

TEXT MINING for PRACTICE

Python을 활용한 비정형 데이터 분석 - WEEK 07
동시출현분석 & 단어 네트워크 분석 & 페이지 랭크

연세대학교 | 서중원

지난 시간...

단어 주머니 (Bag of Words)

- ▶ 단어들 사이의 동시 출현을 연관성으로 취급하여, 단어의 연관성을 파악하는 방법
- ▶ 상용 소셜 미디어 분석 솔루션에서 제공하는 가장 기본적인 분석 형태 (Social Matrix)
- ▶ 활용분야 : 브랜드 이미지 조사, 트렌드 분석, 여론조사, 마케팅 모니터링

TF-IDF (TF-Inverse Document Frequency)

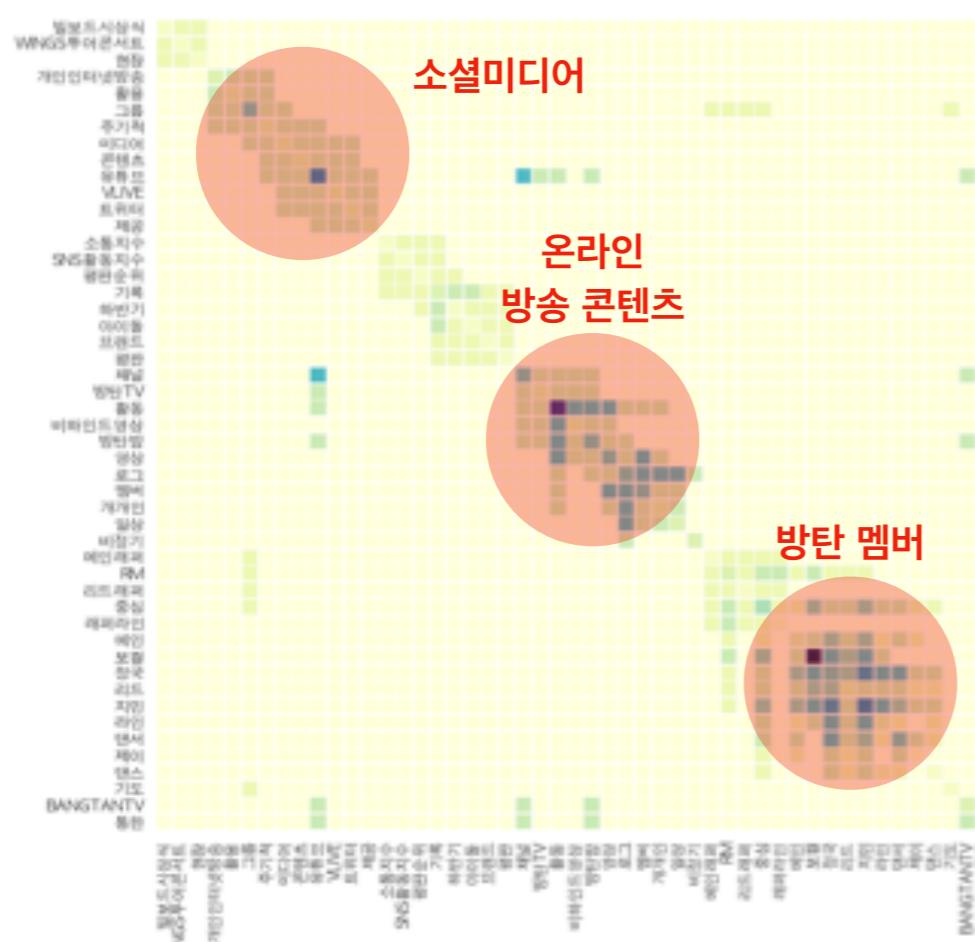
- ▶ 단어가 나온 문서의 수가 적을수록 단어가 문서에 중요할 가능성이 더 큼
- ▶ 단어의 희박성(sparseness)을 역문서빈도(IDF)로 측정
- ▶ 전체 문서 수가 고정된 체로 단어 t 가 출현하는 문서가 많을수록 중요도가 감소
 - ▶ $TF(t,d)$ 가 높다 = 문서 d 에 단어 t 가 많이 출현한다.
 - ▶ $DF(t)$ 가 높다 = 단어 t 가 포함된 문서가 많다. (IDF는 반대)
 - ▶ $TF-IDF(t,d)$ 가 높다 = 단어 t 가 문서 d 에는 많이 출현하지만, 단어 t 를 포함하는 문서는 적다.

동시출현분석

(Co-occurrence Analysis)

단어의 동시출현 빈도를 바탕으로 가중치를 계산하는 분석방법

- ▶ 단어들 사이의 동시 출현을 연관성으로 취급하여, 단어의 연관성을 파악하는 방법
- ▶ 상용 소셜 미디어 분석 솔루션에서 제공하는 가장 기본적인 분석 형태 (Social Matrix)
- ▶ 활용분야 : 브랜드 이미지 조사, 트렌드 분석, 여론조사, 마케팅 모니터링



| 표1 '아쿠로트 아줌마' 연관어 변화 | | | | | | | | | | | | |
|----------------------|-----|---------|-------|-----|-------|-------|-----|-------|-------|-----|---------|-------|
| | 연관어 | 연금 비중 | | 연관어 | 연금 비중 | | 연관어 | 연금 비중 | | 연관어 | 연금 비중 | |
| 2013년 | 1 | 아쿠로트 | 21.3% | 1 | 아쿠로트 | 26.3% | 1 | 아쿠로트 | 26.6% | 1 | 아쿠로트 | 13.1% |
| | 2 | 막다 | 4.9% | 2 | 건강 | 4.5% | 2 | 집 | 4.7% | 2 | 풀드브루 | 8.2% |
| | 3 | 아침 | 4.4% | 3 | 아침 | 4.0% | 3 | 아침 | 4.6% | 3 | 커피 | 7.4% |
| | 4 | 엄마 | 4.2% | 4 | 집 | 3.8% | 4 | 맛 | 3.9% | 4 | 맛 | 6.6% |
| | 5 | 집 | 3.5% | 5 | 제품 | 3.4% | 5 | 막다 | 3.4% | 5 | 파리 | 5.7% |
| | 6 | 오다 | 2.8% | 6 | 엄마 | 3.3% | 6 | 사다 | 2.8% | 6 | 치즈 | 5.3% |
| | 7 | 사다 | 2.7% | 7 | 맛 | 2.7% | 7 | 주다 | 2.8% | 7 | 과자 | 5.0% |
| | 8 | 주다 | 2.5% | 8 | 같다 | 2.6% | 8 | 다니다 | 2.7% | 8 | 아메리카노 | 4.1% |
| | 9 | 구입하다 | 2.4% | 9 | 우유 | 2.6% | 9 | 엄마 | 2.6% | 9 | 막다 | 3.3% |
| | 10 | 아이 | 2.4% | 10 | 주다 | 2.2% | 10 | 우유 | 2.1% | 10 | 크림치즈 | 3.1% |
| | 11 | 아쿠로트 주다 | 2.3% | 11 | 막다 | 2.2% | 11 | 만나다 | 2.1% | 11 | 라떼 | 2.8% |
| | 12 | 배달하다 | 2.3% | 12 | 만나다 | 2.0% | 12 | 제품 | 2.0% | 12 | 만나다 | 2.7% |
| | 13 | 수입 | 2.3% | 13 | 사다 | 1.9% | 13 | 사진 | 2.0% | 13 | 가격 | 2.4% |
| | 14 | 다니다 | 2.1% | 14 | 맙다 | 1.9% | 14 | 나오다 | 2.0% | 14 | 찾다 | 1.9% |
| | 15 | 알려하다 | 2.0% | 15 | 배달하다 | 1.8% | 15 | 왔다 | 1.9% | 15 | 아침 | 1.8% |
| | 16 | 살다 | 2.0% | 16 | 다니다 | 1.8% | 16 | 지나가다 | 1.8% | 16 | 10월 | 1.8% |
| | 17 | 제품 | 2.0% | 17 | 하루이체 | 1.7% | 17 | 하나 | 1.7% | 17 | 엄마 | 1.5% |
| | 18 | 세븐 | 1.8% | 18 | 나누다 | 1.7% | 18 | 판매 | 1.7% | 18 | 우유 | 1.4% |
| | 19 | 가다 | 1.8% | 19 | 지나가다 | 1.6% | 19 | 일하다 | 1.6% | 19 | 팔다 | 1.3% |
| | 20 | 자녀 | 1.8% | 20 | 세븐 | 1.5% | 20 | 오다 | 1.6% | 20 | 발간하다 | 1.3% |
| | 21 | 만나다 | 1.8% | 21 | 수입 | 1.5% | 21 | 찾다 | 1.6% | 21 | 사다 | 1.2% |
| | 22 | 마시다 | 1.7% | 22 | 찾다 | 2.3% | 22 | 용료 | 1.5% | 22 | 인기 | 1.2% |
| | 23 | 유산균 | 1.7% | 23 | 노인 | 1.4% | 23 | 마시다 | 1.4% | 23 | 편의점 | 1.2% |
| | 24 | 힐하다 | 1.7% | 24 | 마시다 | 1.4% | 24 | 길 | 1.4% | 24 | 파리돌렌크원자 | 1.1% |
| | — | | — | | | | — | | | — | | |
| | 29 | 왔다 | 1.4% | 29 | 왔다 | 1.3% | 29 | 배달하다 | 1.2% | 29 | 구입하다 | 1.0% |

■ 상승 기록 ■ 하락 기록 ■ 신규 기록

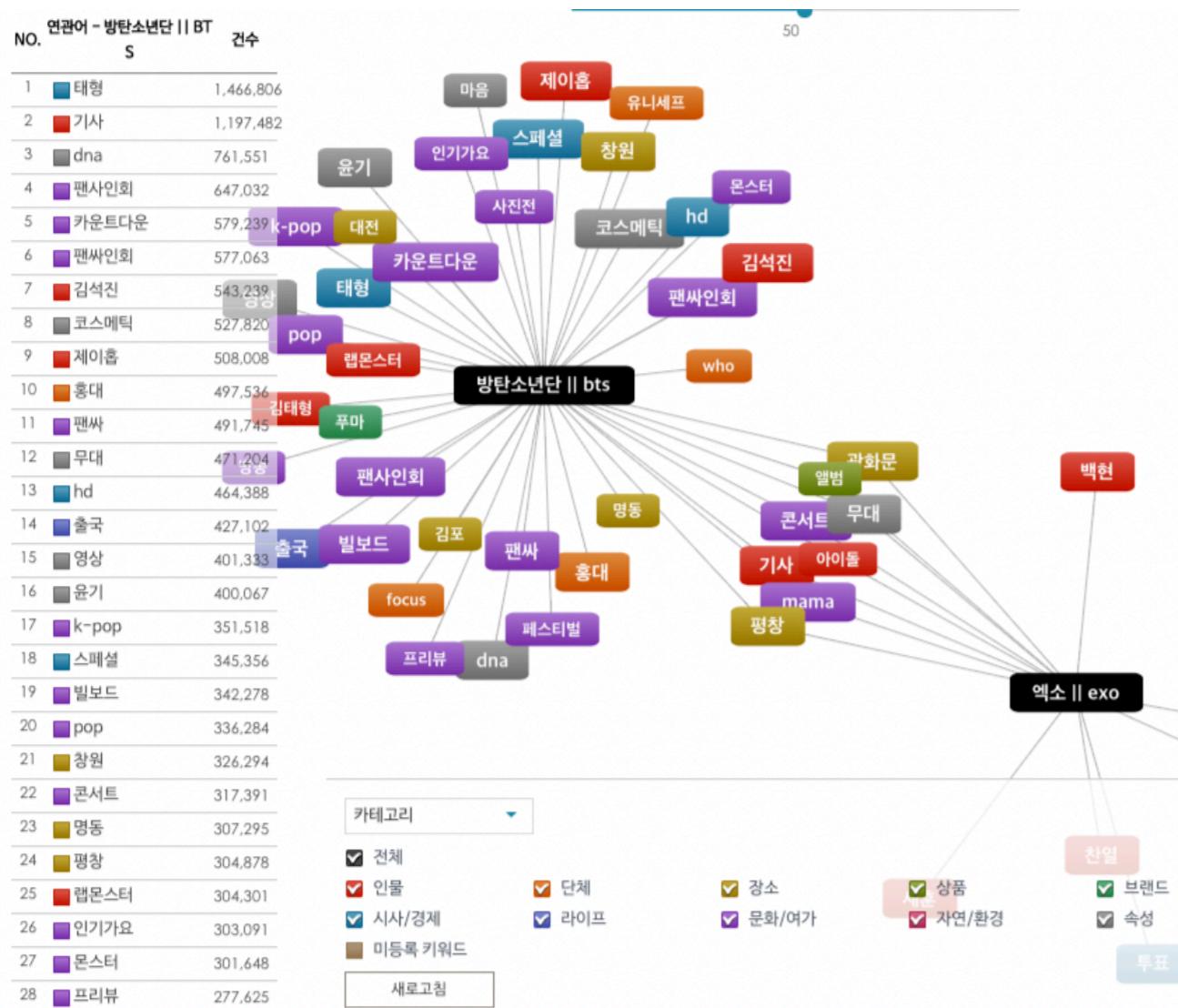
* Source : 전병진, 신한은행 파이썬으로 시작하는 데이터분석: 텍스트 마이닝 기초, 2018.12.12.

** Source : 백경혜(DBR), “매력을 소비하는 나는 덕후! 즐거움을 위해 기꺼이 지갑을 연다”, 2017.1., http://dbr.donga.com/article/view/1203/article_no/7935/.

동시출현분석 (Co-occurrence Analysis)

연관어분석 (Co-word Analysis)

- ▶ 단어들 사이의 동시출현 빈도 중 빈번하게 사용되는 특정 단어를 기준으로 연관성을 파악하는 방법
 - ▶ 연관어(공기어, Co-word) : 같은 문맥 안에서 함께 나타나 서로 밀접한 의미 관계를 갖는 단어



동시출현분석

(Co-occurrence Analysis)

감성 키워드 순위



긍정 감성어



기간별 연관어 순위 : 방탄소년단 || BTS

2017/10/03 ~ 2017/11/03

전체 트위터 블로그 커뮤니티 인스타그램 뉴스

확인

일별 주별 월별 분기별



카테고리 ▾

- 전체
- 인물
- 단체
- 장소
- 상품
- 브랜드
- 라이프
- 시사/경제
- 문화/여가
- 자연/환경
- 속성
- 미등록 키워드

새로고침

| 순위 | 2017/10/03~2017/10/07 | | 2017/10/08~2017/10/14 | | 2017/10/15~2017/10/21 | | 2017/10/22~2017/10/28 | | 2017/10/29~2017/11/03 | |
|----|-----------------------|---------|-----------------------|---------|-----------------------|---------|-----------------------|---------|-----------------------|-----------|
| | 연관어 | 건수 |
| 1 | 방탄소년단 | 826,236 | 방탄소년단 | 987,650 | 방탄소년단 | 798,031 | 방탄소년단 | 549,312 | 방탄소년단 | 1,481,373 |
| 2 | 태형 | 449,263 | 태형 | 552,652 | 기사 | 260,201 | 김석진 | 147,539 | 평창 | 289,317 |
| 3 | 코스메틱 | 404,925 | 홍대 | 428,101 | 태형 | 217,125 | 태형 | 141,507 | 광화문 | 250,964 |
| 4 | 기사 | 260,803 | 기사 | 423,304 | 출국 | 205,573 | hd | 108,600 | 콘서트 | 214,887 |
| 5 | 명동 | 253,605 | dna | 362,319 | 푸마 | 187,739 | 무대 | 106,730 | 기사 | 205,638 |
| 6 | 팬싸인회 | 207,776 | 카운트다운 | 355,310 | 김석진 | 165,320 | 타이페이 | 95,120 | 유니세프 | 205,279 |
| 7 | 팬싸인회 | 200,843 | 팬싸인회 | 329,385 | dvd | 128,894 | 캠 | 92,580 | 캠페인 | 131,943 |
| 8 | 마음 | 198,856 | 영상 | 297,121 | hd | 127,530 | 제이홉 | 77,808 | 무대 | 119,974 |
| 9 | who | 186,415 | 팬싸인회 | 283,711 | dna | 124,366 | 콘서트 | 77,700 | 스페셜 | 119,423 |
| 10 | dna | 184,025 | 팬싸 | 222,826 | 제이홉 | 117,073 | 윤기 | 76,071 | 리허설 | 113,370 |
| 11 | 팬싸 | 175,731 | 출국 | 197,729 | mama | 107,045 | 대만 | 66,871 | 올림픽 | 109,404 |

동시출현분석

(Co-occurrence Analysis)

응용 가능한 예

- ▶ Roses are red.
- ▶ The sky is blue.
- ▶ iPhone has a rose gold color.

| | rose | red | sky | blue | iPhone | gold | color |
|--------|------|-----|-----|------|--------|------|-------|
| rose | | 1 | | | 1 | 1 | 1 |
| red | 1 | | | | | | |
| sky | | | | 1 | | | |
| blue | | | 1 | | | | |
| iPhone | 1 | | | | | 1 | 1 |
| gold | 1 | | | | 1 | | 1 |
| color | 1 | | | | 1 | 1 | |



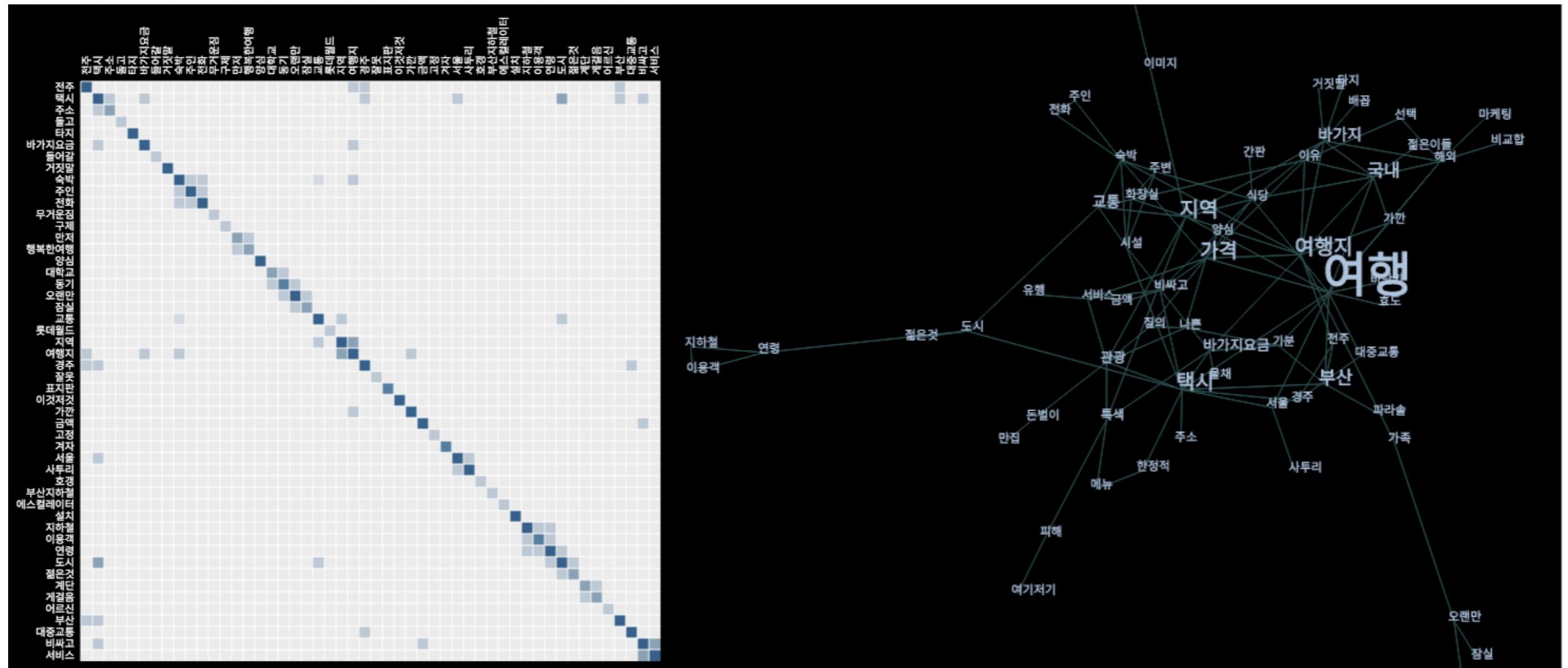
“어쩌면 *rose gold* 는 *red gold*? ”

동시출현분석

(Co-occurrence Analysis)

단어 네트워크 분석 (Word Network Analysis)

- ▶ 텍스트 마이닝 분야에 네트워크 분석, 그래프 이론을 적용한 분석방법
 - ▶ 단어의 연관성(동시출현)을 단어와 단어 사이의 관계로 정의하고 네트워크 분석 방법론 적용
 - ▶ 문서에서 단어와 단어 사이 관계를 파악하고 정량화하여 분석하기 위해 사용

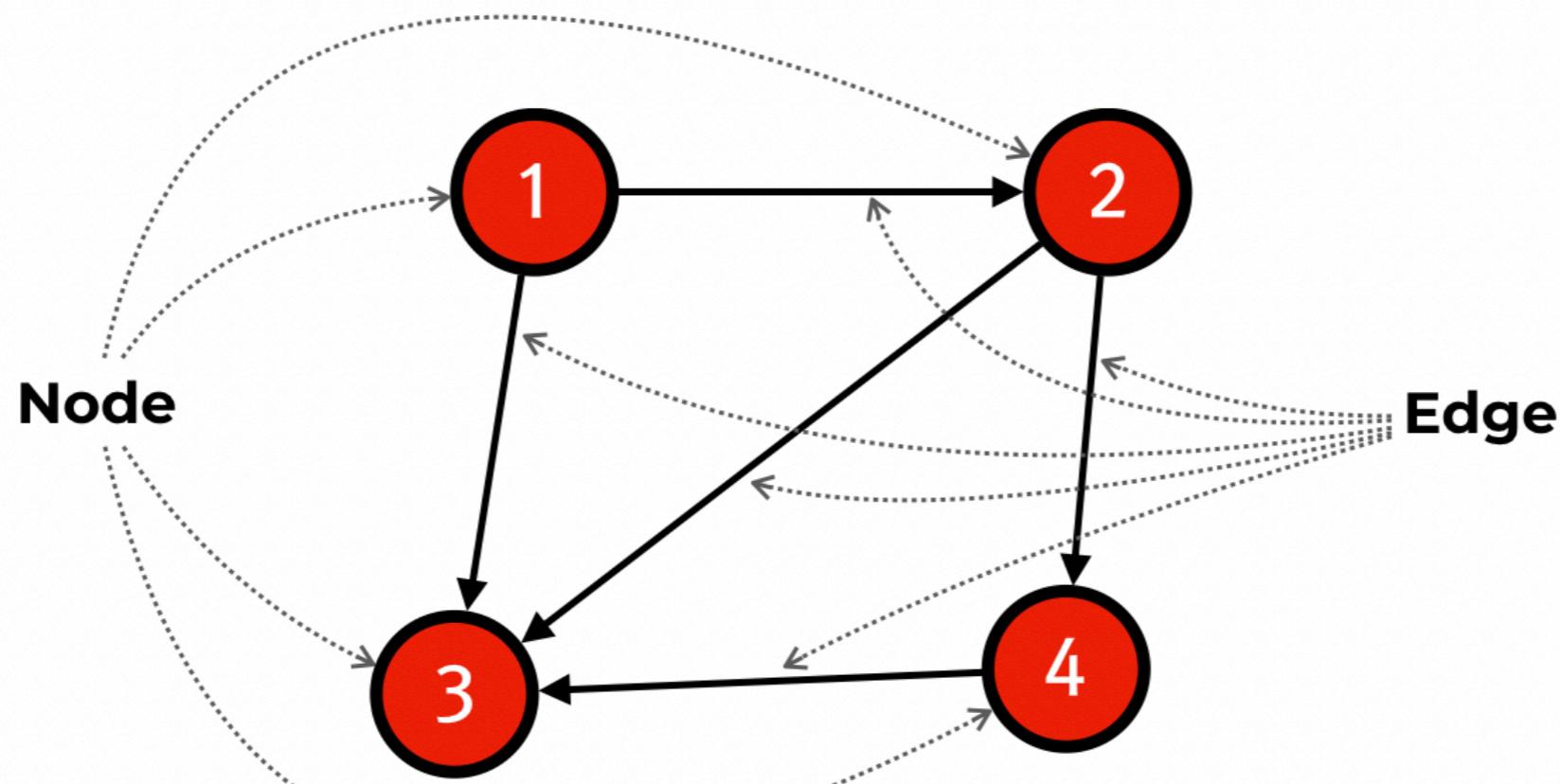


단어 네트워크 (Word Network)

그래프 (Graph) 기본개념

▶ 기본용어

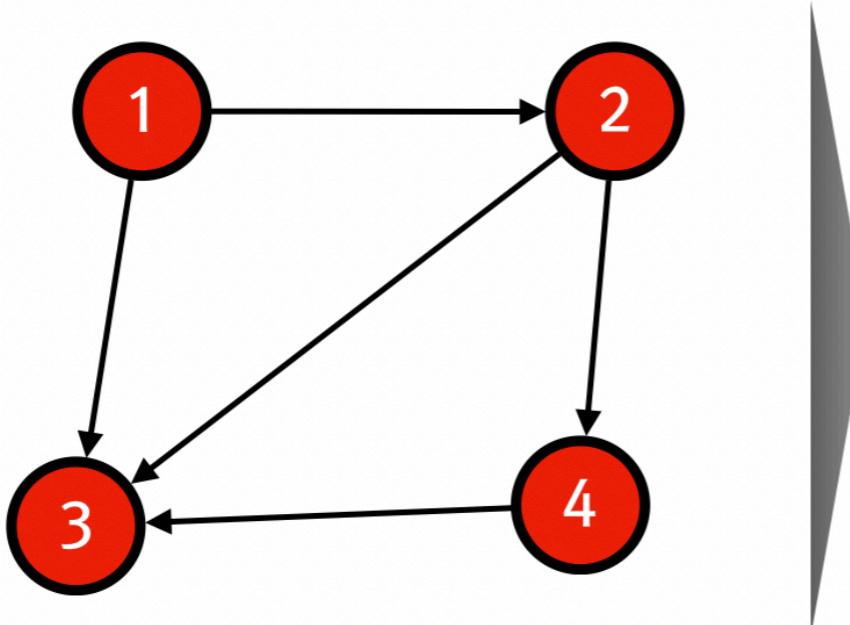
- 노드 (node, vertex, point) : 관계를 가지는 그래프 요소
- 엣지 (edge, line, arc) : 관계로 연결된 한 쌍의 노드
- 방향성 그래프 (directed graph) : 화살표를 이용해 방향이 표시된 그래프
- 비방향성 그래프 (undirected graph) : 방향성이 없는 그래프



단어 네트워크 (Word Network)

그래프 (Graph) 기본개념

방향성 그래프 (*directed graph*)



엣지리스트 (*edge list*)

| Vertex | Vertex |
|--------|--------|
| 1 | 2 |
| 1 | 3 |
| 2 | 3 |
| 2 | 4 |
| 3 | 4 |

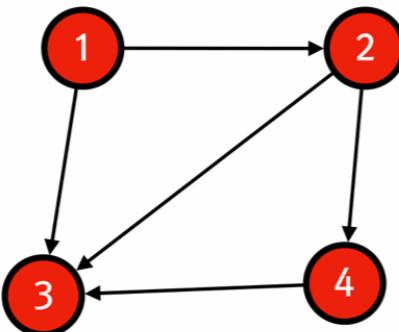
인접행렬 (*adjacency matrix*)

| Vertex | 1 | 2 | 3 | 4 |
|--------|---|---|---|---|
| 1 | - | 1 | 1 | 0 |
| 2 | 0 | - | 1 | 1 |
| 3 | 0 | 0 | - | 0 |
| 4 | 0 | 0 | 1 | - |

단어 네트워크 (Word Network)

그래프 (Graph) 기본개념

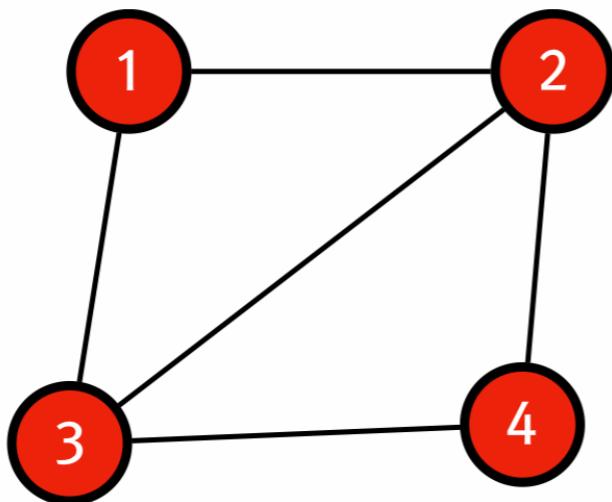
방향성 그래프 (*directed graph*)



엣지리스트 (*edge list*)

| Vertex | Vertex |
|--------|--------|
| 1 | 2 |
| 1 | 3 |
| 2 | 3 |
| 2 | 4 |
| 3 | 4 |

비방향성 그래프 (*undirected graph*)



인접행렬 (*adjacency matrix*)

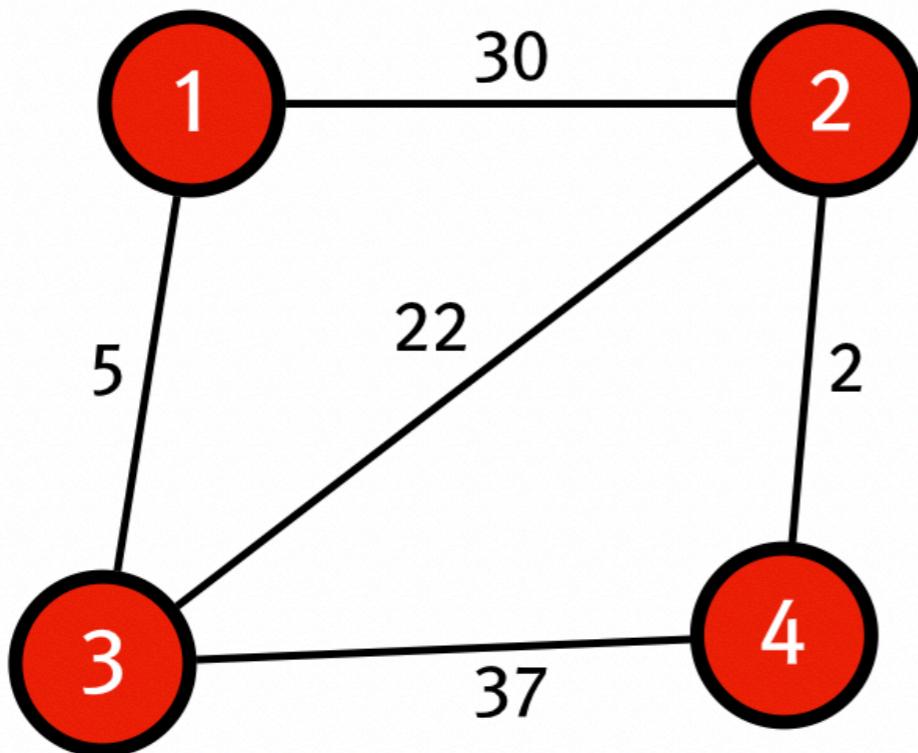
| Vertex | 1 | 2 | 3 | 4 |
|--------|---|---|---|---|
| 1 | - | 1 | 1 | 0 |
| 2 | 1 | - | 1 | 1 |
| 3 | 1 | 1 | - | 1 |
| 4 | 0 | 1 | 1 | - |

단어 네트워크 (Word Network)

그래프 (Graph) 기본개념

▶ 기본용어

- 경로 (path) : 간선에 의하여 연결된 노드들의 순차적 배열
- 최단 경로 (shortest path) : 그래프의 두 노드 간의 가장 짧은 경로
- 엣지 리스트 (edge list) : 노드와 노드 관계(경로)를 짹지어 목록으로 만든 것
- 가중치 (weight) : 네트워크에서 연결 관계의 강도를 나타내는 값



엣지리스트 (edge list)

Vertex Vertex Weight

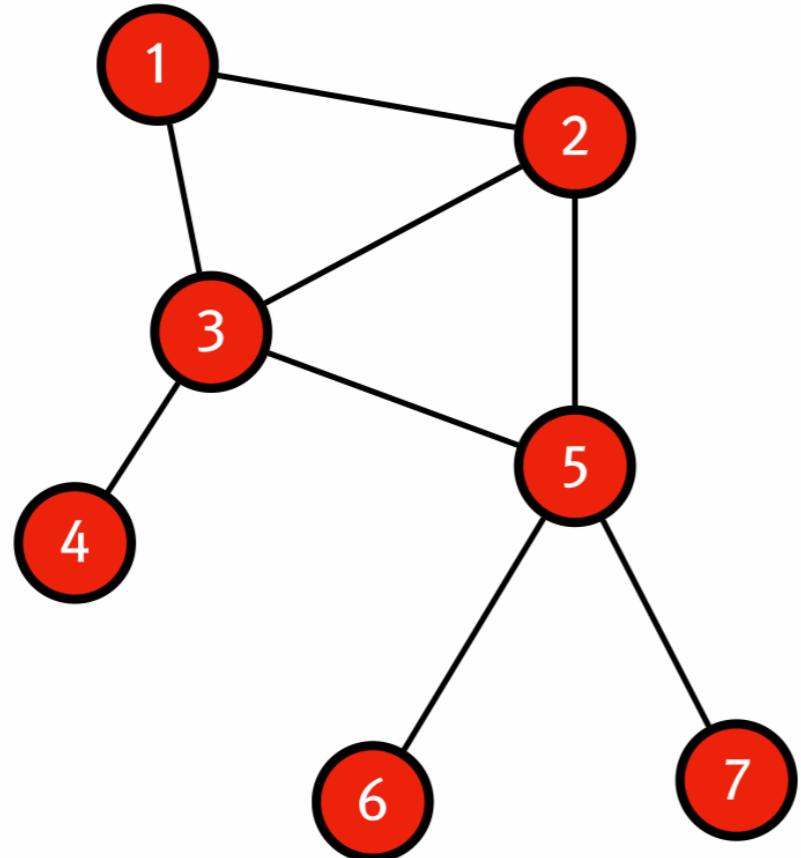
| Vertex | Vertex | Weight |
|--------|--------|--------|
| 1 | 2 | 30 |
| 1 | 3 | 5 |
| 2 | 3 | 22 |
| 2 | 4 | 2 |
| 3 | 4 | 37 |

단어 네트워크 (Word Network)

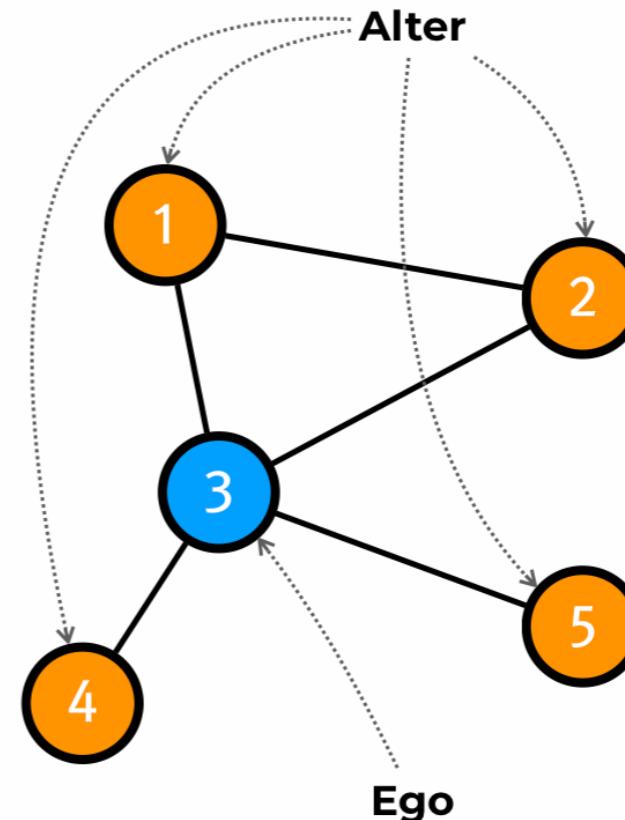
그래프 (Graph) 기본개념

▶ 기본용어

- 에고 네트워크 (ego network) : 한 노드를 중심으로 다른 노드와의 연결관계를 표현한 네트워크



전체 네트워크 (Whole Network)

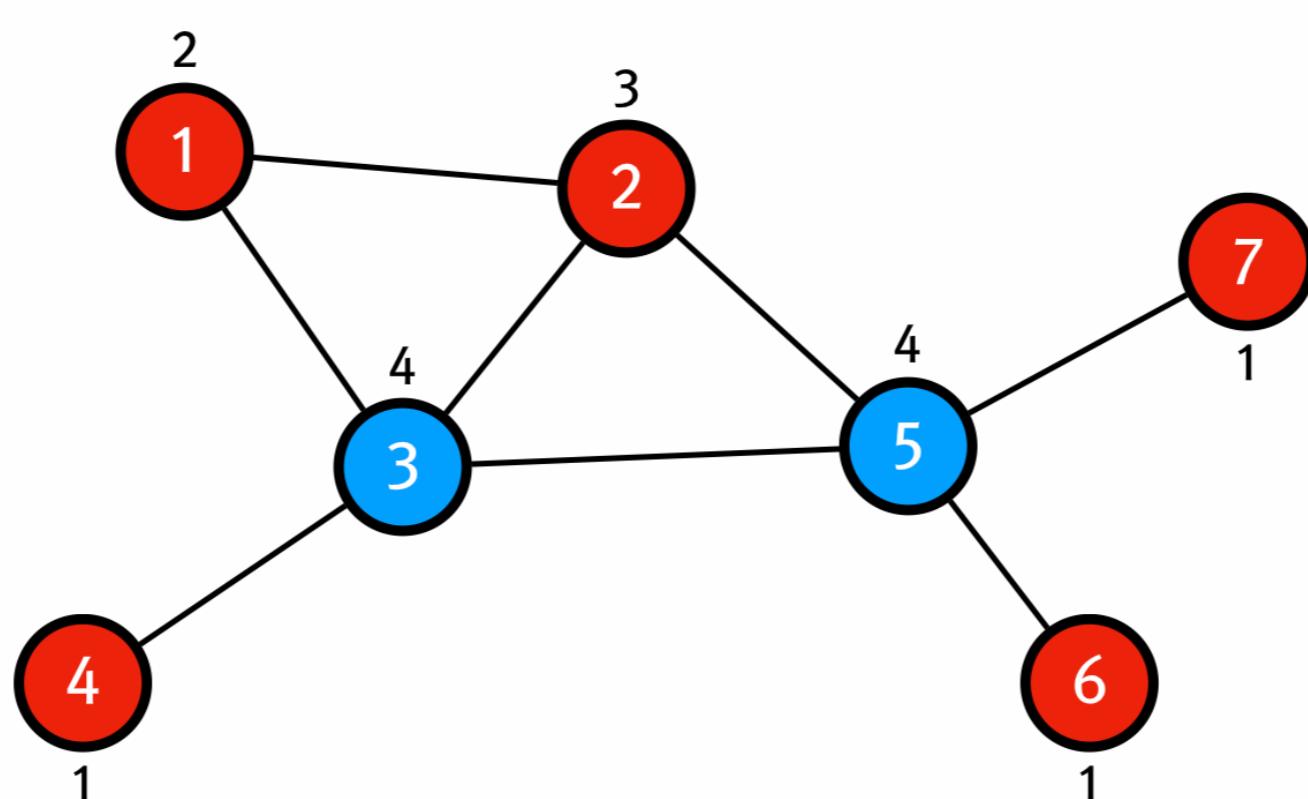


노드 3의 에고 네트워크

단어 네트워크 (Word Network)

그래프 중심성 척도: 연결 중심성 (Degree Centrality)

- ▶ 어떤 단어가 가장 많은 단어들과 같이 쓰였는가에 대한 척도
- ▶ 한 노드가 다른 노드와 연결된 엣지의 개수
- ▶ 비방향성 그래프에서는 한 노드로 연결될 수 있는 경로의 수
- ▶ 영향력 또는 인기도를 측정할 때 노드의 연결 정도의 척도로 사용
- ▶ 정보의 확산과 관련해 어느 노드가 중심이고, 다른 이웃 노드들에게 영향을 미치는지 평가할 때 사용

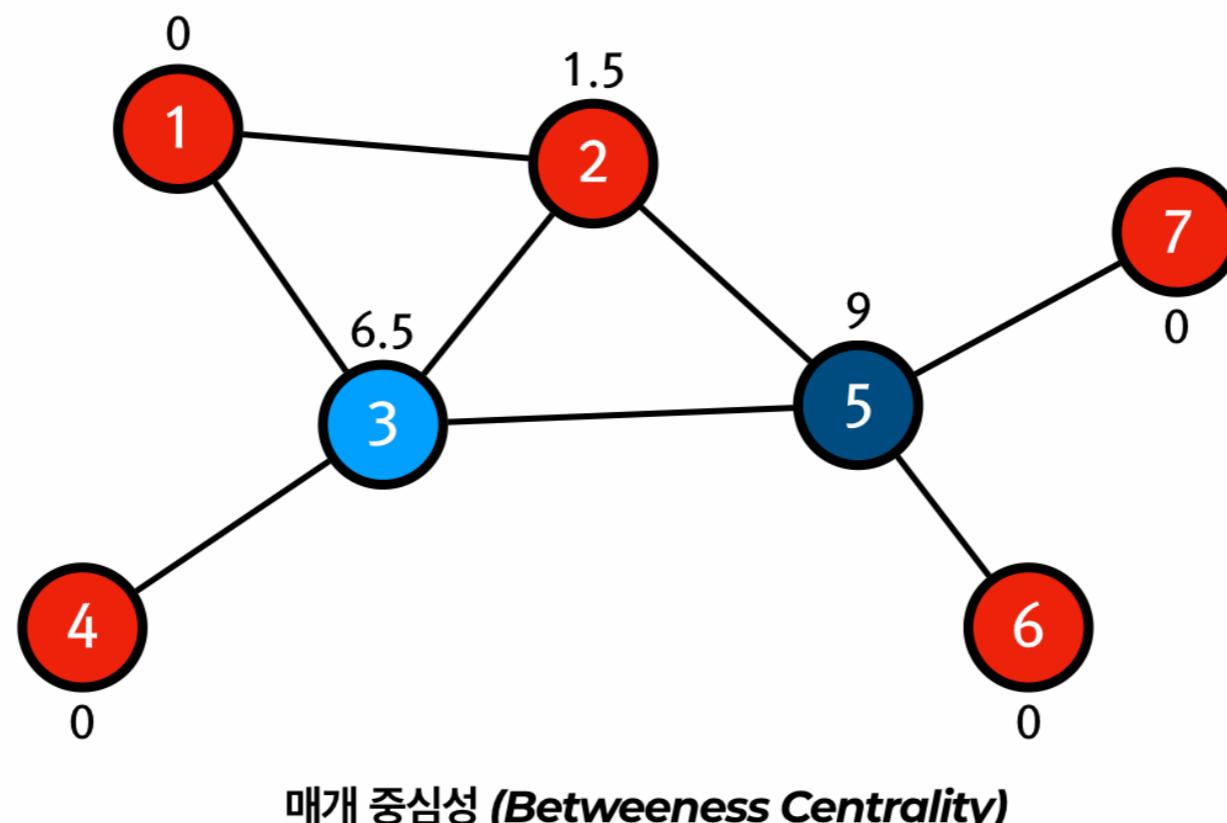


단어 네트워크 (Word Network)

그래프 중심성 척도: 매개 중심성 (Betweenness Centrality)

- ▶ 어떤 단어가 다른 단어들 사이의 연결고리 역할을 하는가에 대한 척도
- ▶ 네트워크 내에서 한 노드가 다른 노드들 사이의 경로에 위치하는 정도
- ▶ 각 노드가 다른 노드들 간의 최단거리 (shortest path)에 등장하는 빈도

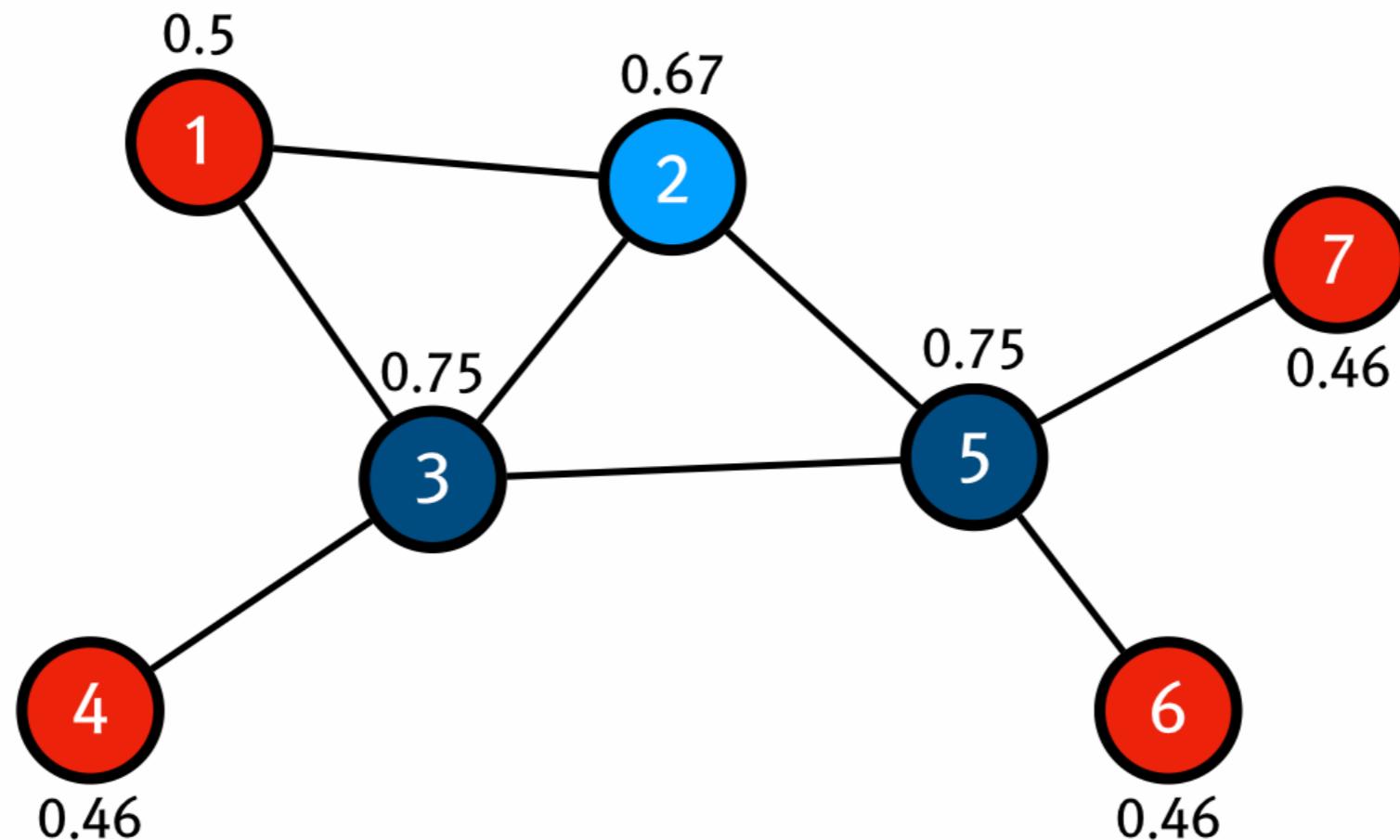
$$C_B(v) = \frac{i\text{와 } j \text{ 간의 최단경로 중 } v\text{를 지나는 경로의 수}}{i\text{와 } j \text{ 간의 최단경로의 수}} \quad i, j, v : \text{노드}$$



단어 네트워크 (Word Network)

그래프 중심성 척도: 근접 중심성 (Closeness Centrality)

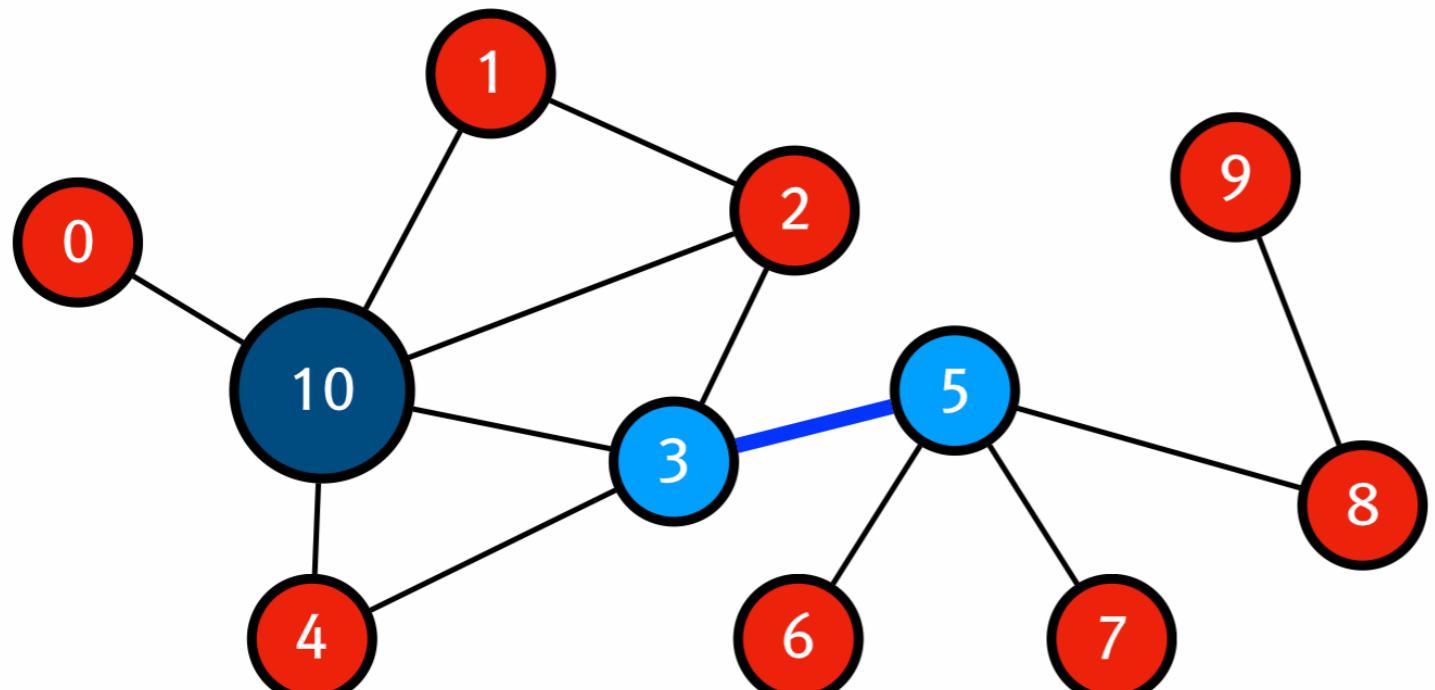
- ▶ 어떤 단어가 다른 단어들과의 가장 가까운 거리에 있는가에 대한 척도
- ▶ 한 노드에서 다른 모든 노드까지 모든 최단 경로의 평균(또는 이의 역수)
- ▶ 모든 다른 노드에 도달하는데 까지 평균 소요 시간



근접 중심성 (*Closeness Centrality*)

단어 네트워크 (Word Network)

그래프 중심성 척도



Sample Graph

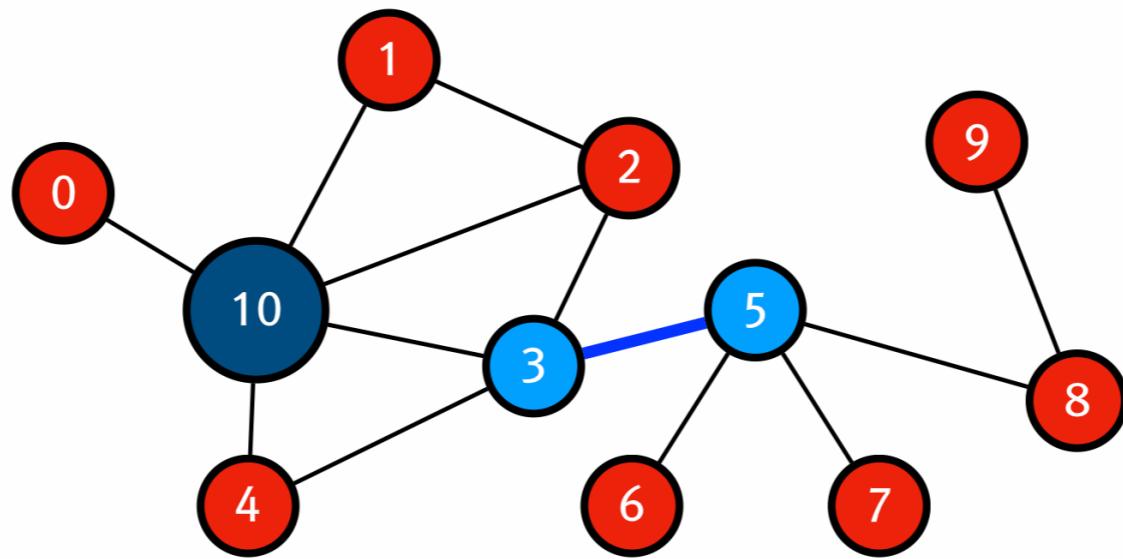
노드리스트 (*node list*)

| Node | Degree Centrality | Betweenness Centrality | Closeness Centrality |
|------|-------------------|------------------------|----------------------|
| 0 | | | |
| 1 | | | |
| 2 | | | |
| 3 | | | |
| 4 | | | |
| 5 | | | |
| 6 | | | |
| 7 | | | |
| 8 | | | |
| 9 | | | |
| 10 | | | |

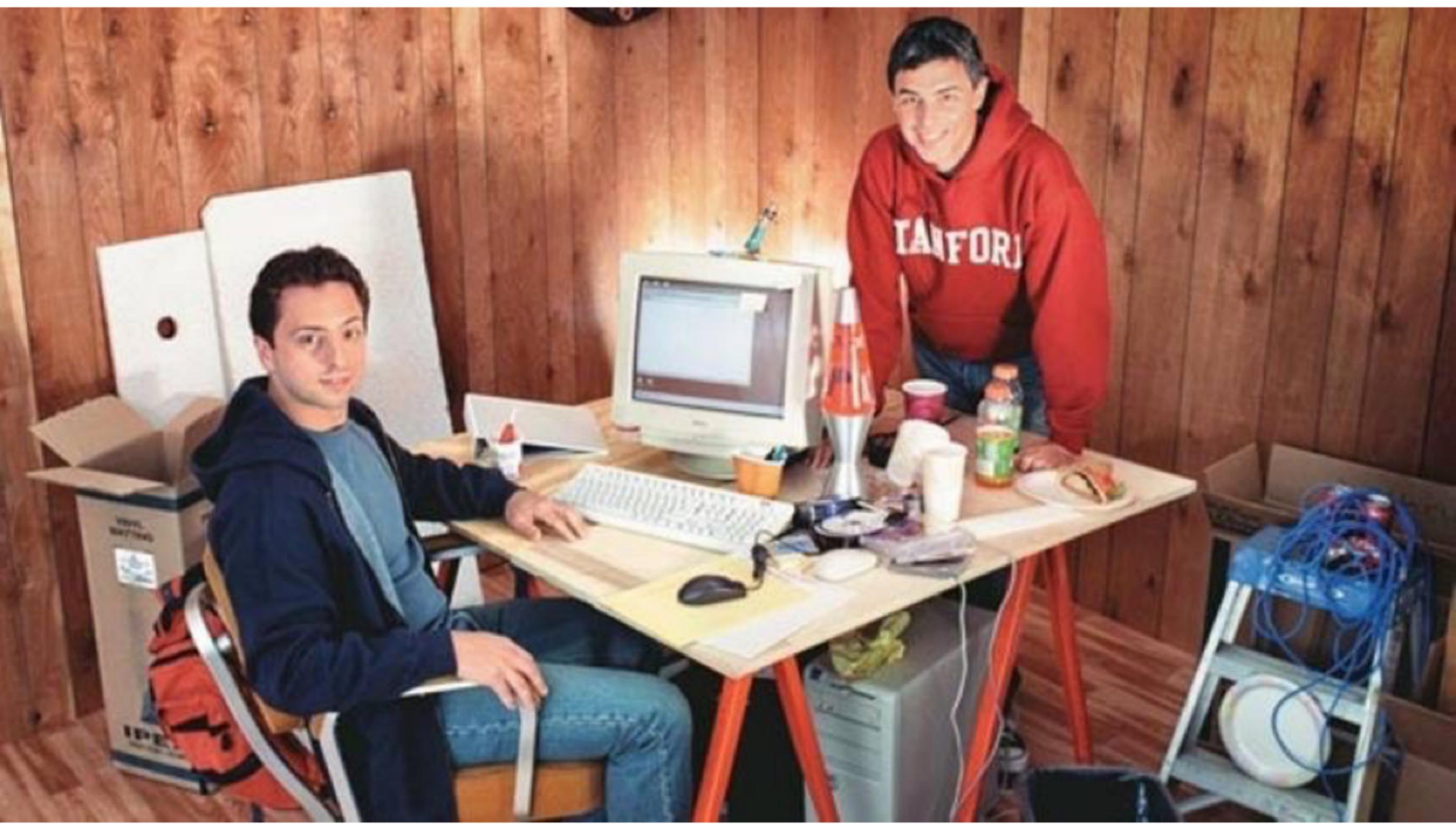
단어 네트워크 (Word Network)

그래프 중심성 척도

- ▶ 분석의 목적에 따라 척도를 다르게 적용하여 분석에 활용 (중심 노드를 노드의 뮝음으로 고려해도 됨)
- ▶ 중심 노드 선정의 예
 - 10은 연결 중심성 측면에서 가장 중심
 - 3과 5는 매개 중심성 측면에서 10 보다 더 중심
 - 또한 3과 5 사이의 관계는 네트워크가 분리될 수 있는 중요한 연결
 - 다른 조건들이 동일할 때, 3과 5는 10보다 네트워크의 중심



Sample Graph





Search the web using Google!

Special Searches

[Stanford Search](#)

[Linux Search](#)

[Help!](#)
[About Google!](#)
[Company Info](#)
[Google! Logos](#)

Get Google!
updates monthly:

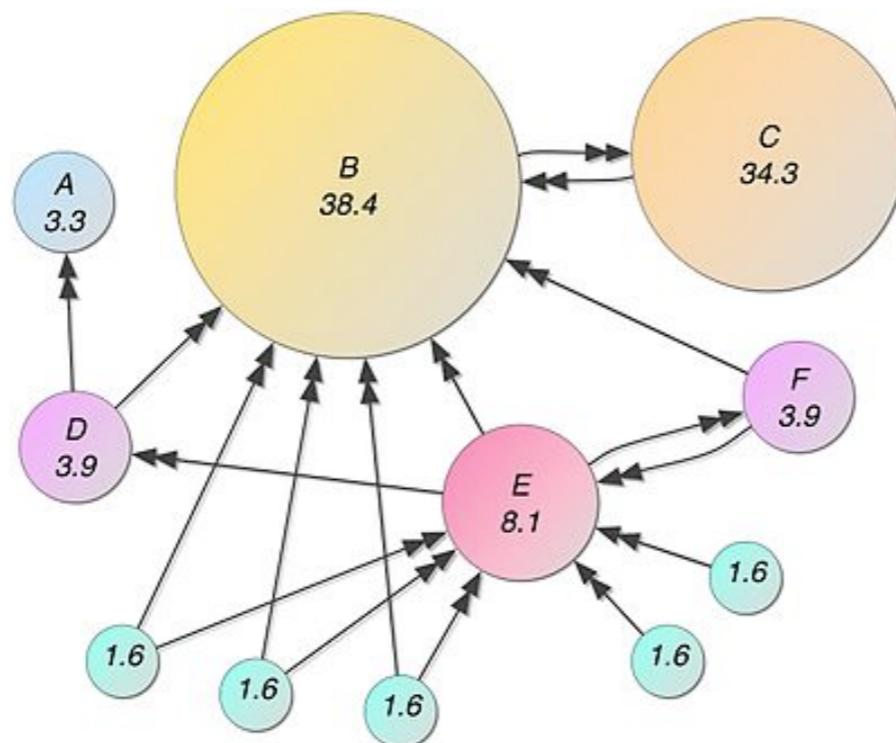
[Archive](#)

Copyright ©1998 Google Inc.

페이지 랭크 (Page Rank)

웹 페이지의 유명도에 따라 랭킹을 매기는 방식

- ▶ 구글 창업자 세르게이 브린과 레리 페이지가 1998년도에 발표
- ▶ 논문 : A web page is important if it is pointed to by other important web pages
- ▶ 웹상에 존재하는 웹 페이지들의 중요도를 나타내는 수치적 척도
- ▶ 하나의 웹 페이지가 다른 웹 페이지를 포함하는 (e.g., **anchor tag**) 형태를 투표 간주
- ▶ 더 많은 투표는 더 높은 중요도를 의미
- ▶ 각 표의 중요도는 그 페이지의 현재 PageRank에 의해 달라짐



페이지 랭크 (Page Rank)

Random Surfer Model

- ▶ 기본적으로 페이지 랭크는 다음과 같은 가정을 함.
- ▶ 사용자는 현재 페이지 a라는 곳에 있다.
 - ▶ 1-q의 확률로 페이지 a에 있는 링크를 타고 다른 페이지로 이동한다.
 - ▶ q의 확률로 임의의 웹페이지로 이동한다. (내부에 링크가 없는 페이지에서 멈추는 것을 막기위해)
 - ▶ 위의 과정을 반복한다

$$PR(a) = \frac{q}{T} + (1 - q) \sum_{i=1}^n \frac{PR(p_i)}{L(p_i)}$$

1-q의 확률로 페이지 a에 있는 링크를 타고
이동함

페이지 p_i 의 페이지랭크

페이지 a의 페이지랭크

웹 그래프에 있는
전체 페이지의 수

페이지 p_i 의 있는 링크의 수

페이지 a를 가리키는
페이지들

페이지 a의 페이지랭크

q의 확률로 임의의 웹페이지로
이동함

페이지 랭크 (Page Rank)

Random Surfer Model

- ▶ 기본적으로 페이지 랭크는 다음과 같은 가정을 함.
- ▶ 사용자는 현재 페이지 a라는 곳에 있다.
 - ▶ 1-q의 확률로 페이지 a에 있는 링크를 타고 다른 페이지로 이동한다.
 - ▶ q의 확률로 임의의 웹페이지로 이동한다. (내부에 링크가 없는 페이지에서 멈추는 것을 막기위해)
 - ▶ 위의 과정을 반복한다

$$PR(a) = \frac{q}{T} + (1 - q) \sum_{i=1}^n \frac{PR(p_i)}{L(p_i)}$$

1-q의 확률로 페이지 a에 있는 링크를 타고
이동함

페이지 p_i 의 페이지랭크

페이지 a의 페이지랭크

웹 그래프에 있는
전체 페이지의 수

페이지 p_i 의 있는 링크의 수

페이지 a를 가리키는
페이지들

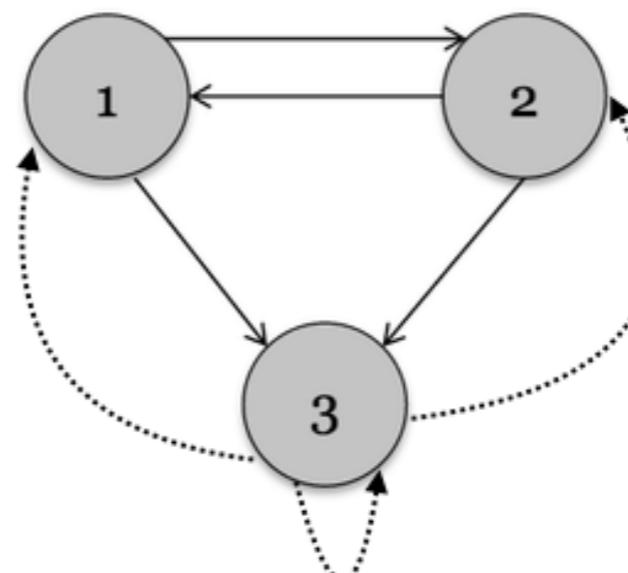
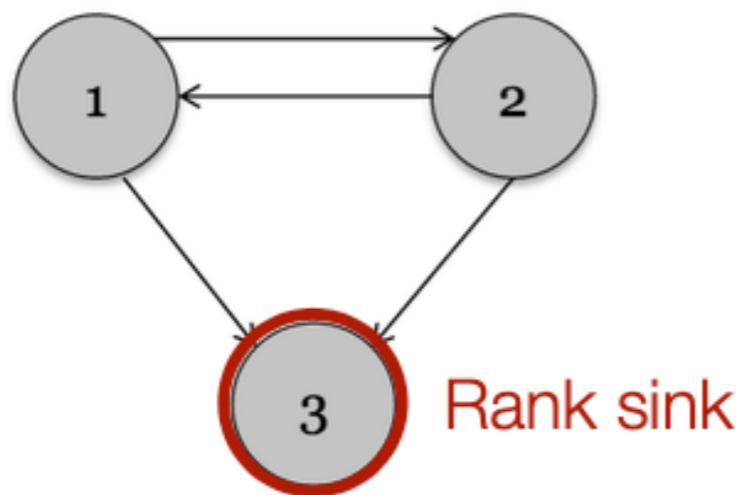
페이지 a의 페이지랭크

q의 확률로 임의의 웹페이지로
이동함

페이지 랭크 (Page Rank)

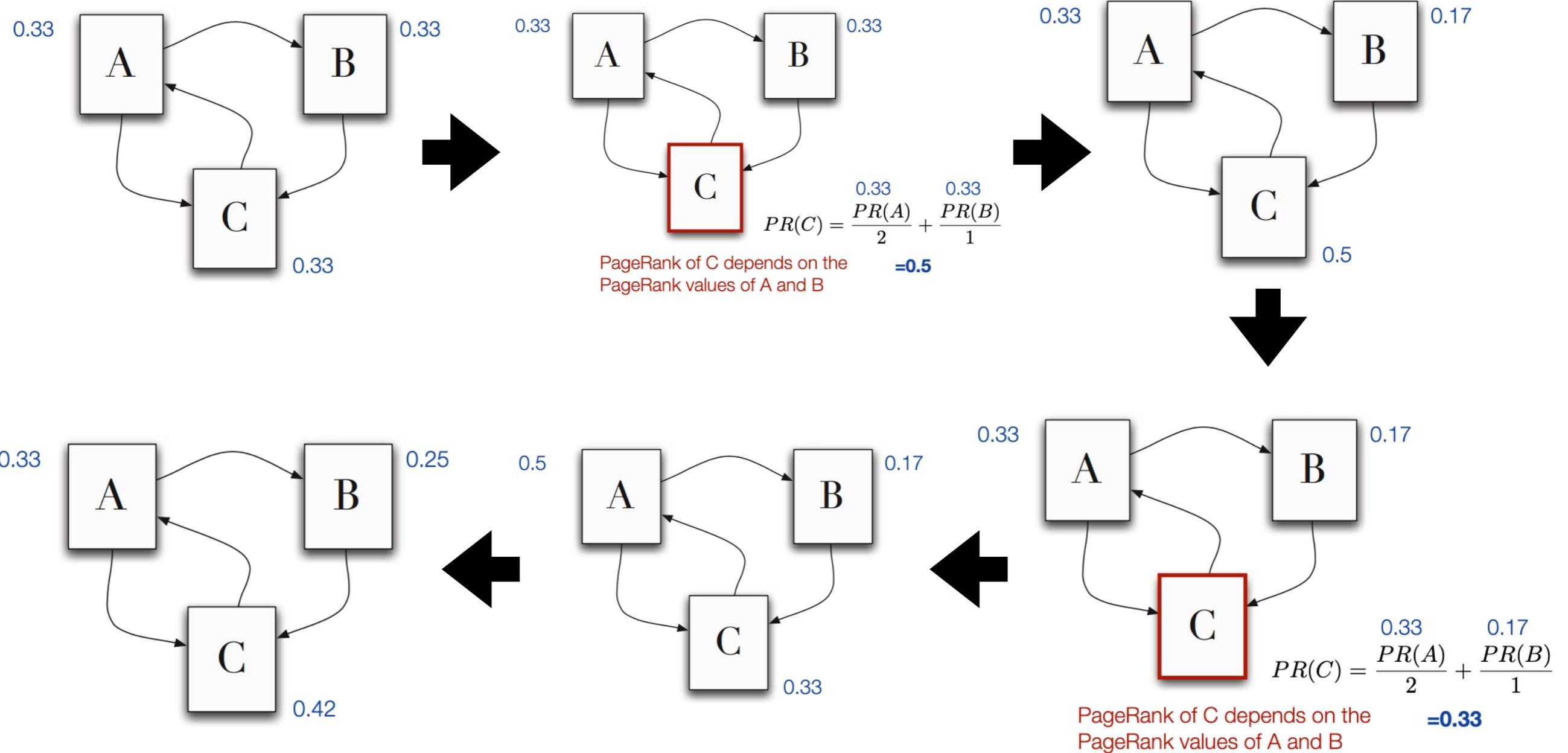
기술적 문제

- ▶ 첫 랭크는 어떻게 결정하나?
 - ▶ 최초 랭크는 모든 페이지에 균등하게 분배한다 = $1/T$
- ▶ 몇 번을 반복해야하나?
 - ▶ 랭크 계산을 여러 사이클을 돌게 되면, 랭크 변화 차이가 크게 나지 않는 지점이 있다.
- ▶ Outgoing link가 없는 페이지는?
 - ▶ 자기 자신을 포함한 모든 페이지에 링크가 있다고 가정한다.
 - ▶ 예) 모든페이지가 10,000개면 outgoing link의 가중치는 $1/10,000$ 가 되기 때문에 큰 영향을 미치지 않는다.



페이지 랭크 (Page Rank)

Example





Stephen Robertson, SIGIR'17 keynote

So how did Google do so well?

My guesses:

Good crawling

A good sense of the variety of types of web search

Good basic NL analysis

Good use of traditional ranking clues

Good use of phrases / proximity / fields

Good use of anchor text

Maybe a bit of PageRank!

Plus good testing

(and, later, good learning from users)

E.O.D