Reddit Hyperlink Network Analysis

Ravi Chepuri
Physics undergraduate
Nonlinear Dynamics 465-0

Reddit hyperlink network

- Reddit: an online forum where users' posts are organized by subject into boards/communities called "subreddits"
- In posts to one subreddit, users can reference another subreddit by including a hyperlink in the title or body of the post
- Reddit hyperlink network: a directed network with subreddits as nodes and hyperlinks from one subreddit to another as directed links
- Analysis of hyperlink network can yield insight into dynamics of online community interaction

Dataset

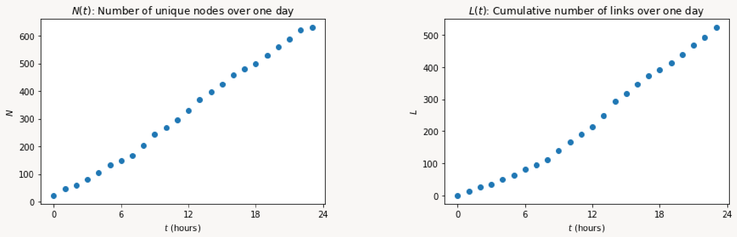
- From Kumar et al., 2018
- 55,863 nodes, 858,490 links gathered from Jan '14 – Apr '17
- Directed, temporal, attributed
- Edge list data format:

```
SOURCE_SUBREDDIT tab TARGET_SUBREDDIT
tab POST_ID tab TIMESTAMP tab
POST_LABEL tab POST_PROPERTIES
```

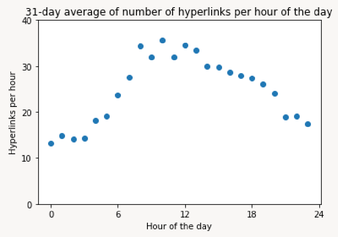
Data size reduction

- Restricted data to only hyperlinks posted Jan. 1 2017 (N = 630 nodes, L = 530 links), except where otherwise noted

N(t), L(t)

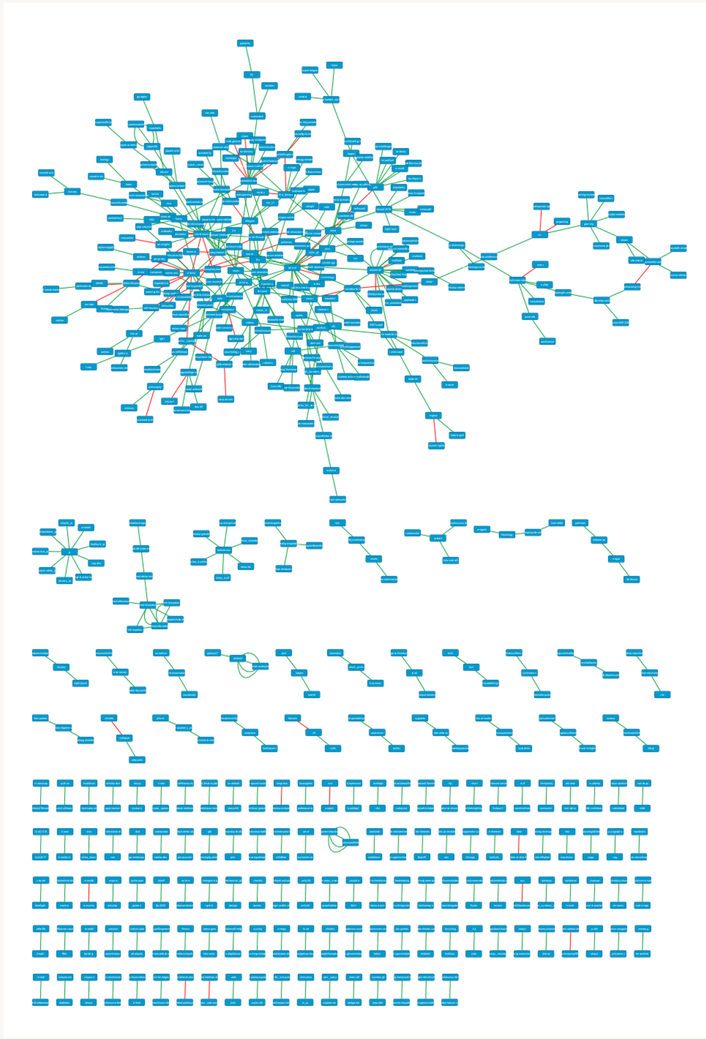


Total number of nodes N and number of links L over the course of the day



Averaging the number of new nodes every hour over a 31-day period, a daily periodicity is observed

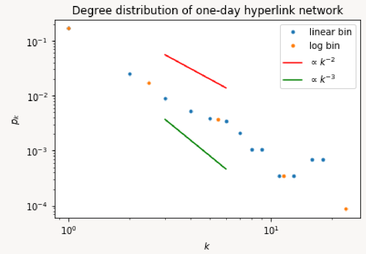
Network visualization and characteristics



Visualization of the Jan. 1, 2017 reddit hyperlink network. The network is directed, temporal, and attributed, with attributes including post sentiment (green = positive or neutral, red = negative)

- N = 630 nodes, L = 530 links
  - Relatively sparse
- 151 connected components / 1 giant component
- Characteristic path length: <l>=1.600
- Clustering coefficient: C = 0.004
  - Greater than C_{rand} = 2.61e-3

Degree distribution



Total (in and out) degree distribution of the Jan. 1, 2017 hyperlink network. Appears to follow a power law distribution, suggesting scale-free topology and the presence of preferential attachment

Preferential attachment?

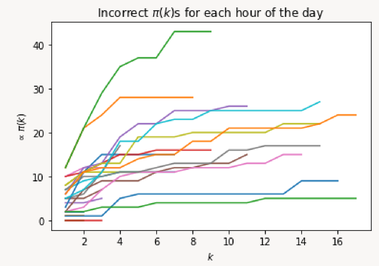
- Scale-free structure suggests preferential attachment; we can test using temporal data
- Probability of new node attaching to node *i* with degree *k_i* can be estimated by

$$\Pi(k_i) \sim \frac{\Delta k_i}{\Delta t}$$

- Cumulative preferential attachment function:

$$\pi(k) = \sum_{k_i=0}^{k} \Pi(k_i)$$

- Expect $\pi(k) \sim k^2$ if linear preferential attachment; $\pi(k) \sim k$ if no preferential attachment



Incorrect attempt to determine $\pi(k)$ from different hours of the data.

- Issue: Preferential attachment assumes we start with a set of nodes all with degree at least 1, but here we start we nodes with degree 0

Further work

- Finish preferential attachment analysis
- Account for directed nature of network in degree distribution/preferential attachment analysis
- Investigate positive and negative sentiment

Tools used: Python (networkx, matplotlib, pandas), Cytoscape

Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). Community Interaction and Conflict on the Web. Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18, 933–943. https://doi.org/10.1145/3178876.3186141