情報システム実験B: 大規模ソーシャルデータ分析（KC-14）

ソーシャルデータ分析の基礎

担当教員：塩川 浩昭

ソーシャルデータ分析とは何か？

- **ソーシャルデータ（ソーシャルネットワーク）分析**

> *Social network analysis (SNA) is the process of investigating social structures through the use of networks and graph theory.[1] It characterizes networked structures in terms of nodes (individual actors, people, or things within the network) and the ties, edges, or links (relationships or interactions) that connect them.*

Wikipedia (https://en.wikipedia.org/wiki/Social_network_analysis)より引用．取得日 2017/10/11.

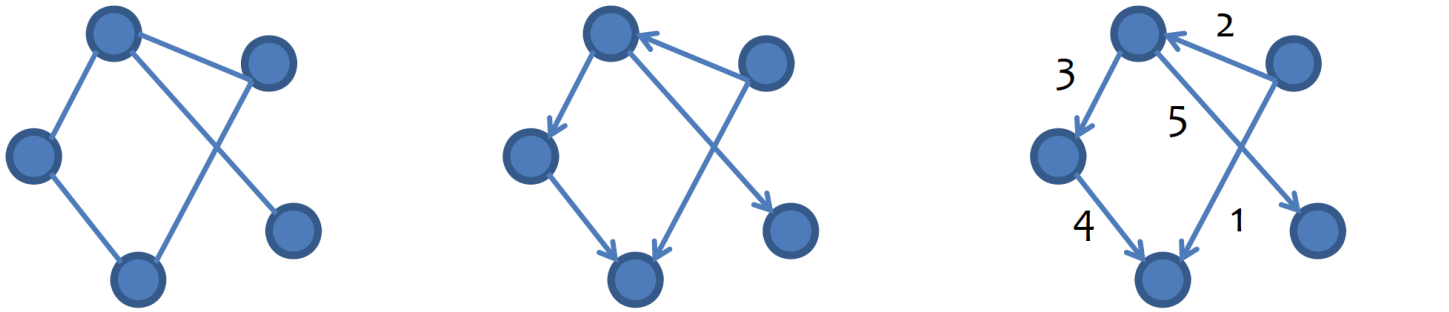
（意訳）ソーシャルデータ分析は，人や物を表すノードとそれらの関係性を表すリンク（エッジ）からなるネットワーク（グラフ）構造を用いて，社会構造を調査する手続きである．

すなわち，ソーシャルデータ分析では，グラフ構造を分析することで，そこから社会的な構造を明らかにすることが課題である．

グラフ

- **Data Entity間の関係性を表現するためのデータ構造**
  - ノード（頂点・点）とエッジ（辺）の集合

無向グラフ　　有向グラフ　　重み付き有向グラフ



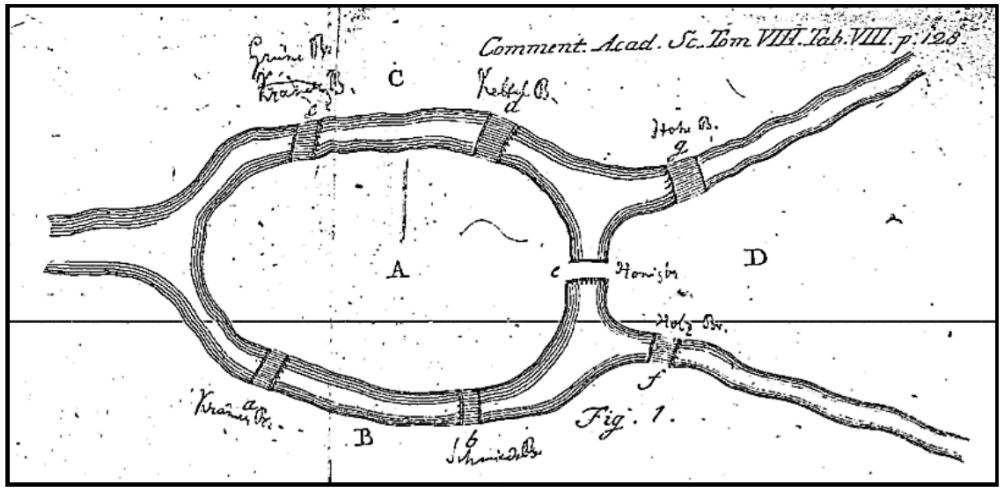
  - 非常にシンプルなデータ構造だが，情報科学の分野で幅広く応用されている

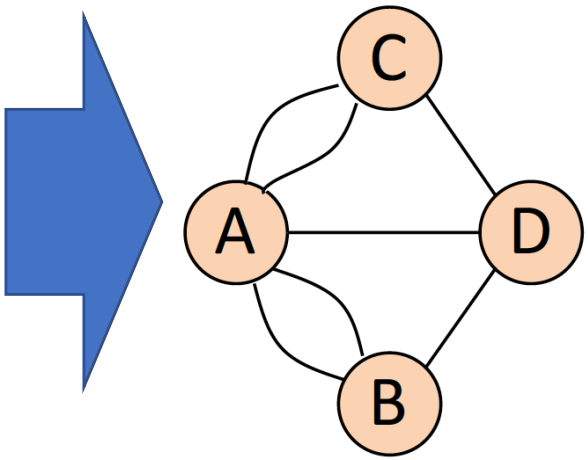
グラフ理論の元祖

『ケーニッヒスベルクの橋』の問題

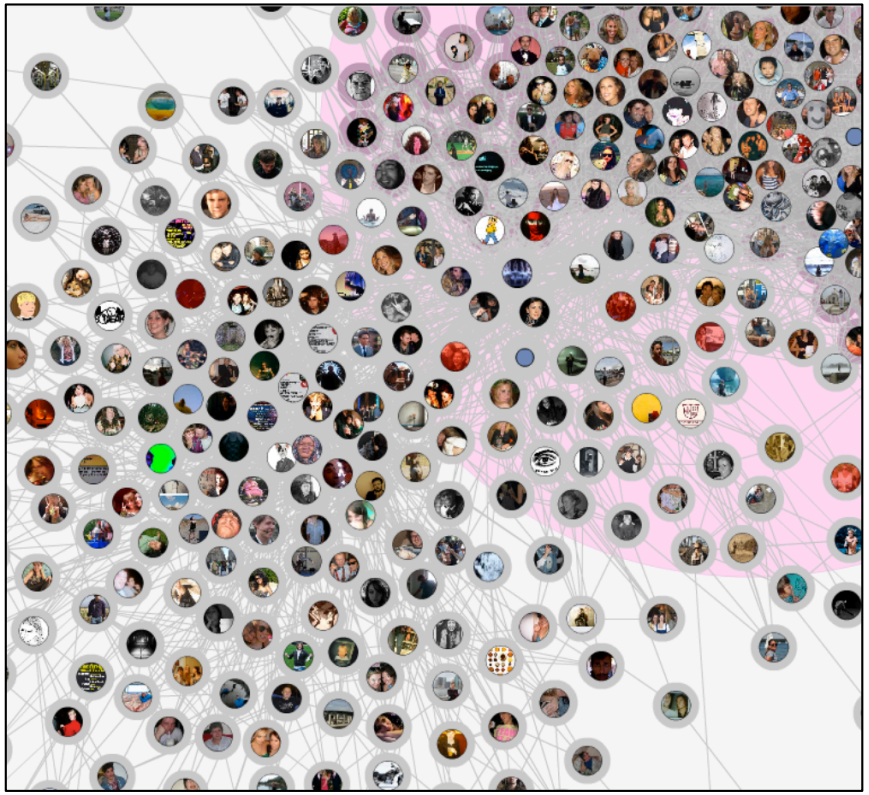
ケーニヒスベルクの７つの橋を
それぞれ１回だけ渡って歩くことはできるか？

レオンハルト・オイラー
（1707～1783）



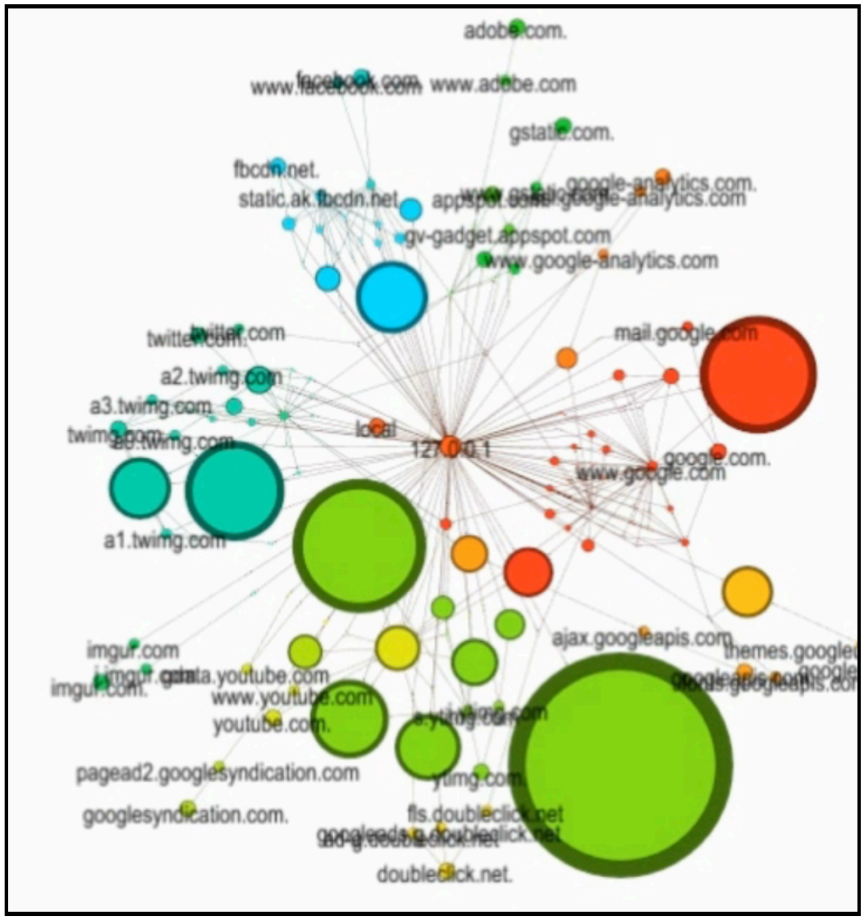


色々なグラフ



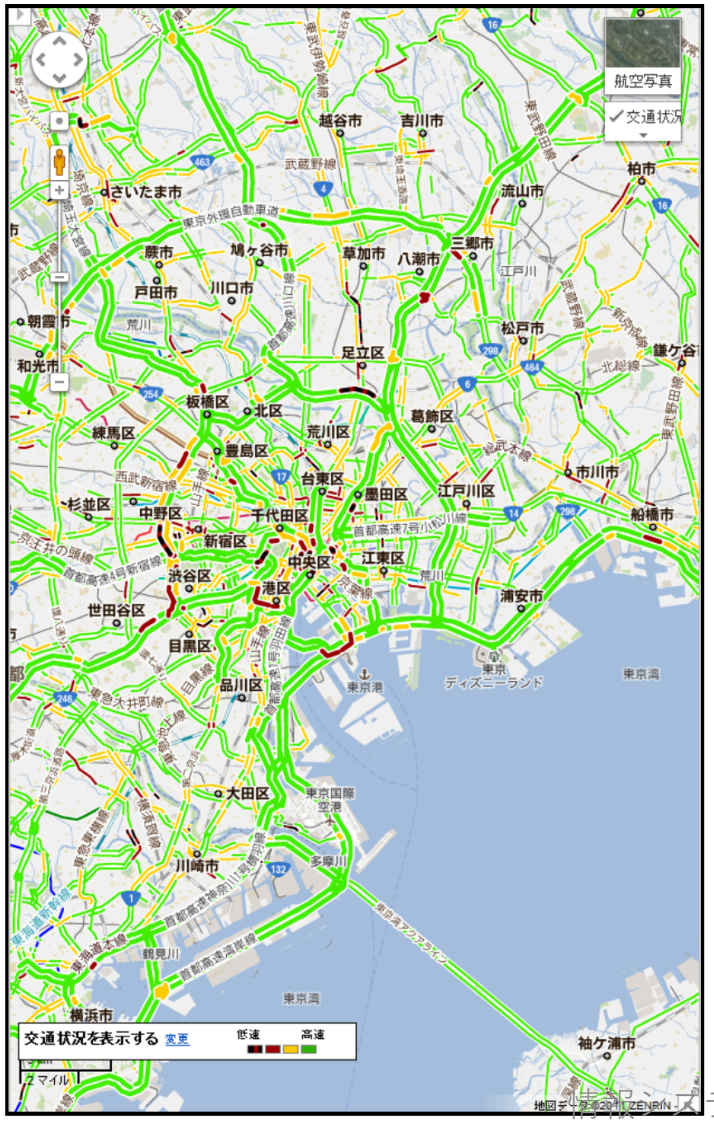
- **マイクロブログサービス**
  - 人間関係をグラフで表現
    - TwitterやInstagramの
      フォロー関係など
  - ノード：人の集まり
  - エッジ：友達関係・フォロー・
    メッセージのやりとり
- **応用例**
  - コミュニティ解析
  - 情報拡散・マーケティング
  - 「知り合いかも？」など

色々なグラフ



- **Webページ**
  - Webページのハイパーリンクをグラフで表現
  - ノード：Webページ
  - エッジ：リンク
- **応用例**
  - 検索
  - 情報推薦，商品推薦
  - ニュース配信
  - 話題発見
  - など

色々なグラフ



- **道路・交通ネットワーク**
  - 駅・交差点のつながりを
    グラフとして表現
  - ノード：駅・交差点
  - エッジ：線路・道路

- **応用例**
  - 交通案内，交通規制
  - 輸送ルートの解析
  - 災害時の避難ルートの誘導
  - など

グラフ生成モデル

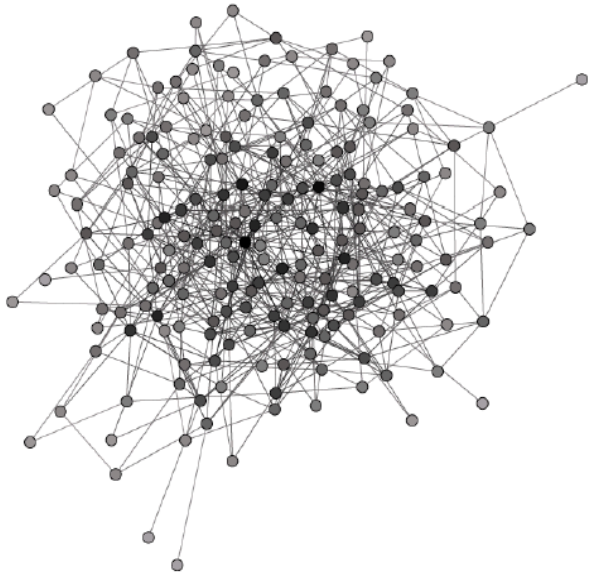
- **グラフの種類は様々**
  - マイクロブログ
  - Webページ
  - 道路・交通ネットワーク
  - 購買履歴
  - …
- **グラフの構造をより理解・解析しやすくするために，実世界のグラフ構造が生成されるモデルがこれまで提案されてきている**

ER (Erdős–Rényi)モデル

- **ランダムなネットワークの生成モデル**
  - 最もシンプルなモデル
  - ノード間に確率pでエッジを張ったグラフ
    - 1959年にポール・エルディッシュとアルフレッド・レーニィによって考案されたグラフモデル
    - ランダムなエッジの接続構造を持つことからランダムグラフの拡張とも言われる

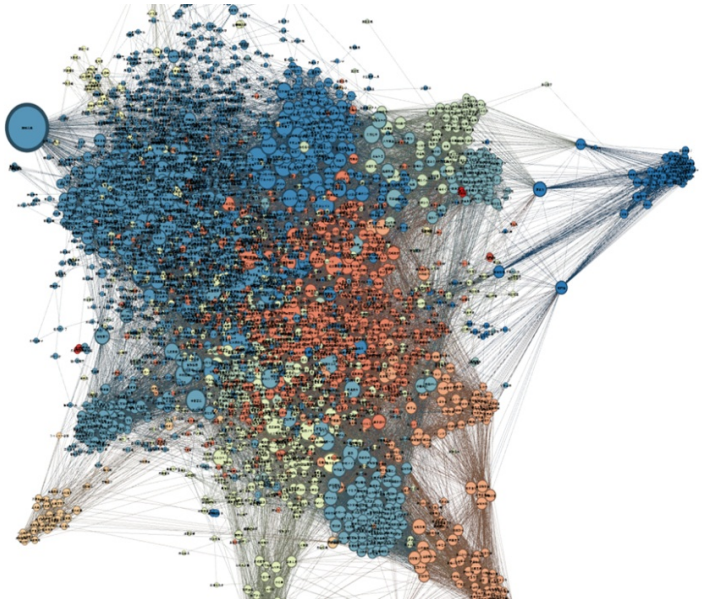
  - ERモデルでノード数n, エッジ数mのグラフが生成される確率は

$$p^m(1-p)^{\binom{n}{2}-m}$$



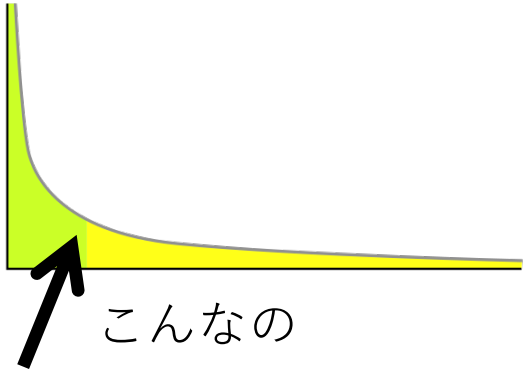
現実のグラフはランダムか？

- **ソーシャルネットワークがもつ３つの性質**
  - 一般的にソーシャルネットワークは，ランダムグラフとは異なり，以下の性質をもつ傾向にあることが知られている
    - **スケールフリー性**
      - 次数の分布が極端に偏る
    - **クラスタ性**
      - 3部クリークが含まれる比率が高い
    - **スモールワールド性**
      - 平均最短経路長が短い



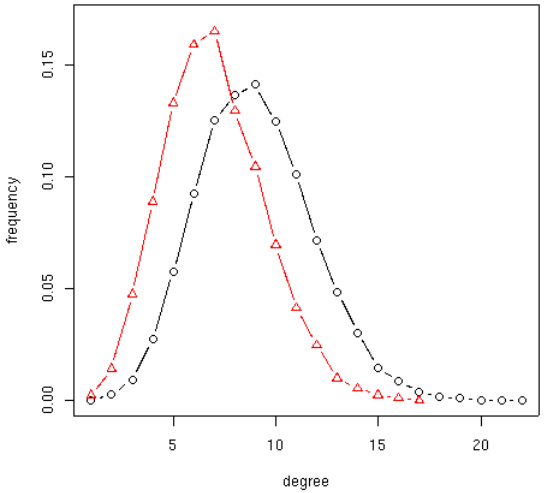
ERモデル（ランダムグラフ）と比較して，現実のグラフでこれらの特性がどのように現れているかを評価することが分析において重要

性質1:スケールフリー性



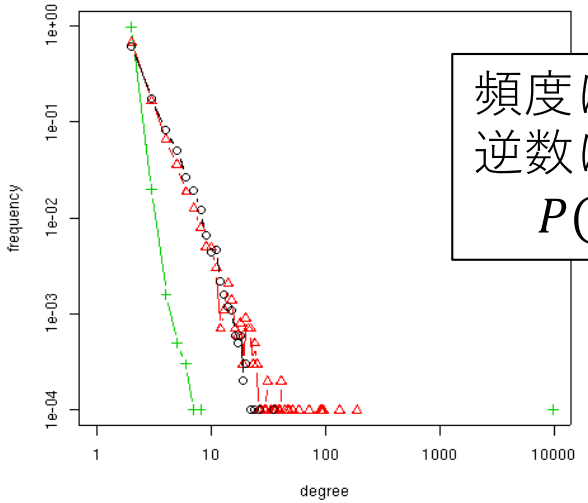
- **ソーシャルネットワークは次数分布が<u>べき乗則</u>に従う**
  - すなわち，「エッジ数の少ないノードが極めて多く存在し，エッジ数の多いノードが稀である」ということ

ランダムグラフ



次数分布は二項分布になる

スケールフリーなグラフ



頻度は次数の逆数に比例

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

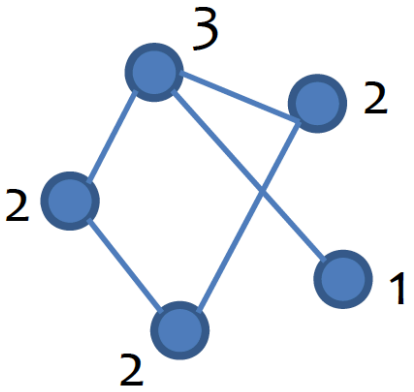
次数分布は**<u>両対数</u>**グラフにおいて線形

次数分布

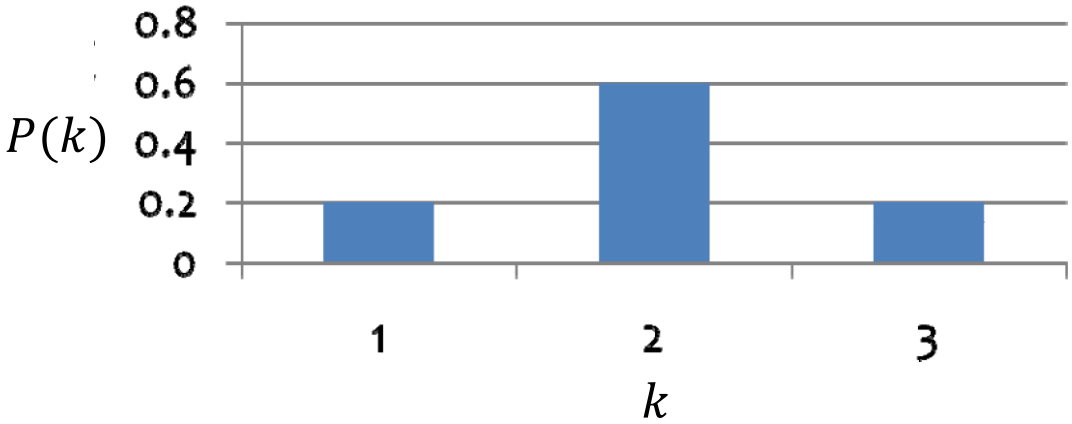
- **特定の次数kを持つノードの出現頻度分布**

$$\text{頻度 } P(k) = \frac{\text{次数が}k\text{のノード数}}{\text{全ノード数}}$$

サンプルグラフ



サンプルグラフの次数分布

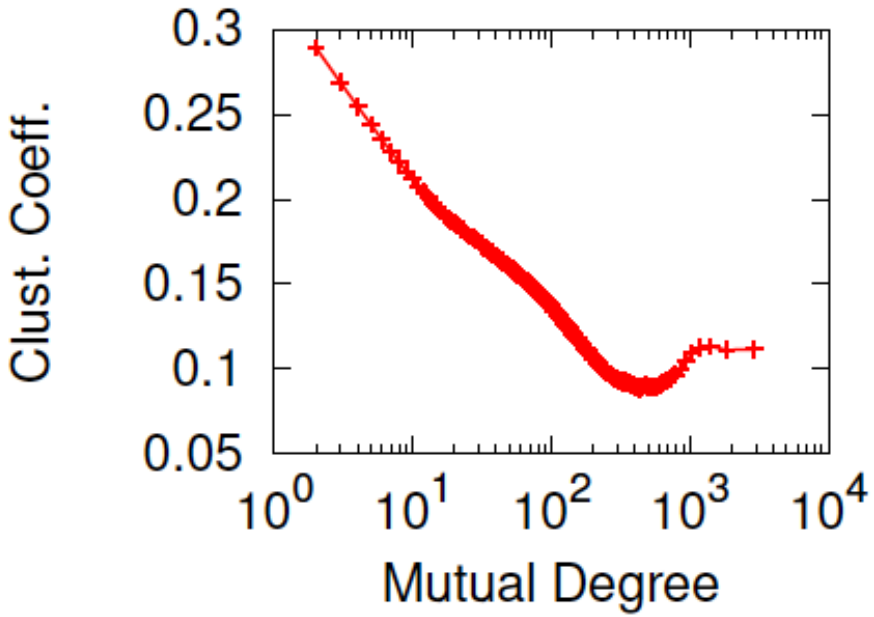


余談：Twitterはソーシャルネットワークか？

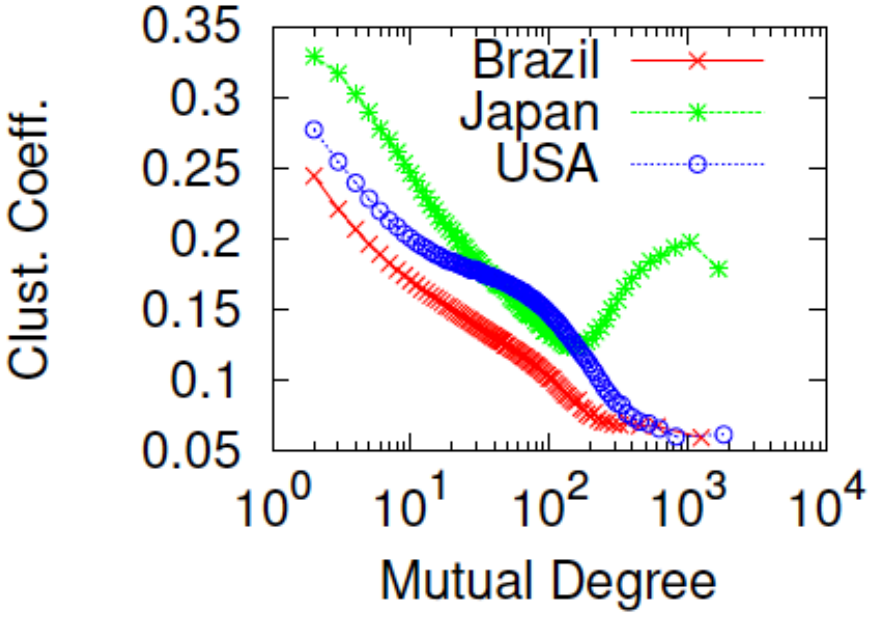
- Twitterでの検証 [Myers et al., WWW2014]
  - 典型的なソーシャルネットワークはノードの次数とクラスタ係数の間に負の相関がなりたつ
  - しかし，Twitter全体ではソーシャルネットワークの性質を満たせないかもしれない
    - 日本人のTwitterユーザが原因
    - 日本人を除くときれいに性質を満たす

S. A. Myers et al., ”Information Network or Social Network? The Structure of the Twitter Follow Graph,” In Proc. WWW 2014.

余談：Twitterはソーシャルネットワークか？



(a) Entire Graph



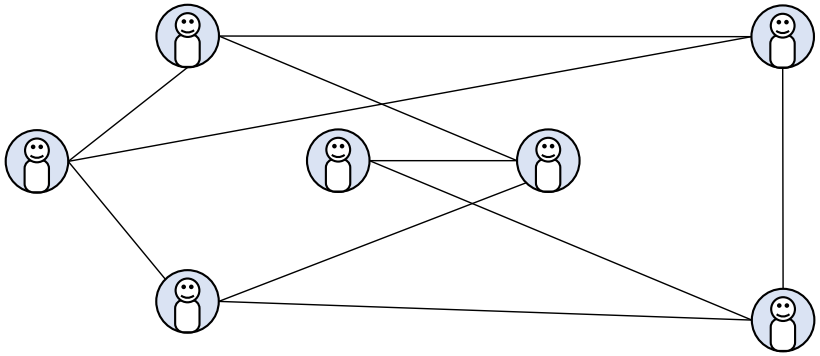
(b) Selected Countries

図は論文から引用

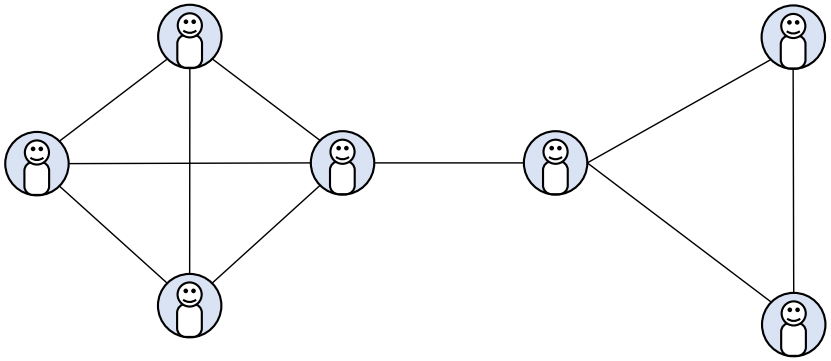
性質2: クラスタ性

- **ソーシャルネットワークは平均クラスタ係数が高い**
  - すなわち，グラフの中に3部クリーク(triangle)が多く含まれる
    = コミュニティ構造が存在する

ランダムグラフ



クラスタ性の高いグラフ

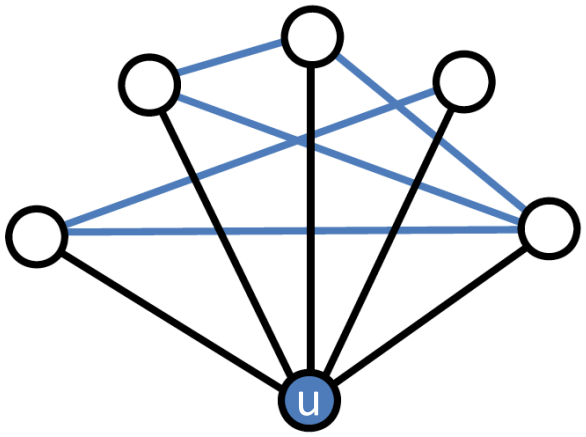


クラスタ係数

- **あるノードuの隣接ノード間でエッジが張られる確率**
  - =ノードuと隣接ノードが作る3部クリーク（triangle）の割合

クラスタ係数 $C_u = \frac{M_u}{\binom{k_u}{2}} = \frac{2M_u}{k_u(k_u-1)}$

- M_uはノードuの隣接ノード間に張られたエッジ数=triangleの数
- k_uはノードuの次数



$$C_u = \frac{5}{\binom{5}{2}} = \frac{2*5}{5*4} = 0.5$$

スモールワールド性

- **ソーシャルネットワークは平均最短経路長が小さい**
  - すなわち，任意の2ノードはわずかな数のノードを経由するだけでたどり着くことができる=情報の伝達速度が早い

Nature **393,** 441 (1998)

Table 1 Empirical examples of small-world networks

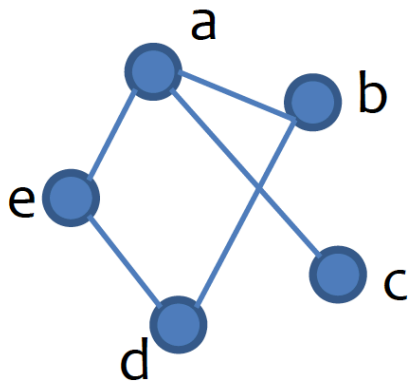
	L_{actual}	L_{random}	C_{actual}	C_{random}
Film actors	3.65	2.99	0.79	0.00027
Power grid	18.7	12.4	0.080	0.005
C. elegans	2.65	2.25	0.28	0.05

性質3:平均最短経路長

- **全ての2ノード間の最短経路長の平均値**

$$L = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} d(i,j)$$

Nはノード数，$d(i,j)$は右の距離行列



距離行列

	a	b	c	d	e
a	-	1	1	2	1
b	1	-	2	1	2
c	1	2	-	3	2
d	2	1	3	-	1
e	1	2	2	1	-

余談：スモールワールド実験

スタンレー・ミルグラム

- **ミルグラムの実験（1967年）**
  - 世界最初のスモールワールドを確認するための実験
  - ミルグラムは最初にカンザス州に住む60人に手紙を送った
  - 彼らには(1)この手紙をマサチューセッツ州に住むある女性Xに転送すること，(2)転送は個人的な知り合いに手渡しで行うように指示された
  - つまり，女性Xに直接，または，女性Xに手渡せそうな友人に転送する必要がある

実験結果

- 無事に女性Xに到達したのは手紙の5%，うち1通は4日で到達
- 最初に手紙を受け取った人物から女性Xまでの平均仲介人数は6人だった → **6次の隔たり**

最近の動向

- 2008年：mixiでの実験
  - スモールワールド性の検証実験
  - 6人たどれば全体の95%に到達できることが明らかになった
- **2016年：Facebookでの実験**
  - 15.9億人分のソーシャルネットワークを分析
  - 全世界のユーザでは4.57人，米国内では3.5人で到達可能

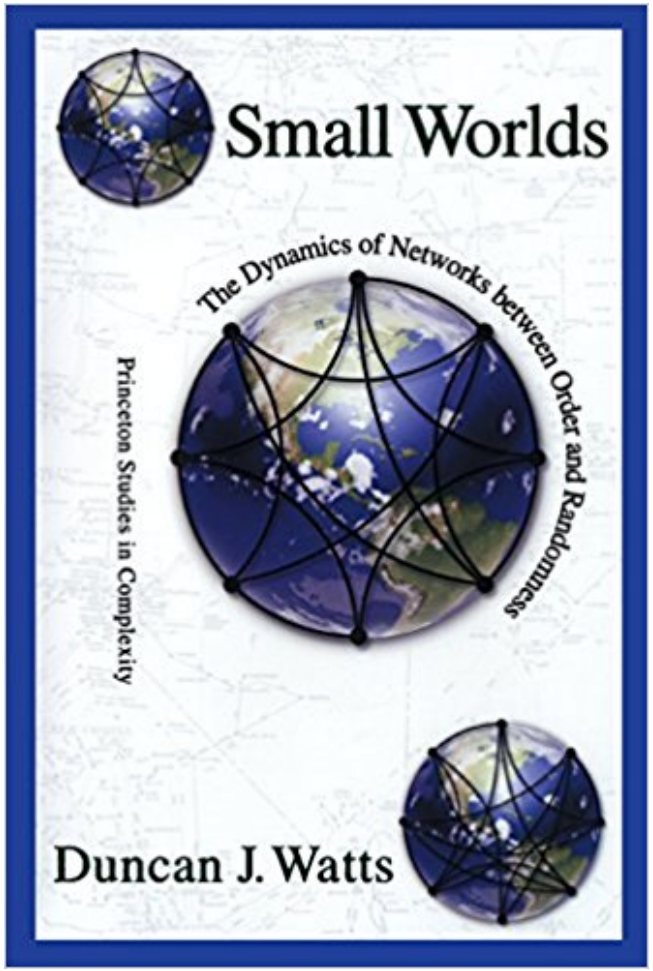
まとめ

- **ソーシャルネットワークが持つ性質**
  - スケールフリー性
    - 次数の分布が極端に偏る
  - クラスタ性
    - 3部クリークが含まれる比率が高い
  - スモールワールド性
    - 平均最短経路長が短い

参考図書

- Duncan J. Watts, “**Small Worlds: The Dynamics of Networks Between Order and Randomness**,” Princeton Studies in Complexity (**ISBN-13:** 978-0691117041)

- 日本語版もあります
  - **ISBN-13:** 978-4501540708



参考図書

- 増田直紀、今野紀雄, **“複雑ネットワーク 基礎から応用まで”**近代科学社 **(ISBN-13: 978-4-7649-0363-0)**