1 Data pre-Description

The data consists of seven data files.

1. *byFollowed.csv*
2. *byFollowers.csv*
3. *IDandDate.csv*
4. *list_whole.csv*
5. *personProducts.csv*
6. *persons.csv*
7. *products.csv*

The files "*byFollowed*.csv" and "*byFollowers.csv*" are practically the same, where both demonstrate the direction of an interaction and its recorded time.

For example the *byFollowe*r.csv looks like below:

	follower	followed	relationDate
1	0000005	sushie	2008-04-22
2	000001	nataliezee	2008-03-15
3	00001020	sassoo	2008-05-17
⋮			
335120	__beeba	kabiri_victoria	2008-03-19
335121	___mike__	shopbop_stylist	2007-06-03
335122	_____	Shopbop_Michelle	2008-08-05

Table 1: *byFollowers.csv* data sample

The two equivalent files, *byFollowers.csv* and *byFollowed.csv*, consists of *79280* unnique individuals interacting. However the file *persons.csv* has only *38266* unique indidivuals. The effective number could also be lower, as some users recorded *"last login date"* and "*member since date"* are the same, and alongside other information they could be just removed. The file *persons.csv* looks as below:

	id	lastLogin	memberSince	...	profileId
1	00001020	2008-05-18	2008-05-17	...	363767
2	000123	2010-02-28	2010-02-28	...	607508
3	000309	2008-02-25	2008-02-20	...	215179
⋮					
38264	_marie_	2010-05-21	2010-05-21	...	618023
38265	_wildhorse	2009-01-11	2009-01-02	...	524881
38266	_xXSwim_a_holicXx_	2009-08-03	2009-08-03	...	577465

Table 2: *Persons.csv* sample data

More importantly *byFollowers.csv/byFollowed.csv* and *persons.csv* have only *18863* individuals in common together. This might be a bit of a concern regarding making of the network, and relating it to the products in the *personProducts.csv*.

On the other hand the data *personProduct.csv* also contains *54002* unique number of individuals and all the individuals in the *Persons.csv* are contained in this data as well. Although a better situation, the overlap between the individuals in the *byFollower.csv* and the *personProducts.csv* is *22119*.

	id	personId	productId	hivedSince
1	1	rocketrobyn	231	70m
2	2	chrisl	231	76m
3	3	mcarrier69	233	76m
⋮				
2737774	2836554	zjl_work	674513	38m
2737775	2836555	ZUBURBIA	216017	56m
2737776	2836556	ZUBURBIA	294724	55m

Table 3: *personProducts.csv* sample data

Next we look into the *list_whole.csv* which contains 19820 unique *originators(*yet to be known what it is, probably just a one who creates a bookmark). The originators are almost all included in the *personProducts.csv*. The overlap between *list_whole.csv* and *Persons.csv* is *16516* and only *9310* with *byFollowers.csv* unique individuals.

2 Network

The network data is represented in either of the *byFollowers.csv* or *byFollowed.csv*. The network should be very sparse and should have some unconnected components. The nodes come from the 79820 individuals for now we consider a sender/source and a reciever/sink and later we can decide whether we want to keep the directed or the undirected network.

The network file is constructed in the Network.csv.

Another issue would be to keep the network data in memory. As the network matrix is sparse and large, we should use other more efficient methods to save the data in memory. Maybe for now though we would need adjacency list or some use of hash maps.