

# Data Mining

## **Disciplina: Machine Learning**

## **Tema da Aula: Sistemas de Recomendação**

### **Coordenação:**

Prof. Dr. Adolpho Walter  
Pimazzi Canton

Profa. Dra. Alessandra de  
Ávila Montini

# **Prof. Carlos Eduardo Martins Relvas**

# Currículo

- Bacharel em Estatística, Universidade de São Paulo.
- Mestre em Estatística, Universidade de São Paulo.
- Itaú, 2010-2015. Principais atividades:
  - Consultoria estatística para várias áreas do banco com foco principal em melhorias no processo de modelagem de risco de crédito.
  - De 2013 a 2015, participação do projeto Big Data do banco usando tecnologia Hadoop e diversas técnicas de machine learning. Desenvolvemos diversos algoritmos em MapReduce usando R e Hadoop streaming, criando uma plataforma de modelagem estatística no Hadoop.
- Nubank, desde 2015. Principais atividades:
  - Equipe de Data Science, responsável por toda a parte de modelagem da empresa, desde modelos de crédito a identificar motivos de atendimento.

# Conteúdo da Aula

- Market Basket Analysis
- Sistemas de Recomendação
- SNA

# Market Basket Analysis



- Quais produtos são comprados em conjunto?
- Objetivo: Encontrar associação e correlações entre itens comprados.

# Market Basket Analysis

- Temos uma base de transações T.
- Cada transação  $T_i$  apresenta um conjunto de itens  $I_i = \{i_1, \dots, i_m\}$
- O objetivo é encontrar regras do tipo  $X \rightarrow Y$ .

TID	Items
T1	bread, jelly, peanut-butter
T2	bread, peanut-butter
T3	bread, milk, peanut-butter
T4	beer, bread
T5	beer, milk

## Examples:

- $\text{bread} \Rightarrow \text{peanut-butter}$
- $\text{beer} \Rightarrow \text{bread}$

# Market Basket Analysis

- Support Count ( $\sigma$ )
  - Contagem de co-ocorrência
  - $\sigma(\text{bread}, \text{peanut} - \text{butter}) = 3$
  - $\sigma(\text{beer}, \text{bread}) = 1$
- Support ( $s$ )
  - Fração de ocorrência
  - $s(\text{bread}, \text{peanut} - \text{butter}) = 3/5$
  - $s(\text{beer}, \text{bread}) = 1/5$

$$s = \frac{\sigma(X \cup Y)}{\text{\# of trans.}}$$

TID	Items
T1	bread, jelly, peanut-butter
T2	bread, peanut-butter
T3	bread, milk, peanut-butter
T4	beer, bread
T5	beer, milk

# Market Basket Analysis

- Confidence (c)
  - Mesma ideia do suporte, mas removendo o viés de itens muito frequentes (itens que todo mundo compra).
  - $s(\text{bread}, \text{peanut} - \text{butter}) = 3/4$
  - $s(\text{beer}, \text{bread}) = 1/2$

$$c = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

TID	Items
T1	bread, jelly, peanut-butter
T2	bread, peanut-butter
T3	bread, milk, peanut-butter
T4	beer, bread
T5	beer, milk



# Market Basket Analysis

TID	s	c
bread $\Rightarrow$ peanut-butter	0.60	0.75
peanut-butter $\Rightarrow$ bread	0.60	1.00
beer $\Rightarrow$ bread	0.20	0.50
peanut-butter $\Rightarrow$ jelly	0.20	0.33
jelly $\Rightarrow$ peanut-butter	0.20	1.00
jelly $\Rightarrow$ milk	0.00	0.00

TID	Items
T1	bread, jelly, peanut-butter
T2	bread, peanut-butter
T3	bread, milk, peanut-butter
T4	beer, bread
T5	beer, milk

# Market Basket Analysis

- Lift (l)
  - O lift de uma regra é uma medida ponderada pela força esperada da regra.
  - $\text{Lift} = \text{confidence} / \text{expected confidence}$
  - $\text{Lift}(X \rightarrow Y) = \text{Support}(X+Y) / \text{Support}(X) * \text{Support}(Y)$
  - Lift maior do que 1 indicam que X e Y aparecem juntos mais do que o esperado, ou seja, a compra de X tem um efeito positivo na compra de Y.

TID	Items
T1	bread, jelly, peanut-butter
T2	bread, peanut-butter
T3	bread, milk, peanut-butter
T4	beer, bread
T5	beer, milk

# Cesto de Compras – Laboratório

- Groceries dataset
- Base de dados real com um mês com todas as compras de um mercado.
- 9.835 transações com 169 categorias de produtos.
- Vamos encontrar regras de associação.
- Quem compra isso, também compra...



# Cesto de Compras – Exercício



Getting Started Prediction Competition

## Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics



Kaggle · 5,857 teams · 3 years to go

[Overview](#)

[Data](#)

[Kernels](#)

[Discussion](#)

[Leaderboard](#)

[More](#)

[Submit Predictions](#)

# Cesto de Compras – Exercício

1. Construa regras de associação usando o arquivo 'Titanic.raw'.
2. Filtre para regras que explique sobrevivência ou não sobrevivência.
3. Qual regra apresenta o maior lift?

# Sistemas de recomendação

- Conjunto de algoritmos cujo objetivo é prever como um usuário irá avaliar um certo produto ou experiência.

## Frequently Bought Together



Total price: \$918.90

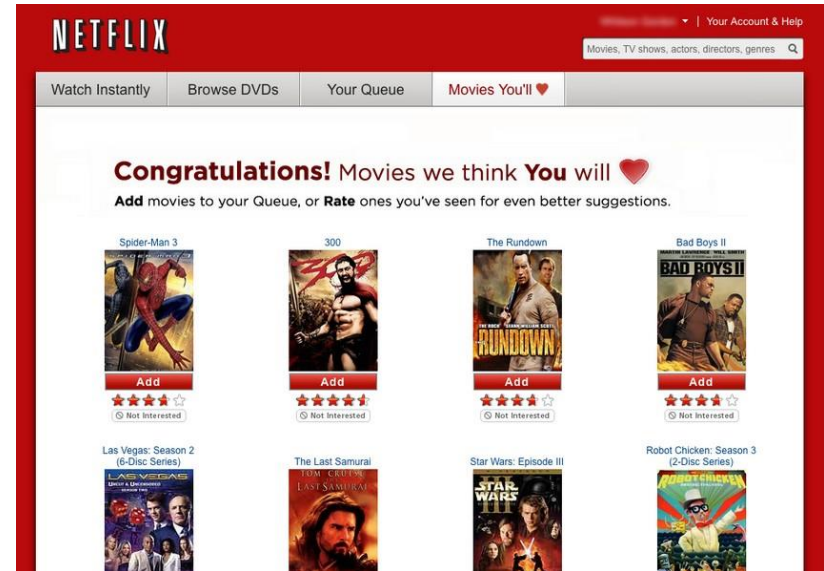
[Add both to Cart](#)

[Add both to List](#)

☒ This item: Apple MacBook Pro MD101LL/A 13.3-Inch Laptop \$910.00

☒ Case Logic Display Sleeve LAPS-113, 13.3-Inch, Black \$8.90

## Customers Who Bought This Item Also Bought



NETFLIX

Watch Instantly | Browse DVDs | Your Queue | Movies You'll ♥

**Congratulations!** Movies we think **You** will ♥

Add movies to your Queue, or **Rate** ones you've seen for even better suggestions.

Spider-Man 3	300	The Rundown	Bad Boys II
<a href="#">Add</a>	<a href="#">Add</a>	<a href="#">Add</a>	<a href="#">Add</a>
★★★★★	★★★★★	★★★★★	★★★★★
<a href="#">Not Interested</a>	<a href="#">Not Interested</a>	<a href="#">Not Interested</a>	<a href="#">Not Interested</a>

Las Vegas: Season 2 (6-Disc Series)	The Last Samurai	Star Wars: Episode III	Robot Chicken: Season 3 (2-Disc Series)

# Algoritmos de recomendação

- Não personalizado
- Content Based
- Filtro Colaborativo
- Outros

# Não personalizado

- Dados de comunidades externas (best sellers, mais popular, etc).
- Agregação de dados dos usuários (Média das avaliações → Cuidado com médias).

Fire, 7" Display, Wi-Fi, 8 GB - Includes Special Offers, Black

by Amazon

★★★★☆ 46,123 customer reviews | 1000+ answered questions

#1 Best Seller in Computers & Accessories



**Hyatt Ziva Cancun** ★★★★★👍  
Cancún

Reservado 2 vezes hoje

**Excepcional 9,6**  
77 avaliações  
👍 230

[Visualizar preços](#)

\$\$\$



**Le Blanc Spa Resort- All Inclusive - Adults Only** ★★★★★👍  
Cancún

Última reserva: há 19 horas

**Excepcional 9,6**  
82 avaliações  
👍 733

[Visualizar preços](#)

\$\$\$



# Content based

- Recomendar baseado em itens que o usuário gosta (avaliou bem anteriormente, clicou, leu, etc).
- Quem compra isso, também compra... (TFIDF).
- Vector Space Model

# Content based

✓ 1 item added to Cart



**NewAir AI-100BK 28-Pound Portable Ice Maker, Black**

\$189.99

Only 18 left in stock.

☐ Protect this product with a 2-year warranty \$16.99

☐ This will be a gift

**Order subtotal: \$189.99**

1 item in your Cart

NewAir AI-100BK 28-Pound Portable Ice Maker, Black

Edit your Cart

**Proceed to checkout**

**Your order qualifies for free shipping!**

Select FREE Super Saver Shipping at checkout. (Some restrictions apply)

**6**  
Month  
Financing

Your cart is eligible for a financing offer

**6 Month Special Financing on orders \$149 or more**  
with the **Amazon.com Store Card**. See details and restrictions.

Apply now

**Recommended for You Based on NewAir AI-100BK 28-Pound Portable Ice Maker, Black**



**NewAir AI-100R 28-Pound Portable Icemaker, Red**

★★★★☆ (116)

~~\$274.98~~ **\$187.98**



**Ice Scoop, Stainless Steel**

★★★★☆ (44)

**\$3.95**

13 New from \$0.01



**NewAir AI-100S 28-Pound Portable Ice Maker, Silver**

★★★★☆ (105)

~~\$274.98~~ **\$187.98**



**The Survival Medicine Handbook**

by Joseph Alton M.D.  
Paperback

★★★★☆ (2)

# Filtro colaborativo

- User-user
  - Medida de concordância intra usuário (correlação, vetor coseno, distâncias, etc).
  - Recomenda itens que usuários parecido aprovaram.
    - Quantos usuários parecidos?
    - Como fazer a predição (média, média ponderada, etc).
    - Diferentes usuários apresentam diferentes escalas (Tudo 1 ou 5, ou tudo entre 2 e 4) → Normalizar
- Item-item → baseado na concordância entre itens (performance computacional tende a ser melhor).

# Filtro colaborativo

Grant, Welcome to Your Amazon.com (If you're not Grant Ingersoll, click here.)

## Today's Recommendations For You

Here's a daily sample of items recommended for you. Click here to [see all recommendations](#).



[Principles of Data Mining \(A...](#)   
by David J....  
★★★★☆ (17) \$52.00



[Python in a Nutshell, Secon...](#)   
by Alex Mart...  
★★★★☆ (40) \$26.39



[Introductory Statistics wit...](#)  
by Peter Dal...  
★★★★☆ (20) \$48.56

**NETFLIX**

Watch InstantlyBrowse DVDsYour QueueMovies You'll ♥

**Congratulations!** Movies we think **You** will ♥

Add movies to your Queue, or **Rate** ones you've seen for even better suggestions.

Spider-Man 3

Add

★★★★☆  
Not Interested

300

Add

★★★★☆  
Not Interested

The Rundown

Add

★★★★☆  
Not Interested

Bad Boys II

Add

★★★★☆  
Not Interested

Las Vegas: Season 2  
(6-Disc Series)

Add

★★★★☆  
Not Interested

The Last Samurai

Add

★★★★☆  
Not Interested

Star Wars: Episode III

Add

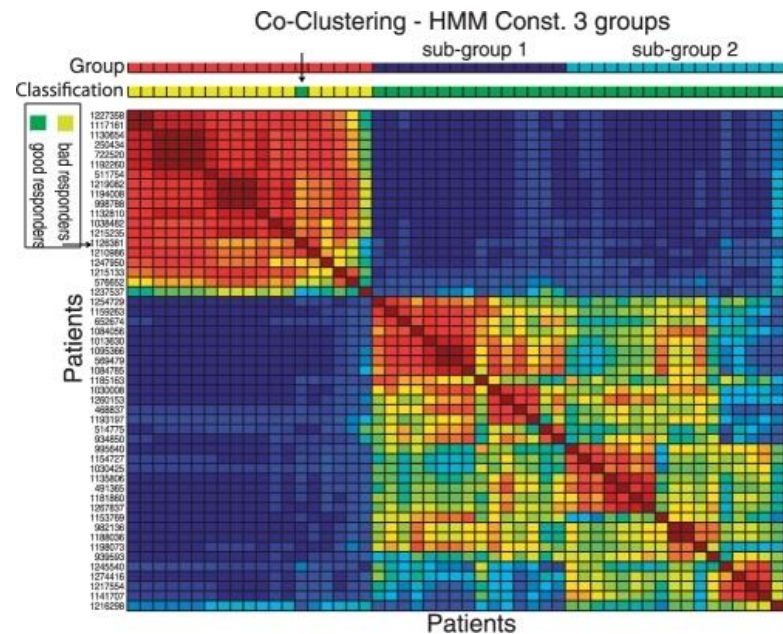
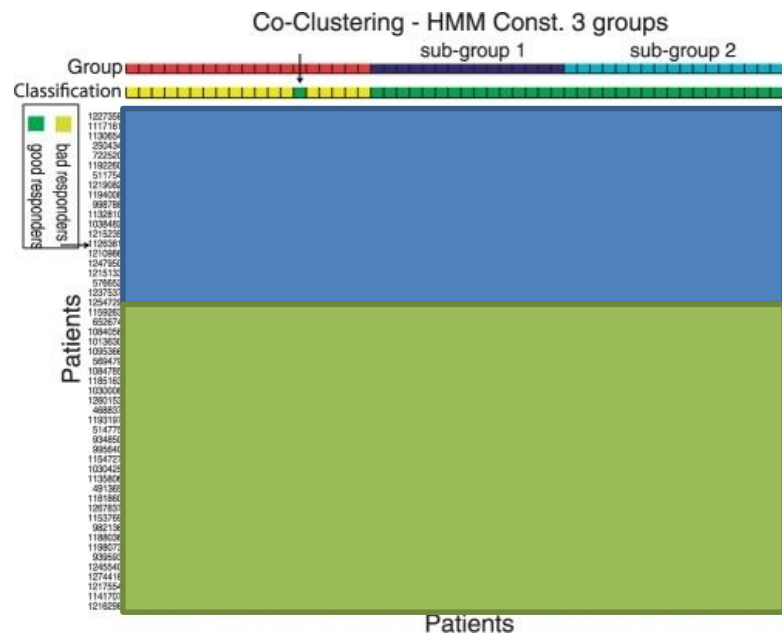
★★★★☆  
Not Interested

Robot Chicken: Season 3  
(2-Disc Series)

Add

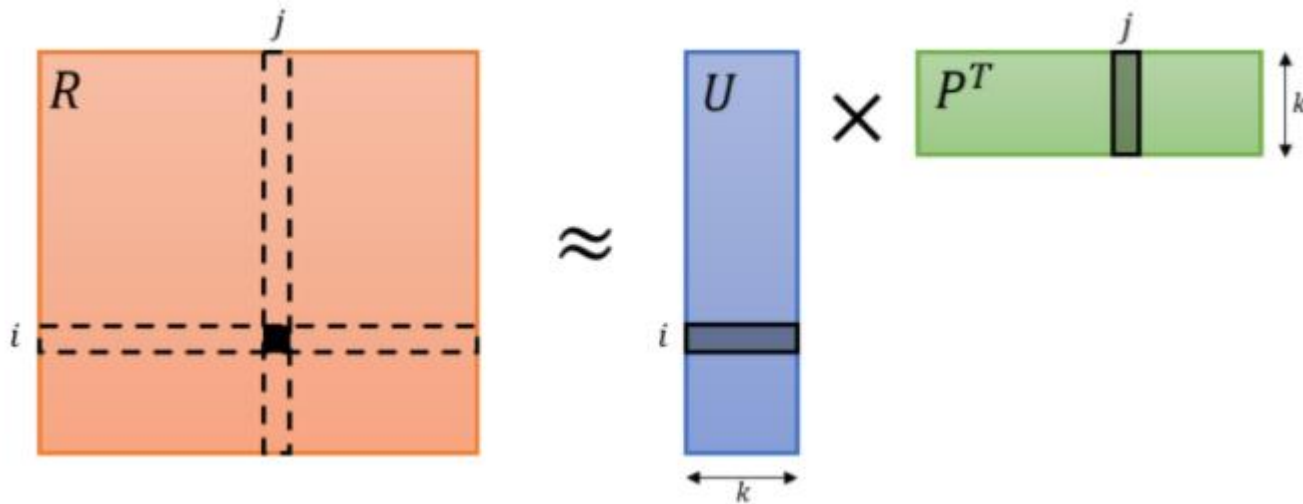
★★★★☆  
Not Interested

# Filtro colaborativo – Co-clustering



# Filtro colaborativo – Matrix Factorization

A ideia consiste em fatorar a matriz de ratings  $R$  ( $m \times n$ ) em duas matrizes,  $U$  ( $m \times k$ ) e  $P$  ( $n \times k$ ), tal que  $R$  é aproximadamente  $U \times P$



$K$  representa o rank da fatora  o.  $R_{ij} \approx \mu_i p_j$ .

# Filtro colaborativo – Matrix Factorization

Como encontrar U e P? A ideia é a mesma de um modelo de regressão. Encontrar P e U tal que o erro seja mínimo. Logo, minimizamos a seguinte função:

$$J = ||R - U \times P^T||_2 + \lambda (||U||_2 + ||P||_2)$$

↑  
Erro quadrático

↑  
Regularização  
(evitar overfitting)

Repare que temos dois parâmetros para escolher ( $k$  e  $\lambda$ ), que podemos escolher por cross-validation.

# Filtro colaborativo – Alternating Least Squares

- A otimização de  $U$  e  $P$  simultaneamente é uma função não complexa, o que traz vários problemas no processo de estimação.
- Mas repare que a otimização somente de  $U$  ou  $P$  com a outra matriz fixada é equivalente ao caso da regressão linear.
- Alternating Least Squares faz exatamente isso por meio de um processo iterativo de duas etapas. Consiste em fornecer uma estimativa inicial para  $U$  e otimizar a função para a  $P$ . Com esta estimativa de  $P$ , otimizar a função para  $U$ .
- Este processo é repetido até a convergências das estimativas.



# Filtro colaborativo – Alternating Least Squares

- Garante a convergência para um mínimo local que depende das condições iniciais (podemos executar com diferentes condições iniciais e utilizar a melhor).
- Cada  $u_i$  é independente dos outros  $u_j$ 's, este algoritmo é facilmente paralelizado.

# Filtro colaborativo – Alternating Least Squares

## Recomendação

- Construímos  $U$  e  $P$  a partir da matriz  $R$  incompleta, em que vários usuários não avaliaram vários itens.
- Agora podemos reconstruir a matriz  $R$  usando  $U$  e  $P$  estimadas e agora nossa matriz  $R$  estimada será completa.
- Recomendamos para cada usuário os top  $K$  itens com maiores nota estimada.

# Filtro colaborativo – Vantagens em relação ao content based

- Não é necessário ter informações do item (não é necessário saber quem são os autores do filme, diretor e categoria).
- O interesse do usuário pode mudar com o tempo.
- Mais fácil de explicar para o usuário.
- Pode capturar itens não relacionados (comprar banana e desodorante).

# Filtro colaborativo – Desvantagens em relação ao content based

- Esparsidade da matriz de preferências. Problemas de processamento.
- Necessário maior cuidado com sinônimos (itens iguais com nomes diferentes).
- Cuidado com usuários “aleatórios”, que parecem não concordar com ninguém.
- Problema com pessoas fanáticas por marcas.

# Métricas

- Métricas para previsões:
  - MAE → Mean absolute error
  - MSE → Mean squared error
  - RMSE → Root mean squared error
- Métricas para recomendações:
  - O que é uma recomendação errada? Toda vez que uma avaliação ruim aparece no top-n é um erro?
  - Métricas de negócio (venda, cliques, etc).
  - Precision → percentual de itens selecionados que são “relevantes”.
  - Recall → percentual de itens relevantes que são selecionados.

# Evaluation

- Cross-Validation
- Base de teste
- Testes (Alguns usuários recebem recomendações “placebo”).

# Cold Start Problem

- Como recomendar algo que ainda não temos dados?
- Item cold start: um novo item foi adicionado (por exemplo, um novo filme) e ninguém avaliou ainda
- User cold start: um novo usuário se cadastra. O que recomendamos para ele?
- Em geral, utilizamos características intrínsecas dos itens / usuários para fornecer as recomendações iniciais.

# Outros

- Modelos mais complexos que envolvem características pessoais dos usuários.
- Modelos baseados em críticas de especialistas.
- Modelos híbridos (combinação de sistemas de recomendação).
- ....



# Sistemas de recomendação – Laboratório

- Last FM dataset
- 1757 usuários e a informação de compra ou não compra de 285 bandas.
- O objetivo é recomendar novas compras para os usuários com base nas bandas que ele costuma ouvir.

The logo for last.fm, featuring the text "last.fm" in a bold, red, lowercase sans-serif font.

# Sistemas de recomendação – Laboratório

## MovieLens: 5 star movie ratings

web site:



My Neighbor  
Totoro

となりのトトロ



MovieLens predicts for you  
4.61 stars

Average of 7,324 ratings  
4.13 stars



dataset:

userId,movieId,rating,timestamp

1,2,3.5,1112486027

1,29,3.5,1112484676

1,32,3.5,1112484819

1,47,3.5,1112484727

1,50,3.5,1112484580

1,112,3.5,1094785740

1,151,4.0,1094785734

1,223,4.0,1112485573

1,253,4.0,1112484940

...

138493,69644,3.0,1260209457

138493,70286,5.0,1258126944

138493,71619,2.5,1255811136

# Sistemas de recomendação – Laboratório

- MovieLens
- Base de dados mais famosas para estudo de sistema de recomendações.
- Há versões de diferentes tamanhos.
- 6040 usuários e 3676 filmes

# Sistemas de recomendação – Exercício

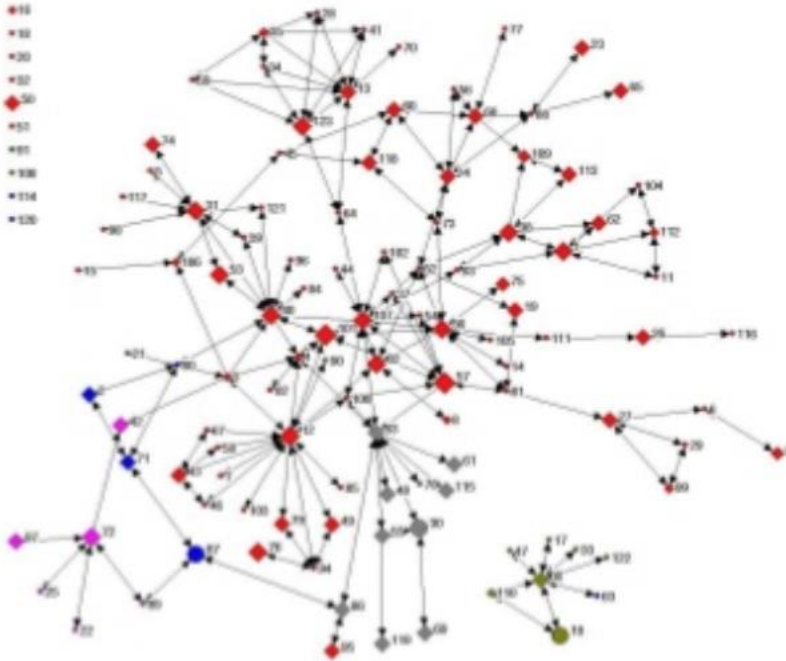


- Jester Online Joke Recommendation.
- Base de dados com 1.7 milhões de ratings contínuos (de -10 a 10) de 150 piadas de 59.132 usuários.

# Sistemas de recomendação – Exercício

1. Utilizando o pacote "recommenderlab", construa um Item Item based recommendation.
2. Faça o mesmo utilizando o ALS.

# SNA – Social Network Analysis



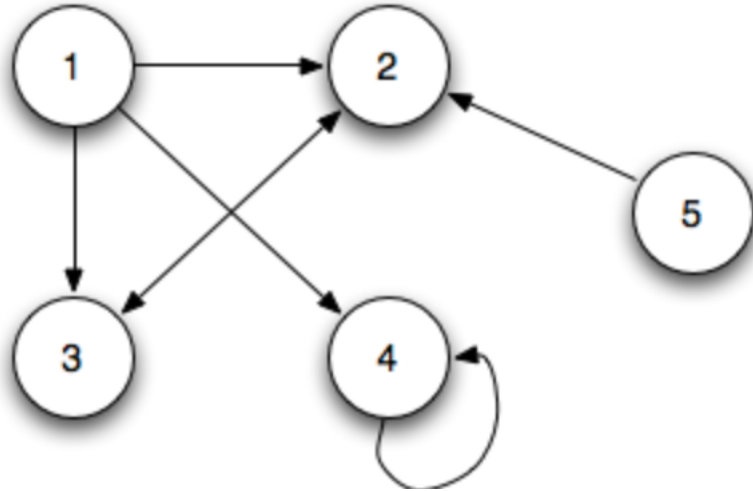
- Pessoas, organizações, fraudadores, etc, estão conectados uns com ou outros por diversas formas.
- Análise de rede social nos permite estudar como estas relações acontecem e como podem impactar o mundo real.
- Será que as pessoas com quem transfiro dinheiro mensalmente tem o mesmo risco que eu?

# SNA – Social Network Analysis

- Com o uso de técnicas de análise de redes sociais, temos dois principais objetivos em Machine Learning:
  1. A partir da estrutura da rede, criar informações (variáveis) que podem ser utilizadas em algoritmos de classificação / regressão, como uma regressão linear ou random forest.
  2. Estudar propriedades das redes para responder perguntas de interesse. Por exemplo, quem é a pessoa mais influente da rede de funcionários de uma empresa?

# SNA – Social Network Analysis

Representação: um grafo represente como objetos (nós ou vértices) se relaciona com outros objetos. Quando a ligação existe, digamos que há uma aresta (edge) entre eles. Esta ligação pode ser direcionada ou não direcionada.

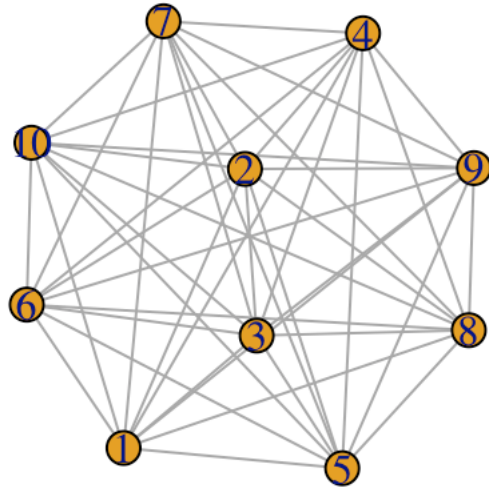


	1	2	3	4	5
1	0	1	1	1	0
2	0	0	1	0	0
3	0	1	0	0	0
4	0	0	0	1	0
5	0	1	0	0	0

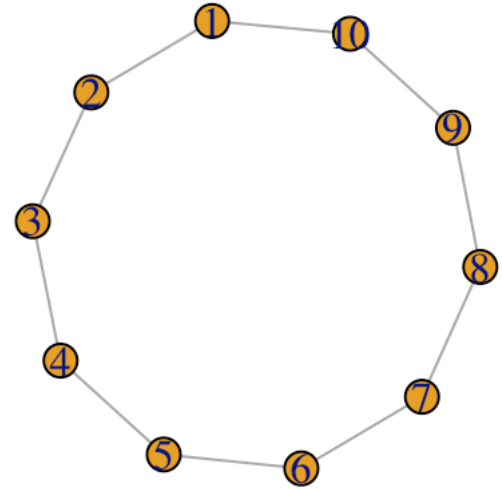


# Propriedades de grafos

Alguns tipos de redes:



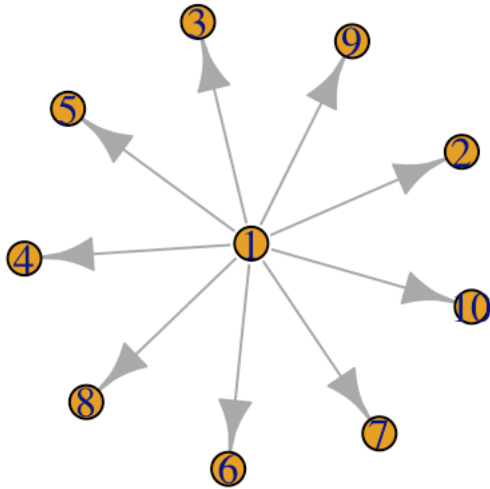
Totalmente conectada



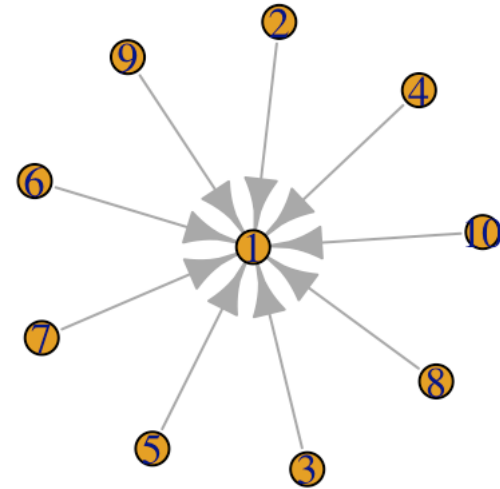
Anel

# Propriedades de grafos

Alguns tipos de redes:



Estrela Out



Estrela In

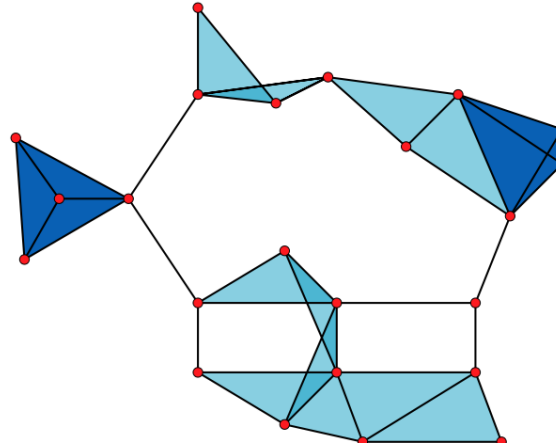
# Propriedades de grafos

- **Densidade:** razão do número de arestas pelo número total de arestas.
- **Transitividade:** se x tem relação com y, e y tem relação com z, então x tem ligação com z? Proporção de vezes que isto acontece.
- **Diâmetro:** maior menor caminho entre dois elementos da rede.
- **Degree:** número de arestas que chegam a cada vértice.
- **Closeness:** medida de centralidade de um nó. Representa uma distância média para todos os outros nós da rede.

$$C(x) = \frac{N}{\sum_y d(y, x)}.$$

# Propriedades de grafos

- **Betweenness:** número de menores caminhos entre todos os menores caminhos que passam por cada nó.
- **Mean distance:** distância média entre todos os nós.
- **Clique:** um subgrafo completo.



# Algoritmos em grafos

- Há diversos algoritmos em grafos, entre eles, podemos citar:
- Algoritmos para encontrar hubs (vértices mais "importantes", onde a informação sempre passa).
- Community detection. Consistem em clusterizar os nós baseado em propriedades do grafo.

# SNA – Social Network Analysis

Laboratório: Utilizamos o pacote igraph do software R.

O igraph é o pacote mais famoso e é bem completo para análise de grafos.

Há alternativas mais parrudas para lidar com grafos grandes como Neo4j, GraphX (Spark), Titan, entre outros.

# SNA – Exemplos Famosos

6 graus de separação:

- Experimento de Stanley Milgram conduzindo em 1961.
- Cartas foram enviadas aleatoriamente para pessoas em Kansas e Nebraska com informações básicas do experimento e do verdadeiro destinatário.
- O experimento consistia em enviar a carta para o verdadeiro destinatário caso ele fosse conhecido ou para algum conhecido que pudesse conhecer o destinatário.
- 296 cartas foram enviadas.



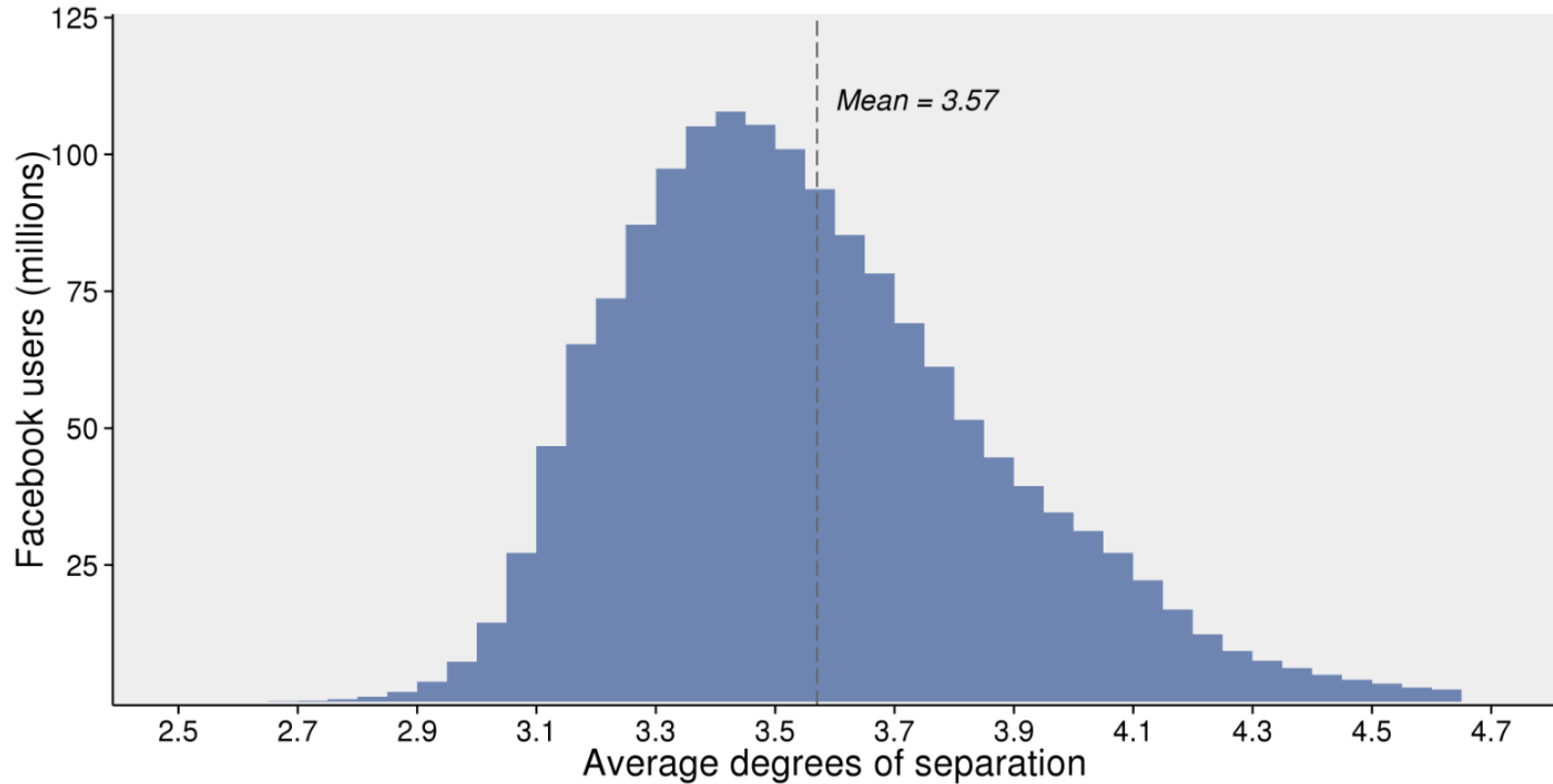
# SNA – Exemplos Famosos

6 graus de separação. Resultados:

- 232 cartas nunca chegaram.
- 64 cartas de fato chegaram e o número médio de passos foi 5,5/6.
- Por isso, os autores descreveram que o grau de separação médio entre pessoas dos Estados Unidos eram 6 pessoas.
- Facebook usando sua própria rede calculou que a distância média entre qualquer pessoa do mundo é 3,57.  
<https://research.fb.com/three-and-a-half-degrees-of-separation/>



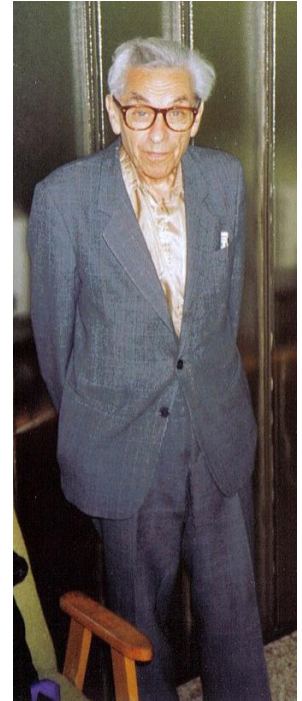
# SNA – Exemplos Famosos



# SNA – Exemplos Famosos

Erdos number:

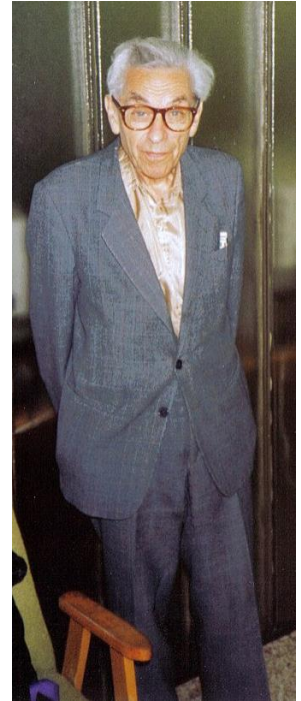
- Paul Erdos (1913-1996) era um matemático húngaro muito famoso. Paul viajava em volta ao mundo colaborando com matemáticos de diferentes faculdades, sendo o matemático com mais publicações em toda a história (1525).
- O Erdos number é a distância que uma pessoa está de Paul Erdos via co-autoria de artigos acadêmicos.
- Se Paulo publicou um artigo com o Erdos, Paulo tem um Erdos number de 1. Já se João publicou um artigo com Paulo e não com Erdos, João é Erdos number 2.



# SNA – Exemplos Famosos

Erdos number:

- Grandes matemáticos tem Erdos number baixos
- A mediana do Erdos number de medalhistas Fields é 3.
- Bacon number é a generalização para distância em filmes de atuação com Kevin Bacon. Também há o Erdos Bacon number sendo a soma dos dois .
- Natalie Portman tem um Erdos number de 7.



# Referências Bibliográficas

- Hastie, T., Tibshirani, R. & Friedman, J.H. (2001) “The Elements of Statistical Learning”
- Bishop, C.M. (2007) “Pattern Recognition and Machine Learning”
- Mitchell, T.M. (1997) “Machine Learning”
- Abu-Mostafa, Y., Magdon-Ismail, M., Lin, H.T (2012) “Learning from data”
- Theodoridis, S., Koutroumbas, K., (2008) “Pattern Recognition”
- Kuhn, M., Johnson, K., (2013) “Applied Predictive Modeling”