Cartography and evolution of the Reddit landscape using graph-based methods

Nora Marcoux

Thesis submitted for the degree of
Master in Applied Computer and Information Technology (ACIT)
30 credits

Department of Computer Science
Faculty of Technology, Art and Design

OSLO METROPOLITAN UNIVERSITY

Spring 2022

Cartography and evolution of the Reddit landscape using graph-based methods

Nora Marcoux

Cartography and evolution of the Reddit landscape using graph-based methods

http://www.oslomet.no/

Printed: Oslo Metropolitan University

Abstract

People are spending more time than ever on social media sites and platforms, and a vast amount of information is collected with every click. All of this information can be used to map how users behave and interact, along with their interests. In this short thesis, we attempt to shed light on how the communities on Reddit evolve over time. Two different community detection algorithms are applied to the networks created from the user activity on Reddit in the time span of 2005 to 2019. These communities are further explored to determine the quality of the detected communities, the dynamics that occur over time, and if real-world events are reflected in the user activity on Reddit.

Acknowledgments

I want to take this moment to thank all the people that supported me throughout this thesis.

First, thank you to my supervisor Daniel Thilo Schröder for helping me shape this project and helping me throughout the process. Also a big thank you to my supervisor Pedro Lind for all the constructive feedback, insight, and kind words. Furthermore, I also have to thank Johannes Langguth for his support.

I would like to thank the Simula Research Laboratory for allowing me to work on this project, as well as for providing me with the necessary tools.

The research presented in this paper has benefited from the Experimental Infrastructure for Exploration of Exascale Computing (eX3), which is financially supported by the Research Council of Norway under contract 270053.

Last, but not least, Nicolai and Mika, thank you for all the support and making life a little more enjoyable during a stressful time.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the last decade, more and more of our daily lives are moving from the physical world over to the internet. People gather on social media or forum pages to connect with like-minded individuals, leading to an increased interest in online social networks among the research community. The behavior of individuals or groups on sites like Twitter, Facebook, Instagram, and Reddit can be analyzed to tell us how people behave on these social networks, how stories spread, and how different communities evolve over time. An example of this from recent times is the Russian government accounts on Twitter, with 7 million followers spread between 74 accounts, which would all retweet each other's tweets within 60 minutes (Clayton, 2022). This coordinated retweet network can be used to purposely spread misinformation.

As these social networks grow in both size and number, they can become the starting point for actual world events or phenomena. Examples of this can be seen in the subreddits `r/wallstreetbets` and `r/antiwork`, which gained a lot of traction through Reddit and have become social movements (Kelly, 2021). This has made Reddit even more of an interesting topic for researchers, especially considering all the generated data on Reddit that is available for public use. Another social network that has piqued the interest of researchers is Twitter, but it is a much harder site to gather data from. However, Reddit is a challenging site to gather information from, due to the vast amount of data points that is available.

One of the aspects that sets Reddit apart from other social networks is that it is anonymous and based on interests, meaning that the user signs up to Reddit to be able to join one or more subreddits with a topic they find interesting. This separates Reddit from other platforms, which are more oriented around connecting with people, rather than topics. These subreddits form communities where people who share

characteristics and interests can find each other. The fact that the whole social network is based around communities, and users are allowed to hide behind the safety of a username might lead to very authentic content as people might not be as afraid of getting judged. Furthermore, they also find validation of their opinions from other like-minded people.

Reddit has a tree-like structure that is organized in sub-communities. These sub-communities are on Reddit called for subreddits and are mostly user-created and -moderated. As real communities, they come in all shapes, from a few people in a subreddit to the biggest subreddit with over 140 million members (for Reddit, 2022). When users choose to join a subreddit, they are signaling that they are interested in whatever topic the subreddit is about. Here it is essential to point out that all subreddits are separate entities. This means that there is no direct connection between subreddits because there is nothing that groups or links them together, and Reddit does not show which subreddit the users have joined. The only way to connect users to a subreddit is by looking through the posts and comment history.

In June 2021, Andreas Huber, a master's student at the University of Oslo built a framework for large-scale mining of Reddit data (Huber, 2021). He generated temporal interaction networks on a massive scale from Reddit's Pushift datasets (Baumgartner et al., 2020) that could be used for further investigation. The network he created contains the top 10.000 subreddits and is based on the number of unique users that contribute to a subreddit. A contribution to a subreddit can either be a post or a comment.

As previously mentioned, the network is a temporal network. The temporal network come in the form of time slices and give a time-based view of Reddits evolution. One time slice consists of a sub-network, and there is one for each year, from 2005 until 2019. The fact that this network is dynamic will make it possible to see how the subreddits evolve and how the relationships between subreddits change over the years. From this point on the yearly sub-networks will be mentioned as time slices and all the time slices put together creates the Reddit network.

The aim of this thesis is to look at how communities in the Reddit landscape evolve over time and if these dynamics also can be correlated to real-world events.

1.1 Research Questions

The topic of interest in this thesis is the use of community detection on Reddit data and how the Reddit landscape has evolved over time. This is done using graph-based methods, and the usage of community detection in order to get a better view of how

interests and communities evolve over time. The research questions for this thesis are

> *- Given the temporal dynamic network underlying Reddit's relation, can we explain its dynamics?*
> *-Is it possible to detect behavioral transitions with quantitative changes in the Reddit time slices?*

As we will see, we can answer positively to both questions. During the process of answering the research questions, the following challenges will be handled:

Network Size As previously mentioned, the time slices for the work in this thesis have already been created by Huber through his Reddit Dataset Stream Pipeline. The time slices for the top 10.000 subreddits consist of one time slice for each year in the time span from 2005 to 2019. There were also time slices created for the top 100 and top 10, but these were not created in a yearly interval. The top 10.000 subreddits had a high enough number of time slices to experiment and discover trends.

Number of Edges Many of the time slices have an extremely high number of edges and if the number is not lowered, the community detection algorithms could need several hours to complete. Finding the best criteria for lowering the number of edges, without compromising the result could streamline the process and provide better results.

1.2 Contributions

The goal of this thesis is to propose possible solutions and new findings to the research question described in the section above. These solutions and findings will make contributions in these areas:

Conduct research on a never seen dataset, the dataset used in this thesis has never before been used for extensive research.

The use of thresholds in large complex networks, one of the challenges with the Reddit dataset is the large number of edges that are present. This thesis suggests a threshold that can be applied to remove weak edges in large fully connected networks.

Increase the understanding of Reddit dynamics, there has been little research done on Reddits entire time span. Most research available focuses on a time span of a few years, excluding much of the available Reddit data. To the best of my knowledge, this is the first time community detection has been used on any type of Reddit dataset with

a time span of 2005 to 2019.

Community evolution on Reddit, increased insight into how the communities on Reddit evolve over time. The large time span can allow for a broader view of how communities evolve on platforms, such as Reddit.

1.3 Outline

The outline of the rest of the thesis is as follows:

Chapter 2 is the background chapter, and it is intended to introduce the reader to Reddit and related concepts such as network theory, community detection, different tools that are used, and related works. After reading this chapter, the reader should be familiar with and gained enough knowledge about the mentioned concepts to read the rest without a problem.

Chapter 3 presents data and method. In this chapter, the data that is being utilized is presented and described. As well as previous work done by Andreas Huber to create the Reddit Dataset Stream Pipeline. The time slices for 2007, 2008, and 2009 that were produced by Huber are visualized. The approach for the experiment is also described at the end of the chapter.

Chapter 4 is the findings and analysis chapter and presents the evaluation of the community detection, as well as, the results from the experiments.

Chapter 5 is where the results from the experiments are discussed and also connected to real-world observations.

Chapter 6 concludes the thesis and contains the summary of the results, in addition to limitations and areas of further work that could be of interest.

Additionally there are three appendixes provided:

Appendix A contains larger and more readable versions of the visualized time slices for 2007, 2008, and 2009 that were created in Gephi.

Appendix B is the systematic result for the number of edges.

Appendix C contains the results from the Louvain algorithm.

Chapter 2

Background

In this chapter, the goal is to give an introduction to relevant material. It can be useful to have a basic understanding of the topics to fully comprehend the scope of the thesis.

2.1 Reddit in a Nutshell

If you are a person that spends much time on the internet, then you have most likely heard about Reddit, but what exactly is Reddit? A short answer would be that it is an enormous collection of forums where people can share whatever they want, from current news to cute animal pictures. Reddit describes itself as a network of communities that want to enable conversations and authentic human conversation (RedditInc, 2022a). The site was founded by college friends Steve Huffman, who is the current CEO, and Alexis Ohanian in 2005.

The content on the site is generated by registered members that submit it to user-created boards that we call "subreddits". Posts can then be upvoted or downvoted by other members. The most popular posts are shared on the site's main homepage, where their positioning will be based on the age of the post and the number of interactions, such as up-and-down votes and comments the post has received. It is important to note that there are two different homepages, one where the user is not logged in and one where the user is logged in and will get a personal homepage with posts from the subreddits they are a member of. According to Pew Research Center, Reddit was one of the ten most popular social media websites in the US in 2021 (Auxier and Anderson, 2021).

There are several components that Reddit is built upon and features that can be useful to know. The most valuable and necessary to know are described below.

Subreddits are user-created niche forums that collectively make up the Reddit landscape. Subreddits can be viewed as forums with a unique name, their own URL, and based on a specific topic. The URL of a subreddit can be accessed through `/r/nameofsubreddit`, meaning that, for example, the subreddit named dogs would be accessible under the URL `reddit.com/r/dogs`. Users are free to create new subreddits, join and post as they wish. The subreddits are mostly open for everybody to access and join, but users can be blocked, or the subreddit can be made private. Some subreddits will have a set of community guidelines or rules that the users must adhere to if they want to post in that particular subreddit. There are currently more than 3,2 million subreddits that have been created, but as of October 2021, only about 100.000 of those were active subreddits (Brian, 2021).

Posts are something all Reddit members can do. The user will submit a post to a subreddit, and the users can post in a subreddit their not subscribed to, as long as it is an open subreddit. There are essentially three post types on Reddit: text, image, and links.

Comments are something all registered members can leave on a post or to an already existing comment by replying to it. These comments inside of a post will form a discussion tree (Singer et al., 2014). A post with numerous comments and replies to those comments, meaning that the post has a high engagement, will often be shown on the front-page where popular posts are displayed.

Account. Having an account is required for any interaction on Reddit. Most users will use a synonym and keep their identity unknown, which is generally encouraged by the Reddit communities. It is also common for users to create a "throw-away" account when posting in a specific subreddit or making a more personal post that they do not want to be connected to their "main" account. All the members' posts and comment history is exposed on their profiles, which is done to increase transparency among the members on the site. To create an account on the site, the user must come up with a username that is not already in taken, a password, and provide their e-mail.

Voting is done on either a post or a comment and is something all the members on Reddit can do. The voting functionality is in the form of a down-vote or up-vote and directly affects Reddit's ranking system. The voting and a time variable determine the placement for both posts and comments inside a post.

Karma is used on Reddit to establish credibility due to the site's promotion of anonymity and adds to users' credibility. Every profile on the site will display the user's karma and is a representation of the user's activity and how many upvotes they

have received. It is often the case that users with more karma are perceived as more trustworthy and more trusted members of the community. There are two ways of earning karma, one is by the number of up-votes received on a post, and the second is by the number of votes received on a comment. These two are shown separately. It is also possible to lose karma by getting downvotes on a post or comment.

Awards are something that members can give each other if they find their content good or feel like that member deserves recognition. The award will be given to a particular post or comment, and the member that created the post or comment is rewarded. The benefits of receiving a reward can be Reddit coins, which they can use on the site, or a few days of an ad-free Reddit. Reddit coins are a type of virtual currency that can be used to buy awards for other members of the site.

The different types of rewards that can be gifted are reaction awards, which as the name suggests, represents a feeling or reaction to the post. The second type of award is Medals, it started as silver, gold, and platinum, but the medals available now are Silver, Gold, Platinum, Argentium, and Ternion All-Powerful. Silver is the medal that costs the least to gift, and Ternion All-Powerful is the most expensive. There are also premium awards that are only available to members with a premium membership, which costs a monthly fee. The last type of award is a community award created by moderators for their subreddit and is only available in that community. A certain percentage of the cost of the award, in the form of Reddit coins, will be given back to the subreddit (RedditInc, 2022b).

Administrators or admins are Reddit site administrators that are employed at Reddit. These users will have a red icon next to their usernames, and the icon will also be visible on the admins' posts and comments. The admins on Reddit have several functions connected to their role. They are there to enforce the platform's rules, and keep everything up and running, including content policy, and communication aspects on the platform (RedditInc, 2021).

Moderators are Reddit members that have volunteered their time to help inside a subreddit. They can be the creator of the subreddit or have been added on as a moderator later. The moderator's role is to maintain a subreddit and make sure the rules inside that community are being upheld, as well as Reddit's own community guidelines. The moderator has the power to remove posts, ban members that are not upholding the rules of the subreddit, make official posts inside the subreddit as moderators and add other members as moderators (RedditInc, 2022c).

2.2 Network Theory

Mark Newman (2018) defines networks in their simplest form as a collection of nodes or vertices connected by edges or lines, composing a network G defined as:

$$Regular\ network : G = \{V, E\} \tag{2.1}$$

where V is the nodes and E is the edges. The edge represents the interrelations between the nodes and shows to which degree the nodes are "related." The type of systems made up of nodes and coupled by edges are all around us and can be modeled as networks. The structure of a network describes how the graph is built up and makes it easier to understand while also making it easier to optimize and predict the nature of the system (Holme and Saramäki, 2012). Networks and graphs can be viewed as mathematical tools that can be used to model interactions. The terms are often used interchangeably, but a common distinction is that a network can be defined as a graph where the nodes and edges have been given attributes.

2.2.1 Important Types of Networks

Networks are often divided into three different categories based on the characteristics of the edges:

Undirected Networks Undirected networks will have edges without a direction, meaning that the edges can traverse in both directions and have a two-way relationship.

Directed Networks Directed networks will have edges with a direction, meaning that the edge can only be traversed in one direction and reveal a one-way relationship.

Multinetworks Multinetworks are networks that can have multiple edges, also called parallel edges, connecting two nodes.

As already stated, there are several types of networks, and they can be found in many academic disciplines. They are so interdisciplinary because of the flexibility of what can be considered a node, for example, cities, television shows, or stores. Below are some of the most relevant networks for this thesis described:

Temporal Networks
Temporal networks, also called dynamic networks, dynamic graphs, time-ordered networks, and evolving networks are network structures that evolve. This means

that parts of the network might appear and disappear as time progresses. They have a general network structure that contains time information about the action of the members along with the interaction that occurs between them (DiBrita et al., 2021).

Temporal networks can be defined as:

$$G_T = \{V, E_1, \ldots, E_n\} \tag{2.2}$$

where V represents the nodes in the network and E_n is the set of edges that change in each time step.

These types of networks can be quite complex, and because of this, they can be difficult to study. However, these traits make them prime candidates when studying dynamic systems (Li et al., 2017). It is the time information in these networks that can depict how communities and information spread, and narrate the movement of members in communities.

Interaction Networks Interaction Networks are networks that are built upon the users interacting with each other. A good example that illustrates this type of network is how users interact on Twitter when replying to each other's tweets. Here the users on the platform interact by writing to each other directly on the platform from their user accounts. The network grows more extensive as more and more users tweet at new people. In this network, the people will be the nodes, and the users' interaction in a subreddit will create the edges.

Participation Networks Participation Networks are networks where people are required to participate to be a part of the network. These types of networks are similar to interaction networks but differ from them in the context that the user has to find the place where the interactions occur. An example of this type of network is Reddit, the users interact with each other inside of the subreddits, but they can only do so inside the subreddit. An act of participation has to occur before the interaction, and it is the participation that creates the edges in the network.

An example of a participation network that can occur on Reddit:

Subreddit	User
News	A
Music	A
Music	B
Books	B
News	C
Books	C

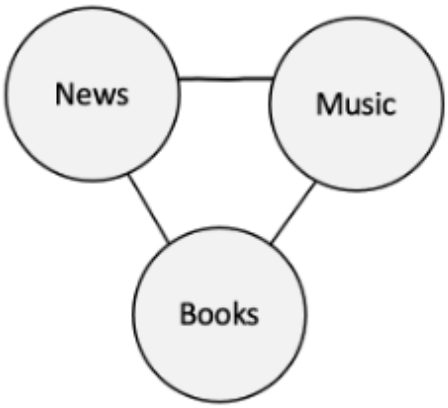


Figure 2.1: Participation Network

The participation network is created by the posting activity of the users and an edge can be drawn between each subreddit a user posts in. The network created in figure 2.1 would be created if:

1. User A →posted in the subreddits News and Music.
2. User B →posted in the subreddits Music and Books.
3. User C →posted in the subreddits News and Books.

2.2.2 Network Terminology

In network theory, there is certain terminology that should be known and shows up repeatedly in the following chapters.

Neighbour: two nodes are neighbours if an edge connects them.

Degree: a degree of a specific node is equal to the number of edges that are connected to it. There is in-degree, out-degree, and regular degree.

Singleton: also called isolates are nodes without any neighbours (0 degrees).

Density: is the ratio of the actual number of edges to the number of possible edges in a network.

Path: is a sequence of nodes connected by edges.

Path length: is the number of edges in a path or the sum of their weight, if the edges have a certain weight.

Cycle: a path that starts and ends at the same node. Note that all cycles are paths, but not all paths are cycles.

Connectivity: two nodes are connected if there is a path between them. In the context of a network, the network is connected if all the nodes are connected, e.g. there is a path between all pairs of nodes.

Connected Components: is a subset of nodes of the network that are connected.

Conductance: is used to measure how well-knit a network or community is and can be used to look at the surface area to volume ratio (Griesemer and Schlessinger, 2018).

2.3 What is a Community?

Most of us are probably familiar with the concept of a community, Wasserman and Faust (1994) defined a community as a *"subset of actors among whom there are relatively strong, direct, intense, frequent or positive ties"*. This is a good definition of a community, but it might not be fully applicable to communities in social networks.

There is no proper universally agreed-upon definition of a community in network theory, and the definition will often vary depending on the system or application at hand. However, a factor that is agreed upon is that a community needs to display the property of connectedness and that there needs to be a path between each pair of nodes (Fortunato, 2010). Another main feature is that the edges within one community typically have a higher density than the edges outside of the community.

2.4 Community Detection

Community detection is used for the purpose of dividing a network into communities. This is to discover divisions and relationships that exist in the network. These communities will have edges that share common properties and have a similar role in the network, meaning that they are most likely connected to each other than to members of other groups, but different patterns are also possible (Fortunato and Hric, 2016).

Real-world networks exhibit inhomogeneities, meaning that they reveal a high level of order and organization and will often have a community structure (Fortunato, 2010).

2.4.1 The Concept of Partition

The purpose of community detection is often to divide the network by disjointing the set of nodes. The action of disjointing the set of nodes is called a Partition

$$P = C_1, ..., C_i$$

and serves the purpose of dividing the network into communities to ensure that each node belongs to one cluster. When a network is divided into overlapping communities, it gets called a cover and not a partition (Fortunato, 2010).

2.4.2 Modularity-Based Methods

In 2004, Newman and Girvan introduced the concept of modularity as a way of measuring the quality of networks that have been divided into community structures (M. E. Newman and Girvan, 2004). Networks with a high degree of modularity will display densely connected communities with a small number of connections between them (M. E. Newman and Girvan, 2004). The range for modularity spans from $-\frac{1}{2}$ to 1, where one is a highly modular network (Brandes et al., 2007).

The modularity in a network can be altered by moving the nodes to other communities (M. E. Newman, 2006). Moving a node is done by placing the node from one community to another, basically assigning the node with a new community ID.

Modularity essentially compares the total edges inside a community with the number one would expect to be present in a random network with the same number of nodes. The higher the number within the community and the lower the number is outside the communities, the higher is the modularity in the network. Modularity does, however, have one flaw, it can have a non-trivial bias when it comes to identifying communities of different sizes. (Fortunato and Barthelemy, 2007). This bias is present in all of the algorithms that use modularity, and this bias affects the edges connected to small communities. For example, two small communities can have fewer expected edges than one community if the network is large, which could lead to a merge between the two communities if there is even one edge connecting them.

This problem causes the algorithms that use modularity to find communities as the networks become more extensive and are called the resolution limit (Fortunato and Barthelemy, 2007). This means that smaller communities are being fused into large networks, while on the contrary, this would not happen if it was a smaller network. This bias is hard to avoid and should be somewhat expected when using modularity.

It can also be essential to note that the larger the network, the higher the modularity

tends to become (Good et al., 2010). This should be taken into consideration when comparing the modularity of different networks, as it might not be wise to compare the modularity value when trying to determine which of the networks has a more robust community structure.

Louvain Method

The Louvain method for community detection is one of the most frequently used algorithms to detect communities. The algorithm requires network data, where for example, the people are the nodes, and the edges are their connections to each other (Blondel et al., 2008). However, it is essential to note that the Louvain method can be applied to any other dataset where there is a connection that forms a community.

The Louvain Method uses the concept of modularity to measure the strength of a proposed community. Modularity is, as earlier mentioned, used as a measurement in networks and is used to measure how dense internal nodes are compared to the nodes outside of the community. High modularity means that the interior nodes are dense and that the nodes in the modules are connected. Low modularity implies that there is little to no connection between the nodes in the modules (Blondel et al., 2008).

The algorithms approach for maximizing modularity for the networks is very straightforward. Initially, each node is assigned its community, the nodes that are alone in their community are called singletons, and if there is a partition where all the given nodes are singletons, it is called a singleton partition.

The nodes are identified by their node ID and are assigned to a community, where the nodes with the same node ID form together to a community. The name of the community will get the same ID as the member's node ID.

The Louvain Method has two steps:

Local moving is where the algorithm tries to maximize the modularity, and this is done by eagerly choosing different moves. The Louvain algorithm is visualized in figure 2.2. The nodes are moved one by one until the most significant degree of modularity is achieved. If there are no more moves that increase the modularity available for a node, then the node is ignored. If a node can be moved to increase the modularity, all the nodes will be considered again until no more moves are left.

Aggregation occurs where the new network is built. One node is inserted for every community, and a weighted edge will be placed for all the edges from the first network. The weighted edge will be placed between the belonging nodes and have the same

weight for both. This leads to the edges inside the same community becoming loops and is why loops are counted twice.

This loop can cause multiple edges between nodes but adding an edge's weight to an already existing edge, as an alternative to inserting a new one, is a possible way of steering clear from this.

Modularity is not affected by aggregation, seeing that the sum weight of the edges and size of the community stays the same. However, the aggregation step does allow the local moving step to move several nodes, seeing that the nodes in the course network constitute several nodes in the original network (Traag et al., 2019).

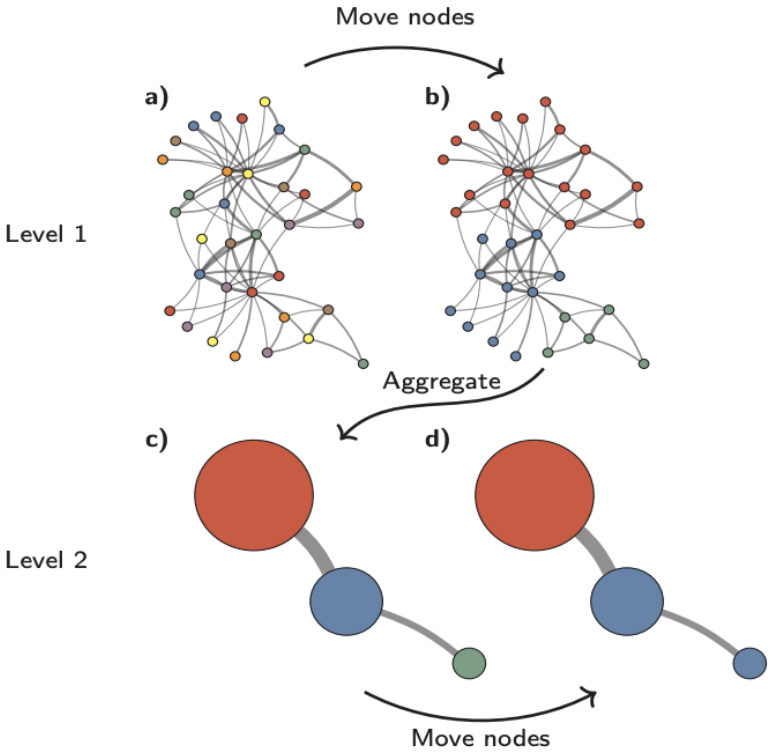


Figure 2.2: Illustration of The Louvain Algorithm
Image taken from Traag et al. (2019). a) The nodes start as singletons. b) The communities are identified by local moving. c) Aggregating the new network. d) The process of b and c are repeated until there are no more changes.

Leiden Method

The Leiden method is considered to be an improved version of the Louvain method. Traag et al. (2019) concluded in their article that the Louvain algorithm has a flaw that can lead to the finding of poorly connected communities or, in the worst case, disconnected communities. Nevertheless, they managed to fix this flaw in the algorithm by introducing another step, the refinement step.

The refinement step aims to optimize the partitioning done in the local moving step, and the refinement step is placed in the middle of the local moving and aggregation step, see figure 2.3. This step splits the communities into multiple sub-communities so that those sub-communities can be independently moved in the next step. Introducing the refinement step increases the chance for more beneficial moves.

While the aggregated network in the Leiden algorithm is based on the refined partition, the first partition done by the aggregated step is the one used by the local moving step. This is done so that the aggregated network can contain numerous nodes belonging to the same community and not just one node representing an entire community as it is possible with the Louvain algorithm (Traag et al., 2019).

Traag et al (2019) also display a minor change to the local moving step that increases the performance of the Leiden method. In the Louvain algorithm, the local moving step will re-evaluate all the nodes when it has been through a loop, and changes have occurred, while the Leiden algorithm will only re-evaluate some of them. All the nodes in the local moving step will be put in a queue to ensure that all of them are evaluated at least once. If a node is sorted into a community, then all the node's neighbours that are not in the community will be added to the queue unless the node has already been added. This step allows the neighbouring nodes to follow the sorted node and prevents the nodes with unchanged neighbours from repeatedly being considered by the algorithm, as they are unlikely to be moved. However, it should be pointed out that even though the neighbouring nodes have not been moved does not mean that the optimal community for a node has not changed (Brandes et al., 2007).

In the partitioning done in the refinement step, every node will start out as a singleton and will only be evaluated at the most one time. Only the nodes that are singletons can be moved, and this is to make sure that communities will not be disconnected, as edges incident to singletons can not act as a bridge between components in a community (Traag et al., 2019). Also, it is only the nodes that are strongly connected within their community from the first partition that can be moved. If these two requirements are met, then the modularity for strongly connected neighbouring communities is

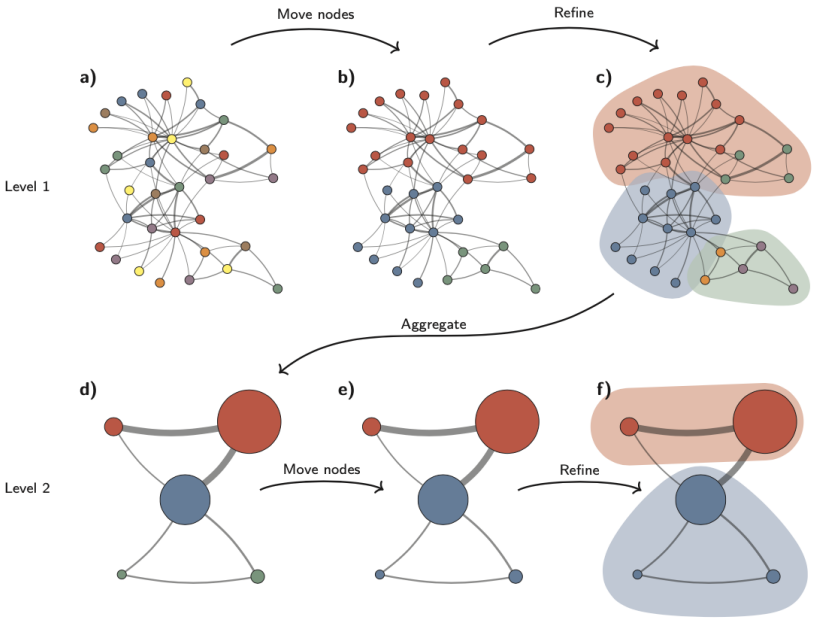


Figure 2.3: Illustration of the Leiden Algorithm
Image taken from Traag et al. (2019). The refinement step divides the green and red communities into two respectively sub-communities.

calculated in the refinement partition.

Unlike the Louvain algorithm, which should only be run once, as it is shown to worsen and yield less connected communities or disconnected communities, the Leiden algorithm can and should be run numerous times with the latest partition from the iteration as the starting partition for the next run. Furthermore, Traag et al. showed that the Leiden algorithm continuously improves the partition with each run until the point where there are no more changes that can occur.

2.5 Pushshift Data Sets

Reddit is open-source and allows for the use of the content on its site as long as it does not violate its guidelines. They also have an API that developers or users can use to obtain data from Reddit. The drawback of using this API is that it only allows for 100 submission requests (Singer et al., 2014). This means that if large amounts of data are needed, it is not a feasible solution for most people. Doing it would be very time-consuming and this is where Pushshift comes in.

Pushshift is a platform created in 2016 to provide and create more functionality and search options for comments and posts on Reddit. The platform collects all the Reddit data and makes it available through an API. One of the benefits of using the

Pushshift API is to make querying easier, which also includes a full-text search against comments, posts, and submissions (Huber, 2021).

Today, Pushshift is maintained by one of the creators, Jason Michael Baumgartner. The platform is popular in the research community and has between 2016 and 2019 been used in more than 100 published peer-reviewed papers (Baumgartner et al., 2020).

2.6 Experimental Infrastructure for Exploration of Exascale Computing

The enormously increased demand for computing power in recent years has incited the development and research of the futures exascale computers, and these machines will have the capability of performing at least billion billion (10^{18}) floating-point operations per second (eX3, n.d.). The next generation of supercomputers, also called high-performance computing (HPC), will consist of heterogeneous state-of-the-art processing nodes with large cores, accelerators, and deep memory hierarchies. Furthermore, all the levels of hardware are expected to be heterogeneous and organized in an advanced communication topology. An example of this is in order for nodes that can include a set of algorithms or machine learning method that is tailored for a specific processor.

The Experimental Infrastructure for Exploration of Exascale Computing also called eX^3, is a collaborative project between Simula Research Laboratory and its partners and is an essential tool for HPC researchers in Norway. The goal is to better national resources and introduce researchers to the power that is available in exascale computing while also staying up to pace with the latest research in exascale computing.

The eX^3 infrastructure is set up to support activities oriented around exascale research. It acts as an experimental testing ground, where the goal is to assess the latest computing devices and interconnect solutions. Some of the equipment that is available for usage are different kinds of processing nodes, each of them consisting of different processors from manufacturers with either ARM or x86 architectures. These can be set up by the user with primary and secondary storage with configurations that are unique.

The users of eX^3 have the responsibility for the data they put on or generate on eX^3 resources and to make sure it adheres to the relevant regulations and that the proper security measures are taken to protect it. The total data storage capacity is 500 TB based on SDDs and spinning disks. To keep as much of the storage capacity available,

users are encouraged to only temporarily store their resources on eX^3 in connection to current experiments, seeing that eX^3 does not offer long-term storage options. All the details about eX^3 in this section are found in the eX^3 documentation, for more details see the eX^3 documentation (eX3, n.d.).

2.7 Gephi

Gephi is open-source software that is used for graph exploration and network analysis. The software supports almost all network types; this includes temporal networks, dynamic networks, complex networks, and hierarchical networks (Khokhar, 2015). It provides interactive visualization that allows the user to identify and discover unknown features about their network and assist in the exploratory phase by providing an extensive toolbox with extensive features. Gephi also has functionality for changing the format and altering the size and colour of the nodes, edges, and labels based on multiple criteria such as edge weight, degree, or modularity. Users of the software are also allowed to create their own layout algorithms or filters through Gephi's plugin option.

Gephi's graphical user interface (GUI), which can be seen in figure 2.4has the graph visualization module in the center, the layout module to the right on the screen, and filtering, statistics, context, algorithms, and clustering on the left side of the screen. The GUI can be seen in the figure below:

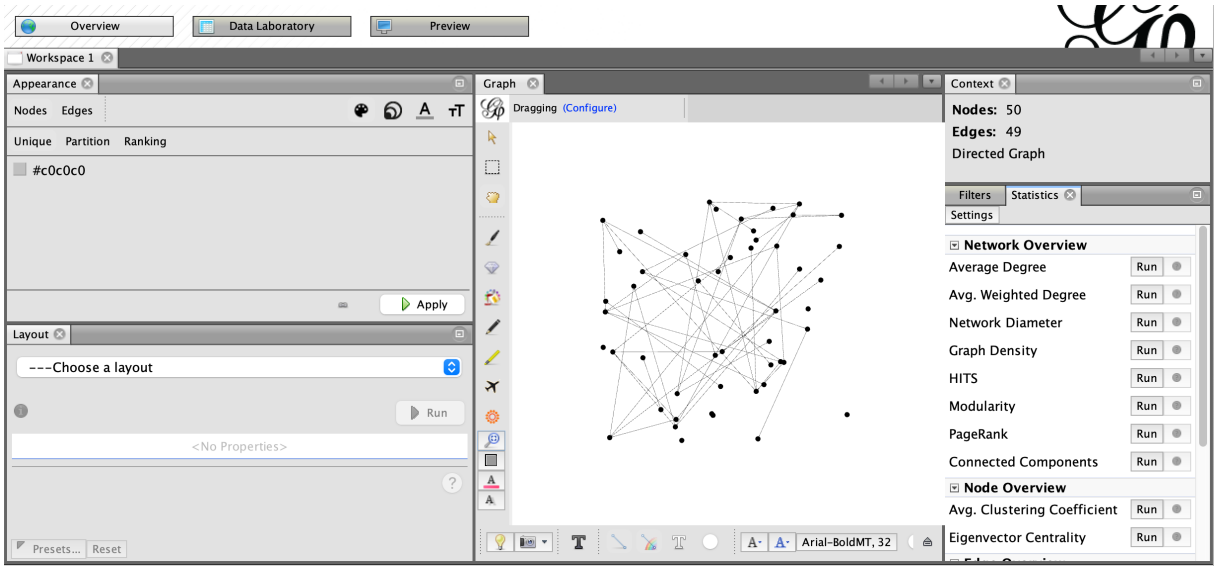


Figure 2.4: The Gephi user interface.

Gephi also has three different modes for the user to choose from, these modes can be seen in the upper left corner of the figure2.4. The three tabs, as seen in the figure above,

are overview, data laboratory, and preview. The first screen that appears to the users when starting Gephi is the overview mode which can be seen in figure2.4. The features for the overview mode have already been described in the section above and are all oriented around graph exploration and manipulation. The data laboratory mode is used for accessing and manipulating the network information in a tabular format. The last mode, preview, is used for altering the look of the final network and exporting an image of the network into the wanted format such as PNG or PDF.

The visualization of the graph is done with a 3D rendering engine that provides real-time visualizations when displaying, altering, filtering, etc., are done to the network. This is possible in Gephi since the software uses the GPU to display the graph, allowing the CPU to handle the other computational tasks. Gephi does additionally support the visualization of dynamic networks where the nodes and edges will disappear and reappear over time, as well as gain new properties over time.

2.8 Related Works

This chapter discusses and summarises some of the most relevant published works related to this thesis.

Singer et al. (2014) present their research where they examined how user submissions on Reddit have changed over the period of five years. They also looked at commenting and voting patterns and how they evolved. To research these topics, they used all Reddit submissions from January 2008 until December 2012 and an additional succinct user survey for supplement. Their research suggests that Reddit has transformed from a gateway to the web to its own self-referential community. Inside this self-referential community, an internal focal point rewards content generated from the users, such as images, videos, and textual content over content from external sources.

Buntain and Golbeck (2014) want to identify user roles inside the individual sub-communities (subreddits) on Reddit and to do this, they explore the posting behavior inside of these sub-communities. The objectives for their research were to answer if the answer-person role was present, if they could automatically identify this role and if there was any notable multi-communication participation on Reddit. They used existing research and built their own toolkit to crawl chosen communities on Reddit and extract posting statistics. Their research concluded that the answer-person role is indeed present on Reddit and that users mainly only participate inside of their own community and rarely venture outside of it.

Mensah et al. (2020) investigate factors that could possibly characterize how communities evolve on Reddit. Their research was oriented around identifying the different patterns that are present in the evolution of the communities and how these patterns differentiate from the different patterns that are identified. The data they used was from subreddits created between 2014 and 2015, and they focused on subreddits that would only allow text posts. They discovered that the linguistic and interaction style of the users in the community could highly affect the different stages of the evolution in the community when it comes to the number of active users. Their investigation also showed that there are three evolution patterns: (1) communities that increase in the number of active users, (2) communities that decrease in the number of active users, and (3) communities that alternate between increasing and decreasing in numbers of active members. It was also discovered that communities' ability to attract new users was highly connected to the number of active users within the community.

Olson and Neal (2015) combine the work from Serrano & Vespignani (2009) in network backbone extraction and Blondel et al. (2008) work in community detection to map

user interest on Reddit. The data used for this study was mined through the open-source API provided by Reddit. In total, 876,961 user names from active users from the period between January 2013 and August 2013 were collected, along with 15,122 subreddits. In addition, all of the users' 1,000 most recent posts or comments were collected and registered as interested if they posted or commented in the same subreddit at least ten times. The result of their work was the discovery that Reddit has a modular community structure that is small-world and scale-free. They also made their network dataset available for further research.

Griesemer and Schlessinger (2018) used the network dataset made available by Olson and Neal to build an interest network and analyze user activity. To find communities on Reddit, they utilized three different community detection algorithms. These were the Louvain method, spectral clustering, and K-means clustering. To determine the quality of the clusters produced by the three algorithms, they used modularity and conductance. Their results were that the Louvain method had the highest modularity and relatively balanced conductances, meaning that the Louvain method produced the best clusters out of the three different methods.

Khandelwal and Xu (n.d.) focus their research on language-based community detection and how well language can be used to find communities on Reddit. They used prediction tasks that were supervised and community detection to determine the quality of the work. Their research concluded that language could, to some degree, be used for modeling user interactions and suggested that further work should be done in this field of research.

Datta and Adar (2019) research focuses on the conflicts between subreddits and the dynamics of community-to-community conflicts on Reddit. To find these community-to-community conflicts, a conflict network was constructed and created by using all the comments on Reddit from 2016, in a total of 743 million comments from 9,75 million unique users. Their approach to creating conflict graphs was described and how the conflict edges were constructed. Their research provided much insight into conflicts in the Reddit landscape, such as that 77,2% of conflicts between subreddits are reciprocated. In addition, their usage of temporal patterns revealed that subreddits that would target multiple subreddits at the same time would more often shift their focus to another conflict.

Medvedev et al. (2017) complete a survey where they have mapped the main academic research related to Reddit. The research has been divided based on whether it focuses on posting or users. The main research areas that are listed in connection to posts in the survey are popularity prediction and generative models for discussion trees. For

research related to the users, the main research was about activity patterns, community loyalty, trolling and hate speech, and inherent networks of communities. The goal of the survey was not to list all research that has been completed about Reddit, but rather to provide a representative sample.

All of the related works use Reddit data in their research and focus on community-related topics, whereas this thesis differentiates itself by the length of time. The time span of this thesis is from 2005 until 2019. The research article with the longest time span in the related works was by Singer et al. and had data collected from 2008 to 2012.

Chapter 3

Data and Methodology

3.1 The Data

There are two sets of CSV files that make up the graph, one for the nodes and one for the edges. The nodes consist of the subreddits and the edges are created when a user either subscribes or posts in two different subreddits. The time span is from 2005 to 2019, which means that there is a total of 28 CSV files. When the datasets were created by Huber (2021), they were filtered by users in subreddits and contains the top 10.000 subreddits for each year.

The variables of such datasets include: **Node Scores**

- **U**: is the sum of all the edges U_{i_j} connected to a node. U_{i_j} represents the number of users that posted in both of the subreddits. i and j represent two connected nodes.
- **Local clustering coefficient**: is as the name suggests the local clustering coefficient and is used to measure how connected the nodes are in the graph.
- **Degree**: is the sum of the edges connected to the node.
- **Weighted degree**: is defined as U/degree.
- **Degree degree**: is the sum of the edges connected to the neighbors of the node.
- **Weighted degree degree**: is defined as sum(U)/ degree degree.

Edge Scores

- **The number of contributions in two different subreddits**: The number of users that contribute in i and j. U_{i_j} is the number of users that posted or commented in both of the subreddits at least one time.
- **Source and target degree**: the degree of distribution for i and j.
- **Weighted source and target degree**: is the sum of U connected to the source i and the target j.
- **Average weighted source and target weight**: is the number of users connected to i or j node divided by the degree of the opposite node.
- **Edge weight**: is the number of contributors that are contributed in the connected subreddits in relation to the number of users connected to all the neighbors of i and j.

3.1.1 Time Slice Values

The network consists of several time slices and the values in each of the time slices are listed in table 3.1. Each of the time slices represents a year in the Reddit Network.

Table 3.1: Network Information

Year	Node	Edges	Max. Weighted Edge	Min. Weighted Edge	Connected
2005	1	0			
2006	35	413	18.354490980813875	0.0099183197199533	True
2007	43	373	12.20067243852107	0.0018309701909121	True
2008	4 358	156 943	220.49239536474843	0.0133954597601711	False
2009	9 221	710 661	387.795204861323	0.0117590501755794	False
2010	10 000	2 128 264	480.2393782487155	0.0059659785832522	False
2011	10 000	6 973 083	472.639634425336	0.0024225509895823	False
2012	10 000	16 487 889	396.5042531732811	0.0011151542117903	True
2013	10 000	28 053 528	434.61167931809206	0.0007518488392034862	True
2014	9 999	38 992 606	292.4090687336958	0.0006287716155965297	True
2015	9 999	43 416 539	351.62276884446646	0.0005161096432701868	False
2016	9 999	44 129 304	387.0991772862282	0.0004608075747856256	True
2017	9 999	49 415 020	284.3704879642765	0.00053410326191156	True
2018	10 000	49 803 235	411.6544884897699	0.0004701613381037132	True
2019	10 000	49 907 181	0.0099183197199533	0.0006143149233932052	True

When the time slices were created by Huber (2021), they were all intended to have

10.000 nodes. This turned out to not be possible for all of them, as before 2010 there were not yet enough edges connecting 10.000 nodes. From 2014 to 2017 it was not possible to create a fully connected graph structure with 10.000 nodes for the same reason.

3.1.2 Degree Distribution

As previously mentioned, the degree of a node is the number of connections the nodes have to other nodes. The degree distribution is formed by counting the degree for each node in the network. Most networks will have a skewed degree distribution, meaning an asymmetric result. A left-skewed degree distribution means that there are many nodes with a high degree and few with a low degree. In a right-skewed degree distribution, there are many nodes with a low degree and few nodes with a high degree.

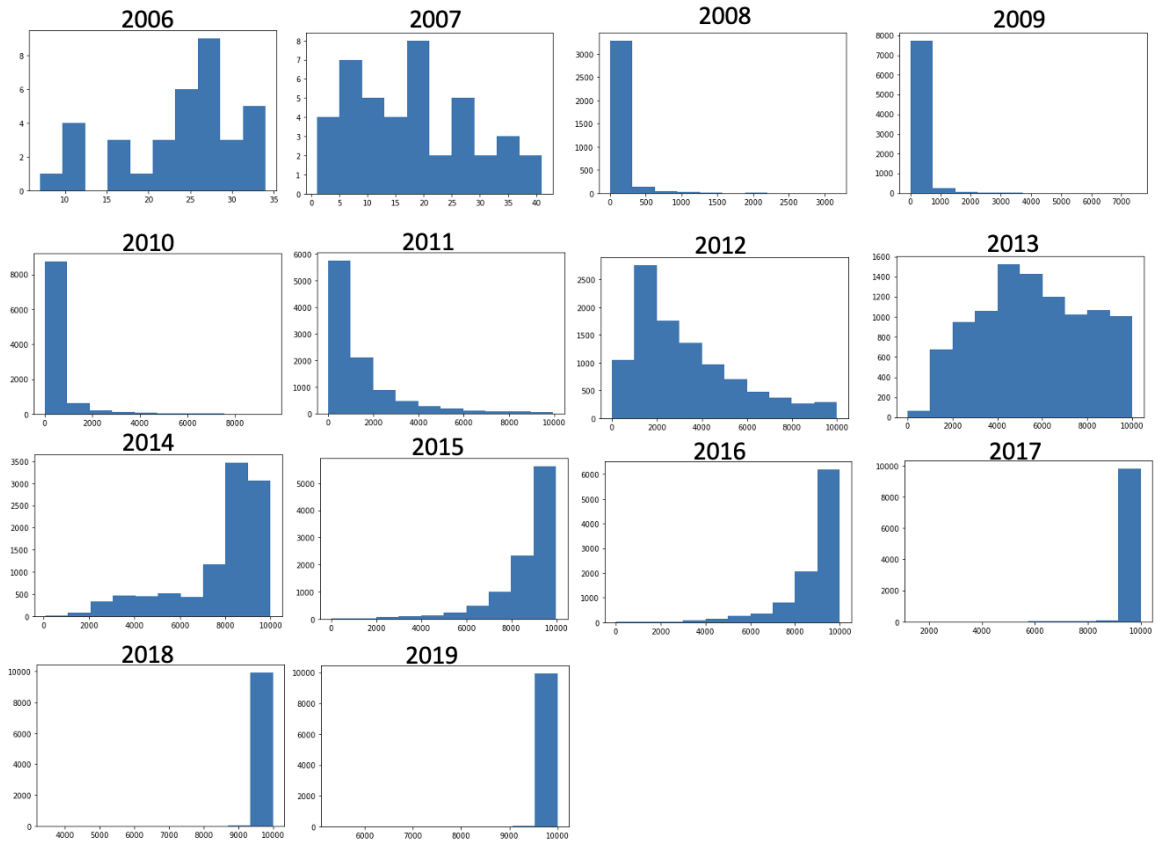


Figure 3.1: Degree of Distribution from 2006 - 2019.

The histograms display the degree distribution for all the nodes in the network. The x-axis displays the degree and y-axis the frequency.

In figure 3.1, we can see that the degree distribution for 2006 was slightly skewed to

the left and in 2007 it was slightly skewed to the right. This indicates that there are fewer nodes with a high degree in 2007, something that is supported by the decrease in the number of edges and increase in nodes in table 3.1. From 2008 to 2012, the degree distribution was right-skewed and becomes more symmetric for each year until 2013. The degree distribution is considered symmetric if the degree distribution is evenly spread around the mean. From 2014 to 2017 the degree distribution becomes increasingly more left-skewed, and in the time span of 2017 to 2019 the degree distribution remains completely left-skewed.

The degree distribution and how it evolves over time imply that more and more people are joining Reddit and that there was an increase in activity throughout the years. Spanning 2017 to 2019, essentially every subreddit was connected. However, the degree distribution does not take the weight of the edge into consideration and every edge is counted equally.

3.1.3 Ethical Assessment

When doing research, it is important to consider the ethical implications, especially when the work is centered around collected data. The research in the thesis is using publicly available data and does not look at the behavior of the individuals on the platform, but rather at how the different communities interact internally and with each other. The RDSP does not collect any type of personal or traceable information about the users on Reddit, such as the IP address of the user, email, browsing history, etc. The data that was used to create the time slices do not have any identifiable properties like those just mentioned. The dataset contains information and attributes about the network, so the name of the subreddit, number of posts, and comments. To be able to create the edges between the subreddits, the posting activity of the members in the subreddits has also been collected. It is important to note that none of the content in the posts was collected and was never used. The focus of the collected Reddit data is overall user activity and not the behaviour of any specific group or person.

Violation of the community

As already mentioned, users can often feel a strong connection to their subreddit and attach the image of their community to their own self-image. Identifying the connection between subreddits could lead to revelations about the members and the subreddits they visit. This could potentially lead to them feeling personally insulted or attacked if there are any "negative" discoveries in their eyes about their subreddit. This new information could to a member of a subreddit feel intruding and invasive, especially if they have a strong connection to their subreddit.

3.2 Building an Interaction Network from Reddit's Database

3.2.1 Reddit Dataset Stream Pipeline

The Reddit Dataset Stream Pipeline (The RDSP) is a program created by Andreas Huber in connection to his master thesis and collaboration with the UMOD project (Huber, 2021). The idea was that the program should be able to read, filter, and transform the dataset from Reddit in parallel. The Reddit data is gathered from the Reddit Pushshift API, which allows for faster and easier querying than Reddit's own API. The program is also adept at providing the dataset as a stream for other programs by using the UNIX named pipes. The RDSP is set up using the Akka Streams standard and this was chosen as it seemed like a good option for making sure the data could be processed now and in the future.

The RDSP can be accessed with a command-line interface with the Scopt command-line parsing library. This allows for all possible parameters to be used and all the command-line arguments are written to a case class. Scopt also supports the use of arguments, options, and commands, while also allowing for the specification of rules for input validation. The available commands are; the dataset directory, concurrency settings, file inclusion, exclusion filters, and an option for compressing the output. Furthermore, the user must choose whether the model should run in statistic mode or pass-through mode (Huber, 2021).

Statistic mode was created to determine which of the subreddits are important, this will allow for some limitations of the network size. It has two methods for this, one is by counting the number of users on Reddit that contributed to at least one of the subreddits and the other method is by counting all the user contributions that occur per subreddit. The data that is gathered through the statistic method is stored as a CSV file.

The pass-through mode was created for the functionality of streaming the whole dataset through other programs via UNIX named pipes. The pass-through mode makes it possible to pass the entire dataset as one stream to another program if it works on the dataset. The pass-through mode uses all the available CPU cores to efficiently decompress the dataset while it is being streamed.

It is possible through the pass-through mode to emit a small number of subsets to a CSV file, but it is also possible to emit, without size restraint, the original NDJSON. In this mode, it is also possible to only get the wanted subreddits by filtering and ignoring

authors that are deleted. The reason deleted authors need to be ignored is that they would all be stored under the same name "deleted", which would lead to a false result, seeing that a large number of subreddits have posts written by now-deleted users. The filtering of rows is the only form of aggregation of the entities in the dataset. In this mode, it is possible to select which entities, meaning the authors, submissions, and comment, that is going to be processed. The pass-through mode allows for the selection of multiple entities, which would lead to multiple streams being started.

3.2.2 Network Building

The network used in this thesis was built from CSV files that were provided by the RDSP and these datasets contain information about the users' activity and which subreddits they posted in. The figure below shows the process that was needed when making the network.

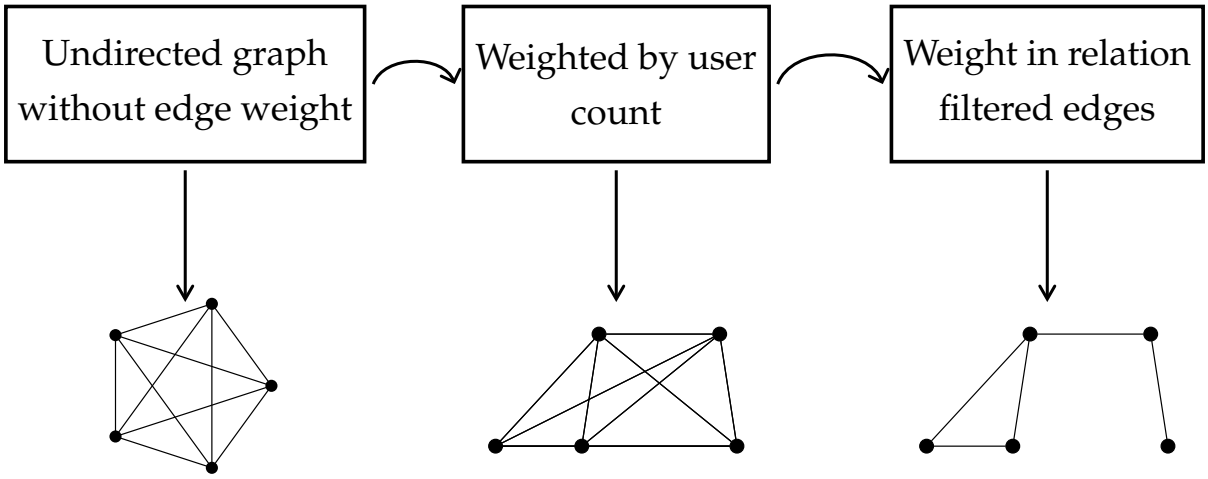


Figure 3.2: Evolution of the weighting function.
Taken from Huber (2021.)

The figure above is an illustration of the network building process created by Huber and can be used to simply explain the building process of the time slices. Huber used the terms graphs and time slices interchangeably in his thesis, as seen in figure 3.2.

The first step in the process demonstrates an undirected subreddit time slice in its simplest form. An edge is created when a user has posted in two subreddits and the probability of a fully connected network increases with n number of included top subreddits. This is because the more subreddits included the more likely it is that at least one user posted in one subreddit or more.

The second step is to weigh the edges, this is done by counting the number of users that posted in the two subreddits where the edge is connected. This is a simple form

of edge weight, but it does provide information about which of the subreddits belong together.

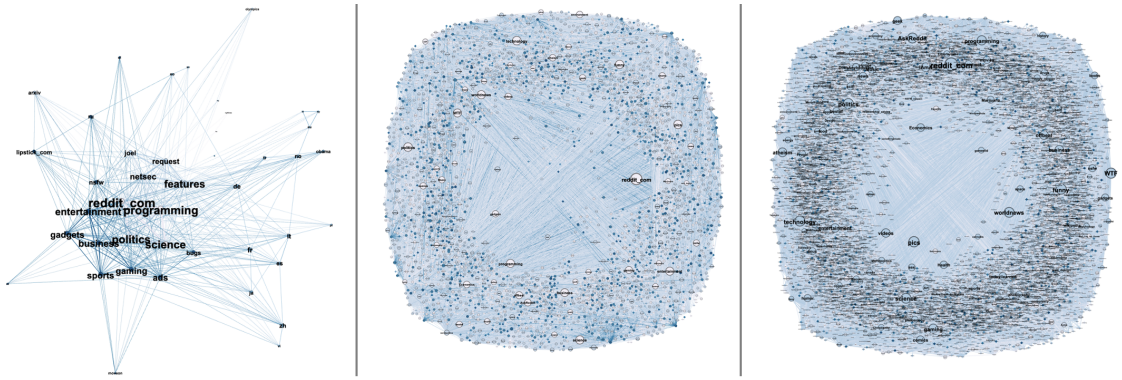
The third step is to determine the importance of neighbors. This is done by filtering out the neighbors with low weights. When this is done, the network will be able to display which subreddits that have a strong connection and the interaction between the different subreddits.

3.3 Time Slice Visualization

To get a better understanding and display the growth in the time slices that occur over the succeeding years, they have been visualized with Gephi, see 3.3. The time slices from 2007, 2008, and 2009 have been used as they nicely display the dynamics that occur with time, and the time slices for the year above require more than 6 GB of RAM to process on Gephi. The time slices were visualized in Gephi before any data processing had been executed on them.

The layout for the time slices is ForceAtlas2, which is a force-directed layout used for spatialization and is, in this case, used for aesthetic and readability. The node's colours are ranked by modularity, meaning the darker the blue, the higher the modularity. The edges have the same colour as the nodes they connect together. The size of the node, along with the text size, is ranked by the degree of the node.

Larger and individually shown images of each of the time slices below can be seen in Appendix I: Time Slice Visualizations.



2007: 43 nodes and 373 edges **2008:** 4358 nodes 156 943 edges **2009:** 9221 nodes 710 661 edges

Figure 3.3: The time slices for 2007 - 2009 in Gephi.

The time slices of the earlier years were also chosen as they easily visualize the evolution that occurs in the number of nodes and edges over the time span. Later time slices contain such a high number of nodes and edges that the increase is difficult to visualize in Gephi.

3.4 Data Processing

3.4.1 Data Cleaning

There is not much cleaning needed on the network data. However, some of the time slices are not fully connected and this has to be fixed. This is to ensure that all the time slices are connected and all uniform in structure. The time slices that have been cleaned are updated in the table 3.2, and the updated content of the table is marked in bold.

Table 3.2: New Network Information

Year	Node	Edges	Max. Weighted Edge	Min. Weighted Edge	Connected
2005	1	0			
2006	35	413	18.354490980813875	0.0099183197199533	True
2007	43	373	12.20067243852107	0.0018309701909121	True
2008	***3 523***	***156 918***	220.49239536474843	0.0133954597601711	***True***
2009	***8 151***	***710 623***	387.795204861323	0.0117590501755794	***True***
2010	***9 805***	***2 128 262***	480.2393782487155	0.0059659785832522	***True***
2011	***9 981***	***6 973 082***	472.639634425336	0.0024225509895823	***True***
2012	10 000	16 487 889	396.5042531732811	0.0011151542117903	True
2013	10 000	28 053 528	434.61167931809206	0.0007518488392034862	True
2014	9 999	38 992 606	292.4090687336958	0.0006287716155965297	True
2015	***9 995***	43 416 539	351.62276884446646	0.0005161096432701868	***True***
2016	9 999	44 129 304	387.0991772862282	0.0004608075747856256	True
2017	9 999	49 415 020	284.3704879642765	0.00053410326191156	True
2018	10 000	49 803 235	411.6544884897699	0.0004701613381037132	True
2019	10 000	49 907 181	0.0099183197199533	0.0006143149233932052	True

It is the updated time slices that will be used in the experiments.

3.4.2 Filtering Edges

Before any community detection algorithm is applied to the network, filtering some of the edges is needed. This is mainly because of the sheer size and to remove the weaker edges that hold no significant information of interest. Several techniques and types of thresholds can be used for filtering edges, and those that were considered in this case are described below:

Highest iteration of connected components: It could be interesting to know at which point the time slices have the highest number of connected components and take that into consideration when filtering edges.

Largest connected component: One possible threshold to use is to find the threshold where the time slices splits from one fully connected component to several connected components. These components can then all be considered their own sub-network in the time slice, and the component containing the highest number of nodes will be stored and used.

Filtering max number of edges with low weight to keep the network connected: Seeing that the network has many edges with low weight, an idea is to find the threshold for cutting the lowest weighted edges while still keeping the graph connected as one connected component. The idea is to keep all the bridges when cutting from the lowest weighted edge and going upwards in weight. The final thresholds for each of the time slice ca be seen in 3.3.

Table 3.3: Filtering Thresholds: Network Information

Year	Filtering point	Nodes	Edges	Edges Removed
2006	0.28	35	257	156
2007	0	42	373	0
2008	0.0136	3 523	156 865	53
2009	0	8 151	710 623	0
2010	0.006	9 805	2 128 259	3
2011	0.0041	9 981	6 972 237	845
2012	0.207	10 000	9 370 164	7 117 725
2013	0.184	10 000	19 917 646	8 135 882
2014	0.37	9 999	16 947 048	22 045 558
2015	0.014	9 995	43 403 780	12 759
2016	0.111	9 999	40 755 015	3 374 289
2017	0.616	9 999	16 699 932	32 715 088
2018	0.872	10 000	6 224 035	43 579 200
2019	0.362	10 000	31 399 453	18 507 728

Number of Edges: Plotting the number of edges that appear at specific thresholds throughout the distribution can point to where the weighted edges should be filtered. The threshold is 50 intervals of the weight from the edges, and the x- and y-axis are displayed with a logarithmic scale. As figure 3.4 displays, the results follow a power-law distribution.

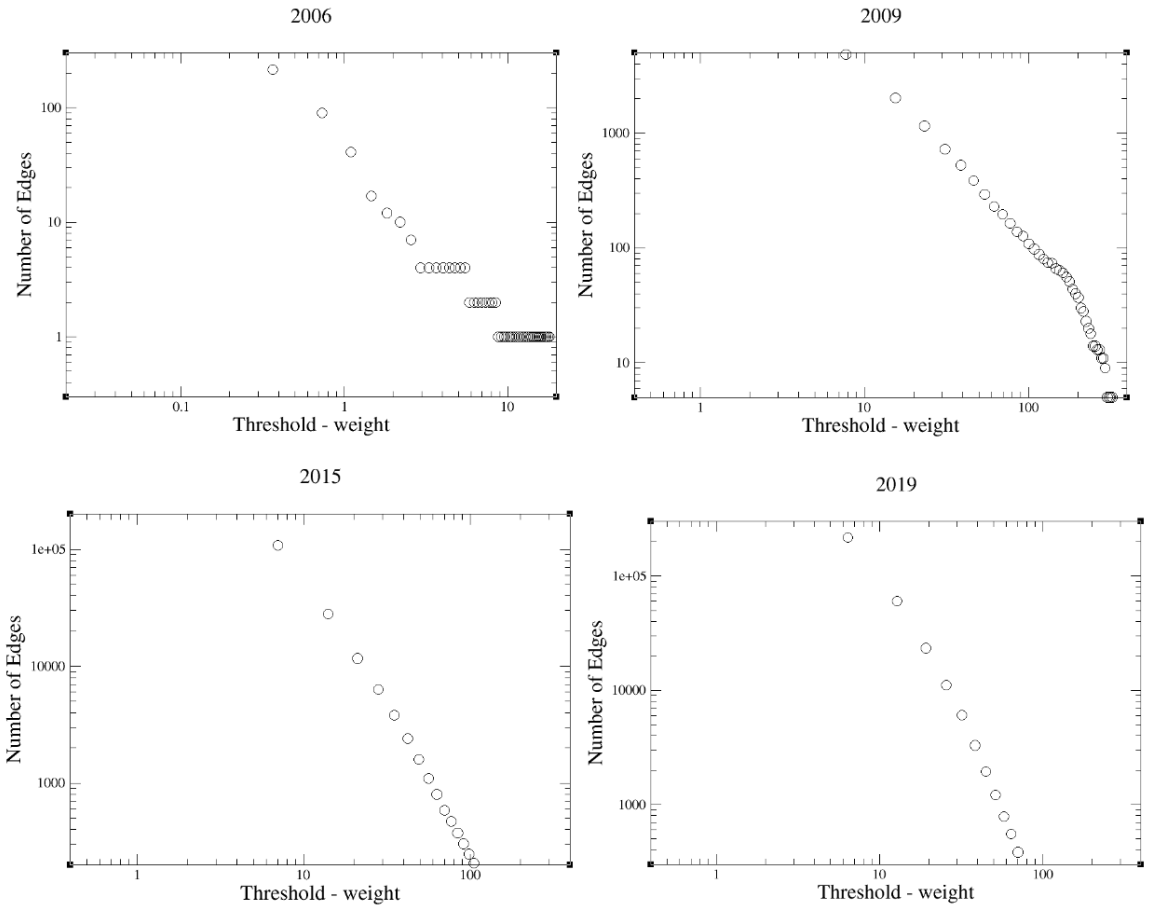


Figure 3.4: Number of edges 2006, 2009, 2015 and 2019.

The rest of the plots for the remaining years can be viewed in appendix B.

Filter by degree: Nodes with a low combined degree can be removed to reduce the network size. This will entail setting a limit of a certain number where the nodes with a less or equal number and the associated edges will be removed.

Minimum Spanning Tree: One potential algorithm that could be used for the minimum spanning tree is the prim algorithm. This algorithm is suitable for weighted undirected networks. However, this option is the least optimal for further experiments as it does not consider the weight of the edges when finding the shortest path.

3.5 Outline of Experiment

This section discusses the steps taken to prepare the time slices for each experiment and details their execution. The results of the experiments will be presented and analyzed in chapter 4.

Applying the two community detection algorithms on the time slices are considered two separate experiments, but will be described in the same section as the process for both of them are the same upon the use of the algorithm.

3.5.1 Create NetworkX Graph

The first step to constructing a networkX graph for all the time slices is to create an empty NetworkX graph structure for each of them. The nodes and edges are then added from the CSV files to the networkX graph structure.

3.5.2 Filtering the time slices

The edges in the network need to be filtered before the community detection algorithms can be applied to the time slices. The threshold we decided to use was filtering the maximum number of edges with low weight to keep the graph fully connected, see table 3.3 for each threshold.

3.5.3 Community detection algorithms

The last step of the experiment is to run the Louvain and Leiden algorithms on each of the networks. The Louvain and Leiden algorithms used are from the library CDlib, short for Community Discovery Library. Cdlib also has a number of different functions for evaluating and benchmarking community detection algorithms.

3.5.4 Source Code

The code is available at: `https://github.com/NoraMarcoux/master-thesis`

Chapter 4

The Evolution of Reddit's Communities

This chapter consists of an evaluation of the community detection algorithms that were conducted on the Reddit time slices. The results of the Louvain and Leiden algorithms were very similar and displaying the findings for both would be redundant. From this point on, only the Leiden results will be mentioned and displayed. The findings in this chapter apply to both community detection algorithms. If it is of interest to the reader, the results for the Louvain algorithm can be found in Appendix C.

For each of the evaluation functions, there is a figure to visualize the result and make it easier to analyze and evaluate the communities, along with the quality of the community detection. The functions that are used to evaluate the Louvain and Leiden algorithms are described above each of the figures.

Each of the nodes in the network represents a subreddit and the name of the subreddit is the node ID. This allowed us to look up the different subreddits and also see if any real-world events or movements are reflected in the top 10.000 subreddits and in which community they are placed in.

4.1 Leiden Visualization

After the community detection algorithm has been used on the network, each of the time slices can be visualized. Every one of the detected communities in a time slice will be assigned a colour and all the nodes in a community will share colour. Separating the communities by colour allows for an easy overview of the detected communities.

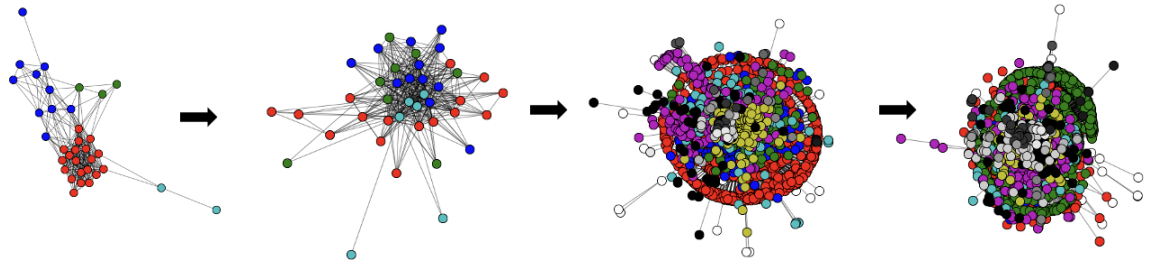


Figure 4.1: The Leiden community detection visualized for 2006 to 2009.

The time slices that are displayed in figure 4.1 are the same ones as displayed in figure 3.3, including the time slice for 2006. The time slice for 2006 starts on the left of the figure and ends with the time slice for 2009.

From figure 4.1 we can see how the community landscape evolves in the years between 2006 to 2009. There are four communities detected in the first two years, with a sudden increase in the number of communities in 2018. There is both an increase in nodes in the network and the number of communities, which could indicate that there a fewer internal edges compared to the number of external edges. Whether or not the number of communities changed in 2009 is difficult to determine based on the visualizations alone. This is one of the reasons why should use other methods for evaluating the community detection algorithms, since visualizations alone do not provide enough information.

4.2 Community Structure Evolution

Threshold values

The figure 4.2 shows the values presented in table 3.3. Seeing them visualized can help identify trends connected with the use of the thresholds.

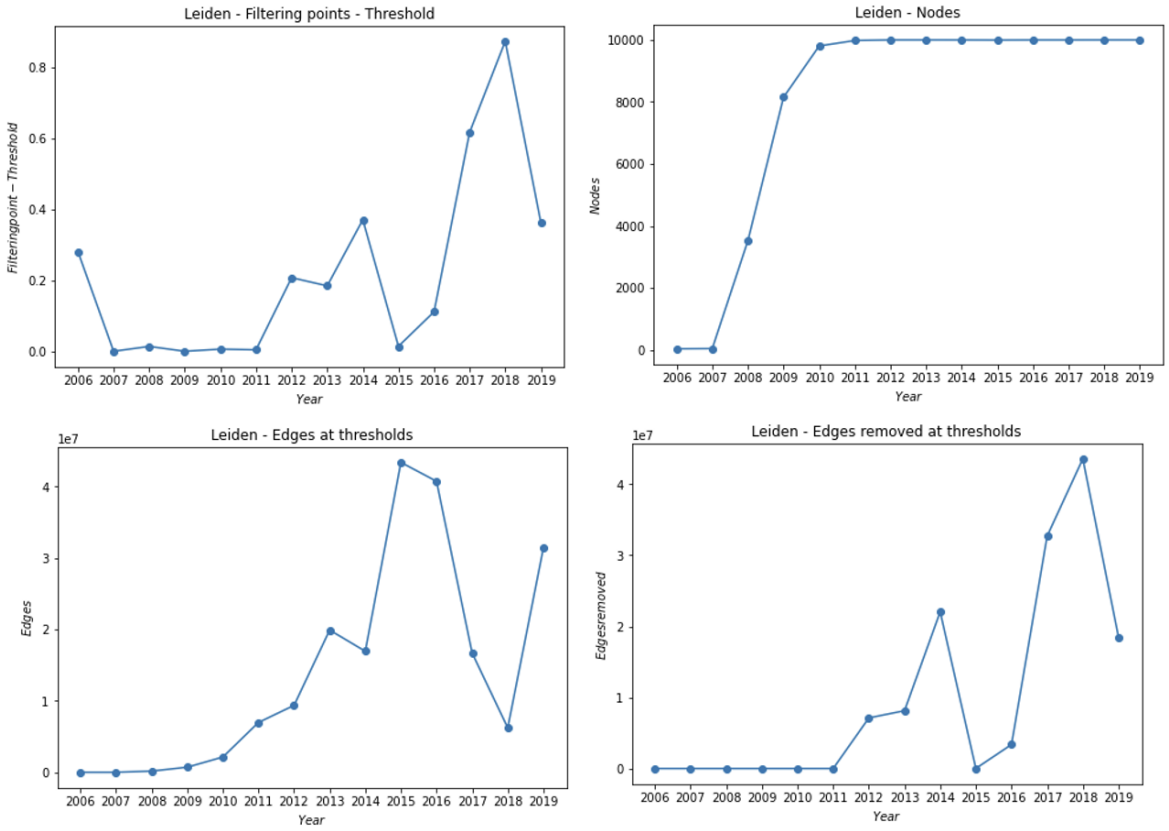


Figure 4.2: The threshold values for each year.

The first plot in the figure 4.2 displays the thresholds used on each of the time slices. These thresholds are individual for each year. The goal was to keep the time slices fully connected, and the number of edges we could remove varied for each year. The years with the lowest threshold are 2007 and 2009, because in those two years all the edges are needed to keep the graph fully connected. The years with the highest threshold are 2014, 2017, and 2018. The high thresholds in those years indicate that a higher number of edges could be removed without splitting the fully connected network.

The second and third plot in figure 4.2 is the nodes and the edges for each of the time slices after the threshold has been applied. We can see that there is a high increase in the number of nodes up until 2010. The first year that achieves 10.000 nodes is 2012.

The last plot displays the number of removed edges after the thresholds have been applied to the time slices. It is worth noting that the graph displaying the thresholds and the graph for the removed edges look very similar. This is because when increasing the threshold the number of edges removed will also naturally increase. The only exception is in 2006 when the threshold is relatively high, but only 156 edges. The time slice for 2006 originally had 413 edges, which means that 38% of the edges were

removed after the threshold was applied.

Number of Communities and the largest community for each time-slice
The number of communities for each time slice was plotted to visualize how the number of communities changed over time, and the largest community is the community that contains the highest number of nodes.

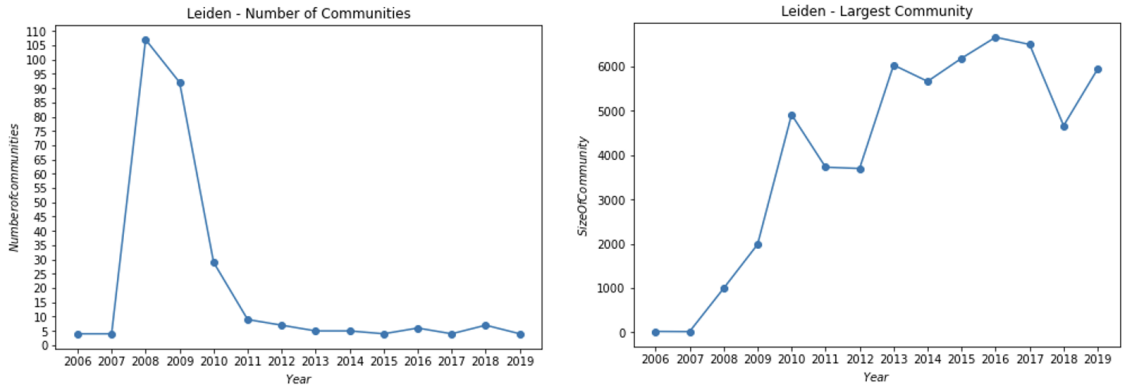


Figure 4.3: The number of communities and the largest community for each year.

Figure 4.3 displays the number of communities on the right and the size, as in the number of nodes, of the largest community for each of the years. The years with the highest number of communities are:

- 2008: with 107 detected communities.
- 2009: with 92 detected communities.
- 2010: with 29 detected communities.

The rest of the years have under ten detected communities, and the lowest number of communities detected was in 2006, 2007, 2015, 2017, and 2019 with four communities.

The size of the largest community increased in the years between 2007 to 2010. After that, there was a dip in 2011 and 2012. From 2012 to 2013, there was an exponential increase in the size of the largest community, from 3927 nodes in 2012 to 6030 nodes in 2013. In 2014 there was a slight drop in size. It increased again in 2015, and the peak of the largest community was in 2016, with a size of 6633 nodes. 2017 has a low decrease in size and a significant drop in 2018 to 5714 nodes. In 2019 there was again an increase in the number of nodes in the largest community.

Largest community divided by the total number of nodes
Dividing the number of nodes in the largest community by the total number of nodes

in the belonging time slice will give the percentage of how many of the total nodes are placed in the largest community. To get the result as a percentage, the values on the y-axis have to be multiplied by 100.

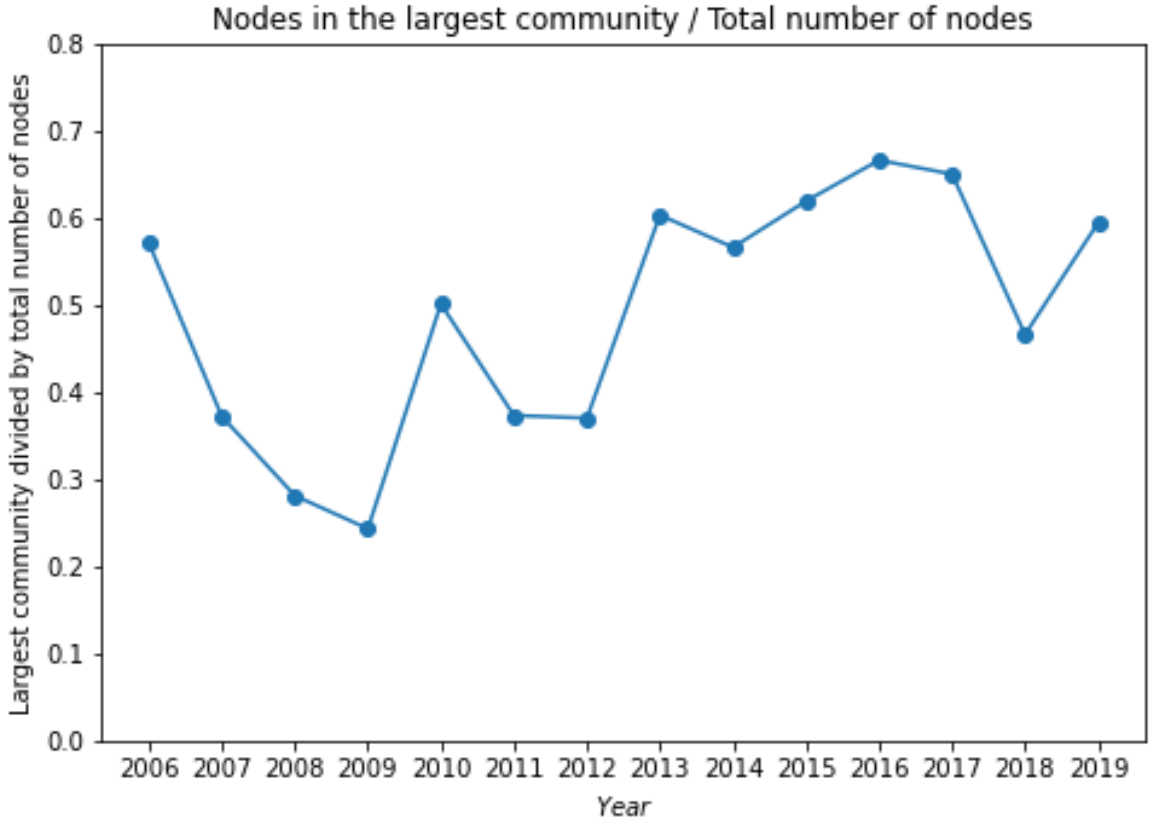


Figure 4.4: The number of nodes in the largest community divided by the total number of nodes in each time slice.

As seen in 4.4, the communities with the highest percentage of total nodes are 2015, which has 62%, 2016 with 66%, and 2017 with 65%. During those three years, the majority of the total number of nodes are located in the largest community.

The year where the largest community has the lowest percentage of total nodes was 2008 and 2009. In 2008 the largest community had 28% of the total nodes, and in 2009 the largest community had 24% of the total nodes in the time slices. These two years also have the highest number of detected communities.

Modularity

Modularity is often used to evaluate a graph's structure as it measures the connection density in a community. The library Cdlib has multiple modularity-based evaluation functions but does not offer standard modularity. We used the Newman-Girvan modularity to evaluate the modularity for the community detection done in the

experiments. The Newman-Girvan modularity is another version of modularity created to improve the hurdle of the resolution limit.

Newman-Girvan modularity is based on the difference between the fraction of edges inside a community and the expected number of edges present in a null model (CDlib, n.d.). Newman-Girvan modularity also allows for the comparison of partitions made of different modules (M. E. Newman and Girvan, 2004). The possible score for the Newman-Girvan modularity ranges from 0 to 1, where 1 indicates a strong community structure. If the modularity score is 0, then the edges within the community are no better than the edges in the null model. A modularity score of 1 means that the nodes within the communities have a high-density connection and a weak connection between nodes in the other communities.

The use of the Newman-Girvan modularity to evaluate the time slices will give a measure of how connected the nodes are within their community.

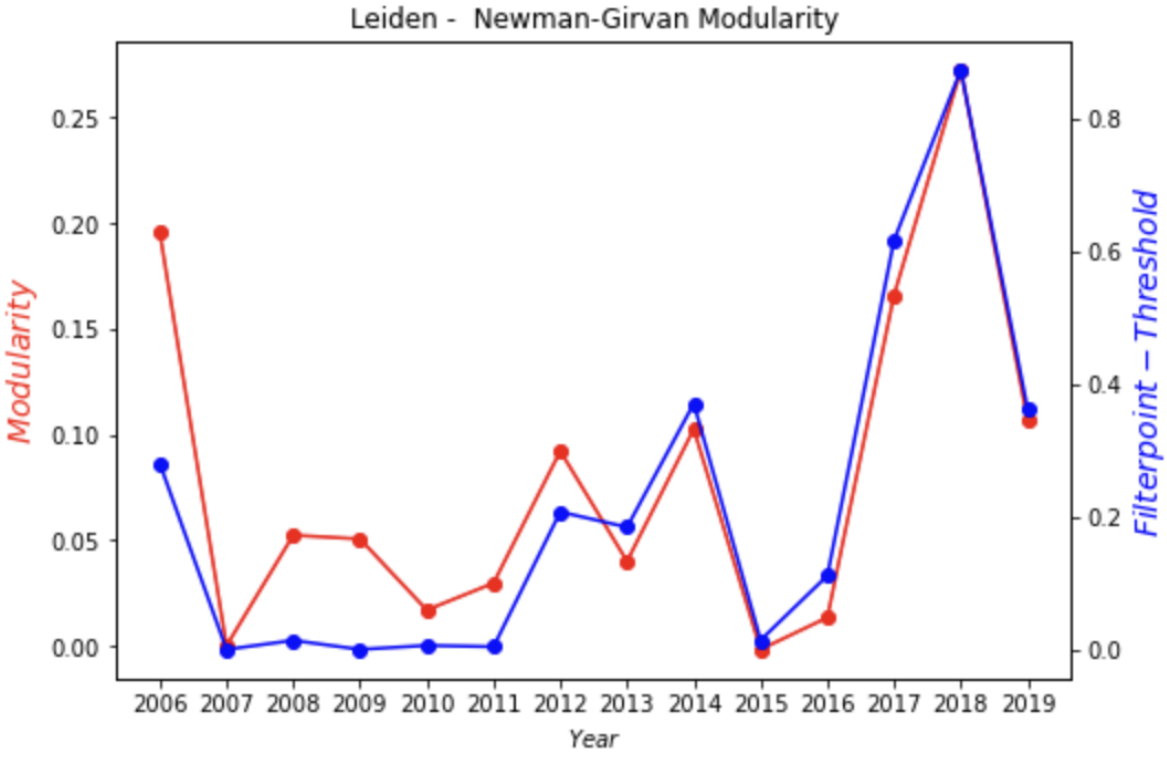


Figure 4.5: The modularity score for each year and the applied thresholds for each time slice.

Newman states in the article (M. E. Newman and Girvan, 2004) that a large network with a high community structure will usually achieve a modularity score between 0.3 and 0.7. As seen in figure 4.5, the highest modularity score achieved in this experiment is 0.28 in the time slice for 2018.

The modularity score throughout the years varies from almost 0 to the highest score in 2018. The highest scores are in 2006 with 0.20, 2017 with 0.17, and as already mentioned, in 2018 with 0.28. All the other time slices score under 0.15, which would indicate a lower measure of modularity.

In figure 4.5 the threshold values used for each time slice are also shown. The modularity scores and the threshold values behave similarly. When the applied threshold is high, the modularity is high. As well as the other way around, a low threshold produces low modularity. This could mean that the specific threshold can be used as a measure of modularity.

Conductance

Conductance is the ratio of edges that leave the community to the edges located internally in the community (Yang and Leskovec, 2015). The use of conductance can be used to measure the quality of the communities created by the community detection algorithm. The conductance score ranges from 0 to 1. The lower the score, the higher the number of edges located internally in the community is compared to the edges leaving, meaning that the community can be considered more "well-knit."

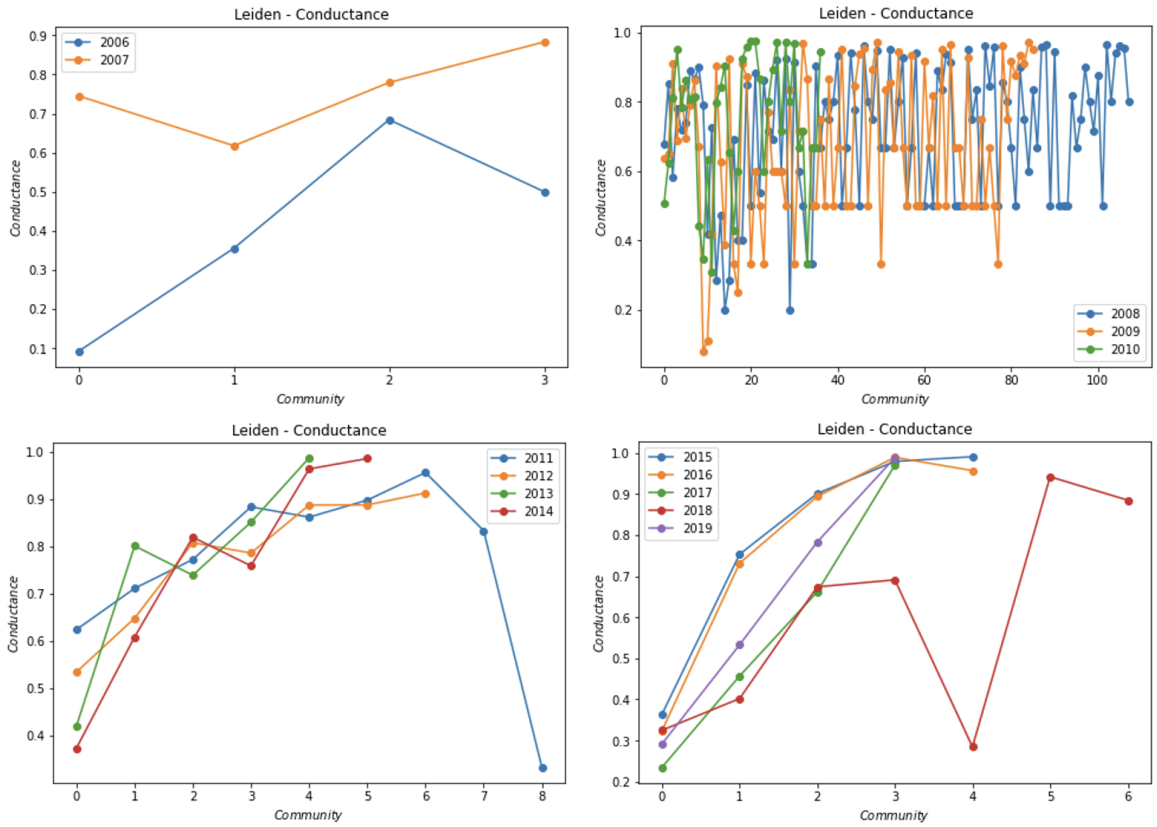


Figure 4.6: The conductance score for each of the detected communities for each of the years.

In figure 4.6 the community conductance score for each of the communities in the time slices is displayed. The conductance score varies significantly for the communities in each time slice, but the time slices that have communities with the lowest conductance score are 2006, 2009, 2011, and 2017. There was at least one community with a conductance score of 0,1, except for the community in 2017 with a conductance score of 0.23.

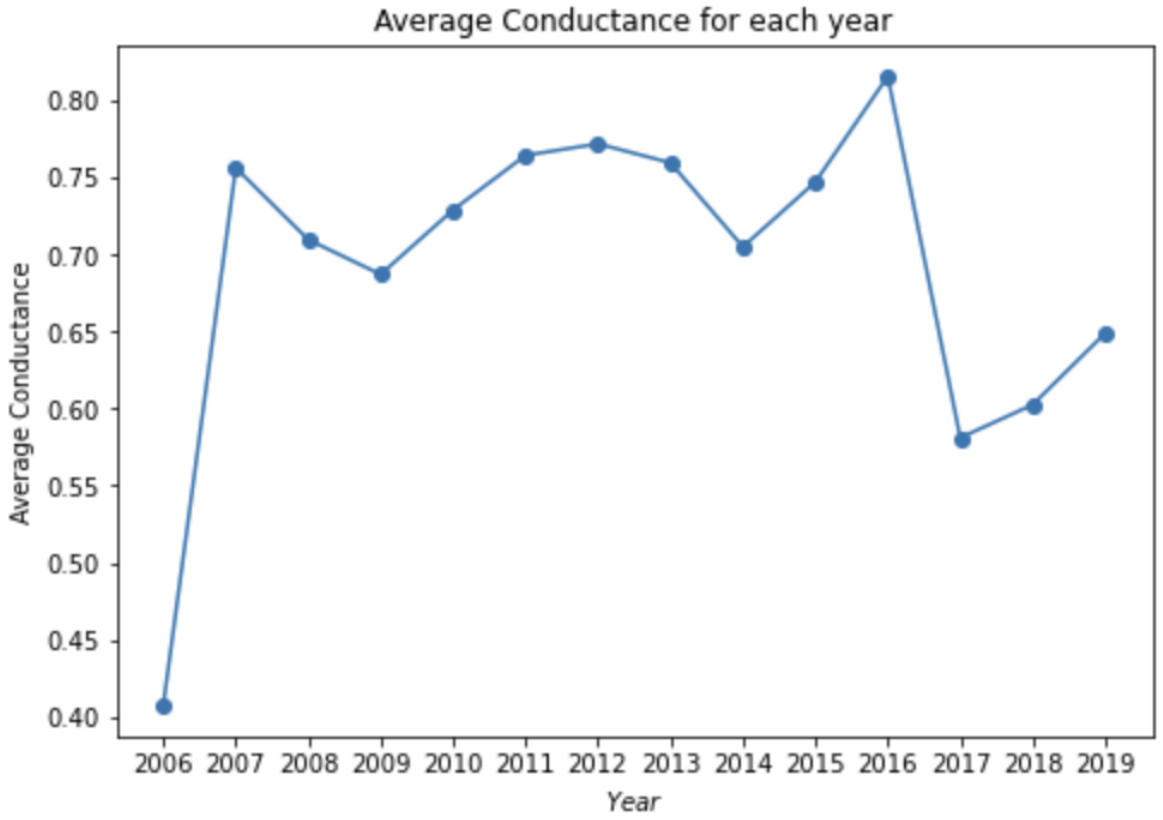


Figure 4.7: The average conductance score for each of the years.

The average conductance score for each time slice can be seen in figure 4.7. The years with the highest average conductance score are 2007 with 0,77 and 2016 with 0,82. The lowest average conductance score are 2006 with 0,42, 2017 with 0,58, and 2018 with 0,59. This means that the communities in those three years are clustered more separately than in the rest of the years.

Community Size and Edges

The size of the community is equal to the number of nodes in the community, and the number of edges represents the number of edges located internally in each community. A plot represents each community.

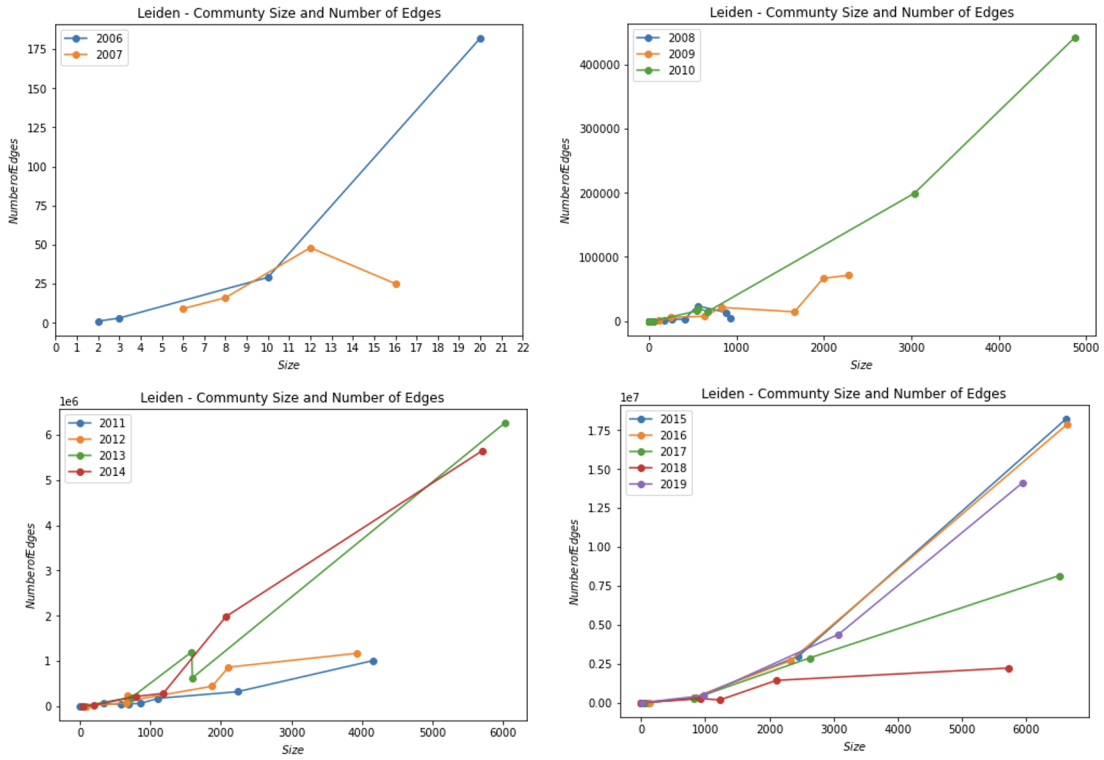


Figure 4.8: The visualized result of the community sizes and the number of edges inside each community.

As seen in figure 4.8 a common characteristic for for all of the largest communities was that there was one community with a much higher number of nodes and internal edges. The exception was for 2009, where there was one community with the highest number of nodes and another community with the highest number of internal edges. These two communities are also much closer together than in the other time slices.

4.3 Community Behaviour

Center of the largest and second-largest community

The center of a community is the node with eccentricity equal to the radius of the community. This means that the center node is the node closest to all the other nodes in the community. As each of the nodes in the time-slices represents a subreddit, the name of the center node could potentially reveal something about the topic of interest in the communities. The center node is also often used to name the whole community.

Table 4.1: The Center of the Largest and Second-largest Community for Each Year

	2006	**2007**	**2008**
Largest community	nl	Entertainment	Reddit.com
Second-largest community	Request	Request	Health
	2009	**2010**	**2011**
Largest community	Travel	Newsreddit	Newsreddit
Second-largest community	Newsreddit	Education	Foodforthought
	2012	**2013**	**2014**
Largest community	Gardening	ABDL	Lebanon
Second-largest community	LinkinPark	Education	FindaReddit
	2015	**2016**	**2017**
Largest community	BabyBumps	UMD	ReBBl
Second-largest community	AMA	Aww	FindaReddit
	2018	**2019**	
Largest community	PS4Planetside2	KinFoundation	
Second-largest community	wtfstockphotos	redditrequest	

Most of the subreddits listed in table 4.1 provide little context and require additional information. Each of the subreddits can be accessed by typing in the link `Reddit.com/r/nameOfSubreddit/` and is also shortly summarised in appendix D.

In table 4.1, each of the centers for the largest and second-largest communities for each year is listed. There seem to be some trends that repeat themselves over the years, `r/NewsReddit` appears three times on the list. The two other subreddits that appear more then once is `r/Request` and `r/FindaReddit`. However, based on the subreddits alone, it is impossible to say that the largest communities stay the same based on the centers. The topic range is wide and includes everything from news and gardening to different games.

Chapter 5

Discussion

The findings that are being evaluated and discussed in this chapter are divided into three different categories. The first category is community structure, the findings in this category are based on the graphs presented and explained in chapter 4.

The second category is community behavior and was based on observations from the community detection algorithms. The community detection algorithms create lists with the sorted communities and all of the nodes belonging to the detected communities. These lists allow for a more thorough look at dynamics that take place in the community behavior.

The last category is Real-world events and the findings in this category are related to real-world behavior and events that have gained attention in recent years. The lists from the community detection will have the names of the nodes, which are subreddit names, and can be used for exploration to form more topic or interest-oriented observations related to real-world events.

5.1 Community Structure

Based on the evaluation done in section 4.2, Reddit does show evidence of having communities and should not be viewed as one homogeneous entity. However, the communities do show medium to low quality, meaning they can probably not be classified as strong communities. This was likely due to the number of edges that were kept, if more edges had been removed, higher quality communities could possibly have been identified. This is also supported by the fact that 2018 was the year with the highest applied threshold and the year that had the highest modularity and second-lowest conductance score.

However, our purpose was to explore the dynamics in the Reddit network and not solely to identify strong communities. Our findings imply that even with a high number of different weighted edges there was a presence of community structure on Reddit. Also, a high number of edges was needed for the network to remain fully connected, which was the basis for the thresholds.

The decided threshold, which was to filter away the maximum number of the lowest weighted edges to keep the network fully connected, can be used as a measure of modularity. As seen in figure 4.5 the threshold and the modularity behave almost correlatively.

Another possible finding was that the Louvain algorithm and Leiden algorithm perform almost identically on large complex networks. The results from the two algorithms produced a very similar result. The Leiden algorithm is considered to be an improved version of the Louvain algorithm but is known to be suited for smaller networks. In this case, the communities detected by the Leiden algorithm did receive better modularity and conductance score, but not enough to be considered a significant improvement.

5.2 Community Behaviour

The number of communities drastically increases in 2008 and immediately start decreasing until it stabilizes in 2011. The drastic increase takes place the same year Reddit allows for user-generated subreddits. The largest community in 2008 had a size of 929 nodes, which means that the remaining 2549 nodes are spread between 108 communities.

Most of these small communities disappear over time, but many of the nodes in these communities are absorbed into bigger communities. The larger the community is the broader the range of topics, which makes it hard to identify if it is the same community that stays the largest throughout the years.

The largest community does, based on the exploration of the communities, seem to shift from year to year. A high number of nodes disappear and some reappear again in other time slices. An example was the node for the subreddit `r/Elephants` that first appears in 2010 and then reappears in 2012 and 2014.

5.3 Real-world Events

As seen in figure 4.3 2008 was the year Reddit saw a significant increase in the number of communities. The growth correlates with the year Reddit allowed its users to create their own user-generated subreddit (Essert, 2014). This also correlates with the first significant increase in the number of nodes in the largest community, which can be seen in 4.2. The growth continues until 2010 when there was enough activity across subreddits to create a connected network with the top 10.000 subreddits.

Controversial Subreddits

The subreddit `r/WallStreetBets` was created on January 31, 2012, ('wallstreetbets', n.d.) and the same year it appeared in the largest community, meaning that the subreddit gained popularity at a fast rate. In the period 2012 to 2019, `r/WallStreetBets` was either placed in the largest or second-largest subreddit. Another Wall Street related subreddit that appeared one year before `r/WallStreetBets` is the subreddit `r/OccupyWallStreet` ('Occupy Wall Street: We Are The 99% Fighting Back', n.d.), which was created on July 15, 2011, and it also remained in the top 10.000 subreddits up until 2019. Occupy wall street was a protest movement that officially started in September 2011 (Editors, 2021), when the protests on Wall Street occurred. This could indicate that the Occupy Wall Street movement got some of its traction on Reddit since it had a presence on the platform before the official protest started.

Another subreddit that has gained a lot of notice in recent years is `r/Antiwork`. This was also reflected in the fact that `r/Antiwork` appears in the third-largest community in 2018 and the second-largest community in 2019. The subreddit was started on August 14, 2013 ('Antiwork: Unemployment for all, not just the rich!', n.d.), but started gaining mainstream popularity as a movement in 2020 - 2021 (Whang, 2022).

A subreddit that gained a lot of attention in recent years was the subreddit `r/The_Donald`. The subreddit was created for supporters of Donald Trump during his 2016 election campaign. The subreddit was known for its racist, sexist, homophobic, and right-wing content (Zakrzewski, 2020). In 2016 and up until 2019, the subreddit was placed in the second-largest community. There were also several other subreddits, such as `r/BannedFromThe_Donald` and `r/AskThe_Donald` in the same community. This indicates that the Pro Trump-movement had a rather active following on Reddit during that period. `r/The_Donald` was removed from Reddit in 2020 after breaking Reddit's content policy several times (Lima, 2020). At the time it was banned, the subreddit had more than 790,000 members.

An ideology that has formed a following on Reddit is that of the online subculture

of incels. Incels are most typically men who are self-proclaimed involuntary celibates (Ging, 2019). The most popular subreddit for incels was r/incels, which was placed in the second-largest community in 2016 and 2017. The subreddit received criticism for its content, which encouraged violence against women, and there was even a petition created on change.com asking for the subreddit to be banned (Van Brunt and Taylor, 2020). The subreddit r/Braincels gained popularity among the incels on Reddit, especially after r/incels was removed in 2017. r/Braincels was placed in the largest community in 2017 and second-largest community in 2018 and 2019. Elliot Rodger was a self-proclaimed incel and in May 2014, he went on a killing spree, murdering 10 people (Binder, 2019). The members of r/Braincels showed support to Elliot Roger, even praising him for his actions. r/Braincels was eventually also banned in 2019, for the same reasons as r/incels. At the time of the banning, the subreddit had almost 20,000 members.

The subreddits r/DarkNetMarkets and r/DarkNetMarketsNO were, as of March 2018 banned on Reddit (Brackett, 2018), but both of them were in the top 10.000 subreddits for several years before being banned. The subreddits acted as discussion forums for illegal markets on the dark web. r/DarkNetMarkets was placed in the largest community in 2013, and second-largest community up until 2017. In 2018, the subreddits was placed in a smaller community, which makes sense, since it was removed in March 2018. The members were not discreet about their actions and Reddit was even subpoenaed by the Department of Homeland Security in 2015 for information about some of the members.

The subreddit r/DarkNetMarketsNO was the Norwegian version of r/DarkNetMarkets. r/DarkNetMarketsNO appears in the largest community in both 2016 and 2017, which is an interesting finding, seeing that the members of the subreddit communicated in Norwegian. This could indicate that there were a number of very active members from Norway on Reddit.

The Boston Bombing was a domestic terror attack in the United States that took place on April 15, 2013, during the annual Boston Marathon (Fielding et al., 2014). In the aftermath of the attack, the subreddit r/BostonBombing was created to keep up with any updates related to the attack. The subreddit was placed in the largest detected community for 2013, meaning a high number of users participated in the subreddit. Reddit actually played a significant role in the aftermath of the Boston Bombing when a number of users on the site decided to try and identify the people responsible for the attack. The subreddit r/FindBostonBombers was created for this purpose and wrongly identified several people. However, r/FindBostonBombers was never in the top 10.000 subreddits.

The most famous case was of the Student Sunil Tripathi that had been missing for over a month when his picture was spread on Reddit as one of the perpetrators. A Reddit user had seen his missing person information online and thought it resembled one of the Boston bombers. Several news outlets picked up on this, including a Twitter account connected to the hacktivist group Anonymous. In a matter of hours, Sunil Tripathi's family started receiving threats, even though it was confirmed that he was not a suspect by the authorities. Reddit later made a statement where they apologized for the platform's role in spreading the false accusation (Shontell, 2013). Sunil Tripathi was later found dead, he had committed suicide by drowning.

Based on the mentioned subreddits, there does seem to be a reflection of real-world events and movements on the platform. As we can see, all the subreddits mentioned in this section does to some extent, some more then others, have real-world implications.

Chapter 6

Conclusions

In this thesis we have explored the dynamics that have occurred on Reddit spanning 2006 to 2019, including the behavioral transitions that took place in the time slices. We have findings that could potentially support the fact that Reddit should not be considered homogeneous, but rather consisting of evolving communities. We were able to use the Louvain and Leiden algorithms to explore the evolution of communities in the Reddit landscape, as well as connecting subreddits to real-world events.

6.1 Challenges and limitations

The two main challenges we encountered will working on this project was:

Graph Size

The time slices from 2005 to 2009 were manageable to work with, but even when using eX^3 there were challenges with the datasets from 2010 to 2019, because of the size. It was the surge of edges that caused the increased size in the time slices after 2009. The Jupyter notebook took up to 8 hours to run each of the community detection algorithms.

Visualization of the Graph

The graphs with the community detection would have been interesting to visualize with Gephi, but due to the size of the time slices from 2010 to 2019 and the capacity restriction of Gephi, this proved to be difficult without any dedicated hardware.

6.2 Future Work

The dataset used in this thesis has never been used or explored to the extent as it has in this thesis. This means that there are potentially a lot of interesting discoveries left in the dataset. Some areas of further work that could be of interest are:

Other community detection algorithms
The Louvain and Leiden algorithms displayed very similar results on the Reddit data. There are other community detection algorithms, and it could be interesting to see if they will yield the same results on such a large complex network.

Dynamic community detection
In this thesis, each of the time slices has been processed as a stand-alone graph and then put together to evaluate the effects of time. It is possible to create dynamic behavior and use algorithms that take previously found communities into consideration when identifying new ones. This would allow for a more comprehensive investigation into the behavior inside the detected communities.

Increase the number of time slices
Increasing the number of time slices to monthly slices for each year, instead of yearly, would allow for more in-depth discoveries on the community evolution and real-world events in the Reddit network.

Include years after 2019
Reddit has seen an increase in activity and new users after the Covid-19 pandemic. It could be interesting to see if and how this increase has affected Reddit users' community behavior.

Predictions
The time slices created from the RDSP could potentially be used to predict the future behavior of the Reddit network. Depending on the type of prediction, it could potentially be done by using a Long Short-Term Memory Network, which is a type of Recurrent Neural Network. However, it might require some fine-tuning of the time slices and probably some additional information for the learning algorithm's training data.

Bibliography

Antiwork: Unemployment for all, not just the rich! (n.d.). https://www.reddit.com/r/antiwork/

Auxier, B. & Anderson, M. (2021). *Social media use in 2021*. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M. & Blackburn, J. (2020). The pushshift reddit dataset. *Proceedings of the international AAAI conference on web and social media*, *14*, 830–839.

Binder, M. (2019). Reddit changes its harassment policy and bans major incel community. https://mashable.com/article/reddit-harassment-braincels

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Brackett, E. (2018). Reddit bans communities dedicated to illegal goods. https://www.digitaltrends.com/social-media/reddit-bans-illegal-communities/

Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z. & Wagner, D. (2007). On modularity clustering. *IEEE transactions on knowledge and data engineering*, *20*(2), 172–188.

Brian, D. (2021). Reddit user and growth stats. https://backlinko.com/reddit-users

Buntain, C. & Golbeck, J. (2014). Identifying social roles in reddit using network structure. *Proceedings of the 23rd international conference on world wide web*, 615–620.

CDlib. (n.d.). Cdlib.evaluation.z_modularity. https://cdlib.readthedocs.io/en/latest/reference/eval/cdlib.evaluation.z_modularity.html#cdlib.evaluation.z_modularity

Clayton, J. (2022). How kremlin accounts manipulate twitter - bbc news. https://www.bbc.com/news/technology-60790821

Datta, S. & Adar, E. (2019). Extracting inter-community conflicts in reddit. *Proceedings of the international AAAI conference on Web and Social Media*, *13*, 146–157.

DiBrita, N., Eledlebi, K., Hildmann, H., Culley, L. & Isakovic, A. F. (2021). Temporal graphs and temporal network characteristics for bio-inspired networks during optimization.

Editors, H. (2021). Occupy wall street begins - history [(Accessed on 10/05/2022)]. https://www.history.com/this-day-in-history/occupy-wall-street-begins-zuccotti-park

Essert, M. (2014). Here's the cool graph reddit fans should show haters who still claim it's not a legit news site. https://www.mic.com/articles/78735/here-s-the-cool-graph-reddit-fans-should-show-haters-who-still-claim-it-s-not-a-legit-news-site

eX3. (n.d.). *Home*. https://www.ex3.simula.no/

Fielding, R., Bashista, R., Ahern, S. A., Duggan, C., Giacobbe, C., Lawn, M., MacMillan, P., Albanese, D., Grieb, J., Crawford, R. et al. (2014). After action report for the response to the 2013 boston marathon bombings.

for Reddit, M. (2022). */r/announcements metrics*. https://frontpagemetrics.com/r/announcements

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, *486*(3-5), 75–174.

Fortunato, S. & Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the national academy of sciences*, *104*(1), 36–41.

Fortunato, S. & Hric, D. (2016). Community detection in networks: A user guide. *Physics reports*, *659*, 1–44.

Ging, D. (2019). Alphas, betas, and incels: Theorizing the masculinities of the manosphere. *Men and Masculinities*, *22*(4), 638–657.

Good, B. H., De Montjoye, Y.-A. & Clauset, A. (2010). Performance of modularity maximization in practical contexts. *Physical Review E*, *81*(4), 046106.

Griesemer, S. & Schlessinger. (2018). Detection and analysis of subreddit communities. *Scientific reports*.

Holme, P. & Saramäki, J. (2012). Temporal networks. *Physics reports*, *519*(3), 97–125.

Huber, A. (2021). *Observing reddit's interaction network-a stream-based approach for large scale network analysis on reddit* (Master's thesis).

Kelly, J. (2021). Reddit's year in review reveals antiwork and wallstreetbets as top 10 discussions. https://www.forbes.com/sites/jackkelly/2021/12/08/reddits-year-in-review-reveals-antiwork-and-wallstreetbets-as-top-10-discussions/

Khandelwal, U. & Xu, S. (n.d.). *How well does language-based community detection work for reddit?*

Khokhar, D. (2015). *Gephi cookbook*. Packt Publishing Ltd.

Li, A., Cornelius, S. P., Liu, Y.-Y., Wang, L. & Barabási, A.-L. (2017). The fundamental advantages of temporal networks. *Science*, *358*(6366), 1042–1046.

Lima, C. (2020). Twitch, reddit crack down on trump-linked content as industry faces reckoning. https://www.politico.com/news/2020/06/29/reddit-bans-pro-trump-forum-in-crackdown-on-hate-speech-344698

Medvedev, A. N., Lambiotte, R. & Delvenne, J.-C. (2017). The anatomy of reddit: An overview of academic research. *Dynamics on and of Complex Networks*, 183–204.

Mensah, H., Xiao, L. & Soundarajan, S. (2020). Characterizing the evolution of communities on reddit. *International Conference on Social Media and Society*, 58–64.

Newman, M. (2018). *Networks*. Oxford university press.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, *103*(23), 8577–8582.

Newman, M. E. & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, *69*(2), 026113.

Occupy wall street: We are the 99% fighting back. (n.d.). https://www.reddit.com/r/occupywallstreet/

Olson, R. S. & Neal, Z. P. (2015). Navigating the massive world of reddit: Using backbone networks to map user interests in social media. *PeerJ Computer Science*, *1*, e4.

RedditInc. (2021). *What is an admin?* https://www.reddithelp.com/hc/en-us/articles/204546259-What-is-an-admin-

RedditInc. (2022a). *About: Dive into anything*. http://https://www.redditinc.com/

RedditInc. (2022b). *About: Dive into anything*. http://https://www.redditinc.com/

RedditInc. (2022c). *What's a moderator?* https://www.reddithelp.com/hc/en-us/articles/204533859-What-s-a-moderator-

Serrano, M. Á., Boguná, M. & Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences*, *106*(16), 6483–6488.

Shontell, A. (2013). Reddit wrongly accuses sunil tripathi of boston bombing. https://www.businessinsider.com/reddit-falsely-accuses-sunil-tripathi-of-boston-bombing-2013-7?r=US&IR=T

Singer, P., Flöck, F., Meinhart, C., Zeitfogel, E. & Strohmaier, M. (2014). Evolution of reddit: From the front page of the internet to a self-referential community? *Proceedings of the 23rd international conference on world wide web*, 517–522.

Traag, V. A., Waltman, L. & Van Eck, N. J. (2019). From louvain to leiden: Guaranteeing well-connected communities. *Scientific reports*, *9*(1), 1–12.

Van Brunt, B. & Taylor, C. (2020). *Understanding and treating incels: Case studies, guidance, and treatment of violence risk in the involuntary celibate community*. Routledge.

Wallstreetbets. (n.d.). https://www.reddit.com/r/wallstreetbets/

Wasserman, S., Faust, K. et al. (1994). Social network analysis: Methods and applications.

Whang, O. (2022). Hating your job is cool. but is it a labor movement?

Yang, J. & Leskovec, J. (2015). Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, *42*(1), 181–213.

Zakrzewski, C. (2020). The technology 202: New study reveals extent of hate speech on reddit in right-leaning forums. https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2020/06/15/the-technology-202-new-study-reveals-extent-of-hate-speech-on-reddit-in-right-leaning-forums/5ee6ab4c602ff12947e8c19a

Appendix A

Appendix I: Time Slice Visualizations

The Figures below A.1, A.2 and A.3 are the visualized time slices as shown in 3.3, but shown individually such that the node names are readable.

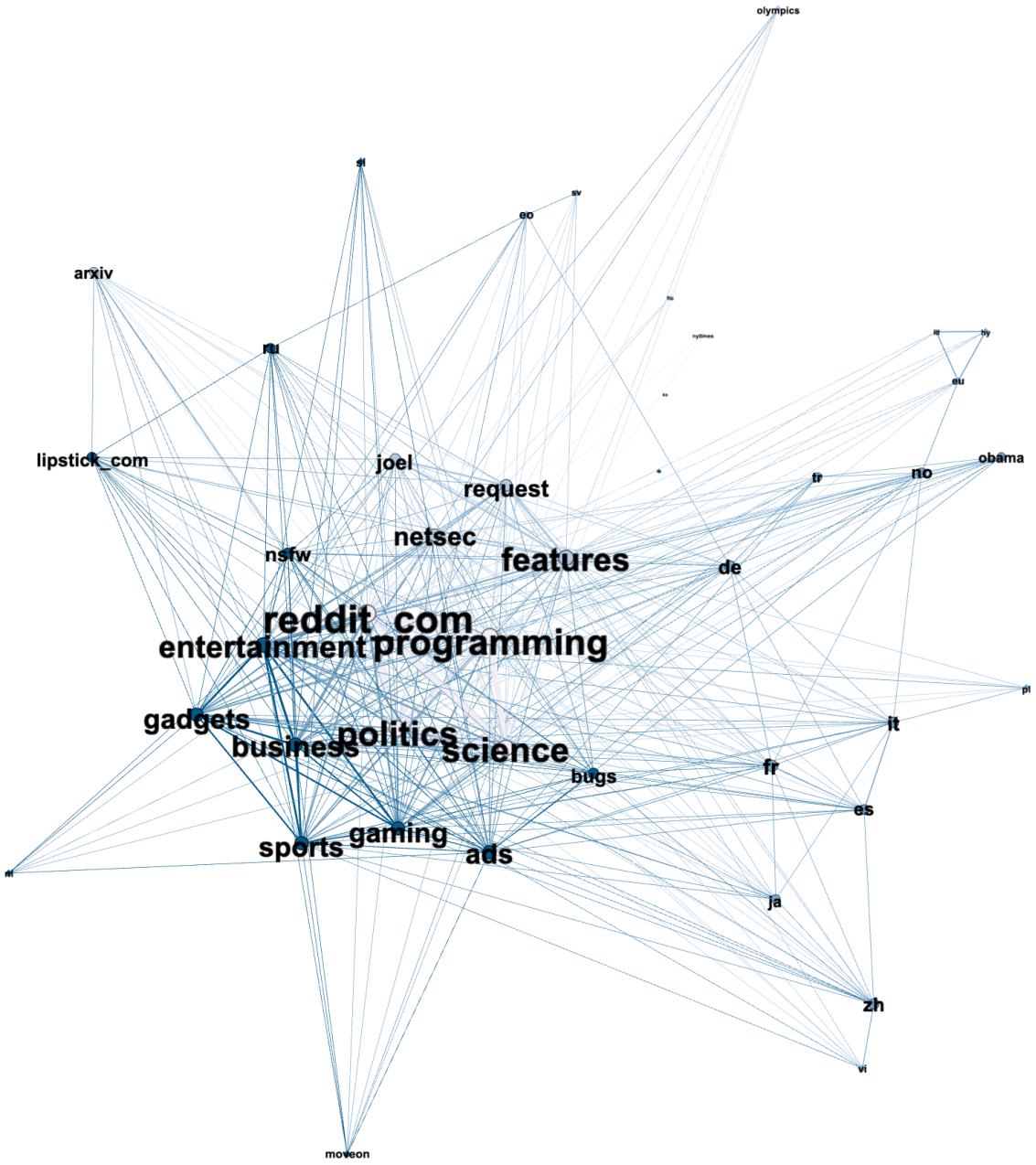


Figure A.1: The time slice for 2007 in Gephi.
The time slice contains 43 nodes and 373 edges.

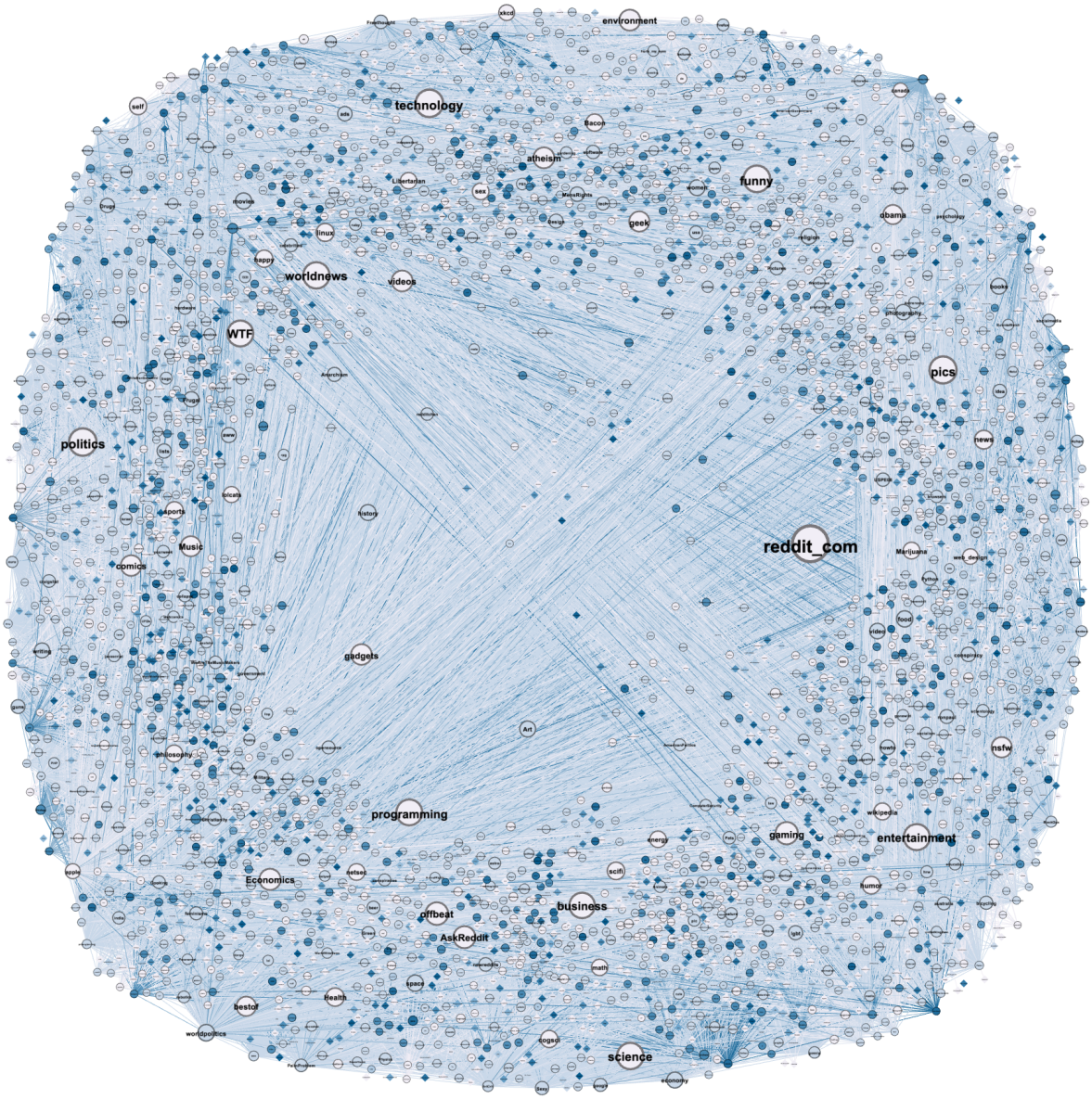


Figure A.2: The time slice for 2008 in Gephi.
The time slice contains 4358 nodes and 156943 edges.

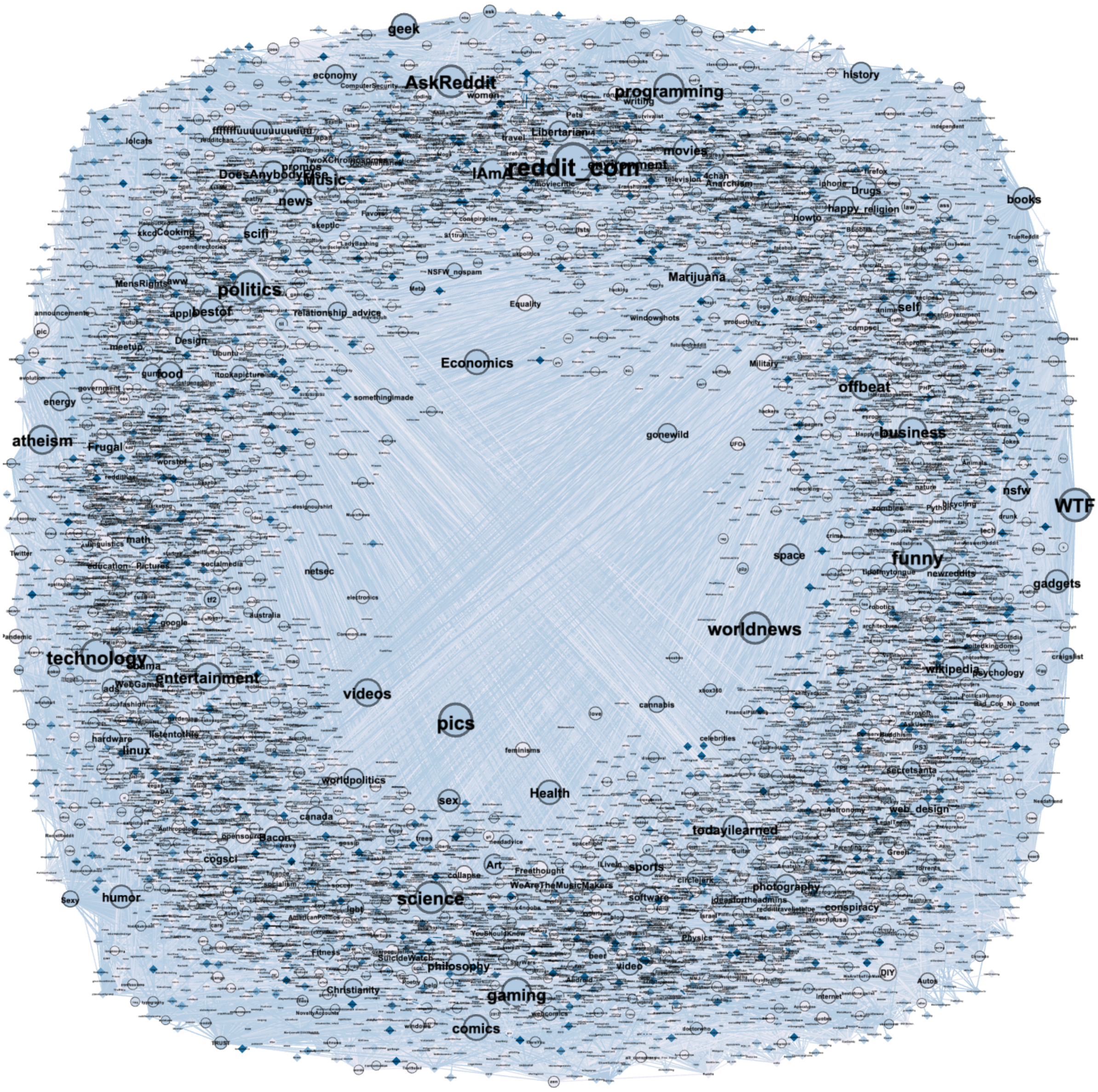


Figure A.3: The time slice for 2009 in Gephi.

The time slice contains 9221 nodes and 710661 edges.

Appendix B

Appendix II: Systematic Results for the Number of Edges

This appendix contains the plots that were not shown in Chapter 3, as seen in3.4, displaying the number of edges at different thresholds with the x- and y-axis using a logarithmic scale. The result display that the edges follow a power-law distribution, meaning that the edge distribution can not be used as a threshold.

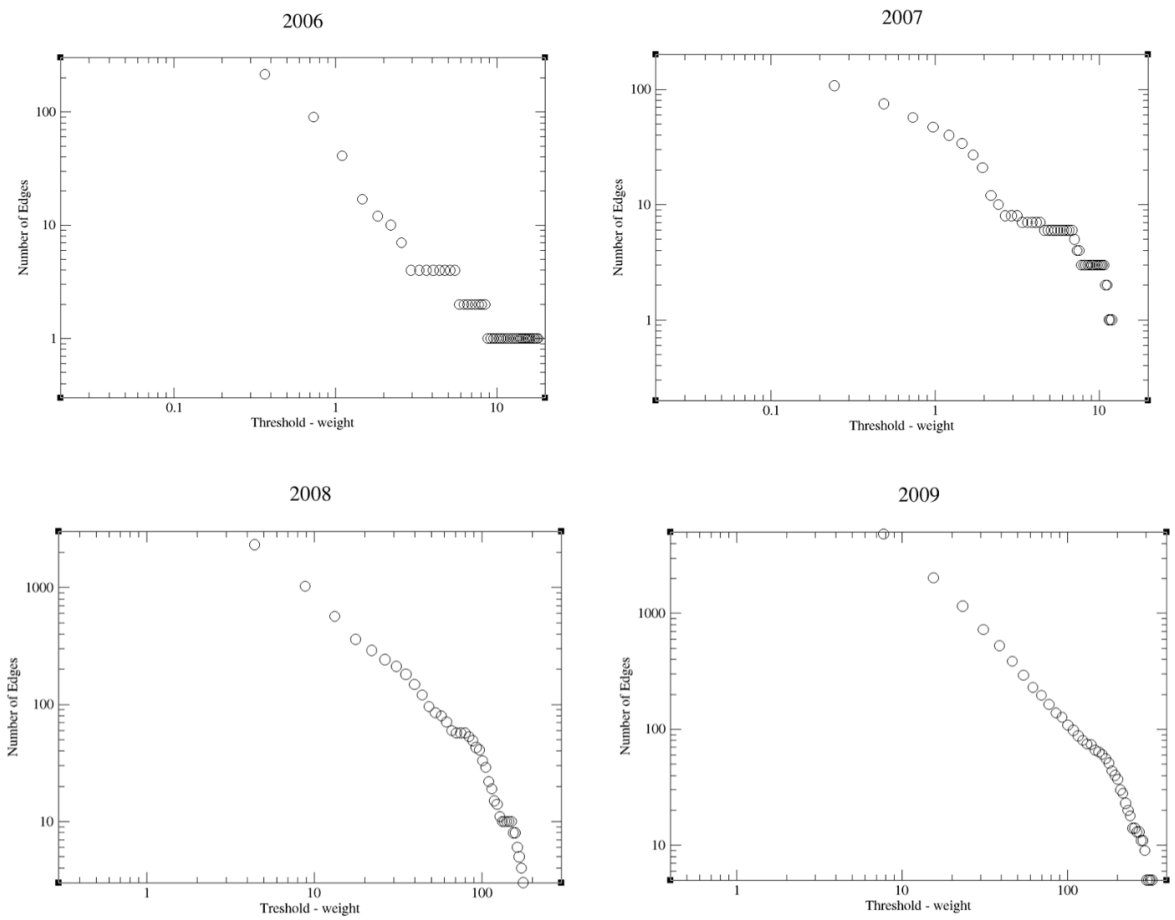


Figure B.1: Number of edges 2006, 2007, 2008 and 2009

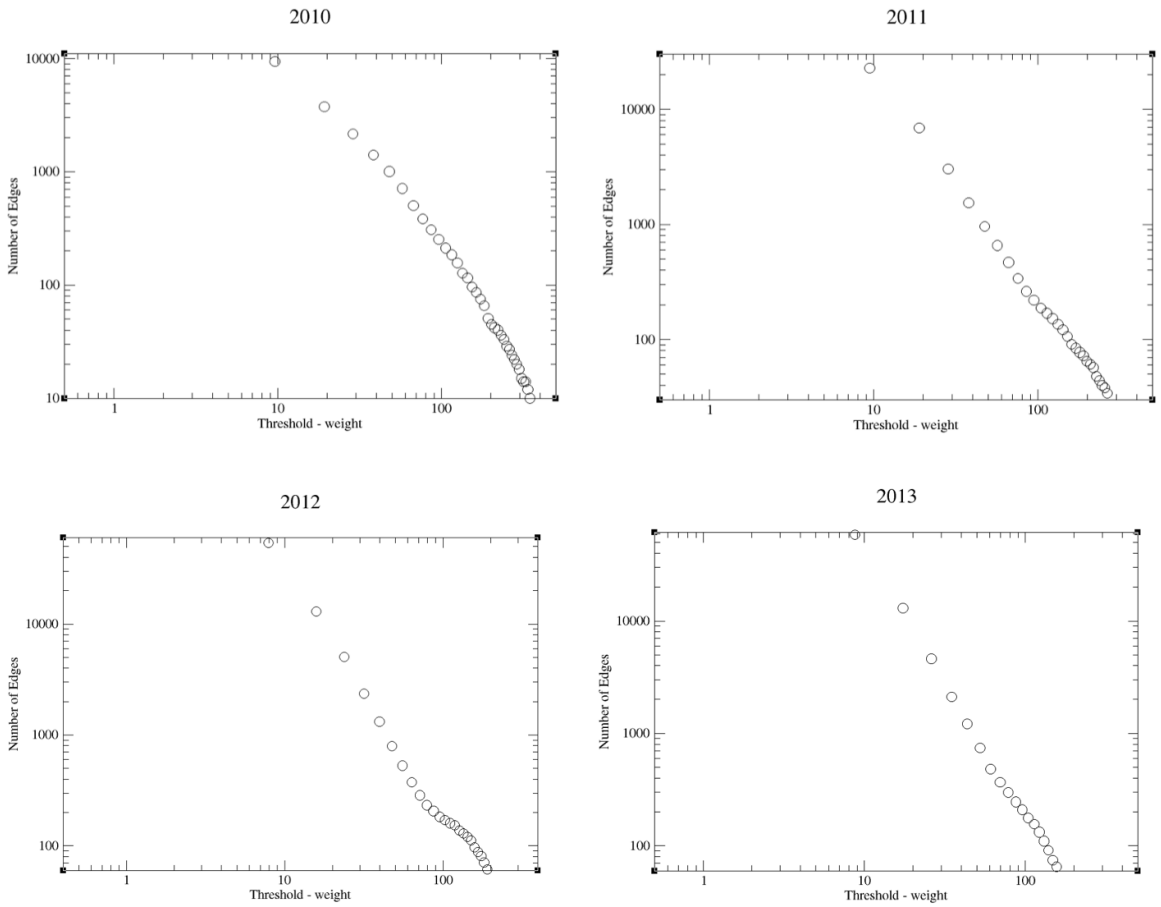


Figure B.2: Number of edges 2010, 2011, 2012, 2013

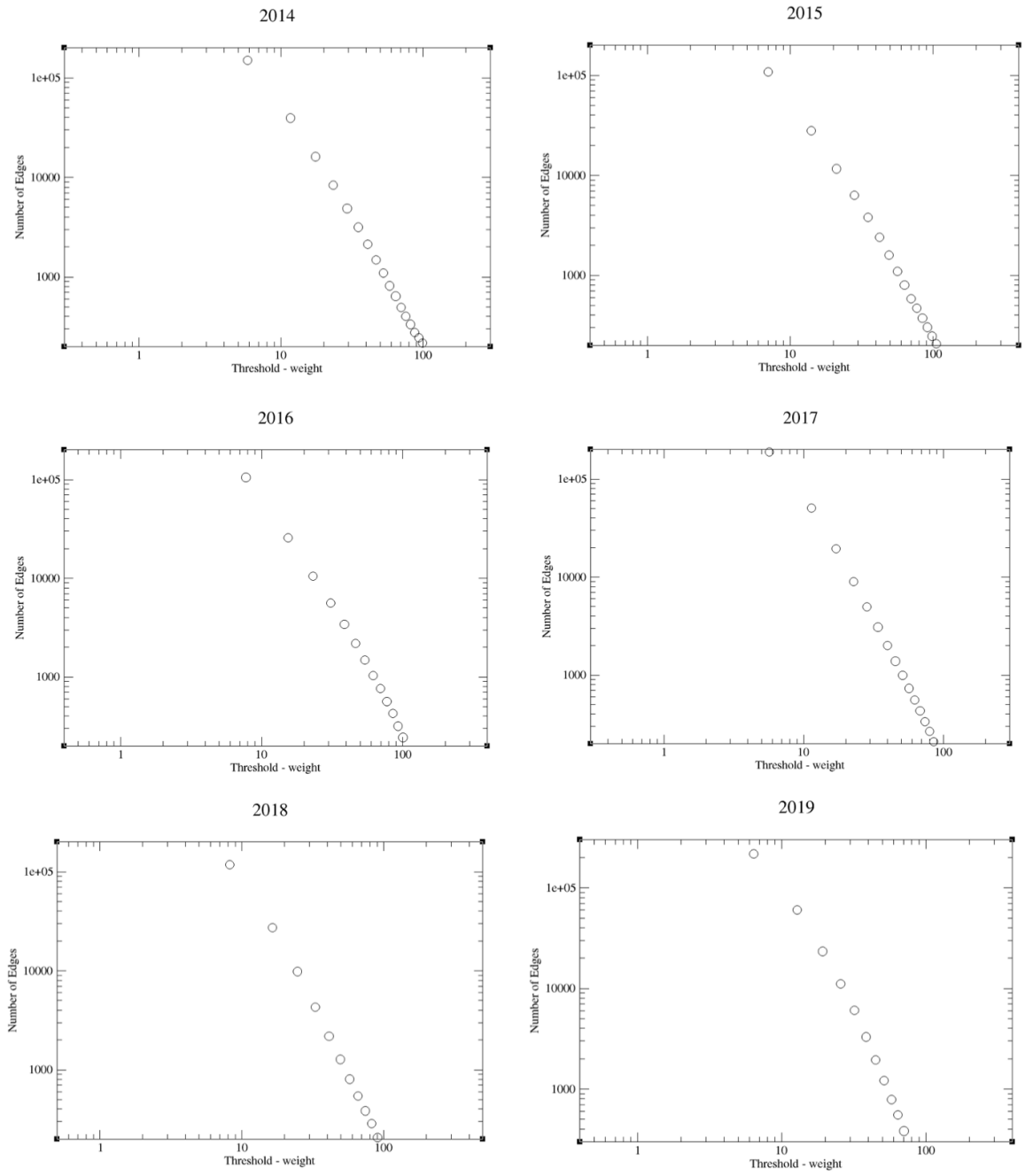


Figure B.3: Number of edges 2014, 2015, 2016, 2017, 2018 and 2019

Appendix C

Appendix III: Results with the Louvain algorithm

Louvain Visualization

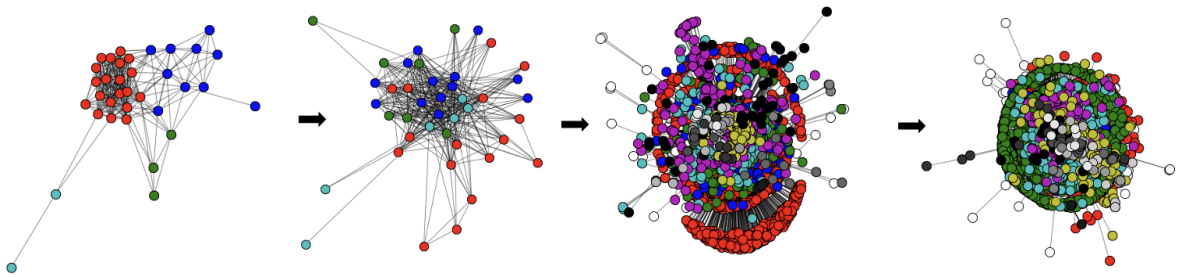


Figure C.1: The Louvain community detection visualized for 2006 to 2009.

Experiment: Louvain Community Detection

This appendix contains the graphs that were displayed in chapter 4, but for the Louvain algorithm. The graphs were excluded from the chapter and added as an appendix due to the fact that there was very little difference between the results from the two community detection algorithms.

Number of Communities and the largest community for each time-slice

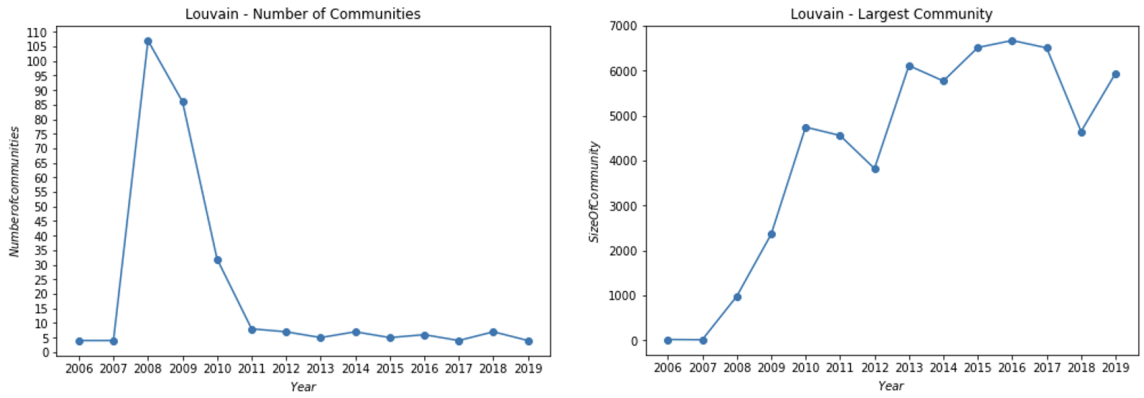


Figure C.2: The number of communities and the largest community for each year

Largest community divided by the total number of nodes

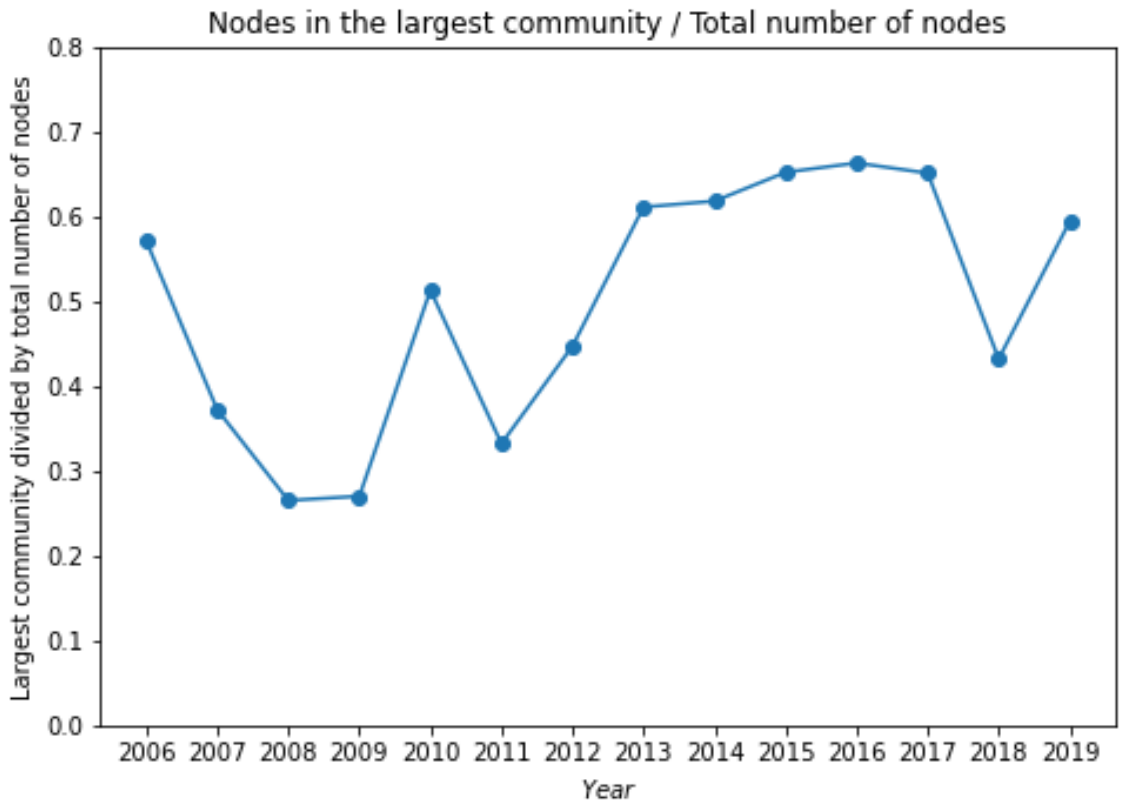


Figure C.3: The number of nodes in the biggest community over time is divided by the total number of nodes in each time slice.

Modularity

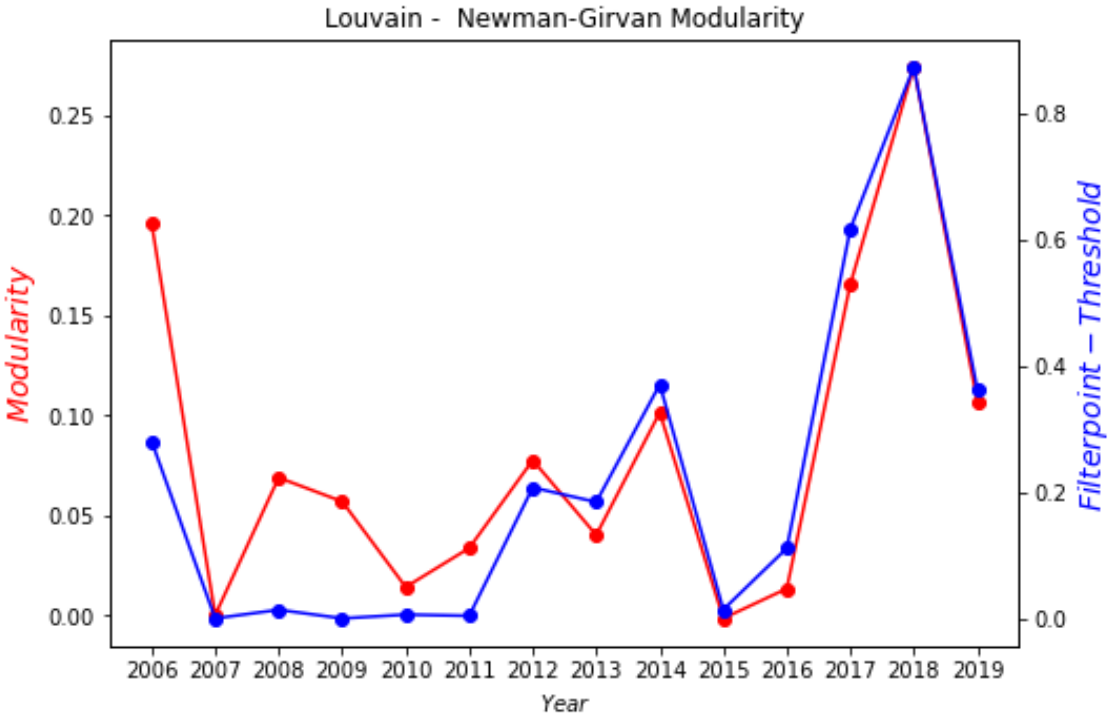


Figure C.4: The modularity score for each year and the applied thresholds for each time slice.

Conductance

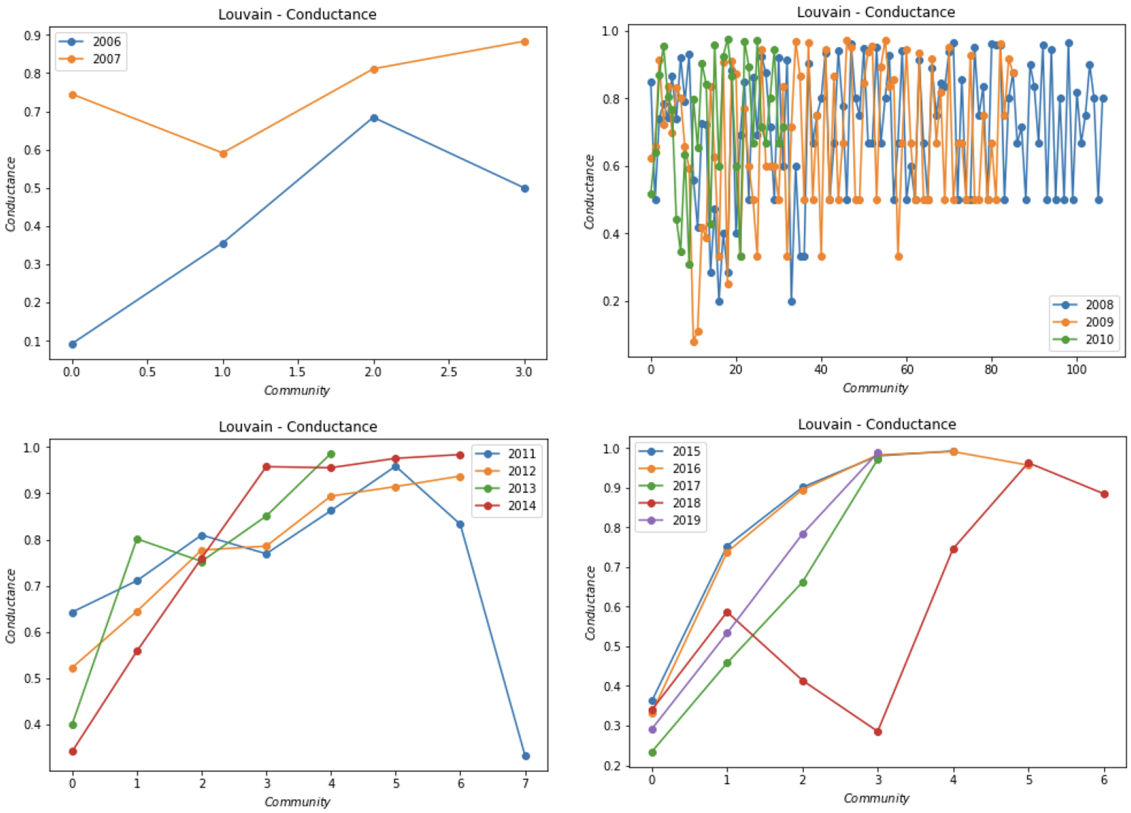


Figure C.5: The conductance score for each of the detected communities for each of the years.

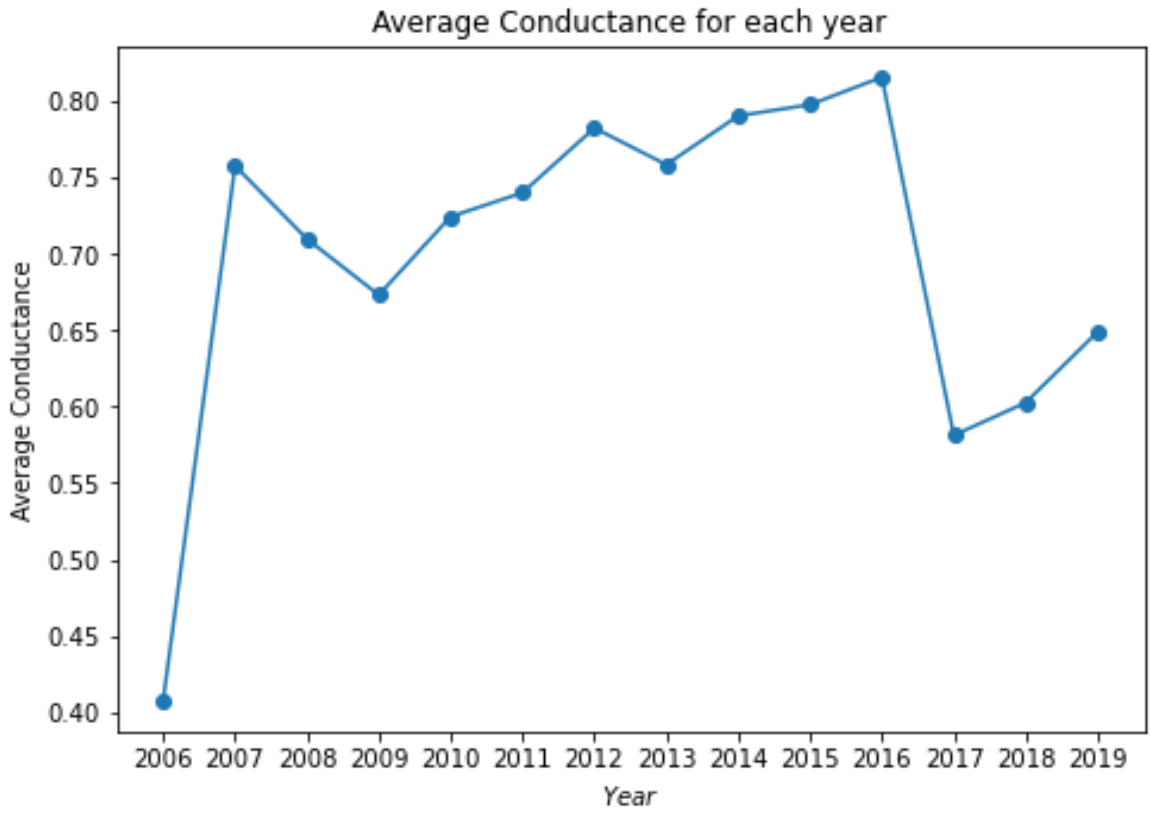


Figure C.6: The average conductance score for each of the detected communities for each of the years.

Community size and Edges

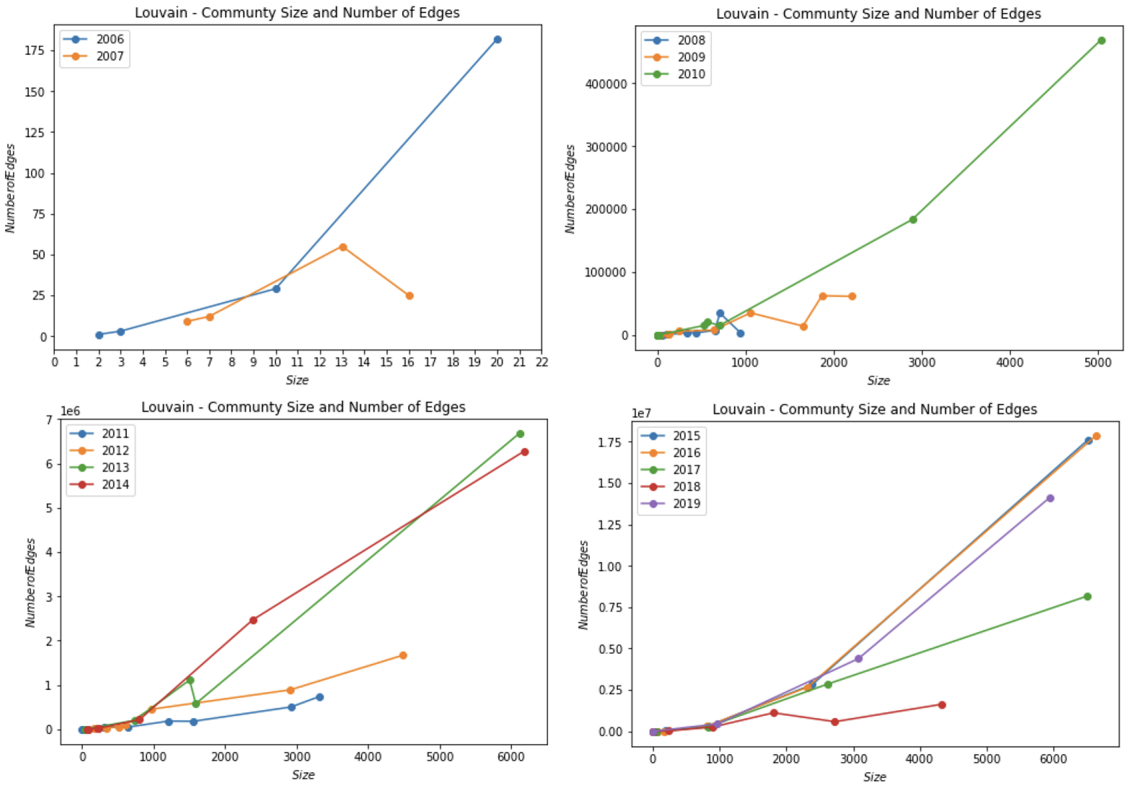


Figure C.7: The visualized result of the community sizes and the number of edges inside each community.

Appendix D

Appendix IV: Description of center subreddits

A short description of the subreddits that are the center in the largest and second-largest communities in the time-span of 2006 to 2019.

`r/nl`: a subreddit for Dutch-speaking people. The subreddit is no longer actively used.

`r/Request`: a subreddit where people can ask for favors.

`r/Entertainment`: a subreddit used for entertainment news. Today it has 3 million members.

`r/Reddit.com`: one of the original subreddits. The subreddit has been archived, meaning it can only be viewed.

`r/Health`: promotes itself as a science-based community that focuses on health and shares health-focused research articles.

`r/Travel`: is a subreddit for sharing travel experiences and asking travel-related questions.

`r/NewsReddit`: this community is at the current date (10.05.22) a private community.

Can only be accessed by approved members.

r/Education: is a subreddit focused on providing a space for discussing all things educational, such as educational policies, research, and technology.

r/FoodForThought: wants to create a community for people that want to share thought-provoking articles and essays.

r/Gardening: is a subreddit with gardening tips and plant care.

r/LinkinPark: this subreddit is for fans of the band Linkin Park.

r/ABDL: is a subreddit for adults that enjoy wearing diapers. This subreddit has an age policy of 18+.

r/Lebanon: is a subreddit about all Lebanon-related topics, from politics to cuisine.

r/FindAReddit: in this subreddit, people post what type of content they are looking for and get subreddit recommendations.

r/BabyBumps: is a subreddit for everything pregnancy-related.

r/UMD: the official subreddit for the university of Maryland.

r/AMA: is a subreddit called "ask me anything" and is used for asking questions.

r/ReBBl: is subreddit oriented around the fantasy football game Blood Bowl.

r/Aww: this subreddit is used for sharing cute pictures and videos.

r/PS4Planetside2: is a subreddit for the people interested in the PlayStation game Planetside 2.

r/WTFstockPhotos: this subreddit is used for sharing funny stock photos.

r/KinFoundation: is a subreddit for the cryptocurrency Kin.

r/RedditRequest: this is an official subreddit, meaning Reddit employees are the moderators of the subreddit. The subreddit is used for requesting abandoned subreddits or removing inactive moderators.