1. Note to the Markers:

Hello and welcome. For your benefit, I've pasted the marking rubric here and added notes as applicable.

Thanks for your time and consideration,
Peter Cebo

Obtaining data (8 points total):

- 2 pts: documentation of how data was obtained (newly assembled or existing data) + source code if applicable; **AND** 2 pts: explanation of criteria for inclusion of nodes and edges
    - All below in the "Data Collection, Cleaning and Exploratory Analysis. Steps Taken:" section.
- 4 pts: subjective interestingness/originality of the subject of data collection
    - Subjective, as the title says, but I personally thing that this is a very interesting topic and that the way the data was manipulated to make it possible to analyze from a network perspective was original. If you've never check out reddit.com, I really recommend it! But, be careful, it can eat up a lot of your time! ☺

Data analysis (17 points total):

- 3 pts: were at least 3 metrics/methods from the course applied to the data?
    - I find this to be slightly subjective as well, but as examples, I used or applied: the directed edges concept; the concept of weighted edges; the degree metric; and various exploratory graphing methods in Gephi. I realize that these are the simpler/easier concept taught in this course, but I still see them as important and applicable.
- 4 pts: were they applied/interpreted appropriately?
    - I believe so. ☺
- 2 pts: was at least one additional technique, not covered in the course materials, applied to the data?
    - As you will see in section 2, I used various MS Excel methods (e.g. excel formulas, pivot tables, etc.) to perform initial data analysis and exploration. Further, I used data plotting techniques in R that were not covered in the course.
- 5pts Visualization: Did the visualization add to your comprehension of the data?
    - Your call. I attempted to ensure that was the case.
- 3 pts: was code/step by step instructions provided such that one could replicate the methods?
    - As noted throughout, everything is available on github here: https://github.com/pcebo/SNA_Project. I've also done my best to describe any steps that are not provided in the code/files.

Interpretation (5 points):

- 2pts: were limitations of the data correctly addressed?
    - See footnotes throughout and considerations in the conclusion.
- 3pts: did the analysis yield new insights (subjective)
    - Subjective, as the rubric states.

Thanks again and best of luck.

2. Introduction and Brief Background:

For my Coursera - Social Network Analysis - Peer Assessment Project, I've chosen Option 1: empirical network analysis. (find data, analyze data (and visualize it), then interpret.)

The data used was taken from a Stanford web site and is "a collection of 132,308 reddit.com submissions. Each submission is of an image, which has been submitted to reddit multiple times. For each submission, we collect features such as the number of ratings (positive/negative), the submission title, and the number of comments it received. We also include the html of the comment pages themselves." (http://snap.stanford.edu/data/web-Reddit.html)

The data was used in some analysis on the factors that affect the popularity of a social post, and there are some interesting insights that are presented in the resulting paper here:

H. Lakkaraju, J. J. McAuley, J. Leskovec What's in a name? Understanding the interplay between titles, content, and communities in social media. ICWSM, 2013.

The authors, however did not perform any network analysis on the data and the dataset had to be further manipulated to allow for the building of a network model/graph. The next section describes the steps that were taken to accomplish this, using various tools, including MS Excel, R and Gephi.

Note that the hypothesis/hypotheses are intentionally omitted in this intro, staying true to the way that this analysis was performed – i.e. exploratory analysis was performed first to identify what further interesting insights could be uncovered in the data.

For this analysis, MS Excel 2010, R version 3.1.1 and Gephi 0.8.2 were used.

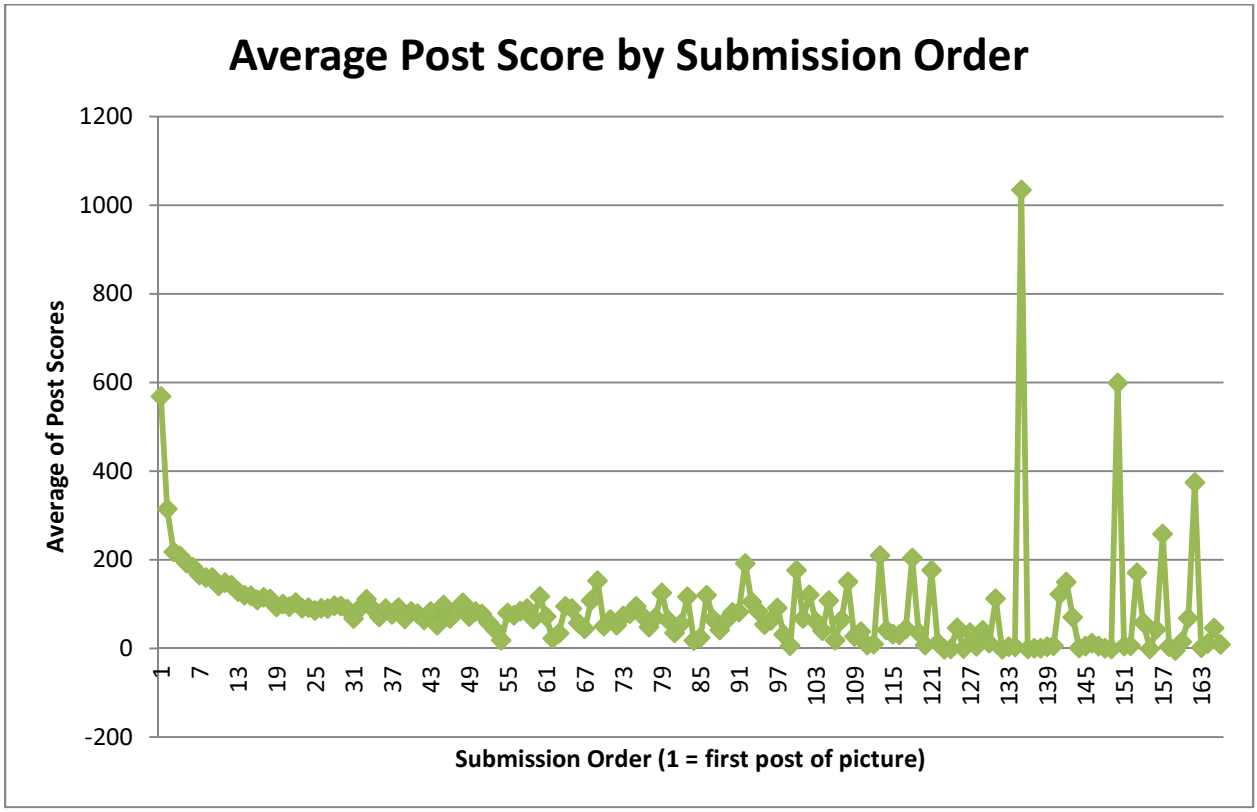
3. Data Collection, Cleaning and Exploratory Analysis. Steps Taken:

1) Download data set from: http://snap.stanford.edu/data/redditSubmissions.csv.gz
2) Unzip and open in Excel.
3) Sort the data by two levels: (a) #image_id; and (b) unixtime.

- This accomplishes the task of putting each submission in order from first (initial, original post) to last repost (as of the end of the time period this dataset covers)[1]

4) Insert a column to the right of column A (#image_id); Insert the following formula into the first cell of this new column (B2): =IF(A2=A1,B1+1,1)
    - This accomplishes the task of numbering the chronological order in which each picture was posted.
5) Add another column, and using a similar formula to the one used in number four, calculate the amount of elapsed time for each post since the last post of the same picture.
    - This results in a column with the value "first_post" or the amount of hours elapsed since the previous post of a "repost". The time elapsed is given in hrs.
6) Lastly, we add another column with a "Boolean" value, simply telling us if that row is a "first_post" or not (i.e. a "repost).

Next, we used pivot tables to plot a few quick charts.



As one would expect (and as the earlier mentioned paper also confirms), earlier/initial posts receive higher scores in general. However, we do see some reposts on the high end of the spectrum with very high scores. **Is there anything we can use from a network analysis standpoint to explain these outliers?** The research paper attributes the variance to two main (groups of) factors – "community" factors (submission order, subreddit, time of post) and "title" (the title/words used in the post). One question

[1] Note that we are in fact told that the initial posts are original on page 2 of the corresponding paper.

we should try to answer is: **Can we further improve their model by using concepts from Social Network Analysis?**

The Excel file with the above data, pivot table, chart, as well as the steps taken to create the dataset for the below (omitted for brevity) is available at: https://github.com/pcebo/SNA_Project, should the reader be interested. Note that some additional exploratory analysis (pivot table and charts was performed, but it omitted here as is it already covered in the research paper related to this data set.)

4. Further Graphing and Analysis in R and Gephi:

Analyses and Hypotheses/Questions:

- Perform graphing of the created network in Gephi to see if anything interesting arises.
    - **Are there distinct communities of reddit members who are likely to repost each other's content?**
- Calculate the degree (see below) of each poster of any original content and see if it has an influence on the popularity of that users posts.
    - **Are users who have their content reposted more likely to post popular content?**

To perform analysis of the "network" we need to first define what an edge is. Using the concept of a "directed edge" we first create a table (.csv file, available on github) that has a list of all initial posters (users) and the reposters (also users) that reposted content. We will define someone who has had their "original" content reposted as an influencer and their (out^2) degree in the network will represent their influencer score[3].

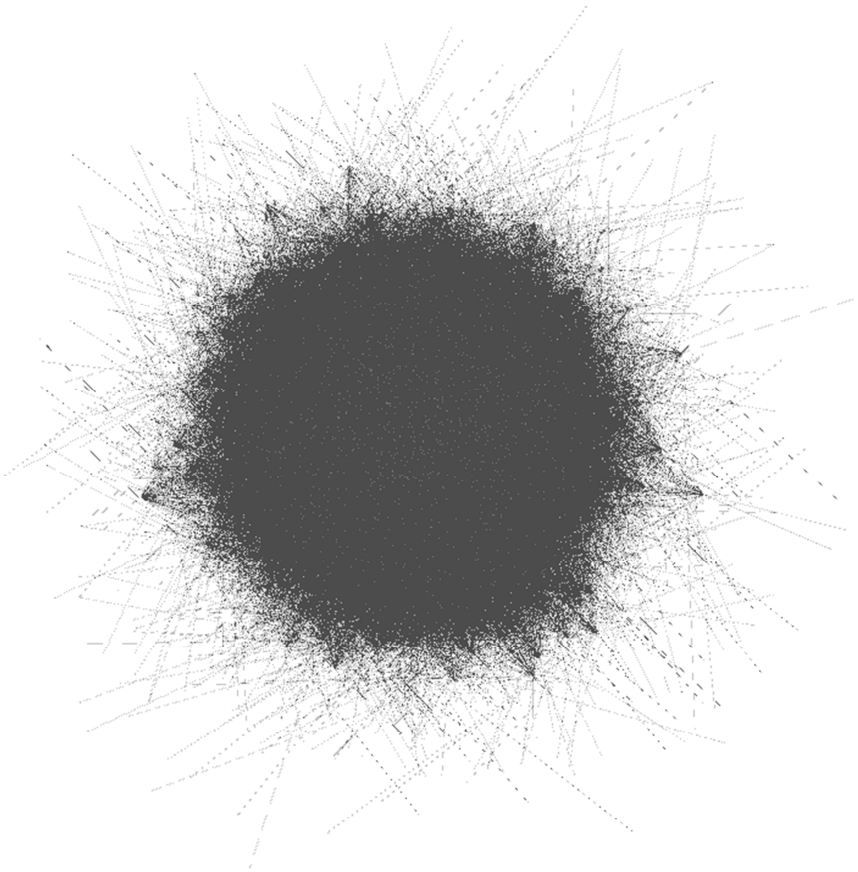
Importantly, note that naturally those users that have more original posts will have a higher degree (i.e. a user may get a higher influencer score for posting more bad quality content, vs. a user that posts fewer pieces of content that are higher quality). Since these users are posting *original* content however (fairly rare on reddit), they are not penalized in our analysis for doing so.

Note that any rows with a missing user name were omitted here, reducing the size of the dataset slightly.

I also create a second (set of) .csv file(s) (also available on github) with "edges" as I have defined them and different variations of edge weights: a) edge weight = the submission order number; b) the inverse of a; and c) no edge weights. All three were laid out in Gephi to explore for interesting patterns or clusters. Different layout methods were tested and ForceAtlas 2 was the method that was chosen.

---

[2] We could have chosen in or out degree, but we will have the arrows point in the direction of the "content flow".
[3] Naturally, just because someone reposted content doesn't mean that they saw the initial (or any previous) post and "stole" it. However, since reddit does have a very good search feature, we can assume that at least a relevant portion of reposts were done so knowingly, and even those that weren't should still give credit to the original poster for being first.

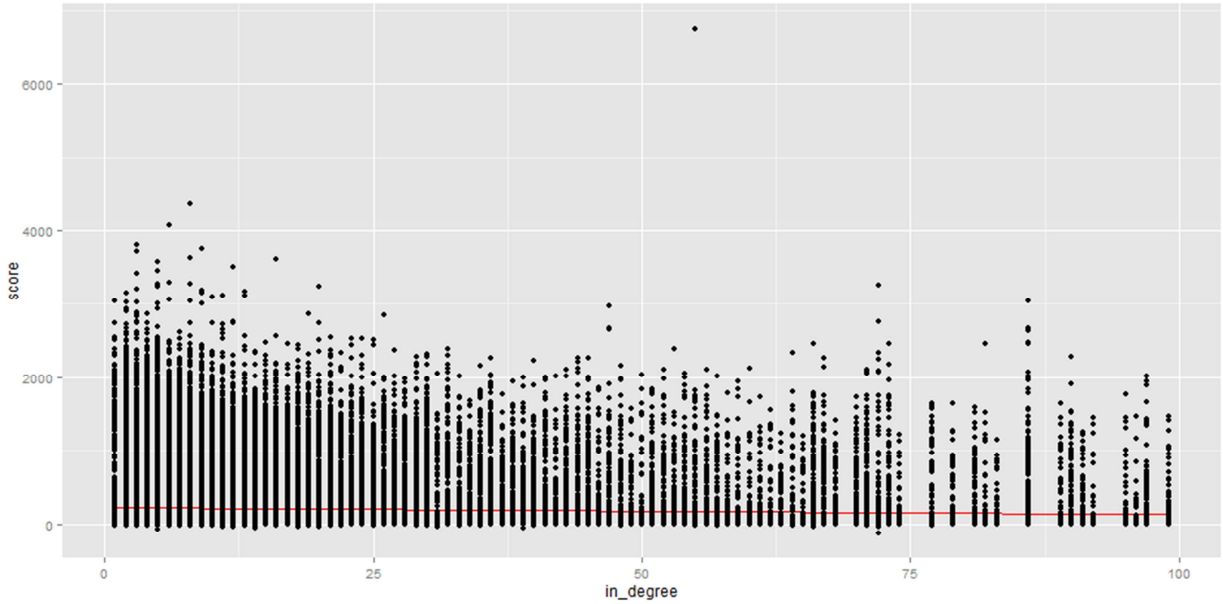


Above is a screenshot of the network, laid out using the weighted data set. As can be seen, this did not produce any discernable communities or interesting data, unfortunately.

Finally, we move to R and attempt to model out a relationship, if one exists between the in degree of a user (i.e. the "influence score) and the popularity of their posts. A few exploratory charts were modeled. As with other items, the R code is available on github.

Surprisingly (and interestingly?!), we find that a user's network in degree is not correlated (or perhaps even slightly negatively correlated) with the scores of their reposts. Note that the red line is the fitted

linear model in the below chart.



5. Conclusion and Lessons Learned:

- **Ultimately, I was not able to conclusively find a network related metric that would further improve the post success model that was presented in the research paper. However, it was slightly surprising to find that reddit users don't have any "credibility" or "influencer" power as one might expect, at least not in terms of how I chose to measure it and its impact.**
- This present an interesting question. On observation, reddit does not make it immediately clear if the content that has been posted is coming from a "credible" user by any means. Further, it is unknown if more "credible" users have their content float to the top more easily.
    - **Should reddit implement a (secret?) user rating system that pushes credible user content to the top pages (it seems they don't as of yet)?**
    - **Should reddit add a user rating beside each user so that consumers of the content can see what the user's reputation is before clicking (this could be based on newness of content submitted by the user, amongst other things.**

Other thoughts:

- The dataset that was chosen (created) was very large (55641 nodes and 76173 edges). In retrospect, perhaps the project would have been easier to execute if a subset of the data was chosen (many tasks took a long time).