Graph Structure in the Web

* 원 자료 : http://www.cis.upenn.edu/~mkearns/teaching/NetworkedLife/broder.pdf

- 2억개의 웹페이지와, 15억개의 링크를 분석함.

연구목적 :

- 더 낳은 웹크로울 전략 개발
- 웹에서의 컨텐트 생성 사회학
- 링크 관계를 이용하는 웹 알고리즘 이해
- 웹 구조의 진화 이해

주요 결과 :

1. The power law for in-degree : the probability that a node (page) has in-degree (incoming hyperlink, pointers) i is proportional to $1/i^x$ for some x>1 (2.1)

Probability [in-degree = i] = $k \cdot 1/i^{2.1}$ (k = positive number)

즉, Pr[1] = 1 [assuming k=1, x=2]

Pr[2] = 1/4

Pr[3] = 1/9

Pr[4] = 1/16

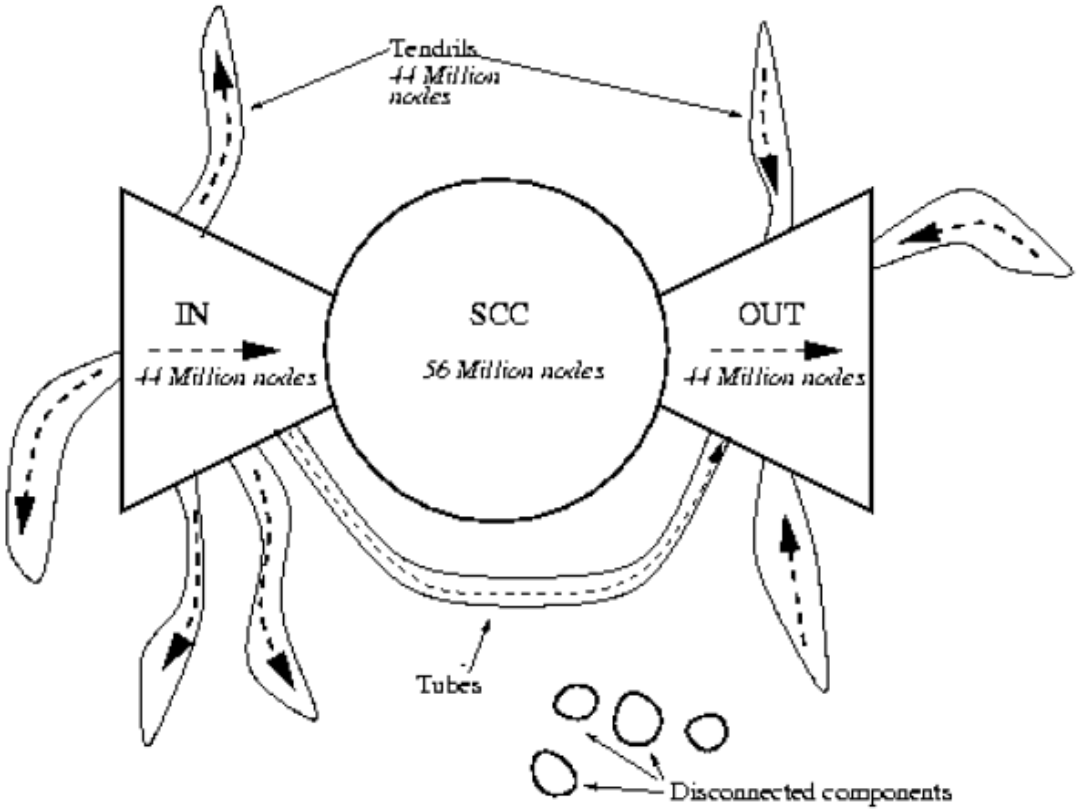
Pr[5] = 1/25

Pr[100] = 1/10000

Pr[10000] = 1/100000000

* long tail 을 갖는다

2.



- 2억개의 노드들의 약 90%가 하나의 connected component에 속한다 (위의 이상한 오징어 나비 모양). 단, 이때 링크들을 undirected edge로 간주한다. 실제로, hyperlink는 directed edge임을 상기. 나머지 10% 미만들의 노드들이 아래의 몇개의 disconnected component로 분류된다.

- 위 오징어/나비 모양의 connected component를 조금 더 자세히 분석해 본다; 가운데 SCC (strongly connected component) 가 먼저 눈에 띈다. 여기에 약 5600 만개의 노드가 뭉쳐있는데, 여기에 속한 노드들은 서로 다른 노드로 가는 directed edge를 통한 path가 존재한다. 즉, 여기에 있는 페이지들은 hyperlink를 통해 그 어느 페이지에게라고 갈 수 있어 서로서로 다 통하는 길 (directed edge)이 있다.

IN ; 4400만개로 구성되었는데, 여기에 있는 페이지들은 path를 통해 SCC에 갈 수 있지만, SCC에 있는 페이지들이 IN 페이지로 가지는 못한다. 즉, IN에 있는 페이지

는 SCC에 있는 모든 페이지들로 갈 수 있다.

OUT : 4400만개의 노드들이 여기 속하는데, SCC의 페이지들이 이 쪽에 속한 페이지로 갈 수 있지만, OUT에 속한 페이지들이 SCC내의 페이지들로 갈 수 있는 path가 없다.

- Tendrils : SCC로 연결되는 길 (가거나 오거나) 이 없는 페이지들로 구성.

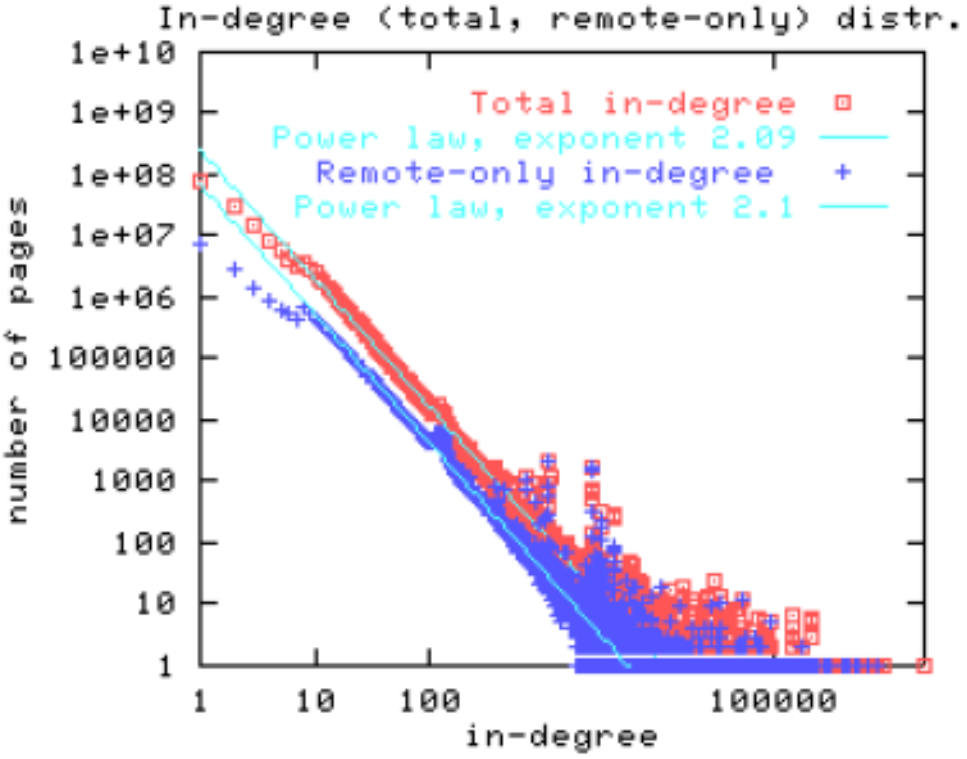
- SCC의 지름 (노드간 평균 거리) 은 최소 28

- 어떤 2개의 페이지를 임의의 선정해서, 2개 사이를 잇는 path가 존재할 확률은 24%

- 만약 2 페이지간 directed path가 존재한다면, 그 평균거리는 16

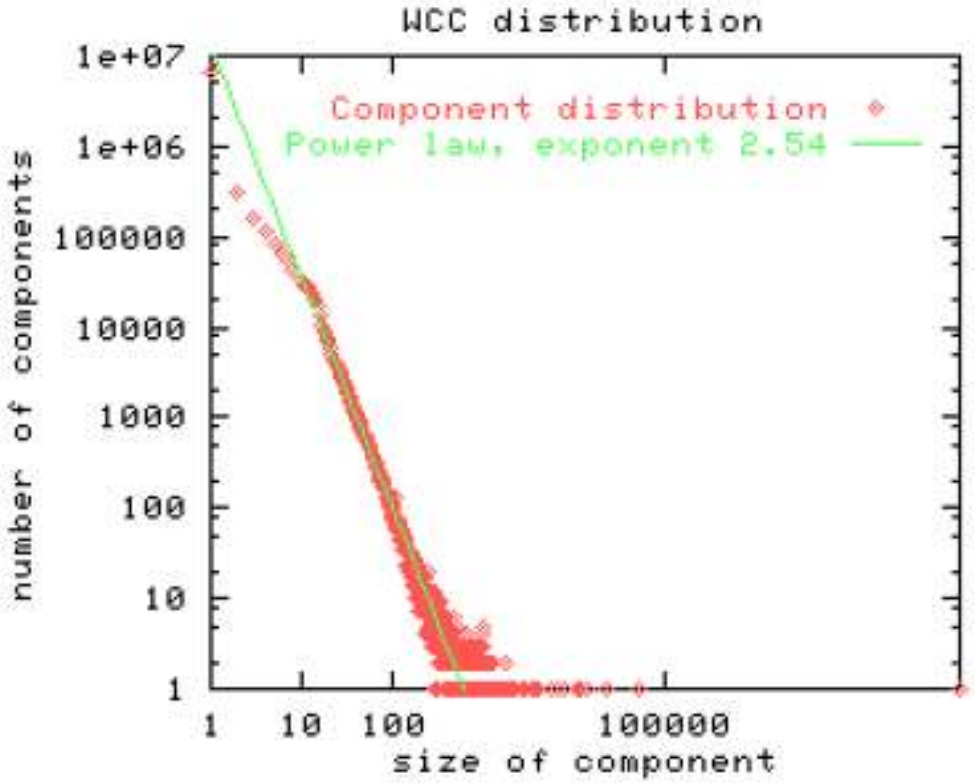
■ **실험 결과**

1. In-degree power law



- 페이지의 in-degree가 10배 늘면, 그런 페이지 개수가 원래보다 약 100배 이상 준다.

2. Connected Components



- 웹그래프를 undirected graph로 해석해서, hyperlink의 방향성에 무관하게 connected component 크기를 따져보면, 약 91%의 페이지 (1억 8천6백/2억) 들이 하나의 큰 component에 속해 있음을 알 수 있다. 그 다음 큰 규모의 component는 15만개의 페이지로 구성

- component 규모에 따른 분포도 power law를 유사하게 나타낸다. 즉, component 크기가 10배 늘면, 10배 크기의 component 개수는 원래보다 $1/10^{2.54}$가 된다.

k	1000	100	10	5	4	3
Size (millions)	177	167	105	59	41	15

Table 1: Size of the largest surviving weak component when links to pages with in-degree at least k are removed from the graph.

- SCC에서 BFS를 통해 얼마만한 깊이로 모든 페이지로 가는 path를 찾을 수 있는가 in-link를 따라가면서 보았다. 어떤 노드에서 시작한 가에 따라 최소 475 step, 최대 503 step이 걸림.

- Diameter and average connected Distance

Edge type	In-links (directed)	Out-links (directed)	Undirected
Average connected distance	16.12	16.18	6.83

* 핵심 내용
- 웹은 앞의 오징어 나비 그림과 같이 부분으로 나뉘어져 있다. 약 1/4은 url간 서로 연결되는 SCC에 속하고, 1/4이 약간 안되는 것이 IN에 그리고 또 1/4이 약간 안되는 것이 OUT 그룹에 속해 있다.
- 이렇게 웹이 나뉘는 이유는 웹의 edge가 hyperlink라 방향성이 있기 때문이다. 따라서, 웹이 계속 커지더라도 (url이 늘더라도) 계속 이런 모양일 것이라 생각된다.
- 이런 방향성에 따른 나누어짐 때문에 웹에서 링크를 따라 모든 웹페이지를 방문하는 것은 불가능 하다. 그나마 가장 많이 방문할 가능성이 있는 초기 페이지는 어떤 IN의 한 url에서 시작해서, 따라갈 수 있는 IN의 모든 url 찾고, SCC에 있는 모든 노드들은 방문가능하고, 역시 OUT에 있는 모든 url들 방문가능하다. 그러니, 최소한 SCC와 OUT 에 있는 모든 노드 방문가능하고, 일부 IN 노드 방문가능하다.
- IN에는 아마도 새로 추가된 웹페이지, 인기가 없는 웹페이지, 그러니까 대부분의 개인페이지, 또는 블로그가 여기에 속할 지도 모르겠다. OUT에는 회사 웹페이지들이 많이 있을 것이고. 그런데, tendril에 있는 녀석들은 어떤 특징이 있을까?
-url 간에 path, 즉 링크를 따라 길이 있다면 2억 개 되는 웹에서는 평균거리가 16이다. 생각보다 가까운가, 먼가?
- url로 들어오는 링크들의 in-degree가 power law를 따른다. 또, 컴포넌트 내부

의 노드 수 등 많은 웹의 여러 특성들이 power law를 따른다.
- Power law는 자연스런 현상이 아니다. 즉, 웹은 자연스럽게 성장/발전하는 것이 아니라 기저에 어떤 규칙/법칙/의도 에 따라 된 것이다.
- 컴포넌트 수와 컴포넌트내의 url 개수가 power law를 보인다는 것은, 가장 큰 컴포넌트가 SCC이니, 이 녀석이 위의 그림에서 가장 오른편에 있는 녀석일 것이고, 이 보다는 작은 녀석들이 있을 것이고, 1개 보다는 많을 것이라는 얘기이다. 그런데, 그림을 보면 컴포넌트 크기가 어느 이상 되는 녀석들은 각각 몇 개 되지 않는 것 같다. 엄밀히 power law는 따르지 않는다. 컴포넌트 크기가 작을수록 그런 컴포넌트들은 power law에 따라 증가한다. 그러면, 이런 컴포넌트들은 어디에 있을까? SCC는 아니고. 물론, IN과 OUT 그리고 tendtril 에 있겠지.
- 우리나라 웹지도를 생각해 보자.