El concepto de profundidad en Estadística

Carlos Cuevas Covarrubias

UNIVERSIDAD ANÁHUAC • VINCE IN BONO MALUM • MÉXICO NORTE

50 AÑOS Facultad de Ciencias Actuariales

Cuando decimos que algo es profundo ¿qué queremos decir exactamente?

¿Cuál es el punto de mayor profundidad en el planeta?

¿Cuál es el punto más superficial?

El concepto de profundidad suele referirse a la característica de una región en el espacio. Con frecuencia corresponde a un espacio tridimensional, pero

¿es posible extrapolar este concepto a espacios de dimensión diferente a tres?

¿La profundidad se refiere siempre a un espacio continuo o puede aplicarse a un espacio discreto?

Tenemos una muestra observada proveniente de la siguiente función de densidad:

$$\boldsymbol{f(x) = \frac{1}{2} e^{|x-\mu|}}$$

¿Cómo encontrar el estimador máximo verosímil para $\boldsymbol{\mu}$ **?**

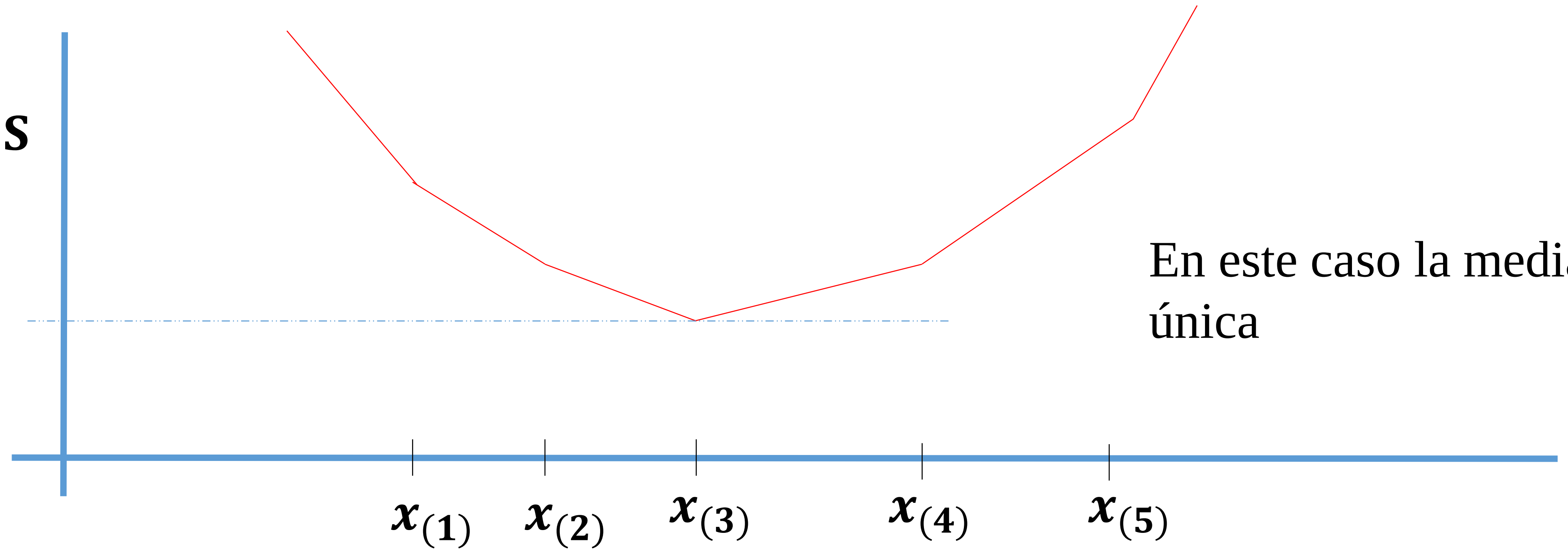
Sabemos que el estimador máximo verosímil es la mediana muestral; es decir:

$$\hat{\mu} = \begin{cases} x_{(k+1)} & si \quad n = 2k+1 \\ \dfrac{x_{(k)} + x_{(k+1)}}{2} & si \quad n = 2k \end{cases}$$

La función de verosimilitud es la siguiente

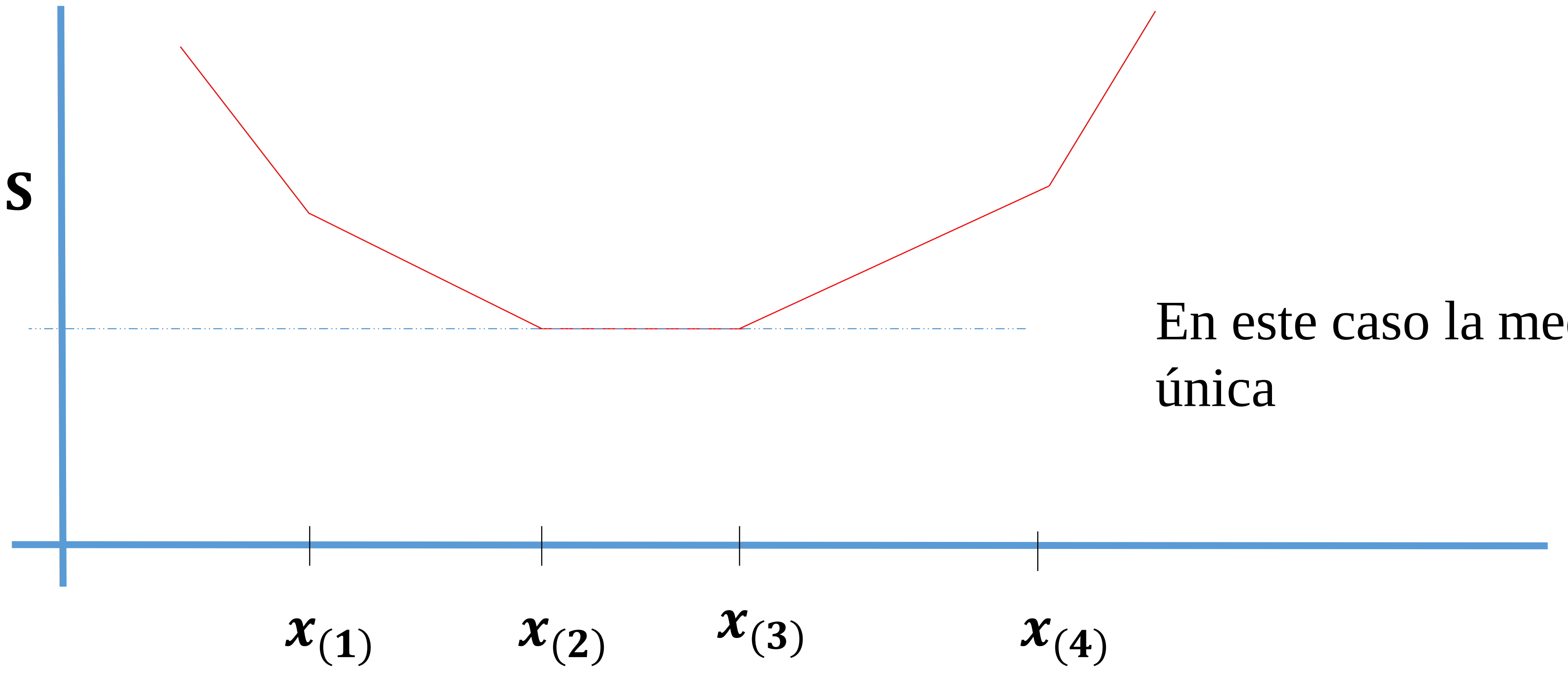
$$L(\mu) = e^{-\sum|x_i-\mu|}$$

La función de verosimilitud será máxima cuando la suma $S = \sum|x_i - \mu|$ sea mínima. Supongamos una muestra hipotética de tamaño 5; la gráfica de S sería de la forma siguiente:



En este caso la mediana es única

Ahora, supongamos que la muestra es de tamaño 4



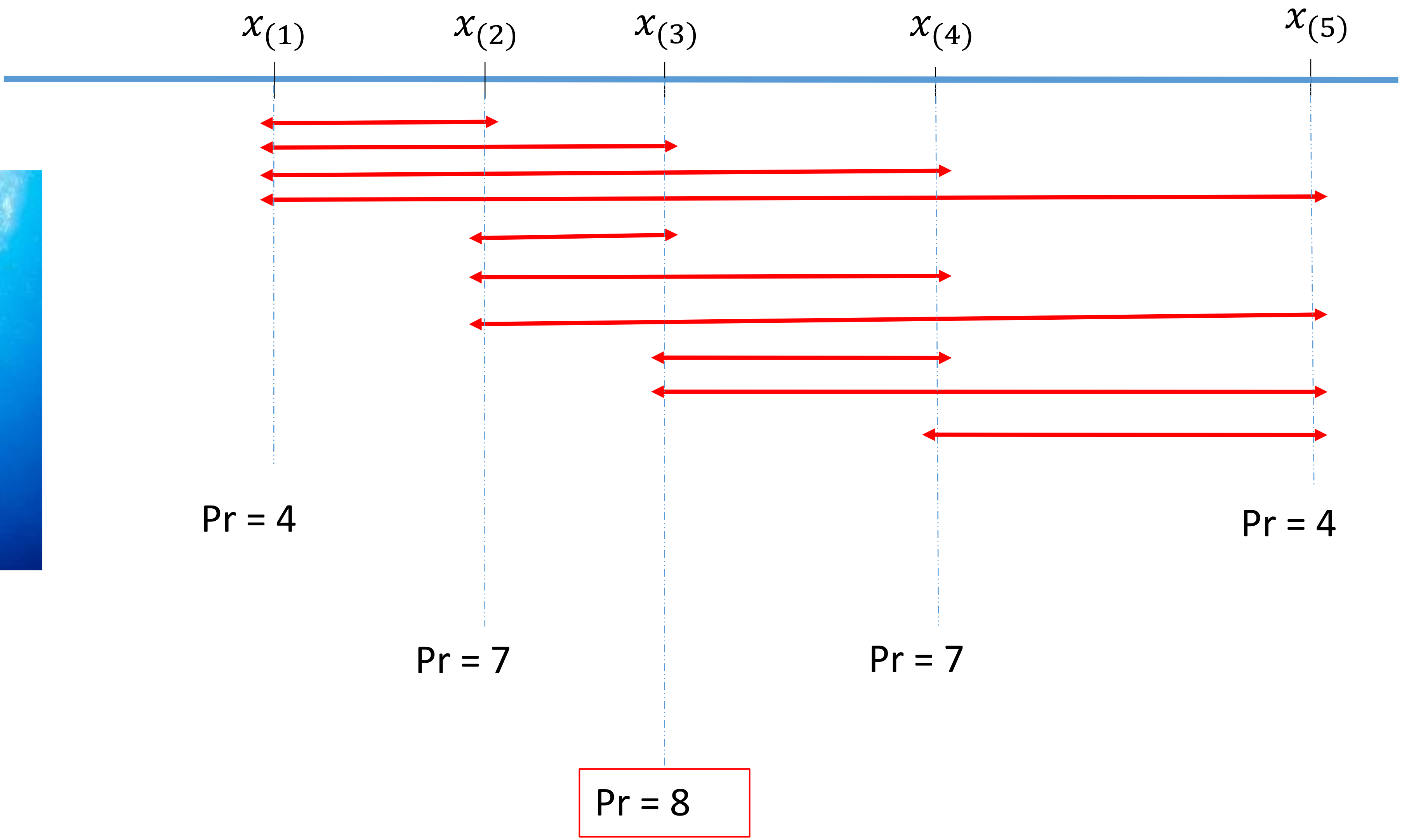
En este caso la median no es única

Entonces, la mediana no es única. El estimador máximo verosímil tampoco

La mediana muestral suele definirse como aquel elemento en la muestra ordenada que deja al 50% de las observaciones por abajo y al 50% restante pos arriba.

$$\boldsymbol{x_{(1)} \leq x_{(2)} \leq \ldots\, x_{(k-1)} \leq x_{(k)} \leq {\color{red}x_{(k+1)}} \leq x_{(k+2)} \ldots x_{(2k)} \leq x_{(2k+1)}}$$

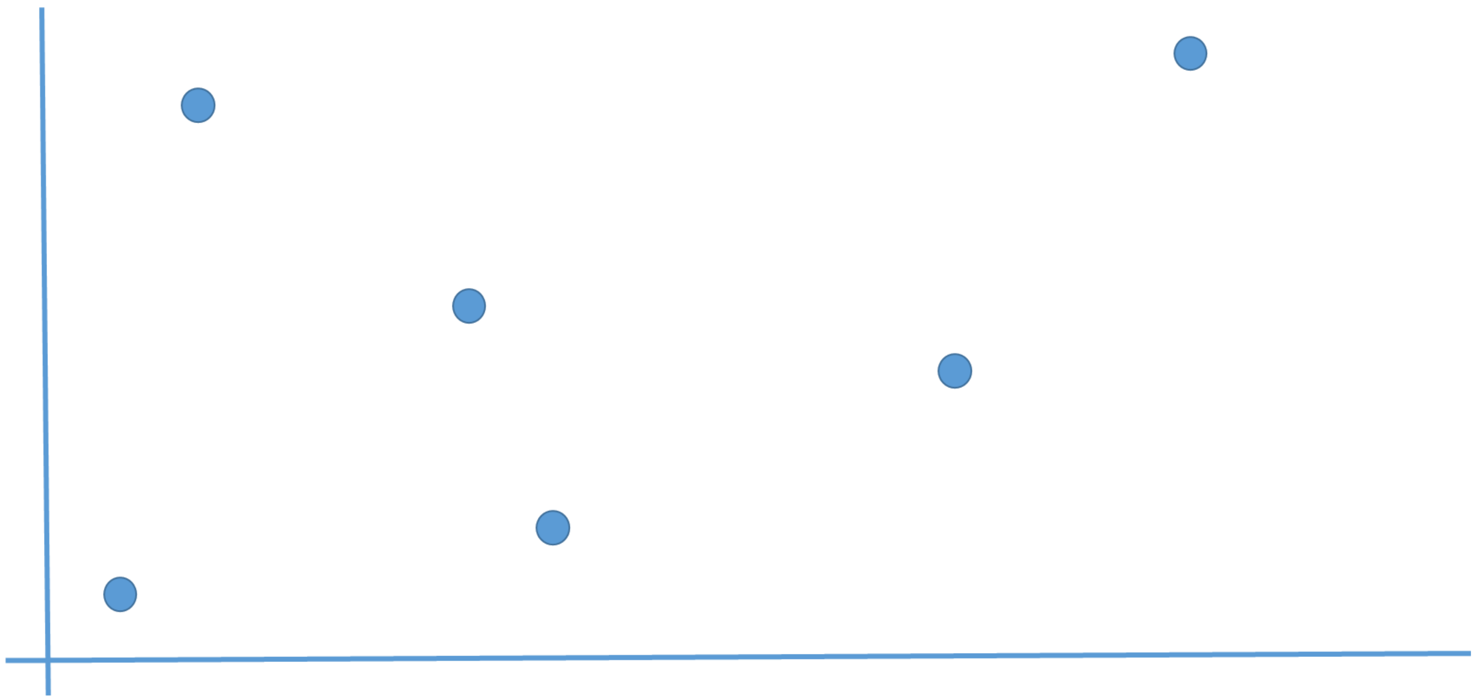
También es un punto de <u>máxima profundidad</u> en la muestra



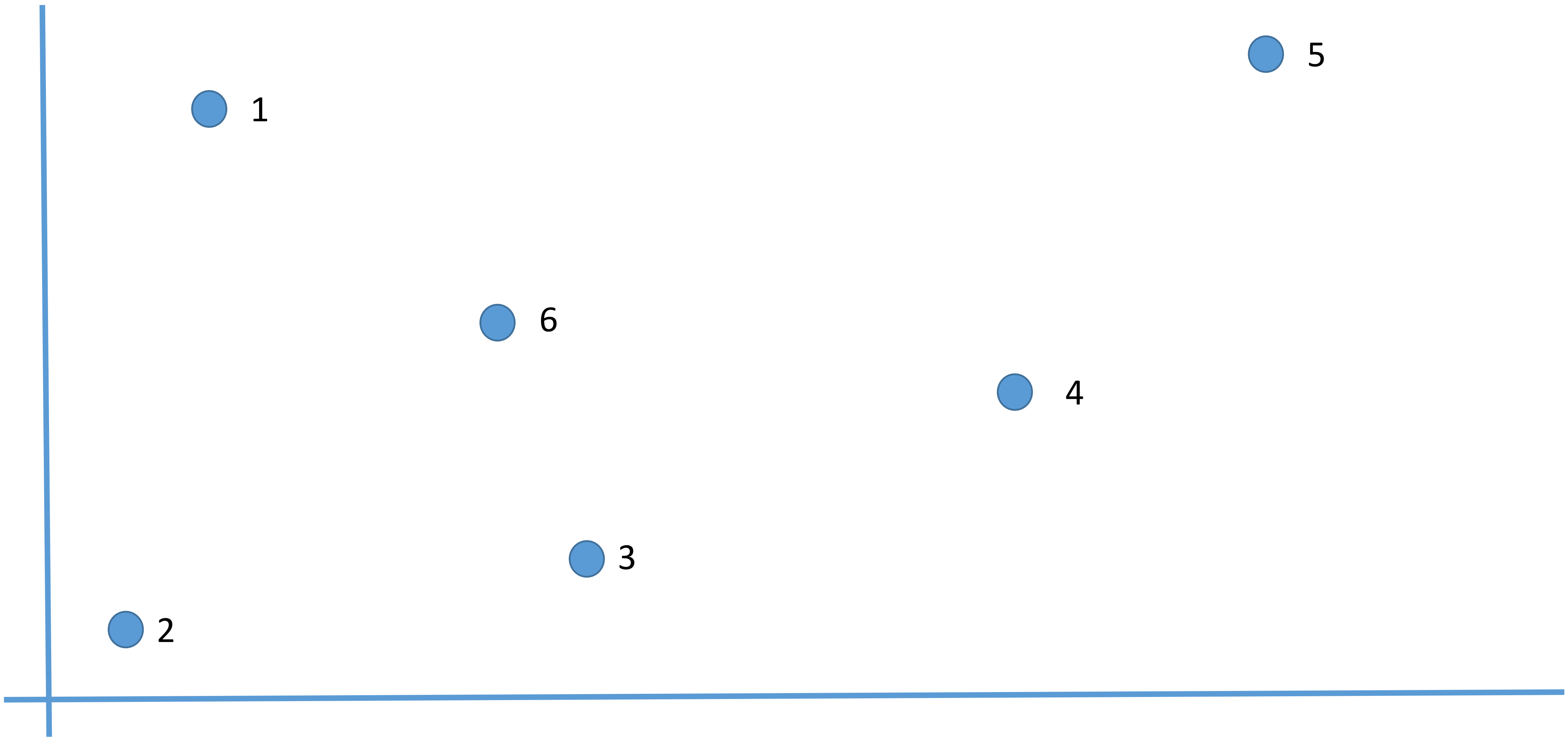
El punto de máxima profundidad resulta ser la mediana.

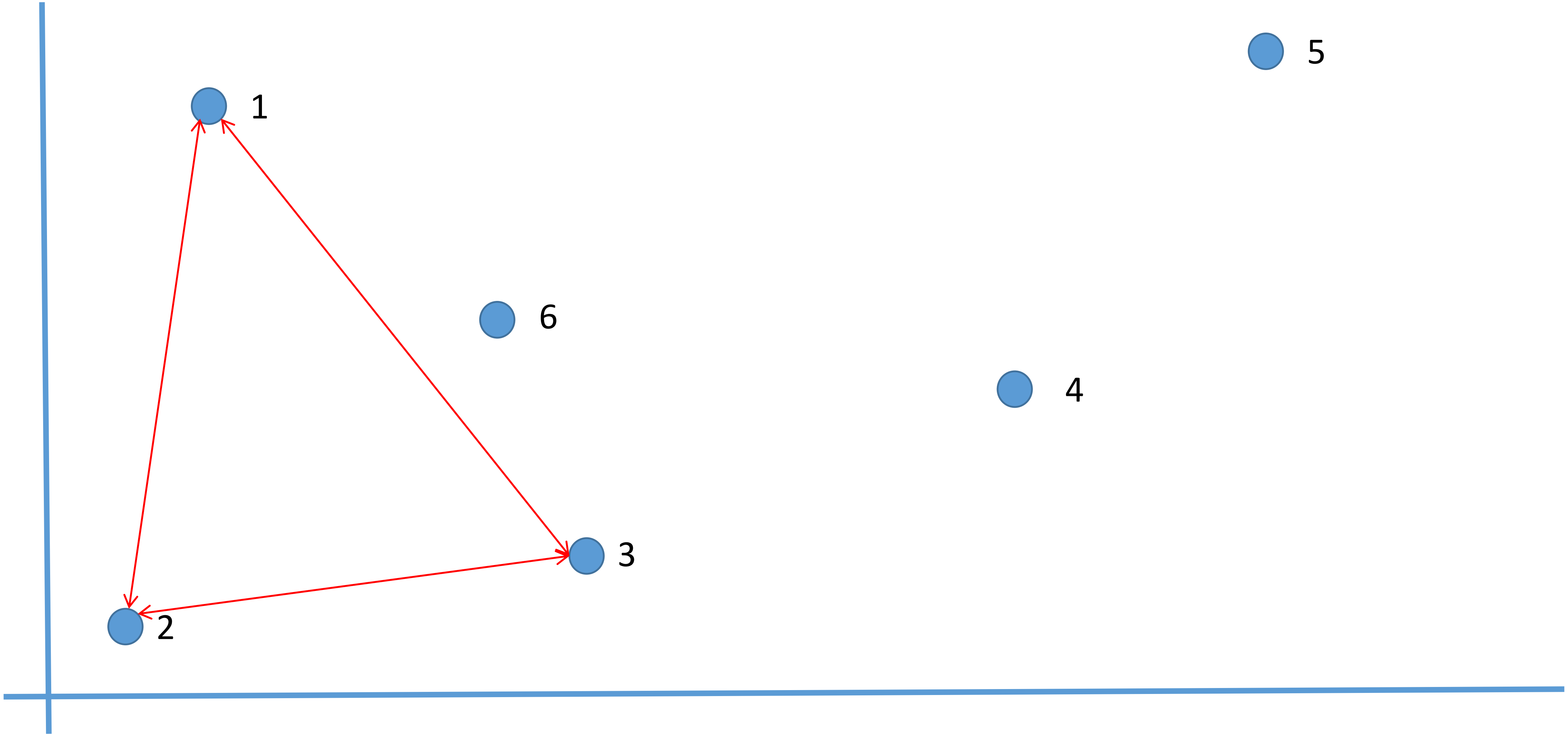
Este concepto puede generalizarse para calcular definir la mediana en espacios de 2 o más dimensiones.

Considere el siguiente arreglo de 6 puntos en el plano

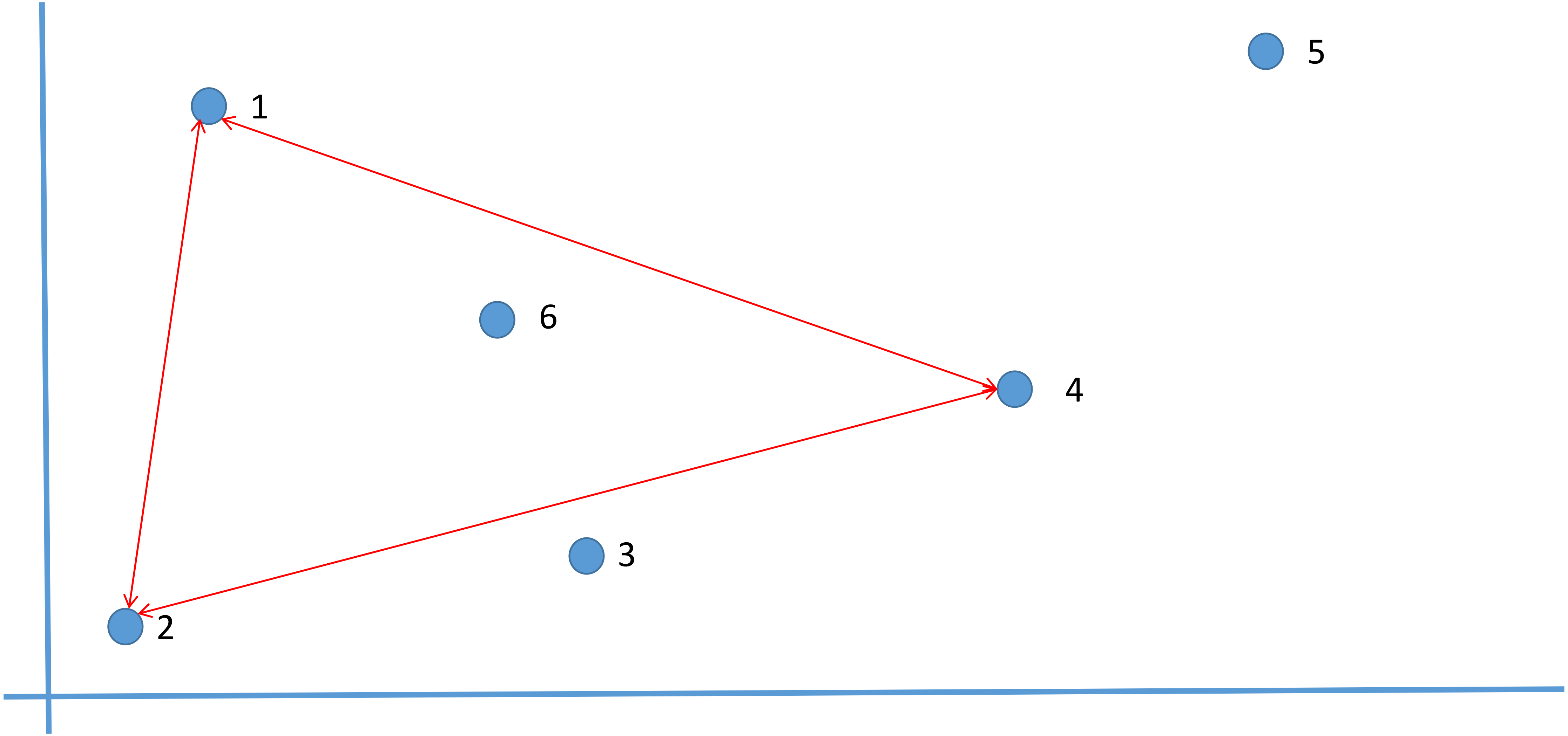


Tratemos de identificar la mediana uniendo ternas de puntos. El punto de máxima profundidad deberá estar contenido en el mayor número de puntos.

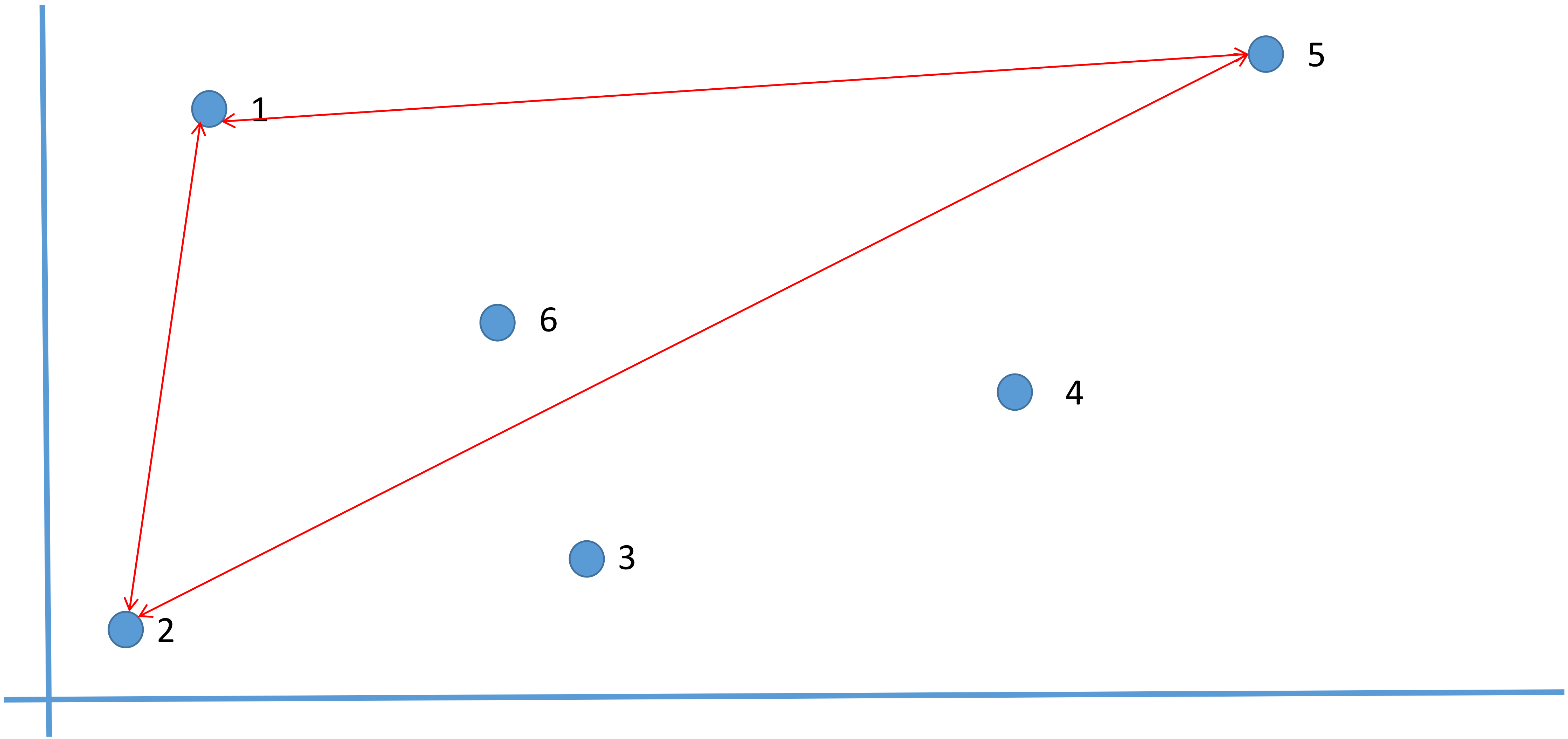

5
1
6
4
3
2

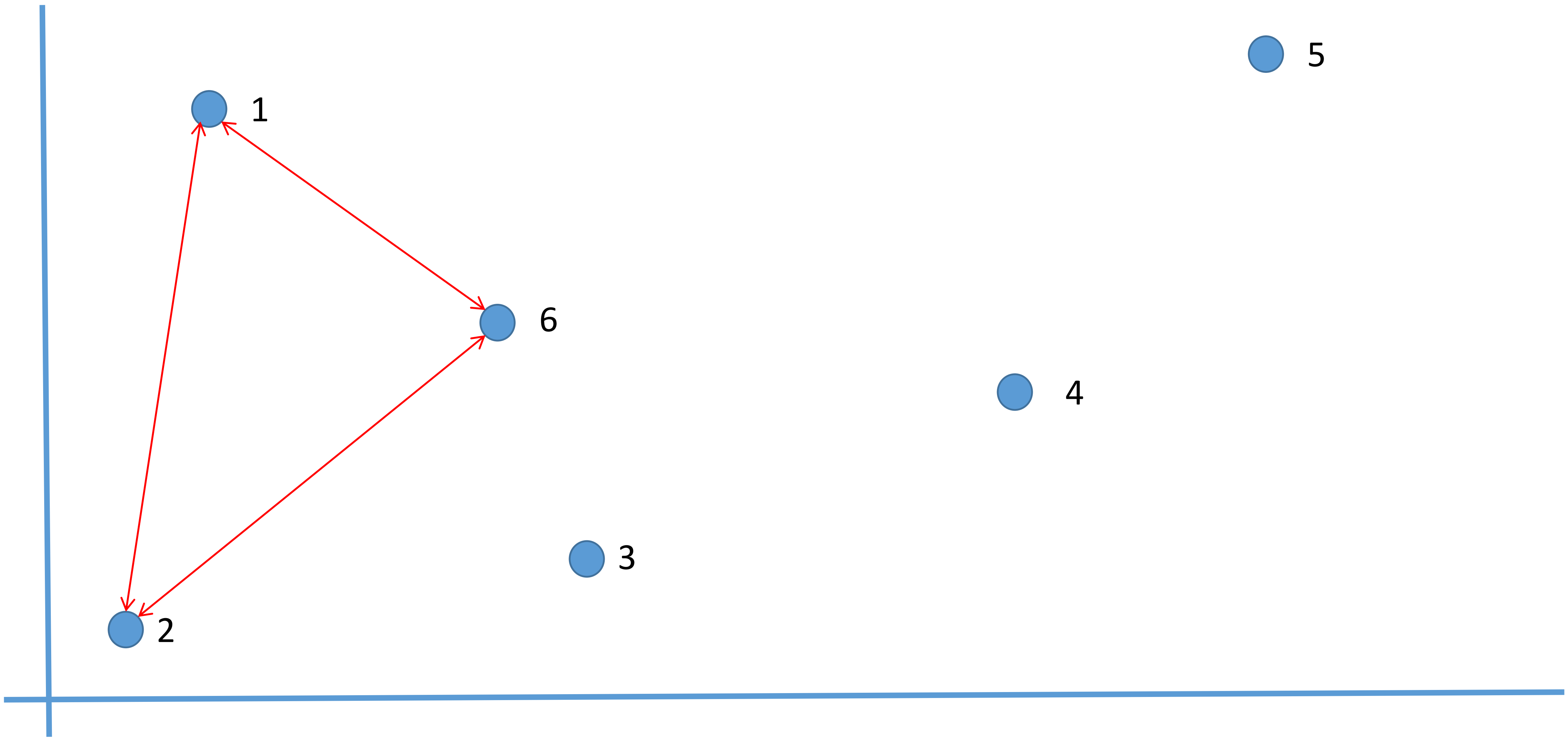


5
1
6
4
3
2

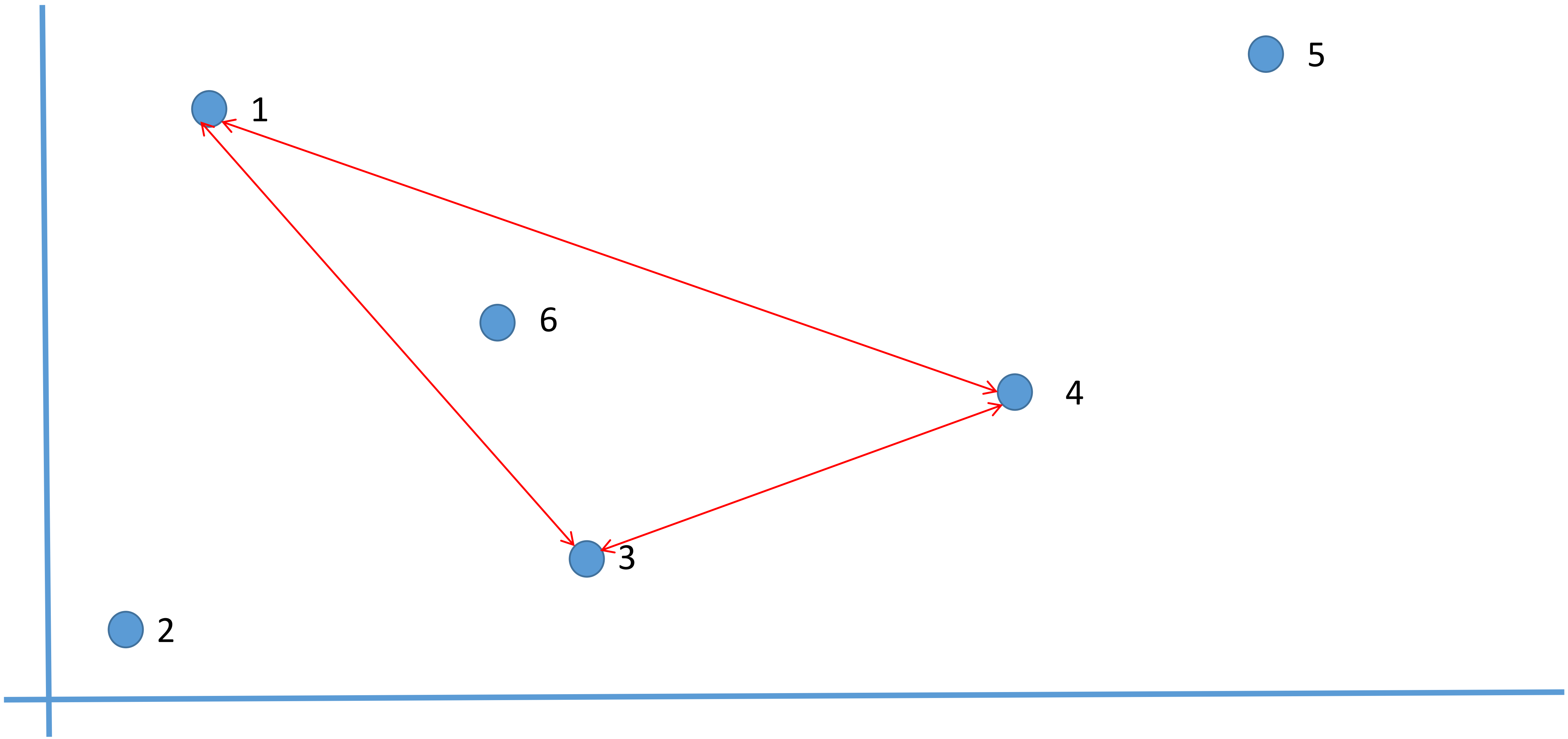


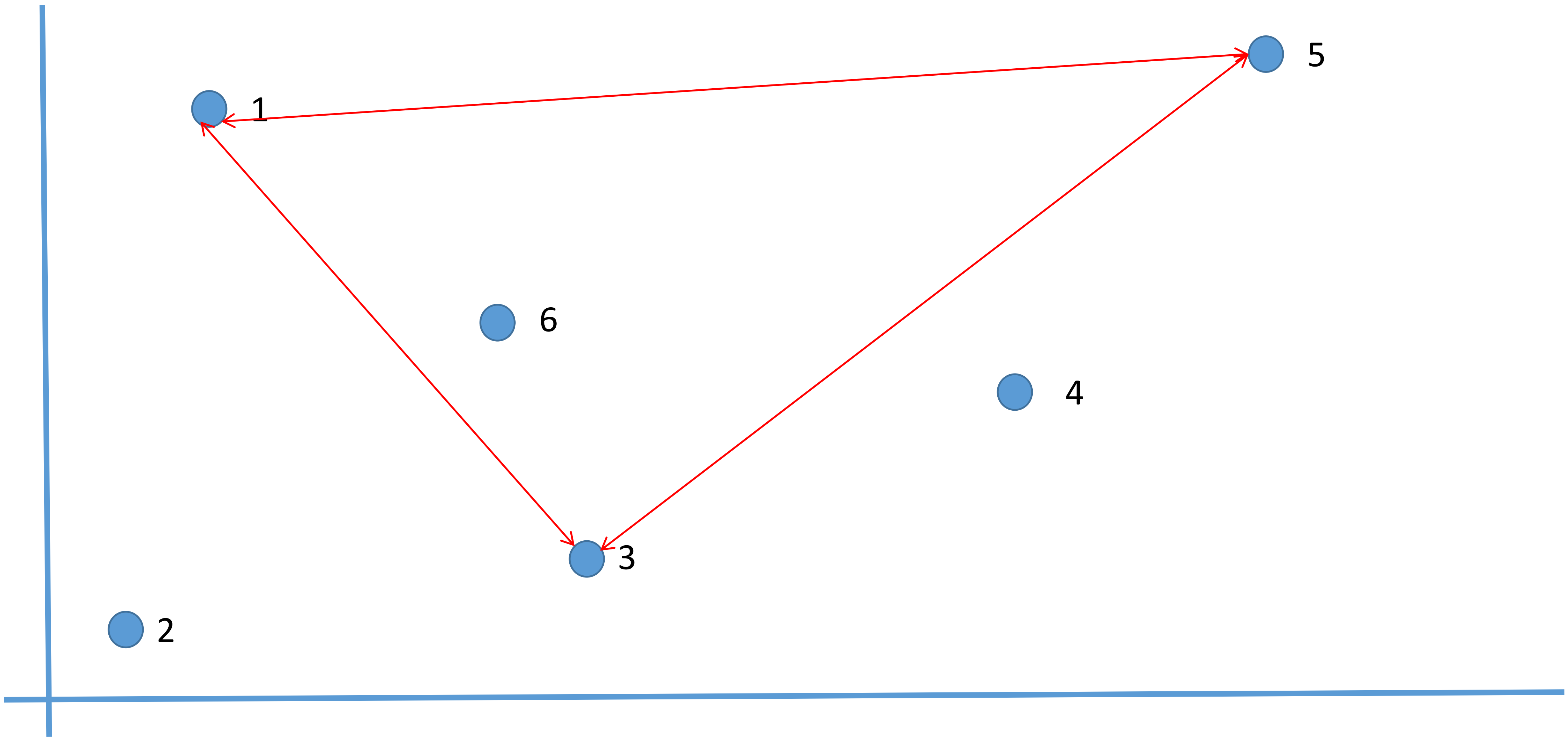
1
2
3
4
5
6

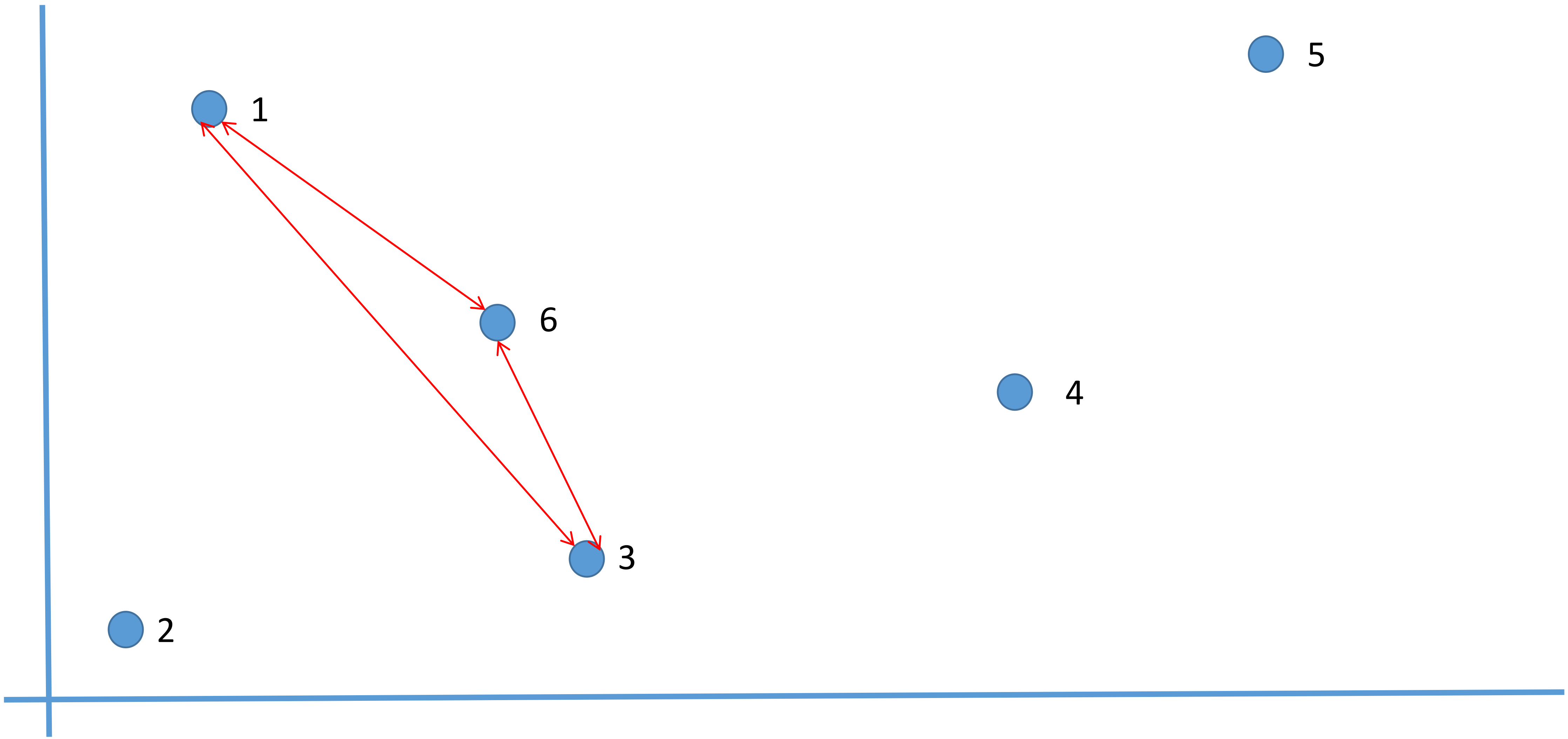

1
5
6
4
3
2

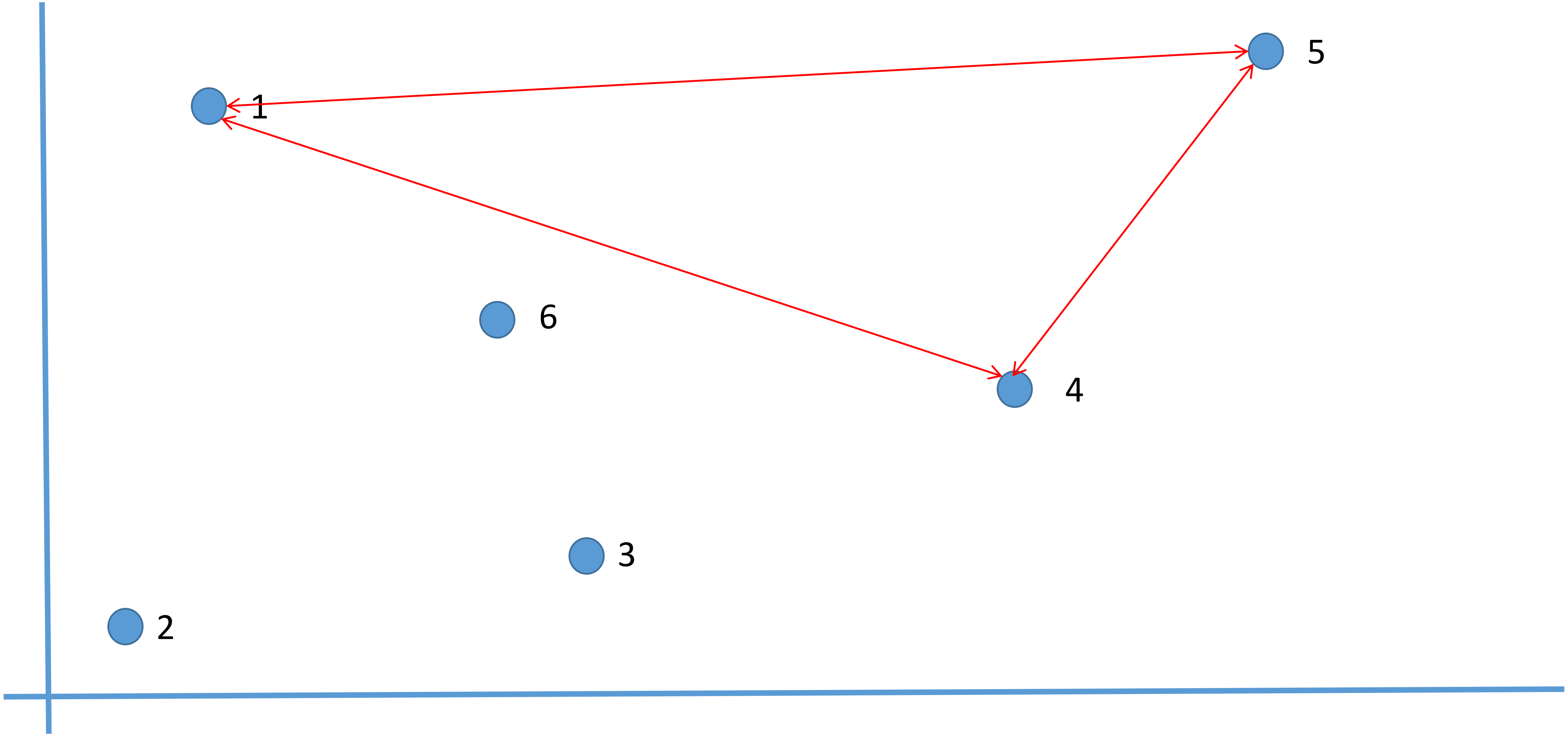


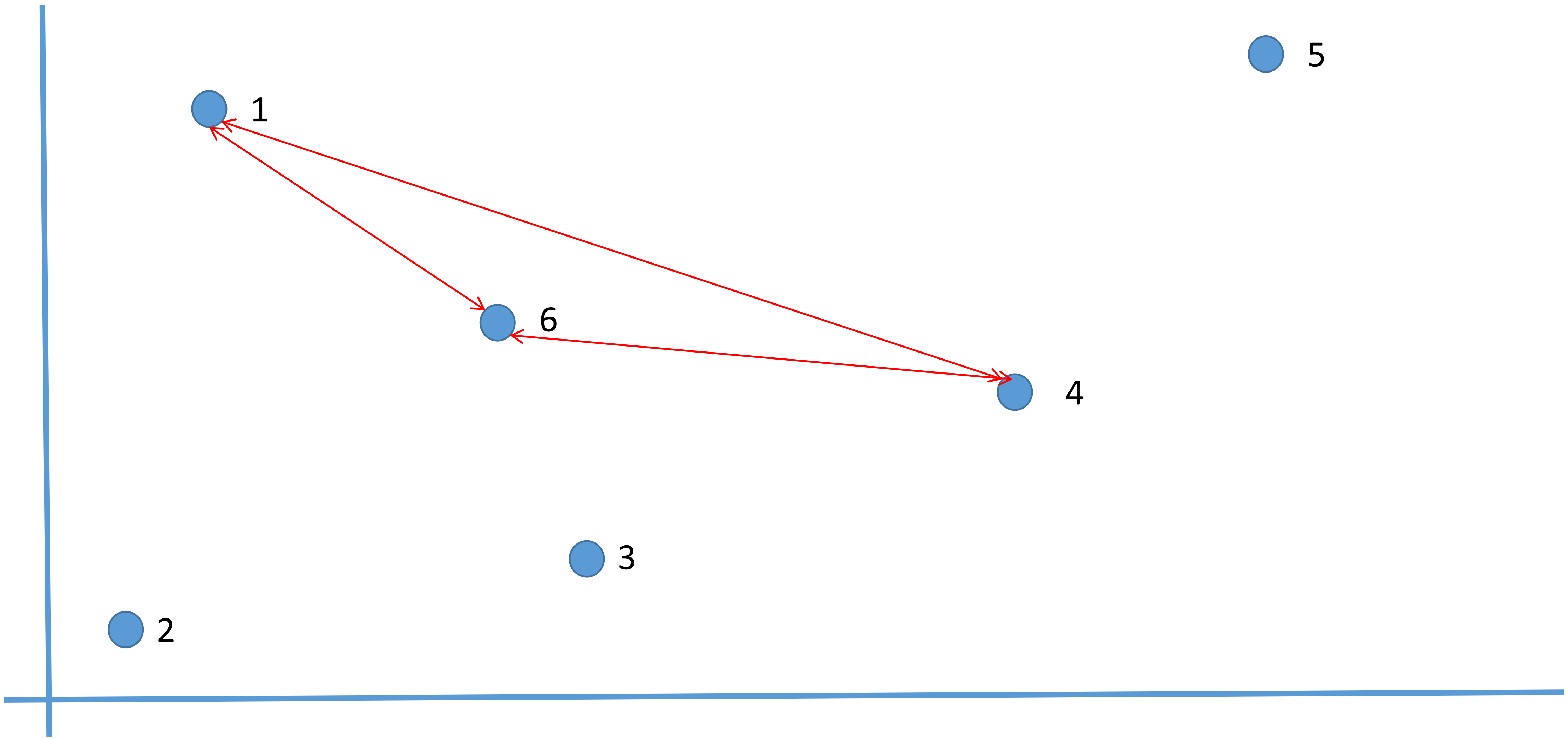
1
2
3
4
5
6

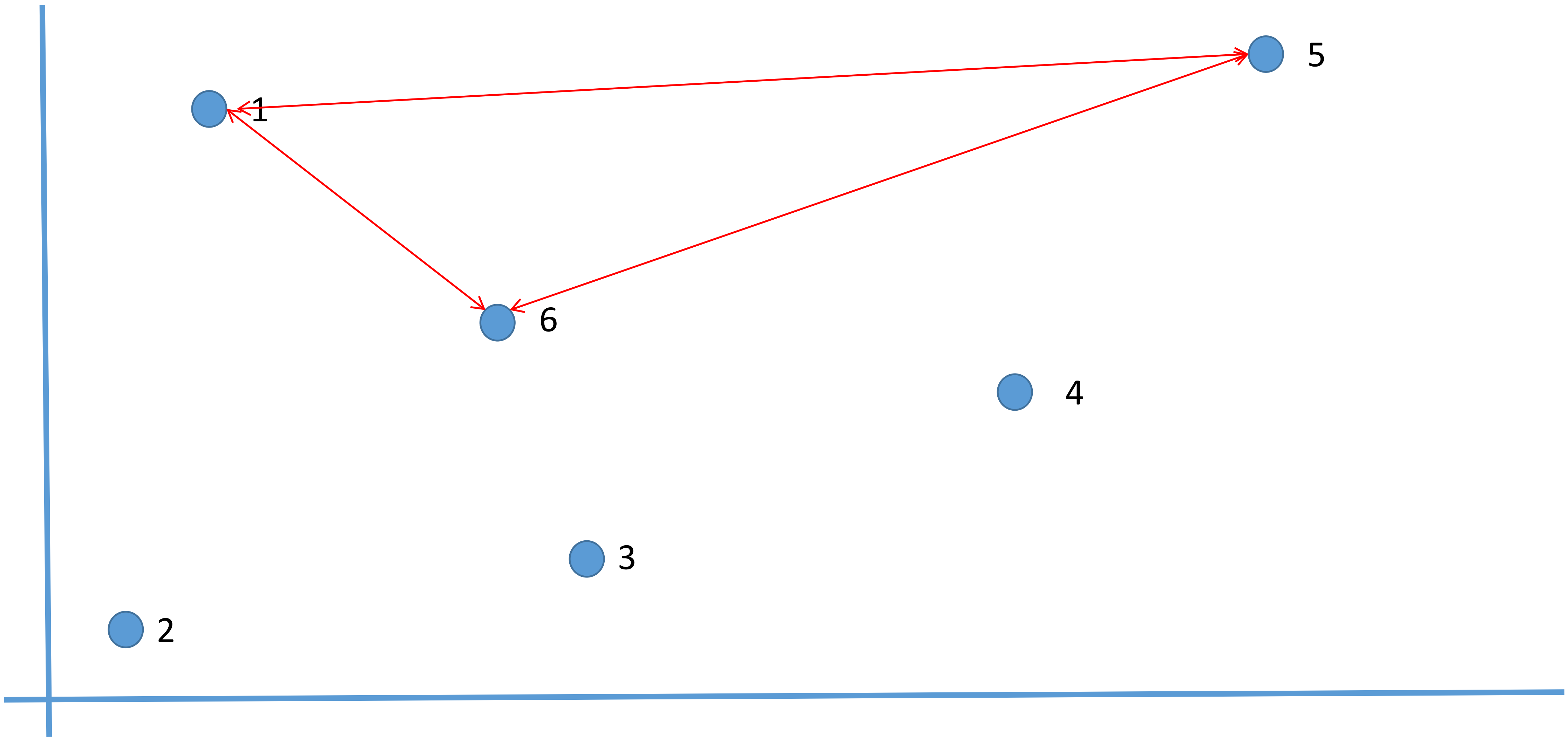

1
2
3
4
5
6

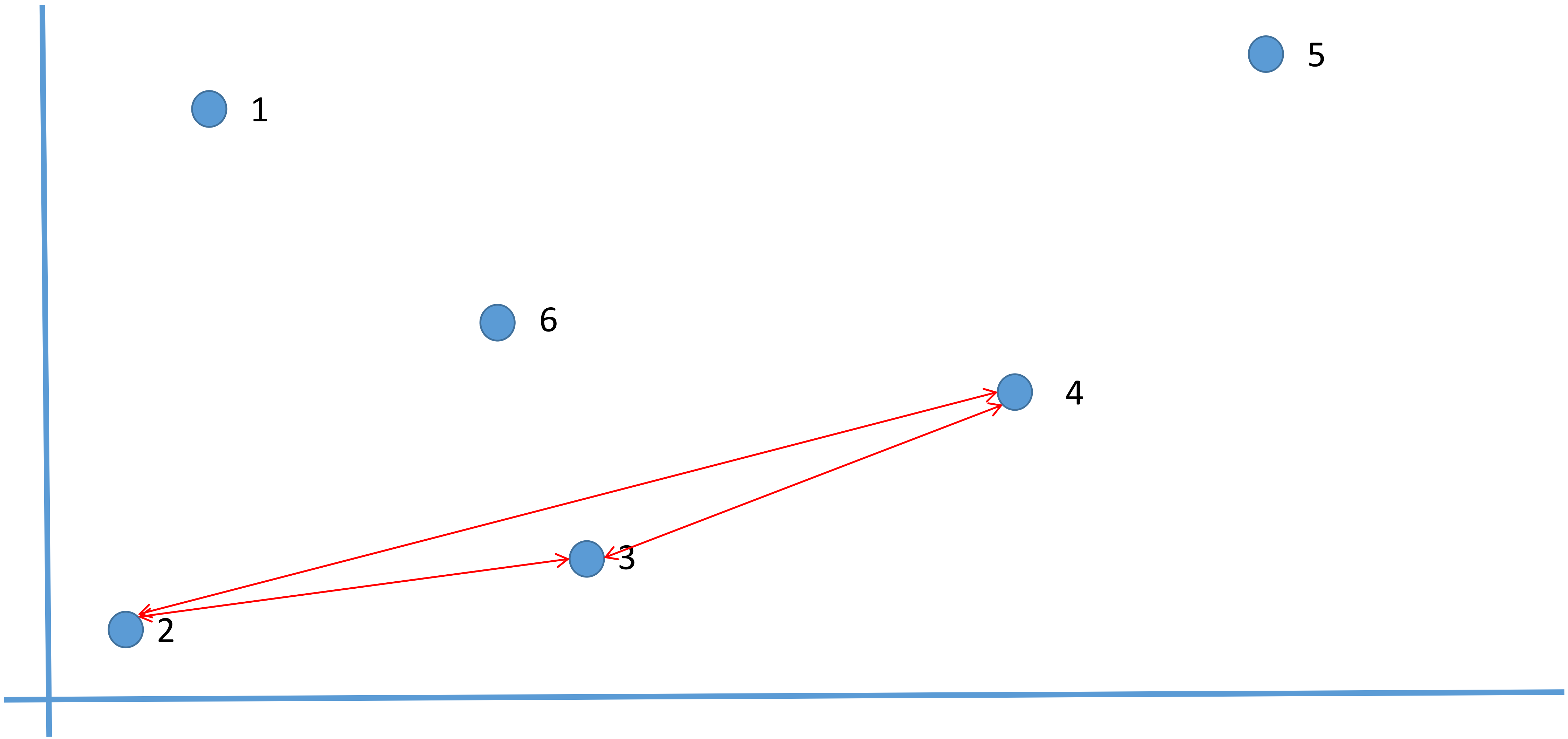

5
1
6
4
3
2

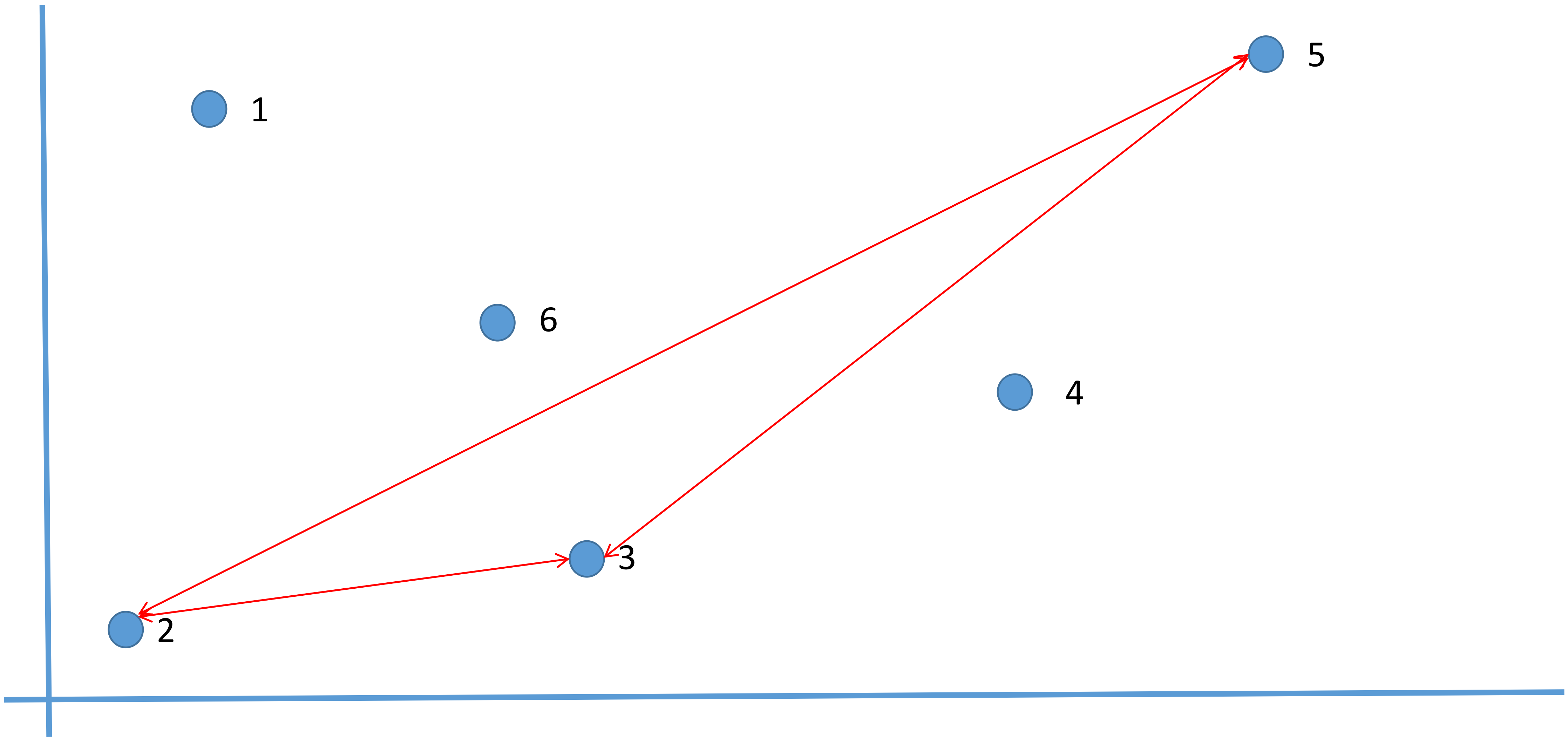

5
1
6
4
3
2

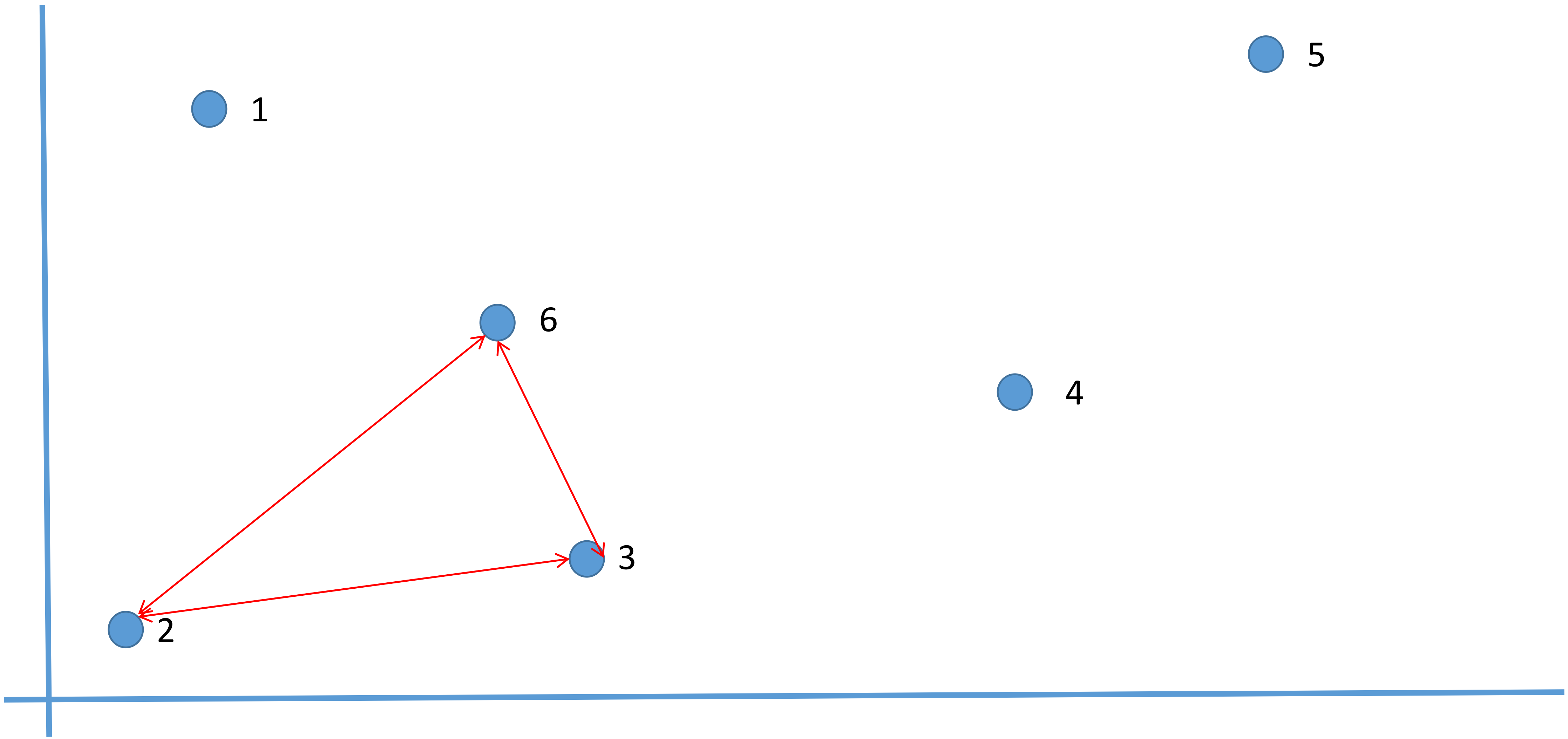

1
5
6
4
3
2

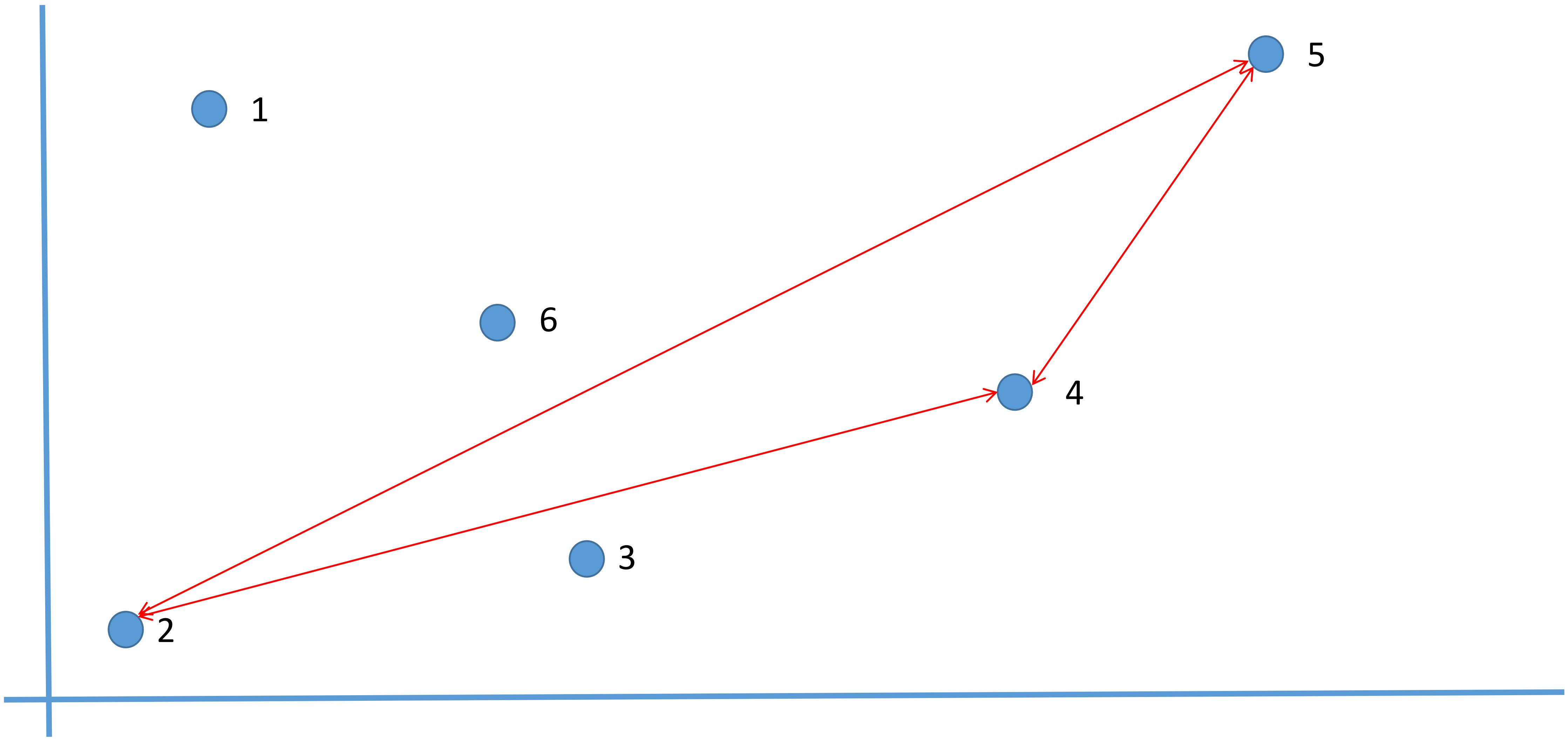

1
2
3
4
5
6

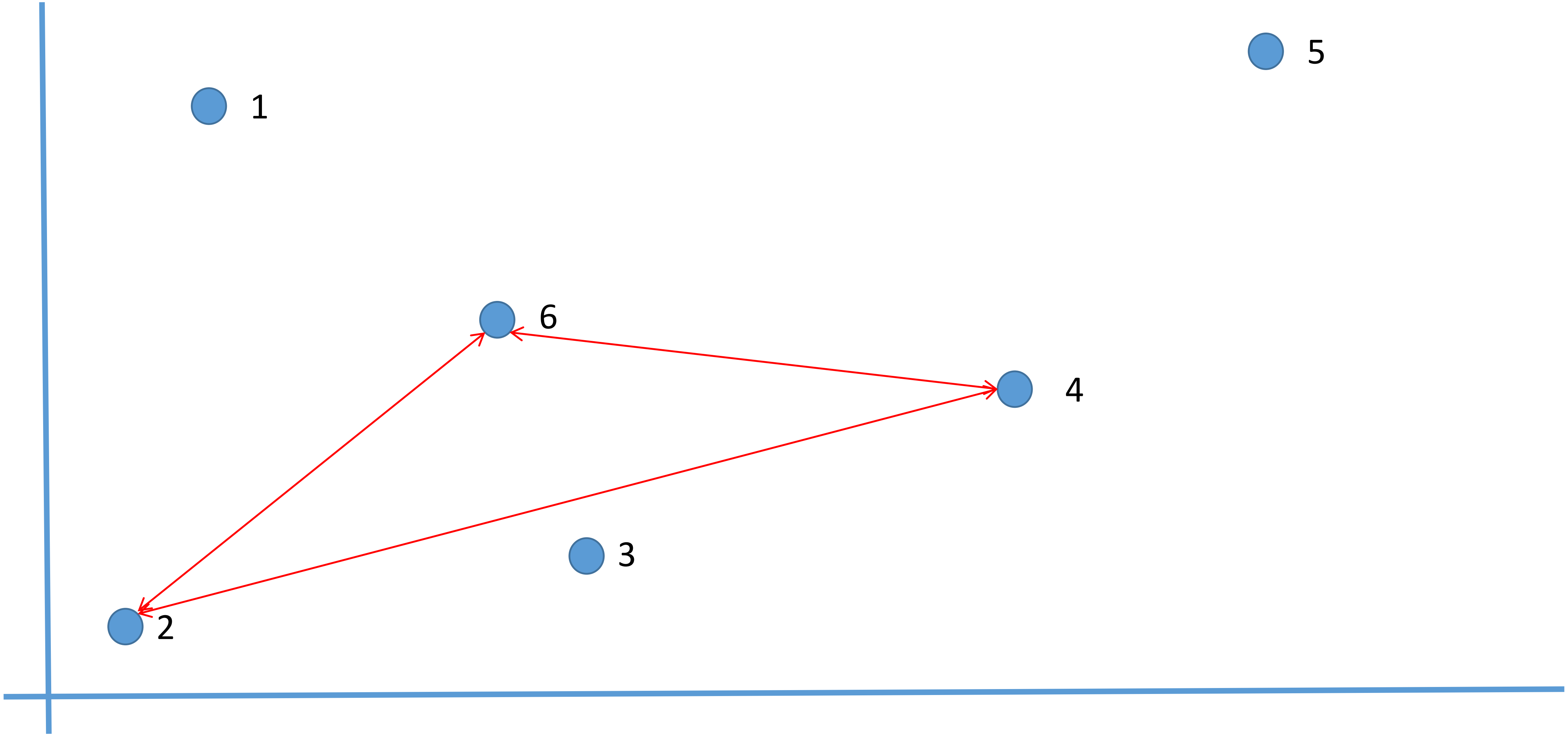

5
1
6
4
3
2

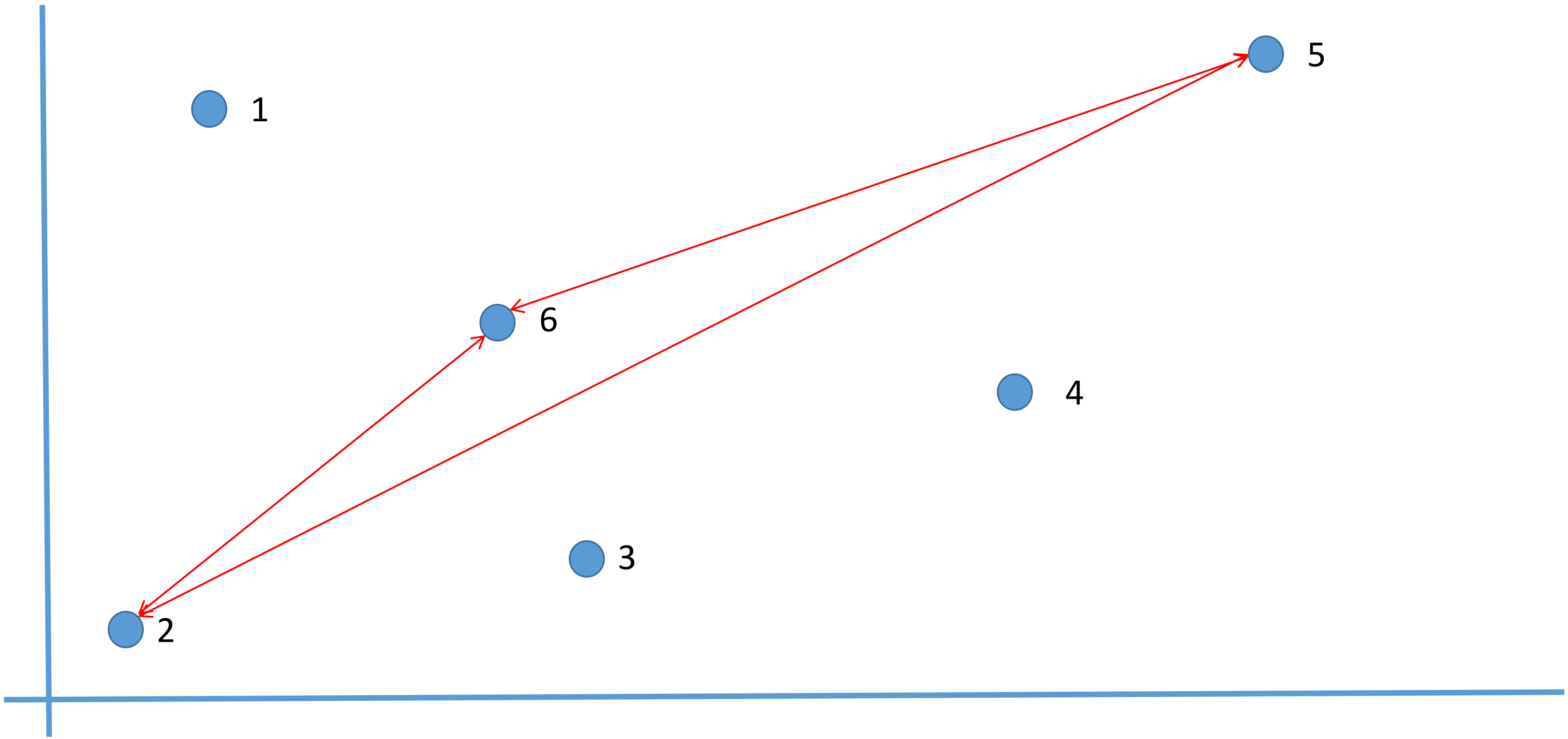

1
2
3
4
5
6

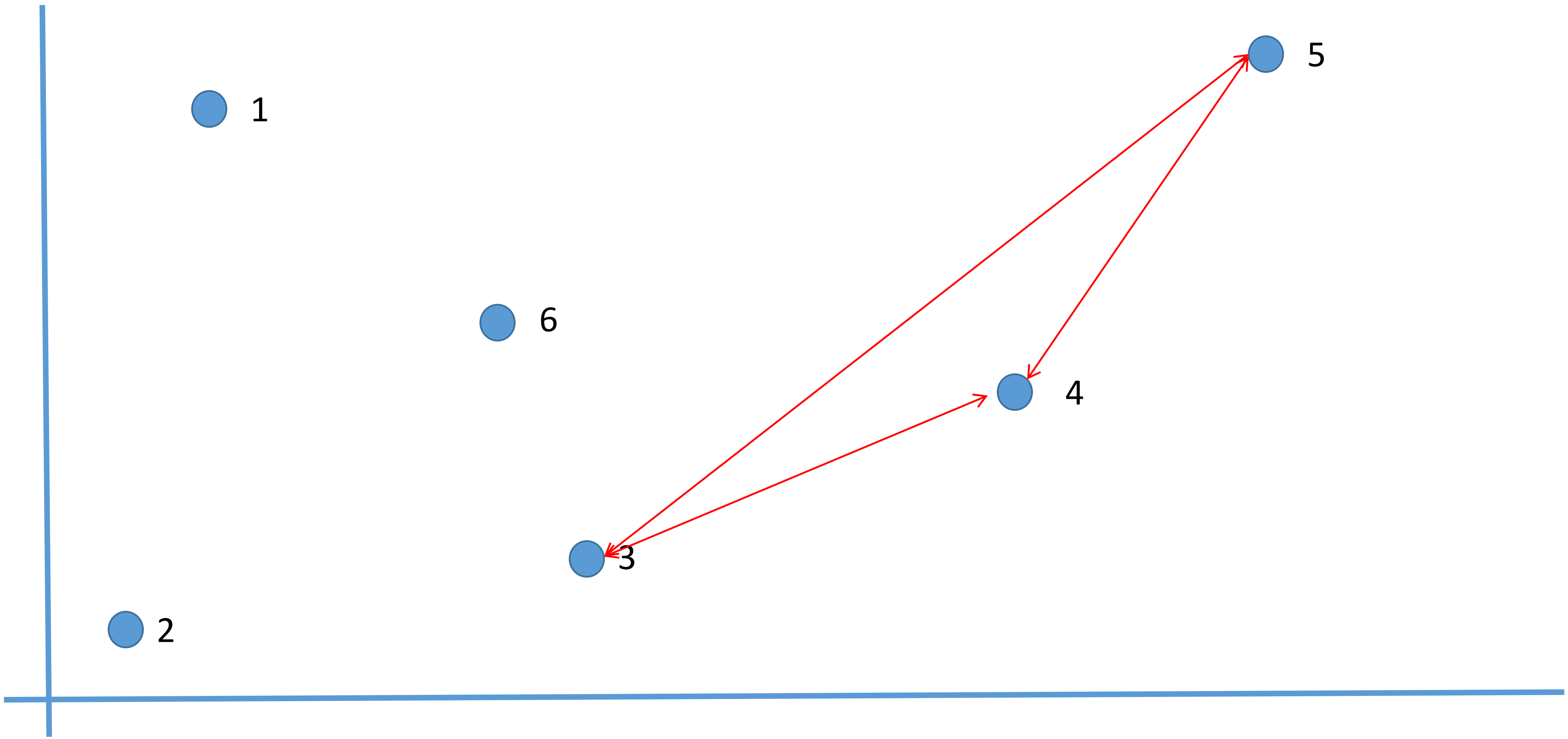

5
1
6
4
3
2

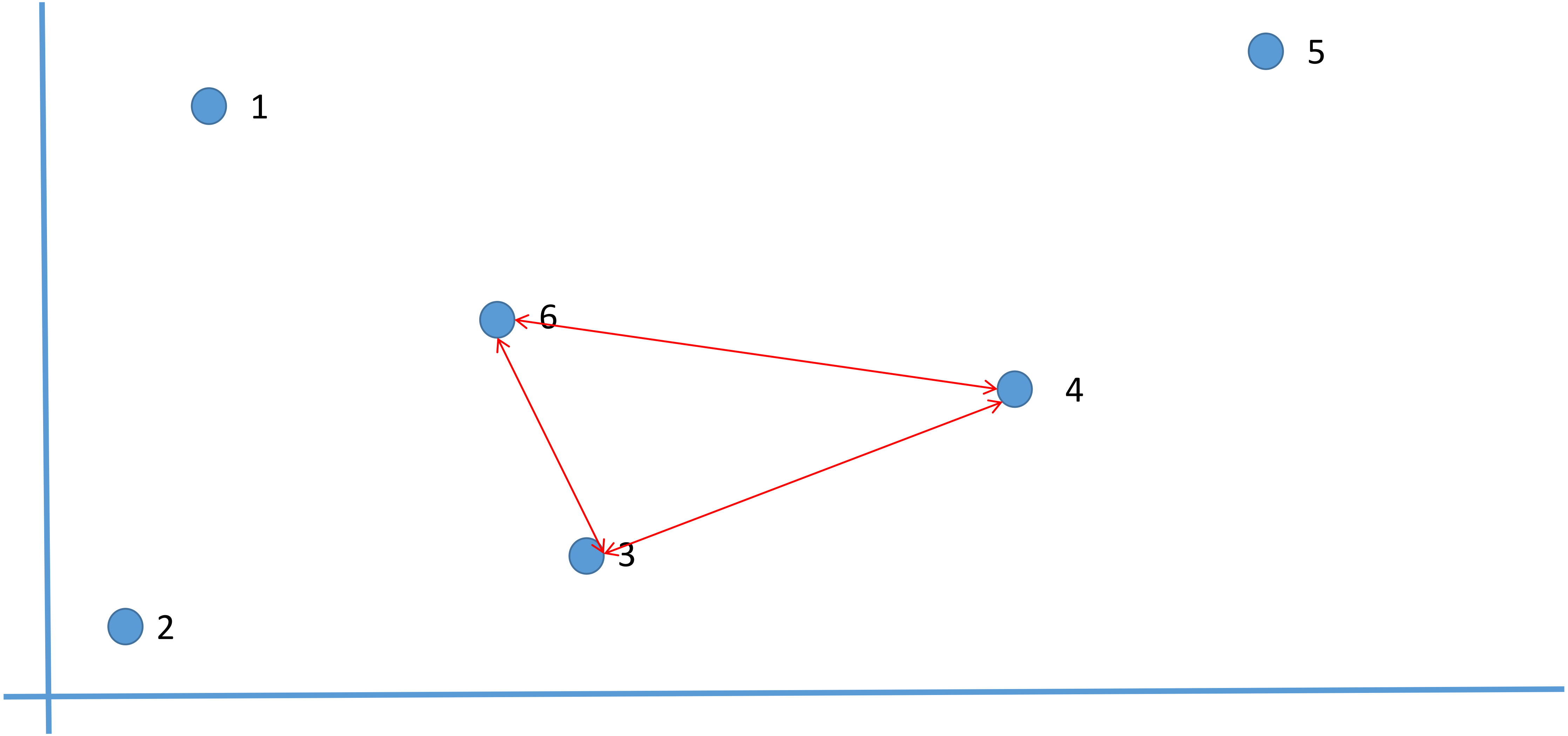

1
2
3
4
5
6

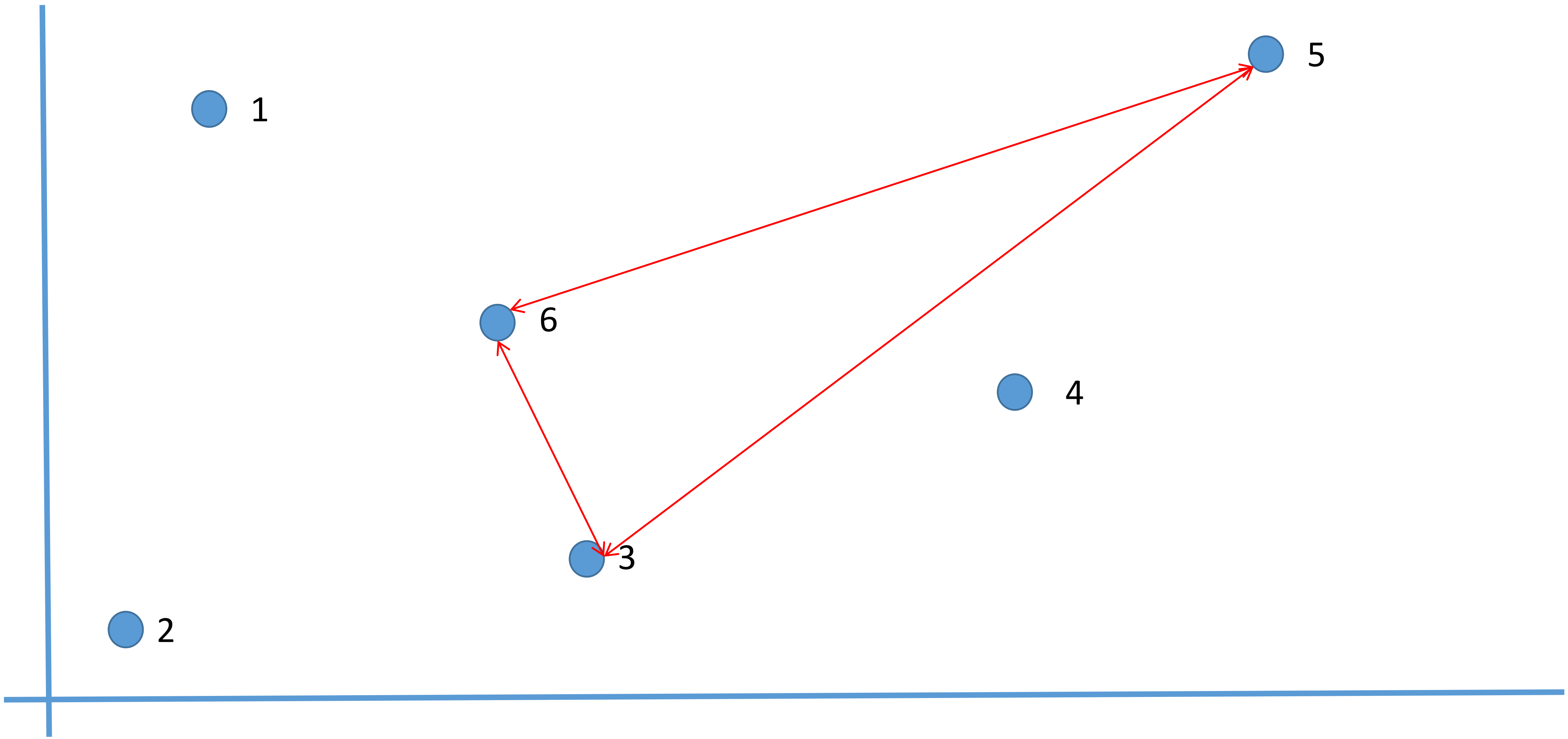

1
2
3
4
5
6

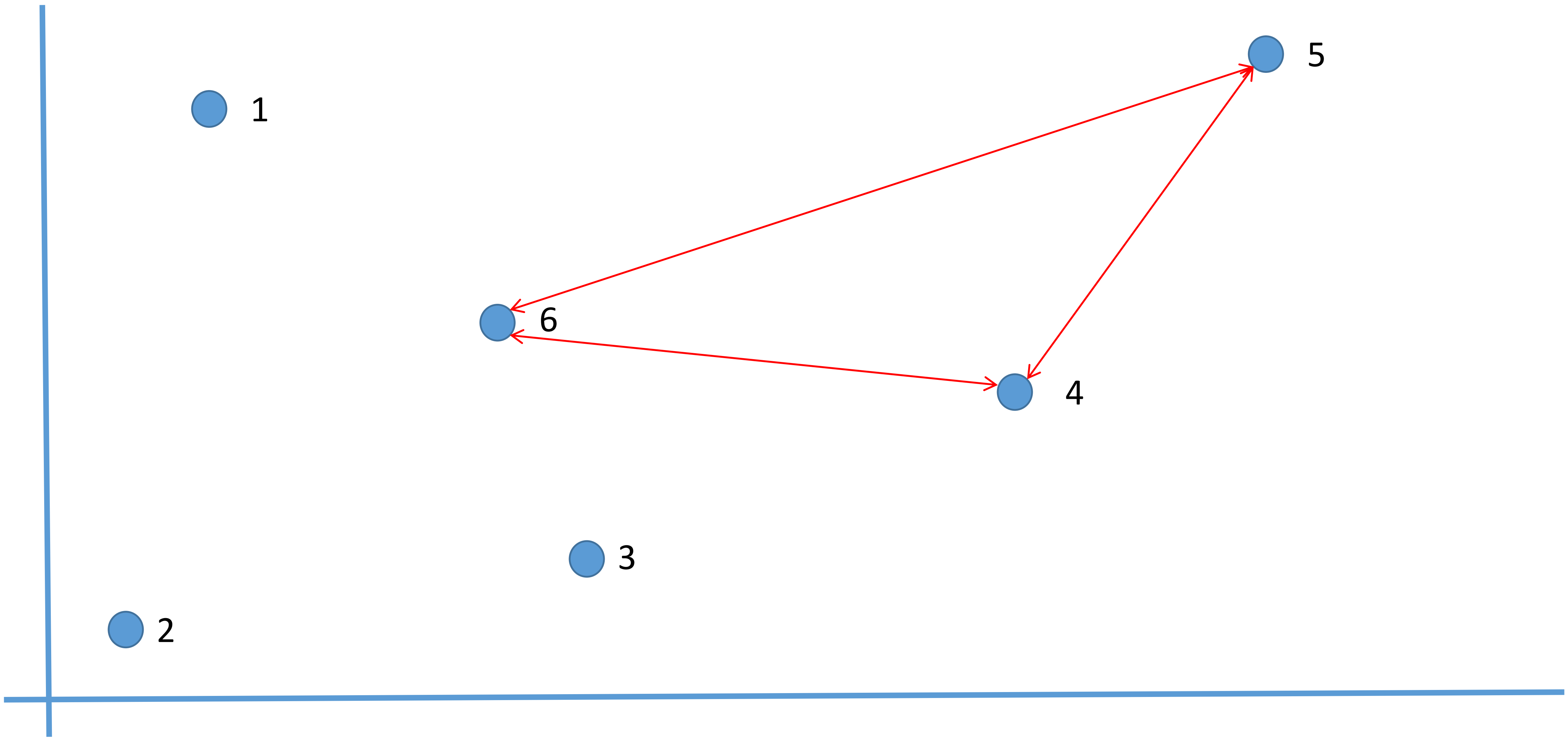

1
5
6
4
3
2

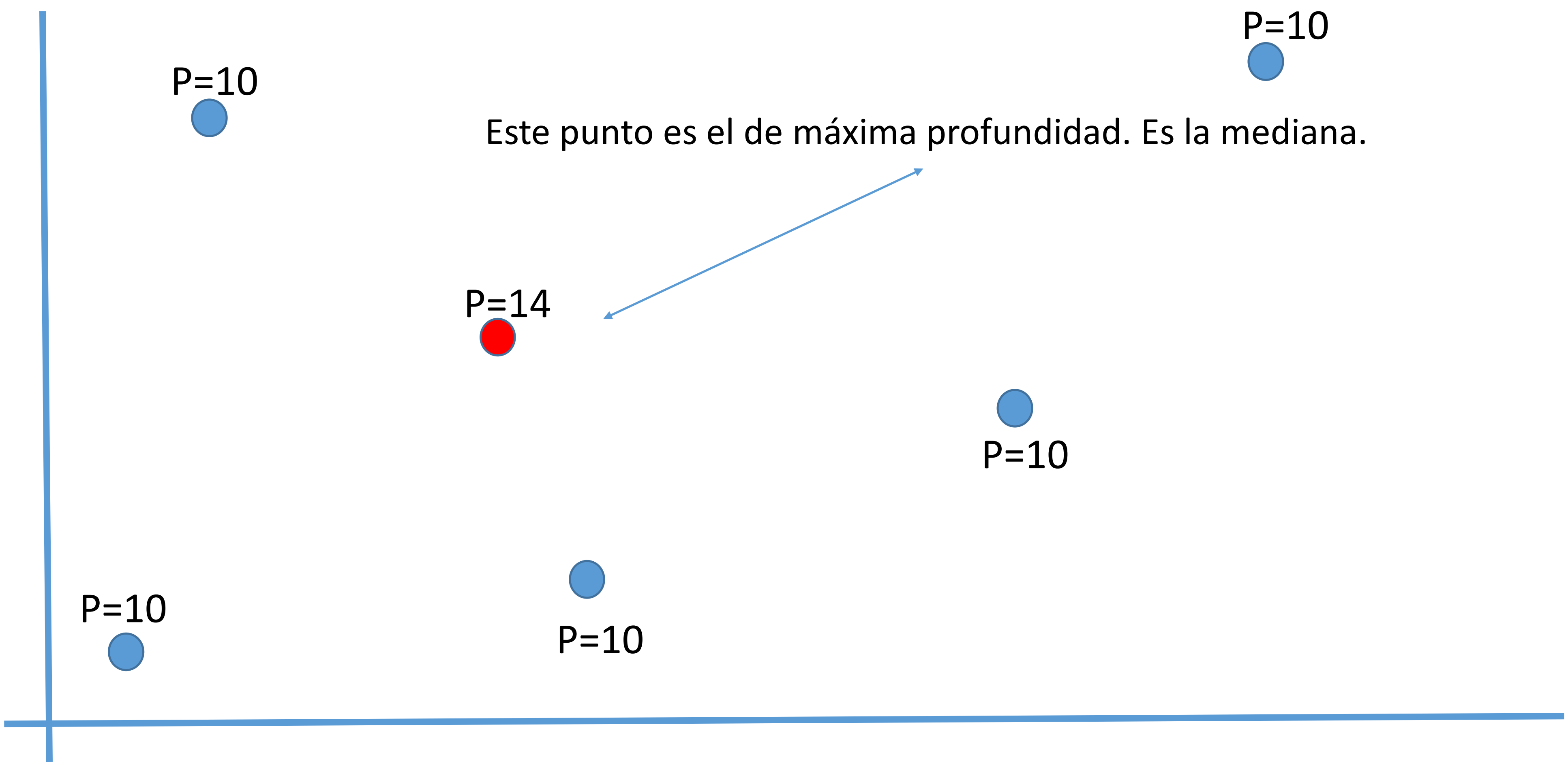

1
2
3
4
5
6


1
2
3
4
5
6


1
5
6
4
3
2


1
2
3
4
5
6


1
5
6
4
3
2


P=10
P=10
Este punto es el de máxima profundidad. Es la mediana.
P=14
P=10
P=10
P=10

La mediana muestral es una medida de tendencia central importante.

Es robusta ante *outliers*.

La mediana es el elemento más profundo en un conjunto finito de números reales

Este concepto de profundidad de datos permite generalizar la mediana para muestras observadas en espacios de dimensión mayor a 1.

Considere el siguiente problema estadístico

- **Contamos con medidas antropométricas de 88 estudiantes de licenciatura seleccionados al azar en el campus norte de la Universidad Anáhuac.**

- **42 estudiantes son mujeres y el resto son hombres**

- **Para cada individuo se tomaron medidas de 6 variables antropométricas: talla, masa corporal, tórax, cintura, cadera y circunferencia del cráneo**

- **¿Podríamos proponer un criterio que, con base en esta información, permita clasificar a una persona como hombre o mujer a partir de estas variables?**

¡Veamos!

- **La profundidad de un punto x con respecto a una muestra observada $x_1, x_2, \ldots$ x_n es una función : $P\colon R^p \to R$**

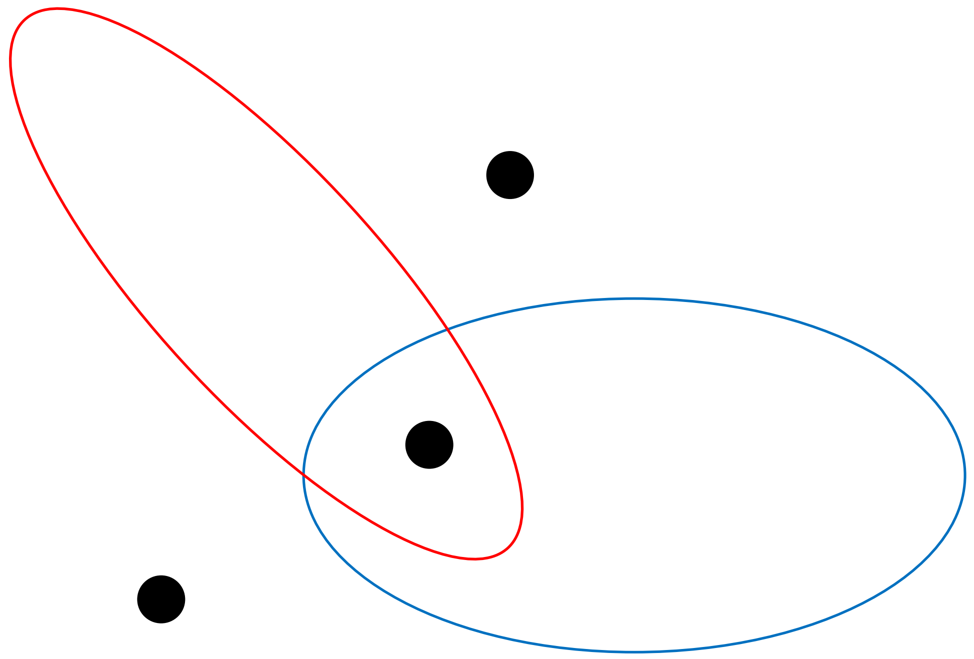
- **indica qué tan inmerso está el punto x con respecto a la nube definida por la muestra.**

- **La Profundidad de Mahalanobis se define como:**

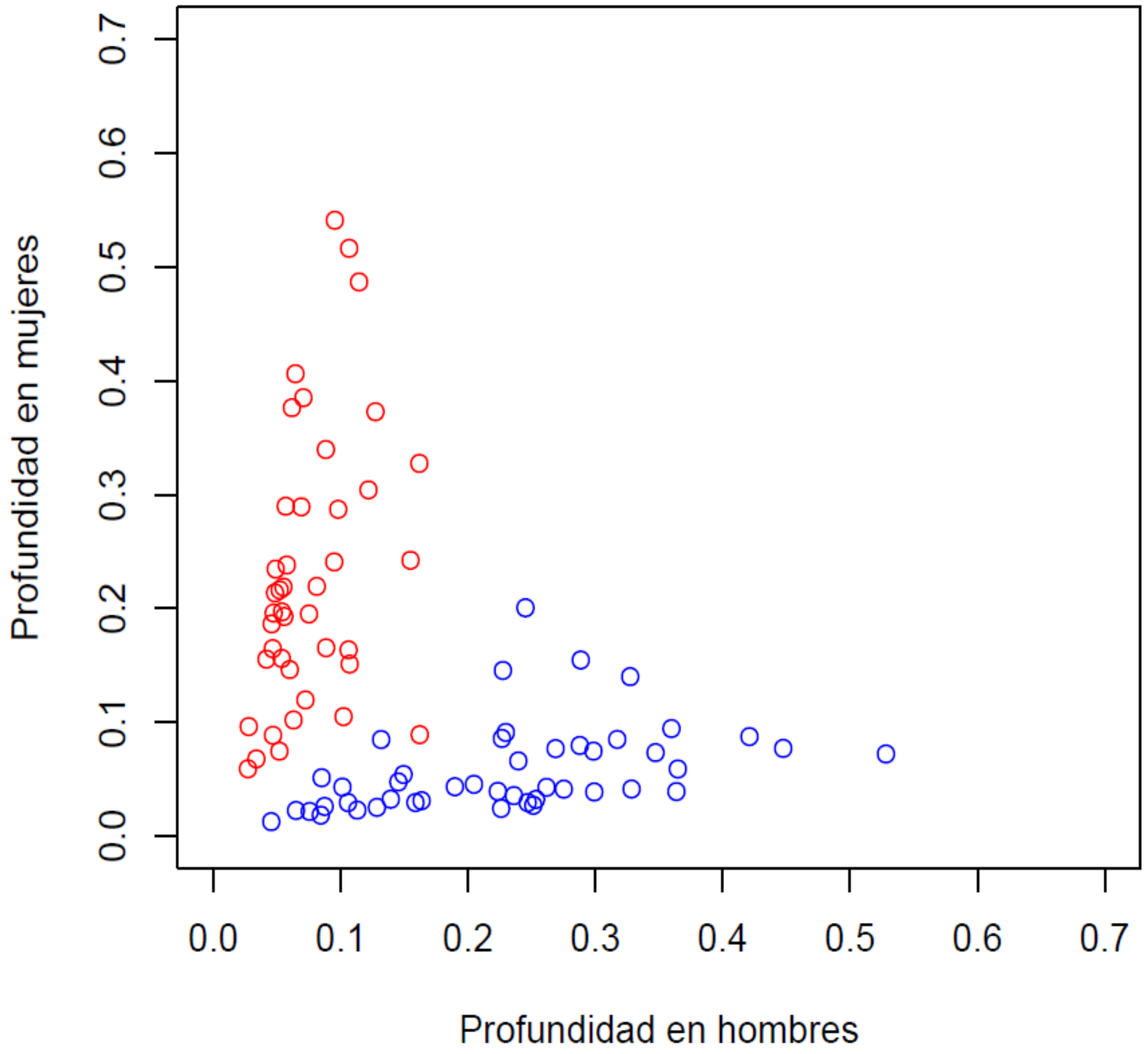
$$\boldsymbol{P(x) = \frac{1}{1 + (x - \overline{x})^t \Sigma^{-1} (x - \overline{x})}}$$

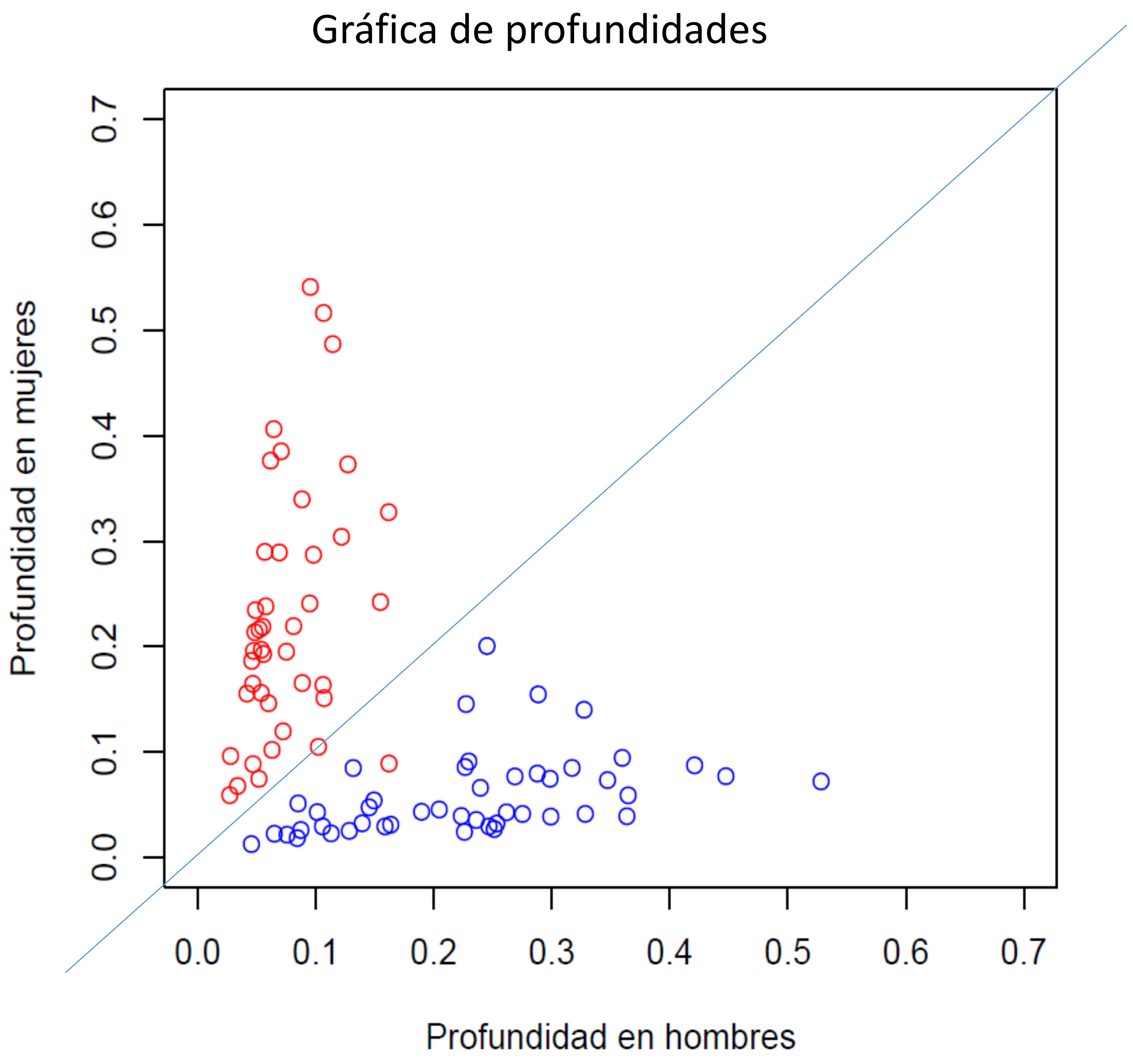
- **Dado un nuevo individuo, podemos medir su profundidad respecto de cada muestra.**

- **Con base en esta información podemos clasificar en la categoría la que alcanza su máxima profundidad de Mahalanobis.**



Gráfica de profundidades




Gráfica de profundidades
Profundidad en mujeres
Profundidad en hombres
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

```
function (X,Y)
{
################################################
##
## ESTA FUNCIÓN CONSTRUYE DIAGRAMAS DE PROFUNIDAD
##
## x = matriz de datos de la población Omega0
## Y = matriz de datos de la población Omega1
##
###############################################

Inicialización de variables
px<-ncol(X)
nx<-nrow(X)

py<-ncol(Y)
ny<-nrow(Y)

Dx<-matrix(nrow = nx, ncol = 2)
Dy<-matrix(nrow = ny, ncol = 2)
```

```
## Cálculo de medias y matrices de covarianza

mx<-vector(length=px)
for(i in 1:px){mx[i]<-mean(X[,i])}

my<-vector(length=py)
for(i in 1:px){my[i]<-mean(Y[,i])}

Sx<-cov(X)
Sy<-cov(Y)

################################################## Cálculo de profundidades; distacia de Mahalanobis

for(i in 1:nx){Dx[i,1]<- 1/(1+mahalanobis(X[i,],mx,Sx,inverted=FALSE)); Dx[i,2]<-
1/(1+mahalanobis(X[i,],my,Sy,inverted=FALSE))}
for(i in 1:ny){Dy[i,1]<- 1/(1+mahalanobis(Y[i,],mx,Sx,inverted=FALSE)); Dy[i,2]<-
1/(1+mahalanobis(Y[i,],my,Sy,inverted=FALSE))}

#############################################
## Presentación del diagrama

plot(Dx[,1], Dx[,2], xlim = c(0,0.7), ylim = c(0,0.7), type = "p", col = "blue", xlab = "Profundidad en
hombres", ylab = "Profundidad en mujeres")
lines(Dy[,1], Dy[,2], xlim = c(0,0.7), ylim = c(0,0.7), type = "p", col = "red", xlab="", ylab="")


}
```

- **La profundidad de datos permite construir criterios de clasificación supervisada, con ptencial aplicativo en Ciencias Actuariales. Esta solución es libre de distribución.**

- **Ofrece soluciones no paramétricas con una interpretación geométrica clara e intuitiva**

- **Constituyen una herramienta útil para la toma de decisiones basadas en información empírica (*data driven*).**

- **Por sus características, el *dd-plot* no sólo es un criterio de clasificación supervisada, sino un medio visualización para describir gráficamente observaciones en espacios de dimensiones mayores a dos.**

¡Muchas gracias a todos por su atención!

UNIVERSIDAD ANÁHUAC
VINCE IN BONO MALUM
MÉXICO NORTE

50 AÑOS
Facultad de Ciencias Actuariales