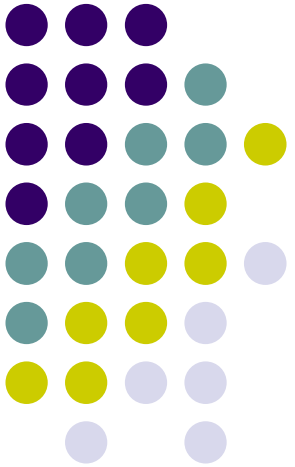
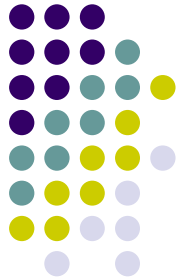
Algoritmo CLIQUE (**Cl**ustering **I**n **QUE**st)

Marcelo Camacho de Souza

Nº USP: 3199616

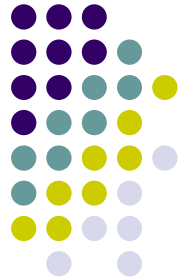


Roteiro



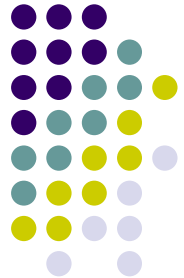
- *Algoritmo CLIQUE*
- *Exemplo Prático*
- *Complexidade Computacional*
- *Vantagens e Desvantagens*

CLIQUE (AGRAWAL *et al.*, 1998)



- *Autores (IBM Almaden Research Center):*
  - **Rakesh Agrawal**
  - **Johannes Gehrke**
  - **Dimitrios Gunopulos**
  - **Prabhakar Raghavan**
- *Publicação:*
  - **Proceedings of the ACM SIGMOD Conference on Management of Data, 1998.**

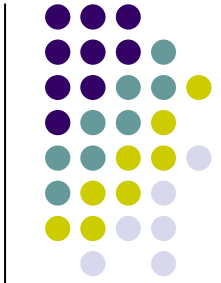
CLIQUE (AGRAWAL *et al.*, 1998)



- *Integra métodos de clustering baseados em:*
  - **Grade**
  - **Densidade**
  - **Subespaço**

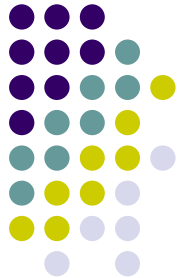
- *Necessita de Parâmetros de Entrada*:
  - **Tamanho da Grade**
  - **Limiar de Densidade**

CLIQUE (AGRAWAL *et al.*, 1998)



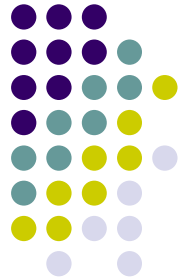
- *Indicação*:
  - Grandes conjuntos de dados com **elevada dimensionalidade**.
  - Identifica **unidades densas** em subespaços de alta dimensionalidade de dados, e usa esses subespaços para prover clusters de melhor eficiência.
  - Capaz de identificar clusters com **formas arbitrárias**.

CLIQUE (AGRAWAL *et al.*, 1998)



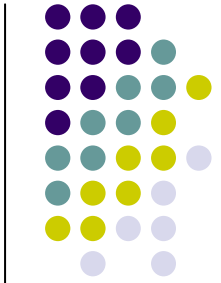
- *Algumas definições*:
  - **Dimensão**: É um Atributo.
  - **Subespaço**: Um subconjunto de dimensões.
  - **Unidade**: São obtidas pela partição de cada dimensão em intervalos de igual tamanho (parâmetro de entrada).

CLIQUE (AGRAWAL *et al.*, 1998)



- *Algumas definições:*
  - **Densidade**: Uma unidade é densa se a quantidade de pontos de dados excede ao limiar de densidade (parâmetro de entrada).
  - **Cluster**: É definido como o maior conjunto de unidades densas conectadas.

CLIQUE (AGRAWAL *et al.*, 1998)

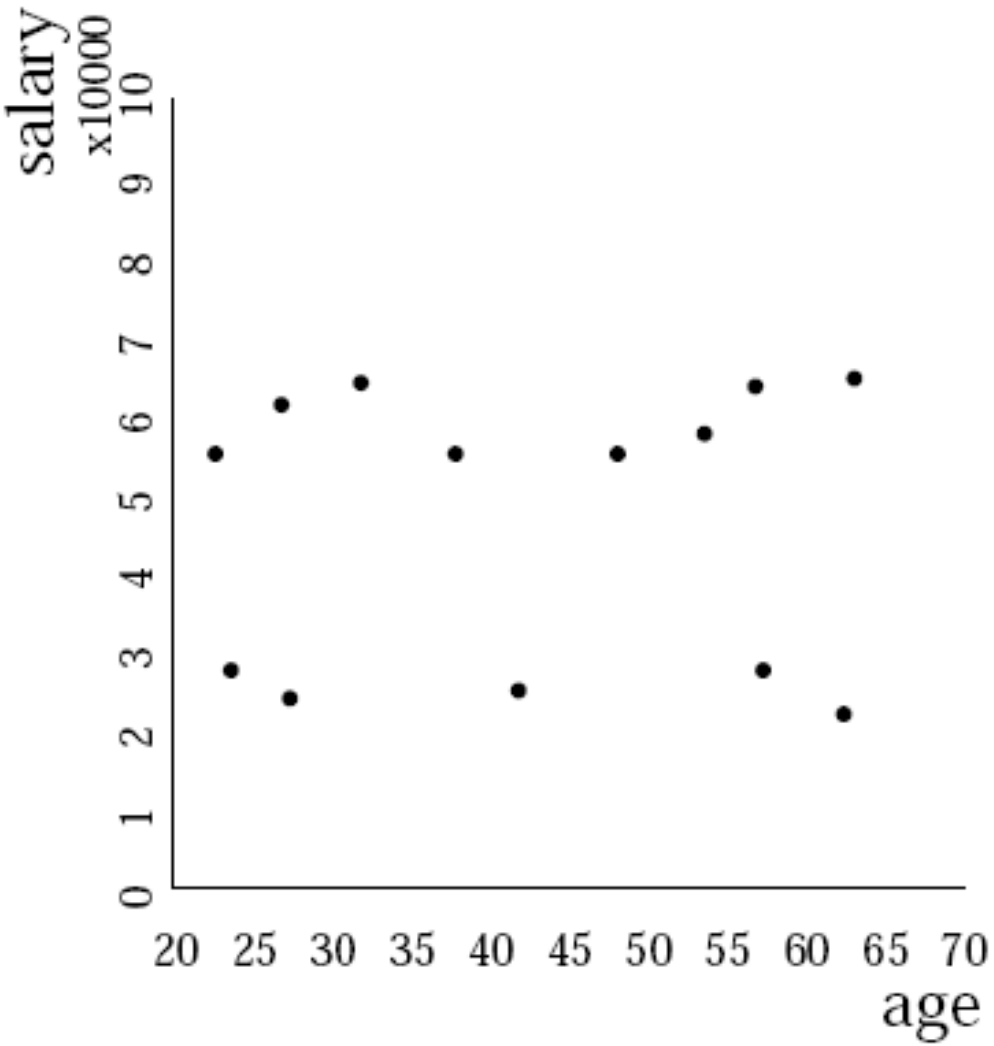


- *Etapas do Algoritmo:*
  1. Identificação de subespaços que contém clusters;
  2. Identificação dos clusters;
  3. Geração de descrição mínima para os clusters.

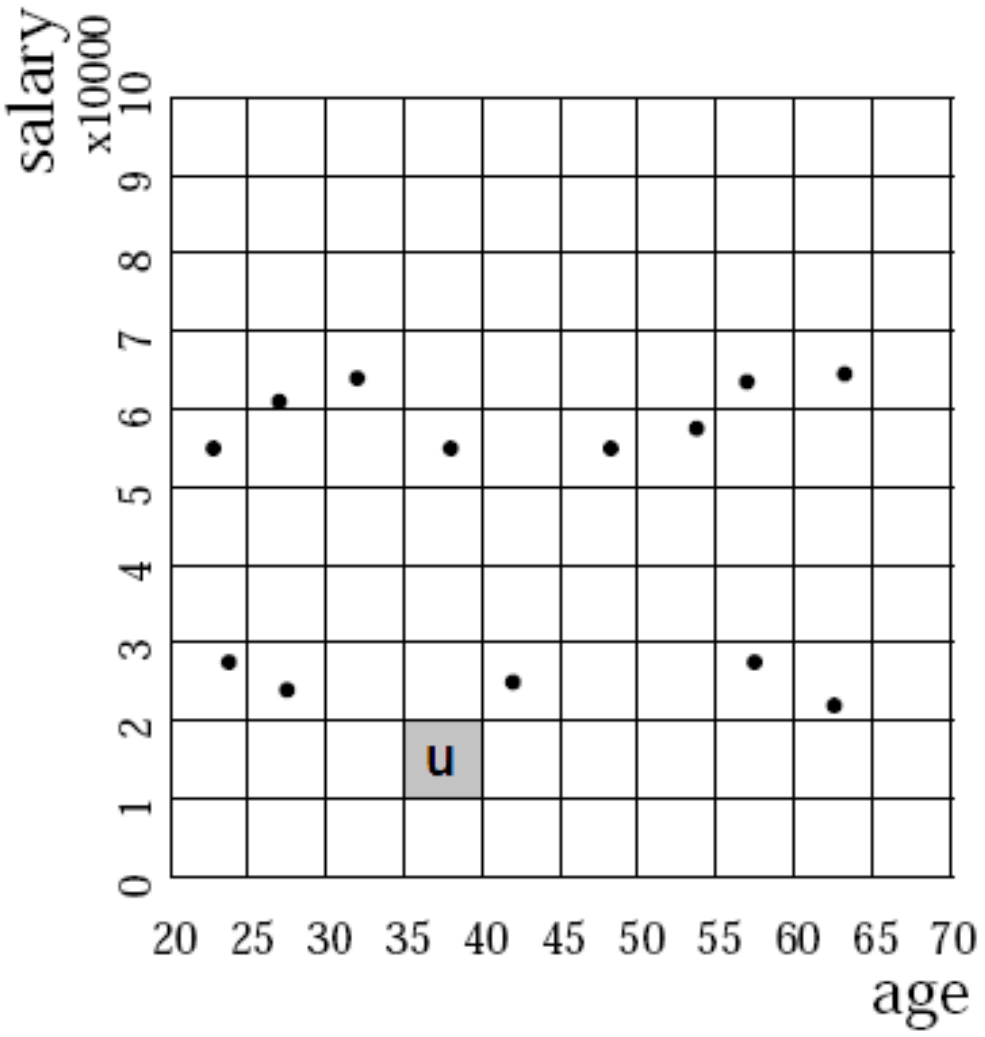
CLIQUE (AGRAWAL *et al.*, 1998)

- *Algoritmo – Parte 1*:
  - Particiona-se cada dimensão de interesse em um número de intervalos de **igual tamanho** (parâmetro de entrada).
  - Identifica-se as áreas escassas e **densas**.
  - Uma unidade é **densa** se o total de pontos de dados contidos nela excede o limiar de densidade (parâmetro de entrada).
  - Determina-se as **unidades densas conectadas** em todos os subespaços de interesse.

CLIQUE (AGRAWAL *et al.*, 1998)

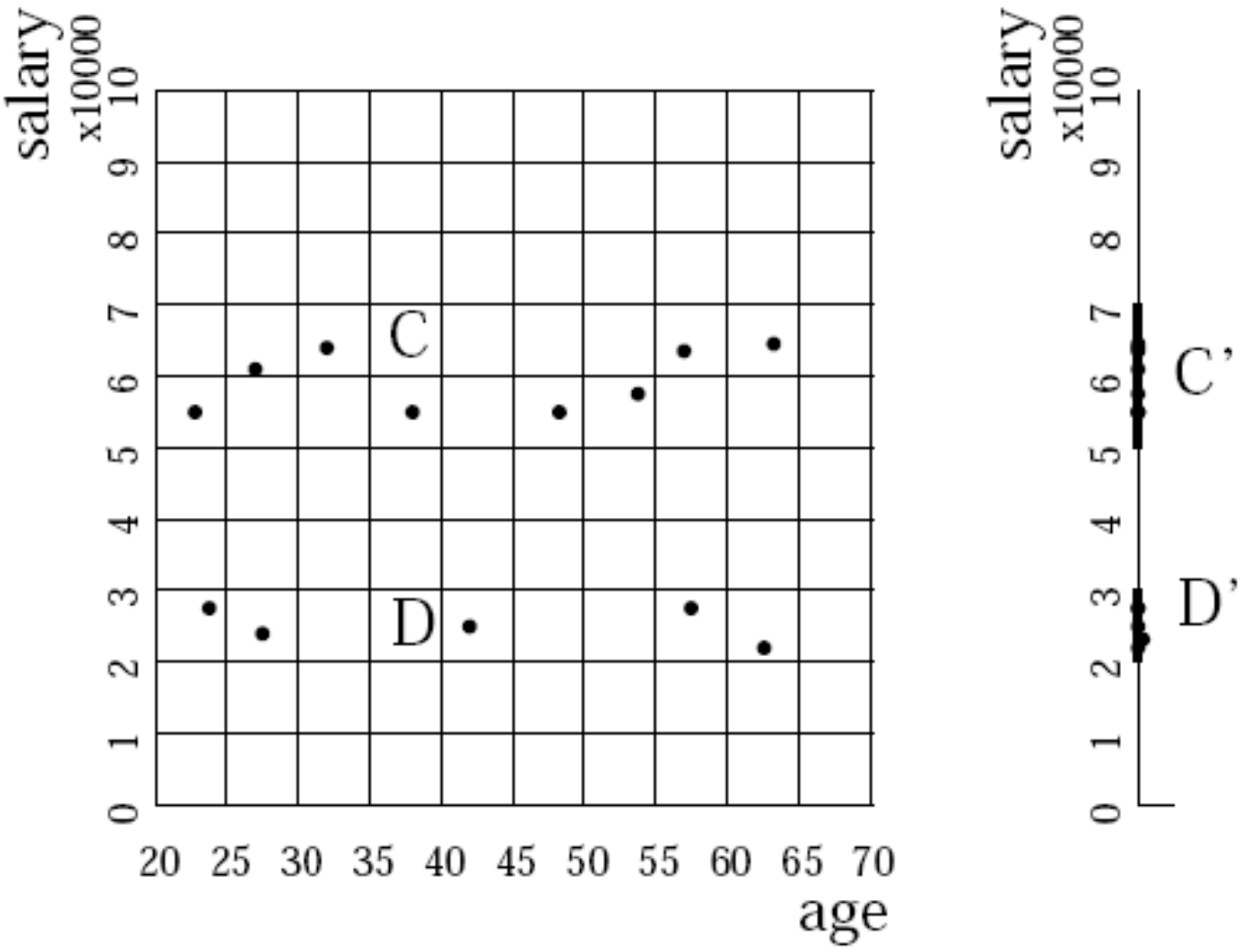


CLIQUE (AGRAWAL *et al.*, 1998)

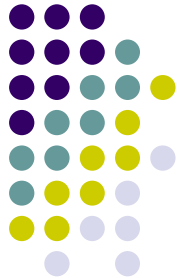


Grade = 10

CLIQUE (AGRAWAL *et al.*, 1998)

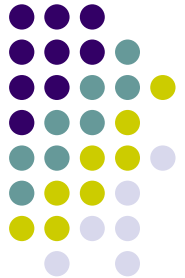


CLIQUE (AGRAWAL *et al.*, 1998)



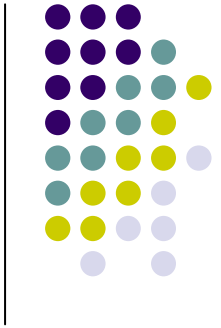
- *Algoritmo – Parte 2*:
  - Uma vez que os subespaços apropriados são encontrados, a tarefa é encontrar clusters nas **projeções** correspondentes.
  - Os clusters são **uniões** de **unidades de alta densidade** conectadas em um subespaço.
  - Os clusters são formados pela união de **retângulos sobrepostos** entre subespaços projetados.

CLIQUE (AGRAWAL *et al.*, 1998)

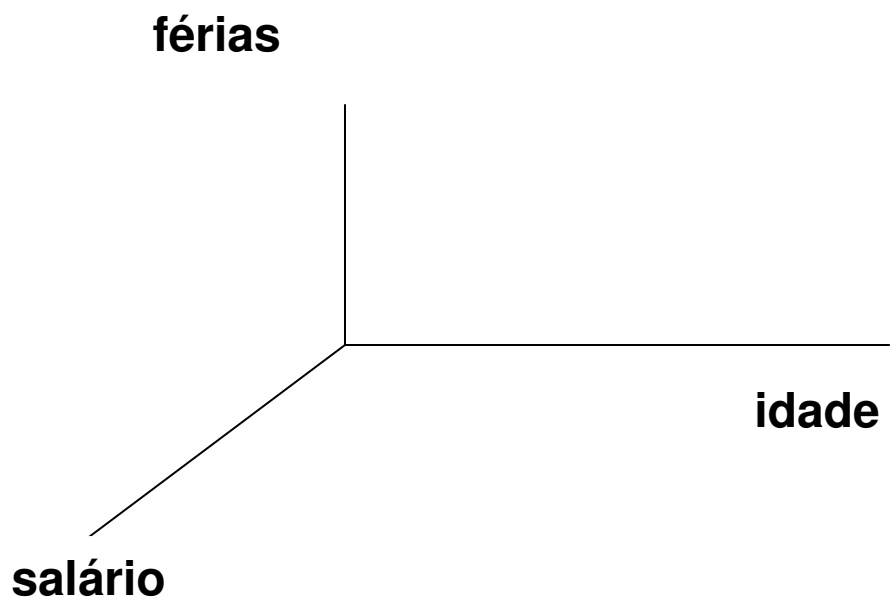


- *Algoritmo – Parte 3*:
  - Gera-se uma **descrição compacta** para cada cluster por meio da região máxima que cobre o cluster de células densas conectadas.

CLIQUE (AGRAWAL *et al.*, 1998)

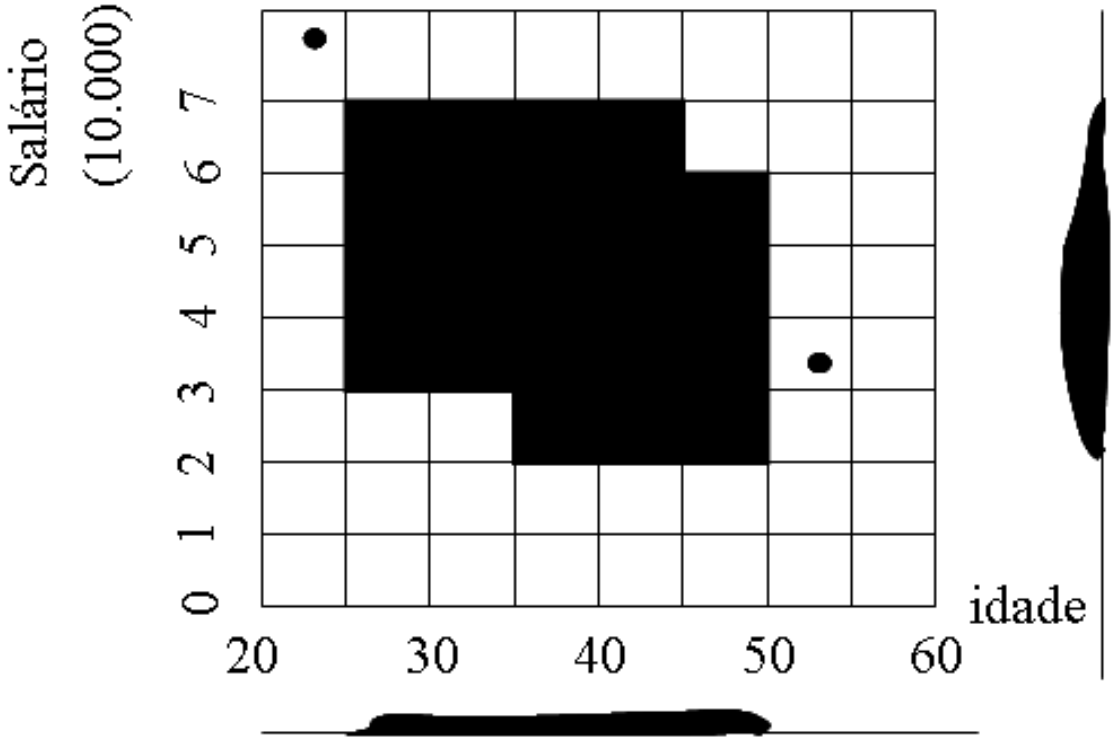


- *Exemplo:*
  - Suponha que você quer analisar um conjunto de dados que possui três atributos: salário, férias e idade.
  - Espaço 3D:



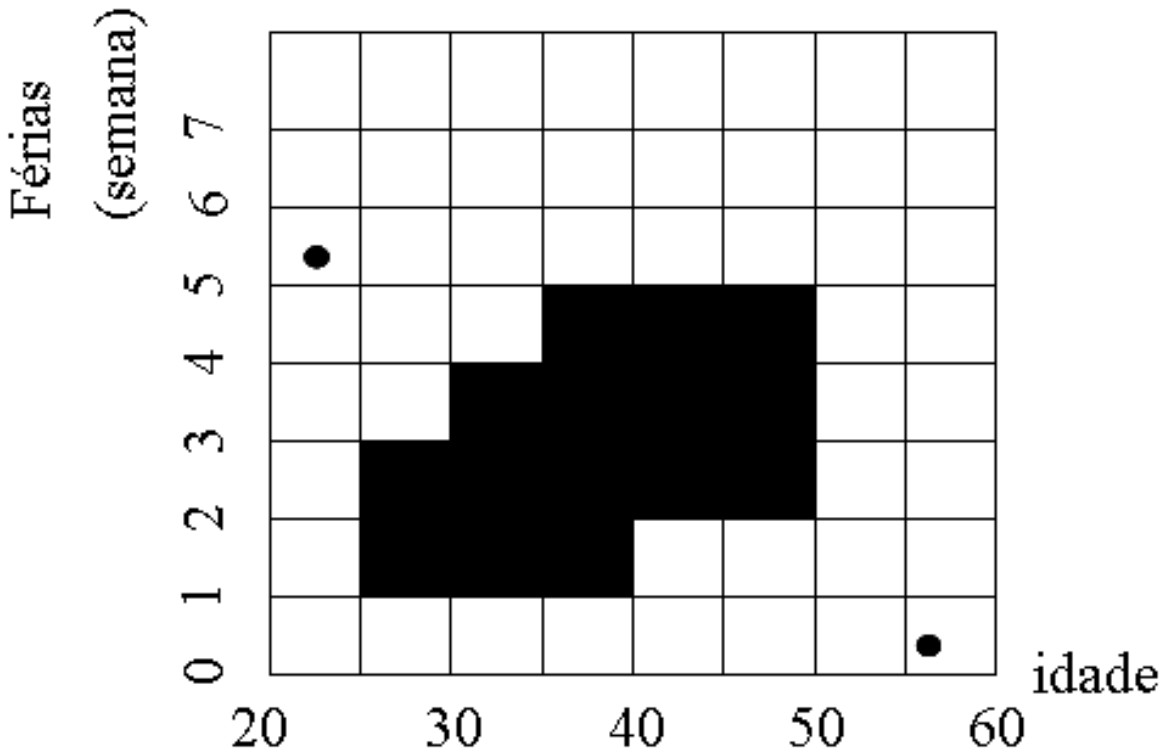
CLIQUE (AGRAWAL *et al.*, 1998)

- *Exemplo:*
  - Projetar cluster no subespaço: “salário e idade”



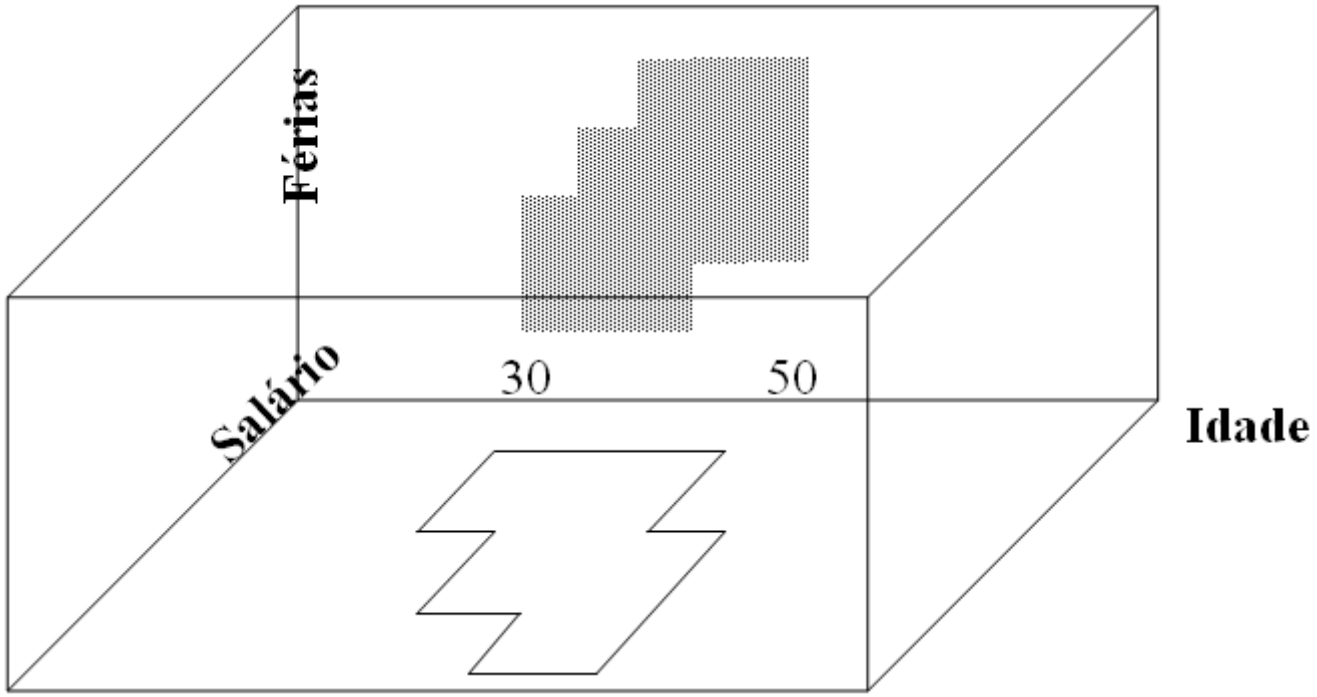
CLIQUE (AGRAWAL *et al.*, 1998)

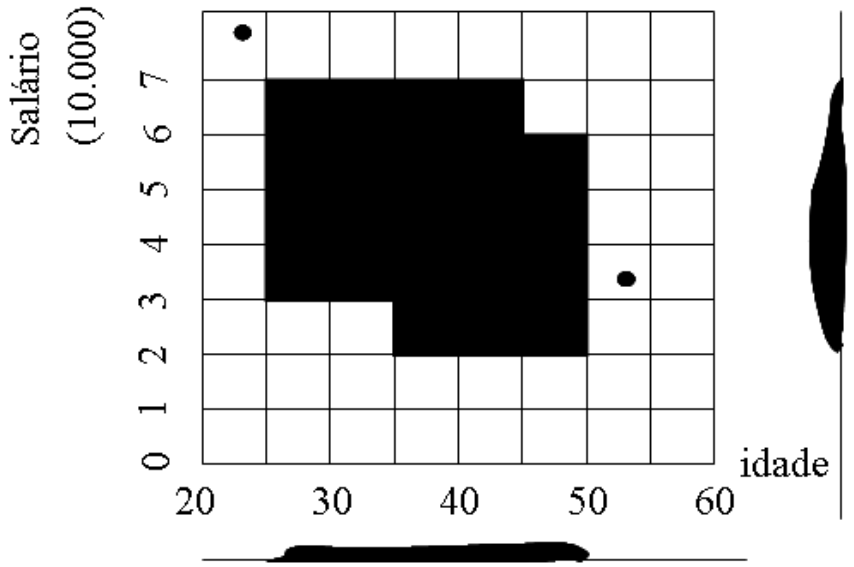
- *Exemplo:*
  - Projetar cluster no subespaço: “férias e idade”

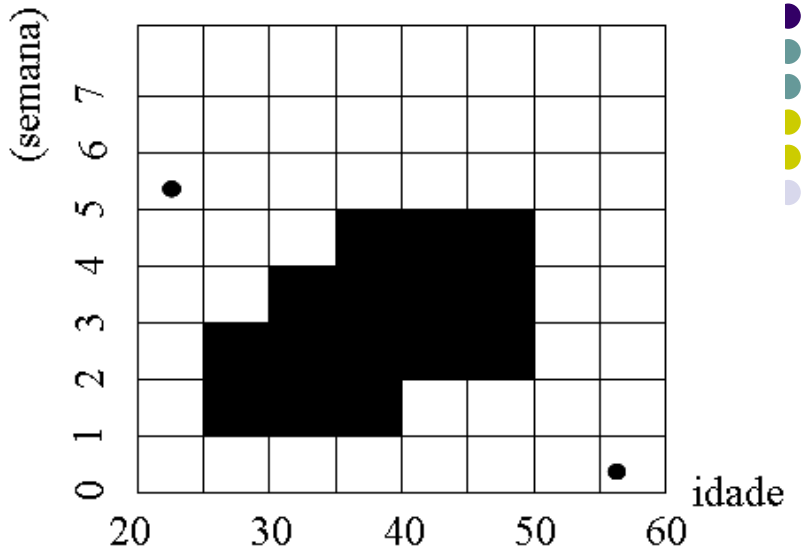


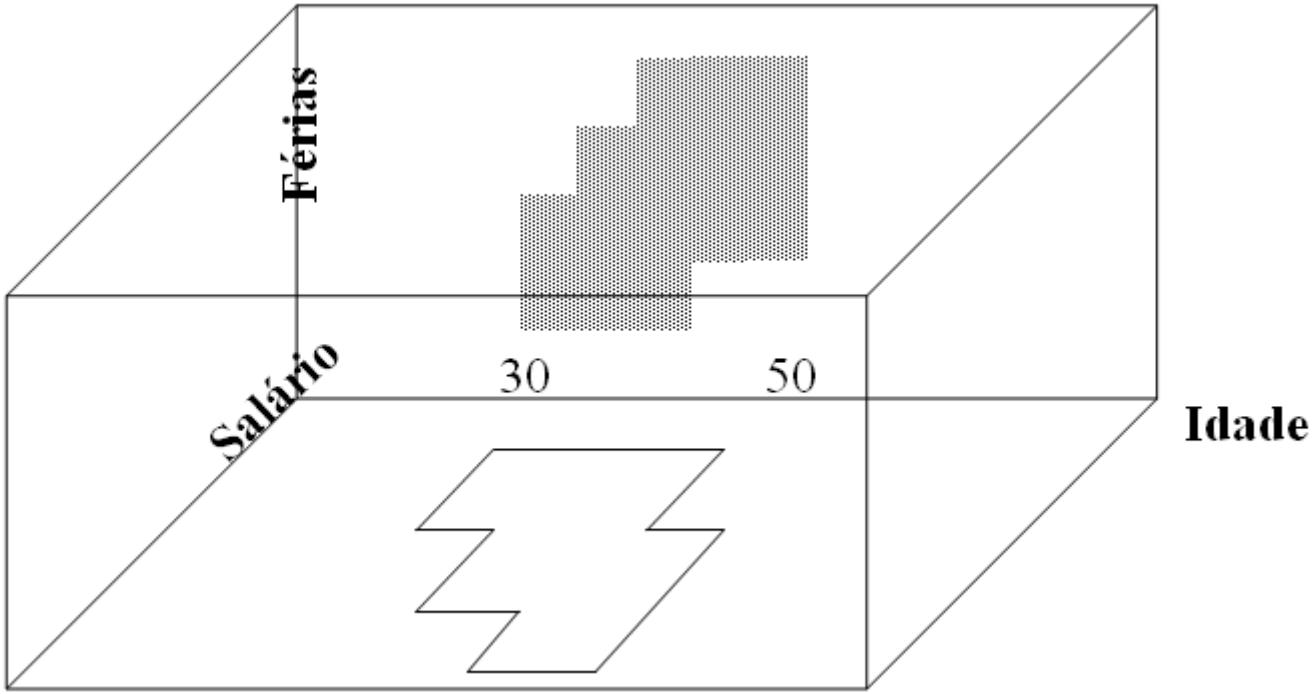
CLIQUE (AGRAWAL *et al.*, 1998)

- *Exemplo:*
  - Projetar cluster entre os subespaços “salário e idade” e “férias e idade”.

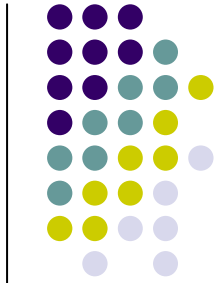



Salário
(10.000)
0 1 2 3 4 5 6 7
20 30 40 50 60
idade
Férias


Férias
(semana)
0 1 2 3 4 5 6 7
20 30 40 50 60
idade


Férias
Salário
30
50
Idade

CLIQUE (AGRAWAL *et al.*, 1998)



- *Exemplo Prático*

Laboratório de Análises de Dados

Prof. Yaling Pei e Prof. Osmar Zaiane

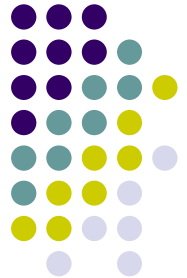
Departamento de Ciências da Computação

Universidade de Alberta - Canadá

Disponível em:

http://webdocs.cs.ualberta.ca/~yaling/Cluster/

CLIQUE (AGRAWAL *et al.*, 1998)



- CLIQUE restringe a sua busca às unidades densas de maiores dimensionalidades na **interseção** das unidades densas entre subespaços
- Utiliza propriedade **Apriori** de **mineração de dados** para indicar unidades densas *j-dimensionais*, onde *j* é o nº de dimensões.

CLIQUE (AGRAWAL *et al.*, 1998)

- Se uma unidade ***j* dimensional** é densa, então suas projeções no espaço **(*j* - 1) dimensional** também são densas.
- Para um candidato a **unidade densa *j* dimensional** é necessário checar suas **(*j* - 1) ésima unidades de projeção.** Se alguma unidade não é densa, então essa unidade densa não será *j*-ésima dimensional.

Análise: **Nº de Registros**

- *Complexidade computacional*:
  - **Tempo linear em relação ao número de objetos**, isto é, o tempo de processamento *independe do número de objetos*, pois *depende apenas do número de células* a ser utilizadas no agrupamento.

O(mk), onde:

m é o número de pontos,

k é o número de dimensões.

Análise: **Nº de Registros**

- Grid = 10, Dimensões = 50 e Clustering em cada diferente subespaço de 5 dimensões.

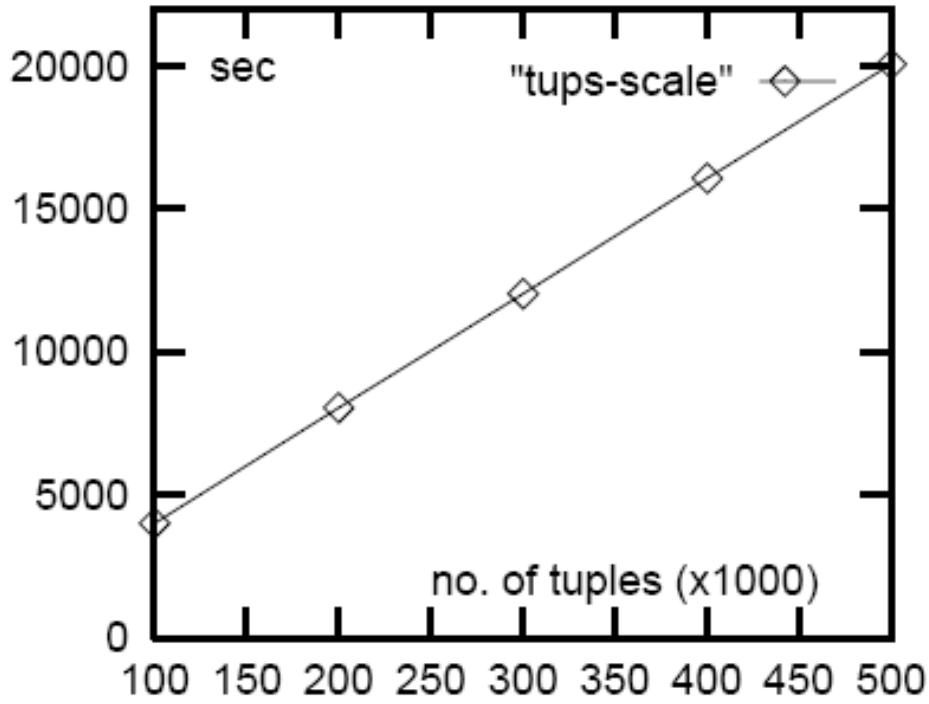
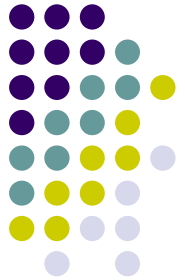


Figure 6: Scalability with the number of data records.

Análise: **Nº de Dimensões**



- *Complexidade computacional*:
  - **Tempo quadrático em relação ao número de dimensões**.

$O(c^k)$, onde:

c é o número de cluster,

k é o número de dimensões.

Análise: **Nº de Dimensões**

- Grid = 10, Registros = 100.000 e Clustering em cada diferente subespaços de 5 dimensões

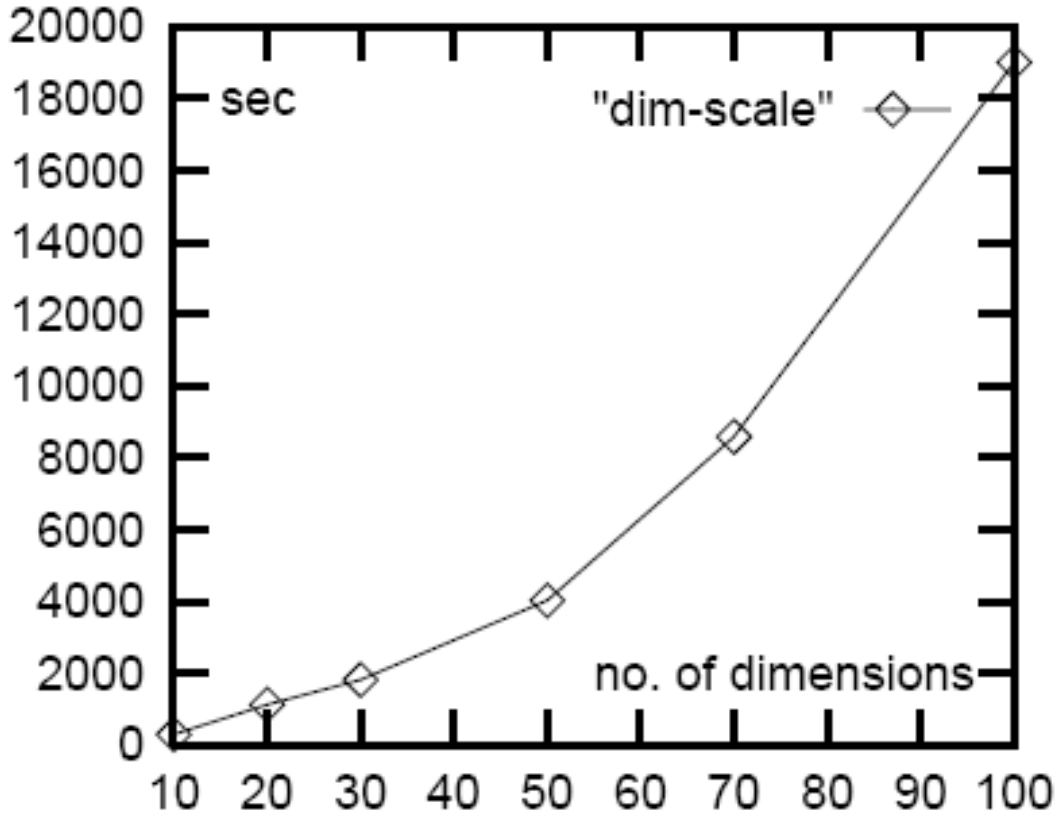
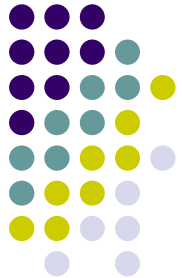


Figure 7: Scalability with the dimensionality of the data space.

CLIQUE (AGRAWAL *et al.*, 1998)



- *Complexidade computacional total*:

$$O(mk + c^k)$$

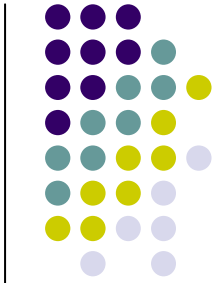
onde:

m : Número de Pontos de Entrada

k : Número de Dimensões Analisadas

c : Número de Clusters

CLIQUE (AGRAWAL *et al.*, 1998)

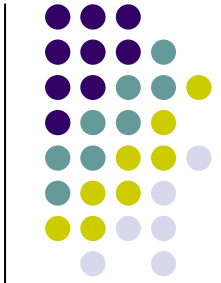


- *Vantagens*:
  - **Insensível** à ordem de entrada de dados.
  - **Escalável linearmente** com o tamanho da entrada.
  - **Bom escalonamento** conforme aumenta o número de dimensões (atributos).
  - Sucesso razoável no tratamento de **ruídos**.
    - Porém se densidade for baixa, unidades com ruídos serão consideradas densas e portanto entrarão em algum grupo.

CLIQUE (AGRAWAL *et al.*, 1998)

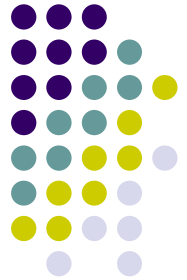
- *Desvantagem*:
  - **Precisão** do *clustering* pode ser degradada a custo da **simplicidade** do método.
  - O parâmetro de densidade é **constante** mesmo quando aumenta o número de dimensões.
  - Em experimentos, AGRAWAL *et al.* encontraram que em média **metade** dos pontos de clusters são considerados **ruídos** pelo método CLIQUE (motivo: eixos paralelos).

CLIQUE (AGRAWAL *et al.*, 1998)



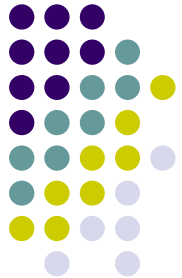
- *Desvantagem*:
  - **Não produz *clustering*** conforme a definição aceita da palavra, pois os pontos não estão particionados em grupos distintos. Pelo contrário, existe grande sobreposição entre as regiões densas descritas, devido ao fato que para uma dada região densa todas suas projeções nos subespaços de dimensionalidades mais baixas são também densas.

Referência Bibliográfica



- AGRAWAL, R., GEHRKE, J., GUNOPULOS, D., et al., 1998, “Automatic Subspace Clustering on High Dimensional Data for Data Mining Applications”, In: *Proceedings of the ACM SIGMOD Conference on Management of Data*, pp. 94-105, Seattle, Washington, USA, June.
- ILANGO, M.R., MOHAN, V., 2010, “A Survey of Grid Based Clustering Algorithms”, In: International Journal of Engineering Science and Technology, pp. 3441-3446, Vol. 2, Issue 8, India, August.

Referência Bibliográfica



- Pei, Yaling., Zaïane, Osmar. “Data Clustering Analysis”, Department of Computing Science, University of Alberta, Canada. Disponível em: http://webdocs.cs.ualberta.ca/~yaling/Cluster/ Acesso em 28/11/2010.