Universiteit Leiden

Master Computer Science

A network-driven understanding of user interaction in topical online communities

Name: [Niek Buwalda]
Student ID: [s1560700]

Date: [14/08/2020]

Specialisation: [Advanced Computer Science and Data Analitics]

1st supervisor: [F.W. Takes]
2nd supervisor: [C. Mattsson]

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Reddit is a large social news website with over 200 million unique users. The aim of this project is to analyse the way users interact within different communities of Reddit. Reddit has no explicit friendship links between users, such as Facebook has, so the social network is inferred from user commenting activity. In this thesis different Reddit communities will be compared based on three different approaches. Descriptive network metrics are used to analyse the static aspect of the networks. To include a temporal aspect of user interaction, the occurrences of temporal motifs within the networks are measured. The flow of information through the networks is modelled using a SIR model. Findings show several distinctions in interaction patterns and behaviour of users within the different communities.

1 Introduction

Everywhere on the internet, community driven discussions can be found. The forming of these communities and interaction between them have been studied extensively. An excellent source for data regarding these online communities and therefore a centre of studies, is the website Reddit. Reddit is a social news aggregation website. The platform houses a lot of serious discussions. On the other hand, also many less serious subjects can be found on Reddit, such as pictures of beautiful landscapes but videos of pets can also be shared on the site.
The content of the site consists of posts uploaded by a user. This user is referred to as OP (original poster). On every post on Reddit, users can post comments, either in response to the post or to an already existing comment. This can lead to large, complicated comment threads. To categorise the content, Reddit is divided into so-called 'subreddits', or simply 'subs'. Often, these subreddits are regarded as communities of Reddit. This is also the case within this thesis. The subreddits are created by the users. Every user can follow subreddits of their choice and receives content of all subreddits they follow on their feed. Subs are referred to by using the prefix *'r/'* followed by the sub name. There are over half-a-million different subreddits on Reddit, of which their purposes vary. Some subs contain serious discussions, for example *'r/politics'*, where news in US politics is discussed, or *'r/personalfinance'*, where users can ask financial advise from other users. Other subs are purely for entertainment, such as *'r/funny'*, where funny pictures and videos are posted, or *'r/FoodPorn'* where food enthusiast share pictures of good looking food. This indicates that subreddits are used differently, and therefore the interaction between users could differ per subreddit.

Figure 1 reinforces the suspicion of the differences in user interaction. With the purpose of introducing new users to platform, every new user is automatically subscribed to a selection of subreddits. In Figure 1, the ratio of votes and comments for the subreddits that selection of subreddits is shown. The average of comments per post differs significantly over the subreddits. This can be explained by the difference in purposes of the subs. The sub with the lowest ratio is *'r/AskReddit'*. On this sub users post all sorts of questions for everyone to answer. Due to, for example differing opinions, these questions can generate discussion, which leads to many comments per post. On the other hand, *'r/aww'* has the highest ratio. This sub is meant for sharing images or videos of animals doing 'cute' things. Such content is

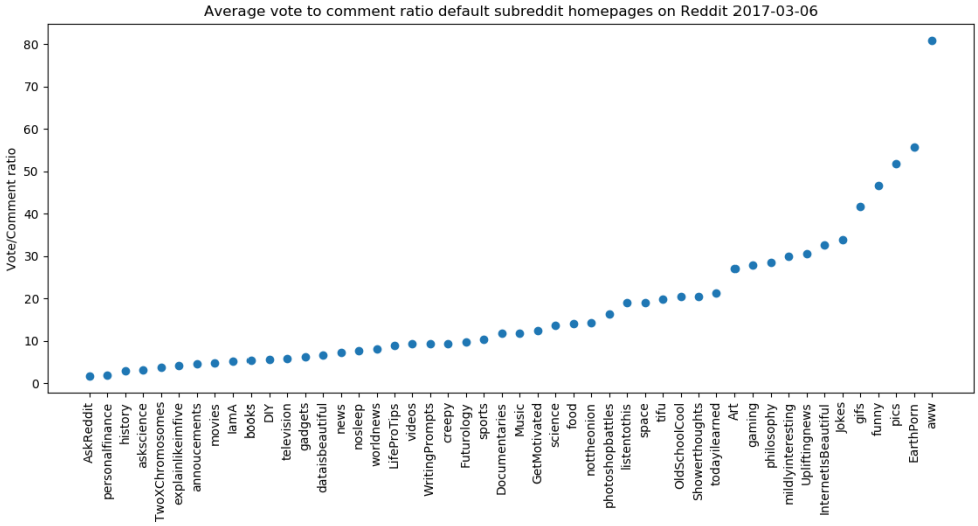


Figure 1: Vote to comment ratio of different subreddits

created to scroll past without commenting extensively, and thus provides less material for discussion.

As mentioned before, online communities on Reddit have been a popular topic of past research. One particular area of interest is the behaviour of users within online communities, such as loyalty to communities and how long users stay active within a online community [8]. Communities might also share a niche when, for example a new community is formed around an older already existing niche, or a new community is created as a counterpart to a already existing one. This leads to another area of interest, thee interaction between online communities, for example conflict between similar communities [10].

A less researched area is the interaction between users within an online community. Nevertheless, a study in the context of identifying roles within a online community has been conducted [4]. This paper suggest that users can be classified into certain roles based on their interaction with other users in a subreddit. This suggests there exists an underlying structure in these communities, but does not fully explain the interaction patterns within Reddit communities.

The aim of this paper is to better understand the interaction within the communities of Reddit. Therefore, the research question of this thesis is: "How can a network-driven understanding of the differences in user-level interaction within topical online communities be obtained?"

To answer this question a network science approach will be used. Network science is a broad field, which studies many aspects of complex networks such as structural network properties [12], the modelling of networks [5], network visualisation and modelling spread of content within networks [9]. The aim of this thesis is to explore three aspects of the Reddit communities, as there are many aspects that influence the interaction between users. The first aspect that is explores are network properties. Using static network metrics an overview of the structure of the network is established. Second, the temporal aspect of the user interaction is analysed using temporal motifs. Temporal motifs are small patterns that occur over a time period within a network. When the occurrences of different interaction patterns are measured, they provide insight to the dynamics of interaction between users. Subreddits that are dominated by discussion are expected to have more reciprocal edges, than for example a subreddit that is meant for sharing pictures. Third, the information flow within the communities of Reddit is studied, using the SIR model, a compartmental model designed to simplify the mathematical modelling of infectious diseases. The information spread gives an indication of the connectedness of a community, as the information will spread easier through a highly connected community than a more loosely connected community.

This results in three different subquestions:

1. "What can static network metrics contribute to understanding user-level interaction within an online community?"

2. "What can interaction patterns, derived from temporal motifs, contribute to understanding the dynamics of user-level interaction within an online community?"

3. "What can the information flow through an online community, modelled by a SIR model contribute to understanding user-level interaction in that community ?"

The rest of this thesis is structured as follows; first, the data used for this thesis is discussed in Section 2, followed by statistics of the data and a description of the relevant attributes. Also, a short description for used subreddits for this project follows. In Section 3, the methods to answering the three subquestions are discussed in detail. The results of result of the three subquestions will be presented in Section 4. Finally, the findings will be summarised and discussed in Section 5.

2 Data

This chapter describes the data source for this project and gives an overview of what the data contains. Afterwards, a short description follows of the subreddits used in this project along with an explanation of the selection criteria. To finalise this chapter the method to create networks from the data is explained.

2.1 Dataset

Reddit offers an API to retrieve everything posted on Reddit along with a vast amount of metadata. One of Reddit's users has created a repository of all posts and comments from December 2005 up until the present [1]. The comments are sorted by the time posted and are grouped per month in separate files. Each comment is in JSON format and contains a lot of information, such as the time of posting, to which message it was a response and the original post. Table 1 gives a description of the relevant data attributes. For this project the time period of 1 January 2016 up until 31 December 2016 is chosen. The data in this time slot is bigger than earlier years, as the user base of Reddit still grows, but not to large to prevent long computational time for the experiments. Also, this time slot is still relatively close to the present and is relatively free from subreddit being banned from Reddit. The size of the data from this time slot is around 50.4 gigabytes. However, after extracting the content of only the specific subreddits used for this project, the size of the data is reduced to 562.9 megabytes.

Attribute	**Explanation**
id	Unique identifier of the comment
author	Username of the poster of the comment
subreddit	Name of the subreddit the comment was posted on
created_utc	Timestamp of when the comment was posted
body	Content of the comment
parent_id	Id of the parent comment or post

Table 1: Description of used data attributes

2.2 Selection of Subreddits

There are over a million of subreddits on Reddit. However, only small subset is actively used by a large amount of users. Therefore, as a first selection criterion, selection has been made based on the amount of users that follow a subreddit. All subreddits used in this project are in the top 500 of all subreddits at moment of publication, according to `http://redditlist.com/`. They also nearly all have more than 20.000 active users in the chosen time period that was used. This ensures there are enough interactions between users. However, there is also one smaller community in this projects to investigate whether the methods used are suitable to analyse smaller communities as well. The second criterion is general purpose of the subreddits. As mentioned before, the purpose of the different subreddits varies. With the aim to identify characteristics of the different types of subs, the selection contains diverse as well as similar subreddits. Table 2 contains an overview statistics of the subreddits studied in this project within the chosen time period.

Subreddit name	**Number of users**	**Number of comments**
'r/AskMen'	71,586	640,521
'r/AskWomen'	59,294	504,685
'r/aww'	307,962	797,072
'r/totalwar'	20,578	196,010
'r/AmItheAsshole'	2,245	6,556
'r/IAmA'	332,872	834,993
'r/formula1'	31,162	503,838
'r/programming'	48,345	302,145

Table 2: Statistics of relevant subreddits

The first two subreddits used in this project, *'r/AskMen'* and *'r/AskWomen'* have a similar purpose. The *'r/Ask...'* format is very popular on Reddit and appears in many variations on the site. The idea behind these subreddits is that everyone can ask a question to a certain group of people, for instance, scientists, the whole of Reddit or in this case men or women. These subreddits are meant for discussion, mainly between the answering group. Also, the difference of interaction between men and interaction between women could be interesting.
The third subreddit is *'r/aww'*. This subreddit mainly consists of pictures

and videos of cute animals. The amount of discussion is expected to be limited on this subreddit, as the main goal is to share photos and short videos. However, this subreddit is very popular, with millions of followers.
Next, a community with a more specific niche is included, namely of people that play the same type of game, *'r/totalwar'*. Total War is a strategy game series. This subreddit is for example for discussions, strategies, stories of any of the games in the series. It is expect that the community is relatively tight and discussions to be more intense.
The following subreddit has a special voting system within the comments of post. In *'r/AmItheAsshole'* users can post situations which they have experienced in which other parties did not agree with their choices, which often resulted in conflict. Readers of the posts can put a designated tag in their comment side with the OP or the other parties. The idea is to help determine the OP if their actions where socially acceptable and/or justified. This subreddit is expected to produce a lot of intense discussions as they are about moral questions. This is the smallest subreddit in this project.
The next subreddit, *'r/IAmA'* also has a special format. In this subreddit, the OP offers to answer any questions other users might have about their profession, hobby or any other particular special feature of their life. This might lead to unique patterns in the interaction of users.
To include a subreddit that deals with actualities and news, *'r/formula1'* is also used for this project. It deals with news and stories from everything that revolves around the Formula 1 races. The content on this sub might be more time sensitive. As new events happen, post about older events become less relevant for users to comment on. For example, a new race comping up will make older races less relevant.
The last subreddit in this project is *'r/programming'*. Although programming itself could be considered a niche, the sub itself has a relative broad spectrum. The nature of the sub might not immediately promise unique discussions, but perhaps discussion patterns of technical topics differ from for example moral questions.

2.3 Network creation

For all of the subreddits above directed weighted networks are constructed. The data itself does not have a direct notion of interaction between users, but it contains the information necessary to create a network based on implicit user interaction. The *parent_id* of a comment, can be used to link the

author of the current comment (A) to another post or comment. When the information of the author of that post or comment (B) is combined with the previous information, two users can be linked to each other. In other words an edge can be constructed between them in a network, as they interacted. The direction of the edge is from the responding user (A) to the other one (B). The weight of the edge between two users indicates how often they interacted. The timestamp of interaction can also be extracted from the data, which is needed to derive the temporal motifs.

3 Approach

In this thesis three different aspects of networks and their role in user interaction will be studied. This chapter describes the three aspects in detail and provides necessary theoretical background. First all the static network metrics will be explained and defined. Second, this chapter will go over the details of the temporal motifs and the difference in the requirements for the networks to utilise the temporal motifs. Third, the SIR model is described along with the transformation to apply it to the networks in this thesis.

3.1 Static network metrics

From the created networks a lot of information can be extracted that might give insight in user interaction. For the comparison with real-world networks, the properties that are generally defining for real-world networks [12] are examined. The following metrics will be examined:

- Density
- Degree distribution
- Average distance
- Average clustering coefficient
- Transitivity
- Size of giant component
- Average degree connectivity

- Modularity
- Closeness centrality
- Average shortest path distribution

3.1.1 Density

The density of a graph is defined as the ratio between edges that occur in the graph and all possible edges, when considering the graph as unweighted. This gives an impression of the interaction intensity on a subreddit. For example, the graph of Figure 2 has 5 nodes, so the maximal possible edges is 10. Therefore the density is $\frac{5}{10} = \frac{1}{2}$.

3.1.2 Degree distributions

Real-world networks have fat tailed power law degree distribution, meaning they have a lot of nodes with a low degree, and few nodes, so called hubs, with a very high degree. The node degree is the number of connections this nodes has with other nodes. Because the subreddit networks are directed graphs, a distinction must be made between indegree and outdegree. Therefore, the distributions of the indegree and outdegree of the subreddits will be compared. From these distributions can be concluded whether a community revolves around a smaller group of influential users, or if all users contribute equally. Also, the exponent of the power law that approximates the distribution is measured, referred to as γ.

3.2 Average Distance

Another property of real-world networks is the small world phenomenon; from every node, every other node can be reached via a relatively short path. To investigate if this the case in the subreddit networks, the average distance between the nodes of the network is calculated. A low average distance indicates that users interact with users all over the subreddit, instead of communicating in smaller isolated groups, when the average distance is high.

3.2.1 Average clustering coefficient

Another common property of real-world networks is a high clustering coefficient. Real-world networks tend to have many triangles compared to ran-

domly generated graphs. For the different subreddit networks the average clustering coefficient is calculated. This is the average of the local clustering coefficients[12] of all nodes in the network. For example, see Figure 2. Node 3 has only one closed triangle, while there are 6 possible triangles, and therefore has a local clustering coefficient of $\frac{1}{6}$. The local clustering coefficient of the nodes is $[1, 1, \frac{1}{6}, 0, 0]$ corresponding to the node indices, resulting in $\frac{13}{30}$ on average. The clustering coefficient indicates how likely users form smaller groups with the community that interacts among themselves.

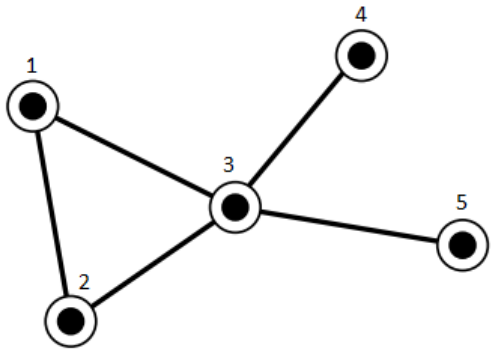


Figure 2: Example graph

3.2.2 Transitivity

A metric related to the clustering coefficient is transitivity [12]. Although they are similar, transitivity values high degree nodes more than the average clustering coefficient. Transitivity is defined as the fraction of three times triangles in a graph to all triplets of connected nodes. To illustrate the difference between the average clustering coefficient and the transitivity, the transitivity of the graph in Figure 2 is $\frac{3}{8}$. There is only one triangle in the graph. And there are 8 triplets of connected nodes:
$(1, 2, 3), (1, 3, 2), (1, 3, 4), (1, 3, 5), (2, 1, 3), (2, 3, 4), (2, 3, 5), (4, 3, 5)$. The same as the clustering coefficient, the transitivity indicates the forming of smaller groups within in the community.

3.2.3 Size of giant component

Real-world networks usually have one connected component that is very big, and that therefore is referred to as the giant component (GC) of the network.

A connected component is a group of nodes, wherein from every node, all other nodes can be reached. For this project, the percentage of nodes and edges that belong to the biggest connected component for each subreddit is measured. From this measurement, some insight can be gained into the fracturing of the community. If there exists smaller exclusive groups within the community, the giant component will not be big as when users interact freely with each other.

3.2.4 Modulartity

Randomly generated graphs usually have a different structure than real-world networks. Modularity [13] is the fraction of the edges that fall within a cluster minus the fraction of edges that would be generated if the graph was random. Optimising this value as is done by popular community algorithms gives an indication of how clustered the network is as opposed to a randomly generated network. The value of the modularity is in the range between -1 and 1. When the value is positive, the cluster has more edges than when it would be generated at random. To measure the modularity of our networks, first the networks are partitioned into clusters using the Louvain algorithm [2]. Then, the modulartity is calculated based on the obtained clustering. The modulatity score also gives a notion of the existence of smaller groups within a community.

3.2.5 Closeness centrality

To find important nodes in a network there are numerous different centrality measures [6]. In this project closeness centrality is used. Closeness centrality of a node is the reciprocal of the sum of the shortest path distances to all other nodes. Since the sum of distances depends on the number of nodes in the graph, closeness is normalised by number of other nodes. For all nodes in the giant component of the subreddit networks the closeness score is calculated to create distributions. As this is computational expensive, for the subreddits *'r/aww'* and *'r/IAmA'* a sample will be used to approximate the distributions [3].
The distributions of the closeness centrality can says something about the import users within the communities. These users have a high centrality value. There might hand full of very active users that play a key roll in the community or a larger group of users that all contribute small amounts.

3.2.6 Average shortest path distribution

To further examine small world phenomenon and the influence of hubs, distributions of the average shortest path length will be created for the subreddit communities. A sample of 20 percent of nodes of the giant component is made, which is a significant enough to approximate the network [3]. For this sample, the shortest path to all other nodes in the giant component will be calculated, as exhaustively calculating the average shortest path for every node to every other node is very computationally expensive. These distributions can show if users interact with each other all over the community or rather within smaller groups. A community of the first scenario will have a lower average distance, a community of the second scenario a higher average degree. Also, the average path length says something about the presence of hubs. Hubs lower the average shortest paths within a community, as they interconnect large portions of the network, due to their high degree.

3.3 Temporal motifs

he measurements mentioned above are all done on a network that is assumed to be static. However, the time between posts and the dynamics between users can say a lot about the nature of interaction between users. Therefore, all three edge temporal motifs of the network are investigated. A temporal motif [14] is a specific pattern in a graph with timestamped edges. This pattern is a directed multigraph, with the edges ordered based on the timestamps. For all possible three edge motifs, see figure 3. Within this Figure, the arrows represent the interaction of users. The interaction occurs in the direction in which the arrow is pointing. The numbers beside the arrows represent the order in which the interaction occurs, from 1 to 3. When dealing with temporal motifs, some changes to the network are needed. Edges are still directed, but should not be weighted. Instead, each interaction is stored as unique edge with a timestamp as attribute, resulting in the required multigraph.
Eventually every post will stop receiving new comments. But the interaction patterns on a post might be different in the first hour after posting, compared to after a day. This difference can be visualised using a parameter, that sets an interval that is allowed to pass between the possible interactions.

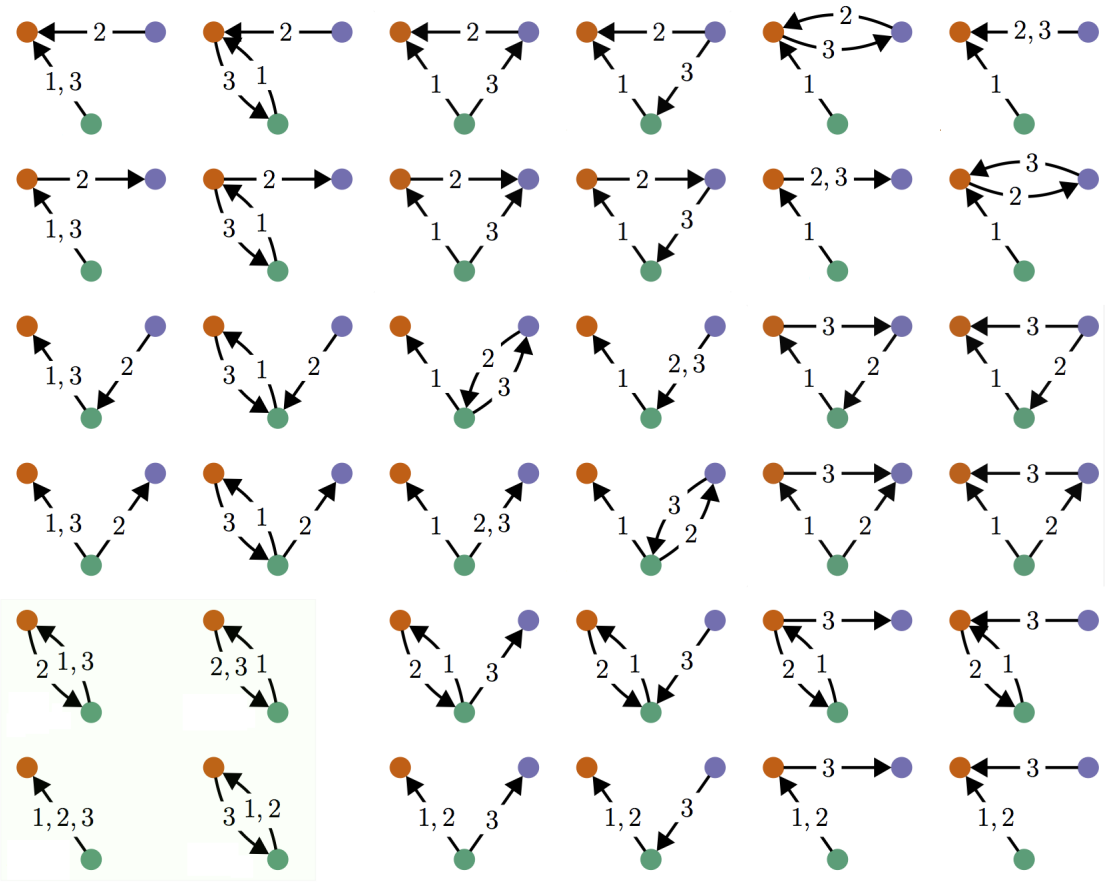


Figure 3: All three edge temporal motifs (figure obtained from [11])

3.4 SIR

The SIR model is a mathematical model to model infectious disease spread [9]. In this model nodes are labelled with one of three states; (S)uscepitble, (I)nfected or (R)emoved. In the most basic model, see Figure 4, it is only possible for an individual to move from the Susceptible to the Infected state and from the Infected state to the Removed state. These transitions are determined by probabilities. The transmission rate, β, determines the change of an infection of a susceptible individual, upon interacting with an infected individual and the recovery rate, γ, which indicates the chance of an infected individual to recover and transition to the recovered state. In the most basic model, when an infected individual recovers, they become immune to

the disease and will remain in the recovered state. In graphs with a large number of nodes this model is defined by the following nonlinear differential equations: $\frac{dS}{dt} = -\beta IS$, $\frac{dI}{dt} = \beta IS - \gamma I$, $\frac{dR}{dt} = \gamma I$ where $S(t)$, $I(t)$, and $R(t)$ are the fractions of the population in each of the three states and t the current timestep.

Figure 4: Most basic SIR model

This model has proven to work when applied to social network graphs [7]. To be able to use the model on the subreddit graphs a few adjustments need to be made. The nodes in our network represent the Reddit users, which are the individuals for the SIR model. However, in the model motioned above contact between individuals is modelled randomly. Every susceptible individual can interact with every other individual from the entire population. Also, the transmission rate is the same for every interaction. To adapt this model to networks, a few adjustments are made. First of all individuals can only interact other individuals if there exists an edge in the correct direction. Second, the weight of the edges influences the infection rate. Nodes that are strongly connected have a higher chance to be infected than loosely connected nodes. The transition rules are as follows: If a node is infected, its susceptible neighbours have a chance to be infected. The probability of infection is influenced by the weight of the edges from the infected node to the neighbours. A higher weight indicates a stronger connection, and therefore a higher probability of infection. If a neighbour is infected it switches to the Susceptible label at the start of the next timestep. All nodes begin in the Susceptible state, except for one or group of nodes, which is initially infected. During the simulation the state of each node is kept track of. Which each timestep, the transition rules is applied to every node, and every node is updated accordingly. Only the giant component subgraph of the networks is considered in these experiments to guarantee the initial infected node is not an isolated node or is in a isolated community. This enables the spread to reach the entire subgraph. This model provides insight into how easily a

topic spreads through a community. When it spreads easily, the subreddit is probably quite strongly connected, and a large portion of the subreddit picks up on that particular topic. On the other hand, within a fractured subreddit topic are unlikely to spread to a large portion of all the users that subreddit.

4 Experiments

In this Chapter the findings of the experiments will be discussed. Every metric is analysed separately, based on expectations and anomalies compared to other subreddit. But first the experimental setup is discussed.

4.1 Experimental Setup

For this project, all code is written in Python 3.6.4 using the Networkx package for graph construction and calculating metrics. With the exception of the temporal motifs. To create the motifs, code was used provided by the Stanford Network Analysis, SNAP[11]. Calculation of the metrics and other calculations were done on the data science lab of LIACS.

4.2 Static metrics

Table 3 gives an overview of the numeric metrics of the studied subreddits. In the following subsections they are discussed in detail individually. Subsequently, the relevant distributions are visualised in graphs and are discussed in the subsections 4.2.7, 4.2.8 and 4.2.9.

	AskMen	*AskWomen*	*totalwar*	*aww*	*AmItheAsshole*	*IAmA*	*formula1*	*programming*
Nodes	71586	59294	20578	307962	2245	332872	31162	48345
Edges	640521	504685	196010	797072	6556	834993	503838	302145
Density	0.00012	0.00014	0.00046	0.0000084	0.00130	0.0000075	0.00052	0.00013
Clustering coef.	0.06044	0.07491	0.08213	0.01532	0.04696	0.04156	0.09839	0.04683
Transitivity	0.04508	0.05035	0.05044	0.00218	0.02344	0.00278	0.074395	0.02424
Average degree	8.9476	8.5116	9.5252	2.5882	2.9203	2.5085	16.1683	6.2498
Nodes in GC(%)	99.3	99.3	99.1	94.8	97.1	98.5	99.6	98.7
Edges in GC(%)	99.9	99.9	99.9	98.7	99.3	99.6	100	99.9
Modularity	0.21926	0.21728	0.28255	0.50261	0.53810	0.60227	0.17657	0.28089

Table 3: Statistics of all studied subreddits

4.2.1 Density

All subreddits have a fairly low density. However, two subreddits have a significant lower density, namely *'r/aww'* and *'r/IAmA'*. This could be explained by the nature of these subreddits, which are not discussion. Therefore, users might interact less with other users than in the other subreddits, resulting in relatively fewer edges.

4.2.2 Average Clustering coefficient

Again *'r/aww'* scores very low, meaning on this subreddit subreddit seldom interact within groups of users, which is not very surprising. On the other hand, *'r/formula1'* by far has the highest average clustering coefficient. This is surprising as this sub has no special format that stimulates clustering. It could mean that *'r/formula1'* is more tight community than excepted.

4.2.3 Transitivity

From the transitivity a similar observation can be made as from the density. *'r/aww'* and *'r/IAmA'* score significantly lower the the other subs, which reinforces the assumption that the nature of the subreddit has a considerable influence in user interaction. Interesting to note that is that although transitivity and the average clustering coefficient both measure a form of clustering, some different observations can be made from the transitivity. The fact that *'r/IAmA'* score relatively low on transitivity, but does not have a that low average clustering coefficient, suggest that there no really important hubs.

Due to the nature of this sub, one person answering questions, this could have the case.

4.2.4 Average degree

From the average degree the same conclusion can be drawn as the previous metrics, the discussing topics generally score higher. The exception is *'r/AmItheAsshole'*. This might be due the number of users in this subreddit, which is significantly lower compared to the other subs.

4.2.5 Size of largest connected component

The percentage of nodes and edges in the giant component are relatively high in all subreddits. Because post of subreddits do not have to be interesting to for everyone, or have to be even seen everyone on the subreddit, this is surprising. Apparently contributors of a sub interact enough to form a big giant component.

4.2.6 Modularity

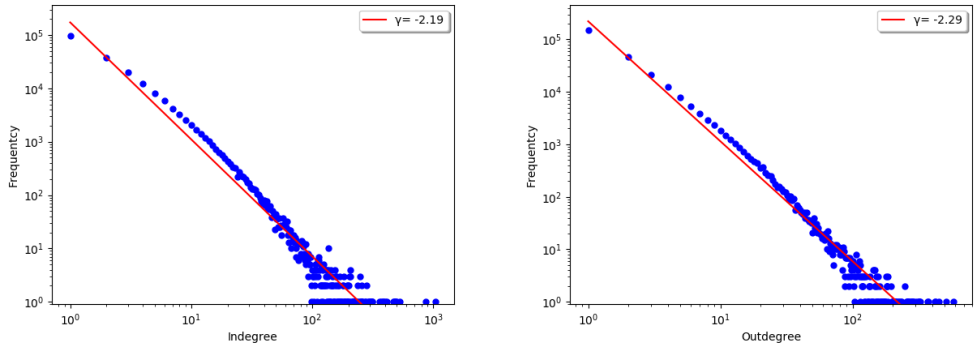
Looking at the modularity of the subreddits, two groups can be distinguished, although all subreddits have a positive modularity score, which means they are more clustered than randomly generated graphs. The first group, *'r/AskMen'*, *'r/AskWomen'*, *'r/totalwar"* and *'r/programming'*, have a significant lower modularity score than the second group. This could be due to the fact the range of the discussion on these subs is very broad and more based on personal experience and preference, rather right and wrongs. The second group, *'r/aww'*, *'r/AmItheAsshole'* and *'r/IAmA'*, have modularity score then the first group. For *'r/AmItheAsshole'* the tendency to cluster more might be explained by the nature of the discussions. The users can decide who is wrong and who is right in the stories that are posted. This might leads to groups of users with similar opinions agreeing with each other, or groups of users with different opinions constantly arguing. Summarising these findings, it could be that discussion in a subreddit leads to a less clustered community, which is quite interesting.

4.2.7 Degree distributions

For this experiment a distribution for the indegree and outdegree is created for all subreddits. In Table 4 the exponent of the lines that approximate the distributions is shown. Again, *'r/aww'* stands out, as the exponent is lowest for both the indegree and outdegree. Meaning it resembles a scale-free network the most, having relatively more low degree users and fewer high degree users. This falls in line with the expectations and many other results. A second observation is that degree distribtions of *'r/aww'* follow a powerlaw, as shown in Figure 5. All other subreddit networks also have degree distributions that follow a powerlaw, as can be seen in Appendix A.

	AskMen	*AskWomen*	*totalwar*	*aww*	*AmItheAsshole*	*IAmA*	*formula1*	*programming*
γ for indegree	-1.49	-1.40	-1.53	-2.19	-1.76	-1.62	-1.43	-1.71
γ for outdegree	-1.60	-1.56	-1.50	-2.29	-1.81	-1.92	-1.43	-1.65

Table 4: Degree exponents of the in- and out-degrees, respectively



(a) Indegree (b) Outdegree

Figure 5: Degree distributions of *'r/aww'*

4.2.8 Closeness centrality

The closeness distributions of the subreddit networks show similarities, see Figure 6. The average score is very low over all networks, and lies in general around 0.2 to 0.25. For *'r/aww'*, in Figure 6c, and *'r/IAmA'*, in Figure 6f the score seems to be even lower on average, however this is an approximation using a sample. Therefore, it seems none of the subreddit networks have key users, that are in the centre of the network.

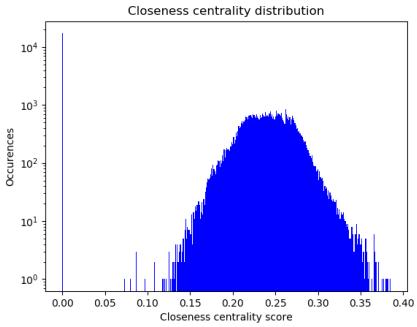
4.2.9 Average shortest path distribution

The average path length for all subreddits is within the range of 4 to 6. What immediately stands out is the different ranges occurring values values. The broad range of *'r/aww'* and *'r/AmItheAsshole'* are not surprising. These subreddits are not meant for large groups of users interacting with each other, hence the distance between users is expected to be higher.

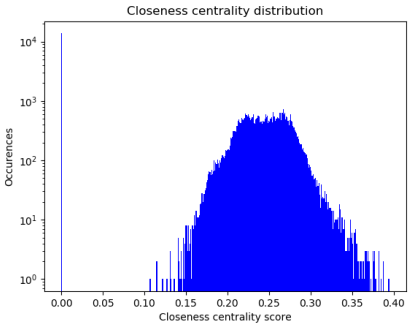
Also, these results are in line with the modularity scores. The more modular networks in general have more distance between nodes. Another possible explanation could be that subreddits with a tighter community are able to retain users for extended periods of time, than subreddits with a loose community. If users remain active within a community for a while, it increases the odds of them interacting with each other of the community.

Also interesting is the difference in range of occurring distances between *'r/AskMen'* and *'r/AskWomen'*. For most other metrics these two subreddits perform similar. Apparently there are *'r/AskWomen'* fewer outlying users, although for example the average distance is very similar.

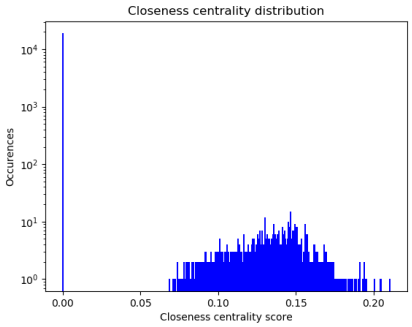
Also unexpected is the relatively narrow range of *'r/formula1'*. The average shortest path distribution is most similar to that of *'r/totalwar'*, which was expected to be the most closely connected subreddit.



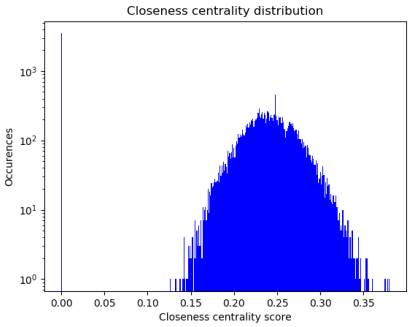
(a) AskMen



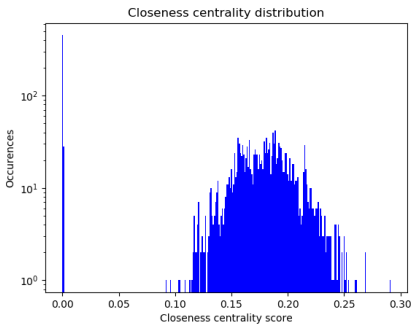
(b) AskWomen



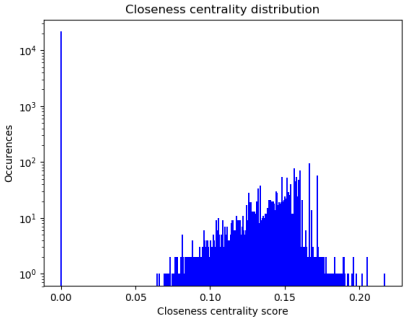
(c) aww



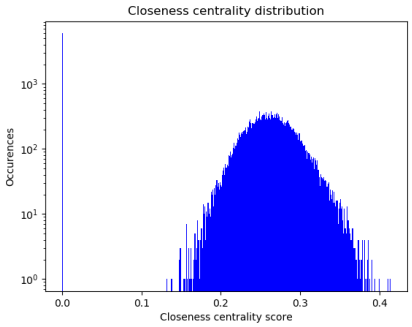
(d) totalwar



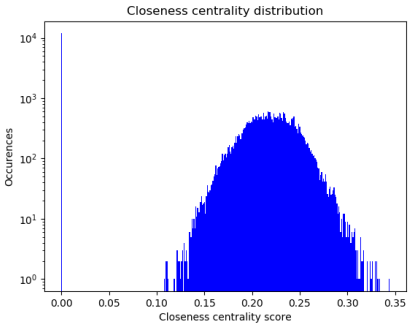
(e) AmItheAsshole



(f) IAmA



(g) formula1



(h) programming

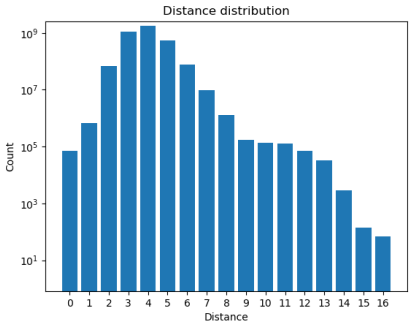
Figure 6: Closeness Distributions

4.3 Temporal motifs

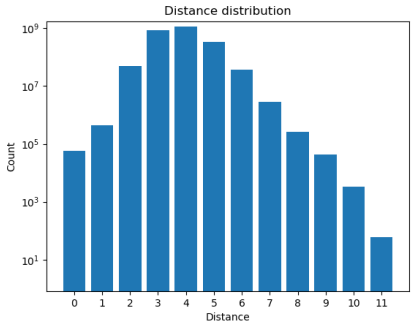
As discussed in chapter 3.3, to perform experiments a parameter is required that determines the time that is allowed to pass between interactions. To gain insight into the effect this parameter has on the occurring motifs, a experiment is performed on 'r/AmItheAsshole' with eight different settings of the parameter. The results of this experiment are shown in Figure 8. To designate different motifs, the indices of Figure 3 will be used as reference, using a prefix M with the corresponding numbers of the row and column, e.g. $M_{3,5}$ represents the motif in the fourth row and sixth column. From Figure 3 several observations can be made. Interesting to see is that all heatmaps contain the same eight fields, which are $M_{0,2}$, $M_{0,3}$, $M_{1,2}$, $M_{1,3}$ and $M_{2,4}$, $M_{2,5}$, $M_{3,4}$, $M_{3,5}$, that always have low values. Upon closer inspection, it makes sense for these motifs to appear less frequent. These motifs deal with triangular interactions. As the comment section on Reddit is structured as a tree, these type of interactions have a smaller chance of happening naturally. A user might respond to multiple other users, but those other users have no particular reason to interact. Also, a transition is occurring over the different time intervals. When the time interval between the interaction is 15 minutes, the most prominent motif is $M_{0,4}$. This motif describes an back and forth between two users. This is a expected result, as it is normal that for example the original user answers any comments made on a fresh post for example. As the time interval increases, other motifs rise in prominence. These motifs describe patterns with three users. Apparently, nearly all discussions involve more than two users. This is also not surprising as the comment structure on Reddit is meant to branch out and include many users.

After the twenty four hour time interval all prominent patterns no longer include only two users. Also, the heatmap for this time interval show little changes in comparison to the heatmap of the twelve hour interval. Therefore, a heat map for the other subreddit networks is created using a twenty four hour heat map. The result can be found in Figure 9.

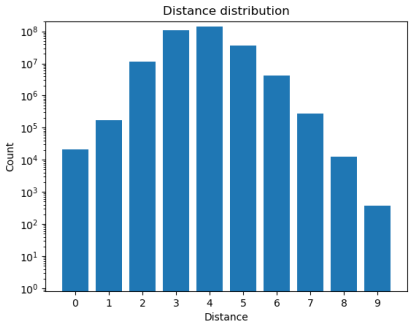
There are several patterns that appear in all twenty four hour heatmaps. The motifs $M_{0,4}$, $M_{2,2}$ especially, $M_{3,3}$, $M_{4,3}$, $M_{4,4}$ and $M_{4,5}$ always appear as prevalent motifs, and $M_{2,3}$, $M_{3,2}$ and $M_{4,2}$ to a lesser extent. These motifs are not unexpected as they fit into the Reddit comment section structure. For example $M_{2,2}$ is a excellent example of user interaction in the comment section of a post. The orange user has posted a comment. Turquoise comment this comment, for example a question about orange's comment(1). Another



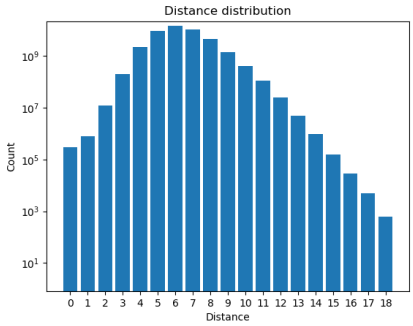
(a) AskMen



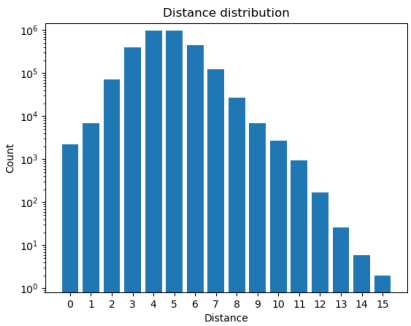
(b) AskWomen



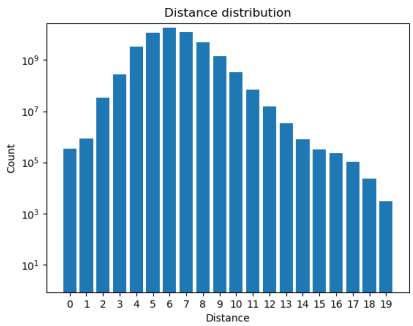
(c) totalwar



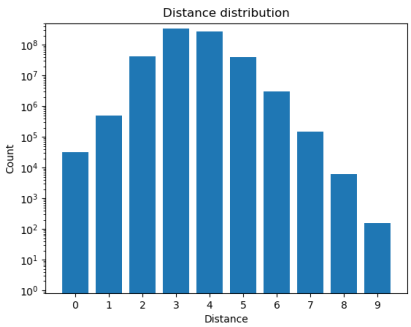
(d) aww



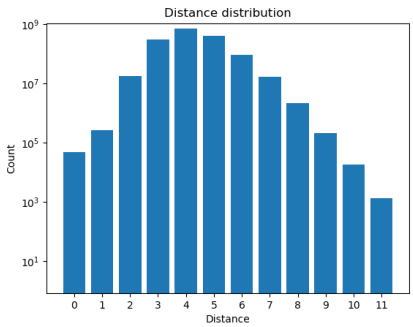
(e) AmItheAsshole



(f) IAmA



(g) formula1



(h) programming

Figure 7: Average shortest path distributions

users, purple, answers turquoise's comment(2). Turquoise thanks lilac or asks a followup question(3). However, it is interesting that other motifs, such as $M_{1,0}$, $M_{1,1}$ and $M_{1,2}$ are very infrequent in all subreddits, while seem very viable in the structure of the comment section.

'r/AskMen', *'r/AskWomen'*, *'r/totalwar'*, *'r/formula1'* and *'r/programming'*, as seen in Figure 9a, 9b, 9c, 9g and 9h respectively, have fairly fairly similar heatmaps. *'r/aww'* also has a resembling heatmap, however a number of motifs appear relatively less frequent, mainly $M_{2,3}$, $M_{2,4}$ and $M_{3,2}$. For *'r/AITA'*, see Figure 9e, the motifs $M_{4,0}$ and $M_{5,4}$ are not as prevalent as in the previously discussed heatmaps. These motifs both deal with two user interaction with the same third user. It could be that the discussion on this subreddit are slightly more personal, and therefore are more between users. However other motifs as $M_{2,2}$ do appear, so it is a minor difference. The most unique heatmap is the one of *'r/IAmA'*, as seen in Figure 9f. The motif $M_{5,5}$ has a distinctive presences. Furthermore, $M_{1,5}$ is also relatively more present than in the other subs. This can be easily explained by the purpose of the sub, one users answering questions from others. These motifs both contain a user that receives more comments than that user writes as a response. A scenario where a post on this subreddit becomes popular and the answering user can not keep up with the number of questions asked, is definitely realistic.

4.4 SIR model

Before the simulation of the SIR model can be executed for the subreddit networks, the transmission rate and recovery rate have to be determined. For these experiments the transmission rate is set to $\frac{1}{3}$ and the recovery rate to $\frac{1}{8}$. From the simulation two important results are reported in Table 5. The first results is the percentage of users who remained susceptible. This indicates how much of the network was affected by the information. The second is the is the percentage of infected that occurred in the simulation, which indicates how fast the information spreads. The results are not surprising. The subreddits with the highest modularity, *'r/aww'* and *'r/IAmA'*, have the least spread throughout the network. It is logical that information does not spread quickly through a clustered community. Also, the networks with low modularity, *'r/totalwar'* and lower average distance, *'r/forula1'* and *'r/programming'*, provide the highest percentage of infected users and the least uninfected users. In Appendix B all the plots of the simulations of

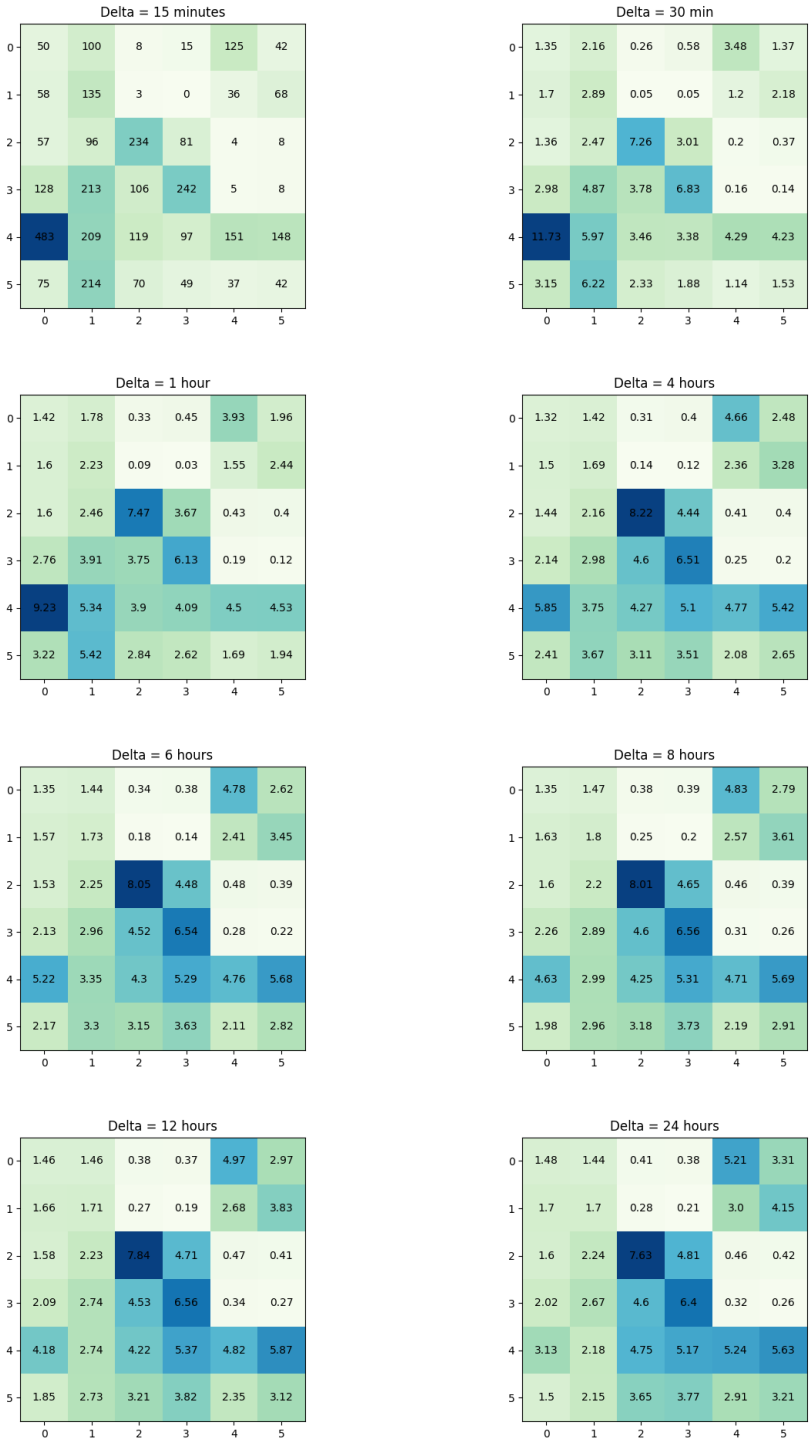
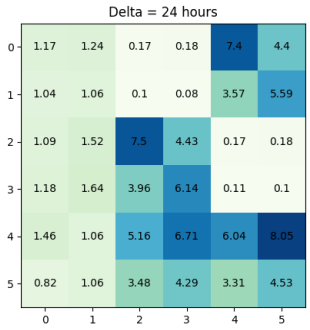
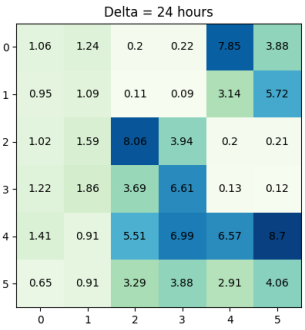


Figure 8: Motif counts of *'r/AmItheAsshole'* for different time intervals in percentages

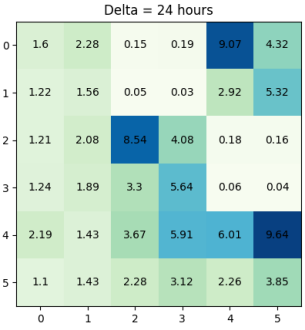


(a) AskMen

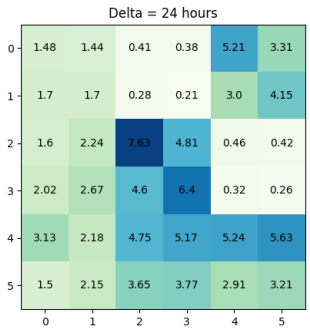


(b) AskWomen

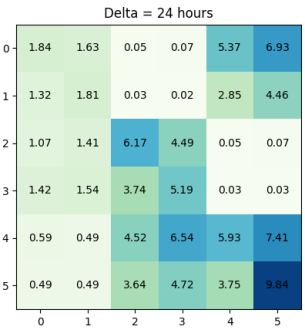
(c) totalwar



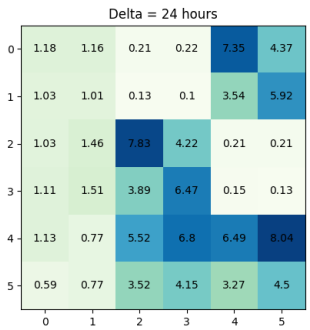
(d) aww



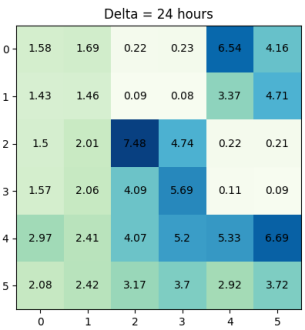
(e) AmItheAsshole



(f) IAmA



(g) formula1



(h) programming

Figure 9: Motif counts of subreddits networks using a 24 hour time interval

all subreddit networks can be found.

	AskMen	*AskWomen*	*totalwar*	*aww*	*AmItheAsshole*	*IAmA*	*formula1*	*programming*
% lowest susceptible	43.7	43.2	30.3	66.0	44.3	67.1	31.5	37.6
% highest infected	55.8	56.3	69.1	37.4	55.1	32.3	67.9	61.8

Table 5: Results of SIR simulations

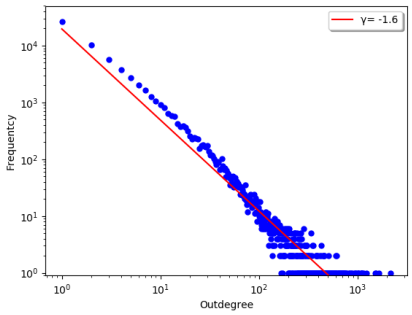
5 Conclusion

In this thesis three network-driven methods have been used to gain insight into the user interaction within communities on Reddit. Combining these measures, several important conclusions can be drawn. There are similarities between the subreddit networks. The degree distributions of the communities all follow a power law distribution. Also, several temporal motifs appear prominently in most heatmaps and some motifs did not appear prominently in any of the heatmap, although they could be very possible.

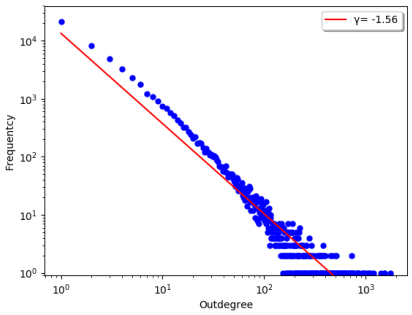
However, user interaction differs significantly over different subreddits. From the static metric observations can be made into the structure of the user interaction networks. For example, subreddits that are expected to have limited discussion, tend to have a high modularity. The temporal motif counts of the subreddit networks show differences user interaction patterns, as well over time as between different networks. Therefore, it appears users behave differently within different subreddits. The simulations of the SIR model show variation in the ability to reach large parts of the network and spread content to users with the network. This might be explained by, for instance, the difference in modularity.

The subreddits that are used in this thesis are chosen to be fairly different. To better generalise the results, many more communities should be analysed. There are many more subreddits that might have a unique network structure. Also, there are many more properties that could be examined to gain more insight into user interaction, such as different centrality measures, larger motifs and different models for information spread. It then might be possible to create a number of characteristic profiles in which the majority of subreddits on Reddit could fit.

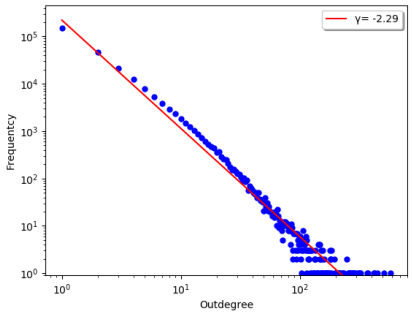
A Degree distributions



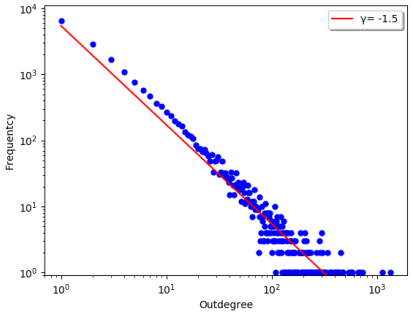
(a) AskMen



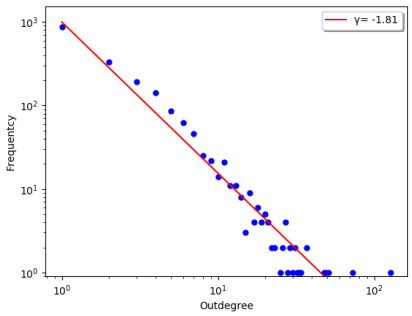
(b) AskWomen



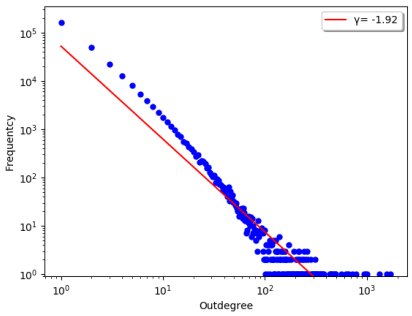
(c) aww



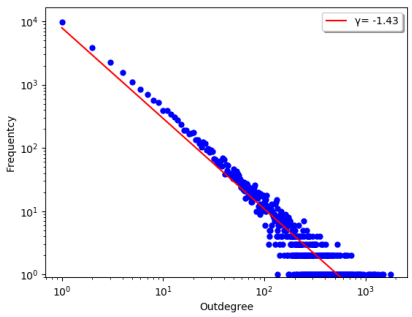
(d) totalwar



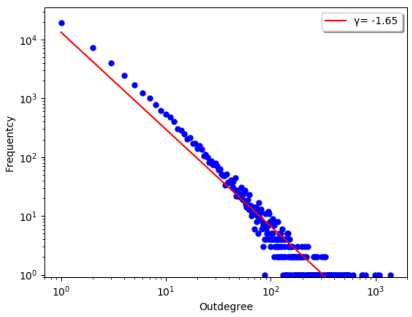
(e) AmItheAsshole



(f) IAmA

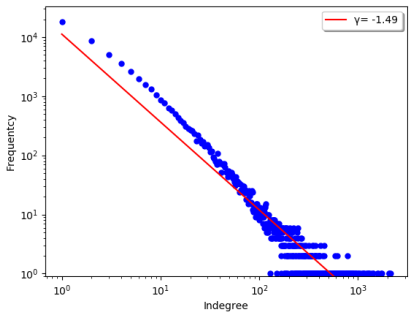


(g) formula1

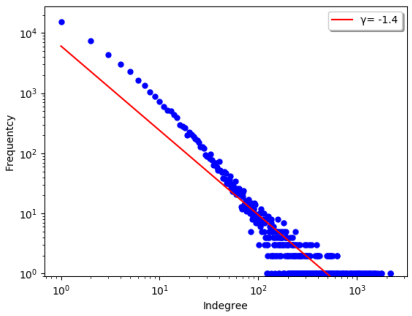


(h) programming

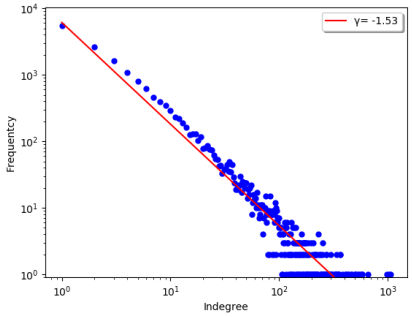
Figure 10: Outdegree distributions



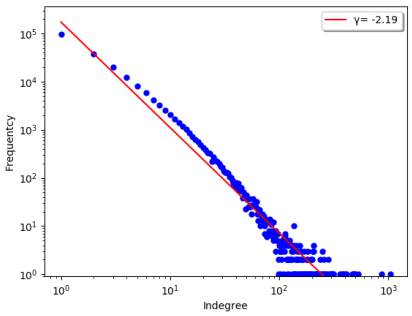
(a) AskMen



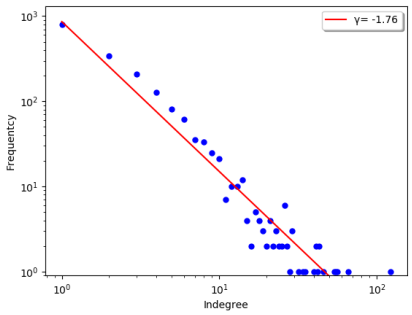
(b) AskWomen



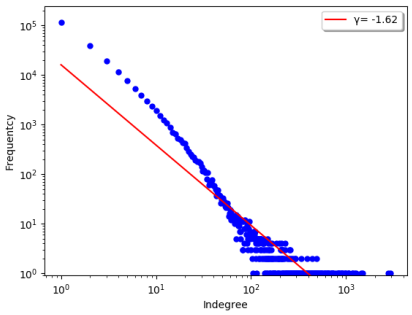
(c) totalwar



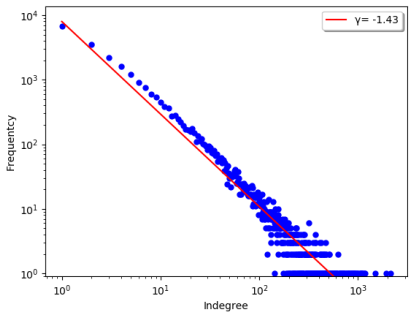
(d) aww



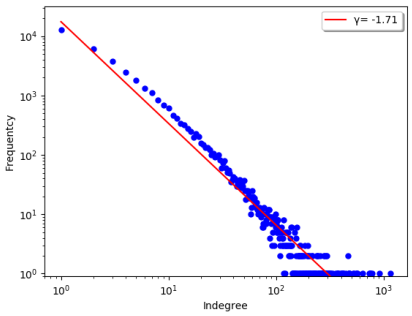
(e) AITA



(f) IAmA



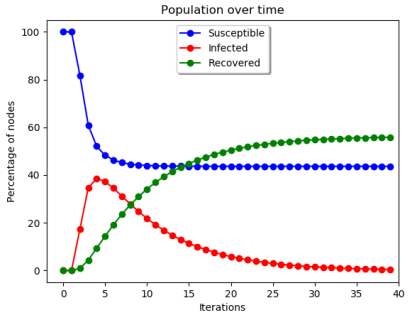
(g) formula1



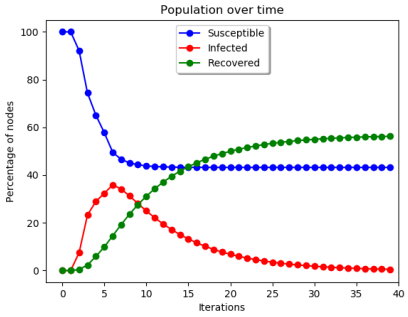
(h) programming

Figure 11: Degree distributions

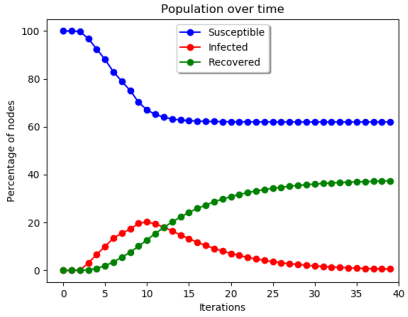
B SIR plots



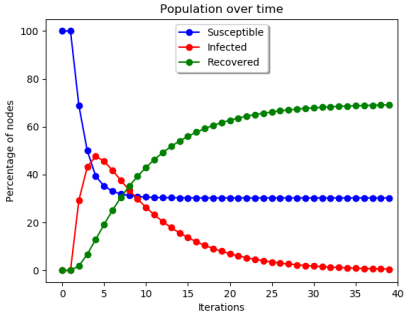
(a) AskMen



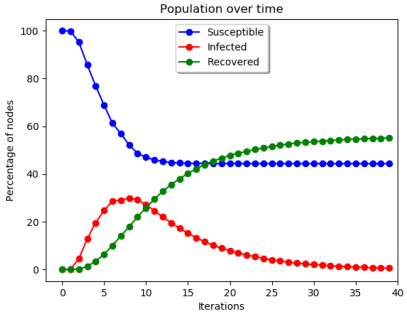
(b) AskWomen



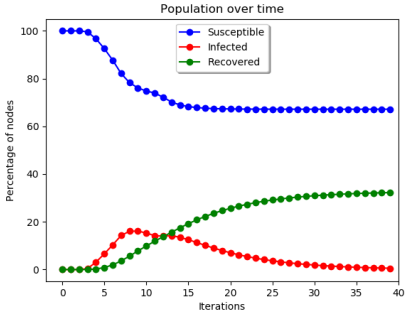
(c) aww



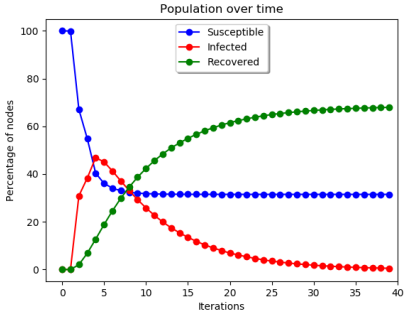
(d) totalwar



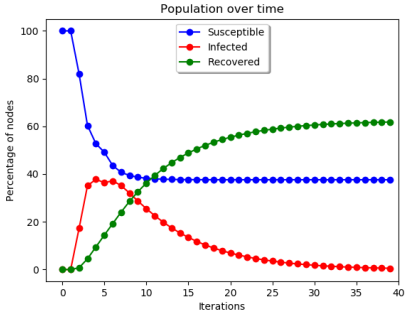
(e) AmItheAsshole



(f) IAmA



(g) formula1



(h) programming

Figure 12: SIR plots for different subreddits

References

[1] Jason Baumgartner et al. “The pushshift reddit dataset”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 830–839.

[2] Vincent D Blondel et al. “Fast unfolding of communities in large networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008.

[3] Ulrik Brandes and Christian Pich. “Centrality estimation in large networks”. In: *International Journal of Bifurcation and Chaos* 17.07 (2007), pp. 2303–2318.

[4] Cody Buntain and Jennifer Golbeck. “Identifying social roles in reddit using network structure”. In: *Proceedings of the 2018 World Wide Web Conference*. 2014, pp. 615–620.

[5] Paul Erdos and Alfred Renyi. “On the evolution of random graphs”. In: *Publ. Math. Inst. Hungar. Acad. Sci* 5 (1960), pp. 17–61.

[6] Linton C Freeman. “Centrality in social networks conceptual clarification”. In: *Social Networks* 1.3 (1978), pp. 215–239.

[7] Daniel Gruhl et al. “Information diffusion through blogspace”. In: *Proceedings of the 2018 World Wide Web Conference*. 2004, pp. 491–501.

[8] William L Hamilton et al. “Loyalty in online communities”. In: *Eleventh International AAAI Conference on Web and Social Media*. 2017.

[9] "W. Ogilvy Kermack and A. G. McKendrick". “"A contribution to the mathematical theory of epidemics"”. In: *"Proceedings of the Royal Society of London"* 115.772 (1960), pp. 700–721.

[10] Srijan Kumar et al. “Community interaction and conflict on the web”. In: *Proceedings of the 2018 World Wide Web Conference*. 2018, pp. 933–943.

[11] J. Lesckovec. *Motifs in temporal networks*. URL: `https://snap.stanford.edu/temporal-motifs/`. (accessed: 14.08.2020).

[12] Mark EJ Newman. “The structure and function of complex networks”. In: *SIAM review* 45.2 (2003), pp. 167–256.

[13] Mark EJ Newman and Michelle Girvan. “Finding and evaluating community structure in networks”. In: *Physical review E* 69.2 (2004), p. 026113.

[14] Ashwin Paranjape, Austin R Benson, and Jure Leskovec. "Motifs in temporal networks". In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 2017, pp. 601–610.