隠れた名品を推薦するための グラフマイニング手法の提案

佐 藤　哲 司

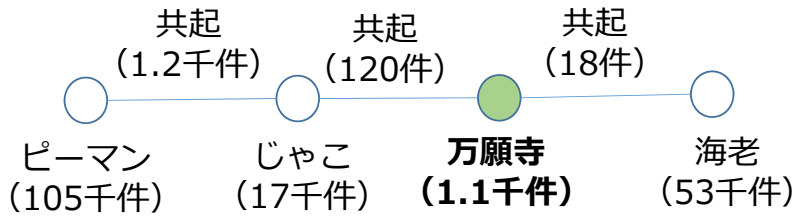
筑波大学

隠れた名品

万願寺とうがらし

- 15年ほど前に京都の料理屋で始めて「何じゃこりゃ？！」
  - 唐辛子？ ピーマン？
    果肉は厚くて柔らかく甘みが深い
  - クックパッドで検索：1,112品
    - そのまま焼いて食べても美味しいが，じゃこ，海老 と相性が良いらしい.

（ざるの直径は27cm）
出典：wikipedia

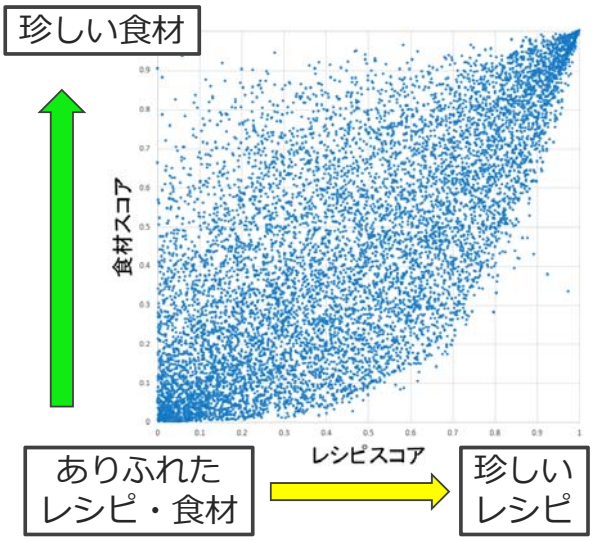


レシピマップ

- レシピの出現する食材
  - Cookpadデータ(NIIから提供）のレシピ集合の解析
  - 食材の出現頻度（レシピ数）は，きれいな『べき乗則』

- レシピと食材の定番度
  - ありふれた/珍しい食材
  - HITSアルゴリズムを適用
    - Authority: 食材，
      Hub: レシピ
    - ありふれた食材だけで
      調理したレシピ
    - 滅多に使用されない食材
      を使ったレシピ

  → 容易にレパートリー拡大できるレシピ推薦の順序 [中岡, DEIM2015]

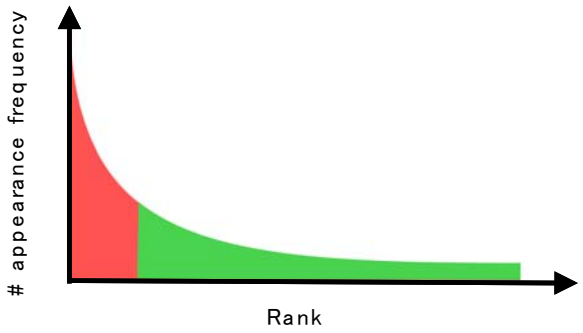


大規模データに対するマイニング

- たいていのデータはべき乗則
  - Cookpadは285万件 → 全数を使った解析が可能
  - 楽天市場のレビュー記事：6,500万件 → 全数はちょっと･･･
- 頻出パターン分析
  - Short Head: 高頻度の領域から典型的なパターンを抽出
  - Long Tail: データ件数は多いがノイズも多い･･･

  → 計算機能力が上がっても，闇雲に解析しても知見は得られない
  低頻度だけど『万願寺とジャコの相性が良い』に相当する関係を知りたい

マイニングの対象領域を
・出現頻度ではない方法で
・局所的な偏りを手がかりに
切り出したい



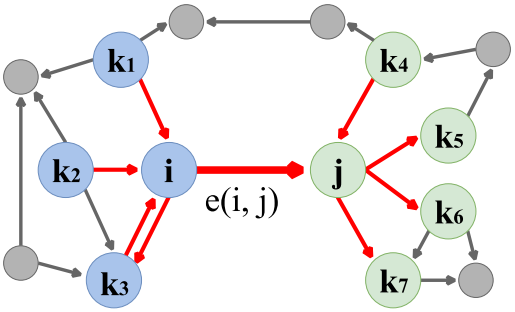
アイデア！

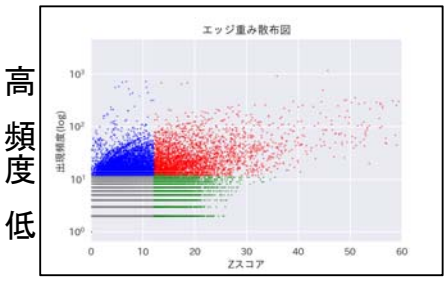
- Ｚスコア
  - 値を標準化して比較可能とする手法を共起グラフに適用
  - あるノードの周辺エッジのうちで，高頻度のエッジに高い値
- 分析対象グラフ
  - Ｚスコアの値と出現頻度の組み合わせで抽出

エッジの出現頻度 *周辺エッジの平均*

$$Zscore\big(e(i,j)\big) = \frac{w\big(e(i,j)\big) - \mu(e(i,j))}{\sigma(e(i,j))}$$

周辺エッジの標準偏差





大規模な実データに適用

- 楽天市場6,500万レビュー
  - 情報学研究データリポジトリ（楽天株式会社, NII）
- レビュー履歴 ≒ 購買履歴 とするための前処理
  - 「投稿者が一意に判別できる」レビューを抽出
  - 「購入の有無が確認できない」「投稿日時が欠落している」レビューを除外

244万ユーザによる1,780万レビュー

購買順序グラフ（ＰＨＧ）の構築

ノード: 商品

エッジ: 個々の消費者が
連続して購入した2商品の順序

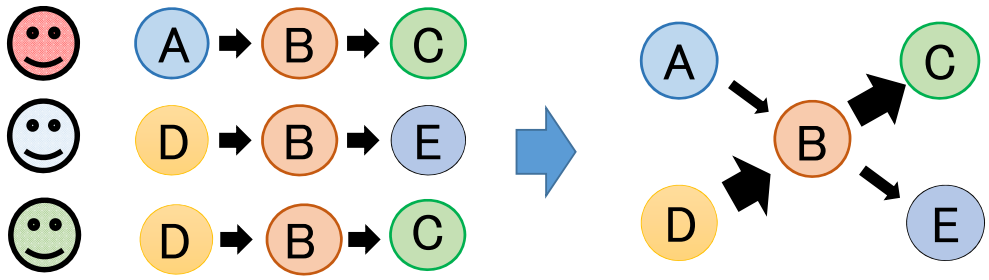
エッジの重み

$w\big(e(A,B)\big) = 1$

$w\big(e(B,C)\big) = 2$

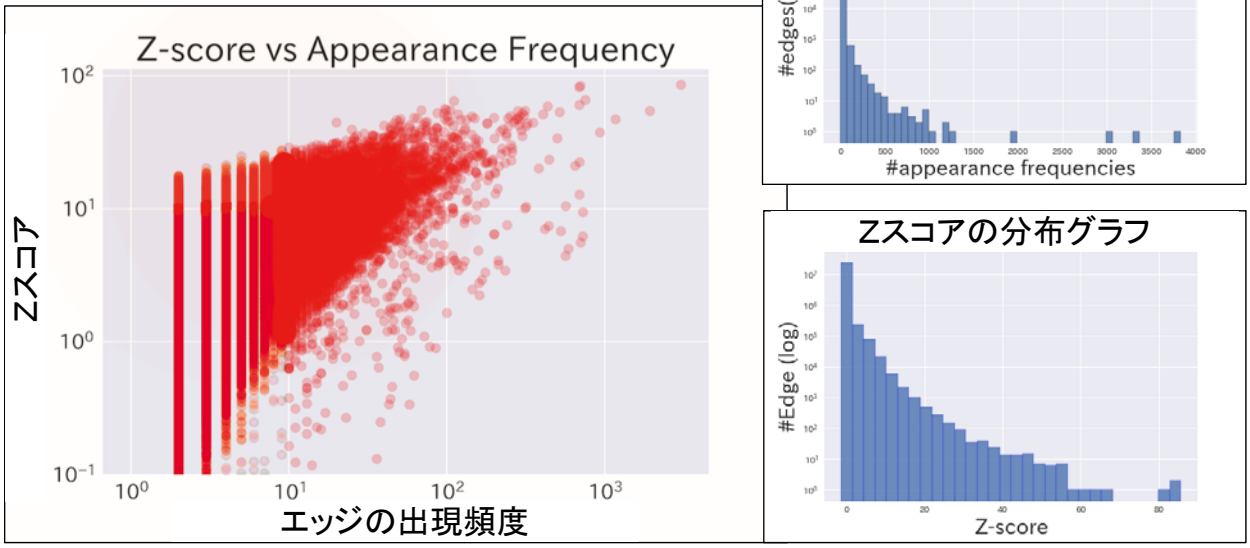
$w(e(D,B)) = 2$

$w(e(B,E)) = 1$



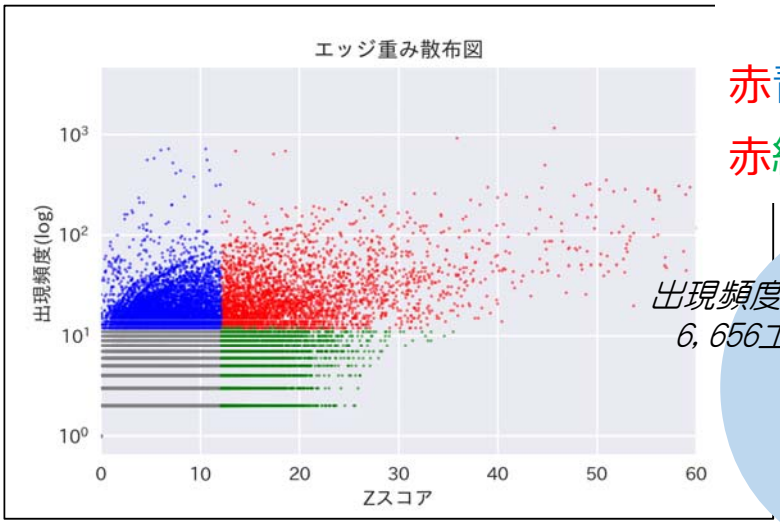
購買行動グラフの特性

- 1,780万レビューから構築したPHG
  - 6,721,476 item nodes
  - 24,325,221 edges

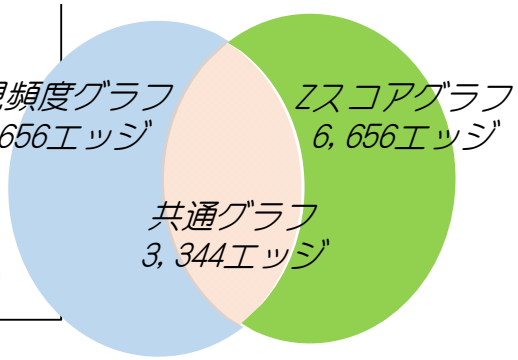


サブグラフの抽出と特性比較

- 出現頻度順 / Zスコア順に
  上位10,000エッジからサブグラフを抽出



赤　：共通グラフ
　　高頻度・高Zスコア
赤青：出現頻度グラフ
赤緑：Zスコアグラフ



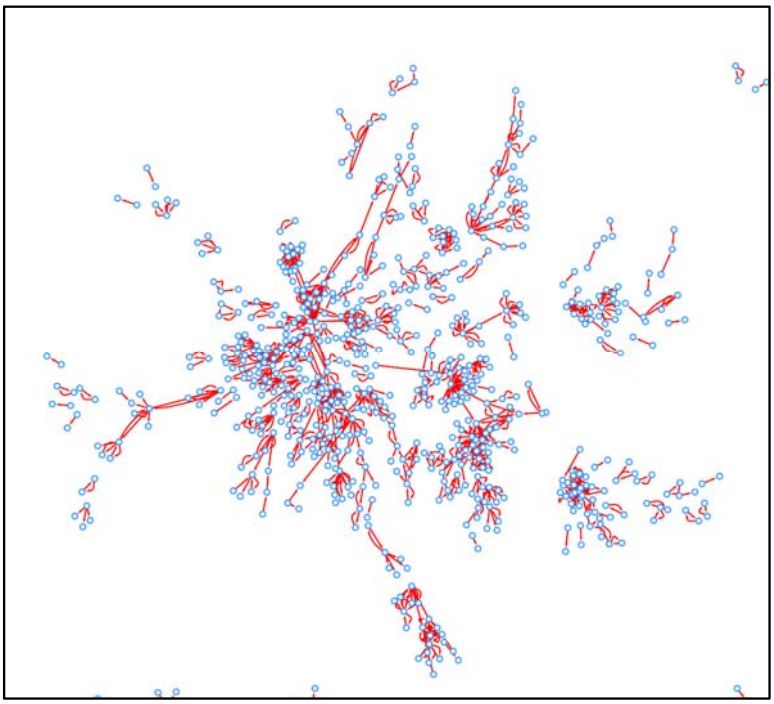
ノード数・エッジ数の比較

- エッジ抽出前と抽出後の3グラフを比較
  - 抽出グラフのエッジ数は10,000で同一なのに対して,
    Ｚスコアグラフのノード数が2倍以上
  - 広範囲のノード（アイテム）を含むグラフを抽出

グラフ種別	ノード数	エッジ数
初期グラフ	6,721,476	24,325,221
出現頻度グラフ	6,932	10,000
Zスコアグラフ	15,468	10,000
共通グラフ	3,874	3,344

抽出グラフの可視化（1）

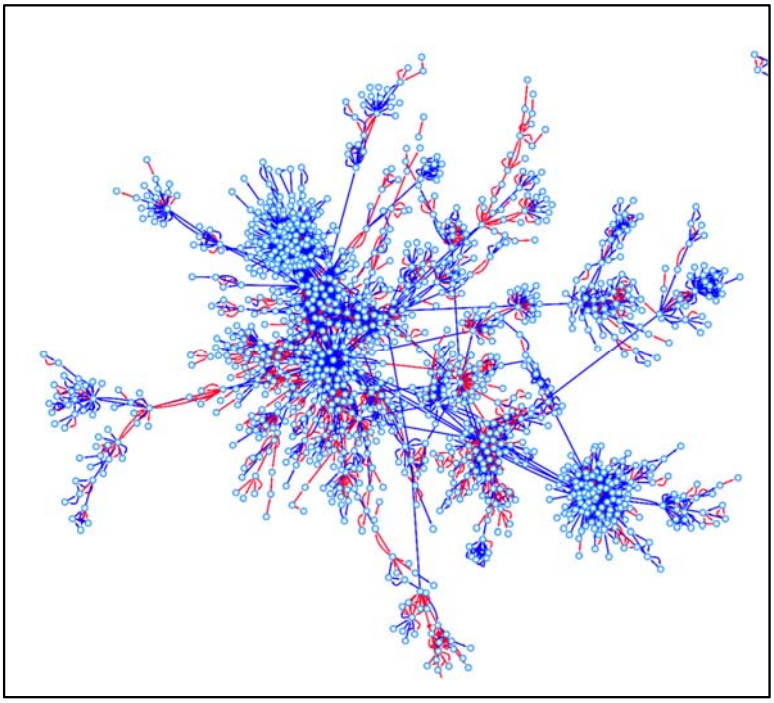
- 共通グラフ（一部）
  - 出現頻度もＺスコアも高いグラフ
  - （従来の）頻出パターン抽出でも重要な構造



抽出グラフの可視化（2）

- 出現頻度グラフ（一部）
  - 共通グラフを補強
    ノード数の増：小
    エッジ数の増：大

グラフの
高密度化

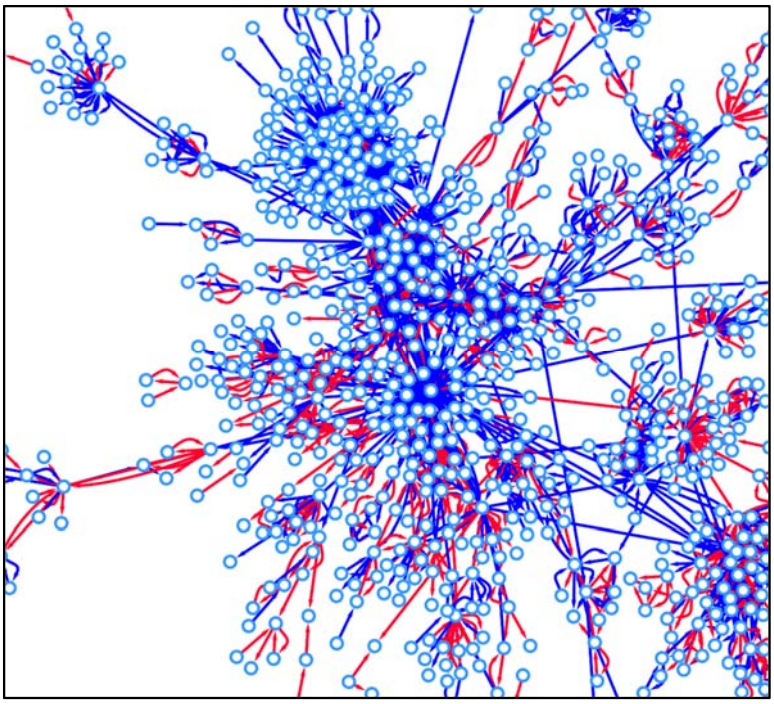


抽出グラフの可視化（3）

- 出現頻度グラフ（一部）
  - 共通グラフを補強
    ノード数の増：小
    エッジ数の増：大

グラフの
高密度化

中心部を拡大

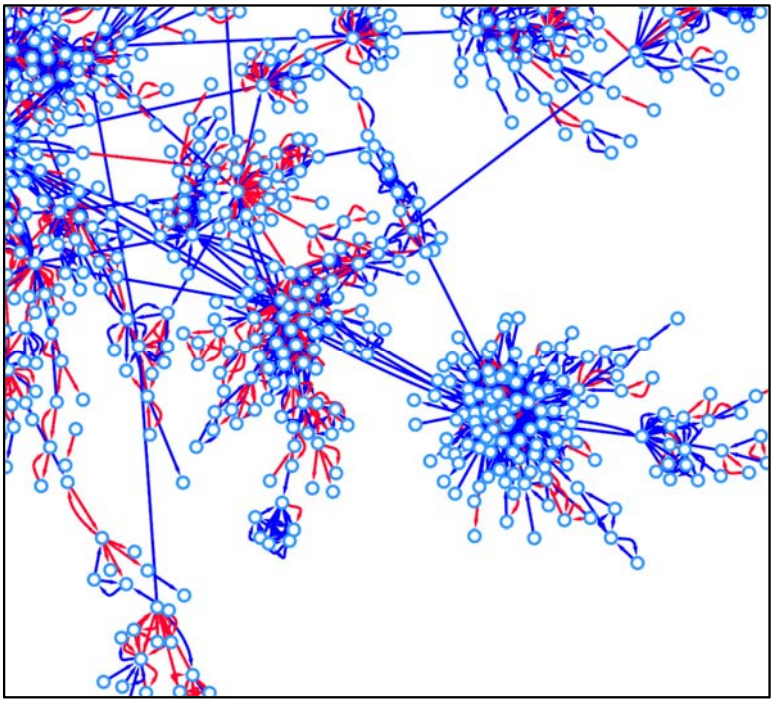


抽出グラフの可視化（4）

- 出現頻度グラフ（一部）
  - 共通グラフを補強
    ノード数の増：小
    エッジ数の増：大

グラフの
高密度化

周縁部を拡大

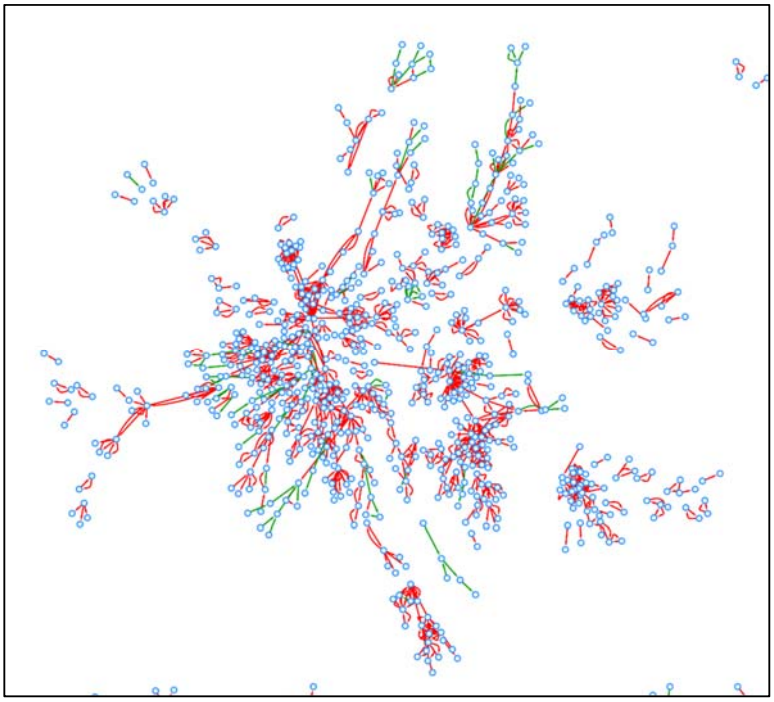


抽出グラフの可視化（5）

- Ｚスコアグラフ（一部）
  - 共通グラフを拡張
    ノード数の増：大
    エッジ数の増：小

↓

グラフの
周縁部拡張

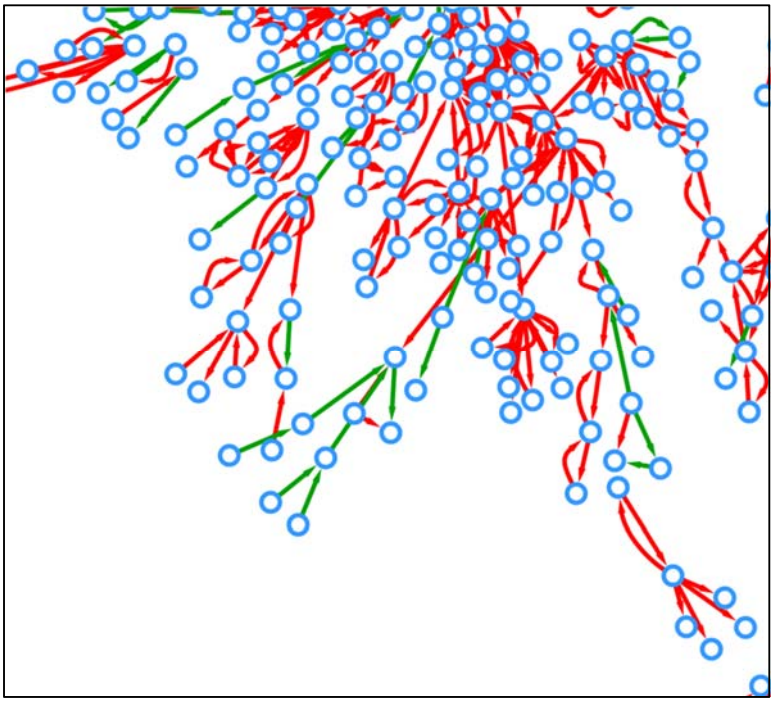


抽出グラフの可視化（6）

- Ｚスコアグラフ（一部）
  - 共通グラフを拡張
    ノード数の増：大
    エッジ数の増：小

↓

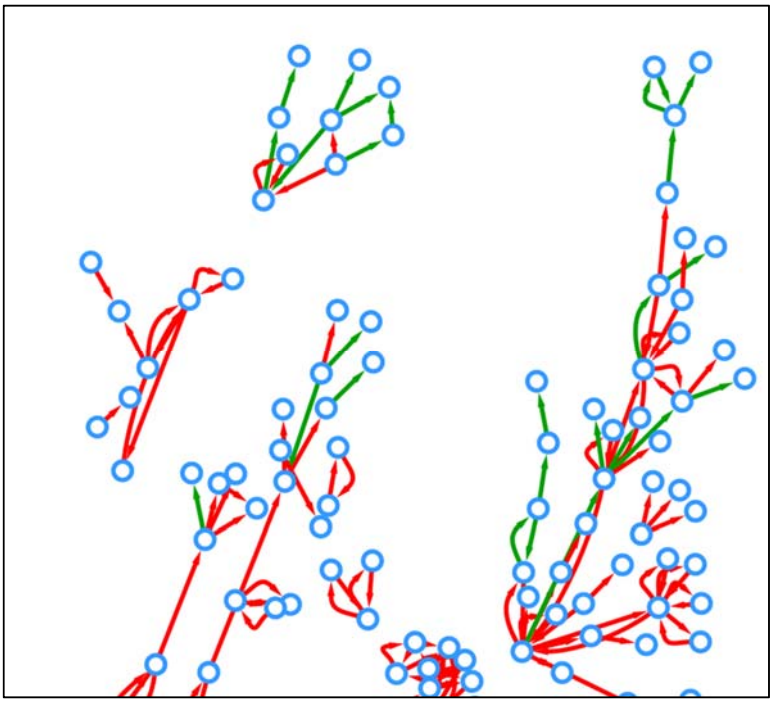
グラフの
周縁部拡張



抽出グラフの可視化（7）

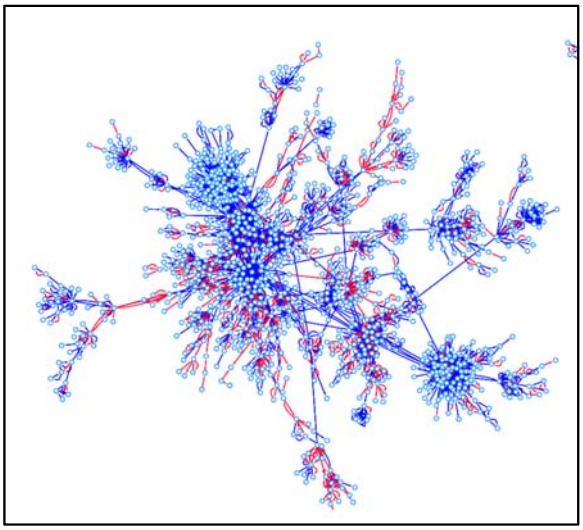
- Ｚスコアグラフ（一部）
  - 共通グラフを拡張
    ノード数の増：大
    エッジ数の増：小

グラフの
周縁部拡張

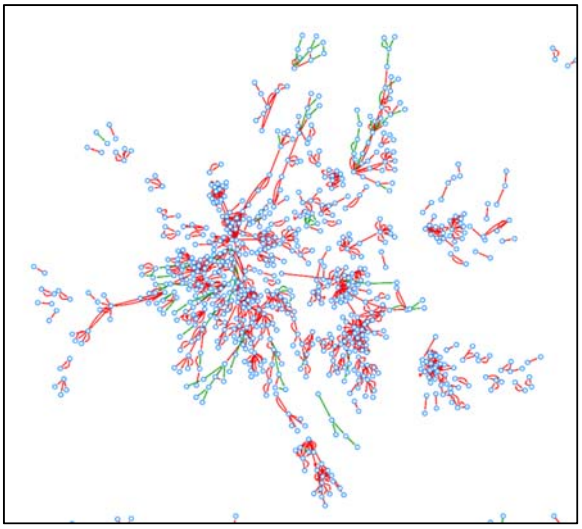


抽出グラフの比較

- 出現頻度（従来法）vs. Ｚスコア（提案法）



出現頻度グラフ
内部の高密度化

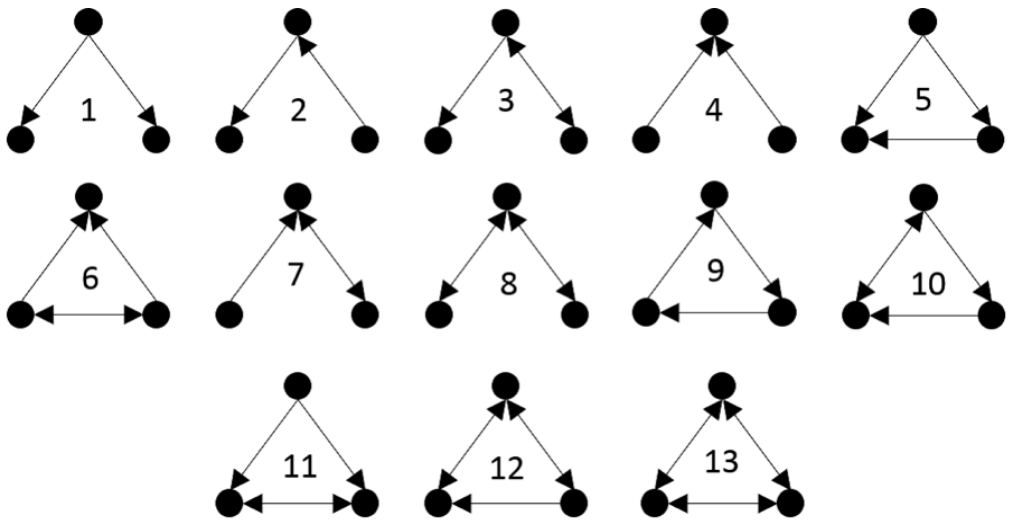


Zスコアグラフ
周縁部の拡張

モチーフ分析による成長分析

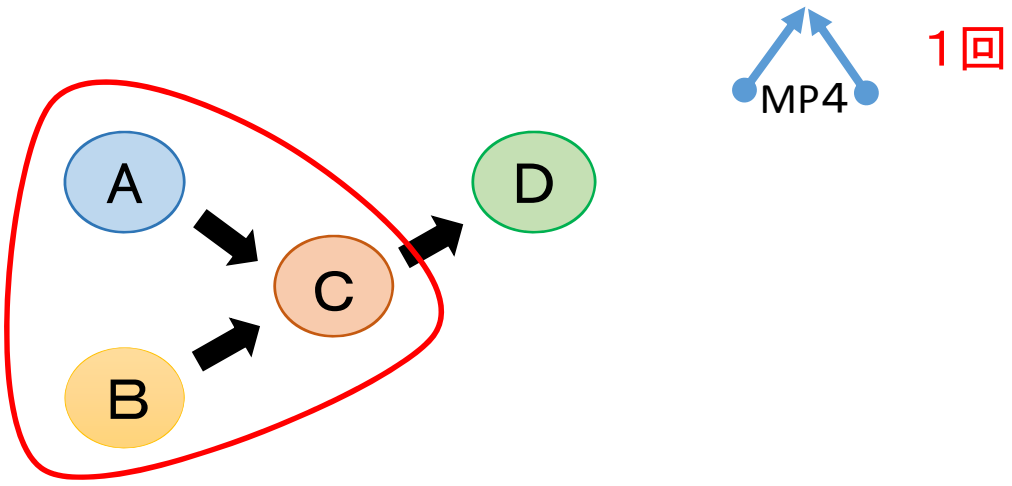
3ノードモチーフ

- 13種類のモチーフ
  - ３個のノード間で考えられる全ての順序関係
  - モチーフの数え上げで購買順序グラフ(PHG)の特徴を抽出



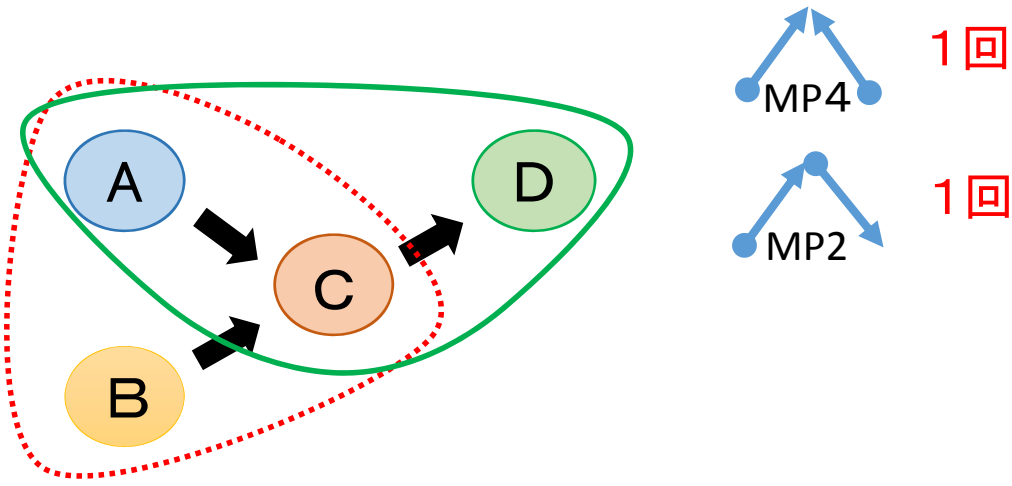
モチーフパターンの例

- 購買順序グラフ(PHG)の特徴



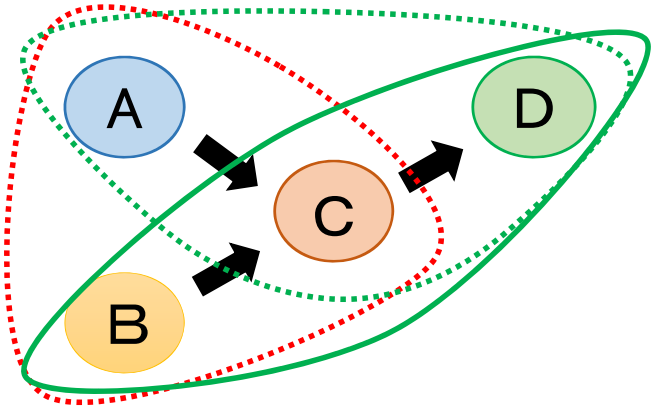
モチーフパターンの例（つづき）

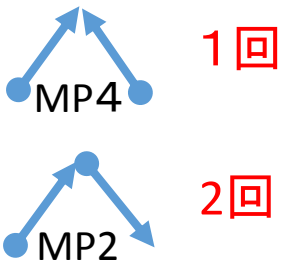
- 購買順序グラフ(PHG)の特徴



モチーフパターンの例（正規化）

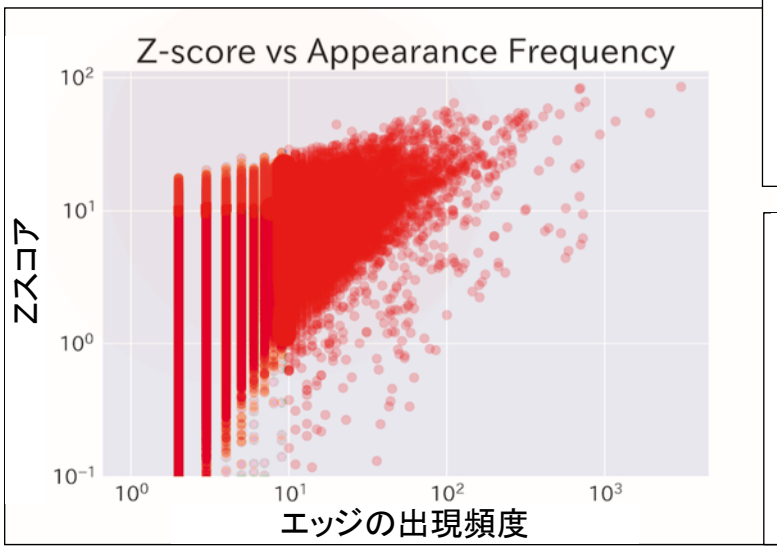
- 購買順序グラフ(PHG)の特徴
  - モチーフベクトル＝（ 0, 2, 0, 1, 0, …, 0 ）
  - 正規化ベクトル　＝（ 0, 2/3, 0, 1/3, 0, …, 0 ）

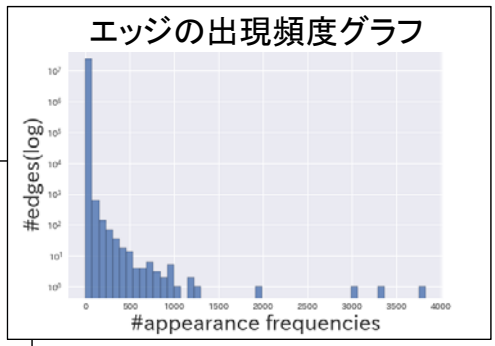


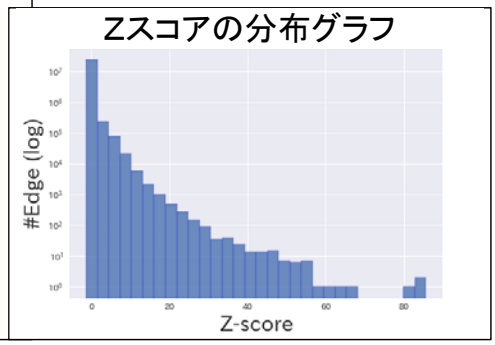


購買行動グラフの評価

- 1,780万レビューから構築したPHG
  - 楽天市場のレビュー記事から構築
  - 6,721,476 item nodes
  - 24,325,221 edges







評価用サブグラフの抽出

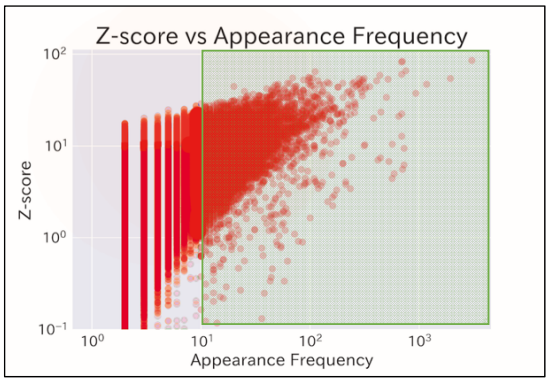
- Ｚスコア vs 出現頻度
  - 緑領域：エッジの出現頻度 ≧ 10
  - 赤領域：Ｚスコア ≧ 10

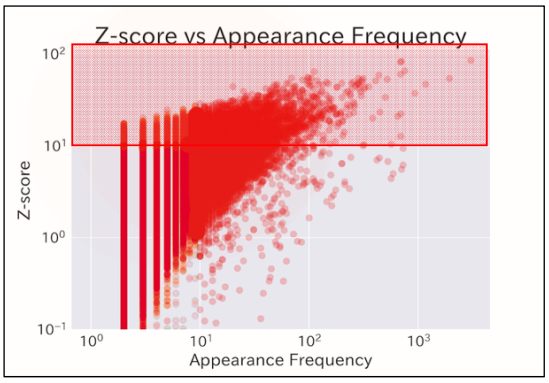
出現頻度が高いサブグラフ
- 10,266 ノード
- 13,939 エッジ

1.36 倍

Ｚスコアが高いサブグラフ
- 18,914 ノード
- 20,600 エッジ

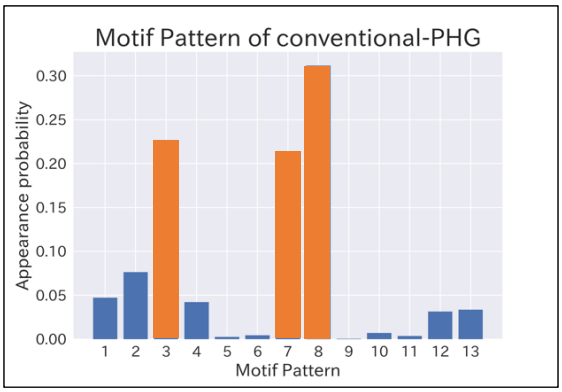
1.09 倍

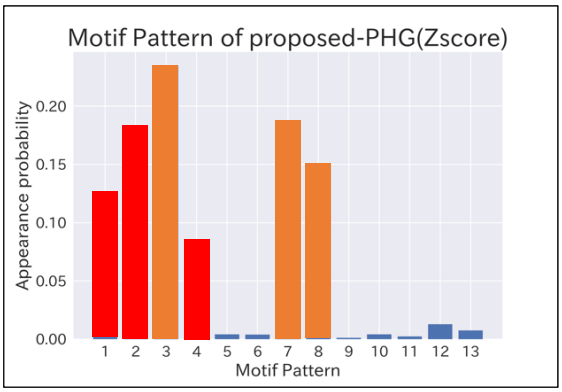


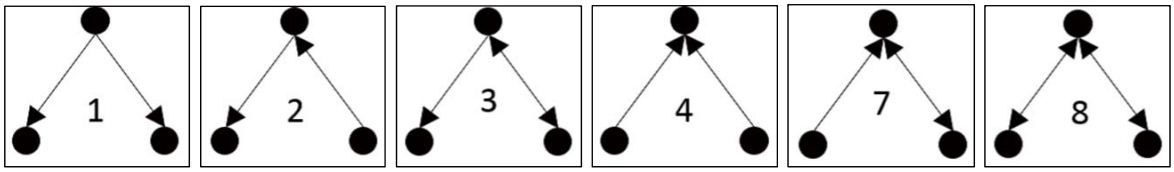


モチーフパターンの比較

- 出現比率の高いパターン
  - 高エッジ頻度： MP 3, 7, 8
  - 高Ｚスコア　： MP 1, 2, 3, 4, 7, 8

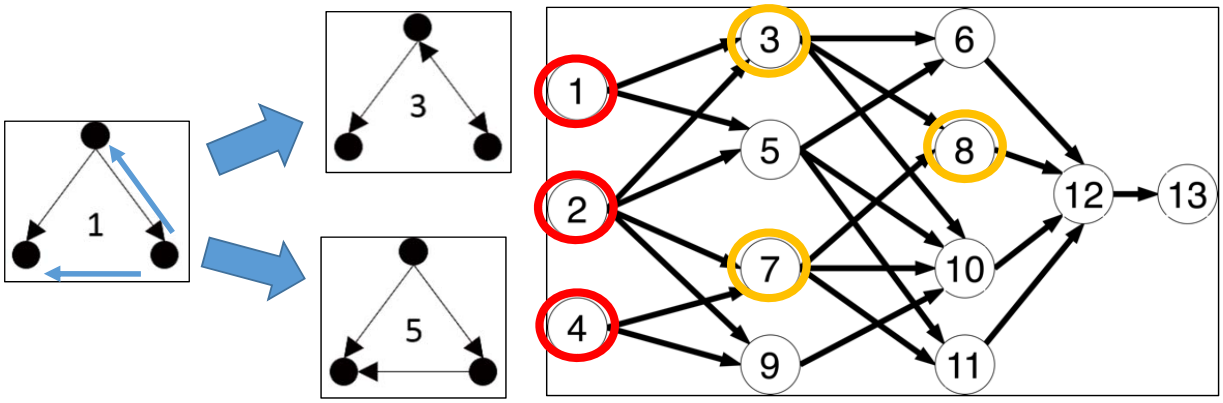






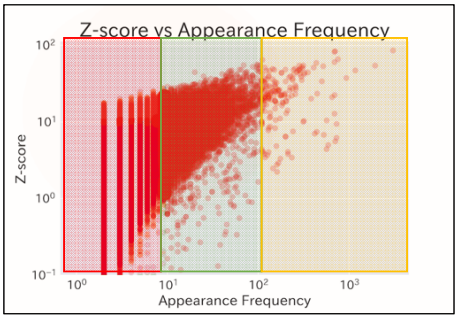
モチーフの成長遷移

- 成長： エッジが追加されることでモチーフが遷移
  - MP1に逆向きエッジを追加 → MP 3
  - MP1に新たなエッジを追加 → MP 5
- 購買順序グラフの成長
  - 高エッジ頻度： MP 3, 7, 8
  - 高Ｚスコア　： MP 1, 2, 3, 4, 7, 8 ･･･ **成長前のモチーフを抽出**



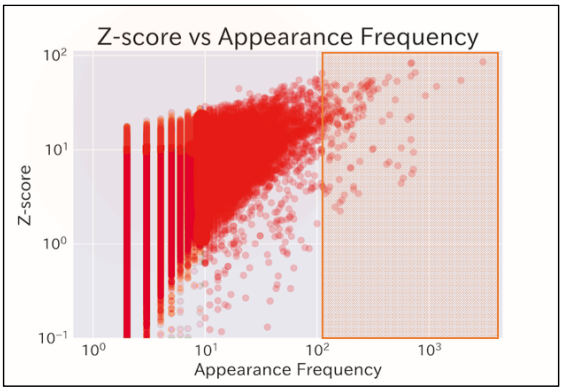
Ｚスコアに対するモチーフパターンの遷移

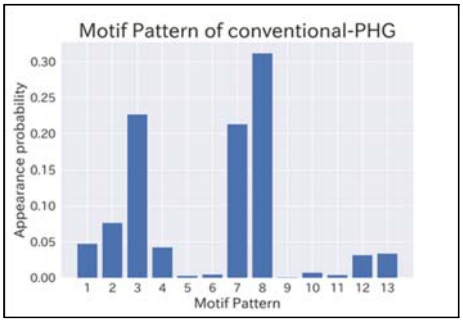
- エッジの出現頻度で評価領域を区分
  - 低頻度域：　　出現頻度 ＜ 10
  - 中頻度域：　10 ≦ 出現頻度 ＜ 100
  - 高頻度域： 100 ≦ 出現頻度
- モチーフパターンの遷移
  - 着目するモチーフ：出現比率が高い MP 1, 2, 3, 4, 7, 8
  - Ｚスコアの閾値を変えて，
    各モチーフの出現比率をトレース

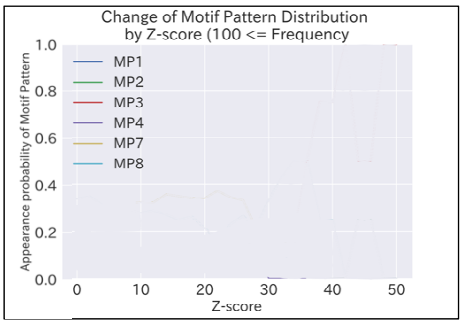


高頻度域のモチーフパターン変化

- 高頻度域 全域（Ｚスコアの閾値＝0）
  - 頻度で足切りする従来法に相当
    - 出現頻度 ≧ 100
  - MP 3, 7, 8 が高確率で出現

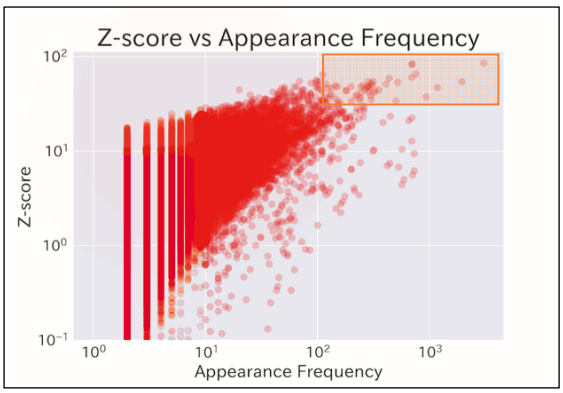


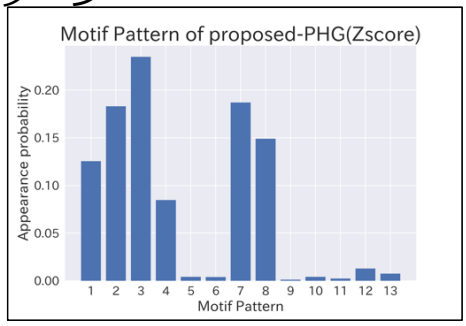


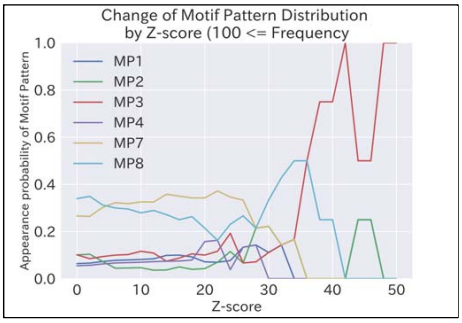


高頻度域のモチーフパターン変化

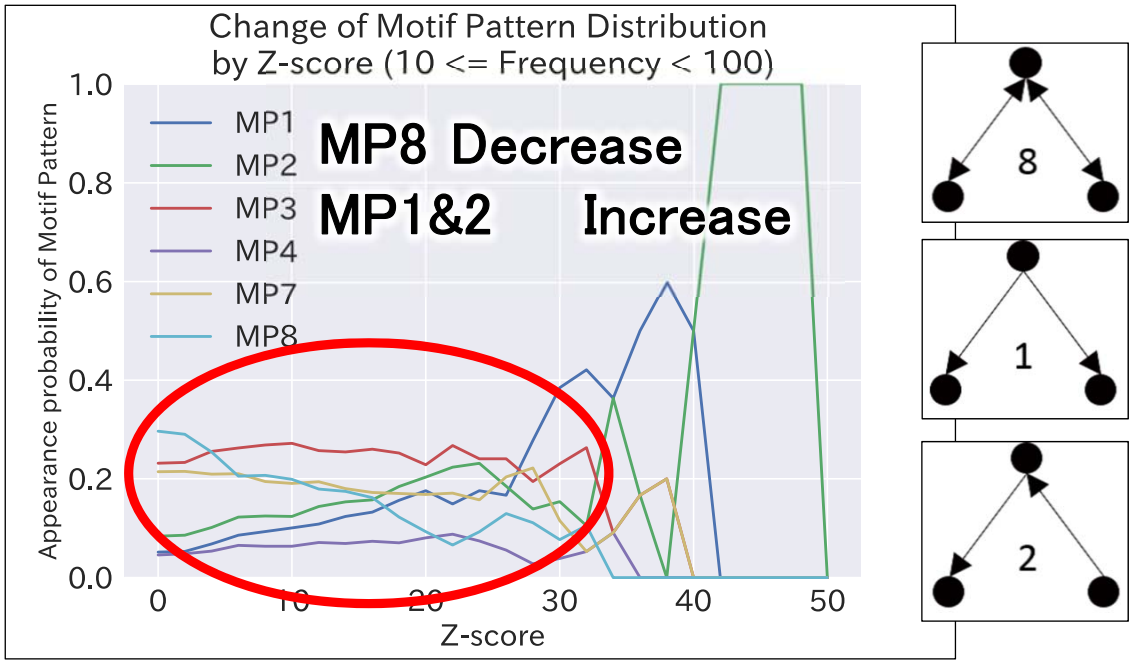
- 高頻度域 でＺスコアの閾値を上昇
  - 右上図は Zscore ≧ 20 のモチーフパターン
  - MP 1, 2, 3, 4, 7, 8 が高確率で出現





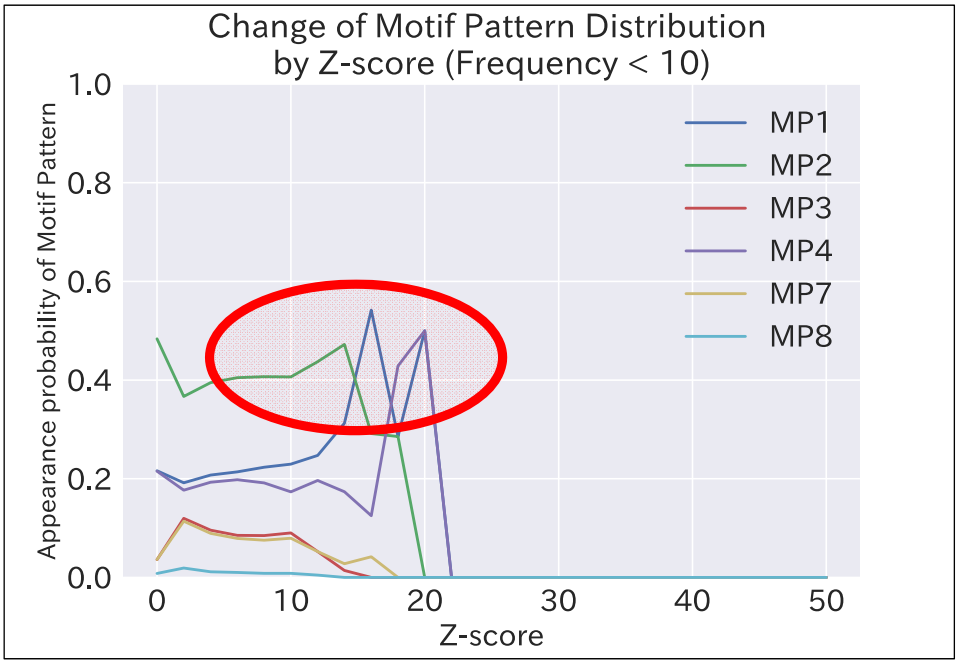


中頻度域でのモチーフパターンの変化



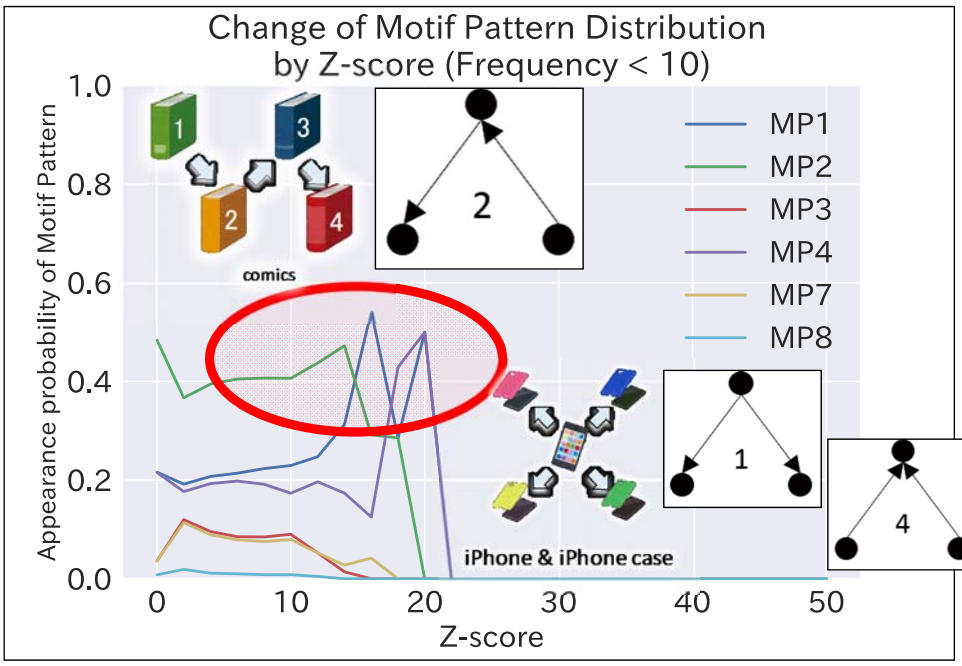
低頻度域でのモチーフパターンの変化

- Ｚスコアの閾値を上げるとMP2が減，MP1, 4が増
- Ｚスコア 20以上のエッジは存在しない



低頻度域でのモチーフパターンの変化

- MP2: 連続購入･･･ シリーズもののコミックなど
- MP1, 4: ハブ型購入･･･本体とカバー，周辺機器など



まとめ

- 大規模な社会ネットワーク分析
  - 大規模 ≒ べき乗則に従ったデータ分布
  - 従来法：Short Head における頻出パターンに着目
    提案法：Long Tail における局所的に得意なパターンに着目
- 分析対象グラフの抽出
  - 分析の前処理に相当： Long Tail に存在する『隠れた名品』
  - **Ｚスコア**の適用：**ノードの周辺で優位なエッジ**を抽出
- ネットショッピングの履歴データで検証
  - 弱連結成分分解したグラフの可視化
    - 従来法：グラフの強度を重視
    - **提案法：グラフの拡がり（周縁部の拡張）**を重視
  - モチーフ分析
    - **成長前のモチーフ抽出**に成功･･･『隠れた名品』の可能性大
    - Ｚスコアの閾値･･･異なる特徴のモチーフを抽出

関連論文

- 稲福 和史, 伏見 卓恭, 佐藤 哲司: 近傍エッジとの関係に着目したグラフマイニング手法の提案と評価, DEIM2018, J2-5 (Mar. 4-6, 2018).
- Kazufumi Inafuku, Takayasu Fushimi, Tetsuji Satoh: Graph Generation Method Based on Relative Value of Neighbor Edges, iiWAS2017, pp. 358-368 (Dec. 4 - 6, 2017).
- 稲福 和史, 伏見 卓恭, 佐藤 哲司: ECサイトにおける購買行動の成長分析, DICOMO2017, pp. 1107 - 1113 (June 28-30, 2017).
- 稲福 和史, 伏見 卓恭, 佐藤 哲司: レビュー順序グラフに基づく購買行動パターンの分析, DEIM2017, A3-4 (Mar. 3 - 9, 2017).
- Kazufumi Inafuku, Takayasu Fushimi, Tetsuji Satoh: Extraction Method of Typical Purchase Patterns Based on Motif Analysis of Directed Graphs, iiWAS2016, pp. 88-97 (Nov. 28 - 30, 2016).
- 中岡 義貴, 佐藤 哲司: 定番度に基づくレシピ推薦システムの提案, DICOMO2013, pp. 1083 - 1089 (July, 2013).
- 中岡 義貴, 佐藤 哲司: 調理レパートリー拡大のためのレシピ推薦手法の提案と評価, DEIM2015, C1-1 (Mar. 2015).