Twitter Data Report

**Join**

The two queries are:

1. Producing the list and the count of user ID pairs which represent "true friends". (they follow one another).

2. Producing the list of top 10 "celebrities", sorted by (in-degree - out-degree)-difference in descending order (from more famous to less famous).

The relational algebra equivalent is:

1a)(list)

1b) (count)

2)

The SQL equivalent is:

1a) (list)

CREATE VIEW true\_friend AS  
SELECT R.uid1, R.uid2  
FROM followers AS R, followers AS S  
WHERE R.uid1 = S.uid2 AND R.uid2 = S.uid1 AND R.uid1 > R.uid2;

1b) (count)

SELECT COUNT(\*)  
FROM true\_friend;

2)

CREATE VIEW out\_deg AS  
SELECT uid1 AS uid, count(uid2) AS out\_degree  
FROM followers  
GROUP BY uid1;

CREATE VIEW in\_deg AS  
SELECT uid2 AS uid, count(uid1) as in\_degree  
FROM followers  
GROUP BY uid2;

SELECT out\_deg.uid, in\_degree, out\_degree, (in\_degree – out\_degree) AS difference  
FROM out\_deg, in\_deg  
WHERE out\_deg.uid = in\_deg.uid  
ORDER BY difference DESC  
LIMIT 10;

Results from the query:

* total number of users: 11316811
* total number of “true friend” pairs: 21776094
* top 10 celebrity:

|  |  |  |  |
| --- | --- | --- | --- |
| **UID** | **In degree** | **Out degree** | **Difference** |
| 5994113 | 564512 | 292 | 564220 |
| 7496 | 350885 | 6035 | 344850 |
| 1439110 | 341963 | 1472 | 340491 |
| 1629776 | 172231 | 2120 | 170111 |
| 8121005 | 155967 | 34 | 155933 |
| 2041453 | 152689 | 620 | 152069 |
| 797152 | 118826 | 74 | 118752 |
| 6623784 | 116002 | 183 | 115819 |
| 645019 | 107914 | 275 | 107639 |
| 3403 | 102877 | 4946 | 97931 |

Summary:

On average, each user has around true friend, which is fairly small considering how large the twitter dataset is (think of how many friends you have on facebook, I am pretty sure you have more than two). Of course this is just an average; there are users with more followers and there are users that follow a lot of other people.

As for the top 10 celebrities, there is a clear trend that the number of people the celebrity follows is a lot less than the number of people that follow them. For the number 1 user, it also has the most followers out of the whole dataset (see part 1 of this report). If we plot the popularity (fame) vs. the rank of each celebrity, the graph would look like:

There is a slight trend of logarithmic relation between the two variables, but in order to tell, we would have to plot more data points.

Other potential question we can ask about the dataset:

* Perhaps the fame vs rank relation is also a power-log relation. Can we plot more data points to verify?
* If we plot the twitter dataset into a directed graph with the users being the node and each direct edge signifying one user following another, what’s the minimum distance from one node to another for all nodes? (Eg. someone might claim that in at most 6 links, you can reach from one user to any other user. Is this true for all users?)
* Is there any user that doesn’t follow anyone? If so how many?
* Is there any user that follows itself? If so how many?