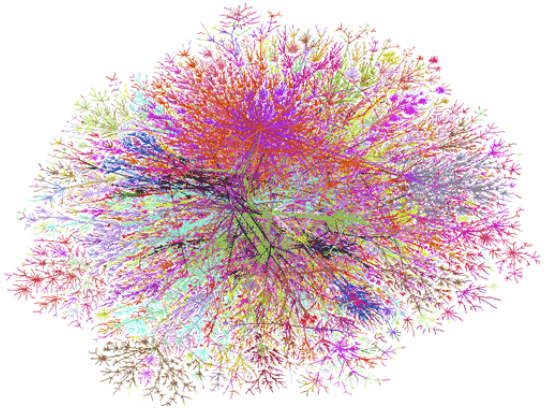
複雑ネットワーク（complex network）

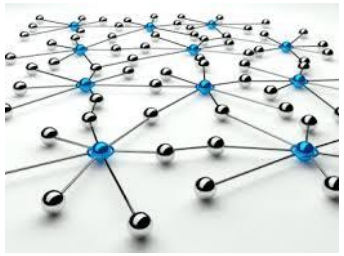
グラフ理論、ネットワーク、中心性、次数中心性、近接中心性、媒介中心性、クラスタ係数



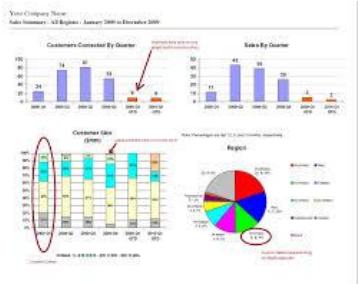
Webマイニングの手法

- Webコンテンツマイニング
  - Webの内容（contents）を分析
  - 主にテキストマイニングといった手法

- **Web構造マイニング**
  - Webサイトのリンク関係、生成・消滅、分裂・統合、といった構造を分析
  - 主に**グラフマイニング**（ネットワーク分析）

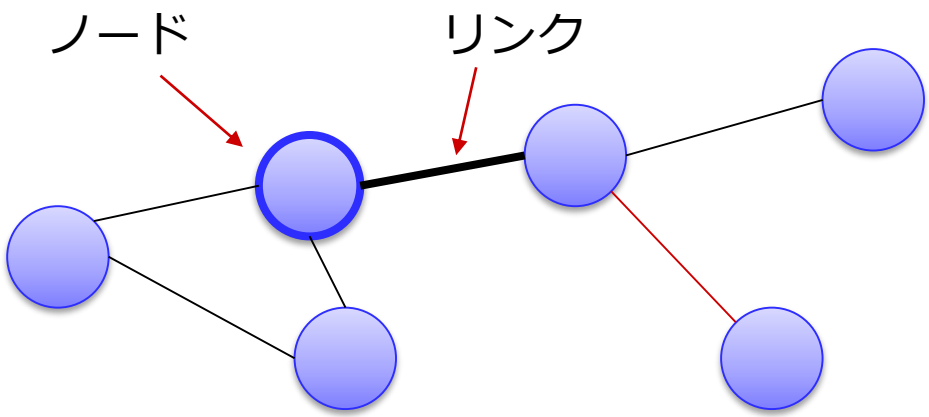


- Webログマイニング
  - Web上の行動ログ（アクセスログ、検索履歴）の統計情報にもとづいて、コミュニティや個人の思考を分析
  - ログ解析、統計分析、協調フィルタリング



ネットワークとは？

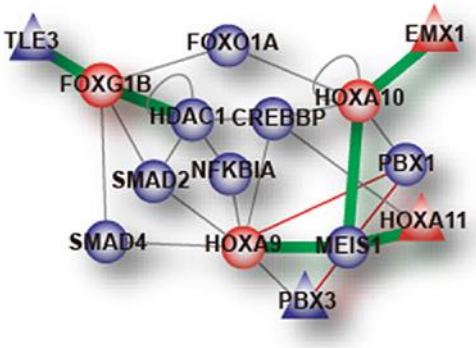
- 頂点（ノード）と辺（リンク）から構成される構造



ネットワーク

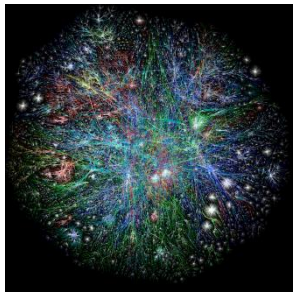
- 世の中にはたくさんのネットワークが存在
  - 人間関係、自然界、人工物、…

ソーシャル・ネットワーク



細胞

地下鉄網

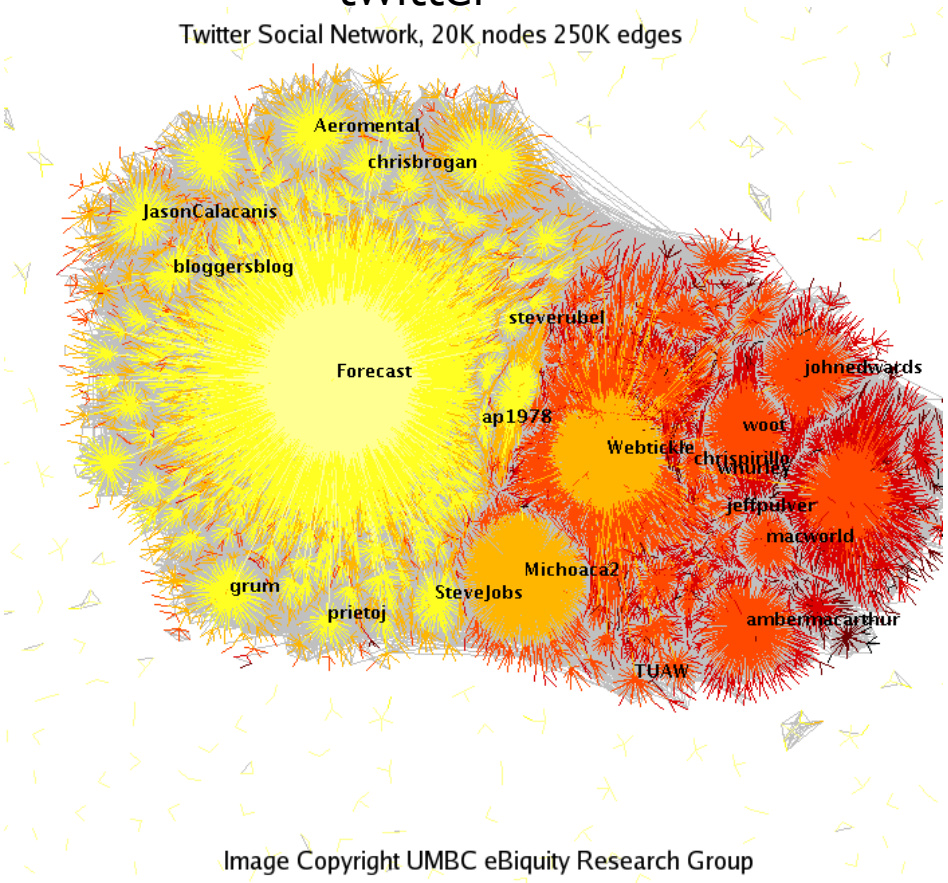


インターネット

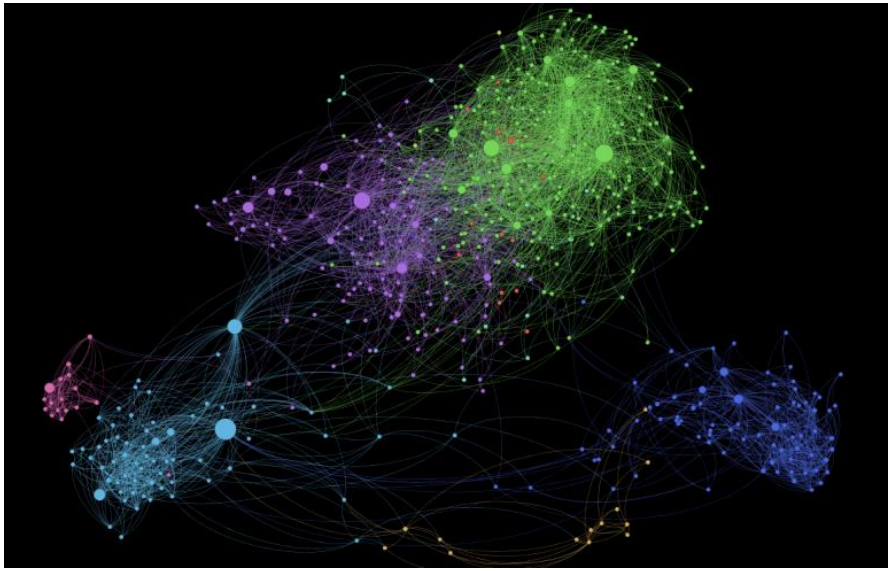
航空網

（例）ソーシャル・ネットワーク

twitter



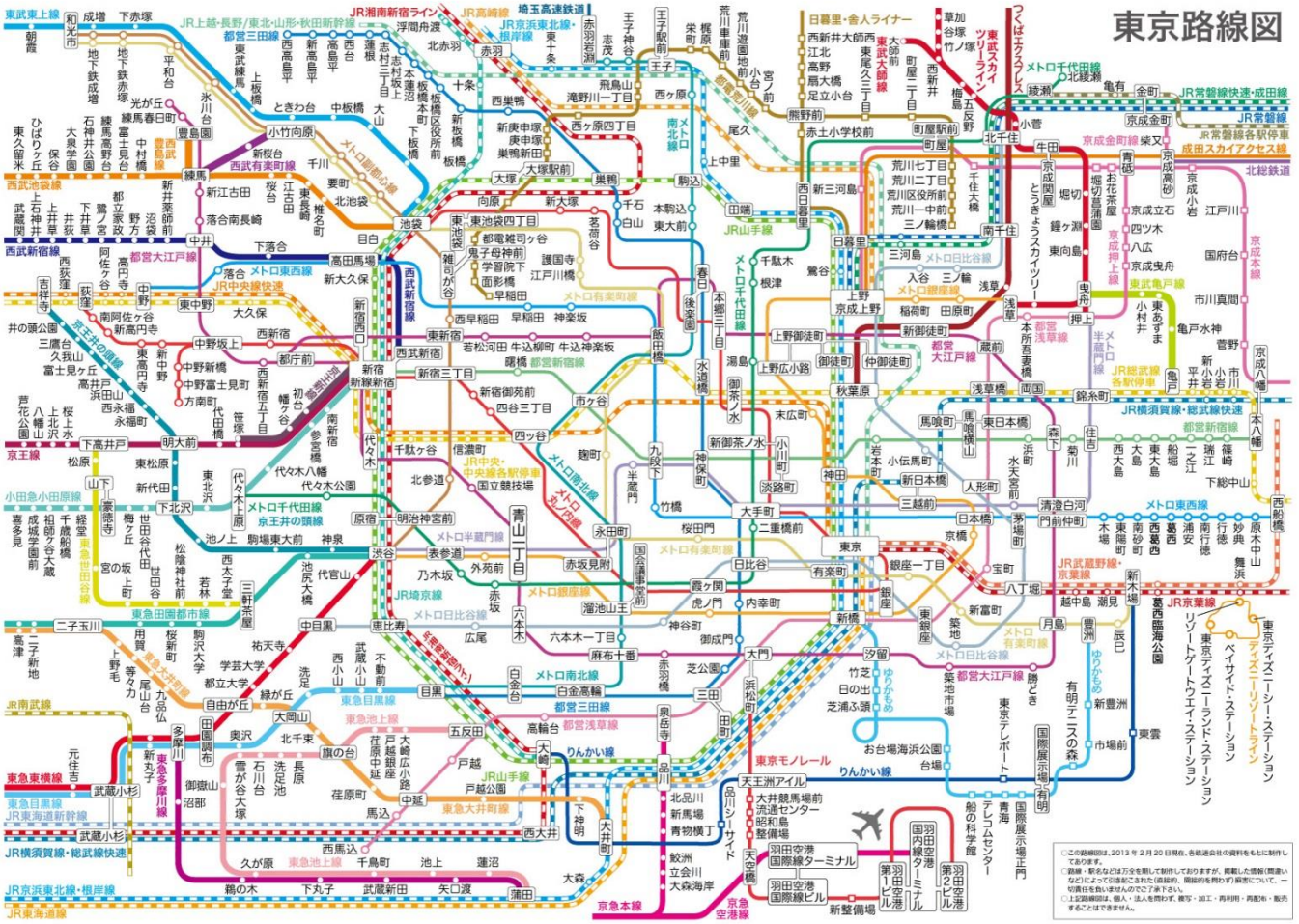
Facebook



FACEBOOK NETWORK VISUALIZATION
OF +1000 FRIENDS

- ノード＝人
- リンク＝フレンド

（例）路線図（東京）



- ノード＝駅
- リンク＝路線

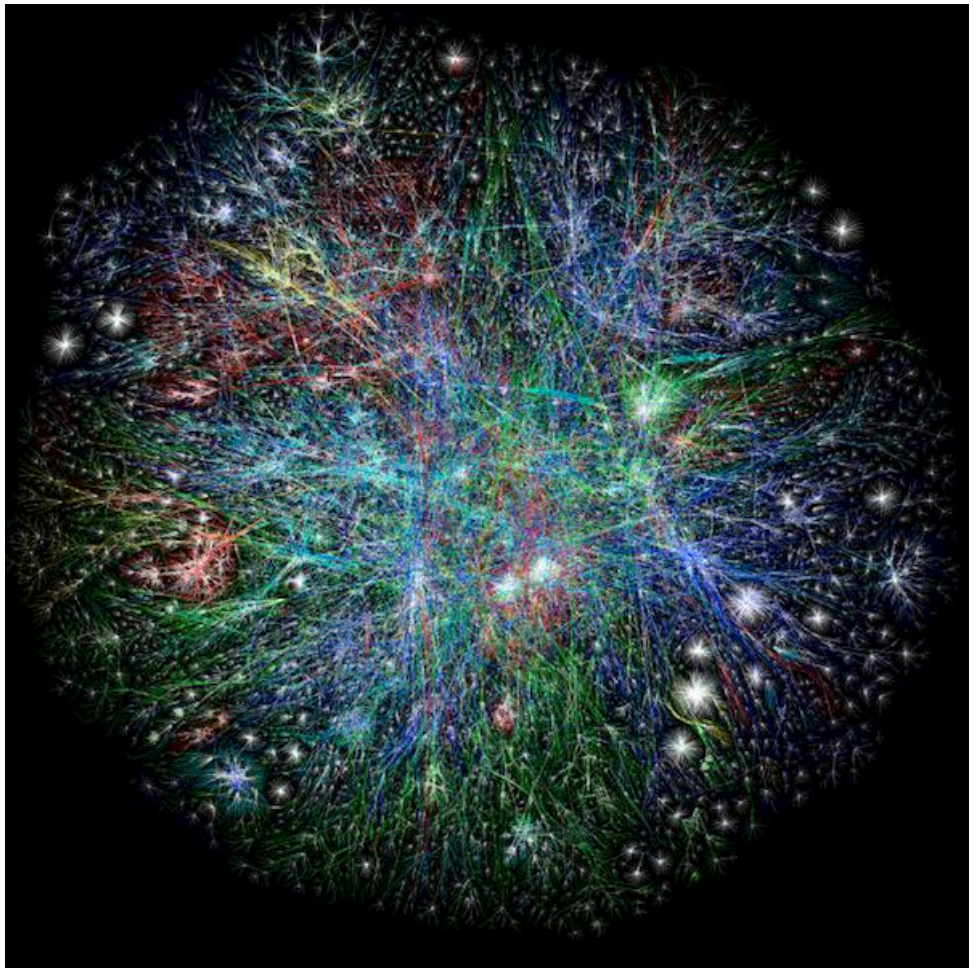
（例）航空ネットワーク

The Openflight.org airport network where the line colour is based on the number of routes (accessed on August 12, 2011). The code to replicate this image can be found at the end of this page.

- ノード＝空港
- リンク＝経路

（例）インターネット

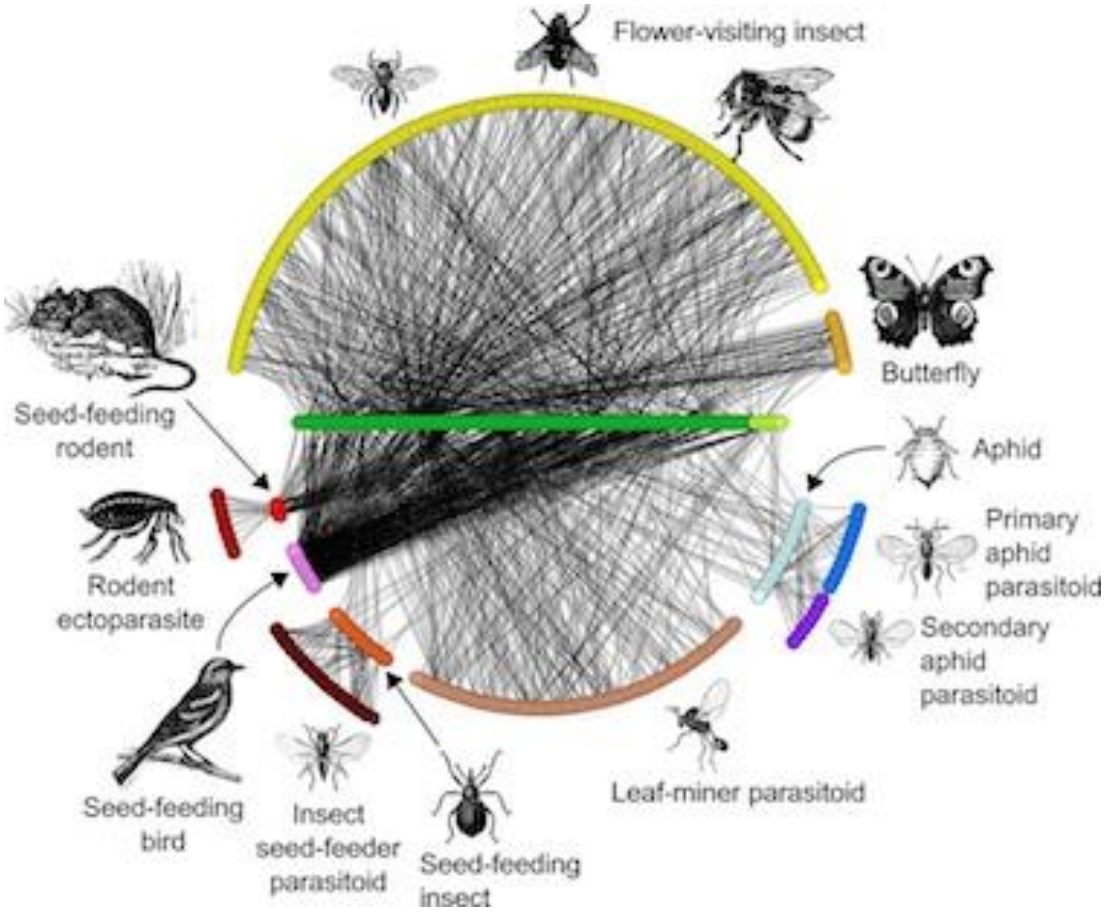
14,000,000,000 Web pages (14 billion)



- ノード＝ページ
- リンク＝ページリンク

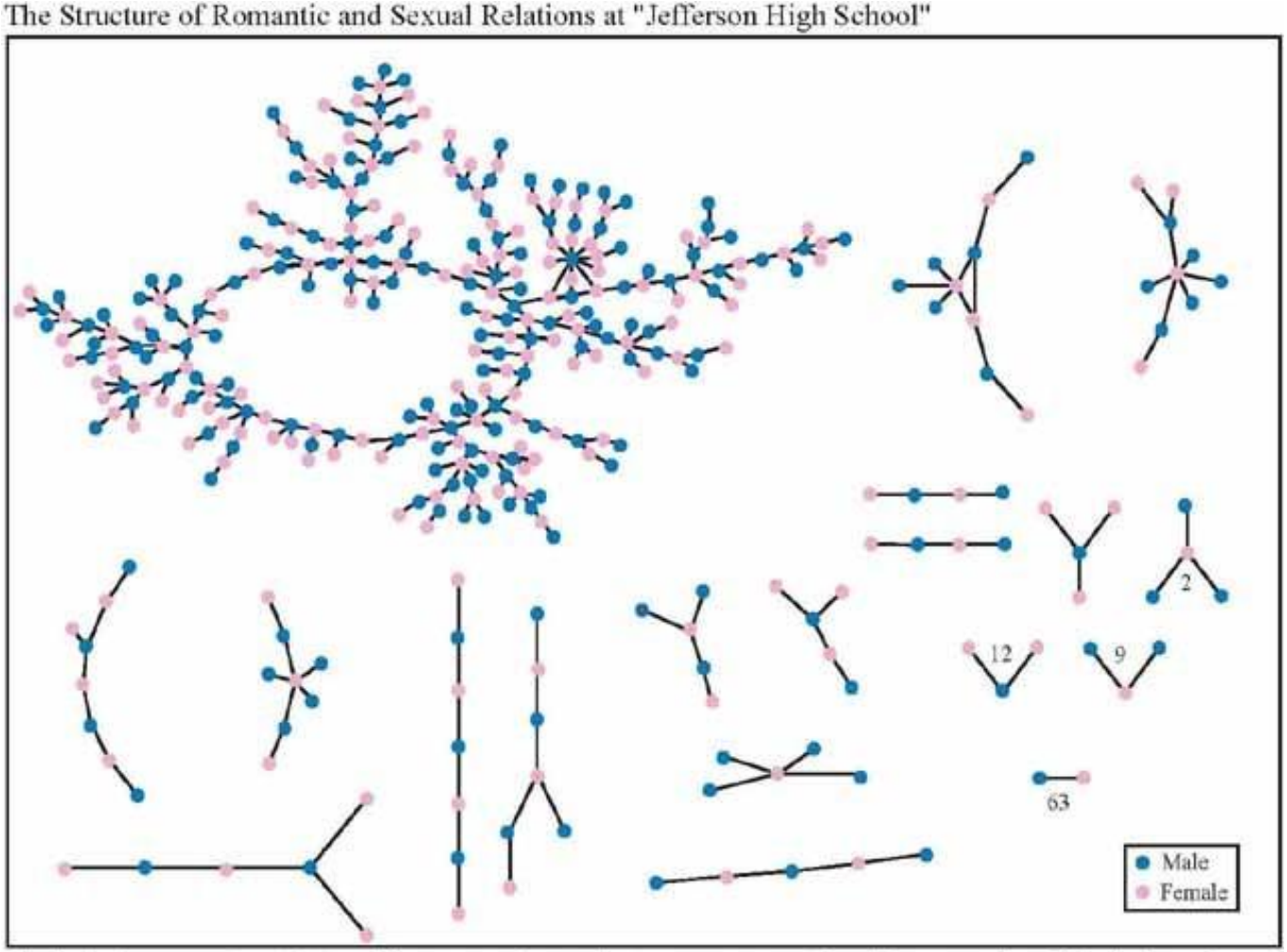
http://nonesnotes.com/2013/02/20/fourteen-billion-web-pages-but-you-can-get-from-any-one-to-any-other-in-19-clicks-or-less/

（例）生態系ネットワーク



ノード＝種
リンク＝相互作用

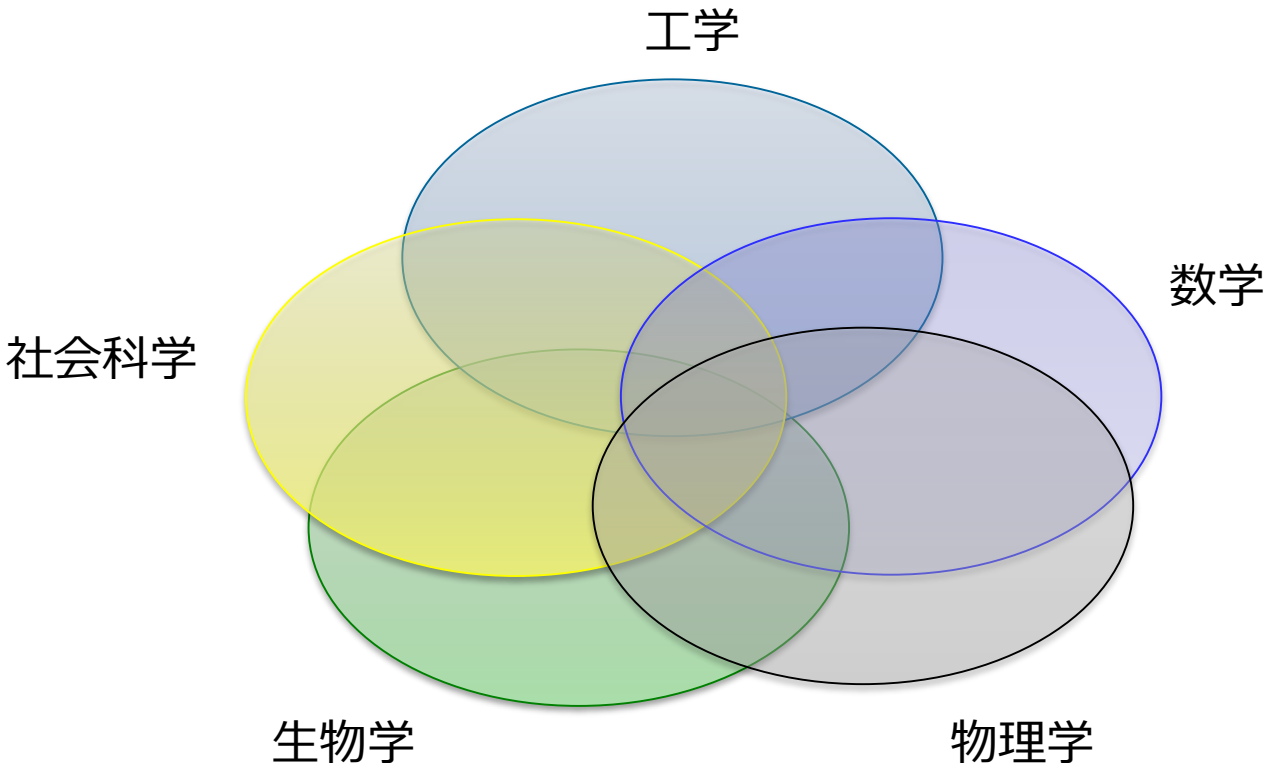
（例）高校生の恋人関係ネットワーク



Each circle represents a student and lines connecting students represent romantic relations occuring within the 6 months preceding the interview. Numbers under the figure count the number of times that pattern was observed (i.e. we found 63 pairs unconnected to anyone else).

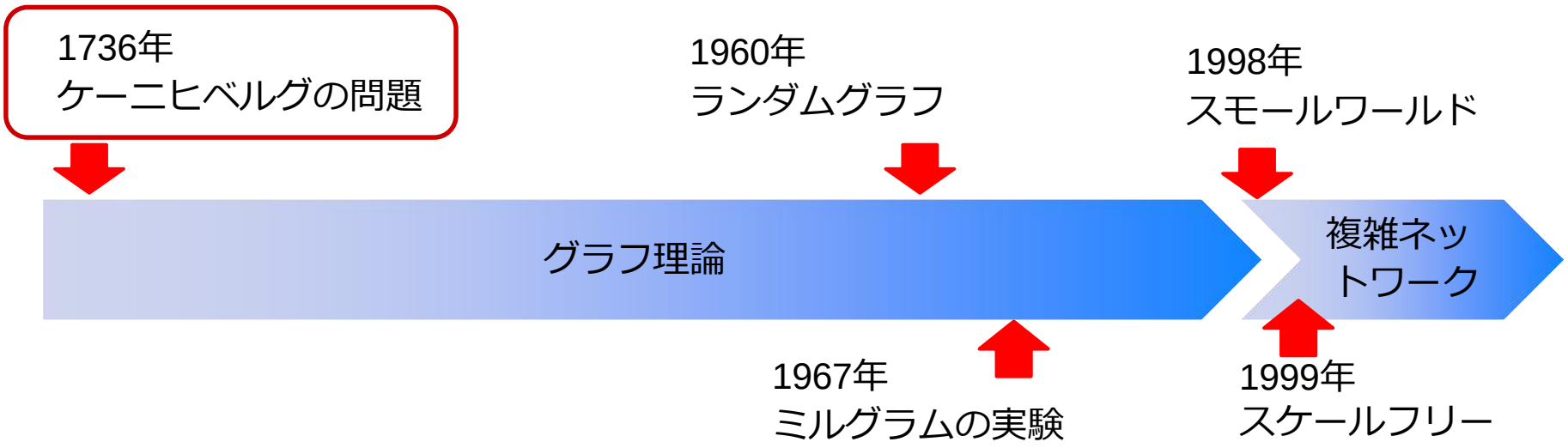
ネットワークの研究

- 実は多くの分野で研究されている



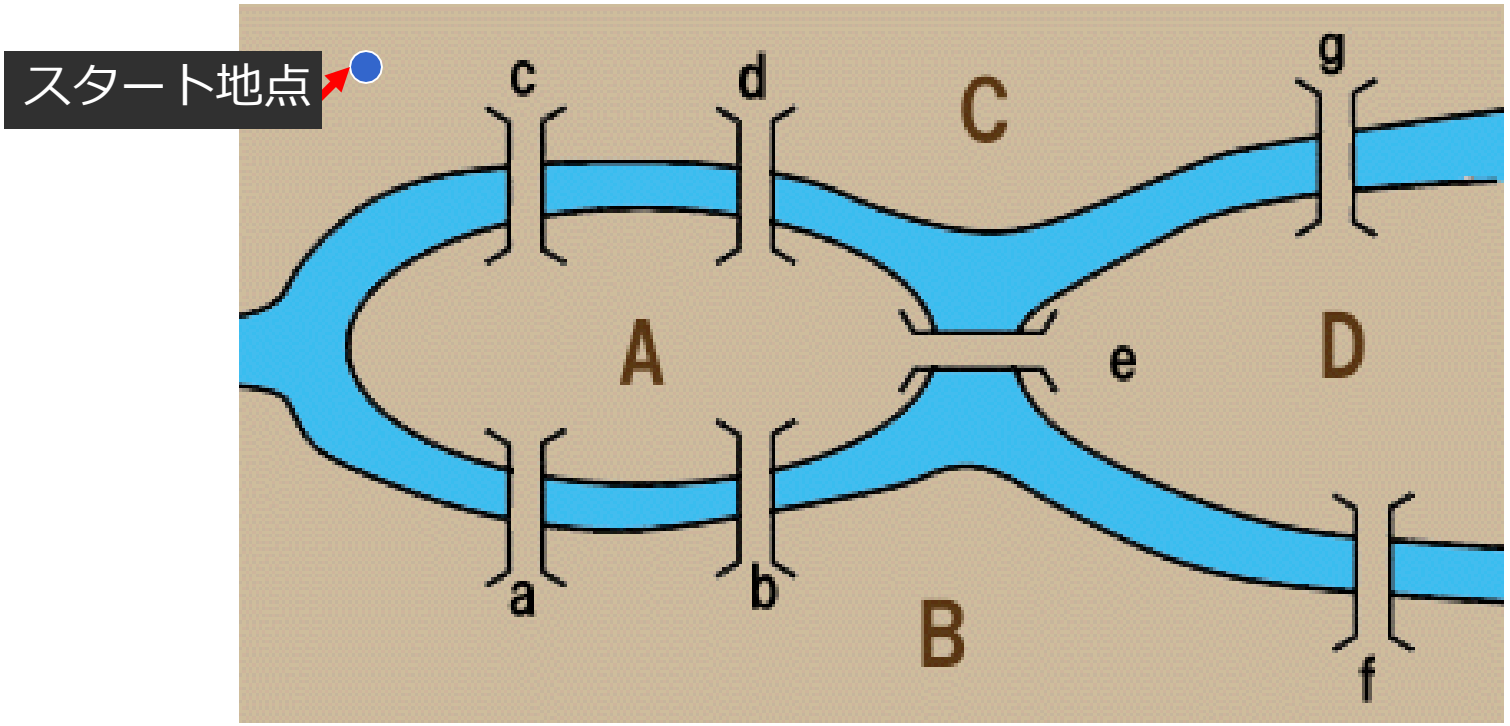
ネットワーク研究の歴史

ネットワーク研究の歴史



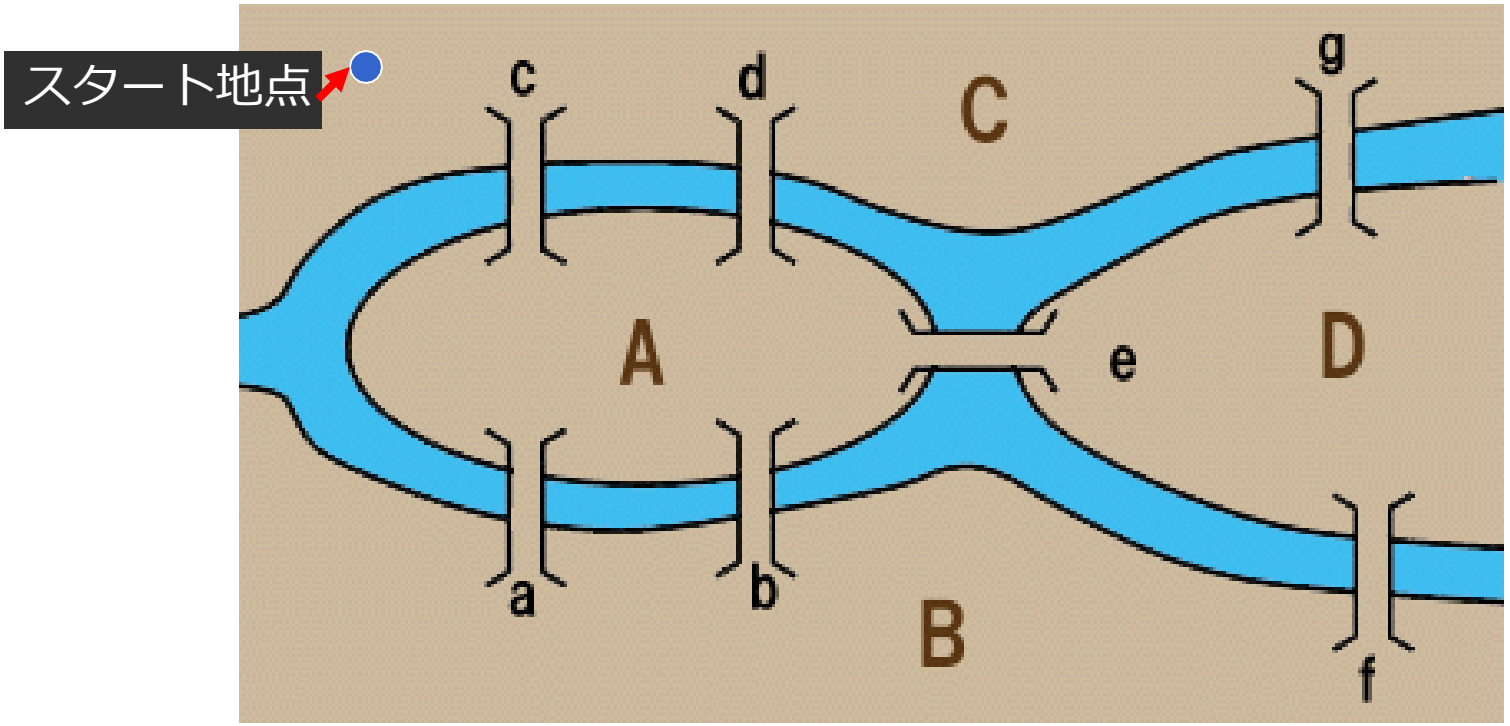
- グラフ理論
- ランダムグラフ
- スモールワールド
- スケールフリー

グラフ理論：ケーニヒスベルクの問題



- ケーニヒスベルクという町（現：ロシア連邦カリーニングラード）
- a～gという7つの橋が架かっている

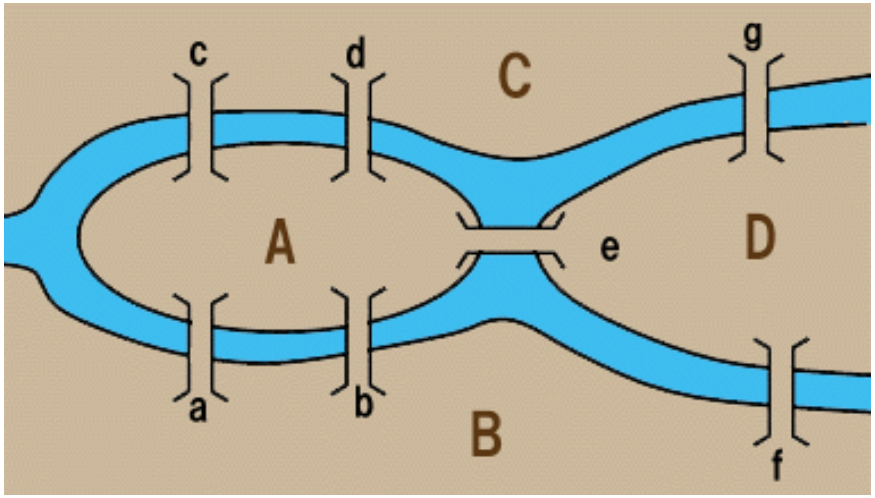
グラフ理論：ケーニヒスベルクの問題



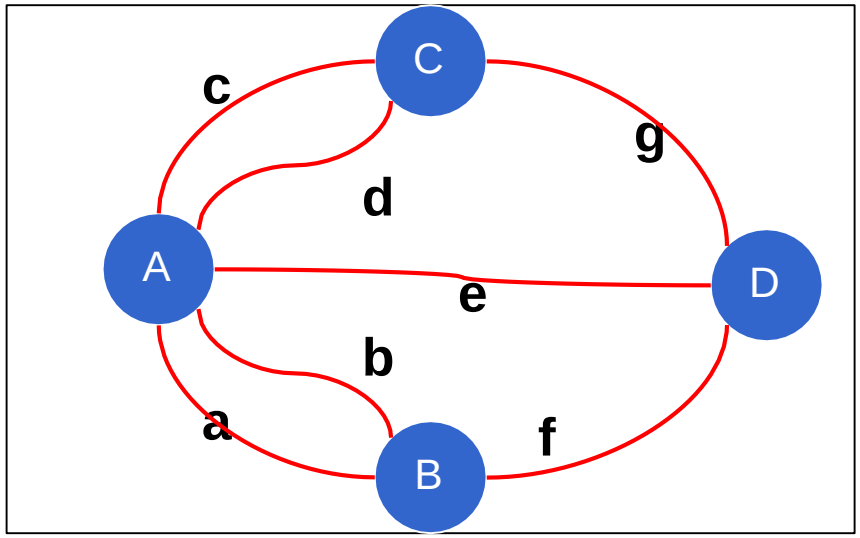
- ルール
  - スタート地点から出発して、スタート地点に戻る
  - 橋は一度しか渡れない

- ルールに従ってスタートからスタートへ戻れるか？

グラフ理論：ケーニヒスベルクの問題



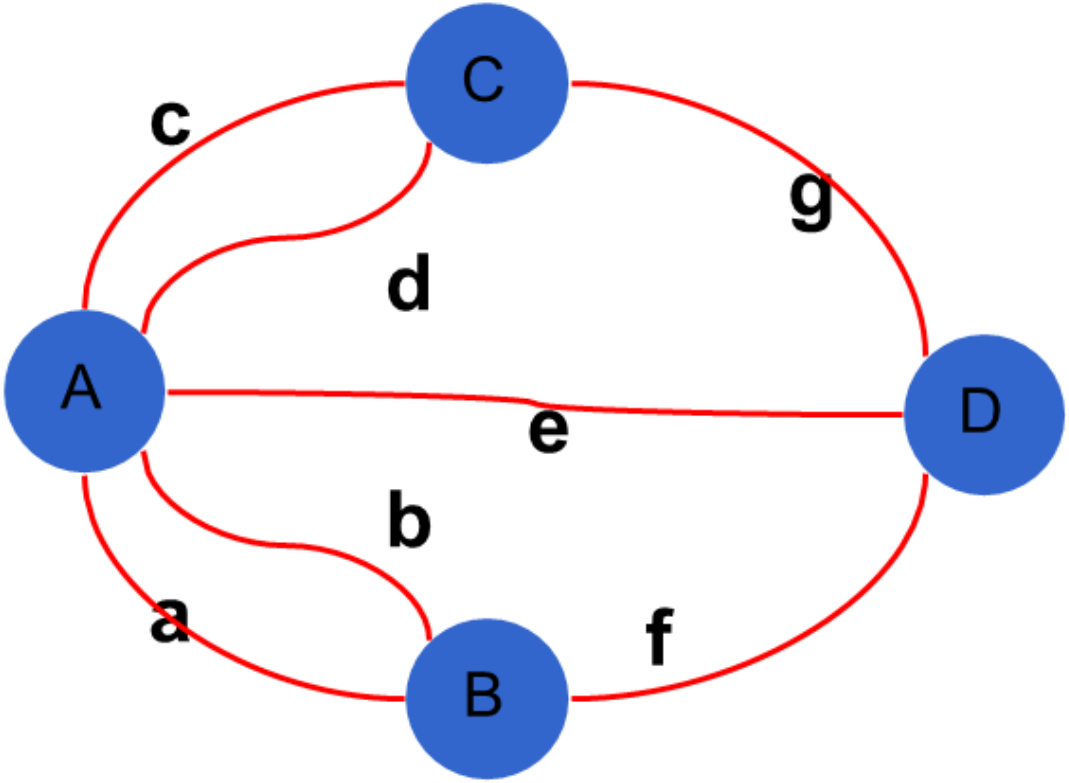
置き換え



オイラー（数学、物理学、天文学）
1707-1783

- このグラフが一筆書き可能であれば、ケーニヒスベルクの橋を全て1度ずつ通って戻ってくるルートが存在する
  ➔ 証明
- 1736年、「ケーニヒスベルクの問題」に対してオイラーが解法を示したのが起源

（ちなみに）一筆書き出来ますか？

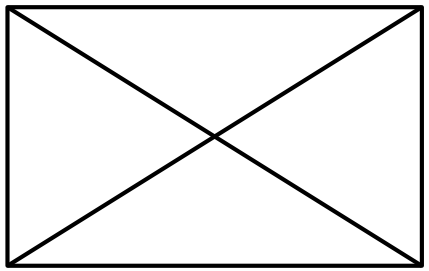


- 一筆書きの一般則（by オイラー）
  - 奇数個の線が出ている頂点が，0 or 2 個なら一筆書き可能

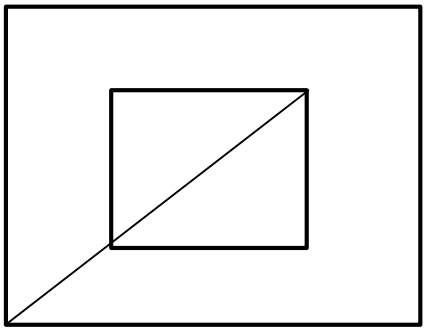
（演習）

- 以下のグラフは一筆書き出来ますか

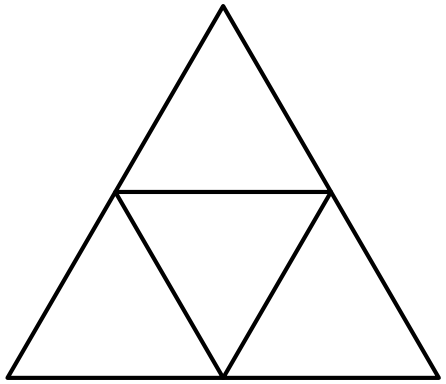
(a)



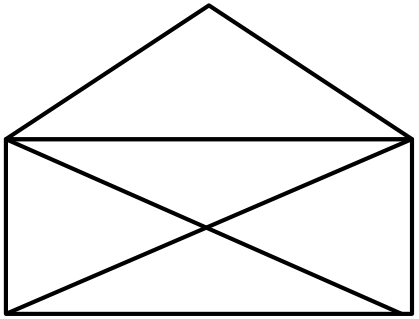
(b)



(c)

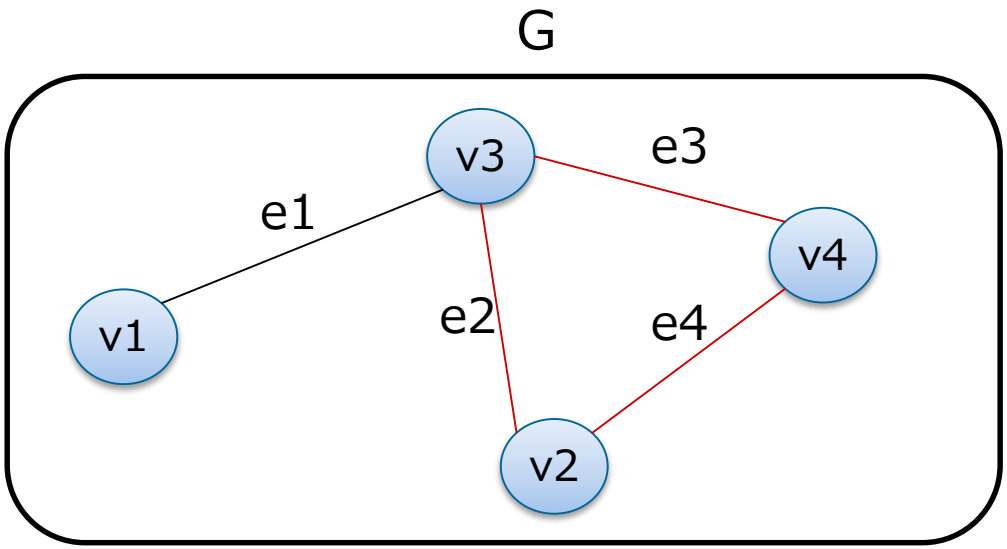


(d)



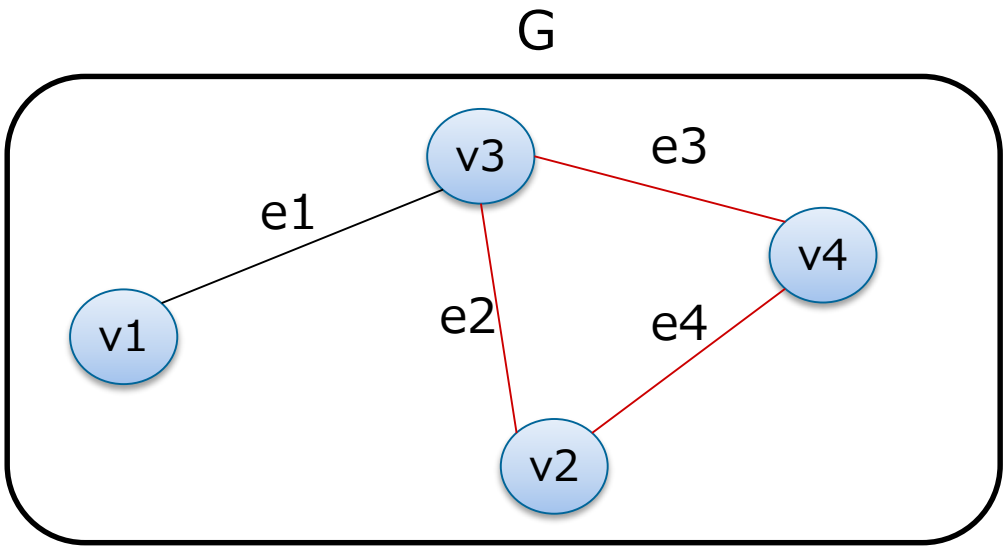
グラフ理論の基礎

- グラフとは
  - – V: 頂点 (vertex) からなる集合
  - – E: 点をつなぐ辺 (edge) からなる集合
  - – G: V と E の集合



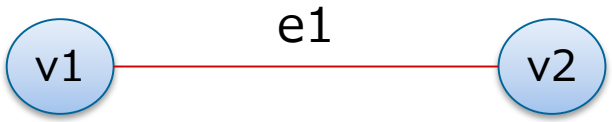
グラフの表現

- $V = \{v_1, v_2, v_3, v_4\}$
- $E = \{e_1 = (v_1, v_3),\ e_2 = \{v_2, v_3\},\ e_3 = \{v_3, v_4\},\ e_4 = \{v_2, v_4\}\}$
- $G = \{V, E\}$



隣接関係

- 点の隣接関係
  - v_1, v_2は隣接した点

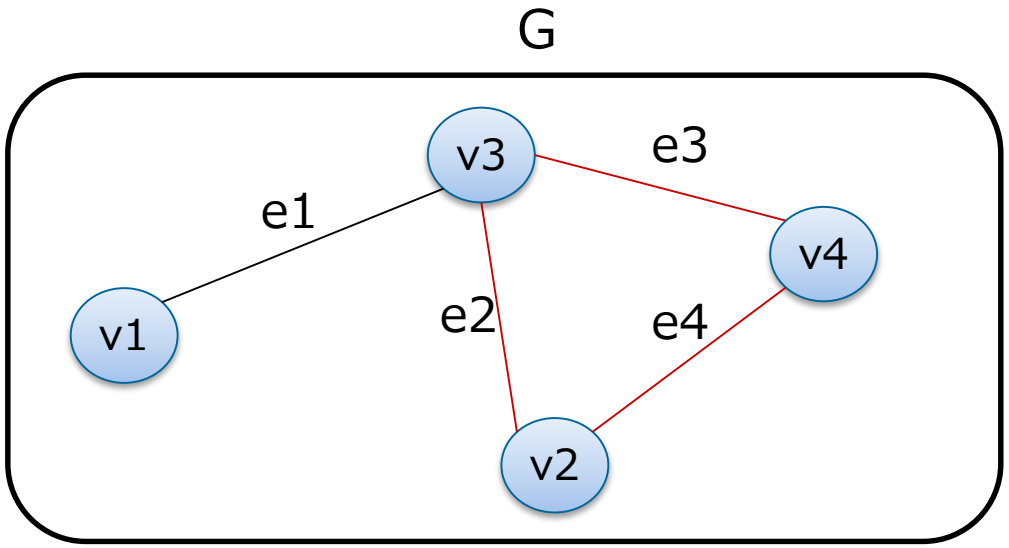


- 辺の隣接関係
  - e_1, e_2は隣接した辺

(確認問題)

- このグラフにおける隣接頂点，隣接辺をすべて述べてください

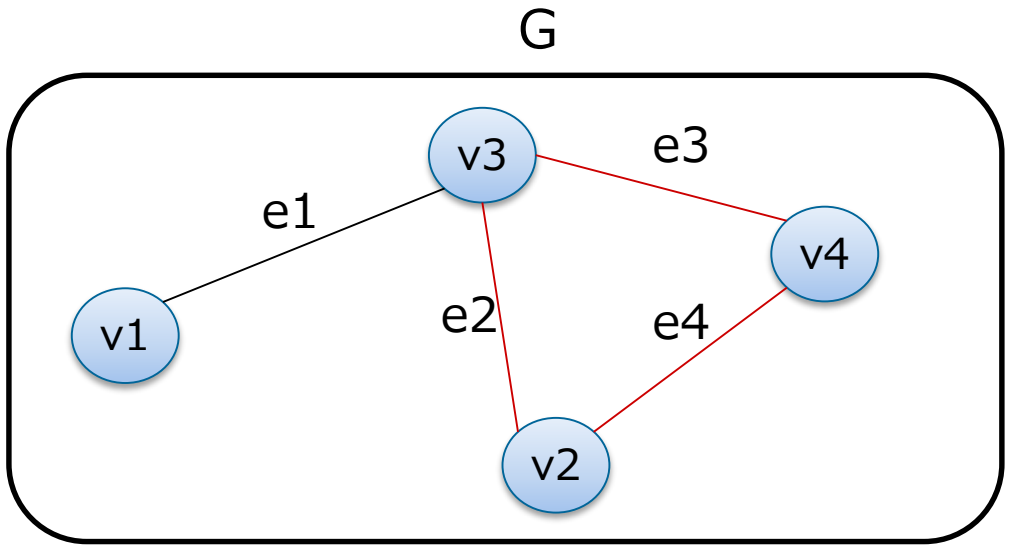
隣接頂点
(v1, v3), (　　), (　　), (　　)
隣接辺
(e1, e2), (　　), (　　), (　　) , (　　)



次数

- 次数（degree）
  - 頂点viが接続している辺の数
  - deg vi

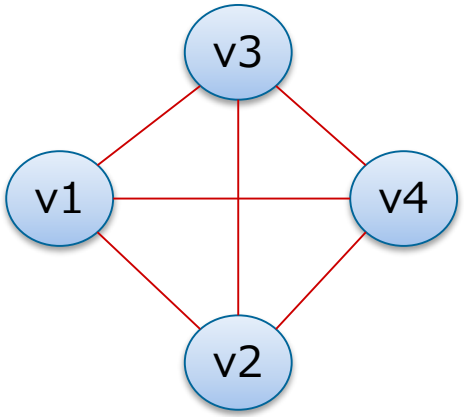
- 各頂点の次数はいくつか？

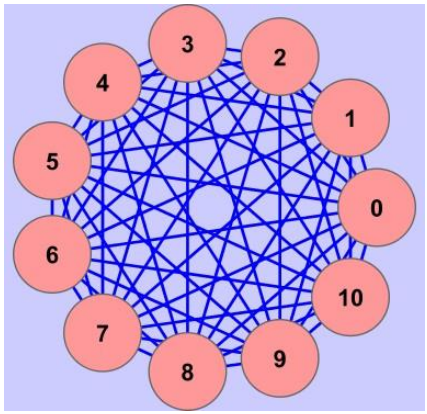


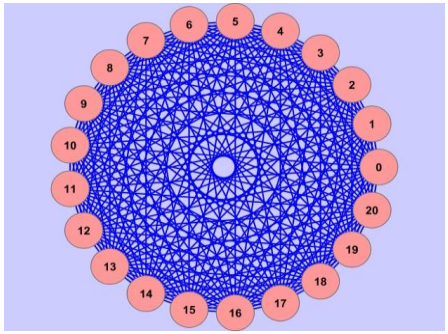
	$\mathbf{degree}(v_i)$
v_1	1
v_2	2
v_3	3
v_4	2

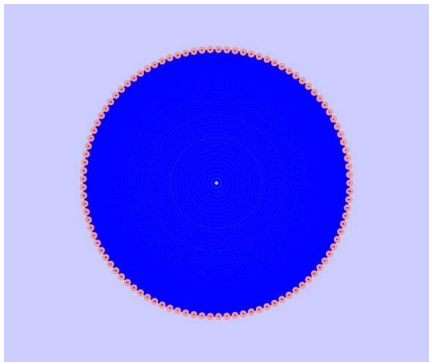
完全グラフ

- すべての頂点が辺でつながれているグラフ









問題

- 頂点の数がnの完全グラフにおける辺の数はいくつか？

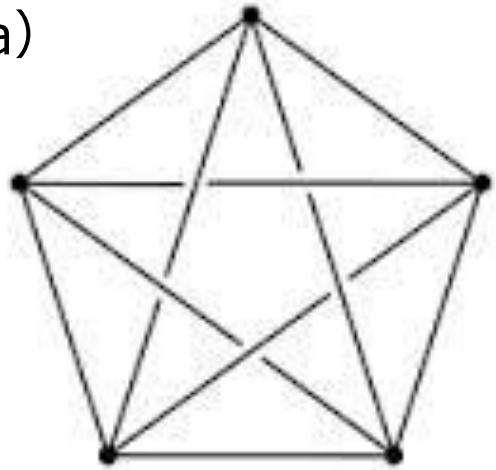
頂点v_1に接続する辺は$n-1$本 従って，全頂点から接続する辺の総数は$n(n-1)$ ただし，一つの辺は2頂点と接続する

$$\therefore \mathrm{N}_e = \frac{n(n-1)}{2}$$

(演習)

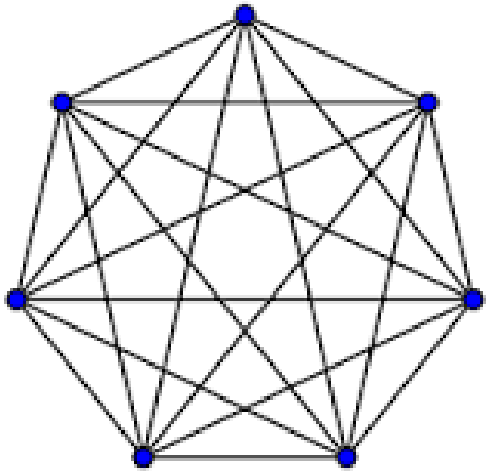
- 以下の完全グラフの辺の数 Ne を求めてください

(a)



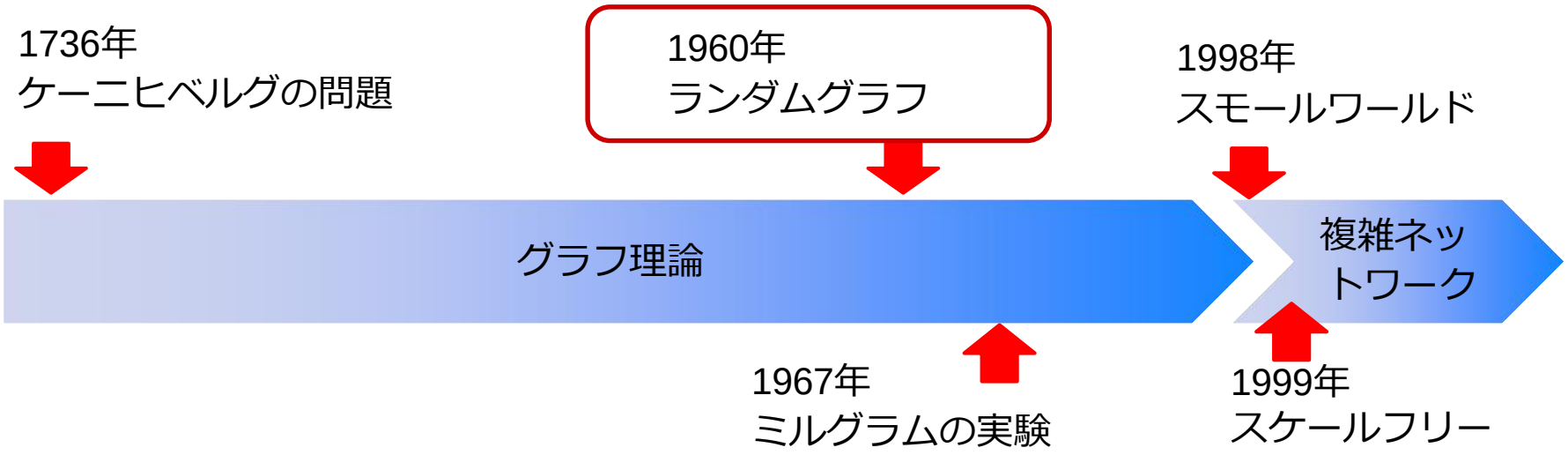
Ne =

(b)



Ne =

ネットワーク研究の歴史



ランダムグラフ

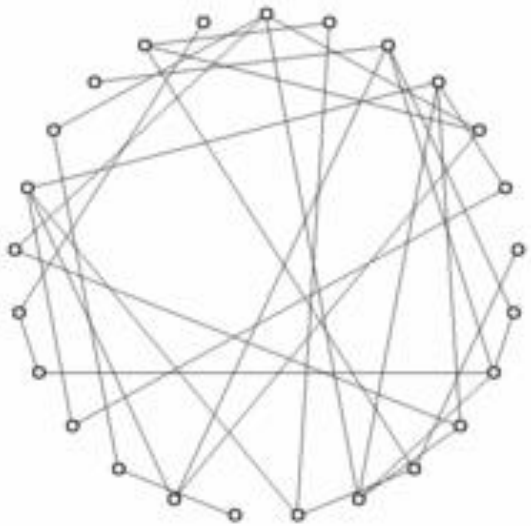
- ランダムグラフ
  - 1959年にエルディシュとアルフレッド・レーニィが考案（ERモデル）
  - 「複雑さとはランダムということである」と仮定
  - 各種の興味深い性質を有し、グラフの解析的な取り扱いを大きく進歩させた

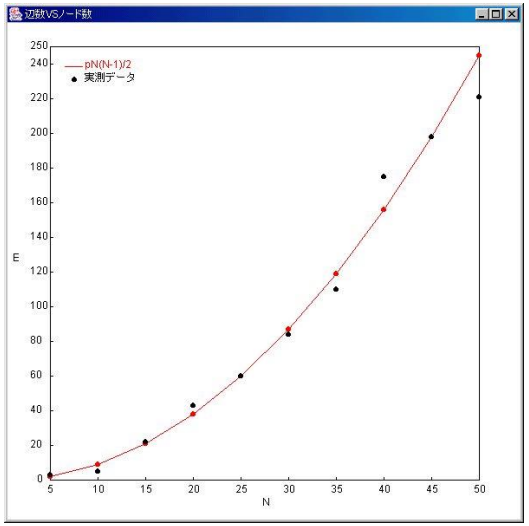
ポール・エルディシュ
（数学者）
1913-1996

- 生涯に約1500篇の論文を書いた

ランダムグラフ

- n個のノードがあるとき、2頂点間に確率pでリンクをつなぐ
- 確率pが小さいとリンクが少なくネットワークが分断され、pが大きいとリンクが多くてすぎてネットワークが密になる
- ある程度のpなら、現実のネットワークのようにリンクの数も中等度（ある程度疎）で乱雑なものとなる



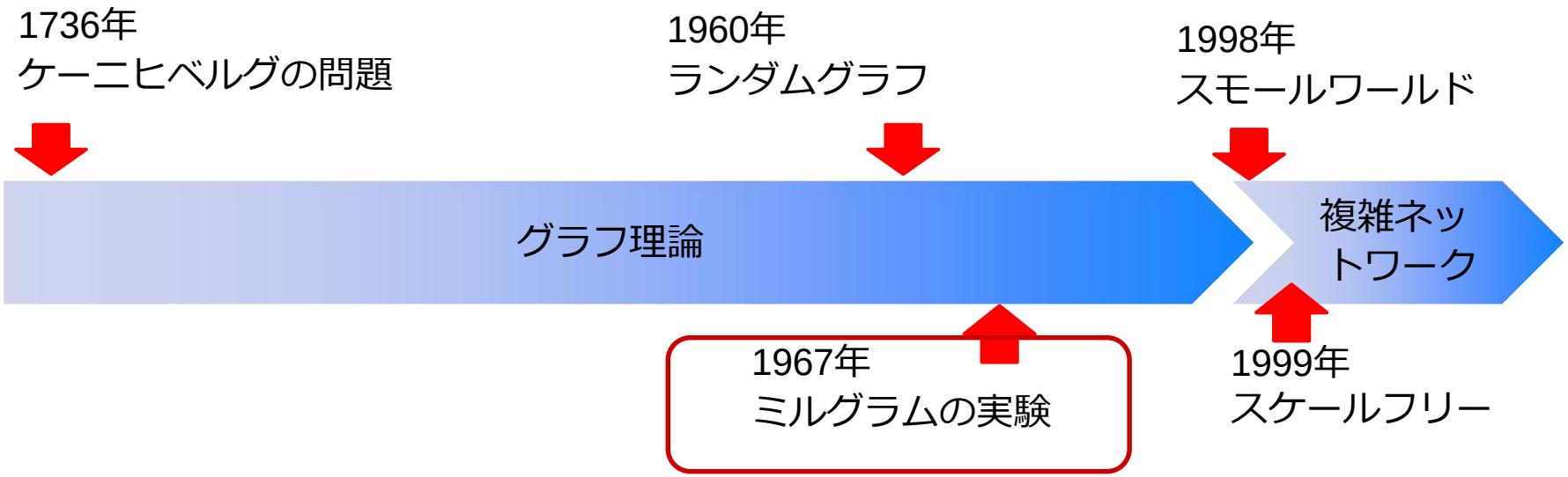


分布： $\frac{pN(N-1)}{2}$

p:辺連結確率

N:ノード総数

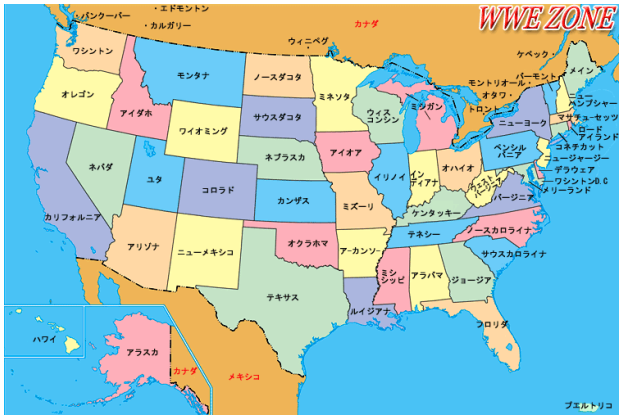
ネットワーク研究の歴史



- 現実のネットワークにおいて、ノードのリンクはランダムで平等か？
- ERモデルは妥当性に欠ける

ミルグラムの実験

- ミルグラムの実験（1967年）
  - 手紙をスタートからゴールまで届ける
    - スタート：カンザス州
    - ゴール：マサチューセッツ州

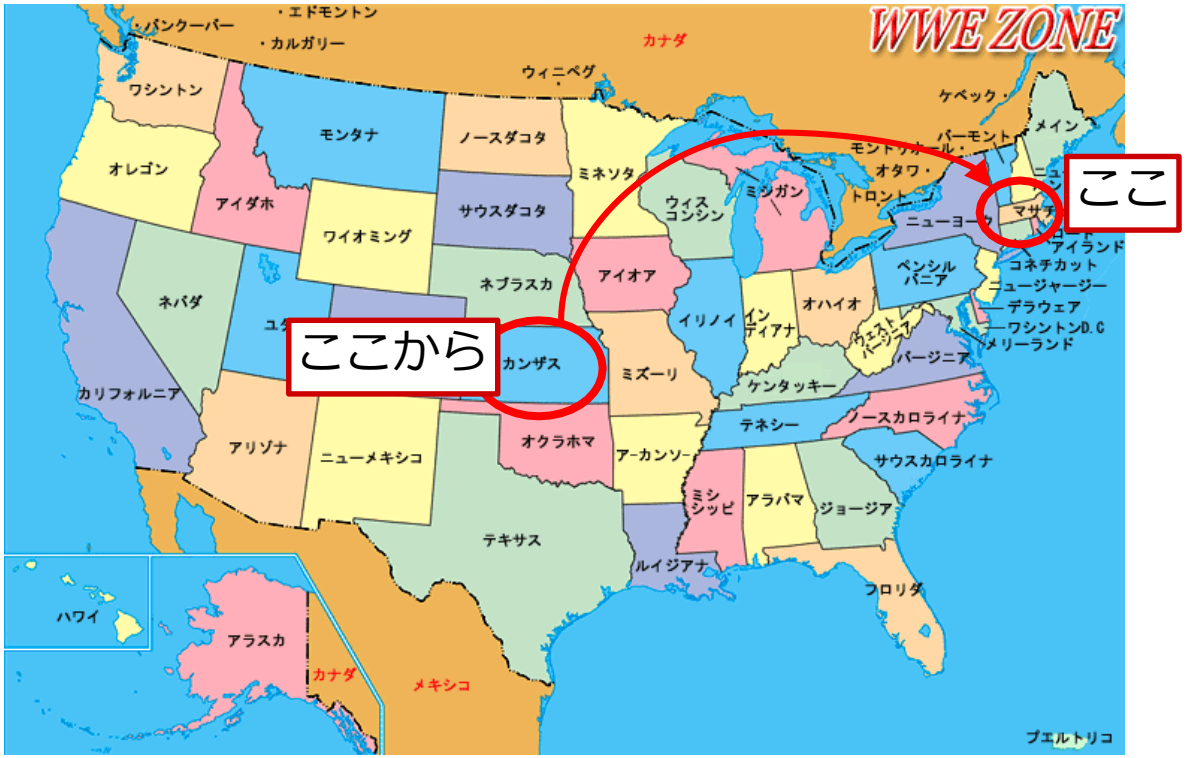


スタンレー・ミルグラム
（社会心理学）
1993-1984

  - 中継できるのはファーストネームで呼び合う仲のいい人だけ

ミルグラムの実験

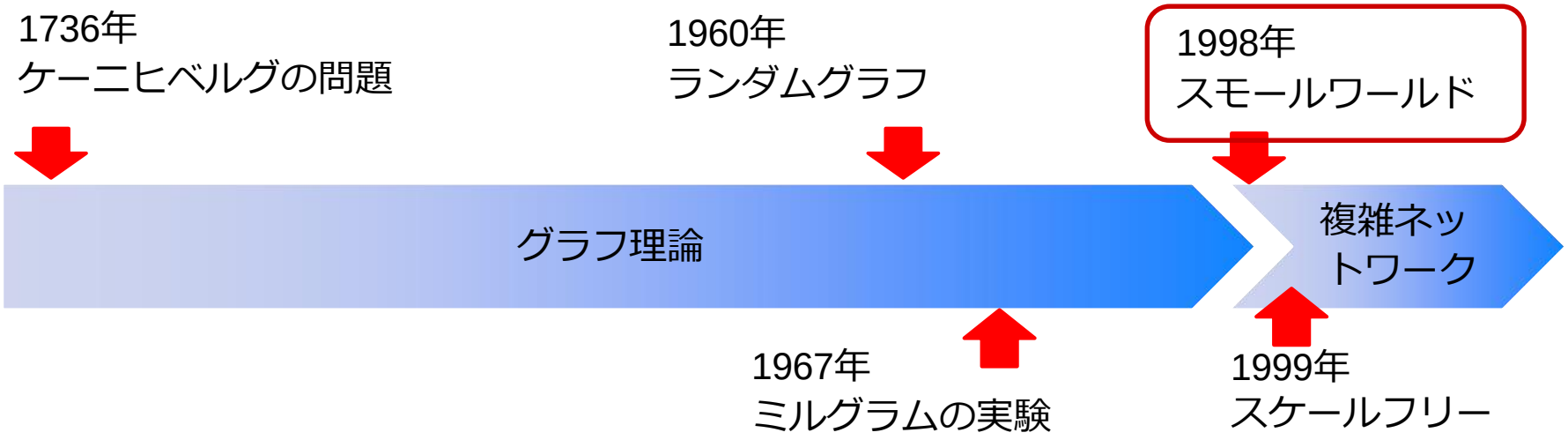
- 直接あるいは「その友人の友人」を通して最終的な対象に到達できそうな知人を選ぶ
- マサチューセッツまで何人で到達できるか？



ミルグラムの実験（結果）

- 160通のうち、42通が到着
- 平均の知人の輪：5.5人（➔6次の隔たり）
- 当時のアメリカの人口：2億人
- 人間関係は思っているよりもずっと小さな世界(small world)

ネットワーク研究の歴史

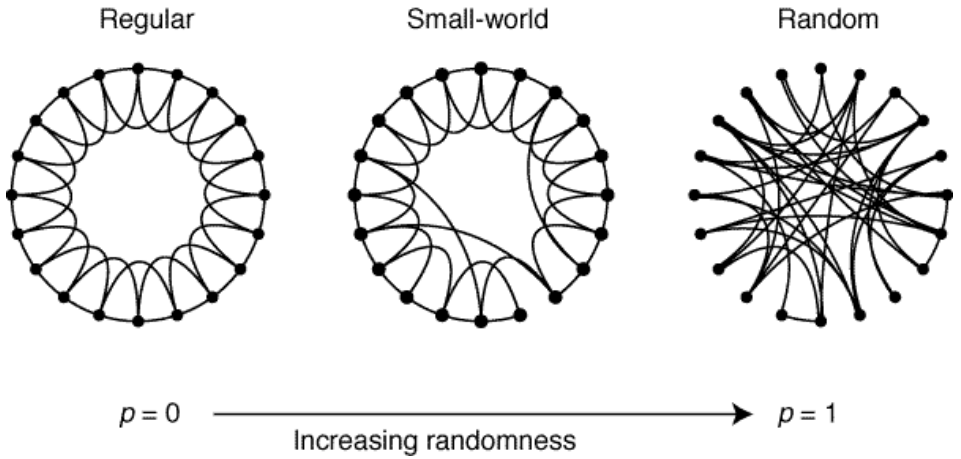


- 従来のネットワークは、完全に規則的か完全にランダムとされてきた
- しかし、現実の社会的ネットワークや自然界のネットワークはこの両極端の間にあるのではないか？

スモールワールド・ネットワーク

- 現実世界のスモールワールド性
  - 「世界は狭く固まっている」
  - 論文の共著関係でクラスター化があるかを実証
    - 平均距離が小さく(小さい世界であり)
    - クラスター性を持つ
  - 現実のネットワークに近いモデルとしてスモールワールドネットワーク（ワッツ・ストロガッツモデル)を提唱

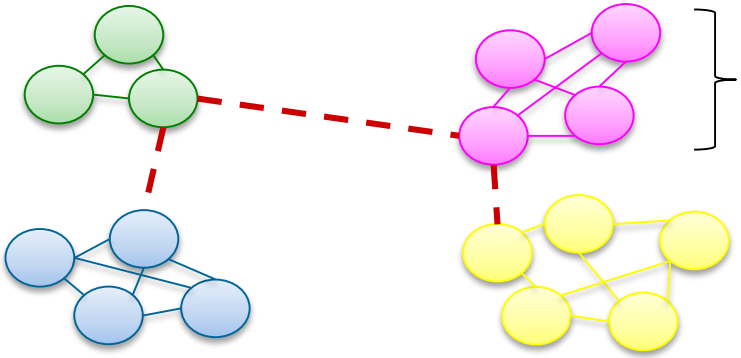
ダンカン・ワッツ
1971-



（補足）クラスター

- 転職に関するグラノヴェッターの研究（1969年）
  - 直接応募：20%
  - 転職会社：18.8%
  - 知り合いの紹介：56%
    - しばしば合う人：16.7%
    - たまに会う人：55.6%
    - ほとんど会わない人：28%
- 助けてくれるのは少し離れた人
  - 親しい友人間のネットワーク（強いつながり、クラスタ）だと孤立
  - 異なるコミュニティを結びつける人の存在（弱いつながり）
  - **「弱い紐帯（ちゅうたい）の強さ」**

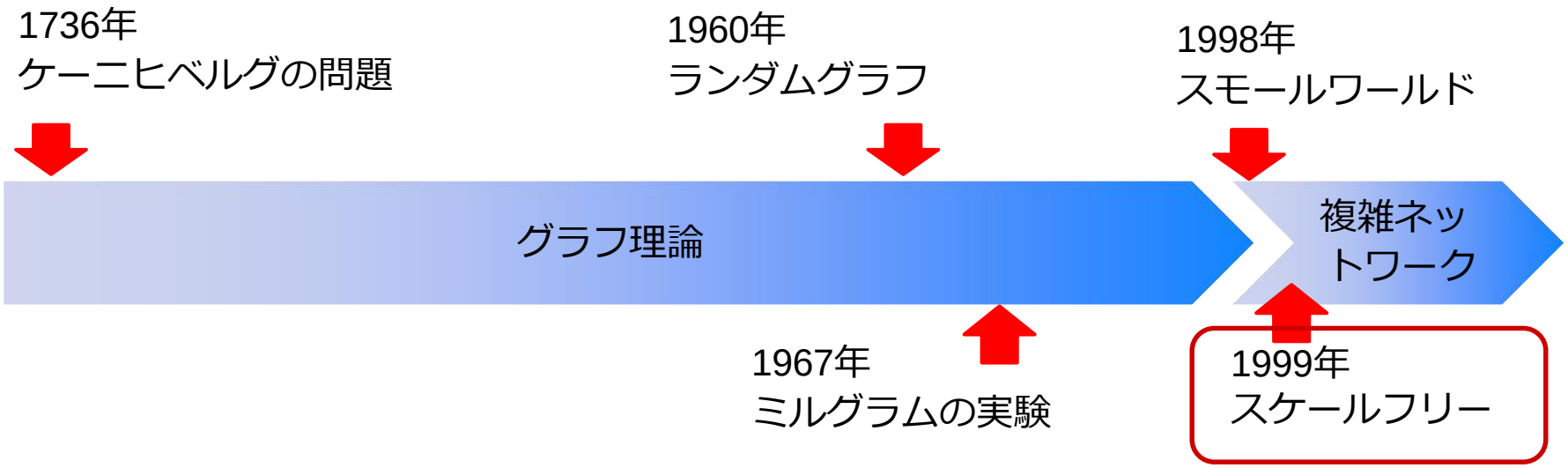
マーク・グラノヴェッター
（社会学者）
1943年-



クラスタ

ワッツはクラスター性を定量化（クラスター係数、クラスタリング係数）

ネットワーク研究の歴史

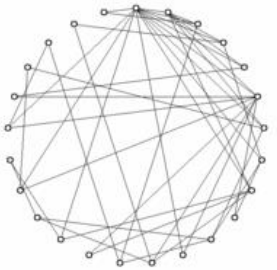


- 現実のネットワークにはハブ(リンクの数が非常に大きい頂点)が存在し、これはスモールワールド・ネットワークでは説明できない

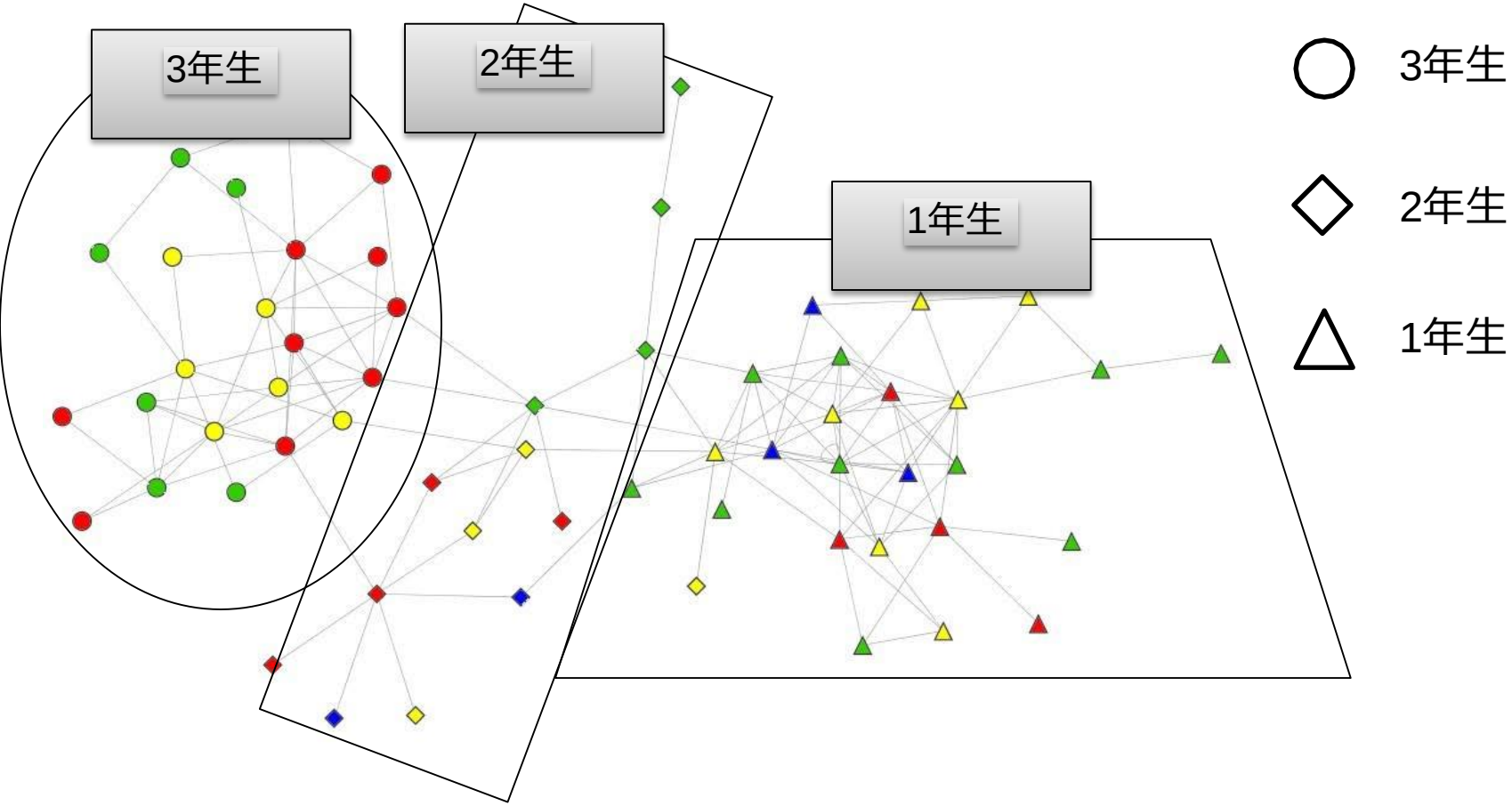
スケールフリー・ネットワーク

- 現実世界のスケールフリー性
  - 「世界は不平等だ」
  - ハブの存在
    - 大多数の人は友人の数は数名だが、「友人の数がずば抜けて多い」人物が何人かはいる
  - べき法則
    - 「収入分布は“べキ法則”にしたがう」
      - イタリアの経済学者ヴィルフレード・パレート
  - スケールフリー・ネットワーク
    - リンクの数とノードの数がべキ法則に分布＝スケールフリー・ネットワーク

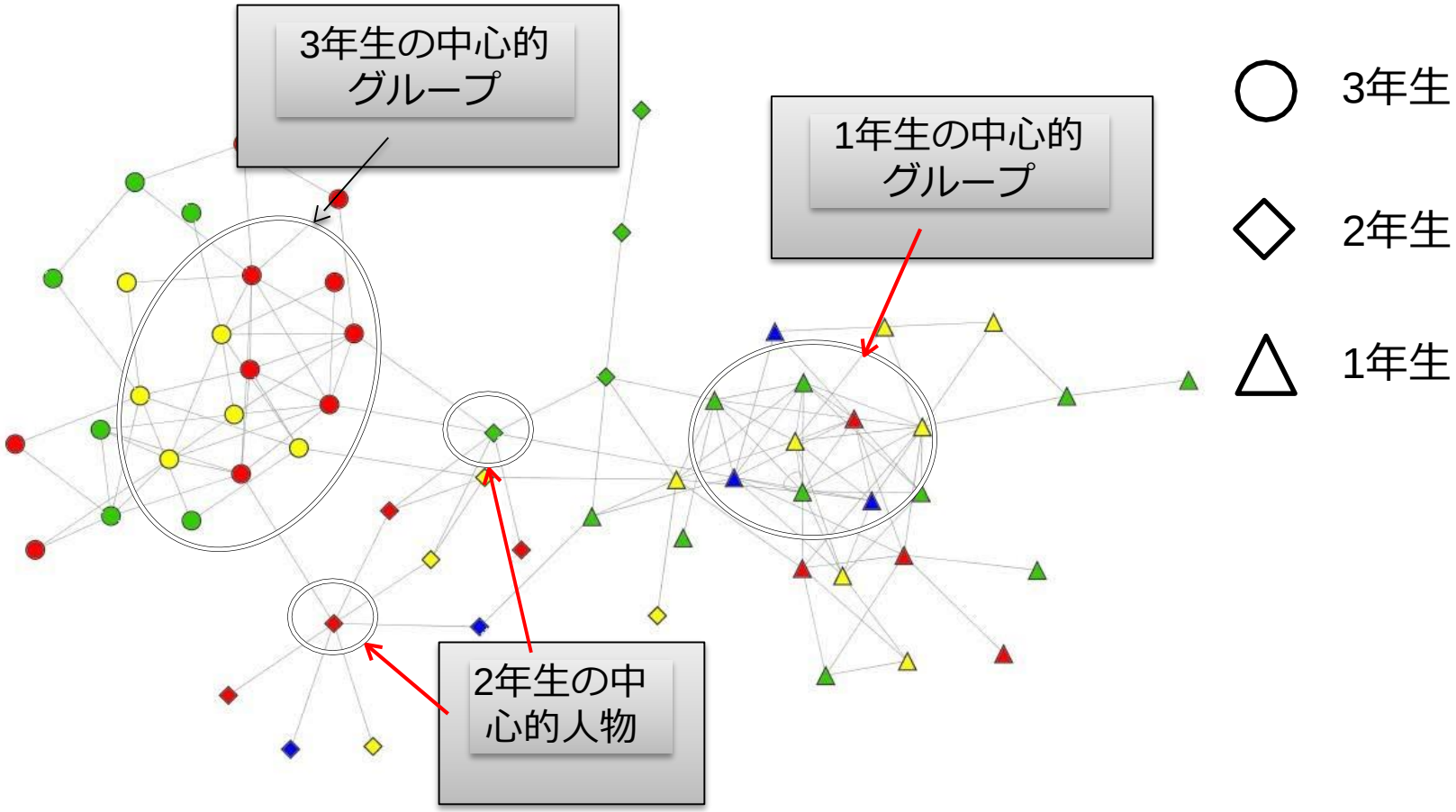
アルバート=ラズロ・バラバシ
1967-



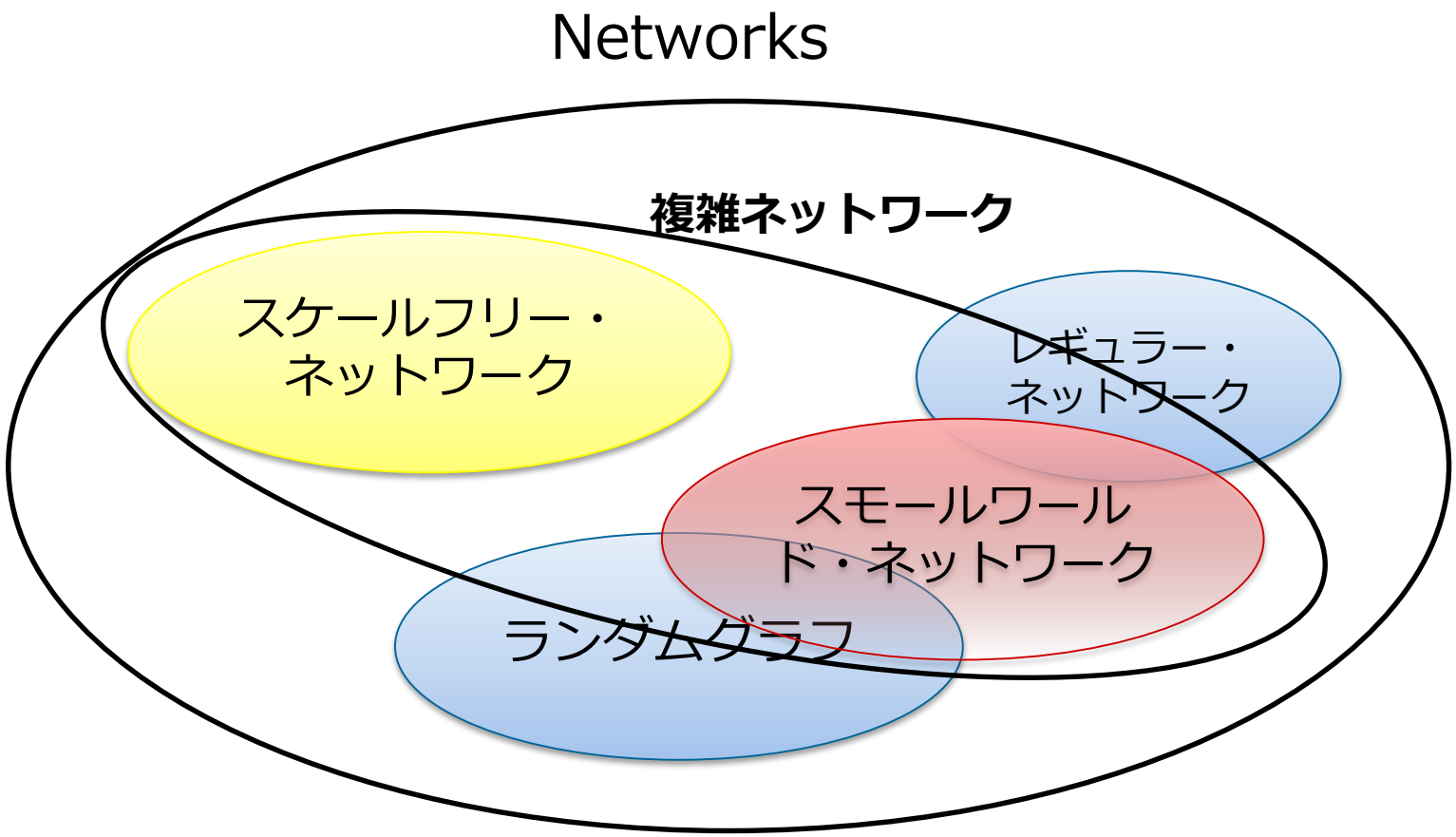
大学運動部の人間関係



大学運動部の人間関係



ネットワークのクラス

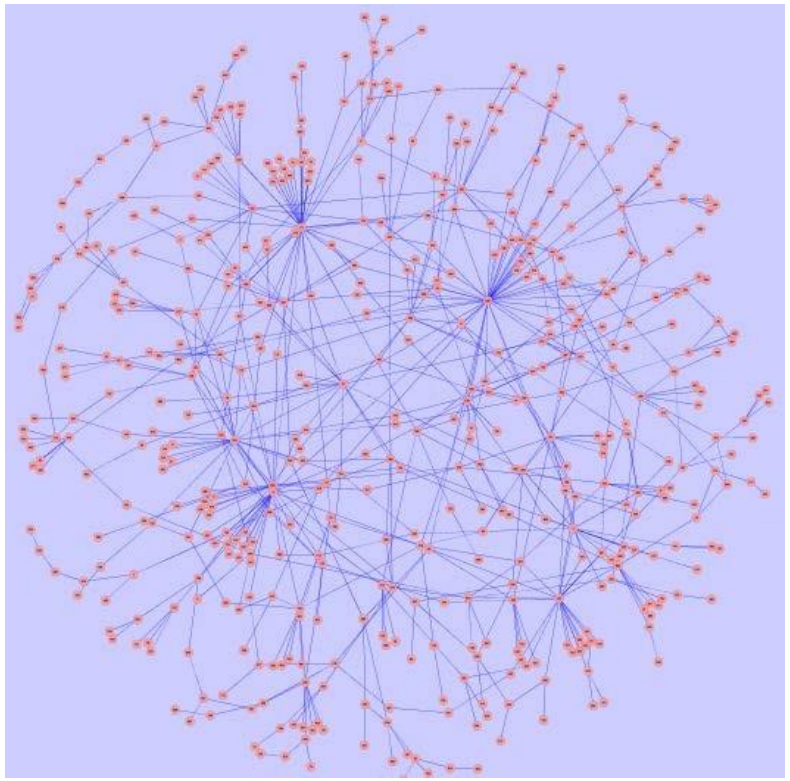

Networks
複雑ネットワーク
スケールフリー・
ネットワーク
レギュラー・
ネットワーク
スモールワール
ド・ネットワーク
ランダムグラフ

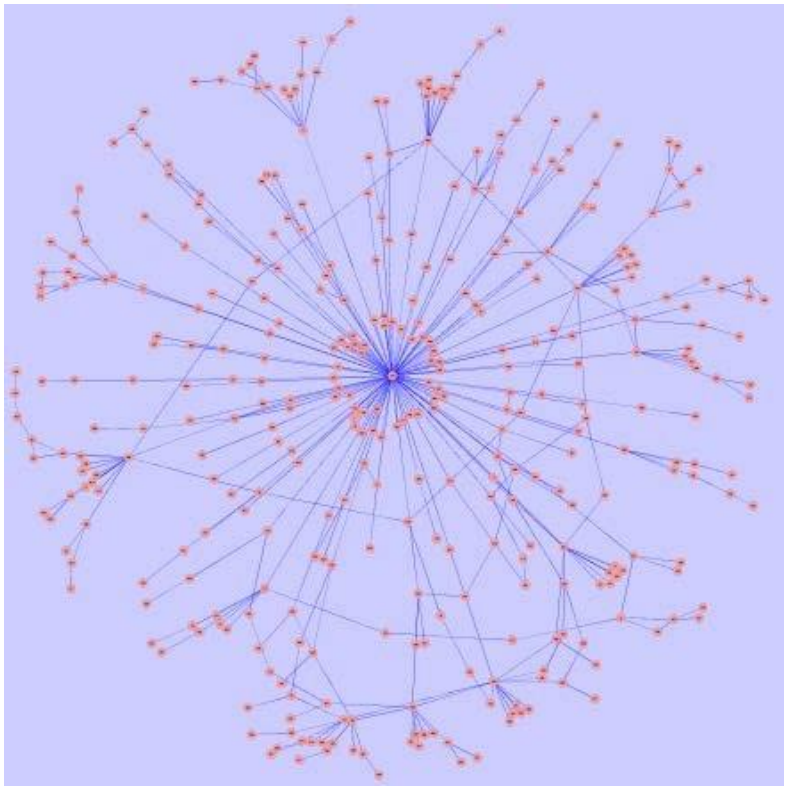
それから…

- 世界的に複雑系ネットワークの研究が盛んに
  - さまざまな現象が明らかに
- Webの発達により大規模なネットワークが出現
  - 電子データの利用
    ➔ 従来より簡単に巨大なネットワークを分析可能に
- ネットワーク分析に注目
  - ビッグデータ
  - ソーシャルメディア

ネットワーク構造、
ネットワーク指標

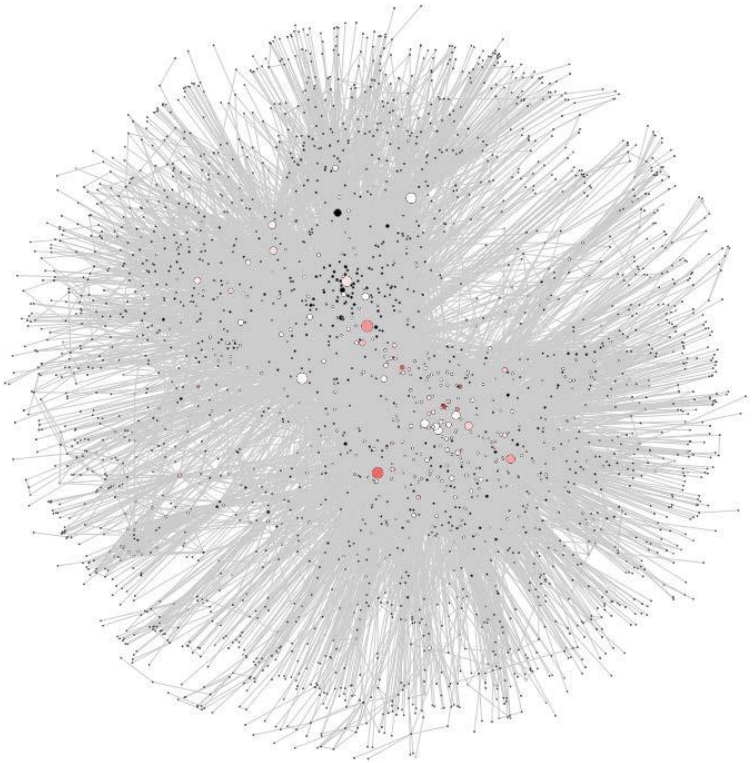
ネットワークの違い

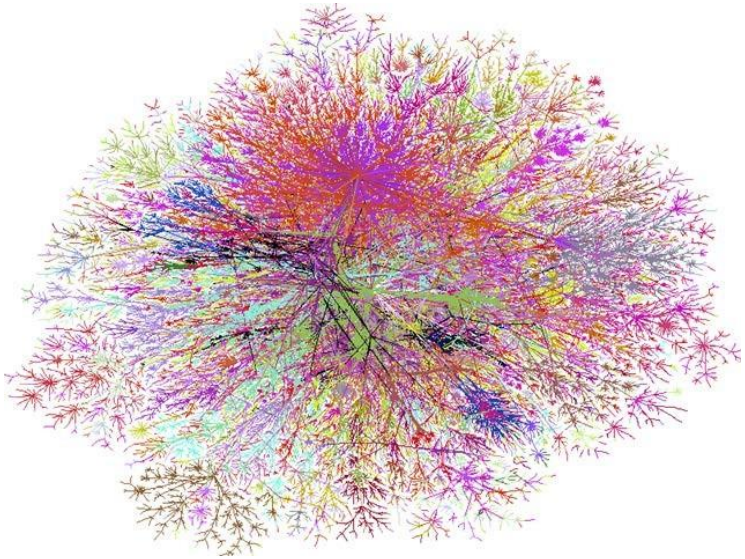




- 確かに違いがある
  - 具体的には何が違うのか？

- 大規模になると人の目では判断不可能





ネットワークの分析

- 直接眺めても良く分からない

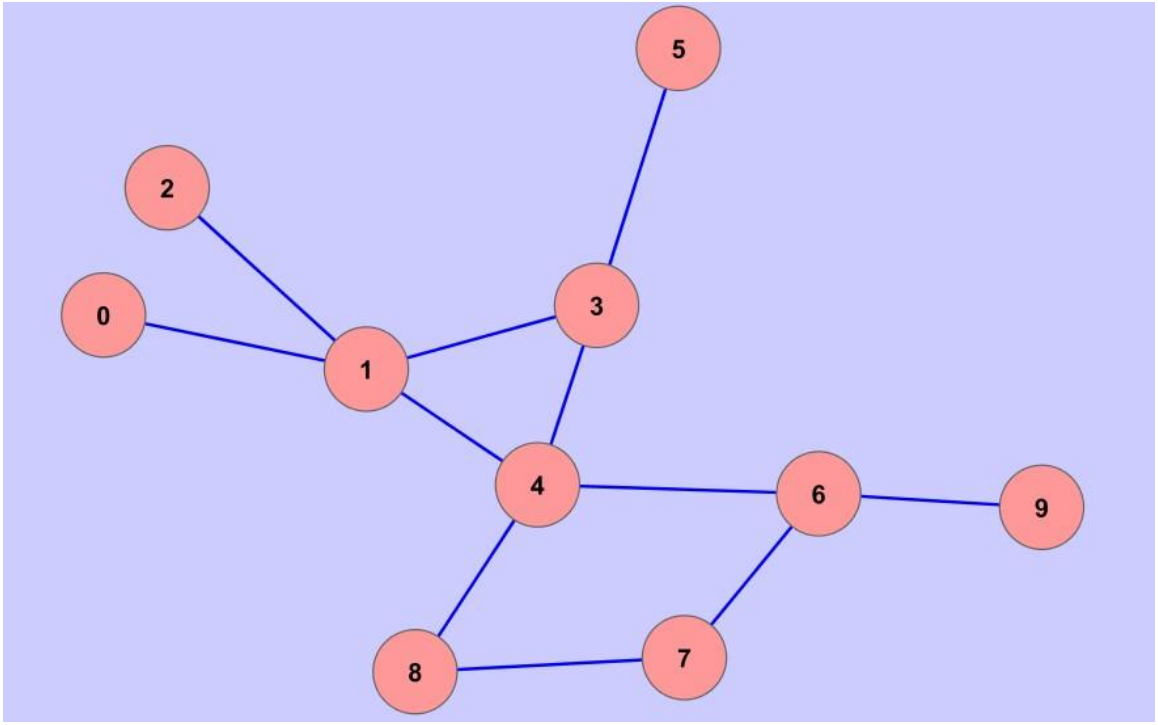
- なんらかの指標で評価する
  - ネットワーク特徴量の利用
  - このネットワークは大きい
  - このネットワークは密度が高い
  - このノードは重要だ
  - etc…

ネットワークの特徴量

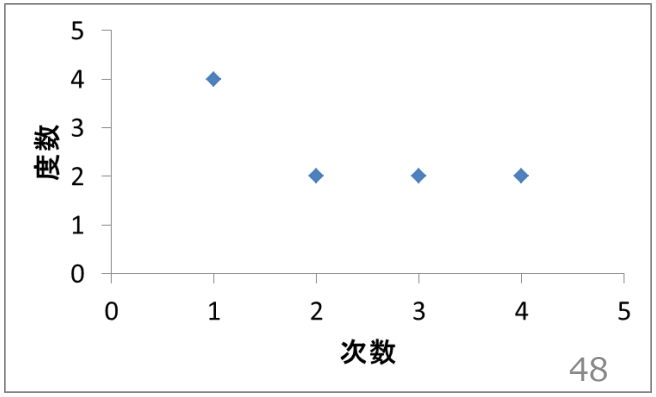
- (1) 次数分布
- (2) 平均経路長
- (3) クラスタ係数
- (4) 次数相関
- (5) 中心性
  - 次数
  - 近接中心性
  - 媒介中心性

(1) 次数分布

- 各ノードの次数がいくつあるか，その分布



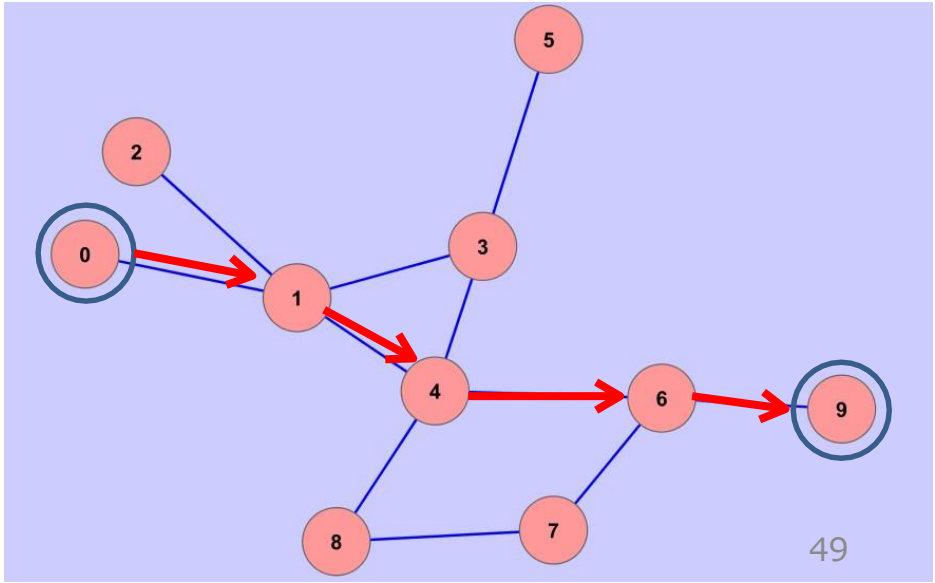
次数	ノード数	割合
1	4	0.4
2	2	0.2
3	2	0.2
4	2	0.2



(2) 平均経路長

- 経路長
  - あるユーザからあるユーザまで到達するのに必要なステップ数
  - ステップ数・距離・距離など呼ぶ

- ネットワークのおおよその大きさを表現可能

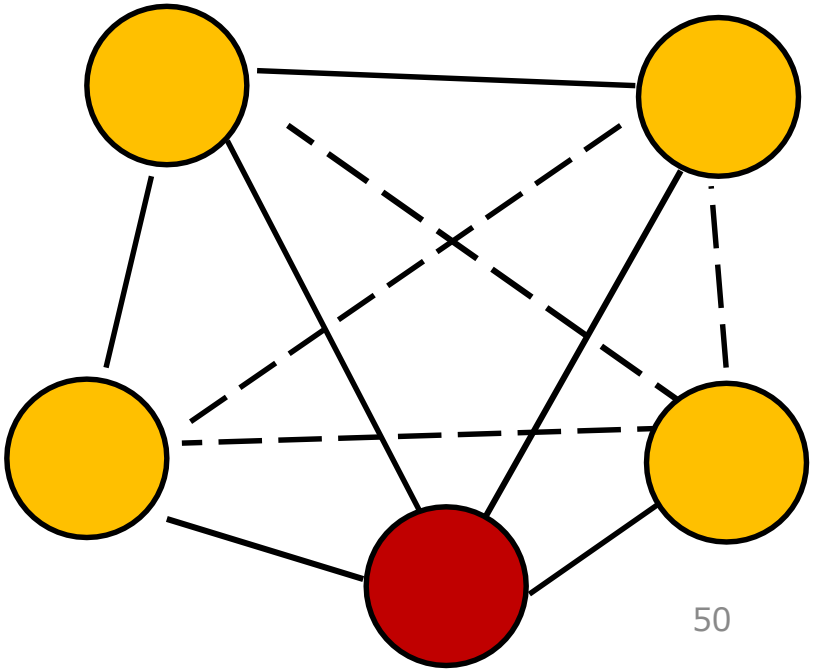


(3) クラスタ係数

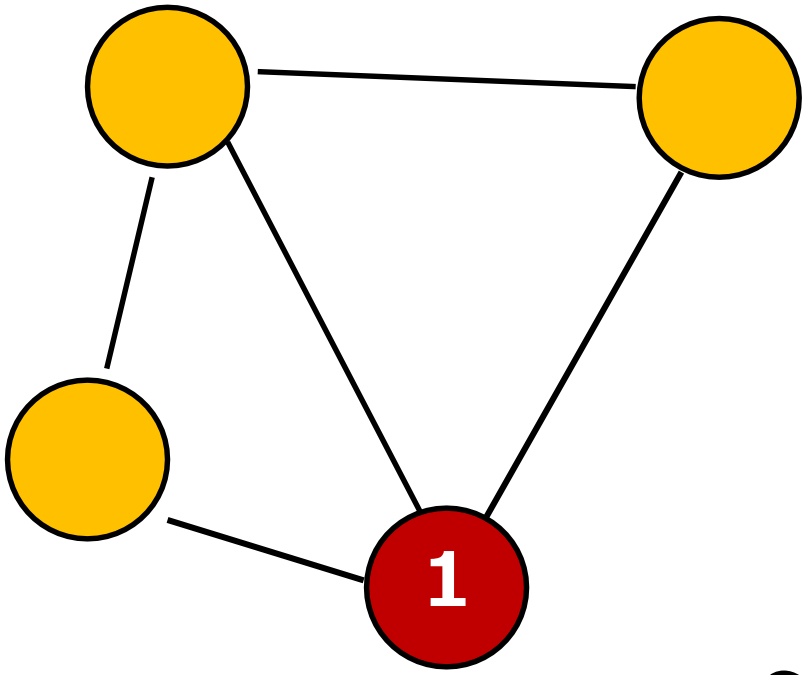
- あるノードの隣接ノード同士が隣接ノードである割合
  - 全ノードのクラスタ係数の平均がネットワークのクラスタ係数
- ネットワークの凝集性を表現
- クラスタ係数が高いネットワークは関係の密度が高い

$$C_i = \frac{v_i \text{を含む三角形の数}}{k_i (k_i - 1)\ /2}$$

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i$$



クラスタ係数の計算例（C_1）



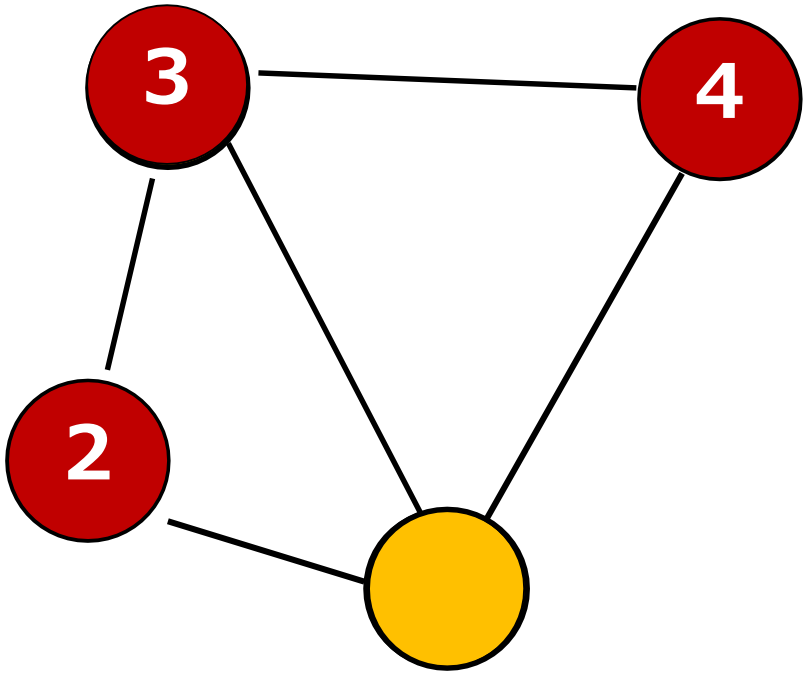
$$C_i = \frac{v_i \text{を含む三角形の数}}{k_i(k_i - 1)/2}$$

$$C = \frac{1}{N}\sum_{i=1}^{N} C_i$$

C_1 = 2 / {3(3 - 1) / 2} = 0.66

（演習）
C_2 C_3 C_4 と C を計算してください

$$C_i = \frac{v_i \text{を含む三角形の数}}{k_i(k_i - 1)\ /2}$$



$$C = \frac{1}{N}\sum_{i=1}^{N} C_i$$

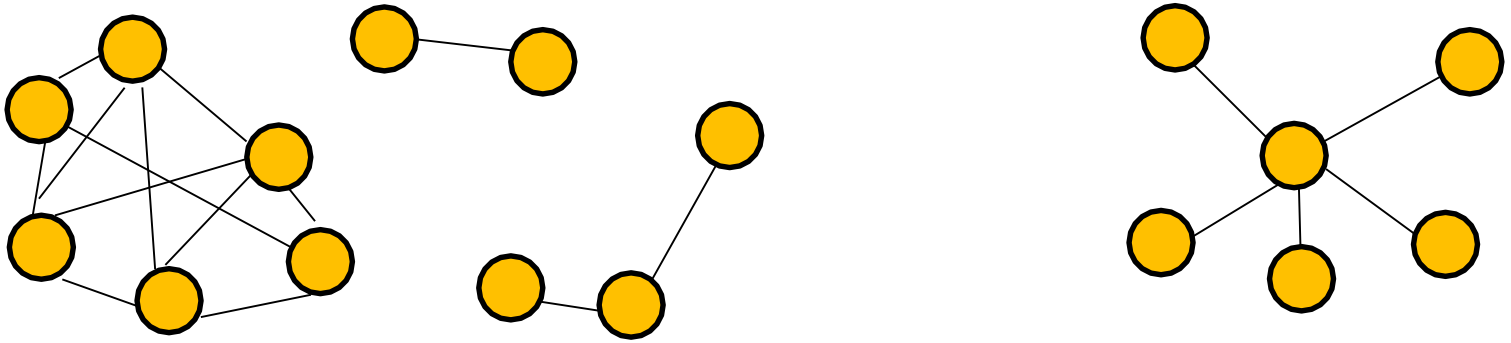
C_1 = 0.66
C_2 =
C_3 =
C_4 =

C =

ki：ノードviの次数

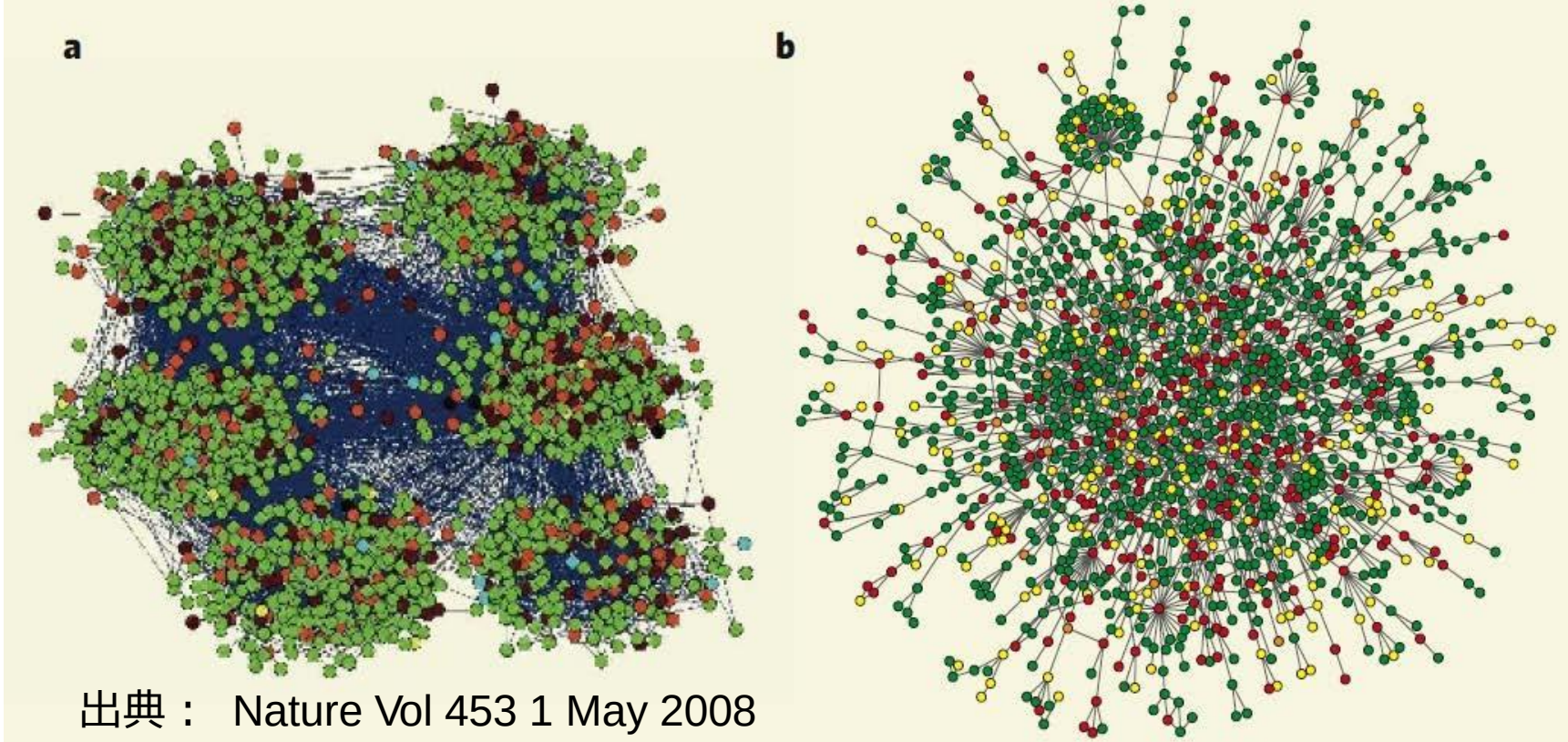
(4) 次数相関（同類選択性）

- 接続ノード間の次数の相関係数
- 正の数
  - 高次数ノードは，高次数ノードと隣接しやすい
- 負の数
  - 高次数ノードは，低次数ノードと隣接しやすい



正

負



出典： Nature Vol 453 1 May 2008

人間関係、共著関係など

タンパク質、インターネット

(5) 中心性

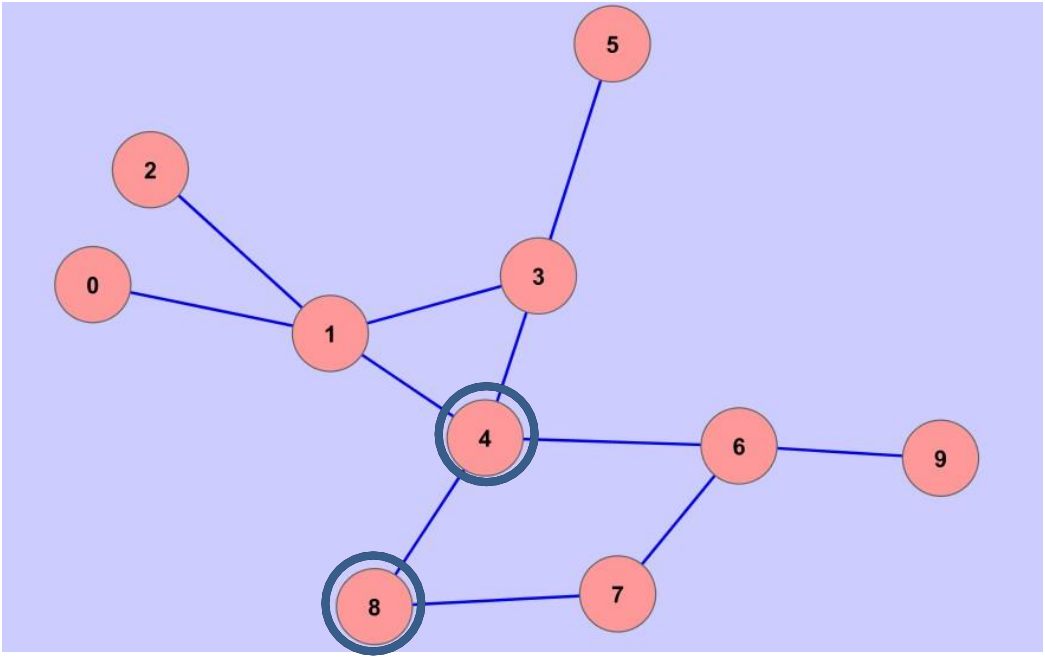
- ノードの特徴を調べる特徴量
  - ネットワークではなく個々のノードに着目
  - 局所性の分析

- 次数中心性
  - 次数そのもの

- 近接中心性
  - すべてのノードからの距離

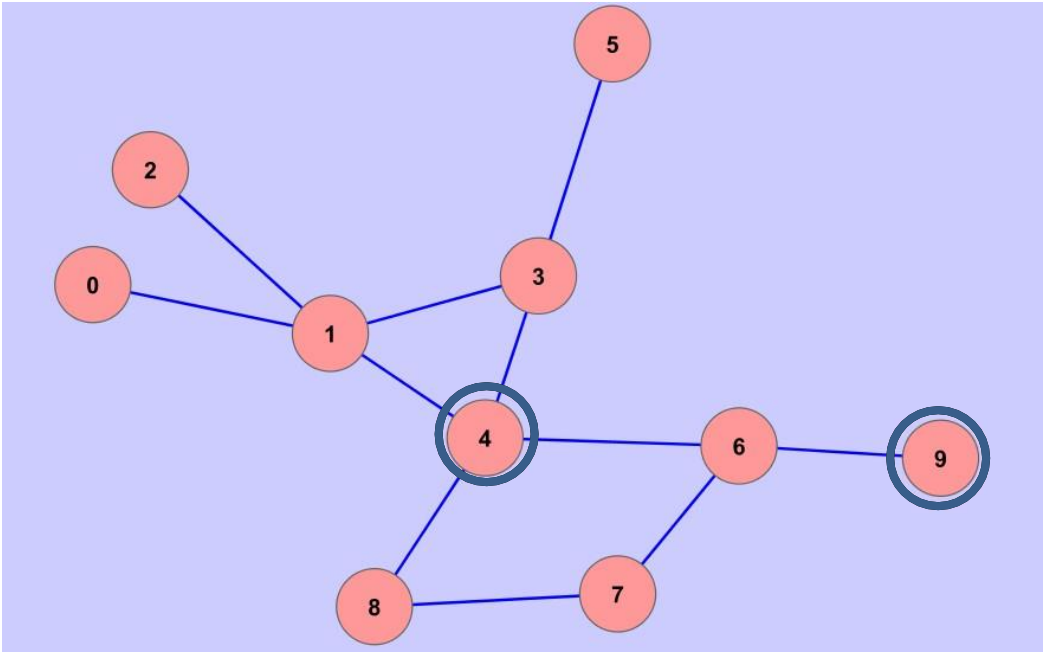
- 媒介中心性
  - 他のノードにたどり着くために当該ノードを通らなければいけない割合

次数中心性



- 次数(リンク数)そのもの
  - ノード4の次数中心性 = 4
  - ノード8の次数中心性 = 2

近接中心性

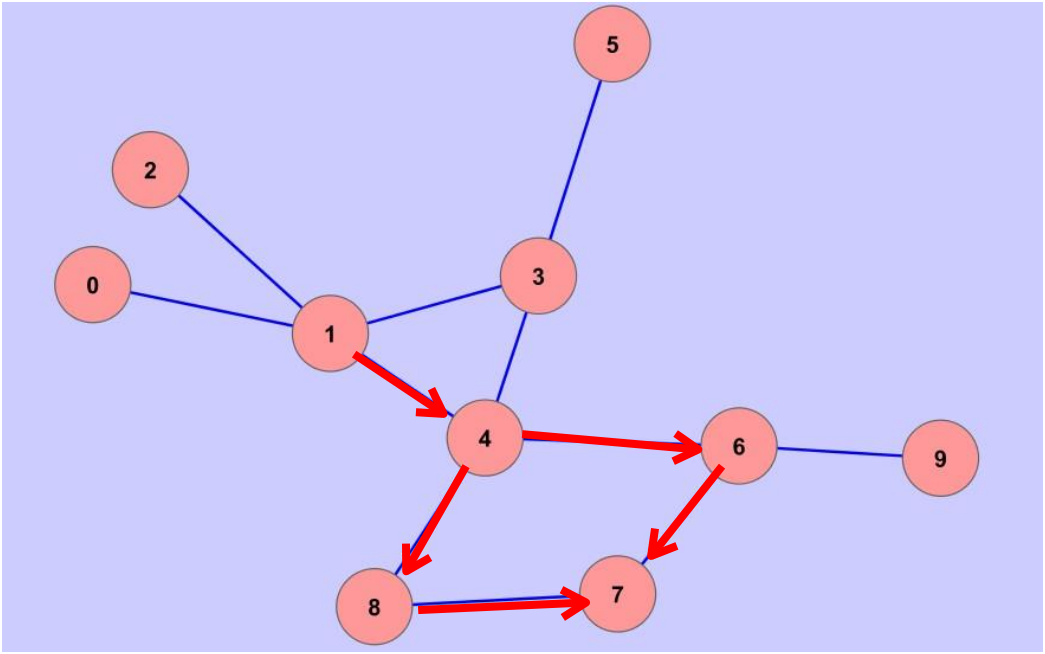


- 各ノードの経路長の逆数

ノードiの近接中心性　＝　(ノードの数 - 1) / (他ノードとノードiの距離の総和)

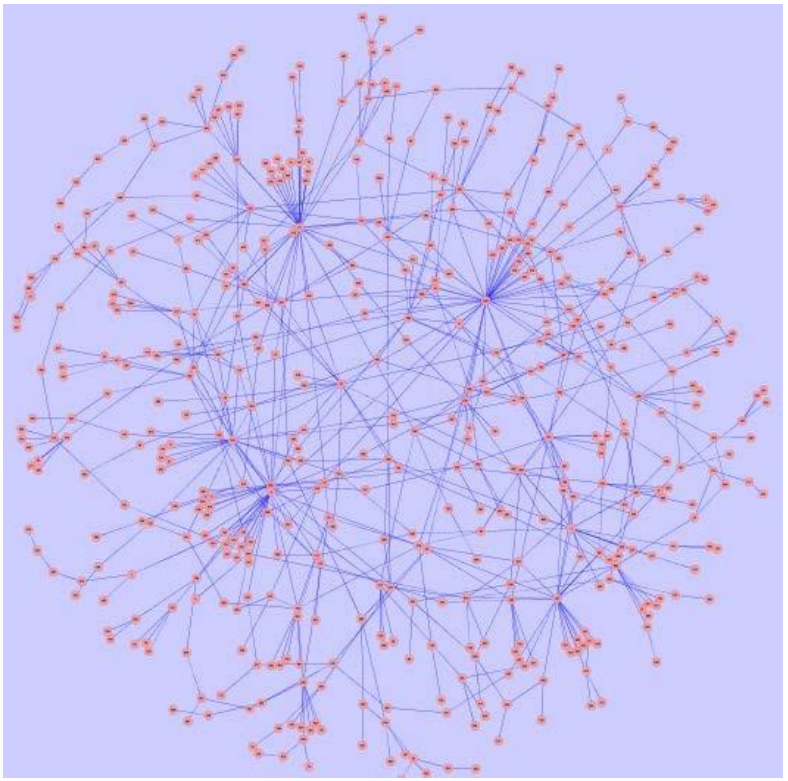
  - ノード9の近接中心性=9/26=0.346
  - ノード4の次数中心性=9/14=0.64

媒介中心性

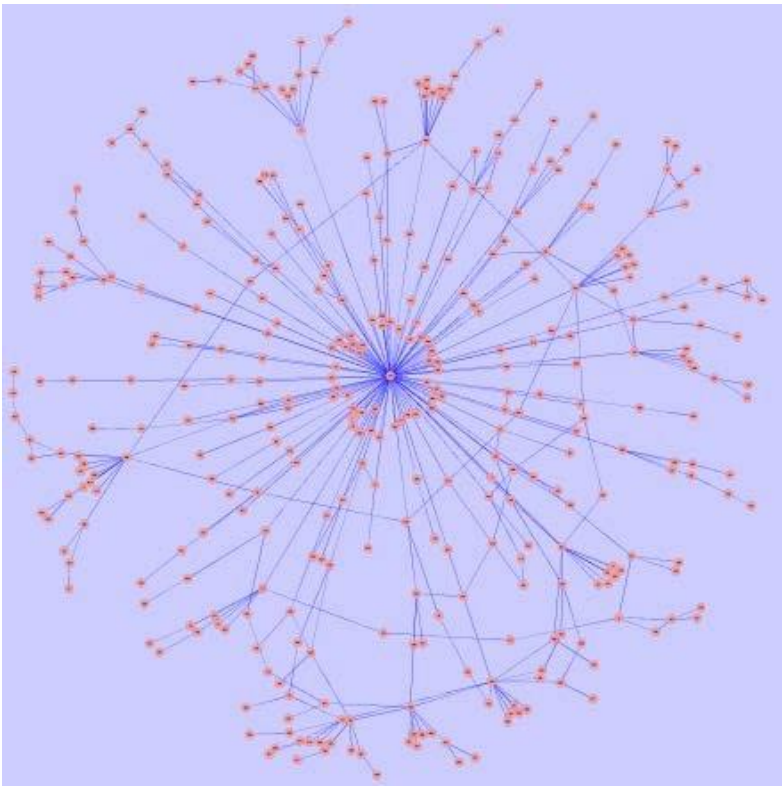


- 全最短経路中ノードv_iを通る経路の割合
- 情報の流れの中心性をはかるのに利用
  - このノードがないとネットワークがバラバラになるとき に高い値となる
- ノード1→7の最短経路は2パターンある
  - – 2回ともノード4を通る
  - – 1回はノード6を通る
  - – ノード4のほうが媒介中心性は高い

ネットワークの違い



- 平均経路長：6.91
- クラスタリング係数：0.113
- 次数相関：-0.0198



- 平均経路長：5.42
- クラスタリング係数：0.0905
- 次数相関：-0.0500

（演習）

・以下のネットワークのネットワーク指標を求めて下さい

ネットワーク指標	
最大次数	
平均次数	
クラスタリング係数C