

Evolução de grafos de genealogia acadêmica

Arthur V. Kamienski
Orientador: Jesús P. Mena-Chalco

Universidade Federal do ABC
Centro de Matemática, Computação e Cognição
Bacharelado em Ciência da Computação

Agenda

1. Introdução
2. Referencial teórico
3. Conjunto de dados utilizado
4. Método
5. Resultados
6. Considerações finais
7. Referências bibliográficas

Agenda

1. Introdução
2. Referencial teórico
3. Conjunto de dados utilizado
4. Método
5. Resultados
6. Considerações finais
7. Referências bibliográficas

- Genealogia acadêmica
 - Relações de orientação entre membros da comunidade científica
 - Fluxo de conhecimento científico

- Informação temporal
 - Geralmente apresentada de forma estática
 - Desenvolvimento da comunidade ao longo do tempo
 - I. Análise do passado
 - II. Previsões para o futuro

- Desenvolvimento de um método computacional de predição do crescimento de grafos de genealogia acadêmica
 - Gerar conhecimento sobre a maneira como essa evolução é dada
 - Estimar um possível cenário futuro de uma comunidade de pesquisadores

- Identificar características de evolução para comunidades de pesquisadores
- Explorar diferentes técnicas de análise de evolução de grafos, buscando identificar aquela mais adequado à tarefa proposta
- Verificar relações entre diferentes áreas do conhecimento e a forma como evoluem ao longo do tempo
- Estimar o cenário científico brasileiro para diferentes momentos no futuro

Agenda

1. Introdução
2. Referencial teórico
3. Conjunto de dados utilizado
4. Método
5. Resultados
6. Considerações finais
7. Referências bibliográficas

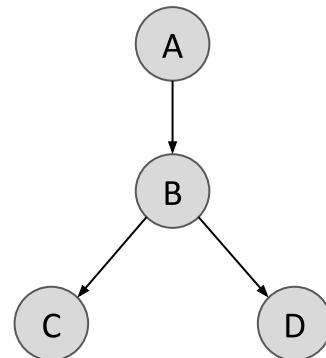
- Modelagem

- Pesquisadores → vértices
 - I. Atributos: nome, área, ano de formação, etc.
- Orientações acadêmicas → arestas direcionadas
 - I. Atributos: tipo, ano de início, ano de término, etc.

Grafos de genealogia acadêmica

- Exemplo:

- Vértices → pesquisadores: A, B, C e D
- Areias → orientações (A, B); (B, C) e (B, D)

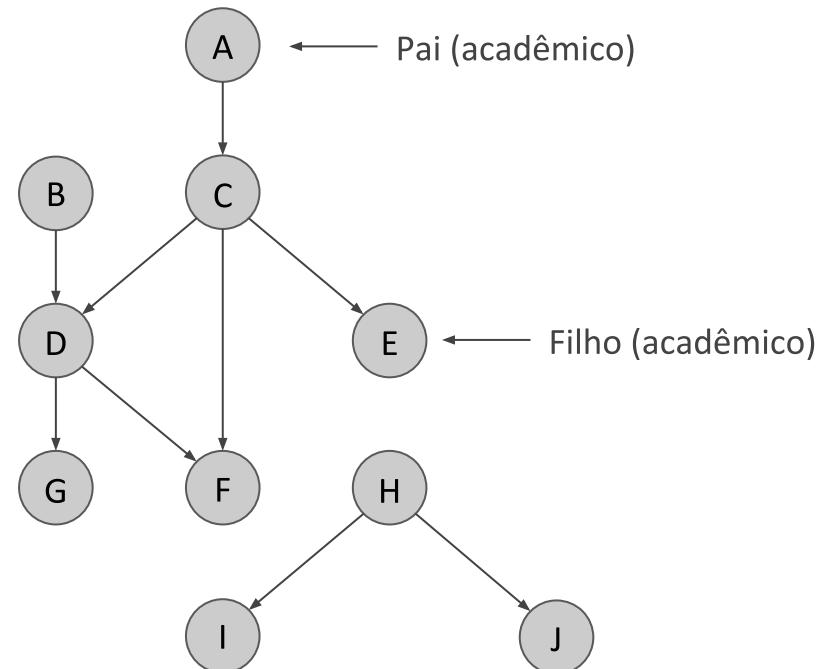


Grafos de genealogia acadêmica

- Estrutura

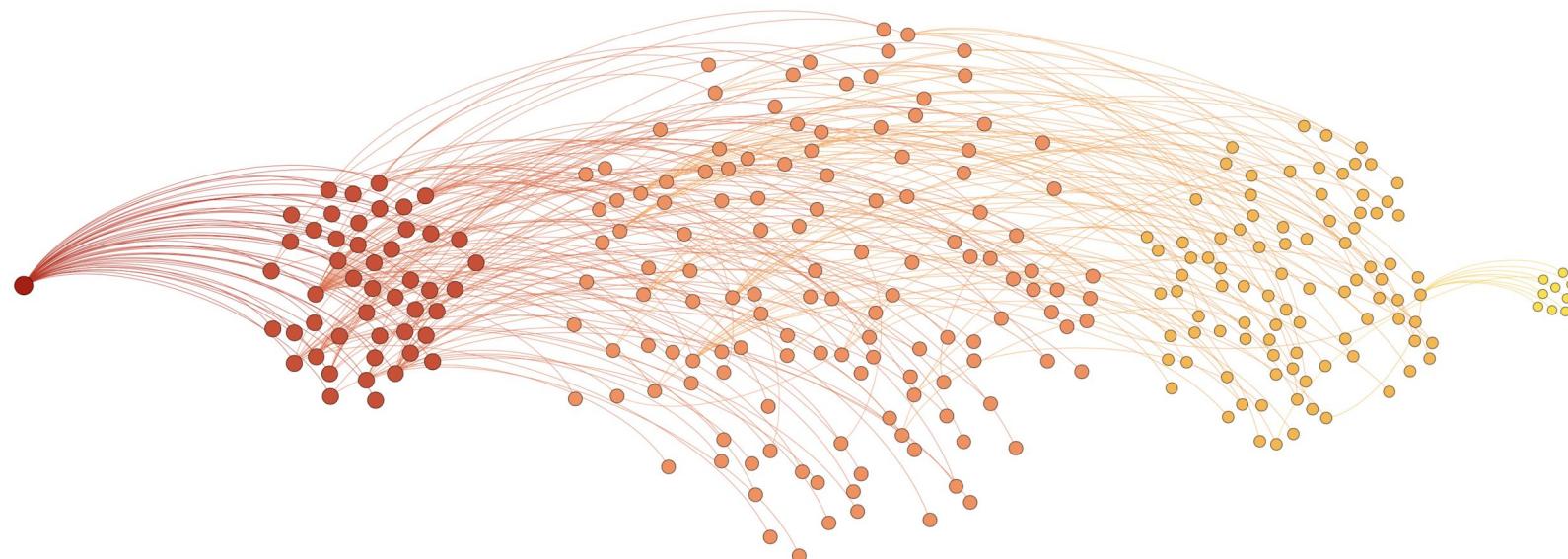
- Similar a uma árvore, porém:
 - I. Grafo desconexo
 - II. Podem existir ciclos

- Uso de termos da genealogia



Grafos de genealogia acadêmica

- Exemplo:
 - Descendentes de Jacob Palis



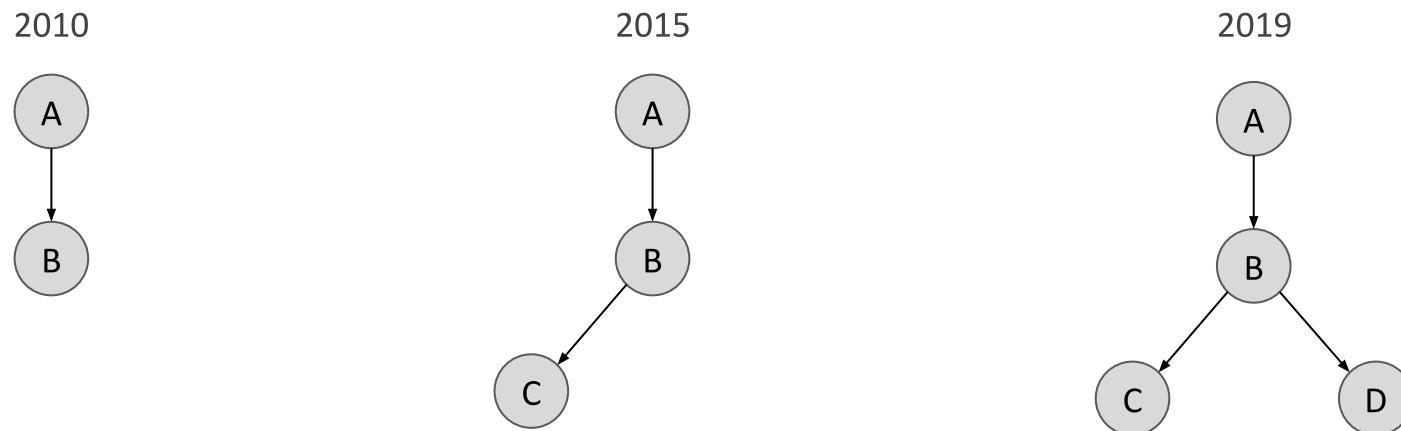
- Informação temporal presente nos grafos
 - Vértices → pesquisadores
 - I. Atributos: nome, área, **ano de formação**, etc.
 - Arestras → orientações acadêmicas
 - I. Atributos: tipo, **ano de início**, **ano de término**, etc.

- Ordem de acontecimento de eventos
 - Surgimento ou desaparecimento de vértices e arestas
 - Alteração do formato do grafo
 - Evolução temporal

Grafos evolutivos

- Exemplo:

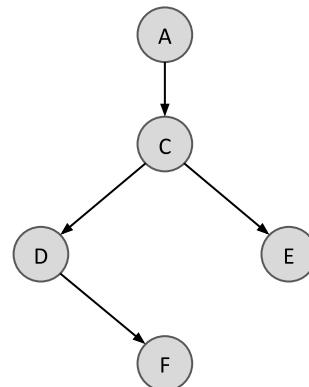
- Vértices → pesquisadores: A, B, C e D
- Areias → orientações (A, B) em 2010; (B, C) em 2015 e (B, D) em 2019



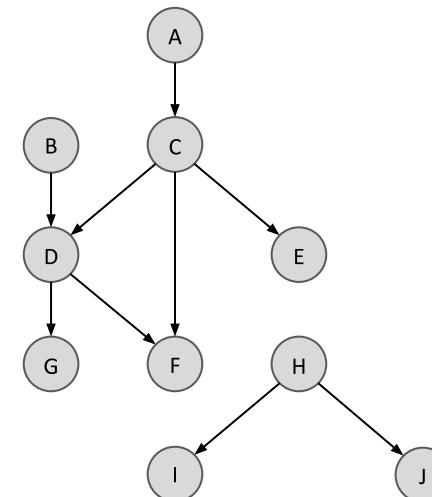
Grafos evolutivos

- Exemplo:

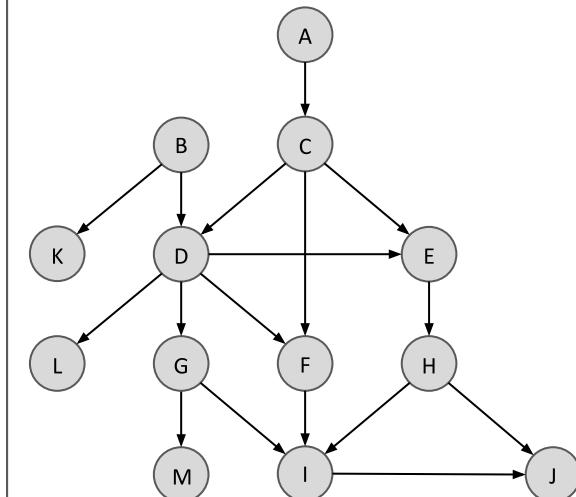
$t = 0$



$t = 1$

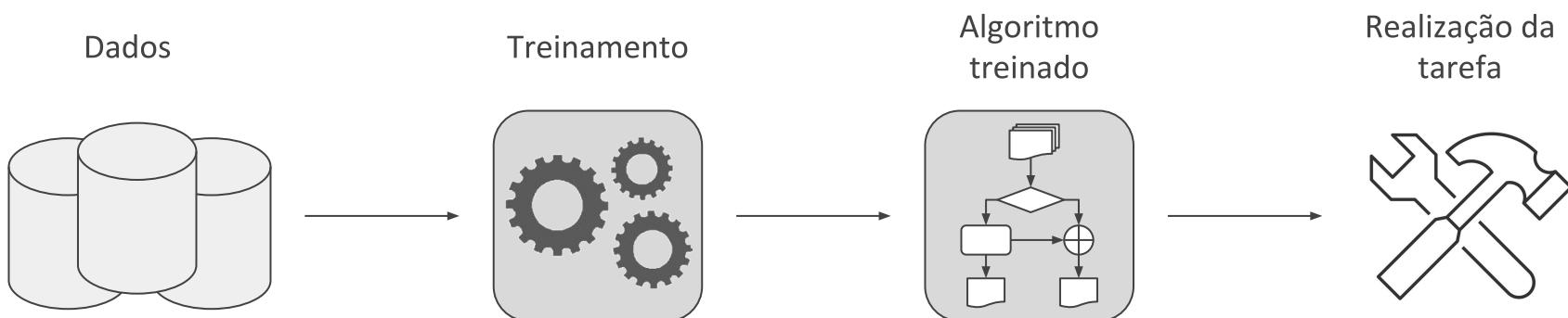


$t = 2$



- Algoritmos que aprendem com a experiência
 - Utilizar informações do passado para realizar uma tarefa
- Principais elementos:
 - Tarefa bem definida
 - Fonte de informações
 - Medida de desempenho

Aprendizado de máquina



- Tipos de aprendizado
 - Supervisionado
 - I. Classificação
 - II. Regressão
 - Não-supervisionado
 - Por reforço

- Árvores de decisão
 - Algoritmo de aprendizado supervisionado para classificação*
 - Criação de regras (sim/não) baseadas nos dados de entrada
 - I. Melhores decisões possíveis para a separação das classes

Árvores de decisão

| Altura | Pelo | Cor | Raça |
|--------|-------|--------|------------|
| 57cm | Curto | Preto | Rottweiler |
| 65cm | Curto | Preto | Rottweiler |
| 60cm | Longo | Preto | Collie |
| 56cm | Longo | Branco | Collie |
| 25cm | Curto | Bege | Pug |
| 23cm | Curto | Preto | Pug |
| 22cm | Curto | Branco | Shih Tzu |
| 28cm | Longo | Branco | Shih Tzu |



Agenda

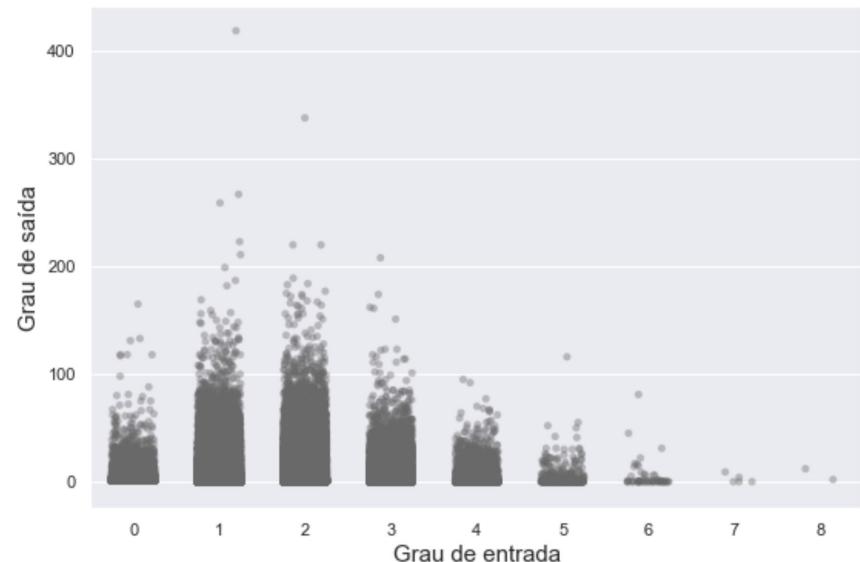
1. Introdução
2. Referencial teórico
3. Conjunto de dados utilizado
4. Método
5. Resultados
6. Considerações finais
7. Referências bibliográficas

- Plataforma Acácia

- <http://plataforma-acacia.org>
- 6,300,000 currículos de 1900 à 2019
- 1,272,590 pesquisadores
- 1,404,109 orientações
 - I. 995,807 mestrado
 - II. 371,074 doutorado
 - III. 37,228 pós-doutorado

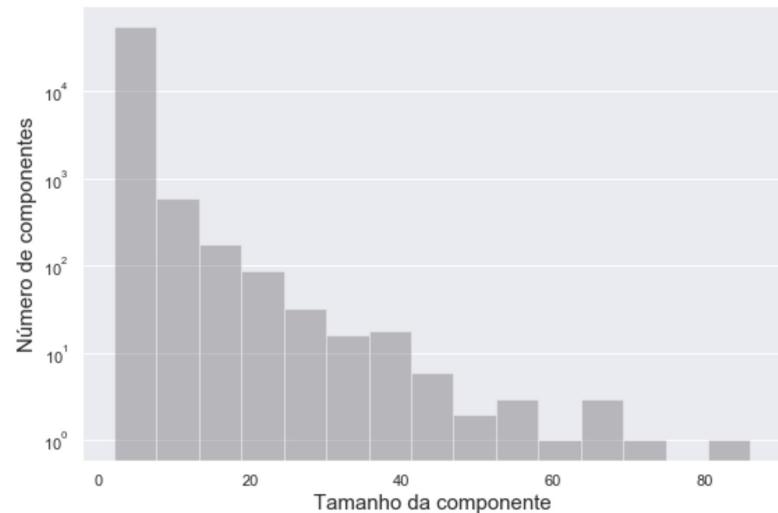
Grafo de genealogia acadêmica da Plataforma Acácia

- Vértices → 1,272,590
 - 130,266 raízes
- Arestas → 1,404,109
 - 1.1 arestas por vértice
 - Grau de entrada = 1 → 914,932
 - Grau de entrada = 2 → 197,208



Grafo de genealogia acadêmica da Plataforma Acácia

- 56,263 componentes conexas
 - Uma grande componente → 89.2%
 - 46,868 com apenas 2 vértices



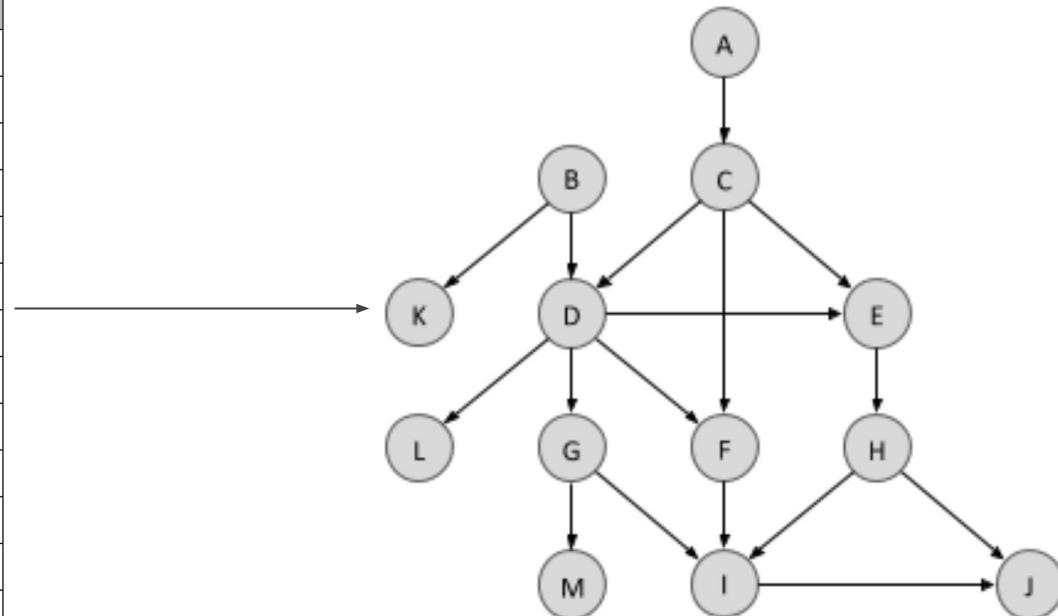
Agenda

1. Introdução
2. Referencial teórico
3. Conjunto de dados utilizado
4. Método
5. Resultados
6. Considerações finais
7. Referências bibliográficas

Processamento do grafo

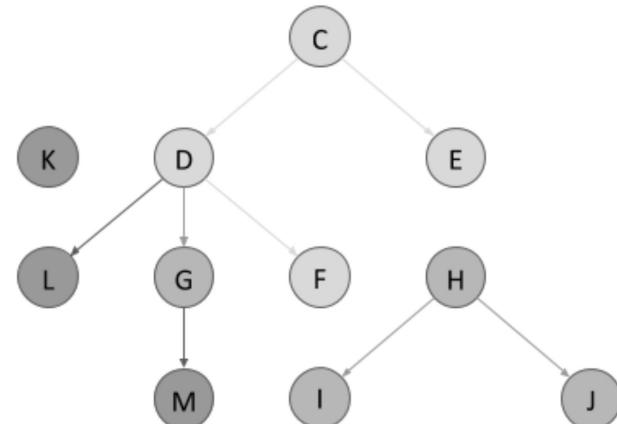
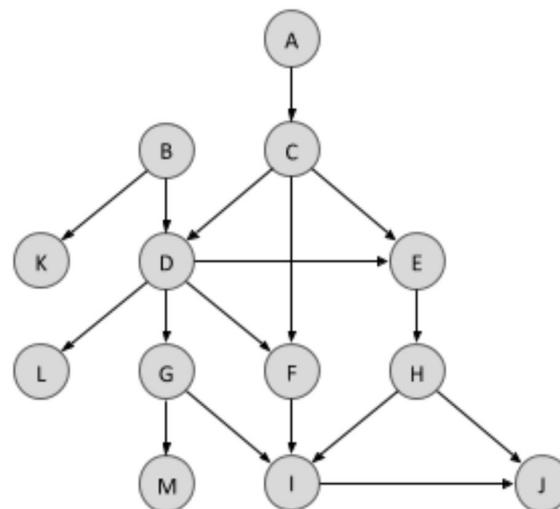
- Representação do grafo
 - Entidades separadas para vértices e arestas

| Vértice | Arestas (vértice, tempo) |
|---------|--------------------------------|
| A | (C, 0) |
| B | (D, 1), (K, 2) |
| C | (D, 0), (E, 0), (F, 1) |
| D | (F, 0), (G, 1), (E, 2), (L, 2) |
| E | (H, 2) |
| F | (I, 2) |
| G | (M, 2), (I, 2) |
| H | (I, 1), (J, 1) |
| I | (J, 2) |
| J | |
| K | |
| L | |
| M | |



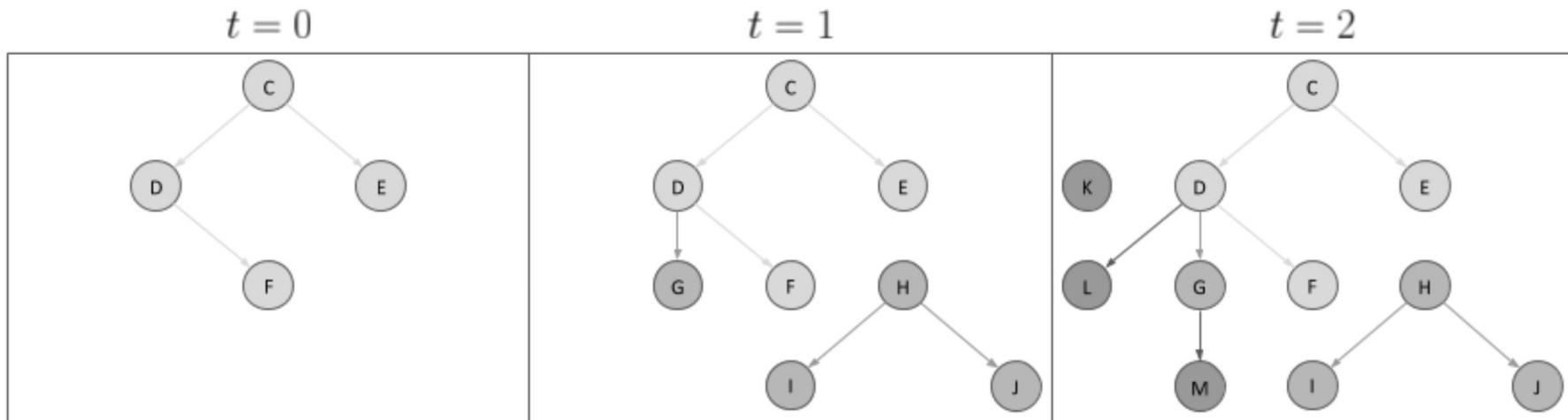
Processamento do grafo

- Remoção de arestas "secundárias"
 - Manter arestas que geram novos vértices
- Remoção de raízes



Processamento do grafo

- Extração da informação temporal
 - Observação do grafo em diferentes momentos
 - Extração de informações para cada vértice



Processamento do grafo

- Extração da informação temporal
 - Observação do grafo em diferentes momentos
 - Extração de informações para cada vértice

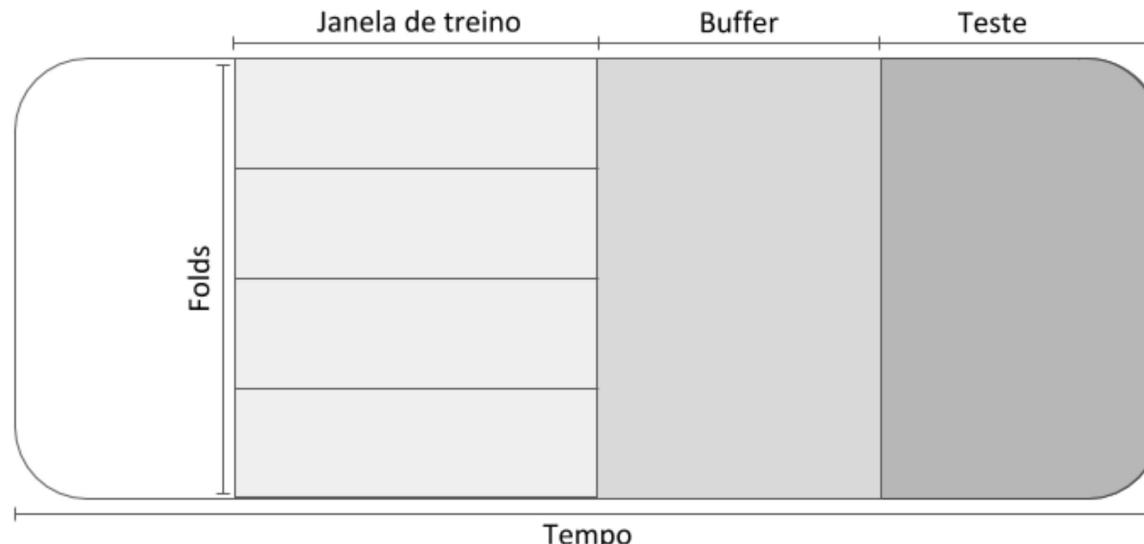
| Vértice | Tempo | Pai | Idade | Filhos | Irmãos | Descendentes |
|---------|-------|-----|-------|--------|--------|--------------|
| C | 0 | - | 0 | 2 | 0 | 3 |
| D | 0 | C | 0 | 1 | 1 | 1 |
| E | 0 | C | 0 | 0 | 1 | 0 |
| F | 0 | D | 0 | 0 | 0 | 0 |
| C | 1 | - | 1 | 2 | 0 | 4 |
| D | 1 | C | 1 | 2 | 1 | 2 |
| E | 1 | C | 1 | 0 | 1 | 0 |
| F | 1 | D | 1 | 0 | 1 | 0 |
| G | 1 | D | 0 | 0 | 1 | 0 |
| H | 1 | - | 0 | 2 | 0 | 2 |
| I | 1 | H | 0 | 0 | 1 | 0 |
| J | 1 | H | 0 | 0 | 1 | 0 |
| C | 2 | - | 2 | 2 | 0 | 6 |
| D | 2 | C | 2 | 3 | 1 | 4 |
| E | 2 | C | 2 | 0 | 1 | 0 |
| F | 2 | D | 2 | 0 | 2 | 0 |
| G | 2 | D | 1 | 1 | 2 | 1 |
| H | 2 | - | 1 | 2 | 0 | 2 |
| I | 2 | H | 1 | 0 | 1 | 0 |
| J | 2 | H | 1 | 0 | 1 | 0 |
| K | 2 | - | 0 | 0 | 0 | 0 |
| L | 2 | D | 0 | 0 | 2 | 0 |
| M | 2 | G | 0 | 0 | 0 | 0 |

- Definição do problema
 - Aprendizado supervisionado para classificação
 - Prever se um pesquisador terá novos filhos
 - I. Futuro k anos distante

Treinamento de modelos preditivos

- Treinamento

- Separação por tempo do conjunto de dados em treino e teste
- Treino dividido em *folds* para validação cruzada
- Zona de *buffer* para evitar vazamento de informações
- Janela de treino para o passado

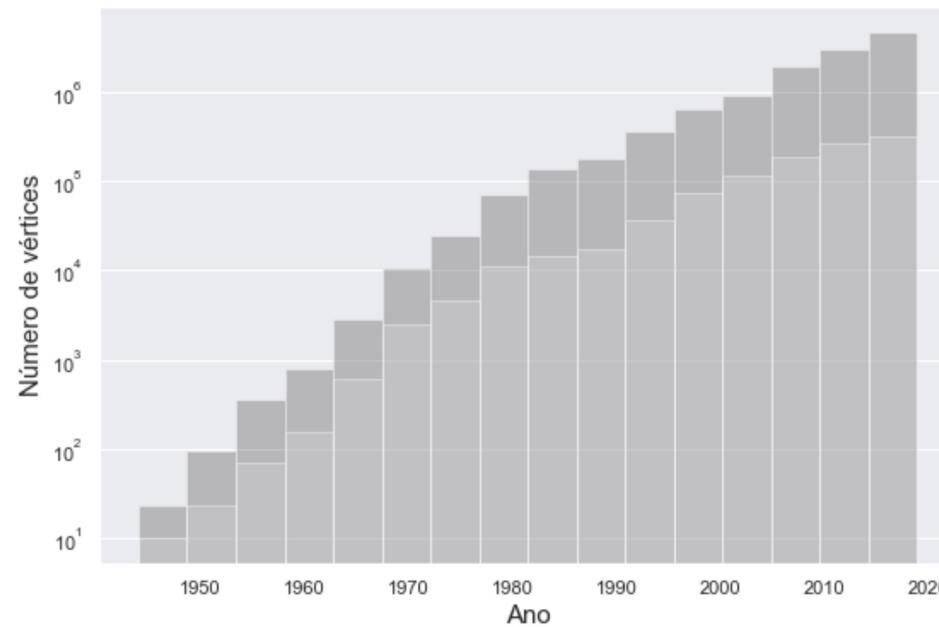


Agenda

1. Introdução
2. Referencial teórico
3. Conjunto de dados utilizado
4. Método
5. Resultados
6. Considerações finais
7. Referências bibliográficas

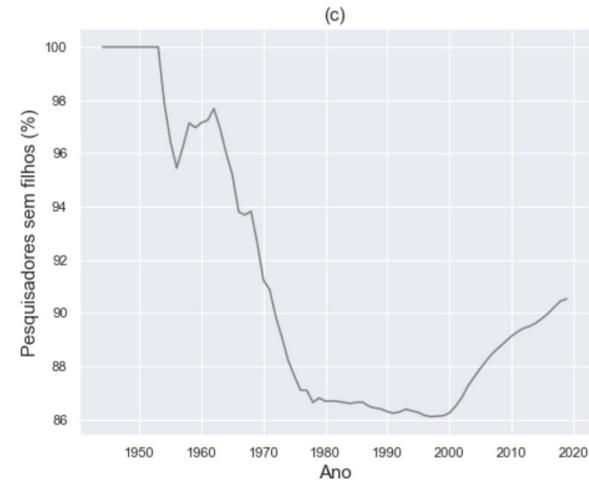
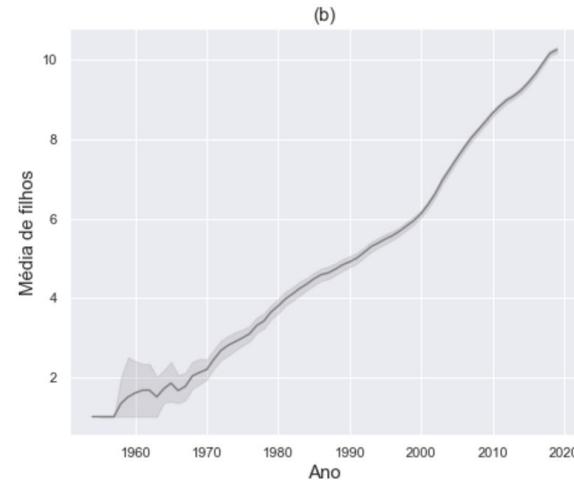
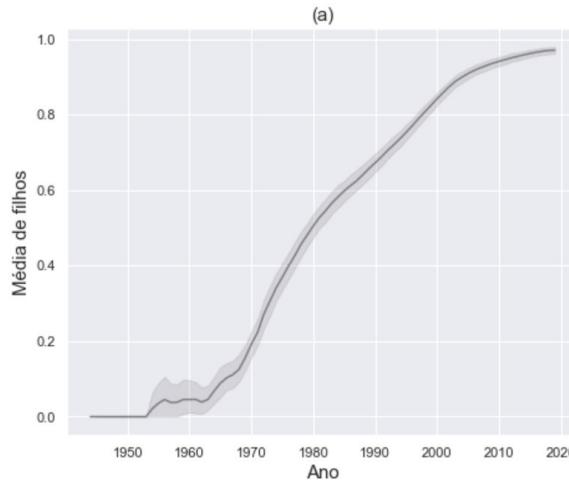
Análise do crescimento

- Crescimento exponencial
 - Diminuição da produção nos últimos anos



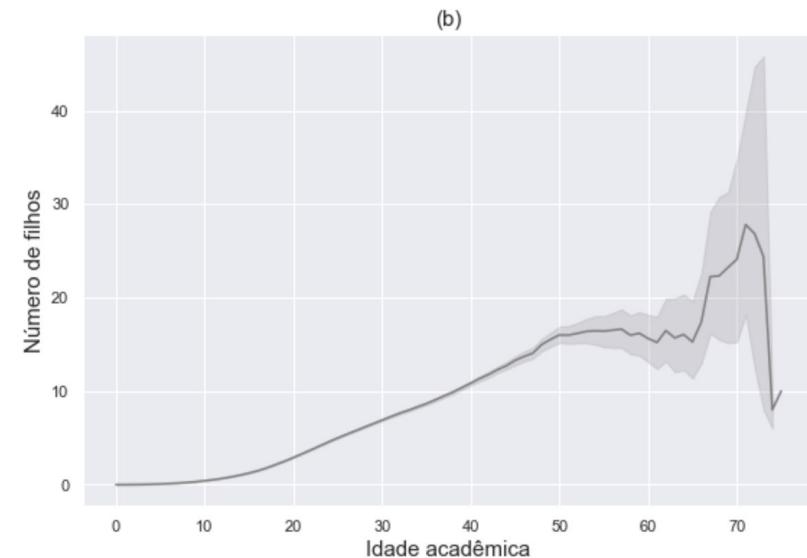
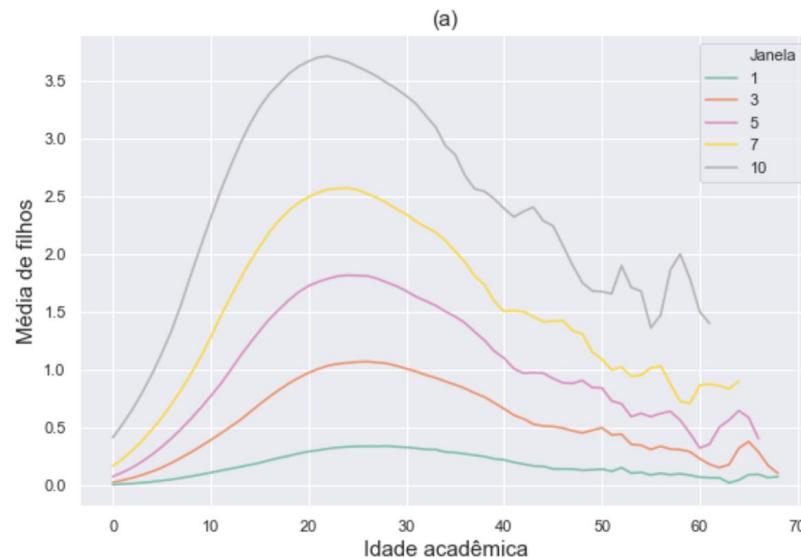
Análise do crescimento

- Crescimento exponencial
 - Diminuição da produção nos últimos anos



- Idade acadêmica

- Crescimento constante a partir de 10 anos
- Produção máxima 25 anos depois do ano de formação



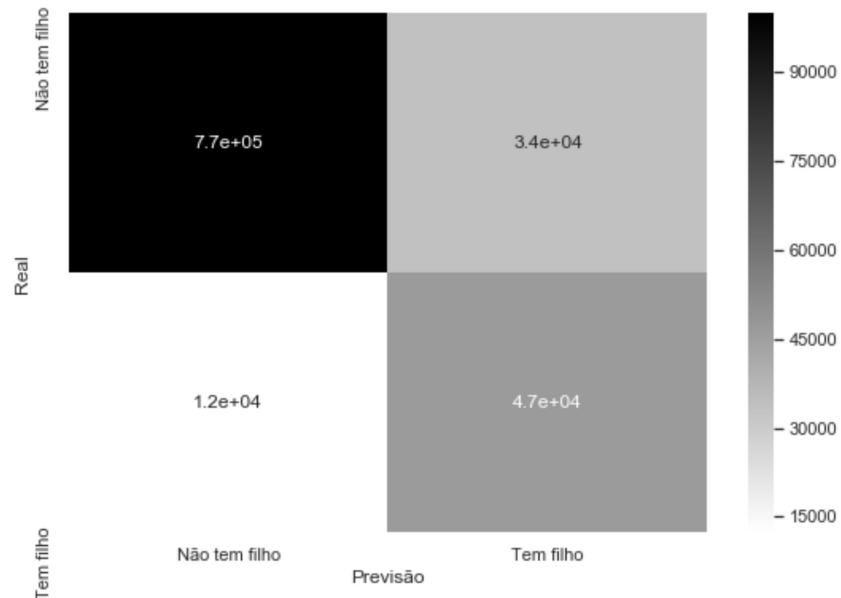
- Conjunto de treinamento
 - Base de treino
 - I. 87.2% do total
 - II. 11.5% de amostras "tem filho"
 - Base de teste
 - I. 6.5% de amostras "tem filho"

- Árvore de decisão
 - Validação cruzada:
 - I. Profundidade = 6
 - II. Número máximo de folhas = 15
 - III. Número mínimo de amostras por folha = 2%

Modelo preditivo

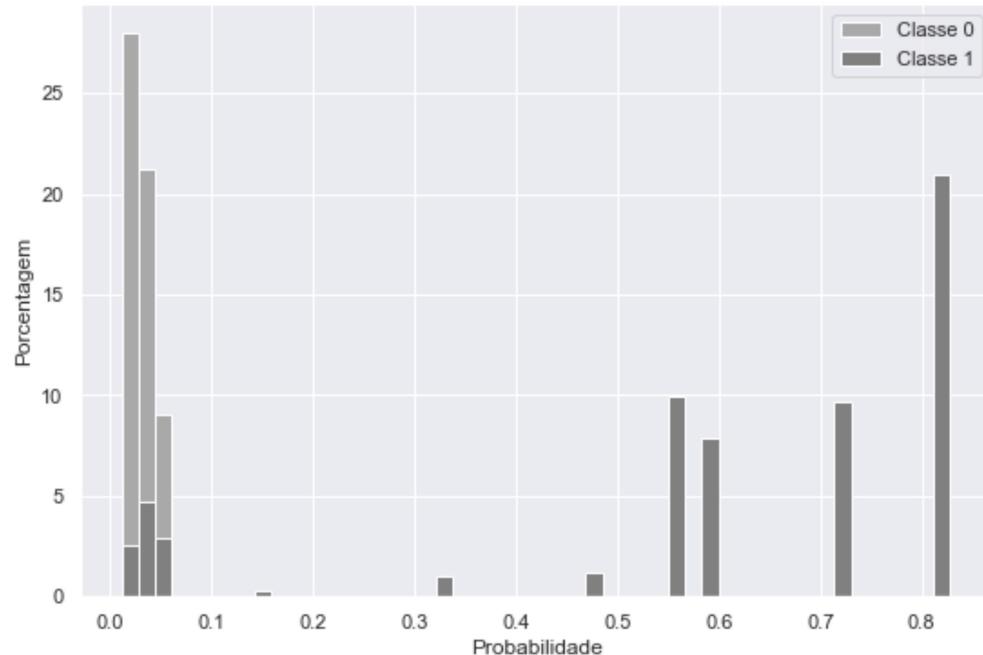
- Desempenho

- 94.5% de acurácia
- 58% de precisão
- 79% de revocação



Modelo preditivo

- Desempenho
 - 0.92 de AUC ROC



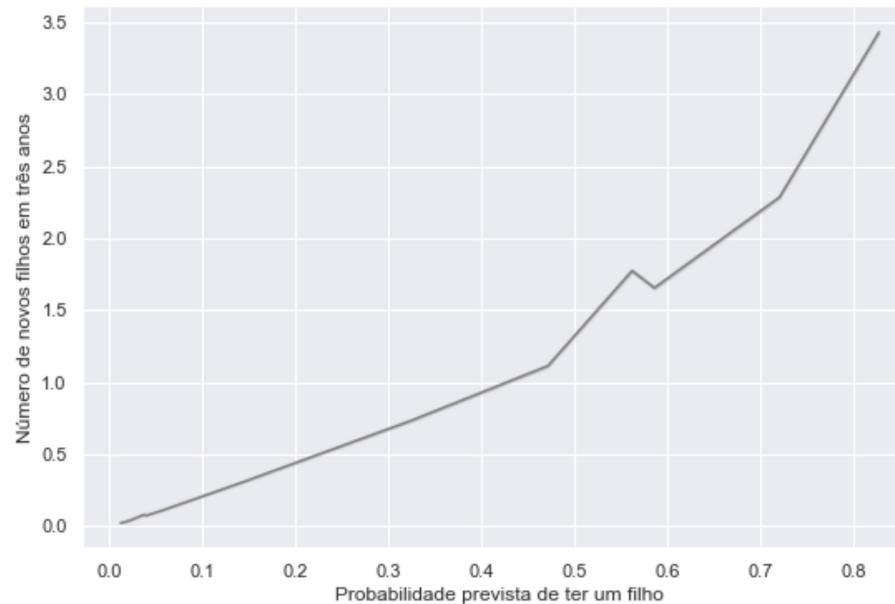
Característica de pais acadêmicos

- Número de filhos
- Idade

| Filhos | Ascendentes | Idade | Sobrinhos | Irmãos | % do total | % classe “tem filho” |
|--------------|-------------|---------------|-----------|-----------|------------|----------------------|
| >5 | - | ≤ 25 | - | - | 3.3 | 82.8 |
| >2, ≤ 5 | - | ≤ 25 | - | - | 2.4 | 72.1 |
| ≤ 2 | - | - | - | >0 | 2.0 | 58.6 |
| >2 | - | >25 | - | - | 2.2 | 58.2 |
| ≤ 2 | - | - | - | 0 | 2.0 | 47.2 |
| 0 | 0 | >7 | - | - | 2.6 | 32.6 |
| 0 | 0 | ≤ 7 | - | - | 3.5 | 14.3 |
| 0 | >0 | $>4, \leq 19$ | >0 | ≤ 27 | 11.7 | 5.5 |
| 0 | >0 | $>6, \leq 19$ | 0 | - | 12.7 | 3.9 |
| 0 | >0 | $>1, \leq 4$ | >0 | - | 5.7 | 3.8 |
| 0 | >0 | $>4, \leq 19$ | >0 | >27 | 6.0 | 3.6 |
| 0 | >0 | $>4, \leq 6$ | 0 | - | 7.3 | 2.6 |
| 0 | >0 | ≤ 1 | >0 | - | 3.7 | 2.3 |
| 0 | >0 | >19 | - | - | 4 | 1.6 |
| 0 | >0 | ≤ 4 | 0 | - | 30.8 | 1.2 |

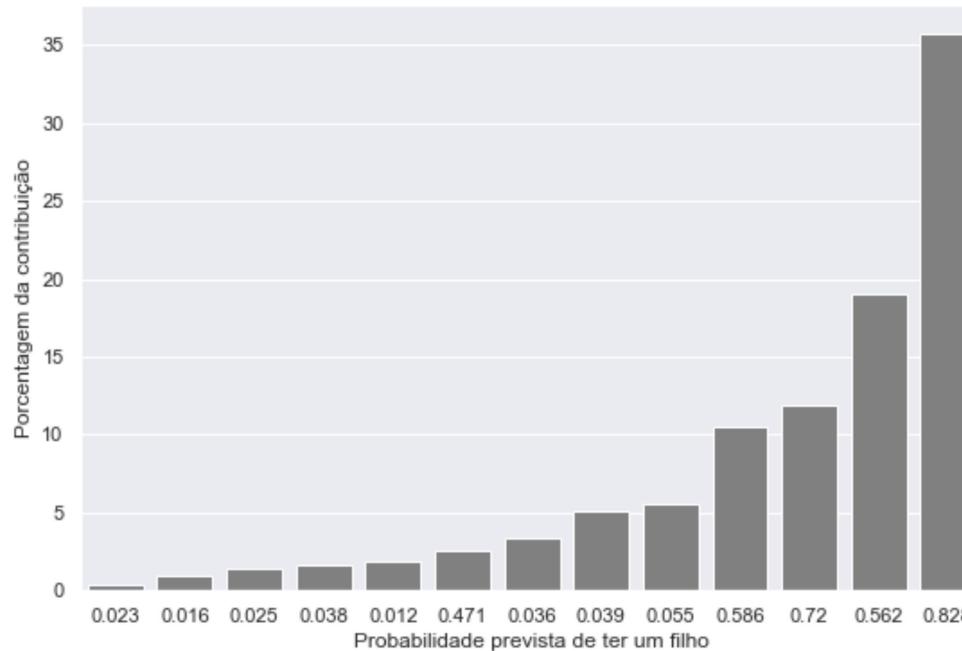
- 95 mil pesquisadores geram novos filhos
 - Probabilidade de gerar um filho * número de pesquisadores

- 280 mil novos pesquisadores
 - Probabilidade de gerar um filho * número de pesquisadores * média de filhos



Crescimento futuro

- Pesquisadores com menos de 25 anos e mais de um filho são os que mais contribuem



Agenda

1. Introdução
2. Referencial teórico
3. Conjunto de dados utilizado
4. Método
5. Resultados
6. Considerações finais
7. Referências bibliográficas

- Método para análise e predição de crescimento
- Características de pesquisadores que produzem novos pesquisadores
- Cenário da comunidade acadêmica em 3 anos

Limitações e pesquisas futuras

- Uso de todas as arestas
- Diferenciação das orientações
- Uso de outros tipos de variáveis
- Métodos de predição com maior desempenho
- Tarefa de regressão
- Previsões para outros anos futuros

Agenda

1. Introdução
2. Referencial teórico
3. Conjunto de dados utilizado
4. Método
5. Resultados
6. Considerações finais
7. Referências bibliográficas

Referências Bibliográficas

- Aggarwal, Charu, and Karthik Subbian. "Evolutionary network analysis: A survey." ACM Computing Surveys (CSUR) 47.1 (2014): 10.
- Damaceno, Rafael JP, et al. "The Brazilian academic genealogy: evidence of advisor–advisee relationships through quantitative analysis." Scientometrics 119.1 (2019): 303-333.
- Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.



UFABC