# Analysis of Large Scale Socil Networks

# Exercise Session 3: Distributed Graph Processing

## Software

* <http://spark.apache.org>
* <http://spark.apache.org/graphx/>
* <https://zeppelin.apache.org/>

## GrapX Documentation

## <http://spark.apache.org/docs/latest/graphx-programming-guide.html>

* <http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.graphx.package>
* <https://github.com/apache/spark/tree/v2.1.0/graphx/src/main/scala/org/apache/spark/graphx>

## Data Sets

1. Collaboration network between scientists:  
   <https://snap.stanford.edu/data/ca-HepPh.html>

s3n://alssn-bucket/CA-HepPh.txt

1. Citation network in High Energy Physics  
   <https://snap.stanford.edu/data/cit-HepPh.html>  
   s3n://alssn-bucket/Cit-HepPh.txt

## Zeppelin Notebook

* <http://ec2-54-229-149-197.eu-west-1.compute.amazonaws.com:8080>
* Notebook is .json file available from Toledo

## Calculate these properties for data set 2.

1. *Number of nodes (cited and citing paper) and relations.*
2. *The in- and out-degree for each node*
3. *Number of Strongly Connected Components*
4. *Number of Weaky Connected Components*
5. *Size of largest Strongly Connected Component*
6. *Find the paper with the highest pagerank*