그래프를 이용한 기계 학습

#4 그래프를 바이럴 마케팅에 어떻게 활용할까?

신기정
(KAIST AI대학원)

boostcamp AI Tech

1. 그래프를 통한 전파의 예시

2. 의사결정 기반의 전파 모형

3. 확률적 전파 모형

4. 바이럴 마케팅과 전파 최대화 문제

5. 실습: 전파 모형 시뮬레이터 구현

1. 그래프를 통한 전파의 예시

1.1 그래프를 통한 정보의 전파
1.2 그래프를 통한 행동의 전파
1.3 그래프를 통한 고장의 전파
1.4 그래프를 통한 질병의 전파

1.1 그래프를 통한 정보의 전파

온라인 소셜 네트워크를 통해 다양한 정보가 전파됩니다

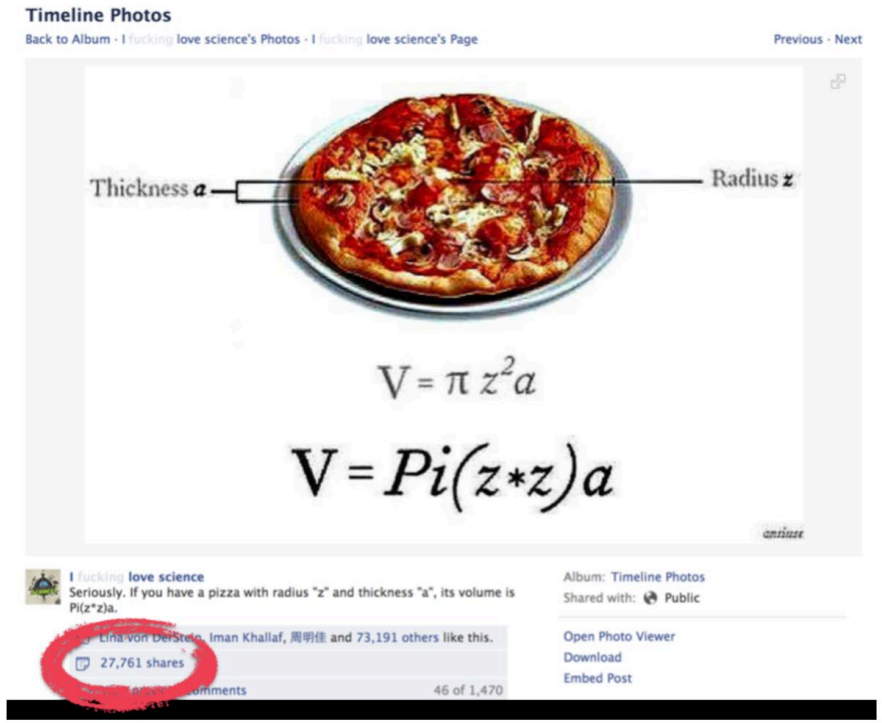
2011년 스페인의 15M 운동에 대한 정보는 트위터를 통해 전국적으로 알려졌습니다
덕분에 주류 언론이 침묵하는 상황에서도, 시민들이 정부의 부정부패에 맞서 연대할 수 있었습니다

×

1.1 그래프를 통한 정보의 전파

온라인 소셜 네트워크를 통해 다양한 정보가 전파됩니다

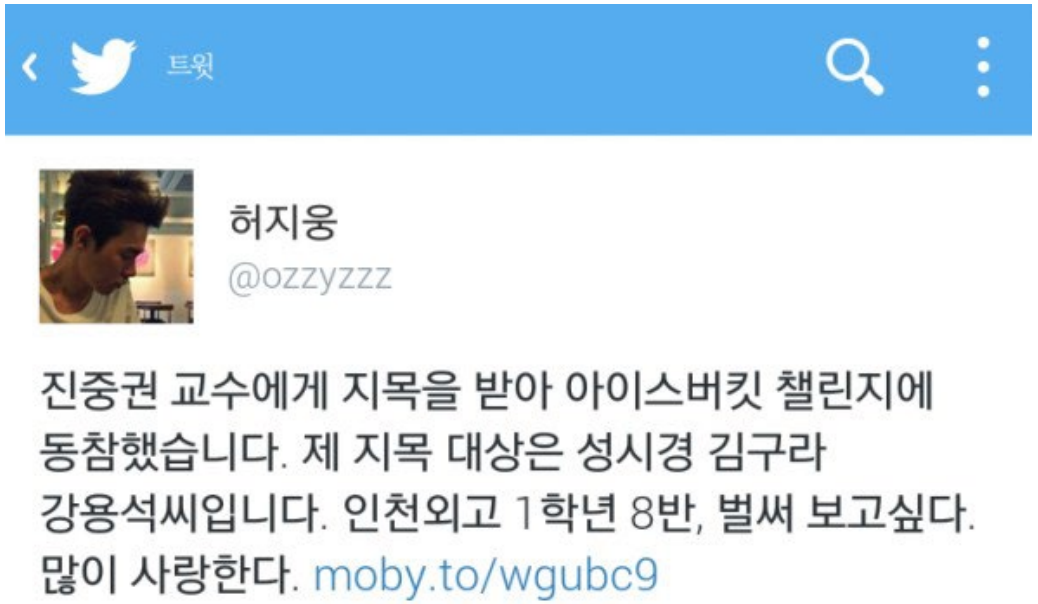
유용한 과학적 정보가 전파되기도 합니다

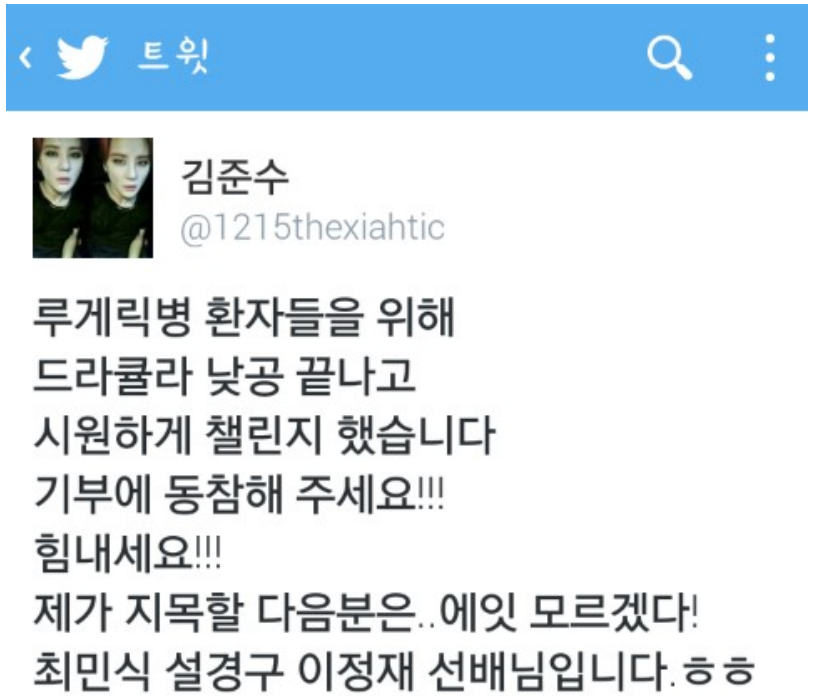


1.2 그래프를 통한 행동의 전파

온라인 소셜 네트워크를 통해 다양한 행동도 전파됩니다

아이스 버킷 챌린지, 펭귄 문제 등이 대표적인 예시입니다





1.2 그래프를 통한 행동의 전파

온라인 소셜 네트워크를 통해 다양한 행동도 전파됩니다

아이스 버킷 챌린지, 펭귄 문제 등이 대표적인 예시입니다

[펭귄 문제]

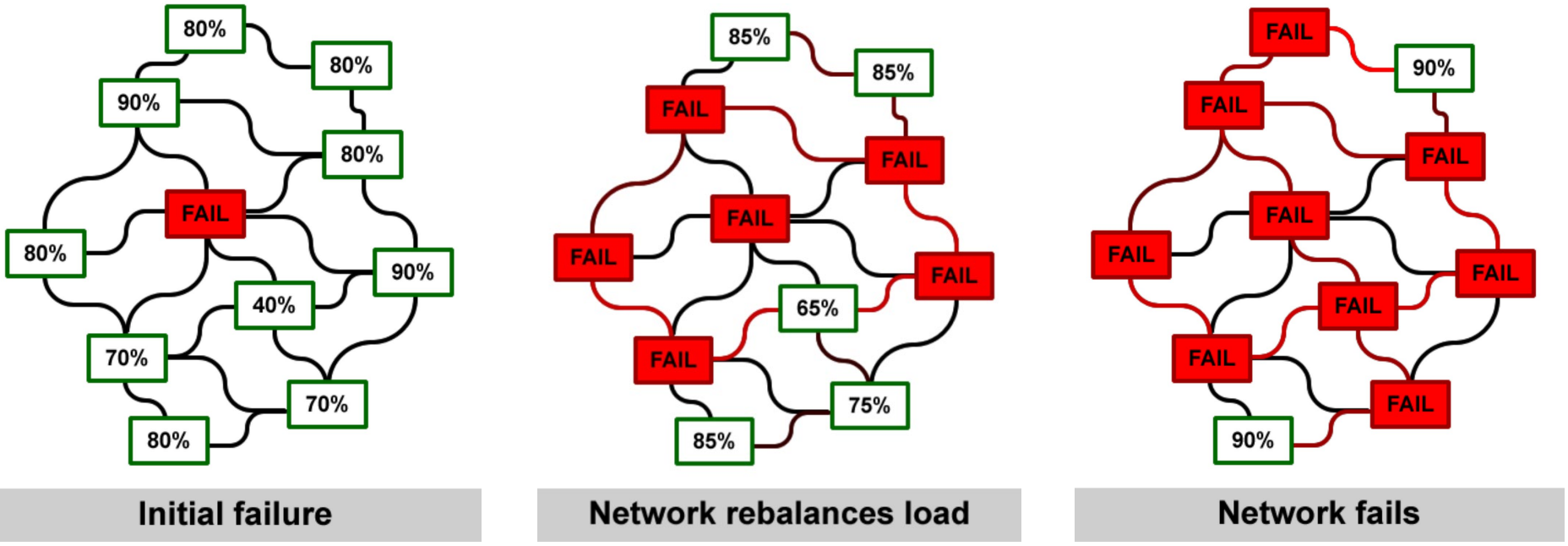
틀리면 3일간 펭귄 프사로 살아야 합니다.
정답은 아무에게도 말하지 마세요.

문제. 부대찌개 3인분을 먹으면 부대찌개 1인분을 서비스로 제공하는 식당이 있다. 부대찌개 20인분을 시키면 몇인분을 먹을 수 있는가?

1.3 그래프를 통한 고장의 전파

컴퓨터 네트워크에서의 일부 장비의 고장이 전파되어 전체 네트워크를 마비시킬 수 있습니다

일부 장비의 고장이, 다른 장비의 과부화로 이어지기 때문입니다



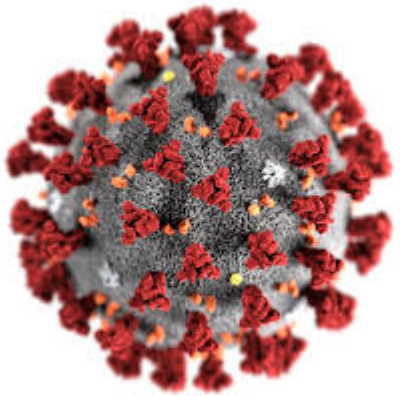
1.3 그래프를 통한 고장의 전파

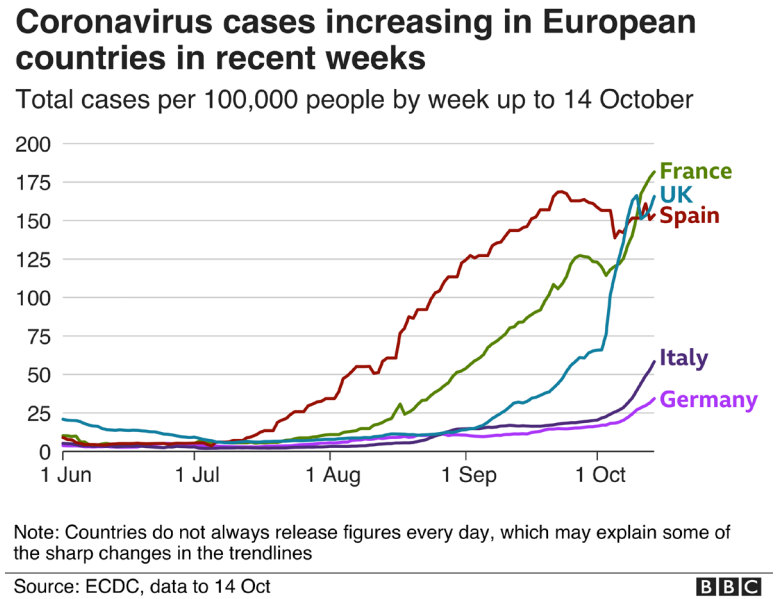
파워 그리드에서의 정전이 전파되는 과정도 유사합니다

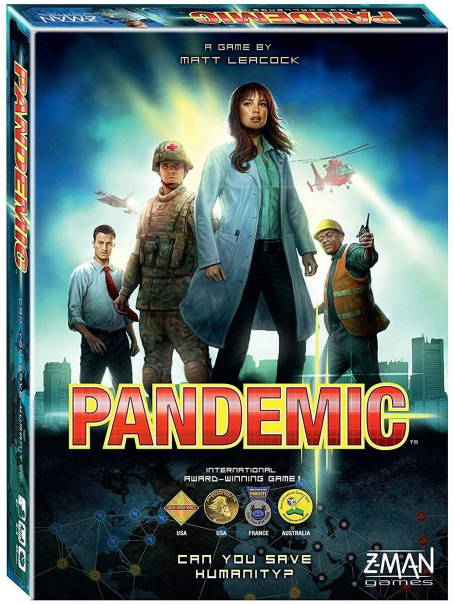
1.4 그래프를 통한 질병의 전파

사회라는 거대한 소셜 네트워크를 통한 질병의 전파도 빠뜨릴 수 없습니다

코로나-19, 메르스, 사스 등이 그 예시입니다







1.4 그래프를 통한 질병의 전파

전파 과정은 다양할 뿐 아니라 매우 복잡합니다
이를 체계적으로 이해하고 대처하기 위해서는 수학적 모형화가 필요합니다

본 수업에서는 전파 과정을 위한 수 많은 모형 중 두 가지를 소개합니다

2. 의사결정 기반의 전파 모형

2.1 언제 의사결정 기반의 전파 모형을 사용할까?

2.2 선형 임계치 모형

2.1 언제 의사결정 기반의 전파 모형을 사용할까?

1970년대에는 VHS와 Betamax라는 호환되지 않는 두 종류의 비디오 유형이 있었습니다

어떤 유형의 비디오 플레이어를 구매하시겠습니까?
의사 결정을 위해 어떤 정보를 참고해야 할까요?

2.1 언제 의사결정 기반의 전파 모형을 사용할까?

그러면 카카오톡과 라인 중에 어떤 것을 사용하시고 있나요? 왜 그런 결정을 하셨나요?

2.1 언제 의사결정 기반의 전파 모형을 사용할까?

두 경우 모두 주변 사람들의 의사결정이 본인의 의사결정에 영향을 미칩니다

친구들이 대부분 라인을 쓴다면, 카카오톡을 사용하는 것이 불편합니다
친구들이 대부분 VHS 유형을 쓴다면, 같은 유형을 써야 서로 비디오를 빌려줄 수 있습니다

이렇듯 주변 사람들의 의사결정을 고려하여 각자 의사결정을 내리는 경우에
의사결정 기반의 전파 모형을 사용합니다

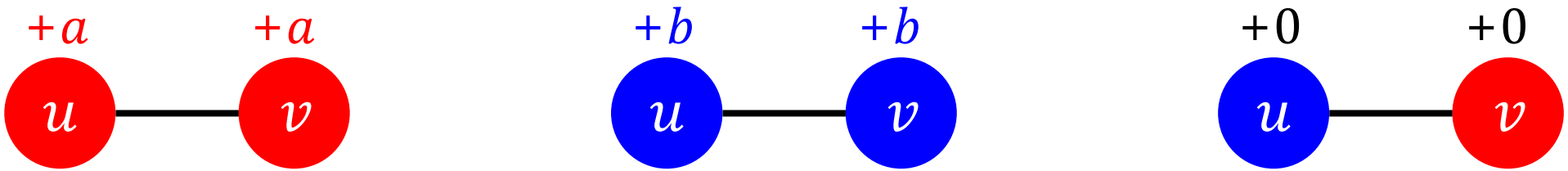
본 수업에서는 가장 간단한 형태의 의사결정 기반의 전파 모형인
선형 임계치 모형(Linear Threshold Model)을 소개합니다

2.2 선형 임계치 모형

이 상황을 수학적으로 추상화 해봅시다

친구 관계의 두 사람 u와 v를 가정합시다
둘은 두 개의 호환되지 않는 기술 A와 B 중에서 하나를 선택합니다

둘 모두 A 기술을 사용할 경우, 행복이 a만큼 증가합니다
둘 모두 B 기술을 사용할 경우, 행복이 b만큼 증가합니다
하지만, 둘이 서로 다른 기술을 사용할 경우, 행복이 증가하지 않습니다



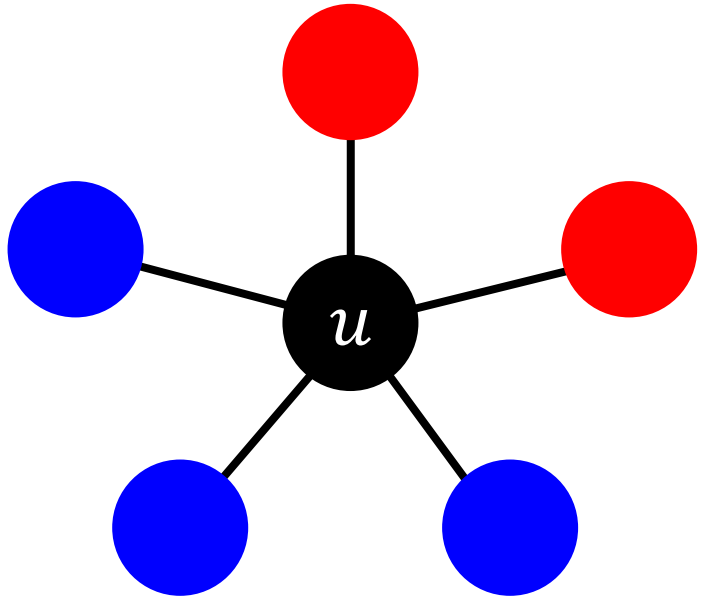
2.2 선형 임계치 모형

소셜 네트워크를 고려해봅시다

우리는 동시에 여러 사람과 친구 관계를 맺습니다
각각의 친구, 즉 **소셜 네트워크 상의 이웃**과의 사이에서 발생하는 행복을 고려해야합니다

오른쪽 예시에서 u가 A를 선택할 경우
행복이 $2a$ 만큼 증가합니다

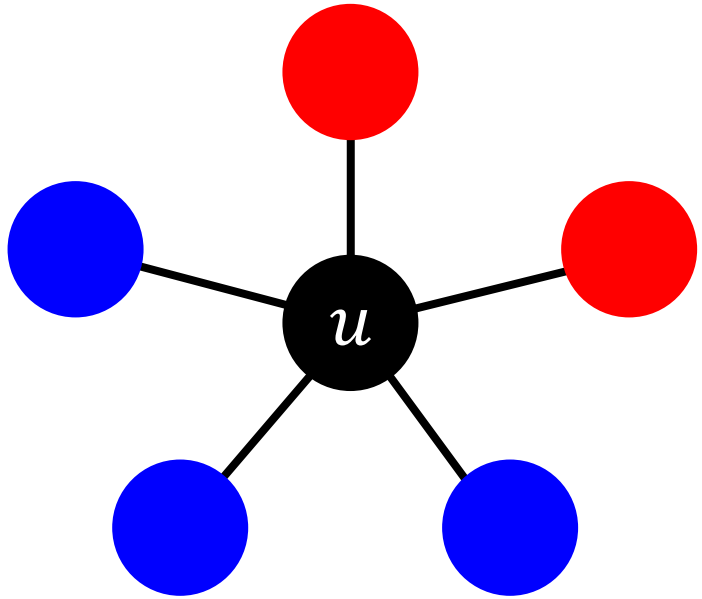
오른쪽 예시에서 u가 B를 선택할 경우
행복이 $3b$ 만큼 증가합니다



2.2 선형 임계치 모형

각자가 행복이 최대화되는 선택을 한다고 가정해봅시다

만약, $2a > 3b$ 라면 u는 A를 택할 것입니다
반면, $2a < 3b$ 라면 u는 B를 택할 것입니다
편의상 $2a = 3b$ 라면 u는 B를 택한다고 합시다



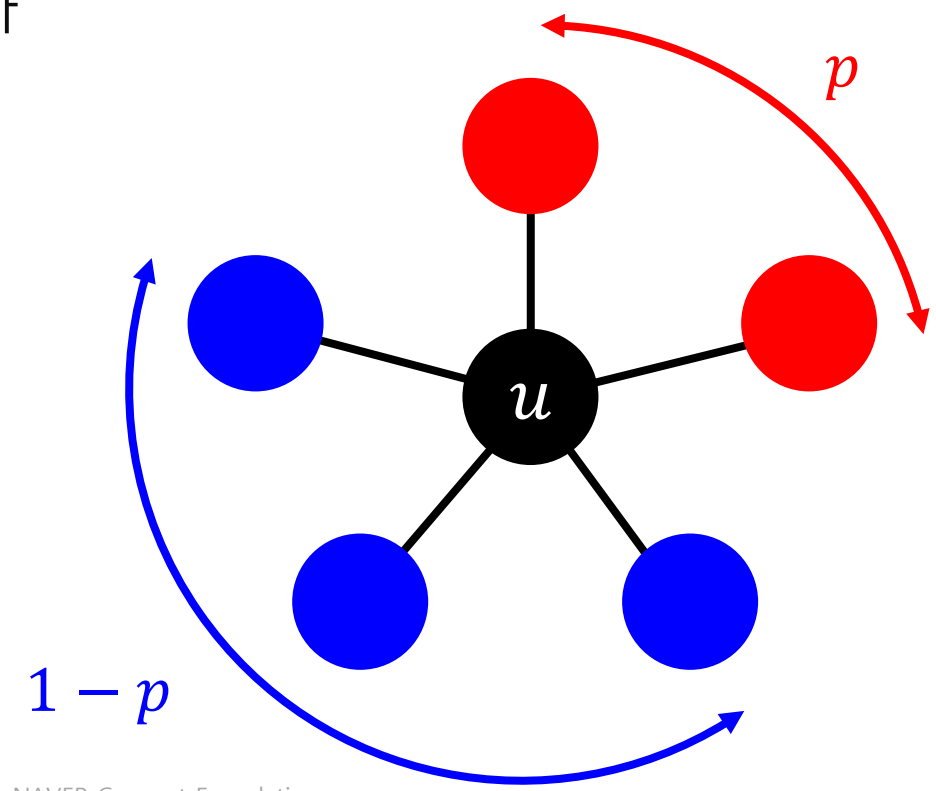
2.2 선형 임계치 모형

좀 더 일반화 해봅시다

p 비율의 이웃이 A를 선택했다고 해봅시다
즉, $1-p$ 비율의 이웃이 B를 선택했습니다

언제 A를 선택할까요?
$ap > b(1-\mathrm{p})$ 일 때입니다

정리하면
$p > \frac{b}{a+b}$ 일 때입니다
편의상 $\frac{b}{a+b}$ 를 임계치 q라고 합시다



2.2 선형 임계치 모형

이 모형을 선형 임계치 모형(Linear Threshold Model)이라고 합니다

각 정점은 이웃 중 A를 선택한 비율이 임계치 q를 넘을 때만 A를 선택합니다

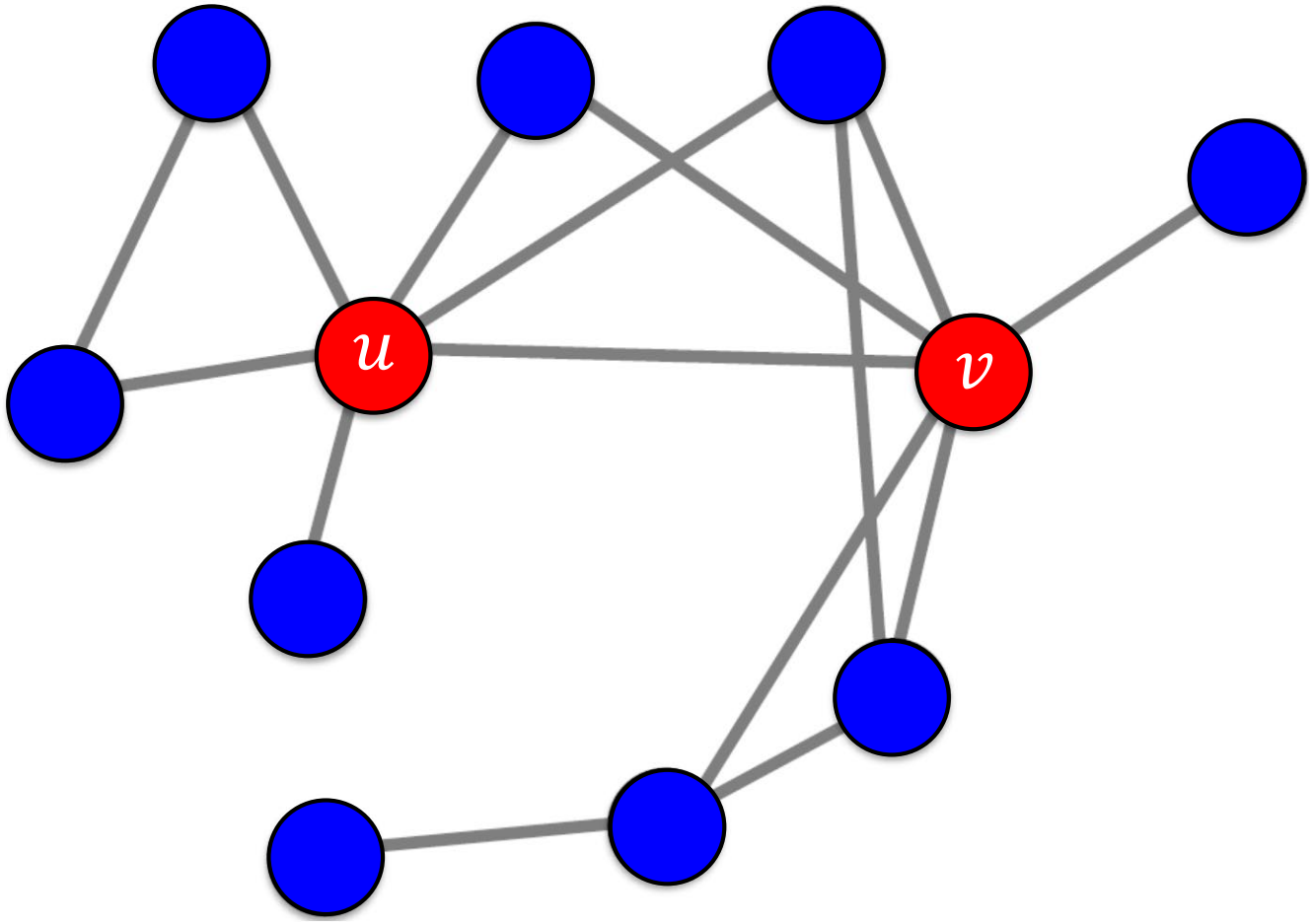
이 모형은 전부 B를 사용하는 상황을 가정합니다
그리고 처음 A를 사용하는 얼리 어답터들을 가정합니다
시드 집합(Seed Set)이라고 불리는 얼리 어답터들은 항상 A를 고수한다고 가정합시다

2.2 선형 임계치 모형

예시를 고려합시다

임계치 q는 55%를 가정합시다
u와 v가 시드 집합인 상황입니다

u와 v의 선택을 고려하여
각자 다시 기술을 선택합니다

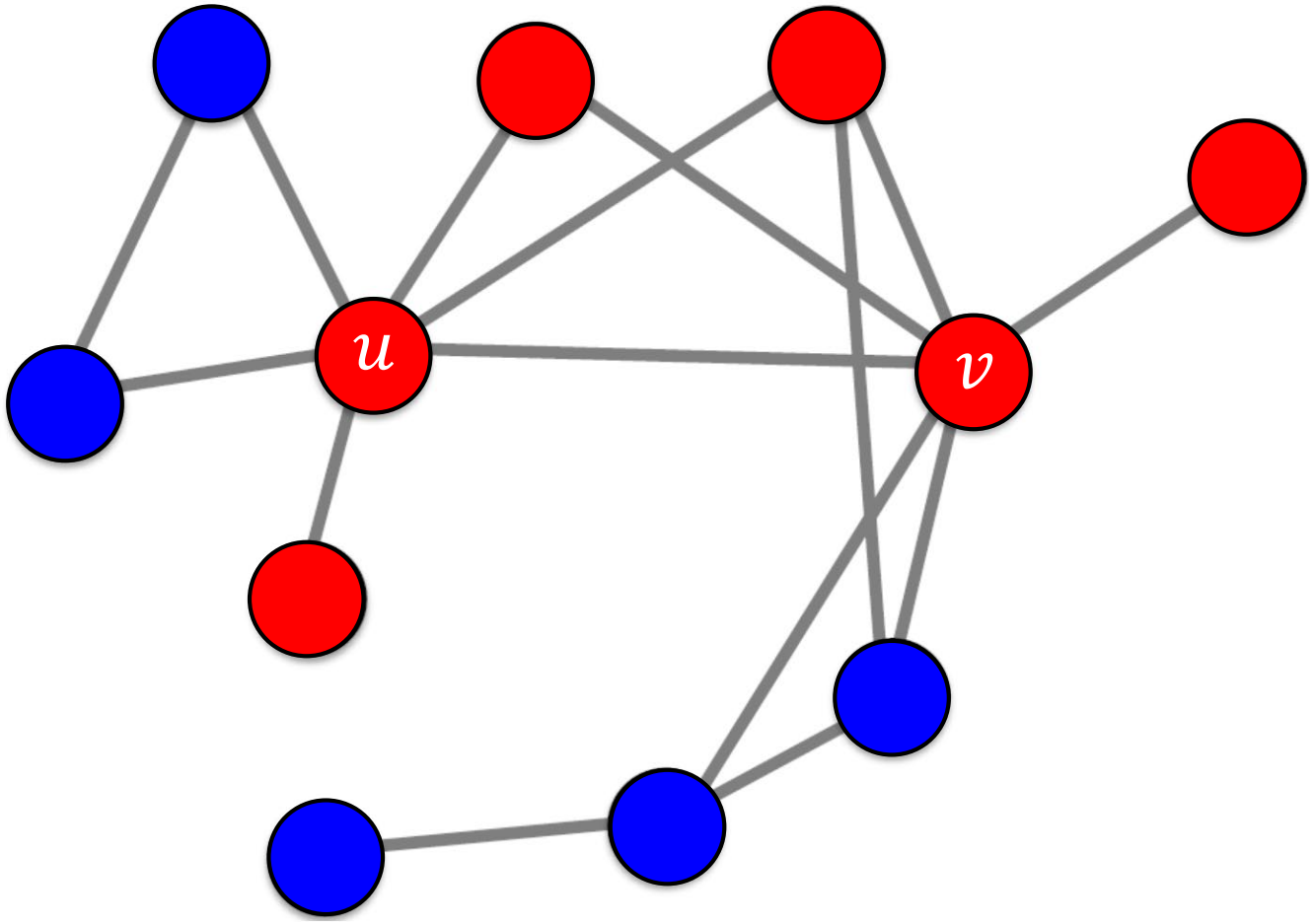


이미지 출처: https://web.stanford.edu/class/cs224w/slides/12-cascades.pdf

2.2 선형 임계치 모형

추가로 4명이 A를 선택했습니다

이웃 중 A를 선택한 비율이
임계치 55%를 넘었기 때문입니다

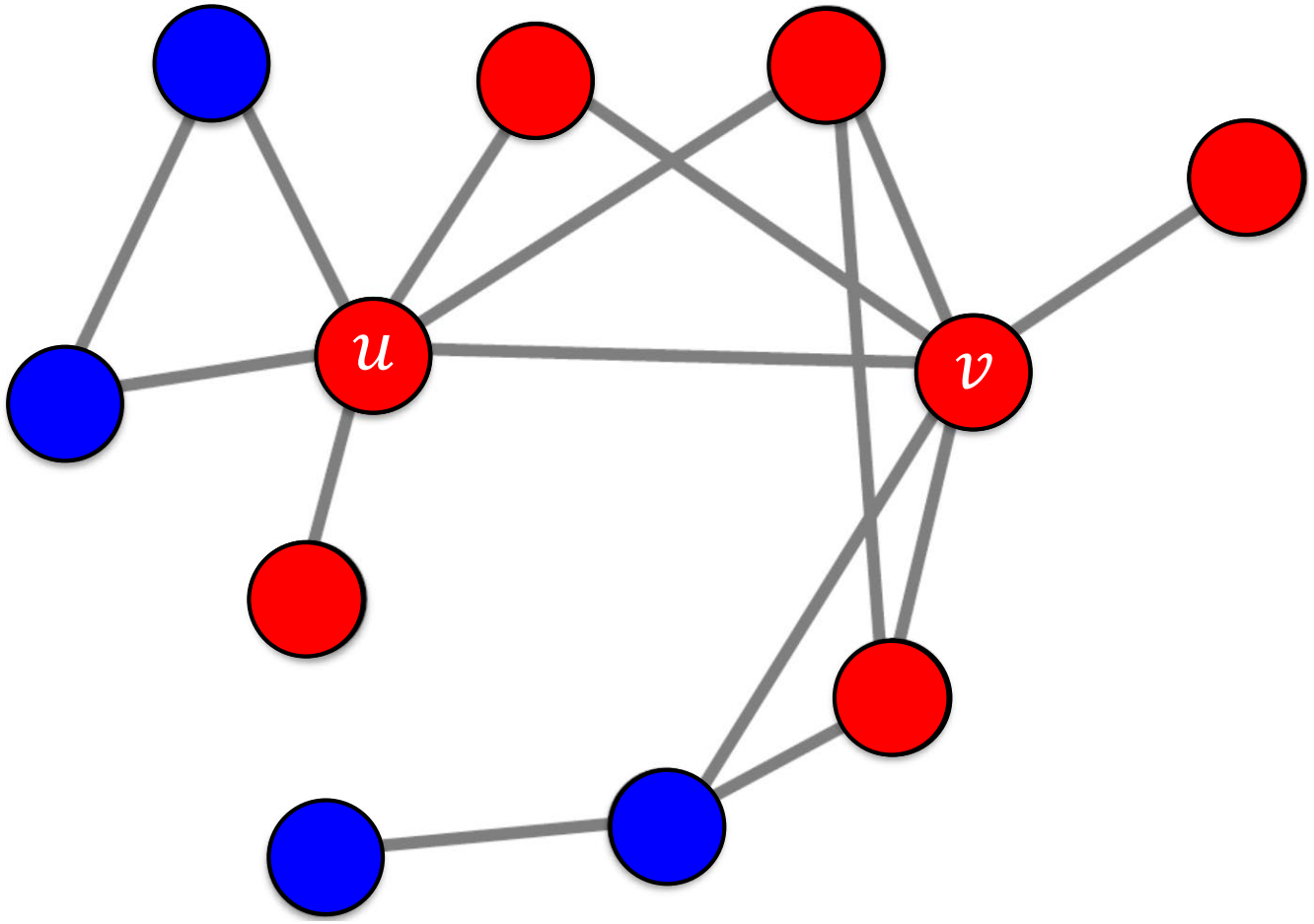


이미지 출처: https://web.stanford.edu/class/cs224w/slides/12-cascades.pdf

2.2 선형 임계치 모형

추가로 1명이 A를 선택했습니다

이웃 중 A를 선택한 비율이
임계치 55%를 넘었기 때문입니다

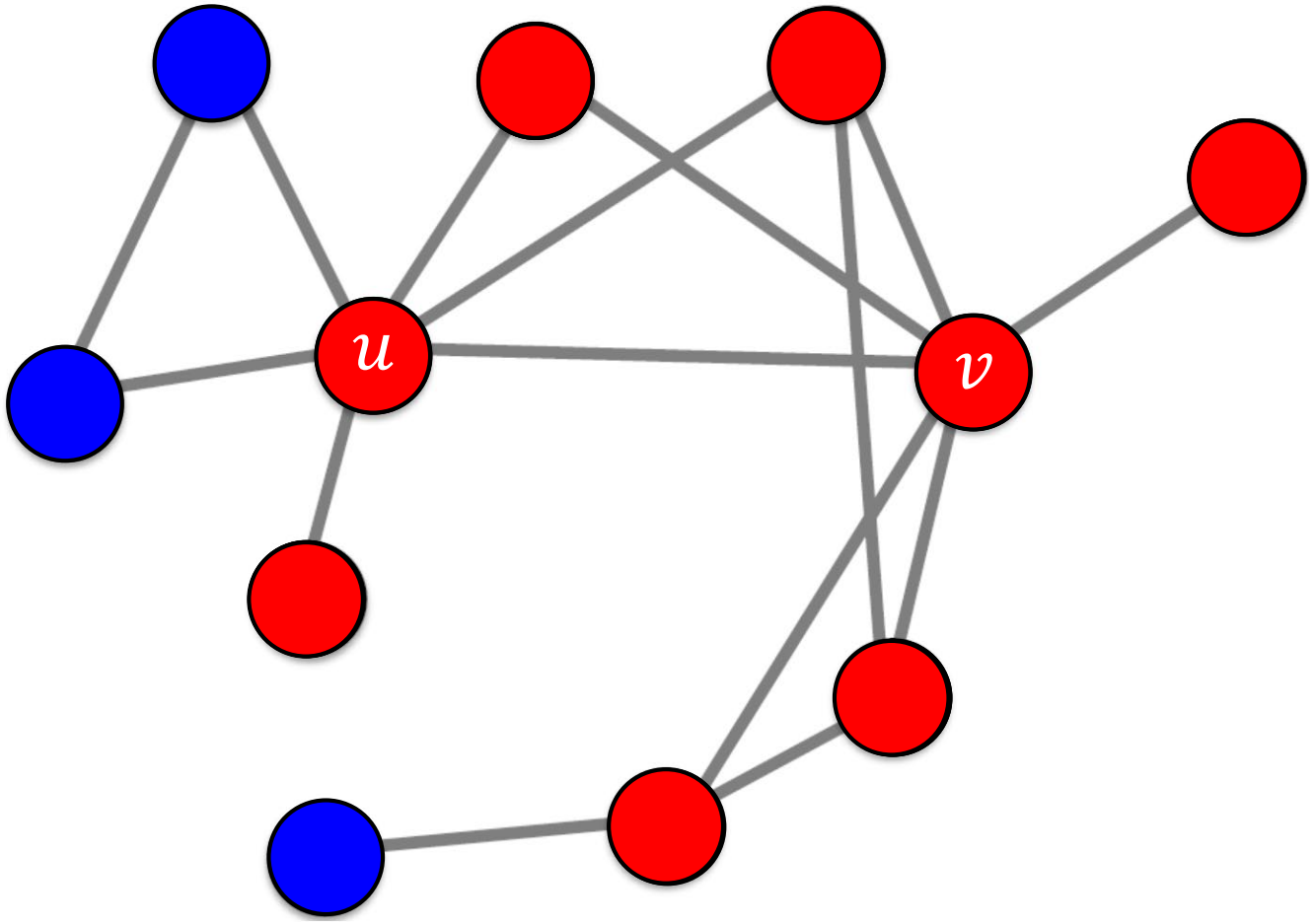


이미지 출처: https://web.stanford.edu/class/cs224w/slides/12-cascades.pdf

2.2 선형 임계치 모형

추가로 1명이 A를 선택했습니다

이웃 중 A를 선택한 비율이
임계치 55%를 넘었기 때문입니다

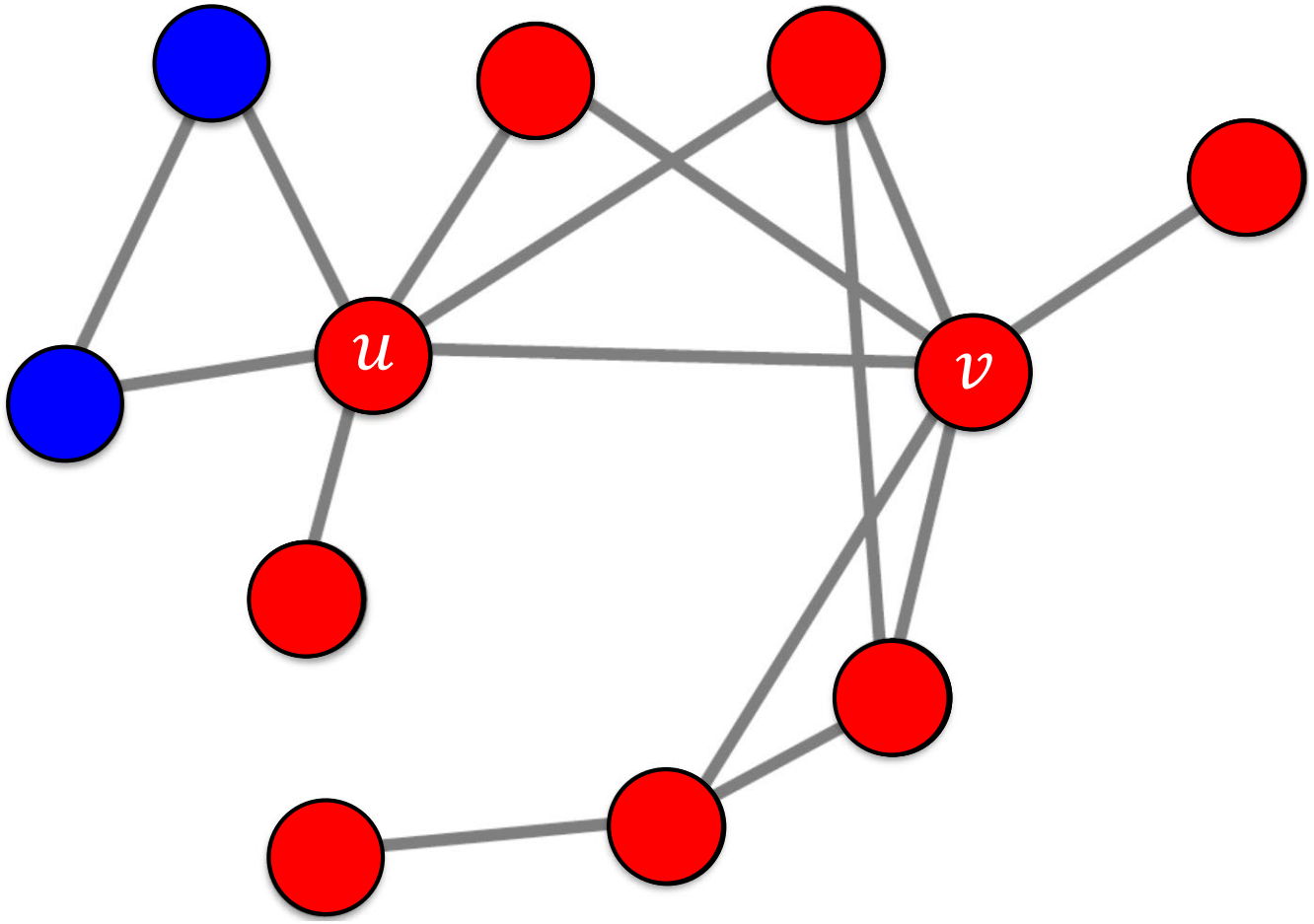


이미지 출처: https://web.stanford.edu/class/cs224w/slides/12-cascades.pdf

2.2 선형 임계치 모형

추가로 1명이 A를 선택했습니다

이웃 중 A를 선택한 비율이
임계치 55%를 넘었기 때문입니다

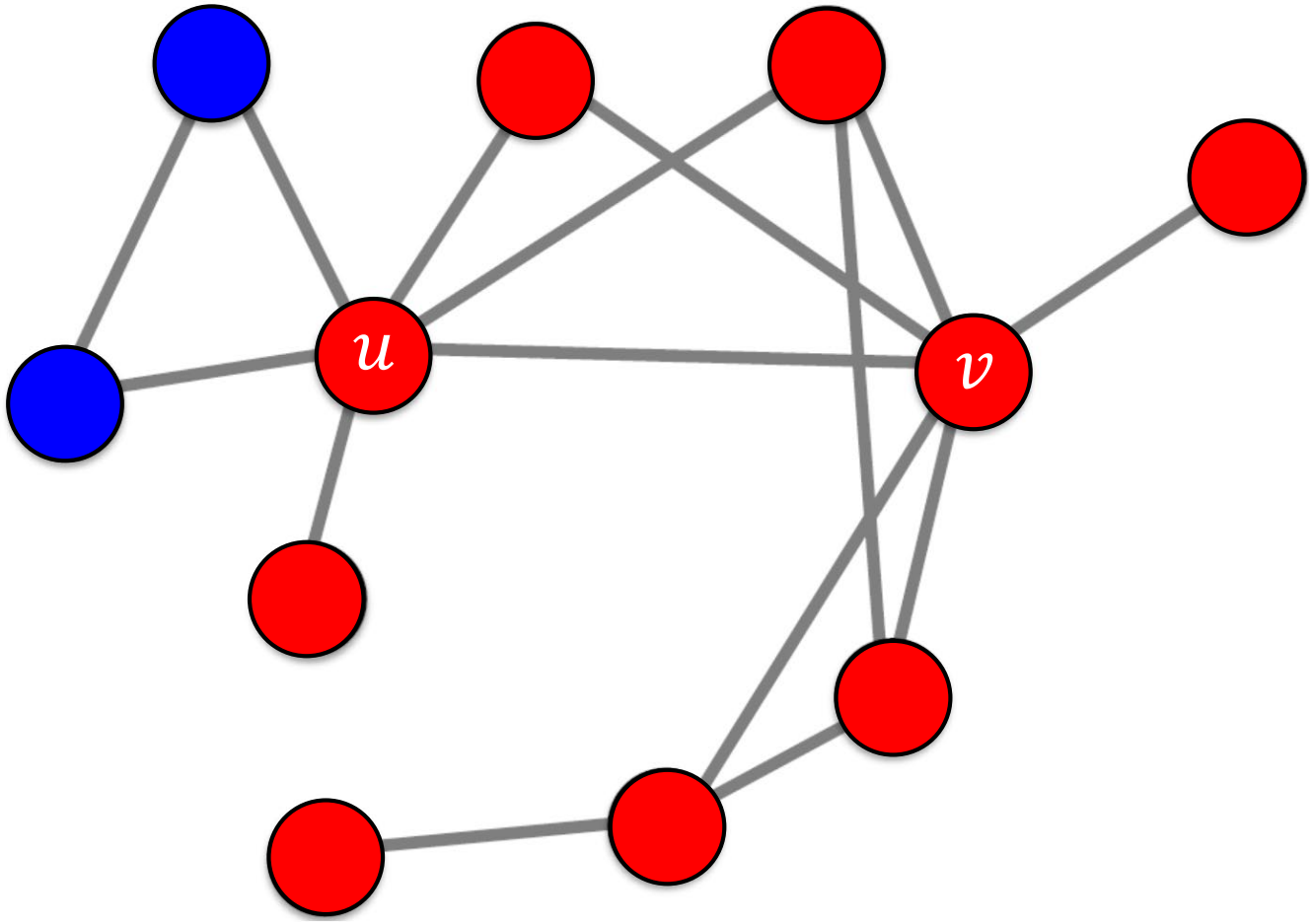


이미지 출처: https://web.stanford.edu/class/cs224w/slides/12-cascades.pdf

2.2 선형 임계치 모형

전파가 멈추었습니다

A를 택하지 않은 정점 중,
이웃 중 A를 선택한 비율이
임계치 55%를 넘는 경우가
없기 때문입니다



이미지 출처: https://web.stanford.edu/class/cs224w/slides/12-cascades.pdf

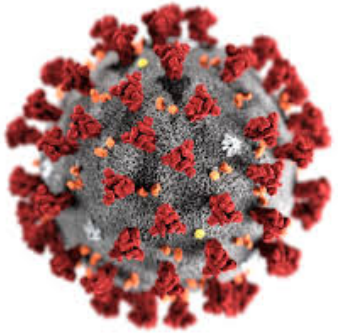
3. 확률적 전파 모형

3.1 언제 확률적 전파 모형을 사용할까?

3.2 독립적 전파 모형

3.1 언제 확률적 전파 모형을 사용할까?

코로나의 전파 과정을 수학적으로 추상화해봅시다



의사결정 기반 모형은 적합하지 않습니다
누구도 코로나에 걸리기로 '의사결정'을 내리는 사람은 없기 때문입니다

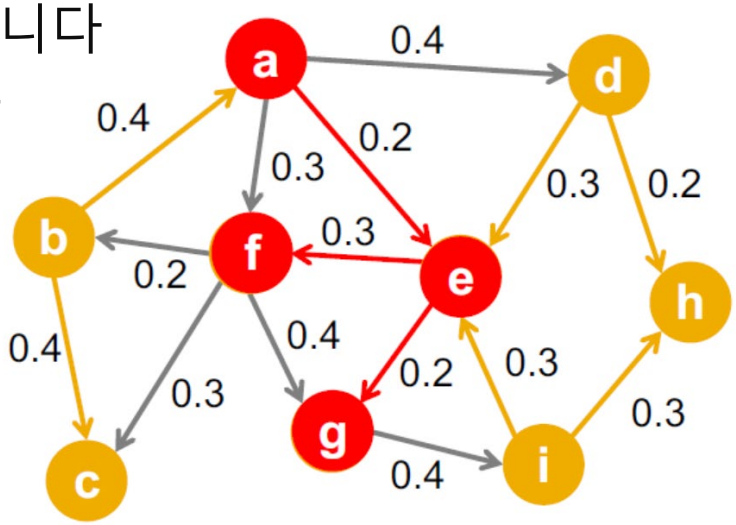
코로나의 전파는 확률적 과정이기 때문에 확률적 전파 모형을 고려해야합니다

본 수업에서는 가장 간단한 형태의 확률적 전파 모형인
독립 전파 모형(Independent Cascade Model)을 소개합니다

3.2 독립적 전파 모형

방향성이 있고 가중치가 있는 그래프를 가정합시다

각 간선 (u, v)의 가중치 p_{uv}는 u가 감염되었을 때
(그리고 v가 감염되지 않았을 때) u가 v를 감염시킬 확률에 해당합니다
즉, 각 정점 u가 감염될 때마다, 각 이웃 v는 p_{uv}확률로 전염됩니다

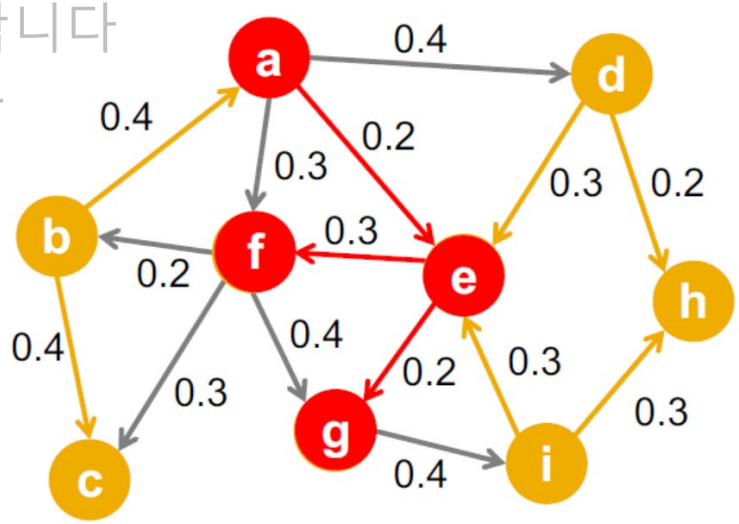


이미지 출처: https://web.stanford.edu/class/cs224w/slides/12-cascades.pdf

3.2 독립적 전파 모형

방향성이 있고 가중치가 있는 그래프를 가정합시다

각 간선 (u, v)의 가중치 p_{uv}는 u가 감염되었을 때
(그리고 v가 감염되지 않았을 때) u가 v를 감염시킬 확률에 해당합니다
즉, 각 정점 u가 감염될 때마다, 각 이웃 v는 p_{uv}확률로 전염됩니다



서로 다른 이웃이 전염되는 확률은 독립적입니다
u가 감염되었을 때 u가 v를 감염시킬 확률은
u가 감염되었을 때 u가 w를 감염시킬 확률과 독립적입니다
u가 감염되었을 때 u가 v를 감염시킬 확률은
w가 감염되었을 때 w가 v를 감염시킬 확률과 독립적입니다

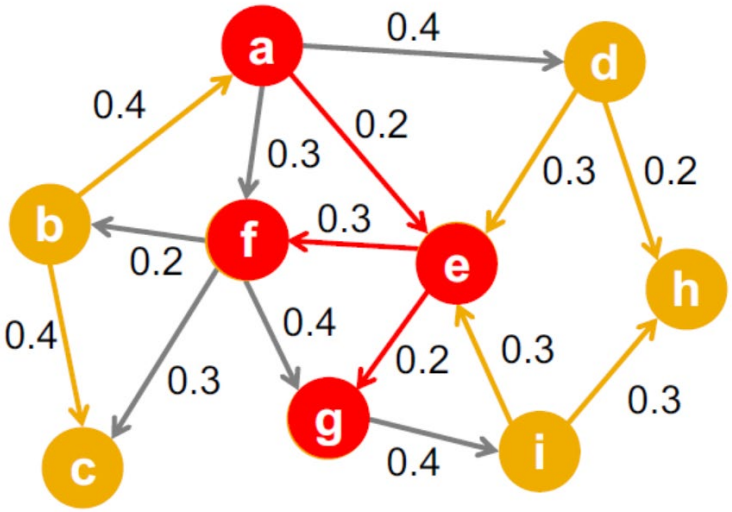
이미지 출처: https://web.stanford.edu/class/cs224w/slides/12-cascades.pdf

3.2 독립적 전파 모형

모형은 모델은 최초 감염자들로부터 시작합니다

이전 모형과 마찬가지로 첫 감염자들을 시드 집합(Seed Set)이라고 부릅니다

각 최초 감염자 u는, 각 이웃 v에게 p_{uv}확률로 병을 전파합니다
위 과정을 새로운 감염자 각각에게 반복합니다
위 과정을 새로운 감염자 각각에게 반복합니다
...
더 이상 새로운 감염자가 없으면 종료합니다



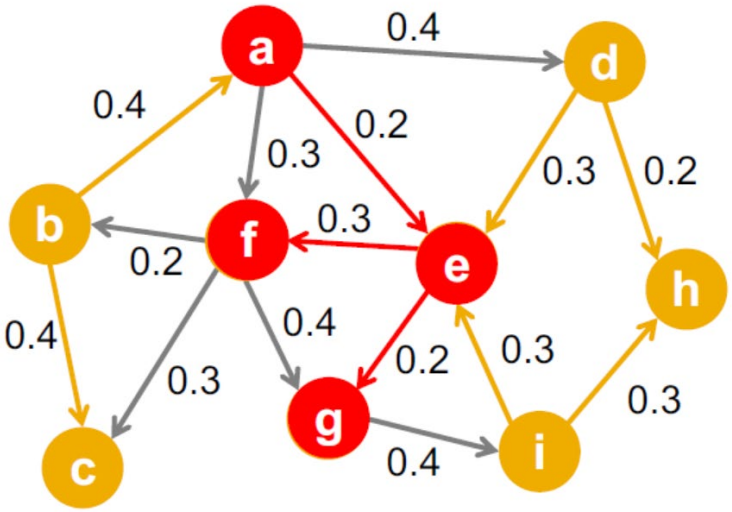
이미지 출처: https://web.stanford.edu/class/cs224w/slides/12-cascades.pdf

3.2 독립적 전파 모형

모형은 모델은 최초 감염자들로부터 시작합니다

이전 모형과 마찬가지로 첫 감염자들을 시드 집합(Seed Set)이라고 부릅니다

각 최초 감염자 u는, 각 이웃 v에게 p_{uv}확률로 병을 전파합니다
위 과정을 새로운 감염자 각각에게 반복합니다
위 과정을 새로운 감염자 각각에게 반복합니다
...
더 이상 새로운 감염자가 없으면 종료합니다



감염자는 계속 감염자 상태로 남아있는 것을 가정합니다
감염자의 회복을 가정하는 SIS, SIR 등의 다른 전파 모형도 있습니다

이미지 출처: https://web.stanford.edu/class/cs224w/slides/12-cascades.pdf

4. 바이럴 마케팅과 전파 최대화 문제

4.1 바이럴 마케팅이란?
4.2 시드 집합의 중요성
4.3 전파 최대화 문제
4.4 정점 중심성 휴리스틱
4.5 탐욕 알고리즘

4.1 바이럴 마케팅이란?

바이럴 마케팅은 소비자들로 하여금 상품에 대한 긍정적인 입소문을 내게 하는 기법입니다

바이럴 마케팅이 효과적이기 위해서는 소문의 시작점이 중요합니다
시작점이 어디인지에 따라서 입소문이 전파되는 범위가 영향을 받기 때문입니다
소셜 인플루언서(Social Influencer)들이 높은 광고비를 받는 이유입니다

이미지 출처: https://web.stanford.edu/class/cs224w/slides/12-cascades.pdf

4.2 시드 집합의 중요성

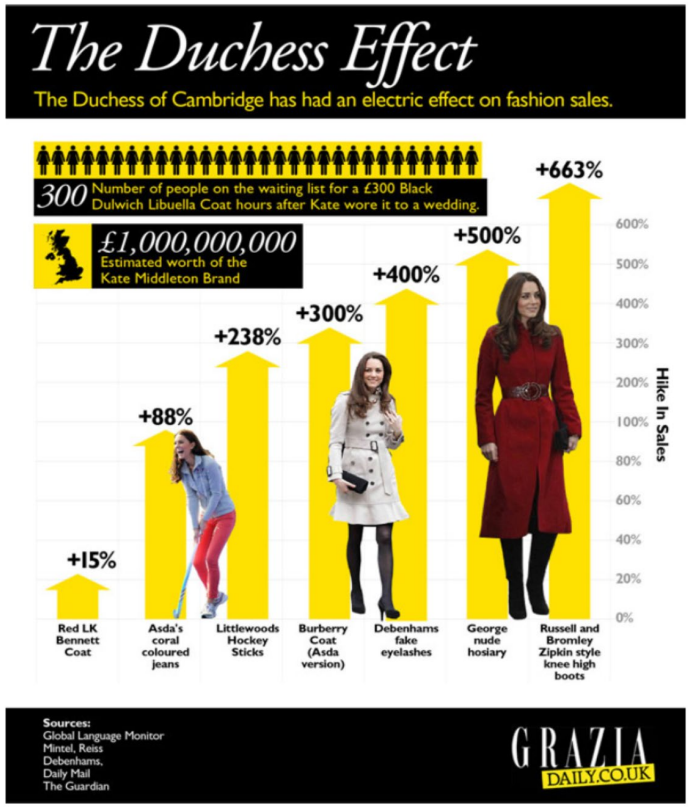
대표적인 소셜 인플루언서에는 영국 윌리엄 왕자의 부인 케이트 미들턴이 있습니다

'미들턴 효과'라는 말이 생겨날 정도입니다

"Kate Middleton effect

The trend effect that Kate, Duchess of Cambridge has on others, from cosmetic surgery for brides, to sales of coral-colored jeans."

WIKIPEDIA
The Free Encyclopedia

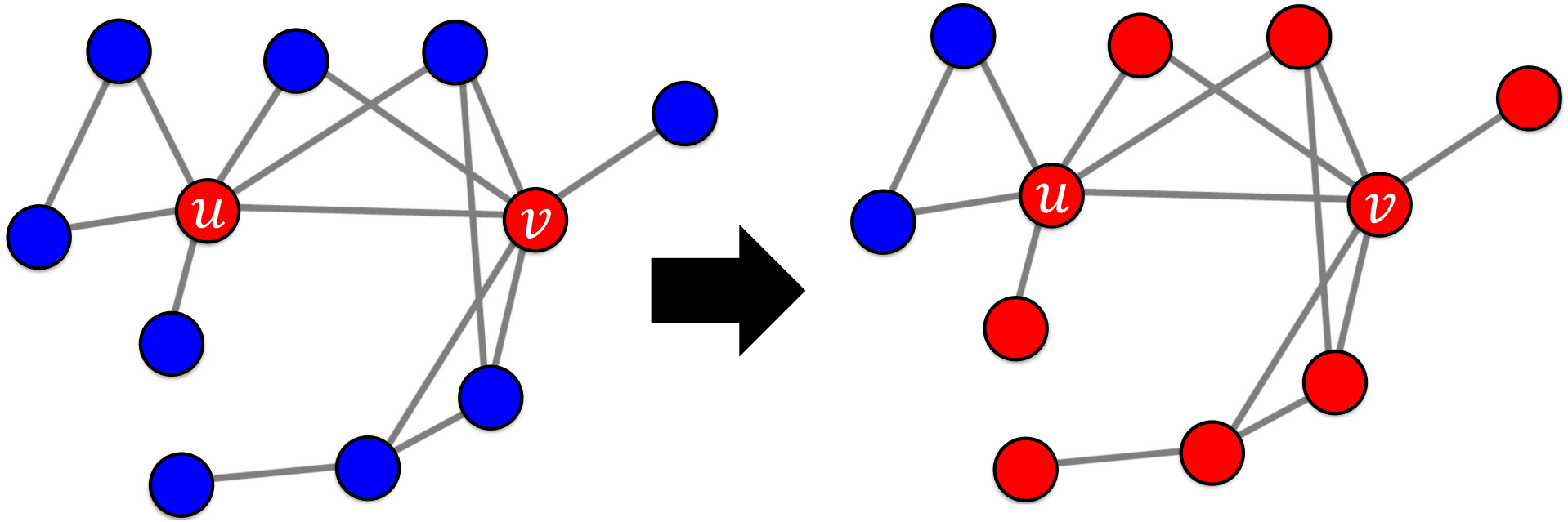


이미지 출처: https://web.stanford.edu/class/cs224w/slides/12-cascades.pdf

4.2 시드 집합의 중요성

앞서 소개한 전파 모형들에서도 시드 집합이 전파 크기에 많은 영향을 미칩니다

선형 임계치 모형의 예시에서 시드 집합으로 u와 v를 선택했을 때, 총 9명이 A를 선택했습니다

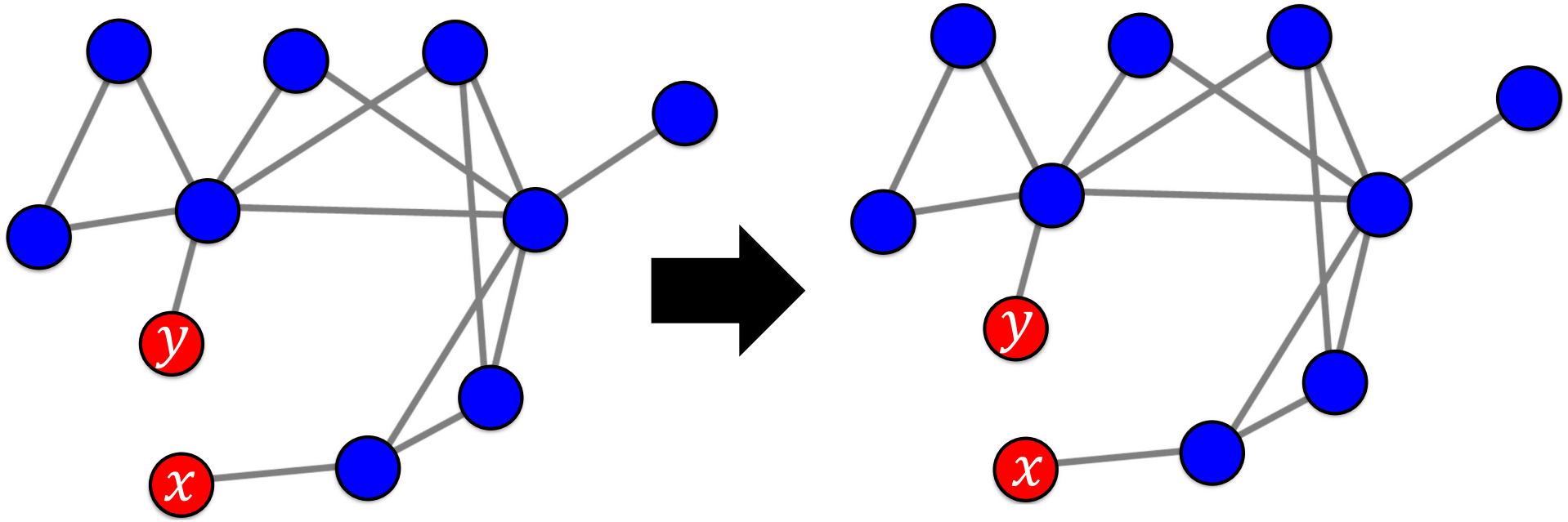


이미지 출처: https://web.stanford.edu/class/cs224w/slides/12-cascades.pdf

4.2 시드 집합의 중요성

앞서 소개한 전파 모형들에서도 시드 집합이 전파 크기에 많은 영향을 미칩니다

시드 집합으로 x와 y를 선택했다면, 추가 전파는 발생하지 않습니다. 2명만이 A를 선택했습니다



이미지 출처: https://web.stanford.edu/class/cs224w/slides/12-cascades.pdf

4.3 전파 최대화 문제

시드 집합을 우리가 선택할 수 있다면, 누구를 선택하시겠습니까?

그래프, 전파 모형, 그리고 시드 집합의 크기가 주어졌을 때
전파를 최대화하는 시드 집합을 찾는 문제를
전파 최대화(Influence Maximization) 문제라고 부릅니다

4.3 전파 최대화 문제

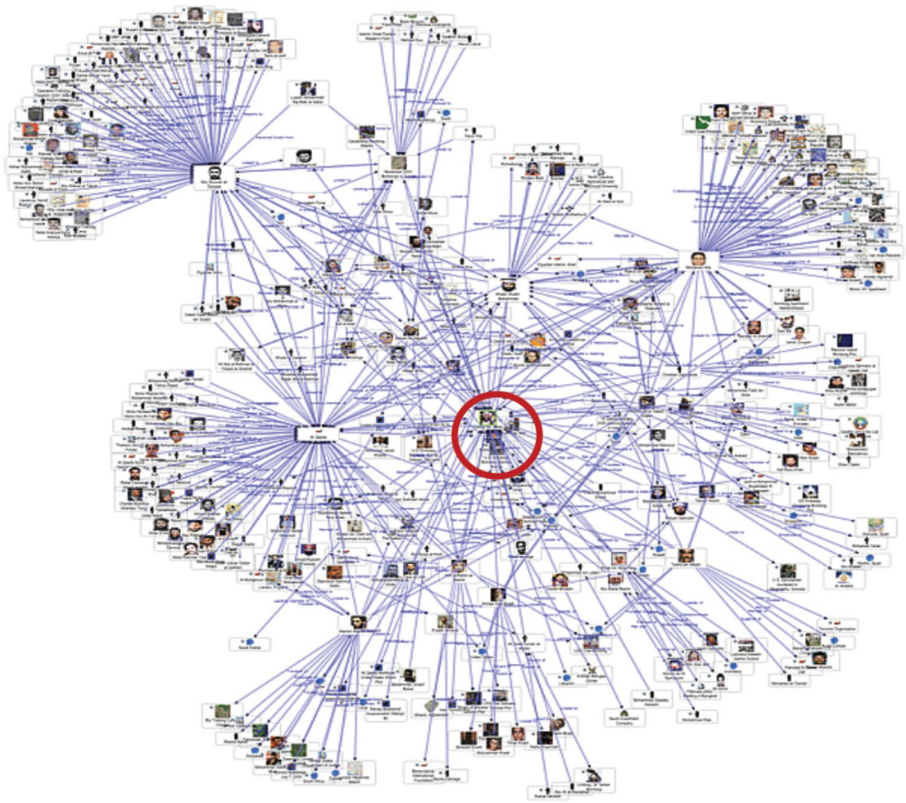
시드 집합을 우리가 선택할 수 있다면, 누구를 선택하시겠습니까?

그래프, 전파 모형, 그리고 시드 집합의 크기가 주어졌을 때
전파를 최대화하는 시드 집합을 찾는 문제를
전파 최대화(Influence Maximization) 문제라고 부릅니다

전파 모형으로는 앞서 배운 선형 임계치 모형, 독립 전파 모형을 포함
다양한 모형을 고려할 수 있습니다

4.3 전파 최대화 문제

전파 최대화 문제는 방대한 그래프, 즉 소셜 네트워크로부터,
‘케이트 미들턴’, 즉 영향력 있는 시드 집합을 찾아내는 문제입니다



이미지 출처: https://web.stanford.edu/class/cs224w/slides/12-cascades.pdf

4.3 전파 최대화 문제

전파 최대화 문제는 굉장히 어려운 문제입니다

그래프에 $|V|$개의 정점이 있을 경우, 시드 집합의 크기를 k개로 제한하더라도
경우의 수는 $\binom{|V|}{k}$개 입니다

정점이 10,000개, 시드 집합의 크기를 10으로 고정합시다
경우의 수는 무려 2,743,355,077,591,282,538,231,819,720,749,000개입니다

이론적으로 많은 전파 모형에 대하여
전파 최대화 문제는 NP-hard임이 증명 되었습니다

최고의 시드 집합을 찾는 것은 포기합시다

이미지 출처: https://web.stanford.edu/class/cs224w/slides/12-cascades.pdf

4.4 정점 중심성 휴리스틱

대표적 휴리스틱으로 정점의 중심성(Node Centrality)을 사용합니다

즉, 시드 집합의 크기가 k개로 고정되어 있을 때,
정점의 중심성이 높은 순으로 k개 정점 선택하는 방법입니다

정점의 중심성으로는 페이지랭크 점수, 연결 중심성, 근접 중심성, 매개 중심성 등이 있습니다

합리적인 방법이지만, 최고의 시드 집합을 찾는다는 보장은 없습니다

이미지 출처: https://web.stanford.edu/class/cs224w/slides/12-cascades.pdf

4.5 탐욕 알고리즘

탐욕 알고리즘(Greedy Algorithm) 역시 많이 사용됩니다

탐욕 알고리즘은 시드 집합의 원소, 즉 최초 전파자를 한번에 한 명씩 선택합니다
즉, 정점의 집합을 $\{1, 2, \dots, |V|\}$라고 할 경우 구체적인 단계는 다음과 같습니다

집합 $\{1\}, \{2\}, \dots, \{|V|\}$를 비교하여, 전파를 최대화하는 시드 집합을 찾습니다
이 때, 전파의 크기를 비교하기 위해 시뮬레이션을 반복하여 평균 값을 사용합니다
뽑힌 집합을 $\{x\}$ 라고 합시다

집합 $\{x, 1\}, \{x, 2\}, \dots, \{x, |V|\}$를 비교하여, 전파를 최대화하는 시드 집합을 찾습니다
뽑힌 집합을 $\{x, y\}$ 라고 합시다

4.5 탐욕 알고리즘

탐욕 알고리즘(Greedy Algorithm) 역시 많이 사용됩니다

집합 $\{x, y, 1\}, \{x, y, 2\}, \ldots, \{x, y, |V|\}$를 비교하여, 전파를 최대화하는 시드 집합을 찾습니다
뽑힌 집합을 $\{x, y, z\}$ 라고 합시다

위 과정을 목표하는 크기의 시드 집합에 도달할 때까지 반복합니다

즉, 탐욕 알고리즘은 최초 전파자 간의 조합의 효과를 고려하지 않고
근시안적으로 최초 전파자를 선택하는 과정을 반복합니다

4.5 탐욕 알고리즘

탐욕 알고리즘은 얼마나 정확한가요?

독립 전파 모형에 경우, 이론적으로 정확도가 일부 보장됩니다
항상, 즉 입력 그래프와 무관하게 다음 부등식이 성립합니다

탐욕 알고리즘으로 찾은 시드 집합의 의한 전파의 (평균) 크기

$$\geq \left(1 - \frac{1}{e}\right) * \text{최고의 시드 집합에 의한 전파의 (평균) 크기}$$

$$\approx 0.632 * \text{최고의 시드 집합에 의한 전파의 (평균) 크기}$$

다시 말해, 탐욕 알고리즘의 최저 성능은 수학적으로 보장되어 있습니다

5. 실습: 전파 모형 시뮬레이터 구현

5.1 독립적 전파 모형 시뮬레이터

5.2 선형 임계치 모형 시뮬레이터

5.1 독립적 전파 모형 시뮬레이터

실습에 사용할 방향성이 있고 가중치가 있는 그래프를 파일에서 읽어 불러옵니다

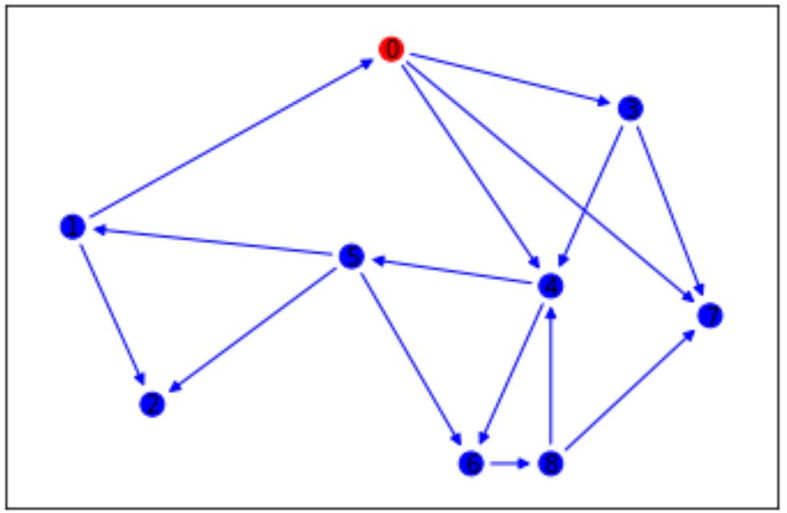
```
G = nx.DiGraph()
data = osp.abspath(osp.join(os.getcwd(), 'drive/MyDrive/data/simple/simple_weighted_directed_graph.txt'))
f = open(data)
for line in f:
    line_split = line.split()
    src = int(line_split[0])
    dst = int(line_split[1])
    w = float(line_split[2])
    G.add_edge(src, dst, weight=w)
```

5.1 독립적 전파 모형 시뮬레이터

시드 집합, 즉 최초 전염자를 지정합니다

0번 정점을 최초 전염자로 지정하였습니다
감염자는 붉게 표시됩니다

```
affected = set()
affected_new = set({0})
used_edge = set()
draw(G, affected | affected_new, used_edge)
```



5.1 독립적 전파 모형 시뮬레이터

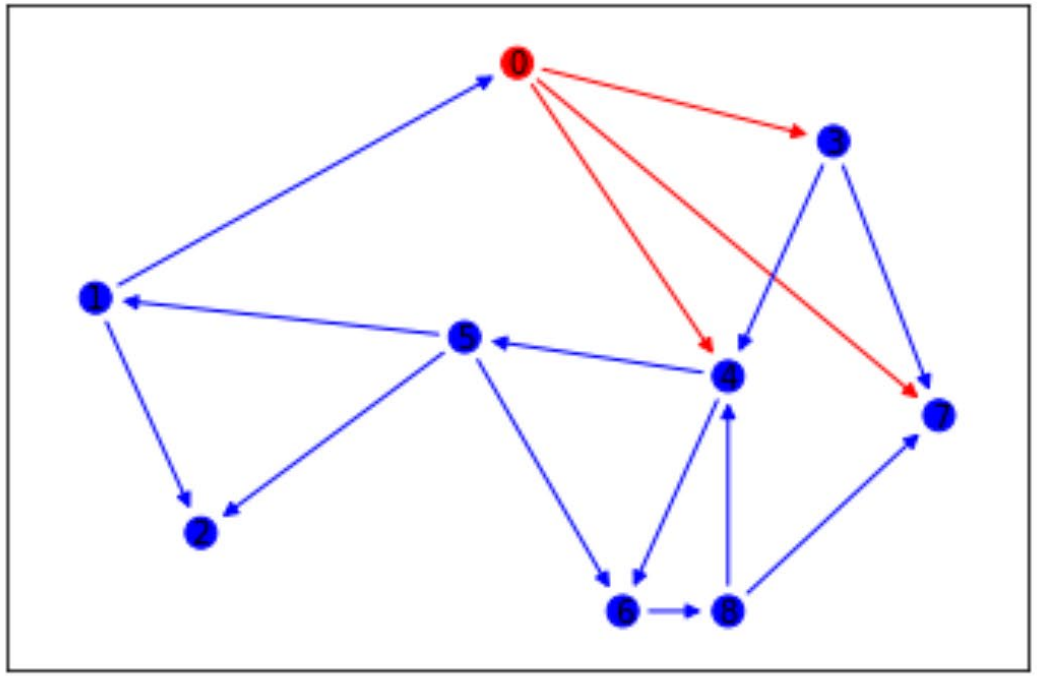
독립적 전파 모형을 시뮬레이션 합니다

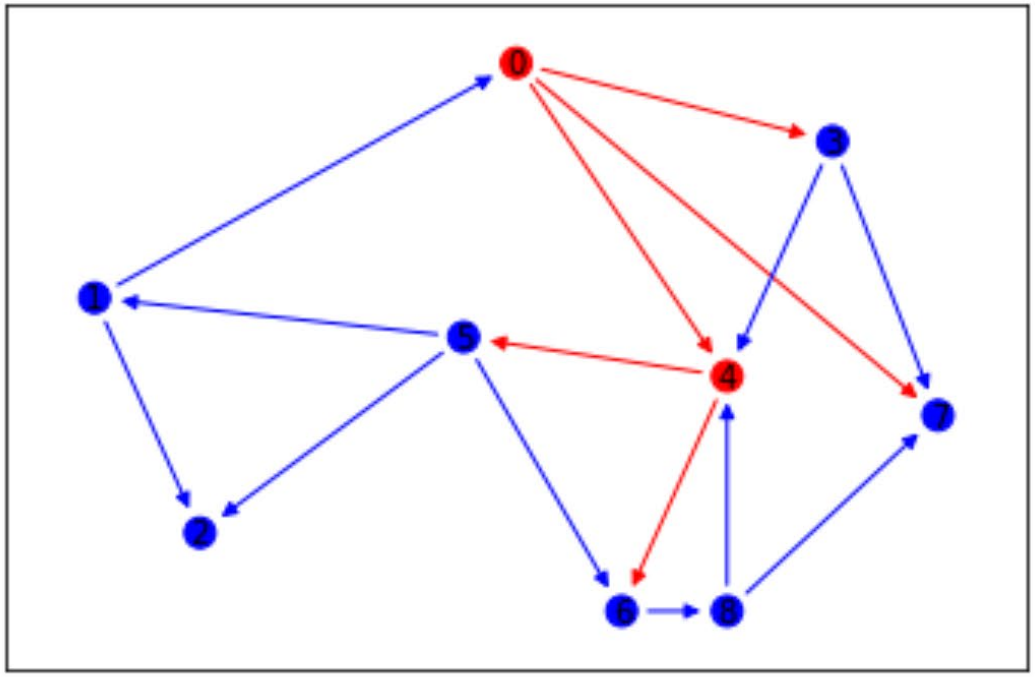
```
while len(affected_new) != 0:
    temp = set()
    for src in affected_new:
        neighbors = G.neighbors(src)
        for dst in neighbors:
            if (dst not in affected) and (dst not in affected_new):
                p = random.uniform(0, 1)
                if p < G.edges[src, dst]["weight"]:
                    temp.add(dst)
                used_edge.add((src, dst))
    affected = affected | affected_new
    affected_new = temp
    draw(G, affected, used_edge)
```

5.1 독립적 전파 모형 시뮬레이터

전파 과정이 그림으로 출력됩니다

전염에 사용된 간선은 붉게 표시됩니다





5.2 선형 임계치 모형 시뮬레이터

실습에 사용할 방향성이 없고 가중치도 없는 그래프를 읽어 불러옵니다

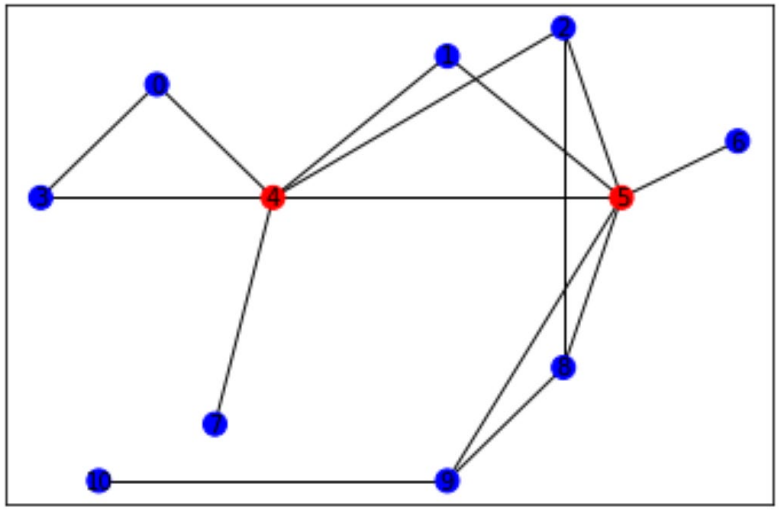
```
[ ]  G = nx.Graph()
     data = osp.abspath(osp.join(os.getcwd(), 'drive/MyDrive/data/simple/simple_undirected_graph.txt'))
     f = open(data)
     for line in f:
         line_split = line.split()
         src = int(line_split[0])
         dst = int(line_split[1])
         G.add_edge(src, dst)
```

5.2 선형 임계치 모형 시뮬레이터

시드 집합, 즉 얼리 어답터를 지정합니다

4번, 5번 정점을 얼리 어답터로 지정하였습니다
A를 선택한 정점들을 빨간색으로 표시됩니다

```
team_A = set({4, 5})
team_B = set([v for v in G.nodes if v not in team_A])
draw(G, team_A)
```



5.2 선형 임계치 모형 시뮬레이터

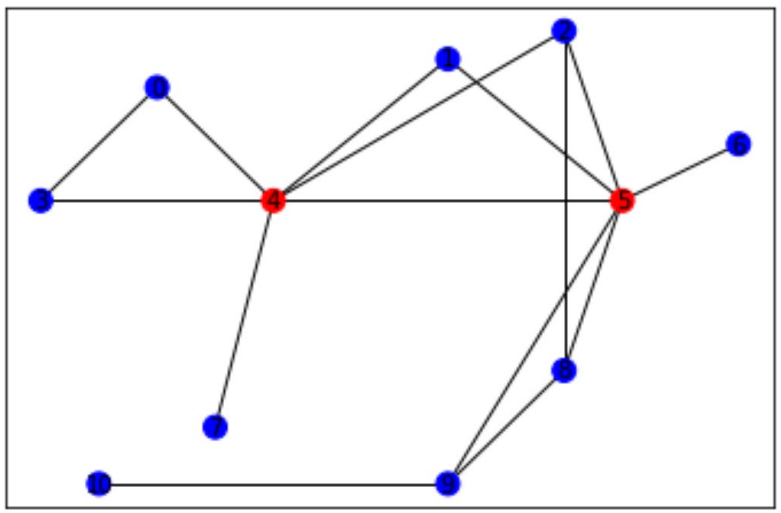
선형 임계치 모형 시뮬레이션 합니다

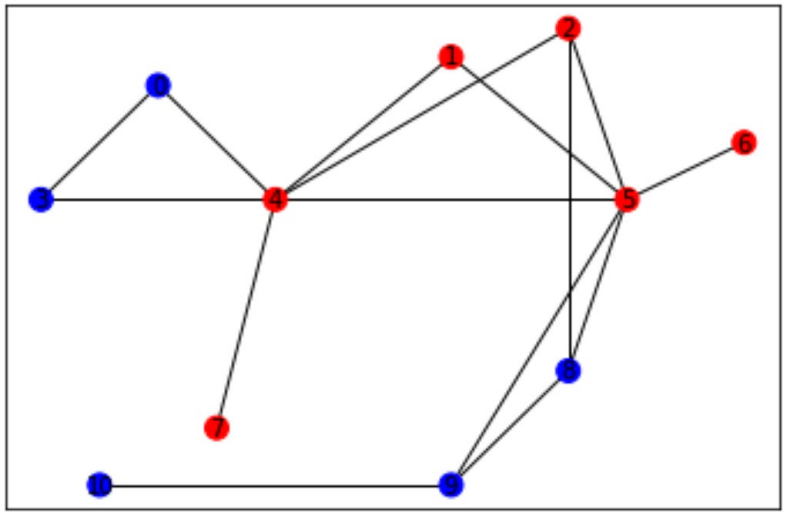
```
threshold = 0.5
while True:
    new_A = set()
    for v in team_B:
        neighbors = list(G.neighbors(v))
        neighbors_A = [v2 for v2 in neighbors if v2 in team_A]
        if len(neighbors_A) / len(neighbors) > threshold:
            new_A.add(v)
    if len(new_A) == 0:
        break
    team_A = team_A | new_A
    team_B = team_B - new_A
    draw(G, team_A)
```

5.2 선형 임계치 모형 시뮬레이터

전파 과정이 그림으로 출력됩니다

A를 선택한 정점들을 빨간색으로 표시됩니다





4강 정리

1. 그래프를 통한 전파의 예시

- 정보, 행동, 고장, 질병 등

2. 의사결정 기반의 전파 모형

- 각각의 정점이 개인의 행복을 최대화하도록 의사결정하는 상황을 모형화
- 대표 예시: 선형 임계치 모형

3. 확률적 전파 모형

- 질병의 전파 등 확률적 과정을 모형화
- 대표 예시: 독립적 전파 모형

4. 바이럴 마케팅과 전파 최대화 문제

- 전파를 최대화하는 시드 집합을 찾는 문제
- 정점 중심성 휴리스틱, 탐욕 알고리즘 등

5. 실습: 전파 모형 시뮬레이터