Social Media Mining (CSE598)

Project Phase-1

-Karthik Vivekanandan (1207683239)

**Introduction**

Network analysis tools used: Snap.py, Networkx

Included in the submission is a requirements.txt file, which contains a simple list of all the packages used in my environment, and their respective versions. To re-create the environment to install the same packages using the same versions, one can use “pip install –r requirements.txt” to run these submitted programs.

**Twitter crawling**

Implemented a crawler to visit 1012 twitter users, through a user’s friend network. Those users with access level set to ‘private’ were skipped over setting their out degree to 0

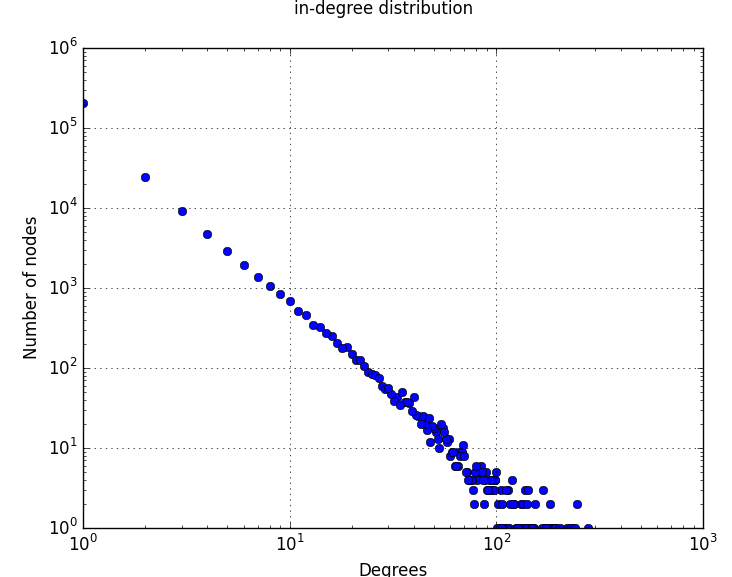
Total nodes in dataset: 258509

Total edges in dataset: 481582

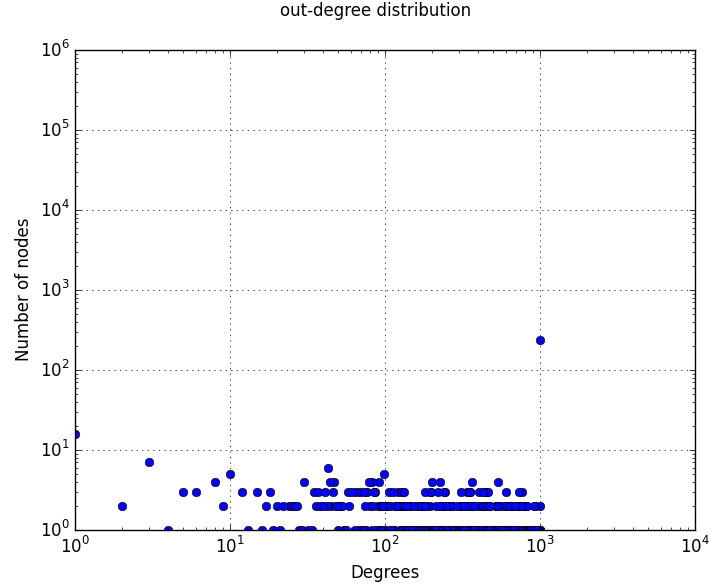
The dataset contains the Twitter user\_id field, which is anonymized and used for the subsequent steps.

**P2**

**In-degree distribution**



**Out-degree distribution**



**Power law**

Calculating best minimal value for power law fit, I got these values:

Alpha exponent of in degree distribution: -2.23090496139

Alpha exponent of out degree distribution: -1.84436355593

**Bridges**

Tried with the bridge detection algorithm mentioned in the text book, but aborted the execution after close to 22 hours of continuous processing without any results.

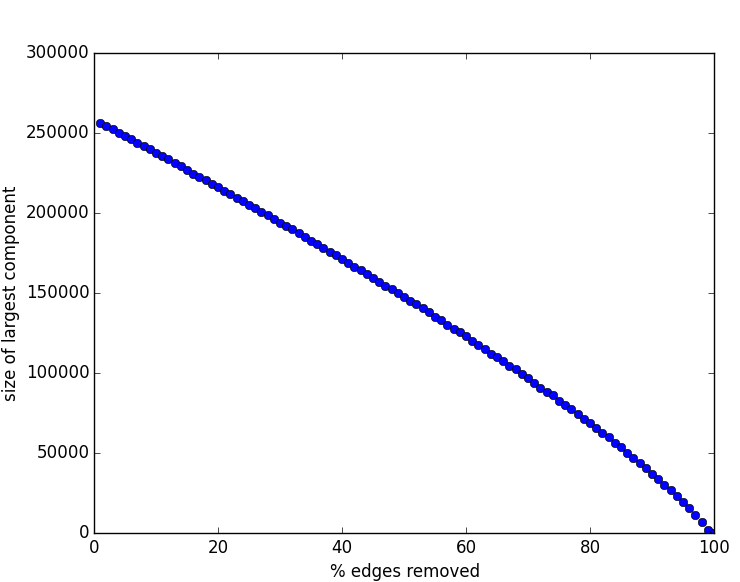
**3-cycles**

Total number of 3-cycles: 258509

**Diameter**

The sampled diameter of the graph is 7

**Visualizing effect of random edge removal on largest connected component size**



**P3 Network Measures**

Clustering coefficients of crawled network:

Local clustering coefficient = 0.067029

Global clustering coefficient = 0.010442

**Centrality**

**Page Rank**

|  |  |  |
| --- | --- | --- |
| No | Node ID | Page Rank Centrality |
| 1 | 678 | 7.69E-06 |
| 2 | 681 | 7.68E-06 |
| 3 | 9708 | 7.67E-06 |
| 4 | 692 | 7.49E-06 |
| 5 | 558 | 7.49E-06 |
| 6 | 315 | 7.38E-06 |
| 7 | 37 | 7.36E-06 |
| 8 | 297 | 7.30E-06 |
| 9 | 298 | 7.28E-06 |
| 10 | 24419 | 7.25E-06 |

**Degree centrality**

|  |  |  |
| --- | --- | --- |
| No | Node ID | Degree centrality |
| 1 | 605 | 0.000947746 |
| 2 | 297 | 0.000947746 |
| 3 | 692 | 0.000932273 |
| 4 | 401 | 0.000912931 |
| 5 | 134 | 0.000881984 |
| 6 | 2 | 0.000847169 |
| 7 | 645 | 0.000789144 |
| 8 | 61 | 0.000754329 |
| 9 | 1455 | 0.000746592 |
| 10 | 1449 | 0.000738855 |

**Eigen Vector Centrality**

|  |  |  |
| --- | --- | --- |
| No | Node ID | Eigenvector centrality |
| 1 | 605 | 0.094956883 |
| 2 | 476 | 0.089509002 |
| 3 | 2 | 0.08719305 |
| 4 | 965 | 0.08532927 |
| 5 | 692 | 0.084466795 |
| 6 | 134 | 0.081707412 |
| 7 | 1010 | 0.079477697 |
| 8 | 1057 | 0.07692632 |
| 9 | 410 | 0.076492721 |
| 10 | 695 | 0.076085917 |

**Rank correlation – Spearman**

*Columns: dc pc ec*

[[ 1. 0.65818536 0.43262553]

[ 0.65818536 1. 0.22420601]

[ 0.43262553 0.22420601 1. ]]

**Jaccard similarity**

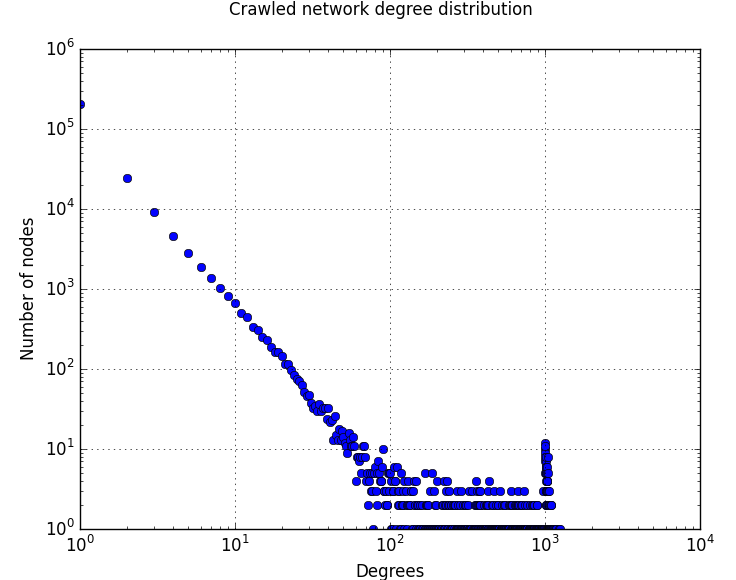
**P4 Network Models**

**Crawled Graph**

Local clustering coefficient = 0.067029

Global clustering coefficient = 0.010442

Average path length = 3.986000

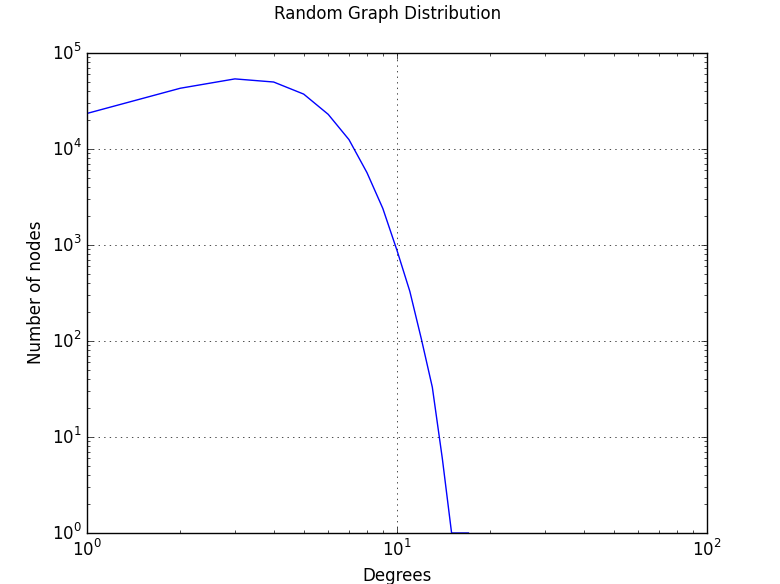


**Random Graph Model**

Local clustering coefficient = 0.000012

Global clustering coefficient = 0.000015

Average path length = 10.294550

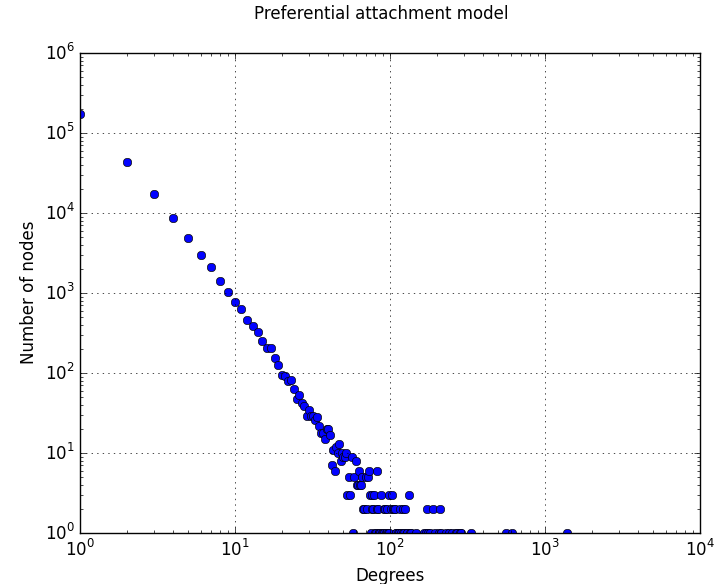


**Preferential Attachment Model**

Local clustering coefficient = 0.000377

Global clustering coefficient = 0.000106

Average path length = 6.185000

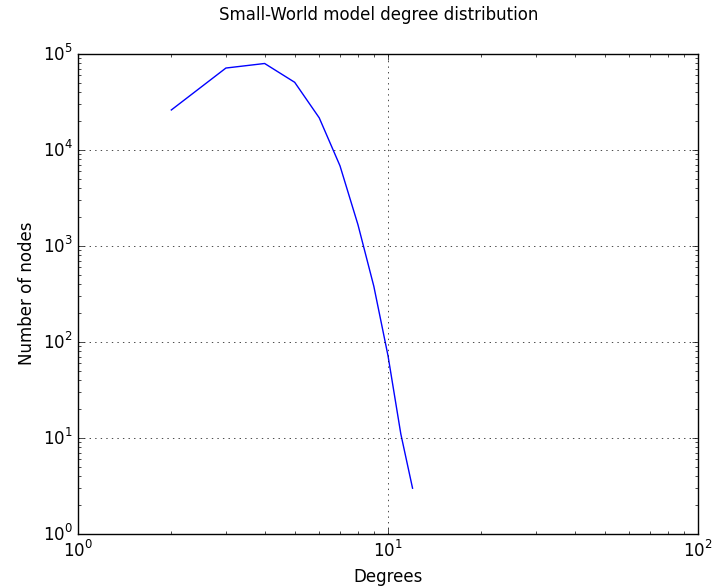


**Small World Model**

Local clustering coefficient = 0.047956

Global clustering coefficient = 0.039214

Average path length = 10.308500



**Comparison**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Local CF** | **Global CF** | **Average Path Length** |
| Crawled Graph | 0.067029 | 0.010442 | 3.986000 |
| Random Graph Model | 0.000012 | 0.000015 | 10.294550 |
| Small World Model | 0.047956 | 0.039214 | 10.308500 |
| Preferential Attachment Model | 0.000377 | 0.000106 | 6.185000 |