つながりのビッグデータ解析

人間関係ネットワークの科学と活用

国立情報学研究所 特任助教

秋葉 拓哉

2016/06/22　@ 国立情報学研究所 市民講座

自己紹介

秋葉 拓哉

所属：国立情報学研究所 特任助教 (2015-)

専門：大規模グラフ解析の高速アルゴリズム開発

最新論文

- IJCAI 2016：グラフの全域木中心性の高速計算
- VLDB 2016：グラフの媒介中心性の高速計算
- KDD 2016：グラフの乱択スケッチの省スペース化

目次

1. グラフ解析入門

- どのようなグラフデータが有るのか？
- なぜそれを解析をするのか？

2. グラフ解析の標準的手法

- グラフはどういった形をしている？
- 中心性，関連度，コミュニティ，全体の性質

3. 大規模グラフ解析の課題と研究動向

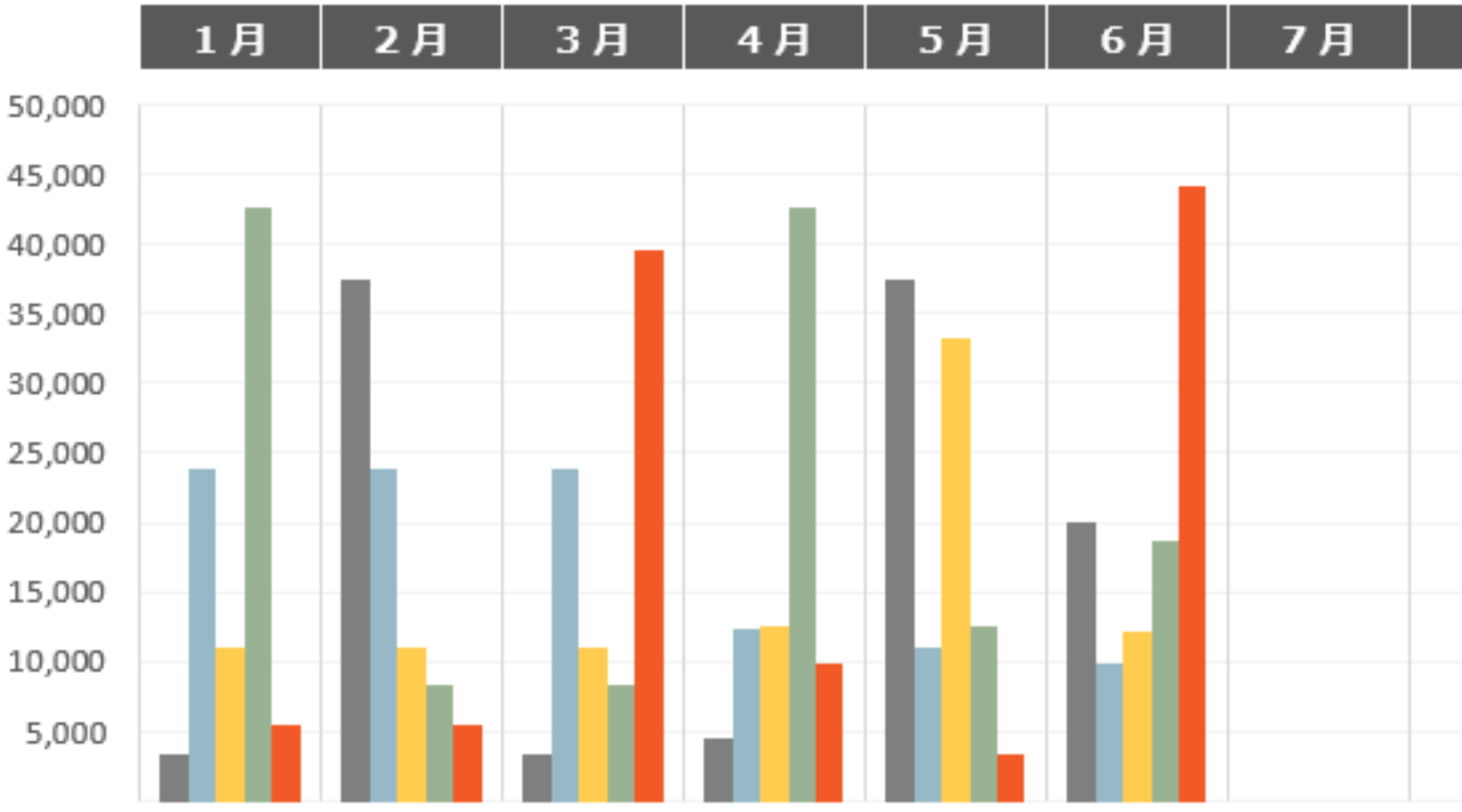
1

グラフ解析入門

どのようなグラフデータが有るのか？
なぜそれを解析するのか？

グラフ

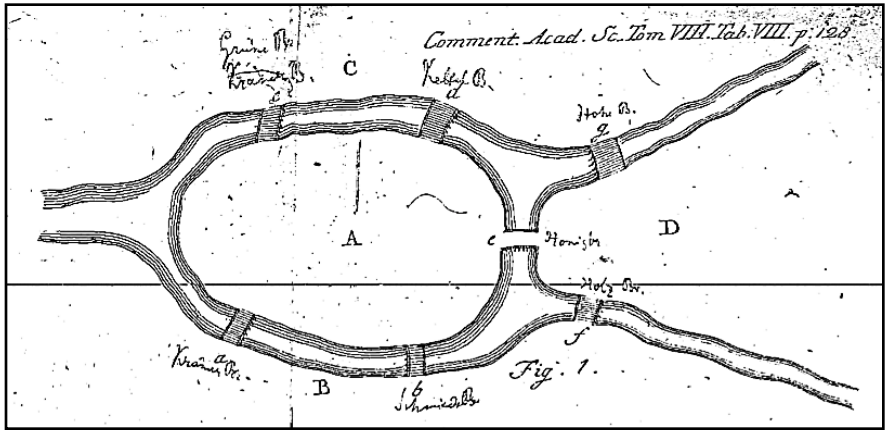
グラフとは……

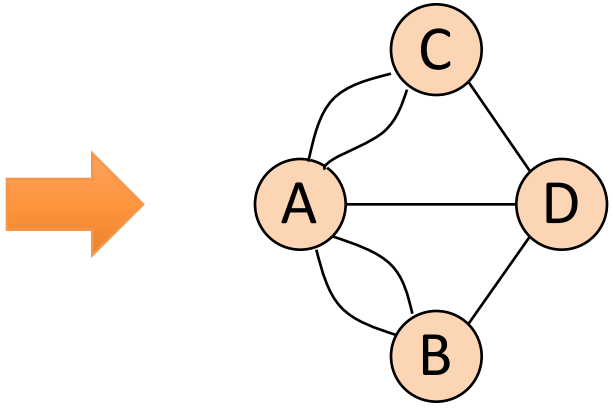


支出	1月	2月	3月	4月	5月	6月	7月	8
支出 1	3,300	37,500	3,300	4,500	37,500	20,100	0	
支出 2	23,800	23,800	23,800	12,300	11,100	9,800	0	
支出 3	11,000	11,000	11,000	12,500	33,300	12,200	0	

支出	1月	2月	3月	4月	5月	6月	7月	8
支出 1	3,300	37,500	3,300	4,500	37,500	20,100	0	
支出 2	23,800	23,800	23,800	12,300	11,100	9,800	0	
支出 3	11,000	11,000	11,000	12,500	33,300	12,200	0	

グラフ＝関係のネットワークを抽象化するモデル



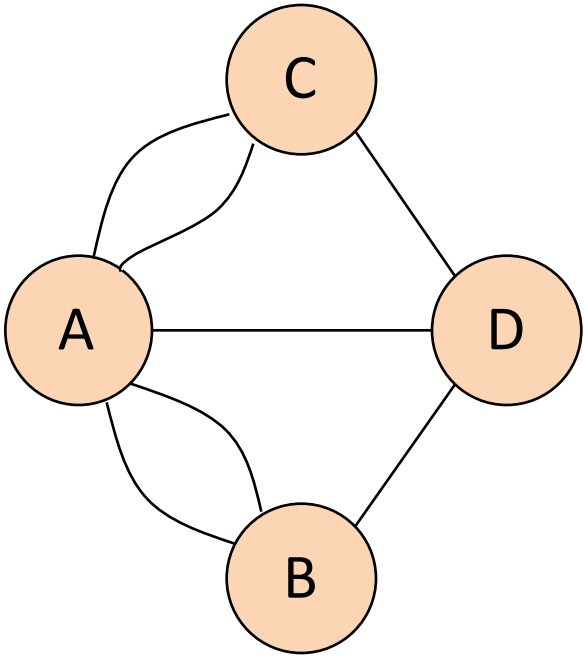


レオンハルト・オイラーによるグラフ理論の始まり

[Commentarii Academiae Scientiarum Petropolitanae, vol.8, p.129, 1741]

グラフ

$G = (V, E)$



V: 頂点集合

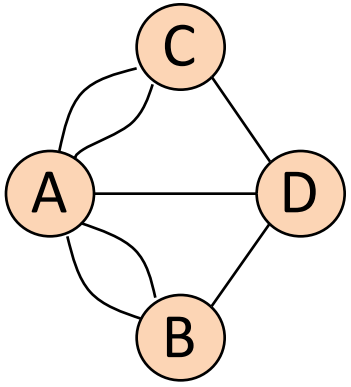
E: 辺集合

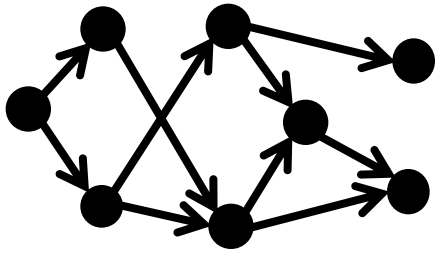
グラフ

無向 / 有向

辺に方向があるか無いか

- 無向：Facebook, LinkedIn
- 有向：Twitter, 文献引用





重み有り / 無し

辺に数値が付加されているか否か

- 重み有り：交通グラフの移動時間

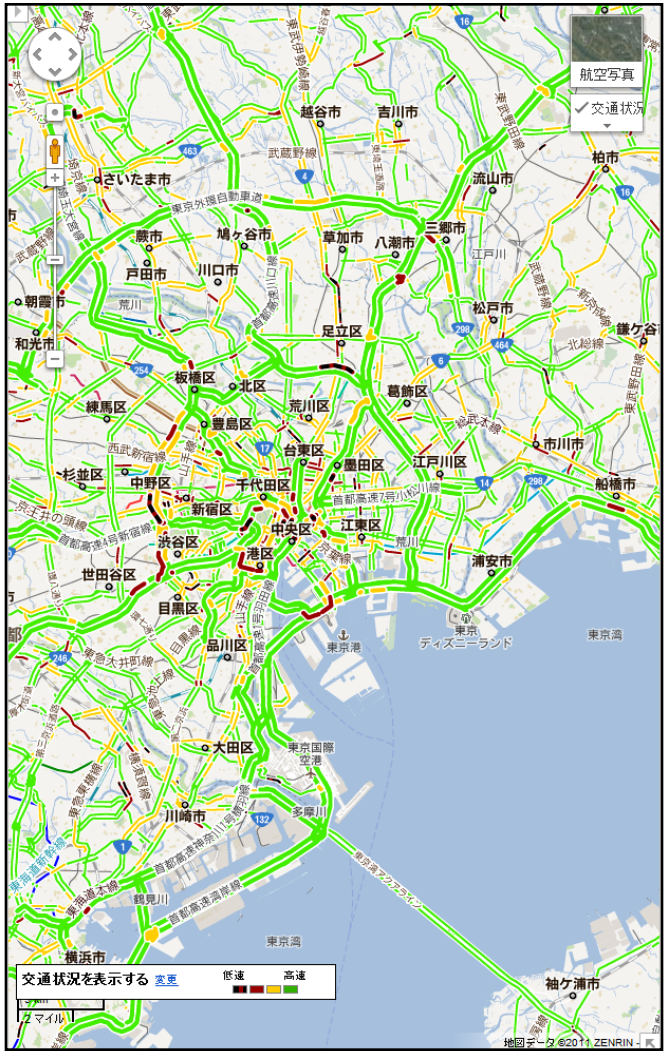
基本

基本：

物事の関係が現れる → グラフ

- グラフ = 物事の関係を表す最も普遍的なモデル
  - バリエーション：無向/有向，重み無し/有り，時間情報，ハイパーグラフ，……
- 物事の関係は，あらゆる所に存在
- それに伴ってグラフデータが扱われる

いろんなグラフ



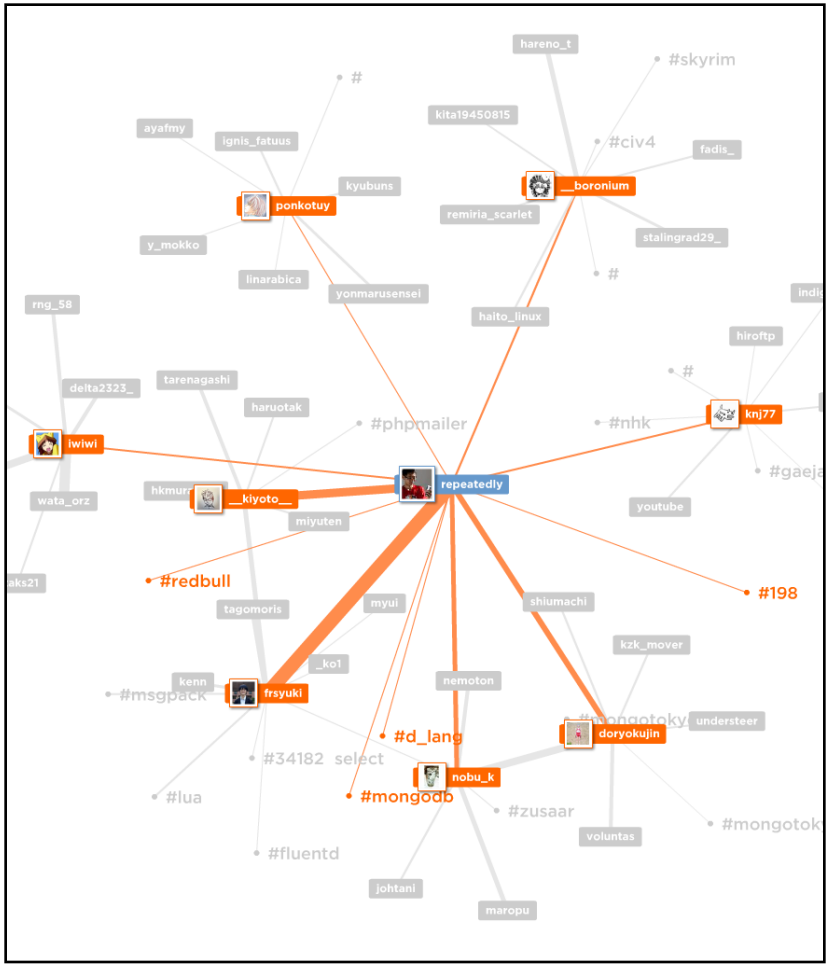
道路・交通ネットワーク

- 頂点：交差点，駅など
- 辺：道，路線など

やりたいことの例

- 案内，交通管制
- 輸送や災害のための解析
- 地理情報と絡めたサービス
- …

いろんなグラフ



(MentionMap で作成)

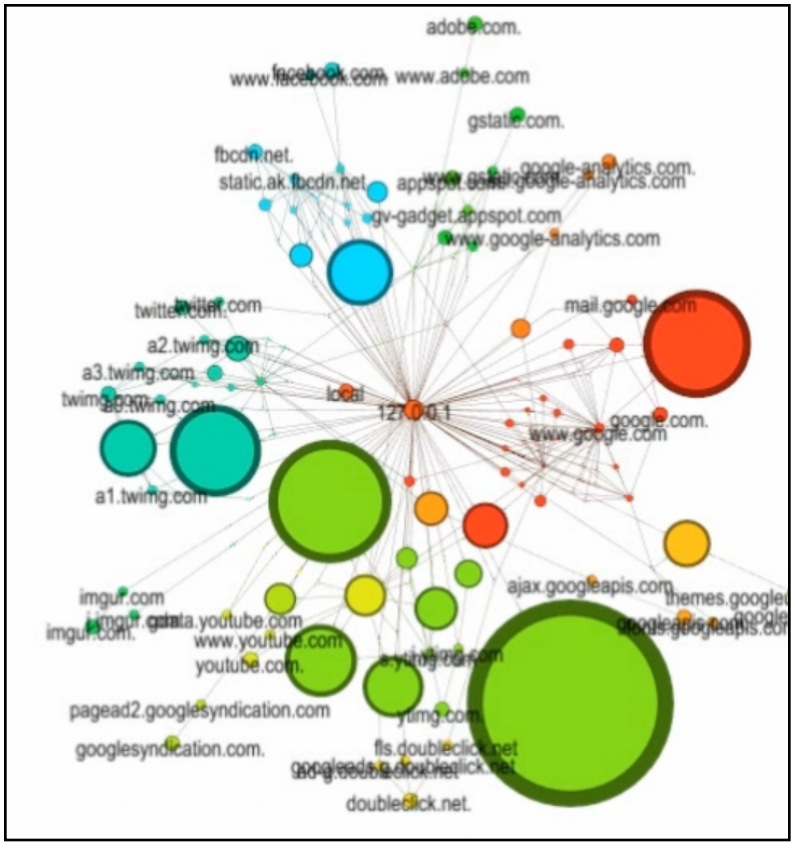
ソーシャルネットワーク

- 頂点：人
- 辺：人間関係

やりたいことの例

- 「知り合いかも？」とか
- 重要度・影響度の解析
- コミュニティ解析
- 情報の伝播力の解析
- …

いろんなグラフ



(Gephi HTTP Graph)

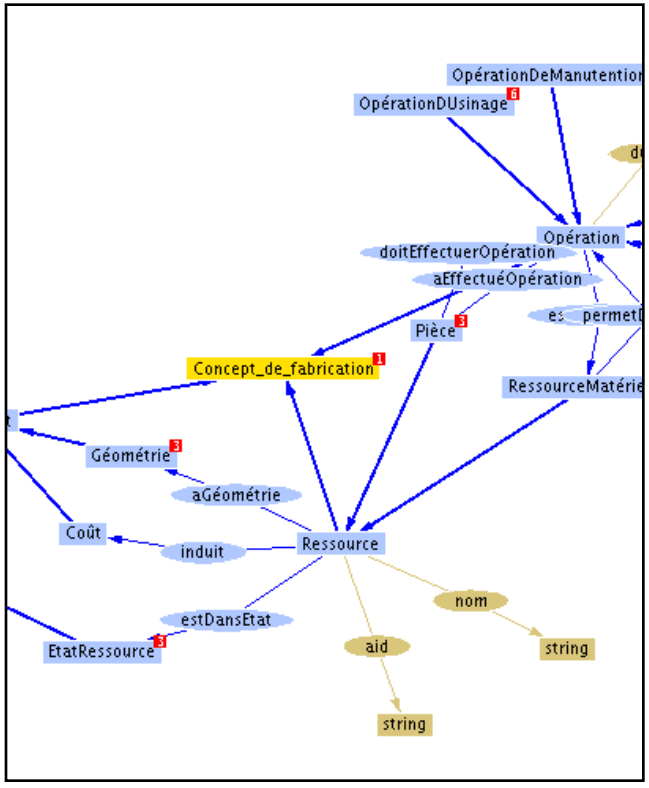
ウェブグラフ

- 頂点：Web ページ
- 辺：リンク関係

やりたいことの例

- PageRank, HITS
- Web ページの関連性
- …

いろんなグラフ



オントロジー

- 頂点：概念
- 辺：is-a 等の関係

やりたいことの例

- チェック，推論
- …

グラフ解析のモチベーション

1. 科学的な関心

2. 有用な情報を得たい

グラフ解析のモチベーション１

モチベーション1: 科学的な関心

- 人間同士はどのように関係しているのか？
- 関係はどうやって形成されるのか？
- ネットワークはどういった「形」？

「ネットワーク科学」 (network science)

物理とコンピュータ科学の境界領域

グラフ解析のモチベーション２

モチベーション2: 有用な情報を引き出す

- 友達の提案
- スパムの判定
- バイラルマーケティング，ワクチン投与

ソーシャルメディアの発達による盛り上がり

Facebook, Twitter, LinkedIn, Instagram, ……

グラフ解析のモチベーション

1. 科学的な関心

2. 有用な情報を得たい

実際には，1 で解明されるモデル等は
2 で活かされるので，明確な区別は微妙

2

グラフ解析の標準的手法

2-1. グラフはどういった形をしている？

この章で対象とするネットワーク

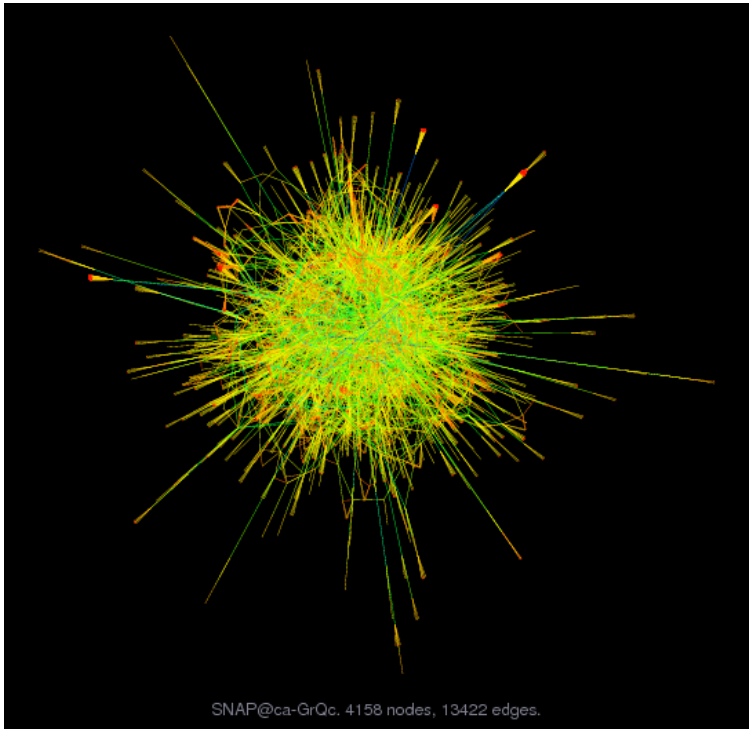
- ソーシャルネットワーク
- ウェブグラフ
- コンピュータネットワーク
- 生物情報学のネットワーク
- ……

こういったネットワークは，一見あまり関係がないのに，
かなり**共通した形**を持っている

なぜ！？
（→ 物事の関係の本質……？？？）

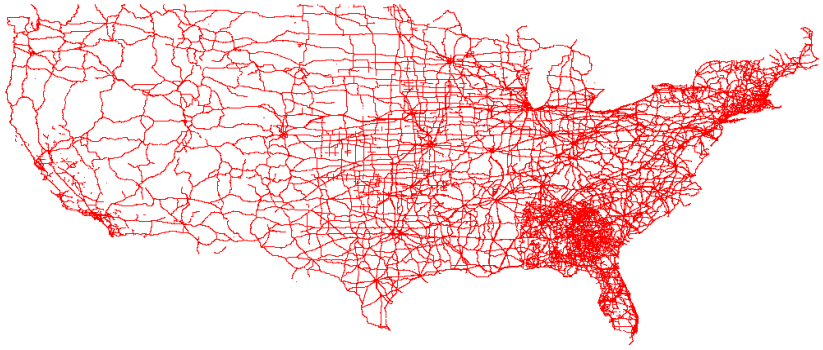
この章で対象とするネットワーク

こういうやつ



共著ネットワーク
[http://www.cise.ufl.edu/research/sparse/matrices/SNAP/ca-GrQc.html]

こうじゃないやつ



道路ネットワーク（アメリカ）

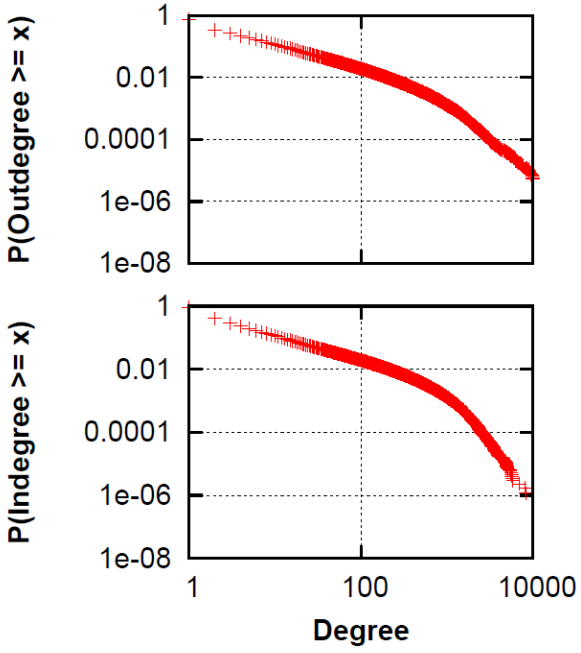
1. 次数分布 (Power Law, Scale Free)

次数分布：冪乗則 (Power Law) に従う

$$p(k) \propto k^{-\gamma}$$

- k：次数
- $p(k)$：次数が k の頂点の割合
- γ：定数 (べき指数)
  - 典型的には $2 < \gamma < 3$

両対数でプロットすると
直線っぽくなる
(累積でプロットしたほうが良い)



[Mislove+'09,Fig.2]

1. 次数分布 (Power Law, Scale Free)

冪乗則

- 単語の使用頻度（ジップの法則）
- 正規分布とかと違い，かなり大きな値が存在する
  - 一部の人がすごくお金持ち，大半の人は平均以下
  - ネットワークでも，凄く次数が高い頂点がちらほらある

次数が冪乗則に従っているか？

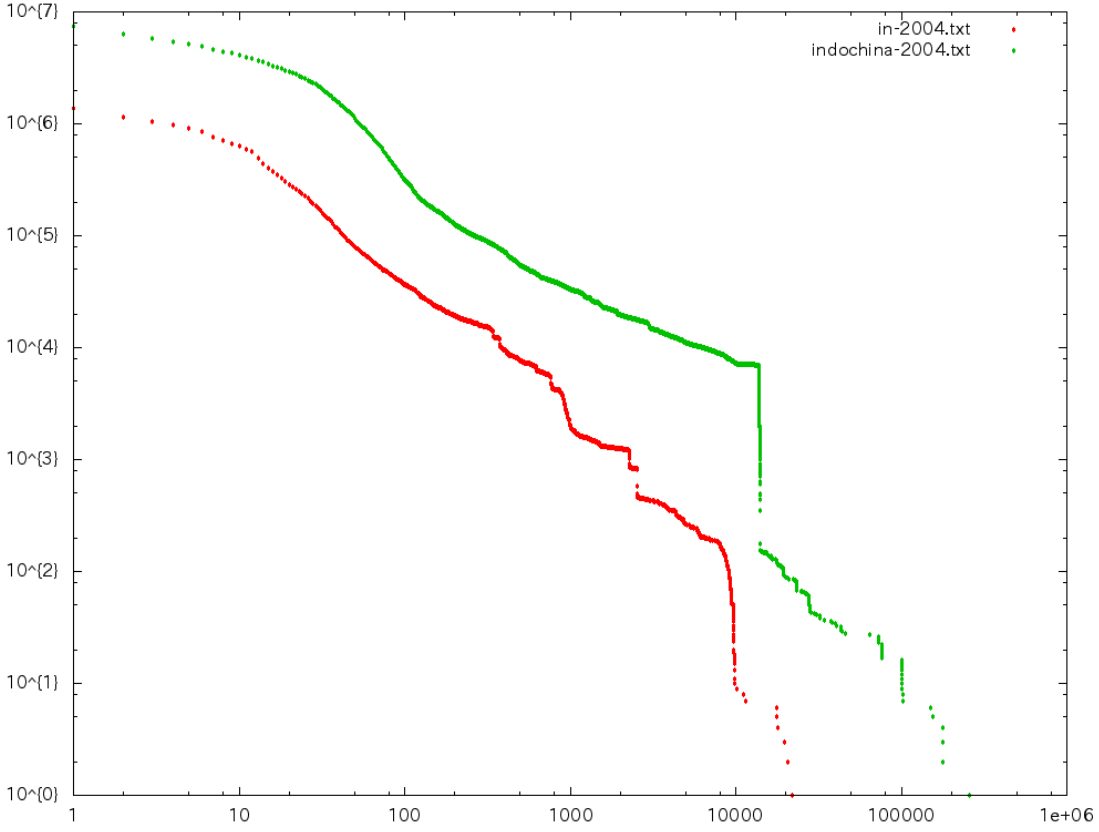
1. 雑：両対数でそれっぽくなっているか
2. 真面目：統計モデルで γ を推定，検定

低い次数の部分では従わないことが多い

アルゴリズム的には，凄い次数が大きい頂点がちょっとある，とかそういうぐらいの事実が重要な気がする

1. 次数分布 (Power Law, Scale Free)

ウェブグラフの累積次数分布



う，うーん……

1. 次数分布 (Power Law, Scale Free)

べき分布の何がすごい？

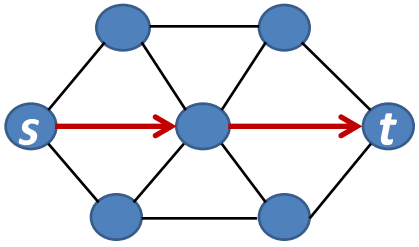
- 普通に独立な分布が重なりあう → 大数の法則・中心極限定理によって，正規分布
- 従って，**大域的な現象**が存在する

べき分布を再現するモデル

- 「次数に比例する確率で新たな辺を獲得する」 (BA model)
- Rich gets richer

2. 距離 (Small World)

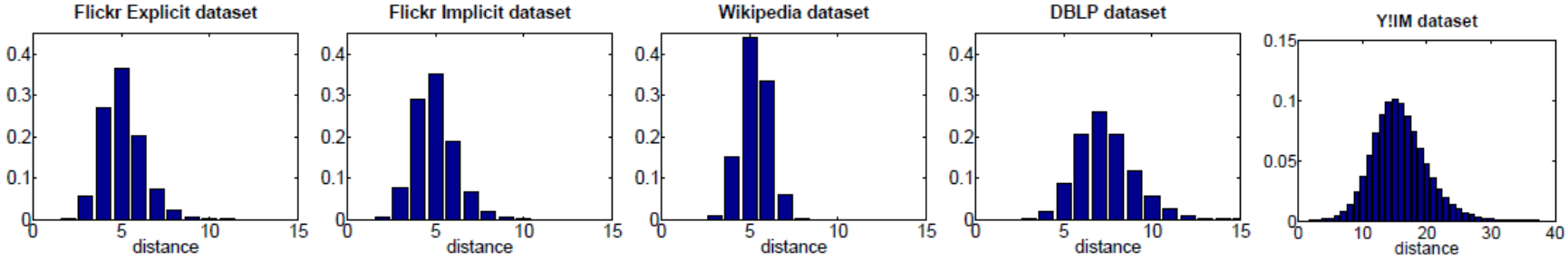
平均距離：短い



- 距離：二頂点の間の最短路の長さ
- 平均：全二頂点対

短いとは：$O(\log n)$ である，ランダムに繋ぎ変えても有意に小さくならない等

距離の分布の例



[Potamias+,CIKM'09,Fig.2]

2. 距離 (Small World)

実際の人間を通した実験

- 1960 年代，ミルグラム「平均距離 6 だ!!!」
  - 手紙を転送して貰って目的の人物に届ける
  - 現代では，破棄の考慮や始点に疑問の声

- 2002 年，ワッツ「似たような感じだった!!!」
  - 電子メール
  - 始点の数や位置，破棄を今度は考慮しているらしい

コンピュータによるネットワークの解析

- カジュアルに計算されてる（やっぱ小さい）
- 2011 年，Backstrom (Facebook社)「4.74 だ!!!」
  - Facebook のネットワーク (721 M users / 69 B links)
  - [Backstorm+'11] http://arxiv.org/pdf/1111.4570v3.pdf

2. 距離 (Small World)

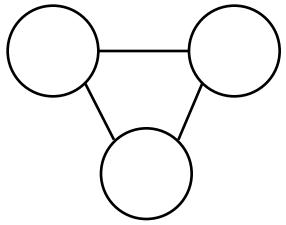
なぜ距離は小さくなる？

逆に，よっぽど恣意的に作らない限り，
グラフの平均距離は勝手にかなり小さくなる

- 平均距離が大きいグラフの例：パス，グリッド
- そういうグラフでも，ほんのちょっと，ランダムな辺を加えてやるだけで，平均距離はすぐに小さくなる

3. クラスター係数 (Small World)

クラスター係数：大きい



$$\text{クラスター係数 } C = \frac{\text{三角形の個数}}{\text{繋がってる三頂点の組の個数}}$$

大きいとは：同じスケールのランダムに作るグラフより有意に大きい

- 友達の友達は友達
- **局所性**のようなものがある

「スモールワールド性」：
小さい平均距離と大きいクラスター係数
（距離だけで言うこともある）

2

グラフ解析の標準的手法

2-2. グラフから有益な情報の獲得

グラフ解析の道具

- **中心性**：ある頂点の重要度
- **関連度**：２頂点の関連度
- **コミュニティ**：関連の強そうな頂点集合
- ……

中心性とは

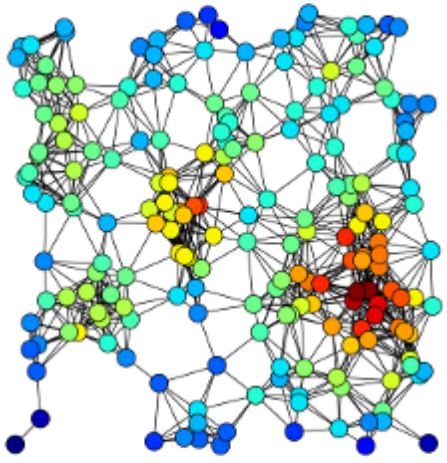
中心性 = 重要度の推定値

色々な考え方がある

- 次数中心性 (Degree Centrality)
- PageRank, HITS
- 近接中心性 (Closeness Centrality)
- 媒介中心性 (Betweenness Centrality)

次数中心性

次数の大きさ = 重要度



- 一番簡単な中心性
- **悪意**に簡単に負ける
- が，それでも，実際に使ってみると
  予想をはるかに超えて役に立つ
  - 「まずは次数中心性を使っておこう」というレベル
    では全然アリ，思ったよりずっと使える
  - べき分布のお陰で，大きな開きがあるから

PageRank

Google 検索で有名な手法

Page はウェブページのページではなく人名

ランダムサーファーモデル

- ランダムな頂点からスタート
- 確率 α で：ランダムな隣接点を選んで移動
- 確率 $1-\alpha$ で：ランダムな頂点にワープ

$\alpha = 0.85$ が「Google 定数」

各頂点の滞在確率が PageRank.

「重要なウェブページからリンクされていると重要」

PageRank

PageRank の計算：反復法

- 全員の**仮 PageRank** を $1/n$ から開始
- 繰り返す：
  - 各頂点の仮 PageRank を，周りの頂点の仮 PageRank を使って更新

これは，線形方程式のヤコビ法に対応.

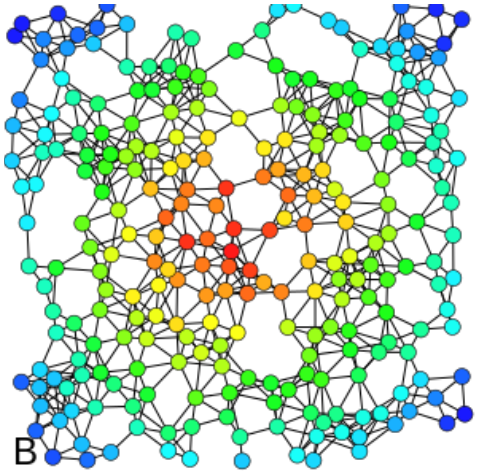
$O(\alpha^t)$ で収束. （t：反復回数）

高速なアルゴリズム [Maehara-Akiba-Iwata-Kawarabayashi, VLDB'14]

近接中心性

距離に基づいた中心性

- $C(v) = \sum_{u \in V} \frac{1}{d_G(v,u)}$
- $C(v) = \sum_{u \in V} 2^{-d_G(v,u)}$



http://en.wikipedia.org/wiki/Centrality

一般に

u から v への距離

$$C_\alpha(v) = \sum_{u \in V} \alpha\big(d_G(v,u)\big)$$

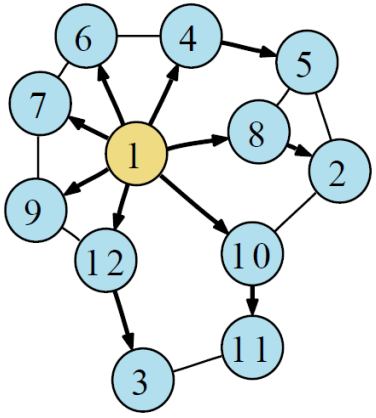
α は単調非減少，　$\alpha(d) \in [0,1]$

（他の定義もあるけどこれで大部分はカバーできてる）

近接中心性のナイーブな計算

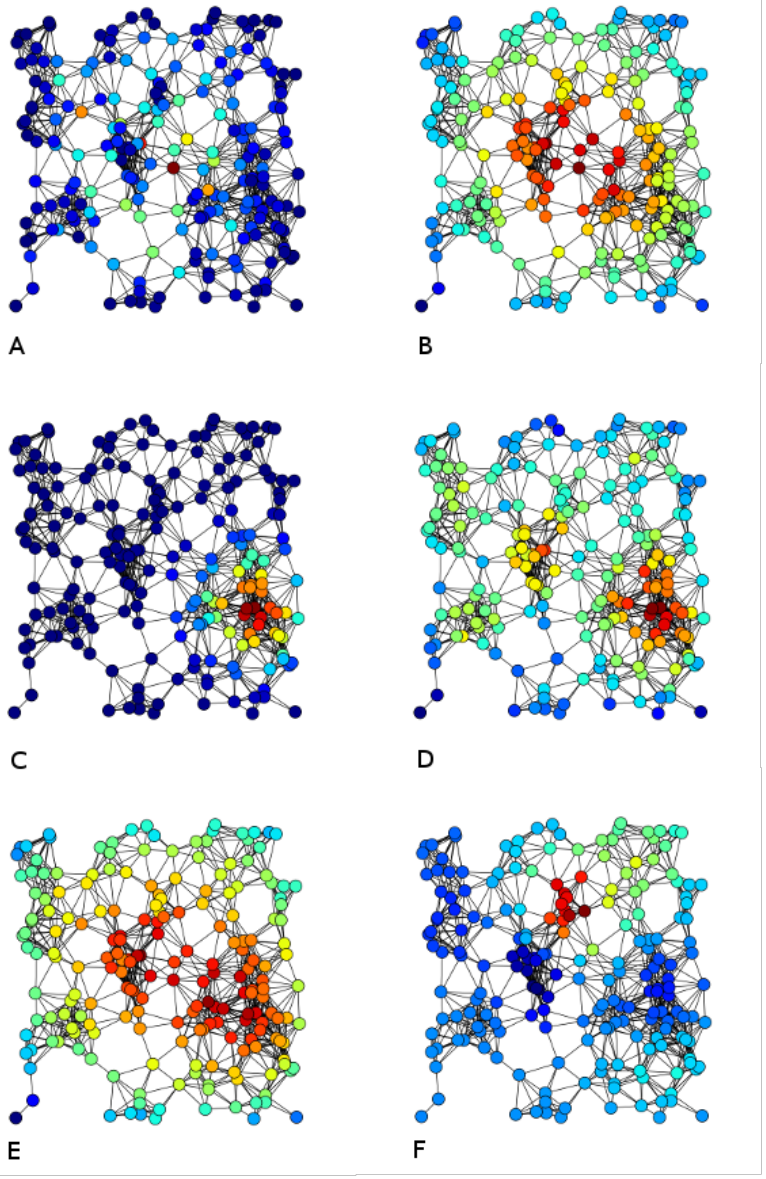
最短経路アルゴリズム

- 幅優先探索 (重みなし)
  - $O(m+n)$ 時間



- Dijkstra のアルゴリズム (重み有り)
  - $O(m+n\log n)$ 時間

色々な中心性



色々な中心性，それぞれ重視するものが異なる

大抵のグラフ解析ソフトウェアはポピュラーなものをどれも用意している

（ただし，左図のような toy では挙動が違っても，実大規模データではなんやかんやどの値も殆ど同じ挙動を示すという話も少なくない）

関連度

頂点間関連度の指標も一杯ある

ローカルなもの

- 共通する近傍の数，Jaccard 係数，AA, ……

グローバルなもの

- 距離
- Random walk with Restart (Personalized PageRank)
- SimRank, Katz, ……
- Closeness Similarity, Co-Betweenness, REL, ……

関連度：Random Walk with Restart

PageRank の，ジャンプ先が一頂点固定の場合

- 頂点 v からスタート
- 確率 α で：ランダムな隣接点を選んで移動
- 確率 $1-\alpha$ で：頂点 v にワープ

ある頂点 u に居る確率を，v と u の関連度とみなす

欠点

- 例え無向グラフでも，(u,v) と (v,u) の値が違う
  （気持ち悪い）

計算法

- PageRank と同じく反復法
- Push-PPR [Andersen Chung Lang 06]

コミュニティ分析

一番有名な思想：モジュラリティ $Q = \sum_i (e_{ii} - a_i^2)$

- グラフの分割の「コミュニティっぽさ」の指標
- コミュニティ内の辺の数を比べる感じ
  - 実際の辺の数
  - 次数のみを利用して，「行き先完全ランダム」と仮定して推定した数

モジュラリティに基づいたコミュニティ分割手法

- Girvan-Newman, CNM, Brondel, ……
- 塩川さん (NTT) のアルゴリズムが多分最速 [AAAI'13]

最適化は NP-hard なのでヒューリスティクス的に分割

3

大規模グラフ解析の課題と研究動向

大規模グラフ処理の難しさ

一言で言うなら（個人的な感想ですが）

「ランダムなアクセスが本質的に要求されるから」

- データや処理が分割できない
  - ある部分の処理がどこに関連するか，というのがメチャクチャ
- ツリーのような構造がフィットしない
  - 二分探索木，Trie 木のように良いデータ構造が作れない

大規模グラフ処理の難しさ

1: (約) 線形時間でできる事が非常に限られる

線形時間でできること

- 単純なグラフ探索（１回）
- Power iteration (PageRank)

$\Theta(n^2)$ とかかかってしまおうものなら，大規模なグラフはもう処理できない

大規模グラフ処理の難しさ

2: ランダムアクセスを要求する性質

- 線形時間で処理できるものでも，超大規模なグラフを扱うのはなお難しい
- 分散処理をしようとしても，MapReduce のようなフレームワークに乗りにくい
- On-disk 処理も同様にかなり難しい

研究動向

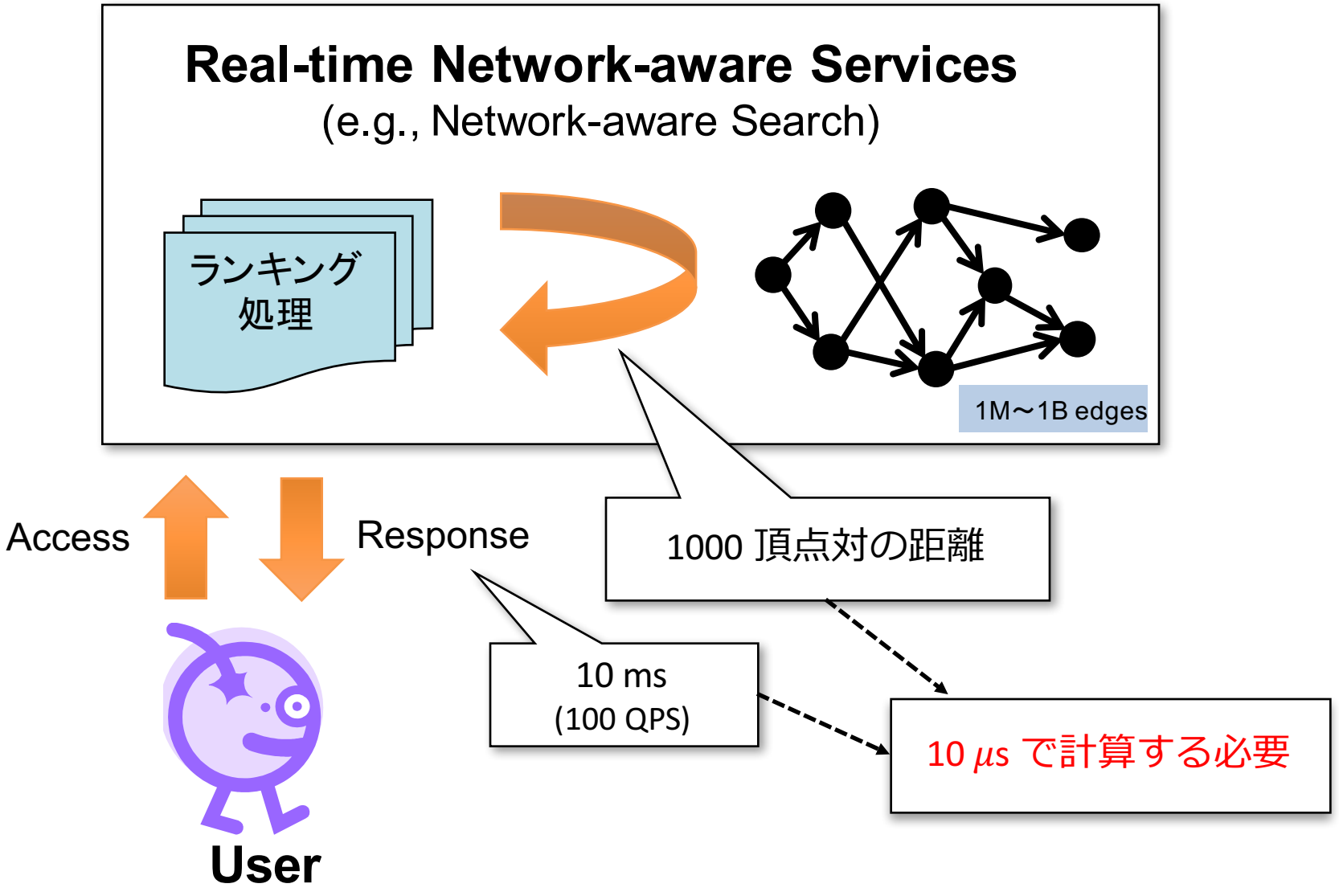
大規模グラフ解析の手法に関する論文が出るのは:

- データベース系: VLDB, SIGMOD, ICDE, …
- データマイニング系: KDD, ICDM, SDM, (AAAI?) ……
- Web 系: WWW, WSDM, ICWSM, …

論文の傾向

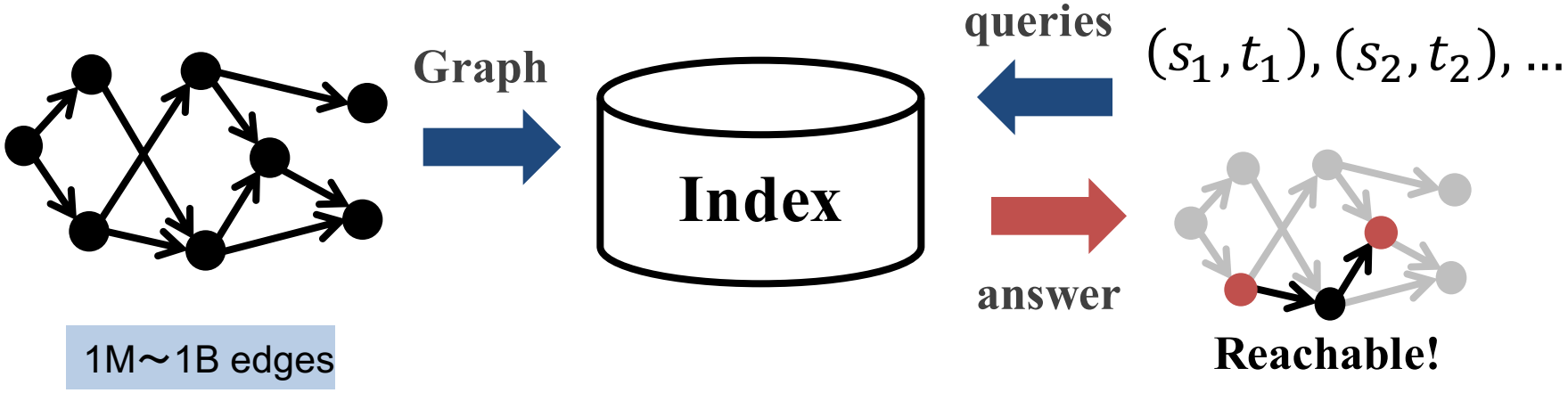
- 大規模グラフ関連に絞ると，どこも近いトピック
- DB 系はスケーラビリティの要求が高い
- マイニング系はより直接的な応用を要求する

索引付け手法の必要性


Real-time Network-aware Services
(e.g., Network-aware Search)
ランキング
処理
1M～1B edges
Access
Response
1000 頂点対の距離
10 ms
(100 QPS)
10 μs で計算する必要
User

グラフ上のクエリ処理

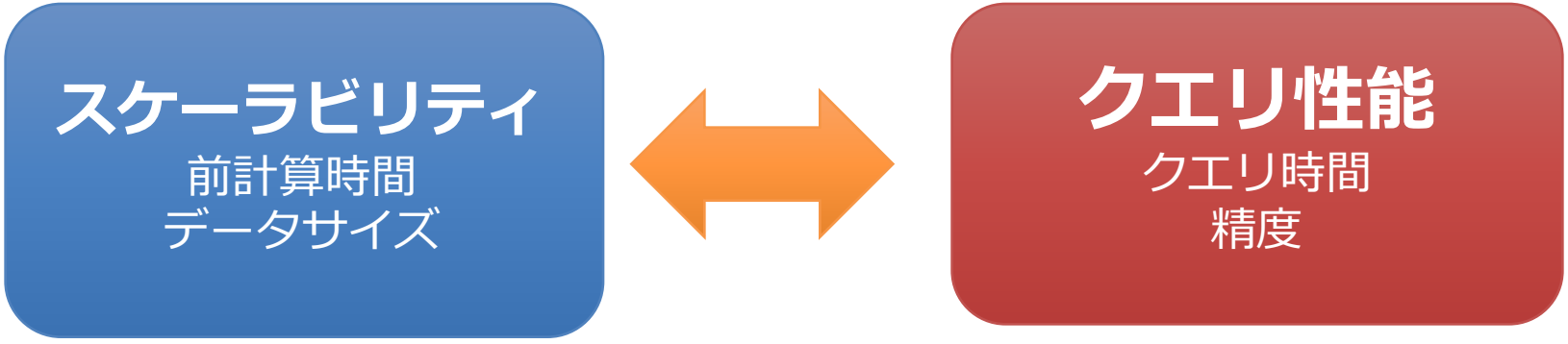
1. **索引構築**
   - グラフからデータ構造を前計算しておき
2. **クエリ応答**
   - それを用いて 2 点間のクエリを高速に答える



グラフ上のクエリ処理

1. **索引構築**
   - グラフからデータ構造を前計算しておき
2. **クエリ応答**
   - それを用いて 2 点間のクエリを高速に答える

目標: 良好なトレードオフ（実性能）



有名なグラフクエリ問題

Distance Queries on Social and Web Graphs

- Landmark [CIKM'09] [WSDM'10] [CIKM'10] [ICDE'12], TD [SIGMOD'10] [EDBT'12], 2-Hop [SIGMOD'12] [ESA'12], ……

Distance Queries on Road Networks

- CH [WEA'08], CHASE [JEA'10], TNR [ALENEX'07], HL [SEA'11], ……

Reachability Queries on Citation and XML Graphs

- GRAIL [VLDB'11], PWAH [SIGMOD'11], SCARAB [SIGMOD'12], TF-Label [SIGMOD'13], ……

僕の研究（宣伝）

独立に研究されていた関連問題に対し

統一的なアプローチでより優れた手法を提案

既存手法と対象グラフの性質の深い解析により「何が統一できるのか」を突き詰めた

これまで不可能であった
億スケールでの厳密解を達成（**100 倍**の改善）

複雑ネットワークでの最短経路

Landmark [CIKM'09] [WSDM'10] [CIKM'10] [ICDE'12], TD [SIGMOD'10] [EDBT'12], 2-Hop [SIGMOD'12] [ESA'12], ……

枝刈りラベリング法
[Akiba-Iwata-Yoshida, SIGMOD'13]
[Akiba-Iwata-Yoshida, WWW'14]

有向無閉路グラフでの到達可能性

GRAIL [VLDB'11], PWAH [SIGMOD'11], SCARAB [SIGMOD'12], TF-Label [SIGMOD'13], ……

枝刈りラベリング法
[Yano-Akiba-Iwata-Yoshida, CIKM'13]

交通ネットワークでの最短経路

CH [WEA'08], CHASE [JEA'10], TNR [ALENEX'07], HL [SEA'11], ……

枝刈りラベリング法
[Kawata-Akiba-Iwata-Yoshida, ALENEX'14]

グラフの索引付け（最新動向）

- 基礎的な問題に対する索引付けは少し落ち着いてきた
  - 最短路，到達可能性，…… 一時期はすごい勢いで論文が出ていた
- 動的な更新のサポート
- 少し複雑なクエリ ・・・ グラフとキーワード検索を絡めたり

研究動向

コミュニティ検出（最新動向）

- 手法の良し悪しを比較するのが未だに難しい
- 乱立，皆自らの指標で優位性を主張
- 最も有名な **modularity clustering** は，大規模データではとても大きな塊が出てしまう等，微妙らしい
- **stochastic block model** が最近は人気（と聞きました）

研究動向

分散 vs シングルマシン

- **グラフ分散処理エンジン**の研究・開発が進む
  - 「並列分散を活用しないのは時代錯誤」
- 一方で，**でかいメモリを積んだシングルマシン**を使うのが大抵は一番いいでしょ，という声明・事例も実は（有名チームから）相次いでいる
  - Jure Leskovec チーム SIGMOD'15 (Best Demonstration Award)
  - "Graph Structure in the Web－Revisited" WWW'14
    1×10^{11} 辺のグラフを 1 台のマシンで解析

通信のコストが本質的に大きく，分散でできる事は限られている？
グラフのデータだけなら圧縮と組み合わせて，
なんやかんや殆ど現代的ワークステーションのメモリに乗る.

話したこと

1. グラフ解析入門

- どのようなグラフデータが有るのか？
- なぜそれを解析をするのか？

2. グラフ解析の手法と古典的アルゴリズム

- グラフはどういった形をしている？
- 中心性，関連度，コミュニティ，全体の性質

3. 大規模グラフ解析の課題と研究動向