Variational EM and K-Means

Machine Learning
CSx824/ECEx242
Bert Huang
Virginia Tech

Outline

- Bounding the log marginal likelihood
- Families of variational distributions
  - Fully factorized distributions
  - Point distributions
- K-means as “hard”-EM

Marginal Likelihood

$$p(X|\theta) = \int_Z p(X, Z|\theta) dZ$$

$$\sum_Z p(X, Z|\theta)$$

e.g., $X = \{x_1, \ldots, x_n\}$

$\theta = \{\mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, \ldots p(c)\}$

$Z = \{z_1, \ldots, z_n\}$ (cluster memberships)

$$p(X, Z|\theta) = \prod_{i=1}^{n} p(z_i)\mathcal{N}(x_i|\mu_{z_i}, \Sigma_{z_i})$$

$$\log p(X|\theta) = \log \sum_Z p(X, Z|\theta)$$

log marginal likelihood

$$\underset{\theta}{\text{argmax}} \log \sum_Z p(X, Z|\theta)$$

learning objective

Jensen's Inequality

For any convex function φ,

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$$

For any concave function ϕ,

$$\phi(\mathbb{E}[X]) \geq \mathbb{E}[\phi(X)]$$

$$\log(\mathbb{E}[X]) \geq \mathbb{E}[\log X]$$

Variational Bound

$$\log(\mathbb{E}[X]) \geq \mathbb{E}[\log X]$$

$$\log \sum_Z p(X, Z|\theta) = \log \sum_Z \frac{q(Z)}{q(Z)} p(X, Z|\theta) \qquad \sum_Z q(Z) = 1$$

$$= \log \sum_Z q(Z) \frac{p(X, Z|\theta)}{q(Z)}$$

$$\geq \sum_Z q(Z) \log \frac{p(X, Z|\theta)}{q(Z)}$$

$$= \sum_Z q(Z) \log p(X, Z|\theta) - \sum_Z q(Z) \log q(Z)$$

Variational Bound

expectation entropy

$$\log \sum_Z p(X, Z|\theta) \geq \sum_Z q(Z) \log p(X, Z|\theta) - \sum_Z q(Z) \log q(Z)$$

We can pick any ***q*** distribution and the bound holds

$$\underset{\theta, q \in Q}{\text{argmax}} \sum_Z q(Z) \log p(X, Z|\theta) - \sum_Z q(Z) \log q(Z)$$

$$q(Z) = \prod_{i=1}^{n} q(z_i) \qquad \sum_{z_i} q(z_i) = 1$$

Fully Factorized Variational Family

$$\underset{\theta, q \in Q}{\text{argmax}} \sum_Z q(Z) \log p(X, Z|\theta) - \sum_Z q(Z) \log q(Z)$$

$$q(Z) = \prod_{i=1}^{n} q(z_i) \qquad \sum_{z_i} q(z_i) = 1$$

$$\underset{\theta, q \in Q}{\text{argmax}} \sum_{i=1}^{n} \sum_{z_i} q(z_i) \log p(x_i, z_i|\theta) - q(z_i) \log q(z_i)$$

Point Distributions

$$\underset{\theta, q \in Q}{\text{argmax}} \sum_Z q(Z) \log p(X, Z|\theta) - \sum_Z q(Z) \log q(Z)$$

$$q(Z) = \prod_{i=1}^{n} q(z_i) \qquad q(z_i) = \begin{cases} 1 & \text{if } z_i = \hat{z}_i \\ 0 & \text{otherwise} \end{cases}$$

$$\underset{\theta, q \in Q}{\text{argmax}} \sum_{i=1}^{n} \sum_{z_i} q(z_i) \log p(x_i, z_i|\theta) - q(z_i) \log q(z_i)$$

$$\underset{\theta, q \in Q, \hat{Z}}{\text{argmax}} \sum_{i=1}^{n} \log p(x_i, \hat{z}_i|\theta)$$

point distributions are often easier to compute, but less robust

Point Distributions for GMMs

$$\underset{\theta, q \in Q, \hat{Z}}{\text{argmax}} \sum_{i=1}^{n} \log p(x_i, \hat{z}_i | \theta)$$

$$\sum_{i=1}^{n} \log \mathcal{N}(x_i | \mu_{\hat{z}_i}, \Sigma_{\hat{z}_i})$$

$$\hat{z}_i \leftarrow \underset{z}{\text{argmax}} \ \log \mathcal{N}(x_i | \mu_z, \Sigma_z)$$

$$\mu_z \leftarrow \frac{\sum_{i;\hat{z}_i=z} x_i}{\sum_{i;\hat{z}_i=z} 1} \qquad \Sigma_z \leftarrow \frac{\sum_{i;\hat{z}_i=z} (x_i - \mu_i)(x_i - \mu_i)^\top}{\sum_{i;\hat{z}_i=z} 1}$$

K-means

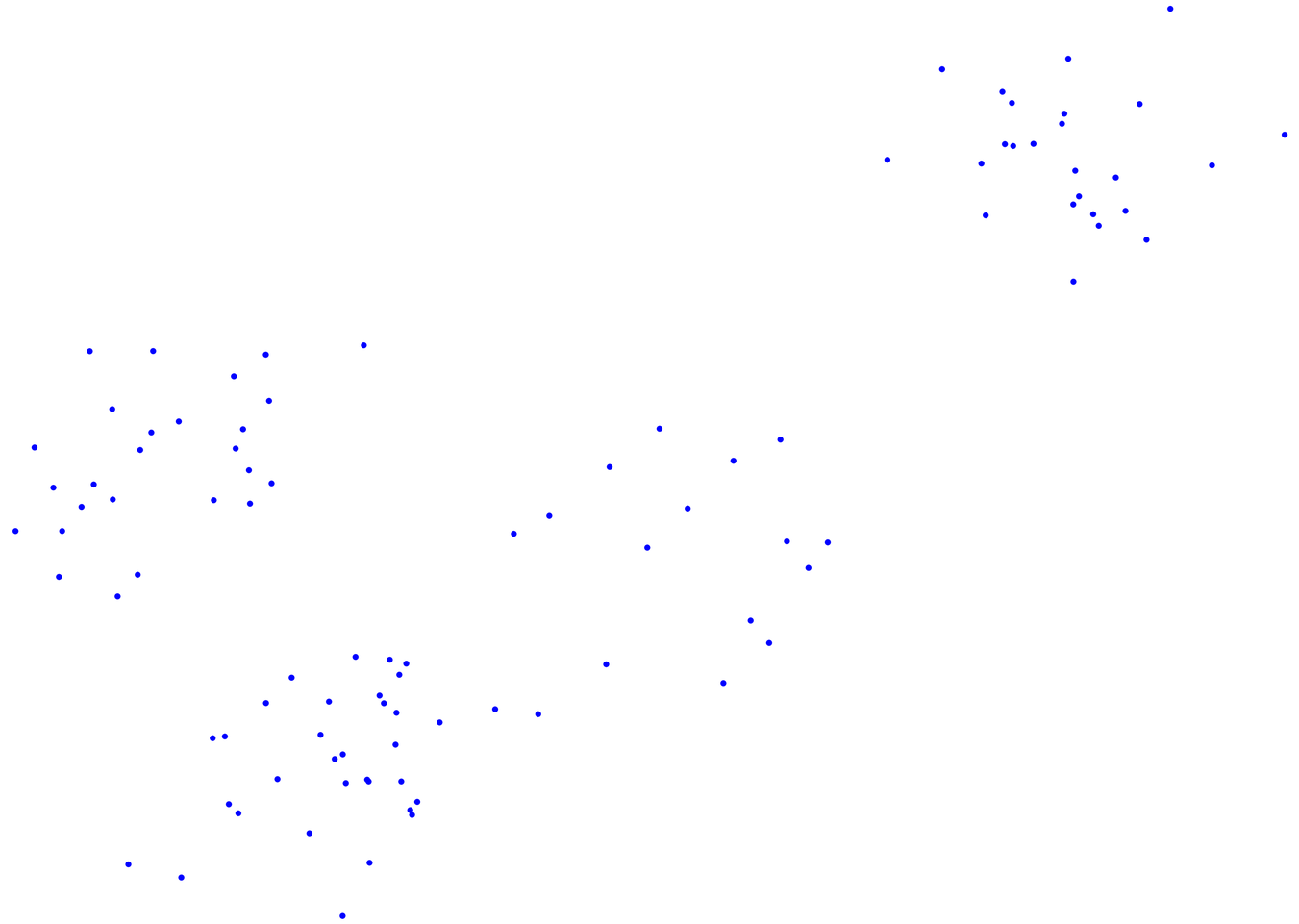
$$\hat{z}_i \leftarrow \underset{z}{\text{argmin}} \, ||x_i - \mu_z||$$

assign points to closest mean

$$\mu_z \leftarrow \frac{\sum_{i;\hat{z}_i=z} x_i}{\sum_{i;\hat{z}_i=z} 1}$$

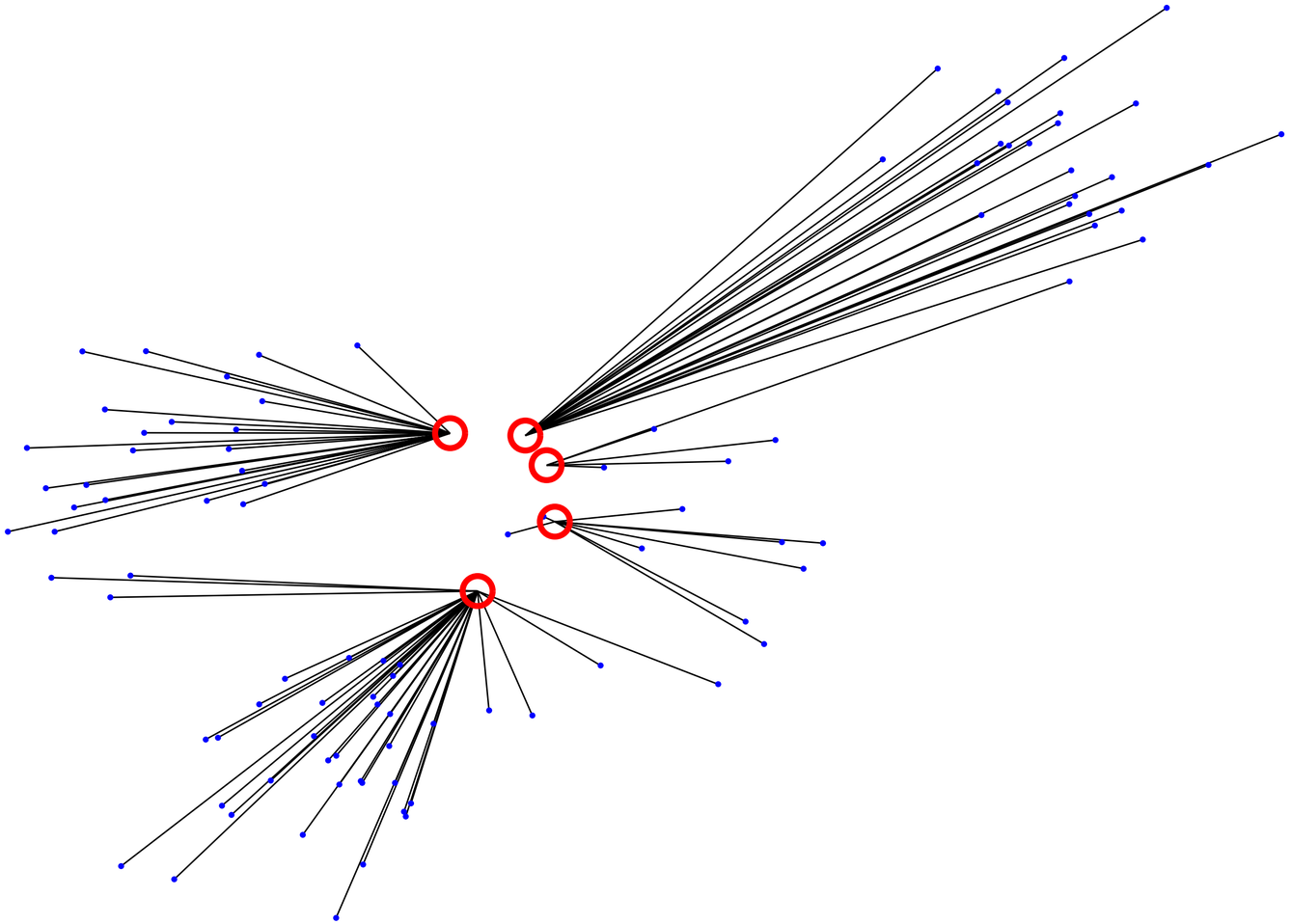
set means to average of points in cluster

Example



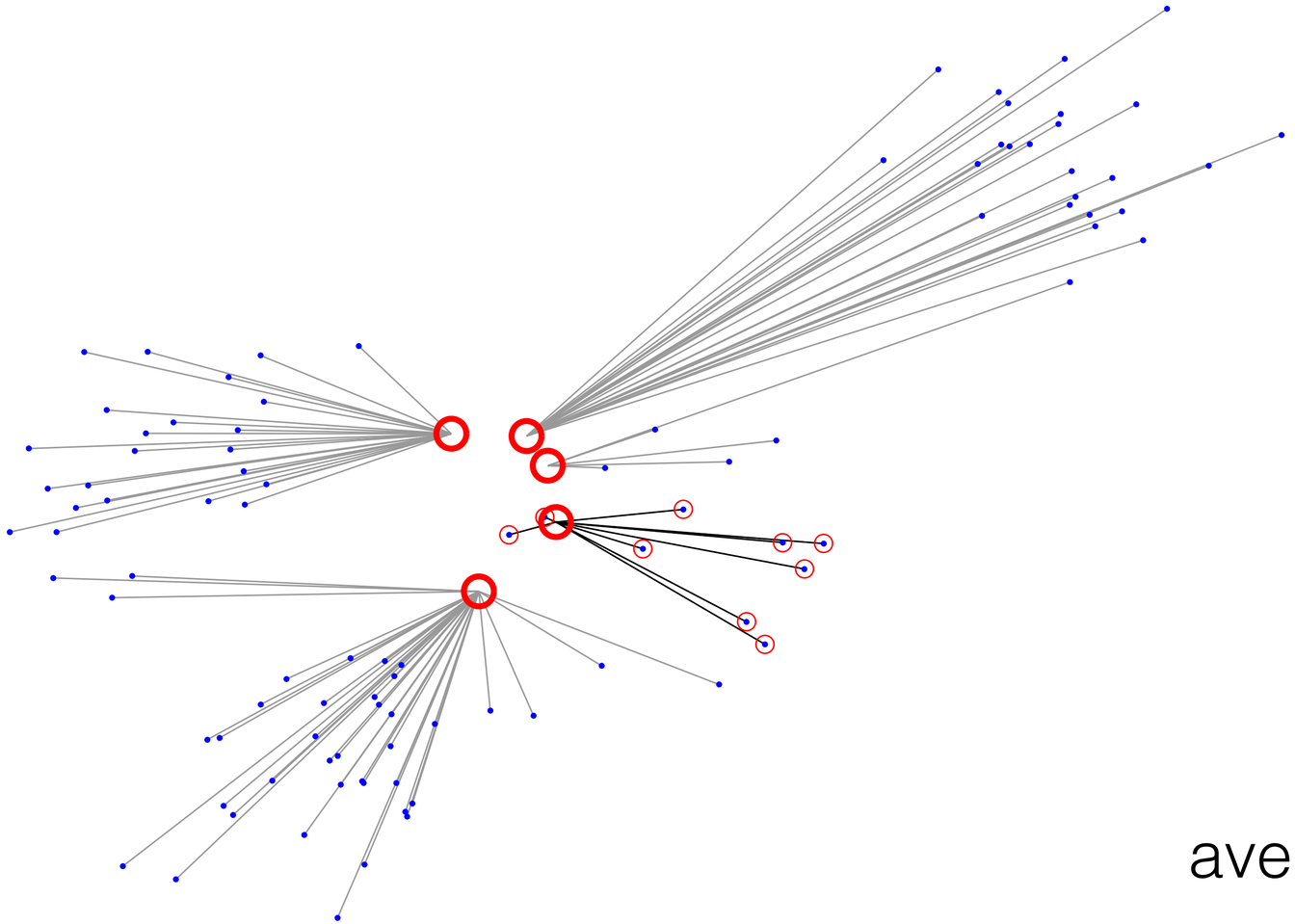
input data

Example



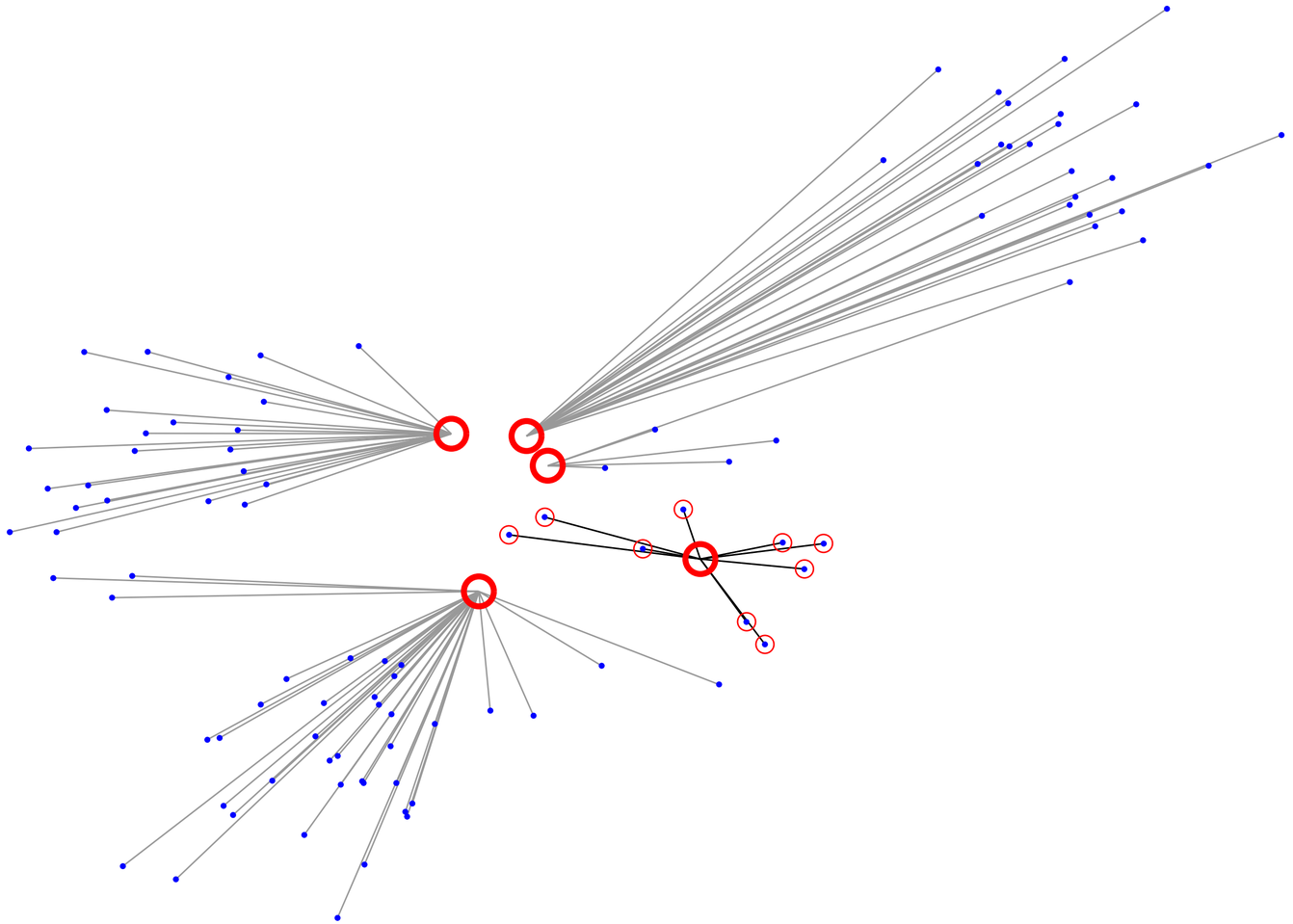
assign points to
initialized means

Example



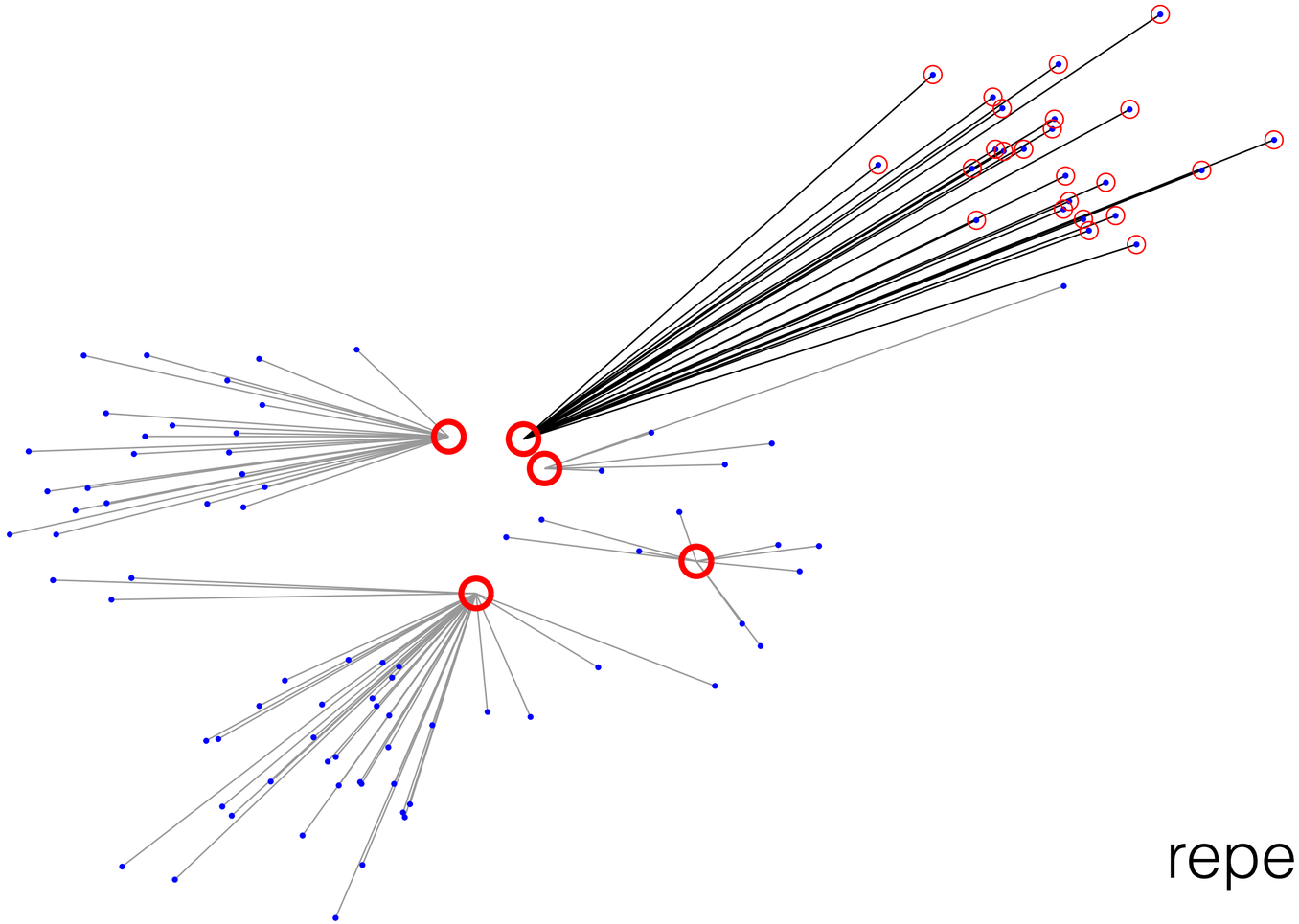
average each cluster

Example



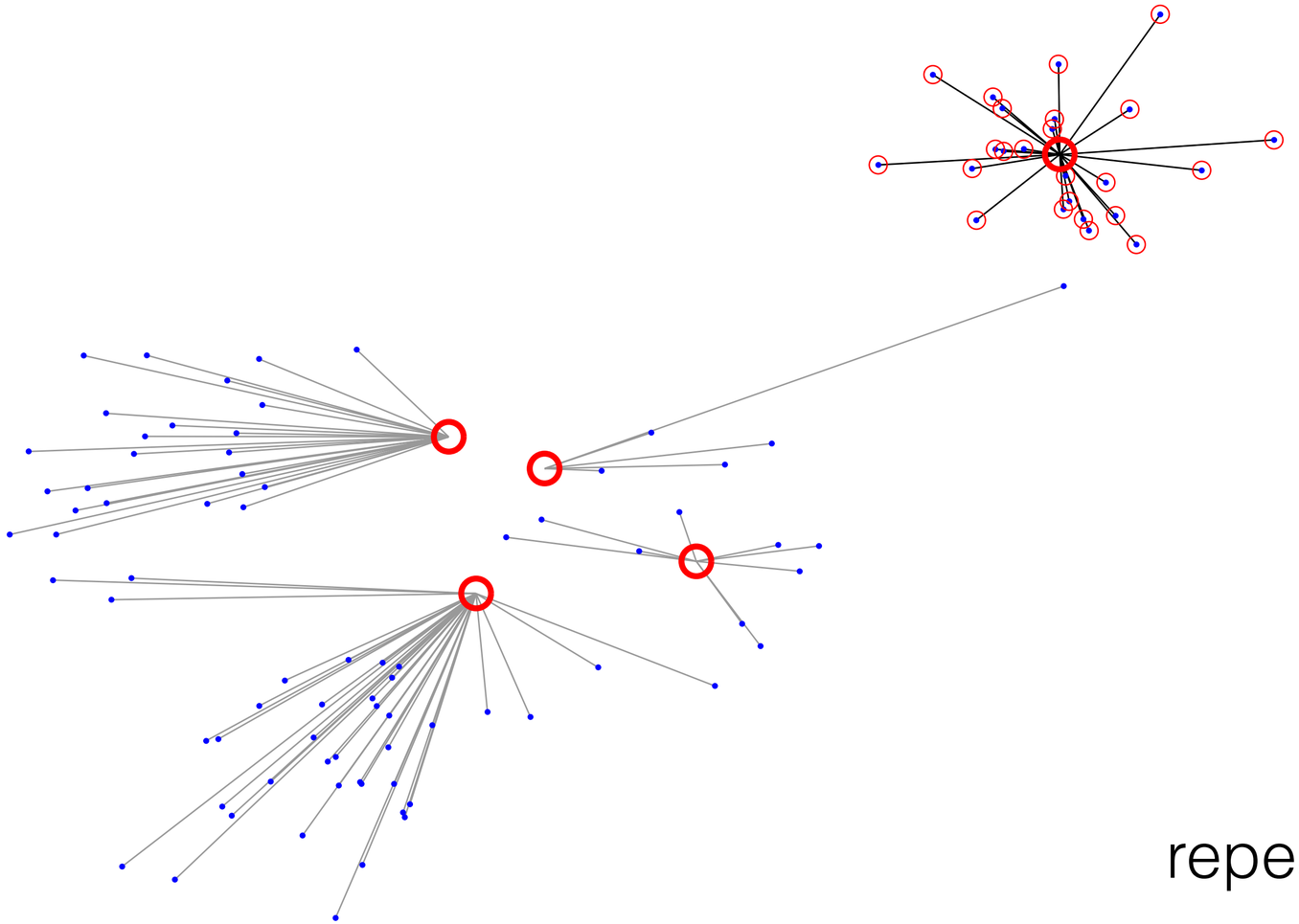
update mean

Example



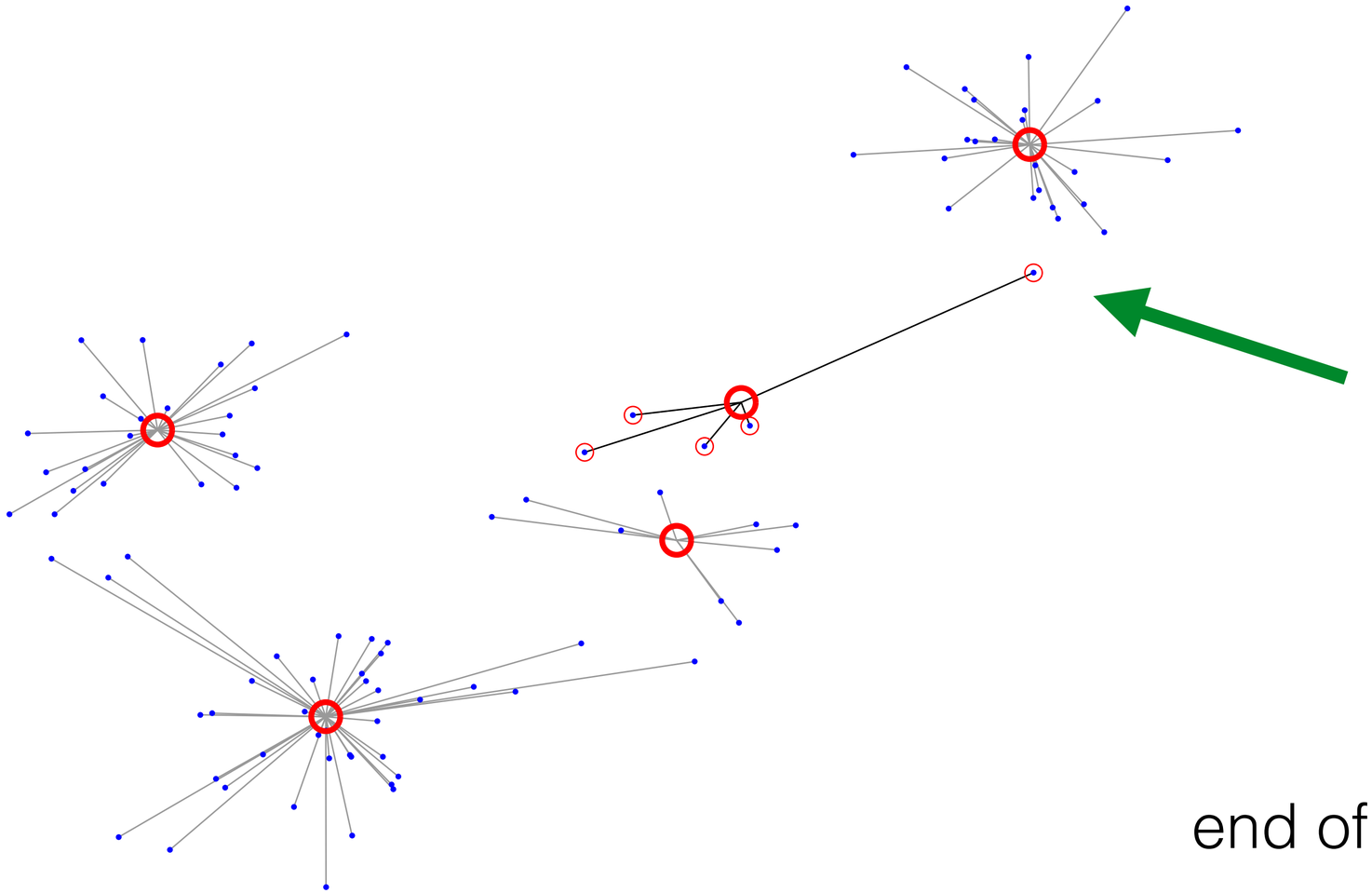
repeat for all clusters

Example



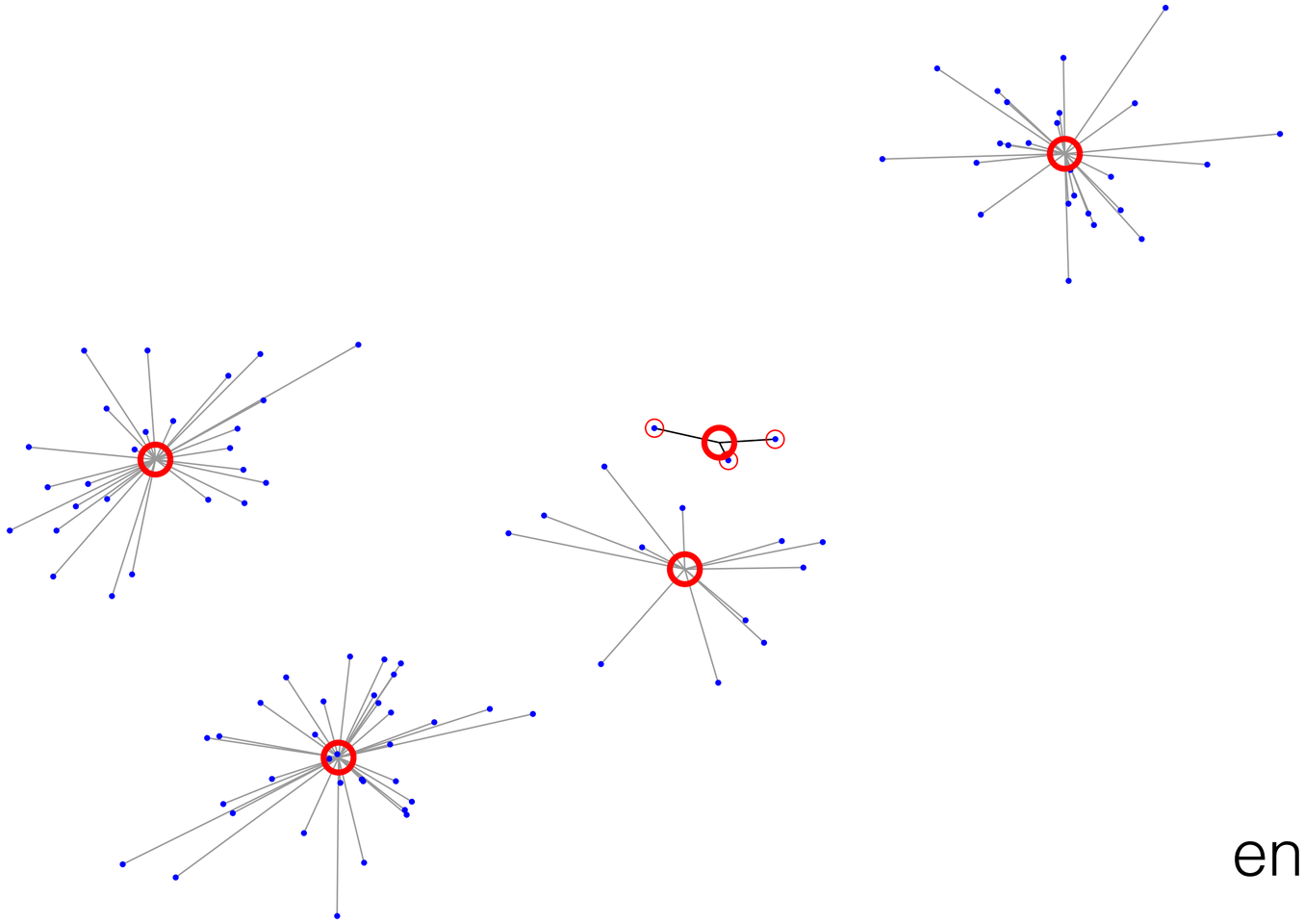
repeat for all clusters

Example



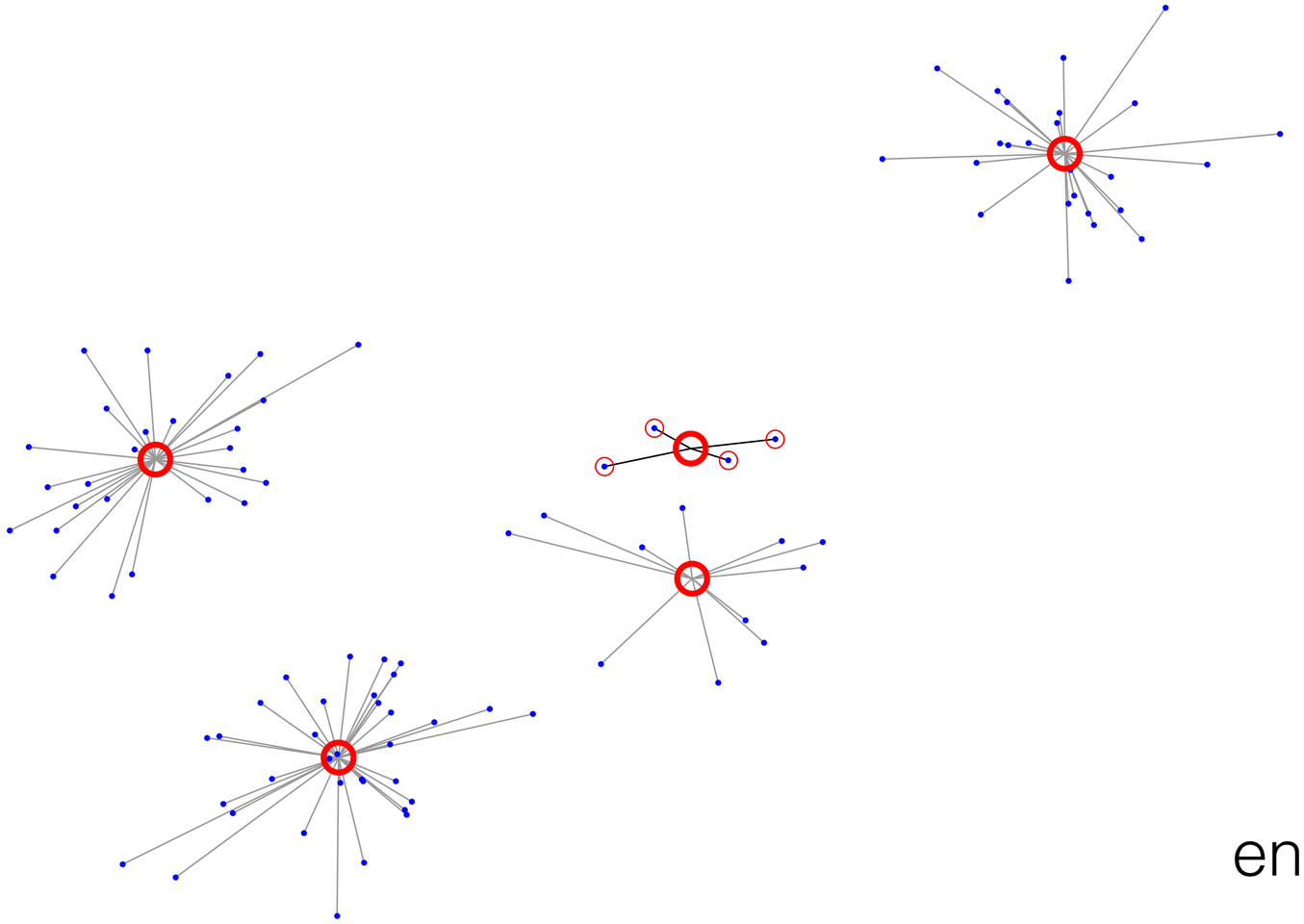
end of iteration 1

Example



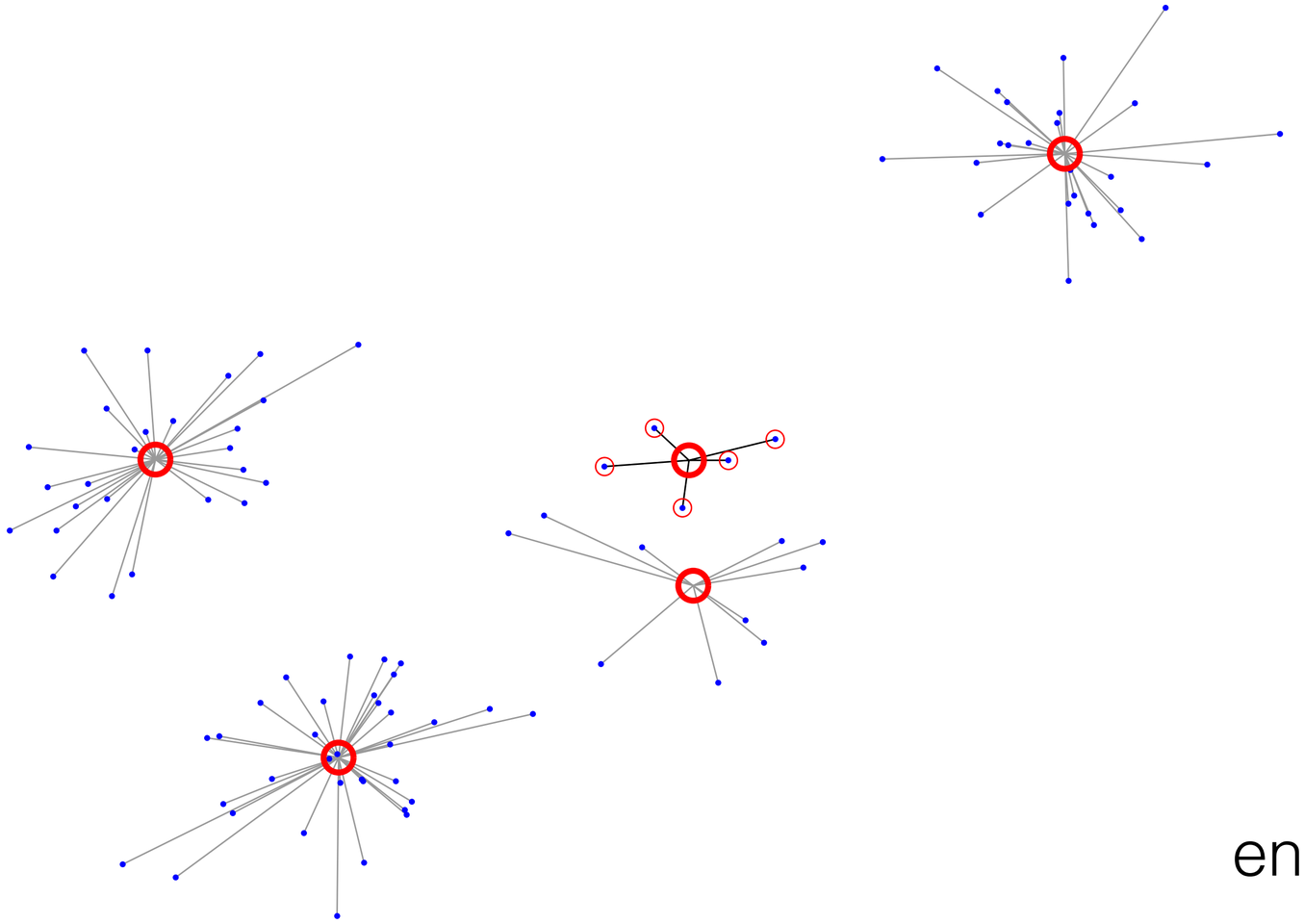
end of iteration 2

Example



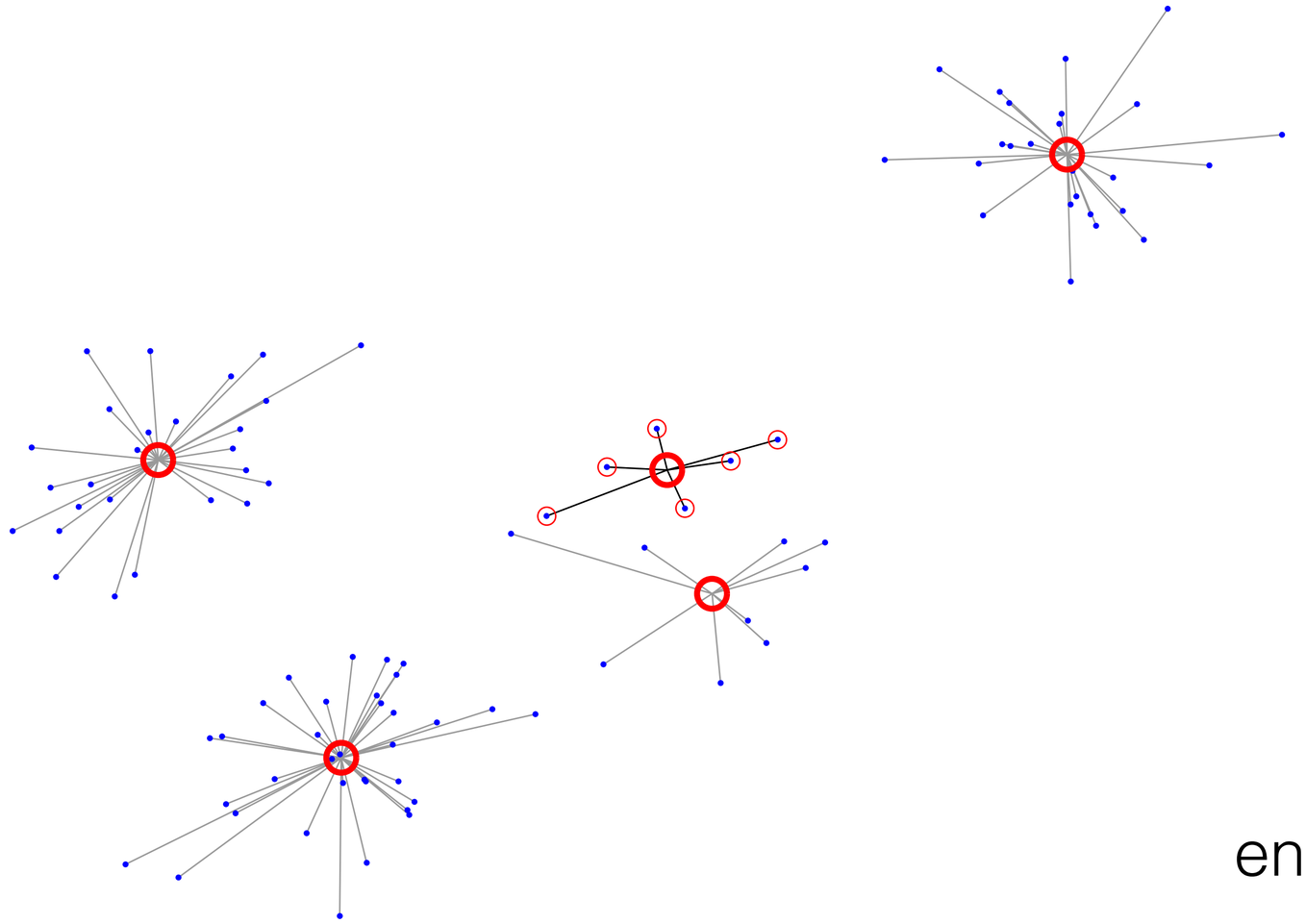
end of iteration 3

Example



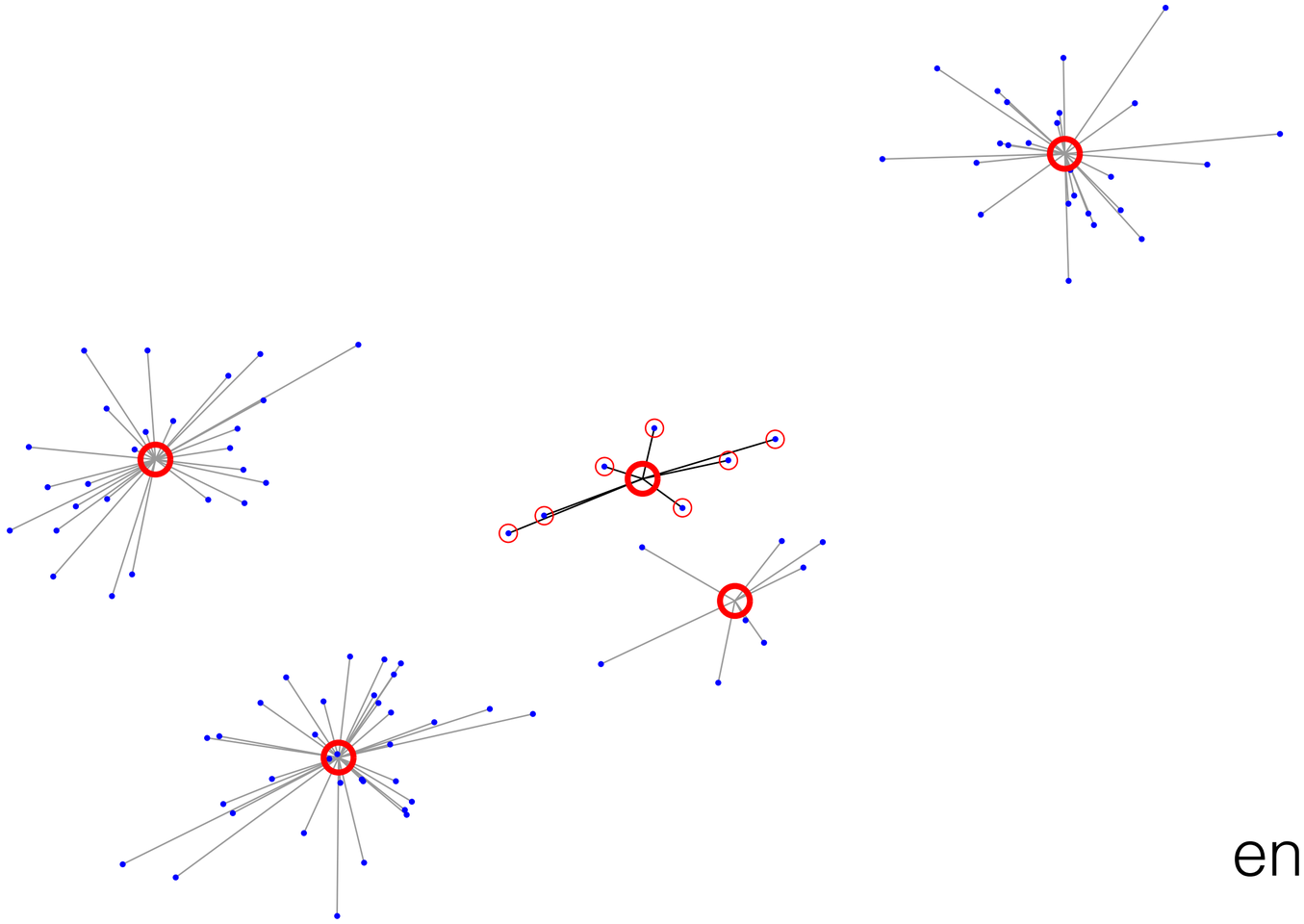
end of iteration 4

Example



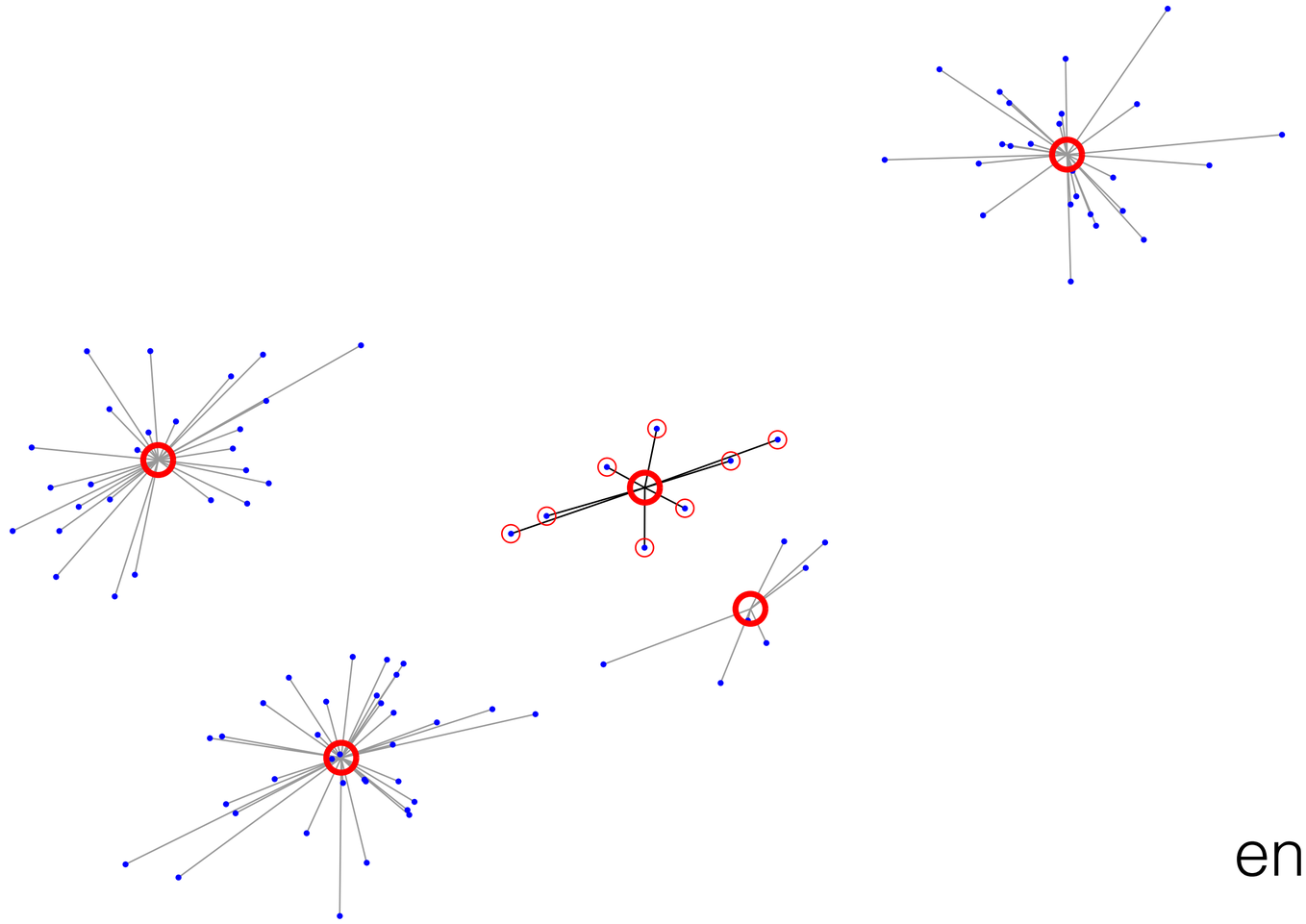
end of iteration 5

Example



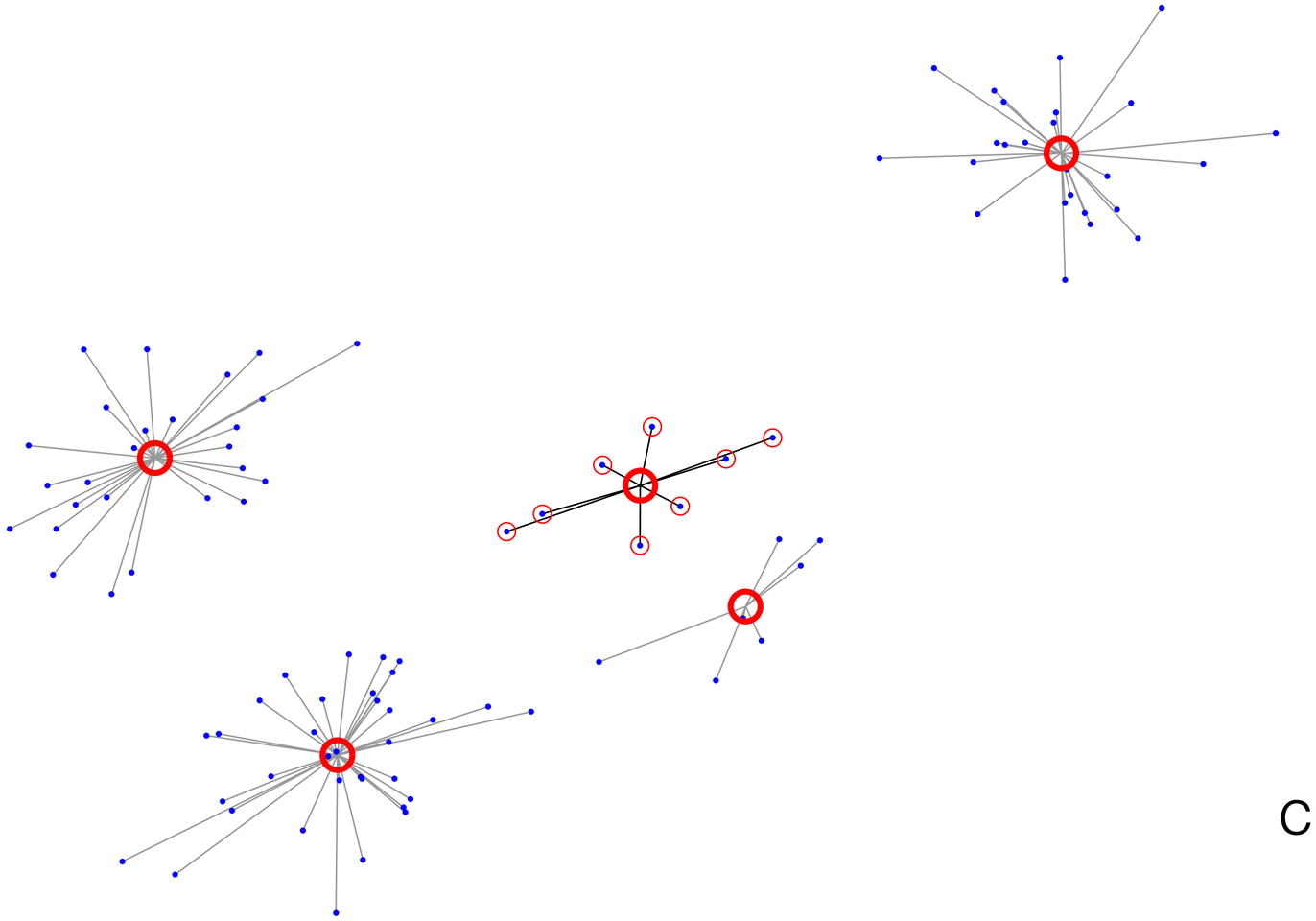
end of iteration 6

Example



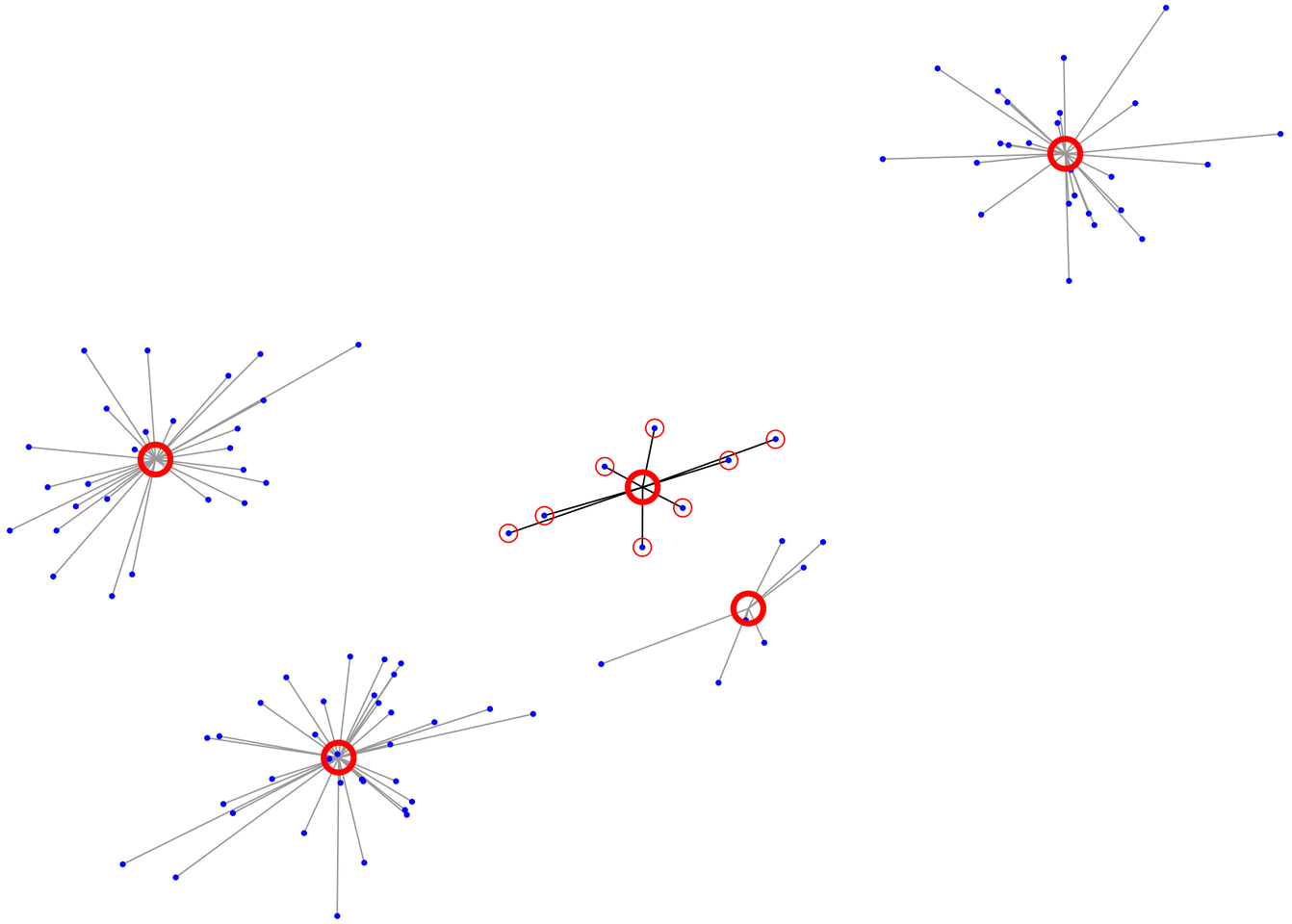
end of iteration 7

Example



converged at
iteration 8

Example



Summary

- Used Jensen's inequality to derive lower bound on log marginal likelihood
- Bound uses variational distribution **q**. We get to choose what family of **q** distributions to consider
- Using fully-factorized multinomial distributions for **q** gets EM
- Fully-factorized point distributions gets "hard"-EM, and using fixed, spherical covariance gets K-means