NETFLIX

PRIZE

Análise de redes de interesses em filmes

Fábio Salomão Vinco e Silva
Juan Sebastián Beleño Díaz
Samuel Gomes Fadel

O “Netflix Prize”

- Desafio lançado pela **Netflix** em 2006 (concluído em 2009)
- Estimar a avaliação de filmes por um usuário
  - A partir do histórico de avaliações deste
  - Melhora de pelo menos 10% (em relação ao algoritmo original)
- O conjunto de dados
  - ~17 mil filmes
  - ~480 mil usuários
  - ~100 milhões de avaliações — números de 1 a 5 (coletadas de 1998 a 2005)

Hipótese

- As avaliações refletem interesses por gêneros de filmes

Problema

Quais são as características de uma rede de interesses em filmes?

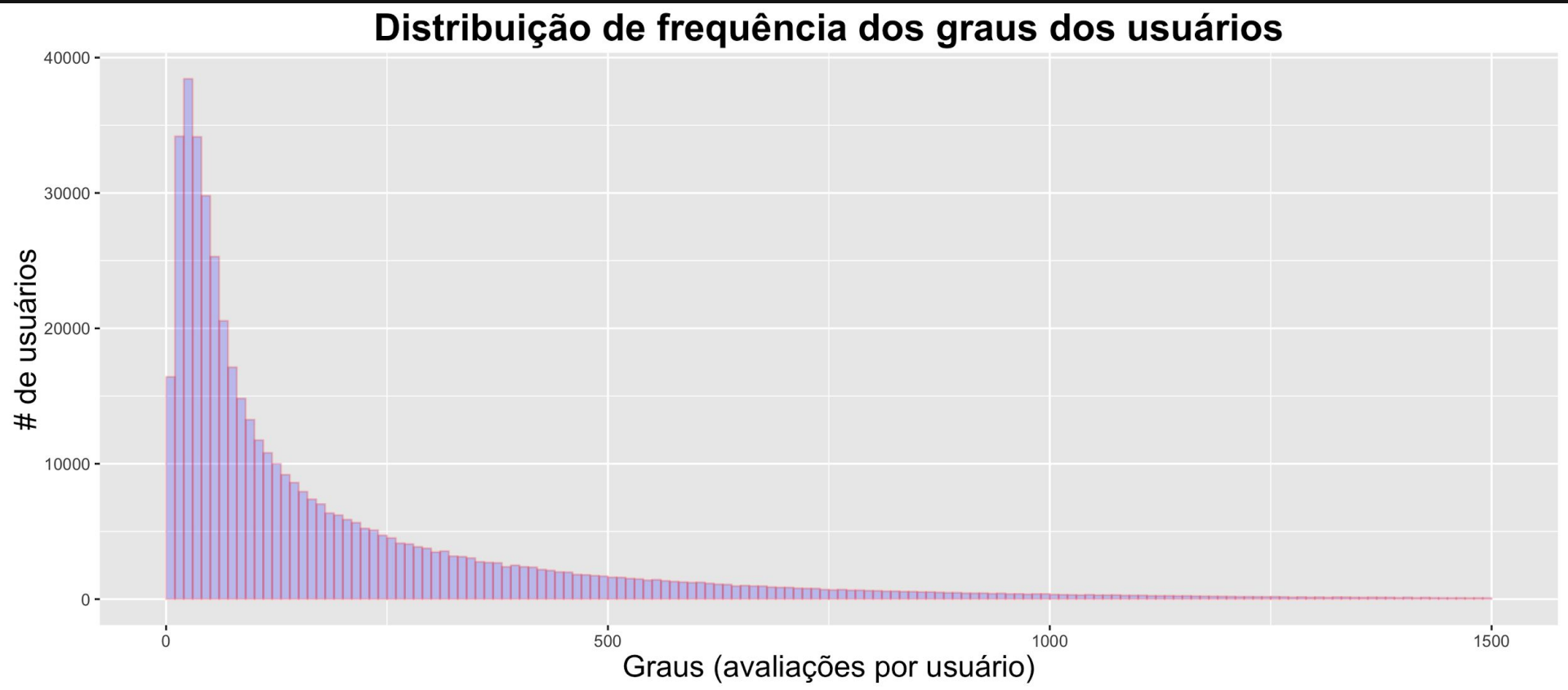
- Estruturas de comunidades
- Relação dos gêneros com a formação de comunidades

Rede de interesses

- Rede de interesses baseada em gêneros
- Gêneros obtidos da Open Movie Database (OMBD)
  - Gêneros de 12.440 dos 17.770 filmes (68,8%)
  - ~89 milhões de avaliações das ~100 milhões

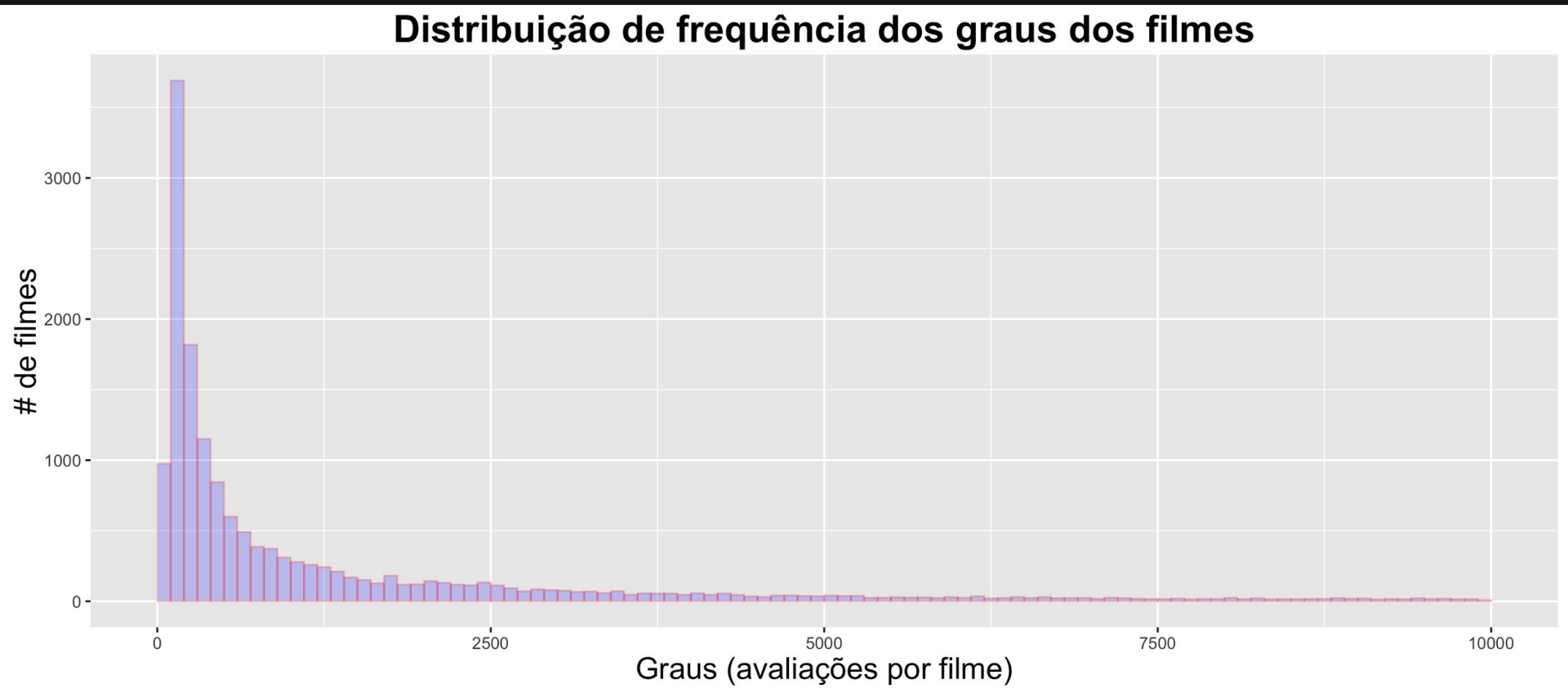
Rede de interesses

- Histograma de número de avaliações por usuário



Rede de interesses

- Histograma de número de avaliações por filme



Rede de interesses

- Duas abordagens
  - Informação agregada de avaliações por gênero
  - *Collaborative filtering*

- Investigar
  - Formação de comunidades
  - Comportamentos de avaliações dentro das comunidades

Rede de interesses

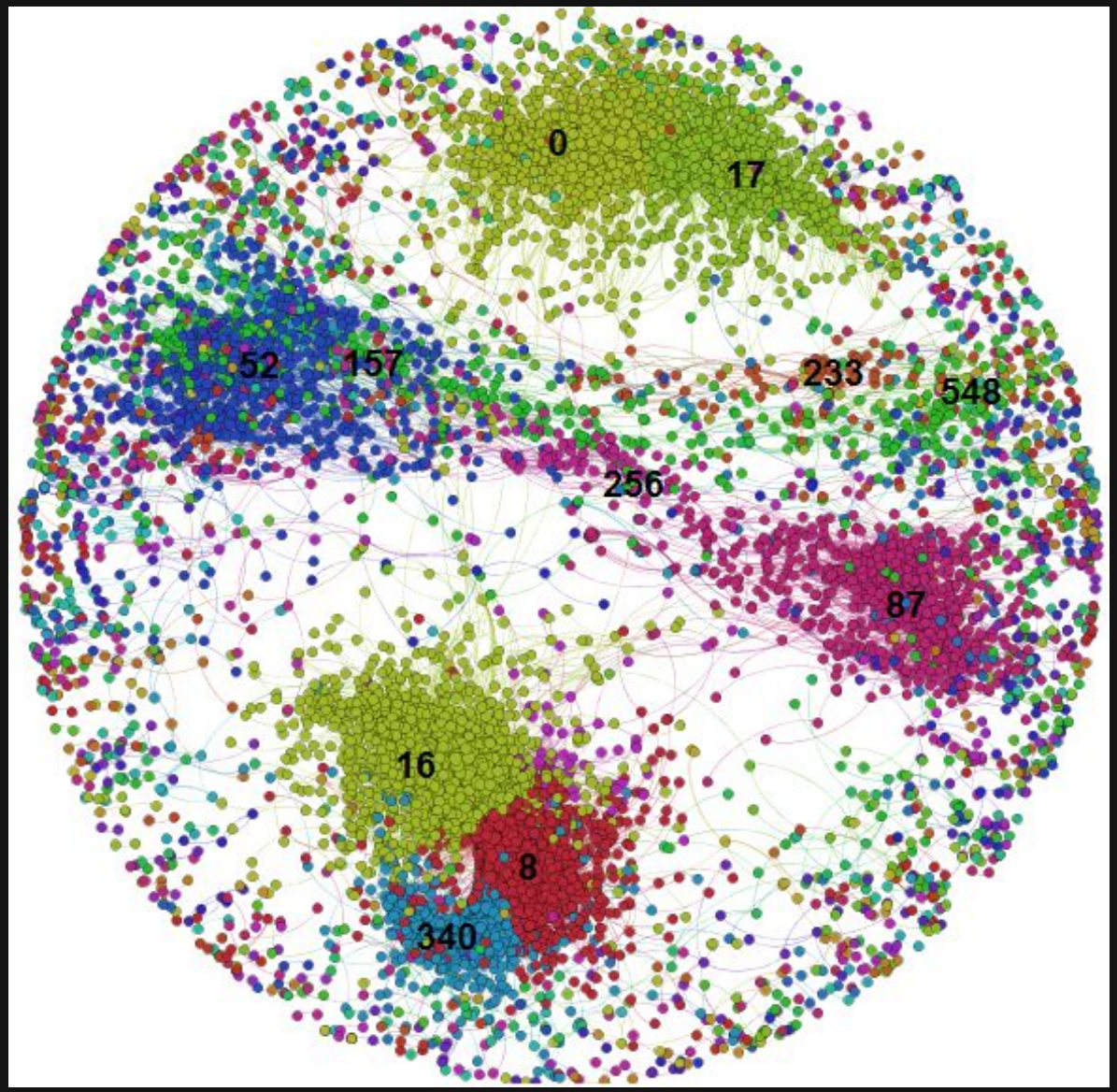
- Selecionados 50 mil usuários do total (~480 mil)
  - Selecionados uniformemente
  - Distribuição de avaliações similar à original

- Rede construída a partir das médias de avaliação
  - Cada usuário se torna um “vetor” de 22 dimensões
  - Cada dimensão é a média das avaliações para aquele gênero

- Ligação entre dois usuários com base em um limiar empírico
  - Dissimilaridade entre dois usuários < 15% da média das dissimilaridades
  - ~623 mil arestas

- Implementação em Python (+ algumas análises em R)

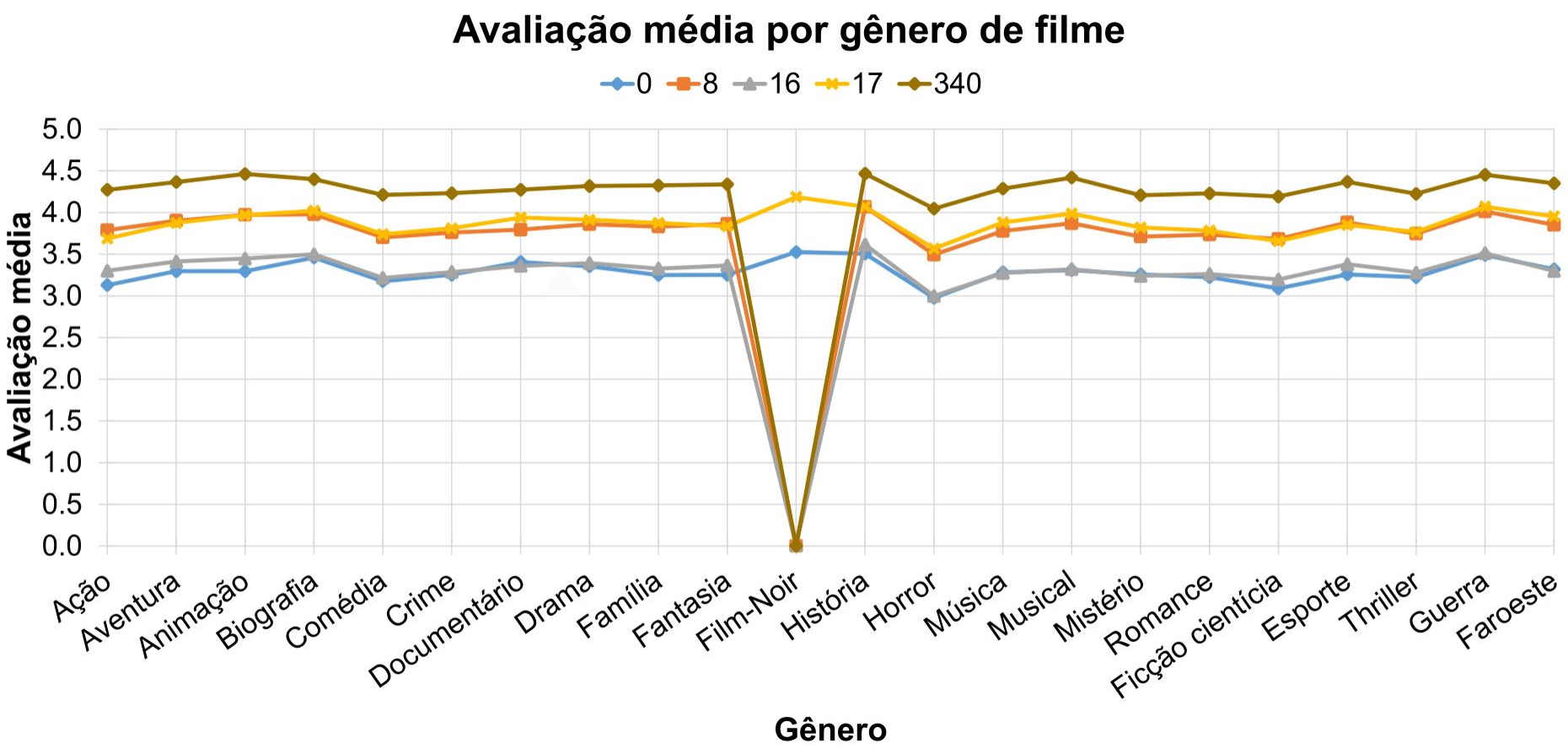
Rede de usuários — Avaliações Agregadas por Gênero



Rede de usuários — Avaliações Agregadas por Gênero

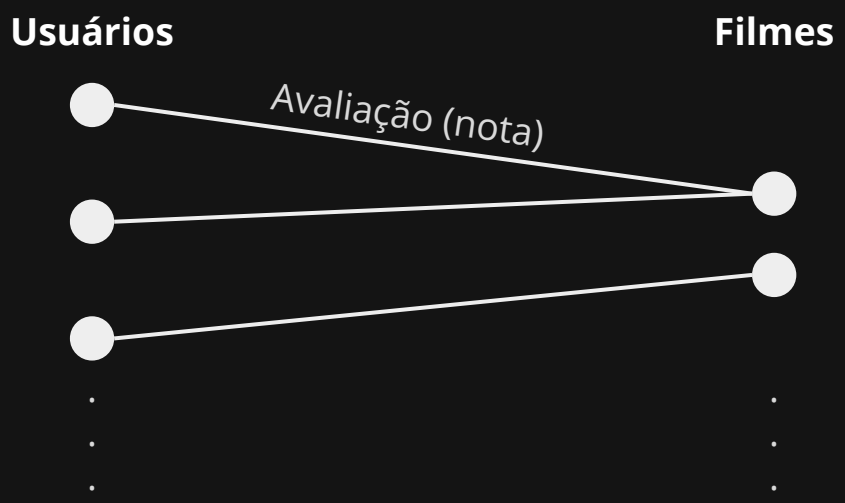
Comunidade	Avaliações/Usuário	# de usuários
16	1.234,42	2.949
18	1.089,77	2.283
0	1.752,35	2.192
17	1.430,44	1.907
340	1.032,08	956
Total (729 comunidades)	467, 26	50.000

Rede de usuários — Avaliações Agregadas por Gênero

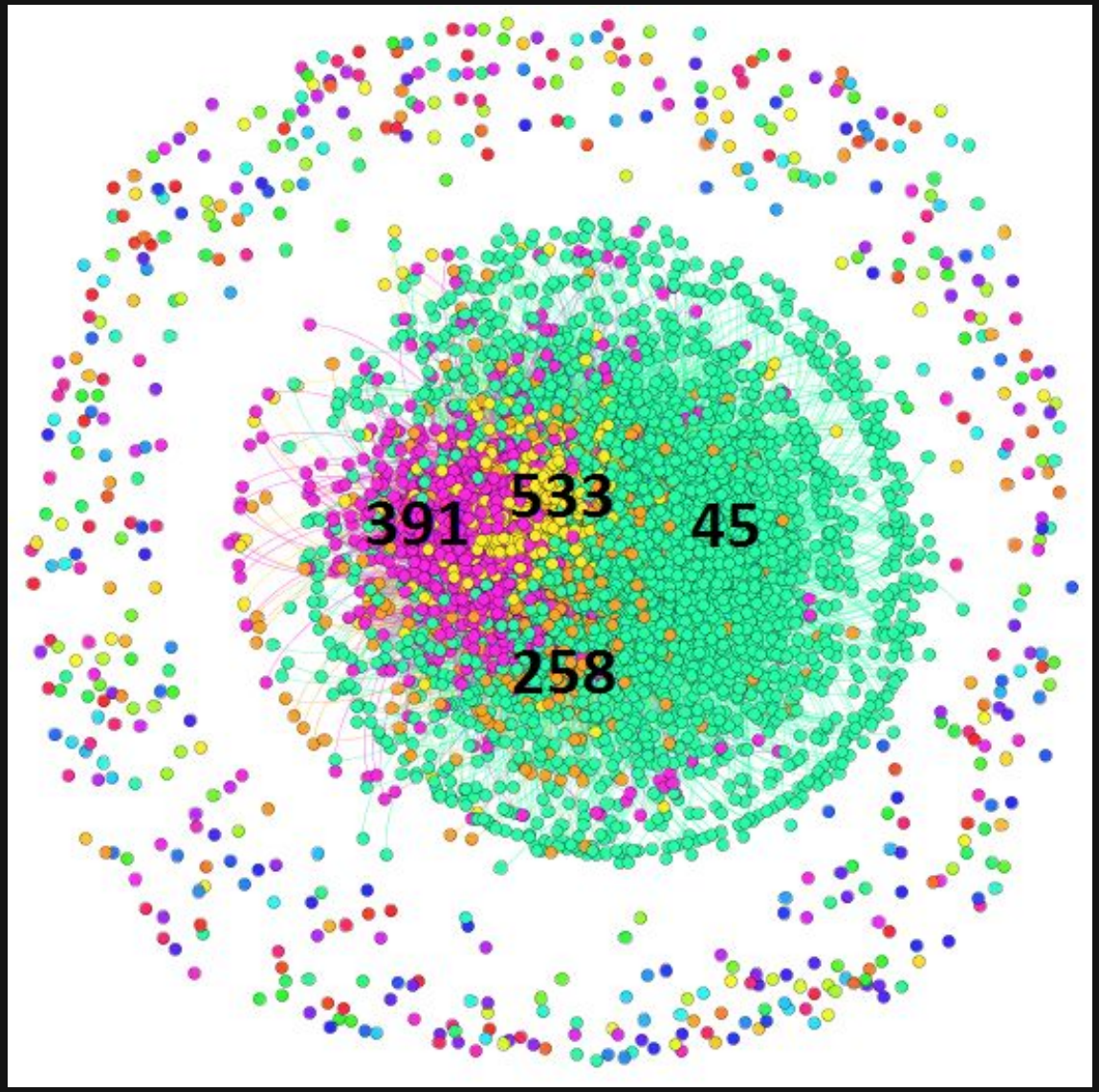


Rede de usuários — Filtragem colaborativa

- Abordagem baseada em *collaborative filtering*
- Parte de um grafo bipartido
  - Pesos de arestas desconsiderado
- Encontra dois grafos a partir do original
  - Usuários
  - Filmes



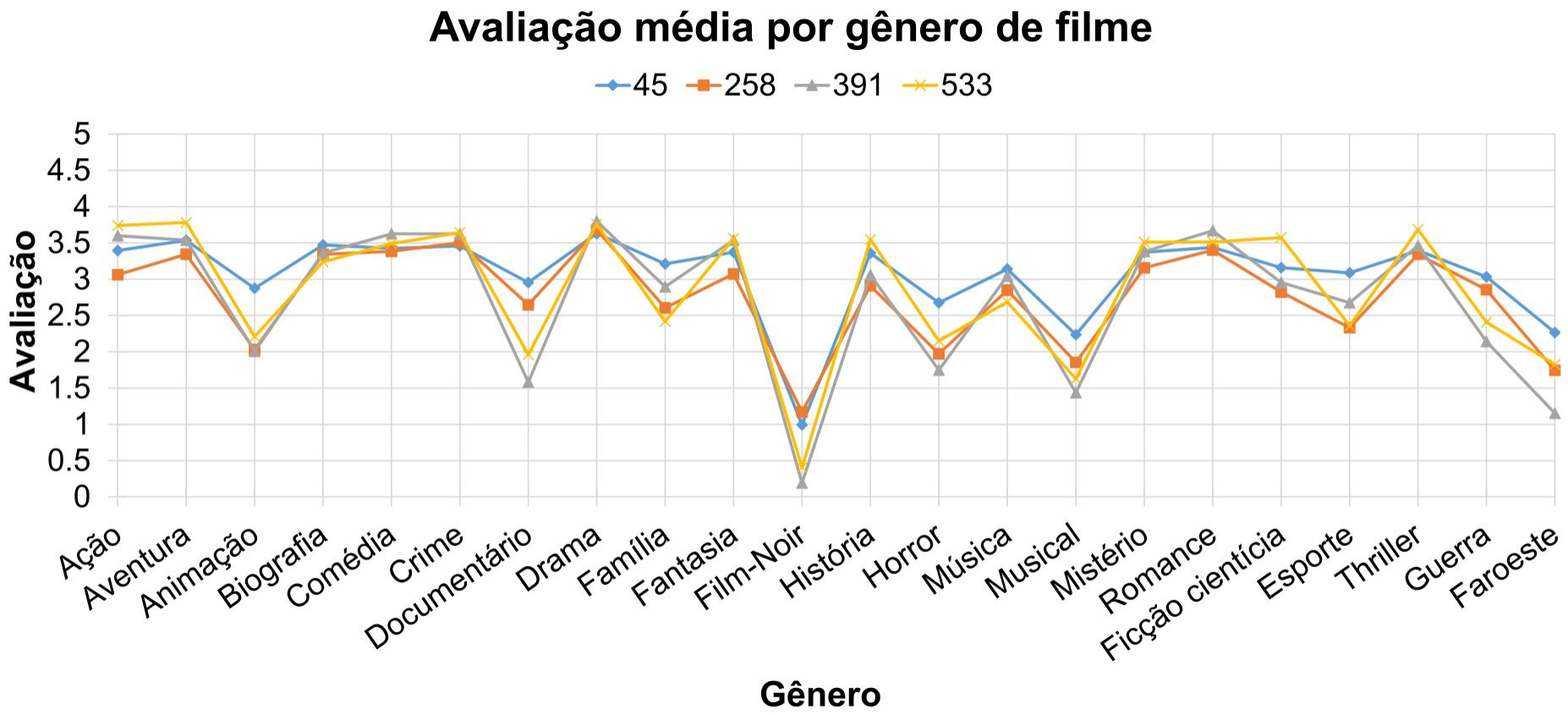
Rede de usuários — Filtragem colaborativa



Rede de usuários — Filtragem colaborativa

Comunidade	Avaliações/Usuário	# de usuários
45	509,09	2.548
391	343,18	973
533	648,12	648
258	702,09	283
Total (551 comunidades)	463,05	5.000

Rede de usuários — Filtragem colaborativa



Considerações finais

- Volume de dados trouxe várias limitações
- O formato dos títulos do Netflix nem sempre era o mesmo do OMDB
  - “Lilo and Stitch” ≠ “Lilo & Stitch”
  - “Beyonce: Live at Wembley” o OMDB não considera um filme
  - “Lord of the Rings: The Return of the King: Extended Edition: Bonus Material”
- Banco relacional
  - Baixo desempenho em algumas consultas
- Neo4j
  - 25 milhões das 89 milhões de arestas resultou em um banco de ~15 GB
- ~6h para computar a rede de interesses para 50 mil usuários (!)