【WWW2012勉強会】

Session 12: Community Detection in Social Networks

担当： 馬場 雪乃　（東京大学）

Community Detection in Social Networks

- **ソーシャルネットワーク上のコミュニティや重要なユーザをグラフ構造（+α）を使って見つけましょうというセッション**
  - ソーシャルネットワークの分析・利用に役立つ

- 12-1 Using Content and Interactions for Discovering Communities in Social Networks
  - ユーザのinterestも考慮したコミュニティ発見

- 12-2 Community Detection in Incomplete Information Networks
  - 部分的なエッジ情報しかないときのコミュニティ発見

- 12-3 QUBE: a Quick algorithm for Updating BEtweenness centrality
  - ノードの重要性の指標”Betweenness Centrality”を、グラフが頻繁に更新される場合でも効率的に計算

ユーザのinterestも考慮したコミュニティ発見

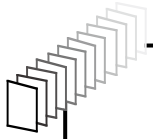
- ユーザ間のインタラクショングラフを利用
  - グラフ構造だけでは解決できない
    - 「今までやり取りがないけどinterestは似ている」こともある
  - メッセージの中身を見ないといけない

- 嬉しい例：
  広告を提示するのに効果的なコミュニティを発見

- ポイント：以下全てを考慮したコミュニティ発見
  - **Content**（メッセージの潜在トピック）
  - **Link**（グラフ構造．誰から誰にメッセージが送られるか）
  - **Interaction type**（例．Broadcast tweet, reply, RT）

メッセージの生成モデルを提案

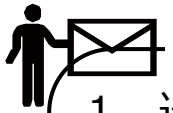
- **Topic User Recipient Community Model** を提案



トピックごと
単語の分布

コミュニティごと
Interaction typeの分布
受信者の分布

送信者ごと
コミュニティの分布
送信者が各コミュニティにメッセージを送るときのトピック分布



送信者が各メッセージ（本文，interaction type，受信者）を生成

1. 送信者のコミュニティ分布に従って、メッセージを送るコミュニティを決定
2. コミュニティの受信者分布に従って、**受信者**を決定(複数人)
3. コミュニティのInteraction type分布に従って、**type**を決定
4. 送信者がコミュニティにメッセージを送るときのトピック分布に従って、**トピック**を決定
5. トピックの単語分布にもとづいて、**メッセージ本文**を決定

- 各分布のパラメータを推定すればコミュニティを発見できる

既存モデルより質の高いコミュニティ発見を達成

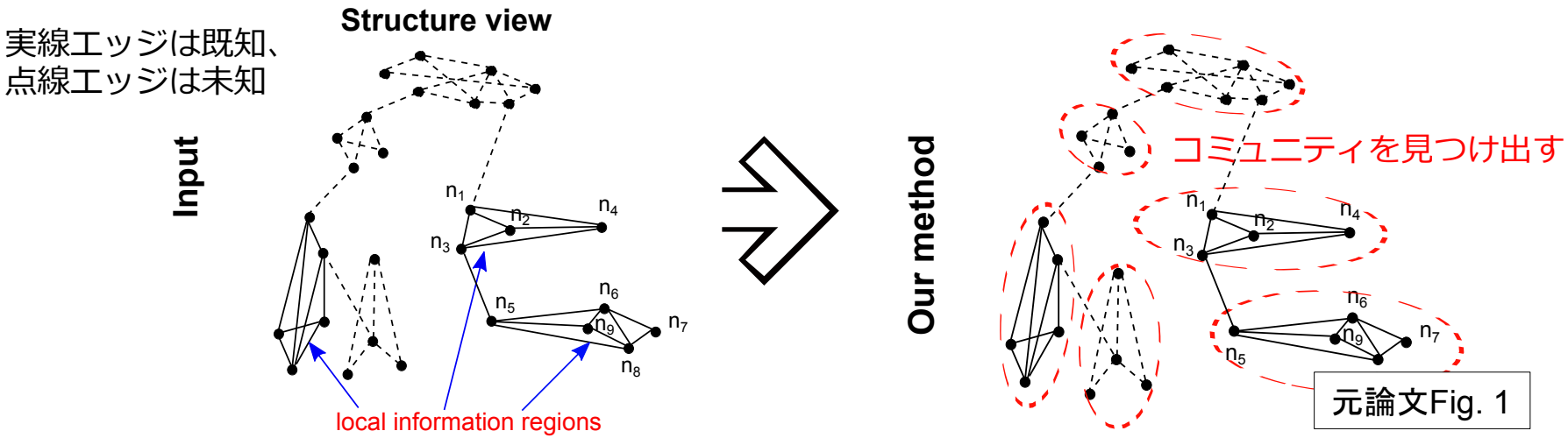
- Twitter, Enron（メール）データセットを用いて実験
- 提案モデルで発見したコミュニティの質を既存モデルと比較
  - ２つの指標で既存モデルと比較し提案モデルの優位性を確認
    - CUT (Community-User-Topic): グラフ情報・Interaction Typeなし
    - CART (Community-Author-Recipient):Interaction Typeなし
  - 評価指標１：Fuzzy Modularity
    - ネットワークの分割の質を測る指標Modularity（「コミュニティ内のエッジ数」ー「エッジをランダムに張った場合の期待値」）を、分割が確率的な場合にも対応できるよう拡張
  - 評価指標２：Perplexity
    - 言語モデルの評価指標としてよく利用される
    - テストデータに対してモデルがどれだけ当てはまっているか評価

部分的なエッジ情報しかないときのコミュニティ発見

NEW!

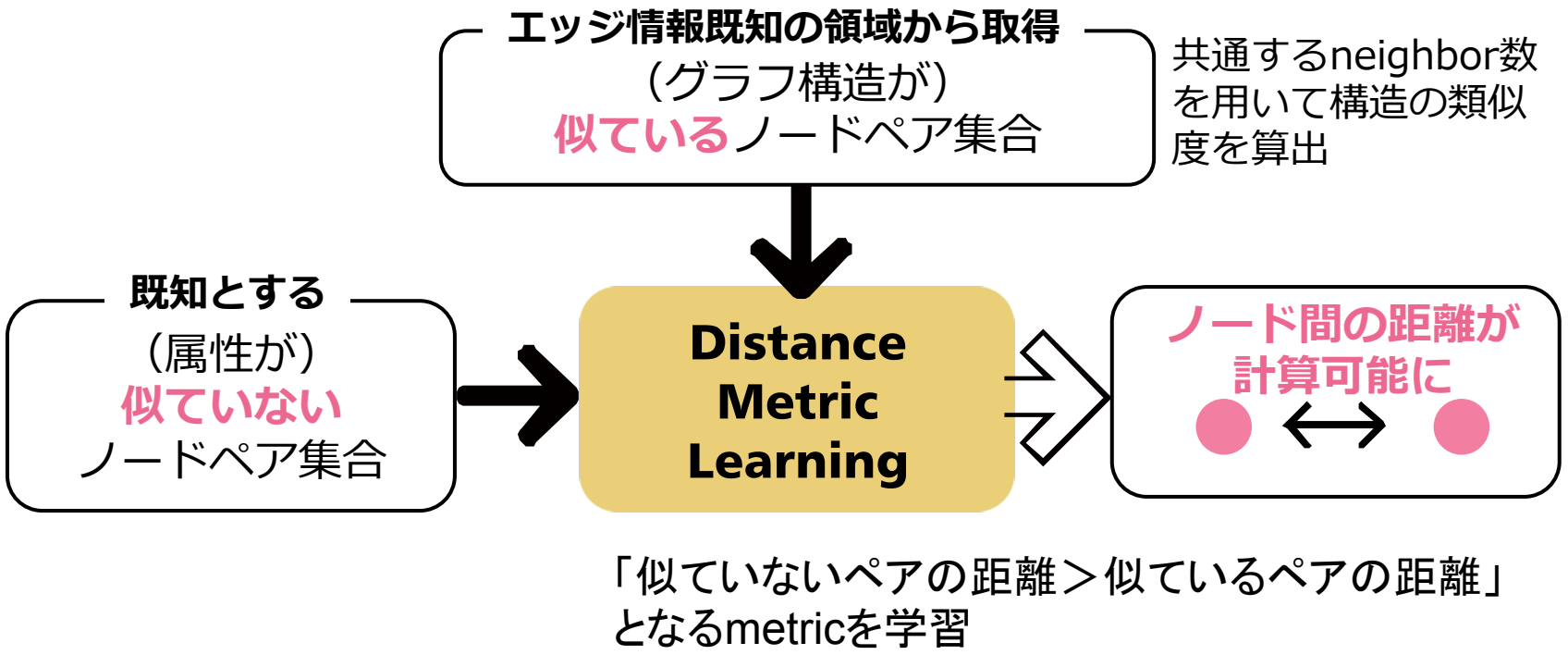
- エッジ情報が不足しているグラフが対象
  - 既存研究：エッジ情報が全て既知という前提
    - エッジの有無が全てわかっている小さい領域がいくつかある

- 例：Terrorist-attack network
  - ノード：テロ攻撃。同じ組織による攻撃のときエッジを張る
  - 捜査が進まないと同じ組織の攻撃かどうかわからない



エッジ情報既知の部分からDistance Metricを学習

- エッジ情報既知の領域内部の構造を利用して
  ノード間のDistance Metricを学習

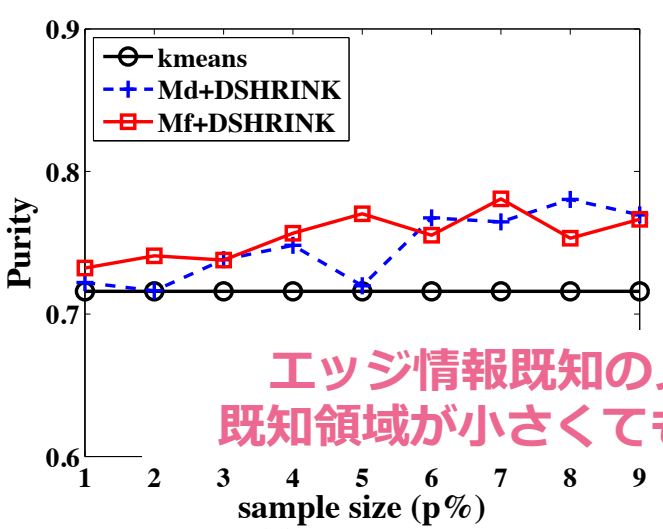


距離を利用してクラスタリング

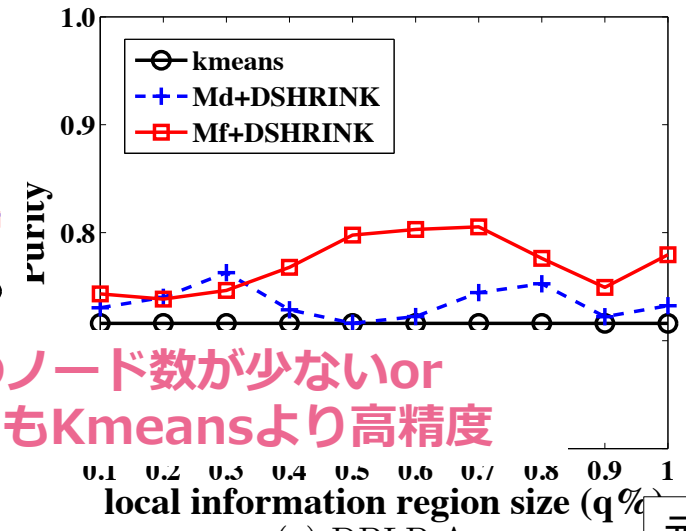
エッジ情報が少なくても高精度でコミュニティを発見可能

- DBLPのデータを利用
  - グラフ：共著関係，ノードの属性情報：論文中の単語
  - いくつかの領域を選択、残りのエッジ情報は捨てる

- 評価指標：提案手法で発見した**コミュニティの精度**（＝コミュニティ内の同じ研究分野の人の割合）
  - エッジ情報を残すノードの数，領域の大きさを変化させて評価



(a) DBLP-A



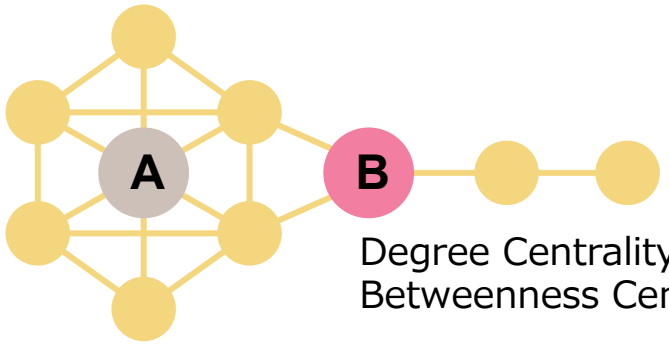
(a) DBLP-A

エッジ情報既知のノード数が少ないor
既知領域が小さくてもKmeansより高精度

元論文Fig. 4, 5

グラフ更新時のBetweenness Centrality計算

- **Betweenness Centrality (BC)**:
  ノードの重要性の指標
  - 全ノード間の最短経路が対象ノードを経由する回数に基づいて算出



Degree Centralityが高いのはA,
Betweenness Centralityが高いのはB

（http://www.orgnet.com/sna.htmlから例を引用）

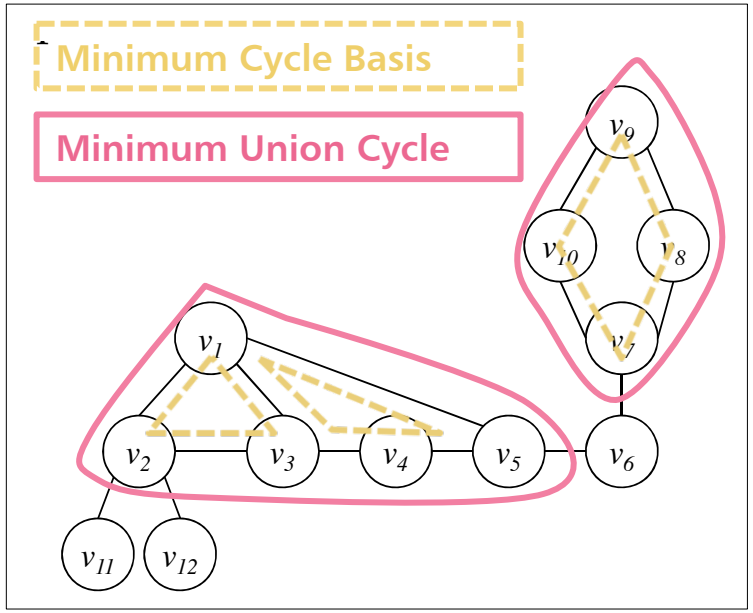
NEW! **グラフの更新**がある場合を想定

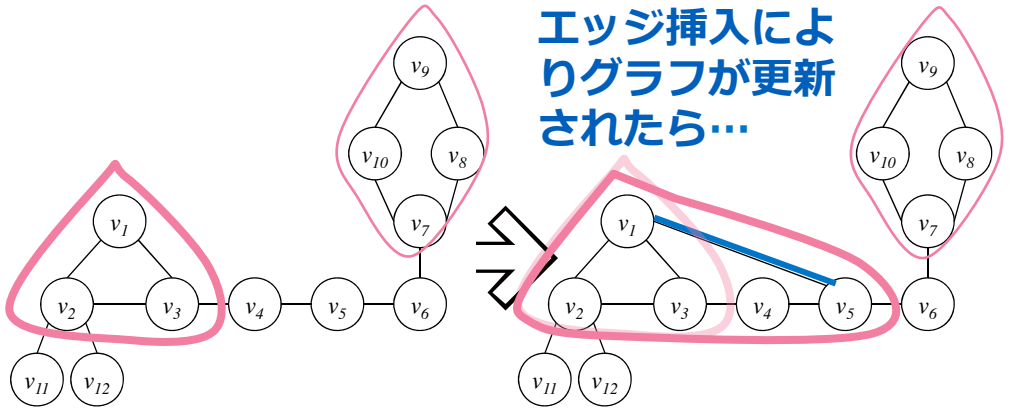
- ソーシャルネットワークのグラフは更新が多い
- BCの算出にはノード間の最短経路が必要
  →更新の度に全ノード間の経路を再計算するのは大変

BCの更新可能性があるノードを見つけて再計算

- グラフが更新されたときに全ノードから**BCの更新可能性があるノード**を見つける
  - 更新可能性があるノードだけBCを再計算→計算量削減

- グラフ更新によって変更があったMinimum Union Cycle (MUC)内のノードは「更新可能性がある」





変更があったMUC内のノード間の最短経路を再計算してBCを更新

元論文Fig. 1に色線を追加

最速アルゴリズムの２～2000倍の速度でBC更新を実現

- 人工データと8種類の実データで実験
  （最大：ノード数11604, エッジ数65441）

- **グラフ更新時のBCの再計算時間**を評価
  - 厳密解を得る既存アルゴリズムのうち最速のものと比較
  - （グラフが更新される度に全て再計算）

- 8種類の実データ中
  - **2～7倍**の高速化が4件
  - 13～40倍の高速化が3件
  - **2400倍**の高速化が1件
    （更新可能性があるノード数に依存）