A Quantitative Analysis on the 17-19th Century Korean Culinary Manuscripts

Kihwang Lee 李基皇
Yonsei University

Jae Yun Lee 李在允
Kyonggi University

2010.11.30

Introduction

Purpose

- Quantitative similarity analysis on 25 culinary manuscripts in late Chosun period
- Boost the research on hand-written Hangeul manuscripts
- Promote the quantitative text analysis on historic data

Rationale

- Dual structure of culinary text
  - Macro-structure: list of recipes
  - Micro-structure: internal structure of recipe
- Key elements of culinary text: cooking methods and ingredients
- Distant reading

Data I

General Properties of the Data

- 25 hand-written culinary manuscripts
- 18-20 century texts (17 century text was not available)
- One large text $\rightarrow$ two texts — 26 texts
- Written in *Hangeul*
- Recipes for general foods and alcoholic beverages
- Number of recipes are varied
- Influential relations between manuscripts are not clear
- Each text was morphologically analyzed and annotated with semantic tags for food types, cooking methods, ingredients, etc.
- Old Korean was translated to modern Korean
- Costruction: 2-year large-scale project
- Encoded in XML

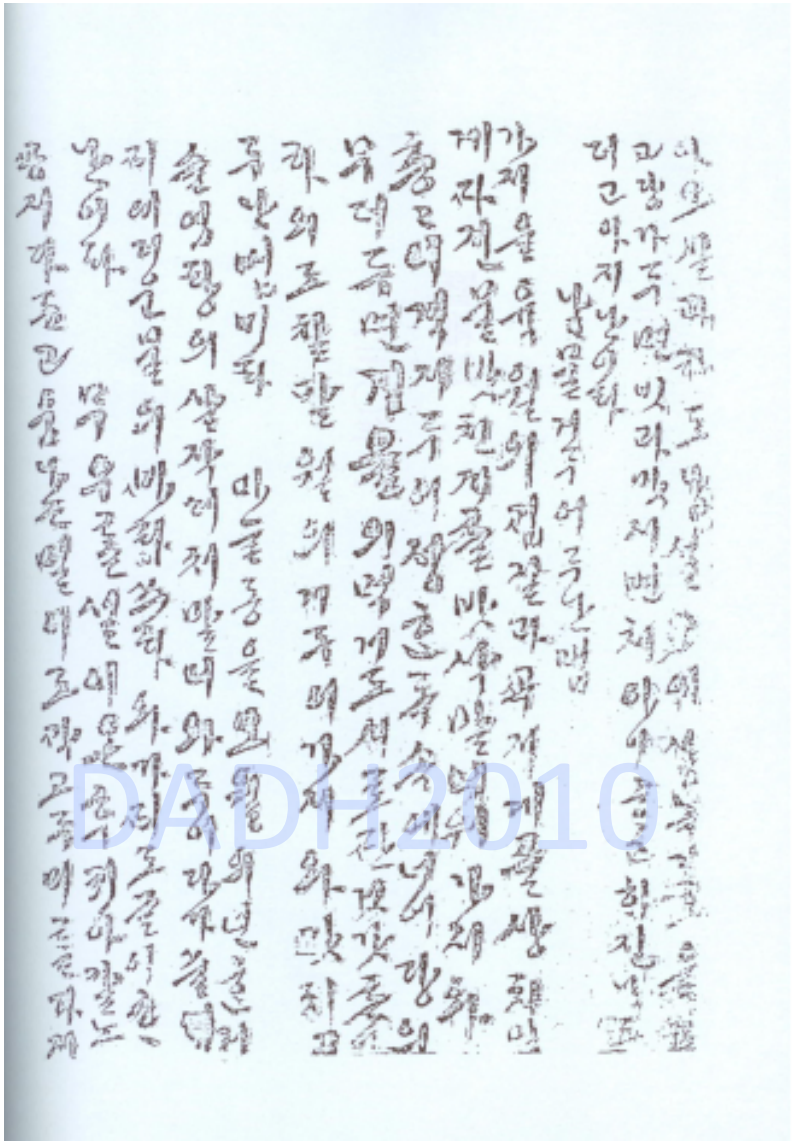
Data II

Title	Year	Author	Num. of Rcp.
酒方文鈔(JBMC)	1700s	unknown	5
Sulmandeuneunbeob(SMDNB)	1700s	unknown	47
飮食譜(ESB)	1700s	unknown	35
蘊酒法(OJB)	1786	unknown	71
酒食方(JSB)	1795	unknown	33
酒方(JB)	early 1800s	unknown	38
曆雜錄(YJL)	1830s	unknown	6
尹氏飮食法(YSESB)	1854	unknown	122
貞一堂雜識(JIDJJ)	1856	unknown	29
金承旨宅厨方文(GJBM)	1860	unknown	30
飮食冊(ESC)	1838/1898	Danyangdaeg	49
閨壺要覽(GGYL)	1896	unknown	31
閨閤叢書鄭良婉本(GHCSJ)	1800s	Bingheogag	243
閨閤叢書東京大本(GHCST)	1800s	Bingheogag	135
閨閤叢書嶺南大本(GHCSY)	1800s	unknown	20
飮食方文(ESBM)	mid 1800s	unknown	67
Sulbijneunbeob(SBNB)	late 1800s	unknown	30
李氏飮食法(ISESB)	late 1800s	Issi	52

Data III

Title	Year	Author	Num. of Rcp.
是議全書(SEJS)	late 1800s	unknown	256
酒食方文老稼齋公本(JSBMK)	late 1800s	unknown	89
酒食是儀(JSSE)	late 1800s	宋氏家	99
禹飮諸方(UEJB)	late 1800s	宋氏家	24
飮食方文NES(ESBMN)	late 1800s	unknown	4
酒食方文丁未年本(JSBM)	1907	unknown	46
寶鑑錄(BGL)	1927년	unknown	35

Data IV



Data V

```
<?xml version="1.0" encoding="utf-8" ?>
 <fc:corpus id="F0001" xmlns:fc="http://www.knu.ac.kr/FoodCorpus"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.knu.ac.kr/FoodCorpus FoodCorpus.xsd">
 <header>
 <bibl>
  <title>음식디미방</title>
  <author>Unknown</author>
  <authorYear>1670</authorYear>
  <desc />
  </bibl>
 <revision>
 <change>
  <date>2007-12-18</date>
  <worker>안의정</worker>
  <memo>기본 마크업</memo>
  </change>
  </revision>
  </header>
 <body>
 <foodItem origFoodName="면" transFoodName="면" page="1a" filename="음식디미방_1a.jpg">
  <head>음식디미방</head>
  <head>면병뉴</head>
  <head>면</head>
 <origText>
  <s no="1">것모밀를 씨어 하 ᄆᆞ이 ᄆᆞᆯ뇌디 말고 알마초 ᄆᆞᆯ뢰여</s>
  <s no="2">ᄲᆞᆯ을 조히 아아 디흘 제 미리 물 품겨 축축이 ᄒᆞ야 듯다가</s>
  <s no="3">디흘 제 녹도 거피ᄒᆞᆫ ᄲᆞᆯ 조히 시어 건져</s>
  <s no="4">믈 쐬거든 모밀ᄲᆞᆯ 닷되예 물 부ᄅᆞᆫ 녹두 ᄒᆞᆫ 복ᄌᆞ식 섯거 지ᄒᆞᄃᆡ 방하를 ᄀᆞ만ᄀᆞ만 지허 것ᄀᆞᆯ를 처 ᄇᆞ리고 키로 퍼
  ᄇᆞ리고 키 그테 흰 ᄲᆞᆯ이 나거든 그를 뫼화 다치 ᄒᆞ면 그 ᄀᆞᆯ리 ᄀᆞ장 희거든</s>
  <s no="5">면 ᄆᆞᆯ 제 더운물에 눅게 ᄆᆞ라 누ᄅᆞ면 비치 희고 조ᄒᆞᆫ 면이 되ᄂᆞ니라</s>
  <s no="6">교틱ᄂᆞᆫ 싀면 교틱 ᄀᆞ치 ᄒᆞ라</s>
  </origText>
```

Data VI

```
<s tid="0">
    <tok seq="0">
      <orth>음식디미방</orth>
      <mor pos="NNG">음식디미방</mor>
      <memo>음식디미방</memo>
    </tok>
  </s>
  <s tid="1">
    <tok seq="0">
      <orth>면병뉴</orth>
      <mor pos="NNG">면병뉴</mor>
      <memo>면병류</memo>
    </tok>
  </s>
  <s tid="2">
    <tok seq="0">
      <orth>면</orth>
      <mor pos="NNG">면</mor>
      <memo>면</memo>
    </tok>
  </s>
  <s tid="3">
    <tok seq="0">
      <orth>것모밀를</orth>
      <mor pos="NNG">것모밀</mor>
      <mor pos="JKO">를</mor>
      <memo>겉메밀+을</memo>
    </tok>
  </s>
```

Data VII

<unit sid="0" tokseq="0" senseClass1="음식명" senseClass2="술" senseClass3="발효주">사졀쥬</unit>
<unit sid="1" tokseq="0" senseClass1="음식재료" senseClass2="곡류">빅미</unit>
<unit sid="1" tokseq="2" senseClass1="계량단위" senseClass2="용량단위">말</unit>
<unit sid="1" tokseq="3" senseClass1="조리정도시기" senseClass2="조리정도">졍셰</unit>
<unit sid="1" tokseq="4" senseClass1="조리방법" senseClass2="비가열조리" senseClass3="가루내기">작말ᄒᆞ</unit>
<unit sid="1" tokseq="5" senseClass1="음식재료" senseClass2="물">물</unit>
<unit sid="1" tokseq="7" senseClass1="계량단위" senseClass2="용량단위">말</unit>
<unit sid="1" tokseq="8" senseClass1="조리정도시기" senseClass2="조리정도">범벅</unit>
<unit sid="1" tokseq="9" senseClass1="조리방법" senseClass2="비가열조리" senseClass3="섞기">긔</unit>
<unit sid="2" tokseq="0" senseClass1="음식재료" senseClass2="발효재">누룩</unit>
<unit sid="2" tokseq="2" senseClass1="계량단위" senseClass2="용량단위">되</unit>
<unit sid="2" tokseq="3" senseClass1="계량단위" senseClass2="용량단위">가옷</unit>
<unit sid="2" tokseq="4" senseClass1="조리방법" senseClass2="비가열조리" senseClass3="섞기">셕</unit>
<unit sid="2" tokseq="5" senseClass1="조리정도시기" senseClass2="조리정도">익</unit>
<unit sid="2" tokseq="6" senseClass1="음식재료" senseClass2="곡류">빅미</unit>
<unit sid="2" tokseq="8" senseClass1="계량단위" senseClass2="용량단위">말</unit>
<unit sid="2" tokseq="9" senseClass1="조리방법" senseClass2="비가열조리" senseClass3="씻기">빅셰ᄒᆞ</unit>
<unit sid="2" tokseq="10" senseClass1="조리방법" senseClass2="가열조리" senseClass3="끓이기">더우</unit>
<unit sid="2" tokseq="11" senseClass1="음식재료" senseClass2="물">물</unit>
<unit sid="2" tokseq="13" senseClass1="계량단위" senseClass2="용량단위">말</unit>

Data VIII

Cooking Method Annotation

- Heating: SIMMERING, PARBOILING, BOILING DOWN, STEAMING, PANFRYING, STIR FRYING, FRYING, GRILLING, SMOKING, DECOCTING, COOKING
- Non-heating: WASHING, PEELING, SCRAPING, SOAKING, STRAINING, DRYING, PICKLING, FLOURING, SHREDDING, SQUEEZING, PRESSING, MIXING[1], CUTTING, GRINDING, CRUSHING, FINE CHOPPING, TEARING, POUNDING, MIXING[2], COOLING, SHAPING, COATING, KNEADING, SEASONING
- Others: PICKLING, BREWING

Ingredient Annotation

WATER, ROOTS, MUSHROOMS, GRAINS, BEANS, VEGETABLES, FRUITS, MEAT, FISH, SEAWEEDS, EGGS, HERBS, SUGAR, STARTER, PETALS/LEAVES, OTHERS

Methods I

Network Analysis

- Co-occurrence network
- Pathfinder network (Schvaneveldt 1990)
- Proximity data $\rightarrow$ underlying organization of the data
- Co-occurrence similarity $\rightarrow$ text similarity

Cluster Analysis

- Assignment of a set of observations into subsets (*clusters*) $\rightarrow$ observations in the same cluster are similar
- Parallel Nearest Neighbor Clustering (PNNC) strategy (Lee 2006a)
- Link individual nodes on a network to their nearest neighbors in hierarchical way

Methods II

Centrality Analysis

- Determine the relative importance of a node on a network
- Relation between a node and other nodes linked from the node both directly and indirectly
- Scope of the centrality $\rightarrow$ local centrality vs. global centrality (Lee 2006b)

Text Similarity Analysis: Food Name Agreement I

Fundamentals

- Culinary texts are composed of recipes
- Texts containing common food names in recipe titles $\rightarrow$ similar texts

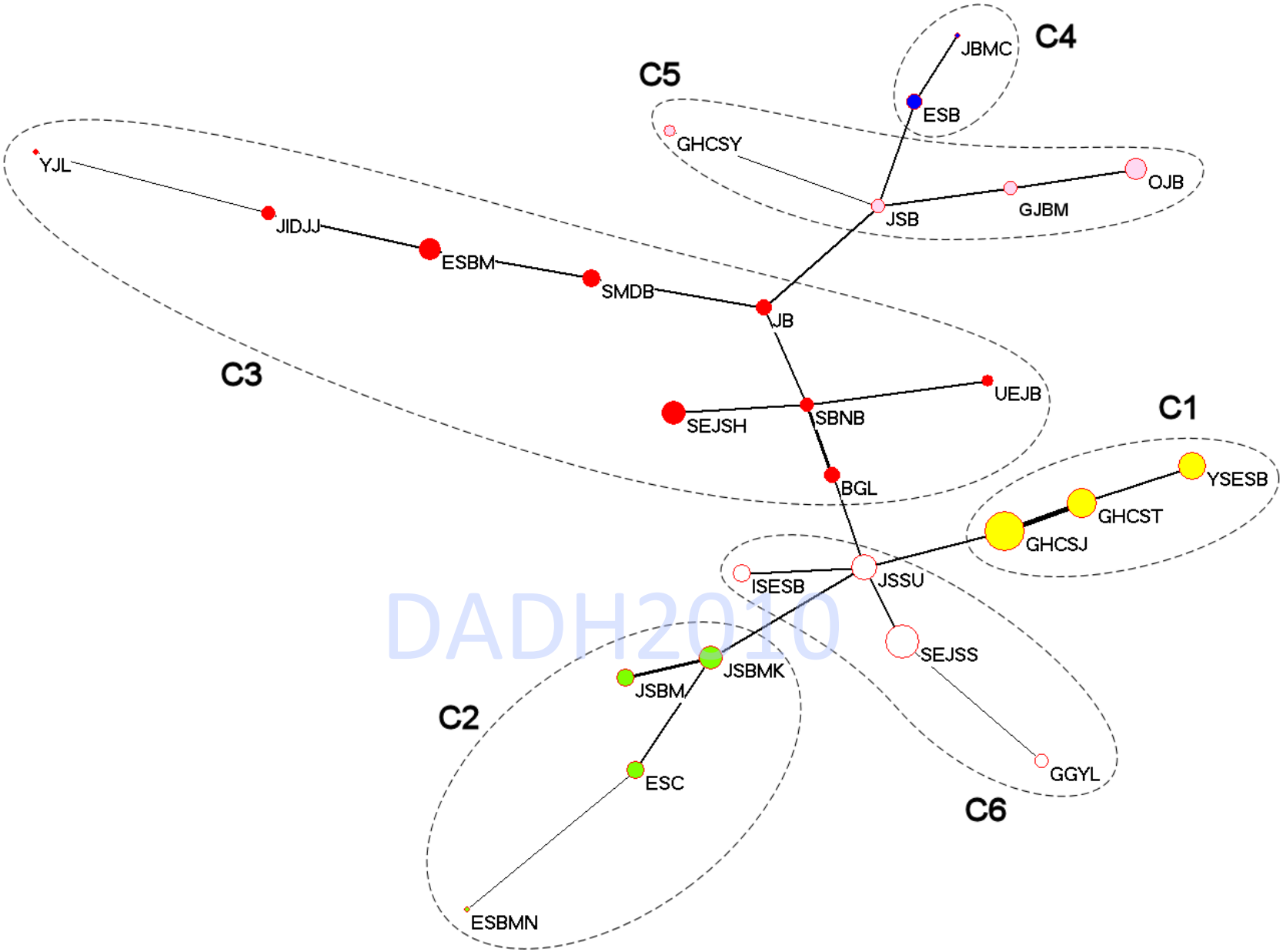
Procedure

1. Pre-process the recipe titles to get food names $\rightarrow$ 1,579 food names (1,043 types)
2. Construct a food name agreement count matrix for 26 texts
3. Normalize the agreement counts using the Jaccard coefficient:

$$J(T_1, T_2) = \frac{n(T_1 \cap T_2)}{n(T_1) + n(T_2) - n(T_1 \cap T_2)}$$

4. Produce a Pathfinder network over the matrix and apply the PNNC clustering

Text Similarity Analysis: Food Name Agreement II



Text Similarity Analysis: Food Name Agreement III

Results

- Size of a node: number of recipes in a text
- Distance: relative texts similarity
- C3, C4, C5: alcoholic beverage dominant clusters:
  eg. JBMC — brewing methods only
- C1, C2, C6: general food dominant clusters:
  eg. ESMBN — general cooking methods only
- GHCST and CHCSJ vs. GHCSY
- JBMC, YJL, and ESBMN containing small numbers of recipes

Limitations

- Food name agreement is not sufficient for similarity analysis
- Food names can be identical, but the actual recipes can be very different and vise versa

Cluster Analysis: Cooking Methods I

Fundamentals

- Cooking methods frequently co-occur in multiple recipes are closely related
- Examine the distribution of 37 cooking method tags in 1,598 recipes

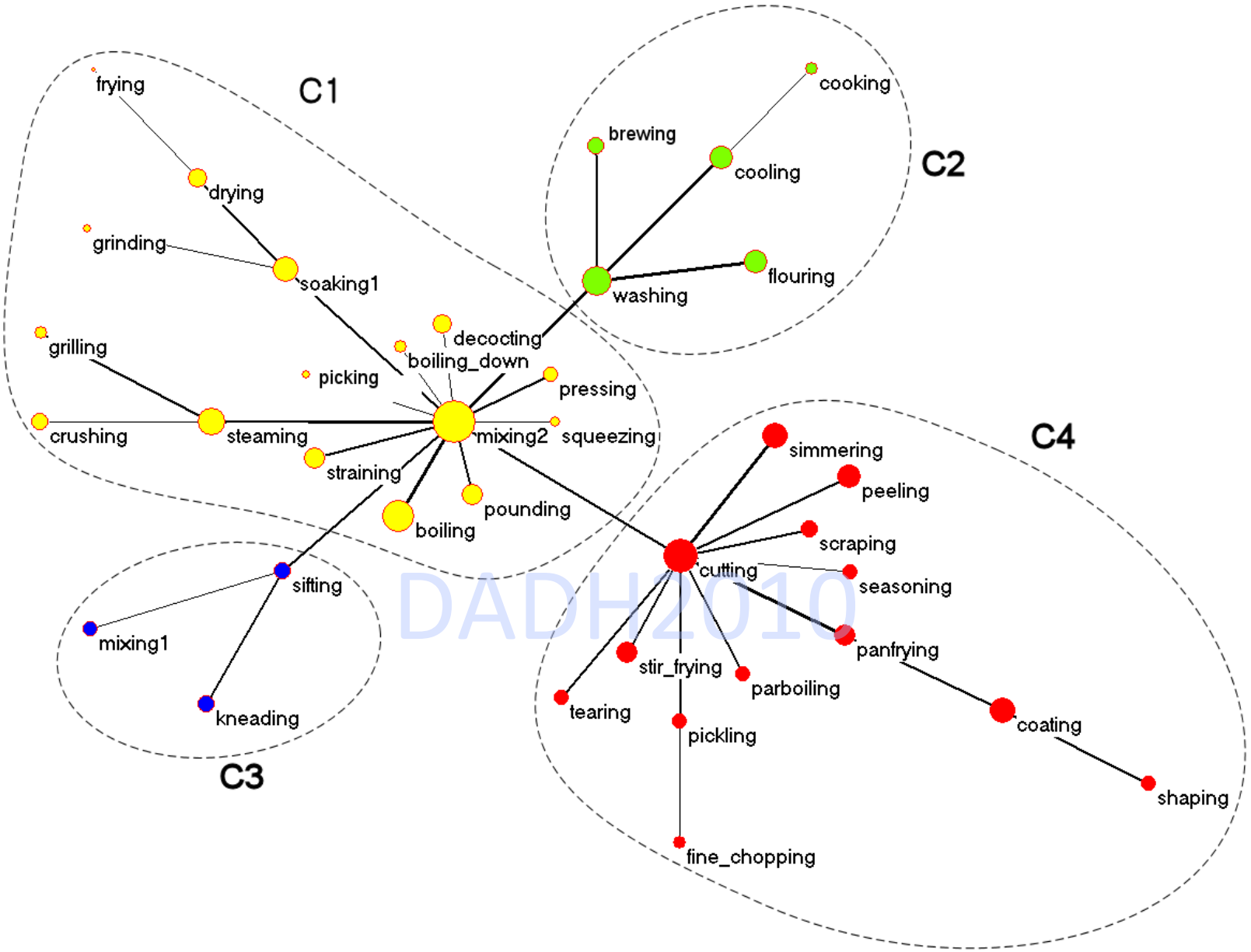
Procedure

1. Generate 37 cooking method vectors containing occurrence counts in 1,598 recipes
2. Compute cosine similarities between vectors:

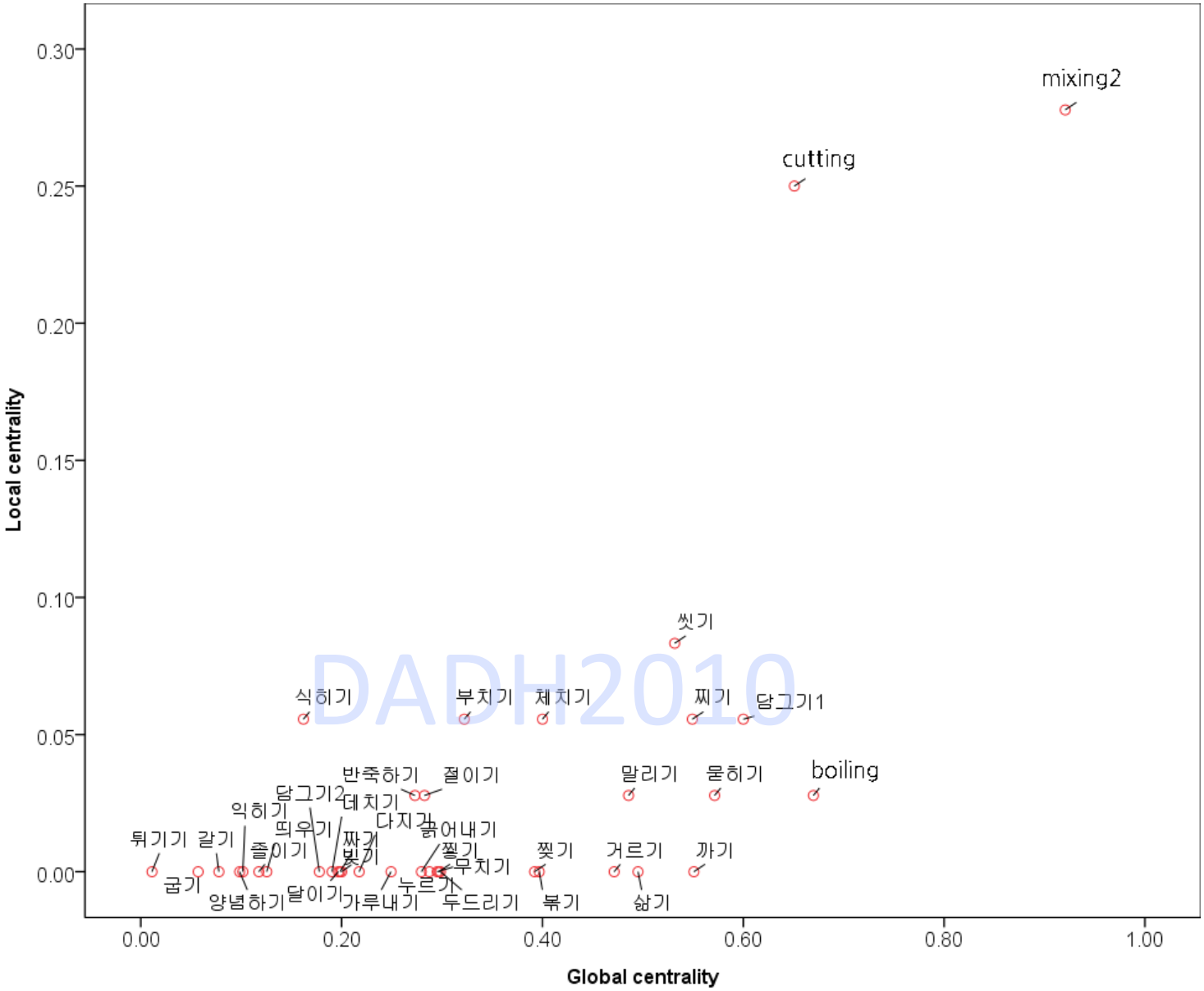
$$Sim_{\cos}(M_1, M2) = \frac{M_1 \cdot M_2}{||M_1|| ||M_2||}$$

3. Construct a Pathfinder network and apply the PNCC clustering
4. Apply centrality analysis on the network

Cluster Analysis: Cooking Methods II



Cluster Analysis: Cooking Methods III



Cluster Analysis: Cooking Methods IV

Results for Cluster Analysis

- MIXING^2, WASHING, and CUTTING are the most frequent and important cooking methods
- C1: largest cluster containing *intensive cooking methods* eg. MIXING^2, BOILING, and STEAMING
- C4: second largest cluster containing *light cooking methods* eg. CUTTING, SIMMERING, STIR FRYING
- C2: brewing methods, C3: preparation methods

Results for Centrality Analysis

- MIXING^2 and CUTTING have high global and local centralities $\rightarrow$ selectively used, discriminate recipes
- BOILING has high global centrality but low local centrality $\rightarrow$ widely used, background cooking method

Text Similarity Analysis: Cooking Methods I

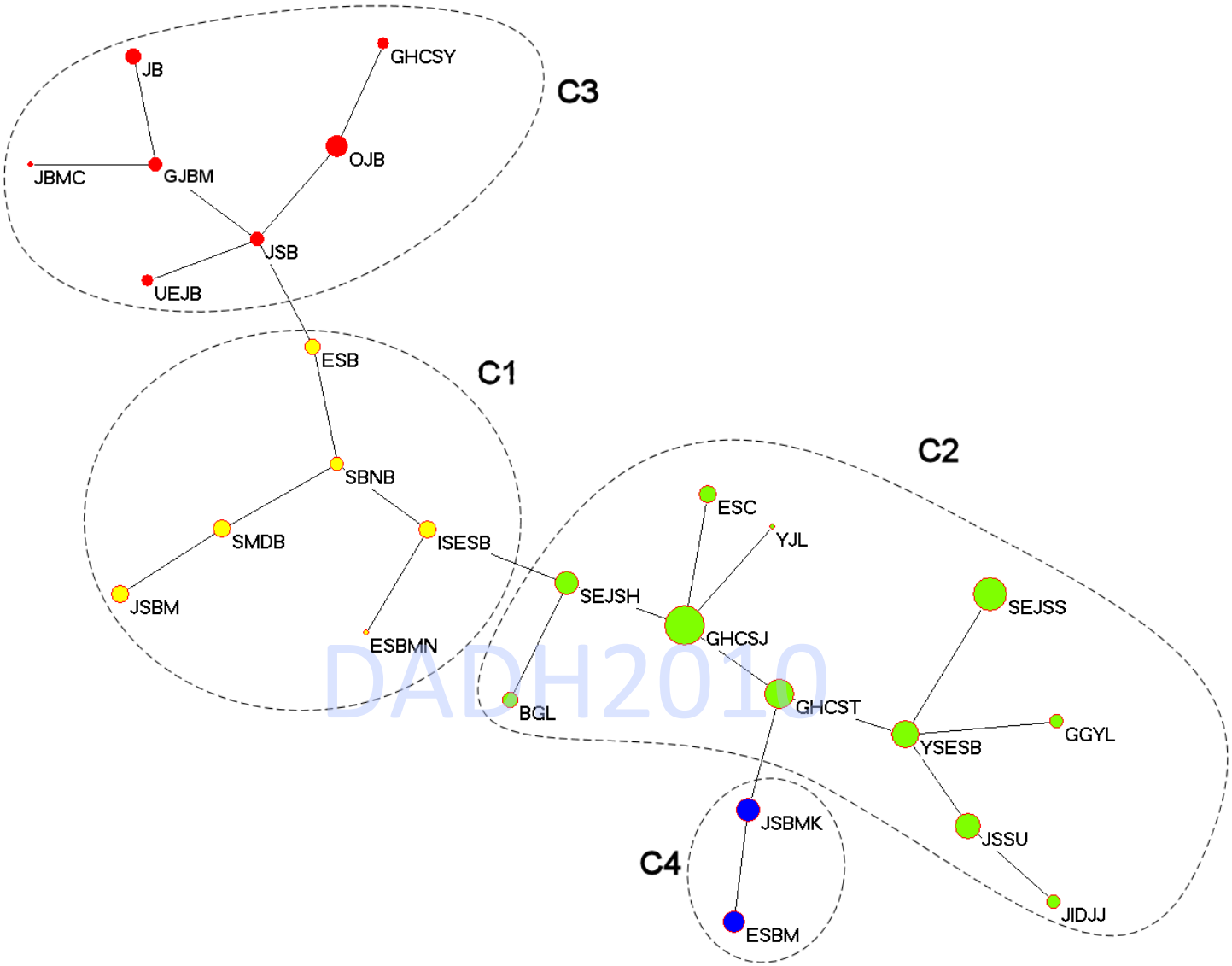
Fundamentals

- Texts containing similar set of cooking methods are similar texts
- Group 26 texts using the distribution of 37 cooking methods

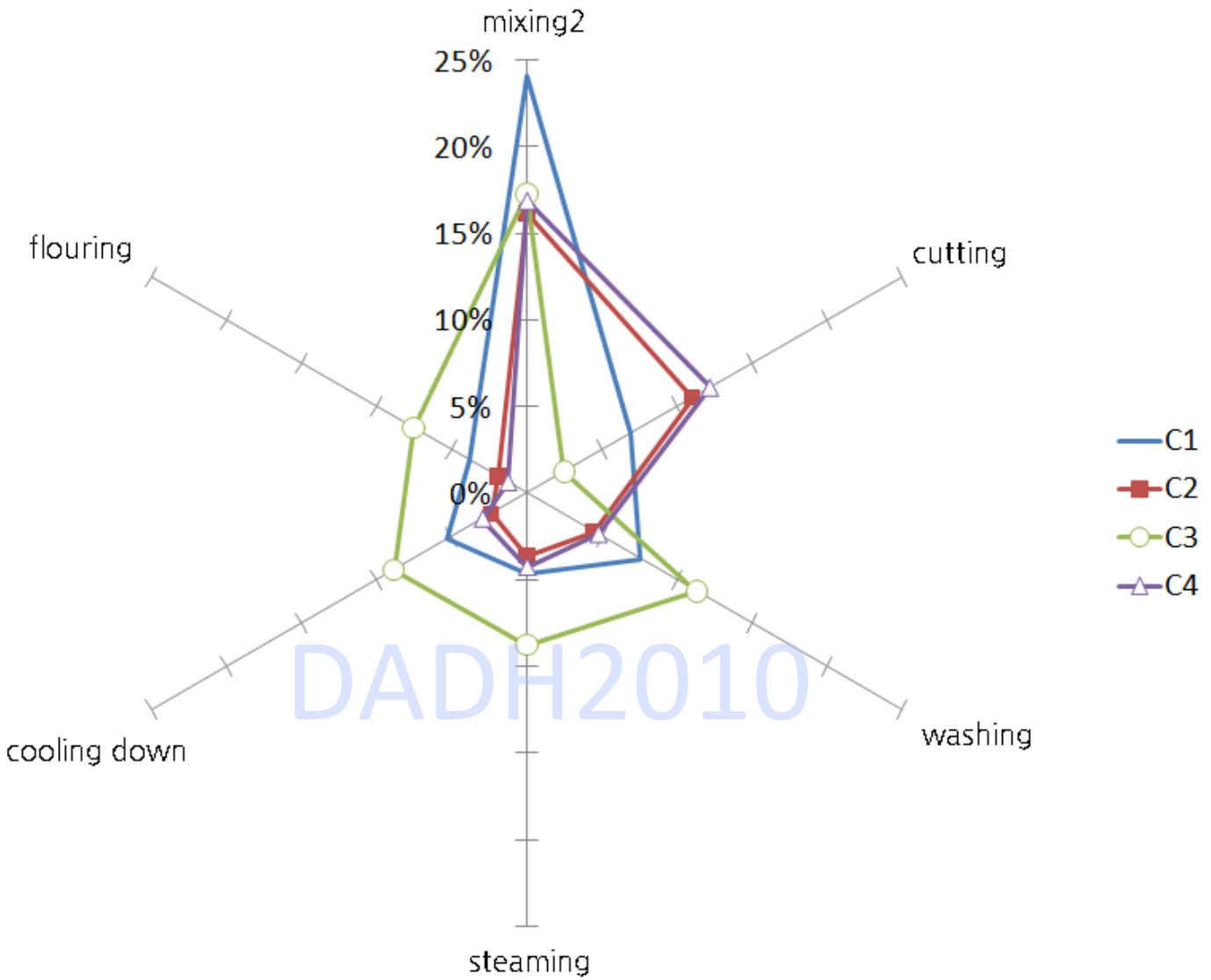
Procedure

1. Generate 26 text vectors containing occurence counts of 37 cooking methods
2. Compute cosine similarities between vectors
3. Construct a Pathfinder network and apply the PNCC clustering
4. Examine the distribution of key cooking methods

Text Similarity Analysis: Cooking Methods II



Text Similarity Analysis: Cooking Methods III



Text Similarity Analysis: Cooking Methods IV

Results for Cluster Analysis

- C1: general cooking methods and brewing methods
- C2: dominantly general cooking methods
- C3: dominantly brewing methods
- C4: only two texts, unclear
- Roughly equivalent to the previous clustering result except JIDJJ

Results for Methods Distribution Examination

- 6 key cooking methods: MIXING^2, CUTTING, WASHING, STEAMING, COLLING DOWN, FLOURING
- C1: high proportion of MIXING^2 $\rightarrow$ main dishes
- C2, C4: high proportion of CUTTING $\rightarrow$ side dishes
- C3: high proportions of WASHING, STEAMING, COLLING DOWN, and FLOURING $\rightarrow$ alcoholic beverages

Cluster Analysis: Ingredients I

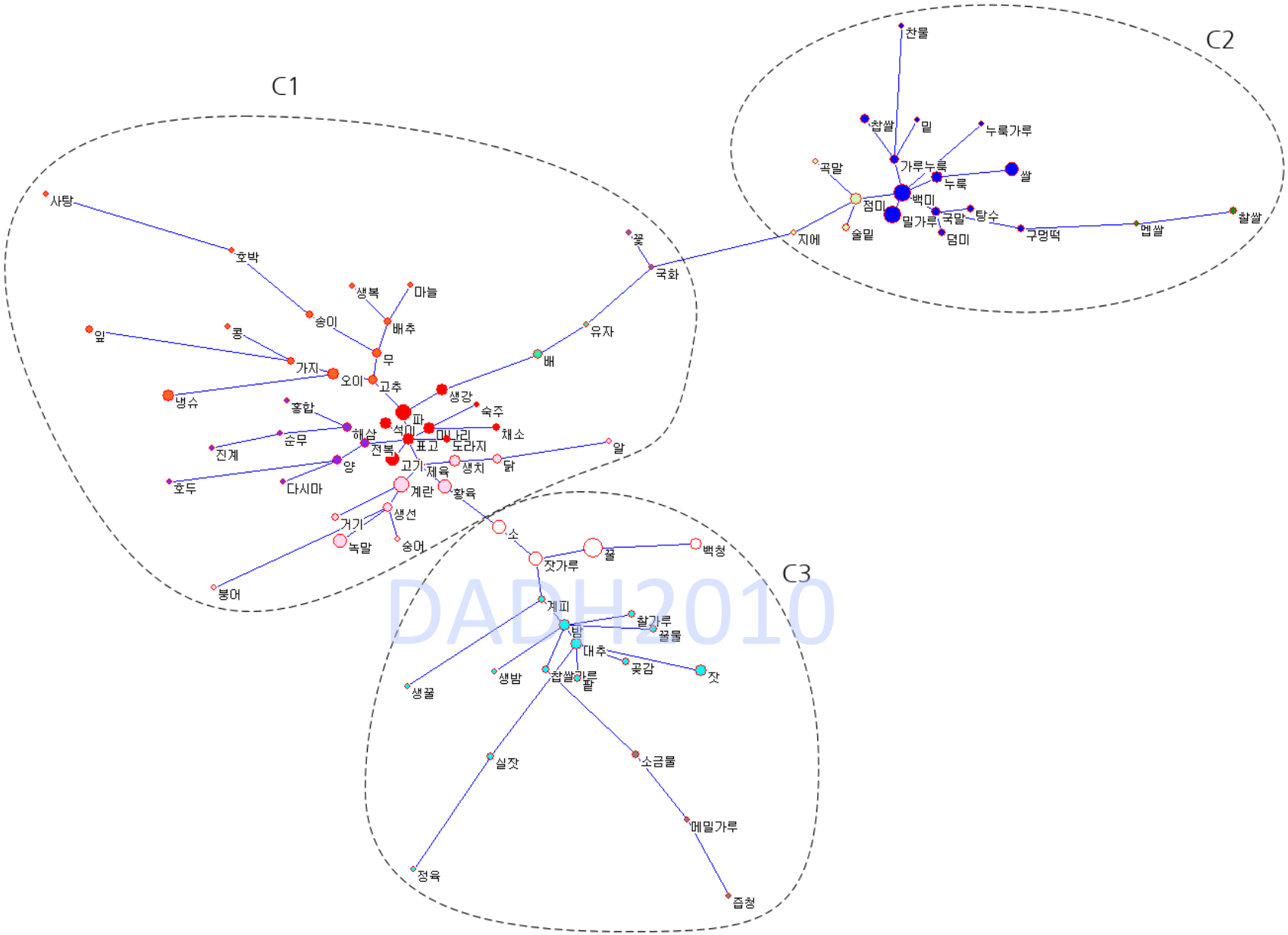
Fundamentals

- Ingredients frequently co-occurring in multiple recipes are closely related
- Examine the distribution of 84 frequent ingredients (excl. 'water') in 1,598 recipes

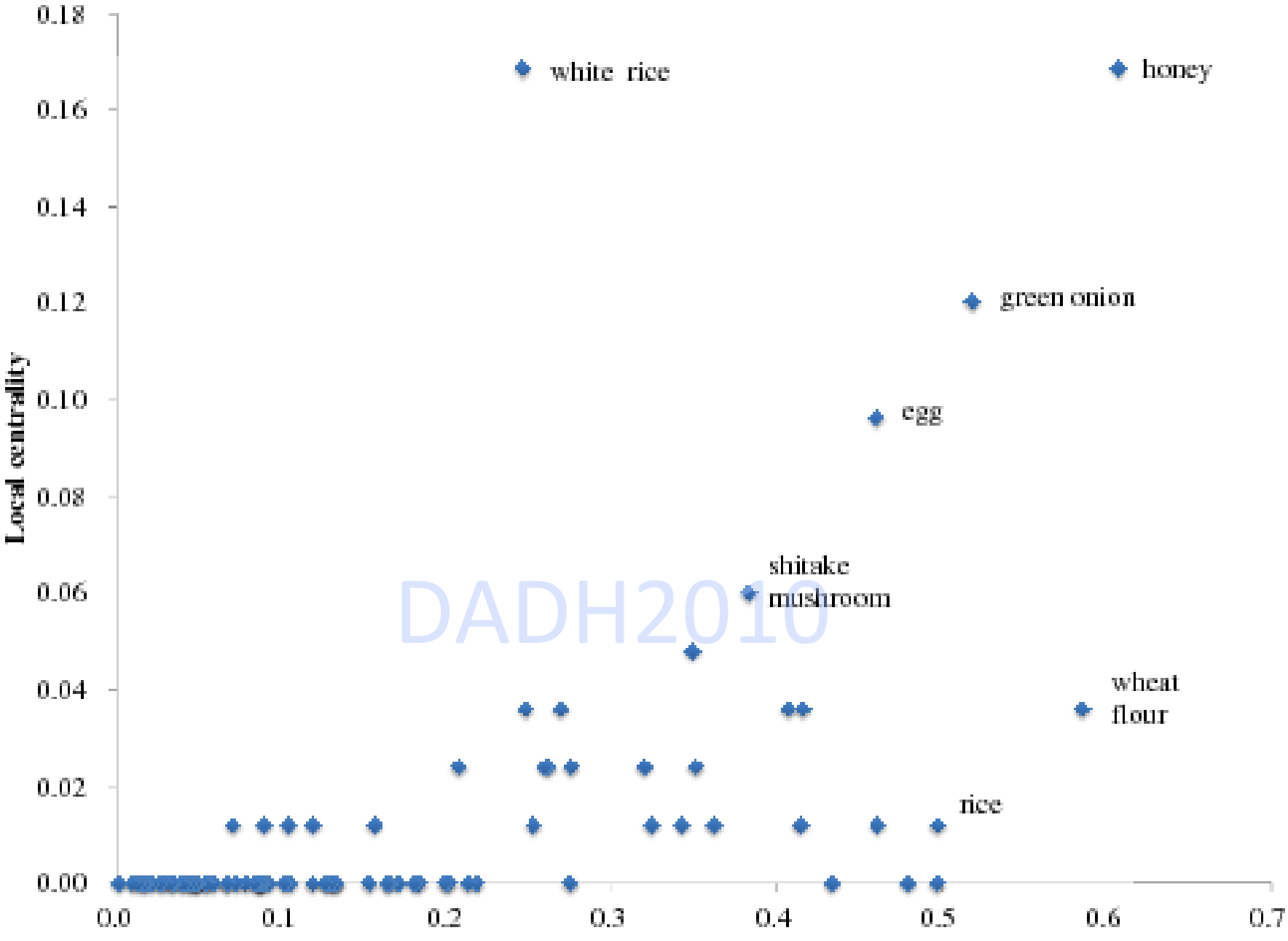
Procedure

1. Form a matrix of 84 ingredients and 1,598 recipes recording occurrence counts of the ingredients
2. Normalize the counts using the Jaccard coefficient
3. Construct a matrix of a Pearson correlation coefficients of the 84 ingredients
4. Generate a Pathfinder network and apply PNCC clustering
5. Apply centrality analysis on the network

Cluster Analysis: Ingredients II



Cluster Analysis: Ingredients III



Cluster Analysis: Ingredients IV

Results: Cluster Analysis

- C1: main ingredients — ‘meat’, ‘fish’, ‘pork’, ‘green onion’, ‘red pepper’
- C2: alcoholic beverage ingredients — ‘white rice’, ‘malt’
- C3: sweet ingredients — ‘chestnut’, ‘jujube’, ‘cinnamon’
- C1-C2: ‘flower’, ‘chrysanthemum’
- C2-C3: ‘stuffing’

Results: Centrality Analysis

- ‘white rice’: high local centrality, low global centrality
  $\rightarrow$ used in limited recipes, mostly alcoholic beverages
- ‘wheat flour’: high global centrality, low local centrality
  $\rightarrow$ used as a background ingredient in many recipes
- ‘honey’: clearly distinguishes recipes of sweets

Text Similarity Analysis: Ingredients I

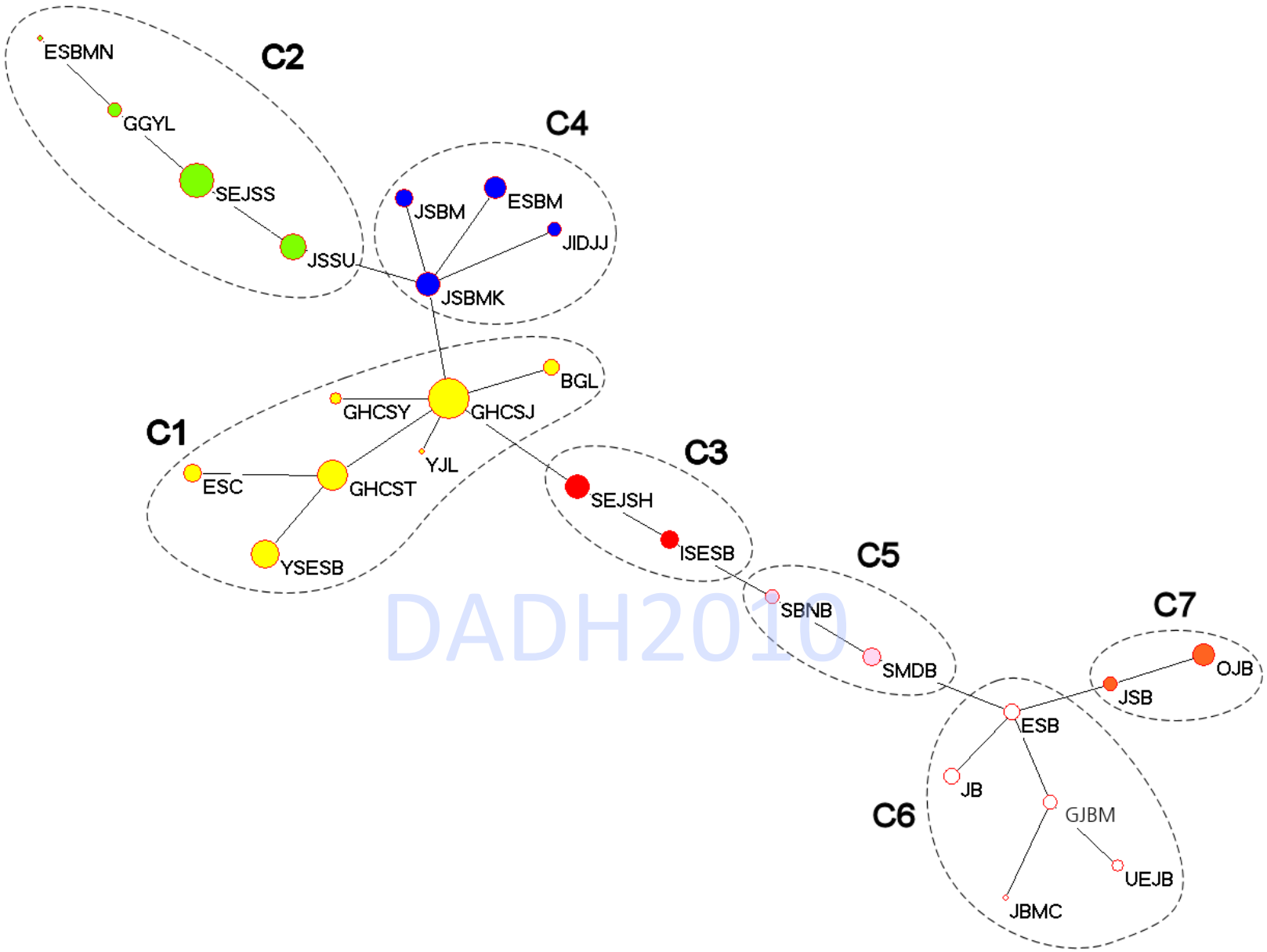
Fundamentals

- Texts containing similar set of ingredients are similar texts
- Group 26 texts using the distribution of 85 ingredients

Procedure

1. Generate 26 text vectors containing proportions of 85 ingredients
2. Compute cosine similarities between vectors
3. Construct a Pathfinder network and apply the PNCC clustering

Text Similarity Analysis: Ingredients II



Text Similarity Analysis: Ingredients III

Results

- C2: dominantly general food recipes
- C1, C4: mostly general food recipes, some brewing methods
- C3, C5: proportion of alcoholic beverage recipes increase
- C6, C7: dominantly alcoholic beverage recipes
- Similar result to the previous similarity analysis result: main axis — YSESB, GHCST, GHCSJ, SEJSH, ISESB, SBNB
- GHCSY moved close to GHCST and GHCSJ

Discussion and Conclusions I

Summary

- Cooking methods and ingredients effectively exhibited properties of culinary texts
- Network analayses revealed similarities of culinary texts
- Centrality analysis identified key cooking methods and ingredients

Usefulness of Network Analysis

- Word/tag occurrence can be represented in a network
- Cluster and centrality analysis can be applied to a network
- Extensive network analysis methods in social network analysis can also be adapted

Discussion and Conclusions II

Importance of High Quality Data Construction

- Annotations are useful, but sometimes not usable
- Standardization/translation is essential for analysis of historic texts
- Results can be fed back to improve consistency of the annotations

Future Directions

- Similarity analysis on individual recipes according to cooking methods and ingredients
- Incorporate the occurring order of cooking methods and ingredients
- Co-occurrence analysis on cooking methods and ingredients: simple and grammatical
- Collaboration with food historians and old Korean experts