

Incluyen una amplia moratoria impositiva en

LA NACION
Jueves 2 de junio de 2016 | lanacion.com

a en la Costanera Norte



**EUROPA, UNIDA
POR UN TUNEL**

Atreviase los
Alpes y es el más
largo del mundo

EL MUNDO DE HOY

**DESCUBRIR
LA FILOSOFÍA**

**INCIDENTE EN EL CÚPULO
EN EL SUPLENIMIENTO
ESPECIALIZADO**

LEVANTOS

HOY ENTREGA 49

ACIÓN

Más T + más
Noticias y opiniones
para el día a día

El mundo de hoy

de 2016 | lunacion.com

la

Eximen de la suba

Eximen de la suba vastos en el Sur

[illegible]

El puente de la ciudad de San Francisco, California, es el más largo del mundo. Fue construido en 1937 y mide 2.123 metros de largo. El puente de la ciudad de San Francisco, California, es el más largo del mundo. Fue construido en 1937 y mide 2.123 metros de largo.

[illegible][illegible]

Una playa en la Costanera Norte

Se podría ir a la universidad para ser **escritor**

Página 20 | [Inicio](#) | [Ayuda](#) | [Contacto](#) | [Sobre nosotros](#) | [Política de privacidad](#) | [Términos de uso](#)

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).

Métodos de detección:

- K-means + PCA.
- Detección de comunidades en redes complejas.

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 conjunto de notas conocido:

- Descripción vectorial de las notas (tf-idf).

Métodos de detección:

- K-means + PCA.
- Detección de comunidades en redes complejas.

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 conjunto de notas conocido:

- **Descripción vectorial de las notas (tf-idf).**

Métodos de detección:

- **K-means + PCA.**
- **Detección de comunidades en redes complejas.**

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 conjunto de notas conocido:

- **Descripción vectorial de las notas (tf-idf).**

Métodos de detección:

- K-means + PCA.
- Detección de comunidades en redes complejas.

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 descrito 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.

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

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

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

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

$$v_3 = \begin{matrix} & \text{la} & \text{casa} & \text{de} & \text{seba} & \text{julia} & \text{cristian} \\ \left[\begin{matrix} 1, & 1, & 2, & 1.69, & 0, & 0 \end{matrix} \right] \end{matrix}$$

“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

N documentos

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

Dim = N x F

Term frequency – Inverse document frequency (Tf-idf)

Parámetros que se pueden regular y valores escogidos:

- **Max_df: descartar términos por muy ocurrentes:** 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.
- **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.**

Term frequency – Inverse document frequency (Tf-idf)

Parámetros que se pueden regular y valores escogidos:

- 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

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 conjunto de notas conocido:

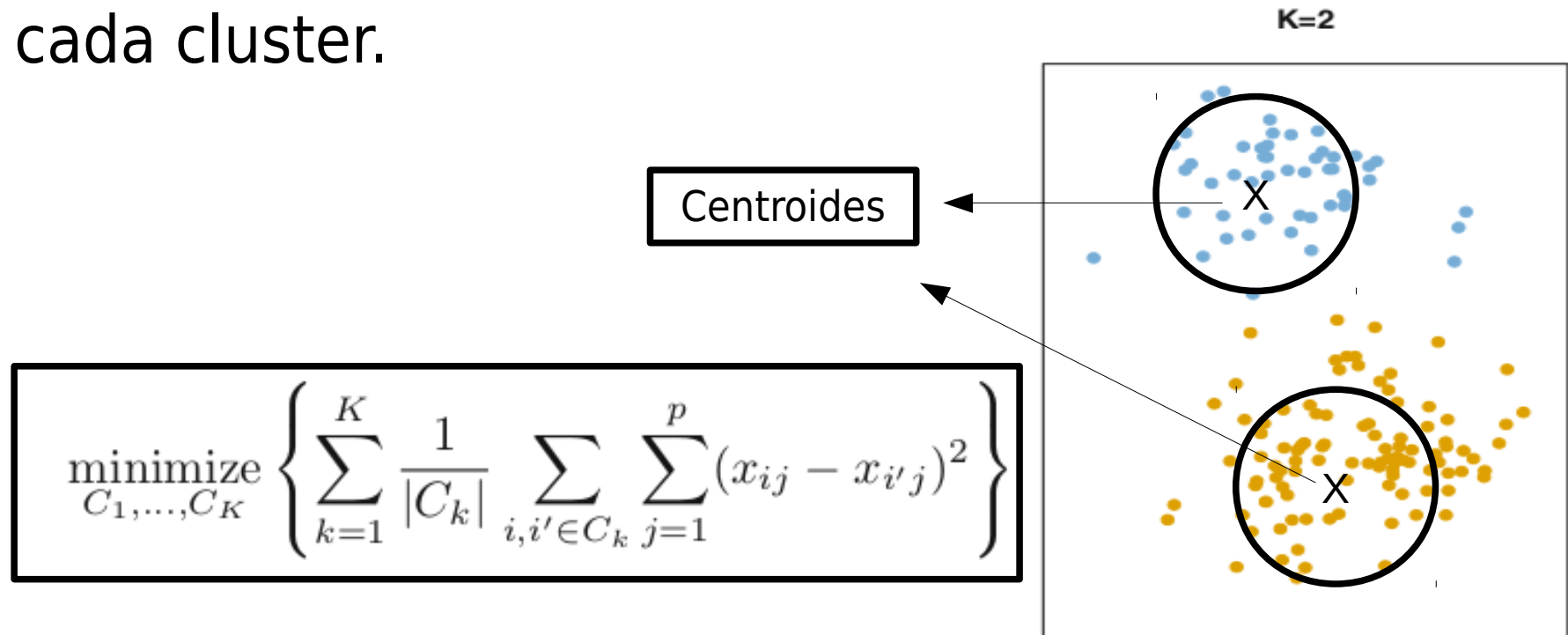
- Descripción vectorial de las notas (tf-idf).

Métodos de detección:

- **K-means + PCA.**
- Detección de comunidades en redes complejas.

K-means

- Busca la mejor partición del espacio multidimensional basado en **minimizar la suma de las varianzas** de cada cluster.



- La **cantidad de particiones K** es un **parámetro del algoritmo** → se debe buscar un criterio para elegir la mejor.

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 \rightarrow 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° puntos - 1).

K-means sobre la base de datos de prueba

Particiones propuestas ordenados según el silhouette:

K (#clusters)	Silhouette
4	0.13
5	0.12
10	0.12
9	0.11
8	0.11



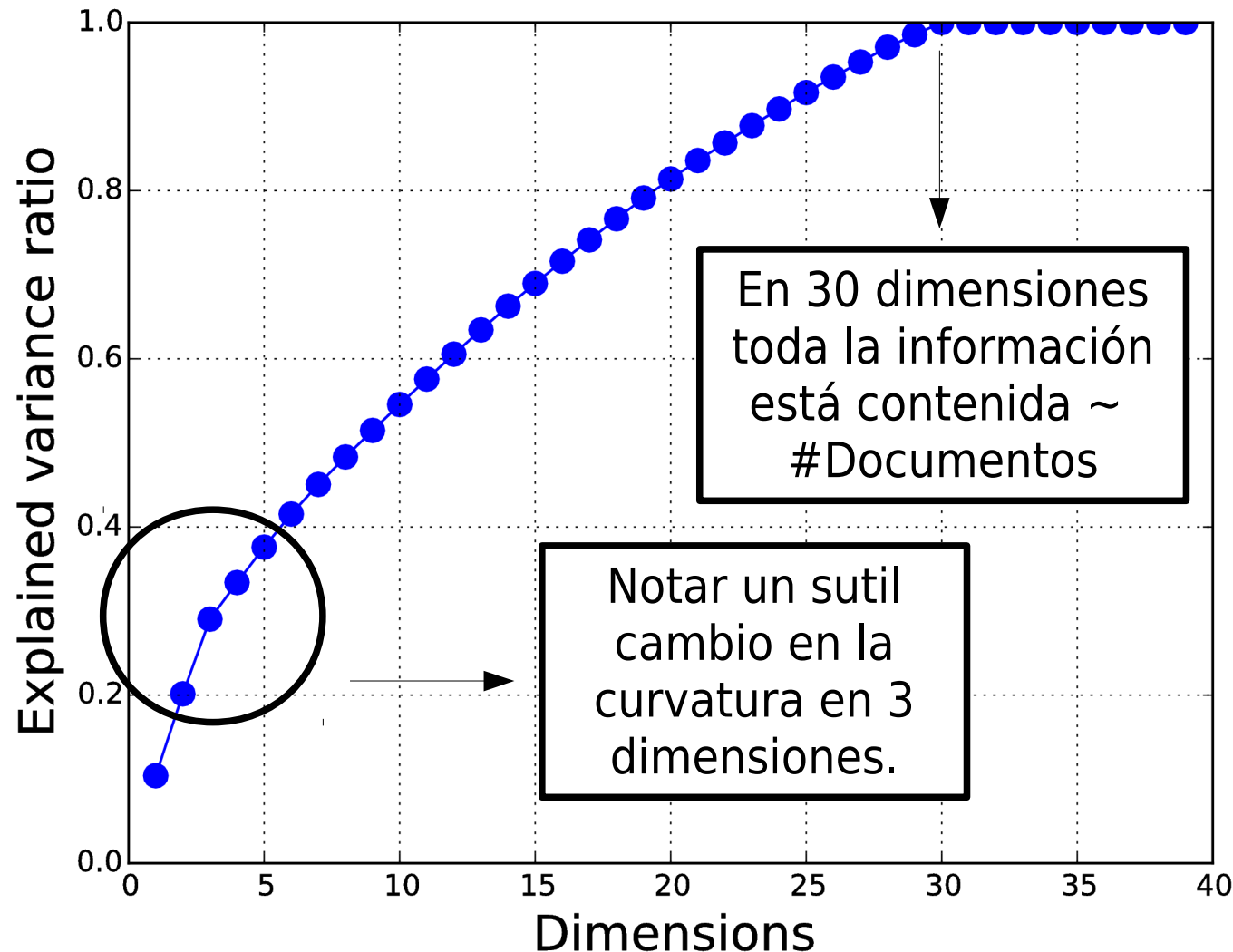
El algoritmo encuentra que la mejor partición coincide con los tópicos esperados.

Sin embargo el coeficiente de silhouette parece ser muy bajo.

K-means + reducción de la dimensionalidad (PCA)

- **PCA** (*principal components analysis*) busca **direcciones** en el espacio multidimensional que **mejor expliquen la variabilidad de los datos**.
- Las nuevas direcciones son una **combinación lineal de los términos** que describen el espacio original.

Información contenida



Fracción de varianza explicada en función de la cantidad de dimensiones utilizadas para explicar los datos. Con 3 dimensiones se explica ~ 33% de la información.

K-means + reducción de la dimensionalidad (PCA)

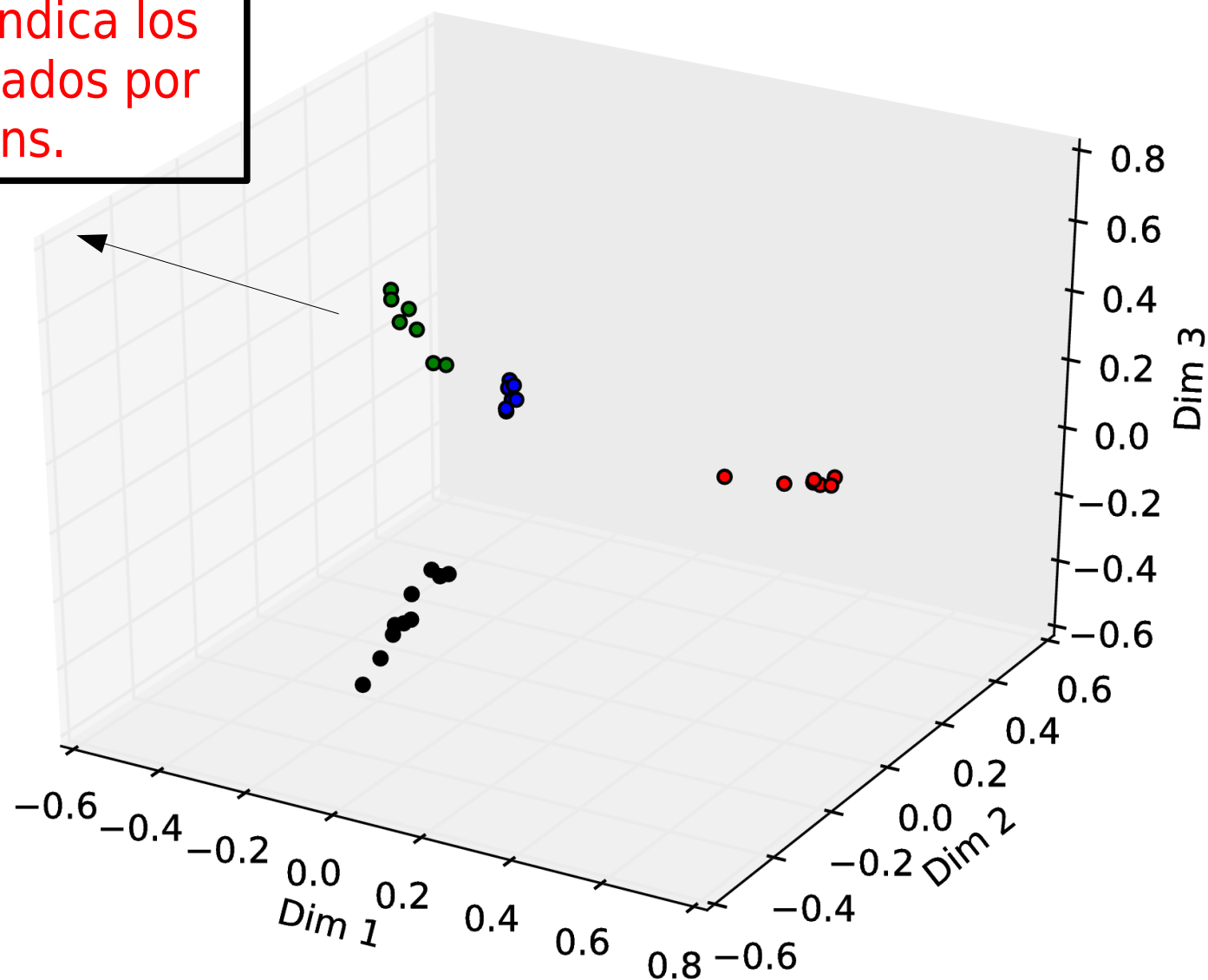
- Barrido en cantidad de dimensiones y número de clusters propuestos:

#Dimensiones	K (#clusters)	Silhouette
3	4	0.86
2	3	0.84
1	2	0.83
1	3	0.78
1	4	0.77

**Aumenta considerablemente el
coeficiente de silhouette.**



Cada color indica los
clusters hallados por
K-means.



Proyección en el espacio tridimensional

Conclusiones K-means

- *El enfoque correcto parece ser K-means*
+ *PCA*: la **reducción de la dimensionalidad** aporta cierto grado de **abstracción** necesaria para detectar tópicos.

Pérdida de información →
Abstracción

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 conjunto de notas conocido:

- Descripción vectorial de las notas (tf-idf).

Métodos de detección:

- K-means + PCA.
- **Detección de comunidades en redes complejas.**

Enfoque redes complejas

- Se interpreta cada documento como un nodo en una red compleja.
- Podemos construir una red pesada, definiendo los pesos entre los nodos *i* y *j* como:

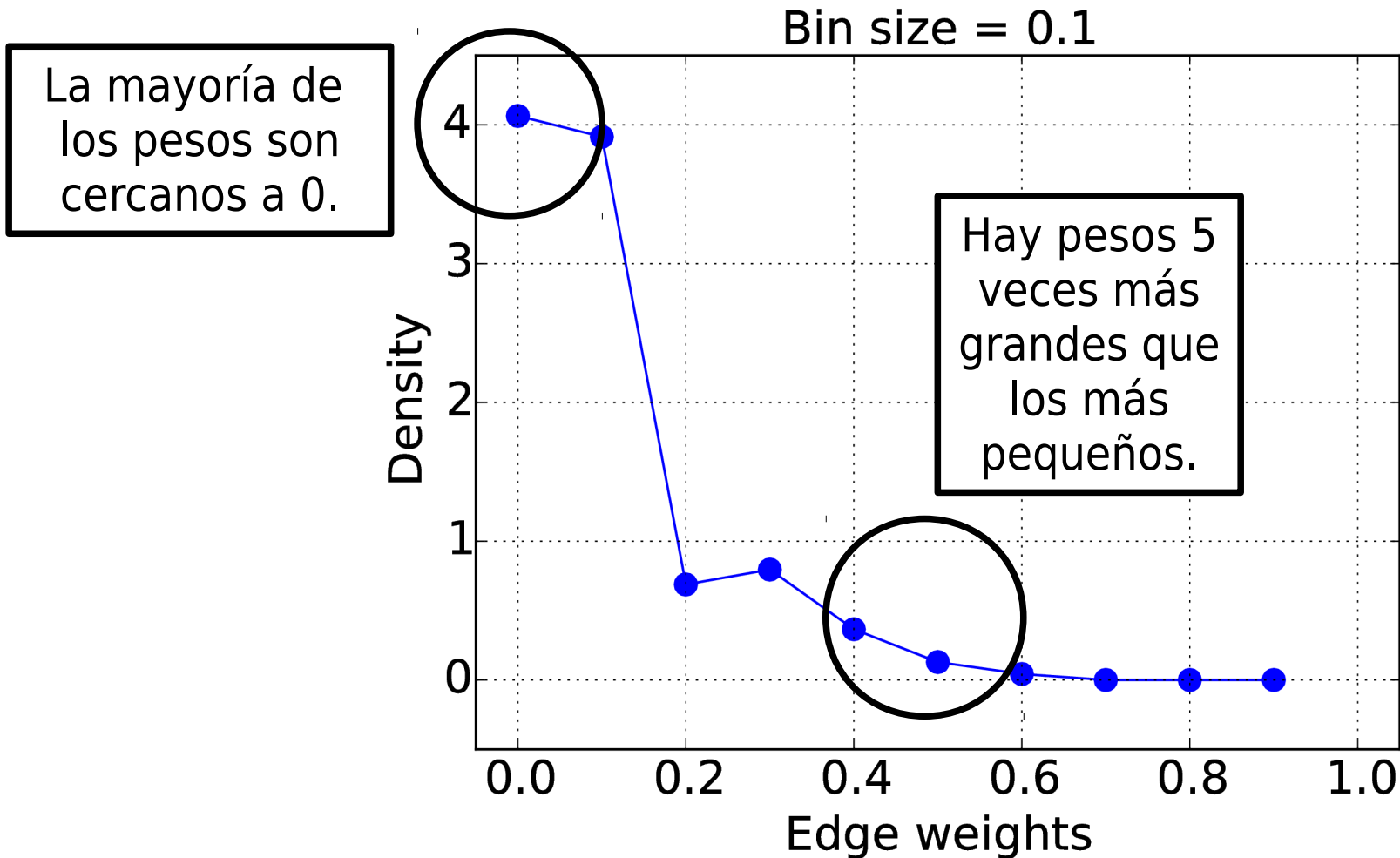
$$w_{ij} = \vec{v}_i \cdot \vec{v}_j = \cos(\theta)$$

Vectores documentos con norma 1

Medida de similaridad

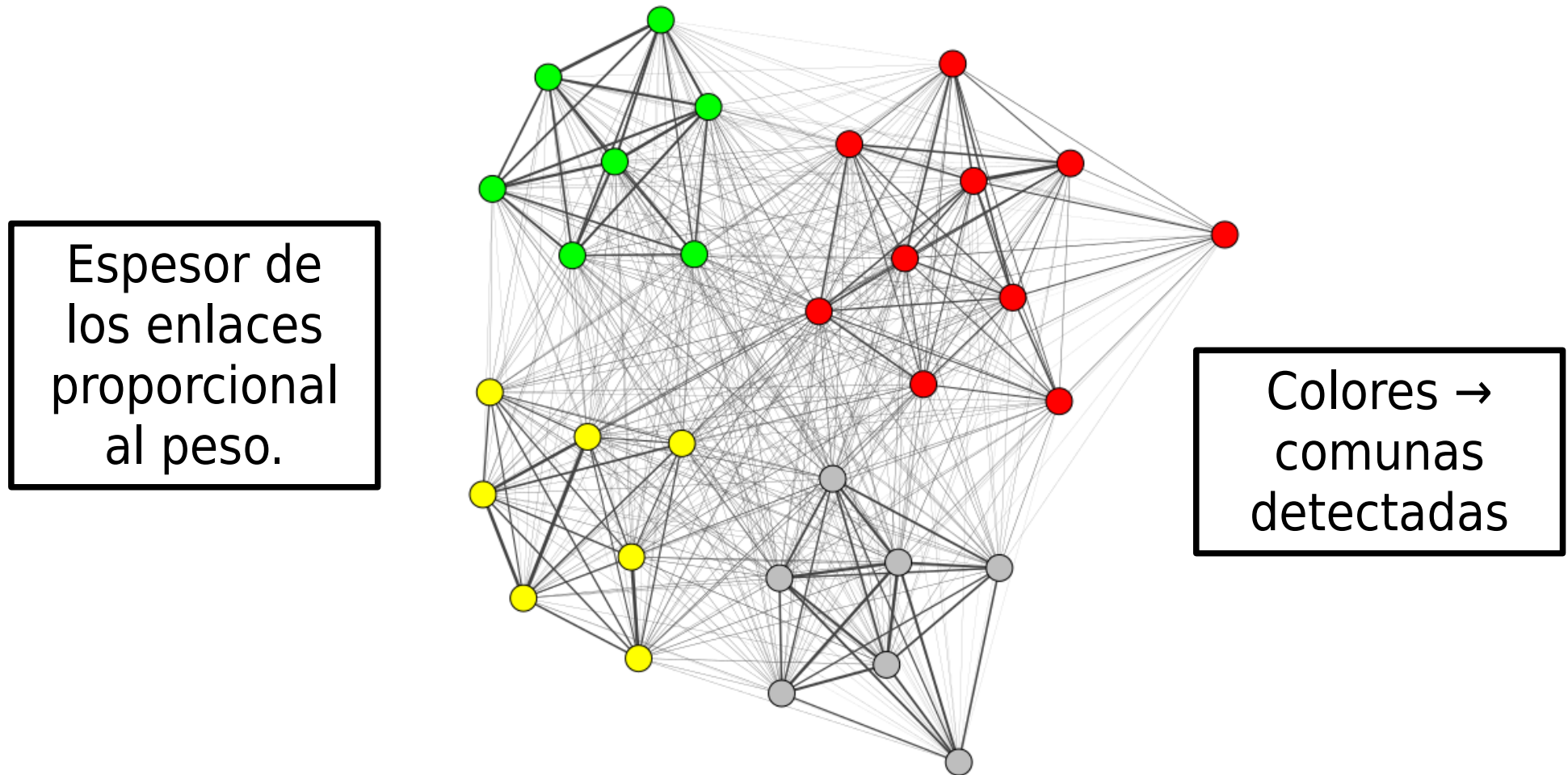
$0 < w_{ij} < 1$

Redes sobre la base de datos de prueba



Histograma de los pesos de los enlaces en la red

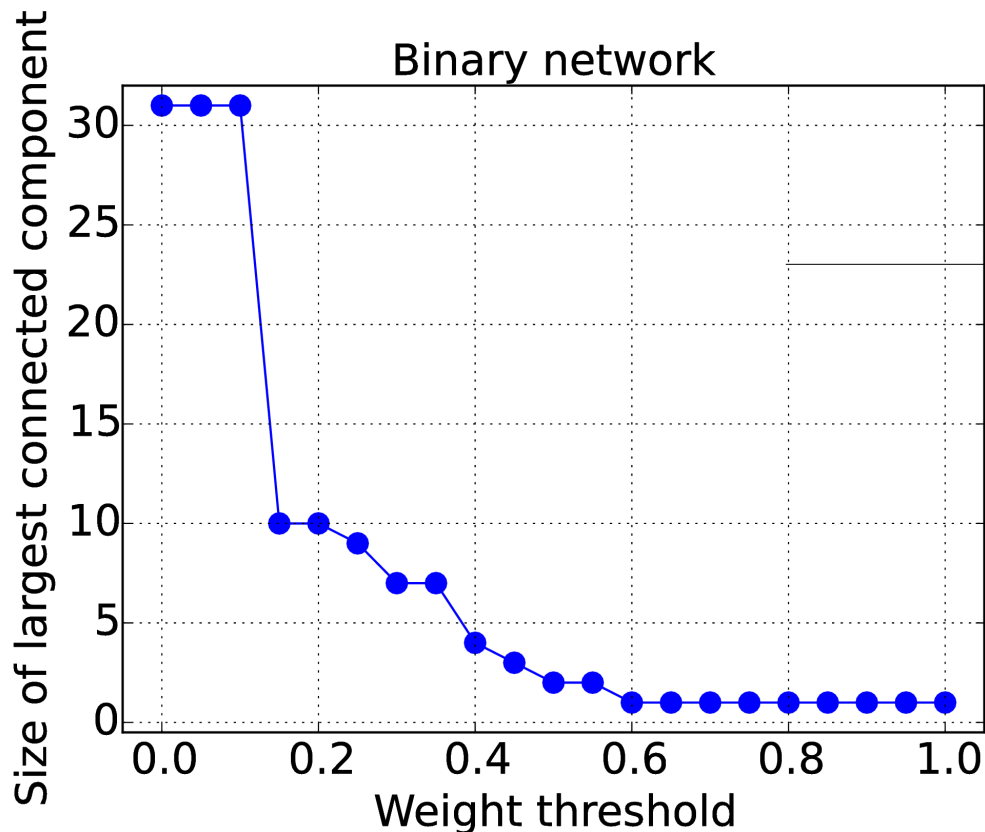
Layout pesado



- Es robusto ante el cambio de algoritmo: *infomap*, *fastgreedy*, y *label propagation*.
- Falla con *edgebetweenness*: detecta un único cluster.

Binarización de la red

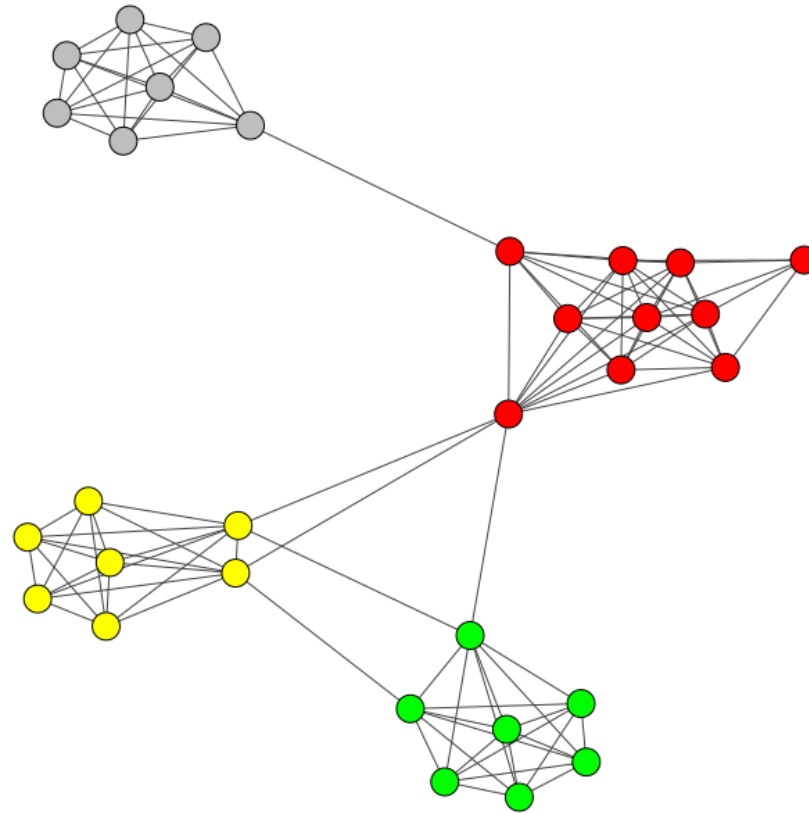
Otro enfoque consiste en binarizar la red: si $w_{ij} > \text{umbral} \rightarrow a_{ij} = 1$ (matriz de adyacencia).
Convertimos la red pesada en una red no pesada.



► **Componente conectado más grande en función del umbral.**

Para los documentos de prueba, no hay nodos aislados para un umbral < 0.10 .

Binarización de la red



Binarización + detección de comunas

Las comunas coinciden con la red pesada.
Al binarizar se favorece los enlaces con más peso.

Conclusiones Redes complejas

- *Redes pesadas parecen reproducir la estructura en tópicos de la red.*
- *La binarización provee una alternativa en caso que los pesos de los enlaces sean muy similares.*