COMPUTATIONAL SCIENCES

구글은 과연 모든 것을 알고 있을까?

2017. 12. 27.

이주영

수학에서 그래프라고 불리는 네트워크는 주어진 노드와 그 노드들 사이의 상호작용을 나타내 는 링크의 집합으로 정의할 수 있다. 엄청난 양의 데이터가 쏟아지고 있는 최근에 특히 주목받고 있는 네트워크 과학은 복잡한 시스템의 구성요소(노드)와 그들 간의 상호작용(링크)을 전체적으로 이해하려고 연구하는 학문이다. 복잡계 연구에는 여러 가지 측면이 있겠지만 여기에서는 community(혹은 module) detection에 대하여 소개하고자 한다.

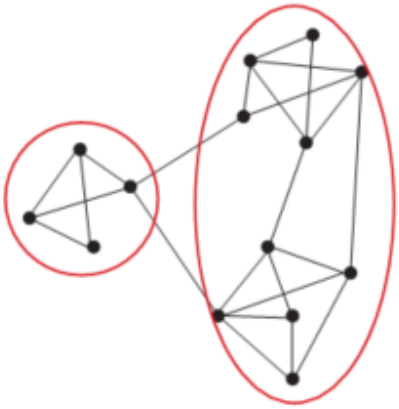
우선 이해를 돕기 위해 소셜 네트워크 중의 하나인 Zachary's Karate Club을 소개한다. 이 네트워크는 1970년대 미국 한 대학의 친분관계를 기초로 만들어진 것으로 34명의 회원, 즉 34개의 노드와 그들 사이의 관계(링크)로 구성되어 있다. 원래 하나이던 Club이 회원들 사이의 불화로 인 하여 2개의 Club으로 나누어지게 된다. 사회학자들은 회원들과의 인터뷰를 통하여 회원들 사이 에 친분관계(링크)를 할당하여 네트워크를 만들었고 이 네트워크를 이해함으로써 이 Club이 2개 의 소그룹으로 나누어지게 된 현상을 이해하려하였다. 즉, 주어진 네트워크의 숨겨진 노드 사이 의 community를 네트워크의 위상정보를 이용하여 파악하려 한 것이다.

네트워크의 링크에는 다양한 종류가 있지만 여기서는 가장 간단한 종류의 링크를 고려한다. 즉 모든 링크에 1이라는 같은 가중치를 부여하고 링크의 방향성이 없다고 가정하자. 예를 들면 <그림 1>과 같은 네트워크가 주어졌다면 과연 이 네트워크의 community를 어떻게 구할 수 있을 까? Girvan과 Newman은 다음과 같이 정의된 Modularity(Q)라는 양을 극대화하는 partition 을 사용하자고 제안하였다.[1]

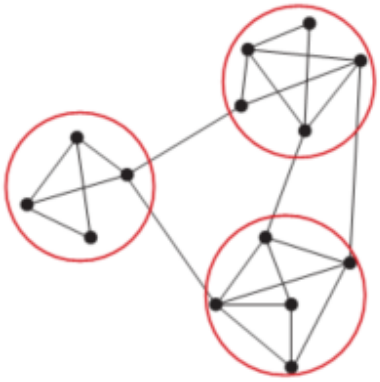
$$Q = \sum_{s=1}^{r}\left[\frac{l_s}{L} - \left(\frac{d_s}{2L}\right)^2\right] \quad Eq.\ (1)$$

여기서 s는 community index이며, 주어진 네트워크의 노드를 r개의 community로 중첩 없이 나누는 경우를 생각한다. L은 네트워크의 전체 링크의 숫자, ls는 community s 안에 존재하는 링크의 수, ds는 community s에 속하는 노드의 연결수의 합으로 정의된다. 여기서 노드의 연결수는 그 노드에 연결된 링크의 수이다. 위의 식에서 첫째 항은 전체 링크 중에서 같은 community 안에서 연결된 링크의 비율이고, 둘째 항은 주어진 네트워크의 링크를 무작위로 재분배하였을 때 의 첫째 항의 기대치이다. 모든 노드를 하나의 community로 할당할 경우에 $Q=0$임을 쉽게 알 수 있다.

<그림1>



<그림 2>



<그림 3>

또한 각각의 노드가 하나의 community로 할당되는 partition은 음수의 Q 값을 갖게 된다. 참고로, <그림 2>나 <그림 3>처럼 나누는 partition도 생각할 수 있는데 이 경우 Q값은 각각 0.292, 0.489이 된다(각자 확인 바람). Modularity 최대화 문제는 모든 가능한 partition 중에서 Q의 값이 최대가 되는 해를 찾는 combinatorial optimization 문제로 전형적인 NP-hard 문제이다.

*Eq. (1)*과 관련하여 두 가지 고민해야할 문제가 있다. 첫 번째 문제는 비교적 정의가 명확한 수학적인 문제이다. 즉 주어진 네트워크에서 Q를 최대화하는 문제이다.[2]를 참고하면 여러 분야에서 자주 연구되어 온 인기 있는 네트워크 정보를 얻을 수 있는데 여기에는 위에서 언급한 Zachary's Karate Club과 같이 작은 네트워크를 비롯하여 수 만 개 이상의 노드로 구성된 것들도 포함되어 있다.

이와 같은 네트워크의 Modularity를 최대화하는 연구는 지난 10여 년간 활발히 진행되어 왔는데, 주로 새로운 heuristic 방법의 개발, 또는 잘 알려진 Simulated Annealing 방법을 적용하여 진행되어 왔다. 주어진 네트워크의 Modularity Optimization 문제는 좀 더 효율적이고 강력한 sampling/search 방법의 개발로 이어져 왔다. 필자의 연구실에서 관심을 갖고 연구하는 주제 중의 하나가 복잡하고 어려운 multiple minima 문제들을 효율적으로 풀어나가는 것으로, Conformational Space Annealing 방법을 개발하여 이러한 문제들에 적용해 왔다.[3] 놀랍게도 이러한 연구를 통하여 기존에 알려진 Modularity 최대 해를, 다

섯 문제의 경우, 새로운 해로 갈아치울 수 있었다.[4] 그렇다면, 과연 더 좋은 Modularity 해를 얻는다는 것은 무슨 뜻일까?

*Eq. (1)*과 관련된 두 번째 문제는, 과연 더 좋은 Modularity 해가 community에 대해 더 유익한 정보를 줄 수 있겠는가 하는 것이다. 그리고 과연 네트워크의 위상적 정보만으로 노드들 사이의 직접적인 관계로부터는 명백하지 않은 추가적인 정보를 얻을 수 있을 것인가 질문할 수 있다. 간단히 결론만 말하자면 기존의 연구 결과, 발표와는 달리 이러한 community 정보가 매우 유용 하다. 아래에서 그 두 가지 예를 들도록 하겠다.

복잡계 네트워크의 새로운 영역으로 생물학적 네트워크가 중요한 연구 영역으로 대두되고 있다. 이는 지난 10여 년간 유전체 규모의 자동화된 실험 방법들이 개발되면서 대규모의 상호작용 정보가 제공되면서 더욱 가속화된 바 있는데, 예를 들면, 생체 대사 네트워크, 단백질-단백질 네트워크 등이다.

단백질들 사이의 상호작용 정보는 대규모로 제공되는 반면, 각각의 단백질의 기능에 대한 정보는 일일이 알아내야하는 어려움이 있다. 이미 기능이 밝혀진 단백질의 정보를 이용하여 아직 기능이 밝혀지지 않은 단백질의 기능을 유추할 수 있다면 단백질 기능 연구에 들어가는 시간과 경비를 크게 줄일 수 있다. 기존의 연구 결과에 따르면 community 정보가 기대와는 달리 오히려 단백질의 기능을 유추하는데 방해가 된다고 알려져 왔다.[5,6] 이에 반하여, 최근에 제대로 된 Modularity optimization과 적절한 community 정보를 이용함으로써 단백질의 기능 예측을 향상시킬 수 있다는 결과가 얻어지고 있다.

두 번째 예로 필자가 올해 3월에 일본에서 개최된 학회에 참석한 초청연사 네트워크를 만들어 봄으로써 생긴 일화를 소개한다. 이 학회에서 필자는 단백질-단백질 네트워크 연구에 대한 발표의 서론을 전개하기 위해서 초청연사 23명의 네트워크를 먼저 보여주고 마지막에 이 네트워크의 community detection 결과도 보여주었다. 23명이 세 그룹 (11+9+3)으로 나뉘었는데 6명의 이론/계산 연사가 모두 11명으로 구성된 community에 속하게 되었고 4명의 서구권 연사 중에 3명 이 하나의 community에 배정되었다.

필자의 발표 후에 몇몇의 연사가 찾아와, 두 번째 community 9명에 한명의 리더 교수와 20여 년간 이 교수의 연구실에서 학위 및 연구원 과정을 거친 5명이 모두 속해 있어서 본인들이 깜짝 놀랐다고 알려주었다. 참고로 이 네트워크의 링크는 현재 NIH 연구원인 이주용 박사가 Google 검색으로 생성한 가중치로 구성한 것이었다. Google이 모든 것을 알고 있다는 말인가? 섬뜩한 기분이 들면서, Modularity의 최대화를 이용한 community detection 연구가 여러 분야에서 유용하게 쓰일 가능성과 파급효과를 점치게 된다.

참고문헌

1. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America 99, 7821- 7826 (2002).
2. http://www-personal.umich.edu/~mejn/netdata/
3. Lee, J., Lee, I.-H. & Lee, J. Unbiased Global Optimization of Lennard-Jones Clusters for N≤201 Using the Conformational Space Annealing Method. Physical Review Letters 91, 080201 (2003).
4. Lee, J., Gross, S. P. & Lee, J. Modularity optimization by conformational space annealing. Physical Review E, 056702 (2012).
5. Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. Molecular Systems Biology 3, 88 (2007).
6. Song, J. & Singh, M. How and when should interactome-derived clusters be used to predict functional modules and protein function? Bioinformatics 25, 3143-3150 (2009).
7. Lee, J & Lee, J. Hidden information revealed by optimal community structure from a protein-complex network improves protein function prediction. PLOS ONE (2013).