


The PageRank citation ranking: Bringing order on web

A complex network graph with numerous nodes and edges, rendered in light blue and grey. The nodes are represented by small circles, and the edges are thin lines connecting them. The graph is dense and interconnected, with some nodes having more connections than others. The background is a light, textured grey.

Seoul National University of Science and Technology

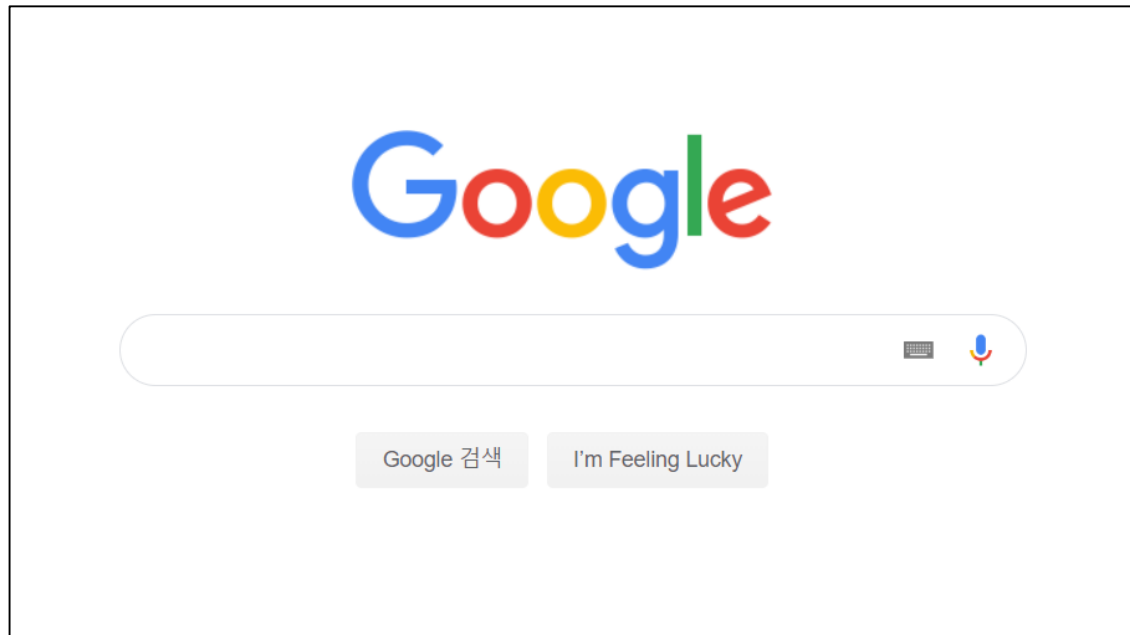
Dept. of Computer Science and Engineering

Cryptography and Information Security Lab.

2019-05-30

Introduction

Google's Search Engine Algorithm



Introduction

Ordering algorithm

- ❖ 인터넷 상에 존재하는 정보의 과다로 인해 중요한 정보를 획득하는 것이 어렵게 됨
- ❖ 정보의 중요도, 신뢰성에 대한 검증이 필요함
- ❖ 정보가 중요하다는 것을 어떻게 판단할 수 있는가?

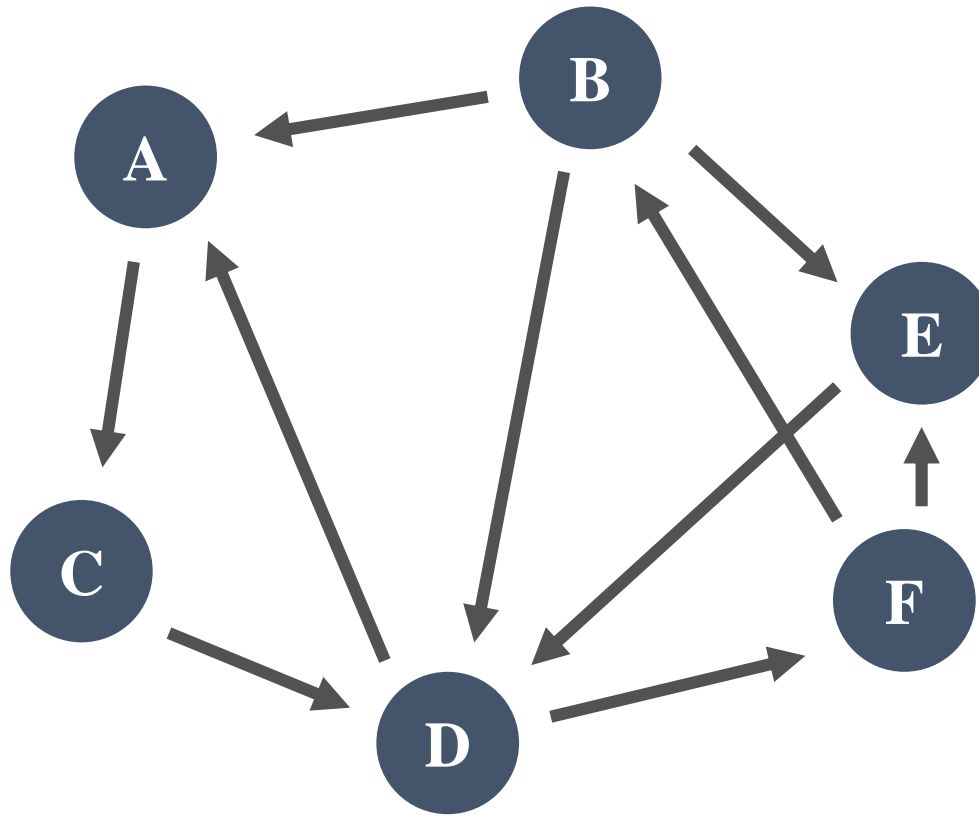
Hypothesis: “많이 참조(reference, citation)된 정보는 중요한 정보일 것이다”

PageRank

Concept

➤ PageRank:

다른 정보(웹 페이지)들이 특정 정보를 얼마나 많이 참조(link)하고 있는가를 기준으로 정보에 중요도를 부여하는 알고리즘

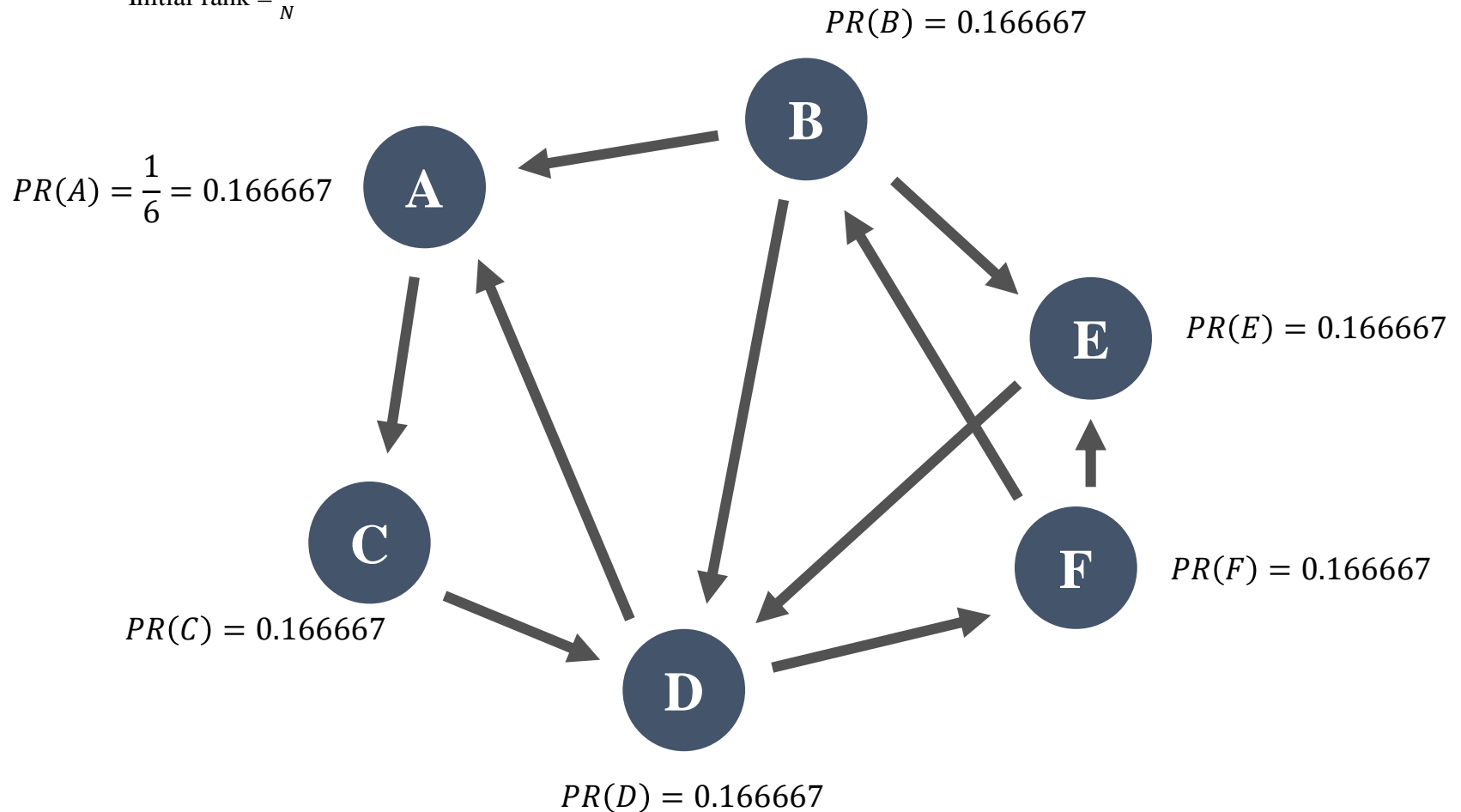


PageRank

Step 1

➤ Initialization

- Number of data $N = 6$
- Initial rank $= \frac{1}{N}$

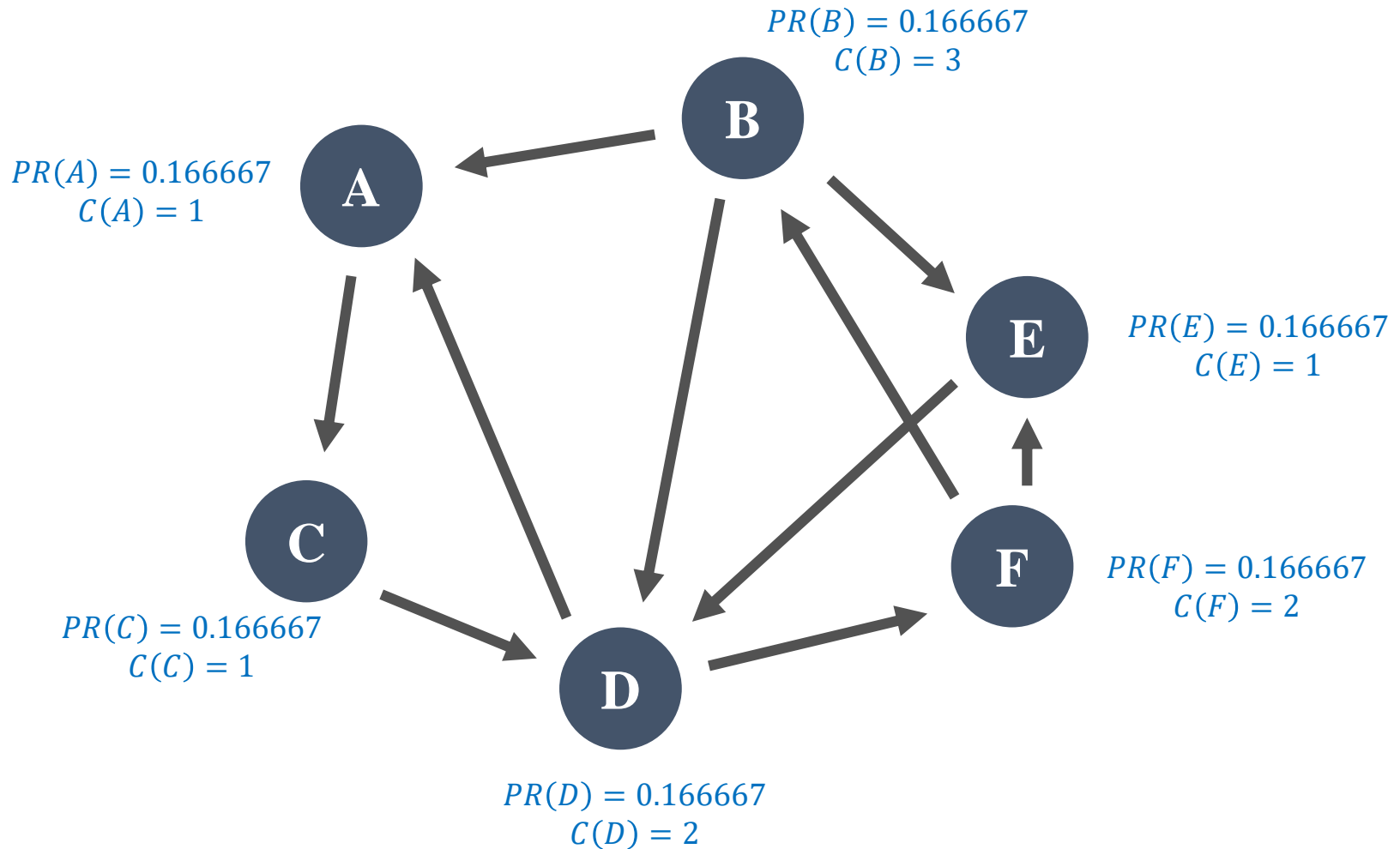


PageRank

Step 2

➤ Link Evaluation

- $PR(i)$: i 정보의 rank value
- $C(i)$: i 정보가 가지고 있는 Out-bound Link들의 개수

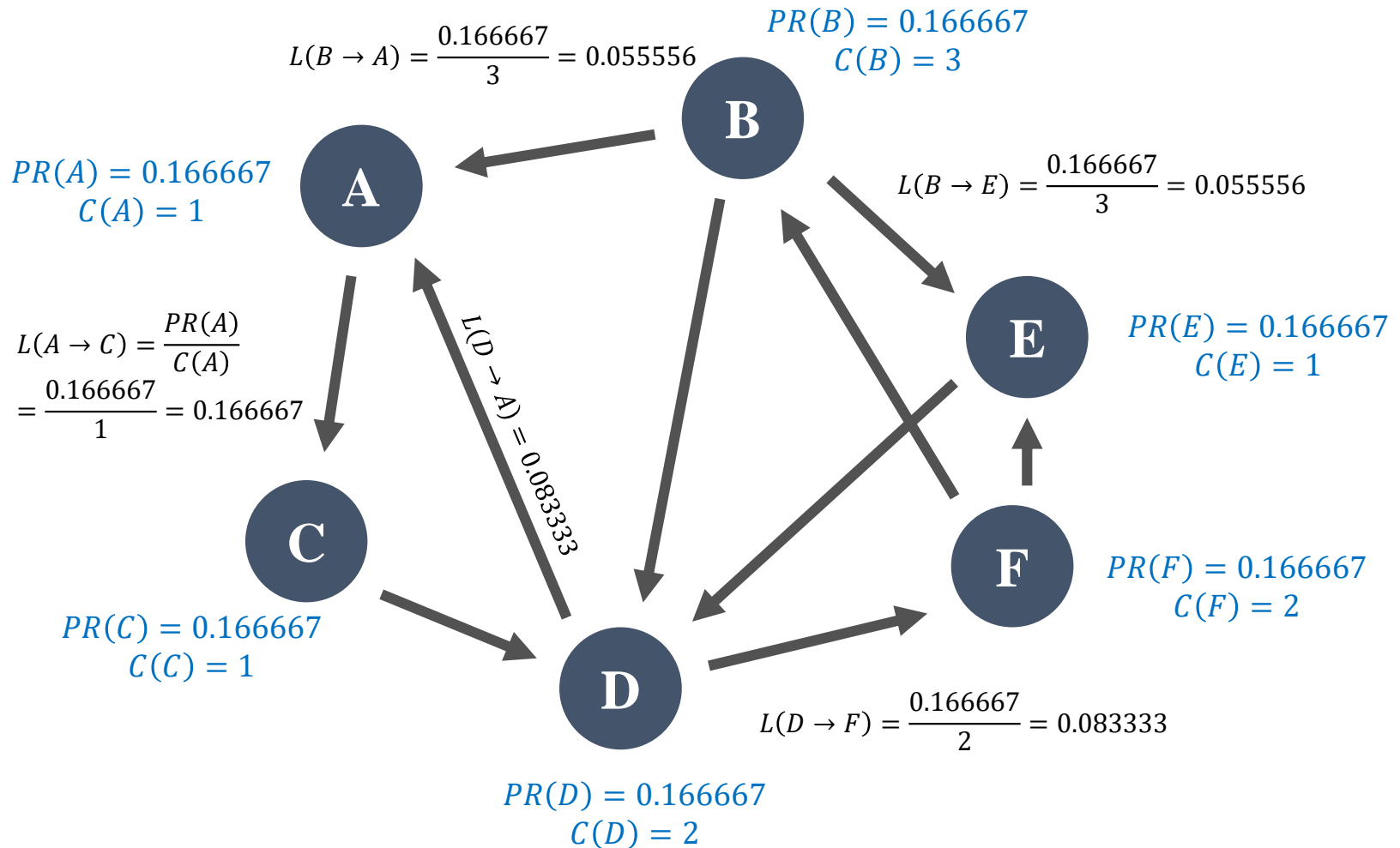


PageRank

Step 2

➤ Link Evaluation

- $PR(i)$: i 정보의 rank value
- $C(i)$: i 정보가 가지고 있는 Out-bound Link들의 개수

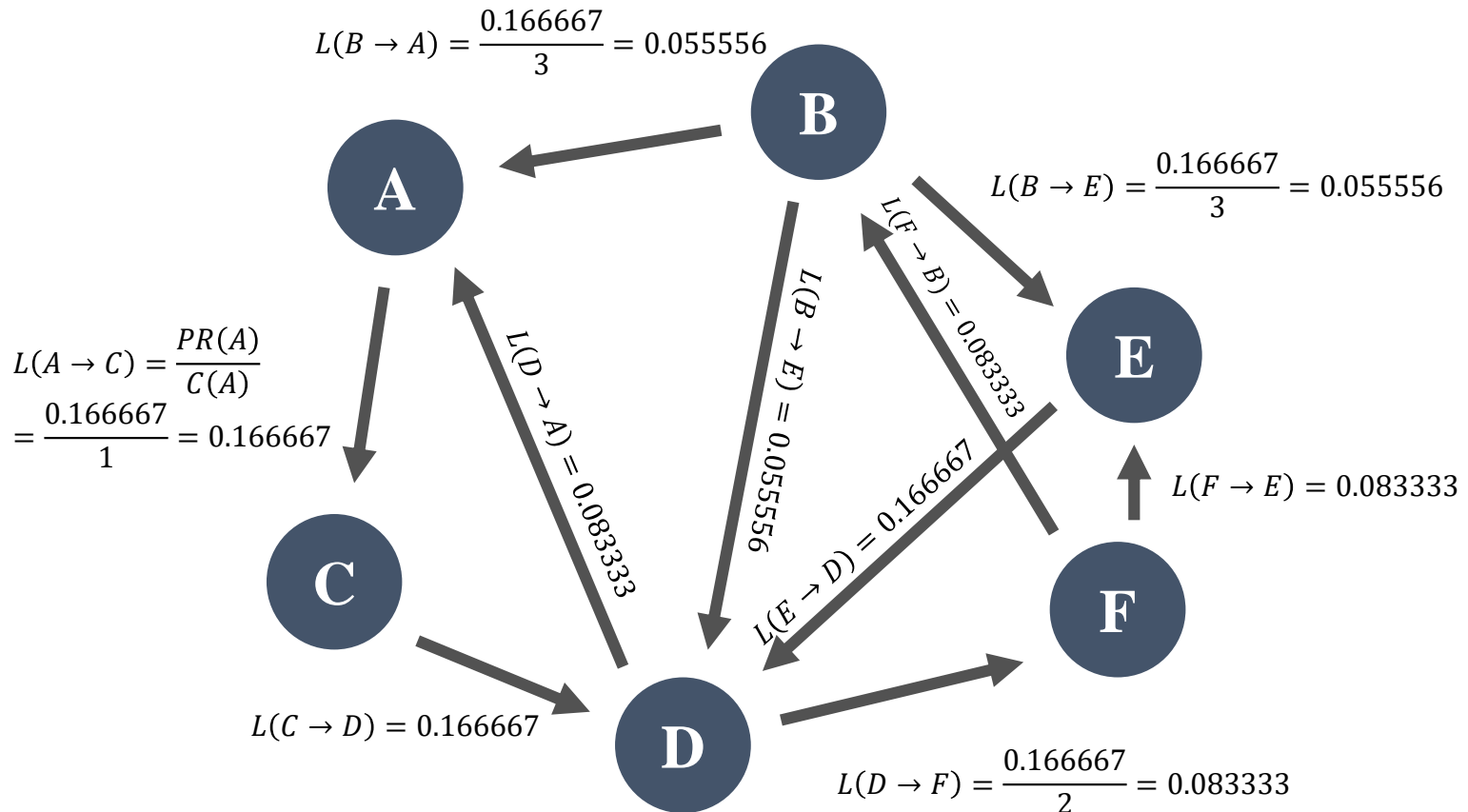


PageRank

Step 2

➤ Link Evaluation

- $PR(i)$: i 정보의 rank value
- $C(i)$: i 정보가 가지고 있는 Out-bound Link들의 개수



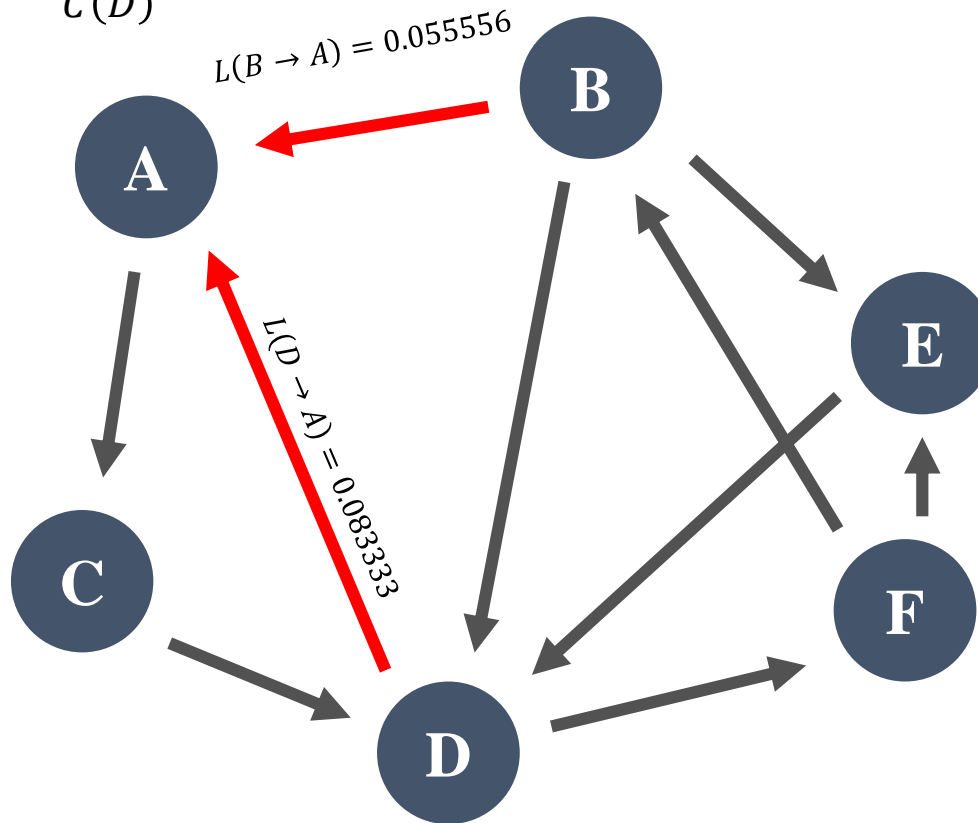
PageRank

Step 3

➤ Rank Computation

$$\blacksquare \quad PR(i) = \frac{PR(A)}{C(A)} + \frac{PR(B)}{C(B)} + \dots + \frac{PR(n)}{C(n)}$$

$$PR(A) = \frac{PR(B)}{C(B)} + \frac{PR(D)}{C(D)} = 0.055556 + 0.083333 = 0.138889$$

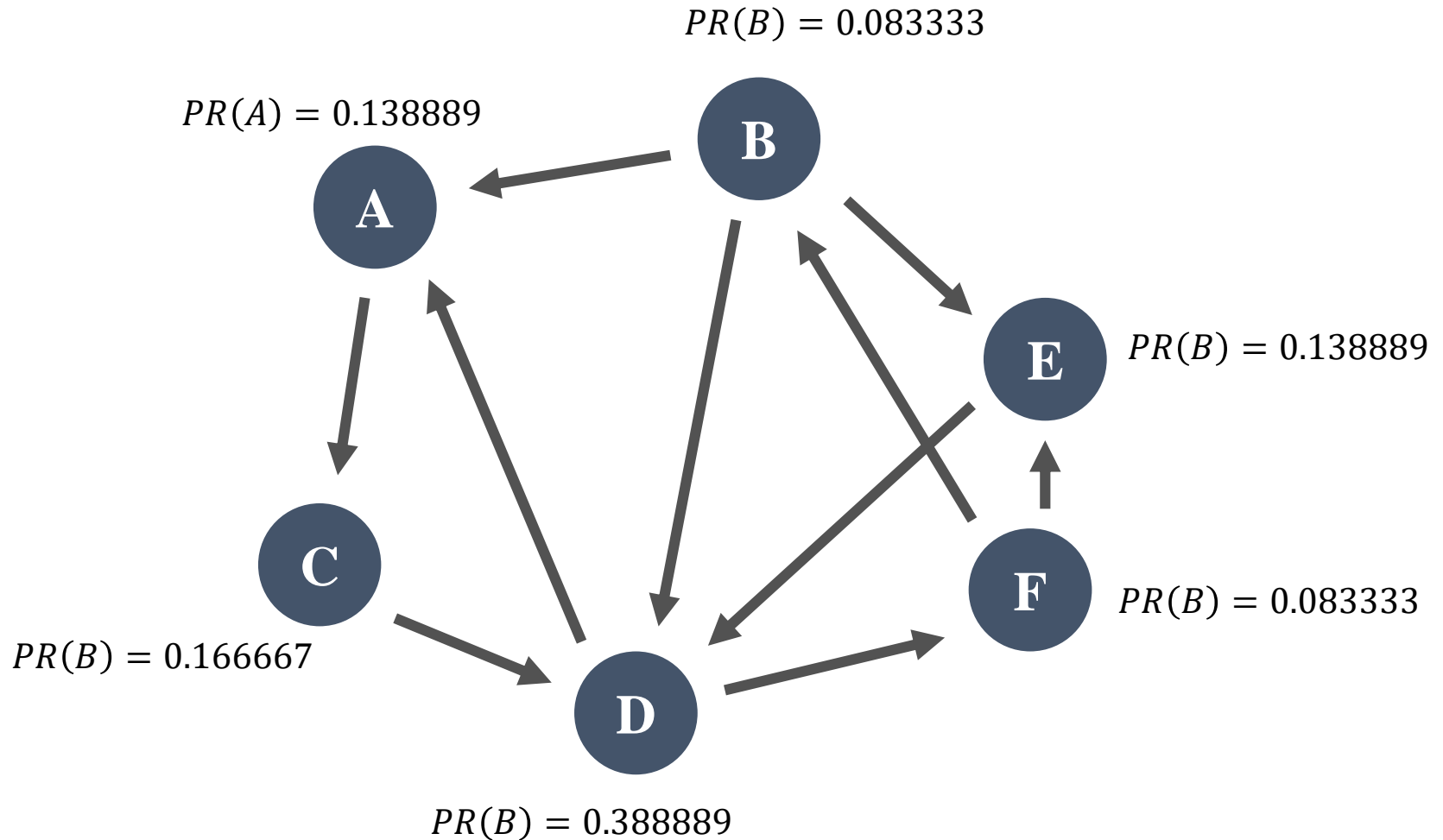


PageRank

Step 3

➤ Rank Computation

- $PR(i) = \frac{PR(A)}{c(A)} + \frac{PR(B)}{c(B)} + \dots + \frac{PR(n)}{c(n)}$



PageRank

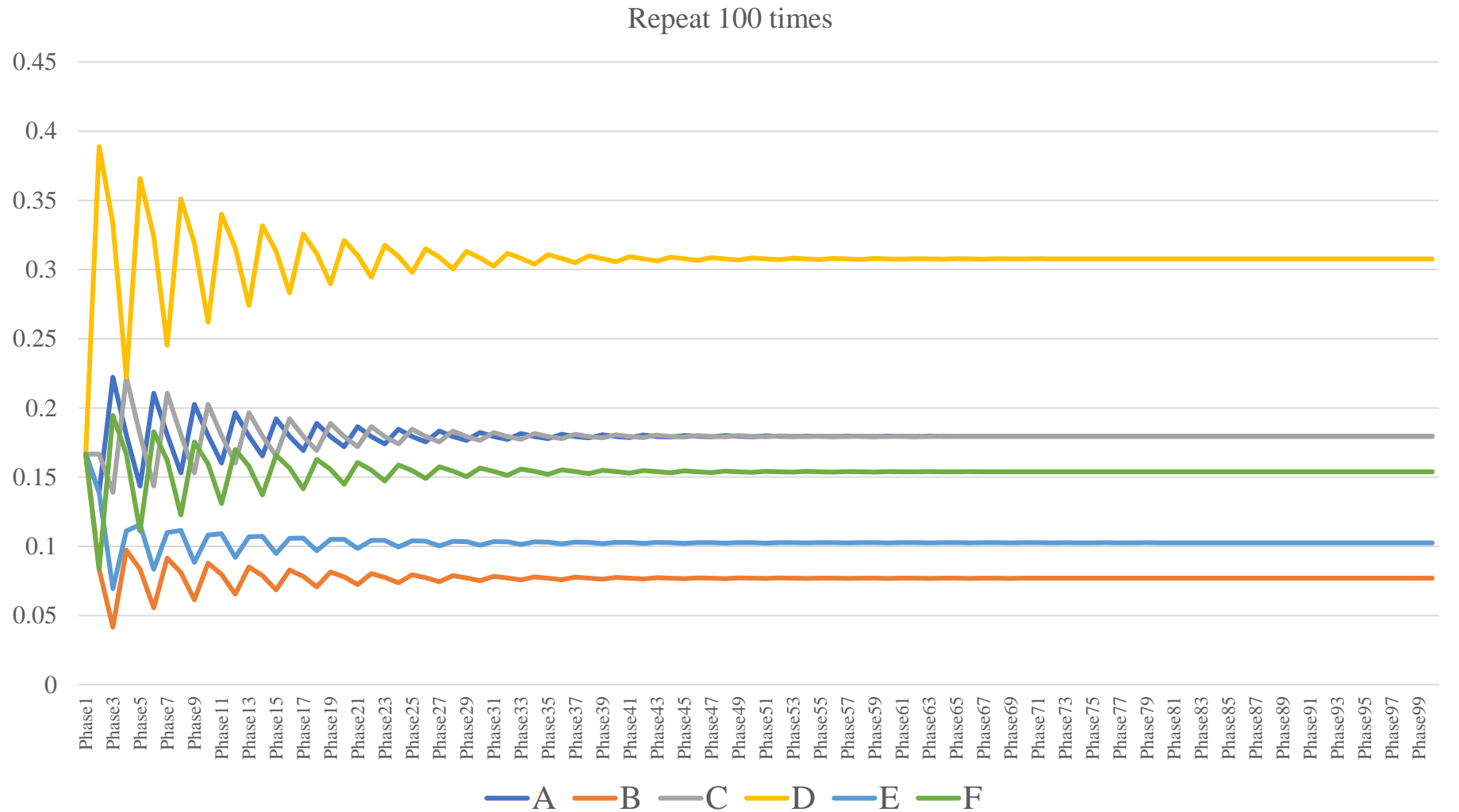
Algorithm Process

- Step 1: Initialize
- Step 2: Link value evaluation
- Step 3: Rank value computation
- Repeat Step 2~3 until the result converges

PageRank

Step 3

➤ Rank Evaluation



Limitations

computation time and impact of original data

- ❖ 데이터의 규모가 커질 경우, 각 정보들의 Rank 값이 수렴(**convergence**)하는데까지 많은 시간과 연산이 요구됨
- ❖ 정보들이 가진 Link에 의해서만 Rank 값을 산출할 경우, 정보 자체가 가지고 있는 중요도가 무시될 수 있음
- **Damping(감쇠) Factor**를 설정하여 결과 수렴 속도와 정보 자체가 가지고 있는 의미 (웹 사이트의 경우, 사용자가 검색 결과에 만족하지 않고, 웹 서핑을 계속해서 진행할 확률)를 부여함

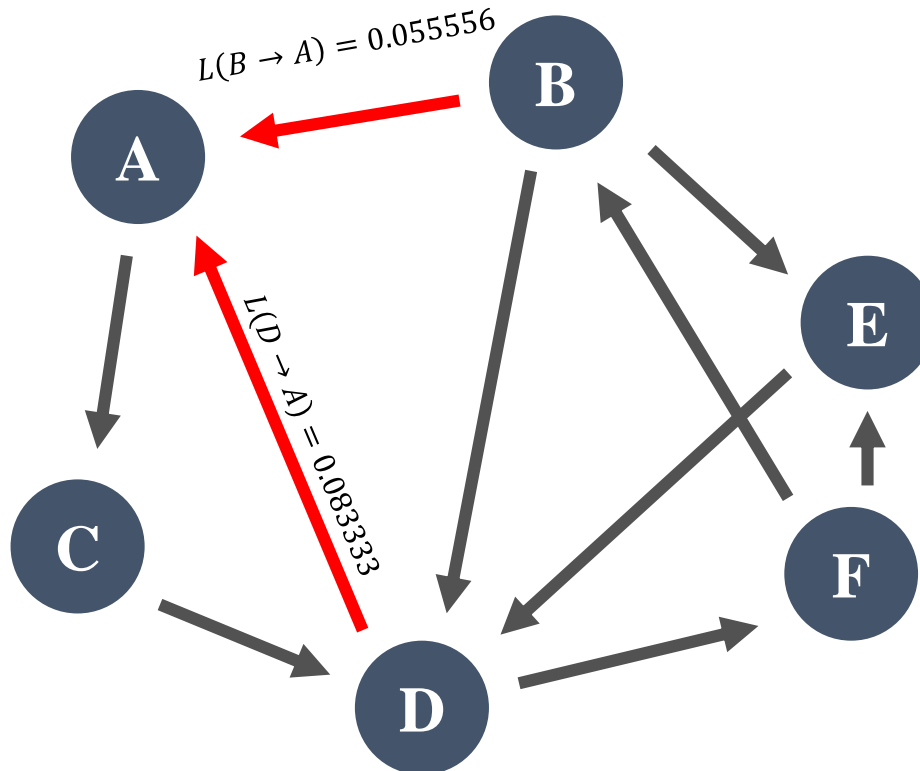
PageRank

Step 3 with damping factor d

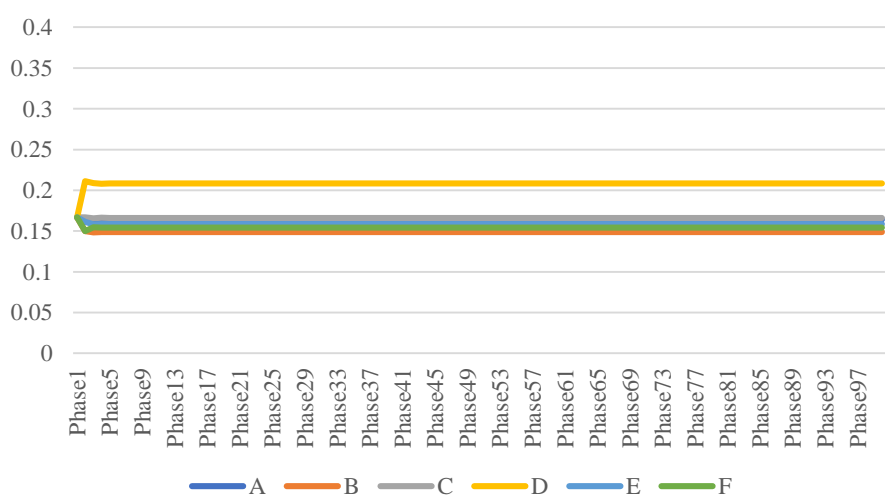
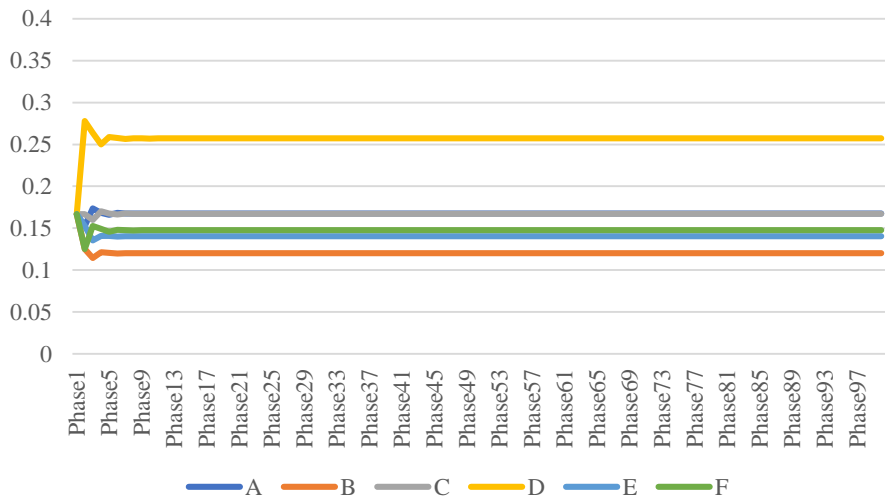
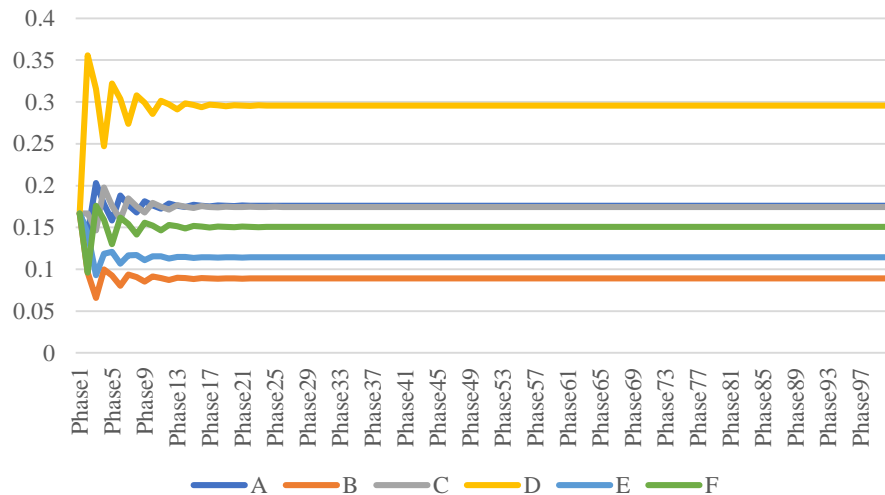
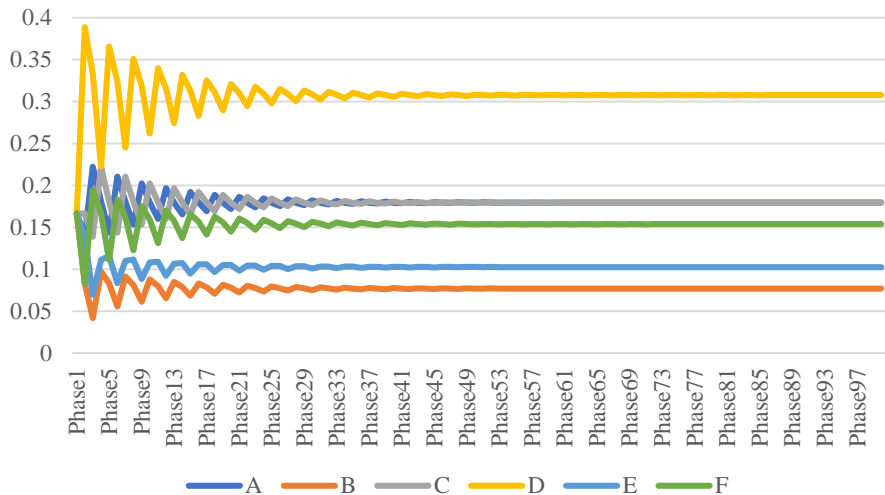
➤ Rank Computation

- $PR(i) = \frac{1-d}{N} + d \left(\frac{PR(A)}{C(A)} + \frac{PR(B)}{C(B)} \dots \frac{PR(n)}{C(n)} \right)$
- $d = 0.85$

$$PR(A) = \frac{1-d}{N} + d \left(\frac{PR(B)}{C(B)} + \frac{PR(D)}{C(D)} \right) = \frac{1-0.85}{6} + 0.85(0.055556 + 0.083333) = 0.14306$$

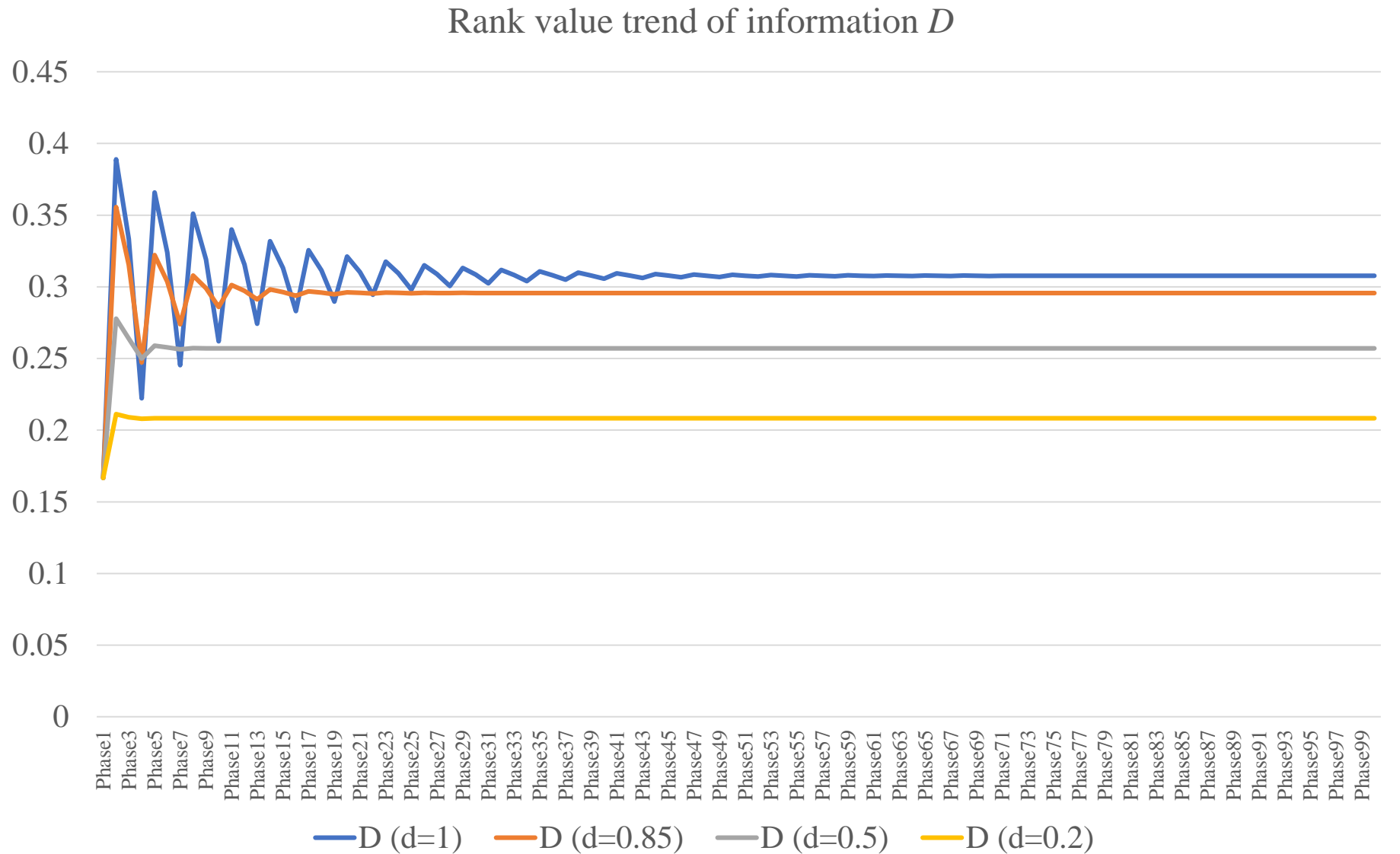


Impact of damping factor



PageRank

Impact of damping factor



Practice

What is the most risk IP address?

Data:

위협 가능성이 존재한다고 볼 수 있는 IP 들 간의
연관 관계 데이터 10,495건

Setting:

damping factor = 0.99

Print Format:

가장 높은 rank를 가진 10개의 IP를 출력,
반복횟수, 총 걸린 시간(초) 출력

Print Example:

1.1.1.1

2.2.2.2

3.3.3.3

...

10.10.10.10

iteration: 100

time: 2.3(s)

	A	B	C	
1	141.8.224.72	188.226.194.251		
2	174.128.255.230	183.136.132.177		
3	174.128.255.232	183.136.132.177		
4	50.117.115.89	183.136.132.177		
5	174.128.255.230	183.136.132.177		
6	112.121.175.139	185.27.134.126		
7	8.5.1.32	180.178.50.242		
8	180.178.63.75	182.16.28.60		
9	183.129.177.209	183.129.177.209		
10	183.129.177.209	183.129.177.209		
11	183.129.177.209	183.129.177.209		
12	183.129.177.209	183.129.177.209		
13	183.129.177.209	183.129.177.209		
14	183.129.177.209	183.129.177.209		
15	183.129.177.209	183.129.177.209		
16	183.129.177.209	183.129.177.209		
17	183.129.177.209	183.129.177.209		
18	183.129.177.209	183.129.177.209		
19	183.129.177.209	183.129.177.209		
20	125.39.187.2	183.129.177.209		
21	122.143.8.173	183.129.177.209		

Thank you

