Social Network Analysis Final Report

Group 4: Kira Luo, Grace Xie, Judy Zhu, Sydney Li

Executive Summary

Reddit, a leading social media platform, comprises thousands of interconnected subreddits catering to diverse themes and interests. Using social network analysis (SNA), this project uncovers relationships between subreddits by analyzing hyperlink interactions. The goal is to identify influential subreddits, explore community connectivity, and examine sentiment dynamics to promote healthier online engagement.

In Depth Questions

Influence Within the Subreddit Network : Identifying influential subreddits is key to understanding network dynamics. Subreddits with high degree centrality serve as hubs for information sharing, while those with high betweenness centrality bridge disconnected communities.

Connectivity and Cross-Community Engagement: Reddit's structure thrives on hyperlink interactions binding subreddits into a cohesive network. Analyzing these interactions reveals how communities connect and form clusters based on shared themes, such as gaming or politics. Mapping these networks uncovers the size, structure, and interconnectedness of communities, providing insights into how engagement flows across the platform.

Sentiment Dynamics Across Subreddit Connections: Sentiment analysis reveals the tone of interactions between subreddits. Positive sentiment often dominates hobby and support subreddits, while contentious topics like politics tend to see more negativity. Identifying sentiment clusters and trends helps moderators pinpoint areas of potential conflict or toxicity, promoting healthier interactions and platform integrity.

Exploratory Data Analysis (EDA)

The analysis covers Reddit’s growth and interaction patterns from 2014 to 2017. Subreddits increased from 191,700 in 2014 to 292,500 in 2016, with hyperlink connections peaking at 36,400. However, 2017 saw a decline, indicating possible shifts in platform activity. Degree distribution identifies central hubs, while sentiment analysis quantifies positive, neutral, and negative interactions.

Visualizing the network using tools like Gephi reveals central subreddits and thematic clusters. Community detection algorithms highlight subreddits with shared interests, offering insights into the platform's structure. A sentiment heatmap identifies positive and negative subreddit pairs, while sentiment trends over time expose shifts linked to major events or controversies.

Centrality analysis highlights influential subreddits, with degree centrality identifying hubs and betweenness centrality uncovering bridges between communities. Cross-community interaction reveals clusters formed around shared themes.

Toxicity analysis pinpoints subreddit pairs with the most negative sentiment, offering actionable insights for moderators to reduce toxicity and promote positive interactions, supporting meaningful discussions.

Overall, this analysis clarifies Reddit's network structure, sentiment dynamics, and connectivity, equipping moderators to enhance platform integrity and foster vibrant online communities.

Analysis

Network Structure and Influence

We conducted a social network analysis using subreddit hyperlink data to identify influential subreddits driving discussions. Focusing on five centrality metrics—**in-degree**, **out-degree**, **eigenvector centrality**, **PageRank**, and **betweenness**—we analyzed yearly trends from 2014 to 2017, identifying the top 5 subreddits for each metric.

The analysis revealed key insights: Subreddits like **"subredditdrama"** and **"bestof"** (high out-degree) drove discussions, while **"AskReddit"** (high in-degree) served as a central hub of attention. Subreddits such as **"AskReddit"** and **"funny"** (high betweenness) acted as bridges connecting disparate communities. Notably, **"AskReddit"** dominated across metrics, highlighting its pivotal role in Reddit's ecosystem. However, a decline in centrality by 2017 suggests decentralization, possibly due to shifting user behavior and the rise of niche subreddits. Subreddits like **"funny"** and **"worldnews"** maintained a stable influence, emphasizing their enduring relevance. These findings demonstrate the value of centrality metrics in understanding Reddit's evolving dynamics.

Cross-Community Engagement

The study analyzed the subreddit hyperlink network for November 2016, a significant period during the U.S. presidential election, to explore heightened activity and cross-community interactions among subreddits. The analysis focused on reciprocity, clustering, and transitivity to understand patterns of user engagement and content dissemination. The dataset was filtered for interactions between November 1 and 30, 2016, removing duplicates to simplify the network of directed edges between subreddits. Exponential Random Graph Models (ERGMs) were used to examine tie formation, with a simple model capturing reciprocity and a complex model incorporating clustering effects (GWESP) for transitivity and triadic closure. Network metrics such as reciprocity, clustering coefficient, and betweenness centrality were also computed. All performances are attached to fig6-8 in the appendix.

ERGMs were chosen for their ability to model structural dependencies and quantify the significance of observed patterns, essential for understanding cross-community engagement. Results showed that reciprocity, captured in the simple ERGM, emphasized the importance of bidirectional ties, fostering collaboration and debates between subreddits like r/politics and r/The_Donald. Clustering, identified in the complex model, highlighted triadic closure and segmentation within topic-specific groups, reflecting shared interests or collaboration. Central subreddits, revealed through betweenness centrality, acted as bridges, disseminating content across communities. These findings underscore how reciprocity and clustering drive user engagement and how central hubs facilitate content dissemination, providing valuable insights into cross-community dynamics.

Sentiment Dynamics

T-Test on Text Complexity Features: We conducted t-tests to compare text complexity features—Number of Characters, Number of Words, and Average Word Length—between positive and negative sentiment posts. Results showed significant differences across all metrics ($p < 0.001$). Negative sentiment posts were longer in characters and words and had slightly higher average word lengths, suggesting that users expressing dissatisfaction tend to elaborate more and use complex language. These differences highlight the behavioral disparity between sentiment groups, with negative posts reflecting greater effort and engagement, useful for engagement modeling.

Centrality vs Sentiment: We analyzed degree and betweenness centrality for subreddits to assess their network prominence and bridging roles, respectively. Aggregating average sentiment scores per subreddit, scatterplots with regression lines revealed relationships between degree centrality and sentiment, providing insights into whether highly connected or influential subreddits exhibit distinct sentiment patterns.

Findings

Network Structure and Influence

Firstly, we find subreddits driving discussions **(Out-Degree Centrality, shown by Fig.1 in Appendix):** Subreddits like "subredditdrama" and "bestof" actively drove cross-community discussions by frequently linking to others. "Subredditdrama" peaked in 2015 with an out-degree centrality of ~10,000, highlighting its role as a major driver, while "bestof" maintained consistent influence by curating top Reddit comments.

Furthermore, we also have insights on subreddits receiving attention (**In-Degree Centrality, Fig.2**) like "AskReddit," "funny," and "pics" consistently ranked highest, reflecting popularity. "AskReddit," a hub for thought-provoking questions, peaked in 2016 with ~8,500 in-degree centrality, highlighting dominance. "Funny" and "pics" also stayed key hubs of attention.

Moreover, we explored influence via quality of connections **(Eigenvector Centrality, Fig.3):** "AskReddit" emerged as the most influential subreddit with an eigenvector centrality of 1.0

across multiple years, connected to other high-impact subreddits. "Funny" and "politics" also demonstrated consistent influence within their clusters.

Next, we investigated hub-like influence **(PageRank, Fig.4):** Subreddits like "AskReddit", "funny", "pics", and "iama" served as central hubs for discussions. "AskReddit" peaked in 2014 with a PageRank score of ~0.025, showcasing its prominent role in driving platform-wide conversation.

Lastly, we found insights on bridging communitie**s (Betweenness Centrality, Fig.5):** Subreddits such as "AskReddit", "politics", and "worldnews" acted as bridges connecting otherwise distinct communities. "AskReddit" showed significant bridging behavior in 2015, fostering cross-community interactions.

These findings stress the diverse roles subreddits play in driving discussions, attracting attention, and connecting communities, emphasizing their critical importance in Reddit's ecosystem.

Cross-Community Engagement

All performances are attached to Fig.6-8 in the appendix, showing simple, complex, and comparison of the 2 models. The comparison of two ERGM models for the subreddit hyperlink network in November 2016 highlights the strengths and trade-offs between simplicity and complexity. The simple model, with terms for edges and mutual (reciprocity), achieves a significantly lower AIC and BIC, indicating a better fit with fewer parameters. In contrast, the complex model, which includes terms for edges, mutual, and gwesp (transitivity), adds complexity without substantial improvement in fit, as evidenced by its higher BIC.

The simple model reveals strong reciprocity effects, with a highly significant mutual term ($p < 0.001$) and an odds ratio of 146.2, indicating that reciprocal ties are 146 times more likely than unidirectional ties. However, the baseline probability of a tie is very low (0.04%), reflecting the network's sparsity. The network's low overall reciprocity score (0.0498) suggests that most links are unidirectional, but when reciprocity occurs, it is a dominant structural feature. While the complex model captures significant clustering effects (gwesp), multicollinearity renders the mutual term insignificant ($p = 0.379$), emphasizing that reciprocity is better captured in the simpler model.

The low reciprocity and lack of clustering in the simple model suggest hierarchical information flow, with certain subreddits acting as hubs for content dissemination. This sparsity, combined with significant local reciprocity, reflects strong collaboration among specific subreddits rather than widespread bidirectional interactions. Ultimately, the simple model is preferred for its clarity and efficiency, offering insights into targeting influential hubs and collaborative subreddits to maximize engagement and content flow.

Sentiment Dynamics

T-Test on Text Complexity Features: Posts with negative sentiment are significantly longer in terms of characters, with a mean difference of approximately 130 characters (95% CI: [59.34, 200.21]), indicating that users expressing dissatisfaction tend to elaborate more. Similarly, negative sentiment posts contain an average of 25 more words (95% CI: [13.68, 35.69]) than positive ones, highlighting greater verbosity. Additionally, the average word length in negative sentiment posts is 0.16 characters longer (95% CI: [0.1331, 0.1890]), reflecting the use of more formal or complex language. These findings suggest that negative sentiment is associated with greater complexity.

Centrality vs Sentiment Analysis: Subreddits with higher degree centrality, indicative of more connections, show a slight negative correlation with sentiment. These highly active subreddits tend to host more critical or contentious discussions, as reflected in their lower average sentiment scores. This trend was visualized in a scatterplot, where a downward-sloping regression line highlighted the relationship. While these subreddits maintain extensive connections within the network, they may also act as hubs for spreading more negative sentiments.

Implications

Network Structure and Influence

Our analysis highlights key areas for Reddit's community management. High in-degree subreddits (e.g., "AskReddit", "worldnews") serve as hubs for positive interactions and inclusivity, while high out-degree subreddits (e.g., "subredditdrama") drive cross-subreddit discussions but require monitoring for conflicts. Bridging subreddits (e.g., "AskReddit", "funny") connect communities, reducing echo chambers but needing effective moderation. Thematic clusters (e.g., politics, entertainment) reflect natural boundaries, enabling tailored moderation and content strategies.

We recommend Reddit take the following steps to strengthen community dynamics and foster meaningful interactions. Firstly, leverage key subreddits by promoting campaigns in high in-degree subreddits like "AskReddit" and monitoring high out-degree subreddits like "subredditdrama" to address conflicts. Secondly, support bridging subreddits such as "funny" to encourage cross-community dialogue. Additionally, tailor moderation strategies by enforcing stricter policies in political clusters to combat misinformation while fostering creativity in entertainment-focused subreddits. Finally, use high-centrality subreddits to share positive stories and promote constructive engagement across the platform.

Cross-Community Engagement

Reddit can enhance future engagement by leveraging insights from cross-community patterns. To boost subreddit engagement, identifying subreddits with strong bidirectional ties using the reciprocity metric can highlight collaborative communities for fostering deeper engagement. For example, r/politics and r/worldnews, which share mutual links during events like elections, can be targeted for joint campaigns, shared events, or discussion threads. Promoting cross-posting and shared moderation policies can strengthen these ties and increase user retention. Additionally, leveraging clustering and transitivity (captured by GWESP) can improve subreddit recommendations, enhancing user experience through personalized content discovery and encouraging cross-community participation.

For selective targeting, advertising can be optimized by focusing on tightly connected subreddits with high clustering. For instance, tech advertisements placed in r/technology could spill over into related subreddits like r/programming or r/gadgets, maximizing ROI. Personalized recommendations can also be tailored using mutual ties and betweenness centrality, connecting users to niche subreddits aligned with their interests. For example, a user engaged with r/science could be guided to r/askscience or r/medicine, enhancing satisfaction and retention.

Sentiment Dynamics

The differences in text characteristics between positive/neutral and negative sentiments provide actionable insights for platforms. Automated content moderation tools can leverage these findings by prioritizing the review of longer posts with complex wording, which are more likely to exhibit negativity, while promoting concise, engaging positive content. Moreover, the association of sentiment types with distinct communication styles can guide targeted content creation. Creators and advertisers can tailor their messaging to align with audience expectations, optimizing engagement based on platform or subreddit context. For businesses, these patterns enhance sentiment analytics, allowing sentiment detection models to capture nuances in language complexity across sentiment classes for improved accuracy and actionable insights.

The negative correlation between degree centrality and sentiment emphasizes the need for community engagement strategies. Platforms can design initiatives to foster positive sentiment in highly central subreddits, recognizing their significant influence on the network's overall tone. Understanding how sentiment relates to network structure further enables administrators to predict sentiment flows, addressing toxicity in critical hubs or amplifying positivity in influential nodes. These insights also inform the design of recommendation systems, ensuring that new or infrequent users are guided toward supportive, positive communities. By steering users away from negative hotspots, platforms can enhance user satisfaction and retention.

Reflections

For network structure and influence, our social network analysis using R addressed key questions on Reddit's network structure and influence. Subreddits like "subredditdrama", "bestof", and "circlebroke2" (high out-degree) drive discussions, while "AskReddit", "worldnews", and "funny" (high in-degree) serve as central hubs of attention. Additionally, subreddits like "AskReddit" and "funny" (high betweenness) bridge disconnected communities, fostering cross-topic engagement. Natural clustering suggests thematic groupings (e.g., politics, entertainment), though deeper analysis is needed. Notable findings include "AskReddit's" consistent dominance across all metrics, highlighting its dual role as a hub and bridge, and the decentralization of influence by 2017, likely reflecting shifting user behavior and the rise of niche subreddits. These results emphasize the evolving dynamics of Reddit and the critical role of central subreddits.

In terms of cross-community engagement, using the ERGM model, user engagement is primarily driven by reciprocal interactions and localized community dynamics. Subreddits with bidirectional ties act as key hubs, fostering collaboration and strong user retention, though these ties are not widespread. The low global reciprocity score (0.0498) highlights selective engagement, with most interactions being unidirectional. While the significance of reciprocity for engagement was expected, the exceptionally low global reciprocity score was surprising, indicating a more hierarchical structure than anticipated.

For content dissemination, unidirectional ties dominate, with central hubs like r/news serving as key bridges that distribute information across disconnected network segments. These hubs amplify content flow during events like elections. While the role of hubs in dissemination was expected, the limited clustering between hubs and niche subreddits was surprising. Despite strong transitivity (captured by GWESP), interactions largely remain confined within isolated clusters, with minimal cross-cluster engagement.

For sentiment analysis, it taught us how sentiment and network connectivity interplay within online communities, emphasizing the value of addressing dataset imbalance for unbiased analysis. By examining centrality measures, we found how influential nodes shape sentiment flows, with highly connected subreddits often exhibiting negative sentiment due to visibility pressures. Visualizing these patterns enhanced our ability to interpret and communicate insights, while identifying clusters with concentrated negativity highlighted opportunities for targeted interventions. Overall, we learned how sentiment analysis and network theory can uncover actionable insights and deepen understanding of online community dynamics.

Visualizations

The first visualization shown in fig. 9 presents a network graph of subreddit hyperlink interactions, with nodes as subreddits and edges as hyperlinks. Key observations reveal a **dense core** of highly connected subreddits, such as r/politics and r/news, that act as interaction hubs and bridges between communities. **Peripheral nodes** on the outer ring, primarily niche subreddits, have fewer connections and mainly link to central hubs, reflecting limited cross-community interactions. The **clustered communities** within the core suggest echo chambers or topic-based groups, such as those centered on political discussions. These patterns highlight opportunities to engage peripheral subreddits and leverage bridge nodes to reach untapped or niche audiences through targeted outreach strategies.

After balancing the dataset, the second visualization shown in fig.10 highlights a clear negative relationship between degree centrality and average sentiment, showing that subreddits with higher connectivity tend to exhibit more negative sentiment. The plot effectively illustrates this trend, with a significant downward slope and a confidence interval that reinforces the robustness of the finding.

Code and References

This is the source of dataset: https://snap.stanford.edu/data/soc-RedditHyperlinks.html

For detailed code, please refer to the zip file attached in the canvas submission of “SNAP Final Report”.

Appendix

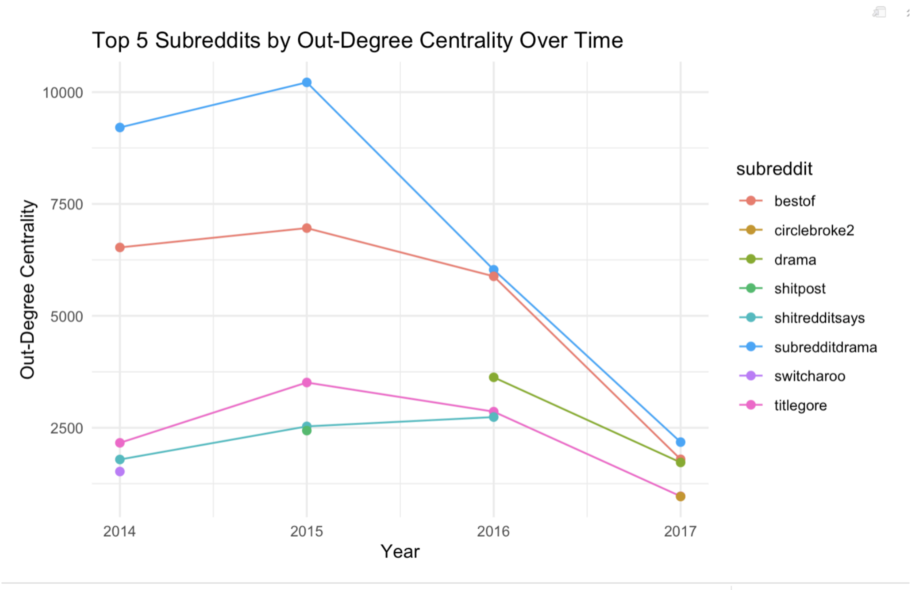


Fig 1

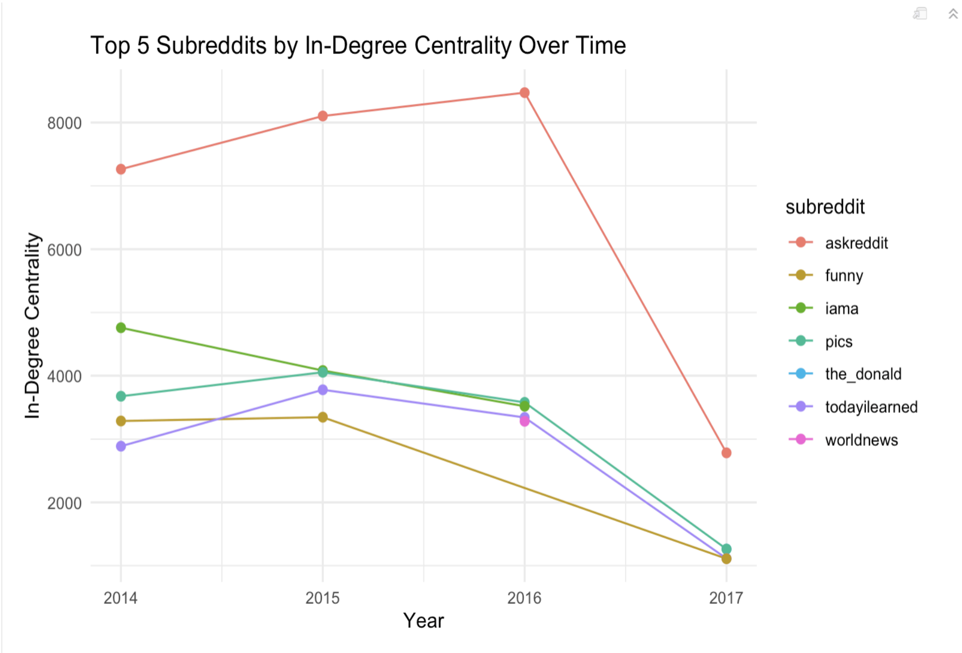


Fig 2

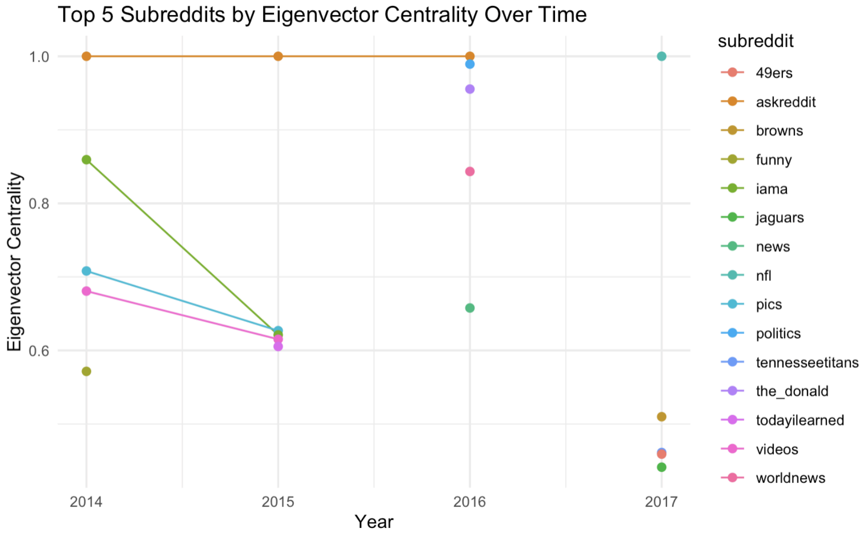

Top 5 Subreddits by Eigenvector Centrality Over Time
Eigenvector Centrality
1.0
0.8
0.6
2014
2015
2016
2017
Year
subreddit
49ers
askreddit
browns
funny
iama
jaguars
news
nfl
pics
politics
tennesseetitans
the_donald
todayilearned
videos
worldnews

Fig 3

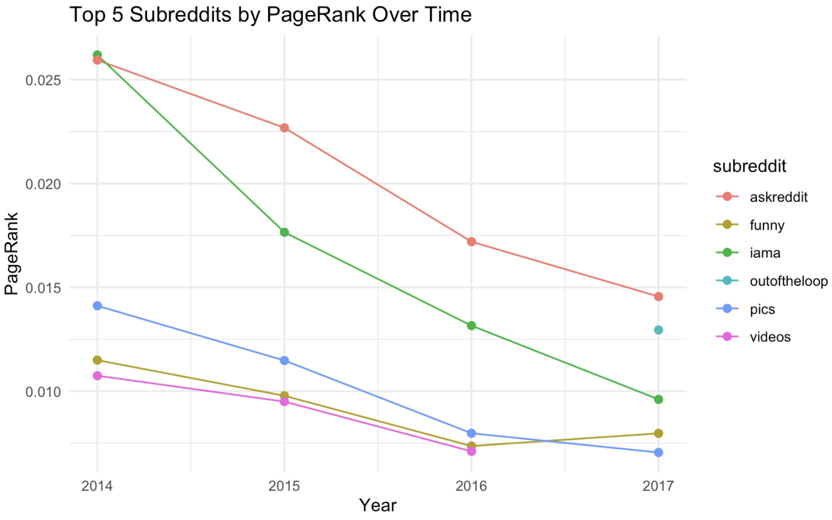

Top 5 Subreddits by PageRank Over Time
PageRank
0.025
0.020
0.015
0.010
2014
2015
2016
2017
Year
subreddit
askreddit
funny
iama
outoftheloop
pics
videos

Fig 4

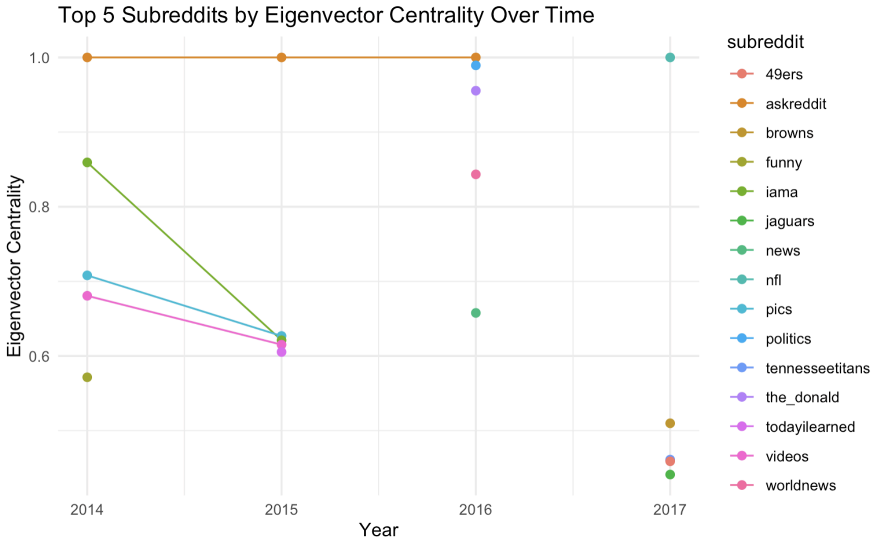


Fig.5

Simple Model

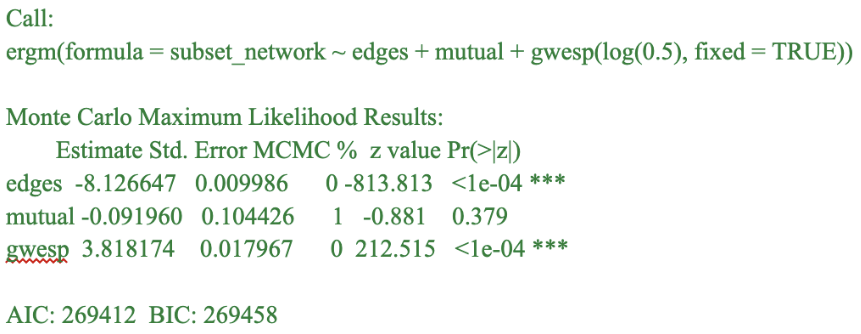
```
Call:
ergm(formula = subset_network ~ edges + mutual)

Monte Carlo Maximum Likelihood Results:
       Estimate Std. Error MCMC %  z value Pr(>|z|)
edges  -7.894782   0.009392      0 -840.59   <1e-04 ***
mutual  4.986986   0.061868      0   80.61   <1e-04 ***

AIC: 214430  BIC: 214460
```

Fig.6

Complex Model

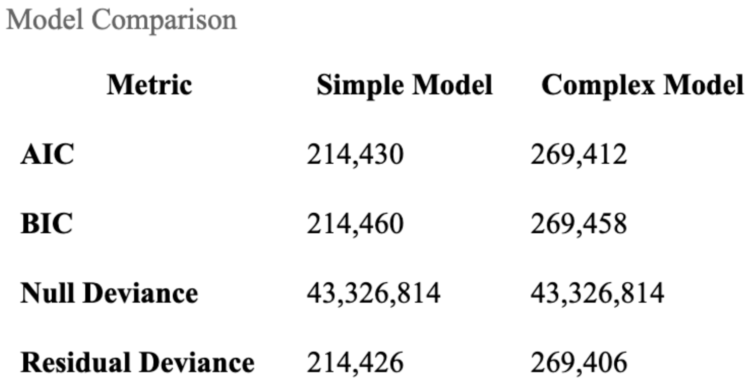


```
Call:
ergm(formula = subset_network ~ edges + mutual + gwesp(log(0.5), fixed = TRUE))

Monte Carlo Maximum Likelihood Results:
       Estimate Std. Error MCMC %  z value Pr(>|z|)
edges  -8.126647   0.009986      0 -813.813   <1e-04 ***
mutual -0.091960   0.104426      1   -0.881    0.379
gwesp  3.818174    0.017967      0  212.515   <1e-04 ***

AIC: 269412  BIC: 269458
```

Fig.7



Model Comparison

Metric	Simple Model	Complex Model
AIC	214,430	269,412
BIC	214,460	269,458
Null Deviance	43,326,814	43,326,814
Residual Deviance	214,426	269,406

Fig.8

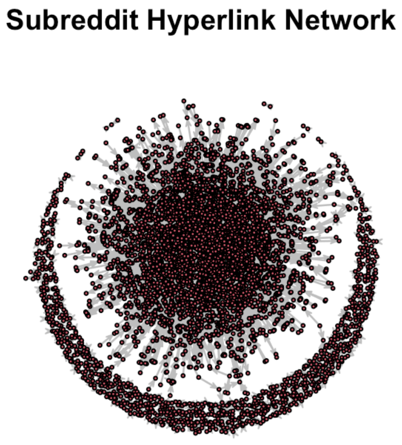


Fig.9

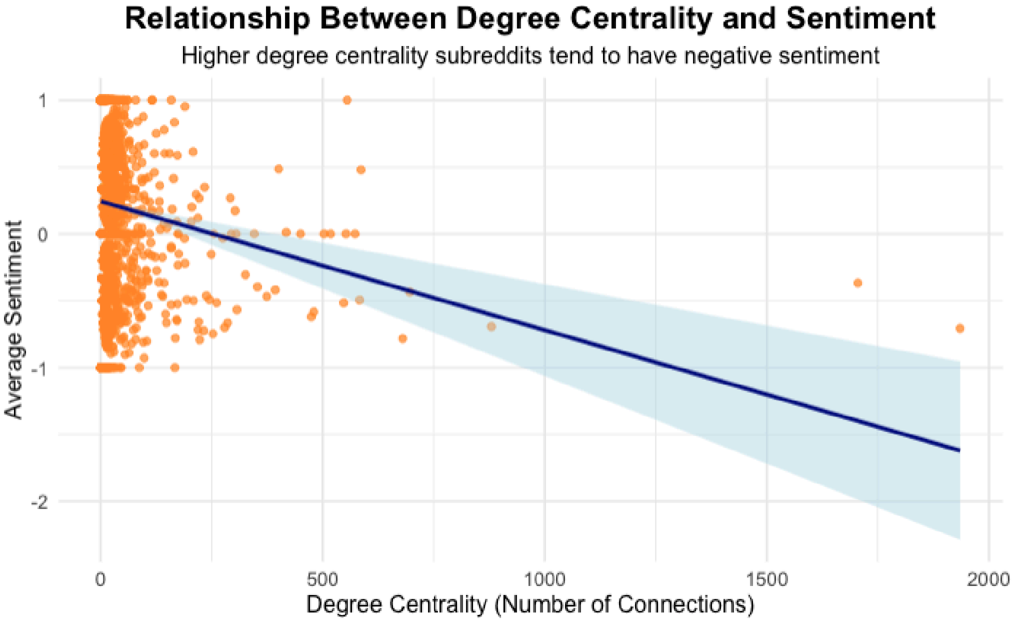

Relationship Between Degree Centrality and Sentiment
Higher degree centrality subreddits tend to have negative sentiment
Average Sentiment
1
0
-1
-2
0
500
1000
1500
2000
Degree Centrality (Number of Connections)

Fig.10