



# Detección de tópicos en un conjunto de notas del Diario La Nación.



# Detección de tópicos

**Objetivos: detectar de forma semi-automática los tópicos en un conjunto de notas surgidas del diario.**

Presentación: análisis de métodos sobre un corpus conocido:

- Descripción vectorial de las notas (tf-idf).
- Detección de tópicos: Non-negative matrix factorization (NMF).
- Evaluación de tópicos encontrados.

# Detección de tópicos

Objetivos: detectar de forma semi-automática los tópicos en un conjunto de notas surgidas del diario.

**Presentación: análisis de métodos sobre un corpus conocido:**

- Descripción vectorial de las notas (tf-idf).
- Detección de tópicos: Non-negative matrix factorization (NMF).
- Evaluación de tópicos encontrados.

# Detección de tópicos

Objetivos: detectar de forma semi-automática los tópicos en un conjunto de notas surgidas del diario.

Presentación: análisis de métodos sobre un corpus conocido:

- **Descripción vectorial de las notas (tf-idf).**
- **Detección de tópicos: Non-negative matrix factorization (NMF).**
- **Evaluación de tópicos encontrados.**

# Detección de tópicos

Objetivos: detectar de forma semi-automática los tópicos en un conjunto de notas surgidas del diario.

Presentación: análisis de métodos sobre un corpus conocido:

- **Descripción vectorial de las notas (tf-idf).**
- Detección de tópicos: Non-negative matrix factorization (NMF).
- Evaluación de tópicos encontrados.

# Term frequency - Inverse document frequency **(Tf-idf)**

- **Describo las notas como vectores en un espacio multidimensional.**
- El espacio vectorial son, en principio, **todos los términos de los documentos (notas)**. Además de palabras, se pueden incluso incorporar **n-gramas** (conjunto de palabras. Ej: “casa rosada”).
- Cada documento es descripto por un vector, cuyas componentes son la **multiplicación** entre la cantidad de ocurrencias en el documento de un dado término (**tf**), multiplicado por una valorización (**idf**).  
**Idf(t)** cuantifica qué tan específico es un término en un conjunto de documentos.

# Term frequency - Inverse document frequency (Tf-idf)

Vector  
documento

$$v = [\dots, tf(t) \cdot idf(t), \dots]$$

Cantidad de veces que aparece el término  $t$  en el documento.

$$tf(t) = 1 + \log \left( \frac{1+N}{1+n_t} \right)$$

Para  $n_t$  más grandes,  $idf(t)$  es más chico → términos más específicos de un documento tienen mayor valorización.

$N$ : # total de documentos.

$n_t$ : # documentos donde aparece el término  $t$ .

# Ejemplo

## **Documentos:**

- 1) La casa de Julia
- 2) La casa de Cristian
- 3) De la casa de Seba

Espacio de 6 dimensiones.

julia

seba

cristian

la      casa

de

$$idf = 1 + \log\left(\frac{1+3}{1+1}\right) = 1$$

$$idf = 1 + \log\left(\frac{1+3}{1+1}\right) = 1.69$$

$$v_3 = [1, 1, 2, 1.69, 0, 0]$$

“de” es importante porque **es más frecuente** en el documento.  
“seba” es importante por **específico** del documento.

Conjunto de notas →  
Matriz documentos x términos

*F términos*

$$M = \begin{matrix} & \text{\color{blue} N documentos} \\ \text{\color{black} N} & \end{matrix}$$

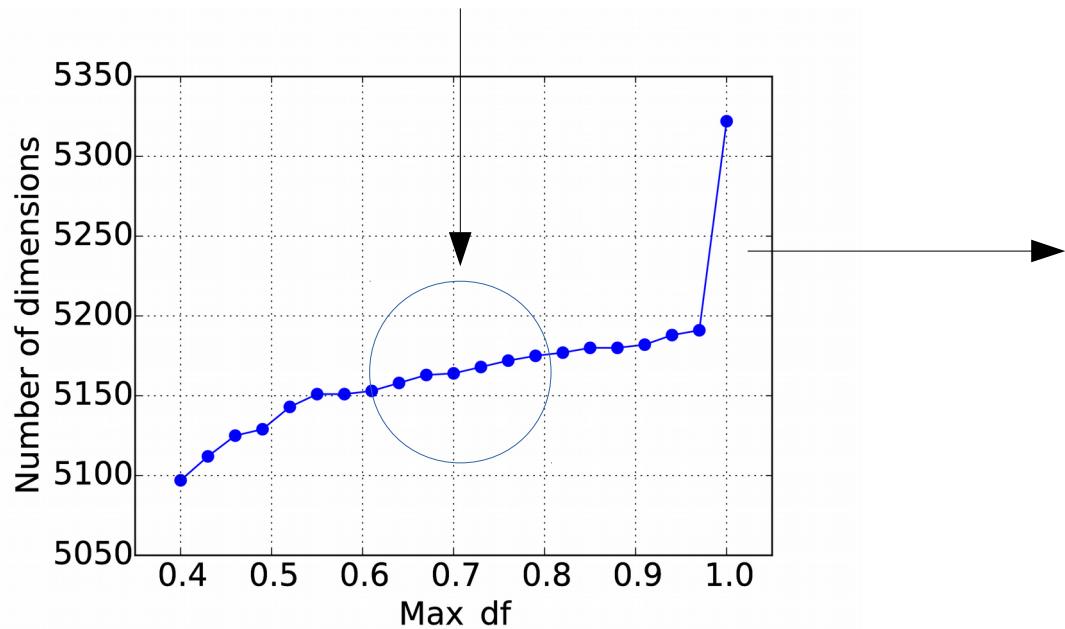
$V_{11}$	$V_{12}$	$V_{13}$	.....	$V_{1F}$
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
$V_{N1}$	$V_{N2}$	$V_{N3}$	.....	$V_{NF}$

$$\text{Dim}(M) = N \times F$$

# Term frequency - Inverse document frequency (Tf-idf)

## Parámetros que se pueden regular y valores escogidos:

- **Max\_df: descartar términos por muy occurrentes:** Podemos fijar un límite superior en la cual si un término aparece en más documentos que ese límite lo descartamos. Tomo los **términos que aparezcan en menos del 70% de los documentos**.



Solo se observa un **cambio abrupto** en la dimensión del espacio **cuando se descartan los términos que aparecen** en prácticamente **todos los documentos**.

# Term frequency - Inverse document frequency (Tf-idf)

## Parámetros que se pueden regular y valores escogidos:

- **Min\_df: descartar términos por raros:** Podemos fijar un límite inferior, es decir, solo tomamos términos que aparezcan en más de un cierto dado de documentos. Tomo los **términos que aparezcan en dos o más documentos**.
- **N-gram\_range:** podemos tomar, además de palabras sueltas, bigramas, trigramas, etc. **Tomo de 1 a 3 gramas.**
- **Los vectores pueden estar o no normalizados:** los vectores documentos pueden estar normalizados a norma 1 → vectores dentro de la esfera unitaria multidimensional, o bien no estarlo. En principio la mayor parte de los cálculos lo hago con vectores normalizados.

# Descripción de corpus de prueba

**Documentos = 31 notas** del Diario La Nación donde claramente podemos identificar **4 tópicos**:

- 10 notas sobre ***Tarifas***.
- 7 notas sobre ***Tratado de paz en Colombia***.
- 7 notas sobre ***Campaña de Trump***.
- 7 notas sobre ***Acuerdo Malvinas***.

Las notas están descriptas en un espacio de **5164 dimensiones**.

**Recordar:** esto surge de tomar todos los **1-gramas, 2-gramas y 3-gramas**, y **descartar** aquellos que aparezcan en más del 70% de las notas y **en menos de dos notas** (es decir aquellos que aparezcan en solo una nota).

# Detección de tópicos

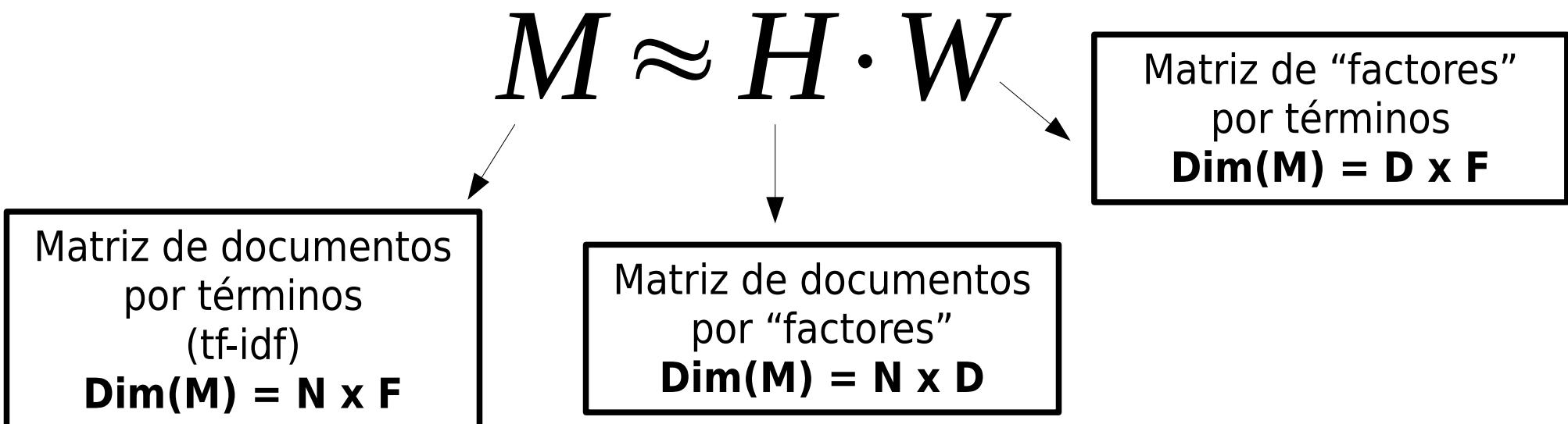
Objetivos: detectar de forma semi-automática los tópicos en un conjunto de notas surgidas del diario.

Presentación: análisis de métodos sobre un corpus conocido:

- Descripción vectorial de las notas (tf-idf).
- **Detección de tópicos: Non-negative matrix factorization (NMF).**
- Evaluación de tópicos encontrados.

# Reducción de dimensionalidad **NMF**

**Non-negative matrix factorization:** técnica de reducción de dimensionalidad, descompone en forma aproximada una matriz con componentes no-negativos como la **multiplicación de dos matrices con componentes no-negativos.**



D es un parámetro a elegir

# Reducción de dimensionalidad **NMF**

## Ventajas:

- La interpretación puede ser más natural que en **PCA** (*principal component analysis*), ya que todos los documentos están representados con vectores de componentes positivas.
- La identificación de conjuntos de documentos surge naturalmente, sin necesidad de aplicar un algoritmo de clusterización, ej K-means.

## Desventajas:

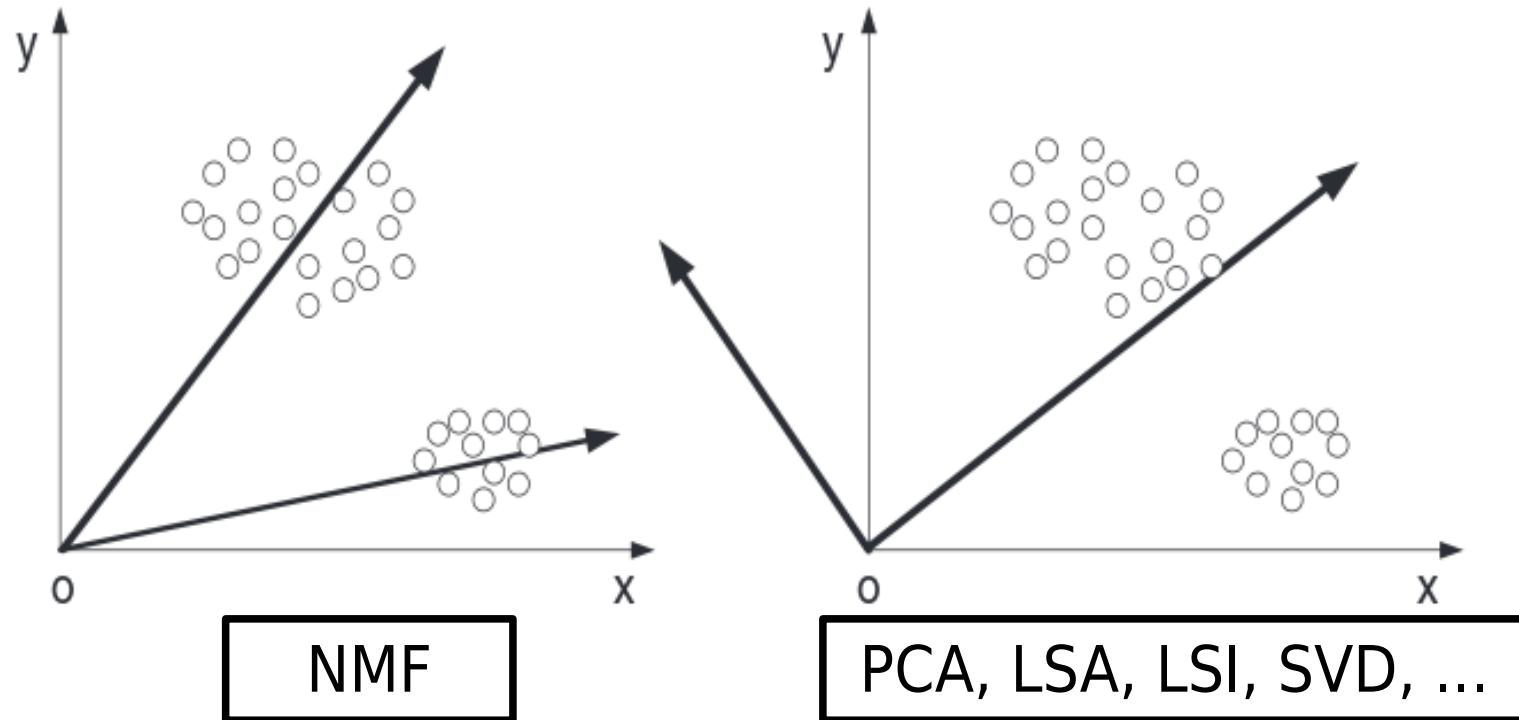
- Es un método aproximado.
- Hay que elegir  $D$  con un criterio.

$$M \approx H \cdot W$$



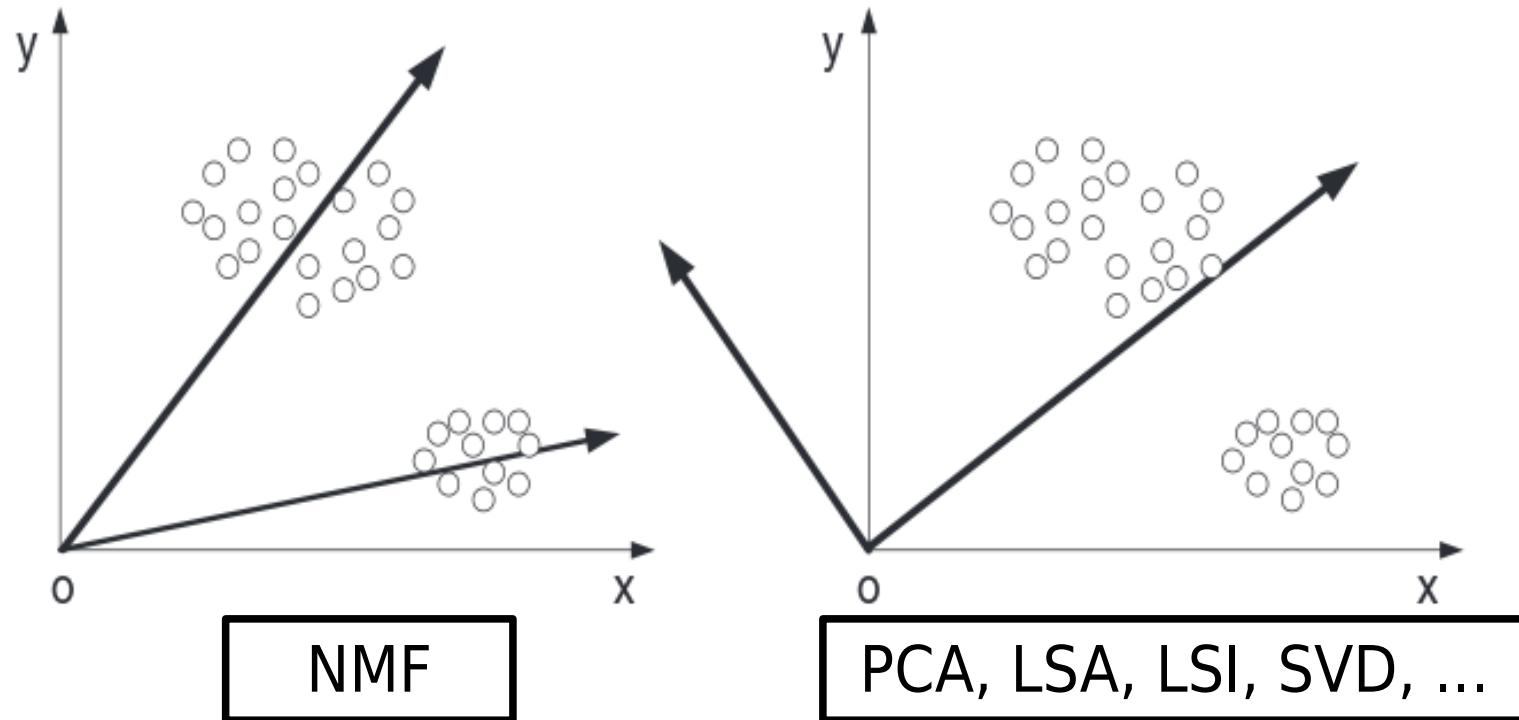
Matriz de documentos  
**reducida**

# Reducción de dimensionalidad **NMF**



**NMF** encuentra nuevas **direcciones no-ortogonales**, donde **cada dirección representa en si misma un tópico**.

# Reducción de dimensionalidad **NMF**



**El tópico de cada documento será el componente más pesado en el espacio reducido.**

# Reducción de dimensionalidad **NMF**

## Metodología:

- **El resultado de NMF depende de la condición inicial** (matrices aleatorias con componentes no negativas). Por lo tanto, al calcular distintos observables muestro el promedio de 1000 inicializaciones.
- Al mostrar un **resultado particular**, tomo la **descomposición NMF con el mínimo error** de un conjunto de condiciones iniciales.

# Enfoque redes complejas

- Se interpreta cada documento como un nodo en una red compleja. Este enfoque me va a permitir definir ciertas métricas de evaluación de particiones.
- Podemos construir una red pesada, definiendo los pesos entre los nodos  $i$  y  $j$  como:

$$w_{ij} = \vec{v}_i \cdot \vec{v}_j = \cos(\theta)$$

Vectores  
documentos con  
norma 1

Medida de  
similitud

$$0 < w_{ij} < 1$$

# Reducción de dimensionalidad **NMF**

**NMF** permite **mantener** el criterio de **construcción de los pesos de la red** como producto escalar entre los vectores documentos (previamente normalizados). Con *PCA*, los componentes negativos de los vectores obligarían a modificar el criterio.

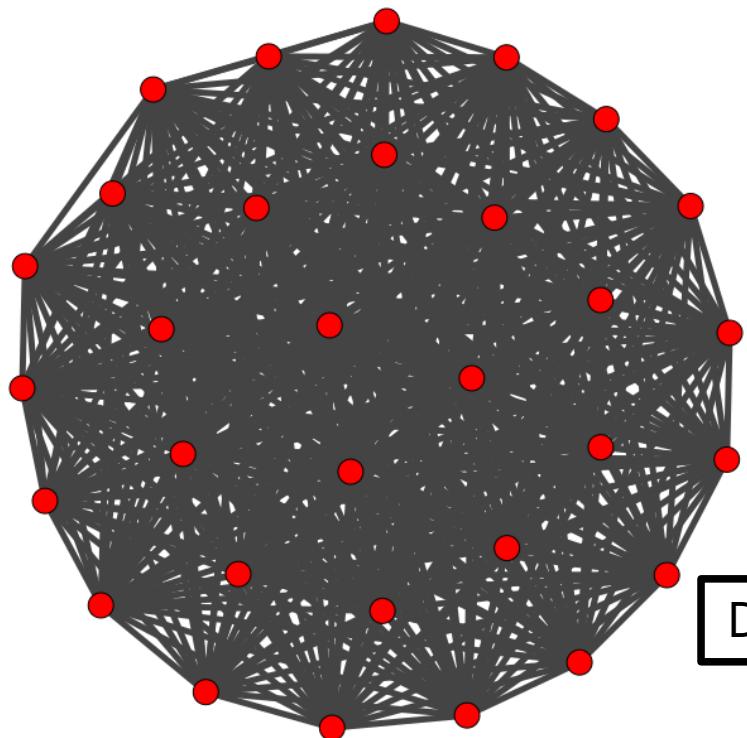
$$w_{ij} = \vec{v}_i \cdot \vec{v}_j = \cos(\theta)$$



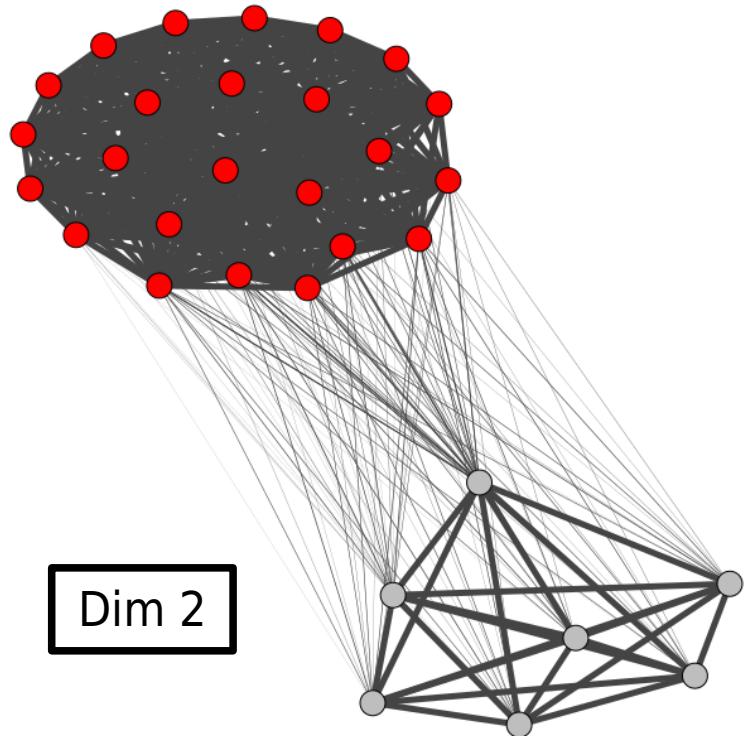
Matriz de  
documentos  
**reducida**

Matriz de  
adyacencia  
**pesada**

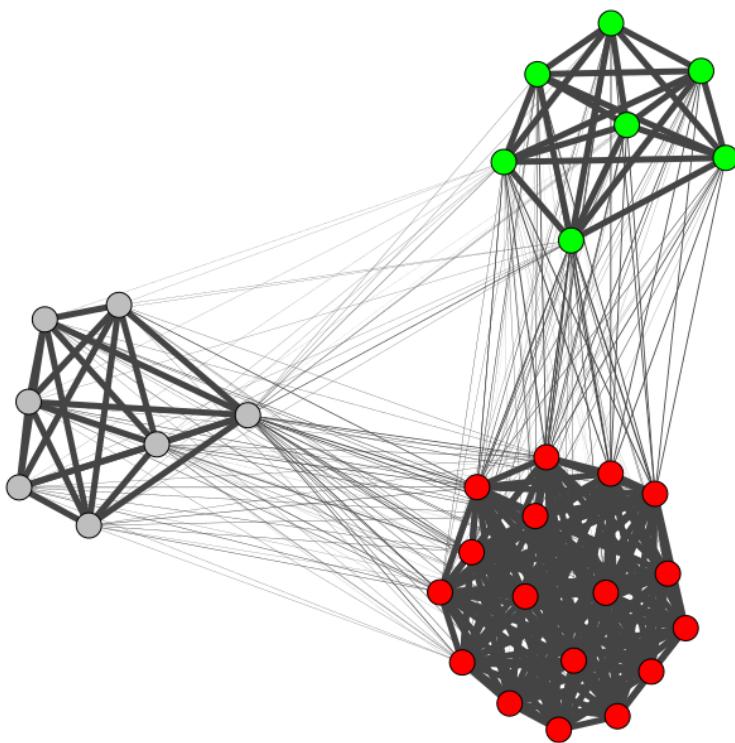
Mantengo las etiquetas dadas por *NMF*, no utilizo ningún algoritmo de detección de comunas en la red.



Dim 1



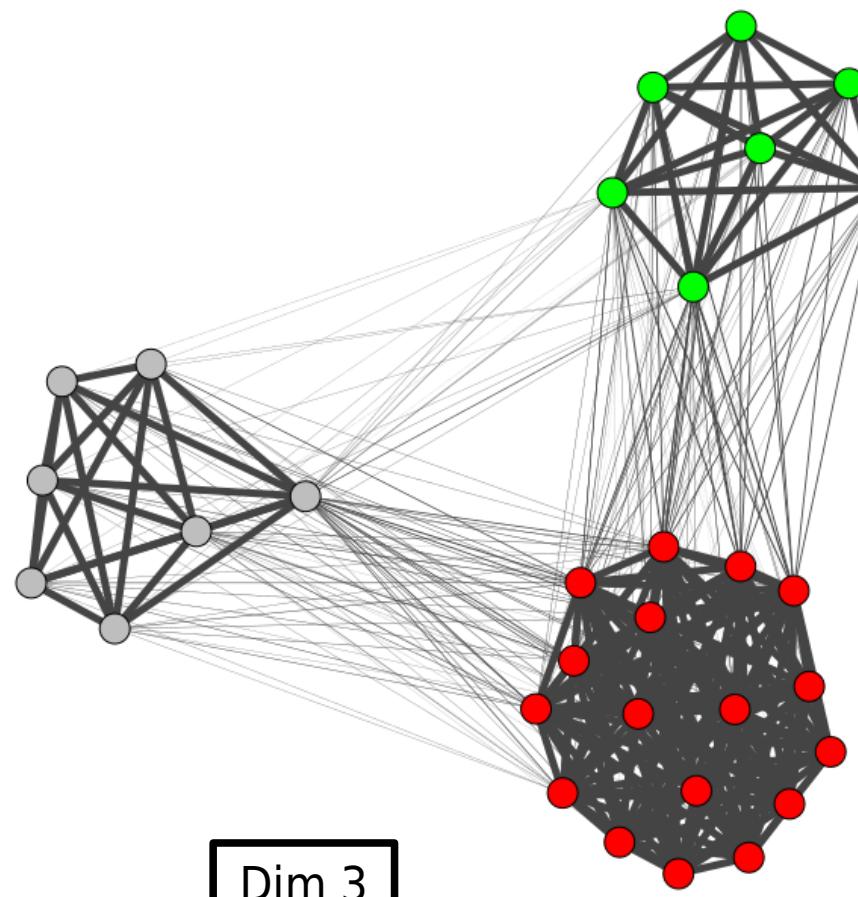
Dim 2



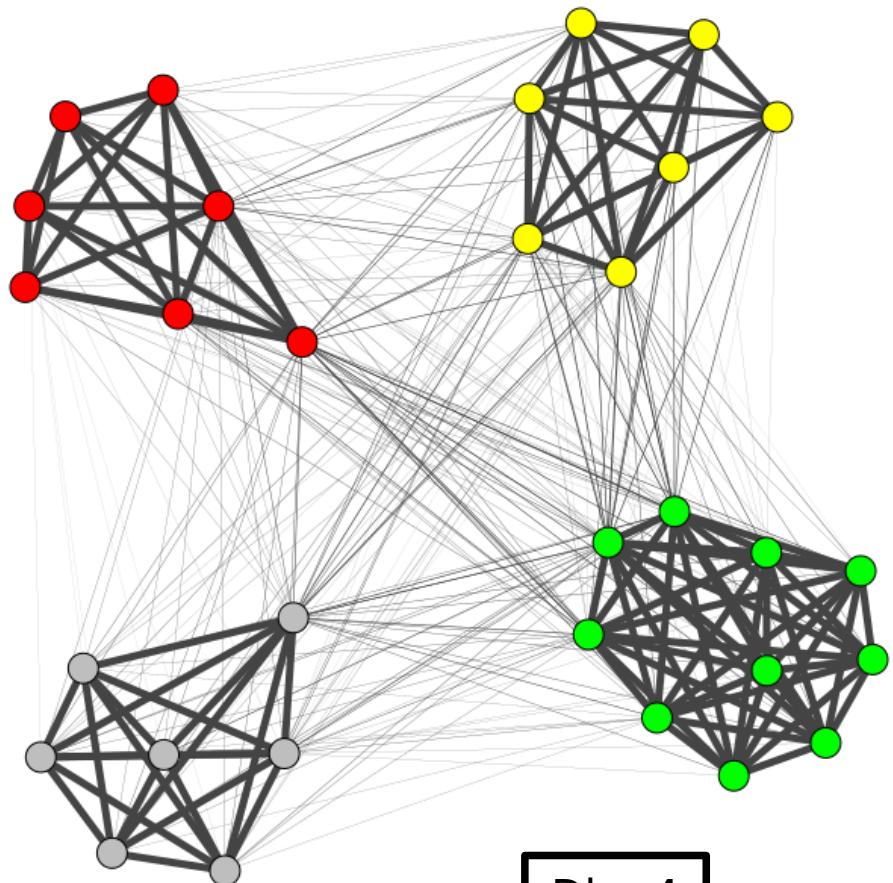
Etiqueta de  
comuna  
dada por  
NMF

Dim 3

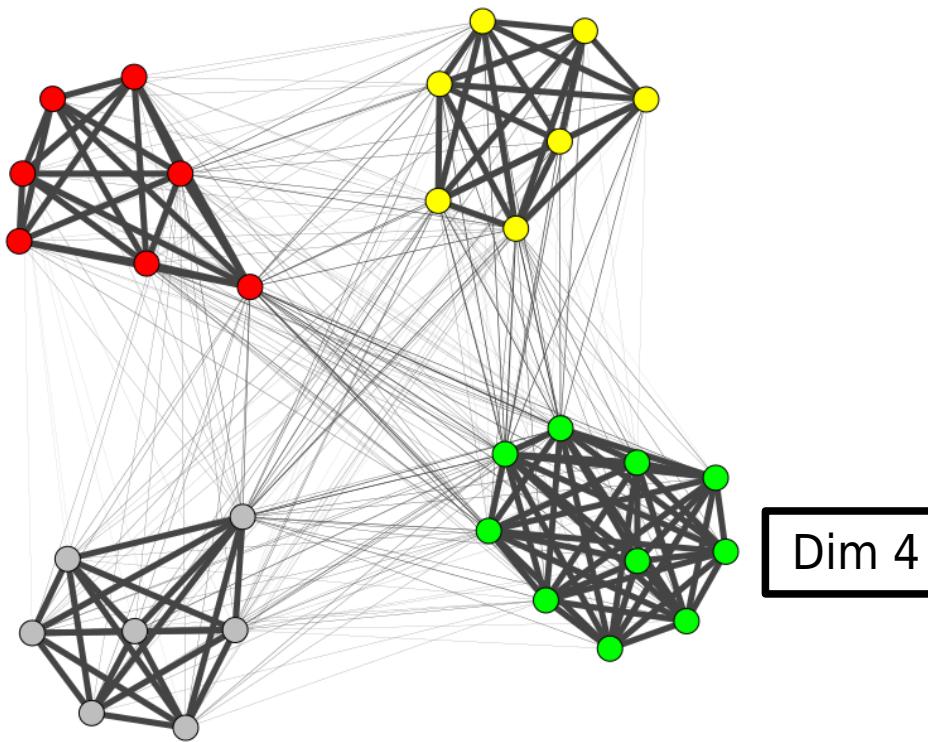
Los colores pueden  
no mantenerse de  
layout a layout



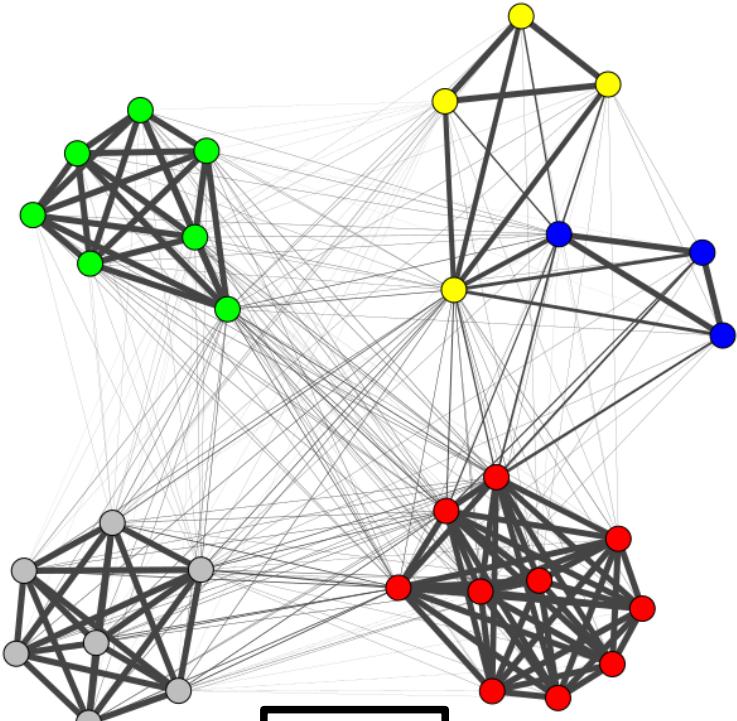
Dim 3



Dim 4

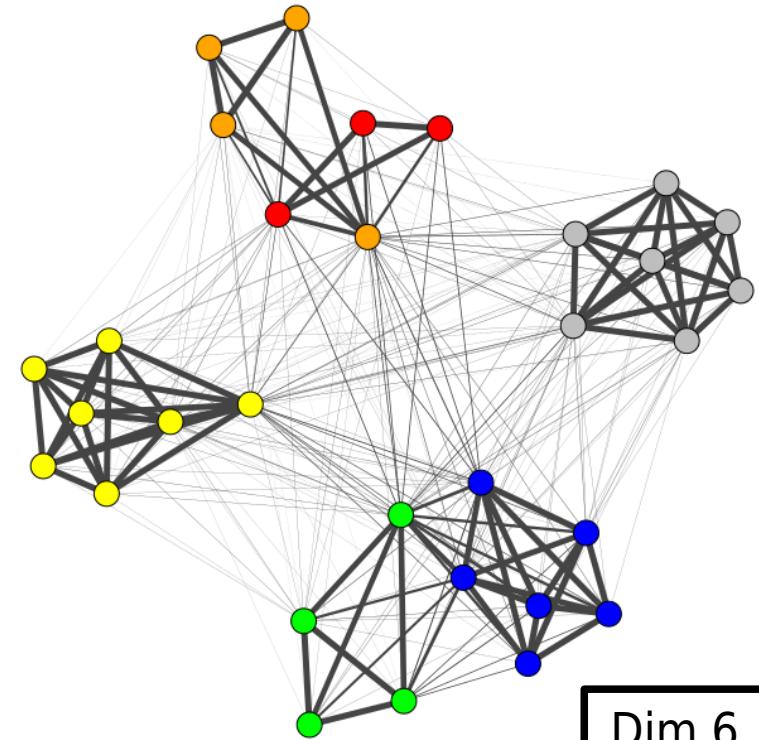


Dim 4



Dim 5

Mejores  
resultados para  
100 condiciones  
iniciales.



Dim 6

# Detección de tópicos

Objetivos: detectar de forma semi-automática los tópicos en un conjunto de notas surgidas del diario.

Presentación: análisis de métodos sobre un corpus conocido:

- Descripción vectorial de las notas (tf-idf).
- Detección de tópicos: Non-negative matrix factorization (NMF).
- **Evaluación de tópicos encontrados.**

# Evaluación de las particiones

- Coeficiente de *silhouette* de un punto  $i$ :

Distancia media entre el punto  $i$  y los puntos del cluster más cercano (mínima distancia media).

Distancia media entre el punto  $i$  y los puntos dentro del mismo cluster.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$-1 < s(i) < 1$

Un  $s(i)$  cercano a 1 → punto  $i$  más cercano a los puntos de su mismo cluster, y/o más alejado de los puntos de los clusters vecinos.

**Para evaluar las particiones se toma el valor medio de  $s$ .**  
El número de particiones van de 2 a ( $N^{\circ}$ puntos - 1).

# Evaluación de las particiones

- Coeficiente de *silhouette* de un punto  $i$ :

Distancia entre nodo  $i$  y  $j$  ~ Disimilitud entre  $i$  y  $j$ :

$$dis(i,j) = 1 - w_{ij}$$

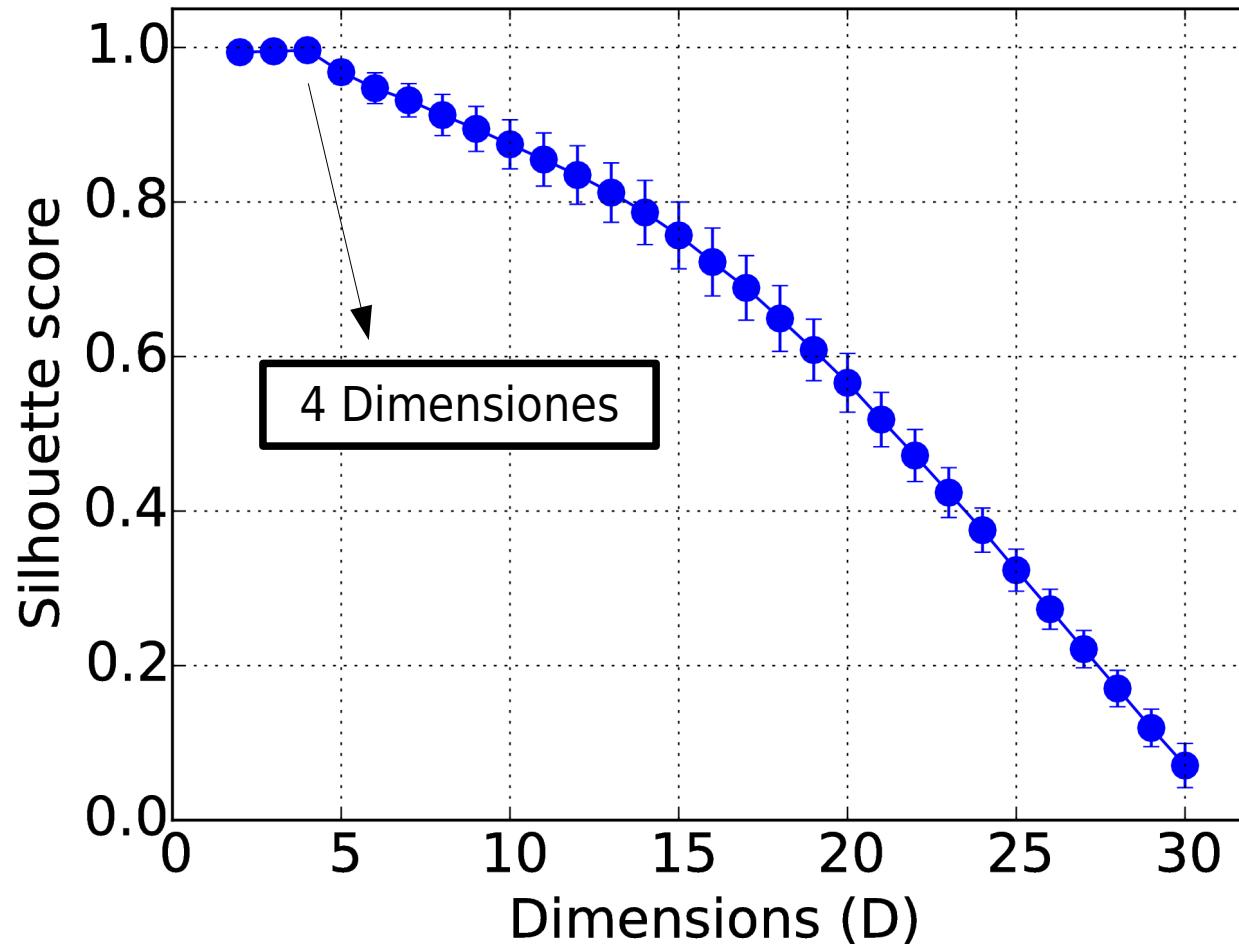


~ distancia entre dos nodos en una red completa y pesada.

Peso del enlace entre 0 y 1



# Evaluación de las particiones



**Coeficiente de silhouette en función del número de dimensiones empleadas:**

A partir de 4 dimensiones, el coeficiente comienza a decaer.

# Evaluación de las particiones

- Modularidad (hasta que encuentre la implementación de otro coeficiente...):

Matriz de adyacencia pesada.

Suma de los pesos sobre el nodo  $i$

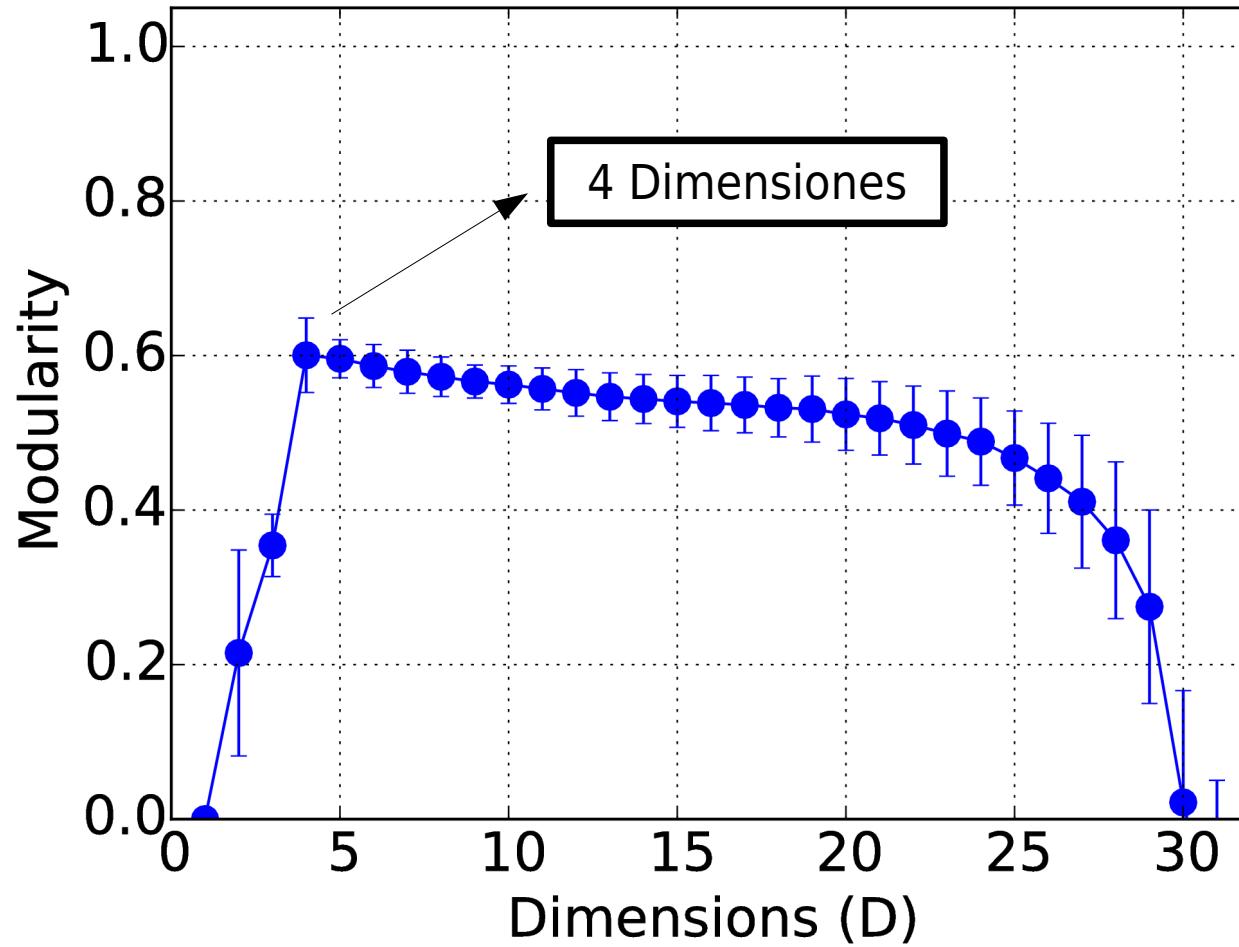
N.<sup>º</sup> de enlaces en la red.

Etiquetas de los comunas.

$$Q = \frac{1}{2m} \sum_{ij} \left[ A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

$$-1 < Q < 1$$

# Evaluación de las particiones



**Modularidad en función del número de dimensiones empleadas:**  
Con 4 dimensiones la modularidad trepa a aproximadamente 0.60.  
**Allí se produce una alta ganancia en modularidad.**

# Evaluación de las particiones

- Weak merit factor ( $Q_w$ ): (*implementado más adelante*)

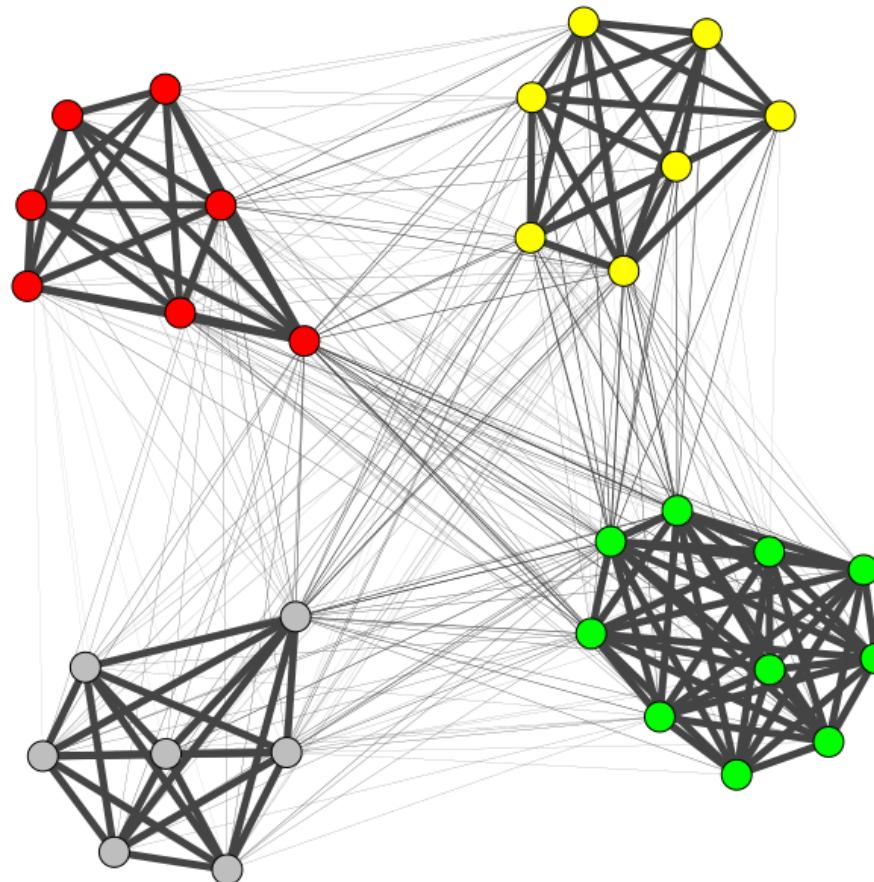
Peso del enlace entre i y los nodos dentro del mismo cluster ( $in$ ) y de clusters vecinos ( $out$ ).

$$Q_W = \sum_{j=1}^m S(C_j) = \sum_{j=1}^m \sum_{i \in C_j} \frac{k_i^{in} - k_i^{out}}{2L(C_j)}$$

Community strength

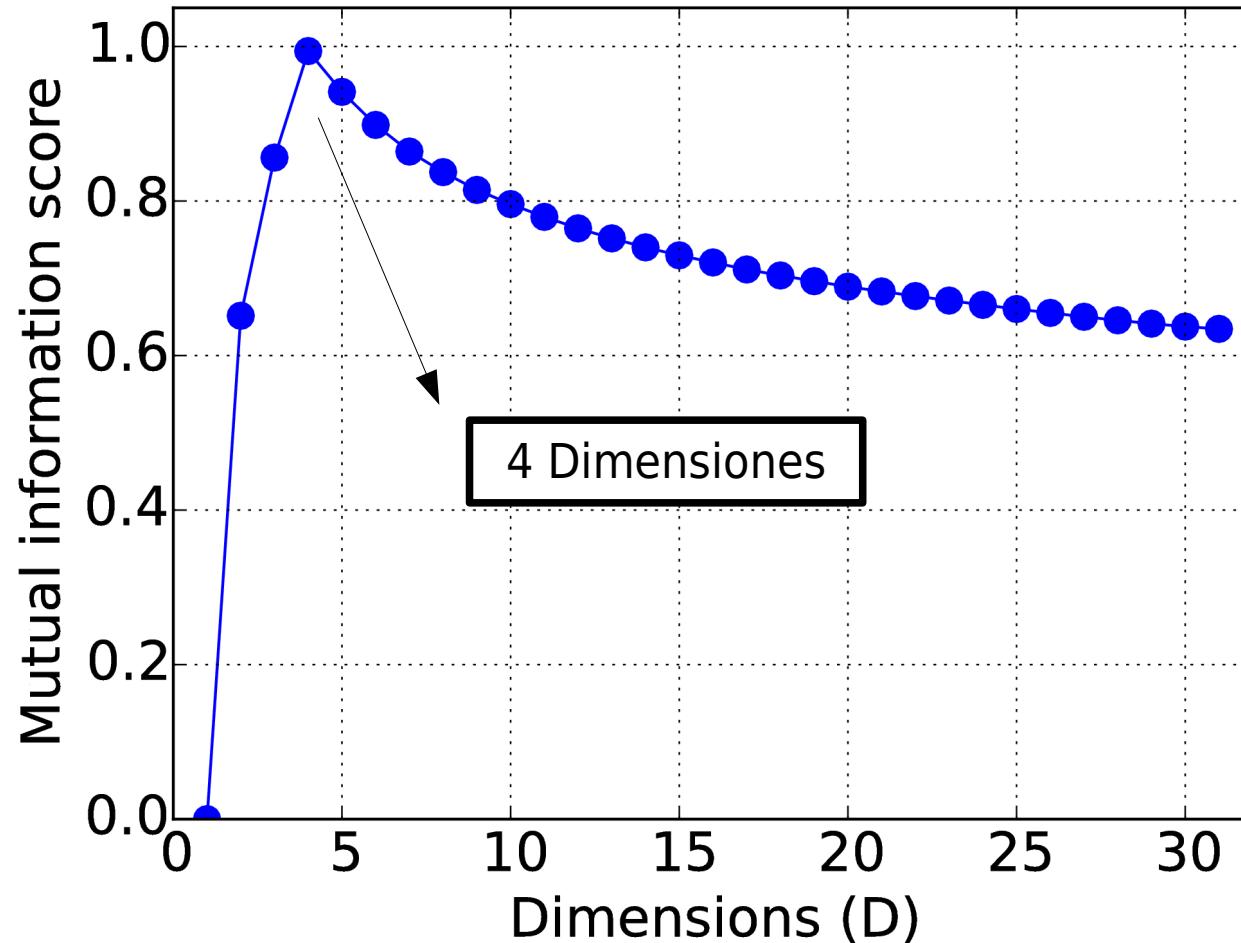
$$L(C_j) = (1/2) \sum_{i \in C_j} k_i$$

# Mejor partición → 4 tópicos



**Layout de la red pesada**, luego de procesar los datos con **NMF** y **D = 4**.

# Evaluación de las particiones



**Correlación entre los tópicos encontrados y los tópicos esperados.**  
Los 4 tópicos encontrados en la mejor partición se corresponde con lo esperado.

# Interpretación NMF

Las dimensiones surgidas de **NMF** son vectores con componentes no negativos, representados en el espacio surgido de **tfidf**.

Componentes con mayor peso, ordenados de mayor a menor:

- **Dimensión 1:** *farc, las farc, paz, acuerdo, colombia, santos, la paz, guerrilla, de las farc, colombiano.*
- **Dimensión 2:** *trump, clinton, hillary, campaña, republicano, de trump, ryan, magnate, demócrata, hillary clinton.*

# Interpretación NMF

Las dimensiones surgidas de **NMF** son vectores con componentes no negativos, representados en el espacio surgido de **tfidf**.

Componentes con mayor peso, ordenados de mayor a menor:

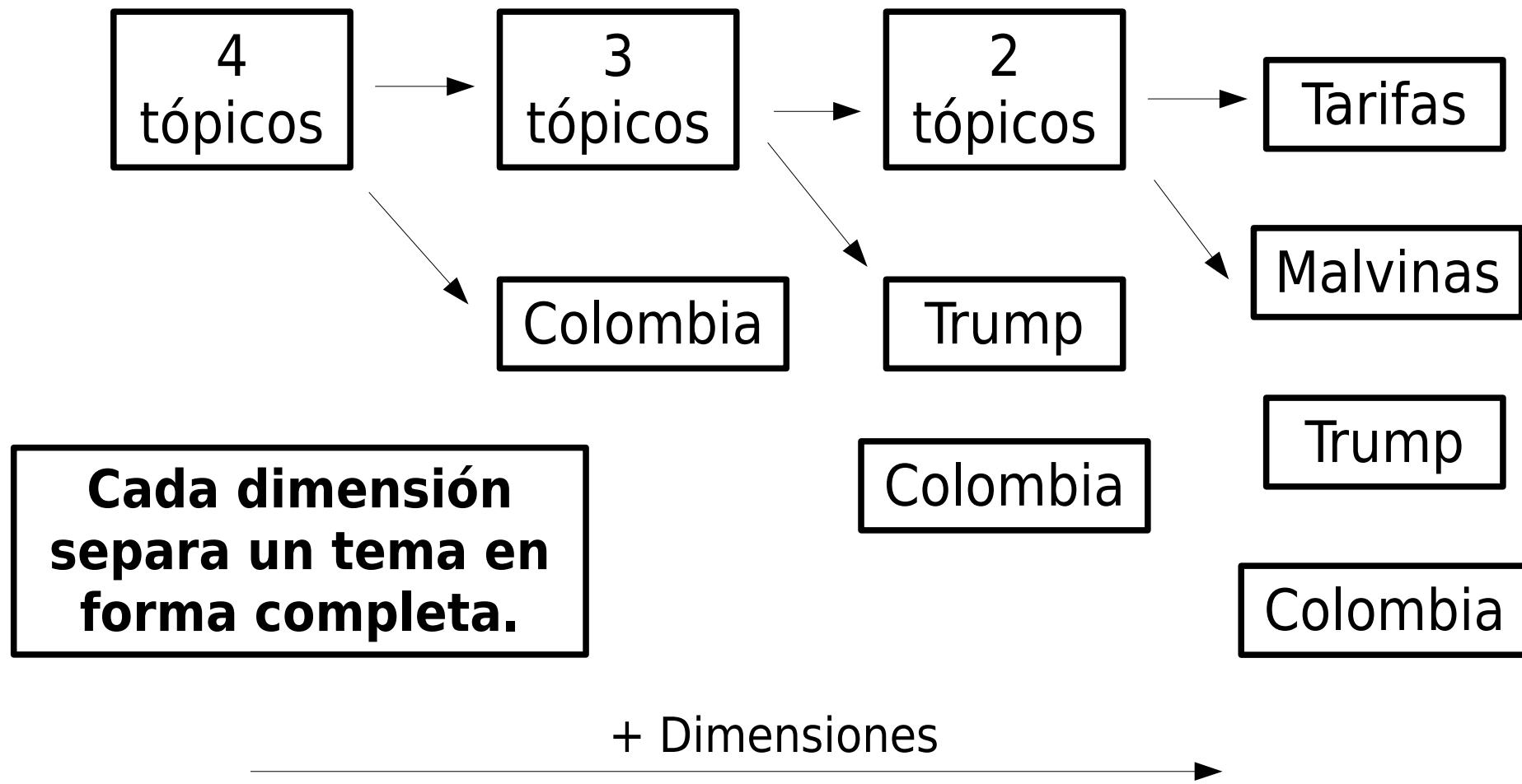
- **Dimensión 3:** *tarifas, corte, la corte, gas, aumentos, los aumentos, aumento, luz, fallo, las tarifas.*
- **Dimensión 4:** *malvinas, islas, las islas, argentina, malcorra, la argentina, vuelos, macri, soberanía, acuerdo.*

# Conclusiones

- **NMF** es una técnica muy poderosa en la **detección de tópicos**.
- Construir una red compleja **permite aplicar distintas medidas de evaluación** de las particiones.

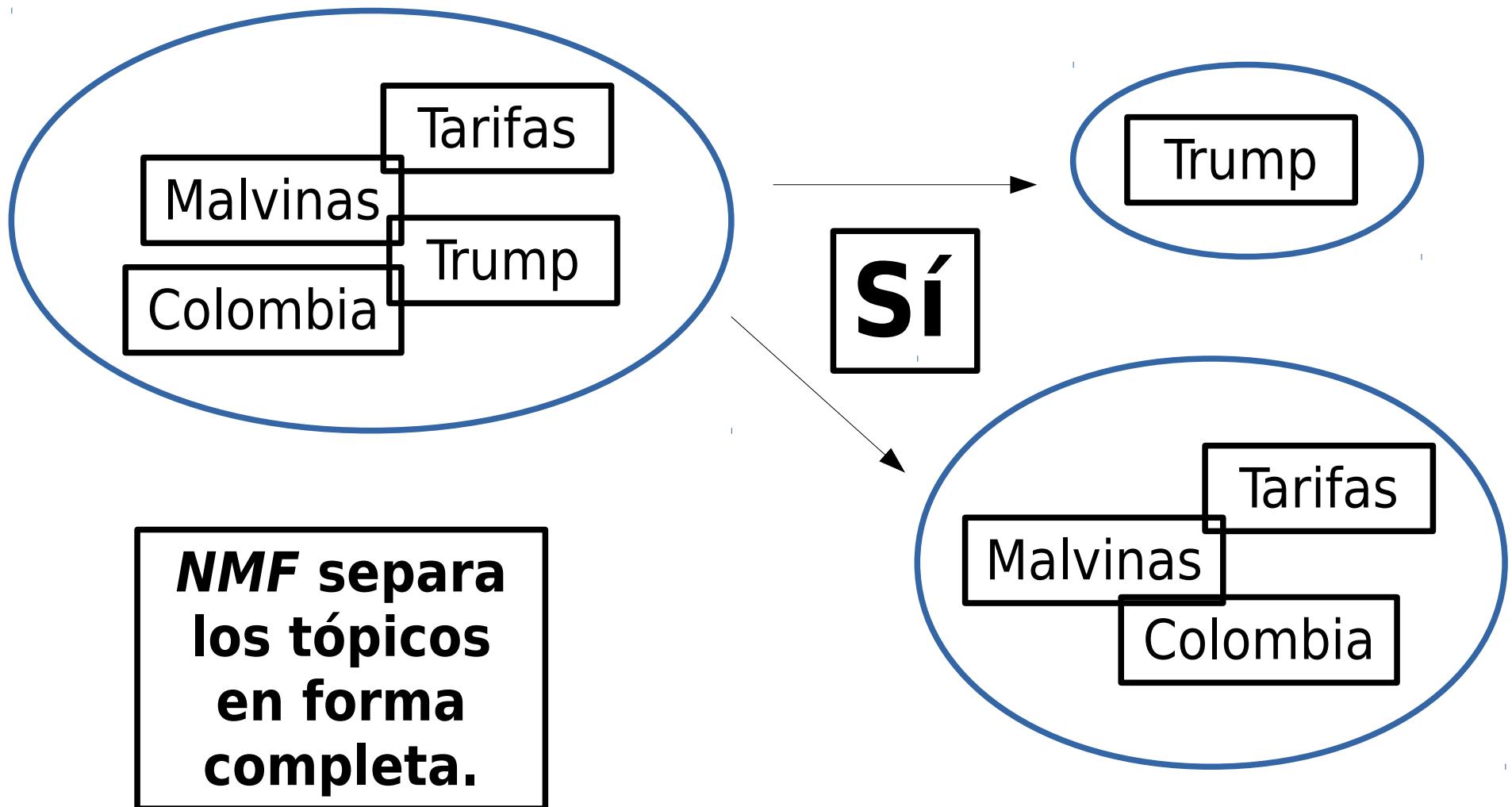
# Discusión de resultados parciales

Evolución de tópicos al aumentar las dimensiones:



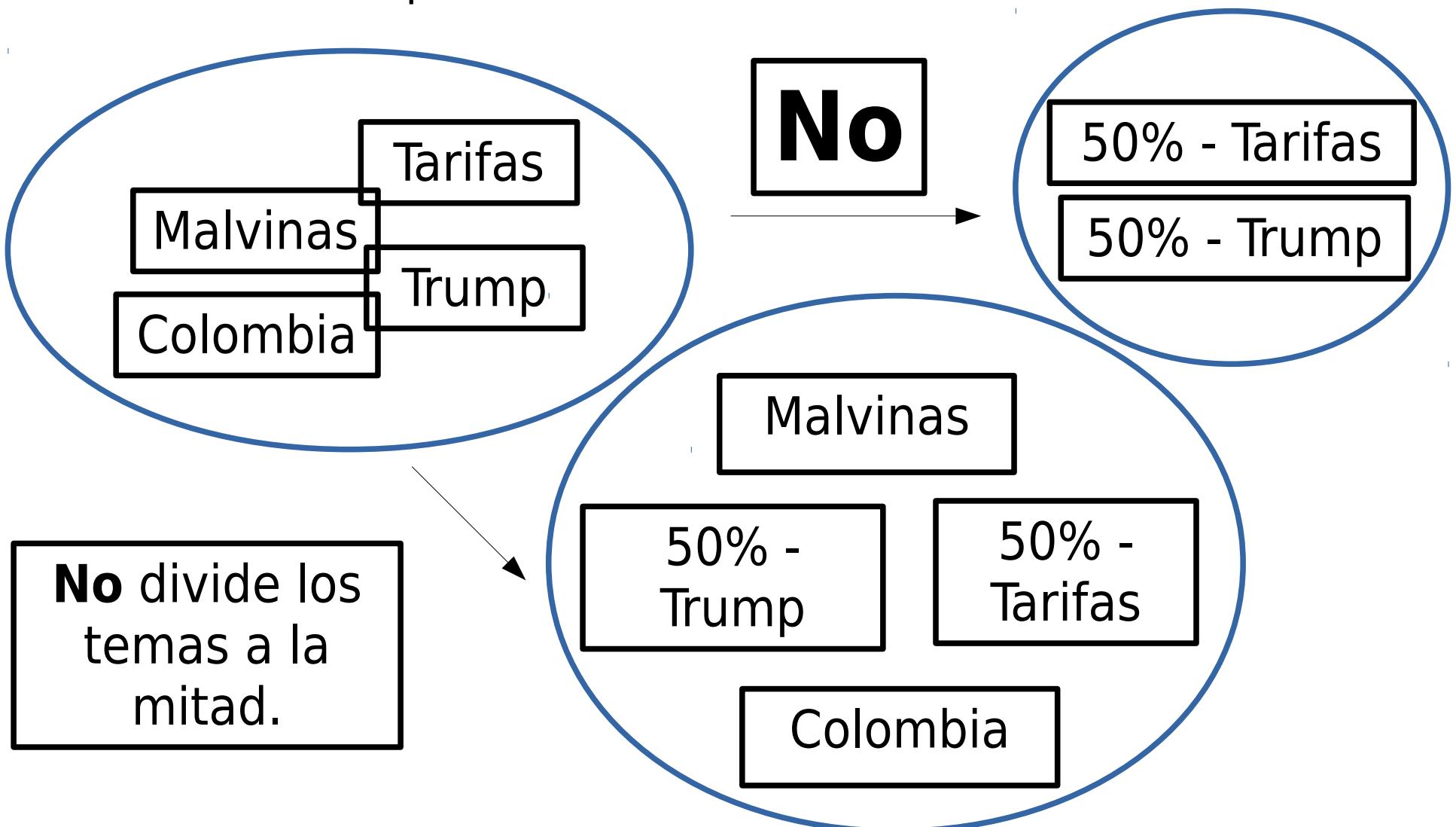
# Discusión de resultados parciales

Evolución de tópicos al aumentar las dimensiones:

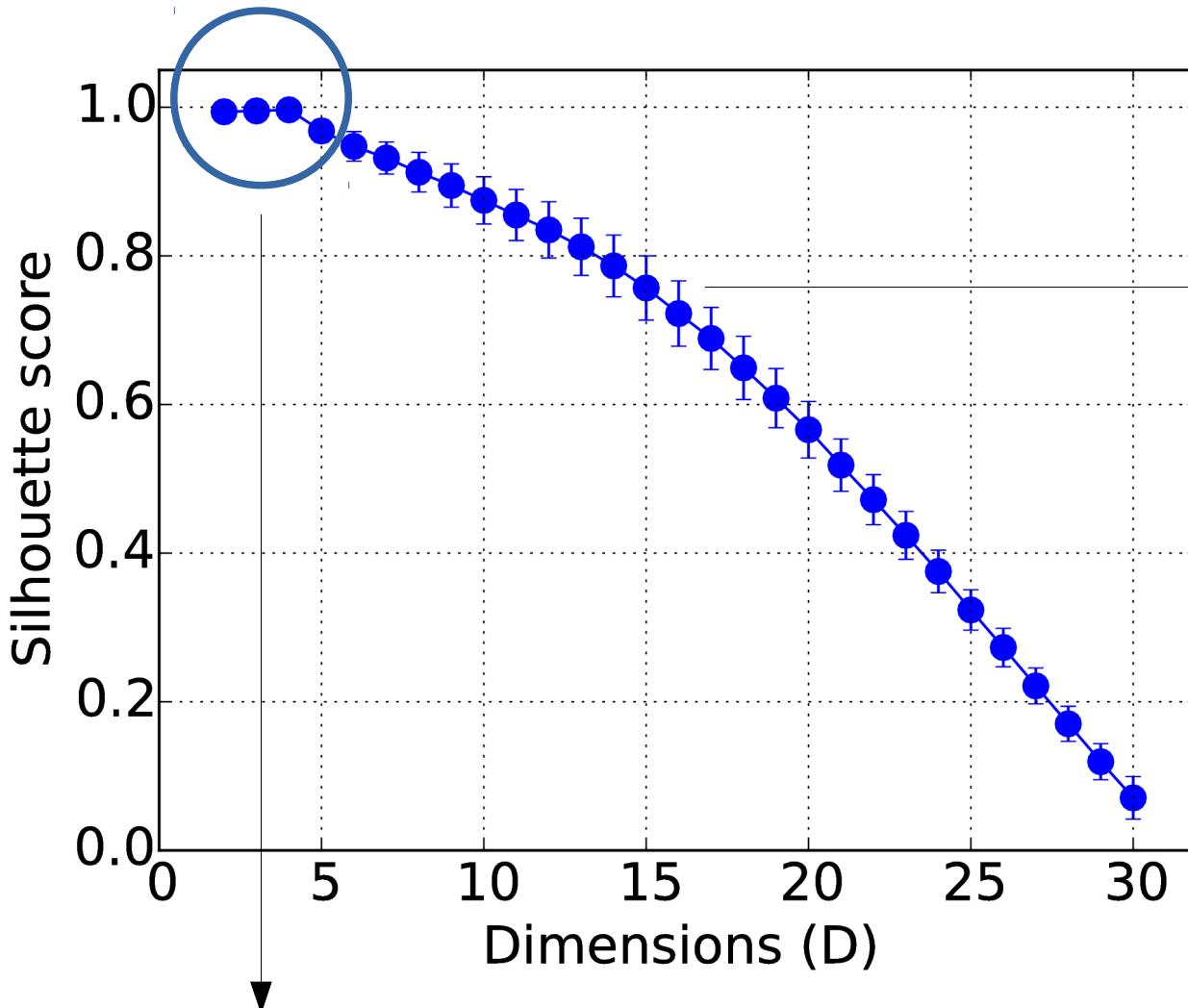


# Discusión de resultados parciales

Evolución de tópicos al aumentar las dimensiones:



# Discusión de resultados parciales

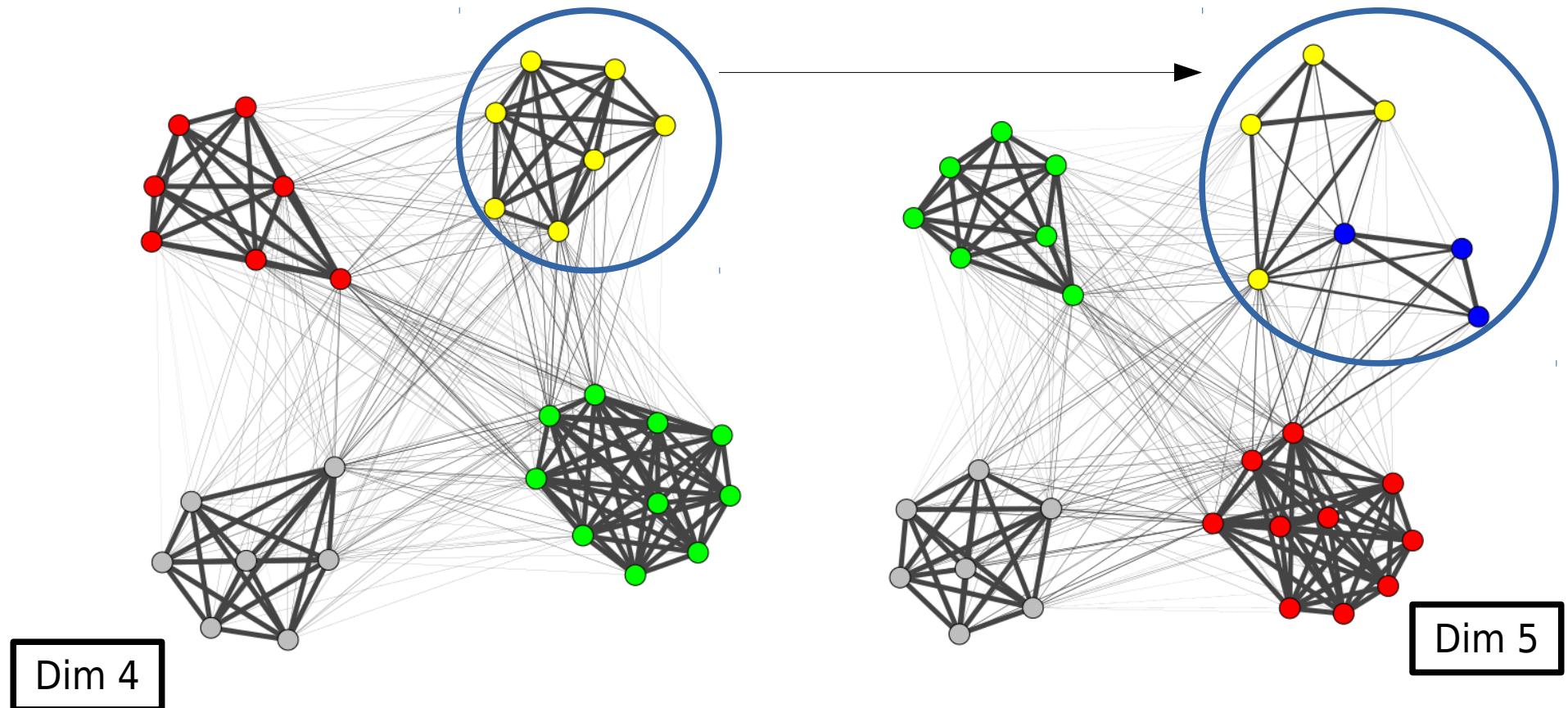


Las barras de error puede deberse a que diferentes condiciones iniciales proponen distintas particiones, lo cual puede hablar de los forzado de particionar los documentos.

El *silhouette* se mantiene constante ya que la resolución del *NMF* en esta zona no es suficiente para discriminar bien los tópicos.  
**Los macro-tópicos están constituidos por tópicos completos, sin fraccionar.**

# Discusión de resultados parciales

Evolución de tópicos al aumentar las dimensiones:



Cómo sigue dividiendo los tópicos?

# Interpretación NMF

Componentes con mayor peso, ordenados de mayor a menor:

Descomposición del tópico **Malvinas**:

- Dimensión 4a: *malvinas, las islas, islas, argentina, may, vuelos, malcorra, la argentina, macri, los vuelos.*

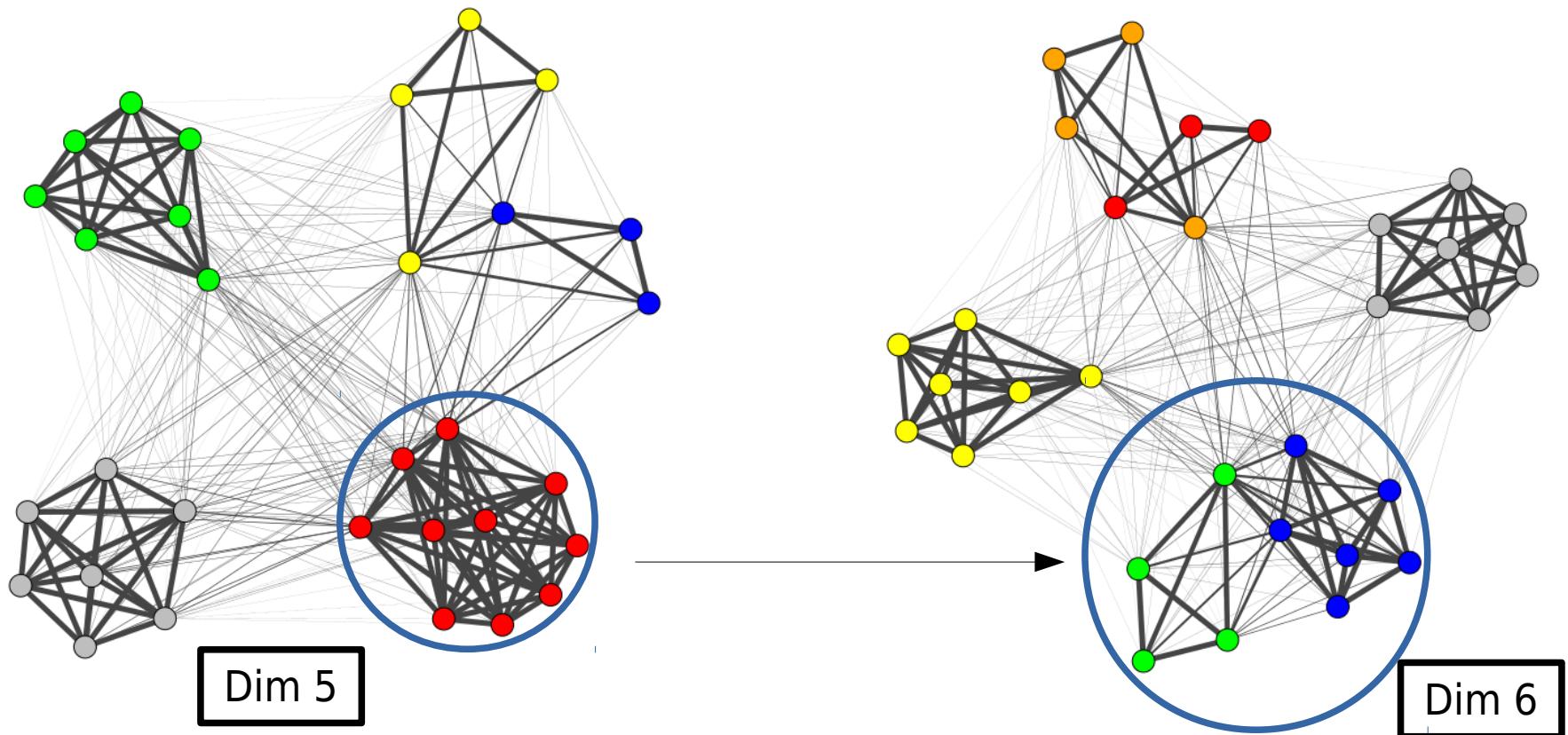
**Vuelos a Malvinas**

- Dimensión 4b: *malvinas, acuerdo, soberanía, malcorra, argentina, la soberanía, de soberanía, la argentina, cooperación, cuestión.*

**Discusión sobre soberanía**

# Discusión de resultados parciales

Evolución de tópicos al aumentar las dimensiones:



Cómo sigue dividiendo los tópicos?

# Interpretación NMF

Componentes con mayor peso, ordenados de mayor a menor:

Descomposición del tópico **Tarifas**:

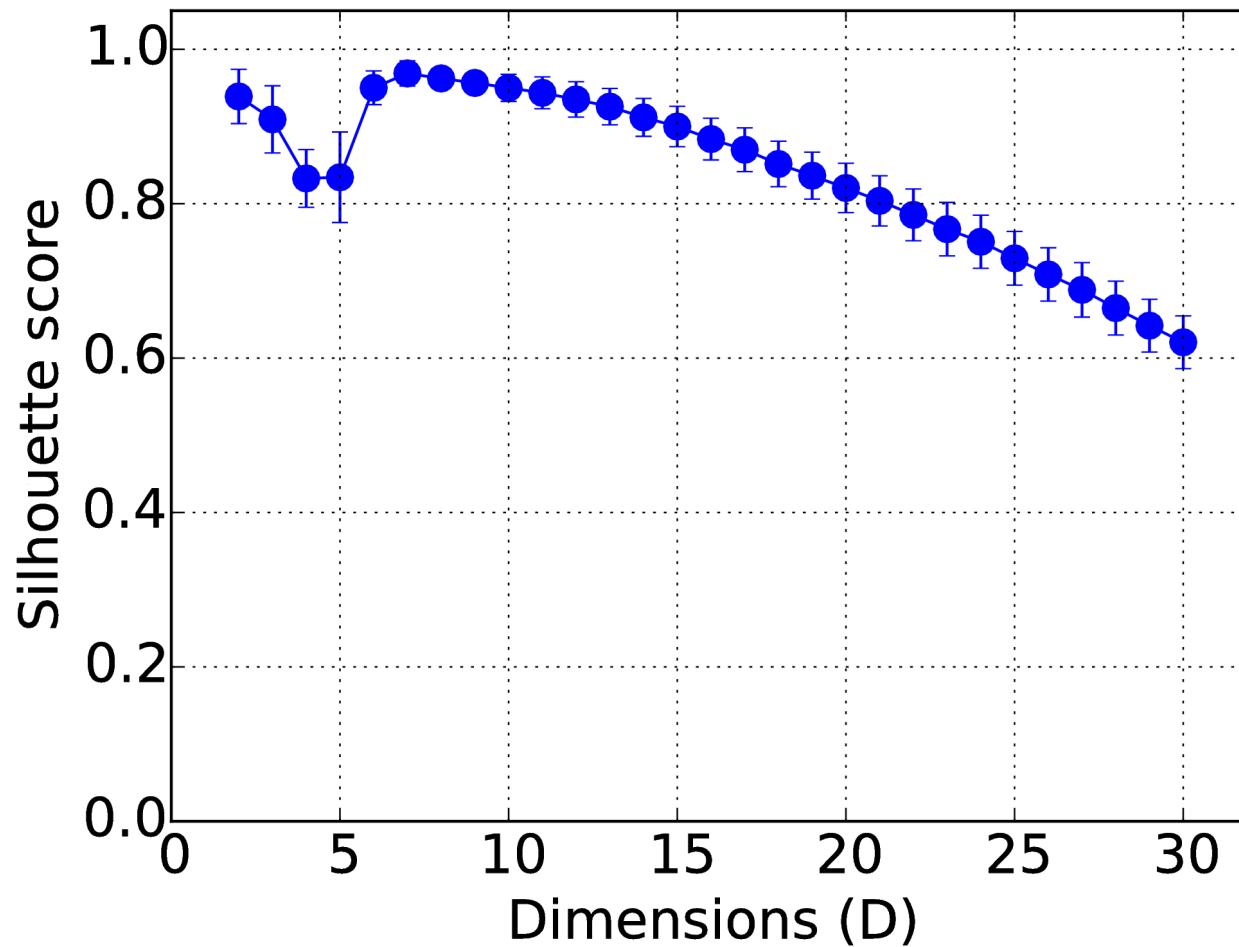
- Dimensión 3a: *la corte, corte, tribunal, plata, la plata, cámara, recurso, de la plata, la cámara, aumento.*

**Fallo de la Corte**

- Dimensión 3b: *tarifas, aumentos, los aumentos, audiencias, luz, gas, energía, las tarifas, precios.*

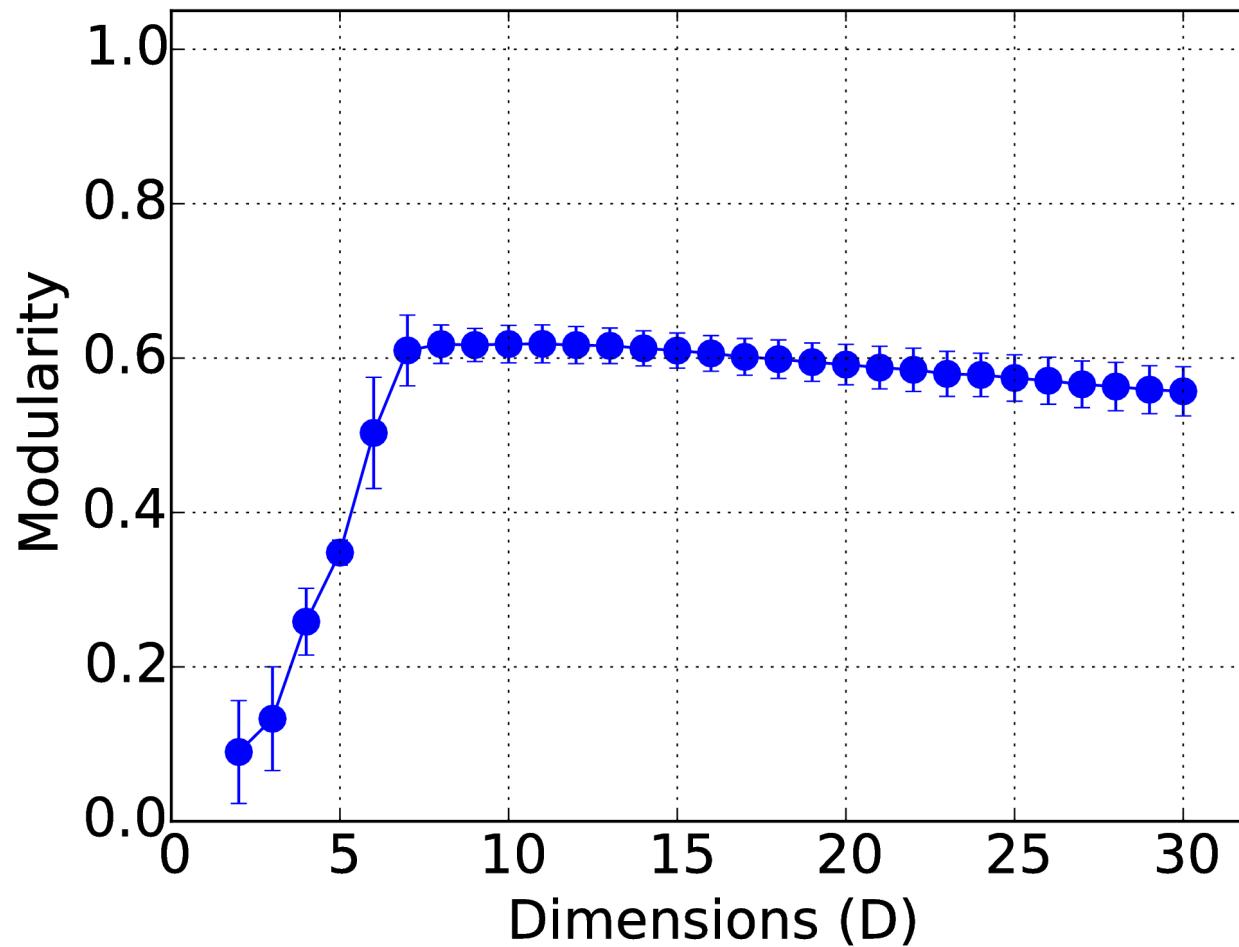
**Cortes, audiencias,  
aumentos.**

# Prueba sobre un corpus más grande



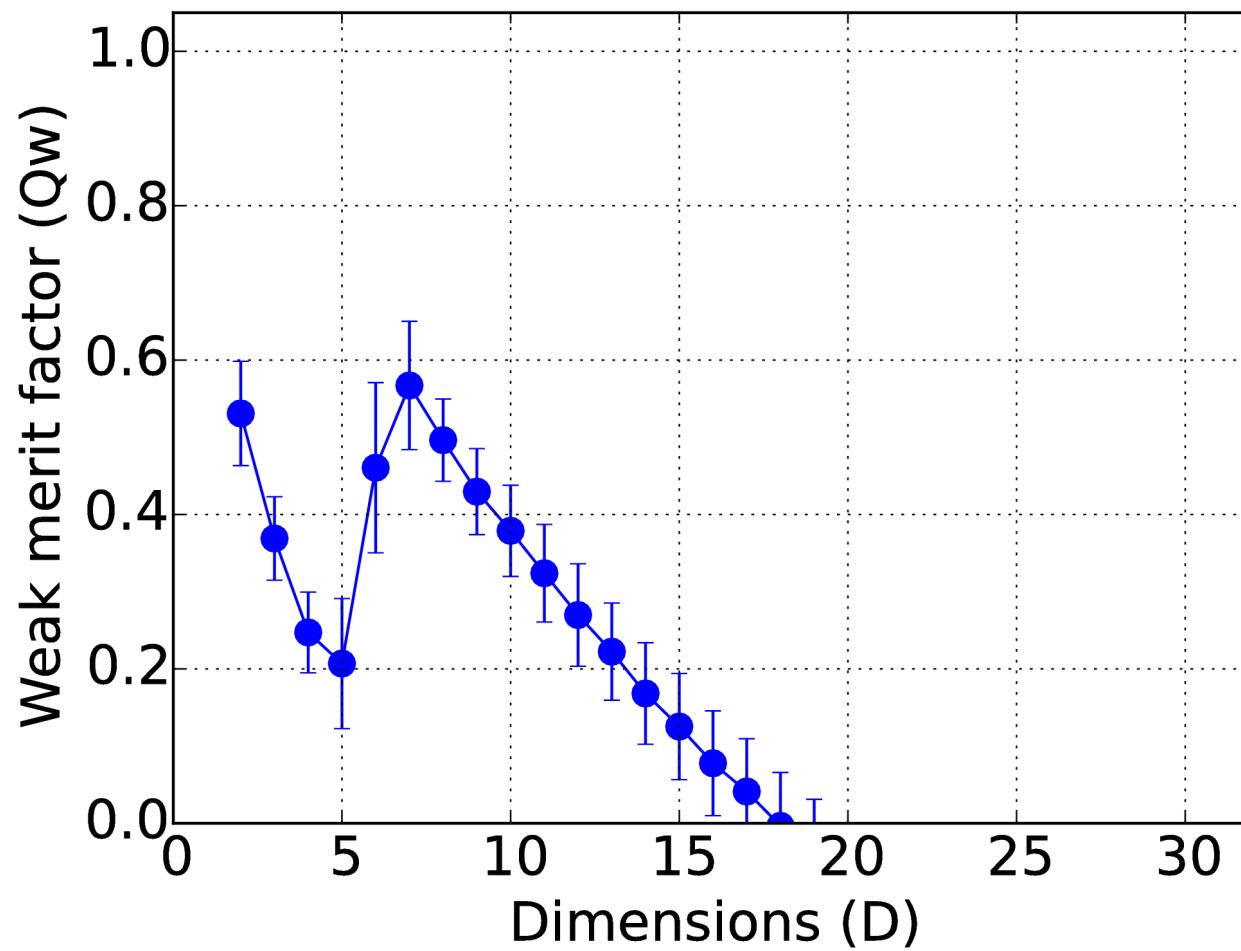
Los observables indican aproximadamente 7 tópicos

# Prueba sobre un corpus más grande



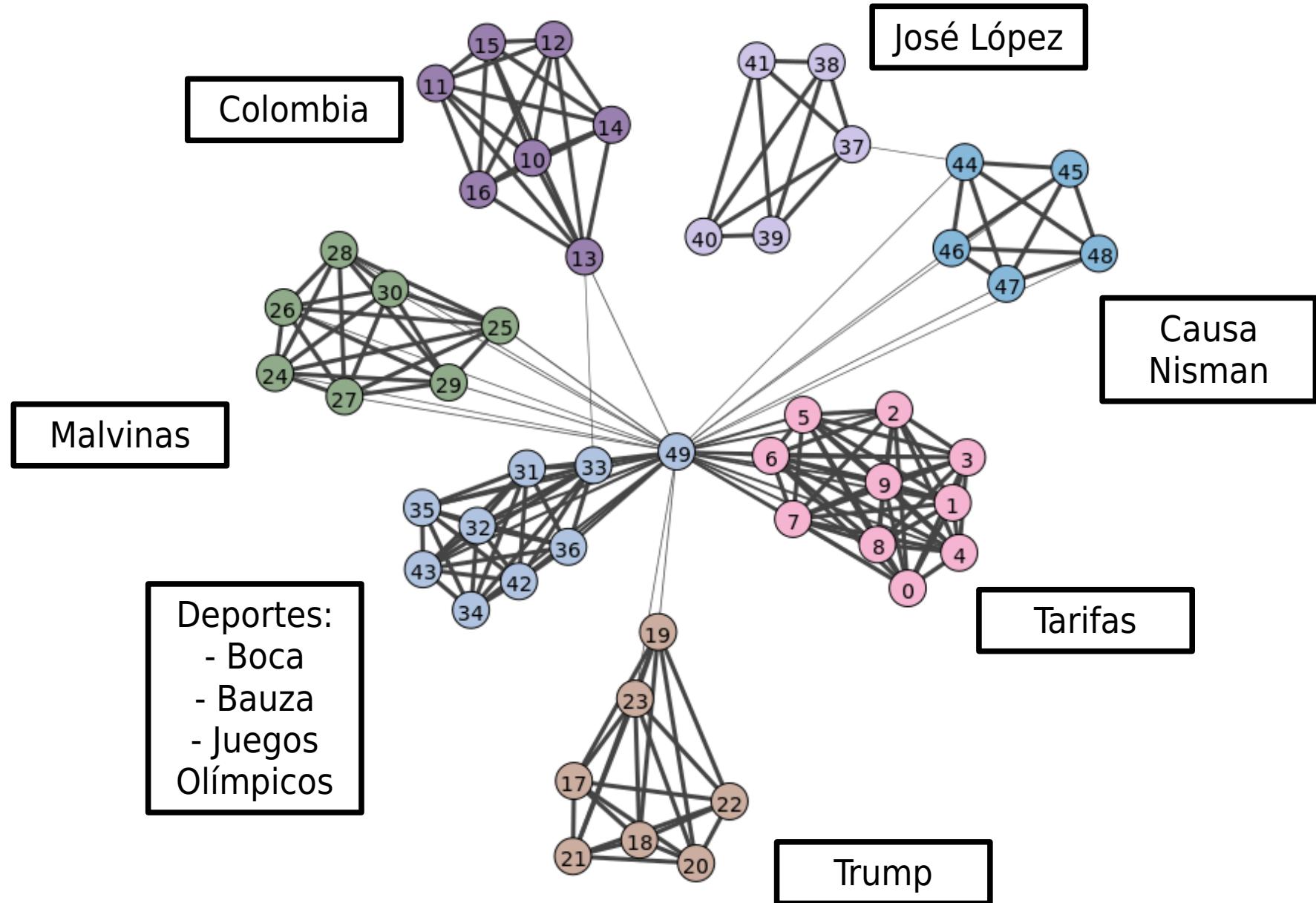
Los observables indican aproximadamente 7 tópicos

# Prueba sobre un corpus más grande

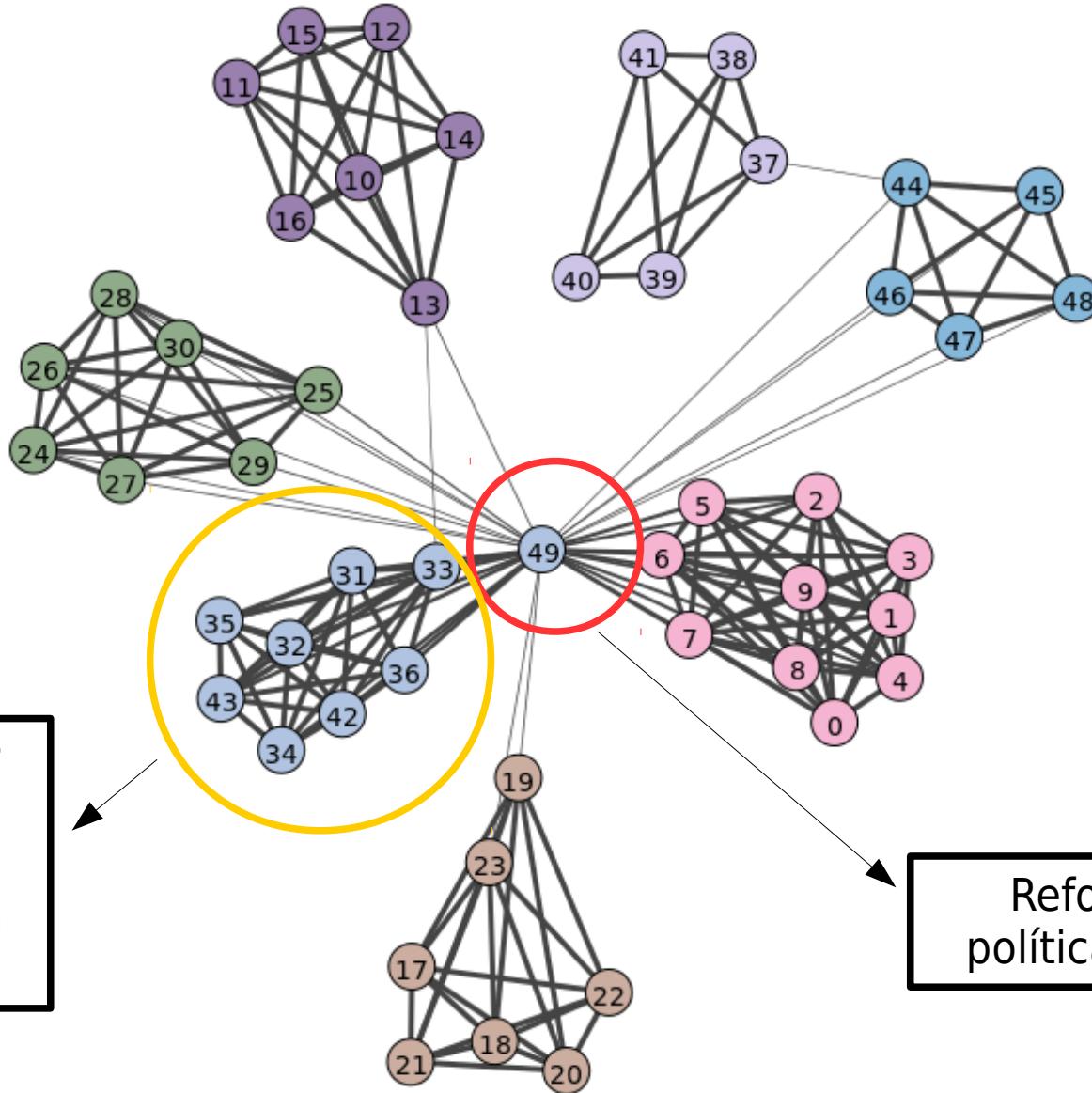


Los observables indican aproximadamente 7 tópicos

# Prueba sobre un corpus más grande



# Prueba sobre un corpus más grande



# Prueba sobre un corpus más grande

Cómo detectar tópicos cuando tengo solo una nota:

- El parámetro  $\min\_df$  del algoritmo tiene que ser 1, es decir debe permitir aquellos términos que aparezcan en un solo documento, para poder definir ese tema.
- Podemos definir una medida de calidad de pertenencia de un nodo a una comuna:
  - viendo el vector de etiquetas asociado a ese nodo (si el peso está distribuido entre varias etiquetas o no).
  - viendo medida de centralidad del nodo en la red.

# Trabajo actual

- Probando el caso de detección de tópicos aislados.
- Estudiando la sección política en un período temporal de 1 mes.