Modelling subreddit interactions by activity overlap

N. Aleksić[1], E. Pajić[1], P. Obradović[1], W. Power[2], M. Mišić[1,*], Z.Obradović[2,*]

[1] University of Belgrade, School of Electrical Engineering, Bulevar Kralja Aleksandra 73, Belgrade 11020, Serbia
Email: an203015m@student.etf.bg.ac.rs, pe203013m@student.etf.bg.ac.rs, predrag.obradovic@etf.bg.ac.rs, marko.misic@etf.bg.ac.rs
[2] Data Analytics and Biomedical Informatics Center, Temple University, 1925 N. 12th St, Philadelphia, PA 19122, United States
Email: tug00038@temple.edu, zoran.obradovic@temple.edu
[*] Corresponding authors

Abstract

Machine learning methods often require using an appropriate graph representation to discover useful knowledge in large social networks. Here, a method is proposed of inducing a Reddit graph model composed of subreddit nodes, and edges representing a measure of user-activity overlap. In the proposed process a bipartite user-subreddit network is generated and a novel projection is applied to induce a subreddit-subreddit network. An approximate version of the Bipartite Configuration Model (WABiCM) is described and used to extract the backbone of the projected networks. This is done separately for comments and submissions and the constructed networks are analyzed.

Key words: social network analysis, bipartite network projection, backbone extraction

1 Introduction

Reddit is a well-known social networking website where users can create submissions, and post comments, images, and videos. It is organized as a set of communities, called *subreddits*, where people gather and discuss a common topic of interest. Several approaches to create useful Reddit networks have been discussed in open literature. Hamilton et al. model the graph by looking at cross-posting between subreddits [1]. Olson and Neal [2] create the graph using user activity overlap. Datta et al. [3] use user overlap and textual similarity and compare both approaches in their work. All of these approaches begin by considering a bipartite subreddit-user network and projecing it to a unipartite subreddit-subreddit network. In certain publications [2][4], the post count for each user are mostly ignored and the focus was to look if the user posted on the subreddit, while there are also approaches (e.g. [3]) that looks at the user post count on each subreddit and normalizes it in TF-IDF fashion. A common step in network analysis after bipartite graph projection is to prune the statistically insignificant edges or extract the important ones. However, many methods assume that the starting bipartite graph is unweighted, which is not the case if we want to consider per-subreddit user post count.

Our study investigates the use of an aggregation strategy for edge weights based on the Two-pass bipartite network projection [5] and we propose a weighted approximate version of the Bipartite Configuration Model [6] (WABiCM) to prune the edges. We used a high performance computing system with 2 Intel Xeon Silver 4214 CPUs and 192 GB of RAM. The large amount of RAM was used to store the graphs during compute.

2 Methodology

2.1 Network modelling

To obtain Reddit data, we use the publicly available Pushshift Reddit API [7] which collects daily Reddit data. Though it is possible to use this to extract the most recent site activity, this work makes use of the provided historical monthly data dumps to accommodate resource limitations. We constrain our analysis to January 2021, where there were 258,647 active subreddits and around 10M active users according to comments data, and 622,336 active subreddits and around 5M active users, according to submissions data. Many subreddits don't have any comments, and many users only post comments. 4 We represent Reddit as a bipartite graph of m subreddits and n users. Let $B \in \mathbb{R}^{m \times n}$ be a matrix where each element $B_{i,j}$ is the number of posts by the j-th user on the i-th subreddit. Our goal is to project this matrix into a $m \times m$ matrix A, an adjacency matrix between subreddits. The projection function should have the following properties when considering subreddits *S1* and *S2*: (a) if two subreddits have more common users (i.e. a larger fraction of users of *subreddit1* are common with *subreddit2*), edge weight between *S1* and *S2* should be larger; (b) if common users have a larger fraction of their posts on the two subreddits, the edge weight between *S1* and *S2* should be larger (*specialized users*); (c) edge weights should be normalized to be similar for different subreddit sizes.

Two-pass aggregation [5] redistributes weight to connected users and back to subreddits, splitting it proportionally to edge weights and therefore incorporates the assumptions we mentioned. Another appealing property of the two-pass aggregation approach is that edges take weights between 0 and 1. Based on this data, a graph built from comments has over 100M directed edges, while the one constructed from submission has over 19M directed edges. Other than being cumbersome and computationally complex to analyze due to sheer size, the obtained networks are very noisy. Therefore, backbone extraction is needed to denoise the data and extract information from the networks.

2.2 Backbone extraction by bipartite validation

To extract the statistically significant edges, we propose a weighted approximate version of the Bipartite Configuration Model (WABiCM). Statistical validation methods in network theory detect network properties that deviate from a benchmark null model. The null model is obtained by fixing bipartite graph node degrees and randomizing everything else. The statistical significance (p-value) of a unipartite edge weight w_{ij} is the probability that it is smaller than a weight obtained by picking a random graph from the null model and computing the edge weight with two-pass aggregation in it.

Some models, like the Bipartite Configuration Model (BiCM) [6] use a null model where graph node degrees are not hard constrained but have a mean equal to the actual degrees. We propose a model where we draw each bipartite weight from the Binomial distribution with suitable parameters. The weight distribution between a node i with degree N_i from the first set and a node a with degree M_a from the second set of nodes will be $B(N_i, P_a = \frac{M_a}{T})$, where T is the sum of all degrees from one set. It can be easily shown that ensemble mean degrees are equal to the actual degrees. For two-pass aggregation and our null model, edge weight distribution is a sum of $B(N_i, P_k) * B(N_j, P_k)$, where the sum goes over all nodes from the second set of nodes. This is hard to compute, hence we use empirical estimates. It should be noted that the above factor is the same for nodes from the second set that have the same degree, hence we can estimate the sum for these nodes using a normal distribution.

In our case, the distribution of user degrees is close to Pareto or Lévy, hence the above approximation speeds up computation for smaller degrees significantly. To prune the edges, we use statistical significance $p = 0.001$, calculated separately for comments and subreddits. Figure 1 shows thresholds depending on the subreddit size. Figure 2 shows the backbone of the graph after applying the derived threshold. After applying the thresholds, comments graph has about 10M directed edges, while submissions graph has about 3M directed edges, resulting in 91% and 84% reduction, respectively.

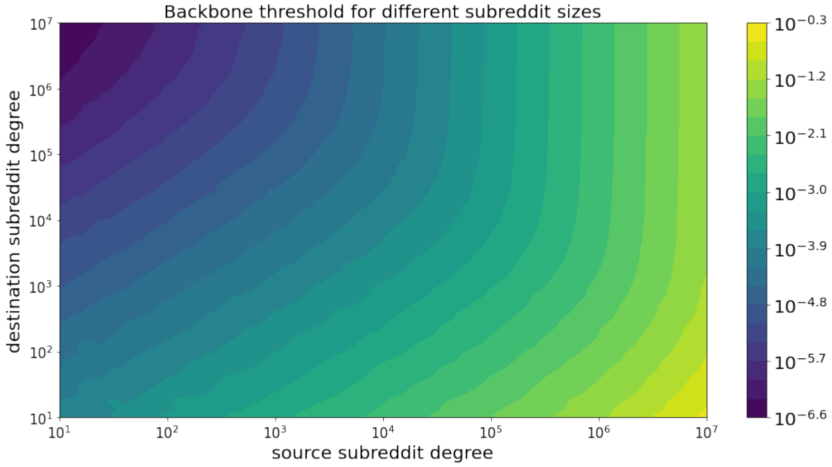

Figure 1: Threshold for edge weights in the subreddit network built from comment data, derived by the proposed Weighted Approximate Bipartite Configuration Model (WABiCM) for statistical significance of $p = 0.001$, plotted for different sizes of the source and destination subreddits. Subreddit size is measured by the degree of the corresponding node in the bipartite network. The degree of the source subreddit impacts the threshold more significantly.

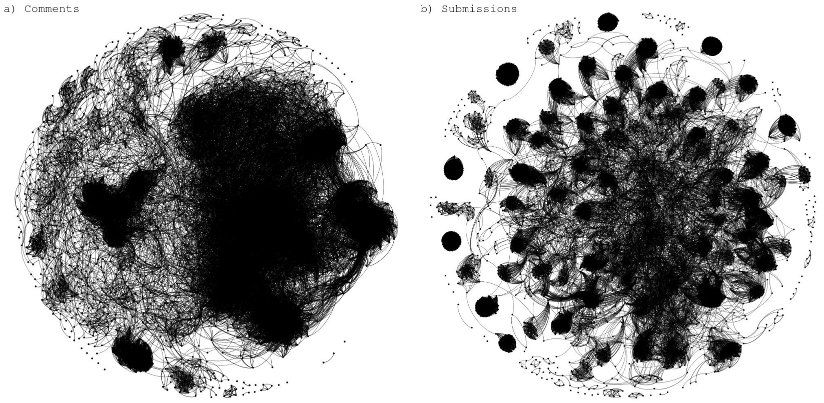

Figure 2: Reddit network models induced from comment data (left) and submissions data (right). Given the size of the data, thresholds were multiplied by 10 and only nodes with degrees greater than 50, were plotted. We can notice clusters in both graphs, but they are expressed more clearly in the submissions graph. We used Gephi[8] and the Fruchterman-Reingold layout[9] to visualize the graphs.

3 Analysis

Table 1 shows some graph metrics for both undirected backbone graphs created from comments and from submission activity. We have an edge in the undirected graph if there is at least one of the directed graph edges.

Metrics	Comments graph	Submissions graph
number of edges	5 393 213	1 807 372
number of connected components	231	2654
average node degree	85	29
max node degree	3856	1613
graph density	0.00067	0.00022
clustering coefficient	0.351	0.458

Table 1: Graph metrics of subreddit networks constructed from comments and submissions (around 126 000 non-isolated nodes in each).

4 Conclusions

We model a subreddit network using user activity overlap based on two pass aggregation and novel WABiCM network projection and derived thresholds. In an analysis of the impact of sizes of the source and destination subreddits on the threshold of the edge that connects them we found that the size of the source subreddit influences the threshold much more strongly. We construct and analyze separately two networks based on submission and comment data. The obtained results provide evidence that submissions network expresses clustering more strongly, howbeit with an order of magnitude more connected components.

References

[1] Hamilton WL, Ying R, Leskovec J, Inductive representation learning on large graphs, Proc. 31st Int'l Conf. Neural Information Processing Systems:1025–1035, 2017

[2] Olson RS, Neal ZP, Navigating the massive world of reddit: Using backbone networks to map user interests in social media, PeerJ Computer Science, 1(e4), 2015

[3] Datta S, Phelan C, Adar E, Identifying misaligned inter-group links and communities, Proc. ACM Conf. Human-Computer Interaction:1-23, 2017

[4] Martin T, community2vec: Vector representations of online communities encode semantic relationships, Proc. 2nd Work. NLP and Comp. Social Sci:27-31, 2017

[5] Zhou T, et al., Bipartite network projection and personal recommendation, Physical Review E, 76(4):046115, 2007

[6] Squartini T, Garlaschelli D, Analytical maximum-likelihood method to detect patterns in real networks, New Journal of Physics:13 083001, 2011

[7] Baumgartner J et al, The pushshift reddit dataset, Proc. Int'l AAAI Conf. Web and Social Media:830-839, 2020

[8] Bastian M, Heymann S, Jacomy M, Gephi: an open source software for exploring and manipulating networks, Proc. Int'l AAAI Conf. Weblogs and Social Media, 2009

[9] Fruchterman TMJ, Reingold EM, Graph drawing by force-directed placement, Software—Practice and Experience, vol. 21(1 1): 1129-1164, 1991