# ネットワークのコミュニティ分析とブートストラップ法

東京工業大 情報理工学研究科
永田 晴久　　下平 英寿

# ネットワークのコミュニティ抽出

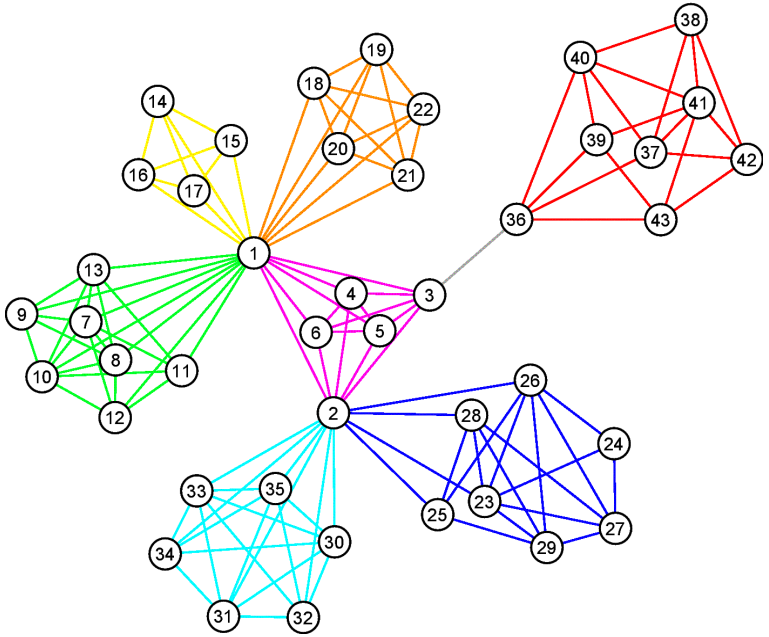
関わりの強いノードのまとまりを見つけたい！

- ノード自体の情報を使わず、リンク構造のみから推定

- 階層型クラスタリング
  - コミュニティはデンドログラムのサブツリーで表される
  - ノード間またはリンク間で距離を定義
  - Newmanの方法 (2004)
  - Ahnの方法 (2010)

# Ahnの方法

- 2本のリンクに対して、「リンクの近さ」Sを定義
  - リンクに対する階層型クラスタリング
  - コミュニティの重なりを表現できる




$$S(e_{ij}, e_{ik}) = \frac{|n_+(j) \cap n_+(k)|}{|n_+(j) \cup n_+(k)|}$$

$$S(e_{ij}, e_{ik}) = \frac{\boldsymbol{a}_i \cdot \boldsymbol{a}_j}{|\boldsymbol{a}_i|^2 + |\boldsymbol{a}_j|^2 - \boldsymbol{a}_i \cdot \boldsymbol{a}_j}$$

$n+(i)$　：ノード$i$に隣接するノードの集合（$i$を含む）

$ai$　：隣接行列$A$の第$i$列ベクトル

共通のノードを持たないリンクに対しては$S=0$

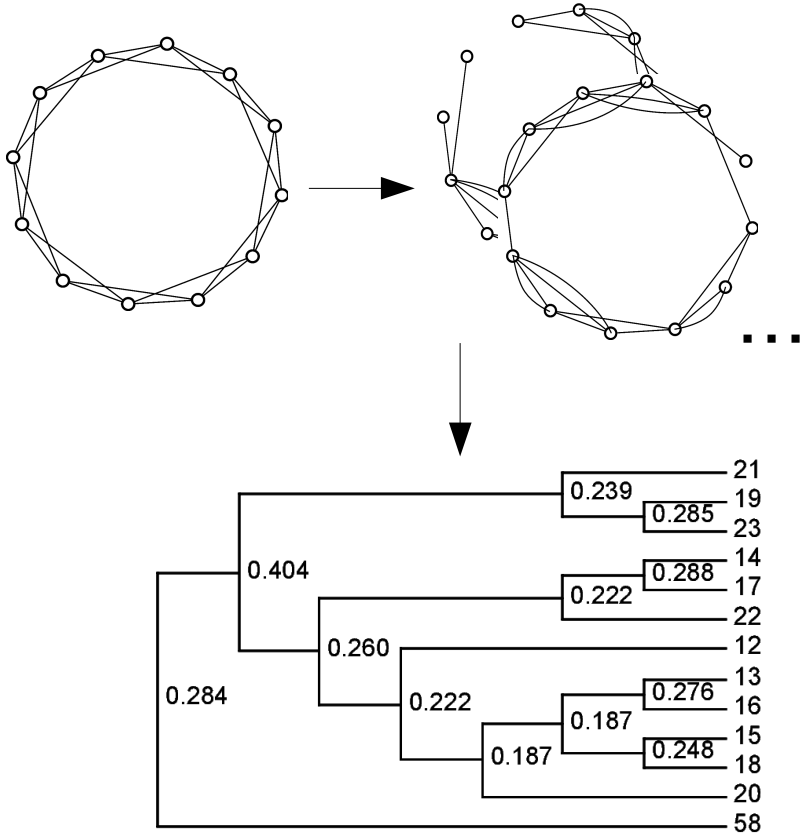
# 問題点とアイディア

- 階層型クラスタリングは、意味のないコミュニティも数多く拾ってしまう
  - 検出されるコミュニティ数は(要素数−1)
  - 現状では、木を適当な高さで切って解とする
    - 階層型クラスタリングの利点を失ってしまう

→ ブートストラップ法を利用して、意味のあるコミュニティを選択する

# ブートストラップ法の利用

1. グラフ構造をリサンプリング
   - 元データの構造によく似た、別の形のグラフを生成
2. リサンプリングデータからデンドログラムを生成
3. 1, 2を多数繰り返す
4. クラスタの出現頻度を計算




→ 頻度の高いクラスタほど、構造の変化に強い

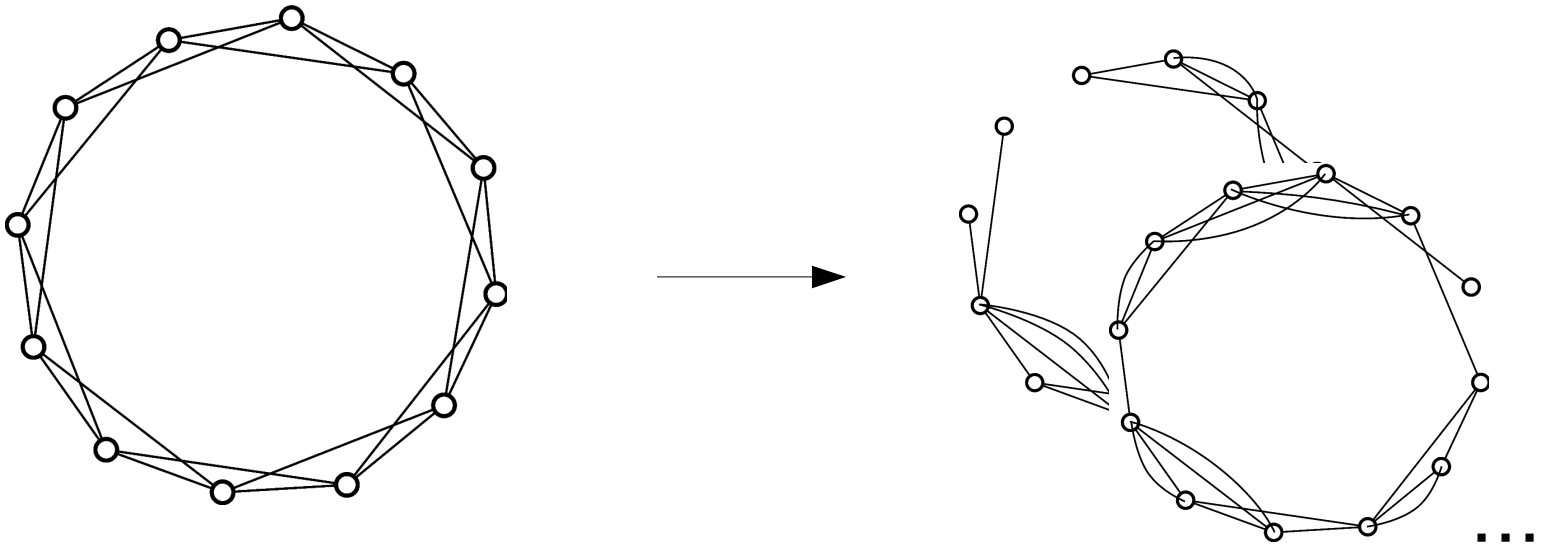
# リサンプリング方法

- リンク集合をデータとしてリサンプリング
  - リンクの重複を許す … multiple edge

$$E = \{e_1, \ldots, e_M\}$$

$$E^* = (e_1^*, \ldots, e_M^*)$$

$$e_i^* \in E,\ |E^*| = M$$



# ブートストラップ確率

- ひとつのクラスタに注目し、複製されたグラフから作ったデンドログラムでの出現回数をカウントする

$$\mathrm{bp}_T = \frac{\sum_{i=1}^{B} h(T, D_i^*)}{B}$$

$$h(T, D^*) = \begin{cases} 1 & (^{\exists}T^* \subset D^* \text{ s.t. } l(T) = l(T^*)) \\ 0 & (\text{otherwise}) \end{cases}$$

$B$ : リサンプリング回数

$T \subset D$ : 木$T$は木$D$のサブツリー

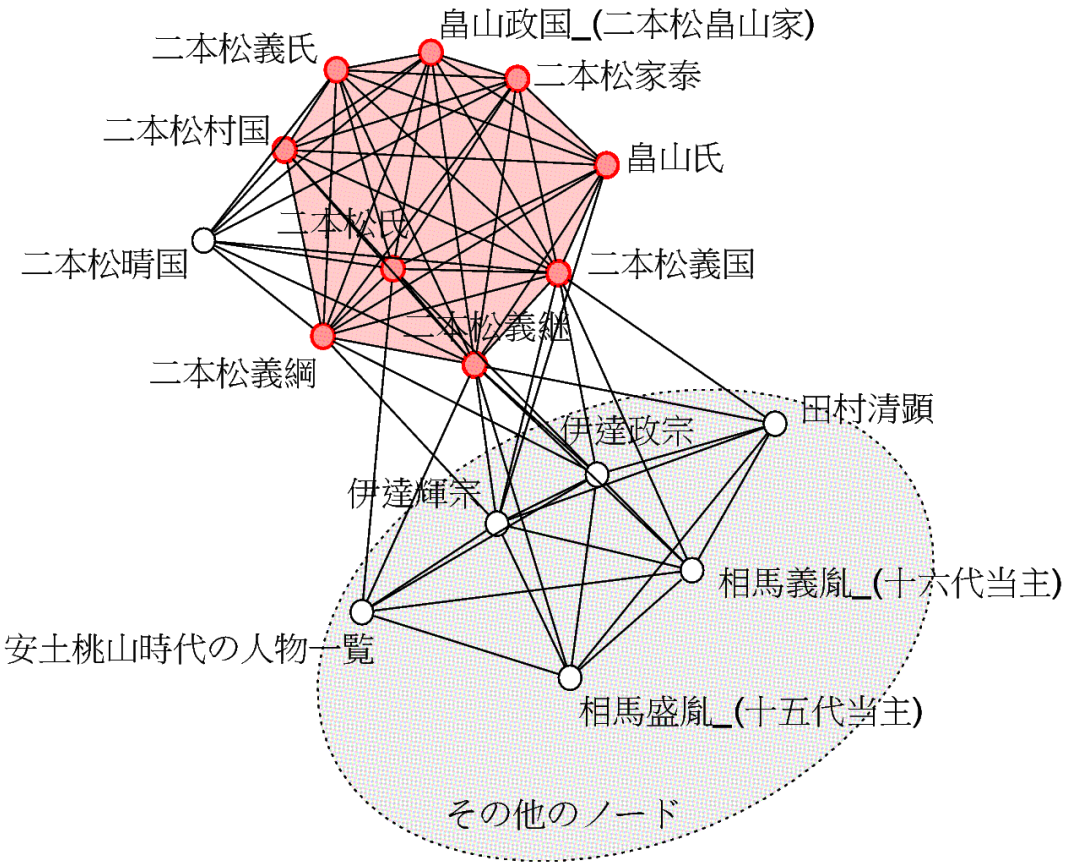
$l(T)$ : 木$T$の葉ノード集合

# 実データへの適用

- Wikipedia「戦国大名」カテゴリの記事とリンク集合
  - |V|=626, |E|=5,341
  - 基本的な複雑ネットワークの性質を満たす
    - スケールフリー性
    - スモールワールド性

# 実験結果

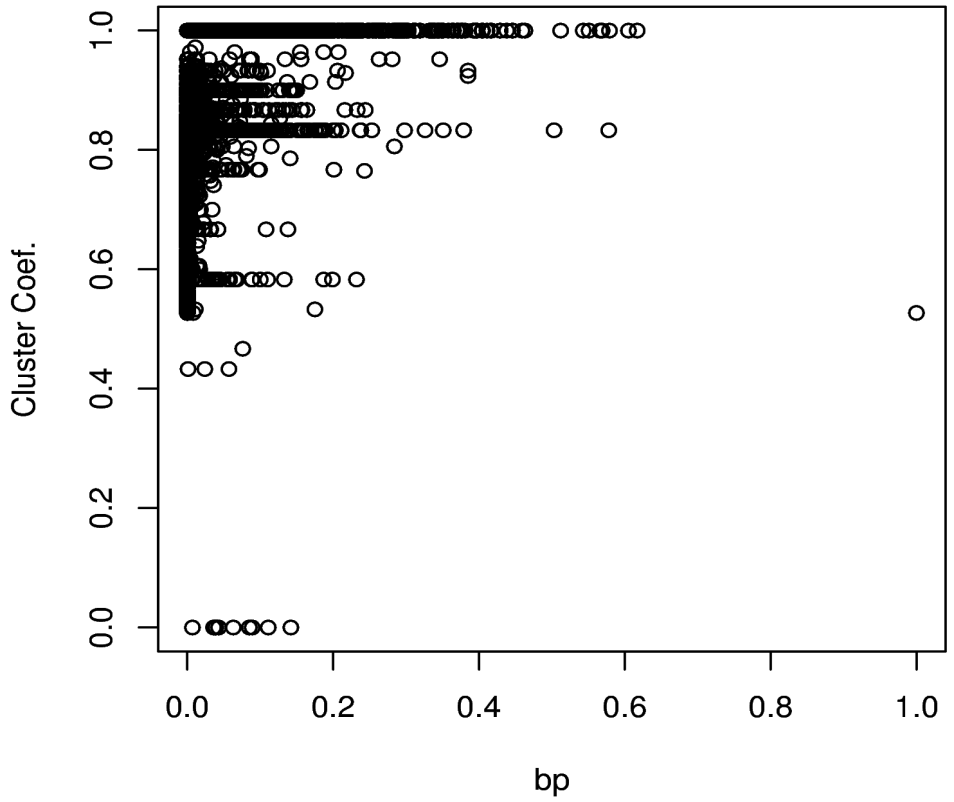
- コミュニティ周辺のサブネットワーク




「二本松氏」のコミュニティ bp=0.431

# 実験結果

- 各クラスタにおけるクラスタ係数との比較
  - bpが大きいクラスタは、「クラスタらしさ」を持っている



# 次のステップ

- 元のグラフ構造に近いデータを生成するには？
  - 今回は簡単のために、リンク集合をリサンプリングした
    - 解釈は？
  - リンクをランダムに張り替えてデータを複製
    - 張り替え確率を小さく設定して、元の構造を残す
    - 張り替え本数はランダムにしない（割合などから計算）
  - ネットワークの生成モデルを入れる
    - BA/WSモデルなど、複雑ネットワークで使われるモデル
    - 構造の変化のモデル化

# 参考文献


- M. Girvan and M. E. J. Newman,

  “Community structure in social and biological networks”,

  *Proc. Natl. Acad. Sci. USA* 99, 7821-7826 (2002)
- M. E. J. Newman,

  ″Modularity and community structure in networks″,

  *Proc. Natl. Acad. Sci. USA* 103 (23): 8577-8582 (2006)
- Y. Y. Ahn,

  “Link communities reveal multiscale complexity in networks”,

  *Nature*, 466, 761-764 (2010)