Social Network Analysis at Scale in Cloud

Project Proposal

Xiaolan Cai

Problem statement

Social network analysis has drawn a lot attention for the past decades. Initially sociologist have started to study how social networks is formed by and how people are connected. Nowadays social networks like Facebook, Twitter is used spread information like news, ideas. And also for marketing purpose, for example we may also want to study how to maximize the influence of advertisement by choose proper targets.
Social networks normally can be modeled as graph structure model, vertices/node represent an individual and edge represents the relationship between individuals.

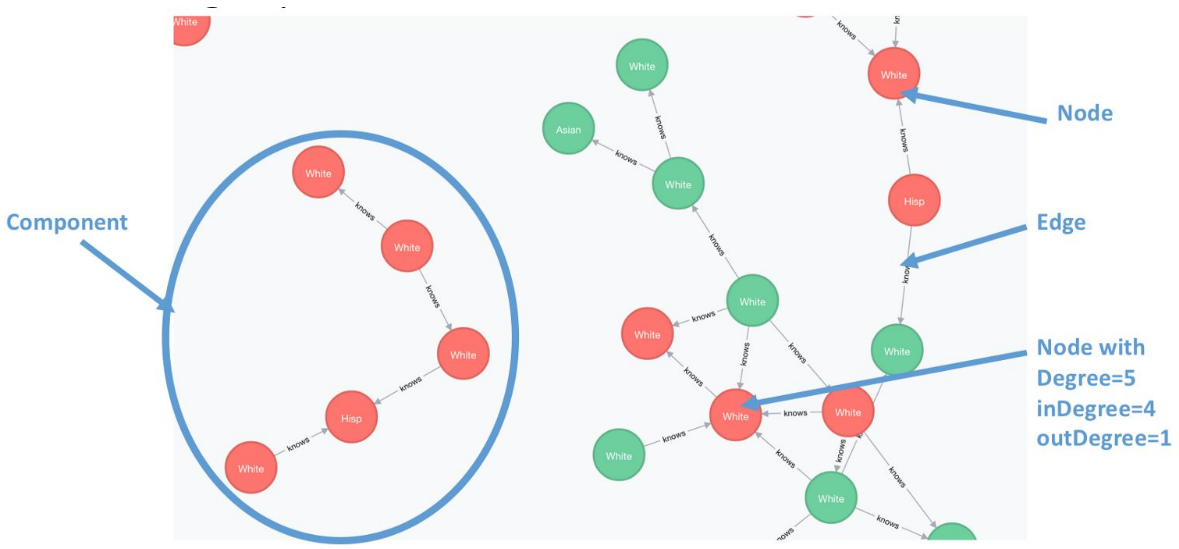


Figure-1: [5]

Meaning time big scale data mining tools is well developed, distributed computing significantly improving the speed of mining data by parallel execute subtasks divided from one single big problem with distributed systems in scalable ways. And social network analysis normally includes massive amount graph-structured data to processing, how to leverage scalable distributed computing work to handle graphed structured data is indeed a interesting topic.

In this work we are interested in a specific problem of SNA, called influence maximization [1], and we want use currently start of art distributed computing tools to exam how well we can processing the graph-structured data in terms of quality and speed.

Max influence in Social network is a interesting topic of SNA (social network analysis). In [1] this problem is formally defined in [1]. Given a Graph N (V,E), and choose an ignition k number of vertices set S, what is the maximum number of influence we can achieve? In [1], two Basic Diffusion Models are considered specifically Linear Threshold and Independent Cascade models. In this work, we considering only Independent cascade models for simplicity. In this model, initially start with a set of active nodes A0, in each step a active nodes v will have a chance to flip his any inactive neighbor w into active one with probability Pvw. In the whole process node v only have a single chance to active any of his inactive node w, meaning he cannot make any further attempts no matter he is successful in flip w or not. And the process stops when there is no mare activation are possible [1].

Finding the optimal solution is NP-hard problem, and in [1], an approximation solution is proposed, it use greedy algorithm to achieve slightly better than 63% of the optimal solution. However, the computation is still pretty heavy, because they need to run Monte Carlo simulation in each update step?

It is challenging task to processing massive data not even to say graph structured data. There are two kind of graph processing framework.GraphLab: share state in parallel vertex computation. Pregel: passage passing, graphx

We need modified existing algorithm to make it suitable for vertex programing, which is request by most graph based commutating framework. We provide a vertex program that is run by each vertex in parallel.

So our question is how influence maximization problem can be processed in cloud computing environment with scalable setups? What performance need to be sacrificed in terms of accuracy to archive speed? (what's the trade off between quality and speed?)

Solutions for processing

One option is to use spark graphx which use message passing between vertices (e.g. Pregel). We first store our data in HDFS and use graphx to build graph by create vertices and edge from it. As normal spark programing use. map to process the source data and then doing targeted programing computing on individual graph objects.

Another option is to combine spark with graph based database such as Neo4j, we import the dataset into Neo4j first. Then we need distribute the data in neo4j to HDFS, after it spark will do parallel computing and put result in HDFS and followed by exporting back to neo4j.

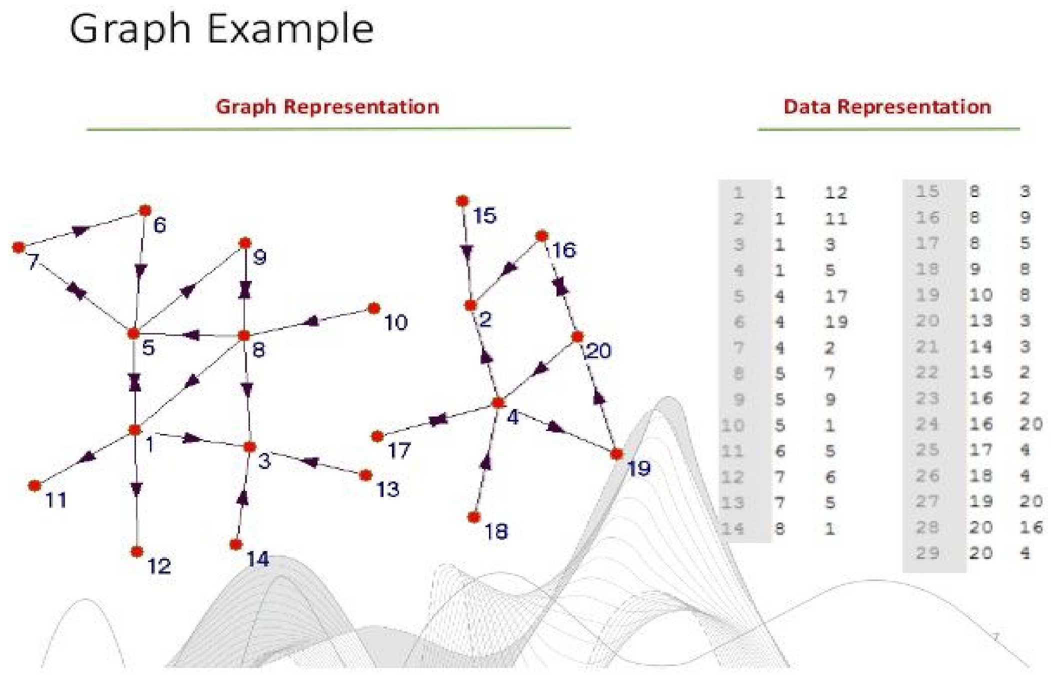
Beside message passing based graph computing framework, we can also use stat shared tool, such as graphlab, we can also explore this tool to for comparing.

Another orthogonal method of improvement the speed of computing is edge sampling, for example we only delete some of the edge of a vertex that have high degrees to reduce the computation with approximation [3].

Setups and data sets:

We will setup our experiment in Cloudlab.

Dataset, we will use undirected social network in snap datasets, such as “ego-facebook”, “gemsec-facebook” [2]. And the data comprise lists of pair of node-ids, which represents adges of graph. The data file is in txt or csf format. We will store the data in HDFS in our cloudlab setups.



1	1	12
2	1	11
3	1	3
4	1	5
5	4	17
6	4	19
7	4	2
8	5	7
9	5	9
10	5	1
11	6	5
12	7	6
13	7	5
14	8	1
15	8	3
16	8	9
17	8	5
18	9	8
19	10	8
20	13	3
21	14	3
22	15	2
23	16	2
24	16	20
25	17	4
26	18	4
27	19	20
28	20	16
29	20	4

Figure-2: [6]

Timelines of the project:

We can divided our project in following phases

Phase 1 (1 week)
We will setup the experiment in AWS/Cloudlab for 1 week, including spark graphx alone setup and neo4j+spark setup.

Phase 2 (1.5 weeks)
We will measure spark alone setup

Phase 3 (1.5 weeks)
Measure neo4j + spark setup

Phase 4 (1 week)
Report write up.

Challenges:

To adapt the influence maximization problem for spark liked framework is difficult, as we need to defined vertex based programing for parallel with approximation and not sacrifice the accuracy too much.

Issues:

- Time constraint, since I am doing this project only, I have concern about I may not able to finish all the measurement I have planned to, but at least two feasible scalable setup should be build and measured.
- Spark graphx use Scala for programing and I need to learn scala for this.

Reference:

[1]. Maximizing the Spread of Influence through a Social Network
[2]. https://snap.stanford.edu/data/

[3]. Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris (2003). statnet: Software tools for the Statistical Modeling of Network Data. URL http://statnetproject.org
[4]. Bridging the GAP: Towards Approximate Graph Analytics
[5].https://medium.com/mobileforgood/social-network-analysis-using-apache-spark-and-neo4j-1ccba3c8af9a
[6]. https://www.slideshare.net/GhulamImaduddin1/social-network-analysis-with-spark