Keyword-based Correlated Network Computation over Large Social Media

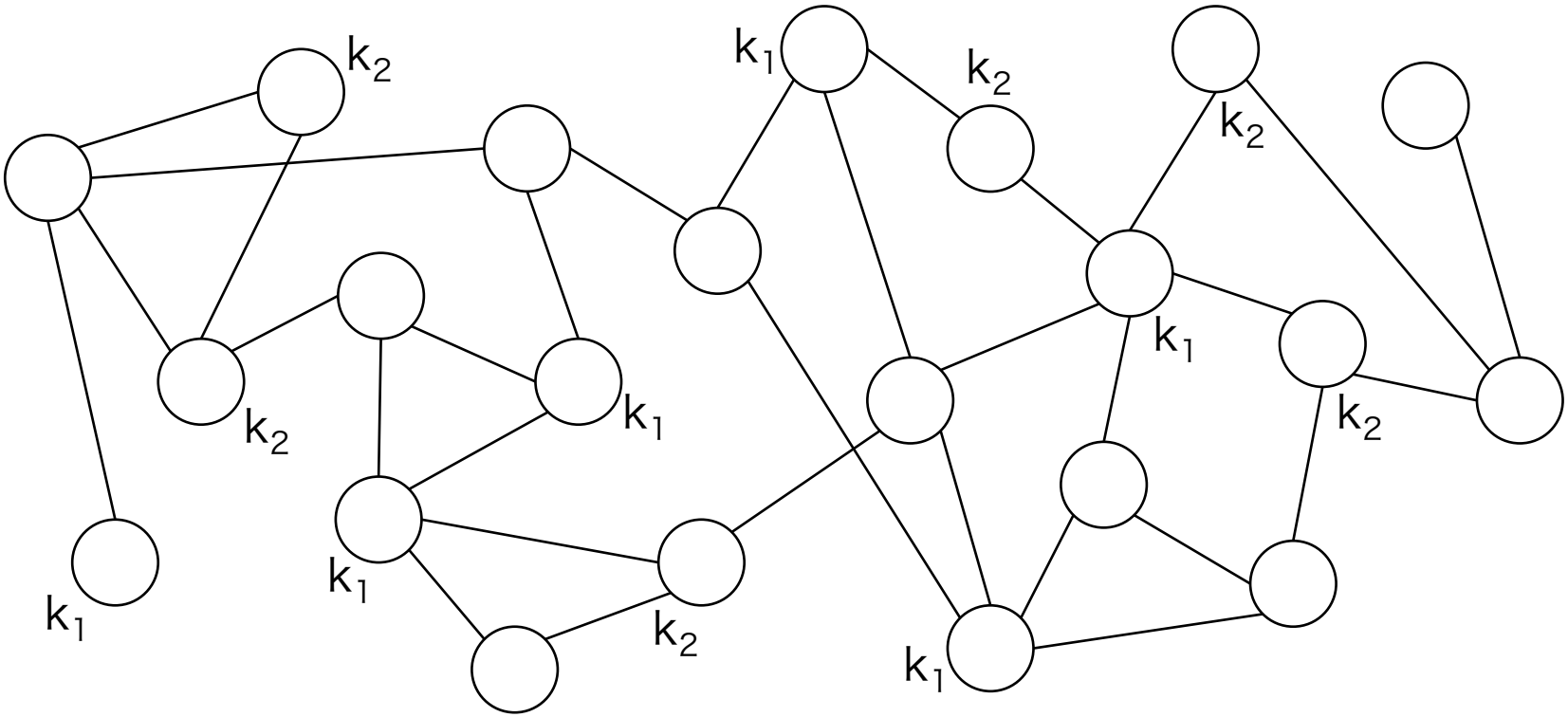
岡田莉奈(筑波大)

背景と研究目的

- 背景
  - ソーシャルメディアデータが増えており，
    これらを大きなグラフとみなすことができる.
    - ノード：実体，エッジ：実体間の関係
  - 研究の動機
    - 何かイベントが起こったときに，キーワード検索をかける.
      欲しい情報は，そのキーワードを持つノードの情報
      だけでなく，そのノードの周辺情報も欲しい.
- 研究目的
  - 大きなグラフからキーワードに基づいた
    相互関係のあるネットワークを見つけ出すこと.
    - サブグラフ(コンポーネント+ネイバー)の抽出.

相互関係のあるネットワークとは？

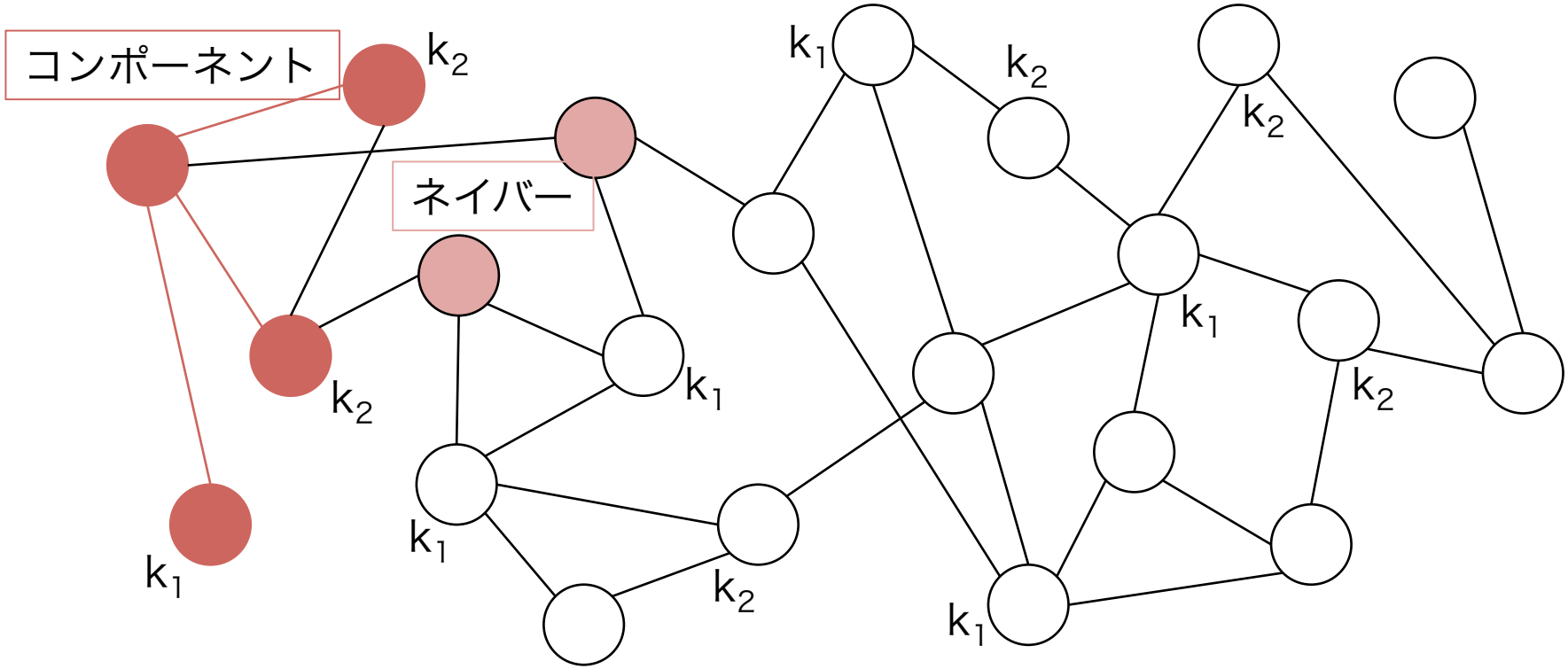
(例) 検索キーワード：{k_1, k_2}
コンポーネント内のキーワード間の最大距離 γ=2



ソーシャルメディアデータ

相互関係のあるネットワークとは？

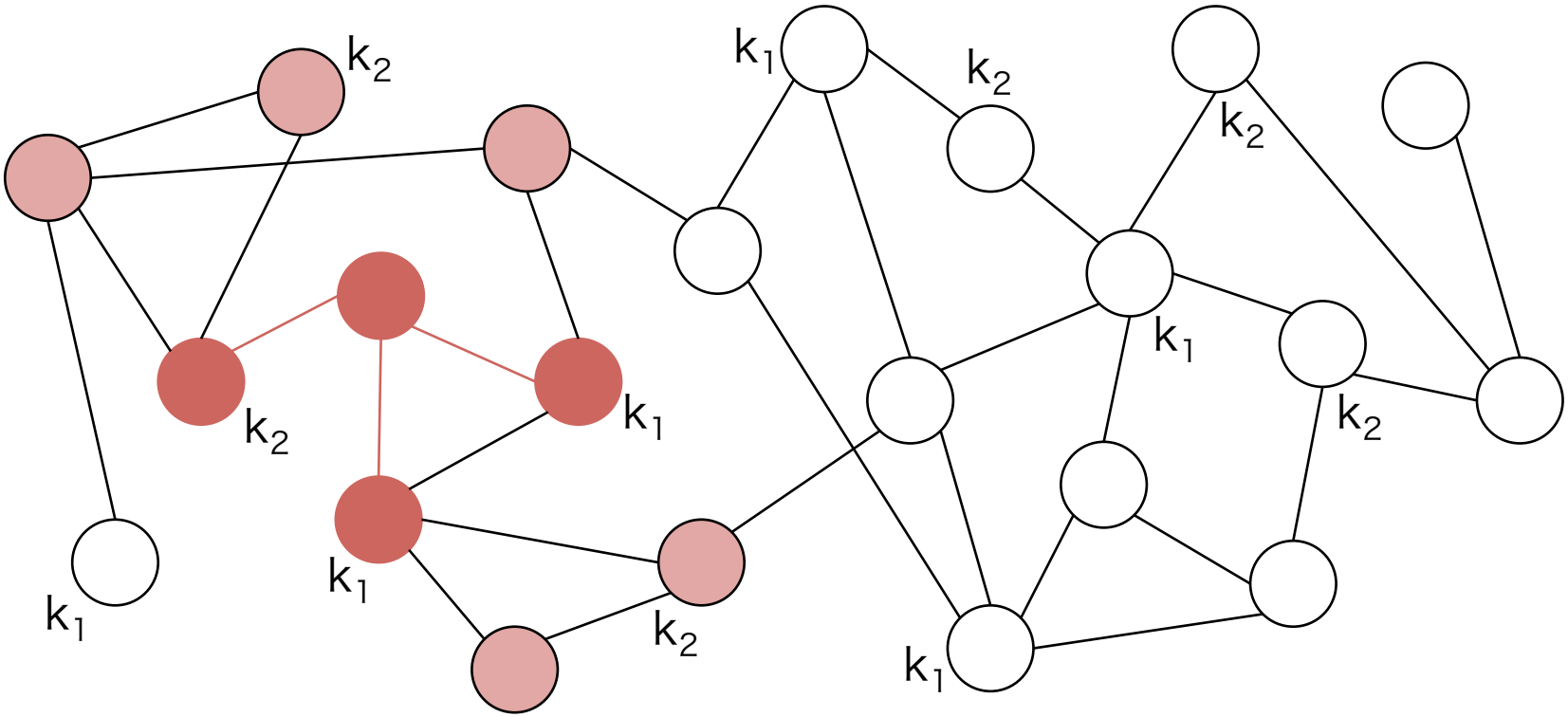
(例) 検索キーワード：{k_1, k_2}
コンポーネント内のキーワード間の最大距離 γ=2



ソーシャルメディアデータ

相互関係のあるネットワークとは？

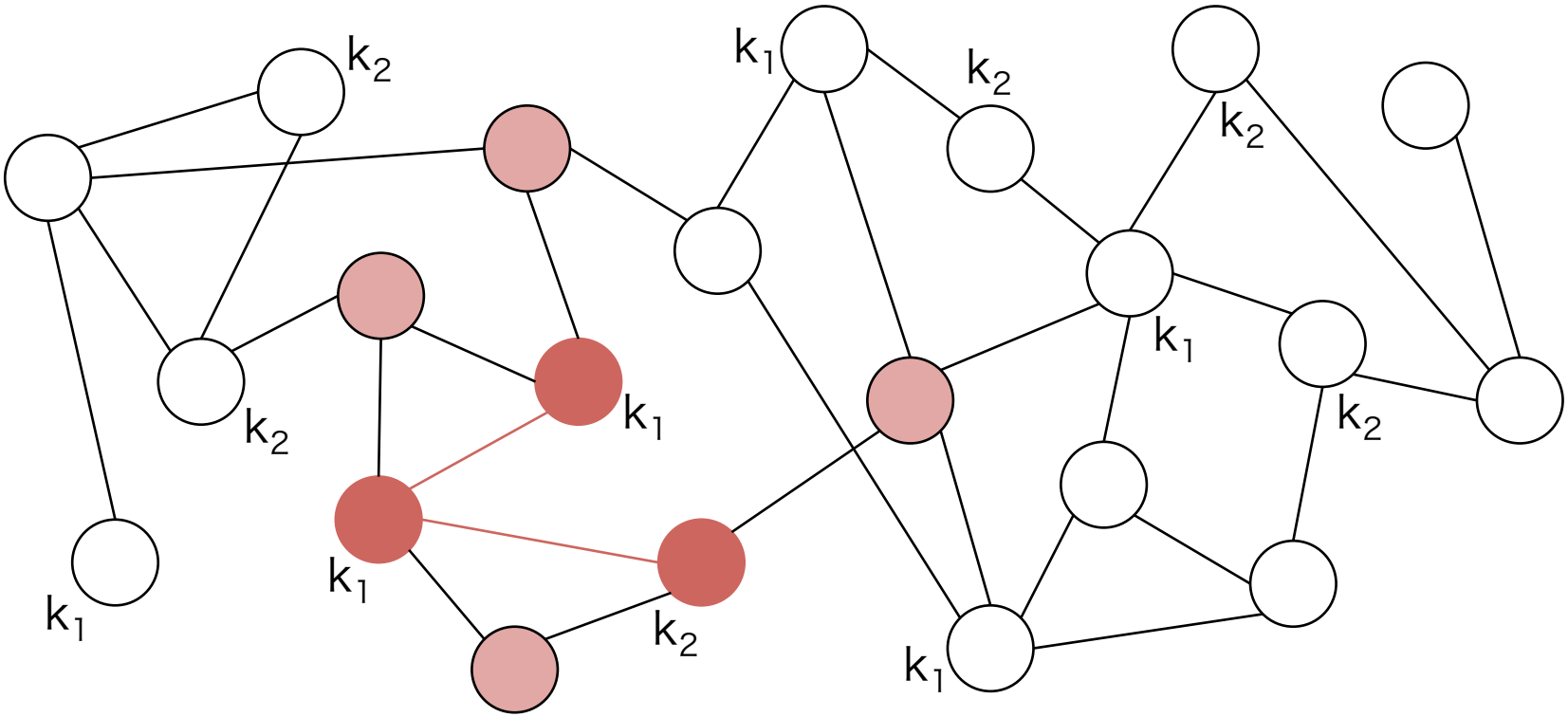
(例) 検索キーワード：{k_1, k_2}
コンポーネント内のキーワード間の最大距離 γ=2



ソーシャルメディアデータ

相互関係のあるネットワークとは？

(例) 検索キーワード：{k_1, k_2}
コンポーネント内のキーワード間の最大距離 γ=2



ソーシャルメディアデータ

相互関係のあるネットワークとは？

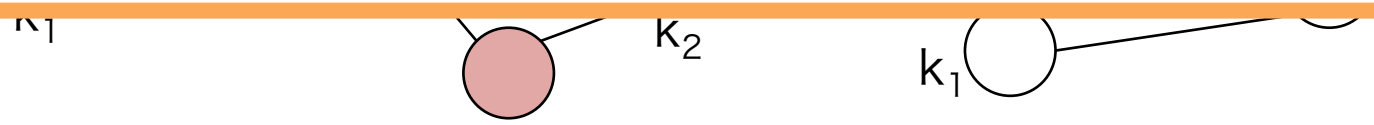
(例) 検索キーワード：{k_1, k_2}

$$\frac{\sum\{weight(v,G_1)*weight(v,G_2)|v\in G_1'\cap G_2'\}}{|G_1'\cup G_2'|}$$

ただし，

$$\begin{cases} weight(v',G_1)=2^{-minDist(v',G_1)} & \text{v'がネイバーのとき} \\ 1 & \text{otherwise} \end{cases}$$

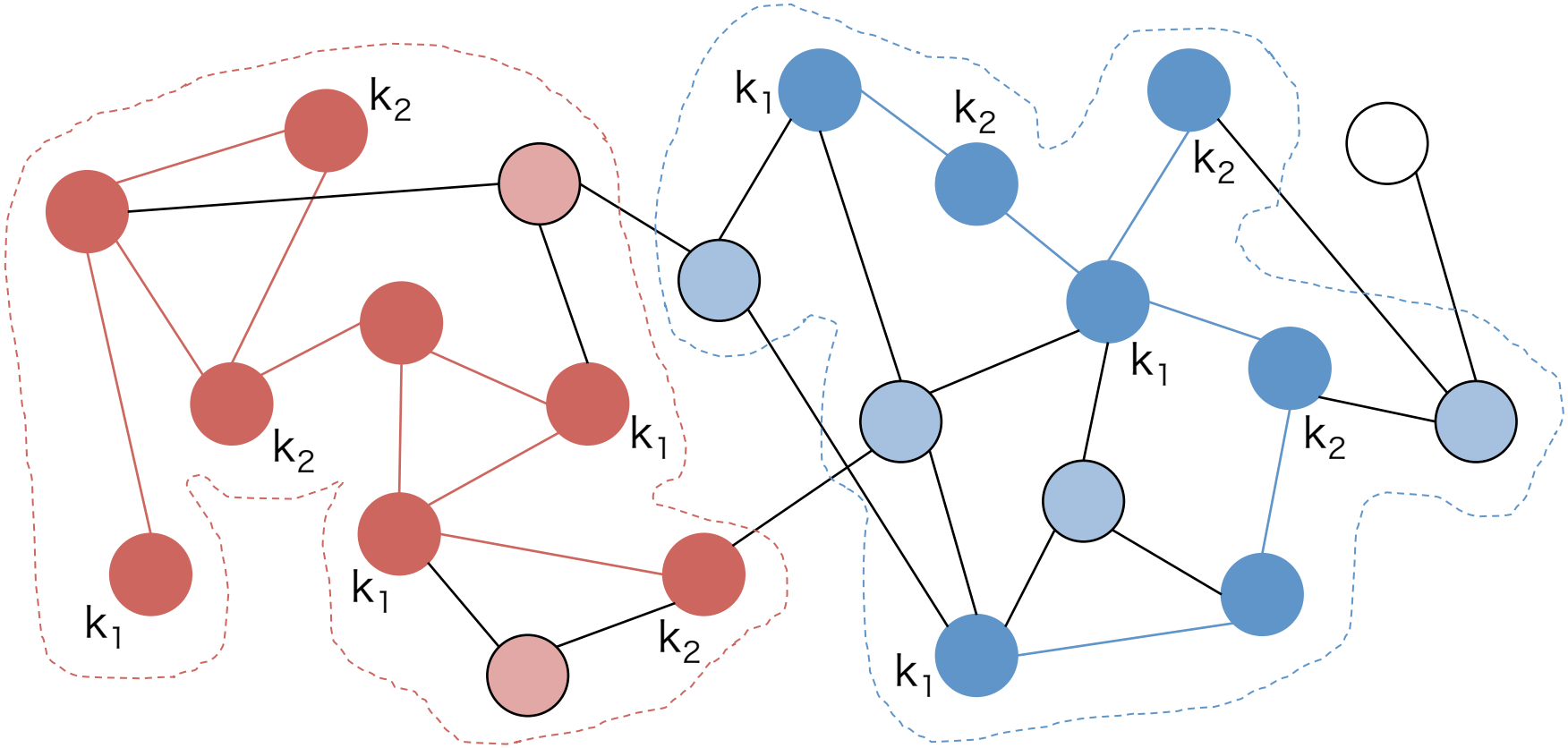
という指標を用いて，
密に関連しているコンポーネントを探す．



ソーシャルメディアデータ

相互関係のあるネットワークとは？

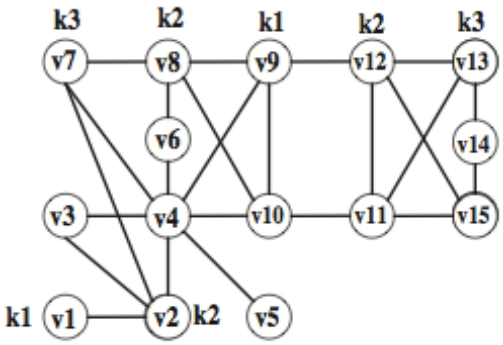
(例) 検索キーワード：{k_1, k_2}
コンポーネント内のキーワード間の最大距離 γ=2



ソーシャルメディアデータ

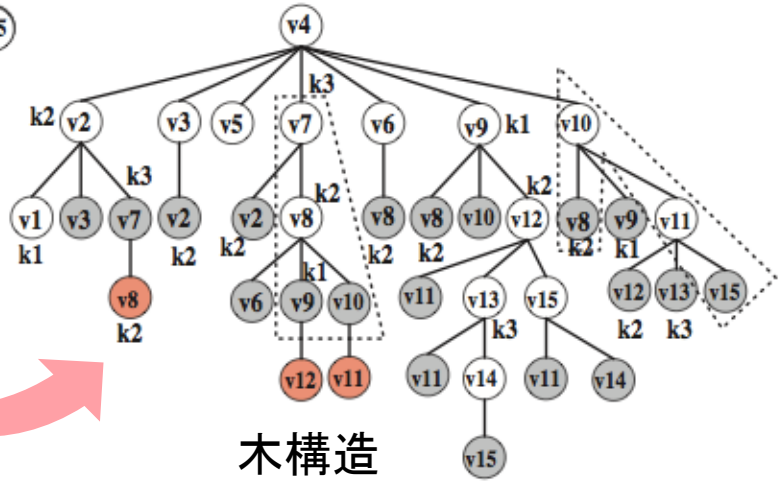
提案

- コンポーネントを
  効率的に見つけ出す方法を提案する.
  - 具体的には，新たな木構造を提案している.



グラフ

パラメータγによって,
木構造の深さをコントロールする
ことができる.



木構造

- 利点
  - 距離を記憶した
    木構造をメイン
    メモリ上に展開
    することができる.
  - 木構造を作らずに
    グラフから
    相互関係のある
    コンポーネントを
    探すと,
    ペアワイズ比較
    (N^2の比較)を
    要する.
    しかし，この
    木構造を作れば
    比較はない.

実験

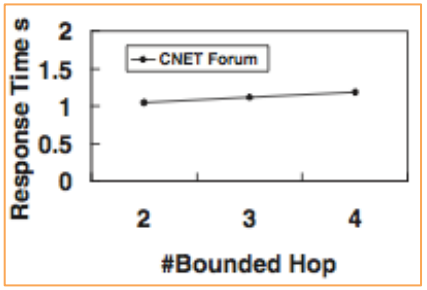
- データセット

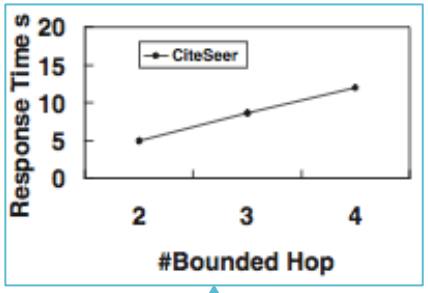
Dataset Name	Size	Number of Nodes	Number of Edges
CNET Forum	7MB	106,436	79,217
CiteSeer	89MB	230,470	406,784
Wikipedia	1,141MB	5,716,808	65,080,196

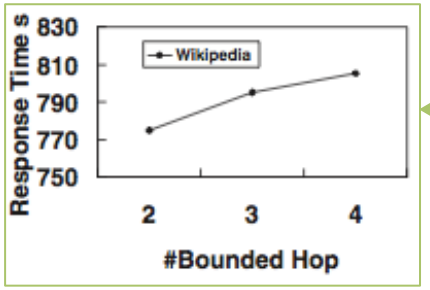
どちらも気にならない程度の時間や使用メモリ領域なので，これを使うのは効果的である.

- 実験結果

①グラフから木への変換にかかる時間

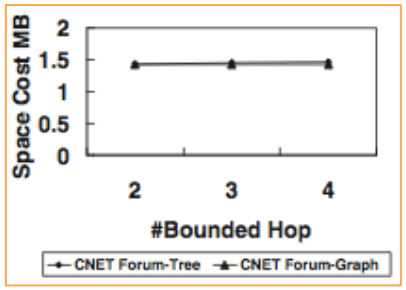


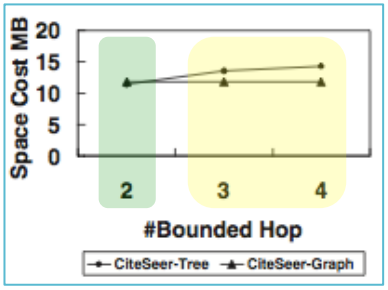


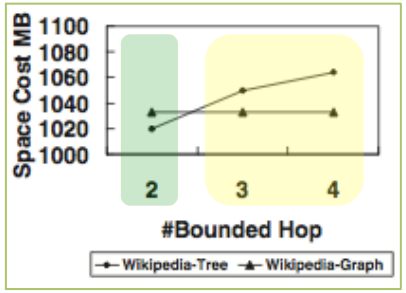


平均次数が10以上のときは増加傾向にある.

②木とグラフの使用メモリ領域







r=2のときは木の方がコストがかからない.
$r \geqq 3$のときは木の方がコストがかかるが，気にならない程度.