Engenharia de Sistemas e Computação
PESC/Coppe

UFRJ
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

Agrupando Redes com Atributos usando as Divergências de Bregman

Felipe Schreiber Fernandes

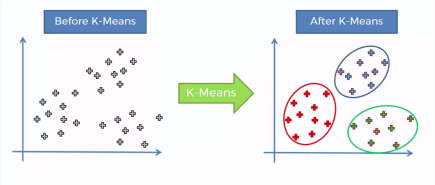
Orientadores:
Daniel Ratton Figueiredo (UFRJ)
Maximilien Dreveton (EPFL, Suíça)

January 22, 2025

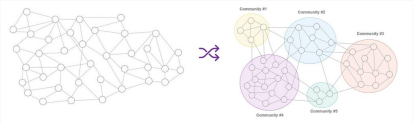
Engenharia de Sistemas e Computação PESC/Coppe

Clusterização

- Metodo para agrupar dados com similaridade.
- Problema estudado há mais de 40 anos (Kmeans 1956)
- Como encontrar agrupamentos?



- E se tivéssemos um grafo? (Louvain 2008)



Clusterização de Redes

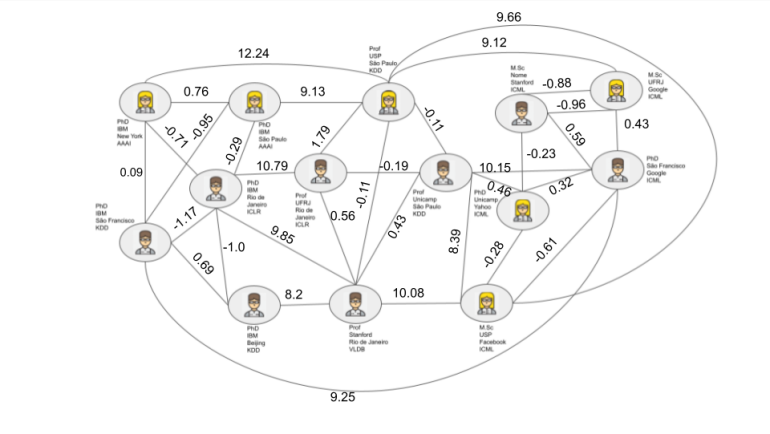


Figure 1: Rede de citações

Clusterização de Redes

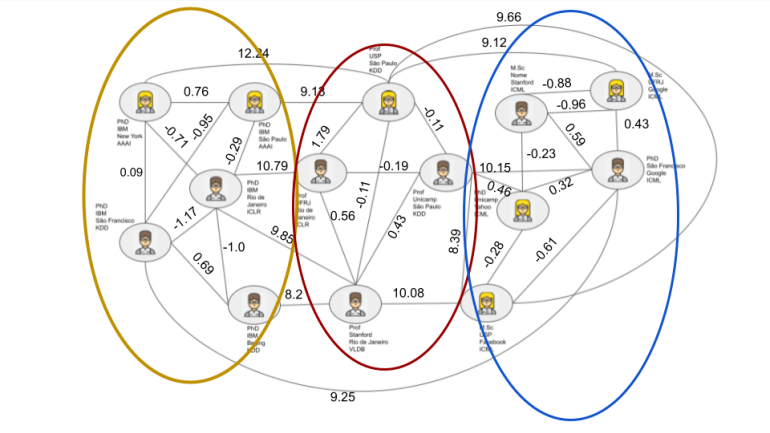


Figure 2: Rede de citações

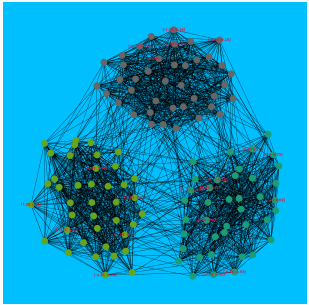
Engenharia de Sistemas e Computação PESC/Coppe

Nossa contribuição

- Modelo matemático para redes esparsas com atributos
- Algoritmos baseados em inferência estatística para obter o agrupamento
- Novos métodos de inicialização
- Avaliação e comparação do algoritmo em redes sintéticas e reais

Modelo proposto

- Assumimos a independência condicional entre atributos e rede
- Para a geração dos dados:
  - Gerar os atributos dos vértices $\mathbb{P}(Y|z) = \mathbb{P}(Y|z, \nu)$
  - Gerar as arestas e pesos da rede $\mathbb{P}(X|z) = \mathbb{P}(X|z, p, \mu)$



$$P = \begin{bmatrix} 0.4 & 0.02 & 0.02 \\ 0.02 & 0.4 & 0.02 \\ 0.02 & 0.02 & 0.4 \end{bmatrix} \qquad \mu = \begin{bmatrix} 10 & 9 & 9 \\ 9 & 10 & 9 \\ 9 & 9 & 10 \end{bmatrix}$$

$$\nu = \begin{Bmatrix} 1. & 0. \\ -0.5 & 0.86 \\ -0.5 & -0.86 \end{Bmatrix}$$

Modelo proposto (Cont.)

- A verossimilhança tem a seguinte forma:

$$\mathbb{P}(X, Y|z) = \prod_{1 \le i < j \le n} f_{z_i z_j}(X_{ij}) \prod_{i=1}^{n} h_{z_i}(Y_i)$$

- $f_{k\ell}(x)$ denota a probabilidade de dois vértices em blocos k e l possuírem uma relação $x \in X$. [a]
- $h_k(y)$ denota a probabilidade de um vértice no bloco $k \in [K]$ possuir um atributo $y \in Y$.
- Z é um vetor contendo as classes para cada vértice

A log-likelihood da densidade $p_{\psi,\theta}$ de uma distribuição da família exponencial se relaciona com a divergência de Bregman por:

$$\log p_{\psi,\theta}(x) = -d_{\psi^*}(x, \mu) + \psi^*(x),$$

onde $\mu = \mathbb{E}_{p_{\psi,\theta}}(X)$ é a média da distribuição.

[a] $f_{kl}(x) = (1 - p_{kl})\delta_0(x) + p_{kl} f^*_{kl}(x)$

Hard Clustering

Algorithm 7 Bregman hard clustering of node attributed SBM

Entrada: : Matriz de adjacências $X \in \mathcal{X}^{n \times n}$, atributos dos vértices $Y \in \mathcal{Y}^n$, funções convexas ψ^*, ϕ^*, labels iniciais $Z_{init} \in \mathcal{Z}_{n,K}$.

Computar p, μ, ν de acordo com as equações (4.11)

Faça

  Para cada $i = 1, \cdots, N$ **faça**

    Encontrar as comunidades de cada vértice i dados os parâmetros (4.10)

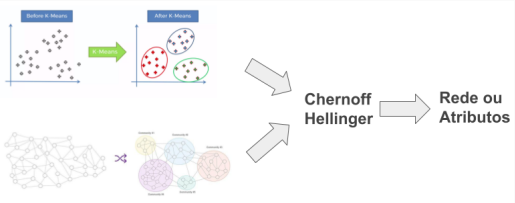
  Atualizar p, μ, ν de acordo com as equações (4.11)

Enquanto Não convergir

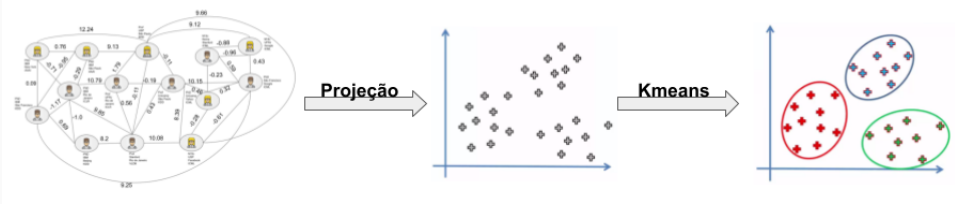
Retornar as comunidades Z

Inicialização

- Alternativa 1: Utilizar ou a rede ou os atributos



- Alternativa 2: Utilizar ambas as fontes de informação

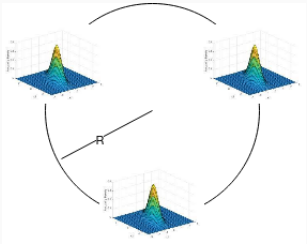


Avaliação

- Precisamos avaliar e comparar algoritmos em diferentes cenários.

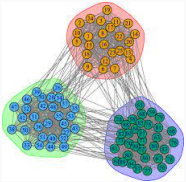
Atributos

- Especificar os centros das distribuições sobre um círculo de raio R



Rede

- Obter a rede pelo modelo SBM. $p_{in} = a\frac{\log n}{n}$, $p_{out} = b\frac{\log n}{n}$
- Especificar os pesos da rede w_{in} e w_{out}



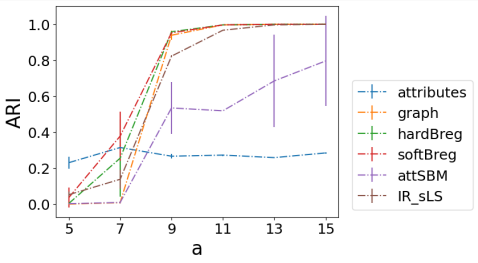
Comparação de Algoritmos

- Apenas atributos - GMM (Modelo de Mistura Gaussiana)
- Apenas a rede - Leiden
- Hard clustering com ambas as fontes de informação e Divergências de Bregman
- Soft clustering com ambas as fontes de informação e Divergências de Bregman
- Attributed Stochastic Block Model [1]
- Contextual Stochastic Block Model (IR_sLS) [2]

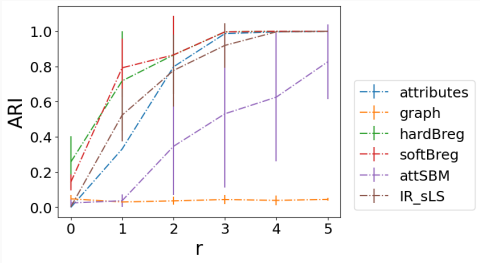
[1] STANLEY, N., BONACCI, T., KWITT, R., et al. "Stochastic Block Models with Multiple Continuous Attributes" - Applied Network Science 2019

[2] BRAUN, G., TYAGI, H., BIERNACKI, C. "An iterative clustering algorithm for the Contextual Stochastic Block Model with optimality guarantees" - ICML 2022

Resultados



(a) Aumentando a informação da rede



(b) Aumentando a informação dos atributos

Robustez a Distribuição

- Um dos parâmetros é a distribuição dos dados
- Entretanto a metodologia é robusta nesse sentido, de não utilizar a distribuição exata

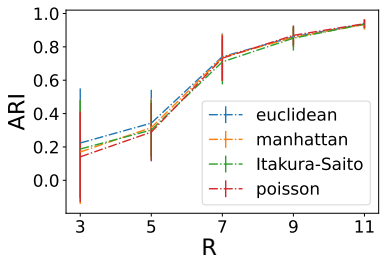


Figure 5: Várias divergências d_{ψ^*}. Poisson é o modelo correto.

Dados Reais

- **Cora**: $n = 2708$, $m = 10556$, $d = 1433$, $K = 7$;
- **Citeseer**: $n = 3327$, $m = 9104$, $d = 3703$, $K = 6$;
- **Wiscosin**: $n = 251$, $m = 515$, $d = 1703$, $K = 5$;
- **Texas**: $n = 183$, $m = 325$, $d = 1703$, $K = 5$;
- **Cornell**: $n = 183$, $m = 298$, $d = 1703$, $K = 5$;

n - Número de vértices; m - Número de arestas; d - Dimensão dos atributos; K - Número de comunidades

ARI	ARI_std	algorithm	dataset
0.23	0.0	both_soft+SC	CiteSeer
0.20	0.0	both_hard+SC	CiteSeer
0.16	0.0	attSBM	CiteSeer
0.02	0.0	leiden	CiteSeer
0.26	0.0	both_soft+SC	Cora
0.12	0.0	both_hard+SC	Cora
0.08	0.0	attSBM	Cora
0.18	0.0	leiden	Cora
0.44	0.0	both_soft+SC	Cornell
0.48	5.5e-17	both_hard+SC	Cornell
0.46	0.0	attSBM	Cornell
0.00	0.0	leiden	Cornell
0.45	0.0	both_soft+SC	Texas
0.32	0.0	both_hard+SC	Texas
0.44	0.0	attSBM	Texas
0.00	0.0	leiden	Texas
0.43	0.0	both_soft+SC	Wisconsin
0.40	0.0	both_hard+SC	Wisconsin
0.43	0.0	attSBM	Wisconsin
0.00	0.0	leiden	Wisconsin

Conclusão

- Nosso método generaliza vários algoritmos permitindo qualquer distribuição exponencial a priori e grafos esparsos com pesos
- Discutimos novos métodos de inicialização
- Avaliações empíricas demonstram a superioridade do nosso modelo ao comparar com algoritmos competitivos
- O método foi capaz de lidar bem com ambas as fontes de informação
- Artigo publicado na Neurips 2023

OBRIGADO

Trabalhos futuros

- Construir um algoritmo escalável, e comparar com outros benchmarks da literatura.
- Explorar outras divergências, de forma a generalizar para distribuições além da família exponencial, ex. t-student e Cauchy.
- Métodos de inicialização mais eficientes

Formulação (Cont.)

- Assumimos que as redes são esparsas, portanto:

$$f_{ab}(x) = (1 - p_{ab})\delta_0(x) + p_{ab} f^*_{ab}(x),$$

- Finalmente, assumimos que as distribuições $\{f^*_{ab}\}$ e $\{h_a\}$ pertencem à família exponencial:

$$f^*_{ab}(x) = e^{<\theta_{ab}, x> - \psi(\theta_{ab})} \quad \text{and} \quad h_a(y) = e^{<\eta_a, y> - \phi(\eta_a)}$$

- E.g. (gaussian): $f^*_{ab}(x) = e^{-\frac{(x-\mu_{ab})^2}{2\sigma^2}}$

Soft Clustering

Recall that ommiting the constants that depends on x we have:

- $f_{ab}(x) \;=\; \exp\{-\mathrm{d_{KL}}(A_{ij}, p_{ab}) - A_{ij} d_{\psi^*}(X_{ij}, \mu_{ab})\}$
- $h_{ab}(x) \;=\; \exp\{-d_{\phi^*}(x, \mu_a)\}$

Now we rewrite the Likelihood in the exponential form:

$$\begin{aligned}\mathbb{P}(X, Y \mid z) \;&=\; \prod_{1\le i<j\le n} f_{z_i z_j}(X_{ij}) \prod_{i=1}^{n} h_{z_i}(Y_i) \\ &=\; \exp\{-\sum_{i,j}\left[\mathrm{d_{KL}}(A_{ij}, p_{z_i,z_j}) + A_{ij} d_{\psi^*}(X_{ij}, \mu_{z_i,z_j})\right] - d_{\phi^*}(Y_i, \nu_{z_i})\}.\end{aligned}$$

Update rule

- **E-step**:

$$\tau_{ia} = p(z_i = a|X, Y, z_{-i}) \propto p(X, Y|z_{-i}, z_i = a)p(z_i = a)$$

$$\propto \pi_a \exp\left\{ -\sum_{i,j} \left[\mathrm{d_{KL}}(A_{ij}, p_{a,z_j}) + A_{ij} d_{\psi^*}(X_{ij}, \mu_{a,z_j}) \right] - d_{\phi^*}(Y_i, \nu_a) - c_i \right.$$

In practice, in order to have a stable exponent, we simply add a constant c_i for every node:

$$c_i = \min_a \{ \sum_j [d_{net}(j)] + d_{\phi^*}(Y_i, \nu_a) \}$$

- **M-step**:

$$\hat{\pi}_k = \frac{1}{n} \sum_i \hat{\tau}_{ik}, \quad \hat{\mu}_{k\ell} = \frac{\sum_{i \neq j} \hat{\tau}_{ik} \hat{\tau}_{j\ell} X_{ij}}{\sum_{i \neq j} \hat{\tau}_{ik} \hat{\tau}_{j\ell}} \quad \text{and} \quad \hat{\nu}_k = \frac{\sum_i \hat{\tau}_{ik} Y_i}{\sum_i \hat{\tau}_{ik}}.$$

Inicialização

Usamos a divergência de Chernoff-Hellinger para medir qual fonte dos dados nos dá maior informação:

- Com Z_{rede} para calcular:

$$C_{rede} = \min_{a \neq b} \sup_{t \in (0,1)} (1-t) \sum_{c=1}^{K} \pi_c \, \mathrm{D}_t \left(f_{bc} \| f_{ac} \right)$$

- Com $Z_{atributos}$ para calcular:

$$C_{atributos} = \min_{a \neq b} \sup_{t \in (0,1)} (1-t) \left[\frac{1}{n} \mathrm{D}_t \left(h_a \| h_b \right) \right]$$

- Retornar Z_{rede} se $C_{rede} > C_{atributos}$, do contrário retornar $Z_{atributos}$.

Inicialização Espectral

1. Construir duas matrizes de similaridade, uma para os atributos e outra para a rede
2. Obter o Laplaciano
3. Computar os K menores autovetores de cada, deixando primeiro de fora
4. Concatenar os autovetores e agrupar no espaço projetado

Inicialização Espectral

Algorithm 9 Spectral clustering on concatenated matrix

Require: : Observed network data X, attributes Y, a (symmetric) kernel function Φ, number of clusters K.

1: Do some preprocessing on X to obtain $\tilde{X}$ (e.g., compute Jaccard similarity between the neighbourhood of different nodes, normalize by degrees, or do nothing);

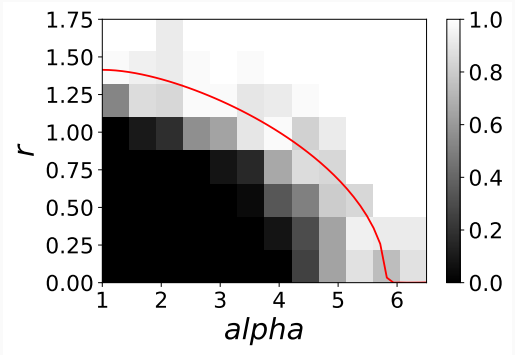
2: Let $\tilde{Y} \in \mathbb{R}_+^{n \times n}$ such that $\tilde{Y}_{ij} = K(Y_i, Y_j)$

3: (i) Let the eigendecomposition of $\tilde{X}$ be $\tilde{X} = \sum_{i=1}^{n} \lambda_i u_i u_i^t$, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and eigenvectors $u_1, \cdots, u_n$. Denote $U = (u_1, \cdots, u_K) \in \mathbb{R}^{n \times K}$ the leading eigenspace, and $\Lambda = (\lambda_1, \cdots, \lambda_K) \in \mathbb{R}^{K \times K}$ the leading eigenvalues.

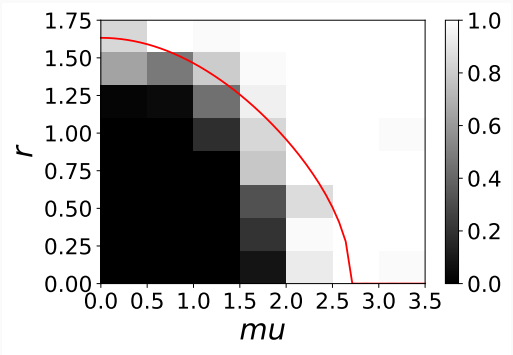
4: (ii) Similarly, denote let $\tilde{Y} = \sum_{i=1}^{n} \sigma_i v_i v_i^T$, with $\sigma_1 \geq \cdots \geq \sigma_n$. Denote $V = (v_1, \cdots, v_K)$ and $\Sigma = (\sigma_1, \cdots, \sigma_K)$.

5: (iii) Apply k-means on the rows of $[U\Lambda, V\Sigma] \in \mathbb{R}^{n \times 2K}$, where $[\cdot, \cdot]$ denotes the concatenation between two matrices.

6: Return estimated clusters $\hat{Z}$ obtained by k-means at step (iii).



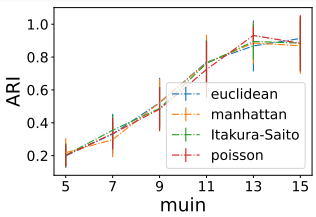
(a) Binary network with Gaussian attributes. Varying the p_{in} and r.



(b) zero-inflated Gaussian weights. We fix $w_{out} \sim \mathcal{N}(0,1)$. Varying $w_{in} \sim \mathcal{N}(\mu,1)$ and r

Figure 6: Phase transition of exact recovery. Each pixel represents the empirical probability that Algorithm 1 succeeds at exactly recovering the clusters (over 50 runs), and the red curve shows the theoretical threshold.

Resultados



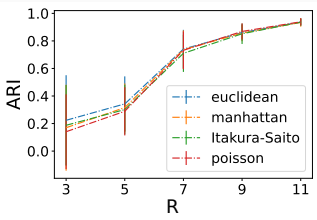
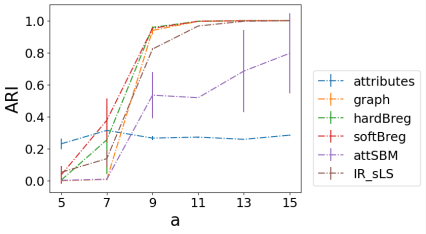


Figure 7: Adjusted Rand Index (ARI) averaged over 20 realisations
(a) $n = 400$, $k = 4$, $f_{in} = (1-p)\delta_0(x) + p\mathrm{Poi}(\mu_{in})$ and $f_{out} = (1-q)\delta_0(x) + q\mathrm{Poi}(5)$, with $p = 0.04$ and $q = 0.01$. Attributes are 2d-Gaussians with unit variances and mean equally spaced the circle of radius $r = 2$.
(b) $n = 400$, $k = 2$, $f_{in} = (1-p)\delta_0(x) + p\mathrm{Nor}(2,1)$ and $f_{out} = (1-q)\delta_0(x) + q\mathrm{Nor}(0,1)$, with $p = 0.04$ and $q = 0.01$. Attributes are Poisson with means ν_1 (for nodes in cluster 1) and 3

Resultados



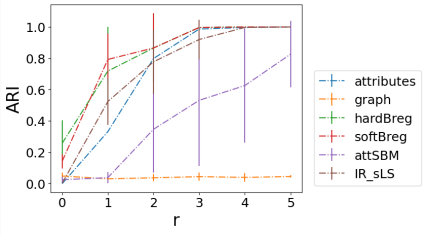


Figure 8: (a)Fixamos $p_{out} = 5\frac{\log n}{n}$ e $r = 1$. Variamos $p_{in} = a\frac{\log n}{n}$
(b) Fixamos $p_{out} = 5\frac{\log n}{n}$, e $p_{in} = 8\frac{\log n}{n}$. Variamos o raio R.
Ambos experimentos: atributos Gaussianos, grafo sem peso

Contextual Stochastic Block Model [3]

- It also assumes Stochastic Block Model for the network and Gaussian Mixtures for the attributes.
- Doesn't rely on MLE.
- Optimization procedure is done via iterative refinement.

[3] An iterative clustering algorithm for the Contextual Stochastic Block Model with optimality guarantees [**braun2022iterative**] - ICML 2022

S&S	S&S_std	CC	CC_std	algorithm	dataset
0.53	0.00	0.06	0.01	both_hard	CiteSeer
0.60	0.00	0.20	0.00	both_hard+SC	CiteSeer
0.52	0.00	0.04	0.01	both_soft	CiteSeer
0.62	0.00	0.24	0.00	both_soft+SC	CiteSeer
0.55	0.00	0.10	0.00	kmeans	CiteSeer
0.54	0.01	0.07	0.01	both_hard	Cora
0.57	0.00	0.13	0.00	both_hard+SC	Cora
0.62	0.04	0.24	0.07	both_soft	Cora
0.63	0.00	0.26	0.00	both_soft+SC	Cora
0.53	0.00	0.05	$6.94 \cdot 10^{-18}$	kmeans	Cora
0.51	0.00	0.03	0.01	both_hard	Cornell
0.75	0.00	0.50	0.00	both_hard+SC	Cornell
0.54	0.02	0.08	0.04	both_soft	Cornell
0.73	0.00	0.46	0.00	both_soft+SC	Cornell
0.68	0.00	0.36	0.00	kmeans	Cornell
0.57	0.04	0.14	0.07	both_hard	Texas
0.66	0.00	0.33	0.00	both_hard+SC	Texas
0.67	0.04	0.34	0.07	both_soft	Texas
0.73	0.00	0.45	0.00	both_soft+SC	Texas
0.74	0.00	0.48	$5.55 \cdot 10^{-17}$	kmeans	Texas
0.58	0.02	0.16	0.03	both_hard	Wisconsin
0.70	0.00	0.40	0.00	both_hard+SC	Wisconsin
0.64	0.01	0.28	0.02	both_soft	Wisconsin
0.72	0.00	0.44	$5.55 \cdot 10^{-17}$	both_soft+SC	Wisconsin
0.69	0.00	0.37	0.00	kmeans	Wisconsin