On Modularity Clustering

Presented by:
Group III
(Ying Xuan, Swati Gambhir &
Ravi Tiwari)

Modularity

- A quality index for clustering a graph G=(V,E)

$$q(\mathcal{C}) := \sum_{C \in \mathcal{C}} \left[\frac{|E(C)|}{m} - \left(\frac{|E(C)| + \sum_{C' \in \mathcal{C}} |E(C, C')|}{2m} \right)^2 \right]$$

This is equivalent to:

$$q(\mathcal{C}) := \sum_{C \in \mathcal{C}} \left[\frac{|E(C)|}{m} - \left(\frac{\sum_{v \in C} \deg(v)}{2m} \right)^2 \right]$$

Contribution of the Paper

- Integer Linear Program Formulation
- Fundamental Observations & Counterintuitive Behavior
- NP- Completeness of Maximizing Modularity Problem
- A Greedy Algorithm
- Optimality Results

Maximizing Modularity via Integer Linear programming

Given a graph G=(V,E), n=|V| nodes, n^2 decision variables X_{uv} $=\{0,1\}$

$$Max : q(C) := \frac{1}{2m} \sum_{(u,v)\in V^2} \left[E_{uv} - \frac{\deg(u)\deg(v)}{2m} \right] X_{uv}$$

$$St :$$

$$\forall u : X_{uu} = 1$$

$$\forall u,v : X_{uv} = X_{vu}$$

$$\forall u,v,w : X_{uv} + X_{vw} - 2X_{uw} \le 1$$

$$\forall u,v : X_{uv} \in \{0,1\}$$

Fundamental Observations

If **G** *is an undirected and un-weighted graph and* **C** *is a clustering then:*

$$-\tfrac{1}{2} \leq q(c) \leq 1$$

- *When all the edges are inter-cluster q(C)=-1 / 2, eg: Bipartite graph G=(X:Y,E)with cluster X andY*

- *When all the clusters cliques with no inter-cluster edges q(C)=1, when number of clusters are infinite*

Fundamental Observations(Contd)

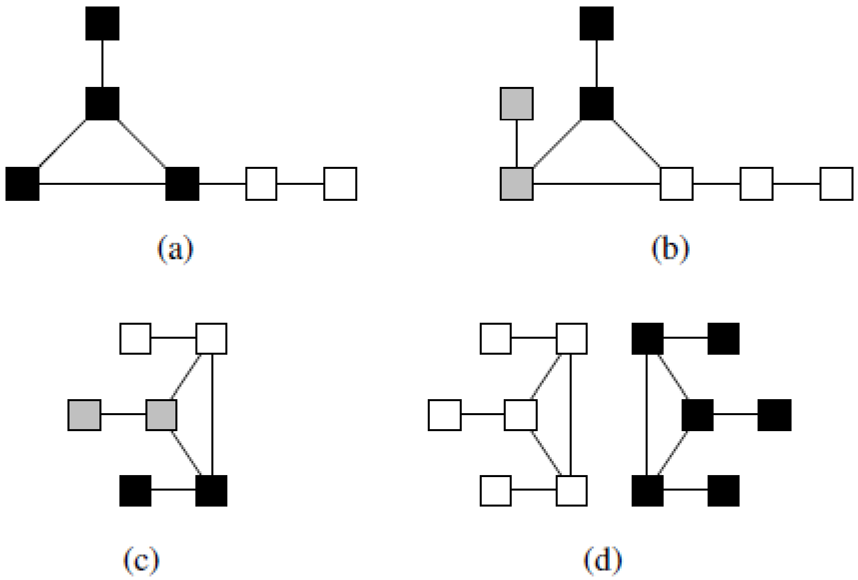
Clustering with maximum modularity has no cluster with single node having degree 1.

Fundamental Observations(Contd)

In clustering with maximum modularity each cluster consist of a connected sub-graph

Counterintuitive Behavior

- Non-locality
- Sensitivity to Satellite
- Scaling Behavior



NP-Completeness

- *Problem 1(Modularity):*

  Given a graph ***G*** *and a number* ***K*** *is there a clustering* ***C*** *of* ***G,*** *for which* $q(C) >= K$

- *Problem 2(3-Partition):*

  Given 3k positive integers numbers $a_1, a_2, \ldots, a_{3k}$ *such that the sum* $\sum_{i=1}^{3k} a_i = kb$ *and* $b/4 < a_i < b/2$ *for an integer* ***b*** *and for all* $i = 1, 2, \ldots 3k$ *is there a partition of these numbers into k sets, such that the numbers in each set sums upto b?*

NP-Completeness (contd)

- An instance $\boldsymbol{A}=\{a_1, a_2, \ldots, a_{3k}\}$ of 3-Partition can be transformed in to an instance (G(A), K(A)) of Modularity
- G(A) has a clustering with modularity at least $\boldsymbol{K(A)}$, if and only if $a_1, a_2, \ldots, a_{3k}$ can be partitioned into k set of sum $b = \frac{1}{k}\sum_{i=1}^{3k} a_i$

NP-Completeness (contd)

- Construct a graph ***G(A)*** with ***k*** cliques $H_1, H_2, \ldots H_k$ of size $a = \sum_{i=1}^{3k} a_i$ each.
- For each element $a_i \in A$ introduce a single element node. And connect it to a_i nodes in each of the ***k*** cliques.
- Each member of clique is connected to exactly one element node.
- Each clique node has degree ***a***, each element node $a_i \in A$ has a degree $\mathbf{ka_i}$.
- The number of edges in ***G(A)*** is $m=(k/2)a(a+1)$

NP-Completeness (contd)

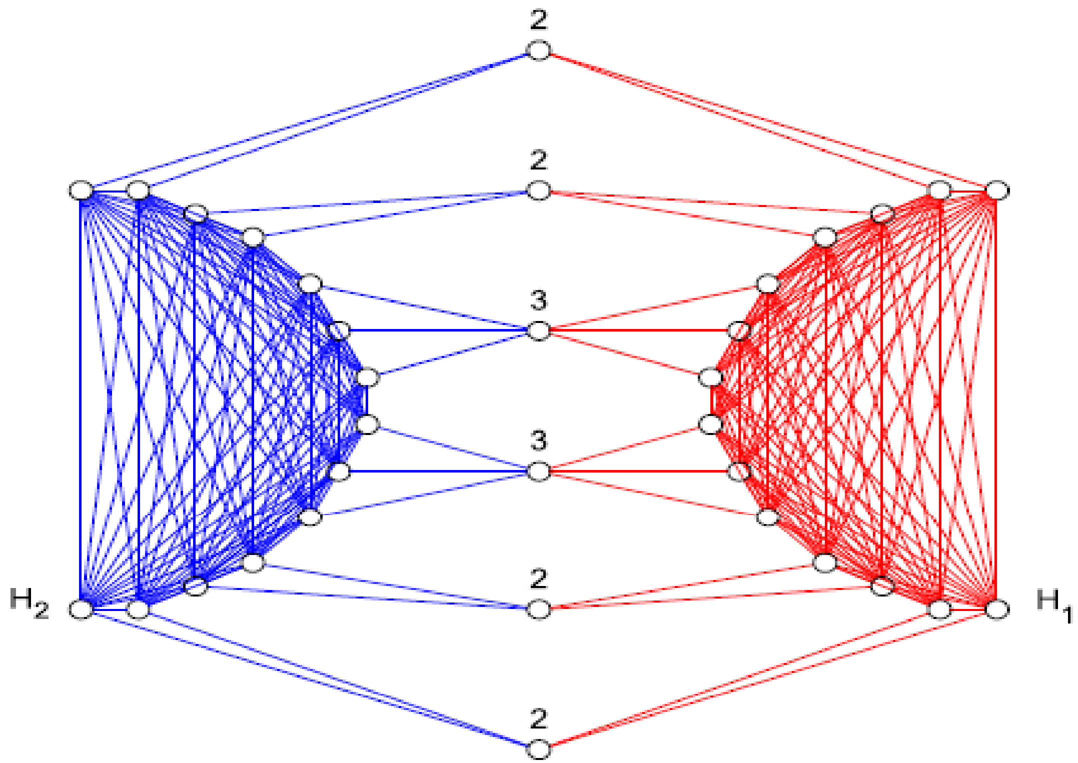


Fig. 2. An example graph $G(A)$ for the instance $A = \{2, 2, 2, 2, 3, 3\}$ of 3-PARTITION. Node labels indicate the corresponding numbers $a_i \in A$.

NP-Completeness (contd)

Lemma 4.1:In maximum modularity clustering of $G(A)$ none of the cliques $H_1, H_2, \ldots H_k$ split.

- Clustering C splits a clique $H \in \{H_1, H_2, \ldots H_k\}$.
- $\{C_1, C_2, \ldots C_r\} \in C$ are clusters that contains the nodes of H.
- n_i is the number of nodes of H contained in cluster C_i
- $m_i = |E(C_i)|$ is the number of edges between nodes in C_i.
- f_i is the number of edges between the nodes of H and C_i.
- d_i is the sum of degree of all nodes in C_i.

NP-Completeness (contd)

The contribution of $\{C_1, C_2, \ldots C_r\} \in C$ to ***q(C)*** is:

$$\frac{1}{m}\sum_{i=1}^{r} m_i - \frac{1}{4m^2}\sum_{i=1}^{r} d_i^2$$

Rearranging nodes in $\{C_1, C_2, \ldots C_r\} \in C$ into clusters ***C', C₁', C₂', … Cᵣ'*** such that:

- ***C'*** contains the nodes of H
- And each C_i' contains the remaining nodes of C_i

NP-Completeness (contd)

- The contribution of $C', C_1', C_2', \ldots C_r'$ to $q(C')$ is given as

$$\frac{1}{m}\sum_{i=1}^{r}\left(m_i + \sum_{j=i+1}^{r} n_i n_j - f_i\right) - \frac{1}{4m^2}\left(a^4 + \sum_{i=1}^{r}(d_i - n_i a)^2\right) .$$

Setting $\Delta := \mathsf{q}(\mathcal{C}') - \mathsf{q}(\mathcal{C})$, we obtain

$$\begin{aligned}
\Delta &= \frac{1}{m}\left(\sum_{i=1}^{r}\sum_{j=i+1}^{r} n_i n_j - f_i\right) \\
&\quad + \frac{1}{4m^2}\left(\left(\sum_{i=1}^{r} 2d_i n_i a - n_i^2 a^2\right) - a^4\right) \\
&= \frac{1}{4m^2}\left((4m\sum_{i=1}^{r}\sum_{j=i+1}^{r} n_i n_j - 4m\sum_{i=1}^{r} f_i \right. \\
&\quad \left. + \left(\sum_{i=1}^{r} n_i\left(2d_i a - n_i a^2\right)\right) - a^4\right) .
\end{aligned}$$

NP-Completeness (contd)

- Substituting $2\sum_{i=1}^{r}\sum_{j=i+1}^{r} n_i n_j = \sum_{i=1}^{r}\sum_{j\neq i} n_i n_j$ and $m = \frac{k}{2}a(a+1)$

$$\begin{aligned}\Delta &= \frac{a}{4m^2}\Bigg(-a^3 - 2k(a+1)\sum_{i=1}^{r} f_i \\ &\quad + \sum_{i=1}^{r} n_i \Big(2d_i - n_i a + k(a+1)\sum_{j\neq i} n_j \Big)\Bigg) \\ &\geq \frac{a}{4m^2}\Bigg(-a^3 - 2k(a+1)\sum_{i=1}^{r} f_i \\ &\quad + \sum_{i=1}^{r} n_i \Big(n_i a + 2k f_i + k(a+1)\sum_{j\neq i}^{r} n_j \Big)\Bigg)\end{aligned}$$

The fact that $n_i a + k f_i \leq d_i$ is used

NP-Completeness (contd)

- Combining n_i and one of the $\sum_{j \neq i} n_j$

$$
\begin{aligned}
\Delta \quad \geq \quad & \frac{a}{4m^2}\left(-a^3 - 2k(a+1)\sum_{i=1}^{r} f_i \right) \\
& + \frac{a}{4m^2}\left(\sum_{i=1}^{r} n_i \left(a\sum_{j=1}^{r} n_j + 2kf_i \right.\right. \\
& \qquad \left.\left. +((k-1)a+k)\sum_{j\neq i}^{r} n_j \right)\right) \\
= \quad & \frac{a}{4m^2}\left(-2k(a+1)\sum_{i=1}^{r} f_i \right. \\
& \left. + \sum_{i=1}^{r} n_i \left(2kf_i + ((k-1)a+k)\sum_{j\neq i}^{r} n_j \right)\right) \\
= \quad & \frac{a}{4m^2}\left(\sum_{i=1}^{r} 2kf_i(n_i - a - 1)) \right. \\
& \left. +((k-1)a+k)\sum_{i=1}^{r}\sum_{j\neq i}^{r} n_i n_j \right) \\
\geq \quad & \frac{a}{4m^2}\left(\sum_{i=1}^{r} 2kn_i(n_i - a - 1) \right. \\
& \left. +((k-1)a+k)\sum_{i=1}^{r}\sum_{j\neq i}^{r} n_i n_j \right) \quad ,
\end{aligned}
$$

Here, n_i<=a-1, so n_i-a-1<0,
Hence, f_i increases
modularity difference
decreases and f_i<=n_i

NP-Completeness (contd)

- Rearranging and using a>=3k

$$\begin{aligned}
\Delta &\geq \frac{a}{4m^2}\sum_{i=1}^{r} n_i\left(2k(n_i - a - 1) + ((k-1)a + k)\sum_{j\neq i}^{r} n_j\right) \\
&= \frac{a}{4m^2}\sum_{i=1}^{r} n_i\left(-2k + ((k-1)a - k)\sum_{j\neq i}^{r} n_j\right) \text{, As } n_i + \sum_{j\neq i} n_j = a \\
&\geq \frac{a}{4m^2}((k-1)a - 3k)\sum_{i=1}^{r}\sum_{j\neq i}^{r} n_i n_j \\
&\geq \frac{3k^2}{4m^2}(3k - 6)\sum_{i=1}^{r}\sum_{j\neq i}^{r} n_i n_j \quad .
\end{aligned}$$

- Assuming k>2, we see $\Delta > 0$

NP-Completeness (contd)

Lemma 4.2: In a maximum modularity clustering of G(A), every cluster contains at most one of the cliques

$\boldsymbol{H_1, H_2, \ldots H_k}$

- Cluster **C** contains $l > 1$ cliques completely and some element nodes a_j with $j \in J \subseteq \{1, 2, \ldots 3k\}$.
- Inside l cliques $la(a-1)/2$ edges are covered and degree sum is $l\, a^2$.
- For each element node a_j, $l\, a_j$ edges are covered and degree sum is $k \sum_{j \in J} a_j$.

NP-Completeness (contd)

- The contribution of **C** to ***q(C)*** is:

$$\frac{1}{m}\left(\frac{l}{2}a(a-1)+l\sum_{j\in J}a_j\right)-\frac{1}{4m^2}\left(la^2+k\sum_{j\in J}a_j\right)^2$$

- Clustering ***C'*** in which **C** is split into **C$_1$'** and **C$_2$'**.
- **C$_1$'** completely contains a single clique **H**.
- The contribution of **C$_1$'** and **C$_2$'** to ***q(C')*** is:

$$\frac{1}{m}\left(\frac{l}{2}a(a-1)+(l-1)\sum_{j\in J}a_j\right)$$
$$-\frac{1}{4m^2}\left(\left((l-1)a^2+k\sum_{j\in J}a_j\right)^2+a^4\right)$$

NP-Completeness (contd)

- Considering the difference :

$$
\begin{aligned}
\mathsf{q}(\mathcal{C}') - \mathsf{q}(\mathcal{C}) &= -\frac{1}{m}\sum_{j\in J} a_j \\
&\quad + \frac{1}{4m^2}\left((2l-1)a^4 + 2ka^2\sum_{j\in J} a_j - a^4\right) \\
&= \frac{2(l-1)a^4 + 2ka^2\sum_{j\in J} a_j}{4m^2} \\
&\quad - \frac{4m\sum_{j\in J} a_j}{4m^2} \\
&= \frac{2(l-1)a^4 - 2ka\sum_{j\in J} a_j}{4m^2} \\
&\geq \frac{9k^3}{2m^2}(9k-1) \\
&> 0,
\end{aligned}
$$

As ***k>0*** for all instance of 3-Partition

NP-Completeness (contd)

Lemma 4.3: In maximum modularity clustering of $\boldsymbol{G(A)}$, there is no cluster composed of element nodes only.

- Element node $\boldsymbol{v_i}$, corresponds to the element $\boldsymbol{a_i}$, which is not a part of any clique cluster. The node $\boldsymbol{v_i}$ forms a singleton cluster $\mathbf{C}=\{\boldsymbol{v_i}\}$.
- $\mathbf{C_{min}}$ is the clique cluster, for which the sum of degrees is minimal.
- $\mathbf{C_{min}}$ contains all nodes from clique $\boldsymbol{H}$, and some other element node $\boldsymbol{a_j}$.

NP-Completeness (contd)

- The contribution of **C** and **C**$_{\mathbf{min}}$ to ***q(C)*** is:

$$\frac{1}{m}\left(\frac{a(a-1)}{2}+\sum_{j\in J}a_j\right)-\frac{1}{4m^2}\left(\left(a^2+k\sum_{j\in J}a_j\right)^2+k^2a_i^2\right)$$

- Joining **C** and **C**$_{\mathbf{min}}$ to form a new cluster **C'**, now the contribution of **C'** to ***q(C')*** is:

$$\frac{1}{m}\left(\frac{a(a-1)}{2}+a_i+\sum_{j\in J}a_j\right)-\frac{1}{4m^2}\left(a^2+ka_i+k\sum_{j\in J}a_j\right)^2$$

NP-Completeness (contd)

Now,

$$\begin{aligned}
\mathsf{q}(\mathcal{C}') - \mathsf{q}(\mathcal{C}) &= \frac{a_i}{m} - \frac{1}{4m^2}\left(2ka^2a_i + 2k^2a_i\sum_{j\in J}a_j\right) \\
&= \frac{1}{4m^2}\left(2ka(a+1)a_i - 2ka^2a_i\right. \\
&\qquad \left. -2k^2a_i\sum_{j\in J}a_j\right) \\
&= \frac{a_i}{4m^2}\left(2ka - 2k^2\sum_{j\in J}a_j\right).
\end{aligned}$$

As for C_{min},

$$\sum_{j\in J} a_j \le \tfrac{1}{k}(a - a_i) < \tfrac{1}{k}a$$

Hence, $q(C') - q(C) > 0$

NP-Completeness (contd)

Theorem 4.4: Modularity is strongly NP-complete

- Polynomial time check $Modularity \in NP$
- Transformation of an instance of 3-Partition problem $\boldsymbol{A}=\{a_1, a_2, \ldots, a_{3k}\}$ to an instance of Modularity (G(A),K(A)).
- Clustering G(A) follows the properties derived in previous lemmas.
- Any clustering yields ***(k-1)a*** inter-cluster edges, so the edge coverage is:

$$\sum_{C \in \mathcal{C}} \frac{|E(C)|}{m} = \frac{m-(k-1)a}{m}$$

$$= 1 - \frac{2(k-1)a}{ka(a+1)} = 1 - \frac{2k-2}{k(a+1)}$$

NP-Completeness (contd)

- The clustering with maximum modularity must minimize

$$\deg(C_1)^2 + \deg(C_1)^2 + \ldots + \deg(C_k)^2$$

- This depends upon the distribution of element nodes, for the optimum case the distribution should be as even as possible.
- In the optimum case to each cluster assign element nodes that sums to $b = \frac{1}{k} a$
- In this case the sum of degrees of element nodes in each cluster is equal to $k \frac{1}{k} a = a$.
- This yields: $\deg(C_1)^2 + \ldots + \deg(C_k)^2 \geq k(a^2 + a)^2 = ka^2(a+1)^2$

NP-Completeness (contd)

- This yields: $\deg(C_1)^2 + \ldots + \deg(C_k)^2 \geq k(a^2 + a)^2 = ka^2(a+1)^2$
- Equality holds only if an assignment of ***b*** is possible to every cluster.
- If the clustering ***C*** with ***q(C)*** of at least:

  The clust $K(A) = 1 - \frac{2k-2}{k(a+1)} - \frac{ka^2(a+1)^2}{k^2a^2(a+1)^2} = \frac{(k-1)(a-1)}{k(a+1)}$ ***k*** clique clusters.
- The assignment of element nodes in ***k*** clique clusters is also the solution to the 3-Partition problem.
- Hence, this choice of ***K(A)*** the instance ***(G(A),K(A))*** of Modularity is satisfied only if the instance ***A*** of 3-Partition is satisfied, and vice-versa.

On Modularity Clustering (part 2)

Ying Xuan

March 3, 2009

Table of contents

Two Intrinsic Characteristics For Community Structure Detection:

- No knowledge over the size of each community structure;
- No knowledge over the number of community structures of the input graph.

Definition

k-Modularity

Given a graph G and a number K, is there a clustering $\mathcal{C}$ of G into exactly/at most k clusters, for which $q(\mathcal{C}) \geq K$?

Reduce from Minimum Bisection for Cubic Graphs.

Reduction

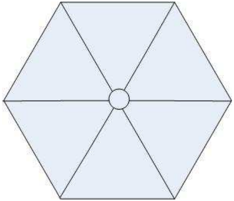
Minimum Bisection for Cubic Graphs(MB3)

Given a 3-regular graph G with n nodes and an integer c, is there a clustering into two clusters of $n/2$ nodes each such that it cuts at most c edges?

This problem is strongly NP-complete.

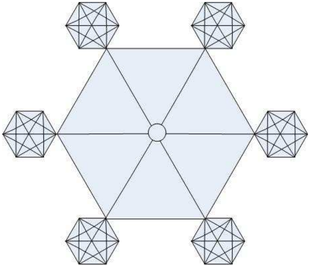
MB3 Instance

Construct a 2-Modularity instance from a MB3 Instance



2-Modularity Instance

Construct a 2-Modularity instance from a MB3 Instance



The following is to prove:

Give a bound K such that MB3 instance has a bisection cut of size at most c iff the corresponding graph has 2-modularity at least K.

Existence of such a clustering of two clusters?

Lemma 1

For every graph constructed from a MB3 instance, there exists a clustering $\mathcal{C} = \{C_1, C_2\}$ such that $q(\mathcal{C}) > 0$.

Proof:

- $C_1 = cliq(v)$ for some $v \in V$;
- $C_2 = V \setminus C_1$
- $q(\mathcal{C}) = 1 - \frac{3}{m} - \frac{(n(n-1)+3)^2 + ((n-1)(n(n-1)+3))^2}{4m^2} > 0$ $(n \geq 4)$

$cliq(v)$ are all in one cluster?

Lemma 2

For every node $v \in V$ there exists a cluster $C \in \mathcal{C}^*$ such that $cliq(v) \subseteq C$.

Proof:

By contradiction, assume a clique $cliq(v)$ is split into two clusters:

- $\mathcal{C}' = \{C_1 \setminus cliq(v), C_2 \setminus cliq(v)\}$;
- $\Delta = q(\mathcal{C}') - q(\mathcal{C})$
  - $v \in C_2 \Longrightarrow \Delta \geq 0$
  - $v \in C_1 \Longrightarrow \Delta \geq 0$
- Clique moved into one cluster makes modularity larger.

Same size for each clusters in 2-Modularity Optimum?

Lemma 3

In $\mathcal{C}^*$, each cluster contains exactly $n/2$ complete node cliques.

Proof:

By contradiction, assume a cluster C_1 has $l_1 \leq n/2 - 1$ cliques.

- $q(C^*) = \frac{m'}{m} - \frac{l_1^2(n(n-1)+3)^2}{4m^2} - \frac{(n-l_1)^2(n(n-1)+3)^2}{4m^2}$;
- $\mathcal{C}' :=$ move one complete clique from C_2 to C_1;(lose at most 3 edges covered by inner-modularity)
  $q(C^*) = \frac{m'-3}{m} - \frac{(l_1+1)^2(n(n-1)+3)^2}{4m^2} - \frac{(n-l_1-1)^2(n(n-1)+3)^2}{4m^2}$;
- $q(\mathcal{C}') \geq q(\mathcal{C})$
  (assume n to be even and $l_1 \leq \frac{n}{2} - 1$);

Strongly NP-complete

Strongly NP-completeness

A problem is said to be NP-hard in the strong sense (strongly NP-hard), if it remains so even when all of its numerical parameters are bounded by a polynomial in the length of the input.

Example: Bin Packing VS 0-1 Knapsack Problem

- Bin Packing: find a minimum integer B, such that n given items with relatively size $a_1, a_2, \cdots, a_n$ can be packed into B bins, where each bin has a given size V.
- 0-1 Knapsack Problem: given n items, each with value p_i and weight w_j for $i \in [1, n]$, maximize the total value of items that can be packed into a bag, on condition that the maximum weight carried in the bag is W. $\Longleftrightarrow$ can be solved by dynamic programming in $O(nW)$ time, note that, W is not polynomial of n. Otherwise, this is not NP-hard problem.

Strongly NP-complete

Theorem 1

2-Modularity is strongly NP-complete.

Proof:

- $K = \frac{1}{2} - \frac{c}{m}$;
- $\sum_{v \in C_i} deg(v) = m$;
- if $q(\mathcal{C}^*) \geq K = \frac{1}{2} - \frac{c}{m} = \frac{m-c}{m} - \frac{m^2}{4m^2} - \frac{m^2}{4m^2}$;

  # of inter-cluster edges can be at most c;
- Optimum $\mathcal{C}^* \iff$ balanced partition cutting at most c edges.

Discussion

- 2-Modularity $\leq_p$ k-Modularity?
- Corresponding algorithms?
- at least k clusters?? wired?

Contributions

- Algorithm by greedily merging clusters;
- Approximation ratio is at least 2.

Greedy Algorithm

Algorithm 1

Input: graph $G = (V, E)$

Output: clustering $\mathcal{C}$ of G

$\mathcal{C} \leftarrow$ singletons;

symmetric matrix Δ with $\Delta_{i,j} = q(\mathcal{C}_{i,j}) - q(\mathcal{C})$;

$\triangleright$ $\mathcal{C}_{i,j}$ is by merging clusters i, j;

while $|\mathcal{C}| > 1$ and there exists $\Delta_{i,j} > 0$ **do**

  merge clusters i and j where $\Delta_{i,j}$ is maximum;

  $\triangleright$ arbitrarily select one if multiple maximum exist;

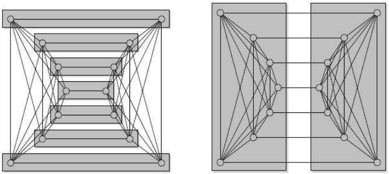
  update matrix Δ;

return clustering with the highest modularity

Worst Case

Theorem 2.1

No finite approximation factor for the greedy algorithm for finding clusterings with maximum modularity.



(a) (b)

- worst case clustering: $q(\mathcal{C}_a) = \frac{2}{n} - \frac{n}{2} \cdot \frac{4n^2}{n^4} = 0$
- one better clustering: $q(\mathcal{C}_b) = \frac{n(n-2)}{n^2} - 2\frac{4n^2}{16n^2} = \frac{1}{2} - \frac{2}{n} \leq opt$

Lower Bound

If re-mapping modularity interval from $[\frac{1}{2}, 1]$ to $[0, 1]$, the greedy algorithm in this instance can get an approximation ratio 2, but not for general case.

$$(q(\mathcal{C}_a) = \frac{1}{3};\ q(\mathcal{C}_b) = \frac{2}{3})$$

Theorem 2.2

The approximation factor of the greedy algorithm is at least 2.

Optimality Results

Deal with two specific structures:

- Complete graph with n nodes – clique;
- Simple cycle with n nodes;

Both these two structures can be abstracted to d-regular graph with $|E| = \frac{d|V|}{2}$

$$\rightarrow q(\mathcal{C}) = \frac{|E(C)|}{dn/2} - \frac{1}{n^2}\sum_{i=1}^{k}|C_i|^2$$

where $\mathcal{C} = \{C_1, \cdots, C_k\}$

n-Clique

Theorem 3

Let k and n be integers, K_{kn} be the complete graph on $k \cdot n$ nodes and $\mathcal{C}$ a clustering such that each cluster contains exactly n elements.

$$q(\mathcal{C}) = (-1 + \frac{1}{k}) \cdot \frac{1}{kn - 1}$$

Observations:

- $k > 1$, $n \to \infty$, $q(\mathcal{C}) \to 0^-$;
- $k = 1$, $q(\mathcal{C}) = 0$ is the global maximum;
- No further things??

n-Cycle (Simple 2-regular cycle with n nodes)

- Define $F(x) = 1 - q(\mathcal{C})$ where $x \in D^{(k)}$ is the vector of size for k clusters; Minimize $F(x)$;
- $F(x) = \frac{k}{n} + \frac{1}{n^2} \sum_{i=1}^{k} x_i^2$ has global minimum at $x^* = \lfloor \frac{n}{k} \rfloor$ or $x^* = \lceil \frac{n}{k} \rceil$, i.e. evening cluster size decreases F.

n-Cycle (Simple 2-regular cycle with n nodes)

Based on these conclusions, we can have:

Lemma 2

Let C_n be a simple cycle with n nodes, $h : [1, \cdots, n] \rightarrow \mathcal{R}$ a function defined as

$$h(x) := x \cdot n + n + \lfloor \frac{n}{x} \rfloor (2n - x \cdot (1 + \lfloor \frac{n}{x} \rfloor))$$

and k^* be the argument of the global minimum of h. Then every clustering of C_n with maximum modularity has k^* clusters.

Key:

$$h(k) = F(x^*)$$

n-Cycle

Theorem 4

Let n be an integer and C_n a simple cycle with n nodes. Then every clustering $\mathcal{C}$ with maximum modularity has k cluster of almost equal size, where

$$k \in [\frac{n}{\sqrt{n+\sqrt{n}}} - 1, \frac{1}{2} + \sqrt{\frac{1}{4} + n}]$$

Proof:

- k is fixed by the distribution of cluster sizes with maximum modularity (even cluster size);
- with k outside this interval, function h is either monotonically increasing or decreasing;
- at most 3 possible values for large $n \Longleftarrow k \in (\frac{n}{\sqrt{n}} - 1, \frac{1}{2} + \sqrt{\frac{1}{4} + n}]$

Any Questions

*http://www.hardcore-stress-management.com/images/ManHoldingQuestionMarkSmallCropped.jpg