Cascading Behaviors and Information Propagation On Reddit Network

Sunjeet Jena, Rajath Manjunath, Ankit Sharma, Aayush Sharma, Oluwaseun Talabi, Akhil Mittal

Arizona State University, Tempe, AZ.

1. Description

Reddit has grown to be one of the most popular discussion and news aggregator websites [1]. Reddit is broken up into more than a million communities known as “subreddits,” each of which covers a topic, but at the same time are linked to each other because of overlapping subjects and common users (Fig 1). They become a channel of propagation of information when a post at one subreddit is linked or tagged in another subreddit. How are these posts shared between subreddits? How does the popularity of the reddit posts change over time? How does a subreddit in turn influence another subreddit? We attempt to answer these questions by building a model to predict the propagation probability of a post. This will allow us to generate realistic cascades that will help us to assess how posts would link together between different subreddits as well as how their popularity changes over time. There is a strong analogy between a post on subreddit and a post on a blog and how a post on a subreddit links to another subreddit is very similar to how posts are shared between different blogs. Due to this analogy, we will extend and validate the cascade generation model described in paper Cascading Behavior in Large Blog Graphs[3] on the Reddit Hyperlink Network dataset[2].

Motivated by the work done by Leskovec et al. [3] we model two bi-directional graphs based on Social Network: Reddit Hyperlink Network dataset [2]. The first graph is modelled in analogous to the “Blogosphere” graph in [3]. For the purpose of readability and analysis we name our graph as “Subreddit-Sphere”. This graph models all hyperlink connections among all the subreddits in our dataset. Figure 2(a) represents an abstract model of the “Subreddit-Sphere”. In particular, the directed edge connection represents the hyperlink connection between a post in source blog pointing to posts in destination blog. Each colored dotted sphere represents a post in the subreddit. For example in figure 2(a) subreddit SR_1 has a post that hyperlinks to one post in subreddit SR_2 and SR_4 has one post that hyperlinks to a post in SR_1.

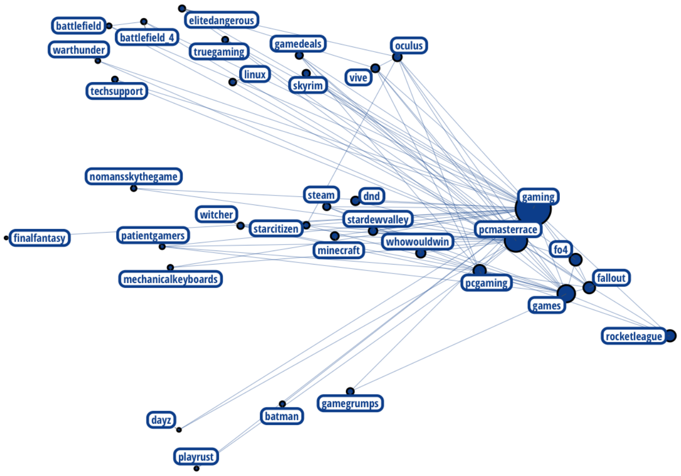

Fig. 1 An example graph of different subreddits representing hyperlink connections between them. It's clearly seen that most of the subreddits have some hyperlinks to the ‘Gaming’ subreddit.[4]

We also model another graph, again analogous to the “Blog-network” graph in [3]. For the purpose of readability and analysis we name our graph as “Subreddit-network” Fig 2(b). Nodes (squares) in our graph represent a subreddit, a directed and weighted edge represents the sum of all references/hyperlinks from the source blog pointing to the destination blog. An edge weight represents the sum of the number of hyperlinks between two subreddits. For example: if subreddit B references x number of posts in subreddit A, then directional edge $E(B -> A)$ has weight x Fig 2 (b) .

We are investigating the topological pattern of the references between subreddits and the probability of a new post in a subreddit (node) being referenced by other nodes (direct or indirect). The previous work[3], used different values of probabilities by trial-and-error approach to validate their model. We plan to use statistical machine learning to learn this susceptible probability (probability of a link) between two subreddits and generate new cascading patterns that shall be validated against the reddit hyperlink dataset. In the previous work done by Leskovec et al.[3], this susceptible parameter is constant irrespective of edge weights and the starting point but we wish to implement a method which incorporates edge-weights as a parameter/feature for the calculation of the susceptible probability.

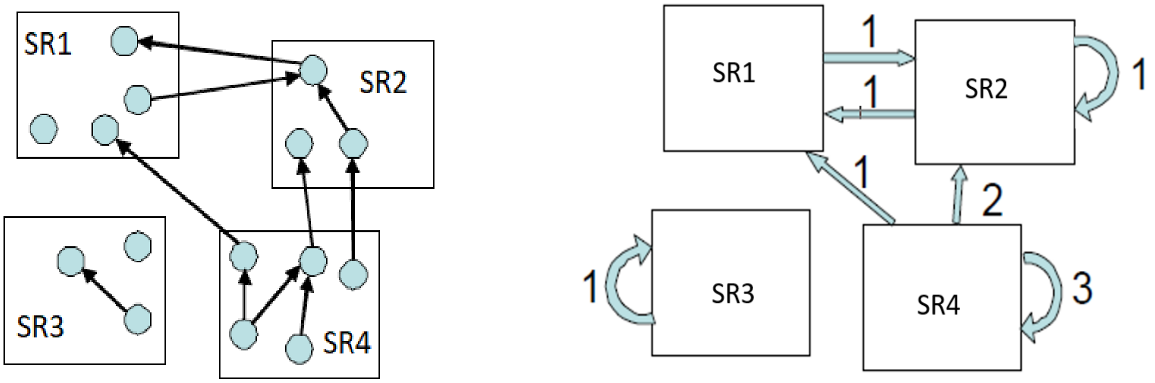

Fig. 2 (a) Subreddit-Sphere

Fig. 2 (b)Subreddit-network

Network of different posts from different subreddits. Each square represents a subreddit, each dot represents a post in a subreddit and the edge direction represents the hyperlinks to the post. The edge weight represents the total number of hyperlinks referenced by one subreddit from another subreddit.

Given a single root node that represents a subreddit, we wish to predict the probability of susceptibility that the subreddits which had earlier hyperlinked any post in the past shall also hyperlink a new post in the root-subreddit. In particular we uniformly pick a node (subreddit) ***'s'*** in the "Subreddit-Network" as a starting point of the cascade and set its state to as infected (i.e. a new post) and that add that new node to the cascade graph. This subreddit, which is now in an infected state, infects each of its neighbouring uninfected nodes **(s_{11} , s_{12} , s_{13} .... s_{1n})** with some susceptible probability **β**. We add this new set of infected nodes to our cascade and set the state of the node (subreddit) ***'s'*** to uninfected .We continue this method recursively until no new nodes are added to the cascade.

Please note that the above algorithm for the cascade generation is heavily motivated from the algorithm mentioned in earlier work by Leskovec et al. [3] . Our main contribution in this project shall be to experiment with the susceptible probability **β** and use standard machine learning techniques to predict this susceptible probability for a given dataset and incorporate additional parameters/features such as edge weights which was considered as an invariant in the work done by Leskovec et al. [3]

2. Preliminary Plans (Milestones) :

1) Data collection (October 1st - October 6th).
2) Data preprocessing (October 7th - October 13th).
3) Data Visualisation using Graph Generation. (October 14th - October 21st)
4) Additional research on Machine Learning Algorithms. (October 22nd - October 26th)
5) Code Implementation of the model. (October 27th - November 4th)
6) Training of the model.(November 5th- November 7th)
7) Validation of the model. (November 7th - November 10th)
8) Making possible improvements/revisions to the model (November 10th- November 12th)
9) Report Writing (November 12th - November 20th).

3. Team Members and Roles:

Sunjeet Jena - Did initial research on the topic, brainstormed new possible applications of the algorithm, and shall be implementing the machine learning algorithm for predicting the susceptible parameter.

Rajath Manjunath - Did initial research on the topic and shall be working on the data preprocessing and graph generation.

Ankit Sharma - Shall be working on data analysis and observations of patterns in the graph.

Aayush Sharma - Did initial research on the topic and shall be working on the data analysis and data preprocessing.

Oluwaseun Talabi - Shall be working on data collection and data processing.

Akhil Mittal - Shall be working on data collection and data processing.

4. References:

[1] https://en.wikipedia.org/wiki/Reddit

[2] Social Network: Reddit Hyperlink Network https://snap.stanford.edu/data/soc-RedditHyperlinks.html

[3] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. Cascading Behavior in Large Blog Graphs. *Proceedings of the 2007 SIAM International Conference on Data Mining* (SDM). 2007, 551-556

[4]https://minimaxir.com/2016/05/reddit-graph/group-008_hub305a5241915c816f5c26d5026f16c55_288782_1200x1200_fit_gaussian_2.png