出

可視化

⇒ Kamada&Kawai バネモデル
⇒ Fruchterman&Reingoldモデル

有向グラフ

ランキング

⇒ PageRank
⇒ HITS

二部グラフ、有向グラフなど、より複雑な関係データへの拡張

共起ネットワーク

⇒ 二部→一部

ブロックモデル

⇒ IRM

[トップページ](#)
[ごあいさつ](#)
[プログラム](#)
[アクセス](#)
[お問い合わせ](#)
[個人情報の取り扱いについて](#)

What's New

2008/4/14

[ごあいさつ](#)、[プログラム](#)、[どんな展示があるの?](#)、[会場案内](#)を公開致しました。

2008/3/17

オープンハウス×未来想論2008のホームページを開設致しました。

NTT コミュニケーション科学基礎研究所

オープンハウス × 未来想論2008

5月29日(木) 12:00 - 17:30

5月30日(金) 10:00 - 17:30

革新と調和を目指すコミュニケーション科学

会場:NTT 京阪奈ビル
(けいはんな学研都市)

入場無料・事前登録不要

オープンハウス×未来想論2008とは

NTT コミュニケーション科学基礎研究所では、皆様に最新の研究成果を体感し身近に感じていただく機会として「オープンハウス×未来想論2008」を開催いたします。革新と調和をキーワードに、最新の研究成果を体感していただく展示と、コミュニケーションの未来を皆様とともに議論する「未来想論」を行います。皆様ぜひご来場いただき、コミュニケーション科学の持つ可能性に触れてください。

[NTT コミュニケーション科学基礎研究所 公式ホームページ](#)

オープンハウス×未来想論2008って

どんな展示があるの？

会場案内

会場をマップで紹介

参加方法

入場無料・事前登録不要

※お名刺を2枚お持ち下さいますようお願い申し上げます。(うち1枚は名刺として使用させていただきます。)

所長講演&未来想論
5/29,30 ストリーミングLive配信
詳細は近日公開予定 **決定!!**

所長講演



感・知・心で極める
コミュニケーションの未来

外村 佳伸

5/29 13:00-13:40

未来想論

テーマ展示

場とコミュニケーション

- 環境知能
- 裏世界コミュニケーションシーンを理解する音響映像技術

メディアとコンテンツ

- メディアコンテンツを縦断に特定
- 世界中のメディアで語る、学ぶ、楽しむ