Text Mining

Lecture 09

Text Network Analysis

Keungoui Kim

awekim@handong.edu

Handong Global University, 2024-2.

Network Structure

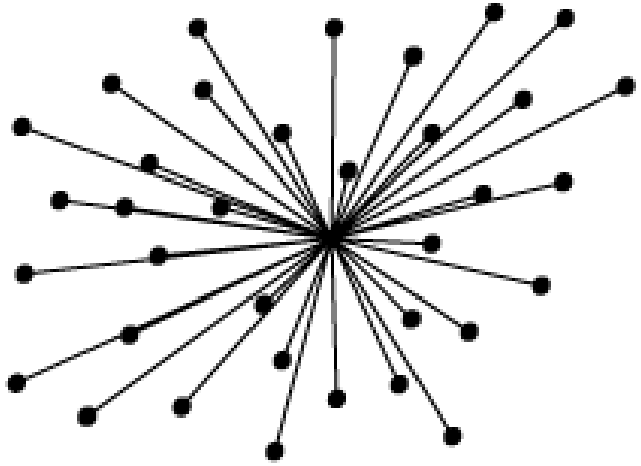
- Society
  - A group of individuals involved in persistent social *interaction*, or a large social group *sharing* the same spatial or social territory, typically subject to the *same* political authority and dominant cultural expectations [Wikipedia]
  - Relationship: the way two or more people are connected, or the way they behave toward each other [Cambridge Dictionary]
  - Father – Daughter, Professor – Student, Employer – Employee, etc.
  - “Man is by nature a social animal”

- When understanding our society, the structure of “relationship” is an important aspect to consider
  - Structure of relationship ~ Network structure

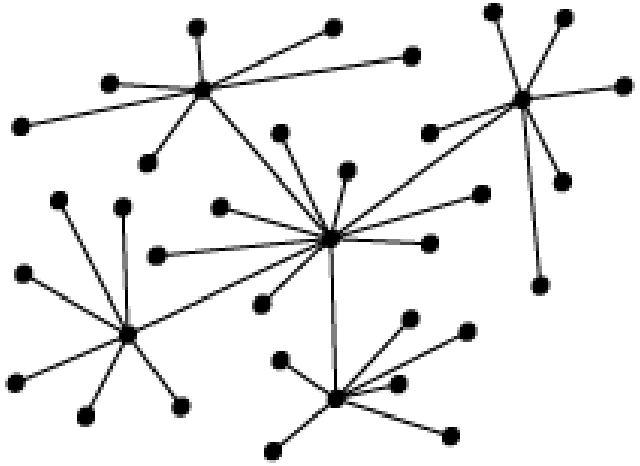
Types of Network Structure

https://blog.jinbo.net/silviogesell/101

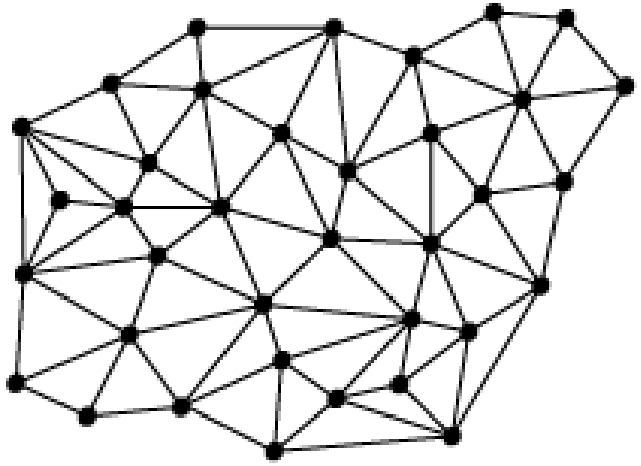
- Three different types of network structure 네트워크구조 3가지
  - Centralized network: A single, centralized server/master node, which handles all major data processing and stores data and user information that other users can access
  - Decentralized network: Distributes information-processing workloads across multiple devices instead of relying on a single central server
  - Distributed network: Forgoes a single centralized master server in favor of multiple network owners



Centralized network



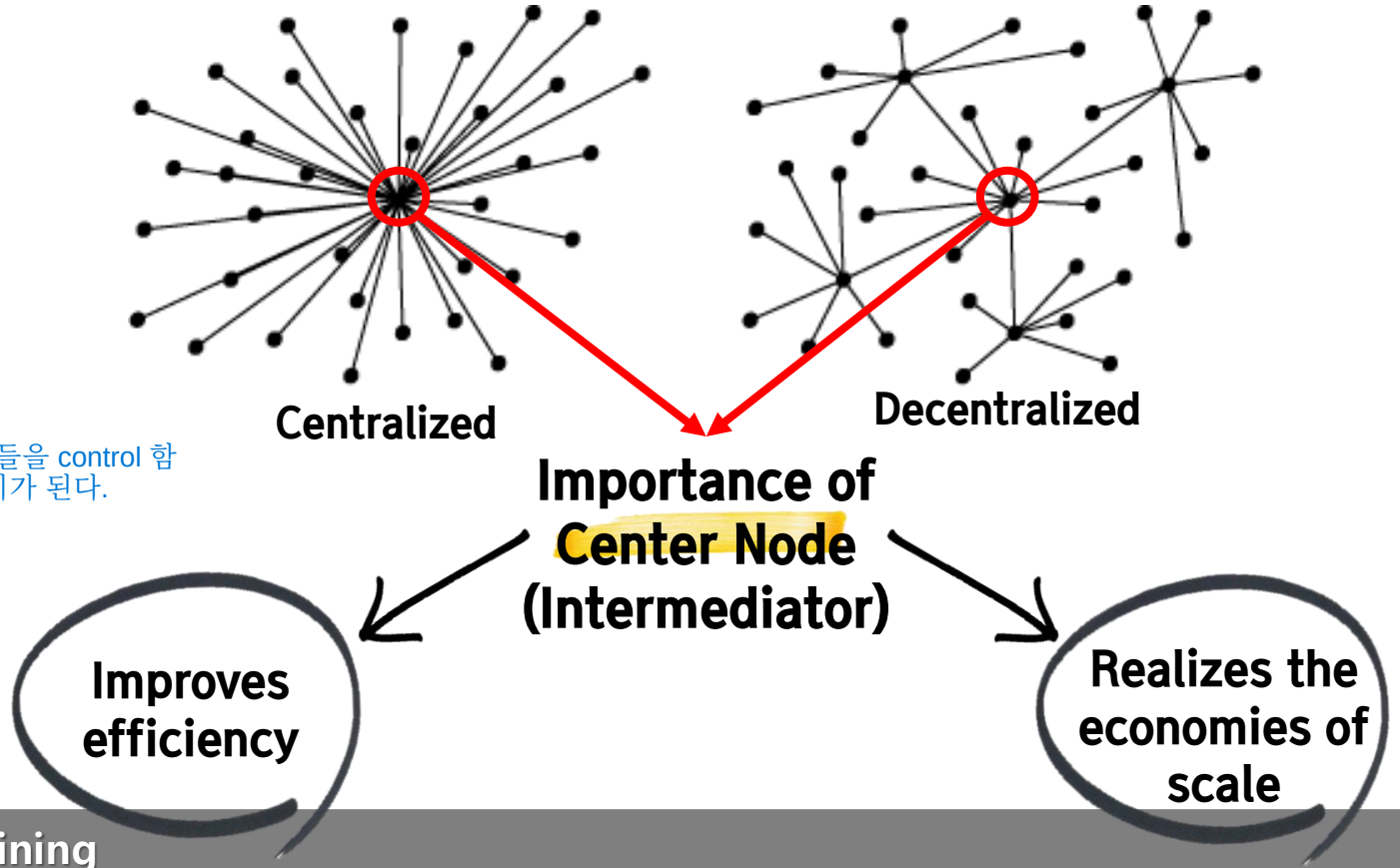
Decentralized network



Distributed network

Conventional Network Structures

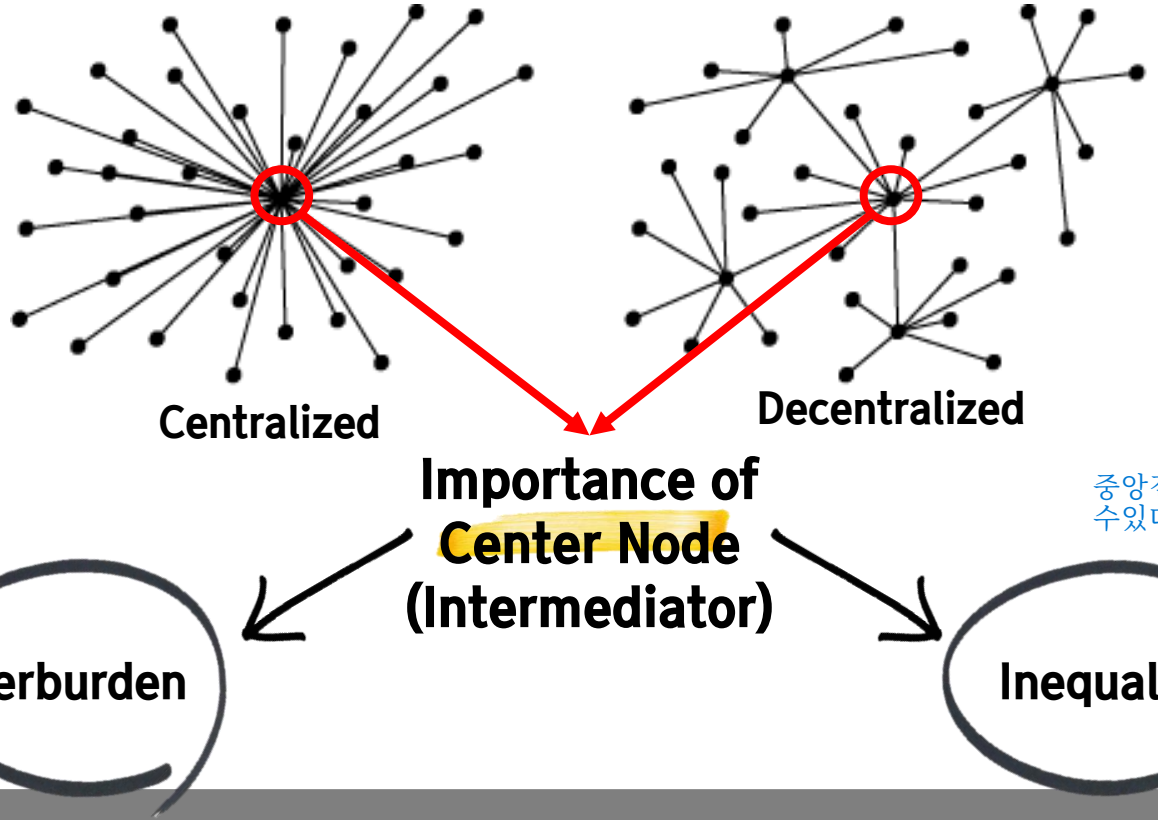
- Positive sides of centralized & decentralized networks
  - Autonomy: Can operate without a certain node
  - Improved Performance: Overcome the overburden issue
  - Economic Feasibility: Reduce investment for a certain node



특정권한을 중앙노드에게 주고 어떤것들을 control 함
치명적인 단점 중앙이 마비되면 다 마비가 된다.

Conventional Network Structures

- Negative sides of centralized & decentralized networks
  - Complexity: Difficult to control all nodes
  - Security: Security issues for all nodes
  - Management: Difficult to maintain overall management



중앙집중이 아니면 공평하지만 정말 비효율적인 일이 발생할 수있다. 너무 세분화

- Needs data management capability
- Requires high level of security cost

- Creates an unequal power balance → Strong dependency of weaker nodes

New Network Structure

- Distributed networks are starting to be recognized as the future network structure of our society
  - A distributed structure can not only overcome the limitations of a centralized structure, but also enhance the value of social capital.
  - However, the problem is that a distributed structure cannot be realized until these limitations are solved.
  - How can this problem of a distributed network be solved? → *Blockchain Technology*

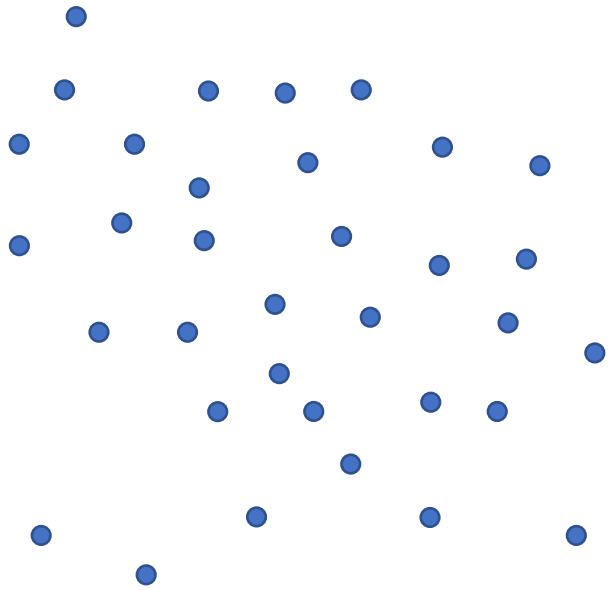
블록체인이라는 암호와 알고리즘을 통해 모두가 연결되어있으면서 모두가 이 안전하게 보호할수있는것이 나온다.
블록체인 기술은 우리가 이상으로 생각했던 사회구조를 도와주는 기술의 등장
네트워크 연결성에 대해 이야기했다 기본적으로 우리 사회 우리가 다루는 대다수의 데이터는 중앙집중형이다
특정소수가 많이 연결되고 많이 가지고있는 그것들에 대해 장단점이 있었고 다른 것들도 장단점이 있었다.

Network Analysis Setting

Network Analysis

- Network analysis is a method that depicts the relationship among the analysis targets in the data
  - Ex) Who is the most influential student in the class? Who collaborates more often with John? … etc.

데이터 안에서 내가 설정한 분석의 대상 들을 간의 관계를 묘사하고 설명하고 이해하기위해 네트워크분석이란것을 사용한다.
Ex)이수업에서 가장 영향력있는 학생은 누굴까?
그런것들을 네트워크분석을 볼수있는데 하나의 관찰자(observation)로 봐도 무방하다
관계적인 structural features
분석의 대상
분석의 대상과의 관계를 어떻게 정의해야할지 잘 표현한다.



Data

Network analysis

Information on structural features such as relationships and patterns between objects can be obtained

Network Analysis: Setting

- Setting
  - Node (*V*, vertices): Target of interest
  - Edge (*E*): Link between targets
  - Weighted(*W*): Weighted? Or unweighted?
    - Weighted meaning all the weights for each edge are measured differently based on the frequency of edges
    - Unweighted meaning all edges have the same weight (default 1)
  - Directed: Directed? Or undirected?
    - Direct meaning when an edge has a direction (A → B): when a direction exists between targets
    - Indirect meaning when an edge has no direction (A – B)
- The setting of a network depends on the data source and the goals that you are aiming for.

$$G = G(V, E, W)$$

네트워크 분석세팅
노드: 내가 분석하고자 하는 대상
(너가 학부가 노드라는것은 애가 경영경제,ict)
관계를 도대체 어떻게 정의할것인가?
0 0 이것두개가 어떻게 연결되었는지를 정의하는것
관계라는것은 이 두개를 어떻게 연결시킬것인가+ 연결 구조를 통하여 어떤것을 주로 말하고싶은가 (수업? 전공? 같은 교회?)

노드와 엣지 를 알면 50% 완성
Weighted(가중치) 연결에 있어 가중치를 넣을지 말지.

예) 두사람이 같이 듣는 수업이 3개인데 가중치를 둔다면 줄을 3개 연결, 그러나 가중치를 두지 않으면 선 하나로 통일

Directed 방향성이 있냐 없냐 (수업을 듣는것으로 네트워크 두고싶다 서로 방향성이 없다 같이 듣는것! , 누구한테 문자를 보낸다 그사람이 읽씹 문자를 보냈다 방향성이 있다. 현차가 삼전에 1억원을 팔았다 방향성이 있는것 관계에서도 방향성이 있는것이 있고 없는것이 있다) 이런것들을 누가 정해준다 (분석하는 사람이)

Network Analysis: Data Structure

- Data structure (undirected network)
  - Matrix or Pair
  - ex) Number of international branches
    - Apple: Paris, Dublin, London
    - Nike: Dublin, London
    - Louis Vuitton: Paris
    - Adidas: Paris, London

Dublin-Paris: Apple
Dublin-London: Apple, Nike
Paris-London: Apple, Adidas

Matrix table

	Dublin	Paris	London
Dublin	-	1	2
Paris	1	-	2
London	2	2	-

Pair table

From	To	Weight
Dublin	Paris	1
Dublin	London	2
Paris	London	2

Network Analysis: Data Structure

- Data structure (directed network)
  - ex) Movement of local branches to foreign market
    - AIB (Dublin): London, Paris
    - Guinness (Dublin): London
    - Louis Vuitton (Paris): London, Dublin
    - L'oreal (Paris): Dublin
    - Burberry (London): Paris

Dublin → Paris: AIB
Paris → Dublin: L'oreal, Louis Vuitton
Dublin → London: AIB, Guinness
London → Dublin: -
London → Paris: Burberry
Paris → London: Louis Vuitton

Matrix table

from \ to	Dublin	Paris	London
Dublin	-	1	2
Paris	2	-	1
London	0	1	-

Pair table

From	To	Weight
Dublin	Paris	1
Paris	Dublin	2
London	Dublin	0
Dublin	London	2
London	Paris	1
Paris	London	1

Network Analysis Indicators

Network Analysis: Measures

- Global level (network level)
  - Network size
  - Network density
  - …

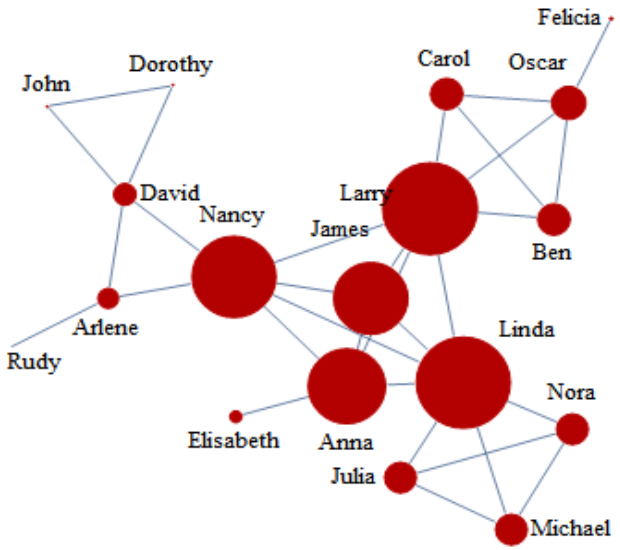
- Individual level (node or edge)
  - Degree centrality
  - Betweenness centrality
  - Closeness centrality
  - Eigenvector centrality
  - …

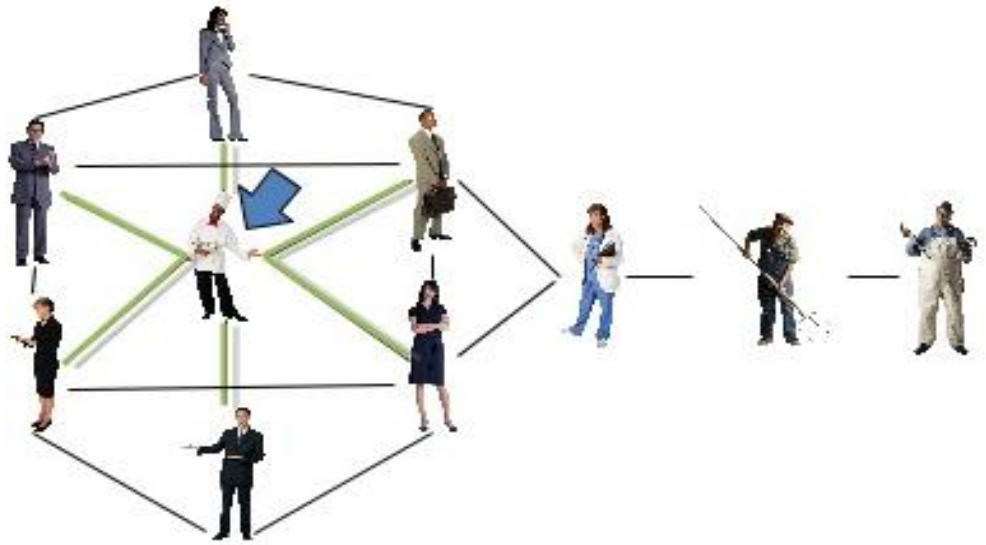
어느것이?
Individual level.
이 더 좋다

Global level 하면 많은 노드수

Network Analysis: Measures

- Degree centrality
  - Calculates how many edges a node is connected to
  - ex) people who have relationships with a large number of friends
  - $C_D(v) = \deg(v)$, v: node

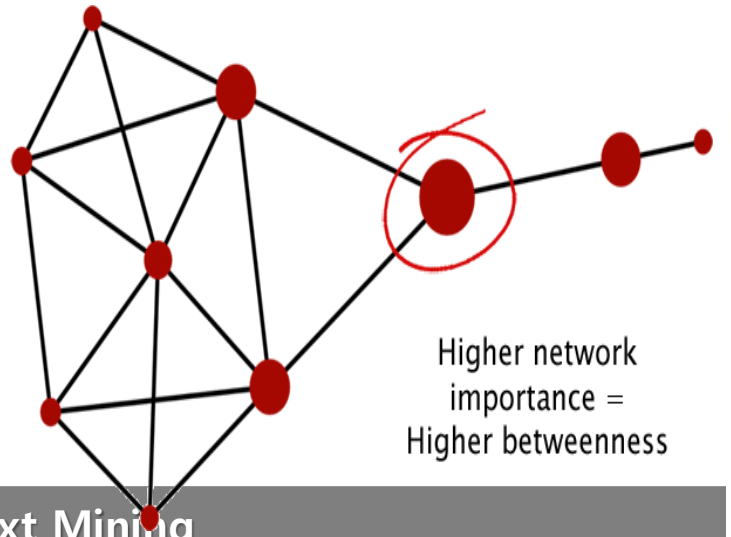


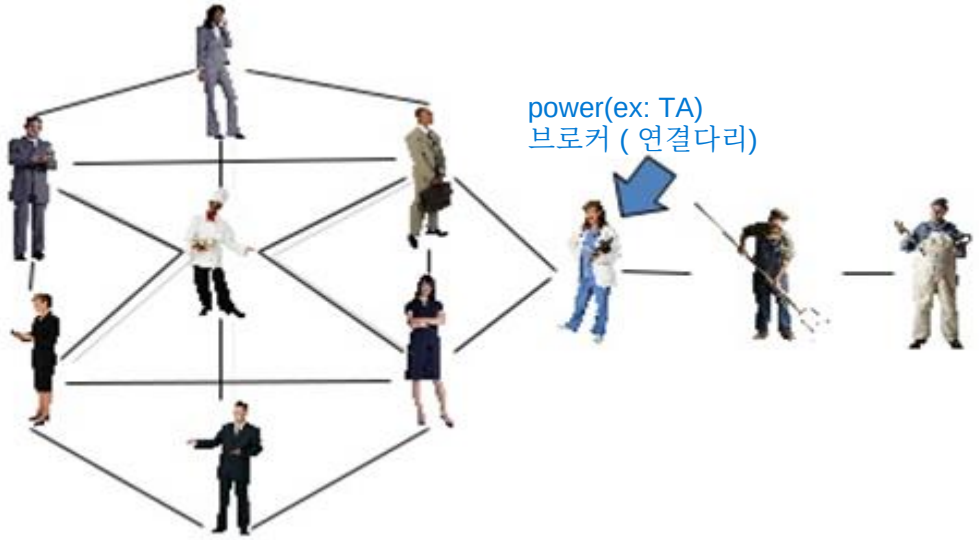


Network Analysis: Measures

- Betweenness centrality (Brokerage)
  - Determines the degree to which a node controls or brokers relationships between other nodes not directly connected

  → Calculate the brokering role in the network
  - ex) The person who gets the most contact when announcing a meeting (Teaching assistant)
  - $Betweenness\ centrality\ (v) = \sum_{s \neq v \neq t \epsilon V} \frac{\sigma_{st}(v)}{\sigma_{st}}$
  - v,s,t: node, V: node set, $\sigma_{st}(v)$: number of paths that pass v, σ_{st}: number of all paths

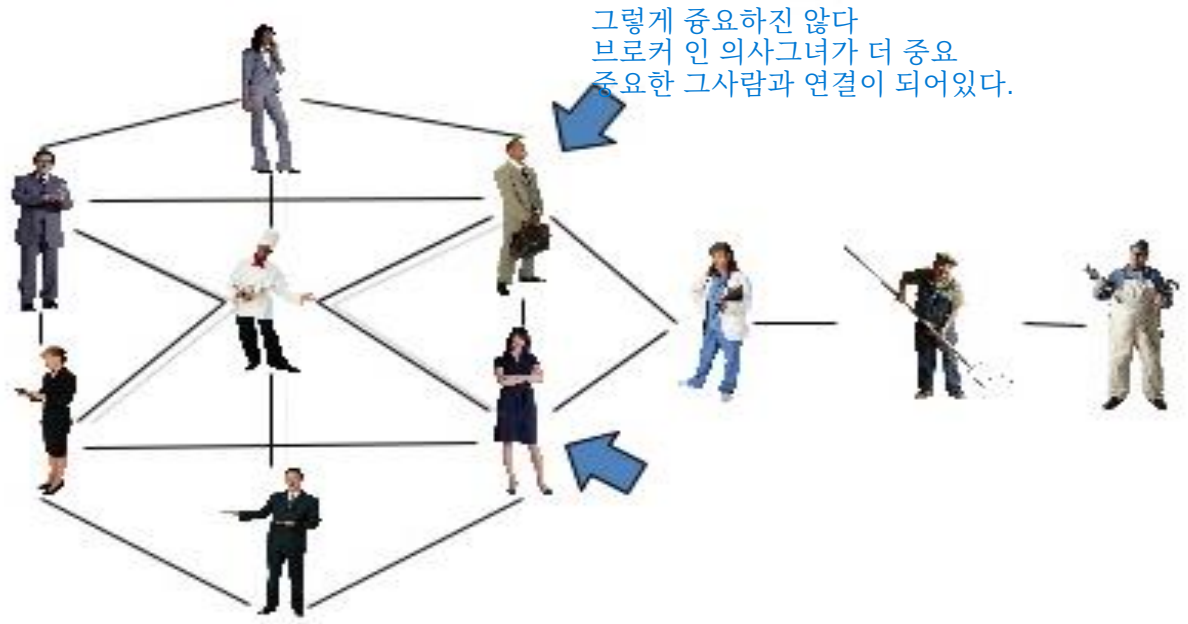




Network Analysis: Measures

- Closeness Centrality
  - Calculates the distance between a node and other nodes to understand whether it is linked closely to others.

  → Calculate whether a node is located closely to other nodes
  - ex) easy to get in touch with others in the center
  - $Closeness\ centrality\ (v) = \frac{1}{\sum_y d(y,v)}$
  - v,x: node, d: distance



Text Network Analysis

Text Network Analysis

- Text network analysis
  - Analyzing the relationship between the texts
  - Finding hidden text patterns

Start	Stop	Line	Color
Oberlaa	Neulaa	1	red
Neulaa	Alaudagasse	1	red
Alaudagasse	Altes Landgut	1	red
Altes Landgut	Troststrasse	1	red
Troststrasse	Reumannplatz	1	red
Reumannplatz	Keplerplatz	1	red
Keplerplatz	Suedtiroler Platz - Hauptbahnhof	1	red
Suedtiroler Platz - Hauptbahnhof	Taubstummengasse	1	red
Taubstummengasse	Karlsplatz	1	red

Network relation between subway stations

<Vienna Subway Data>
https://www.kaggle.com/datasets/lenapiter/vienna-subway-network

아담과 제임이 3번정도 만난다 (스크립트를 통해 알수있음)

Source	Target	Type	weight	book
Addam-Marbrand	Jaime-Lannister	Undirected	3	1
Addam-Marbrand	Tywin-Lannister	Undirected	6	1
Aegon-I-Targaryen	Daenerys-Targaryen	Undirected	5	1
Aegon-I-Targaryen	Eddard-Stark	Undirected	4	1
Aemon-Targaryen-(Maester-Aemon)	Alliser-Thorne	Undirected	4	1
Aemon-Targaryen-(Maester-Aemon)	Bowen-Marsh	Undirected	4	1
Aemon-Targaryen-(Maester-Aemon)	Chett	Undirected	9	1
Aemon-Targaryen-(Maester-Aemon)	Clydas	Undirected	5	1
Aemon-Targaryen-(Maester-Aemon)	Jeor-Mormont	Undirected	13	1

Network relation between characters

<Game of Throne Data>
https://www.kaggle.com/code/mmmarchetti/game-of-thrones-network-analysis/data

- Any case where a column is filled with multiple elements has the potential to be used for network analysis

StudentID	Major	Community	Age
Kim_001	Economics – AI	DAL; Chayochayo	19
Yoon_001	AI – Computer Science	Chamber; DAL; Milan	24
Ryu_001	AI – Computer Science	Doonamis; Milan	33
Oh_001	ICT – Life Science	Milan; Chayochayo	21
Lee_001	ICT – Counselling Psychology	Chamber; Milan	22

Major Co-occurrence Network

Community Co-occurrence Network

Major1	Major2	Weight
Economics	AI	1
AI	Computer Science	2
ICT	Life Science	1
ICT	Counselling Psychology	1

Community1	Community2	Weight
DAL	Chayochayo	1
Chamber	DAL	1
DAL	Milan	1
Chamber	Milan	2
Doonamis	Milan	1
Milan	Chayochayo	1

Text Network Analysis

- Setting
  - Node: unit of text (word or n-gram words)
  - Edge: link between texts (co-occurrence)
  - In most cases, an undirected network is used

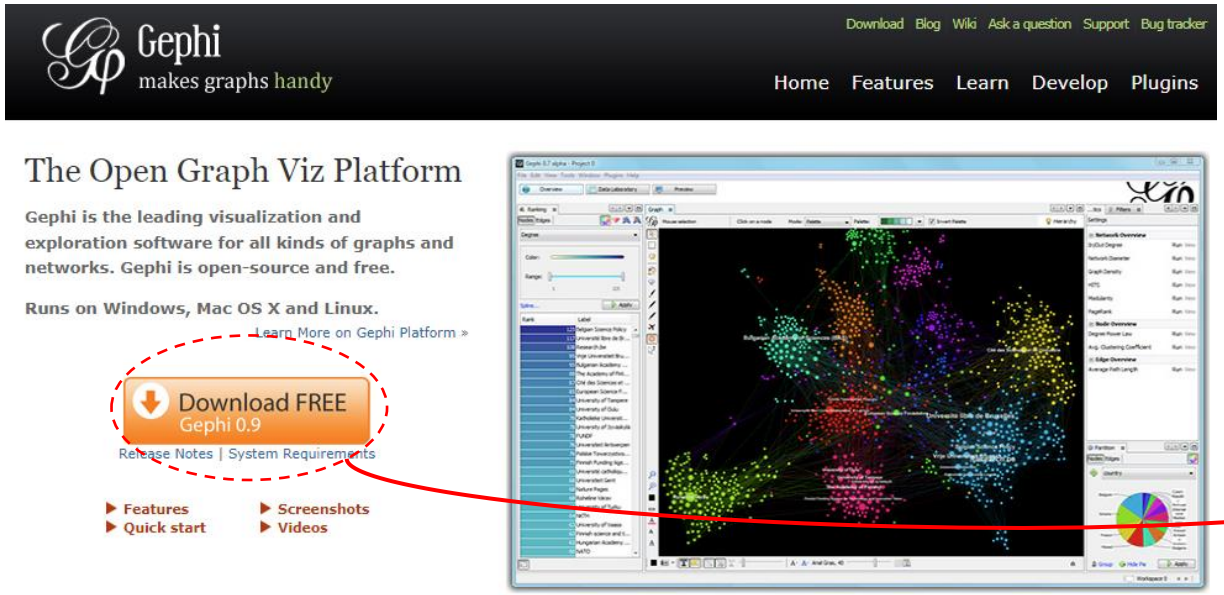
  → Gives greater value to the words that are more frequently used together (in a sentence, in a paragraph, or in a document)

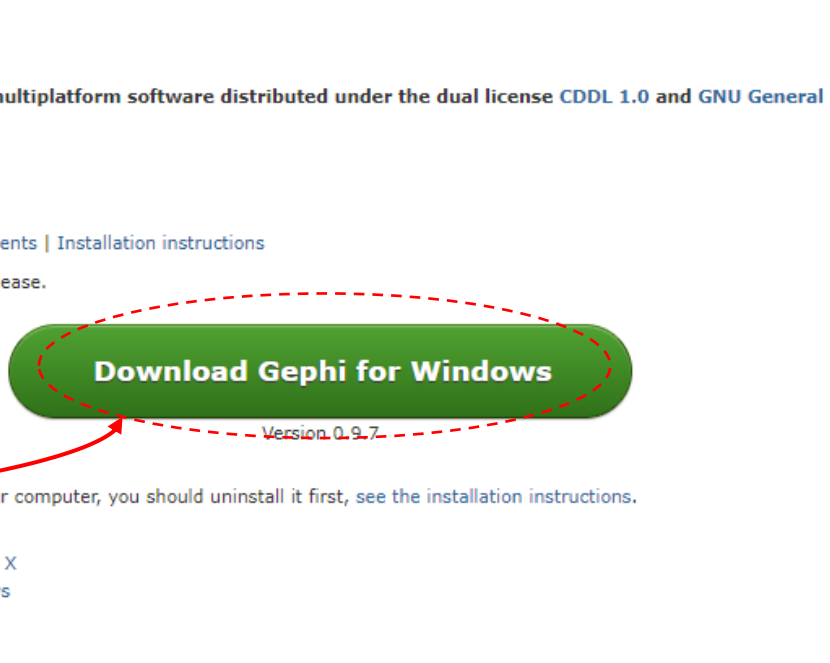
- Steps
  - Create a node table → Used for classifying the nodes
  - Create an edge table → Used for building the network
  - Create a network object → Used for the main network analysis
  - Visualize or measure network indicators

- igraph package
  - A collection of network analysis tools with an emphasis on efficiency, portability, and ease of use.
  - Possible to use in R, Python, C/C++

- igraph::graph_from_data_frame(*d*, *vertices*, directed=TRUE or FALSE)
  - Creates an igraph object from the data frame
  - d: edge table
  - vertices: node table
  - directed: TRUE for directed network and FALSE for undirected network

- Gephi
  - Visualization and exploration software for network analysis
  - Open-source and free
  - Useful for generating a high-quality network visualization
  - https://gephi.org/





Text Network Analysis – Data Preparation

Data Preparation: Harry Potter Script

- Import Harry Porter script to R

```
> hp.script <- read.csv(file="R file/R file_LEC09/hp_script_ed.csv")
> hp.script %>% head(2)
  ID_number scene     character_name                                                dialogue
1         1     1     Albus Dumbledore I should have known that you would be here, Professor McGonagall.
2         2     1 Minerva McGonagall   Good evening, Professor Dumbledore. Are the rumours true Albus?
> hp.script %>% nrow
[1] 793
> hp.script$ID_number %>% unique %>% length
[1] 793
> hp.script$scene %>% unique %>% length
[1] 34
> hp.script$character_name %>% unique %>% length
[1] 41
> hp.script$dialogue %>% unique %>% length
[1] 758
```

- *ID_number: the unique ID number of each line of dialogue*
- *scene: the scene number as stated on the DVD/Blu-ray*
- *character_name: name of the character speaking the line of dialogue*
- *dialogue: the dialogue of the character*

Data Preparation: Harry Potter Script

- Based on the frequentist approach, who are the top 3 main characters in the Harry Potter movie?

- Based on the frequentist approach, find the main characters in each scene.

Data Preparation: Harry Potter Script

- Based on the frequentist approach, who are the top 3 main characters in the Harry Porter movie?

```
> hp.script %>%
+   group_by(character_name) %>%
+   summarize(count=length(ID_number)) %>%
+   arrange(desc(count))
# A tibble: 41 x 2
   character_name     count
   <chr>              <int>
 1 Harry Potter         230
 2 Ron Weasley          120
 3 Hermione Granger      92
 4 Rubeus Hagrid         81
 5 Minerva McGonagall    31
 6 Albus Dumbledore      24
 7 Vernon Dursley        23
 8 Dudley Dursley        17
 9 Quirinus Quirrell     17
10 Neville Longbottom    14
# ... with 31 more rows
```

- Based on the frequentist approach, find the main characters in each scene.

```
> hp.script %>%
+   group_by(scene, character_name) %>%
+   summarize(count=length(ID_number)) %>%
+   arrange(desc(count)) %>%
+   slice(1)
`summarise()` has grouped output by 'scene'.
# A tibble: 34 x 3
# Groups:   scene [34]
   scene character_name   count
   <int> <chr>            <int>
 1     1 Albus Dumbledore     9
 2     2 Dudley Dursley      11
 3     3 Vernon Dursley      12
 4     4 Rubeus Hagrid       17
 5     5 Rubeus Hagrid        6
 6     6 Rubeus Hagrid        6
 7     7 Harry Potter         5
 8     8 Rubeus Hagrid        6
 9     9 Molly Weasley        5
10    10 Ron Weasley         16
# ... with 24 more rows
```

Data Preparation: Harry Potter Script

- Create a co-occurrence matrix

```
> hp.mat <- hp.script %>%

> hp.mat
                    character_name
character_name       Albus Dumbledore Argus Filch Bloody Baron Class
  Albus Dumbledore                   0           0            1     0
  Argus Filch                        0           0            0     0
  Bloody Baron                       1           0            0     0
  Class                              0           0            0     0
  Doris Crockford                    0           0            0     0
```

Data Preparation: Harry Potter Script

- Create a co-occurrence matrix

*crossprod:
A = m x n → m: 장면 수, n: 캐릭터 수
A T A → (n x m) x (m x n)

```
> hp.mat <- hp.script %>%
+   select(scene, character_name) %>%
+   table %>%
+   crossprod
> diag(hp.mat) <- 0
> hp.mat
                    character_name
character_name       Albus Dumbledore Argus Filch Bloody Baron Class
  Albus Dumbledore                  0           0            1     0
  Argus Filch                       0           0            0     0
  Bloody Baron                      1           0            0     0
  Class                             0           0            0     0
  Doris Crockford                   0           0            0     0
```

Setting 0 for self co-occurrence

diag(hp.mat)<-0 이거 잊지말기

Data Preparation: Harry Potter Script

- Create a node table

```
> hp.mat.node <- hp.mat %>%
```

```
> hp.mat.node %>% head
                     character_name freq
Albus Dumbledore Albus Dumbledore  226
Argus Filch           Argus Filch  121
Bloody Baron         Bloody Baron   23
Class                       Class   71
Doris Crockford   Doris Crockford   14
Draco Malfoy         Draco Malfoy  402
```

- Create an edge table

```
> hp.mat.edge <- hp.mat %>%
```

```
> hp.mat.edge %>% head
                from                  to Frequency
1 Albus Dumbledore Albus Dumbledore         NA
2      Argus Filch Albus Dumbledore         NA
3     Bloody Baron Albus Dumbledore          1
4            Class Albus Dumbledore         NA
5  Doris Crockford Albus Dumbledore         NA
6     Draco Malfoy Albus Dumbledore         NA
```

NA if frequency is 0

Data Preparation: Harry Potter Script

- Create a node table

```
> hp.mat.node <- hp.mat %>%
+   as.data.frame %>%
+   mutate(character_name = rownames(.),
+          freq = rowSums(.)) %>%
+   select(character_name, freq)
> hp.mat.node %>% head
                    character_name freq
Albus Dumbledore Albus Dumbledore  226
Argus Filch           Argus Filch  121
Bloody Baron        Bloody Baron   23
Class                      Class   71
Doris Crockford   Doris Crockford   14
Draco Malfoy         Draco Malfoy  402
```

- Create an edge table

```
> hp.mat.edge <- hp.mat %>%
+   as.data.frame %>%
+   mutate(from = rownames(.)) %>%
+   gather(to, Frequency,
+          'Albus Dumbledore':Voldemort) %>%
+   mutate(Frequency =
+            ifelse(Frequency == 0, NA, Frequency))
> hp.mat.edge %>% head
                from                to Frequency
1 Albus Dumbledore Albus Dumbledore        NA
2      Argus Filch Albus Dumbledore        NA
3     Bloody Baron Albus Dumbledore         1
4            Class Albus Dumbledore        NA
5  Doris Crockford Albus Dumbledore        NA
6     Draco Malfoy Albus Dumbledore        NA
```

NA if frequency is 0

Text Network Analysis – Network Analysis (igraph)

Text Network Analysis (igraph): Harry Potter Script

- Create an igraph object

```
> library(igraph)
> hp.graph <-
+   graph_from_data_frame(
+     d=hp.mat.edge %>% filter(is.na(Frequency)==FALSE),
+     vertices=hp.mat.node,
+     directed = FALSE)
> hp.graph
IGRAPH 1f1c35f UN-- 41 494 --
+ attr: name (v/c), freq (v/n), Frequency (e/n)
+ edges from 1f1c35f (vertex names):
 [1] Albus Dumbledore--Bloody Baron         Albus Dumbledore--Fred Weasley
 [3] Albus Dumbledore--George Weasley       Albus Dumbledore--Harry Potter
 [5] Albus Dumbledore--Hermione Granger     Albus Dumbledore--Minerva McGonagall
 [7] Albus Dumbledore--Nearly Headless Nick Albus Dumbledore--Neville Longbottom
 [9] Albus Dumbledore--Percy Weasley        Albus Dumbledore--Quirinus Quirrell
[11] Albus Dumbledore--Ron Weasley          Albus Dumbledore--Rubeus Hagrid
[13] Albus Dumbledore--Seamus Finnigan      Albus Dumbledore--Sorting Hat
[15] Albus Dumbledore--The Fat Lady         Argus Filch     --Draco Malfoy
+ ... omitted several edges
```

Exclude Frequency with NA

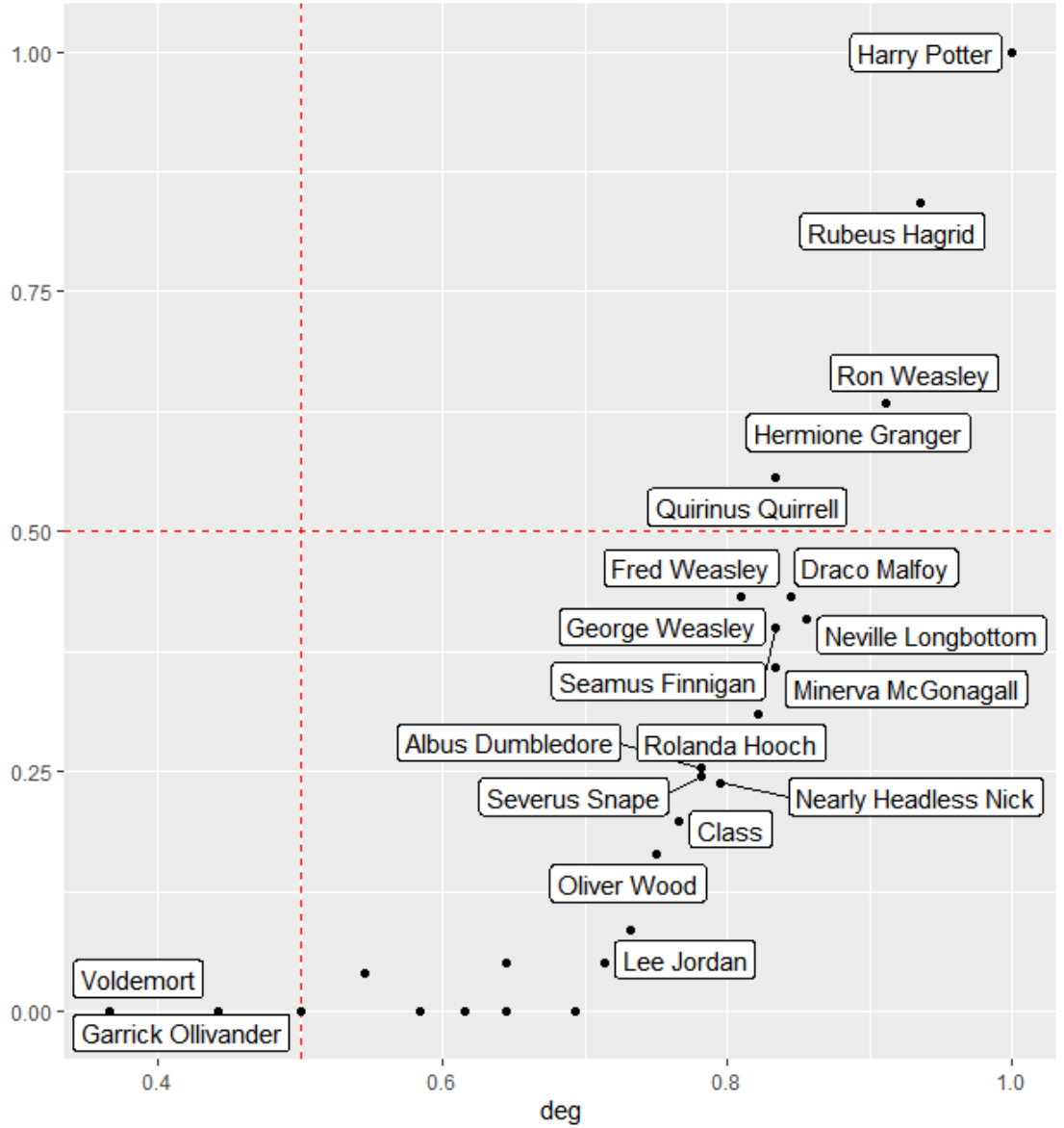
undirectied network

Text Network Analysis (igraph): Harry Potter Script

- Measure network centrality indicator

```
library(igraph)
hp.graph.tbl <-
  data.frame(
    node = V(hp.graph) %>% names,
    deg = hp.graph %>% degree,
    bet = hp.graph %>% betweenness,
    clo = hp.graph %>% closeness
  )
library(ggplot2)
library(ggrepel)
hp.graph.tbl %>%
  mutate(deg=log(deg+1)/max(log(deg+1)),
         bet=log(bet+1)/max(log(bet+1))) %>%
  ggplot(aes(x=deg, y=bet)) +
  geom_vline(xintercept=0.5, linetype='dashed', color='red') +
  geom_hline(yintercept=0.5, linetype='dashed', color='red') +
  geom_point() +
  geom_label_repel(aes(label=node))
```

Betweenness Centrality

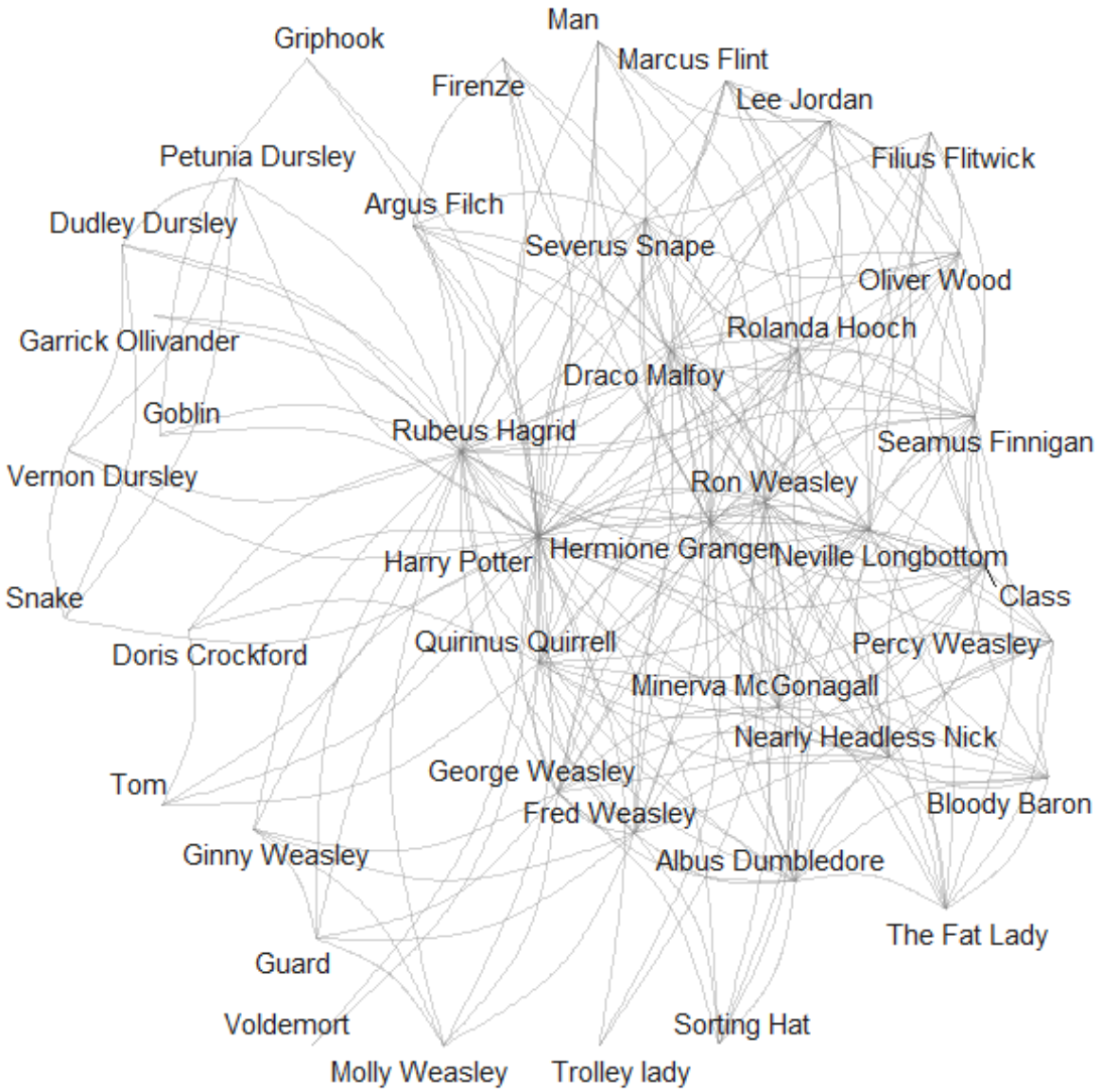


Degree Centrality

Text Network Analysis (igraph): Harry Potter Script

- Visualize text network

```
library(ggraph)
hp.graph %>%
  ggraph(layout = "auto") +
  geom_edge_arc(colour= "gray50",
                strength = .2, alpha = .2) +
  geom_node_text(aes(label = name),
                 repel = TRUE,
                 colour="gray10") +
  theme_graph(background = "white")
```



Text Network Analysis – Network Analysis (Gephi)

Text Network Analysis (Gephi): Harry Potter Script

- Visualize with Gephi
  - Create a proper .csv file for Gephi

```
write.csv(hp.mat.node %>%
            rename(Id=character_name) %>% mutate(Label=Id) %>%
            select(Id, Label, freq),
          file="R file/R file_LEC09/hp.mat.node.csv",
          row.names=FALSE)
```

Additional label information can be added…

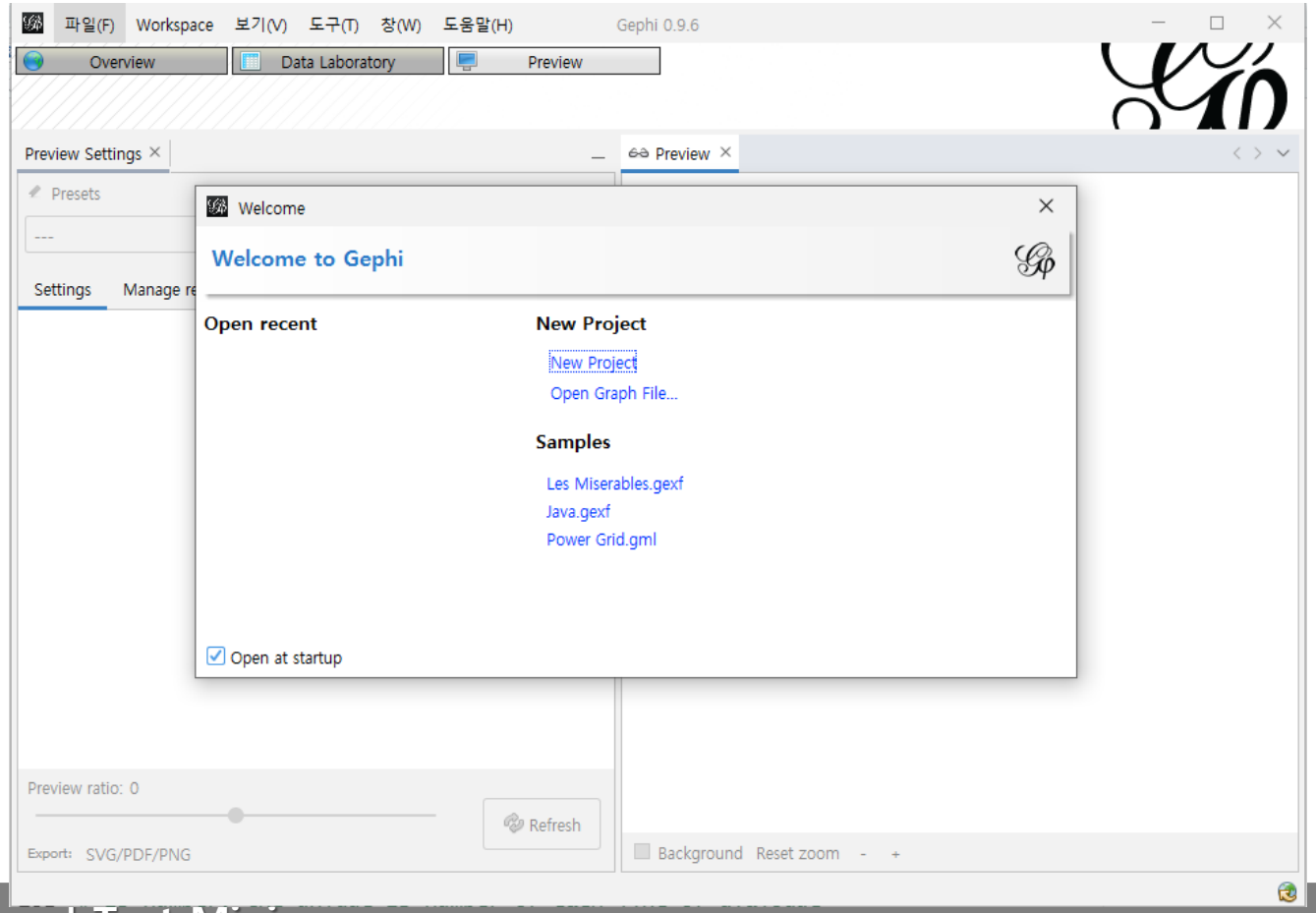
	A	B	C
1	Id	Label	freq
2	Albus Dumbledore	Albus Dumbledore	226
3	Argus Filch	Argus Filch	121
4	Bloody Baron	Bloody Baron	23
5	Class	Class	71
6	Doris Crockford	Doris Crockford	14
7	Draco Malfoy	Draco Malfoy	402
8	Dudley Dursley	Dudley Dursley	418
9	Filius Flitwick	Filius Flitwick	80
10	Firenze	Firenze	125

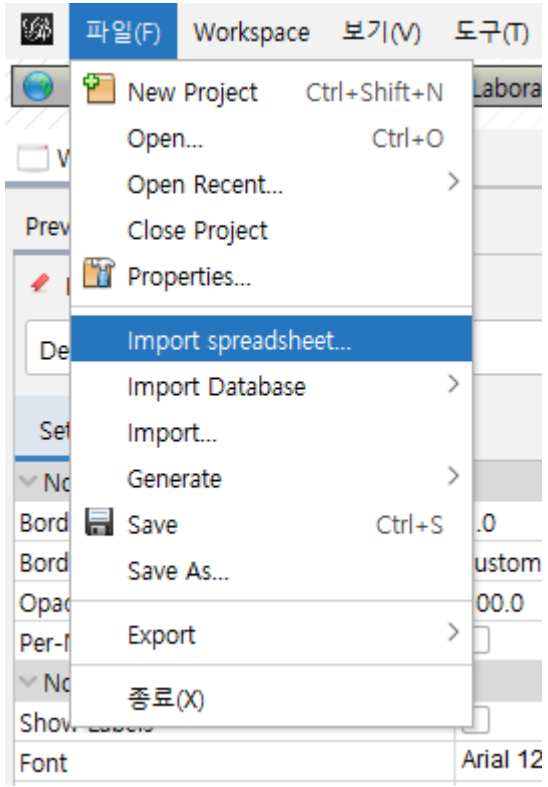
```
write.csv(hp.mat.edge %>% filter(is.na(Frequency)==FALSE) %>%
            rename(Source=from, Target=to) %>%
            rename(Weight=Frequency) %>%
            select(Source, Target, Weight),
          file="R file/R file_LEC09/hp.mat.edge.csv",
          row.names=FALSE)
```

	A	B	C
1	Source	Target	Weight
2	Bloody Baron	Albus Dumbledore	1
3	Fred Weasley	Albus Dumbledore	1
4	George Weasley	Albus Dumbledore	1
5	Harry Potter	Albus Dumbledore	92
6	Hermione Granger	Albus Dumbledore	13
7	Minerva McGonagall	Albus Dumbledore	55
8	Nearly Headless Nick	Albus Dumbledore	4
9	Neville Longbottom	Albus Dumbledore	2
10	Percy Weasley	Albus Dumbledore	6

Text Network Analysis (Gephi): Harry Potter Script

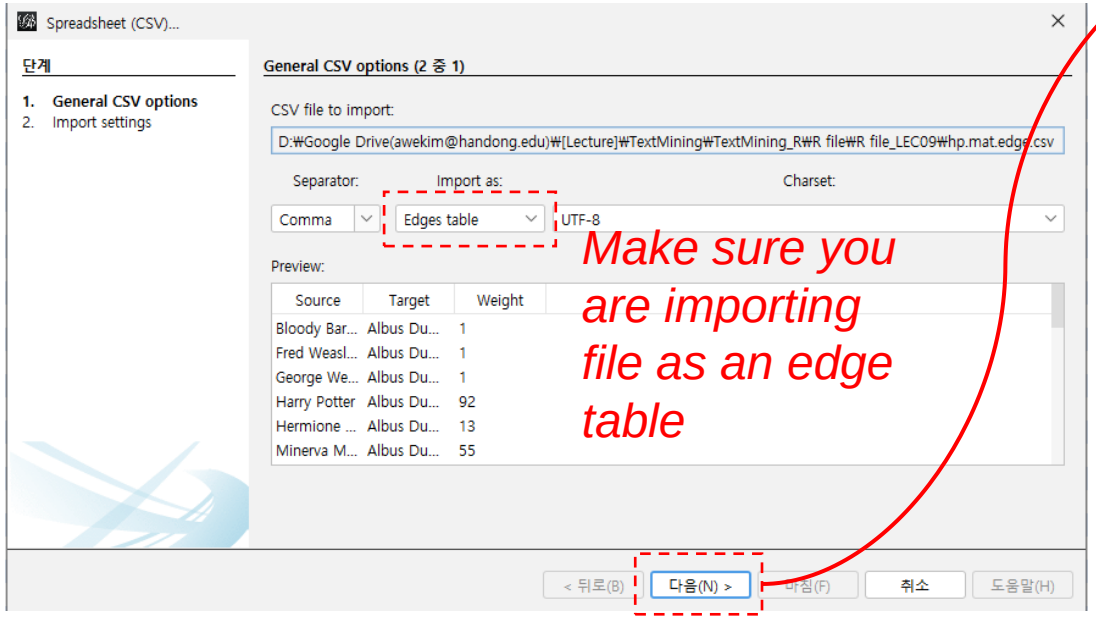
- Visualize with Gephi
  - Create a new project & save it as “hp_gephi”
  - Import node table & edge table

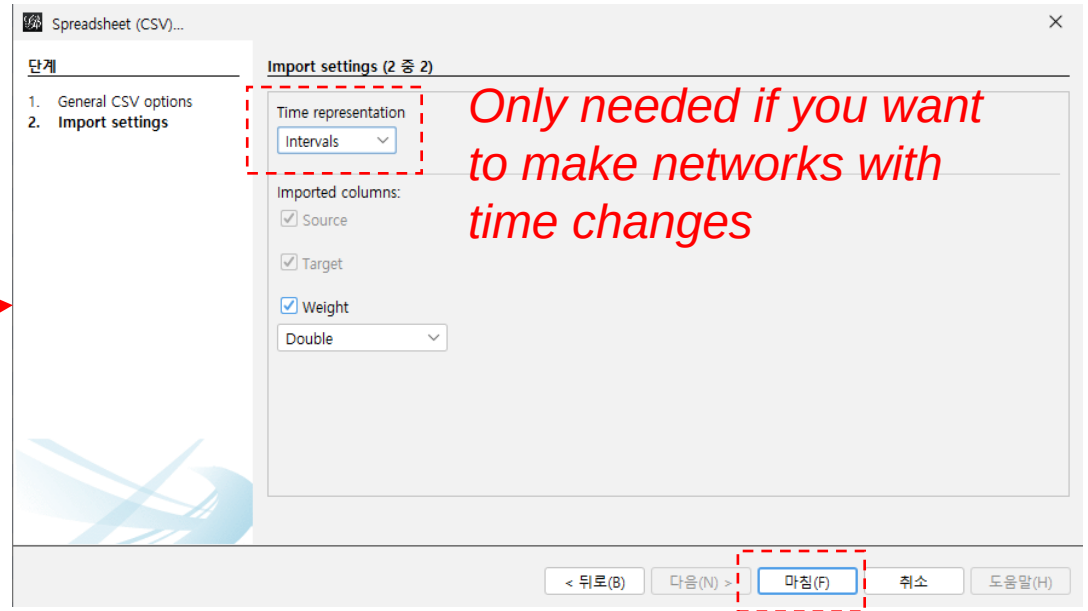


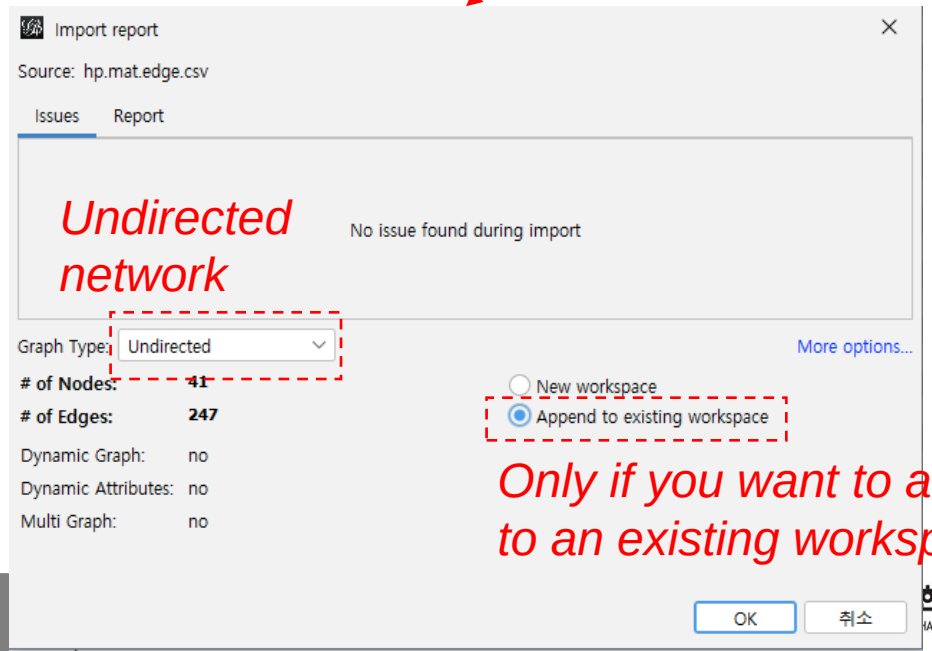


Text Network Analysis (Gephi): Harry Potter Script

- Visualize text network
  - Import edge table

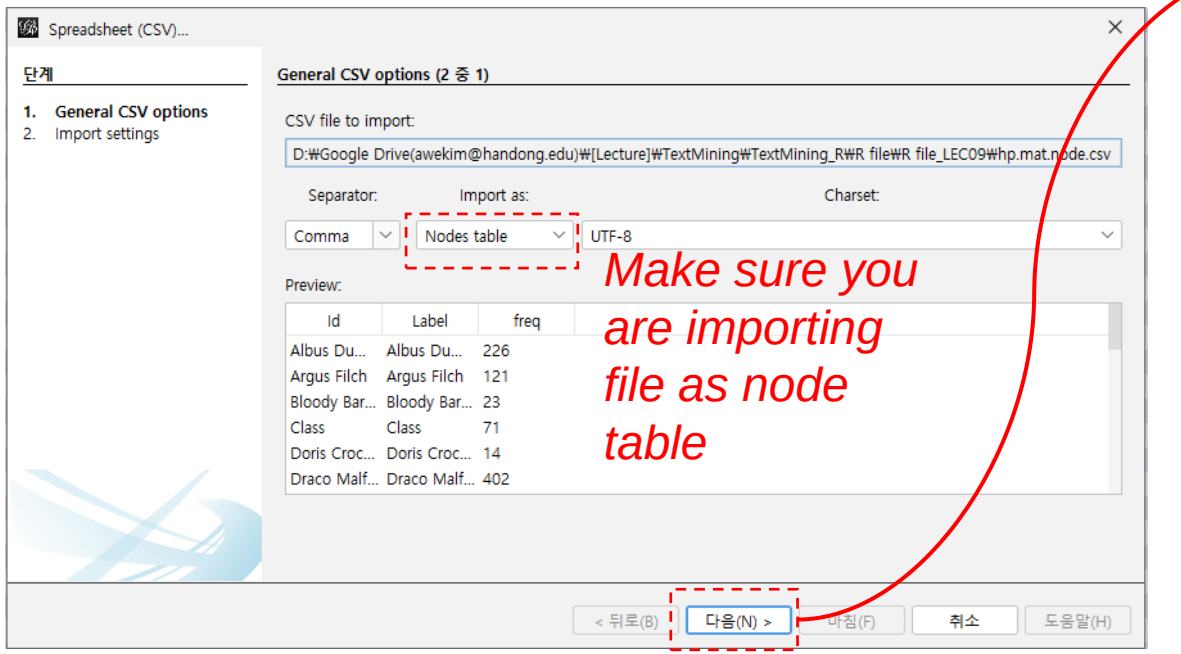




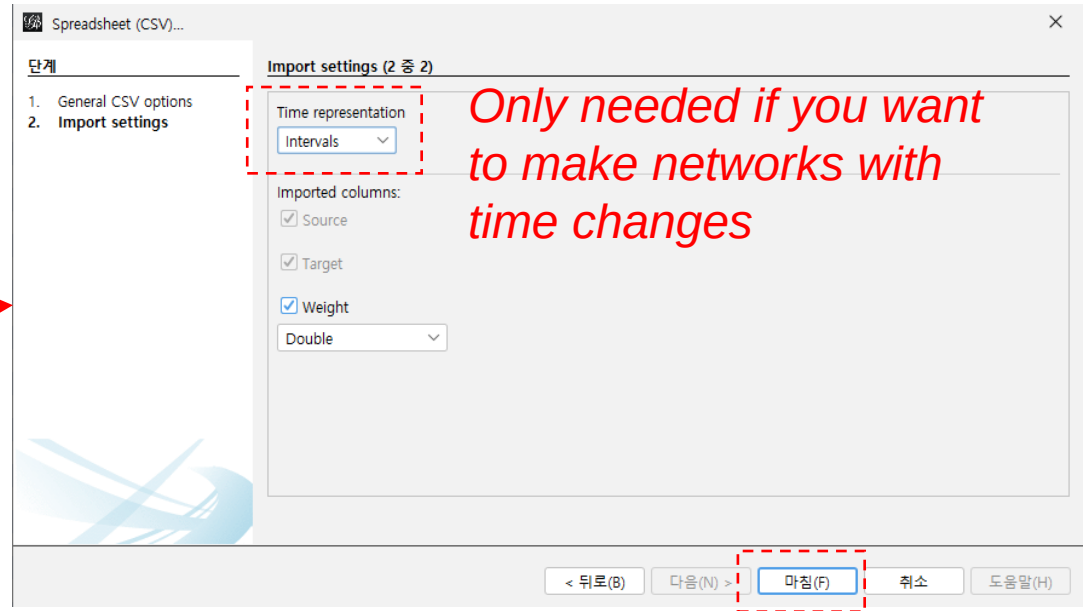


Text Network Analysis (Gephi): Harry Potter Script

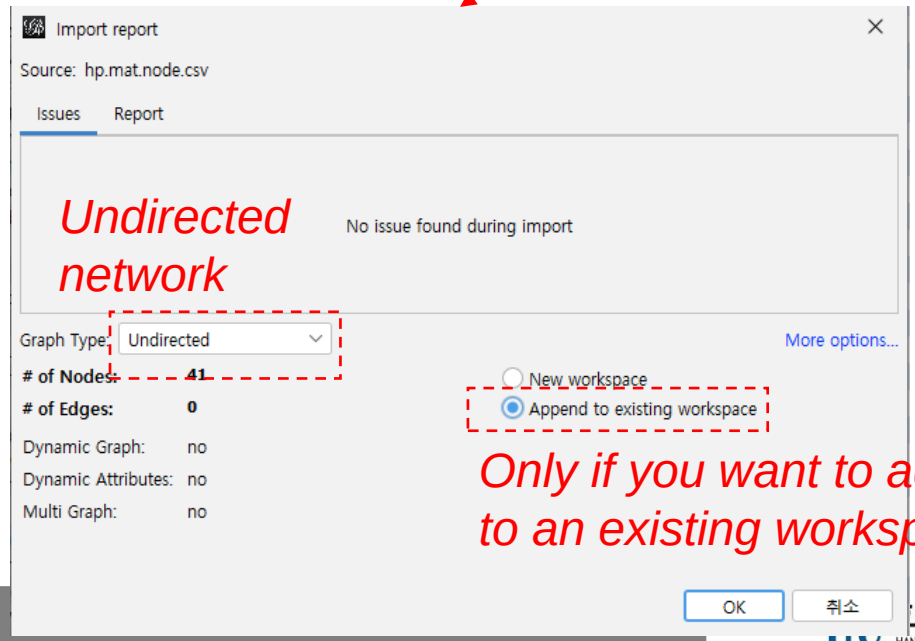
- Visualize text network
  - Import node table



Make sure you are importing file as node table



Only needed if you want to make networks with time changes



Undirected network

Only if you want to add table to an existing workspace

Text Network Analysis (Gephi): Harry Potter Script

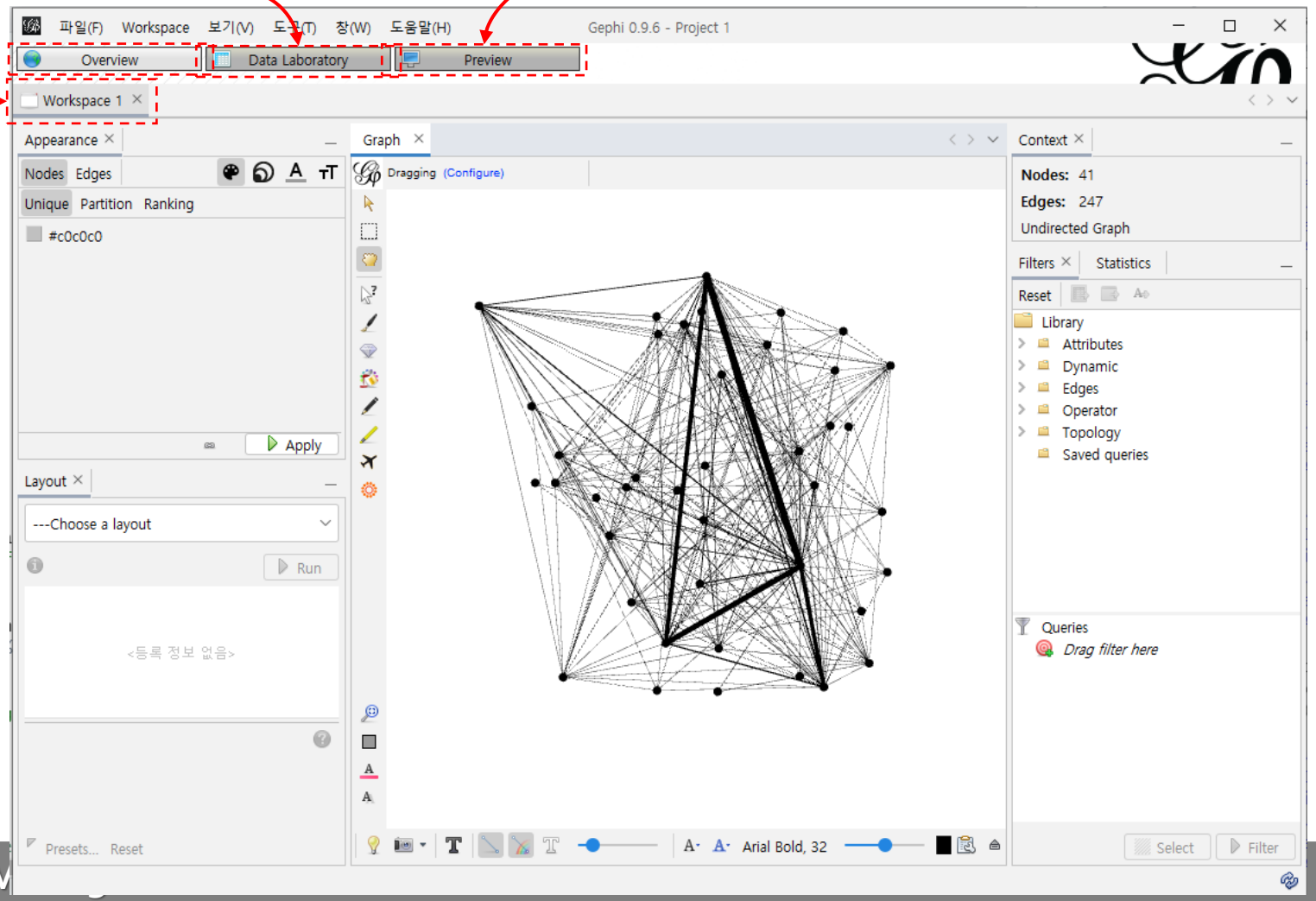
- Visualize text network

For you to work on the network visualization

Check the data

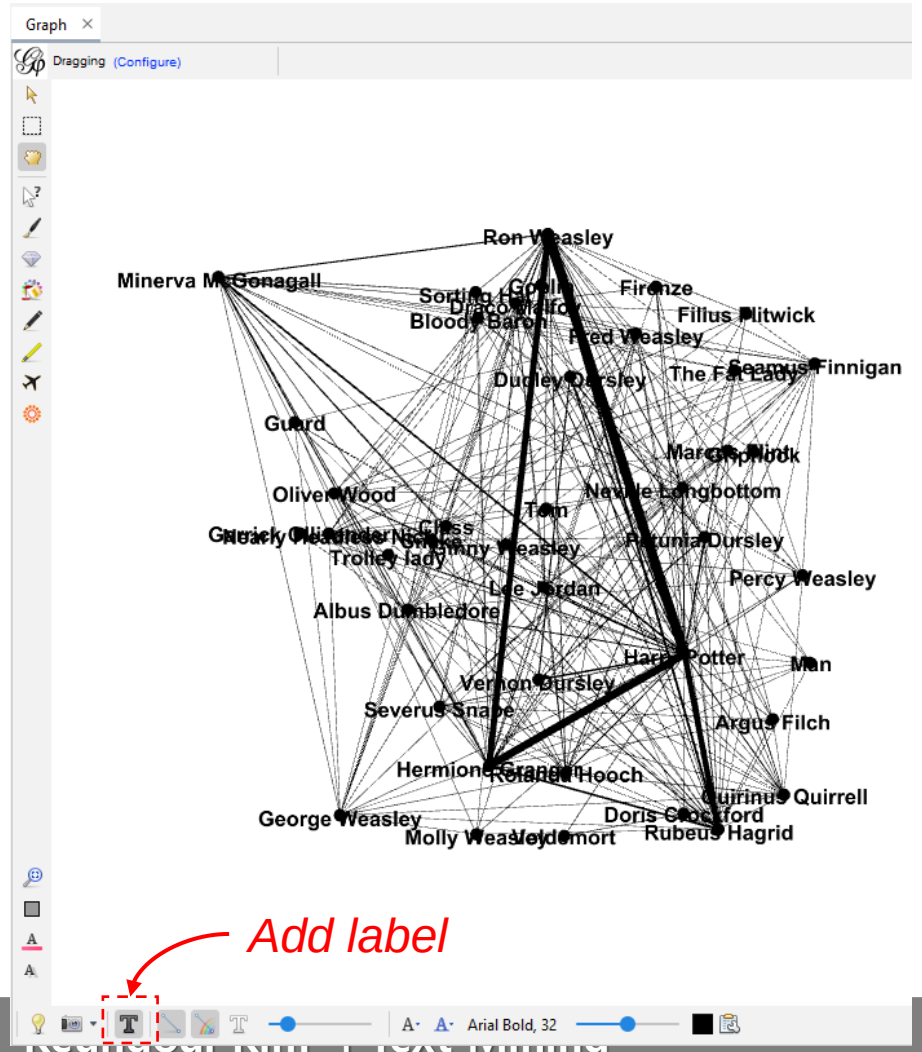
Check work

Workspace. To change the name, Workspace → Rename

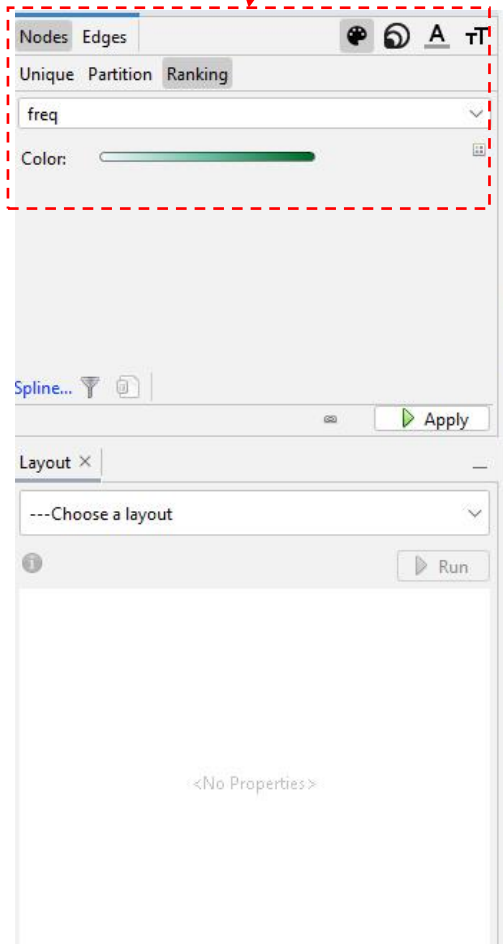


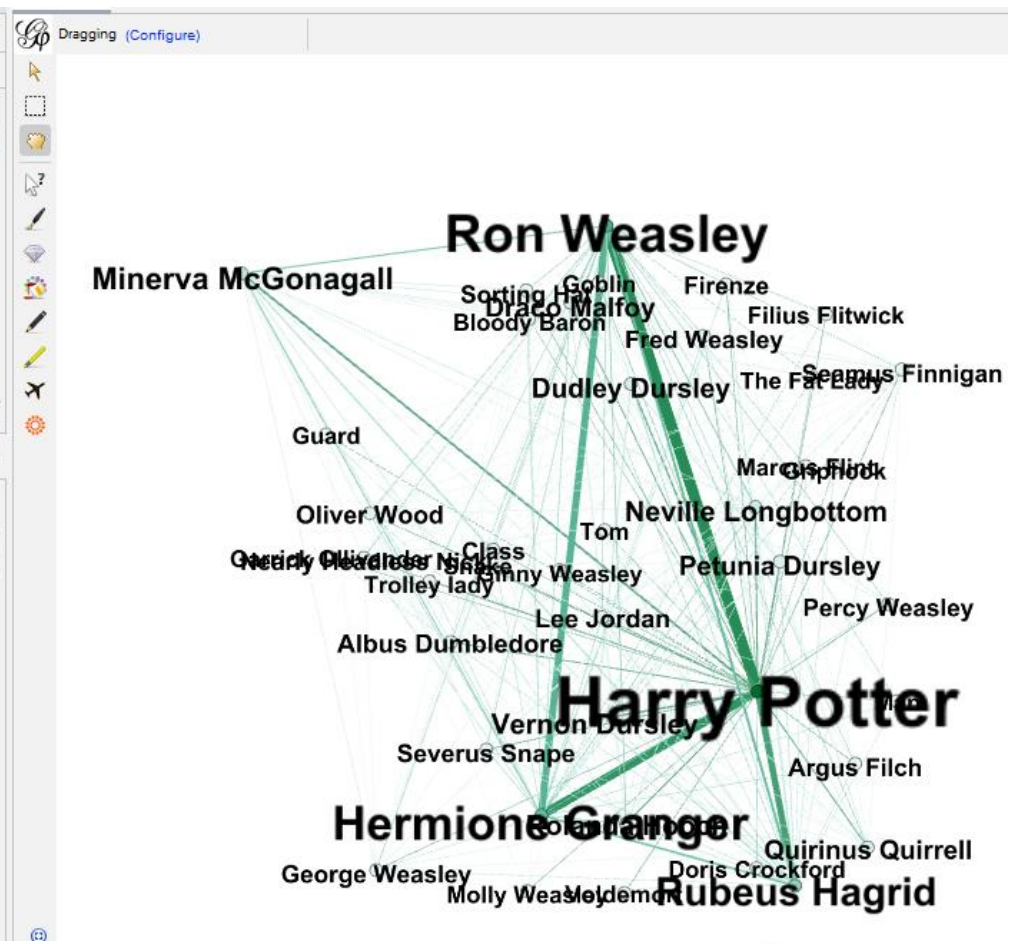
Text Network Analysis (Gephi): Harry Potter Script

- Visualize text network
  - Add node information



Change color and text size

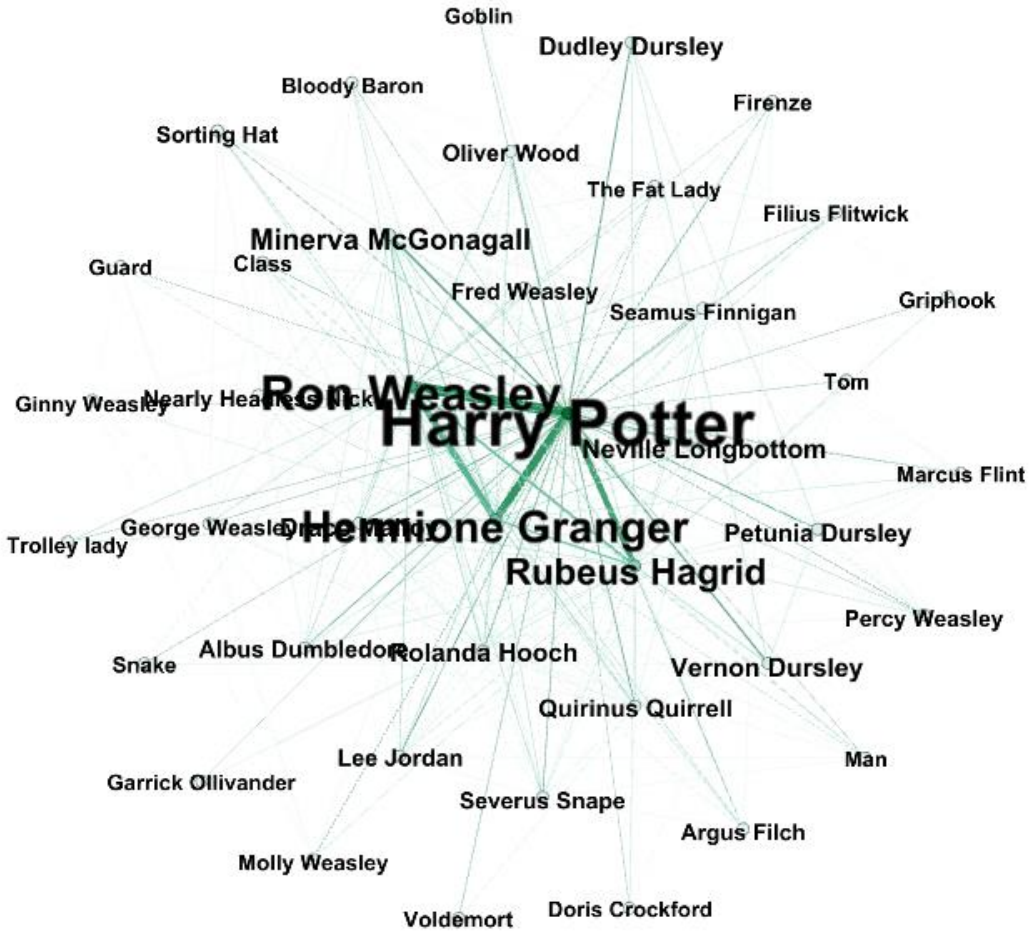




Text Network Analysis (Gephi): Harry Potter Script

- Visualize text network
  - Change layout

Fruchterman Reingold



Yifan Hu