# Reddit: Discovering structure in comment networks

Shriyak Sridhar, Department of Statistics

### Introduction

Reddit is the fourth most popular website in the world [1] and with an average time of 16 minutes per day spent by a user on the site, it is the most addictive of the top twenty websites. A repository of every single Reddit comment posted is available at [2] and this project describes the structure of the comment networks. Users often comment on several subreddits; niche forums with a dedicated topic and similar topics of discussion are present on the site.

Attempts to cluster Reddit communities have been performed by Bustain and Golbeck [3] but they focus solely on user interactions. Weninger et al [4] look at the discussions to understand if it can be useful for search engines but none of the methods look at comments.. This work reinforces the findings from interactions by looking at the comment contents and various topics that arise in comments. Using comment networks, I draw a map of reddit and describe what the different parts of the site are discussing.

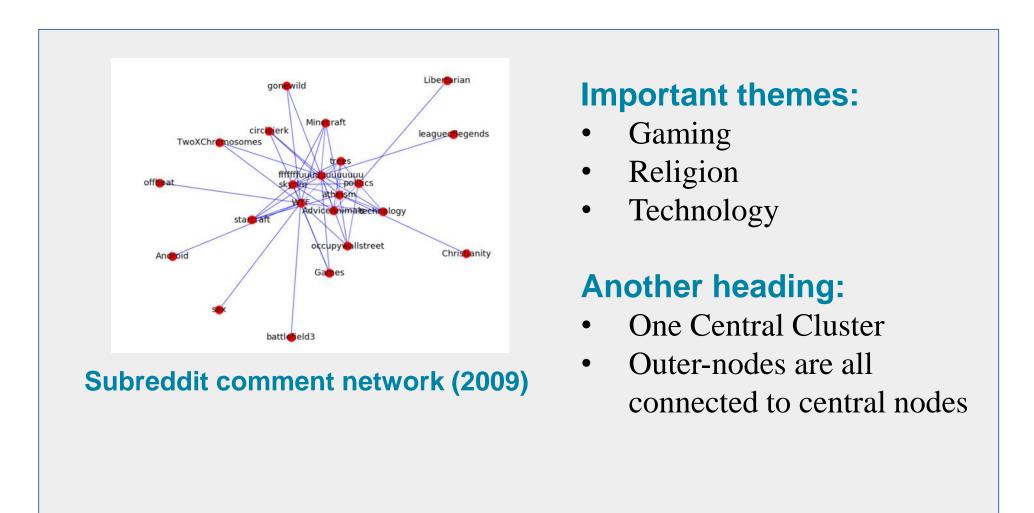
# Methods

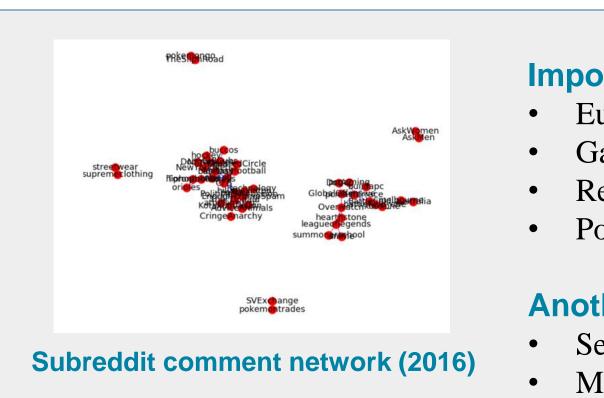
Interactions between different friends of a social network is often studied in but comment networks are a relatively unstudied. With Reddit, since there is no formal social circle for every user, we analyze the interactions of different users on the various subreddits they frequent. To analyze the interactions between two forums, we first look at the set of distinct users each of them contain. Once we have their user base, we look at the members common to both forums.

To obtain the final pairwise intergroup scores we use to map the distances between these groups, we also study the frequency of comments made by these set of users on their respective communities. This forms the foundation for the measure of interactions between two communities.

# Results

- As the site evolves, we see the emergence of certain cliques: Europe, Gaming, Sports, Politics, Fashion, Advice, Pokemon etc.
- In the 2016 comment network, the outer nodes are isolated and self-sustained.
- Using Guassian Mixture Models, we identify the presence of 7 comment clusters. These are depicted to the right.





<b>Important</b>	themes:
Furono	

- Politics

#### **Another heading:**

- Several Central Cluster
- Many isolated outer-nodes

Topic 1 - politics		Topic 2 - gaming		Topic 3 - <i>sports</i>		Topic 4 - other	
people	0.0057	good	0.0059	games	0.0036	lol	0.0083
trump	0.0045	play	0.0053	post	0.0031	f*ck	0.0059
make	0.0025	game	0.0051	play	0.0029	god	0.0040
money	0.0021	team	0.0046	time	0.0029	awesome	0.0039
clinton	0.0017	wow	0.0032	yeah	0.0028	man	0.0036
country	0.0016	kill	0.0026	work	0.0026	damn	0.0034
vote	0.0016	Damage	0.0025	years	0.0023	guy	0.0033

# 0.381756 AdviceAnimal 0.186261 0.222244

#### **Topic Modelling and Clustering**

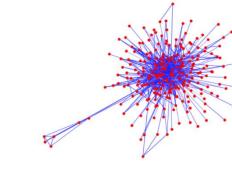
- Insert your text here. You can place your organizations logos on either side of the title of the poster. Insert your text here.
- Remember to size your font to fit your information into the space. The larger your font, the easier it will be for others to read your poster.
- Insert your text here. Remember to size your font to fit your information into the space.
- The larger your font, the easier it will be for others to read your poster. Insert your text here. You can place your organizations logos on either side of the title of the poster. Insert your text here. Insert your text here. Remember to size your font to fit your information

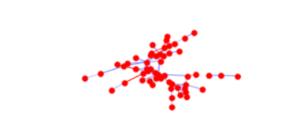
Training Program Components

# **Network Interactions**

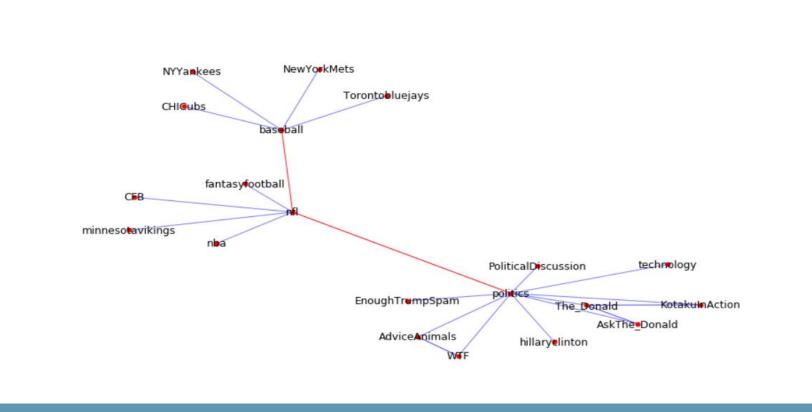
We have to finetune the level of interactions to capture using our weighted graph and as we reduce the threshold, more subreddits are represented. This might lead to a densely connected network without capturing the nuanced connections or outer clusters.







Looking at different geodesic paths, weidentify the set of nodes that appear frequently to find the centrality of a nodes; this shows its importance and role. In addition to the degree, closeness and betweenness centralities, we also observe the Katz centrality measure. Nfl, baseball and politics dominate the centrality scores and this shows their important role in the network as evidenced by the graph. They have scores of .71, .67 and .38 respectively as tabulated in the numerical studies.



#### References

- [1] https://www.alexa.com/topsites
- [2] https://bigquery.cloud.google.com/table/fh-
- bigquery:reddit\_comments.2017\_02?tab=details&pli=1
- [3] https://dl.acm.org/citation.cfm?id=2579231
- [4] https://ieeexplore.ieee.org/abstract/document/6785761
- [5] https://dl.acm.org/citation.cfm?id=2576943
- [6] https://www.ranks.nl/stopwords
- [7] https://github.com/ShriyakSridhar/RedditCommentNetwork