GraphFrames intro

GraphFrames:

1. A graph processing library for Apache Spark

GraphFrames:

1. A graph processing library for Apache Spark
2. API available from Scala, Java and Python

GraphFrames:

1. A graph processing library for Apache Spark
2. API available from Scala, Java and Python
3. Are built on top of Spark DataFrames:

GraphFrames:

1. A graph processing library for Apache Spark
2. API available from Scala, Java and Python
3. Are built on top of Spark DataFrames:
   - powerful queries

GraphFrames:

1. A graph processing library for Apache Spark
2. API available from Scala, Java and Python
3. Are built on top of Spark DataFrames:
   - powerful queries
   - saving & loading graphs

Creating GraphFrames

From vertex and edge DataFrames

Creating GraphFrames

From vertex and edge DataFrames

- a vertex DataFrame should contain a special column named "id"

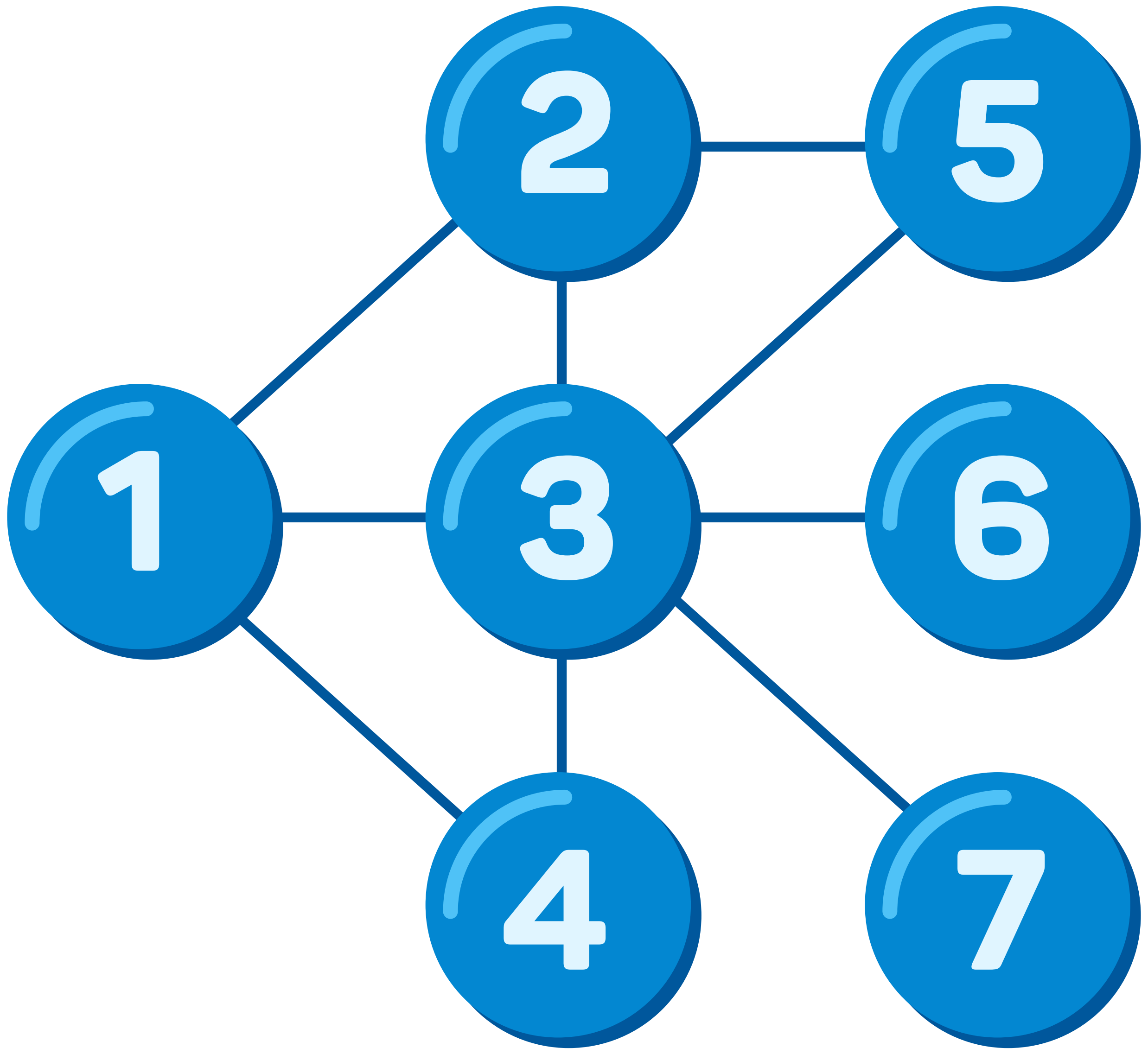
Creating GraphFrames

From vertex and edge DataFrames

- a vertex DataFrame should contain a special column named "id"

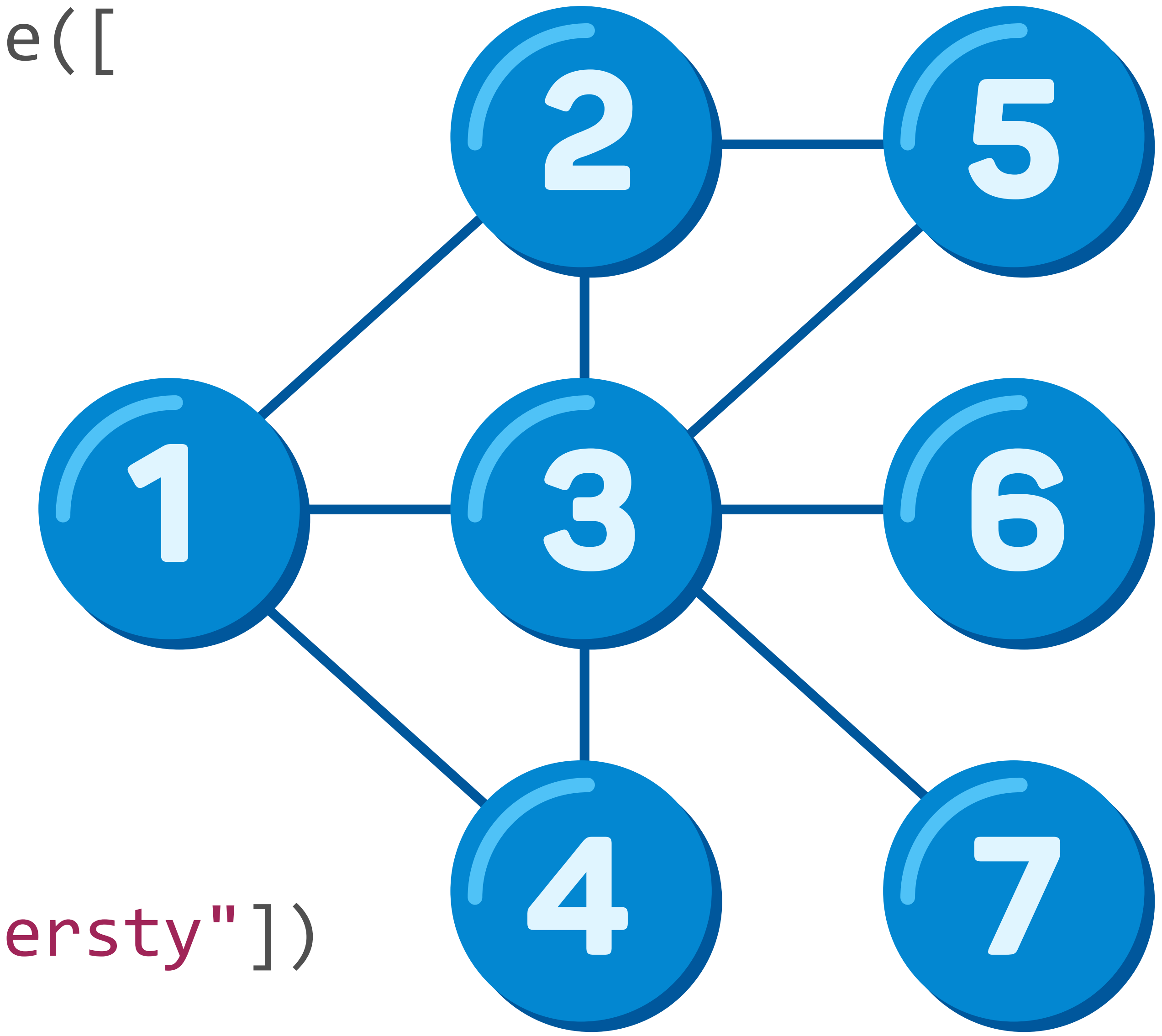
- an edge DataFrame should contain two special columns: "src" and "dst"

Mini social graph



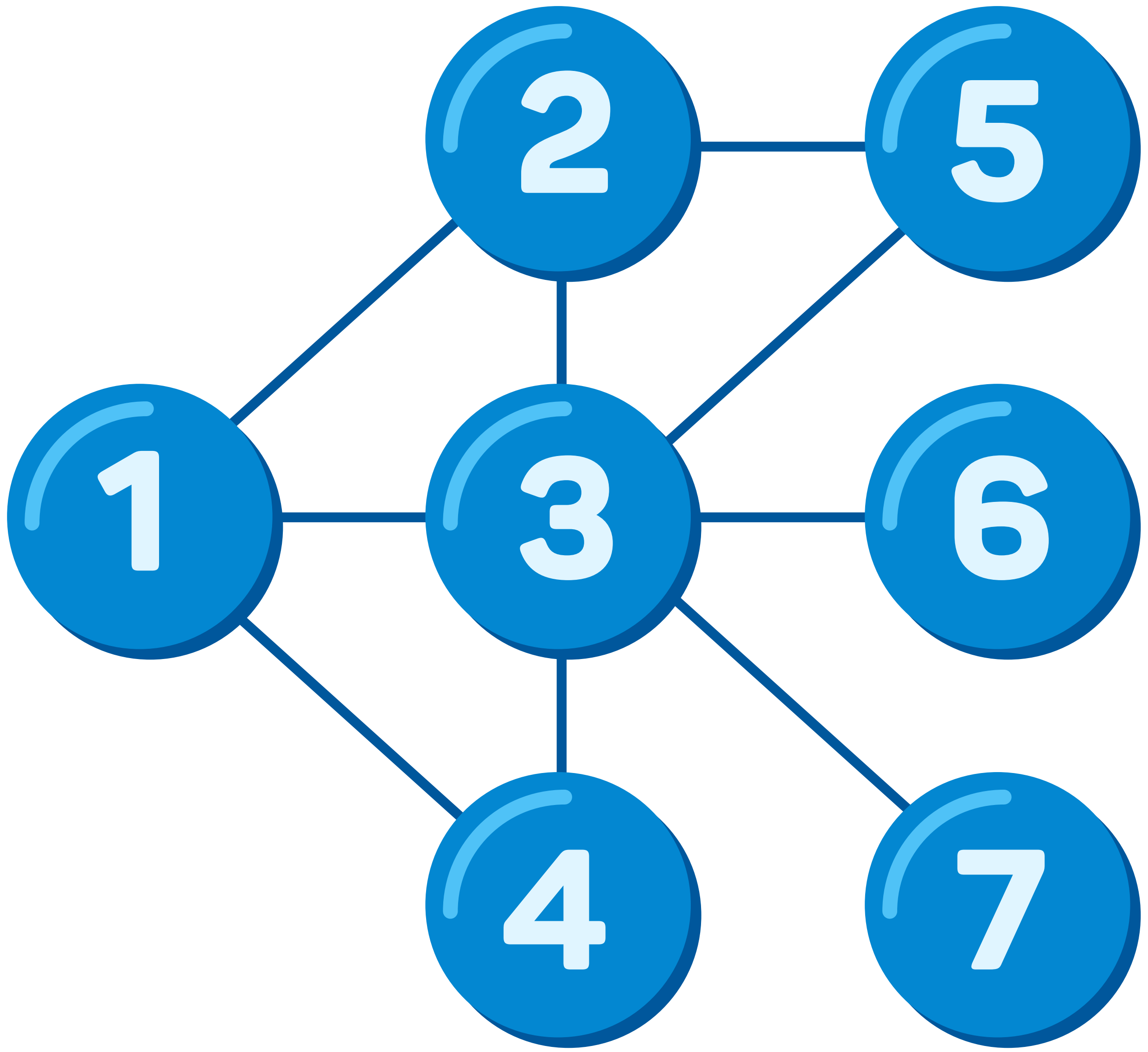
Mini social graph

```
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","universty"])
```



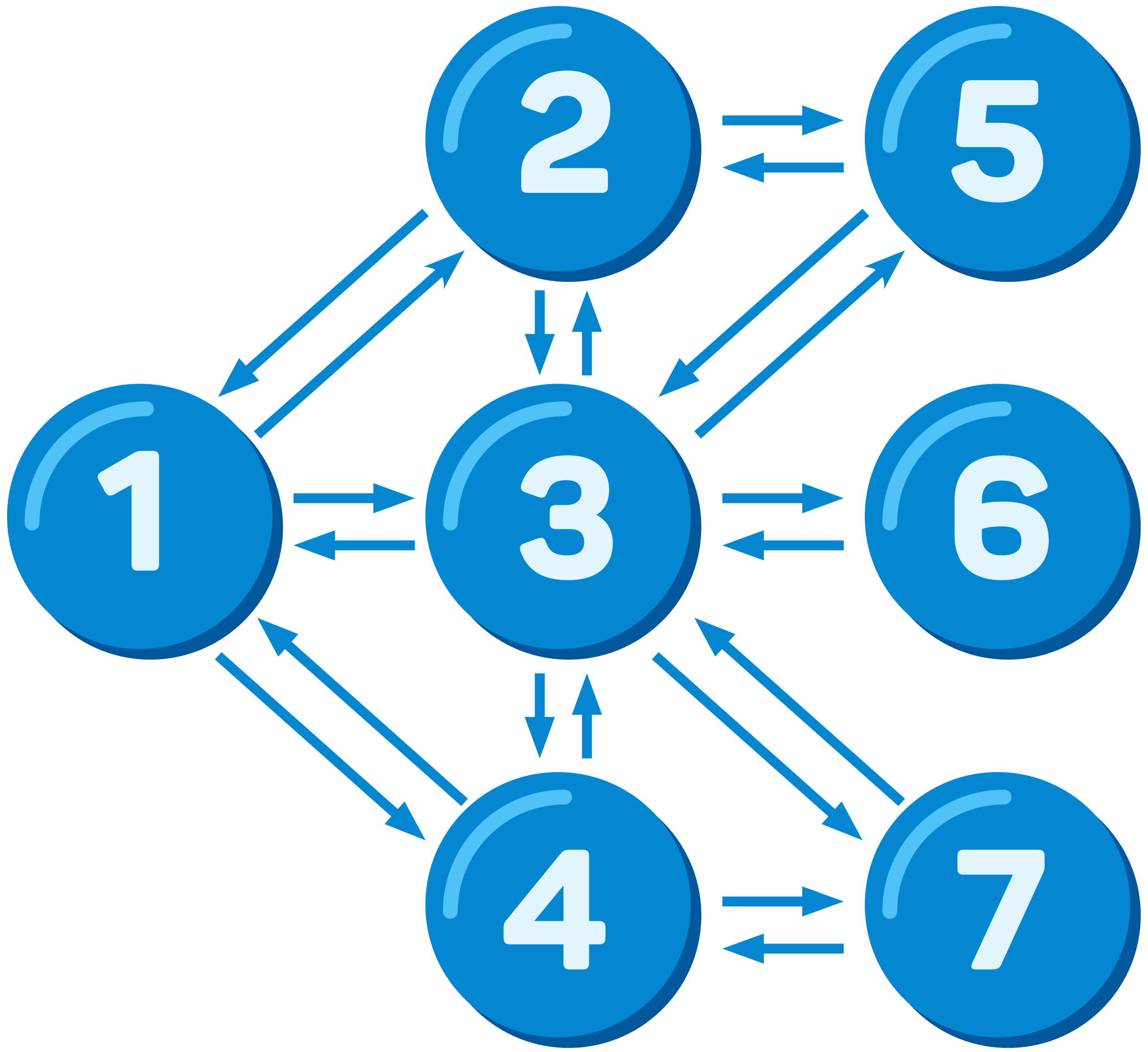
Mini social graph

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

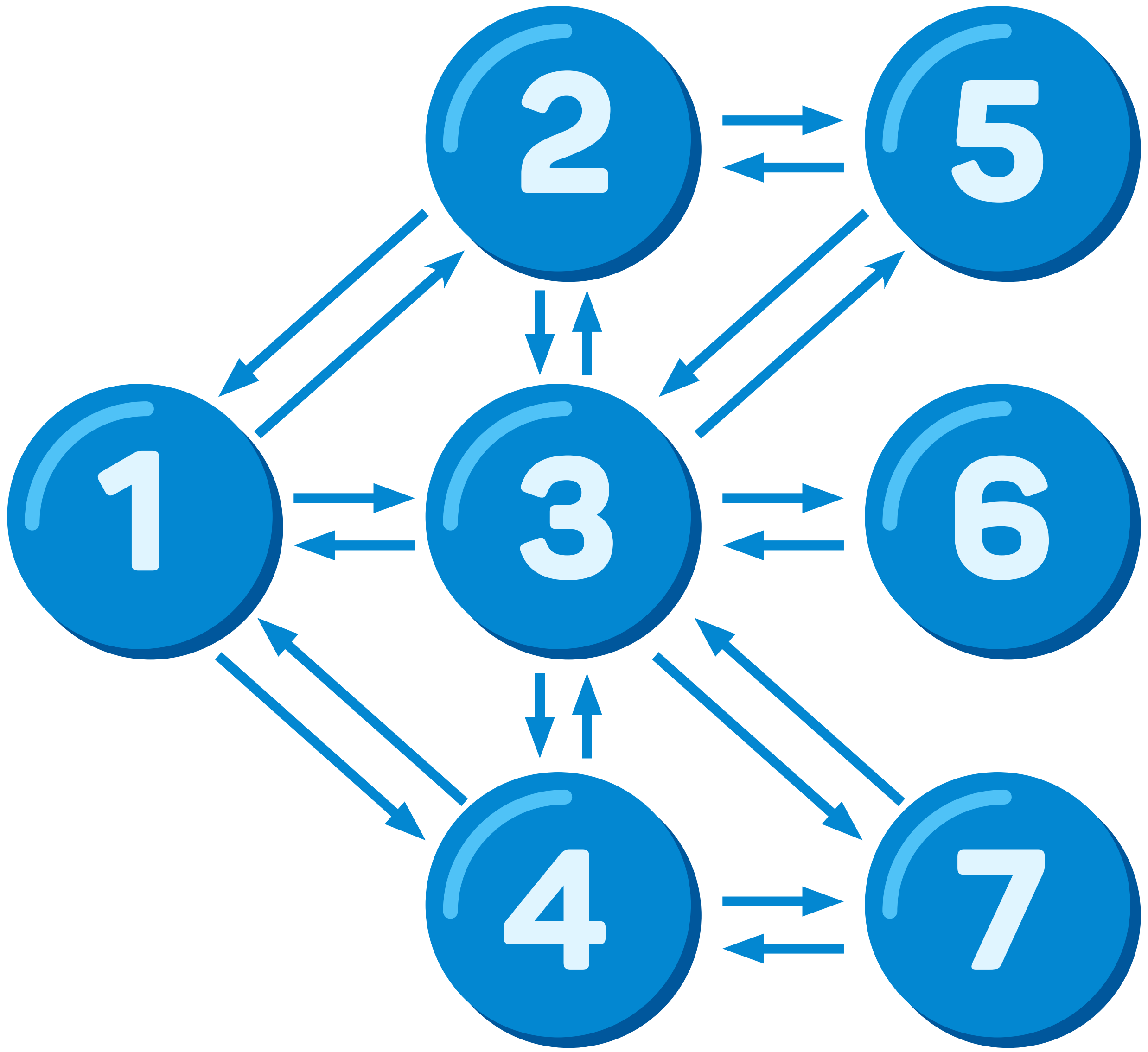


Mini social graph

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



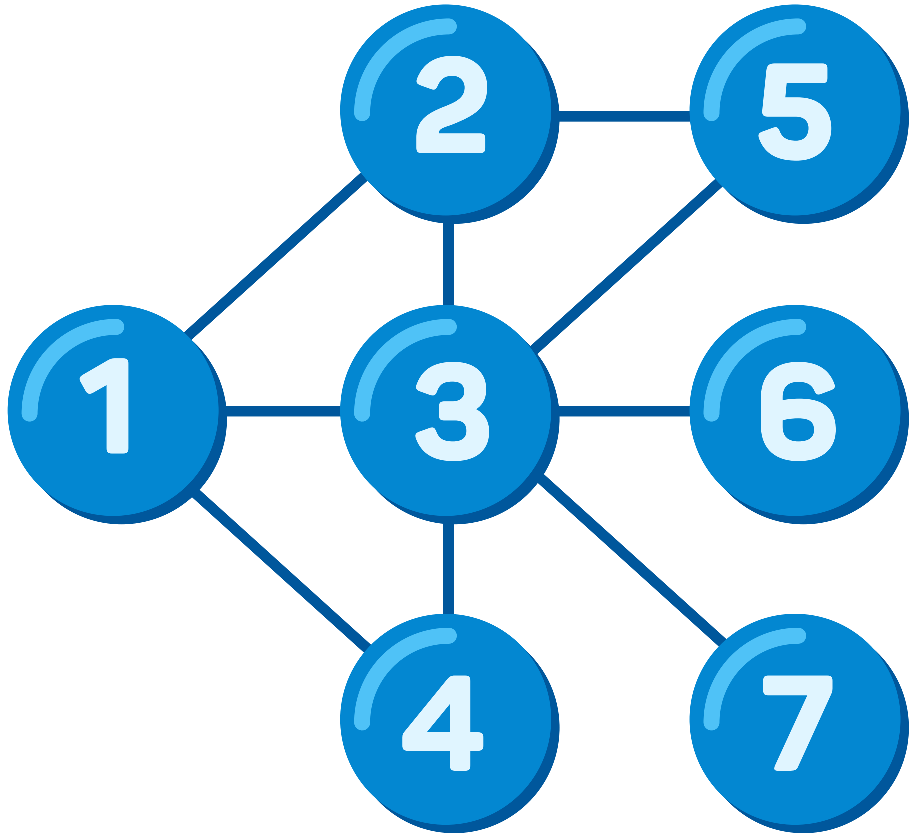
Mini social graph



```
g = GraphFrame(vertices, edges)
```

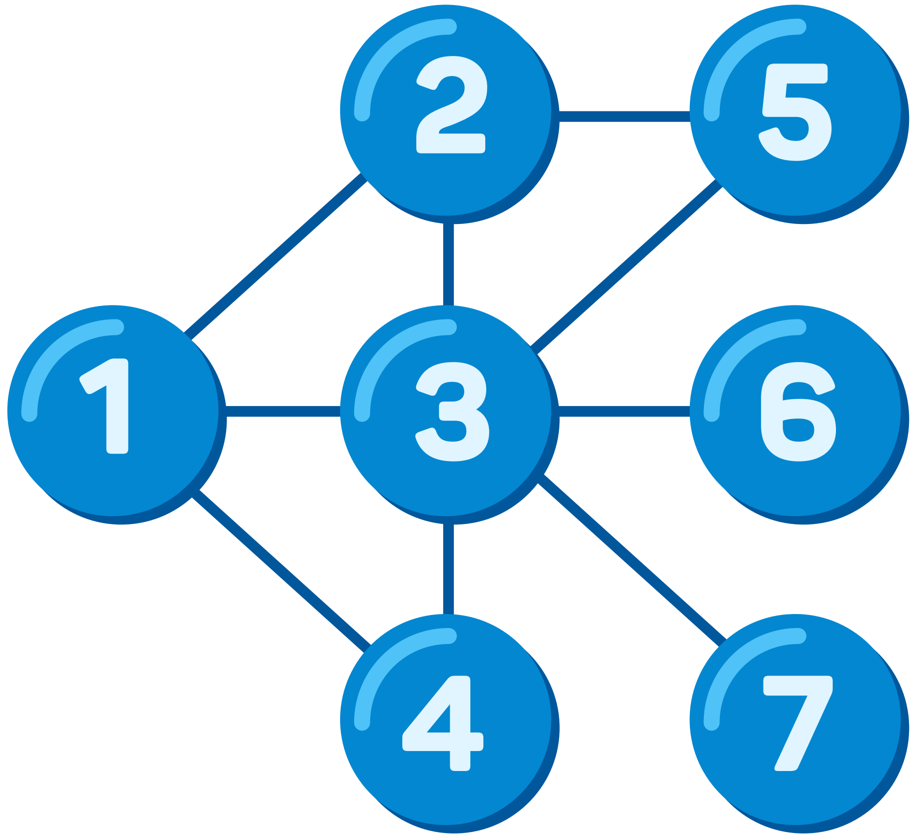
Basic graph and DataFrame queries

Basic graph and DataFrame queries



How many users in our mini social network have "age" > 30?

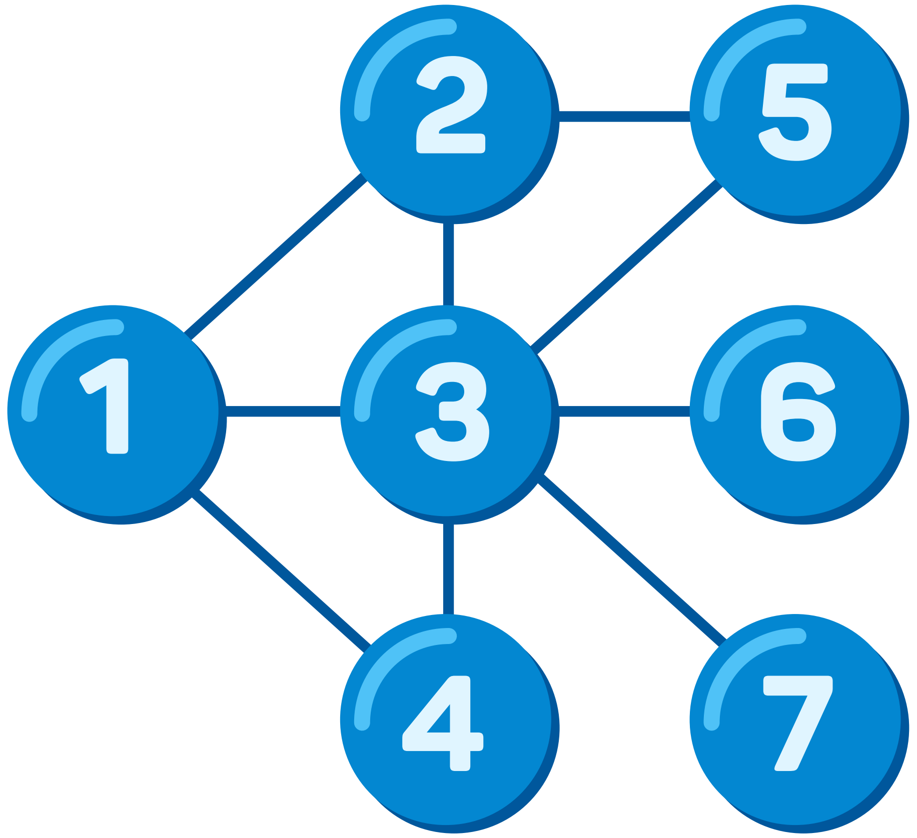
Basic graph and DataFrame queries



How many users in our mini social network have "age" > 30?

How many users have at least 2 friends?

Basic graph and DataFrame queries

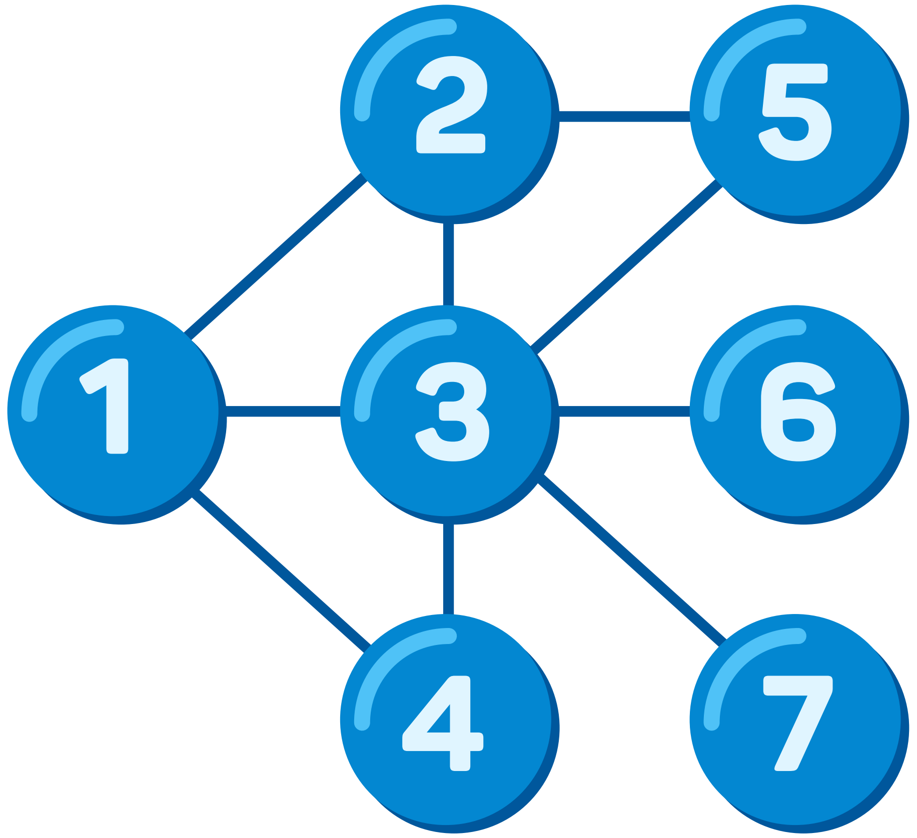


Example:

How many users in our mini social network have "age" > 30?

```
g.vertices.filter("age > 30")
```

Basic graph and DataFrame queries

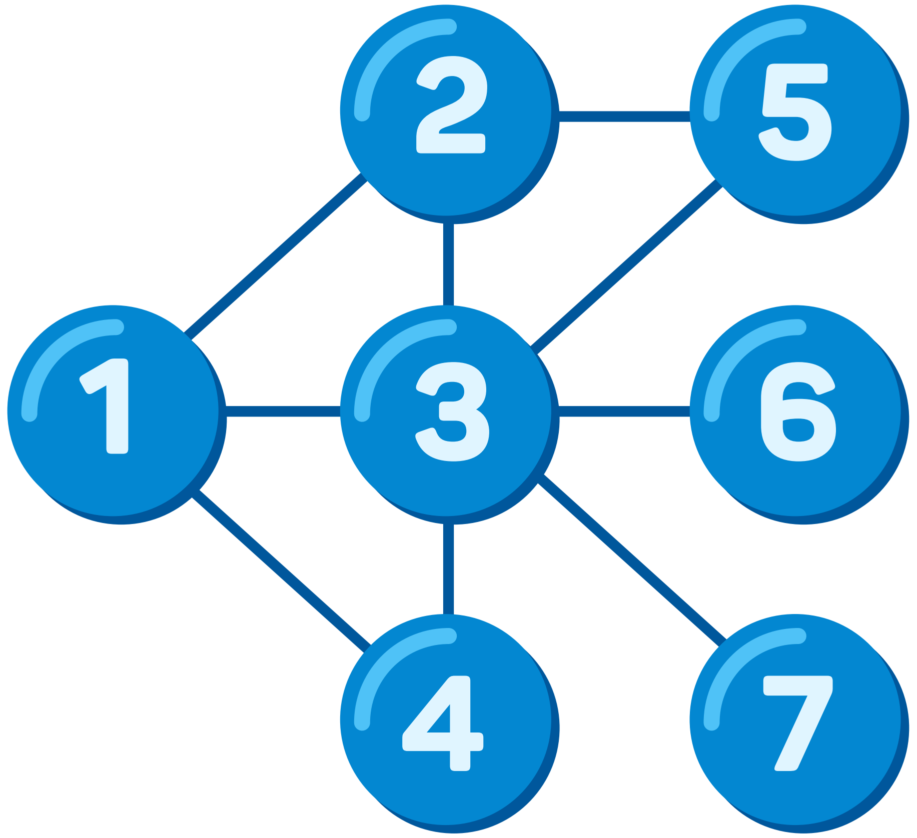


Example:

How many users have at least 2 friends?

```
g.inDegrees.filter("inDegree >= 2")
```

Basic graph and DataFrame queries



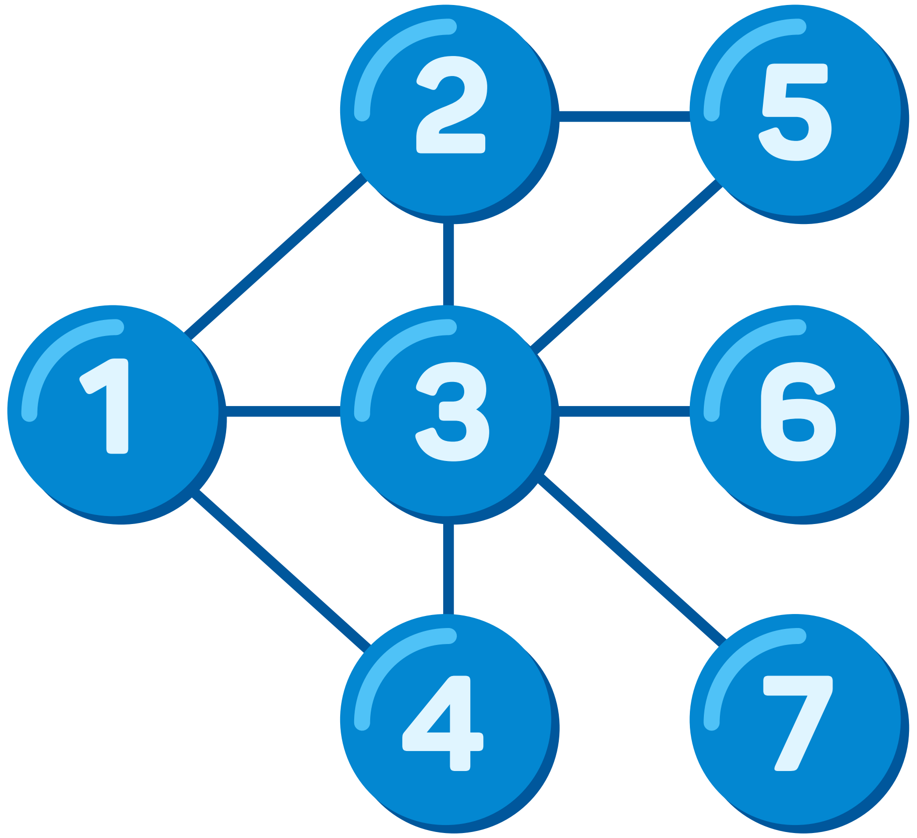
Example:

How many users have at least 2 friends?

```
g.inDegrees.show()
```

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

Basic graph and DataFrame queries



Example:

How many users have at least 2 friends?

```
g.inDegrees.show()
```

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

```
g.inDegrees
    .filter("inDegree > 2")
    .show()
```

```
+---+--------+
| id|inDegree|
+---+--------+
|  1|       3|
|  2|       3|
|  3|       6|
+---+--------+
```

Summary

- How to create GraphFrame

Summary

- How to create GraphFrame
- How to do basic queries to it