The interrelationship of knowledge structure across language groups in communal data sets

Jisung Yoon[1,2], Jinhyuk Yun[3] and Woo-Sung Jung[1,4,5]

1. Dept. of Industrial and Management Engineering, POSTECH, Pohang, Korea
2. School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA
3. Department of Scientometric Research, Korea Institute of Science and Technology Information(KISTI), Korea
4. Dept. of Physics, POSTECH, Pohang, Korea
5. Asia Pacific Center for Theoretical Physics(APCTP), Korea

-> Future Technology Analysis Center

Conference on Complex System 2019
Oct 1st, 2019

POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Table of Contents

1. 발표자료가 전체적으로 불친절
→ 발표자료로 어느정도 이해될 수 있도록

2. 관사, 단복수 한번 더 체크 …

Introduction: Knowledge Structure

- Research background
- Research question

Data and methods

- Wikipedia data: communal data set
- Calculate similarity between knowledge structure

Results

- Community detection results
- Factor analysis

Summary

Knowledge is a familiarity, awareness, or understanding of someone or something

which is acquired through **experience** or **education** by perceiving, discovering, or learning.

-Wikipedia

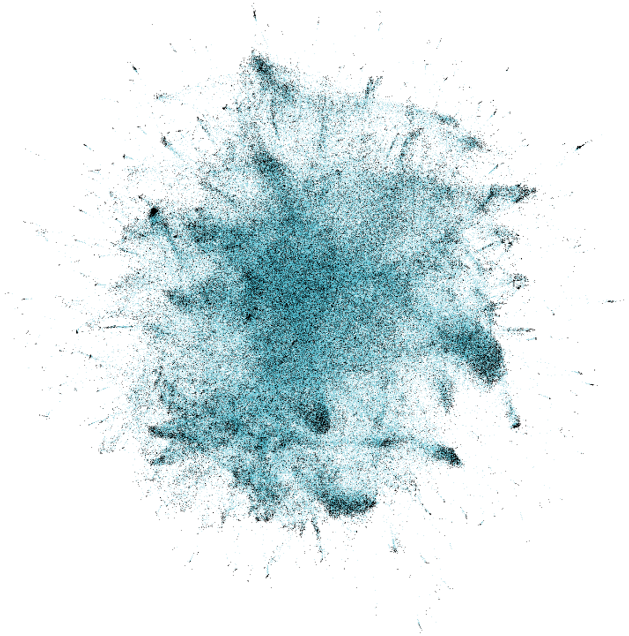
Human understanding is root of the general laws of nature that organize all experience

칸트 사진이라도(?) -Immanuel Kant

Knowledge Structure

보는 방식은 3(~)

- Knowledge structure can be varied by **personality**, **living country** or **linguistic profile** based on the social structure and education system



Language Affect Knowledge

- Language Socialization (Schieffelin, B. B., & Ochs, E.,1986)
  - Socialization(acquiring knowledge) through the use of language
  - Schieffelin, B. B., & Ochs, E. (1986). Language socialization. Annual review of anthropology, 15(1), 163-191.

- Language and Knowledge (Code, L., 1980)
  - Language and knowledge are mutually influential
  - Perception and knowledge are organized by language from the flux of sensory experience.

Research Questions

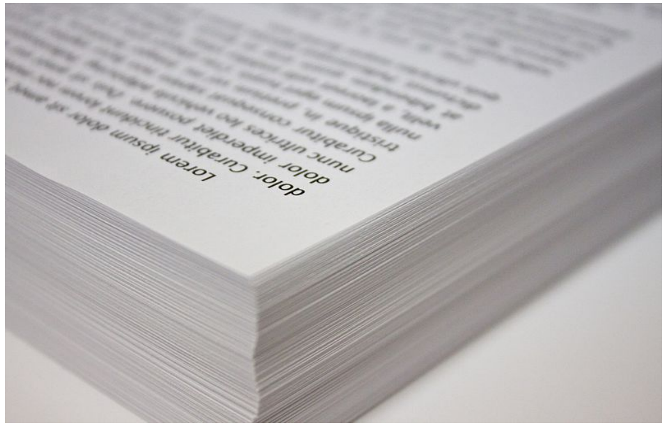
- What are major factors influencing the similarity of knowledge structure across the language group?

1. How can we construct a knowledge structure of a language group?

2. How can we ~~compute~~ derive similarity among the obtained knowledge structure?

3. What are major factors influencing the similarity?

Knowledge Database



The United States of America

The Commissioner of
Patents and Trademarks

Has received an application for a patent for a new and useful invention. The title and description of the invention are enclosed. The requirements of law have been complied with, and it has been determined that a patent on the invention shall be granted under the law.

Therefore, this 5,860,492

United States Patent

Grants to the person(s) having title to this patent the right to exclude others from making, using, offering for sale, or selling the invention throughout the United States of America or importing the invention into the United States of America for the term set forth below, subject to the payment of maintenance fees as provided by law.

If this application was filed prior to June 8, 1995, the term of this patent is the longer of seventeen years from the date of grant of this patent or twenty years from the earliest effective U.S. filing date of the application, subject to any statutory extension.

If this application was filed on or after June 8, 1995, the term of this patent is twenty years from the U.S. filing date, subject to any statutory extension. If the application contains a specific reference to an earlier filed application or applications under 35 U.S.C. 120, 121 or 365(c), the term of the patent is twenty years from the date on which the earliest application was filed, subject to any statutory extension.

Q. Todd Dickinson

Acting Commissioner of Patents and Trademarks

Attest Marjorie V. Turner

a

Not proper to construct knowledge structure of specific language group

Communal data set - Wikipedia

- Internet encyclopedia is edited by users that use specific language
  - Result of a collective intelligence
- 294 active language editions (April, 2019)
- Possible to get **knowledge structure** of a **specific language group**

WIKIPEDIA

Data

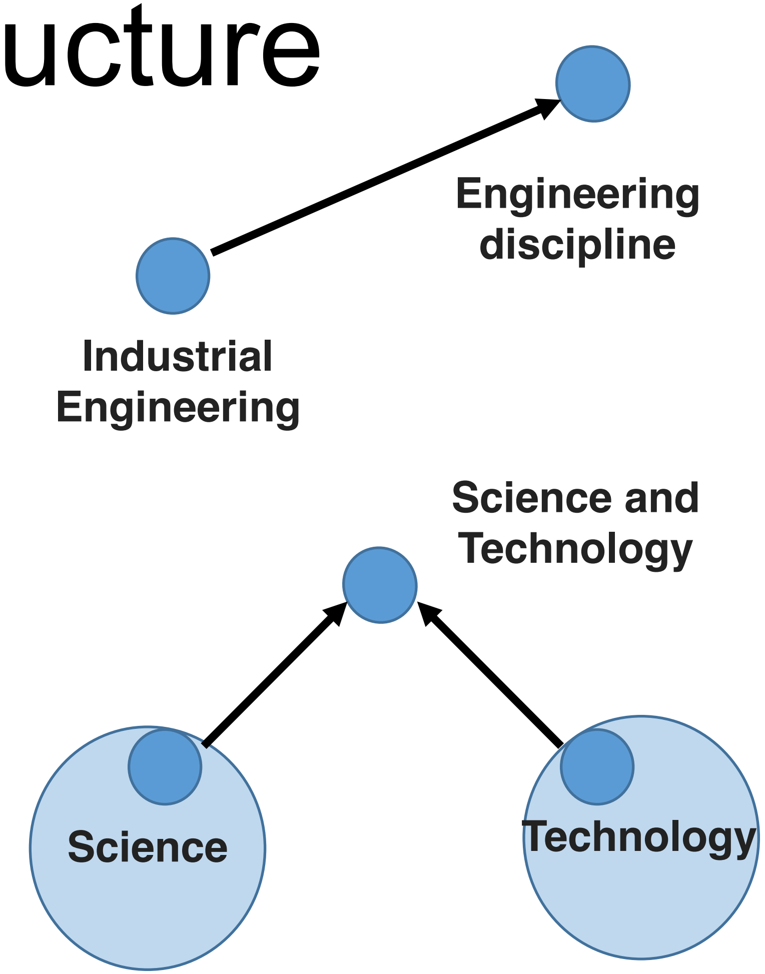
at(?)

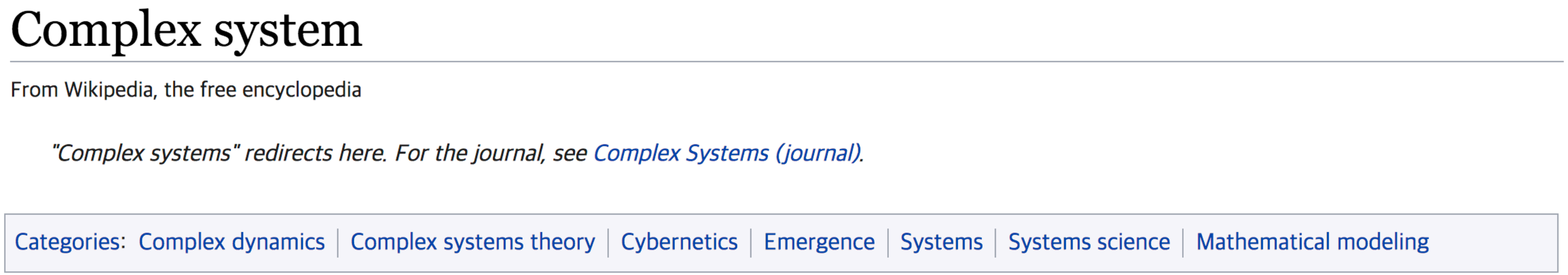
- Dump data of 59 different language editions of Wikipedia on August 20, 2018
  - Category-link data set: relation between a category and other items
    - For constructing a knowledge network
  - Language-link data set: bridge data between another language editions of Wikipedia items that same meaning
    - For comparing knowledge structure of different language edition

Construct Knowledge Structure

- Knowledge Network
  - **One for each language**
  - **Node**: each category or page
  - **Link**: directed If node A refer node B, node A -> node B

  - Sub-network with artificial root assigned as a common parent node “**Science**” and “**Technology**”
    - To get fine-grained form of knowledge network
    - Science covers all branch of science
      - Applied sciences, Formal sciences, Natural Sciences, Social sciences





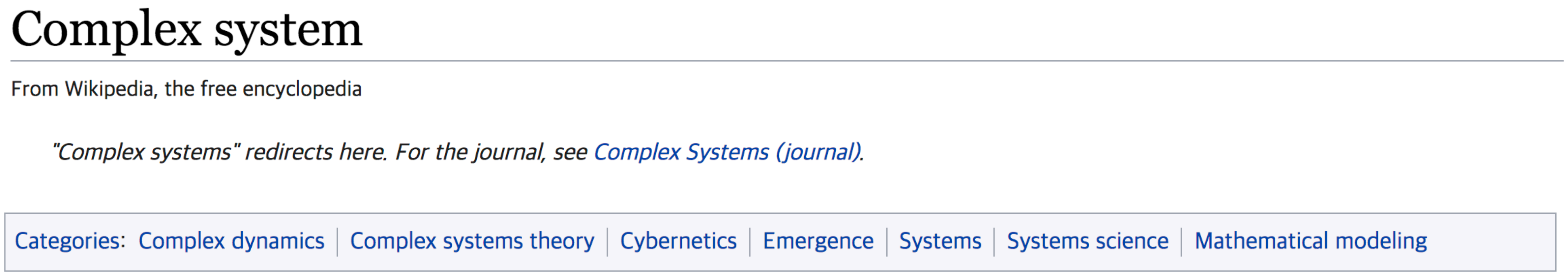
Complex system

From Wikipedia, the free encyclopedia

"Complex systems" redirects here. For the journal, see Complex Systems (journal).

Categories: Complex dynamics | Complex systems theory | Cybernetics | Emergence | Systems | Systems science | Mathematical modeling

Construct Knowledge Structure

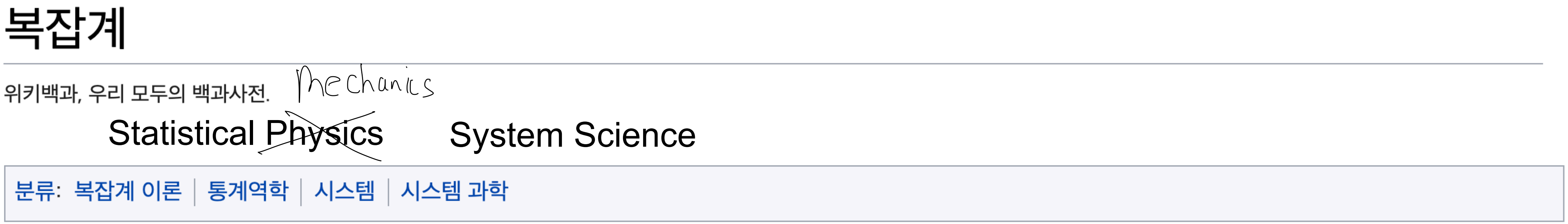


Complex system

From Wikipedia, the free encyclopedia

"Complex systems" redirects here. For the journal, see Complex Systems (journal).

Categories: Complex dynamics | Complex systems theory | Cybernetics | Emergence | Systems | Systems science | Mathematical modeling



복잡계

위키백과, 우리 모두의 백과사전.

mechanics

Statistical ~~Physics~~ System Science

분류: 복잡계 이론 | 통계역학 | 시스템 | 시스템 과학

Complex system Theory

System

Similar, but slightly different

Then, how can we calculate similarity between knowledge structure?

Calculate knowledge structure similarity

→ mapping(?) / pairing(?)

→ translate는 조금 안 맞는듯...

- Calculate subject similarity first.
  - **Characterize** with genealogy vector, and **Translate** to target language with **language-link data set,** and **compare!**
    - **[Characterize]** Genealogy vector of a given node as a Personalized Page Rank (Jeh, G., & Widom, 2003) of a subject in network.
    - **[Translate]** Matching with language link data set
      - E.g.) Republic of Korea (en) => 한국 (ko)
    - **[Compare]** Calculate between translated genealogy vector and target genealogy vector
      - We use 1 - Euclidian distance as similarity

- Then, knowledge structure similarity is average value of all subject similarity

→ We define the similarity between two knowledge structure as the average value of all subject similarity between the languages.

Calculate knowledge structure similarity

- For example, *Complex System* and *복잡계* (English to Korean)
  - **Characterize**

$$X_{\mathrm{Complex\ System}} = [\frac{1}{5},\frac{1}{5},\frac{1}{5},0,\frac{1}{5},\frac{1}{5},\dots] \in \mathrm{R}^{\mathrm{N_E}}$$

$$Y_{복잡계} = \left[\frac{1}{4},\frac{1}{4},0,\frac{1}{4},\frac{1}{4},\dots\right] \in R^{N_k}$$

  - **Translate**

$$Y_{Complex\ System} = \left[\frac{1}{5},\frac{1}{5},0,0,\frac{1}{5},\frac{1}{5},\dots\right] \in \mathrm{R}^{\mathrm{N_K}}$$

  - **Compare**

$$S_{Complex\ System\ -복잡계} = 1 - d\left(Y_{Complex\ System}, Y_{복잡계}\right)$$

- Knowledge structure similarity English to Korean can calculate with averaging over all the subject.

↳ Structural similarity of knowledge from

(예) 클러스터 이름 색을 클러스터 색과 같게 or ● Cluster 1 형태의 회색 가능하게

(예) 잘 와닿지 않는 이름 ... → intercontinental (?)

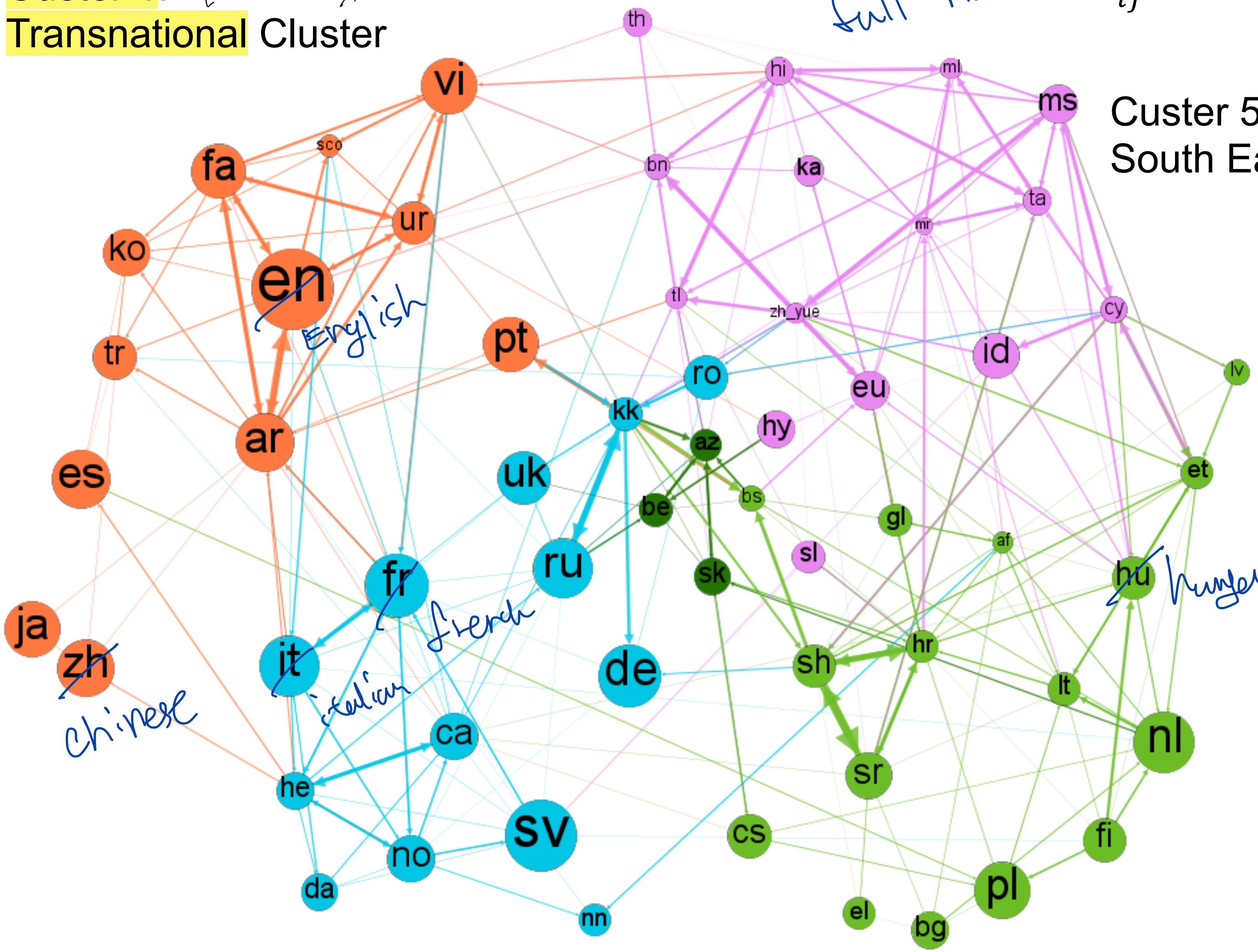
Custer 1:
Transnational Cluster

* 중요한 멀게 없는 → 클러스터별 멀게 정도 ... full name으로

$$r_{ij} = Strength_{total} * \frac{s_{ij}}{strength_{out}\ of\ i\ * strngth_{in}\ of\ j}$$

역사학자의 comment도 좋은 것 같음..

이런 부분들 설명을 위해
↓ Ancient History를 조금 가져와 보기 (?)
기원 | 보고서 | 기원없이 | 조언 등등..



Custer 5:
South East Asia Cluster

Custer 4:
Northeastern Europe Cluster

Northern / Eastern

↳ 먼나라 이웃나라 (?) ㅋㅋ

Custer 2:
Western Europe Cluster

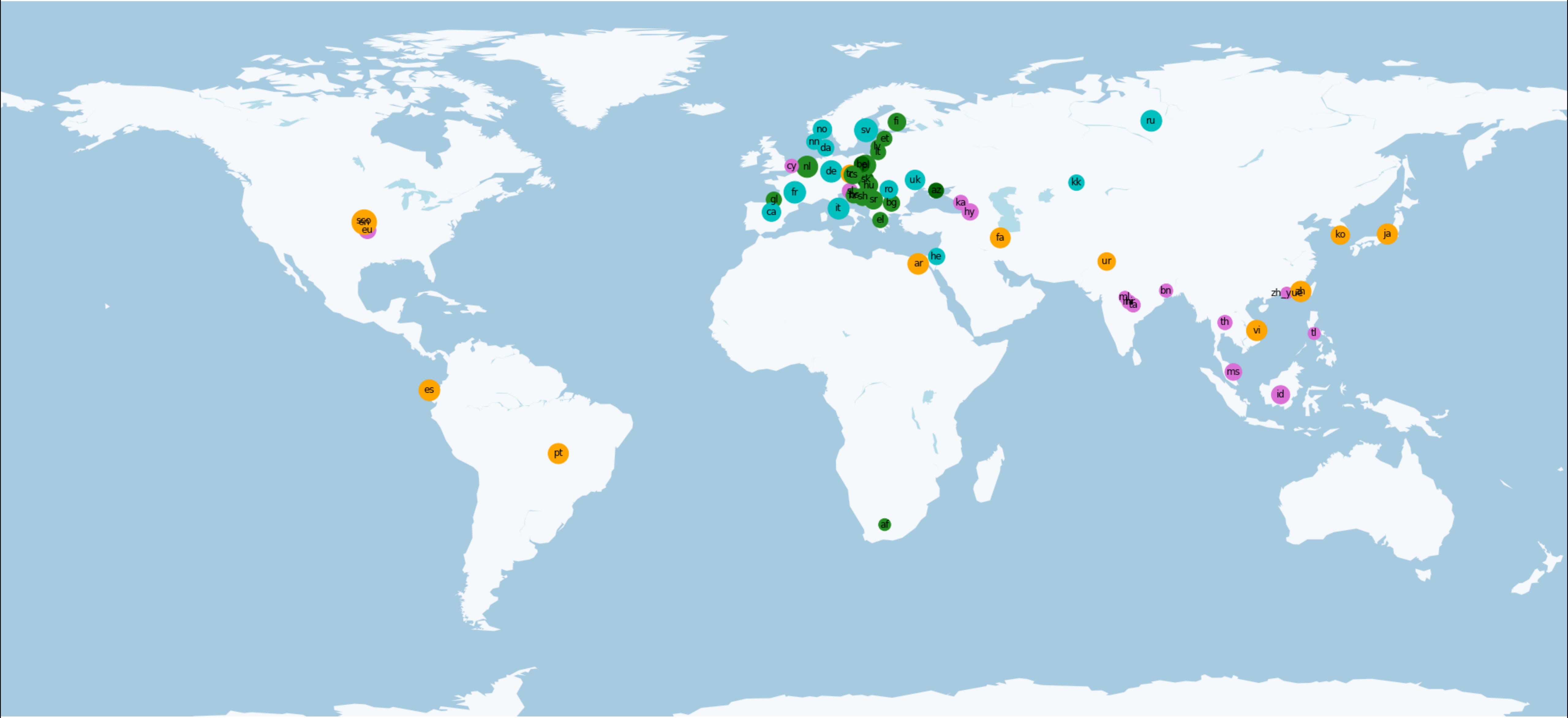
Custer 3:
Eastern Europe Cluster

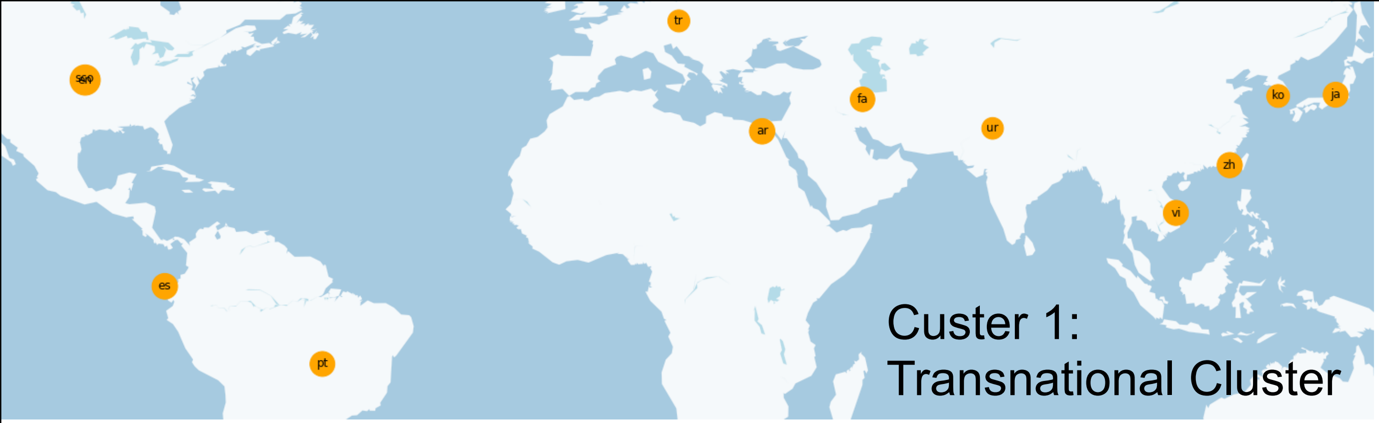
Threshold = 1.04

Resolution parameter = 1

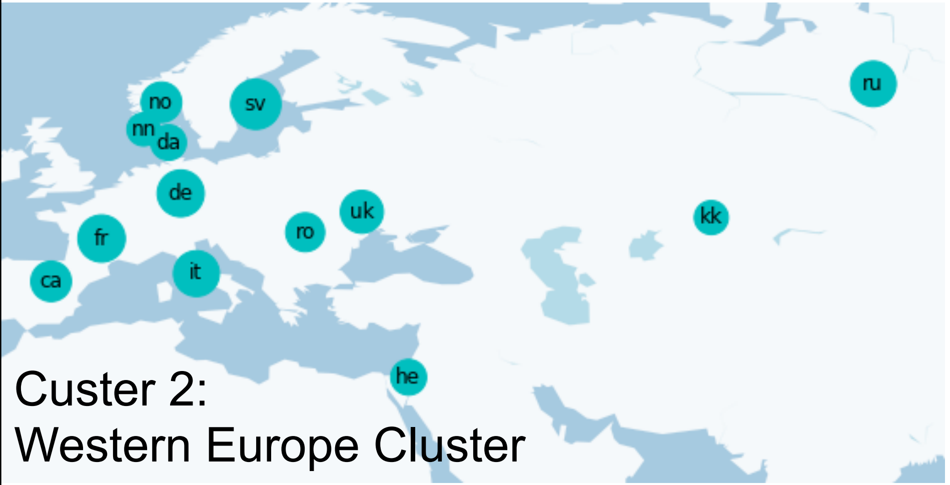
Modularity = 0.42

Community result on map

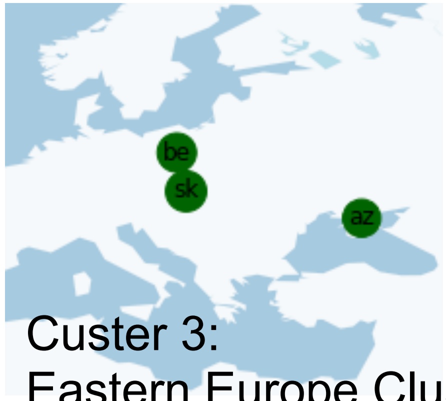




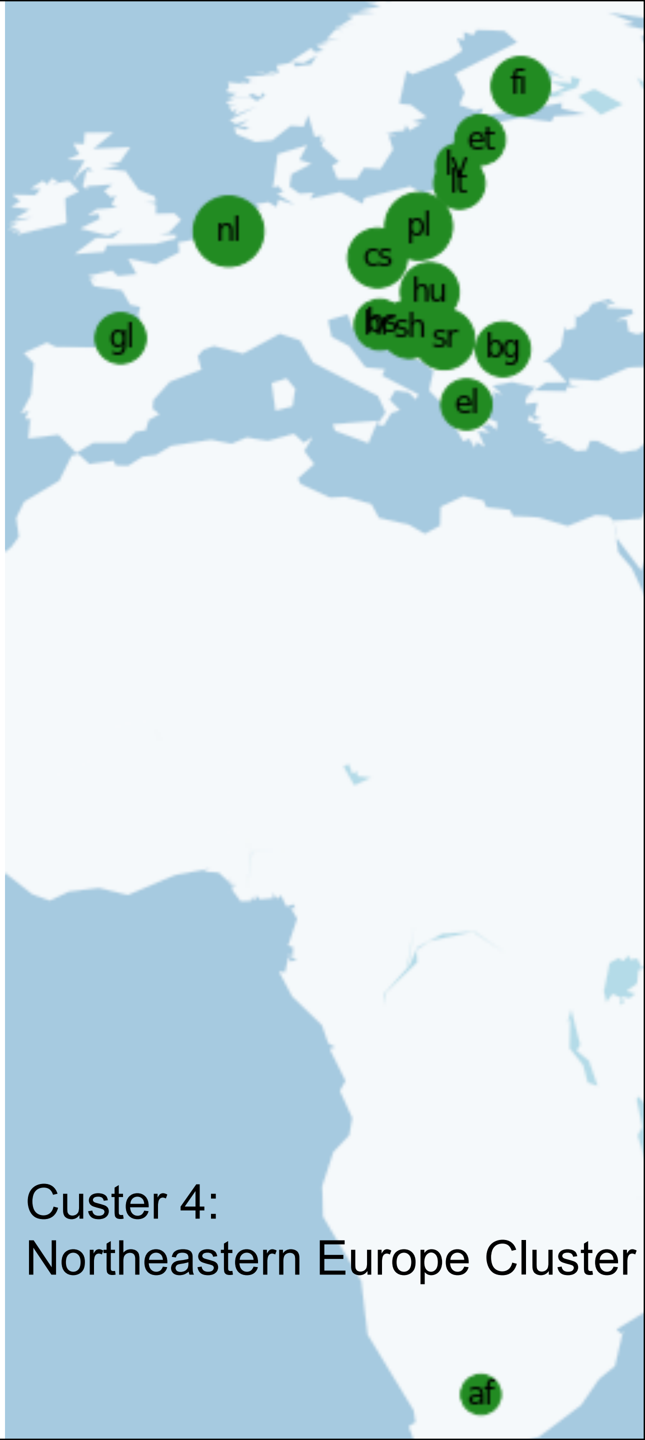
Custer 1:
Transnational Cluster



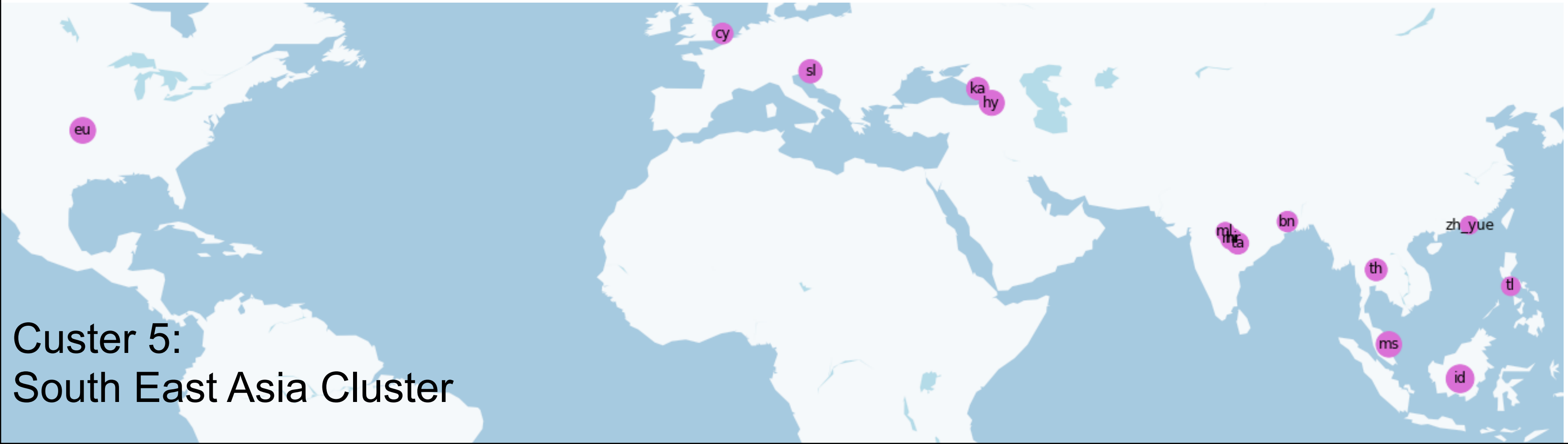
Custer 2:
Western Europe Cluster



Custer 3:
Eastern Europe Cluster

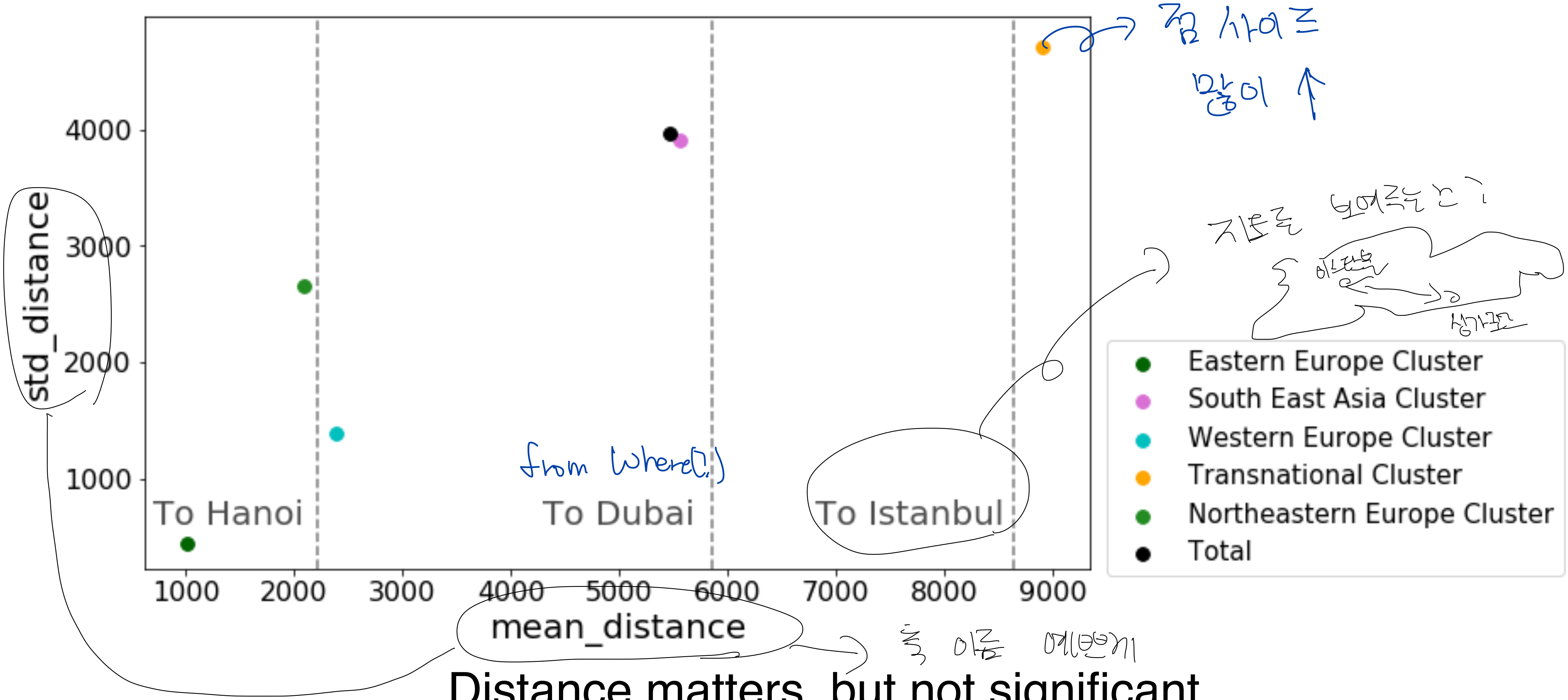


Custer 4:
Northeastern Europe Cluster



Custer 5:
South East Asia Cluster

Community result on map



점 사이즈 많이 ↑

지도로 보여주는 건?

from Where(?)

축 이름 예쁘게

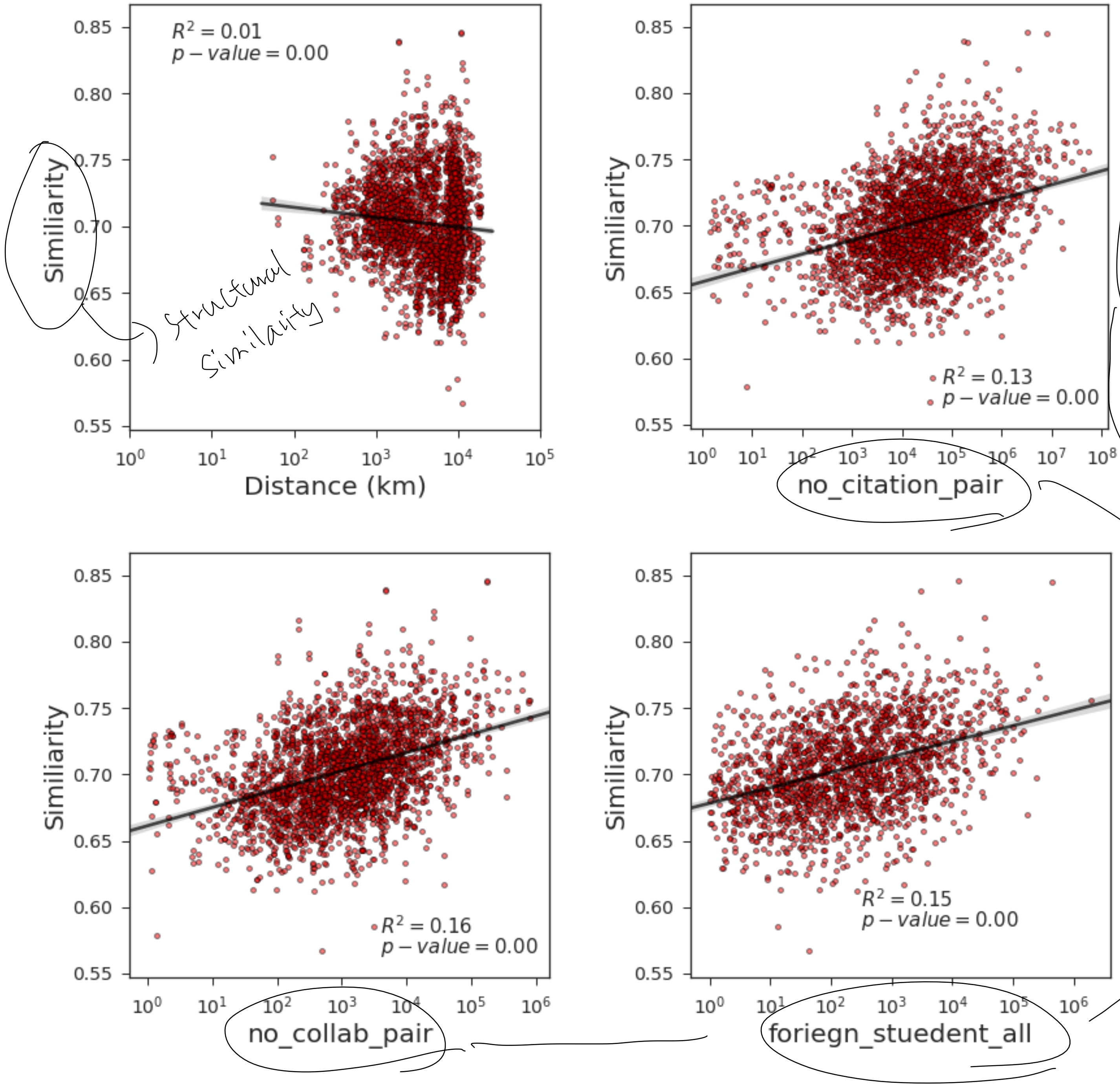
Distance matters, but not significant

Factor Analysis

- 4 Factors to analysis
  1. **Physical Distance** - distance between language
  2. **Weak knowledge transfer*** – number of citation (paper)
  3. **Strong knowledge transfer*** – number of collaboration (paper)
  4. **Soft Power Mobility*** – number of foreign students

↳ Movement of the soft power.

* These data are from SCOPUS and OECD. Basically, there are county to country data. We projected to language to language dimension with country to language data set (Ronen et al, 2014).



Negative

Physical Distance

Positive

Weak knowledge Transfer

Strong Knowledge Transfer

Soft Power Movement

이건 당연한 부분

아래의 3개는 associativeness에

축이름 고치기

Summary

- We use Wikipedia dump data of 59 different language editions and construct the Knowledge Network to compare the knowledge structure between languages.
- We found 5 geo-locational clusters, but physical distance is not a a significant factor for some clusters.
- We conduct factor analysis to identify knowledge structure similarity between languages, and find ~~a pattern~~ correlations for :
    1. **Physical Distance** - distance between language
    2. **Weak knowledge transfer*** – number of citation (paper)
    3. **Soft Power Mobility*** – number of foreign students
    4. **Strong knowledge transfer*** – number of collaboration (paper)
- It helps to understand fact that knowledge structure has been affected by language groups.

↳ 나만 여기 폰트 이상해 보이나요...

To do에 대해 언급하셔도 ...

References

- Schieffelin, B. B., & Ochs, E. (1986). Language socialization. Annual review of anthropology, 15(1), 163-191.
- Code, L. (1980). Language and knowledge. word, 31(3), 245-258.
- Jeh, G., & Widom, J. (2003, May). Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web* (pp. 271-279). Acm
- Ronen, S., Gonçalves, B., Hu, K. Z., Vespignani, A., Pinker, S., & Hidalgo, C. A. (2014). Links that speak: The global language network and its association with global fame. *Proceedings of the National Academy of Sciences*, *111*(52), E5616-E5622

Thank you for attention

Jisung.yoon@postech.ac.kr

Appendix 1. Literatures on Wikipedia

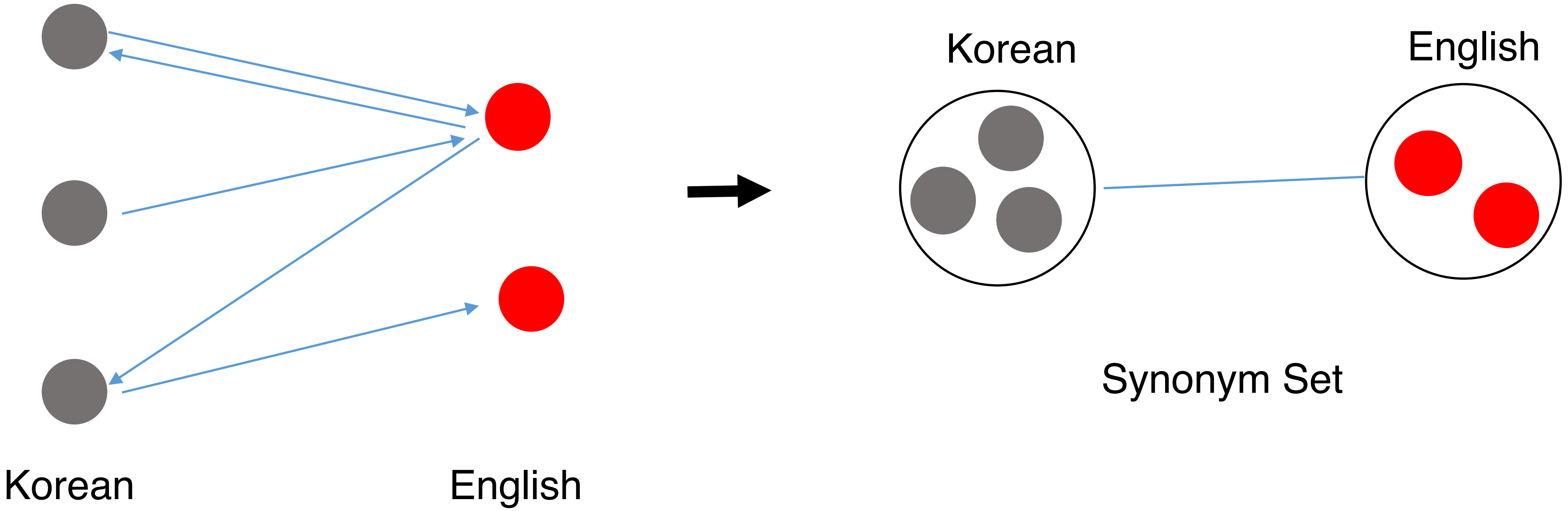
- Dynamics on editing Wikipedia
  - Dynamics and pattern of modification (Yasseri et al., 2012a; Yasseri et al., 2012b),
  - Mechanistic model for intellectual interchanges (Yun et al., 2016)

- Credibility of Wikipedia data
  - TBA

- Data analysis with Wikipedia data
  - Extracting knowledge structure of Wikipedia (Ponzetto andNavigli;2009;Gabella,2017)
  - Clustering of languages across the wikipedia growth (Ban, 2017)

Appendix 2. Matching with language link data set

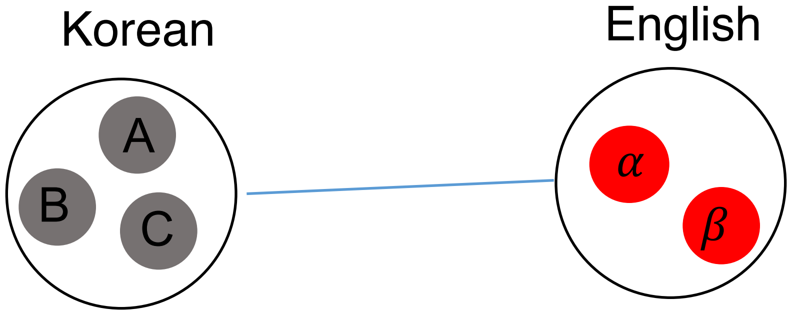
- For more general case (Many to Many)
  - Pairwise between two different language editions
  - **Node**: each category or page
  - **Link**: directed If A is connected as same documents B, A->B
    - E.g.) Republic of Korea (en) => 한국(ko)
  - Remove direction and merge after construction (likes synonym set)



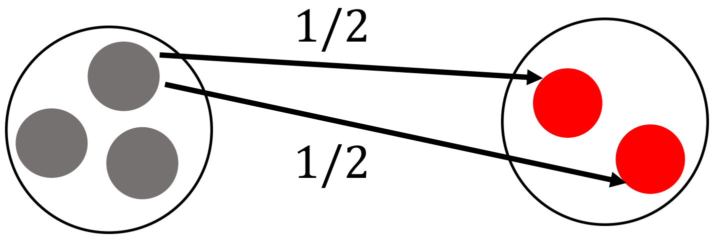
Appendix 3. Calculate subject distance

- For general case, many to many

Korean genealogy vector of node A, $\mathrm{X}_A \in \boldsymbol{R}^{N_k}$
English genealogy vector of node B, $\mathrm{Y}_\alpha \in \boldsymbol{R}^{N_E}$
Transition matrix, $\boldsymbol{T}_{K\rightarrow E} \in \boldsymbol{R}^{N_k * N_E}$



Synonym Set 1

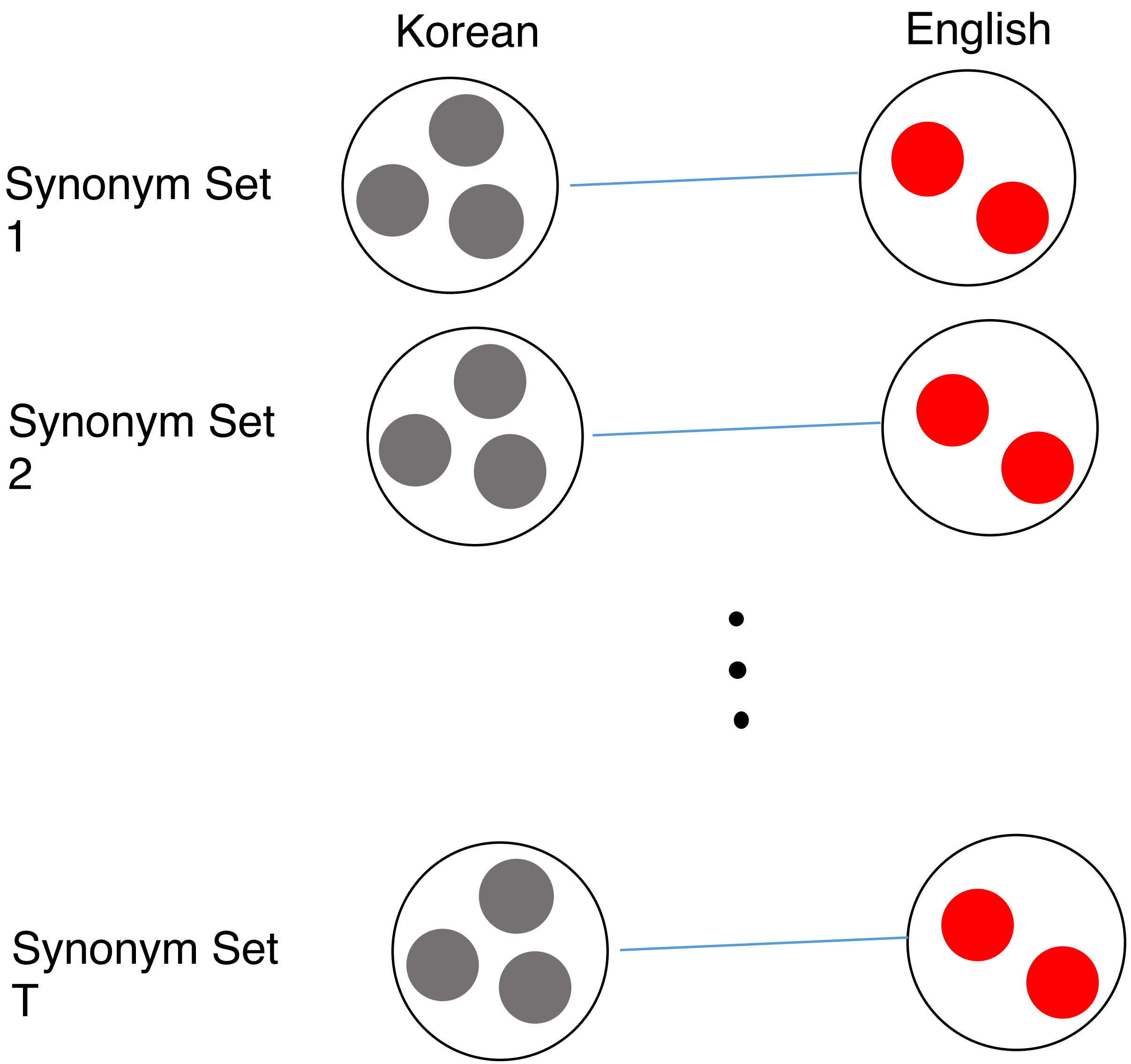


$$d_{A\alpha}^{\boldsymbol{k}\rightarrow\boldsymbol{E}} = D(X_A * T_{k\rightarrow E}, Y_\alpha \ )$$

Calculate Euclidean distance between English version of Korean genealogy vector and English genealogy vector

$$\boldsymbol{s}_{\boldsymbol{1}}^{\boldsymbol{k}\rightarrow\boldsymbol{E}} = \boldsymbol{1} - \frac{1}{N * M} * \sum_{i=1}^{N} \sum_{j=1}^{M} d_{ij}^{\boldsymbol{k}\rightarrow\boldsymbol{E}}$$

Appendix 4. Calculate overall similarity



Korean

English

Synonym Set 1

Synonym Set 2

Synonym Set T

$$s^{K\rightarrow E} = \frac{1}{T} * \sum_{i=1}^{T} s_i^{K\rightarrow E}$$

$$s^{E\rightarrow K} = \frac{1}{T} * \sum_{i=1}^{T} s_i^{E\rightarrow K}$$

Appendix 5. Extracting Backbone of similarity network

- Extracting Backbone of similarity network
    - Relative similarity

$$r_{ij} = \frac{\frac{s_{ij}}{\sum_j s_{ij}}}{\frac{\sum_j s_{ij}}{\sum_i \sum_j s_{ij}}} = \sum_i \sum_j s_{ij} * \frac{s_{ij}}{\sum_j s_{ij} * \sum_j s_{ij}}$$

$$r_{ij} = Strength_{total} * \frac{s_{ij}}{strength_{out}\ of\ i\ * strngth_{in}\ of\ j}$$

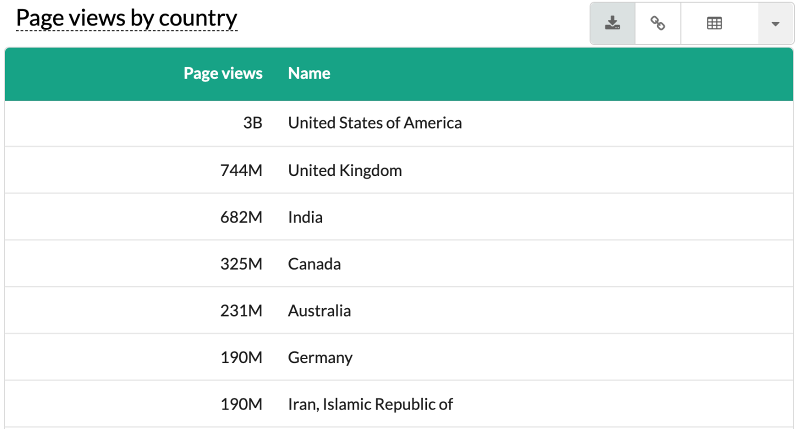
- Select edges that higher than threshold
    - we select 1.04 which network fully connected to one weakly-connected component

Appendix 6. Language Code

Code	Name	Code	Name	Code	Name	Code	Name	Code	Name
af	Afrikaans	**el**	Greek	**hr**	Croatian	**ms**	Malay	**sr**	Serbian
ar	Arabic	**en**	English	**hy**	Armenian	**nl**	Dutch	**sv**	Swedish
az	Azerbaijani	**es**	Spanish	**id**	Indonesian	**nn**	Norwegian nynorsk	**ta**	Tamil
be	Belarussian	**et**	Estonian	**it**	Italian	**no**	Norwegian	**th**	Thai
bg	Bulgarian	**eu**	Basque	**ja**	Japanese	**pl**	Polish	**tl**	Tagalog
bn	Bangla	**fa**	Persian	**ka**	Georgian	**pt**	Portuguese	**tr**	Turkish
bs	Bosnian	**fi**	Finnish	**kk**	Kazakh	**ro**	Romanian	**uk**	Ukrainian
ca	Catalan	**fr**	French	**ko**	Korean	**ru**	Russian	**ur**	Urdu
cs	Czech	**gl**	Galician	**lt**	Lithuanian	**sco**	Scots	**vi**	Vietnamese
cy	Welsh	**he**	Hebrew	**lv**	Latvian	**sh**	Serbo-croatian	**zh**	Chinese
da	Danish	**hi**	Hindi	**ml**	Malayalam	**sk**	Slovak	**zh_yue**	Cantonese
de	German	**hu**	Hungarian	**mr**	Marathi	**sl**	Slovenian		

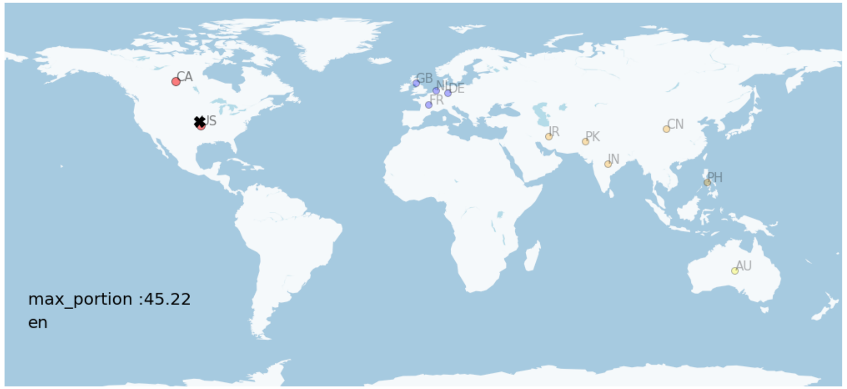
Appendix 7. Location of Language

- Each Wikipedia has a page view by country statistics.
  1. Get centroid locations of each country
  2. Conduct geo-location clustering, and get max portion cluster
     - To reduce noise
  3. Get weighted centroid of max portion cluster



Page views by country

Page views	Name
3B	United States of America
744M	United Kingdom
682M	India
325M	Canada
231M	Australia
190M	Germany
190M	Iran, Islamic Republic of



Appendix 8. Language Projection Method

- Basically, socio-economic data are county to country data.
- For our analysis, we develop a method that projects county to country data to language to language data.
- Language projection method
  - $Y_{l \to L} = A_{L \to C}^{T} * X_{C \to C}{}^{*} A_{L \to C}$, Language projected data
    - $X_{C \to C} \in R^{N_C * N_C}$, Country to country socio-economic data
    - $A_{L \to C} \in R^{N_C * N_L}$, Country to language matching matrix(Ronen et al, 2014)
      - e.g.) South Korea → 100% Korean
      - e.g.) United States → 82.1% English, 10.7% Spanish