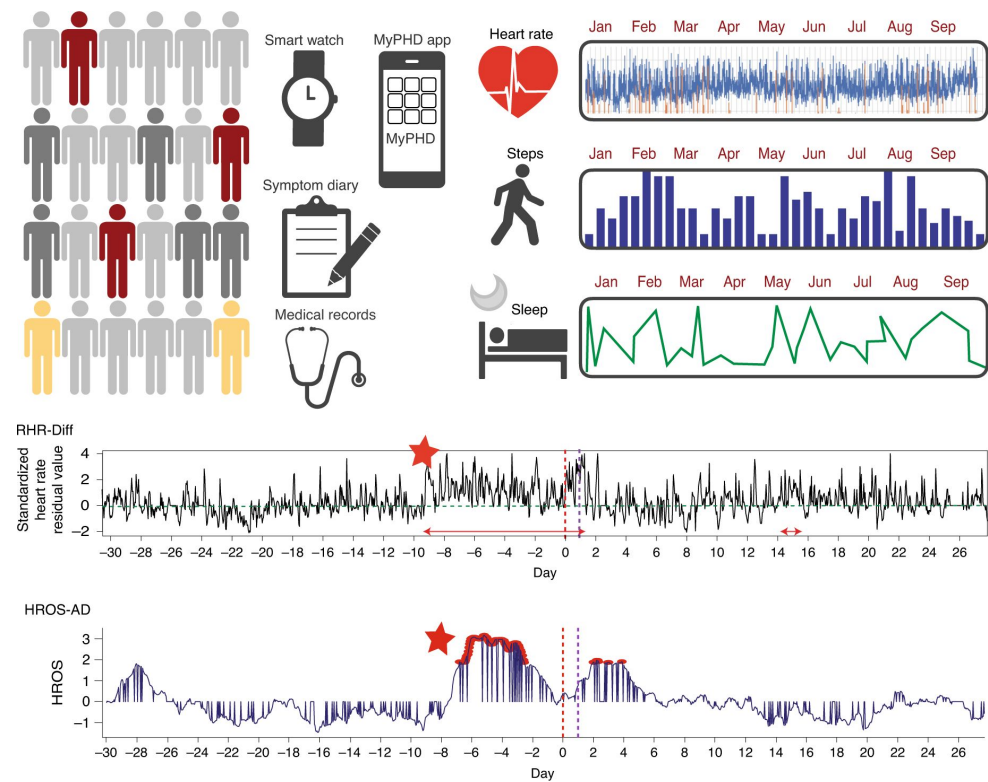


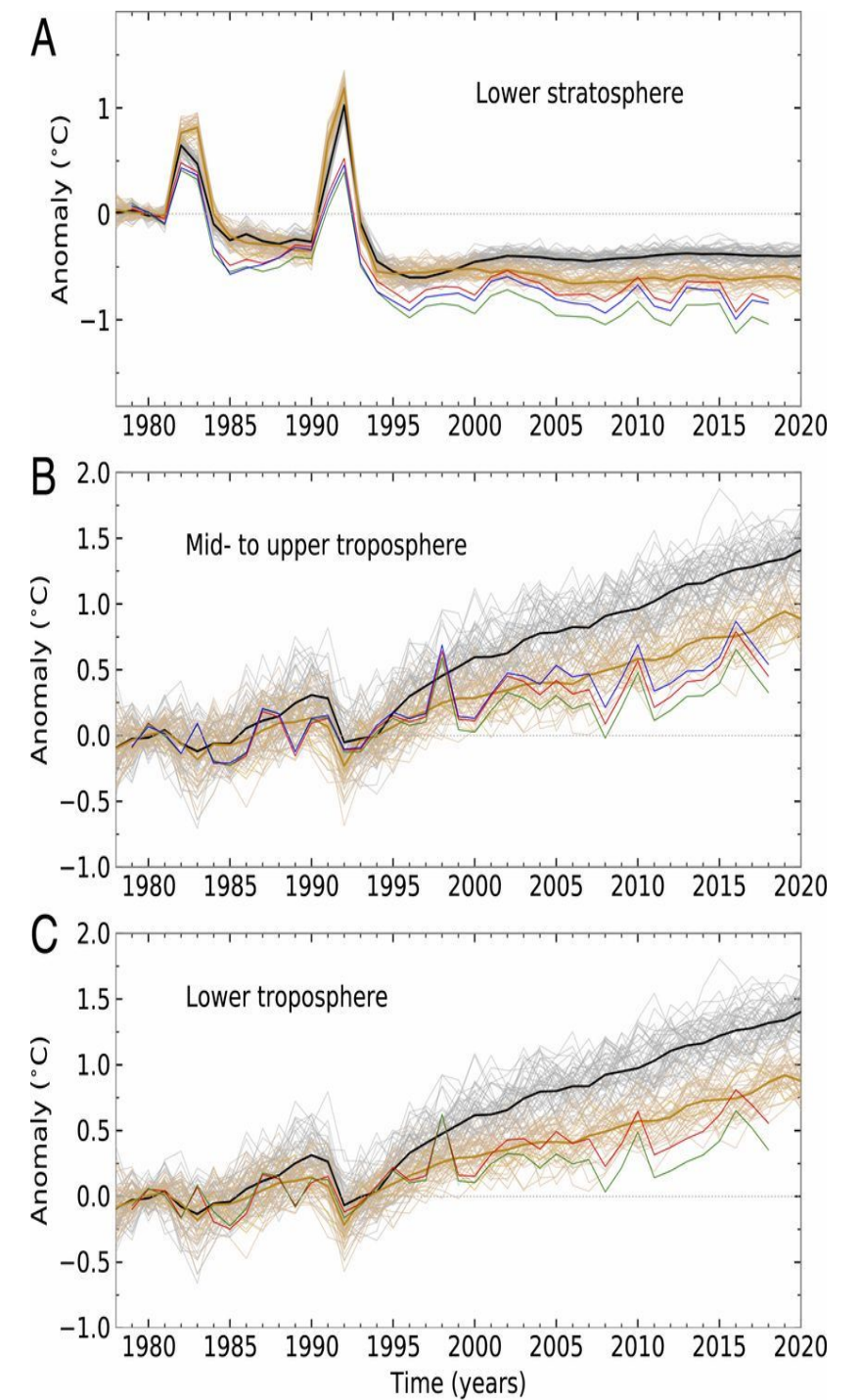
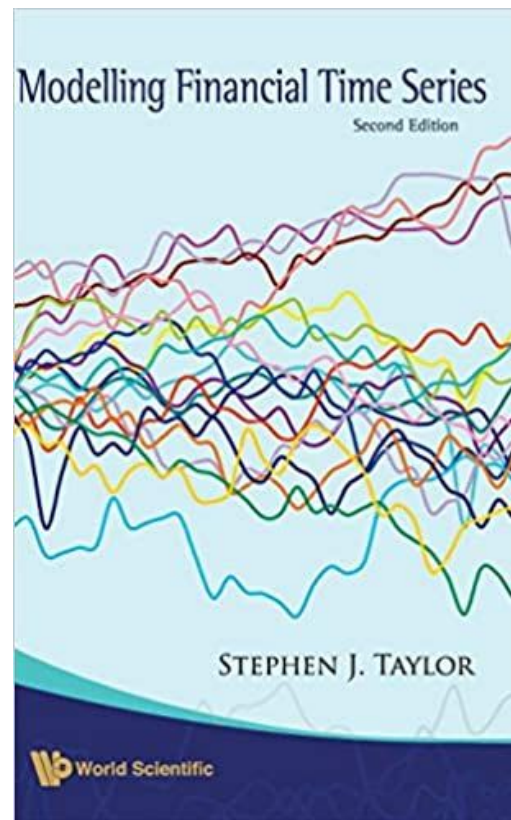


# Clustering de Series de Tiempo

Andrés Abeliuk - Hernán  
Sarmiento



Mishra, T.. et al. Pre-symptomatic detection of COVID-19 from smartwatch data. *Nat Biomed Eng* (2020).



Santer, Benjamin D., et al. "Quantifying stochastic uncertainty in detection time of human-caused climate signals." *PNAS* (2019)

# Clustering basado en la forma usando el algoritmo K-means

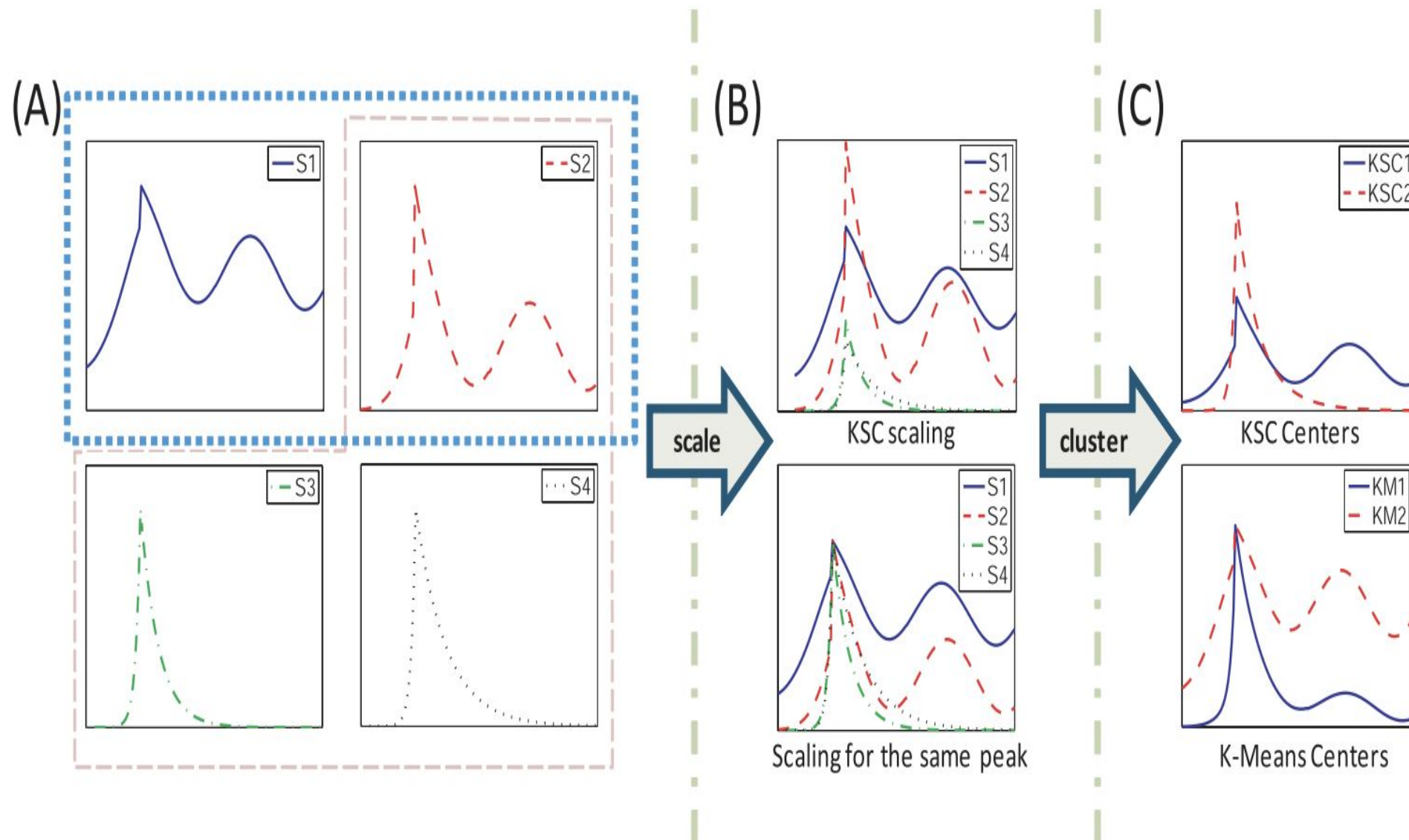


Figure 2: (A) Four time series,  $S1, \dots, S4$ . (B) Time series after scaling and alignment. (C) Cluster centroids. K-Means wrongly puts  $\{S1\}$  in its own cluster and  $\{S2, S3, S4\}$  in the second cluster, while K-SC nicely identifies clusters of two vs. single peaked time series.

[Yang, Jaewon, and Jure Leskovec. "Patterns of temporal variation in online media." \*ACM international conference on Web search and data mining\*. 2011](#)

# Medida de distancia

$$\hat{d}(x, y) = \min_{\alpha, q} \frac{\|x - \alpha y_{(q)}\|}{\|x\|}$$

$$F = \sum_{k=1}^K \sum_{x_i \in C_k} \hat{d}(x_i, \mu_k)^2.$$

$$\mu_k^* = \arg \min_{\mu} \sum_{x_i \in C_k} \hat{d}(x_i, \mu)^2.$$

# Cálculo de Centroide

$$\mu_k^* = \arg \min_{\mu} \sum_{x_i \in C_k} \min_{\alpha_i, q_i} \frac{\|\alpha_i x_{i(q_i)} - \mu\|^2}{\|\mu\|^2}$$

Finally, substituting  $\sum_{x_i \in C_k} (I - \frac{x_i x_i^T}{\|x_i\|^2})$  by  $M$  leads to the following minimization problem:

$$\mu_k^* = \arg \min_{\mu} \frac{\mu^T M \mu}{\|\mu\|^2}. \quad (4)$$

Resultado de álgebra lineal: La solución de este problema es el vector propio (eigenvector) u correspondiente al valor propio más pequeño  $\lambda$  de la matriz M



# Extensión: Series de tiempo multidimensionales

---

**Algorithm 1** m-kSC Algorithm

---

**Input:**  $\{\mathcal{X}, K\}$  where  $\mathcal{X} \in \mathbb{R}^{N \times D \times M}$  is the tensor containing  $N$  multidimensional time series and  $K$  is number of clusters.

**Output:**  $\{\mathcal{C}, S\}$  where  $\mathcal{C} \in \mathbb{R}^{K \times D \times M}$  is the tensor of cluster centroids and  $S$  contains each cluster assignments.

```
1: Initialize cluster assignments  $S$  randomly
2: while  $S$  changes on every iteration do
3:   for  $k = 1 : K$  do
4:     for  $d = 1 : D$  do
5:        $M = \sum_{\mathbf{x}_n \in S_k} (I - \frac{\mathbf{x}_n(d, :)\mathbf{x}_n(d, :)^T}{\|\mathbf{x}_n(d, :)\|^2})$ 
6:        $\mathbf{C}(k, d, :) =$  Smallest eigenvector of  $M$ .
7:     end for
8:   end for
9:   for  $n = 1 : N$  do
10:     $k = \underset{k=1, \dots, K}{\operatorname{argmindist}}(\mathbf{c}_k, \mathbf{x}_n)$  using Eq. 1
11:     $S(n) = k$ 
12:   end for
13: end while
```

---

Patrones temporales de la evolución de los 1.000 repositorios de GitHub más populares.

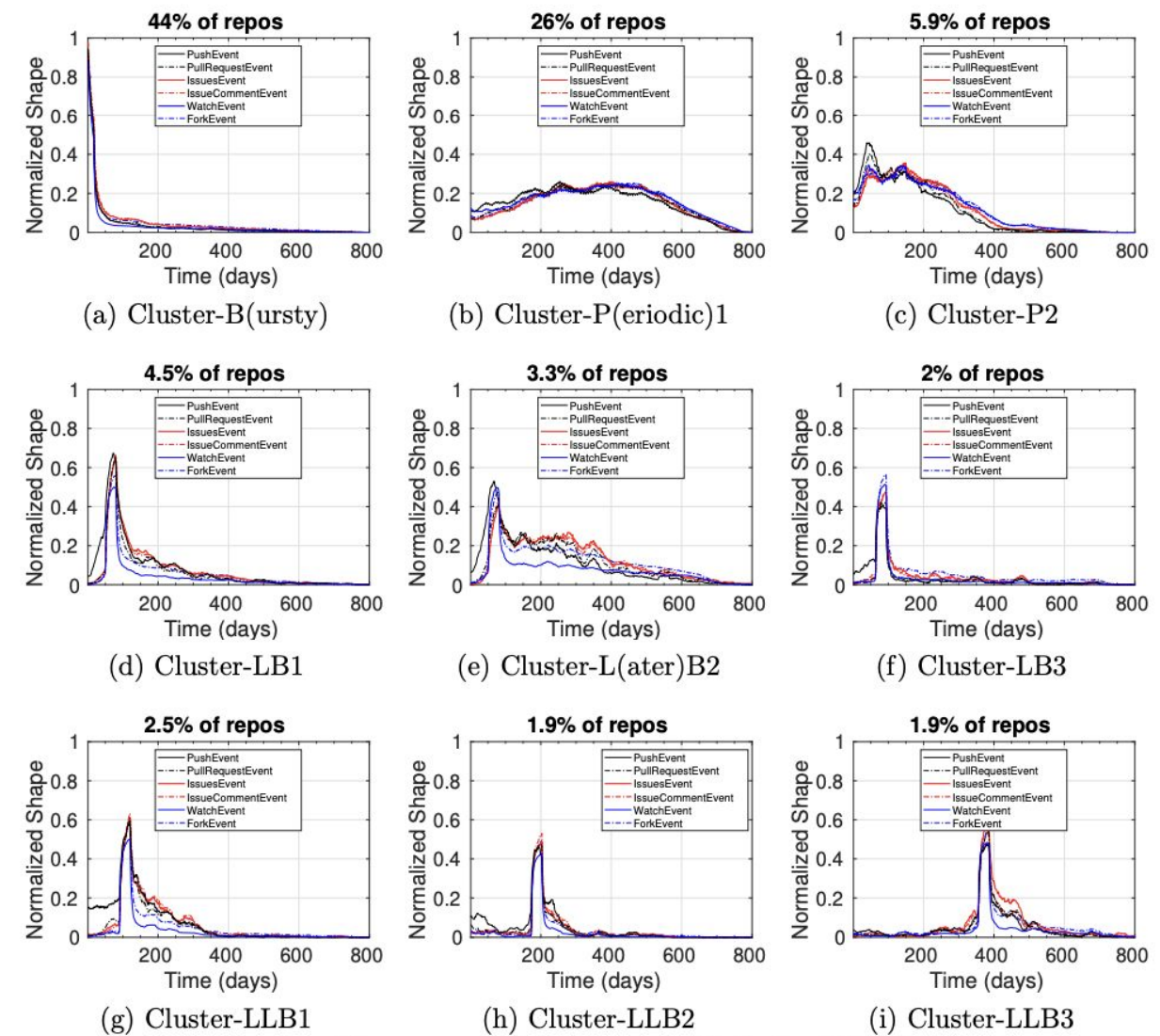


Figure 3: Shapes of the uncovered cluster centroids in the GitHub dataset.

Ozer, M., Sapienza, A., Abeliuk, A., Muric, G., & Ferrara, E. (2020). Discovering patterns of online popularity from time series. *Expert Systems with Applications*.

# Patrones de popularidad en Twitter

Análisis de la línea de tiempo de los top mil hashtags más populares de Twitter.

Datos desde el 14 de febrero hasta 6 de marzo 2018, relacionado con el tiroteo en la escuela de Parkland.

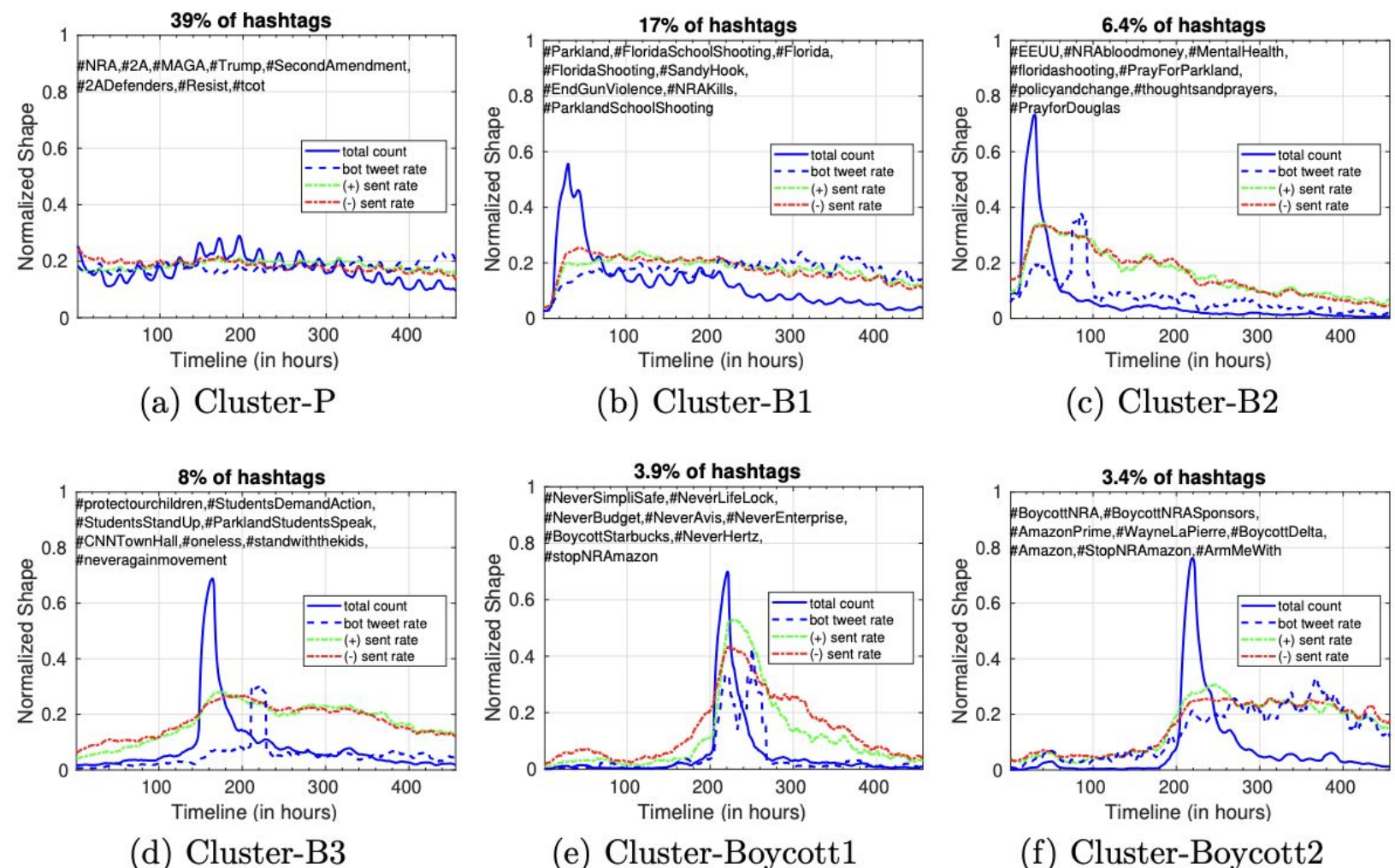


Figure 7: Shape of the uncovered clusters of Twitter





**dcc**

CIENCIAS DE LA COMPUTACIÓN  
UNIVERSIDAD DE CHILE

[www.dcc.uchile.cl](http://www.dcc.uchile.cl)

f @ in  / DCCUCHILE