**MS Special Problem Report**

* **Guided by Prof. Eric Gilbert and Tanushree Mitra**

**Objective:** The project deals with studying the /r/conspiracy and the /r/conspiracyAMA subreddits.

**Motivation:** This topic was motivated after looking at the huge number of conspiracy theories that occur on the internet. In the background study it was observed that individuals were more likely to endorse conspiracy theories if they were inclined in participating in such alleged conspiracies. There is a certain trend observed among the conspiracists wherein they willingly agree to the labelling of the “conspiracy theory” to people who have beliefs opposite to their own. Conspiracy theories, defined as allegations that powerful people or organizations are plotting together in secret to achieve sinister ends through deception of the public. This has been an element of popular discourse. These usually hinge largely on the broader beliefs of the world. Some of these conspiracy theories maybe related which is why they garner the same audience, while some very different conspiracy theories may have the same set of followers. There are many important areas where these conspiracy theories develop such as theories around the presidential assassinations. As social scientists it is important to find how these conspiracy theories are developed, how do the disperse in the network, what impact do they have on the people who read these theories etc. This will help curb the spread of data that is not credible or falsified and prevent people from being fooled into something that may not be true.

**Tools:**

* Bigquery: Most of the subreddits have their data posted here. Since it is a public forum it is easy to access the data. The problem faced here there is no direct way of storing the data onto the local machine. Incase if the data being queried is less then this can be stored directly to the bigquery cloud storage. However, if the amount of data being stored is large then you have to first store it as a database table and then this table should be stored to the cloud storage. The CSV’s from the cloud storage have to be manually downloaded. Querying the data is similar to writing SQL statements.
* Bigquery cloud storage: Was primarily used to store all the data collected from the google big query before downloading it onto the local machine’
* Python: This was the programming language used to code most of the metrics that were used to study this data. Netwokx was used while analyzing the thread structure and to see how the posts and comments are structured with respect to each other. Pandas and numpy were used for most of the other metrics calculations.
* Flashmob servers: All the data collected is present in the flashmob00 and flashmob01 servers. The code files are also present on these servers. Vim editor was used to run the scripts on the server.
* Github: All code, with detailed description of what is being computed along with the data files collected is stored on git. This is the [link](https://github.gatech.edu/shiremath9/Social-Computing-Project).

**Work Done** (representation in points here; for more details, please visit the github account)**:**

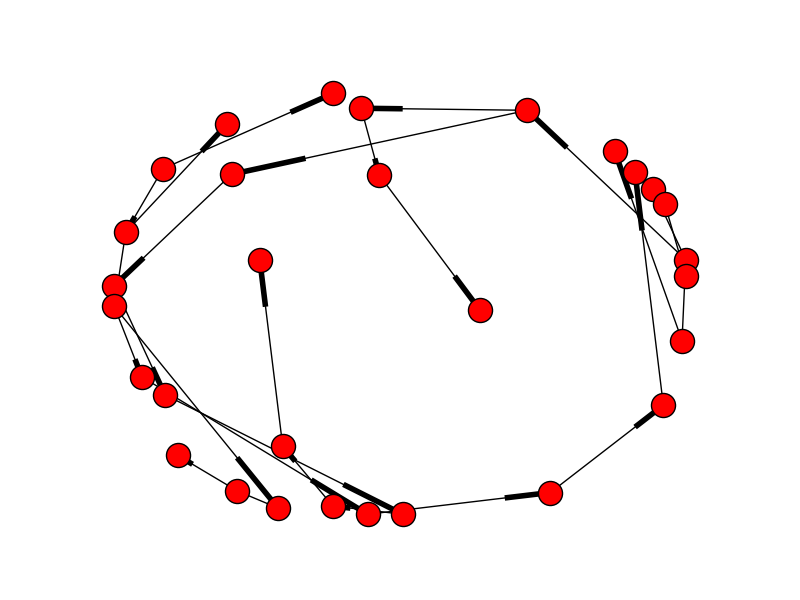
* Reading through the papers in the reference section
* Familiarising with biquery and bigquery cloud storage. Identifying the public datasets to use and obtaining data from them. Creating my own database on bigquery; storing the obtained values into the tables in this database.
* Obtaining unique users from among the comments and submissions respectively.
* Identifying the number of posts and comments per user for both posts and comments.
* Getting a better representation of data for every post, in the format dict [submission :comments]. So now each submission is the key and the values are the comments made

for that submission.

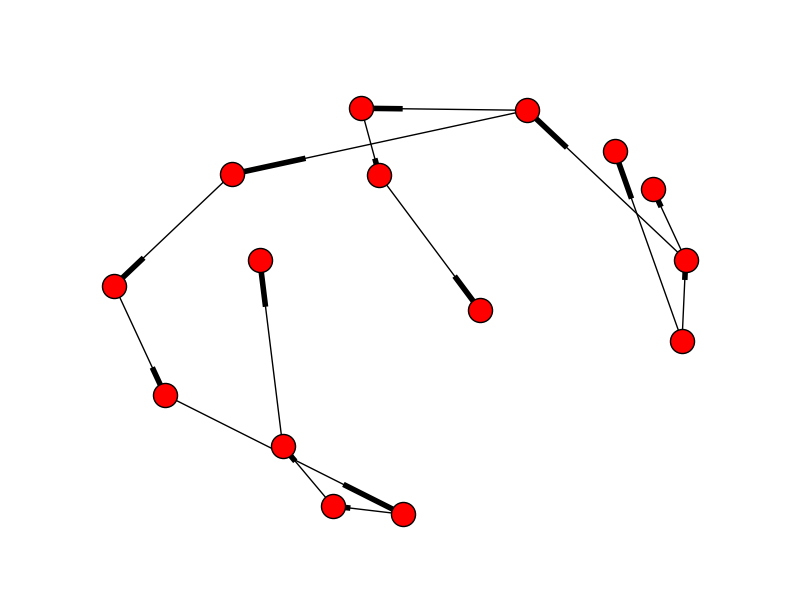
* Identifying the unigrams and the bigrams from the submissions and the comments using

the NLP tokenizer

* Find the reddit thread structure for this subreddit.
* Find the reddit thread depths
* Obtaining user timelines – this data was obtained from reddit comments: 2009- 2016-08. This data was also obtained from the full corpus
* Implemented finding the thread length
* Implemented finding the lifetime of a thread
* Implemented the ratio of the thread structure.

****

**Fig 1:** The depth first search structure obtained on starting with ‘czywyuz’ as the root node.



**Fig 2:** The breadth first search structure obtained on starting with ‘czywyuz’ as the root node.

**Problems Faced:**

* The major roadblock faced was while collecting data for user timelines. I had to spend around 50-60 hours those 2 weeks to get all the data from 2009-2016. The problem here was that bigquery did not have a command line statement or any other method to directly download all the files stored on the google cloud storage.
* I was not aware of the networkx package. It took me some time to figure out that python has something so useful built to be used. It was a good learning experience.

**Future Work:**

* What were the terms/concepts/theories discussed and during what time frames on the subreddit? Example: If a conspiracy aroused, what was the time frame from its birth till its extinction? What aspect of the conspiracy was most discussed?
* Were there any external factors that took place to give rise to such theories?
* What were the discussions like? Can we predict these discussions given a situation where a new but similar conspiracy were to be faced?
* Since I have collected the user timelines and have this data, the study can be geared towards the users as well – how do they behave on other forums? Does the information they gather elsewhere affect the kind of discussions held on this subreddit? Does information percolate?
* I have not worked much with NLP, but I guess there could be some methodologies that could be used to make deductions from text collected through comments; something that I am not aware of.

**Conclusion:**

* I learnt a number of things through this project, and most of my learning was in the space of utilizing the different theories and methodologies present in the existing research papers and using them to study this particular subreddit.
* Since I have collected so much data and resources I am motivated to work on this and take it further.
* Some of the metrics I have implemented do not seem to be in sync with the general observations. For example, the re-entry of the participants is giving me a score for every user (and this is generally not the case). When I investigated this issue I found that all the users were indeed unique in the data that was obtained. This was not true on verification from the subreddit. This draws suspicions to the source where this data was obtained from.

**References:**

1. [Does it take one to know one? Endorsement of conspiracy theories is influenced by personal willingness to conspire](http://onlinelibrary.wiley.com/doi/10.1111/j.2044-8309.2010.02018.x/full)
2. [“What about building 7?” A social psychological study of online discussion of 9/11 conspiracy theories](http://journal.frontiersin.org/article/10.3389/fpsyg.2013.00409/full)
3. [The popularity of conspiracy theories of presidential assassination: A Bayesian analysis.](http://psycnet.apa.org/journals/psp/37/5/637/)
4. [Belief in Conspiracy Theories and Susceptibility to the Conjunction Fallacy](http://onlinelibrary.wiley.com/doi/10.1002/acp.2995/epdf)
5. [Mental Health Discourse on reddit: Self-disclosure, Social Support, and Anonymity](http://www.munmund.net/pubs/icwsm_14_1.pdf)
6. [Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions](https://arxiv.org/pdf/1602.01103.pdf)