

Redes complexas – MO804 (MC908)

Modularidade e Deteccão de Comunidades

Prof. Dr. Ruben Interian

Instituto de Computação, UNICAMP

Resumo

- 1 Revisão do conteúdo
- 2 Modelo de configuração
- 3 Modularidade
- 4 Detecção de comunidades
- 5 Aplicações

Resumo

- 1 Revisão do conteúdo
- 2 Modelo de configuração
- 3 Modularidade
- 4 Detecção de comunidades
- 5 Aplicações

Revisão do conteúdo

Homofilia – tendência dos vértices de ter vínculos com outros vértices semelhantes.

Revisão do conteúdo

Homofilia – tendência dos vértices de ter vínculos com outros vértices semelhantes.

O que é “semelhantes”?

- Representação **mais simples**: dois vértices são “semelhantes” se pertencem ao mesmo grupo.
- Os vértices possuem **atributos numéricos**. Podemos usar a “similaridade” entre os atributos dos vértices.

Revisão do conteúdo

Homofilia – tendência dos vértices de ter vínculos com outros vértices semelhantes.

O que é “semelhantes”?

- Representação **mais simples**: dois vértices são “semelhantes” se pertencem ao mesmo grupo.
- Os vértices possuem **atributos numéricos**. Podemos usar a “similaridade” entre os atributos dos vértices.

Como avaliar a homofilia em uma rede real?

- Para **um vértice**?
- Para **a rede inteira**?

Revisão do conteúdo

Como avaliar a **homofilia de um vértice**?

- Seja $G = (V, E)$ um grafo, e seja $A = \{A_1, \dots, A_k\}$ uma coleção de k grupos de nós que formam uma partição do conjunto V .
- A homofilia de um nó v pode ser definida como a razão $h(v) = \frac{d_i(v)}{d(v)}$ entre o número de vizinhos de v que estão no seu grupo, e o grau de v .
- Se os grupos não influenciam na estrutura de arestas, então $h(v) \approx \frac{|A_i|}{|V|} = w_i$ é o valor **esperado** para $h(v)$.
- Qual deve ser a **diferença** entre $h(v)$ e w_i para que possamos afirmar que há homofilia? – Analisamos a significância estatística de ter $d_i(v)$ adjacentes do mesmo grupo dentre os $d(v)$ adjacentes.

Revisão do conteúdo

Como avaliar a **homofilia da rede inteira**?

Atributos categóricos:

- Avaliar a **diferença** entre o **número de arestas entre vértices de diferentes grupos**, e o **valor esperado** do número de arestas entre vértices de diferentes grupos em uma rede sem homofilia.
- Se uma aresta é colocada **aleatoriamente** no grafo, então a chance de:
 - Escolher dois vértices de X é w_x^2 , e escolher dois vértices de Y é w_y^2 .
 - Escolher vértices de grupos diferentes é $2w_x w_y$.
- Se a proporção de arestas intra-grupo é significativamente **maior** do que $w_x^2 + w_y^2$, então **há evidência** de homofilia.

Revisão do conteúdo

Como avaliar a **homofilia da rede inteira**?

Atributos numéricos:

- Usar o **coeficiente de assortatividade** numérico.

$$r = \frac{\frac{1}{m} \sum_{(u,v)} (x_u - \mu)(x_v - \mu)}{\sigma^2}.$$

- O coeficiente de assortatividade calcula o **coeficiente de correlação de Pearson** dos **valores associados** aos **dois extremos** das arestas do grafo.
- O valor de r varia de -1 a 1: um valor próximo de 1 indica **homofilia**. O valor zero indica que não há homofilia. Um valor próximo de -1 indica **heterofilia**.

Revisão do conteúdo

Pendências. No caso de **atributos categóricos**:

Pergunta 1: A diferença entre os números real e esperado de arestas intra-grupo na rede é **significativa, suficiente** para afirmar que há homofilia?

Pergunta 2: Podemos criar um **indicador simples** (um número) para a homofilia de uma rede?

Pergunta 3: O que fazer se temos **mais de 2 grupos**?

Resumo

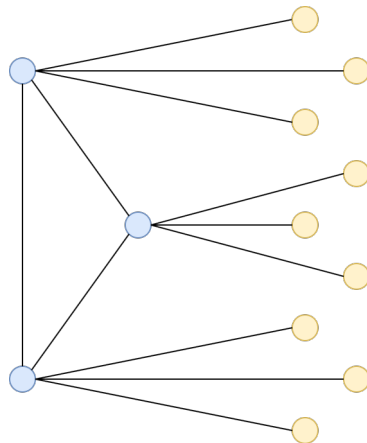
- 1 Revisão do conteúdo
- 2 Modelo de configuração**
- 3 Modularidade
- 4 Detecção de comunidades
- 5 Aplicações

Homofilia *revisited*

Revisitando a homofilia:

- Número de arestas dentro / entre os grupos:
3 arestas no grupo azul, 9 entre os grupos.

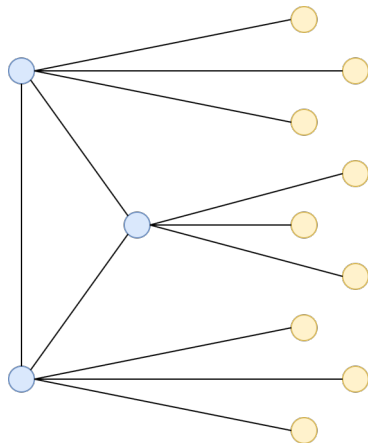
:



Homofilia *revisited*

Revisitando a homofilia:

- Número de arestas dentro / entre os grupos:
3 arestas no grupo azul, 9 entre os grupos.
- **Número esperado de arestas no grupo azul se não houvesse homofilia (rede aleatória):**



Homofilia *revisited*

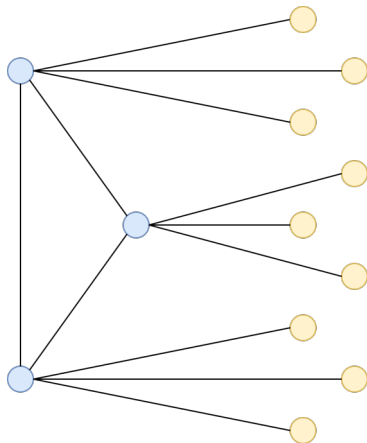
Revisitando a homofilia:

- Número de arestas dentro / entre os grupos:
3 arestas no grupo azul, 9 entre os grupos.

- **Número esperado de arestas no grupo azul se não houvesse homofilia (rede aleatória):**

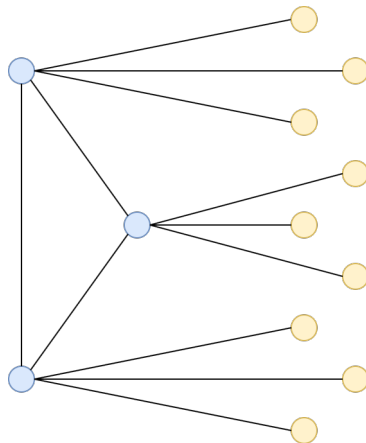
$$w_{\text{azul}}^2 \approx \left(\frac{1}{4}\right)^2 = \frac{1}{16}$$

⇒ o esperado é $\frac{12}{16} = \frac{3}{4}$ arestas no grupo azul.



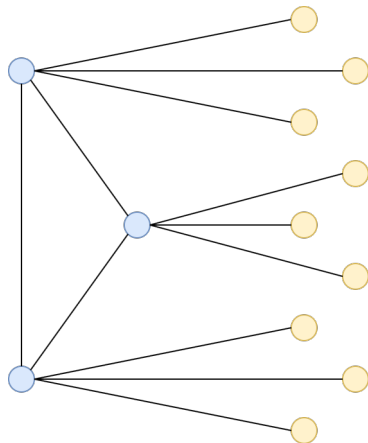
Homofilia *revisited*

- O esperado, em um grafo aleatório, é $\frac{3}{4}$ de aresta no grupo azul.
- Mas, se temos **dois vértices de grau 5** em um grafo com 12 vértices, **qual é a chance deles estarem conectados?**...



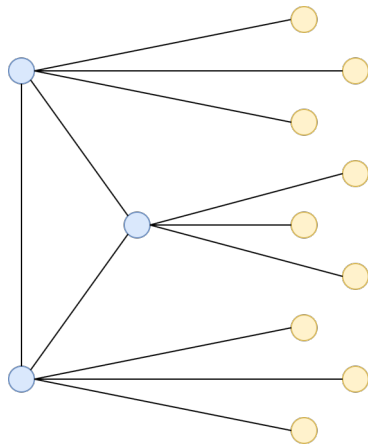
Homofilia *revisited*

- O esperado, em um grafo aleatório, é $\frac{3}{4}$ de aresta no grupo azul.
- Mas, se temos **dois vértices de grau 5** em um grafo com 12 vértices, **qual é a chance deles estarem conectados?**...
- A chance deles **não estarem conectados** é **pequena**. Entre 2 vértices azuis, **é bem mais provável a aresta existir do que não existir!**

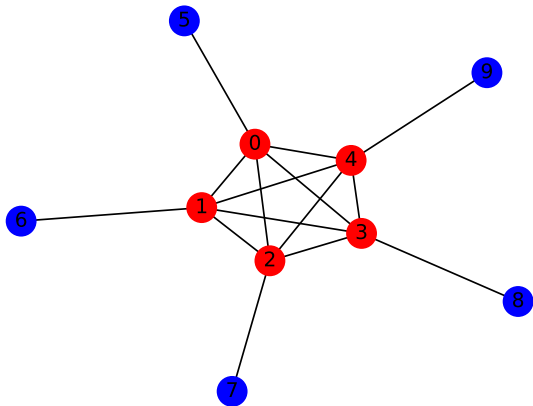


Homofilia *revisited*

- O esperado, em um grafo aleatório, é $\frac{3}{4}$ de aresta no grupo azul.
- Mas, se temos **dois vértices de grau 5** em um grafo com 12 vértices, **qual é a chance deles estarem conectados?**...
- A chance deles **não estarem conectados** é **pequena**. Entre 2 vértices azuis, **é bem mais provável a aresta existir do que não existir!**
- Vértices de grau maior e menor precisam de uma **análise diferenciada**.
Precisamos considerar isso no nosso modelo!



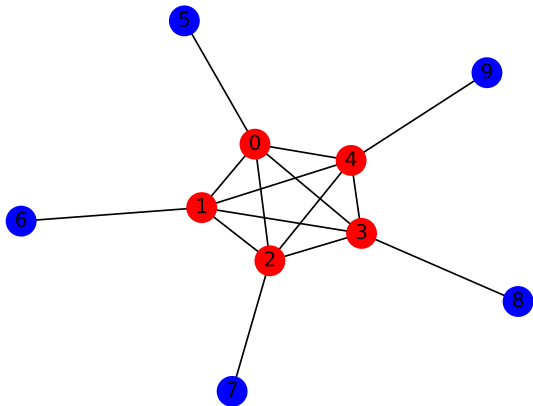
Homofilia *revisited*



Nesta rede:

- Número de arestas dentro dos grupos:

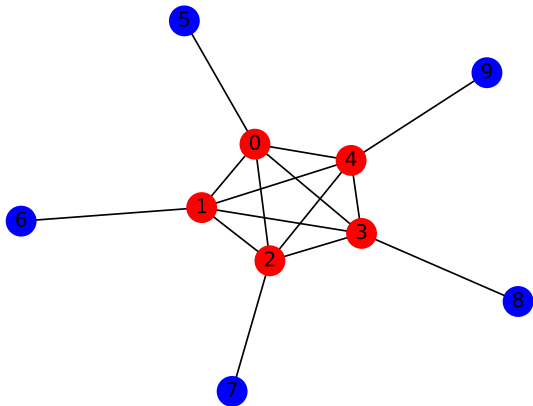
Homofilia *revisited*



Nesta rede:

- Número de arestas dentro dos grupos: **10**.
- Número esperado de arestas dentro dos grupos:

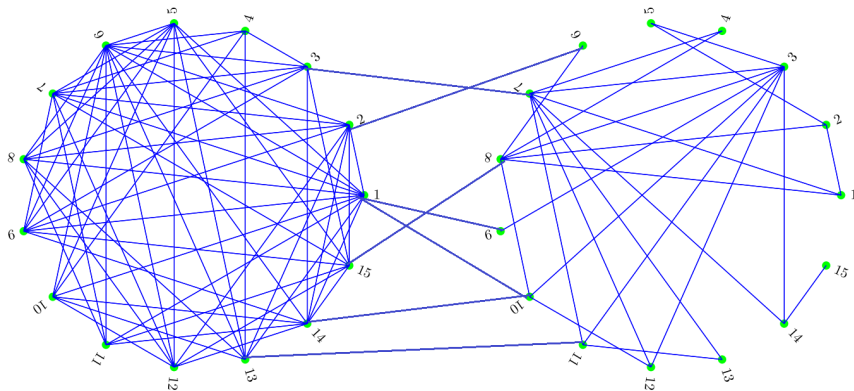
Homofilia *revisited*



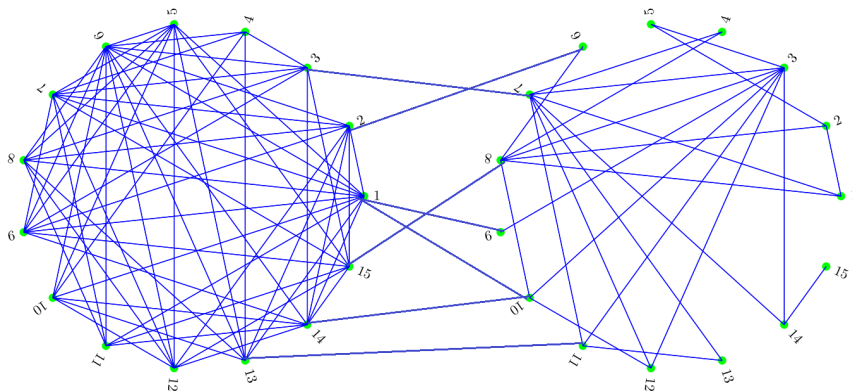
Nesta rede:

- Número de arestas dentro dos grupos: 10.
- Número esperado de arestas dentro dos grupos: $\approx 7,5$ (6,7).
- Na verdade, este tipo de estrutura é esperada, se considerados os graus. . .

Homofilia *revisited*



Homofilia *revisited*



Não parece correto aleatorizar todas as arestas!

Homofilia *revisited*

Objetivo: Criar um **indicador numérico simples** para a homofilia de uma rede, que contempla o caso de ter **um número arbitrário de grupos**, e que considere **os diferentes graus dos vértices** dessa rede.

Para isso, vamos estudar o **modelo de configuração**.

Modelo de configuração

Modelo de configuração:

O que é uma **rede aleatória** na qual os **graus** dos vértices estão previamente definidos?

- A **distribuição dos graus** é fixa. Cada vértice v_i na rede tem grau $d(v_i) = k_i$.

Temos: $\sum_i k_i = 2|E| = 2m$.

Modelo de configuração

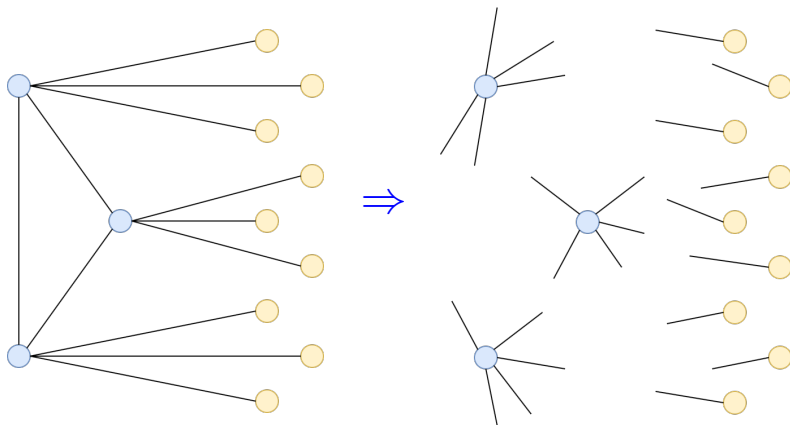
Modelo de configuração:

O que é uma **rede aleatória** na qual os **graus** dos vértices estão previamente definidos?

- A **distribuição dos graus** é fixa. Cada vértice v_i na rede tem grau $d(v_i) = k_i$.
Temos: $\sum_i k_i = 2|E| = 2m$.
- Vamos a associar a cada vértice v_i uma quantidade k_i de **objetos** chamados *stubs*, ou meia-arestas (“metades de arestas”). Haverá $2m$ stubs ao todo.

Modelo de configuração

Associamos a cada vértice v_i uma quantidade k_i de objetos chamados stubs.



Modelo de configuração

Procedimento:

- Escolher dois **stubs**, dentre os $2m$, e conectá-los, criando uma aresta.
- Depois, escolher **outro par** dentre os $2m - 2$ **stubs** restantes.
- ... Assim até todos os **stubs** acabarem.

Resultado:

Modelo de configuração

Procedimento:

- Escolher dois **stubs**, dentre os $2m$, e conectá-los, criando uma aresta.
- Depois, escolher **outro par** dentre os $2m - 2$ **stubs** restantes.
- ... Assim até todos os **stubs** acabarem.

Resultado: um grafo no qual o **conjunto de arestas** é gerado a partir de um **emparelhamento** dos *stubs*, há m arestas, e cada vértice v_i possui grau k_i .

Modelo de configuração

Procedimento:

- Escolher dois **stubs**, dentre os $2m$, e conectá-los, criando uma aresta.
- Depois, escolher **outro par** dentre os $2m - 2$ **stubs** restantes.
- ... Assim até todos os **stubs** acabarem.

Resultado: um grafo no qual o **conjunto de arestas** é gerado a partir de um **emparelhamento** dos *stubs*, há m arestas, e cada vértice v_i possui grau k_i .

Em particular, poderíamos **fazer as escolhas aleatoriamente**:

- Cada sequência de escolhas aleatórias gera um grafo.
- Cada *stub* tem a mesma probabilidade de se conectar a qualquer outro.

Modelo de configuração

O **modelo de configuração de um grafo** é o **espaço** (conjunto) **de todos os emparelhamentos** possíveis entre os *stubs*. Cada emparelhamento define um conjunto de arestas E , e um grafo $G = (V, E)$.

Modelo de configuração

O **modelo de configuração de um grafo** é o **espaço** (conjunto) **de todos os emparelhamentos** possíveis entre os *stubs*. Cada emparelhamento define um conjunto de arestas E , e um grafo $G = (V, E)$.

Observações:

- A rede gerada dessa forma pode conter um **laço** ou uma **aresta múltipla**!

Modelo de configuração

O **modelo de configuração de um grafo** é o **espaço** (conjunto) **de todos os emparelhamentos** possíveis entre os *stubs*. Cada emparelhamento define um conjunto de arestas E , e um grafo $G = (V, E)$.

Observações:

- A rede gerada dessa forma pode conter um **laço** ou uma **aresta múltipla**!
- Evitar laços e arestas múltiplas leva a problemas!

Modelo de configuração

O **modelo de configuração de um grafo** é o **espaço** (conjunto) **de todos os emparelhamentos** possíveis entre os *stubs*. Cada emparelhamento define um conjunto de arestas E , e um grafo $G = (V, E)$.

Observações:

- A rede gerada dessa forma pode conter um **laço** ou uma **aresta múltipla**!
- Evitar laços e arestas múltiplas leva a problemas!
- Felizmente, a densidade de laços e arestas múltiplas geradas **tende a zero** quando o número de nós na rede $n \rightarrow \infty$. Portanto, podemos negligenciar este fato: consideraremos **apenas os grafos simples** que um modelo de configuração define.

Modelo de configuração

Probabilidade de uma aresta entre dois vértices v_i e v_j no modelo de configuração:

- **Chance** de um stub específico se conectar com outro stub específico: $\frac{1}{2m-1}$.

Modelo de configuração

Probabilidade de uma aresta entre dois vértices v_i e v_j no modelo de configuração:

- **Chance** de um stub específico se conectar com outro stub específico: $\frac{1}{2m-1}$.
- **Chance** de um stub de v_i se conectar com um dos stubs de v_j : $\frac{k_j}{2m-1}$.

Modelo de configuração

Probabilidade de uma aresta entre dois vértices v_i e v_j no modelo de configuração:

- **Chance** de um stub específico se conectar com outro stub específico: $\frac{1}{2m-1}$.
- **Chance** de um stub de v_i se conectar com um dos stubs de v_j : $\frac{k_j}{2m-1}$.
- **Chance** de um vértice v_i com k_i stubs se conectar com o vértice v_j : $\frac{k_i k_j}{2m-1}$.

Modelo de configuração

Probabilidade de uma aresta entre dois vértices v_i e v_j no modelo de configuração:

- **Chance** de um stub específico se conectar com outro stub específico: $\frac{1}{2m-1}$.
- **Chance** de um stub de v_i se conectar com um dos stubs de v_j : $\frac{k_j}{2m-1}$.
- **Chance** de um vértice v_i com k_i stubs se conectar com o vértice v_j : $\frac{k_i k_j}{2m-1}$.
- **No limite** (se m é grande), podemos ignorar o valor -1 do denominador:

$$p_{ij} = \frac{k_i k_j}{2m}.$$

Resumo

- 1 Revisão do conteúdo
- 2 Modelo de configuração
- 3 Modularidade**
- 4 Detecção de comunidades
- 5 Aplicações

Modularidade

A **modularidade** é uma forma estatisticamente mais precisa de avaliar a **homofilia de uma rede**.

- A modularidade é baseada no **modelo de configuração**.
- É um **indicador numérico** para a homofilia de uma rede.
- Não assume que há um número específico de grupos.

Modularidade

- Seja $A = \{A_1, \dots, A_k\}$ uma coleção de k grupos de nós, e s_i é o grupo do nó v_i .
(A é uma partição do conjunto de vértices V .)

Modularidade

- Seja $A = \{A_1, \dots, A_k\}$ uma coleção de k grupos de nós, e s_i é o grupo do nó v_i . (A é uma partição do conjunto de vértices V .)
- No modelo de configuração, a chance de uma aresta entre v_i e v_j existir é: $\frac{k_i k_j}{2m}$.
Qual é o valor esperado do **número total** de arestas entre vértices que pertencem ao **mesmo grupo**?

Modularidade

- Seja $A = \{A_1, \dots, A_k\}$ uma coleção de k grupos de nós, e s_i é o grupo do nó v_i . (A é uma partição do conjunto de vértices V .)
- No modelo de configuração, a chance de uma aresta entre v_i e v_j existir é: $\frac{k_i k_j}{2m}$. Qual é o valor esperado do **número total** de arestas entre vértices que pertencem ao **mesmo grupo**?

$$\frac{1}{2} \sum_{ij} \frac{k_i k_j}{2m} \delta(s_i, s_j),$$

onde $\delta(s_i, s_j) = 1$ se $s_i = s_j$, e $\delta(s_i, s_j) = 0$ se $s_i \neq s_j$ (delta de Kronecker).

Modularidade

- **Número esperado** de arestas entre vértices no mesmo grupo:

$$\frac{1}{2} \sum_{ij} \frac{k_i k_j}{2m} \delta(s_i, s_j).$$

Modularidade

- **Número esperado** de arestas entre vértices no mesmo grupo:

$$\frac{1}{2} \sum_{ij} \frac{k_i k_j}{2m} \delta(s_i, s_j).$$

- **Número real** de arestas entre vértices no mesmo grupo?

Modularidade

- **Número esperado** de arestas entre vértices no mesmo grupo:

$$\frac{1}{2} \sum_{ij} \frac{k_i k_j}{2m} \delta(s_i, s_j).$$

- **Número real** de arestas entre vértices no mesmo grupo?

$$\sum_{(v_i, v_j) \in E} \delta(s_i, s_j) = \frac{1}{2} \sum_{ij} a_{ij} \delta(s_i, s_j),$$

onde $a_{ij} = 1$ se há aresta entre v_i e v_j .

Modularidade

Diferença entre o **número real** e **número esperado** de arestas dentro dos grupos:

$$\frac{1}{2} \sum_{ij} a_{ij} \delta(s_i, s_j) - \frac{1}{2} \sum_{ij} \frac{k_i k_j}{2m} \delta(s_i, s_j) = \frac{1}{2} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(s_i, s_j).$$

Modularidade

Diferença entre o **número real** e **número esperado** de arestas dentro dos grupos:

$$\frac{1}{2} \sum_{ij} a_{ij} \delta(s_i, s_j) - \frac{1}{2} \sum_{ij} \frac{k_i k_j}{2m} \delta(s_i, s_j) = \frac{1}{2} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(s_i, s_j).$$

Normalizando (dividindo pelo número de arestas m):

$$Q = \frac{1}{2m} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(s_i, s_j).$$

Modularidade

Diferença entre o **número real** e **número esperado** de arestas dentro dos grupos:

$$\frac{1}{2} \sum_{ij} a_{ij} \delta(s_i, s_j) - \frac{1}{2} \sum_{ij} \frac{k_i k_j}{2m} \delta(s_i, s_j) = \frac{1}{2} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(s_i, s_j).$$

Normalizando (dividindo pelo número de arestas m):

$$Q = \frac{1}{2m} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(s_i, s_j).$$

O valor Q é chamado de **modularidade**.

Modularidade

$$Q = \frac{1}{2m} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(s_i, s_j)$$

A **modularidade** Q avalia, dado um grafo G e uma partição A dos vértices, o número real de arestas intra-grupo existentes.

A modularidade **subtrai** do número real de arestas intra-grupo, o número esperado dessas arestas em uma rede com os **mesmos vértices**, os **mesmos grupos** e os **mesmos graus** dos vértices, mas com as arestas colocadas aleatoriamente.

Modularidade

$$Q = \frac{1}{2m} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(s_i, s_j)$$

A **modularidade** Q avalia, dado um grafo G e uma partição A dos vértices, o número real de arestas intra-grupo existentes.

A modularidade **subtrai** do número real de arestas intra-grupo, o número esperado dessas arestas em uma rede com os **mesmos vértices**, os **mesmos grupos** e os **mesmos graus** dos vértices, mas com as arestas colocadas aleatoriamente.

A **modularidade** foi proposta por **Mark Newman**, da Universidade de Michigan, autor de vários trabalhos de grande impacto no século XXI na área de redes complexas.

Modularidade

Resumo da fórmula:

$$Q = \frac{1}{2m} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(s_i, s_j), \text{ onde}$$

- $m = |E| > 0$ é o número de arestas na rede;
- $a_{ij} = 1$ se há aresta entre v_i e v_j , senão 0 ;
- $k_i = d(v_i)$ é o grau do vértice $v_i \in V$;
- $\delta(s_i, s_j) = 1$ se v_i e v_j estão no mesmo grupo, senão 0 .

Modularidade

Resumo da fórmula:

$$Q = \frac{1}{2m} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(s_i, s_j), \text{ onde}$$

- $m = |E| > 0$ é o número de arestas na rede;
- $a_{ij} = 1$ se há aresta entre v_i e v_j , senão 0 ;
- $k_i = d(v_i)$ é o grau do vértice $v_i \in V$;
- $\delta(s_i, s_j) = 1$ se v_i e v_j estão no mesmo grupo, senão 0 .

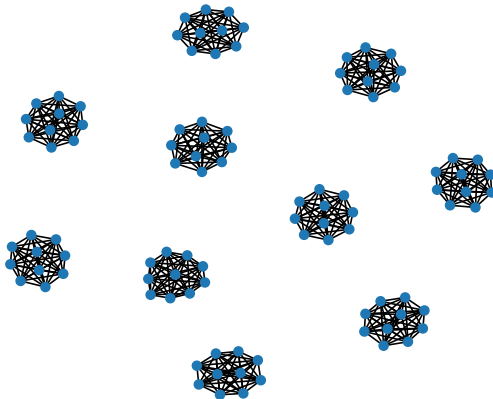
Os valores da modularidade estão no intervalo $[-\frac{1}{2}, 1)$.

Modularidade

Pergunta: Para qual rede e partição dos nós o valor da modularidade pode ser ≈ 1 ?

Modularidade

Pergunta: Para qual rede e partição dos nós o valor da modularidade pode ser ≈ 1 ?



Modularidade

Pergunta: Para qual rede e partição dos nós o valor da modularidade pode ser ≈ 0 ?

Modularidade

Pergunta: Para qual rede e partição dos nós o valor da modularidade pode ser ≈ 0 ?

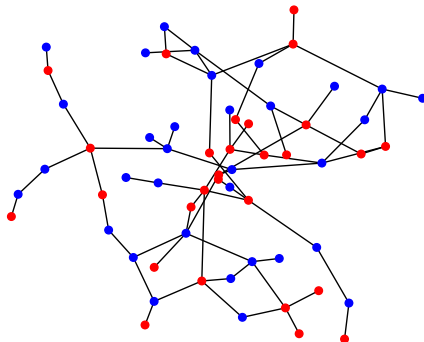
Resposta 1: Se a partição contém um único grupo, $Q = 0$. Por quê?

Modularidade

Pergunta: Para qual rede e partição dos nós o valor da modularidade pode ser ≈ 0 ?

Resposta 1: Se a partição contém um único grupo, $Q = 0$. Por quê?

Resposta 2: Se o número de arestas nos grupos não é maior do que “o esperado”.

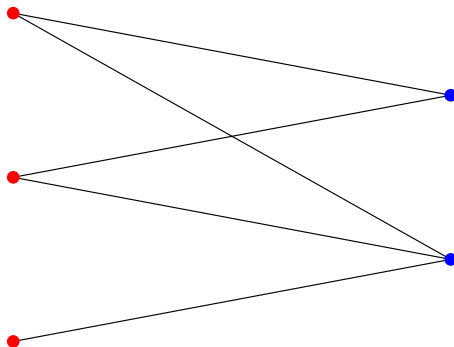


Modularidade

Pergunta: Para qual rede e partição dos nós o valor da modularidade pode ser $\approx -\frac{1}{2}$?

Modularidade

Pergunta: Para qual rede e partição dos nós o valor da modularidade pode ser $\approx -\frac{1}{2}$?



Modularidade

Qual é a **complexidade computacional** do cálculo da modularidade pela fórmula?

$$Q = \frac{1}{2m} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(s_i, s_j)$$

Modularidade

Qual é a **complexidade computacional** do cálculo da modularidade pela fórmula?

$$Q = \frac{1}{2m} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(s_i, s_j)$$

A complexidade é $O(n^2)$. O custo é alto para redes esparsas!

Modularidade: versão eficiente

Como calcular a modularidade de forma eficiente?

Modularidade: versão eficiente

Como calcular a modularidade de forma eficiente?

Solução: fórmula “**centrada em comunidades**”. Primeiro, vamos **reescrever** a fórmula:

$$Q = \frac{1}{2m} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(s_i, s_j)$$

$$Q = \frac{1}{2m} \sum_k \sum_{i,j \in A_k} \left(a_{ij} - \frac{k_i k_j}{2m} \right)$$

Modularidade: versão eficiente

$$Q = \frac{1}{2m} \sum_k \sum_{i,j \in A_k} \left(a_{ij} - \frac{k_i k_j}{2m} \right)$$

Modularidade: versão eficiente

$$Q = \frac{1}{2m} \sum_k \sum_{i,j \in A_k} \left(a_{ij} - \frac{k_i k_j}{2m} \right)$$

$$Q = \frac{1}{2m} \sum_k \sum_{i,j \in A_k} a_{ij} - \frac{1}{2m} \sum_k \sum_{i,j \in A_k} \frac{k_i k_j}{2m}$$

$$Q = \sum_k \sum_{i,j \in A_k} \frac{a_{ij}}{2m} - \frac{1}{4m^2} \sum_k \sum_{i,j \in A_k} k_i k_j$$

Modularidade: versão eficiente

$$Q = \frac{1}{2m} \sum_k \sum_{i,j \in A_k} \left(a_{ij} - \frac{k_i k_j}{2m} \right)$$

$$Q = \frac{1}{2m} \sum_k \sum_{i,j \in A_k} a_{ij} - \frac{1}{2m} \sum_k \sum_{i,j \in A_k} \frac{k_i k_j}{2m}$$

$$Q = \sum_k \sum_{i,j \in A_k} \frac{a_{ij}}{2m} - \frac{1}{4m^2} \sum_k \sum_{i,j \in A_k} k_i k_j$$

O valor $\sum_{i,j \in A_k} \frac{a_{ij}}{2m}$ é $2 \times$ o número de arestas no grupo A_k , dividido por $2m$. Ou seja, é a fração real de arestas dentro do grupo A_k . Vamos chamar este valor de e_k :

$$\sum_{i,j \in A_k} \frac{a_{ij}}{2m} = e_k.$$

Modularidade: versão eficiente

$$Q = \sum_k e_k - \frac{1}{4m^2} \sum_k \sum_{i,j \in A_k} k_i k_j$$

$$\sum_{i,j \in A_k} k_i k_j = \sum_{i \in A_k} \sum_{j \in A_k} k_i k_j = \sum_{i \in A_k} k_i \sum_{j \in A_k} k_j = \left(\sum_{i \in A_k} k_i \right)^2$$

Modularidade: versão eficiente

$$Q = \sum_k e_k - \frac{1}{4m^2} \sum_k \sum_{i,j \in A_k} k_i k_j$$

$$\sum_{i,j \in A_k} k_i k_j = \sum_{i \in A_k} \sum_{j \in A_k} k_i k_j = \sum_{i \in A_k} k_i \sum_{j \in A_k} k_j = \left(\sum_{i \in A_k} k_i \right)^2$$

$$Q = \sum_k e_k - \frac{1}{4m^2} \sum_k \left(\sum_{i \in A_k} k_i \right)^2 = \sum_k e_k - \sum_k \left(\frac{\sum_{i \in A_k} k_i}{2m} \right)^2$$

Modularidade: versão eficiente

$$Q = \sum_k e_k - \frac{1}{4m^2} \sum_k \sum_{i,j \in A_k} k_i k_j$$

$$\sum_{i,j \in A_k} k_i k_j = \sum_{i \in A_k} \sum_{j \in A_k} k_i k_j = \sum_{i \in A_k} k_i \sum_{j \in A_k} k_j = \left(\sum_{i \in A_k} k_i \right)^2$$

$$Q = \sum_k e_k - \frac{1}{4m^2} \sum_k \left(\sum_{i \in A_k} k_i \right)^2 = \sum_k e_k - \sum_k \left(\frac{\sum_{i \in A_k} k_i}{2m} \right)^2$$

O valor $\frac{\sum_{i \in A_k} k_i}{2m} = a_k$ é o grau total do grupo, dividido por $2m$.

Outra interpretação: é a **fração de stubs** incidentes aos vértices no grupo k .

Modularidade: versão eficiente

Forma mais fácil de calcular a modularidade:

$$Q = \sum_k (e_k - a_k^2).$$

- e_k : número de arestas no grupo A_k , dividido por m .
Representa a fração real de arestas dentro dos grupos.
- a_k : grau total do grupo k , dividido por $2m$ (ou fração de stubs incidentes aos vértices do grupo k no modelo de configuração).

Modularidade: versão eficiente

Forma mais fácil de calcular a modularidade:

$$Q = \sum_k (e_k - a_k^2).$$

- e_k : número de arestas no grupo A_k , dividido por m .
Representa a fração real de arestas dentro dos grupos.
- a_k : grau total do grupo k , dividido por $2m$ (ou fração de stubs incidentes aos vértices do grupo k no modelo de configuração).

Complexidade computacional do cálculo da modularidade pela fórmula “eficiente”?

Modularidade: versão eficiente

Forma mais fácil de calcular a modularidade:

$$Q = \sum_k (e_k - a_k^2).$$

- e_k : número de arestas no grupo A_k , dividido por m .
Representa a fração real de arestas dentro dos grupos.
- a_k : grau total do grupo k , dividido por $2m$ (ou fração de stubs incidentes aos vértices do grupo k no modelo de configuração).

Complexidade computacional do cálculo da modularidade pela fórmula “eficiente”?

– Complexidade $O(n + m)$. O custo é linear!

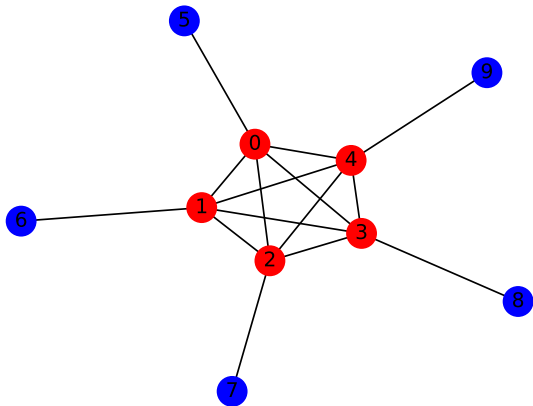
Modularidade: versão eficiente

Exemplo, se temos três grupos:

$$Q = \sum_k (e_k - a_k^2) = (e_1 - a_1^2) + (e_2 - a_2^2) + (e_3 - a_3^2).$$

Cada um desses termos representa a **contribuição de um grupo** à modularidade da rede inteira!

Modularidade



Nesta rede:

- Número de arestas dentro dos grupos: 10.
- Número esperado de arestas dentro dos grupos (*old*): 7,5.
- Valor da modularidade: -0,05.

Revisão do conteúdo

Pendências:

? **Pergunta 1:** A diferença entre os números real e esperado de arestas intra-grupo na rede é **significativa**, **suficiente** para afirmar que há homofilia?

- Um valor de modularidade **a partir de 0,3** geralmente é considerado um indicador da existência de uma estrutura de comunidades na rede.

✓ **Pergunta 2:** Podemos criar um **indicador simples** (um número) para a homofilia de uma rede?

✓ **Pergunta 3:** O que fazer se temos **mais de 2 grupos**?

Modularidade

E em grafos com pesos nas arestas? – Não muda nada!

$$Q = \frac{1}{2W} \sum_{ij} \left(w_{ij} - \frac{w_i^T w_j^T}{2W} \right) \delta(s_i, s_j),$$

onde:

- w_{ij} são os pesos das arestas;
- W é o peso total de todas as arestas;
- w_i^T são os pesos totais das arestas incidentes ao vértice.

Resumo

- 1 Revisão do conteúdo
- 2 Modelo de configuração
- 3 Modularidade
- 4 **Detecção de comunidades**
- 5 Aplicações

Detecção de comunidades

O que é detecção de comunidades?

- Suponha agora que temos **apenas** o grafo $G = (V, E)$. Não temos uma partição dos vértices definida!

Detecção de comunidades

O que é detecção de comunidades?

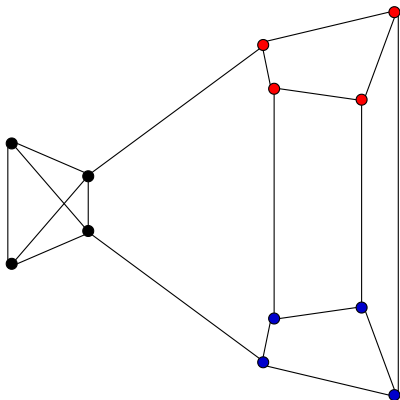
- Suponha agora que temos **apenas** o grafo $G = (V, E)$. Não temos uma partição dos vértices definida!
- Existem **muitas** formas de **particionar** (dividir) **os vértices** em grupos. É impossível checar todas elas!

Detecção de comunidades

O que é detecção de comunidades?

- Suponha agora que temos **apenas** o grafo $G = (V, E)$. Não temos uma partição dos vértices definida!
- Existem **muitas** formas de **particionar** (dividir) **os vértices** em grupos. É impossível checar todas elas!
- Algumas divisões (partições) são **melhores** do que outras. Como avaliar qual é a melhor? $\Rightarrow$ **Modularidade!**

Detecção de comunidades



Modularidade:

- Se juntamos os nós vermelhos e azuis no mesmo grupo:

$$Q \approx 0,36.$$

- Se consideramos os 3 grupos:

$$Q \approx 0,40.$$

Detecção de comunidades

Ideia de todos os métodos de detecção de comunidades:

Devemos ter muitas arestas entre vértices do mesmo grupo, e comparativamente poucas arestas entre vértices de grupos diferentes.

Ideias: um “bom grupo” seria aquele no qual a **densidade** das arestas dentro do grupo é grande? No qual **os cortes** entre os grupos seriam os menores?

Detecção de comunidades

Ideia de todos os métodos de detecção de comunidades:

Devemos ter muitas arestas entre vértices do mesmo grupo, e comparativamente poucas arestas entre vértices de grupos diferentes.

Ideias: um “bom grupo” seria aquele no qual a **densidade** das arestas dentro do grupo é grande? No qual **os cortes** entre os grupos seriam os menores?

Os modelos mais utilizados são aqueles baseados na otimização da **modularidade**.
Melhor partição – aquela para a qual a função **Q** possui o maior valor!

- O número de grupos não é fixo. Ele pode variar de **0** até **$|V|$** !

Santo Fortunato: “Community detection in graphs” (2010).
→ +8900 citações no Scopus, +13600 citações no Google Scholar.

Detecção de comunidades

A modularidade permite **avaliar** se uma dada divisão dos vértices em grupos é **apropriada**:

- Enquanto **maior** o valor da modularidade Q , **melhor** é a partição proposta.

A modularidade é a função objetivo a ser maximizada no problema de encontrar a melhor partição dos vértices em comunidades!

Detecção de comunidades

Encontrar uma partição dos vértices em grupos que maximize o valor de modularidade é um problema **NP-difícil!**

U. Brandes et al., “On Modularity Clustering”, *IEEE Transactions on Knowledge and Data Engineering*, v. 20, pp. 172-188 (2008).

Detecção de comunidades: pergunta de controle

Em quais tipos de grafos, **independentemente da partição dos vértices escolhida**, o valor da modularidade será **zero**?

Detecção de comunidades: pergunta de controle

Em quais tipos de grafos, **independentemente da partição dos vértices escolhida**, o valor da modularidade será **zero**?

- Em grafos completos.

Resumo

- 1 Revisão do conteúdo
- 2 Modelo de configuração
- 3 Modularidade
- 4 Detecção de comunidades
- 5 Aplicações

Aplicações

A **modularidade** é usada no contexto da detecção de comunidades nas **mais diversas áreas** para identificar estruturas ou grupos de vértices presentes nessas redes:

- Nas redes de interação de proteínas, as comunidades agrupam **proteínas que têm a mesma função específica** dentro da célula;
- Na World Wide Web, os grupos correspondem a **páginas relacionadas aos mesmos tópicos**, interesses ou negócios;
- Identificar grupos de **clientes com interesses semelhantes** permite criar bons sistemas de recomendação (www.amazon.com).

Material bibliográfico

M. Newman, M. Girvan: “Finding and evaluating community structure in networks” (2004).

S. Fortunato: “Community detection in graphs” (2010).

S. Fortunato, D. Hric: “Community detection in networks: A user guide” (2016).

Dúvidas

Dúvidas?