Community Detection

Community Structure: $V(G)$ can be partitioned into $A_1, \dots, A_\ell$ where each A_i is "densely internally connected." Informally, $G[A_i]$ has more edges than we might expect.

Assumptions: $A_1 \dots A_\ell$ is a partition of $V(G)$. Communities do not overlap.

Some measures:

Let C be a subset of $V(G)$

For $v \in C$, $\deg^{int}(v) = \deg_{G[C]}(v) = |N_G(v) \cap C|$

$$\deg^{ext}(v) = \deg_G(v) - \deg^{int}(v) = |N_G(v) \setminus C|.$$

Strong Community: For each $v \in C$, $\deg^{int}(v) > \deg^{ext}(v)$.

Weak Community: $\sum_{v \in C} \deg^{int}(v) > \sum_{v \in C} \deg^{ext}(v)$

Drawbacks: Ignores the size of C.

- Data scientists (millions) versus Graph Product Structure Theory Researchers (10-20).

Community: For each $v \in C$

$$\frac{\deg^{int}(v)}{|C|} > \frac{\deg^{ext}(v)}{|V \setminus C|}$$

[I have lots (100) of friends.

- I am friends with all 20 graph product structure researchers in the world.
- Most of my friends are not studying graph product structure.

$|C| = 20$.]

$$\frac{20}{20} > \frac{80}{8 \text{ billion}}$$

Ranking Community Members.

$A_1, \dots, A_\ell$ is a partition of $V(G)$

Normalized-within-module degree:

$$z(v) = \frac{\deg^{int}(v) - \mu(v)}{\sigma(v)}.$$

] average internal degree of members of v's community

$$\sum_{w \in A_i} \deg^{int}(w) \Big/ |A_i|$$

z-test: how many standard deviations is v away from the average?

standard deviation of $\left(\deg^{int}(w)\right)_{w \in A_i}$

Participation Coefficient

$$p(v) = 1 - \sum_{i=1}^{\ell} \left(\frac{|N_G(v) \cap A_i|}{\deg_G(v)} \right)^2 .$$

$p(v) = 0$ if and only if $N_G(v) \subseteq A_i$ for one $i \in \{1, \dots, \ell\}$.

$p(v) = 1 - \frac{1}{\ell}$ if $N_G(v)$ is evenly distributed over $A_1 \dots A_\ell$.

$$p(v) = 1 - \sum_{i=1}^{\ell} \left(\frac{\deg(v)/\ell}{\deg(v)} \right)^2 = 1 - \ell \cdot \left(\frac{1}{\ell}\right)^2 = 1 - \frac{1}{\ell} .$$

$p(v) \approx 0$: v fits in to one or more communities.

$p(v) \approx 1$: v doesn't fit particularlly well in any community.

In economics, $1 - p(v)$ is called the Herfindahl-Hirschman index

ecology : Simpon diversity index

physics : Inverse participation ratio.

politics : Effective number of parties.

Anomaly Score:

$$cd(v) = \frac{deg(v)}{deg^{int}(v)}$$

requires $deg^{int}(v)$

how attached is v to it's community?

$cd(v) = 1$ iff $\exists N_G(v) \subseteq A_i$ ~~where~~ $v \in A_i$.

Community Association Strength

$$cas(v) = \frac{|N_G(v) \cap A_i|}{deg_G(v)} - \boxed{\frac{\sum_{w \in A_i} deg_G(w) - deg_G(v)}{2m}}.$$

Expected # edges from v to A_i in the Chung-Lu model.

Community Distribution Distance.

$$cdd(v) = \left(\sum_{i=1}^{\ell} \left(\frac{|N_G(v) \cap A_i|}{deg_G(v)} - \frac{\sum_{w \in A_i} deg(w)}{2m} \right)^2 \right)^{\frac{1}{2}}$$

Normalized Euclidean distance between $\left(|N_G(v) \cap A_i|\right)_{i \in [\ell]}$ and Chung-Lu model

Community Finding.

Community finding is hard. The number of partitions of $V(G)$ ~~into 2 parts~~ is huge.

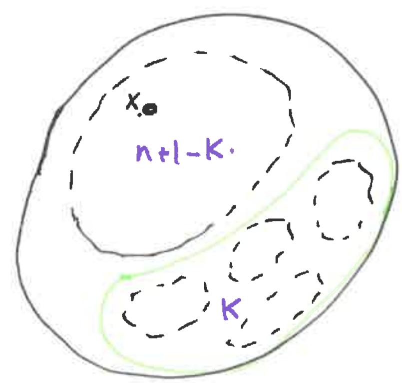
- There are 2^n-1 ways to partition $V(G)$ into two non-empty parts.
- There are $3^n - 3\cdot(2^{n}-1) - 3$ ways to partition $V(G)$ into three parts

⋮

If B_n is the number of partitions into n non-empty parts, then $B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k$

Fix x.

[one part of size $n+1-k$ contains x.
the remaining k elements are partitioned in one of B_k ways]



$$B_n = \left(\frac{n}{(e+o(1))\ln n} \right)^n$$

For almost any reasonable measure of goodness, finding an optimal partition $A_1 \ldots A_\ell$ is (at least) NP-hard. Even if we fix $\ell=2$.

Evaluating Quality of Clustering.

Ground Truth: We are given the "true" community structure.

E.g. We know which group each member of a social network belongs to.

In this case, we measure distance between $A_1 \dots A_\ell$ and the ground truth.

Let $U_1, \dots, U_u$ and $W_1, \dots, W_w$ be two partitions of $V(G)$.

$$P_U(i) = \frac{|U_i|}{n} \qquad P_W(j) = \frac{|W_j|}{n} \qquad P_{UW}(i,j) = \frac{|U_i \cap W_j|}{n} = \frac{n_{ij}}{n}.$$

$\Pr(\text{random } v \text{ is in } U_i)$ $\quad$ $\Pr(\text{random } v \in W_j)$ $\quad$ $\Pr(\text{random } v \in U_i \cap W_j)$.

Mutual Information
$$MI(U,W) = \sum_{i \in [u],\, j \in [w] : P_{UW}(i,j) > 0} P_{UW}(i,j) \cdot \log_2\left(\frac{P_{UW}(i,j)}{P_U(i) \cdot P_W(j)}\right)$$

Measures how much information knowing U gives us about W.

~~If~~ ~~$\Pr(v \in U_i$~~ If the events $v \in U_i$ and $v \in W_j$ are independent, then $P_{UW}(i,j) = P_U(i) \cdot P_W(j)$, $\log \frac{P_{UW}(i,j)}{P_U(i) \cdot P_U(j)} = \log 1 = 0$

So $MI(A,W)$ tells us how much information A gives about the "ground truth" W

If $U_i = W_i$ for all i, then

$$MI(U,W) = \sum_{i,j} \underbrace{P_{UW}(i,j)}_{\text{zero when } i \neq j} \cdot \log \frac{P_{UW}(i,j)}{P_U(i) \cdot P_W(j)} = \overbrace{\underbrace{\sum_i P_U(i) \cdot \log \frac{1}{P_U(i)}}_{\text{Entropy of the distribution } P_U(1), \dots, P_U(n)}}^{H(U)}$$

In this case U gives all the information about W.

Normalized Mutual Information

$$NMI(U,W) = \frac{MI(U,W)}{(H(U) + H(W))/2}$$

Just like $MI(U,W)$, but is always between 0 and 1.

0 : U and W are "independent"

1: $U = W$

Graph Modularity.

- MI, RI, AMI are useful when you have some ground truth.

Recall the Chung-Lu Model.

- Degree sequence $\underline{d} = d_1, \dots, d_n$ of a graph. $2m = \sum_{i=1}^{n} d_i$

$$\Pr(ij \in E(G_{\underline{d}})) = \begin{cases} \dfrac{d_i \cdot d_j}{2m} & \text{if } i \neq j. \\ \dfrac{d_i \cdot d_j}{4m} & \text{if } i = j. \end{cases}$$

Modularity Function: Take $\underline{d} = (\deg_G(v_1), \deg(v_2), \dots, \deg_G(v_n))$

$\mathcal{A} = A_1, \dots, A_\ell$.

$$q_G(\mathcal{A}) = \frac{1}{m} \sum_{i=1}^{\ell} |E(G[A_i])| - \mathbb{E}(E(G_{\underline{d}}[A_i])).$$

We can compute $\mathbb{E}(E(G_d[A_i]))$ easily.

$$\mathbb{E}(E(G_d[A_i])) = \sum_{a \in A_i} \sum_{b \in A_i \setminus \{a\}} \frac{d_a \cdot d_b}{2m} + \sum_{a \in A_i} \frac{d_a^2}{4m} =$$

$$= \frac{1}{4m} \left(\sum_{v \in A_i} \deg_G(v) \right)^2 = m \cdot \left(\frac{\sum_{v \in A_i} \deg_G(v)}{2m} \right)^2$$