# Network Theory and Basic Statistic

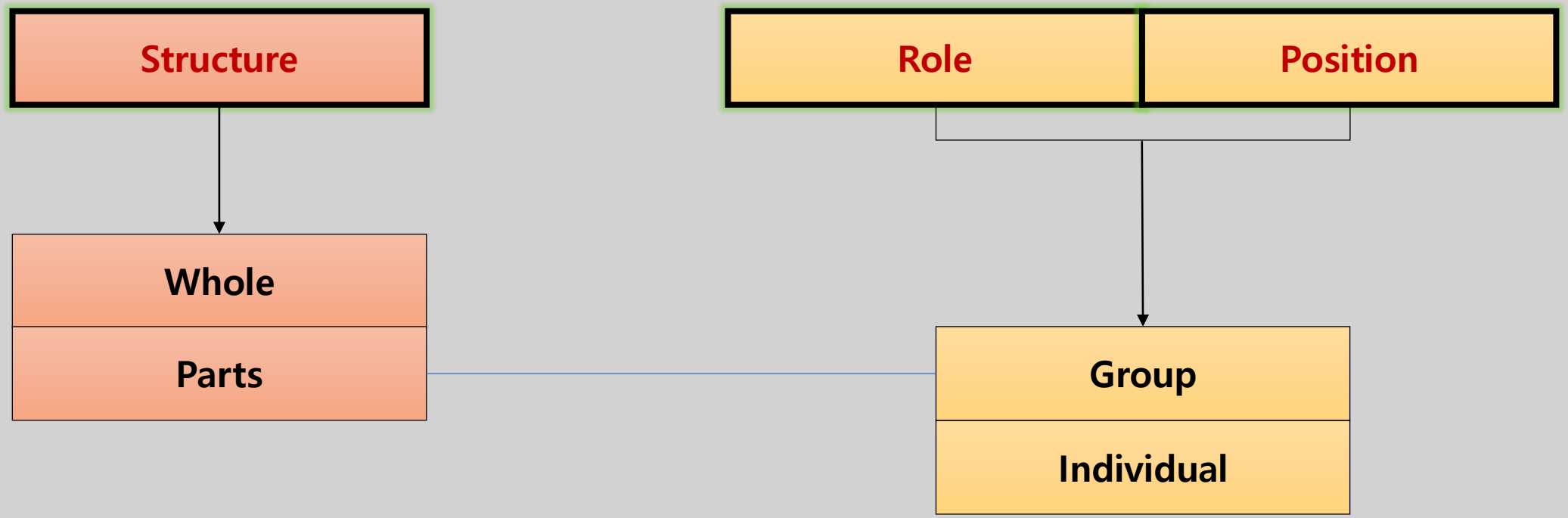
# 1. 네트워크의 개념

# 1. SNA

- 네트워크 분석할 때 고려해야 할 점
  - 사회 연결망 분석방법으로 잘 보여줄 수 있는 주제인가?
  - 사회 연결망 분석방법이어야만 하는가?
- 네트워크 연구는 insight가 부족하다?
  - 대부분 네트워크 연구는 다음과 같은 형식으로 기술
    - 어떤 노드의 연결 중심도가 높다.
    - 네트워크의 밀도가 높다.
  - 이와 같은 기술방법은 통계분석에서 기초통계 내용만 기술하는 수준
    - 어떤 노드의 연결 중심도가 높고, 이 노드의 연결 중심도가 높음으로써 나타나는 결과는 무엇인가?

# 2. SNA의 목적

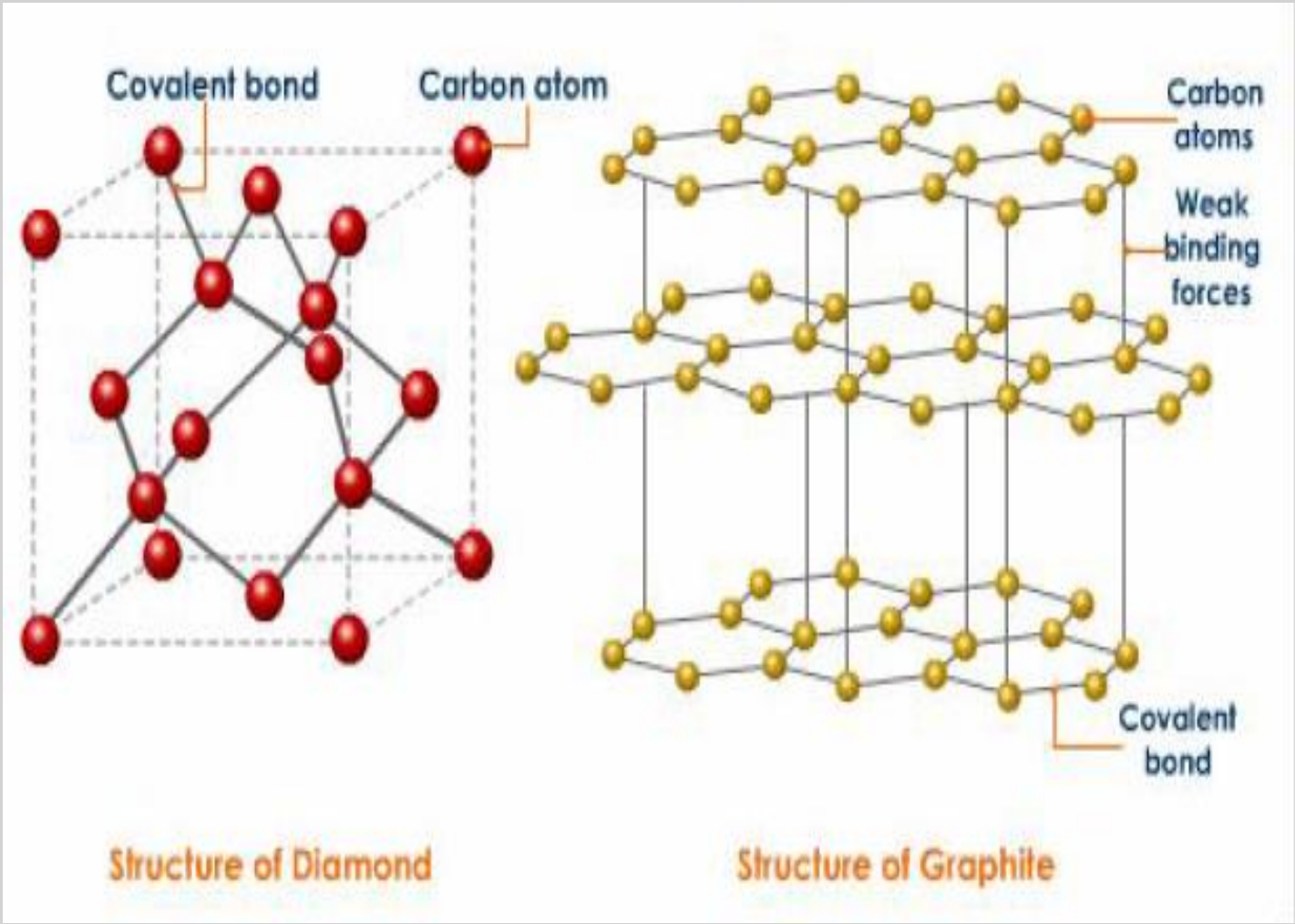
- 네트워크 분석의 목적
  - 구조(Structure)
  - 노드의 역할(Role)과 위치(position)



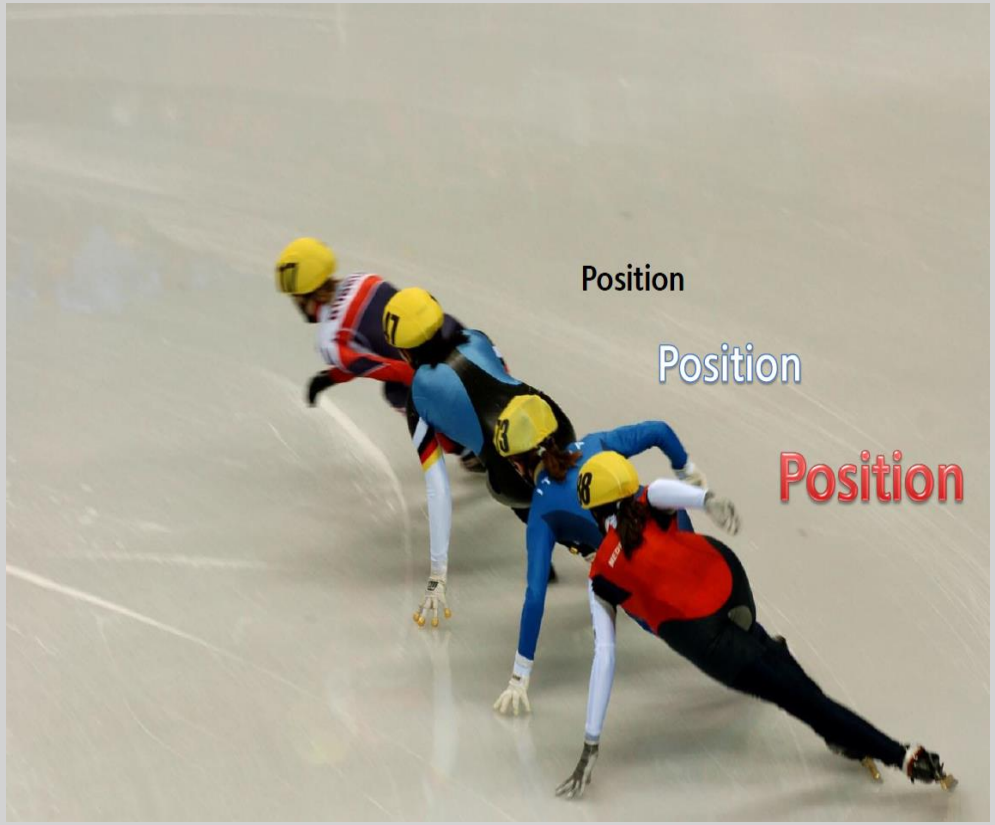
# 3-1. Unpacking SNA with three keys

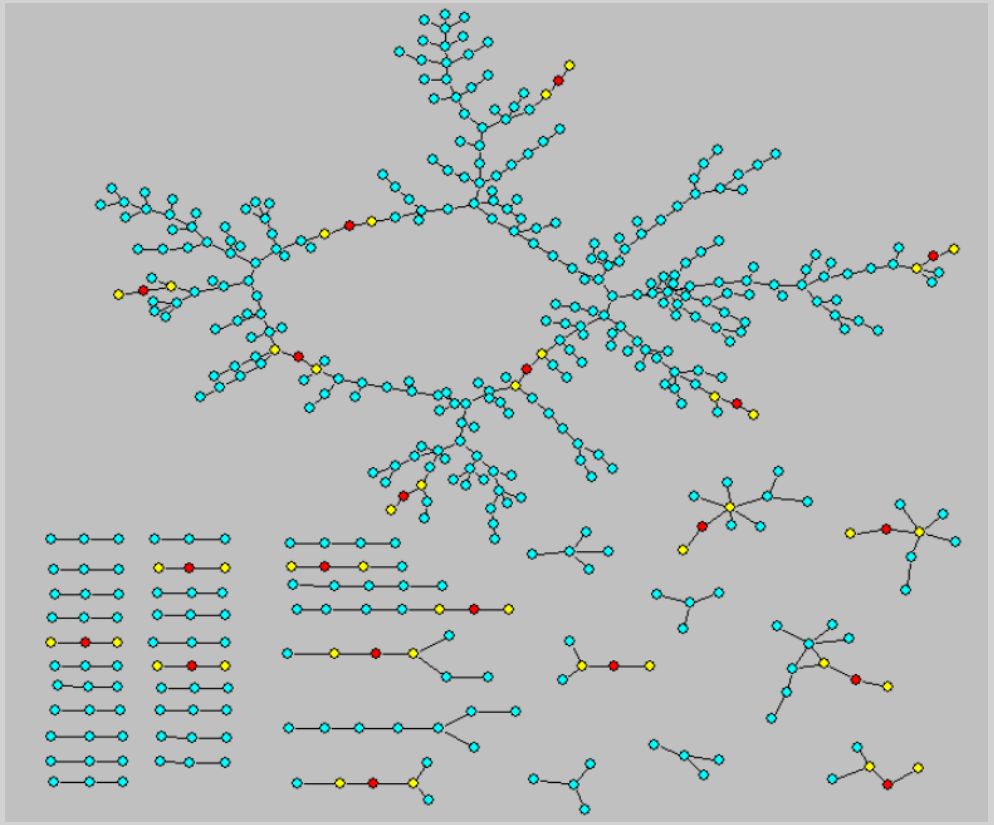
- 사람들은 연결망에 배태(embedded)되어 있음
- 연결망이란 '구조(Structure)'를 살펴보는 것
- 구조(Structure)내에서 노드(node) 간의 관계
  - 삶은 여러가지 사회적인 역할(role)이 교차되는 것
    - (어머니, 아내, 친구, 과장, 후배, 선배, 동생, 학생)
  - 역할(role)은 사람들간의 상호작용에 의해 나타나며, 대개 일정한 패턴(pattern)을 갖음
    - (위계적, 평등적, 일방향, 쌍방향, 송신, 수신 등)

# 3-2. Unpacking SNA with three keys - 구조

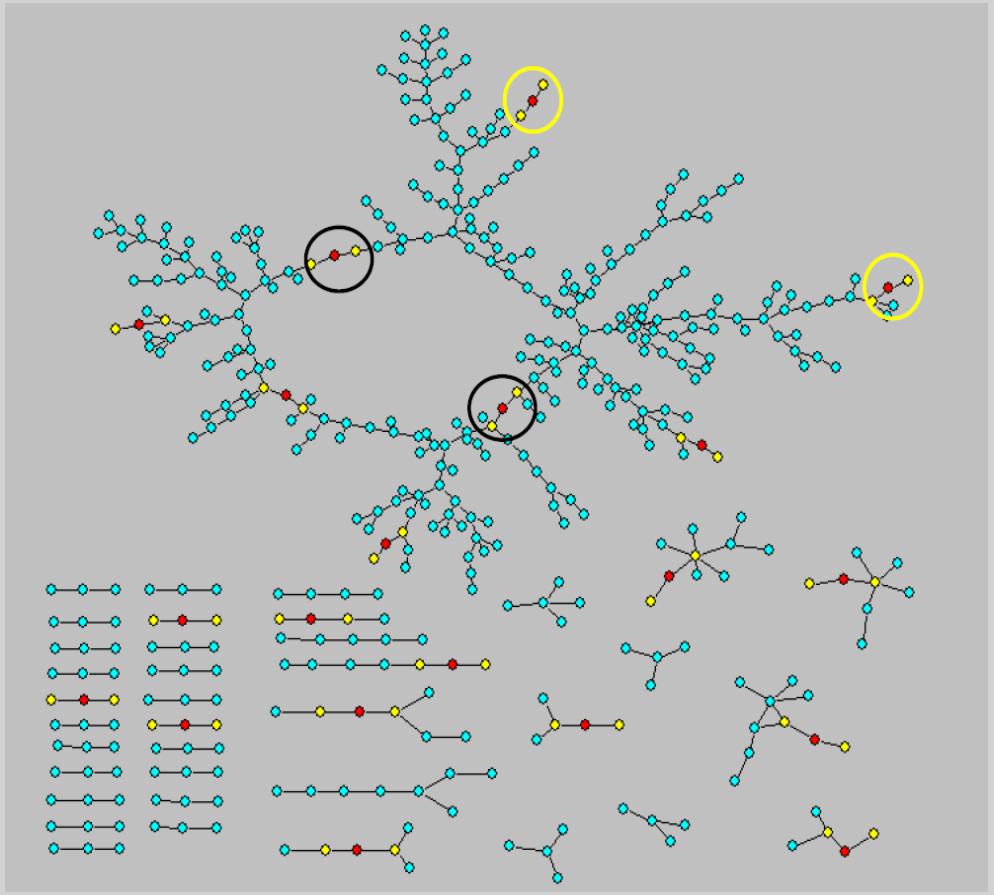


# 3-3. Unpacking SNA with three keys - 위치



# 3-4. Unpacking SNA with three keys - 역할

# 4-1. 자연과학&사회과학의 SNA 네트워크 구조(Structure)

- 자연과학 법칙으로서 네트워크 구조 : 데이터의 종류나 분석 맥락에 상관없이 발견되는 자연 법칙
  - Scale-free Network(Power law distribution)
  - Small World
- 사회과학 현상으로서 네트워크 구조 : 데이터의 종류나 분석 맥락에 따라 분석 바업을 달리해야 하는 내용
  - 연결망간 비교(시공간, 시간변화, 반복 측정)

# 4-2. SNA 구조(Structure) – Power law distribution

## Scale-free Networks & Power-law distribution

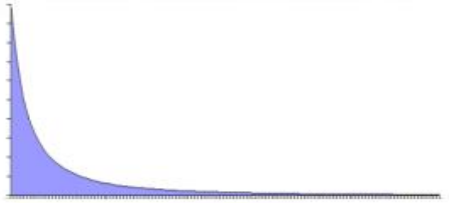
THE "NEW" SCIENCE OF NETWORKS

Duncan J. Watts
*Department of Sociology, Columbia University, New York, NY 10027; Santa Fe Institute, Santa Fe, New Mexico 97501; email: djw24@columbia.edu*

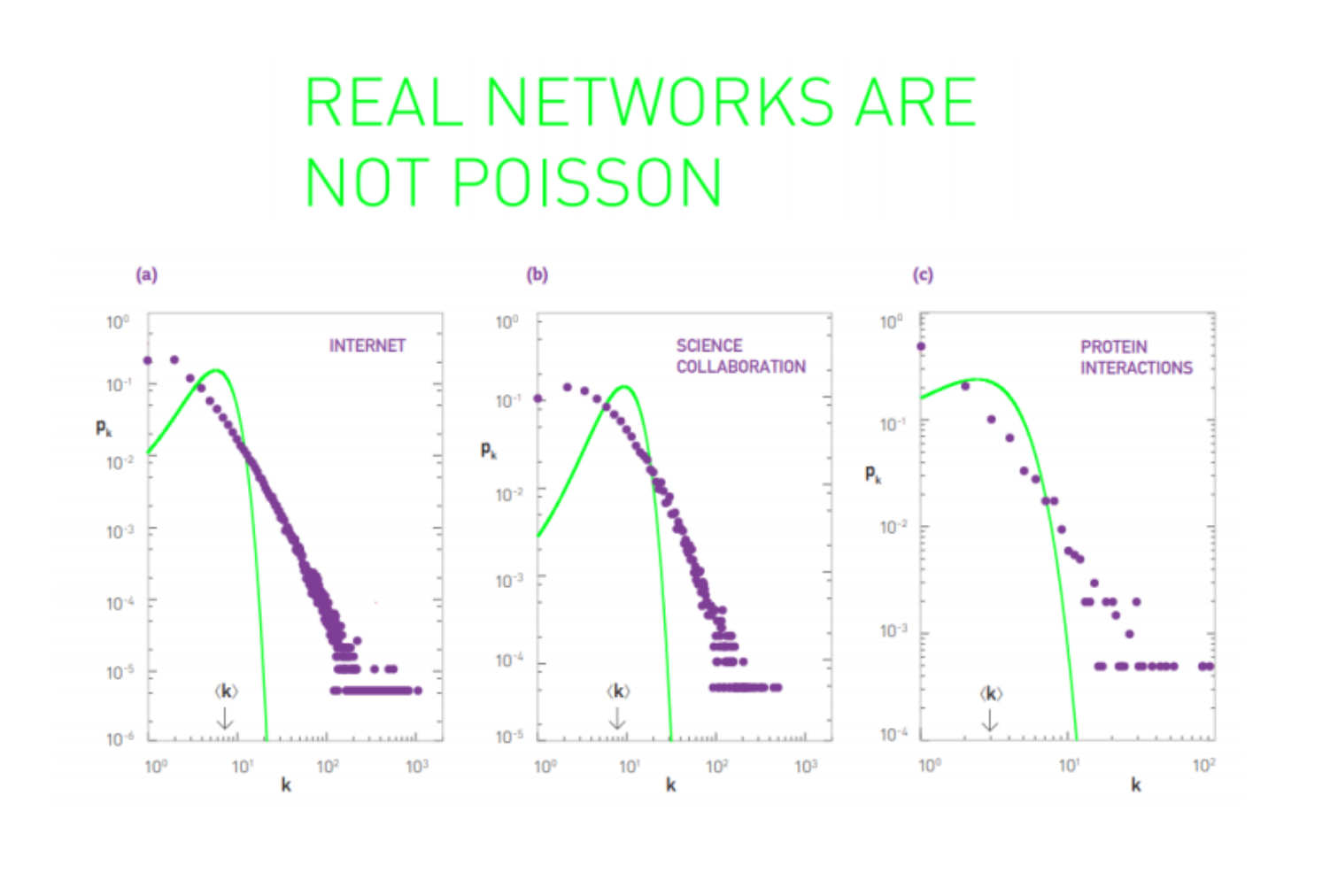
http://www.cornell.edu/video/emergence-of-network-science

### Scale-Free Networks

A separate development in the recent literature on networks has been the growing realization that in many real-world networks, the distribution of the number of network neighbors—the degree distribution—is typically right-skewed with a "heavy tail," meaning that a majority of nodes have less-than-average degree and that a small fraction of hubs are many times better connected than average. This qualitative description can be satisfied by several mathematical functions, but a particularly popular one in the current literature is a power law (Barabasi & Albert 1999), which has the asymptotic form



# 4-3. SNA 구조(Structure)
# – Power law distribution



# 4-4. SNA 구조(Structure)
# – Small world

## Small-world Networks

The "New" Science of Networks
Author(s): Duncan J. Watts
Source: *Annual Review of Sociology*, Vol. 30 (2004), pp. 243-270
Published by: Annual Reviews
Stable URL: http://www.jstor.org/stable/29737693

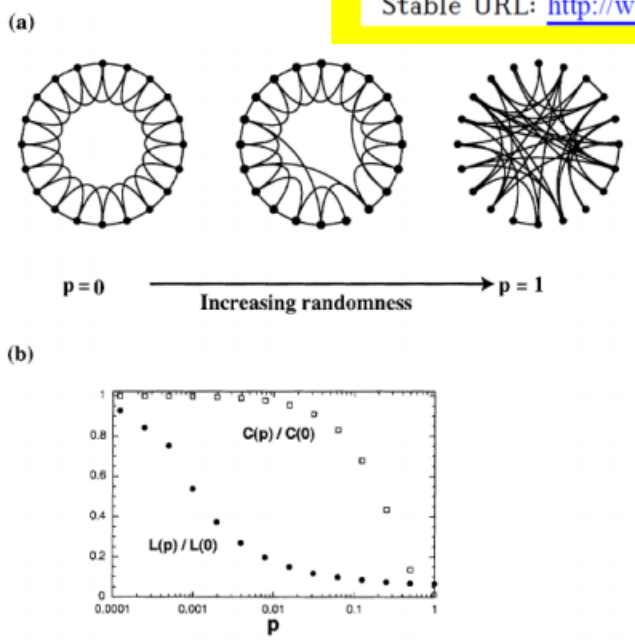



**Figure 1** (*a*) Schematic of the Watts-Strogatz model. (*b*) Normalized average shortest path length $L$ and clustering coefficient $C$ as a function of the random rewiring parameter $p$ for the Watts-Strogatz model with $N = 1000$, and $\langle k \rangle = 10$.

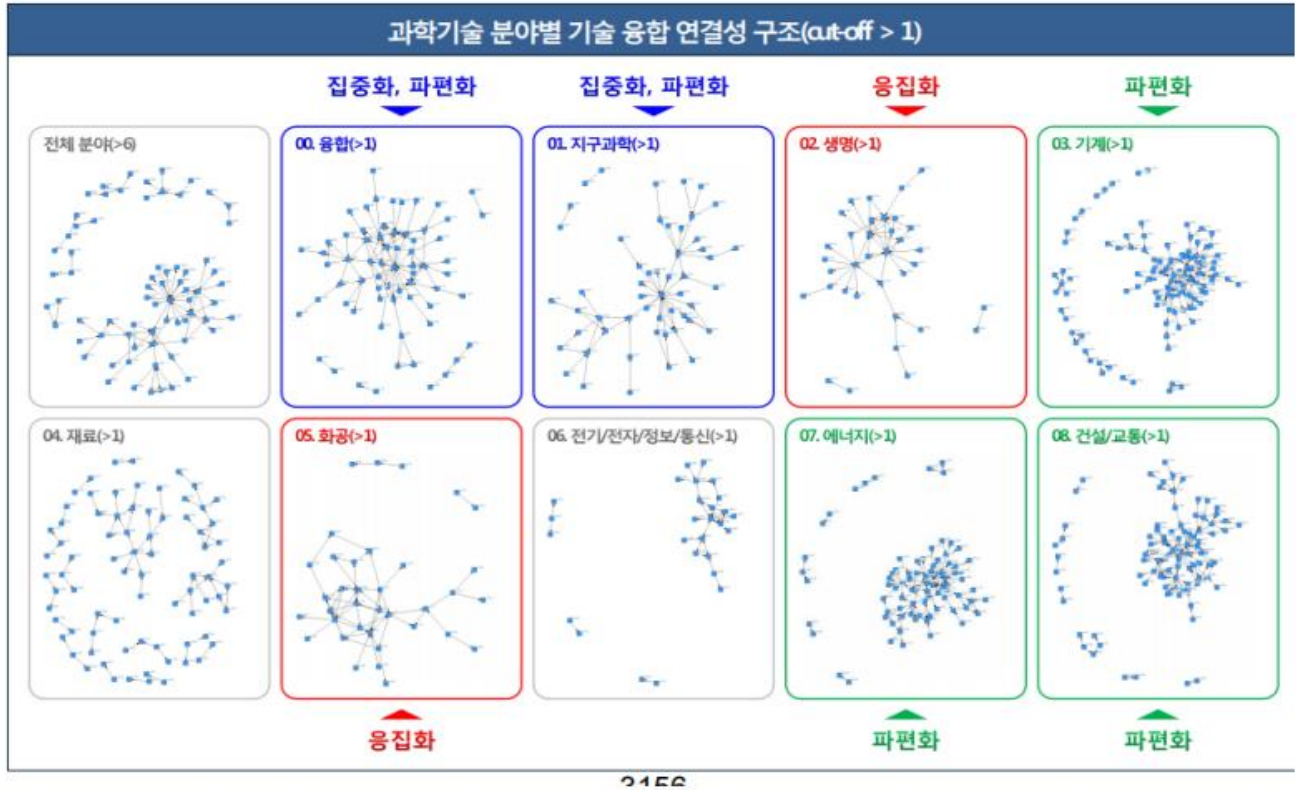
- √ Short Global Path Lengths
- √ High Local Clustering

# 4-5. SNA 구조(Structure)분석의 예시

정부출연연구기관 기술융합 연결망 구조의 동태적 분석

[DBpia(누리미디어)] Author 김경수, 조남욱 Publication 대한산업공학회 Date 2019/11/08 Journal 대한산업공학회 추계학술대회 논문집
Pages 3145-3164 Type 학술대회자료

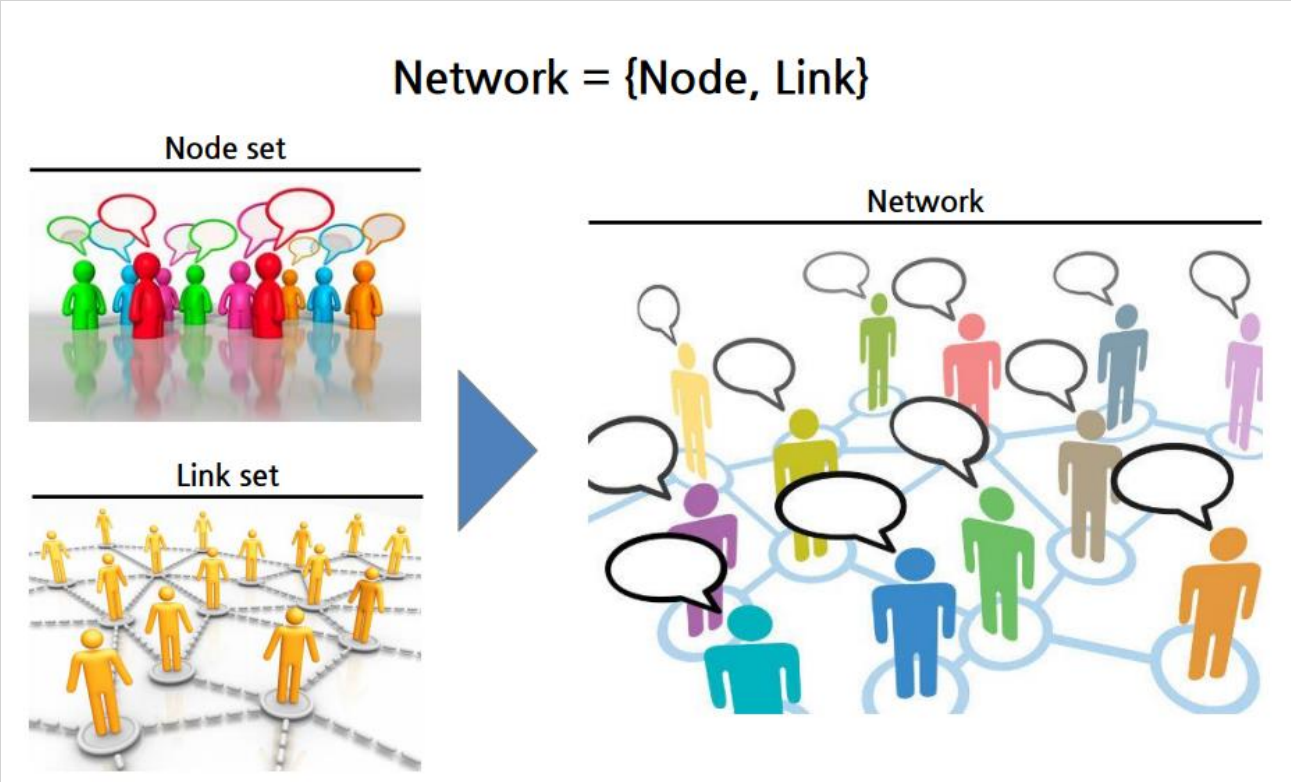
FullText | LSL | 결과저장 |



# 2. 네트워크의 구성 (노드와 링크)

# 1. 네트워크의 구성

- 네트워크는 노드와 링크로 구성되어 있음



# 2. 노드

- 노드(node)는 고유식별이 가능해야함
  - ID는 Text와 숫자 모두 가능
  - 단, 사람 이름이나 키워드를 ID로 할 경우에는 동명이인이나 동음이의어가 없는지 확인해야 함
- 노드(node)의 속성(attribute)정보를 포함할 수 있음
  - 분석 목적에 따라 다양하게 구성할 수 있음
  - 성별, 지역, 연령, 소속조직, 만족도, 찬/반 태도 등

# 3-1. 링크의 방향성

- **방향성(Direction)**
  - 링크는 방향성을 갖을 수 있음
  - 방향성이 있는 경우 그 데이터는 Source Node(행)와 Target Node(열)로 구분
    - 대각선이 일치하면 방향성 X
    - 대각선이 일치하지 않으면 방향성 O
    - Source node는 In-degree node의 수
    - Target node는 Out-degree node의 수
  - **Source node와 Target node가 동일한 Link(self loop) : 대각 행렬(Diagonal)**
  - 필요에 따라 방향성이 있는 데이터를 방향성이 없는 데이터로 변환할 수 있음(Symmetrize)

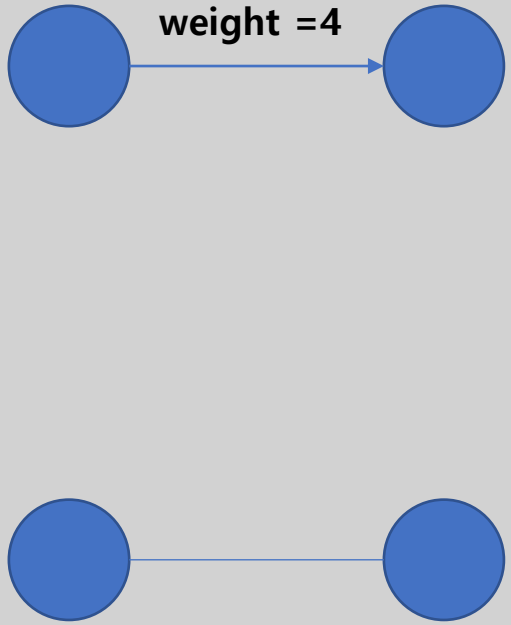


| 1 mode network | | | | | |
|---|---|---|---|---|---|
| | M1 | M2 | M3 | M4 | M5 |
| M1 | | | 1 | | |
| M2 | 1 | | | | 1 |
| M3 | | | | | |
| M4 | | 1 | | | |
| M5 | 1 | | 1 | | |

| 1 mode network | | | | | |
|---|---|---|---|---|---|
| | M1 | M2 | M3 | M4 | M5 |
| M1 | | 1 | 1 | | 1 |
| M2 | 1 | | | 1 | 1 |
| M3 | 1 | | | | |
| M4 | | 1 | | | 1 |
| M5 | 1 | 1 | | 1 | |

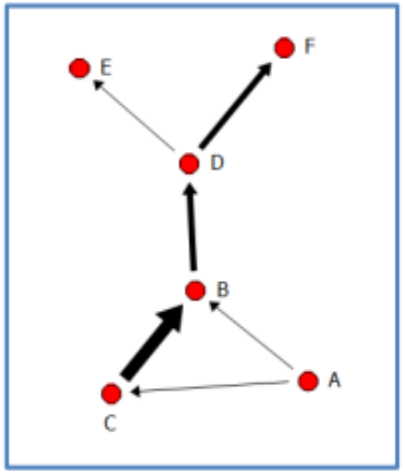
# 3-2. 링크의 가중치

- 가중치(Weight)
  - 가중치가 없는 데이터는 1과 0으로만 구분함
  - 필요에 따라 가중치가 있는 데이터를 가중치가 없는 데이터로 변환할 수 있음(Dichotomize)



# 4. 노드 간 링크를 표현방법

Network Map




Matrix

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | 0 | 1 | 1 | 0 | 0 | 0 |
| B | 0 | 0 | 0 | 2 | 0 | 0 |
| C | 0 | 3 | 0 | 0 | 0 | 0 |
| D | 0 | 0 | 0 | 0 | 1 | 2 |
| E | 0 | 0 | 0 | 0 | 0 | 0 |
| F | 0 | 0 | 0 | 0 | 0 | 0 |

Edge List

| Source | Target | Weight |
|---|---|---|
| A | B | 1 |
| C | B | 3 |
| A | C | 1 |
| B | D | 2 |
| D | E | 1 |
| D | F | 2 |

Linked List

| Source | Target1 | Target2 |
|---|---|---|
| A | B | C |
| B | D | |
| C | B | |
| D | E | F |

**Matrix**
- 네트워크 분석의 기초 형식임
- 데이터가 많은 경우에는 Matrix 형식으로 데이터를 구성하기 어려움

**Edge List**
- 대부분의 네트워크 데이터가 Edge List 형태로 수집되므로 별도의 데이터 가공 과정이 필요 없어 데이터 입력이 쉬움

**Linked List**
- 데이터 입력이 쉽지만, 링크의 Weight를 표현할 수 없음

# 5. 네트워크 분석시 고려할 점

- 링크의 방향성 : 방향성이 있는 네트워크인가? 방향성이 없는 네트워크 인가?
- 링크의 가중치 : 가중치를 고려할 것인가? 가중치를 고려하지 않을 것인가?
- 네트워크 입력 형태 : matrix type, edge list type, linked list type
- 연결망 관계 형태 : 1-mode network인가 2-mode network인가?
  - 1-mode : 사람과 사람 간의 관계, 기관과 기관 간의 관계, 키워드와 키워드 간의 관계
  - 2-mode : 사람과 클럽 간의 관계, 사람과 물품 간의 관계, 사람과 기관 간의 관계, 문서와 키워드 간의 관계

# 3. 네트워크 기초분석

# 1. 네트워크 기초분석
# - Density

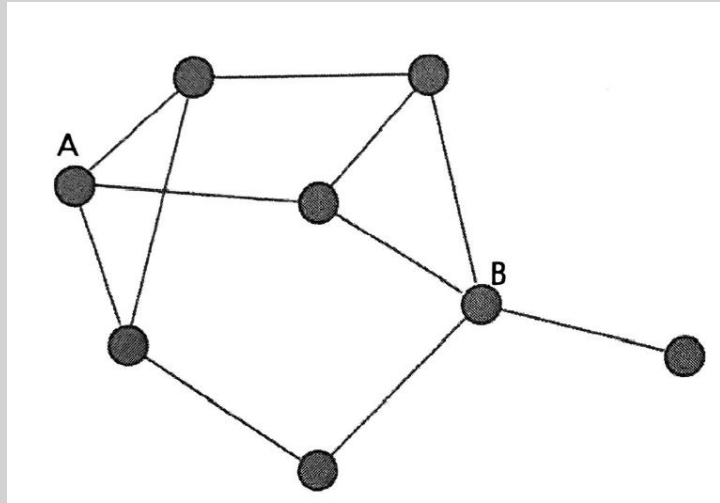
- **Density**
  - **Node size : 네트워크 내 전체 노드 수**
  - **Link size : 네트워크 내 전체 링크 수**
  - **Density : 네트워크 내 방생가능한 링크 중 존재하는 링크의 비율 / 노드 간의 전반적인 연결 정도의 수준**
  - **방향성이 있는 네트워크 :** $\frac{\#\ of\ links}{n(n-1)}$
  - **방향성이 없는 네트워크 :** $\frac{\#\ of\ pairs}{n(n-1)/2}$

# 2. 네트워크 기초분석
# - Degree

- **Degree**
  - **Degree : 하나의 노드가 직접 연결되어 있는 이웃 노드의 수**
    - 방향성이 있는 네트워크의 경우, 방샹성을 고려해 In-degre와 out-degree르 각각 측정
    - 가중치가 있는 네트워크의 경우, 링크값을 모두 합한 값으로 측정
  - **Average Degree : 네트워크 내 모든 노드 Degree의 평균값**
  - **Ego network : 하나의 노드(Ego node) 및 해당 노드와 직접 연결된 이웃 노드만으로 구성된 네트워크**
    - Alter(neighbor) : Ego node와 연결되어 있는 이웃 노드를 의미
    - Ego network size : alter의 개수를 의미

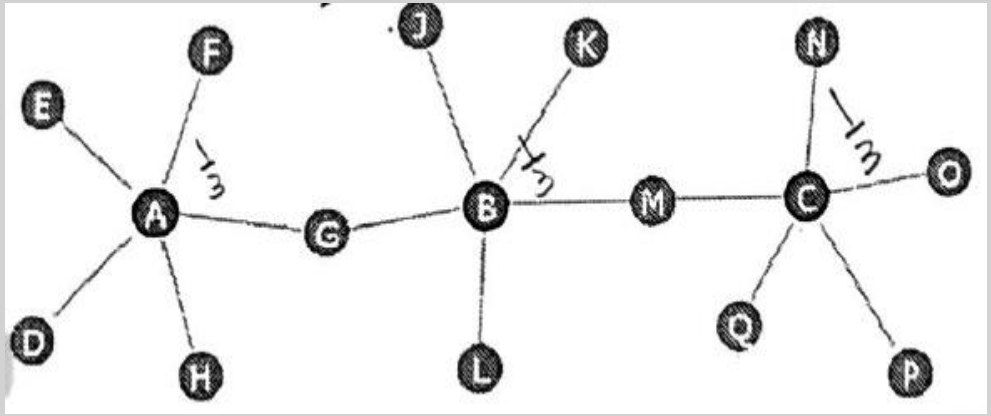
# 3. 네트워크 기초분석
# - Path & Distance

- **Path & Distance**
  - **Shortest Path : 두 노드가 연결될 있는 경로 중 가장 잛은 경로**
    - 경로의 거리는 지니가는 링크의 개수로 측정(A에서 B까지의 최단경로)
  - **Mean Distance : 모든 Shortest path의 평균**
    - 모든 Shortest path의 합 / Shortest path의 개수
      - 네트워크 A는 평균적으로 2.5단계로 연결
    - 네트워크 내에서 확산/전파 속도를 짐작할 수 있음
  - **Diameter**
    - 네트워크 내 모든 노드 간 Shortest path Distance중 가장 큰 값
    - 네트워크 내 크기를 짐작할 수 있는 지표
    - 끝에서 끝까지의 거리가 어느 정도가 되는지 파악하는데 필요



# 4. 중심성
# - Degree Centrality

- 특징
  - 해당 노드와 연결된 이웃 노드의 수
  - 해당 노드의 직접적인 영향력 파악
  - 방향성이 있는 네트워크의 경우 In/Out이 각각 측정됨
- 산출방법
  - $\frac{\text{해당노드가 가진값}}{\text{생성가능한 최대값}} = \frac{\text{해당노드의 } Degree}{\text{전체노드수}-1} = \frac{d(n_i)}{g-1}$
  - g = 전체 노드의 수
  - d(ni) = 노드 n의 degree




- A, B, C의 Degree = 5/(16-1) = 1/3
- 그렇지만 B가 더 중요
  - Degree centrality는 이러한 것을 반영할 수 없다는 한계를 갖고 있음

# 4. 중심성
# - Closeness Centrality

- 특징
  - 노드 간의 거리를 바탕으로 중심성을 측정하는 방식
  - 중요한 노드일 수록 다른 노드까지 도달하는 경로가 짧을 것이라고 가정
  - 직접 연결되어 있는 노드들을 통해 접근할 수 있는 간접적인 노드 간의 관계까지 파악하기 때문에, 연결중심도와 비교할 때 네트워크 전체적인 범위 내에서 노드의 중심도를 파악
  - 방향성이 있는 네트워크의 경우 In/Out이 각각 측정됨
- 산출방법
  - 거리가 짧을수록 중심성이 크기 때문에, 측정값의 역수
  - 최대값으로 정규화하며, 자신을 제외한 모든 노드와 연결되어 있을 때, 최대값을 가짐
  - $\left[\frac{\text{해당노드가 가진값}}{\text{생성가능한 최대값}}\right]^{-1} = \left[\frac{\text{해당노드와 다른노드와의 거리 총합}}{\text{전체노드수}-1}\right]^{-1} = \frac{\sum_{j=1}^{g} d(n_i n_j)}{g-1}^{-1} = \frac{g-1}{\sum_{j=1}^{g} d(n_i n_j)}$

# 4. 중심성
# - Betweenness Centrality

- 특징
  - 네트워크 내 하나의 노드가 다른 노드들 사이에 위치하는 정도
- 산출방법
  - 다른 노드와의 최단경로에서 해당 노드를 거칠 확률을 고려
  - $\frac{\text{해당노드가 가진값}}{\text{생성가능한 최대값}} = \frac{(\text{해당노드를 경유하는 최단경로} * \text{경유할 확률})\text{의 총합}}{\text{네트워크 내 모든 경로 수}} = \frac{\sum_{j>k} g_{jk}(n_i)/g_{jk}}{[\frac{(g-1)(g-2)}{2}]}$
  - $g_{jk}$ : 노드 j와 k를 연결하는 최단경로의 개수
  - $g_{jk}(n_i)$ : 노드 j와 k를 연결하는 최단경로 중 노드i를 거치는 경로의 수
  - $\sum_{j>k} g_{jk}(n_i)/g_{jk}$ : 각각의 node pair의 최단경로상 노드i가 포함될 확률의 누적합
  - $[\frac{(g-1)(g-2)}{2}]$ : ni를 포함하지 않는 모든 노드 쌍의 수

# 4. 중심성
# - Eigenvector Centrality

- 특징
  - degree centrality의 원리에서 확장된 것
  - 이웃이 많을 뿐만 아니라, 중심성이 높은 이웃과 연결되어 있을수록 중심성이 높음

# 2mode-network & Text network

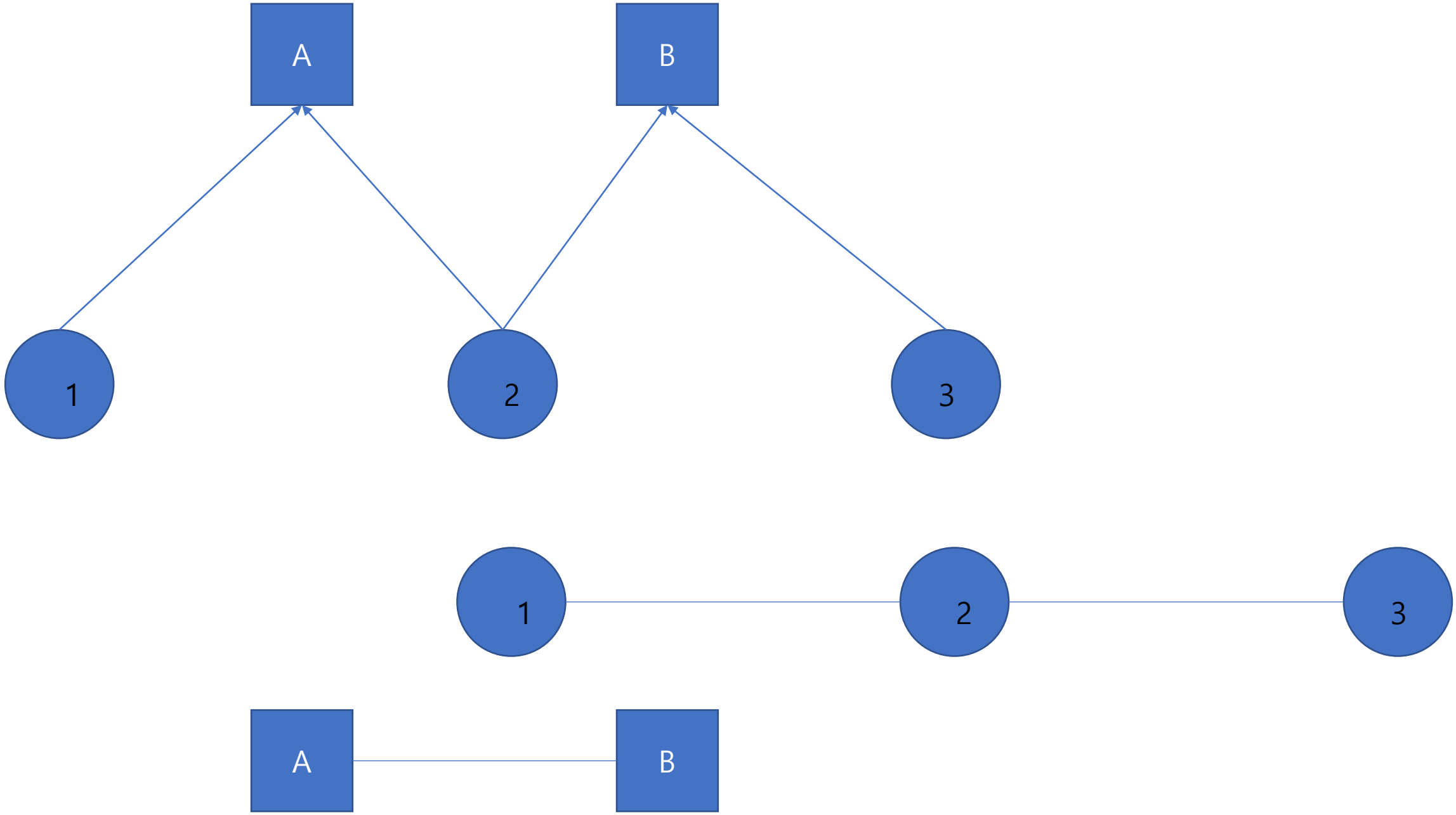
# 1. 2mode network

# 1. 2mode network의 구조

| 1 mode network | | | | | |
|---|---|---|---|---|---|
| | M1 | M2 | M3 | M4 | M5 |
| M1 | | | | | |
| M2 | | | | | |
| M3 | | | | | |
| M4 | | | | | |
| M5 | | | | | |

| 2 mode network | | | |
|---|---|---|---|
| | N1 | N2 | N3 |
| M1 | | | |
| M2 | | | |
| M3 | | | |
| M4 | | | |
| M5 | | | |

| Bipartite network(이분 그래프) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | M1 | M2 | M3 | M4 | M5 | N1 | N2 | N3 |
| M1 | | | | | | | | |
| M2 | | | | | | | | |
| M3 | | | | | | | | |
| M4 | | | | | | | | |
| M5 | | | | | | | | |
| N1 | | | | | | | | |
| N2 | | | | | | | | |
| N3 | | | | | | | | |


A
B
1
2
3
1
2
3
A
B

# 2. 2mode network의 특징

- 2-mode network는 관계의 방향성(direction)을 정의하기 곤란한 경우가 대다수
- 2-mode network에서 링크의 강도(weight)는 정의할 수 있음
- 2-mode network는 2개의 1-mode network로 변환할 수 있음
  - 2-mode network에서 1-mode network로 변환시 링크의 의미의 변화
  - Connection -> Conversion -> Similarity
  - A = 사람 X 책
  - A의 전치행렬 = 책 X 사람
  - A * A의 전치행렬 = (사람 X 책)(책 X 사람) = (사람 X 사람)
  - A 의 전치행렬 * A = (책 X 사람)(사람 X 책) = (책 X 책)

# 3. 2mode network의 접근방법

- 변환 접근(Coversion approach)
  - 학생 X 이벤트로 구성된 매트릭스
    - 학생 X 학생 매트릭스
    - 이벤트 X 이벤트 매트릭스
  - 한 모드가 다른 모드에 비해 더 큰 연구적 관심일 때
    - 예를 들어, 이벤트가 여성들 간의 관계를 탐구하는 데 활용될 수 있음
  - 그러나 변환으로 인한 정보 상실에 대한 우려는 존재
- 직접 접근(Direct approach)
  - 여성과 이벤트를 모두 노드로 취급
  - 두 모드에 동일하게 연구적 관심이 있거나 모드 간 대응에 관심이 있는 경우

# 4. 2mode network의 변환방법

- N * M -> N * N / M * M
- Match method
  - Jaccard coefficient : 키워드가 문서에 등장했는지 여부만 고려하여 두 키워드의 유사성을 측정. 문서에 등장한 횟수는 고려하지 않음 / 0~1사이의 값으로 나타나며, 1일 경우 두 키워드가 문서에 항상 함께 등장했다는 것을 의미
  - Jaccard coefficient = $\frac{A\text{와 }B\text{가 함께 등장한 문서 수}}{A\text{ 또는 }B\text{가 등장한 문서수}} = \frac{a}{a+b+C}$
  - Ochiai = $\frac{A\text{와 }B\text{가 함께 등장한 문서 수}}{(A\text{가 등장한 문서수} * B\text{가 등장한 문서수})\text{의 제곱근}} = \frac{a}{\sqrt{(a+b)*(a+c)}}$

# 4. 2mode network의 변환방법

- N * M -> N * N / M * M (전치행렬을 이용하여 2mode를 1mode로 변경)
  - Jaccard coefficient : 키워드가 문서에 등장했는지 여부만 고려하여 두 키워드의 유사성을 측정. 문서에 등장한 횟수는 고려하지 않음 / 0~1사이의 값으로 나타나며, 1일 경우 두 키워드가 문서에 항상 함께 등장했다는 것을 의미
  - Jaccard coefficient = $\frac{A\text{와 }B\text{가 함께 등장한 문서 수}}{A\text{ 또는 }B\text{가 등장한 문서수}} = \frac{a}{a+b+C}$
  - Inner product : 두 키워드가 문서에 등장한 횟수를 고려하여 우사성을 측정 / 0이상의 정수로 나타남 / 두 키워드가 동시에 등장한 문서가 많아야 값이 커지며, 같은 문서에 등장횟수가 많아야 값이 커짐
  - Inner product = (A의 등장횟수 * B의 등장횟수)의 합

# 4. 2mode network의 변환방법

- N X M -> N X N / M X M (전치행렬을 이용하여 2mode를 1mode로 변경)
  - Cosine similarity : en 키워드가 문서에 등장한 횟수를 고려하여 유사성을 측정 / 0~1사이의 값으로 나타나며, 1일 경우 두 키워드가 똑같은 문서에 똑같은 빈도로 함께 등장했다는 것을 의미
  - Cosine Similarity = $\frac{(A\text{의 등장횟수} * B\text{의 등장횟수})\text{의 합}}{(A\text{의 등장횟수 제곱합의 제곱근}) * (B\text{의 등장횟수 제곱합의 제곱근})}$
  - = $\frac{\sum_{k=1}^{n} C_{ik} C_{jk}}{\sqrt{\sum_{k=1}^{n} C_{ik}^2} \sqrt{\sum_{k=1}^{n} C_{jk}^2}}$

# 5. 텍스트 네트워크 분석

# 1. 네트워크로서의 텍스트

## History [ edit ]

Social network analysis has its theoretical roots in the work of early sociologists such as Georg Simmel and Emile Durkheim, who wrote about the importance of studying patterns of relationships that connect social actors. Social scientists have used the concept of "social networks" since early in the 20th century to connote complex sets of relationships between members of social systems at all scales, from interpersonal to international. In the 1930s Jacob Moreno and Helen Jennings introduced basic analytical methods.[10] In 1954, John Arundel Barnes started using the term systematically to denote patterns of ties, encompassing concepts traditionally used by the public and those used by social scientists: bounded groups (e.g., tribes, families) and social categories (e.g., gender, ethnicity). Scholars such as Ronald Burt, Kathleen Carley, Mark Granovetter, David Krackhardt, Edward Laumann, Anatol Rapoport, Barry Wellman, Douglas R. White, and Harrison White expanded the use of systematic social network analysis.[11] Even in the study of literature, network analysis has been applied by Anheier, Gerhards and Romo,[12] Wouter De Nooy,[13] and Burgert Senekal.[14] Indeed, social network analysis has found applications in various academic disciplines, as well as practical applications such as countering money laundering and terrorism.

단어와 단어의 연결구조

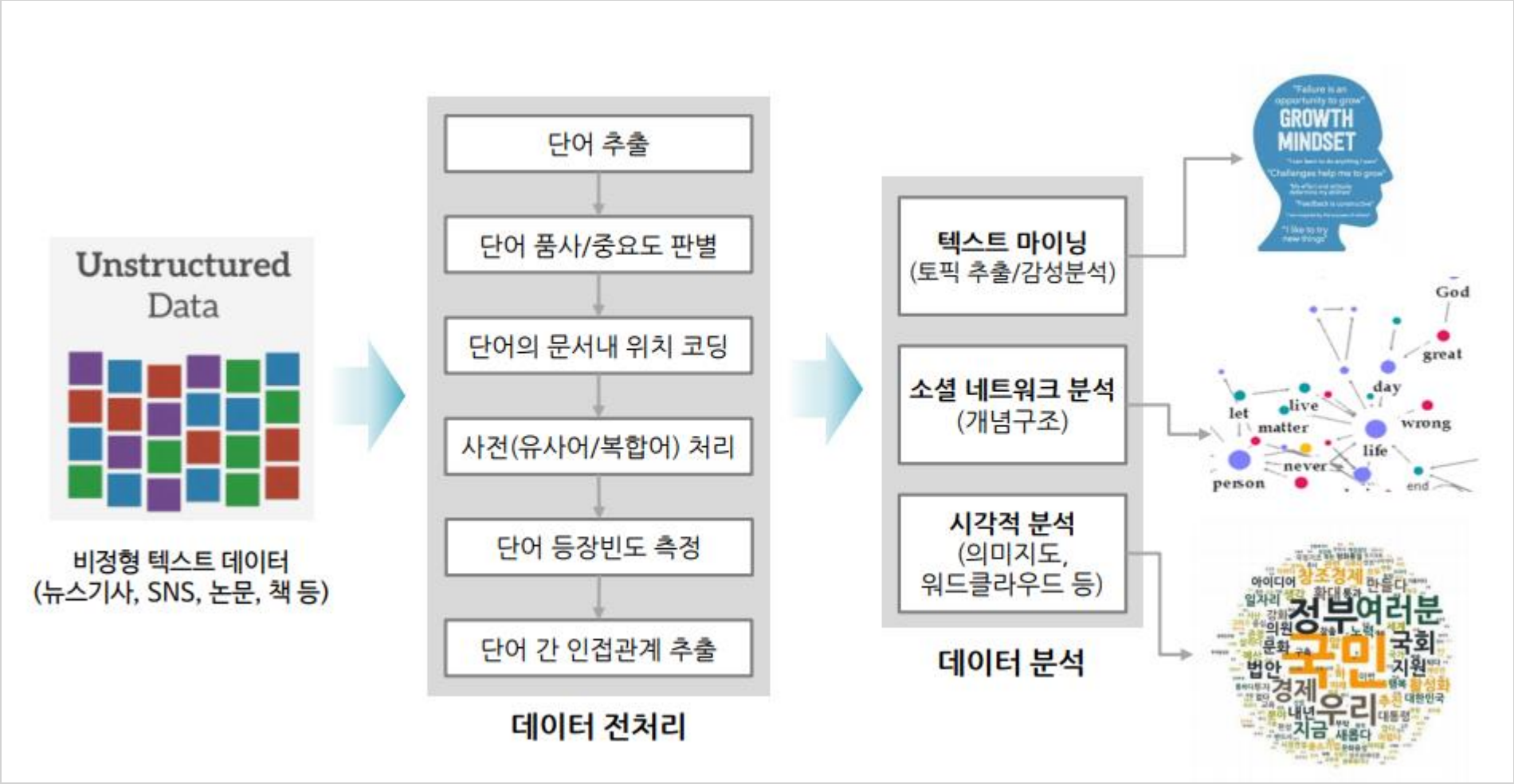

구조의 재현

중심구조와 응집구조

# 2. 텍스트 네트워크 분석의 기본 가정

- 텍스트에 내재된 언어와 지식은 단어와 그들간의 관계 네트워크로 모델링 될 수 있음
- 텍스트 네트워크 내에서 개념들의 위치와 연결패턴을 통해 텍스트의 의미 또는 중요한 주제에 대해 이해할 수 있음

# 3. 텍스트 네트워크 분석의 절차

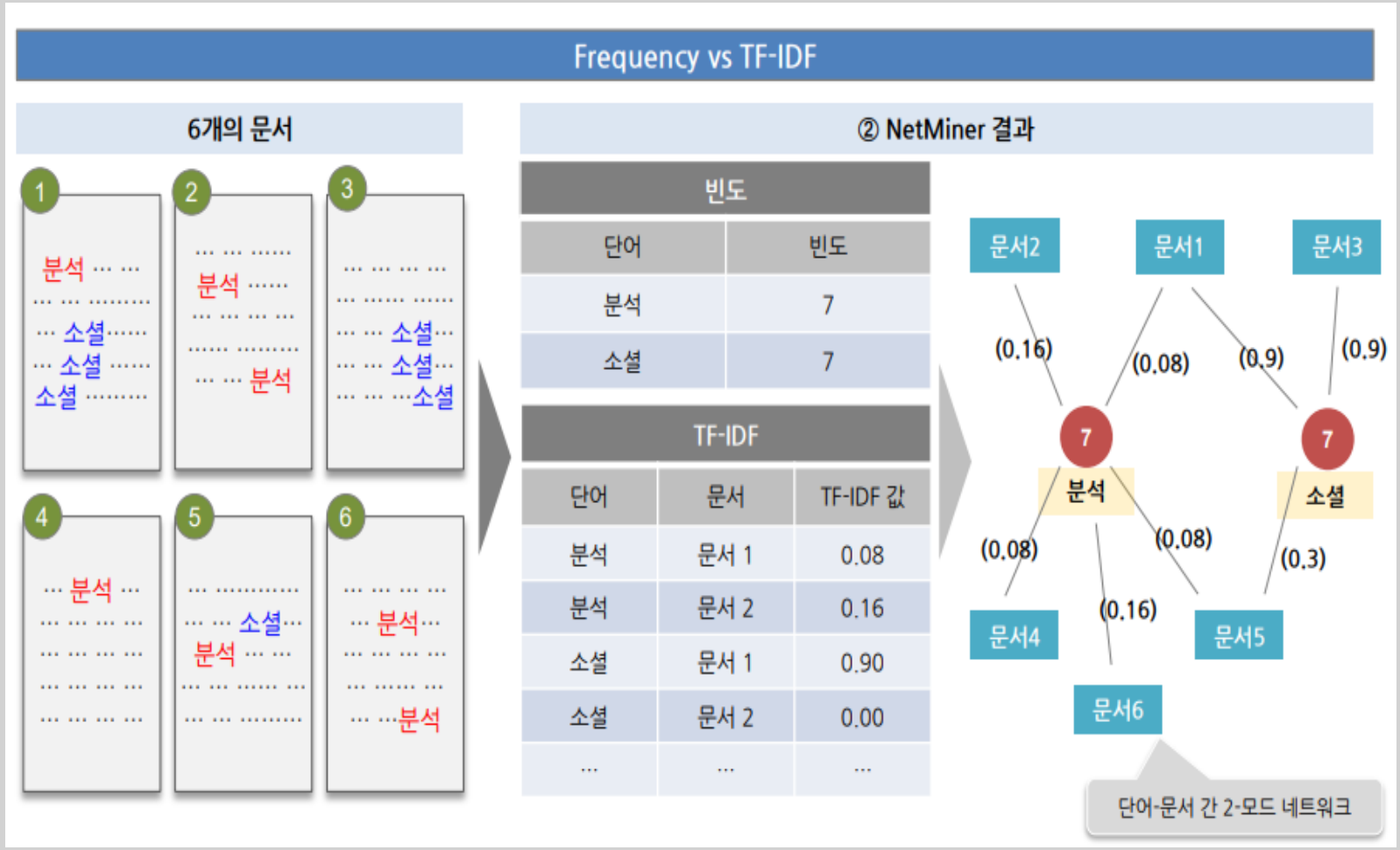


# 4. 단어의 중요도 측정

- 단어의 중요도는 문서 내 또는 문서 간 등장 빈도, 총 등장 빈도 등 다양한 방법으로 측정할 수 있음
- 그 중 TF-IDF(Term Frequency – Inverse Document Frequency) 는 '어떤 문서에서의 특정 단어의 중요도'를 측정할 수 있어 매우 널리 활용되는 방법
  - TF-IDF는 어떤 단어가 특정 문서에서 사용된 정도(TF)와 이 단어가 사용된 문서의 수(DF)를 이용
  - DF 값의 역수가 바로 IDF 입니다.
  - TF-IDF 는 TF와 IDF 를 곱한 값
  - 어떤 단어가 특정 문서에 가지는 TF-IDF는 TF가 높을 수록, DF가 낮을 수록 높게 나타납니다. 그리고 TF가 낮고, DF가 높으면 낮게 나타남
  - A라는 단어가 문서(가)에서만 많이 사용되었다면, 이 단어는 문서(가)에서 TF가 높지만 DF는 낮으므로, 문서(가)에서의 단어 A의 TF-IDF 는 높게 나타남
  - 여기에서 A는 문서(가)에서만 사용되어 다른 문서와 문서(가)를 구분 지어줄 수 있는 특징적인 단어로 볼 수 있음
  - 그래서 각 문서에서 TF-IDF 가 높은 단어는 그 문서의 특징적인 내용을 이해할 수 있는 단어라고 해석
- 특정 문서에서의 단어 빈도가 높을 수록, 그리고 전체 문서 중 그 단어가 등장한 문서가 적을 수록 TF-IDF 값이 높게 나타남
- 즉, TF-IDF 값이 높은 단어는 해당 문서 내에서 핵심적인 메시지를 담고 있을 확률이 높음

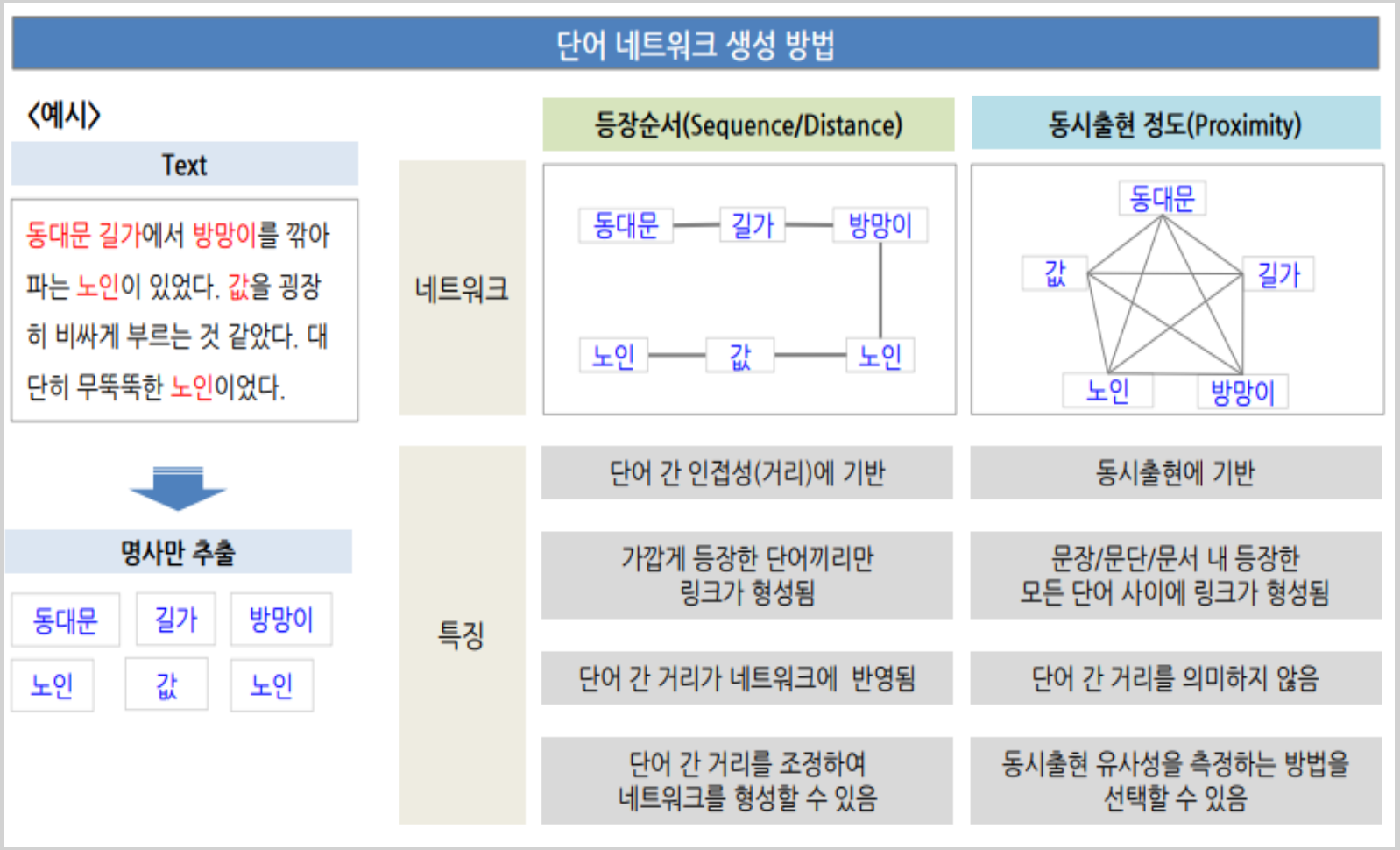
# 4. 단어의 중요도 측정

- 빈도(Frequency)는 전체 문서에서 특정 단어가 등장한 횟수이며, 단어마다 각각의 고유 값이 산출
- TF-IDF는 문서별 단어의 가중치이며, 같은 단어라도 문서에 따라 다른 값이 산출



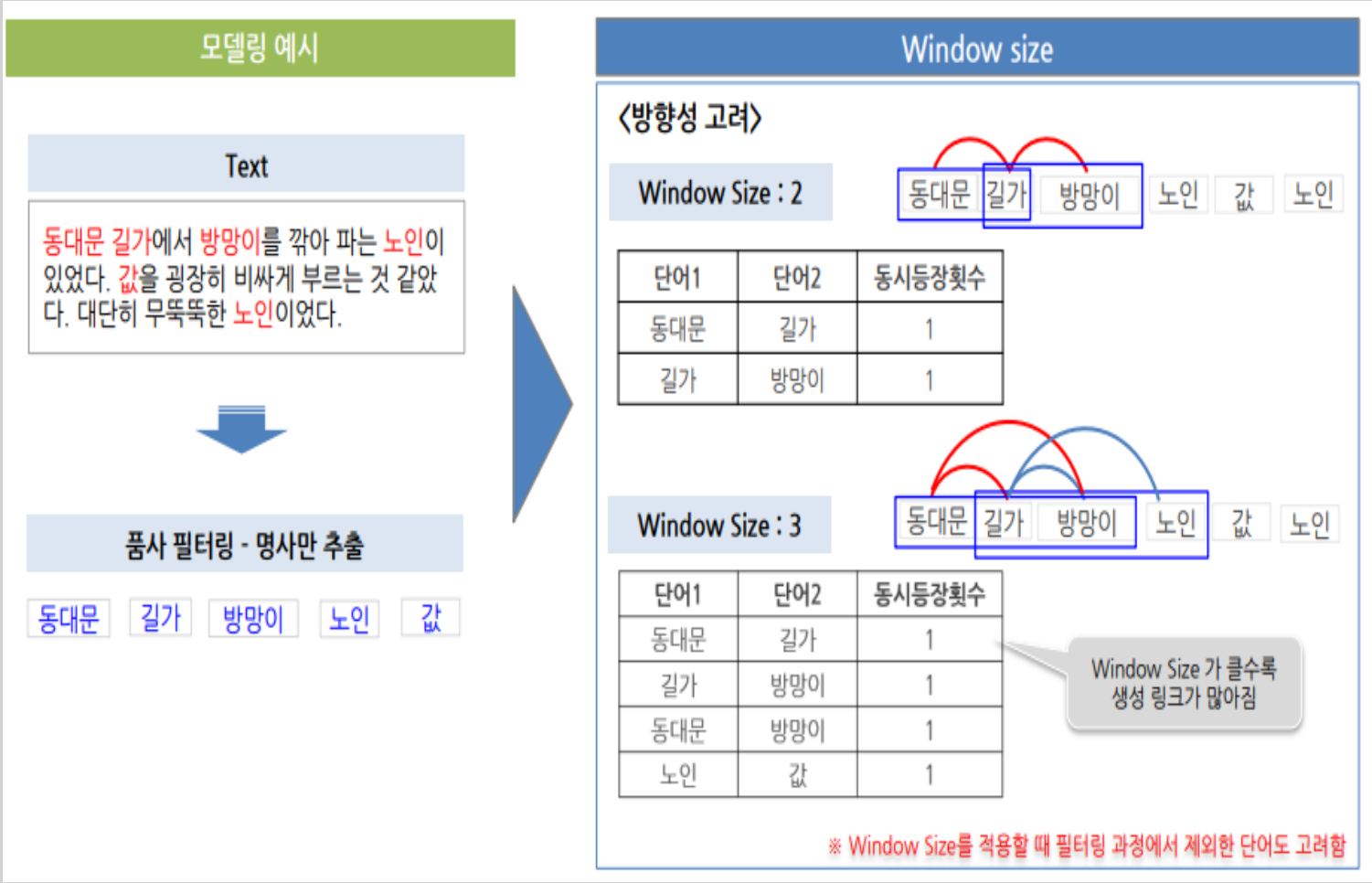
# 4. 단어의 중요도 측정

- 텍스트에서 추출한 단어들로 네트워크를 구성하는 방법은 두 가지로 구분할 수 있음
- 등장순서에 기반한 방법과 동시출현에 기반한 방법이 있음



# 4. 단어의 중요도 측정

- ① 등장순서에 기반한 모델링
- Window Size: 얼마나 인접한 단어 사이에 링크를 형성할지 결정할 수 있음
- 방향성: 등장 순서에 방향성을 부여할 수도 있음
- 문서의 스토리를 반영하여 네트워크를 형성하는 경향이 있음



# 4. 단어의 중요도 측정

- 모델링을 할 때, 단어간 링크를 형성하는 단위를 문장/문단/문서 중 선택해야 함

**Co-occurrence Unit**

**문서(Document) 구분**

<txt 파일> 파일 1개가 1개의 문서

문서1
문서2
문서3

<엑셀 파일> 행 1개가 1개의 문서

| | Abstract |
|---|---|
| 2 | It is important to have a good cost estimate in order to budget a |
| 3 | Information retrieval algorithms attempt to match a user's desc |
| 4 | In this thesis, I present three novel heuristic algorithms for lean |
| 5 | Recent years have seen a growing recognition of the role of go |
| 6 | This study explores national security intelligence education with |
| 7 | A particularly successful line of research for numerical optimiz |
| 8 | Computational natural language understanding and generation |
| 9 | Limited research exists concerning which machine learning alg |
| 10 | Following World War II, President Truman disbanded the Offic |

문서1
문서2

**문단(Paragraph), 문장(Sentence) 구분**

- 문단 구분 : 줄바꿈 표시(개행)
- 문장 구분 : 마침표, 물음표, 느낌표 등 문장을 끝맺는 문장 부호

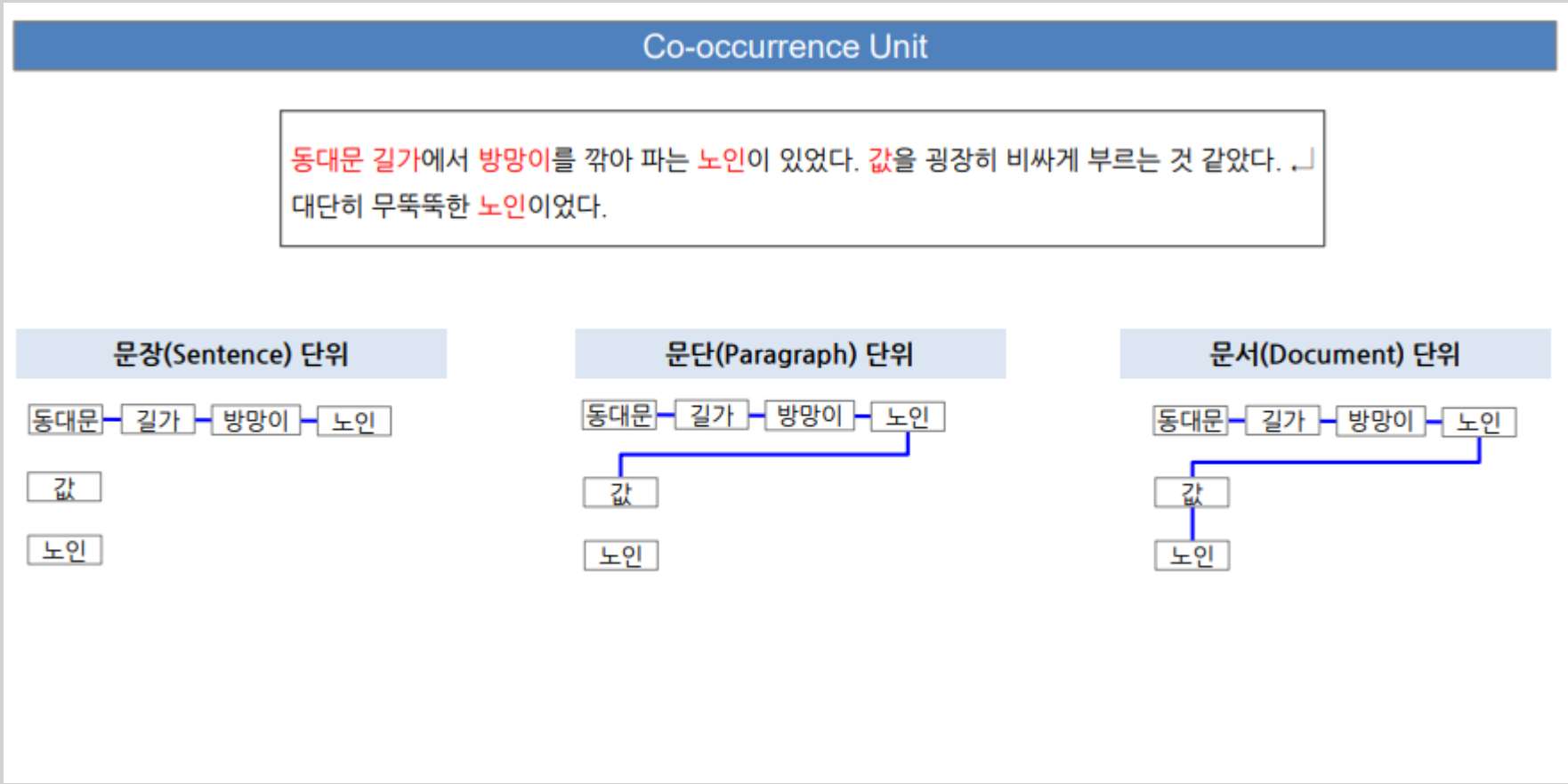
<예시>

동대문 길가에서 방망이를 깎아 파는 노인이 있었다. 값을 굉장히 비싸게 부르는 것 같았다. ↵
대단히 무뚝뚝한 노인이었다.

**문장 : 3개**
**문단 : 2개**
**문서 : 1개**

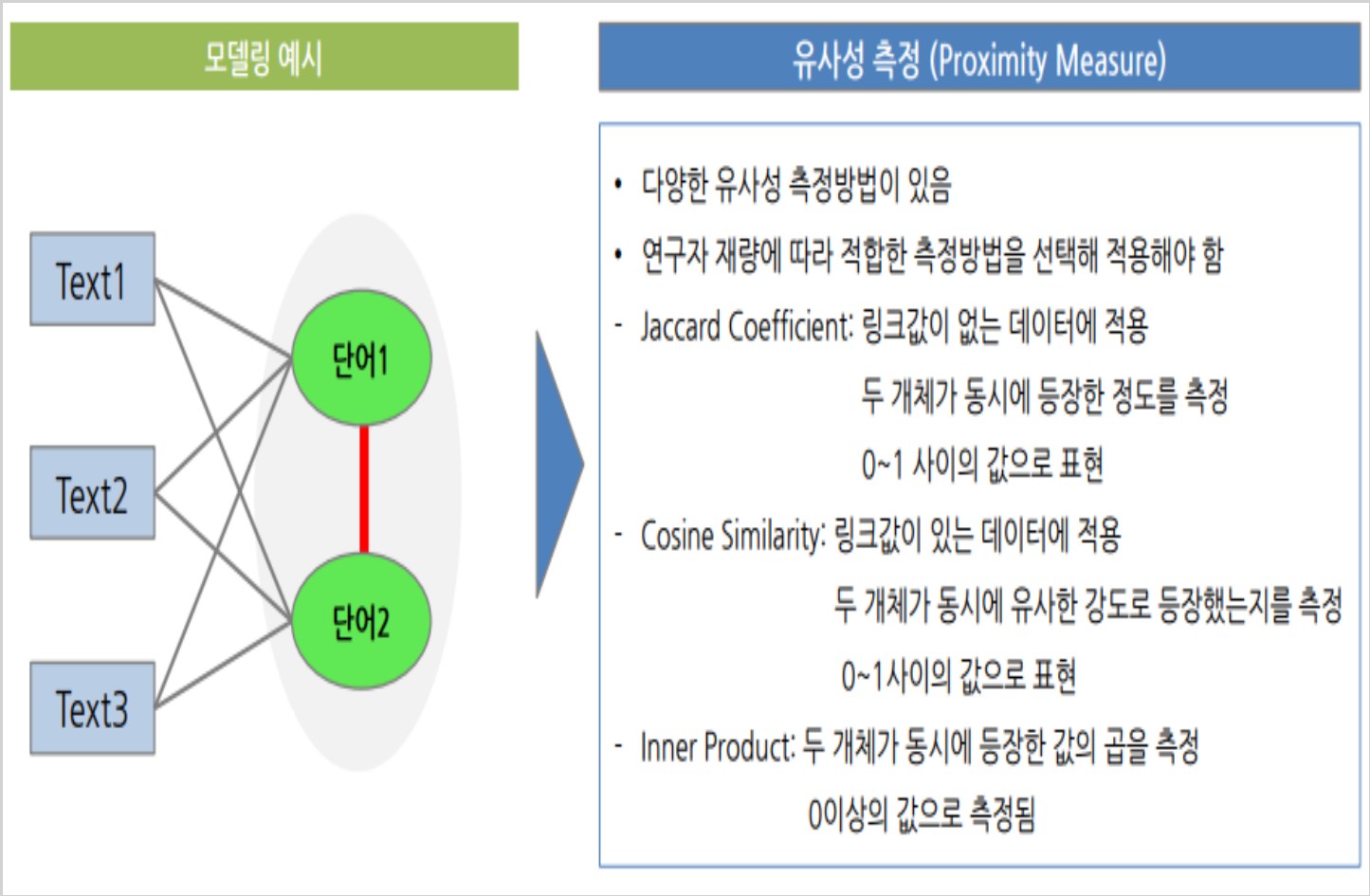
# 4. 단어의 중요도 측정

- 전체 문서의 길이와 형식, 주제를 담고 있는 단위, 분석 목적 등을 고려하여 선택해야 함



# 4. 단어의 중요도 측정

- ② 동시출현 유사성에 기반한 모델링
- 동시출현 단위를 문장/문단/문서로 구분하여 각각 측정할 수 있음
- 단어-문장/문단/문서 간 2모드 네트워크 데이터를 1모드 유사성 네트워크로 변환하는 방법
- 비슷한 주제를 의미하는 단어들끼리 링크가 형성되는 경향이 있음



# 감사합니다.