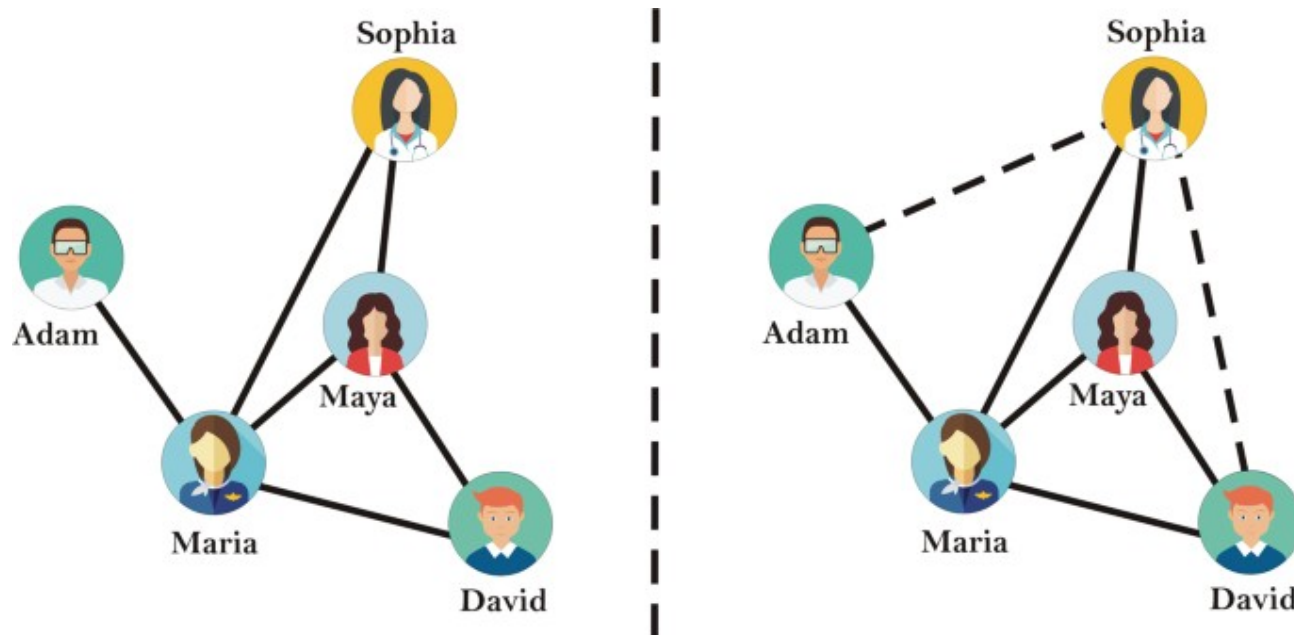


Ciência das Redes

Link prediction

Ricardo Luders
Thiago H Silva

Dada uma rede incompleta, preveja se é provável que dois nós tenham um link.



Aplicações:

- Recomendação de amigos nas redes sociais
- Recomendação de produtos em comércio eletrônico
- Previsão de interação em redes biológicas

- Rede observada: estado atual
- **Previsão de link:**
 - Pode aparecer no futuro (previsão de link futuro)
 - Pode ter sido perdido (previsão de link perdido)

Previsão de link com base nas propriedades da rede:

- **Local:** Alto agrupamento (amigos dos meus amigos se tornarão meus amigos)
- **Global:** dois *hubs* não relacionados com maior probabilidade de ter links que pequenos nós não relacionados
- **Organização em meso escala:** dois nós na mesma comunidade ...

A previsão do link também pode ser baseada nas propriedades do nó

- Combinando com o aprendizado de máquina normal

“Amigos dos meus amigos são meus amigos”

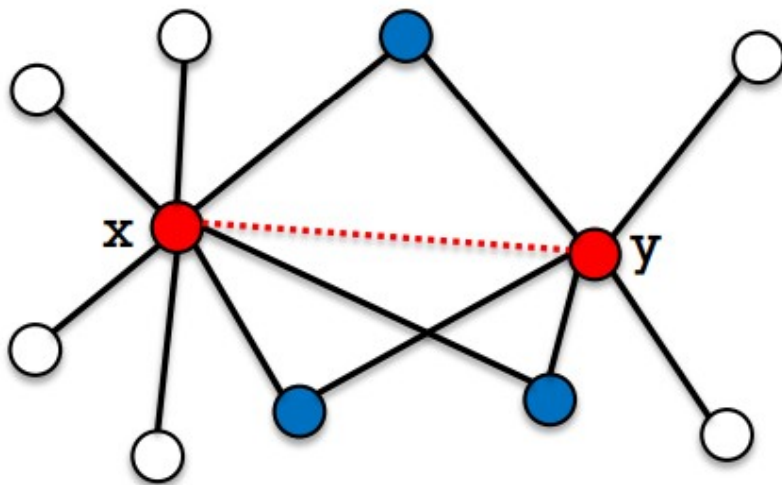
Alto *clustering* na maioria das redes

Quanto mais amigos em comum, maior a probabilidade de tornam-se amigos

$$CN(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

Common Neighbors

$\Gamma(x)$ o conjunto de vizinhos do nó x no grafo



x e y provavelmente têm um link se eles têm muitos vizinhos em comum.

Como prever links com base em vizinhos comuns?

Como prever links com base em vizinhos comuns?

- Para cada par de nós não conectados, calcule CN
Lista ordenada de pares de mais provável para menos provável

Um ponto fraco dessa abordagem é que ela não leva em consideração o número relativo de vizinhos comuns.

A medida Jaccard aborda o problema de vizinhos comuns, calculando o número relativo de vizinhos em comum:

$$JC(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

Jaccard Coefficient

Intuição:

Duas pessoas que conhecem apenas 4 pessoas, mas apenas 1 não compartilhada:

- alta probabilidade

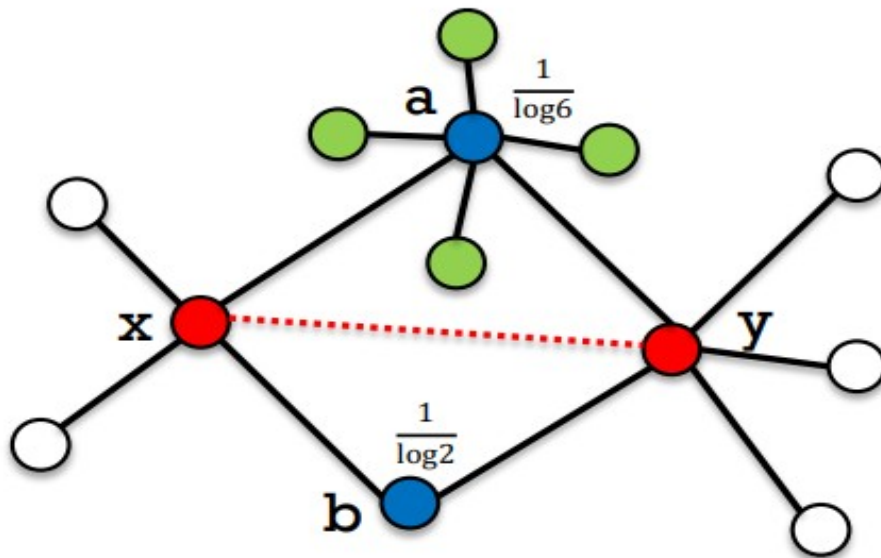
Duas pessoas que conhecem 1000 pessoas, apenas 3 em comuns:

- probabilidade mais baixa

Para medidas anteriores: todos os nós em comuns têm o mesmo valor

Adamic-Adar index

$$AA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$$



Vizinhos comuns ponderados;
Vizinhos comuns populares
contribuem menos.

Semelhante ao Adamic Adar, penaliza graus mais elevados

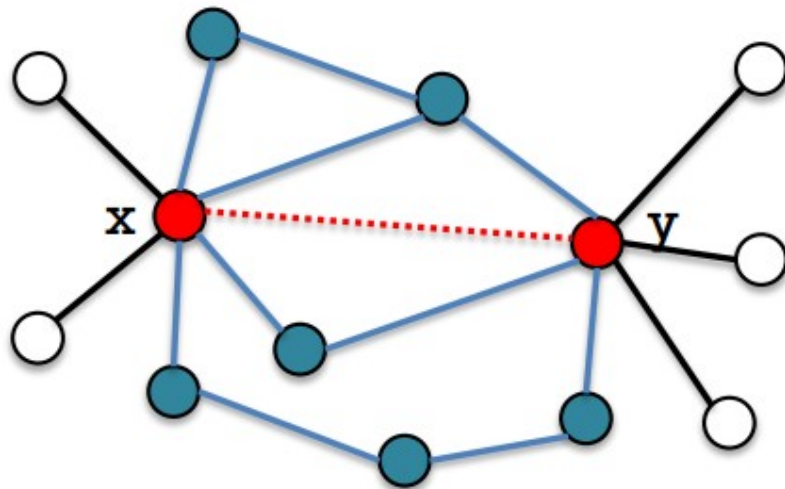
$$\text{RA}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|}$$

Resource Allocation index

Os métodos baseados em vizinhos podem ser eficazes quando o número de vizinhos é grande, mas esse não é o caso em grafos esparsos.

Nessas situações, é apropriado usar métodos que levem em consideração caminhos mais longos (similaridade Global):

$$\sum_{l=1}^{\infty} \beta^l \cdot |\text{paths}_{xy}^{(l)}|$$



Soma todos os caminhos entre x e y;

Cada caminho com desconto de β^l

$\beta < 1$ é o fator de desconto

l é o comprimento de um caminho

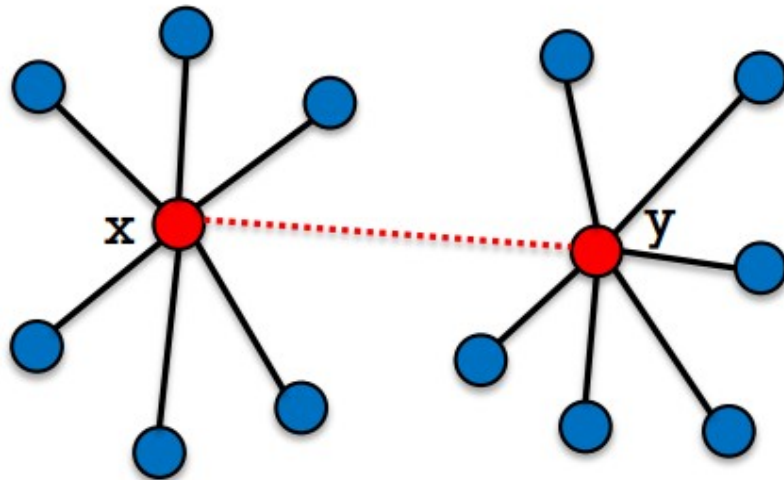
Caminhos mais longos contribuem menos.

Modelo de crescimento da rede baseado na ideia de que os “ricos ficam mais ricos”

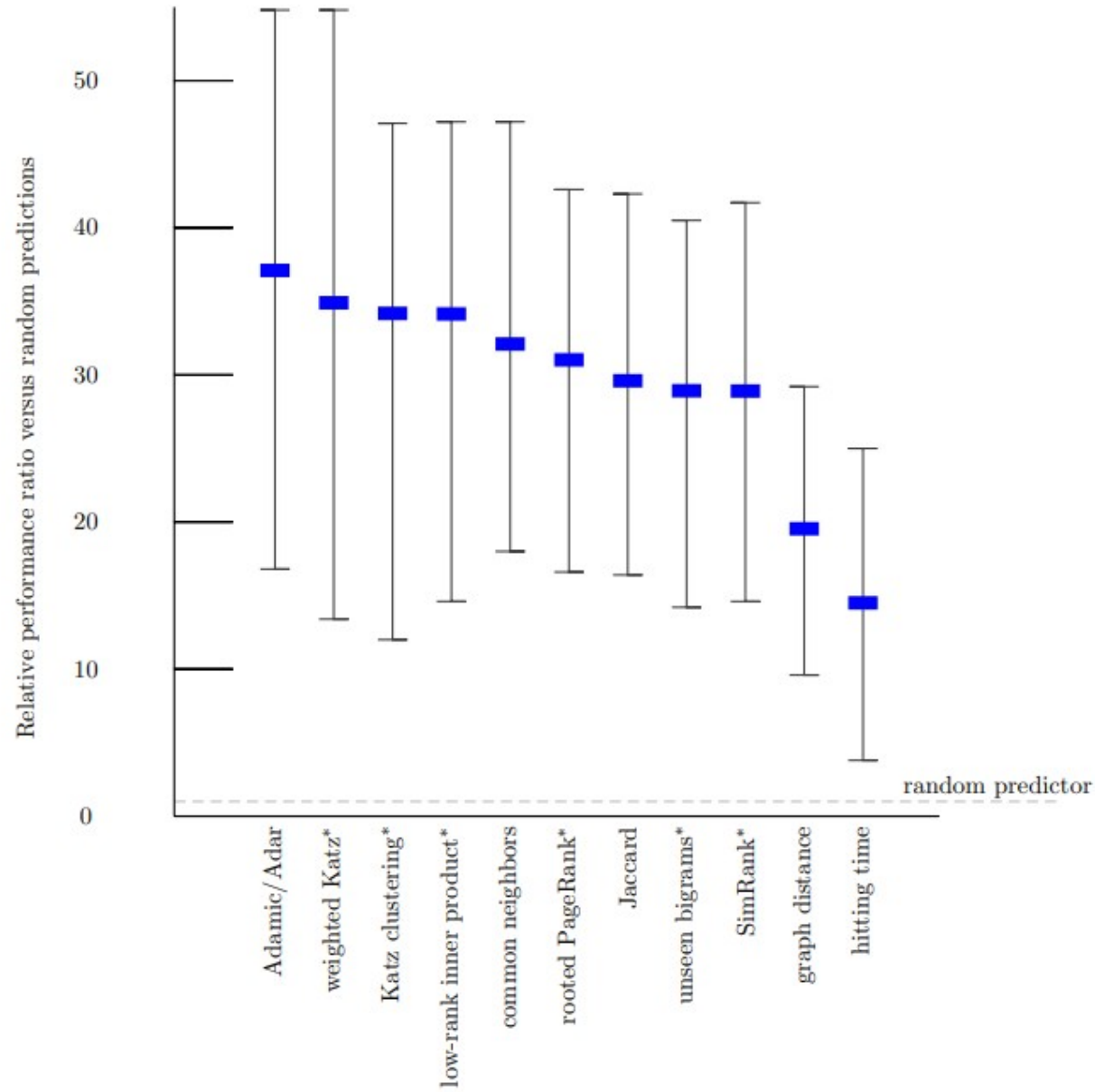
Cada vez que um nó se junta à rede, ele cria um link com nós com probabilidade = grau atual

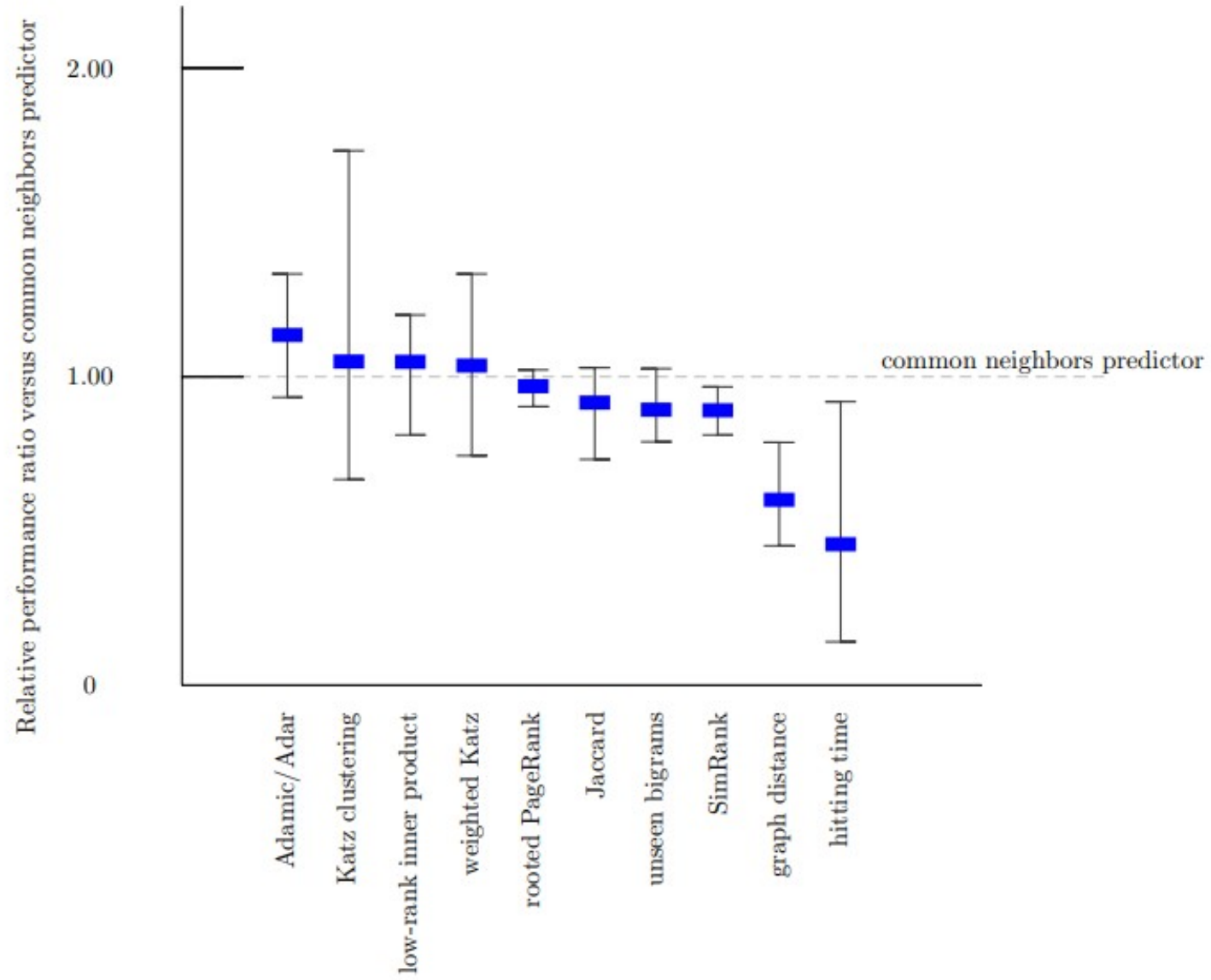
Gera distribuição de graus de lei de potência

$$|\Gamma(x)| * |\Gamma(y)|$$



x prefere se conectar a y se y for popular.





Vantagens

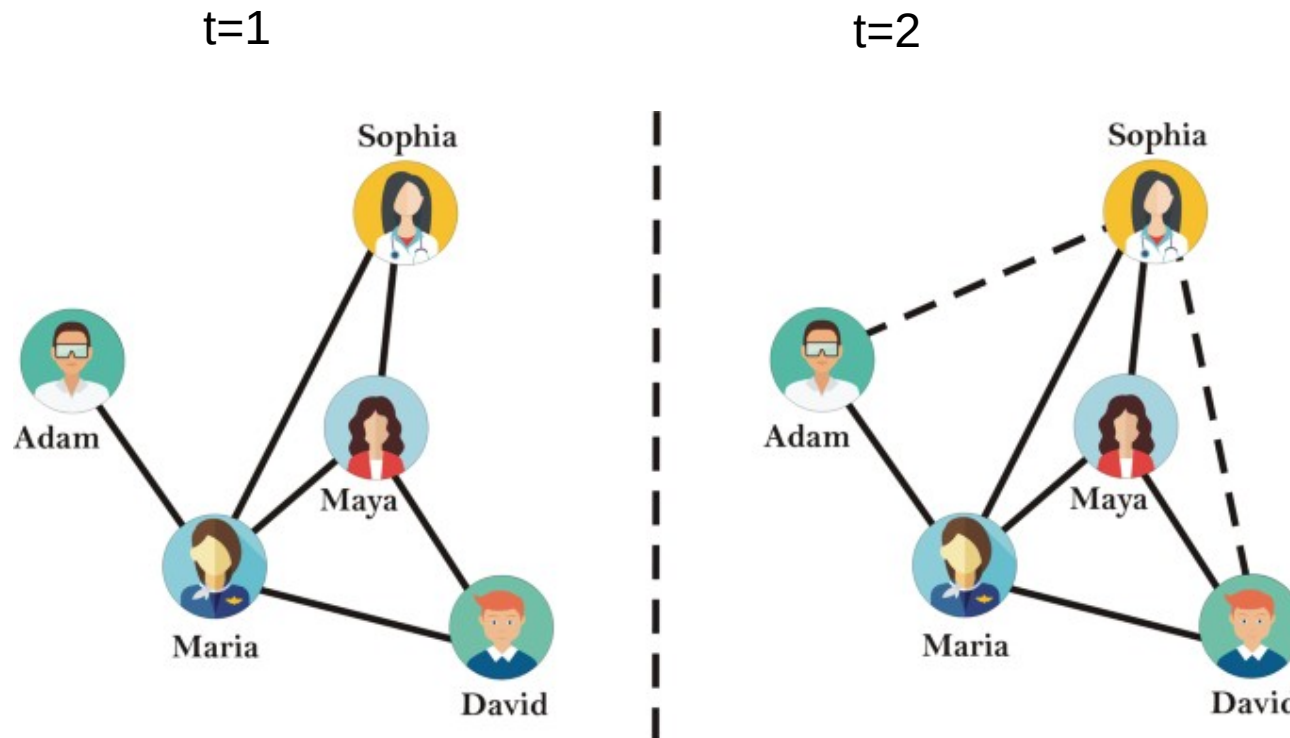
Fácil de computar

Alta interpretabilidade

Desvantagens

Features de estrutura da rede, como as mostradas, não são gerais.
Tem fortes suposições sobre os mecanismos de formação de links.
Funcionam bem apenas em algumas redes.

Aprendizado supervisionado poderia ser aplicado ao problema de *link prediction*?

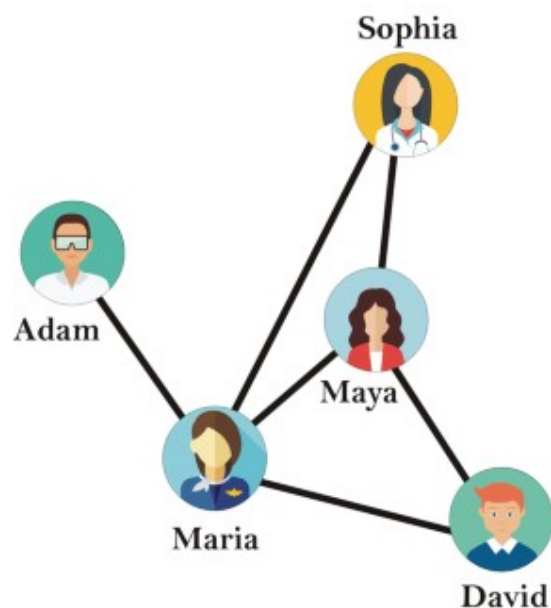


Nosso objetivo é prever se
existirá uma aresta dois nós
não conectados

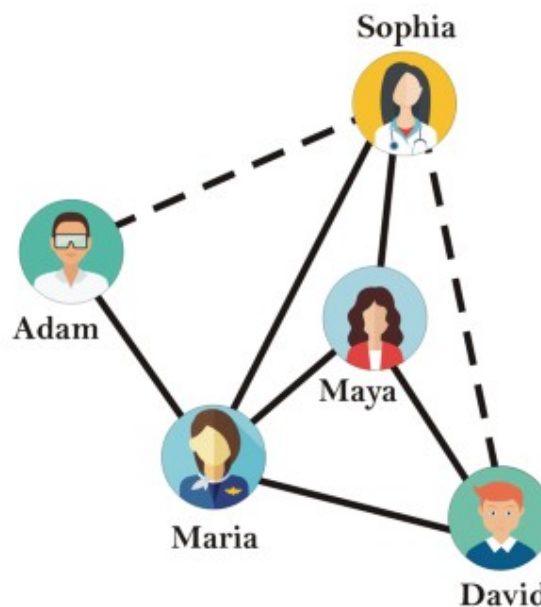
Sofia-Adam = ?
Sofia-David = ?
David-Adam = ?

...

t=1



t=2



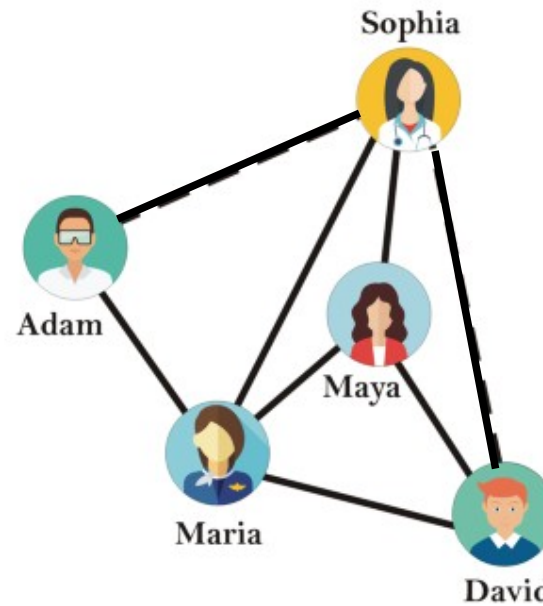
Nosso objetivo é prever se existirá uma aresta dois nós não conectados

Sofia-Adam = ?
Sofia-David = ?
David-Adam = ?
...

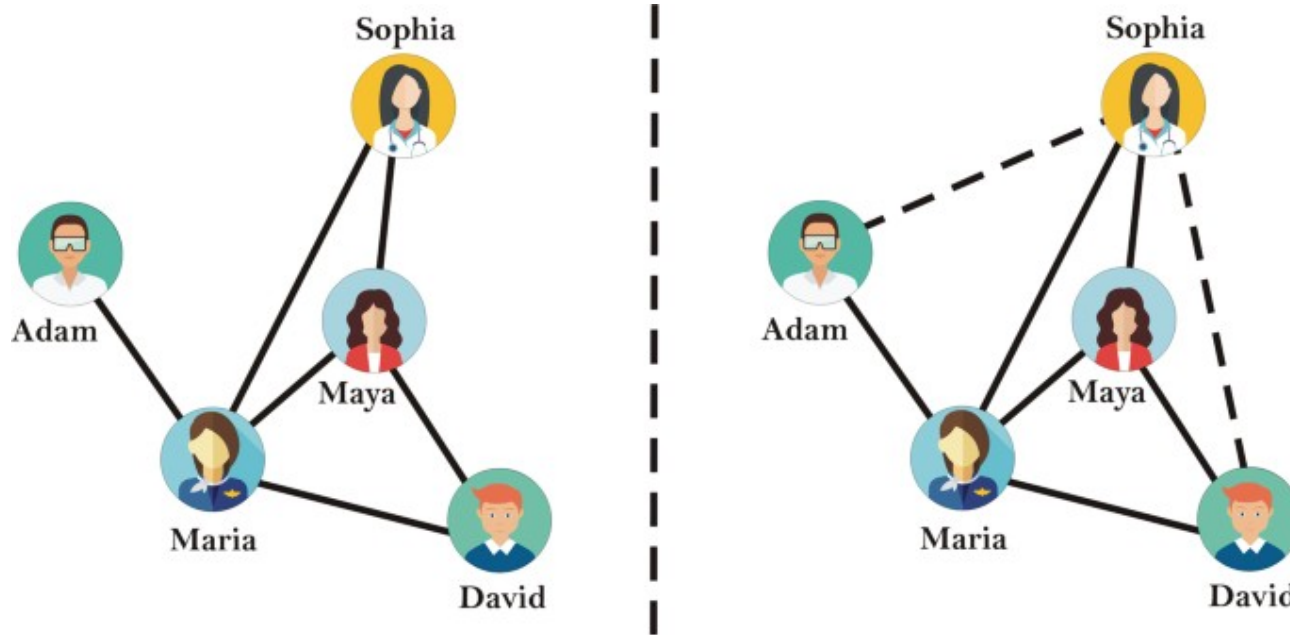
Existem várias estratégias para a extração de *features*

Olhando no grafo t=2

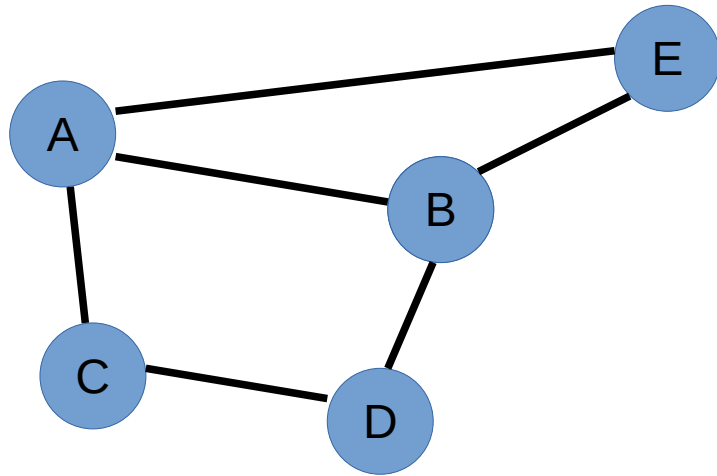
Features	Rótulo
Features do par Sofia-Adam	1
Features do par Sofia-David	1
Features do par David-Adam	0



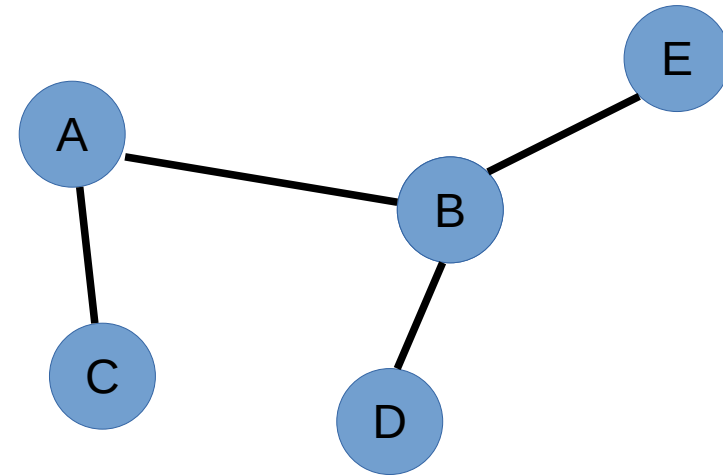
Tipicamente só temos um snapshot



Mas podemos simular isso



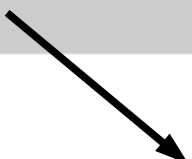
Rede original



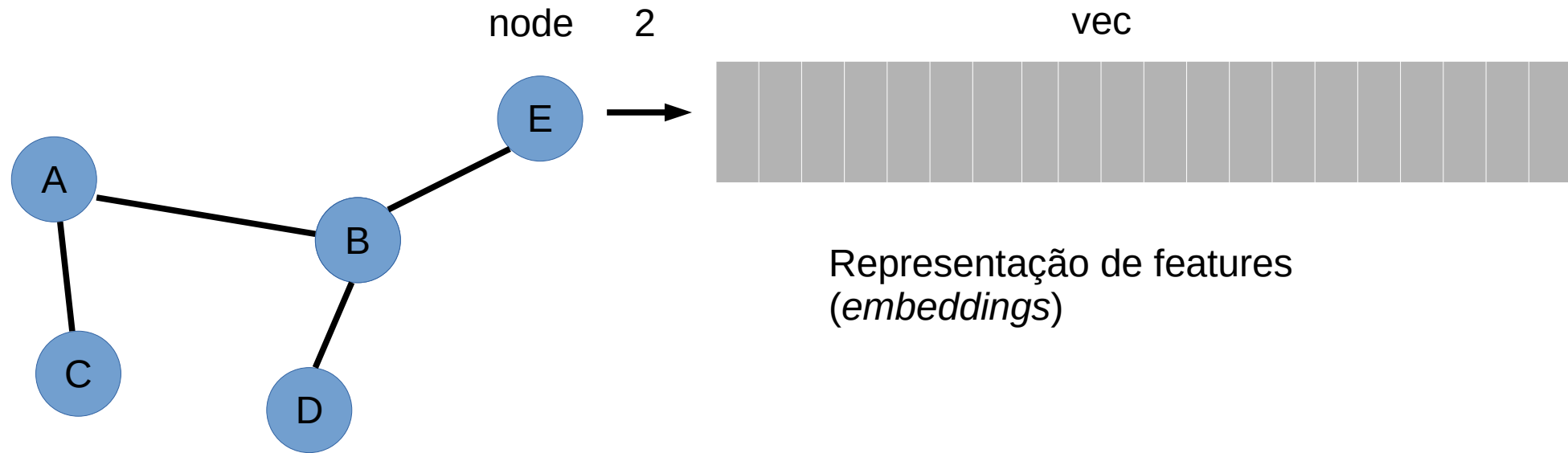
Rede com arestas removidas e reservadas.
Assim podemos conferir as previsões

Features	Rótulo
Features do par A-E	1
Features do par C-D	1
Features do par A-D	0
Features do par B-C	0
Features do par D-E	0

<i>Features</i>	Rótulo
Features do par A-E	1
Features do par C-D	1
Features do par A-D	0
Features do par B-C	0
Features do par D-E	0



O que poderiam ser features nesse problema?

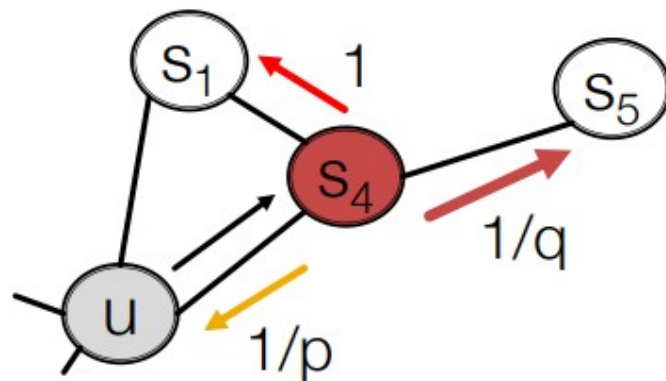


Mapeamos cada nó da rede em um espaço n-dimensional de features

A similaridade entre os nós indica a força do link

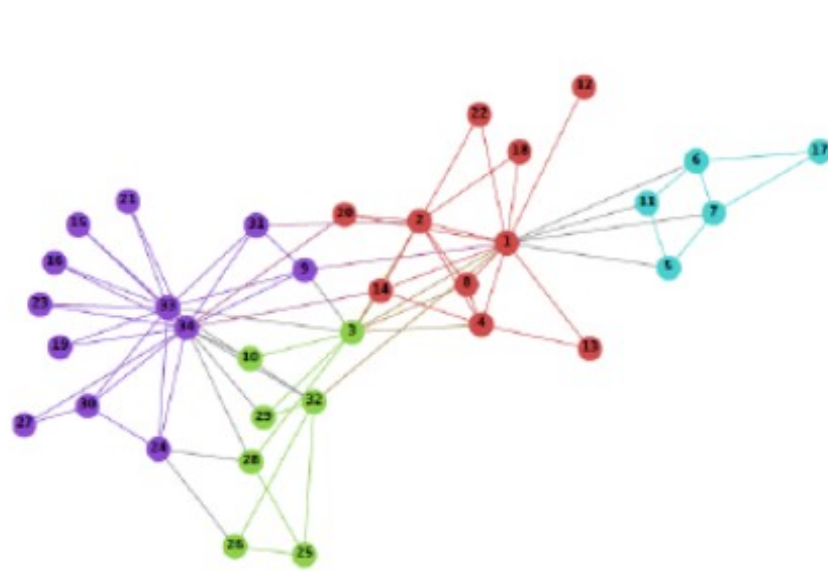
Baseado em *random walks*

Usa essa ideia para computar os *embeddings*

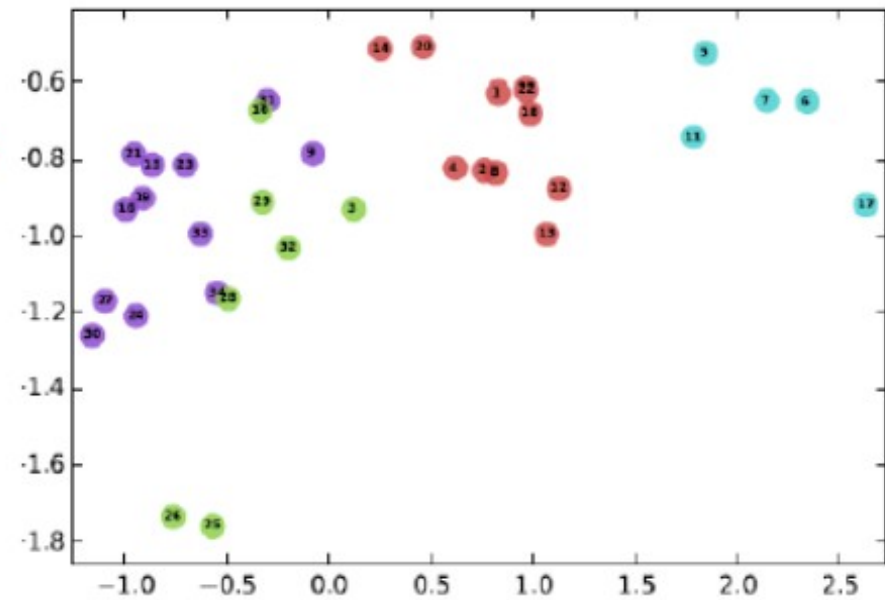


$u \rightarrow s_4 \rightarrow ?$





Input



Output

