# Detection and Analysis of Subreddit Communities

**Sam Griesemer** SAMGRIESEMER@WUSTL.EDU
Washington University in St. Louis

**Louis Schlessinger** LSCHLESSINGER@WUSTL.EDU
Washington University in St. Louis

## Abstract

The social media platforms of today are rife with massive amounts of user data. Information is collected about every click made in order to better suggest new content and people to connect with. As users of Facebook, Twitter, and Reddit, we often don't think about the methods by which new content is recommended and how our interests are mapped. In this work, we attempt to shed light on this by building a subreddit interest network from user activity data on the Reddit platform. We then analyze the properties of this network and compare the performance of several community detection algorithms. We further explore the quality of the produced communities and visualize them to better understand the intricacies of user activity on Reddit.

## 1. Introduction

Reddit is a popular social media platform where users can share links, stories, photos, videos, and various other forms of media. Each content submission is uploaded to a particular *subreddit*, a specialized community created and maintained by Reddit users devoted to a certain topic. Users then have the ability to leave comments and vote on posts submitted to these subreddits.

Subreddits are natural communities within the Reddit platform, providing a place for discussion among users with similar interests. Many users are active across a number of different subreddits, ranging from general subreddits like *r/videos* to niche groups like *r/MinnesotaCamping*. These cross-community interactions reveal an intricate network of related user interests, from which we can extract higher-level subreddit relations.

In this project, we analyze user activity from a large dataset of active user data, containing information on over 850,000 active users across more than 15,000 subreddits. We use this data to build a subreddit-to-subreddit interest network, where two subreddits are connected if a large portion of one subreddit's members are also active in the other. We then apply several community detection algorithms to this network to find clusters of similar subreddits across Reddit.

## 2. Data

As mentioned previously, we make use of a dataset from (Olson & Neal, 2014) containing user activity from over 876,961 Reddit users across 15,122 subreddits. Each entry in this dataset provides a user ID and a list of subreddits within which that user is active. A user is considered active in a particular subreddit if at least 10 of their 1,000 most recent posts or comments were made in that subreddit. This ensures our graph does not get bogged down by millions of unhelpful edges representing inactive users.

It is worth noting that this dataset was created in mid-2013. As such, the subreddits present in our graph do not necessarily reflect the state of the subreddits in 2018. Further detail regarding the differences between Reddit in 2013 and Reddit in its current form are discussed in the further work section below.

## 3. Network Construction

In order to better analyze the information in the dataset described above, we use it to construct a subreddit-to-subreddit interest network. We'll use $S$ to denote the set of subreddits and $U$ to denote the set of users present within the dataset. We first constructed a binary matrix $X \in \mathbb{R}^{|S|\times|U|}$, where entry $X_{i,j}$ is 1 if user $j$ is active in subreddit $i$ and 0 otherwise. $X$ compactly summarizes the dataset, and is implemented as a sparse matrix due to the vast majority of entries being 0's. We then compute the subreddit overlap matrix $G = XX^T$, where entry $G_{i,j}$ is an integer representing the number of users that are active in both subreddit $i$ and subreddit $j$. This matrix alone is enough to describe a subreddit graph with edges weighted by the number of overlapping users. The resulting network includes all 15,122 subreddits as nodes, with a total of 4,520,054 interconnect-

*Figure 1.* Visualization of the subreddit network *before* applying a threshold to the edges. This graph contains over 4.5M edges, and each node represents a subreddit.

ing edges.

A large number of these subreddits have a very small number of active users, and thus are not well connected with the rest of the graph. We want to ensure that an edge only exists between two subreddits if a large enough percentage of users in *both* subreddits are also active in the other subreddit. To enforce this constraint, we first compute a *percentage overlap* matrix $P$, where the $P_{i,j}$ entry gives the percentage of active users in subreddit $i$ that are also active in subreddit $j$. Unlike $G$, $P$ is not a symmetric matrix. We then threshold $P$ by 0.05, replacing entries that are less than 0.05 with 0. Finally, we compute the element-wise product of P with its transpose, $P \otimes P^T$, to eliminate one-sided edges. This has the effect of removing an edge between two subreddits if less than five percent of users in one of the subreddits are active in the other. This makes our graph edges more meaningful, in the sense both subreddits have users involved with the other. The resulting graph has 6,372 nodes and 14,791 edges, eliminating roughly 58% of the original subreddits and nearly 99.7% of the original edges. We attempt to visualize the subreddit network before (figure 1) and after (figure 2) thresholding the edges to emphasize the stark difference in structure.

The degree distribution for the thinned network is shown in figure 3. The shape of this distribution is indicative of a scale-free network; it appears linear on a log-log plot. We found the power-law exponent estimate for this distribution to be $\alpha = 2.892$, which is typical of large social networks (usually $2 < \alpha < 3$).

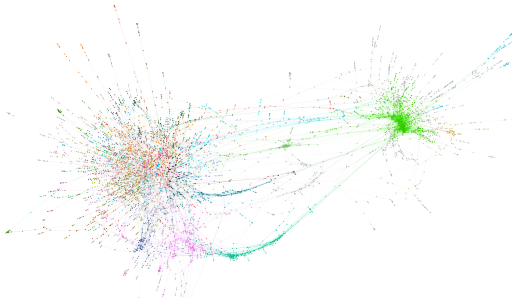

*Figure 2.* Visualization of the subreddit network *after* applying a threshold to the edges. The new graph now contains only 14,791 edges, a tiny fraction of the original edges.

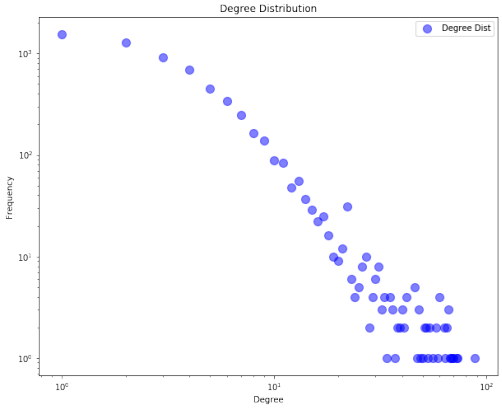


*Figure 3.* A log-log plot showing the degree distribution of the thinned subreddit network. We estimate a power-law exponent for this distribution of $\alpha = 2.892$.

## 4. Model

After building and thinning the network of subreddit relationships, we wanted to find communities of similar subreddits within the graph. Here we make use of several community detection algorithms to find these clusterings: the Louvain method, spectral clustering, and k-means clustering. These group-based community detection algorithms come from three different classes of methods, and each is theoretically well understood.

### 4.1. Louvain Method

The Louvain method for community detection (Blondel et al., 2008) is a greedy, agglomerative optimization method that aims to maximize modularity. It is similar to the modularity maximization method earlier proposed by (Clauset et al., 2004). We use the implementation of the Louvain method provided by Gephi to find communities on our subreddit interest network. After initializing each node to a cluster, the algorithm repeats two steps until the modularity converges. In the the first step, the algorithm continually merges communities until a local maximum is found. In the second

step, it creates a new graph using the communities previously computed and aggregates edges across clusters.

### 4.2. Spectral Clustering

Spectral clustering is an algorithm that aims to solve the optimal cut problem by using the graph Laplacian $L$. This can be solved by computing the eigenvectors of $L$. We use an implementation of unnormalized spectral clustering proposed by (Bauckhage, 2016), which uses the eigenvectors corresponding to the $k$ smallest eigenvalues and computes a $k$-means clustering on them.

### 4.3. K-means Clustering

The k-means algorithm aims to find clusters so that points within clusters are closer to each other and points within other clusters are more distant. This is a non-convex, discontinuous optimization problem. We use the implementation provided by the Python library scikit-learn (Pedregosa et al., 2011), using Lloyd's algorithm (Lloyd, 1982). To solve this problem, the algorithm performs a two-step, alternating minimization. After initializing the cluster centers, the first step updates cluster assignments and the second step updates cluster centers. In order to cluster nodes, we use the edge matrix of $P$ as the data matrix on which to compute the clustering.

## 5. Results

After running the three community detection algorithms described above on our network, we analyze the quality of the produced clusters using *modularity* and *conductance*. These metrics help measure the degree to which the chosen clusters are "good" communities within the network. Modularity considers edges within communities whereas conductance considers both edges within communities and outside them. We outline interpretations of these metrics below and present the quantitative results in figures 4 and 5.

### 5.1. Modularity

Modularity $Q$ is a measure of mesoscopic network structure designed to capture the quality of a clustering for a given graph. It is defined as:

$$Q = \frac{1}{2m} \sum_{ij} A_{ij} - \frac{k_i k_j}{2m} \delta(c_i, c_j)$$

Intuitively, it may be thought of as the difference of the fraction of edges within communities and those expected in a random graph model.

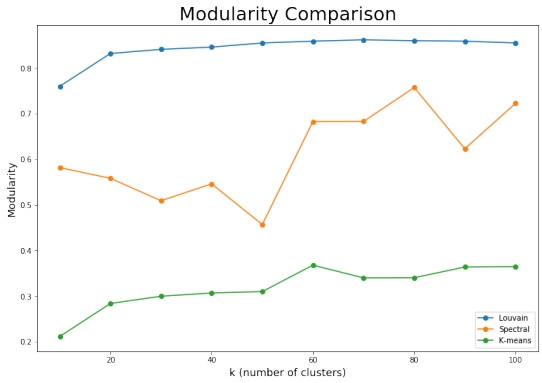


*Figure 4.* A comparison of modularity scores for each community detection algorithm

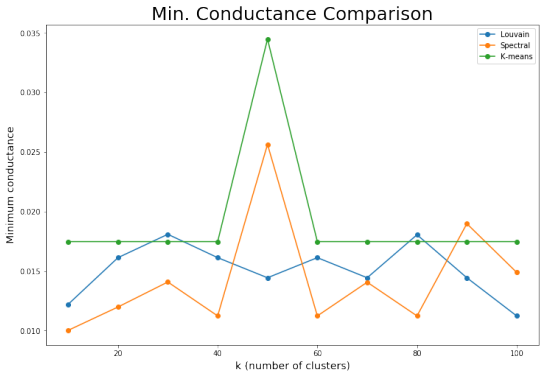


*Figure 5.* A comparison of minimum conductance scores for each community detection algorithm

### 5.2. Conductance

Conductance $f(S)$ is a simple measure of cluster quality. It is the ratio of edges leaving the community to the number of edges in the community and is defined to be:

$$f(S) = \frac{c_S}{2m_S + c_S}$$

Where $S$ is a set of nodes, $c_S$ is the number of inter-community edges, and $m_S$ is the number of intra-community edges. It can be thought as of a surface area-to-volume ratio.

## 6. Discussion

### 6.1. Analysis

We notice a number of different trends when comparing the modularity and conductance scores across community detection algorithms. In figure 4, we can see a clear distinction between the modularity scores of each algorithm when applied to the subreddit interest network. For all observed values of $k$ (number of clusters), we find the Louvain al-

gorithm consistently provides the highest modularity, and maintains a relatively stable score as $k$ varies. The k-means algorithm here serves as something of a baseline for performance due to its simplistic approach, and significantly underperforms the other methods as we might expect. The spectral clustering algorithm produces modularity scores strictly between the those of k-means and the Louvain algorithm, acting as a sort of "middle-ground" method. Additionally, it's clear that spectral clustering is the most sensitive to the value of $k$; its modularity scores vary significantly more than the other approaches as the number of clusters changes. Overall, we observe the highest modularity score from the spectral clustering and Louvain algorithms when the graph is clustered into 70-80 communities.

The conductance scores shed light on a slightly different aspect of the clusters produced by each algorithm. As mentioned above, the conductance is the ratio of edges *leaving* the community to the number of edges *within* the community. Thus, a smaller conductance on a clustering generally indicates that the clustering is more "community-like", as there are a relatively large number of edges within the cluster and less going out. Intuitively, this implies that such a group of nodes is distinctly separated from the rest of the nodes in the graph. In figure 5, we plot the minimum observed conductance over all clusters at the given number of clusters $k$. In this plot, we find the k-means algorithm consistently has a larger minimum conductance, whereas spectral clustering tends to report the lowest conductance values. The Louvain algorithm mostly reports conductances between those of k-means and spectral clustering until $k = 90$, after which we see its conductance drop. Similar to the modularity plot, the Louvain algorithm maintains more consistent conductances as $k$, the number of clusters, varies.

Due to its superior modularity scores and relatively stable conductances, the Louvain algorithm arguably produces the best clusters according to our quantitative results. We further justify this statement with a qualitative exploration and series of network visualizations to better understand the produced clusters.

### 6.2. Network Exploration

In order to better understand Reddit user activity and the relations between similar subreddits, we attempt to visualize the subreddit interest network using Gephi. Figure 6 shows the entire subreddit network with a number of manually annotated clusters indicating some of the intuitive communities we could identify. The graph clusters shown in this image were produced using the Louvain algorithm with a total number of 70 communities.

The teal colored community labeled "learning" is clearly the largest cluster of nodes in the entire graph. The cluster encompasses a large number of the most popular subreddits within Reddit, including r/pics, r/AskReddit, and many others. We label this cluster "learning" primarily because of its inclusion of subreddits devoted to fields of study (i.e r/Economics, r/Philosophy, r/science, etc) as well as other intellectual subreddits (i.e. r/AskHistorians, r/YouShouldKnow, r/geek). A more detailed close up of this community can be seen in figure 7.

Additionally, we find many instances of mainstream subreddits connecting to smaller, more niche groups of related subreddits. An example of this is given in figure 8, where the more mainstream subreddit r/bicycling connects via a single edge to a more niche, distinct community of specific bicycling subreddits.

Overall, we observed many instances of clear, distinct, and intuitive clusters of related subreddits within our network. This helps to back up some of the claims made using our quantitative results, and provides visual insight into how interests are mapped across the Reddit platform.

### 6.3. Further Work

There are many avenues of future work that could be explored from the findings in our project. For example, the subreddit interest network could be used to build a subreddit recommendation system. By analyzing a user's active subreddits, it would be possible to place them in distinct communities of subreddits and make suggestions using neighboring subreddits. In addition, it would be interesting to explore and model "intra-subreddit" structure, in opposition to our focus on "inter-subreddit" connections. This could perhaps be performed using the Network Community Profile introduced by (Leskovec et al., 2010). Lastly, it would be interesting to collect data from active Reddit users of today and compare the network structure to that of Reddit in 2013.

## 7. Conclusion

Through the application and analysis of several community detection algorithms, we were able to identify clusters of related subreddits from user activity data. Metrics like modularity and conductance allowed us to understand the scale ($k$) at which good communities exist and evaluate which algorithms produce clusters of the highest quality. As (Olson & Neal, 2014) found, this further corroborates the view of Reddit as a very diverse set of communities and falsifies the view of Reddit as a single homogeneous entity.

The GitHub repository with our code can be found here: https://github.com/samgriesemer/community-detection

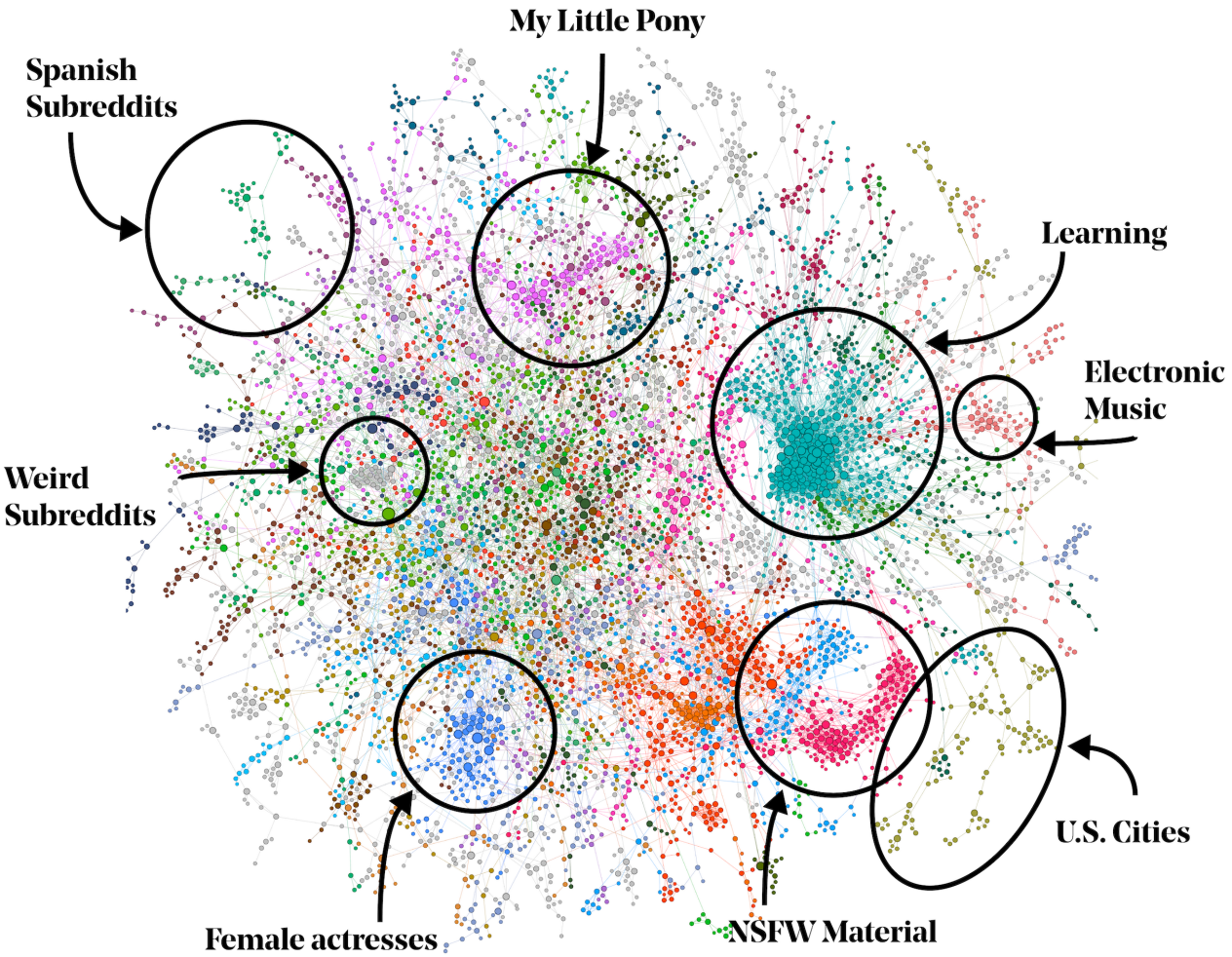


*Figure 6.* A visualization of the entire subreddit interest network with manually annotated clusters of nodes. The plot was produced with Gephi using the OpenOrd layout. Nodes are sized according to their degree and colored according their community.

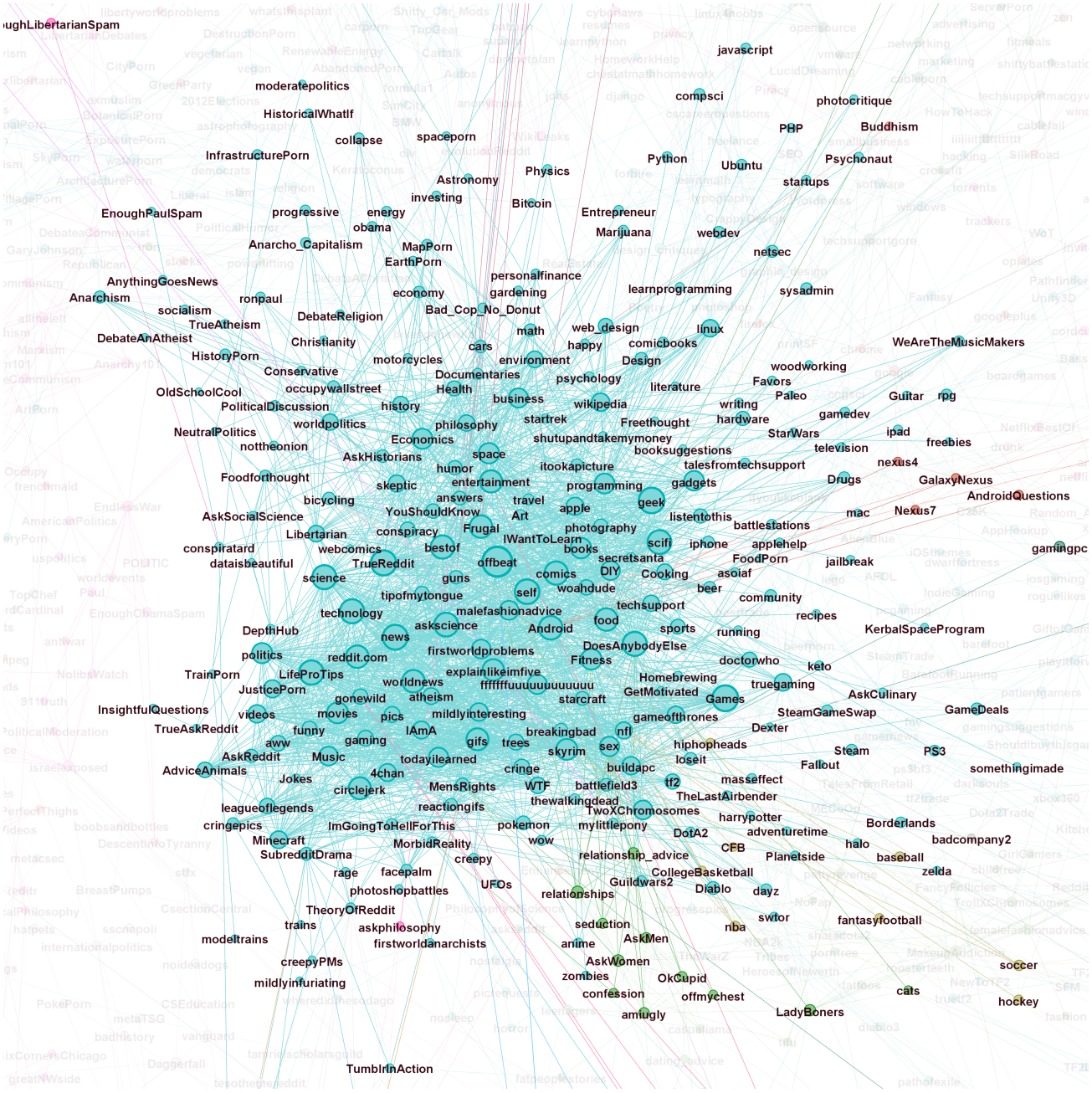


*Figure 7.* A close up of the largest detected community in the subreddit network.

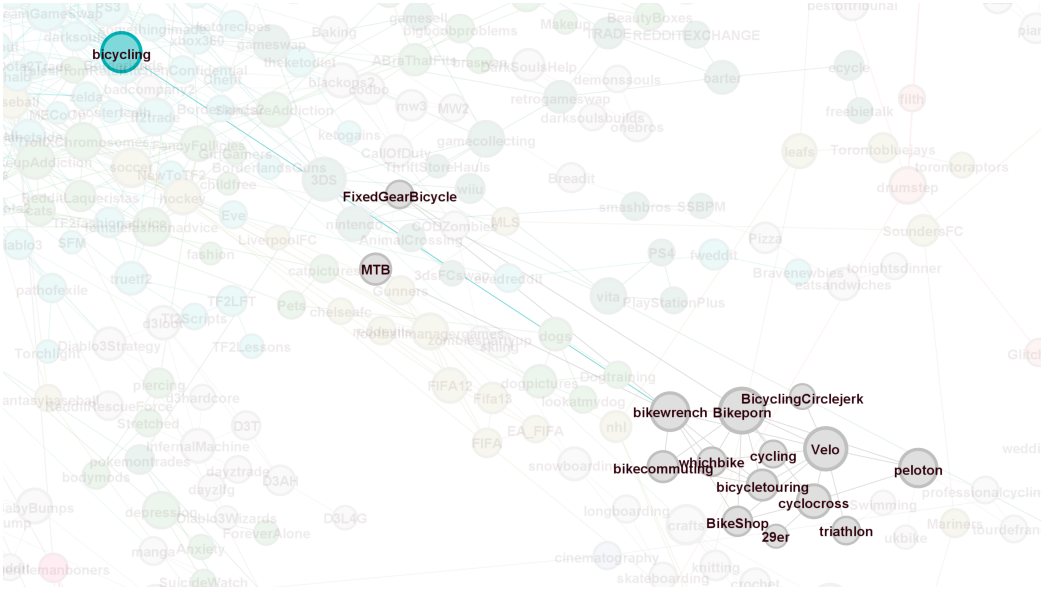


*Figure 8.* A close up of the more popular r/bicycling subreddit connecting to a niche group of biking subreddits.

## References


Bauckhage, Christian, 10 2016. Numpy / scipy / networkx recipes for data science: Spectral clustering.

Blondel, Vincent D., Guillaume, Jean-Loup, Lambiotte, Renaud, and Lefebvre, Etienne, 2008. Fast unfolding of communities in large networks.

Clauset, Aaron, Newman, Mark EJ, and Moore, Cristopher. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

Leskovec, Jure, Lang, Kevin J., and Mahoney, Michael W., 2010. Empirical comparison of algorithms for network community detection.

Lloyd, Stuart. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

Olson, Randal S. and Neal, Zachary P., 2014. Navigating the massive world of reddit: Using backbone networks to map user interests in social media.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.