Introdução às Redes de Interação – MO804 (MC908)
Homofilia, Modularidade e Detecção de Comunidades

Prof. Dr. Ruben Interian

Instituto de Computação, UNICAMP

Resumo

Resumo

Homofilia

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

Homofilia

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

O que é "semelhantes"?

- Representação **mais fácil**: dois vértices são "semelhantes" se pertencem ao mesmo grupo.
- **Outras** representações: cada nó possui um vetor de atributos, podemos usar a "similaridade" entre estes vetores.

Homofilia

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

O que é "semelhantes"?

- Representação **mais fácil**: dois vértices são "semelhantes" se pertencem ao mesmo grupo.
- **Outras** representações: cada nó possui um vetor de atributos, podemos usar a "similaridade" entre estes vetores.

Como avaliar a homofilia em uma rede real?

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

Homofilia

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

Homofilia

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

- Seja $G = (V, E)$ um grafo, e seja $A = \{A_1, \ldots, A_k\}$ uma coleção de k grupos de nós que formam uma partição do conjunto V.

Homofilia

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

- Seja $G = (V, E)$ um grafo, e seja $A = \{A_1, \ldots, A_k\}$ 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.

Homofilia

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

- Seja $G = (V, E)$ um grafo, e seja $A = \{A_1, \ldots, A_k\}$ 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 $h(v)$.

Homofilia

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

- Seja $G = (V, E)$ um grafo, e seja $A = \{A_1, \ldots, A_k\}$ 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.

Homofilia

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

Homofilia

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

- **Primeira abordagem**: usar a média dos valores de $h(v)$ de cada vértice. Útil apenas em casos simples.

Homofilia

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

- **Primeira abordagem**: usar a média dos valores de $h(v)$ de cada vértice. Útil apenas em casos simples.
- **Abordagem um pouco melhor**: avaliar quão longe está a quantidade de arestas entre vértices de diferentes grupos do valor esperado.

Homofilia

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

- **Primeira abordagem**: usar a média dos valores de $h(v)$ de cada vértice. Útil apenas em casos simples.
- **Abordagem um pouco melhor**: avaliar quão longe está a quantidade de arestas entre vértices de diferentes grupos do valor esperado.
- 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 $2w_x w_y$.

Homofilia

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

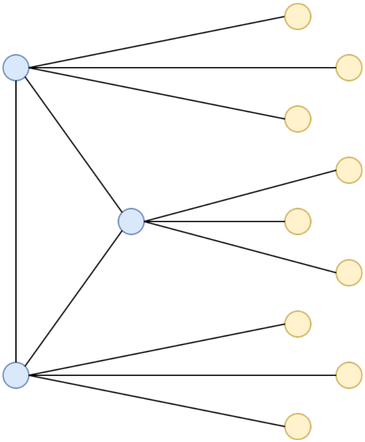
- **Primeira abordagem**: usar a média dos valores de $h(v)$ de cada vértice. Útil apenas em casos simples.
- **Abordagem um pouco melhor**: avaliar quão longe está a quantidade de arestas entre vértices de diferentes grupos do valor esperado.
- Se uma aresta é colocada **aleatoriamente** no grafo, então 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 evidencia** de homofilia.

Resumo

Homofilia *revisited*

Revisitando a homofilia.

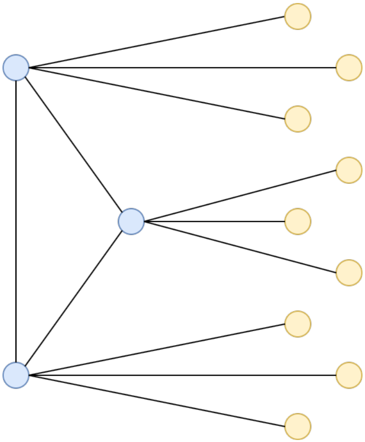
- **Quantas arestas inter-grupo e intra-grupo há na rede?**



Homofilia *revisited*

Revisitando a homofilia.

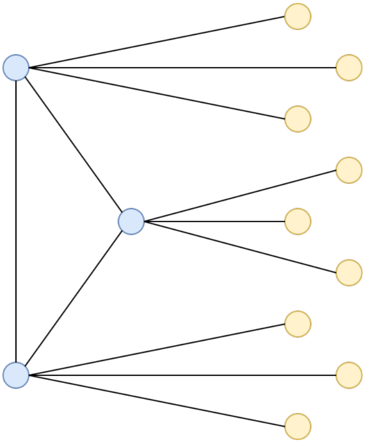
- **Quantas arestas inter-grupo e intra-grupo há na rede?**
  9 arestas inter-grupo, 3 arestas no grupo azul.



Homofilia *revisited*

Revisitando a homofilia.

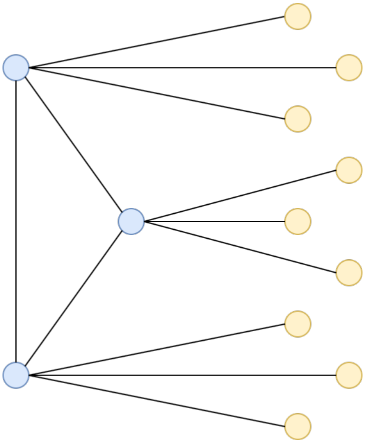
- **Quantas arestas inter-grupo e intra-grupo há na rede?**
  9 arestas inter-grupo, 3 arestas no grupo azul.
- **Qual é o número esperado dessas arestas se não houvesse homofilia?**



Homofilia *revisited*

Revisitando a homofilia.

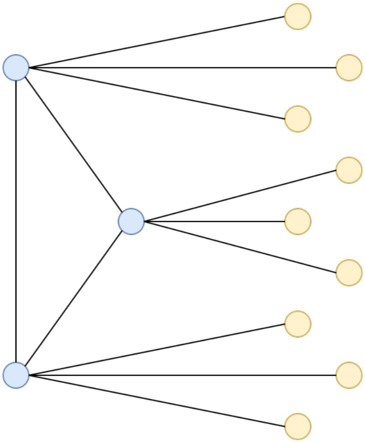
- **Quantas arestas inter-grupo e intra-grupo há na rede?**
  9 arestas inter-grupo, 3 arestas no grupo azul.
- **Qual é o número esperado dessas arestas se não houvesse homofilia?**
  $2w_x w_y = 2\frac{1}{4}\frac{3}{4} = \frac{3}{8}$
  $\Rightarrow$ o esperado é $12\frac{3}{8} = 4.5$ arestas inter-grupo.



Homofilia *revisited*

Revisitando a homofilia.

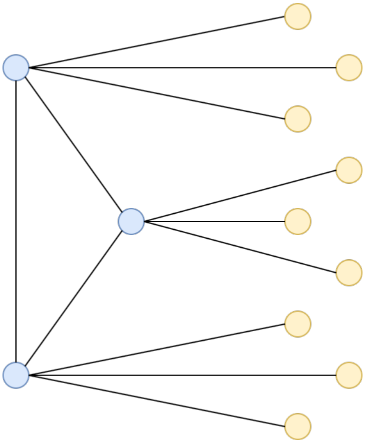
- **Quantas arestas inter-grupo e intra-grupo há na rede?**
  9 arestas inter-grupo, 3 arestas no grupo azul.
- **Qual é o número esperado dessas arestas se não houvesse homofilia?**
  $2w_x w_y = 2\frac{1}{4}\frac{3}{4} = \frac{3}{8}$
  $\Rightarrow$ o esperado é $12\frac{3}{8} = 4.5$ arestas inter-grupo.
  $w_x^2 = (\frac{1}{4})^2 = \frac{1}{16}$
  $\Rightarrow$ o esperado é $12\frac{1}{16} = \frac{3}{4}$ arestas no grupo X.



Homofilia *revisited*

Revisitando a homofilia.

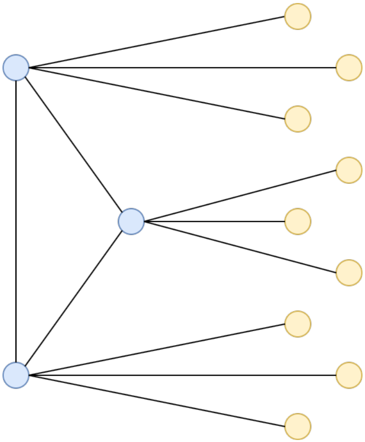
- O esperado, em um grafo aleatório, é $\frac{3}{4}$ de aresta no grupo azul. Mas, se temos 2 vértices de grau 5 em um grafo com 12 arestas, qual é a chance deles estarem conectados?...



Homofilia *revisited*

Revisitando a homofilia.

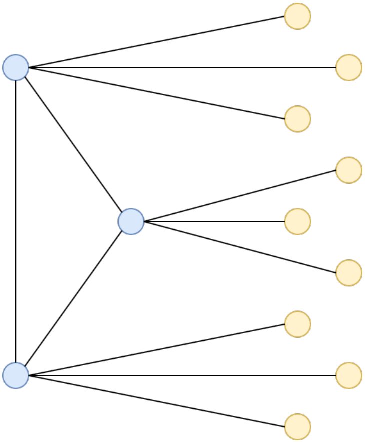
- O esperado, em um grafo aleatório, é $\frac{3}{4}$ de aresta no grupo azul. Mas, se temos 2 vértices de grau 5 em um grafo com 12 arestas, qual é a chance deles estarem conectados?...
- A chance deles **não estarem conectados** é **pequena**. Entre 2 vértices azuis, é bem mais provável a aresta existir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



Homofilia *revisited*

Revisitando a homofilia.

- O esperado, em um grafo aleatório, é $\frac{3}{4}$ de aresta no grupo azul. Mas, se temos 2 vértices de grau 5 em um grafo com 12 arestas, 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 tem **mais chance de se conectar** do que vértices de grau baixo... Precisamos considerar isso no nosso modelo!



Modelo de configuração

Como **modelar corretamente** uma rede aleatória na qual **os graus** dos vértices estão **previamente definidos**?

Modelo de configuração

Como **modelar corretamente** uma rede aleatória na qual **os graus** dos vértices estão **previamente definidos**?

- Se que cada vértice v_i na rede tem grau $d(v_i) = k_i$, temos: $\sum_i k_i = 2|E| = 2m$.

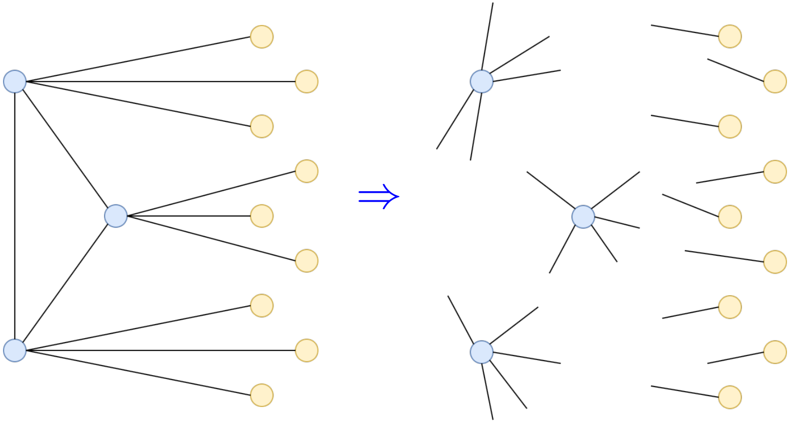
Modelo de configuração

Como **modelar corretamente** uma rede aleatória na qual **os graus** dos vértices estão **previamente definidos**?

- Se que 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).

Modelo de configuração

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



Modelo de configuração

- Podemos escolher um par de **meia-arestas** / **stubs** aleatoriamente dos $2m$ **stubs**, e por cada par criar uma aresta. Depois, escolhemos **outro par** dentre as $2m - 2$ meia-arestas restantes, e assim até todos os **stubs** acabarem.

Modelo de configuração

- Podemos escolher um par de **meia-arestas** / **stubs** aleatoriamente dos $2m$ **stubs**, e por cada par criar uma aresta. Depois, escolhemos **outro par** dentre as $2m - 2$ meia-arestas restantes, e assim até todos os **stubs** acabarem.
- O **resultado** é uma rede na qual o **conjunto de arestas** é gerado a partir de um emparelhamento dos *stubs*, e cada vértice possui exatamente o grau k_i.

Modelo de configuração

- Podemos escolher um par de **meia-arestas** / **stubs** aleatoriamente dos $2m$ **stubs**, e por cada par criar uma aresta. Depois, escolhemos **outro par** dentre as $2m - 2$ meia-arestas restantes, e assim até todos os **stubs** acabarem.
- O **resultado** é uma rede na qual o **conjunto de arestas** é gerado a partir de um emparelhamento dos *stubs*, e cada vértice possui exatamente o grau k_i.
- O **modelo de configuração** é o **espaço** (conjunto) **de todos os emparelhamentos**, onde cada um deles tem a mesma chance de aparecer. Cada emparelhamento define um **conjunto de arestas** E.

Modelo de configuração

- Podemos escolher um par de **meia-arestas** / **stubs** aleatoriamente dos $2m$ **stubs**, e por cada par criar uma aresta. Depois, escolhemos **outro par** dentre as $2m - 2$ meia-arestas restantes, e assim até todos os **stubs** acabarem.
- O **resultado** é uma rede na qual o **conjunto de arestas** é gerado a partir de um emparelhamento dos *stubs*, e cada vértice possui exatamente o grau k_i.
- O **modelo de configuração** é o **espaço** (conjunto) **de todos os emparelhamentos**, onde cada um deles tem a mesma chance de aparecer. Cada emparelhamento define um **conjunto de arestas** E.
- No modelo de configuração, cada **stub** tem a mesma probabilidade de se conectar a qualquer outro.

Modelo de configuração

Outras características do modelo de configuração:

- A rede gerada pode contar **laços** e **arestas múltiplas**.

Modelo de configuração

Outras características do modelo de configuração:

- A rede gerada pode contar **laços** e **arestas múltiplas**.
- **Evitar laços e arestas múltiplas leva a problemas!**

Modelo de configuração

Outras características do modelo de configuração:

- A rede gerada pode contar **laços** e **arestas múltiplas**.
- **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 $n \to \infty$, portanto podemos negligenciar esse fato.

Modelo de configuração

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

Modelo de configuração

Probabilidade de uma aresta entre vértices não isolados 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 vértices não isolados 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 vértices não isolados 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 vértices não isolados 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: $\frac{k_i k_j}{2m-1}$.
- **Tecnicamente falando**, o valor $\frac{k_i k_j}{2m-1}$ é, na verdade, a média de arestas que haverá entre v_i e v_j, mas no limite, quando m cresce, a média e a **chance** de termos a aresta são iguais, e podemos ignorar o valor -1 do denominador:

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

Resumo

Modularidade

A **modularidade** é uma forma mais precisa e matematicamente correta de avaliar a **homofilia de uma rede**, e é baseada no modelo de configuração.

Modularidade

A **modularidade** é uma forma mais precisa e matematicamente correta de avaliar a **homofilia de uma rede**, e é baseada no modelo de configuração.

- Seja $A = \{A_1, \ldots, A_k\}$ uma coleção de k grupos de nós, e s_i é o grupo do nó v_i.

Modularidade

A **modularidade** é uma forma mais precisa e matematicamente correta de avaliar a **homofilia de uma rede**, e é baseada no modelo de configuração.

- Seja $A = \{A_1, \ldots, A_k\}$ uma coleção de k grupos de nós, e s_i é o grupo do nó v_i.
- No modelo de configuração, a chance de uma aresta entre v_i e v_j existir é: $\frac{k_i k_j}{2m}$.

Modularidade

A **modularidade** é uma forma mais precisa e matematicamente correta de avaliar a **homofilia de uma rede**, e é baseada no modelo de configuração.

- Seja $A = \{A_1, \ldots, A_k\}$ uma coleção de k grupos de nós, e s_i é o grupo do nó v_i.
- 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 **total de arestas esperadas** entre vértices que pertencem ao **mesmo grupo**?

Modularidade

A **modularidade** é uma forma mais precisa e matematicamente correta de avaliar a **homofilia de uma rede**, e é baseada no modelo de configuração.

- Seja $A = \{A_1, \ldots, A_k\}$ uma coleção de k grupos de nós, e s_i é o grupo do nó v_i.
- 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 **total de arestas esperadas** 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

- **Quantidade esperada** de arestas entre vértices no mesmo grupo:

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

Modularidade

- **Quantidade esperada** de arestas entre vértices no mesmo grupo:

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

- **Quantidade real** de arestas entre vértices no mesmo grupo ($a_{ij} = 1$ se há aresta):

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

Modularidade

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

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

Geralmente, não queremos a diferença absoluta, mas a **diferença relativa**:

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

Modularidade

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

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

Geralmente, não queremos a diferença absoluta, mas a **diferença relativa**:

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

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, dados os grupos de nós em uma rede, o número de arestas intra-grupo existentes, calculando o número real de arestas intragrupo menos o número esperado dessas arestas em uma rede com os mesmos nós, os mesmos grupos e os mesmos graus dos nós, 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, dados os grupos de nós em uma rede, o número de arestas intra-grupo existentes, calculando o número real de arestas intragrupo menos o número esperado dessas arestas em uma rede com os mesmos nós, os mesmos grupos e os mesmos graus dos nós, mas com as arestas colocadas aleatoriamente.

A **modularidade** é a forma apropriada de avaliar a **homofilia de uma rede**.

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, dados os grupos de nós em uma rede, o número de arestas intra-grupo existentes, calculando o número real de arestas intragrupo menos o número esperado dessas arestas em uma rede com os mesmos nós, os mesmos grupos e os mesmos graus dos nós, mas com as arestas colocadas aleatoriamente.

A **modularidade** é a forma apropriada de avaliar a **homofilia de uma rede**.

A **modularidade** foi proposta por Mark Newman, da Universidade de Michigan, cujos trabalhos estão entre aqueles com maior impacto (em ciências exatas) no século XXI.

Modularidade

Resumo da fórmula:

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

- $m = |E|$ 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:

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

- $m = |E|$ 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

Forma mais fácil de calcular a modularidade:

$$Q = \sum_i (e_{ii} - a_i^2).$$

Modularidade

Forma mais fácil de calcular a modularidade:

$$Q = \sum_i (e_{ii} - a_i^2).$$

- O valor e_{ii} é a **fração de stubs** de arestas dentro do grupo i. (2× o número de arestas no grupo i dividido por $2m$) e_{ii} representa a fração real de arestas dentro dos grupos.

Modularidade

Forma mais fácil de calcular a modularidade:

$$Q = \sum_i (e_{ii} - a_i^2).$$

- O valor e_{ii} é a **fração de stubs** de arestas dentro do grupo i.
  (2× o número de arestas no grupo i dividido por $2m$)
  e_{ii} representa a fração real de arestas dentro dos grupos.
- O valor a_i é a **fração de stubs** de arestas incidentes a vértices no grupo i.
  (inclui as anteriores, e as arestas que vão para outros grupos)
  a_i^2 representa a chance esperada de uma aresta ficar dentro dos grupos no modelo de configuração.

Modularidade

Observe, se temos dois grupos:

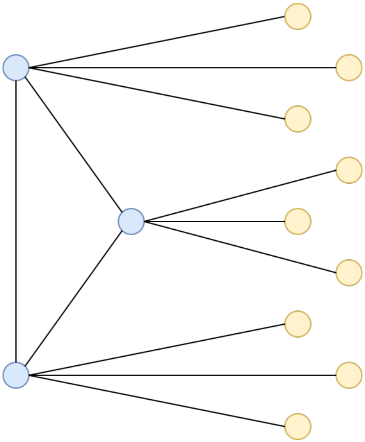
$$Q = \sum_i (e_{ii} - a_i^2) = (e_{11} - a_1^2) + (e_{22} - a_2^2).$$

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

Modularidade

Voltando ao exemplo anterior:

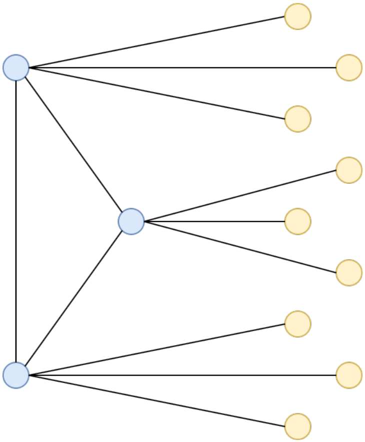
- Qual é a **fração de arestas** real e esperada no grupo azul no modelo de configuração?



Modularidade

Voltando ao exemplo anterior:

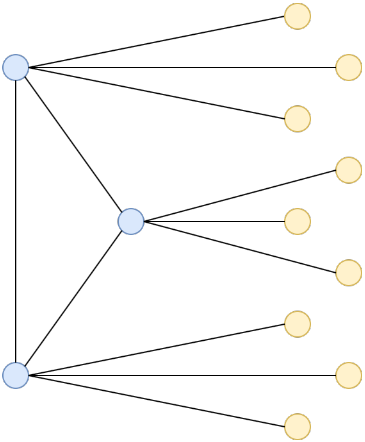
- Qual é a **fração de arestas** real e esperada no grupo azul no modelo de configuração?
  Exercício.



Modularidade

Voltando ao exemplo anterior:

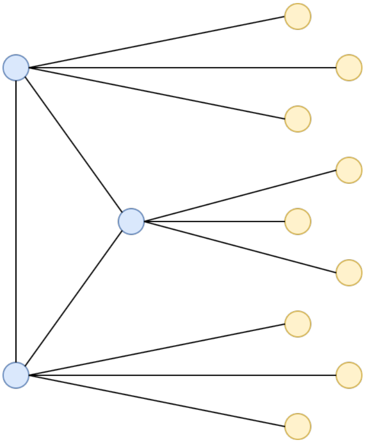
- Qual é a **fração de arestas** real e esperada no grupo azul no modelo de configuração? **Exercício.**
- Qual é o **número de arestas** real e esperado no grupo azul no modelo de configuração?



Modularidade

Voltando ao exemplo anterior:

- Qual é a **fração de arestas** real e esperada no grupo azul no modelo de configuração? **Exercício.**
- Qual é o **número de arestas** real e esperado no grupo azul no modelo de configuração? **Exercício.**



Modularidade

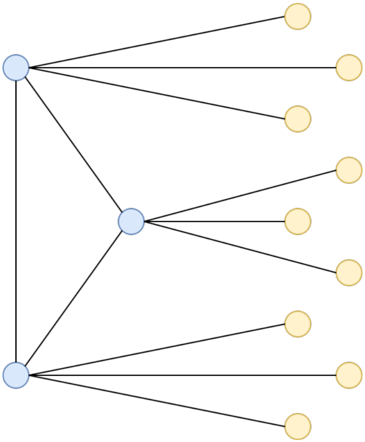
Voltando ao exemplo anterior:

- Qual é a **fração de arestas** real e esperada no grupo azul no modelo de configuração? **Exercício.**

- Qual é o **número de arestas** real e esperado no grupo azul no modelo de configuração? **Exercício.**

  Dica: calcular os valores e_{ii}, a_i^2 usando a fórmula do slide anterior.

  Compare os valores com os obtidos no modelo que não considera os graus.



Resumo

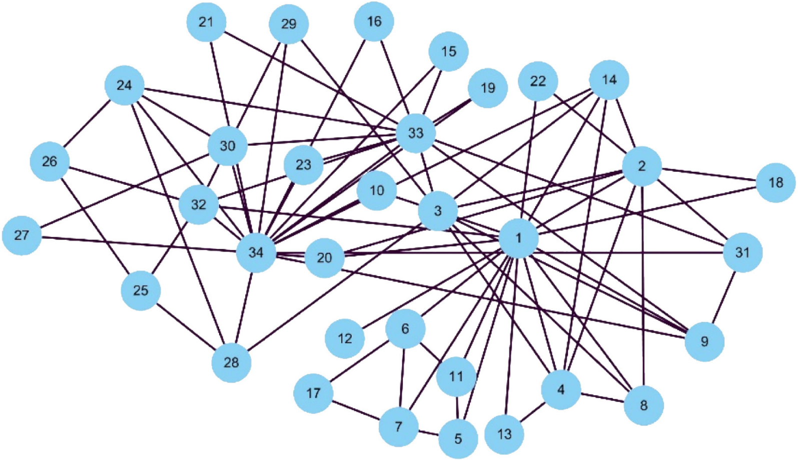
Detecção de comunidades

O que é detecção de comunidades?

- Suponha agora que temos **apenas** o grafo $G = (V, E)$.
- Existem muitas formas de **dividir os vértices** em grupos: $2^{|V|}$ divisões diferentes!
- Algumas divisões (partições) são **melhores** do que outras.

Detecção de comunidades: clube de karatê do Zachary

Algumas divisões (partições) são **melhores** do que outras.



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.

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 e abordagens para a detecção de comunidades:

- Um bom grupo seria aquele no qual a **densidade** das arestas no grupo é maior.

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 e abordagens para a detecção de comunidades:

- Um bom grupo seria aquele no qual a **densidade** das arestas no grupo é maior.
- *Hierarchical clustering*: os **grupos** podem **conter outros grupos**.

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 e abordagens para a detecção de comunidades:

- Um bom grupo seria aquele no qual a **densidade** das arestas no grupo é maior.
- *Hierarchical clustering*: os **grupos** podem **conter outros grupos**.
- Métodos baseados na **modularidade**: a melhor partição é aquela para a qual a função Q da modularidade possui o maior valor!

Santo Fortunato: "Community detection in graphs" (2010).
$\rightarrow$ 8264 citações no Scopus, 12681 citações no Google Scholar.

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

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).

Mesmo assim, a modularidade é muito importante para os métodos de **detecção de comunidades**, sendo considerada como um *gold standard* para boas comunidades.

- Ela 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 divisão dos vértices em grupos!

Detecção de comunidades

Exemplo de aplicação: método de agrupamento hierárquico **guloso**.

1. Não sabemos quantos grupos temos. Então, iniciamos com todos os vértices em grupos diferentes;
2. Em cada passo, escolhemos e adicionamos ao grafo uma aresta que junte dois grupos de forma que a **modularidade** aumente no maior valor possível;
3. Finalizamos o algoritmo com um único grupo, e escolhemos a melhor das $n-1$ divisões encontradas durante a execução.

Complexidade em grafos esparsos: $O(n^2)$.
Melhorado usando estruturas de dados eficientes, passa a $O(n\log^2(n))$.

$\rightarrow$ "Fast algorithm for detecting community structure in networks"

Detecção de comunidades em redes dinâmicas

Comunidades dinâmicas: o que significa identificar comunidades em redes dinâmicas (onde cada aresta ou vértice possui um momento de tempo associado)?

Detecção de comunidades em redes dinâmicas

Comunidades dinâmicas: o que significa identificar comunidades em redes dinâmicas (onde cada aresta ou vértice possui um momento de tempo associado)?

Algoritmos precisam identificar as **etapas do ciclo de vida** das comunidades:

- Nascimento,
- Expansão ou contração,
- Fusão com outra comunidade,
- Divisão em duas ou mais comunidades,
- Morte.

É um problema mais complexo e uma ativa área de pesquisa!

Resumo

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** nas redes de clientes e produtos 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?