〈링크 – A.-L. 바라바시〉(2002) 정리

– 조항현 h2jo23@gmail.com

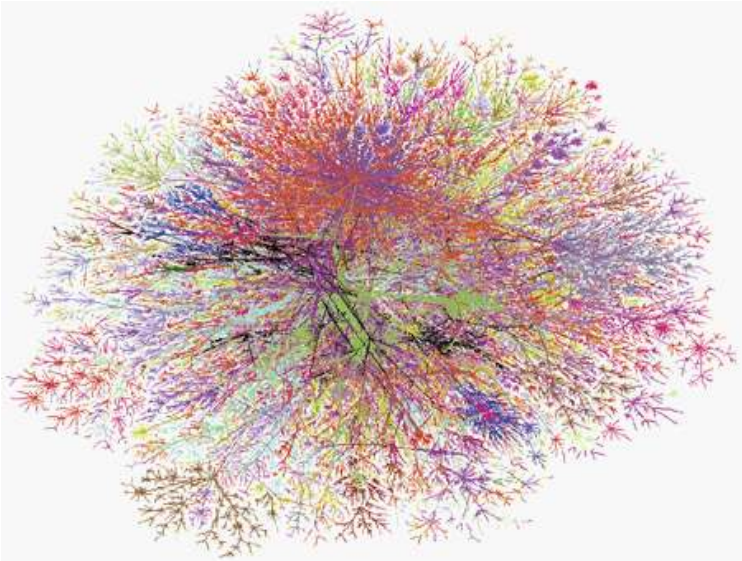

그림 1 인터넷 [http://www.research.ibm.com/nips03workshop/]

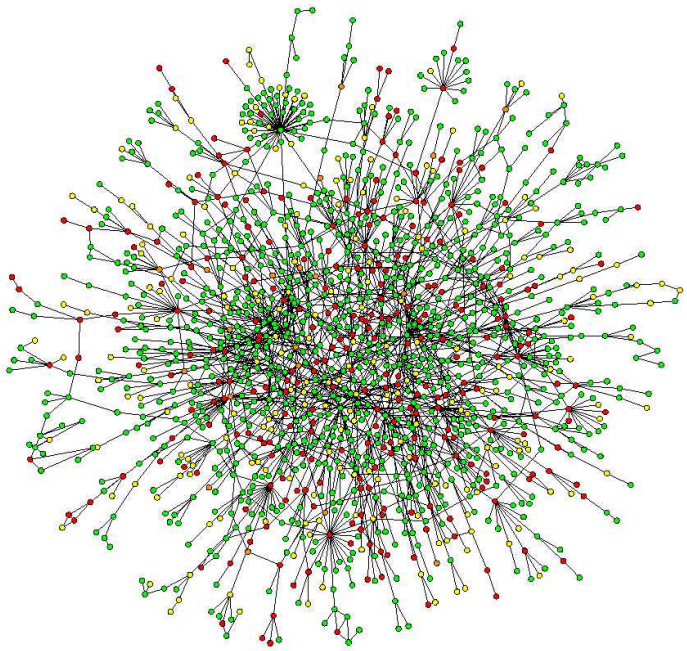

그림 2 단백질 상호작용 네트워크 [출처: H. Jeong *et al.*, Nature **411**, 41 (2001)]

1. 서론

최근에도 뉴스에 나온 디도스(DDoS; distributed denial of service) 공격에 관한 얘기가 책의 첫머리를 장식하고 있습니다. 2000년 2월 7일 야후 사이트를 디도스 공격한 15세 소년(일명 마피아보이)에 관한 얘기입니다. 그 소년은 어떻게 그런 큰 일을 벌일 수 있었을까? 간단히 말해서 네크워크의 힘이라는 것이죠. 또 다른 예로 바울의 기독교 전파에 관한 이야기. 바울이 사회적 네트워크를 효과적으로 사용할 줄 알았기에 가능한 일이었다는 얘기를 합니다.

2. 무작위의 세계

레온하르트 오일러가 쾨니히스베르크의 다리 문제를 해결하면서 그래프 이론이 시작됩니다. 20세기에는 폴 에르되스와 알프레드 레니의 무작위 그래프에 관한 연구가 큰 자리를 차지하고 있습니다. 그들은 복잡한 그래프가 지닌 "다양성을 의도적으로 논외로 하고 자연이 따를 수 있는 가장 단순한 해결책"으로 '무작위 연결'을 제안했다고 합니다.
노드들이 주어져 있으면 아무거나 두 개를 골라서 일정한 확률로 연결을 시키는 거죠. 연결시킬 확률이 낮으면 연결된 노드들, 즉 덩어리(cluster; 책에서는 '클러스터'로 표기)가 별로 없거나 있더라도 그 덩어리 크기(연결된 노드의 개수)가 작습니다. 그 확률이 높다면 거의 대부분의 노드가 커다란 덩어리에 포함되겠죠. 아주 소수의 노드만이 따로 작은 덩어리를 형성하고 있을 겁니다. 그 확률이 너무 작지도 너무 크지도 않다면, 다시 말해서 거의 대부분의 노드를 포함하는 커다란 덩어리가 나타나기 시작하는 순간의 확률은? 이게 네트워크에서 스미기(percolation) 문제입니다.

무작위 연결 이론에 의해 사람들은 "복잡성과 무작위성을 동일시"하는 영향을 받았다고 합니다. 물론 그렇지 않다는 게 네트워크 이론뿐 아니라 다른 분야에서도 제기되었고 또한 연구되고 있습니다.
하나만 더 밑줄을 긋고 가겠습니다. "그들(파티에 초대된 사람들)은 이내 서로 이야기하기 시작할텐데, 이는 만나서 서로 알고자 하는 인간의 태생적인 욕구 때문이다."

3. 여섯 단계의 분리

헝가리 문학계의 천재로 여겨진 카린시가 1929년에 단편소설집을 냈는데, 그 중 〈연쇄〉라는 글에는 "지구상의 15억 주민들 중 아무나 한 사람의 이름을 뽑았을 때, 다섯 명 이하의 지인의 연쇄적인 친분관계를 통해 자신이 그에게 연결할 수 있다고 장담했다"는 구절이 있다고 합니다. 이게 바로 "여섯 단계의 분리(six degrees of separation)"로 우리에게 잘 알려진 개념을 처음 공식적으로 출판한 것이라네요.
1967년 스탠리 밀그램 교수는 미국에서 실제 실험을 통해 이를 증명해보이죠. 우리는 이런 "좁은 세상(small world)"에서 살고 있습니다. 이를 수학적으로 표현하면, N개의 노드로 이루어진 네트워크에서 임의의 두 노드 사이의 평균 거리가 log N에 비례한다고 합니다.

예를 들어 모든 사람이 각각 보통 100명의 친구(직장 및 가족 포함)를 갖고 있다고 하면 친구의 친구는 대략 10,000명이고 친구의 친구의 친구는 1,000,000명입니다. 이렇게 세 단계만 거쳐도 백만 명과 연결되며, 다섯 단계를 거치면 100억 명, 즉 지구상의 모든 사람과 연결될 수 있습니다. 한 사람당 친구의 수를 k, 단계 수를 l, 전체 인구를 N이라고 하고 이를 수식으로 나타내면 다음과 같습니다.

$$k^l = N \rightarrow l = \log_k N \propto \log N$$

이건 인터넷 네트워크에서도 나타나는 현상인데요, 당시 연구 결과 19단계만 거치면 수억 개의 웹페이지들이 연결된다고 합니다. 하지만 우리는 실제로 더 빨리 원하는 문서를 찾곤 합니다. 모든 링크를 따라가기보다는 단서들을 적절하게 '해석'하기 때문이지요. 여기서도 밑줄 하나 그으면, "끊임없이 서로 접촉하고자 하는 사람들의 추구"에 의해 좁은 세상이 되었다고 합니다.

4. 좁은 세상

1960년대 후반에는 하버드 대학의 사회학에서 네트워크에 대한 관심이 일기 시작했다네요. 그라노베터는 "약한 연결의 힘(The strength of weak ties)"이라는 논문에서 노동시장에서 직업을 구하는데 친구가 아니라 그냥 아는 사람의 도움이 크다는 결과를 발표합니다. "약한 연결들은 외부 세계와 의사소통을 하려고 할 때 결정적인 역할을 한다. (중략) 그들(가장 친한 친구들)은 나와 같은 서클에 있으므로 대개 동일한 정보를 갖고 있을 경우가 많기 때문이다. 새로운 정보를 얻고 싶으면 우리는 약한 연결을 사용해야 한다."라고 합니다.

그의 주장은 무작위 네트워크와는 달리 완전연결 그래프(모든 노드가 서로 연결된 덩어리)들이 약하게 연결되어 전체를 이루는 것으로 봅니다. 던컨 와츠와 스티븐 스트로가츠는 덩어리 계수(clustering coefficient)라는 양을 제시하고 이를 여러 실제 네트워크에 대해서 측정함으로써 에르되스-레니의 무작위 네트워크가 실제 네트워크의 덩어리 성질을 반영하지 못한다는 것을 보입니다. 이들은 1998년에 〈네이처〉에 좁은 세상 효과와 덩어리 구조를 모두 갖는 모형을 제시하여 네트워크 연구를 촉발시키게 되죠.

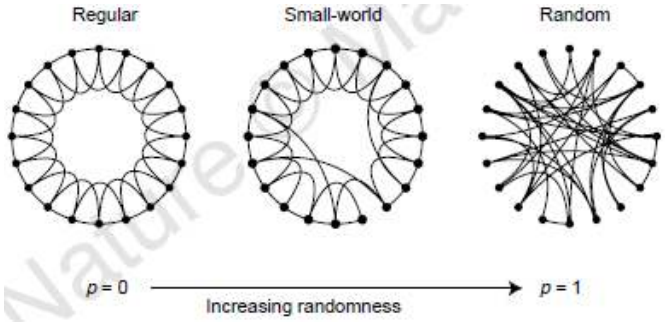

그림 3 좁은 세상 네트워크 [출처: D. Watts, S. Strogatz, Nature **393**, 440 (1998)]

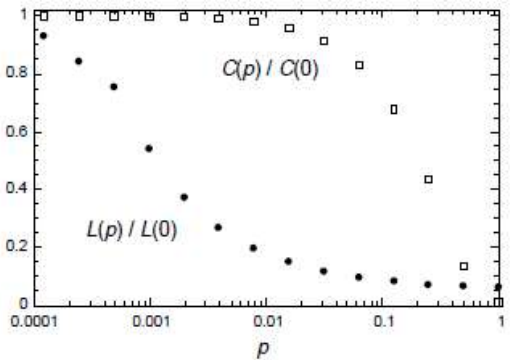

그림 4 좁은 세상 네트워크의 결과

여기서도 밑줄 하나. "사람들은 친근함, 안전, 익숙함 등을 주는 파벌(clique)과 클러스터(cluster)를 형성하고자 하는 태생적인 욕구를 갖고 있다."

5. 허브와 커넥터

"우리의 웹 지도 만들기 프로젝트의 결과 중 가장 흥미로운 점은 웹에는 민주주의, 공정성, 평등성이 완벽하게 존재하지 않는다는 것이다. 웹의 위상구조는 저기에 널려 있는 수십억의 문서들을 모두 똑같이 볼 수만은 없게 만든다."

저자는 웹의 구조를 연구하면서 소수의 웹페이지(허브/커넥터)만이 매우 많은 들어오는 링크를 가지며, 대부분의 웹페이지는 매우 적은 수의 들어오는 링크를 갖는다는 사실을 확인합니다. 헐리우드 배우 네트워크에서도 허브들이 존재하며, 이들로 인해 배우들 사이의 거리가 매우 짧아집니다. 또한 많은 영화에 출연한 배우일수록 다른 배우들과의 평균거리가 짧아집니다. 그런데 그게 항상 옳은 것은 아닌데, 포르노 배우의 경우 엄청나게 많은 영화에 출연해도 네트워크 전체에서는 중요한 위치를 차지할 수 없습니다. "네트워크 전체에서 진정으로 중심적인 위치를 차지하는 것은 여러 개의 큰 클러스터들에 동시에 속해 있는 노드들이다."

이런 허브의 존재는 에르되스-레니의 무작위 네트워크뿐 아니라 와츠-스트로가츠의 좁은 세상 네트워크에도 발견될 수 없는 현상이어서, 새로운 모형을 요구하게 됩니다. 그래서 바라바시 교수가 척도 없는 네트워크 모형을 제시합니다.

6. 80/20 법칙

드디어 거듭제곱 법칙(power law)이 나오네요. 관련하여 파레토가 제시한 80/20 법칙도 지금은 너무나 잘 알려져 있죠. (그가 "80/20"이라는 표현을 사용한 적은 없다네요.) 다양한 실제 네트워크에서 각 노드가 갖고 있는 링크의 개수, 즉 이웃수의 분포를 그리면 거듭제곱 꼴이 나옵니다. 무작위 네트워크나 좁은 세상 네트워크에서는 대부분의 노드가 비슷한 이웃수를 갖습니다. 그 네트워크의 이웃수 분포에 특정한 규모, 즉 잘 정의된 평균값이 있다고 말합니다.

하지만 대부분의 노드는 이웃수가 매우 작고, 소수의 노드는 엄청나게 큰 이웃수를 갖는 네트워크에서는 대개 그런 '평균' 이웃수는 별 의미가 없습니다. 예를 들면, 국민소득 1만 달러라고 해도 여전히 인구의 대부분이 가난에 허덕이는 경우가 있겠죠. 평균보다는 편차가 현상을 이해하는데 더 효과적인 경우입니다. 거듭제곱 분포에는 특정한 규모/척도(scale)가 없으므로, 이웃수가 거듭제곱 분포를 보이는 네트워크를 척도 없는 네크워크라고 부릅니다. 수식으로 쓰면 다음과 같습니다.

$P(k) \sim k^{-\gamma}, 2 < \gamma < 3$

거듭제곱 법칙은 질서 상태와 무질서 상태 사이의 상전이에서 일반적으로 발견되는데, 그래서 특히 통계물리학자들이 지대한 관심을 갖고 연구해왔습니다. 그래서 거듭제곱 꼴의 이웃수 분포는 뭔가 재미있는 일이 벌어지고 있다는 신호로 보이는 거죠.

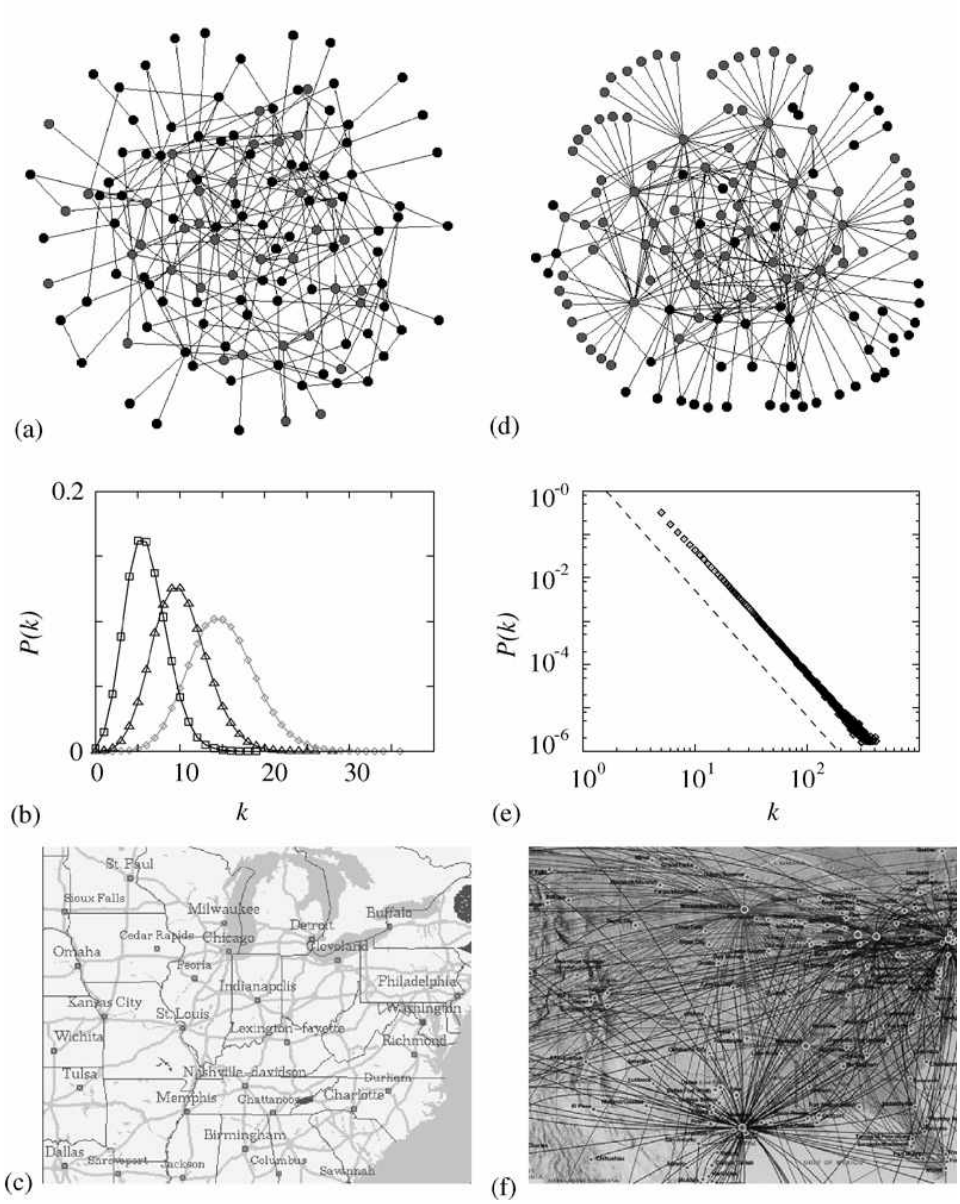

그림 5 랜덤 네트워크(왼쪽)와 척도없는 네트워크(오른쪽)의 비교 [출처: H. Jeong, Physica A **321**, 226 (2003)]

7. 부익부 빈익빈

이런 실제 척도 없는 네트워크를 위한 가장 단순한 모형을 제시하기 위해 바라바시는 성장과 선호적 연결을 도입합니다. 실제 네트워크에서 노드는 끊임없이 생성됩니다. (물론 사라지기도 하죠.) 그리고 새 노드는 기존 노드에 무작위로 연결되지 않고 이웃수가 많은 노드에 연결되려는 선호(즉 부익부)를 갖습니다. 이 두 요소를 도입함으로써 척도 없는 네트워크 모형이 탄생됩니다. 처음부터 노드 개수가 고정되는 모형들과 달리 성장 모형은 네트워크의 '진화'에 대해서도 통찰할 수 있게 해줍니다.

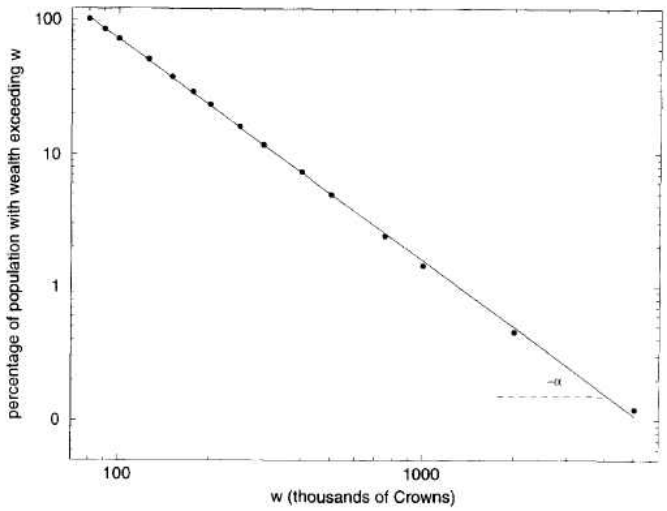

Fig. 2. Measurement of the percentage of the population with wealth exceeding different wealth levels. This measurement was done in Sweden. A power-law distribution of wealth yields a straight line on a log–log scale, with slope $-\alpha$. The empirical data is represented by dots, the solid line is the power-law fit. The empirical distribution is in excellent agreement with the power-law fit (Source: Steindl, 1965).

그림 6 스웨덴 부의 분포 [출처: M. Levy, S. Solomon, Physica A **242**, 90 (1997)]

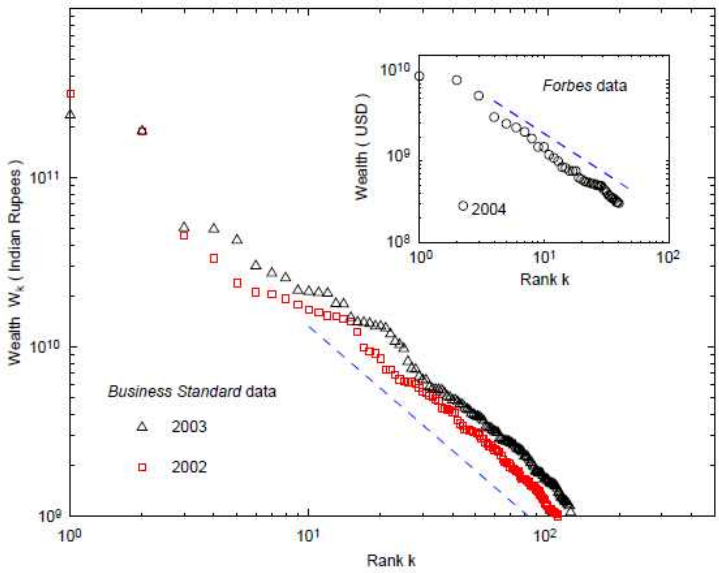

Fig. 1. Rank ordered plots of the wealth of the richest Indians during the period 2002–2004 on a double-logarithmic scale. The main figure shows the wealth of the kth ranked richest person (or household) against the rank k (with rank 1 corresponding to the wealthiest person) as per two surveys conducted by *Business Standard* in Dec 31, 2002 (squares) and Aug 31, 2003 (triangles). The broken line having a slope of −1.23 is shown for visual reference. The inset shows the rank ordered plot of wealth based on data published by *Forbes* in Dec 10, 2004, with the broken line having a slope of −1.08.

그림 7 인도 부의 분포 [출처: S. Sinha, Physica A **359**, 555 (2006)]

8. 아인슈타인의 유산

바라바시와 알버트의 척도 없는 네트워크 모형(Barabasi-Albert scale-free network; 줄여서 BA 모형)에서는 먼저 생성된 노드일수록 점점 더 많은 이웃을 갖게 되어 허브가 됩니다. 하지만 실제로 오래된 노드가 사라지거나, 새로운 노드가 기존 노드보다 더 많은 이웃을 갖는 허브로 발전하는 경우도 많습니다.

BA 모형은 모든 노드가 언제 생성되는지 외에 모두 똑같다고 가정합니다. 앞서 말한 후발주자의 장점을 고려하기 위해 각 노드에 랜덤한 적합도(fitness)를 부여합니다. 그리고 새로운 노드가 기존 노드에 연결될 확률을 기존 노드의 이웃수와 적합도의 곱에 비례하도록 합니다. 이런 방법을 통해 뒤늦게 생겼으나 적합도가 높은 노드가 먼저 생겼으나 적합도가 낮은 노드보다 더 많이 링크될 가능성이 있습니다. 이걸 책에서는 '적익부(fit-get-rich)'라고 하네요. "하지만 적합성은 가상의 양으로서 각 노드에 대해 적합성을 정확하게 측정하는 방법과 도구는 아직 개발되어 있지 않다." 비슷한 얘기를 얼마 전에 한 적이 있죠.

이제, 각 노드에 주어진 적합도(η)를 그 노드의 에너지(ε)로 해석하고($\epsilon = -kT\log\eta$; 여기서 k는 볼츠만 상수, T는 절대온도), 그 노드의 이웃수를 그 에너지를 갖는 입자의 개수로 해석합니다. 그러면 한 노드가 거의 모든 링크를 독식하는 현상을 양자역학의 보즈-아인슈타인 응축(Bose-Einstein condensation)으로 이해할 수 있습니다. 사실 대부분의 물리, 생물, 사회, 기술 네트워크들과 양자역학은 아무런 상관이 없습니다. 그래서 '해석'이라고 썼습니다. 다만 보즈-아인슈타인 응축에 관한 수학적 틀이 "적합도가 높은 노드의 링크 독식"에 그대로 적용될 수 있다는 걸 짚으려는 것입니다.

밑줄 하나. "노드들은 항상 연결을 위해서 경쟁한다. 상호 연관된 세계에서 링크는 곧 생존을 의미하기 때문이다."

9. 아킬레스 건

이 '링크'(장)에서는 네트워크의 견고성(robustness)을 다룹니다. 전력망에서 한 노드의 고장이 연쇄반응을 일으켜 전체 시스템을 마비시키기도 하고(1996년 미국 서부 정전 사태), 생태계에서 어떤 종들의 멸종이 생태계 전체에 영향을 주기도 하지요.

좀더 구체적인 질문으로 바꿔봅니다: 척도 없는 네트워크에서 노드를 무작위로 없앴을 때 언제까지 전체 구조가 유지될까? 척도 없는 네트워크에서 허브를 공격하면 어떻게 될까? 그런 (무작위 또는 허브에 먼저 가해지는) 공격으로부터 안정적인 네트워크 구조는 어떤 모양이어야 할까?

척도 없는 네트워크에서 무작위로 노드를 없애나가면, 대부분의 경우 이웃수가 매우 작은 노드들이 선택될 확률이 높으므로 전체 구조에 영향을 주기 힘듭니다. 반대로 이웃수가 많은 허브를 먼저 없애기 시작하면 전체 구조가 빠르게 무너집니다. 이게 바로 네트워크의 '아킬레스 건'입니다. 요컨대 무작위 공격에 대해 '견고'하지만 허브 공격에는 '취약'한 특징을 모두 보여줍니다.

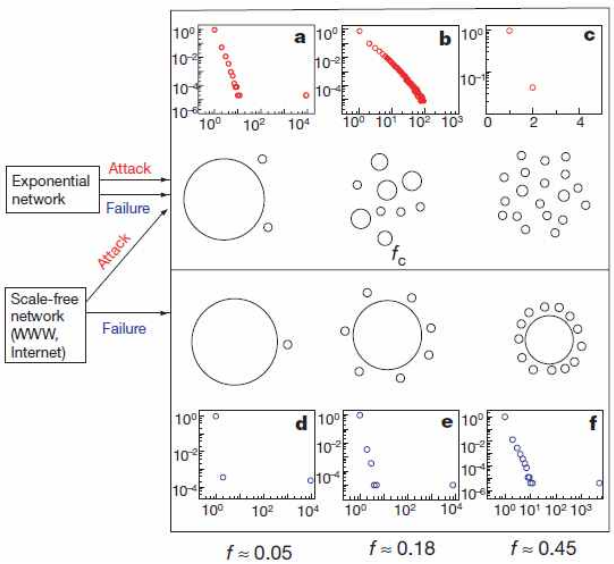

그림 8 무작위 고장(failure)과 허브 공격(attack)에 의한 네트워크 쪼개짐 [출처: R. Albert, H. Jeong, A.-L. Barabasi, Nature **406**, 378 (2000)]

10. 바이러스와 유행

다음으로 척도 없는 네트워크에서 바이러스나 유행이 어떻게 퍼지는가를 소개합니다. 척도 없는 네크워크가 아닌 경우, 한 노드에서 발생한 바이러스가 시스템 전체에 영향을 미치기 위해서는 그 바이러스의 전염성이 어떤 문턱값보다 커야 합니다. 하지만 척도 없는 네트워크에서는 그 문턱값이 0이 됩니다. 즉 일단 발생한 바이러스는 웬만해서는 사라지지 않는다는 말이죠.
결국 여기서도 허브의 존재가 중요합니다. 전염성을 한 노드에서 다른 노드로 병이 전염될 확률로 정의합시다. 허브가 감염되면 전염성이 매우 낮더라도 이웃수가 너무 많아서 결국 바이러스가 퍼지게 됩니다. 매우 단순하게 말했는데, 결국 이 얘기겠죠.
많은 사회연결망이 척도 없는 성질을 보인다는 사실은 전염병을 막기 위해 허브를 먼저 공략해야 한다는 교훈을 주지만, 감염된 사람 중에 누군가를 '선택'해야 한다는 면에서 윤리적인 문제 또한 갖고 있습니다. 반대로 마케팅을 하는 입장에서는 이런 구조가 매우 반가울 것 같네요.

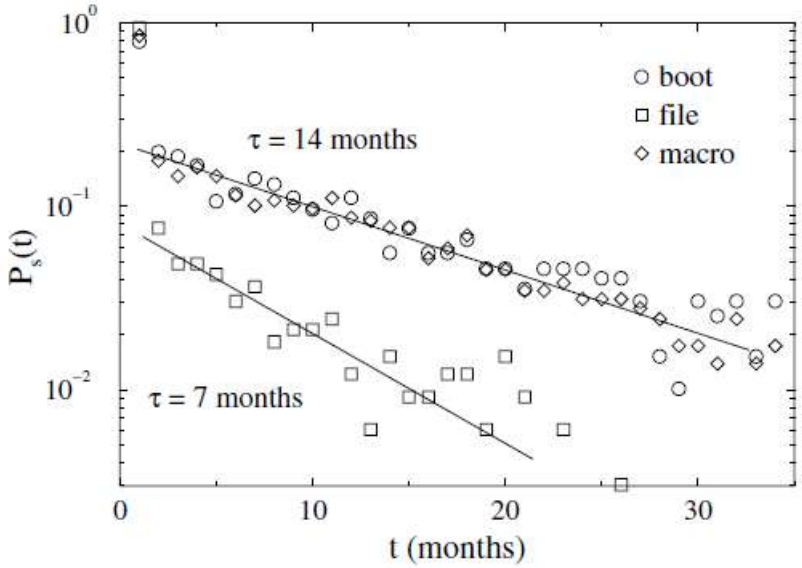

그림 9 컴퓨터 바이러스의 생존확률 [출처: R. Pastor-Satorras, A. Vespignani, Physical Review Letters 86, 3200 (2001)]

11. 인터넷의 등장

냉전 시대에 외부의 공격에 대해 견고한 통신 시스템을 만들고자 하는 시도가 있었습니다. 이때 폴 배런은 중앙집중형(1개의 허브에 지국들이 연결)이나 탈집중형(몇 개의 허브 구조들이 연결)이 아니라 분산형을 제안했는데 무시당했다네요.
여튼 그보다 나중인 1960년대에 컴퓨터들을 연결시킨다는 개념이 발전하여 비로소 인터넷이 생겼다고 합니다. 첫 연결은 1969년 UCLA와 스탠퍼드 사이에서 이루어집니다. 그해 10월과 11월에 UC 산타 바바라와 유타 대학에 3, 4번째 노드가 만들어지고... 1971년 말까지 인터넷을 이루는 노드는 15개로 늘어납니다. 지금은 몇 개나 될까요.

12. 웹의 분화 현상

웹페이지 사이의 연결은 방향성이 있습니다. 들어오는 링크가 적은 웹페이지는 발견되기 힘들며 심지어 존재하는지조차 알 수 없는 경우도 많지요. 이 방향성에 의해 웹의 구조는 4개의 '대륙'으로 나뉜다고 합니다.
우선 '중심핵'의 페이지들은 서로 연결되는 경우가 많습니다. 많은 포탈 사이트나 요즘으로 치면 유명한 블로그 등이 여기 포함되겠죠. 다음으로 IN대륙과 OUT대륙인데요, IN대륙에서 출발하면 중심핵으로 갈 수 있지만 그 반대로는 갈 수 없습니다. 이를

테면 개인 웹페이지 정도 되겠네요. 중심핵에서 OUT대륙으로 갈 수 있지만 그 반대로는 갈 수 없습니다. 기업의 홍보용 웹사이트들이 OUT대륙에 많다고 합니다. 마지막 대륙은 IN대륙에서 중심핵을 거치지 않고 OUT대륙으로 연결시켜주는 덩굴과 다른 대륙들로부터 고립된 섬들로 이루어집니다.

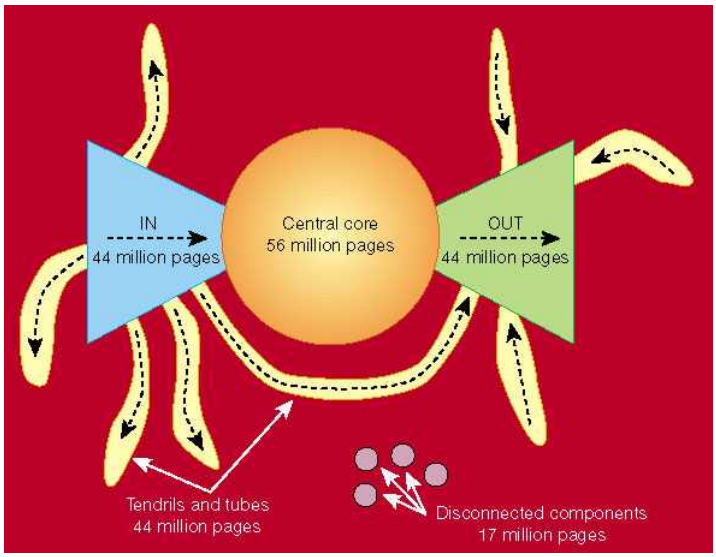

그림 10 [출처: Nature **405**, 113 (2000)]

웹의 방향성은 피할 수 없으므로, 웹의 전체 지도를 그린다는 목표는 쉽게 달성될 수 없다는 것도 분명합니다. 물론 웹뿐 아니라 방향성이 있는 네트워크라면 이런 분화 현상이 일반적으로 나타날 것입니다.

"사이버스페이스를 진정으로 이해하기 위해서는 코드와 아키텍처를 보다 면밀하게 구분할 필요가 있다는 것이 나의 생각이다. 코드 또는 소프트웨어는 사이버스페이스를 구성하는 벽돌과 시멘트에 해당한다. 한편 아키텍처는 코드를 벽돌 삼아 쌓아올린 구조물을 의미하는 것이다. (중략) 웹의 아키텍처는 중요한 두 가지 층위인 코드와 그 코드를 이용하는 인간의 집합적 행동이 함께 작용해 얻어진 산물이다. 전자는 얼마든지 법원과 정부 또는 기업의 규제 대상이 될 수 있다. 그렇지만 후자는 일개 사용자나 기관에 의해 만들어지는 것이 아니다. (중략) 웹은 수백만 명에 달하는 사용자들이 취하는 개별적 행동에 모두 영향을 받으면서 진화하며, 그 결과로 나타난 아키텍처는 부분들의 단순한 총합 이상의 의미를 지닌다."

13. 생명의 지도

생명현상을 이해하는데도 네트워크가 중요하다는 얘기를 합니다. "여러 단계의 세포 간의 생화학 반응 과정"으로 구성된 신진대사 네트워크라든지, "유전자와 DNA에서 해독된 정보로부터 만들어진 단백질"을 노드로 하고, "이러한 단백질 사이에서 일어나는 생화학 반응"을 링크로 하는 조절 네트워크를 이해할 필요가 있다는 말입니다.

43개의 유기체에 대한 신진대사 반응을 정리해놓은 웹사이트로부터 자료를 내려받아 분석한 결과, 3단계 분리라는 좋은 세상 효과를 확인했다고 합니다. 더 놀라운 건 네트워크 크기가 서로 다른 43개 유기체에서 모두 일정한 평균거리가 나타났다는 사실입니다. 허브 역할을 하는 분자는 ATP, ADP, 물이라네요. 그리고 척도 없는 네크워크라는 것도 확인합니다.

단백질 상호작용 네트워크가 어떻게 척도 없는 네크워크가 되었는지에 대한 설명으로 단백질의 복제과정이 성장과 선호적 연결을 모두 만족시키고 있다고 합니다.

14. 네트워크 경제

기업의 이사회에 소속된 사람들은 여러 기업의 이사를 맡는 경우가 있습니다. 이사들의 활동을 분석한 결과, 79%의 사람들은 한 회사에만 몸담고 있었고, 14%는 두 회사, 7%는 세 회사에 몸담고 있었습니다. 그중 버논 조르단은 10개 회사의 이사였다고 합니다. 이런 현상도 일종의 '선호적 연결'이 작용한 결과입니다. "회사는 여러 회사에 걸쳐 있는 영향력 있는 이사들을 영입하고 싶어"하는데, "그 회사로서는 상호 간에 걸쳐 있는 이사를 통해 다른 회사의 경험을 이용할 수 있기 때문"이랍니다.

이외에도 바이오 산업 네트워크의 성장에서 허브의 역할을 하는 회사가 존재한답니다. 회사 간의 제휴와 협력 역시 네트워크를 통해 분석될 수 있고, 1997년 아시아 금융 위기 등도 책에 언급됩니다.

15. 거미 없는 거미줄

끝입니다.

각 링크마다 제가 '밑줄'을 그은 건 보면 아시겠지만, 네트워크의 "노드들은 왜 다른 노드를 링크하는가?"에 대한 답을 얻기 위해서입니다. 사회 네트워크의 경우, "당연한 욕구" 외에는 "살아남기 위해서" 정도가 눈에 띕니다. 사회 네트워크를 제대로 이해하기 위해서는 이 부분에 대한 미시적인 설명이 더 필요하며, 이건 게임이론 등 미시경제학에서 이루어져온 네트워크 연구로부터 배워야 하겠습니다.

이후의 네트워크 연구 동향:
1. 공동체 찾기 알고리즘 개발
2. 행위자와 네트워크의 공진화