Reddit Network Analysis

Aditi Jorapur & Amrit Randev

Project Objective

- Analyze sets of the different nodes from the Reddit dataset to find the most important node (subreddit community) in the dataset

THE SUBREDDIT HYPERLINK NETWORK

This network represents the directed connections between 2 subreddits

- ***Reddit*** - Social news and discussion website for different communities
- ***Subreddits*** - niche groups/forums on Reddit
- Data Range - Reddit data of 2.5 years from Jan 2014 to April 2017
    - Data analyzed: January 1st - January 15th, 2014

How it works

- Based on posts that create a hyperlink from one Subreddit to another
- The Source Subreddit is where the hyperlink originates from and links to the Target Subreddit
- Directed and signed network

Reddit Hyperlink Network: Sample Set

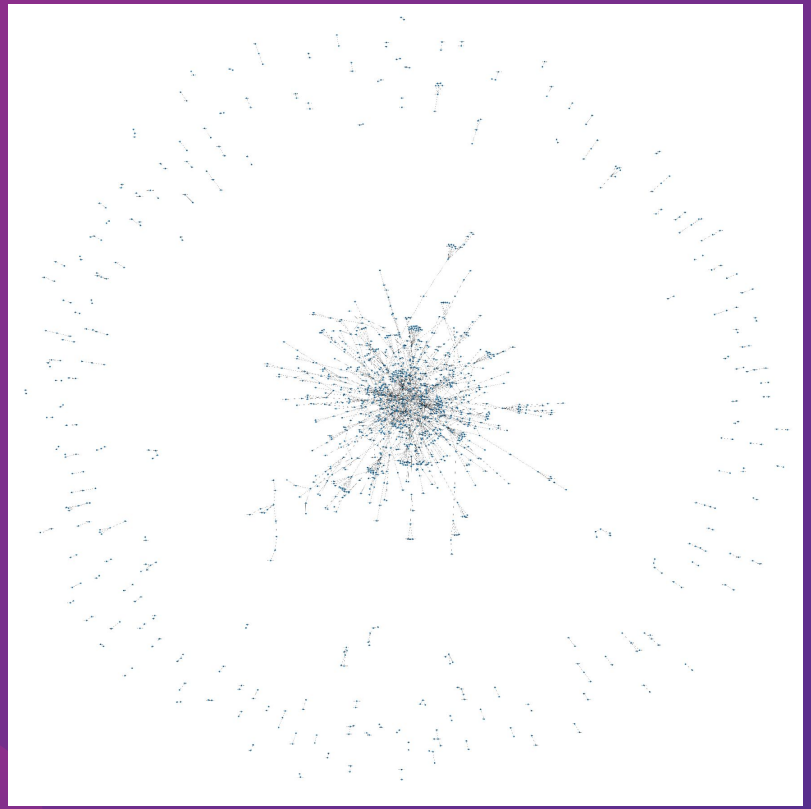
55,863
Nodes

858,490
Edges

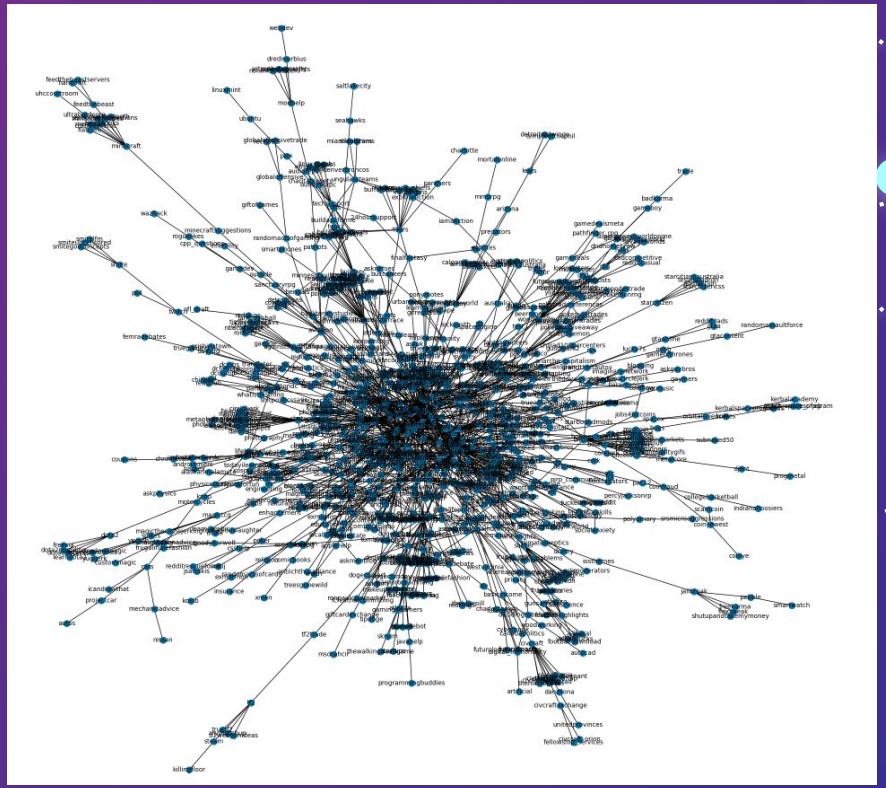
- Dataset was too big, so created a sample set: January 1st 2014 - January 15st 2014
- Each hyperlink is annotated with three properties: the timestamp, the sentiment of the source community post towards the target community post, and the text property vector of the source post.
- **Categories Used:** Timestamp, Source Community Post, Target Community Post, Sentiment of the Source Community post Towards Target Community Post
- Two datasets provided: Used Network of subreddit-to-subreddit hyperlinks extracted from hyperlinks in the **body** of the post.

Why This Dataset?

- Wanted to learn more about how the subreddits on Reddit are intertwined with each other
- Learn about how different interactions lead to hyperlinks between subreddits



Original Network Graph



Largest Connected Component Graph

Network Characterization

1454
Nodes

1617
Node Edges

181
Connected Components

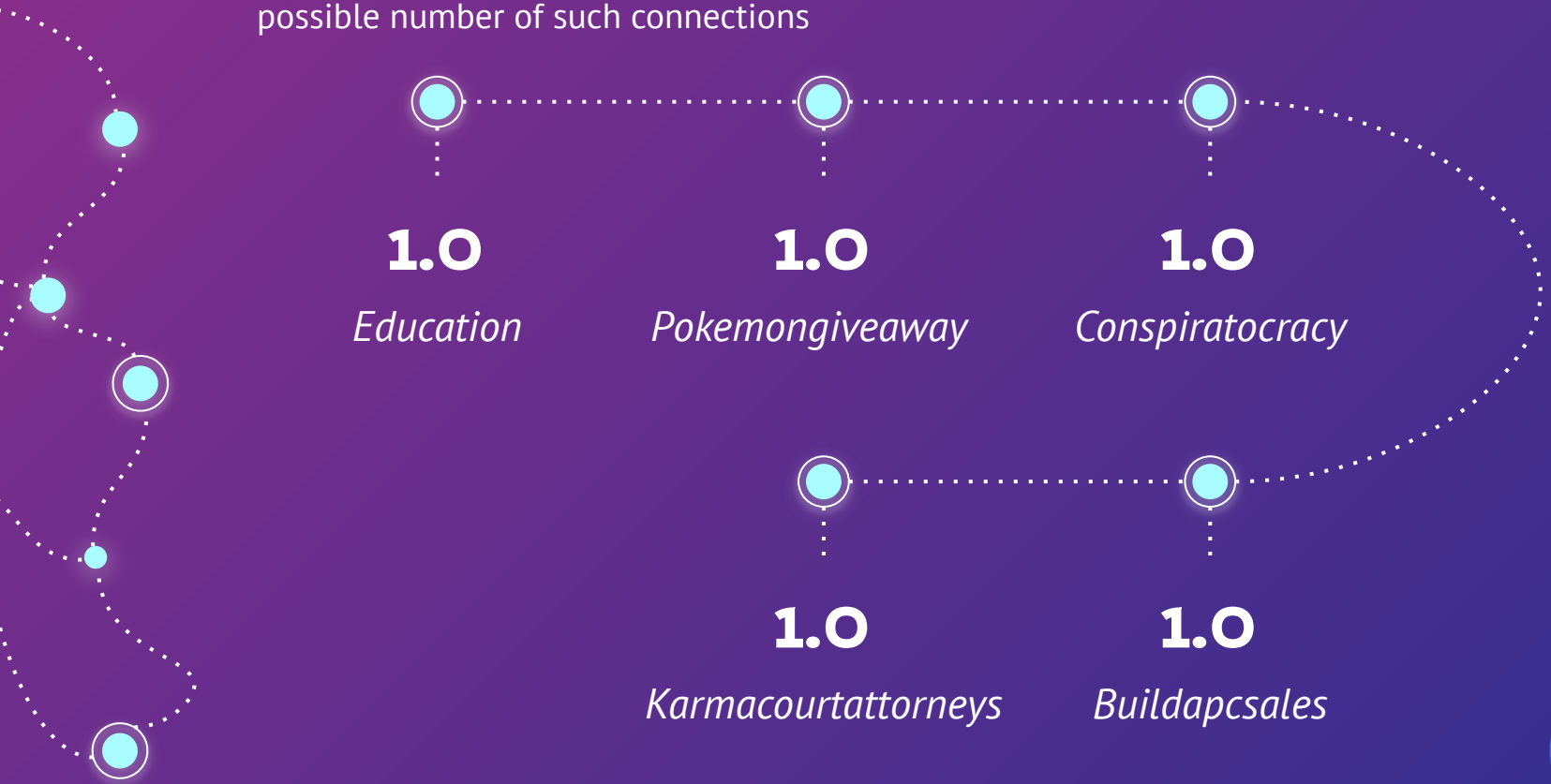
13
Diameter

1003
Diameter Nodes

1339
Diameter Edges

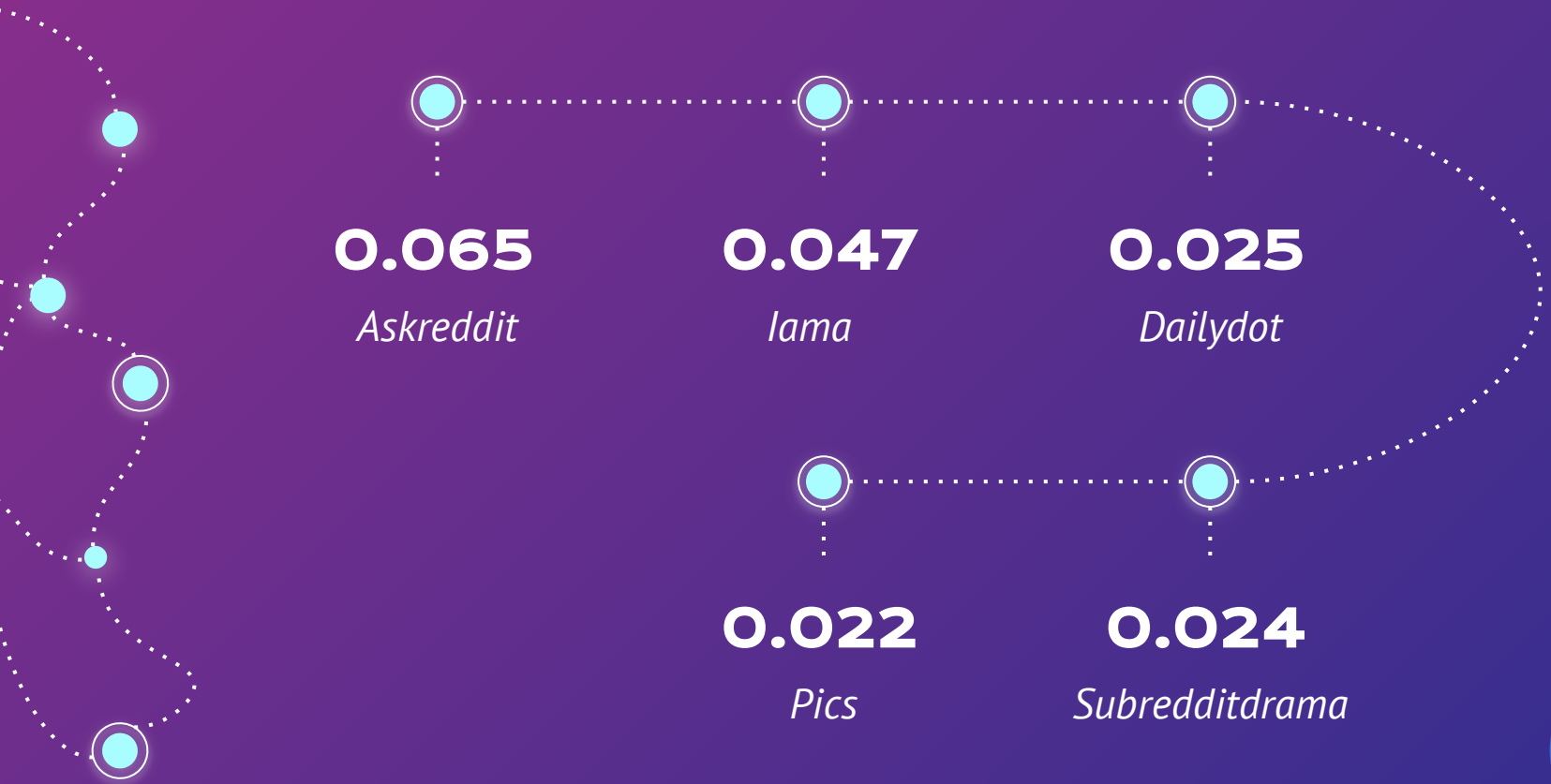
Highest Clustering Coefficients

- quantify the degree to which nodes in a network tend to cluster or form groups
- ratio of the number of connections between a node's neighbors to the maximum possible number of such connections



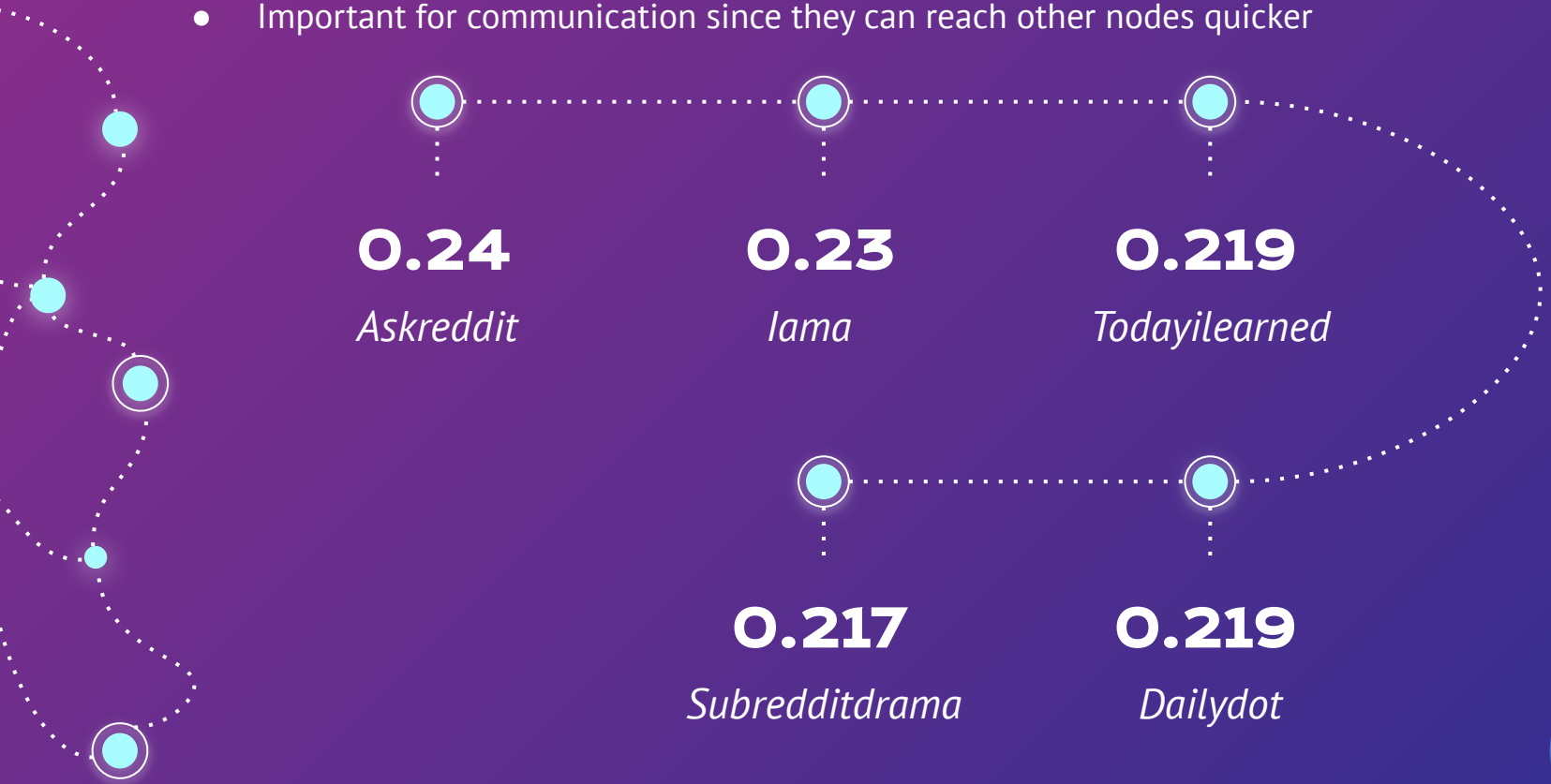
Highest Degree Centrality

- These represent the subreddits that are most connected to other subreddits
- These are the most central nodes and have the most incoming/outgoing links



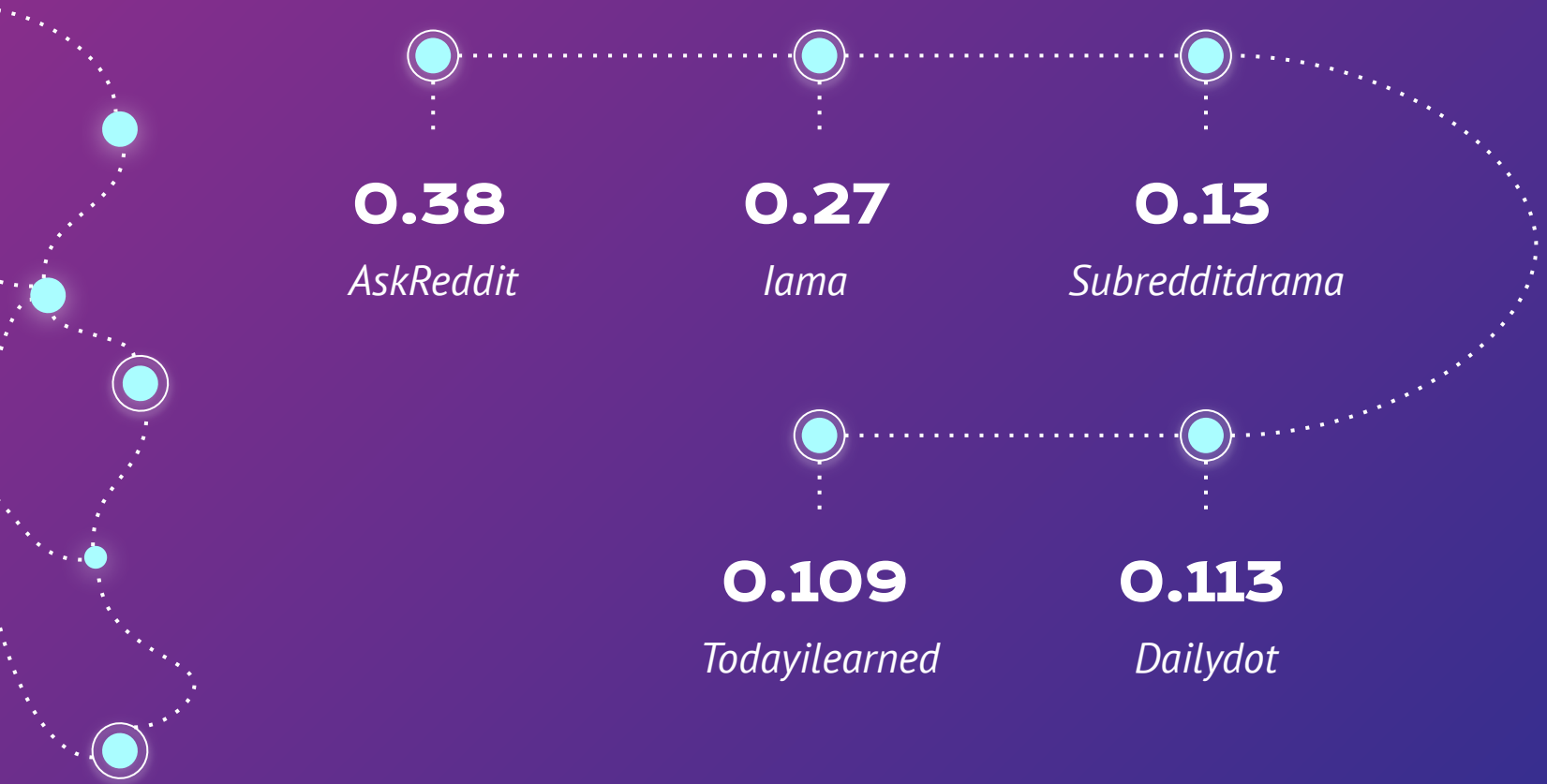
Highest Closeness Centrality

- These nodes have the shortest average distance to all the other nodes in the network
- Important for communication since they can reach other nodes quicker

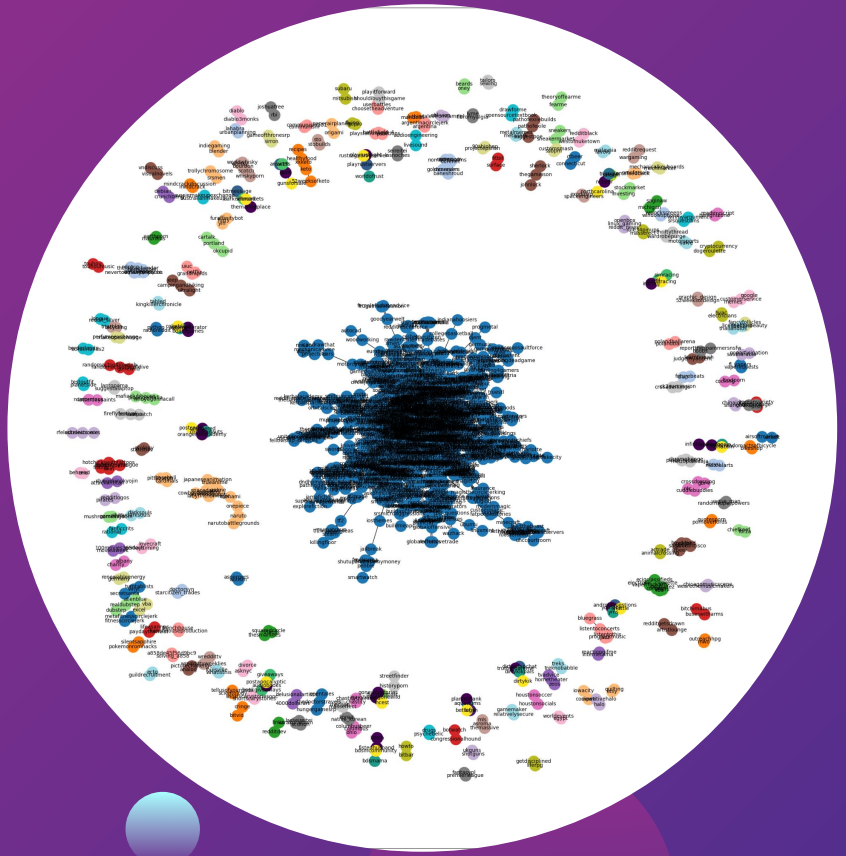


Highest Betweenness Centrality

- number of shortest paths in the network that pass through a given node
- identify important nodes in a network that may play critical roles
- these subreddits were mentioned the most before reaching the target subreddit



COMMUNITY DETECTION



182 Communities

- Detects communities by removing nodes from original network through Girvan-Newman

ANALYSIS

- The communities represent a group of subreddits that have a higher likelihood of being connected and interacting with each other
- One large community and hundreds of smaller communities consisting of 3 - 4 nodes
- In the largest community, there are 975 nodes

Conclusion

- Learn about the structure of subreddits
- Able to identify the most important subreddits and see how subreddits interacted with each other

References

Link to the Reddit Hyperlink Subreddit Network: https://snap.stanford.edu/data/soc-RedditHyperlinks.html

Slides

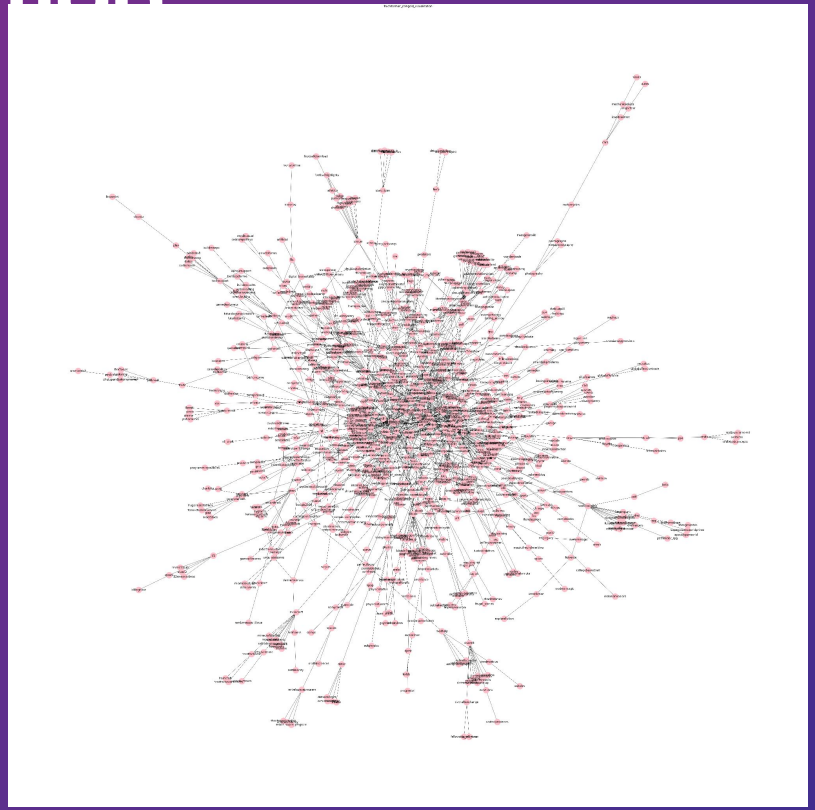
- Project Intro - what it is
- Characteristics
  - Number of nodes: 1454
  - Number of edges: 1617
  - Number of connected components: 181
  - The diameter (longest shortest path). 13 (longest shortest path of the largest connected component nodes: 1003, edges: 1339)
  - The five nodes with the highest clustering coefficients. (DOUBLE CHECK)
    - education 1.0
    - pokemongiveaway 1.0
    - conspiratocracy 1.0
    - buildapcsales 1.0
    - karmacourtattorneys 1.0
  - The five nodes with highest betweenness centrality (node betweenness)
    - askreddit 0.3821071304476218
    - iama 0.26955903209063353
    - subredditdrama 0.12945525499014546
    - dailydot 0.11342460425497487
    - todayilearned 0.10864249952731021
- Community detection algorithm
- Conclusion/takeaways

- Include visualizations
- Source nodes are connected to target node

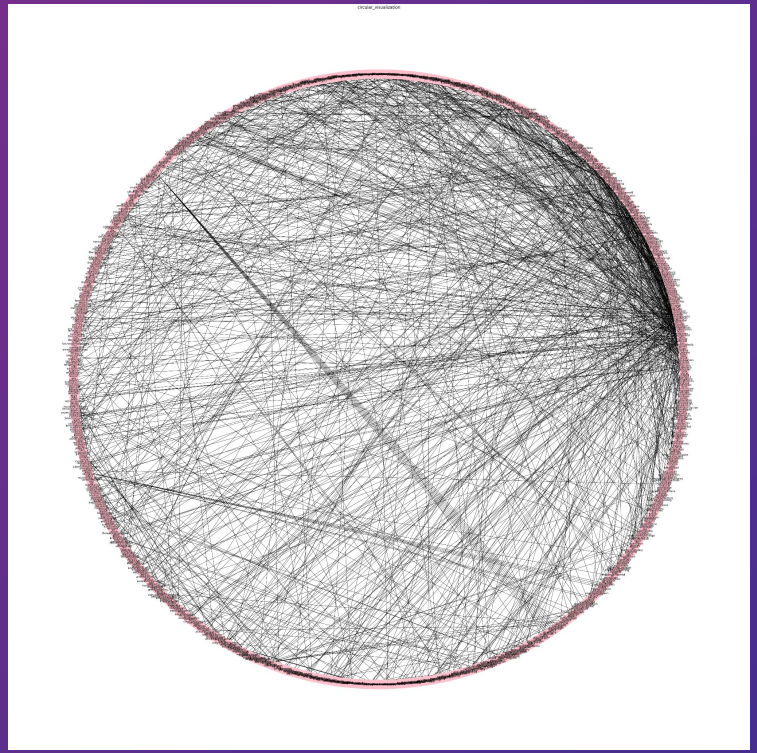
Slides

- Number of communities - 182

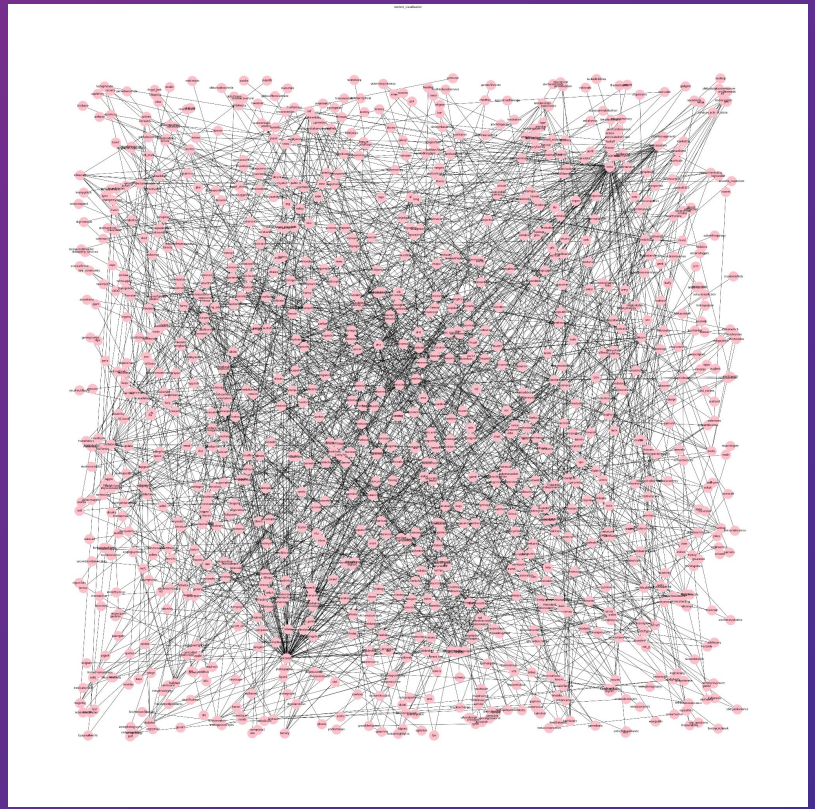
Fruchterman Reingold Visualization



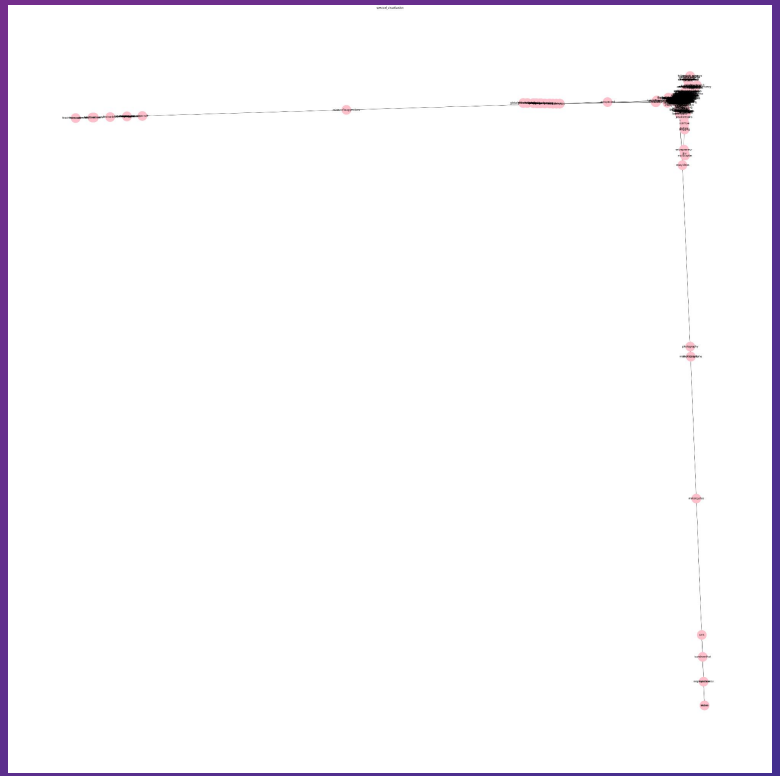
Circular Visualization



Random Visualization



Spectral Visualization



Spring Visualization