On Detection of Users and Content of Interest on Reddit*

Rollen S. D'Souza[1]

Abstract— **The online social networking platform Reddit has gained large popularity over the last decade. The news aggregation and discussion site is organized into subreddits whose existence is focussed on a given topic: from political rambling to serious discussion of novel research to homework help. Content on these subreddits is sorted using a voting scheme: users can either upvote or downvote a given submission. It is, however, not easy to identify engaging content or engaged users over long spans of time. I propose a systematic framework using the Katz centrality measure to identify content that was most engaging over a large period of time that does not rely on simply looking at upvotes. The technique both ranks users for their engagement as well as the content that received the most engagement. I also investigate using spectral clustering to reason about behaviour on subreddits.**

I. INTRODUCTION

The goal of this project was to investigate how one can use the tools developed in class for the purpose of analyzing social networks. Social networks have a natural inherent graph-like structure that lend themselves to the tools provided by graph theory. A commonly used social network is Reddit. The Reddit platform provides users a place to submit content, either personal or externally linked, to a sub-community for the purpose of discussion. These communities are called subreddits. In this paper, I investigate an application of the Katz centrality measure to the graph structure of submissions, comments and users found on a given subreddit within a fixed time-frame in order to better understand the key content and players engaging in the community.

II. MODELLING

A. Assumptions

In order to reason about interesting content without temporal information or content analysis, we make the following simplifying assumptions:

Assumption II.1. Users are interested (positively or negatively) by content they read.

Assumption II.2. Users comment only on content they have read.

Assumption II.3. Content is made interesting by the content creator.

Assumption II.4. Deleted content and users are not interesting nor generate interesting content and do not affect other users in a meaningful way.

*This work was not supported by any organization

[1]Rollen S. D'Souza is a graduate student in the Department of Electrical and Computer Engineering, Faculty of Engineering, University of Waterloo, rollen.dsouza@uwaterloo.ca

Assumption II.5. Active engagement of users (through comments or submissions) is a good measure of interest.

Assumptions II.1 and II.2 are the most contentious. They effectively assert that users engage with the platform in an intellectually curious manner. This is, at face value, absurd. On the other hand, the case can be made that even if users don't fully read content that they interact with, they at, the very minimum, skim and grasp the general gist of the content. This relaxation of the assumption is sufficient but doesn't lose the essential structure of interest implied by the assumptions. Specifically these two assumptions imply that we can infer a user is interested in content when we detect a comment made by the user. As a result, we take the first two assumptions as reasonable. Assumption II.3 is easy to see as reasonable. Even if the content posted is not owned by the user — that is, a link to some other content — the user is still introducing the content to the community and, as a result, acts as the influencing agent. It is not apparent at first glance that Assumption II.4 is reasonable. The primary problem with deleted users and content is that the data only shows these with names "[deleted]". This results in a node with very high connectivity where there may not be any. For example, a user on post A may delete her comment while user B deletes his comment on another post. Both these comments will have author "[deleted]", translating what was two distinct authors into a single author. This can skew any analysis of the graph. The consequence is mentioned in Section IV.

B. The Structure of Reddit

Reddit is a platform for users to submit content, named submissions, and to have other users comment on them. These submission come equipped with a title and either a body text, media or link to external content. These submissions are organized into large communities, known as subreddits, that which users may subscribe to. The structure can be visualized as in Figure 1.

Reddit also relies on a voting system to rank content. These are known as upvotes (or downvotes for negative votes). These votes can be placed by users on submissions or comments in order to inform other users about the value of said content. Of course, this is not the only purpose (or even the one most intended) it is used for. Often users will downvote or upvote simply as a way of expressing disagreement or agreement. The resulting score of a given content is arrived at by a sum of the upvotes and downvotes, and various "streams" are constructed to sort based on these values.

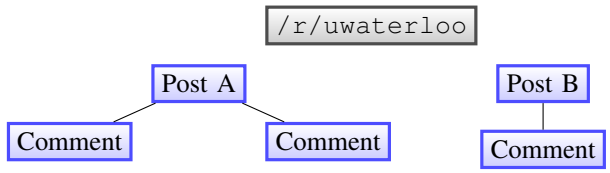

Fig. 1. Graph-like structure of the Reddit platform.

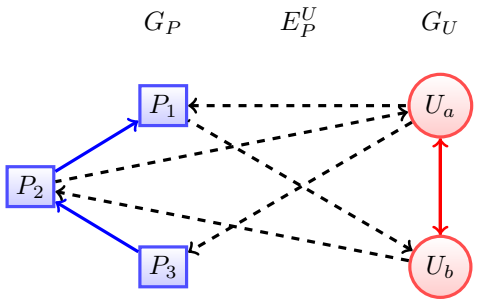

Fig. 2. Graph structure showing the relationship between users and submissions/comments. G_P are those nodes and edges that are blue. G_U are those nodes and edges that are red. E_P^U is the set of edges denoted by broken lines.

C. Graph Model

Two models are used for analysis. The first models how influential content or users are on a given subreddit as a connection between two graphs. A visualization of the structure can be found in Figure 2. Graph G_P is the graph that captures the inherent structural relationship between submissions/comments. Assumptions II.1 and II.2 imply that a given submission is influenced by its comments. More concretely, any parent content is made more interesting by its children. These are visualized as the blue edges in Figure 2. This is augmented using edges that relate users to submissions/comments. The second model attempts to capture which users converse with each other. This graph, G_U, can be thought of as being generated by the graph created above.

1) Graph Model for Centrality Analysis: Denote the set of users by U and the set of submissions and comments by P. We first define the existence of the digraph $G_P = (P, E_P)$ that denotes the relationship between submissions and comments. E_P is constructed in the manner described in Algorithm II.1. We give edges in E_P a universal weight of 1.

Algorithm II.1: Constructing E_P.

forall `parent` $\in P$ **do**
  for `child` $\in P \cap ChildrenOf(P)$ **do**
    $E_P \leftarrow E_P \cup \{(\texttt{parent}, \texttt{child})\}$;

Weighting using upvotes was considered but was determined to yield results that were not desireable. For one, it was not clear that upvotes were a meaningful method of ranking interesting content. Further complexity arises when considering controversial content, which has a mix of upvotes and downvotes. A highly controversial topic may be highly interesting, which contradicts the idea that upvotes correlate with interest. Possibly given knowledge of the number of voters this problem could be alleviated. However, since not all subreddits release data about how many users voted, it is simpler to reduce the complexity by eliminating any non-trivial weighting method. Observe that G_P is a forest where the submissions act as roots of the trees.

This structure is not interesting by itself and we need to extend it in order to facilitate analysis. The set of edges E_P^U, that connects submissions/comments to users, is what brings the significant cross-post structure. E_P^U is generated in the manner described by Algorithm II.2. The rules are that: a comment author is interested in a submission (root in G_P) and a comment author makes a comment interesting. A more convoluted structure may be derived by instead having the comment author be interested in the immediate parent comment. This structure was not investigated and could yield differing results. The argument for choosing the structure defined here is that authors may bring the whole context of the submission to a given comment, and not just the parent comment. The directed edge from the author to the submission allows us to model this sort of behaviour. That is, the user is interested by all the content on the submission page. This does ignore temporal effects, and so could yield incorrect results. However, it leaves out the ability to model any sort of larger contextual awareness the user may have. On the other hand, the model not investigated is not as susceptible to temporal effects as it directly captures the temporal structure. The structure described above may be

Algorithm II.2: Constructing E_P^U.

forall `user` $\in U$ **do**
  for $p \in P$ *and* `user` $= AuthorOf(p)$ **do**
    `root` $\leftarrow RootOf(p)$;
    $E_P^U \leftarrow E_P^U \cup \{(p, \texttt{user}), (\texttt{user}, \texttt{root})\}$;

augmented with self-edges for all users and aperiodicity of the resulting graph can be guaranteed.

Claim II.1. Let the set of vertices be $N = P \cup U$. We also let $E_N = E_P \cup E_P^U \cup \hat{E}$ be a set of edges, where $\hat{E}$ is the set of directed self-edges for all users in U. Then $G' = (N, E_N)$ is an aperiodic graph.

Proof. Easy to see by the fact that cycles of any length may be constructed by including the users in the cycles. The simplest cycle may be constructed (see Figure 2) using a commenting user and the associated submission. □

2) Graph Model for Cluster Analysis: Submissions and comments are not directly significant for identifying subgroups within a community of users. The model used here is restricted to a weighted digraph $G_U = (U, E_U, W_U)$. However, we desire a method of constructing relationships between users that capture "interaction" between users. Algorithm II.3 defines one such construction. It captures the

idea that a user Bob is interested in the content user Jenn creates if Bob comments a lot on content Jenn submits. The algorithm adds an additional requirement: that the interactions are unique modulo the root submission. This additional restrition attempts to count the number of "conversations" taking place between users without the need for explicit start and stop points.

Algorithm II.3: Constructing E_U.

```
A_{:,:} ← 0;
forall p ∈ RootsOf(P) do
    conversations ← ∅;
    parentauthor ← AuthorOf(p);
    for p_2 ∈ ChildrenOf(p) do
        author ← AuthorOf(p_2);
        if (author, parentauthor) ∉
         conversations then
            A_{author,parentauthor} =
             A_{author,parentauthor} + 1;
            conversations ← conversations ∪
             {(author, parentauthor)};
```

III. IMPLEMENTATION

There are two methods of collecting Reddit submission and comment data. There is a live application programming interface (API) available for use, that provides live up-to-date data. The most natural API request available for downloading submissions is limited to the last 1000 submissions on a given subreddit. This substantially limits the amount of data available for collection; for some subreddits this constitutes only a few weeks of data. Other methods to getting this old data exist, but involve using the API in a convoluted manner. Baumgartner addressed this problem and collated all data on Reddit for over ten years for the purpose of data analytics. The data can be found online, freely available for use, in an archived format [1]. Submissions and comments are stored in separate archives, organized by months or days. The data is stored in these files as compressed, line-delimited Javascript Objects (JSON).

The analysis focusses on data for the time frame starting at November 2017 and ending in February 2018. I developed a Python script that filters out unnecessary data from these archives and inserts the useful data into a Redis database, a fast in-memory key-value database [2]. The software developed for this project and used for analysis is publicly available on Github [3].

A. Centrality Analysis

For measuring the interest of any given user or content, I propose using the Katz centrality measure. Centrality measures provide a measure of how "central" a node in a graph is. The simplest of measures simply counts the in-degree of a node. The Katz centrality measure is defined as,

Definition III.1. Let $A \in \mathbb{R}^{n\times n}$ be an adjacency matrix for an aperiodic graph G. Let $0 < \alpha < \rho(A)$. The Katz centrality measure for node i is defined as,

$$c_i = \sum_{k=1}^{\infty}\sum_{j=1}^{n} \alpha^k \left(A^k\right)_{j,i}.$$

This measure counts the number of paths from all nodes in the graph to a given node i, scoring paths of shorter length with a higher weight. The graph G' was shown in Claim [**?**] to be aperiodic and so we can use the Katz centrality measure on the induced adjacency matrix. Since G' is unweighted, we choose the natural weighting of 1 for all edges in the adjacency matrix. The implementation is a custom script that employs an iterative technique converging to the Katz centrality measure for all nodes [4].

B. Cluster Analysis

Spectral clustering is a family of techniques used for identifying disconnected components in a graph or, more generally, clusters of highly connected nodes using the graph Laplacian matrix. A survey of these techniques can be found in [5]. I instead follow a variation of the simplest technique described in [6], [7]. We recall significant results used for spectral clustering now,

Definition III.2. Let $A \in \mathbb{R}^{n\times n}$ and $D \in \mathbb{R}^{n\times n}$ be the adjacency and out-degree matrix for the graph G_U. The graph Laplacian is defined as,

$$L = D - A.$$

The normalized graph Laplacian is defined as,

$$L_{rw} = I - D^{-1}A.$$

Theorem III.1 (Properties of the Graph Laplacian [4])**.** The graph Laplacian and the normalized graph Laplacian enjoy the following properties,

(a) Say graph G_U has n nodes with k connected components. Then

$$\operatorname{rank} L_{rw} = \operatorname{rank} L = n - k.$$

(b) The vector of all ones, $\mathbf{1} \in \mathbb{R}^n$, is in $\ker L_{rw} = \ker L$. This vector evalutes the membership of all nodes to the graph.

(c) The eigenvalues λ_i of L and L_{rw} can be arranged in the following manner,

$$0 = \Re\{\lambda_1\} \leq \Re\{\lambda_2\} \leq \cdots \leq \Re\{\lambda_n\}.$$

The methodology to find a cluster within a graph is to bootstrap a clustering technique used in $\mathbb{R}^m$, the k-means algorithm, by the eigenvectors of L_{rw}. This relies on the observation that the eigenvectors associated with eigenvalues at 0 are those that denote membership to a given connected component of the graph. Moreover, eigenvectors near zero can be thought of as measuring a partial membership to a cluster. As a result, we wish to find the eigenvectors of L_{rw} whose eigenvalues are near 0. This is not the best approach due to numerical instability of the eigenvalue algorithms used

for sparse matrices. The choice to use L_{rw} alleviates the problems introduced by having a poor condition number, however we need to also employ a technique discussed by Frederix and Barel. They suggest instead looking at the matrix [8],

$$\hat{L}_{rw} = D^{-1}A.$$

The matrix $\hat{L}_{rw}$ has the same eigenvectors as L_{rw} while the eigenvalues shift. The eigenvalues instead move to be less than, or equal to, 1. This reformulation turns out to be much better suited to the iterative numerical algorithms implemented to find eigenvalues of large sparse matrices.

IV. ANALYSIS

One subreddit of particular interest to myself is the `/r/uwaterloo` subreddit. This community consists to date of over 26 thousand subscribers. The subreddit is famous for a goose-related subculture that resulted in a man receiving a goose tattoo on an unsavoury location of his body [9]. It also is home to a top comment created by lecturer Dave Tompkins that was made in response to a very strong criticism of his teaching style [10]. Both posts made it to other more public subreddits, `/r/all` and `/r/bestof`. The latter comment was made in the time frame downloaded for the purpose of this analysis and so we expect either the comment, submission or user to show up in the interest analysis. Two other subreddits are evaluated: `/r/science` and `/r/CanadianInvestor`. The former is a large community where users post links to novel research for users to discuss. The latter is a more niche community where users can discuss investment strategies and get advice.

A. Interest Analysis

Figure 3 contains the top ten nodes — scored by the Katz centrality measure — in graph G' as constructed for each individual subreddit of concern. First, we observe that t3_7w0dgv is the post that sparked the response by Dave Tompkins; this, paired with personal observation of familiar users with significant reach in the community, confirms that the methodology used is leading to reasonable results.

Observe that `/r/science` top most interesting content is actually a submission, unlike the other two subreddits. This is easily explained by the nature of the community. The community is designed around submitting links to journal articles that can be discussed by other interested users. The community is large, with over 19 million subscribers. Not only this, but the sheer number of posts is breath-takingly large. So a user would have to be highly active in order to not be washed out by the number of other users and content interacting in the subreddit. As it turns out, the only user that made the top 10 is actually a moderator for this forum, `/u/mvea`. Unsurprisingly, the post that received the highest score is t3_7e1jo1, which is a post that discussed the implications of the new Trump tax policy on graduate students. This content garnered 100 thousand upvotes with nearly 11 thousand comments. The vast majority of active users on the subreddit were interested in this post, which isn't surprising; its content was pointed and of great relevance to the future of the community.

Notably a bot, the `AutoModerator`, made it to the top of the ranking for `/r/science`. This may be because it comments a lot on a number of posts. It may be of value to detect bots and remove them from analysis and the author is displeased that he forgot about them.

Now we address Assumption II.4. It can be verified that the analysis including deleted users does not substantially permute the above ranking. In fact, for the top 10, no changes besides a shift in rank is observed for all three subreddits. Had there been a significant change in the ranking and it can be argued any results may be deemed uninsightful due to a lack of important information.

B. Cluster Analysis

Figures 4, 5, 6 show the results of applying the spectral clustering analysis discussed in Section III-B. Qualitatively we observe that the Waterloo subreddit has subclusters but no valuable insight can be gained on their nature. The most interesting of insights can be made in Figure 6. In Section IV-A we noted how the top 10 nodes were dominated by users. This is reflected in the spectral cluster, where a small group of users interact with almost every other user involved in the community. The spectral clusters for the science subreddit also yield interesting, very tight clusters but this warrants further investigation.

ACKNOWLEDGEMENT

The author thanks Professor John Simpson for offering the special topics course on Networked and Distributed control, thereby providing the essential background for this analysis. The author also gives a special thanks to Jason Baumgartner for uploading archived Reddit API data to [1].

REFERENCES

[1] J. Baumgartner, "Directory contents." [Online]. Available: https://files.pushshift.io/reddit/

[2] "Redis," July 2017. [Online]. Available: https://redis.io/

[3] R. D'Souza, "rollends/red-prism." [Online]. Available: https://github.com/rollends/red-prism

[4] F. Bullo, *Lectures on Network Systems.*, 1st ed. CreateSpace, 2018. [Online]. Available: http://motion.me.ucsb.edu/book-lns

[5] M. C. Nascimento and A. C. de Carvalho, "Spectral methods for graph clustering – a survey," *European Journal of Operational Research*, pp. 221–231, 2011. [Online]. Available: doi:10.1016/j.ejor.2010.08.012

[6] R. Horaud, "A short tutorial on graph laplacians, laplacian embedding, and spectral clustering." [Online]. Available: https://csustan.csustan.edu/~tom/Clustering/GraphLaplacian-tutorial.pdf

[7] U. von Luxburg, "A tutorial on spectral clustering." *Statistics and Computing*, vol. 17, no. 4, 2007.

[8] K. Frederix and M. V. Barel, "Sparse spectral clustering method based on the incomplete cholesky decomposition." *Journal of Computational and Applied Mathematics*, Jan. 2012. [Online]. Available: doi:10.1016/j.cam.2012.07.019

[9] `/u/SsseCn8jx`. Fuck it, if this post receives 2000 upvotes i will tattoo a goose on my ass. [Online]. Available: https://www.reddit.com/r/uwaterloo/comments/5ibr95

[10] `/u/[deleted]`. Dave tompkins is overrated. [Online]. Available: https://www.reddit.com/r/uwaterloo/comments/7w0dgv/

/r/uwaterloo	/r/science	/r/CanadianInvestor
supersonic63	t3_7e1jo1	johnnychi
kw2002	AutoModerator	EyesOnTsx
t3_7f2c1d	mvea	jnf_goonie
t3_7w0dgv	t3_7vxito	langlois44
t3_7i045q	t3_7s6a9z	t3_7ii8af
microwavemasterrace	t3_7my58a	t3_7fapx7
uwsmile	t3_7okc7u	t3_7pawu6
t3_7ld2jb	t3_7hs6ro	helkish
mywaterlooaccount	t3_7sv8vb	Ginhisf
honhonhonFRFR	t3_7qs5xz	MarketStorm

Fig. 3. Top scoring nodes under the Katz centrality measure for the `/r/uwaterloo`, `/r/science` and `/r/canadianinvestor` subreddits. Nodes prefixed with `t3_` and `t1_` are submissions and comments respectively. Other nodes are users. To translate these id's to a link, simply follow the pattern `https://reddit.com/r/{subreddit}/comments/{id}`.

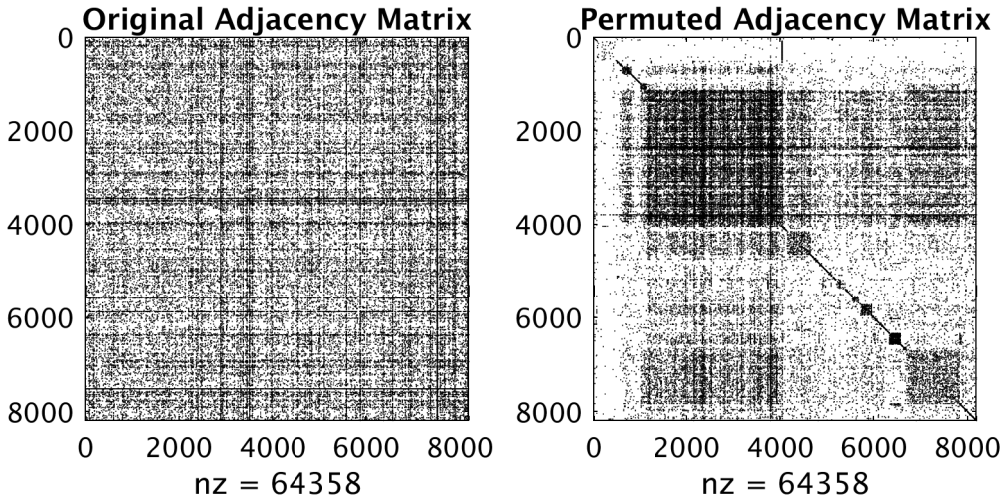

Fig. 4. Spectral clustering applied to reduced adjacency matrix for `/r/uwaterloo`.

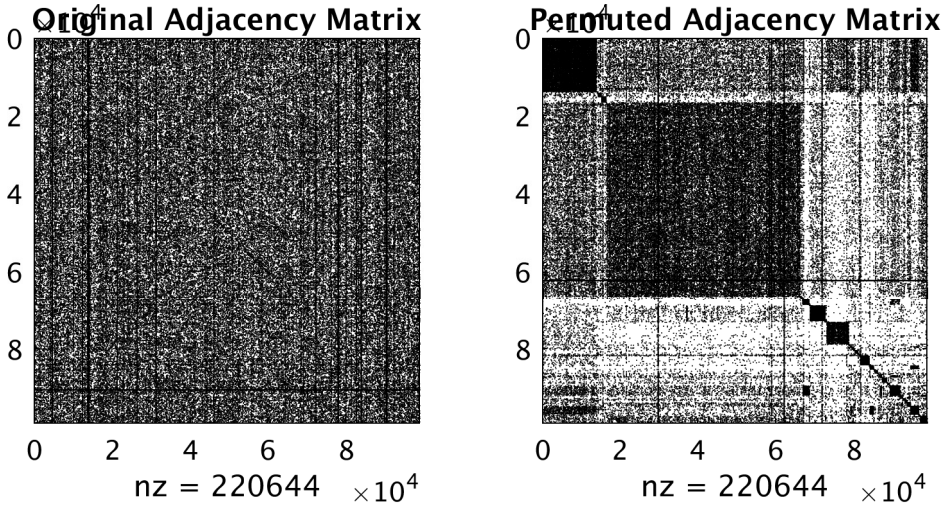

Fig. 5. Spectral clustering applied to reduced adjacency matrix for `/r/science`.

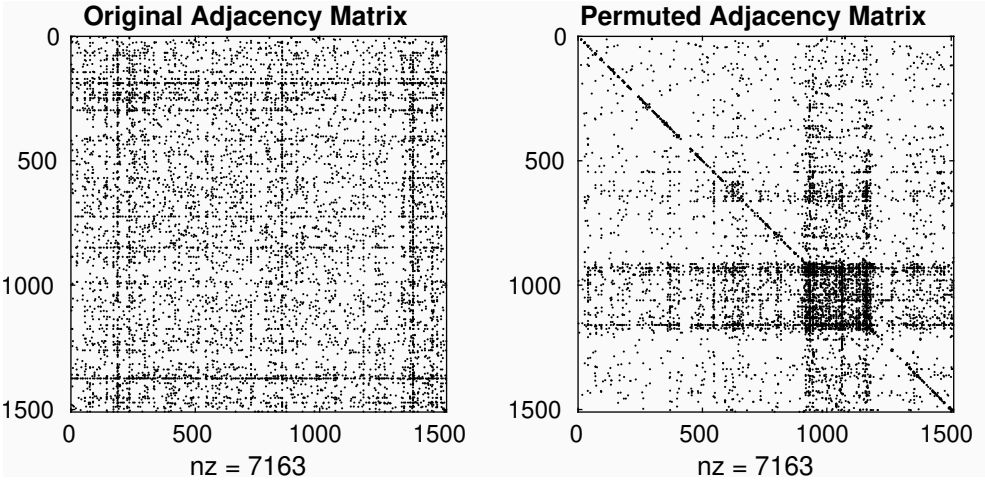

Fig. 6. Spectral clustering applied to reduced adjacency matrix for `/r/CanadianInvestor`.