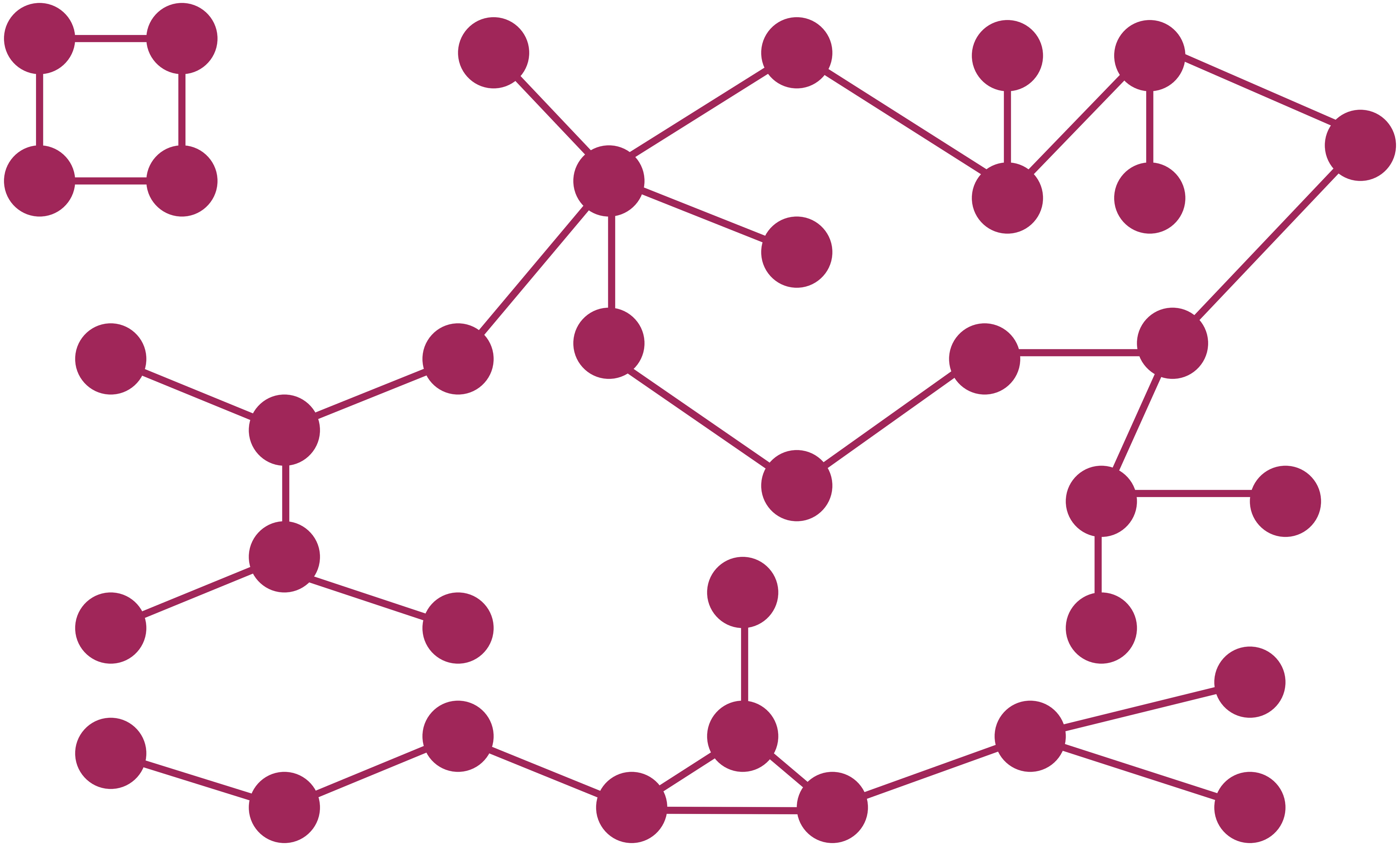
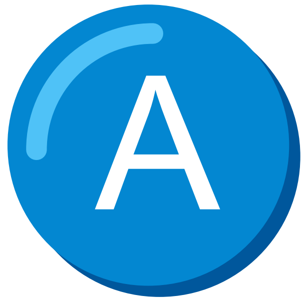
Connected Components

A **connected component** of an undirected graph is a subgraph with any two vertices connected to each other by paths, whereas the subgraph itself is connected to no additional vertices in the supergraph.

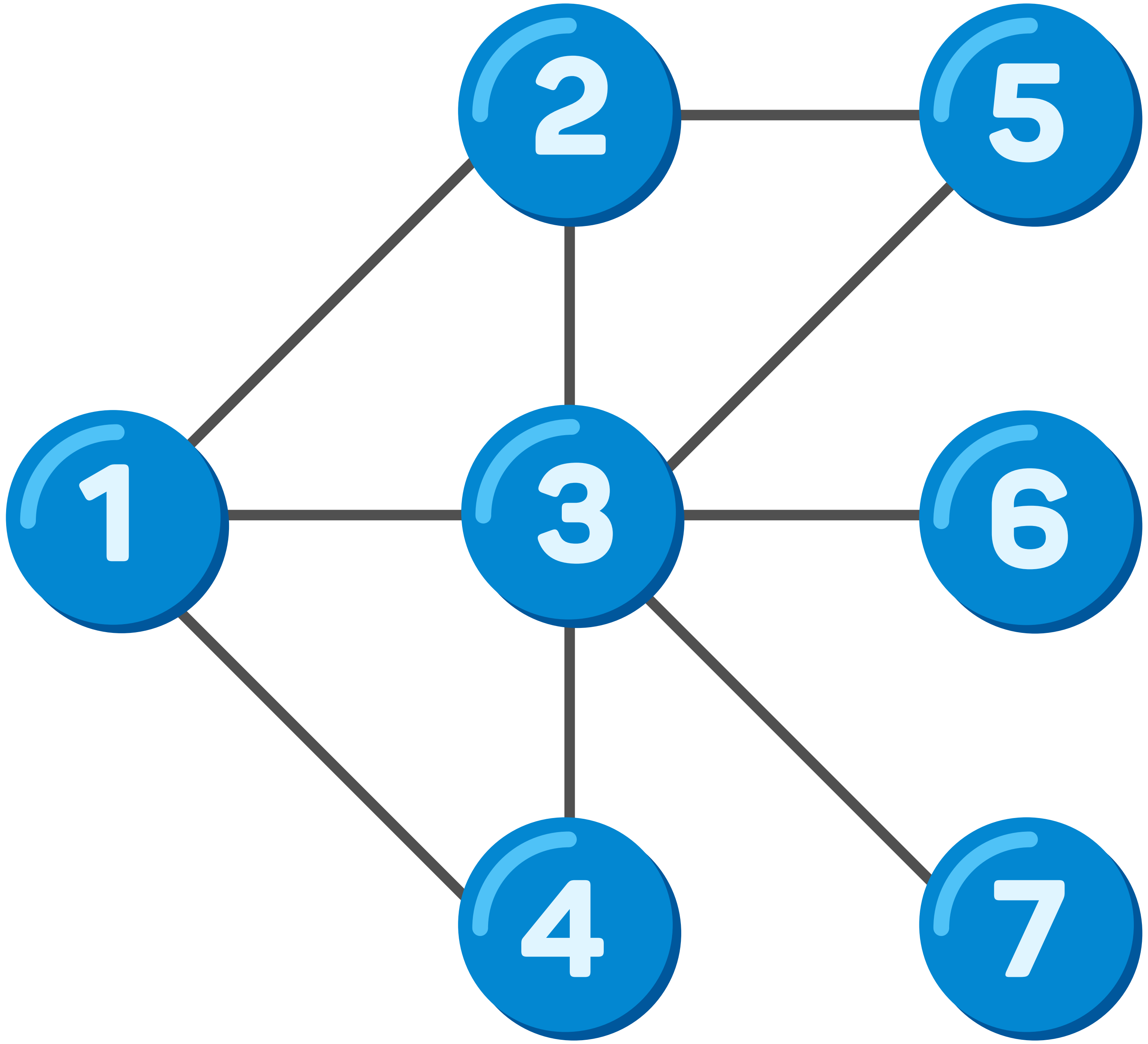
3 Connected Components



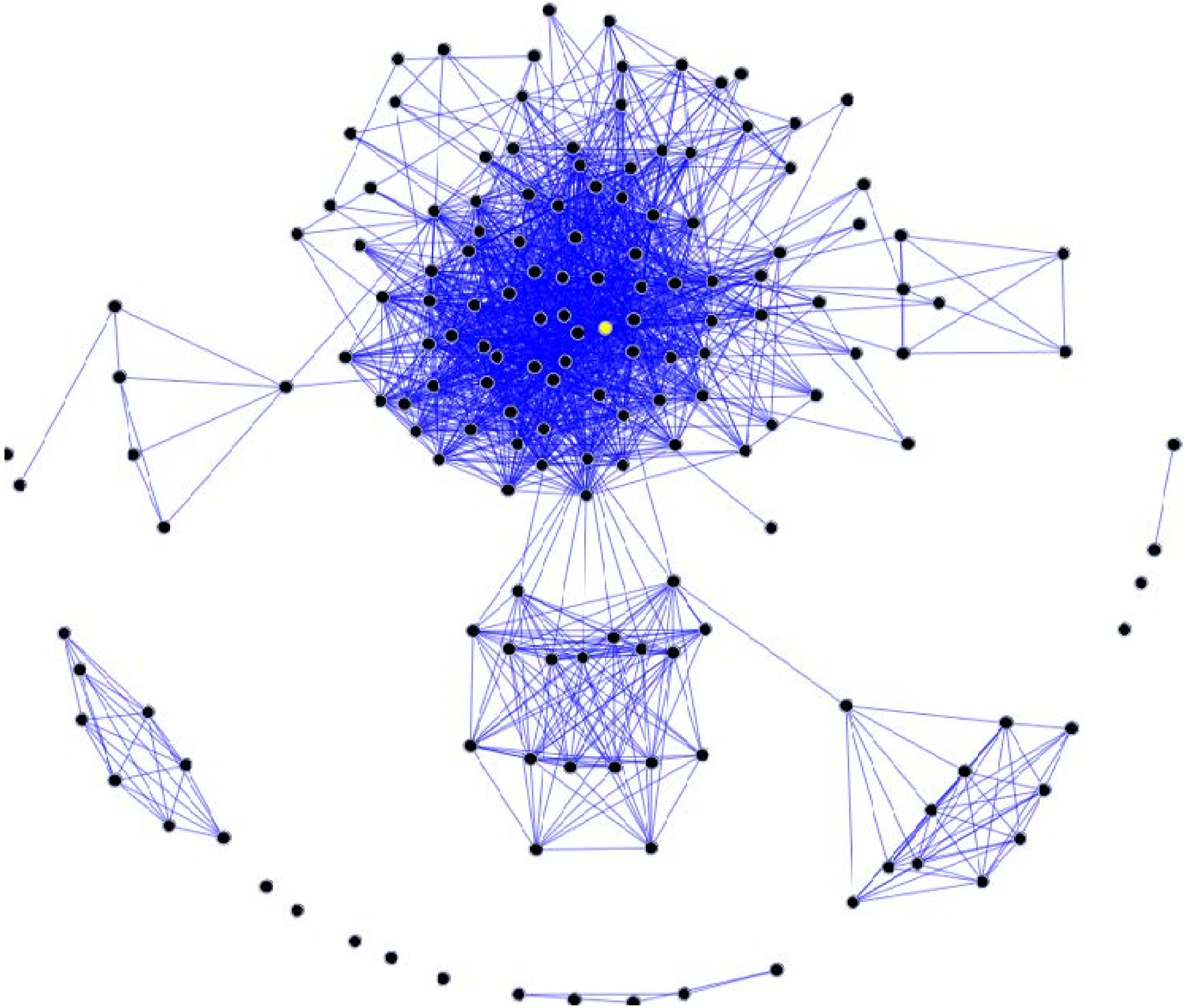
A vertex with no incident edges



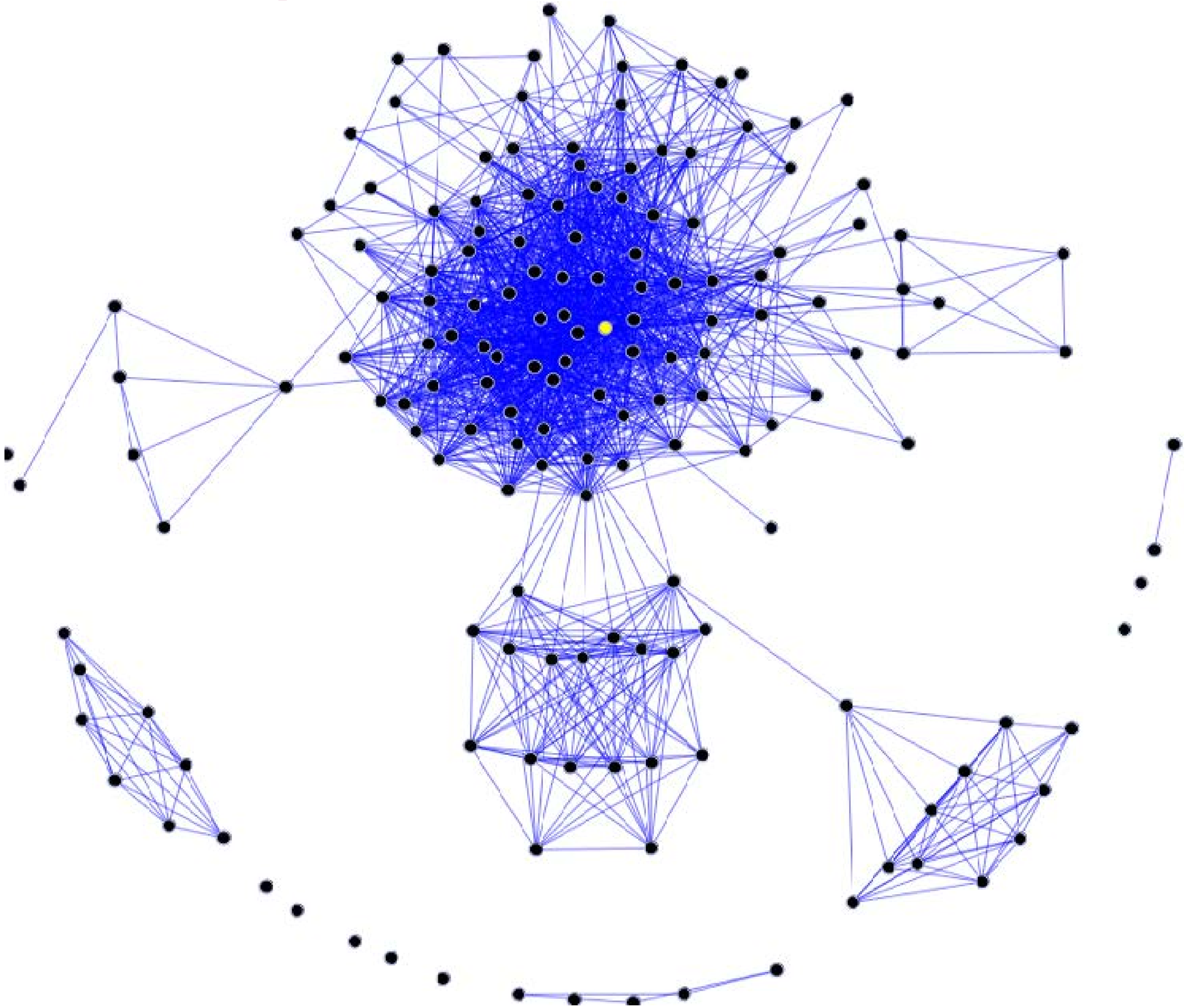
Mini social graph - 1 connected component



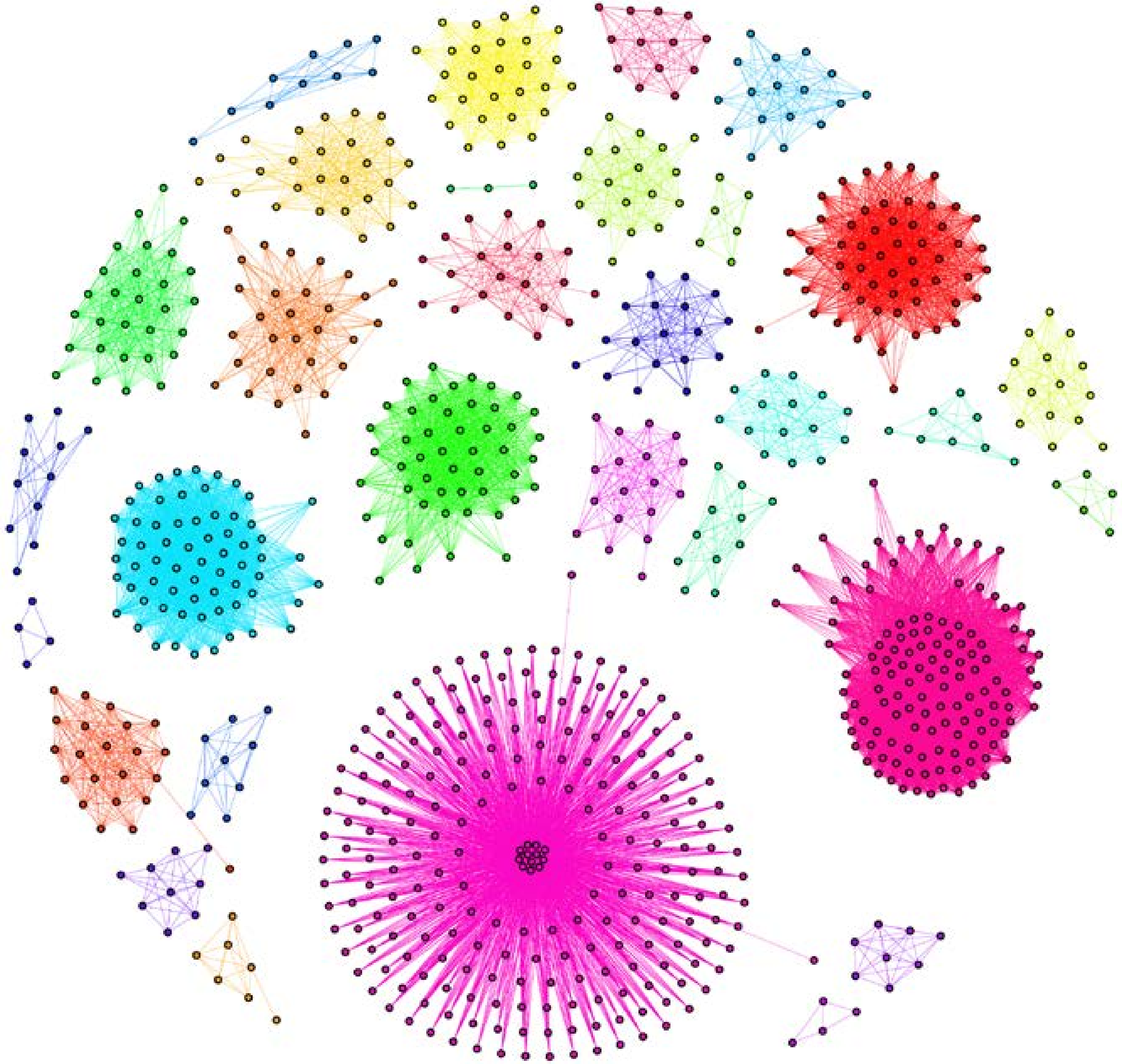
Is graph connected?

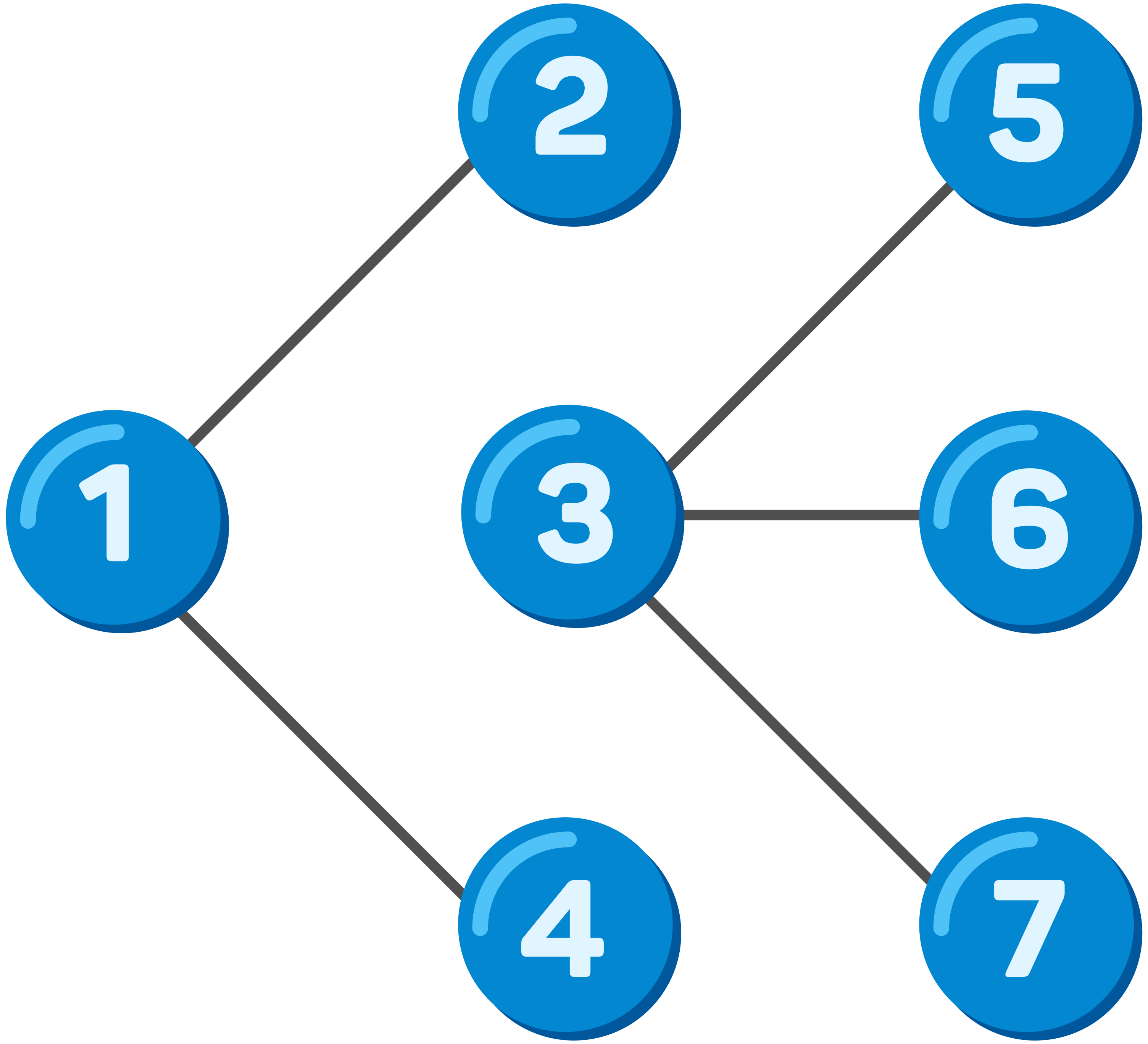


Graph description



Graph clustering




1
2
3
4
5
6
7

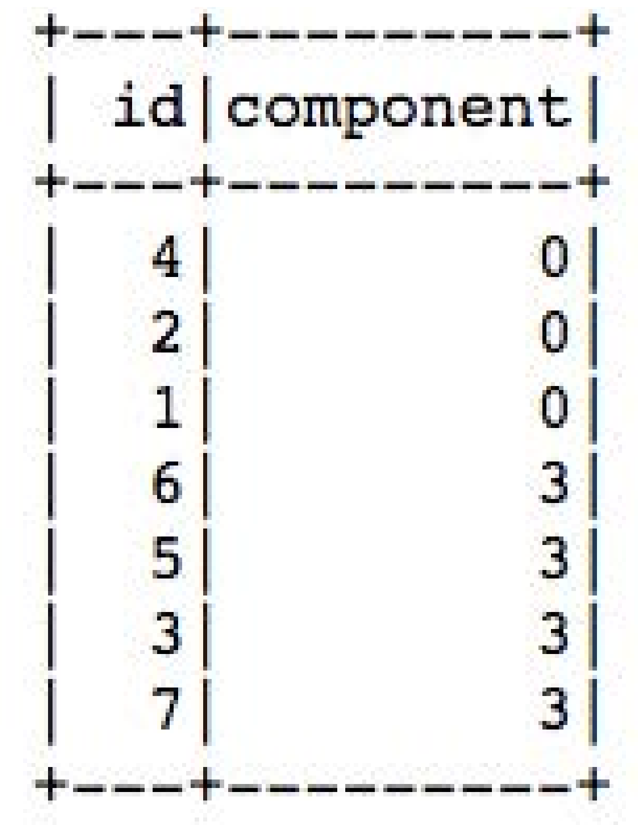
```
from graphframes import *

vertices = sparkSession.createDataFrame([
   ("1"."Alex", 28, "M","MIPT"),
   ("2","Emeli", 28, "F","MIPT"),
   ("3","Natasha", 27, "F","SPbSU"),
   ("4","Pavel", 30, "M","MIPT"),
   ("5","Oleg", 35, "M","MIPT"),
   ("6","Ivan", 30, "M","MSU"),
   ("7","Ilya", 29, "M","MSU")], ["id","name","age","gender","university"])

edges = sparkSession.createDataFrame([
   ("1","2","friend")
   ("1","4","friend")
   ("3","5","friend")
   ("3","6","friend")
   ("3","7","friend")
], ["src", "dst" , "type"])

g = GraphFrame(vertices, edges)
```

```
result = g.connectedComponents()
result.select("id", "component").orderBy("component").show()
```



```
+---+---------+
| id|component|
+---+---------+
|  4|        0|
|  2|        0|
|  1|        0|
|  6|        3|
|  5|        3|
|  3|        3|
|  7|        3|
+---+---------+
```

Algorithm

“graphframes” - “Connected Components in MapReduce and Beyond” by Raimodas Kiveris et al.

“graphx” - GraphX

Checkpoint interval - number of iterations of connected components algorithm:

- helps recover from failures

Checkpoint interval - number of iterations of connected components algorithm:

- helps recover from failures
- clean shuffle files

Checkpoint interval - number of iterations of connected components algorithm:

- helps recover from failures
- clean shuffle files
- shorten the lineage of the computation graph

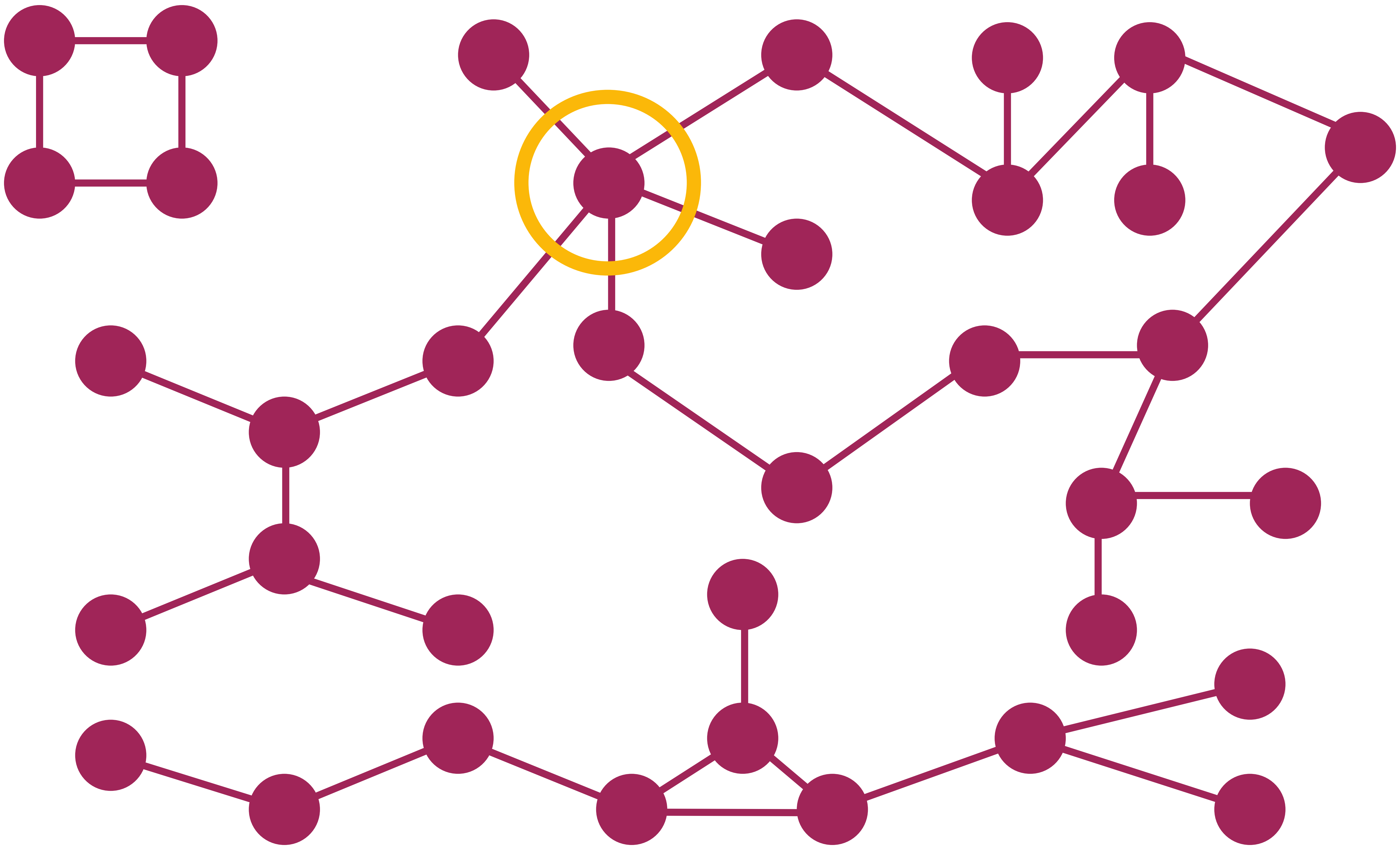
Checkpoint interval - number of iterations of connected components algorithm:

- helps recover from failures
- clean shuffle files
- shorten the lineage of the computation graph
- reduce the complexity of plan optimization

Checkpoint data:

- saved under `org.apache.spark. SparkContext.getCheckpointDir` with prefix "connected-components".
- If the checkpoint directory is not set, this throws a `java.io.IOException`.
- a nonpositive value to disable checkpointing.

Broadcast threshold



Summary

- What is the connected component of the graph

Summary

- What is the connected component of the graph
- How to find all connected components of the graph using GraphFrames API