Chapter 7 Structural Information

Angsheng Li

Institute of Software
Chinese Academy of Sciences

Advanced Algorithms, U CAS
18th, April, 2016

Outline

1. Backgrounds
2. The challenges
3. Overall ideas
4. Structural information
5. Three-dimensional gene map
6. Resistance
7. Theory

Shannon's information

1949:
Given a distribution $p = (p_1, p_2, \cdots, p_n)$, the Shannon's information is

$$H(p) = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i. \tag{1}$$

p_i is the probability that item i is chosen, $-\log_2 p_i$ is the "self-information" of item i.
This metric and the associated notions of noises form the foundation of information theory.

Shannon's question, 1953

Shannon's information fails to support communication network.
Given a communication network G,

1. (De-structuring) Let p be a distribution computed from G, degree distribution, or distance distribution, and so on.
   This discards the interesting properties of G.
2. Define $H(p)$ to be the information of G.
   This number $H(p)$ does not tell us anything about the interactions and communications occurred in G.

The question is hence:
What is the information embedded in a graph?

Physical systems

Given a physical system G, the information embedded in G should determine and decode the *essential structure* of G. For example, for a car and a boat, the essential structures of the two objects should be different, and the essential structures of a car and a boat should be determined by the information embedded in the car and the boat respectively.

Question: What is the essential structure of a physical system?

Evolving network

Given a network G that is evolved in nature by two mechanisms:

1. The rules, regulations and laws of the objects
2. Perturbations by noises and random variations

In this case, the information embedded in G should determine and decode the structure of G that is formed by the rules, regulations and laws in which the noises and random variations occurred in G are excluded.

Noisy data

Given a structured noisy data G, the information embedded G should determine and decode the structure T of G that excludes the noises occurred in G.

Dynamical complexity of a network

Given a network G, the dynamical complexity of G should be the measure of complexity of the interactions, operations and communications occurring in G. This is different from the static complexity such as the number of nodes, the number of edges etc.

What is the measure of dynamical complexity of a network?

Natural structure and natural rank

In Nature and Society, individuals form natural structures and follow some natural ranking.
This is different from the current-generation search engine based on PageRank.
What is the natural rank?

Hierarchical thesis

- The natural structure of a physical system is a hierarchical structure
- The natural structure of a network evolving in Nature and Society is a hierarchical structure
- The true structure of a structured noisy data is a hierarchical structure

Decoding the truth

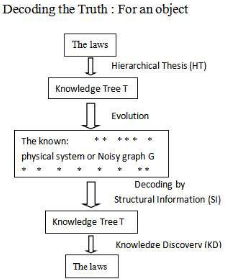

Figure: Decoding the truth by structural information.

Decoding ECC

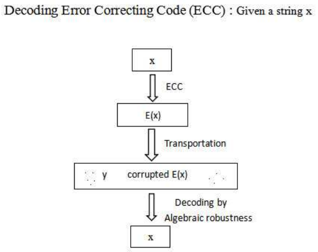

Figure: Decoding error correcting code.

One-Dimensional structural information

Definition

(One-dimensional structural information) Given a connected graph $G = (V, E)$ with n nodes and m edges, for each node $i \in \{1, 2, \cdots, n\}$, let d_i be the degree of i in G, and let $p_i = \frac{d_i}{2m}$. We define the one-dimensional structural information or positioning entropy of G by using the entropy function H as follows:

$$\mathcal{H}^1(G) = H(\mathbf{p}) = H\left(\frac{d_1}{2m}, \ldots, \frac{d_n}{2m}\right) = -\sum_{i=1}^{n} \frac{d_i}{2m} \cdot \log_2 \frac{d_i}{2m}. \quad (2)$$

Intuition of $\mathcal{H}^1(G)$

- The Shannon information for graphs
- It is the number of bits required to determine the code of the node that is accessible from random walk in G

Structural information by partition

Definition

(Structural information of networks by a partition) Given a graph $G = (V, E)$, suppose that $\mathcal{P} = \{X_1, X_2, \cdots, X_L\}$ is a partition of V. We define the *structural information of G by $\mathcal{P}$* as follows:

$$\begin{aligned}\mathcal{H}^{\mathcal{P}}(G) &:= \sum_{j=1}^{L} \frac{V_j}{2m} \cdot H\left(\frac{d_1^{(j)}}{V_j}, \ldots, \frac{d_{n_j}^{(j)}}{V_j}\right) - \sum_{j=1}^{L} \frac{g_j}{2m} \log_2 \frac{V_j}{2m} \\ &= -\sum_{j=1}^{L} \frac{V_j}{2m} \sum_{i=1}^{n_j} \frac{d_i^{(j)}}{V_j} \log_2 \frac{d_i^{(j)}}{V_j} - \sum_{j=1}^{L} \frac{g_j}{2m} \log_2 \frac{V_j}{2m}, \quad (3)\end{aligned}$$

where L is the number of modules in $\mathcal{P}$, n_j is the number of nodes in X_j, $d_i^{(j)}$ is the degree of the i-th node of X_j, V_j is the volume of X_j which is the sum of degrees of nodes in X_j, and g_j is the number of edges with exactly one endpoint in X_j.

Understanding $\mathcal{H}^{\mathcal{P}}(G)$

1. $\frac{V_j}{2m}$: the probability that random walk in G arrives at X_j
2. $-\sum_{i=1}^{n_j} \frac{d_i^{(j)}}{V_j} \log_2 \frac{d_i^{(j)}}{V_j}$: the positioning information in X_j
3. $\frac{g_j}{2m}$: the probability that random walk going into X_j from nodes outside X_j
4. $-\log_2 \frac{V_j}{2m}$: self-information of X_j
5. $\mathcal{H}^{\mathcal{P}}(G)$: the number of bits required to determine the two-dimensional code of the node v that is accessible from random walk

Telephone call

- Local number
- Area codes

Two-dimensional structural information

Definition

(Two-dimensional structural information of networks) Let G be a connected graph.

(1) Define the two-dimensional structural information of G as follows:

$$\mathcal{H}^2(G) = \min_{\mathcal{P}}\{\mathcal{H}^{\mathcal{P}}(G)\}, \tag{4}$$

where $\mathcal{P}$ runs over all the partitions of G.

(2) We say that a partition $\mathcal{P}$ of the vertices of G is a *natural structure* of G, if:

$$\mathcal{H}^{\mathcal{P}}(G) = \mathcal{H}^2(G). \tag{5}$$

Partitioning tree

Definition

(Partitioning tree of graphs) Let $G = (V, E)$ be an undirected and connected network. We define the *partitioning tree* $\mathcal{T}$ *of* G as a tree $\mathcal{T}$ with the following properties:

(1) For the root node denoted λ, we define the set $T_\lambda = V$.

(2) For every node $\alpha \in \mathcal{T}$, the immediate successors of α are $\alpha\hat{}\langle j\rangle$ for j from 1 to a natural number N ordered from left to right as j increases.
Therefore, $\alpha\hat{}\langle i\rangle$ is to the left of $\alpha\hat{}\langle j\rangle$ written as $\alpha\hat{}\langle i\rangle <_{\mathrm{L}} \alpha\hat{}\langle j\rangle$, if and only if $i < j$.

(3) For every $\alpha \in \mathcal{T}$, there is a subset $T_\alpha \subset V$ that is associated with α.
For α and β, we use $\alpha \subset \beta$ to denote that α is an initial segment of β. For every node $\alpha \neq \lambda$, we use α^- to denote the longest initial segment of α, or the longest β such that $\beta \subset \alpha$.

Partitioning tree - II

(4) For every i, $\{T_\alpha \mid h(\alpha) = i\}$ is a partition of V, where $h(\alpha)$ is the height of α (note that the height of the root node λ is 0, and for every node $\alpha \neq \lambda$, $h(\alpha) = h(\alpha^-) + 1$).

(5) For every α, T_α is the union of T_β for all β's such that $\beta^- = \alpha$; thus, $T_\alpha = \cup_{\beta^- = \alpha} T_\beta$.

(6) For every leaf node α of $\mathcal{T}$, T_α is a singleton; thus, T_α contains a single node of V.

Structural information by partitioning tree

Definition

(Structural information of a graph by a partitioning tree) For an undirected and connected network $G = (V, E)$, suppose that $\mathcal{T}$ is a partitioning tree of G. We define the structural information of G by $\mathcal{T}$ as follows:

(1) For every $\alpha \in \mathcal{T}$, if $\alpha \neq \lambda$, then define

$$H^{\mathcal{T}}(G; \alpha) = -\frac{g_\alpha}{2m} \log_2 \frac{V_\alpha}{V_{\alpha^-}}, \tag{6}$$

where g_α is the number of edges from nodes in T_α to nodes outside T_α, V_β is the volume of set T_β, namely, the sum of the degrees of all the nodes in T_β.

Definition

(2) We define the structural information of G by the partitioning tree $\mathcal{T}$ as follows:

$$\mathcal{H}^{\mathcal{T}}(G) = \sum_{\alpha \in \mathcal{T}, \alpha \neq \lambda} H^{\mathcal{T}}(G; \alpha). \tag{7}$$

K-dimensional structural information

Definition

(K-dimensional structural information) Let $G = (V, E)$ be a connected network.

(1) We define the K-dimensional structural information of G as follows:

$$\mathcal{H}^K(G) = \min_{\mathcal{T}}\{\mathcal{H}^{\mathcal{T}}(G)\}, \tag{8}$$

where $\mathcal{T}$ ranges over all of the partitioning trees of G of height K.

(2) Given a K-level partitioning tree $\mathcal{T}$ of G, we say that $\mathcal{T}$ is the K-*dimensional knowledge tree* of G, if:

$$\mathcal{H}^{\mathcal{T}}(G) = \mathcal{H}^K(G). \tag{9}$$

Cell sample network

Suppose that $v_1, v_2, \cdots, v_n$ are n samples of cells and that $g_1, g_2, \cdots, g_N$ are N genes. For every pair (i, j), let $a(i, j)$ be the expression profile of gene g_i in sample v_j. Then, for every j from 1 to n, a vector $(a(1, j), a(2, j). \cdots, a(N, j))$ occurs and represents the gene expression profiles of the sample v_j, denoted P_j. For every pair (j, j'), let $W_{j,j'}$ be the Pearson correlation coefficient between P_j and $P_{j'}$, the gene expression profiles of samples v_j and $v_{j'}$, respectively.
A cell sample network $G = (V, E)$ is constructed on the basis of the gene expression profiles by the following algorithm, denoted $\mathcal{G}$.
Algorithm $\mathcal{G}$ works with a fixed natural number k, and proceeds as follows:

(1) The vertices of G are the cell samples $v_1, v_2, \cdots, v_n$, that is, let $V = \{v_1, v_2, \cdots, v_n\}$; and

(2) For every j, suppose that $u_1, u_2, \cdots, u_k$ are the cell samples such that $W(v_j, u_1), W(v_j, u_2), \cdots, W(v_j, u_k)$ are the highest k weights among the weights $W(v_j, u)$ for all of the samples u, where $W(v_j, u)$ is the Pearson correlation coefficient between the gene expression profiles of samples v_j and u. For every i from 1 to k, create an edge (v_j, u_i) with weight $W(v_j, u_i)$.

This constructs the weighted graph $G = (V, E)$.

Structuring of gene expression profiles

Algorithm $\mathcal{C}$ proceeds as follows:

(1) (Noise amplifying) Fix a *noise amplifier* σ. Let W be the average wight among all the pairs of cell samples. Let $M = \sigma \cdot W$ be the modifier. Let H be the weighted graph of the cell samples such that for every pair (i, j) of cell samples, there is a weight $W'(i, j) = W(i, j) + M$. This step amplifies the noise for all the weights. The roles of this step are two-fold: if the weight $W(i, j)$ between cell samples i and j is nontrivially high, then the modified weight $W'(i, j) = W(i, j) + M$ is approximately the original weight $W(i, j)$ since the modifier M is small, and if the weight $W(i, j)$ is trivial or noisy, then the modified weight $W'(i, j) = W(i, j) + M$ is significantly amplified, which allows our algorithm to better filter the noise or trivial weights from the highly nontrivial weights.

(2) For every k, let H_k be the weighted graph obtained from H as follows:
– The modifier M is kept for every edge.
– For every cell sample i, keep the weighted edges of the top k weights, and delete all the other weights.

(3) For each k, let $H(k)$ be the one-dimensional structure entropy of the weighted graph H_k. We say that k is a *stable point*, if both $H(k-1) > H(k)$ and $H(k+1) > H(k)$ hold.

(4) (Minimisation of non-determinism or uncertainty) Define k to be the k' that achieves the least one-dimensional structure entropy among all the stable points. That is, k is a stable point, and $H(k)$ is the least among the $H(k')$ for all the stable points k'.
This step ensures that the chosen k generates a network structure with minimum uncertainty or non-determinism.

Lymphomas Real

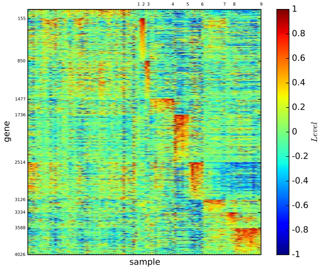

Figure: Gene map of true types of the lymphomas.

Lymphomas: Two-dimensional structural information

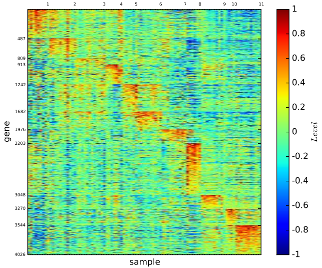

Figure: Gene map of types of the lymphomas found by $\mathcal{E}^2$.

Lymphomas: Three-dimensional structural information

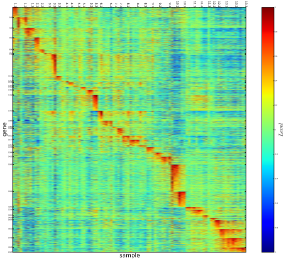

Figure: Gene map of types of the lymphomas found by $\mathcal{E}^3$.

Clinical data analysis

(1) The DLBCL samples in each of the submodules 2.2, 3.1, 3.3, 4.1, 4.3, 5.1, 6.1, 6.2 and 8.1 are similar to one another in survival times, survival indicators and IPI scores.

(2) However, the DLBCL samples in submodules 3.2, 7.1, 7.2, 8.2 and 8.3 are divergent in survival times, survival indicators and IPI scores.

(3) The overall survival times, survival ratios and IPI scores in most of the submodules are distinguishable.

Therefore, many of the submodules of the DLBCL samples identified by $\mathcal{E}^3$ are interpretable by the similarity of survival times, survival indicators and IPI scores for the cell samples within the same submodule, and distinguishable by overall survival times, survival ratios and IPI scores for different submodules.

Resistance

Definition

Given a connected network $G = (V, E)$, let $\mathcal{P}$ be a partition of G. We define the *resistance of G given by* $\mathcal{P}$ as follows:

$$\mathcal{R}^{\mathcal{P}}(G) = -\sum_{j=1}^{L} \frac{V_j - g_j}{2m} \log_2 \frac{V_j}{2m}, \tag{10}$$

where V_j is the volume of the j-th module X_j of $\mathcal{P}$, and g_j is the number of edges from X_j to nodes outside X_j.

In Equation (10), for the j-th term $-\frac{V_j - g_j}{2m} \log_2 \frac{V_j}{2m}$, $\frac{V_j - g_j}{2m} = \frac{V_j - g_j}{V_j} \cdot \frac{V_j}{2m}$ is the probability that a random walk goes to the j-th module X_j and fails to escape from the j-th module X_j, and $-\log_2 \frac{V_j}{2m}$ is the number of bits to determine the code of the j-th module.

Resistance law

For the resistance of graph G by $\mathcal{P}$, we have the following *resistance principle*:
Let $G=(V,E)$ be a connected graph. Suppose that $\mathcal{P}$ is a partition of V with the notations the same as that in the definitions of $\mathcal{H}^1(G)$ and $\mathcal{H}^{\mathcal{P}}(G)$. Then the positioning entropy of G, $\mathcal{H}^1(G)$, and the structure entropy of G by given $\mathcal{P}$, i.e., $\mathcal{H}^{\mathcal{P}}(G)$, satisfy the following properties:

(1) (Additivity of $\mathcal{H}^1(G)$) The positioning entropy of G satisfies:

$$\mathcal{H}^1(G) = -\sum_{j=1}^{L} \frac{V_j}{2m} \sum_{i=1}^{n_j} \frac{d_i^{(j)}}{V_j} \log_2 \frac{d_i^{(j)}}{V_j} - \sum_{j=1}^{L} \frac{V_j}{2m} \log_2 \frac{V_j}{2m}. \tag{11}$$

Resistance law - II

(2) (Local resistance law of networks)

$$\mathcal{R}^{\mathcal{P}}(G) = -\sum_{j=1}^{L} \frac{V_j - g_j}{2m} \log_2 \frac{V_j}{2m} = \mathcal{H}^1(G) - \mathcal{H}^{\mathcal{P}}(G) \quad (12)$$

(3) Assume that for each j, $V_j \leq m$, for $m = |E|$. Then

$$\mathcal{R}^{\mathcal{P}}(G) = -\sum_{j=1}^{L} (1 - \Phi(X_j)) \frac{V_j}{2m} \log_2 \frac{V_j}{2m} = \mathcal{H}^1(G) - \mathcal{H}^{\mathcal{P}}(G) (13)$$

where $\Phi(X_j)$ is the conductance of X_j in G.

Resistance law - III

Now, we are ready to define the *resistance of a graph G* as follows:

$$\mathcal{R}(G) = \max_{\mathcal{P}}\{\mathcal{R}^{\mathcal{P}}(G)\}, \tag{14}$$

where $\mathcal{P}$ runs over all the partitions of G.

By the definition of the resistance of G, the local resistance law in (2) above and the definition of the two-dimensional structure entropy, we have the following:

Global resistance law of networks: For a network G, we have

$$\mathcal{R}(G) = \mathcal{H}^1(G) - \mathcal{H}^2(G). \tag{15}$$

One-dimensional structural information - I

Theorem

(Lower bound of positioning entropy of simple graphs) Let $G = (V, E)$ be an undirected, connected, and simple graph with m edges, i.e., $|E| = m$. Then:

$$\mathcal{H}^1(G) \geq \frac{1}{2}(\log_2 m - 1).$$

One-dimensional structural information - II

Theorem

(Lower bound of positioning entropy of graphs of balanced weights) Let $G = (V, E)$ be a connected graph with weight function w. Let $m = |E|$ be the number of edges. If the ratio of maximum weight and minimum weight is at most m^{ϵ}, that is $\frac{\max_{e\in G}\{w(e)\}}{\min_{e\in G}\{w(e)\}} \leq m^{\epsilon}$, for some constant $\epsilon < 1$, then:

$$\mathcal{H}^1(G) \geq \frac{1}{2}\left[(1-\epsilon)\log_2 m - 1\right].$$

Locality

Theorem

(Locality theorem) Given a connected graph G, let $\mathcal{P}$ be the partition of nodes of G such that each module X of $\mathcal{P}$ contains a single node of V, and let $\mathcal{Q}$ be the partition of G containing only one module of the whole set V. Then, we have

$$\mathcal{H}^{\mathcal{P}}(G) = \mathcal{H}^{\mathcal{Q}}(G). \tag{16}$$

Separation

Theorem

(Separation theorem) Let $G = (V, E)$ be a connected graph. Suppose that $\mathcal{P}$ is a partition of V, and X and Y are two modules of $\mathcal{P}$. Let $Z = X \cup Y$. Let $\mathcal{Q}$ be the partition consisting Z and all the modules of $\mathcal{P}$ other than X and Y. If there is no edge between the nodes in X and the nodes in Y, then, we have:

$$\mathcal{H}^{\mathcal{P}}(G) \leq \mathcal{H}^{\mathcal{Q}}(G). \tag{17}$$

Basic principle

Theorem

(Structural information principle) For any graph G, the structural information of G follows:

$$\mathcal{H}^2(G) \geq \Phi(G) \cdot \mathcal{H}^1(G), \tag{18}$$

where $\Phi(G)$ is the conductance of G, and $\mathcal{H}^1(G)$ is the positioning entropy of G.

Lower bounds - I

Foe simple graphs, we have

Theorem

(Lower bounds of two-dimensional structural information of simple graphs) Let $G = (V, E)$ be an undirected, connected and simple graph with number of edges $|E| = m$. Then the two-dimensional structural information of G satisfies

$$\mathcal{H}^2(G) = \Omega(\log_2 \log_2 m). \tag{19}$$

Lower bounds - II

For the graphs with balanced weights, we have

Theorem

(Lower bound of two-dimensional structural information of graphs with balanced weights) Let $G = (V, E)$ be a connected graph with weight function w. Let $m = |E|$ be the number of edges. If the ratio of maximum weight and minimum weight is at most $\log_2^{\epsilon} m$, that is $\frac{\max_{e \in G}\{w(e)\}}{\min_{e \in G}\{w(e)\}} \leq \log_2^{\epsilon} m$, for some constant $\epsilon < 1$, then the structural information of G satisfies

$$\mathcal{H}^2(G) = \Omega(\log_2 \log_2 m). \tag{20}$$

Trees

Theorem

(Upper bounds of structural information of trees) Let T be a complete binary tree of depth H and thus of size $n = 2^H - 1$. Then the structural information of T satisfies

$$\mathcal{H}^2(T) \leq \log_2 \log_2 n + 4 + o(1). \tag{21}$$

Grids

Theorem

(Upper bound of two-dimensional structural information of grid graphs) Let $G = (V, E)$ be an $n \times n$ grid graph. Then the two-dimensional structural information of G satisfies

$$\mathcal{H}^2(G) \leq 2\log_2 \log_2 n + O(1). \tag{22}$$

Expanders

For expander graphs, we have

Theorem

(Expanders) Let $\{G_n\}$ be a family of expanders, each of which is either a simple graph or a graph with balanced weights on edges. Then for each $G = G_n$, we have that

$$\mathcal{H}^2(G) = \Omega(\log n). \tag{23}$$

New direction: We could define expander by $\mathcal{H}^2(G) = \Omega(\log_2 n)$, giving a new class and an information theoretical characterisation of expanders.

Phase transition in a small world

Theorem

(Phase transition theorem of two-dimensional structural information of networks of the small world model) Let G be a network generated from the small world model with parameter $r \geq 0$. Then the two-dimensional structural information has a sharp phase transition at the point $r = 2$. That is,

(1) if $r \geq 2$, then with probability $1 - o(1)$, $\mathcal{H}^2(G) = O(\log \log n)$;

(2) if $r < 2$, then with probability $1 - o(1)$, $\mathcal{H}^2(G) = \Omega(\log n)$.

New directions More phase transition results are possible.

Black hole - I

Theorem

(Black hole theorem - necessity) Let $G = (V, E)$ be a connected weighted graph of size $n = |V|$ and weight function $w : E \to \mathbb{R}^+$.

(1) *If there is a subset $S \subseteq V$ of size s and volume $\mathrm{vol}(S) = \rho \cdot \mathrm{vol}(G)$ for some $0 < \rho \leq 1$, then both positioning entropy $\mathcal{H}^1(G)$ and structural information $\mathcal{H}^2(G)$ of G are at most $H(1-\rho, \rho) + (1-\rho)\log_2(n-s) + \rho \log_2 s$.*

(2) *If $s = \log^{o(1)} n$ and $\rho \geq 1 - \frac{1}{\log n}$, then $\mathcal{H}^2(G) \leq \mathcal{H}^1(G) = o(\log\log n)$.*

Black hole - II

Theorem

(Black hole theorem - sufficiency) Let $G = (V, E)$ *be a connected graph of size* $n = |V|$ *and volume* $\mathrm{vol}(G)$*. If* $\mathcal{H}^2(G) = o(\log\log n)$*, then we have the following conclusions.*

(1) *If* $\mathcal{H}^1(G) = o(\log n)$*, then there is a subset* $S \subseteq V$ *in* G *whose size is* $n^{o(1)}$ *and whose volume is* $(1 - o(1)) \cdot \mathrm{vol}(G)$*.*

(2) *Otherwise, there is a subset* $S \subseteq V$ *in* G *whose volume is* $\mathrm{vol}(S) \geq \rho \cdot \mathrm{vol}(G)$ *for some constant* $0 < \rho < 1$*, and each node in* S *belongs to a subset of size* $\log^{o(1)} n$ *and conductance* $O(1/\log^{1-o(1)} n)$ *(understood as a black hole, that is,* S *is composed by black holes). For the complement* $\overline{S}$ *of* S*, either its volume is* $o(\mathrm{vol}(G))$*, in which case, the complement of* S *consists of only "tiny dusts" and it is trivial, or there is a subset* $U \subseteq \overline{S}$ *with size* $|U| = n^{o(1)}$*, volume* $\mathrm{vol}(U) = (1 - o(1)) \cdot \mathrm{vol}(\overline{S})$ *and conductance* $\Phi(U) = o(1)$*, in which case,* U *corresponds to a black hole.*

Small community phenomenon - I

Theorem

(Small community phenomenon – necessity) Let $G = (V, E)$ be a connected and balanced graph of size $n = |V|$. Then both (1) and (2) below hold:

(1) *If there is a set of modules A satisfying*

- (i) $vol(A) = (1 - o(1)) \cdot vol(G)$, *where* $vol(A)$ *is the sum of the weighted degrees of all the nodes in the modules in A;*
- (ii) *For each module $X \in A$, its size $|X| = n^{o(1)}$;*
- (iii) *For each module $X \in A$, its conductance $\Phi(X) = o(1)$,*

then the two-dimensional structural information of G is $\mathcal{H}^2(G) = o(\log n)$.

(2) *If there is a set of modules A satisfying*

- (i) $vol(A) = \left(1 - O\left(\frac{\log\log n}{\log n}\right)\right) \cdot vol(G)$;
- (ii) *For each module $X \in A$, $|X| = \log^{O(1)} n$;*
- (iii) *For each module $X \in A$, $\Phi(X) = O\left(\frac{\log\log n}{\log n}\right)$,*

then $\mathcal{H}^2(G) = O(\log\log n)$.

Small community phenomenon - II

Theorem

(Small community phenomenon – sufficiency) Let $G = (V, E)$ be a graph of number of edges $m = |E|$ and volume $\mathrm{vol}(G)$ without isolated nodes. Let $w : E \to \mathbb{R}^+$ be the weight function satisfying $\frac{\max_{e \in G}\{w(e)\}}{\min_{e \in G}\{w(e)\}} \leq W$, for some constant $W \geq 1$. If $\mathcal{H}^2(G) \leq c \log_2 \log_2 m$ for some constant $0 < c \leq 1$ and sufficiently large m, then for any $\varepsilon > 0$, and sufficiently large m, there is a set of modules of nodes, denoted by A, satisfying

(1) $vol(A) \geq (1 - 2\varepsilon) \cdot \mathrm{vol}(G)$;

(2) For each module $X \in A$, $|X| \leq \log^{3c/\varepsilon} m$;

(3) For each module $X \in A$, $\Phi(X) \leq 2\varepsilon/(1 - \varepsilon)$.

New directions

1. Algorithmic theory of structural information
2. Foundations for communication networks
3. Foundations for knowledge discovering
4. Theory of data processing
5. Structures and algorithms for big data
6. Natural ranking and smart searching

Great thesis

Structural information minimisation is the mathematical measure of natural selection.

References

1. A. Li, Y. Pan, Structural Information and Dynamical Complexity of Networks, To appear.
2. A, Li, X. Yin and Y. Pan, Three-dimensional gene map of cancer cell types: Structural entropy minimisation principle for defining tumour subtypes. Scientific Reports, **6**: 20412 (2016).
3. Brooks, F. P., Jr. Three great challenges for half-century-old computer science. Journal of the ACM, **50** (1), pp 25 - 26 (2003).