On Detection of Influential Users and Content on Reddit*

Rollen S. D'Souza[1]

***Abstract*—**

I. INTRODUCTION

II. BACKGROUND

Talk about graph theory, centrality and Laplacians.

III. MODELLING

A. Assumptions

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

Assumption III.1. *Users are influenced (positively or negatively) by content they read.*

Assumption III.2. *Users comment only on content they have read.*

Assumption III.3. *Content is influenced by the content creator.*

Assumption III.4. *Deleted users and content are not influential and do not affect other users in a meaningful way.*

Assumptions III.1 and III.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 influence implied by the assumptions. Specifically these two assumptions imply that we can infer a user is influenced by content when we detect a comment made by the user. As a result, we take the first two assumptions as reasonable. Assumption III.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 III.4 is reasonable. We discuss this later in the work and will establish its validity.

*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

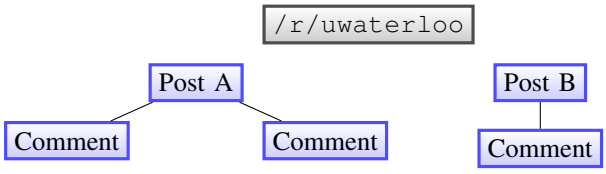

Fig. 1. Graph-like structure of the Reddit platform.

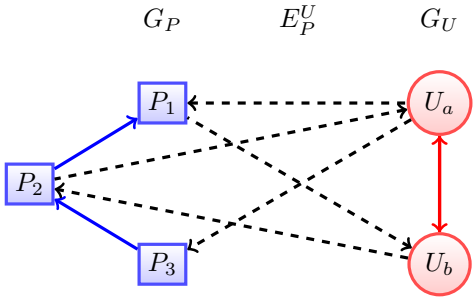

Fig. 2. Graph structure showing the relationship between users and submissions/comments. G_P are those nodes and edges that are blue. G_U are those nodes and edges that are red. E_P^U is the set of edges denoted by broken lines.

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. Explain other possible related structure. Upvotes and Downvotes. Controversial

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 III.1 and III.2 imply that a given submission is influenced by its comments. More generally, any parent content is directly influenced 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 III.1. We give edges in E_P a universal weight

Algorithm III.1: Constructing E_P.

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

of 1. 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 influence. Further complexity arises when considering controversial content, which has a mix of upvotes and downvotes. A highly controversial topic may be highly influential, which contradicts the idea that upvotes correlate with influence. 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 III.2. The rules are that: a comment author is influenced by the submission (root in G_P) and a comment author influences the comment. A more convoluted structure may be derived by instead having the comment author be influenced by 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 influenced by all content on the submission page. This does ignore temporal effects, and so could yield incorrect results. On the other hand, the model not investigated is not as susceptible to temporal effects as it directly captures the temporal structure. However, it leaves out the ability to model any sort of larger contextual awareness the user may have. The structure described above may be augmented with self-edges for all users and aperiodicity of the resulting graph can be guaranteed.

Algorithm III.2: Constructing E_P^U.

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

Claim III.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 III.3 defines one such construction. It captures the idea that a user Bob is influenced strongly by another user Jenn 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. Complete explanation of the graph used for cluster analysis.

Algorithm III.3: Constructing E_U.

forall $p \in RootsOf(P)$ **do**
  `conversations` $\leftarrow \emptyset$;
  for $p_2 \in ChildrenOf(p)$ **do**
    ;

IV. IMPLEMENTATION

The algorithms used

V. ANALYSIS

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 Reddit user who did this for uploading archived Reddit API data to website please. [1]

REFERENCES

[1] *Telemetry Channel Coding*, ser. Blue Book, No. 4, Consulative Committee for Space Data Systems (CCSDS) Recommendation for Space Data System Standard 101.0-B-4, May 1999. [Online]. Available: http://www.ccsds.org/documents/pdf/CCSDS-101.0-B-4.pdf